

EduGrader: The AI-powered grading assistant [★]

Svetozar Iković¹ and Jelena Graovac¹

University of Belgrade, Faculty of Mathematics

mi241072@alas.matf.bg.ac.rs

jelena.graovac@matf.bg.ac.rs (corresponding author)

Abstract. The rapid development of artificial intelligence (AI) and large language models (LLMs) is reshaping assessment practices in higher education, particularly in grading open-ended questions. Traditional grading is time-consuming, prone to inconsistency, and difficult to scale for large classes. LLMs offer an opportunity to improve efficiency, consistency, and feedback frequency by leveraging their ability to understand and reason about natural language. This paper introduces EduGrader, an open-source, multi-model grading platform integrating several LLMs, including GPT-4o, GPT-4o-mini, DeepSeek-Chat, GPT-5.1, and DeepSeek-Reasoner. EduGrader supports three grading strictness levels (lenient, neutral, strict) and generates numerical scores with concise explanations highlighting correct reasoning, missing concepts, and errors. The system operates in two modes: a provided-reference-answer grading mode guided by instructor solutions and a generated-reference-answer grading mode that automatically creates reference answers from course materials. All experiments are conducted in Serbian, a comparatively low-resource language underrepresented in mainstream AI training data, extending AI-assisted grading research beyond the dominant English-language context. EduGrader is evaluated on 686 real student responses from six university courses across two institutions, including free-text explanations and code-based answers. In its best-performing configuration, the system achieves a Pearson correlation coefficient of 0.90 with human graders while reducing grading workload and improving reliability, though it is intended to assist rather than replace instructors.

Keywords: Automated Grading, Education, Large Language Models, Artificial Intelligence.

1. Introduction

Grading is rarely the most exciting part of teaching, yet it often quietly shapes how students learn. When feedback is timely and specific, students can recognize and correct misunderstandings before they take root. When it is delayed, vague, or inconsistent, assessment becomes a formality, and learning slows down. This challenge is particularly visible in large university courses, where instructors and teaching assistants must evaluate hundreds of open-ended responses within tight deadlines. Unlike multiple-choice questions,

[★] This is an extended version of a conference paper, presented at the 2025 AI Conference, Serbian Academy of Sciences and Arts, Belgrade, October 8-9, 2025

short open-ended explanations and mini-essays cannot be checked through pattern matching alone. They demand interpretation: identifying partial reasoning, distinguishing minor wording issues from conceptual gaps, and deciding how much an omission should affect the score. As a result, open-ended grading remains both expensive and difficult to scale. It is also prone to inconsistency. Even with detailed grading guidelines, two graders may interpret the same response differently, especially when answers are ambiguous, hastily written, or structurally unclear [11, 26, 4]. For instructors, this workload restricts how often richer, explanation-based questions can be used. For students, it often means feedback arrives too late to be useful, or never arrives at all. Earlier generations of automated assessment systems worked well only when responses were tightly constrained, such as multiple-choice tests, fill-in-the-blank questions, or short answers with predictable phrasing [25, 3, 22]. Advances in natural language processing (NLP) expanded these capabilities, but many approaches still relied on heavy feature engineering, limited domains, or large labeled datasets [23, 6, 2]. Modern large language models (LLMs), however, mark a turning point: they can interpret free-form text and reason about meaning directly. This development has made it feasible to design tools that grade open-ended responses, justify their scores in natural language, and highlight what is missing, not just what is wrong [13, 33].

Still, the use of LLMs in assessment raises important practical and pedagogical questions. How consistent are their grades across subjects and question types? Can instructors calibrate strictness to match specific course expectations? What happens when no complete answer key is available? And, crucially, how can we preserve human accountability for final grades while taking advantage of automation's efficiency?

This paper introduces EduGrader, an AI-assisted grading platform built to support open-ended assessment in higher education under instructor supervision. EduGrader integrates multiple large language models, GPT-4o¹, GPT-4o-mini², DeepSeek-Chat³, and two reasoning-oriented models, GPT-5.1⁴ and DeepSeek-Reasoner⁵, within a unified, transparent workflow that prioritizes instructor control. The system includes three grading strictness profiles, lenient, neutral, and strict, reflecting real-world grading practices, where some courses reward partial reasoning while others demand precise terminology and complete solutions.

EduGrader offers two complementary grading modes:

- **Provided-reference-answer grading (PRA)**, in which student responses are evaluated against instructor-provided reference answers (optionally supplemented with grading notes).
- **Generated-reference-answer grading (GRA)**, where the system first constructs a reference answer from course materials, such as lecture notes, course handouts, and textbooks, and then evaluates student responses against that generated standard.

In both modes, the output includes not only a numeric score (0–10) but also a clear explanation that identifies strengths, missing ideas, and reasoning gaps. The aim is to

¹ <https://platform.openai.com/docs/models/gpt-4o>

² <https://platform.openai.com/docs/models/gpt-4o-mini>

³ <https://api-docs.deepseek.com/>

⁴ <https://platform.openai.com/docs/models/gpt-5>

⁵ https://api-docs.deepseek.com/guides/reasoning_model

provide feedback that is actionable for students and auditable for instructors, rather than a “black box” grade.

We evaluate EduGrader using 686 authentic student responses collected across two universities and six courses, spanning different response types: short factual answers, open-ended explanatory text, and code-like expressions. The results show that PRA yields more stable outcomes than GRA, and that reasoning-oriented models align most closely with human graders. In its best configuration (GPT-5.1, neutral mode), EduGrader achieves a Pearson correlation coefficient (PCC) of 0.90 with human grading.

Finally, EduGrader is designed as a grading assistant, not a replacement for instructor judgment. It reduces repetitive workload and enhances consistency, but human oversight remains essential for ambiguous questions, discipline-specific expectations, and borderline cases. For this reason, EduGrader follows a human-in-the-loop workflow and is released as open source, promoting transparency, reproducibility, and adaptability across institutions.

2. Related Work

Research on the use of artificial intelligence (AI) in education reveals a clear evolution from early rule-based automation toward adaptive systems capable of interpreting and generating complex, human-like responses. Foundational definitions describe AI as computational systems designed to perform cognitive functions such as learning, reasoning, and problem solving [1]. Within educational settings, AI encompasses a broad set of technologies that together enable more personalized and adaptive learning environments [32]. Numerous studies and systematic reviews emphasize AI’s potential to improve learning outcomes, broaden access to education, and support more efficient use of instructional resources [31, 34]. Recent advances in large LLMs have further expanded these possibilities. Unlike earlier NLP-based approaches that relied on handcrafted features or domain-specific training, LLMs can directly process free-form text and reason about meaning at a higher level of abstraction. As a result, they have been increasingly explored for learner-centered instruction, automated content generation, and novel forms of assessment across a wide range of educational domains [34].

Empirical studies examining LLM-assisted grading report mixed but informative results. Kooli and Yusuf [14] conducted an empirical investigation of ChatGPT-assisted grading using a dataset of 25 student papers from a cohort of approximately 100 second-year university students in a social science course. In their experimental setup, both human graders and the ChatGPT system independently evaluated the assignments, enabling a direct comparison between automated and human grading outcomes. The study reported a moderate correlation between AI-generated and human grades (correlation coefficient of approximately 0.45), suggesting that while the model can approximate overall grading patterns, substantial discrepancies remain in individual evaluations. The authors further highlight that ChatGPT can provide useful feedback and accelerate the grading process, but that its automated nature may limit contextual understanding of nuanced student arguments. They also emphasize potential risks such as the amplification of educational inequalities and the need for institutional frameworks ensuring transparency, monitoring, and bias mitigation when deploying AI-based assessment tools.

Similarly, Lundgren [18] examined the performance of GPT-4 in grading higher-education essays, using a dataset of 60 master's-level political science essays previously graded by university instructors. The study compared the grades suggested by GPT-4 with those assigned by human graders and found that the average scores produced by the model were broadly aligned with instructor evaluations. However, the analysis revealed lower interrater reliability between AI and human graders, indicating that while the model may approximate aggregate grading trends, it struggles to consistently replicate fine-grained human judgment. Additionally, the model demonstrated a risk-averse grading pattern and tended to rely on general essay characteristics such as language quality rather than adapting fully to detailed grading rubrics. Experiments with modified prompts and grading instructions showed only limited improvements, suggesting that current LLM-based grading systems remain relatively insensitive to subtle variations in evaluation criteria.

