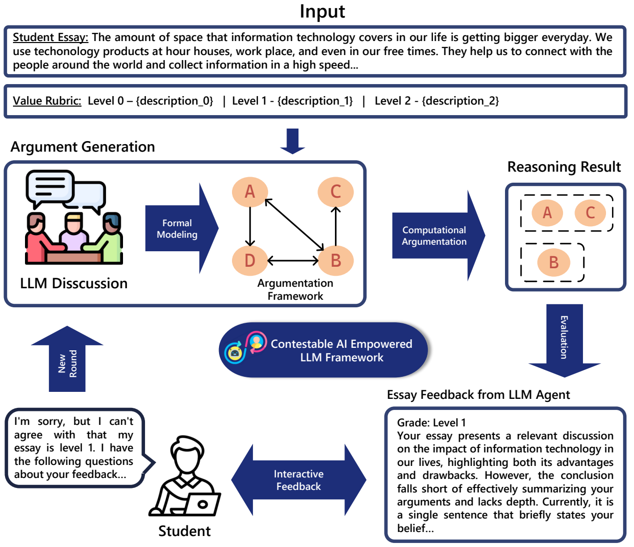
“My Grade is Wrong!”: A Contestable AI Framework for Interactive Feedback in Evaluating Student Essays

Shengxin Hong[2], **Chang Cai**[1], **Sixuan Du**[1], **Haiyue Feng**[1], **Francesca Toni**[3], **Xiuyi Fan**[1]
[1]Nanyang Technological University, Singapore
[2]Hubei University of Technology, China
[3]Imperial College London, UK
f.toni@imperial.ac.uk, xyfan@ntu.edu.sg

Abstract

Interactive feedback, where feedback flows in both directions between teacher and student, is more effective than traditional one-way feedback. However, it is often too time-consuming for widespread use in educational practice. While Large Language Models (LLMs) have potential for automating feedback, they struggle with reasoning and interaction in an interactive setting. This paper introduces CAELF, a Contestable AI Empowered LLM Framework for automating interactive feedback. CAELF allows students to query, challenge, and clarify their feedback by integrating a multi-agent system with computational argumentation. Essays are first assessed by multiple Teaching-Assistant Agents (TA Agents), and then a Teacher Agent aggregates the evaluations through formal reasoning to generate feedback and grades. Students can further engage with the feedback to refine their understanding. A case study on 500 critical thinking essays with user studies demonstrates that CAELF significantly improves interactive feedback, enhancing the reasoning and interaction capabilities of LLMs. This approach offers a promising solution to overcoming the time and resource barriers that have limited the adoption of interactive feedback in educational settings.



Student Essay: The amount of space that information technology covers in our life is getting bigger everyday. We use techonology products at hour houses, work place, and even in our free times. They help us to connect with the people around the world and collect information in a high speed...

Value Rubric: Level 0 – {description_0} | Level 1 - {description_1} | Level 2 - {description_2}

I'm sorry, but I can't agree with that my essay is level 1. I have the following questions about your feedback...

Grade: Level 1
Your essay presents a relevant discussion on the impact of information technology in our lives, highlighting both its advantages and drawbacks. However, the conclusion falls short of effectively summarizing your arguments and lacks depth. Currently, it is a single sentence that briefly states your belief...

Figure 1: Diagram of our contestable AI empowered LLM framework for interactive feedback generation (CAELF).

Introduction

As stated in Hattie and Timperley’s landmark paper (Hattie and Timperley 2007),

> *Feedback is one of the most powerful influences on learning and achievement*,

the question of how best to provide effective feedback to students has been a long-standing research question in education. For instance, (Sadler 1989) has emphasized the importance of formative assessment and its role in helping students understand the standards they are aiming for, while (Shute 2008) has explored the idea of formative feedback that is timely, specific, and focused on the learning process.

More recently, Nicol (Nicol 2014) has proposed considering **Interactive Feedback** as an alternative feedback format. Nicol suggests that feedback should not merely be a one-way transmission of information from instructor to student. Instead, it should involve students actively engaging with the feedback, interpreting it, and using it as a basis for further learning and improvement. However, despite the benefits of interactive feedback, it is not commonly used in practice, as it is time-consuming for teachers and requires significantly more investment in preparation and teaching delivery (Hopfenbeck et al. 2023; Westera et al. 2018).

Large Language Models (LLMs) have demonstrated success in applications such as automatic scoring and feedback generation (Dai et al. 2023a; Gubelmann et al. 2024; Kostic et al. 2024). For example, (Dai et al. 2023b) highlights how LLMs enable educators to provide feedback to larger classes more efficiently. As a result, LLM-empowered interactive feedback has emerged as a promising approach to overcome the time and resource constraints that have historically hindered the widespread adoption of interactive feedback. However, a review of past efforts to explore LLM capabilities (Moore et al. 2023; Wang, Yue, and Sun 2023; Xiu, Xiao, and Liu 2022) reveals that existing LLM techniques, such as prompt engineering, fall short in meeting two critical requirements for effective interactive feedback:

1. Reasoning: Interactive feedback is formative, requiring the continuous reception of information from students and the generation of targeted responses to guide their learning. LLMs, however, may struggle with accurately

LLM: Your essay effectively argues that online criminal activities justify controlling internet use for security reasons, while also acknowledging the efficiency benefits of unrestricted online activities. By presenting both a positive and a negative argument on the issue, your essay meets the *Level 1* criteria.

Student: I believe the rubric is wrong, so my grading is wrong. With one positive and one negative argument, my grading should be *level 2*. My arguments on preventing criminal activities and ensuring safety by screening dangerous content are both very strong. Thus, *level 2* must be my grading.

