# From Vision to Practice: AI-Supported Feedback As the Key to Scalable Competence Orientation in University Teaching

Armin Egetenmeier[0009−0001−3944−3277], Tamara Rachbauer[0000−0002−0178−5917]
and Sven Strickroth[0000−0002−9647−300X]

**Abstract** E-Portfolios provide a flexible and competence-oriented assessment format that supports students' reflection processes. However, in larger courses individual feedback is not scalable. The use of artificial intelligence (AI), for example in the form of specific chatbots, can support teachers in providing formative feedback. This study examines the perceived usefulness and pedagogical quality of Generative AI-supported feedback systems as reported by instructors and students. The results demonstrate how AI-generated feedback can serve as a bridge between theoretical vision of personalized, scalable education and its practical implementation. The results indicate that AI-generated feedback is perceived as well-structured and helpful by participants. However, its usefulness is dependent on careful prompt engineering and adaptation to specific disciplinary contexts. This study utilises an analysis of quantitative metrics and qualitative assessments from educators and students to indicate that AI-supported feedback reduces instructor workload (e. g. in feedback provision) and supports competence development when implemented within a well-designed pedagogical framework. The integration of AI within competence-oriented education has the potential to enhance its scalability. However, to ensure contextual sensitivity, the incorporation of a human-in-the-loop approach remains a critical component in such endeavours.

Armin Egetenmeier
LMU Munich, Oettingenstr. 67, 80538 Munich, Germany e-mail: armin.egetenmeier@ifi.lmu.de

Tamara Rachbauer
University Passau, Innstraße 25, 94032 Passau, Germany e-mail: tamara.rachbauer@uni-passau.de

Sven Strickroth
LMU Munich, Oettingenstr. 67, 80538 Munich, Germany e-mail: sven.strickroth@ifi.lmu.de

# 1 Motivation

An increased application of AI-based tools can be seen in higher education to support scalable teaching methods, for example in Computer Science (CS) [1]. Large classes, which are common in CS, pose challenges such as a very heterogeneous audience and limited interaction options to provide personal feedback. This makes it difficult to address individual learning needs and often reduces feedback opportunities to automated or peer-based approaches (cf. [2, 3]). Providing meaningful feedback in competency-based teaching formats such as E-Portfolios in a scalable way is a key challenge, especially in large courses. Although E-Portfolios traditionally promote students' reflective skills, the high level of support required for manual formative feedback often pushes teaching staff to their capacity limits. AI-based tools such as customized chatbots offer innovative solutions for this case (see [4]). These tools not only enable students to engage individually with course content in greater depth, but also automatically generate feedback on student work (i. e., E-Portfolio submissions) – thus significantly relieving the burden on teaching staff. Chatbot assistance systems have already been used successfully in various projects (e. g., [5]), and automated feedback is also being provided on external (commercial) platforms in compliance with data protection regulations [6]. This helps to overcome logistical hurdles and opens up new opportunities for personalized learning processes in higher education.

The present study aims to analyse the perception and acceptance conditions of AI-generated feedback by educators and students, and to investigate its quality in direct comparison with human expert feedback from educators. Here, acceptance is inspired by the Technology Acceptance Model (TAM, [7]), focusing on perceived usefulness and ease of use. Specific prompt designs were iteratively improved within a course to increase the informative value of the AI-based feedback. The findings provide insight into the potential and limitations of AI tools, as well as concrete recommendations for their didactically meaningful use. This chapter focuses on the following research questions (RQ):

- (RQ1) How do AI and expert feedback differ in terms of correctness, comprehensiveness, and context sensitivity?
- (RQ2) Under what conditions is AI feedback rated as helpful by both students and educators?
- (RQ3) How can the quality of feedback be improved by making targeted adjustments to the prompts?

This chapter examines the implementation and perceived usefulness of AI-supported feedback systems in a lecture series on artificial intelligence with different lecturers by students and educators. The contribution of the chapter is threefold. Based on a case study conducted at LMU Munich, Germany, this research demonstrates how AI-generated feedback differs from expert feedback, how the stakeholders, both students and experts, perceive this feedback, and how the quality of the feedback can be improved. For practitioners, the chapter provides insights into how AI can support educators in providing scalable feedback, how prompts need to be adjusted in order to support educators and students, and what aspects need to be considered to integrate scalable feedback options in E-Portfolio approaches. For researchers, the chapter shows how AI-based responses differ from expert feedback based on specific disciplinary contexts and how educators and students reflect on the use of AI-based feedback, which can enhance the adoption process of such feedback.

