Advances in Artificial Intelligence and Machine Learning; Research 4 (4) 3097-3113   Received 5-10-2024; Accepted 21-12-2024; Published 28-12-2024

Evaluating Students’ Open-ended Written Responses with LLMs: Using the RAG Framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large

Jussi S. Jauhiainen jusaja@utu.fi
Department of Geography and Geology,
University of Turku, Vesilinnantie 5, FI-20014 Turku,
Finland
Institute of Ecology,
and the Earth Sciences, University of Tartu
Estonia.

Agustín Garagorry Guerra
Department of Geography and Geology,
University of Turku, Vesilinnantie 5, FI-20014 Turku,
Finland

Corresponding Author: Jussi S. Jauhiainen

Abstract

Evaluating open-ended written examination responses from students is an essential yet time-intensive task for educators, requiring a high degree of effort, consistency, and precision. Recent developments in Large Language Models (LLMs) present a promising opportunity to balance the need for thorough evaluation with efficient use of educators’ time. We explore LLMs—GPT-3.5, GPT-4, Claude-3, and Mistral-Large—in assessing university students’ open-ended responses to questions about reference material they have studied. Each model was instructed to evaluate 54 responses repeatedly under two conditions: 10 times (10-shot) with a temperature setting of 0.0 and 10 times with a temperature of 0.5, expecting a total of 1,080 evaluations per model and 4,320 evaluations across all models. The RAG (Retrieval Augmented Generation) framework was used to make the LLMs to process the evaluation. Notable variations existed in studied LLMs consistency and the grading outcomes. There is a need to comprehend strengths and weaknesses of using LLMs for educational assessments.

Keywords: LLM, GPT, Claude, Mistral-Large, Education, Evaluation, Open-ended responses, RAG

1. INTRODUCTION

Large Language Models (LLM) are being integrated into workflows by a variety of users, institutions, and stakeholders since the launch of ChatGPT-3.5 in November 2022. Subsequently, this and other LLMs have been adopted for many kinds of uses. The educational sector, in particular,

Citation: Jussi S. Jauhiainen and Agustín Garagorry Guerra. 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. 2024;4(4):177.

has seen significant impact in this. Educators are using LLMs for adapting and creating content and evaluating students' performance, and students leverage these tools for assistance in writing essays and completing assignments. Despite their benefits, the adoption of LLMs also presents challenges. These include LLMs generating inaccurate ("hallucinated") content, security challenges with user data leaking for model training and ethical issues for using the models inappropriately. LLMs contain biases that stem from their training data, architecture, and hyperparameter settings [1–4].

A promising application within education is to use LLMs to assess students' open-ended responses. Evaluating open-ended examination responses is time-consuming and labor-intensive task for educators. However, if LLMs can effectively handle this task or at least support teachers in it, they could significantly reduce teacher workload, and potentially enhance their job satisfaction and results. This could, in turn, improve the overall quality of the learning environment, benefiting both teachers and students. However, implementing LLMs in educational settings demands a thorough understanding of their capabilities and limitations to ensure that their integration helps to achieve educational goals while mitigating potential risks [5, 6]. Research on the use of LLMs in education is still developing.

In this article, we explore the effectiveness of LLMs in assessing open-ended written examination responses from university students. Our study provides a comparative analysis of various LLMs' performance in educational contexts, focusing on integrating these tools in intelligent, secure, transparent, and cost-efficient ways. We introduce a methodological framework to enhance the effectiveness of LLMs in education.

We utilized the respective APIs for all models included in this study: gpt-3.5-turbo, gpt-4-0125-preview, claude-3-opus-20240229, and Mistral-Large-large-latest. These models will be referred in this article to as Gpt3.5, Gpt4, Claude3, and Mistral-Large, respectively. Our investigation focuses on the feasibility of implementing these models in educational settings, including the accuracy of their grading, consistency of grading results, processing speed, control over the model, and the costs associated with computational resources.

Our investigation is structured around four primary research questions: What are the main characteristics of the evaluation process when open-ended written responses are evaluated and graded with Gpt3.5, Gpt4, Claude3, and Mistral-Large?; What are the differences between the grades assigned by these LLMs?; How consistent are the grades assigned by these LLMs?; What were the processing times of these LLMs to perform the evaluation?

2. LLM IN EDUCATIONAL ENVIRONMENT

2.1 Elements Behind Llms' Capacity to Evaluate Written Texts

In general, LLMs are sophisticated deep learning algorithms and models excelling in tasks such as summarization, recognition, translation, prediction, and content generation, built on vast datasets for training [7]. The educational sector has shown keen interest in LLM advancements, experiencing a blend of beneficial and challenging impacts. Three pivotal developments have catalyzed the rise of LLMs, particularly in the evaluation of written educational texts.

The first element is the advancement in machine translation, where models utilize architectures that feature the ability to "soft-search" for relevant parts of a source sentence to predict the desired outcome. This breakthrough has significantly enhanced the model's ability to focus on relevant text segments, improving context processing capabilities [8].

The second significant advancement is the development of the Transformer architecture by Google DeepMind [9]. Transformers employ multi-head attention layers that enable the model to process various word characteristics simultaneously, thus enhancing both efficiency and performance. Unlike models based on recurrent or convolutional layers, Transformers can be trained more rapidly and are more scalable. LLMs based on this architecture use mechanisms like Temperature, Top-k and Top-p to generate diverse text outputs. The temperature parameter influences the randomness of predictions—lower values like 0.0 produce more predictable text, while higher values like 0.5 introduce greater creativity in text prediction. The Top-k sampling restricts the model to only consider the top 'k' probable next words, while the Top-p sampling uses a cumulative probability threshold to select the next words, adding flexibility and nuance to text generation [10].

