Journal of Information Technology Education: Innovations in Practice

An Official Publication of the Informing Science Institute InformingScience.org

JITEiip.org

Volume 24, 2025

EDUCATIONAL EVALUATION WITH LARGE LANGUAGE MODELS (LLMS): CHATGPT-4 IN RECALLING AND EVALUATING STUDENTS' WRITTEN RESPONSES

Jussi S. Jauhiainen*	Department of Geography and Geology, University of Turku, Turku, Finland *and* Institute of Ecology and the Earth Sciences, University of Tartu, Tartu, Estonia	jusaja@utu.fi
Agustin Bernardo Garagorry Guerra	Department of Geography and Geology, University of Turku, Turku, Finland	abgagu@utu.fi

* Corresponding author

ABSTRACT

Aim/Purpose	This article investigates the process of identifying and correcting hallucinations in ChatGPT-4's recall of student-written responses as well as its evaluation of these responses, and provision of feedback. Effective prompting is examined to enhance the pre-evaluation, evaluation, and post-evaluation stages.
Background	Advanced Large Language Models (LLMs), such as ChatGPT-4, have gained significant traction in educational contexts. However, as of early 2025, systematic empirical studies on their application for evaluating students' essays and open-ended written exam responses remain limited. It is important to consider pre-evaluation, evaluation and post-evaluation stages when using LLMs.
Methodology	In this study, ChatGPT-4 recalled 10 times 54 open-ended responses submitted by university students, making together almost 50,000 words, and assessing and offering feedback on each response.
Contribution	The findings emphasize the critical importance of pre-evaluation, evaluation, and post-evaluation stages, and in particular prompting and recalling when utilizing LLMs for educational assessments.

Accepting Editor Stamatis Papadakis │ Received: October 30, 2024│ Revised: December 29, 2024; January 4, 2025 │ Accepted: January 6, 2025.
Cite as: Jauhiainen, J., S., & Garagorry Guerra, A. B. (2025). Educational evaluation with Large Language Models (LLMs): ChatGPT-4 in recalling and evaluating students' written responses. *Journal of Information Technology Education: Innovations in Practice, 24,* Article 2. https://doi.org/10.28945/5433

Findings	Using systematic prompting techniques, such as Chain of Thought (CoT), ChatGPT-4 can be effectively prepared to accurately recall, evaluate, and provide meaningful, individualized feedback on students' written responses, following specific instructional guidelines.
Recommendations for Practitioners	Proper implementation of pre-evaluation, evaluation and post-evaluation stages and testing of recall accuracy are important when using ChatGPT-4 for evaluating students' open-ended responses and providing feedback.
Recommendations for Researchers	Recall accuracy needs to be tested, and the prompting process carefully revealed when using and researching LLMs like ChatGPT-4 for educational evaluations.
Impact on Society	As LLMs continue to evolve, they are expected to become valuable tools for assessing student essays and open-ended responses, offering potential time and resource savings for educators and educational institutions.
Future Research	Future research should explore the use of various LLMs across different academic fields and topics to better understand their potential and limitations in educational evaluation.
Keywords	ChatGPT-4, education, recalling, LLM, evaluation

INTRODUCTION

Evaluating and grading student essays and open-ended exam responses, along with providing feedback, are essential yet time-intensive tasks for educators. These responsibilities require substantial commitment and resources, highlighting the significant workload placed on teachers and educational institutions. There is a growing interest in utilizing generative AI and advanced Large Language Models (LLMs) such as ChatGPT, Claude, Cohere, Gemini, LLaMa and Mistral to streamline evaluation and feedback tasks to assist teachers (Adiguzel et al., 2023; Baidoo-Anu & Owusu Ansah, 2023; Fütterer et al., 2023; Jauhiainen & Garagorry Guerra, 2023, 2024a, 2024b; Jeon & Lee, 2023; Lo, 2023; Michel-Villarreal et al., 2023; Steiss et al., 2024; J. Su & Yang, 2023; Xia et al., 2024).

To effectively use generative AI and LLMs for evaluating student essays and open-ended exam responses, many prerequisites need to be met. First, it must be ensured that the model precisely reproduces reading materials, questions, and student responses. Given that LLMs primarily function as text prediction models, there is a risk of them generating nonfactual or irrelevant content, known as hallucinations. For assessment and grading accuracy, it is crucial that the model follows the educational institution's specified criteria. The model needs also to provide relevant and constructive feedback to students. Additionally, considerations of security, privacy, and ethics are crucial to ensure that student data, evaluation results, and feedback are properly handled and not used to train models without appropriate consent (Wu et al., 2024).

In 2024, ChatGPT was the most popularly employed interactive chat-based LLM. ChatGPT utilizes Generative Pre-trained Transformer (GPT) architecture to deliver human-like text generation that has accurate response capabilities and can perform a wide range of linguistic tasks. It is trained on expansive text-based datasets that allow users to customize conversations, adjusting for desired output length, style, detail, and language. As LLMs evolve, they are transitioning to function as general-purpose solvers for complex tasks (Zhao et al., 2023).

This advancement underlines the importance of structuring the use of LLMs to cover the entire evaluation process, from pre-evaluation to evaluation and post-evaluation. This article addresses this research gap by examining the comprehensive evaluation process using ChatGPT-4, focusing on identifying and correcting hallucinations in recalled student responses, analyzing how the model evaluates these responses based on questions on learning materials, and examining the feedback it offers. This approach aims to fully leverage LLM capabilities in educational settings by optimizing prompt design.