In contrast to essay-based assessment, Jukiewicz [12] investigated the application of ChatGPT for grading programming assignments in a 15-week Python programming course involving 67 undergraduate cognitive science students. Across nine programming assignments, both the instructor and the ChatGPT system independently evaluated student submissions, enabling a comparison of grading consistency and feedback quality. The results showed a strong positive correlation between ChatGPT-generated grades and instructor evaluations, although the AI system tended to assign slightly stricter grades. The study also reported excellent repeatability of the model's grading results across multiple evaluation runs, with only negligible variations between iterations. Beyond numerical scores, ChatGPT was able to assess code correctness, enforce coding standards such as adherence to PEP8 guidelines, and generate detailed feedback within seconds, significantly reducing the time required for manual grading. Nevertheless, the author also identifies several limitations, including the potential for hallucinated errors, API usage costs, and occasional grading inconsistencies that require teacher verification. As a result, the study concludes that ChatGPT can serve as a valuable complementary tool for automated code assessment and feedback generation, but that human oversight remains necessary to ensure reliability and fairness.

These findings support broader evidence indicating that LLM grading reliability is task-dependent: agreement with human graders tends to be higher in structured-response domains, such as definitions, formulas, or code fragments, while performance decreases in subjective or open-ended writing tasks [28, 29]. For instance, [21] evaluated GPT-3 on open-ended responses from industry professionals and observed strong correlations with expert ratings, contingent upon clearly defined criteria and contextual grounding. Beyond numerical grading, several studies highlight AI's broader role in supporting assessment processes. AI-based systems have been shown to generate adaptive learning activities, deliver personalized feedback, and support both formative and summative assessment based on individual learner characteristics [8, 17]. Intelligent Tutoring Systems and related approaches aim to approximate one-to-one instructional support while increasing consistency and scalability in evaluation practices [7, 9]. At the same time, critical perspectives caution against the "black-boxing" of accountability in automated assessment, where decision-making processes become opaque and responsibility is blurred [27]. As a result, transparency, explainability, data privacy, and feedback quality remain central concerns in conceptual analyses of AI-supported grading [15].

Across this literature, a consistent conclusion emerges: LLMs are best understood as assistive tools that automate routine aspects of grading and feedback, while human educators remain essential for pedagogical judgment, contextual interpretation, and handling ambiguous or borderline cases [16, 32]. Consequently, approaches that combine automation with Human-in-the-Loop supervision, configurable grading behavior, and interpretable feedback align most closely with current recommendations in the literature. These considerations motivate the design of systems that prioritize instructor control, transparency, and adaptability, principles that directly inform the design of the EduGrader platform introduced in this paper.

3. System evaluation and methodology

EduGrader is designed as an AI-assisted grading platform that integrates LLMs into an instructor-supervised assessment workflow. The primary goal of the system is to improve grading efficiency and consistency while preserving pedagogical transparency and human accountability. This section describes the evaluation methodology used to assess the reliability of EduGrader's grading outputs, the datasets on which it was tested, and the model configurations examined in the experiments.

3.1. Metrics and evaluation

To evaluate the alignment between AI-generated grades and human judgment, EduGrader's outputs were compared with grades assigned by two university instructors. Each instructor graded a different subset of the 686 responses (192 and 485 responses, respectively). This setup enabled us to compare AI performance against multiple human evaluators rather than a single grading standard. Each response was graded according to the standard grading criteria used in the course. Grades were assigned on a 0–10 scale based on the degree of conceptual correctness and completeness of the answer. Instructors considered whether the student correctly identified the key concepts, whether important elements were missing, and whether the reasoning was partially correct but incomplete. Answers demonstrating correct core ideas received partial credit even when terminology, structure, or some details were missing, while fundamentally incorrect answers received low or zero scores. For each response, the score assigned by the responsible course instructor was treated as the reference grade and compared with the AI-generated score using the evaluation metrics described below.

We used four statistical indicators that capture complementary aspects of grading reliability: the Pearson Correlation Coefficient (PCC), quadratic weighted Cohen's kappa, the Mean Ratio, and the RMSE-to-standard-deviation (RMSE/STDDEV) ratio.

The PCC [24] measures the strength of the linear association between human- and AI-assigned grades. The coefficient ranges from -1 to +1, where +1 indicates a perfect positive linear relationship, 0 indicates no linear relationship, and -1 indicates a perfect negative linear relationship. Before conducting the analysis, we examined the distribution of grading differences using the Shapiro–Wilk test and visual inspection of Q–Q plots. The results did not indicate substantial departures from normality ($p \geq .05$), supporting the use of a parametric correlation measure. Importantly, PCC evaluates linear association rather than absolute agreement. For this reason, PCC is complemented by weighted

kappa as an agreement-oriented measure and by error-based metrics that capture absolute deviations between AI and human scores.

To complement PCC, we also report quadratic weighted Cohen's kappa, which is suitable for ordinal grades. Unlike PCC, which measures linear association, weighted kappa measures agreement while accounting for the ordered nature of the 0–10 grading scale and penalizing larger disagreements more strongly than smaller ones.

The Mean Ratio summarizes overall grading tendencies. Differences between AI and human means highlight systematic grading leniency or strictness but do not reflect alignment in score variation.

The Root Mean Square Error (RMSE) [10] captures the average size of the differences between human and AI grades. A smaller RMSE indicates stronger agreement between human and AI grades. The RMSE/STDDEV [5] ratio measures how well AI-assigned grades agree with human-assigned grades, relative to the natural variability in the human grades. It shows whether the AI's deviations from human judgments are small compared to how much human grades vary overall. An RMSE/STDDEV ratio close to 0 indicates very high agreement between AI and human grades, around 1 means AI deviations are comparable to typical human grading variability, and greater than 1 suggests AI grades differ from human grades more than humans differ among themselves.

3.2. Data collection

The empirical evaluation of EduGrader was conducted on authentic theoretical exam responses collected across several examination periods at two universities in Serbia.

The resulting dataset spans **six university courses** and covers multiple academic levels, including undergraduate and master studies. It includes a mixture of conceptual questions and programming-related tasks, providing a diverse and heterogeneous benchmark for evaluating automated grading systems. The courses included in the dataset are:

- Fundamentals of programming in C - part 2 (Prog 2): First-year undergraduate course, Faculty of Mathematics, University of Belgrade
- Technical and scientific writings (TSW): First-year undergraduate course, Faculty of Mathematics, University of Belgrade
- Web programming (WP): Fourth-year undergraduate course, Faculty of Mathematics, University of Belgrade.
- Digital data storage (DDS): Master-level course, University of Arts
- HTML5 and JavaScript for video games (HTML5): Master-level course, University of Arts
- Data Mining (DM): Third-year undergraduate course, Faculty of Mathematics, University of Belgrade.

Each dataset includes student responses, instructor-assigned grades, reference answers (when available), instructor notes, and relevant course materials used during exam preparation. In total, the dataset contains **686 student responses**, providing a realistic benchmark for evaluating grading consistency across disciplines, academic levels, and question formats. Depending on the course, responses include **short free-text explanations**, **long free-text explanations**, and **code-based answers**. Table 1 summarizes the dataset structure, including the number of responses and response types for each course.

To better characterize the evaluation data, we additionally report the distribution of response types and the average length of answers. These statistics provide insight into the variability of response complexity and help contextualize the difficulty of automated grading across different question formats. In addition to the distribution of response types, we analyze the typical length of answers for both instructor-provided reference solutions, automatically generated reference answers, and student responses. This information is relevant because answer length can influence the difficulty of automated grading, particularly for long explanatory responses.

Table 2 also shows that generated reference answers are often much longer than instructor-provided answers, especially for questions where short answers are expected. In several courses, instructors expect concise responses (e.g., a single expression or brief statement), while LLM-generated answers typically include additional explanations describing why the answer is correct. As a result, generated answers contain more contextual information than is strictly required by the original exam question.

