Article

Not peer-reviewed version

Can ChatGPT Replace the Teacher in Assessment? A Review of Research on the Use of Large Language Models in Grading and Providing Feedback

Marcin Jukiewicz [*,†] and Michał Wyrwa [*,†]

Posted Date: 15 September 2025

doi: 10.20944/preprints202509.1233.v1

Keywords: generative AI; education; grading; assessment

Article

Can ChatGPT Replace the Teacher in Assessment? A Review of Research on the Use of Large Language Models in Grading and Providing Feedback

Marcin Jukiewicz *,† **and Michał Wyrwa** †

Faculty of Psychology and Cognitive Science, Adam Mickiewicz University, Wieniawskiego 1, Poznan, 61-712, Poland
* marcin.jukiewicz@amu.edu.pl
† These authors contributed equally to this work

Abstract

This article presents a systematic review of empirical research on the use of large language models (LLMs) for automated grading of student work and providing feedback. The study aimed to determine the extent to which generative artificial intelligence models, such as ChatGPT, can replace teachers in the assessment process. The review was conducted in accordance with PRISMA guidelines and predefined inclusion criteria; ultimately, 42 empirical studies were included in the analysis. The results of the review indicate that the effectiveness of LLMs in grading is varied. These models perform well on closed-ended tasks and short-answer questions, often achieving accuracy comparable to human evaluators. However, they struggle with assessing complex, open-ended, or subjective assignments that require in-depth analysis or creativity. The quality of the prompts provided to the model and the use of detailed scoring rubrics significantly influence the accuracy and consistency of grades generated by LLMs. The findings suggest that LLMs can support teachers by accelerating the grading process and delivering rapid feedback at scale, but they cannot fully replace human judgment. The highest effectiveness is achieved in hybrid assessment systems that combine AI-driven automatic grading with teacher oversight and verification.

Keywords: generative AI; education; grading; assessment

1. Introduction

In recent years, there has been a rapid increase in the number of publications addressing the use of generative artificial intelligence (GenAI), including large language models (LLMs), in education. This literature depicts numerous scenarios of positive transformation in educational practices [1]. Personalized and adaptive systems powered by LLMs are expected to adapt tailor educational content and learning pace to individual student needs, in effect acting as an intelligent tutor [2–5]. Such GenAI-based assistants start to provide psychological and emotional support to students, interacting with them directly but also monitoring their condition and alerting staff and relevant adults in need, thus potentially relieving overburdened school counseling services and other forms of institutional support [6–8]. Technological companies responsible for GenAI revolution promise the expansion and democratization of access to knowledge. With features like automatic translations and simplifying content's complexity, and the possibility of presenting the same message in alternative formats nearly immediately, GenAI tools can cater to student with diverse learning requirements [9–11].

This assistive role can relieve teachers and administration of educational institutions as well, aiding with documentations, preparing classes or feedback for students [12–15]. It enables the rapid generation of various types of questions and exercises, automatic adjustment of their difficulty level, personalization of educational materials, and helping with the automatic assessment of student progress [16,17]. Beyond these convenient enhancements, reseachers anticipate that engaging with GenAI could help develop digital competencies and more universal cognitive functions like creativity and analytical thinking. Being literate in digital technology, and GenAI in particular, is thought to help prepare students for future job market, and simultaneously improve their metacognitive skills. In the classroom, GenAI is said to inspire students by offering multiple perspectives and problem solutions, thus encouraging creative thinking. Teachers are already experimenting with image and text generators in artistic and humanities courses to spark imagination and increase student engagement [18–22].

But despite these broad expectations and hopes, experimental validation of GenAI usage in education remains limited. Existing work is often theoretical, focusing on speculation, hopes, conceptual analyses, or preliminary exploratory studies rather than systematic empirical investigations. The amount of opinions accompanied by the scarcity of results can potentially distort our perception of what are actual benefits, opportunities, and limitations of GenAI in education.

It could also be argued that many peer-reviewed studies on GenAI educational usage make little reference to the technical benchmarks and performance tests found in model cards. Such benchmarks are created for different purposes, engineering metrics and competitive evaluations, so their relevance to classroom practice is only partial. The end result, however, is that educational research often lags behind rapid model developments, assessing earlier versions while industry moves on to reasoning and agentic models. This gap highlights the constant need for systematic reviews of academic research.

This article aims to bridge this gap by systematically reviewing empirical studies focused on one specific aspect of technology impact on education: using GenAI to assess and grade student work. To map empirical studies on this topic, we followed the PRISMA protocol [23–25].

2. Methods

2.1. Eligibility criteria and the search

We were interested in the studies that:

1. Focused on students across different educational levels, i.e., both elementary and higher education, as well as specific educational contexts like studies on only medical students or only English as a Foreign Language (EFL) students.
2. Involved the use of LLM type of generative AI, both commercially available (like ChatGPT, Gemini, or Claude) as well as open sourced or custom LLMs. The usage of GenAI needed to be within the context of student grading (e.g., automatic scoring, response sheets, or descriptive grades). So, for instance, we were not interested in studies that involved using GenAI solely to provide feedback to students without the grading context.
3. Were empirical. Other reviews and purely opinion-based pieces were excluded from our review.
4. Were published in peer-review journals and conference proceedings after the official launch of ChatGPT in late 2022.

We didn't exclude studies that didn't involve direct comparisons of GenAI and human, as we expect a vast variety of empirical designs are being used to explore the usefulness of GenAI in student grading contexts. Also, while quantitative metrics give us a chance to compare different models and GenAIs directly—similarly as do different measures of human vs GenAI agreement—given how quickly such metrics and other benchmarks change, we decided to neither include nor exclude phrases directly linked to outcome metrics.

The full list of search criteria is presented in Table 1. We searched for articles in four databases: Scopus, Web of Science, PubMed, and EBSCO. All queries were performed in March 2025.

2.2. Screening and quality assessment

Our queries identified 1806 articles. After removing duplicates, we screened all records based on title and abstract to assess their relevance. We then retrieved and assessed the full texts of potentially eligible studies against our inclusion and exclusion criteria.

Table 1. Search Strings by Inclusion and Exclusion Criteria

Category	Search Terms / Filters
Inclusion: AI Technologies	"ChatGPT" OR "Claude" OR "Gemini" OR "Copilot" OR "Deepseek" OR "generative AI" OR "generative artificial intelligence" OR "large language model*" OR "LLM*" OR "text-to-text model*" OR "multimodal AI" OR "transformer-based model*"
Inclusion: Education Contexts	"adult education" OR "college*" OR "elementary school*" OR "graduate" OR "higher education" OR "high school*" OR "K-12" OR "kindergarten*" OR "middle school*" OR "postgrad*" OR "primary school*" OR "undergrad*" OR "vocational education"
Inclusion: Assessment Contexts	"assessment" OR "grading" OR "scoring" OR "feedback" OR "AI grading" OR "automated assessment" OR "automated scoring" OR "automated grading" OR "automated feedback" OR "automatic assessment" OR "automatic scoring" OR "automatic grading" OR "automatic feedback" OR "machine-generated feedback"
Exclusion: Study Types	NOT "bibliometrics analysis" OR "literature review" OR "rapid review" OR "scoping review" OR "systematic review"
Exclusion: Time Frame	AND PUBYEAR > 2022

All screening steps were performed independently by two reviewers. Discrepancies were resolved through discussion. The number of records identified, screened, assessed for eligibility, and included in the final synthesis is presented in the PRISMA flow diagram (Figure 1).

In the end, 42 papers were included in the review. Their methodological quality was assessed using the Mixed Methods Appraisal Tool [26]. Following its guidelines, no quality score was calculated. Instead, criteria were individually assessed per study. Most studies met at least three criteria relevant to their respective designs. The most common issue was the use of small, convenience-based, or otherwise unrepresentative samples. Additionally, several studies lacked sufficient reporting on analytical procedures. Full details of the quality appraisal are provided in the Supplementary Materials.

