Paradigm
reference-global.com

RISTAL 8/2025

Research in Subject-matter Teaching and Learning

Special Edition:
Reading, writing and interpreting texts in digital environments

Edited by
Stefan D. Keller & Volker Frederking

Citation:
Frederking, V. (2025). Reliability of AI in the domain of literary literacy and literary text interpretation: an empirical study. RISTAL, 8*(2)*, 70-87.

DOI: https://doi.org/10.2478/ristal-2025-0007

ISSN 1863-0502

Paradigm
reference-global.com

RISTAL 8(2)/2025, 70–87.

Reliability of AI in the domain of literary literacy and literary text interpretation: an empirical study

*Volker Frederking**

Abstract

The article presents the results of an experimental study in the field of L1 education that used literary literacy and the interpretation of literary texts as examples to examine the reliability of generative AI in helping learners in literature classes to overcome complex challenges. Due to the ambiguity of literary texts, the reliability of AI faces particular challenges here. Based on an empirically verified model of literary literacy, two chatbots considered particularly powerful—ChatGPT-5 from Open AI and Claude Sonnet-4.5 from Anthropic—were systematically tested. After explaining the study design, the initial results of the experimental study are presented and discussed.

Keywords: AI, reliability of AI, literary literacy, ChatGPT-5, Claude Sonnet-4.5

***Volker Frederking**
University of Erlangen-Nuernberg, Department Subject Didactics, Chair for German Language and Literature Education Research, Regensburger Straße 160, 90478 Nuernberg.
E-Mail: volker.frederking@fau.de

Author Note
This research received funding from Federal Ministry for Education, Family Affairs, Senior Citizens, Women and Youth (former Federal Ministry of Education and Research) in the project “Digital Sovereignty as Goal of Forward-Thinking Professional Development for Teachers for Languages, Social Sciences, and Economics in the Digital World (DiSo-SGW)”.

Correspondence concerning this article should be addressed to Prof. Dr. Volker Frederking, Universität Erlangen-Nürnberg, Department Fachdidaktiken, Lehrstuhl für Didaktik der deutschen Sprache und Literatur, Regensburger Straße 160, 90478 Nürnberg

Introduction

Since ChatGPT was presented as a technological innovation on November 30, 2022, generative AI has permeated all areas of social, political, and scientific life. This also applies to the field of education. The US Department of Education's “Artificial Intelligence (AI) Guidance” (2025) and the “EU Artificial Intelligence Act” (2025) have established a framework for using AI in educational processes. The KMK, the Standing Conference of the Ministers of Education and Cultural Affairs of the German Federal States, has announced concrete steps for the research-led use of AI in mathematics, German, and foreign languages. However, its focus is on basic skills (KMK, 2024, p. 5). While this is a pragmatic approach, It does not fully account for the opportunities and limitations of generative AI's capabilities. After all, ChatGPT and other chatbots are already being used

across the entire spectrum of school teaching, i.e., not only in basic but also in more complex application contexts. Therefore empirical evidence is urgently needed here, too.

This is particularly true with regard to the reliability issue examined in this article. It presents findings from the LL-AI-3 study, an experimental study in the field of L1 education (first language education), which used the example of literary literacy and literary text interpretation to investigate the reliability of generative AI in supporting learners in literature classes to overcome complex challenges. The first section examines digital writing, reading and communication in, with, and by text-generating AI from media-cultural and computational linguistic perspectives in order to highlight theoretical foundations for using generative AI in schools (1). Against this backdrop, the question of how teachers and learners can use generative AI effectively and reflectively in language and literature teaching comes into focus. A research overview highlights opportunities and challenges. The question of AI reliability plays an important role here. This is particularly pertinent in the context of literary learning. Due to the ambiguity of literary texts, AI reliability faces particular challenges here (2). In this context, research gaps become apparent, which the presented study can help to close. For the first time, two chatbots currently considered particularly powerful, Open AI's 'ChatGPT-5' and Anthropic's 'Claude Sonnet-4.5', were investigated in terms of their performance in the field of literary literacy. First, the experimental study's survey design is explained, then initial findings are presented and discussed (3). On this basis, the potential of generative AI as a tutorial system in the field of L1 didactics and subject-specific teaching and learning in general is reflected upon (4).

1 Digital writing, reading, and communicating in, with, and by text-generating AI

Digital writing, reading, and communication in, with, and by text-generating AI are part of media culture history and have foundations in computational linguistics. Both aspects will be examined below as they contribute to a better understanding of the opportunities and limitations of generative AI in teaching and learning at school.

1.1 Generative AI from a media-cultural and linguistic perspective

It is a remarkable fact that digital writing and reading emerged from efforts to develop a program-controlled calculating machine (Hiebel et al., 1998, p. 227). The effect that computers have also established themselves as writing and information media is thanks to Alan Turing's (1937) ingenious insight that the computer is not only a calculating machine at its core, but a universal simulation machine which is able to imitate all symbol-mediated interactions by 'semiotically dissolving and reassembling numbers, writing, and images, but also sound and haptics' (Gramelsberger, 2023, p. 126). In this way, the 'digital reconstruction of the world' (ibid., pp. 155–156) is possible – based on computer language codes, the reception and production of which is part of the basic inventory of the 'cultural technique of the digital' (ibid., p. 126).

Generative AI is one of these new cultural techniques of the digital. It complements the existing spectrum of digital reconstruction of the world by imitating written and oral interaction, creating a new quality of human-machine interaction. In this interaction,

generative AI is not only a tool, but also a quasi-autonomous actor of digital writing, reading, and communication processes. In other words, digital writing, digital reading, and digital communication now also take place in, with, and by generative AI.

