Jauhiainen and Garagorry Guerra *Research and Practice in Technology Enhanced Learning* (2026) 21:27
https://doi.org/10.58459/rptel.2026.21027

Research and Practice in Technology Enhanced Learning

RESEARCH Free and Open Access

Generative AI in educational processes: ChatGPT-4 in providing feedback to students' written responses

Jussi S. Jauhiainen [1 2] * and Agustín Garagorry Guerra [1]

*Correspondence:
jusaja@utu.fi
[1] Department of Geography and Geology,
University of Turku,
Vesilinnantie 5, Turku, 20014,
Finland
[2] Institute of Ecology and the Earth Sciences,
University of Tartu,
Vanemuise 46, Tartu, 50009,
Estonia
Full list of author information is available at the end of the article

Abstract

This article examined the use of ChatGPT-4 in offering written feedback to students regarding their open-ended responses in written exams. Proper feedback is crucial for learning, helping students to understand their strengths, identify areas for improvement, and devise strategies and practices for future learning. However, crafting detailed and systematic feedback is a time-intensive task for teachers. Consequently, there is growing interest in educational circles to leverage generative AI and Large Language Models (LLMs) like ChatGPT for facilitating feedback provision as part of adaptive learning environments. For this article, ChatGPT-4 was employed to generate feedback for university students' written exams. It familiarized itself with evaluation guidelines, three short articles as learning materials in English and related 54 student responses, which varied in length from 24 to 256 words in English. Then it evaluated these responses and provided each student with personalized feedback that was on average 64 words long. The findings suggest that ChatGPT-4 has the potential for providing systematic and constructive feedback on students' written exam responses. ChatGPT-4 need to be instructed with precisely crafted prompts to ensure the feedback is precise and consistent and aligns with teacher's and educational institution's objectives to support students in learning.

Keywords: ChatGPT, Feedback, Education, Assessment, Generative AI

Introduction

Adaptive learning platforms, driven by generative AI, are at the forefront of personalized education. They provide customized delivery of learning materials, assessments, and feedback to meet the unique needs of each student. By adjusting the content and pacing based on individual performance and requirements, these platforms support engaging and

motivating educational experiences (Adiguzel et al., 2023; Bahroun et al., 2023; Baidoo-Anu & Owusu Ansah, 2023; Dai et al., 2023; Fagbohun et al., 2024; Jauhiainen & Garagorry Guerra, 2023, 2024c; Kaplan-Rakowski & Grotewold, 2023; Lo, 2023; Yu & Guo, 2023).

Feedback in education, crucial for guiding student performance and improvement, can be formative, offered during the learning process, or summative, provided after a learning period. It helps students recognize their strengths, pinpoints improvement areas, and devises strategies for future learning. Without constructive feedback, students may struggle to effectively reflect on and adjust their learning behaviors (Du Toit, 2012; Hattie & Timperley, 2007; Henderson et al., 2019; Morris et al., 2021; Panadero et al., 2023).

Educators continuously seek cost-effective methods to enhance student performance through feedback. However, providing detailed and systematic feedback is resource-intensive. The challenge intensifies as teachers must cater to many students with diverse needs and contents (Donovan et al., 2021; Morris et al., 2021; Wang et al., 2023). Human-generated feedback also presents challenges, such as variability due to subjectivity, inconsistency, or evaluator fatigue (Demszky et al., 2023; Mizumoto & Eguchi, 2023; Wang et al., 2023). Thus, there is an increasing interest in using generative AI and Large Language Models (LLMs) like ChatGPT to aid feedback provision (Jukiewicz, 2024; Steiss et al., 2024; Wilson & MacArthur, 2024; Xie et al., 2024). Despite rapid advancements of LLMs, most current insights are based on less capable models like GPT-3.5 (Steiss et al., 2024). Very few systematic evaluations have been conducted on newer, more powerful and accurate LLMs (Jauhiainen & Garagorry Guerra, 2024c).

ChatGPT-4's ability to provide feedback on students' written texts relies on processes powered by complex neural networks and deep learning algorithms, alongside an understanding of nuanced natural language. Central to this mechanism of being able to provide feedback is GPT's transformer architecture, a type of neural network adept at processing sequential data like text. This architecture is built on hundreds of billions of parameters, trained on diverse and extensive text collections, enhancing the model's performance and introducing new capabilities. Distributed training algorithms facilitate effective learning from large datasets (Zhao et al., 2023).

For depth and relevance in feedback, GPT is trained with broad spectrum of texts. The pretraining involves semi-supervised learning, where predicting word sequences provides implicit learning cues. This is complemented by supervised fine-tuning and reinforcement learning from human feedback, which refine the model's functions for specific tasks. It corrects inaccuracies, combining broad linguistic exposure and targeted human input to enhance language generation capabilities (Wu et al., 2023; Zhao et al., 2023). LLMs can effectively adapt to tasks specifically to a certain domain, such as automatic scoring in education, using the fine-tuning optimization technique (Latif & Zhai, 2024). This enables

the model to interpret sentence meanings and contextual nuances, offering a possibility to generate coherent and relevant feedback that is comprehensible to students and teachers. Fine-tuned LLMs can be made to do particular detailed evaluation tasks (Wu et al., 2023).

In previous articles, we explored the role of ChatGPT at different stages of the educational process. Initially, we examined ChatGPT's ability to customize learning materials for study purposes of diverse students. We also investigated, during on-going lectures, its capability to instantaneously tailor and personalize the learning content through formative assessment based on each student's on-going performance. The results indicated that the students could continue learning with adapted personified learning materials. Furthermore, we evaluated ChatGPT's proficiency in distinguishing correct and incorrect answers in both multiple-choice and open-ended questions in assessing students' learning (Jauhiainen & Garagorry Guerra, 2023, 2024a). We verified its accuracy in recalling lengthy written responses, before examining the model's effectiveness and consistency in evaluating students' open-ended responses (Jauhiainen & Garagorry Guerra, 2024b), and its ability to follow the evaluation criteria set by educational institutions when it assigned grades. Finally, we compared the most commonly used LLMs for their capacity in educational evaluation (Jauhiainen & Garagorry Guerra, 2024c).

This article explores the use of ChatGPT-4 for providing written feedback on university students' open-ended responses. Feedback provision is a key element in teaching, however, demanding substantial time and other resources from educators. In this study, students engaged with learning materials and responded to related questions in written form. The article discusses in detail the LLM-based evaluation and feedback processes, and in particular that of ChatGPT-4, highlighting the nature of the feedback the model provided to students' responses. Additionally, it analyzes the clarity of ChatGPT-4's feedback by assessing how accurately the feedback provided by the model alone can predict the students' overall grades.

The article addresses the following research questions: How does ChatGPT-4 provide feedback on students' open-ended written responses? What kind of feedback it gives to students' open-ended written responses? How well ChatGPT-4 is able to reconstruct its evaluation grades from the feedback it gave to of student responses?