3. Results

3.1. Characteristics of the analyzed articles

The studies analyzed in this review cover 2023–2025, reflecting the dynamic growth of interest in using LLMs for educational assessment. In 2023, the first pilots and case studies were published (5 studies). The year 2024 saw the most significant increase in publications (29 studies), including GPT-4 implementations and experiments with assessment across various subjects. In the first half of 2025 (8 studies), analyses emerged that extended the use of GenAI to assessing creativity, spoken responses, and high-stakes examinations, indicating the rapid maturation of this technology within education.

The studies analyzed in this article focus primarily on higher education—as many as 28 out of 42 publications concern using large language models to assess student work at universities and technical colleges. The remaining studies address education at the secondary school level (4 studies), primary school level (5 studies), and those categorized as K-12 (5 studies), i.e., studies marked as K-12, meaning they simultaneously cover both primary and secondary education.

In terms of academic disciplines and school subjects, the most frequently represented areas in the analyzed studies are:

- **Computer science** – 7 studies (mainly automatic assessment of programming tasks and code),
- **Foreign languages** – 7 studies (assessment of essays and written assignments in foreign languages),
- **Mathematics** – 5 studies (calculation tasks, mathematical reasoning),

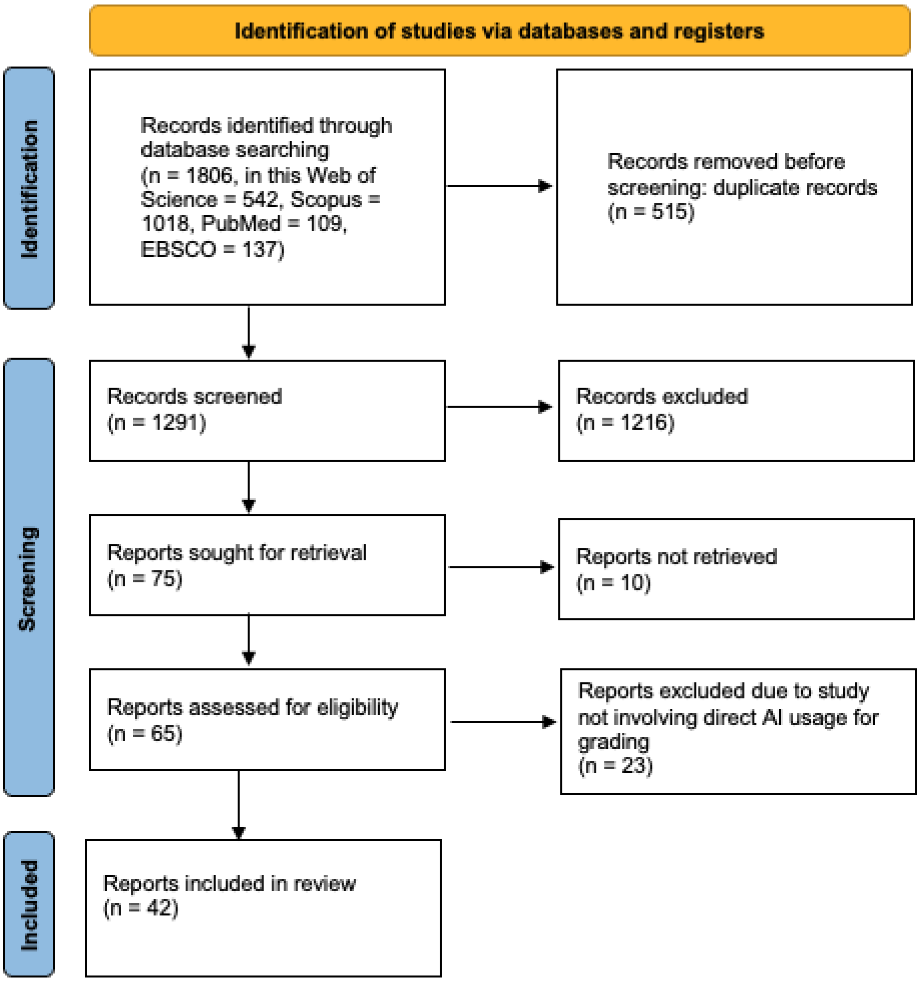
Identification of studies via databases and registers
Identification
Screening
Included
Records identified through database searching (n = 1806, in this Web of Science = 542, Scopus = 1018, PubMed = 109, EBSCO = 137)
Records removed before screening: duplicate records (n = 515)
Records screened (n = 1291)
Records excluded (n = 1216)
Reports sought for retrieval (n = 75)
Reports not retrieved (n = 10)
Reports assessed for eligibility (n = 65)
Reports excluded due to study not involving direct AI usage for grading (n = 23)
Reports included in review (n = 42)

Figure 1. PRISMA Flow Diagram

- **Medicine** – 5 studies (exam questions, open-ended tasks in medical education),
- **Engineering** – 3 studies (engineering projects, open-ended tasks),
- **Pedagogy, social sciences, and humanities** – single studies (assessment of essays, reflections, creativity).

Among the analyzed publications, the most frequently used model was GPT-4, which appeared in 23 studies. This was followed by GPT-3.5 (20 studies), the classic BERT model (7 studies), and the latest GPT-4o (3 studies). Considering the time needed for the peer-review process, this distribution indicates that researchers primarily focused on the most recent generative language models, while also conducting comparisons with earlier solutions.

Table 2. Selected LLM grading studies

Study	Level	Field	LLM type	Task type	Key findings
[27]	university or college	medicine	gpt4, gemini 1.0 pro	short answers	GPT-4 assigns significantly lower grades than the teacher, but rarely incorrectly gives maximum scores to incorrect answers (low "false positives"). Gemini 1.0 Pro gives grades similar to those given by teachers—their grades did not differ significantly from human scores. Both GPT-4 and Gemini achieve moderate agreement with teacher assessments, but GPT-4 demonstrates high precision in identifying fully correct answers.
[28]	secondary school	foreign language	gpt3.5, gpt4, iFLYTEK, and Baidu Cloud	essay	LLM models (GPT-4, GPT-3.5, iFLYTEK) achieved results very close to human raters in identifying correct and incorrect segments of text. Their accuracy at the T-unit level (syntactic segments) was about 81%, and at the sentence level 71–77%. GPT-4 stood out with the highest precision (fewest false positives), meaning it rarely incorrectly marked incorrect answers as correct.
[29]	university or college	math	other	calculation tasks	Even with advanced prompting (zero/few-shot), GPT-3.5 performs noticeably worse than AsRRN. GPT-3.5 improves when reference examples are provided, but it does not reach the effectiveness of the relational model.
[30]	university or college	economics	gpt4	short answers	GPT-4 grades very consistently and reliably: agreement indices (ICC) for content and style scores range from 0.94 to 0.99, indicating nearly "perfect" consistency between repeated ratings on different sets and at different times.
[31]	university or college	computer science	gpt3.5	coding tasks	The model correctly classified 73% of solutions as correct or incorrect. Its precision for correct solutions was 80%, and its specificity in detecting incorrect solutions was as high as 88%. However, recall—the ability to identify all correct solutions—was only 57%. In 89–100% of cases, AI-generated feedback was personalized, addressing the specific student's code. This was much more detailed than traditional e-assessment systems, which usually rely on simple tests.

