Exploring LLM-Generated Feedback for Economics Essays: How Teaching Assistants Evaluate and Envision Its Use

Xinyi Lu(✉), Aditya Mahesh, Zejia Shen, Mitchell Dudley, Larissa Sano, and Xu Wang

University of Michigan, Ann Arbor, MI 48109, USA
{lwlxy,mahesha,jasnshen,mrdudley,llubomud,xwanghci}@umich.edu

Abstract. This project examines the prospect of using AI-generated feedback as suggestions to expedite and enhance human instructors' feedback provision. In particular, we focus on understanding the teaching assistants' perspectives on the quality of AI-generated feedback and how they may or may not utilize AI feedback in their own workflows. We situate our work in a foundational college Economics class, which has frequent short essay assignments. We developed an LLM-powered feedback engine that generates feedback on students' essays based on grading rubrics used by the teaching assistants (TAs). To ensure that TAs can meaningfully critique and engage with the AI feedback, we had them complete their regular grading jobs. For a randomly selected set of essays that they had graded, we used our feedback engine to generate feedback and displayed the feedback as in-text comments in a Word document. We then performed think-aloud studies with 5 TAs over 20 1-h sessions to have them evaluate the AI feedback, 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 highlights the importance of providing detailed rubrics for AI to generate high-quality feedback for knowledge-intensive essays. TAs considered that using AI feedback as suggestions during their grading could expedite grading, enhance consistency, and improve overall feedback quality. We discuss the importance of decomposing the feedback generation task into steps and presenting intermediate results, in order for TAs to use the AI feedback.

Keywords: Automated feedback generation · Large-language models · Human-AI partnership

1 Introduction

Extensive research has shown that feedback is important for learning [4,9,10,12,16], yet high quailty feedback requires expertise and efforts to write [24,25]. Since the rise of generative AI, research communities around AI and Education have explored using large language models (LLMs) to generate tutoring responses and

A. Mahesh and Z. Shen—Contributed equally to this work.

A. I. Cristea et al. (Eds.): AIED 2025, LNAI 15878, pp. 392–406, 2025.
https://doi.org/10.1007/978-3-031-98417-4_28

feedback. A number of studies showed promising results that when instructed well, LLMs can generate high quality feedback that is similar to human feedback [5,11,14,29]. However, many studies observe problems in LLM feedback such as they can be too general [14], cannot capture nuanced differences in students' answers [11,14,28], hallucinate and make mistakes [14]. In this work, 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?

This work aims to address key gaps in our understanding of LLMs' capabilities in generating effective feedback. First, prior work on LLM-based feedback generation has primarily focused on single-turn, single-rubric evaluations of short answers in subjects such as math and programming. Several studies have examined using LLMs to generate tutoring moves for math problems [11,22,27,29], while others have examined their use in providing feedback to human tutors on tutoring strategies, such as encouraging praise [15,32]. Some research has investigated feedback generation for longer texts, including essays for English learners [7,28] and project reports [14]. However, these studies primarily assess language features, with limited exploration of LLMs' effectiveness in evaluating knowledge accuracy in longer essays. To address this gap, this work examines LLM-generated feedback for knowledge-intensive essays in a college-level introductory economics class, where students write essays to explain economic concepts and phenomena. Second, existing approaches have focused on developing techniques and pipelines to align LLM-generated feedback with human feedback using quantitative metrics such as accuracy, recall, and linguistic overlap [2,17]. In contrast, this study investigates the potential for human-AI collaboration, exploring whether AI-generated feedback – despite its imperfections – can enhance instructors' grading experiences. Third, generating feedback for a whole essay presents unique challenges beyond short-answer evaluation, particularly in the context of human-AI interaction. Prior research has shown that instructors prefer to critically review AI-generated content before using it, and seek to understand the rationale behind the AI-generated content [20]. This work also examines the UI challenges of visualizing AI-generated feedback within essays, aiming to enhance usability and instructor trust.

We conducted our study in a college-level Introduction to Economics (ECON101) course. It has frequent short-essay assignments, such as "identify an economic phenomenon involving market failure and analyze the market failure in it". We consider these as knowledge-intensive essays, as they require students to demonstrate a precise understanding of economic concepts through their writing. These assignments also have well-defined rubrics, e.g., "correctly explain the definition of market failure: the market fails to allocate an efficient quantity without government intervention." Extensive prior research has shown that rubrics improve feedback quality [25,33], a finding that also holds when using LLMs to generate feedback [14,27–29,32]. Given this, we chose this course –ECON101– with well-established rubrics to examine an ideal scenario: When detailed rubrics are provided, how does LLM provide feedback on a knowledge-intensive essay?