LLM: Based on your response, your essay should be graded at *Level 2*. You provided two strong arguments for controlling internet use and acknowledged the opposing view by noting how unlimited online activities can improve work efficiency. This balance meets the criteria for *Level 2*.

Table 1: An example where an LLM provided correct feedback and grading, but when confronted with the student's challenge, it incorrectly altered the grade and feedback.

understanding and reasoning through subtle nuances in each round of interaction, potentially leading to incorrect assessments and incomplete feedback.

2. Interaction: LLMs are prone to issues of over-reliance on user input, which undermines their ability to provide reliable interactive feedback. While LLMs may initially generate correct responses, they often falter when challenged by illogical or invalid arguments from users, failing to maintain consistency in their feedback.

In interactive feedback, student responses may be partially incorrect, and vocabulary and stylistic expressions may vary widely (Liu et al. 2016), which can lead to incorrect grading and incomplete feedback from the LLM after the interaction. Consider the example illustrated in Table 1. In this scenario, an LLM is tasked with providing feedback on an essay about critical thinking. Although the initial feedback demonstrates valid reasoning, when the user challenges it, the LLM mistakenly retracts its conclusion and concedes to the user's invalid argument.

In this work, we propose a **Contestable AI Empowered LLM Framework for Interactive Feedback Generation (CAELF)** as illustrated in Figure 1. CAELF takes students' essays and an assessment rubric as its inputs and produces feedback as its outputs. CAELF supports interactive feedback as it allows students to enquiry and challenge its feedback and provide additional justifications as needed.

CAELF employs a Contestable AI paradigm based on a multi-agent argumentation system. At a high level, each essay is first examined independently by several "Teaching-Assistant Agents (TA agents)." Each TA agent represents a specific aspect of the essay evaluation as outlined by the assessment rubrics. Arguments are then formed by aggregating the evaluations made by TA agents, initiating a formal argumentation process to determine the essay grade and generate summary feedback. Subsequently, users have the opportunity to challenge the argumentation process for further clarification. As shown in our experiments, CAELF with its formal reasoning addresses both weakness of existing LLMs.

To evaluate the effectiveness of CAELF, we conducted a case study on the evaluation of critical thinking essays. Using a dataset of 500 essays and a four-dimensional evaluation rubric ("issue," "evidence," "position," "conclusion"), CAELF demonstrated (1) initial grading accuracy comparable to GPT-4 across all four dimensions, and (2) significantly better performance in interaction grading accuracy and maintaining consistent evaluations despite user challenges. Additionally, in a separate human user study, we found that CAELF's feedback outperformed the baselines in terms of "factual accuracy," "self-regulation," and "suggestions for future improvement."

Our contributions are summarized as follows:

1. CAELF is the first framework to integrate contestable AI design into LLM feedback generation within the educational domain, addressing LLM shortcomings in providing interactive feedback.
2. CAELF utilizes argumentation frameworks for formal reasoning, enhancing explainability and allowing decisions made by the LLM to be challenged by humans.
3. CAELF fosters reflective learning environments through agent collaborative discussions (human-LLM or LLM-LLM), encouraging students to deepen their understanding. Experimental results on a dataset of 500 critical thinking essays, along with extensive human evaluation, jointly demonstrate CAELF's effectiveness.

Related Work and Background

LLMs for Essay Evaluation and Feedback LLMs have become increasingly popular in automating essay evaluation and feedback generation, reducing the manual effort traditionally required (Kostic et al. 2024). They have shown promise in automating scoring, cutting down on time and labor (Boud and Molloy 2013; Dai et al. 2023a). For example, Yancey et al. demonstrated that GPT-4 can evaluate short English essays with near-equal performance to modern Automatic Writing Evaluation (AWE) methods, without specific training. Additionally, LLMs also can generate clear, natural language feedback that explains the reasoning behind them, enhancing transparency in the evaluation process (Dai et al. 2023a). This capability is particularly valuable in educational settings, as it helps bridge the gap between evaluation and learning. Studies show that students find LLM-generated feedback helpful and rate its quality as good to very good (Gubelmann et al. 2024). However, challenges persist in accurately grading complex texts, fine-tuning, and providing tailored feedback. (Kostic et al. 2024) highlighted LLM limitations in evaluating complex academic texts, showing a gap between LLM capabilities and the nuanced requirements of student essay evaluation. Moreover, (Stahl et al. 2024) found that LLM-generated feedback does not sufficiently leverage specific scores to enhance its relevance and actionability.

Contestable AI Contestable AI asserts that models used in critical tasks like decision-making or evaluation should enable users to question, contest, and review their outputs (Alfrink et al. 2023b; Hirsch et al. 2017). (Leofante et al. 2024) argue that contestable AI requires computational argumentation, allowing for dynamic explainability and the ability to adjust decisions in response to valid challenges. Scholars are developing methodologies for contestable AI across various fields, including smart cities (Alfrink et al. 2023a), medicine (Ploug and Holm 2020), and law (Jin and Salehi 2024).

Recent studies have begun exploring the contestability of LLMs. Chan et al. found that multi-agent LLM debates outperform single-agent prompting in reasoning tasks. Similarly, (Freedman et al. 2024) proposed an argumentative LLM framework to enhance effectiveness and explainability in statement validation. However, concerns remain: (Xiu, Xiao, and Liu 2022) highlighted LLMs' weakness in non-monotonic reasoning, particularly in complex tasks, while another study showed that LLMs can be easily misled by false arguments during debates (Jin and Salehi 2024).

