

Didaktik Deutsch

Writing with AI

Practices – Literacies – Pedagogies

Special Issue

Edited by Katrin Lehnert and Torsten Steinhoff

Contents

EDITORIAL

Katrin Lehen & Torsten Steinhoff Between ghostwriting and cowriting: AI as a transformative force across writing practices, literacies, and pedagogies.....	1
---	---

STATEMENTS

Bradley Robinson Resisting the efficiency imperative in writing education in the age of AI	8
Jin Joy Mao Writing in the GenAI era: Assessing thinking or picturing a mirage?.....	13
Cerstin Mahlow Writing versus automated text production by generative AI: Distinctions, disruptions, and the need for new models	19
Isabel Pedersen Reconceptualizing postsecondary adaption to GenAI within the turn toward digital life	29
Justin Lewis Euphoria, alienation, transformation: Vignettes of friction in AI writing	35

EMPIRICAL RESEARCH ARTICLES

Afra Sturm, Valentin Unger & Fabian Grünig Human versus machine: Comparing human and large-language-model assessments of students' writing through benchmark ratings	42
Maurice Fürstenberg Measuring the quality of AI-generated feedback: From theoretical modelling to empirical evidence.....	65
Maxi Kupetz "I did write 'huge'. Why does he say 'enormous'?" Text revision processes with an AI-based chat tool and classroom interaction in a German-as-a-Second-Language lesson....	97
Alyssa Kutzner Cross-linguistic writing with AI: Investigating transfer effects in L1/L2 and FL contexts	123

METHODOLOGICAL ARTICLES

Nina Schnatz, Aline Bieri & Cerstin Mahlow Exemplifying methodological triangulation in GenAI-supported writing research: A course-based study	158
Gloria Goller & Nadine Anskeit Empirical research on AI-assisted writing: Mixed-methods approaches, long-term effects, and ethical-technological conditions	166

Daniela Matz & Carolin Führer

Methodological approaches to research on writing processes in the context of AI:

The example of creative and literary writing 171

Editorial

Between ghostwriting and cowriting: AI as a transformative force across writing practices, literacies, and pedagogies

New media and technologies of writing have always been met with skepticism. The fear was that they would alter fundamental dimensions of human expression, cognition, subjectivity, and embodiment. It was claimed, for example, that print diminished the vitality of language use (Rousseau), the typewriter undermined the individual character of writing (Heidegger), and the computer strained the body (Auster; see Wilder & Lordick, 2026). Are widespread fears that generative AI impairs human thinking and learning, then, merely the recurrence of a media-historical figure of thought whose earlier manifestations now appear exaggerated?

No, because AI is far more disruptive. It brings to an end the human monopoly on language processing in general and text production in particular. Unlike all earlier writing media and technologies, texts are no longer merely shaped; they are generated (Robinson, 2023). The integration of AI further reduces the need for physical text production by enabling users to produce text without having to type or dictate it themselves. In this sense, AI partially assumes a function that was previously associated with ghostwriters, but in an automated and widely accessible form. Second, and more fundamentally, automation brings with it a detachment from the intellectual production of an individual person. AI thus appears less as a writing tool, such as a keyboard or word processor, than as a content-generating “participant” (Hirschauer, 2005) in writing practices, thereby calling established notions of human authorship into question. More than that, AI surpasses human writing in speed and often in quality as well. Compounding this, it is being developed with considerable rapidity, is backed by substantial financial investment, and is attracting a remarkably broad social response. Venture-capital investment in AI reached nearly USD 260 billion in 2025 alone (OECD, 2026a), and ChatGPT, at the time of writing this editorial, has more than 900 million weekly users worldwide (OpenAI, 2026).

What further distinguishes AI from other writing media and technologies is that its use invariably also raises serious ethical, political, and legal concerns including environmental costs, discriminatory harms such as sexist and racist stereotypes, concentration of power, surveillance, and issues surrounding data protection and copyright (Huang et al., 2026). These concerns arise within a field shaped by economic and political interests, including the strategic rivalry among some of the world’s largest technology companies and most powerful states whose interests are not necessarily compatible with those of education or the public sphere. Writing with AI is thus not merely a matter of writing, but also of power, control, governance, and ethical responsibility. From a didactic perspective, one question that this raises is how the use of AI in the classroom can be connected to forms of critical reflection on its output. At the same time, it remains a pedagogical challenge to operationalize the notion of critical

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 1–7

DOI: 10.21248/dideu.948

Copyright Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

thinking—a term that surfaces almost reflexively in these discussions—in ways that translate into concrete ideas and practical measures for classroom teaching.

This challenge has become more urgent, because writing with AI has already turned into an everyday practice for teachers and learners in schools and universities in many countries, as international surveys suggest (Digital Education Council, 2025; UNESCO, 2025). This uptake has outpaced regulatory oversight, institutional norm formation, curricular adaptation, and didactic reflection. As a result, writing with AI often takes place in a gray zone (“Shadow AI”) (Silic et al., 2025): widely practiced; variously tolerated, ignored, or underestimated; and only partially governed (Freeman, 2025; OECD, 2026b). This is precisely what makes the current situation so consequential. Educational institutions are confronting a new paradigm of writing that remains informal, opaque, and pedagogically underdefined. The pressing question, then, is not only under what conditions this already widespread configuration of writing can be pedagogically integrated in a meaningful, transparent, and accountable way, but also whether, and in what form, writing should remain a central educational aim at all.

What is at stake, first of all, is writing itself. Writing is a central medium of the individual’s broader literate socialization across school, university, and numerous professional domains (Graham et al., 2026). In planning, formulating, and revising texts, individuals learn to communicate across time and space and acquire extensive linguistic, worldly, disciplinary, and aesthetic resources. Second, writing supports the development of other core language practices on which education depends—above all reading, but also speaking and listening—and, in so doing, fosters the acquisition of academic language. Third, AI bears not only on writing itself, but also on the learning culture of contemporary education more broadly, including its assessment culture. Until now, writing is embedded in everyday practices of schools and universities: Students read and annotate texts, take notes during lectures, combine written language with images and video in presentations, compose source-based assignments, and articulate understanding in writing across subjects.

What happens to all of this under conditions in which writing is increasingly shaped by, or shared with AI? If the role of writing in education changes, this will reshape not only how students write, but also how they read, study, demonstrate understanding, and are assessed. Without an adequate response from education policy and didactics, it may foster forms of “cognitive surrender”—that is, the uncritical adoption of AI-generated outputs in place of independent judgment and deliberation (Shaw & Nave, 2026); and, in the longer term, it may contribute to “deskilling” by reducing opportunities to practice central cognitive and linguistic operations. Moreover, one increasingly visible effect is the unprecedented pressure placed on both students and teachers. Students report being asked with growing frequency to verify the authenticity of their work, especially when it is judged to be good. Teachers, in turn, are reading student writing with increasing doubts about authorship and heightened sensitivity to possible rule violations and attempts at deception.

In general, three scenarios seem conceivable here (Steinhoff, 2026). One is “delegation”: Writing is increasingly handed over to AI—for example, when the human activity is reduced to prompting, selecting, and editing—and gradually abandoned as an educational aim. Another is “conservation”: Writing continues to be cultivated, in conscious distinction from AI, much as before—for example, by privileging handwriting or by more tightly regulating assessments. A third is “transformation”: Writing is reworked under the conditions of AI and redefined as an educational aim accordingly, with writing with AI increasingly becoming the new default form of writing—for example, by shifting attention from finished products to writing processes, from isolated performances to portfolios, and from individual text production to more explicit forms of collaboration.

Taken together, these scenarios challenge writing research to reconsider its objects, models, methods, and instructional designs. Most fundamentally, however, the scenario of transformation requires approaching AI as a reconfiguration of writing. Its objects of inquiry thus extend beyond human writing processes and texts to include human–AI writing constellations and writing practices that emerge from them, and even forms of automated text production in which human involvement may be minimal or absent. In doing so, research should take into account not only classical forms of text production, but also more interactive writing processes such as chats with so-called “AI companions,” which, in 2025, were used regularly by 52% of 13- to 17-year-olds in the United States (Robb & Mann, 2025, pp. 2–4).

One way of conceptualizing these expanded objects of inquiry is offered by Steinhoff and Lehnen (2025) who propose a model of roles and practices: AI may act as “ghost,” “partner,” or “tutor,” while the human counterpart appears as “client,” “explorer,” or “learner.” These role constellations correspond to different communicative practices—“chat-to-generate,” “chat-to-create,” and “chat-to-chat”—which make clear that writing with AI ranges from delegated text production to quasidialogical cocreation and tutorial support. The boundaries between these roles and practices are not rigid; they overlap in various ways and may shift within a single writing process. Especially relevant in this context is how these roles shift when writing becomes collaborative and several writers must negotiate how a text should be formulated. In such constellations, roles are no longer simply distributed between human and AI, but are reconfigured through processes of attribution, coordination, and negotiation among the participants. This raises important questions about authorship, authority, and responsibility when decisions about content, wording, and structure are no longer made by a single writer. It also invites a reconsideration of collaborative writing in school—a practice with a well-established didactic tradition—under conditions in which AI enters the process not merely as a tool, but as an active contributor.

This makes it all the more crucial to recognize that assessments of the opportunities and risks of AI for human writing are often implicitly tied to the “ghost” role. Its pedagogical potential, by contrast, is more likely to emerge in constellations in which AI is configured as a “partner” or “tutor”—roles that may, in a broader sense, be subsumed under “cowriting.” Beyond this, the question arises whether and how writing with AI can be incorporated into models of competence structure and competence development, how writing with AI relates to basic writing and reading competencies, and whether AI may begin to take on quasisocializing roles in literacy development that were previously conveyed primarily by human interlocutors. This possibility seems all the more plausible given that AI use extends well beyond educational contexts and has long since become embedded in everyday practices, fulfilling a broad range of needs and functions—in some cases, even as an object of emotional attachment.

Methodologically, this calls not for a replacement of established approaches, but for their extension (Anson & Cole, 2026). Product analyses, process studies, and experimental designs remain indispensable. Accordingly, methodological pluralism is required. Established designs therefore need to be complemented by approaches capable of capturing the coactivity of human and machine more directly—for example, through screen-based process data, ethnographic observation, case studies, and mixed-method designs that treat writing as an embedded, multimodal, and interactional practice. Alongside this, research also needs to examine the professionalization of writing educators in this field, including the beliefs, orientations, and competencies that shape how AI is understood, adopted, and pedagogically integrated into writing instruction.

The reorientation of writing research also concerns the role of researchers themselves. The call that “academics need to wake up on AI” (Kustov, 2026, title) points to a broader shift: As AI increasingly contributes to literature synthesis, data analysis, interpretation, and textual formulation, it may also transform academic socialization. This is particularly evident in doctoral education, where established standards of authorship, independence, and scholarly performance may no longer be taken for granted; and where findings themselves risk becoming rapidly outdated as models, interfaces, and practices evolve.

With regard to instructional design, research needs to explore how AI can be used in ways that support learning—for example, in the roles of partner and tutor such as for teachers’ task design, feedback within secondary and tertiary students’ writing processes, or the personalized support of learners in multilingual and inclusive educational contexts. One promising option is that of “didactic agents” that extend the tradition of pedagogical agents into the context of AI (Corvacho del Toro et al., 2025). Didactic agents are AI applications that teachers develop for specific learning objectives and learner groups by supplementing, on suitable learning platforms, the system prompts of models such as ChatGPT, Claude, or Gemini with prompts of their own. In doing so, they do not merely adapt these models, but align them with specific instructional purposes and pedagogical concepts, delimit their affordances through the narrowing and shaping of possible outputs, and specify the informational basis on which those outputs are to draw.

Because AI has become a globally pervasive, conceptually challenging, empirically complex, and pedagogically consequential force in writing, it forms the subject of this journal’s first international special issue. It approaches this development at the interconnected levels of practices, literacies, and pedagogies: as a shift in writing practices; as a challenge to established literacies; and as a task for the pedagogies through which writing is taught, learned, and assessed. The issue serves both as a forum for international voices in the debate and as an opportunity for research from the German-speaking world to present itself to an international audience.

Moreover, this English-language edition marks a first attempt to experiment with new formats while combining them with more established ones. The issue brings together **invited statements** in which authors from different disciplines reflect on questions they consider especially relevant to the present situation and its likely future trajectories, **empirical research articles** that present ongoing work in the field of schooling and share initial findings, and **methodological articles** that address how writing with AI can be studied.

The issue opens with five **statements** ranging from assessment, pedagogy, and writing instruction to automated text production, digital life, and the promises and risks of writing with AI. **Bradley Robinson** emphasizes the pedagogical value of friction and resistance in writing processes. **Jin Joy Mao** argues that writing should remain a medium of learning and assessment even under AI. **Cerstin Mahlow** insists on a clearer distinction between writing and automated text production. **Isabel Pedersen** frames AI as part of students’ digital life and stresses AI and privacy literacy in postsecondary education. And **Justin Lewis** highlights both the emancipatory promise and the alienating risks of AI-supported writing.

The issue then turns to four **empirical research articles** from the German-speaking context. **Afra Sturm, Valentin Unger, and Fabian Grünig** address the question of AI-based text assessment in elementary education. Using benchmark-rated narrative and instructive texts written by students in Grades 4 and 5, the article examines whether large language models can be employed for the evaluation of writing in psychometrically meaningful ways. **Maurice Fürstenberg** examines AI-generated

feedback in writing education. Combining conceptual work on feedback quality with an empirical comparison of teacher- and AI-generated feedback on texts from 6th-grade students, the article investigates how generative AI can be described and analyzed as a source of feedback in educational contexts. **Maxi Kupetz** investigates the use of an AI-based application for text revision in a German-as-a-second-language classroom for 11th-grade students at a secondary school. Based on video data from secondary-school support lessons, the article reconstructs how revising a self-written text with AI is interactionally accomplished. And **Alyssa Kutzner** investigates AI-assisted writing in multilingual secondary education. Drawing on a mixed-methods study with 9th-grade students writing in German and English, the article explores how AI-mediated writing strategies are used across languages and how they relate to proficiency, strategy transfer, and metacognitive awareness.

A third format is provided by the **methodological articles** that reflect on how writing with AI can be adequately described, modeled, and investigated. **Nina Schnatz, Aline Bieri, and Cerstin Mahlow** argue for methodological triangulation in AI-supported writing research, combining student texts, prompting journals, surveys, and case studies. **Gloria Goller and Nadine Anskeit** address methodological challenges in empirical research on AI-assisted writing, emphasizing mixed methods, long-term effects, and transparent ethical-technological conditions. **Daniela Matz and Carolin Führer** discuss qualitative approaches to writing-process research under conditions of AI, using creative and literary writing to show how individual and collaborative processes can be reconstructed.

Taken together, the contributions to this special issue make clear that writing with AI is not simply a new tool to be integrated, but a development that forces us to rethink what writing is, what it is for, and how it should be studied, taught, and assessed.

We would like to thank all authors for their contributions to this special issue, the reviewers for their careful and constructive feedback, the editors of the journal—Ricarda Freudenberg, Friederike Kern, Susanne Riegler, and Dorothee Wieser—for their trust in and commitment to this project, Tiffany Brugger for her thoughtful editorial support, and Jonathan Harrow for kindly assisting with the translation and style. They have all helped bring this first international special issue of the journal to publication.

Katrin Lehnen/Torsten Steinhoff

References

- Anson, C. M., & Cole, K. (2026). Empirical studies of writing and generative AI: Introduction to the special issue. *Journal of Writing Research*, 17(3), 371–385. <https://doi.org/10.17239/jowr-2026.17.03.01>
- Corvacho del Toro, I., Fuhlrott, M., & Steinhoff, T. (2025). Didaktische Agenten: KI als Lehr-/Lern-partner im Deutschunterricht im Forschungsprojekt KIMADU. In M. Fürstenberg & H.-G. Müller (Eds.), *DeutschGPT – Deutschunterricht im Dialog mit künstlicher Intelligenz* (pp. 65–86). Frank & Timme. https://link.springer.com/content/pdf/10.57088/978-3-7329-8796-2_3
- Digital Education Council. (2025). *Global AI faculty survey*. https://cdn.prod.website-files.com/65f1d299b87bcc50550a6398/678a9e0aa38a44fdac1d53a8_Digital%20Education%20Council%20Global%20AI%20Faculty%20Survey%202025-1.pdf
- Freeman, J. (2025, February). *Student generative AI survey 2025* (HEPI Policy Note 61). Higher Education Policy Institute & Kortext. <https://www.hepi.ac.uk/wp-content/uploads/2025/02/HEPI-Kortext-Student-Generative-AI-Survey-2025.pdf>
- Graham, S., Skar, G. B., & Myhill, D. (2026). Why writing? *Scientific Studies of Reading*, 30(2), 95–104. <https://doi.org/10.1080/10888438.2026.2614774>
- Hirschauer, S. (2005). On doing being a stranger: The practical constitution of civil inattention. *Journal for the Theory of Social Behaviour*, 35(1), 41–67. https://www.theorie.soziologie.uni-mainz.de/files/2019/09/2005_TheSoBe_On-Doing-Being-a-Stranger.pdf
- Huang, Y., Arora, C., Huang, W. C., Kanij, T., Madugalla, A., & Grundy, J. (2026). Ethical concerns of generative AI and mitigation strategies. *Applied Soft Computing*, 193, Article 114789. <https://doi.org/10.1016/j.asoc.2026.114789>
- Kustov, A. (2026). *Academics need to wake up on AI. Ten theses for folks who haven't noticed the ground shifting under their feet*. Popular by Design. <https://www.popularbydesign.org/p/academics-need-to-wake-up-on-ai>
- OECD. (2026a). *Venture capital investments in artificial intelligence through 2025*. https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/venture-capital-investments-in-artificial-intelligence-through-2025_3bcb227f/a13752f5-en.pdf
- OECD. (2026b). *OECD digital education outlook 2026: Exploring effective uses of generative AI in education*. <https://doi.org/10.1787/062a7394-en>
- OpenAI. (2026, February 27). *Scaling AI for everyone*. <https://openai.com/index/scaling-ai-for-every-one/>
- Robinson, B. (2023). Speculative propositions for digital writing under the new autonomous model of literacy. *Postdigital Science and Education*, 5, 117–135. <https://doi.org/10.1007/s42438-022-00358-5>
- Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.common Sense Media.org/sites/default/files/research-report/talk-trust-and-trade-offs_2025_web.pdf
- Shaw, S. D., & Nave, G. (2026). *Thinking fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender* [Wharton School research paper]. https://doi.org/10.31234/osf.io/yk25n_v1
- Silic, M., Silic, D., & Kind-Trüller, K. (2025). From shadow IT to shadow AI: Threats, risks and opportunities for organizations. *Strategic Change*. Advance online publication. <https://doi.org/10.1002/jsc.2682>

Steinhoff, T. (2026). Zur Zukunft des Schreibens in Schule und Universität: Delegation, Konservation oder Transformation? In S. Ballestracci, M. Foschi, M. Lefèvre, & T. Steinhoff (Eds.), *Digitale Wissenschaftskommunikation in Deutschland, Frankreich und Italien am Beispiel der germanistischen Linguistik* (pp. 233–250). *Lingue e Letterature d'Oriente e d'Occidente*.

<http://dx.doi.org/10.36253/lea-1824-484x-10>

Steinhoff, T., & Lehnen, K. (2025). Schreiben mit Künstlicher Intelligenz: Das GPT-Modell (Ghost, Partner, Tutor). *Leseräume*, 12(11), 1–14. <https://xn--leserume-4za.de/wp-content/uploads/2025/06/Steinhoff-Lehnen-2025-LR-JG12-H11.pdf>

UNESCO. (2025). *UNESCO dedicates the International Day of Education 2025 to artificial intelligence*. <https://www.unesco.org/en/articles/unesco-dedicates-international-day-education-2025-artificial-intelligence>

Wilder, N., & Lordick, N. (2026). *Werkzeuge, Techniken und ihre Kränkungen des Menschen: Versuch einer historischen Rekonstruktion von Schreib- und Denkwerkzeugen vor dem Hintergrund des aktuellen Diskurses um künstliche Intelligenzen*. Kieler Universitätsverlag.

<https://doi.org/10.38072/2751-1359/v28>

Authors' addresses:

Katrin Lehnen, Justus-Liebig-Universität Gießen, Otto-Behaghel-Str. 10 B, 35394 Gießen, Germany

katrin.lehnen@germanistik.uni-giessen.de

Torsten Steinhoff, Universität Siegen, Hölderlinstr. 3, 57076 Siegen, Germany

steinhoff@germanistik.uni-siegen.de

Bradley Robinson

Resisting the efficiency imperative in writing education in the age of AI

“There are no limits on how many ways you can ask ChatGPT to help you during finals.” This was OpenAI’s promise in an advertisement that appeared across the social media platform Reddit during final exam season in spring 2025, echoing similar campaigns by Google’s Gemini and xAI’s Grok. Whether students are writing literary analyses, coding algorithms, or preparing lab reports, the vision remains the same: effortless productivity with “no limits.”

What strikes me about the advertisement is not just the way it targets students during a high-pressure time of the school year, but rather the phrase “no limits.” For me, this seemingly banal marketing copy signals something fairly profound about the erosion of boundaries that have long structured writing and writing education: those between teacher and student, between process and product, between thought and expression. But perhaps more troubling is how AI has enacted entirely new boundaries that we must now patrol: between human effort and machine generation, between authentic struggle and its mere appearance, between genuine articulation and algorithmic production.

1 From “not-yetness” to here and now

When I wrote about what I termed the “new autonomous model of literacy” in late 2022 (Robinson, 2023)—an emerging construct of literacy at the intersection of expression, automation, and platform capitalism—I was engaging with what Ross (2017) referred to as the “not-yetness” of emerging technologies in education, a sense that what we now call generative AI loomed on the horizon but had not yet transformed literacy practices or learning processes. ChatGPT had just launched the same month my article was published online, an article in which I attempted to undertake a sort of speculative critique of natural language processing technologies in digital writing education ahead of any uncritical integration or hypercritical rejection.

That not-yetness, as I am sure most would agree, has definitively ended. The Reddit ads targeting students during finals week represent something more than a convenient service; they underscore how a powerful efficiency imperative has been materialized within commercial AI chatbots that are actively shaping educational practices at essentially all levels of education (Hsu, 2025). The speculative has become actual.

This shift reflects what I see as a broader tech-powered efficiency imperative underwriting so much of contemporary social life, education included. Langdon Winner (1977) conceptualized this phenomenon as *reverse adaptation*, or the process by which “people come to accept the norms and standards

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 8–12

DOI: 10.21248/dideu.951

Copyright Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

of technical processes as central to their lives as a whole” (p. 229). Winner argued that when the lives of individuals become entangled with technological systems, “efficiency, speed, precise measurement, rationality, productivity, and technical improvement become ends in themselves applied obsessively to areas of life in which they would have previously been rejected as inappropriate” (p. 229). Is this not precisely what we see happening today with respect to AI platforms and education—writing education included? The efficiency imperative represents the systematic prioritizing of speed and effort reduction as an organizing principle that shapes our educational practices, technological choices, and understanding of what it means to think and write.

What makes this especially concerning, in my view at least, is how this technical logic has spread beyond its original domain. As Winner observed, efficiency, or “the quest for maximum output per unit” (p. 229), becomes “a universal maxim for all intelligent conduct” (p. 229), rather than remaining appropriately confined to the technical systems where it belongs. In other words, unlike business and industry, where optimizing for speed and output volume drives profit margins, education has traditionally valued the process of learning itself—including its inherent struggles, iterations, and reflective pauses—as much as, if not more than, the final product. But the efficiency imperative now embedded in the design of AI chatbot platforms such as ChatGPT, Claude, and Gemini encourages what Winner described as a “subtle but comprehensive alteration” (p. 229) in how we think about learning and writing. Students and educators find themselves increasingly adapting their goals and practices to match the technical capabilities of AI systems—and are even admonished that failing to do so will imperil their future economic well-being (see Singer, 2025)—rather than demanding that these systems serve educational ends.

2 The efficiency imperative in action

Winner’s (1977) process of reverse adaptation helps explain the fundamental shift in compositional practices we are currently witnessing, and which I have observed in my own teaching. The tension between efficiency and meaningful struggle plays differently across disciplines—such as when and how to use a calculator in mathematics—but it seems particularly acute in writing in which the messy work of articulation is often where the most significant learning happens (e.g., Watts et al., 2023). When students describe, as mine have, typing up their initial ideas in a disorganized manner and using AI platforms to “clean it up”—that is, to convert their exploratory, emergent ideas into language that appears coherent and academic—they are describing a crucial shift, or in Winner’s words, a reverse adaptation, in compositional practices that has serious implications for the practice of writing.

This shift effectively alters the relation between thought and expression. That messy middle where ideas take shape, where writers struggle to articulate their thoughts and, in doing so, bring them into coherence, is precisely the space being optimized away in the name of efficiency. Students are adapting their writing processes to match what AI systems do well rather than preserving the struggle that builds their own capacities. What we risk losing is not just writing as a skill or practice, but rather the possibility for the kind of thinking and intellectual growth that is uniquely nurtured through sustained engagement with difficulty.

3 Word factories versus word gyms

Winner’s (1977) analysis of reverse adaptation also helps explain why the design choices embedded in AI writing platforms matter so much for education. These design choices shape whether we treat

writing as mere production or as intellectual development, a distinction Booten (2024) captures in his distinction between “word factories” and “word gyms.” This metaphor clarifies a critical difference in how we might approach AI writing systems in educational contexts, and it offers a helpful framework for resisting the reverse adaptation that Winner described.

The difference is critical: Factories optimize for output, gyms for capacity. In Booten’s (2024) formulation, the word factory model situates AI writing platforms as industrial assembly lines for language, optimizing for speed, volume, and standardization of output, with writing becoming a form of production rather than expression. Consider the user interface for ChatGPT, for example, everything about which reinforces this factory logic: the minimalist design, the “Ask Anything” prompt, and especially the follow-up suggestions that nudge users toward delegating more of their intellectual work to the machine, prompting them with options such as “Would you like for me to draft an essay based on this outline?” What makes Booten’s factory metaphor so apt is how it captures the industrial, essentially Fordist logic underwriting the design of many AI writing platforms. Just as factories break down complex processes into simplified, repeatable steps that can be performed with minimal skill, AI writing platforms break down the complex processes of composition into simplified outputs that can be generated with minimal thought.

Booten’s (2024) word gym metaphor, by contrast, is less concerned with efficient output than it is with development through resistance. The weights one lifts in a gym are not obstacles to be eliminated, but challenges to be embraced precisely because of the resistance they offer. Similarly, the challenges of writing—finding the right words, restructuring a clumsy sentence, clarifying a complex argument—are not inefficiencies to be eliminated. They are the means by which we become stronger writers and thinkers. The word gym model, then, helps us view writing as a practice that should develop rather than simply generate, which, in turn, allows us to better appreciate why AI-powered writing platforms present such a challenge. They offer a form of optimization that, as Winner (1977) explained, encourages us to adapt our ends to match the available means rather than the reverse. And students may become dependent on technological assistance, not through conscious choice, but through the subtle pressures of systems designed for efficiency above else, a form of what Fourcade and Johns (2020) have described as “dependency by design” (p. 821).

The question students and educators now face in the age of AI feels existential: Is struggle valuable, or is it simply inefficient? The answer may determine how we think just as much as how we write. A recent Microsoft-funded study found that higher confidence in generative AI correlates with less critical thinking, whereas higher self-confidence correlates with more critical thinking (Lee et al., 2025). In other words, the more people trust AI to do their thinking for them, especially when they do not feel confident in their own abilities, the less they actually think when using it. In the context of writing education, this creates the possibility of a concerning feedback loop: experienced writers with high self-confidence tend to be more critical of AI assistance and therefore exert more control over their writing, whereas less confident writers may be more likely to defer to the authority they invest in AI, despite its tendency to produce writing that, as Tressie McMillan Cottom (2025) put it, could best be described as “mid.”

4 Toward resistive pedagogy

In 2022, I proposed what I called a resistive digital writing pedagogy, an approach that acknowledges how deeply the efficiency imperative reaches into our educational institutions while working to create

spaces in which other values can flourish (Robinson, 2023). At the time, this was largely theoretical, but ongoing developments suggest we need to move beyond such gestures toward concrete practices that actively resist the force of reverse adaptation in writing education. In my view, however, a resistive writing pedagogy needs to be about more than simply pressuring educators to design better assignments, as if individual pedagogical heroics were the answer.

Instead, a resistive pedagogy must recognize that we are all—students, instructors, administrators, and so on—caught within an efficiency-obsessed society that, as Winner (1977) described, “tends to arrange all situations of choice, judgment, or decision in such a way that only instrumental concerns have any true impact” (p. 232). In other words, institutional discussions about AI in education, at least in my experience, tend to focus on questions of implementation strategies, learning outcomes, and competitive advantage rather than on deeper questions about the kind of thinking and learning we want to preserve and why, or why we teach writing in the first place. Of course, recognizing this does not necessarily commit us to rejection or acceptance of AI in writing education broadly, but rather suggests that we need to actively resist how these systems, and the marketing campaigns that accompany them, encourage us to frame educational choices in their often purely instrumental terms.

At its core, then, a resistive pedagogy is one that insists that institutional policies, technological choices, and pedagogical practices work to preserve meaningful struggle where and when it matters. What might this look like in practice? It could involve assignments that ask students to document their thinking processes and not just their final products, showing how ideas evolved through struggle rather than emerged fully formed (Giles, 2010). It could involve students in critically analyzing AI-generated synthetic text alongside human writing to examine what patterns and values each one embodies and what different modes of thinking each represents (Levine et al., 2025). It might also involve institutional policies that center the value of process alongside product (Conference on College Composition and Communication, 2022), alongside institutional demands that AI writing platforms be designed for development rather than dependency—word gyms, not word factories. And it must involve classroom practices in which AI becomes a critical object of inquiry rather than a neutral tool, those that enable students to analyze how these systems work, what values they encode, and how they shape writing and thinking (Burris & Leander, 2024).

5 Conclusion

Recall the ChatGPT advertisement I began with: “There are no limits on how many ways you can ask ChatGPT to help you during finals.” Thinking with Winner’s (1977) reverse adaptation, we can now see this promise for what it is: not just a convenience service offered to students but the efficiency imperative in action, positioning struggle as wasteful and efficiency as the ultimate goal. Yet what matters most in education—and indeed in life—often emerges from the spaces where we resist this imperative, where we choose to engage with difficulty rather than avoid it. The value of meaningful struggle, however, cannot be erased entirely. Students will continue experiencing moments of genuine insight when wrestling with difficult ideas, and, as writers, they will still discover what they think through the often awkward, laborious act of articulation. But such moments of intellectual growth will only persist if we actively work to preserve the conditions that make them possible.

The future of writing education will be determined by the values we choose to prioritize, not by technological change alone. In his work, Winner showed us that we have choices to make in how we relate to technology, that we can resist simply adapting ourselves to whatever systems corporations design

for us. By collectively reclaiming meaningful struggle in our lives, classrooms, schools, and universities, we can work toward a future in which technology serves human flourishing rather than merely human convenience.

References

- Booten, K. (2024). Build word gyms, not word factories. *Critical AI*, 2(1).
- Burriss, S. K., & Leander, K. (2024). Critical posthumanist literacy: Building theory for reading, writing, and living ethically with everyday artificial intelligence. *Reading Research Quarterly*, 59(4), 560–569.
- Conference on College Composition and Communication. (2022, April). *Writing assessment: A position statement*. National Council of Teachers of English. <https://cccc.ncte.org/cccc/resources/positions/writingassessment>
- Fourcade, M., & Johns, F. (2020). Loops, ladders and links: The recursivity of social and machine learning. *Theory and Society*, 49(5), 803–832.
- Giles, S. L. (2010). Reflective writing and the revision process: What were you thinking? In C. Lowe & P. Zemliansky (Eds.), *Writing spaces 1: Readings on writing* (pp. 191–204). Parlor Press.
- Hsu, H. (2025, June 30). What happens after A.I. destroys college writing? *The New Yorker*. <https://www.newyorker.com/magazine/2025/07/07/the-end-of-the-english-paper>
- Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025, April). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–22).
- Levine, S., Beck, S. W., Mah, C., Phalen, L., & Pittman, J. (2025). How do students use ChatGPT as a writing support? *Journal of Adolescent & Adult Literacy*, 68(5), 445–457.
- McMillan Cottom, T. (2025, March 29). The tech fantasy that powers A.I. is running on fumes. *The New York Times*. <https://www.nytimes.com/2025/03/29/opinion/ai-tech-innovation.html>
- Robinson, B. (2023). Speculative propositions for digital writing under the new autonomous model of literacy. *Postdigital Science and Education*, 5(1), 117–135.
- Ross, J. (2017). Speculative method in digital education research. *Learning, Media and Technology*, 42(2), 214–229.
- Singer, N. (2025, June 7). Welcome to campus. Here's your ChatGPT. *The New York Times*. <https://www.nytimes.com/2025/06/07/technology/chatgpt-openai-colleges.html>
- Watts, F. M., Dood, A. J., Shultz, G. V., & Rodriguez, J. M. G. (2023). Comparing student and generative artificial intelligence chatbot responses to organic chemistry writing-to-learn assignments. *Journal of Chemical Education*, 100(10), 3806–3817.
- Winner, L. (1977). *Autonomous technology: Technics-out-of-control as a theme in political thought*. MIT Press.

Author's address:

Bradley Robinson, Texas State University, 301 Woods St., San Marcos, Texas 78666, USA
bradrobinson@txstate.edu

Jin Joy Mao

Writing in the GenAI era: Assessing thinking or picturing a mirage?

Since the arrival of ChatGPT in November 2022, numerous resources have explored the opportunities and challenges of generative artificial intelligence (GenAI) tools such as ChatGPT, Gemini, Midjourney, DALL-E, GitHub Copilot, Synthesia, and ElevenLabs, to name a few. They allow users to generate text, code, images, and audio and video content within seconds. Among the challenges discussed regarding GenAI tools, major concerns include faculty fears that students will use large language models (LLMs) and numerous AI-powered homework services to plagiarize or cheat on assignments and exams, and the faculty's inability to define appropriate or inappropriate uses of artificial intelligence (AI) in student work. These concerns emerged shortly after ChatGPT was released (e.g., Cotton et al., 2023; Mao et al., 2024), and they have remained the same since then (e.g., An et al., 2025; Kovari, 2025). As educators and researchers quickly found out that AI detection tools are ineffective in detecting plagiarism, and as more students and researchers started using such tools, concerns intensified over plagiarizing AI-authored content, violating academic integrity, and disrupting research standards (e.g., Balalle & Pannilage, 2025; Malik et al., 2024).

This statement paper continues the academic discourse on these concerns and follows the call of this international special issue to focus specifically on how the arrival of GenAI tools changes perceptions of writing as an academic endeavor. One central question that prompts this statement about the issues with ChatGPT and other GenAI tools is: Will these tools lead to the end of our abilities to write and think? In this statement paper, I shall focus on this central question by reflecting on the role of writing in educational assessment, particularly as a powerful and widely used tool for evaluating and fostering higher-order thinking and learning. This reflection begins with the history of writing. It then considers various functions of writing in human thinking, education, human knowledge, and intelligence; AI's disruptive impact; and, finally, what we should do to prepare for writing in the AI era.

1 Writing as human thinking

Writing is human thinking, and it demonstrates and records the evolution of human thinking. Historically, written work has been an important way to convey meaning and record knowledge, tradition, and culture since the very beginning of language development. Writing is "humankind's principal technology for collecting, manipulating, storing, retrieving, communicating and disseminating information" and may have been invented independently in the Near East, China, and Mesoamerica (Schmandt-Besserat, 2014, p. 1). Borrowed by many other cultures, the unique Greek invention of the alphabet in

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 13–18

DOI: 10.21248/dideu.953

Copyright Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

western culture is the most representative stage in the history of writing systems following picture writing, word-based writing, and sound-based syllabic writing (Olson, 2025). Regardless of cultural differences, writing was a privilege reserved for the rich and powerful in ancient history, and it has remained one of the most important literacy skills in contemporary history.

Britannica describes writing as a skill that must be taught to children and to every generation, unlike spoken or signed language as a basic human characteristic and competence that can be acquired without systematic instruction (Olson, 2025). However, writing is more than just a mechanical skill that must be taught as a literacy skill. According to Ong (1992), it is a technology for thinking and a tool to structure cognition, and it is an indispensable part of the oral and the literate mind because: “Without writing, the literate mind would not and could not think as it does, not only engaged in writing but even when it is composing its thoughts in oral form” (p. 293).

2 Writing as a major learning outcome

Writing has been considered a significant learning outcome in formal education, although it existed long before formal schooling was established in history. Since the 1950s, Bloom’s taxonomy has served as the foundation for classifying cognitive processes and higher-thinking skills. Given the apparent challenge of assessing thinking directly, writing has been considered a significant learning outcome and a frequently used assessment technique for measuring higher-order thinking skills at all levels of education. Writing across the curriculum is a representative movement that demonstrates the sweeping influence and widely recognized importance of writing in education, not only as an assessment technique but also a means of learning. Regardless of educational context, writing in various formats has been widely used to demonstrate learning and performance status, assess higher-order thinking and learning outcomes, and communicate formative or summative feedback on student work in both traditional face-to-face classrooms and blended or online environments.

3 Writing as human knowledge and intelligence

Writing is fundamental to human knowledge and intelligence, and therefore, it is widely used in educational assessment to demonstrate higher-order thinking and learning skills. The definitions of knowledge vary by different theoretical foundations such as behaviorism, cognitivism, constructivism, and connectivism. According to Wilson and Myers (2000), “knowing, learning, and cognition are social constructions, expressed in actions of people interacting within communities” (p. 59). In line with this social and cultural perspective on knowledge within situated cognition, writing is one of the actions that capture and demonstrate human interactions with the communities and with the environment.

To summarize, writing has contributed to the accumulation of human knowledge and collective intelligence since its invention. Writing is essential to any meaning-making processes and products, within the broader social contexts that both shape human experience and are shaped by human activity. Writing is an indispensable part of human knowledge, intelligence, culture, and civilization.

4 AI and its disruptive impact

Unlike other learning technologies that enter education gradually, the abrupt, high adoption of GenAI tools has had a disruptive, immense impact on education. Serious concerns emerged about GenAI-created content, whether for writing or scientific research, due to a lack of human authenticity and

creativity. Consequently, the loss of trust in human-to-human interactions and the crisis of authenticity have added immediate pressure to the tensions already present in educational tests and assessments (Mao et al., 2024). Can writing still be a trustworthy assessment technique with widespread, but untrained and unintended, GenAI adoption among students? Should research and publications devoid of human authenticity continue to be reliable sources and indicators of human thinking, intelligence, and productivity?

Increasingly, the unconstrained and unethical use of GenAI-created content cycled from limited training data will lead to the vicious cycle of a falsified, fabricated, erroneous knowledge base, completely deprived of the trust expected in the human knowledge system and intelligence. By that time, writing will no longer be human thinking, and can no longer be fundamental to human knowledge or intelligence. Nevertheless, there are still optimistic expectations for positive changes in education, with the assumption that AI and GenAI tools are developing rapidly and that such issues may disappear with the upcoming versions.

5 Writing with GenAI as a new tool

Inevitably, AI and GenAI are developing and advancing rapidly, and their pervasive, disruptive impact will continue to appear in education. It is impossible to ban the use of AI and GenAI tools in educational contexts or to limit how students or researchers use them, due to the challenges of defining and providing guidelines for appropriate and inappropriate use in the first place. The history of educational technology has shown that banning new technologies (e.g., calculators to cell phones) in schools has never resolved the issues they raise.

The AAC&U report (Watson & Rainie, 2025) finds that most colleges and universities are unprepared for the challenges posed by AI and GenAI, and that graduates are not adequately prepared to work in the AI-flooded job market. Thus, one pressing need is to provide opportunities to improve AI literacy and to acquire the relevant expertise and skills. Besides the practical needs of workforce readiness, writing educators require the teaching of GenAI tools specifically for writing. Wang and Tian (2025) pointed out that AI is a technology just like other tools for writing that must be taught to students because of “the inherently social and political nature of both writing and technology that compels us, as writing educators, to engage students with the reality of an increasingly AI-mediated world” (p. 2).

Undoubtedly, the writing process can be more efficient with accelerated search and research processes. In this regard, the transition to using GenAI tools for writing is similar to the transition from using a traditional library and searching through hard copies of books to using digital library databases for electronic searching using keywords. The speed of searching and summarizing with AI and prompt engineering should save writers much time in the research process, requiring the ability to write effective prompts and to critically evaluate the results generated by AI tools. It is ultimately the users’ or writers’ responsibility to determine how to assign meaning to the search results and to use them with critical thought relevant to human experience, the purpose of the writing, and the context for the writing. Reflecting on the role of writing in assessment and for the human knowledge system, I would summarize the following recommendations for transitioning to the use of writing with GenAI.

6 Develop ethical guidelines

Many colleges and universities have developed AI and GenAI policies for teaching, learning, assessment, and research. Educators have started including class policy language in their syllabi and incor-

porating GenAI into assignments. In addition to such policies at different levels, developing detailed ethical guidelines to help students understand and learn responsible use of AI and GenAI tools should be a crucial first step in the transition and learning process. Researchers and authors could also benefit from the clear ethical guidelines from academic journals and professional organizations to delineate the fine line between acceptable and unacceptable uses of AI tools or AI-powered services in the final writing products.

7 Restore and rebuild trust

In his interview with the *Practical Wisdom* podcast, Chomsky named using AI tools as high-tech plagiarism. The use of GenAI tools has destroyed trust between educators and students, raising concerns about educators' reliance on writing as an assessment approach and, conversely, about students' lack of appreciation for the feedback they receive on their work. Using AI or GenAI as real tools rather than a substitute for human effort can help restore and rebuild trust in human-to-human interactions. To achieve this, we need to develop a mindset for trust, a willingness to trust, and the ability to trust through learning and improving the overall AI literacy for everyone involved.

8 Improve AI and GenAI literacy

It is imperative to improve AI and GenAI literacy and learn how to use them as real tools, not replacement agents. With the necessary guidelines and trust in place, everyone involved in the use of AI and GenAI tools should have access to professional learning and practice opportunities to expand AI literacy through first-hand experiences. AI literacy should include not only knowledge of AI in education and skills in using AI tools, but also the soft skills needed for writing and researching with AI tools. Such soft skills should include the ability to reflect on one's writing process and the metacognitive skills that enable one to be self-aware of the process together with the integration of AI tools.

9 Make the process human-centered

It is crucial to intentionally make writing processes human-centered, especially given the argument developed earlier in this paper: Writing is human thinking and is fundamental to the human knowledge system and intelligence. Educators need to develop creative pedagogical approaches to teaching students, especially young students, to take human agency in the writing process, and to engage themselves as active inquiry initiators in writing. Taking an active agency role will help students cultivate beneficial habits and mindsets that make writing processes human-centered from a young age. Additionally, learning and implementing design thinking may be another way to improve human agency in writing.

10 Redefine standards and terms

It is necessary to redefine academic standards and terminologies such as plagiarism, academic dishonesty and integrity, authorship, copyright, ethics, and many others. The redefinition may vary by institutional values, policies, subject areas, or learning outcomes. Undoubtedly, the redefinition requires updated AI literacy, and the definition of certain terms may affect others. For example, a lack of AI literacy can lead to confusing definitions of ethical or unethical uses of AI tools; the meaning of authorship may be affected by how ethics are defined; and the understanding of ethics can decide the scope of misuse.

11 Support writing with GenAI tools

Appropriate support for writing with GenAI tools is essential to help educators and students adapt to the changes. Besides the above recommendations to serve as foundational support, which require institutional involvement, updating assessment techniques and culture in order to focus more on processes rather than products is a direct measure to prepare both educators and students for the changed writing culture. Banning students from using GenAI tools for writing and replacing it with requiring them to write by hand within a set time and location instead is an ineffective and counter-productive approach. Furthermore, it is critical to encourage writing *about* AI, writing *with* AI tools for the process, and reflecting on both.

12 Conclusion

Given its unique importance to human thinking and knowledge systems, writing should continue to be a way of learning and an assessment technique for developing higher-order thinking skills. Essentially, it is more important than ever to learn how to write with new tools and to write within the fast-changing larger environment for writing. Both educators and students are faced with the challenges and opportunities as they learn to write with GenAI tools. Educators will need to intentionally teach students to use AI tools ethically, responsibly, and effectively for writing. Improved human agency and trust are key factors that can help students avoid the unethical, inappropriate use of AI tools for writing or work right from the start. Further, without meaning-making and authentic human experience, sharing content generated by AI or machines in the name of writing is just like picturing a mirage, leading to false, unattainable outcomes. That will make us lose not only the trust in a longstanding assessment technique, but also our ability to think, know, experience, and create as human beings.

References

- An, Y., Yu, J. H., & James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education*, 22(10), Article 10. <https://doi.org/10.1186/s41239-025-00507-3>
- Balalle, H., & Pannilage, S. (2025). Reassessing academic integrity in the age of AI: A systematic literature review on AI and academic integrity. *Social Sciences & Humanities Open*, 11, Article 101299. <https://doi.org/10.1016/j.ssaho.2025.101299>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Kovari, A. (2025). Ethical use of ChatGPT in education—Best practices to combat AI-induced plagiarism. *Frontiers in Education*, 9, Article 1465703. <https://doi.org/10.3389/feduc.2024.1465703>
- Malik, M. A., Amjad, A. I., Aslam, S., & Fakhrou, A. (2024). Global insights: ChatGPT's influence on academic and research writing, creativity, and plagiarism policies. *Frontiers in Research Metrics and Analytics*, 9, Article 1486832. <https://doi.org/10.3389/frma.2024.1486832>
- Mao, J., Chen, B., & Liu, J. C. (2024). Generative artificial intelligence in education and its implications for assessment. *TechTrends*, 68, 58–66. <https://doi.org/10.1007/s11528-023-00911-4>
- Olson, D. R. (2025, September 5). *History of writing systems*. Britannica. <https://www.britannica.com/topic/writing/History-of-writing-systems>

- Ong, W. J. (1992). Writing is a technology that restructures thought. In P. A. Downing, S. D. Lima, & M. Noonan (Eds.), *The linguistics of literacy* (pp. 293–320). John Benjamins Publishing Company.
- Schmandt-Besserat, D. (2014). The evolution of writing. In J. Wright (Ed.), *International encyclopedia of social and behavioral sciences* (pp. 1–15). Elsevier. <https://sites.utexas.edu/dsb/tokens/the-evolution-of-writing/>
- Wang, C., & Tian, Z. (2025). Introduction. In C. Wang & Z. Tian (Eds.), *Rethinking writing education in the age of generative AI* (pp. x–xii). Routledge.
- Watson, C. E., & Rainie, L. (2025). *Leading through disruption: Higher education executives assess AI's impact on teaching and learning*. AAC&U. <https://www.aacu.org/research/leading-through-disruption>
- Wilson, B. G., & Myers, K. M. (2000). Situated cognition in theoretical and practical context. In D. H. Jonassen & S. M. Land (Eds.), *Theoretical foundations of learning environments* (pp. 57–88). Erlbaum.

Author's address:

Jin Joy Mao, Dr. Richard Abbas Alley School of Education, College of Health and Education, Wilkes University, 84 W. South Street, Wilkes-Barre, PA 18766, USA
jinjoy.mao@wilkes.edu

Cerstin Mahlow

Writing versus automated text production by generative AI: Distinctions, disruptions, and the need for new models

1 Introduction

In written societies, literacy (reading and writing) is essential for participation and has always been intertwined with technology. In the 1990s, the Internet moved reading and writing into the digital space. Now, large language models (LLMs) and generative AI (GenAI) are transforming literacy once more.

AI-based language technology offers unprecedented opportunities for inclusion and writing support such as automatic plain language translation (Ebling et al., 2022) and written language acquisition (Mahlow, 2023a). Today's computing power, mature natural language processing (NLP) technology, and LLMs make it feasible to realize innovative ideas for adaptive human-centered writing support (Mahlow, 2023b) and adaptive writing tools (Mahlow & Piotrowski, 2009). However, we lack an understanding of *how* GenAI fundamentally alters writing processes. This paper first explicates the context and defines text generation and automated text production in contrast to writing. This leads to the proposal of rigorous revisions of writing models as part of an extensive research agenda to reflect the integration of GenAI and its cognitive impact.

2 Context: The bigger picture

With the appearance of ChatGPT in November 2022, the general public was confronted with impressive developments in the field of language technology and artificial intelligence. Additionally, various professional fields are being called into question.

The embedding of GPT3.5 in a simple chat interface has led to astonished as well as enthusiastic reactions. The sometimes euphoric assessments that anyone can now create professional-looking texts by a simple request to a chatbot have shaken professional fields such as communication, technical writing, copywriting, marketing, and translation: If these automatically produced texts, as Mahlow and Dale (2014) call them, are as good as those of professional translators and copywriters, what does this mean for the entire industry? And downstream, the question: What does such an upheaval in fields of practice mean for corresponding courses of study and further education?

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 19–28

DOI: 10.21248/dideu.956

Copyright Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

There was a similar upheaval in the field of translation and interpreting a few years before: the power of automatic machine translation based on LLMs created by neural learning (Bahdanau et al., 2014) and its general availability in the mid-2010s in online services such as Google Translate (Wu et al., 2016) and DeepL posed enormous challenges for the industry itself (cf. Bouillon et al., 2018) as well as for training concepts and institutions (cf. Cotelli Kureth et al., 2023). In retrospect, the adaptation to this has been successful—well-trained professional translators and interpreters continue to be used for many areas.

Both chatbots and language models have been research topics in human–machine communication and computational linguistics for quite a long time. Chatbots based on GenAI are not restricted in terms of language—both for interaction and for the desired text—and text type. However, upon closer examination, it turns out that both the interaction with such chatbots and the quality of the texts produced depend on the training material for the language models, and this is mostly not disclosed. In addition, users do not interact directly with the language models, but by means of a chat interface that filters and expands both input and output.

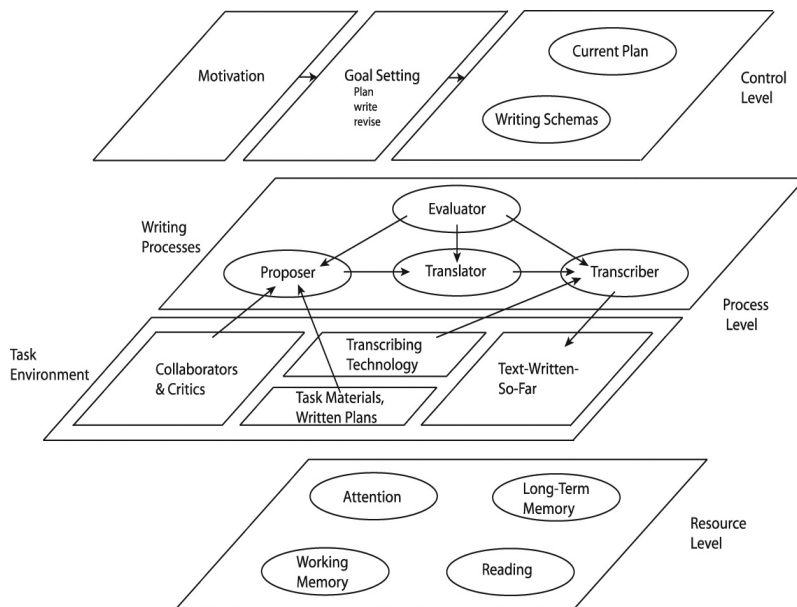
Since ChatGPT’s release in late 2022, we have witnessed the impact of GenAI on language professions. In fields such as communication, copywriting, and translation, GenAI challenges established roles and processes. Mastering these tools and using them successfully requires more than “prompt engineering”—which is a misnomer because no proper engineering is involved. It demands evaluating the quality of AI-produced texts and using them responsibly in professional contexts. GenAI for text production is also reshaping how we teach and practice writing (cf. Anson et al., 2025).

Current studies on GenAI in writing focus on stand-alone tools (see, e.g., Bedington et al., 2024; Cummings et al., 2024; Tarchi et al., 2024), but the trend is moving toward integrating GenAI directly into writing tools. To make best use of it, we need to understand how automated text production by GenAI works. If we are to help writers interact with GenAI efficiently and effectively, the writing research community must understand what GenAI can and cannot do. As I argued in my SIG Writing 2024 keynote (*On the road to language-aware writing support—are we there yet with large language models?*), the scientific community currently lacks a comprehensive writing model to account for GenAI.

I am specifically interested in understanding writing on a linguistic level (Mahlow et al., 2024) and work on respective models (Ulasik et al., 2025) based on keystroke-logging experiments. Referring to Hayes’ comprehensive writing model (Hayes, 2012; see Figure 1), our studies focus on the interplay of Transcriber (process level), Transcribing Technology and Text-Written-So-Far (task environment) for drafting and revising.

Figure 1

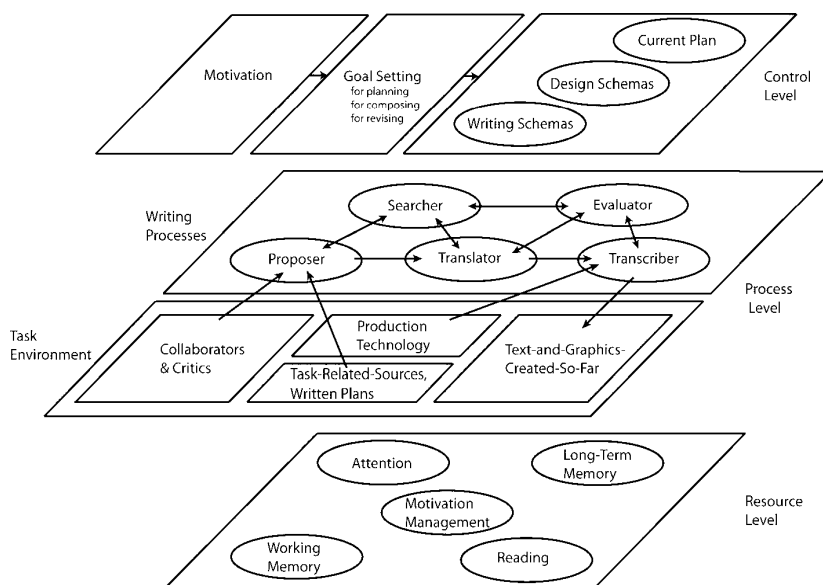
Hayes' 2012 Model of the Writing Process (Hayes, 2012)



Writing models aim to capture how people actually write. However, these models sometimes fail (Bazerman, 2018), and need to be extended or adjusted. For example, Leijten et al. (2014) extend Hayes' model to account for skilled professional communication (see Figure 2), explicitly including collaboration and nontextual creation resulting in adjustments on all levels (Schrivver, 2013). As GenAI changes the way people write, clearly writing models need revisions.

Figure 2

Extended Model by Leijten et al. (2015) to Cover Professional Communication



Additionally, writing models should be implementable as running programs (Alamargot & Chanquoy, 2001; Hayes, 2012) following the practice of creating running programs to simulate human actions in cognitive science (Newell & Simon, 1956). They are considered to be an implementation of a theory (Weizenbaum, 1976) that serve as a basis for designing tools to support writing (Cookson, 1989). So far, we, as the writing research community, have not succeeded in this matter—probably because we did not tackle this task seriously, except for a few attempts to implement some aspects of children writing in Python by Hayes (2012).

3 Automated text production versus natural language generation versus writing

To understand the impact on writing, we need to answer the question: When LLMs produce text, what is that, really? The output of applications using LLM is often called “generated text,” using a term that already has a specific meaning in the field of natural language generation (NLG). NLG is the production of “understandable texts in . . . human languages from some underlying non-linguistic representation of information” (Reiter & Dale, 1997, p. 57) and has been an active research field for several decades. In analogy to text-to-speech, it has also been named data-to-text (Schneider et al., 2022).

Data (e.g., measure points, facts in knowledge bases, entire databases), as nonlinguistic representation of information, is the explicit context of the text to be generated and the most important aspect of NLG systems. This context is made available in structured form to the machine that generates the text. Additionally, all information about intended audience, genre, text length, and so forth has to be made explicit as well, and is used as features and parameters for generation. Usually, these systems involve human curation or are (at least partially) based on human-generated rules. Dale (2020) gives an overview of the (commercial) state of the art on NLG; Schneider et al. (2022) describe categories of NLG systems. They emphasize that:

data-to-text systems in real-world applications still require such a share of human configuration and control and the creative contribution share of the software . . . is still so limited that it would not be adequate to claim creative autonomy of the software in the process. (p. 1)

Current language technology such as NLG is thus not an autonomous creative agent taking part in the writing process. There is no *interaction* between a human writer and a text-producing machine, only very sophisticated *interactivity* (cf. Rafaeli, 1988).

Language technology is used for writing support in settings of *automated text production* as defined by Mahlow and Dale (2014) with the deliberate emphasis that, in these cases, it is not the author that is the writer, but the machine. They see NLG as one aspect of automated text production and also mention summarization or machine translation: A new text is produced without human intervention, based on existing text or explicit prompts.

In a revision of the work by Mahlow and Dale (2014), I suggest to distinguish *text generation* from *automated text production*:

1. *Text generation* covers NLG. It is characterized by relying on nonlinguistic representations of information and involving some kind of text planning. Planning concerns the text as a whole—length, main information, structure, and so forth—as well as single sentences.
2. *Automated text production* is the output of current LLMs as fast continuous predictive texting—known from input-support on mobile phones (Ganslandt et al., 2009)—by design. No facts and information determine the next words or sentences, but only available co-texts of already existing words and sentences—when writing from left to right, there is always only

left co-text. Note that what computer science and NLP call “context” as used in “size of context window” to describe the power of an LLM is, in linguistics, called “co-text,” whereas “context” refers to everything outside the text.

Such systems are operated by humans through communication using unrestricted natural language. There is no planning of either the text as a whole or individual sentences; and, more importantly, no revision and editing take place. The text is produced as a plausible continuation of already existing text. Hence the similarity to predictive texting. The resulting text is correct from a linguistic point of view: it does not contain spelling and grammar issues; it is most probably coherent and consistent on the sentence level. But it is plausible and acceptable only when focusing on the *language* produced. Automated text production is linear or actually *incremental* both in time *and* in space: One token after the other is added to the edge of the text produced so far (TPSF).

Writing, in contrast, can be characterized as involving planning—the degree can of course vary—and more importantly: Writing is linear in time, but never linear in space. Writers can and do go back to text already written (TPSF) at any point in time and delete or add something; they reorder text parts for various reasons—style, change in plans, and so forth. They consider left *and* right co-text as well as context including the writing task and their own planning. None of these movements will be visible in the final product.

Dale and Viethen (2021) provide an overview on writing assistance based on state-of-the-art NLP resources (i.e., language models) and approaches. In 2021, GPT-3 was already available and had been incorporated into various tools aimed at supporting writers by co-authoring and not only (copy-)editing. These applications addressed specific genres such as blog posts and poetry, as well as specific writing tasks such as expanding, rewriting, and shortening texts (Dale & Viethen, 2021). However, they were not used widely and did not trigger the same discussions that we see now.

The more powerful LLMs get, the more users forget that texts automatically produced by any application using these resources are just plausible extensions of existing co-text. These co-texts are either existing (parts of) texts, which are then extended, or prompts the system seems to “react” to: “GPT-3 is still very capable of generating nonsense, but on the whole it’s more plausible nonsense; and with appropriate fine tuning and prompting, the texts it generates can be eerily convincing” (Dale & Viethen, 2021, p. 516).

All information that seems to be included in the text is merely due to frequent co-occurrences of words and thus not trustworthy (cf. Kalai et al., 2025). No references to any context (i.e., the world outside the text) can be made; no conclusions on underlying knowledge, understanding, or intention of an assumed author can be drawn. Depending on the genre the LLM is asked to produce, a text may contain what looks like facts (e.g., dates, places, or even correctly formatted bibliographical references). But again, these are only *plausible*, they are not true; they could, but most often do not, exist. The LLM just produces sequences of words that follow general patterns for bibliographical entries. While references produced by an LLM are thus most probably nonexistent, one can very confidently use them to check whether the *format* of a list of bibliography entries adheres to a certain citation style.

