Taylor & Francis
Taylor & Francis Group

Innovations in Education and Teaching International

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/riie20

Generative AI in education: ChatGPT-4 in evaluating students' written responses

Jussi S. Jauhiainen & Agustín Garagorry Guerra

To cite this article: Jussi S. Jauhiainen & Agustín Garagorry Guerra (01 Nov 2024): Generative AI in education: ChatGPT-4 in evaluating students' written responses, Innovations in Education and Teaching International, DOI: 10.1080/14703297.2024.2422337

To link to this article: https://doi.org/10.1080/14703297.2024.2422337

Published online: 01 Nov 2024.

INNOVATIONS IN EDUCATION AND TEACHING INTERNATIONAL
https://doi.org/10.1080/14703297.2024.2422337

Routledge
Taylor & Francis Group

Generative AI in education: ChatGPT-4 in evaluating students' written responses

Jussi S. Jauhiainen [a,b] and Agustín Garagorry Guerra [a]

[a]Department of Geography and Geology, University of Turku, Turku, Finland; [b]Institute of Ecology and the Earth Sciences, University of Tartu, Tartu, Estonia

ABSTRACT

The study highlights ChatGPT-4's potential in educational settings for the evaluation of university students' open-ended written examination responses. ChatGPT-4 evaluated 54 written responses, ranging from 24 to 256 words in English. It assessed each response using five criteria and assigned a grade on a six-point scale from fail to excellent, resulting in 3,240 evaluations. Verification-based chain-of-thought prompting with the RAG framework ensured ChatGPT-4's accurate recall of responses and secure alignment in the university's evaluation criteria. ChatGPT-4's grading showed good consistency with the teacher's grading. Mistakes in recalls and discrepancies between ChatGPT-4 and teacher assessments could be reduced. The results suggest a promising potential for using LLMs like ChatGPT-4 in academic written response evaluations.

KEYWORDS

ChatGPT; education; evaluation; generative AI; LLM; assessment

Introduction

Education in the 2020s faces a complex landscape with significant challenges and opportunities worldwide. Teachers are under increasing pressure due to rising administrative tasks, the integration of new educational technologies and tight budgets. Continuous student evaluations across all educational levels consume a substantial portion of teachers' time, contributing to a stressful work environment. This stress has led to a decline in interest in teaching careers and difficulties in retaining current educators, raising concerns about a potential future shortage of education staff (Skaalvik & Skaalvik, 2021).

Scholarly performance is declining in many OECD countries, as various evaluations highlight. However, the moderate introduction of digital devices can positively impact learning (OECD, 2023). Advanced digitalisation and generative AI platforms, including Claude, Cohere, Gemini, GPT, LLaMa and Mistral, present both opportunities and challenges for education. Maximising the benefits of these technologies requires a strategic approach, ensuring ethical and equitable practices, data security and privacy protection.

CONTACT Jussi S. Jauhiainen jusaja@utu.fi Department of Geography and Geology, University of Turku, Vesilinnanmäki 5, Turku 20014, Finland

Successful integration of generative AI and LLMs in education involves collaboration among administrators, managers, students and teachers (Adiguzel et al., 2023; Baidoo-Anu & Owusu Ansah, 2023; Farrokhnia et al., 2023; Holmes & Miao, 2023; Jauhiainen & Garagorry Guerra, 2023; Jeon & Lee, 2023; Lo, 2023; Su & Yang, 2023; Yu & Guo, 2023).

At the level of education management, implementing AI regards a more knowledgeable, efficient and optimised use of educational resources aligned with curriculum objectives. Following ethical data mining principles, AI can analyse vast amounts of data to identify students´ learning patterns, assess their performance and provide insights to teachers. Digitalisation, AI in general and generative AI in particular will increasingly influence curriculum design (Bahroun et al., 2023). This enables informed data-driven decision-making for curriculum development, teaching strategies and student support systems (Chiu, 2023).

At the teacher level, various practical applications such as chat-based LLM tools present numerous advantages. They provide easy access to information (Farrokhnia et al., 2023). Additionally, for example, Copilot and other tools can be utilised to automate routine administrative tasks, such as managing schedules and curriculum design, allowing educators to prioritise teaching in dynamic learning environment over its administration (Baidoo-Anu & Owusu Ansah, 2023; Kaplan-Rakowski & Grotewold, 2023). They may streamline student assessment, grading and feedback provision through automated processes, freeing up time for more interactive teaching (Mizumoto & Eguchi, 2023). It is important for both teachers and students to understand the possibilities and pitfalls of these technologies as they become integral to educational processes. Generative AI and LLM tools are expected to complement and enhance, rather than replace, the valuable human element in education. Educators need training to effectively integrate these tools into their teaching methods, ensuring a harmonious and productive fusion of human and AI-driven instruction (Jeon & Lee, 2023).

Open access to generative AI platforms, such as LLMs, has created new possibilities to enhance students' learning experiences. Adaptive learning platforms can be used to assess student performance, deliver personalised motivating learning materials, lessons and feedback, catering to the unique requirements of each student without forgetting teamwork and collaborative learning (Baidoo-Anu & Owusu Ansah, 2023; Chiu, 2023). Advanced AI tools, like LLM-supported chatbots and virtual assistants, engage students in interactive experiences, offering instant feedback, responding to queries and encouraging active participation (Fui-Hoon Nah et al., 2023). Generative AI introduces personalised learning paths, tailoring content to individual needs and accommodating diverse learning styles (Y. Dai et al., 2023; Farrokhnia et al., 2023). Furthermore, chat-based LLM tutoring systems provide personalised assistance both inside and outside the classroom and lecture halls, aiding students in reinforcing their understanding of concepts and bridging learning gaps. Although access to digital tools is increasing globally, remaining digital divides could short-term increase educational inequalities due to unequal Internet and AI access (Mannuru et al., 2023).