The research questions are: How can ChatGPT-4 be prompted for efficient evaluation of written responses? How can differences between students' written responses and ChatGPT-4's recall be identified and corrected, particularly regarding potential hallucinations? How does ChatGPT-4 evaluate and grade students' written responses and provide feedback on those responses?

As of early 2025, most of the research on generative AI and LLMs in the field of education still focuses on their potential rather than their actual systematic application in real educational settings. Initial studies explored how these technologies can support teaching and enhance student learning experiences (Baidoo-Anu & Owusu Ansah, 2023; Lo, 2023). Experiments placing generative AI in the role of a student have shown that it can competently handle open-ended questions across various subjects and perform on par with advanced human students in examinations (Chalkidis, 2023; Guerra et al., 2023; Jung et al., 2023; Vázquez-Cano et al., 2023).

Empirical research on using LLMs for evaluating student essays and open-ended responses is limited but expanding rapidly. Initial studies suggest that educators generally concur with LLM assessment results (Dai et al., 2023; Steiss et al., 2024; L. Wang et al., 2024). However, much of this research focuses on the older GPT-3.5 model, which has shown limitations in evaluation tasks. As the field advances and newer models like GPT-4 are employed, the findings from studies utilizing earlier versions have become less relevant (Dai et al., 2023; Elkhatat, 2023; Elkhatat et al., 2023; Kooli & Yusuf, 2024; Lu et al., 2024; Y. Su et al., 2023). More advanced LLMs offer improved accuracy and reliability in educational assessment (Guerra et al., 2023; Jauhiainen & Garagorry Guerra, 2024a; Xia et al., 2024).

Besides grading, ChatGPT also offers substantial capabilities in providing detailed feedback on student assignments. This includes feedback on arguments and specific comments tailored to each student's submission. Studies show a general agreement among educators on the model's utility for feedback provision (Dai et al., 2023; Steiss et al., 2024; L. Wang et al., 2024). Xia et al. (2024) further highlight the potential of generative AI, noting three major advantages: perceived unbiased feedback, immediate and diverse feedback, and self-assessment.

Much of the research on feedback results derive from ChatGPT-3.5, yielding mixed results. For instance, Steiss et al. (2024) found that human feedback typically surpassed that of the model in various formative assessment aspects, though the overall quality and time-saving benefits between humans and the model were comparable. Contrarily, L. Wang et al. (2024) found students were unaware of whether the feedback was generated by ChatGPT or a human, and they found the model's feedback more useful, with 68–78% requiring little to no modification by teachers.

Research on the use of LLMs in educational settings has mostly overlooked the crucial recalling process and the challenges associated with it. Additionally, studies have generally not paid enough attention to the temperature settings of GPT models, which significantly affect the randomness of word predictions, thereby impacting LLM recall and performance in educational evaluations (Hackl et al., 2023; Jauhiainen & Garagorry Guerra, 2024b). There is also a lack of comparative analysis between different LLMs within such research frameworks (Jauhiainen & Garagorry Guerra, 2024a).

This article proposes a comprehensive framework that incorporates generative AI, prompt engineering, and text analysis to address issues such as hallucinations and minor inaccuracies commonly encountered when using LLMs for educational purposes (Bai et al., 2023). The framework emphasizes the importance of sophisticated prompting techniques to improve the accuracy of LLM outputs, aligning with multiple studies that highlight the critical role of prompt engineering, in particular the use of Chain-of-Thought (CoT) approach, in optimizing the performance of LLMs in educational evaluations (Guerra et al., 2023; Kojima et al., 2022; M. Wang et al., 2024; Yao et al., 2023; Zhao et al., 2023).

The empirical study tested the suggested framework evaluating the effectiveness of ChatGPT-4 in handling open-ended student responses from university courses based on English reading materials. This involved the model conducting 10 cycles of recalling, evaluating, and providing feedback for each of 54 student responses, culminating in 540 actions across each category. To ensure data security and integrity, ChatGPT-4 operated on a secure platform, which is crucial for training the model without compromising privacy. The study concludes with key findings and proposes directions for future research into the integration of generative AI and LLMs in educational assessment practices.

LANGUAGE PROCESSING AND PROMPTING WITH CHATGPT

Humans communicate with natural language using diverse vocabulary, varying sentence structures and free-form expressions. Such language is essentially a complex system of human expressions governed by grammatical rules. For this, generative AI needs to capture nuances and complexities of natural language to understand the meaning of sentences and the communication context, in order to generate coherent responses.

GPT employs neural networks within a transformer architecture that contains hundreds of billions of parameters, which are trained on massive text-based datasets. Such scaling largely improves the capacity of GPT-based LLMs and enhances their emergent abilities. The transformer is efficient in handling sequential data, making it well-suited for natural language processing (NLP) tasks. The transformer can capture long-range dependencies in the text, so GPT can focus on relevant information and generate contextually appropriate and semantically meaningful responses (Vaswani et al., 2017). OpenAI has been the key developer of GPT models having evolved over several generations. The initial version of ChatGPT was released in 2018. In May 2024, ChatGPT-4o was released.

PROMPT ENGINEERING WITH CHATGPT

Effective prompt engineering is essential for optimizing the performance of LLMs, such as ChatGPT, as the quality of prompts significantly impacts their output (M. Wang et al., 2024; White et al., 2023). Prompting involves crafting specific instructions or questions that guide the model, such as directing ChatGPT on how to evaluate students' responses. The choice of prompting technique varies depending on the task and the intended interaction with the model. Users typically refine their prompts through iteration, testing various phrasings and structures until they achieve clear, contextually relevant, and effective instructions. More detailed prompting has been shown to outperform simpler approaches (Liu & Shah, 2023), as iterative refinement helps align the model's responses with user expectations.

