Journal of Geography in Higher Education

ISSN: 0309-8265 (Print) 1466-1845 (Online) Journal homepage: www.tandfonline.com/journals/cjgh20

Generative AI in assessing written responses of geography exams: challenges and potential

Jussi S. Jauhiainen, Agustín Gagagorry Guerra, Tua Nylén & Sanna Mäki

To cite this article: Jussi S. Jauhiainen, Agustín Gagagorry Guerra, Tua Nylén & Sanna Mäki (05 Dec 2025): Generative AI in assessing written responses of geography exams: challenges and potential, Journal of Geography in Higher Education, DOI: 10.1080/03098265.2025.2593484

To link to this article: https://doi.org/10.1080/03098265.2025.2593484

JOURNAL OF GEOGRAPHY IN HIGHER EDUCATION
https://doi.org/10.1080/03098265.2025.2593484

Generative AI in assessing written responses of geography exams: challenges and potential

Jussi S. Jauhiainen[a,b], Agustín Gagagorry Guerra[a], Tua Nylén[a] and Sanna Mäki[a]

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

ABSTRACT

This article examines the application of Large Language Models (LLM) – GPT-4, Claude, Cohere, and Llama – to assess students' open-ended responses in Geography exams. The models' assessment scores were compared to assessment and scores by the original multi-stage human assessment as well as two additional human expert scoring. The case study considers the high-stakes national matriculation exam in Finland. The exam results play a crucial role in determining individuals' eligibility for higher education, including a study right in Geography at the university. We selected 18 essays that had originally been given 5 (basic), 10 (good) and 15 (excellent) points on a scale from 0 to 15 points. Findings show variability between LLMs and notable differences between LLM and human evaluations. The language of responses and grading instruction influenced LLM performance. These results highlight the potential and complexities of integrating generative AI today in learning assessments to score open-ended responses. Precise control of prompts and LLM settings proved crucial for the LLM to align with original assessment scores more closely.

ARTICLE HISTORY
Received 18 January 2025
Accepted 29 October 2025

KEYWORDS
Geography education; learning assessment; high-stakes; matriculation exam; LLM; generative AI

Introduction

In the rapidly evolving field of educational technology, the integration of generative artificial intelligence offers both opportunities and challenges in education-related practices from schools to universities, particularly to assess students' open-ended responses and essays. These technologies are gaining awareness and popularity in the 2020s, so it is crucial to investigate whether generative AI, and especially Large Language Models (LLM), can be effectively utilized in learning assessments, identify the conditions under which their potential is maximized, and explore ways to mitigate possible risks. LLMs are anticipated to develop to a capacity at which they can be systematically employed in the evaluation of students' written examinations and essays. Wilby and Esson (2024) have noted that AI and LLMs such as ChatGPT have transformative potential in Geography education and knowledge production but their critical usage is needed.

Assessment is a key to recognize learning differences among students (Hattie, 2008). Human and LLM assessments of student writing offer distinctive benefits and challenges. Human assessments excel in interpreting nuanced responses, recognizing creativity, and providing customized feedback, crucial for addressing individual learning needs. However, detailed assessments conducted by teachers are time-consuming, costly, susceptible to subjective bias, and inconsistency. Grading fatigue, personal preferences, and preconceived notions sometimes affect human scoring (Barrot, 2024). For instance, teachers might subconsciously give higher scores to students they view more favorably, while penalizing others due to student behavior, personal writing style or other subjective factors (Ferman & Fontes, 2022).

Anonymous learning assessment helps overcome some of this subjectivity, but not all. The teacher's concentration may also be momentarily lost, or their assessment line may develop or fluctuate due to human causes. This suggests a need for mechanisms in learning assessment that mitigate subjective biases

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

and inconsistency, to ensure that the evaluation is based on student performance alone. LLMs could potentially be suitable tools for this, even though they are not free from biases either.

Automated essay scoring (AES) and open-ended response evaluation using computers has been practiced for decades in various academic fields (Barrot, 2024), yet automatic essay scoring of open-ended responses remains one of the greatest challenges in natural language processing (Beseiso & Alzahrani, 2020). LLM-based systems offer speed and scalability, allowing rapid and consistent assessment of large volumes. When appropriately prompted, these models can approximate grading standards and provide automated feedback. However, ensuring their fairness and accuracy involves complex quality control. The models often struggle with the nuances in student responses due to their reliance on pre-trained data, which may not fully adapt to specific educational contexts. Better performance can be achieved by fine-tuning LLMs with domain-specific data, such as relevant learning materials and assessment guidelines, and by carefully designing prompts that align with the expected criteria for evaluating written responses (Latif & Zhai, 2023).

The release of ChatGPT-3.5 by OpenAI in November 2022 marked a significant advancement in the field of LLMs, facilitating easy natural language interactions through the chat version of the GPT model. This development ignited interest across various educational sectors, encouraging widespread use and further research into LLM applications (Bewersdorff et al., 2023; Bui & Barrot, 2025; Henkel et al., 2024). Turning into the mid-2020s, newer more advanced versions of GPT and other models have been introduced by different companies, each offering more refined capabilities and features, thus continually expanding the possibilities and applications for learning assessment purposes (Jauhiainen & Garagorry Guerra, 2024c).

This article examines the use of LLMs in assessing open-ended student responses from the official national matriculation examination at the end of secondary education, comparing the models' outcomes with scores assigned by human evaluators. The study is anchored in Finland's national matriculation exam in Geography, which rigorously assesses students' understanding of both Physical and Human Geography at the end of upper secondary school. The results of this exam are critical, as they influence university admissions. Additionally, the study examines students' responses in Finnish – a low-resource language (LRL) for LLMs – and in English, the most commonly used training language for these models and a high-resource language (HRL).

The case of the Finnish matriculation exam in Geography is very relevant, as the grading of the exam is rigorous, involving a systematic multi-step process to ensure fairness and high consistency. Multiple trained teachers and evaluation experts determine the final score of each of nine questions of the exam. This extensive assessment process is labor-intensive requiring significant human and financial resources.