The basis of human-machine interaction in, with, and by generative AI are, in computer linguistic terms, "prompts," i.e., commands from the computer to a user and from a user to the computer or AI. From a linguistic point of view, however, this is a communicative-dialogical simulation based on illocutionary speech acts (Austin 1962; Searle 1965; 1969). In linguistic theory, speech acts are considered the smallest complete units of human linguistic communication (Searle, 1969). A chatbot initiates interaction with a user by simulating an illocutionary speech act. For example: When accessing ChatGPT the question "How can I help you?" appears on the screen, along with the prompt "Ask any question" (https://chatgpt.com). Claude Sonnet-4.5 asks a similar question: "How can I help you today?" (https://claude.ai/new). Both speech acts used by the chatbots aim to elicit responses from human interaction partners in the form of illocutionary speech acts. However, this "dialogical" framing of human-machine interaction should not be misunderstood. A LLM model cannot think independently or act intentionally in the medium of language. Human-machine interaction is not real communication between two autonomous actors, but rather a simulation thereof. Generative AI merely mimics communication from a large set of programmed speech acts resulting from an intersection of plausible linguistic options. In the words of John Searle (1980, 427): "The formal symbol manipulations [of a computer] by themselves don't have any intentionality; they are quite meaningless [...]. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output."

1.2 Generative AI from a computational linguistic background

Although AI is capable of receiving and producing all facets of multimodal semiotics in principle, the associated processes are language-based. Regardless of whether a chatbot is asked to produce written, visual, auditory, or audiovisual signs or sign combinations in the form of an image, an audio text, or a video, this always takes place in the form of speech acts in natural language (questions, requests, commands etc.). This natural language, which appears on the surface level of human-machine interactions, is processed at the deep level of programming language in so-called 'tokens'. In computational linguistics, tokens refer to the smallest units with which AI models such as ChatGPT process text. ChatGPT 'understands' and 'generates' (Porschen, 2024) text at the token level, not at the word or letter level. However, the term 'understanding' is misleading or inaccurate in the context of LLMs. In fact, chatbots do not 'understand' what they write, read, or communicate (Krämer, 2024, p. 303). LLMs use tokens to extract 'meaning structures' or 'attributions of meaning' from questions, statements, texts, or prompts, and thus generate plausible-sounding statements based on statistical probability.

Functional chatbots existed long before the advent of ChatGPT, Claude, Gemini etc. However, early chatbots were not yet based on tokens. One famous example is 'ELIZA', a language model developed by Joseph Weizenbaum (1966) at MIT. While the thesaurus-based ELIZA model could only reproduce basic linguistic patterns unchanged, a Google

research group (Vaswani et al. 2017) achieved a technological breakthrough with ‘a new simple network architecture’ (ibid., p.1). The so-called transformer architecture enabled the variable use and generation of linguistic interaction modules. With “Duplex,” Google succeeded in creating an application in the form of a digital voice assistant for everyday verbal communication (Metz, 2018), while Open AI developed “Generative Pre-trained Transformers (GPT)” based on the transformer architecture using tokens. RLHF (= “reinforcement learning from human feedback”) played a decisive role in this process (Open AI, 2022). RLHF enabled human feedback on AI-generated responses to be used to optimize the model's performance. On this “communicative-interactive” computer linguistic basis, Open AI has been developing various GPT versions since 2018, before ChatGPT was presented to the global public on November 30, 2022. Since then, numerous other chatbot systems have emerged, such as ‘Claude’ from Anthropic, ‘Gemini’ from Google, ‘Perplexity’ from Perplexity AI.

2 Generative AI in the context of language and literature teaching and learning. A research overview

The media-cultural and computational linguistic background outlined in the previous section is important for understanding the opportunities and challenges of generative AI in the field of school teaching and learning. It forms the basis for reflective and responsible use. Education policymakers in Germany have responded relatively quickly to the opportunities and challenges presented by generative AI. The KMK, for example, has emphasized the need to embed skills in dealing with AI as an integral part of all three phases of teacher training and to teach and research pedagogical and didactic application scenarios for AI in subject teaching (KMK, 2024, p. 8). In doing so, the KMK highlights the use of generative AI as tutorial systems, for feedback, and as writing assistants to promote basic skills in German, mathematics, and foreign languages (ibid., p. 5). In its recommendations for the use of generative AI in schools, the SWK (=Standing Scientific Commission of the Conference of Ministers of Education and Cultural Affairs), has identified potential in particular with regard to text creation, internal and subject differentiation, adaptive learning, and individual feedback (SWK, 2024, pp. 10-12). In contrast to the KMK, however, the SWK recommends the use of generative AI in schools only for secondary levels I and II. In current research, these age groups are indeed a focus alongside academic use, as the following research overview shows. At the same time, gaps in research are becoming apparent.

2.1 Opportunities for language and literature teaching and learning through generative AI

In the field of first language education (L1) and second language education (L2), relevant research has been conducted on the use of AI applications in relation to the subjects of German and foreign languages mentioned by the KMK. The focus is on the importance of generative AI for English and German language learning and teaching, but with regard to higher levels of schooling (Lee, Jeon, Mckinley & Rose, 2025; Führer & Gerjets, 2024). Potential is seen, for example, in generative AI as a writing assistant, for digital writing processes (Steinhoff 2023), as a ghostwriter, writing tutor, and writing partner (Steinhoff, 2025, p. 85) or for AI feedback (Führer 2025), for “argumentative writing” (Su, Lin & Lai, 2023) or automated feedback (Fleckenstein, Liebenow & Meyer, 2023; Meyer et al., 2024;

Jansen et al., 2024). There is also relevant research on feedback on spelling, grammar, text coherence, argumentation structure, and content quality of texts (Bewersdorff et al., 2023; Fang et al., 2023). Insightful empirical research in the field of teacher professionalism has also been conducted e.g. with regard to the usage and beliefs of student teachers towards artificial intelligence in writing (Helm & Hesse, 2024; Hesse & Helm, 2025).

2.2 The problem of the unreliability of generative AI as a challenge for language and literature teaching and learning

Challenges for language and literature teaching arise from a problem for which the anthropomorphic and euphemistic term “hallucinations” has become established. When ChatGPT was released, OpenAI posted the following notice on its website: “ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers” (Altman, 2022a). Data scientist Teresa Kubacka already clarified on December 6, 2022, how justified these warnings, which are now displayed during every interaction with ‘ChatGPT’ or ‘Claude’, are: "Today I asked ChatGPT about the topic I wrote my PhD about. It produced reasonably sounding explanations and reasonably looking citations. So far so good – until I fact-checked the citations. And things got spooky when I asked about a physical phenomenon that doesn't exist." Similar warnings could be found early in medical and psychological literature (Emsley, 2023, p.2; Else, 2023, p. 423; Smith, 2023; Athaluri et al., 2023; Rawte, Sheth & Das, 2023; Bergener et al., 2023).

