11th International Conference on Higher Education Advances (HEAd'25)
Universitat Politècnica de València, València, 2025
DOI: https://doi.org/10.4995/HEAd25.2025.20152

A Framework for Automated Student Grading Using Large Language Models

Josep Domenech[1], Pau Carles-Vega[1], Alicia Martínez-Varea[2]
[1]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain,[2]University and Polytechnic La Fe Hospital, Spain.

How to cite: Domenech, J.; Carles-Vega, P.; Martinez-Varea, A. (2025). A Framework for Automated Student Grading Using Large Language Models. In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17–20 June 2025. https://doi.org/10.4995/HEAd25.2025.20152

Abstract
This paper proposes a systematic framework for integrating large language models (LLMs) into the evaluation of student work. The framework addresses challenges inherent in automated grading, such as ensuring validity, reliability, and minimizing bias, by outlining a structured process that includes prompt design, model selection, evaluation, calibration, and iterative refinement. The approach is designed to be adaptable across diverse educational contexts, supporting both formative and summative assessment needs. This work contributes to the growing literature on AI-driven education, offering practical guidelines and highlighting the need for careful design and continuous validation for high-stakes educational applications.

Keywords: *Large Language Models; automatic asssessment; framework.*

1. Introduction

The advent of large language models (LLMs) has marked a significant milestone in artificial intelligence. These models, capable of understanding and generating human-like text, have evolved beyond mere linguistic tasks to demonstrate abilities in contextual understanding, reasoning, and problem-solving (Budnikov et al., 2024). Additionally, with their multimodal capacities, emerging LLMs are increasingly able to process and integrate inputs from diverse formats such as text, images, and structured data (Yin et al., 2024). These advancements position LLMs as transformative tools across various domains, including education, where their potential for evaluating student performance has garnered considerable interest (Domenech, 2023; Kortemeyer, 2023; Latif & Zhai, 2024).

Education systems worldwide continue to seek scalable, accurate, and efficient methods for evaluating student performance (Timotheou et al., 2023). Traditional assessment approaches, while well-established, often require substantial time and effort from educators (Brown, 2022; Dikli, 2006), a challenge that has been acknowledged in early work on automated grading

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

systems (Page, 1966). LLMs offer a promising alternative to automate the analysis of student submissions and generating feedback. With their ability to contextualize responses and adapt to diverse prompts, LLMs can potentially streamline grading processes and provide individualized, text-based feedback to enhance learning outcomes.

However, the application of LLMs in student evaluation is not without challenges. One critical concern is the validity of these evaluations, particularly whether they measure constructs comparable to those assessed by human raters (Attali, 2013). Can LLMs accurately assess student work across diverse disciplines and levels of complexity? Are the scores and feedback generated by these models reliable and unbiased? Ensuring the validity, reliability and fairness of LLM-based grading is paramount to their adoption in educational settings. Addressing these concerns requires robust validation frameworks that align with established educational standards and account for the depth inherent in human judgment.

Prior research on integrating LLMs into student evaluation has adopted a variety of approaches, ranging from direct application of general-purpose models to fine-tuning them for specific educational tasks. In some cases, educators have manually entered student responses into tools like ChatGPT to solicit grading suggestions or feedback (Floden, 2024; Sreedhar et al., 2024). Others have experimented with optimization techniques, such as prompt engineering, to elicit more accurate and contextually relevant model outputs (Wan & Chen, 2024). Validation of these methodologies has been equally diverse, drawing on quantitative alignment with human graders (Pack et al., 2024; Sreedhar et al., 2024), qualitative analysis of feedback (Almasre, 2024; Wan & Chen, 2024), and psychometric evaluations to assess reliability (Pack et al., 2024; Yavuz et al., 2025). Despite these efforts, a comprehensive framework that systematically integrates LLM capabilities into educational evaluation, while addressing the challenges of validity and scalability, remains an ongoing challenge.

The objective of this paper is to bridge this gap by proposing a general framework for employing LLMs in the evaluation of student work. This framework encompasses key dimensions, including model selection, input preparation, evaluation design, and validation mechanisms. Additionally, it aims to provide guidelines for educators and researchers on optimizing LLM usage while mitigating biases and ensuring alignment with pedagogical goals. By synthesizing existing approaches and addressing current limitations, this work seeks to advance the understanding and application of LLMs in education, ultimately contributing to more equitable, efficient, and insightful assessment practices.

This paper is organized as follows. Section 2 reviews related literature and approaches to LLM-based evaluations. Section 3 outlines the proposed framework, detailing its components and discussing validation techniques. Finally, Section 4 draws some concluding remarks.

