The Promises and Pitfalls of LLMs as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback[1]

[1]Both authors contributed equally to this work.

Lucas Jasper Jacobsen[2] & Kira Elena Weber[3]

[2]Leuphana Universität Lüneburg

[3]Universität Hamburg

Abstract

Artificial intelligence (AI) in higher education (HE) is reshaping teaching and learning, and feedback provided by large language models (LLMs) seems to have an impact on student learning. However, few empirical studies have compared the quality of LLM feedback with the feedback quality of real persons. Therefore, this study addresses the following questions: What prompts are needed to ensure high-quality LLM feedback in HE? How does feedback from novices, experts, and LLMs differ in terms of quality and content accuracy? We developed a learning goal with three errors and a theory-based manual to evaluate prompt quality. Specifically, three prompts of varying quality were created and used to generate feedback from ChatGPT-4. We provided the highest-quality prompt to novices and experts. Our results showed that only the best prompt produced consistently high-quality feedback. Additionally, LLM and expert feedback were significantly better than novice feedback, with LLM feedback being both faster and better than expert feedback in the categories of explanation, questions, and specificity. This suggests that LLM feedback can be a high-quality and efficient alternative to expert feedback. However, we postulate that prompt quality is crucial, highlighting the need for prompting guidelines and human expertise.

Keywords: AI; Feedback; Prompt Engineering; Teacher Education

1. Introduction

AI is transforming not only numerous industries but also the education sector, leveraging technologies like machine learning and natural language processing to improve individual learning experiences (Harry & Sayudin, 2023). Accordingly, the use of Artificial Intelligence (AI) in higher education (AIEd) has grown significantly over the past five years (Crompton & Burke, 2023) and is now reshaping the academic environment. This development has sparked extensive debates regarding its impact on the future of teaching and learning (von Garrel & Mayer, 2023). For example, several studies have examined the impact of automated writing evaluation feedback tools on students' writing performance (for a meta-analysis, see Fleckenstein et al., 2023). In general, feedback is considered an integral part of educational processes in higher education (HE), but as Henderson et al. (2019) noted, it raises issues as well:

> *Feedback is a topic of hot debate in universities. Everyone agrees that it is important. However, students report a lot of dissatisfaction: they don't get what they want from the comments they receive on their work and they don't find it timely. Teaching staff find it burdensome, are concerned that students do not engage with it and wonder whether the effort they put in is worthwhile*. (p. 3)

The provision of feedback represents both an opportunity and a challenge, as it can also have negative consequences for learning processes (Kluger & DeNisi, 1996). Therefore, high-quality feedback is needed in HE, characterized by features such as concreteness, activation, and empathy (Prilop et al., 2019). Unfortunately, the human and financial resources required to provide high-quality feedback are often lacking (Demszky et al., 2023), which is why AI feedback is promising, offering the potential to optimize teaching and learning processes in

HE. According to the United Nations Educational, Scientific and Cultural Organization (UNESCO) (2024), AI tools should be used to enhance the professional development of teachers, allowing them “to practice skills and receive feedback” (p. 42).

Following up on these ideas, this study will specifically look at the potential of Large Language Model (LLM-)based feedback in HE by comparing it to novice and expert feedback. For teachers to use AI tools, they need adequate “application skills” (UNESCO, 2024, p. 22). Consequently, we present a theory-driven and evidence-based manual for prompt engineering that can facilitate teachers’ use of AI and improve their ability to apply it in the educational context. We address the following questions in this study: 1) What kinds of prompts are required to ensure high-quality AI feedback? 2) What are the differences between novice, expert, and AI feedback in terms of feedback quality?

2. Theoretical Background

2.1 Artificial Intelligence in Higher Education

Although AIEd has been around for about three decades, educators struggle with understanding how to use it for pedagogical purposes and its implications for teaching and learning in HE (Zawacki-Richter et al., 2019). However, the field of AIEd is growing and developing rapidly, and there is an urgent need to improve academic understanding of its potential and limitations (Crompton & Burke, 2023). It has been argued that AIEd, including LLMs like ChatGPT, can increase student engagement, collaboration, and the accessibility of education (Cotton et al., 2023). It has been used for various purposes, including assessment/evaluation, prediction, AI assistance, intelligent tutoring, and managing student learning (Crompton & Burke, 2023). Nevertheless, empirical research on this new technology, particularly in the context of HE, is still in its infancy, and more research is needed. This paper contributes to the ongoing discourse by developing a theory-driven manual for analyzing the quality of prompts, thereby ensuring high-quality output. Furthermore, we critically examine the quality of LLM feedback in HE, especially in comparison to novice and expert feedback.

2.1.1 Prompt Engineering for Large Language Models in Higher Education

To effectively use LLMs in HE, it is crucial to recognize the importance of prompt engineering. Research must answer the question of how to write prompts that yield high-quality output. In simple terms, prompt engineering is the process of designing effective questions or stimuli, known as "prompts," for LLMs. The aim is to get clear, relevant answers. Essentially, the process involves fine-tuning questions for LLMs to produce the desired results. Although prompt engineering is a fairly new research topic, findings have consistently suggested that the quality of the output of LLMs is not merely determined by their foundational algorithms or training data. Equally crucial is the clarity and accuracy of the prompts they are given (Bsharat et al., 2023; Lo, 2023; Zamfrescu-Pereira et al., 2023).

Studies have highlighted different aspects of prompt engineering (e.g., ChatGPT & Enkin 2023; Lo 2023; Zamfrescu-Pereira et al., 2023). For example, Kipp (2023) noted that four primary elements (*context*, *question*, *format*, and *examples*) should serve as modular guidelines for constructing effective prompts. Enkin (2023) proposed five factors that influence prompt selection: *user intent*, *model understanding*, *domain specificity*, *clarity* and *specificity*, and *constraints*. In addition, Lo (2023) developed the CLEAR framework, which comprises five key elements of effective prompts: *concise*, *logical*, *explicit*, *adaptive*, and *reflective*.