Providing feedback on student's performance and generative AI

State-of-art in LLM-based feedback provision for students' open-ended responses

Early research on computer-generated feedback for students' written work primarily concentrated on automated essay scoring (AES) systems. These AES offered faster evaluations compared to human assessors but were often criticized for their limitations,

including providing overly general, lengthy, and sometimes inaccurate or irrelevant feedback.

LLMs, such as ChatGPT-4, i.e., a GPT in a chat version, have been applied not only for automated scoring but also as tools for facilitating more dynamic interactions in the writing assessment process (Latif & Zhai, 2024). The integration of LLMs into educational contexts represents a significant leap forward, fostering more inclusive and adaptive learning environments, including in evaluation and feedback processes (Dai et al., 2023; Jacobsen & Weber, 2024; Mizumoto & Eguchi, 2023; Wang et al., 2023; Xie et al., 2024).

Emerging research highlights the transformative potential of LLMs in enhancing the quality of feedback provided to students. Unlike earlier AES, LLMs are capable of offering personalized and nuanced responses by swiftly analyzing student written responses and essays to identify both strengths and areas for improvement. They can deliver targeted suggestions, examples, and strategies to help students refine their writing (Dai et al., 2023; Escalante et al., 2023; Steiss et al., 2024; Wang et al., 2023).

Studies by Bernius et al. (2022) and Mizumoto and Eguchi (2023) have highlighted the consistency of feedback generated by LLMs. LLMs excel in delivering feedback that is accurate, consistent, and rich in detail, setting a new standard for digital assessment tools (Dai et al., 2023; Escalante et al., 2023; Wang et al., 2023). This capability positions LLMs as essential instruments in the evolution of educational feedback systems, aligning with the broader shift toward personalized and technology-enhanced learning environments.

Steiss et al. (2024) conducted an in-depth analysis of feedback provided by ChatGPT-3.5, focusing on its adherence to evaluation criteria, accuracy, prioritization of key features, clarity in improvement suggestions, and use of a supportive tone. Their findings revealed that the differences between ChatGPT-3.5 and human evaluators were relatively modest in terms of overall feedback quality, while the time-saving benefits offered by the model were substantial.

From the perspective of students who have received feedback from ChatGPT, Dai et al. (2023) reported that the feedback was widely perceived as highly precise and valuable. Students appreciated the clarity and specificity of the feedback, which provided them with a better understanding of their academic progress and areas for improvement. Moreover, the feedback offered actionable and concrete advice on skill enhancement, which many students found helpful for their learning journey. Such feedback has the potential to enhance students' self-awareness and confidence in their academic abilities, fostering a more effective and personalized learning experience.

Teachers have noted ChatGPT's effectiveness in generating detailed, fluent, and coherent feedback, often surpassing human evaluations in comprehensiveness (Dai et al., 2023; Guo & Wang, 2023; Mizumoto & Eguchi, 2023). It can recognize nuances that human evaluators might overlook and provides feedback quickly and consistently across content,

organization, and language aspects. This capability not only reduces teachers' grading workload but also allows them to dedicate more time to direct student interaction and personalized instruction, ultimately enriching the overall learning experience.

However, the use of LLMs for providing feedback has faced substantial criticism. Ethical concerns have been risen about the possibility of learning materials and feedback inadvertently becoming part of the training data for these models. Ensuring the security, privacy, and ethical handling of students' responses and feedback is critical to maintaining confidentiality and preventing their use in future model training during assessment processes (Wu et al., 2023). Furthermore, the lack of transparency in how LLMs process and generate feedback has contributed to skepticism among scholars, who question whether these models can consistently provide feedback that is as comprehensive and insightful as that of experienced human evaluators (Steiss et al., 2024). We argue that much of this criticism arises from the lack of well-implemented LLM applications for feedback provision. Furthermore, the LLMs commonly employed, such as ChatGPT-3.5, were often not advanced enough to effectively manage complex tasks, such as thoroughly evaluating open-ended student responses and delivering detailed, nuanced feedback on them.

To address these challenges, we propose a three-stage approach to ensure that feedback provided by LLMs is accurate, consistent, and constructive. As detailed in Figure 1, discussed in the following sections, and implemented in the empirical study for this article, the process begins with the pre-evaluation stage, where the model is primed with relevant educational guidelines and prompted with detailed instructions to ensure adequate preparation. In the evaluation stage, the model processes accurate and comprehensive information to perform its assessment tasks effectively. Finally, in the post-evaluation stage, the model delivers meaningful, individualized feedback tailored to students' needs, optimizing the overall quality and impact of the feedback process.

Pre-evaluation stage

The first stage, *pre-evaluation*, is critical for establishing a reliable LLM-based evaluation process, including the feedback provision (Figure 1). It starts with prompt engineering that involves designing and formulating appropriate input prompts for LLMs like ChatGPT. This process encompasses several key steps: defining the task with specific natural language instructions for the LLM to follow, supplying input data (presenting scenarios for the LLM to respond to), incorporating contextual information (providing background documents), and determining the prompt style, which includes crafting detailed questions or instructions with prefixes and/or suffixes (Jacobsen & Weber, 2024; White et al., 2023; Zhao et al., 2023).