2. Background

The evolution of automated assessment systems in education has a rich history, beginning with early efforts to mechanize grading processes (Page, 1966) and progressing through various technological advances. This section reviews key developments and current approaches in automated student evaluation, with particular focus on the emergence of LLM-based methods.

2.1. Traditional Automated Assessment

Early automated grading systems primarily focused on multiple-choice questions and other structured response formats that could be evaluated through pattern matching (Forsythe & Wirth, 1965; Page, 1966). These systems gradually evolved to incorporate more sophisticated natural language processing (NLP) techniques, enabling the assessment of short-answer questions and essays. Traditional automated essay scoring (AES) systems typically relied on extracting linguistic features such as vocabulary usage, syntactic complexity, and discourse coherence to predict human-assigned scores (Attali & Burstein, 2004).

Despite their utility, these conventional approaches faced several limitations. They often struggled with semantic understanding, context sensitivity, and the ability to provide detailed, constructive feedback. Additionally, their effectiveness was largely confined to specific domains and question types, requiring substantial effort to adapt to new contexts (Balfour, 2013; Dikli, 2006).

2.2. LLM-based Assessment Approaches

Advances in deep learning have led to the development of large language models that overcome some of the challenges faced by traditional AES systems. Leveraging extensive training on varied text corpora, these models are capable of understanding and generating contextually rich responses, which makes them suitable for assessing open-ended student work (Hu et al., 2025; Mondal et al., 2024). Research such as that by Yavuz et al. (2025) and Latif and Zhai (2024) has demonstrated that LLM-based approaches can yield evaluations that correlate with human judgments.

LLM-based methods provide several benefits compared to conventional approaches. Their capacity to capture semantics and generate detailed feedback supports formative assessment practices that can inform both teaching and learning (Morris et al., 2024). In addition, the flexibility inherent to these models facilitates their application across diverse disciplines and question formats without requiring extensive modifications for each new context (Bruscia et al., 2024; Wang & Zhang, 2024).

However, these methods are not without challenges. The sensitivity of LLM outputs to prompt formulation can result in variability in assessments (Liu et al., 2023; Wan & Chen, 2024).

Moreover, ensuring that evaluations produced by LLMs are unbiased and adhere to educational standards remains an area of concern (Attali & Burstein, 2004). In response, researchers have begun to develop calibration techniques and robust validation protocols that integrate quantitative measures with qualitative insights (Pack et al., 2024; Sreedhar et al., 2024).

While LLM-based assessment approaches hold promise for enhancing the scalability of automated student evaluation, further investigation is necessary to ensure their reliability and fairness. The framework proposed in this paper builds on these developments by outlining systematic procedures for model selection, input preparation, evaluation design, and validation.

3. Framework

This section presents a comprehensive framework for implementing LLM-based student evaluation systems. The framework encompasses the essential components and processes required to implement reliable and valid automated assessment systems based on LLMs. Figure 1 illustrates the primary components and their interactions. Although described sequentially, the process is not strictly linear. Outcomes from the evaluation stage can motivate changes in prompt design and model configuration.

System Overview. The framework starts with a student's text, such as a short answer or essay, and produces a final score or grade. Optionally, it can also generate textual feedback. Four key elements structure the approach: prompt design, the application of the LLM itself, evaluation, and an optional calibration step.

Prompt design defines how the student's submission and grading instructions are presented to the LLM. The model then processes the prompt and student text to produce raw outputs. The evaluation stage verifies the reliability of the LLM's responses and assesses their alignment with human-based grading. If biases or systematic deviations are found, an optional calibration step can adjust the model's outputs. Finally, the process can iterate, incorporating insights from evaluation into refined prompt designs or updated model parameters.

Prompt Design. Prompt design specifies the task and grading criteria for the LLM. It should include a structured rubric detailing performance levels and specific criteria relevant to the assignment. The system may also incorporate examples that demonstrate correct and incorrect responses, known as few-shot prompting, to guide the model's reasoning. Additionally, the prompt may also instruct the LLM to produce intermediate reasoning steps (chain-of-thought) before arriving at a final judgment. This design can be further adjusted by assigning a particular role to the model, for instance, instructing it to act as a teacher or domain expert. Since LLMs can be sensitive to variations in how the prompt is phrased, prompt design often involves iterative experimentation and refinement to achieve reliable outputs.