In the educational context, the SWK (2024, p. 13-14) drew attention to the problem of the unreliability of generative AI at an early stage. The problem has also been addressed in the research discourse on linguistic and literary teaching and learning (Frederking, 2023; Maiwald, 2023; Müller & Fürstenberg, 2023). However, systematic empirical research has been the exception so far. Examples include studies on improving the reliability and transparency of ChatGPT for educational question answering (Wu, Y. et al., 2023), on reading comprehension exercises generated by LLMs (Xiao et al., 2023), and on a Tutor.AI tool developed in teacher training in the subject of German (Bach et al., 2025).

Two research projects funded by the former Federal Ministry of Education and Research (BMBF) on the topic of digital sovereignty in L1 German language and literature teaching have also conducted more systematic empirical studies on the reliability of generative AI (Ascherl, 2025; Brüggemann et al., 2025a; 2025b; Frederking, 2025). These studies—LL-AI-1 and LL-A2-2—focused on testing AI in the domain of literary literacy, as literary texts pose particular challenges for comprehension due to their ambiguity (Eco, 1962; 1990) and self-referential language use. For this reason, they are a good indicator of the reliability of generative AI in a demanding learning area. The above-mentioned research has provided evidence that ChatGPT has its limitations when dealing with literary texts, which poses particular challenges for teacher training (Brüggemann et al., 2025a; 2025b). However, the findings are based on ChatGPT version 4o. Further research is therefore needed to clarify whether the results are also confirmed in tests with ChatGPT-5, the latest version of OpenAI. Furthermore, it needs to be investigated whether other generative AI models, such as ‘Claude Sonnet-4.5’ from Anthropic, lead to better results than ChatGPT version 4o. The LL-AI-3 study presented below should help to close these research gaps—with the knowledge that follow-up studies will be necessary as new and improved chatbots are introduced.

3. The LL-AI-3 experimental study

3.1 Research design and research questions

In the LL-AI-3 study, the technical accuracy, reliability, and dimensionality of the AI-generated interpretations produced by ChatGPT-5 were tested using test tasks for grades 8-10, which were also used in the LL-AI-2 study. The test tasks are based on the construct of 'literary literacy', which was theoretically modeled and empirically tested in studies on literary text comprehension (LUK) funded by the German Research Foundation (DFG) between 2007 and 2013 (Frederking et al., 2012; 2016; Meier et al., 2017). Five sub-dimensions of literary literacy were distinguished.

1. *Semantic comprehension* (= the ability to understand the content, meaning structures, and scope for interpretation of literary texts).
2. *Idiolectal comprehension* (= the ability to grasp the formal specifics of literary texts and their aesthetic function).
3. *Literary knowledge* (= the ability to apply background information to literary texts).
4. *Aesthetic awareness* (= the ability to recognize linguistic and stylistic features of literary texts).
5. *Understanding of emotions intended in literary texts* (= the ability to grasp the intended emotional effects of a literary text) (cf. Frederking, 2022a).

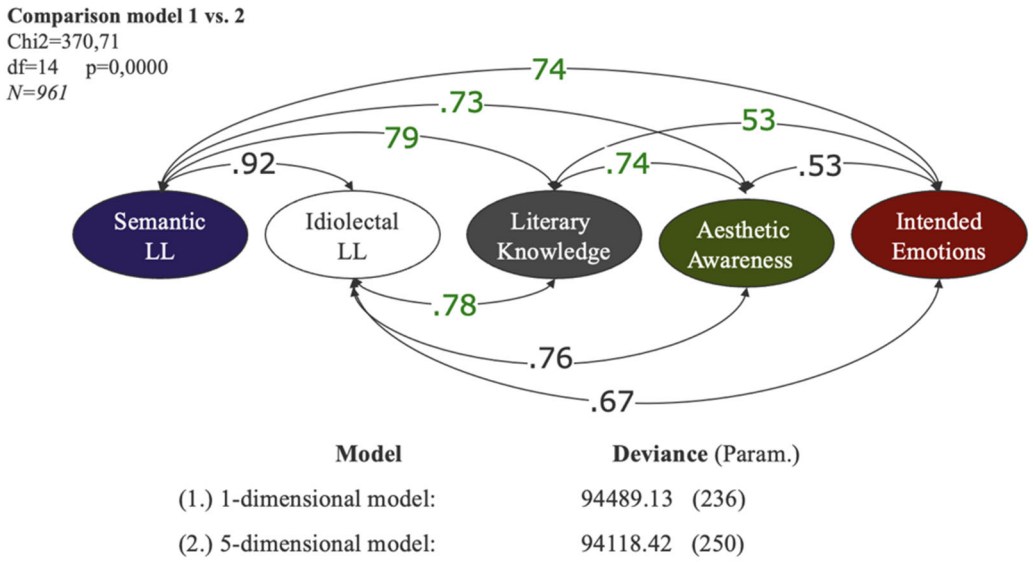

Fig. 1: Five Dimensions of Literary Literacy (Frederking et al., 2016)

All five sub-dimensions have been empirically proven to be clearly distinguishable (Fig. 1). A significant empirical separability had also emerged between literary literacy and reading literacy (Fig. 2).