As mentioned above, there is a growing need in leveraging LLMs for practical implementation in schools and universities. The number of scholarly articles discussing the potential of LLMs and in particular ChatGPT has boomed since 2023. However, the vast majority of articles discuss about the potential of ChatGPT rather than testing and using it in educational contexts (Adiguzel et al., 2023; Baidoo-Anu & Owusu Ansah, 2023;

Farrokhnia et al., 2023; Jeon & Lee, 2023; Lo, 2023; Su & Yang, 2023; Yu & Guo, 2023). Furthermore, the bulk of research published so far have focused on GPT model 3.5. However, the GPT model 3.5 has substantially weaker performance than that of the model 4, especially when it needs to be used systematically (Bewersdorff et al., 2023; Jauhiainen & Garagorry Guerra, 2024). Therefore, we will be focusing on ChatGPT-4 in this article.

There is growing interest in using ChatGPT to assess and provide feedback on students' written texts, such as essays and open-ended exam responses, as these tasks are time-consuming for teachers (Y. Dai et al., 2023). While ChatGPT-4 accurately identifies many fundamental student errors, detecting more complex errors remains challenging (Bewersdorff et al., 2023). Using ChatGPT for evaluation requires a systematic process to ensure accurate recall of students' responses. It must adhere to educational evaluation guidelines, perform consistent and reliable grading, and provide systematic, grounded feedback to both teachers and students on the evaluation process and final grade.

As regards the use of ChatGPT-4 in educational evaluation, our earlier research focused on technical aspects of LLMs' recalling of students´ written texts. Forthcoming research deals with ChatGPT-4's feedback on students' performance. In the current article, we use OpenAI´s GPT-4 model in the chat format. Instead of comparing various LLMs, we focus on ChatGPT-4 that is globally widely used, and address the following research questions: (1) How and how consistently does ChatGPT-4 evaluate open-ended written responses from students?; (2) Where do disparities emerge between ChatGPT-4 and human evaluation of students´ written responses?; (3) How the potential disparities between ChatGPT-4 and human grading of students responses could be solved?

Evaluating students´ written texts with LLMs

Evaluating students´ open-ended written responses to exams through the use of LLMs such as GPT, Cohere, Claude, Gemini, LLaMa or Mistral entails a systematic process of pre-evaluation, evaluation and post-evaluation, as it will be discussed below (Figure 1). Each student´s responses need to be assessed according to the criteria established by the educational institution (Baidoo-Anu & Owusu Ansah, 2023; Gimpel et al., 2023). The evaluation criteria range from a basic pass/fail system to more intricate scales that span

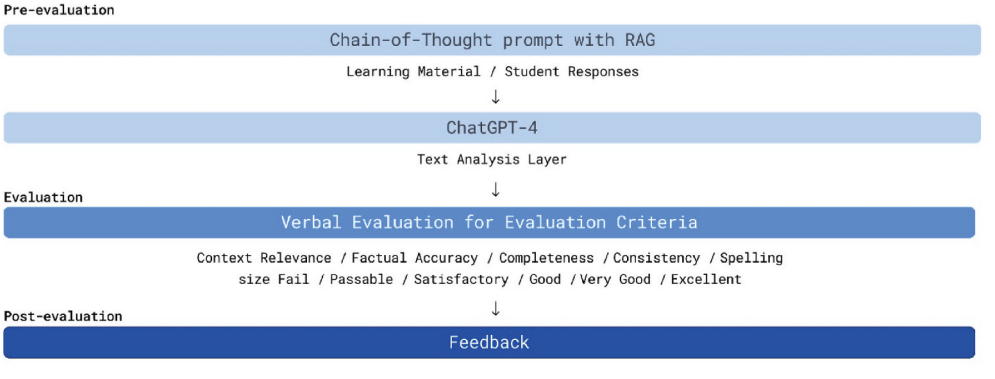

Figure 1. Educational evaluation with ChatGPT-4.

from poor to excellent performance, and may use either numeric or text-based marking methods (Irvine, 2021).

Responding to examinations, students may be tasked with remembering or expanding on a specific topic from reading material, applying knowledge to solve problems, or synthesising new insights beyond factual content. Therefore, the answer may require knowledge from different taxonomy levels (Irvine, 2021). Traditionally, teacher designs examination questions but these can be also designed automatically with LLMs (U. Lee et al., 2024). Requirements for students' responses may encompass general expectations of learning outcomes or specifics such as response length, coherence, factual accuracy and even writing style and grammatical correctness. Teachers need to adhere to the evaluation criteria consistently. Ultimately, the evaluation process culminates in assigning a verbal (word or number) grade to each response. In exams with multiple questions, each individual response is considered, possibly weighted, and they contribute to the final grade. This comprehensive evaluation approach ensures a thorough and nuanced assessment of students´ performance in examinations.

Pre-evaluation with LLM

The first-stage students´ written response evaluation, pre-evaluation, include input on students´ responses, analysing the similarity between the original student's response and its recalled texts for the use of LLM, and sometimes reshooting of the response until correct recall is achieved (see Yasunaga et al., 2023). Accurate recalled students´ written responses are fundamental for their evaluation in LLMs, otherwise their evaluation cannot be correct.