The ability to develop prompts is crucial to support future skills in an increasingly AI-influenced world (UNESCO, 2024). However, creating effective prompts can be challenging and may lead to unexpected results (Zamfrescu-Pereira et al., 2023). To the best of our knowledge, there are no manuals for analyzing the quality of prompts within HE, and no investigations have been performed to determine whether such guidelines actually improve the output of LLMs. This study aims to develop such a manual and investigate whether there are differences in output when feeding an LLM (ChatGPT-4) with different kinds of prompts.

2.2 Feedback

Feedback is widely recognized as an integral component of individual and institutional learning and developmental processes (Prilop et al., 2019; Weber et al., 2019) and thus as a crucial component in HE (Henderson et al., 2019). Feedback is defined as information offered to an individual concerning their current performance to facilitate improvement in future endeavors (Narciss, 2013), and individuals often struggle to effectively reflect on, manage, and adjust their actions or tasks in the absence of appropriate feedback (Weber et al., 2018). In the context of teacher education, pre-service teachers receive feedback after actual classroom practice or specific skill training, which could be from peers with a similar knowledge base (novices) or from experts with knowledge authority (Weber et al., 2018; Lu, 2010). However, while the incorporation of feedback sessions in teacher education is becoming increasingly prevalent (Weber et al., 2018; Kraft et al., 2018), feedback designs are often compromised, as feedback from novices is not as high in quality as expert feedback (Weber et al., 2019). In addition, educators (experts) frequently express concerns about a lack of time for delivering high-quality feedback (Demszky et al., 2023).

2.2.1 Feedback Quality

Ericsson et al. (1993) underscored that substantial enhancements in performance are achievable only through high-quality feedback. Similarly, Prilop et al. (2021) showed that the quality of feedback is crucial to its acceptance and for facilitating the continuous development of professional competencies among teachers. Regarding the quality of feedback, Prilop et al. (2019, 2021) provided criteria for effective feedback for teachers based on various studies in other domains (e.g., Gielen & De Wever, 2015; Prins et al., 2006). Summarizing these criteria, effective feedback should consistently be *specific*, *empathetic*, and *engaging* (Prilop et al., 2019; Prilop & Weber, 2023). On a cognitive level (specific and engaging), numerous studies (e.g., Strijbos et al., 2010) have suggested that effective feedback should incorporate both evaluative and tutorial components. Therefore, individuals providing feedback should assess a particular situation with a firm emphasis on content, offer and explain alternative actions, and pose engaging questions. At the affective-motivational level (empathetic), the delivery of feedback is crucial. Ultimately, according to Prins et al. (2006), effective peer feedback should be presented in first person. This perspective suggests that feedback is subjective and open to dialogue rather than an indisputable fact. In our previous research (Prilop et al., 2021), we found that critiques should always be counterbalanced by positive evaluations. Regarding the criteria for high-quality feedback, a few studies (Prins et al., 2006; Weber et al., 2019) have examined the impact of expertise on feedback quality by comparing the feedback provided by novices and experts.

2.2.2 Novice and Expert Feedback

Hattie and Timperley (2007) emphasized that feedback can be provided by different agents, such as experts or novices. The disparity in the quality of feedback given by experts and novices has been systematically examined in a few studies. Prins et al. (2006) compared expert and novice feedback in medical education, finding that experts utilized more criteria, provided more situation-specific comments and positive remarks, and frequently adopted a first-person perspective style. They also observed that a significant portion of novices either did not pose any reflective questions (59%) or failed to offer alternative suggestions (44%). Similar observations were made in the domain of teacher education by Weber et al. (2019). Specifically, they reported that expert feedback was more specific, question-rich, and first-person-perspective-oriented than pre-service teachers' feedback at the bachelor level. Pre-service teachers seldom included specific descriptions of teaching situations in their feedback and rarely utilized activating questions. In sum, expert feedback seems to be of higher quality than novice feedback. However, the provision of adaptive feedback is resource intensive if done manually for every learner's task solution, and accordingly, experts in HE often struggle to provide high-quality feedback due to insufficient resources (Henderson et al., 2019). LLM feedback offers a potential solution (Sailer et al., 2023), but it remains unclear whether LLM feedback is qualitatively equivalent to expert feedback in HE.

2.2.3 Large Language Models as Feedback Providers

The integration of AI into education is changing teaching methods, curriculum planning, and student engagement (Wang et al., 2024). Recent studies have investigated the use of LLMs to generate adaptive feedback. For example, in their meta-analysis, Fleckenstein et al. (2023) established that the utilization of automated feedback could enhance students' writing progress. Zhu et al. (2020) examined an LLM-powered feedback system in a high school climate activity task and found that it helped students refine their scientific arguments. Sailer et al. (2023) investigated the impact of adaptive feedback on pre-service teachers' diagnostic reasoning, showing that while it improved justification quality in written assignments, it did not enhance diagnostic accuracy. In contrast, static feedback negatively affected learning in dyads. Additionally, Bernius et al. (2022) used natural language processing models to generate feedback for student responses in large courses, reducing grading effort by up to 85% and being perceived as highly precise. Kasneci et al. (2023) highlighted how LLMs can assist university and high school teachers with research and writing tasks, improving efficiency and reducing the time spent on personalized feedback (Kasneci et al., 2023). In a recent study, Dai et al. (2024) investigated the ability of two GPT model versions (GPT-3.5 and GPT-4) to provide feedback on students' open-ended writing assignments. The feedback generated by GPT-3.5 and GPT-4 was compared to that of human instructors, evaluating three key aspects: readability, the presence of effective feedback components, and reliability in assessing student performance. The results indicated that (1) both GPT-3.5 and GPT-4 consistently produced more readable feedback than human instructors, (2) GPT-4 outperformed GPT-3.5 as well as human instructors by delivering feedback enriched with crucial components such as feeding-up, feeding-forward, and self-regulation strategies, and (3) GPT-4 exhibited superior feedback reliability compared to GPT-3.5. Considering the results of these previous studies, LLMs appear to be promising feedback givers. However, there is still a lack of empirical evidence in the context of teacher education as well as on the quality of feedback in terms of the criteria for effective feedback (*specific*, *empathetic*, and *engaging*). Moreover, our study addresses the importance of prompt engineering when using LLMs as feedback providers.