These systems are not creative: They do not invent anything, because they have no agency. They just react to any input—be it a prompt from a human or previously produced text. All creativity, all surprise is only in the mind of human readers interpreting these texts, ignoring authorship, and not being aware of the circumstances (i.e., the process) of their production.

4 A research agenda

Writing research as a scientific community needs to closely observe and carefully analyze current writing and text production practices. Writing models were always intended to explain what happens during writing. With new tools and new affordances come new practices currently not covered by existing models. As a field, we need to work on respective topics and most probably need to develop new approaches.

4.1 Research topics

I pose four research topics that need to be addressed through interdisciplinary research:

1. Seeing GenAI as co-author, as Armitage (2024) and Kleebayoon and Wiwanitkit (2023) do, requires the extension of collaborative writing models to include nonhuman authors.
2. GenAI intervenes at all stages of the writing process and on all levels of existing models. No single generic GenAI writing tool exists, but an ever-growing number of GenAI-based tools tailored to specific uses (or GenAI-based features integrated into existing tools for specific purposes such as Grammarly or Antidote) are being developed and used that allow elements in writing models to be linked to distinct GenAI features.
3. GenAI-based automated text production as part of writing fundamentally disrupts writing as a practice as well as a concept, requiring new models to account for its role in creative tasks and cognitive load.
4. The adequate and deliberate use of GenAI shifts focus to higher-level tasks such as planning. Which would be a good thing, if true.

The revision of writing models will generate further research questions like those addressing Research Topic 4 for psycholinguistic experiments. First observations suggest that GenAI could be used as Transcriber. In Hayes' (2012) model, Transcription is the actual production of visible text by handwriting or typing, and it involves spelling (Berninger & Winn, 2006). Higher-level activities such as planning require a significant amount of cognitive capacity. The more writers automate lower-level tasks (e.g., master touch-typing), the more they can focus on planning, and so forth (Hayes & Chenoweth, 2006; Kellogg, 1999; Olive, 2022; Olive et al., 2009; Quinlan et al., 2012), because limited working memory must be distributed to different tasks (Olive, 2014).

4.2 Advance methodology

Writing research always incorporated methods and approaches from different fields and is truly interdisciplinary. To address the research topics presented above, writing research needs to look at current developments in NLP, human–computer interaction, and the neurosciences (cf. Zock, 2020). First studies looking at not only the efficacy of integrating GenAI into writing courses by triangulation of product and process data but also research methods are already under way (e.g., Schnatz & Bieri, 2025).

For Research Topic 4 specifically, there is a need to adapt methods from Bouriga and Olive (2021) and Liu et al. (2024) for controlled experiments focusing on producing running text from outlines by calling specific GenAI features in editors with integrated GenAI features. Established methodologies are insufficient. A new approach is needed to capture how GenAI reshapes cognitive load and higher-order tasks such as planning. Collaboration with psycholinguistics is needed to adapt double and triple task paradigms (cf. Olive et al., 2002). GenAI's integration into writing involves dynamic and unpredictable

interactions, emphasizing the need to develop experimental protocols that account for these nuances, including how GenAI influences the distribution of cognitive resources during the writing process.

4.3 What to expect

The proposed research ideas and topics add significant value to writing research and writing instruction as well as to society at large. By exploring how GenAI can enhance cognitive aspects of writing, they address key issues at the intersection of AI, writing processes, and education.

On a societal level, this research agenda addresses the pressing issue of (digital) literacy in a world soon to be imbued with AI—one might even argue that this has already happened. Writing remains an essential skill involving cognitive processes, but it is undergoing radical shifts due to new GenAI-based text production tools and the integration of GenAI into existing writing technology. This research agenda will provide insights into how these shifts can be harnessed. It will directly inform teaching practices by proposing improved methods for teaching writing (both learning-to-write and writing-to-learn).

GenAI-based tools or GenAI-driven features in writing tools promise to improve writing efficiency and efficacy: There is a clear need to validate and operationalize these claims by transferring them into practice. Methodological development of AI-assisted writing models will contribute to writing research as a scientific field and to frameworks for digital education that transform how writing is taught in universities and professional settings.

5 Conclusion

The integration of generative AI (GenAI) into writing represents a paradigm shift that calls for a reevaluation of writing models, pedagogical approaches, and digital literacy. This paper tries to explicate the roles of language technology in text production and proposes a research agenda addressing this shift through empirical studies and interdisciplinary research to develop new writing models that account for GenAI as collaborator and cognitive scaffold.

References

- Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing: With commentaries by Ronald T. Kellogg & John R. Hayes* (Vol. 9). Springer. <http://www.worldcat.org/isbn/0792371593>
- Anson, C. M., Delorme Benites, A., Ganobcsik-Williams, L., Mahlow, C., & McKee, H. A. (2025). Reflections on writing and generative AI: A symposium. *Journal of Academic Writing*, 15(S2), 1–14. <https://doi.org/10.18552/joaw.v15iS2.1121>
- Armitage, R. (2024). Generative AI in medical writing: Co-author or tool? *British Journal of General Practice*, 74(740), 126–127. <https://doi.org/10.3399/bjgp24X736605>
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv*. <https://doi.org/10.48550/arXiv.1409.0473>
- Bazerman, C. (2018). What does a model model? And for whom? *Educational Psychologist*, 53(4), 301–318. <https://doi.org/10.1080/00461520.2018.1496022>
- Bedington, A., Halcomb, E. F., McKee, H. A., Sargent, T., & Smith, A. (2024). Writing with generative AI and human–machine teaming: Insights and recommendations from faculty and students. *Computers and Composition*, 71, Article 102833. <https://doi.org/10.1016/j.compcom.2024.102833>

- Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In *Handbook of writing research* (pp. 96–114). The Guilford Press.
- Bouillon, P., Girletti, S., Estrella, P., Mutal, J., Bellodi, M., & Bircher, B. (2018). Integrating MT at Swiss Post's language service: Preliminary results. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 281–286). <http://hdl.handle.net/10045/76086>
- Bouriga, S., & Olive, T. (2021). Is typewriting more resources-demanding than handwriting in undergraduate students? *Reading and Writing*, 34(9), 2227–2255. <https://doi.org/10.1007/s11145-021-10137-6>
- Cookson, S. (1989). Designing computational writing tools within a linguistic model of the writing process. In N. Williams & P. O. Holt (Eds.), *Computers and writing: Models and tools* (pp. 17–21). Intellect.
- Cotelli Kureth, S., Delormes Benites, A., Haller, M., Noghrechi, H., & Steele, E. (2023). “I looked it up in DeepL”: Machine translation and digital tools in the language classroom. In *Human translation and natural language processing towards a new consensus?* Fondazione Università Ca' Foscari. <https://doi.org/10.30687/978-88-6969-762-3/006>
- Cummings, R. E., Monroe, S. M., & Watkins, M. (2024). Generative AI in first-year writing: An early analysis of affordances, limitations, and a framework for the future. *Computers and Composition*, 71, Article 102827. <https://doi.org/10.1016/j.compcom.2024.102827>
- Dale, R. (2020). Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, 26(4), 481–487. <https://doi.org/10.1017/S135132492000025X>
- Dale, R., & Viethen, J. (2021). The automated writing assistance landscape in 2021. *Natural Language Engineering*, 27(4), 511–518. <https://doi.org/10.1017/S1351324921000164>
- Ebling, S., Battisti, A., Kostrzewa, M., Pfütze, D., Rios, A., Säuberli, A., & Spring, N. (2022). Automatic text simplification for German. *Frontiers in Communication*, 7, Article 706718. <https://doi.org/10.3389/fcomm.2022.706718>
- Ganslandt, S., Jörwall, J., & Nugues, P. (2009). Predictive text entry using syntax and semantics. *Proceedings of the 11th International Conference on Parsing Technologies*, 37–48.
- Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3), 369–388. <https://doi.org/10.1177/0741088312451260>
- Hayes, J. R., & Chenoweth, N. A. (2006). Is working memory involved in the transcribing and editing of texts? *Written Communication*, 23(2), 135–149. <https://doi.org/10.1177/0741088306286283>
- Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. <https://arxiv.org/abs/2509.04664>
- Kellogg, R. T. (1999). Components of working memory in text production. In M. Torrance & G. Jeffery (Eds.), *The cognitive demands of writing: Processing capacity and working memory effects in text production* (Vol. 3, pp. 42–61). Amsterdam University Press.
- Kleebayoon, A., & Wiwanitkit, V. (2023). Use of ChatGPT as a co-author. *ESMO Open*, 8(4). <https://doi.org/10.1016/j.esmoop.2023.101620>
- Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. R. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research*, 5(3), 285–337. <https://doi.org/10.17239/jowr-2014.05.03.3>

- Leijten, M., Van Waes, L., & Horenbeeck, E. V. (2015). Analyzing writing process data: A linguistic perspective. In G. Cislaru (Ed.), *Writing(s) at the crossroads: The process/product interface* (pp. 277–302). John Benjamins Publishing Company. <https://doi.org/10.1075/z.194.14lei>
- Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, 211, Article 104977. <https://doi.org/10.1016/j.compedu.2023.104977>
- Mahlow, C. (2023a). Large language models and artificial intelligence as tools for teaching and learning writing. *Osnabrücker Beiträge zur Sprachtheorie*, 101, 175–196. <https://doi.org/10.17192/OBST.2023.101.8607>
- Mahlow, C. (2023b). Writing tools: Looking back to look ahead. <https://arxiv.org/abs/2303.17894>
- Mahlow, C., & Dale, R. (2014). Production media: Writing as using tools in media convergent environments. In E.-M. Jakobs & D. Perrin (Eds.), *Handbook of writing and text production* (Vol. 10, pp. 209–230). De Gruyter Mouton.
- Mahlow, C., & Piotrowski, M. (2009). LingURed: Language-aware editing functions based on NLP resources. *Proceedings of the International Multiconference on Computer Science and Information Technology*, 4, 243–250. <https://annals-csis.org/proceedings/2009/PTI/pliks/101.pdf>
- Mahlow, C., Ulasik, M. A., & Tuggener, D. (2024). Extraction of transforming sequences and sentence histories from writing process data: A first step towards linguistic modeling of writing. *Reading and Writing*, 37, 443–482. <https://doi.org/10.1007/s11145-021-10234-6>
- Newell, A., & Simon, H. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory*, 2(3), 61–79. <https://doi.org/10.1109/TIT.1956.1056797>
- Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research*, 6(2), 173–194. <https://doi.org/10.17239/jowr-2014.06.02.4>
- Olive, T. (2022). How do novice and skilled writers engage working memory? In J. W. Schwieter & J. Cheng (Eds.), *The Cambridge handbook of working memory and language* (pp. 504–526). Cambridge University Press. <https://doi.org/10.1017/9781108955638.028>
- Olive, T., Alves, R. A., & Castro, S. L. (2009). Cognitive processes in writing during pause and execution periods. *European Journal of Cognitive Psychology*, 21(5), 758–785. <https://doi.org/10.1080/09541440802079850>
- Olive, T., Kellogg, R. T., & Piolat, A. (2002). The triple task technique for studying the process of writing. In T. Olive & C. M. Levy (Eds.), *Contemporary tools and techniques for studying writing* (Vol. 10, pp. 31–59). Kluwer.
- Quinlan, T., Loncke, M., Leijten, M., & Van Waes, L. (2012). Coordinating the cognitive processes of writing. *Written Communication*, 29(3), 345–368. <https://doi.org/10.1177/0741088312451112>
- Rafaeli, S. (1988). Interactivity: From new media to communication. *Advancing Communication Science: Sage Annual Review of Communication Research*, 16, 110–134.
- Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1), 57–87. <https://doi.org/10.1017/S1351324997001502>
- Schnatz, N. I., & Bieri, A. (2025). Refining an AI-integrated scientific writing course for engineers: Insights from a design-based research approach. *EDULEARN25 Proceedings*, 2394–2406. <https://doi.org/10.21125/edulearn.2025.0663>

- Schneider, A., Madsack, A., Heininger, J., Chen, C.-Y., & Weißgraeber, R. (2022). Data-to-text systems as writing environment. *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, 1–10. <https://doi.org/10.18653/v1/2022.in2writing-1.1>
- Schrivier, K. (2013). What do technical communicators need to know about information design? In J. Johnson-Eilola & S. Selber (Eds.), *Solving problems in technical communication* (pp. 386–427). University of Chicago Press.
- Tarchi, C., Zappoli, A., Casado Ledesma, L., & Brante, E. W. (2024). The use of ChatGPT in source-based writing tasks. *International Journal of Artificial Intelligence in Education*, 35, 858–878. <https://doi.org/10.1007/s40593-024-00413-1>
- Ulasik, M. A., Mahlow, C., & Piotrowski, M. (2025). Sentence-centric modeling of the writing process. *Journal of Writing Research*, 16(3), 463–498. <https://doi.org/10.17239/jowr-2025.16.03.05>
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman & Co.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., . . . Dean, J. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv*. <https://doi.org/10.48550/arXiv.1609.08144>
- Zock, M. (2020). AI at the crossroads of NLP and neurosciences. *Journal of Cognitive Science*, 21(1), 1–14.

Author’s address:

Cerstin Mahlow, ZHAW Zurich University of Applied Sciences, School of Applied Linguistics, Theaterstraße 17a, 8401 Winterthur, Switzerland
cerstin.mahlow@zhaw.ch

Isabel Pedersen

Reconceptualizing postsecondary adaption to GenAI within the turn toward digital life

1 Introduction

Given its capacity to personalize writing practices, generative AI (GenAI) can be a productive tool for postsecondary student writers. Moreover, the conversational nature of some of the interfaces such as AI chatbots makes students feel as if they are personally interacting with a writing service meant for them: an AI as a personal editor or AI tutor. Yet, the use of GenAI apps in learning environments is still debated in higher education. On the one hand, human–AI teaming is lauded for its benefits. One article indicates that OpenAI’s ChatGPT has a “positive impact on improving learning performance” although with several caveats, including that “ChatGPT should be flexibly integrated into teaching as an intelligent tutor, learning partner, and educational tool” (Wang & Fan, 2025, para. 1). Conversely, research findings on GenAI’s negative implications for higher education are myriad: “Concerns around GenAI have included academic integrity and diminished development of evergreen skills like critical thinking, problem solving, communication and creativity” (Matsui, 2025, p. 8). Deeply concerning are “confusing, and harmful GenAI outputs on social media, turning AI safety into a public concern” due to risk of bias and discrimination (Singh et al., 2025, p. 6). Researchers have suggested that postsecondary educators ought to introduce novel literacy practices in their curricula to ensure ethical adoption (or rejection) of GenAI for writing (Duin & Pedersen, 2021; McKee & Porter, 2020). One route to that recommendation is to respond to GenAI as a personal technology that is increasingly geared to “digital-life” experiences.

2 Definitions

AI literacy

AI literacy is defined “as a set of competencies that enable individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace” (Long & Magerko, 2020, p. 598). The emphasis on competencies is important. AI literacy is about what one can *do* as well as what one knows. It is becoming apparent that educators are moving beyond a strict binary of including or excluding GenAI in curricula; and the importance of AI literacy has come to the fore in higher education. Education scholars frame AI literacy as a means to encourage students “to critically examine ethics, bias, privacy, inclusion, data collection and explainability in popular AIED [AI in Education] tools” (Wilton et al., 2022, p. 183). From these criteria,

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 29–34

DOI: 10.21248/dideu.957

Copyright Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

AI literacy affords students the agency to evaluate AI tools personally “at home” and at “work” (Long & Magerko, 2020, p. 598) as well as while using AI EdTech tools “in the classroom” (Wilton et al., 2022, p. 182). Another criterion by Woods and Johnson (2024) grounds GenAI usage in “*privacy literacy* as an essential component of reconceptualizing our digital lives” (p. 87). They argue that “moving from a binary of ‘literate = protected and private’ and ‘illiterate = unprotected and public’ toward multidimensional layered literacies is essential for developing critical privacy literacy in the age of genAI” (p. 88).

Digital life

GenAI is also changing how people live. A digital-life theoretical perspective looks more closely at the individual, the *flip side* of the conversation. One often privileges either AI’s hype and technological feats or the organizations desperately seeking governance models to deal with the chaos of GenAI adoption. Instead, one ought to recognize the many ways that digital technologies become inter-meshed in people’s lives, that is, take an ontological perspective. “Digital life fronts individuals’ lived experiences rather than the capabilities or possibilities of technologies” (Stambler, Veeramoothoo, & Davis, 2024, p. 7). AI industries are increasingly encroaching on people to adopt self-centric AI tools promising personalized, human–AI assistant interaction for everyday tasks (Pedersen, 2025a). Yet, the adoption for *personal* uses is sometimes ignored in higher education. Postsecondary students are adapting to using personal, noneducational GenAI tools. They use these tools embedded in platforms to automate the composition of personal content to make their lives easier. It helps them craft résumés, generate videos for social media posts, plan personal events, and share them with others. They use GenAI to explain all kinds of practical things that confuse them in life such as renting a car for the first time, signing a lease, or accepting a job offer. If Google search is a textbook, GenAI is not only an AI tutor, but also a one-on-one life assistant. Moreover, GenAI chatbot technologies behave as personal interlocutors and companions (Farmer, Smakman & Birtwistle, 2025). The point is not to marginalize this cohort from their everyday-life uses of GenAI. The point is to acknowledge that they and other actors in postsecondary spheres, namely professors, instructors, staff members, administrators and adjacent professionals, are being normalized to its adoption as well.

This short article goes into greater detail on these themes and makes four arguments to support the call to reconceptualize GenAI adoption in higher education using a sociotechnical, “digital-life” point of view.

3 Reconceptualizing postsecondary adaption to GenAI

First, the growth of GenAI technology adoption is exponential; and, thus, using these tools is becoming normalized. As of August 2025, ChatGPT has 800 million active weekly users, doubling from 400 million in February 2025 (Singh, 2025). According to a survey involving 16 countries, the majority of postsecondary students (86%) said they use artificial intelligence in their studies (Kelly, 2024). AI writing tools are becoming more accurate and better at assisting student writers. This deployment is dynamic and AI development is rapidly entering its second wave (Schwarcz et al., 2025). Retrieval-augmented-generation (RAG) models tailored to specific content are “a technique for enhancing the accuracy and reliability of generative AI models with information fetched from specific and relevant data sources” (Merritt, 2025, para. 9). For example, research supports the claim that law students using RAG models in commercial AI apps are benefiting from them. Results “led to substantial, statistically significant improvements in the speed with which participants completed legal tasks”

(Schwarcz et al., 2025, p. 7). The point here is not to assume that the speed and accuracy of these tools necessarily improve learning outcomes for all postsecondary students, nor that educators should adopt them blindly. The point is that first-wave GenAI tools and the original large language models (LLMs) that students use now are advancing. AI literacy competencies need to accommodate constant, dynamic, and substantial technological change.

Second, AI-literacy approaches need to reframe GenAI as an aspect of society and personal life, as introduced above. With the mass adoption of GenAI apps in nonacademic spheres, *adaptation* to AI assistance is undergoing societal assimilation (Cardillo, 2025). People are using it for communication in their everyday social lives, exchanging highly personal information with chatbots (Elon University, 2025). Work-based practices are also evolving to accept it. “Fifty percent of U.S. CEOs say their companies have already automated content creation with it, according to a 2024 Deloitte survey” (Tools for why CEOs, 2025, p. 24). A simplified “sociotechnical perspective means viewing society and technology together as one coherent system. In other words, it is not possible to understand the ‘social’ without the ‘technical,’ nor the ‘technical’ without the ‘social.’” (Chen & Metcalf, 2024, para. 3). AI tools in education require assessment as an aspect of society and personal life.

Consumer GenAI writing technologies function in the social ecosystems of postsecondary students, instructors, and administrators. Digital-life researchers have identified that “digital literacy is not just about using tools; it also attends to the socio-technical environments in which technical communicators perform their work,” which amounts to their life worlds (Stambler et al., 2024, p. 220). A digital-life perspective supports a dialogue on privacy “for actively protecting one’s private data” beyond the time students spend in school (Woods & Johnson, 2024, p. 87). Reconceptualizing AI writing tools in higher education means accepting the process of their enculturation occurring in the lived realities of those who use it.

Third, universities are integrating good guidelines for sanctioned usage of AI tools, but they are not making room for the way users, across postsecondary, incorporate AI platforms in their lifeworlds. For example, Harvard University integrates a good range of AI tools including OpenAI’s ChatGPT Edu, Google Gemini, and Adobe Firefly in its “AI for General Use” category with access to them controlled by faculty or unit. Harvard University (2025a) defines general use as the:

category of tools [that] includes “chatbots” and AI assistants for general use and productivity. They are designed to understand and generate human-like responses to text-based, natural language prompts. They can generate text, code, and images, translate languages, write different kinds of creative content, or integrate with productivity and collaboration tools. (para. 4)

It also discusses “security and privacy protections that ensure the tools are appropriate for use with certain types of confidential data, and assurances that the data entered will not be used to train vendor models” (Harvard University, 2025b, para. 8). Harvard University offers an exemplary policy framework for a GenAI toolset usage to meet expectations for first-wave GenAI.

When GPT technologies first deployed in 2022, people went to websites to use them, and they still do. However, AI content-generating chatbots are on students’ phones, in earbuds, and they are “working” on very personal devices. Wearable AI is defined as a type of computing consisting of AI technologies embedded in wearable computer devices such as AI glasses. Postsecondary students are early adopters of current wearables (Pedersen, 2025b). Problematically, Harvard’s policy treats an AI toolset as if a university can control its usage, or that *its university contracted platforms* will be the *only* ones used by students. It does not acknowledge the integrated ways that people use these tools regularly for other purposes, for communicating in myriad ways.

Fourth, the final and most important point is that students need to be included as co-designers and critical evaluators in every act of AI adoption to achieve ethical reframing. Education is not the only sector calling for a reevaluation of AI adoption. The Data & Society Research Institute & AI Risk and Vulnerability Alliance express it well:

we must reframe the relationship between AI and society from adversarial to co-constitutive. By co-constitutive framing, we mean that AI is always already embedded within society. AI emerges from social practices, responds to social cues, mediates social practices, and is built from historically shaped, socially produced knowledge and data. This framing draws our attention to how AI produces and mediates particular social relations; how existing social relations shape certain AI applications and evaluation strategies; and how AI harms are experienced in everyday life. (Singh et al., 2025, p. 59)

The paradigm shift of AI's positioning from adversarial to co-constitutive is not a simple task, nor should it be. However, this article is arguing that all actors who might experience harm ought to be included in evaluation strategies to determine whether it should *or should not* be used. Along with evaluation, disclosure of tools' usage should also always be incorporated for student, instructor, or administrative use (Weaver, 2024).

There are many ways to involve students in evaluation strategies as a facet of AI literacy. Getting them to validate facts and sources from outputs to determine the feasibility of using AI tools or to reject them empowers them. One more structured method is red teaming, whereby teams "use adversarial methods to identify vulnerabilities in systems" (Singh et al., 2025, p. 4). Whereas red teaming is controversial, "practitioners and policymakers alike have pointed to *red-teaming* as an integral part of their strategies to identify and address related risks, with the goal of ensuring some notion of alignment with human and societal values" (Feffer et al., 2024, p. 421). Proposed by UNESCO, "public red teaming" is one that invites participation from "everyday users who interact with AI in their daily lives" (Chowdhury et al., 2025, p. 14). The crux of this point is to include students in decisions about GenAI tools, involving them in determining inaccuracy or harm based on their interactions with instructors and their own training.

4 Conclusion

Although difficult to prove, it might well be that postsecondary students are using GenAI to compose, write, represent meaning, and narrate their lives more than ever before. One can speculate that writing practices are changing as a result at home, in school, and in work scenarios. Along these lines, this article argues that GenAI tools impact not only writing in specific academic contexts, necessitating AI literacies, but that they are also changing students' lives. Reconceptualizing GenAI adoption in higher education will help students adapt to it from a "digital-life" point of view for future contexts as writers across a multitude of genres.

References

- Cardillo, A. (2025, May 1). 40+ chatbot statistics (2025). *Exploding Topics*.
<https://explodingtopics.com/blog/chatbot-statistics#chatbot-statistics-by-region>
- Chen, B. J., & Metcalf, J. (2024, May 28). *A sociotechnical approach to AI policy*. Data & Society Research Institute. <https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/>

- Chowdhury, R., Skeadas, T., Lakshmi, D., & Amos, S. (2025). Red teaming artificial intelligence for social good: The playbook. In D. Cliche & S. Andrews (Eds.), *UNESCO Division for Gender Equality & Humane Intelligence*.
- Duin, A. H., & Pedersen, I. (2021). *Writing futures: Collaborative, algorithmic, autonomous*. Springer.
- Elon University. (2025). Survey: 52% of U.S. adults now use AI large language models like ChatGPT. Today at Elon. <https://www.elon.edu/u/news/2025/03/12/survey-52-of-u-s-adults-now-use-ai-large-language-models-like-chatgpt/>
- Farmer, H., Smakman, J., & Birtwistle, M. (2025, November 11). *The dilemmas of delegation*. Adalovelaceinstitute.org. <https://www.adalovelaceinstitute.org/report/dilemmas-of-delegation/>
- Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C. & Heidari, H. (2025). Red-teaming for generative AI: Silver bullet or security theater? *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 421–437). AAAI Press. <https://dl.acm.org/doi/10.5555/3716662.3716698>
- Harvard University. (2025a). Generative AI tool comparison. *Harvard University*. <https://www.huit.harvard.edu/ai/tools>
- Harvard University. (2025b). Guidelines for the use of generative AI tools at Harvard. *Harvard University*. <https://www.huit.harvard.edu/ai/guidelines>
- Kelly, R. (2024, August 28). Survey: 86% of students already use AI in their studies. *Campus Technology*. <https://campustechnology.com/articles/2024/08/28/survey-86-of-students-already-use-ai-in-their-studies.aspx?utm%20source=chatgpt.com>
- Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376727>
- Matsui, K. (2025, June 9-13). *AI in Education: Mapping the Present, Exploring the Future Conference*, University of Guelph, Guelph, ON, Canada. <https://otl.uoguelph.ca/facultyinstructor-programming/2025-teaching-ai-conference>
- McKee, H. A., & Porter, J. E. (2020). Ethics for AI writing: The importance of rhetorical context. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)* (pp. 110–116). Association for Computing Machinery. <https://doi.org/10.1145/3375627.3375811>
- Merritt, R. (2025, January 31). What is retrieval-augmented generation, aka RAG? *NVIDIA Blogs*. <https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/>
- Pedersen, I. (2025a). Want an advanced AI assistant? Prepare for them to be all up in your business. *The Conversation*. <https://doi.org/10.64628/aam.ca36vuf95>
- Pedersen, I. (2025b). AI agents, wearable computing and the future of postsecondary learning. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1651453>
- Schwarcz, D., Manning, S., Barry, P. J., Cleveland, D. R., Prescott, J. J., & Rich, B. (2025, March 2). AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice. *Minnesota Legal Studies Research Paper No. 25-16*; University of Michigan Public Law Research Paper No. 24-058. SSRN. <https://doi.org/10.2139/ssrn.5162111>
- Singh, R., Blili-Hamelin, B., Anderson, C., Tafesse, E., Vecchione, B., Duckles, B., & Metcalf, J. (2025, February). *Red-teaming in the public interest*. Data & Society Research Institute & AI Risk and Vulnerability Alliance. <http://doi.org/10.69985/VVGP4368>
- Singh, S. (2025, August 14). *ChatGPT statistics (2025) – Daily & monthly active users*. DemandSage. <https://www.demandsage.com/chatgpt-statistics/?utm%20source=chatgpt.com>

- Stambler, D. M., Ranade, N., Hocutt, D. L., Fonash, S., Campbell, J. L., Duin, A. H., Pedersen, I., Tham, J., Veeramoothoo, S., & Verhulsdonck, G. (2024). Beyond digital literacy: Investigating threshold concepts to foster engagement with digital life in technical communication pedagogy. *Technical Communication Quarterly*, 34(2), 218–238. <https://doi.org/10.1080/10572252.2024.2388038>
- Stambler, D. M., Veeramoothoo, S. (Chakrika), & Davis, K. (2024). Toward digital life: Embracing, complicating, and reconceptualizing digital literacy in communication design. *Communication Design Quarterly*, 12(2), 5–10. <https://doi.org/10.1145/3655727.3655728>
- Tools for why CEOs should think twice before using AI to write messages. (2025, May/June). *Harvard Business Review*, 103(3), 23–27.
- Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12, Article 621. <https://doi.org/10.1057/s41599-025-04787-y>
- Weaver, K. D. (2024). The artificial intelligence disclosure (AID) framework: An introduction. *C&RL News*, 85(10), 407–411. <https://doi.org/10.48550/arXiv.2408.01904>
- Wilton, L., Ip, S., Sharma, M., & Fan, F. (2022). Where is the AI? AI literacy for educators. *23rd International Conference, AIED Proceedings 2022*, Durham, UK, July 27–31, 2022, 180–188. https://doi.org/10.1007/978-3-031-11647-6_31
- Woods, C., & Johnson, G. (2024). (Re)designing privacy literacy in the age of generative AI. *Communication Design Quarterly*, 12(2), 86–97. <https://dl.acm.org/doi/10.1145/3655727.3655736>

Author's address:

Isabel Pedersen, Ontario Tech University, 2000 Simcoe Street North, Oshawa, Ontario L1G 0C5, Canada

Isabel.pedersen@ontariotechu.ca

Justin Lewis

Euphoria, alienation, transformation: Vignettes of friction in AI writing

1 Introduction: Writing instruction and the US community college context

In December 2022, the advent of ChatGPT launched me on a journey from pedagogical euphoria to institutional panic, to a renewed, conditional hope for writing instruction. As a writing instructor in an open-access community college in the US, I work with a diverse student population in a resource-constrained environment where the promise of truly individualized AI assistance initially felt like a democratic dream fulfilled. Twenty-four months later, I find myself sad and hopeful about the prospects of teaching writing in an age of artificial intelligence.

Each phase of this journey reveals a different dimension of “friction”¹ between AI and writing: first, as pedagogical euphoria; then, as alienation from the cognitive processes that make writing pedagogically valuable; and finally, as a site of potential transformation. The three vignettes that follow trace this evolution through lived experience, examining how friction between human writers and artificial intelligence might reshape rather than simply threaten the future of composition pedagogy.

2 Vignette 1: Euphoria

The digital “stack” of portfolios sat unopened in my Canvas LMS when I opened ChatGPT for the first time in a work context. It was nearly 9 p.m., twenty-five essays awaited feedback, and I was exhausted. In a moment of desperation, I typed my first work-related prompt: “Help me write feedback for a student essay about community college transfer challenges.”

The response appeared in seconds—not generic platitudes, but genuinely thoughtful feedback addressing the student’s specific arguments about transfer credit policies. The AI had identified strengths (clear thesis, relevant examples) and offered constructive suggestions (stronger transitions, more evidence). More remarkably, the tone struck the balance I struggled to maintain after hours of grading: encouraging yet analytical, specific yet accessible. By offloading the laborious aspects of response, the AI was smoothing away friction in the assessment process; and, at first this, felt like a pedagogical breakthrough.

¹ In this essay, I draw heavily on Collin Gifford Brooke’s (2025) notion of “friction” in the context of generative AI as discussed here.

A week later, I introduced ChatGPT to my research writing course. Twenty-five students, ages 15–56, representing three first languages. We projected the interface and brainstormed together: “Should community college be free?” The room erupted as ideas cascaded across the screen—economic arguments about workforce development, social justice frameworks about educational equity, counter-arguments about funding mechanisms. Students who typically struggled to generate thesis statements were suddenly debating which angle would be most compelling. Maria, who had told me she “wasn’t a writer,” was explaining why the AI’s suggestion about student debt didn’t capture her experience as a working mother. The tool had become a thinking partner, reducing the blank-page friction that often paralyzed novice writers.

This seemed to fulfill the vision of process pedagogy (Flower & Hayes, 1981; Murray, 1972) by externalizing brainstorming and revision processes, finally making Murray’s dream of truly individualized, iterative writing instruction feasible. What Flower and Hayes had theorized about making complex mental operations visible was happening before my eyes. The AI wasn’t replacing student thinking; it was scaffolding it, giving students access to rhetorical moves that expert writers internalize but novices struggle to reach.²

The implications for multilingual students felt particularly profound. Students such as Carlos, whose insights about immigration policy were obscured by surface-level errors, could now focus on developing arguments while AI helped with syntactic complexity. Fatima, whose rich knowledge about educational systems in Eritrea was difficult to express in academic English, found in the AI a patient translator bridging linguistic communities. This wasn’t replacing their voices, but amplifying them by providing access to linguistic resources that native speakers took for granted.

By spring quarter, the democratic promise felt tangible. Students would draft with AI support, then bring those texts to peer review workshops. The recursive, social, meaning-focused writing process that process pedagogy had long envisioned finally seemed within reach for every student regardless of linguistic background or preparation level. The friction that had historically constrained both instructor feedback and student ideation was dissolving.

There was only one small discomfort, barely noticeable then. When I used AI to draft my own research proposals that winter, the writing flowed more smoothly than usual . . . perhaps too smoothly. But that unease was easy to dismiss amid the excitement of pedagogical possibility. In hindsight, I realize my winter euphoria masked deeper questions about what happens when the effortful processes that make writing pedagogically valuable become frictionless. The very ease that felt so liberating would soon reveal cracks in our understanding of what writing instruction actually accomplishes.

3 Vignette 2: Alienation

The call for a regional conference on AI and pedagogy sat open in one browser tab while ChatGPT waited in another. Six months after my initial discovery, AI had become routine for everything from email responses to assignment sheets. But this felt different—this was my own scholarship, my voice in the academic conversation. The 500-word abstract needed to capture months of thinking about AI’s impact on student agency while positioning my work within debates about digital literacies and writing pedagogy.

² This reminds me of Stuart Selber’s call for multiliteracies and challenges us to consider that tools do not simply expand fluency but reshape what counts as functional, critical, and rhetorical literacy in digital spaces (2004).

I crafted a prompt: “Help me draft an abstract for a conference presentation about AI writing tools in community college courses. Focus on student agency and democratic access” The response was technically excellent—coherent academic register, relevant theoretical frameworks, compelling research questions. It mentioned Freire’s (2018) critical pedagogy, referenced digital divide scholarship, and positioned the work appropriately. Yet something felt profoundly wrong. The prose was smooth—too smooth. It sounded like everyone and no one, a perfectly adequate academic voice that could have emerged from any scholar in our field.

I stared at the screen with growing unease. This wasn’t the excitement I had felt watching students discover ideas through AI collaboration. This was a hollowing sensation, as if my scholarly voice were being translated through some institutional algorithm that preserved form while evacuating substance. The abstract was good enough to submit, which was precisely the problem.

This anxiety demanded a different theoretical lens. As a graduate student, I had embraced Roland Barthes’ declaration of the author’s death with theoretical enthusiasm. His 1967 essay had liberated me from anxieties about originality by arguing that all writing draws from an “immense dictionary” of existing language—the author merely a “scriptor” arranging pre-existing materials rather than creating *ex nihilo* (145–146). What felt theoretically liberating in graduate school felt personally devastating in practice. ChatGPT seemed to perfectly embody Barthes’ vision, yet confronting this algorithmic realization of poststructural theory felt nothing like liberation. Instead, I experienced what Foucault might have recognized as the return of the author-function in its most anxious form.

Jody Shipka’s work on embodied literacies offered vocabulary for what was being lost. In *Toward a Composition Made Whole* (2011), Shipka argues that writing is never purely cognitive but always “a dynamic, embodied, multimodal whole—one that both shapes and is shaped by the environment” (p. 26). The physical act of composing, including pauses between sentences, false starts, the tactile experience of revision, constitutes meaning-making itself. Likewise, Kristie Fleckenstein’s (2003) account of embodied literacies extends this point: Composing emerges not only through words but through images, gestures, and the sensory imagination she calls “imageword.” What AI erases is not just effort but this embodied friction: the tactile, imaginal dimension that makes writing feel lived.

Taken together, Shipka’s and Fleckenstein’s accounts remind us that composing is always textured, bodily, and imaginal . . . the precise conditions that AI’s smoothing begins to strip away. Likewise, as Brooke notes, friction serves as a defining condition for intellectual growth inasmuch as it constitutes the inherent effort, resistance, and negotiation required to learn and create. As he puts it: “we don’t improve our physical or our mental abilities by removing the friction from our lives,” (para. 7) for we “learn by initiating friction (by encountering and engaging things we don’t know) and then by negotiating it, through reading, reflecting, note taking, memorization, associating it with other things we do know, etc.” (para. 25). Yet the promise of AI is precisely to offload that difficulty: “It’s the tyranny of convenience: in any given moment, it’s easier to ‘shove a few sloppy words’ than it is to take the time and effort to develop them. And as you rely more and more upon that convenience, the alternative becomes increasingly distant” (para. 5). In my own writing, the sentence-level struggle I had previously regarded as laborious was not simply a precursor to meaning but integral to it. With AI assuming this syntactic and conceptual labor, I had shifted from sustained structural engagement to superficial prompt optimization. I was becoming, in Brooke’s terms, less writer than prompt engineer.

Back in the classroom, I noticed similar patterns. An, whose personal narrative about overcoming addiction had been raw and powerful in early drafts, submitted a final revision that felt curiously flat:

grammatically perfect but emotionally distant. He admitted to “cleaning up” his draft with AI to “sound more academic.” The algorithm had indeed made his prose conventionally scholarly while smoothing away the very experiences that gave his argument force. This wasn’t the democratic access story I had celebrated; instead of amplifying An’s voice, the AI had normalized it, translating lived experience into academic lingua franca that met institutional expectations while diminishing its distinctive power.

Looking back, one might argue it was predictable that such unbridled enthusiasm would meet this reality check. Indeed, the initial “frictionless” promise carried the seeds of this alienation. Yet even as I grappled with these losses, I wasn’t ready to abandon AI entirely. I still remembered the genuine improvements it had offered, the students who had found their voices through its scaffolding.

I returned to my conference abstract with new awareness. Ultimately, I deleted the AI-generated text and painstakingly rewrote it myself, trading polished fluency for authentic voice, creating a draft that was rougher but genuinely mine. What had felt like personal inadequacy was revealing itself as institutional transformation, a crisis of authorship that extended far beyond my individual struggle. As the new academic year began, my private epiphany about friction’s importance was about to become a very public reckoning about academic integrity itself. What I experienced as a hollowing of voice, my colleagues were now experiencing as a hollowing of the assessment process. The friction that once made writing instruction meaningful was no longer just my problem; rather, it had become a crisis for the profession.

4 Vignette 3: Transformation

I spent Fall quarter 2024 listening to colleagues discover what I had begun to suspect: Our writing pedagogy was a broken system. Faculty emails filled my inbox with concerns about student work that “didn’t sound right.” Through either willful ignorance or administrative paralysis, the college provided no guidance on AI policy, leaving us to address instances of what we dubbed “Algiarism” case by case, instructor by instructor.

Beneath procedural questions around “Can we prove it?” and “What’s the policy?” ran deeper professional anxiety. These were colleagues who had spent decades perfecting their pedagogy, who prided themselves on meaningful assignments. Now they sounded like detectives, trading theories about linguistic tells. “I’m using Google Docs revision history to review composing processes,” one colleague told me, looking drained. “They’re just copying and pasting. A full third of the class is going to fail.” Another colleague, preparing materials for Student Affairs, discovered there was “no one at the wheel. We can’t reliably know if essays are composed by robots or students. The institution doesn’t care about the reliability of our assessments.”

This crisis felt familiar because it was. Academic integrity has always been haunted by questions of authentic authorship. The surveillance model that had dominated institutional responses since the 1950s treated plagiarism as character defect rather than complex negotiation of textual practices. Rebecca Moore Howard’s (1992, 1995, 1999) research on “patchwriting” had revealed that what instructors labeled plagiarism was often evidence of students learning to engage academic discourse (1992).³ Chris Anson’s *We never wanted to be cops* (2008) captured writing teachers’ frustration about being cast as academic police rather than literacy educators. Yet here we were again, defaulting to

³ Rebecca Moore Howard’s work on patchwriting (1995, 1999) showed us that what we call plagiarism often marks developmental writing, not deceit—a reminder that our responses to AI risk repeating old errors under a new name.

surveillance when faced with new textual technology. This seemed to me to be evidence that our approach to writing assessment was fundamentally broken.

To address our collective anxiety, a group of faculty formed an ad hoc AI Academic Integrity Workgroup during winter and spring 2025. Our report deliberately resisted surveillance models, instead drawing on decades of scholarship challenging punitive approaches to academic integrity. Where traditional responses focused on catching and punishing, we advocated systemic reform addressing the ecological conditions of assessment. We called for alternative assessment methods, labor-based grading systems that prioritize learning processes over policing products (following Asao Inoue's, 2022, work),⁴ and critical AI literacy courses. In short, we channeled Kathleen Blake Yancey's (2004) vision of writing in the 21st century to use assessment ecologies as evolving networks in which multimodal and process-driven practices better reflect how students actually compose today. When we recommended protected faculty time to develop ethical frameworks for AI engagement, we echoed Anson's argument that teachers should be educators, not cops.

In essence, we were acknowledging that students using AI to write traditional essays wasn't breaking our system—rather, it was revealing how broken the system already was. AI didn't introduce these cracks; it illuminated them. The crisis had forced recognition of pedagogical possibilities that required institutional disruption to become visible: If multimodal, process-focused, labor-based approaches were more effective for student learning, why had we defaulted to traditional essays for so long?

At heart, this raised fundamental questions about education's purpose, concerns Paulo Freire (2005) had articulated in his critique of banking education. In Freire's banking model, students are passive recipients of knowledge deposits, with teachers making "deposits which the students patiently receive, memorize, and repeat" (p. 72). Generative AI threatens to perfect this model, offering students the ultimate deposit system: not just receiving knowledge, but having it instantly transformed into seemingly original work. Where Freire saw banking education as dehumanizing, AI-mediated learning risks being posthuman altogether.

But learning requires friction. Echoing Collin Brooke's (2025) argument, writing builds capacity through resistance. As he notes, you cannot develop stronger muscles by having machines lift weights for you, and you cannot develop critical thinking by having AI do your thinking. The struggle with ideas, wrestling with language, the frustration of multiple drafts are not inefficiencies to optimize away. They are the mechanisms through which learning occurs. When students who have never constructed thesis statements ask ChatGPT to generate them, they're not learning argumentation. Instead, they're learning to outsource intellectual labor.

Kenneth Bruffee's (1984) work on collaborative learning offers a pathway forward, but only if we resist including AI as collaborator before students develop their own voices. True collaboration requires participants who bring knowledge, experience, perspective to conversation. When novice students collaborate with AI, they're not engaging in knowledge construction—they're being acted upon by a system with no investment in their learning, no commitment to their intellectual growth.

⁴ Asao Inoue's (2022) arguments about antiracist assessment clarify what is at stake here: Grading systems that privilege polished products over labor and process inevitably marginalize those for whom friction is greatest.

Ironically, the same technology that sparked panic about academic integrity is pushing us toward more ethical, more pedagogically sound assessment practices.⁵ The AI crisis has accelerated conversations about equity and authenticity that have simmered in composition studies for decades, forcing innovation in assignment design and grading practices. My hope for AI in writing education is conditional. It depends on our commitment to preserving productive difficulties that make learning possible, on designing assignments that maintain necessary friction, on understanding that dialogic pedagogy cannot include AI as dialogue partner until students develop their own capacity for dialogue. Taken together, these episodes of initial euphoria, to alienation of voice, to institutional reckoning suggest that friction is not simply an obstacle to overcome but the very condition of meaningful writing education.

The institutional response remains uncertain. Without administrative support for the labor-intensive reforms our report advocated—including process-based grading, multimodal assessment, and critical AI literacy—I suspect colleagues will default to surveillance approaches. But the crisis has at least clarified what’s at stake: not just preventing cheating, but ensuring educational environments in which students develop capacity to think, struggle, grow. We cannot prevent the AI revolution, but we can ensure that in this brave new world of writing, the friction that makes us human, and learners, is not optimized away.

References

- Anson, C. (2008). We never wanted to be cops: Plagiarism, institutional paranoia, and shared responsibility. In R. M. Howard & A. E. Robillard (Eds.), *Pluralizing plagiarism: Identities, contexts, pedagogies* (pp. 140–157). Boynton/Cook-Heinemann.
- Barthes, R. (1977). *Image-Music-Text*. Fontana Press.
- Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Utah State University Press.
- Brooke, C. G. (2025). *Stronger than friction?* Brookeshelves. <https://cgbrooke.substack.com/p/stronger-than-friction>
- Bruffee, K. A. (1984). Collaborative learning and the “conversation of mankind.” *College English*, 46(7), 635–652. <https://doi.org/10.2307/376924>
- Fleckenstein, K. S. (2003). *Embodied literacies: Imageword and a poetics of teaching*. Southern Illinois University Press.
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365–387.
- Freire, P. (2005). *Pedagogy of the oppressed: 30th Anniversary Edition*. Continuum.
- Freire, P. (2018). *Pedagogy of the oppressed: 50th Anniversary Edition*. Bloomsbury.

⁵ In this sense, Bob Broad’s (2003) model of *dynamic criteria mapping* (DCM) becomes instructive. His work challenges teachers and students to co-construct evaluative values rather than relying on static, surveillant standards. DCM rejects the idea that assessment criteria can or should be fixed in advance, instead treating them as contingent, negotiated, and embedded in the lived practices of particular classrooms and institutions. In the context of AI, where machine-produced text tempts institutions toward rigid rubrics or detection protocols, Broad’s model highlights the pedagogical value of making evaluation itself a site of dialogue. If friction is what sustains learning, then assessment must also preserve friction: the push-and-pull of perspectives between faculty and students about what makes writing meaningful, what counts as labor, and how quality is defined. DCM reframes evaluation from a top-down policing mechanism into a generative practice that resists automation, reminding us that the criteria by which we judge writing should evolve in concert with the very communities most affected by AI’s disruptions.

- Inoue, A. B. (2022). *Labor-based grading contracts: Building equity and inclusion in the compassionate writing classroom* (2nd ed.). The WAC Clearinghouse; University Press of Colorado.
- Moore Howard, R. (1992). A plagiarism pentimento. *Journal of Teaching Writing*, 11(2), 233–245.
- Moore Howard, R. (1995). Plagiarisms, authorships, and the academic death penalty. *College English*, 57(7), 788–806.
- Moore Howard, R. (1999). *Standing in the shadow of giants*. Bloomsbury Publishing.
- Murray, D. (1972). Teaching writing as a process not product. *The Leaflet*, 71(3), 11–14.
- Selber, S. (2004). *Multiliteracies for a digital age*. Southern Illinois University Press.
- Shipka, J. (2011). *Toward a composition made whole*. University of Pittsburgh Press.
- Yancey, K. B. (2004). Made not only in words: Composition in a new key. *College Composition and Communication*, 56(2), 297–328.

Author's address:

Justin Lewis, Olympic College, 1600 Chester Avenue – Bldg. 4, Room 330, Bremerton, WA 98337, USA,
jlewis2@olympic.edu

Afra Sturm, Valentin Unger & Fabian Grünig

Human versus machine: Comparing human and large-language-model assessments of students' writing through benchmark ratings

Abstract

For teachers to effectively use large-language-model(LLM)-based ratings in the formative or summative assessment of texts, it is essential to ensure that such ratings can assess student writing in a valid and reliable manner. This study investigates whether a validated human text-rating procedure (benchmark rating) can be replicated by an LLM-based rating procedure. We tested the replication with two genres of elementary school students' text—narrative and instructive—using nine LLMs from three providers (OpenAI, Anthropic, Mistral). Each LLM generated three independent scores per text via structured, benchmark-aligned prompts that were then aggregated into a consensus score. Results showed that intrarater reliability was high to excellent, $ICC(3, k) \approx .68-.97$, and alignment with human ratings ranged from moderate to strong, $ICC(3, 1) \approx .47-.85$, with larger models consistently outperforming smaller ones. Systematic bias patterns emerged, varying by model and genre, indicating a need for calibration. Increasing output token windows and reducing temperature parameters mitigated truncation and schema-related failures. Although LLM-based benchmark ratings can approximate expert judgments and reduce the need for labor-intensive human triple coding, limitations remain regarding cost (for larger models), genre- and task specificity, and sensitivity to text presentation and student grade level—factors that constrain immediate classroom use, particularly for formative feedback.

Keywords: text rating • large language models • narrative texts • instructive texts • elementary school

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 42–64

DOI: 10.21248/dideu.963

Copyright Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

1 Introduction

The effectiveness of feedback in writing instruction is well documented (Graham et al., 2011a; Scherer et al., 2024). Moreover, in both summative and formative feedback contexts, large language models (LLMs) hold considerable promise for supporting text rating.¹ Several arguments support their use:

- Text rating, whether for formative or summative purposes, is time-intensive. LLMs could help reduce teacher workload (Peltzer et al., 2025).
- Teacher-generated ratings often vary considerably (Birkel & Birkel, 2002). In contrast, LLMs generate responses based on probabilistic, distributional predictions (Zhao et al., 2023), potentially yielding less situationally influenced and more consistent assessments. By reducing susceptibility to subjective bias, LLM-based scoring may offer greater objectivity and rating quality.
- Process-oriented writing instruction and feedback on text revision—whether from teachers or peers—remain rare in classroom practice (Müller et al., 2023). Surveys among teachers and observations show that students predominantly receive summative feedback (Applebee & Langer, 2011; Roos, 2017) that focuses on final products rather than supporting students' writing during the writing process (Busse et al., 2022). This imbalance is probably related to the high demands on teacher time and effort. LLM-based feedback may offer a scalable way to provide timely, revision-oriented feedback that is integrated into the writing process.

To fully harness the potential of LLM-based feedback, however, LLM-based rating must meet the same requirements for objectivity, reliability, and validity as traditional text ratings. Otherwise, scores from LLM-based ratings will be too elastic or not suitable for teachers to make sound decisions about students' writing and their need to make progress as writers.

Following a brief overview of the current state of research, we present a study examining whether an established benchmark rating procedure can be replicated through structured prompting of multiple LLMs. The resulting ratings are compared with human ratings and evaluated using psychometric criteria (e.g., ICCs and mean-difference tests). Results indicate that LLM-based benchmark scoring is technically feasible following prompt and pipeline refinements, with virtually no unrecoverable failures. Across repeated runs, the models demonstrated high to excellent within-model consistency and moderate to strong overall alignment with human ratings. However, notable systematic biases emerged, varying by model and text genre—particularly for narrative texts and smaller models.

2 Empirical background

A recent meta-analysis on algorithm-based writing feedback found no statistically significant effects on either surface-level or deep-level writing outcomes. However, this is partly due to the limited number of eligible studies. Notably, this meta-analysis did not include studies on LLM-based feedback (Scherer et al., 2024). Early studies on the text-rating capabilities of LLMs present mixed results, with a slight overall trend toward high-quality text-rating performance (Altamimi, 2023; Bui & Barrot, 2024; Evmenova et al., 2024; Lu et al., 2024; Parker et al., 2023; Seßler et al., 2025; Shabara et al., 2024;

¹ In general, feedback and particularly feedback on texts also encompasses the evaluation of texts and the assessment of text quality. As used here, the term “feedback” subsumes summative and formative feedback as well as the assessment and rating of texts.

Steiss et al., 2024). However, the quality of these ratings has been shown to depend strongly on the structure and specificity of the prompt input (Jacobsen & Weber, 2025; Liu et al., 2023).

The existing studies address different aspects that can be categorized into distinct thematic areas. From these categories, several research gaps emerge that have yet to be explored systematically: consistency between LLM-based and human ratings; type of assessment; sensitivity to student characteristics, genres, and grade level; and single versus multiple LLMs.

(a) Consistency between LLM-based and human ratings

From both scientific and educational practice perspectives, it is highly relevant to determine the extent to which LLM-based ratings align with human ratings that themselves meet established standards of scoring validity and reliability. To understand the potential of LLM-based feedback, it is necessary to examine its performance through two lenses: the internal consistency of repeated iterations and the degree of correspondence with validated human ratings.

Seßler et al. (2025) report consistently good reliability of repeated evaluations using GPT-3.5, GPT-4, and GPT-o1, whereas LLaMA-3-70B and Mixtral-8×7B showed weaker internal agreement. However, only GPT-o1 achieved high correlations between LLMs and human raters. More variability was present when it came to content-related (deep-level) features, prompting the authors to suggest that LLMs may be more suitable for feedback on surface-level features. Whereas Scherer et al. (2024) and Seßler et al. (2025) assume, based on single studies, that algorithmic or LLM-based feedback is particularly accurate for assessing surface-level aspects, other researchers (e.g., Yan et al., 2024) argue that LLM-based text rating is also suitable for assessing deep-level features. However, it is important to note that interrater reliability among human raters is not always reported (e.g., Seßler et al., 2025), and this complicates interpretation of LLM–human alignment.

(b) Type of assessment

According to Schipolowski and Böhme (2016) assessments can follow different approaches: analytic (separate evaluation of various characteristics using criteria), holistic (overall evaluation of the text as a whole), or semiholistic (evaluation of distinct dimensions, but not fully criterion-separated). LLMs, however, do not analyze texts in a strictly criterion-based analytic manner—that is, they do not reliably decouple their evaluations along distinct, independently weighted criteria (Murugadoss et al., 2025). As a result, LLM-based ratings, particularly when prompted with multiple criteria, are best described as semiholistic. Even when individual criteria are specified in the prompt, the resulting rating is not analytical.

A new alternative is comparative assessment, which includes the benchmark rating approach (Bouwer et al., 2023, p. 305). In benchmark rating, texts are holistically positioned on a rating scale anchored by predefined benchmark texts. Each benchmark text exemplifies a specific quality level and is, for that purpose, enriched with textual descriptors. Empirical studies have shown that benchmark ratings meet the central psychometric quality standards of objectivity, reliability, and validity (Bouwer et al., 2023). To date, comparative or benchmark ratings using LLMs remain a desideratum. Given that benchmark ratings have been shown to be valid, reliable, and objective—and due to their clear operationalization that aligns well with structured LLM prompting—such procedures can be assumed to have high potential to be adapted for LLM ratings.

(c) Sensitivity to student characteristics

Previous studies have largely relied on relatively simple prompting strategies, occasionally including grade level and other student characteristics (e.g., second-language [L2] learners or struggling writers). This raises another desideratum: Evmenova et al. (2024) showed that LLM-based feedback does not always take information about students' abilities or needs into account, even though this information was provided in the prompt. It should be noted that Evmenova et al. (2024) used ChatGPT 3.5 and 4, and presumably also the smaller version of the model. With larger models, potentially better results could be achieved.

(d) Genre sensitivity

The studies cited above concentrated on one single text genre. Given that LLMs, as probabilistic language models, encode latent genre knowledge (Bubenhof, 2025), it is essential to evaluate whether their rating performance remains consistent across different genres. However, prior research has largely been limited to single genres in isolation, most commonly focusing on persuasive opinion-based essays, argumentative texts, or source-based writing (Bui & Barrot, 2024; Evmenova et al., 2024; Lu et al., 2024; Steiss et al., 2024); and sometimes also narrative texts (Seßler et al., 2025).

(e) Grade level

Most studies to date focused on postsecondary writing, leaving other educational levels underexplored. Notable exceptions include the studies by Evmenova et al. (2024) and Seßler et al. (2025) that examine school-level texts. As shown by Evmenova et al. (2024), LLM-based rating shows a limited capacity to differentiate between grade levels when assessing student texts. This issue is particularly evident in the fact that alignment of LLM-based with human ratings is lower when evaluating lower grade level texts.

(f) Single versus multiple LLMs

Variations in training data composition, model architecture, and scale suggest that different LLMs probably exhibit divergent performance in text assessment tasks. Recent findings by Seßler et al. (2025) indicate that larger and more sophisticated models—albeit with higher operational costs—generally tend to outperform smaller models. However, a systematic comparison of multiple LLMs within a unified benchmark rating framework is still lacking. Such comparative analyses are essential to determine which models are most suitable for educational use—particularly in terms of reliability, cost-efficiency, and genre or age-group sensitivity.

3 Research questions

To address the research gaps identified above, especially regarding (a) consistency, (b) type of assessment, (d) several genres, (e) lower grade levels and (f) multiple LLMs, the present study tests a benchmark-aligned prompting procedure across two text genres (narrative and instructive), lower grade levels (4 and 5), and multiple LLMs from different providers (see below). The study aims to answer the following research questions:

- (1) Can a comparative rating procedure, such as benchmark rating, be successfully implemented into an LLM-based assessment system?

- (2) Do LLM-generated benchmark ratings demonstrate sufficient intrarater reliability across repeated assessments, and is this reliability consistent across different model providers and architectures?
- (3) To what extent do LLM-generated benchmark scores correlate with human expert ratings, and are there systematic differences in scoring patterns between human raters and LLMs?

These questions are central not only to understanding the general capability of LLMs for educational text evaluation but also to evaluating their potential for use in teaching and learning settings (e.g., model solutions in the classroom, automated formative feedback, and integration into digital learning environments).

4 Method

4.1 Sample

The present study—conducted as part of the DEEP consortium (<https://deep-consortium.ch/de>) and funded by the Jacobs Foundation—uses data from another project funded by the Swiss National Science Foundation.² This project was designed as an intervention study with three measurement points and aimed to investigate whether a writing-fluency training, implemented in addition to a process-oriented writing program, enhances students' writing competencies (Grade 4/5), and whether students with German as a second language (L2) benefit more than students with German as a first language (L1) (for further details on the intervention study, see Sturm & Schneider, 2021). The data used in the present study stem from all three measurement points.

To ensure transparency of the chosen approach, Section 4.2 provides a detailed outline of the benchmark rating procedure, and Section 4.3 outlines the design of the LLM-based system and the technical implementation, including information on the selected models.

4.2 Material and rating procedure

Students completed two writing tasks—one narrative and one instructive—each within a 30-min timeframe. The narrative writing task was based on a prompt adapted from Schneider et al. (2012): “It was a rainy Sunday. Family Racconti was still asleep.”³ Students were asked to continue the story and write an exciting narrative text. The instructive writing task required students to write step-by-step directions for drawing a specific picture. They were shown an image of a snowman standing on a meadow in the sun and instructed to trace the image so that someone unfamiliar with it could reproduce it based on the instructions.

To eliminate potential presentation effects (Graham et al., 2011b),⁴ the handwritten texts were transcribed and corrected for spelling and grammar. The resulting dataset from all three measurements comprised 873 narrative texts (mean text length $M = 820.52$ characters, $SD = 412.07$ characters,

² “Fostering writing fluency,” grant number 100019_159311

³ The story prompt was given in German: “*Es war ein verregneter Sonntag. Familie Racconti schlief noch.*” The story prompt was slightly modified at each measurement point: “There was thick fog outside. The Racconti family were having dinner.” “It was snowing heavily outside. The Racconti family were having lunch.” Despite these changes, the prompts remained comparable in structure and intent, each setting the stage for a narrative involving the Racconti family.

⁴ Surface-level features such as spelling or grammar errors may influence raters' judgments about text quality.

Min = 62 characters, Max = 2,586 characters) and 859 instructive texts (mean text length $M = 459.95$ characters, $SD = 243.35$ characters, Min = 56 characters, Max = 1,502 characters).