We propose a feedback engine that decomposes the feedback-generation task into three subtasks. For each rubric, the feedback engine follows these steps: 1)

AI identifies sentences in the essay relevant to the rubric; 2) AI makes a judgment on whether the essay satisfies this rubric; 3) AI generates a feedback message to guide the student toward achieving the rubric without explicitly providing the correct answer. We conducted the study in Fall 2024 in ECON101, where teaching assistants (TAs) provided feedback on students' essays. To create an authentic environment for TAs to critically compare their feedback with the AI's feedback, we first asked TAs to complete their grading tasks as usual. After grading, we invited them to participate in think-aloud sessions where they reviewed AI-generated feedback on the same set of graded essays. During these sessions, TAs compared their feedback with the AI's and reflected on whether they would incorporate AI-generated suggestions into their workflow. The study spanned 4 writing assignments, with 5 TAs participating in a total of 20 one-hour sessions.

Through this study, we aimed to answer the following research questions: 1) How do TAs perceive the differences between AI-generated feedback and their own feedback? 2) What are the prospects of using AI-generated feedback as suggestions in the TA grading process? Specifically, can AI feedback be helpful and how should it be presented to enhance grading effectiveness?

Here is a summary of our findings: 1) AI feedback exhibits more characteristics of effective feedback than human feedback, including the use of praise, explanations, and guiding questions. 2) AI-generated feedback aligns more closely with the rubrics, which can be a double-edged sword– it ensures consistency but may be misleading when the AI applies the rubric too rigidly. 3) AI feedback can be overly fragmented due to rubric-based structuring. This may overlook the need for holistic evaluations that could better support students' learning. However, for knowledge-intensive essays, AI cannot generate accurate feedback without detailed rubrics. 4) Despite its errors, TAs found AI mistakes easy to correct, especially when intermediate AI outputs–such as in-text highlights–were visualized. TAs responded positively to a human-AI collaborative approach, where AI-generated feedback serves as suggestions. They viewed this approach as a way to expedite grading, enhance consistency, and improve overall feedback quality.

2 Related Work

2.1 Existing Work on Evaluating AI Feedback

Recent work has shown the potential of using AI to generate feedback comparable to that of human experts' [2,7,17]. Researchers showed that AI-generated feedback could have higher coherence than human-crafted ones [5], and maintains objectivity [1] across large volumes of evaluations. However, these objective metrics lack insights from both the experts and the students. Jia et al. explored students' and teachers' perspectives on AI feedback. They found limitations in AI feedback including hallucination and vague content, and suggested the irreplaceable nature of human feedback [14]. In this project, we aim to explore the potential for human-AI collaboration in providing feedback, and investigate whether AI-generated feedback can assist with the process despite its limitations.

2.2 Capabilities of LLMs in Modular Tasks

Recent work has shown the capacity of LLMs in a wide range of modular natural language tasks. They demonstrated strong ability in retrieving relevant content and producing highly fluent text. Moreover, with advanced prompting techniques like Chain-of-Thought (CoT) [30], LLMs show notable reasoning abilities and produce rational decisions [6]. However, LLMs are limited by their lack of domain-specific knowledge. And even worse, when they lack relevant knowledge, instead of expressing uncertainty or declining to respond, they often generate inaccurate content, leading to hallucinations, especially for complex tasks [18]. One approach to mitigate hallucinations is to improve AI explainability [13,21]. Researchers developed algorithms leveraging explainability to improve truthfulness in generations [19]. Other work showed that exposing the source of the generated text helps the users better identify hallucinations [26]. In this work, instead of using LLM to perform a single-turn, end-to-end feedback generation, we decompose the task into such modular sub-tasks that LLM has shown high performance on and are easy to validate. To better support users in evaluating the generated content, we present the results from all the intermediate steps.

3 Method

3.1 Tasks

We conducted our study in a college-level Introduction to Economics (ECON101) course, which has frequent knowledge-intensive essay assignments requiring a precise understanding of economic concepts. These assignments also have well-defined rubrics. Additionally, the course instructors provided historical feedback in response to each rubric, which is a feedback message. And it is suggested to provide to the students, if the student doesn't meet this rubric.