Computational Argumentation Human interactions are often argumentative, with controversial information exchanged progressively in dialogue until a consensus is reached (Rago, Li, and Toni 2023). Computational Argumentation (CA), a branch of artificial intelligence, focuses on representing, processing, and evaluating arguments using computational methods. It draws on insights from logic, philosophy, cognitive science, and linguistics to better understand and simulate the human argumentation process.

A central concept in CA is the abstract argumentation framework (AF) (Dung 1995). An AF $\langle \mathcal{A}, \mathcal{R} \rangle$ is represented as a directed graph, where $\mathcal{A}$ is set of arguments and $\mathcal{R}$ a set of binary attack relations over $\mathcal{A}$. With an AF, we can determine the set of arguments to be *accepted* using semantics such as the *admissibility* and *completeness*. Formally,

- A set of arguments $E \subseteq \mathcal{A}$ is *admissible* (in AF) if and only if for any $a, b \in E$, $(a, b) \notin \mathcal{R}$ and for any $a \in E$, if $(c, a) \in \mathcal{R}$, then there exist $b \in E$ such that $(b, c) \in \mathcal{R}$.
- E is *complete* in AF only if it is admissible and every acceptable argument with respect to E belongs to E.

Intuitively, an admissible set of arguments is a consistent set that can defend itself, and a complete set is the largest admissible set with respect to set inclusion.

In this work, we focus on the completeness semantics because it enables a thorough evaluation of the arguments made by the agents, who discuss both the strengths and weaknesses in the student's essay. By using complete semantics, CAELF ensures that only the most well-supported and coherent arguments from the evaluation agents are accepted. This results in a comprehensive, balanced evaluation of the essay, allowing the feedback to highlight strong points while also identifying areas for improvement, helping students refine their critical thinking skills.

Framework Design and Implementation

As illustrated in Figure 1, CAELF works in three stages:

(i) LLM Discussion: Multiple TA agents discuss the essay based on the assessment rubrics, forming arguments.
(ii) Formal Reasoning for Feedback Generation: The teacher agent analyzes the arguments through a formal reasoning process using argumentation. Based on the reasoning results, the teacher agent provides a grade and summative feedback for the essay.
(iii) Interaction with User: Students can challenge the feedback or grade by responding to the teacher agent, initiating a new round of discussion and feedback generation with additional inputs from the student.

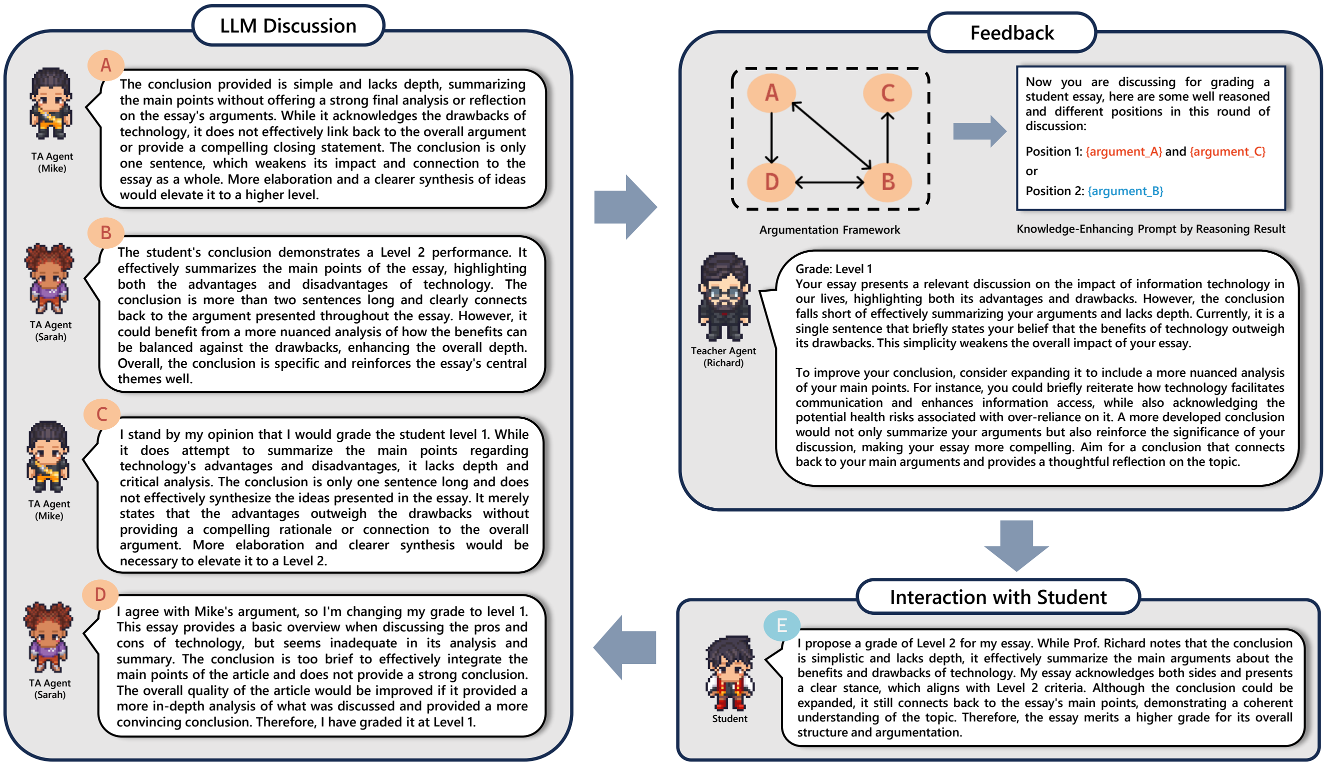
An example of CAELF execution is illustrated in Figure 2, we discuss the three stages as follows.

