Front. Digit. Educ., 2026, 3(2): 17
https://doi.org/10.1007/s44366-026-0091-1

RESEARCH ARTICLE

Evaluating Open-Ended High-Stakes Examinations with LLMs: Alignment Between ChatGPT-4o and Human Grading in High- and Low-Resource Languages

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

[a] Department of Geography and Geology, University of Turku, Turku 20014, Finland

[b] Institute of Ecology and the Earth Sciences, University of Tartu, Tartu 51003, Estonia

Abstract Large language models (LLMs) are increasingly used to grade written responses, yet large-scale benchmarks against human expert evaluation remain scarce, especially across languages with differing resource levels. This study evaluates ChatGPT-4o using a reranked retrieval-augmented generation framework to grade Finland's national high-stakes matriculation examination based on 1,016 students' open-ended responses. We examined GPT-4o's agreement with official grades, its recognition of grading-relevant keywords, and the effect of translated responses from a low-resource language (Finnish) into a high-resource language (HRL) (English). Using descriptive statistics and correlation analyses, the results show that GPT-4o's grades on a 0–15 scale closely matched human expert evaluations; 75.00% of scores were within ±2 points of official grades, with only 3.00% being severe outliers. The translated responses into English improved this alignment to 85.00%. While the model generally identified relevant keywords effectively, occasional misinterpretations of contextual usage reduced grading reliability in a few cases. Overall, the findings demonstrate both the promising and current limitations of LLM-based assessment. There is a significant potential to use LLMs as supplementary grading tools, particularly in HRLs, but they do not yet match the consistency or interpretative depth of human expert evaluators. The study illustrates the need for human oversight, rigorous validation, and careful consideration of language effects when deploying LLMs in high-stakes educational assessments.

Keywords ChatGPT, evaluation, large language models (LLMs), grading, retrieval-augmented generation (RAG)

Received November 27, 2024; revised December 15, 2025; accepted February 13, 2026

Jussi S. Jauhiainen (✉)
E-mail: jusaja@utu.fi

1 Introduction

Grading student essays and open-ended examination responses has long posed a challenge for education systems, as the process is time-consuming, burdens teachers and institutions, may vary across evaluators, and is prone to bias. In the mid-2020s, generative AI (GenAI) and large language models (LLMs) emerged as potential tools to address these challenges, increasing attention for their ability to streamline and scale assessments. The prospect that LLMs, such as GPT, could revolutionize educational assessment has generated much interest (Bewersdorff et al., 2023; Jauhiainen & Garagorry Guerra, 2024; Jauhiainen & Garagorry Guerra, 2025b; Jauhiainen et al., 2026; Jukiewicz, 2024; Su & Yang, 2023). However, automatic essay scoring (AES), particularly in high-stakes examinations, remains one of the most challenging problems in natural language processing (Beseiso & Alzahrani, 2020).

LLMs promise speed, consistency, and impartiality. They offer a relief from heavy grading workloads while improving the efficiency of formative and summative assessments (Bui & Barrot, 2025; Fütterer et al., 2023; Hackl et al., 2023; Henkel et al., 2024; Jukiewicz, 2024; Mao et al., 2024; Pinto et al., 2023). LLMs demonstrate accuracy, fairness, and adaptability comparable to, or exceeding, that of human expert evaluators. In addition to the automatic grading, LLMs ensure reliability across large cohorts of students without amplifying inequities or biases. However, challenges remain. LLMs continue to struggle with complex reasoning tasks and require careful oversights (Bewersdorff et al., 2023). Teachers and students also have misconceptions and concerns about AI (Bewersdorff et al., 2023). Moreover, most existing studies rely on small-scale experiments (Bui & Barrot,

2025; Mao et al., 2024), which indicates the need for large-scale evaluative assessments with LLMs covering large cohorts of students to assess their true capacity in high-stakes educational contexts.

Real-world LLMs applications in educational assessments for non-English, low-resource languages (LRLs) remain insufficiently documented and analyzed. A primary constraint concerns data. Foundational models, such as GPT, are trained predominantly on high-resource languages (HRLs), which limits performance and validity when evaluating work produced in other languages or based on materials in other languages. These models underrepresent perspectives from LRL communities, as they have not been trained on such data. Ethical and practical issues further complicate the adoption. Data privacy and governance are unresolved, including the risk that assessment data could be repurposed to train models without consent. Baseline mistrust persists due to hallucinations, as incorrect outputs can undermine reliable scoring and feedback, particularly when hallucinations cannot be systematically recognized and eliminated during the evaluation process (Huang et al., 2023; Jauhiainen & Garagorry Guerra, 2025b; Nazaretsky et al., 2022; Ouyang et al., 2024). Because LLMs are trained largely on Internet texts, they may reproduce social biases (such as racism, sexism, and ethnocentrism). Substantial energy demands raise sustainability-related questions and motivate calls for "greener" LLMs (Stojkovic et al., 2024). Digital divides, such as unequal access to devices, internet connectivity, and institutional capacity, limit equitable uptake and risk, widening educational inequalities (Lee & Cha, 2024; Lee et al., 2024). While these challenges are significant, LLM technology and its safeguards are evolving rapidly, and mitigation strategies are improving over time. Meanwhile, robust evidence for LRL assessment contexts is still emerging.

One example of a high-stakes examination is a national matriculation examination that differentiates higher-performance students from their peers. Such examinations assess the mastery of curriculum across an entire cohort of students, and the results greatly influence students' academic and career trajectories. However, the task of assessing open-ended responses from hundreds of thousands of upper secondary-school students annually is highly labor-intensive and time-consuming. Stringent quality-control measures are required to ensure consistent grading standards among a large number of human expert evaluators. Grading fatigue and the influence of personal biases, preferences, and preconceived notions can sometimes affect human scoring results (Barrot, 2024; Palermo, 2022). As a response to these challenges and due to a shortage of teaching resources, some countries (such as China, the United States, and India), are opting for multiple-choice tests, which are simpler to score automatically. The disadvantage is that these tests fail to assess the wide spectrum of skills that essay writing can demonstrate (Ramesh & Sanambudi, 2022).