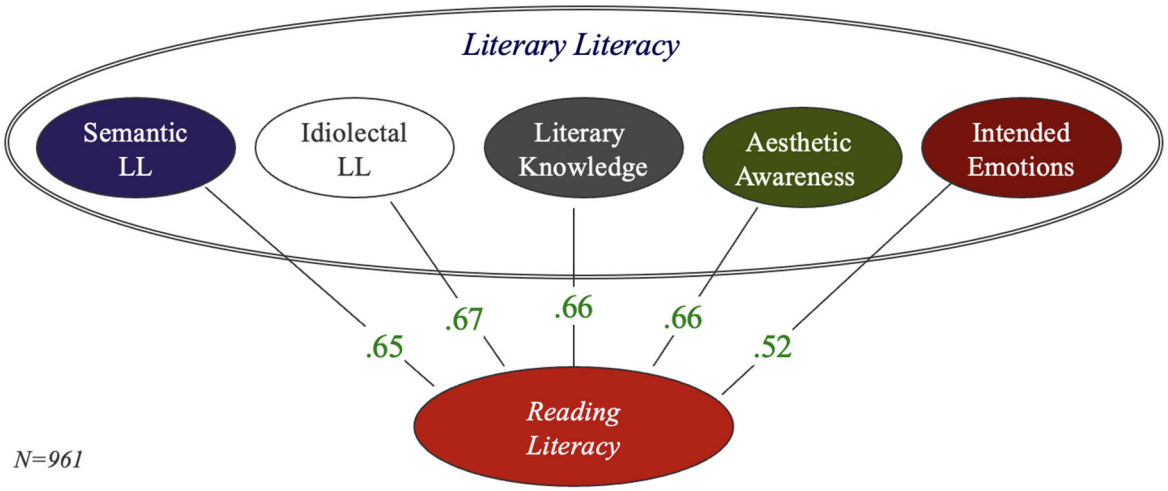

Fig. 2: Literary literacy und reading literacy (ibid.)

The LL-AI-3 study was based on this five-dimensional model of literary literacy. As in LL-AI-2, two lyrical texts and 82 empirically proven LUK test items and subitems (open, semi-open, MC, and FC) were used. The LUK test tasks provide reliable measurement instruments that have been systematically evaluated in three DFG projects in terms of their validity, reliability, and objectivity. As in LL-AI-2, the investigations in LL-AI-3 were conducted in four experimental runs, but with ChatGPT-5 and Claude Sonnet-4.5.

Experiment 1: ChatGPT-5 processed poem 1 and 51 items in five separate complete runs (= 5 x 51 solutions = 255 solutions)

Experiment 2: ChatGPT-5 processed poem 2 and 31 items in five separate complete runs (= 5 x 31 solutions = 155 solutions).

Experiment 3: Claude Sonnet-4.5 processed poem 1 and 51 items in five separate complete runs (= 5 x 51 solutions = 255 solutions).

Experiment 4: Claude Sonnet-4.5 processed poem 2 and 31 items in five separate complete runs (= 5 x 31 solutions = 155 solutions).

On this basis, both chatbots processed a total of 820 items across the two units. In each case, the literary stimulus text (poem 1 or 2) was first made available to the two chatbots by entering it in the dialog box. A note was then given that questions about the text would follow. The questions or items were entered one after the other and answered separately by the two chatbots. After completion, a completely new chat was started in each case to avoid interference, dependencies, or influences. The item solutions were evaluated by an expert using the coding grid successfully used in LUK research. Forced choice items were evaluated separately in each case.

Research questions

Five research questions formed the basis of the LL-AI-3 study:

Research question 1:	Are the solutions to test tasks on literary literacy generated by ChatGPT-5 and Claude Sonnet-4.5 correct?
Research question 2:	Are there empirically measurable differences in the quality of the solutions formulated by ChatGPT-5 and Claude Sonnet-4.5 to the test tasks for the five dimensions of literary literacy?
Research question 3:	Are the same solutions presented when the identical items are processed five times, or are there deviations?
Research question 4:	Are there any differences in the quality of the solutions formulated by ChatGPT-5 and Claude Sonnet-4.5 compared to ChatGPT-4o from the LL-AI-2 study?
Research question 5:	Are there differences in the quality of the solutions formulated by ChatGPT-5 and Claude Sonnet-4.5 in the five LUK dimensions compared to ChatGPT-4o?

3.2 Results

The data and results obtained in the LL-AI-3 study are presented and explained below in direct relation to the five research questions.

Results for research question 1

There are slight differences in the number of correct solutions. The total percentage of correctly solved items for the two poems was 88.9% in the experiments with ChatGPT-5 and 87.4% with Claude Sonnet-4.5 (see Fig. 3).

	ChatGPT 5		Claude Sonnet 4.5	
	Unit 1	*Unit 2*	*Unit 1*	*Unit 2*
	%	*%*	*%*	*%*
Semantic comprehension	77,27	100	67,27	100
	88,64		83,64	
Idiolectal comprehension	68	88	80	74
	78		77	
Literary Knowledge	100	85,71	100	100
	92,86		100	
Aesthetic Awareness	100	100	100	100
	100		100	
Intended Emotions	90	80	87,5	65
	85		76,25	
Summe	**88,9**		**87,4**	

Fig. 3: Results from ChatGPT-5 and Claude Sonnet-4.5 for the two LUK units

Results for research question 2

The data from the four experiments reveal similarities, but also differences in the quality of the solutions formulated by the two chatbots for the five dimensions of literary literacy. A overvieew table with an example item for each category is included in the appendix (see table 1 in the appendix).

1. For the items on *semantic comprehension*, 88.64% of the solutions formulated by ChatGPT-5 are correct, and 83.64% for Claude Sonnet-4.5. There are also clear differences with regard to the two units. While both chatbots achieve 100% correct solutions for Unit 2, ChatGPT-5 only achieves 77.27% for Unit 1 and Claude Sonnet-4.5 only 67.27%.

2. In terms of *idiolectal comprehension*, both chatbots are close together in their total scores. ChatGPT-5 achieves 78%, Claude Sonnet-4.5 77%. Clear differences can be seen when looking at the two units. ChatGPT-5 achieves only 68% correct solutions in Unit 1, while Claude Sonnet-4.5 achieves 80%. The situation is exactly the opposite in Unit 2. Here, ChatGPT-5 is ahead with 88%, while Claude Sonnet-4.5 only achieves 74%.