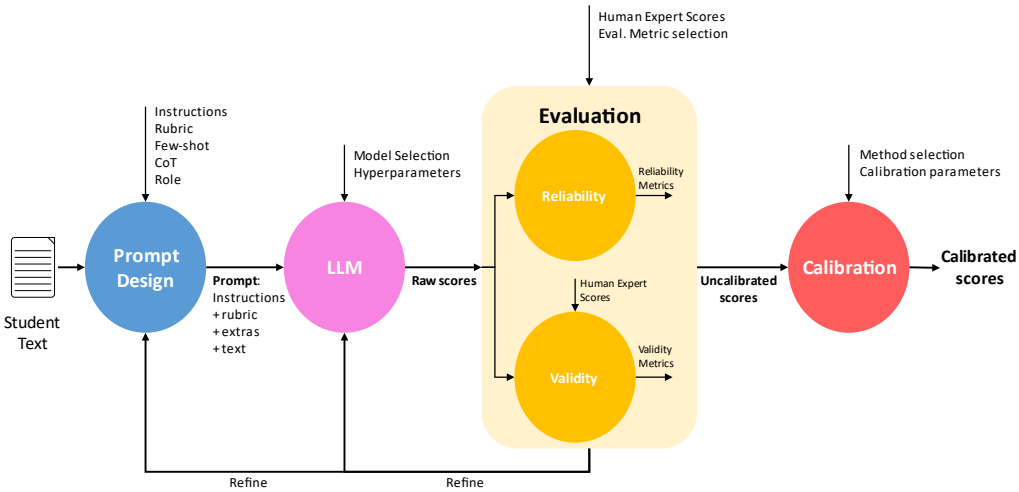

Figure 1. Overview of the proposed LLM-based grading framework. The system receives a student's text and produces raw scores using a prompt design and model selection process. The scores are then evaluated for reliability and validity, optionally undergoing calibration to correct biases or misalignments with human expert judgments. Iterative refinements can be made at each stage to improve the overall quality and reliability of the assessments.

Large Language Model. Once the prompt is formulated, it is fed into the chosen LLM together with the student text. Model selection is a key decision, as LLMs differ in their training data, size, and domain specialization. Hyperparameters such as temperature or maximum output length can further shape the model's style and reliability. The raw output from the LLM typically includes a proposed score or grade, and it can also contain explanatory feedback. However, these initial results may not fully align with human evaluations, and they can vary across repeated queries, which motivates the subsequent evaluation stage.

Evaluation. Evaluation assesses whether the model's outputs are both valid and reliable. Validity checks involve comparing the model's outputs with human-assigned scores. These comparisons may rely on statistical metrics, such as Pearson correlation coefficients (r), Spearman's rank correlation (ρ), and mean squared error (MSE) between LLM and human scores, or more qualitative analyses of the feedback provided. The goal is to identify potential biases or patterns of error and to establish how closely the model's grading aligns with established standards.

Reliability assessment complements validity by examining the consistency of the model's outputs. This involves testing whether the LLM produces similar results when presented with the same or slightly modified prompts. Metrics such as inter-rater reliability, intraclass correlation coefficient (ICC), and Cohen's or Fleiss' kappa are applied to evaluate consistency across different instances of scoring. Large fluctuations may indicate a need to adjust either the prompt design or hyperparameters. The relationship between reliability and validity is

particularly important in LLM-based assessment, as unreliable measurements inherently limit the potential validity of the system.

Calibration. When systematic deviations are observed, an optional calibration step can correct for biases in the LLM's grading. If, for example, the model consistently assigns higher scores than human raters, a calibration function can shift the raw outputs to more closely match human scores. Simple approaches might use linear transformations based on observed differences, whereas more complex methods can incorporate additional contextual variables. Calibration is especially useful in high-stakes assessments, where grading reliability is paramount.

Iterative Refinement. Although the framework is presented as a series of steps, it is often applied in a cyclic manner. Insights from the evaluation and calibration stages can inform adjustments to the prompt design or changes to model parameters. Over multiple iterations, the system can converge on a more stable, valid, and transparent grading process.

4. Conclusions

This paper has presented a systematic framework for implementing LLM-based student evaluation systems, addressing key challenges in automated assessment while leveraging the advanced capabilities of modern language models. The proposed framework integrates essential components including prompt design, model selection, evaluation protocols, and calibration mechanisms to create a robust foundation for educational assessment applications.

Several important insights emerge from this work. First, the effectiveness of LLM-based assessment systems depends heavily on careful prompt engineering and model selection. The framework emphasizes the importance of structured rubrics and clear evaluation criteria, while acknowledging that prompt design often requires iterative refinement to achieve optimal results. Second, the evaluation process must balance multiple objectives, including alignment with human judgment, consistency across assessments, and the generation of meaningful feedback. The proposed validation mechanisms provide a structured approach to measuring and improving these aspects of system performance.