For this study, we utilized CoT, a well-known and widely used prompting technique (Zhao et al., 2023). It employs a sequential structure (prompts are presented in a sequential order, with each prompt building upon the previous one), context maintenance (each prompt contributes to the overall context of the conversation so that responses stay relevant and connected to preceding inputs), logical procession (each instruction logically follows from the previous one), exploration depth (a series of related questions result in a more in-depth exploration of the topic), flexibility (experimenting with different aspects of a topic), enhanced output quality (eliciting more detailed, contextually relevant, and logically coherent responses), and iterative process (experimenting with adjusting the phrasing or structure) (Bai et al., 2023; Chen et al., 2023).

When using the CoT technique, ChatGPT can break down complex tasks into simpler steps. This method involves intermediate prompts that guide the model through logical reasoning paths before it delivers a final answer, thus enhancing its zero-shot reasoning capabilities. Studies by Chen et al., (2023) and Kojima et al. (2022) suggest that CoT significantly improves the performance of zero-shot tasks compared to other methods when using the same prompt template. Additionally, even minimal but accurate human supervision can increase the accuracy of ChatGPT's outputs (Chen et al., 2023). Proper prompt tuning allows the model to adapt to new tasks without explicit examples, improving its generalization ability (Zhao et al., 2023). As LLMs evolve, it is anticipated that prompt engineering will become more intuitive and easier for users to implement to positively impact model effectiveness.

Evaluation with ChatGPT

One aspect in generative AI transformation of education is the application of LLMs to evaluate students' essays and open-ended written exam responses as well as provide feedback on their performance (Dai et al., 2023). Leveraging LLMs such as ChatGPT for this purpose holds promise, but specific prerequisites in pre-evaluation, evaluation and post-evaluation stages must be addressed (Figure 1). All three stages – material provision and recalling, consistent and accurate evaluation, and constructive tailored feedback – are integral to comprehensively assessing students' written performance.

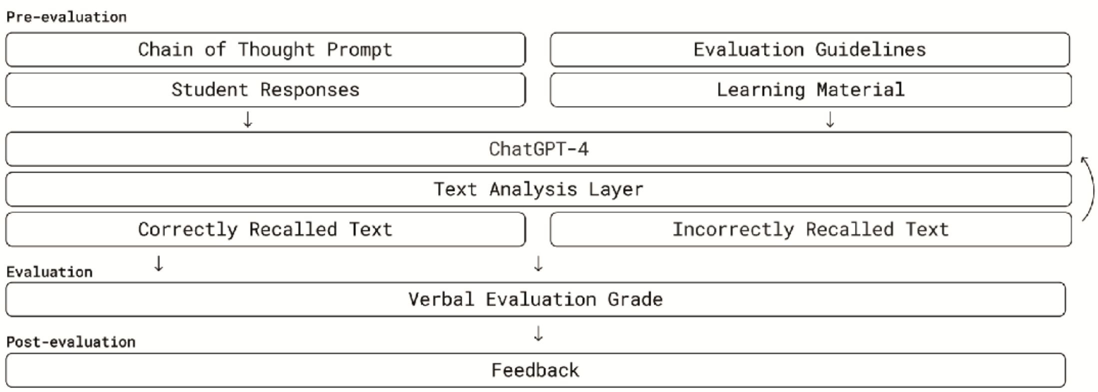

Figure. 1. Educational evaluation with ChatGPT-4

The first stage is pre-evaluation (Figure 1). This stage involves providing input material to the LLM for evaluation. The length and content of examination material varies significantly. As of early 2024, ChatGPT-4t was capable of processing approximately 300 pages of written text at once (OpenAI, 2024). However, the newest versions outperform the older ones.

The use of the Retrieval-Augmented Generation (RAG) technique improves the LLM's performance with specific, diverse and factual tasks such as fact checking (Jauhiainen & Garagorry Guerra, 2024a; Piktus et al., 2021). This includes informing the LLM about its evaluation tasks (prompting) and the material it will use for evaluation. As discussed earlier, using CoT, the LLM will respond to queries with reference to a specified set of documents given to it. For example, one can input written learning materials that students are required to study before their exam, insert teacher's exam questions covering the reading materials, and transfer students' written responses to the model. Finally, ChatGPT recalls these texts, i.e. reproduces digitally students' responses and other related material. However, the quality of recalled text needs to be systematically verified to identify possible hallucinations or truncated outputs, and erroneously recalled texts corrected.

Evaluation is the second stage in the LLM-based assessment of students' written responses (Figure 1). It regards LLM processing and assessing each student's written response according to criteria set by the educational institution, which are outlined in the prompts. Different knowledge taxonomies (Bloom's, revised Bloom, the SOLO taxonomy, Webb's depth of knowledge, etc.) may be required in students' responses to categorize their levels of understanding (Irvine, 2021). Students may be asked to repeat the reading material information, elaborate on a specific topic, apply their knowledge to solve a problem, or synthesize new insights beyond the material's facts.

Grading criteria used by the LLM ranges from simple pass/fail to detailed scales that judge performance from poor to excellent. Each response receives a grade (Irvine, 2021). Specific detailed criteria for evaluation are used, and these may include general performance expectations as well as more detailed aspects such as response length, coherence, factual accuracy, writing style, and grammatical correctness. These criteria are based on the regulations set by the educational institution. Ultimately, each response is assigned a verbal grade, either as a word or a number. For exams with multiple questions, each response may be individually scored and weighted which the LLM needs to compute the final grade.