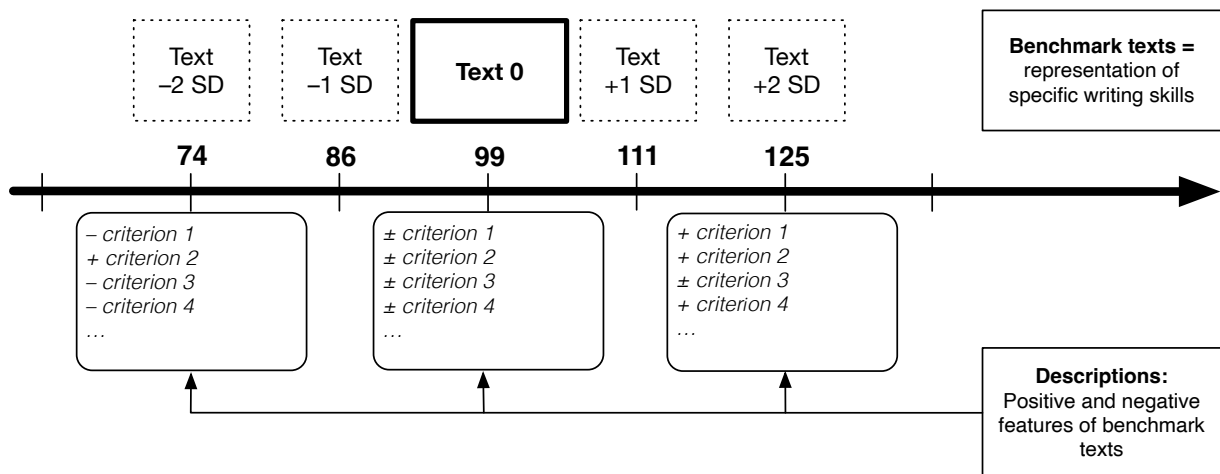
The quality of the texts was assessed using a benchmark rating. To this end, a scale of five benchmark texts representing varying quality levels was developed following a five-step procedure outlined by Feenstra (2014). Steps 1–4 involved the random selection of 48 narrative and 48 instructive texts from the full sample.

- (1) A group of three experienced raters formulated genre- and task-specific criteria deemed essential for assessing text quality. These criteria were aligned with widely accepted dimensions of writing assessment, content, structure, and language; and they tailored the two task types: For the narrative task, criteria focused on the central idea of the story, the narrative structure, and the use of linguistic devices to create tension or to differentiate characters or other narrative elements. For the instructive task, criteria included the reader-oriented clarity of the instructions, the descriptive quality of the image, the logical sequencing of the instructional steps, and the use of cohesive linguistic markers.
- (2) Based on these criteria from Step 1, the raters selected a representative “average” benchmark text from the sample that demonstrated both strengths and weaknesses across the defined dimensions. This text was assigned an arbitrary score of 100, accompanied by a qualitative rationale that detailed why the text was considered of average quality with reference to the task-specific criteria.
- (3) The raters from Step 1 then scored the remaining randomly selected student texts by comparing each one directly to the average benchmark text. Each text received a numerical score indicating how it compared to the benchmark: Thus, if a text seemed only half as good as the benchmark text, it received a score of 50; if it seemed twice as good, it was assigned a score of 200. Interrater reliability for this comparative prerating was good⁵ (narrative: .89 for ICC[3, 1]; instructive: .86 for ICC[3, 1]). To control for individual rater leniency and severity, all scores were z-standardized. Next, an average score was computed per text across all three judgments. These means were then transformed to a new scale with a mean of 100 and a standard deviation of 15 to facilitate the construction of a benchmark scale.
- (4) Based on the transformed scores, five texts per genre were selected to represent different levels of writing quality with (mean) ratings around 70, 85, 100, 115, and 130. Each of the five selected benchmark texts was supplemented with a descriptor summarizing its key strengths and weaknesses according to the established criteria. Figure 1 illustrates this for the narrative benchmark texts (see also Appendix A with the narrative benchmark Text 0 and accompanying descriptions).
- (5) Using the newly constructed benchmark scale, an independent group of three trained raters (different from those in Step 3) rated all texts from the three measurement points in randomized order. Each text was systematically compared to the benchmark texts and assigned a score that places it along the scale. The text under evaluation was first compared with Text 0; if it differed from Text 0, it was then compared with the next benchmark text at +1 SD or –1 SD , and, if necessary, at +2 SD or –2 SD .

⁵ Koo and Li (2016) classify values above .75 as “good”. In the context of complex educational writing assessments, such coefficients are often regarded as sufficient for consequential use of ratings (Graham et al., 2012).

Figure 1

Schematic Representation of the Five Narrative Benchmark Texts with Their Respective Scores



Most raters involved in Step 5 were student teachers, although a smaller number were research assistants participating in Sturm and Schneider's (2021) intervention study. All raters received a half-day training session supported by a detailed rating manual. During training, raters practiced on a selection of texts of varying quality and discussed their judgments with the trainer to align on the benchmark-based rating approach. The three raters then independently scored a set of 50 additional texts. Provided that interrater reliability reached an acceptable level, they proceeded to rate the remaining texts independently. For the instructive texts, a second training was required due to uncertainties in handling student texts that were purely descriptive or narrative, thereby deviating from the task expectations. Interrater reliability of the final human ratings was estimated using the intraclass correlation, yielding an excellent reliability of $ICC(3, 1) = .99$. For all subsequent analyses, the mean score across the three raters was computed for each text.

4.3 Design of the LLM-based system

To translate the benchmark rating procedure into an LLM-based system, we designed genre-specific system prompts and individualized user prompts for each student text. The system prompt employed several evidence-based prompt engineering strategies to enhance model performance and reliability.

Specifically, the system prompt was structured to incorporate multiple complementary strategies. First, it used human-aligned instruction design by mirroring both the benchmark descriptors and the training manual used to calibrate human raters, thereby anchoring the LLM to established expert standards as described in Section 4.2. Second, it implemented a step-by-step reasoning specification via an explicit, hierarchical outline of the assessment criteria, effectively using a chain-of-thought (CoT) approach.⁶ Third, the prompt incorporated explicit reasoning extraction, instructing the model to articulate and return its rationale for each score. Finally, the system prompt employed structured output formatting, requiring all benchmark scores to be in a machine-readable JSON format, facilitating

⁶ Chain-of-thought prompting is a technique that enables LLMs to solve complex tasks by decomposing them into a series of intermediate reasoning steps. This process mimics human-like problem solving, improving the model's accuracy and transparency in decision making (Wei et al., 2023).

downstream processing, and minimizing ambiguity in score extraction. The full system prompt for scoring narrative texts is provided in Appendix B.

Rather than submitting requests synchronously, we employed batch processing⁷ across multiple LLM providers via each provider's native batch API. This asynchronous, provider-specific batch handling reduced costs while enabling high-throughput processing. For each model–corpus combination, we generated a provider-specific JSONL batch file⁸ containing the system prompt, the instantiated user prompt, and structured schema constraints. The user prompt template for the narrative text scoring is shown in Appendix C. Responses that failed schema validation or resulted in API errors were logged separately, along with corresponding error messages, allowing for subsequent error analysis.

To answer the first research question, we implemented this proof of concept with three LLM providers: OpenAI and Anthropic as well-established commercial providers, and Mistral, representing a European-based provider. In total, we evaluated student texts across nine distinct LLM variants, spanning these three providers and representing different performance levels (see Table 1). Notably, we did not use reasoning-augmented models in this study, because these entail substantially higher computational costs than standard language models, and our study prioritizes cost-effective applications suitable for research and educational contexts. The step-by-step reasoning specification we provided in the system prompt already enables LLMs to articulate abbreviated reasoning processes and justifications for text assessment.

Table 1

Providers and Models of Different Performance Levels

Provider	Large	Medium	Small
Mistral	mistral-medium-2505	open-mistral-nemo-2407	ministral-8b-2410
Anthropic	claude-sonnet-4-20250514	claude-3-5-haiku-20241022	claude-3-haiku-20240307
OpenAI	gpt-4.1-2025-04-14	gpt-4.1-mini-2025-04-14	gpt-4.1-nano-2025-04-14

All requests were executed using standardized generation parameters. The maximum token⁹ output was set to 1,024 tokens, and the temperature¹⁰ was set to standard values used in the chat interface for end users, enabling consistent behavior across evaluations. The temperature parameter controls the randomness of token selection by scaling the probability distribution of output tokens, thereby affecting the determinism of model responses (Li et al., 2025).

To answer the second research question, we refined both the prompt design and processing pipeline based on insights from the initial proof-of-concept implementation and identified failure modes. While

⁷ Unlike synchronous processing (real-time interaction), batch processing allows a large number of requests to be sent simultaneously and processed by the provider at a lower priority. This reduces costs.

⁸ JSONL (JSON Lines) is a text format in which each line is a valid JSON object, allowing for efficient processing of large datasets because each record can be handled independently.

⁹ A token is the basic unit of text processed by an LLM, representing a sequence of characters, a word, or part of a word.

¹⁰ The temperature parameter ranges typically from 0 to 1 (or 2). Lower values (e.g., < 0.3) make the model more deterministic and focused, whereas higher values increase creativity and diversity but also the risk of “hallucinations” or inconsistencies.

retaining the same set of LLM providers and models, we enhanced the system prompt to elicit more explicit comparisons between student texts and benchmark exemplars. We increased the maximum token output to 2,048 tokens to prevent truncation, and the temperature was fixed at 0.3, allowing for more deterministic responses that align with the structured output schema.¹¹

To assess the intrarater reliability of LLM scoring, each student text was evaluated three independent times by each of the nine models. We then computed ICCs using a two-way mixed effects model for mean-rating consistency—that is, $\text{ICC}(3, k)$ with $k = 3$ raters.

To address the third research question concerning the validity of automated LLM-based scoring, we compared LLM-generated scores to human benchmark ratings. For each text, an aggregate LLM score was calculated by averaging the results from the three independent scoring runs. To handle missing data, we applied a systematic protocol: A missing score in a single evaluation run was excluded from that run's contribution, and the aggregate score was calculated from the remaining two evaluations. Texts with two or more missing scores across the three runs were excluded from further analysis due to insufficient evaluation data. These aggregate LLM scores were then compared to the human expert benchmark scores using a two-way mixed-effects model for consistency of a single rater—that is, $\text{ICC}(3, 1)$. To examine systematic differences between human and LLM scoring, we computed the mean difference between LLM and human scores for each text, along with corresponding standard deviations. Paired-samples *t*-tests were conducted to test for significant mean differences at the aggregate level and Cohen's *d* was reported as a measure of effect size to characterize the magnitude of observed differences.

5 Results

First, we report findings from the initial implementation of our LLM-based benchmark scoring system across nine LLM variants (three providers, three models per provider, see Table 1). The aim was to assess feasibility and identify any systematic failure pattern in schema validation and output generation. Second, we present reliability analyses based on three independent scoring runs per text, quantifying intrarater consistency using $\text{ICC}(3, k)$ with 95% confidence intervals (CI). Third, we evaluate criterion validity by comparing aggregate LLM-generated scores against human expert benchmark ratings using $\text{ICC}(3, 1)$ and paired-samples *t*-tests, with effect sizes (Cohen's *d*) characterizing the magnitude of any systematic differences.

The LLM-based benchmark rating system was shown to be technically feasible across all nine model variants, despite notable heterogeneity in initial performance. During the full proof-of-concept phase, the system successfully generated scoreable outputs for the majority of student texts. Based on the error analysis from the proof-of-concept phase, two critical issues emerged that required methodological adjustments in the subsequent phase: (a) The primary failure mode across all LLM providers was output truncation at the 1,024-token limit. In 89.81% of these cases, extraction errors could be resolved through post hoc automated recovery. Nonetheless, the underlying root cause was systematic: Models tended to generate reasoning justifications that exceeded the token budget. (b) The error patterns also revealed that higher temperature (default chat settings) contributed to verbose,

¹¹ Although lower temperature values are known to reduce output randomness and promote more deterministic model behavior, specific temperature recommendations in the literature are limited. The value of 0.3 was selected based on common practice in applications requiring consistent, structured output, though we note this reflects practitioner guidance rather than empirically derived thresholds from academic research.

unfocused reasoning, which often failed to align with the structured output schema. To address these issues, two adjustments were implemented: The maximum token limit was increased to 2,048 tokens, and the temperature was reduced to 0.3 to promote more deterministic responses. In the refined scoring phase, the system yielded just 0.02% unrecoverable failures, which were treated as missing values in subsequent analyses.

To address our second research question, we evaluated the consistency and reliability of LLM-based scoring across three repeated independent evaluation runs (see Table 2). Following Koo and Li (2016), results revealed moderate to excellent intrarater reliability across all nine LLM variants and both text genres.

- Narrative texts: ICC(3, k) values ranged from .77 to .95 ($M = .88$), falling within the good to excellent range. The most consistent models were OpenAI GPT-4.1 and Claude Sonnet.
- Instructive texts: Reliability was comparable or higher, with ICC(3, k) values between .68 and .97 ($M = .87$), again with OpenAI GPT-4.1 achieving the highest reliability. Smaller models fell into the moderate category.

These findings confirm that the refined prompt design combined with increased token budget and reduced temperature successfully enhanced the determinism and reproducibility of the scoring system.

Table 2

Intraclass Correlation of Three Individual Scores Produced by the Same LLM by the Benchmark Rating Procedure

Corpus	Model	Intraclass correlation ICC(3, k) and 95% CI		
		Estimate	Lower limit	Upper limit
Instructive	gpt-4.1-2025-04-14	.973	.969	.976
	claude-sonnet-4-20250514	.955	.949	.960
	claude-3-5-haiku-20241022	.949	.943	.955
	mistral-medium-2505	.942	.935	.949
	gpt-4.1-mini-2025-04-14	.936	.928	.943
	claude-3-haiku-20240307	.877	.862	.891
	open-mistral-nemo-2407	.765	.736	.791
	ministral-8b-2410	.744	.712	.773
	gpt-4.1-nano-2025-04-14	.684	.644	.720
Narrative	claude-sonnet-4-20250514	.952	.947	.958
	gpt-4.1-2025-04-14	.952	.946	.957
	claude-3-5-haiku-20241022	.924	.915	.932
	gpt-4.1-mini-2025-04-14	.907	.896	.917
	mistral-medium-2505	.896	.883	.907
	claude-3-haiku-20240307	.876	.861	.890
	open-mistral-nemo-2407	.859	.842	.875
	gpt-4.1-nano-2025-04-14	.772	.744	.797
	ministral-8b-2410	.766	.738	.792

Note. For each corpus the models are sorted by the point estimate of the ICC(3, k) value in descending order.

To address our third research question, we compared aggregate LLM scores (like in human-based rating—cf. Section 4.2—, averaged across three independent runs) with human expert benchmark ratings (see Table 3). In 14 cases, the aggregate LLM score was based on only two valid runs due to missing values. For narrative texts, the alignment of LLM-generated with human ratings ranged from poor to good: ICC(3, 1) = .47 to .85 ($M = .71$). OpenAI GPT-4.1 demonstrated the highest alignment with human benchmarks, followed by OpenAI GPT-4.1-mini and Claude Sonnet. In contrast, the smaller Mistral Ministral-8b model showed substantially lower alignment. For instructive texts, alignment was similarly variable: ICC(3, 1) = .61 to .85 ($M = .75$), ranging from moderate to good. Mistral Medium, Claude Sonnet, and OpenAI GPT-4.1 achieved the highest validity for instructive texts, whereas smaller models again showed reduced alignment with human ratings.

Table 3

Intraclass Correlation Between Human and LLM-Based Mean Scores Produced by the Benchmark Rating Procedure

Corpus	Model	Intraclass correlation ICC(3) and 95% CI		
		Estimate	Lower limit	Upper limit
Instructive	mistral-medium-2505	.848	.828	.865
	claude-sonnet-4-20250514	.830	.808	.849
	gpt-4.1-2025-04-14	.819	.796	.840
	gpt-4.1-mini-2025-04-14	.801	.776	.824
	claude-3-haiku-20240307	.793	.767	.817
	claude-3-5-haiku-20241022	.725	.691	.755
	ministral-8b-2410	.683	.646	.718
	open-mistral-nemo-2407	.618	.575	.658
	gpt-4.1-nano-2025-04-14	.608	.564	.649
Narrative	gpt-4.1-2025-04-14	.854	.835	.871
	gpt-4.1-mini-2025-04-14	.808	.784	.830
	claude-sonnet-4-20250514	.787	.760	.811
	mistral-medium-2505	.770	.741	.796
	gpt-4.1-nano-2025-04-14	.712	.678	.743
	claude-3-5-haiku-20241022	.682	.645	.716
	open-mistral-nemo-2407	.638	.597	.676
	claude-3-haiku-20240307	.637	.595	.674
	ministral-8b-2410	.471	.418	.521

Note. For each corpus the models are sorted by the point estimate of the ICC(3) value in descending order.

Paired-samples *t*-tests and Cohen's *d* effect sizes revealed substantial systematic bias patterns across the evaluated models (see Table 4). For instructive texts, four models, with small to medium performance levels, showed negligible differences from human ratings. Claude Sonnet showed a small underestimation of text quality compared to the human ratings, whereas OpenAI GPT-4.1 showed a large overestimation. In contrast, the other small models, OpenAI GPT-4.1-nano and Mistral Ministral, showed substantial systematic underestimation of text quality.

Table 4

Mean Difference Analysis of Human and LLM-Based Mean Scores Produced by the Benchmark Rating Procedure

Corpus	Model	Difference between human and LLM-based mean scores		Paired sample t-test			Cohen's <i>d</i>
		<i>M</i>	<i>SD</i>	<i>t</i>	<i>df</i>	<i>p</i>	
Instructive	gpt-4.1-nano-2025-04-14	18.976	9.774	54.913	799	< .001	1.941
	ministral-8b-2410	11.589	8.205	39.825	794	< .001	1.412
	claude-sonnet-4-20250514	1.290	6.918	5.355	824	< .001	0.186
	claude-3-haiku-20240307	0.600	7.340	2.300	790	.022	0.082
	claude-3-5-haiku-20241022	-0.219	9.855	-0.630	806	.529	-0.022
	gpt-4.1-mini-2025-04-14	-0.194	8.274	-0.671	822	.502	-0.023
	open-mistral-nemo-2407	-0.457	11.587	-1.133	824	.258	-0.039
	mistral-medium-2505	-3.276	6.306	-14.920	824	< .001	-0.519
	gpt-4.1-2025-04-14	-5.562	7.754	-20.614	825	< .001	-0.717
Narrative	ministral-8b-2410	17.642	10.596	48.940	863	< .001	1.665
	gpt-4.1-nano-2025-04-14	8.348	9.633	25.472	863	< .001	0.867
	claude-3-5-haiku-20241022	6.473	10.183	18.683	863	< .001	0.636
	gpt-4.1-mini-2025-04-14	5.603	9.157	17.988	863	< .001	0.612
	open-mistral-nemo-2407	5.299	11.035	14.114	863	< .001	0.480
	claude-3-haiku-20240307	5.011	10.574	13.922	862	< .001	0.474
	gpt-4.1-2025-04-14	2.068	7.436	8.176	863	< .001	0.278
	claude-sonnet-4-20250514	2.248	10.230	6.458	863	< .001	0.220
	mistral-medium-2505	-1.862	8.943	-6.119	863	< .001	-0.208

Note. A positive value for the mean difference between human and LLM-based mean scores indicates that humans assigned higher scores to the student texts than the LLM-based scoring system. For each corpus, models are sorted by the effect size in descending order.

For narrative texts, systematic bias was more pronounced. All models showed significant differences to those of human ratings. The majority of models significantly underestimated text quality. Mistral Ministral showed the largest bias followed by OpenAI GPT-4.1 nano—both representing large effect sizes. Other models also substantially overestimated text quality, though with small to medium effects. The closest alignment with human judgments was found in Claude Sonnet, Open AI GPT-4.1, and Mistral Medium—all showing small but statistically significant differences. Notably, Mistral Medium was the only model to systematically overestimate the quality of narrative texts.

6 Discussion

This study examined whether an established human rating procedure (benchmark rating) can be replicated using LLMs, and whether such replication meets key psychometric criteria. Specifically, the latter question must be resolved affirmatively before an LLM-based assessment of this kind can be authorized for implementation in schools. Overall, findings indicate that LLM-based benchmark rating is feasible, demonstrates good alignment with human ratings, and shows acceptable reliability and validity. However, several challenges remain before LLM-based assessment can be reliably integrated into educational practice.

In response to our first research question, results confirm that LLM-based benchmark rating is operationally viable across a diverse set of models and providers. However, successful implementation required prompt refinement and parameter optimization—specifically, increased output windows and reduced temperature settings—to reduce schema failures and ensure consistency. Without such technical adjustments, initial failure rates were substantial.

When texts are rated through comparative benchmark rating, there is generally less room for subjective interpretation than in analytical or (semi)holistic approaches (Lindauer & Sommer, 2018). Consequently, interrater reliability among human raters tends to be very high. This makes comparative benchmark rating highly relevant for educational use. Regarding our second research question, findings show that the consistency of LLM-based ratings across all nine LLM variants and both text genres was satisfactory to excellent. As expected, larger, more sophisticated models (OpenAI GPT-4.1, Claude Sonnet) outperformed models of medium and small performance levels and demonstrated higher consistency.

Regarding our third research question, LLM-generated scores were found to correlate meaningfully with human expert ratings across models and genres. However, systematic differences were also observed: Larger models demonstrated closer alignment with human standards and smaller effect sizes in deviation from human scores. The smallest models (OpenAI GPT-4.1 nano and Mistral minstral-8b variants) showed larger, consistent biases. Furthermore, genre-specific patterns emerged: Instructive texts were often underestimated (or scored conservatively) by midsize and large models, whereas narrative texts tended to be overestimated, particularly by smaller models. These findings suggest that genre and model-specific calibration is essential when applying LLM-based assessments in practice.

It should be noted that the LLMs used in this study—provided by OpenAI, Anthropic, and Mistral—are trained on broad but nonspecific datasets. As the SWK (2023) emphasizes, transparency regarding training data is critical in educational contexts. The report also highlights the potential of “model solutions” to improve LLM-based feedback. In our study, we operationalize this concept by using comparisons with benchmark texts as a model solution. These serve as transparent reference standards that communicate task- and level-specific performance expectations to the LLM. However, a model solution is more than a tool for achieving high statistical alignment of human with LLM-based ratings. It acts as an anchor that makes performance expectations explicit and verifiable. This shifts the evaluative focus from correlational agreement to alignment with educational standards. In our study, the benchmark texts represent a specific stage in the learning trajectory, namely from Midgrade 4 to Midgrade 5. It follows that different grade levels or genres, such as argumentative writing, would require their own distinct benchmark texts to ensure the model solution accurately reflects appropriate expectations.

Whereas previous studies have focused largely on academic or argumentative texts (e.g., persuasive opinion essays or source-based writing), our findings demonstrate that LLM-based benchmark ratings can be successfully applied to both narrative and instructive texts. Notably, results yield slightly better validity for instructive texts than for narrative texts (e.g., higher average alignment with human ratings, ICC(3, 1) with $M = .75$ and range .61–.85 versus $M = .71$ and range .47–.85). This probably reflects that the instructive task is more tightly specified: The writing goal and audience are explicit, and the task permits less variation in acceptable solutions than a narrative task. In contrast, the narrative prompt is more open-ended, leading to highly diverse responses: Some students wrote adventure stories; others fantasy, crime, or even a drama with an unhappy ending. Human raters appear somewhat better equipped to handle such open-ended tasks. To enable broad implementation of LLM-based benchmark rating across genres in educational settings, further research with additional genres is required.

It remains unclear to what extent the presentation effect arises in LLM-based feedback in a manner analogous to human rating, or whether it can be substantially reduced. This question would arise particularly if—unlike in the present study—uncorrected texts were to be assessed using an LLM-based approach. As Sturm et al. (2024) show, presentation effects can also be demonstrated in human-based benchmark rating. When we use the same texts in their uncorrected form, we find initial indications that LLM-based feedback is likewise subject to this effect (Sturm et al., 2026).

The various providers tested are comparable in terms of assessment quality. However, larger models both perform better and appear less susceptible to distortions. This complicates the use of LLM-based ratings in school settings, because these models are substantially more expensive than smaller LLMs. Based on these findings, efforts by some German federal states to develop their own LLMs should be viewed with caution, because such models are unlikely to match the capabilities of the large international systems. If schools nevertheless opt to use low-cost or unverified models, it can be expected that students' writing performance will not be assessed reliably, and that instructional decisions based on such assessments may likewise be inappropriate.

Evidence from pre-LLM-automated writing evaluation indicates that such feedback is most effective when part of a larger instructional program in which, for example, teachers provide strategy instruction, iterative practice, or individual support (Nunes et al., 2022). Although evidence is lacking for LLM-based feedback, it is reasonable to assume that these pedagogical principles are transferable to this new generation of feedback tools.

The development of comprehensive instructional programs and materials that integrate the potential of LLMs therefore appears relevant. Although LLMs can replicate text assessment, the ultimate responsibility for assessing texts should remain with teachers (SWK, 2023). Depending on which LLM-based tools are available and actually used—whether lower-cost or higher-cost, validated for the respective grade level, and validated for various genres—teachers would need to pay closer attention particularly to students with lower writing performance.

Limitations

As mentioned above, a key limitation of the current study is that only two text types have been evaluated so far. Future work should extend the approach to additional genres. A further limitation, particularly complicating use in educational contexts, is that benchmark ratings have thus far been restricted to the closed setting of specific writing tasks for which benchmark texts are available. Developing such benchmarks remains a manual, time-intensive process. Furthermore, our approach currently provides only summative feedback. Generating formative feedback—which would ultimately provide real relief

for teachers—still requires substantial development. This would require linking total scores to pedagogically useful descriptors that identify text-specific strengths and areas for improvement in a learner-appropriate way. Benchmark ratings appear promising in this regard, because they are inherently descriptor-based and align with defined performance expectations. Initial attempts in this area are already promising. In addition to intended use in teaching and learning settings, it is already conceivable that LLM-based benchmark rating could also serve as a testing tool in research projects—for instance, in cross-sectional assessments of students' writing skills, similar to DESI (Neumann & Lehmann, 2008) or NAEP (2017).

References

- Altamimi, A. B. (2023). Effectiveness of ChatGPT in essay autograding. In *2023 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 102–106). <https://doi.org/10.1109/iCCECE59400.2023.10238541>
- Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal*, 100(6), 14–27. <https://doi.org/10.58680/ej201116413>
- Birkel, C., & Birkel, P. (2002). Wie einzig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, 49, 219–224.
- Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., & De Maeyer, S. (2023). Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement. *Journal of Writing Research*, 15(3), 497–518. <https://doi.org/10.17239/jowr-2024.15.03.03>
- Bubenhofer, N. (2025). Vektorisierte Praktiken: Ein linguistisches Verständnis von Large Language Models. *Leseräume*, 11, 1–16.
- Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, 30, 2041–2058. <https://doi.org/10.1007/s10639-024-12891-w>
- Busse, V., Müller, N., & Siekmann, L. (2022). Wirksame Schreibförderung durch diversitätssensibles formatives Feedback. In V. Busse, N. Müller, & L. Siekmann (Eds.), *Schreiben fachübergreifend fördern. Theoretische Grundlagen und Praxisanregungen für Schule, Unterricht und Lehrerinnen- und Lehrerbildung* (pp. 114–133). Klett Kallmeyer.
- Evmenova, A. S., Regan, K., Mergen, R., & Hrisseh, R. (2024). Improving writing feedback for struggling writers: Generative AI to the rescue? *TechTrends*, 68(4), 790–802. <https://doi.org/10.1007/s11528-024-00965-y>
- Feenstra, H. (2014). *Assessing writing ability in primary education: On the evaluation of text quality and text complexity* (Doctoral dissertation). Universiteit Twente.
- Graham, S., Harris, K. R., & Hebert, M. (2011a). *Informing writing: The benefits of formative assessment*. Carnegie Corporation of New York.
- Graham, S., Harris, K. R., & Hebert, M. (2011b). It is more than just the message: Presentation effects in scoring writing. *Focus on Exceptional Children*, 44(4), 1–12.
- Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform. <https://files.eric.ed.gov/fulltext/ED532068.pdf>

- Jacobsen, L. J., & Weber, K. E. (2025). The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback. *AI*, 6(35), 1–17. <https://doi.org/10.3390/ai6020035>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- Li, L., Sleem, L., Gentile, N., Nichil, G., & State, R. (2025). Exploring the impact of temperature on large language models: Hot or cold? *arXiv*. <https://doi.org/10.48550/arXiv.2506.07295>
- Lindauer, N., & Sommer, T. (2018). Verfahren der Textbeurteilung: Merkmale und Vorzüge eines holistischen Benchmarkratings. *Leseräume*, 5, 1–14.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), Article 195, 1–35. <https://doi.org/10.1145/3560815>
- Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education*, 49(5), 616–633. <https://doi.org/10.1080/02602938.2024.2301722>
- Müller, N., Utesch, T., & Busse, V. (2023). Qualität statt Quantität? Zum Zusammenhang von Schreibförderungs- und Feedbackpraktiken mit Textqualität unter Berücksichtigung von migrationsbedingter Mehrsprachigkeit. *Unterrichtswissenschaft*, 51(2), 169–198. <https://doi.org/10.1007/s42010-023-00173-2>
- Murugadoss, B., Poelitz, C., Drosos, I., Le, V., McKenna, N., Negreanu, C. S., Parnin, C., & Sarkar, A. (2025). Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18), 19589–19597. <https://doi.org/10.1609/aaai.v39i18.34157>
- NAEP. (2017). *Writing framework for the 2017 National Assessment of Educational Progress*. National Assessment Governing Board.
- Neumann, A., & Lehmann, R. H. (2008). Schreiben Deutsch. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (pp. 89–103). Beltz.
- Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, 38(2), 599–620. <https://doi.org/10.1111/jcal.12635>
- Parker, J. L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education*, 62(12), 721–727. <https://doi.org/10.3928/01484834-20231006-02>
- Peltzer, K., Lira Lorca, A., Krause, U.-M., Graham, S., Panadero, E., & Busse, V. (2025). How to support at-risk writers: Differential effects of formative feedback on argumentative writing and motivation. *Reading and Writing*, 39, 965–1000. <https://doi.org/10.1007/s11145-025-10652-w>
- Roos, M. (2017). *QUIMS-Bericht: Entwicklungen von 2014 bis 2016 (provisorische Fassung)*. spectrum GmbH.
- Scherer, S., Graham, S., & Busse, V. (2024). How effective is feedback for L1, L2, and FL learners' writing? A meta-analysis. *Learning and Instruction*, 93, Article 101961. <https://doi.org/10.1016/j.learninstruc.2024.101961>

- Schipolowski, S., & Böhme, K. (2016). Assessment of writing ability in secondary education: Comparison of analytic and holistic scoring systems for use in large-scale assessments. *L1-Educational Studies in Language and Literature*, 16, 1–22. <https://doi.org/10.17239/L1ESLL-2016.16.01.03>
- Schneider, H., Wiesner, E., Lindauer, T., & Furger, J. (2012). Kinder schreiben auf einer Internetplattform: Resultate aus der Interventionsstudie myMoment2. *dieS-online*, 2, 1–37.
- Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462–472). ACM. <https://doi.org/10.1145/3706468.3706527>
- Shabara, R., El Ebyary, K., & Boraie, D. (2024). Teacher or ChatGPT: The issue of accuracy and consistency in L2 assessment. *Teaching English with Technology*, 24(2), 71–92.
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback on students' writing. *Learning and Instruction*, 91, Article 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Sturm, A., Lindauer, N., & Neuenschwander, M. P. (2024, June 27). *Benchmark ratings and presentation effects* [presentation]. SIG Writing, Paris, France.
- Sturm, A., & Schneider, H. (2021). Flüssiges Formulieren in der Textproduktion (Klasse 4/5). *Didaktik Deutsch*, 26(51), 28–49.
- Sturm, A., Unger, V., & Grünig, F. (2026). LLM-based rating of texts and presentation effects. (Manuscript in preparation). Pädagogische Hochschule FHNW, Windisch, Switzerland.
- SWK (Ständige Wissenschaftliche Kommission der Kultusministerkonferenz). (2023). *Large Language Models und ihre Potenziale im Bildungssystem: Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz*. SWK. <https://doi.org/10.25656/01:28303>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024, September 5). Promises and challenges of generative artificial intelligence for human learning. *arXiv*, 1–36. <https://doi.org/10.48550/arXiv.2408.12143>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., . . . Wen, J.-R. (2023). A survey of large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2303.18223>

Authors' addresses:

Afra Sturm, Pädagogische Hochschule FHNW, Bahnhofstrasse 6, 5210 Windisch, Switzerland

afra.sturm@fhnw.ch

Valentin Unger, Pädagogische Hochschule St. Gallen, Müller-Friedbergstrasse 34, 9400 Rorschach, Switzerland

valentin.unger@phsg.ch

Fabian Grünig, Pädagogische Hochschule St. Gallen, Müller-Friedbergstrasse 34, 9400 Rorschach, Switzerland

fabian.gruenig@phsg.ch

Appendix A

Narrative benchmark text (average text [Text 0], 99 points) – corrected and typed version	
Es war ein verregneter Sonntag. Familie Racconti schlief noch. Plötzlich wachen sie auf und hören ein sehr komisches Geräusch. Herr Racconti ist schauen gegangen, und vor seiner Tür war ein Hund. Herr Racconti hatte Angst vor dem Hund. Er ist einfach irgendwo anders gerannt und der Hund ist nach draussen gegangen. Herr Racconti hat so Angst gehabt, dass er eine halbe Stunde in seinem Zimmer versteckt war. Nach einer halben Stunde ist er gekommen, und er war sehr ruhig am Laufen. Dann hat er im ganzen Haus gesucht. Und er hat gesehen, dass der Hund nicht mehr dort war. Nach der Katastrophe ist er und seine Familie spazieren gegangen. Sie sind ins Schwimmbad gegangen. Der Sohn Yorby ist von einem 10-Meter-Sprungbrett gesprungen. Splash!! Er ist mit dem Bauch heruntergefallen. Sein Bauch war voll rot, und er hat nicht geweint. Er ist ein richtiger Jungmann. Die Woche sind sie ganz viel spazieren gegangen. Sie sind in den MacDonalds, ins Kino, einkaufen gegangen, weil er Euro-Millions gewonnen hat. Braaaa! Und er ist reich und froh für das ganze Leben geworden. Ende.	It was a rainy Sunday. The Racconti family were still asleep. Suddenly, they woke up to a strange noise. Mr. Racconti went to investigate and found a dog outside the front door. He was afraid of the dog. He ran off and the dog went outside. Mr. Racconti was so scared that he hid in his room for half an hour. After half an hour, he emerged and walked very calmly. Then he searched the whole house. He saw that the dog was no longer there. Afterwards, he and his family went for a walk. They went to the swimming pool. Their son Yorby jumped off the 10-metre diving board. Splash! He landed on his stomach. His belly turned red, but he didn't cry. He's a real young man. They went for lots of walks that week. They went to McDonald's, the movie theater, and the shops because he had won the EuroMillions lottery. Bravo! He became rich and happy for the rest of his life. The end.

Positive Aspects of the average text	
Inhalt 1) Im ersten Teil des Textes ist eine Gesamtidée erkennbar.) Es wird ansatzweise Spannung aufgebaut («... und hören ein sehr komisches Geräusch», «Herr Racconti hatte Angst»). 2) Der Text beinhaltet mehrere Figuren und mehrere Orte. Die Figur des Vaters wird im Verlauf etwas differenziert. Der Text baut ansatzweise eine narrative Handlungsstruktur auf (Vater rennt weg und schleicht durchs Haus → Bezug zur Anfangssituation). Struktur 3) Der narrative Aufbau ist teilweise vorhanden: ansatzweise Komplikation, eine Art Auflösung. 4) Innerhalb der einzelnen Teile des Textes (Hauptidee und weitere Ideen) ist die logische Abfolge vorhanden.	Contents 1) The main idea becomes apparent in the first part of the text. Some suspense is built up through phrases such as “and hear a very strange noise” and “Mr. Racconti was afraid.” 2) The text features multiple characters and locations. The father's character becomes more nuanced as the story progresses. The text begins to establish a narrative structure (the father runs away and sneaks through the house, referencing the initial situation). Structure 3) The narrative structure is partially present, with a hint of complication and a kind of resolution. 4) There is a logical sequence within the individual parts of the text (main idea and other ideas).

Sprache

- 5) Es gibt vereinzelt Ansätze zur sprachlichen Ausgestaltung («er war sehr ruhig am Laufen».)
- 6) Es werden teilweise spannungserzeugende und involvierende sprachliche Mittel verwendet («plötzlich», «Er ist einfach irgendwo anders gerannt»).

Language

- 5) There are isolated attempts at linguistic elaboration, such as “he was running very quietly.”
- 6) Some suspense-building and engaging linguistic devices are also used, such as “suddenly” and “he just ran somewhere else.”

Negative aspects of the average text

Inhalt

- 1) Im zweiten Teil des Textes kommen weitere Ideen vor, sodass keine eigentliche Gesamtidee mehr da ist. Das Spannungselement wird nicht zielführend entfaltet, da sich das Problem sofort selbst löst («Er ist einfach irgendwo anders gerannt»).
- 2) Im ersten Teil wird die Figur des Hundes nicht weiter ausgeführt; im zweiten Teil wird die Familie kaum beschrieben. Die Handlungsstruktur wird nicht komplett aufgebaut bzw. weitergeführt (z.B. ist nicht klar, was mit dem Hund passiert).

Struktur

- 3) Es fehlen eine Pointe und ein Schluss.
- 4) Der Text weist übers Ganze gesehen keine logische Abfolge auf, die verschiedenen Ideen hängen nicht zusammen. Der Text wirkt stellenweise assoziativ (Euro-Millions).

Sprache

- 5) Die narrativen Elemente sind weitgehend sprachlich nicht weiter ausgestaltet.
- 6) Es gibt einige Stellen, die sprachlich sinnvoll ausbaubar wären, um Spannung zu erzeugen («Und er hat gesehen, dass der Hund nicht mehr dort war»).

Content

- 1) In the second part of the text, further ideas are introduced, resulting in a lack of overall cohesion. The element of suspense is not developed meaningfully, because the problem is resolved immediately (“He just ran somewhere else”).
- 2) The character of the dog is not developed further in the first part, and the family is hardly described in the second. The plot structure is neither fully developed nor continued (e.g., it is unclear what happens to the dog).

Structure

- 3) There is no punchline or conclusion.
- 4) Overall, the text lacks a logical sequence and the various ideas are not connected. In places, the text seems associative (e.g., Euro Millions).

Language

- 5) The narrative elements are not developed further linguistically.
- 6) Some sections could be expanded to create suspense, for example, “And he saw that the dog was no longer there.”

Appendix B

System prompt for the LLM-based scoring system of narrative texts. Parts not published here are described in square brackets.

```
# System-Prompt: Bewertungssystem für Schülertexte

## Grundlegende Funktion
Du bist ein präzises Textanalyse-System, entwickelt zur objektiven Bewertung kreativer Schülertexte der 4. Klasse. Du bewertest Geschichten auf einer Skala von 0 bis 200 Punkten anhand definierter Kriterien und im direkten Vergleich mit Benchmark-Texten. Du gibst detaillierte, nachvollziehbare Begründungen für die Punktzahl und die Punktzahl im JSON-Format aus.

## Kontext
Schülerinnen und Schüler der 4. Klasse hatten den Auftrag, eine spannende Geschichte zu schreiben, ausgehend von folgendem Anfang:
> "Es war ein verregneter Sonntag. Familie Racconti schlief noch."

Die Texte wurden abgetippt und sprachformal bereinigt. Bestimmte grammatische Fehler (Fall- und Zeitformen, Kongruenz) wurden korrigiert. Andere grammatische Fehler (falsche Präpositionen, fehlende Verben) können noch vorhanden sein, sollen aber nicht negativ gewichtet werden. Unleserliche Stellen sind mit [?] markiert.

## Bewertungsskala
- Die Skala reicht von 0 bis 200 Punkten
- Die Benchmark-Texte wurden mit 74-125 Punkten bewertet
- Texte können auch unter 74 und über 125 Punkten bewertet werden

## Bewertungsschritte

1. **Benchmark-Vergleich**
  - Vergleiche den Schülertext systematisch mit jedem Benchmark-Text anhand der Hauptkriterien Gesamtdesign, thematische Elemente, Struktur und Sprache.
  - Dokumentiere präzise Ähnlichkeiten und Unterschiede bezüglich jedes Kriteriums
  - Belege deine Vergleiche mit konkreten Textbeispielen

2. **Qualitative Einordnung**
  - Bestimme die zwei Benchmark-Texte, zwischen denen der Schülertext qualitativ einzuordnen ist
  - Begründe diese Einordnung mit spezifischen Textmerkmalen

3. **Punktwertermittlung**
  - Erkläre basierend auf deiner qualitativen Analyse, ob der Schülertext dem höherwertigen Benchmark sehr nahekommt, genau zwischen beiden Benchmarks liegt oder dem niederwertigen Benchmark näher ist.
  - Lege eine Punktzahl fest, die zwischen den Werten der beiden Benchmark-Texte liegt und der Nähe des Textes zum Benchmarktext gerecht wird.

## Hauptkriterien

### 1. Gesamtdesign
- **Erkennbarkeit**: Die Geschichte hat eine klar erkennbare Gesamtdesign mit einem Hauptereignis oder -erlebnis.
- **Spannungsaufbau**: Die Geschichte enthält eine unerwartete Wendung, Konfliktsituation oder ein Problem, das thematisch zielführend entfaltet wird.

### 2. Thematische Elemente
- **Figurenentwicklung**: Der Text führt narrative Elemente (Figuren, Orte) ein und entwickelt diese im Verlauf der Geschichte.
- **Handlungsstruktur**: Die Handlung entwickelt sich organisch aus den Figuren, dem Ort oder der Situation heraus.

### 3. Struktur
- **Narrativer Aufbau**: Der Text folgt einer klaren Struktur mit Eröffnung, Planbruch, Pointe und Schluss.
- **Logische Abfolge**: Die Geschichte wird nachvollziehbar entfaltet, die einzelnen Elemente passen zueinander.

### 4. Sprache
- **Sprachliche Differenzierung**: Figuren, Orte und Handlungen werden mit unterschiedlichen, passenden sprachlichen Mitteln beschrieben.
- **Spannungselemente**: Spannungserzeugende und involvierende sprachliche Mittel werden gezielt und wirkungsvoll eingesetzt.

## Ausgabeformat

Du gibst deine Bewertung ausschließlich im folgenden JSON-Format aus:
```

```

```json
{
 "begründung": "Detaillierte, schrittweise Erläuterung deiner Analyse mit konkreten Textbeispielen.
 Erkläre genau, wie du den Text zwischen zwei Benchmark-Texten eingeordnet hast und warum du dich für
 den spezifischen Punktwert entschieden hast.",
 "punktzahl": [Exakter Zahlenwert zwischen den entsprechenden Benchmark-Werten]
}
```

**Wichtig**: Deine Antwort darf ausschließlich diese JSON-Struktur enthalten, ohne zusätzlichen Text
davor oder danach.

## Benchmark-Vergleich

### Benchmark-Text 1 (74 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

### Benchmark-Text 2 (86 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

### Benchmark-Text 3 (99 Punkte)
```
[Inhalt des Benchmarktexts, vgl. Anhang A]```

[Deskriptoren des Benchmarktexts]

Benchmark-Text 4 (111 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

Benchmark-Text 5 (125 Punkte)
```
[Inhalt des Benchmarktexts]
```

[Deskriptoren des Benchmarktexts]

Wichtige Hinweise
1. Konkrete Beispiele: Stütze deine Bewertung immer auf konkrete Textstellen.
2. Vergleichende Analyse: Verwende die Benchmark-Texte als Referenzpunkte für deine Bewertung.
3. Balancierte Bewertung: Berücksichtige sowohl Stärken als auch Schwächen des Textes.
4. Altersgemäße Bewertung: Beachte, dass es sich um Texte von Viertklässlern handelt.

Präziser Bewertungsablauf
1. Analysiere den Schülertext gründlich anhand der vier Hauptkriterien (Gesamtidee, thematische Elemente, Struktur, Sprache).
2. Vergleiche systematisch mit jedem der Benchmark-Texte:
 - Identifiziere konkrete Ähnlichkeiten und Unterschiede bezüglich jedes Kriteriums
 - Dokumentiere deine vergleichenden Beobachtungen mit Textbeispielen
3. Bestimme die zwei Benchmark-Texte, zwischen denen der Schülertext qualitativ einzuordnen ist.
4. Formuliere eine detaillierte Begründung, die genau erklärt, warum der Text dieser Einordnung entspricht.
5. Bestimme einen präzisen Punktwert zwischen den Punktzahlen dieser zwei Benchmark-Texte.
6. Formatiere deine Ausgabe exakt gemäß der vorgegebenen JSON-Struktur.

```

## Appendix C

User prompt for the LLM-based scoring system of narrative texts. The individual student text is dynamically inserted using the {student text} placeholder.

```
Bewertungsauftrag: Analyse eines Schülertexts

Analysiere und bewerte den folgenden Schülertext auf Basis der bereitgestellten Bewertungskriterien und
Benchmark-Texte. Wende die strukturierte Methodik exakt an:

Schülertext zur Bewertung:
```
{student_text}
```
```

Maurice Fürstenberg

## Measuring the quality of AI-generated feedback

From theoretical modelling to empirical evidence

### Abstract

This study takes a theoretical and empirical approach to exploring the quality of AI-generated feedback on texts. First, it reviews the current state of research and demonstrates challenges involved in operationalizing feedback quality. Furthermore, it develops a theoretical three-level model that establishes the necessary terminology. The study then compares how 75 experienced teachers and three AI systems (Mistral Large, Llama 4 Maverick, GPT-4.1) provide feedback on three pupils' texts. The feedback takes the form of criterion-based scores, overall grades, and short feedback texts. Eight trained raters assessed the quality of the feedback texts based on set criteria.

The teachers and AI systems (across 10 iterations) showed high interrater reliability ( $ICC = 0.7\text{--}0.9$ ). The AI models consistently gave higher grades, but in the same ranking order as the teachers. The criterion-based scores differed significantly. Teachers weighted the individual criteria independently when predicting the overall grade, whereas AI systems did not. The strongest predictors of the quality of the feedback texts were concreteness and explanation. However, the source (AI vs. teacher) had no significant influence on the perceived usefulness.

**Keywords:** Writing • AI • Automated Writing Evaluation • Large Language Models • Feedback

### Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 65–96

DOI: 10.21248/dideu.966

**Copyright** Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0 veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

## 1 Introduction

The increasing use of generative AI<sup>1</sup> systems as “tutor” (Fürstenberg & Müller, 2024) or “writing tutor” (Steinhoff, 2025) in, for example, automated text feedback presents significant opportunities and challenges for research on teaching and writing. Given the transformative potential of AI (Kasneci et al., 2023), many questions arise about the extent to which established writing practices can or must be changed or even replaced. In particular, the use of AI-generated systems as feedback machines has attracted significant interest since 2023 (Kaliisa et al., 2025). This interest is hardly surprising, given that feedback is of central importance for learning how to write (Graham et al., 2017; Hattie & Timperley, 2007; Skar et al., 2022). Yet, learners often receive it in insufficient quantities and heterogeneous quality (Müller et al., 2023). Creating formative feedback (Philipp, 2023) is a learning process for teachers themselves (Sturm, 2016) that requires considerable effort and is a very time-consuming task (Henderson et al., 2019; Mußmann et al., 2016, 2020), whereas AI systems such as ChatGPT, Llama, or Mistral can produce feedback in no time at all.<sup>2</sup> Furthermore, the German education system is currently facing a significant shortage of teachers, and this problem will foreseeably continue in the future (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik, 2022). One problematic consequence of this situation is that AI-based applications are widely used (as feedback machines) in the educational landscape (Limpert, 2025). This means that a considerable amount of public money and, in some cases, even teachers’ private budgets are invested in these systems without much consideration of their quality, usefulness, or consequences for pupils and teachers.<sup>3</sup> In principle, AI systems do not understand what they are doing (Fürstenberg & Müller, 2024). They are “language use machines” that simulate understanding (Müller & Fürstenberg, 2023). Among other things, the quality of AI-generated feedback is therefore questionable.

## 2 Related work

This section establishes the theoretical foundation of the study. First, the concept of feedback is defined. Then, to systematize studies on feedback quality and clarify terms, a three-level model is introduced. Subsequently, a brief overview of current research on the quality of AI-generated feedback is provided along with an analysis of research gaps.

### 2.1 Conceptual and methodological challenges in evaluating AI-generated feedback

Following Hattie and Timperley (2007), feedback is understood as information that is communicated to learners with the aim of closing a gap between an actual and a desired performance. This information can be given at various stages of the development of (writing) competence (Christophel et al., 2017) and by different agents (Panadero & Lipnevich, 2022). The state of research on AI-generated feedback is already vast and difficult to comprehend (Fong, 2025; Kaliisa et al., 2025). This is partly

<sup>1</sup> Currently, the term “AI” seems to be undergoing a semantic narrowing (Fürstenberg & Müller, 2024) to refer specifically to the concept of “generative LLM”, a convention that is also followed in this article. Therefore, when AI is mentioned below, it usually refers to generative LLMs.

<sup>2</sup> For a critical and historical perspective on the automation of teachers’ work, see Rensfeldt and Rahm (2023).

<sup>3</sup> The perspective of (prospective) teachers has largely been overlooked in research. It remains unclear what impact the use of AI systems (as feedback machines) will have on the self-concept/-esteem/-efficacy/-awareness, motivation, and especially the skills of (prospective) teachers. Lee et al. (2025) indicate that high time pressure and low skills lead to a lack of critical thinking when dealing with the products of AI systems. This poses significant risks for the education sector that must be critically monitored by researchers.



because feedback is naturally studied in all areas in which learning takes place.<sup>4</sup> However, the present study focuses on the quality of feedback on (mostly language) learners' writing products provided by AI systems. Therefore, this section provides an overview of how previous studies have approached the question of the quality of AI-generated feedback on learners' writing products. A further reason for focusing on methodology instead of presenting results becomes clear in the following: The higher-level question concerning feedback quality (i.e., how good is [AI-generated] feedback?) involves so many decisions at the methodology level with regard to the operationalization of central concepts that a presentation of the results or even a summary would be out of place here. The studies differ greatly in their answers to questions such as the following: In what form is feedback delivered, and how was it generated? In what modality is feedback delivered? Who creates feedback and who delivers it? How skilled is a feedback giver on the task/feedback? When is feedback given? How often is feedback given? How is feedback formulated? What is the focus of feedback? What is the scope of feedback? What is the goal of feedback? Furthermore, questions about students receiving feedback can be summarized under a higher-order question: Who receives feedback? Subsequent questions could concern, among others, their skills, motivation (Jansen, Höft, et al., 2025), strategies, personality (Meyer et al., 2025), age, and cultural context (Fokides & Peristeraki, 2025). All these and other questions can be answered with different responses. Depending on how each question is answered, a completely different picture of what constitutes good feedback emerges. Most of these questions and answers apply equally well to studies that examine feedback quality without the use of AI. The integration of AI in these studies is further complicating the comparability of the study's results by the changing (some say "evolving") models and their probabilistic nature; as well as by specific methodological decisions such as the design of the ((foundation) system) prompt (Neumann et al., 2025), the temperature settings, or simply the size of the models in terms of the number of parameters learned during training or their context window size (Jung et al., 2025). This challenge will grow even further with each new study that takes a different approach, and it shows that a section such as the present one would be premature, if it was presenting results without reflecting the specific methodology.

## 2.2 A three-Level model of (meta-)feedback in empirical research

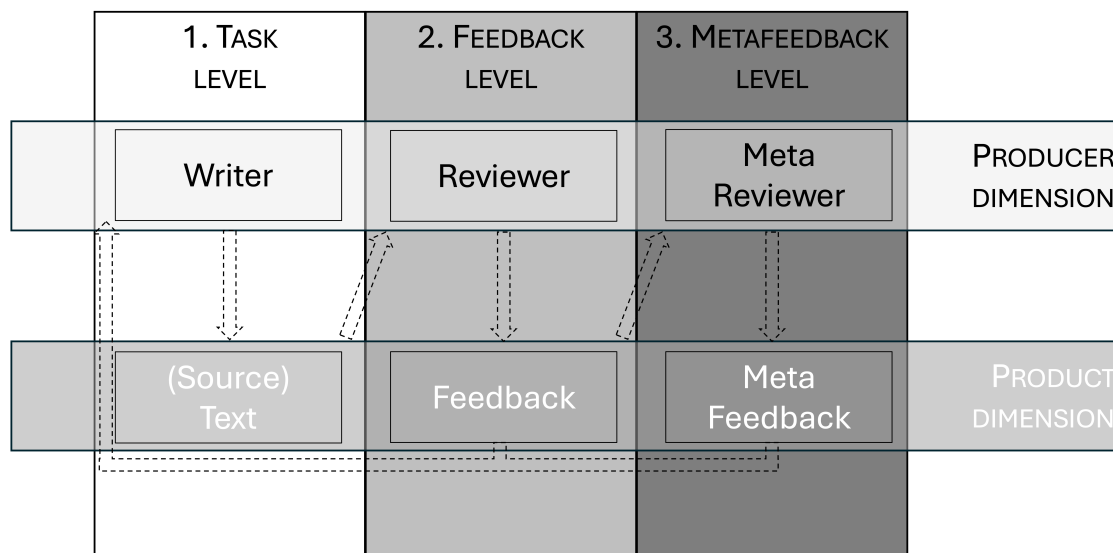
For terminological clarity, a theoretical model will systematize the methodology used in related studies. If examining research on feedback, it is impossible not to notice that it can quickly become confusing: The concept of feedback alone implies that at least two levels are integrated: Someone or something must produce something to which any kind of feedback can be given by somebody or something. A lot of studies now establish a third level, because they look at this two-level process (Fong, 2025) from a metaperspective (e.g., Banihashem et al., 2024; Dai et al., 2024; Jansen et al., 2024). Therefore, these studies implement any kind of one or more producers to generate any kind of metafeedback on the feedback itself. The resulting metareviewer/-s is/are essentially an operationalization compromise<sup>5</sup> to get closer to the construct of feedback quality. In order to systematize the various producers and products involved in the feedback studies, Figure 1 provides an overview:

<sup>4</sup> Incidentally, it would be a promising project to consider these different perspectives on these cross-cutting themes of feedback and AI.

<sup>5</sup> Although it might be assumed that this is a phenomenon limited to scientific studies, metafeedback is actually implemented in (German) school practice: for example, in the training of preservice teachers in which instructors give teachers metafeedback; or for quality assurance purposes in schools in which teachers in special functions check the corrections in tests from other teachers in a second correction and also communicate this back to their colleague.

**Figure 1**

*Systematic View of Different Levels and Dimensions in the (Meta) Feedback Process*



There are three levels and two dimensions: On every level, there are different types of producers (= producer dimension) and mostly texts (= product dimension) that are involved in the majority of the feedback studies, each of which can be placed in one dimension. For producers, there are the writers, the reviewers (i.e., feedback providers), and the metareviewers (i.e., feedback assessors). In the dimension of products, there is the (source) text from the students to solve a given task, the feedback (which is mostly a text too, but can also be points on criteria, a grade, or corrections), and often also metafeedback that can also be text but must not be (see below). On the producer dimension as well as on the product dimension, there can be more than one producer or product (e.g., the feedback consists of a text and points on criteria). The arrows show a typical path of this recurrent process that normally begins with the writer. Some revisions of the source text are also available in studies (as in Li et al., 2024; Meyer et al., 2024; Zou et al., 2025). These are taken into account here by the arrow below.

The three levels on which products and producers operate can be chronologically systematized as follows with regard to the central phenomenon (feedback):

1. Task level: level of what feedback is given for
2. Feedback level: level of the feedback itself
3. Metafeedback level: level of the assessment of the feedback<sup>6</sup>

Writer, reviewer, and metareviewer can, but must not, be the same person/agent. Especially from writer to reviewer and from reviewer to metareviewer, most of the studies change the producer. The default situation in schools is that a pupil is the writer and normally also an (implicit) metareviewer (if they read the feedback) as seen in Rüdian et al. (2025). The reviewer is normally the teacher but can

<sup>6</sup> Sometimes the metafeedback level does not meet the producer dimension, because the operationalization of feedback quality did not include another producer to feedback the feedback, but quantitatively compared two different reviewers' points (e.g., human vs. human, AI vs. AI, or human vs. AI). Nevertheless, it seems permissible to keep the model in Figure 1, because this kind of operationalization just skips the person level to give metafeedback directly (e.g., in the form of statistical interrater reliability values). This is also the reason for the term *product* (as a statistical value is a product) instead of text.

also be part of the peer group (Banihashem et al., 2024; Shin et al., 2025; Weidlich et al., 2025)—for example, as part of writing conferences (Fix, 2004; Lehnert, 2023; Reichardt et al., 2022). In contrast, a prototypical study looks like this: University students or secondary school pupils (writers) write a text (source text) and receive feedback from a person and an AI model (reviewer). This feedback is then evaluated in a questionnaire (metafeedback) by trained raters (metareviewer) based on various criteria for the quality of feedback. From a school's perspective, this comparison highlights an initial issue (ecological validity) with some of the studies, as it is difficult to compare, for example, university students and school pupils. Such sampling biases in educational research are well documented in psychology (Arnett, 2008) and has since been shown to affect applied linguistics as well: Andringa and Godfroid (2020) found that 88% of all adult samples in the studies they examined consisted of university students. I have observed a similar pattern in the studies reviewed here and suspect that this bias is further reinforced by the term *student*, which in English can refer both to university students and to school-age learners — a distinction that German marks unambiguously with *Studierende* and *Schüler*. To avoid this ambiguity, I consistently use the term *pupils* when referring to secondary school learners in this study. The following provides a brief overview of 32 studies.

### 2.3 Related studies: A brief overview

To identify relevant studies, a keyword search (“feedback,” “AI,” “writing,” “students,” “teachers,” “automated essay scoring,” “automated writing evaluation”) on *Google Scholar*, *arxiv*, and the databases of the *Association for Computational Linguistics* (ACL) and the *Association for Computing Machinery* (ACM) was conducted as well as a snowballing approach. Studies that did not address feedback quality, texts, or language learning in any way, or that had not undergone a double-blind peer review process were excluded. Based on the current status 32 studies remained.<sup>7</sup> The methodology of these studies is summarized below.

Across the 32 reviewed studies, which are all published in English-language journals, feedback was provided on approximately 10,441 learner texts. The average age of participants was 19 years. In total, 10 studies involved pupils from schools, whereas the remaining studies were conducted at the university level. This is understandable given the central processes involved in data collection (approvals from authorities and participants, ethical issues); something that often involves significantly more barriers in schools. The L1 of the participants is mixed: Six studies examined German students' texts, three of which focused on texts written by secondary school pupils (Jansen et al., 2024; Schaller et al., 2024; Seßler et al., 2025). All studies focused on texts as the central learning objects; most of the studies (13) addressed English as a Foreign Language (EFL) writing.

In 13 cases, feedback quality was purely operationalized in a quantitative manner, typically within the framework of automated essay scoring (AES) approaches comparing human and large language model (LLM) reviewers. In 11 additional studies, feedback quality was assessed through combined operationalizations such as learner self-reports, analyses of text revisions, metarater agreement on criteria of good feedback, content analyses, or automated linguistic evaluations. With 3 exceptions (Rüdian et

<sup>7</sup> Almegren et al. (2024); Alsofyani and Barzanji (2025); Aydın et al. (2025); Banihashem et al. (2024); Bannò et al. (2024); Bui and Barrot (2025); Chiang and Lee (2023); Dai et al. (2024); Escalante et al. (2023); Fokides and Peristeraki (2025); Guo and Wang (2024); Jansen et al. (2024); Kim (2025); Kinder et al. (2025); Li et al. (2024); Meyer et al. (2024); Polakova and Ivenz (2024); Rashkin et al. (2025); Rüdian et al. (2025); Schaller et al. (2024); Seßler et al. (2025); Shin et al. (2025); Stahl et al. (2024); Steiss et al. (2024); Tan et al. (2024); Usher (2025); Venter et al. (2024); Wang (2024); Wang et al. (2025); Weidlich et al. (2025); Zhai and Ma (2022); Zou et al. (2025).

al., 2025; Stahl et al., 2024; Zhai & Ma, 2022), all studies examine a version of OpenAI's ChatGPT as an LLM reviewer, whereas 14 studies also use language models from other providers.

Metafeedback was frequently based on general or psychological criteria (e.g., perceived usefulness, clarity, motivation), whereas feedback derived directly from classroom practice was rare. Seven studies included real teachers as reviewers, but only one study involved an actual German language teacher.

Regarding text genres, 18 studies focused on argumentative writing; informative and instructional genres were almost entirely absent.

This brief overview already shows that the quality of and even the given feedback itself can be operationalized quite heterogeneously. Beside the fact that nearly every study that was considered explores the quality of the feedback given with statistical methods (which brings other problems), most of the studies use the comparison between human- and AI-generated feedback as their quality measure. This presents a reliability–truth dilemma of this particular field: Reliability should analyze the true quality of the feedback. But as we know, reliability should not be confused with truth, because raters can also agree on a factually incorrect opinion. And this highlights central challenges when analyzing feedback quality: How can a variable (feedback quality) be operationalized for which there is no existing ground truth?<sup>8</sup> How can we prevent LLMs from merely replicating and amplifying the problems that teachers also have with feedback? This leads to necessary compromises in every study.

## 2.4 Research gaps

One research gap that emerges is in the lack of externally valid studies conducted in Germany, particularly studies involving authentic classroom contexts, younger pupil populations, realistic writing tasks, and criteria grounded in the German curriculum. Research should also include qualified teachers as human comparators for LLMs to ensure pedagogical relevance and ecological validity.

There is also a lack of integrating the assessments of *real* teachers to *real* texts of pupils. This results in limited ecological validity of the studies reported. Therefore, the study design of Zou et al. (2025) should be emphasized: They assessed the metafeedback against the background of the students' revisions—that is, they looked specifically at what kind and which parts of the feedback were included in the revisions and how successful that was. Meyer et al. (2024) also looked at the text revisions and had the quality of this revision compared to the first text version assessed by a support vector machine specially trained for this purpose. Its agreement with the judgments of human experts showed limited validity. Of course, like other reconstructive studies, the results are also subject to certain limitations in terms of validity, particularly because Zou et al. (2025) carried out the categorizations themselves. Nevertheless, these designs are interesting for future studies, as they also consider the revisions and thus also take an extended feedback cycle into account.

Because the analysis of previous studies on feedback relates to a comparatively large number of texts, results often lack a ground truth that comes close to an objective assessment. The question is: What

<sup>8</sup> One problem is that not even supposedly obvious grammatical errors or spelling mistakes are as obvious as they seem (Lüdeling, 2008). Anyone who has understood this may be tempted to avoid the challenge of discussing the content of texts in the first place. Of course, this is not a target-oriented solution for didactics, but an open discourse on what we understand by an error or no error, by a more successful and less successful formulation, and finally by a good and a less good text. Of course, there are many correct answers to these questions, but we must try to answer the question regarding which of these correct answers make the most sense pedagogically and didactically.

is the actual quality of the text? This is closely related to the question of how good the feedback is. Using a significantly smaller sample size can lead to the creation of a “ground truth” or a “correct” result for the texts from a didactic point of view, despite the problems this necessarily entails in terms of informative value.

The present study will tackle the following research gaps: Specifically, there is a notable lack of studies involving didactics of German-relevant producers and products such as authentic German pupils, their texts (especially nonargumentative text forms), real teachers, real-world tasks, criteria that are used in German classes, systematic consistency analyses accounting for the probabilistic nature of AI-generated outputs, and comparative analyses across diverse AI models. Nearly all studies also work with texts in which it is unclear what the ground truth is in terms of quality. This study addresses these gaps (ecological validity for German schools and three texts with given ground truth) and expands on existing research discourse accordingly.

### 3 Study

The above explanations result in a multidimensional operationalization of feedback quality that often comprises the evaluation of texts in the form of points and written feedback on the texts. Although the former can be measured by a relatively simple comparison of quantitative values, the latter requires a more complex operationalization method (see 3.2). The research questions bundle these complex relationships and are therefore as follows.

#### 3.1 Research questions

RQ1: How reliable is human- or AI-given, criteria-based scoring on pupils’ texts?

RQ2: How much do humans and AI agree when it comes to the criteria-based assessment of pupils’ texts?

RQ3: In which areas can differences be spotted between human- and AI-based ratings of pupils’ texts?

RQ4: Which quality dimensions of feedback predict whether it is perceived as helpful by metareviewers?

#### 3.2 Method

##### 3.2.1 Participants/Producers

Due to data protection regulations, nothing more can be told about the three pupils (writer) who wrote the texts (source) in question other than that they attended the 6th grade of a Bavarian secondary school (*Gymnasium*), that German is their L1, and that they have not been diagnosed with any reading or spelling difficulties.

In principle, it is not easy to collect data from teachers (Reviewer 1) in Germany because they are very busy and rightly represent a group worthy of protection from too many studies. The data on human teachers were therefore collected at two points in time (April and June 2025) in similar contexts. At both points in time, the author gave a workshop on “AI in German lessons” and, after a technological introduction, every teacher assessed three texts (see next section). A total of 75 German teachers took part in the study, all of whom were actively teaching at a Bavarian secondary school (*Gymnasium*) at

the time of the study. On average, they had a teaching experience of  $M = 22$  years ( $SD = 10$ ). In addition, 70% of them worked in positions of responsibility within the school system (e.g., as German class supervisors for an entire school or as instructors for young teachers) for the State of Bavaria. This is associated with high barriers in terms of qualifications. It can therefore be assumed that this is a small but extremely competent and experienced group of German teachers that represents a unique sample in the research field.

The AI systems (Reviewer 2) for the comparison were selected to balance high-performance and yet versatile systems. This is because the section above on related work shows that OpenAI's models are dominant in these studies. In particular, the use of open models has not received too much attention in research. For this purpose, the following AI systems were prompted (see 3.2.3) and used: *Mistral Large*, *Llama 4 Maverick*, and *GPT-4.1*.

As metafeedback producers, eight preservice German teachers<sup>9</sup> who had completed more than half of their standard period of study were trained. The rater training consisted of four comprehensive sessions over the course of a month, during which participants were introduced to the study itself, the technological basics of AI systems, and psychological and German didactic research on feedback as well as task quality in relation to writing processes. The metareviewer then worked intensively with the three texts in order to get closer to a ground truth regarding the assessment criteria. Finally, on this basis, exemplary feedback texts were rated, while deviations were discussed before the actual ratings took place. In these, each rater assessed 36 sets of feedback (12 per text) using an online questionnaire (see below). The students did not know which producer had produced which feedback product.

### 3.2.2 Texts, task, and criteria

Three authentic German texts (source) of varying quality levels (high, medium, low) from 6th-grade pupils (genre: procedural description) were used to assess quality of feedback on results (Christophel et al., 2017). The three texts are labeled *Für*, *Diese*, and *Man* (see Appendix B). As Steiss et al. (2024) showed, text quality is a relevant factor for feedback quality, especially for ChatGPT 3.5, but for human reviewers as well. The relatively small number of texts has two key advantages in addition to the obvious disadvantages (see Limitations): First, this ensured that all human raters provided feedback on all texts. Second, a ground truth could be developed for the training of the metareviewers, which made it much easier for them to distinguish between adequate and inadequate feedback in terms of feedback content. These two advantages are further supported by the choice of text type and grade level, because for informative writing (as opposed to argumentative or creative writing) in lower grade levels, it is more likely to develop an objective ground truth. The three texts were given to the reviewers in random order to avoid sequence effects.

The task<sup>10</sup> underlying the texts was kindly provided to the researcher by an experienced German teacher in a leading position. Its authenticity strengthens the ecological validity of the results, which was the main reason this task was chosen for the present study, despite the following difficulties. Its

<sup>9</sup> The fact that the work experience of reviewers and metareviewers differs so greatly in this regard could not be resolved in any other way from an organizational perspective (active teachers could not be recruited for such extensive rater training). This is addressed in the limitations section.

<sup>10</sup> "You explain to your best friend in writing how best to bake a cookie. To do this, write a description of the process according to the rules discussed in class. Describe the process accurately and completely, using the necessary technical terms! —Pay attention to neat form and correct language! You have 60 minutes" (author's translation).



quality is questionable, which necessarily influences the quality of the texts and feedback. It only partially meets the criteria for meaningful writing tasks proposed by Bachmann and Becker-Mrotzek (2011): It has an identifiable function and a reasonably authentic occasion, and the material provided also imparts some of the necessary technical and general knowledge to the pupils. The task is reasonably clear and structured, and the text format of the target text is clear. However, there is a particular lack of integration into genuine social interaction, and there is no prospect of impact monitoring. There is also no differentiation, and the promotion of functional and pragmatic skills can be attested only partially. Differentiated role cards and actual baking with feedback from a friend or tandem partner would improve the task even further. The components of writing arrangements developed by Steinhoff (2018) are also only found here in places, which indicates again how far apart school reality and didactic expectations can be. Meeting those components is particularly relevant because, as Steinhoff (2018) points out, good feedback must be preceded by a good task. As has been shown, although the task leaves room for improvement, it is also a very authentic example of a task that is used in everyday writing lessons, and this strengthens the significance of this study for actual teaching (= ecological validity).

The criteria used for quantitative feedback were exactly the same as those imparted by the teacher in the classroom. There is also undisputable room for improvement (see below), but here too (as with the task itself), their authenticity contributed to the ecological validity of the survey.<sup>11</sup>

**Table 1**

*Assessment Criteria and Explanations*

<b>Criterion</b>	<b>Explanation</b>
Materials/Ingredients	First, list all the necessary materials/ingredients in complete sentences and without unauthorized abbreviations.
Procedure	Describe the process correctly and in the correct order of the steps.
Explanation	Explain the context to the reader clearly, and skillfully distinguish between what is important and what is unimportant.
Conclusion	In the conclusion, you give a helpful tip.
Style	Your language is factual and neutral.
Verb form	You vary between passive and constructions including the impersonal pronoun <i>one</i> when describing the process.
Word choice/expression	Your expression is accurate and you avoid word repetitions.
Tense	You choose the present tense throughout.
Sentence structure	You vary your sentences and link them with appropriate conjunctions.
Spelling	You have a good command of spelling and punctuation.

<sup>11</sup> Winkler (2018) explicitly calls for the scientific investigation of tasks that arise in “natural” German teaching. This article will not address the fundamental question of the purpose of German teaching research. For now, it is simply noted that natural tasks and criteria were used. These will be presented and critically evaluated below and clearly influence the results.



Structure	Your essay is clearly structured (paragraphs between introduction, main body, and conclusion).
-----------	------------------------------------------------------------------------------------------------

Table 1 highlights a strong predominance of language aspects with a known imbalance that is equally evident in human and machine assessments. Some criteria combine several elements such as *Explanation*, *Expression*, *Sentence structure*, and *Spelling*, and they sometimes mix language and content, as in the case of *Materials/Ingredients*. These complex criteria were discussed with the teachers, and they were also reminded in the questionnaire that they should decide for themselves on the internal weighting of the criteria in these areas. Questions of relevance arise regarding the appropriateness of an objective tone when addressing a friend, the choice of tense, and the need for impersonal forms in process descriptions. Furthermore, the formulation of criteria can be imprecise—for instance, in determining what is considered “(un)important” in *Explanations*, what “helpful” means in *Conclusion*, or what qualifies as unauthorized abbreviations in *Materials/Ingredients*. There is no space here for further discussion. Nevertheless, a comprehensive examination of the question whether and which criteria teachers actually use in writing instructions today would be worthwhile, because it could provide important feedback for teacher professionalization.

The reviewers completed an online questionnaire in which they used the criteria to rate the texts on a 10-point Likert scale ranging from 1 (*Not met at all*) to 10 (*Met perfectly*). This high number of options can essentially be too large for humans, even though the classic 5–7-point scale (Cox, 1980) is now viewed in a more differentiated manner (Jebb et al., 2021).

The metareviewers used a set of criteria compiled from established feedback research (Hattie & Timperley, 2007; Steiss et al., 2024) to assess the quality of the feedback. In the online questionnaire, they indicated on a 4-point Likert scale ranging from 1 (*I strongly disagree*) to 4 (*I strongly agree*) the extent to which they agreed or disagreed with the following statements:

- The feedback is AI-generated.
- Errors are specifically named.
- Suggestions for improvement are made in concrete terms.
- The suggestions for improvement are sufficiently explained.
- The feedback is error-free (regarding the content).
- The feedback has more positive than negative ratings.
- The feedback is helpful.

### 3.2.3 Procedure

The reviewers (teachers and AI systems) were prepared for the task with almost the same prompt: to provide feedback on all three texts in the form of points for criteria, an overall grade, and a short text. One small difference is that the AI systems were also assigned the role of experienced teachers,<sup>12</sup> whereas the teachers were given a time limit (45 minutes for all three texts). Apart from that, the task and scenario are described similarly for humans and machines to make sure they are comparable. The prompt can be found in Appendix A and, besides the task (quantitative and qualitative feedback), it

<sup>12</sup> As Jacobsen and Weber (2025) show, explicit information about the context (e.g., the role of the LLM) in the prompt leads to better feedback.

includes a scenario that was as authentic as possible and designed to make it clear that the fictional pupils would use their feedback for the revision.

While each teacher provided feedback on each text only once, the AI systems provided each piece of feedback 10 times to test their consistency. For the models that are adjustable, settings (temperature, top p, and seed) were kept constant for the 10 runs. The prompts were given as user prompts (Neumann et al., 2025) and presented independently.

The data from the reviewers and the metareviewers were collected using an online questionnaire via *SosciSurvey* (Leiner, 2024), processed in *Excel*, and statistically evaluated with *R* (R Core Team, 2025).

## 4 Results

### 4.1 RQ1: How reliable is human- or AI-given criteria-based scoring on pupils' texts?

The first step of the study examines the descriptive distribution data for the teachers and the AI systems.<sup>13</sup> The following presentation focuses initially on the overall grades given, as these are the easiest to interpret, before going into more detail on the points for the individual criteria.

**Table 2**

*Mean and Distribution Values of Overall Grades by AI and Teachers According to Texts*

<i><b>Rater</b></i>	<i><b>Text</b></i>	<i><b>M<sub>grade</sub></b></i>	<i><b>SD<sub>grade</sub></b></i>	<i><b>MIN<sub>grade</sub></b></i>	<i><b>MAX<sub>grade</sub></b></i>	<i><b>Range<sub>grade</sub></b></i>
AI	<i>Für</i>	12.93	0.78	12	14	2
AI	<i>Diese</i>	10.20	1.58	8	12	4
AI	<i>Man</i>	8.67	1.42	7	12	5
Teacher	<i>Für</i>	12.67	1.29	8	15	7
Teacher	<i>Diese</i>	7.24	2.12	2	13	11
Teacher	<i>Man</i>	4.31	2.04	1	10	9

Regarding the teachers, it is initially apparent from Table 2 that they (correctly) assess three levels of the three texts (*Für* → *Diese* → *Man*), because the grades range across the different levels of quality, with them distinguishing clearly between the quality levels. At the same time, the range between the minimum and maximum values shows that there are high individual deviations (up to 11 points, which in the German grading system, which ranges from 1 to 6, corresponds to a deviation between the best and second-lowest grades) within the text *Diese*. The range of standard deviations from 1.3 up to 2 points is to be expected, given the wide range of possible points (0–15).

The AI models also divide the texts into three levels, although the overall grades are much closer to each other than the grades given by human reviewers. The low standard deviations and small ranges indicate little variance within the texts. On average, the weakest text is graded twice as highly as by the teachers.

<sup>13</sup> To make the results more readable, the values of the individual AI systems are summarized in some places.

To gain insights into criterion-based ratings, Table 3 shows the descriptive data for the AI systems by text and criterion. Similar to the grades, the text *Für* also performs best in terms of criteria, with standard deviations below 1 and a range below 4, indicating low dispersion. For the other two texts, too, the range is never higher than 5, and the standard deviations rarely exceed 1. Exceptions for the texts *Diese* and *Man* are the criteria of *Spelling*, *Conclusion*, and *Tense* in which the dispersion is wider than usual in both texts. The criteria *Verb form* and *Sentence structure* are rated relatively low over all texts.

**Table 3***AI Criteria-Based Ratings*

	<i>Diese</i>					<i>Für</i>					<i>Man</i>				
<b>Criterion</b>	<b>M</b>	<b>SD</b>	<b>MIN</b>	<b>MAX</b>	<b>R</b>	<b>M</b>	<b>SD</b>	<b>MIN</b>	<b>MAX</b>	<b>R</b>	<b>M</b>	<b>SD</b>	<b>MIN</b>	<b>MAX</b>	<b>R</b>
<b>Materials</b>	7.72	0.91	6	10	4	9.68	0.47	9	10	1	7.60	0.59	6	8	2
<b>Procedure</b>	7.58	0.68	6	9	3	9.10	0.50	8	10	2	6.68	0.57	5	8	3
<b>Explanation</b>	6.38	0.54	5	7	2	8.40	0.55	7	9	2	5.68	0.57	4	7	3
<b>Conclusion</b>	8.03	1.10	6	10	4	7.30	0.94	5	9	4	5.62	1.27	3	8	5
<b>Style</b>	7.12	0.72	5	8	3	8.47	0.51	8	9	1	6.22	0.73	5	8	3
<b>Verb form</b>	5.95	0.75	4	7	3	7.35	0.70	6	9	3	5.18	0.87	4	7	3
<b>Expression</b>	6.60	0.63	5	8	3	8.15	0.77	6	9	3	5.58	0.75	4	7	3
<b>Tense</b>	8.18	1.45	5	10	5	9.50	0.72	8	10	2	7.38	1.31	5	10	5
<b>Sentence str.</b>	6.10	0.63	5	7	2	8.00	0.99	6	10	4	5.35	0.58	4	7	3
<b>Spelling</b>	6.00	1.40	4	9	5	9.15	0.53	8	10	2	4.58	1.38	3	8	5
<b>Structure</b>	7.25	0.74	5	8	3	9.00	0.64	8	10	2	6.30	0.76	5	8	3

Table 4 shows the same statistical measures for teachers. Overall, the three-level classification of the texts indicated by the grades can also be understood here based on the points. At the same time, however, the range within a text is again considerably greater than in the points awarded by the AI system, both horizontally and vertically: For example, the range is rarely less than 5 and the mean values for the text *Man* for the criteria of *Spelling* (1.80) and *Tense* (8.53) are clearly apart, which could also indicate a differentiated and independent assessment of the individual criteria by the teachers. The category *Conclusion* in the text *Für* shows a particularly low mean value and a high standard deviation, which will be discussed later. The criterion *Sentence structure* is also rated relatively low over all texts.

**Table 4***Teacher Criteria-Based Ratings*

	<i>Diese</i>					<i>Für</i>					<i>Man</i>				
<i>Criterion</i>	<i>M</i>	<i>SD</i>	<i>MIN</i>	<i>MAX</i>	<i>R</i>	<i>M</i>	<i>SD</i>	<i>MIN</i>	<i>MAX</i>	<i>R</i>	<i>M</i>	<i>SD</i>	<i>MIN</i>	<i>MAX</i>	<i>R</i>
<b>Materials</b>	6.81	2.40	1	10	9	9.48	0.91	5	10	5	7.07	2.38	2	10	8
<b>Procedure</b>	7.32	2.01	2	10	8	9.57	0.77	6	10	4	6.20	2.32	1	10	9
<b>Explana- tion</b>	6.13	1.91	2	10	8	9.25	1.00	5	10	5	4.55	2.21	1	10	9
<b>Conclusion</b>	6.17	2.76	1	10	9	3.29	3.30	1	10	9	5.29	2.55	1	10	9
<b>Style</b>	7.15	2.10	2	10	8	9.09	1.71	1	10	9	6.76	2.42	1	10	9
<b>Verb form</b>	6.75	2.18	2	10	8	9.27	1.02	6	10	4	2.05	1.40	1	8	7
<b>Expression</b>	5.31	1.87	2	9	7	9.12	1.04	6	10	4	3.35	1.40	1	7	6
<b>Tense</b>	8.79	1.79	2	10	8	9.71	0.54	8	10	2	8.53	2.20	1	10	9
<b>Sentence str.</b>	4.80	2.03	1	9	8	8.91	1.28	5	10	5	2.72	1.31	1	7	6
<b>Spelling</b>	4.27	1.83	1	8	7	9.53	0.81	5	10	5	1.80	0.92	1	5	4
<b>Structure</b>	7.11	2.53	1	10	9	7.04	1.94	2	10	8	4.60	2.65	1	10	9

To examine the reliability of the grading (overall grade) in a more statistically sound manner, a two-way mixed-effects intraclass correlation coefficient for consistency, single measures (ICC (C,1)) was used (see Table 5).

**Table 5***ICC Values of the Reviewers' Grades*

<i>Rater</i>	<i>N<sub>ratings</sub></i>	<i>Texts</i>	<i>ICC</i>	<i>F test (df1, df2)</i>	<i>P</i>	<i>95% Confidence Interval (CI)</i>
Teacher	75	3	0.875	F(2, 148) = 528	< .001	[0.649, 0.996]
GPT-4.1	10	3	0.920	F(2, 18) = 117	< .001	[0.711, 0.998]
Mistral L	10	3	0.919	F(2, 18) = 115	< .001	[0.707, 0.998]
Llama 4	10	3	0.750	F(2, 18) = 31.1	< .001	[0.368, 0.992]

Interrater reliability among the 75 teachers who evaluated the three texts yielded an ICC of 0.875, indicating a high degree of consistency across the reviewers' grades. The associated *F* test was significant. The 95% confidence interval (CI) from the teachers [0.649, 0.996] suggests that the true level of consistency in the population is at least moderate and potentially near-perfect. These results indicate

that, although individual teachers may have different absolute scoring levels, there is strong agreement in the relative ranking of the texts.

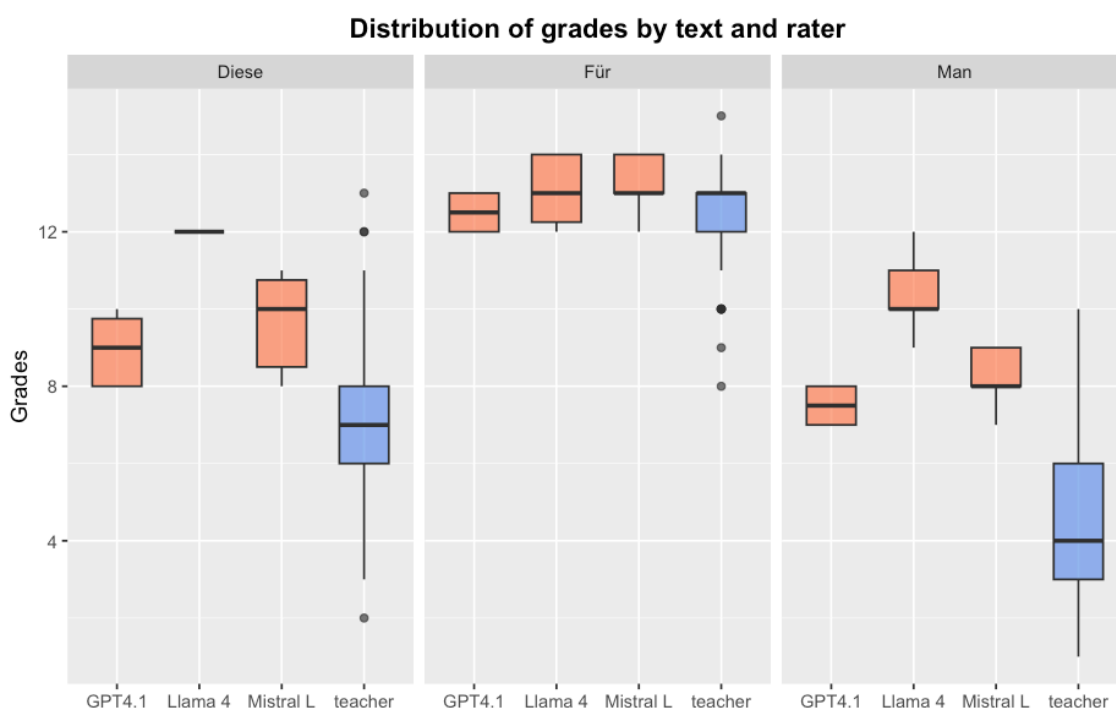
For the 10 AI ratings of the three texts, the ICC values indicate very high interrater consistency except for Llama 4. All  $F$  tests were significant, and the 95% CIs suggest that the true “population” consistency is at least good and may be near perfect, again except for Llama 4.

#### 4.2 RQ2: How much do humans and AI agree when it comes to the criteria-based assessment of pupils’ texts?

First, a circumstance that was already indicated by the descriptive data of the grades from RQ1 is examined graphically, namely that the teachers distinguished more clearly between the three texts. To answer RQ2, the box plots in Figure 2 show how the overall grades for the texts are distributed for each of the three texts separately according to reviewer (group).

**Figure 2**

*Box Plots of Grades by Text and Rater*



It can be noted that the medians for the AI raters are mostly closer together overall; and, for two of the three texts, they are significantly higher than for the teachers. Teachers distinguish more clearly between the three texts and are thus closer to the ground truth. For the *Für* text, the medians are roughly the same. In terms of dispersion, the height of the boxes varies, but is lowest for the best text, except for Llama4’s scores for the text *Diese*, which scored 12 points 10 times and was thus significantly higher than the average scores of the other raters. The whisker lines and outlier points take into account the fact that the dispersion among teachers is generally higher for a text than among AI systems.

To gain detailed insights into the criteria-based ratings, Table 6 shows the descriptive data for the teachers' and AI systems' ratings (mean values and differences) as well as a  $p$  value of a Wilcoxon rank-sum test with a Holm–Bonferroni correction for multiple comparisons.

**Table 6**

*Criteria-Based Assessments From AI and Teachers*

	<i>Diese</i>				<i>Für</i>				<i>Man</i>			
<i>Criterion</i>	<i>M<sub>hu-man</sub></i>	<i>M<sub>AI</sub></i>	<i>Diff</i>	<i>p<sub>holm</sub></i>	<i>M<sub>hu-man</sub></i>	<i>M<sub>AI</sub></i>	<i>Diff</i>	<i>p<sub>holm</sub></i>	<i>M<sub>hu-man</sub></i>	<i>M<sub>AI</sub></i>	<i>Diff</i>	<i>p<sub>holm</sub></i>
<b>Materials</b>	6.81	7.72	-0.91	0.85	9.48	9.68	-0.20	0.49	7.07	7.60	-0.53	0.99
<b>Procedure</b>	7.32	7.58	-0.25	1.00	9.57	9.10	0.47	0.00***	6.20	6.68	-0.47	0.62
<b>Explanation</b>	6.13	6.38	-0.24	1.00	9.25	8.40	0.85	0.00***	4.55	5.68	-1.13	0.00***
<b>Conclusion</b>	6.17	8.03	-1.85	0.01**	3.29	7.30	-4.01	0.00***	5.29	5.62	-0.33	0.85
<b>Style</b>	7.15	7.12	0.02	1.00	9.09	8.47	0.62	0.00***	6.76	6.22	0.54	0.49
<b>Verb form</b>	6.75	5.95	0.80	0.03*	9.27	7.35	1.92	0.00***	2.05	5.18	-3.12	0.00***
<b>Expression</b>	5.31	6.60	-1.29	0.00***	9.12	8.15	0.97	0.00***	3.35	5.58	-2.23	0.00***
<b>Tense</b>	8.79	8.18	0.61	0.03*	9.71	9.50	0.21	0.26	8.53	7.38	1.16	0.00***
<b>Sentence str.</b>	4.80	6.10	-1.30	0.00**	8.91	8.00	0.91	0.00***	2.72	5.35	-2.63	0.00***
<b>Spelling</b>	4.27	6.00	-1.73	0.00***	9.53	9.15	0.38	0.00***	1.80	4.58	-2.78	0.00***
<b>Structure</b>	7.11	7.25	-0.14	1.00	7.04	9.00	-1.96	0.00***	4.60	6.30	-1.70	0.00***

Overall, 22 of the 33 averaged differences show significant discrepancies between the criteria-based judgments of humans and LLMs. The judgments for the text *Für* deviate significantly from one another. The largest difference in relation to the *Conclusion* criterion is explained in the next section. The differences also vary in terms of magnitude and direction. This is because teachers assess the two weaker texts more strictly in relation to most criteria, resulting in mostly negative differences, whereas they assess the best text more positively. Although the overall boxplots for this text suggest close agreement between human and AI ratings at the global level, the criterion-level analysis reveals many statistically significant differences. This apparent paradox may arise because both groups arrive at similar total scores through different weighting of the criteria: AI systems and human raters emphasize or de-emphasize different aspects of the text, leading to compensating effects that mask criterion-specific

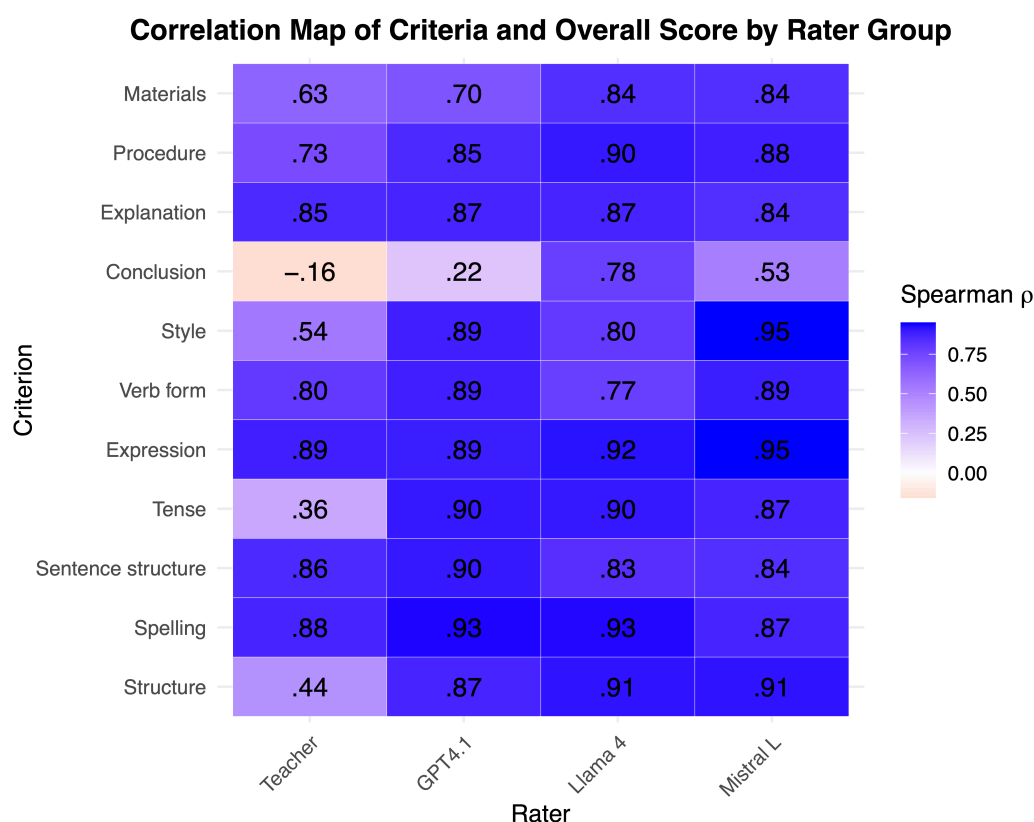
divergences in the aggregated boxplots. Although the grading tendencies of AI and teachers are similar, there are clear differences in the criteria-based assessment. These differences in grading and criteria-based assessment may be explained by the different weightings of the criteria and their influence on the overall grade. This leads directly to RQ3.

#### 4.3 RQ3: In which areas can differences be spotted between human- and AI-based ratings of pupils' texts?

The Spearman correlation plots for teachers and AI models (see Figure 3) show how highly the individual assessment categories correlate with the overall grade:

**Figure 3**

*Correlation Map of Criteria and Overall Score by Rater Group*



Almost all values for the three AI models correlate highly with the overall score as well as with each other, although this cannot be shown in the figure. Among all raters, the *Conclusion* category correlates particularly weakly and even negatively with the overall score. Of course, this may be because the conclusion of the text is considered less important by the reviewers; but with regard to the text *Für*, another explanation suggests itself: This text, which otherwise scores very well in almost all criteria, does not contain a conclusion. This explains the highest variance values among all assessment categories in Table 4 and to the highest difference between AI and humans in Table 6.



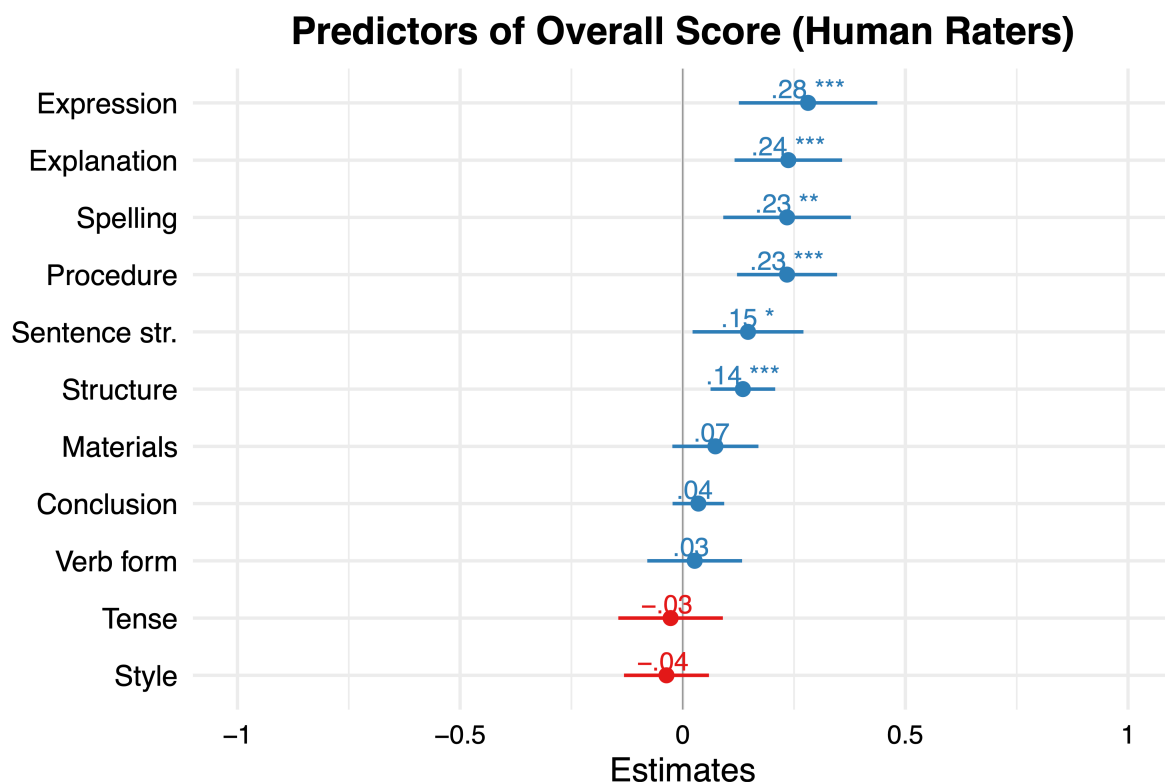
The picture is more nuanced when it comes to teachers. In addition to high scores in the language categories of *Expression*, *Sentence structure*, and *Spelling*,<sup>14</sup> the *Explanation* correlates particularly highly in terms of content. In contrast to the models, however, other criteria such as *Tense* or *Structure* do not show a high correlation with the overall grade among teachers.

The linear mixed-effects models employed here provide a more statistically robust analysis of RQ3. For the three AI models, these models could not be estimated meaningfully due to singular fits and perfect or near-perfect multicollinearity among predictors. Figure 3 already indicates high correlations between the rating criteria and the overall score; and also exploratory analyses revealed very high correlations ( $r > .9$ ) between all rating criteria, resulting in overfitted models ( $R^2 \approx .97$ ) with nonsignificant individual predictors. Because the rating criteria were not statistically independent within the AI data, no interpretable regression coefficients are reported. This indicates that the AI systems produced highly consistent, undifferentiated evaluations across criteria, effectively assigning global quality judgments rather than criterion-specific ratings.

In contrast, the teachers showed sufficient variability across criteria to allow for stable model estimation, indicating more differentiated evaluative behavior.

**Figure 4**

*Estimates of the Fixed Effects From the Mixed-Effects Model for Human Raters*



<sup>14</sup> The dominance of language criteria in essay assessment has already been pointed out by Birkel (2003) or more recently by Vögelin et al. (2019) and, for AI models, by Seßler et al. (2025).

Figure 4 displays the fixed effects from the linear mixed-effects model (detailed results in Appendix C) predicting overall text scores given the rating criteria. The model showed a marginal  $R^2$  of .82 and a conditional  $R^2$  of .90 (Nakagawa et al., 2017), indicating that the fixed effects (i.e., the rating criteria) accounted for most of the variance in overall scores, whereas random variation between raters and texts contributed modestly to the total explained variance. Positive coefficients indicate that higher scores on a criterion are associated with higher overall evaluations. Significant positive predictors included *Expression*, *Explanation*, *Spelling*, *Procedure*, *Sentence structure*, and *Structure*, whereas *Style* and *Tense* showed small, nonsignificant negative associations. All other predictors (*Materials*, *Conclusion*, *Verb form*) had negligible effects.

The advantage of this model compared to the simple correlation analysis shown in Figure 3 is that the criteria and random effects are examined in parallel in terms of their influence on the overall score. In the model, the influence of the categories is lower for *Sentence structure* and even insignificant for *Verb form*, whereas in Figure 3, they appear to have a huge influence. Another specific advantage is evident in the criterion *Procedure*, which is considerably more relevant here than in simple correlation analysis.

Random-effect variances indicated modest variability between raters ( $SD = 0.39$ ) and somewhat greater variability between texts ( $SD = 0.87$ ), suggesting that while raters differed slightly in their general scoring tendency, text-level differences contributed more strongly to variation in overall grades. This rests on only three texts and should not be interpreted substantively (modeling text as a fixed factor left the criterion coefficients essentially unchanged).

Human raters demonstrated differentiated weighting of content, linguistic, and structural criteria in their overall evaluations, whereas the AI systems produced highly uniform, globally consistent ratings across all criteria without meaningful distinctions.

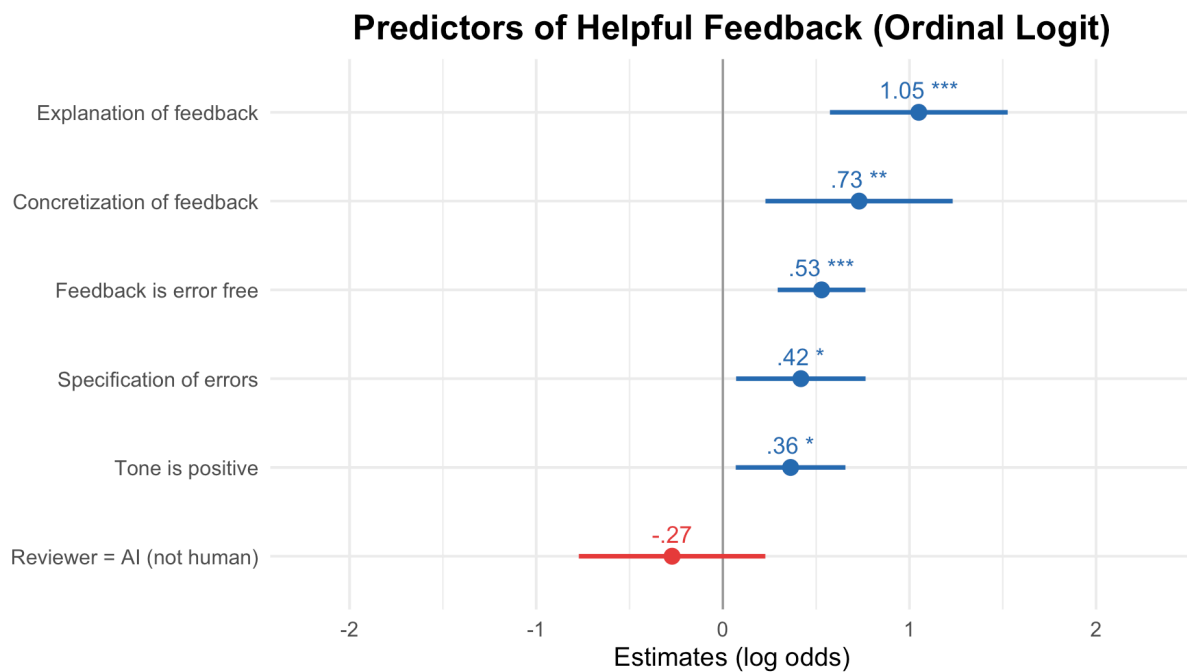
#### 4.4 RQ4: Which quality dimensions of feedback predict whether it is perceived as helpful by meta-reviewers?

To examine which feedback characteristics predicted perceived helpfulness by the metareviewers, ordinal logistic regression analyses were conducted (proportional odds model,  $AIC = 536.434$ ,  $\chi^2 = 25.09$ ,  $df = 16$ ,  $p = .07$ ,  $R^2 = .30$ )<sup>15</sup> with *helpfulness* of the feedback as the dependent variable. Predictors included all other feedback criteria (see Figure 5), as well as reviewer (AI, teacher) and student text quality (high, medium, low). Odds ratios and  $p$  values are used to assess the relative contribution of each predictor (see Figure 5).

<sup>15</sup> Note that McFadden's pseudo  $R^2$  is not directly comparable to the  $R^2$  in other regression models (e.g., RQ3). Values between .2 and .4 are typically considered indicative of a moderate to good model fit (McFadden, 1977; Menard, 2000). The global test of the proportional odds assumption was not significant, suggesting that the assumption was met.

**Figure 5**

*Estimated Effects (Log Odds) from the Ordinal Logit Mixed Model Predicting Helpfulness of the Feedback*



The strongest predictors of perceived helpfulness by the metareviewers were the presence of sufficiently explained ( $OR \approx 2.86$ ,  $p < .001$ ), concrete ( $OR \approx 2.08$ ,  $p = .004$ ), and error-free feedback ( $OR \approx 1.70$ ,  $p < .001$ ). Each of these dimensions substantially increased the likelihood that feedback would be rated as helpful. Additional significant predictors were the specific naming of errors ( $OR \approx 1.52$ ,  $p = .018$ ) and a predominantly positive tone ( $OR \approx 1.44$ ,  $p = .015$ ), both of which contributed to higher helpfulness ratings. In contrast, AI as reviewer compared to teachers as reviewers did exert a negative effect, but it is not significant ( $OR \approx 0.76$ ,  $p = .287$ ). Likewise, the quality level of the underlying student text showed a nonsignificant effect for the text with the medium quality (*Diese*). However, the likelihood that the metareviewers would rate the feedback on the weak text (*Man*) as helpful is significantly higher ( $OR \approx 2.33$ ,  $p = .007$ ) than for the good text (*Für*).

## 5 Discussion

When it comes to predicting an essay's grade, AI systems show a high level of agreement with average grades from experienced teachers. This is consistent with the results of other studies that demonstrate a moderate (Li et al., 2024; Stahl et al., 2024) up to high (Seßler et al., 2025) level of alignment between the grades assigned by humans and AI systems. Although this could lead to a recommendation in favor of preassessment, the European Union's AI Act is critical of the use of AI systems for grading (European Parliament, 2024)—and with good reason. In the debate on the use of AI systems in education, the discrepancy between legal classification and technical capability must be addressed. Additionally, the analysis of the criterion-based judgments also shows that the teachers' judgments differ significantly from the AI criterion-based ratings.

When evaluating the texts using grades or criteria-based points, it becomes apparent that teachers show significantly greater variation than the tested AI systems between different criteria, not only

within a category but also within a text. Although this may appear problematic in terms of objectivity within a category,<sup>16</sup> the difference between the criteria within a text in predicting the overall grade (see Figure 3) shows that teachers are able to evaluate individual criteria more independently of one another. The tested AI systems were less able to distinguish between criteria due to their iterative inferencing. As a result, AI systems used for criteria-based scoring would need even more information, and examples and would also have to assess the criteria separately from one another. To return to the *reliability–truth dilemma* mentioned at the outset: Although the reliability of AI systems (except for Llama 4) is always higher, it remains questionable whether this is also closer to the truth, because they consistently gave higher grades.