## 2 Related Research

Competence orientation represents a paradigm shift in higher education from traditional content-focused teaching to learning outcomes-oriented approaches that emphasize students' ability to apply knowledge in complex situations [8]. According to the European Qualifications Framework for Lifelong Learning [8, 9], competence focuses on students' demonstrated ability to apply knowledge, skills, and methodological abilities in new, authentic contexts. This competence-orientation prioritizes critical reflection, transfer of knowledge, and active problem-solving rather than the simple reproduction of facts. This definition highlights the multidimensional nature of competence, which extends beyond cognitive knowledge to include practical skills, attitudes, and the ability to transfer learning across contexts. Structured formative feedback – supported by both educators and AI tools – directly addresses these capacities by challenging students to make explicit links between theory and their own disciplinary practice. The implementation of competence-oriented teaching requires careful attention to what Biggs termed ”constructive alignment” — the coherent relationship between intended learning outcomes, teaching activities, and assessment methods [10, 11, 12]. This alignment ensures that students understand what is expected of them and that their learning activities are directly connected to the competencies they are expected to develop. However, achieving such alignment at scale presents significant challenges, particularly in the provision of timely, personalized feedback that supports competence development [13].

Research has identified several key components essential for effective competence-oriented education (cf. [13, 9, 14, 15]): First, learning objectives must be clearly articulated in terms of what students should be able to do upon completion of their studies, moving beyond simple knowledge recall to application, analysis, and synthesis (cf. Bloom's taxonomy, [16]). Second, assessment methods must be designed to evaluate these higher-order competencies rather than just factual knowledge. Third, feedback mechanisms must support students' reflection processes and help them understand how to improve their performance in authentic contexts. The emergence of artificial intelligence technologies, particularly large language models (LLM), has opened new possibilities for addressing scalability challenges inherent in competence-oriented education [1, 17, 18, 19, 20]. AI-supported feedback systems leverage natural language processing and machine learning algorithms to analyze student work and provide personalized responses that can guide learning and development. These systems offer several potential advantages over traditional feedback methods, including consistency, scalability, and the ability to provide immediate responses to student work. Recent research has demonstrated the potential of AI feedback systems across various educational contexts [19, 21, 22].

However, there are typical problems of AI-based feedback that need to be considered. For example in Computer Science education, different LLM systems have been evaluated to give feedback on programming tasks. Research shows that LLM systems can be useful to identify student problems (e. g., in programming tasks [22]) and provide helpful feedback. However, in a study by Azaiz et al., the authors discovered that open AI models (such as Llama) may struggle to provide accurate feedback, with some responses being entirely incorrect [23]. Based on these experiences, the decision for a proper AI model has a significant impact on the results.

Studies have shown that AI-generated feedback can achieve high levels of accuracy when compared to human expert assessments [21, 24]. Furthermore, AI feedback systems have been shown to provide more structured and detailed responses than human feedback in certain contexts, while offering concrete suggestions for improvement that students find actionable [15]. However, the integration of AI feedback systems into educational practice is not without challenges. Several limitations of current AI feedback systems has been identified, including their tendency toward generic responses, limited contextual sensitivity, and difficulties in addressing domain-specific nuances [15, 19]. Thus, prompt engineering has gained increased attention with the rise of LLM models [25, 26, 27]. For example, Donkor proposed a cross-domain evaluation framework including strategies to refine prompts and reduce biases in responses based on stereotype reinforcement [28].

The successful implementation of AI feedback systems depends not only on their technical capabilities, but also on their acceptance by both students and educators. TAM provides a useful framework for understanding the factors that influence the adoption of new technologies in educational contexts [7, 29]. According to TAM, technology acceptance is primarily determined by two factors: perceived usefulness and perceived ease of use. These factors, in turn, influence attitudes toward the technology and ultimately determine whether it will be adopted and used effectively [29]. Research examining student acceptance of AI feedback systems has produced mixed results [15, 30]. While some studies report generally positive attitudes toward AI assistance in learning, acceptance tends to decrease when students are asked about specific implementations in their own courses [21]. This suggests that successful implementation requires careful attention to how AI systems are introduced and integrated into existing pedagogical practices. Factors that appear to influence acceptance include the perceived quality and relevance of AI feedback, the extent to which it complements rather than replaces human interaction, and students' prior experience with AI technologies.

In summary, there is still research necessary in analyzing AI-based feedback in different domains, especially in comparison with a human expert. In particular, in competence-oriented educational settings where AI feedback is less common.

## 3 Research Context

Our research was conducted within a lecture series titled ”AI in Science and Society” during the winter semester 2024/25 at LMU Munich, which is part of a major minor in Artificial Intelligence (for bachelor students). The course structure consisted of eleven lecture sessions, with nine delivered in person by educators from different disciplines (e. g., radiology, cosmology, social science, computer science) and two presented through recorded conference presentations. The course was designed as an interdisciplinary introduction to artificial intelligence, targeting students from humanities and social sciences backgrounds. The design of the lecture series was chosen to expose students to diverse perspectives on AI applications from different fields of study and their implications. Each week, an educator from a different field held a lesson about an AI topic in his/her research area, either in German or English. Covered topics included, for example, history of AI, AI Ethics in Practice, and Bias and Fairness in AI. Additionally, immediately after the lesson, the instructor offered a seminar to deepen the understanding of the topics covered. The design of the lecture and seminar was the responsibility of the respective educator.