In the third post-evaluation stage, the LLM provides feedback on students' written responses (Figure 1). This ranges from a simple numerical or verbal grade to detailed tailored feedback that assesses how well an individual student met or exceeded expectations. The model can also identify gaps and offer suggestions to improve students' understanding of the topic in order to enhance the learning process.

MATERIAL AND METHODS

MATERIAL

The data analyzed in this article includes university students' open-ended written responses in English, based on questions made by a teacher on assigned reading material. The data also covers ChatGPT-4's recall of responses, its evaluation and grading, and the feedback provided. This setup mimics a real educational scenario where a teacher would use an LLM, in this case, ChatGPT-4, to assess student responses. The scenario reflects a typical university exam environment, utilizing academic materials from a university course and corresponding questions to which students answered in writing. In this case, ChatGPT-4 processed each of the 54 student responses through 10 cycles of recall, evaluation, and feedback, resulting in 540 instances each of recall, evaluation, and feedback.

Students took part in three tests during related lectures at the university in which this article was conducted. The participants were informed about the test and were able to withdraw at any time if they wished to do so. Participation in the test did not have an impact on the student's curriculum. All participating students were adults and remained anonymous throughout the test; no names or other identifying information was collected from them. By writing and submitting responses for further analysis, participants gave informed consent to take part in the research.

Three reading materials that were part the students' Master of Science study curriculum courses were selected for the tests with one per test. The first author of this article wrote the texts from excerpts of published articles. One reading material was about irregular migration at the EU borderland (2,543 words), another was about irregular migration during the Covid-19 pandemic (3,734 words), and the third was about the knowledge creation processes (1,816 words). Three questions were constructed for each reading material.

In the first test, the length of students' responses to the three questions regarding the first reading material varied between 31 and 256 words (responded by six students), in the second test between 24 and 103 words (responded by six students), and in the third test between 62 and 102 words (responded by six students). The total length of answers was 4,261 words. In addition, recalling, evaluating and feedback provision was repeated 10 times (10-shot), so the total number of evaluated student responses (recall) was 540, together representing almost 50,000 words and 300,000 characters (exactly 46,871 and 298,618).

METHODS

The methods were designed for using ChatGPT-4 in educational settings across pre-evaluation, evaluation, and post-evaluation stages (Figure 1). ChatGPT-4t was operated in platform mode, ensuring that utilized data did not leak into general training of the model, which provides a secure environment for educational applications. This setup is ideal for teachers evaluating student responses. It does, however, limit control over LLM performance parameters such as temperature and sampling methods, although many teachers may not be fully aware of these parameters. Alternatively, using an

API (Application Programming Interface) could offer more control, as it employs a set of rules and additional tools to allow different software applications to communicate with each other. Yet this requires programming skills that many educators may also lack.

Pre-evaluation

During the pre-evaluation stage, ChatGPT-4 was calibrated to accurately recall student responses and consistently evaluate them using institutional criteria (Figure 1). The CoT prompting sequences guided the model through each evaluation step, utilizing self-generated reasoning exemplars without manual intervention. This method, aligned with analogical prompting strategies from AI research by Google DeepMind and Princeton University (Vaswani et al., 2017; Yasunaga et al., 2023), optimizes ChatGPT-4's performance, ensuring effective operation within educational evaluation frameworks.

The second step involved manually uploading educational materials for ChatGPT-4's assessment (Figure 1). This material comprised three articles that formed the basis for the questions asked and to which students responded. Three questions per article were provided, resulting in a total of 18 responses from students for each piece of material, amounting to 54 responses altogether.

The third step involved ensuring the accuracy of ChatGPT-4's recall by testing each individual recalled student response. If the recalled response deviated from the original student-written response, the response was re-inputted (re-shot) into ChatGPT-4 until the recalled version closely matched the original response (Figure 1). This iterative process guaranteed the precision of the data used for further evaluation.

To identify inaccuracies between the original student responses and versions recalled by ChatGPT-4, we utilized similarity testing. Initially, a basic software tool was employed to count the words in each original and recalled response. This straightforward method effectively highlighted major discrepancies, where the word count differed significantly between the two versions.

For a detailed analysis of the similarities between original student responses and their recalled versions by ChatGPT-4, complementing techniques were employed including Sentence Transformers Semantic Similarity (STSS) that provides a measure of contextual similarity and the Levenshtein index that provides a granular look at the textual changes in recalled responses, ensuring a comprehensive analysis of recall accuracy.

STSS uses contextual similarity to evaluate text. It involves embedding reference texts and performing semantic similarity analysis using the cosine similarity score. This metric is commonly used in LLM-related analyses but may not always detect hallucinations, as the vector representations of the original and hallucinated texts can be similar.

To address potential shortcomings of the STSS, the Levenshtein similarity index, also known as the Edit Distance index, was used. This index measures the minimum number of single-character edits (insertions, deletions, or substitutions) needed to make two strings identical. It effectively identifies whether ChatGPT-4 made any small changes, such as minor grammar and spelling corrections, or more significant alterations like hallucinations.

Evaluation

During the evaluation stage, ChatGPT-4 systematically assessed and graded students' responses based on established educational criteria. To enhance accuracy and minimize potential hallucinations in the evaluations, a 10-shot evaluation method was employed, where ChatGPT-4 reviewed each response ten times. This approach provided a comprehensive overview of the model's performance capabilities. The final grading for each response was determined by selecting the most frequently occurring grade (the mode) from these ten evaluations, allowing for a more reliable and consistent assessment outcome (see Figure 1).