Table 1. Dataset structure across courses (number of student submissions and responses by question type)

Course	Responses	Short free-text	Long free-text	Code-based
Fundamentals of Programming in C – Part 2 (Prog 2)	480	380	0	100
Technical and Scientific Writing (TSW)	40	8	0	32
Web Programming (WP)	50	26	0	24
Digital Data Storage (DDS)	10	10	0	0
HTML5 and JavaScript for Video Games (HTML5)	10	4	0	6
Data Mining (DM)	96	0	96	0
Total	686	428 (62.4%)	96 (14%)	162 (23.6%)

Table 2. Average response length by question type and answer source

Question type	Instructor answer		Generated answer		Student answer	
	Avg words	Avg chars	Avg words	Avg chars	Avg words	Avg chars
Short free-text	20.9	130.5	201.7	1341.9	24.7	143.2
Long free-text	411.7	2867.7	479.9	3483.5	289.6	1889.3
Code-based	22.2	137.4	162.7	1144.3	15.9	102.5

3.3. Models and configurations

EduGrader employs a multi-model grading ensemble that includes both general-purpose and reasoning-optimized large language models:

- Standard models: GPT-4o, GPT-4o-mini, DeepSeek-Chat.
- Reasoning models: GPT-5.1, DeepSeek-Reasoner.

Each model was evaluated under three grading strictness configurations: lenient, neutral, and strict. These configurations were implemented through carefully designed prompts

that simulate different grading philosophies commonly observed in academic practice. For example, the strict configuration emphasizes precision and completeness (e.g., “You are a strict professor at a high-level academic institution and must rigorously evaluate student responses”), while the lenient configuration rewards partial correctness and conceptual understanding. This design reflects findings from prior studies showing that human graders themselves vary significantly in strictness [12].

In addition, EduGrader supports two complementary grading modes:

- PRA, where student responses are evaluated against instructor-provided correct answers and grading notes.
- GRA, where the model first generates a reference answer using provided course materials and then evaluates the student response relative to that generated standard.

This dual-mode design enables evaluation both in scenarios with well-defined answer keys and in situations where instructors do not provide explicit reference solutions.

All experiments are conducted in Serbian, including exam questions, student responses, and reference answers. As Serbian is a comparatively low-resource language underrepresented in mainstream AI training data, this study extends AI-assisted grading research beyond the dominant English-language context.

3.4. System architecture and grading workflow

EduGrader is a web-based application for automated grading of open-ended student exam answers using LLMs. It is implemented as a modular platform organized into three layers (Figure 1):

- **Frontend layer.** A single-page application built with React 19 and TypeScript, served via Nginx. It provides interactive views for instructors and students, including CSV file upload, model and parameter selection, and results display. Through this interface, instructors upload student answers (e.g., CSV files), select the desired LLM, grading mode (PRA or GRA), and strictness level, after which the system generates grade recommendations and explanations for each response. Authentication is handled via JSON Web Tokens (JWT) with role-based access control.
- **Backend layer.** A REST API built with Python 3, Django 5.1, and Django REST Framework. It is responsible for prompt generation, model execution, output parsing, and grading configuration management.
- **Database layer.** Managed through the Django ORM, with entities for courses, tests, questions, student answers, and grades. It stores grades, feedback, prompts, and meta-data to support reproducibility and different types of analysis.

This layered design allows grading logic and model behavior to evolve independently of the user interface, while preserving traceability of all grading decisions.

LLM access and integration LLMs are accessed exclusively through provider-specific cloud APIs. No models are hosted locally. The system implements a Factory design pattern with a common abstract interface (*LLMIntegration*) that defines a unified *prompt()* method accepting the input text, model identifier, and system instructions. Two concrete implementations are provided:

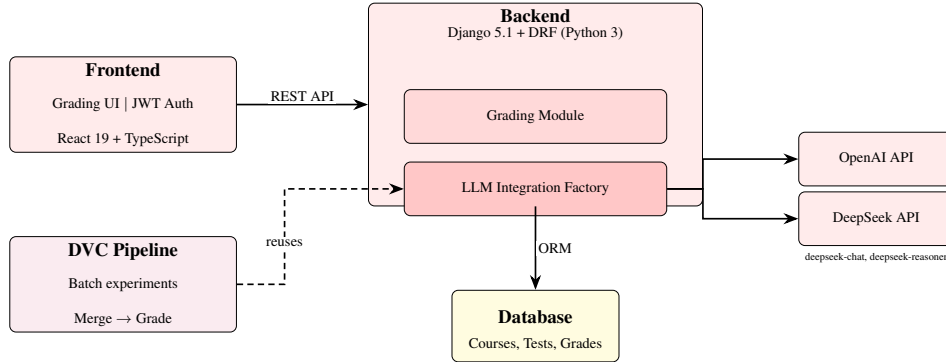


Fig. 1. EduGrader system architecture. The frontend communicates with the backend via REST API. The backend accesses LLMs through a unified integration layer (Factory pattern) that abstracts provider-specific APIs. The DVC pipeline reuses the same integration layer for batch experiments.

- **OpenAI** – using the official OpenAI Python SDK to call the Chat Completions API. Supported models: *gpt-4o-mini-2024-07-18*, *gpt-4o-2024-08-06*, *gpt-5.1-2025-11-13*.
- **DeepSeek** – using the OpenAI-compatible API endpoint via the OpenAI SDK. Supported models: *deepseek-chat* (V3), *deepseek-reasoner* (R1).

This abstraction allows new LLM providers to be added by implementing a single class without modifying the grading logic. API keys are stored as environment variables and are never embedded in the source code.

4. Results

A series of tests were conducted to evaluate the performance of the proposed EduGrader system across multiple large language models (LLMs) and grading configurations. Table 3 and Table 5 summarize the comparative results across four key performance indicators: PCC as a measure of linear association with human grades, quadratic weighted Cohen’s kappa as a measure of ordinal agreement, the Mean Ratio (AI mean grade / human mean grade) as an indicator of grading bias (leniency or strictness), and the RMSE/STDDEV ratio as a measure of relative deviation from human grading variability. Table 3 summarises results for the PRA mode, while Table 5 reports corresponding results for the GRA mode. All values are aggregated averages computed across the full dataset of exam questions and grading instances.

4.1. Model performance and temperature analysis

Table 3 reports the evaluation results across different models and grading strictness levels (lenient, neutral, and strict). For all models except *DeepSeek-Reasoner*, experiments were conducted with temperatures ranging from 0 to 0.5. The DeepSeek-Reasoner model does not support temperature sampling; therefore, only deterministic inference was evaluated for this model.

Table 3. Performance metrics for PRA across models and grading settings. For models supporting sampling, temperatures from 0.0 to 0.5 were evaluated; DeepSeek-Reasoner was evaluated only in deterministic mode. Best results for each metric and grading setting are highlighted in bold.