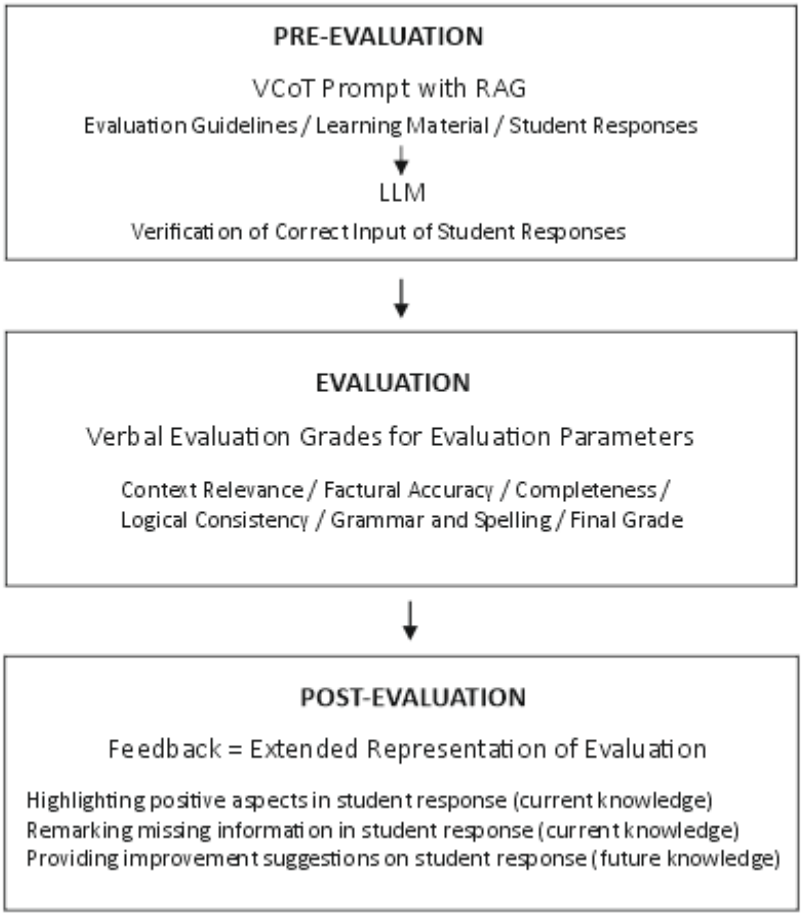

Fig. 1 Three-stage approach for evaluation process with LLM

Effective prompting is essential for guiding the LLM. There are several prompting techniques so that LLMs would produce comprehensive reasoning processes. The use of Verification-based Chain-of-Though (VCoT) prompting is to perform explicit and rigorous deductive reasoning during the prompting process. It is to ensure the trustworthiness of the reasoning process through self-verification (Ling et al., 2023). This method, which aligns with the concept of analogical prompting from Google DeepMind and Princeton University researchers, uses prompts to guide the LLM without needing manually labeled reasoning examples for each specific task. VCoT is a sophisticated, two-tiered querying method suitable for complex reasoning tasks, which prompts ChatGPT to articulate intermediate reasoning steps for each query, fostering a logical flow and deeper exploration of topics (Zhao et al., 2023).

VCoT prompting is integral to deploying ChatGPT-4 in educational assessments (for its use for this article, see below). It resembles the IDEE framework, which emphasizes

identifying desired outcomes, determining the right level of automation, upholding ethical standards, and assessing effectiveness in educational applications (Su & Yang, 2023). Furthermore, effective prompting strategies should follow the CLEAR framework by being concise, logical, explicit, adaptive, and reflective, optimizing the structure for educational evaluations (Lo, 2023).

The prompting follows with providing LLM with full access to the examination materials, including the learning content and the questions derived from it. This ensures evaluations are grounded in the content students were expected to study and address. These materials serve as the sole reference for evaluating student responses, with questions being teacher-designed, sourced from existing resources, or generated by the LLM itself (Lee et al., 2023). To assign accurate grades, the LLM requires a clear understanding of the evaluation criteria, which necessitates detailed and explicit instructions. These criteria encompass aspects such as accuracy, clarity of expression, and inclusion of key insights, ensuring alignment with the educational objectives of the examination (Irvine, 2021).

Finally, the material to be evaluated, such as students' written responses and essays, need to be input into the model. Following this, it is essential to assess the similarity between the original texts and the LLM's recalled versions. Advanced LLM models like GPT-4, trained on extensive datasets, reconstruct rather than automatically replicate responses verbatim. This can result in inaccuracies or "hallucinations," where novel content is introduced to original texts during recall (Luo et al., 2024; Su & Yang, 2023; Zhao et al., 2023). Such errors compromise the reliability and fairness of evaluations, making it vital to ensure precise recalls without omissions or distortions. Furthermore, in the recalling process, LLMs may correct spelling and grammar in responses. This may not impact the content of evaluation and feedback if the evaluation is not specifically about linguistic matters.

Mitigation strategies to prevent the presence of mistakenly recalled texts include iterative processing, rigorous testing, and verification methods, such as comparing word counts and conducting semantic analyses (Jauhiainen & Garagorry Guerra, 2024b; Luo et al., 2024; McIntosh et al., 2024). Enhanced training on targeted educational datasets and feedback mechanisms can improve accuracy, though practical implementation may be limited due to a lack of teacher awareness of it or not possessing skills to do it properly (Latif & Zhai, 2024; Luo et al., 2024). Manual checks are also necessary to address over-corrections in grammar and spelling of received texts (Fang et al., 2023; Wei et al., 2022). This stage ensures that evaluations and feedback are based on accurate representations of student work.

To maintain the content-relatedness and security in the process, one may use the Retrieval-Augmented Generation (RAG) framework. It is to ensure that LLM evaluations, such as those by ChatGPT, are accurate, secure, and ethically sound. RAG enables the

model to access specific reference materials, such as exam content and rubrics, during assessments (see Figure 1), ensuring feedback and grades align with educational objectives. By anchoring the model's output to these materials, RAG minimizes inaccuracies or "hallucinations," enhancing the reliability and relevance of feedback (Lewis et al., 2021; Wei et al., 2022; Yasunaga et al., 2023).

Additionally, RAG protects the confidentiality of sensitive data by preventing student responses or reference materials from being incorporated into the model's training data. Materials are used only temporarily, ensuring compliance with ethical standards and reducing concerns about data misuse. By grounding evaluations in external references, RAG allows LLMs to provide precise, contextually relevant assessments while addressing key privacy and reliability challenges in education.

Evaluation stage

In the second stage, *evaluation*, an LLM like ChatGPT performs a comprehensive assessment of each student's response, ultimately assigning grades (either as a word or number) for individual responses and the overall examination (Figure 1). This stage involves several critical steps to ensure a thorough, fair, and accurate evaluation process. The evaluation process with LLMs is complex. It is not discussed in detail in this article as it has been addressed elsewhere (Jauhiainen & Garagorry Guerra, 2024b, 2024c).

Shortly, the evaluation begins with the LLM accessing the examination materials and student responses. Responses are then assessed against predefined criteria such as content accuracy, argument coherence, clarity of expression, and use of evidence. A systematic grading methodology is critical, as it not only ensures fair evaluations but also significantly impacts student motivation (Chamberlin et al., 2023).

The LLM may assign grades holistically, generating a single final grade for each response, or evaluate specific criteria individually. By assessing various aspects of a response, the LLM can identify strengths and weaknesses. This is necessary for LLMs to be able to provide detailed feedback that is contextually relevant to what students were expected to understand and analyze so that the model can highlight areas for improvement and support student learning with relevant feedback.

Post-evaluation stage

The third stage, *post-evaluation*, in using LLMs like ChatGPT for educational assessment emphasizes delivering personalized feedback strictly based on student performance and evaluation guidelines (Figure 1). This stage builds on the evaluation insights to provide detailed, nuanced feedback aimed at enhancing student learning and development.

The feedback includes specific textual insights for each student, highlighting strengths, areas for improvement, and actionable suggestions (Chamberlin et al., 2023; Donovan et

al., 2021; Hattie & Timperley, 2007). By evaluating student responses against multiple criteria, the LLM generates targeted feedback for each, pinpointing how well the response aligns with the assessment standards of the educational institution and offering clear guidance for skill enhancement. This ensures feedback is not only constructive but also directly applicable, encouraging self-reflection and learning.