Study	Level	Field	LLM type	Task type	Key findings
[32]	primary school	na	gpt3	other	AI grades were in very high agreement with the ratings of a panel of human experts (r = 0.79–0.91 depending on the task type), meaning that AI can generate scores comparable to subjective human judgments. The model demonstrates high internal consistency of ratings (reliability at H = 0.82–0.89 for different subscales and overall originality), making AI a reliable tool in educational applications.
[33]	secondary school	biology	bert	short answers	The alephBERT-based system clearly outperforms classic CNN models in automatic grading of students' short biology answers, both in overall and detailed assessment (for each category separately). The model showed surprisingly good performance even in "zero-shot" mode—that is, when grading answers to new questions about the same biological concepts it had not seen during training. This means that AI can be used to automatically grade new tasks, as long as they relate to similar concepts/rubrics.
[34]	university or college	foreign language	gpt4, gpt3.5	essay	ChatGPT 4 achieved higher reliability coefficients for holistic assessment (both G-coefficient and Phi-coefficient) than academic teachers evaluating English as a Foreign Language (EFL) essays. ChatGPT 3.5 had lower reliability than the teachers, but was still a supportive tool. In practice, ChatGPT 4 was more consistent and predictable in its scoring than humans. Both ChatGPT 3.5 and 4 generated more and more relevant feedback on language, content, and text organization compared to the teachers.

Study	Level	Field	LLM type	Task type	Key findings
[35]	university or college	mixed	gpt4, gpt3.5	short answers	GPT-4 performs well in grading short student answers (QWK $\geq$ 0.6, indicating good agreement with human scores) in 44% of cases in the one-shot setting, significantly better than GPT-3.5 (21%). The best results occur for simple, narrowly defined questions and short answers; open and longer, more elaborate answers are more challenging. ChatGPT (especially GPT-4) tends to give higher grades than teachers—less often awarding failing grades and more often scoring higher than human examiners.
[36]	university or college	foreign language	Infinigo ChatIC	essay	Texts "polished" by ChatGPT (infinigoChatIC) receive much higher scores in automated grading systems (iWrite) than texts written independently or with classic machine translators. The average score increase was about 17%. AI effectively eliminates grammatical errors and significantly increases the diversity and sophistication of vocabulary (e.g., more complex and precise expressions). This results in higher scores for both language and technical/structural aspects of the text.
[37]	university or college	na	gpt4	multiple choice tasks	Generative AI (GPT-4) achieved high agreement with human expert ratings when grading open tutor responses in training scenarios (F1 = 0.85 for selecting the best strategy, F1 = 0.83 for justifying the approach). The Kappa coefficient indicates good agreement (0.65–0.69), although the AI scored somewhat more strictly than humans and was more sensitive to answer length (short answers were more often marked incorrect). Overall, AI grading was slightly more demanding than human grading (average scores were about 25% lower when graded by AI).
[38]	university or college	engineering	gpt4o	short answers	GPT-4o (ChatGPT) showed strong agreement with teaching assistants (TAs) in grading conceptual questions in mechanical engineering, especially where grading criteria were clear and unambiguous (Spearman's p exceeded 0.6 in 7 out of 10 tasks, up to 0.94; low RMSE).

Study	Level	Field	LLM type	Task type	Key findings
[39]	k12	mixed	gpt3.5	essay	Essay scores assigned by ChatGPT 3.5 show poor agreement with those assigned by experienced teachers. Agreement was low in three out of four criteria (content, organization, language), and only moderate for mechanics (e.g., punctuation, spelling). On average, ChatGPT 3.5 gave slightly higher grades than humans in each category, but the differences were not statistically significant (except for mechanics, where AI was noticeably more "lenient").
[40]	primary school	math	gpt3.5, gpt4	calculation tasks	A system based on GPT-3.5 (fine-tuned) achieved higher agreement with teacher ratings than "base" GPT-3.5 and GPT-4, especially for clear, unambiguous criteria.
[41]	university or college	computer science, medicine	gpt4	essay, coding tasks	ChatGPT grading is more consistent and predictable for high-quality work. With weaker or average work, there is greater variability in scores (wider spread), especially in programming tasks.
[42]	secondary school	mixed	gpt4	short answers	GPT-4 using the "chain-of-thought" (CoT) method and few-shot prompting achieved high agreement with human examiners when grading short open-ended answers in science subjects (QWK up to 0.95; Macro F1 up to 1.00). The CoT method allowed AI not only to assign a score but also to generate justification (explanation of the grading decision) in accordance with the criteria (rubrics).
[43]	k12	mixed	bert	short answers	LLM models (here: BERT) achieve high agreement with human ratings in short open-ended response tasks in standardized educational tests (QWK ≥ 0.7 for 11 out of 13 tasks).
[44]	university or college	medicine	gpt3.5	essay	ChatGPT 3.5 could grade and classify medical essays according to established criteria, while also generating detailed, individualized feedback for students.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1233.v1

Study	Level	Field	LLM type	Task type	Key findings
[45]	university or college	medicine	gpt3.5	essay	ChatGPT 3.5 achieves comparable grading performance to teachers in formative tasks (including critical analysis of medical literature). Overall agreement between ChatGPT and teachers was 67%, and AUCPR (accuracy index) was 0.69 (range: 0.61–0.76). The reliability of grading (Cronbach's alpha) was 0.64, considered acceptable for formative assessment.
[46]	primary school	math	gpt4	calculation tasks	GPT-4, used as an Automated Assessment System (AAS) for open-ended math responses, achieved high agreement with the ratings of three expert teachers for most questions, demonstrating the potential of AI to automate the grading of complex assignments, not just short answers.
[47]	university or college	computer science	gpt4	coding tasks	ChatGPT-4, used as part of the semi-automatic GreAIter tool, achieves very high agreement with human examiners (98.2% accuracy, 100% recall, meaning no real errors in code are missed). Despite this high effectiveness, AI can be overly critical (precision 59.3%)—it generates more false alarms (incorrectly detected errors) than human examiners.
[48]	university or college	chemistry	gpt3.5, bard	short answers	Generative AI (e.g., ChatGPT, Bard) can significantly improve the grading of open-ended student work in natural sciences, especially thanks to data augmentation and more precise mapping of students' reasoning levels.
[49]	university or college	computer science	gpt3	short answers	AI (GPT-3, after fine-tuning) achieved very high agreement with human ratings in classifying the specificity of statements (97% accuracy) and sufficient accuracy in assessing effort (83%). Zero-shot models (without training) were much less accurate—automatic grading performance increases after fine-tuning on human-graded data.

Study	Level	Field	LLM type	Task type	Key findings
[50]	university or college	education	bert	essay	Pretrained models on large datasets, such as BigBird and Longformer, achieved the highest effectiveness in classifying the level of reflection in student assignments (77.2% and 76.3% accuracy, respectively), clearly outperforming traditional "shallow" ML models (e.g., SVM, Random Forest), which did not exceed 60% accuracy.
[51]	k12	mixed	gpt3.5turbo, claude3.5, gpt4turbo, spark	essay	Generative models such as GPT-4 often incorrectly assess the topic relevance of essays and generate feedback that is too general, impractical, or even misleading. This phenomenon results partly from "sycophancy" (excessive leniency) and LLM hallucinations—the models tend to give overly optimistic evaluations.
[52]	university or college	chemistry	bert, gpt3.5, gpt4	other	Language models (LLMs, e.g., GPT-4) perform worse in grading long, multifaceted, factual answers than in grading short responses—even with advanced techniques and rubrics, F1 for GPT-4 was only 0.69 (for short responses, 0.74). Using a detailed rubric (i.e., breaking grading into many small criteria) improves AI's results by 9% (accuracy) and 15% (F1) compared to classical, general scoring. Rubrics help the model better identify which aspects of the answer are correct or incorrect. GPT-4 achieves the highest results among tested models (F1 = 0.689, accuracy 70.9%), but still does not perform as well as for short answers and often misses nuances in complex, multifaceted student answers.
[53]	university or college	computer science	other	essay	ChatGPT should not be used as a standalone grading tool. AI-generated grades are unpredictable, can be inconsistent, and do not always match the criteria used by teachers.