Post-evaluation

During the post-evaluation stage, ChatGPT-4 was specifically tasked with providing written feedback on student responses, a process repeated ten times (10-shot) for each response to ensure thoroughness and consistency (Figure 1). While examples from both the evaluation and post-evaluation stages are discussed in this article, an in-depth analysis of these stages extends beyond the scope of this article, which primarily focuses on the pre-evaluation phase. Such detailed examination of entire educational cycle would necessitate a more extensive discussion than the article length limitation allows.

RESULTS

PRE-EVALUATION: PROMPTING AND INPUT

We evaluated various methods to enhance the accuracy of ChatGPT-4's recall capabilities. Initially, the model's recall was conducted without specific prompts, leading to inconsistencies and unclear accuracy. To refine this process, we conducted experiments with various prompts, adjusting them based on the outcomes observed. This approach led to the identification of effective practices that yielded successful and consistent results. Instead of focusing on numerous unsuccessful attempts, this discussion will concentrate on the methods that proved effective in improving the precision of the model's performance.

The use of the CoT prompting technique, as suggested by Chen et al. (2023), helped guide ChatGPT-4 through intermediate problem-solving steps. Finally, we identified a stable prompt that consistently yielded structured outcomes. This final prompt defined ChatGPT-4's role as a university professor, tasked with evaluating responses using university criteria and familiarizing itself with reference materials, and finally providing related feedback (Figure 2). We instructed the model to minimize randomness in its recalls and evaluations to reduce hallucinations—instances of unexpected or implausible content that could distort the alignment between original responses and generated outputs. Detailed instructions and examples also clarified the format of questions posed to students as well as the task to reference learning materials when conducting the evaluation (Figure 2).

Prompt example for recalling and evaluating student answers

```
You are a University Professor that evaluates the student answer according to the evaluation
guidelines provided and contrast the facts with the reference material provided as knowledge.

Use the uploaded PDF as a reference material for your evaluation.

You will be provided with a series of questions and answers looking like this:

"""
Student ID: write the student ID
Question:  here goes a question
Answer: here goes student answer
Question:  here goes a question
Answer: here goes student answer
Question:  here goes a question
Answer: here goes student answer
"""

//When writing the student answer, ALWAYS write it in the same way without any modification or
shortening, write the full student answer.

//Do it following the format guideline:

"""
Student ID: Write student ID
Question: Write the question being answered
Answer: Write the student answer to the question
Student Answer Feedback: Write the Feedback
Context Relevance:
Factual Accuracy:
Completeness: Fail | Passable | Satisfactory | Good | Very Good | Excellent
Depth of Understanding: Fail | Passable | Satisfactory | Good | Very Good | Excellent
Logical Consistency: Fail | Passable | Satisfactory | Good | Very Good | Excellent
Grammar and Spelling: Fail | Passable | Satisfactory | Good | Very Good | Excellent
Grade Answer: Fail | Passable | Satisfactory | Good | Very Good | Excellent
"""
```

Figure 2. Example of CoT prompting in the pre-evaluation of students' written responses

Pre-Evaluation: Similarity Tests

In addition to the discussion on guidance regarding prompt generation, this article focuses on ways to identify discrepancies and hallucinations between original student responses and the versions recalled by ChatGPT-4. A critical preliminary step in any LLM-based evaluation of written texts involves detecting these inconsistencies.

Initially, discrepancies were identified by comparing the word count of the original and recalled responses; any variance in word count would signal non-identical texts. Substantial differences often indicate hallucinations, which may involve additions or deletions of text. Without these checks, there is a risk that such hallucinations could go unnoticed, affecting the reliability of the evaluation.

In a one-shot recall test involving 54 texts, 87.0% maintained identical word lengths with minor discrepancies in the remainder: 7.4% differed by one word, 1.9% by two or three words, and 3.7% by four or more words. Consequently, 94.4% of the recalled responses were either identical or had a minimal one-word difference. In a 10-shot dataset, which involves repeating the recall process ten times, the fidelity decreased slightly, particularly notable in the third shot where discrepancies were more prevalent. In an unfortunate case in which that third shot would have been the only one utilized, and without measuring the accuracy of the recalled responses, the entire evaluation process would have failed. Overall, 88.2% of the 540 recalled responses had the exact same word count, with smaller variations in others: 4.3% had a one-word difference, 2.8% a two-word difference, 0.2% a three-word difference, and 2.8% more than a three-word discrepancy. This variability indicates the challenges of maintaining recall accuracy across multiple iterations and highlights the need for careful monitoring to ensure consistency.

Later we conducted more in-depth analysis of similarities and differences between the original and recalled texts. Maintaining the same word count between an original response and its recalled version does not guarantee identical content. To determine if the texts conveyed the same meaning, we conducted similarity tests using Levenshtein similarity index as well STSS (for explanation, see methods). For this, we utilized a one-shot scenario conducting the analysis only once regarding all student responses.

The Levenshtein index showed all recalled texts having a high similarity score of at least 0.969. Specifically, 38.9% of responses were identical, 61.1% had minor differences, and none showed major discrepancies. Using STSS, 48.2% of recalled responses were completely identical, 50.0% had minor variations, and 1.9% indicated a potential hallucination with a lower similarity score of 0.780, suggesting significant content alteration. These tests help confirm how well the recalled texts mirror the originals and when corrections are needed.

