Avalon: A Human-in-the-Loop LLM Grading System with Instructor Calibration and Student Self-Assessment

Derek Armfield[1], Eason Chen[1], Asilbek Omonkulov, Xinyi Tang[1], Jionghao Lin[2], Erik Thiessen[1], and Kenneth Koedinger[1]

[1] Carnegie Mellon University, United States
[2] The University of Hong Kong, Hong Kong

Abstract. Open-ended assessments promote deeper learning but pose significant challenges for timely, high-quality feedback—often burdening instructors with lengthy grading processes and resulting in underutilized student feedback. We introduce Avalon, a human-in-the-loop AI grading system that integrates (1) an iterative calibration phase to align AI graders with instructor expectations and (2) a novel student self-grading and discrepancy reporting mechanism. Through rubric calibration, instructors provide corrective feedback on AI-graded samples, ensuring consistent application of grading criteria. After the AI grades all submissions, students assess their own work using the same rubric, then compare their scores with the AI's. They submit short "discrepancy reports" for any mismatches, distinguishing between accepted differences and genuine disputes. In a pilot with 102 undergraduates, Avalon reduced instructor grading time by focusing manual review on a small subset - fewer than 16% of submissions - that were disputed by students. Moreover, students show high engagement with the feedback process, as self-grading compelled them to revisit rubric criteria and reflect on their submissions more deeply. Furthermore, the system uncovered misconceptions that might otherwise have gone undetected—prompting targeted instructor intervention. Although additional validation and larger-scale studies are needed, current preliminary findings suggest Avalon's hybrid approach can reduce grading workloads, improve feedback effectiveness, and enhance student engagement and learning outcomes. The Avalon platform can be access at `https://avalonlearn.com`

Keywords: Large Language Models, Grading System, Feedback

1 Introduction and Related Works

Large enrollment courses and online learning platforms (e.g., MOOCs) [1] offer unprecedented access to education, but they also place immense grading burdens on instructors [6]. Open-ended assignments—ranging from short answers to lengthy essays or proofs—are particularly time-consuming to grade[11,13]. Ironically, despite the rich feedback instructors give, many students do not fully

read or act on these comments [18,16,5,23]. Prior research shows that only 58% of feedback gets accessed[16], implying lost opportunities for deeper learning.

Automated grading systems promise to reduce grading overhead [17] and deliver timely and personalized feedback [14,3], especially with recent advances in large language models (LLMs) [2,7,8,4]. However, completely delegating grading to AI raises concerns over accuracy, transparency, and accountability [22,10]. Even though researchers report that AI-generated grades can align human scores moderately well, doubts persist regarding whether AI can consistently evaluate nuanced student thinking [8]. These concerns highlighted the need for a *hybrid* approach that combines LLM capabilities with instructor oversight and more engaging mechanisms for students to receive feedback and raise disputes.

Recent research has shown that LLMs can achieve high agreement with human instructors on short-answer and essay grading [8,9], yet fully automated systems often lack transparency, ignore student agency, and cannot detect shared misunderstandings between AI and student. Recent hybrid approaches involving instructor–AI collaboration (e.g., rubric refinement [19]) and AI-assisted code grading with TA verification [20] show promise, but typically omit student involvement. Similarly, self-assessment tools have been shown to improve metacognition [21,12], but are rarely integrated with AI grading.

To address these gaps, we propose Avalon, a human-in-the-loop AI grading system that combines instructor calibration, large language model scoring, and structured student self-assessment. Rather than treating evaluation as something that happens to the student and their work, Avalon reimagines grading as a collaborative process with the student—transforming them from a passive consumer of feedback and evaluation into a key, active participant who both deepens their engagement and strengthens the accuracy of assessment.