GPT models are based on data going through neural networks to understand and generate natural language. GPT recreates students´ responses as its ability to directly copy them is limited (Su & Yang, 2023). Such prediction of words in text increases a possibility of hallucinations so that recalled student responses contain text that was not initially there. Hallucinations are one challenge in using LLMs for evaluation and their presence needs to be systematically tested and eliminated (McIntosh et al., 2023).

Evaluation with LLM

In the second stage, evaluation, GPT assesses each student's response and assigns final grades for each question and the overall exam (Figure 1). To do this, GPT needs to access the reference material students study for the exam and questions based on these materials. LLM-based evaluation of students' understanding of the exam topic often requires extensive reading materials, intricate questions and detailed responses. However, this ideal scenario may not always be feasible, especially with concrete topics, exams with multiple questions and time constrained exams leading to brief student responses. When the content is insufficient for LLM to accurately gauge cognitive abilities and advanced knowledge, it is recommended to segment the evaluation and grading based on specific criteria parameters that the LLM can be instructed to assess (Wei et al., 2022).

For evaluation, exam questions can be made by the teacher or GPT, or they can be available otherwise (U. Lee et al., 2024). For assigning the final grade, GPT needs to have assessment instructions and criteria. It can provide the final grade with or without subdividing it along parameter criteria. Initially, designing such prompt (instruction) takes some time, but educational institutions and teachers can later easily and quickly adjust the prompts to suit their specific needs. Furthermore, as LLMs continue to develop rapidly, they are expected to understand instructions even better.

Various knowledge taxonomies, including Bloom´s, revised Bloom´s, SOLO and Webb´s depth of knowledge, have been employed to categorise different levels of student understanding (see, for example, Anderson et al., 2001; Biggs & Collis, 2014). These taxonomies serve as frameworks for educators to design assessments that align with the intended learning outcomes. By categorising the levels of cognitive complexity, teachers can create questions that assess students´ understanding, applying, analysing, synthesising and evaluating skills.

The choice of a taxonomy depends on the educational goals and the depth of students´ understanding desired in the assessment. The consideration of knowledge taxonomies is crucial for the evaluation performance of LLMs as these models need to be instructed (prompted) which level of knowledge is expected from responding students in a specific subject: whether students need to simply repeat parts of the study material by exactly remembering what was in there or demonstrate more advanced comprehension by applying it into new contexts and creating novel perspectives on it (Irvine, 2021).

Relevant prompt engineering is enough to instruct LLMs in the lowest knowledge taxonomy levels. Evaluating remembering is relatively easy, as it pertains to the factual accuracy of students' texts in relation to the learning material. However, evaluating higher levels of the knowledge taxonomy requires more nuanced prompts. These prompts must guide LLMs to identify how to apply learning material to new situations, solve problems in unfamiliar contexts and create novel knowledge by synthesising information to generate new ideas, products, or concepts (Anderson et al., 2001; Biggs & Collis, 2014). Potentially, LLMs might need examples from which they would learn which elements indicate students´ higher knowledge levels.

In addition, exam instructions should guide students to provide concrete evidence of their learning levels in their written responses. This would facilitate LLMs in identifying students´ knowledge levels in their written text. Conceptual knowledge would be identifiable if factual knowledge elements are connected in the response. However, when LLM inspects the evidence of metacognitive level, there is a risk of hallucinations as this level goes beyond explicitly expressed in the reference material (Wei et al., 2022). It is difficult to LLM to verify whether students' examples of higher knowledge level are relevant or not.

To ensure the consistency of evaluation outcomes and facilitate a systematic analysis of the responses, the use of the Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2021; Wei et al., 2022; Yasunaga et al., 2023) has been found useful, including to provide reference materials and evaluation guidelines to ChatGPT-4. In this framework, employed also in this article, ChatGPT-4 is prepared for evaluation task by prompting it with instructions that explicitly outline its role. Here, it was a university professor tasked with evaluating written responses from Master-level students, grounded in the provided reading materials. This verification-based chain-of-thought prompting uses also examples

of tasks created by the system itself rather than by humans. There is no need for teachers to manually label or annotate examples of reasoning for each task. This makes the process more efficient and less labour-intensive (Yasunaga et al., 2023). The use of these self-generated prompts significantly improves the accuracy of evaluations performed by LLMs (G. Lee et al., 2024). Furthermore, this approach aligns with the IDEE framework for the use of LLMs in education, namely to identify the desired outcomes, determine the appropriate level of automation, ensure ethical considerations and evaluate the effectiveness (Su & Yang, 2023).

Post-evaluation with LLM

The third stage, post-evaluation, involves LLM delivering individualised feedback to students (Figure 1). It considers the examination reading material and addresses details related to each evaluation criteria parameter along which the students were evaluated. Teachers often lack the time to provide detailed, personalised written feedback to each student. However, LLMs can offer systematic feedback on each student's response without time and resource constraints. W. Dai et al. (2023) indicated that compared with teachers, ChatGPT was capable of summarising students´ performance fluently and coherently generating more detailed textual feedback than a teacher usually can deliver. Teachers agreed highly with the feedback that ChatGPT gave on students´ assignments. This feedback benefits students and helps them to develop their learning skills further.