This study investigates the use of GPT-4o in chat-version for high-stakes educational assessment by evaluating its ability to grade large volumes of open-ended student responses. GPT-4o was selected as one of the most advanced and widely deployed LLMs at the time of the study. We examine how closely GPT-4o's scoring agrees with human expert evaluation, how consistent the model is as a grader, and whether performance differs between HRLs and LRLs. While traditional grading of such responses requires weeks of human expert labor and substantial costs, GPT-4o can complete the same task within a single day. This raises the question of whether LLM-based assessment can reliably supplement or partially even replace human expert evaluators.

The case study draws on Finland's high-stakes national matriculation examination, a compulsory high-stakes examination taken after 12 years of schooling. We analyze responses to one geography examination question, requiring students to describe 3 types of rainfall, their geographic distribution, and rainfall formation processes. The task involves concise, essay-style answers (typically 100–200 words) that assess students' understanding rather than advanced expertise. Such tasks are common in the social sciences and humanities disciplines, as well as in other disciplines. The dataset comprises all 1,016 Finnish responses submitted in the autumn 2023 examination session, providing a complete national cohort and a challenging test case because Finnish is an LRL for LLMs.

The study addresses 3 research questions: (1) How do GPT-4o's grades compare with those produced by a systematic human evaluation process? (2) How effectively does the model identify grading-relevant keywords in students' responses? (3) Does translating responses from LRL (Finnish) to HRL (English) improve LLM-based assessment outcomes?

Methodologically, the study discusses and builds on prior work in automated scoring and introduces 2 reliability enhancements in the LLM-based context: a retrieval-augmented generation (RAG) framework and a 5-shot evaluation. The results show that GPT-4o closely approximates human expert evaluation. An important point is that, translating responses from LRL (Finnish) to HRL (English) improved agreement, with the percentage of alignment with official scores rising from 75.00% to 85.00% within ±2 points on a 16-point scale. GPT-4o and human expert evaluators also largely agreed on key content elements, though the model occasionally exhibited narrower keyword coverage and minor misinterpretations. The study concludes by outlining the feasibility and limits of LLM-based assessment in high-stakes contexts as well

as directions for future research on scalable, multilingual educational assessments.

2 Automated Educational Evaluation and the Emergence of GenAI

The LLM-based assessment of students' responses and essays has become a widely discussed and practiced topic in the mid-2020s. However, the automation of students' performance evaluation and feedback has been a focus in educational technology for decades. The need to streamline the time-consuming tasks of grading students' written responses generated early innovations in AES, as noted in the related overviews and discussions (Barrot, 2024; Beseiso & Alzaharni, 2020).

2.1 | Early Development of AES in Educational Assessment

The origins of AES can be traced back to the 1960s, when Professor Ellis Batten Page pioneered the use of technology to evaluate essays based on the quantifiable text features, such as word count and sentence length, through systems such as Project Essay Grade. These early systems used simple linear regression to mimic human expert scoring, but were limited to superficial textual analysis due to the computational constraints of the time (Page, 2003). The 1990s marked a significant evolution in AES with the introduction of Intelligent Essay Assessor (IEA) by Professor Peter Foltz's research team. IEA moved past surface analysis by employing latent semantic analysis to assess the semantic content of essays, thus enhancing the evaluation of content relevance and coherence (Foltz et al., 1999; Landauer et al., 1998).

Subsequent developments integrated machine learning technologies, such as bidirectional encoder representation from transformers, which considers the context of words within text and allows for more refined analyses of essay content, structure, and semantics. Nevertheless, Singla et al. (2023) argued that they behaved like bag-of-words models. Biases in training datasets could skew these models' sensitivity to certain widely used words, which impacts their grading accuracy (Beseiso & Alzaharni, 2020; Ramesh & Sanambudi, 2022). Further refinements in AES employed models such as random forest, which analyzed complex features, such as word choice and sentence complexity, to provide detailed evaluations. This approach improved grading accuracy and supported educators by streamlining the evaluation process, thus freeing up time for personalized teaching (Lee et al., 2024). However, challenges, such as domain specificity, indicates the need for careful training data selection to avoid biases that could affect grading fairness (Leech et al., 2024).

2.2 | LLMs in Educational Evaluation and Related Challenges

The latest advancements in educational evaluation technologies for AES employ LLMs, which uses deep learning to process and generate texts. These models, trained on vast datasets, excel in evaluating open-ended responses holistically, considering the context, tone, and intent of students' written responses. LLMs adapt to various writing styles and subjects for tailored suggestions (Mao et al., 2024). This flexibility, due to their access to extensive and diverse training materials, is a significant advancement over older systems, such as AES and IEA. These capabilities allow LLMs to be highly adaptable across different educational settings assisting tailored suggestions in evaluating students' open-ended responses. More advanced LLMs can handle different evaluation systems and guidelines without extra training, which makes LLMs promising educational technology tools for evaluation (Jukiewicz, 2024; Pinto et al., 2023).

On the one hand, LLMs represent a continuity of long-standing educational technology goals, but they bring new opportunities for enhancing educational processes with deeper semantic understanding and richer feedback. On the other hand, LLMs raise old concerns, such as bias, fairness, and validity, which must be addressed to responsibly integrate LLMs into high-stakes educational settings. LLMs often reflect biases in the data training process, favoring dominant languages, cultural norms, and standard writing styles, which disadvantages students from diverse linguistic or socioeconomic backgrounds. The performance of LLMs varies across languages, which potentially reinforces educational inequalities, and they may privilege fluency or keyword presence over genuine conceptual understanding. Moreover, limited transparency and occasional misinterpretation of context undermine the validity and contestability, which makes fully automated LLM-based assessment still challenging in high-stakes educational assessments (Bewersdorff et al., 2023; Hackl et al., 2023; Henkel et al., 2024; Jukiewicz, 2024; Zhai, 2025).

Carefully designed prompts are important for successful LLM-based assessments in education. Research indicates that the effectiveness of LLMs depends on domain-specific knowledge, the type of scoring systems used, and factors, such as model parameters and prompt clarity (Bewersdorff et al., 2023; Hackl et al., 2023; Henkel et al., 2024; Mao et al., 2024; Jukiewicz, 2024; Scarlatos et al., 2024; Zhai, 2025). Detailed and well-structured prompts generally yield better performance by guiding LLMs through the

complexities of the tasks set to perform (Avnat et al., 2024; Cain, 2024; Hackl et al., 2023). When LLMs are applied to subject-specific areas, domain-specific prompts can further improve grading accuracy (Hu & Zhou, 2024; Jauhiainen & Garagorry Guerra, 2024). However, prompt variation can generate divergent evaluation results (Mao et al., 2024).