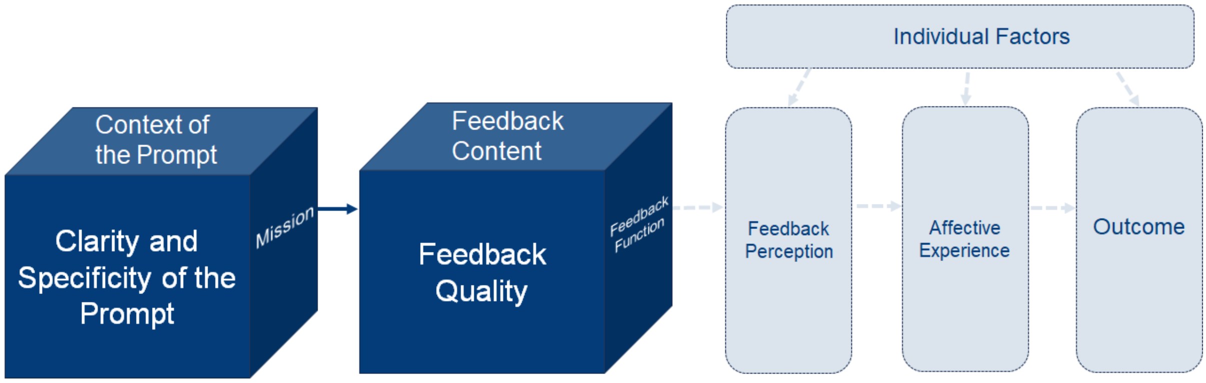
3. Aim of the Study

Our paper seeks to enhance the current understanding of LLM-based feedback in HE by addressing the following research questions:

a. What kinds of prompts are required to ensure high-quality LLM feedback?
b. What are the differences between novice, expert, and AI feedback in terms of feedback quality (*specific*, *empathetic*, and *engaging*)?

Figure 1 shows our heuristic working model, which includes the quality of prompts, the quality of feedback, and potential outcomes that should be investigated in future studies.

Figure 1: *Heuristic Working Model adapted from Narciss (2008) and Pekrun et al. (2023).*



4. Method

4.1 Development of a Theory-Driven Prompt Manual

We developed a theory-driven coding manual to analyze prompt quality for LLMs, integrating various prompt engineering approaches. Our design followed Kipps' (2023) four key elements of prompt engineering and considered five factors influencing prompt selection highlighted by ChatGPT and Enkin (2023). Lastly, we applied Lo's (2023) CLEAR framework to refine each prompt module. This resulted in a manual with eight distinct categories of prompt quality (see Table 1).

Table 1: *Prompt manual to ensure the development of high-quality prompts.*

Category	Subcategory	Good	Code	Average	Code	Suboptimal	Code

Context	Role	The role of the LLM and of the person asking the question is explained	2	Only one of the roles is explained	1	Neither the role of the LLM nor the role of the person asking the question is explained	0
	Target audience	There is a clearly defined and described target audience	2	The target audience is roughly described	1	The target audience is not specified	0
	Channel	The channel is clearly described	2	The channel is roughly described	1	The channel is not mentioned	0
Mission	Mission/Question	The mission of the LLM is clearly described	2	The mission of the LLM is roughly described	1	The mission of the LLM is not clear	0
Clarity and specificity	Format and constraints	Stylistic properties as well as length specifications are described	2	Either stylistic properties are described or a length specification is given	1	Neither stylistic properties nor length specifications are given	0
	Conciseness	The prompt contains only information that is directly related and relevant to the output. It is clear and concise	2	The prompt is concise with little superfluous information	1	The prompt contains a lot of information that is irrelevant to the mission/question	0

Domain specificity	Technical terms are used correctly and give the LLM the opportunity to refer to them in the answer	2	Technical terms are used sporadically or without explanation	1	No specific vocabulary that is relevant to the subject area of the question is used	0
Logic	The prompt has a very good reading flow, internal logical coherence, a very coherent sequence of information, and a clearly understandable connection between the content and mission	2	The prompt fulfills only some of the conditions of the coding "2"	1	The prompt is illogically constructed	0

Subsequently, we developed three prompts of different quality (poor, medium, good) using our prompting manual. Following Wittwer et al. (2020), we formulated a learning goal with three types of errors (learning goal: students will recognize a right triangle and understand the Pythagorean theorem [type of errors: *no activity verb*; *instructional rather than learning goal*; and *multiple learning goals in a single statement*]) and asked ChatGPT-4 to provide feedback on the abovementioned learning goal.

4.2 Assessment of Feedback Quality

To analyze the quality of LLM feedback and answer our first research question, we conducted a quantitative feedback analysis. We adapted the coding scheme of Prilop et al. (2019) based on the feedback quality index developed by Prins et al. (2006). Each feedback instance served as a unit of analysis and enabled a thorough content evaluation. The original scheme comprises six categories: evaluation criteria, specificity, suggestions, questions, first-person perspective, and valence (positive/negative). The feedback is assigned a rating of "2" for high quality, "1" for average, and "0" for suboptimal. A detailed explanation of this process can be found in Prilop et al. (2020). We added three categories: errors, explanations, and explanations of suggestions. The error category was necessary due to the tendency of LLMs to hallucinate (Alkaissi & McFarlane, 2023; Ji et al., 2022), with points deducted in this area. Hallucination in LLMs refers to the generation of information or responses that appear plausible but are factually incorrect or not based on the given input or data. The category explanation was based on the manual by Wu and Schunn (2021). Finally, suggestions were divided into two categories: *presence of suggestion* and *explanation of suggestion* to improve coding accuracy (see Table 2 for the coding manual and inter-coder reliability).