The teachers show a (surprisingly) high degree of agreement on the grades ( $ICC = 0.875$ ), apart from a few outliers. This result broadens the scope of research into the reliability of teachers' assessments of German essays, which is still a relatively underresearched area (Birkel, 2003; Mathiebe et al., 2024; Schröter et al., 2022). The high level of agreement may be due to the combination of short, informative texts and highly experienced teachers employed in this study. Notably, although the criteria used for this study were contestable from a didactic perspective, they performed very well operationally: In the mixed model for RQ3, they accounted for the vast majority of the variance in teachers' holistic judgments (marginal  $R^2 = .82$ ). In other words, these ecologically valid criteria, derived from one teacher's own practice, accurately reflected the factors that influenced the highly experienced teachers' overall grades, despite not representing optimal didactic constructs.

The cumulative link mixed model used indicates several strong predictors of perceived helpfulness by the metareviewers (e.g., Explanation, Concretization or Error-free) that are plausible and fit into the feedback model from Hattie and Timperley (2007). The fact that the model predicts that, in terms of the helpfulness of the feedback, it makes no difference to the metareviewers whether it was generated by AI or by teachers does not prove that the quality is similarly high. To ensure this, experts and the pupils themselves would also have to evaluate the feedback both before and after it was implemented in the text. Nevertheless, it can be said that the metareviewers, who recognized AI-generated feedback as such in 70% of cases, found no significant difference in terms of helpfulness. This result also seems to apply to pupils, for whom it is probably irrelevant who created the feedback message as long as it helps them. What is more relevant here is who communicates it, how they do so, and what relationship the receiving person has with the communicating person/agent. Immediacy and scalability will remain the biggest advantages of AI-generated feedback, whereas the question of the conditions for trust in the reviewer will need to be examined more closely in the future. That the feedback (no matter if AI- or human-based) is perceived as more helpful for the weakest text seems obvious, because there is much more room for improvement.

These results could help to limit the algorithm aversion bias (Jussupow et al., 2024) of teachers and pupils, which, due to exaggerated expectations, sometimes leads to the nonacceptance of supposed and actual errors made by AI systems, whereas errors are also made every day in assessments by peers or teachers. In exploratory focus group interviews, Holtzman and Nimrod (2025, p. 6) note: "Acceptance [of technology faults] is also affected by the user's expectations regarding the technology." The completely exaggerated expectations of AI systems could be adjusted to reality with the help of

<sup>16</sup> I believe it is far from certain that objectivity in the assessment of a pupil's text should be viewed as a purely positive construct. From a pedagogical and psychological perspective, a teacher's subjectivity can contribute to feedback being significantly more valuable than feedback that is objective and (supposedly) factually correct. This once again highlights the complexity of trying to measure the construct of feedback quality.

better education about how AI systems work. However, this should not obscure the fact that there is still a lot of work to be done in this field to improve (measuring) the quality of AI-generated feedback.

## 6 Limitations

Choosing only three texts for making the comparison is obviously one of the most important limitations. One reason for this limited number is that teachers (and especially those who teach prospective teachers in addition to their duties as teachers) often feel overloaded and have correspondingly little time. A compromise had to be found that allowed time-consuming narrative feedback to be collected in addition to quantitative feedback. This was not conceivable with more than three texts.

Another limitation inherent to the topic is the rapid development of AI models. Shortly after the data were collected, OpenAI introduced GPT-5. It would have been interesting not least to compare whether the improvements also apply to the area discussed here. Also, the models used were not trained especially for this task.

Another limitation is the choice of an authentic task that does not meet the contemporary understanding of what constitutes a good writing task or a writing arrangement that supports learning. This affects not only the texts, but also the feedback and metafeedback provided.

The sample of teachers was very exclusive and certainly does not represent the population as a whole, because it is a particularly experienced group. In contrast, the metareviewers are less experienced in terms of school and teaching. Operationalizing feedback quality by training student raters is not unknown in research, but that does not make it any less critical. They have not witnessed which feedback has helped which student at what level to improve their writing. This missing experience is another important limitation. Attempts were made to compensate for this through intensive rater training, and it is possible that, due to their closeness to pupils in terms of age, trainee teachers are not as unsuitable as they might appear at first glance when it comes to assessing the helpfulness of feedback. Nevertheless, it must also be noted that the selection of metareviewers is a compromise, and this leads to the most important limitation of this study.

Finally, it should also be noted that feedback quality remains a very difficult construct to operationalize, especially if the learners' perspective is not taken into account at all. The current situation, as indicated by Jansen, Horbach, and Meyer (2025) in which approximately half of the pupils do not engage with AI-generated feedback at all, should bring the perspective of feedback users much more to the fore, especially since previous studies have tended to focus on feedback-centered research (Panadero & Lipnevich, 2022). Learners have to engage with the feedback given. Otherwise, the question of how best to operationalize feedback quality remains purely theoretical.

## 7 Summary and Future Work

Overall, measuring the quality of AI-generated feedback and therefore operationalizing feedback quality poses many conceptual and methodological challenges. This article presents a theoretical model (Figure 1) that can help to systematize the various approaches already taken in research and tackles some of these challenges. The AI systems tested demonstrate high interrater reliability values and often agree satisfactorily with experienced teachers when it comes to the overall grade tendency for pupils' texts. Their written feedback also does not differ significantly from expert feedback provided

by the experienced teachers in terms of how helpful it is perceived to be by the trained metareviewers. However, the AI-generated grades were too lenient. Their criteria-based ratings were largely undifferentiated and diverged significantly from the ratings provided by experienced teachers.

So, is the measured AI-generated feedback good? Answering this question with a yes would overstretch the robustness of the study design and reveal an overly simplistic understanding of the processes that take place every day in schools when feedback is given. Feedback is embedded in social processes (Fong, 2025), and so an AI system will (hopefully) not know what happened in the previous PE lesson, what is currently happening at home for a child, how the pupil has developed, or what exactly was taught under what circumstances. As long as we do not provide such context data to AI systems—and I strongly advocate this—AI-generated feedback can be useful for providing teachers or students with initial guidance—for example, during exercises or by correcting orthographic errors, for example. The AI-generated feedback can therefore only be individualized and not adaptive—which is considered a key criterion for effective learning (Corno, 2008; Parsons et al., 2018; Vygotsky, 1978).

Future studies could investigate whether the advantages of AI-generated feedback, such as immediacy and scalability, could compensate for some of these gaps. Additionally, they could examine whether hybrid approaches that combine the strengths of both AI-generated and human feedback represent useful alternatives. Ultimately, AI-driven feedback is here to stay, so measuring its quality will be a demanding but worthwhile task for science, as it should be as effective as possible in supporting learning.

## References

- Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2024). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1–16. <https://doi.org/10.1080/14703297.2024.2437122>
- Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. *Journal of Educational Computing Research*, 63(2), 431–463. <https://doi.org/10.1177/07356331241307297>
- Andringa, S., & Godfroid, A. (2020). Sampling Bias and the Problem of Generalizability in Applied Linguistics. *Annual Review of Applied Linguistics*, 40, 134–142. <https://doi.org/10.1017/S0267190520000033>
- Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *American Psychologist*, 63(7), 602–614. <https://doi.org/10.1037/0003-066x.63.7.602>
- Aydın, B., Kışla, T., Elmas, N. T., & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System*, 133, 103784. <https://doi.org/10.1016/j.system.2025.103784>
- Bachmann, T., & Becker-Mrotzek, M. (2011). Schreibaufgaben situieren und profilieren. In T. Pohl & T. Steinhoff (Eds.), *Textformen als Lernformen* (pp. 191–209). Gilles & Francke.
- Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1), Article 23.
- Bannò, S., Vydan, H. K., Knill, K. M., & Gales, M. J. (2024). Can GPT-4 do L2 analytic assessment? *arXiv*. <https://arxiv.org/abs/2404.18557>



- Birkel, P. (2003). Aufsatzbeurteilung – ein altes Problem neu untersucht. *Didaktik Deutsch*, 8(15), 46–63. <https://didaktik-deutsch.de/index.php/dideu/article/view/189>
- Bui, N. M., & Barrot, J. (2025). Using generative artificial intelligence as an automated essay scoring tool: A comparative study. *Innovation in Language Learning and Teaching*, 1–16. <https://doi.org/10.1080/17501229.2025.2521003>
- Chiang, C.-H., & Lee, H.-Y. (2023). Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. (Volume 1: Long Papers)* (pp. 15607–15631). Association for Computational Linguistics.
- Christophel, E., Baadte, C., Heyne, N., & Schnotz, W. (2017). Diagnostische und didaktische Kompetenz von Lehrkräften zur Förderung der Text-/Bild-Integrationsfähigkeit bei Schülerinnen und Schülern der Sekundarstufe I (DIKOL). In C. Gräsel & K. Trempler (Eds.), *Entwicklung von Professionalität pädagogischen Personals: Interdisziplinäre Betrachtungen, Befunde und Perspektiven* (pp. 263–281). Springer. [https://doi.org/10.1007/978-3-658-07274-2\\_14](https://doi.org/10.1007/978-3-658-07274-2_14)
- Corno, L. (2008). On teaching adaptively. *Educational Psychologist*, 43(3), 161–173. <https://doi.org/10.1080/00461520802178466>
- Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4), 407–422. <https://doi.org/10.1177/002224378001700401>
- Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers & Education: Artificial Intelligence*, 7, 100299.
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), Article 57. <https://doi.org/10.1186/s41239-023-00425-2>
- European Parliament. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)*. <http://data.europa.eu/eli/reg/2024/1689/oj>
- Fix, M. (2004). Textfeedback in der Sekundarstufe I. In G. Bräuer (Ed.), *Schreiben(d) lernen: Ideen und Projekte für die Schule* (pp. 120–132). edition Körber-Stiftung.
- Fokides, E., & Peristeraki, E. (2025). Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek. *Education and Information Technologies*, 30(2), 2577–2621. <https://doi.org/10.1007/s10639-024-12912-8>
- Fong, C. J. (2025). A renaissance in feedback science? Reviewing and reimagining feedback research methods. *Contemporary Educational Psychology*, 83, 102414. <https://doi.org/10.1016/j.cedpsych.2025.102414>
- Fürstenberg, M., & Müller, H.-G. (2024). KI im Deutschunterricht. *Der Deutschunterricht*, 2024(5), 2–13.
- Graham, S., Harris, K. R., Fidalgo Redondo, R., Harris, K., & Braaksma, M. (2017). Evidence-based writing practices: A meta-analysis of existing meta-analyses writing. In R. F. Redondo, K. Harris, & M. Braaksma (Eds.), *Design principles for teaching effective* (pp. 13–37). Brill. [https://doi.org/10.1163/9789004270480\\_003](https://doi.org/10.1163/9789004270480_003)



- Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29(7), 8435–8463. <https://doi.org/10.1007/s10639-023-12146-0>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education*, 44(8), 1237–1252. <https://doi.org/10.1080/02602938.2019.1599815>
- Holtzman, I., & Nimrod, G. (2025). Forgiveness in human-machine interaction. *Frontiers in Computer Science*, 7. <https://doi.org/10.3389/fcomp.2025.1617471>
- Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Comparing generative AI and expert feedback to students' writing: Insights from student teachers. *Psychologie in Erziehung und Unterricht*, 71(2), 80–92.
- Jansen, T., Höft, L., Bahr, J. L., Kuklick, L., & Meyer, J. (2025). Constructive feedback can function as a reward: Students' emotional profiles in reaction to feedback perception mediate associations with task interest. *Learning and Instruction*, 95, 102030. <https://doi.org/10.1016/j.learninstruc.2024.102030>
- Jansen, T., Horbach, A., & Meyer, J. (2025). Feedback from generative AI: Correlates of student engagement in text revision from 655 classes from primary and secondary school. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 831–836). <https://doi.org/10.1145/3706468.3706494>
- Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.637547>
- Jung, J., Lu, M., Benker, S. C., & Darici, D. (2025). How model size, temperature, and prompt style affect LLM-human assessment score alignment. *arXiv*. <https://arxiv.org/abs/2509.19329>
- Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, 48(4), 1575–1590. <https://doi.org/10.25300/misq/2024/18512>
- Kaliisa, R., Misiejuk, K., López-Pernas, S., & Saqr, M. (2025). How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. *Educational Psychology*, 1–32. <https://doi.org/10.1080/01443410.2025.2553639>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
- Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly*, 59. <https://doi.org/10.1002/tesq.3405>
- Kinder, A., Briese, F. J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., & Leßmann, S. (2025). Effects of adaptive feedback generated by a large language model: A case study in teacher education. *Computers & Education: Artificial Intelligence*, 8, 100349. <https://doi.org/10.1016/j.caeai.2024.100349>
- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In Association for Computing Machinery (Ed.),

- CHI'25: *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–23). ACM. *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- Lehnen, K. (2023). Peerfeedback beim schulischen Schreiben: Grundlagen, Methoden, Modellierung. *IDE: Informationen zur Deutschdidaktik*, 47(2), 18–30.
- Leiner, D. (2024). SoSci Survey (Version 3.5.02). <https://www.sosicisurvey.de>
- Li, T., Liu, Z., Matsumura, L., Wang, E., Litman, D., & Correnti, R. (2024). Using large language models to assess young students' writing revisions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 365–380). Association for Computational Linguistics.
- Limpert, A.-L. (2025, June 26). fobizz, schulKI und Co: Welche KI-Tools können Schulen nutzen? *Deutsches Schulportal*. <https://deutsches-schulportal.de/unterricht/fobizz-schulki-und-co-welche-ki-tools-koennen-schulen-nutzen/>
- Lüdeling, A. (2008). Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lerner-korpora. In M. Walter & P. Grommes (Eds.), *Fortgeschrittene Lernervarietäten* (pp. 119–140). Niemeyer.
- Mathiebe, M., Grabowski, J., & Becker-Mrotzek, M. (2024). Reliabilität bei der Beurteilung von Text-qualität. In I. Petersen, R. Reble, & J. Kilian (Eds.), *Texte schreiben in allen Unterrichtsfächern. Textbeurteilung als Grundlage für Schreibförderung und Leistungsbewertung* (pp. 253–276). Waxmann.
- McFadden, D. (1977). Quantitative methods for analysing travel behaviour of individuals: Some re-cent developments. *Cowles Foundation Discussion Paper* (No. 474), 1–47. <https://elischolar.library.yale.edu/cowles-discussion-paper-series/707>
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1), 17–24. <https://doi.org/10.2307/2685605>
- Meyer, J., Jansen, T., Daumiller, M., & Fleckenstein, J. (2025). Understanding individual differences in students' responses to technology-based feedback on a writing task: The role of achievement mo-tives and initial task performance. *Journal of Research on Technology in Education*, 1–31. <https://doi.org/10.1080/15391523.2025.2471765>
- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers & Edu-cation: Artificial Intelligence*, 6, 100199. <https://doi.org/10.1016/j.caeai.2023.100199>
- Müller, H.-G., & Fürstenberg, M. (2023). Der Sprachgebrauchsautomat: Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. *Mitteilungen des Deutschen Germanisten-verbandes*, 70(4), 327–345.
- Müller, N., Utesch, T., & Busse, V. (2023). Qualität statt Quantität? Zum Zusammenhang von Schreibförderungs- und Feedbackpraktiken mit Textqualität unter Berücksichtigung von migrati-onsbedingter Mehrsprachigkeit. *Unterrichtswissenschaft*, 51(2), 169–198. <https://doi.org/10.1007/s42010-023-00173-2>
- Mußmann, F., Hardwig, T., Riethmüller, M., Klötzer, S., & Peters, S. (2020). *Arbeitszeit und Arbeitsbe-lastung von Lehrkräften an Frankfurter Schulen 2020: Endbericht*. <https://doi.org/10.3249/ugoe-publ-7>

- Mußmann, F., Riethmüller, M., Hardwig, T., Peters, S., Parciak, M., Ohms, I. C., & Klötzer, S. (2016). *Niedersächsische Arbeitszeitstudie: Lehrkräfte an öffentlichen Schulen 2015/2016: Ergebnisbericht*. <http://resolver.sub.uni-goettingen.de/purl/?webdoc-3971>
- Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination  $R^2$  and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134), 20170213. <https://doi.org/10.1098/rsif.2017.0213>
- Neumann, A., Kirsten, E., Zafar, M. B., & Singh, J. (2025). Position is power: System prompts as a mechanism of bias in large language models (LLMs). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 573–598). <https://doi.org/10.1145/3715275.3732038>
- Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, 35, 100416. <https://doi.org/10.1016/j.edurev.2021.100416>
- Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., Pierczynski, M., & Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research*, 88(2), 205–242. <https://doi.org/10.3102/0034654317743198>
- Philipp, M. (2023). Formatives Feedback aus der Sicht des selbstregulierten Lernens: Grundlagen und Grundsätze förderlicher Rückmeldungen. *IDE: Informationen zur Deutschdidaktik*, 47(2).
- Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, 11(1), 2410101. <https://doi.org/10.1080/2331186X.2024.2410101>
- R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.2). R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rashkin, H., Clark, E., Huot, F., & Lapata, M. (2025, July). Help me write a story: Evaluating LLMs' ability to generate writing feedback. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 25827–25847.
- Reichardt, A., Kruse, N., & Lipowsky, F. (2022). Textüberarbeitung mit Schreibkonferenz oder Textlupe: Zum Einfluss der Schreibumgebung auf die Qualität von Schülertexten. *Didaktik Deutsch*, 36. <https://didaktik-deutsch.de/index.php/dideu/article/view/449>
- Rensfeldt, A. B., & Rahm, L. (2023). Automating teacher work? A history of the politics of automation and artificial intelligence in education. *Postdigital Science and Education*, 5(1), 25–43. <https://doi.org/10.1007/s42438-022-00344-x>
- Rüdian, S., Podelo, J., Kužilek, J., & Pinkwart, N. (2025). Feedback on feedback: Students' perceptions of feedback from teachers and few-shot LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*.
- Schaller, N.-J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 210–221). Association for Computational Linguistics.
- Schröter, P., Söldner, H., Hoffmann, L., Riemenschneider, A., Jost, J., & Wieser, D. (2022). Wie vergleichbar sind die Bewertungen von Abiturarbeiten im Fach Deutsch? Empirische Studien zu verschiedenen Bewertungsmodellen. In L. Hoffmann, P. Schröter, A. Groß, S. M. Schmid-Kühn, & P.

- Stanat (Eds.), *Das unvergleichliche Abitur. Entwicklungen – Herausforderungen – Empirische Analysen* (pp. 213–250). wbv.
- Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2022). *Lehrkräfteeinstellungsbedarf und -angebot in der Bundesrepublik Deutschland 2021–2035: Zusammengefasste Modellrechnungen der Länder*. <https://www.kmk.org>
- Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In Association for Computing Machinery (Ed.), *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 462–472). Association for Computing Machinery.
- Shin, I., Hwang, S. B., Yoo, Y. J., Bae, S., & Kim, R. Y. (2025). Comparing student preferences for AI-generated and peer-generated feedback in AI-driven formative peer assessment. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*.
- Skar, G. B., Graham, S., & Rijlaarsdam, G. (2022). Formative writing assessment for change – Introduction to the special issue. *Assessment in Education: Principles, Policy & Practice*, 29(2), 121–126. <https://doi.org/10.1080/0969594X.2022.2089488>
- Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*.
- Steinhoff, T. (2018). Schreibarrangements: Impulse für einen lernförderlichen Schreibunterricht. *Der Deutschunterricht*, (3), 2–11.
- Steinhoff, T. (2025). Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner. Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel des Schreibens mit ChatGPT in der 8. Klasse. In C. Albrecht, J. Brüggemann, T. Kretschmann, & C. Meier (Eds.), *Personale und funktionale Bildung im Deutschunterricht. Theoretische, empirische und praxisbezogene Perspektiven* (pp. 85–99). Metzler. <https://doi.org/10.1007/978-3-662-69640-8>
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, 101894.
- Sturm, A. (2016). Beurteilen und Kommentieren von Texten als fachdidaktisches Wissen. *leseforum.ch*, (3), 115–132. <https://doi.org/10.26041/fhnw-1001>
- Tan, M., Mah, C., & Demszyk, D. (2024). Reframing authority: A computational measure of power-affirming feedback on student writing. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*.
- Usher, M. (2025). Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*, 50(6), 912–927.
- Venter, J., Coetzee, S. A., & Schmulian, A. (2024). Exploring the use of artificial intelligence (AI) in the delivery of effective feedback. *Assessment & Evaluation in Higher Education*, 50(4), 516–536.
- Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, 39, 50–63. <https://doi.org/10.1016/j.asw.2018.12.003>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>

- Wang, D. (2024). Teacher- versus AI-generated (Poe application) corrective feedback and language learners' writing anxiety, complexity, fluency, and accuracy. *The International Review of Research in Open and Distributed Learning*, 25(3), 37–56. <https://doi.org/10.19173/irrodl.v25i3.7646>
- Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., & Wang, R. (2025). Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Computers & Education: Artificial Intelligence*, 9, 100481. <https://doi.org/10.1016/j.caeai.2025.100481>
- Weidlich, J., Gotsch, F., Schudel, K., Marusic-Würscher, C., Mazzarella, J., Bolten, H., Bütler, D., Luger, S., Wohlfender, B., & Merki, K. M. (2025). Teacher, peer, or AI? Comparing effects of feedback sources in higher education. *Computers & Education Open*, 9, 100300. <https://doi.org/10.1016/j.caeo.2025.100300>
- Winkler, I. (2018). Aufgaben. In J. M. Boelmann & L. König (Eds.), *Empirische Forschung in der Deutschdidaktik* (pp. 39–52). Schneider Verlag Hohengehren.
- Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, 35(9), 2817–2842.
- Zou, S., Kai, G., Jun, W., & Liu, Y. (2025). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. *Computer Assisted Language Learning*, 1–30. <https://doi.org/10.1080/09588221.2024.2447279>

Author's address:

Maurice Fürstenberg, Ludwig-Maximilians-Universität München, Schellingstraße 3, 80799 München, Germany  
[m.fuerstenberg@lmu.de](mailto:m.fuerstenberg@lmu.de)

## Appendix A

Sie sind Deutschlehrer mit 21 Jahren Berufserfahrung. Es ist Montag, der 23. März 2026. Kurz nach den in einer Woche beginnenden Osterferien schreibt Ihre Klasse 6b, die durchschnittliche Leistungen zeigt, die dritte Schulaufgabe (Klassenarbeit) in diesem Jahr, zum materialgestützten Informieren in Form einer Vorgangsbeschreibung. Die Schülerinnen und Schüler haben zur Vorbereitung eine Backanleitung als Übungsaufsatz geschrieben und hatten dafür eine Backanleitung als Bild zur Vorlage. Diese Aufsätze sollen die Schülerinnen und Schüler nach Ihrem Lehrer-Feedback noch einmal überarbeiten und erneut einreichen, damit Sie die Aufsätze in den Osterferien in Ruhe korrigieren können.

Da Sie die Klasse wegen dringender Dienstaufgaben diese Woche nicht mehr sehen, haben Sie sich dazu entschlossen, Ihnen eine erste Rückmeldung zu ihren Texten zu geben, und zwar in Form von ...

- ... Punkten zu Bewertungskriterien, welche die Klasse aus dem Unterricht kennt,
- ... einer Gesamtnote sowie
- ... einem Feedbacktext.

Auf dieser Grundlage sollen die Schülerinnen und Schüler ihre Übungsaufsätze überarbeiten und Ihnen bis Ende der Woche ins Fach legen.

Sie haben bereits fast alle Übungsaufsätze fertig, nur drei fehlen Ihnen noch. Sie haben aber nur noch wenig Zeit, weil Sie dann zu einem dringenden Arzttermin aufbrechen müssen. Der Rest der Woche ist mit anderen Aufgaben schon voll. Sie müssen das Feedback für die Texte also in den nächsten Minuten fertigstellen.

Ich gebe Ihnen im Folgenden die Aufgabenstellung, die Beurteilungskriterien und die Beschreibung des Bildes der Backanleitung, die die Schüler bekommen haben.

### Aufgabenstellung:

Du erklärst deinem besten Freund/deiner besten Freundin schriftlich, wie man am besten einen Cookie zubereitet.

Fertige dazu nach den im Unterricht besprochenen Regeln eine Vorgangsbeschreibung an. Stelle den Vorgang genau und vollständig dar und verwende die nötigen Fachbegriffe! – Achte auf saubere Form und Sprachrichtigkeit! Du hast 60 Minuten.

Die Beurteilungskriterien lauten wie folgt:

- Material/Zutat: Du nennst zunächst alle benötigten Materialien/Zutaten in ganzen Sätzen und ohne unerlaubte Abkürzungen
- Ablauf: Den Vorgang gibst du korrekt und in der richtigen Reihenfolge der Handlungsschritte wider.
- Erklärung: Dabei erklärst du dem Leser die inhaltlichen Zusammenhänge gut und unterscheidest gekonnt zwischen Wichtigem und Unwichtigem.
- Schluss: Im Schluss gibst du einen hilfreichen Tipp.
- Stil: Deine sprachliche Darstellung ist sachlich und nüchtern.
- Verbform: Du variierst bei der Vorgangsbeschreibung zwischen Passiv- und man-Konstruktionen.
- Wortwahl/Ausdruck: Dein Ausdruck ist treffend und du vermeidest Wortwiederholungen.
- Tempus: Die gewählte Zeitform ist durchgehend das Präsens.
- Satzbau: Du gestaltest die Sätze abwechslungsreich und verknüpfst sie mit passenden Subjunktionen.
- Orthografie: Rechtschreibung und Zeichensetzung beherrschst du sicher.
- Aufbau: Dein Aufsatz ist klar gegliedert (Absätze zwischen Einleitung, Hauptteil und Schluss).

Die Backanleitung als Bild, welche die Schüler bekommen haben, enthielt Folgendes:

Oben links befindet sich der große, zweifarbige Titel: „CHOCOLATE CHIP COOKIES“

Darunter liegt ein blauer Teller mit einem Stapel Zeichentrick-Cookies; kleine Striche deute Duft oder Glanz an.

Es folgt die Überschrift „DU BRAUCHST“ in braun-roter Schrift, rechts daneben der handgeschriebene Zusatz „für ca. 20 Stück“.

Abgebildete Zutaten (von links nach rechts)

- Sack „MEHL“ und Messbecher „250 g“
- Schachtel „BACKPULVER“ mit Löffel „1 TL“
- Text „SALZ“, Salzstreuer und zwei Löffel „2 TL“
- Text „BUTTER“, Messbecher „250 g“, Butterstück, kleiner Hinweis („GESCHMOLZEN, ABER NICHT HEISS“)
- Flasche „VANILLE“ mit Löffel „1 EL“
- Schüssel „SCHOKOSTÜCKCHEN!“ „350 g“
- Glas „ZUCKER“ mit Messbecher „150 g“
- Pappschachtel „BRAUNER ZUCKER“ mit Messbecher „150 g“
- Zwei Eier mit Beschriftung „2 EIER“
- Beutel „KOKOS-RASPELN“ mit Schälchen „80 g“



#### Rechte Bildhälfte – Schritt-für-Schritt-Illustration

1. Gebogener Pfeil und Wort „OFEN“ zeigen auf einen Drehknopf; darüber steht „AUF 180°“.
2. Rote Schüssel; hinein fließen „BRAUNER ZUCKER“, „BUTTER“ und „ZUCKER“. Unter der Schüssel: „VERRÜHREN!“; daneben ein Schneebesen.
3. Gelber Pfeil zu einer zweiten roten Schüssel, in der ein Schneebesen liegt. Oben ein Löffel „VANILLE“ und ein aufgeschlagenes Ei sowie ein zweites Ei; Text auf der Schüssel: „UNTER RÜHREN NACH UND NACH VANILLE UND EIER DAZUGEBEN“.
4. Pfeil zu blauer Schüssel mit Aufschrift „IN EINER ANDEREN SCHÜSSEL“; hinein fallen „MEHL“ und „BACKPULVER“. Daneben steht „1 TL SALZ“ und unter der Schüssel unter einem Teigschaber: „GUT VERRÜHREN!“.
5. Inhalt einer blauen Schüssel mit senkrechtem Wort „TROCKEN“ kippt in rote Schüssel mit Wort „FEUCHT“, in der eine Masse und ein Schneebesen sind; daneben der Text „LANGSAM UNTERRÜHREN“.
6. Päckchen „KOKOS-RASPELN“ und ein gelber Becher mit Schokostückchen werden in eine rote Schüssel geschüttet; daneben ein Teigschaber.
7. Teighäufchen werden aus einer roten Schüssel mit einem Löffel auf ein „UNGEFETTETES BACKBLECH“ gesetzt; unter dem Blech steht handschriftlich „Abstand!“.
8. Eine Hand streut Salz auf ein Teighäufchen; schräg daneben: „AUF JEDEN KEKS EINE WINZIGE PRISE SALZ“.
9. Küchenuhr mit Beschriftung „ETWA 10 MIN.“ (Ziffer 10 auf der Skala). Darunter ein fertig gebackener Cookie in gelbem Strahlenkranz mit Text „BIS SIE perfekt SIND!“.
10. Ganz unten mittig: umkreiste „10“ und handschriftlicher Hinweis „abkühlen lassen!“.

Ihre Aufgabe ist nun, die Schülertexte, die per Prompt eingegeben werden in einer Tabelle nach den Kriterien zwischen 0–10 Punkten zu bepunkten, danach eine Gesamtnote zwischen 0–15 Punkten zu verteilen und einen knappen Feedbacktext zu schreiben. Geben Sie nur diese 3 Dinge (Tabelle mit Punkten (2. Spalte) nach Kriterien (1. Spalte), Gesamtnote und kurzer Feedbacktext) aus.

[text]

## Appendix B

**Diese** Zutaten werden benötigt für 20 Kekse 250 g Mehl, ein Teelöffel Backpulver, zwei Teelöffel Salz, 250 g Butter (sie muss geschmolzen, aber nicht heiß sein), 350 g Schokostückchen, 150 g Zucker, 150 g Brauner Zucker, zwei Eier, 80 g Kokosraspeln und ein Esslöffel Vanilleextrakt.

Diese Küchengeräte werden benötigt ein Ofen, zwei große Schüsseln zwei kleine Schüsseln, ein Schneebesen, ein Schaber und ein Backblech.

Zuerst muss man den Ofen auf 180 Grad vorheizen. Dann wird in einer Schüssel der braune Zucker, die geschmolzene Butter und der Zucker mit einem Schneebesen verrührt. Danach wird unter ständigem rühren zuerst der Teelöffel Vanilleextrakt und dann die zwei Eier dazugegeben. In einer anderen Schüssel wird ein Teelöffel Salz, ein Teelöffel Backpulver und Mehl mit einem Schaber gut verrührt. Und trocknen lassen. Dann gibt man es in die andere Schüssel wo die geschmolzene Butter, der braune Zucker und der normale Zucker vermischt mit den Eiern und dem Vanilleextrakt sind. Das tut man unter ständigem Unterrühren. Nun werden die Kokosraspeln und die Schokostückchen hinzu gegeben. Dann formt man sie zu kleinen Kügelchen und legt sie auf ein ungefettetes Backblech. Streue auf jedes Kügelchen noch eine Prise Salz. Als nächstes kommen sie für etwa 10 Minuten in den Ofen. Danach nur noch abkühlen lassen und genießen.

Hier noch ein Tipp: Lege die Kügelchen in einem Abstand von fünf cm auf das Backblech.

**Man** benötigt für 20 Kekse 250 Gramm Mehl, ein Teelöffel Backpulver, zwei Teelöffel Salz, 250 Gramm Butter (sie muss geschmolzen, aber nicht heiß sein), 350 Gramm Schokostückchen, 150 Gramm Zucker, 150 Gramm Brauner Zucker, zwei Eier, 80 Gramm Kokosraspeln, einen Esslöffel Vanille, einen Ofen, einen Schneebesen, einen Teigschaber und zwei Schüsseln. Als erstes heizt man den Ofen auf 180 Grad Oberunterhitze vor.

Als Nächstes gibt man den Braunen Zucker, die Butter und den Weißen Zucker in eine der beiden Schüsseln und verrührt alles gut.

Zunächst Rührt man noch die Vanille und die Eier unter.

In die andere Schüssel gibt man Mehl, einen Teelöffel Salz und Backpulver und verrührt dies auch gut.

Als fünftes rührt man das Trockene aus der Zweiten Schüssel in die erste.

Daraufhin gibt man die Raspeln und die Schokostücke in den Teig.

Nun muss man nur noch den Teig auf die ungefettete Bleche verteilen und jeden Chocolate chip Cookie mit einer Prise Salz bestreuen, in den Ofen geben und zehn Minuten warten und sie abkühlen lassen. App

Hier noch ein paar Tipps:

Tipp ein: bestreue die Cookies auch noch wenn sie heiß sind mit ein wenig Zucker.

Tipp zwei: das Rezept ist für 20 Chocolate chip Cookies.



**Für 20 Chocolate Chips Cookies** werden 250 Gramm Mehl, ein Teelöffel Backpulver, zwei Teelöffel Salz, 250 Gramm Butter, die geschmolzen, und abgekühlt wurde, 350 Gramm Schokostückchen, 150 Gramm Zucker, 150 Gramm Brauner Zucker, zwei Eier, 80 Gramm Kokosraspeln und ein Esslöffel Vanilleextrakt benötigt.

Als Materialien werden zwei Rührschüsseln, ein Teelöffel, ein Esslöffel, ein Schneebesen, ein Teigschaber und ein ungefettetes Backblech verwendet.

Bevor man anfängt zu backen, wird der Ofen auf 180 Grad vorgeheizt. Zuerst werden weißer Zucker, brauner Zucker und Butter in die Rührschüssel geschüttet und mit Hilfe eines Schneebesens miteinander vermischt. Unter ständigen Rühren werden auch die Eier und das Vanilleextrakt in die Masse gegeben und so vermengt, dass der Teig zu einer homogenen Masse wird. In die andere Schüssel füllt man nun das Mehl und verrührt es mit einem Teelöffel Salz und dem Backpulver mit Hilfe eines Teigschabers. Nun wird die Mischung der trockenen Zutaten langsam in die feuchte Masse untergehoben. Jetzt werden auch die Kokosraspeln und die Schokoladenstückchen hineingerieselt und mit dem Teigschaber in der Masse untergerührt, sodass die Stücke gleichmäßig verteilt sind. Der Teig wird nun in kleinen Portionen auf das Backblech verteilt. Hierfür sollte man einen Teelöffel verwenden, da ausreichend Abstand zwischen den Teigkügelchen verhindern kann, dass diese während des Backvorgangs zusammenkleben. Vor dem Backen wird auf jeden noch ungebackenen Keks eine winzige Prise Salz gestreut. Zum Schluss schiebt man die Bleche mit den Keksen in den Ofen. Sie werden für etwa zehn Minuten gebacken - und anschließend abgekühlt. Jetzt kann man die Chocolate Chip Cookies genießen.

## Appendix C

**Table C1**

*Random Effects*

<i>Group</i>	<i>Variance</i>	<i>SD</i>
Rater	0.16	0.38
Text	0.75	0.86

*Fixed Effects (Human Raters)*

<i>Term</i>	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
(Intercept)	-0.22	0.76	5.75	-0.29	.78
Materials	0.07	0.05	211.70	1.49	.14
Procedure	0.23	0.06	210.44	4.11	.00
Explanation	0.24	0.06	211.59	3.87	.00
Conclusion	0.04	0.03	212.84	1.19	.23
Style	-0.04	0.05	209.68	-0.76	.45
Verb form	0.03	0.05	193.08	0.50	.62
Expression	0.28	0.08	203.11	3.57	.00
Tense	-0.03	0.06	200.67	-0.46	.65
Sentence str.	0.15	0.06	209.66	2.32	.02
Spelling	0.23	0.07	161.14	3.22	.00
Structure	0.14	0.04	211.47	3.67	.00

**Table C2***Random Effects (Standard Deviations)*

<b>Group</b>	<b>Component</b>	<b>Value</b>
ID	<i>SD</i> (Intercept)	0.39
Text	<i>SD</i> (Intercept)	0.87
Residual	<i>sd__Observation</i>	1.11

*Fixed Effects*

<b>Term</b>	<b>Estimate</b>	<b>OR</b>	<b>SE</b>	<b>p</b>
Explanation of feed-back	1.05	2.86	0.243	<0.001
Concretization of feed-back	0.730	2.08	0.256	0.004
Feedback is error-free	0.529	1.70	0.120	<0.001
Specification of errors	0.418	1.52	0.177	0.018
Tone is positive	0.363	1.44	0.150	0.015
Reviewer = AI	-0.272	0.762	0.255	0.287
Text quality low	0.845	2.33	0.314	0.007
Text quality medium	0.244	1.28	0.305	0.423

Maxi Kupetz

## “I did write ‘huge’. Why does he say ‘enormous’?”

**Text revision processes with an AI-based chat tool and classroom interaction in a German-as-a-Second-Language lesson**

### **Abstract**

This article presents results of a pilot study on the use of a generative tool (*KI Chat*) in a support lesson for German as a second language (GSL) at a secondary school in Saxony (Grade 11). Video data were collected in a learning setting in which tablets were used to work with the GDPR-compliant platform *Fobizz* that includes *KI Chat*. The questions addressed are: How is an L2 learner’s process of revising a self-written text—using *KI Chat*—accomplished? Which specific interactional tasks and intricacies can be observed? What can be learned from these classroom observations for future research and for improving learning settings? The analytic work is based on the principles of conversation analysis (CA), with a specific focus on multimodality. A detailed case analysis shows the reconstruction of the trajectory of a text revision process using *KI Chat* and the concurrent classroom interaction. The article closes by addressing pedagogic considerations such as the need to carefully look out for learning opportunities when working with AI-based tools. This includes thoroughly reading the texts provided by the tool and reflecting on the tool’s role in the writing process.

**Keywords:** classroom interaction • writing processes • AI-based tools • German as a second language (GSL) • conversation analysis (CA) • multimodality

### **Didaktik Deutsch**

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 97–122

DOI: 10.21248/dideu.969

**Copyright** Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0 veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

## 1 Introduction<sup>1</sup>

Current research on AI-based tools in German schools often focuses on teachers' or students' perceptions of ways of using them (cf. Helm et al., 2024). Although classroom research on the use of digital tools was described as a research desideratum a few years ago (Maahs et al., 2023, p. 36; Steinhoff, 2023, pp. 12–13), it is only more recently that empirical studies have investigated actual working processes with AI-based tools (e.g., Knopp & Schindler, 2025; Schneegass, 2025; Schulz & Gunkel, 2025). This article presents first results of a pilot study on the use of an AI-based tool in support lessons for German as a second language (GSL) at a secondary school in Saxony (Grade 11). Video data were collected in five lessons in which students worked with tablets, using the GDPR-compliant platform *Fobizz* that includes generative tools such as *KI Chat*. The data revealed the interactional complexity of the writing and text-revision processes. To facilitate comprehension of this complexity, this article presents a detailed case analysis that reconstructs the trajectory of text revision using *KI Chat* and concurrent classroom interaction. The research questions addressed are: How is an L2 learner's process of revising a self-written text—using *KI Chat*—accomplished? Which specific interactional tasks and intricacies can be observed? What can be learned from these classroom observations for future research and for improving learning settings? The analytic work is based on the principles of conversation analysis (CA) with a specific focus on multimodality and classroom interaction (Gardner, 2013; Pitsch & Ayaß, 2008).

After briefly discussing the current literature on the use of AI-based tools in writing processes (Section 2), I shall present the research desideratum that the present study addresses. Methodological considerations will be presented in Section 3: First, I summarize previous studies showing the potential of CA as an analytic approach to writing (in interaction). Second, I thoroughly describe the video data this study is based on. Section 4 contains a detailed case study of one student's text revision process, revealing specific interactional intricacies, learning potentials, and the student's stances. In Section 5, I discuss the observations in terms of future research and didactic considerations for teachers. This article thus contributes to the Special Issue *Writing with AI. Literacies – Practices – Pedagogies (LPP)* by discussing aspects of the following questions: Which methods and methodologies can be useful for researching the role of AI tools in the (language) classroom? How significant are prompting and multimodality? Which practices of writing with AI and which (new) literacies are didactically relevant?

## 2 The use of AI-based tools for writing in the (language) classroom

Nowadays, reading and writing processes are influenced by digital tools in various ways (Lehnen & Steinhoff, 2024). Often, users may not even realize how their reading and writing processes are linked to programs such as spelling programs running in the background (Lehnen & Steinhoff, 2024, pp. 529–530). Moreover, the strict separation of reading and writing processes is being blurred in digital practices (Lehnen & Steinhoff, 2024, p. 533). Since 2022, the chatbot ChatGPT has been available, an AI-based tool that generates text in response to prompts realized in everyday language (Schindler, 2023; Weiss et al., 2023). In the field of German didactics, there have been attempts to model the use of GPTs (generative pretrained transformers) theoretically: Müller and Fürstenberg (2023) emphasize their nature as language-usage machines (*Sprachgebrauchsautomaten*) that rely on statistics and not on semantics. Focusing on scenarios of usage, Steinhoff and Lehnen (2025) have proposed the GPT model of practices. This describes chat-to-generate practices, chat-to-create practices, chat-to-chat

<sup>1</sup> DeepL Translator Pro tool was used to translate parts of the manuscript and to revise this text linguistically. I would like to thank the Didaktik Deutsch editorial team for arranging an English-language proofread, which greatly improved the quality of the text.

practices—each practice being constituted by specific “roles” taken by the AI-based tool and the human user (Steinhoff & Lehnert, 2025). Table 1 (adapted from Steinhoff & Lehnert, 2025, p. 4) presents the typical constellations.

**Table 1**

*Roles Taken by AI-Based Tool and Human User*

	AI tool	user
r o l e	ghost	client
	partner	explorer
	tutor	learner

The model implies that the questions of authorship and agency in writing processes need to be revisited (Bajohr & Hiller, 2024; Lehnert & Steinhoff, 2024, p. 535). Further research is required to achieve a more profound comprehension of the coactivity of humans with the program, including an examination of not only how tasks are constructed but also the quality of AI-generated texts (Lehnert & Steinhoff, 2024, p. 542). Hence, methods need to be developed to observe machine–human interaction in ongoing writing processes (Steinhoff, 2022, p. 154). The present study contributes to this line of research (see Section 5 for a discussion of the findings against the backdrop of the GPT model).

In terms of teaching and learning a foreign or second language (L2), the potential of AI-based tools has been addressed by many authors. Yang and Li (2024) provide a systematic review of studies on ChatGPT and L2 learning, taking into consideration ChatGPT’s roles and prompt patterns, participants (educational levels, sample sizes, and target languages), objectives (language skills and learning perceptions), adopted theories and research methods, and outcomes in terms of advantages and challenges. From a language-teaching perspective, there are numerous didactic suggestions for using Chat GPT or other chatbots in various phases of L2 writing (see, e.g., Dornicheva & Sulimova, 2025, for Russian; Steinbach, 2025, for English; Markou, 2024, for German). Regardless of the target language, the authors of the aforementioned publications demand openness and, at the same time, critical awareness of the opportunities and limitations of AI-based tools among teachers and learners (Feick, 2025). Which considerations can be derived from the present study for teacher training will be discussed in the last section.

### **3 Methodological approach**

#### **3.1 CA perspectives on classroom interaction and writing processes**

Typically, classroom studies that draw on CA methodology focus on the oral dimensions of classroom interaction—many of them, nowadays, taking materiality into account when interactionally relevant. But there are also studies that focus specifically on combining an interaction analysis perspective with the analysis of writing processes, be it in elementary school, secondary school, or adult education (e.g., Bär, 2021; Beisswenger, 2017; Lehnert, 2000, 2014; Stierwald, 2020; Tadres, 2025). Moreover, for L2 settings, studies have shown that such a methodological approach may yield important insights into the characteristics involved in the interactional accomplishment of writing (e.g., Domenech & Leßmann, 2022; Geist et al., 2019; Lingnau & Mehlem, 2012). Such research on didactic settings that

involve social interaction and writing tasks provides important groundwork for language teacher education (Domenech & Leßmann, 2022, p. 85).

The present study draws on the methodological premises of CA (Gardner, 2013) and multimodality (Hall & Looney, 2019; Pitsch & Ayaß, 2008). In other words, I work with video data of naturally occurring classroom interaction, and I assume that sequentiality is the underlying mechanism of the classroom interaction order while taking all semiotic resources (verbal, vocal, bodily, material, etc.) seriously with regard to their relevance for the overall constitution of classroom activities. Processes of written text production are an integral part of the analysis when this is what the participants themselves orient to. During the course of the present study, first observations on the data showed that positioning actions and stance taking might become of analytical interest. Therefore, I draw on the concept of self- and other-positioning as locally constituted dimensions of actions that ascribe specific social identities to persons in interaction (Lucius-Hoene & Deppermann, 2004, pp. 168–170). Affect and stance taking are analyzed as displays that are locally situated in the interaction (Couper-Kuhlen, 2009; Kupetz, 2014).

### 3.2 The data

This study is based on video recordings of a support course for German as a second language (GSL) held once or twice a week after regular classes at a high school in Saxony (Kupetz, in press). Five double lessons (totaling 7 hours and 20 minutes) were recorded in the winter of 2024. The course was attended regularly by four 11th-grade students. These are students who are accustomed to learning and have advanced German language skills (B2 or C1); their family languages are Arabic or Spanish. In accordance with the “Saxon Concept for the Integration of Migrants” (Saxon State Ministry of Culture, 2021), the students are in the so-called third stage—that is, they are fully integrated into regular school classes and receive additional instruction in GSL with the aim of enhancing their academic language skills. In the lessons recorded, each student works on an individual tablet with prepared tools and tasks. The teacher is an advanced tutor (student for the subjects English, Protestant religion, and GSL). She has been accompanying the group in the support program for about 3 years.

**Figure 1**

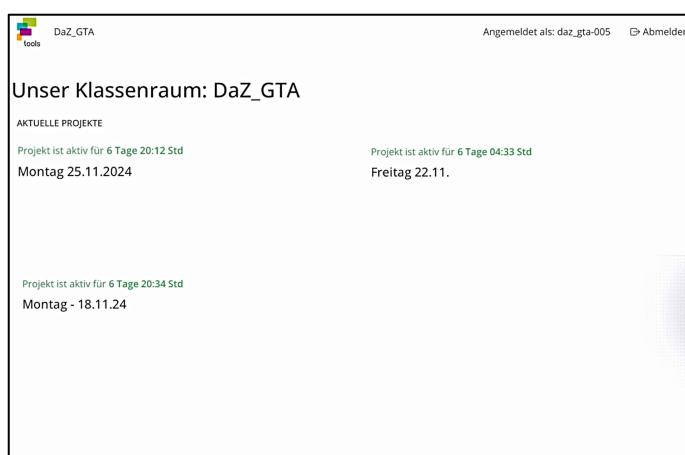
*Classroom Setting*



In addition to linguistic objectives, AI learning objectives were formulated for all lessons. For example, “The students can independently ask ChatGPT questions on given topics and have them answered... The students can use ChatGPT to improve and structure their own texts. The students can critically discuss the use of AI on the basis of reflection questions.”<sup>2</sup> All lessons were conducted using the GDPR-compliant platform *Fobizz* on tablets. A digital classroom was set up in *Fobizz*, with a separate project available for each double lesson in which applications were activated and materials uploaded (see Figures 2, 3, and 4).

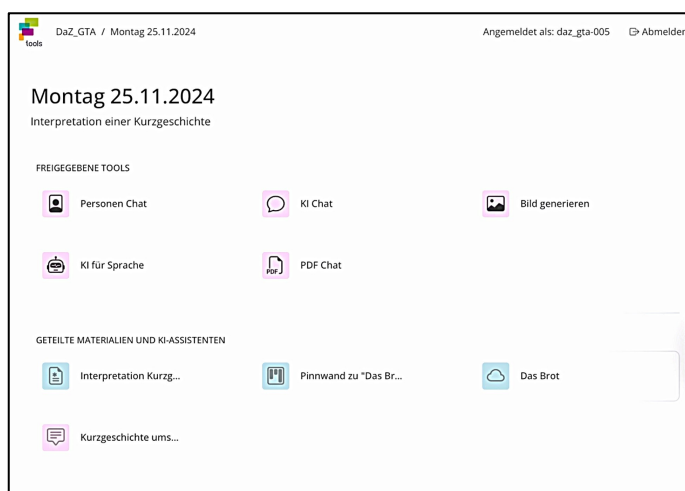
**Figure 2**

### *Digital Classroom*



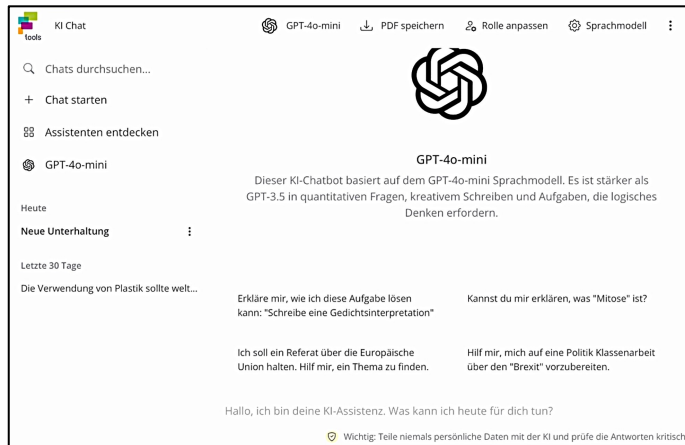
**Figure 3**

### *Project for Each Lesson*

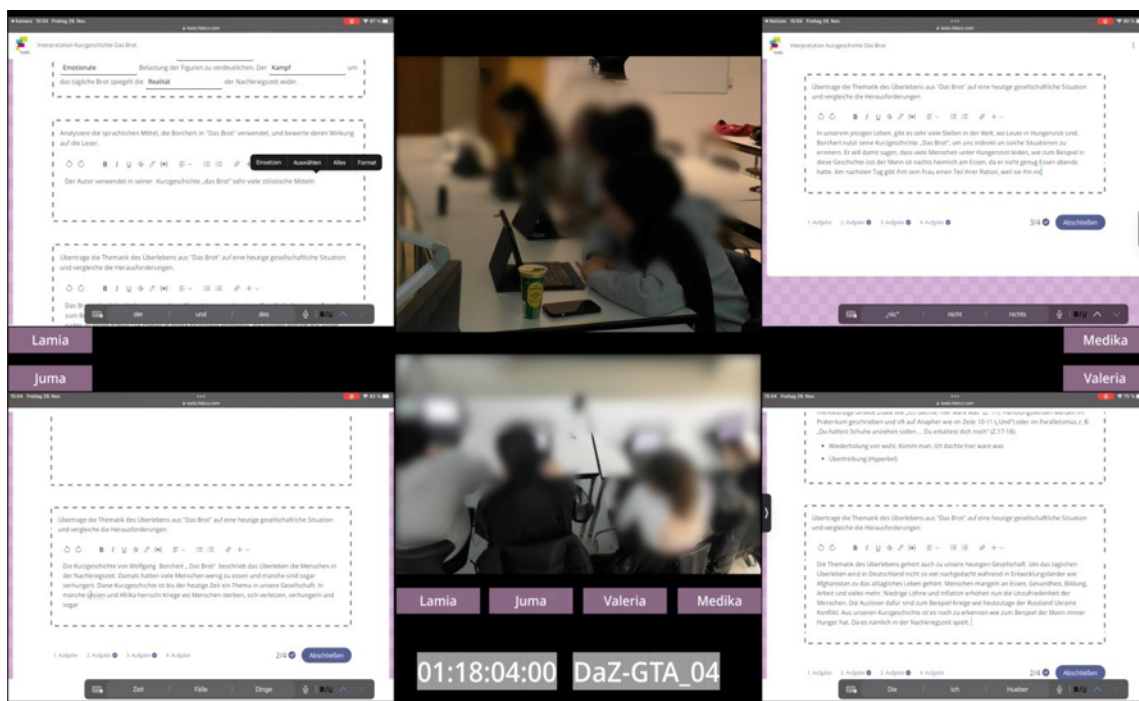


<sup>2</sup> The learning objectives refer to the application ChatGPT. What is meant, however, is the LLM-based chatbot *KI Chat* that is integrated in the *Fobizz* platform.



**Figure 4***Interface of KI Chat (based on GPT-4o-mini)*

Lessons were recorded using four cameras, individual microphones, and table microphones in order to make the students' activities as comprehensively accessible as possible for the analysis. In addition, screen videos of the students' tablets were created in order to access the students' written work. To enable analysis of the material, two camera recordings were synchronized with the screen videos (see Figure 5).

**Figure 5***Example of Edited Videos Combining Camera Perspectives and Screen Recordings*

Two of the recordings were selected for closer analysis. The first lesson was chosen because of the initial work with *KI chat*, and the fourth lesson was chosen because of the intensive work with the same tool. For large parts of selected lessons, minimal transcripts according to GAT 2 (Selting et al.,

2009) were made; for some selected passages, basic transcripts are also available supplemented by notations of visually perceptible actions.<sup>3</sup> The overall aim is to make the data accessible for detailed sequential and multimodal analysis.

For the present study, a specific lesson was analyzed (Lesson 4). In Lesson 3, the group started working on “interpreting the short story *Das Brot* by Wolfgang Borchert”—a topic chosen to prepare the students for their exam in the school subject German. This work is continued in Lesson 4. After several preparatory steps (in the sense of scaffolding), the students are asked to write two texts, each fulfilling a task in a *Fobizz* worksheet (see Figure 6).

**Figure 6**

*Worksheet With Translations on the Right*

Analysiere die sprachlichen Mittel, die Borchert in "Das Brot" verwendet, und bewerte deren Wirkung auf die Leser.   Antwort hier eingeben...	Analyze the linguistic and stylistic devices used by Borchert in <i>Das Brot</i> and evaluate their effect on readers.
Übertrage die Thematik des Überlebens aus "Das Brot" auf eine heutige gesellschaftliche Situation und vergleiche die Herausforderungen.   Antwort hier eingeben...	Apply the theme of survival from <i>Das Brot</i> to a contemporary social situation and compare the challenges faced.

#### 4 Case study: Tracing Lamia's use of the *KI Chat* tool for text revision

The focal point of this analysis is one student's (Lamia's) process of text revision using the *KI Chat* tool.<sup>4</sup> It starts from the teacher's instruction of "revising the text with the help of *KI Chat*" (Kupetz, in press) about 20 minutes before the end of the lesson. Before this, the students have been working on their two tasks (see above) for 45 minutes.

##### 4.1 Lamia's response to the instruction: challenges in using the *KI Chat* tool

In Transcript 01, Lamia's first revision of her text is achieved. It starts with the teacher's instruction. At the moment of the teacher's instruction, Lamia has finished the second text ("Apply the theme of survival . . ."), but only just started the first one.

<sup>3</sup> I would like to express my sincere gratitude to Sophie Oettrich and Lyan Krumpfer for their invaluable assistance in establishing the video corpus and transcribing the data.

<sup>4</sup> Due to the duration and complexity of the student's working process, I decided to carry out a case study. Further research may concentrate on comparing the students' writing processes.

**Transcript excerpt 01: "revise" (DaZ-GTA\_04\_01:18:04) (LEH = Lehrer, i.e., teacher)**

(01:18:04) 0001 **LEH** jetzt würd ich auch Sagen,  
now I would also say

(01:18:05) 0002 ihr kommt jetzt langsam zum ENde,  
you are now slowly coming to the end

(01:18:07) 0003 wenn die aufgaben noch nich FERTig sind==  
if the tasks are not yet finished

(01:18:08) 0004 =eine der beiden habt ihr ja definitiv FERTig,  
you've definitely finished one of the two

(01:18:10) 0005 **LAM** mhm,  
hmm

(01:18:11) 0006 **LEH** dann überarbeitet DIEse mit ka i:.  
then revise this one with AI

(01:18:14) 0007 (1.01)

(01:18:15) 0008 wir achten auf SPRAche und Inhalt.  
we pay attention to language and content

(01:18:17) 0009 ja,  
okay

(01:18:17) 0010 (1.89)

(01:18:18) 0011 **MED** ((ca. 2.36s talking to herself quietly, commenting  
on her actions))

(01:18:19) 0012 **LAM** okay;  
okay

(01:18:20) 0013 also soll ich jEtzt  
so now I'm supposed to

(01:18:21) 0014 (1.03)

(01:18:22) 0015 **MED** ((ca. 5.86s talking to herself quietly, commenting  
on her actions))

(01:18:22) 0016 **LAM** das (-) das tExt was ich geschrIeben habe koPIEren,  
copy the text that I wrote

(01:18:26) 0017 **LEH** mhm,  
hmm

(01:18:28) 0018 (1.22)

(01:18:29) 0019 **LAM** und DANN,  
and then

(01:18:30) 0020 (4.17)

(01:18:34) 0021 **LEH** habt ihr (ja) den ka i: chat,  
you have the KI chat

(01:18:35) 0022 mit dem ihr ARbeiten könnt;  
with which you can work

(01:18:36) 0023 (0.52)

(01:18:37) 0024 **LAM** ka i: CHAT,  
KI chat

(01:18:38) 0025 **LEH** und dann GUCken wir doch ma:l,  
and then we'll see

(01:18:40) 0026 **VAL** [aber soll man (den TEXT ge]ben,) ]  
but should one give the text

(01:18:40) 0027 **LAM** [und soll ich (dann in ] ]  
and should I

(01:18:41) 0028 (0.45)

((The teacher continues by talking to Valeria.))

The students' responses to the teacher's instruction (Segments 01–09) lack clear displays of understanding. Instead, Lamia produces a candidate understanding on what to do (Segments 13, 16). About a minute and a half later, Lamia takes the floor and addresses the teacher:

**Transcript excerpt 02a: "revise my text I" (DaZ-GTA\_04\_01:20:07)**

(01:20:07) 0001 **LAM** ähm:: nachdem ich dann mein text (.) EIngefügt  
hab,  
uhm after I then pasted my text

(01:20:12) 0002 [was ]mit was soll ich n\_das verGLEICHen,  
what with what should I compare it

(01:20:12) 0003 **LEH** [ ( (X) ) ]

(01:20:14) 0004 (0.71)

(01:20:15) 0005 **LAM** soll ich einfach das SCHIcken,  
should I just send it

(01:20:16) 0006 **VAL** in den rIchtigen <<engl.> CHAT gi pi ti> oder den  
to the real chat gpt or the one  
von [ ( X X X )  
from ]

(01:20:18) 0007 **LEH** [nee; bei <<pointing at VALs Ipad> ka ] i:  
no to KI  
chat> den ihr DA habt;  
chat which you have there

(01:20:20) 0008 (1.49)

(01:20:22) 0009 **VAL** [ ((laughs))

(01:20:22) 0010 **LAM** [und soll ich einfach ihn: das SCHIcken,  
and should I just send it to him

(01:20:24) 0011 (0.39)

(01:20:24) 0012 **LEH** naja;=wie überarbeitest du denn SONST deine tExte  
well how else do you revise your texts  
mit ka I:;  
with A I

(01:20:27) 0013 (3.57)

(01:20:30) 0014 **LAM** AH;  
ah

(01:20:31) 0015 ich muss noch ih(n) SCHREIben-  
I have to write (to him)

(01:20:32) 0016 beab be beArbeite mein: TEXT;=bitte;  
rev revise my text please

(01:20:35) 0017 **LEH** mhm,  
hm

(01:20:35) 0018 geNAU.  
that's right

(01:20:36) 0019 (1.57)

(01:20:38) 0020 **LAM** ((clicks tongue))

(01:20:38) 0021 (0.75)

(01:20:39) 0022 **LAM** der is richtig DUMM manchmal (auch).  
he's really stupid sometimes

(01:20:41) 0023 (0.72)

(01:20:41) 0024 **LEH** ((laughs))

(01:20:43) 0025 WIRKlich; lamia.  
really lamia

(01:20:44) 0026 **LAM** ndja  
well