Material and methods

Material

The dataset consist of ChatGPT-4's evaluation results regarding 54 open-ended student responses, each evaluated along five criteria resulting in the final grade. The original written responses derived from adult Master of Science-level university students to questions related to reading materials pertinent to their respective courses at the University of Turku. Students' reading materials (8,103 words in English) comprised three distinct topics, each accompanied by reading materials sourced from articles in peer-reviewed journals, all written in English by one author of this article. The first reading material on irregular migration at the EU borderland encompassed 2,543 words. The second reading material on irregular migration during the COVID-19 pandemic spanned 3,734 words. Lastly, the third reading material with 1,816 words explored knowledge creation processes.

In practice, the receiving of student responses resembled typical university written examinations. Students responded to one reading material and three questions once in a week during 3 weeks. Students volunteered to take part in these tests and engaged with the learning material, and then received related questions in a lecture room in which they articulated their written responses typing them in English. Despite their non-native English-speaker status, English served as the language of instruction in these courses and formed a predominant component of their curriculum. Anonymity was maintained in responses, no sensitive data was collected, and all identifying markers such as name, gender and age were removed before conducting the analysis. By submitting their

response, students gave their consent to take part in the research. All adult participants could decide independently about their participation, so according to the study country and the educational institution regulations, the need for ethical board evaluation could be waived.

To make ChatGPT-4 to perform the evaluation task, it was provided with the three reading materials used for the test, the criteria for evaluation (context relevance, factual accuracy, completeness, logical consistency, as well as grammar and spelling), the university evaluation grading system consisted of six scales (fail, passable, satisfactory, good, very good, excellent). ChatGPT-4 needed to evaluate each student response along these five criteria and give the final grade.

Evaluating context relevance ensured that students' answers were pertinent to the topic and questions asked. Factual accuracy involved comparing the answer to the reference material to verify the correctness of all statements. Completeness evaluated whether the response met the requirements of the question and any sub-questions. Logical consistency checked that the arguments or statements in the answer were logical and coherent. Grammar and spelling assessed the correctness of the language used in the response. Since grammar and spelling do not necessarily indicate the student's knowledge level, they may be given less weight or omitted from the evaluation criteria.

In the evaluation process, ChatGPT-4's recalled student responses by reproducing these responses textually as accurate as possible. To verify the accuracy between original and recalled responses, text similarity analysis was utilised. If there were differences between students´ responses and ChatGPT-4´s recalled versions, these responses were reshot. Initially, 2.7% of recalls contained hallucinations, but these were removed by multiple shots as well as potential fluctuations in ChatGPT-4´s evaluation.

Finally, ChatGPT-4 evaluated responses, spanning a length from 24 to 256 words and having in total 4,261 words. This diversity in response lengths across the different reading materials and questions adds richness to the overall assessment, capturing varied perspectives of students and how well they understood the topics. Initially, ChatGPT-4 made the evaluation once (one-shot evaluation) regarding each of five elements in each 54 responses as well as the overall grading for these responses. To verify its consistency, this evaluation was repeated 10 times (10-shot evaluation). G. Lee et al. (2024) identified that few-shot practice outperformed zero-shot practice in LLM-based evaluation of students' responses. In total, ChatGPT-4 thus performed 3,240 evaluations with grades, i.e. 10 times each 54 responses that each was evaluated with five criteria and the final grade.

For comparison, the teacher (professor) responsible for the courses performed the same tasks as ChatGPT-4. The teacher was non-native English speaker, however, confident and experienced in using English as language of tuition. Furthermore, he was the sole author of the reading materials used for the test, thus competent to evaluate students´ responses. Usually, teachers provide only the final grade to each response. Howeverin this case, the teacher needed to evaluate responses both regarding the final grade as well as by the agreed five criteria, making in total 324 evaluations.

Methods

Once gathered all data, it was processed in Python 3.11 environment and Excel. For empirical material was employed statistical methods such as calculating the average,

mode and standard deviation on three main research topic aspects. The average values helped in identifying the central tendency of the evaluation data, while the mode indicated the most frequently occurring values. Additionally, the standard deviation offered insights into the variability and consistency of the evaluation that ChatGPT-4 made.

Overall, the article first analysed the consistency of ChatGPT-4's evaluations based on five detailed criteria and the final grades assigned to each student's response. This involved comparing the initial evaluation (one-shot) with 10 repeated evaluations (10-shots) to assess consistency and variability in the LLM's performance. Next, the article identified variations in ChatGPT-4's grading to determine if differences were random or specific to certain responses or students. Finally, it compared the grades assigned by ChatGPT-4 and the teacher for each criterion, using statistical analysis to identify disparities. This included calculating the mode and average grades from ChatGPT-4 and comparing them with the teacher's grades.

Results

Consistency of ChatGPT-4 in student evaluation

This section answers the first research question on how ChatGPT-4 evaluated students' open-ended written responses and how consistent it was in this task. Following the prompt, ChatGPT-4 conducted an initial evaluation (first-shot) for each of the 54 student responses by grading them according to the five specified criteria before assigning a final grade. ChatGPT-4 thus performed a total of 324 first-shot evaluations, which is the sum of 270 evaluations for the five criteria per response plus 54 evaluations for the final grades, across all different responses (Table 1).

To assess the consistency of ChatGPT-4´s evaluations, the model was tasked with repeating the evaluation process 10 times for each response, known as the 10-shot evaluation. Consequently, ChatGPT-4 carried out 2,700 evaluations for the responses based on the specified five criteria (10 evaluations per criterion for each of the 54 responses) and 540 evaluations for the final grades (10 evaluations per response). This resulted in a total of 3,240 evaluations.