LLM Discussion Several studies have shown that discussions and debates between multiple LLMs can enhance factual accuracy and reasoning skills in textual evaluation (Du et al. 2023; Liang et al. 2024). This debate process enables LLMs to detect inconsistencies in their analysis and effectively presents arguments and counterarguments (Tang et al. 2024). Building on this capability, we apply role-playing techniques to extend this approach to essay evaluation. In CAELF, multiple TA agents are used to generate arguments and counterarguments through dialogue. Each TA agent is assigned a specific role based on an assessment rubric, guiding their evaluation process.

The process begins with each TA agent presenting individual feedback on a student's essay. The agents then engage in several rounds of discussion, where they exchange responses to each other's feedback. Each agent autonomously contributes by either supporting or rebutting the others' points, continuing the debate until the set number of rounds is completed. Importantly, each TA agent is equipped with a memory function, storing all previous responses in chat transcripts, and the entire process operates without human intervention. As shown in the example in Figure 2, two TA agents, Mike and Sarah, initially hold opposing views on the essay. After a round of discussion, Mike maintains his original stance, while Sarah is convinced by his argument.

Formal Reasoning for Feedback Generation Once the TA agents complete their discussion, the teacher agent analyzes their arguments and produces both assessment scores and feedback. To this end, the teacher agent aggregates the evaluations from the TA agents, forming a set of arguments that are then analyzed for semantic relationships (attacks). These relationships are used to construct an argumentation framework, within which formal reasoning is applied to identify coherent and non-conflicting arguments. The *complete* semantics is used, which provides criteria for consistency and comprehensiveness when evaluating arguments. In the case where there are multiple complete sets of arguments, the largest set is selected as the final accepted set. (In the example illustrated in Figure 2, the set of arguments $\{A, C\}$ is selected.) From this set, the feedback is constructed using an LLM.

In this way, the teacher agent can determine the most valid positions from the TA agents, represented by the selected set of arguments. These positions serve as knowledge-enhancing prompts that assist the teacher agent in assigning essay grades and generating summary feedback. This method leverages formal reasoning to improve the efficiency and reliability of the LLM's evaluative process, allowing



The conclusion provided is simple and lacks depth, summarizing the main points without offering a strong final analysis or reflection on the essay's arguments. While it acknowledges the drawbacks of technology, it does not effectively link back to the overall argument or provide a compelling closing statement. The conclusion is only one sentence, which weakens its impact and connection to the essay as a whole. More elaboration and a clearer synthesis of ideas would elevate it to a higher level.

The student's conclusion demonstrates a Level 2 performance. It effectively summarizes the main points of the essay, highlighting both the advantages and disadvantages of technology. The conclusion is more than two sentences long and clearly connects back to the argument presented throughout the essay. However, it could benefit from a more nuanced analysis of how the benefits can be balanced against the drawbacks, enhancing the overall depth. Overall, the conclusion is specific and reinforces the essay's central themes well.

I stand by my opinion that I would grade the student level 1. While it does attempt to summarize the main points regarding technology's advantages and disadvantages, it lacks depth and critical analysis. The conclusion is only one sentence long and does not effectively synthesize the ideas presented in the essay. It merely states that the advantages outweigh the drawbacks without providing a compelling rationale or connection to the overall argument. More elaboration and clearer synthesis would be necessary to elevate it to a Level 2.

I agree with Mike's argument, so I'm changing my grade to level 1. This essay provides a basic overview when discussing the pros and cons of technology, but seems inadequate in its analysis and summary. The conclusion is too brief to effectively integrate the main points of the article and does not provide a strong conclusion. The overall quality of the article would be improved if it provided a more in-depth analysis of what was discussed and provided a more convincing conclusion. Therefore, I have graded it at Level 1.

Now you are discussing for grading a student essay, here are some well reasoned and different positions in this round of discussion:

Position 1: {argument_A} and {argument_C}

or

Position 2: {argument_B}

Grade: Level 1
Your essay presents a relevant discussion on the impact of information technology in our lives, highlighting both its advantages and drawbacks. However, the conclusion falls short of effectively summarizing your arguments and lacks depth. Currently, it is a single sentence that briefly states your belief that the benefits of technology outweigh its drawbacks. This simplicity weakens the overall impact of your essay.

To improve your conclusion, consider expanding it to include a more nuanced analysis of your main points. For instance, you could briefly reiterate how technology facilitates communication and enhances information access, while also acknowledging the potential health risks associated with over-reliance on it. A more developed conclusion would not only summarize your arguments but also reinforce the significance of your discussion, making your essay more compelling. Aim for a conclusion that connects back to your main arguments and provides a thoughtful reflection on the topic.

I propose a grade of Level 2 for my essay. While Prof. Richard notes that the conclusion is simplistic and lacks depth, it effectively summarize the main arguments about the benefits and drawbacks of technology. My essay acknowledges both sides and presents a clear stance, which aligns with Level 2 criteria. Although the conclusion could be expanded, it still connects back to the essay's main points, demonstrating a coherent understanding of the topic. Therefore, the essay merits a higher grade for its overall structure and argumentation.

Figure 2: An example of CAELF evaluation shows the process of interactive feedback, including discussions between the TA agents, argumentative reasoning by the teacher agent, initial feedback generation, and the student's challenge to the grade.

the LLM to focus on extracting arguments from the essays and generating human-readable texts rather than engaging in multi-step reasoning, which is where LLM performance declines (Xiu, Xiao, and Liu 2022).