```

(01:20:45) 0027 #1(1.92)((starts typing 'B'))
(01:20:47) 0028 bearb
 edi
(01:20:48) 0029 (1.91)
(01:20:50) 0030 BITte;
 please
(01:20:51) 0031 ich bin sehr Höflich;=
 I am very polite

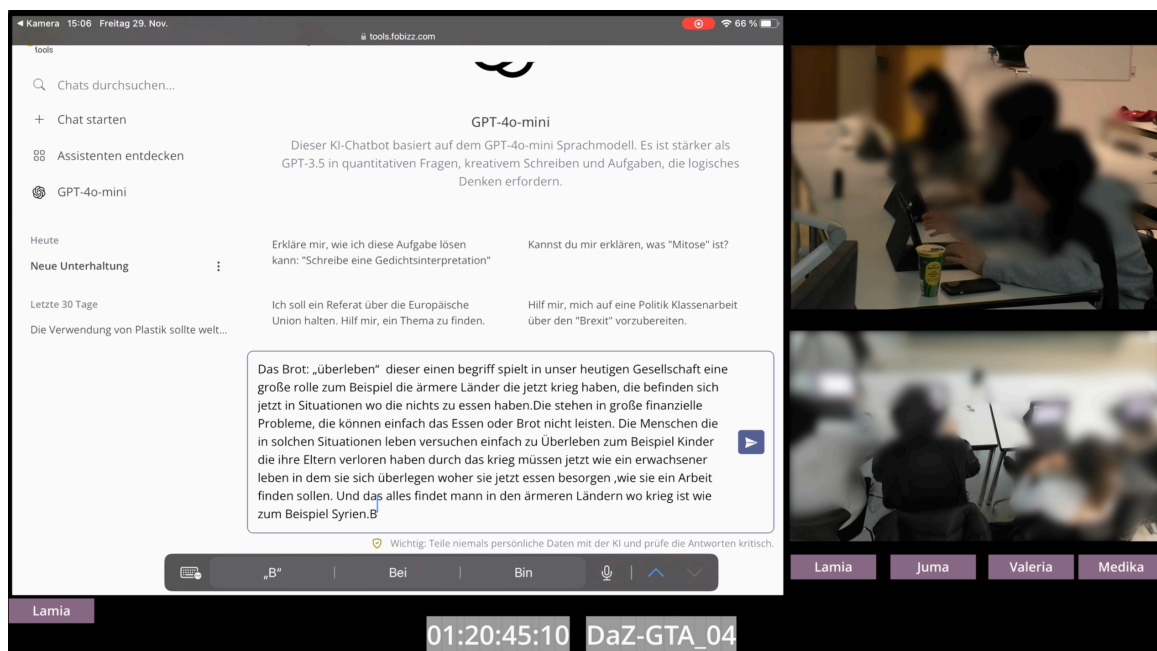
```

((The teacher then explains how texts are saved, followed by a playful side sequence between the students. Lamia then refocuses her tablet while the teacher continues talking to the other students.))

This excerpt shows that Lamia is still displaying uncertainty regarding the proper execution of the task of “revising the text with the help of *KI Chat*.” She offers options (comparing something vs. sending the text) (Segments 02, 05), thereby revealing uncertainties about the role of the chatbot. At the end of Segment 5, a teacher’s response is made relevant. Valeria uses this sequential slot to take the floor (Segment 06), designing her utterance as an increment to Lamia’s turn by using a prepositional phrase, addressing a question about the tool to be used (*ChatGPT* vs. *KI Chat* in *Fobizz*). The teacher responds to this by specifying through the use of multimodal resources that the tool *KI chat* is to be used (Segment 07). With Valeria’s laughter, the insert sequence is brought to an end (Segment 09). Lamia thus repeats her request for specification on what to do (Segment 10). This is taken up by the teacher with an utterance in the format of a question on how Lamia would usually proceed (Segment 12), thereby addressing the student’s practices for text revision (for further discussion, see Kupetz, in press). This leads to Lamia’s realization of a metacomment on what to do (Segment 15), and a formulation of a possible prompt (Segment 16), and this is positively assessed by the teacher (Segments 17–18). After a click and a pause, Lamia comments on the chatbot (Segment 22), referring to the tool with “he,” assessing it as “stupid sometimes.” In this very moment, she ascribes the role of the “incompetent assistant” to the tool (quite in contrast to other comments, see below), thereby creating an interaction mode that oscillates between seriousness and joking. The teacher takes this up with a gentle admonition (Segment 25), orienting toward school rules that prohibit derogatory speech. Lamia then starts working and reading out the beginning of the prompt while typing (Segment 27).

Figure 7

## #1 Video and Screenshot of Lamia 01:20:24 (02a, seg. 27)



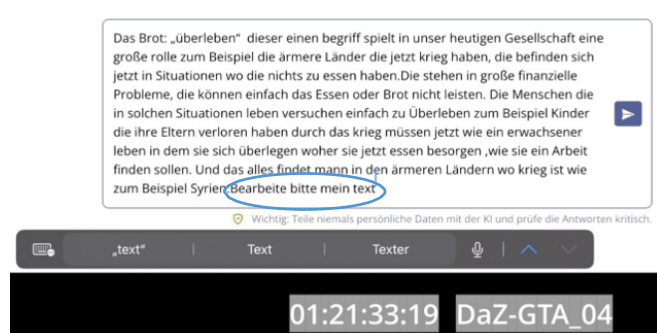
Having verbally produced *bitte*, she produces the meta-comment of being polite (Segment 31), as if to repair the derogatory remark she made earlier. It is to be noted that Lamia seems to be constantly navigating between (trying to) fulfil the task with *KI Chat* and commenting on the tool and/or her actions. Not even a minute later, Lamia reads out loud her prompt:

## Transcript excerpt 02b: “revise my text II” (DaZ-GTA\_04\_01:21:28)

(01:21:28) 0001 LAM <<pp> (X X) bitte mein (---) TEXT.>#2  
(X X) my text please  
(01:21:35) 0002 ((observes screen))

Figure 8

## #2 Screenshot Lamia 01:21:33, Prompt Highlighted: “Please Edit My Text” (02b, seg. 01)



Only a few seconds later, she comments:

**Transcript excerpt 03: “he revised my text” (DaZ-GTA\_04\_01:21:42)**

(01:21:42) 0001 **LAM** #3 er hat mir mein: TEXT beArbeitet.  
he revised my text for me

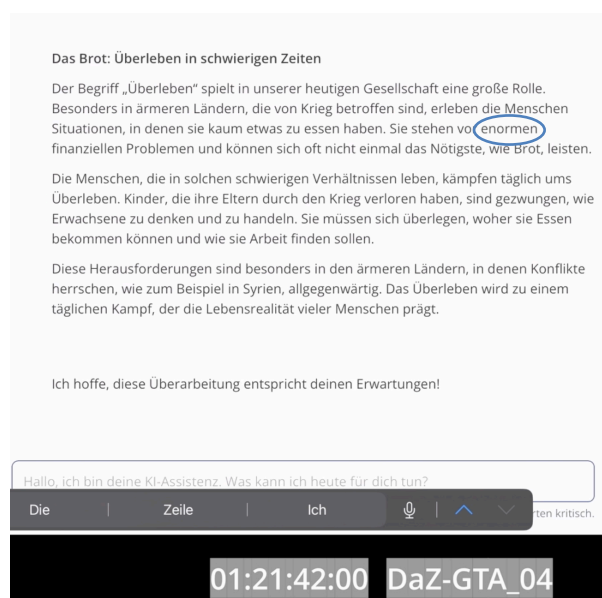
(01:21:44) 0002 (0.53)

(01:21:44) 0003 **LEH** SUper.  
great

((The teacher and Medika talk about what to do next. Meanwhile, Lamia reads her text, whispering while reading.))

**Figure 9**

*#3 Screenshot 01:21:42, Lamia’s Text Revised by AI, enormen Highlighted (03, seg. 01)*

**4.2 Lamia’s observations on the edited text I: “Why does he say ‘enormous’?”**

After Lamia has pointed out that the tool has revised the text (Transcript 03), she starts reading the revised text and then takes the floor after a couple of seconds:

**Transcript excerpt 04a: “enormous I” (DaZ-GTA\_04\_01:21:57)**

(01:21:57) 0001 **LAM** <<p> ah ja.>  
oh I see

(01:21:58) 0002 er hat\_s WIRKlich;> ((nods))  
he really has

(01:22:00) 0003 ((LAM whispers while reading))  
((TEA turns to her own computer))

(01:22:06) 0004 **LAM** DAS kann er vielleicht sEhr gut;  
perhaps he can really do that well

(01:22:07) 0005 oder,  
or

(01:22:08) 0006 hm;

(01:22:08) 0007 (0.73) ((TEA shrugs her shoulders))

(01:22:09) 0008 **JUM** ((whispers, ca. 1.79s))

(01:22:10) 0009 **LEH** <<p> sag DU\_s mir.>  
you tell me



(01:22:11)	0010	LAM	((whispers while reading))
(01:22:13)	0011		was bedeutet eNORMen; what does enormous mean
(01:22:14)	0012		(1.03)
(01:22:15)	0013	VAL	GROßen. huge
(01:22:15)	0014		(0.73)
(01:22:16)	0015	LEH	[JA. ] yes
(01:22:16)	0016	LAM	[WAS?] what
(01:22:16)	0017	VAL	GROß. huge
(01:22:17)	0018		(0.77)
(01:22:18)	0019	LAM	ja; [ich HAB doch] [groß geschrieben;]= but I did write “huge”
(01:22:18)	0020	VAL	[VIEL;                   ] a lot
(01:22:18)	0021		[masSIV;                   ] massive
(01:22:19)	0022	LAM	=warum sagt er eNORMen. why does he say “enormous”
(01:22:21)	0023	VAL	masSIV; massive
(01:22:22)	0024		VIEL; hehe a lot
(01:22:23)	0025		eNORM; enormous
(01:22:24)	0026	LAM	als ob ich das jetzt (.) mh verSTeHe; as if I mh understand that now
(01:22:27)	0027		(1.23)

Lamia’s observations on the edited text are addressed in the interaction between Lamia and the teacher, as well as among the students themselves. In Segments 01 and 02, Lamia acknowledges changes realized by the tool. Both segments are realized with low volume. At this point, it is ambiguous whether the talk is designed as “self-talk” (Goffman, 1978, pp. 787–789) or addressed to the teacher. It is not oriented to by the teacher, and Lamia continues reading (Segment 03). Lamia then produces an assessment of the tool’s abilities (Segments 04–05). It is noticeable that she uses *vielleicht* (perhaps) as a hedge and *oder* as try-marking (Cantarutti, 2023). The utterance’s volume is regular—that is, it is produced as publicly observable and to be answered. A response, however, does not follow immediately. Juma, Lamia’s neighbor, keeps orienting to her text, and no other student takes the floor. The teacher then, while gazing at her own computer, responds in a way that obliges Lamia to find an answer on her own (Segment 09). After 2 seconds, Lamia addresses a request for clarification *was bedeutet eNORMen;* (Segment 11), produced to be heard by the teacher as well as the students. This leads to a long pause (Segment 12), ended by Valeria who takes the floor and provides a synonym: *GROßen* (Segment 13). Lamia’s request is oriented to the use of *enormen* in the revised text. A comparison within the text reveals that *KI Chat* changed the wording from *große* (huge) to *enormen* (enormous) in one of the sentences:

**Figure 10**

*By KI Chat Revised Version (Above) and Lamia's Version (Below)*

Situationen, in denen sie kaum etwas zu essen haben. Sie stehen vor **enormen** finanziellen Problemen und können sich oft nicht einmal das Nötigste, wie Brot, leisten.

jetzt in Situationen wo die nichts zu essen haben. Die stehen in **große** finanzielle Probleme, die können einfach das Essen oder Brot nicht leisten. Die Menschen die

Valeria's response (Segment 13) is not taken up immediately. After a pause, both the teacher and Lamia take the floor in overlap: the teacher assessing Valeria's response as correct (Segment 15), Lamia displaying a lack of understanding (Segment 16), prosodically realized in a marked way making astonishment interpretable (Selting, 1995, p. 298). In the following, she makes explicit what her irritation refers to: *ja; ich HAB doch grOß geschrieben;=* (Segment 19). If *enorm* (enormous) is synonymous to *groß* (huge), why would *KI Chat* use "enormous"? There is overlapping talk by Valeria who treats Lamia's utterance in Segment 16 as a request for clarification, and provides further synonyms (Segments 20, 21). Lamia still keeps the floor, adding through latching another explicit request for information: *=warum sagt er eNORMen.* (Segment 22). This change in format from "what does X mean" to "why" reveals that Lamia does not perceive herself as the only author; what she wrote is changed by "him." This comment on authorship, however, is not oriented to by Valeria: She focuses on vocabulary work by repeating synonyms and, as sequential closure, the target item (Segments 23–25)—this time, aligned with a wide arm gesture as an additional resource to depict "largeness." Although the students are sitting in a row, Lamia may be able to perceive these large gestures in her peripheral view. Nevertheless, Lamia continues by realizing a comment, thereby displaying nonunderstanding (Segment 26). By using an "as-if" format, she makes explicit that her request for clarification with regard to the "why" has not been answered in a way that helps her understand (see Section 4.3 for continuation). Throughout this sequence, the teacher gazes at her laptop and types continuously, thus constructing herself as "not available" at this moment.

From a pedagogical point of view, this sequence is interesting: Whereas Lamia's use of *groß* (huge) is not wrong, the use of *enorm* (enormous) belongs more to academic language use. Having read the text revised by *KI Chat*, Lamia notices the change in wording (from "huge" to "enormous"). She addresses the meaning of the word used. However, her request for an explanation for the reason for the change remains unacknowledged. This sequential position would have presented the opportunity to explicitly address stylistic issues or questions regarding language use in different registers. However, due to the unavailability of the teacher and Valeria's orientation to the word meaning, this potential learning opportunity passes.

#### 4.3 Lamia's observations on the edited text II: Authorship in question

The following two extracts show students' actions that can be interpreted as comments regarding the authorship of the texts. Excerpt 04b is the continuation of Excerpt 04a:

**Transcript excerpt 04b: “enormous II” (DaZ-GTA\_04\_01:22:24)**

(01:22:24) 0026 **LAM** als ob ich das jetzt (.) mh verSTEhe;  
as if I now mh understand that

(01:22:27) 0027 (1.23)

(01:22:28) 0028 als ob jetzt mein: LEHRerin mich so  
as if my teacher like

(01:22:30) 0029 öh he glAubt dass ich das geSCHRIEben hab.  
uh would believe I wrote this

(01:22:31) 0030 **X** ((laughter))

(01:22:35) 0031 **VAL** weißt du ich KENN das wort;  
you know I know this word

(01:22:36) 0032 (0.56)

(01:22:36) 0033 **VAL** [von SPAnisch ]  
from spanish

(01:22:36) 0034 **LEH** [ihr habt jetzt nen] neuen block FREIgegeben;  
there is a new block available/editable now

(01:22:39) 0035 auf dem ARbeitsblatt.  
on your worksheet

(01:22:40) 0036 das heißt die überarbeitung könnt ihr jetzt  
EINFügen.  
that means you can now insert the revision

((Talk about technical problems, website is not loaded. Lamia states that she does not know where to insert the edited text. Then the teacher explains to Juma what to do next (Kupetz, in press)))

After her explication of nonunderstanding (Segment 26), Lamia holds the floor (Segments 28–29), again using the “as-if” format, this time to bring in the potential perspective of the (nonpresent German) teacher. By constructing the teacher as someone who would not believe she wrote the text, she distances herself from her text, constructs herself as “not” being its author. Valeria, however, is still oriented to the action of clarifying the meaning of “enormous” by making explicit that she can infer its meaning from her L1 Spanish (Segments 31–32). This is when the teacher takes the floor in order to instruct the students on where to paste in the revised texts in the digital learning platform (Segments 34–35). At this moment, it is revealed why she was “unavailable” before: She was occupied solving a technical problem with *Fobizz*.

Another moment in which the authorship of texts is dealt with ironically occurs about 2 minutes later. The teacher shows Lamia where to insert her text and explains to Juma what to do with *KI Chat*. In the course of this sequence, Juma reveals that she usually does not ask ChatGPT to revise anything; she would rather ask ChatGPT to fulfill complete tasks (Atanasoska, 2025; Kupetz, in press). Juma is asked by the teacher to start working, and the teacher turns to Medika, when Lamia opens a playful side sequence, by pinching Juma’s cheek:

**Transcript excerpt 05: “smart girls” (DaZ-GTA\_04\_01:24:39)**

(01:24:39) 0001 **LAM** ((knufft Juma in die Wange))  
((pinches Juma’s cheek))

0002 (( X )); ich hab NIE:  
i’ve never

0003 [so schlAue mädchen gesehen;]  
seen such smart girls before

0004 **JUM** [((lacht)) ]  
((laughs))

(01:24:42) 0005 JUM (nie/nee) ;=  
(never)

(01:24:43) 0006 JUM =<<lachend> hast du\_s Alles komplEtt alLEIne  
geschrieben.>  
did you write everything completely by yourself

(01:24:45) 0007 J&L ((lachen laut und klatschen ab))  
((laugh and high five))

((Talk about technical problems, inserting individual texts in Task sheet does not work.))

While laughing, Lamia comments (in a nonidiomatic way) on how she perceives Juma and herself as amazingly smart girls. The whole sequence is multimodally framed as playful and humorous. Juma joins in, by laughter, and she playfully takes up the idea of them being “great” authors by pointing at Lamia’s tablet and saying *hast du\_s Alles komplEtt alLEIne geschrieben.* (Segment 06). By imitating such a “teacher’s remark,” putting emphasis on having written “everything on their own,” Juma takes up this “perspective of others toward the girls’ writing skills.” Lamia and Juma then burst into laughter and the high five, thereby ending this side sequence.

Transcripts 04 and 05 thus show how the students position themselves: They reveal their assumptions on how their texts may be perceived by their teachers—that is, that they would not be considered the text’s author. In a similar way, the following extract (Transcript 06) reveals how Lamia distances herself from the language used in the text. The extract begins with the teacher’s invitation to comment on the text:

#### Transcript excerpt 06: “language-wise” (DaZ-GTA\_04\_01:27:15)

(01:27:15) 0001 LEH was habt ihr denn FESTgestellt;  
what did you find out

(01:27:17) 0002 eure (2.08) bei eurem TEXT.  
your in your text

(01:27:21) 0003 LAM was hab ich (.) FEST[gestellt;  
what did I find out

(01:27:22) 0004 MED [ich konnt\_s ni]ch ve  
I couldn’t com

(01:27:23) [noch nich verGLEICHen. ]  
not yet compare

(01:27:23) 0005 LAM [ich hab FESTgest ] [ellt,]  
I found out

(01:27:24) 0006 MED [°hh ]

(01:27:25) 0007 MED <<covering her mouth> mh;>

(01:27:25) 0008 (0.51)

(01:27:26) 0009 LAM dass öhm  
that uhm

(01:27:27) 0010 er mi:r SPRACHlich öhm alles verbEssert hat,  
that he corrected everything for me language-wise

(01:27:31) 0011 (0.61)

(01:27:32) 0012 LAM [in ] in eine SPRACHE,  
in into a language

(01:27:32) 0013 LEH [mhm,]

(01:27:35) 0014 LAM die ich nich so:  
that I do not

(01:27:36) 0015 (1.79) ((kreist mit den Händen))  
((circular hand movements))

(01:27:38) 0016 **LAM** SPREchen (.) oder SCHREIben (-)  
speak or write

(01:27:40) 0017 **JUM** <<pp, addressed to screen> hä:::? >

(01:27:41) 0018 **LAM** und äh aber  
and uh but

(01:27:43) 0019 er hat nIchts geÄNdert,  
he didn't change anything

(01:27:44) 0020 (1.17)

(01:27:45) 0021 nur öh bEsser zugefügte SÄTze geschrIeben.  
only uh wrote better (composed) sentences

(01:27:49) 0022 (0.65)

(01:27:50) 0023 **LEH** mhm,

(01:27:51) 0024 (0.77)

(01:27:51) 0025 **LEH** also wäre dein STICHwort-  
so your keyword would be

(01:27:53) 0026 SPRACHlich sehr gute überArbeitung==  
language-wise very good revision

(01:27:55) 0027 =INhaltlich eher mh::..  
content-wise more like mh

((Juma takes the floor, stating technical problems. The teacher addresses Valeria, inviting a reflection on how KI Chat revised Valeria's text. This is when a discussion about ChatGPT's abilities starts.))

The transcript starts with the teacher's request to point out what has been observed in terms of changes to the text (Segments 01–02). Lamia takes the floor, by repeating the question (Segment 03), in partial overlap with Medika who states that she has not yet been able to do a comparison (Segment 04). Lamia keeps the floor to state that *KI Chat* made changes to her text “into a language that I am not able to speak or write” (Segments 09–16), adding that no changes were made in terms of content (Segments 18–19). She resumes her turn by rephrasing the changes made as “better (composed) sentences” (Segment 21). This is taken up by the teacher through an acknowledgment token (Segment 23) and an upshot formulation (Segments 25–27) (Heritage & Watson, 1979). The teacher's formulation contains assessments (“language-wise very good revision,” “content-wise more like mh”) that will be the subject of later debate (see Section 4.4). At this point, it is noteworthy that the language-related changes made by *KI Chat* seem to result in Lamia no longer perceiving the text as “her own.”

#### 4.4 Lamia's reflections on the tool's capabilities

In the course of the following interaction, a debate between the teacher and the students develops with regard to the tool's abilities and the role of the prompt. Whereas the teacher underlines that *KI Chat* will not be of help content-wise, the students insist that it depends on the prompts. Remember, that Lamia's prompt was *Bearbeite bitte mein text*. The discussion culminates in the teacher's request to try out another prompt. Lamia inserts a new prompt (Figure 12) and reads the revised text (Figure 13), while the teacher is talking to Valeria about her results.

Figure 12

Screenshot 01:32:11 With New Prompt

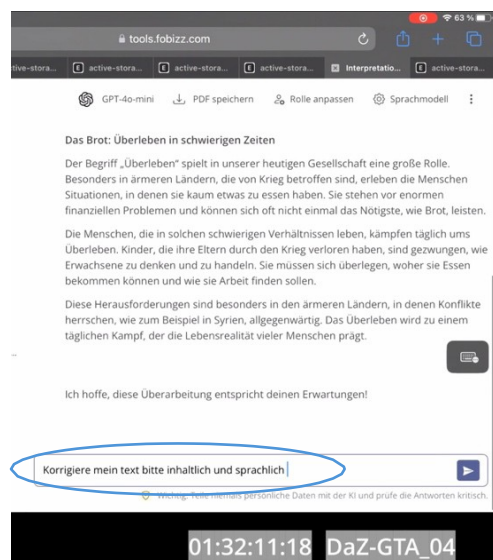
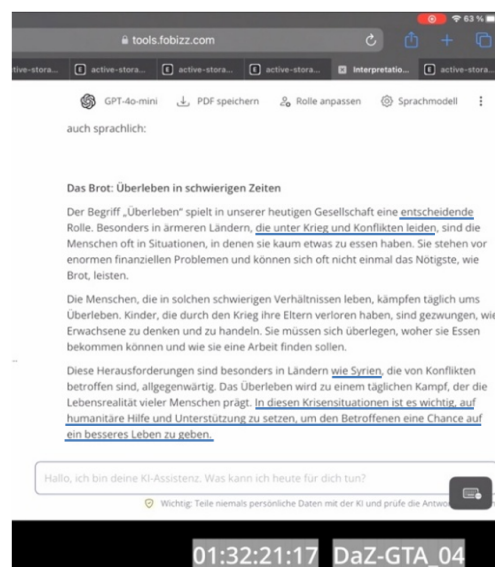


Figure 13

Screenshot 01:32:21 With Revised Text, Changes Highlighted



In the following minutes, the teacher raises concerns regarding the *KI Chat*'s capability to help with the interpretation of short stories, which is debated by Valeria. Note that the teacher and Valeria have been talking about *KI Chat*'s revisions of Valeria's text on the stylistic devices used (Task 1), whereas Lamia constantly orients to the text transposing the theme to the present day (Task 2).<sup>5</sup>

#### Transcript excerpt 07: "we need him" (DaZ-GTA\_04\_01:32:44)

- (01:32:51) 0007 **LAM** er hat nichts gemacht.  
he didn't do anything  
(01:32:52) 0008 (1.63) ((Leh leans over Lamia's tablet))  
(01:32:54) 0009 **LAM** finde ich.  
I think  
(01:32:54) 0010 (0.65)  
(01:32:55) 0011 **LAM** also  
well  
(01:32:56) 0012 (1.03)  
(01:32:57) 0013 **LAM** er hat nur:  
he only  
(01:32:58) 0014 (0.63)  
(01:32:58) 0015 **LAM** paar wörter rein (--) geschrieben.  
wrote a couple of words in it  
(01:33:00) 0016 **LEH** ja;  
yes  
(01:33:01) 0017 **LAM** [die ohne diese wör]ter können auch

<sup>5</sup> This is important, because it reveals how complex it is to negotiate between the classroom and the digital learning platform *Fobizz* and various areas/texts within *Fobizz*. It is not possible to reconstruct with any certainty whether teacher and students are making reference to the same text in the course of their discussion, although, from a German teaching perspective, it would seem important to distinguish between a *KI Chat* text revision on the analysis of a short story's stylistic devices and a revision of a text transposing the theme of hunger (in a postwar society) to the present day.

th- without these words can also

(01:33:01) 0018 JUM [ (X X X ) ]

(01:33:03) 0019 (1.67)

(01:33:04) 0020 LEH also:,  
well

(01:33:05) 0021 (0.85)

(01:33:06) 0022 LAM KANN er das <<p> nich.>  
he cannot do that

(01:33:07) 0023 LEH hhh°

(01:33:07) 0024 (0.46)

(01:33:08) 0025 LEH LEIder.  
unfortunately

(01:33:08) 0026 (0.38)

(01:33:09) 0027 MED aber [ich glaub]  
but i think

(01:33:09) 0028 LAM [aber ]  
but

(01:33:09) 0029 [hh° ]

(01:33:10) 0030 MED [ich glaube er hat]  
I think he

(01:33:10) 0031 ja den [TEXT nich;  
doesn't have the text ]

(01:33:11) 0032 LAM [<<f, lachend> OHne chat gi pi ti]  
without chat gpt

0033 MED [oder, ]  
or

LAM [kAnn ich nich] [LEben.> ]  
I cannot live

(01:33:12) 0034 JUM [<<p> leben.>]  
live

(01:33:13) 0035 LAM ((laughs))

(01:33:13) 0036 VAL ich hab den text [ihn ge][geben]  
I gave him the text

(01:33:14) 0037 MED [°h ] [oh ]

(01:33:15) 0038 MED (hm.)

(01:33:16) 0039 LAM aber er HILFT uns ja.  
but he does help us

(01:33:18) 0040 (0.46)

(01:33:18) 0041 mit zuSAMmenfassen,  
with summarizing

(01:33:19) 0042 [also ] an [mAnchen (--)] STEllen.  
well in some places

(01:33:19) 0043 JUM [ (X) ] [(au weia) ]  
(X) (oh dear)

(01:33:22) 0044 MED mh:;

(01:33:23) 0045 LAM ja; ich WEIß;  
yes I know

(01:33:24) 0046 wir haben auch fEstgestellt dass er auch das  
auch nich KANN,  
we also found that he also cannot do that

(01:33:26) 0047 ((laughs))

(01:33:27) 0048 aber  
but



(01:33:28) 0049 ((laughs))  
 (01:33:28) 0050 **JUM** aber trOtZdem werden wir ihn immer BRAUchen.  
 but still we will always need it  
 (01:33:30) 0051 **LAM** JA.  
 yes  
 (01:33:30) 0052 (0.71)  
 (01:33:31) 0053 **LEH** [mhhm,  
 (01:33:31) 0054 **MED** [((laughs quietly))  
 (01:33:32) 0055 **LEH** medika; was ist dEine erKENNTnis.]  
 medika what is your insight  
 ((Medika states that the chatbot changed a few words. Talk continues about the version of the LLM used - GPT4.0, the same as used by Chat GPT at the time of the recording.))

In Segment 7, Lamia takes the floor to comment on the revised text. Note that she designs her utterance in a way that establishes *KI Chat* as the author, by using *er* (he) and ascribing an action (didn't do anything) while downgrading the impact of her comment by attaching *fInde ich*. (I think) (Segment 09). She continues by specifying the changes she sees (wrote a couple of words in it), downgrading this revision by using *nur* (only) (Segments 13–14). Due to overlap with Juma, the following utterance is hardly intelligible. The overlap talk leads to a pause (Segment 19) that is brought to an end by the teacher who initiates an upshot formulation by using *also* (so) with lengthened final syllable and mid-rising intonation (Segment 20), thus designed as a designedly incomplete utterance (DIU) (Koshik, 2002). Lamia takes this up by syntactically embedding *KANN er das <p> nich.>* (he cannot do that) (Segment 22), thereby referring to the tool's inability to help with the analysis of short stories. The teacher continues with a long outbreath and an assessment (Segment 23), thereby orienting to the (anticipated) students' stance. This is responded to by both Medika and Lamia who start talking in overlap, establishing a kind of *schisming* (Egbert, 1997) in which two focused interactions are created (Medika talking to Valeria; Lamia talking to the teacher). Medika questions the basis of the tool's changes; she seems to assume that the self-written text was not provided to the tool (Segments 27, 30, 31, 33). Valeria orients to this assumption, clarifying that she did insert her text in the tool (Segment 36), and this is responded to by Medika with acknowledgment tokens (Segments 37, 38).

At the same time, Lamia uses *aber* (but) to project an objection (Segment 28) that is then realized in Segment 32: *Ohne chat gi pi ti kAnn ich nich LEben.* (without ChatGPT I cannot live). The utterance is produced while laughing and with high volume, which frames it as nonserious. It has the nature of an extreme case formulation (Pomerantz, 1986) used in this case by Lamia to defend her own objection to the teacher's claim that the AI-based tool was barely useful for revisions in terms of content. Note that Juma joins in through a co-construction (Segment 34), thereby signaling a high degree of agreement with Lamia's stance. Lamia then substantiates her claim, providing an argument for the tool's relevance (Segments 39, 41) that is then immediately restricted (Segment 42). Note that Lamia produces an explicit claim of knowledge on the tool's limits (Segments 45, 46), projecting a further claim of need (Segment 48) as well as laughter. At this point, Juma takes the floor to substantiate Lamia's claim *aber trOtZdem werden wir ihn immer BRAUchen* (but still we will always need it) (Segment 50). Both students thus acknowledge the teacher's objection regarding the tool's abilities, but insist on their perspective, implying a stance that implies the need for the tool. The teacher responds to this with a—notably short—recipency token (Segment 53), thereby not showing explicit agreement. She turns to Medika and asks her to share her observations.

From an analytic point of view, it is noteworthy that Lamia comments on the dimension of language-related help—that is, she states that words were added. An analysis of the text (Figure 13), however, reveals that one sentence was added: *“In diesen Krisensituationen ist es wichtig, auf humanitäre Hilfe und Unterstützung zu setzen, um den Betroffenen eine Chance auf ein besseres Leben zu geben”* (In these crisis situations, it is important to focus on humanitarian aid and support in order to give those affected a chance at a better life). This sentence introduces the topic of humanitarian aid, which was not included in Lamia’s text before. Lamia’s comment thus occurs in reaction to insufficient reading and/or understanding of the revised version of the text. Moreover, it remains unclear whether the participants themselves—Lamia, the teacher, and the other students—are aware of the fact that they refer to and comment on different texts and their revised versions (Valeria commenting on changes made to her Text 1, Lamia commenting on changes made to her Text 2). These observations will be discussed from a pedagogic point of view in the discussion.

## 5 Discussion and summary

In the following, I shall summarize the findings of the present case study and discuss them in the light of research needs (Section 5.1) and pedagogical considerations for working with generative tools in schools (Section 5.2). The present study’s sequential and multimodal analysis is based on video data from an additive GSL classroom setting in which an AI-based chatbot was used. A close analysis of the interactional process of one student’s text revision reveals the complexity of dealing in copresence with digital interaction spheres (Oloff & König, 2024): in this case, with a digital classroom in *Fobizz* including the tool *KI Chat*. A classroom setting in which students work on tablets with AI-based tools can thus be described as a complex semiotic ecology (Goodwin, 2000): Students need to carefully coordinate themselves with the ongoing classroom activities. They need to coordinate their interaction with the teacher, their interaction with each other, and their work on the tablet in which a tool generates new texts that need to be taken into consideration as well. It is an interactional task for all participants to track what is happening in the others’ digital learning environment; and quite often, from an analytical perspective, one cannot state for sure whether the participants are all “on the same page” in the sense of “talking about the same texts.” A fundamental concern for teachers should thus be to carefully consider the multimodal complexity that comes with working in the real-life classroom and in the digital sphere when giving instruction (García Garcá, 2025; Kupetz, 2021).

Furthermore, it has become clear, that the task of “revising an own text” with the help of an AI-based tool is not easily implemented (see also Kupetz, in press). The observations in terms of learning opportunities vary: On the one hand, it becomes obvious that the tool’s revisions are not necessarily fully read and/or understood. On the other hand, the revision process does seem to yield language-learning opportunities—for example, in terms of use of register.

Moreover, it has been shown that a small group setting such as the one analyzed here allows for comments and detailed reflections on the working process; and through this, it is possible to gain an analytical grasp of the students’ affective stance toward the AI-based tool and their positions on authorship and ownership of (revised) text: Lamia’s comments suggest that she does not perceive the revised text as her own (any more). Whereas the teacher may have oriented toward *KI Chat* as a tutoring tool providing feedback to Lamia’s text, Lamia herself treats it as a tutor with high epistemic authority in terms of language use (e.g., Transcript 06) but also as a ghostwriter (e.g., Transcripts 04, 05). The present study’s findings thus support Steinhoff and Lehnen’s (2025) GPT model that describes the different kinds of roles and practices involved in the writing processes of humans using AI-based tools.

Against the backdrop of the observations made, it is crucial to note that the roles and practices described in the model are not mutually exclusive, but that intersections and transitions of roles are mapped in the model (Steinhoff & Lehnert, 2025, p. 4) and may evolve in the course of a writing process.

### **5.1 Further research**

It has proven insightful to combine a CA perspective on the interaction in and about the writing process with text analysis based on classroom video data and screen videos. The present study, however, focused only on the revision phase in which an AI-based tool was used. Further research should address other phases of writing processes as well, and use data from L2 learners with lower language proficiency, because, presumably, other phenomena will be observed. Further research should also focus more on the learning opportunities (or obstacles) that come with the use of AI-based tools in writing (Neumann et al., 2024, p. 234) and take into consideration in a more systematic way the students' ordinary practices of writing with/through/despite chatbots (Kupetz in press). In this line of work, the whole dimension of affective stance, agency, and (perceived) authorship should also be addressed more thoroughly—for example, by drawing on combinations of classroom observations and think-aloud data.

### **5.2 Pedagogical challenges**

The insights gained from reconstructive work on classroom interaction and learners' texts are the basis for pedagogical and didactic considerations that are of a rather normative nature (see Kupetz, 2018, in press, for similar discussions). The following three points address the work with AI-based tools in the language classroom. Although these may not be unexpected to experienced educators, they may offer valuable insights and direction for novice teachers and practitioners inexperienced in working with AI-based tools. First, in the phase of lesson planning, it seems important to come to an understanding of the AI tool's role in the task at hand or in the working process. A possible approach may be to use Steinhoff and Lehnert's (2025) model as an orientation framework for clarifying the tool's role, and use it to address questions about authorship, agency, and so forth. Second, it seems important to carefully plan enough time for reading the texts revised (or produced) by the tool (Philipp, 2025). In the case study (and also in other instances in the corpus), students' comments on the text reveal that the texts are not fully read (or understood), so it is important to make sure that the tool's output can be processed thoroughly. Especially in the context of GSL, meticulous reading needs to be planned in order to take advantage of the language-learning opportunities provided. It might be worthwhile for teachers to consider letting students produce less in quantity, but work on the (revised) texts in more detail instead. Third, systematic and criteria-based classroom talk about the tool's role and the texts provided yields the chance to create meaningful learning moments and to foster discursive and reflective skills on language use and the use of AI-based tools. This is in line with previous studies underlining that students who do not have the opportunity to develop the skills to use AI-based tools for their learning processes and to reflect this usage critically, will lack major competencies in the future (Atanasoska, 2025, p. 260; Schindler, 2023, p. 18; see also Wei Dai et al., 2025, p. 45ff).

## References

- Atanasoska, T. (2025). „Für Matura-Ausarbeitungen war es wirklich ein Lifesaver!“ KI und Schreiben in der Schule. In H.-G. Müller & M. Fürstenberg (Eds.), *DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz* (pp. 245–265). Frank & Timme.
- Bajohr, H., & Hiller, M. (Eds.). (2024). *Das Subjekt des Schreibens. Über Große Sprachmodelle* (Special edition). Edition Text+Kritik.
- Bär, C. (2021). *Textproduktion in Interaktion – Eine Studie zum kollaborativen Schreiben in der Grundschule*. Peter Lang.
- Beisswenger, M. (2017). Sprechen, um zu schreiben: Zu interaktiven Formulierungsprozessen bei der kooperativen Textproduktion. In Y. Ekinici, E. Montanari, & L. Selmani (Eds.), *Grammatik und Variation – Festschrift für Ludger Hoffmann zum 65. Geburtstag* (pp. 161–174). Synchron.
- Cantarutti, M. N. (2023). Try-marking. In A. Gubina & C. W. Raymond (Eds.), *Encyclopedia of Terminology for Conversation Analysis and Interactional Linguistics*. International Society for Conversation Analysis (ISCA). [https://emcawiki.net/Encyclopedia\\_of\\_Terminology\\_for\\_CA\\_and\\_IL](https://emcawiki.net/Encyclopedia_of_Terminology_for_CA_and_IL)
- Couper-Kuhlen, E. (2009). A sequential approach to affect: The case of “disappointment.” In M. Haakana, M. Laakso, & J. Lindström (Eds.), *Talk in interaction – Comparative dimensions* (pp. 94–123). Finnish Literature Society.
- Domenech, M., & Leßmann, A.-C. (2022). Schreib-Interaktionen im DaZ-Förderunterricht der Grundschule – Eine medialitäts- und methodenintegrierende Fallstudie. *Zeitschrift für Interaktionsforschung in DaFZ (ZIAF)*, 2(1), 65–88. <https://doi.org/10.17192/ziaf.2022.2.8515>
- Dornicheva, D., & Sulimova, M. (2025). Chatbots im Schreibunterricht für fortgeschrittene Russisch-lernende. *SlavUn*, 02/2025, 30–56. <https://doi.org/10.20377/slavun-17>
- Egbert, M. (1997). Schisming: The collaborative transformation from a single conversation to multiple conversations. *Research on Language & Social Interaction*, 30(1), 1–51.
- Feick, D. (2025). Aktuelle Forschungstendenzen zum digitalen Wandel in Deutsch als Fremd- und Zweitsprache. In D. Feick & K. Biebighäuser (Eds.), *Digitale Lehr-/Lernressourcen und digitale Kompetenz – Forschung aus dem Hochschulkontext* (pp. 9–42). Erich Schmidt.
- García García, M. (2025). Developing classroom interactional competence: The case of instruction giving. In O. Sert & H. Z. Waring (Eds.), *Conversation analysis and language teacher education* (pp. 69–90). Springer Nature.
- Gardner, R. (2013). Conversation analysis in the classroom. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 593–611). Wiley-Blackwell.
- Geist, B., Kupetz, M., & Glaser, K. (2019). Accounting for spellings: Ambisyllabic consonants in L2 German spelling discussions. *Classroom Discourse*, 10(1), 71–98. <https://doi.org/10.1080/19463014.2019.1567361>
- Goffman, E. (1978). Response cries. *Language*, 54(4), 787–815.
- Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32, 1489–1522.
- Hall, J. K., & Looney, S. D. (2019). Introduction: The embodied work of teaching. In J. K. Hall & S. D. Looney (Eds.), *The embodied work of teaching* (pp. 1–14). Multilingual Matters.
- Helm, C., Große, C., & Österreichischer Bundesverlag Schulbuch (ÖBV). (2024). Einsatz künstlicher Intelligenz im Schulalltag: Eine empirische Bestandsaufnahme. *Erziehung und Unterricht*,

- 2024(3–4), 370–381. [https://www.oebv.at/images/product-images/Einsatz\\_kuenstlicher\\_Intelligenz\\_im\\_Schulalltag\\_Helm\\_Grosse\\_E-U.pdf](https://www.oebv.at/images/product-images/Einsatz_kuenstlicher_Intelligenz_im_Schulalltag_Helm_Grosse_E-U.pdf)
- Heritage, J., & Watson, R. (1979). Formulations as conversational objects. In G. Psathas (Ed.), *Every-day language: Studies in ethnomethodology* (pp. 123–162). Irvington.
- Knopp, M., & Schindler, K. (2025). Schöne neue Textwelt: Vertrauen in KI-generierte Texte. In S. Emmersberger & I. Kammerer (Eds.), *Mediale Praktiken in einer Kultur der Digitalität – Impulse für sprachliche und literarische Bildung* (pp. 139–159). J. B. Metzler.
- Koshik, I. (2002). Designedly incomplete utterances: A pedagogical practice for eliciting knowledge displays in error correction sequences. *Research on Language and Social Interaction*, 35(3), 277–309. [https://doi.org/10.1207/S15327973RLSI3503\\_2](https://doi.org/10.1207/S15327973RLSI3503_2)
- Kupetz, M. (2014). Empathy displays as interactional achievements—Multimodal and sequential aspects. *Journal of Pragmatics*, 61, 4–34. <https://doi.org/10.1016/j.pragma.2013.11.006>
- Kupetz, M. (2018). Gesprächsanalytische Unterrichtsforschung als Möglichkeit einer kasuistischen Lehrer\*innenbildung im Bereich sprachsensibler Fachunterricht. In C. Caruso, J. Hofmann, A. Rohde, & K. Schick (Eds.), *Sprache im Unterricht. Ansätze, Konzepte, Methoden* (pp. 49–67). Wissenschaftlicher Verlag Trier.
- Kupetz, M. (2021). Multimodalität und Adressatenorientierung in Instruktionen im DaZ- und fachintegrierten Unterricht. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, (22), 348–389. <http://www.gespraechsforschung-online.de/fileadmin/dateien/heft2021/si-kupetz.pdf>
- Kupetz, M. (in press). Die Nutzung eines KI-gestützten Chat-Tools als interaktive Herstellungsleistung – Eine exemplarische Rekonstruktion anhand von additivem DaZ-Unterricht. In S. Guldenschuh & K. Ulrich (Eds.), *Digitalität und Deutsch als Zweitsprache*. Waxmann.
- Lehnen, K. (2000). *Kooperative Textproduktion – Zur gemeinsamen Herstellung wissenschaftlicher Texte im Vergleich von ungeübten, fortgeschrittenen und sehr geübten SchreiberInnen* (Doctoral dissertation). Bielefeld University, Germany.
- Lehnen, K. (2014). Gemeinsam schreiben. In H. Feilke & T. Pohl (Eds.), *Schriftlicher Sprachgebrauch. Texte verfassen* (pp. 414–431). Schneider Verlag Hohengehren.
- Lehnen, K., & Steinhoff, T. (2024). Digitales Lesen und Schreiben. In J. Androutsopoulos & F. Vogel (Eds.), *Handbuch Sprache und digitale Kommunikation* (pp. 527–546). De Gruyter.
- Lingnau, B., & Mehlem, U. (2012). Interaktive Entstehung von Wortschreibungen mehrsprachiger Kinder im ersten Schuljahr. In W. Gießhaber & Z. Kalkavan (Eds.), *Orthographie- und Schriftspracherwerb bei mehrsprachigen Kindern* (pp. 143–166). Fillibach.
- Lucius-Hoene, G., & Deppermann, A. (2004). Narrative Identität und Positionierung. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 5, 166–183.
- Maahs, I.-M., Lawida, C., Gutenberg, J., & Drews, K. (2023). Wie nutzen Schülerinnen und Schüler digitale Medien zum sprachlichen Lernen in allen Fächern? Eine Fragebogenstudie mit Jugendlichen zur Nutzung digitaler Medien als (mehr-)sprachliche Hilfen. *Medienimpulse*, 61(2), 1–43. <https://doi.org/10.21243/mi-02-23-15>
- Markou, V. (2024). Chatten mit Chatbots! – Die schriftliche Online-Interaktion mit KI im DaF/DaZ-Unterricht. *Zeitschrift für Interaktionsforschung in DaFZ*, 4, 211–224. <https://doi.org/10.17192/ziaf.2024.4.1.8630>
- Müller, H.-G., & Fürstenberg, M. (2023). Der Sprachgebrauchsautomat: Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. *Mitteilungen des Deutschen Germanistenverbandes*, 70(4), 327–345.



- Neumann, K., Kuhn, J., & Drachsler, H. (2024). Generative künstliche Intelligenz in Unterricht und Unterrichtsforschung – Chancen und Herausforderungen. *Unterrichtswissenschaft*, 52, 227–237. <https://doi.org/10.1007/s42010-024-00212-6>
- Oloff, F., & König, K. (2024). Face-to-face & digital: Die interaktive und multimodale Herstellung digitalisierter Handlungsräume in Kopräsenz. *Zeitschrift für Germanistische Linguistik*, 52(1), 150–182. <https://doi.org/10.1515/zgl-2024-2007>
- Philipp, M. (2025). Die didaktischen Fragen, die KI aus Sicht des Lesens zum Zweck des Lernens aufwirft. In H.-G. Müller & M. Fürstenberg (Eds.), *DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz* (pp. 39–63). Frank & Timme.
- Pitsch, K., & Ayaß, R. (2008). Gespräche in der Schule – Interaktion im Unterricht als multimodaler Prozess. In H. Willems (Ed.), *Lehr(er)buch Soziologie: Für die pädagogischen und soziologischen Studiengänge* (Vol. 2, pp. 959–982). Verlag für Sozialwissenschaften.
- Pomerantz, A. (1986). Extreme case formulations: A way of legitimizing claims. *Human Studies*, 9(2–3), 219–229.
- Saxon State Ministry of Culture (Sächsisches Staatsministerium für Kultus) (2021). *Lehrplan für Vorbereitungsklassen/ Vorbereitungsklassen an allgemeinbildenden Schulen: Deutsch als Zweitsprache*. SMK/Sächsisches Staatsinstitut für Bildung und Schulentwicklung. <https://publikationen.sachsen.de/bdb/artikel/19713/documents/71547>
- Schindler, K. (2023). ChatGPT oder Überlegungen zu den Veränderungen des Schreibens in der Schule. *Medien im Deutschunterricht*, 5(2), 1–22. <https://doi.org/10.18716/ojs/midu/2023.2.5>
- Schneegass, R. (2025). Schreibprozesse im Wandel: Eine Analyse KI-gestützter Schreibworkflows von Studierenden. *Leseräume – Zeitschrift für Literalität in Schule und Forschung*, 11, 1–21.
- Schulz, L., & Gunkel, G. (2025). Der Einsatz generativer KI zur Förderung des Textverstehens im schulischen Fachunterricht. Strategien zur Individualisierung des Lernprozesses. In L. Baum & G. Günay (Eds.), *Künstliche Intelligenz in DaF/DaZ* (pp. 163–175). Frank&Timme.
- Selting, M. (1995). *Prosodie im Gespräch – Aspekte einer interaktionalen Phonologie der Konversation*. Niemeyer.
- Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., . . . Uhmann, S. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion*, 10, 353–402.
- Steinbach, A. (2025). Using AI to master writing styles in English lessons. *Praxis Englisch*, 3, 23–27.
- Steinhoff, T. (2022). Die digitale Transformation des Schreibens. In M. Knopp, N. Bulut, K. Hippmann, S. Jambor-Fahlen, M. Linnemann, & S. Stephany (Eds.), *Sprachliche Bildung in der digitalisierten Gesellschaft – Was wir in Zukunft wissen und können müssen* (pp. 143–159). Waxmann.
- Steinhoff, T. (2023). Der Computer schreibt (mit): Digitales Schreiben mit Word, WhatsApp, ChatGPT & Co. als Koaktivität von Mensch und Maschine. *Medien im Deutschunterricht*, 5(1), 1–16. <https://doi.org/10.18716/ojs/midu/2023.1.4>
- Steinhoff, T., & Lehnen, K. (2025). Schreiben mit künstlicher Intelligenz: Das GPT-Modell (Ghost, Partner, Tutor). *Leseräume – Zeitschrift für Literalität in Schule und Forschung*, 11, 1–14.
- Stierwald, M. (2020). Sprachtransfer innerhalb mehrsprachiger konversationeller Schreibinteraktionen? Eine Konversationsanalyse zu Reparaturen beim interaktiven Formulieren. *Bulletin suisse de linguistique appliquée*, No spécial, 161–182.

- Tadres, J. (2025). aso es isch jo ElGentlich voll okE? – Die interaktionale Organisation von Überarbeitungssequenzen während Schreibkonferenzen auf der Sekundarstufe II. *Didaktik Deutsch*, 30(58), 75–96. <https://doi.org/10.21248/dideu.788>
- Wei Dai, D., Zhu, H., & Chen, G. (2025). How does interaction with LLM-powered chatbots shape human understanding of culture? The need for Critical Interactional Competence (CritIC). *Annual Review of Applied Linguistics*, 45, 28–49. <https://doi.org/10.1017/S0267190525000054>
- Weiss, Z., Woerfel, T., & Meurers, D. (2023). Intelligente digitale Werkzeuge in der sprachlichen Bildung – Chancen und Herausforderungen. In M. Becker-Mrotzek, I. Gogolin, H.-J. Roth, & P. Stanat (Eds.), *Grundlagen der sprachlichen Bildung* (pp. 185–197). Waxmann.
- Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System*, 124, Article 103351. <https://doi.org/10.1016/j.system.2024.103351>

Author's address:

Maxi Kupetz, Universität Leipzig, Herder-Institut, Beethovenstraße 15, 04107 Leipzig, Germany  
[maxi.kupetz@uni-leipzig.de](mailto:maxi.kupetz@uni-leipzig.de)



Alyssa Kutzner

## Cross-linguistic writing with AI

Investigating transfer effects in L1/L2 and FL contexts

### Abstract

This study explores how multilingual 9th-grade students use generative AI as a writing assistant in German (L1/L2) and English (FL) writing contexts by focusing on strategy use, cross-linguistic transfer, and proficiency-related variation. Drawing on a mixed-methods design that combines qualitative and quantitative data, the analysis suggests that generative AI may introduce a new mediating layer in the writing process, enabling learners to externalize cognitive subtasks such as planning, drafting, and revising. Five main AI writing strategies were identified, ranging from planning and linguistic optimization to metacognitive reflection. Overall, findings indicate patterns consistent with cross-linguistic transfer of AI writing strategies in this exploratory sample, highlighting AI's potential to foster metacognitive awareness and cross-linguistic strategy development in multilingual writing. However, proficiency appears to moderate these patterns: More proficient (Level B2) learners of English as a foreign language show greater strategic stability in their AI use, whereas lower-proficiency learners engage with AI in a more situational and cautious manner, a pattern that is also reflected in two illustrative learner cases.

**Keywords:** AI-assisted writing • cross-linguistic transfer • AI writing strategies • multilingual education • language proficiency

### Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 123–157

DOI: 10.21248/dideu.971

**Copyright** Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0 veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

## 1 Introduction and objectives

Writing in school increasingly unfolds at the intersection of cognitive strategy use, multilingual repertoires, and emerging AI tools. The advent of generative artificial intelligence (AI)—particularly large language models (LLMs) such as GPT-3.5, GPT-4.0, the recently released GPT-5.0, or even GPT-5.2—has transformed how writing is conceptualized and taught. Learners now access systems that can generate coherent, context-sensitive text and provide feedback on structure, style, and content, thereby challenging established notions of authorship, strategy use, and language learning. This development necessitates a rethinking of writing instruction: How can AI be meaningfully embedded without undermining autonomy or critical awareness? This question has been discussed in both emerging international research on AI-mediated higher-order regulation and evaluative processes in writing (e.g., Gonsalves, 2024; Shen & Teng, 2024; Song & Song, 2024) as well as in German-speaking contexts (e.g., Fürstenberg & Müller, 2024; Rezat & Schindler, 2025; Steinhoff & Lehnert, 2025). Although research has long shown that effective writers plan, draft, and revise strategically (Harris et al., 2010; Lei, 2016; Manchón & Roca De Larios, 2007; Raoofi et al., 2017; Roca De Larios et al., 2001), and that such strategies can transfer across languages (Cummins, 1980, 2000; García & Wei, 2015; Herdina & Jessner, 2002; Odlin, 1989), less is known about how learners mobilize these resources when a generative AI assistant is available throughout the writing process.

The present contribution is part of a project on the transformative impact of generative AI on learner writing in secondary education. As the second of three interconnected studies, it shifts from learners' L1/L2 (German) writing to multilingual contexts and examines the potential of *AI writing strategies* to support foreign-language (FL) learning and cross-linguistic transfer. Under authentic classroom conditions in a Grade-9 German *Gymnasium* (academic track secondary school), learners were introduced to AI as a writing assistant and encouraged to deploy it autonomously and purposefully across planning, drafting, and revising phases of summary writing in German (L1/L2) and English (FL). Using a mixed-methods design, standardized reading assessments (*LGVT 5-12+*; *Cambridge English B1 Preliminary for Schools*) and student–AI chat logs were analyzed to identify patterns of AI writing strategies and to examine the extent to which these strategies transfer across languages.

Building on models of writing, cross-linguistic influence, and metacognition, AI was treated as a potential scaffold for planning, drafting, and revising—benefits that depend on learners' critical control of AI output. Accordingly, the study addresses three questions:

1. What observable patterns of AI usage emerge in L1/L2 and FL writing contexts?
2. To what extent does transfer of *AI writing strategies* occur across languages?
3. How does the use of *AI writing strategies* relate to language proficiency in L1/L2 and FL contexts?

Preliminary findings from the first study within this research project (Kutzner, in press) indicate that learners develop emerging *AI writing strategies*: flexible repertoires that learners use to strategically decompose writing tasks and use AI for planning, drafting, or revision support. These strategies appear to resemble established writing scaffolds and hold potential for fostering writing development, particularly in multilingual or foreign language contexts. Comparable observations have also been reported in recent international research on AI-supported English as a foreign language (EFL) writing (e.g., Xu et al., 2025). Furthermore, findings suggest that interaction with AI may enhance learners' metalinguistic awareness, aligning with principles of the *noticing hypothesis* (Schmidt, 1990) proposing that *noticing* facilitates language learning by helping learners become aware of gaps between their own output and

target-language input. For instance, learners report paying closer attention to lexical and syntactic patterns suggested by AI (Kutzner, in press); and this, in turn, appears to affect their linguistic output in both their first and foreign language writing.

By mapping strategy profiles onto language proficiency indicators and quantifying cross-linguistic overlap, the present study contributes: (a) an empirically grounded typology of AI-supported writing behaviors in school contexts; (b) indications on their transferability between L1/L2 and FL; and (c) implications for teaching AI-informed, process-oriented writing. The study raises important questions for the integration of AI in language teaching, particularly in fostering AI literacy (Allen & Kendeou, 2023; Ng et al., 2021) and strategic competence in multilingual learners. It also contributes to understanding how artificial interlocutors can scaffold writing in ways that support noticing, transfer, and cross-linguistic development.

## 2 Theoretical framework

This study is primarily grounded in process-oriented writing theory and models of cross-linguistic transfer. Concepts such as *noticing* and *AI literacy* serve as complementary perspectives to contextualize AI-mediated strategy use. Writing is widely conceptualized as a complex cognitive activity involving planning, text production, and revision (Hayes, 2012; Hayes & Flower, 1980). Skilled writers employ flexible writing strategies that enable them to manage cognitive load and maintain coherence. Research in writing pedagogy demonstrates that explicit instruction in such strategies significantly improves writing quality and self-regulation (Harris et al., 2010; Harris et al., 2011). Writing strategies support the writing process across its various phases (Hayes, 2012; Hayes & Flower, 1980), helping writers to plan, compose, and revise their texts more effectively.

Research on cross-linguistic influence and metacognition provides a comprehensive framework for understanding how linguistic and cognitive processes interact in multilingual learners. Cummins's (1980, 2000) *common underlying proficiency* (CUP) model assumes that bilingual individuals draw on a shared reservoir of cognitive and linguistic abilities, enabling knowledge and skill transfer across languages (see also Swain & Lapkin, 1995) once a sufficient proficiency threshold is reached (*threshold hypothesis*: Cummins, 1976). Complementing this view, Odlin (1989) distinguishes between positive (facilitative) and negative (interfering) transfer effects arising from similarities and discrepancies between languages. Later models such as Herdina and Jessner's (2002) *dynamic model of multilingualism* (DMM) and García and Wei's (2015) concept of translinguaging reconceptualize multilingual competence as an adaptive and fluid system in which learners flexibly draw on their full linguistic repertoire to construct meaning. Within this dynamic perspective, writing in multilingual contexts activates an integrated set of linguistic and metacognitive resources. Strategic skills developed in one language—particularly those related to planning, organization, and revision—can thus transfer to others. This process is mediated by metalinguistic awareness (Jessner, 2006) and *noticing* (Schmidt, 1990) that enable learners to consciously monitor linguistic form, compare output with target input, and refine their writing across languages. Schmidt's (1990) *noticing hypothesis* provides a key theoretical link between awareness and learning: Language acquisition occurs when learners notice the gap between their output and target-language input. In writing, this noticing can be facilitated through comparison, reflection, and feedback.

Generative AI introduces a new mediating layer in writing processes. Through prompting, learners can externalize subtasks such as planning, drafting, and revising (Putri et al., 2025). In doing so, AI reshapes

the ecology of writing (Warschauer et al., 2023), and holds potential for fostering strategy development and metacognitive reflection. However, these benefits depend on learners' ability to regulate and evaluate AI output critically—a central component of AI literacy (Allen & Kendeou, 2023; Ng et al., 2021). Although research on AI-assisted writing has increased since the release of ChatGPT 3.5 in late 2022, significant gaps persist. Most existing studies focus on university students and emphasize the advantages of AI such as its support for idea generation, feedback, and language enhancement (Kim et al., 2025; Li, 2025; Siregar & Hz, 2025; Werdiningsih et al., 2024). Studies involving school-aged learners are fewer, but likewise tend to address isolated stages of the writing process—for example, AI-supported feedback rather than holistic writing development (Castro et al., 2025; Chen et al., 2024; Kangasharju et al., 2022; Zheldibayeva et al., 2025; or, for German L1 contexts, Fürstenberg, 2025; Pitzen, 2025). However, only a few of these studies explore how learners strategically integrate AI across the entire writing process; and this holds particularly for elementary and secondary education (Alghasab, 2025; Chan & Lee, 2025; or, for German L1 contexts, Bodora et al., 2025). Existing research often examines isolated stages such as revision or editing, or raises concerns about potential overreliance on AI without addressing broader, process-oriented strategies (Putri et al., 2025). Yet, such practices may foster noticing and strategic flexibility, especially among multilingual learners who navigate between linguistic systems. This study therefore seeks to advance the understanding of how these strategies are applied and potentially transferred across L1/L2 and FL contexts. Because empirical research in this area is still limited, the study adopts an exploratory design aimed at identifying recurring patterns.

### 3 Context and study design

The study was conducted at a public German *Gymnasium* (academic track secondary school) in the state of North Rhine-Westphalia during the 2024/25 school year with 20 ninth-grade students (aged 14–15). German was the primary language (L1) for most learners, although 13 students reported bilingual backgrounds (e.g., Turkish, Russian, Arabic). English was taught as a compulsory foreign language (FL) from Grade 2 onward, and, as stipulated by the Ministry of School and Education (*Ministerium für Schule und Bildung*, 2019), students in Grade 9 at German *Gymnasien* (G9 track) are expected to achieve at least level B1 of the Common European Framework of Reference for Languages (CEFR). All students had access to personal laptops provided by the school. Although they were familiar with digital learning tools, the integration of generative AI into the writing process was a rather novel experience, at least in classroom contexts. Written consent was obtained from students and parents.

The research followed a comparative, cross-linguistic, mixed-methods design with one assessment each in German (L1/L2) and English (FL) conducted under authentic classroom conditions. Following an introduction to AI's potential as a writing assistant in the writing process, learners were encouraged to use AI autonomously and purposefully throughout the writing process. This design allowed learners to build individual scaffolds to support their planning, composing, and revising activities. Each assessment included a standardized reading test, an AI-supported writing task, and a survey on learners' experiences and attitudes. The entire assessment was conducted over the course of three consecutive days.

Reading proficiency served as a cognitive indicator of language proficiency and was measured using the *LGVT 5-12+* (Schneider et al., 2017) for German (L1/L2)—assessing reading comprehension, speed, and accuracy—and the *Cambridge English* (2018) *B1 Preliminary for Schools* reading test for English (FL). Reading proficiency was selected as an indicator of broader language competence, because AI-

assisted writing requires learners to read, interpret, and evaluate AI-generated output. Reading comprehension integrates decoding, vocabulary knowledge, and higher-order inferencing processes that overlap with cognitive demands of writing. Writing competence, by contrast, is difficult to operationalize independently of surface-level accuracy in small exploratory samples. The Cambridge Reading component was chosen for English due to its CEFR alignment and its separate assessment of reading, ensuring construct clarity and cross-linguistic comparability with the LGVT.

Following the reading component, learners completed a summary-writing task based on a short story (*USE 2030* by Bob Blume, 2025, 1,328 words, for German; *The Text* by Brendan Dunne, 2012, 894 words, for English). Students wrote a summary (in German, an *Inhaltsangabe*)—that is, the English and German realizations of the same genre-specific text pattern. In line with current pedagogical practice in Germany, EFL instruction follows a *task-supported language learning* (TSLL) approach. Unlike fully task-based approaches, TSLL embeds communicative tasks within a syllabus that focuses on particular topics or linguistic structures (Müller-Hartmann & Schocker-von Ditfurth, 2011, p. 24). Within this framework, tasks serve as additional pedagogic methods that can be utilized to achieve various curricular goals such as developing intercultural, literary, and communicative competencies (Summer, 2012, p. 10). Similarly, writing instruction in German (L1/L2) education promotes *profilierete Schreibaufgaben* ([authentic and contextually embedded writing tasks]; Bachmann & Becker-Mrotzek, 2010) that situate writing within authentic, functional, communicative, and social contexts to enhance motivation and writing quality. In both the German and English assessments, the task design and accompanying instructions were aligned with these pedagogical principles, ensuring that the writing tasks were authentic, purpose-oriented, and communicatively meaningful (in accordance with Breen, 1985, pp. 61–69; Ellis, 2009, p. 6; Guariento & Morley, 2001, pp. 349–351). In both contexts, the summary task was embedded in a simulated classroom conference scenario in which students should present a short story to their peers. The summary (in German, an *Inhaltsangabe*) thus functioned not merely as a self-directed writing exercise but as an audience-oriented informational text intended to inform and engage classmates who had not read the story. This framing situates the writing task within a socially and communicatively meaningful classroom practice. Learners were encouraged to use GPT-4o via the GDPR-compliant educational platform *fobizz.com* to assist in planning, drafting, and revising their texts. All interactions were automatically logged.

The dataset includes 40 AI chat logs (20 for German and 20 for English), the corresponding student texts, as well as reading test results. All AI prompts were successfully categorized into the identified AI writing strategies; no prompt fell outside the analytical framework. Quantitative data were analyzed using Excel and JASP for descriptive and correlational statistics, whereas qualitative data were analyzed in MAXQDA using qualitative content analysis (Kuckartz & Rädiker, 2023). The analysis focuses on the relationship between reading performance and prompt-based strategy use in AI-assisted writing.

#### 4 Methods and data analysis

The study was designed as an exploratory mixed-methods investigation combining quantitative and qualitative analyses to examine the relationship between reading proficiency and AI-assisted writing behavior. The study employed standardized reading tests *LGVT 5-12+* (Schneider et al., 2017) for German (L1/L2) and *Cambridge English* (2018) *B1 Preliminary for Schools Reading Test* for English (FL) alongside qualitative content analysis (Kuckartz & Rädiker, 2023) using MAXQDA in order to examine the relationship between patterns of AI strategy use and linguistic competence using reading proficiency as an indicator. All quantitative results are strictly descriptive and serve exploratory pattern

identification rather than inferential validation. Given the small sample size, statistical precision should not be overinterpreted.

The *LGVT* reading test was scored according to the normative tables for Grade 9 *Gymnasium* (academic track secondary school). Raw scores were adjusted using the test-specific timing correction factor; corresponding percentile ranks were determined and subsequently converted into standardized *T* scores. Although percentile ranks represent an individual's relative standing within the normative sample, they are nonlinear and compress differences at the distributional extremes. In contrast, *T* scores ( $M = 50$ ,  $SD = 10$ ) are based on the normal distribution, allowing for meaningful statistical operations, cross-test comparability, and precise interpretation of distance from the mean. According to the *LGVT* manual, values below 30 indicate *very low* performance (PR 0–2.3); 30–39, *below average* (PR 2.3–15.9); 40–49, *lower average range* (PR 15.9–50); 50–59, *upper average range* (PR 50–84.1); 60–69, *above average* (PR 84.1–97.7); and 70–76, *very high* performance (PR 97.7–100).

The *Preliminary* raw results were converted into *Cambridge English Scale Scores*, enabling direct comparability across test versions. According to the *Cambridge English Scale* (2025), scores of 160–170 correspond to *Grade A* (CEFR level B2); 153–159, to *Grade B* (B1); 140–152, to *Grade C* (B1); and 120–139, to *Fail* (A2 level).

Table 1 presents the descriptive statistics (mean, median, standard deviation, range, minimum, and maximum) for the three *LGVT* subtests—reading comprehension (RC), reading speed (RS), and reading accuracy (RA)—alongside the *Cambridge English Scale Score*.

**Table 1**

*Descriptive Statistics for LGVT and Cambridge English Scale Score*

	Reading comprehension (RC) ( <i>Leseverständnis (LV)</i> ) <i>T</i> score	Reading speed (RS) ( <i>Lesegeschwindigkeit (LGS)</i> ) <i>T</i> score	Reading accuracy (RA) ( <i>Lesegenauigkeit (LGN)</i> ) <i>T</i> score	<i>Cambridge Scale Score</i>
<i>M</i>	44.35	42.16	52.75	147.55
<i>Mdn</i>	43.50	41.00	50.50	150.00
<i>SD</i>	8.60	7.36	10.85	14.26
Range	30.00	33.00	40.00	50.00
Min	30.00	27.00	33.00	120.00
Max	60.00	60.00	73.00	170.00

As shown in Table 2, all *LGVT* subtests (reading comprehension, reading speed, and reading accuracy) and the *Cambridge English Scale Score* are significantly intercorrelated, indicating that they assess closely related dimensions of reading competence. This pattern suggests that the tests capture overlapping cognitive processes involved in higher-level reading such as text comprehension, inference generation, and the integration of information across sentences. The strong associations between comprehension, speed, and accuracy imply that efficient decoding and fluent word recognition support overall understanding, whereas comprehension performance in turn reflects the successful



coordination of these subskills. The close relationship between the *LGVT* performance and the *Cambridge English Scale Score* further supports the validity of both instruments as complementary measures of general reading proficiency across L1/L2 and FL contexts. Accordingly, all subsequent analyses considered the three *LGVT* subtests alongside the *Cambridge English Scale Score* to provide a comprehensive and multidimensional account of learners' reading competence.

**Table 2**

*Spearman Correlations ( $r_s$ ) Between LGVT Subtests and Cambridge English Scale Score*

Variable		Reading comprehension (RC)	Reading speed (RS)	Reading accuracy (RA)	Cambridge Scale Score
1. Reading comprehension (RC)	$r_s$	—			
	$p$	—			
2. Reading speed (RS)	$r_s$	.834***	—		
	$p$	< .001	—		
3. Reading accuracy (RA)	$r_s$	.837***	.511*	—	
	$p$	< .001	.021	—	
4. Cambridge Scale Score	$r_s$	.702***	.600**	.630**	—
	$p$	< .001	.005	.003	—

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

A detailed overview of the results is provided in Appendix A in which participants' *Cambridge Preliminary* reading scores are sorted according to their performance and accompanied by the corresponding *LGVT* T-scores for reading comprehension (RC), reading speed (RS), and reading accuracy (RA) for each learner.

Whereas the quantitative measures provided a detailed profile of learners' reading proficiency, the qualitative analysis focused on their interactional and strategic behavior when engaging with AI during writing. The qualitative content analysis was conducted combining inductive and deductive coding procedures. The analytical framework drew on AI writing strategies that were inductively developed in a previous study of similar design (Kutzner, in press). In that earlier research, student–AI chat logs ( $N = 20$ ) were openly reviewed, and prompts were coded inductively in order to identify recurring functional actions in learners' interaction with AI. From this coding process, five categories of AI writing strategies were derived. AI writing strategies are understood as an extension of established writing strategies (e.g., Harris et al., 2010; Lei, 2016; Manchón & Roca De Larios, 2007; Raoofi et al., 2017; Roca De Larios et al., 2001). They regulate not only the learner's own writing process but also the deliberate integration of AI into specific phases of that process. Importantly, the prompts themselves were not conceptualized as strategies. Rather, they were treated as observable operationalizations of underlying strategic decisions made by learners. The strategy labels build directly on the previously coded strategy categories. Each label combines: (a) an orientation descriptor reflecting the learner's strategic orientation (e.g., focused, holistic, functional, differentiated, metacognitive) with (b) a term referring to the function of the strategy in the writing process (e.g., structuring, optimizing, pragmatizing, limiting, questioning), thus describing how and for what purpose learners use AI in writing. The



identified strategy categories included (the example prompts are illustrative model prompts that represent typical patterns identified in the data):

1. *Focally-structuring strategies* aimed at planning and organizing the writing process—for example, by generating outlines or writing plans (“Give me a writing plan to write a summary of a short story”).
2. *Holistically-optimizing strategies* supporting linguistic and stylistic elaboration—for instance, through requests for useful phrases or synonyms (“Tell me useful phrases to write a summary of a short story”).
3. *Functionally-pragmatizing strategies* oriented toward efficient task completion, such as delegating larger text operations to AI (e.g., generating a full summary).
4. *Differentially-limiting strategies* characterized by selective AI use, often limited to one or very few prompts.
5. *Metacognitively-inquiring strategies* serving reflective purposes including feedback-seeking (“Is this a good summary?”), revision support (“Give me a revision plan to revise a summary of a short story.” “Is the following sentence correct: ‘. . .’?”), and clarification of content or language (“What does X mean?”).