The third element is the ChatGPT model that combines the interactive format of chatbots with the generative capabilities of an LLM and the robust processing power of the Transformer architecture. This combination of interactivity, generative capacity, and sophisticated architecture enhances LLMs' utility in contemporary applications, notably in educational settings where they can be leveraged for tasks such as evaluating students' written texts, facilitating adaptive learning environments, and providing automated feedback [1, 2, 4, 7, 11, 12]. These capabilities underline the transformative potential of LLMs in education, necessitating careful consideration of their deployment to maximize benefits while addressing inherent challenges efficiently. Top of Form

2.2 Assessing Open-Ended Written Responses With Llms

The assessment of open-ended responses for examination questions is necessary but burdensome for educators, especially when applied to large groups of students. There are also risks of human errors and subjectivity as well as interpretation differences between human evaluators. To alleviate these challenges, automated computer-assisted evaluation systems have been developed. Historically, the automation of open-ended answer evaluation has incorporated technologies such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Transformers, including implementations such as using BERT for automated scoring [13, 14].

LLMs have proven versatile in various educational settings to create adaptive learning environments that dynamically adjusted learning materials to match the learner's skill level, thereby enhancing student engagement with the learning materials [11, 12]. Most studies so far have focused on discussing the potential of generative AI technologies rather than analyzing systematically LLMs' concrete effects on educational practices [1, 2, 7] or explored ChatGPT-3.5 in responding to open-ended questions [15, 16], or providing detailed feedback for students [17].

Employing LLMs for evaluating students' written texts requires a systematic approach. LLMs need to access the students' examination materials and corresponding questions. They must accurately interpret students' responses and adhere to educational evaluation guidelines. Finally, they need to follow the grading system and assign grades based on correct and consistent grading.

3. MATERIAL AND METHODS

The study was based on 54 open-ended responses from students enrolled in a master's level geography course taught in English at the University of Turku. The professor, who is an expert in the subject matter, selected three scholarly articles to form the basis of the questions designed to assess the students' understanding about the learning material after they read it. Each article was accompanied by three corresponding questions about the content of the learning material. This approach simulates a typical university exam setting where students are required to demonstrate their knowledge within a specified time. For the test, we used only text-based materials to eliminate any potential "noise" that visual elements like images or tables might introduce when using LLMs for evaluation as not all models perform enough well in multimodal contexts.

The test involved analyzing three open-ended responses for questions derived from one reference material, with a total of three reference materials and nine questions used for the test. The responses in English spanned a length from 24 to 256 words with average length of 152 words. Each LLM was tasked with evaluating each of the 54 different student responses 10 times under a 10-shot scenario at temperature settings of 0.0 and another 10 times at the temperature setting of 0.5. Consequently, each LLM conducted a total of 540 + 540 evaluations, cumulating in a grand total of 4,320 (of which 23 were truncated or incomplete) evaluations of student responses having in total 656,640 words.

Our methodology for evaluating the performance of LLMs in assessing student responses involved several structured steps. Firstly, we obtained consent from all participating students to include their responses in the study. We removed all personal identifiers, thereby maintaining anonymity when analyzed with the LLMs. We did not collect sensitive information such as gender, age, or ethnicity.

Secondly, we collected the student responses and utilized the Langchain Open AI Embedding method to convert the text into numerical representations. This was part of our data generation process. Reference texts were first processed using the PyPDF library, which allowed us to segment documents into 500-token chunks with a 20-token overlap. This specific granularity was selected to provide the LLMs with adequate context for accurately evaluating the student responses, while preventing token overflow in model inputs. For each student's response, we calculated cosine similarities with the document chunks, selecting the top five most relevant chunks (k=5). These chunks were then reorganized to optimize the retrieval process, thus enhancing the LLM's efficiency in referencing pertinent information during the evaluation process as outlined by [18]. We followed the RAG (Retrieval-Augmented Generation) technique (FIGURE 1). The use of RAG is useful in applications where the quality of output benefits from specific information, such as in question answering systems, content creation, and advanced chatbot functionalities. We mainly focused on Temperature and Prompt with RAG, which are among the most common approaches when implementing LLMs in development.

Thirdly, to ensure the reliability of our research prompts, we adhered to established standards such as verification-based chain-of-thought (CoT) prompting [19] along with other advanced prompting techniques. This helped us to confirm the consistency and validity of the prompts used for the LLMs' evaluation processes. Various prompts were tested. In our methodology, each LLM was finally provided with the same prompts, reference materials, individual student responses, corresponding questions, and detailed evaluation guidelines. This comprehensive setup equipped the LLMs with the necessary context to accurately assess the students' knowledge (FIGURE 1).

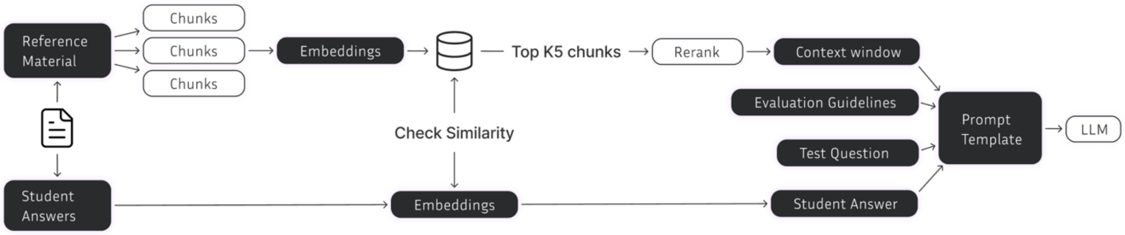

Figure 1: Evaluation process for LLM evaluation with reranked RAG.