3.2 Feedback Generation and Visualization

Instead of adopting an end-to-end generation for feedback, we decomposed the task and developed an LLM-powered pipeline, as shown in Fig. 1. Specifically, the pipeline generates a piece of feedback for each rubric separately. To improve the localization of each feedback, we adopted the idea of Chain-of-Thought [30] and structured the process into three steps. First, we requested the AI model to identify the most relevant sentences in the student's essay that indicate whether the response meets or misses the rubric. These sentences serve as evidence for the subsequent judgment. Next, based on this judgment, the model generates feedback while explicitly providing a rationale.

To better understand the preferable form of AI assistance, we explored two sets of AI-assisted feedback paradigms. In one variant, AI is only used to make judgments, which are used to retrieve the historic feedback. In the other variant, the feedback is generated entirely by AI. To produce feedback aligned with high-quality feedback guidelines, we specified the suggestions by Patchan et al.

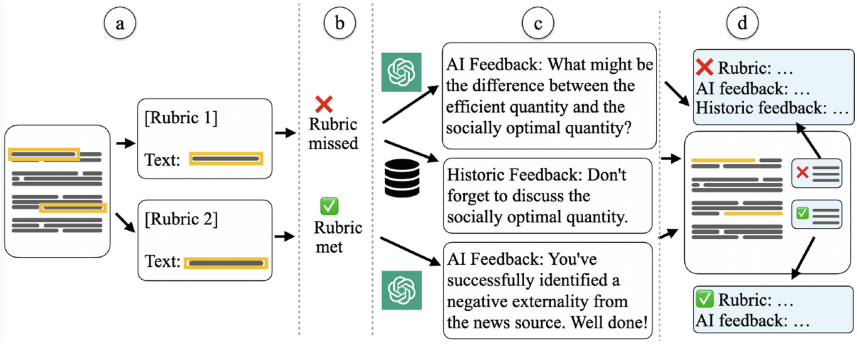

Fig. 1. The feedback generation is decomposed into steps. With the relevant sentences for each rubric identified in (a), AI makes judgments on whether each rubric is met (b). Then the relevant sentences and judgments are used to generate feedback (c). Additionally, if the rubric is missed, historic feedback is retrieved from a set of feedback designed by instructors. The feedback and relevant information will be shown as in-text comments in a Word document (d).

[25] in the system prompt for feedback generation. Specifically, it is specified to (1) use specific and localized language in all feedback; (2) provide praise when the student meets the rubrics; and (3) if the student doesn't meet the rubrics, pinpoint the student's mistake and pose guiding questions without revealing the answer. Additionally, following the suggestion of few-shot learning [3], examples of high-quality feedback created by the instructor were also provided. All the feedback was generated by GPT-4o with a temperature of 0.05 to enhance consistency.[1]

To visualize the feedback, we developed a Word plugin to integrate the generated feedback as in-text comments on a document. To improve the AI explainability of the feedback, the plugin highlighted the relevant sentences identified by AI in the generation. As shown in Fig. 2, each comment includes the following information – the corresponding rubric item, AI judgment, the historic feedback, and the AI feedback. This is to help users better identify hallucinations.

3.3 Study Design

We conducted the study in Fall 2024 in ECON101, where TAs provided feedback to students' essays. The study is IRB-approved. To create an authentic environment for TAs to critically compare their feedback with the AI's feedback, we first asked them to complete their grading tasks as usual. After grading, we invited the TAs to participate in think-aloud sessions where they reviewed AI-generated feedback on the same set of graded essays. During these sessions, TAs compared their feedback with the AI's and reflected on whether they would incorporate

[1] https://github.com/UM-Lifelong-Learning-Lab/AIED2025-Exploring-LLM-Generated-Feedback-for-Economics-Essay.

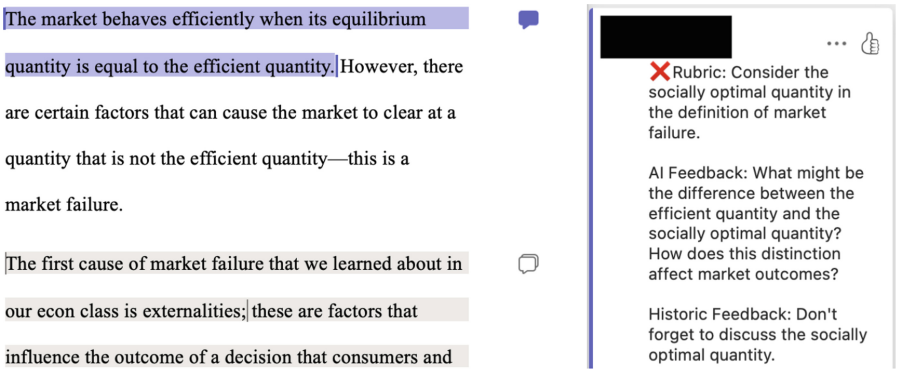

Fig. 2. AI-generated feedback is shown as in-text comments on a Word document, added to the most relevant sentences in the student essay. Each comment contains the rubric, the AI judgment, the AI feedback, and the historic feedback.