Because individual students employed multiple strategies, strategy use was not treated as mutually exclusive or as fixed learner types. Rather, learners shifted between different strategies over the course of the writing process. For example, planning may be characterized by Strategy-1-typical structuring practices; drafting, by Strategy-2-typical linguistic optimization; and revision, by Strategy-5-typical metacognitive inquiry. In line with process-oriented writing models (Hayes, 2012; Hayes & Flower, 1980), learners oscillated between planning, drafting, and revising, while AI assumed different functional roles—such as cognitive offloading, linguistic support, or reflective scaffolding—within this dynamic process. Thus, the strategies are conceptualized as situational and phase-sensitive usage patterns in AI-supported writing rather than as stable user types. In this sense, “phase-sensitive” does not imply a one-to-one correspondence with the classical three-phase writing model (planning, drafting, revising). Although certain strategies can be associated with particular writing phases, not all learners used AI in every phase, and some phases were completed entirely without AI support. Rather, the data indicate that learners selectively integrated AI depending on individual preferences, because the use of AI was neither obligatory nor uniformly distributed across phases. The strategies therefore do not map directly onto the three-phase writing model, but instead reflect selective and context-dependent configurations of AI integration within the broader writing process.

In addition to the five main strategies, the previous study identified three supplementary categories that, while not representing independent strategy types, nonetheless revealed meaningful patterns of interaction with AI. These were *greeting/politeness* (phatic communication such as greetings or expressions of gratitude), *translation* (requests primarily for linguistic support), and *off-task activities* (playful prompts, e.g., asking for jokes or pop-cultural stories, playing Battleship, writing poems, or posing absurd tasks such as “ $2 + 2 = 5$ ”<sup>1</sup>).

In the present study, the five main strategies are applied deductively to an independent sample in order to analyze their occurrence as well as the cross-linguistic transfer of AI writing strategies

<sup>1</sup> Similar patterns of playful prompting have been observed among Grade 4 elementary school students (Kutzner & Schindler, 2026).

between L1/L2 (German) and FL (English). In doing so, the study not only extends a previously developed typology of AI-assisted writing, but also provides an initial, careful validation of this typology by demonstrating its applicability to a new, independent dataset (albeit still rather small at  $N = 20$ ). Whereas the original study combined multiple data sources to contextualize these strategies (student–AI chat logs, questionnaire responses [Likert-scale and open-ended responses], and characteristics of the corresponding student texts; see Kutzner, in press), the present analysis focuses exclusively on strategy use in relation to learners’ reading proficiency. This allows for a more transparent and analytically focused examination of how learners employ AI writing strategies across languages. Selected student texts were additionally consulted to illustrate how these strategies are reflected in actual written products.

The supplementary categories identified in the earlier study were adapted for the present dataset. Because the current data did not include *off-task activities*, this category was dropped. Instead, several learners directly addressed the AI in a derogatory, confrontational, or frustrated manner, leading to the inductive identification of a distinct *AI rejection* category. To maintain conceptual coherence, the earlier *greeting/politeness* category was reframed as *AI affiliation*, capturing the affiliative dimension of anthropomorphic interaction. Additionally, the former *translation* category was integrated into the main strategy *metacognitively-inquiring*, because translation requests typically represent clarification and knowledge retrieval within the writing context in which AI acts as a reflective partner supporting comprehension and self-regulated writing.

## 5 Results and interpretation

### 5.1 AI writing strategies across languages

To gain deeper insights into how learners interacted with AI during the writing process, all prompts were analyzed using qualitative content analysis (Kuckartz & Rädiker, 2023) in MAXQDA. Coding was conducted by the author following an iterative procedure of category development and refinement. To enhance analytical rigor and interpretative reliability, category definitions and selected coding decisions were discussed repeatedly within the research group. Given the exploratory and theory-building character of the study and the limited dataset of the analysis, emphasis was placed on conceptual coherence and transparency rather than statistical intercoder agreement. All AI prompts were successfully categorized into the identified AI writing strategies ranging from text planning and linguistic refinement to metacognitive reflection and social interaction with AI. No interactions fell outside the analytical framework. All prompt examples presented in the following sections represent authentic excerpts from student–AI chat logs. To contextualize these qualitative patterns, strategy use was related to quantitative reading measures from *LGVT* (L1/L2, German) and *Cambridge Preliminary* (FL, English) tests. For each strategy, mean *LGVT* T-scores in reading comprehension, speed, and accuracy, as well as average *Cambridge English Scale Score* were considered. Additionally, the cross-linguistic transfer from L1/L2 (German) and FL (English) of each strategy type was quantified using the Jaccard index, representing the overlap of learners who employed the same strategy in both languages according to the formula:

$$J = \frac{|A \cap B|}{|A \cup B|}$$

where  $A$  and  $B$  represent the sets of *individual learners* who used a given strategy in L1/L2 (German) and FL (English) respectively. The intersection  $|A \cap B|$  was determined based on the actual overlap of

learners across both language contexts. The resulting Jaccard values therefore reflect empirically observed individual-level overlap from the actual intersection of learners. For full transparency, Appendix B reports all learners alongside their reading test results, their respective strategy use in both languages, as well as the resulting individual Jaccard values. Higher values indicate that a larger proportion of the same learners applied a given AI strategy in both languages, reflecting a higher degree of cross-linguistic transfer and strategic stability, whereas lower values suggest language- or context-specific patterns. This measure thus captures how consistently individual learners drew on similar AI-assisted writing strategies across linguistic contexts. The index ranges from 0 (*no overlap*) to 1 (*complete overlap*). Intermediate values can be interpreted as follows: Values below .3 indicate low transfer (mostly distinct strategy use), values between .3 and .6 indicate moderate transfer (some overlap in strategy use), and values above .6 indicate high transfer (substantial consistency in strategy use across languages). Tables 3 and 4 present the resulting strategies and their respective transfer values. For reasons of space and readability, the tables report aggregated frequencies of prompts and users per strategy. Nonetheless, the Jaccard indices reported in the tables are based on the actual learner-level data rather than on marginal frequencies alone. The underlying learner-level distributions and individual strategy constellations across both language contexts are documented in full detail in Appendix B.

**Table 3***Five Main AI Writing Strategies*

Characteristics	Student prompt examples (FL)	Prompts and users		
		L1/L2 (German)	FL (English)	Transfer (Jaccard index)
<b>Strategy 1: focally-structuring</b>     - AI primarily used for text planning and structuring    - AI serves as an organizing tool and a thinking aid	“Help me plan a summary of the short story ‘The Text’ by Brendan Dunne (2012)” (o5b-gpc)     “Give me a writing plan to write a summary of a short story” (lqm-cml)	11 prompts by 8 users  <i>LGVT:</i>     - RC average <math>\varnothing = 48</math>    - RS average <math>\varnothing = 44</math>    - RA average <math>\varnothing = 60</math>	16 prompts by 13 users  <i>Cambridge Scale Score</i> average $\varnothing = 149$ (C–B1)	.62

<b>Strategy 2: holistically- optimizing</b>	• AI used for linguistic fine-tuning: choice of words, style, grammar, sentence structure, and useful phrases     • Prompts asking for synonyms and re-phrasing     • Iterative adjustment of text length and complexity	• “Can you give me useful phrases for the summary” (kzo-wng)     • “another word for but” (sld-dpz)	9 prompts by 7 users  <i>LGVT:</i>      • RC average <math>\emptyset = 45</math>     • RS average <math>\emptyset = 41</math>     • RA average <math>\emptyset = 55</math>	17 prompts by 9 users  <i>Cambridge Scale Score</i> average $\emptyset = 150$ (C–B1)	.60
<b>Strategy 3: functionally- pragmatizing</b>	• AI used to increase efficiency: (supposedly) saves time and reduces effort     • Focus on “quickly getting results”	• “write me a summary for the story of brendan dunne” (sld-dpz)     • “write a summary about ‘the text’ by brendan dunne” (c6a-vgp)	14 prompts by 7 users  <i>LGVT:</i>      • RC average <math>\emptyset = 48</math>     • RS average <math>\emptyset = 46</math>     • RA average <math>\emptyset = 54</math>	10 prompts by 4 users  <i>Cambridge Scale Score</i> average $\emptyset = 156$ (B–B1)	.57
<b>Strategy 4: differentially- limiting</b>	• AI use ends after one or two prompts; the text is largely written independently	• “is this summary good?” (auy-bq3)	7 users  <i>LGVT:</i>      • RC average <math>\emptyset = 41</math>     • RS average <math>\emptyset = 42</math>     • RA average <math>\emptyset = 47</math>	6 users  <i>Cambridge Scale Score</i> average $\emptyset = 138$ (fail–A2)	.44
<b>Strategy 5: metacogni- tively-inquir- ing</b>	• Clarification and knowledge retrieval within the writing context (including translations)     • AI as a reflection partner: assistance with evaluation, obtaining formative feedback	• “What does ‘to jeer’ mean?” (b5b-il8)     • “I have now created a summary of the story, would you be able to read it and point out any mistakes?” (dat-k6n)     • “this is my summary can you please make it better” (kzo-wng)     • “how is this?” (mm4-ysy)	44 prompts by 15 users  <i>LGVT:</i>      • RC average <math>\emptyset = 45</math>     • RS average <math>\emptyset = 42</math>     • RA average <math>\emptyset = 53</math>	56 prompts by 15 users  <i>Cambridge Scale Score</i> average $\emptyset = 146$ (C–B1)	.58

**Table 4***Two Supplementary Categories of AI Interaction*

	Characteristics	Student prompt examples (FL)	Prompts and users		
			L1/L2 (German)	FL (English)	Transfer (Jaccard index)
<b>AI affiliation</b>	Phatic communication in which the AI is greeted, thanked, or addressed in a socially affiliative manner	“hello” (rv5-gjh)     “Thank you very much. I am grateful for the advice, it is of great use” (dat-k6n)	8 prompts by 6 users  <i>LGVT:</i>     - RC average <math>\emptyset = 46</math>    - RS average <math>\emptyset = 43</math>    - RA average <math>\emptyset = 54</math>	4 prompts by 2 users  <i>Cambridge Scale Score</i> average $\emptyset = 164$ (A–B2)	.14
<b>AI rejection</b>	Derogatory, confrontational, or frustrated remarks directed at the AI.	“NO? what are you on vro. never trust ai” (fsl-dda)     “youre useless” (iyv-svk)	8 prompts by 1 user  <i>LGVT:</i>     - RC average <math>\emptyset = 60</math>    - RS average <math>\emptyset = 60</math>    - RA average <math>\emptyset = 54</math>	4 prompts by 2 users  <i>Cambridge Scale Score</i> average $\emptyset = 157$ (B–B1)	.50

Across both languages, the five main strategies reflect functionally distinct forms of AI engagement throughout the writing process, as shown in Table 3. The most prevalent strategy type was *metacognitively-inquiring* (Strategy 5), employed by the majority of learners in both L1/L2 and FL. This indicates that learners primarily perceived AI as a reflective partner that supports comprehension, clarification, and formative feedback. This is reflected in prompts such as “What does ‘to jeer’ mean” (user b5b-il8), “I have now created a summary of the story, would you be able to read it and point out any mistakes?” (user dat-k6n), or “this is my summary can you please make it better” (user kzo-wng). This shows that through *metacognitively-inquiring* behavior, students became aware of both linguistic and conceptual knowledge gaps. By deliberately asking AI for clarification or elaboration, they engaged in *noticing processes*, using AI as a trigger for reflection and knowledge expansion. This aligns with Schmidt’s (1990) *noticing hypothesis* and suggests that AI can act as a catalyst for awareness and self-regulated learning rather than merely as a feedback provider. The moderate Jaccard index (.58) suggests that such reflective strategies are relatively stable across languages, pointing to their generalizability as higher-order cognitive tools that transcend linguistic boundaries. The *holistically-optimizing* strategies (Strategy 2) showed similarly moderate transfer values (.60) as illustrated by prompts such as “Can you give me usefull [sic] phrases for the summary” (user kzo-wng) or “another woord [sic] for but” (user sld-dpz) that target lexical and stylistic refinement. The transfer of *focally-structuring* strategies

(Strategy 1) can be regarded as high (.62), as reflected in prompts such as “Help me plan a summary of the short story ‘The Text’ by Brendan Dunne (2012)” (user o5b-gpc) or “Give me a writing plan to write a summary of a short story” (user lqm-cml). Together, these three strategies represent procedural, linguistic, and cognitive scaffolding functions within AI-assisted writing. Their balanced occurrence in both L1/L2 and FL contexts suggests that learners are able to flexibly integrate AI across different stages of the writing process from planning and linguistic refinement to metacognitive reflection. In contrast, *functionally-pragmatizing* strategies (Strategy 3) reflect an instrumental orientation toward efficiency and task completion. Learners employing this approach delegated writing tasks to the AI, aiming to minimize effort or time investment. This orientation becomes visible in prompts such as “write me a summary for the story of brendan dunne” (user sld-dpz) or “write a summary about ‘the text’ by brendan dunne” (user c6a-vgp). The moderate transfer value (.57) indicates that such instrumental use patterns were relatively stable across languages, suggesting that the tendency to exploit AI for efficiency may represent a general behavioral disposition rather than a language-specific adaptation. However, although both metacognitive and pragmatic strategies transfer across languages, they differ fundamentally in depth of engagement: The former involves reflective regulation of the writing process with queries such as “how is this?” (user mm4-ysy), whereas the latter primarily targets output optimization through direct task delegation such as “write me a summary . . .” (user sld-dpz). *Differentially-limiting* strategies (Strategy 4) were associated with the lowest transfer value (.44) and the weakest reading proficiency (CEFR A2 for English, lowest LGVT means for German). Learners employing these strategies used AI only sparingly, producing only one or two prompts that were functionally assignable to other strategy types. This pattern is reflected in minimal interaction sequences—for example, in prompts such as “is this summary good?” (user auy-bq3), after writing the summary without any AI support. Learners relied primarily on their own writing, indicating cautious engagement and limited trust in AI output. The combination of limited AI use and low transfer implies that weaker learners may not yet have developed a flexible or adaptive repertoire for integrating (AI) writing strategies into their writing process.

The two supplementary categories, *AI affiliation* and *AI rejection*, shown in Table 4, highlight anthropomorphization (e.g., Chaturvedi et al., 2025; Kim & Im, 2023; Li et al., 2025) as a cross-cutting phenomenon in AI use. In both cases, AI is treated as a social interlocutor mirroring human-like communication patterns rather than as a neutral tool. Expressions of greeting, gratitude, or playfulness address constitute affiliative bonding and phatic communication as reflected in prompts such as “hello” (user rv5-gjh) or “Thank you very much. I am grateful for the advice, it is of great use” (user dat-k6n), whereas confrontational or rejecting remarks similarly presuppose a human-like counterpart capable of intention or failure, for example “youre [sic] useless” (user iyv-svk) or “NO? what are you on vro [sic]. never trust ai” (user fsl-dda). Although *AI affiliation* was rare and showed little cross-linguistic transfer (.14), it reveals moments of social alignment and gratitude toward AI. Conversely, *AI rejection* reflects explicit distancing from the system, typically arising from frustration or distrust and showing moderate transfer across languages (.50). Both categories illustrate the interpersonal framing of AI interaction: Affiliative and antagonistic responses alike presuppose a degree of social agency and interpersonal relevance for the AI. Such interactions may reduce affective barriers and foster motivation, but also risk diminishing critical distance and reinforcing naïve assumptions about the AI’s epistemic competence. Consequently, learners may perceive AI not merely as a functional tool but as a seemingly authentic interlocutor—a finding of particular relevance for pedagogical approaches to AI-assisted writing.

Overall, the pattern of Jaccard indices across strategies reveals a moderate degree of cross-linguistic transfer, indicating that learners generally employed comparable AI-supported writing strategies in both L1/L2 and FL. However, the variability across strategies reflects differences in functionality and cognitive depth. Strategies involving planning, linguistic optimization, or metacognitive reflection (Strategies 1, 2, and 5) demonstrate consistent use across languages, suggesting that such cognitively integrated forms of AI engagement operate as transferable writing strategies. *Functionally-pragmatizing* strategies (Strategy 3), although comparable in transfer rate, reflect a more instrumental engagement with AI, indicating a general behavioral tendency toward efficiency rather than a language-specific adaptation. By contrast, *differentially-limiting* strategies (Strategy 4) remain the least transferable, pointing to cautious or inconsistent AI use among less proficient learners. This suggests that the transferability of (AI) writing strategies depends less on language or linguistic factors and more on the cognitive complexity and degree of learner autonomy. Strategies grounded in reflective regulation or deliberate process management generalize more easily across languages, whereas those shaped by avoidance or minimal involvement remain more context-bound and situational.

To complement the strategy-level analysis, an additional measure was applied to capture the degree of strategic transfer between L1/L2 and FL at the group level. A *generalized Jaccard similarity coefficient* based on strategy frequency vectors was calculated to quantify the overall structural similarity of strategy use across languages. This aggregate measure compares the distribution of learners across strategies in both languages and reflects the degree to which similar patterns of AI-supported writing behavior emerged at the group level.

The generalized Jaccard index was calculated using the number of learners who employed each strategy in L1/L2 (German) and FL (English). The index is defined as:

$$J_g = \frac{\sum_i \min(L1_i, FL_i)}{\sum_i \max(L1_i, FL_i)}$$

where  $L1_i$  and  $FL_i$  denote the number of learners who used strategy  $i$  in L1/L2 (German) and FL (English) respectively. Importantly, this measure operates purely on aggregated frequency counts and does not take into account whether the same learners employed a given strategy in both language contexts. Rather than capturing individual-level overlap, it compares the overall distribution of strategy use across languages at the group level. Because learners could employ multiple strategies, the analysis reflects multilabel strategy use rather than mutually exclusive group membership. The distribution of learners across strategies in both language contexts is summarized in Table 5.



**Table 5**

*Number of Learners Employing Each AI Writing Strategy in L1/L2 (German) and FL (English)*

Strategy	L1/L2 users	FL users	Min (L1; FL)	Max (L1; FL)
Strategy 1: focally-structuring	8	13	8	13
Strategy 2: holistically-optimizing	7	9	7	9
Strategy 3: functionally-pragmatizing	7	4	4	7
Strategy 4: differentially-limiting	7	6	6	7
Strategy 5: metacognitively-inquiring	15	15	15	15
AI affiliation	6	2	2	6
AI rejection	1	2	1	2
$\Sigma$			<b>43</b>	<b>59</b>

Based on the aggregated distributions presented in Table 5, the calculation yields:

$$J_g = \frac{\sum_i \min(L1_i, FL_i)}{\sum_i \max(L1_i, FL_i)} = \frac{43}{59} = .73$$

This value indicates a high degree of structural similarity in the overall distribution of AI writing strategies across languages. It suggests that, at the group level, comparable numbers of AI writing strategy occurrences are observed in both contexts, with only moderate differences in relative strategy use.

## 5.2 Cross-linguistic transfer of AI writing strategies across CEFR levels

To further contextualize these quantitative findings on strategic transfer, the distribution of AI writing strategies was examined in relation to learners' CEFR proficiency levels in English (FL). This step allows for a more nuanced understanding of how linguistic competence is associated with the presence of particular strategies across languages. The results of this analysis are later complemented by grouping the individual Jaccard indices by proficiency level, thereby linking the qualitative typology of AI writing strategies with quantitative measures of linguistic proficiency to account for cross-linguistic transfer. When examined across CEFR proficiency levels, distinct patterns emerged in the distribution of AI writing strategies. The values reported in Table 6 represent the number of strategy uses across learners in

the respective language context. Given the small subgroup sizes, however, these analyses are strictly descriptive and exploratory and do not allow inferential generalization. Percentages should be interpreted with caution, because small absolute differences correspond to relatively large percentage shifts due to the limited subgroup sizes.

**Table 6**

*Distribution of AI Writing Strategies by English CEFR Proficiency Level and Language Context (L1/L2, FL)*

Strategy	B2		B1		A2	
	L1/L2	FL	L1/L2	FL	L1/L2	FL
Strategy 1: focally-structuring	2 (18.18%)	2 (20%)	5 (17.24%)	9 (29.03%)	1 (9.09%)	2 (20%)
Strategy 2: holistically-optimizing	1 (9.09%)	2 (20%)	4 (13.79%)	6 (19.35%)	2 (18.18%)	1 (10%)
Strategy 3: functionally-pragmatizing	1 (9.09%)	1 (10%)	6 (20.69%)	3 (9.68%)	0	0
Strategy 4: differentially-limiting	1 (9.09%)	0	3 (10.34%)	3 (9.68%)	3 (27.27%)	3 (30%)
Strategy 5: metacognitively-inquiring	3 (27.27%)	3 (30%)	8 (27.59%)	8 (25.81%)	4 (36.36%)	4 (40%)
AI affiliation	2 (18.18%)	1 (10%)	3 (10.34%)	1 (3.23%)	1 (9.09%)	0
AI rejection	1 (9.09%)	1 (10%)	0	1 (3.23%)	0	0
Total (strategy uses across learners)	11 (100%)	10 (100%)	29 (100%)	31 (100%)	11 (100%)	10 (100%)

At the B2 level, *metacognitively-inquiring* strategies accounted for 3 out of 11 strategy uses in L1/L2 and 3 out of 10 in FL, making this the most frequently observed strategy in both contexts and suggesting a reflective engagement with AI. *Focally-structuring* strategies accounted for 2 occurrences in both L1/L2 and FL; indicating a balanced integration of planning-related support across languages, whereas *holistically-optimizing* strategies appeared slightly more often in the FL (2 occurrences) than in L1/L2 (1 occurrence). This pattern suggests that strategy use at higher proficiency spans a range of functions

with a somewhat stronger reliance on linguistic refinement in the foreign language. Notably, both *AI affiliation* (2 occurrences in L1/L2 and 1 in FL) and *AI rejection* (1 occurrence in both contexts) were observed most frequently at this proficiency level. Whereas affiliative interactions point to moments of social engagement with the AI, the presence of rejection may indicate a more critical stance toward AI output, including the ability to distinguish between helpful and less helpful responses.

At the B1 level, *metacognitively-inquiring* strategies again accounted for the highest number of strategy uses (8 out of 29 in L1/L2 and 8 out of 31 in FL). In the FL, however, *focally-structuring* strategies appeared more frequently (9 occurrences) than in L1/L2 (5 occurrences), suggesting a stronger reliance on AI for structuring support in the foreign language. *Holistically-optimizing* strategies were also observed in multiple instances across both contexts (4 in L1/L2 and 6 in FL), indicating that linguistic refinement plays a relevant role at this proficiency level, particularly in the FL. *Functionally-pragmatizing* strategies, by contrast, were more frequent in L1/L2 (6 occurrences) than in FL (3 occurrences), which may suggest that more instrumental uses of AI are less consistently transferred across languages. Overall, these patterns indicate that at intermediate proficiency levels, learners tend to rely more strongly on AI for text organization and linguistic refinement in the foreign language.

At the A2 level, *metacognitively-inquiring* strategies again showed the highest number of strategy uses (4 out of 11 in L1/L2 and 4 out of 10 in FL). At the same time, a comparatively high number of learners employed *differentially-limiting* strategies (3 occurrences in each context), indicating a tendency toward cautious or minimal engagement with AI support that was more pronounced than at higher proficiency levels. In contrast, strategies associated with planning and linguistic elaboration were less frequent: *Focally-structuring* strategies accounted for only 1 occurrence in L1/L2 and 2 in FL, whereas *holistically-optimizing* strategies appeared in 2 occurrences in L1/L2 and 1 in FL. This pattern suggests a comparatively limited use of planning and refinement-oriented strategies, which may indicate a less developed strategic repertoire for managing the writing process (with AI support). This contrasts with the broader and more flexible distribution of strategies observed among more proficient learners and aligns with previous findings suggesting that lower-proficiency writers tend to employ fewer and less elaborated writing strategies (Harris et al., 2010; Lei, 2016; Manchón & Roca De Larios, 2007; Raoofi et al., 2017; Roca De Larios et al., 2001).

Overall, these observations tentatively suggest an association between linguistic proficiency and the breadth as well as the qualitative nature of AI-supported writing strategies. In addition, the language context appears to influence AI use. In the FL, a slightly higher number of learners rely on structural and linguistic scaffolding, whereas in L1/L2 contexts, reflective and evaluative uses of AI are more widely represented.

The observed distributional patterns across proficiency levels suggest potential differences in how learners engage with AI during writing. However, although these frequency-based results capture which strategies are most characteristic for each group, they do not yet account for the stability of individual strategy use across linguistic contexts. To complement the group-level findings and to obtain a more fine-grained perspective, individual Jaccard indices were calculated for each learner to assess intra-individual transfer of AI writing strategies between L1/L2 (German) and FL (English). The index is defined as:

$$J = \frac{|A \cap B|}{|A \cup B|}$$

where  $A$  and  $B$  represent the sets of strategies used by a single learner in each language. This measure quantifies how consistently individual learners employed similar strategic patterns across linguistic contexts, thereby reflecting the stability or variability of their AI-assisted writing behavior. A complete overview of all individual Jaccard indices for each user is provided in Appendix B. Across all learners, the mean Jaccard index was .57 (see Table 7), indicating a moderate overall degree of strategic transfer between L1/L2 and FL. To further refine this analysis, the individual Jaccard indices were grouped by English CEFR proficiency levels (see Table 7). Linking these transfer measures with proficiency levels allows for a more differentiated perspective on how linguistic competence and language background relate to the stability, variability, and transferability of AI writing strategy use.

**Table 7**

*Descriptive Statistics for the Transfer of AI Writing Strategies (Jaccard Indices)*

	Overall Jaccard indices	B2 Jaccard indices	B1 Jaccard indices	A2 Jaccard indices
<i>M</i>	.57	.67	.54	.55
<i>Mdn</i>	.50	.59	.50	.50
<i>SD</i>	.30	.24	.27	.45
Range	1.00	.50	.75	1.00
Min	.00	.50	.25	.00
Max	1.00	1.00	1.00	1.00

Learners at the B2 level exhibited the highest mean transfer ( $M = .67$ ), indicating a greater degree of consistency in their strategic engagement with AI across languages. In contrast, learners at the B1 ( $M = .54$ ) and A2 ( $M = .55$ ) levels demonstrated moderate transfer, suggesting more situational and language-specific variability in their strategy use.

Taken together, the distributional patterns of strategy occurrences, the group- and individual-level transfer measures, and the comparison of individual Jaccard indices across CEFR proficiency levels suggest that cross-linguistic consistency in AI-assisted writing behavior appears more pronounced when strategies involve planning, linguistic refinement, or metacognitive reflection (i.e., focally-structuring, holistically-optimizing, and metacognitively-inquiring strategies) rather than efficiency-driven or cautious modes of engagement (i.e., functionally-pragmatizing and differentially-limiting strategies). Within this sample, learners with higher reading competence tended to transfer such strategies more consistently and flexibly across languages, which may reflect that proficiency enables more deliberate and self-regulated (Harris et al., 2010; Harris et al., 2011) forms of AI use in writing. In contrast, lower-proficiency learners showed a comparatively greater degree of variability and situational adaptation, relying on narrower, more context-dependent patterns of interaction. Overall, the findings point to a possible relation between linguistic proficiency and strategic stability, suggesting that linguistic proficiency may be linked not only to the qualitative nature of the strategies employed but also to the stability and flexibility of their cross-linguistic application. Nevertheless, given the exploratory nature and limited sample size of the study, these patterns should be interpreted with caution.

### 5.3 Exemplary learner cases of AI writing strategy use

To complement the quantitative and strategy-based analyses, two exemplary student cases are presented to illustrate how AI-supported writing strategies manifest in actual written products across languages. The selected cases are intended to illustrate contrasting patterns of cross-linguistic transfer and proficiency levels and should be interpreted as exploratory, nonrepresentative examples.

An illustrative example of a high-proficiency learner is provided by Student mm4-ysy, whose texts and interaction patterns reflect a consistent and strategically stable use of AI across languages. The learner achieved a Cambridge English Scale Score of 160 (Grade A, CEFR B2), along with LGVT *T* scores in the lower average range for reading comprehension (RC = 46) and reading speed (RS = 41) and in the upper average range for reading accuracy (RA = 56). Notably, the learner shows complete cross-linguistic transfer of AI writing strategies (Jaccard index = 1.0), indicating identical strategy use in L1/L2 and FL contexts.

In the German (L1/L2) task, the learner employed *holistically-optimizing* and *metacognitively-inquiring* strategies throughout the writing process. The interaction with AI is characterized by iterative refinement, including requests for stylistic improvement such as “*nenne mir textbausteine um eine inhaltsangabe für eine kurzgeschichte zu schreiben*” [sic] [Tell me useful phrases to write a summary of a short story] and evaluative feedback on draft versions “*Wie ist das bis jetzt?*” [How is this so far?]. The resulting German summary demonstrates a clear structure and coherent progression of ideas, with a concise presentation of the dystopian scenario and its central conflict (see Figure 1).

**Figure 1**

#### *German Summary by Student mm4-ysy*

##### **Kurzgeschichte von mm4-ysy**

Die Kurzgeschichte „USE 2030“ von Bob Blume, veröffentlicht im Jahr 2025, thematisiert die Konsequenzen politischer Korruption in einer dystopischen Zukunft, in der Elon Musk die Kontrolle über die USA übernommen hat.

Zu Beginn der Geschichte wird der Ich-Erzähler vorgestellt, der aufgrund eines sogenannten „Visual Echo“ gezwungen ist, mit Kuli auf Papier zu schreiben, da er nicht freisprechen kann. Er schildert die Ereignisse der letzten Jahre: Der damalige Präsident wurde vermutlich mithilfe von Ketamin vergiftet, was Elon Musk ermöglichte, die Macht zu übernehmen. Die USA werden nun als USE – Vereinigte Staaten von Elon Musk bezeichnet.

In dieser neuen Ordnung gibt es keine Freiheit mehr. Die Menschen werden überwacht, unter anderem durch Jeeps mit schwarzen Flaggen, die ein weißes „X“ tragen, und dazu ermahnt, in ihren Unterkünften zu bleiben. Alles, was als „unamerikanisch“ gilt, ist streng verboten, um angeblich die Unabhängigkeit des Landes zu sichern. Der Erzähler beschreibt, wie die Abschiebung stark zugenommen hat und dass „regierungsfeindlichen“ Personen ihre Staatsbürgerschaft verlieren.

Er selbst besitzt ein paar europäische Konservendosen, die er wie Schätze hütet, und plant, seinen Brief zu verstecken, um späteren Generationen zu zeigen, wie es wirklich war. Die Spannung steigt, als der Erzähler plötzlich Schritte im Treppenhaus hört.

Die Geschichte endet damit, dass der Erzähler kurz davor ist, wegen dem Visual Echo, dass ihn „erwischt“ hat, entdeckt zu werden. Er erinnert die Leser daran, dass es einst eine Zeit gab, in der die Menschheit frei war.

A similar pattern can be observed in the English (FL) task. Here, the learner again relied on *holistically-optimizing* and *metacognitively-inquiring* strategies, using AI to refine linguistic expression and to evaluate the developing text. Examples include “tell me useful phrases to write a summary of a short story” or “how is this?” The final English summary is likewise well-structured, coherent, and linguistically

precise, additionally including an interpretative element that connects the narrative to the fable *The Boy Who Cried Wolf* (see Figure 2).

**Figure 2**

*English Summary by Student mm4-ysy*

**Summary mm4-ysy**

The short story "The Text" by Brendan Dunne, published in 2012, follows Amy, a girl with asthma who develops feelings for a boy named Kamal. However, when Kamal takes one joke too far, no one believes him in a moment of real danger, which tragically leads to his death.

The short story follows Amy, a girl with asthma who isn't good with technology, and on that day she somehow forgot her inhaler, her best friend Tara, Kamal and his best friend Grant whom neither of the girls really like

Kamal invites Amy and Tara to visit an abandoned house behind the factory to see his art project. However, Kamal and Grant play a prank by pretending Grant is a zombie, which shocks Amy and triggers her asthma. Although Amy brushes it off as paint fumes, the girls are upset and lose trust in the boys.

The following evening at the school disco, Kamal is nowhere to be found. Grant searches for him, but the girls dismiss him. When Tara receives a text from Kamal saying he is trapped in the abandoned house and needs help urgently, she assumes it is another prank and tells others to ignore it. The next morning, the news reports that Kamal had been seriously injured and died after being trapped under rubble.

Overall, the short story emphasizes the dangers of losing trust, much like in the classic fable "The Boy Who Cried Wolf" once trust is broken, even a true cry for help may be ignored, with devastating consequences.

Across both language contexts, the learner demonstrates a highly consistent approach to AI-supported writing. The learner applies the same combination of refinement- and feedback-oriented strategies in both German and English while maintaining primary authorship of the texts. AI is used selectively as a supportive scaffold rather than as a substitute for text production. This stable pattern is reflected in the maximal Jaccard value (1.0) and suggests that the learner draws on a transferable repertoire of AI writing strategies. The case thus exemplifies how higher-proficiency learners may engage with AI as a tool for iterative optimization and reflective evaluation across languages, supporting the broader finding that strategic stability increases with linguistic proficiency.

In contrast, Student auy-bq3 represents a lower-proficiency learner whose AI use appears more selective and less strategically integrated across languages. The learner achieved a Cambridge English Scale Score of 134 (Fail, CEFR A2), alongside LGVT *T* scores in the lower average range for reading comprehension (RC = 42), below average for reading speed (RS = 39), and in the upper average range for reading accuracy (RA = 50). The learner's low cross-linguistic transfer of AI writing strategies (Jaccard index = .25) indicates only limited overlap in strategy use between L1/L2 and FL contexts.

In the German (L1/L2) task, the learner employed *holistically-optimizing* (*anderes wort für repressiven* [another word for repressive]) and *metacognitively-inquiring* (*ist das gut für ein schlussteil* [sic] [Is that a good ending?]) strategies alongside *AI affiliation* (*DANKE* [thank you]), engaging in multiple interactions with the AI to refine wording, structure the text, and obtain evaluative feedback. The resulting German summary is coherent and well-structured, clearly presenting the central elements of the story. The learner effectively uses AI as a supportive scaffold for refining wording and structuring the text, resulting in a linguistically clear and organized summary (see Figure 3).

**Figure 3***German Summary by Student auy-bq3***Kurzgeschichte von Bob Blume Inhaltsangabe**

Der Ich-Erzähler sitzt in seiner Wohnung in Ohio am 23. Februar 2030 und beginnt, all seine Beobachtungen und Gedanken mit Papier und Stift zu schreiben – möglicherweise sein letzter Versuch, Zeugnis abzulegen.

Im Jahr 2030 lebt der Ich-Erzähler in einem totalitären Staat, den „United States of Elon Musk“. Die Demokratie wurde abgeschafft, und die Regierung kontrolliert das Volk durch Überwachung, Ausgangssperren und Unterdrückung. Öffentliche Orte verfallen, viele Produkte fehlen, und die Bevölkerung lebt in Angst. Der Erzähler berichtet von der Abschiebung seiner Nachbarin und sieht sich selbst bedroht. Trotz dieser Umstände versucht er, seine Gedanken schriftlich festzuhalten, um die Geschehnisse für die Zukunft zu dokumentieren.

Am Ende der Geschichte hört der Erzähler Schritte im Treppenhaus und fürchtet, entdeckt zu werden. Obwohl er nicht sicher ist, ob jemand zu ihm kommt, bleibt er entschlossen, mit seinem Text ein Zeichen zu setzen. Er möchte die Erinnerung an eine Zeit bewahren, in der Freiheit noch ein selbstverständliches Gut war.

A different pattern emerges in the English (FL) task. Here, AI use is markedly reduced and limited to only one prompt, focused on general feedback (“is this summary good?”), without extended revision or optimization cycles. This corresponds to a *differentially-limiting* strategy in which AI interaction remains minimal and does not substantially shape the writing process. The resulting English summary, while covering the main plot, is comparatively verbose, less condensed, and shows structural inconsistencies (see Figure 4).

**Figure 4***English Summary by Student auy-bq3***The Text**

The Message, titled *The Text* by Brendan Dunne, tells the story of two boys who enjoy playing jokes but take them too far. In the end, no one believes them anymore. There is a girl named Amy who really likes a boy named Kamal. Kamal invites Amy and her friend Tara to come with him so he can show them some graffiti he made for school. However, the girls don't know that Kamal only wants to prank them with his friend Grant. When they arrive at an abandoned house, Kamal tells them that Grant will come later because he has to buy some things. After a few minutes, Amy starts hearing noises like groaning and moaning, similar to a zombie. Kamal insists it's nothing and that he doesn't hear anything. Suddenly, a bloodstained zombie appears. Amy thinks it's a real zombie, which causes her breathing to worsen due to her asthma. Meanwhile, Tara knows it is actually Grant in disguise and notices Amy's breathing difficulties. The next evening was Halloween, but Tara was still worried about Amy because of what happened the day before. However, Amy insisted she was okay and that it wasn't serious. She stopped caring about Kamal and just enjoyed dancing with Tara. Suddenly, Grant approached them looking worried. He said they should text Kamal because his phone had been stolen. The girls didn't take it seriously and thought it was just another one of Kamal's jokes. But then, Tara's phone beeped with a message from Kamal saying he was trapped and asking her to call an ambulance. Despite this, Tara didn't believe him and told everyone to ignore the message. The next day, Amy's parents were listening to the news when they heard a report about a sixteen-year-old boy who had died. The news asked if anyone had any information about the incident. That's when Amy's dad remembered that Amy had gone to that abandoned house with her friends on Monday. The final news was that the boy who died had lost a lot of blood and was buried under a pile of rubble. He was identified as Kamal Naseer. This story teaches an important lesson about honesty and trust. Because the boys took their jokes too far, no one believed Kamal when he really needed help, which led to a tragic outcome.



This suggests that the learner relies predominantly on their own writing with limited integration of AI as a scaffold for revision or improvement.

Across both language contexts, the learner demonstrates a less stable and only weakly transferable pattern of AI-supported writing behavior. Whereas AI is used more actively in the German task, its role in the English task remains limited and largely peripheral. The low Jaccard value reflects this discrepancy and indicates that AI writing strategies are not applied consistently across languages. A tentative explanation for this pattern may lie in differences in linguistic competence: The learner appears more confident and engaged in the German (L1/L2) context in which the proficiency is in the average range and AI is integrated more frequently and productively. In contrast, lower proficiency in the English (FL) context (Cambridge: fail; CEFR A2) may lead to more cautious and reduced AI use. This case thus exemplifies how lower-proficiency learners may engage with AI more cautiously and inconsistently, with limited integration into the writing process and reduced cross-linguistic transfer of AI writing strategies.

Taken together, the two cases illustrate contrasting patterns of AI-supported writing across proficiency levels and language contexts. Whereas the high-proficiency learner (mm4-ysy) demonstrates consistent and transferable use of AI as a scaffold for refinement and reflection in both languages, the lower-proficiency learner (aui-bq3) shows a more fragmented and context-dependent pattern with uneven integration of AI across L1/L2 and FL. This contrast tentatively suggests that cross-linguistic transfer of AI writing strategies may be associated not only with strategic awareness but also with linguistic competence; and this, in turn, may enable learners to engage with AI more flexibly, consistently, and productively across languages. However, given the illustrative nature of these two cases, such interpretations remain exploratory and require further validation in larger samples.

## 6 Discussion and outlook

The findings of this study provide new insights into how learners engage with AI in the writing process and how these patterns vary across linguistic and proficiency levels. As an exploratory study based on a small-scale classroom sample, the aim was to develop an initial, data-driven understanding of AI-supported writing behavior. The results should therefore be interpreted as preliminary indications rather than generalizable claims, and they warrant further validation in larger and more diverse cohorts.

The qualitative analysis revealed that learners employed a diverse repertoire of AI writing strategies ranging from *focally-structuring* and *holistically-optimizing* approaches to *metacognitively-inquiring* uses in which AI served as a reflective and dialogic partner. Importantly, these strategies were not fixed user types, but rather flexible, phase-dependent writing behaviors, thereby aligning with process-oriented models of writing (e.g., Hayes, 2012). Notably, this process-related character was not imposed by the analytical framework but emerged inductively from the interaction data. Learners moved dynamically between planning, drafting, and revising, using AI as both a cognitive and linguistic scaffold.

The quantitative findings complement this perspective by indicating that such strategic patterns are not confined to a single language context. A high generalized Jaccard similarity coefficient ( $J_g = .73$ ) points to a similarity in the distribution of learners across strategy types in L1/L2 and FL, suggesting that comparable strategic repertoires emerge at the group level. However, this aggregate measure does not capture individual-level transfer, so this was examined separately using individual Jaccard indices. On the individual level, especially the cognitively demanding strategies for planning, drafting,

and revising (i.e., *focally-structuring*, *holistically-optimizing*, and *metacognitively-inquiring* strategies) appear to be more stable and transferable across languages than *pragmatic* or *limiting* strategies. This is reflected in their comparatively higher Jaccard values (.62, .60, and .58 respectively vs. .57 and .44 respectively), suggesting patterns consistent with partial cross-linguistic transfer in AI-supported writing within this exploratory sample. At the same time, meaningful variation between learners was revealed: The language context appears to shape AI engagement, with learners relying more on structural and linguistic scaffolding in FL writing, whereas reflective and evaluative uses prevail in L1/L2 tasks. These differences also align with proficiency levels, as individual Jaccard indices reveal: Students with higher English proficiency (B2) demonstrate greater strategic stability and transferability ( $M = .67$ ), whereas those at B1 ( $M = .54$ ) or A2 levels ( $M = .55$ ) exhibit more situational and context-dependent use. This observation is consistent with prior research on transfer and language awareness (Cummins, 1980, 2000; García & Wei, 2015; Herdina & Jessner, 2002; Odlin, 1989), and may indicate that higher linguistic proficiency is associated with more strategic stability across contexts. However, the observed associations between proficiency and strategic stability should be interpreted with caution, because causal directionality cannot be established, and alternative explanations such as motivational differences or prior AI experience remain possible.

This pattern is further illustrated by the two exemplary cases presented in Section 5.3. The high-proficiency learner demonstrates consistent and transferable use of refinement- and feedback-oriented strategies across both languages, using AI as a scaffold for iterative optimization while maintaining primary authorship. In contrast, the lower-proficiency learner shows a more fragmented and context-dependent pattern, with more frequent and productive AI integration in the L1/L2 context but minimal and peripheral use in the FL. This contrast highlights how differences in linguistic competence may shape not only the extent of AI use but also its functional integration into the writing process, supporting the observed variation in cross-linguistic transfer. Because these cases are illustrative, however, they serve to contextualize rather than generalize the broader quantitative findings.

Interestingly, the identification of *AI affiliation* and *AI rejection* as supplementary categories highlights the anthropomorphic dimension of human–AI interaction. In this dataset, *AI affiliation* involved greeting, thanking, or politely addressing the system, thereby positioning it as a socially responsive interaction partner. Conversely, *AI rejection* did not merely involve refraining from AI use but included direct, confrontational, or evaluative remarks addressed to the system (e.g., insulting or blaming statements). In both cases, the AI was treated as a dialogic counterpart, suggesting that learners attributed a degree of social presence or agency to the AI rather than treating it as a purely instrumental and neutral tool. Although such social framing may lower affective barriers and enhance engagement and motivation, it may also diminish critical distance and foster epistemic naivety toward AI-generated content (see also Dang & Nguyen, 2025; Yan et al., 2025; Zhai et al., 2024). This dual role of AI as both a cognitive and socioaffective agent in the learning process underscores the need for explicit reflection and the fostering of critical AI literacy in educational contexts.

From a pedagogical perspective, these findings indicate that AI interaction may support metalinguistic awareness and self-regulated learning processes, suggesting its pedagogical potential for supporting language acquisition within task-supported and scaffolded learning environments. Furthermore, they point to several implications: First, AI-assisted writing instruction should focus not only on tool competence but also on strategic awareness: helping learners to consciously select, adapt, and evaluate AI use according to the task requirements and writing phase. Second, teachers can scaffold this process through promoting metacognitive awareness and explicit reflection on when and how AI is helpful,

and this may strengthen learners' autonomy and critical engagement. Third, because higher proficiency appears to be associated with greater strategic transfer within this sample, targeted support for lower-proficiency learners may be necessary—for example, through modeling effective prompts, explicit instruction on evaluating AI outputs, and contrastive activities across L1/L2 and FL contexts.

Despite these contributions, the study has several limitations: The small sample size and classroom-based context limit the generalizability of the findings. Future research should therefore aim to validate the identified strategy types and transfer patterns in larger and more diverse samples in order to test their robustness and generalizability across different learner populations. Such studies could examine whether the observed relationships between strategy use, cross-linguistic transfer, and proficiency levels hold under varying educational conditions. Moreover, the focus on a single writing genre (summary writing) and a specific AI tool may have influenced the observed strategy patterns. Future research should therefore also explore how AI writing strategies can vary across genres, modalities (e.g., multimodal or argumentative writing), and AI systems with different affordances. Longitudinal designs could further illuminate how learners' AI literacy (Ng et al., 2021) and strategy repertoires evolve over time and whether cross-linguistic transfer increases with continued AI use.

In sum, the integration of qualitative and quantitative analyses suggests that AI-supported writing may function as a site of both linguistic transfer and metacognitive development. The findings underscore the importance of positioning AI not as a substitute for writing competence but as a catalyst for reflective engagement with language. In this sense, AI can serve as a “mirror” for learners' linguistic and metacognitive development, provided that its use is embedded in pedagogical frameworks that promote autonomy, critical awareness, and linguistic transfer. The challenge for future pedagogy therefore lies not merely in teaching learners how to use AI, but in fostering the reflective capacity to decide when, why, and to what extent AI should be used in the writing process.

## References

- Alghasab, M. B. (2025). English as a foreign language (EFL) secondary school students' use of artificial intelligence (AI) tools for developing writing skills: Unveiling practices and perceptions. *Cogent Education*, 12(1). <https://doi.org/10.1080/2331186X.2025.2505304>
- Allen, L. K., & Kendeou, P. (2023). ED-AI Lit: An interdisciplinary framework for AI literacy in education. *Policy Insights from the Behavioral and Brain Sciences*, 11(1), 3–10. <https://doi.org/10.1177/23727322231220339>
- Bachmann, T., & Becker-Mrotzek, M. (2010). Schreibaufgaben situieren und profilieren. In T. Pohl & T. Steinhoff (Eds.), *Textformen als Lernformen* (pp. 191–209). Gilles und Francke.
- Blume, B. (2025). *USE 2030*. <https://bobblume.de/2025/02/11/kurzgeschichte-use-2030/>
- Breen, M. P. (1985). Authenticity in the language classroom. *Applied Linguistics*, 6(1), 60–70. <https://doi.org/10.1093/applin/6.1.60>
- Bodora, A. L., Decker, L., Fuhlrott, M., Nolden, A., & Steinhoff, T. (2025). Wie schreiben Schüler:innen in der 8. Klasse mit ChatGPT? Einblicke in das Design-Based-Research-Projekt KI-Schreibarrangements. *Leseräume*, 12(11), 1–4.
- Cambridge English (2018). *B1 Preliminary for Schools. Preliminary English Test: Reading Sample Test*. <https://www.cambridgeenglish.org/exams-and-tests/preliminary-for-schools/preparation/>
- Cambridge English (2025). *B1 Preliminary for Schools results*. <https://www.cambridgeenglish.org/exams-and-tests/preliminary-for-schools/results/>

- Castro, V., Nascimento, A. K. D. O., Zheldibayeva, R., Searsmith, D., Saini, A., & Kalantzis, M. (2025). Generative AI in K-12 education: The CyberScholar initiative. *Computers and Society*. <https://doi.org/10.48550/arXiv.2502.19422>
- Chan, C. K. Y., & Lee, K. K. W. (2025). The balancing act between AI and authenticity in assessment: A case study of secondary school students' use of genAI in reflective writing. *Computers & Education*, 238. <https://doi.org/10.1016/j.compedu.2025.105399>
- Chaturvedi, R., Verma, S., Srivastava, V., & Khot, S. S. (2025). Exploring the frontier of anthropomorphism in AI agents: Trends and way forward. *Business and Society Review*, 130(1), 42–80. <https://doi.org/10.1111/basr.70002>
- Chen, B., Bao, L., Zhang, R., Zhang, J., Liu, F., Wang, S., & Li, M. (2024). A multi-strategy computer-assisted EFL writing learning system with deep learning incorporated and its effects on learning: A writing feedback perspective. *Journal of Educational Computing Research*, 61(8). <https://doi.org/10.1177/07356331231189294>
- Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. *Working Papers on Bilingualism* 9, 1–43.
- Cummins, J. (1980). The construct of language proficiency in bilingual education. *Georgetown University Round Table on Languages and Linguistics*, 81–103.
- Cummins, J. (2000). Language, power, and pedagogy: Bilingual children in the crossfire. *Bilingual Research Journal*, 25(3), 405–412.
- Dang, C. T., & Nguyen, A. (2025). Distinguishing fact from fiction: Student traits, attitudes, and AI hallucination detection in business school assessment. *arXiv:2506.00050*. <https://doi.org/10.48550/arXiv.2506.00050>
- Dunne, B. (2012). *The text (B1)*. <https://learnenglishteens.britishcouncil.org/study-break/reading-zone/b1-graded-reading/text-b1>
- Ellis, R. (2009). Task-based language learning and teaching. Oxford UP.
- Fürstenberg, M. (2025). Zur Qualität KI-generierten Feedbacks. Ein explorativer Vergleich menschlicher und künstlicher Intelligenzen. In H. G. Müller & M. Fürstenberg (Eds.), *DeutschGPT 2.0. Deutschunterricht im Dialog mit Künstlicher Intelligenz* (pp. 163–190). Frank & Timme. <https://doi.org/10.26530/20.500.12657/104475>
- Fürstenberg, M., & Müller, H. G. (2024). KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. *Der Deutschunterricht*, 5, 2–13.
- García, O., & Wei, L. (2015). Translanguaging, bilingualism, and bilingual education. In W. E. Wright, S. Boun, & O. García (Eds.), *The handbook of bilingual and multilingual education* (pp. 223–240). Wiley-Blackwell. <https://doi.org/10.1057/9781137385765>
- Gonsalves, C. (2024). Generative AI's impact on critical thinking: Revisiting Bloom's taxonomy. *Journal of Marketing Education*, 48(1), 4–19. <https://doi.org/10.1177/02734753241305980>
- Guariento, W., & Morley, J. (2001). Text and task authenticity in the EFL classroom. *ELT Journal*, 55(4), 347–353. <https://doi.org/10.1093/elt/55.4.347>
- Harris, K., Santangelo, T., & Graham, S. (2010). Metacognition and strategies instruction in writing. In H. S. Waters & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 226–256). Guilford.

- Harris, K. R., Graham, S., MacArthur, C. A., Reid, R., & Mason, L. H. (2011). Self-regulated learning processes and children's writing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 187–202). Routledge/Taylor & Francis.
- Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3), 369–388. <https://doi.org/10.1177/0741088312451260>
- Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum.
- Herdina, P., & Jessner, U. (2002). *A dynamic model of multilingualism. Perspectives of change in psycholinguistics*. Multilingual Matters. <https://doi.org/10.21832/9781853595547>
- Jessner, U. (2006). *Linguistic awareness in multilinguals: English as a third language*. Edinburgh UP. <https://doi.org/10.1515/9780748626540>
- Kangasharju, A., Ilomäki, L., Lakkala, M., & Toom, A. (2022). Lower secondary students' poetry writing with the AI-based Poetry Machine. *Computers and Education: Artificial Intelligence*, 3. <https://doi.org/10.1016/j.caeai.2022.100048>
- Kim, J., & Im, I. (2023). Anthropomorphic response: Understanding interactions between humans and artificial intelligence agents. *Computers in Human Behavior*, 139. <https://doi.org/10.1016/j.chb.2022.107512>
- Kim, J., Yu, S., Detrick, R., & Li, N. (2025). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies*, 30(1), 1265–1300. <https://doi.org/10.1007/s10639-024-12878-7>
- Kuckartz, U., & Rädiker, S. (2023). *Qualitative content analysis: Methods, practice and software* (2nd ed.). SAGE.
- Kutzner, A. (in press). KI-Schreibstrategien: Empirische Einblicke in das Schreiben mit Künstlicher Intelligenz. *ide – informationen zur deutschdidaktik*, 4.
- Kutzner, A. & Schindler, K. (2026). Writing a fairy tale with a little help of ChatGPT – Prompting experiences of fourth-graders. *Journal für Medienlinguistik (jfml)*, 8(2): Talking AI. Linguistic Approaches to Artificial Intelligence, 36–72. <https://doi.org/10.21248/jfml.2026.73>
- Lei, X. (2016). Understanding writing strategy use from a sociocultural perspective: The case of skilled and less-skilled writers. *System*, 60(1), 105–116. <https://doi.org/10.1016/j.system.2016.06.006>
- Li, S. (2025). Generative AI and second language writing. *Digital Studies in Language and Literature*, 2(1), 122–152. <https://doi.org/10.1515/dsll-2025-0007>
- Li, T., Wang, M., & Wang, F. (2025). Anthropomorphism of artificial intelligence service agent and consumer responses: A systematic literature review and future research agenda. *International Journal of Consumer Studies*, 49(3), 1–51. <https://doi.org/10.1111/ijcs.70066>
- Manchón, R. M., & Roca De Larios, J. (2007). On the temporal nature of planning in L1 and L2 composing. *Language Learning*, 57(4), 549–593. <https://doi.org/10.1111/j.1467-9922.2007.00428.x>
- Ministerium für Schule und Bildung (2019). *Verwaltungsvorschriften zur Verordnung über die Ausbildung und die Abschlussprüfungen in der Sekundarstufe I (VVzAPO-S I)*. <https://bass.schule.nrw/Service/12691.htm>
- Müller-Hartmann, A., & Schocker-von Dittfurth, M. (2011). *Task-supported language learning*. Schöningh. <https://doi.org/10.36198/9783838533360>



- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy. An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 1–11.  
<https://doi.org/10.1016/j.caeai.2021.100041>
- Odlin, T. (1989). *Language transfer. Cross-linguistic influence in language learning*. Cambridge UP.  
<https://doi.org/10.1017/CBO9781139524537>
- Pitzen, L. (2025). Peer- vs. KI-Feedback. Wie Schüler\*innen der Sekundarstufe I Textfeedback zu einer Zusammenfassung rezipieren. In K. Schindler, N. Wilder, & Virtuelles Kompetenzzentrum für Künstliche Intelligenz und Wissenschaftliches Arbeiten (VK:KIWA) (Eds.), *Kleine Kieler Beiträge zu Künstlicher Intelligenz. Kiel: Virtuelles Kompetenzzentrum für Künstliche Intelligenz und Wissenschaftliches Arbeiten (VK:KIWA)*. <https://doi.org/10.38071/2025-01281-7>
- Putri, B. G., Suryati, N., Ivone, F. M., & Hayati, N. (2025). The role of AI in academic writing: Unveiling strategies among EFL students. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, 12(2), 774–792. <https://doi.org/10.22219/celtic.v12i2.42039>
- Raoofi, S., Binandeh, M., & Rahmani, S. (2017). An investigation into writing strategies and writing proficiency of university students. *Journal of Language Teaching and Research* 8(1), 191–198.  
<http://dx.doi.org/10.17507/jltr.0801.24>
- Rezat, S., & Schindler, K. (2025). Textgenerierende KI im Deutschunterricht: Texte lesen und schreiben mit KI. *Praxis Deutsch*, 311, 4–13.
- Roca De Larios, J., Marin, J., & Murphy, L. (2001). A temporal analysis of formulation processes in L1 and L2 writing. *Language Learning*, 51(3), 497–538. <https://doi.org/10.1111/0023-8333.00163>
- Schneider, W., Schlagmüller, M., & Ennemoser, M. (2017). *LGVT 5–12+. Lesegeschwindigkeits- und -verständnistest für die Klassen 5–12* (2nd ed.). Hogrefe.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158. <https://doi.org/10.1093/applin/11.2.129>
- Shen, X., & Teng, M.F. (2024). Three-wave cross-lagged model on the correlations between critical thinking skills, self-directed learning competency and AI-assisted writing. *Thinking Skills and Creativity*, 52. <https://doi.org/10.1016/j.tsc.2024.101524>
- Siregar, A. R. F., & Hz, B. I. R. (2025). Undergraduate students' perceptions on utilizing Deepseek AI as a tool in writing journal articles. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, 12(2), 713–729. <https://doi.org/10.22219/celtic.v12i2.41375>
- Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14.  
<https://doi.org/10.3389/fpsyg.2023.1260843>
- Steinhoff, T., & Lehnen, K. (2025). Schreiben mit Künstlicher Intelligenz: Das GPT-Modell (Ghost, Partner, Tutor). *Leseräume*, 11.
- Summer, T. (2012). From method to postmethod. In M. Eisenmann & T. Summer (Eds.), *Basic issues in EFL teaching and learning* (pp. 1–15). Winter.
- Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, 16(3), 371–391.  
<https://doi.org/10.1093/applin/16.3.371>
- Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, 62. <https://doi.org/10.1016/j.jslw.2023.101071>

- Werdiningsih, I., Marzuki, & Rusdin, D. (2024). Balancing AI and authenticity: EFL students' experiences with ChatGPT in academic writing. *Cogent Arts & Humanities*, 11(1). <https://doi.org/10.1080/23311983.2024.2392388>
- Xu, T., Huang, F., & Teo, T. (2025). Understanding EFL learners' strategies in AI-assisted English writing: An activity theory perspective. *Porta Linguarum. An International Journal of Foreign Language Teaching and Learning*, 8, 69–87. <https://doi.org/10.30827/portalin.viXIII.32849>
- Yan, L., Pammer-Schindler, V., Mills, C., Nguyen, A., & Gašević, D. (2025). Beyond efficiency: Empirical insights on generative AI's impact on cognition, metacognition and epistemic agency in learning. *British Journal of Educational Technology*, 56(5), 1675–1685. <https://doi.org/10.1111/bjet.70000>
- Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(37), Article 28. <https://doi.org/10.1186/s40561-024-00316-7>
- Zheldibayeva, R., Nascimento, A. K. D. O., Castro, V., Kalantzis, M., & Cope, B. (2025). The impact of AI-driven tools on student writing development: A case study. *Online Journal of Communication and Media Technologies*, 15(3), Article e202526. <https://doi.org/10.30935/ojcm/16738>

Author's address:

Alyssa Kutzner, University of Wuppertal, Faculty of Humanities and Cultural Studies | German Studies,  
German Language Education, Chair of Prof. Dr. Kirsten Schindler  
[kutzner@uni-wuppertal.de](mailto:kutzner@uni-wuppertal.de)



## Appendix A

*Reading Test Results (Sorted by Performance: Cambridge Scale Score/Grade/CEFR Level and Reading Comprehension)*

Student code	<b>Cambridge Preliminary</b>		<b>Reading comprehension (RC) (Leseverständnis (LV))</b>		<b>Reading speed (RS) (Lesegeschwindigkeit (LGS))</b>		<b>Reading accuracy (RA) (Lesegenauigkeit (LGN))</b>	
	Cambridge Scale Score	grade / CEFR level	T score	Classification	T score	Classification	T score	Classification
dat-k6n	170	A–B2	58	Upper average range	53	Upper average range	63→73	Above average to very high
fsl-dda	167	A–B2	60	Above average	60	Above average	54	Upper average range
lqm-cml	160	A–B2	48	Lower average range	41	Lower average range	56	Upper average range
mm4-ysy	160	A–B2	46	Lower average range	41	Lower average range	56	Upper average range
1jj-px3	157	B–B1	55	Upper average range	50	Upper average range	59	Upper average range
b5b-il8	157	B–B1	54	Upper average range	48	Lower average range	63→73	Above average to very high
z9c-nry	157	B–B1	48	Lower average range	41	Lower average range	63→73	Above average to very high
c6a-vgp	153	B–B1	43	Lower average range	43	Lower average range	46	Lower average range
uqe-0qh	153	B–B1	40	Lower average range	37	Below average	50	Upper average range

8ds-ok0	150	C–B1	47	Lower average range	43	Lower average range	51	Upper average range
o5b-gpc	150	C–B1	42	Lower average range	39	Below average	50	Upper average range
sld-dpz	150	C–B1	41	Lower average range	41	Lower average range	46	Lower average range
iyv-svk	147	C–B1	50	Upper average range	49	Lower average range	52	Upper average range
s9c-vnm	143	C–B1	35	Below average	37	Below average	40	Lower average range
rv5-gjh	143	C–B1	30	Below average	<27	Very low	50	Upper average range
auy-bq3	134	fail–A2	42	Lower average range	39	Below average	50	Upper average range
ovz-kbj	130	fail–A2	31	Below average	34	Below average	37	Below average
ouf-wbf	128	fail–A2	31	Below average	42	Lower average range	33	Below average
qqs-yax	122	fail–A2	42	Lower average range	39	Below average	50	Upper average range
kzo-wng	120	fail–A2	44	Lower average range	39	Below average	56	Upper average range

## Appendix B

Overview of Individual Results for Reading Performance, AI Writing Strategies, and Cross-Linguistic Transfer (Jaccard Indices)

(Sorted by Performance: Cambridge Scale Score/CEFR Level/Grade and Reading Comprehension)

Student code	Cambridge Preliminary		Reading comprehension (RC) (Leseverständnis (LV))		Reading speed (RS) (Lesegeschwindigkeit (LGS))		Reading accuracy (RA) (Lesegenauigkeit (LGN))		AI writing strategies		
	Cambridge Scale Score	Grade/CEFR level	T score	Classification	T score	Classification	T score	Classification	L1/L2 (German)	FL (English)	Transfer (Jaccard index)
dat-k6n	170	A–B2	58	Upper average range	53	Upper Average Range	63→73	Above average to very high	• focally-structuring     • AI affiliation     • differentially-limiting	• focally-structuring     • metacognitively-inquiring     • AI affiliation	.5
fsl-dda	167	A–B2	60	Above average	60	Above Average	54	Upper average range	• functionally-pragmatizing     • metacognitively-inquiring     • AI affiliation     • AI rejection	• functionally-pragmatizing     • AI rejection	.5

lqm-cml	160	A–B2	48	Lower average range	41	Lower Average Range	56	Upper average range	• focally-structuring     • metacognitively-inquiring	• focally-structuring     • holistically-optimizing     • metacognitively-inquiring	.67
mm4-ysy	160	A–B2	46	Lower average range	41	Lower Average Range	56	Upper average range	• holistically-optimizing     • metacognitively-inquiring	• holistically-optimizing     • metacognitively-inquiring	1.0