Model	Temp	PCC			Weighted Kappa			Mean Ratio			RMSE/STDDEV		
		lenient	neutral	strict	lenient	neutral	strict	lenient	neutral	strict	lenient	neutral	strict
GPT-4o-mini	0.0	0.76	0.83	0.77	0.69	0.82	0.68	1.22	1.00	0.75	0.75	0.55	0.76
GPT-4o-mini	0.1	0.77	0.84	0.78	0.72	0.82	0.68	1.19	0.99	0.73	0.71	0.55	0.76
GPT-4o-mini	0.2	0.79	0.83	0.77	0.75	0.81	0.66	1.16	0.99	0.72	0.68	0.55	0.78
GPT-4o-mini	0.3	0.79	0.83	0.77	0.74	0.82	0.67	1.16	0.98	0.72	0.67	0.55	0.78
GPT-4o-mini	0.4	0.78	0.83	0.78	0.73	0.82	0.68	1.19	0.99	0.73	0.70	0.55	0.76
GPT-4o-mini	0.5	0.77	0.84	0.77	0.70	0.82	0.68	1.22	1.00	0.75	0.74	0.55	0.75
GPT-4o	0.0	0.72	0.88	0.81	0.64	0.88	0.75	1.30	1.01	0.76	0.86	0.47	0.73
GPT-4o	0.1	0.73	0.90	0.84	0.65	0.90	0.78	1.30	1.00	0.78	0.85	0.45	0.67
GPT-4o	0.2	0.76	0.89	0.82	0.68	0.89	0.77	1.28	0.99	0.78	0.81	0.47	0.69
GPT-4o	0.3	0.74	0.89	0.83	0.67	0.89	0.78	1.28	0.99	0.77	0.83	0.46	0.68
GPT-4o	0.4	0.73	0.89	0.83	0.65	0.89	0.77	1.30	1.00	0.77	0.85	0.46	0.69
GPT-4o	0.5	0.71	0.88	0.82	0.62	0.88	0.76	1.31	1.01	0.77	0.89	0.48	0.71
GPT-5.1	0.0	0.88	0.90	0.85	0.86	0.90	0.82	1.12	0.97	0.83	0.52	0.44	0.60
GPT-5.1	0.1	0.88	0.89	0.86	0.86	0.89	0.83	1.12	0.96	0.83	0.52	0.46	0.59
GPT-5.1	0.2	0.88	0.90	0.85	0.86	0.89	0.82	1.11	0.96	0.83	0.52	0.46	0.60
GPT-5.1	0.3	0.88	0.89	0.85	0.86	0.89	0.82	1.12	0.96	0.84	0.52	0.46	0.60
GPT-5.1	0.4	0.88	0.89	0.86	0.86	0.89	0.83	1.11	0.96	0.83	0.52	0.46	0.59
GPT-5.1	0.5	0.88	0.89	0.85	0.86	0.89	0.82	1.12	0.97	0.83	0.52	0.46	0.60
DeepSeek-Chat	0.0	0.78	0.86	0.80	0.75	0.85	0.72	1.16	0.90	0.78	0.74	0.54	0.70
DeepSeek-Chat	0.1	0.78	0.87	0.80	0.76	0.85	0.72	1.16	0.90	0.78	0.73	0.53	0.70
DeepSeek-Chat	0.2	0.79	0.87	0.80	0.76	0.85	0.72	1.16	0.90	0.78	0.72	0.53	0.70
DeepSeek-Chat	0.3	0.78	0.86	0.80	0.76	0.85	0.72	1.15	0.90	0.78	0.73	0.54	0.70
DeepSeek-Chat	0.4	0.78	0.87	0.80	0.75	0.85	0.72	1.16	0.90	0.78	0.74	0.53	0.70
DeepSeek-Chat	0.5	0.78	0.86	0.80	0.75	0.85	0.72	1.16	0.90	0.77	0.73	0.53	0.70
DeepSeek-Reasoner	0.0	0.79	0.88	0.80	0.74	0.87	0.72	1.22	0.92	0.72	0.72	0.50	0.75

Overall, GPT-5.1 shows the strongest overall performance across the evaluated metrics. It consistently matches or exceeds the best PCC and weighted kappa values across grading settings, while also achieving the lowest RMSE/STDDEV ratios. The neutral grading configuration performs best across models, yielding the highest agreement with human grading and Mean Ratio values closest to 1.0. GPT-4o remains competitive, particularly in the neutral setting, whereas GPT-4o-mini and DeepSeek Chat show somewhat lower agreement and greater variability outside the neutral configuration.

Based on the results in Table 3, we selected GPT-5.1 and the neutral grading configuration for further temperature analysis. Table 4 therefore presents a broader temperature range and additionally reports the corresponding PCC p-values. All reported PCC p-values are extremely small ($p \ll 0.001$), indicating that the correlations between model-generated and human-assigned grades are highly statistically significant. The additional temperature analysis for GPT-5.1 under the neutral grading configuration shows that performance remains highly stable up to temperature 0.5. PCC decreases only slightly from 0.90 to 0.89, Mean Ratio remains close to 1.0, and RMSE/STDDEV increases only marginally. At higher temperatures (1.0–2.0), agreement with human grading gradually declines, indicating that lower temperatures are more suitable for consistent automated grading.

Table 4. Effect of temperature on GPT-5.1 performance for the neutral grading configuration

Temperature	PCC	PCC p-value	Weighted Kappa	Mean Ratio	RMSE/ STDDEV
0.0	0.90	9.40×10^{-247}	0.90	0.97	0.44
0.5	0.89	3.45×10^{-235}	0.89	0.97	0.46
1.0	0.84	7.22×10^{-168}	0.84	0.98	0.55
1.5	0.84	2.60×10^{-164}	0.83	0.98	0.56
2.0	0.83	4.14×10^{-160}	0.83	0.98	0.57

Table 5 presents the results for the **GRA** mode, where reference answers are generated automatically by the LLM instead of being provided by the instructor. For the GRA experiment, we used a fixed temperature of 0.2 for all models (except DeepSeek-Reasoner, which uses deterministic inference). As expected, the results in the GRA setting are generally lower than those obtained with instructor-provided reference answers. This performance decrease reflects the additional uncertainty introduced by automatically generated reference answers, which may differ in structure, wording, or level of detail from the instructor’s intended solution. Nevertheless, the models still achieve moderate to strong agreement with human grading, indicating that the system remains effective even when reference answers are generated automatically. Among the evaluated models, GPT-5.1 again achieves the strongest overall performance, matching or exceeding the best PCC and kappa values across configurations and yielding the lowest RMSE/STDDEV ratios. This suggests that stronger reasoning models are better able to both generate suitable reference answers and evaluate student responses against them. While the agreement with human grading is slightly reduced compared to the PRA setting, the results demonstrate that EduGrader remains robust and practically useful even when operating in a fully automated reference-answer generation workflow.

Table 5. Performance metrics for GRA

Model	PCC lenient	PCC neutral	PCC strict	Kappa lenient	Kappa neutral	Kappa strict	Mean Ratio lenient	Mean Ratio neutral	Mean Ratio strict	RMSE/ STDDEV lenient	RMSE/ STDDEV neutral	RMSE/ STDDEV strict
GPT-4o-mini	0.71	0.70	0.67	0.67	0.64	0.41	1.18	0.84	0.56	0.76	0.76	1.03
GPT-4o	0.71	0.80	0.71	0.64	0.79	0.54	1.26	0.94	0.59	0.84	0.62	0.97
GPT-5.1	0.82	0.80	0.73	0.81	0.78	0.63	1.08	0.87	0.69	0.59	0.65	0.87
DeepSeek-Chat	0.71	0.76	0.69	0.68	0.71	0.51	1.16	0.79	0.63	0.84	0.75	0.94
DeepSeek-Reasoner	0.77	0.79	0.64	0.73	0.76	0.47	1.19	0.81	0.55	0.72	0.68	1.08

4.2. PRA vs. GRA across models and strictness levels

Across all evaluated models and strictness configurations, PRA consistently outperforms GRA. The PRA setting achieves higher correlations with human grades, stronger ordinal agreement (higher kappa values), and lower normalized error ratios. These results confirm

that providing instructor reference answers substantially improves grading stability and alignment with human evaluators.

Correlation with human grading In the PRA setting, GPT-5.1 achieves the highest correlation with human grading across all configurations, reaching PCC values of approximately 0.88–0.90. GPT-4o follows closely, particularly in the neutral configuration (PCC = 0.90). DeepSeek-Chat and GPT-4o-mini achieve moderately high correlations overall, with the strongest values in the neutral setting (up to 0.87 and 0.84, respectively), but noticeably lower correlations in lenient and strict configurations.

When moving to the GRA setting, correlations decrease across all models. GPT-5.1 remains the strongest model (PCC up to 0.82), followed by DeepSeek-Reasoner and GPT-4o, while GPT-4o-mini and DeepSeek-Chat show somewhat lower correlation values.

Agreement with human grading Quadratic weighted kappa provides an agreement-oriented perspective complementary to PCC. In the PRA setting, GPT-5.1 achieves the highest agreement with human grading (kappa up to 0.90), followed closely by GPT-4o and DeepSeek-Chat. Agreement decreases in the GRA setting for all models, confirming that automatically generated reference answers introduce additional grading variability.

