The Synergy Between Using LLMs to Generate Feedback based on Rubrics and Using LLMs to Moderate Content based on Governance Rules

Xinyi Lu
University of Michigan
Ann Arbor, United States
lwlxy@umich.edu

Xu Wang
University of Michigan
Ann Arbor, United States
xwanghci@umich.edu

Abstract

High-quality feedback is crucial for learning but requires expertise and effort. This project examines the prospect of using AI-generated feedback as suggestions to expedite and enhance human instructors' feedback provision. We situate our work in an introductory Economics class which has short-essay assignments. We developed an LLM-powered feedback engine that generates feedback on students' essays based on the grading rubrics used by the instructors and teaching assistants (TAs), and presented the feedback as in-text comments within student submissions. We then performed think-aloud studies with 5 TAs over 20 1-hour sessions to have them evaluate the AI feedback, and share how they envision using the AI feedback if they were offered as suggestions. Our preliminary findings suggest that AI feedback encouraged TAs to consider more perspectives, fostering critical thinking and rubric alignment. Participants also envisioned AI assistance improving feedback consistency. Then we discuss the generalizability of our findings to policy enforcement in various domains, such as content moderation. We also emphasize the important role of clear rubrics adapted for AI and discuss broader implications in system designs for enhancing AI explainability.

CCS Concepts

• **Applied computing** → **Computer-assisted instruction**.

Keywords

Automated feedback generation, Human-AI partnership, Policy enforcement

ACM Reference Format:
Xinyi Lu and Xu Wang. 2025. The Synergy Between Using LLMs to Generate Feedback based on Rubrics and Using LLMs to Moderate Content based on Governance Rules. In *Proceedings of Workshop on Sociotechnical AI Governance: Opportunities and Challenges for HCI (STAIG CHI '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

STAIG CHI '25, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Extensive research has shown that feedback is important for learning [4, 7, 10, 11, 13, 16]. Studies have suggested including detailed rubrics to improve feedback quality [1, 3, 6, 22, 23, 29], yet high quality feedback requires expertise and efforts to write [21, 22, 24]. Since the rise of generative AI, research communities around AI and Education have explored using large language models (LLMs) to generate tutoring responses and feedback. Studies showed promising results that when instructed well, LLMs can generate high quality feedback comparable to human feedback [9, 12, 14, 19, 26]. However, many studies observe problems in LLM feedback such as they can be overly general [14], cannot capture nuanced differences in students' answers [12, 14, 25], and produce mistakes[15]. Most existing research has focused on developing techniques to align LLM-generated feedback with human feedback using quantitative metrics such as accuracy, recall, and linguistic similarity to human feedback [2, 17]. In contrast, this study investigates the potential for human-AI collaboration in feedback provision. We aim to address the question: even when AI feedback is imperfect, can it be used as suggestions to expedite and enhance human instructors' feedback provision?

We conducted our study in an introductory Economics class in Fall 2024. It has frequent knowledge-intensive short-essay assignments with detailed rubrics developed by the instructor, and Teaching Assistants (TAs) provide feedback to students' essays based on these rubrics. We developed a feedback engine to generate feedback regarding each rubric with decomposed steps, and displayed them as in-text comments in a Word document. We then performed think-aloud studies with 5 TAs over 20 1-hour sessions to have them evaluate the AI feedback, and share how they envision using the AI feedback as suggestions. Our preliminary findings suggest that AI feedback aligns better with the characteristics of effective feedback as instructured in the prompts to the LLM, but might contain mistakes when the rubrics are not well written. Moreover, TAs considered the AI feedback to foster critical thinking, and align more closely with the rubrics. Participants also envision that having AI assessments could help improve the consistency among TAs. Based on these insights, we plan to conduct a deployment study to assess whether having AI assessments while grading would influence the feedback TAs provide.