Approximately 90 students from various fields of study were registered for the course within the learning management system Moodle. Students had access through Moodle to a web-based collaborative editor (Etherpad Lite) and a short weekly survey – for each lecture session. The participation in the survey and Etherpad was voluntary, but emphasized weekly in the sessions. Central to the course design was the integration of the AI-based system, providing feedback to the responses collected using Etherpad. For each lecture session, students were presented with three reflection questions (e. g., concerning relevant aspects for their field of study, the practical application of the discussed concepts, and any open questions resulting from the lecture) unless instructors had incorporated alternative reflection activities or questions into their presentations. This setup embodies a competence-oriented group E-Portfolio approach by encouraging students to reflect regularly on their learning and connect lecture content to real-world or their individual academic context. Providing feedback on the students’ answers further supports this reflection process of the competence-oriented design that fosters critical thinking, deeper engagement, and continuous development through reflective writing. The core competencies targeted by these assignments included, for example, critical reflection, communication skills, and transferring AI concepts into disciplinary practice (e. g., through active participation and peer discussion in the collaborative Etherpad). These competencies were communicated to students. Furthermore, access to a customized chatbot was granted to them for both preparation and review purposes, allowing them to engage with the AI system in addition to feedback provided to the Etherpad records. The interactions of students with the chatbot are not part of this study.

The evaluation focus was on the recurring reflective questions. Students had up to five days following each lecture to contribute their responses to the collaborative document, which was then archived and shared with all participants as a PDF file. After collecting the responses, the document was forwarded to the respective educator with a request for expert feedback/opinion on the students’ answers. Simultaneously, the

AI system was used to generate feedback for each question, which was immediately provided to students with a focus on answer accuracy, level of detail, and suggestions for improvement. Students were provided with expert feedback upon its availability.

The AI feedback component (and customized chatbots) utilized OpenAI's GPT-4o model, accessed through an external platform (fobizz[1]) designed specifically for educational contexts and being compliant with European data protection regulations. The AI system-generated feedback is based on customized chatbots, which use learning material provided by the educator for each session. The provided materials included, for example, research articles or papers, presentation slides, and book chapters differing in scope, length, and language (German or English), and were also accessible to the students through Moodle. The concept of the lecture series and AI-based feedback was presented to the students in the first, introductory session.

## 4 Study Design and Methodology

The study design follows the design based research (DBR) approach [31, 32] with two iterations, allowing for continuous refinement of the course design and AI feedback prompt based on ongoing evaluation. The first iteration lasted till week 8 in the semester (roughly half of the term), after three in-person sessions and two digital lectures (in week 5 and 8). A critical component of the study design involved the collection of feedback from expert instructors on the AI-generated feedback (hereinafter named meta-feedback). After providing their expert feedback on student responses, the AI-generated feedback was sent to the instructors and they were asked to evaluate its quality. In detail, the request was to evaluate the correctness, thoroughness, appropriateness of the suggested improvement, and usefulness of the AI feedback for their work. This meta-feedback provided essential insights into the comparative strengths and limitations of AI versus human feedback from an expert perspective. Overall, the data collection included voluntary expert feedback from six out of nine educators, who held the lecture in person. Two educators did not provide any expert feedback, and in one case, the Etherpad remained unedited. Meta-feedback was collected on five occasions, while in two educators only shared brief opinions, and one educator did not respond to the request at all. Since one Etherpad was left unedited, no meta-feedback was required. Due to the multilingual nature of the course, some responses were given in German and have been translated into English for this chapter. Another component of the research involved the iterative refinement of AI prompts to improve feedback quality. In the first iteration, a generic, relatively simple prompt was used:

> "You are an expert in the field of [insert field]. Based on your knowledge base, please provide feedback on the following responses: '[insert questions and answers]'. In particular, address the following three aspects: the accuracy of the answers, the level of detail in the answers, and suggestions for improvement."

[1] https://fobizz.com/de/was-ist-fobizz/, last accessed 2025-07-31

However, analysis of early feedback revealed opportunities for enhancement. Based on the meta-feedback received from educator reviews and qualitative analysis, the second iteration involved targeted revisions to make the prompt more specific and context-sensitive. These changes were implemented starting in week 8 of the semester and formed the basis for all subsequent feedback cycles. In particular, instructions were added to ensure the AI structured its feedback by question, addressed all posed questions from students explicitly, and emphasized depth and clarity in feedback.

> **Revised prompt:** "You are an expert in the field of [insert field] of Artificial Intelligence. Based on the documents in your knowledge base, please provide feedback on the answers to the following questions: '[insert questions and answers]'. Focus in particular on depth, clarity, and completeness. If any questions are posed within the answers, please respond to them. Structure your feedback into separate paragraphs for each question."

