Closing the Feedback Loop: A Deployment of LLM-driven Question Generation and Automated Grading in a Large-Scale Undergraduate Course

Paper ID: 821

No Institute Given

Abstract. Open-ended questions are a powerful tool for fostering deep learning, but providing timely feedback in large courses is a major logistical challenge. Large Language Models (LLMs) offer a promising solution by automating the entire assessment lifecycle, from question generation to grading. This paper presents the design, deployment, and evaluation of a closed-loop, LLM-powered system for generating questions and providing automated feedback in a 14-week undergraduate database course with 170 students. We conducted a comprehensive, in-vivo study to assess the quality of generated questions, the reliability of automated grading, and the impact on student engagement.

Our findings reveal a nuanced picture. We found that while LLMs can rapidly generate a large number of questions from course materials, significant human oversight is required to ensure pedagogical quality, with rejection rates for generated questions being high due to issues like lack of grounding and excessive difficulty. In grading, LLMs showed high agreement with human experts for correct answers, but struggled with partially correct responses. Interestingly, while our data showed some models being more lenient, over 72% of students perceived the AI grading as overly strict. Despite this, students reported high trust in the AI's feedback. Finally, engagement data showed that students predominantly used the system for last-minute credit rather than for continuous formative learning, highlighting a critical gap between the availability of a tool and its intended pedagogical use. Our study provides key insights into the practical challenges and opportunities of integrating end-to-end AI assessment systems into real-world educational settings.

Keywords: Large Language Models · Automated Short Answer Grading · Question Generation · Formative Feedback · Educational Technology.

1 Introduction

Formative feedback is a powerful lever for improving student learning, with timely and constructive guidance being a cornerstone of effective pedagogy [24, 14]. Among various formative exercises, open-ended short-answer questions are particularly valuable. They compel students to move beyond simple recognition and engage in deeper cognitive processes like recall and synthesis, which are

strongly linked to better long-term retention and comprehension [3, 21]. However, the logistical challenges of using short-answer questions at scale are substantial. Crafting high-quality questions that span different cognitive levels (e.g., Bloom's Taxonomy) is a time-consuming art for educators [2]. More daunting is the subsequent grading workload, which becomes an insurmountable bottleneck in large courses, often leading to significant delays in feedback that can negate its pedagogical value [6]. Consequently, many large courses default to multiple-choice questions, which, despite their scalability, fail to assess higher-order thinking.

The recent surge in the capabilities of Large Language Models (LLMs) offers a potential paradigm shift [4, 27]. LLMs' proficiency in text generation and analysis suggests they could automate the entire lifecycle of formative assessment, from question generation [25, 11] to automated short-answer grading (ASAG) [5, 9]. Indeed, recent studies have explored using frontier models like GPT for grading, showing promising results in controlled settings [22, 8, 15]. However, much of this research focuses on the grading task in isolation, often using existing, static datasets. A critical gap remains: there is a lack of research on the deployment and evaluation of a complete, end-to-end system in a live, large-scale course, where the complexities of real-world student behavior and pedagogical goals come into play.

This study addresses this gap by presenting the design, deployment, and comprehensive evaluation of a dual-pipeline LLM-powered system that both generates questions and provides automated grading and feedback. **Our primary novelty lies in the holistic, in-vivo evaluation of this closed-loop system within a 14-week undergraduate database course with 170 students.** Unlike studies that use static datasets or focus on a single task, we deployed our system for weekly quizzes, collecting thousands of real student answers and, crucially, their direct feedback on the utility of the AI-generated grades and comments. This provides an authentic, large-scale perspective on the practical viability of such a system.

To guide our investigation, we address the following research questions:

1. **RQ1 (Generation Quality):** How do different frontier LLMs compare in generating pedagogically relevant short-answer questions from course materials, in terms of difficulty, cognitive skill level (Bloom's Taxonomy), and overall quality as judged by human experts?
2. **RQ2 (Grading Reliability):** How accurately can various LLMs grade student answers compared to human experts, and how do students perceive the quality of the generated feedback?
3. **RQ3 (Student Impact):** What is the relationship between student engagement with the automated quiz system and their final course performance?