AI-generated suggestions into their workflow. The study spanned four writing assignments, with five TAs participating in a total of 20 one-hour sessions. Participants were compensated with a $25 Gift Card for each study session.

Procedure. After getting participants' consent, we gave an introduction on how the AI feedback and historic feedback were generated. We explained that the feedback is generated based on the highlight and the AI judgment. Then participants were presented with a randomly selected set of essays they had graded, with AI feedback as in-text comments in a Word document as shown in Fig. 2. Participants were asked to evaluate the AI feedback, and contrast the AI feedback with their own feedback on the same essay. We also specifically asked about how they envision using the AI feedback if provided as they graded.

Data Analysis Methods. The recordings were transcribed, de-identified, and analyzed with affinity diagram [23]. Two authors interpreted the transcripts, iteratively grouped their notes, and identified key emerging themes.

4 Findings

4.1 RQ1: How Do TAs Perceive the Differences Between AI-Generated Feedback and Their Own Feedback?

AI Feedback Exhibits More Characteristics of Effective Feedback, Such as Providing Praise, Using Guiding Questions and Explanations. First, AI provides more positive feedback (praises), which is often neglected by the TAs. Many participants mentioned that they would directly incorporate the positive feedback during grading. P5 said, *"I would definitely add that (positive) comment just as it is."* In contrast, TAs often do not prioritize leaving positive comments given time constraints in grading. As P2 explained, *"Honestly,*

I was grading so many of these. If I had time I would have liked to put in nice comments, so I would have definitely used AI's positive comments."

Second, participants value AI feedback for using guiding questions as scaffolds. Many TAs agreed on the importance of providing scaffolds in their feedback when students had an opportunity to revise their work. For instance, P5 found the guiding questions helpful for *"push[ing] the students in a certain direction where they're already kind of like on that page."* At the same time, they found it challenging to craft guiding questions on their own. As P2 said, *"the hardest part is not giving them the answer, but also leading them in the right direction."*

Third, AI feedback often contained explanations of the students' mistakes. For example, when the student mistook fish, which is a rival good, to be non-rival, the TA's feedback was, "Fish is rival instead of non-rival". The historic feedback was also generic: "Discuss the concepts of rivalry in the context of the news article", while the AI feedback provides a detailed rationale that "... tuna, as a finite resource, is rival because one person's consumption reduces availability for others." P5 preferred the AI feedback, stating *"I feel like overall, this comment is probably better than the comment I left ... It gives a good explanation as to why (the student was wrong)."*

AI Feedback Is More Personalized to Students' Responses in Comparison to the Historic Feedback. For example, when the student mentioned that the nearby communities are the "third party", and needed to further analyze the involuntary nature of the nearby communities, AI feedback says, "Consider explaining how the communities near the factories are involuntarily affected by pollution." On the contrary, the historic feedback provided by the instructors is designed to be more generic and may not apply to all students. As P2 said, *"We had (historical) comments that we could use... Even those I'd like to tweak a little bit for the specific essays."* Below is an example contrasting the difference between the AI and the historic feedback.

> **Rubric:** "Explain that the third party in the negative externality is involuntarily affected."
> **Student response:** "In economics, we call this a negative externality; the social costs are not taken on by the producers or consumers but by society."
> **AI Feedback:** "How might the impact on individuals differ if they were voluntary participants in the market? Consider how the concept of choice plays into the definition of negative externalities."
> **Historic Feedback:** "Revisit the definition of an externality and consider how those affected are economically reflected in the market."