The results show that ChatGPT-4 effectively recalled student responses, although it made automatically minor grammar and spelling corrections. The accuracy of recall was influenced by how students framed their responses, such as their use of grammar and logical phrasing. All of the one-shot recalled responses (54/54) and 97.2% of the 10-shot recall responses (524/540) were sufficiently similar to the original responses to proceed with their evaluation. Instances of hallucination were noted, but repeated shots showed that ChatGPT-4 could eventually recall these responses accurately. Three groups emerged:

In the "exact match" category, original and recalled texts were identical, representing an ideal outcome where ChatGPT-4 accurately reproduced responses as initially written following the prompt. This accuracy allowed for direct evaluation without further intervention. In one-shot tests, 38.9% (21 of 54) of recalled responses were perfect matches, while in 10-shot tests, the figure rose to 53.1% (287 out of 540).

In the “minor variations” category, the discrepancies between the original and recalled texts by ChatGPT-4 were insignificant and did not affect the meaning of the responses. Notably, the model corrected common typing errors in the student responses without specific instruction to do so. For example, misspelled words such as "recieved" and "sisable" were corrected to "received" and "sizable." In one-shot tests, 61.1% (33 of 54) of recalled responses fell into this category, while in 10-shot tests, the proportion was 44.1% (238 out of 540).

In the "significant differences" category, substantial disparities emerged between the original responses and those recalled by ChatGPT-4. Despite precise prompts, hallucinations occurred where the model either omitted or added content not present in the original responses. For instance, instead of simply repeating a student's response, in a few cases ChatGPT-4 generated a summary or explanation of the response, substantially altering it. Such changes required reshooting until they were correct. In the one-shot tests, there were no instances (0.0%) of significant discrepancies, but in the 10-shot tests, 15 out of 540 responses (2.8%) included these hallucinations, particularly in longer responses. In addition, hallucinations particularly concentrated into one round of 10 shots. As mentioned above, it was possible to identify these hallucinations with word counts and similarity tests and remove these through repeated shots.

EVALUATION

In this article, the primary focus is on the pre-evaluation stage, setting the groundwork for evaluations conducted via LLMs like ChatGPT-4. Nevertheless, the evaluation and post-evaluation stages with ChatGPT-4 are now briefly discussed.

As explained in detail earlier, ChatGPT-4 was carefully prompted to assess student responses according to the grading scale and evaluation guidelines customary at the university where the study was conducted. Besides overall evaluation, grading and feedback, instructions were provided within prompts so that the model addressed five key evaluation parameters: context relevance, factual accuracy, completeness, logical consistency, and grammar and spelling (see Figure 2).

An experiment conducted demonstrated ChatGPT-4's capability to evaluate student responses, accurately grading them as per the given instructions (see Figure 2). It distinguished differences in the quality of responses and distributed grades accordingly, thus aligning with the educational standards of the institution. In detail, the grading results were the following: 1.3% fail, 2.6% passable, 24.8% satisfactory, 43.0% good, 15.4% very good and 13.0% excellent. Notably, when the evaluation was repeated multiple times (10-shot), ChatGPT-4's grading showed consistency, with final grades remaining stable or deviating minimally across multiple assessments. Of 540 evaluated student responses, the final evaluation grade was the same in 68.7% of cases, it remained within one grade in 96.1% of cases and all were at least within two points from the initial grade. A minor inconsistency was identified in 3.9% of cases. Further details and a deeper analysis of this evaluation process are discussed in another article (Jauhiainen & Garagorry Guerra, 2024b).

POST-EVALUATION

In the post-evaluation phase, ChatGPT-4 was capable of delivering personalized feedback to student responses. It reflected in the feedback content on similarities and differences between the response and the learning material, recognized significant omissions, identified logical inconsistencies in writing, and pinpointed grammatical and spelling errors. It aligned with evaluation parameters that ChatGPT-4 was prompted to follow.

The structure of ChatGPT-4's provided feedback to student response is illustrated and analyzed here (Figure 3). The model effectively identified both strengths and areas for improvement within student responses. In feedback provision, ChatGPT-4 relied on evaluation parameters. Initially, the model recognized the connection between student's response and the learning material along with the predefined evaluation criteria and positively remarked that in the feedback. Figure 3 highlights positive remarks in green. Later, the model recognized where responses lacked required knowledge. Figure 3 highlights critical remarks in red. Overall, ChatGPT-4 provided written feedback with neutral tones, accompanied by constructive suggestions aimed at enhancing student's understanding and response quality. Further discussion of the post-evaluation process is reserved for another article.

Evaluation Process Example

What were the key challenges that the Covid-19 brought to organize innovation events?

Recalled Answer: During the Covid-19 pandemic it was not allowed to organize any meetings because the stakeholders were scared that the virus would spread in crowds. People were not allowed to travel to different cities or countries for example. There still was a need to organize innovation events because new ways to adjust to the situation were needed. That's why the innovation events were organized virtually.

Feedback: The answer correctly identifies the challenges posed by the Covid-19 pandemic for organizing innovation events, such as the prohibition of physical meetings and travel restrictions. The response also correctly points out the necessity of continuing innovation events and their adaptation to virtual formats. However, it could be improved by elaborating on specific challenges related to virtual event organization, like technical issues, participant engagement, and the difference in experience compared to in-person events.