We piloted Avalon in an undergraduate course with 102 students and 507 submissions. *Preliminary findings in the pilot study* indicate that

(a) **78% of students engage in receiving feedback**, as Avalon encourages them to revisit rubric criteria for self-grading, then resolve misalignment with the AI grader—whether by disputing or reflecting on their grading results.
(b) An estimated **reduction in grading time of over 85%** since the instructor only needs to review the fewer than 16% of submissions which were disputed.
(c) **Only 9 out of 1729 of the AI's judgements were overturned** by the instructor due to AI misinterpretation, suggesting strong alignment with the instructor's criteria.

These promising outcomes highlight Avalon's potential to ease instructor workloads while fostering more meaningful student feedback utilization.

2 Avalon System Implementation

Figure 1 illustrates how instructors and students engage with the grading agent and self-grading system. Specifically, they will follow the following steps:

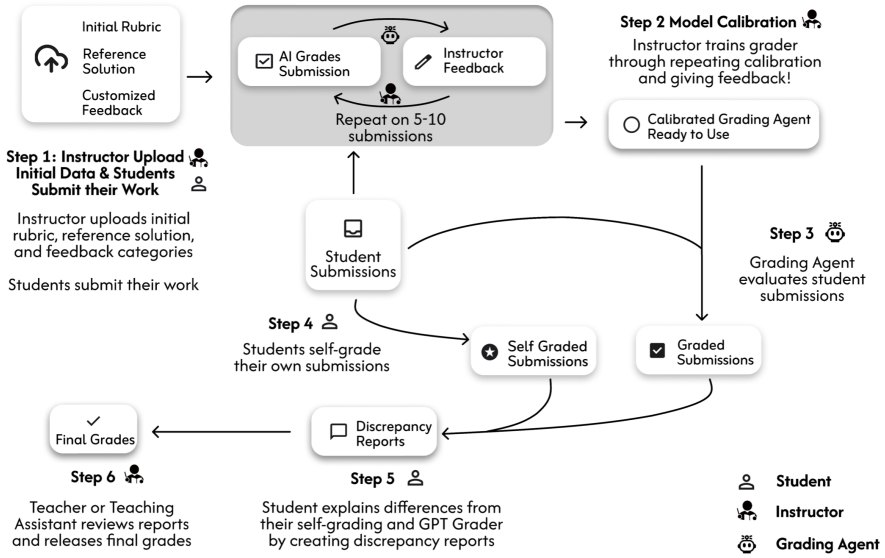

Fig. 1: User flow for Avalon. Please refer section 2 for explanation of each step.

Step 1: Instructor Uploads Material and Students Submit Their Work In order to initiate the grading process, several components must be in place. First, the instructor creates a preliminary rubric consisting of concise item descriptions paired with associated point values. This rubric serves as a tentative hypothesis for how the assignment should be graded and will be refined throughout the calibration procedure. Additionally, the instructor provides a reference solution, which offers the Grading Agent context regarding the expected solution. The instructor may also define custom feedback categories—for example, *strengths*, *weaknesses*, and *improvements*—or adapt these categories to address specific learning objectives. Crucially, students must submit their work so the instructor can begin the calibration process; however, it is possible to do the calibration process using only a subset of the total submissions.

Step 2: Rubric Calibration Rubric calibration ensures that the Grading Agent, powered by GPT-4o, interprets the rubric in accordance with the instructor's intended grading criteria via iterative prompt-engineering. This procedure, inspired by human grading practices often referred to as *normalization*, proceeds one submission at a time. Both the Grading Agent and the instructor independently evaluate the same submission. Whenever a discrepancy arises, the instructor provides corrective feedback. The Grading Agent then proposes adjustments to the rubric to prevent future misinterpretations. The instructor may accept this suggestion or modify the rubric further to align with their vision for how the assignment should be graded. This iterative process continues until the instructor is satisfied that the Grading Agent is consistently applying the rubric in agreement with the instructor.