Our main contributions are:

– The design, implementation, and, most importantly, the **deployment and evaluation of an end-to-end system** for LLM-powered question generation and automated grading in a live, large-scale undergraduate course.

- A comparative analysis of two frontier LLMs for the task of question generation, evaluated by domain experts across multiple pedagogical dimensions including grounding, difficulty, and cognitive skill level.
- A multi-faceted evaluation of the LLM-based grading system, measuring its accuracy against human experts and analyzing over **4,000 pieces of in-situ student feedback** on the quality of the automated feedback.
- A new, annotated dataset of 565 LLM-generated questions and 400 graded student answers to support further research in automated assessment.

This paper proceeds as follows: Section 2 reviews related work. Section 3 describes the course context. Section 4 details our system architecture. Section 5.1 outlines our evaluation methodology. Section 5 presents our results, followed by a discussion in Section 6 and our conclusion in Section 7.

2 Related Work

Our work intersects with two primary areas of research in educational technology: the automatic grading of student answers and the automatic generation of questions. We position our contribution as bridging these two areas with a holistic, in-vivo system deployment.

2.1 Automated Short Answer Grading (ASAG)

The task of automatically grading short, open-ended student answers is a long-standing challenge in AI in Education. Early approaches relied on keyword matching or statistical methods like Latent Semantic Analysis (LSA) to measure the similarity between student and reference answers [19]. A comprehensive survey by Burrows et al. [5] details this history, marking the shift towards more sophisticated machine learning and natural language processing techniques.

Deep learning, particularly transformer-based models as BERT [10], revolutionized the field. By fine-tuning models on domain-specific data, researchers achieved significant improvements in grading accuracy [26]. The use of sentence embeddings, such as those from Sentence-BERT [23], allowed for more nuanced semantic comparisons, moving beyond simple lexical overlap [7]. This line of work has culminated in a variety of datasets and benchmarks for the ASAG task [18, 20, 12].

Recently, the focus has shifted towards LLMs. Several studies have demonstrated that modern LLMs can act as powerful few-shot or even zero-shot graders, achieving high agreement with human experts without task-specific fine-tuning [28, 22, 17]. Studies by Chang and Ginter [8] and Henkel et al. [15] evaluated ChatGPT's grading capabilities in different languages and educational levels, finding promising but not yet perfect performance. Concurrently, research has also focused on not just providing a score, but also generating constructive feedback, a critical component for learning. The EngSAF dataset by Aggarwal et al. [1] is a notable contribution in this direction, providing a benchmark for feedback

generation. Our work builds on these findings by using LLMs for both grading and feedback, but evaluates them in a live, continuous deployment rather than on static datasets.

2.2 Automatic Question Generation (AQG)

Parallel to the work in grading is the field of Automatic Question Generation (AQG). The goal of AQG is to alleviate the burden on educators of creating high-quality assessment materials. Traditional methods for AQG often relied on rule-based systems or templates applied to structured knowledge sources. However, as with ASAG, the rise of LLMs has opened new frontiers.

Modern approaches leverage the generative power of LLMs to create a wide variety of question types from unstructured text, such as course materials [25]. This allows for the generation of questions that can target different levels of Bloom's Taxonomy, from simple recall to more complex analysis and application [2]. Elkins et al. [11] investigated the utility of questions generated by LLMs, finding that while they are often grammatically correct and relevant, their pedagogical quality can vary. This highlights the need for human-in-the-loop systems, where educators can curate and approve AI-generated content, which is the model we adopt in our work.

2.3 Research Gap and Our Contribution

While there is a growing body of work on using LLMs for isolated educational tasks as grading or question generation, few studies have combined these into a single, closed-loop system and deployed it in an educational context. Many evaluations of LLM grading capabilities are performed on existing, static datasets [16] or in controlled, one-off experiments [13]. Our work addresses this gap by:

1. Implementing and deploying an **end-to-end system** that handles both LLM-based question generation and automated grading with feedback.
2. Conducting a **longitudinal, in-vivo study** over 14 weeks in a large-scale undergraduate course, capturing the complexities of real-world usage.
3. Collecting not only student answers but also their **direct feedback** on the quality and usefulness of the AI-generated feedback, providing a unique, student-centered evaluation perspective.