Table 2: *Content analysis of feedback quality: Categories, examples, and inter-coder reliability (Fleiss kappa).*

Category	Good feedback definition	Code	Average feedback definition	Code	Sub-optimal feedback definition	Code	κ	Good feedback example
Assessment criteria	Aspects of a good learning goal are addressed using technical terms/theoretical models	2	Aspects of a good learning goal are addressed without technical terms/theoretical models	1	Aspects of a good learning goal are not addressed	0	.81	"However, the learning goal, as currently stated, has room for improvement. The verb 'recognize' is on the lower end of Bloom's taxonomy; it's

								more about recall than application or analysis." (LLM feedback 3.30)
Specificity	All three error types are named and explicitly explained	2	Two types of errors are named and explicitly explained	1	One type of error is named and explicitly explained	0	.81	"Your goal contains two separate objectives: […] Next, the verbs you've chosen, 'recognize' and 'understand,' are a bit vague in the context of Bloom's taxonomy […] And how do you envision this learning goal relating back to the learner? […]"(LLM feedback 3.28)
Explanation	A detailed explanation is given regarding why the aspects of a good learning goal are relevant	2	A brief explanation is given of why the aspects of a good learning goal are relevant	1	No explanation is given regarding why the aspects of a good learning goal are relevant	0	.86	"According to best practices, it's beneficial to focus on just one learning goal at a time. This makes it clearer for both you and the students, streamlining the assessment

								process." (LLM feedback 3.14)
Presence of suggestions for improvement	Alternatives are suggested in a cognitively stimulating way	2	Alternatives are presented in concrete terms	1	No alternatives are named	0	.86	"A more targeted learning goal will focus on just one of these. Which one is your priority?" (LLM feedback 3.28)
Explanation of suggestions	Alternatives are explained in detail	2	Alternatives are briefly explained	1	Alternatives are not explained	0	.82	"This would align the goal more closely with achieving deeper understanding and skill utilization. […]This goal is learner-centered, contains only one focus, and involves higher-level thinking skills. It also makes the intended learning outcome clear." (LLM feedback 3.30)
Errors	The feedback includes several	–2	The feedback includes one error	–1	Feedback does not include errors	0	.90	

	content errors regarding learning goals		regarding learning goals		regarding learning goals			
Questions	Activating question posed	2	Clarifying question posed	1	No questions posed	0	1.00	“So, what specific skill or understanding are you hoping your students will gain by the end of this lesson?” (LLM feedback 3.28)
First person	Written in first person throughout feedback	2	Occasionally written in first person	1	Not written in first person	0	.90	“I appreciate the effort you’ve put into formulating this learning goal for your future teachers. […]Let me share my thoughts with you. Firstly, I noticed […]” (LLM feedback 3.23)
Valence	Balance between positive and negative feedback	2	Mainly positive feedback	1	Mainly negative feedback	0	.76	“I don’t think this learning goal is well worded. [...]However, I like that your learning goal is formulated in a very clear and structured

way.” (Novice feedback 13)

4.3 Coding of the Feedback

The LLM feedback (20 pieces of feedback from low-quality prompts, 20 from medium-quality prompts, and 20 from high-quality prompts) was coded by three trained student coders. These coders were trained by a member of the research team and initially coded a sample of 20 feedback comments. Any discrepancies were discussed and resolved following the method described by Zottmann et al. (2013). The feedback was then randomly assigned for coding. Fleiss’ kappa (κ) was used to measure agreement between coders, resulting in significant kappa values (see Table 2), indicating reliable coding. Based on the analysis, it became clear which prompt provided better results. Subsequently, the high-quality prompt was presented to 30 pre-service teachers (novices), seven teacher trainers, two educational science professors, one teacher trainer, and one headmaster (experts), who also formulated feedback. This feedback was coded by the same coders.

4.4 Analysis Method

We used our prompt manual to analyze the prompt quality of our three different prompts. We then analyzed differences between LLM feedback ($n = 30$), expert feedback ($n = 11$), and novice feedback ($n = 30$) (independent variables) concerning the different subdimensions of feedback quality (dependent variables) using one-way analyses of variance (ANOVAs), followed by Bonferroni post hoc tests. All statistical calculations were performed using SPSS 26, and we set the significance level at $p < .05$ for all tests.

5. Results

5.1 Differences between Prompts and their Output

Regarding the first research question, we fed ChatGPT-4 with different types of prompts (to see the prompts, please view the supplementary material) and analyzed the outcome in terms of quality as well as the accuracy of the feedback. The first prompt achieved low quality (5 out of 16 points according to the prompt manual). The second prompt contained more details than the first and therefore achieved slightly higher quality (8 out of 16 points). The third prompt had the highest quality, scoring 15 out of a possible 16 points. We generated feedback 20 times for each prompt and coded the results using our feedback quality manual. To compare the feedback, we conducted an ANOVA with Bonferroni post hoc tests. Our results showed significant differences between the prompts regarding feedback quality for all subdimensions except *valence* and *presence of suggestions* (for more details about descriptive data, see Table 3). Bonferroni-adjusted post hoc tests revealed that the feedback generated with prompt 3 (most sophisticated prompt) performed significantly ($p < .001$) better in the subcategory *assessment criteria* than prompt 1 ($M_{Diff} = 1.50$, 95% CI[1.10, 1.90]) and prompt 2 ($M_{Diff} = 0.90$, 95% CI[0.50, 1.30]). We found the same effect for the categories *explanation* (prompt 1: $M_{Diff} = 0.75$, 95% CI[0.41, 1.09], $p < .001$; prompt 2: $M_{Diff} = 0.40$, 95% CI[0.06, 0.74], $p < .05$), *first person* (prompt 1: $M_{Diff} = 1.05$, 95% CI[0.63, 1.47], $p < .001$; prompt 2: $M_{Diff} = 0.95$, 95% CI[0.53, 1.37], $p < .001$), and *questions* (prompt 1: $M_{Diff} = 0.70$, 95% CI[0.28, 1.12], $p < .001$; prompt 2: $M_{Diff} = 1.00$, 95% CI[0.58, 1.42], $p < .001$). Furthermore, the feedback generated with prompt 3 was significantly ($p < .001$) better than that generated with prompt 1 for the categories *explanation of suggestion* ($M_{Diff} = 0.60$, 95% CI[0.23, 0.97]) and *specificity* ($M_{Diff} = 1.25$, 95% CI[0.90, 1.60]). For the category *error*, prompt 2 generated significantly ($p < .001$)