3. With regard to *literary knowledge*, ChatGPT-5 achieves a total of 92.86%, while Claude Sonnet-4.5 reaches 100%. The performances also differ in the two units. Claude Sonnet-

4.5 achieves 100% in both units, while ChatGPT-5 also comes to 100% in Unit 1, but only to 85.71% in Unit 2.

4. In the area of *aesthetic awareness*, ChatGPT-5 and Claude Sonnet-4.5 each achieve 100% in the total score and in both units.

5. ChatGPT-5 succeeds in capturing *textually intended emotions* in 85% of cases, while Claude Sonnet-4.5 achieves only 76.25%. In Unit 1, ChatGPT-5 reaches 90% correct solutions, while Claude Sonnet-4.5 comes to 87.5%. In Unit 2, ChatGPT-5 achieves 80%, while Claude Sonnet-4.5 even only reaches 65%.

Results for research question 3

While processing a unit five times, the two chatbots produce partially different solutions. In Unit 1, the results generated by ChatGPT-5 in the five runs differ for 4 semantic items and 5 idiolectal items. In Claude Sonnet-4.5, differences appear in the five runs for 2 semantic items and 2 idiolectal items. In Unit 2, the results of ChatGPT-5 vary in the five runs for 2 idiolect items and 1 item on intended emotions. With Claude Sonnet-4.5, there is only 1 item in the area of intended emotions in Unit 2 where deviations occur.

Results for research question 4

The results of the LL-AI-3 study show that the two current versions of ChatGPT and Claude perform significantly better in the area of literary literacy than ChatGPT-4o in the LL-AI-2 study. While only 81.5% of the solutions generated by ChatGPT-4o were correct, the figure is 88.9% for ChatGPT-5 and 87.4% for Claude Sonnet-4.5.

Results for research question 5

ChatGPT-5 and Claude Sonnet-4.5 also perform better than ChatGPT-4o in almost all tests relating to the five literary literacy dimensions. While ChatGPT-4o generates only 60% correct solutions semantically, ChatGPT-5 (88.64%) and Claude Sonnet-4.5 (83.64%) achieve improvements of over 20%. An increase of over 20% in the frequency of solutions can also be observed in terms of idiolect. The 55.6% of ChatGPT-4o is contrasted by 78% for ChatGPT-5 and 77% for Claude Sonnet-4.5. The progression is lower in the area of subject knowledge (ChatGPT-4o: 92%; ChatGPT-5: 92.86; ‘Claude Sonnet-4.5’: 100%) and ‘Aesthetic Awareness’ (ChatGPT-4o: 90%; ChatGPT-5: 100%; Claude Sonnet-4.5: 100%). The opposite finding can be seen with regard to intended emotions. Here, Claude Sonnet-4.5 lags behind ChatGPT-4o with 76.25% compared to ChatGPT-4o with 78.8% and ChatGPT 4 with 85%.

	ChatGPT 4o	ChatGPT 5	Claude Sonnet 4.5
	%	%	%
Semantic comprehension	60	88,64	83,64
Idiolectal comprehension	55,6	78	77
Literary Knowledge	92	92,86	100
Aesthetic Awareness	90	100	100
Intended Emotions	78,8	85	76,25
Summe	**81,5**	**88,9**	**87,4**

Fig. 4: Comparison of results from ChatGPT-4o, ChatGPT-5, and Claude Sonnet-4.5

4 Discussion and outlook

The findings of the LL-AI-3 study indicate improved performance of ChatGPT-5 and Claude Sonnet-4.5 in the area of literary literacy compared to the ChatGPT-4o version tested in LL-AI-2. In the discussion of conclusions and limitations, two perspectives must be distinguished.

4.1 Conclusions and limitations I

The aim of the LL-AI-3 study was to help eliminate a desideratum in empirical educational research in the field of L1 education in German language: the empirical verification of the reliability of generative AI using the example of literary literacy. The performance of ChatGPT-5 and Claude Sonnet-4.5 was empirically investigated. The theoretical basis was the five-dimensional model of literary literacy developed and empirically confirmed in the context of LUK research. Using the test instruments developed in this context, ChatGPT-5 and Claude Sonnet-4.5 each worked through two empirically proven units on lyrical texts five times. The results showed that both chatbots performed significantly better than ChatGPT-4o in the area of literary literacy in general and in four of the five dimensions of literary literacy. Only in the area of intended emotions Claude Sonnet-4.5 performed worse than ChatGPT-4o. The persistence of fluctuations in the solutions of an item when it is repeatedly processed is surprising. This problem, already observed in the

LL-AA-2 study with ChatGPT-4o, persists in the LL-AA-3 study with ChatGPT-5 and Claude Sonnet-4.5.

Despite these minor limitations, the findings of the LL-AI-3 study are of immediate practical relevance for L1 literature teaching in German language and other languages. This is partly because the negative results from the LL-AI-1 and LL-AI-2 studies regarding the reliability of generative AI in the context of literary literacy were not confirmed in the LL-AI-3 study. Rather, ChatGPT-5 and Claude Sonnet-4.5 prove to be significantly improved in their ability to interpret literary text compared to previous versions such as ChatGPT-4o. With scores of 87.4 and 88.9 respectively, Claude Sonnet-4.5 and ChatGPT-5 are approaching a good level, but have not yet fully achieved it. In other words, the results provide initial empirical evidence that ChatGPT-5 and Claude Sonnet-4.5 appear to be near to the level of a good learner in the area of literary competence, but not yet that of an excellent learner. While the previous studies LL-AI-1 and LL-AI-2 still suggested great caution with regard to ChatGPT 3 and 4, the results of the LL-AI-3 study provide initial indications that the latest versions of ChatGPT-5 and Claude Sonnet-4.5 can be used in a relatively reliable manner for feedback or practice purposes in literature classes. The use of generative AI as a tutorial system, which has been repeatedly recommended in research for literature teaching (Führer & Gerjets, 2024), has thus found improved foundations in the latest chatbot versions from Open AI and Anthropic. However, further improvements are needed to ensure that AI tools can be truly reliable aids in literature teaching.