This paper discusses the synergy between using AI to generate feedback based on rubrics, and using AI to moderate content based on governance rules. In particular, we discuss insights on what makes a good "rubric", which may have implications on what makes good "governance rules". Policies and rules are established to populate standardization, but the execution of the policy is distributed,

and discrepancies arise both between policymakers and enforcers, and between enforcers themselves [5, 8]. Our findings show the potential of providing AI-generated assessments as suggestions to human moderators, which might improve the consistency of rule enforcement and enhance policy comprehension. We offer recommendations on the system design in enhancing AI explainability, including decomposing the task into subtasks and providing intermediate outputs to improve transparency and user control. Moreover, our findings emphasize the importance of clearly-written rubrics for AI systems and provide suggestions on adapting the existing rubrics written for human to AI to ensure the assessments are aligned with expectations.

2 Methods

We conducted our study in a college-level introductory Economics class in Fall 2024. The course included frequent short-essay assignments with detailed rubrics developed by the instructor, and TAs provide feedback based on the rubrics. For each student response, AI judgements and feedback are generated separately for each rubric provided. The generation method follows the idea of Chain-of-Thought [27], structured into three steps, including 1) identifying the relevant sentences in the student responses to the rubric; 2) making a judgement on whether the student response meets the rubric; 3) constructing feedback. Specifically, we requested a rationale before making judgments and feedback. All the feedback was generated by GPT-4o with a temperature of 0.05 to enhance consistency. To visualize the feedback, we developed a Word plugin to integrate the generated judgments and feedback as in-text comments on a document. We also highlighted the relevant sentences identified by AI in the generation, and as the anchoring text for the comments.

We conducted 20 1-hour think-aloud sessions with 5 TAs. To create an authentic environment for TAs to critically compare their feedback with the AI's feedback, we first asked the TAs to complete their grading tasks as usual. After they finished, we invited them to participate in think-aloud sessions where they were asked to reviewe AI-generated feedback on the essay they graded, contrast the AI feedback with their handwritten feedback, and share how they envision using the AI feedback if they were offered as suggestions. The study is IRB-approved. The sessions were conducted via Zoom, and the de-identified transcripts of the recordings were analyzed with affinity diagram [20]. Participants were compensated with a $25 Gift Card for each study session.

3 Preliminary Findings

3.1 AI feedback aligns more closely with the rubrics, while TA's feedback is more holistic

Participants appreciated that the AI feedback was better aligned with the rubrics and more fine-grained since the feedback engine generates one feedback message per rubric item. They also reported that AI feedback exhibits more characteristics of effective feedback, such as personalized and localized language, providing praise, using guiding questions and explanations. In contrast, TAs often synthesize multiple items into a single, more comprehensive comment. However, this brings about the trade-off that AI feedback could be misleading when the rubrics are not well written.

We will describe two scenarios where AI tends to make mistakes. First, AI struggles with assessing definitions of specialized economic terms not explicitly covered in the rubric. Second, AI tends to rigidly enforce rubric phrasing, often rejecting valid alternative expressions. For instance, a decrease in demand can also be expressed as "a demand curve shifts leftwards" or "a demand curve shifts downwards" or "a reduction in the consumers' willingness to pay for the good". Both P4 and P5 noted that although some students captured the core idea, AI marked them incorrect for deviating from rubric language.

In contrast, TAs adopt a more holistic approach. Rather than addressing rubric items in isolation, they often prioritize higher-level qualities such as conceptual understanding and argument coherence. For example, participants shared cases in which students might be struggling with deeper conceptual issues that go beyond merely missing a single rubric. In such cases, they found the AI feedback focusing on a single rubric to be insufficient. They also noted that some rubric items may not apply to every essay, and some errors fall outside the rubric's scope. For example, the students are required to point out the "third party" in the negative externalities and explain how they don't have a say in the market. However, some students identified animals or the environment as the third party, making it unnecessary to explain why these parties lack a voice in the markets. Since these nuances are challenging to fully capture within rubrics, these edge cases highlight the limitations of AI feedback and the importance of including human judgments.

3.2 Seeing example essays with AI feedback improve understanding in the rubrics

Although no separate learning process is provided, participants find themselves having a deeper understanding in the rubrics as they read through the sample essays and the AI feedback. For example, P4 mentioned that through seeing the AI feedback and the AI highlights, they get a better understanding of the typical order in which students addressed rubric criteria, which helped them identify missing key points when evaluating other student responses later.