This article analyses written responses exemplifying student performance at three levels of competence: basic (5 out of 15 points), good (10 out of 15 points), and excellent (15 out of 15 points), as indicated by the official grading results. The study comprehensively evaluates the capabilities of various LLMs in grading these essays. The anonymized data, devoid of any school or individual identifiers, was provided by the Finnish Matriculation Examination Board. The official assessment scores regarding these 18 responses serve as a benchmark to compare LLMs assessment results.

The responses were independently assessed by four LLMs – Claude, Cohere, GPT-4, and Llama – and two nationally recognized experts in Geography assessment. All evaluators, both human and models, were given the same official assessment guidelines and they had access to Geography textbook material that had been available to students before their exams. Neither the human expert evaluators nor the LLMs were informed of the original scores assigned to these responses, ensuring an unbiased re-evaluation based on the same criteria used in the official assessment process.

The aim of this paper is to promote the appropriate use of LLMs in learning assessment by producing basic information on its potential and challenges. The research questions for this article were: 1) How do the assessment scores of open-ended responses in Geography differ among and between human evaluators and LLMs?; 2) What differences emerge between LLM assessments conducted in a high-resource (English) and low-resource language (Finnish)?; and 3) How do humans and LLMs use keyword frequency and variety in open-ended responses to score them?

This article begins by discussing machine and human learning assessment, with a particular emphasis on initial findings and challenges encountered in research on LLM-based learning assessment. We then explore the rigorous evaluation system utilized in the Geography matriculation exam in Finland, examining its multi-stage process design for ensuring fairness and consistency, before introducing our empirical study.

The assessment results between human evaluators and LLMs are compared, the results after translating responses to English are analyzed, and the impact of keywords on LLMs' assessment is indicated. The article concludes by discussing findings and proposing future research directions.

Generative AI in educational contexts

Since the release of ChatGPT in 2022, the use of generative AI in learning assessments has increased significantly, revealing both potential and challenges. Much of current research comparing generative AI with human evaluators in educational contexts is experimental and based on small samples. These studies primarily assessed models like ChatGPT-3.5 across various tasks – ranging from student interactions and learning material design to assessment of responses and feedback provision from primary school to university levels (Ayeni et al., 2024; Jukiewicz, 2024).

Earlier studies indicate that systematic application of generative AI can significantly enhance teaching efficiency but long-term studies on their impact are lacking (Keppler et al., 2024). The LLMs have capacity of tailoring learning materials to align with students' abilities, receiving positive feedback for these adaptive learning materials and highlighting the potential for personalized learning enhancements (Jauhiainen & Garagorry Guerra, 2023, 2024a; Wen et al., 2024). Early experiments with generative AI in educational settings have shown that LLMs such as ChatGPT-3.5 can simulate different performance levels of typical university students by generating written exam responses (Bui & Barrot, 2025). More recently, designed generative AI agents have advanced generative AI uses in education by automating complex pedagogical tasks and tailoring them for specialized academic fields (Chu et al., 2025).

Generative AI is increasingly employed to assess student exams and provide feedback. Since the 1960s, AES systems have been used to assess student essays, analyzing responses based on textual features like word count and sentence length (Barrot, 2024). LLMs can easily detect correct or incorrect answers in multiple-choice exams. In general, when models are correctly instructed with precise prompts and provided with access to learning material, LLMs can effectively indicate the alignment of students' responses with study materials (Jauhiainen & Garagorry Guerra, 2024b, 2024c; Zhou et al., 2023). LLMs have also been used to provide tailored immediate feedback to students on their essays, responses, and arguments and specific comments tailored to each student's submission (Jukiewicz, 2024; Steiss et al., 2024).

However, LLMs often struggle with more complex assessment tasks (Bewersdorff et al., 2023). It is more challenging to assess open-ended responses or essays, especially those in which students have to demonstrate higher cognitive ability. Such responses are often categorized using education frameworks such as Bloom's Taxonomy and its revised version, the SOLO taxonomy, Webb's Depth of Knowledge, and similar frameworks (Irvine, 2021). Models have assessing limitations when the depth and complexity of students' responses increases beyond students' learning material (Zhou et al., 2023). The models face difficulties when responses require applying knowledge to new situations, synthesizing beyond the provided facts or developing new scenarios.

Furthermore, rarely studies on LLMs in learning assessment have discussed the adjustment of models' default parameter settings that impact their behavior. For example, the parameter "temperature" affect models' creativity and output variability, with higher settings introducing more randomness. The models' default settings often start at 0.5, leading to inherent unpredictability in assessing responses, which can undermine their reliability in learning assessments (Hackl et al., 2023; Jauhiainen & Garagorry Guerra, 2025).

One of the significant challenges in deploying LLMs for learning assessment is ensuring they have access to up-to-date information necessary for accurate task performance. Earlier studies, such as those by Lee et al. (2024), highlight the effectiveness of advanced techniques like Chain-of-Thought (CoT) prompting to ensure logical progression in tasks, crucial for achieving coherent assessment results (Latif & Zhai, 2023; Wei et al., 2023). Moreover, techniques such as Retrieval-Augmented Generation (RAG), a technique critical for enriching LLMs' responses with relevant information, remains underutilized or inconsistently applied in many studies (Piktus et al., 2021).

The use of single-shot iterations in generative AI assessments is problematic due to the inherent randomness in LLMs, which can lead to significant variations in performance from just one evaluation attempt. This makes it difficult to confirm the reliability of results and to distinguish between anomalies and

consistent patterns in evaluations. Implementing multiple iterations (multi-shot), i.e. conducting assessment several times for the same task, is essential to validate the results, ensuring that results are not only stable but also replicable. This approach allows researchers to more accurately discern systematic accuracy from anomalous results, enhancing the reliability of LLM-based assessments in educational settings.

Further complicating the issue, research on generative AI, particularly using models like GPT, has been predominantly model-specific, meaning insights derived from substantially weaker LLM, such as GPT-3.5, do not necessarily apply to mode advanced LLMs, like GPT-4 and GPT-5. While some studies, such as those by Hackl et al. (2023) and Henkel et al. (2024), suggest GPT-4 aligns closely with human assessments, others like Jukiewicz (2024) and Azaiz et al. (2024) advise caution due to inconsistencies in systematic evaluation outcomes. These findings emphasize the need for further studies to ensure accuracy and generalizable results across different LLM models.