To answer the research questions, multiple data sources were used to provide a comprehensive understanding of the AI feedback system's effectiveness. First, the AI generated feedback and the expert opinion on the Etherpad responses were analysed both quantitatively and qualitatively. Second, voluntary weekly surveys were administered to students throughout the semester to monitor their perceptions of both the course content and the feedback mechanisms. These surveys were expanded at the end of the semester to include specific questions about AI and expert feedback, allowing for an analysis of student preferences and perceptions. Third, acceptance was measured via student and instructor surveys assessing perceived usefulness, intention to use, and trust in AI-generated feedback. Finally, the meta-feedback from expert instructors was collected through e-mails and analyzed using qualitative content analysis methods based on Mayring's approach to thematic analysis [33] by two of the authors with the help of an AI system (perplexity). This analysis focused on evaluating AI feedback across several dimensions: correctness, comprehensiveness, context sensitivity, and the quality of improvement suggestions. Categories were incrementally developed based on the RQs and mentioned dimensions.

## 5 Results and Evaluation

The systematic comparison between AI-generated and expert feedback (RQ1) revealed distinct patterns in terms of content focus, structure, and pedagogical approach. AI feedback consistently demonstrated superior performance in several key areas, particularly in terms of structure and provision of concrete improvement suggestions. The AI system typically provided more detailed written responses than human experts, with feedback organized into clear categories and specific recommendations for enhancement (e. g., "Add details about the technical limitations of the Logic Theory Machine (LTM)", AI week-3). Expert feedback, however, showed greater contextual sensitivity and deeper engagement with the specific teaching context and learning objectives of individual sessions (e. g., "Ok. However, I did not explicitly mention the problems that the LTM was able to solve", expert week-3). Human experts demonstrated a more nuanced understanding of the broader pedagogical goals and were more likely to address the appropriateness of student responses within the specific disciplinary context. Quantitative analysis of feedback characteristics revealed interesting patterns in word count and structure (see Table 1). Expert feedback showed considerable variation in length and approach, with some instructors providing brief, targeted comments ("Absolutely right.", expert week-3) or a self-developed 5-point rating system (expert week-11) while others offered extensive elaboration (cf. expert week-7). AI feedback, in contrast, demonstrated greater consistency in length and outline of responses (in both iterations), though this consistency sometimes came at the expense of contextual appropriateness. Thus, certain responses appear less personalized and lack details or reasoning ("This [student] response is well-structured and covers the essential points.", AI week-11).

**Table 1** Weekly AI and expert feedback shown with reflective answers, including structure, word count (average per question), and correctness (based on meta-feedback). Missing feedback excluded.

| Week | Expert structure | AI structure | AI feedback correct? | Expert / AI #words (avg.) |
|---|---|---|---|---|
| 3 | context-sensitive, focus on correctness | bullet points, formal, 3 parts (accuracy, detail, suggestions) | fully | 21 / 88 |
| 6 | no feedback | 2 parts, suggestions/example | – | – / 144 |
| 7 | individualized, reflective | standardized, feedback on writing | – | 193 / 76 |
| 9 | detailed, adaptive, high contextual depth | bullet points, 3 parts, example/solutions | mainly | 83 / 176 |
| 10 | detailed, context-based, reflective | structured, 2 parts, detailed example, question-based | mainly | 169 / 246 |
| 11 | 5 point rating, correctness | short, reflective | partially | 1 / 32 |
| 12 | no feedback | structured, answer question-based | – | – / 116 |

The analysis of improvement suggestions revealed fundamental differences between AI and expert feedback. AI systems tended to provide generic but comprehensive suggestions that could apply broadly to similar types of responses (e. g., "However, the response could be further elaborated by including concrete examples or applications", AI week-10). Expert feedback offered more targeted suggestions that reflected understanding of the specific learning context and individual student needs. In some cases, the expert encouraged deeper student understanding by asking questions or showing appreciation for interesting responses about the lecture session (e. g., "I would really like to know about that myself!", expert week-9).

In terms of acceptance (RQ2), the end of semester student survey results revealed a complex pattern of acceptance and preference regarding AI versus expert feedback. Overall, students showed positive attitudes toward the AI feedback system (including using the customized chatbots). AI based responses already provide suitable feedback from the students' perspective and most of them felt that AI feedback was sufficiently context-specific for their needs. "However, [the Etherpad] was rarely used." as one student noted. Expert feedback was appreciated for its content depth and disciplinary specificity, while AI feedback was valued for its (expected) usefulness in exam preparation and clarification of concepts. Student recommendations for future implementations included stronger integration of AI systems (e. g., chatbots) into individual lectures. All students expressed interest in maintaining access to AI feedback tools, but showed a preference for receiving expert feedback. Notably, all surveyed students reported using the AI chatbots for exam preparation, with varying assessments of their utility. Some students found the AI tools ”definitely useful” for answering questions, deepening understanding, and self-testing, while others noted that the AI feedback ”mainly assessed the texts but contributed little content-wise”.