The findings, including the average and standard deviation of these 10-shot evaluations for each criterion and the final grades, are detailed in Table 2. Out of the final grades, ChatGPT-4 assigned during the 10-shot evaluations, more than two out of three (371, 68.7%) grades were consistent across all evaluations, more than a fourth (148, 27.4%) varied by one grade, a few (21, 3.9%) varied two grades and none showed a difference of three grades when compared to the mode value of the final grade for each response.

Table 1. Final grade given to student responses by LLMs (Gpt3.5, Gpt4, Claude3 and mistral-large) along 10-shot evaluation with temperature 0.0% and 0.5%, number of cases, in total 4,298 evaluations.

	Fail – 0	Passable – 1	Satisfactory – 2	Good – 3	Very Good – 4	Excellent – 5
0.0 (2149)	6.75 (145)	19.64 (422)	35.18 (756)	17.4 (374)	11.40 (245)	9.63 (207)
0.5 (2149)	5.58 (120)	22.43 (482)	33.60 (722)	18.8 (404)	10.7 (230)	8.89 (207)
Total (4298)	6.17 (265)	21.03 (904)	34.39 (1478)	18.1 (778)	11.05 (475)	9.26 (398)

Table 2. Final grade given to student responses by different LLMs (Gpt3.5, Gpt4, Claude3 and mistral-large) along 10-shot evaluation with temperature 0.0% and 0.5%, number of cases, in total 4,298 evaluations.

	Fail – 0	Passable – 1	Satisfactory – 2	Good – 3	Very Good – 4	Excellent – 5
Claude3, 0.0 (540)	1.67 (9)	10.37 (56)	48.33 (261)	18.15 (98)	21.48 (116)	0.00 (0)
Gpt3.5, 0.0 (529)	14.37 (76)	34.59 (183)	8.51 (45)	13.42 (71)	9.07 (48)	20.04 (106)
Gpt4, 0.0 (540)	7.41 (40)	11.85 (64)	49.26 (266)	21.48 (116)	5.37 (29)	4.63 (25)
Mistral-Large, 0.0 (540)	3.70 (20)	22.04 (119)	34.07 (184)	16.48 (89)	9.63 (52)	14.07 (76)
Claude3, 0.5 (540)	1.48 (8)	13.52 (73)	43.33 (234)	22.59 (122)	19.07 (103)	0.00 (0)
Gpt3.5, 0.5 (529)	11.91 (63)	36.86 (195)	10.02 (53)	12.67 (67)	10.96 (58)	17.58 (93)
Gpt4, 0.5 (540)	7.41 (40)	15.37 (83)	46.85 (253)	21.11 (114)	4.63 (25)	4.63 (25)
Mistral-Large, 0.5 (540)	1.67 (9)	24.26 (131)	33.70 (182)	18.70 (101)	8.15 (44)	13.52 (73)

ChatGPT-4 was thus very consistent in final gradings as 96.1% of its evaluations were within one grade. ChatGPT-4 had particular challenges to evaluate consistently responses by one student (A11) as evidenced by standard deviation. Then, the teacher could verify the accuracy of that evaluation by reading the response (Table 2).

Among the five (context relevance, factual accuracy, completeness, logical consistency and grammar and spelling) evaluation parameter criteria, ChatGPT-4 yielded 1,933 out of the 2,700 evaluations (71.6%) with identical grades across all 10 evaluations for the same response. Meanwhile, 557 evaluations (20.6%) experienced a variance of one grade, and few evaluations (183, 6.8%) had a variance of two grades when compared to the mode value. This demonstrates a high level of consistency across the evaluations as ChatGPT-4 as 92.2% of its parameter evaluations were within one grade (Table 2).

Comparison between ChatGPT-4 evaluation and the teacher evaluation

This section answers the second research question on disparities between ChatGPT-4 and human evaluations of students´ written responses by comparing the grading as evaluation results between ChatGPT-4 and the teacher. Initially, the teacher graded all five criteria and the final grade of each student´s response a six-scale spectrum (fail, passable, satisfactory, good, very good, excellent), which reflects the university´s grading system. Subsequently, these grades were compared with the mode values obtained from ChatGPT-4´s 10-shot evaluation for each student´s response, following the same evaluation metrics.

The results highlight similarities and differences between ChatGPT-4 and the given teacher as evaluators (Table 3). Of the 54 final grades assigned by the teacher to student responses, slightly less than three out of four (72.2%) of final grades were same or within one grade difference between the teacher and the LLM used for this evaluation. Approximately a fourth (14, 25.9%) was identical with the 10-shot mode grade given by ChatGPT-4, less than a half (25, 46.3%) was one grade different, almost a fourth (13, 24.1%) was different with two grades, and very few (2, 3.7%) were more than grades different from ChatGPT-4´s final evaluation.

The teacher in this test typically gave slightly higher grades than those assigned by ChatGPT-4, a pattern that was consistent across all questions, evaluation through parameters, and in the calculation of the final grade (Table 4). The closest alignment between ChatGPT-4 and the teacher was observed in the evaluation of students´ responses for

Table 3. Score differences from the benchmark LLM grade for final grade of student responses, 10-shot evaluation with temperature 0.0 and 0.5 variants (%, number of cases, in total 4,298 evaluations; in green the highest performance results, in red the lowest performance results).