Context Relevance:Good
Factual Accuracy:Good
Completeness:Satisfactory
Logical Consistency:Good
Grammar and Spelling:Excellent
Final Grade:Good

** Green highlights feedback where the student positively met the evaluation criteria,*

*** Red highlights sections with a knowledge gap, accompanied by the model's suggestions for improvement*

Figure 3. Example of a student's written response alongside ChatGPT-4's feedback

CONCLUSION

This article evaluated ChatGPT-4's ability to recall, assess, and provide feedback on 54 open-ended responses from Master level university students, processing each response 10 times, resulting in 540 recalled responses, evaluations, and feedback instances. The main results of this article are the following.

First, ChatGPT-4 can be effectively prompted so that it performs consistently across pre-evaluation, evaluation, and post-evaluation stages. The CoT technique was proven useful for creating clear and effective prompts. It is advisable to rigorously test, develop, and iterate the prompt until the LLM, such as ChatGPT, achieves the desired performance. Furthermore, detailed documentation of the prompting process is crucial when reporting on the use of LLMs in educational evaluations as it aids in assessing the quality of observations. Additionally, adjustments to model parameters such as temperature settings should be considered to refine model performance. These have been largely overlooked in previous studies on educational evaluation with LLMs despite their recognized importance by scholars in other fields (Hackl et al., 2023; White et al., 2023; Yao et al., 2023; Zhao et al., 2023). Furthermore, to minimize risks of unwanted data leakage, it is recommended to employ a secure platform version of ChatGPT-4. This ensures that sensitive educational data remains protected while facilitating the evaluation process. These measures are essential to leverage the capabilities of LLMs effectively and ethically in educational evaluation.

Second, when using LLMs like ChatGPT-4 for educational evaluations, particularly for assessing student essays and open-ended responses, it is crucial to initially focus on the model's capabilities of accurate recall. Studies have often overlooked this aspect, yet it is fundamental to reliable evaluation. If the LLM does not accurately recall original student responses, its subsequent evaluations cannot be deemed reliable.

Additionally, given the inherent randomness in how LLMs recall information, it is advisable to conduct tests comparing the similarity between the original texts and their recalled versions. We found word count and the Levenshtein index effective for measuring the alignment between original texts and those recalled by ChatGPT-4. These methods are adept at identifying discrepancies in the text structure as they focus on formal differences rather than semantic content. In contrast, methods based on vector and semantic similarities might overlook inaccuracies, recognizing semantically coherent content that may still be factually incorrect or hallucinated.

When assessing the recall accuracy of ChatGPT-4 in a one-shot test of 54 responses, only 38.9% were identical to the original response. However, 94.4% of the recalled responses differed only minimally and did not affect the overall meaning of the responses. In a more extensive 10-shot evaluation test involving 540 responses, the model created substantial new text or significantly altered the content in 2.8% of the cases, effectively hallucinating. This was easy to identify implementing word count and the Levenshtein similarity index. These methods proved effective, particularly when the responses were written in reasonably correct English and did not exceed 250 words. When recalled responses were identified as hallucinated, a practical solution was to re-input the original responses into ChatGPT-4, repeating the process until the recall was accurate. ChatGPT-4 automatically corrected minor grammatical and spelling errors during the recall process, even though it was prompted to reproduce the responses exactly as submitted. This did not impact the meaning of the responses.

Third, the rapid development of LLMs looks promising regarding their application to education evaluation, including feedback provision. Earlier studies have noted a positive evaluation performance of ChatGPT-3.5, though opinions varied regarding its consistency. Newer models like ChatGPT-4o show significant improvements, making them more suitable for educational evaluation. These advancements respond to several limitations noted in previous studies, including those on feedback provision (Kooli & Yusuf, 2024; Steiss et al., 2024). It is important to approach early conclusions about the role of LLMs in educational evaluations cautiously, particularly if they are based on observations of earlier, less capable models (Jauhiainen & Garagorry Guerra, 2024a).

A small experiment in this article suggests that ChatGPT-4 is capable of adhering to specific prompts for evaluating each student's response along prompted criteria and providing systematic feedback with constructive tone. The model offered tailored feedback that reflected the unique aspects in student's response, effectively identified positive attributes in responses and suggested relevant improvement. This is consistent with remarks in earlier studies in which students appreciated feedback from ChatGPT and often did not differentiate it from feedback provided by human instructors (L. Wang et al., 2024).

The study has limitations as it was based on a small sample size of 54 student responses though ChatGPT-4 performed 540 operations of recall, evaluation, and feedback. The confined use of short reading materials and restriction to text-only responses without figures or tables are limiting factors. Also, the lack of comparative human evaluations limits the ability to gauge the LLMs' evaluation and feedback provision performance comprehensively.

Future research should expand to the evaluation of various LLMs across all stages of educational assessment (Jauhiainen & Garagorry Guerra, 2024a) investigating aspects such as the fairness and quality of LLM-provided feedback, its long-term impact on learning, and the computational costs involved. There is a need to compare generative AI assessment with human evaluation to explore the potential for a hybrid approach that encompasses both human and LLM evaluation.

LLMs hold considerable promise for assisting educators with time-consuming tasks of consistent evaluation and tailored feedback provision. LLMs like GPT-4 have significantly advanced beyond earlier versions such as GPT-3.5, offering enhanced capabilities in evaluating and providing tailored feedback for student assignments. The effectiveness of LLMs in educational settings relies heavily on precise prompt engineering to guide the models systematically through tasks. The initial setup in the pre-evaluation stage is crucial, ensuring that the model's responses align accurately with the text being evaluated. As LLMs use expands, it is essential to align their use with ethical, secure, and transparent educational practices.

Conflict of Interest

The authors indicate no conflict of interest.