Additionally, the prompts were customized to reflect the educational context, indicating roles such as the "University Professor" and specifying the academic level as master's degree (see Appendix). This customization was integral to tailoring the evaluation process to the unique requirements of the educational setting.

The evaluation criteria employed in this study comprised several parameters designed to address different aspects of students' responses as analyzed by the LLMs. These parameters included content completeness, factual accuracy, logical consistency, contextual relevance, and grammar and spelling. Each of these components was crucial for conducting a detailed and comprehensive assessment of the LLMs' performance in educational applications.

The final grade was determined by the combined influence of these parameters, with each parameter assigned in this case an equal weight of 20% of the total grade. The choice and implementation of these evaluation criteria were thoroughly discussed and validated by a diverse group of educators, including teachers, pedagogical experts, and university staff. The objective was not only to assess the overall grading accuracy of the LLMs but also to evaluate specific elements of student-written responses (FIGURE 1).

Fourthly, we adopted a grading scale based on the Finnish higher education system, which includes grades Fail (0), Passable (1), Satisfactory (2), Good (3), Very Good (4), and Excellent (5). This scale was familiar to the participating university, teachers, and students, thereby enhancing the usability and relevance of our results. Each grade was initially presented in verbal form and subsequently converted into numerical values during the data cleaning phase for more straightforward analysis. This grading scale was applied to evaluate both the overall final grade of each response and the specific evaluation parameters used by the LLMs to assess the responses.

Fifthly, we conducted multiple evaluations for each student's answer using a 10-shot scenario, where each LLM assessed each response 10 times. In this context, the term "shot" refers to the repeated, independent processing of a student's response by an LLM, which allowed us to observe and document variability in performance across multiple iterations. This method of repeating several times the evaluation was vital for identifying potential inconsistencies in the models, which could lead to biased or inaccurate evaluation results. This iterative evaluation process helped us to identify patterns and characteristics in each LLM's behavior. This provided useful insights for educators, students, and researchers considering the use of these tools in educational settings.

Sixthly, to quantitatively assess and compare the performance differences among the LLMs, we employed Python programming language and Pandas library. Our statistical analysis primarily focused on descriptive statistics, which included calculating percentage distributions, means, and standard deviations to summarize the data effectively. Additionally, we conducted cross-tabulation analyses to determine the statistical significance of our findings, utilizing Pearson's chi-square test and Spearman's rank correlation.

4. RESULTS

As highlighted in the introduction, a primary objective of our analysis was to delineate the key features of the student open-ended answer evaluation process using Gpt3.5, Gpt4, Claude3, and Mistral-Large. We specifically focused on discerning the differences in the grades assigned by these LLMs, examining the reliability and consistency of grading results, and assessing the processing time required by each LLM during the evaluation.

4.1 Differences in Grading Across LLMs

Four LLMs used for this test detected variations in the quality of students' open-ended responses, resulting in the assignment of different final grades (TABLE 1). The most frequently assigned grade by all LLMs together was Satisfactory (2), attributed to 34.39% of the responses, followed by Passable (1) at 21.03%, Good (3) at 18.1%, and Very Good (4) at 11.05%. Grades at the extremes of the evaluation scale, such as Excellent (5) and Fail (0), were less common, given to 9.63% and 6.15% of student responses, respectively.

Table 1: Share and number of total 4,298 Final Grades assigned by LLMs (Gpt3.5, Gpt4, Claude3, and Mistral-Large) to student responses along 10-shot evaluation with temperature 0.0 (n=2,149) and 0.5 (n=2,149).

	Fail	Passable	Satisfactory	Good	Very Good	Excellent
	grade 0	grade 1	grade 2	grade 3	grade 4	grade 5
	% nr	% nr	% nr	% nr	% nr	% nr
Temperature 0.0	6.75 145	19.64 422	35.18 756	17.40 374	11.40 245	9.63 207
Temperature 0.5	5.58 120	22.43 484	33.90 722	18.80 404	10.70 230	8.89 191
Total	6.17 265	21.03 904	34.29 1478	18.10 778	11.05 475	9.26 398

Using the temperature setting 0.0 tended to yield slightly more often grades at the both ends of the evaluation spectrum (either Fail or Excellent) compared to using the temperature 0.5. Furthermore, the proportion of responses rated as Good (3) was noticeably lower with temperature 0.0 than with temperature 0.5 (TABLE 1). With a higher temperature, LLMs can implement more creativity in their grading rather than following the prompt instructions rigorously. These variations in the grading outcomes across the different models highlight the influence of model settings and selection on the assessment of student responses.

The evaluation results of the studied LLMs in terms of grades assigned to students' responses showed significant disparities (TABLE 2). For instance, as regards the grade Fail (0), Gpt3.5 at temperature 0.0 assigned this grade to 14.37% of student responses, whereas Claude3 at temperature 0.5 did so for only 1.48% of responses. Additionally, Gpt4 issued at least three times as many Fail (0) grades as Mistral-Large, and this discrepancy widened to over seven times when comparing Gpt3.5 with Claude3. Meanwhile, Gpt4 at temperature 0.0 awarded the grade Satisfactory (2) to nearly half (49.26%) of the responses, in stark contrast to Gpt3.5 at the same temperature, which assigned it to just a tenth (10.02%) of the responses. Regarding the highest grade, Excellent (5), Gpt3.5 at temperature 0.0 awarded this grade to almost a fifth of responses (20.04%), while Claude3 awarded no Excellent grades (0.0%) at either temperature setting. Mistral-Large's performance fell between these extremes, demonstrating a more moderate grading pattern compared to the other LLMs (TABLE 2).