While LLMs offer significant potential for AES, they face the following 3 unresolved issues: (1) the opacity of LLMs' decision-making processes; (2) teachers' inadequate skills to use LLMs; and (3) challenges related to the reproducibility of research on using and evaluating LLMs in educational assessments. These all make it difficult to validate benchmarks (Leech et al., 2024).

Regarding LLMs' transparency and opacity in educational assessments, ensuring transparency is vital, as it builds trust among educators and facilitates adoption (Conijn et al., 2023). A lack of transparency regarding how these models function can lead to distrust, which hinders the integration into educational processes (Lee & Cha, 2024; Qin et al., 2020). Nazarevksy et al. (2022) argued that explaining how these systems make decisions and how they can support rather than replace teachers can alleviate concerns and build trust in AI educational technology. This helps clarify the benefits of AI tools and reduces potential resistance to their use. This has led to calls for explainable AI, which seeks to make AI decisions interpretable by clarifying how and why a model produces a particular outcome. It aims to increase transparency, trust, and accountability in AI systems, as well as in educational assessments (Dwivedi et al., 2023). Studies, such as those by Mao et al. (2024), point out that LLMs, despite their advanced capabilities, often show variability in their outputs, especially when they are not properly calibrated with consistent prompts and settings.

Regarding teachers' skills in LLM-based assessments, teachers' characteristics (such as knowledge level) and specific needs for explanations significantly influence the acceptance of AI advice (Su & Yang, 2023). A limited understanding of how LLMs function may slow their adoption in education. Moreover, teachers' skills in operating these models are important in achieving accurate results. In the realm of educational technology, the deployment of LLMs also raises concerns about security, transparency, and the ethical use of AI. However, von Eschenbach (2021) and Ouyang et al. (2024) noted that a limited understanding about these processes can lead to skepticism and algorithm aversion, whereby people distrust and resist algorithmic decisions. Nevertheless, wide access to LLMs encourages many educators to integrate these tools into pedagogical settings.

In terms of the capacity of models, such as ChatGPT, a generative pretrained transformer in a chat version, for educational evaluation, Bui and Barrot (2025) reported scoring discrepancies between ChatGPT and human expert evaluators. These scoring discrepancies stemmed from limitations in ChatGPT's algorithm, which may not fully capture complex writing elements, such as style and creativity. Such challenges can lead to different interpretations and applications of scoring criteria. The model's training data might not comprehensively represent diverse writing qualities, or it could contain biases affecting grading accuracy. In addition, while GPT models can simultaneously identify multiple errors in evaluated texts, their strictness in error detection might result in lower scores compared to human expert evaluators, who might be more lenient with more holistically judge the texts. In their test, Bui and Barrot (2025) concluded that ChatGPT failed to match the rating ability of an experienced human expert evaluators; however, in the test, they utilized GPT-3.5, which is substantially weaker than GPT-4o used in this study, not to mention GPT-5.2 launched in 2025. In contrast, Jukiewicz (2024) found a strong positive correlation between GPT-3.5-Turbo and human expert evaluators which suggested consistency and repeatability in grading. The growing consistent accuracy of LLMs in evaluation tasks also raises questions about variability in human-based assessment (Jauhiainen et al., 2026). Given the rapid pace of LLM development, results from the earlier LLM versions quickly become outdated and need to be interpreted with caution. Model drift further complicates the longitudinal consistency and comparability of evaluation outcomes.

3 Materials and Methods

3.1 | Materials

This study quantitatively examines 1,016 written responses from Finland's national high-stakes matriculation examination in geography. Of 2,515 registered students, 2,372 sat the examination and responded to 9 questions (Matriculation Examination Board, 2024). One question—on the description on 3 rainfall types—was selected for analysis. A total of 1,016 students answered the question in Finnish.

The data, obtained from the National Matriculation Board (research permit number OPH-6154-2023), consisted of students' written responses from the Finnish upper secondary school matriculation examination in geography in autumn 2023. The dataset was fully anonymized, only containing students' written responses and their corresponding grades (0–15), without any information on respondents, schools, or evaluation processes.

The question analyzed asked students to

"Describe all 3 types of precipitation and name one occurrence characteristic for each type." This topic was covered in one chapter of the second textbook for the second geography course of the curriculum, which was not mandatory. Typically, the students who chose to answer this question were strong interested in geography and well prepared for the topic.

Students were assessed by human expert evaluators on their open-ended responses, with the potential to earn 15 points. The official grading allowed for 5 points per rainfall type, allocated as follows: 1 point for correct naming, 1–3 points for formation explanations, and 1 point for accurately identifying a characteristic geographical area. The average response length was 170 words (1,247 characters, without space), with a median of 159 words (1,162 characters, without space). The average grade was 9.2 (median 10). The lowest 20% scored 6 or below, while the highest 20% scored 13 or above.

Finland's high-stakes national matriculation examination in geography involves a multistage evaluation process to ensure accuracy, consistency, and fairness. Geography is an optional component of the compulsory matriculation examination (*Act on the matriculation examination 502/2019*), taken in the final year of upper secondary school at around age of 19. Students interested in geography generally complete 1 to 4 courses, equivalent to 2 to 8 academic credits, which covers the extensive topics in physical and human geography and related methodologies (Finlex, 2019; Finnish National Agency for Education, 2019).

In the pre-evaluation stage, examination questions and preliminary evaluation guidelines are developed collaboratively by geography specialists and nationally recognized experts from various academic fields. These guidelines help standardize the evaluation process.

In the evaluation stage, teachers conduct the preliminary assessment using official guidelines. The final evaluation is carried out anonymously by a team of qualified human expert evaluators, referred to as censors, who may include the question developers and other experienced geography educators. However, current geography teachers are excluded to preserve impartiality. A chairperson of the team is responsible for maintaining quality assurance standards. This multistage process aims to ensure the highest possible quality reachable with human expert evaluators.

In the post-evaluation stage, following the final evaluations, students have an option to request a recheck of their grades if they believe an error has occurred. This request initiates a review by 2 impartial censors who verify and, if necessary, amend the grades. This multistage evaluation process ensures that the grading is equitable and adheres strictly to the established standards. However, it requires a lot of labor, time, and financial resources, and this is becoming less feasible to achieve.