Study	Level	Field	LLM type	Task type	Key findings
[54]	university or college	education	other	essay	ChatGPT 3.5 achieved moderate or good agreement with teacher grades (ICC from 0.6 to 0.8 in various categories), meaning it can effectively score student work based on the same criteria as humans. ChatGPT generated more feedback, often giving general suggestions and summaries, but less frequently providing specific solutions and explanations. AI feedback was less precise, less detailed, and less emotionally supportive, but more "inspiring" for some students. Students implemented teacher suggestions more often (80.2%) than ChatGPT's (59.9%), and AI feedback was more frequently rejected (40% of cases).
[55]	university or college	mixed	gpt4o, claude3.5, gemini1.5pro	other	The newest models (Claude 3.5 Sonnet, GPT-4o, Qwen2 72B) can be a valuable support in grading complex, open-ended tasks (authentic assignments) in higher education, using rubric-based assessment. GPT-4o and Claude 3.5 Sonnet produce results closest to human grading in terms of score distribution and consistency.
[56]	primary school	math	gpt4o	calculation tasks	GPT-4o can effectively support the assessment of tasks requiring logical and spatial reasoning in children, provided that advanced prompting techniques are used. Simple prompting (Zero-Shot/Few-Shot) is insufficient—grading effectiveness by LLMs increases significantly when using advanced methods, such as Visualization of Thought and Self-Consistency (multiple generations with majority voting). In the best settings, GPT-4o achieved 93–100% correct grades on tasks testing logical and spatial skills, such as coloring objects according to math-logical-spatial rules.
[57]	university or college	foreign language	bert, gpt3.5turbo	essay	ChatGPT (GPT-3.5-turbo) performs much worse than dedicated BERT or SIAM models in the task of assessing whether a pair of student essays represents progression (improvement) or regression in language skills. ChatGPT's effectiveness in this task was significantly lower (accuracy ∼56%) than the best models based on BERT architecture.

doi:10.20944/preprints202509.1233.v1

Study	Level	Field	LLM type	Task type	Key findings
[58]	university or college	medicine	gpt4	essay	The results obtained by ChatGPT strongly correlate with human ratings, especially for questions with clearer scope. AI enables immediate grading of large numbers of assignments and provides individual, detailed rubric-based feedback—which is hard for a single teacher to deliver with large groups.
[59]	k12	mixed	gpt4, gpt3.5	short answers	The GPT-4 model using few-shot prompting achieved human-level agreement (Cohen's Kappa 0.70), very close to expert level (Kappa 0.75). AI grades were stable regardless of subject (history/science), task difficulty, and student age.
[60]	secondary school	na	bert, gpt3.5turbo	multiple choice tasks	Tuned ChatGPT outperforms the state-of-the-art BERT model in both multi-label and multi-class tasks. The average grading performance improvement for ChatGPT was about 9.1% compared to BERT, and for more complex (multi-class) tasks even over 10%. The "base" version of GPT-3.5 is not effective enough—tuning is necessary for subject/class/task-specific data.
[61]	university or college	computer science	bert	coding tasks	A fine-tuned LLM can automatically assess student work in terms of higher-order cognitive skills (HOTS) without expert involvement, with greater accuracy and scalability than traditional systems.
[62]	primary school	foreign language	gpt4trubo	essay	Scores assigned by ChatGPT 3.5 were generally comparable to those given by academic staff. The average exact agreement between AI and humans was 67%. For individual criteria, agreement ranged from 59% to 81%. Using ChatGPT for grading reduced grading time fivefold (from about 7 minutes to 2 minutes per assignment), potentially saving hundreds of teacher work hours.
[63]	university or college	engineering	gpt3.5	short answers	ChatGPT 3.5 more often assigns average grades (e.g., C, D, E) and less frequently than humans gives the highest or lowest marks. ChatGPT is consistent—repeated grading of the same work yielded similar results.

doi:10.20944/preprints202509.1233.v1

Study	Level	Field	LLM type	Task type	Key findings
[64]	k12	computer science	other	coding tasks	The average effectiveness of GPT autograder for simple tasks was 91.5%, and for complex ones—only 51.3%. There was a strong correlation between LLM and official autograder scores (r = 0.84), but GPT-4 tended to underestimate scores, i.e., giving students lower results than they actually deserved (prevalence of so-called false negatives).
[65]	university or college	biology	gpt4	short answers	SteLLA achieved "substantial agreement" with human grading—Cohen's Kappa was 0.67 (compared to 0.83 for two human graders), and raw agreement was 0.84 (versus 0.92 for humans). This means the automated grading system is close to human-level agreement, though not equal to it yet.
[66]	university or college	math	gpt4, gpt3.5	calculation tasks	The average grade given by GPT-4 for student solutions is very close to grades assigned by human examiners, but detailed accuracy is lower—random errors occur, equivalent solutions can be missed, and calculation mistakes appear.
[67]	university or college	foreign language	gpt4, bard	essay	The highest reliability was obtained for the FineTuned ChatGPT version (ICC = 0.972), slightly lower for ChatGPT Default (ICC = 0.947), and Bard (ICC = 0.919). The reliability of LLMs exceeded even some human examiners in terms of repeatability. The best match between LLM and human examiners was for "objective" criteria such as grammar and mechanics. Both ChatGPT and Bard almost perfectly matched human scores for grammar and mechanics (punctuation, spelling).
[68]	university or college	engineering, foreign language	gpt4, gpt3.5	essay	ChatGPT (both 3.5 and 4) grades student essays more "generously" than human examiners. The average grades assigned by ChatGPT are higher than those given by teachers, regardless of model version. AI often gives higher grades even if the work quality is not high, leading to a "ceiling effect" (grades clustering near the top end). The agreement between human examiners was high ("good"), while for ChatGPT-3.5 it was only "moderate" and for ChatGPT-4 even "low".

Regarding model integration methods, 20 studies utilized API (enabling automated processing of large datasets). In contrast, in 16 cases, data were entered manually into the ChatGPT interface—a solution widespread in pilot projects and experiments conducted on a smaller scale.

3.2. Role of GenAI in grading

In the analyzed studies, the most frequently assessed types of student tasks by language models were essays [16 studies, e.g., 62,68] and short written answers [11 studies, e.g., 27,35,69]. Programming assignments appeared less often [5 studies, e.g., 31,47], as did multiple-choice questions [60,70]. This distribution shows that researchers are particularly eager to test the potential of GenAI where content analysis and the ability to understand argumentation are essential.

In many analyzed studies, particular emphasis was placed on instructing language models before the assessment process. Most often, this involved providing exemplar solutions or detailed scoring rubrics—the model would receive an answer key or a description of the criteria it should use when evaluating assignments. In 32 cases, prompts included a scoring rubric or a sample correct answer for comparison [e.g., 38,55,67]. Additionally, 8 studies applied holistic assessment criteria, such as overall impression [41,53].

In the included studies, GenAI played an active role in evaluating student work. However, the scope of this role varied. In most studies, artificial intelligence acted as an automatic grader, independently assigning points or grades to student responses with minimal human involvement, e.g., by generating a score for a short exam answer [e.g., 35,59,64]. In some cases, GenAI provided grades and generated feedback: the model assigned a score to the assignment and commented on its strengths and weaknesses, much like a human teacher would [e.g., 31,34,41]. There were also instances in which GenAI supported the teacher by suggesting a grade that the educator could then verify [so-called co-grading, e.g., 47].

3.3. AI vs. human grading quality

Findings on the alignment between GenAI and human graders were mixed. Several studies reported high agreement for specific tasks, including assessments of original thinking [32], macroeconomic and engineering tasks [30,38], programming assignments [31], of biology [69], mathematics [66], and across other courses [35]. Notably, some evidence suggests that LLMs can surpass human raters in terms of reliability in particular contexts, like in EFL essay scoring [34,67]. Nonetheless, longer essays remain challenging for GenAI [39,67]. In many cases, GenAI performance was comparable to human grading. For example, models from OpenAI and from iFLYTEK achieved similar accuracy scores in assessing Chinese writing [28]. Other studies indicate medium to good accuracy when compared to human assessment, e.g., in higher education exams and in medical education [45,63].