Step 3: Grading Agent Evaluates All Submissions Once calibration is complete, the instructor has gained sufficient confidence that the Grading Agent can reliably apply the rubric. At this stage, the Grading Agent proceeds to evaluate all remaining student submissions. Points are awarded, and feedback is generated based on the finalized rubric.
Step 4: Students Submit Their Self-Grading After the Grading Agent has completed its evaluations, students are given the opportunity to grade their own work using the same rubric. Notably, they do not see the Grading Agent's assigned scores or feedback beforehand. This self-assessment requires students to interpret and apply the rubric to their own submissions, encouraging deeper engagement with the criteria used to evaluate their work.
Step 5: Student View Results and Create Discrepancy Report Following the self-grading phase, the Grading Agent's evaluation is revealed. The system compares the student's self-assessment with the agent's grading to identify any discrepancies for each rubric item. The student is encouraged to submit a *discrepancy report* for each identified mismatch. Two primary types of discrepancy reports occur:

1. **Self-Acknowledged**: The student concedes that their initial self-assessment was incorrect. In such cases, the student provides a brief explanation detailing why their original understanding differed from the Grading Agent's evaluation and clarifies what led to their revised perspective.
2. **Disputed**: The student maintains that their original assessment was correct, thus contesting the Grading Agent's evaluation. Here, the student argues why they believe the rubric item should (or should not) have been applied as initially self-assessed. Although rare, it is possible for a student to argue against awarding themselves positive points if they believe the Grading Agent misapplied the rubric.

Step 6: Instructor Review Reports and Release Final Grades Finally, the instructor and teaching assistants review the submitted discrepancy reports. While it may be sufficient to focus on the *disputed* reports, instructors can also examine *self-acknowledged* reports to gain insight into common student misunderstandings and adapt their teaching strategy in the future, such as emphasizing the misunderstanding part. When necessary, the instructor can adjust scores or offer clarifications. After any revisions are finalized, the instructor releases the official grades to the students.

3 Discussion with Preliminary System Evaluation Results

Preliminary system evaluation with 102 students from an undergraduate course at a university suggests several promising outcomes. Our study was approved by the IRB. This section highlights three key observations: (1) reductions in grading time, (2) improved delivery for feedback (3) high agreement for the AI grading result, and (4) increased student engagement facilitated by self-grading and self-reflection on potential misconceptions.

(1) Grading Time Reduction In total, 77.12% of all submissions received were self-graded by students. Among these, 54.48% (equivalent to 42.01% of all submissions) exhibited differences in at least one rubric item compared to the AI grader (note that each submission may have more than one rubric item with discrepancies). Among these differences, 66.20% (27.81% overall) led to the submission of a discrepancy report. Within these reports, 56.74% contained at least one disputed rubric item, implying that only 15.78% of all submissions ultimately required direct instructor intervention.

Because instructors only have to review submissions with a disputed rubric item (approximately 15%), Avalon can potentially reduce instructor grading time by around 85%. In other words, an assignment that would normally take 20 hours of human grading time could now be finished in less than 3 hours. This estimated reduction may even be conservative since instructors typically only have to resolve the disputed rubric items – often just one out of four rubric items – and resolving these reports was reportedly 2-3x faster (from instructor interview) than grading from scratch. That same 20 hours of human grading time could be reduced to just 1 hour. Since calibration's impact on overall grading time is minimal as it does not scale with class size, is already relatively quick (under one hour), and can be expected to become even faster as instructors gain familiarity with the system, it has been omitted from these estimates.

(2) Improved Deliverability of Feedback Previous studies have shown that up to 42% of feedback goes unaccessed when grades can be viewed elsewhere [16]. In contrast, our system requires students to complete a self-grading exercise before viewing any instructor or AI-generated comments. This additional step increases feedback accessed: only 22.28% of the feedback goes unaccessed despite the extra effort required from students.

(3) High Agreement, and Misconception Detection We analyzed all submissions that were graded both by the Grading Agent and by the students themselves. These comprised five questions, a total of 391 submissions, which include 1332 rubric items assessed by both the AI system and the students. We observed an overall alignment of 78% between the student-self-grading and AI Agent's scores on rubrics. Then, when students compared their self-assessments with the AI's evaluations, we observed that 93% of student evaluations agreed with the AI's judgment, resulting in a Kappa of 0.84, indicating strong agreement [15]. That being said, it is possible that both the student and the AI are wrong while in agreement. Future work would then undergo a human review for all submissions to mitigate this issue.