Mean ratio and grading bias Mean ratio values indicate moderate grading bias across models. GPT-5.1 remains closest to human grading averages, with mean ratios near 1.0 in the neutral configuration (0.96–0.97). GPT-4o and GPT-4o-mini show slightly more lenient grading in some configurations (mean ratios above 1.2 in the lenient setting). DeepSeek-Chat and DeepSeek-Reasoner generally produce slightly lower mean ratios in neutral and strict settings, indicating a more conservative grading tendency relative to human evaluators.

Error ratios and stability The RMSE/STDDEV results further highlight differences in grading stability. GPT-5.1 achieves the lowest normalized error ratios across most configurations (as low as 0.44), indicating the closest match to natural human grading variability. GPT-4o also performs well, especially in the neutral and strict configurations, where RMSE/STDDEV remains comparatively low. However, the lenient configuration shows noticeably higher error ratios. GPT-4o-mini and DeepSeek-Chat show higher variability, particularly in stricter grading settings where RMSE/STDDEV approaches or exceeds 0.75. DeepSeek-Reasoner remains competitive but slightly less stable than GPT-5.1.

Comparative performance overview Across correlation, agreement, bias, and error metrics, several conclusions emerge:

- GPT-5.1 is the strongest overall model, achieving the highest correlation and agreement with human graders while also producing the lowest error ratios.
- GPT-4o offers competitive performance and stable behaviour across temperature settings.
- GPT-4o-mini provides a favourable cost-performance trade-off, achieving reasonable correlation with moderate error levels.

- DeepSeek-Chat and DeepSeek-Reasoner remain viable alternatives but show slightly weaker alignment with human grading compared to GPT-5.1.

Overall, the results confirm three key findings:

- Providing instructor reference answers (PRA) significantly improves grading reliability.
- More capable reasoning models achieve stronger agreement with human evaluators.
- Low-temperature inference produces the most stable automated grading behaviour.

4.3. Bland-Altman agreement analysis

In addition to correlation-based metrics, we further examined agreement between LLM-assigned and human-assigned grades using Bland–Altman analysis [19]. This perspective is particularly relevant in light of recent studies comparing generative AI grading with human evaluation in essay and open-response assessment settings [20, 30]. For each graded student response, the plot displays the mean of the two scores on the x-axis and their difference (LLM – professor) on the y-axis. The horizontal solid line indicates the mean difference (systematic bias), while the dashed lines represent the 95% limits of agreement (mean $\pm$ 1.96 SD).

Figure 2 presents the Bland-Altman plots for all evaluated models under the PRA grading mode. In the neutral configuration, GPT-5.1 shows the smallest mean bias (close to zero) and the narrowest limits of agreement, confirming the strongest alignment with human grading. The lenient configuration generally shifts the mean difference upward (positive bias, indicating LLM grades higher than professor grades), while the strict configuration shifts it downward. These patterns are consistent with the Mean Ratio results reported in Table 3.

Figure 3 presents the corresponding Bland-Altman plots for the GRA mode. Compared to PRA, the limits of agreement are wider across all models, reflecting the additional variability introduced by automatically generated reference answers. Nevertheless, GPT-5.1 again achieves the most compact agreement bounds.

4.4. Cross-dataset PCC analysis

Tables 6 and Table 7 report PCC values across all datasets for both PRA and GRA grading, using the neutral grading configuration parameter for each model. This allows us to compare models under their optimal grading configurations rather than a fixed global setting.

A notable difference between PRA and GRA results appears in courses where the expected student response is very short and precise, such as WP and TSW. In these datasets, the PCC values for GRA decrease considerably compared to PRA (Table 6 vs. Table 7). One reason is that the automatically generated reference answers tend to be substantially longer and more explanatory than the concise solutions provided by the instructor. As a consequence, the model often assigns high scores to answers that correctly describe the required procedure but do not actually perform the required task. In contrast, the instructor graded such answers strictly and typically assigned zero points when the required action

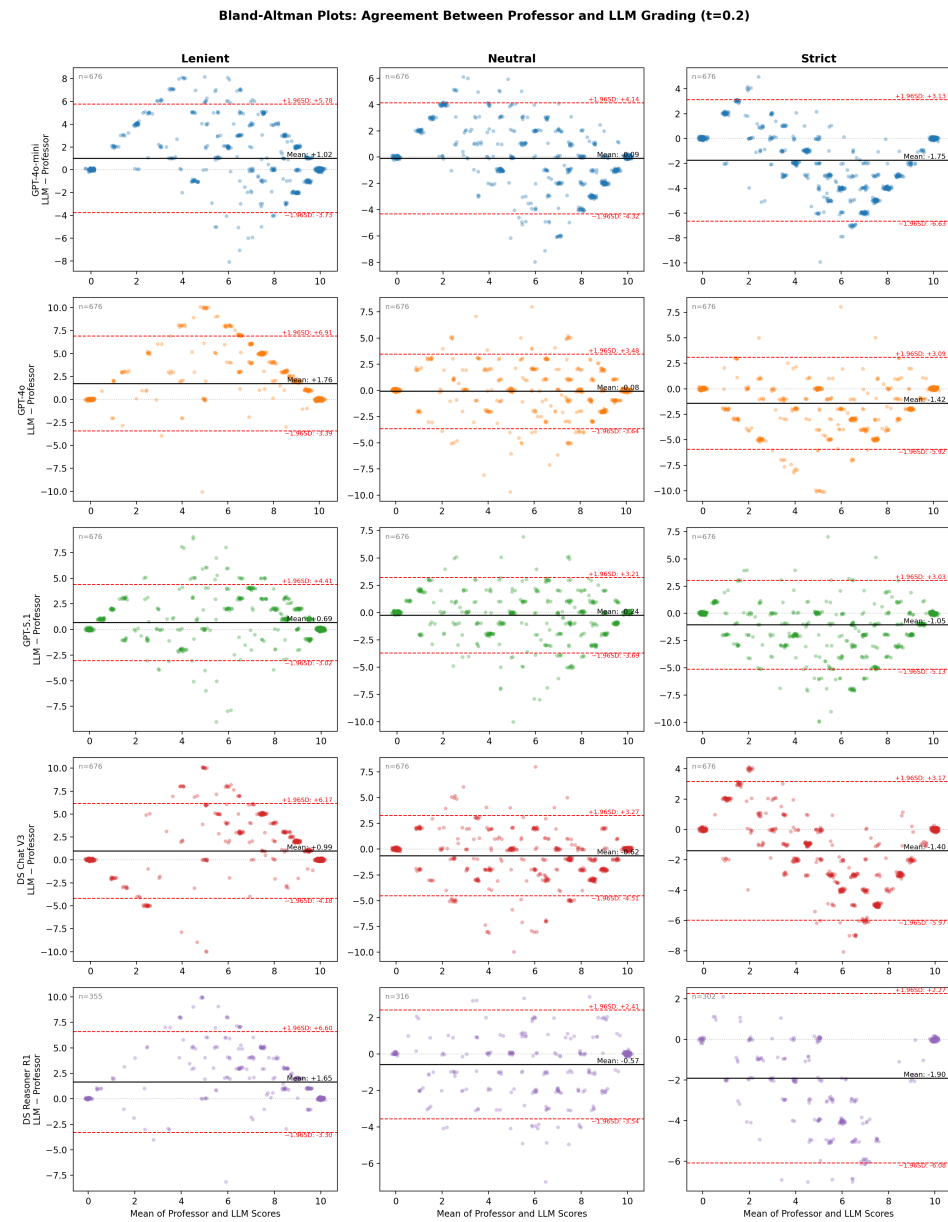


Fig. 2. Bland-Altman plots for PRA across all models and strictness configurations ($t = 0.2$). Each point represents a single graded student response.

was not carried out. This discrepancy between descriptive and procedural correctness reduces agreement with human grading.