Unlike generic and short feedback such as "good" or "well done," that is easy to be done by any evaluator, ChatGPT delivers longer, highly personalized responses tailored to the individual's unique performance and needs (Dai et al., 2023; Mizumoto & Eguchi, 2023; Xie et al., 2024). It highlights both positive aspects and areas requiring improvement, suggesting strategies such as additional reading, argument development techniques, or writing exercises. This approach promotes deeper engagement and helps students prepare for future assessments by identifying concrete steps for improvement (Donovan et al., 2021; Morris et al., 2021).

Immediate feedback, which can be received often within seconds with LLM, further enhances its value by allowing students to promptly revise their work or discuss it with teachers and peers. This transforms the assessment into a dynamic learning process, where feedback serves not as an endpoint but as a foundation for ongoing improvement and active engagement with learning objectives.

Material and methods

For this research, we used ChatGPT-4, which is one of the most widely used LLMs in educational contexts globally. The chat-based version of GPT is particularly popular among teachers for providing feedback to students. However, many other LLMs are also available for this purpose, and their performance varies significantly (Jauhiainen & Garagorry Guerra, 2024c).

Material

The material for this study comprised written feedback generated by ChatGPT-4 in English for 54 Master of Science-level student responses related to their course readings. The student responses varied in length from 24 to 256 words, while the feedback provided ranged from 37 to 99 words, averaging 64 words per response, for a total of 3456 words of feedback.

The university-based tests were conducted in a simulated examination environment. All participants were adults who remained anonymous, and no sensitive information was collected during the study. Participation was voluntary, and students could withdraw at any time. By submitting their responses, students consented to participate in the study and agreed that their responses could be analyzed.

The students engaged with three sets of reading materials, each tied to a specific topic and based on peer-reviewed journal articles authored in English by one of the contributors to this article. The first reading, on irregular migration at the EU border, was 2543 words long. The second, focusing on managing irregular migration during the COVID-19 pandemic, contained 3734 words. The final reading, on knowledge creation processes, totaled 1816 words, bringing the combined word count of the materials to 8103.

The study involved three test rounds. In the first round, six students responded to three questions based on the first reading, with answers ranging from 31 to 256 words. A week later, six additional students addressed three questions on the second reading, with responses between 24 and 103 words. In the third round, another group of six students tackled three questions on the third reading, with responses varying from 62 to 102 words. These variations in response length, influenced by the reading materials and questions, offered valuable insights into the students' diverse perspectives and levels of understanding.

Before providing feedback, ChatGPT-4 evaluated each response and assigned a grade based on the university's six-level grading system: 0 (fail), 1 (passable), 2 (satisfactory), 3 (good), 4 (very good), and 5 (excellent). It also assessed multiple aspects of each response, including contextual relevance, factual accuracy, completeness, logical consistency, and grammar and spelling. Similar criteria have been used also in earlier research by several scholars (see Steiss et al., 2024; Su & Yang, 2023). This systematic approach ensured comprehensive and constructive feedback.

Methods

The methods in this article refer to two dimensions. The first dimension focuses on designing the model to ensure it is capable of effectively performing evaluation and feedback tasks. This involves implementing specific prompts to enhance its ability to assess content and provide constructive feedback. The second dimension pertains to the methodologies used to analyze the feedback generated by the model. This includes evaluating the quality, relevance, and impact of the feedback, as well as exploring systematic approaches to measure its effectiveness in achieving the desired outcomes.

As regards the first methodological dimension, to generate feedback, the earlier discussed three-stage approach was rigorously followed (see Figure 1). The process started with the first pre-evaluation stage, where the model is primed with relevant educational guidelines and prompted with detailed instructions to ensure adequate preparation.

ChatGPT-4 was provided with the reading materials and associated questions using the RAG framework. VCoT was essential to guide ChatGPT-4 in producing logical and contextually appropriate feedback. VCoT was tailored to the specific task and the type of interaction desired with ChatGPT, in this case regarding personalized feedback provision

for each student separately. Refining prompts based on ChatGPT's responses allowed for gradual guidance toward the intended outcomes.

VCoT enhanced the model's logical reasoning and accuracy by breaking tasks into step-by-step processes and verifying each step for consistency and correctness. It included a verification layer to cross-check intermediate prompting steps and final outputs against predefined criteria or reference materials, ensuring factual accuracy and contextual relevance (Wei et al., 2022). Liu and Shah (2023) note that specificity in prompting tends to yield better results than more generic prompts. Employing VCoT prompting enables ChatGPT to deconstruct complex tasks into simpler components, using intermediate steps to guide it toward accurate conclusions by tracing a logical path of reasoning before reaching the final answer. This iterative process minimized errors, prevented hallucinations, and produced more reliable results though it requires more computational resources and careful prompt design.

For this study, the prompts were carefully crafted to define ChatGPT-4's role as analogous to that of a university professor, responsible for providing feedback on written submissions by Master's-level students based on specified reading materials. These prompts were sequentially linked, with each new prompt building on the previous one, ensuring systematic guidance toward logical and coherent reasoning (Wei et al., 2022; White et al., 2023).

In practice, ChatGPT-4 received explicit instructions regarding its tasks, including the use of a standard grading scale from 0 (fail) to 5 (excellent), as employed by the educational institution where the study was conducted. The evaluation was to be based on five sub-criteria—context relevance, factual accuracy, completeness, logical consistency, and grammar and spelling—each weighted equally at 20%. The final grade was to be computed by multiplying each sub-criterion's score by 0.2, summing the results, and rounding to the nearest whole number between 0 and 5. The model then utilized this information to provide feedback.

We conducted prompting as an iterative process refined through our extensive earlier experiences of the use of ChatGPT-4 in various educational contexts. Following VCoT principles, different phrasings and structures were experimented with and verified until the instructions were sufficiently clear and contextually relevant. Verification involved assessing whether ChatGPT-4 could consistently perform the desired tasks; if not, prompts were adjusted for greater precision and consistency. Finely tuned prompts were employed to enhance the reliability and effectiveness of the model's educational assessments (Figure 2).

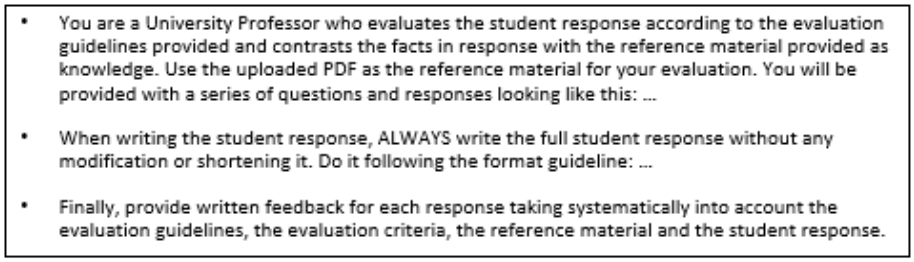

Fig. 2 Example of prompting fragment utilized to make ChatGPT-4 to perform its tasks

The second evaluation stage consisted of ChatGPT-4 processing the inserted accurate and comprehensive information, e.g., the evaluation guidelines, the learning material as well as the student written responses. Then the model executed its task of assessing and grading the responses both on each sub-criteria and the final grade, as explained above in detail.