For this reason, further research is needed:

- The results of the LL-AI-3 study, which are based on the latest chatbot versions, will need to be re-evaluated when newer versions from Open AI or Anthropic become available on the market.
- In follow-up studies with the LUK instruments, it should be examined to what extent the findings obtained with lyrical texts can also be confirmed with epic and dramatic literary texts, for which LUK units and test instruments are also available.
- It should be investigated to what extent prompt engineering can contribute to further optimization of performance. In the LL-AI-2 study, these attempts were only partially successful (Ascherl 2025; Brüggemann et al. 2025a). However, the challenges identified in the LL-AI-3 study in the areas of literary ambiguity, contradictory text signals, or intended emotions could provide starting points here.
- Furthermore, the irregularities in the solution of items during repeated processing by generative AI that became apparent in the LL-AI-2 study and the LL-AI-3 study need to be clarified.
- The results presented are also limited with regard to the underlying versions and types of generative AI. For example, it should be examined whether the new version ‘Claude Opus 4.5’ from Anthropic performs better in the area of literary literacy than Claude Sonnet-4.5. Other offerings such as Gemini 3, Perplexity, NotebookLM, Vobizz, or KAI should also be included in follow-up studies.

4.2 Conclusions and Limitations II

In the context of general subject didactics (Bayrhuber et al., 2017; Rothgangel et al., 2021; Frederking, 2022b), the question also arises as to what extent the results of the LL-AI-3 study are relevant for other subject didactics. In this sense, it must be clarified whether the results generated using the example of literary literacy with a reliability of approximately 88% for 'ChatGPT-5' and 'Claude Sonnet-4.5' can also be confirmed in other subjects or subject didactics for tasks on subject-specific topics. The finding that ChatGPT-5 and Claude Sonnet-4.5 arrive at different results when tasks are repeated is also significant for other subjects and subject didactics. The question is how learners, teachers and researchers get to a level of AI literacy (Long & Magerko 2022) in subject specific contents that allows them to understand principles of GPT and LLM, and understand what they mean for using these tools. For example, there are good arguments, that understanding OpenAI's "Temperature" and "Top_p" Parameters in Language Models (de la Vega, 2023) are a key for digital sovereignity and AI literacy in subject teaching and learning in general. In addition, research at the level of general subject didactics is needed to clarify whether and to what extent precise prompt engineering can significantly improve the performance of generative AI by adding additional material. These questions point to a need for research in joint projects involving several subject didactics.

References

Altman, S. (2022b). ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/index/chatgpt/ [22.11.2025].

Athaluri, S.A., Manthena, S.V., Kesapragada, K.M., Yarlagadda, V., Dave, T. & Duddumpudi, R.S.T. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. (https://doi.org/10.7759/cureus.37432)

Ascherl, C. (2025) 2026. Custom GPTs als didaktisches Supportinstrument: Chatbots konfigurieren, um literarisches Verstehen zu fördern? In L. Jagdschian & F. Hesse (Hrsg.), Literatur // KI // Didaktik. Königshausen & Neumann. (in print)

Austin, J. L. (1955) 1962. How to do things with words. The William James Lectures delivered at Harvard University in 1955. Cambridge (MA): Harvard University Press.

Bach, F., Bernhardt, S., Reuvekamp, S. & Schmiedgen, N. (2025). SHIFT happens. Lernen mit und von textgenerierender KI. In H.-G. Müller & M. Fürstenberg (Hrsg.), DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz (S. 87–105). Berlin: Frank & Timme.

Bayrhuber, H., Abraham, U. Frederking, V., Jank, W., Rothgangel, M. & Vollmer, H. J. (2017). Auf dem Weg zu einer Allgemeinen Fachdidaktik. Allgemeine Fachdidaktik, Band 1, (S. 161–178). (= Fachdidaktische Forschungen, Band 9). Münster: Waxmann.

Bergener, J., Gossen, M., Hoffmann, M. L., Bießmann, F., Veneny, M. & Korenke, R. (2023). Evaluating the Quality of ChatGPT's Climate-related Responses. Wirtschaften - Fachzeitschrift, 38(3), 46–50. (https://doi.org/10.14512/OEW380346).

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E. & Nerdel, C. (2023, 11. August). Assessing Student Errors in Experimentation Using Artificial Intelligence and Large Language Models: A Comparative Study with Human Raters. (https://arxiv.org/pdf/2308.06088.pdf).

Brüggemann, J., Ascherl, C., Okesson, L. & Frederking, V. (2025a). Digitale Souveränität im Umgang mit KI im Deutschunterricht. K. Scheiter, D. Richter & J. Brüggemann (Hrsg.), Digital gestützte Lehrkräftefortbildungen: Schwerpunkt Sprachen, Gesellschaft, Wirtschaft. Waxmann (im Druck).

Brüggemann, J., Ascherl, C., Okesson, L. & Frederking, V. (2025b). Gelingensbedingungen und Grenzen einer KI-gestützten Förderung literarischer Verstehenskompetenz. Befunde aus zwei experimentellen Studien. In MiDU – Medien im Deutschunterricht 7 (im Druck).

Cohen, J. (1988). Statistical Power Analysis for the Social Sciences. Erlbaum.

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D. & Siemens, G. (2024). Impact of AI assistance on student agency. Computers & Education, Vol. 210, Article 104967.

de la Vega, M. (2023). Understanding OpenAI's "Temperature" and "Top_p" Parameters in Language Models.https://medium.com/@1511425435311/understanding-openais-temperature-and-top-p-parameters-in-language-models-d2066504684f

Eco, U. (1962) 1998. Das offene Kunstwerk. Frankfurt am Main: Suhrkamp.

Eco, U. (1972) 2000. Einführung in die Semiotik. 9. Aufl. München: Wilhelm Fink.

Eco, U. (1990) 1999. Die Grenzen der Interpretation. Berlin: Hanser.

Else, H. (2023). Abstracts written by ChatGPT fool scientists. In Nature 613, 423.

Emsley, R. (2023). ChatGPT: these are not hallucinations - they‘re fabrications and falsifications. Schizophrenia (Heidelberg, Germany), 9(1), 52. (https://doi.org/10.1038/s41537-023-00379-4).