By doing so, we provide a more holistic and ecologically valid assessment of the potential for LLMs to augment and support formative assessment cycles in higher education.

3 Course Context and Pedagogical Design

The study was conducted over a 14-week semester in a mandatory undergraduate "Databases" module at a university of applied sciences. The course enrolled approximately 170 students from diverse degree programs, including data

science, industrial engineering, medical informatics, and applied digital life sciences. This created a heterogeneous cohort, with about half the students in their first semester and the other half in their fourth, bringing a wide range of prior knowledge and academic maturity to the course.

The module was delivered in a blended learning format, combining asynchronous self-study with synchronous in-person sessions. The asynchronous components consisted of course slides, supplementary reading materials, and instructional videos. These materials were intended to be reviewed by students prior to the weekly synchronous sessions. The in-person meetings were dedicated to interactive problem-solving, hands-on exercises, and deeper exploration of complex topics, allowing the instructor to address questions and reinforce concepts.

The assessment structure was designed to encourage continuous engagement. A final written, closed-book exam accounted for 80% of the final grade. Three larger, graded assignments, which could be completed in groups of up to three, contributed 15%. The remaining 5% was awarded for the completion of small, weekly formative exercises—the focus of this study. These exercises were individual tasks designed to help students practice and self-assess their understanding of the weekly topics. The low-stakes nature of these quizzes was intentional, aiming to foster a learning-oriented environment where students could freely engage with the material without the pressure of high-stakes evaluation. The LLM-powered system described in Section 4 was deployed to deliver these weekly formative quizzes.

4 System Design

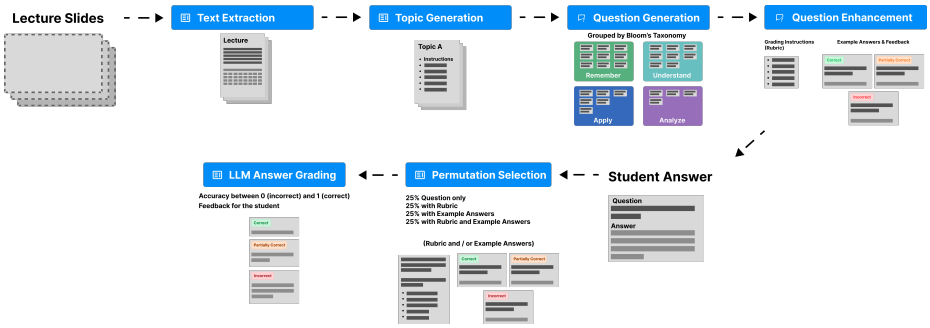

Fig. 1. High-level overview of the dual-pipeline system architecture.

Our system is designed as a dual-pipeline architecture (Figure 1). The **Question Generation Pipeline**, is an offline process that automatically creates relevant questions, grading rubrics, and example answers from course materials. The **Learning and Grading Pipeline**, is an online system that for students, it collects their answers, and provides them with immediate, AI-generated grades and feedback. This section details the design of both pipelines.

4.1 Question Generation Pipeline

The Mistral OCR model is employed to perform optical character recognition on the slides, converting them into machine-readable text; a future iteration of the system will additionally extract embedded images for processing by vision models, though this functionality was not implemented in the current version. The resulting document, referred to as a "source," is subsequently paired with a predefined prompt and processed by either Gemini-2.5-pro or GPT-5 to extract topic objectives that the generated questions should address. Questions are then generated using these topic objectives as contextual guidance. For each question, the large language model is instructed to select a relevant section from the original slide source as context and to assign an appropriate Bloom's taxonomy level prior to generating the question itself. Following question generation, a rubric containing grading instructions and a set of example answers are produced. These example answers are labeled by the language model as "Correct," "Partially Correct," or "Incorrect." Finally, the teacher reviews the generated questions and either approves them for student use or rejects them.

Human Verification New questions were reviewed by one of the authors (a domain expert) on a weekly basis, after which students were free to complete them at their own convenience. During the review, questions were mainly rejected for not being part of the source material. The LLMs used for generation already have a broad knowledge of the subject and so would be able to generate questions without context. However, this is undesirable, since the students would not necessarily be familiar with the concepts and there would be less certainty that the conveyed information is true.