Based on this, the model proceeded to the third post-evaluation stage. There ChatGPT-4 delivered individualized feedback to each students' response. Technically, in the TurkuEval platform (see sites.utu.fi/digileac) in which the test was conducted, designed by the authors, the model presented both the evaluation result as well as the feedback next to each other so that their direct comparison would be easy to execute. In a real-world context, a teacher would review the evaluation and feedback generated, make any necessary modifications, and subsequently enter the final results into the university's grading system while also providing the feedback to students. However, since this was a test scenario, the process did not advance to that stage.

As regards the second methodological dimension, the analysis of ChatGPT-4's feedback on student responses combined quantitative and qualitative methods, focusing on thematic content analysis. The course professor, who authored the reading materials and is an experienced evaluator, provided guidelines for assessing ChatGPT-4's feedback, with two research assistants supporting the evaluation.

Feedback analysis began by measuring its length for each feedback. For this, the word counting was used for examining the relationship between feedback length and the overall grades assigned by ChatGPT-4. Later, the content was analyzed to identify themes aligned with the five evaluation criteria used by ChatGPT-4, with each element related to these criteria counted and analyzed for its influence on overall and criterion-specific grading. ChatGPT-4 was explicitly instructed to calculate the final grade as the sum of five evaluated sub-criteria, each weighted at 20%, as specifically outlined in the prompt. The analysis also investigated whether ChatGPT-4 suggested improvements and how these correlated with assigned grades. Additionally, the tone of the feedback was assessed.

Both absolute and proportional frequencies were calculated to provide statistical insights. Absolute frequencies captured how often specific themes or elements appeared in the ChatGPT-4 feedback, while proportional frequencies contextualized their significance as percentages of the total. Correlations between feedback themes and grades were examined to identify significant relationships, with statistical tests, including p-values, used to determine the strength and significance of these correlations, ensuring the findings reflected true patterns rather than random chance.

Results

ChatGPT-4's feedback for student responses

Before providing feedback, ChatGPT-4 was tasked with evaluating and grading each of the 54 student responses using a standard grading scale from 0 (failed) to 5 (excellent), as used at the educational institution where the study was conducted. As mentioned, alongside the overall grade, ChatGPT-4 assessed each response based on five predefined criteria: context relevance, factual accuracy, completeness, logical consistency, and grammar and spelling. These criteria, commonly employed in other studies on ChatGPT's feedback provision (e.g., Steiss et al. 2024), were given equal weight (20%) in forming the final grade, with the model calculating the overall score as the sum of the scores for each criterion. This was explicitly stated in the prompt.

The prompts provided to ChatGPT-4 were designed to ensure alignment with these criteria, but they did not specify an expected total word count for the feedback. Consequently, ChatGPT-4 generated feedback of varying lengths, ranging from 33 to 99 words per response. Despite this variability, the feedback lengths were relatively consistent, with an average of 64 words and a median of 65 words (Table 1).

Analysis of feedback length relative to the grades assigned revealed no direct correlation between the two. The variation in the grade assigned did not appear to influence feedback length. For instance, responses graded as 5 (excellent) received feedback averaging 58 words, while responses graded 2 (satisfactory) had slightly longer feedback averaging 68 words. Interestingly, responses graded 3 (good) received the longest average feedback at 70 words, highlighting that feedback length alone was not indicative of the grades assigned

Table 1 Feedback length vs. grading in three written examinations

	Overall grade of response / average feedback length in words					
	1	2	3	4	5	N
Exam 1	43	61	73	56	59	18
Exam 2	70	70	69	49	-	18
Exam 3	-	72	69	45	57	18

(Table 1). This suggests that ChatGPT-4 maintained consistent quality and content across different grade levels and their feedback length.

A weak correlation was observed between the length of students' responses and the length of feedback provided by ChatGPT-4, but this relationship was not statistically significant ($p = .233$). The 10 longest feedback entries, averaging 82 words, corresponded to responses with an average overall grade of 2.9 and a median grade of 3.0. Conversely, the 10 shortest feedback entries, averaging 47 words, were associated with slightly higher average grades of 3.3 and a median grade of 3.5.

No statistically significant correlation was found between grades assigned to specific criteria and feedback length. However, longer responses were significantly less likely to receive critical comments from ChatGPT-4. The difference was statistically significant ($p = 0.034$), indicating a likely effect. While initially intending to analyze the relationship between feedback tone, length, and grade, the analysis found ChatGPT-4 predominantly used a neutral tone. Only two feedback instances (3.7%) were interpreted as having a clearly critical tone.

Among the 54 feedback instances, 70.4% included suggestions for improving responses. Responses that received improvement suggestions had an average grade of 2.7, significantly lower than the 3.8 average grade for responses without suggestions. Responses graded as satisfactory (2) were most likely to receive improvement suggestions (84.6%), while only 20.0% of responses graded excellent (5) received such suggestions. All passable (1) responses included improvement suggestions. Feedback with suggestions tended to be longer than those without suggestions. There was a very strong level of statistical significance regarding this difference ($p < .001$). This suggests that the result was highly unlikely to be due to chance.

The distribution of improvement suggestions across the five evaluation criteria revealed distinct patterns. Feedback addressing multiple criteria required more text. There was a very strong level of statistical significance regarding this difference ($p < .001$). In evaluating content relevance, ChatGPT-4 consistently noted that all responses (100%) referred directly to the learning materials, though not always consistently throughout the entire response.

Nearly all feedback (94.4%) mentioned completeness, frequently noting that responses should have included more detailed recollections from the reading materials. Comments on grammar and spelling were rare, occurring in only one instance (1.9%), reflecting the students' overall fluency in English (Table 2), or that the model corrected the language when processing the responses.