Although some students still disagreed with AI's grading and submitted discrepancy reports to dispute it, most of the discrepancy reports were ultimately attributed to student misunderstandings of the rubric by the instructor, like the example illustrated in Figure 2. Out of 1729 AI grading judgments, only 9 were overturned due to AI misinterpretation. This result underscores a high level of consistency between the AI grader and the instructor's assessments.

Furthermore, the self-grading component appeared pivotal in uncovering misconceptions and promoting engagement. By requiring students to explicitly as-

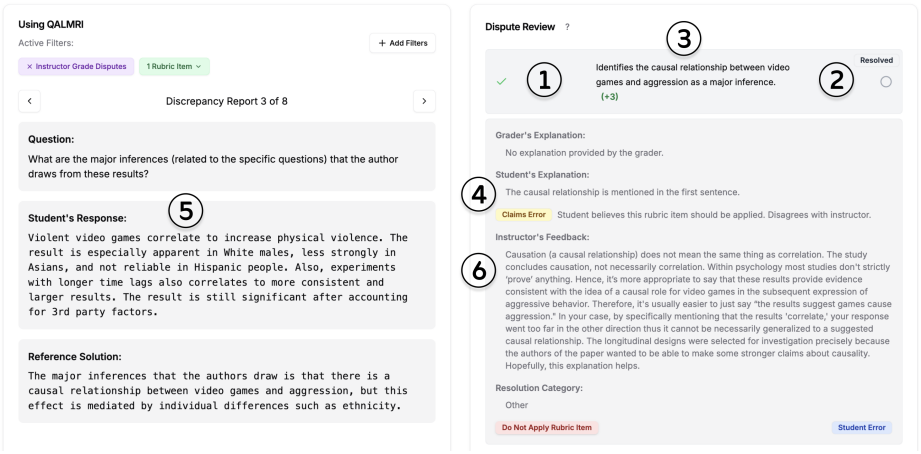

Fig. 2: A screenshot from Avalon's interface illustrates a student's misunderstanding. The green check (1) shows the student believed the rubric item should be applied, whereas the grey circle (2) indicates the Grading Agent did not apply it. Despite this, the student argued their response met the rubric item's criteria (3) of demonstrating a causal relationship as evidenced by their first sentence (4), although it actually only mentioned a correlation (5). The instructor resolved the dispute by not applying the rubric item and providing targeted feedback (6), highlighting how such misunderstandings can be identified and clarified.

sess their own work using the same rubric, this system prompted students to reflect on their reasoning processes and confront any incongruities with the AI's evaluations. Instructors could then focus their efforts on resolving the most substantive disputes, where students genuinely misunderstood the material or rubric criteria, thus enabling more targeted and impactful teaching and feedback.

4 Conclusion

In this paper, we introduce *Avalon* – a platform that seeks to rethink the student's role in feedback and evaluation, reduce time spent on grading open-ended assignments, and provide consistency, transparency, and interpretability to AI grading systems. By having students independently self-grade their own submissions, comparing this self-grade to the AI grading, submitting short explanations on why any item may differ, and then having instructors respond to such reports with tailored feedback, *Avalon* manages to drastically increase student engagement, reflection, and interaction with feedback. The iterative calibration process, having the agent grade and the instructor providing corrective feedback, helps align the AI Grading Agent to human grading standards.