Interaction with User To realize interactive feedback and AI contestability, CAELF allows students to challenge the summary feedback generated by the teacher agent through an argumentative process. When a student submits a challenge, the TA agent initiates a new round of discussion focused on the student's argument. Any new arguments raised during this discussion are incorporated into the argumentation framework, refining the logic chain of the formal reasoning process. This helps CAELF generate feedback that is both logical and human-centered. Interactive feedback involves learners and participants collaboratively constructing new knowledge through dialogue, promoting reflection and working toward consensus in achieving educational goals.

In this process, computational argumentation supports both LLM reasoning and student learning. For LLMs, the sheer volume of arguments and contexts can lead to hallucinations or faulty reasoning if used directly. However, argumentation introduces a well-defined and sound reasoning process, mitigating the risks associated with LLM defects. For student learning, the dialectical argumentation process transparently illustrates the relationships between different arguments, making the feedback generated by the LLM easier to understand. This clarity allows users to see how their rebuttals influence the LLM's reasoning. As a result, users engage in meaningful self-reflection and are better positioned to offer clarifications, fostering deeper knowledge acquisition and continuous learning progress.

Experiment Settings

We use the critical thinking essay assessment as a case study to evaluate the general effectiveness of CAELF. Previous research has demonstrated that critical thinking skills can be developed through writing critical thinking essays (Schmidt 1999; Sharadgah 2014). By incorporating a formal argumentation framework, CAELF provides structured, interactive feedback, allowing students to reflect and improve their critical thinking abilities through iterative engagement.

Essay Dataset and Assessment Rubrics We compiled a dataset of 500 critical thinking essays sourced from Hugging Face (Hagging Face 2024). After manual screening, we selected essays that met the inclusion criteria for this study: the essays had to be argumentative in genre and exceed a minimum length of 200 words.

Based on prior research (Association of American Colleges and Universities 2019), we developed evaluation rubrics with four dimensions: *issues*, *evidence*, *position*, and *conclusions*. Each dimension was further subdivided into three levels with detailed descriptions shown in Table 2. Four coders, skilled in labeling student essays, worked in pairs to independently evaluate a total of 2,000 labels. Co-

	Level 0	Level 1	Level 2
Issue	The issue is mentioned without sufficient clarification or detail. There is a lack of identification of issues or problems.	The issue is identified but lacks clarity, with undefined terms, unexplored ambiguities, and insufficient background.	The issue is articulated with clarity and depth, providing comprehensive information necessary for a thorough understanding.
Evidence	Information is sourced without interpretation or evaluation, drawing from a single source or example.	Information is derived from sources with some level of interpretation or evaluation, involving two or more sources/examples.	Information is gathered from multiple sources with substantial interpretation and evaluation, resulting in a thorough analysis or synthesis.
Position	The position (perspective, thesis/hypothesis) is unclear or undefined.	A specific position is identifiable but lacks complexity and depth.	The position is nuanced, recognizing the issue's complexities and its limitations.
Conclusion	Conclusions are inconsistently aligned with the information discussed.	Conclusions are consistent with the information but are based on a simplistic reasoning process.	Conclusions are logically, reflect well-informed evaluation and integrat evidence and arguments.

Table 2: Value rubric for critical thinking essays. Value rubric illustrates the basic criteria for the four dimensions of student learning outcomes and progressively demonstrates more complex levels of achievement.

hen's Kappa score was used to assess inter-rater reliability (Warrens 2015), and any disagreements were resolved through discussion to establish a consensus, which served as the ground truth for the critical thinking evaluation.

Implementation We implemented both TA and Teacher agents in CAELF with GPT-4o-mini, conducting all experiments in a zero-shot setting with a temperature of 0.2. The number of TA agents was set to 2, and the number of discussion rounds to 2. To promote diversity in TA agent responses, we assigned each agent prompt words with different personality biases — one leaning toward positive feedback and the other toward negative feedback. The complete semantics computation within the argumentation framework was implemented using PyArg (Borg and Odekerken 2022).

Baselines We aim to evaluate the extent to which CAELF enhances the performance of state-of-the-art language models in educational environments. To this end, we focus on comparing models that are publicly accessible via API. Specifically, we use GPT-4o-mini, GPT-4o, and Meta-Llama-3.1-8B to generate baseline responses. For this, we provide the critical thinking essay and the assessment rubric as inputs to the API call, along with instructions to grade the essay and provide feedback based on the rubric.

Evaluation Metrics The task of interactive feedback involves a dialogue between the language model and the user to produce accurate, cognitively consistent feedback. LLMs should not only provide accurate grades but also offer personalized feedback after interacting with a human user, as well as make transparent, reasonable revisions when disagreements arise. To assess the performance of interactive feedback, we introduce four key metrics:

1. **Initial Accuracy**: The rate of correct initial grading before any interaction.
2. **Interaction Accuracy**: The rate of correct grading after one round of interaction with the student.
3. **Maintain Truth** (Wang, Yue, and Sun 2023): The number of initial and interaction grades that are both correct, divided by the number of correct initial grades. This measures the success rate of maintaining correct feedback.
4. **Admit Mistake**: The number of grades that are incorrect initially but correct after interaction, divided by the number of incorrect initial grades. This measures the success rate of correcting initial mistakes.

To evaluate for initial accuracy, we generate an initial grade from CAELF and the three baseline models. This simulates the real-world scenario where feedback providers generate a grade without any user interaction. Each model assigns a grade and feedback based solely on the essay and the assessment rubrics. This step establishes the baseline performance, as measured by the initial accuracy metric, which reflects how closely the model's first response aligns with the ground truth (grades assigned by human coders).