When presented with both the AI feedback and the historic feedback above, P4 preferred the AI feedback because it is more personalized to the student's problem saying *"I like that (the AI feedback) uses the word 'voluntary' here, because that's kind of the direction you're trying to point them in... (The historic feedback) is just talking about the 3rd party which the student already discusses."*

AI-Generated Feedback Aligns More Closely with the Rubrics, Which Can Be a Double-Edged Sword–It Ensures Consistency but May Be Misleading When the AI Applies the Rubric Too Rigidly. 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, while the TAs might give a combined feedback for several rubric items. However, this brings about the trade-off that AI feedback could be misleading when the rubrics are not well written. We will describe scenarios where AI tends to make mistakes.

First, participants found that AI frequently made mistakes on assessing students' definitions of specialized economic terms, when the definitions were not provided in the rubric. For example, on the rubric item "Correctly use the terms quantity demanded vs. demand", AI often misjudged responses because it doesn't have the expert knowledge to differentiate "quantity demanded" and "demand".

Second, AI can be very strict about a rubric item and may reject alternative ways of writing that also satisfy the rubric criterion. For example, 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 some students implicitly conveyed the intended idea in their writing without explicitly using the language in the rubric, yet AI marked them incorrect.

Third, TAs and AI had different requirements for the depth of explanation in the students' writing. For example, for the rubric item "Explain the change from nonbinding to binding price floor", the student correctly stated that the policy posed a binding minimum wage, which was marked as correct by AI. However, P1 expected a deeper explanation, *"The binding minimum wage. That's good. But I also wanted them to describe the difference between binding and non-binding."*

Lastly, participants found that AI could be too stringent by focusing on unnecessary details in the rubric. When a rubric consists of multiple components, TAs can better identify and prioritize key ideas, whereas AI may lose focus during evaluation and flag minor points that are peripheral to the core concepts.

TAs' Feedback Is More Holistic, Extending Beyond the Rubrics to Consider Broader Aspects of Student Understanding and Writing Quality. While AI-generated feedback typically targets individual rubric items, TAs adopt a more holistic approach. They do not treat each rubric item in isolation; rather, they sometimes synthesize multiple rubrics and focus on overarching aspects such as conceptual understanding and the flow of ideas. For example, P2 said they would combine AI-generated feedback for two different rubric items to create a comprehensive feedback message. Here, we summarize the key considerations TAs took into account when handwriting their feedback.

First, TAs shared that there were 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 insuffi-

cient. As P4 mentions, *"I feel like a more specific comment (than the AI feedback or historic feedback) was needed, just because the student was so wrong."* For instance, one rubric item instructs students to "Mention that automation and labor are substitutes for consumption", prompting the student to use economic terms to analyze how changes in automation affect the labor market. The corresponding AI feedback suggested "To strengthen your argument, mention that automation and labor are substitutes in consumption. This will help explain why firms might switch to automation when labor costs increase." However, the student's essay showed no recognition of the relationship between automation and labor, indicating a more fundamental misunderstanding rather than omission of terminology. Because the AI's suggestion did not address this deeper conceptual gap, the TA found it inadequate. As P4 explained: *"So I don't think this comment really helps the student understand what's going on here. I would ... try to guide them in a way to talk about how the demand for automation is shifting... I like that the AI mentioned that they should explain that they're substitutes. That's important. But this (understanding the relationship) part is also important."*

Second, some rubrics do not apply to all the essays, and some students' mistakes are not covered by the rubrics. For example, the students are required to explain why the "third party" in the negative externalities 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, AI feedback that relies strictly on rubrics often fails to identify such edge cases.

Third, the TAs would also evaluate and provide additional feedback on the flow, conciseness and clarity of the essay, which are not required by the rubrics. For example, P2 advised a student to connect their solutions to the described scenario. P5 flagged a confusing sentence in the essay. P2, P3 and P4 all left additional comments on where the student could be more succinct.

Fourth, TAs would emphasize what they want the student to learn when they provide feedback, and make sure that the feedback can be addressed by the content taught in the class. As P3 remarked, AI feedback didn't catch the student analysis being out of scope. P3 said *"They were using concepts that weren't necessarily relevant to the scope of the assignment. But it's hard to kind of know that unless you know what's being taught in class already."* P5 also considered content covered in prior assignments when offering feedback, saying *"The students did not have to (elaborate on) [a rubric], because at this point in the semester, I think they're able to identify that."*

4.2 RQ2: What Are the Prospects of Using AI-Generated Feedback as Suggestions in the TA Grading Process? How Should It Be Presented to Enhance Grading Effectiveness?