	Inaccurate	Minor Deviation	Accurate	Minor Deviation	Inaccurate	Inaccurate
	+2	+1	0	−1	−2	Other
Claude3, 0.0 (540)	3.89 (21)	13.89 (75)	62.78 (339)	10.37 (56)	2.22 (12)	6.85 (37)
Gpt3.5, 0.0 (529)	11.91 (63)	11.34 (60)	24.95 (132)	23.33 (126)	14.93 (79)	13.04 (69)
Gpt4, 0.0 (540)	1.67 (9)	15.19 (82)	58.15 (314)	15.93 (86)	3.89 (21)	5.19 (28)
Mistral-Large,0.0 (540)	4.07 (22)	10.00 (54)	56.67 (306)	17.41 (94)	2.22 (12)	9.63 (52)
Claude3, 0.5 (540)	6.30 (34)	25.19 (136)	48.70 (263)	14.63 (79)	1.67 (9)	3.52 (19)
Gpt3.5, 0.5 (529)	13.23 (70)	10.96 (58)	29.68 (157)	24.01 (127)	13.23 (70)	8.88 (47)
Gpt4, 0.5 (540)	3.89 (21)	13.52 (73)	59.07 (319)	16.48 (89)	2.22 (12)	4.81 (26)
Mistral-Large,0.5 (540)	8.89 (48)	19.26 (104)	43.15 (233)	18.70 (101)	4.26 (23)	5.74 (31)

logical consistency, while the greatest disparity was noted in assessing grammar and spelling (Table 5). The latter, often not emphasised in the university´s content-related assessments, saw the teacher applying stricter standards compared to ChatGPT-4. Conversely, when evaluating the completeness of student responses, ChatGPT-4 was found to be more stringent than the teacher, who did not require that every specific detail from the reading materials would be present in student responses.

Solving evaluation discrepancies between ChatGPT-4 and teacher

This section addresses the third research question: how to resolve the disparities between ChatGPT-4 and human grading of student responses. As mentioned, about three-quarters (72.2%) of gradings showed that ChatGPT-4's and the teacher's final grades were identical or differed by only one grade. However, in 27.8% of evaluations, the difference exceeded one grade, which is significant given the six-grade scale used. This large discrepancy can be approached from two perspectives.

If both the teacher and the LLM evaluated the same responses and significant grading differences emerged only from specific responses, this could indicate a potential error in human evaluation. These responses could be flagged for the teacher to confirm whether the initial human-made grading was correct. Ultimately, the teacher may choose to either concur with ChatGPT-4's assessment and grade or provide an alternative, thus making an informed decision to conclude the grading process.

Conversely, if there was a systematic difference in grading results between the LLM and the teacher, and it was established that the teacher's grading was generally accurate, adjustments could be made. One approach is to refine the instructions given to the LLM so that its evaluation criteria align more closely with the teacher's and the educational institution's expectations. Another approach involves calibrating the LLM's grading with a coefficient that adjusts the initial grading results. For example, in this test, the average final grade awarded by ChatGPT-4 for 54 responses was 3.5, while the teacher's average final grade was 4. The grading difference of 0.5 points between ChatGPT-4 and the teacher could be adjusted with a coefficient to bring their final grades closer together.

If it is determined that the teacher's and LLM's grading aligns on most parameters but diverges on a non-fundamental parameter, that parameter can be disregarded from the

Table 4. Share of cases without variation within their 10 shot-grading regarding the final grade and evaluation parameters (%, number of cases, in total 54 evaluation groups.