more errors than prompt 1 (M_{Diff} = –0.85, 95% CI[–1.34, –0.36]) and prompt 2 (M_{Diff} = –0.95, 95% CI[–1.44, –0.46]).

Table 3: *Quality of the feedback generated using the three different prompts.*

Category	Concreteness											
Subcategory	Assessment criteria				Explanation							
	M	SD	Min.	Max.	M	SD	Min.	Max.				
Prompt 1	0.45	.76	0	2	0.25	.44	0	1				
Prompt 2	1.05	.39	0	2	0.60	.50	0	1				
Prompt 3	1.95	.22	1	2	1.00	.32	0	2				
	Empathy											
	First person				Valence							
	M	SD	Min.	Max.	M	SD	Min.	Max.				
Prompt 1	0.00	.00	0	0	0.85	.56	0	2				
Prompt 2	0.10	.45	0	2	1.00	.00	1	1				
Prompt 3	1.05	.83	0	2	1.00	.00	1	1				
	Activation											
	Questions				Presence of suggestions for improvement				Explanation of suggestions			
	M	SD	Min.	Max.	M	SD	Min.	Max.	M	SD	Min.	Max.
Prompt 1	1.20	.52	0	2	1.15	.75	0	2	0.50	.51	0	1
Prompt 2	0.90	.72	0	2	1.15	.37	1	2	1.25	.55	0	2
Prompt 3	1.90	.31	1	2	1.50	.51	1	2	1.10	.31	1	2
	Correctness											
	Specificity				Errors							
	M	SD	Min.	Max.	M	SD	Min.	Max.				
Prompt 1	0.10	.30	0	1	–0.40	.50n	–1	0				
Prompt 2	1.05	.39	0	2	–1.25	.79	–2	0				

Prompt 3	1.35	.59	0	2	–0.30	.57	–2	0

These quantitative results are illustrated using examples of feedback generated using the low-quality prompt (prompt 1) and the high-quality prompt (prompt 3). The feedback on prompt 1 received zero points for feedback quality from our coders. While it vaguely listed some errors and mentioned aspects of a learning goal, the feedback lacked clarity and would not help a novice improve their learning goal. In contrast, the feedback on prompt 3 received 12 out of 16 points. High-quality sections of this feedback are annotated and explained (comments in italics and gray). The following example illustrates the difference in feedback quality when an LLM is provided with prompts of varying quality, although it is not intended as an in-depth qualitative analysis.

LLM feedback generated using Prompt 1:

The learning goal "Students will recognize a right triangle and understand the Pythagorean theorem" is clear but could benefit from further specificity. Recognizing right triangles and understanding the Pythagorean theorem are two distinct skills, and clarifying the level of understanding and application expected can enhance the goal. While the goal is relevant to mathematics education, it would be beneficial to outline how success will be measured. Breaking the goal into subgoals, such as identification, understanding, and application, could help in creating appropriate assessments. Finally, aligning this goal with students' grade level and setting a specific time frame might provide a more structured educational pathway. This feedback aims to enhance the clarity and measurability of the goal, ensuring it is tailored to the students' needs. *[→ The LLM does not write in first person throughout the feedback; the LLM provides no alternatives, poses no questions, and recognizes only one error]*

LLM feedback generated using Prompt 3:

I appreciate the effort you've put into formulating this learning goal. It's clear that you aim to teach a foundational concept in geometry that has a wide range of applications. Let me share my thoughts with you. *[→ The LLM writes in first person throughout the feedback]*

Firstly, I noticed that the goal contains two distinct objectives: recognizing right triangles and understanding the Pythagorean theorem. *[→ The LLM recognizes the first error: multiple learning goals in a single statement]* This could lead to a scattered focus. *[→ The LLM explains the relevance of learning goals]* How might you refine it to contain just one learning goal? *[→ The LLM suggests alternatives in a cognitively stimulating way and poses a question]*

In terms of action verbs, "recognize" and "understand" are quite broad and don't lend themselves to easy assessment. *[→ The LLM recognizes the second error: No activity verb]* Bloom's taxonomy might help you choose more specific verbs that clarify what you want students to do. *[→ The LLM uses theoretical models to address good learning goals]* Perhaps "identify" for recognizing right triangles and "apply" for using the Pythagorean theorem would be more measurable and actionable? *[→ The LLM presents alternatives]*

The goal is aimed at the learner, which is a positive, but I wonder if it's concrete enough. How will you measure "understanding" in this context? Could you link this more closely to a specific learning outcome, like solving a problem using the Pythagorean theorem? *[→ The LLM recognizes the third error: instructional rather than learning goal; and poses a question]*

Lastly, the connection between content and goal is somewhat present but could be more explicit. Could the goal focus more on what students will actually do with their knowledge about right triangles and the Pythagorean theorem in practice? *[→ The LLM poses a thought-provoking question]*

So, what do you think about these points? How would you modify the learning goal to make it more specific, actionable, and closely related to measurable outcomes? *[→ The LLM poses thought provoking questions]*