More importantly, AI feedback prompted participants to notice overlooked rubrics, and evaluate them more critically. For instance, both P4 and P5 realized they had overlooked certain rubric criteria in their evaluations, which the AI had identified. Consequently, this process led participants to take into account nuanced student examples and different perspectives represented in the AI feedback, improving their understanding of the rubrics. As P1 noted, *"reading through the AI Feedback gave me a better understanding of what I should be looking for."* A notable example involves the rubric item "Point out the third party in the article and explain why they don't have a say." P5 initially missed the first half of this criterion. As they noticed AI marking the student response as incomplete due to the lack of an explicitly identified third party, they spent more time reading the highlighted sentences and reconsidering their judgment. This prompted them to spend more time interpreting the rubric item, ultimately agreeing with the AI's assessment and recognizing their initial misunderstanding. As P5 reflected, *"it also helped me understand ... what was expected out of the assignment."*

3.3 Having AI feedback could help standardize feedback

Although the TAs scheduled regular staff meetings with the instructor to understand the rubrics and ensure they applied the rubrics consistently, there is still inconsistency when applying them, and TAs expressed concerns about that. Participants identified two primary sources of inconsistency: within the same TA and between different TAs.

Firstly, participants expressed concerns about inconsistencies within their own feedback. For example, P1 said, *"I also worry about consistency. Like, if I take 1 point off for one student here. Did I take off 2 points for another student?"*. This issue is particularly pronounced when a student's response implies a correct answer but lacks full accuracy. Several participants found AI useful in helping them verifying their evaluations and consistently considering all the rubrics. As P2 noted, *"It's just like having another set of eyes on the paper."* and P5 said AI feedback *"helpful to have standardization within your own section"*. P5 also mentioned the AI feedback could help them improve consistency in their feedback, as AI would check the same list of rubrics for all the students. P5 said, *"sometimes I would question if I left the same feedback for all of the students like.... (With AI,) I was not having that worry at all, because I was just looking through a list of items and checking to see if students had them. And so it wasn't like subjective in that way anymore."*

Secondly, participants also found inconsistency among different TAs due to various understandings in the requirements. As P5 said, *"Something that we are always concerned about is that one [TA] is grading too leniently versus other [TA] that's grading harshly."* One main reason is that some rubrics lack specificity in the depth and accuracy in analysis needed. For example, one rubric item requires "provides a thoughtful and well-reasoned solution". P5 expected a thorough reasoning of why the solution alleviates the problem, while P2 found pointing out a valid solution to be sufficient. Participants envision that AI feedback could help them build consistency in grading among the TAs, as the AI feedback is capable of providing more objective and consistent judgement and feedback. P5 said, *"I think this would be helpful, and we (TAs) would be making sure that we're taking into account similar things."*

4 Discussion

4.1 Implications on other demains

Although this work is conducted in the context of feedback provision only, our findings provide broader implications for policy development and enforcement beyond Education, especially in areas where human moderators or decision-makers have to interpret and apply predefined rules. Although policies and rules have been established in various domains to standardize decision-making, discrepancies often arise between policymakers and enforcers, leading to unintended leniency or excessive penalties [5, 8]. Additionally, differences in interpretation among multiple enforcers can further undermine standardization, resulting in inconsistent outcomes [8, 23, 30].

One example domain is content moderation on online platforms, such as social media networks and forums, where moderators enforce policies designed to regulate harmful, misleading, or inappropriate content. Similar to providing feedback based on rubrics, moderators work with content guidelines to make case-by-case decisions about whether a post violates community rules. However, discrepancies in interpretation among different moderators can lead to inconsistent enforcement, where similar cases receive varying outcomes. Our finding suggests that although AI suggestions could be misleading when the rubric is not clearly-written, providing AI-generated judgments could provide two potential benefits. Firstly, the AI suggestions provide a more objective set of opinions, which could improve the consistency in moderation both within and between moderators. Secondly, seeing the AI suggestions could encourage moderators to reflect on ambiguous cases, consider alternative perspectives from different stakeholders. Similar to how TAs in our study benefited from engaging with AI suggestions, content moderators might also develop a deeper grasp of policies by critically evaluating AI-generated assessments.