Geography matriculation exam in Finland

In Finland, the Geography exam is an elective part of the compulsory national matriculation examination, as detailed in the law (Act on the Matriculation Examination, 2019). This critical exam is taken by upper secondary school students typically at the age of 19, marking the conclusion of their studies in upper secondary school. The exam assesses their competence and is crucial since more than half of university students are admitted based on their matriculation exam results.

One course in Geography is compulsory for all upper secondary school students. However, students opting for the Geography exam usually complete 1 to 3 courses more, earning maximally 8 academic credits, which translates up to 114 hours of study. These courses encompass a wide range of topics in both Physical and Human Geography and related methodologies, ensuring a comprehensive understanding of the subject (Act on the Matriculation Examination, 2019).

In the Geography exam, students face a structured test comprising three sections with a total of nine questions. They need to answer these in an open-ended format on the exam day. The first question is mandatory for all students, while the subsequent questions, which increase in complexity, are optional though also these generate points for the overall grade of the examination. More challenging questions assess not just the students' factual knowledge but also their ability to apply this knowledge to more complex situations, testing their analytical skills and understanding of geographical concepts in depth (Reaaliaineiden, 2024). Although no fixed length is required, students typically produce open-ended written responses of 150–300 words per question.

The Geography exam in Finland is crafted, reviewed and revised by a team of experts in Geography and other experts under the Matriculation Examination Board. The team integrates diverse expertise in the discipline, pedagogy and assessment, together with experienced in geography teaching across educational levels, to ensure that exam questions and assessment guidelines are fair, consistent and of high quality. The assessment guideline document for guiding the assessment of student responses follows a structure that is applied across examinations and includes general principles that apply to the entire exam and specific assessment criteria tailored for individual questions. Key general principles include structuring responses as open-ended, examining assignments from the geographical point-of-view, and using proper geographical terminology. The guidelines align with broader academic standards in the humanities and natural sciences (Reaaliaineiden, 2024).

The Act on the Matriculation Examination (2019) includes rigorous quality assurance measures implemented by the Matriculation Examination Board seeking to ensure that assessments are consistent and impartial across all student assessments (Matriculation Examination Board, 2024). The assessment process for the Geography exam includes multiple stages seeking to ensure fairness and thorough assessment.

Initially, students' responses are assessed by their own Geography teachers using official preliminary guidelines (Figure 1). This stage relies heavily on teachers' understanding and application of these guidelines to each response. The final assessment process is overseen by a chairperson seeking to ensure adherence to quality standards. The final assessment is conducted anonymously by a group of 10–15 qualified evaluators, referred to as "censors". They may include individuals involved in creating the exam questions as well as other experienced Geography teachers and assessment experts (Act on the Matriculation Examination, 2019).

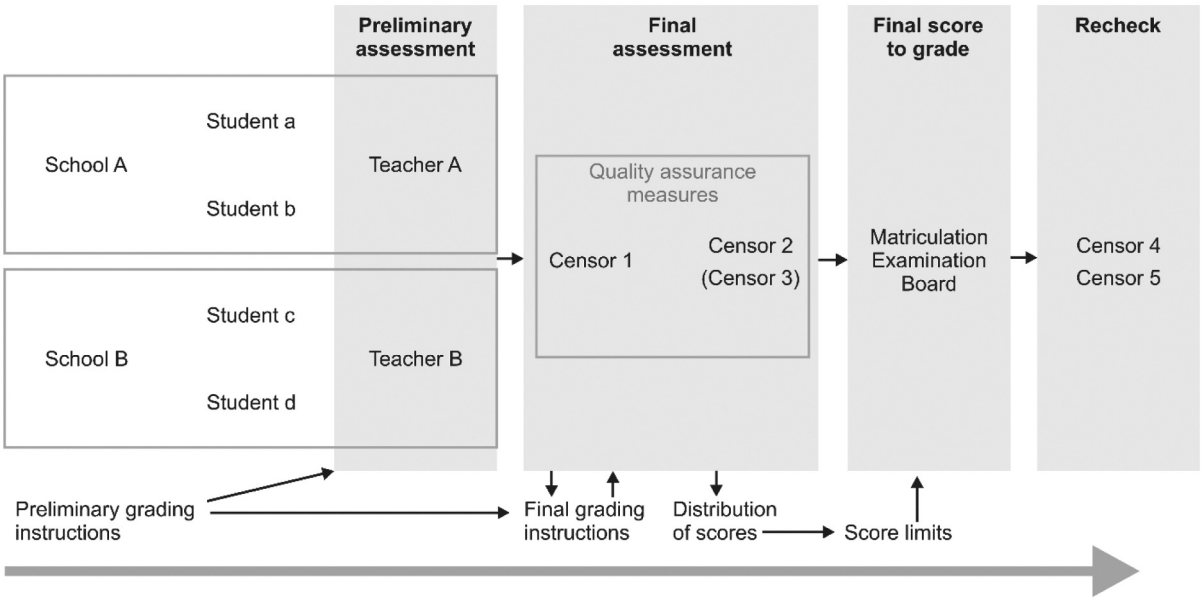

Figure 1. Finland's national matriculation exam assessment process in Geography.

On the day of the exam, censors receive the preliminary assessment guidelines from the Matriculation Examination Board, and detailed instructions from the chairperson. Each censor is tasked with assessing the Geography exams for about 100–300 students. To prevent bias, censors do not know the identity of the students they assess and must declare any potential conflicts of interest with specific schools ahead of time. The design of the evaluation guidelines by censors happens after they have reviewed all student responses. This is crucial for the subsequent censor meeting where evaluators exchange views and finalize assessment guidelines. The process is designed to minimize subjectivity and enhance the reliability of the assessments through a systematic secondary assessment procedure.

During the assessment process, the final scores assigned by the censors are rigorously compared against the preliminary score given by the teachers. This comparison is key to identifying and addressing any discrepancies in scoring. The system employed by the censors includes automated alerts that flag significant deviations in scores to ensure consistency and accuracy, however, after censors conducted their initial grading (Matriculation Examination Board, 2024).