However, some studies also identified instances of GenAI performing worse or facing significant limitations. GenAI tends to perform worse in evaluating open-ended programming and mathematics tasks [46], in recognizing nuanced errors [67], in assignments requiring subjective judgment [67]. Issues also arise regarding the consistency of grading [66], as well as the tendency for hallucinations [31] or for producing overly generic feedback [51]. More specifically, GPT-3.5 struggled with open-ended coding tasks, especially complex ones, sometimes producing false corrections and failing to distinguish basic programming naming conventions like capitalization [31]. GPT-4 had difficulties in visual programming tasks for children [56] and in calculus problem-solving and grading [66]. And while in many instances, self-consistency of GenAI was high [30,35,67, eg.,], in some cases multiple grading attempts yielded varying results, even with the usage of a predefined scoring rubric [53].

There is no one grading tendency identified across all studies. In some cases, GenAI was shown to be more lenient than human raters, assigning higher grades for average to good essays [39], and more positive in its feedback [62]. In others, it was more strict [30]. And finally, in several studies, GenAI displayed a tendency to assign middle scores, avoiding extreme grades [53,55].

LLMs-based GenAIs were found to provide more comprehensive feedback than human graders [34,44]. But while larger in volume, such synthetic feedback was sometimes evaluated as too abstract, generic, or even confusing, with instances when final grade did not align with the generated comments [53].

The analyzed studies yield mixed results when comparing GenAI and human assessment. GenAI performed worse than humans in 11 cases. In 12 instances, its effectiveness was comparable to human ratings. Regarding just the feedback quality, GenAI was worse 5 times, comparable 3 times, and better than humans in 3 cases.

3.4. Factors impacting GenAI grading performance

The effectiveness of using GenAI for grading looks to be influenced by the quality of prompts, specific model of LLM being used, language of assessment, types of tasks, and level of education.

The final outcome is especially heavily influenced by the prompt given to the GenAI. Testing prompts beforehand and focusing on their precision, for instance by having detailed scoring rubrics supplemented with sample correct answers seem to be especially vital for consistent and effective grading [41,46,58]. On the other hand, using specific strategies often recommended by prompting online courses may improve GenAI grading performance but also lead to overfit with simpler sub-problems, as it is in the case of chain-of-thought and tree-of-thought prompting, as well as self-consistency checks [42,46].

Interestingly, only a handful of studies allow us to track the progress of subsequent generations of LLM models. Available publications rarely provide a basis for determining whether the newest models actually outperform their predecessors in grading or feedback quality. Initial results indicate that, for instance, GPT-4 generally outperforms GPT-3.5 in the accuracy of short-answer grading; however, this advantage is not yet consistent across all tasks and domains [34,35,52]. Fine-tuning and customizing models can help increase the reliability and precision of GenAI grading [67].

Language is another determinant. GenAI models perform best in English, although they are increasingly capable of handling other languages. Nevertheless, the effectiveness in detecting errors and linguistic nuances can vary depending on the student's working language [35,62], which highlights the importance of attention when using multilingual models to grade linguistic tasks.

Another critical factor is the nature of the task itself and what the student has written. Artificial intelligence performs very well with short, clearly formulated answers; however, its effectiveness often decreases when it comes to more complex assignments requiring broader context or subjective judgment [35,41,52,67]. Interpreting non-textual works (e.g., handwritten drawings) or very long and complex texts can also be challenging for GenAI [46].

On the other hand, GenAI demonstrates strong effectiveness in grading short-answer responses [35,58,59,63], in generating clear and well-structured feedback (though, as noted in other articles, this was also raised as a concern against GenAI-based assessment) [27], and in reducing grading bias. In some tasks, automated systems outperform human assessors in detecting language errors and in the speed and repeatability of grading [31,45].

3.5. Educational and pedagogical implications

The reviewed studies highlight several implications of using GenAI in educational assessment, from good practices regarding grading and feedback to broader instructional design and ethical considerations.

A recurrent theme is the potential of GenAI to reduce teacher workload while enabling more personalized and timely feedback, particularly in large-scale settings where manual grading is resource-intensive [43,45,56,63,67,68,71]. GenAI-assisted scoring can improve consistency by mitigating human fatigue and bias [41,62,68]. Moreover, formative assessments supported by LLMs can provide detailed explanations behind scores, helping both students and educators identify areas for improvement [31,42,44].

Beyond grading, GenAI contributes to instructional design. Models have been used to generate adaptive exercises tailored to individual learning needs [57], create exam questions and lesson plans, and expand the scope of what is evaluated to include more nuanced reasoning and argumentation

[48,66]. Such generated tasks can help assess not only theoretical knowledge but also practical skills and communication abilities [46,65,66]. Fine-tuning models on discipline-specific literature further aligns feedback with pedagogical frameworks and curricular goals [43]. Such tasks can help assess not only theoretical knowledge but also practical skills and communication abilities [46,65,66].

Despite these opportunities, studies consistently emphasize the necessity of human oversight. Current LLMs struggle with errors, algorithmic biases and loss of coherence, and often lack full contextual understanding [30,39,47,55,62,63,66]. Ethical concerns ranging from privacy and transparency to fairness and accountability are particularly salient for GenAI use in educational settings [27,42,60,62,66,67]. Consequently, GenAI is widely regarded as an assistive rather than autonomous tool, with integration strategies emphasizing hybrid approaches and robust institutional safeguards.

4. Discussion

A systematic review of 42 empirical studies on the use of generative artificial intelligence for grading student work reveals promising opportunities and significant limitations of these technologies. The results are highly heterogeneous—no clear, universal trend exists across all studies. A similar review by [72] found that existing studies on GenAI use in education lack a consistent and systematic approach from the educational perspective. To fill this gap, the authors mentioned above emphasize the key role of increased interdisciplinary, or even transdisciplinary, collaboration that would include educators and education researchers. We agree that only such joint efforts can ensure that implementing AI in education is matched with pedagogical frameworks.

In many cases, GenAI models (especially GPT-4) aligned well with teachers on well-defined tasks: short answers in science, engineering, or structured problems in other domains, where expected responses are clear. Here, GenAI sometimes matched human ratings or even surpassed them in consistency. Several studies confirmed high correlation or comparable accuracy between GenAI and teachers, suggesting that LLMs can imitate human grading in selected domains fairly well. For example, both OpenAI and iFlytek models assessed Chinese essays comparably to instructors. These results indicate that with clearly defined evaluation criteria, contemporary LLMs can serve as teaching assistants for specific tasks.

At the same time, the obtained results indicate that GenAI is far from being a universal or fully reliable grader. Performance dropped below that of human teachers in numerous situations, especially for open-ended, complex, or subjective assignments, particularly those requiring nuanced judgment, contextual understanding, or creativity. Characteristic minor errors, logical inconsistencies, or contextual mistakes, usually detected by a human teacher, may elude GenAI models. Similar concerns were described in another review article [73]. In one study, GPT-3.5 struggled with assessing code correctness and occasionally invented false corrections. Even GPT-4 had difficulties with children's visual logic tasks and advanced mathematics. Consistency was another recurrent issue: in some studies, the same model, using the same criteria, produced different grades across attempts. Phenomena such as GenAI hallucinations and overly generic feedback were also observed. These problems have been noticed in other review articles as well [74]. These limitations show that current LLMs cannot fully replace teachers, especially where in-depth analysis, creativity, or subjective evaluation is required.