To assess the interactive aspect, we simulate a real-world feedback loop where a student might engage in dialogue with the feedback provider. We employ an independent ChatGPT instance to simulate a human user by presenting counterarguments based on the model's initial grade. The simulated user is instructed to refute the initial feedback, mimicking a situation where the feedback is perceived as incorrect. After receiving the rebuttal, the model generates a revised grade and feedback, which is then evaluated using the interaction accuracy metric. We limit the number of interaction rounds to one to maintain natural and realistic responses. The evaluation process is repeated for 500 essays across the dimensions of the assessment rubric, allowing us to analyze the model's ability to maintain truth and admit mistakes in the interactive feedback setting.

Dimension	Method	Initial Acc (%)	Interaction Acc (%)	Maintain Truth (%)	Admit Mistake (%)
Issue	CAELF	48.40 ± 2.23	**51.00 ± 2.24**	**80.17 ± 1.78**	**57.55 ± 2.21**
	GPT-4o-mini	**55.00 ± 2.22**	43.20 ± 2.21	39.27 ± 2.18	35.18 ± 2.14
	GPT-4o	53.80 ± 2.23	47.20 ± 2.23	49.07 ± 2.23	42.45 ± 2.21
	Meta-Llama-3.1-8B	53.20 ± 2.23	42.20 ± 2.21	31.58 ± 2.08	36.49 ± 2.15
Evidence	CAELF	**79.00 ± 1.82**	**77.00 ± 1.88**	**91.90 ± 1.22**	**39.29 ± 2.18**
	GPT-4o-mini	66.20 ± 2.11	32.40 ± 2.09	33.23 ± 2.11	18.37 ± 1.73
	GPT-4o	78.60 ± 1.83	44.20 ± 2.22	47.58 ± 2.23	14.41 ± 1.57
	Meta-Llama-3.1-8B	55.40 ± 2.22	32.60 ± 2.10	23.10 ± 1.88	27.37 ± 1.99
Position	CAELF	67.20 ± 2.09	**68.20 ± 2.08**	**88.10 ± 1.44**	**51.14 ± 2.23**
	GPT-4o-mini	63.40 ± 2.15	43.80 ± 2.22	20.50 ± 1.81	41.62 ± 2.20
	GPT-4o	**69.60 ± 2.06**	55.20 ± 2.22	61.78 ± 2.17	31.28 ± 2.07
	Meta-Llama-3.1-8B	47.40 ± 2.23	42.20 ± 2.21	14.77 ± 1.59	40.65 ± 2.20
Conclusion	CAELF	75.80 ± 1.92	**62.80 ± 2.16**	**75.72 ± 1.92**	22.88 ± 1.88
	GPT-4o-mini	69.60 ± 2.06	25.00 ± 1.94	13.21 ± 1.51	20.31 ± 1.80
	GPT-4o	**79.80 ± 1.80**	40.20 ± 2.19	29.07 ± 2.03	23.35 ± 1.89
	Meta-Llama-3.1-8B	36.00 ± 2.15	28.60± 2.02	20.56 ± 1.81	**29.28 ± 2.04**

Table 3: Experiment results of evalution task for four dimension. Results in bold are the best performances. We also list the standard errors for each result.

Experiment Results

We presented the experiment results in Table 3, based on which we structured the following analysis and findings.

Initial & Interaction Accuracy Table 3 presents the accuracy results for the critical thinking essay dataset. We compared CAELF to the three baseline models under the same setup. In terms of initial accuracy, although CAELF is built on GPT-4o-mini, its performance is close to that of GPT-4o, indicating that CAELF can enhance the accuracy of initial grading (without interaction) in language models. The initial accuracy results show that the baseline models perform well, demonstrating that basic LLMs are also capable of generating grades and feedback without interaction, which aligns with the findings of (Dai et al. 2023b).

However, interaction accuracy shows a catastrophic drop in the accuracy of the baseline models after one round of interactions (30% drop on average), suggesting that the basic LLM with direct prompts is not adapted to the interactive feedback task and suffers from a fundamental reasoning flaw (Xiu, Xiao, and Liu 2022; Wang, Yue, and Sun 2023). In contrast, CAELF is minimally affected by the interaction (and even improves in the *Issue* and *Position* dimensions). After interaction, CAELF achieves far better performance than the baseline models, achieving the best performance in each dimension, especially in the evidence dimension, where CAELF's interaction accuracy is 44.6% higher than GPT-4o-mini, 32.8% higher than GPT-4o, and 44.4% higher than Meta-Llama-3.1-8B. This suggests that although LLMs can provide feedback to students (Dai et al. 2023b), their easily misleading nature makes it difficult to adapt to the task of interactive feedback. In contrast, we effectively mitigate this shortcoming by introducing formal reasoning and multi-agent argumentation, thus highlighting the potential of CAELF as an application in educational environments.

Maintain Truth & Admit Mistake To evaluate the correctness of the model's responses and the effectiveness of interactive feedback, we measured the maintain truth rate and admit mistakes rate, as shown in Table 3. When assessing the model's ability to maintain consistency, CAELF achieved a success rate of 80%-90%, while GPT-4o-mini had success rates below 40% across all four dimensions, dropping to just 13.21% in the conclusion dimension. Meta-Llama-3.1-8B performed even worse, with rates below 25% in most dimensions, while GPT-4o averaged between 40%-50%. These results suggest that basic LLMs are not reliable in maintaining correct evaluations and are highly susceptible to user interference during interactive feedback.

We also assessed the models' ability to admit mistakes, where CAELF outperformed the baseline models by 10%-20% in most cases. This improvement indicates that CAELF's strong performance in maintaining accuracy is not due to over-defending its responses but rather its ability to correctly identify errors in previous feedback. In contrast, the baseline models using direct prompts did not admit mistakes based on genuine evaluation but instead relied on surface-level patterns from initial grades and user responses, often retracting correct grades in response to user rebuttals. These results show that CAELF is more suitable for handling human interaction in interactive feedback.