Consistency of the exam results over years is assured by transforming scores to grades, based on the final score distribution (Figure 1). This relative assessment accounts for the difficulty level of an individual exam. A key to a fair assessment among the students taking one exam is, in turn, a consistency in how competence is rewarded with points. While the assessment process incorporates several measures to minimize subjectivity and human error, recognizing variations of one or two points on a scale from 0 to 15 are typical and not of major concern. If assessment errors are suspected in the final score, students can request a score review, where two unbiased censors reassess the issue. If an error is verified, adjustments are made; otherwise, the original score remains unchanged.

Material and methods

In the national matriculation exam of Finland in the autumn of 2023, 1,016 students responded in Finnish to one optional Geography question that is analyzed in this article. The question asked students to "Describe all three types of precipitation and name one characteristic occurrence area for each type of precipitation." (YLE, 2023a). This topic was covered in the second course of the Geography curriculum. Students who chose to answer this question were likely having a strong interest in Geography and had prepared on this topic before the examination. However, it was not mandatory to respond to this question.

The question does not require very complex level of knowledge but it is enough to repeat the main information from the learning material and apply it concisely. Students could earn a maximum of 15 points

on their open-ended response, a maximum of 5 point for the description of each of the three precipitation type. For each type, 1 point was given for correctly naming it, up to 3 points for accurately explaining its formation, and 1 point for identifying a relevant geographical area.

The data for this study – students' written responses from the Finnish secondary school matriculation examination in Geography in autumn 2023 – were obtained from the National Matriculation Board (research permit nr OPH-6154–2023). The dataset was fully anonymized, containing only the written responses and their corresponding grades (0–15), without any information on the respondents, their schools, or the evaluation process. The exam responses are confidential official documents, so examples of these cannot be provided here.

From all 1,016 students open-ended responses, we selected 18 responses with stratified random sampling method representing the variety of responses. Six responses demonstrated, according to the official assessment, students' basic competence (scoring 5 out of 15 points), six demonstrated good competence (scoring 10 out of 15 points), and six demonstrated excellent competence (scoring 15 out of 15 points). Among all 1,016 students, the lowest 20% scores were 6 points or below, while the top 20% scored at least 13 points, and the average response length was 170 words. In the sample of 18 responses, the basic competence group had shorter responses having on average 120 words in Finnish, the good competence group had 154 words, and the excellent competence group had the longest responses with 256 words. Later these responses were translated to English resulting in having 172, 229 and 366 words in English, respectively. However, response length was not a criterion that directly influenced grading. Human evaluators and LLMs did not know how many points each response had received.

The selected four LLMs (Claude, Cohere, GPT-4, and Llama with respective versions in use in the autumn of 2024, namely Claude 3.5 Sonnet, Cohere Command R, GPT-4o, and Llama 3.2) and two experts (Human Evaluator 1, and Human Evaluator 2) assessed these responses in Finnish and scored them on a scale from 0 to 15. Initially, two experts and four LLMs assessed and scored these responses having all materials in Finnish: the responses, the guidelines (YLE, 2023b) and the learning material (Jauhiainen et al., 2021a). Later, all these materials were translated into English using ChatGPT-4o version, with translation accuracy verified by the authors, and the learning material was available in English (Jauhiainen et al., 2021b). Then LLMs assessed and scored these responses. The two experts did not conduct assessment in English as they could have remembered the student responses in Finnish they had assessed earlier. However, LLMs did not have such remembering capacity. In total, 198 assessments were analyzed: LLMs made 72 assessments in Finnish and 72 in English, two human evaluators made each 18 assessments in Finnish, 36 in total, and the official assessment consisted of 18 assessment results in Finnish. Of the four LLMs examined, Cohere is headquartered in Canada, while the other three are based in California, United States. Both free and paid versions of these models were in use.

The LLMs were configured with a CoT prompt to improve performance on complex assessment tasks by encouraging step-by-step reasoning for the models. CoT improves transparency and reduces the likelihood of oversimplified or inconsistent outputs (Patil & Gudivada, 2024). Text segments were adjusted to fit each model's token limit, ensuring consistent evaluations across models, including smaller ones such as Cohere and LLaMA. The models were operated at a zero temperature (0.0) setting, promoting more consistent output behaviors and reducing variability in their tasks (Hackl et al., 2023). This methodical setup facilitated a uniform assessment process for all LLMs involved.

To confirm the precision of LLMs' recall of student responses, each response was processed through 10 API calls (10-shot). Any discrepancies between the original text and the LLM-recalled versions were examined using word count and Levenshtein distance, a measure for tracking differences between two text sequences. These methods are effective for detecting hallucinations and recall errors, as discussed in earlier research (Jauhiainen & Garagorry Guerra, 2024c). Responses containing errors were resubmitted to the LLM until they were accurately reproduced.

To ensure the consistency of scoring results, each model's assessment of the responses was repeated 10 times (10-shot). If variations in scores appeared across these trials, the mode (most frequently occurring value) of these results was used to determine the final assessment score for each model. This method was chosen to standardize the assessment process and minimize discrepancies. Initial attempts using a few-shot approach were abandoned as they led to inconsistencies in outputs, particularly impacting the performance of smaller models like Cohere and Llama.

The statistical methods employed in this study included descriptive statistics for analyzing the distribution and fraction of scores across the scale from 0 to 15 to observe trends and variations among different evaluators. Additionally, the assessment guidelines suggested certain keywords, or key geographical concepts, to indicate the correct response, and the study examined the influence of the number and variety of these keywords, present in each response, on the score awarded. Moreover, correlation analyses were conducted to measure how closely the LLM-based scores aligned with those of official (Matriculation Examination Board) and additional human evaluators regarding the keyword use. These analyses facilitated a detailed comparison of scoring methods, highlighting both similarities and discrepancies among the evaluators.

Results

Score distribution between human evaluators and LLMs

The first research finding revealed variations in scoring among two human evaluators and four LLMs when assessing 18 Geography exam responses. Assessment results by human evaluators and LLMs tended to converge better between each other than with the scores of the official assessment resulted in the matriculation exam evaluation process.