3.2 | Methods to Set GPT-4o to Evaluate Students' Open-Ended Written Responses

We compared GPT-4o's grades with those of humans, analyzed keyword-based grading behavior, and tested the effect of translating responses from LRL (Finnish) to HRL (English). The evaluation of 1,016 students' written responses followed a 5-step pipeline for preparing and executing the evaluation: learning material input, examination question input, evaluation instruction input, response input and verification, and evaluation execution. We also conducted the analysis using a 3-step pipeline: grade comparison analysis, outlier keyword analysis, and language impact analysis.

First, learning materials, examination questions, official grading guidelines, and students' written responses were provided to the model. The written responses were evaluated using GPT-4o via application programming interface, with the parameter temperature set to 0.0 to minimize output variability (Hackl et al., 2023; Jauhiainen & Garagorry Guerra, 2025a). The workflow was implemented in Python using agents, like LangChain and pandas, for data preprocessing, prompt construction, and result aggregation. To improve the consistency and reasoning quality, we employed step-by-step (chain-of-thought) prompting (Cain, 2024; Chen et al., 2023; Lee et al., 2024; Wei et al., 2023; Yao et al., 2023). Prompts specified the examination context, required the exact recall of student responses, and instructed the model to assign scores on a 0–15 scale based solely on the official evaluation guidelines and the designated textbook materials.

Second, the evaluation employed a reranked-RAG approach, an open-book method in which relevant documents are retrieved and re-ranked to support more accurate LLM-based assessments despite input length limitations (Glass et al., 2022; Yu et al., 2024). As learning materials, evaluation guidelines, and students' written responses were often lengthy, the corpus (1.27 million characters, without space) was split into coherent, overlapping chunks—paragraph-sized text pieces sized to capture complete ideas while fitting within the model's input constraints. The light overlap between chunks ensured that key ideas were not cut at their boundaries. Chunks serve a single purpose: to make specific examination-related textbook passages easy to find for LLMs and to require LLMs to justify grades with those specific passages. This chunk allows GPT-4o to efficiently retrieve and ground grading decisions in specific approved textbook passages rather than relying on their training data. This approach improves accuracy, reduces token uses and costs, prevents

hallucinations by limiting irrelevant input, and avoids LLMs overload or forced processing of mixed-content sections that degrade grading quality.

Third, for each student's written response, we compared vector embeddings and evaluation context against all chunk embeddings to retrieve the top 5 chunks with the highest similarity to student's written response. These were organized by sending the most relevant chunks towards the edges of the context material to prevent important knowledge loss. These top chunks, containing the most pertinent textbook passages and rubric elements, were provided to LLMs alongside students' written responses, ensuring that grading was precisely grounded in key source materials and supported by clear evidence-based reasoning.

Fourth, to ensure that the model assessed each student's written response exactly, we implemented a 5-step pipeline. Despite using a deterministic setting (parameter temperature = 0.0), minor variations still occur depending on the model utilized. Multiple passes mitigated random errors or hallucinations from 1-shot evaluations, ensured accurate recall of each response, and substantially improved grading reliability.

Fifth, to ensure the consistency of the evaluation results, the model evaluated each response 5 times. This guaranteed the reliability of the results in this complex educational evaluation context despite the increased computational costs and total processing time. When the 5 times (shot results) converged, we treated the model's evaluation results as stable. In the case of minor discrepancies, we adopted the most frequent evaluation grade outcome as the grade suggested by the model. Although this approach increased computational costs and total processing time, it fundamentally improved confidence in the LLM-based assessment results, which is essential in all educational contexts. This structured methodology ensured that every aspect of the students' written responses was accurately assessed, mirroring a well-executed human expert evaluation. Together, these methodological choices provided a secure, transparent, and reproducible framework for exploring the role of LLMs in large-scale educational assessment (as shown in Figure 1 and Table 1).

3.3 | Methods to Evaluate Students' Written Responses with GPT-4o

To address the research questions, the analyses examined the grade alignment between the official human expert evaluation and the model, keyword usage, and translation effect, structured into 3 parts (as shown in Table 1).

First, the alignment analysis measured how closely the GPT-4o's grades matched those assigned by human expert evaluators at the end of 4 official human expert evaluation rounds. In addition to calculating the actual grade on a scale of 0–15-point difference, grades were categorized according to the difference between the model and the human expert evaluations: exact match (no difference), moderate variation (0–2-point difference), outlier (3–4-point difference), and severe outlier (⩾ 5-point difference). The distribution of responses across these 4 categories quantified the consistency between GPT-4o and human expert evaluation.

Second, the automatic keyword analysis tracked the presence of 24 keywords, which the official evaluation guidelines identified as indicators of correct responses. Each of the 3 rainfall types discussed in the examination had 8 associated keywords, making a total of 24. The presence of these keywords in the responses was tracked and correlated with grades from both GPT-4o and human expert evaluators, by drawing attention to cases in which the model awarded points despite incorrect keyword usage by human standards.

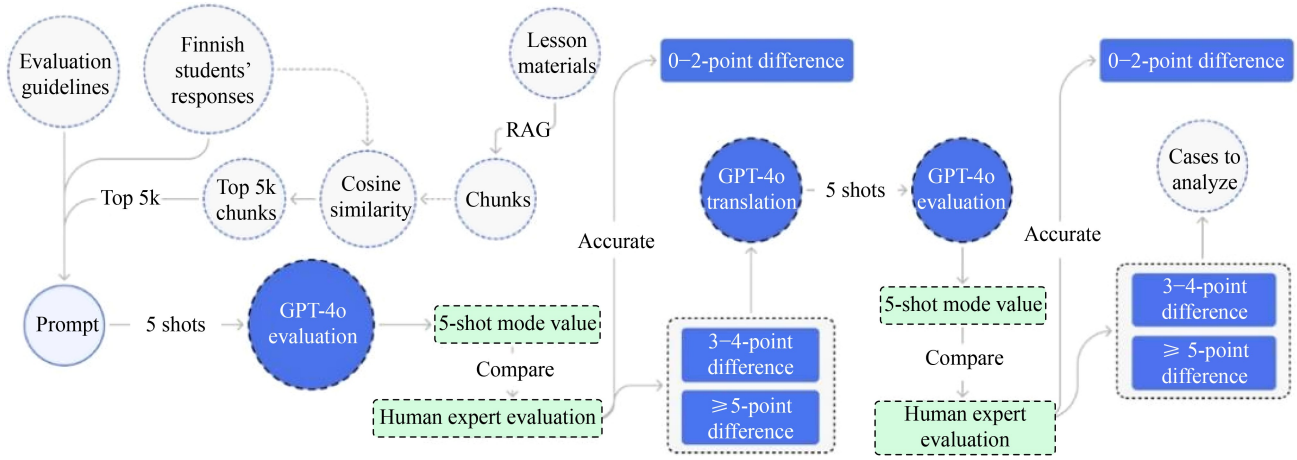

Figure 1 Evaluation process of students' open-ended responses with GPT-4o. RAG: retrieval-augmented generation.