4.2 Learning and Grading Pipeline

Each question was presented to the students individually, requiring students to submit a free-text response in Markdown format before proceeding. Student answers were graded by one of four large language models using one of four distinct prompting strategies, varying in whether they included the rubric and/or example answers. To optimize study efficiency, students could proceed to the next question immediately while awaiting feedback; however, submission of a new answer was delayed until feedback on the previous response had been received. For the purposes of this study, students were required to provide three-tiered feedback (Happy, Okay, or Unhappy) regarding the quality of the LLM-generated feedback before advancing to the subsequent question.

5 Evaluation

The overall goal of our evaluation is to analyse (1) question generation quality, (2) student answer and grading quality and (3) student engagement and learning.

5.1 Evaluation Design

Generated Questions To evaluate the quality of the generated questions, we manually annotate all questions generated from the source material. Questions for the first 9 weeks were generated with Gemini-2.5-pro and the next 5 weeks were generated with GPT-5 to evaluate if there are any differences between these two frontier models when it comes to generating questions. For the course, a total of 555 questions were generated. Of these, 198 (35.68%) were given to the students, while 357 (64.32%) kept back. We separately annotate all generated questions after the course had been finished to create an evaluation dataset to measure *question generation quality*. Specifically, we manually annotate (1) Grounding (i.e., if the question is grounded in the provided context during generation); (2) Duplicate from slides (i.e., if the question has been directly copied from the slide); (3) Redundancy (i.e., if the question redundant inside the same topic); (4) Difficulty (easy, medium, hard) and (5) Rejection reason (Accept, Reject: Redundant, Reject: Difficulty, Reject: Grounding, Reject: Incorrect Question, Reject: Incorrect Rubric, Reject: Unclear Question, Reject: Answer in Question)
Student Engagement and Answer Grading In order to evaluate the quality of the grading, we manually grade 400 answers. We sample the answers in the following manner: First, we select 5 questions per bloom's taxonomy level. This gives us 20 questions. Then, for each question, we select 5 graded answers for every LLM are selected. This gives us 5*4*20=400 question, answer, grade triplets. To measure student engagement, we track student progress over time (from course start to the exam). Further, we conduct an online survey after the conclusion of the teaching period, but before the exam. Overall 54 students participated in the survey.

5.2 RQ1: Quality of Generation (The Teacher Perspective)

The various generated questions were labeled by three human annotators and Figure 2 illustrates the rejection reasons given, which are defined as follows: **Reject: Difficulty** (inappropriate easy/hard level); **Reject: Hallucination** (generated irrelevant or hallucinated content); **Reject: Incorrect Question** (factually wrong or meaningless question); **Reject: Incorrect Rubric** (inappropriate rubric under topic); **Reject: Unclear Question** (ambiguous phrasing in question); **Reject: Answer in Question** (unveiled answer in question or self-answered question) During labeling, questions were also rejected for being redundant, which may be undesirable during a course to avoid having too many questions about the same subject, however for the analysis we treat these questions as "approved". Similarly, questions labeled as "Incorrect" were merged together with questions labeled as "Unclear". Note that on this data, the three human annotators achieved an average agreement rate of *76.37%* regarding whether to accept or reject each question.

Further, The two large language models exhibit noticeable differences when generating questions for an undergraduate course. Figure 3 shows that the perceived level of difficulty is higher for GPT-5, than it is for Gemini-2.5-pro.

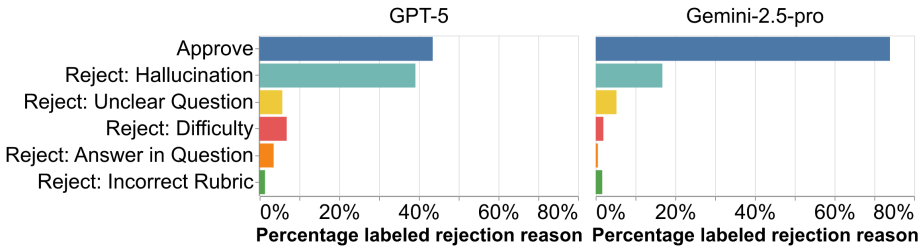

Fig. 2. Reason given by the annotators for why questions should be rejected. The questions generated by GPT-5 are more frequently rejected. This graph is adjusted for the number of questions each LLM has generated.