Highlighting the Relevant Texts for Each Piece of AI Feedback Could Speed up Reading and Assessments. Participants shared that the AI highlights in the Word document helped them more rapidly identify key ideas in students' responses and thus sped up both reading and evaluation. As P4 said,

"Figure[ing] out what the AI picked up... I think that would be faster than me having to read carefully through the essay twice.". Similarly, P3 noted that highlighting helps them understand the students' essays, and ensures important concepts are not overlooked. They said, *"So I think with AI it's able to parse it out a little further to reduce the amount of time to help me understand. If I missed that they're getting the concept because they said it later on, it [AI highlight] would help me point that out and see that."* Participants found the highlights especially useful when specific terms or evidence were required by the rubric, as they could confirm the presence of these terms at a glance. For example, P2 mentioned, *"I sometimes feel like, I'm looking for specific points that they're hitting. And the AI highlights those specific points... So I definitely think it would help save time."*

AI Feedback Could Save TAs' Time Constructing Feedback. Participants found it challenging to construct meaningful feedback for each student. For example, P4 said, *"The main time is writing those comments cause you gotta write them individually for everyone."* P2 further emphasized the difficulty in posing guiding questions in the feedback, *"The hardest part is not giving them the answer, but also leading them in the right direction, because you can't be too explicit."* Participants shared that having AI feedback could facilitate their feedback writing. P4 said, *"I think it's hard for me to write those comments because they can get a little bit wordy, but for the AI to do it, it's quick and it's easy and I think that's really good to use."*. When participants found AI feedback that they could directly adopt, they applauded that it would save a lot of time.

AI Feedback Could Improve Consistency in Grading Within One TA. Several participants worried about the fairness of their own grading. As P1 explained, *"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?"*. P5 also mentioned the same concern, wondering whether they might *"grade certain students harsher than others (unintentionally)."* We did observe an inconsistency in TAs' grading when the student responses implied the correct idea but were imprecise. For example, for the rubric item, "Mention that automation and labor are substitutes for consumption", P4 was okay with one student using the word "alternatives" in one session, but decided to deduct points in a different session. In such uncertain scenarios, participants often wanted a second opinion. P2 noted, *"Sometimes we'll have questions like, is this acceptable? ... So it's nice to have something to consult."* Several participants considered AI feedback to help them verify their evaluation, e.g., P2 said, *"It's just like having another set of eyes on the paper."*

Moreover, several participants shared that the AI feedback made them think a little harder about the rubrics. For instance, both P4 and P5 realized they had overlooked certain rubric criteria in their evaluations, which the AI had identified. P5 commented after reading the AI feedback for several students: *"I'm reading through these (AI) comments. And I'm like, Wow, I feel like I'm*

not catching a lot of things.". Consequently, they found AI feedback *"helpful to have standardization within your own section"*.

AI Feedback Could Help Standardize Grading Among TAs. Although instructors developed the rubrics and held regular staff meetings throughout the semester to help TAs learn and apply them consistently, TAs still observed inconsistency in their application and expressed concerns about this issue. P5 said *"I think something that we are always concerned about is that one [TA] is grading too leniently versus other [TA] that's grading harshly."* Both P1 and P2 noticed that they were more lenient on certain rubrics than other TAs. One main source of inconsistencies lies in the varying interpretations of the rubrics. Some rubrics defined aspects that students should address, and there remains flexibility in the detailedness and depth of analysis. For example, one rubric item requires "a thoughtful and well-reasoned solution", which leaves significant discretion to TAs. P5 expected a thorough explanation of why a solution would be effective, whereas P2 found that a correctly named potential solution was sufficient. Moreover, even when the idea is specified in the rubric, the judgment on students' alternative ways to express the idea also leads to inconsistency. Participants envision that AI feedback could help build consistency in grading among TAs.

Concern About Over-Relying on AI Assistance. Participants are aware of potential AI mistakes. They also shared concerns about missing key points and making mistakes if they fully rely on AI. For example, P3 found reading AI feedback in the middle of reading the student's essay distracting, and P4 shared concerns about overlooking sentences that were not highlighted. Almost all participants mentioned that to ensure all the important points were covered and all the errors were caught, they would read the essay and make their own judgment first, before they read and evaluate the AI judgment and feedback. As P5 said, *"You don't want to like over-rely on it (AI), in case something is inconsistent."*

5 Discussion, Limitation and Design Implications

Our study highlights both the strengths and limitations of AI-generated feedback. While AI feedback offers benefits such as stronger alignment with rubrics and detailed, personalized explanations, it may contain errors and ineffective messages. Using AI feedback as a supplementary tool to enhance TAs' grading processes shows significant promise. In this section, we discuss three key considerations when designing human-AI collaborative systems to support the process: 1) Establish clearly written rubrics; 2) Use highlights to increase transparency on how feedback is generated; 3) Provide intermediate outputs on each AI subtask, so that users can decide which outcomes to consume.