Table 1 Settings of GPT-4o for evaluating students' open-ended responses (with a 5-step pipeline for preparing and executing the evaluation and 3-step pipeline for analysis)

Input category	GPT-4o evaluation setup
Learning material input	GPT-4o was provided with the examination chapter from a geography textbook used by students
Examination question input	The examination question was input into GPT-4o to define the model's task
Evaluation instruction input	GPT-4o received the same evaluation guidelines (250 words or 1,910 characters, without space) as human expert evaluators, instructing the model to grade students' written responses on a scale from 0 to 15, with 0–5 points assigned for each type of rainfall
Response input and verification	A total of 1,016 students' written responses were input into the GPT-4o, with each response recalled 5 times to verify correctness
Evaluation execution	Responses were graded (parameter temperature = 0.0) from 0 to 15 points 5 times, selecting the mode value to ensure evaluation consistency
Grade comparison analysis	GPT-4o-generated grades were compared to human expert evaluators' grades; differences were categorized as exact match (no difference), moderate variation (0–2-point difference), outlier (3–4-point difference), and severe outlier (⩾ 5-point difference)
Outlier keyword analysis	The presence of essential content-related keywords (as indicated in the evaluation guidelines) was analyzed in severe outlier cases (⩾ 5-point difference)
Language impact analysis	The same analyses were conducted in English with outlier grade (⩾ 3-point difference) responses. All learning materials were translated into English, and the translation quality was verified

Third, the translation analysis evaluated whether translating responses from LRL (Finnish) into HRL (English) affected grade alignment. This article's author wrote a textbook chapter in both Finnish and English on which the examination question was based. Each response, originally written in Finnish and later translated into English by a high-end LLM, was carefully reviewed by the author, who is fluent in both languages and an expert on the examination topic. The original and translated versions were fully consistent in content. This expertise provided a strong assurance of accuracy, although additional procedures, such as inter-rater reliability or back-translation checks, were not employed. Responses with at least 3-point difference between human and model grades were selected for detailed examination. Future studies could improve translation validity through inter-rater reliability or back-translation checks.

4 Results

4.1 | GPT-4o's Grade Alignment with Human Expert Evaluators' Grading Results

The study assessed the grade alignment between GPT-4o and official human expert evaluation grades for 1,016 students' written responses. Unlike trained human expert evaluators, who had specific grading experiences and extended evaluation training sessions, GPT-4o relied solely on prompt instructions and evaluation guidelines but enhanced accuracy by repeating the evaluation 5 times, as shown in Figure 1 and Table 1.

The correlation between grades assigned by human expert evaluators and GPT-4o was strong and statistically significant (Pearson's r = 0.87 for 5-shot evaluation and r = 0.85 for 1-shot evaluation; both $p < 0.001$, here r denoting correlation coefficients and p indicating P value). GPT-4o assigned slightly higher scores than human expert evaluators in over half of the cases—53.90% in 5-shot evaluation and 56.50% in 1-shot evaluations. Conversely, it assigned lower scores in 23.30% of cases for both 1-shot and 5-shot evaluations (Table 2). Compared to the 1-shot evaluation, the 5-shot evaluation resulted in a slight increase in exact matches; 18.10% of grades moved closer to the human expert evaluation, 70.30% remained the same, and 11.60% moved further away. Running the evaluation 5 times resulted in only a slight improvement: 5-shot evaluation increased the rate of exact matches by 2.50 points and reduced the average score difference by 0.11 points, thus confirming high consistency.

For scores very close to the human expert evaluation (within ±1-point difference), GPT-4o matched human expert scores 52.90% of the time in the 5-shot evaluation. The best grade alignment occurred for scores 2 (68.20%), 11 (62.40%), and 4 (60.00%), while the weakest grade alignment was for scores 7 (42.00%), 3 (35.00%), and 6 (34.80%). Exact matches occurred in 22.70% (5-shot evaluation) and 20.20% (1-shot evaluation) of cases, with the highest agreement at the low end of the scale (87.80% for 0-point difference; 61.10% for 1-point difference).

A ±2-point difference (approximately 12.50% on a 16-point scale) is comparable to typical human expert evaluators variability. Factors (such as text features, response quality, and grade bands) can affect scoring difficulty, with some features influencing research and writing assessments differently. Prior research has shown that individual evaluators' scoring can vary from one response to another, and even across different days (Palermo, 2022). In GPT-4o's assessment, 75.00% (5-shot evaluation) and 71.60% (1-shot evaluation) of scores fell within this 2-point difference range

Table 2 Comparison of official human expert evaluation grades and GPT-4o grading results (5-shot mode value)