	Range	Context Relevance	Factual Accuracy	Completeness	Logical Consistency	Grammar & Spelling	Final Grade
Claude3, 0.0	0	74.93 (41)	75.93 (41)	74.07 (40)	68.52 (37)	61.11 (33)	70.37 (38)
	1	24.07 (13)	22.22 (12)	24.07 (13)	31.48 (17)	33.33 (18)	27.78 (15)
	2	0.00 (0)	1.85 (1)	1.85 (1)	0.00 (0)	3.70 (2)	1.85 (1)
	3	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.85 (1)	0.00 (0)
	4	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Gpt3.5, 0.0	0	31.48 (17)	20.37 (11)	37.04 (20)	20.37 (9)	16.67 (9)	18.52 (10)
	1	40.74 (22)	44.44 (24)	33.33 (18)	44.44 (24)	44.44 (24)	37.04 (20)
	2	24.07 (13)	27.78 (15)	20.37 (11)	22.22 (17)	31.48 (17)	31.48 (17)
	3	3.70 (2)	5.56 (3)	5.56 (3)	7.41 (3)	5.56 (3)	5.56 (3)
	4	0.00 (0)	1.85 (1)	1.85 (1)	9.26 (1)	1.85 (1)	7.41 (4)
	5	0.00 (0)	0.00 (0)	1.85 (1)	0.00 (0)	0.00 (0)	0.00 (0)
Gpt4, 0.0	0	37.04 (20)	24.07 (13)	37.04 (20)	24.07 (13)	53.70 (29)	35.19 (19)
	1	48.15 (26)	64.81 (35)	46.30 (25)	64.81 (35)	24.07 (13)	55.56 (30)
	2	9.26 (5)	7.41 (4)	14.81 (8)	9.26 (5)	16.67 (9)	5.56 (3)
	3	5.56 (3)	3.70 (2)	1.85 (1)	1.85 (1)	3.70 (2)	0.00 (0)
	4	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.85 (1)	3.70 (2)
	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Mistral-Large, 0.0	0	88.89 (48)	88.89 (48)	87.04 (47)	88.89 (48)	87.04 (47)	83.33 (45)
	1	9.26 (5)	9.26 (5)	12.96 (7)	9.26 (5)	11.11 (6)	16.67 (9)
	2	1.85 (1)	1.85 (1)	0.00 (0)	1.85 (1)	1.85 (1)	0.00 (0)
	3	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
	4	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Claude3, 0.5	0	33.33 (18)	22.22 (12)	33.33 (18)	33.33 (18)	31.48 (17)	12.96 (7)
	1	59.26 (32)	53.70 (29)	53.70 (29)	50.00 (27)	37.04 (20)	70.37 (38)
	2	7.41 (4)	22.22 (12)	12.96 (7)	16.67 (9)	24.07 (13)	16.67 (9)
	3	0.00 (0)	1.85 (1)	0.00 (0)	0.00 (0)	5.56 (3)	0.00 (0)
	4	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.85 (1)	0.00 (0)
	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Gpt3.5, 0.5	0	11.11 (6)	9.26 (5)	12.96 (7)	9.26 (5)	7.41 (4)	12.96 (7)
	1	16.67 (9)	24.07 (13)	44.44 (24)	24.07 (13)	20.37 (11)	20.37 (11)
	2	35.19 (19)	37.04 (20)	18.52 (10)	38.89 (21)	40.74 (22)	33.33 (18)
	3	35.19 (19)	25.93 (14)	16.67 (9)	22.22 (12)	29.63 (16)	20.37 (11)
	4	1.85 (1)	3.70 (2)	7.41 (4)	5.56 (3)	0.00 (0)	12.96 (7)
	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.85 (1)	0.00 (0)
Gpt4, 0.5	0	18.52 (10)	7.41 (4)	16.67 (9)	22.22 (12)	48.15 (26)	20.37 (11)
	1	53.70 (29)	68.52 (37)	59.26 (32)	44.44 (24)	22.22 (12)	59.26 (32)
	2	22.22 (12)	22.22 (12)	12.96 (7)	29.63 (16)	18.52 (10)	16.67 (9)
	3	3.70 (2)	0.00 (0)	9.26 (5)	3.70 (2)	9.26 (5)	3.70 (2)
	4	1.85 (1)	1.85 (1)	1.85 (1)	0.00 (0)	1.85 (1)	0.00 (0)
	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Mistral-Large, 0.5	0	33.33 (18)	29.63 (16)	35.19 (19)	24.07 (13)	25.93 (14)	14.81 (8)
	1	37.04 (20)	38.89 (21)	40.74 (22)	46.30 (25)	44.44 (24)	42.59 (23)
	2	27.78 (15)	31.38 (17)	20.37 (11)	27.78 (15)	27.78 (15)	35.19 (19)
	3	1.85 (1)	0.00 (0)	3.70 (2)	1.85 (1)	0.00 (0)	7.41 (4)
	4	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.85 (1)	0.00 (0)
	5	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)

final grading. For instance, in the test discussed in this article, there was a notable difference between the teacher and ChatGPT-4 regarding grammar and spelling. This parameter could be omitted from the overall evaluation to lessen its impact on the final grade.

This adaptability suggests that ChatGPT-4 can be customised to match the grading style of each teacher and the specific requirements of every exam evaluation, should there be a need. However, it remains crucial that the teacher and the educational

Table 5. Evaluation cases without any variation within their 10-shot evaluation.

	Context Relevance	Factual Accuracy	Completeness	Logical Consistency	Grammar & Spelling	Final Grade
Claude3, 0.0	74.93	75.93	74.07	68.52	61.11	70.37
Gpt3.5, 0.0	31.48	20.37	37.04	20.37	16.67	18.52
Gpt4, 0.0	37.04	24.07	37.04	24.07	53.70	35.19
Mistral-Large, 0.0	88.89	88.89	87.04	88.89	87.04	83.33
Claude3, 0.5	33.33	22.22	33.33	33.33	31.48	12.96
Gpt3.5, 0.5	11.11	9.26	12.96	9.26	7.41	12.96
Gpt4, 0.5	18.52	7.41	16.67	22.22	48.15	20.37
Mistral-Large, 0.5	33.33	29.63	35.19	24.07	25.93	14.81

institution retain ultimate authority over determining the final grades for all student responses.

Conclusions

There is considerable speculation and anticipation regarding the potential of LLMs like ChatGPT to transform educational assessment, particularly in evaluating students' overall performance and their written exams (Bahroun et al., 2023; Baidoo-Anu & Owusu Ansah, 2023; Farrokhnia et al., 2023; Gimpel et al., 2023; Holmes & Miao, 2023; Su & Yang, 2023; Yu & Guo, 2023). Despite these expectations, systematic analyses of LLMs in evaluating open-ended student responses are still sparse.

Before deploying ChatGPT-4 or similar LLMs for evaluating written responses, it is crucial to tailor these models for educational assessments. This involves using verification-based chain-of-thought prompting to align the models with specific evaluation criteria, as suggested by Wei et al. (2022). Initial prompt creation may take time, but teachers can easily modify prompts later, and future LLM developments will simplify this process further. Ensuring accurate recall of student responses, such as through text similarity analysis, is vital, with any discrepancies corrected by revisiting the responses. Testing must be conducted ethically in a secure environment to prevent data leakage. This systematic approach is straightforward but requires user awareness. Training sessions for teachers are essential for effective use of LLMs in student evaluation. In this, the systematic design of LLMs for pre-evaluation, evaluation and post-evaluation supports the assessment of students' open-ended responses.