Moreover, in communities with decentralized moderation models, where policies are established through appeal processes with an open committee of stakeholders, viewing examples and AI suggestions could potentially facilitate stakeholders assess the practical consequences of proposed rules, understand it more critically by considering different cases and perspectives from other stakeholders, and even refine it accordingly.

However, one key premise for our positive findings in feedback provision is that the TAs are familiar with domain knowledge. Thus, they find identifying the key information for making the decision more time-consuming than making the judgement itself. In that case, evaluating an AI feedback when provided with reasonings is easier than creating one from scratch. This assumption may not hold universally across all policy enforcement areas, where moderators may have varies level of expertise. Future research is needed to determine which domains align with this premise.

4.2 Providing intermediate outputs on subtasks to improve transparency and control

In this work, we decomposed the feedback generation task into subtasks including identifying relevant sentences, making judgments and constructing feedback, and provided intermediate outputs of each AI subtask. Many participants reported that these outputs, especially the highlighted text, helped them better understand and verify the AI's reasoning. This decomposition gives users better flexibility and control on what to take from AI. For example, AI might make mistakes on evaluating whether the rubric is satisfied. With the highlighted sentence, the user can easily flip the judgment. As P4 said, "even if it's sometimes incorrect, that's what you can check ... I think that's the easiest part for us to get." Our findings also suggest that participants were aware of AI's potential hallucinations and inaccuracies, and the visibility of intermediate outputs helped them better evaluate the correctness of AI suggestions. Future work on AI systems could also consider decomposing the task, presenting intermediate AI output, and localize the reasonings to enhance AI explainability.

4.3 Establishing clearly written rubrics

Our study highlights the importance of clearly written rubrics for enabling LLMs to generate accurate judgments and high-quality feedback. We found that rubrics designed for human graders often

Table 1: Suggestions for elaborating rubrics in order for LLMs to generate accurate feedback for knowledge-intensive essays.

	Good rubric example	Bad rubric example
Explain the domain-specific knowledge	The student demonstrated an understanding of the Law of Demand, that is, as the price of the good increases, the quantity demanded by the good or service decreases.	The student correctly used the terms quantity supplied/demanded vs. supply/demand.
Include acceptable alternatives	The student demonstrated that farmers demand water, or analyze the influence on farmers as consumers of water.	The student stated that with tax or ban on automation, the demand for labor increases.
Specify the expected depth of the explanation	The student explained why deadweight loss exists and mention it is quite large given that the Government purchased the excess.	Explain the concept of artificially scarce goods conceptually.
Negative behaviors should be explicity called out	The student did not use long direct quotes (more than 1 sentence in one quote) from the article.	Direct in-text references are present.

lack the explicit detail needed for LLMs to interpret evaluation criteria correctly, aligning with findings from [28]. We provide tips in Table 1 on elaborating rubrics to make them more understandable by LLMs. Since LLMs are sensitive to the exact phrasing of the prompt [18], rubric descriptions need to be precise, especially regarding the expected depth of explanation. For example, one rubric requires the students to "point out the third party in the negative externalities", but the instructor accepted implicit understanding, whereas the LLM penalized responses that did not explicitly state who the third party is. As the AI judgement was sticking to the description, it was overly strict. We revised the rubric to read "show understanding of who the third party is in the negative externality," aligning it better with both human and AI interpretations.

References

[1] Sandra Allen and John Knight. 2009. A Method for Collaboratively Developing and Validating a Rubric. *International Journal for the Scholarship of Teaching and Learning* 3, 2 (2009), n2.

[2] Afnan Almegren, Hassan Saleh Mahdi, Abduljalil Nasr Hazaea, Jamal Kaid Ali, and Rehan Megren Almegren. 2024. Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International* (2024), 1–16.