It is important to emphasize that there is no uniform pattern of bias in GenAI grading. In some cases, models were slightly more lenient than teachers (e.g., ChatGPT more often awarded somewhat higher grades, especially for language accuracy), while in others, they were stricter (e.g., GPT-4 was harsher than humans when grading short answers). Some studies indicate that GenAI "avoids extreme grades," tending to cluster around the middle of the scale. This means that the behavior of the models depends on both the model version and the context. Regardless of these differences, about 12 analyzed studies showed comparability of GenAI and human grades without significant deviations. Frequently, feedback generated by GenAI was more extensive than that of teachers. Still, its quality was sometimes questioned: GenAI sometimes wrote too abstract or generic comments, not always matching the actual assessment.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 September 2025 doi:10.20944/preprints202509.1233.v1

The analysis also allows us to distinguish key factors affecting the quality of GenAI grading. One of the most important is prompt quality and the level of detail in the grading rubric. In many studies, prompt engineering and providing example answers or model criteria significantly increased the agreement between GenAI and teacher grades. Well-defined criteria (rubrics) are a condition for stable results. At the same time, more advanced prompting techniques, such as "chain-of-thought," yielded mixed results—they improved grading in some types of tasks but led to redundancy or repetitiveness in others. Another critical factor is the specific model version: although new models (e.g., GPT-4) usually perform better than earlier ones (e.g., GPT-3.5), this advantage is neither universal nor guaranteed across all tasks. How the capabilities of successive, newer models in grading student work are changing requires further exploration. Attention is also drawn to the language of the assessed work—AI performs best in English, with effectiveness declining for other languages, especially those less represented in the training data. GenAI generally works best for short, closed-ended answers where a specific response is expected, but loses precision on open-ended tasks or those requiring contextual understanding. The principle "garbage in, garbage out" is key here, which requires educators to be skilled in "prompt engineering" and aware of the limitations of GenAI.

It can be said, then, that across the reviewed studies, agreement with human grades was strongest when tasks were short, tightly structured, and written in English; when models received detailed rubrics or exemplar answers; and when newer versions such as GPT-4 were used through stable API workflows. As tasks became longer, more open-ended, or non-English, accuracy and consistency often declined—sometimes regardless of model generation—unless prompts contained rich, explicit criteria and a human remained in the loop.

The discussed findings also have broad didactic implications, in agreement with earlier reviews such as [75] and [76]. GenAI can reduce teachers' workload by taking over some of the routine grading, which in large student groups means faster feedback and greater process scalability. LLMs can generate detailed comments for each student and support learning personalization. Automation may also reduce certain forms of human bias and fatigue, improving fairness in large-scale assessments. GenAI also enables the development of new forms of assessment, such as generating quizzes or coding tasks tailored to the student's level. However, nearly all studies emphasize that GenAI should only support the teacher, not replace them. Oversight is necessary to catch errors, address ethical issues such as bias and transparency, and provide the contextual and empathetic judgment that current LLMs cannot replicate. This oversight is especially critical for high-stakes grading contexts and wherever subjective interpretation is required for the assessment.

5. Conclusions

Based on 42 empirical studies published between 2023 and 2025, this review does not support replacing teachers with GenAI in assessment. Under specific conditions—short, well-structured tasks, explicit rubrics or exemplar answers, newer models (e.g., GPT-4), and stable API-based pipelines—LLMs can achieve accuracy and internal consistency comparable to human raters and can scale timely feedback.

These gains are not general. The effectiveness of GenAI-assisted grading strongly depends on the type of task, the quality of prompts and rubrics, the subject domain, and the language used to interact with GenAI. Performance declines for complex, open-ended, or subjective assignments requiring nuanced judgment, and in less-represented languages; issues with between-run consistency, hallucinations, and generic or misaligned feedback persist. No uniform grading bias emerged across studies (sometimes more lenient, sometimes stricter, with a tendency in several cases to avoid extreme scores). Improvements over earlier model generations are measurable but uneven and task-dependent.

The review highlights that LLMs can support teachers by automating routine grading and generating rapid feedback at scale. Still, they cannot fully replace human judgment, especially in high-stakes or subjective assessments. GenAI-based scoring systems should be deployed as assistive tools, integrated within a hybrid, human-in-the-loop framework that maintains transparency, oversight, fairness, and adaptability to evolving educational needs. Human educators remain indispensable for validating grades and GenAI personalized feedback on students' progress.

Finally, this evidence base itself needs tightening. Future work should:

1. adopt common reporting standards and stronger sampling beyond convenience cohorts;
2. track model generations with preregistered protocols;
3. expand multilingual and modality-diverse evaluations; and
4. connect grading metrics to downstream learning outcomes and equity impacts.

Until these gaps narrow and results show more consistent reliability and quality of automated grading, GenAI can help make assessment faster and richer in feedback, but only there where clarity, precision, rubrics, and human judgment meet.

Funding: This work was supported by the Polish National Science Centre (grant no. 2025/09/X/HS6/00038).

Acknowledgments: Our thanks go to A.W. for invaluable remarks that improved this manuscript.

Conflicts of Interest: The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Jukiewicz, M. How generative artificial intelligence transforms teaching and influences student wellbeing in future education. *Frontiers in Education* **2025**, *Volume 10*. https://doi.org/10.3389/feduc.2025.1594572.
2. Gill, S.S.; Xu, M.; Patros, P.; Wu, H.; Kaur, R.; Kaur, K.; Fuller, S.; Singh, M.; Arora, P.; Parlikad, A.K.; et al. Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots. *Internet of Things and Cyber-Physical Systems* **2024**, *4*, 19–23. https://doi.org/10.1016/j.iotcps.2023.06.002.
3. Al Murshidi, G.; Shulgina, G.; Kapuza, A.; Costley, J. How understanding the limitations and risks of using ChatGPT can contribute to willingness to use. *Smart Learning Environments* **2024**, *11*, 36. https://doi.org/10.1186/s40561-024-00322-9.
4. Castillo, A.; Silva, G.; Arocutipa, J.; Berrios, H.; Marcos Rodriguez, M.; Reyes, G.; Lopez, H.; Teves, R.; Rivera, H.; Arias-Gonzáles, J. Effect of Chat GPT on the digitized learning process of university students. *Journal of Namibian Studies : History Politics Culture* **2023**, *33*, 1–15. https://doi.org/10.59670/jns.v33i.411.
5. Hung, J.; Chen, J. The Benefits, Risks and Regulation of Using ChatGPT in Chinese Academia: A Content Analysis. *Social Sciences* **2023**, *12*. https://doi.org/10.3390/socsci12070380.
6. Raile, P. The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social Sciences Communications* **2024**, *11*, 1–8.
7. De Duro, E.S.; Improta, R.; Stella, M. Introducing CounseLLMe: A dataset of simulated mental health dialogues for comparing LLMs like Haiku, LLaMAntino and ChatGPT against humans. *Emerging Trends in Drugs, Addictions, and Health* **2025**, p. 100170. https://doi.org/10.1016/j.etdah.2025.100170.
8. Maurya, R.K.; Montesinos, S.; Bogomaz, M.; DeDiego, A.C. Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. *Counselling and Psychotherapy Research* **2025**, *25*, e12759, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/capr.12759]. https://doi.org/10.1002/capr.12759.
9. Memarian, B.; Doleck, T. ChatGPT in education: Methods, potentials, and limitations. *Computers in Human Behavior: Artificial Humans* **2023**, *1*, 100022. https://doi.org/10.1016/j.chbah.2023.100022.
10. Diab Idris, M.; Feng, X.; Dyo, V. Revolutionizing Higher Education: Unleashing the Potential of Large Language Models for Strategic Transformation. *IEEE Access* **2024**, *12*, 67738–67757. https://doi.org/10.1109/ACCESS.2024.3400164.
11. Hadi Mogavi, R.; Deng, C.; Juho Kim, J.; Zhou, P.; D. Kwon, Y.; Hosny Saleh Metwally, A.; Tlili, A.; Bassanelli, S.; Bucchiarone, A.; Gujar, S.; et al. ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans* **2024**, *2*, 100027. https://doi.org/10.1016/j.chbah.2023.100027.
12. Davis, R.O.; Lee, Y.J. Prompt: ChatGPT, Create My Course, Please! *Education Sciences* **2024**, *14*. https://doi.org/10.3390/educsci14010024.
13. Onal, S.; Kulavuz-Onal, D. A Cross-Disciplinary Examination of the Instructional Uses of ChatGPT in Higher Education. *Journal of Educational Technology Systems* **2024**, *52*, 301–324. https://doi.org/10.1177/00472395231196532.