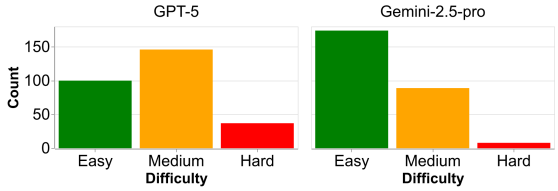

Fig. 3. Question difficulty as labeled by the annotators: GPT-5 generates more medium and hard questions compared to Gemini.

5.3 RQ2: Reliability of Grading (The Grader Perspective)

The four large language models analyzed for grading in this study were "qwen-3-32b" as a small open-source that could be provided on local school infrastructure, gemini-2.5-flash and gemini-2.5-pro to evaluate the effect of model scale on the performance and GPT-5 which was considered the State-of-the-art model with the most public usage at the time of writing. With Figure 4 we see that all models generate a similar distribution of grades, though there are notable differences, especially with Qwen-3-32b, which seems to generate more intermediate grades, rather than focusing on 0 (incorrect) and 1 (correct).

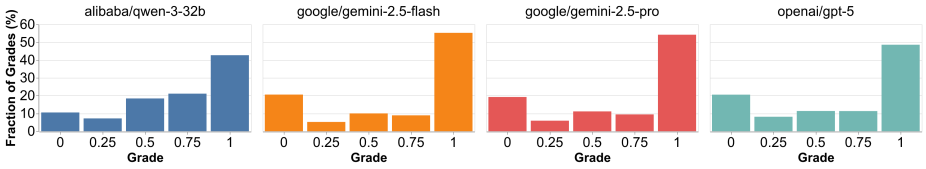

Fig. 4. Distribution of the grades (0-1) given by each LLM as fraction of all grades generated by the respective LLM.

Figure 5 highlights the differences between the human-given grades and the generated ones. We can see that a majority of data points lie on the diagonal or right next to it, indicating decent performance by the LLMs.

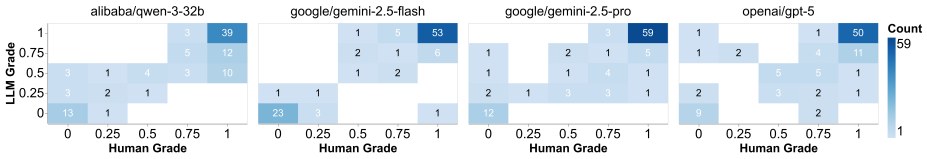

Fig. 5. Human grades given by annotators compared to LLM grades.

5.4 RQ3: Student Engagement & Learning (The Student Perspective)

The daily submission data reveals a distinct pattern in student engagement with the weekly quizzes. Throughout the semester, from September 2025 to early January 2026, submissions show a steady but minimal daily interaction. However, a surge in activity occurred in the final week before the exam in January, with submissions peaking at over 4,000 answers per day. This behavior pattern is often associated with completing assignments for summative credit rather than for continuous, formative learning.

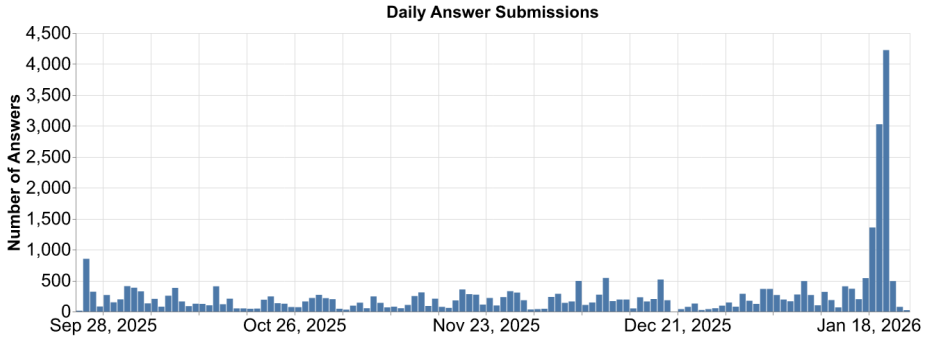

Fig. 6. Student engagement over time. A large number of answers were made shortly before the deadline.