1jj-px3	157	B–B1	55	Upper average range	50	Upper Average Range	59	Upper average range	• holistically-optimizing     • functionally-pragmatizing     • metacognitively-inquiring	• focally-structuring     • holistically-optimizing     • metacognitively-inquiring	.5
b5b-il8	157	B–B1	54	Upper average range	48	Lower Average Range	63→73	Above average to very high	• focally-structuring     • metacognitively-inquiring	• focally-structuring     • holistically-optimizing     • metacognitively-inquiring     • AI affiliation	.5
z9c-nry	157	B–B1	48	Lower average range	41	Lower Average Range	63→73	Above average to very high	• focally-structuring     • holistically-optimizing     • functionally-pragmatizing	• focally-structuring     • holistically-optimizing	.5

									• metacognitively-inquiring		
c6a-vgp	153	B–B1	43	Lower average range	43	Lower Average Range	46	Lower average range	• functionally-pragmatizing     • AI affiliation     • differentially-limiting	• focally-structuring     • functionally-pragmatizing	.25
uqe-0qh	153	B–B1	40	Lower average range	37	Below Average	50	Upper average range	• functionally-pragmatizing     • metacognitively-inquiring     • AI affiliation (1)	• focally-structuring     • functionally-pragmatizing	.25
8ds-ok0	150	C–B1	47	Lower average range	43	Lower Average Range	51	Upper average range	• metacognitively-inquiring     • differentially-limiting	• metacognitively-inquiring     • differentially-limiting	1.0
o5b-gpc	150	C–B1	42	Lower average range	39	Below Average	50	Upper average range	• focally-structuring     • holistically-optimizing     • metacognitively-inquiring	• focally-structuring     • holistically-optimizing     • metacognitively-inquiring	1.0
sld-dpz	150	C–B1	41	Lower average range	41	Lower Average Range	46	Lower average range	• focally-structuring     • functionally-pragmatizing	• focally-structuring     • holistically-optimizing     • functionally-pragmatizing     • metacognitively-inquiring	.5

iyv-svk	147	C–B1	50	Upper average range	49	Lower Average Range	52	Upper average range	• focally-structuring     • functionally-pragmatizing     • metacognitively-inquiring	• focally-structuring     • metacognitively-inquiring     • AI rejection	.5
s9c-vnm	143	C–B1	35	Below average	37	Below Average	40	Lower average range	• holistically-optimizing     • differentially-limiting	• holistically-optimizing     • metacognitively-inquiring     • differentially-limiting	.67
rv5-gjh	143	C–B1	30	Below average	<27	Very Low	50	Upper average range	• metacognitively-inquiring     • AI affiliation	• focally-structuring     • metacognitively-inquiring     • differentially-limiting	.25
auy-bq3	134	fail–A2	42	Lower average range	39	Below Average	50	Upper average range	• holistically-optimizing     • metacognitively-inquiring     • AI affiliation	• metacognitively-inquiring     • differentially-limiting	.25
ovz-kbj	130	fail–A2	31	Below average	34	Below Average	37	Below average	• metacognitively-inquiring     • differentially-limiting	• differentially-limiting	.5

ouf-wbf	128	fail–A2	31	Below average	42	Lower Average Range	33	Below average	• metacognitively-inquiring     • differentially-limiting	• metacognitively-inquiring     • differentially-limiting	1.0
qqs-yax	122	fail–A2	42	Lower average range	39	Below Average	50	Upper average range	• differentially-limiting	• focally-structuring     • metacognitively-inquiring	0
kzo-wng	120	fail–A2	44	Lower average range	39	Below Average	56	Upper average range	• focally-structuring     • holistically-optimizing     • metacognitively-inquiring	• focally-structuring     • holistically-optimizing     • metacognitively-inquiring	1



Nina Schnatz, Aline Bieri & Cerstin Mahlow

## Exemplifying methodological triangulation in GenAI-supported writing research

### A course-based study

#### 1 Introduction

The integration of generative AI (GenAI) into academic writing (Dunnigan et al., 2025; Helm & Hesse, 2024; Kelly et al., 2023; Rapp et al., 2025) is reshaping not only how texts are produced but also how knowledge is constructed and evaluated. This poses profound didactic and epistemic challenges (Anson et al., 2025). Research on writing studies is increasingly conceptualizing scientific writing as an epistemic and genre-based practice situated in disciplinary conventions (Chen, 2019; Kruse, 2013; Pohl, 2007, 2016). In GenAI-supported contexts, these epistemic and metacognitive processes partially become distributed across human and machine agents. Recent work shows that GenAI can boost task performance while weakening metacognitive accuracy (Fernandes et al., 2026), thereby raising new questions about how writers maintain epistemic agency when goal setting, evaluation, and reasoning are shared with AI systems. Alvarado (2023) conceptualizes GenAI systems as “epistemic technologies”—tools that shape how meaning is constructed through reasoning and explanation. Their use in education therefore redefines what it means to “know” or “learn” through writing. Similarly, Parker et al. (2024) identify GenAI-supported writing as a hybrid form of collaboration involving dynamic brainstorming, negotiation of meaning, and comparative evaluation. These interactions unfold across prompts, drafts, and feedback cycles, yet often leave no visible trace in the final product. This invisibility complicates both instruction and research, because textual analysis alone cannot reveal how epistemic decisions are made. Studying writing with GenAI thus requires triangulated methods that capture both written outcomes—the product—and the processes of negotiation, decision making, and revision. This article adopts a method-focused lens, exploring triangulation (Flick, 2023) as a strategy to enhance epistemic transparency in GenAI-supported writing. We first outline key methodological challenges, then present triangulation strategies that address these issues. Drawing on our own iterative course-based study (Schnatz & Bieri, 2024, 2025), we illustrate how student texts, prompting journals, and additional data can be combined to trace how epistemic agency is enacted across writing phases.

#### Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 158–165

DOI: 10.21248/dideu.974

**Copyright** Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

## 2 Current methodological challenges

Since the release of accessible large language models (LLMs), a growing body of research has documented tool use and student behavior (e.g., Cieliebak et al., 2023; Helm & Hesse, 2024; Rapp et al., 2025). These studies offer valuable descriptive insights, yet often stop short of examining how GenAI use affects the quality, structure, or epistemic depths of writing. Evidence for deeper rhetorical or structural development, particularly in non-EFL contexts,<sup>1</sup> remains limited. Pedagogical integration has likewise received only partial attention. Although Kelly et al. (2023) emphasize the need for GenAI tool literacy, curriculum integration, and assessment redesign, most studies stop at offering recommendations rather than going on to engage in empirical analyses of classroom practice. Process-oriented studies do exist (e.g., Cummings et al., 2024), but they remain isolated and methodologically constrained. Taken together, current research leaves several methodological gaps. These can be summarized as three interrelated challenges: ethical concerns and originality, balancing AI and human insight, and assessment difficulties (see Table 1).

**Table 1**

*Methodological Challenges in GenAI-Assisted Writing Research*

Challenge	Description
<b>Ethical concerns + originality</b>	Blurred boundary between student- and AI-generated content complicates authorship, originality, and academic integrity. Reliable capture of distributed agency is difficult without breaching ethical norms.
<b>Balancing AI + human insight</b>	Risk that overreliance on GenAI may reduce opportunities for critical thinking and rhetorical development. Invisible cognitive processes make negotiation of agency hard to trace.
<b>Assessment difficulties</b>	Traditional competence measures fail: Grammar/style gains may stem from tool affordances rather than student ability, thereby complicating fair evaluation.

Addressing these intertwined challenges requires robust triangulated process-focused research designs that can (a) trace how students use GenAI tools during writing; (b) link such use to measurable dimensions of text quality; and (c) situate these analyses within defined teaching contexts by combining textual analysis with data on tool interaction, student reflection, and instructional framing. This article therefore proposes a triangulated approach to researching GenAI-assisted writing. Triangulation refers to the systematic use of multiple methods and/or data sources to investigate one and the same phenomenon. In our project, we apply both data triangulation (integrating student texts, prompting journals, and survey responses) and methodological triangulation (combining text analysis, surveys, QDA, and QCA).

<sup>1</sup> When textual outcomes are assessed, the focus lies predominantly on English as a Foreign Language (EFL) settings (e.g., Song & Song, 2023; Tsai, Lin & Brown, 2024; Tseng & Lin, 2024).

### 3 Methods: Course context, data, and triangulated design

This study takes a mixed-methods approach within a design-based research (DBR) (Reinmann, 2023) framework, aiming (a) to examine how GenAI shapes writing processes and outcomes (see Schnatz & Bieri, 2024, 2025), and (b) to iteratively refine didactic strategies for GenAI-integrated writing instruction.<sup>2</sup> To address the methodological challenges mentioned in Table 1, triangulation was applied at two levels and is displayed in detail in the following Tables 2 and 3.

**Table 2**

#### *Triangulated Data Sources*

Source	Scope	Purpose
<b>Student texts (in German)</b>	Approx. 150 per iteration (2023–2025)	Compare rhetorical quality across Pre-/Post-GenAI cohorts
<b>Prompting journals</b>	Approx. 150 per iteration (2025)	Document tool use, reflection, epistemic decision making
<b>Midterm surveys</b>	$n = 22$ (2024), $n = 24$ (2025)	Capture perceptions of course design and GenAI literacy
<b>Planned (Iteration 3, 2026)</b>	Approx. 10 student texts + approx. 10 journals + pre/post surveys Recordings of coaching sessions	Enrich triangulation with coaching session recordings and longitudinal perception data

To support methodological transparency, several data sources were combined to capture both product- and process-related dimensions of GenAI-supported writing. Entries in the prompting journals frequently documented conceptual evaluations such as noting that suggested terms were “too specific” or that a proposed broader label was “actually a synonym.” Such process notes provided evidence of how writers assessed and refined AI-generated terminology during drafting.

Planned analyses combine (a) comparative text analysis of Pre-/Post-GenAI cohorts, (b) thematic and content analysis of prompting journals, (c) focused case studies of 10 projects, and (d) quantitative + qualitative survey evaluation. This design enables triangulation of product quality, process documentation, and student perceptions within a structured didactic setting. After data collection, the project

<sup>2</sup> The study was conducted in a second-semester scientific writing course at the Zürcher Hochschule für Angewandte Wissenschaften (ZHAW) with approximately 600 engineering students per term. Students write a scientific paper in German in groups of 3–4 students. Across two iterations (2024, 2025), GenAI tools are integrated into each lesson along the phases of writing (e.g., topic selection, drafting, and revision). The 2025 iteration added prompting journals and self-study units (video inputs, tutorials, genre-based exercises such as Swales’ [2014] CaRS model for introductions and Weissberg & Buker [1990] for discussions). The course design fosters procedural competence and epistemic awareness, while also contributing—indirectly—to students’ epistemic agency in hybrid human–AI writing contexts (see Barzilai & Chinn, 2018; Schnatz & Bieri, 2025).

aims to use the following methods detailed in Table 3 to address the methodological challenges described in the previous section.

**Table 3**

*Triangulated Methods*

Approach	Data	Focus
<b>Comparative text analysis</b>	Approx. 150 student texts (in German) per iteration (2023–2025)	Assess structural quality (e.g., CaRS-based intros), framing clarity, rhetorical coherence; later corpus-linguistic analysis of lexical/syntactic patterns
<b>Thematic and content analysis</b>	Approx. 150 prompting journals (2025)	Trace epistemic decision making/agency (goal setting, revision, tool evaluation) via QDA + QCA
<b>Survey analysis</b>	Midterm surveys (2024 $n = 22$ , 2025 $n = 24$ )	Perform a quantitative comparison of closed items; qualitative coding of open responses to assess course design and tool use
<b>Planned (Iteration 3, 2026)</b>	Coaching recordings, pre/post surveys, smaller cohort (approx. 40)	Trace epistemic agency through real-time instructor–student dialogue and longitudinal perception shifts

This design allows us to investigate GenAI’s impact on student writing at multiple levels: outcome quality, process transparency, and pedagogical integration. By combining student texts, prompting journals, surveys, and focused case studies, we can reconstruct how epistemic agency is enacted or compromised across writing phases. Triangulation both strengthens empirical validity and provides a methodological basis for refining instructional strategies that promote epistemic integrity in GenAI-supported academic writing.

#### **4 Methodological insights from a triangulated approach**

The triangulated design offers a promising foundation for researching GenAI-assisted writing by uncovering dimensions of writing that single methods overlook. The preliminary study of prompting journals along with initial thematic coding and the evaluation of the midterm surveys illustrate both the strengths of triangulation and the general challenges that remain. Prompting journals make visible how students interact with GenAI during different phases of writing (see, e.g., Table 4). They show pathways for linking GenAI use to rhetorical development and decision making. Mapping prompts across writing phases appears particularly useful for reconstructing trajectories of goal setting and revision.

**Table 4**

*Illustration of Prompting Journal Entry (Example From a Student Group) Reflecting on the Use of ChatGPT for Literature Research*

AI Tool	Goal	Prompt <sup>3</sup>	Results and Reflection
ChatGPT	Literature research for theoretical foundations chapter	Find me current sources on the topic of metal 3D printing and conventional manufacturing. Find sources that describe the economic efficiency of the manufacturing processes.	Provides few good sources, very often confuses metal 3D printing with normal plastic 3D printing.

Whereas the prompting-journal entry above captures students' moment-to-moment reflections on their interaction with GenAI, the written products themselves provide a different lens on how epistemic decisions are enacted. The following excerpt (Figure 1 and Table 5) demonstrates how these textual traces were coded along our three epistemic categories.

**Figure 1**

*Illustration of Initial Coding in Student Text*

## 5.4. Ausblick für zukünftige Forschung

Um die Verbreitung und Wirkung einzelner Dark Patterns im Nutzungskontext fundiert zu analysieren, sind quantitative Studien notwendig. Methoden wie Click-Tracking oder Eye-Tracking können dabei helfen, die tatsächliche Aufmerksamkeit der Nutzer:innen zu erfassen und Rückschlüsse auf die Effektivität manipulativer Designs zu ziehen.

future research suggestions

methodological justification

Darüber hinaus sind Längsschnittstudien erforderlich, um die langfristigen Auswirkungen solcher Gestaltungsmuster auf Vertrauen, Nutzerbindung und Wiederkaufsabsicht

longitudinal design recommendation

zu untersuchen. Nur durch die Beobachtung über einen längeren Zeitraum lassen sich verlässliche Aussagen über mögliche negative Konsequenzen treffen.

reasoning

Da anzunehmen ist, dass Dark Patterns auf verschiedenen Plattformen unterschiedlich wirken – etwa in sozialen Netzwerken, Smartphone-Apps – erscheint eine plattformübergreifende Analyse sinnvoll. Solche Vergleiche könnten nicht nur differenzierte Einblicke in die Wirkmechanismen geben, sondern auch die Grundlage für gezielte regulatorische Massnahmen schaffen.

contextual differentiation

policy linkage

Gleichzeitig sollte die Forschung zu sogenannten „Bright Patterns“ und effektiven Digital Nudges intensiviert werden. Sie bieten ethisch vertretbare Alternativen zum manipulativen Design und zeigen, dass digitale Gestaltung auch im Sinne der Nutzer:innen möglich ist – etwa durch Transparenz, Nutzerautonomie und förderliches Entscheidungsverhalten.

ethical re-orientation/ outlook

positive vision

<sup>3</sup> Originally written in German; prompt and reflection translated by the authors.

**Table 5***Initial Coding Scheme of Epistemic Agency Traceable in the Student Text*

„Um die Verbreitung und Wirkung ... fundiert zu analysieren, sind quantitative Studien notwendig.“	Decision-making – Future research design suggestion	Proposes methodological next step; procedural foresight.
„Methoden wie Click-Tracking oder Eye-Tracking ...“	Evaluation – Methodological justification / instrument relevance	Evaluates suitability of specific tools for analyzing cognitive attention.
„Darüber hinaus sind Längsschnittstudien erforderlich ...“	Decision-making – Longitudinal design recommendation	Outlines temporal extension; structural proposal.
„Nur durch die Beobachtung ... lassen sich verlässliche Aussagen treffen.“	Reflection – Epistemic reasoning / validity reflection	Argues for longitudinal validity; metacognitive awareness of research logic.
„Da anzunehmen ist, dass Dark Patterns auf verschiedenen Plattformen unterschiedlich wirken ...“	Evaluation – Contextual differentiation / comparative reasoning	Analyzes variation potential; evaluative generalisation.
„Solche Vergleiche könnten ... Grundlage für gezielte regulatorische Massnahmen schaffen.“	Reflection & Justification – Societal implication / policy linkage	Reflects on regulatory utility; normative extrapolation.
„Gleichzeitig sollte die Forschung zu ‚Bright Patterns‘ ... intensiviert werden.“	Reflection & Justification – Ethical re-orientation / constructive outlook	Advocates value-oriented research agenda; transformative epistemic stance.
„Sie bieten ethisch vertretbare Alternativen ... Transparenz, Nutzerautonomie ...“	Reflection – Positive vision / normative synthesis	Concludes with constructive ethical synthesis; high-level reflective agency.

*Note:* blue = decision making, yellow = evaluation, green = self-regulation and reflection.

However, triangulation does not fully resolve key methodological difficulties. Ethical ambiguities remain around authorship, transparency, and the extent of student consent when journals expose cognitive processes. Interpreting GenAI's influence continues to be tricky: Even with prompts and outputs available, the rationale for why students accepted, modified, or ignored suggestions is not always recoverable. Although prompting journals improved epistemic visibility, they introduced issues of standardization: Students varied in how thoroughly and reflectively they documented their interactions. Finally, triangulated designs are resource-intensive, requiring significant time for data collection, coding, and integration across sources.

## 5 Discussion

The analyses are expected to contribute to both writing research and instructional design in GenAI-supported academic contexts. Linking prompting journals, writing-phase mapping, and text features also allows for the design of a transferable course model focusing on teaching writing with GenAI, while supporting and tracing the unfolding of epistemic agency during writing. A second contribution lies in conceptualizing prompting literacy as both a didactic and measurable competence. Prompts reflect epistemic behavior—how students frame questions, assess feedback, and revise accordingly. These practices can be scaffolded through teaching and serve as indicators of reflective engagement. Prompting literacy thus emerges as a key component of academic AI literacy. The triangulated design also enhances reproducibility and international transferability. Embedding a triangulated design in a structured writing curriculum and developing transferable materials aligns with open educational practices and current demands for didactic transparency. Finally, the study addresses a broader pedagogical challenge: how to preserve epistemic agency and integrity in technology-enhanced learning environments. Because of the phenomenon's inherent complexity, a triangulated approach can provide the depth and balance needed to address it in a meaningful way.

## References

- Alvarado, R. (2023). AI as an epistemic technology. *Science and Engineering Ethics*, 29(5), Article 32. <https://doi.org/10.1007/s11948-023-00451-3>
- Anson, Ch. R., Delorme Benites, A., Ganobcsik-Williams, L., Mahlow, C., & McKee, H. A. (2025). Reflections on writing and generative AI: A symposium. *Journal of Academic Writing*, 15(S2), 1–14. <https://doi.org/10.18552/joaw.v15is2.1121>
- Barzilai, S., & Chinn, C. A. (2018). On the goals of epistemic education: Promoting apt epistemic performance. *Journal of the Learning Sciences*, 27(3), 353–389. <https://doi.org/10.1080/10508406.2017.1392968>
- Chen, Y.-C. (2019). Writing as an epistemological tool: Perspectives from personal, disciplinary, and sociocultural landscapes. In V. Prain & B. Hand (Eds.), *Theorizing the future of science education research* (Vol. 49, pp. 115–132). Springer International Publishing. [https://doi.org/10.1007/978-3-030-24013-4\\_8](https://doi.org/10.1007/978-3-030-24013-4_8)
- Cieliebak, M., Drewek, A., Grob, K. J., Kruse, O., Mlynchuk, K., Rapp, C., & Waller, G. (2023). *Generative KI beim Verfassen von Bachelorarbeiten: Ergebnisse einer Studierendenbefragung im Juli 2023*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. <https://doi.org/10.21256/ZHAW-2491>
- Cummings, R. E., Monroe, S. M., & Watkins, M. (2024). Generative AI in first-year writing: An early analysis of affordances, limitations, and a framework for the future. *Computers and Composition*, 71, Article 102827. <https://doi.org/10.1016/j.compcom.2024.102827>
- Dunnigan, J., Kozak, M., Pearce, N., & Rasmussen, H. T. (2025). Introduction to the AI special edition themed issue: The role of generative artificial intelligence (AI) in doctoral research and writing. *Impacting Education: Journal on Transforming Professional Practice*, 10(1). <https://doi.org/10.5195/ie.2025.529>
- Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., & Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, 175, Article 108779. <https://doi.org/10.1016/j.chb.2025.108779>
- Flick, U. (2023). *An introduction to qualitative research* (7th ed.). Sage.
- Helm, G., & Hesse, F. (2024). Usage and beliefs of student teachers towards artificial intelligence in writing. *Research in Subject-Matter Teaching and Learning (RISTAL)*, 7(1), 1–18. <https://doi.org/10.2478/ristal-2024-0001>
- Kelly, A., Sullivan, M., & Strampel, K. (2023). Generative artificial intelligence: University student awareness, experience, and confidence in use across disciplines. *Journal of University Teaching and Learning Practice*, 20(6). <https://doi.org/10.53761/1.20.6.12>
- Kruse, O. (2013). Perspectives on academic writing in European higher education: Genres, practices, and competences. *REDU: Revista de Docencia Universitaria*, 11(1), 37–58. <https://doi.org/10.4995/redu.2013.5591>
- Parker, J. L., Richard, V. M., Acabá, A., Escoffier, S., Flaherty, S., Jablonka, S., & Becker, K. P. (2024). Negotiating meaning with machines: AI's role in doctoral writing pedagogy. *International Journal of Artificial Intelligence in Education*, 35, 1218–1238. <https://doi.org/10.1007/s40593-024-00425-x>
- Pohl, T. (2007). *Studien zur Ontogenese wissenschaftlichen Schreibens*. De Gruyter. <https://doi.org/10.1515/9783110946116>



- Pohl, T. (2016). Studentisches Schreiben in Geschichte und Gegenwart. In A. Hirsch-Weber & S. Scherer (Eds.), *Wissenschaftliches Schreiben in Natur- und Technikwissenschaften* (pp. 179–193). Springer Fachmedien Wiesbaden. [https://doi.org/10.1007/978-3-658-12211-9\\_10](https://doi.org/10.1007/978-3-658-12211-9_10)
- Rapp, C., Jakob, K., Drewek, A., Kruse, O., & Waller, G. (2025). *Generative KI beim Verfassen von Abschlussarbeiten: Ergebnisse einer Studierenden- und Dozierendenumfrage im Juni/Juli 2024*. ZHAW Zürcher Hochschule für Angewandte Wissenschaften. <https://doi.org/10.21256/ZHAW-2549>
- Reinmann, G. (2023). Design-based research in der Hochschuldidaktik: Forschen für Lehrinnovationen. In *Hochschuldidaktik als Wissenschaft: Disziplinäre, interdisziplinäre und transdisziplinäre Perspektiven* (pp. 269–286). transcript. <https://doi.org/10.25656/01:31878>
- Schnatz, N. I., & Bieri, A. (2024). Artificial intelligence in scientific writing – A new course design for undergraduate students. *EDULEARN24 Proceedings* (pp. 5502–5510). <https://doi.org/10.21125/edulearn.2024.1336>
- Schnatz, N. I., & Bieri, A. (2025). Refining an AI-integrated scientific writing course for engineers: Insights from a design-based research approach. *EDULEARN25 Proceedings* (pp. 2394–2406). <https://doi.org/10.21125/edulearn.2025.0663>
- Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1260843>
- Tsai, C.-Y., Lin, Y.-T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and Information Technologies*, 29(17), 22427–22445. <https://doi.org/10.1007/s10639-024-12722-y>
- Tseng, Y.-C., & Lin, Y.-H. (2024). Enhancing English as a foreign language (EFL) learners' writing with ChatGPT: A university-level course design. *Electronic Journal of e-Learning*, 22(2), 78–97. <https://doi.org/10.34190/ejel.21.5.3329>
- Swales, J. (2014). Create a research space (CARS) model of research introductions. In *Writing about writing: A college reader* (2nd ed., pp. 12–15). Bedford/St. Martin's.
- Weissberg, R., & Buker, S. (1990). *Writing up research*. Prentice Hall Regents.

Authors' addresses:

Nina Schnatz, Institute of Language Competence, Zürcher Hochschule für Angewandte Wissenschaften, Theaterstrasse 17, 8401 Winterthur, Switzerland  
[nina.schnatz@zhaw.ch](mailto:nina.schnatz@zhaw.ch)

Aline Bieri, Institute of Language Competence, Zürcher Hochschule für Angewandte Wissenschaften, Theaterstrasse 17, 8401 Winterthur, Switzerland  
[aline.bieri@zhaw.ch](mailto:aline.bieri@zhaw.ch)

Cerstin Mahlow, Institute of Language Competence, Zürcher Hochschule für Angewandte Wissenschaften, Theaterstrasse 17, 8401 Winterthur, Switzerland  
[cerstin.mahlow@zhaw.ch](mailto:cerstin.mahlow@zhaw.ch)

Gloria Goller & Nadine Anskeit

## Empirical research on AI-assisted writing

**Mixed-methods approaches, long-term effects, and ethical-technological conditions**

### 1 Introduction

With the widespread availability of generative AI tools since the end of 2022, the discussion about writing has changed fundamentally. New possibilities for supporting, reflecting on, and individualizing writing processes are opening up (Kasneci et al., 2023; Rezat & Schindler, 2025), while concerns persist that learners may become dependent on AI and that key writing skills may be weakened (Niloy et al., 2023). Against this backdrop, a central question arises: How can the effects of AI use in writing be researched empirically in a valid and reliable manner?

Existing studies often focus on short-term output analyses, whereas sustainable effects, process data, and motivational dimensions remain underexplored. Therefore, this article aims to systematize key methodological problems in empirical research on AI-assisted writing and to identify potential solutions for intervention studies that address three core requirements: mixed-methods approaches, the examination of long-term effects, and transparent ethical and technological conditions.

As recent work shows (Fütterer et al., 2023; Kasneci et al., 2023), methodological questions cannot be separated from broader educational and ethical considerations, as the use of AI in writing directly affects pedagogical practices, assessment, and notions of authorship. These conditions shape the design, interpretation, and comparability of empirical studies and thus form the framework within which robust research on AI-assisted writing must be developed.

### 2 Methodological challenges

One central methodological challenge concerns the validity of empirical measurements. AI-supported writing processes can produce text quality that reflects technical assistance rather than actual competence gains (Rezat & Schindler, 2025). Empirical designs must therefore distinguish AI effects from learning processes through the use of, for example, process data, qualitative reflections, or control-group comparisons.

A second challenge is the insufficient integration of process-oriented perspectives. Classical writing research emphasizes product ratings, yet the interaction between writers and AI—described as human–machine “coactivity” (Lehnen & Steinhoff, 2024)—remains underexamined. Log files, prompt developments, chat histories, or self-reports are therefore necessary to capture cognitive and

#### Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur

Special Issue (2026) – Writing with AI. S. 166–170

DOI: 10.21248/dideu.975

**Copyright** Dieser Artikel wird unter der Creative-Commons-Lizenz CC BY-NC-ND 4.0

veröffentlicht: <https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

metacognitive processes and to complement product analyses, underscoring the need for practice-oriented approaches to AI use in writing that also consider the social dimensions of writing.

A third issue concerns the sustainability of effects. Most studies rely on one-off interventions, although pre–post–follow-up designs are essential for detecting both immediate and delayed learning outcomes (Pissarek & Wild, 2018). This is particularly relevant in AI-assisted writing in which short-term relief may obscure long-term implications for self-regulation and transfer.

Finally, empirical designs must account for ethical and technological conditions. Limited model transparency, data protection issues, and algorithmic biases (Gethmann et al., 2022) influence both research implementation and the interpretation of findings. Transparent reporting of technical parameters and ethical standards is therefore necessary to ensure comparability and replicability.

### **3 International state of research and methodological designs**

International research on writing with AI is characterized by a high degree of methodological and content diversity. The studies discussed here were selected because they represent different methodological orientations (e.g., mixed-methods, quasiexperimental, product-focused, comparative), different writing contexts (EFL, secondary school, creative writing, literary writing, reflective learning settings), and different analytic emphases (output vs. process, short-term vs. long-term effects, human–machine evaluation). Taken together, they illustrate the methodological challenges and opportunities that arise when examining mixed-methods requirements, long-term effects, and ethical and technological conditions.

A prominent example of experimental research is Song and Song's (2023) controlled intervention study on the influence of ChatGPT on the academic writing skills and motivation of EFL students. Their pretest–posttest mixed-methods design, supplemented by qualitative interviews, shows significant improvements in text quality and motivation. Although this triangulation addresses mixed-methods requirements, its short duration also illustrates the need for longitudinal approaches to examine sustainability.

A complementary perspective is offered by Anjarani et al.'s (2024) quasiexperimental study embedded in authentic classroom practice. By focusing on engagement and enjoyment, the study achieves high ecological validity, but offers less experimental control. In comparison to Song and Song, this work highlights a methodological tension between internal validity and ecological validity and between cognitive outcomes and motivational processes.

Product-focused studies such as Bulante et al. (2024) examine AI-assisted and non-AI-assisted writing outputs among secondary school students. Their findings show short-term improvements in linguistic quality, yet without process data or extended time frames, results remain difficult to attribute to actual competence development. Compared to Song and Song or Anjarani et al., Bulante et al.'s study illustrates the limitations of output-only designs and reinforces the need for multimodal data.

The study by Niloy et al. (2023) adds a different perspective by investigating creative writing in a large-scale experiment. AI-assisted texts were more fluent but less original, raising conceptual questions about evaluating creativity under technological influence. Whereas Bulante et al. foreground linguistic quality, Niloy et al. highlight trade-offs between stylistic gains and expressive diversity, thereby introducing an ethical–technological dimension to the necessary evaluation criteria. Führer and Gerjets

(2024) extend this by emphasizing that empirical studies must also account for aesthetic, emotional, and creative dimensions of writing.

From a didactic-methodological standpoint, Rezat and Schindler (2025) argue that empirical research on AI-assisted writing cannot be separated from instructional aims and assessment logics. Their analyses align with approaches that examine the learner perspective, reflective competence, and processual writing development.

A particularly relevant extension of this perspective is found in Graham (2024), who argues that writing remains central in the age of AI because it supports thinking, understanding, and active engagement with content. Accordingly, AI should be integrated in ways that stimulate rather than replace cognitive processes. This theoretical stance frames the design of Goller's (2026) project *Rethinking Writing*, which explicitly investigates how different forms of chatbot interaction can promote writing awareness and argumentative thinking in secondary education.

Goller compares two instructional AI models: a question-guided chatbot that provides prompts and follow-up questions to stimulate deeper reasoning, and a dialogic chatbot that responds to student input without directing it. The study conceptualizes AI not as an optimization tool, but as a didactic catalyst for reflective writing processes. Methodologically, it employs a mixed-methods design: qualitative analyses of chat logs to trace cognitive and metacognitive movements along with quantitative analyses of revised texts focusing on coherence, argumentative structure, and linguistic complexity. Pilot findings (2023) indicate that the question-guided chatbot in particular supports deeper reflective engagement, thereby exemplifying how AI can foster rather than automate thinking. In contrast to product-oriented studies such as Bulante et al., Goller's work foregrounds processual and reflective dimensions, aligning closely with the mixed-methods requirement and ethical–technological considerations regarding the pedagogical purpose of AI use.

Taken together, these studies form an exemplary cross section of international research: They contrast product versus process orientation, short-term versus longitudinal designs, experimental versus practice-based approaches, while also addressing the cognitive, creative, and reflective dimensions of writing. Their methodological differences highlight the need for integrated mixed-methods approaches, sustained perspectives on long-term effects, and transparent ethical–technological reporting standards.

#### **4 Perspectives and outlook**

The studies reviewed here indicate that empirical research on AI-assisted writing is still in an early phase of methodological consolidation. The considerable heterogeneity of existing designs suggests that pragmatic, context-dependent approaches have so far prevailed over theoretically coherent methodological frameworks. At the same time, this diversity has produced several developments that are likely to shape the field in the coming years.

First, there is increasing evidence of a paradigm shift toward mixed-methods research that systematically combines quantitative output analyses with qualitative process and reflection data. Such triangulation is essential for generating valid insights into practices of human–machine interaction and for understanding how writing development unfolds in AI-supported settings.

Second, long-term effects are emerging as a key criterion for evaluating the educational value of AI support. Whereas many studies document short-term improvements or motivational gains, robust

empirical knowledge about delayed effects on self-regulation, transfer, and writing awareness remains limited. Longitudinal and follow-up designs are therefore central for future research.

Third, empirical investigations must systematically address ethical and technological conditions. Limited model transparency, algorithmic bias, and data protection constraints affect not only research practice but also the interpretability and comparability of empirical results. Ethical–technological parameters are thus methodologically relevant variables that must be documented explicitly to ensure validity and reproducibility.

Taken together, these developments indicate a conceptual shift away from asking whether AI promotes writing toward examining how, under what conditions, and with what long-term consequences AI systems influence writing processes and outcomes. Advancing empirical research on AI-assisted writing therefore requires an expanded methodological repertoire that integrates mixed-methods logic, long-term perspectives, and transparent ethical and technological reporting. Only through such reflective and context-sensitive approaches can research adequately capture the complexity of interactions between humans, AI systems, and educational environments and contribute to an internationally coherent methodological discourse.

## References

- Anjarani, S., Furqon, M., & Rakhmawati, F. (2024). *ChatGPT to foster students' engagement in writing class: An intervention study*. *Voices of English Language Education Society*, 8(3), 630–640. <https://doi.org/10.29408/veles.v8i3.27687>
- Bulante, K. M., Morris, D. Q., Jr., Orteza, M. A. T., Balagan, A. M., Taloza, K., & Torres, M. A. (2024). AI-assisted and non-AI-assisted writing performances of Grade 10 students. *Azal Arts and Humanities*, 1(1), 65–94. <https://doi.org/10.55687/aah.v1i1.74>
- Führer, C., & Gerjets, P. (2024). How to understand and write literature with AI? Potenziale und Risiken von KI-Tools für literarisches Lesen und Schreiben. *Medien im Deutschunterricht*, 6(1). <https://doi.org/10.18716/OJS/MIDU/2024.1.3>
- Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. *Scientific Reports*, 13, Article 12018, 1–11. <https://doi.org/10.1038/s41598-023-42227-6>
- Gethmann, C. F., Buxmann, P., Distelrath, J., Humm, B. G., Lingner, S., Nitsch, V., Schmidt, J. C., & Spiecker genannt Döhmann, I. (Eds.). (2022). *Künstliche Intelligenz in der Forschung: Neue Möglichkeiten und Herausforderungen für die Wissenschaft* (Vol. 48). Springer VS. <https://doi.org/10.1007/978-3-662-63449-3>
- Goller, G. (2026). *Schreiben neu denken? Zum Einfluss dialogischer KI-Interaktionen auf das schriftliche Argumentieren von Lernenden der Sekundarstufe I* (Doctoral dissertation in preparation). Department of German Language and Literature, Karlsruhe University of Education.
- Graham, S. (2024). Why should we teach writing in the age of artificial intelligence? In *Inspired Ideas*. McGraw Hill. <https://medium.com/inspired-ideas-prek-12/why-should-we-teach-writing-in-the-age-of-artificial-intelligence-c49c50300584>
- Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education.

*Learning and Individual Differences*, 103, Article 102274.

<https://doi.org/10.1016/j.lindif.2023.102274>

Lehnen, K., & Steinhoff, T. (2024). Digitales Lesen und Schreiben. In J. Androutsopoulos & F. Vogel (Eds.), *Handbuch Sprache und digitale Kommunikation* (pp. 527–546). De Gruyter.

<https://doi.org/10.1515/9783110744163-025>

Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2023). Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning*, 40(2), 919–930.

<https://doi.org/10.1111/jcal.12929>

Pissarek, M., & Wild, J. (2018). Prä-/Post-/Follow-up-Kontrollgruppendesign. In J. M. Boelmann (Ed.), *Empirische Forschung in der Deutschdidaktik* (Vol. 1, pp. 215–236). Schneider.

[https://phbl-opus.phlb.de/frontdoor/deliver/index/docId/590/file/Boelmann\\_Empirische+Forschung+in+der+Deutschdidaktik\\_Bd.1\\_Grundlagen.pdf](https://phbl-opus.phlb.de/frontdoor/deliver/index/docId/590/file/Boelmann_Empirische+Forschung+in+der+Deutschdidaktik_Bd.1_Grundlagen.pdf)

Rezat, S., & Schindler, K. (2025). KI-gestützte Textproduktion im Deutschunterricht. *Praxis Deutsch*, 52(311), 4–11.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14,

Article 1260843. <https://doi.org/10.3389/fpsyg.2023.1260843>

Authors' addresses:

Gloria Goller, Karlsruhe University of Education, Bismarckstraße 10, 76133 Karlsruhe, Germany  
[gloria.goller@gmail.com](mailto:gloria.goller@gmail.com)

Nadine Anskeit, Karlsruhe University of Education, Bismarckstraße 10, 76133 Karlsruhe, Germany  
[nadine.anskeit@ph-karlsruhe.de](mailto:nadine.anskeit@ph-karlsruhe.de)

Daniela Matz & Carolin Führer

## Methodological approaches to research on writing processes in the context of AI

### The example of creative and literary writing

#### 1 Introduction

Under the conditions of generative AI, writing processes have become fundamentally invisible. Unlike traditional writing in which teachers can observe drafting stages, revision marks, and visible struggle, AI-mediated writing collapses the visible evidence of cognitive work. This poses a crucial educational and empirical challenge: Educators and researchers can no longer assume that a finished text reflects the student's actual intellectual labor (Hau, 2025). This is precisely what makes writing-process research so essential, because it makes the invisible visible again. In empirical writing research, the reconstruction of the writing processes themselves is underrepresented compared to the evaluation of the (learners') writing products. In contrast, recent empirical writing-process research (Singh et al., 2023) reveals the need to better understand the relationship between the processes and the products of writing. Especially within the context of AI research, the role of AI in learners' writing development has not yet been studied sufficiently and therefore cannot be assessed accurately. Currently, international writing research is trying to design different AI-support approaches to promote writers' sense of agency and depth of knowledge transformation (Siddiqui et al., 2025). However, in our view, it is imperative to first gain a more profound comprehension of the impact of this technology on the act of writing itself. This understanding is paramount for the establishment of appropriate and effective writing-process support within the context of future writing instruction in educational institutions.

The present article discusses methodological issues in L1 writing-process research by referring to two selected studies. These focus on reconstructing individual (LiSKIDS) and collaborative (WR-AI-TING) writing processes while integrating AI. The former project, LiSKIDS, is funded by the federal state of Baden-Württemberg and explores literary writing.<sup>1</sup> The latter, WR-AI-TING, is funded by the German Federal Ministry of Research, Technology and Space and focuses on creative writing with AI.

After briefly introducing the research questions in each study, the methodological discussion compares and contrasts the two studies under the following rubrics: research questions (2), the writing phases (3), the writing environment and the form of interaction (4), the experimental setup (5), the analysis

<sup>1</sup> "Literary writing" is understood as the process of composing fictional, poetic, or otherwise aesthetically shaped texts. In classroom settings, it is supposed to foster students' awareness of narrative structures, stylistic devices, and the artistic potential of language (Abraham et al., 2023).



and interpretation of the collected data (6), and finally the conclusions (7). The purpose of this comparative presentation is to highlight the methodological issues involved in qualitative research processes in the field of writing.

## 2 Research questions

The first LiSKIDS study (LiSKIDS 01) focuses on the evaluative processes of writers during AI-assisted literary writing and on the influence of these evaluations on the composition. To address the research questions, the aim is to get as close as possible to the cognitive and emotional processes—that is, to examine them in the very moment of their emergence. LiSKIDS aims at answering the following research questions:

1. How can literary writing processes with and without artificial intelligence be modeled?
2. What influence do literary evaluations exert on the direction and development of writing processes in both settings?
3. Which narrative-related normative values shape the writing processes in both settings?

The central research question guiding WR-AI-TING concerns the manner in which students revise creative texts in co-constructive writing processes both with and without the utilization of AI. The following subquestions are posed:

1. What (peer) interaction practices shape the writing processes with and without AI?
2. What orientational frameworks for creativity shape the writing processes with and without AI?
3. What metareflections on writing become visible in AI-supported and non-AI-supported collaborative settings?

## 3 Writing phases

Human–machine collaborative writing currently dominates the discussion around meaningful AI applications (Steinhoff, 2025). However, outsourcing too much work and learning to AI systems raises the risk of deskilling (Ferdman, 2025; Führer & Gerjets, 2024). It is particularly the early stages of writing that may be impacted by an intense collaboration with AI, because these stages are characterized by openness: Writers must gather ideas and come up with initial formulations. In order to better understand these dynamics that arise from collaboration with AI in relation to storytelling, LiSKIDS 01 focuses on an early phase of the writing process. It aims to gain a better understanding of the specific effects of human–machine collaboration on storytelling, especially among schoolchildren. What narrative patterns are implemented in such collaborative processes, with which storytelling functions are they associated, and how does human–machine collaboration change the writer’s relationship to storytelling and to their products?

In contrast, WR-AI-TING focuses on later stages of the writing process. Its main emphasis is on revision in creative writing. According to Hayes (2012, p. 376), revisions should not be understood as separate writing processes, but rather as a writing activity that focuses on the writing goal. Here, the function of writing processes as problem-solving processes becomes particularly clear: Revisions are initiated by identifying problems in the existing draft text and then replacing or modifying those text passages that are found to be problematic (Hayes, 2012, p. 376). The reader’s perspective comes to the fore, promoting both text awareness and abstract thinking; and this, in turn, simplifies the formulation of

writing goals. Because inconsistencies in one's own texts are very difficult to detect, it is advisable to involve readers who did not write the text in the revision process (Philipp, 2015, p. 50). In this context, passively accepting AI suggestions without critical evaluation would pose a key risk to the revision process. The authority of AI feedback might tempt students to stop questioning whether the suggestions make sense for their text or whether, conversely, they undermine their original intention. Hence, it is necessary to empower students to evaluate AI results independently. Focusing on learners' discussions about AI feedback in empirical research will help to develop appropriate didactic interventions to counter this temptation, because the metareflection on AI feedback is the decisive learning moment.

#### 4 Writing environment: Interactions and tasks

Writing tasks and the social setting in which they are carried out are central to writing processes. What dynamics shape writing with peers, and how do such environments differ from working alone? What opportunities do learners have to distance themselves from AI output in each of these two settings?

LiSKIDS 01 examines the writing processes of 16 learners in the 9th and 10th grades of intermediate and higher secondary schools (8 from each school type). The writing task involved selecting one of three optional beginnings for short stories that the learners were then asked to continue. Half of the students had the opportunity to use AI and were given general instructions on its use. All students had 50 min to complete the writing assignment. Revisions were not an explicit part of the assignment, but students could make them independently during the writing process. Students worked alone, which is the case in many learning situations, without a teacher, tutor (human or machine), or other students.

The WR-AI-TING study explored the writing discourse between secondary school learners ( $n = 118$ ) who wrote creative texts together in groups with and without AI support. As a first step, learners took part in a brief workshop on Franz Kafka's writing habits, for example, on the characteristically extensive revision processes he performed on his work. Next, they watched a custom-made instructional video about creative writing with AI. In this video, the researchers used thinking aloud to model how AI can be integrated into creative writing processes. Following this, the students had 20 min to write a creative text continuing the first sentence of Kafka's *The Trial*. Finally, the learners were asked to decide which of the two individually created drafts they found to be the more creative, and then to revise it jointly. In addition, the students received structuring tips for revision based on Torrance's (1974) psychological concept of creativity together with more recent, well-regarded models (Sternberg, 2003; Sternberg & Lubart, 1995). These guidelines focused on four key dimensions: fluency (i.e., ability to generate many ideas), flexibility (i.e., ability to produce different ideas), originality (i.e., ability to produce unique and statistically rarer ideas), and elaboration (i.e., ability to elaborate on ideas and develop them in detail). Students were free to decide whether to incorporate these tips into their revision or to ignore them.

#### 5 Experimental setup

Both studies used a writing tool developed specifically for research purposes in collaboration with the Leibniz-Institut für Wissensmedien (IWM) in Tübingen. This AI-integrated tool uses ChatGPT (version 4.0) as its underlying language model and embeds the chat interface directly into the writing environment. Whereas the students' writing conversations (WR-AI-TING) or their think-aloud verbalizations (LiSKIDS 01) were captured via screen and audio recording, their writing activities and chat protocols

were logged automatically. The students worked on 12.9-inch tablets equipped with external keyboards.

**Figure 1**

*IWM Writing Tool With ChatGPT Integration (Version 4.0)*

The primary data source in LiSKIDS 01 consists of think-aloud protocols that capture participants' verbalizations of working memory contents while performing a primary task (Stark, 2010, pp. 58–59). Think-aloud data are often collected for research on social practices (e.g., Noushad et al., 2024), but have played only a minor role in research on writing processes since the work of Hayes and Flower (1980a, 1980b). Criticism of this method includes the potential for reactivity—that is, the effect verbalizing thoughts may have on the processing of the primary task (see Bräuer & Kubik, 2024, p. 301, for a discussion)—as well as the risk of cognitive overload (Kellogg, 2001; Russo et al., 1989). However, researchers disagree on whether the phenomenon of reactivity actually exists. A meta-analysis of 94 studies found that concurrent thinking aloud does not affect the underlying thought process itself (Fox et al., 2011), thereby confirming initial assumptions by Ericsson and Simon (1984) about Level-1 (inner speech) and Level-2 verbalizations (putting nonverbal thoughts into words). LiSKIDS-01 addressed this uncertain situation with the following strategies:

- (a) Preceding data collection, participants received individualized training to familiarize themselves with both the task of thinking aloud and the writing tool. This included viewing a custom-made video demonstrating thinking aloud on a related task and practicing thinking aloud on a similar writing task. Individual training was essential to prevent imitation of peer

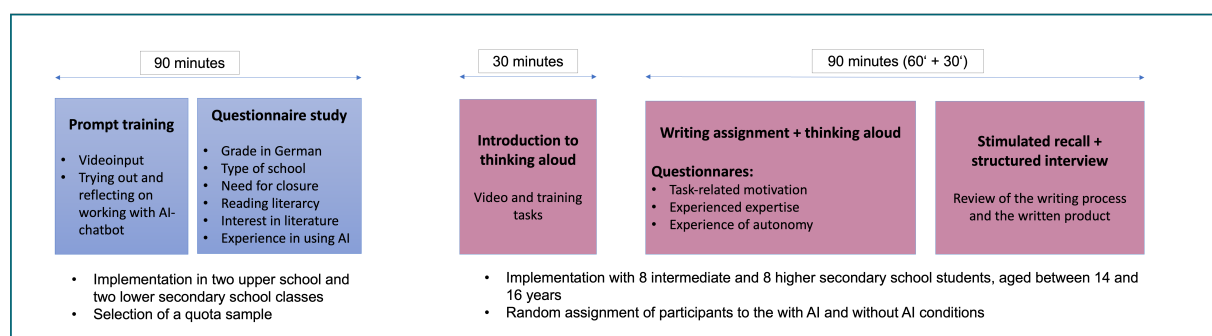
strategies and to enhance participants' confidence in verbalizing their thoughts (Aguado et al., 2013).

- (b) To avoid any additional strain on their cognitive resources, the instructions contained no additional demands (Fox et al., 2011, p. 319).
- (c) In addition to concurrent thinking aloud, subsequent verbal thinking in combination with interrupting the primary writing task was also used when this supported the test subjects, and stimulated recall data (Messmer, 2015) were also collected.
- (d) Although meta-analysis revealed no reactivity for Level-1 and Level-2 verbalizations, it did indicate that metareflective writing strategies and explanations addressed to others could significantly alter results (Fox et al., 2011, p. 333). Therefore, introspection, prompts, or questions that incite test subjects to explain their actions were avoided (Dannecker, 2018; Fox et al., 2011).
- (e) It cannot be ruled out that cognitive and emotional processes can be attributed to the study situation. This aspect of the data must be considered when categorizing and interpreting the results.

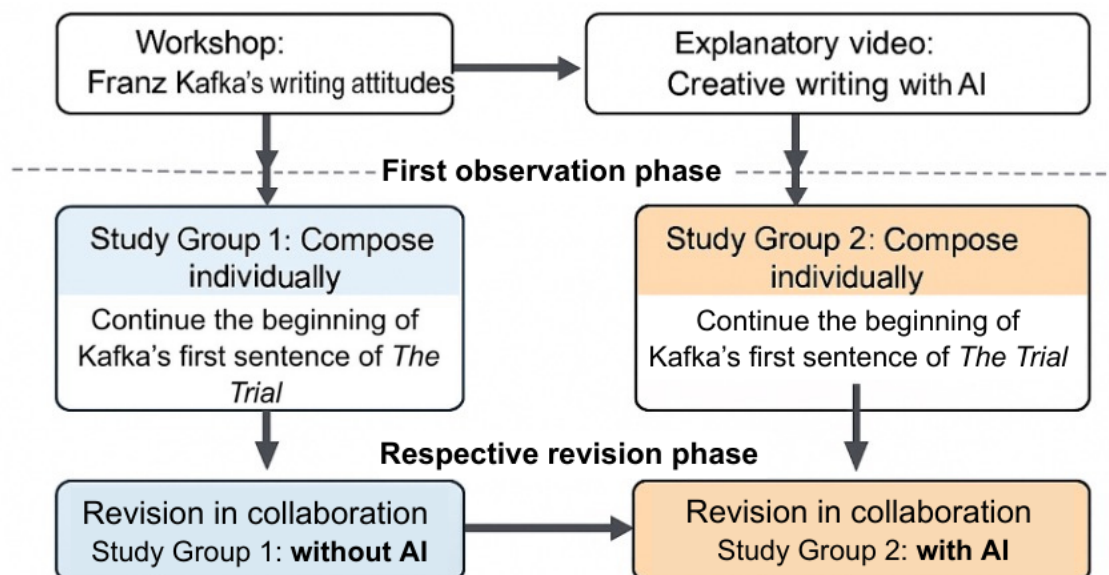
Working with think-aloud data requires further discussion, particularly regarding the epistemological stance of the researchers. Even within the field of cognitive research, it is necessary to consider, for example, whether researchers adhere to the paradigm of traditional information-processing models or to more recent approaches such as the concept of embodied cognition that place greater emphasis on “emotional, social, and institutional dimensions of human information processing” (Schnell, 2013, p. 96, translation D.M.).

**Figure 2**

*Study Design of LiSKIDS 01*



In WR-AI-TING, the writing groups used tablets either with or without access to the AI tool. Students in the AI group were free to decide how and when to use the AI tool. Students did not wear headsets because this might have created a feeling of distance and thus distorted results. Several pretests had confirmed that the audio quality was good enough to enable transcription of the revision discussions. The aim was to interfere as little as possible in the students' conversations and to ensure that they could hear each other. Through exchanging their own text drafts, the reflective writing process became authentically visible.

**Figure 3***Study Design of WR-AI-TING*

## 6 Data analysis and interpretation

The LiSKIDS 01 data are evaluated using the paradigm of constructive and reflective grounded theory (Breuer et al., 2019; Charmaz, 2025). The general aim of grounded theory is to establish new theories in a thematic field (Glaser & Strauss, 1967). Therefore, the goal is to expand existing knowledge through hermeneutic methods. This presents researchers with the challenge of deciding how to engage in inductive interpretation processing of the collected data while also applying existing theoretical concepts in a specific research process (Schatzman, 1991). Such research processes should be designed strategically and transparently. Moreover, own perspectives can be expanded through working in interpretative groups. In LiSKIDS 01, the conditions of data collection and evaluation are taken into account, because they constitute a social constellation: (a) First, the situation of data collection reflects a specific social situation constituted by the writing task, the collaboration of an individual with an AI, and the collection situation. (b) Second, the interpretative standpoint of the researchers has to be considered. This does not necessarily have to be understood as a deficit, but can also bear potential as long as the social situatedness of the data and of their interpretation are always taken into account (Bräuer & Kubick, 2024, p. 302). This refers to the important epistemological assumption of LiSKIDS that “social reality is multiple, processual, and constructed” (Charmaz, 2025, p. 12). The aim of the interpretative process is therefore to understand the role and nature of the evaluations that determine the choice of content and formal text design options in early stages of the writing process under different social and communicative conditions. Are these, for example, emotional reactions, intuitions, aesthetic experiences, or specific text norms? Further data collection will be needed to explore literary writing processes (with AI) as a cultural practice (Abraham et al., 2023).

For WR-AI-TING, chat interactions between AI and peers are evaluated by using the documentary method to reconstruct the learners’ conjunctive orientational frameworks (Bohnsack et al., 2013). As

a qualitative method for reconstructing implicit knowledge and collective frames of reference (Bohnsack et al., 2013, p. 12), this approach captures not only the explicit content of the conversations but also the underlying orientations. The underlying assumption is that the orientational frameworks of a group can be reconstructed by analyzing the formal structure of a conversation. It is assumed that this can be documented in the form and manner of an interaction (Asbrand & Martens, 2018, p. 55). Analyses based on the documentary method distinguish two types of knowledge: communicative and conjunctive. The first step is to identify communicative knowledge. This is available to the speakers and is made explicit. Conjunctive knowledge lies one level below this communicative knowledge, and has to be inferred from it (Asbrand & Martens, 2018, p. 50). It is based on conjunctive spaces of experience—that is, people’s shared a priori spaces of experience (Przyborski, 2004, p. 23). The purpose of interpreting data with the documentary method is to reveal the conjunctive understanding of writing processes and especially of revisions—something that is not explicitly apparent from the conversation. The aim is to evaluate the writers’ implicit, collectively shared knowledge, structures, and agencies. Hence, the focus of the data analysis is on communicative and conjunctive knowledge (and the gaps between the two) about creativity, peer cooperation, and interaction with AI as well as metareflections on writing (in terms of ownership, authorship, text production, literariness, etc.).

## 7 Conclusion

Reconstructive approaches to teaching and learning research, as exemplified by LiSKIDS and WR-AI-TING, provide significant methodological advantages for writing-process research. They enable the comprehensive and nuanced examination of writing processes as situated, social, and meaning-laden events within authentic classroom or learning contexts. Intermediate stages of writing processes, detours, and breaks can be captured accurately and almost authentically (Pelchat, 2022).

The primary objective of LiSKIDS 01 is to formulate theoretical assumptions about the unique features of literary writing processes (with and without AI collaboration) extending beyond existing process models. This encompasses aspects such as idea generation, decision making related to narrative text worlds and their concrete elaboration (such as choice of narrative situation, characters, plot, setting, and time), as well as the negotiation of creative options. The methodological design facilitates the collection of data that at least partially capture the writing process. Alongside think-aloud protocols, this includes drafts and chat transcripts, all of which must be systematically correlated during analysis. Both explicit and implicit literary evaluations can thus be reconstructed by comparing working sequences, thereby enabling the inductive development of theory grounded in the data.

Microprocesses of peer interaction in collaborative writing activities, as investigated in the WR-AI-TING study, can be analyzed by looking at the meaning structures of the actors. The documentary method also takes into account how materialities (digital writing surfaces) and medialities (with or without AI) change and structure writing processes. Through case-specific reconstructions and the comparison of several cases, both individual characteristics of writing processes and typical patterns and structures in collaborative writing settings can be identified. Video-based observations of screen activity and verbal discourse enable researchers to analyze revised text passages as well as the conversation sequences, gestures, and referential practices through which students negotiate meaning in texts. In this way reconstructive classroom research provides systematic access to precisely this link between interaction and text development.

**Table 1***Overview*

	<b>LiSKIDS</b>	<b>WR-AI-TING</b>
Test subjects	Aged 14–16 years Intermediate and upper secondary school students	Aged 17–19 years Upper secondary school students
Number of test subjects	16	118
Survey situation	Browser-based online writing tool with ChatGPT integration (Version 4.0)	
	Concurrent and retrospective thinking aloud	Writing discourse
Data analyses	Grounded theory	Documentary method
Writing environment	Individual	Collaborative
	Early writing stage	Revision
conditions	Writing with AI and writing without AI	

## References

- Abraham, U., Bräuer, C., & von Brand, T. (Eds.). (2023). *Literarisches Schreiben als kulturelle Praxis: Deutschdidaktische, schreibpädagogische und medienkulturelle Impulse für den Unterricht*. Klett Kallmeyer.
- Aguado, K., Heine, L., & Schramm, K. (Eds.). (2013). *Introspektive Verfahren und qualitative Inhaltsanalyse in der Fremdsprachenforschung*. Peter Lang. <https://doi.org/10.3726/978-3-653-02548-4>
- Asbrand, B., & Martens, M. (2018). *Dokumentarische Unterrichtsforschung*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-10892-2>
- Bohnsack, R., Nentwig-Gesemann, I., & Nohl, A.-M. (2013). *Die dokumentarische Methode und ihre Forschungspraxis: Grundlagen qualitativer Sozialforschung* (3rd ed.). VS Verlag für Sozialwissenschaften.
- Bräuer, C., & Kubik, S. (2024). Lautes Denken als Verbalisationsverfahren literarischer Rezeption in soziokultureller Perspektive – neue Grenzen, neue Chancen. In M.-O. Carl, M. Jörgens, & T. Schulze (Eds.), *Literarische Texte lesen – Texte literarisch lesen* (pp. 281–310). Springer. [https://doi.org/10.1007/978-3-662-67816-9\\_17](https://doi.org/10.1007/978-3-662-67816-9_17)
- Breuer, F., Muckel, P., & Dieris, B. (2019). *Reflexive grounded theory: Eine Einführung für die Forschungspraxis* (4th ed.). Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-22219-2>
- Charmaz, K. (2025). *Constructing grounded theory*. Sage.
- Dannecker, W. (2018). Lautes Denken. In J. M. Boelmann (Ed.), *Empirische Forschung in der Deutschdidaktik* (Vol. 2, pp. 131–137). Pädagogische Hochschule Ludwigsburg.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. MIT Press.



- Ferdman, A. (2025). AI deskilling is a structural problem. *AI & Society*.  
<https://doi.org/10.1007/s00146-025-02686-z>
- Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, 137(2), 316–344. <https://doi.org/10.1037/a0021663>
- Führer, C., & Gerjets, P. (2024). How to understand and write literature with AI? Potenziale und Risiken von KI-Tools für literarisches Lesen und Schreiben. *Medien im Deutschunterricht*, 6(1), 1–18. <https://doi.org/10.18716/OJS/MIDU/2024.1.3>
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory*. Aldine.
- Hau, M. F. (2025). Writing with machines? Reconceptualizing student work in the age of AI. *Frontiers in Communication*, 10. <https://doi.org/10.3389/fcomm.2025.1598988>
- Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29(3), 369–388. <https://doi.org/10.1177/0741088312451260>
- Hayes, J. R., & Flower, L. S. (1980a). Identifying the organization of writing processes. In L. W. Gregg (Ed.), *Cognitive processes in writing* (pp. 3–30). Lawrence Erlbaum Associates.
- Hayes, J. R., & Flower, L. S. (1980b). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg (Ed.), *Cognitive processes in writing* (pp. 31–50). Lawrence Erlbaum Associates.
- Kellogg, R. T. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, 114(2), 175–191. <https://doi.org/10.2307/1423513>
- Messmer, R. (2015). Stimulated recall als fokussierter Zugang zu Handlungs- und Denkprozessen von Lehrpersonen. *Forum Qualitative Sozialforschung*, 16(1).
- Noushad, B., Van Gerven, P. W. M., & De Bruin, A. B. H. (2024). Twelve tips for applying the think-aloud method to capture cognitive processes. *Medical Teacher*, 46(7), 892–897. <https://doi.org/10.1080/0142159X.2023.2289847>
- Pelchat, L. (2022). *Kollaboratives Schreiben: Eine rekonstruktive Studie zu Problemlösepraktiken im Französischunterricht*. Narr Francke Attempto.
- Philipp, M. (2015). *Schreibkompetenz: Komponenten, Sozialisation und Förderung*. A. Francke Verlag.
- Przyborski, A. (2004). *Gesprächsanalyse und dokumentarische Methode: Qualitative Auswertung von Gesprächen, Gruppendiskussionen und anderen Diskursen*. VS Verlag für Sozialwissenschaften. <https://doi.org/10.1007/978-3-531-90347-7>
- Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, 17(6), 759–769. <https://doi.org/10.3758/BF03202637>
- Schatzman, L. (1991). Dimensional analysis: Notes on an alternative approach to the grounding of theory in qualitative research. In D. R. Maines (Ed.), *Social organization and social process* (pp. 303–314). Aldine de Gruyter.
- Schnell, A. K. (2013). Lautes Denken im Licht der embodied cognition: Neue Grenzen, neue Möglichkeiten in der Schreibprozessforschung? In K. Aguado, L. Heine, & K. Schramm (Eds.), *Introspektive Verfahren und qualitative Inhaltsanalyse in der Fremdsprachenforschung* (pp. 92–115). Peter Lang.
- Siddiqui, M. N., Feliciano, V., Pea, R., & Subramonyam, H. (2025). AI in the writing process: How purposeful AI support fosters student writing. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S.

- Isotani (Eds.), *Artificial intelligence in education* (Lecture Notes in Computer Science, Vol. 15880). Springer. [https://doi.org/10.1007/978-3-031-98459-4\\_14](https://doi.org/10.1007/978-3-031-98459-4_14)
- Singh, N., Bernal, G., Savchenko, D., & Glassman, E. L. (2023). Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer–Human Interaction*, 30(5), 1–57. <https://doi.org/10.1145/3511599>
- Stark, T. (2010). Lautes Denken in der Leseprozessforschung: Kritischer Bericht über eine Erhebungsmethode. *Didaktik Deutsch*, 29, 58–83.
- Steinhoff, T. (2025). Künstliche Intelligenz als Ghostwriter, writing tutor und writing partner: Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel des Schreibens mit ChatGPT in einer 8. Klasse. In C. Albrecht, J. Brüggemann, T. Kretschmann, & C. Meier (Eds.), *Personale und funktionale Bildung im Deutschunterricht: Theoretische, empirische und praxisbezogene Perspektiven* (pp. 219–244). Metzler.
- Sternberg, R. J. (Ed.). (2003). *Wisdom, intelligence, and creativity synthesized*. Cambridge University Press.
- Sternberg, R. J., & Lubart, T. (1995). An investment approach to creativity: Theory and data. In S. M. Smith, T. B. Ward, & R. A. Finke (Eds.), *The creative cognition approach* (pp. 262–293). MIT Press. <https://doi.org/10.7551/mitpress/2205.003.0017>
- Torrance, E. P. (1974). *Torrance tests of creative thinking: Verbal tests, forms A and B; figural tests, forms A and B; norms-technical manual*. Personnel Press.

Authors' addresses:

Daniela Matz, University of Tübingen, Deutsches Seminar, Wilhelmstr. 50, 72074 Tübingen, Germany  
[daniela.matz@uni-tuebingen.de](mailto:daniela.matz@uni-tuebingen.de)

Carolin Führer, University of Tübingen, Deutsches Seminar, Wilhelmstr. 50, 72074 Tübingen, Germany  
[carolin.fuehrer@uni-tuebingen.de](mailto:carolin.fuehrer@uni-tuebingen.de)