[3] Heidi L Andrade and Ying Du. 2005. Student perspectives on rubric-referenced assessment. (2005).

[4] John Bitchener, Stuart Young, and Denise Cameron. 2005. The effect of different types of corrective feedback on ESL student writing. *Journal of second language writing* 14, 3 (2005), 191–205.

[5] Paul Cairney, Siabhainn Russell, and Emily St Denny. 2016. The 'Scottish approach'to policy and policymaking: what issues are territorial and what are universal? *Policy & Politics* 44, 3 (2016), 333–350.

[6] Maria Asun Cantera, María-José Arevalo, Vanessa García-Marina, and Marian Alves-Castro. 2021. A rubric to assess and improve technical writing in undergraduate engineering courses. *Education Sciences* 11, 4 (2021), 146.

[7] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[8] MacKenzie F Common. 2020. Fear the reaper: How content moderation rules are enforced on social media. *International Review of Law, Computers & Technology* 34, 2 (2020), 126–152.

[9] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 323–325.

[10] John Hattie and Shirley Clarke. 2018. *Visible learning: feedback*. Routledge.

[11] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[12] Hasnain Heickal and Andrew Lan. 2024. Generating feedback-ladders for logical errors in programming using large language models. *arXiv preprint arXiv:2405.00302* (2024).

[13] Catherine M Hicks, Vineet Pandey, C Ailie Fraser, and Scott Klemmer. 2016. Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 458–469.

[14] Qinjin Jia, Jialin Cui, Haoze Du, Parvez Rashid, Ruijie Xi, Ruochi Li, and Edward Gehringer. 2024. LLM-generated Feedback in Real Classes and Beyond: Perspectives from Students and Instructors. In *Proceedings of the 17th International Conference on Educational Data Mining*. 862–867.

[15] Qinjin Jia, Jialin Cui, Ruijie Xi, Chengyuan Liu, Parvez Rashid, Ruochi Li, and Edward Gehringer. 2024. On Assessing the Faithfulness of LLM-generated Feedback on Student Assignments. In *Proceedings of the 17th International Conference on Educational Data Mining*. 491–499.

[16] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[17] Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence* 6 (2024), 100210.

[18] Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine* 388, 13 (2023), 1233–1239.

[19] Zhiping Liang, Lele Sha, Yi-Shan Tsai, Dragan Gasevic, and Guanliang Chen. 2024. Towards the automated generation of readily applicable personalised feedback in education. In *International Conference on Artificial Intelligence in Education*. Springer, 75–88.

[20] Bill Moggridge and Bill Atkinson. 2007. *Designing interactions*. Vol. 17. MIT press Cambridge.

[21] Melissa M Nelson and Christian D Schunn. 2009. The nature of feedback: How different types of peer feedback affect writing performance. *Instructional science* 37 (2009), 375–401.

[22] Melissa M Patchan, Christian D Schunn, and Richard J Correnti. 2016. The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology* 108, 8 (2016), 1098.

[23] Stefan K Schauber, Anne O Olsen, Erik L Werner, and Morten Magelssen. 2024. Inconsistencies in rater-based assessments mainly affect borderline candidates: but using simple heuristics might improve pass-fail decisions. *Advances in Health Sciences Education* 29, 5 (2024), 1749–1767.

[24] Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, and Christoph Meinel. 2016. Improving the peer assessment experience on MOOC platforms. In *Proceedings of the third (2016) ACM conference on Learning@ Scale*. 389–398.

[25] Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. 2024. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction* 91 (2024), 101894.

[26] Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2174–2199.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[28] Xuansheng Wu, Padmaja Pravin Saraf, Gyeong-Geon Lee, Ehsan Latif, Ninghao Liu, and Xiaoming Zhai. 2024. Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring. *arXiv preprint arXiv:2407.18328* (2024).

[29] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1005–1017.

[30] Mengxue Zhang, Neil Heffernan, and Andrew Lan. 2023. Modeling and Analyzing Scorer Preferences in Short-Answer Math Questions. *arXiv preprint arXiv:2306.00791* (2023).