The findings of this article demonstrate the potential of ChatGPT-4 as an effective tool for evaluating students' written texts, essays and exams. ChatGPT-4 successfully assessed the factual accuracy of responses against reference learning materials, performing well at least at lower levels of the knowledge taxonomy. However, higher-level evaluations require precise instructional prompts, detailed student responses and explicit instructions for students to demonstrate their higher-level knowledge, which were not requires in this study. Using LLMs systematically to assist in evaluating and marking written exams and essays could save teachers time, allowing them to focus on enhancing student learning. Integrating such an evaluation system into higher-order assessment processes in educational institutions can be achieved by using and designing evaluation platforms that incorporate LLMs.

A crucial aspect of using ChatGPT-4 is the consistency of its evaluations. It is recommended to conduct multiple evaluations on each response since initial assessments may yield lower performance than subsequent ones. In our findings, 31.2% of the initial grades changed upon re-evaluation, with 25.7% changing by more than one grade level. Repeated evaluations help identify inaccuracies and inconsistencies in LLM's grading. In a specific test where 54 responses were evaluated ten times each, ChatGPT-4's final grade to the answer was consistent in 68.7% of cases, varied by one grade in 27.4%, and by two grades in only 3.9%. For a five-criteria evaluation involving 2,700 gradings, the consistency was 71.6%, with variations of one and two grades at 20.6% and 6.8%, respectively.

Discrepancies between human (teacher) evaluations and LLM (ChatGPT-4) assessments were noted, indicating potential errors in human grading or inaccuracies in LLM grading. Who were right or wrong was not addressed in this article. Systematic differences between LLM and human gradings can be addressed by refining the prompts used, adjusting the LLM's grading results, or eliminating unnecessary evaluation criteria that contribute to grading discrepancies.

This study has limitations that indicate areas for further research. Future studies should evaluate the performance of various LLMs, not just one, across different subjects and academic disciplines, languages and educational settings. It is crucial to examine how well these models assess higher cognitive skills and knowledge levels according to various educational standards. This research is essential for tailoring LLMs to meet the specific evaluation criteria of educational institutions and teachers. Proper comparing of the evaluation performance between LLMs and human evaluators requires large datasets and a systematic longitudinal approach.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the FAPI (University of Turku) [26004172].

Notes on contributors

Jussi S. Jauhiainen is Professor at the University of Turku (Finland) and Visiting Professor at the University of Tartu (Estonia). His research interests regard innovations, generative AI, LLMs in education and the Metaverse. He is developing an LLM-based evaluation platform TurkuEval (see sites.utu.fi/digileac).

Agustín Garagorry Guerra is Research Assistant at the University of Turku (Finland). His research interests regard generative AI and LLMs in education. He is developing an LLM-based evaluation platform TurkuEval (see sites.utu.fi/digileac).

ORCID

Jussi S. Jauhiainen http://orcid.org/0000-0001-8095-8240

References

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

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman.

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

Bewersdorff, A., Sessler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, *5*, 100177. https://doi.org/10.1016/j.caeai.2023.100177

Biggs, J., & Collis, K. (2014). *Evaluating the quality of learning: The Solo taxonomy (Structure of the observed learning outcome)*. Academic Press.

Chiu, T. (2023). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, 1–17. https://doi.org/10.1080/10494820.2023.2253861

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. *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323–325). Orem, UT, USA. https://doi.org/10.1109/ICALT58122.2023.00100.

Dai, Y., Liu, A., & Lim, 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

Farrokhnia, M., Banihashem, S., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and ai-human collaboration. *Journal of Information Technology Case & Application Research*, *25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Maedsche, A., Röglinger, M., Ruiner, C., Schoch, M., Schoop, M., Urbach, N., & Vandirk, S. (2023). *Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers*. Preprint. https://doi.org/10.13140/RG.2.2.20710.09287/2

Holmes, W., & Miao, F. (2023). *Guidance for generative AI in education and research*. UNESCO. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Irvine, J. (2021). Taxonomies in education: Overview, comparison, and future directions. *Journal of Education and Development*, *5*(2), 1. 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. (2024). *Evaluating students' open-ended written responses with LLMs: Using the RAG framework for GPT-3.5, GPT-4, Claude-3, and mistral-large*. arXiv: 2405:05444. https://doi.org/10.48550/arXiv.2405.05444

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*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

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.

Lee, G., Latif, E., Wu, X., Liu, N., & Zhai, Z. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, *29*(9), 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*. arXiv:2005.11401. https://doi.org/10.48550/arXiv.2005.11401

Lo, C. (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

Mannuru, N., Shahriar, S., Teel, Z., Wang, T., Lund, B., Tijani, S., Pohboon, C., Agbaji, D., Alhassan, J., Galley, J., Kousari, R., Ogbadu-Oladapo, L., Saurav, S., Srivastava, A., Tummuru, S., Uppala, S., & Vaidya, P. (2023). Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. *Information Development*. https://doi.org/10.1177/02666669231200628

McIntosh, T., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. (2023). 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 essays scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

OECD. (2023). *PISA 2022 results (volume I): The state of learning and equity in education, PISA*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Skaalvik, E., & Skaalvik, S. (2021). Teacher burnout: Relations between dimensions of burnout, perceived school context, job satisfaction and motivation for teaching. A longitudinal study. *Teachers & Teaching*, *26*(7–8), 602–616. https://doi.org/10.1080/13540602.2021.1913404

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

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*. arXiv:2201.11903v6: https://doi.org/10.48550/arXiv.2201.11903

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., Chi, E., & Zhou, D. (2023). *Large language models as analogical reasoners*. arXiv:2310.01714v2. 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