Establish Clearly Written Rubrics. Our study highlights the importance of rubrics for AI to generate high quality feedback. We found that further explanation and clarification are needed beyond the original rubrics TAs use for the LLM to make accurate evaluations, in line with [31]. We provide tips in Table 1 on elaborating rubrics to make them more understandable by LLMs.

Table 1. Suggestions for elaborating rubrics in order for LLMs to generate accurate feedback that is aligned with expectations 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 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	The student did not use long direct quotes (more than 1 sentence in one quote) from the article	Direct in-text references are present

Use Highlights to Increase Transparency on How Feedback is Generated. Many participants shared that even when AI makes mistakes in determining whether a rubric item is satisfied, the highlighted sentences in the essay made it easy for them to understand and correct AI's judgment. For example, 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."* After grading 2 essays, P4 found a pattern in what rubric AI might make mistakes. They would be more careful and mainly relied on their own judgments. On the other hand, they appreciate AI for providing detailed and personalized feedback and giving more praise, as well as guiding questions, which are time-consuming for the TAs to write themselves. Our findings indicate that highlighting the sentences used for the AI outputs is an effective way to visualized the AI's rationales without requiring users to read additional explanatory text.

Provide Intermediate Outputs on Each AI Subtask, So That Users Can Decide Which Outcomes to Consume. Our findings showed that the three-step feedback engine was effective in generating high quality feedback for knowledge-intensive essays. For each rubric, the three steps include 1) identify relevant sentences in the essay for this rubric; 2) make a judgment on whether this rubric is satisfied or not; 3) generate a feedback message. This decomposition of the feedback generation task gives users the flexibility to decide whether they should take the AI results. For example, AI might make mistakes on the second step, i.e., evaluating whether the rubric is satisfied. With the highlighted sentence, the user can locate the mistake more easily and flip the judgment. AI might also generate suboptimal feedback, e.g., revealing the correct answer. With

the highlighted sentences and the suggested historic and AI feedback, users can decide to either use the historic feedback, write their own feedback, or combine both feedback to craft a better message. Our findings 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.

In this work, we concentrated on knowledge-intensive essays. Such essay assignments are widely adopted in STEM courses to support student learning in various domains, including Materials Science and Engineering, Organic Chemistry, Introductory Statistics, etc. [8] Our work looks into the TA's perspective and highlights the promising prospects of AI-generated feedback. TAs found AI feedback more detailed and aligned with the rubrics, while they usually adopt a more wholistic approach. Future work could explore whether such point-to-point feedback is preferable. Moreover, as our findings show positive prospects of AI feedback, future work could further investigate how feedback created with AI feedback as suggestions would impact students' learning outcomes.

6 Conclusion

This study aims to examine the prospect of using AI-generated feedback as suggestions to expedite and enhance human instructors' feedback provision. We performed the study in a college-level introductory economics class, which often assigns knowledge-intensive essays, where students need to demonstrate a precise understanding of economic concepts in their writing. Despite providing detailed and clearly-written rubrics, and using a carefully designed feedback engine, there are inevitable hallucinations, mistakes, and ineffectiveness in AI feedback. This study argues for a future where AI feedback can be provided as suggestions during human instructors' grading and feedback provision process. Participants responded positively to the use of AI suggestions and perceived them to accelerate grading and enhance feedback quality. Moreover, we highlight the importance of providing detailed and comprehensive rubrics when using AI to provide feedback on knowledge-intensive essays.