At the lower end of scores, no evaluators assigned scores lower than 4 points, suggesting a consensus recognition of baseline response quality (Table 1). Four responses officially scored 5 points (basic competence), both human evaluators scored them slightly higher, suggesting a lenient interpretation of assessment criteria. LLMs showed variability, assigning 4–5 points to about half of these responses. Conversely, for responses officially awarded 15 points (excellent competence), the scoring by both Human Evaluators and all LLMs tended to be more conservative than the official one. Only two responses received full marks from one human evaluator whereas another human evaluator did not give any of those, indicating a stricter assessment at the higher end of the scale (Table 1).

The official scoring averaged 10.0 points per response, with each of the six responses in the basic (5 points), good (10 points), and excellent (15 points) categories. Compared with the official scores, both Human Evaluators gave on average slightly less points to responses while each LLM gave more points to these responses. The scoring thus diverged between humans and the models.

In detail, Human Evaluator 1 was the closest to the official assessment results assigning on average 0.3 points (1.7% of the scoring scale) less than the official scoring. The scores varied between 4 and 15 having 13 as the mode value. Human Evaluator 2 was slightly stricter in assessment and gave on average 1.0 point (5.6%) less having scores between 5 and 14, and the mode value was 9 (Table 1).

Table 1. Scores to student responses given by LLMs, human evaluators and the official assessment in Finnish ($N = 18$).

Score	Claude		Cohere		GPT-4		Llama		Human_1		Human_2		Official	
Mean	11.3		12.7		10.7		10.6		9.7		9.0		10.0	
score	C	%	C	%	C	%	C	%	C	%	C	%	C	%
0														
1														
2														
3														
4					1	5.6	1	5.6	1	5.6	2	11.1		
5	1	5.6			1	5.6	2	11.1	1	5.6	2	11.1	**6**	**33.3**
6	1	5.6					2	11.1	2	11.1	2	11.1		
7	2	11.1					1	5.6	2	11.1				
8	1	5.6			2	11.1			1	5.6				
9					2	11.1	**4**	**22.2**	3	16.7	**4**	**22.2**		
10	1	5.6			2	11.1	3	16.7			2	11.1	**6**	**33.3**
11			3	16.7					2	11.1				
12	**4**	**22.2**	3	16.7	**7**	**38.9**	**4**	**22.2**			3	16.7		
13	2	11.1	**9**	**50.0**			1	5.6	**4**	**22.2**	2	11.1		
14	**4**	**22.2**	2	11.1							1	5.6		
15	**2**	11.1	1	5.6	3	16.7			2	11.1			**6**	**33.3**

For LLMs, these numbers for Claude were on average 1.3 points (7.2%) more, scores between 5 and 15 and having bimodal distribution of 12 and 14. Cohere deviated substantially from the official assessment results giving on average 2.7 points (15.0%) more, variating between 11 and 15, and having 13 as the mode value. GPT-4 delivered 0.7 points (3.9%) more, variating between 4 and 15 with a mode of 12. Llama was the closest of LLMs to the official assessment results, and it gave 0.6 points (3.3%) more, variating between 4 and 13, and having a bimodal distribution of 9 and 12 (Table 1).

Alignment between human evaluator and LLM assessments

The second research finding indicated that no perfect assessment alignment was observed between two human evaluators, four LLMs, and the official assessment. However, given the sample size of 18 responses, these findings should be interpreted with caution. Assessments were conducted in Finnish, and each evaluator provided different distribution of scores.

The scoring differences were categorized into four levels for clearer analysis. An exact match indicated a perfect alignment with the official scores and up to two-point difference was considered an acceptable variation (2 points or 13.4% difference from the official scores) in a scale from 0 to 15 points. Three or four points difference indicated an outlier, a significant discrepancy of at least 20.0% from the official scores, and a severe outlier was scores different at least 5 points (33.3%). This structured approach allowed systematic assessment of the alignment and variations in scoring standards across different evaluators (Table 2).

Each evaluator produced a different distribution of scores, and the official scoring process tended to amplify results, producing more responses at both the highest and lowest ends. When differences were categorized, small deviations were common, but significant and severe discrepancies also appeared, especially among LLMs. Human evaluators generally aligned more closely with the official scores, with most of their ratings falling within a narrow range of variation. In contrast, the LLMs displayed greater inconsistency: while they occasionally matched official scores, their results often diverged markedly, sometimes by wide margins.

In detail, for the four LLMs, comparing with the official scores, exact matches (5.6–22.2%) were almost at par with human evaluators (16.7–22.2%), depending on the model. However, allowing for a two-point deviation increased their alignment but compared to human evaluators (83.3–88.8%), LLM scores within this acceptable variation were clearly fewer (38.9–66.7%), thus LLMs having more outliers or severe outliers (Table 2).

At the individual model level, Claude slightly over-scored but remained relatively consistent with human evaluators, with no instances of scoring responses lower than Human Evaluators 1 or 2. It matched exactly with the official scores in rather few cases (11.1%), was within acceptable two points in 66.7% of scores and outliers regarded 27.8% of responses. It matched exactly with Human Evaluator 1 in 27.8% of scores and had 33.3% outliers, and with Human Evaluator 2, these shares were 11.1% and 44.5%, respectively (Table 2).