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

Selecting one or another LLMs for evaluating students' responses can lead to significant discrepancies in grading outcomes. There is a need for careful consideration in selecting both the LLM and the temperature settings to ensure fairness and accuracy in grading student responses.

Claude3 grading tended to be homogeneous (TABLE 2). Between 48.33% (at temperature 0.0) and 43.33% (at temperature 0.5) of its evaluations categorized as the lower grade of Satisfactory (2). After this, the model exhibited a further tendency towards assigning grades in the center, with 18.15% (at temperature 0.0) to 22.59% (at temperature 0.5) of the evaluations falling into the Good (3) category. Claude3 generally assigned lower-middle-range grades, and notably, it never assigned the grade Excellent (5).

Gpt3.5 displayed a distinct grading pattern, significantly different from the other models (TABLE 2). On the one hand, a considerable proportion of Gpt3.5's evaluations—ranging from 34.59% (at temperature 0.0) to 36.86% (at temperature 0.5)—fell into the very low category of Passable (1). On the other hand, its grading spanned all grades from Fail (0) to Excellent (5), with a substantial distribution across each category.

Gpt4 consistently assigned from lower to mid-range grades, primarily between Satisfactory (2) and Good (3), with almost half of its grades being Satisfactory (2)—49.26% (at temperature 0.0) to 46.85% (at temperature 0.5) (TABLE 2). This indicates a grading pattern that was somewhat similar to that of Claude3.

Mistral-Large demonstrated a broader distribution of grades across the spectrum (TABLE 2). It particularly assigned lower grades of Satisfactory (2) and Passable (1). Specifically, 34.07% and 33.70% of its grades fell into the Satisfactory (2) category across temperatures 0.0 and 0.5.

4.2 Comparison of LLMs Along Their Scoring Criteria

Our research assesses the grading performance of various LLMs in comparison to one another. The potential influence of prompt engineering, which could affect model performance, was controlled in our study through strict adherence to advanced prompting techniques and consistent research standards. All models received identical prompts, reference materials, and student responses. The uniformity of prompt instructions and the context provided to the LLMs suggest that the observed variation in model performance is likely due to inherent characteristics and settings of each model rather than the prompts themselves, this also explains specific cases like Gpt3.5 not respecting the output format.

Setting fully objective and transparent grading criteria is inherently complex. Ultimately, determining the correct and fair grade for each response requires judgment, which can be subjective. Comparing the grading performance of large language models (LLMs) with a single human evaluator does not necessarily resolve this issue, as human evaluators are also prone to errors or biases, potentially assigning grades that are either too lenient or overly harsh. This comparison between human and LLM-based evaluations had been addressed in greater detail in another article [12]. Additionally, the behavior of LLMs in evaluation can vary significantly depending on whether they are applied to high-resource or low-resource languages, further complicating their consistency in grading. Overall, our earlier study indicated that up to 75% of more than 1,000 open-ended responses evaluated with LLM, in that case ChatGPT-4o, matched exactly with that of expert human evaluators or remained within acceptable small deviation from it [20].

In this article, we implemented a process in which all four LLMs evaluated and graded each response. In this process, each model evaluated each responses 10 times at temperature 0.0 and 10 times at temperature 0.5. We selected then as the reference standard grade the most frequently occurring grade (mode value) for each response.

With this “accurate” reference grade established as a benchmark, we analyzed the performance of each LLM against this standard using three key metrics: Accurate (where the LLM-generated grade perfectly matches the benchmark grade that was the most commonly given grade by all LLMs studied), Small Deviation (grades given by LLMs that were within ±1 of the benchmark grade), and Inaccurate (grades given by LLMs that were more distant than 1 grade from the benchmark grade). Grades deemed Accurate or within a Small Deviation range can be considered acceptable, reflecting variability that might occur also among human evaluators.

We found that all LLMs did not always find whether one grade or the grade next to it should be the correct one. These were the cases in which the most commonly suggested grade got at least 40%

of all grades suggested by all LLMs and the second most common grade got at least 30% of all grades suggested by all LLMs. The share of such "undecisive" grades was rather low being 9.3% at temperature 0.0 (five cases). At temperature 0.5, they were 18.5% of cases (ten cases).

The primary aim of our analysis is to identify LLMs that not only maximize accuracy their grading but also minimize the incidence of significant grading discrepancies, thereby enhancing the reliability and fairness of automated grading systems with LLMs (TABLE 3). Overall, of the studied LLMs, the share of grades considered inaccurate, i.e. deviating more than 1 grade from the benchmark grade at 0.0 temperature was 12.96% for Claude-3, 15.92% for Mistral-Large, 10.75% for Gpt4, and 39.88% for Gpt3.5.

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

Claude3 at temperature 0.0 demonstrated a good alignment with the LLM benchmark grade with 62.78% of its grades falling into the Accurate category and a very high 87.04% share was within Small Deviation (±1 of the benchmark). However, at temperature 0.5, Claude3's performance became more distant from the benchmark grade with the share of accurate grades lowering to 48.70%, though those within ±1 of the benchmark was at 88.52%. Overall, Claude3's grading performance at temperature 0.0 can be considered good as barely 12.96% of evaluations fell more than one point out of the benchmark grade (Table 3).