Acknowledgments. This work was funded by NSF Grants IIS-2302564. The findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Almasre, M.: Development and evaluation of a custom GPT for the assessment of students' designs in a typography course. Educ. Sci. **14**(2), 148 (2024)
2. Almegren, A., Mahdi, H.S., Hazaea, A.N., Ali, J.K., Almegren, R.M.: Evaluating the quality of AI feedback: a comparative study of AI and human essay grading. Innovations Educ. Teach. Int., 1–16 (2024)

3. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
4. Chi, M.T., Wylie, R.: The ICAP framework: linking cognitive engagement to active learning outcomes. Educ. Psychol. **49**(4), 219–243 (2014)
5. Dai, W., et al.: Can large language models provide feedback to students? A case study on ChatGPT. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 323–325. IEEE (2023)
6. Eigner, E., Händler, T.: Determinants of LLM-assisted decision-making. arxiv 2024. arXiv preprint arXiv:2402.17385
7. Escalante, J., Pack, A., Barrett, A.: AI-generated feedback on writing: insights into efficacy and ENL student preference. Int. J. Educ. Technol. High. Educ. (20), 55 (2023)
8. Finkenstaedt-Quinn, S.A., Watts, F.M., Shultz, G.V., Gere, A.R.: A portrait of MWrite as a research program: a review of research on writing-to-learn in stem through the MWrite program. Int. J. Scholarsh. Teach. Learn. **17**(1), 18 (2023)
9. Hattie, J., Clarke, S.: Visible learning: Feedback. Routledge (2018)
10. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007)
11. Heickal, H., Lan, A.: Generating feedback-ladders for logical errors in programming using large language models. arXiv preprint arXiv:2405.00302 (2024)
12. Hicks, C.M., Pandey, V., Fraser, C.A., Klemmer, S.: 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, pp. 458–469 (2016)
13. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)
14. Jia, Q., et al.: LLM-generated feedback in real classes and beyond: perspectives from students and instructors. In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 862–867 (2024)
15. Kakarla, S., Borchers, C., Thomas, D., Bhushan, S., Koedinger, K.R.: Comparing few-shot prompting of GPT-4 LLMs with BERT classifiers for open-response assessment in tutor equity training. arXiv preprint arXiv:2501.06658 (2025)
16. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance robust student learning. Cogn. Sci. **36**(5), 757–798 (2012)
17. Latif, E., Zhai, X.: Fine-tuning ChatGPT for automatic scoring. Comput. Educ. Artif. Intell. **6**, 100210 (2024)
18. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474 (2020)
19. Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M.: Inference-time intervention: eliciting truthful answers from a language model. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
20. Lu, X., Fan, S., Houghton, J., Wang, L., Wang, X.: ReadingQuizMaker: a human-NLP collaborative system that supports instructors to design high-quality reading quiz questions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–18 (2023)
21. Luo, H., Specia, L.: From understanding to utilization: a survey on explainability for large language models. arXiv preprint arXiv:2401.12874 (2024)

22. McNichols, H., Lee, J., Fancsali, S., Ritter, S., Lan, A.: Can large language models replicate its feedback on open-ended math questions? arXiv preprint arXiv:2405.06414 (2024)
23. Moggridge, B., Atkinson, B.: Designing Interactions, vol. 17. MIT Press, Cambridge (2007)
24. Nelson, M.M., Schunn, C.D.: The nature of feedback: how different types of peer feedback affect writing performance. Instr. Sci. **37**, 375–401 (2009)
25. Patchan, M.M., Schunn, C.D., Correnti, R.J.: The nature of feedback: how peer feedback features affect students' implementation rate and quality of revisions. J. Educ. Psychol. **108**(8), 1098 (2016)
26. Rashkin, H., et al.: Measuring attribution in natural language generation models. Comput. Linguist. **49**(4), 777–840 (2023)
27. Scarlatos, A., Smith, D., Woodhead, S., Lan, A.: Improving the validity of automatically generated feedback via reinforcement learning. In: International Conference on Artificial Intelligence in Education, pp. 280–294. Springer (2024)
28. Steiss, J., et al.: Comparing the quality of human and ChatGPT feedback of students' writing. Learn. Instr. **91**, 101894 (2024)
29. Wang, R., Zhang, Q., Robinson, C., Loeb, S., Demszky, D.: 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), pp. 2174–2199 (2024)
30. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837 (2022)
31. Wu, X., Saraf, P.P., Lee, G.G., Latif, E., Liu, N., Zhai, X.: Unveiling scoring processes: dissecting the differences between LLMs and human graders in automatic scoring. arXiv preprint arXiv:2407.18328 (2024)
32. Xu, C., Lin, J., Wu, T., Aleven, V., Koedinger, K.R.: Improving automated feedback systems for tutor training in low-resource scenarios through data augmentation. arXiv preprint arXiv:2501.09824 (2025)
33. Yuan, A., Luther, K., Krause, M., Vennix, S.I., Dow, S.P., Hartmann, B.: 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, pp. 1005–1017 (2016)