14. Li, M. The Impact of ChatGPT on Teaching and Learning in Higher Education: Challenges, Opportunities, and Future Scope. *Encyclopedia of Information Science and Technology, Sixth Edition* **2025**, pp. 1–20. https://doi.org/10.4018/978-1-6684-7366-5.ch079.
15. Lucas, H.C.; Upperman, J.S.; Robinson, J.R. A systematic review of large language models and their implications in medical education. *Medical Education* **2024**. https://doi.org/10.1111/medu.15402.
16. Jukiewicz, M. The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity* **2024**, *52*, 101522. https://doi.org/10.1016/j.tsc.2024.101522.
17. Kıyak, Y.S.; Emekli, E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. *Postgraduate medical journal* **2024**. https://doi.org/10.1093/postmj/qgae065.
18. Ali, M.M.; Wafik, H.A.; Mahbub, S.; Das, J. Gen Z and Generative AI: Shaping the Future of Learning and Creativity. *Cognizance Journal of Multidisciplinary Studies* **2024**, *4*, 1–18. https://doi.org/10.47760/cognizance.2024.v04i10.001.
19. Ahmed, A.R. Navigating the integration of generative artificial intelligence in higher education: Opportunities, challenges, and strategies for fostering ethical learning. *Advances in Biomedical and Health Sciences* **2025**, *4*. https://doi.org/10.4103/abhs.abhs_122_24.
20. Reimer, E.C. Examining the Role of Generative AI in Enhancing Social Work Education: An Analysis of Curriculum and Assessment Design. *Social Sciences* **2024**, *13*, 648. https://doi.org/10.3390/socsci13120648.
21. Tzirides, A.O.O.; Zapata, G.; Kastania, N.P.; Saini, A.K.; Castro, V.; Ismael, S.A.; You, Y.l.; dos Santos, T.A.; Searsmith, D.; O'Brien, C.; et al. Combining human and artificial intelligence for enhanced AI literacy in higher education. *Computers and Education Open* **2024**, *6*, 100184.
22. Wu, D.; Chen, M.; Chen, X.; Liu, X. Analyzing K-12 AI education: A large language model study of classroom instruction on learning theories, pedagogy, tools, and AI literacy. *Computers and Education: Artificial Intelligence* **2024**, *7*, 100295.
23. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*. https://doi.org/10.1136/bmj.n71.
24. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*. https://doi.org/10.1136/bmj.n160.
25. Rethlefsen, M.L.; Kirtley, S.; Waffenschmidt, S.; Ayala, A.P.; Moher, D.; Page, M.J.; Koffel, J.B.; Blunt, H.; Brigham, T.; Chang, S.; et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Systematic Reviews* **2021**, *10*, 39. https://doi.org/10.1186/s13643-020-01542-z.
26. Hong, Q.N.; Fàbregues, S.; Bartlett, G.; Boardman, F.; Cargo, M.; Dagenais, P.; Gagnon, M.P.; Griffiths, F.; Nicolau, B.; O'Cathain, A.; et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for information* **2018**, *34*, 285–291.
27. Grévisse, C. LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education* **2024**, *24*, 1060.
28. Jiang, Z.; Xu, Z.; Pan, Z.; He, J.; Xie, K. Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages* **2023**, *8*, 247.
29. Li, Z.; Lloyd, S.; Beckman, M.; Passonneau, R.J. Answer-state recurrent relational network (AsRRN) for constructed response assessment and feedback grouping. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 3879–3891.
30. Hackl, V.; Müller, A.E.; Granitzer, M.; Sailer, M. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education* **2023**, *8*, 1272229. https://doi.org/10.3389/feduc.2023.1272229.
31. Azaiz, I.; Deckarm, O.; Strickroth, S. Ai-enhanced auto-correction of programming exercises: How effective is gpt-3.5? *arXiv preprint arXiv:2311.10737* **2023**.
32. Acar, S.; Dumas, D.; Organisciak, P.; Berthiaume, K. Measuring original thinking in elementary school: Development and validation of a computational psychometric approach. *Journal of Educational Psychology* **2024**, *116*, 953–981. https://doi.org/10.1037/edu0000844.
33. Schleifer, A.G.; Klebanov, B.B.; Ariely, M.; Alexandron, G. Transformer-based Hebrew NLP models for short answer scoring in biology. In Proceedings of the Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 2023, pp. 550–555.

34. Li, J.; Huang, J.; Wu, W.; Whipple, P.B. Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications* **2024**, *11*, 1268. https://doi.org/10.1057/s41599-024-03755-2.
35. Chang, L.H.; Ginter, F. Automatic short answer grading for finnish with chatgpt. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 23173–23181.
36. Dong, D. Tapping into the pedagogical potential of InfinigoChatIC: Evidence from iWrite scoring and comments and Lu & Ai's linguistic complexity analyzer. *Arab World English Journal (AWEJ) Special Issue on ChatGPT* **2024**.
37. Thomas, D.R.; Lin, J.; Bhushan, S.; Abboud, R.; Gatz, E.; Gupta, S.; Koedinger, K.R. Learning and ai evaluation of tutors responding to students engaging in negative self-talk. In Proceedings of the Proceedings of the Eleventh ACM Conference on Learning@ Scale, 2024, pp. 481–485.
38. Gao, R.; Guo, X.; Li, X.; Narayanan, A.B.L.; Thomas, N.; Srinivasa, A.R. Towards Scalable Automated Grading: Leveraging Large Language Models for Conceptual Question Evaluation in Engineering. In Proceedings of the Proceedings of Machine Learning Research. FM-EduAssess at NeurIPS 2024 Workshop, 2024.
39. Jackaria, P.M.; Hajan, B.H.; Mastul, A.R.H. A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning, Teaching and Educational Research* **2024**, *23*.
40. Jin, H.; Kim, Y.; Park, Y.S.; Tilekbay, B.; Son, J.; Kim, J. Using large language models to diagnose math problem-solving skills at scale. In Proceedings of the Proceedings of the eleventh ACM conference on learning@ scale, 2024, pp. 471–475.
41. Li, J.; Jangamreddy, N.K.; Hisamoto, R.; Bhansali, R.; Dyda, A.; Zaphir, L.; Glencross, M. AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology* **2024**, *40*, 56–72.
42. Cohn, C.; Hutchins, N.; Le, T.; Biswas, G. A Chain-of-Thought Prompting Approach with LLMs for Evaluating Students' Formative Assessment Responses in Science. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 23182–23190. https://doi.org/10.1609/aaai.v38i21.30364.
43. Zhang, M.; Johnson, M.; Ruan, C. Investigating Sampling Impacts on an LLM-Based AI Scoring Approach: Prediction Accuracy and Fairness. *Journal of Measurement and Evaluation in Education and Psychology* **2024**, *15*, 348–360. https://doi.org/10.21031/epod.1561580.
44. Shamim, M.S.; Zaidi, S.J.A.; Rehman, A. The revival of essay-type questions in medical education: harnessing artificial intelligence and machine learning. *Journal of the College of Physicians and Surgeons Pakistan* **2024**, *34*, 595–599. https://doi.org/10.29271/jcpsp.2024.05.595.
45. Sreedhar, R.; Chang, L.; Gangopadhyaya, A.; Shiels, P.W.; Loza, J.; Chi, E.; Gabel, E.; Park, Y.S. Comparing Scoring Consistency of Large Language Models with Faculty for Formative Assessments in Medical Education. *Journal of General Internal Medicine* **2025**, *40*, 127–134.
46. Lee, U.; Kim, Y.; Lee, S.; Park, J.; Mun, J.; Lee, E.; Kim, H.; Lim, C.; Yoo, Y.J. Can we use GPT-4 as a mathematics evaluator in education?: Exploring the efficacy and limitation of LLM-based automatic assessment system for open-ended mathematics question. *International Journal of Artificial Intelligence in Education* **2024**, pp. 1–37.
47. Grandel, S.; Schmidt, D.C.; Leach, K. Applying Large Language Models to Enhance the Assessment of Parallel Functional Programming Assignments. In Proceedings of the Proceedings of the 1st International Workshop on Large Language Models for Code, New York, NY, USA, 2024; LLM4Code '24, p. 102–110. https://doi.org/10.1145/3643795.3648375.
48. Martin, P.P.; Graulich, N. Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence* **2024**, *7*, 100265.
49. Menezes, T.; Egherman, L.; Garg, N. AI-grading standup updates to improve project-based learning outcomes. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*; 2024; pp. 17–23.
50. Zhang, C.; Hofmann, F.; Plößl, L.; Gläser-Zikuda, M. Classification of reflective writing: A comparative analysis with shallow machine learning and pre-trained language models. *Education and Information Technologies* **2024**, *29*, 21593–21619.