Looking ahead, the integration of LLMs into educational assessment holds significant promise for scaling personalized feedback and reducing educator workload. However, successful implementation requires careful attention to validity, reliability, and fairness. The framework presented here provides some ground for developing such systems, while acknowledging the ongoing need for human oversight and validation in educational assessment.

References

Almasre, M. (2024). Development and Evaluation of a Custom GPT for the Assessment of Students' Designs in a Typography Course. Education Sciences, 14 (2), 148. https://doi.org/10.3390/educsci14020148

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 181-198). Taylor & Francis.

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater v.2.0. ETS Research Report Series, 2004 (2). https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Balfour, S. P. (2013). Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer ReviewTM. Research & Practice in Assessment, 8, 40-48. https://eric.ed.gov/?id=EJ1062843

Brown, G. T. L. (2022). The past, present and future of educational assessment: A transdisciplinary perspective. Frontiers in Education, 7. https://doi.org/10.3389/feduc.2022.1060633

Bruscia, M., Manduzio, G. A., Galatolo, F. A., Cimino, M. G., Greco, A., Cominelli, L., & Scilingo, E. P. (2024). An Overview On Large Language Models Across Key Domains: A Systematic Review. 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE), 125-130. https://doi.org/10.1109/MetroXRAINE62247.2024.10797032

Budnikov, M., Bykova, A., & Yamshchikov, I. P. (2024). Generalization potential of large language models. Neural Computing and Applications, 1-25.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. The Journal of Technology, Learning, and Assessment, 5 (1).

Domenech, J. (2023). ChatGPT in the Classroom: Friend or Foe? 9th International Conference on Higher Education Advances (HEAd'23), 339-347. https://doi.org/10.4995/head23.2023.16179

Floden, J. (2024). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. British Educational Research Journal. https://doi.org/10.1002/berj.4069

Forsythe, G. E., & Wirth, N. (1965). Automatic grading programs. Communications of the ACM, 8 (5), 275-278. https://doi.org/10.1145/364914.364937

Hu, B., Zhu, J., Pei, Y., & Gu, X. (2025). Exploring the potential of LLM to enhance teaching plans through teaching simulation. NPJ Science of Learning, 10 (1), 7. https://doi.org/10.1038/s41539-025-00300-x

Kortemeyer, G. (2023). Toward AI grading of student problem solutions in introductory physics: A feasibility study. Physical Review Physics Education Research, 19 (2). https://doi.org/10.1103/physrevphyseducres.19.020163

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. Computers and Education: Artificial Intelligence, 6, 100210. https://doi.org/ 10.1016/j.caeai.2024.100210

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys, 55 (9), 1-35. https://doi.org/10.1145/3560815

Mondal, H., De, R., Mondal, S., & Juhi, A. (2024). A large language model in solving primary healthcare issues: A potential implication for remote healthcare and medical education. Journal of Education and Health Promotion, 13, 362. https://doi.org/10.4103/jehp.jehp_688_23

Morris, W., Crossley, S., Holmes, L., Ou, C., Dascalu, M., & McNamara, D. (2024). Formative Feedback on Student-Authored Summaries in Intelligent Textbooks Using Large Language Models. International Journal of Artificial Intelligence in Education. https://doi.org/10.1007/s40593-024-00395-0

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 47 (5), 238-243.

Sreedhar, R., Chang, L., Gangopadhyaya, A., Shiels, P. W., Loza, J., Chi, E., Gabel, E., & Park, Y. S. (2024). Comparing Scoring Consistency of Large Language Models with Faculty for Formative Assessments in Medical Education. Journal of General Internal Medicine. https: //doi.org/10.1007/s11606-024-09050-9

Timotheou, S., Miliou, O., Dimitriadis, Y., Sobrino, S. V., Giannoutsou, N., Cachia, R., Mon es, A. M., & Ioannou, A. (2023). Impacts of digital technologies on education and factors influencing schools' digital capacity and transformation: A literature review. Education and information technologies, 28 (6), 6695- 6726.

Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. Physical Review Physics Education Research, 20 (1). https://doi.org/10.1103/physrevphyseducres.20.010152

Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. Artificial Intelligence Review, 57 (11), 299. https://doi.org/10.1007/s10462-024-10921-0

Yavuz, F., Çelik, Ö., & Yavas Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology 56, 150–166.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. National Science Review, 11 (12), nwae403. https://doi.org/10.1093/nsr/nwae403