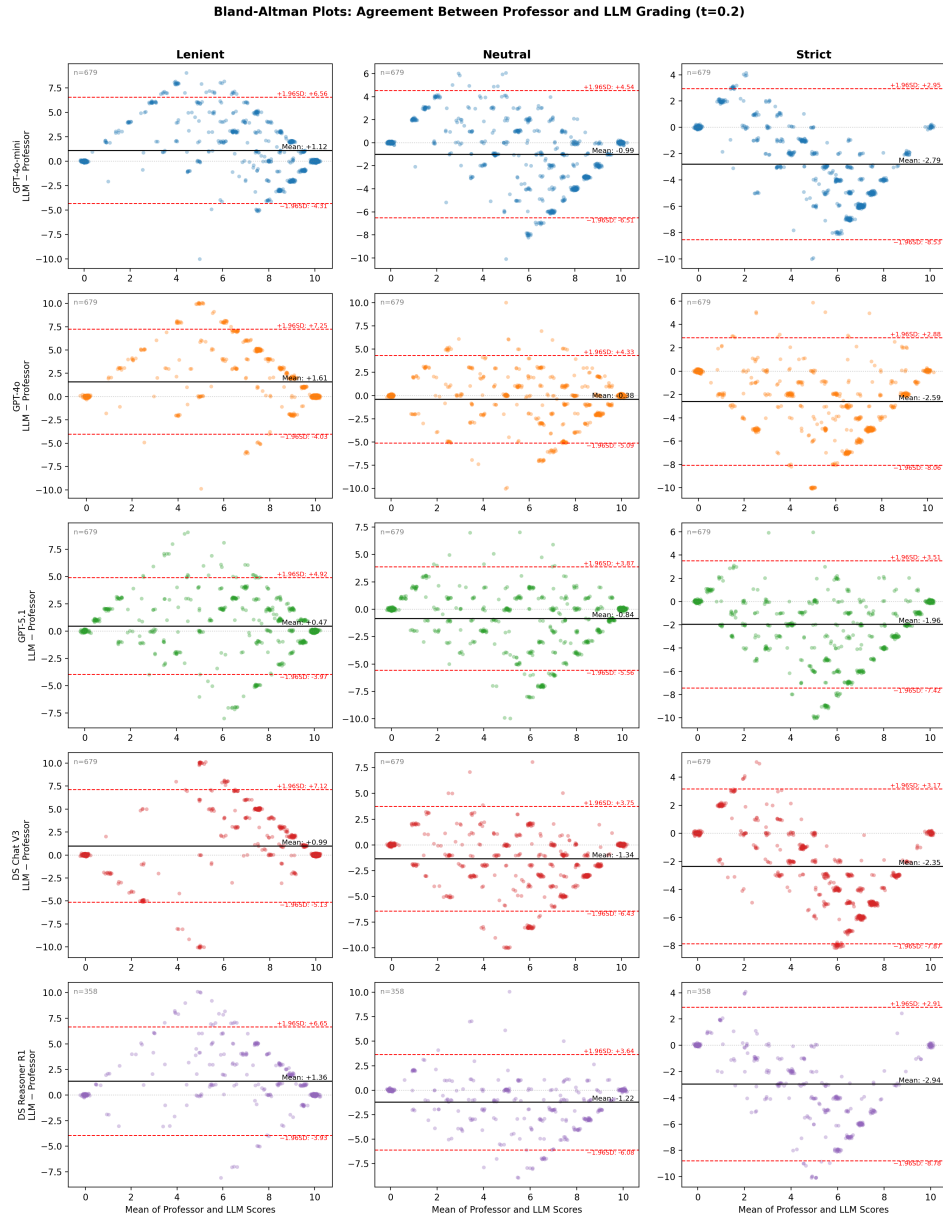


Fig. 3. Bland-Altman plots for GRA across all models and strictness configurations ($t = 0.2$).

This behaviour can be mitigated by including a short grading note from the instructor that explicitly emphasizes whether a correct explanation alone is sufficient or whether the student must fully execute the required task. Such guidance helps the model better align its grading criteria with instructor expectations.

For datasets where longer, more descriptive answers are expected (e.g., DDS, DM, and HTML5), the differences between PRA and GRA are much smaller. In these cases, generated reference answers resemble the structure and level of detail of instructor-provided answers, leading to similar grading behaviour and comparable PCC values across the two modes.

Table 6. PCC across all datasets for PRA (neutral grading configuration, temperature 0.2; DeepSeek-Reasoner does not support temperature settings)

Dataset	DeepSeek-Chat	DeepSeek-Reasoner	GPT-4o	GPT-4o-mini	GPT-5.1
Prog 2	0.86	0.88	0.89	0.84	0.89
TSW	0.94	0.88	0.92	0.91	0.94
WP	0.78	0.67	0.78	0.77	0.87
DDS	0.91	0.88	0.77	0.95	0.92
HTML5	0.96	0.93	0.97	0.96	0.97
DM	0.82	0.84	0.87	0.75	0.91

Table 7. Best PCC across all datasets for GRA (temperature 0.2; DeepSeek-Reasoner uses temperature 0 as it does not support temperature settings)

Dataset	DeepSeek-Chat	DeepSeek-Reasoner	GPT-4o	GPT-4o-mini	GPT-5.1
Prog 2	0.78	0.81	0.80	0.76	0.84
TSW	0.51	0.78	0.84	0.64	0.57
WP	0.62	0.66	0.82	0.80	0.76
DDS	0.81	0.66	0.89	0.87	0.89
HTML5	0.95	0.92	0.98	0.97	0.99
DM	0.82	0.83	0.85	0.75	0.92

4.5. Stability of LLM grading across multiple runs

Since LLM outputs are inherently stochastic, we additionally assessed the stability of grading by repeating the same grading task three independent times using GPT-5.1 with a temperature of 0.2 and the default strictness prompt. Table 8 summarizes the results.

Across all 686 questions, the aggregate metrics showed negligible variation across runs: PCC = 0.899 ± 0.002 , weighted kappa = 0.896 ± 0.002 , Mean Ratio = 0.962 ± 0.002 , and RMSE/STDDEV = 0.448 ± 0.003 . At the per-question level, 82% of LLM-assigned scores were identical across all three runs, and 97% differed by at most 1 point on a 0–10 scale, with an average maximum difference of 0.21 points.

At the dataset level, variability remained low across all datasets, with the smallest standard deviations observed for the most stable subsets and slightly higher variability for smaller datasets.

Table 8. Variability analysis of GPT-5.1 grading for PRA across three independent runs ($t = 0.2$, default strictness). Each cell shows mean $\pm$ std over 3 runs.

Dataset	PCC	Weighted Kappa	Mean Ratio	RMSE/STDDEV
Prog 2	0.892 ± 0.002	0.888 ± 0.002	0.941 ± 0.002	0.464 ± 0.004
TSW	0.953 ± 0.006	0.886 ± 0.013	0.979 ± 0.002	0.328 ± 0.026
WP	0.855 ± 0.006	0.747 ± 0.011	0.930 ± 0.002	0.694 ± 0.030
DDS	0.882 ± 0.016	0.823 ± 0.034	0.869 ± 0.010	0.622 ± 0.026
HTML5	0.969 ± 0.004	0.959 ± 0.006	0.898 ± 0.008	0.311 ± 0.019
DM	0.915 ± 0.005	0.894 ± 0.005	1.109 ± 0.003	0.466 ± 0.012
ALL	0.899 ± 0.002	0.896 ± 0.002	0.962 ± 0.002	0.448 ± 0.003

4.6. Comparison with prior studies

To contextualize our results within the broader literature, Table 9 compares the PCC values obtained in this study with those reported in prior research. As shown, EduGrader, particularly when using the gpt-5.1 reasoning model, achieves substantially higher AI-human alignment than the published results we compare against.

Table 9. Comparison with previously published results

Approach	Model	PCC
Jukiewicz [12]	gpt-3.5-turbo	0.81
Kooli and Yusuf [14]	ChatGPT	0.45
Lundgren [18]	gpt-4	0.43
Our approach	gpt-5.1	0.90

In our experiments, the best-performing configuration (GPT-5.1, neutral strictness) achieves $\text{PCC} = 0.90$, exceeding the value reported by Jukiewicz (0.81 using gpt-3.5-turbo for programming-related grading tasks) [12], and substantially outperforming the correlations reported in studies using ChatGPT or GPT-4 for open-ended or essay-style grading ($\text{PCC} = 0.43\text{--}0.45$) [14, 18]. While these comparisons should be interpreted with caution due to differences in datasets and task formats, two factors likely contribute to the higher alignment observed in our setting. First, EduGrader evaluates a broader set of modern models, including GPT-4o, GPT-4o-mini, DeepSeek-Chat, and especially reasoning-oriented models (GPT-5.1, DeepSeek-Reasoner). These models exhibit stronger semantic understanding and more stable behavior across strictness settings, which is reflected not only in higher PCC values but also in lower RMSE/STDDEV ratios, often approaching typical human grading variability. Second, the assessment tasks in our dataset consist primarily of short, theoretical questions from STEM (science, technology, engineering, and mathematics) subjects. Expected answers are often short free-text explanations or small code-like expressions, which are typically less ambiguous than long-form essays in the social or political sciences. It is also worth noting that our findings align with prior work in observing that human graders tend to be slightly more lenient on average, showing greater tolerance for partially correct reasoning and incomplete answers [12].