Table 2. Alignment of LLM scores (share of point deviation) with the official, human evaluators 1 and 2 assessments of students' responses in Finnish and English. In bold the most aligned llm and in *cursive* the least aligned llm results.

llm	Human	Match (%) (fin)	Match (%) (eng)	±2%) (fin)	±2%) (eng)	3–4%) (fin)	3–4%) (eng)	5- (%) (fin)	5- (%) (eng)
claude	Official	11.1	11.1	66.7	50.0	16.7	38.9	16.7	11.1
claude	Hum_1	**27.9**	22.2	66.7	77.8	22.2	16.7	11.1	5.6
claude	Hum_2	11.1	11.1	55.5	66.7	27.8	22.2	16.7	11.1
cohere	Official	5.6	**22.2**	44.5	50.0	22.2	11.1	33.3	*38.9*
cohere	Hum_1	5.6	5.6	44.5	50.0	22.2	33.3	33.3	16.7
cohere	Hum_2	11.1	5.6	38.9	38.9	22.2	22.2	*38.9*	*38.9*
gpt-4	Official	22.2	**22.2**	50.0	44.4	38.9	33.3	11.1	22.2
gpt-4	Hum_1	16.7	16.7	62.3	61.1	27.8	38.9	0.0	0.0
gpt-4	Hum_2	11.1	5.6	55.5	50.0	44.4	38.9	0.0	11.1
llama	Official	16.7	**22.2**	38.9	66.7	44.4	33.3	16.7	0.0
llama	Hum_1	22.2	16.7	72.2	77.8	22.2	22.2	5.6	0.0
llama	Hum_2	22.2	11.1	61.1	72.2	33.3	27.8	5.6	0.0
Official	Hum_1	16.7	–	88.8	–	5.6	–	5.6	–
Official	Hum_2	22.2	–	83.3	–	5.6	–	11.1	–
Hum_1	Hum_2	33.3	–	83.3	–	16.7	–	0.0	–

Cohere consistently gave higher scores, rarely matched others, and produced the largest share of outliers, thus its score alignment was lowest of the studied LLMs both regarding the official assessment ($r = 0.587$, $p = .011$) as well as Human Evaluator 1 ($r = 0.480$, $p = .044$) and Human Evaluator 2 ($r = 0.447$, $p = .063$). The model never gave less than 11 points, and it matched only 5.6% of cases with the official assessment and Human Evaluator 1 and 11.1% with Human Evaluator 2. The majority of Cohere's scores were outliers regarding the scores of the official assessment (55.6%), Human Evaluator 1 (55.6%) and Human Evaluator 2 (61.1%) (Table 2).

GPT-4 achieved the strongest overall alignment with both the official assessment ($r = 0.7618$, $p < .001$) as well as those by Human Evaluator 1 ($r = 0.8532$, $p = .000$) and Human Evaluator 2 ($r = 0.8423$, $p = .000$), though it still produced many substantial deviations. It matched exactly with the official scores in 22.2% of cases and was within acceptable two points 44.4% of the cases (in English), but outliers made another 55.5% of responses. With Human Evaluator 1, the model matched exactly in slightly fewer cases (16.7%) but outliers were substantially fewer (27.8%). Matching and being close with Human Evaluator 2 scores happened in fewer cases (11.1% and 44.5%) (Table 2).

Llama showed moderate consistency, aligning more closely with Human Evaluator 1 ($r = 0.771$, $p = .000$) than from Human Evaluator 2 ($r = 0.666$, $p = .003$) or the official assessment ($r = 0.649$, $p = .004$). It matched exactly with the official, Human Evaluator 1 and Human Evaluator 2 scores in 22.2% of cases. It was within acceptable two points from official scores in 38.9% of cases but outliers were 61.1%. The share of outliers was substantially fewer regarding Human Evaluator 1 (27.8%) and Human Evaluator 2 (38.9%) (Table 2).

LLMs' scoring across assessment languages

Student responses and assessment guidelines were translated to English to test the impact of using a high resource language (HRL), one with abundant digital data, tools, and resources, making it easier to process with LLMs, including in assessment contexts. Translation was conducted with ChatGPT-4. The quality of translations matched with the original responses, verified by one of the authors.

After translation from Finnish to English, the LLMs' scores changed substantially. Of 72 scores assigned by LLMs – 18 for each model – only 29 (40.3%) remained the same across language translation. ChatGPT-4 retained the same scoring in 50.0% of cases, indicating a moderate level of consistency. Other models showed lower consistency, with both Cohere and Llama matching 38.9% of their original scores and Claude in 33.3% of cases. The language of assessment, in this case English as an HRL and Finnish as a low-resource language (LRL), significantly influenced the scoring outcomes of the LLMs. Linguistic resource disparities shape model performance. This underlines the need to account for language-specific factors when applying LLMs in educational assessment to ensure validity, reliability, and equity.

Comparing scores across languages illustrated varied levels of consistency among LLMs (Table 2). For example, GPT-4 exhibited the most stable performance, maintaining a 22.2% exact match rate with the official scoring in both Finnish and English. In contrast, Cohere's performance fluctuated notably, with only a 5.6% exact match in Finnish but improving to 22.2% in English. Further analysis within an acceptable up to 2-point difference range showed divergent patterns: Claude aligned more closely with official scoring in Finnish (66.7%) compared to English (50.0%), whereas Llama performed better in English, showing a significant improvement in agreement after translation. It was within acceptable two points in 38.9% of responses in Finnish and 66.7% in English.

Presence of keywords and assessment results

The official assessment guidelines suggested the presence of specific keywords, such the rain type name (convectional, orographic and frontal precipitation), place names (such as Amazon, Norway, subtropical Africa, etc.) and key features (such as saturation, polar front, etc.) in student responses that would indicate correct elements regarding the question about Geography, namely about each three rain types. Both humans and models scores were connected to these keywords present in responses. The correlation analysis revealed nuanced differences in how human evaluators and

Table 3. Pearson correlation coefficients between keyword presence and score assigned in official assessment, by human evaluators 1 and 2 and LLMs.

Evaluator	Unique Keywords (r)	Unique Keywords (p)	Total Keywords (r)	Total Keywords (p)
Human_1	0.890	0.0000	0.856	0.0000
Human_2	0.828	0.0000	0.688	0.0016
Official	0.824	0.0000	0.802	0.0001
Claude	0.787	0.0001	0.639	0.0043
GPT-4	0.755	0.0003	0.703	0.0011
Llama	0.692	0.0015	0.703	0.0011
Cohere	0.398	0.1019	0.602	0.0082

models weighted these factors. However, the sample size was rather small for major generalizations beyond the sample.