51. Zhuang, X.; Wu, H.; Shen, X.; Yu, P.; Yi, G.; Chen, X.; Hu, T.; Chen, Y.; Ren, Y.; Zhang, Y.; et al. TOREE: Evaluating topic relevance of student essays for Chinese primary and middle school education. In Proceedings of the Findings of the Association for Computational Linguistics ACL 2024, 2024, pp. 5749–5765.
52. Sonkar, S.; Ni, K.; Tran Lu, L.; Kincaid, K.; Hutchinson, J.S.; Baraniuk, R.G. Automated long answer grading with RiceChem dataset. In Proceedings of the International Conference on Artificial Intelligence in Education. Springer, 2024, pp. 163–176.
53. Huang, C.W.; Coleman, M.; Gachago, D.; Van Belle, J.P. Using ChatGPT to Encourage Critical AI Literacy Skills and for Assessment in Higher Education. In Proceedings of the ICT Education; Van Rensburg, H.E.; Snyman, D.P.; Drevin, L.; Drevin, G.R., Eds., Cham, 2024; pp. 105–118. https://doi.org/10.1007/978-3-031-48536-7_8.
54. Lu, Q.; Yao, Y.; Xiao, L.; Yuan, M.; Wang, J.; Zhu, X. Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education* **2024**, *49*, 616–633.
55. Agostini, D.; Picasso, F.; Ballardini, H.; et al. Large Language Models for the Assessment of Students' Authentic Tasks: A Replication Study in Higher Education. In Proceedings of the CEUR Workshop Proceedings, 2024, Vol. 3879.
56. Tapia-Mandiola, S.; Araya, R. From play to understanding: Large language models in logic and spatial reasoning coloring activities for children. *AI* **2024**, *5*, 1870–1892.
57. Lin, N.; Wu, H.; Zheng, W.; Liao, X.; Jiang, S.; Yang, A.; Xiao, L. IndoCL: Benchmarking Indonesian Language Development Assessment. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 4873–4885.
58. Quah, B.; Zheng, L.; Sng, T.J.H.; Yong, C.W.; Islam, I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education* **2024**, *24*, 962. https://doi.org/10.1186/s12909-024-05881-6.
59. Henkel, O.; Hills, L.; Boxer, A.; Roberts, B.; Levonian, Z. 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 Proceedings of the Eleventh ACM Conference on Learning@ Scale, 2024, pp. 300–304.
60. Latif, E.; Zhai, X. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence* **2024**, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210.
61. Xiao, X.; Li, Y.; He, X.; Fang, J.; Yan, Z.; Xie, C. An assessment framework of higher-order thinking skills based on fine-tuned large language models. *Expert Systems with Applications* **2025**, *272*, 126531.
62. Fokides, E.; Peristeraki, E. Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek. *Education and Information Technologies* **2025**, *30*, 2577–2621.
63. Flodén, J. Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British educational research journal* **2025**, *51*, 201–224.
64. Gonzalez-Maldonado, D.; Liu, J.; Franklin, D. Evaluating GPT for use in K-12 Block Based CS Instruction Using a Transpiler and Prompt Engineering. In Proceedings of the Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1, New York, NY, USA, 2025; SIGCSETS 2025, p. 388–394. https://doi.org/10.1145/3641554.3701910.
65. Qiu, H.; White, B.; Ding, A.; Costa, R.; Hachem, A.; Ding, W.; Chen, P. Stella: A structured grading system using llms with rag. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData). IEEE, 2024, pp. 8154–8163.
66. Gandolfi, A. GPT-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *International Journal of Artificial Intelligence in Education* **2025**, *35*, 367–397.
67. Yavuz, F.; Çelik, Ö.; Yavaş Çelik, G. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology* **2025**, *56*, 150–166.
68. Manning, J.; Baldwin, J.; Powell, N. Human versus machine: The effectiveness of ChatGPT in automated essay scoring. *Innovations in Education and Teaching International* **2025**, pp. 1–14. https://doi.org/10.1080/14703297.2025.2469089.
69. Gurin Schleifer, A.; Beigman Klebanov, B.; Ariely, M.; Alexandron, G. Transformer-based Hebrew NLP models for Short Answer Scoring in Biology. In Proceedings of the Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023); Kochmar, E.; Burstein, J.; Horbach, A.; Laarmann-Quante, R.; Madnani, N.; Tack, A.; Yaneva, V.; Yuan, Z.; Zesch, T., Eds., Toronto, Canada, 2023; pp. 550–555. https://doi.org/10.18653/v1/2023.bea-1.46.

70. Thomas, D.R.; Lin, J.; Bhushan, S.; Abboud, R.; Gatz, E.; Gupta, S.; Koedinger, K.R. Learning and AI Evaluation of Tutors Responding to Students Engaging in Negative Self-Talk. In Proceedings of the Proceedings of the Eleventh ACM Conference on Learning @ Scale, New York, NY, USA, 2024; L@S '24, p. 481–485. https://doi.org/10.1145/3657604.3664700.
71. Lu, Q.; Yao, Y.; Xiao, L.; Yuan, M.; Wang, J.; Zhu, X. Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education* **2024**, *49*, 616–633. https://doi.org/10.1080/02602938.2024.2301722.
72. Zhu, M.; Zhang, K. Artificial Intelligence for Computer Science Education in Higher Education: A Systematic Review of Empirical Research Published in 2003–2023: M. Zhu, K. Zhang. *Technology, Knowledge and Learning* **2025**, pp. 1–25.
73. Pereira, A.F.; Mello, R.F. A Systematic Literature Review on Large Language Models Applications in Computer Programming Teaching Evaluation Process. *IEEE Access* **2025**.
74. Albadarin, Y.; Saqr, M.; Pope, N.; Tukiainen, M. A systematic literature review of empirical research on ChatGPT in education. *Discover Education* **2024**, *3*, 60.
75. Lo, C.K.; Yu, P.L.H.; Xu, S.; Ng, D.T.K.; Jong, M.S.y. Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. *Smart Learning Environments* **2024**, *11*, 50.
76. Dikilitas, K.; Klippen, M.I.F.; Keles, S. A systematic rapid review of empirical research on students' use of ChatGPT in higher education. *Nordic Journal of Systematic Reviews in Education* **2024**, *2*. https://doi.org/10.23865/njsre.v2.6227.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.