From the educators’ point of view, the AI feedback is partially convincing. Five instructors rated AI feedback as requiring only minor or no modification before sharing it with students, suggesting a high quality and appropriateness. Nevertheless, instructors repeatedly noted that although AI feedback was generally factually correct, the AI system had difficulty adapting their feedback to the specific learning objectives or contexts. This contextual awareness was particularly evident in expert opinions (as a replacement for the meta-feedback) that questioned the assessment criteria themselves when dealing with open-ended reflection questions (e. g., “In this specific scenario I do not find any benefit in using the LLM output, quite the opposite.”, expert week-6) or reflects awareness of their disciplinary limitations, highlighting the importance of subject-matter expertise (e. g., “I’m of course not an expert in the domain of music, so I don’t really feel qualified to correct the student’s text or suggestions.”, expert week-12). This limitation was particularly evident in cases where reflection questions were deliberately open-ended or when multiple valid approaches to a topic existed. In addition, the AI feedback was not rated helpful especially when feedback is given on specific instructional activities such as role-plays (week 15) or short questions already answered in the session (week 11). The meta-feedback also highlighted the tendency of AI systems to provide generic praise and suggestions that, while not incorrect, might not be helpful to individual students. Human experts were more likely to identify subtle misunderstandings or to redirect student attention toward more fundamental conceptual issues that the AI system might miss. Despite these limitations, experts acknowledged the potential value of AI feedback as a supplementary tool, particularly for providing initial feedback that could be refined through human review (“Where I saw potential for a human expert response to be better than the AI, I noted it.”, expert week-9). The structured nature and consistency of AI feedback were viewed as valuable characteristics that could support human expertise in feedback provision. However, reactions among some educators were mixed and they felt that the AI feedback was too impersonal or even ”soulless” (expert week-6).

Targeted adaptations of the AI instructions according to meta feedback from the educators were made to increase the level of feedback detail (RQ3). Beginning in week 8 of the semester, prompts were revised to include more specific instructions about feedback structure, depth, and format (see Sect. 4 for the prompts). This modification marked the second iteration. A side-by-side comparison of AI feedback before and after prompt adaptation illustrates improvements across multiple dimensions. The revised prompt explicitly requested attention to depth, comprehensibility, and comprehensiveness. In addition, any questions posed must be answered and the feedback organized by individual questions. This led to more detailed, structured, and context-aware AI feedback, as evidenced both quantitatively by increased word count and qualitatively by greater specificity and relevance of suggestions.

The quantitative analysis showed an increase in average feedback length, with average word count rising from approximately 102 to 142 words per response (see Table 1). In addition, the structure of AI-generated feedback became more consistent, with clearly separated paragraphs addressing different evaluation criteria and more concrete examples included. Qualitative analysis further highlighted improvements in the relevance and expected usefulness of feedback. Revised prompts resulted in AI responses that referenced domain-specific content, included practical improvement suggestions, provided nuanced critique, and performed more sophisticated analysis of student responses. For instance, AI feedback began including specific references to relevant technologies, methodological approaches, and practical applications that were absent in earlier versions (e. g. week-10). A preliminary analysis of keyword usage patterns showed that terms related to comprehensibility (e. g. clarity, precision, structure) appeared more frequently.

## 6 Discussion

Regarding RQ1, the results showed that AI-supported feedback systems can contribute noticeably to the scalability of feedback, for example, in competence-oriented education. By providing structured, detailed feedback on reflection questions and other competence-development exercises, AI systems can help address one of the primary barriers to implementing competence-oriented approaches in large-scale educational settings. However, the study reveals tendency of AI systems toward generic feedback and limited contextual sensitivity, as seen in other research (e. g., [15, 19]). This suggests that significant human oversight and intervention remain necessary to ensure that AI feedback supports rather than hinders competence development. Future developments in AI technology may address some of these limitations, but current implementations require careful integration with human expertise. Usually, AI-generated feedback is longer than the expert feedback and has a typical structure including each dimension requested (e. g. correctness, comprehensiveness). In terms of the length, this can be considered appropriate, as the usefulness of feedback does not solely depend on its extent. Research suggests that overly detailed feedback or excessively long outputs can overwhelm students [22] or may be only superficially reviewed by instructors, and thus reducing its effectiveness. The AI system offers more generic but comprehensive suggestions, which could be applied to a wide range of responses. Students might perceive AI suggestions more actionable due to their typical and familiar structure, which makes them easier to understand and somehow consistent. This may be partly explained by the involvement of multiple instructors, each providing feedback of varying quality and quantity.