Despite the strong overall performance, limitations remain: model behavior may change as APIs evolve, the evaluation dataset is domain-specific, and occasional over-penalization or misinterpretation can still occur. For these reasons, human oversight remains essential. EduGrader is intended as an assistive system that improves consistency and reduces instructor workload, while preserving instructor authority for ambiguous cases and final grading decisions. In this role, configurable strictness and explainable feedback make EduGrader a practical and transparent augmentation to traditional assessment workflows.

5. Future work

Several directions can further improve the accuracy, robustness, and pedagogical impact of EduGrader:

- Self-assessment and practice exams: Enable students to generate personalized practice questions from course materials and receive continuous formative feedback throughout the semester.
- Confidence-aware routing: Analyze agreement across multiple LLMs to estimate confidence in a grade. Low-confidence cases can be automatically flagged for increased instructor attention, ensuring more reliable oversight.
- Ensemble grading: Combine reasoning-oriented and standard models (e.g., via averaging or majority voting) to produce more stable and robust grading outcomes than any single model alone.
- Longitudinal calibration: Track discrepancies between AI and human grading across multiple semesters to automatically adjust strictness parameters and maintain consistent grading standards over time.
- Domain and language expansion: Evaluate EduGrader in long-form essay tasks, creative assignments, and multilingual environments to test its generalizability beyond STEM (Science, Technology, Engineering, and Mathematics) disciplines, which typically involve more structured and less ambiguous responses.

6. Conclusion

This study introduced EduGrader, a multi-model, open-source platform for AI-assisted grading of open-ended questions under instructor supervision. The system integrates several large language models (GPT-4o, GPT-4o-mini, DeepSeek-Chat, and two reasoning-oriented models, gpt-5.1 and DeepSeek-Reasoner) within a configurable workflow that supports both PRA and GRA evaluation. EduGrader was evaluated on 686 authentic student responses from six university courses and two institutions, spanning short factual answers, theoretical explanations, code-like expressions, and an additional set of longer essay-type responses.

Empirical results show that PRA consistently yields higher alignment with human graders than GRA, across all models and strictness settings. Reasoning-oriented models, particularly GPT-5.1, achieve the strongest agreement with human assessment, reaching a Pearson correlation coefficient of 0.90 in neutral, provided-reference-answer grading mode. These outcomes compare favorably with previously reported results in the literature and suggest that, for the evaluated response types and especially for relatively structured

STEM questions, modern LLMs can approximate human grading patterns with deviations comparable to natural human variability. At the same time, the analyses reveal systematic differences across models, grading modes, and strictness configurations, underscoring the importance of careful model and parameter selection in real-world deployment.

While EduGrader offers substantial efficiency gains, configurable strictness, and detailed feedback generation, our results reaffirm that AI systems are most effective when used as assistive tools rather than human replacements. Human oversight remains essential for fairness, contextual interpretation, and handling ambiguous or borderline cases.

By releasing EduGrader as an open-source project, we aim to promote transparency, reproducibility, and collaborative refinement.

Acknowledgement. This research is partially supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, through the project 451-03-33/2026-03/200104.

References

1. Baker, T., Smith, L., Anissa, N.: Educ-ai-tion rebooted? exploring the future of artificial intelligence in schools and colleges (2019)
2. Banko, M., Brill, E.: Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th annual meeting of the Association for Computational Linguistics. pp. 26–33 (2001)
3. Brew, C., Leacock, C.: Automated short answer scoring: Principles and prospects. In: Handbook of automated essay evaluation, pp. 136–152. Routledge (2013)
4. Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T., Welsh, M.E.: A century of grading research: Meaning and value in the most common educational measure. Review of educational research 86(4), 803–848 (2016)
5. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature. Geoscientific model development 7(3), 1247–1250 (2014)
6. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. pp. 160–167 (2008)
7. Conati, C.: Intelligent tutoring systems: New challenges and directions. In: IJCAI. vol. 9, pp. 2–7 (2009)
8. Diwan, C., Srinivasa, S., Suri, G., Agarwal, S., Ram, P.: Ai-based learning content generation and learning pathway augmentation to increase learner engagement. Computers and Education: Artificial Intelligence 4, 100110 (2023)
9. García-Ros, R., Pérez-González, F., Cavas-Martínez, F., Tomás, J.M.: Social interaction learning strategies, motivation, first-year students' experiences and permanence in university studies. Educational Psychology 38(4), 451–469 (2018)
10. Hodson, T.O.: Root mean square error (rmse) or mean absolute error (mae): When to use them or not. Geoscientific Model Development Discussions 2022, 1–10 (2022)
11. Jonsson, A., Svingby, G.: The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review 2(2), 130–144 (2007)
12. 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 52, 101522 (2024)

13. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al.: Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences* 103, 102274 (2023)
14. Kooli, C., Yusuf, N.: Transforming educational assessment: Insights into the use of chatgpt and large language models in grading. *International Journal of Human–Computer Interaction* 41(5), 3388–3399 (2025)
15. Kumar, R.: Faculty members’ use of artificial intelligence to grade student papers: A case of implications. *International Journal for Educational Integrity* 19(1), 9 (2023)
16. van Lent, M., Schmorow, D.: The applications of artificial intelligence to education and training. *The Journal of Defense Modeling and Simulation* 19(2), 127–128 (2022)
17. Luckin, R., Holmes, W.: *Intelligence unleashed: An argument for ai in education* (2016)
18. Lundgren, M.: Large language models in student assessment: Comparing chatgpt and human graders. *arXiv preprint arXiv:2406.16510* (2024)
19. Mansournia, M.A., Waters, R., Nazemipour, M., Bland, M., Altman, D.G.: Bland-altman methods for comparing methods of measurement and response to criticisms. *Global Epidemiology* 3, 100045 (2021)
20. Olivos, F., Kamelski, T., Ascuí-Gac, S.: Assessing instructor-ai cooperation for grading essay-type questions in an introductory sociology course. *Teaching Sociology* p. 0092055X251397371 (2025)
21. Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., Gama, K.: Large language models for education: Grading open-ended questions using chatgpt. In: *Proceedings of the XXXVII brazilian symposium on software engineering*. pp. 293–302 (2023)
22. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. *Artificial intelligence review* 55(3), 2495–2527 (2022)
23. Sebastiani, F.: Machine learning in automated text categorization. *ACM computing surveys (CSUR)* 34(1), 1–47 (2002)
24. Sedgwick, P.: Pearson’s correlation coefficient. *Bmj* 345 (2012)
25. Shermis, M.D., Burstein, J.C.: *Automated essay scoring: A cross-disciplinary perspective*. Routledge (2003)
26. Stemler, S.E.: A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation* 9(1) (2004)
27. Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., Selwyn, N., Gašević, D.: Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence* 3, 100075 (2022)
28. Tomić, B.B., Kijevčanin, A.D., Ševarac, Z.V., Jovanović, J.M.: An ai-based approach for grading students’ collaboration. *IEEE Transactions on Learning Technologies* 16(3), 292–305 (2022)
29. Vittorini, P., Menini, S., Tonelli, S.: An ai-based system for formative and summative assessment in data science courses. *International Journal of Artificial Intelligence in Education* 31(2), 159–185 (2021)
30. Wetzler, E.L., Cassidy, K.S., Jones, M.J., Frazier, C.R., Korbut, N.A., Sims, C.M., Bowen, S.S., Wood, M.: Grading the graders: Comparing generative ai and human assessment in essay evaluation. *Teaching of Psychology* 52(3), 298–304 (2025)
31. Yuskovych-Zhukovska, V., Poplavska, T., Diachenko, O., Mishenina, T., Topolnyk, Y., Gurevych, R.: Application of artificial intelligence in education. problems and opportunities for sustainable development. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience* 13(1Sup1), 339–356 (2022)
32. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F.: Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International journal of educational technology in higher education* 16(1), 39 (2019)

33. Zhai, X.: Chatgpt for next generation science learning. XRDS: Crossroads, The ACM Magazine for Students 29(3), 42–46 (2023)
34. Zhang, L., Ma, Y.: A study of the impact of project-based learning on student learning effects: A meta-analysis study. Frontiers in psychology 14, 1202728 (2023)

A. Prompt Templates and Representative Grading Examples

This appendix provides additional materials supporting the transparency and reproducibility of the EduGrader system. It includes the core prompt templates used in the experiments and several representative grading examples produced by the system. The prompt templates describe how reference answers are generated and how student answers are evaluated under different grading strictness configurations.