EU Artificial Intelligence Act (2025). https://artificialintelligenceact.eu .

Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S. & Zhang, Y [Yue]. (2023). Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation. (https://doi.org/10.48550/ARXIV.2304.01746).

Frederking, V. (2022a). Literarische Verstehenskompetenz erfassen und fördern. In S. Gailberger & F. Wietzke (Hrsg.), Handbuch Kompetenzorientierter Deutschunterricht. (S. 146–177). 2. Aufl. Weinheim: Beltz Juventa.

Frederking, V. (2022b). Allgemeine Fachdidaktik. In M. Harring, C. Rohlfs & M. Gläser-Zikuda (Hrsg.), Handbuch Schulpädagogik (S. 550–566). Münster: Waxmann.

Frederking, V. (2023). Von Fake News bis ChatGPT. Digitale Textsouveränität als ethisch-politische Bildungsaufgabe in der digitalen Welt. Medien im Deutschunterricht 5(2), 1–27. (https://doi.org/10.18716/OJS/MIDU/2023.2.4).

Frederking, V. (2025) 2026. Digitale Textsouveränität und KI. Theoretische Modellierungen und experimentelle Erprobungen am Beispiel vom ‚Prolog im Himmel‘ in Goethes ‚Faust‘. In F. Hesse & L. Jagdschian (Hrsg.): Literatur – Künstliche Intelligenz – Didaktik. Würzburg: Königshausen & Neumann. (in print)

Frederking, V., Brüggemann, J. & Hirsch, M. (2016). Das Fünfdimensionale Literary Literacy-Modell. In K. Lehmann, M. Werner & S. Zabold (Hrsg.), Historisches Denken jetzt und in Zukunft. (S. 211–234). Münster: LIT Verlag.

Frederking, V., Henschel, S., Meier, C., Roick, T., Stanat, P. & Dickhäuser, O. (2012). Beyond Functional Aspects of Reading Literacy: Theoretical Structure and Empirical Validity of Literary Literacy. (Special issue guest edited by Irene Pieper & Tanja Janssen). L1- Educational Studies in Language and Literature, 12, 35–56.

Führer, C. (2025). KI- Feedback mit Peer-Kooperation verbinden. Textüberarbeitung mit LLMS fördern. Deutsch 5-10 84 (2025), 45–51.

Führer, C. & Gerjets, P. (2024). How to understand & write literature with AI? Potentiale und Risiken von KI-Tools für Literarisches Lesen und Schreiben. MiDU - Medien im Deutschunterricht, 6(1), 1–18. (https://doi.org/10.18716/OJS/MIDU/2024.1.3).

Goddard, J. (2023). Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers. In The American Journal of Medicine, Vol. 136, Issue 11 November 2023. (https://doi.org/10.1016/j.amjmed.2023.06.012).

Gramelsberger, G. (2023). Philosophie des Digitalen zur Einführung. Hamburg: Junius.

Helm, G. & Hesse, F. (2024). Usage and beliefs of student teachers towards artificial intelligence in writing. RISTAL, *7*, 1–19. DOI: https://doi.org/10.2478/ristal-2024-0001

Hesse, F. & Helm, G. (2025). Writing with AI in and beyond teacher education: Exploring subjective training needs of student teachers across five subjects. Journal of Digital learning in teacher eDucation 2025, Vol. 41, no. 1, 21–36 (https://doi.org/10.1080/21532974.2024.2431747).

Hiebel, H. H., Hiebler, H., Kogler, K. & Walitsch, H. (1998). Die Medien. Logik – Leistung – Geschichte. Stuttgart: UTB.

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.

KMK (2024). Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit Künstlicher Intelligenz in schulischen Bildungsprozessen. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10-Handlungsempfehlung-KI.pdf [22.11.2025].

Krämer, S. (2024). "Die Nicht-Vernunft der Chatbots: Was macht auf Large Language Models beruhende Künstliche Intelligenz philosophisch interessant?". In R. Adolphi, S. Alpsancar, S. Hahn & M. Kettner (Hrsg.), Philosophische Digitalisierungsforschung: Verantwortung, Verständigung, Vernunft, Macht. (S. 297–314). Bielefeld: transcript Verlag. (https://doi.org/10.1515/9783839474976-011).

Kubacka, T. (2022). Today I asked ChatGPT about the topic I wrote my PhD about. 6.12.2022. https://x.com/paniterka_ch?lang=deI [22.11.2025].

Lee, S., Jeon, J., Mckinley, J. & Rose, H. (2025). Generative AI and English language teaching: A global Englishes perspective. September 2025 Annual Review of Applied Linguistics 45:1-24. (https://doi.org/10.1017/S0267190525100184).

Long, D. & Magerko, B. (2020). "What is AI Literacy? Competencies and Design Considerations". Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM. pp. 1–16. doi:10.1145/3313831.3376727. ISBN 978-1-4503-6708-0.

Maiwald, Klaus (2023). Digitale Textkompetenz als Bildungsaufgabe in einer Kultur der Digitalität und künstlichen Intelligenz. Mitteilungen des Deutschen Germanistenverbandes, 70(4), 359-372.

Meier, C., Roick, T., Henschel, S., Brüggemann, J., Frederking, V., Rieder, A., Gerner, V. & Stanat, P. (2017). An extended model of literary literacy. In D. Leutner, J. Fleischer, J. Grünkorn & E. Klieme (Hrsg.), Competence assessment in education (pp. 55-74). Heidelberg: Springer. (https://doi.org/10.1007/978-3-319-50030-0_5).

Metz, R. (2018). Google demos Duplex, its AI that sounds exactly like e very weird, nice human. https://www.technologyreview.com/2018/06/27/141823/google-demos-duplex-its-ai-that-sounds-exactly-like-a-very-weird-nice-human/ [22.11.2025].

Müller, H.-G. & Fürstenberg, M. (2023). Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In Mitteilungen des Deutschen Germanistenverbandes 70(4), 327–345.