Grade	Official human grading		GPT-4o grade			Average distance (points)	Original Finnish		English translation	
	n	Point	Same (%)	Higher (%)	Lower (%)		Outlier (%)	Correct (%)	Outlier (%)	Correct (%)
0	41	4.00	**87.80**	*12.20*	*0.00*	*0.20*	*2.40*	**97.60**	*0.00*	**100.00**
1	18	*1.80*	**61.10**	27.80	11.10	*0.94*	16.70	83.30	16.70	83.00
2	22	*2.20*	27.30	59.10	13.60	1.32	18.20	81.80	*2.40*	**97.60**
3	20	*2.00*	20.00	75.00	*5.00*	**2.10**	**35.00**	*65.00*	8.00	92.00
4	25	2.50	16.00	72.00	12.00	1.36	*12.00*	**88.00**	8.00	92.00
5	40	3.90	*10.00*	**80.00**	10.00	1.88	**35.00**	*65.00*	**32.50**	*67.50*
6	69	6.80	17.40	**82.60**	*0.00*	**2.30**	**44.90**	*55.10*	**36.00**	*64.00*
7	74	7.30	*13.50*	**81.10**	5.40	1.96	33.80	66.20	**23.00**	*77.00*
8	86	8.50	*10.50*	75.60	14.00	**1.99**	34.90	65.10	22.10	77.90
9	88	8.70	14.80	78.40	6.80	1.90	29.50	70.50	15.90	84.10
10	112	**11.00**	14.30	68.80	17.00	1.73	25.00	75.00	8.90	91.10
11	93	9.20	21.50	61.30	17.20	1.45	23.70	76.30	14.10	85.90
12	104	**10.20**	29.80	41.30	28.80	*1.29*	*13.50*	**86.50**	*4.80*	**95.20**
13	98	**9.60**	20.40	26.50	**53.10**	1.45	16.30	83.70	6.10	93.90
14	62	6.10	21.00	*9.70*	**69.40**	1.56	19.40	80.60	12.90	87.10
15	64	6.30	**34.40**	*0.00*	**65.60**	1.61	28.10	71.90	17.20	82.80
Total	1,016	100.10	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Average	N/A	N/A	22.70	53.90	23.30	1.63	25.00	75.00	15.00	85.00

Notes. Correct = 0–2-point difference between human expert evaluators and GPT-4o, outlier = at least 3-point difference from human expert evaluator grade, bold = 3 highest scores, italics = 3 lowest scores.

when the original Finnish was used for evaluation (Table 2).

However, wider scoring discrepancies also occurred: outliers with 3–4-point difference appeared in 22.00% of cases (5-shot evaluation) and 24.50% (1-shot evaluation). Severe outliers, with a difference of 5 or more points, occurred in 3.00% of 5-shot and 3.90% of 1-shot evaluations. GPT-4o faced greater challenges in aligning with human expert scores for responses officially graded in the lower range of a scale of 0–15-point difference: specifically, those receiving 6 points (44.90% of grades were outliers), 5 points (35.00%), and 3 points (35.00%). Conversely, grade alignment was better, with fewer severe outliers, for responses at the extreme of the grading scales: 0-point difference (2.40%), 4-point difference (12.00%), and 12-point difference (13.50%) (Table 2). The grades most commonly assigned by the model (12 and 11) were very close to those frequently assigned by human expert evaluators (10 and 12).

4.2 | Effect of LRL and HRL on GPT-4o Grading

Of the 1,016 responses originally analyzed in Finnish, 254 (25.00%) showed scoring discrepancies of at least 3-point difference (a grade difference of 18.80% or higher) between human and GPT-4o. To test for potential differences between LRL and HRL contexts, the original learning materials, evaluation guidelines, and outlier students' written responses were provided to GPT-4o in HRL (English). Given that much of the model's reinforcement learning from human feedback has been conducted in English, it was expected that using HRL (English) would lead to a better grade alignment. Moreover, translation can simplify keyword recognition because the complex morphology and compound words in LRL (Finnish) can obscure key term recognition by an LLM. The translation quality was meticulously verified before analysis.

The results indicated that translation into HRL (English) significantly improved the grade alignment between human and LLMs. The model's grading of English-translated responses corresponded more closely to human expert evaluators. First, when evaluating the quality of the English translations, 85.00% of the grades assigned by GPT-4o were within 2-point difference of the official human grading, which means the proportion of outliers decreased from 25.00% to 15.00% (as shown in Table 2). Second, after translation, 52.70% of the scores shifted closer to the official human grading, indicating a better consistency, while 15.00% diverged further. Third, translation into HRL (English) had a marked positive impact on severe outliers (⩾ 5-point

difference), with 76.70% of these cases showing reduced scoring discrepancies. To conclude, these results demonstrate the potential advantages of using HRL (English) with GPT-4o compared to an LRL (Finnish).

4.3 | Impact of Keyword Use on GPT-4o Evaluation

Official evaluation guidelines identified 24 essential keywords (8 per rainfall type) in the geography textbook to support grading. The presence of these keywords assisted both human expert evaluators and GPT-4o in the evaluation task and grading of responses.

Human expert evaluators and GPT-4o were closely matched in recognizing keyword presence, with strong correlations between students' written response length, keyword variety, and awarded grade. Typically, longer responses, including more keywords consistently received higher scores from both human expert evaluators and GPT-4o. Longer responses are tended to include a wider variety of keywords ($r = 0.676$, $p < 0.001$) and received higher grades both from human ($r = 0.685$, $p < 0.001$) and GPT-4o ($r = 0.555$, $p < 0.001$). Conversely, students' written responses with fewer keywords received lower grades from both. Moreover, responses with a wider range of keywords received higher grades both from human expert evaluators ($r = 0.832$, $p < 0.001$) and the model ($r = 0.817$, $p < 0.001$), while those with a narrower range received lower grades. This indicates the importance of keyword presence in evaluations and grading by both human and the model.

GPT-4o showed high reliability in the vast majority of cases, with 96.60% free from keyword misinterpretation. Scoring discrepancies occurred in 3.40% of students' written responses. In 1.80%, human assigned lower grades despite many keywords and detected incorrect usage that GPT-4o missed; in 1.60%, human awarded higher grades despite fewer keywords and recognized deeper understanding overlooked by the model. These differences stem from GPT-4o's reliance on keyword presence as probabilistic signals, suggesting a lack of nuanced contextual judgment of human expert evaluators who assess simultaneously both the accuracy and appropriateness of keyword use (Baidoo-Anu & Ansah, 2023; Pinto et al., 2023). The model primarily recognized the presence of relevant keywords within responses and used these as indicators of students' knowledge. This approach generally works well for students' written responses in which students use the keywords correctly and logically, as specified in the evaluation guidelines and the learning materials for this examination. However, LLMs can occasionally fail to recognize incorrect applications, favoring responses based more on keyword presence than on correct contextual usage. The model's reliance on approximations may cause it to miss the nuanced judgment humans apply in evaluating complex responses (Schlangen, 2024). This shortcoming points to the need to refine LLM training to better capture domain-specific subtleties for reliable educational assessments.

In contrast, human expert evaluators draw on their training and expertise to assess not only the presence of keywords but also the appropriateness and accuracy of keyword usage within the educational context, often considering subtle textual elements. This nuanced judgment allows for more accurate assessments, particularly where students misuse keywords or include them inappropriately.