Students' acceptance (RQ2) of the AI based feedback is based on various factors. In this case, acceptance reflects subjective evaluations along perceived usefulness, intention to use, and trust (in accordance with the TAM framework) and does not represent actual behavioral adoption or long-term engagement. Three of the students who participated in the survey perceived the AI feedback as a valuable aid for deepening understanding, while one student regarded it as offering only superficial commentary with limited value. This variation in student experience highlights the importance of clear expectations about AI feedback systems. However, the willingness to continue using AI feedback tools despite acknowledged limitations suggests that students recognize their potential value when appropriately implemented. Furthermore, students appreciate feedback from both experts and AI, recognizing the unique strengths of each (i. e., consistency, context sensitivity). To address the limitations of AI-generated feedback (e. g., domain adaption), however, human involvement remains crucial. This finding suggests that effective, scalable feedback systems may require integration of both AI and human elements rather than replacement of one with the other (cf. [15]). This highlights the value of an AI-supported human-in-the-loop approach in this setting. The meta-feedback of the experts offered interesting insights into the practical utility of AI-generated feedback from a pedagogical perspective. The mixed reactions from educators' point of view about using AI feedback show their concerns e. g. in terms of the capabilities of AI systems or appropriateness in the specific teaching contexts. Full acceptance among educators is therefore not guaranteed, but may only develop through practical experience and targeted domain-specific adjustments. Moreover, providing guidance and training is considered essential for fostering acceptance and effective use for both educators and students - as emphasized in [1]. This educational component is crucial for maximizing the learning benefits of AI feedback systems and avoiding potential misunderstandings or misuse.

The improvements in feedback quality following prompt engineering modifications (RQ3) highlight the importance of this aspect of AI system implementation. The study demonstrates that the (perceived) usefulness of AI feedback depends not only on the underlying educational technology, but also on the skill and precision with which they are instructed. The success of prompt engineering requires an understanding of both the capabilities of available AI systems and the pedagogical goals of the educational context. This interdisciplinary collaboration is essential for developing prompts that can achieve the dual goals of technical effectiveness and pedagogical appropriateness. Although valuable feedback can be provided with a simple prompt, more effective prompts must strike a balance: they need to be specific enough to guide AI responses appropriately, yet flexible enough to accommodate diverse student needs and answer types. This balance requires iterative refinement based on empirical evidence of feedback quality and perceived usefulness. Therefore, involving opinions from all relevant stakeholders could be useful for further development. The study's findings suggest that prompt engineering should be viewed as an ongoing process rather than a one-time implementation task (cf. [21]). As AI systems evolve and as understanding of effective feedback practices develops, prompts may need regular updating and refinement. The differences shown in the second iteration suggest that AI systems can be effectively guided to emphasize particular aspects of feedback when given appropriate direction. This requires institutional investment in technical expertise and pedagogical development to ensure appropriate system performance. Continuous fine-tuning is essential to ensure high-quality and personalized feedback across disciplines and underscores the need for precise, context-aware instructions.

The findings reveal interesting implications for institutions considering the implementation of AI-supported feedback systems. Successful integration requires investment in suitable AI technologies as well as training for both educators and students to establish realistic expectations regarding the benefits and limitations of AI use. While AI systems can provide valuable support for feedback provision, their success depends on human expertise (e. g., in prompt engineering) and requires at the same time careful integration with other pedagogical strategies to achieve meaningful learning outcomes. These efforts are also associated with various costs. In the study, the AI system and chatbots were based on a commercial Generative AI model (GPT 4o), which caused additional costs for the university. However, the use of open and freely available models, such as Llama or DeepSeek, could be considered as alternatives [23]. But selecting an open model may need some more testing to decide which ones are appropriate. Even if these are not as good as commercial models, they still could be a sufficient alternative. Especially when combined with a human-in-the-loop approach, such models may still offer a scalable and cost-effective solution for providing feedback.

A noteworthy response from one educator indicated an active reflection process (”The feedback was helpful in that I realized I should have been clearer at certain points.“, expert week-3). By reviewing the AI-generated feedback in the context of the meta feedback request, the educator gained new insights into their own teaching practice. This illustrates how AI feedback systems can support not only student learning but also instructor reflection on teaching practices, aligning with the principles of the Scholarship of Teaching and Learning [34].

## 7 Threats to Validity

The study has several limitations, including the relatively small number of participants (up to 8 participants of approx. 90 students enrolled) who actively provided responses via Etherpad Lite. Only a small number of students actively engaged with the reflective questions (with a median of $n = 2$ responses per week). As a result, the findings of the AI feedback on the Etherpad results may reflect the perspectives of only a few individuals. However, this aligns with the anticipated E-Portfolio approach and provides initial indications that AI-supported feedback can be effectively applied to individual responses. Additionally, the weekly survey only included up to two responses, with the exception of the final week with ($n = 4$) responses. Despite the low number, these responses showed valuable insights which are used to improve the lecture series and the provided AI tools. Furthermore, the overall cohort size was relatively small. Nevertheless, this setting is suitable for initial testing, especially considering that student enrollment in this lecture series is expected to grow in future semesters as this is part of a newly introduced study program.