In the official assessment, high correlations existed between the presence of unique keywords and their total occurrences in responses reflecting both keyword diversity and their repetition (Table 3). Human Evaluator 1 seemed to take into account strongly both diversity and repetition, as evidenced by the highest correlations with keyword variety and frequency among all evaluators to explain the final scores. In contrast, Human Evaluator 2, while valuing unique keyword presence, demonstrated a significantly lower correlation between scores and keyword frequency compared to both the official assessment and Human Evaluator 1. While human evaluators typically relied more on both presence and variety of relevant keywords to inform scoring, this relationship was less clear in LLMs, and keyword presence and variety had lower correlation to explain the models' scores (Table 3).

In detailed analysis of LLMs (see Table 3), Claude showed a moderate correlation between unique keyword presence and scores, yet it had a notably lower correlation for total keyword occurrences. Cohere displayed the most significant gaps in its keyword-based assessment of students' responses, with very weak correlations for unique keywords and lower than other models regarding total occurrences of keywords explaining the scores it provided. Keyword variety had minimal impact on its outcomes. Compared with Claude, GPT-4 displayed a somewhat lower correlation for unique keywords but a higher one for total occurrences, suggesting a lesser focus on keyword variety in providing scores. Llama demonstrated lower correlations across unique keywords and the same level for keyword metrics as GPT-4.

In one particular student response, the final score by all three human assessments was unanimously lower than by LLMs. This response included 9 unique keywords, representing 37.5% of all possible keywords, and had 13 total occurrences of these keywords. However, the student had used many keywords incorrectly in the response. All human evaluators gave it a consistent score of 5 points (basic competence) despite evidence of keyword presence and variety. However, the average score from LLMs was 9.50 (good competence), with a standard deviation of 2.52, indicating a significant variability. In scoring of this response, LLMs emphasized the presence of keywords rather than holistic understanding how these keywords were used. Overall, LLM scores were clearly higher and along wider range. Claude awarded this response 7, GPT-4 and Llama each 9, and Cohere 13.

Conclusions and discussion

This study examined assessment in Finland's national matriculation exam, focusing on the detailed analysis of scoring of open-ended Geography responses by official human assessment, two independent human evaluators, and four LLMs. The responses were selected through purposeful stratified sampling to represent basic, good, and excellent performance levels, six responses in each. The analysis highlights the degree of consistency and alignment within and between human evaluators and LLMs, with implications for the use of LLMs in assessing open-ended responses in geography and other disciplines.

As the first contribution, this article advances understanding of how to conduct LLM-based assessments rigorously and how to compare them with human evaluations of open-ended responses. Besides security and ethical issues, this study emphasizes the importance of careful model pre-configuration to ensure technical accuracy, ethical compliance, and secure practices. Employing multiple recall and scoring rounds mitigates the inherent randomness of LLM outputs, while adjusting the model's parameter settings

(temperature) improves focus and consistency. Nevertheless, LLMs may still reproduce biases rooted in their training data or developers' interventions (Wilby & Esson, 2024; Zhou et al., 2023), raising concerns about fairness and accuracy. It is, however, anticipated that rapid model development will alleviate many of these limitations in the near future. Few prior studies have systematically examined systematic comparison between humans and LLMs in assessment while carefully verifying the conditions required for reliable LLM-based assessment, such as model recall accuracy and parameter settings (Hackl et al., 2023; Jauhiainen & Garagorry Guerra, 2024a). Furthermore, earlier research often relied on less advanced models, such as GPT-3.5, and rarely compared multiple LLMs simultaneously.

The second contribution concerns the impact of assessment language on LLM-based assessment. Previous research has largely concentrated on evaluations in English, as LLMs are predominantly trained on HRLs such as English (Wilby & Esson, 2024; Zhou et al., 2023). This study extends the discussion by directly comparing assessments conducted in both Finnish (LRL) and English, demonstrating that language influences LLM-based scores. Importantly, translating responses from LRL such as Finnish into English does not uniformly improve model performance. While some LLMs aligned more closely with official assessments after translation, others diverged further, suggesting that such variation is model-specific rather than strictly language-dependent. The small sample size, however, limits definitive conclusions. Educational institutions need to select LLMs carefully in relation to their language requirements and consider whether conducting assessments universally in English could enhance consistency, given the predominance of English in LLM training data. At the same time, translation may obscure nuances in student writing, highlighting the importance of optimizing LLM use in multilingual assessment contexts.

The third contribution concerns the challenges of LLM-based assessing a broad and complex discipline such as Geography. In this field, students' appropriate use of keywords and geographical concepts is central to evaluating the accuracy of their responses. Higher-scoring written responses generally contain more of these keywords, often specified in the assessment guidelines, indicating stronger subject knowledge. Our results show that both human evaluators and LLMs assign higher scores to responses with a greater number of both individual and total keywords. However, compared to human evaluators, LLMs appear to emphasize keyword frequency over diversity, which affects scoring outcomes, and a low sensitivity to recognize incorrect use of concepts. There is a need for further tuning and calibration of LLMs to better reflect the nuanced judgments of experienced human evaluators, particularly in academic contexts where diverse keyword usage demonstrates deeper understanding and mastery of subject matter in open-ended responses.

A broader reflection regards grading open-ended responses and essays overall. In this study, clear grading differences occurred between the official score and those of the two human evaluators, between two human evaluators, between human evaluators and LLMs, and among the LLMs themselves. The independent human evaluators tended to assign slightly lower scores than those involved in the national multi-stage process, which may reflect the collective calibration built into that system. The discrepancies of at least 10–15% in score matching were common. This challenges what counts as the "correct" grade in a high-stakes setting such as the national matriculation exam: should final scores reflect a consensus among expert humans and top-performing LLMs, or remain the outcome of multi-stage human review that may be shaped by administrative pressures across evaluation rounds? Grading open-ended responses is inherently subjective, so further research is needed to define correctness. In our study, no score was treated as ground truth; we compared differences only. Future scoring systems might combine human rating plus suggestions from the best LLMs to harness human nuance and LLM consistency. Even now, LLMs could flag potential human errors: large human – model scoring discrepancies should trigger a targeted recheck though the error might be in the LLM.