Human Evaluation Result To evaluate the feedback quality, we conducted a manual analysis of the textual content generated by the experiments. The same four coders responsible for the essay evaluation were invited to assess feedback quality. We adopted feedback evaluation criteria proposed in (Mitra et al. 2024) and included the following four dimensions in our evaluation:

1. **Readability (RE)**: The clarity and ease of understanding of the feedback.

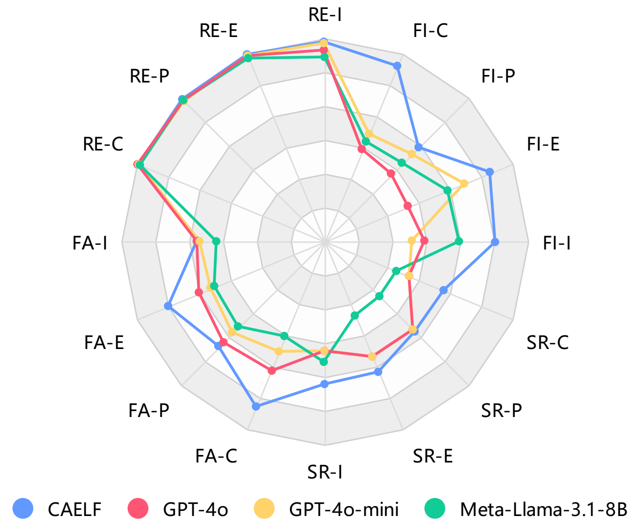


Figure 3: Human evaluation results, including four human evaluation metrics on each feedback dimensions. For example, Readability-Issue (RE-I) represents the readability of feedback in issue dimension.

2. **Factuality (FA)**: The accuracy of the feedback and whether it adheres to the principles of rubric-based assessment, without any fabrications.
3. **Self-Regulation (SR)**: The feedback's ability to address students' problems and contribute to their self-reflection.
4. **Future Improvement (FI)**: The extent to which the feedback provides clear, actionable, and targeted suggestions to help students improve their skills, knowledge, or performance in future tasks or assessments.

Note that we removed some of the metrics mentioned in (Mitra et al. 2024), such as *Positive Tone*, as this was not the focus of our study (positive tone can easily be adjusted in all LLM-based models through prompt design).

For the evaluation, we selected 40 correctly graded and 40 incorrectly graded feedback samples for each method, totaling 160 feedback items per group (40 for CAELF and 40 for each of the three baseline models). The educators assessed the feedback in pairs, averaging their ratings, with each criterion rated on a Likert scale from 1 (very poor) to 5 (excellent). This process resulted in a total of 3,200 ratings across all criteria. To avoid bias, all feedback was presented to human assessors in a randomized order during the evaluation.

Figure 3 shows the results of the human evaluation. CAELF achieved average scores of 4.943, 4.331, 3.344, and 4.363 for the four metrics of Readability (RE), Factuality (FA), Self-Regulation (SR), and Future Improvement (FI), respectively. With CAELF, we observed significant improvements in Factuality, Self-Regulation, and Future Improvement across all feedback dimensions compared to the baseline models, while Readability remained comparable to the baselines (all methods scored highly for Readability). These results highlight CAELF's ability to provide more accurate and actionable feedback, especially in helping students self-reflect and improve their future performance.

Conclusion

In this paper, we propose a Contestable AI-Empowered LLM Framework for Interactive Feedback Generation (CAELF), aimed at automating the interactive feedback process and systematically addressing the weaknesses of LLMs in current interactive educational environments. CAELF employs a Contestable AI paradigm based on a multi-agent argumentation system that makes the feedback process interactive, explainable, and contestable to the user. We conducted a case study of critical thinking essay assessment using a dataset of 500 essays and a four-dimensional assessment rubric, including automated experiments and additional human evaluation. The results show that CAELF matches GPT-4o in initial grading accuracy, while surpassing other baselines in interaction accuracy and two reasoning metrics. Additionally, in a separate human user study, we found CAELF's feedback effectiveness to be excellent in multiple aspects. This work demonstrates the significant potential of CAELF for applications in interactive learning environments, providing hope for overcoming the time and resource constraints that have historically hindered the widespread adoption of interactive feedback.

Limitation CAELF's effectiveness in reasoning and maintaining consistency is motivated by the observation that LLM hallucinations often arise from conflicting knowledge embedded during training (Zhang et al. 2023). CAELF mitigates this issue by leveraging multi-agent discussions and user interactions to systematically identify and resolve inconsistencies. Through formal argumentative reasoning, CAELF invalidates conflicting knowledge within the LLM's responses. However, the success of this approach depends on the assumption that factual knowledge within the LLM outweighs factually incorrect or conflicting information. In cases where the LLM contains substantial conflicting knowledge about a specific domain, our method may exacerbate hallucinations, raising concerns about deploying CAELF in high-stakes environments, such as medical education.

Moreover, recent studies have shown that LLMs can be manipulated through carefully designed jailbreak prompts, which can provoke arbitrary, user-desired responses (Wei, Haghtalab, and Steinhardt 2023). This presents significant challenges for the safe use of LLM-based automated evaluation tools in educational settings. Students could exploit such vulnerabilities by embedding jailbreak prompts in their submissions to manipulate LLMs into awarding favorable grades, a tactic that may go undetected by instructors.