LLMs are probabilistic models and, therefore, may provide a different response for every request. This might affect the quantitative analysis, especially the response length, which only included one response from the AI feedback. However, the one-shot approach of the study should be acceptable, since the feedback is complemented by human expert feedback. Additionally, the results and responses may depend on the language, quantity or quality of the provided material to customize the AI system. There was no restriction on the material, which resulted in different documents provided for the database. However, the quality of the LLM feedback remained relatively consistent, highlighting the robustness of the AI system despite the variability in input materials. The wording of the prompts might work in the used scenario, however, a generalization or transfer to other AI models might be difficult. Thus, the refinement of prompts is still necessary.

Design-based research faces several inherent challenges such as the strong context dependence which also limits the generalizability of findings beyond the specific setting. The iterative approach makes replication difficult. Nonetheless, this methodology was particularly appropriate given the exploratory nature of the research and the need to optimize AI prompt design based on empirical evidence.

The data collection from the experts included only half of the educators. However, this already captured a good range of academic disciplines and diverse perspectives on AI-generated feedback. A limitation of the collected meta-feedback is the lack of standardization, although the request e-mail was formulated identical. This resulted in diverse responses from the expert with different focus showing their individual priorities and disciplinary perspective. In order to provide more comparable results, the use of a structured approach could be helpful. For example, a scoring system applied by expert in week-11, even if it may limit the variety and complexity of expert responses.

## 8 Conclusion and Future Work

This study of AI-supported feedback in competence-oriented university teaching provides valuable insights into both the potential and limitations of educational technologies. The research demonstrates that AI feedback systems can indeed contribute to the scalability of personalized education and feedback, but their usefulness depends critically on careful implementation, prompt engineering, and integration with human expertise.

The results show that AI feedback systems are effective in providing structured responses to student work, which is particularly valuable in large-scale educational settings where personalized, timely feedback has been difficult to achieve. The ability of Generative AI systems to provide detailed improvement suggestions and maintain consistency across multiple evaluations represents a significant advancement in addressing the scalability challenges of competence-oriented education. However, the study also identifies important limitations in current AI feedback systems, particularly regarding contextual sensitivity and pedagogical appropriateness. LLMs tend to provide generic responses that, while factually correct, may not address important learning opportunities for individual students or align optimally with learning goals.

The importance of prompt engineering emerged as a key finding, with systematic improvements in feedback quality following more specific and structured prompt design. This highlights the need for institutional investment in technical expertise and ongoing system refinement rather than simple adoption of existing AI tools. Successful implementation requires collaboration between technical and pedagogical experts to develop context-specific solutions that align with learning goals.

The study's findings support a human-in-the-loop approach that leverages the complementary strengths of AI and human expertise. AI feedback systems demonstrated clear advantages in terms of structure and scalability, while human expert feedback is effective in delivering contextual sensitivity and pedagogical appropriateness. However, the development of effective approaches requires careful consideration of how AI and human feedback can be integrated to maximize their respective strengths. From a practical perspective, the research provides evidence that students and educators can accept and benefit from AI feedback systems when they are appropriately implemented and positioned as supplements to rather than replacements for human expertise. Student appreciation for the timely availability and structured nature of AI feedback, combined with their continued preference for expert feedback in certain contexts, suggests that AI-supported human-in-the-loop approaches may be most successful.

Looking forward, several areas need further investigation. First, research is needed to explore integration strategies for AI-supported human-in-the-loop feedback models, including timing, sequencing, and presentation of AI and human feedback. Second, investigation of domain-specific adaptations of AI feedback systems could help address current limitations in contextual sensitivity. Third, longitudinal studies examining the impact of AI feedback on actual competence development (e. g. AI literacy) and learning outcomes would provide valuable evidence of educational effectiveness. Additionally, research into the training and support needs of faculty members implementing AI feedback systems could help optimize institutional adoption and usefulness. Understanding how to best prepare educators to work effectively with AI feedback systems and how to integrate such systems into existing pedagogical practices represents an important area for future investigation. The study suggests the need for continued investigation of prompt engineering techniques and their impact on feedback quality, and also different AI models, especially open ones. As AI technologies continue to evolve, understanding how to optimize their performance for educational applications will remain an important area of research and development.

In conclusion, this research demonstrates that AI-supported feedback systems represent a promising tool for addressing the scalability challenges of competence-oriented university teaching, but their success depends on thoughtful implementation that recognizes both their potential and limitations. The future of feedback in higher education likely lies not in choosing between AI and human expertise, but in developing sophisticated integration strategies that leverage the strengths of both approaches to support student learning and competence development at scale. The transformation from vision to practice in AI-supported feedback requires sustained institutional commitment, ongoing technical and pedagogical development, and realistic expectations about the role of technology in education. When implemented thoughtfully, these systems can help scale competence-oriented education and improve accessibility, while still ensuring the level of quality and personalization essential for educational effectiveness.