Gpt4 at temperature 0.0 had the second highest performance with 58.15% of its grades falling into the Accurate category and of its grades, 89.27% were within Small Deviation (±1 of the benchmark). At temperature 0.5, it became slightly more aligned toward the benchmark grade. The share of accurate grades increased to 59.07%, and slightly decreased within ±1 of the benchmark to 89.07%. Overall, Gpt4's grading performance can be considered particular: it managed to grade slightly less often fully accurate grades, especially at 0.0 temperature level, but almost nine out of ten of its grades were within small deviation from the benchmark grade (TABLE 3).

Gpt3.5 showed substantially weaker grading accuracy compared to other LLMs studied in this article. At temperature 0.5, less than a third (29.68%) of its grades fell into the Accurate category and within the category Small Deviation (±1 of the benchmark) were 64.65% of its grades. The performance of Gpt3.5 at temperature 0.0 did not improve, with only a quarter (24.95%) of its grades falling into the Accurate category and of its grades, within small deviation (±1 of the benchmark) were only 59.62%. Comparing the grading by Gpt3.5 to other LLMs studied, its performance was substantially inferior. Additionally, 13.04% evaluations were "wildly" different, deviating from the benchmark grade by more than 2 points, which at a 6-point grading scale is significant. In this test, Gpt3.5 was not found to be a reliable tool for grading student's open-ended responses (TABLE 3).

Mistral-Large (at a temperature of 0.0) performed well, achieving an accuracy of 56.67% of its grades being the same as the benchmark grade. Of its grades, 84.08% were within ±1 of the benchmark (Small Deviation). However, at temperature 0.5, Mistral-Large's performance became less accurate with the share of accurate grades dropping to 43.15% and those within ±1 of the benchmark at 81.11%. In both scenarios the share of 'Inaccurate cases' and 'Inaccurate plus bigger difference' increased when a temperature of 0.5 was assigned to 15.92% at temperature 0.0 and 18.89% at temperature 0.5 (TABLE 3).

4.3 Consistency of LLMs on Evaluating Student Open-ended Responses

Consistency is a critical attribute in the evaluation performance of LLMs for several reasons. Primarily, a high variance in the evaluation results of LLMs, such as the grades assigned, undermines the models' reliability for being consistent in their grading of students' open-ended responses and assigning final grades to their performance. To ascertain a model's consistency, it is necessary to run multiple evaluations of the same answer.

In this study, we opted for a 10-shot scenario, evaluating each answer 10 times (FIGURE 2). Our approach involved analyzing whether the grades assigned by each LLM were consistent across all 10 shots, both for the final grade and for each parameter used in the evaluation. Consistency was defined as having identical grades within a 10-shot series for a particular student response analyzed at both 0.0 and 0.5 temperature settings. If all grades within a series were the same then, the LLM demonstrated full consistency for that answer. Conversely, any variation within the series indicated a lack of full consistency in grading by the LLM. This method allowed us to systematically determine the reliability of each LLM in maintaining grading standards across multiple evaluations.

Mistral-Large at 0.0 temperature demonstrated the highest grading consistency among the models, with 83.33% (45 out of 54) of its 10-shot gradings showing no variation within the evaluations. Claude3 showed considerable consistency as well, with 70.37% (38 out of 54) of its 10-shot gradings displaying no internal variation. Gpt4 and Gpt3.5 had lower consistency rates, with 35.19% (19 out of 54) and 18.52% (10 out of 54) respectively, showing uniformity in the grades assigned.

When the temperature setting was increased to 0.5, the consistency of the LLMs noticeably decreased, highlighting a sensitivity to temperature changes. At this higher temperature, Gpt4 achieved the highest level of consistency, albeit reduced to 20.37% (11 out of 54). Mistral-Large's consistency significantly declined to 14.81% (8 out of 54). Both Claude3 and Gpt3.5 demonstrated low