When any of the five evaluation criteria were rated low (grades 1 or 2), 85.0% of the feedback included improvement suggestions. In contrast, for criteria rated as very good or excellent (grades 4 or 5), a smaller proportion (65.2%) of feedback included such

Table 2 Feedback provided by ChatGPT-4 to 54 student responses (%)

Grade	1	2	3	4	5	average	N
Context relevance	100.0	100.0	100.0	100.0	100.0	100.0	54
Factual accuracy	100.0	53.8	87.5	80.0	100.0	79.6	43
Completeness	100.0	100.0	100.0	80.0	80.0	94.4	51
Logical consistency	50.0	61.5	75.0	90.0	100.0	75.9	41
Grammar and spelling	0.0	0.0	4.2	0.0	0.0	1.9	1
Improvement suggestions	100.0	84.6	87.5	30.0	20.0	70.4	38
Feedback words (average)	57	66	70	57	58	62	54
Received feedback	2	13	24	10	5	-	54

1=passable, 2=satisfactory, 3=good, 4=very good, 5=excellent

suggestions. This trend indicates that ChatGPT-4 focused improvement suggestions on responses with lower grades, where there was more scope for enhancement.

The findings revealed a notable instance of variability in ChatGPT-4's assessment results in one case. It highlighted inconsistencies in grading a single 104-word student response. In three separate evaluations, the same response received different overall grades: satisfactory (2), good (3), and very good (4). Correspondingly, the grades assigned to the five evaluation criteria also varied, resulting in distinct feedback for each evaluation (Figure 3). This variability is comparable to a scenario where three different human teachers assess the same response, each assigning different grades and providing unique feedback based on their individual evaluations.

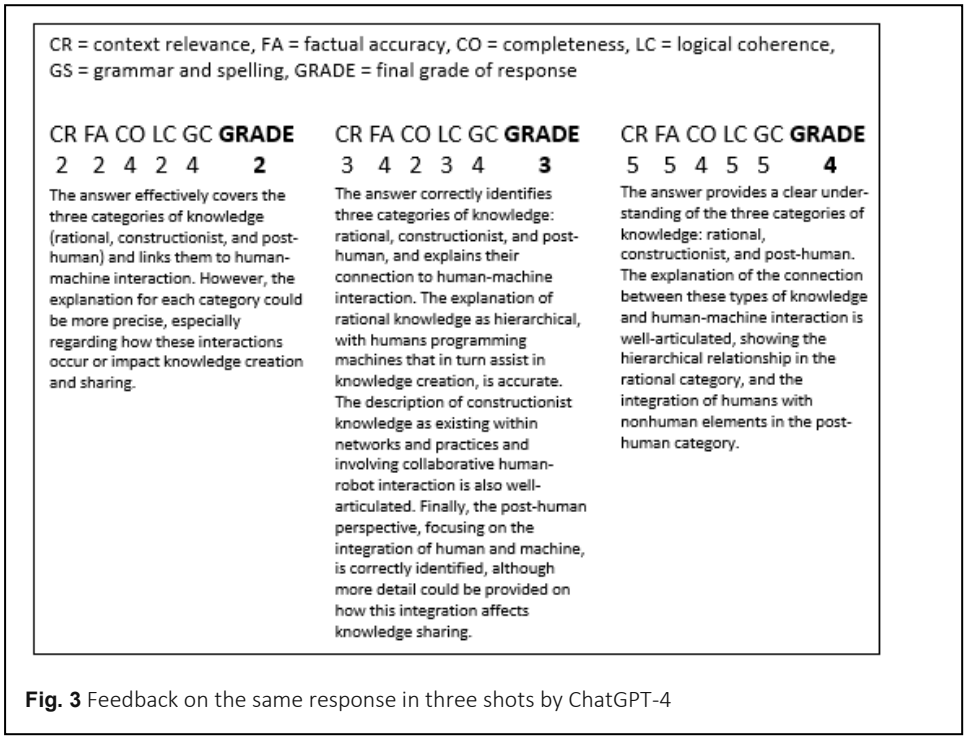

Fig. 3 Feedback on the same response in three shots by ChatGPT-4

When ChatGPT-4 graded the response as satisfactory (grade 2), the feedback was 41 words long, with individual criteria grades ranging from 2 to 4, and included suggestions for improvement. For the good (grade 3) evaluation, the feedback expanded to 86 words, maintaining criteria grades between 2 and 4, and also included improvement suggestions. However, when the response was rated as very good (grade 4), the feedback was shorter, at 57 words, with higher criteria grades of 4 or 5 and no improvement suggestions (Figure 3).

This analysis underlines that differences in ChatGPT-4's grading can significantly affect the length and content of the feedback. It also highlights the potential for variability in evaluation results across multiple assessments of the same response. This variability reflects the inherent complexity of how LLMs like ChatGPT-4 process reading materials, questions, and student responses through neural networks. While grading variability is a known limitation of LLMs, a deeper examination of this issue is beyond the scope of this article but has been addressed in other research (Jauhiainen & Garagorry Guerra, 2024b).

To conclude our analysis, we conducted an experiment to test whether ChatGPT-4 could accurately predict the overall grade of a student's response based solely on the feedback it had generated, without access to the original response or reading material. The rationale was that if the recreated grade matched or closely approximated the actual grade, it would demonstrate that the feedback content logically reflected the evaluation criteria and was internally consistent, at least to ChatGPT-4.

The results showed that ChatGPT-4 exhibited a strong ability to predict the quality of a student's response based purely on its feedback. It effectively incorporated its comments into a coherent judgment about the response's quality. For the 54 evaluated responses, ChatGPT-4's predicted grades matched or were within one grade of the actual grades in 96.3% of cases, with an exact match in 53.7% of cases (Table 3). When we modified the criteria by excluding grammar and spelling and assigning equal 25% weights to the remaining four criteria, 87.0% of the recreated grades exactly matched the actual grades, while 12.0% were one grade higher. These findings highlight ChatGPT-4's consistency in aligning its feedback with its evaluation outcomes.

Table 3 Student responses' overall grade recreation by ChatGPT-4 from feedback (%)

Original grade	Recreated grade						
	0	1	2	3	4	5	N
0	**0.0**	0.0	0.0	0.0	0.0	0.0	0
1	0.0	**0.0**	100.0	0.0	0.0	0.0	2
2	0.0	0.0	**53.8**	46.2	0.0	0.0	13
3	0.0	0.0	8.3	**70.8**	16.7	4.2	24
4	0.0	0.0	0.0	20.0	**20.0**	60.0	10
5	0.0	0.0	0.0	20.0	20.0	**60.0**	5
N	0	0	9	26	7	10	54

0=fail, 1=passable, 2=satisfactory, 3=good, 4=very good, 5=excellent

Table 4 Student response evaluation criteria recreation by ChatGPT-4 from feedback (%)

Original grade	Recreated grade						
	0	1	2	3	4	5	N
0	**0.0**	100.0	0.0	0.0	0.0	0.0	1
1	0.0	**36.8**	57.9	0.0	0.0	5.3	19
2	0.0	5.8	**57.7**	28.8	0.0	7.7	52
3	0.0	0.0	9.4	**54.2**	30.2	6.3	96
4	0.0	0.0	0.0	15.4	**30.8**	53.8	26
5	0.0	0.0	1.3	5.3	7.9	**85.5**	76
N	0	11	52	75	44	88	270

0=fail, 1=passable, 2=satisfactory, 3=good, 4=very good, 5=excellent

We further examined how accurately ChatGPT-4 could reconstruct student grades for each of the five evaluation criteria it used. To assess this, the mode grade from 10 evaluations (10-shot) was utilized. The findings reveal that ChatGPT-4 demonstrates a strong ability to consistently and accurately reconstruct grades based on specific evaluation criteria, though the accuracy varied across criteria and grade levels.