As of 2025, the most advanced LLMs already demonstrate considerable promise for the systematic evaluation of open-ended responses, but their use must be carefully calibrated and validated. As the models develop, they are able to assess longer responses and essays with the same precision than shorter ones, however, not with the same financial costs as longer assessment require more energy and tokens. From an academic perspective, this study highlights a pressing research agenda: how to conceptualize correctness, fairness, and reliability in grading when generative AI enters assessment systems. From a practical perspective, LLMs are best positioned – at least for now – as supplementary tools to support teachers as evaluators of students' written texts, particularly in broad subjects like Geography. This combined approach would harness the scalability and consistency of LLMs together with human expertise and contextual

judgment, aiming toward an assessment system that is not only more efficient but also balanced, fair, and pedagogically sound.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Act on the Matriculation Examination. (2019). Act on the matriculation examination 502/2019. Retrieved November 27, 2024, from https://www.finlex.fi/en/laki/kaannokset/2019/en20190502.pdf

Ayeni, O., Mohd Al Hamad, N., Chisom, O., Osawaru, B., & Adewusi, O. (2024). AI in education: A review of personalized learning and educational technology. *GSC Advanced Research and Reviews*, *18*(2), 261–271. https://doi.org/10.30574/gscarr.2024.18.2.0062

Azaiz, I., Kiesler, N., & Strickroth, S. (2024). Feedback-generation for programming exercises with GPT-4. *TiCSE 2024: Proceedings of the 2024 on Innovation and Technology in Computer Science Education*, *1*, 31–37. https://doi.org/10.1145/3649217.3653594

Barrot, J. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*, *29*(6), 7155–7179. https://doi.org/10.1007/s10639-023-12083-y

Beseiso, M., & Alzarani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*, *11*(10). https://doi.org/10.14569/IJACSA.2020.0111027

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

Bui, N., & Barrot, J. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education Information Technology*, *30*(2), 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P., & Wen, Q. (2025). Llm agents for education: Advances and applications. *arXiv*. https://doi.org/10.48550/arXiv.2503.11733

Ferman, B., & Fontes, L. (2022). Assessing knowledge or classroom behavior? Evidence of teachers grading bias. *The Journal of Public Economics*, *216*, 104773. https://doi.org/10.1016/j.jpubeco.2022.104773

Hackl, V., Müller, A., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education*, *8*, 2308.02575. https://doi.org/10.3389/feduc.2023.1272229

Hattie, J. (2008). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Henkel, O., Hills, L., Boxer, A., Roberts, B., & Lovonian, Z. (2024). Can large language models make the grade? In *L@S '24: Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 300–304). Atlanta. https://doi.org/10.1145/3657604.3664693

Irvine, J. (2021). Taxonomies in education: Overview, comparison, and future directions. *International Journal of Educational Development*, *5*(2), 1. https://doi.org/10.20849/jed.v5i2.898

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

Jauhiainen, J., & Garagorry Guerra, A. (2024a). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, *62*(4), 1–18. https://doi.org/10.1080/14703297.2024.2422337

Jauhiainen, J., & Garagorry Guerra, A. (2024b). Generative AI and education: Dynamic personalization of pupils' school learning material with ChatGPT. *Frontiers in Education*, *9*, 1–18. https://doi.org/10.3389/feduc.2024.1288723

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

Jauhiainen, J., & Garagorry Guerra, A. (2025). Large language models in educational evaluation: ChatGPT-4 in recalling and evaluating students' written responses. *Journal of Information Technology Education-Innovations in Practice*, *24*, 002. https://doi.org/10.28945/5433

Jauhiainen, J., Salminen, J., Tolvanen, S., & Veistola, S. (2021a). *Tellus, 2: Sininen planeetta*. E–Oppi.

Jauhiainen, J., Salminen, J., Tolvanen, S., & Veistola, S. (2021b). *Globe, 2: The Blue Planet*. E-OPPI.

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

Keppler, S., Sinchairsi, W. P., & Snyder, C. (2024). Backwards planning with generative AI: Case study evidence from US K12 teachers. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4924786

Latif, E., & Zhai, X. (2023). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

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

Matriculation Examination Board. (2024). Assessment of the examination. Retrieved September 12, 2024, from https://www.ylioppilastutkinto.fi/en/assessment-and-certificates/assessment-examination

Patil, R., & Gudivada, V. (2024). A review of current trends, techniques, and challenges in large language models (LLMs). *Applied Sciences*, *14*(5), 2074. https://doi.org/10.3390/app14052074

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

Reaaliaineiden kokeiden määräykset ja ohjeet. (2024). https://www.ylioppilastutkinto.fi/fi/tutkinnon-toimeenpano/maaraykset-ja-ohjeet/koekohtaiset-maaraykset-ja-ohjeet/reaaliaineiden. Accessed 10.5.2024.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Booth, C. (2024). Comparing the quality of human and ChatGPT feedback of students writing. *Learning & Instruction*, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, S. (2023). Chain-of-thought prompting elicits reasoning in large language models. *ArXiv*, *2201*, 11903. https://doi.org/10.48550/arXiv.2201.11903

Wen, Q., Liang, J., Sierra, C., Luckin, R., Tong, R., Liu, Z., Cui, P., & Tang, J. (2024). AI for education (AI4EDU): Advancing personalized education with LLM and adaptive learning. In *KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6743–6744). Barcelona. https://doi.org/10.1145/3637528.3671498

Wilby, N., & Esson, J. (2024). AI literacy in geographic education and research: Capabilities, caveats, and criticality. *The Geographical Journal*, *190*(1), e12542. https://doi.org/10.1111/geoj.12548

YLE. (2023a). Describe all three precipitation types and name one characteristic occurrence area for each precipitation type (original in Finnish). https://yle.fi/plus/abitreenit/2023/syksy/maantiede/index.html#question-nr-2

YLE. (2023b). Grading instruction for geography in the Finnish matriculation exam, in the autumn of 2023 (original in Finnish). https://tiedostot.ylioppilastutkinto.fi/kokeet/2023-09-21_GE_fi/grading-instructions.html#question-nr-2

Zhou, T., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via jailbreaking: Bias, robustness, reliability and toxicity. *ArXiv*, 2301.12867. https://doi.org/10.48550/arXiv.2301.12867