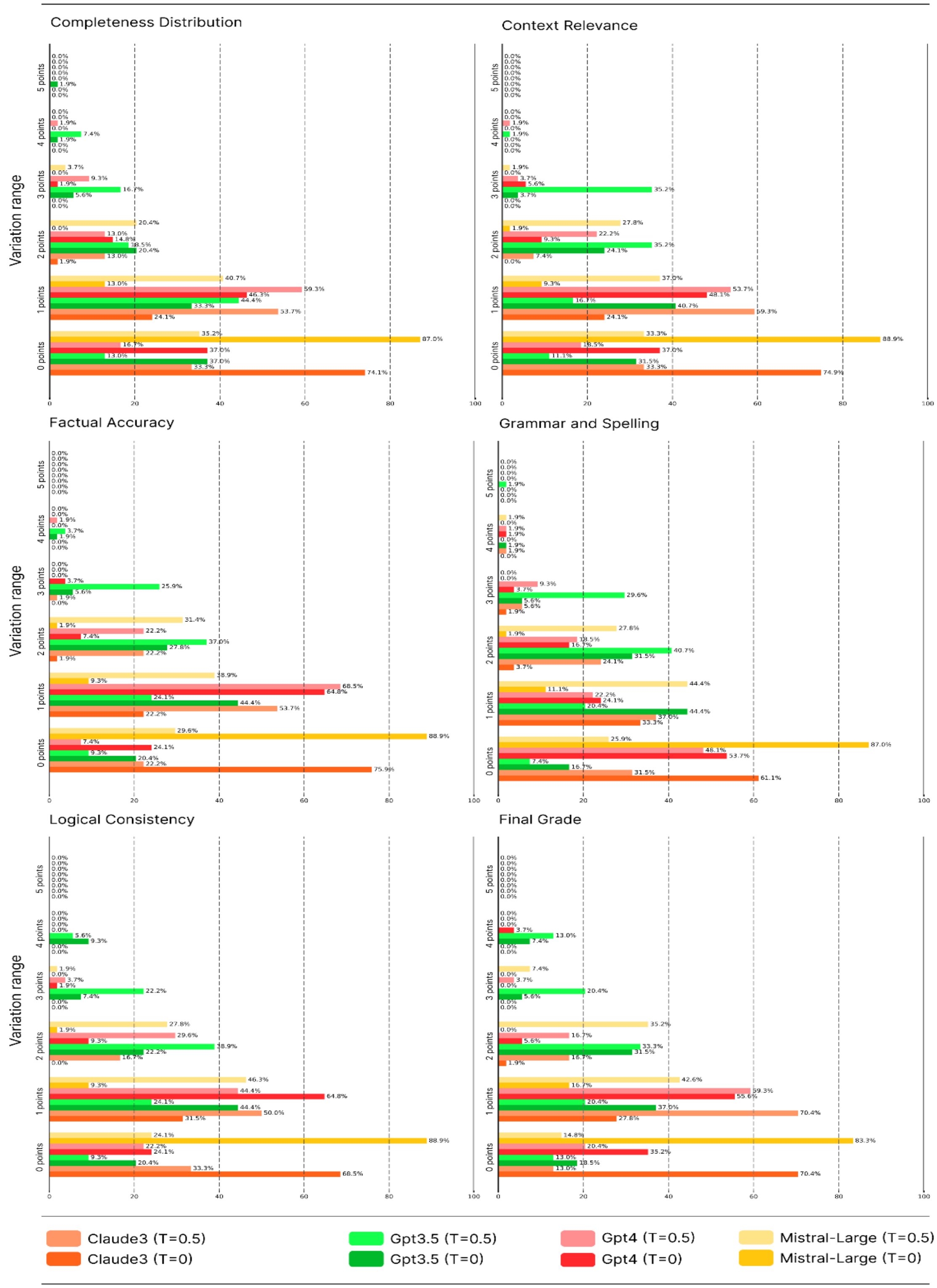

Figure 2: Share of cases without variation within their 10 shot-grading regarding the final grade and evaluation parameters (%, number of cases, in total 54 evaluation groups.

consistency, with 12.96% (7 out of 54) and 14.81 (8 out of 54) of their gradings being internally consistent, indicating substantial internal variation in more than 85% of their evaluations.

Looking at the evaluation of various parameters that assess different aspects of student responses—including context relevance, factual accuracy, completeness, logical consistency, and grammar and spelling—the patterns of consistency largely mirrored those observed in the final grade assessments (FIGURE 2). At 0.0 temperature, Mistral-Large exhibited very high consistency across all parameters, with rates ranging from 87.04% to 88.89%. Claude3 at 0.0 temperature was the second most consistent, with rates between 68.52% and 75.93%. In contrast, at 0.5 temperature, Gpt3.5 demonstrated notably lower consistency, ranging from 7.41% to 12.96% across the parameters. Gpt4 at 0.5 temperature showed higher variability in consistency rates, from 7.41% to 48.15%, though it performed notably better in the specific evaluation of grammar and spelling, reaching a consistency rate of 48.15%, although still lower compared to Claude3 and Mistral-Large.

In terms of grading variability at 0.0 temperature, Mistral-Large and Claude3 displayed high consistency between 98% and 100% of consistency of its gradings within one grade point. Gpt4's consistency in final grades was also high when including one grade deviations reaching 90.75% consistency, yet it included notable outliers where the grade assignation was up to 4 points difference. Gpt3.5 demonstrated significant inconsistency with only almost half (46.3%) of its final grade evaluations within a one-point deviation (TABLE 4).

Table 4: Evaluation cases without any variation within their 10-shot evaluation (%).

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

With temperature 0.5, grading variability increased across all models. In final grades, Claude3 maintained the highest consistency with 83.33% of its gradings falling within one grade point, while Gpt3.5 exhibited significant inconsistency with only 33.33% of its gradings within the same range, pointing to substantial fluctuations in its grading of student responses across all parameters (TABLE 4).

4.4 Differences in Processing Time in Evaluation Tasks among Different LLMs

Processing time is an essential consideration when deploying LLMs in educational settings, particularly for the efficient and timely evaluation of large volumes of student assessments. In our analysis of LLMs' ability to grade open-ended responses, we define processing time as the duration each model requires to analyze context information and evaluate student responses thoroughly. This is crucial for minimizing errors such as hallucinations and ensuring well-informed grading decisions.

To maintain the integrity of our data, we excluded substantial outliers from our analysis. In these outliers represented cases of which evaluation and grading that required exceptionally long processing times—more than 10 times the average or exceeding 150 seconds, with one instance surpassing 240 seconds. While these cases, very few in total, were not considered in the primary analysis, their potential impact on computational costs could be significant, especially when considering the large-scale deployment of LLMs in academic assessments.

Our study found considerable differences in processing times both between different LLMs and between temperature settings within the same model. We observed that increasing the temperature setting from 0.0 to 0.5 typically resulted in a 4% to 12% increase in processing time for depending on the model. However, the correlation between processing times and the grade performance assigned by LLMs was relatively low, indicating that quicker processing does not necessarily translate to more accurate or consistent grading (TABLE 5).