5 Conclusions

This study evaluated the accuracy of GPT-4o in grading 1,016 open-ended responses from Finland's national high-stakes matriculation examination. GPT-4o was selected as one of the most advanced and widely deployed LLMs in 2024, with demonstrated potential for educational assessments. The primary empirical and practical contribution of this study is to provide a benchmark of GPT-4o against high-standard human evaluation in an official, population-level, high-stakes examination. The analysis focused on one high-stakes matriculation examination question, which typically requires 100–200 words essay-style responses on the rainfall types within the geography curriculum. Human grades were produced through multiple rounds of expert reviews under strict national criteria, providing a highly reliable reference.

The first contribution of this study is empirical: GPT-4o closely matched human grading, substantially outperforming earlier findings based on GPT-3.5 (Bui & Barrot, 2025; Mao et al., 2024). When grading responses directly in LRL (Finnish), 75.00% of GPT-4o's scores fell within ±2 points of human grades on the 16-point scale, well within typical inter-rater variation among expert human evaluators. Overall, GPT-4o and human expert evaluators showed a strong agreement in identifying grading-relevant keywords; responses containing a wider range of required terms consistently received higher scores from both. However, in 3.80% of cases, GPT-4o misinterpreted the contextual use of keywords, leading to grading discrepancies. LLM evaluations based on keywords primarily may overestimate response quality, as surface-level lexical matches can mask weak or incorrect conceptual understandings. While human expert evaluators can distinguish meaningful use from superficial or incorrect mentions, the model showed limitations in this respect. Excluding these cases, 96.20% of GPT-4o's evaluations

reliably identified the presence and breadth of essential concepts.

The second contribution is methodological: The effective use of LLMs in educational assessment depends on careful system design. Previous studies noted a range of ethical and security challenges, including risks to data privacy, potential leakage of training data, and general mistrust of model outputs (Huang et al., 2023; Nazaretsky et al., 2022; Ouyang et al., 2024). Precise prompting with access to official evaluation guidelines and relevant learning materials is essential as indicated in the earlier work. Setting the model's temperature parameter to 0.0 minimizes output variability (Hackl et al., 2023; Jauhiainen & Garagorry Guerra, 2025a). The application programming interface version and RAG setups help prevent data leakage. Repeating the evaluation process response recall and evaluation multiple times (such as 3-shot or 5-shot evaluations) further reduces potential randomness and hallucinations, substantially improving grading consistency and reliability (Jauhiainen & Garagorry Guerra, 2025a). These methods offer a replicable framework that increases the reliability and interpretability of LLM-based assessment at scale, even in a high-stakes examination context involving the written responses from over 1,000 students.

The third contribution is both conceptual and empirical: The study demonstrates that LLM-based assessment can be made more accurate by using HRLs (such as English), provided translations are faithful. Translating responses from LRL (Finnish) into HRL (English) improved grading accuracy, which reflects the HRL (English)-centric training of most LLMs. Here, translating responses from LRL (Finnish) into HRL (English) increased the grading alignment with human scores from 75.00% to 85.00%. This gain likely reflects GPT-4o's stronger semantic representations, vocabulary coverage, and reasoning reliability in HRL (English). While translation into a HRL improves the model's performance, it also reinforces linguistic inequalities between English and non-English users. Addressing this risk requires developing and fine-tuning LLMs directly in LRLs to ensure equitable assessment.

Advanced LLMs have substantial potential as supportive tools for educational assessment. They can significantly accelerate the large-scale educational assessment of open-ended responses and reduce educators' workload. Even currently, high-performing models can be used to audit human grading by flagging potential inconsistencies or outliers. This study, together with prior comparative research, shows that even state-of-the-art LLMs still fall short of the accuracy and reliability required for fully autonomous, high-stakes evaluation. As of 2026, fully replacing human expert evaluators is not yet feasible. Responsible adoption, therefore, requires a human-in-the-loop approach, in which LLMs may provide draft assessments or consistency checks while final decisions remain with educators. Full automation risks compromising fairness, transparency, and student trust.

Educator involvement in designing and validating LLM-based tools is essential for the usability and pedagogical relevance (Nazaretsky et al., 2022; Ouyang et al., 2024; Zhai, 2025). Although the current model implementation remains complex, rapid advances suggest that reliable evaluation pipelines will become easier to deploy. At present, only a small number of platforms, such as TurkuEval, meet the requirements for secure, scalable, and user-friendly LLM-based assessments. Looking ahead, task-specific LLMs constrained by curricula, guidelines, and student responses offer a promising path toward higher assessment efficiency with lower computational costs and total processing time. Apart from grading, LLMs also show the potential for generating structured and constructive feedback for students (Jauhiainen & Garagorry Guerra, 2026).

Persistent ethical challenges, such as transparency, bias, and the data security of LLMs, must be addressed to build trust and legitimacy in their use in educational assessments. Future research should examine how AI-assisted grading affects teacher and student motivation and perception of educational equity, which are critical for the sustainable integration of LLMs into assessments. Moreover, more recent models, such as GPT-5.2, Claude Sonnet 4.5, Llama 3, and Gemini 3, are already more advanced than GPT-4o. Their performance in assessment needs to be tested, as earlier studies indicate significant differences between models and their versions (Jauhiainen & Garagorry Guerra, 2024). Future research should evaluate LLM-based assessments across educational levels, expansive subjects, and demographic groups, with particular emphasis on multilingual contexts. Longitudinal comparisons between LLM-based and traditional assessments are needed to understand the impacts on learning outcomes and to support the development of reliable, fair, and globally inclusive AI-based evaluation systems.

Acknowledgments The reviewers are acknowledged for their constructive criticism of the original manuscript. Open Access funding was provided by the University of Turku.

Author Contributions The authors worked equally to develop the original manuscript, except that the first author revised the original manuscript and resubmitted the final version. The authors have approved the version to be published and agree to be accountable for all aspects of the work to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of Interest The authors declare that they have no conflicts of interest related to the content of this article.

Ethics Statements The authors declare that their Institutional Ethics Committee confirmed that no ethical review was required for this study. Written informed consent for participation was not required because all participants' data were anonymized before the statistical analyses were done.

Data Availability Statements The data supporting the findings of the study are not available due to the restrictions by the Finnish National Agency for Education.