The analysis of student feedback on the weekly quizzes, as depicted in Figure 7, reveals a generally moderate but varied perception. On a 6-point Likert scale, most aspects, including comprehension (M=3.12, SD=1.41), exam preparation (M=3.44, SD=1.72), clarity (M=3.42, SD=1.27), and the quality of explanations (M=3.47, SD=1.19), received mean scores slightly above the neutral midpoint. This suggests a moderately positive student sentiment. Notably, students perceived the quiz difficulty to be on the harder side (M=4.10, SD=1.04), while also finding the questions to be well-aligned with the lecture content (M=3.96, SD=1.43). The standard deviations indicate a considerable spread in student opinions, particularly regarding the utility of the quizzes for exam preparation.

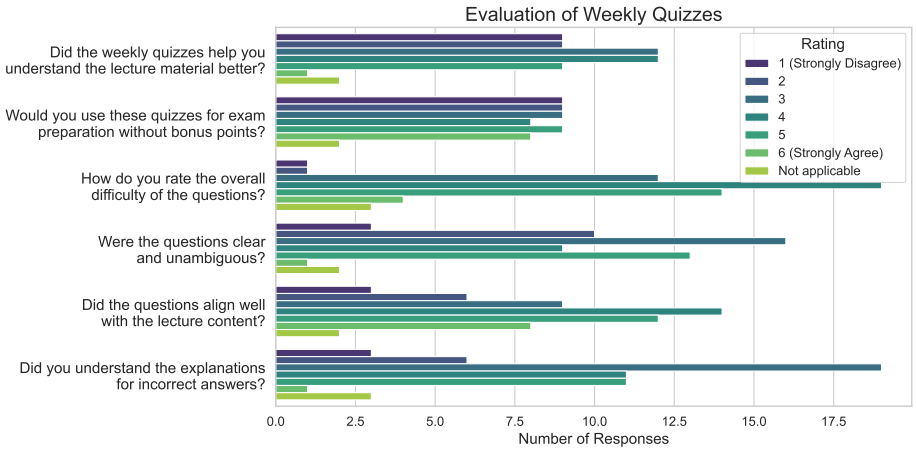

Fig. 7. Distribution of student responses to six questions evaluating the weekly quizzes. The questions cover comprehension, exam preparation, difficulty, clarity, content match, and the quality of explanations for incorrect answers. Responses were given on a 6-point Likert scale, where 1 represents strong disagreement and 6 represents strong agreement.

Figure 8 presents student feedback specifically on the AI-generated grading and feedback. The results indicate a high level of trust in the AI's accuracy. A significant majority of students reported that they perceived the AI feedback to be incorrect only Occasionally" (38.9%) or Very rarely" (35.2%). Only a small minority (1.9%) found the feedback to be Almost always incorrect." Regarding the grading standard, the AI was predominantly perceived as being stricter than human correctors. Over 72% of students found the grading to be either Somewhat too strict" (53.7%) or Much too strict" (18.5%), while only 16.7% of respondents felt the grading was Just right / Fair." No students reported that the grading was too lenient. This suggests that while the AI's feedback is trusted, its grading calibration may be stricter than what students are accustomed to from human graders.

6 Discussion

Our in-vivo deployment of an LLM-powered assessment system reveals critical insights into the practical challenges of bridging AI capabilities with pedagogical goals.

6.1 RQ1: The Need for Human-in-the-Loop in Question Generation

While LLMs can generate questions at scale, our findings underscore that ensuring pedagogical quality is not an automated process. A significant portion of generated questions were rejected for issues such as being ungrounded from