References

1. Baturay, M.H.: An overview of the world of moocs. Procedia-Social and Behavioral Sciences **174**, 427–433 (2015)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Chen, E., Lee, J.E., Lin, J., Koedinger, K.: Gptutor: Great personalized tutor with large language models for personalized learning content generation. In: Proceedings of the Eleventh ACM Conference on Learning@ Scale. pp. 539–541 (2024)
4. Chen, E., Wang, D., Xu, L., Cao, C., Fang, X., Lin, J.: A systematic review on prompt engineering in large language models for k-12 stem education. arXiv preprint arXiv:2410.11123 (2024)
5. Crisp, B.R.: Is it worth the effort? how feedback influences students' subsequent submission of assessable work. Assessment & evaluation in higher education **32**(5), 571–581 (2007)
6. Dunlap, J.C.: Workload reduction in online courses: Getting some shuteye. Performance Improvement **44**(5), 18–25 (2005)
7. Figueras, C., Farazouli, A., Cerratto Pargman, T., McGrath, C., Rossitto, C.: Promises and breakages of automated grading systems: a qualitative study in computer science education. Education Inquiry pp. 1–22 (2025)
8. Flodén, J.: Grading exams using large language models: A comparison between human and ai grading of exams in higher education using chatgpt. British Educational Research Journal (2024)
9. Gobrecht, A., Tuma, F., Möller, M., Zöller, T., Zakhvatkin, M., Wuttig, A., Sommerfeldt, H., Schütt, S.: Beyond human subjectivity and error: a novel ai grading system. arXiv preprint arXiv:2405.04323 (2024)
10. Idris, M.D., Feng, X., Dyo, V.: Revolutionising higher education: Unleashing the potential of large language models for strategic transformation. IEEE Access (2024)
11. Jerrim, J., Sims, S.: Teacher workload and well-being. new international evidence from the oecd talis study. Teaching and Teacher Education (2020)
12. Khojasteh, L., Kafipour, R., Pakdel, F., Mukundan, J.: Empowering medical students with ai writing co-pilots: design and validation of ai self-assessment toolkit. BMC Medical Education **25**, 159 (2025). `https://doi.org/10.1186/s12909-025-06753-3`
13. Kreuzfeld, S., Felsing, C., Seibt, R.: Teachers' working time as a risk factor for their mental health-findings from a cross-sectional study at german upper-level secondary schools. BMC Public Health **22**(1), 307 (2022)
14. Lin, J., Chen, E., Han, Z., Gurung, A., Thomas, D.R., Tan, W., Nguyen, N.D., Koedinger, K.R.: How can i improve? using gpt to highlight the desired and undesired parts of open-ended responses. arXiv preprint arXiv:2405.00291 (2024)
15. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)
16. Mensink, P.J., King, K.: Student access of online feedback is modified by the availability of assessment marks, gender and academic performance. British Journal of Educational Technology **51**(1), 10–22 (2020)
17. Messer, M., Brown, N.C., Kölling, M., Shi, M.: Automated grading and feedback tools for programming education: A systematic review. ACM Transactions on Computing Education **24**(1), 1–43 (2024)

18. Murtagh, L., Baker, N.: Feedback to feed forward: Student response to tutors' written comments on assignments. Practitioner research in higher education **3**(1), 20–28 (2009)
19. Myint, P.Y.W., Lo, S.L., Zhang, Y.: Harnessing the power of ai-instructor collaborative grading approach: Topic-based effective grading for semi open-ended multipart questions. Computers and Education: Artificial Intelligence **7**, 100339 (2024). https://doi.org/10.1016/j.caeai.2024.100339
20. Nagakalyani, G., Chaudhary, S., Apte, V., Ramakrishnan, G., Tamilselvam, S.: Design and evaluation of an ai-assisted grading tool for introductory programming assignments: An experience report. In: Proceedings of the 56th ACM Technical Symposium on Computer Science Education (SIGCSE) (2025)
21. Nieminen, J.H., et al.: Self-assessment design in a digital world: centring student agency. Assessment & Evaluation in Higher Education (2025). https://doi.org/10.1080/02602938.2025.2467647
22. Quttainah, M., Mishra, V., Madakam, S., Lurie, Y., Mark, S., et al.: Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education: Narrative review and qualitative study. JMIR AI **3**(1), e51834 (2024)
23. Woods, D.: Students viewing of feedback: An exploration of technology-mediated learning. Journal of Educational Technology Systems **51**(1), 46–62 (2022)