Ong, W.J. (1982). Oralität und Literalität. Die Technologisierung des Wortes. Opladen: Westdeutscher Verlag.

Porschen, H. (2024). OpenAI GPT Token: Ein Einblick in die DNA von Texten in ChatGPT. https://hubertusporschen.com/openai-gpt-tokens/# [22.11.2025].

Rawte, V., Sheth, A. & Das, A. (2023). A Survey of Hallucination in "Large" Foundation Models. (https://doi.org/10.48550/arXiv.2309.05922).

Rothgangel, M., Abraham, U., Bayrhuber, H., Frederking, V., Jank, W. & Vollmer, H.J. (2020) 2021. Lernen im Fach und über das Fach hinaus. Bestandsaufnahmen und Forschungsperspektiven aus 17 Fachdidaktiken im Vergleich. Allgemeine Fachdidaktik, Band 2. (2. Aufl.). Münster: Waxmann.

Searle, J. R. (1965). „What is a speech act?“. In M. Black (Hrsg.), Philosophy in America. (pp. 221–239). London: G. Allen & Unwin.

Searle, J. R. (1969). Speech acts. An essay in the philosophy of language. Cambridge University Press.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences 3 (3), 417–457.

Searle, J. R. (1992) 1994. The rediscovery of the mind. First MIT Presspaperback edition, Massachusetts Institute of Technology.

Smith, C.S. (2023). AI Hallucinations Could Blunt ChatGPT's Success. IEEE Spectrum, 13. März 2023 (https://spectrum.ieee.org/ai-hallucination).

Steinhoff, T. (2023). Der Computer schreibt (mit). Digitales Schreiben mit Word, WhatsApp, ChatGPT & Co. als Koaktivität von Mensch und Maschine. MiDU – Medien im Deutschunterricht 5(1), 1–16.(https://doi.org/10.18716/ojs/midu/2023.1.4).

Steinhoff, T. (2025). Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner. In C. Albrecht, J. Brüggemann, T. Kretschmann, A. Krommer & C. Meier (Hrsg.), Personale und funktionale Bildung im Deutschunterricht (S. 85–99). Heidelberg: Metzler.

Su, Y., Lin, Y. & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. Assessing Writing, 57, 100752. (https://doi.org/10.1016/j.asw.2023.100752).

SWK (=Ständige Wissenschaftlichen Kommission der Kultusministerkonferenz) (2024). Large Language Models und ihre Potenziale im Bildungssystem https://www.swk-bildung.org/content/uploads/2024/02/SWK-2024-Impulspapier_LargeLanguageModels.pdf [22.11.2025].

Turing, A. (1937). ‚On Computable Numbers, with an application to the Entscheidungsproblem'. Proceedings of the London Mathematical Society, ser. 2. vol. 42, pp 230-265 (1936/7); corrections, Ibid, vol 43, 544–546.

U.S. Department of Education (2025) (Eds.). "Artificial Intelligence (AI) Guidance". https://www.ed.gov/about/ed-overview/artificial-intelligence-ai-guidance [22.11.2025].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., L. Jones, L., A. N. Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is All you Need. In Curran Associates, Inc., 2017 (https://doi.org/10.48550/arXiv.1706.03762).

Weizenbaum, J. (1965) 1966. ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine. In Communications of the ACM. Band 9, Nr. 1, Januar 1966, (https://doi.org/10.1145/365153.365168).

Appendix

Tab.1: Examples of Items in the five dimensions of literary literacy

Semantic comprehension

Der Autor des Gedichts „AUS!“ erfand für sich und seine Frau das gemeinsame Kind Ludolf, da für ihn ein Kind das größtmögliche Gemeinsame in einer Beziehung zwischen Mann und Frau darstellt.

a) Nenne eine Textstelle des Gedichts „AUS!“, die durch diese biografischen Informationen neue Deutungsmöglichkeiten erhält.

Textstelle: ✏️ Vers(e) .....

b) Wie lässt sich die Textstelle mit Hilfe dieser Zusatzinformation deuten? Notiere zwei Aspekte.

1. Aspekt: ✏️.....
2. Aspekt: ✏️.....

Idiolectal comprehension

Ein Schüler behauptet, dass die Satzzeichen am Versende oft die Aussage des Textes stützen.
Suche bitte eine Textstelle heraus, für die diese Behauptung zutrifft, und erkläre deine Wahl.

Textstelle: .....

Erklärung: .....

Literary Knowledge

In welchem der folgenden Texte spielt das Ende einer Liebesbeziehung eine wesentliche Rolle?

Kreuze bitte an, was nicht zutrifft und was zutrifft. (Bitte in jeder Zeile nur ein Kreuz)

		Trifft nicht zu	*Trifft zu*	*Kenne ich nicht*
a)	Romeo und Julia (William Shakespeare)	☐	☐	☐
b)	Das Tagebuch der Anne Frank (Anne Frank)	☐	☐	☐
c)	Erebos (Ursula Poznanski)	☐	☐	☐
d)	Corpus Delicti (Juli Zeh)	☐	☐	☐

Aesthetic Awareness

Lies dir die Verse 11 und 12 noch einmal aufmerksam durch.

„Ihr wart euer Kind. Jede Hälfte sinkt nun herab –:
ein neuer Mensch."

Dieser Textauszug ist sprachlich ungewöhnlich gestaltet. Notiere eine Auffälligkeit und erläutere kurz deine Wahl.

Auffälligkeit 1: .....

Erläuterung: .....

Intended Emotions

Welche emotionalen Reaktionen sollen durch den Text nach der Lektüre beim Leser/bei der Leserin entstehen?

Kreuze bitte an, was nicht zutrifft und was zutrifft. (Bitte in jeder Zeile nur ein Kreuz)

		Trifft nicht zu	*Trifft eher zu*
a)	Schuldgefühle	☐	☐
b)	Schadenfreude	☐	☐
c)	Betroffenheit	☐	☐
d)	Distanz zu Gefühlen	☐	☐
e)	Ärger	☐	☐
f)	Gelassenheit	☐	☐
g)	Ernüchterung	☐	☐
h)	Trauer	☐	☐