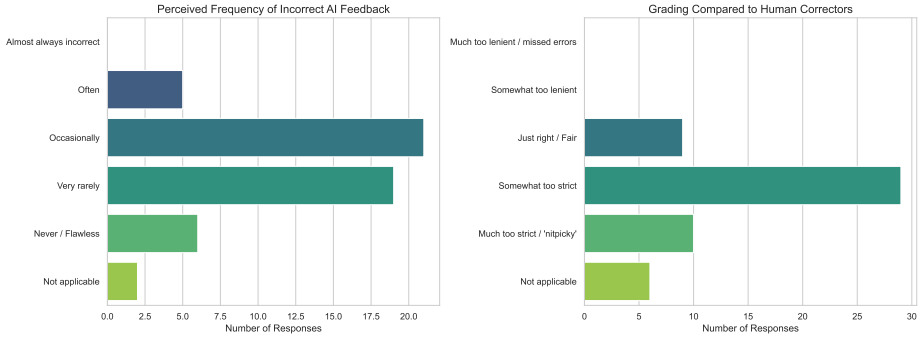

Fig. 8. Student feedback on the AI-generated grading. The left plot shows the perceived frequency of incorrect AI feedback. The right plot shows a comparison of the AI's grading strictness relative to human correctors.

course material, unclear, or inappropriately complex. We observed a trade-off where one model (GPT-5) produced more cognitively demanding questions but had a higher rejection rate, while another (Gemini-2.5-pro) created simpler, more reliable questions. This highlights the continued necessity of a human-in-the-loop to curate content, balancing cognitive challenge with curriculum alignment and clarity.

6.2 RQ2: The Disconnect Between AI Grading and Student Perception

Automated grading showed a nuanced reliability. While consistent with human experts on correct answers, LLMs struggled with the ambiguity of partially correct responses—a critical area for student learning. This inconsistency was mirrored in a disconnect between our data and student experience. Although some models tended to over-grade, over 72% of students perceived the AI as "too strict." This suggests that student perception is highly sensitive to even occasional harshness. Despite this, students reported high trust in the feedback's content, differentiating the score from the explanation. This, along with the observed "AI vs. AI" behavior of students using LLMs to answer questions, points to a complex and evolving interaction between students and automated assessment tools.

6.3 RQ3: The Failure of Formative Intent

Perhaps the most critical finding is the gap between our formative intent and actual student behavior. The overwhelming concentration of activity just before the exam deadline (Fig. 6) indicates that students treated the system as a summative hurdle, not a continuous learning tool. This "cramming" behavior, likely driven by the perceived difficulty of questions and strictness of grading, defeats the pedagogical purpose of retrieval practice. It demonstrates a crucial

tension: an effective automated formative system must not only be accurate but also be perceived as fair and supportive enough to encourage voluntary, regular engagement. Without this, even a technologically advanced system can fail to foster meaningful learning.

7 Conclusion

We deployed and evaluated an end-to-end LLM-based system for question generation and automated grading in a large undergraduate course. Our in-vivo study provides critical insights into the practical application of AI in formative assessment, revealing a significant gap between technological capability and pedagogical effectiveness.

While LLMs can automate the creation and grading of quizzes at scale—a task manually prohibitive—our findings highlight key challenges. The quality of generated questions was inconsistent, requiring human oversight to ensure relevance and clarity. In grading, LLMs struggled with partially correct answers, and while students trusted the feedback's content, they perceived the scoring as overly strict. Most importantly, student engagement patterns showed a "cramming" behavior for credit, rather than the intended continuous formative learning. This suggests that the system's design and integration into the course structure are as crucial as its technical accuracy.

Limitations Our findings are from a single course, which may not generalize. The study also lacked a human-in-the-loop for grading and did not include qualitative student interviews for deeper insights into their motivations.

Future Work Future work should focus on improving question quality control, developing hybrid human-AI grading models, and designing pedagogical strategies that encourage formative engagement. Investigating the "AI vs. AI" scenario, where students use LLMs to answer AI-generated questions, also remains a critical research direction.

References

1. Aggarwal, D., Sil, P., Raman, B., Bhattacharyya, P.: "i understand why i got this grade": Automatic short answer grading (asag) with feedback. In: International Conference on Artificial Intelligence in Education. pp. 304–318. Springer (2025)
2. Anderson, L.W.: A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives. Longman, New York, complete ed. edn. (2001)
3. Brown, P.C., Roediger III, H.L., McDaniel, M.A.: Make it stick: The science of successful learning. Harvard University Press (2014)
4. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 (2023)

5. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. International journal of artificial intelligence in education **25**, 60–117 (2015)
6. Butler, R.: Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. British Journal of Educational Psychology **58**(1), 1–14 (1988)
7. Camus, L., Filighera, A.: Investigating transformers for automatic short answer grading. In: Artificial Intelligence in Education: 21st International Conference, AIED 2020. pp. 43–48. Springer (2020)
8. Chang, L., Ginter, F.: Automatic short answer grading for finnish with chatgpt. Proceedings of the AAAI Conference on Artificial Intelligence **38**(21), 23173–23181 (Mar 2024). https://doi.org/10.1609/aaai.v38i21.30363, https://ojs.aaai.org/index.php/AAAI/article/view/30363
9. del Gobbo, E., Guarino, A., Cafarelli, B., Grilli, L., Limone, P.: Automatic evaluation of open-ended questions for online learning. a systematic mapping. Studies in Educational Evaluation **77**, 101258 (2023). https://doi.org/https://doi.org/10.1016/j.stueduc.2023.101258, https://www.sciencedirect.com/science/article/pii/S0191491X2300024X
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
11. Elkins, S., Kochmar, E., Cheung, J.C.K., Serban, I.: How useful are educational questions generated by large language models? (2023)
12. Filighera, A., Parihar, S., Steuer, T., Meuser, T., Ochs, S.: Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8577–8591. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.587, https://aclanthology.org/2022.acl-long.587
13. Grévisse, C.: Llm-based automatic short answer grading in undergraduate medical education. BMC Medical Education **24**(1), 1060 (2024). https://doi.org/10.1186/s12909-024-06026-5
14. Hattie, J., Timperley, H.: The power of feedback. Review of educational research **77**(1), 81–112 (2007)
15. Henkel, O., Vankadara, L.C., Nussbaumer, A., Diederich, J.: Can large language models make the grade? an empirical study evaluating llms ability to mark short answer questions in k-12 education. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale. pp. 184–195 (2024)
16. Lai, P., Zhang, K., Lin, Y., Zhang, L., Ye, F., Yan, J., Xu, Y., He, C., Wang, Y., Zhang, W., et al.: Sas-bench: A fine-grained benchmark for evaluating short answer scoring with large language models. arXiv preprint arXiv:2505.07247 (2025)
17. Meyer, G., Breuer, P., Fürst, J.: Asag2024: A combined benchmark for short answer grading. In: Proceedings of the 2024 on ACM Virtual Global Computing Education Conference V. 2. pp. 322–323 (2024)
18. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Lin, D., Matsumoto, Y., Mihalcea, R. (eds.) Proceedings of the 49th Annual Meeting

of the Association for Computational Linguistics: Human Language Technologies. pp. 752–762. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://aclanthology.org/P11-1076
19. Mohler, M., Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading. In: Proceedings of the 12th Conference of the ACL (EACL 2009). pp. 567–575. Association for Computational Linguistics, Athens, Greece (Mar 2009), https://aclanthology.org/E09-1065
20. Nielsen, R.D., Ward, W.H., Martin, J.H., Palmer, M.: Annotating students' understanding of science concepts. In: International Conference on Language Resources and Evaluation (2008), https://api.semanticscholar.org/CorpusID:12938607
21. Ozuru, Y., Briner, S., Kurby, C.A., McNamara, D.S.: Comparing comprehension measured by multiple-choice and open-ended questions. Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale **67**(3), 215 (2013)
22. 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)
23. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
24. Shute, V.J.: Focus on formative feedback. Review of educational research **78**(1), 153–189 (2008)
25. Steuer, T.: Automatic Question Generation to Support Reading Comprehension of Learners - Content Selection, Neural Question Generation, and Educational Evaluation. Ph.D. thesis, Technische Universität Darmstadt, Darmstadt (2023). https://doi.org/https://doi.org/10.26083/tuprints-00023032, http://tuprints.ulb.tu-darmstadt.de/23032/
26. Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., Arber, R.: Pre-training bert on domain resources for short answer grading. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 6071–6075 (2019)
27. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., hsin Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. ArXiv **abs/2206.07682** (2022), https://api.semanticscholar.org/CorpusID:249674500
28. Zhao, C., Silva, M., Poulsen, S.: Language models are few-shot graders. In: International Conference on Artificial Intelligence in Education. pp. 3–16. Springer (2025)