References

Avnat, E., Levy, M., Herstain, D., Yanko, E., Ben Joya, D., Tzuchman Katz, M., Eshel, D., Laros, S., Dagan, Y., Barami, S., & et al. (2024). Performance of large language models in numerical vs. semantic medical knowledge: Benchmarking on evidence-based Q&As. *arXiv Preprint*, arXiv:2406.03855.

Baidoo-Anu, D., & Ansah, L. O. (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.

Barrot, J. S. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*, *29*(6), 7155–7179.

Beseiso, M., & Alzaharni, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*, *11*(10), 204–210.

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.

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*(2), 2041–2058.

Cain, W. (2024). Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends*, *68*(1), 47–57.

Chen, J. H., Chen, L. C., Huang, H., & Zhou, T. Y. (2023). When do you need chain-of-thought prompting for ChatGPT? *arXiv Preprint*, arXiv:2304.03262v2.

Conijn, R., Kahr, P., & Snijders, C. C. P. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, *10*(1), 37–53.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z. Y., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 194, 1–33.

Finlex. (2019, December 4). *Act on the matriculation examination (502/2019)*. Available from Finlex website.

Finnish National Agency for Education. (2019, July 11). *Lukion opetussuunnitelman perusteet 2019*. Available from Opetushallitus website.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2).

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X. B., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. *Scientific Reports*, *13*(1), 15310.

Glass, M., Rossiello, G., Chowdhury, F. M., Naik, A., Cai, P. S., & Gliozzo, A. (2022). Re2G: Retrieve, rerank, generate. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle: ACL, 2701–2715.

Hackl, V., Müller, A. E., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education*, *8*, 1272229.

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Levonian, Z. (2024). Can large language models make the grade? An empirical study evaluating LLMs ability to mark short answer questions in K-12 education. In: *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. Atlanta: ACM, 300–304.

Hu, T. J., & Zhou, X.-H. (2024). Unveiling LLM evaluation focused on metrics: Challenges and solutions. *arXiv Preprint*, arXiv:2404.09135.

Huang, H. Y., Tang, T. Y., Zhang, D. D., Zhao, X., Song, T., Xia, Y., & Wei, F. R. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: ACL, 12365–12394.

Jauhiainen, J. S., & 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. *Advances in Artificial Intelligence and Machine*

Learning, *4*(4), 3097–3113.

Jauhiainen, J. S., & Garagorry Guerra, A. (2025a). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, *62*(4), 1377–1394.

Jauhiainen, J. S., & Garagorry Guerra, A. (2025b). Educational evaluation with large language models (LLMs): ChatGPT-4 in recalling and evaluating students' written responses. *Journal of Information Technology Education: Innovations in Practice*, *24*, 2.

Jauhiainen, J. S., & Garagorry Guerra, A. (2026). Generative AI in educational processes: ChatGPT-4 in providing feedback to students' written responses. *Research and Practice in Technology Enhanced Learning*, *21*, 27.

Jauhiainen, J. S., Garagorry Guerra, A., Nylén, T., & Mäki, S. (2026). Generative AI in assessing written responses of geography exams: Challenges and potential. *Journal of Geography in Higher Education*, *50*(2), 210–222.

Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity*, *52*, 101522.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Lee, C., & Cha, K. (2024). Toward the dynamic relationship between AI transparency and trust in AI: A case study on ChatGPT. *International Journal of Human–Computer Interaction*, *41*(13), 8086–8103.

Lee, G.-G., Latif, E., Wu, X. S., Liu, N. H., & Zhai, X. M. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213.

Leech, G., Vazquez, J. J., Kupper, N., Yagudin, M., & Aitchison, L. (2024). Questionable practices in machine learning. *arXiv Preprint*, arXiv:2407.12220v2.

Mao, R., Chen, G. Y., Zhang, X. L., Guerin, F., & Cambria, E. (2024). GPTEval: A survey on assessments of ChatGPT and GPT-4. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. Torino: ELRA and ICCL, 7844–7866.

Matriculation Examination Board. (2024). *Ilmoittautuneet eri kokeisiin tutkintokerroittain 2015–2024*. Available from Ylioppilastutkintolautakunta website.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, *53*(4), 914–931.

Ouyang, S. M., Yun, H. Y., & Zheng, X. J. (2024, May). *How ethical should AI be? How AI alignment shapes the risk preferences of LLMs*. Available from HAI LAB website.

Page, E. B. (2003). Project essay grade: PEG. In: Shermis, M., & Burstein, J., eds. *Automated essay scoring: A cross-disciplinary perspective*. New Jersey: Lawrence Erlbaum Associates Publishers, 43–54.

Palermo, C. (2022). Rater characteristics, response content, and scoring contexts: Decomposing the determinates of scoring accuracy. *Frontiers in Psychology*, *13*, 937097.

Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023). Large language models for education: Grading open-ended questions using ChatGPT. In: *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*. New York: ACM, 293–302.

Qin, F., Li, K., & Yan, J. Y. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from China. *British Journal of Educational Technology*, *51*(5), 1693–1710.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527.

Scarlatos, A., Smith, D., Woodhead, S., & Lan, A. (2024). Improving the validity of automatically generated feedback via reinforcement learning. In: *The 25th International Conference on Artificial Intelligence in Education*. Zug: Springer, 280–294.

Schlangen, D. (2024). LLMs as function approximators: Terminology, taxonomy, and questions for evaluation. *arXiv Preprint*, arXiv:2407.13744.

Singla, Y. K., Parekh, S., Singh, S., Li, J. J., Shah, R. R., & Chen, C. (2023). Automatic essay scoring systems are both overstable and oversensitive: Explaining why and proposing defenses. *Dialogue & Discourse*, *14*(1), 1–32.

Stojkovic, J., Choukse, E., Zhang, C. J., Goiri, I., & Torrellas, J. (2024). Towards greener LLMs: Bringing energy-efficiency to the forefront of LLM inference. *arXiv Preprint*, arXiv:2403.20306.

Su, J. H., & Yang, W. P. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, *6*(3), 355–366.

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622.

Wei, J., Wang, X. Z., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 1800, 24824–24837.

Yao, S. Y., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 517.

Yu, S., Cheng, M. Y., Yang, J. Q., & Quyang, J. (2024). A knowledge-centric benchmarking framework and empirical study for retrieval-augmented generation. *arXiv Preprint*, arXiv:2409.13694v1.

Zhai, X. M. (2025). Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices. *Journal of Science Education and Technology*, *34*(6), 1323–1333.