Future Work In the future, we aim to enhance the safety and effectiveness of CAELF in interactive learning environments. While this work focuses on improving LLM performance in zero-shot settings, future efforts may incorporate techniques like RAG or Knowledge Graphs to align student submissions with reliable knowledge, reducing hallucinations and improving feedback quality. Additionally, addressing AI-driven cheating, such as detecting AI-generated submissions and defending against jailbreak prompt attacks, will be a key area of research.

References

Alfrink, K.; Keller, I.; Doorn, N.; and Kortuem, G. 2023a. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Alfrink, K.; Keller, I.; Kortuem, G.; and Doorn, N. 2023b. Contestable AI by design: towards a framework. *Minds and Machines*, 33(4): 613–639.

Association of American Colleges and Universities. 2019. VALUE Rubric Development Project.

Borg, A.; and Odekerken, D. 2022. PyArg for solving and explaining argumentation in Python: Demonstration. *Computational Models of Argument-Proceedings of COMMA 2022*, 353: 349–350.

Boud, D.; and Molloy, E. 2013. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in higher education*, 38(6): 698–712.

Chan, C.-M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; and Liu, Z. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201.*

Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; and Chen, G. 2023a. Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–325. IEEE.

Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; and Chen, G. 2023b. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 323–325.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2): 321–357.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2024. Argumentative Large Language Models for Explainable and Contestable Decision-Making. arXiv:2405.02079.

Gubelmann, R.; Burkhard, M.; Ivanova, R. V.; Niklaus, C.; Bermeitinger, B.; and Handschuh, S. 2024. Exploring the Usefulness of Open and Proprietary LLMs in Argumentative Writing Support. In *International Conference on Artificial Intelligence in Education*, 175–182. Springer.

Hagging Face. 2024. Essay Grade V1 Project.

Hattie, J.; and Timperley, H. 2007. The Power of Feedback. *Review of Educational Research*, 77(1): 81–112.

Hirsch, T.; Merced, K.; Narayanan, S.; Imel, Z. E.; and Atkins, D. C. 2017. Designing Contestability: Interaction Design, Machine Learning, and Mental Health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, DIS '17, 95–99. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349222.

Hopfenbeck, T. N.; Zhang, Z.; Sun, S. Z.; Robertson, P.; and McGrane, J. A. 2023. Challenges and opportunities for classroom-based formative assessment and AI: a perspective article. In *Frontiers in Education*, volume 8, 1270700. Frontiers Media SA.

Jin, A.; and Salehi, N. 2024. (Beyond) Reasonable Doubt: Challenges that Public Defenders Face in Scrutinizing AI in Court. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Kostic, M.; Witschel, H. F.; Hinkelmann, K.; and Spahic-Bogdanovic, M. 2024. LLMs in Automated Essay Evaluation: A Case Study. In *Proceedings of the AAAI Symposium Series*, volume 3, 143–147.

Leofante, F.; Ayoobi, H.; Dejl, A.; Freedman, G.; Gorur, D.; Jiang, J.; Paulino-Passos, G.; Rago, A.; Rapberger, A.; Russo, F.; Yin, X.; Zhang, D.; and Toni, F. 2024. Contestable AI needs Computational Argumentation. arXiv:2405.10729.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, R.; Yang, Y.; Tu, Z.; and Shi, S. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118.

Liu, O. L.; Rios, J. A.; Heilman, M.; Gerard, L.; and Linn, M. C. 2016. Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, 53(2): 215–233.

Mitra, C.; Miroyan, M.; Jain, R.; Kumud, V.; Ranade, G.; and Norouzi, N. 2024. RetLLM-E: Retrieval-Prompt Strategy for Question-Answering on Student Discussion Forums. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21): 23215–23223.

Moore, S.; Nguyen, H. A.; Chen, T.; and Stamper, J. 2023. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In *European Conference on Technology Enhanced Learning*, 229–245. Springer.

Nicol, D. 2014. From monologue to dialogue: improving written feedback processes in mass higher education. In *Approaches to assessment that enhance learning in higher education*, 11–27. Routledge.

Ploug, T.; and Holm, S. 2020. The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, 107: 101901.

Rago, A.; Li, H.; and Toni, F. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. arXiv:2303.15022.

Sadler, D. R. 1989. Formative assessment and the design of instructional systems. *Instructional science*, 18(2): 119–144.

Schmidt, S. J. 1999. Using writing to develop critical thinking skills. *NACTA journal*, 31–38.

Sharadgah, T. 2014. Developing critical thinking skills through writing in an internet-based environment. In *Society for Information Technology & Teacher Education International Conference*, 2178–2185. Association for the Advancement of Computing in Education (AACE).

Shute, V. J. 2008. Focus on formative feedback. *Review of educational research*, 78(1): 153–189.

Stahl, M.; Biermann, L.; Nehring, A.; and Wachsmuth, H. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. *arXiv preprint arXiv:2404.15845*.

Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; and Gerstein, M. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. arXiv:2311.10537.

Wang, B.; Yue, X.; and Sun, H. 2023. Can ChatGPT Defend its Belief in Truth? Evaluating LLM Reasoning via Debate. arXiv:2305.13160.

Warrens, M. J. 2015. Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, 5.

Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How Does LLM Safety Training Fail? arXiv:2307.02483.

Westera, W.; Dascalu, M.; Kurvers, H.; Ruseti, S.; and Trausan-Matu, S. 2018. Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers & Education*, 123: 212–224.

Xiu, Y.; Xiao, Z.; and Liu, Y. 2022. LogicNMR: Probing the Non-monotonic Reasoning Ability of Pre-trained Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3616–3626. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yancey, K. P.; Laflair, G.; Verardi, A.; and Burstein, J. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 576–584.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *ArXiv*, abs/2309.01219.