5.2 Differences between Novice, LLM, and Expert Feedback

To compare LLM feedback with novice and expert feedback, we provided the highest-quality prompt (prompt 3) to pre-service teachers and experts (see Section 4.3 for expert details). An ANOVA with Bonferroni post hoc tests revealed significant differences among the groups in feedback quality across all subdimensions except *empathy*, *valence*, and *first person* (see Table 4 for descriptive data). The Bonferroni-adjusted post hoc tests confirmed previous findings (Weber et al., 2019, Prilop et al., 2021), indicating that expert feedback was more *concrete*, *activating*, and *correct* but not more *empathetic* than that of novices. Expert feedback showed significantly higher quality ($p < .001$) in the subcategories *assessment criteria*, *explanation*, *questions*, *presence of suggestions*, *explanation of suggestions*, and *specificity.* The comparison between novice and LLM feedback showed that LLM feedback outperformed novice feedback in all subcategories except *valence* and *first person*. Regarding the difference between LLM and expert feedback, the Bonferroni adjusted post hoc tests revealed that the LLM feedback had higher quality than expert feedback in the subcategories *explanation* ($M_{Diff} = 0.46$, 95% CI [0.17, 0.74], $p < .001$)*, questions* ($M_{Diff} = 0.50$, 95% CI [0.07, 0.93], $p < .05$), and *specificity* ($M_{Diff} = 0.96$, 95% CI [0.52, 1.41]).

Table 4: *Quality of the novice, expert, and LLM feedback.*

	Concreteness	
	Assessment criteria	Explanation

	M	SD	Min.	Max.	M	SD	Min.	Max.				
Peers	0.63	.81	0	2	0.10	.31	0	1				
Experts	1.64	.51	1	2	0.55	.52	0	1				
ChatGPT-4	1.97	.18	1	2	1.00	.26	0	2				
	Empathy											
	First person				Valence							
	M	SD	Min.	Max.	M	SD	Min.	Max.				
Peers	1.10	.71	0	2	1.10	.30	1	2				
Experts	1.18	.60	0	2	1.25	.50	1	2				
ChatGPT-4	1.10	.76	0	2	1.00	.39	0	2				
	Activation											
	Questions				Presence of suggestions for improvement				Explanation of suggestions			
	M	SD	Min.	Max.	M	SD	Min.	Max.	M	SD	Min.	Max.
Peers	0.17	.38	0	1	0.87	.82	0	2	0.30	.54	0	2
Experts	1.36	.81	0	2	1.73	.47	1	2	0.82	.60	0	2
ChatGPT-4	1.86	.44	0	2	1.57	.50	1	2	1.13	.35	1	2
	Correctness											
	Specificity				Errors							
	M	SD	Min.	Max.	M	SD	Min.	Max.				
Peers	0.17	.38	0	1	–.73	.87	–2	0				
Experts	0.64	.67	0	2	–.18	.60	–2	0				
ChatGPT-4	1.60	.56	0	2	–.17	.46	–2	0				

6. Discussion

The findings of this study offer compelling insights into the utility and effectiveness of LLM-based feedback in HE. Currently, novice feedback, in the form of peer feedback, is often used in HE, but it is not always conducive to learning (Kluger & DeNisi, 1996). Moreover, it is challenging for experts to provide high-quality feedback in HE due to a lack of human and financial resources (Demszky et al., 2023). LLM feedback can provide an enriching and economical alternative. A particularly promising result of our study is that feedback generated by the LLM surpassed novice feedback in quality and even rivaled that of experts. Accordingly, our results align with those of Dai et al. (2024) while underlining the importance of prompting when using LLMs.

Our first research question addressed what kinds of prompts are needed to generate high-quality LLM feedback. One key finding of our study was the importance of prompt quality in determining the quality of LLM-based feedback. While LLMs can generate high-quality feedback, the output is dependent on the *context*, *mission*, *specificity*, and *clarity* of the prompts provided. The study revealed that only the prompt with the highest quality could induce the LLM to generate consistent high-quality feedback. When considering the category *error*, prompt 2 was revealed to be a wolf in sheep's clothing, having good stylistic properties but resulting in significantly more errors than prompt 1 and more errors than any other prompt or feedback provider in this study. This illustrates the potential of LLMs to hallucinate (Alkaissi & McFarlane, 2023; Ji et al., 2022) and underscores the importance of careful, theory-driven prompt design. The ability to craft high-quality prompts is a skill that educators need to master (e.g. Zamfrescu-Pereira et al., 2023; UNESCO, 2024), necessitating a manual or guidelines. In our study, we designed a prompt manual which could and should be used by educators who work with LLMs.

With regard to research question 2, our study supports previous findings (Prilop et al., 2021; Weber et al., 2019) showing that expert feedback is of higher quality than novice feedback. We found that experts outperformed pre-service teachers in the categories *concreteness, activation*, and *correctness* but not in the category *empathy*. The same was true when we compared LLM and novice feedback. By comparing LLM feedback with expert feedback, we complement these findings, providing new insights regarding feedback processes in HE. Our results show that LLM feedback can outperform expert feedback in the categories *explanation, questions*, and *specificity*. This attests to the transformative potential of LLMs in educational settings, offering the promise of scalable, high-quality feedback that could revolutionize the way educators assess student work. Furthermore, the LLM-based feedback was produced in significantly less time than expert feedback (in our study, ChatGPT-4 produced an average of 49 pieces of feedback in the same amount of time that an expert produced one), heralding efficiency gains that could free up educators for more personalized or creative pedagogical endeavors. However, considering our proposed heuristic model, future studies should investigate how LLM-based feedback is perceived by students and whether students' learning experiences and learning gains can be enhanced by LLM feedback.