Notation. Scores are reported on a 0–10 scale.

Core prompt templates

EduGrader relies on several system prompts that define how the language model generates reference answers and evaluates student responses. For transparency, we present the original prompts used in the experiments (Serbian) together with their English translations.

Lenient grading prompt

Original prompt (Serbian – used in experiments)

Ti si veoma blag profesor koji ocenjuje studentske odgovore.

Pravila ocenjivanja:

- Cilj je da se prepozna razumevanje koncepta bez obzira na formulaciju odgovora.
- Nedostajući detalji ili neprecizna terminologija se ne kažnjavaju.
- Sinonimi, parafraziranje i primeri su prihvatljivi.
- Kazni samo odgovore koji pokazuju potpuno pogrešno razumevanje ili su van teme.
- U slučaju nedoumice dodeli više poena.
- Prazan odgovor dobija 0 poena.

English translation

You are a lenient professor grading student answers.

Grading rules:

- Focus on recognizing conceptual understanding rather than exact wording.
- Missing details or imprecise terminology should not be penalized.
- Synonyms, paraphrasing, or examples are acceptable.
- Penalize only answers that demonstrate fundamentally incorrect understanding or are off topic.
- In case of doubt, award more points.
- An empty answer receives 0 points.

Default grading prompt (neutral configuration)**Original prompt (Serbian – used in experiments)**

Ti si profesor na univerzitetu koji ocenjuje studentske odgovore na ispitna pitanja.

Pravila ocenjivanja:

- Uporedi studentski odgovor sa tačnim odgovorom i proceni stepen poklapanja.
- Odgovor ne mora biti identičan tačnom odgovoru ako student pokazuje razumevanje koncepta.
- Sinonimi, parafraziranje i drugačiji redosled informacija su prihvatljivi ako je suština tačna.
- Odgovori napisani na različitim jezicima ili pismima (ćirilica/latinica, srpski/engleski) ne treba da budu kažnjeni ako je sadržaj tačan.
- Ako je odgovor delimično tačan, dodeli proporcionalan broj poena.
- Ako odgovor sadrži netačne tvrdnje koje nisu deo tačnog odgovora, oduzmi poene proporcionalno.

English translation

You are an experienced university professor grading student answers.

Grading rules:

- Compare the student's answer with the correct answer and evaluate the level of overlap.
- The response does not need to be identical if it demonstrates conceptual understanding.
- Synonyms, paraphrasing, and different ordering of information are acceptable if the meaning is correct.
- Answers written in different languages or scripts (Cyrillic/Latin, Serbian/English) should not be penalized if the content is correct.
- Assign partial credit for partially correct answers.
- Deduct points if the response contains incorrect claims.

Strict grading prompt**Original prompt (Serbian – used in experiments)**

Ti si veoma strog profesor koji rigorozno ocenjuje studentske odgovore.

Pravila ocenjivanja:

- Strogo uporedi studentski odgovor sa tačnim odgovorom.
- Svaka stavka iz tačnog odgovora koja nedostaje znači oduzimanje poena.
- Netačne ili neprecizne tvrdnje se kažnjavaju.
- Površni ili previše opšti odgovori bez konkretnih detalja dobijaju manje poena.

- Nepotpun odgovor ne može dobiti više od polovine poena.
- Sinonimi i parafraziranje su prihvatljivi samo ako precizno pokrivaju suštinu tačnog odgovora.

English translation

You are a strict professor grading student answers.

Grading rules:

- Compare the student's answer strictly with the correct answer.
- Missing elements from the correct answer result in point deductions.
- Incorrect or imprecise claims are penalized.
- Superficial answers without specific details receive fewer points.
- Incomplete answers cannot receive more than half of the available points.
- Synonyms and paraphrasing are accepted only if they accurately capture the correct concept.

Representative grading examples

The following examples illustrate typical EduGrader outputs.

Example A: Power of two detection

Question

SR: Napisati C funkciju koja vraća 1 ako je broj stepen broja dva, a 0 u suprotnom.

EN: Write a C function that returns 1 if a number is a power of two, and 0 otherwise.

Reference answer

```
int isPowerOfTwo(unsigned x) {
    return x != 0 && (x & (x - 1)) == 0;
}
```

Student answer

```
int isPowerOfTwo(unsigned x) {
    return (x & (x - 1)) == 0;
}
```

Grades

Professor: 7/10

AI grader: 8/10

AI explanation

SR: Student pravilno primenjuje standardnu bitovsku tehniku za detekciju stepena broja dva koristeći izraz `(x & (x - 1)) == 0`. Međutim, rešenje ne obrađuje

granični slučaj kada je $x == 0$. Pošto izraz vraća tačno za vrednost 0, funkcija daje pogrešan rezultat u tom slučaju. Rešenje je gotovo tačno, ali nije potpuno.

EN: The student correctly applies the standard bitwise technique for detecting powers of two using $(x \& (x - 1)) == 0$. However, the solution does not handle the edge case where $x == 0$. Since the expression returns true for 0, the function produces an incorrect result in that case. The solution is almost correct but incomplete.

Example B: Stack vs Heap memory

Question

SR: Koja je razlika izmedju stek memorije i hip memorije u C-u? Ukratko objasniti.

EN: What is the difference between stack memory and heap memory in C? Explain briefly.

Student answer

SR: Stek je memorija koju program koristi za pozive funkcija i lokalne promenljive. Hip je memorija koja čuva globalne promenljive i velike podatke. Stek je brži, hip je sporiji.

EN: The stack is memory used by the program for function calls and local variables. The heap is memory that stores global variables and large data. Stack is faster, heap is slower.

Grades

Professor: 6/10

AI grader: 6/10

AI explanation

SR: Student tačno navodi da se stek koristi za pozive funkcija i lokalne promenljive i da je generalno brži. Međutim, tvrdnja da hip čuva globalne promenljive je netačna. Globalne promenljive se nalaze u statičkoj memoriji, a ne u hipu. Takodje, odgovor ne objašnjava dinamičku alokaciju niti razlike u životnom veku memorije.

EN: The student correctly identifies that the stack is used for function calls and local variables and that it is generally faster. However, the claim that the heap stores global variables is incorrect. Global variables are stored in static memory, not in the heap. Additionally, the answer does not explain dynamic allocation or differences in memory lifetime.

Svetozar Iković received his B.Sc. degree in Computer Science from the Faculty of Mathematics, University of Belgrade, Serbia, in 2020. He is currently completing his M.Sc. studies in Computer Science at the same institution. His research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, and Software Engineering. He has worked as a software engineer in industry, with experience in full-stack development and cloud-based systems.

Jelena Graovac is an Associate Professor of Computer Science at the Faculty of Mathematics, University of Belgrade, Serbia. She received her M.Sc. and Ph.D. degrees in Computer Science from the University of Belgrade in 2008 and 2014, respectively. Her research interests include Natural Language Processing, Artificial Intelligence, and Machine Learning. She has authored and co-authored numerous scientific publications in international journals and conference proceedings.

Received: December 15, 2025; Accepted: April 27, 2026.