References

Adiguzel, T., Kaya, M., & Cansu, F. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Education Technology, 15*(3), 429. https://doi.org/10.30935/cedtech/13152

Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. *Brain Behavior, 1*(3). https://doi.org/10.1002/brx2.30

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, 2023. https://doi.org/10.2139/ssrn.4337484

Chalkidis, I. (2023). ChatGPT may pass the bar exam soon, but has a long way to go for LexGLUE benchmark. *arXiv*:2304.12202**.** https://doi.org/10.48550/arXiv.2304.12202

Chen, J., Chen, L., Huang, H., & Zhou, T. (2023). When do you need Chain-of-Thought prompting for ChatGPT? a*rXiv*:2304.03262v2 https://doi.org/10.48550/arXiv.2304.03262

Dai, Y., Liu, A., & Pim, C. (2023). Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education. *Procedia CIRP, 119*, 84–90. https://doi.org/10.1016/j.procir.2023.05.002

Elkhatat, A. (2023). Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. *International Journal for Educational Integrity, 19*, 15. https://doi.org/10.1007/s40979-023-00137-0

Elkhatat, A., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity, 19,* 17. https://doi.org/10.1007/s40979-023-00140-5

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. *Science Reports, 13,* 15310. https://doi.org/10.1038/s41598-023-42227-6

Guerra, G., Hofmann, H., Sobhani, S., Hofmann, G., Gomez, D., Soroudi, D., Hopkins, B., Dallas, J., Pangal, D., Cheok, S., Nguyen, V., Mack, W., & Zada, G. (2023). GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. *World Neurosurgery, 179*, e160–e165. https://doi.org/10.1016/j.wneu.2023.08.042

Hackl, V., Müller, A., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education, 8*, 1272229. https://doi.org/10.3389/feduc.2023.1272229

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

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). 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*(4), 177. https://doi.org/10.54364/AAIML.2024.44177

Jauhiainen, J., & Garagorry Guerra, A. (2024b). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International.* https://doi.org/10.1080/14703297.2024.242

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies, 28*, 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Jung, L., Gudera, J., Wiegand, T., Allmendinger, S., Dimitriadis, K., & Koerte, I. (2023). ChatGPT passes German state examination in medicine with picture questions omitted. *Deutsches Ärtzeblatt, 120*(21), 373–374. https://doi.org/10.3238/artzebl.m2023.0113

Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are zero-shot reasoners. *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 1613, 22199-22213. https://doi.org/10.48550/arXiv.2205.11916

Kooli, C., & Yusuf, N. (2024). Transforming educational assessment: Insights into the use of ChatGPT and Large Language Models in grading. *International Journal of Human-Computer Interaction*, 1–12. https://doi.org/10.1080/10447318.2024.2338330

Liu, R., & Shah, N. (2023). ReviewerGPT? An exploratory study on using Large Language Models for paper reviewing. *arXiv*:2306.00622v1. https://doi.org/10.48550/arXiv.2306.00622

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

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., & Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education, 49*, (5), 616–633. https://doi.org/10.1080/02602938.2024.2301722

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D., Thierry-Aguilera, R., & Gerardou, F. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13*(9), 856. https://doi.org/10.3390/educsci13090856

OpenAI. (2024). New models and developer products announced at DevDay. (last visited January 12th, 2024). https://openai.com/blog/new-models-and-developer-products-announced-at-devday

Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäshcel, T., & Kiela, D. (2021). Retrieval-Augmented Generation for knowledge-intensive NLP tasks. *arXiv*:2005.11401. https://doi.org/10.48550/arXiv.2005.11401

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschasuer, 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

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv*:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Vázquez-Cano, E., Ramírez-Hurtado, J., Sáez-López, J., & López-Meneses, E. (2023). ChatGPT: The brightest student in the class. *Thinking Skills and Creativity, 49,* 101380. https://doi.org/10.1016/j.tsc.2023.101380

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity*, *51*, 101440. https://doi.org/10.1016/j.tsc.2023.101440

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., & Yin, M. (2024). Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*, *17*, 629–641. https://doi.org/10.1109/TLT.2023.3324714

White, J., Fu, C., Sanborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv*:2302.11382. https://doi.org/10.48550/arXiv.2302.11382

Wu, X., Duan, R., & Ni, J. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Intelligent Information Systems, 2*(2), 102–115. https://doi.org/10.1016/j.jiixd.2023.10.007

Xia, Q., Weng, X., Ouyang, F., Lin, T., & Chiu, T. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal in Education Technology and Higher Education* 21, 40. https://doi.org/10.1186/s41239-024-00468-z

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate problem solving with Large Language Models. *arXiv:2305.10601*. https://doi.org/10.48550/arXiv.2305.10601

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

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., ... Wen, J. (2023). A survey of Large Language Models. *arXiv:*2303.18223. https://doi.org/10.48550/arXiv.2303.18223

AUTHORS

Dr. Jussi S. Jauhiainen is Professor at the University of Turku (Finland) and Visiting Professor at the University of Tartu (Estonia), in which he is also Doctor honoris causa. His main research interests are in human geography, innovation studies and generative AI, in particular the use of LLMs for educational evaluation and feedback provision. He is author of some 200 scientific publications and developer of an LLM-based evaluation-feedback platform, TurkuEval (https://sites.utu.fi/digileac).

Mr. Agustín Bernardo Garagorry Guerra is Research Associate at the University of Turku (Finland). His main research interest is generative AI, in particular the use of LLMs for educational evaluation and feedback provision. He has published several articles on these topics. He is also developer of an LLM-based evaluation-feedback platform, TurkuEval (https://sites.utu.fi/digileac).