Overall, our findings support the results of Dai et al. (2024) and lend credence to the promise of LLMs as a viable alternative to expert feedback in HE. However, we must also consider the scope and limitations of LLMs. While they can quickly analyze and generate feedback based on set parameters, LLMs lack the nuanced understanding of individual learners' psychology, needs, and the socio-cultural context within which learning occurs. LLMs seem to perform particularly well with task-related feedback (Dai et al., 2024), which corresponds to the feedback-level observed in this study. Nevertheless, it is crucial to recognize that expertise is not solely a function of accurate or quick feedback. Experts bring a depth of experience, professional judgment, and a personal touch to their interactions with students. These qualities are currently beyond the reach of AI systems and may prove irreplaceable in educational settings that value not only the transfer of knowledge but also the building of relationships and character. Even if efficiency and quality are the only benchmarks, there was one outlier with multiple errors among the 20 feedback comments generated by the highest-quality prompt. Thus, we posit that experts are still needed but that their tasks should be shifted from providing feedback to monitoring and revising LLM feedback.

Future studies should investigate how the quality of expert feedback can be enhanced by using LLMs and how this intertwined approach is perceived by students and educators in HE. Going beyond the promise of efficiency and quality, and considering Russel and Norvig's (2010) warning that every researcher in the field of AI should be aware of the ethical ramifications of their projects, it becomes evident that the ethical and data-related dimensions of LLMs cannot be ignored in HE. While LLMs are not subjectively biased, the data on which they are trained does have inherent biases. Moreover, there are potential concerns about data security, privacy, and intellectual property, particularly in a learning environment where sensitive information may be discussed. As educators and policymakers consider implementing LLMs in HE, these ethical questions need careful attention and possibly regulatory oversight. In sum, we come to the same conclusion as Zawacki-Richter et al. (2019): "We should not strive for what is technically possible, but always ask ourselves what makes pedagogical sense" (p. 21).

6.2 Limitations and Implications

This study takes an in-depth look at the efficacy of LLMs as a tool for generating feedback in HE. An important limitation of our study that warrants discussion is the restricted focus on a single learning goal and a limited set of errors for which feedback was generated. This narrow scope may limit the generalizability of our findings. While we found that the LLM outperforms both novices and experts in providing high-quality feedback for the specific errors we examined, it remains an open question whether these findings would hold true across a broader range of academic subjects and tasks in HE. Educational settings are diverse, encompassing a wide array of subjects, each with their own unique types of content and forms of assessment. Therefore, it would be risky to assume that the efficacy of an LLM in our context would be universally applicable across all educational environments. Future research should aim to diversify the types of tasks and the corresponding feedback. This would provide a more comprehensive understanding of where LLM-based feedback can be most effectively and appropriately utilized in HE. Until such broader research is conducted, the application of our findings should be considered preliminary and best suited for contexts similar to the one we studied. For this reason, we are conducting a study in which we compare the feedback quality of three different LLMs giving feedback to 153 pre-service teachers regarding their learning goals in their first teaching practicum.

Another practical implication of this study is that the relevance of prompt engineering may create a barrier to entry for educators less familiar with the nuances of designing effective prompts, thus necessitating further training or guidance.

6.3 Conclusion

In conclusion, there is compelling evidence supporting the use of LLMs as tools for feedback in HE, including its quality and efficiency. However, its application is not without pitfalls. Overall, we find that LLMs have the potential to be useful tools, but educators must be skilled in prompt engineering and adept at utilizing the tool to achieve optimal results. As Azaria et al. (2023) emphasized in the title of their article, “ChatGPT is a Remarkable Tool - For Experts.” The dependence on prompt quality, ethical challenges, and the irreplaceable nuanced inputs from human experts make it a tool to be used cautiously. Future research should explore these dimensions in more detail, possibly leading to a balanced hybrid approach that combines the strengths of both LLM and human expertise in educational feedback mechanisms. The endeavor to incorporate LLMs in HE is not a question of replacement but of augmentation. How we navigate this balance will determine the efficacy of such technological solutions in truly enriching the educational landscape.

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Azaria, A., Azoulay, R., & Reches, S. (2023). *ChatGPT is a remarkable tool–For experts*. https://doi.org/10.48550/arXiv.2306.03102

Baidoo-Anu, D., & Ansah, L. O. (2023). *Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning*. https://www.researchgate.net/publication/369385210

Bernius, J. P., Krusche, S., & Bruegge, B. (2022). Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, *3*, 100081. https://doi.org/10.1016/j.caeai.2022.100081

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2024). Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4 [Preprint]. arXiv. https://arxiv.org/abs/2312.16171

ChatGPT, & Ekin, S. (2023). Prompt engineering for ChatGPT: A quick guide to techniques, tips and best practice. https://doi.org/10.36227/techrxiv.22683919

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00392-8

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

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., Gasevic, D., Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation:

An evaluation study. *Computers and Education: Artificial Intelligence*, *7*, 100299. https://doi.org/10.1016/j.caeai.2024.100299

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2023). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 483-505. 016237372311692. https://doi.org/10.3102/01623737231169270

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406.

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, *6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Gielen, M., & De Wever, B. (2015). Structuring peer assessment: Comparing the impact of the degree of structure on peer feedback content. *Computers in Human Behavior*, *52*, 315–325.

Harry, A., & Sayudin, S. (2023). Role of AI in education. *Interdisciplinary Journal and Hummanity (INJURITY)*, *2*(3), 260–268.