Across 270 evaluations, ChatGPT-4's reconstructed grades were within one grade of the original in 94.1% of cases, with an exact match in 60.0% of cases (Table 4). Prediction accuracy varied by grade, with grade '5' being the most accurately reconstructed (85.5% exact matches), while grade '4' was the least accurately predicted (30.8% exact matches). Accuracy also varied across criteria: context relevance and factual accuracy each had a 57.4% exact match rate, completeness had a 50.0% match rate, logical consistency was slightly higher at 64.8%, and grammar and spelling achieved the highest match rate at 70.4%. These results highlight ChatGPT-4's strong, albeit variable, capability to reconstruct evaluation grades based on its criteria.

Conclusions

Receiving personalized feedback is essential for student learning, offering targeted insights that promote development and motivation. Constructive and systematic feedback helps students recognize their strengths and areas for improvement, guiding their learning process. However, providing such detailed feedback is time-intensive for teachers, often resulting in inconsistent delivery. This inconsistency hampers student progress, as numerical grades alone lack the depth needed for meaningful improvement.

Feedback in education, crucial for guiding student performance and improvement, can be formative, offered during the learning process, or summative, provided after a learning period. It helps students recognize their strengths, pinpoints improvement areas, and devises strategies for future learning. Without constructive feedback, students may struggle

to effectively reflect on and adjust their learning behaviors (Du Toit, 2012; Hattie & Timperley, 2007; Henderson et al., 2019; Morris et al., 2021; Panadero et al., 2023).

This article examines how LLMs, particularly ChatGPT-4, can enhance education by generating feedback on students' open-ended written responses based on learning materials. It introduces a systematic framework for LLM-based feedback, covering the pre-evaluation, evaluation, and post-evaluation stages. Practical recommendations are also highlighted, such as the importance of well-crafted prompts and using verbal grading terms like "poor" or "excellent" instead of numerical scores, as LLMs process explicit language more effectively. These insights aim to improve the relevance and effectiveness of LLM-generated feedback for student needs.

Building on a limited but growing body of research on ChatGPT's feedback capabilities (e.g., Bernius et al., 2022; Dai et al., 2023; Demszky et al., 2023), this article contributes to understanding the potential of LLMs in education. While earlier studies often relied on less advanced models like ChatGPT-3.5 (e.g., Dai et al., 2023; Steiss et al., 2024), which struggled with the complexities of educational evaluation, the findings here highlight the improvements and efficiencies offered by newer versions, ChatGPT-4 in this case. However, at the development stage of LLMs in the mid-2020s, evaluation inconsistency remains a significant challenge, closely tied to feedback generation. If a model cannot accurately evaluate a response, it cannot produce precise or reliable feedback. However, continued advancements are expected to make LLMs increasingly reliable and valuable tools across diverse educational contexts.

The key findings of the article are the following:

Firstly, the model demonstrated the ability to provide constructive suggestions, particularly for lower-graded responses that required improvement. This adaptability in tailoring feedback based on the quality of the response suggests that ChatGPT-4 can serve as a valuable teaching aid, helping students refine their work in a targeted and effective manner. The quality of feedback generated by ChatGPT-4 did not strongly correlate with feedback length.

Secondly, ChatGPT-4 showed a high level of accuracy in predicting student response grade based solely on the feedback it had previously generated, even without access to the original responses or reading materials. This indicates that ChatGPT-4 can effectively internalize and apply grading criteria in relation to feedback, a key factor for its potential integration into educational assessment systems.

Thirdly, the study identified instances of variability in the grades and feedback provided by ChatGPT-4 when evaluating a particularly complex response multiple times. While the model was generally reliable, such inconsistencies are typical of generative AI systems dealing with nuanced human language. This variability highlights the importance of employing multiple evaluation attempts (multi-shot approaches) to enhance reliability

when using LLMs for grading open-ended responses. Although advancements in LLMs are expected to reduce grading variability over time, at their current developmental stage, it is strongly recommended that teachers review and verify ChatGPT's evaluations and feedback. This ensures alignment with educational standards and allows adjustments where necessary to maintain consistency and accuracy in assessment.

Building on these conclusions, future research can expand in several key directions. As generative AI, such as LLMs, becomes increasingly embedded in educational processes, ongoing studies must address ethical concerns, including privacy, transparency, and the impact of biases in data and algorithms.

Expanding the application of LLM-based feedback to a wider range of subjects and educational levels is essential to explore the versatility and adaptability of generative AI tools across diverse learning environments and varying content complexities. Comparative studies are also needed, both to evaluate the feedback capabilities of different LLMs (see Jauhiainen & Garagorry Guerra, 2024c) and to systematically compare human-generated and LLM-generated feedback, analyzing their similarities, differences, and effectiveness.

Addressing algorithmic aversion is another critical area for future research. Teachers' and students' skepticism, fear, or optimism toward generative AI-based assessment may impact their willingness to engage with feedback provided by LLMs. Surveys of teachers and students could help investigate how such perceptions influence the acceptance and effectiveness of AI-generated feedback.

Future studies could also focus on calibration methods to align LLM-based feedback with individual teachers' evaluation styles, tones, and preferences. Additionally, researchers should analyze how to improve the accuracy and consistency of these tools when handling written materials of varying quality and length. Experimental use of LLM-generated synthetic responses alongside real student-written content could further enhance understanding of these tools' capabilities and limitations.

Finally, while feedback on student responses is a central focus, future research could explore how LLMs might suggest personalized learning strategies based on individual students' needs and styles. Enhancing personalized feedback would make LLM-powered educational tools more effective and supportive throughout the learning process, ultimately benefiting both students and educators.

Abbreviations

AES: Automated Essay Scoring; LLM: Large Language Model; RAG: Retrieval-Augmented Generation; VCoT: Verification-based Chain-of-Though.

Authors' contributions

Jussi S. Jauhiainen took part in the research design, data collection and analysis, and wrote and revised the manuscript. Agustín Garagorry Guerra took part in the research design, data collection and analysis, and commented on the manuscript.

Authors' information
Dr. Jussi S. Jauhiainen is Professor of Geography at the University of Turku (Finland) and Visiting Professor at the University of Tartu (Estonia) where he has also honorary doctorate. Mr. Agustín Garagorry Guerra is research associate at the University of Turku. Both have studied and developed platforms for the use of generative AI in education (https://sites.utu.fi/digileac).