**Acknowledgment:** This research was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DHBKI013 (AIM@LMU).

## References

[1] J. Prather, J. Leinonen, N. Kiesler, J.G. Benario, S. Lau, S. MacNeil, N. Norouzi, S. Opel, V. Pettit, L. Porter, B.N. Reeves, J. Savelka, D.H. Smith IV, S. Strickroth, D. Zingaro, in *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 2* (ACM, New York, NY, USA, 2024), ITiCSE 2024, p. 771–772. DOI 10.1145/3649405.3659534

[2] D. Nicol, in *Approaches to assessment that enhance learning in higher education* (Routledge, 2014), pp. 11–27

[3] S. Strickroth, in *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2023)* (Association for Computing Machinery, 2023), pp. 498–504. DOI 10.1145/3587102.3588802

[4] T. Schmohl, A. Watanabe, *Künstliche Intelligenz in der Hochschulbildung: Chancen und Grenzen des KI-gestützten Lernens und Lehrens* (Springer VS, 2022)

[5] P. Zhang, G. Tur, Open Praxis **16**(3), 429 (2024)

[6] T. Rachbauer, fnma Magazin (2023)

[7] V. Venkatesh, F.D. Davis, Management Science (2000)

[8] European Commission. Eqf level descriptors (2023). URL https://www.ehea.info/Upload/TPG_A_QF_RO_MK_1_EQF_Brochure.pdf

[9] European Qualifications Framework. Lifelong learning competence framework (2023). URL https://www.ehea.info/Upload/TPG_A_QF_RO_MK_1_EQF_Brochure.pdf

[10] J. Biggs, C. Tang, *Teaching for Quality Learning at University*, 5th edn. (Open University Press, 2022)

[11] J. Biggs, HERDSA (1999)

[12] Queen Mary University of London. Implementing constructive alignment (2016). URL https://www.qmul.ac.uk/queenmaryacademy/educational-development/curriculum/implementing-constructive-alignment/

[13] J. Biggs (ed.), *Teaching for Quality Learning at University: What the student does*, 1st edn. (Open Univ. Press, Buckingham, 1999)

[14] J. Biggs, *Constructive Alignment: From Theory to Practice* (Open University Press, 2022)

[15] T. Rachbauer, fnma Magazin **1**, 30 (2025). Schwerpunkt: KI-Zugänge für Hochschulen. Peer-Reviewed

[16] D.R. Krathwohl, Theory into Practice **41**(4), 212 (2002)

[17] X. Li, Y. Wang, W. Zhang, J. Liu, Computers & Education: Artificial Intelligence **4**, 100155 (2023). DOI 10.1016/j.caeai.2023.100155

[18] W. Holmes, I. Tuomi, M. Bond, UNESCO Digital Library (2022). URL https://unesdoc.unesco.org/ark:/48223/pf0000376705

[19] Y. Xu, C. Chen, M. Li, Frontiers in Education **9**, 1203456 (2024). DOI 10.3389/feduc.2024.1203456

[20] K. Connect. Ai and competency-based education: Scaling personalized learning with large language models. White Paper (2025). URL https://www.k20connect.org/ai-cbe-whitepaper

[21] T. Rachbauer, in *KI-Use-Cases in der Hochschullehre* (Hochschulforum Digitalisierung, 2025). URL https://hfd.digital/ki-use-case-katalog

[22] I. Azaiz, N. Kiesler, S. Strickroth, in *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (Association for Computing Machinery, New York, NY, USA, 2024), ITiCSE 2024, p. 31–37. DOI 10.1145/3649217.3653594

[23] I. Azaiz, N. Kiesler, S. Strickroth, A. Zhang. Open, small, rigmarole – evaluating llama 3.2 3b's feedback for programming exercises (2025). URL https://arxiv.org/abs/2504.01054

[24] OpenAI, Gpt-4 technical report. Tech. rep. (2023). URL https://cdn.openai.com/papers/gpt-4.pdf

[25] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Advances in neural information processing systems **33**, 1877 (2020)

[26] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, ACM computing surveys **55**(9), 1 (2023)

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, et al., Advances in neural information processing systems **35**, 24824 (2022)

[28] P. Donkor, Proceedings of the AAAI Conference on Artificial Intelligence **39**(28), 29573 (2025). DOI 10.1609/aaai.v39i28.35329. URL https://ojs.aaai.org/index.php/AAAI/article/view/35329

[29] F.D. Davis, Management Science (1989)

[30] e-teaching.org. Constructive alignment in curriculum design (2024). URL https://www.e-teaching.org/didaktik/konzeption/constructive-alignment

[31] T. Anderson, J. Shattuck, Educational researcher **41**(1), 16 (2012)

[32] D.B.R. Collective, Educational researcher **32**(1), 5 (2003)

[33] P. Mayring, Forum Qualitative Sozialforschung / Forum: Qualitative Social Research **1**(2) (2000)

[34] S. Minocha, T. Collins. Impact of scholarship of teaching and learning: A compendium of case studies. DOI 10.21954/OU.RO.000155C0