Table 5: Processing time of LLMs for evaluating student responses by one shot after another.

	mean	std	min	25%	50%	75%	max	Pearson	Spearman
temperature 0.0									
Claude	20.61	3.82	12.70	18.33	20.12	22.12	66.25	0.028	0.034
Gpt3.5	3.89	4.79	1.50	2.94	3.51	4.12	112.20	−0.041	−0.006
Gpt4	18.51	9.28	7.80	13.54	16.78	21.46	120.09	0.009	0.096
Mistral-L	11.17	2.91	5.92	9.17	10.64	12.40	27.57	0.177	0.145
temperature 0.5									
Claude3	21.55	4.09	14.44	18.81	20.89	23.54	48.55	−0.007	−0.027
Gpt3.5	3.52	3.38	1.42	2.68	3.19	3.74	78.38	0.093	0.070
Gpt4	20.72	6.74	9.25	15.84	19.74	24.21	64.30	0.132	0.135
Mistral-L	11.95	3.05	6.16	9.87	11.36	13.80	34.08	0.247	0.255

The fastest average processing time per answer was recorded at 3.52 seconds by Gpt3.5 at temperature 0.5, followed closely by the same model at temperature 0.0 (3.89 seconds). In stark contrast, the slowest average processing times were significantly longer, with Claude3 at temperature 0.5 taking 21.55 seconds and Gpt4 at temperature 0.5 taking 20.72 seconds. This discrepancy implies that computational costs and associated energy consumption can vary more five times between the fastest and slowest models (TABLE 5).

Evaluating 100 student responses with a single shot at time from the fastest model would take approximately 6.5 minutes, and using a 10-shot scenario would extend this time to about 1 hour and 5 minutes. Conversely, using the slowest model for a one-shot evaluation of the same number of responses would take nearly 36 minutes, and a 10-shot scenario would require almost 6 hours, if shots would be analyzed one after another. However, there are techniques to analyze several shots and responses–even up to 20 responses–simultaneously, so these total processing times are only examples of extremely rare circumstances.

Additionally, preparing the evaluation process, including setting up learning materials, evaluation guidelines, and scoring scales, typically takes about 10–15 minutes for one exam. In this case, the total time needed to evaluate 54 responses of this test would have been 25–30 minutes but, besides the general grade assigned, it could have included also detailed grading of each response along several parameters and personalized written feedback and improvement suggestions for each student on their response.

The time required by LLMs for evaluation is much less that a human could do. An experienced and trained human evaluator, such as a teacher with expertise in the topic, could have assessed a single response of this test and assign a general grade for it within 30–60 seconds. However, as mentioned, in addition to assigning an overall grade, LLM can assess within the same time frame various detailed aspects of each response, such as coherence, consistency, comprehensiveness, and grammar, etc., and also generate personalized feedback, if prompted so ([12]). For a human evaluator, this would take minutes for each response. Moreover, one human evaluator cannot maintain consistent and uninterrupted grading performance over extended periods without fatigue, which raises the risk of inconsistencies. This limitation does not affect LLMs, which can maintain consistent performance regardless of the volume or duration of evaluations.

The variability in processing times was the largest in Gpt4 (both temperatures) and the smallest in Mistral-Large (both temperatures), as illustrated by the standard deviations. This signifies the necessity of a balanced approach in selecting LLMs for educational use. While faster processing times can improve the scalability of LLM applications, they must not compromise the accuracy and consistency of the outcomes. Therefore, educational institutions need to weigh both the efficiency and reliability of LLMs when integrating these technologies into their grading systems.

5. CONCLUSIONS

This study demonstrated that LLMs like Gpt3.5, Gpt4, Claude3, and Mistral-Large can effectively evaluate and grade students' open-ended responses, especially with the Retrieval Augmented Generation (RAG) framework. This framework enables efficient, ethical, and secure grading by using tools like the LangChain OpenAI Embedding method to process text into numerical representations.

Grading with LLMs requires significant computational resources, energy, and tokens, emphasizing the importance of model consistency. For reliable grading, each response should be evaluated multiple times, with 10 evaluations (shots) recommended for consistency, though future improvements might reduce this to 5 shots. Evaluating LLM performance requires a reliable benchmark grade, established through either consensus from multiple human experts or agreement among high-performing LLMs, favoring the mode grade from multiple evaluations. Temperature settings significantly impact grading performance, with a setting of 0.0 providing more consistent results.

Variability in grading outcomes reflects differences in LLM training, architecture, and hyperparameters, highlighting the need for careful selection of LLMs and temperature settings (preferably 0.0) for consistent results. Performance comparisons should precede the implementation of any LLM in grading to ensure reliability and accuracy. Speed is also one aspect. Gpt3.5, while the fastest, displayed significant grading inconsistencies, with a broader range of grades and a higher proportion of inaccuracies. Due to these limitations, Gpt3.5 is not recommended for systematic evaluation of students' open-ended responses.

In this test, the benchmark grade was defined as the grade most frequently selected by the four studied LLMs during their evaluations. Grades matching the benchmark were considered accurate, while those deviating by one grade on a six-point scale were deemed acceptable. Grades differing by two or more levels were classified as inaccurate. Temperature settings significantly influenced grading consistency, though the effect varied across models. At a 0.0 temperature, the share of inaccurate grades was 12.96% for Claude3, 15.92% for Mistral-Large, 10.75% for Gpt4, and 39.88% for Gpt3.5. Claude3 demonstrated the highest accuracy, with 62.78% of its grades matching the benchmark, followed by Gpt4 (58.15%), Mistral-Large (56.67%), and Gpt3.5 (24.44%).

Assessing the strengths and limitations of LLMs in evaluation is essential, along with comparative analyses to identify models best suited to specific educational needs. Criticisms from scholars and educators on LLMs in educational evaluation have primarily targeted the inconsistent and insecure evaluation performance of LLMs, often focusing on Gpt3.5. However, so far only limited amount of studies have systematically analyzed better-performing models and they rarely have incorporated the use of frameworks like RAG.