Funding
The research did not receive external funding.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
Jussi S. Jauhiainen, Department of Geography and Geology, University of Turku, and Institute of Ecology and the Earth Sciences, University of Tartu. Agustín Garagorry Guerra, Department of Geography and Geology, University of Turku.

Received: 21 June 2024 Accepted: 30 April 2025
Published online: 1 January 2026 (Online First: 8 October 2025)

References

Adiguzel, T., Kaya, M., & Cansu, F. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Education Technology*, *15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Bahroun, Z., Anane, C., Ahmed, V., & Zacca, A. (2023). Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability*, *15*(17), 12983. https://doi.org/10.3390/su151712983

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bernius, J., 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

Chamberlin, K., Yasué, M., & Chiang, I. (2023). The impact of grades on student motivation. *Active Learning in Higher Education*, *24*(2), 109–124. https://doi.org/10.1177/1469787418819728

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y., Gasevic D., & Chen, G. (2023). Can Large Language Models provide feedback to students? A case study on ChatGPT. In *Proceedings of 2023 IEEE International Conference on Advanced Learning Technologies* (pp. 323–325). IEEE. https://doi.org/10.1109/ICALT58122.2023.00100

Demszky, D., Liu, J., Hill, H., 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), 016237372311692. https://doi.org/10.3102/01623737231169270

Donovan, B., den Outer, P., Price, M., & Lloyd, A. (2021). What makes good feedback good? *Studies in Higher Education*, *62*(2), 318–329. https://doi.org/10.1080/03075079.2019.1630812

Du Toit, E. (2012). Constructive feedback as a learning tool to enhance students' self-regulation and performance in higher education. *Perspectives in Education*, *30*(2), 32–40. https://doi.org/10.38140/pie.v30i2.1757

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*, 57. https://doi.org/10.1186/s41239-023-00425-2

Fagbohun, O., Iduwe, N., Abdullahi, M., Ifaturoti, A., & Nwanna, O. (2024). Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artificial Intelligence, Machine Learning & Data Science*, *2*(1), 1–8. https://doi.org/10.51219/JAIMLD/oluwole-fagbohun/19

Fang, T., Yang, S., Lan, K., Wong, D., Hu, J., Chao, L., & Zhang, Y (2023). *Is ChatGPT a highly fluent grammatical error correction system? A comprehensive evaluation*. https://doi.org/10.48550/arXiv.2304.01746

Guo, K., & Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

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

Irvine, J. (2021). Taxonomies in education: Overview, comparison, and future directions. *Journal of Education and Development*, *5*(2). https://doi.org/10.20849/jed.v5i2.898

Jacobsen, L., & Weber, K. (2024). The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback. *AI*, *6*(2), 35. https://doi.org/10.3390/ai6020035

Jauhiainen, J., & Garagorry Guerra, A. (2023). Generative AI and ChatGPT in school children's education. Evidence from a school lesson. *Sustainability*, *15*(18), 14025. https://doi.org/10.3390/su151814025

Jauhiainen, J., & Garagorry Guerra, A. (2024a). Generative AI and education: Dynamic personalization of pupils' school learning material with ChatGPT. *Frontiers in Education*, *5*, 1288723. https://doi.org/10.3389/feduc.2024.1288723

Jauhiainen, J., & Garagorry Guerra, A. (2024b). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, 1–18. https://doi.org/10.1080/14703297.2024.2422337

Jauhiainen, J., & Garagorry Guerra, A. (2024c). Evaluating students' open-ended written responses with LLMs: Using the RAG framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large. *Advances in Artificial Intelligence and Machine Learning*, *4*, 3097–3013. https://dx.doi.org/10.54364/AAIML.2024.44177

Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity*, *52*, 101522. https://doi.org/10.1016/j.tsc.2024.101522

Kaplan-Rakowski, R., & Grotewold, K. (2023). Generative AI and teachers' perspectives on its implementation in education. *Journal of Interactive Learning Research*, *34*(2), 313–338. https://www.learntechlib.org/primary/p/222363

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Mon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technology*, *29*, 11483–11515. https://doi.org/10.1007/s10639-023-12249-8

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan & H. Lin (Eds.), *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp.9459–9474). Curran Associates Inc.

Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., & Su, H. (2023). Deductive verification of chain-of-thought reasoning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt & S. Levine (Eds.), *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 36407–36433). Curran Associates Inc.

Liu, R., & Shah, N. (2023). *ReviewerGPT? An exploratory study on using Large Language Models for paper reviewing*. https://doi.org/10.48550/arXiv.2306.00622

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Science*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., & Dukek, G. (2024). *Hallucination detection and hallucination mitigation: An investigation*. https://doi.org/10.48550/arXiv.2401.08358

McIntosh, T., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. (2024). A culturally sensitive test to evaluate nuanced GPT hallucination. *IEEE Transactions on Artificial Intelligence*, *5*(6), 2739–2751. https://doi.org/10.1109/TAI.2023.3332837

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292. https://doi.org/10.1002/rev3.3292

Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review*, *35*(4),113. https://doi.org/10.1007/s10648-023-09823-4

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Booth Olson, C. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying Generative AI in education. *ECNU Review of Education*, *6*(3), 355–366. https://doi.org/10.1177/20965311231168423

Wang, P., Lei, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, T., & Sui, Z. (2023). Large Language Models are not fair evaluators. In L.-W. Ku, A. Martins & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9440–9450). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.511

Wei, J., Want, K., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems* (pp. 24824–24837). Curran Associates Inc.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. In J. Yoder, R. Gabriel, V. Vranić & K.

Brown (Eds.), *Proceedings of the 30th Conference on Pattern Languages of Programs* (pp. 1–31). The Hillside Group.

Wilson, J., & MacArthur, C. (2024). Exploring the role of automated writing evaluation as a formative assessment tool supporting self-regulated learning and writing. In M. Shermis & J. Wilson (Eds.) *Routledge international handbook of automated essay evaluation*. Routledge. https://doi.org/10.4324/9781003397618-14

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Xie, W., Niu, J., Xue, C., & Guan, N. (2024). *Grade like a human: Rethinking automated assessment with Large Language Models*. https://doi.org/10.48550/arXiv.2405.19694

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., Chi, E., & Zhou, D. (2023). *Large Language Models as analogical reasoners*. https://doi.org/10.48550/arXiv.2310.01714

Yu, H., & Guo, Y. (2023). Generative artificial intelligence empowers educational reform: Current status, issues, and prospects. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1183162

Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, J., Dong, Z., Du, Y., Yan, C., Chen, Y., Chen, Z., Jian, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. (2023). *A survey of Large Language Models*. https://doi.org/10.48550/arXiv.2303.18223

Publisher's Note