Hammerness, K. M., Darling-Hammond, L., Bransford, J., Berliner, D. C., Cochran-Smith, M., McDonald, M., & Zeichner, K. M. (2005). How teachers learn and develop. In L. Darling-Hammond, J. Bransford, P. LePage, K. Hammerness, & H. Duffy (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 358–389). Jossey-Bass.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Henderson, M., Ajjawi, R., Boud, D., & Molloy, E. (Eds.). (2019). *The impact of feedback in higher education: Improving assessment outcomes for learners*. Springer International Publishing. https://doi.org/10.1007/978-3-030-25112-3

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, J., Dai, W., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys.*, *55*(12), 1–38. https://doi.org/10.48550/arXiv.2202.03629

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103, https://doi.org/10.1016/j.lindif.2023.102274

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, https://doi.org/10.1016/j.lindif.2023.102274

Kipp, Michael (2023). *Wie sag ich's meiner KI? Hintergründe und Prinzipien zum #Prompting bei #ChatGPT* [video]. https://www.youtube.com/watch?v=cfl7q1llkso&t=2382s

Krause, G. (2019). Training zur Förderung von Kompetenzen für die Arbeit mit Videofeedback. In G. Uhde & B. Thies (Eds). *Kompetenzentwicklung im Lehramtsstudium durch professionelles Training* (pp. 83–108). https://doi.org/10.24355/dbbs.084-201901231126-0

Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering, *The Journal of Academic Librarianship*, *49*(4). https://doi.org/10.1016/j.acalib.2023.102720.

Lu, H.-L. (2010). Research on peer-coaching in preservice teacher education – A review of literature. *Teaching and Teacher Education*, *26*(4), 748–753.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–144). Lawrence Erlbaum Associates.

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive feedback model. *Digital Education Review*, *23*.

Pekrun, R., Marsh, H. W., Elliot, A. J., Stockinger, K., Perry, R. P., Vogl, E., Goetz, T., van Tilburg, W. A. P., Lüdtke, O., & Vispoel, W. P. (2023). A three-dimensional taxonomy of achievement emotions. *Journal of Personality and Social Psychology*, *124*(1), 145–178. https://doi.org/10.1037/pspp0000448

Prilop, C. N., Weber, K., & Kleinknecht, M. (2019a). *Entwicklung eines video- und textbasierten Instruments zur Messung kollegialer Feedbackkompetenz von Lehrkräften* [Development of a video-and text-based instrument for the assessment of teachers' peer feedback competence]. In T. Ehmke, P.Kuhl, & M. Pietsch (Eds.), Lehrer. Bildung. Gestalten: Beiträge zur empirischen Forschung in der Lehrerbildung (pp. 153-163). Weinheim Basel: Beltz Juventa Verlag.

Prilop, C. N., Weber, K. E. & Kleinknecht, M. (2019). How digital reflection and feedback environments contribute to pre-service teachers' beliefs during a teaching practicum. *Studies in Educational Evaluation, 62*, 158–170. https://doi.org/10.1016/j.stueduc.2019.06.005

Prilop, C. N., Weber, K. E., & Kleinknecht, M. (2020). Effects of digital video-based feedback environments on pre-service teachers' feedback competence. *Computers in Human Behavior, 102*, 120–131. https://doi.org/10.1016/j.chb.2019.08.011

Prilop, C. N., Weber, K. E. & Kleinknecht, M. (2021). The role of expert feedback in the development of pre-service teachers' professional vision of classroom management in an online blended learning environment. *Teaching and Teacher Education, 99*(4). https://doi.org/10.1016/j.tate.2020.103276

Prilop, C. N., & Weber, K. E. (2023). Digital video-based peer feedback training: The effect of expert feedback on pre-service teachers' peer feedback beliefs and peer feedback quality. *Teaching and Teacher Education, 127*, 104099. https://doi.org/10.1016/j.tate.2023.104099

Prins, F., Sluijsmans, D., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles and preferences. *Advances in Health Sciences Education*, *11*, 289–303.

Russel, S., & Norvig, P. (2010). *Artificial intelligence – A modern approach*. Pearson Education.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, Julia., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learning and Instruction*, *83*. https://doi.org/10.1016/j.learninstruc.2022.101620

Strijbos, J.W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*(4), 291–303.

Tsai, Y.-S., Rates, D., Moreno-Marcos, P. M., Muñoz-Merino, P. J., Jivet, I., Scheffel, M., Drachsler, D., Kloos, C. D., Gašević, D. (2020). Learning analytics in European higher education—Trends and barriers. *Computers & Education*, *155.* https://doi.org/10.1016/j.compedu.2020.103933

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2024). *AI competency framework for teachers*. https://www.unesco.org/en/open-access/cc-sa

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, *252*(Part A), 124167. https://doi.org/10.1016/j.eswa.2024.124167

Weber, K. E., Gold, B., Prilop, C. N. & Kleinknecht, M. (2018). Promoting pre-service teachers' professional vision of classroom management during practical school training: Effects of a structured online- and video-based self-reflection and feedback intervention. *Teaching and Teacher Education, 76*, 39-49. https://doi.org/10.1016/j.tate.2018.08.008

Weber, K. E., Prilop, C. N. & Kleinknecht, M. (2019). Effects of blended and video-based coaching approaches on preservice teachers' self-efficacy and perceived competence support. *Learning, Culture and Social Interaction, 22*, 103-118. https://doi.org/10.1016/j.lcsi.2019.100318

Wittwer, J., Kratschmayr, L., & Voss, T. (2020). Wie gut erkennen Lehrkräfte typische Fehler in der Formulierung von Lernzielen? *Unterrichtswissenschaft*, *48*(1), 113–128. https://doi.org/10.1007/s42010-019-00056-5

Wu, Y., & Schunn, C. D. (2021). From plans to actions: A process model for why feedback features influence feedback implementation. *Instructional Science*, *49*(3), 365–394.

Zamfrescu-Pereira, J. D., Wong, R., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany (pp. 1–21). ACM. https://doi.org/10.1145/3544548.3581388

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouveneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1). https://doi.org/10.1186/s41239-019-0171-0

Zhu, M., Liu, O, L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*. https://doi.org/10.1016/j.compedu.2019.103668

Zottmann, J. M., Stegmann, K., Strijbos, J.-W., Vogel, F., Wecker, C., & Fischer, F. (2013). Computer-supported collaborative learning with digital video cases in teacher education: The impact of teaching experience on knowledge convergence. *Computers in Human Behavior*, *5*, 2100–2108.