This study highlights the substantial potential of LLMs in educational evaluation. While LLMs are not flawless evaluators, they can significantly enhance the evaluation process, offering also individualized feedback for students that will help the time consuming tasks of educators. Comparative analyses are essential to identify the most reliable and efficient LLMs to ensure consistency in automated grading. Additionally, refining LLM grading through precise prompting, calibrating results, and adjusting algorithms can improve outcomes.

Future research should focus on exploring newer LLM versions, alternative prompting strategies, task-specific training, and the application of these tools across diverse cultural and linguistic contexts. A comprehensive approach will promote responsible and effective use of LLMs in education, improving learning outcomes, supporting teachers, and ensuring fairness and accuracy in evaluating students' exam responses, essays and other written performances.

References

[1] Baidoo-Anu D, Ansah O, L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. J AI. 2023;7:52-62.

[2] Dai Y, Liu A, Lim CP. Reconceptualizing ChatGPT and Generative AI as a Student-Driven Innovation in Higher Education. Procedia CIRP. 2023;119:84-90.

[3] https://www.econstor.eu/handle/10419/270970

[4] Lo CK. What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. Educ Sci. 2023;13:410.

[5] Adiguzel T, Kaya MH, Cansu FK. Revolutionizing Education With AI: Exploring the Transformative Potential of ChatGPT. Contemp Educ Technol. 2023;15:429.

[6] Bahroun Z, Anane C, Ahmed V, Zacca A. Transforming Education: A Comprehensive Review of Generative Artificial Intelligence in Educational Settings Through Bibliometric and Content Analysis. Sustainability. 2023;15:12983.

[7] Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z et al. Is ChatGPT a Good Nlg Evaluator? A Preliminary Study. 2023. Arxiv preprint https://arxiv.org/pdf/2303.04048

[8] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2015. ArXiv: https://arxiv.org/pdf/1409.0473

[9] Vaswani A, Shazeer N, Parmar N, Uskoreit J, Jones L, et al. Attention Is All You Need. 2017. Arxiv preprint https://arxiv.org/pdf/1706.03762v7

[10] https://cohere.com/blog/llm-parameters-best-outputs-language-ai

[11] Jauhiainen JS, Garagorry Guerra A. Generative AI and ChatGPT in School Children's Education: Evidence From a School Lesson. Sustainability. 2023;15:14025.

[12] Jauhiainen JS, Garagorry Guerra A. Generative AI in Education: ChatGPT-4 in Evaluating Students' Written Responses. Innov Educ Teach Int. 2024:1-18.

[13] Milsakaki E, Kukich K. Evaluation of Text Coherence for Electronic Essay Scoring Systems. Nat Lang Eng. 2002;10:25-55.

[14] Beseiso M, Alzharni S. An Empirical Analysis of Bert Embedding for Automated Essay Scoring. Int J Adv Comput Sci Appl. 2020;11.

[15] Xia Q, Weng X, Ouyang F, Lin TJ, Chiu TK. A Scoping Review on How Generative Artificial Intelligence Transforms Assessment in Higher Education. Int J Educ Technol High Educ. 2024;21:40.

[16] Vázquez-Cano E, Ramírez-Hurtado J, Sáez-López J, López-Meneses E. ChatGPT: The Brightest Student in the Class. Think Skills Create. 2023;49.

[17] Bewersdorff A, Seßler K, Baur A, Kasneci E, Nerdel C. Assessing Student Errors in Experimentation Using Artificial Intelligence and Large Language Models: A Comparative Study With Human Raters. Comput. 2023;5.

[18] Lee GG, Latif E, Wu X, Liu N, Zhai X. Applying Large Language Models and Chain-Of-Thought for Automatic Scoring. Comput. 2024;6.

[19] Wei J, Want K, Schuurmans D, Bosma M, Xia F, et al. Chain of Thought Prompting Elicits Reasoning in Large Language Models. 2022. Arxiv preprint https://arxiv.org/pdf/2201.11903v6

[20] Jauhiainen J, Garagorry Guerra A. Evaluating Open-Ended Responses of National Matriculation Exam With ChatGPT. 2024;40.

Appendix

Examples of prompts for guiding LLMs in the evaluation process

“““Your role is of a University Professor, responsible for evaluating students.

/1. Evaluate the student’s response based on the provided documents.
###Reference material: {context_documents}.
—————————————-
###Evaluation Guideline: {evaluation guidelines =>
Evaluate the student’s response based on the provided context information.
Evaluate the student’s response based on these criteria: Context Relevance, Factual Accuracy, Completeness, Logical Consistency, Grammar and Spelling, Grade Answer.
Context Relevance: Is the student’s response relevant to the context provided in the reference material?
Factual Accuracy: Is the student’s response factually accurate?
Completeness: Is the student’s response complete?
Logical Consistency: Is the student’s response logically consistent?
Grammar and Spelling: Does the student’s response have any grammar or spelling errors?
Grade Answer: Grade the student’s response based on the criteria above.}
—————————————-
###Expected Knowledge from the student: {knowledge_level}
—————————————-
###Question: {student_question}
###Answer: {student_answer}
—————————————-
/2. How well does the student answer the question?
###Output format: {output_format
Question: Write the question being answered
Answer: Write the student answer to the question
Student Answer Feedback: Write Feedback
Context Relevance:Fail | Passable | Satisfactory | Good | Very Good | Excellent
Factual Accuracy:Fail | Passable | Satisfactory | Good | Very Good | Excellent
Completeness: 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}”””