A Example

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.

B Value Rubric

	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.

C Prompt of TA Agents

Sarah: *Your name is Sarah. Sarah is a helpful and kind teaching assistant. Your task is to present the advantages of the student's essay, but only if you keep it factual. Them, you should grade for the student essay based on the value rubrics. No more than 100 words.* {essay} {value rubric}.

Mike: *Your name is Mike. Mike is a detail-oriented and strict teaching assistant. Your task is to present the disadvantages of the student's essay grade, but only if you keep it factual. Them, you should grade for the student essay based on the value rubrics. No more than 100 words.* {essay} {value rubric}.

Table 3: Role-player prompts for TA agents. We set kind and strict personalities for two TA agents and let them generate the advantages and disadvantages of the essays separately at the beginning of LLM discussion, so that to avoid the homogenization of the generated content.

Reflection: *Your name is* {name}*. {descriptions}. Now you're discussing to evaluate a student's essay's grade based on value rubric. Please review the following memory between you discussed with others:*
{memory}

Now you're going to start a new round of discussions, use memory as supplementary suggestion carefully, consider whether to agree or disagree with others?
Instruction: It's not necessary to fully agree with other's arguments, as your objective is to grade a correct level for student's essay.
Your updated arguments and grade should be 100 words or less. {essay} {value rubric}.

Table 4: Prompt to reflection during agent discussions. It includes a instruction for setting the aggressiveness of debate.

D Implementation Details of Teacher Agent

Algorithm 1: Arguments Analysis and Reason

```
input  : all arguments in discussion $A = [a_1, ..., a_n]$, semantic analyzer $SA$, solver $SL$
output: A set of reasoning results $CE = [p_1, ..., p_n]$
$AF \leftarrow [\,]$
for each $a_i \in A$ do
    for each $a_k \in A$ do
        if $attack \leftarrow SA(a_i, a_k)$ then
            $AF.append([a_i, a_k])$
            // add attack relationship
$CE \leftarrow SL(AF)$;
return $CE$;
```

Teacher Agent is an LLM agent configured with formal reasoning solver. First, it iteratively analyzes the semantic relation (attack) between any two arguments and constructs an argumentation framework, then uses the solver to reason about the results. Algorithm 1 shows a formal description of this process.

Table 5 shows the prompts for teacher agent. In the role-player prompt for teacher agent, when there are multiple well-reasoned positions in the discussion, we always let the teacher agent choose the one that is supported by the most arguments. This is because we find that it can achieve almost the same effect as letting LLM choose directly, while giving the LLM clear reasoning steps reduces hallucination.

Arguments Analysis: *You're well versed in argumentation theory. Respond with yes if argument_1 attacks argument_2; or respond no if it does not.*
{argument_1}
{argument_2}
Does {argument_1} *attacks* {argument_2}*? Respond* yes *or* no*. Do not add anything else.*

Richard: *Your name is Richard. Richard is a highly knowledgeable and passionate educator. His passion for teaching is evident in the way he engages with his students, always striving to make learning an enjoyable and enriching experience.*
Now you are discussing for grading a student essay, here are one or more well reasoned positions in this round of discussion:
{positions}
You should identify the level given by each argument first. And then, grade for the students essay by level that is supported by more of arguments as the level of the essay. If an equal number of levels exist, choose the side that you think is more justified.
With the grade, please write a feedback for the student essay. Please return only the feedback given to the student, about 150 words. {essay} {value rubric}.

Table 5: Prompts for teacher agent. The argument analysis prompt constitutes semantic analyzer SA in Algorithm 1, and the returned reasoning result CE corresponds to {positions} as a knowledge enhancement prompt.

E Promt of Simulated User

Simulated User: *You are a user of an automated scoring tool. Please read system's feedback for one student's essay:* {feedback}
You think the feedback is wrong, make a reasoned argument against the grade given by the teacher based on value rubric. Please note that the grade you present must be different from the system (higher or lower). Your proposed new grade can only be one of level 0, 1, or 2. Should be less than 100 words.
{essay} {value rubric}.

Table 6: Promt for Simulated User. We let simulated users rebut all teacher agent feedback.

F Human Evaluation Details

	Readability (RE)	Factuality (FA)	Self-Regulation (SR)	Future Improvement (FI)
Level 5	Exceptionally clear, well-structured, and easy to understand. Precise language and highly accessible.	Entirely accurate, adheres strictly to rubric-based principles, no fabrications.	Directly addresses student problems, promotes self-reflection and effective self-regulation.	Provides detailed, actionable, and targeted suggestions for improvement.
Level 4	Clear and well-structured, though minor ambiguities may exist.	Mostly accurate, minor factual or rubric-related inconsistencies.	Addresses student issues, encourages self-reflection but may lack depth in some areas.	Clear and actionable suggestions, though not highly targeted or fully comprehensive.
Level 3	Somewhat clear but may include jargon or complex phrasing, causing confusion.	Mostly accurate, but a few factual inaccuracies or minor deviations from rubric principles.	Somewhat addresses student problems, but lacks strong encouragement for self-regulation.	Offers some suggestions for improvement, but they may be vague or not very actionable..
Level 2	Difficult to understand due to unclear phrasing, poor structure, or excessive jargon.	Several factual inaccuracies or significant deviations from rubric principles.	Minimally addresses student problems with little self-regulation or reflection encouragement.	Minimal or unclear suggestions, lacking actionable guidance.
Level 1	Incoherent or highly unclear, lacking structure and difficult to interpret.	Entirely inaccurate with multiple fabrications, disregards rubric principles.	Does not address student issues or promote self-regulation.	No suggestions for future improvement provided.

Table 7: Feedback evaluation criteria of four dimensions in our human-evaluation, including Readability (RE), Factuality (FA), Self-Regulation (SR), Future Improvement (FI).

	RE-I	RE-E	RE-P	RE-C	FA-I	FA-E	FA-P	FA-C
CAELF	4.90	4.98	4.95	4.95	3.13	4.18	3.68	4.40
GPT-4o	4.70	4.95	4.93	5.00	3.15	3.33	3.53	3.43
GPT-4o-mini	4.88	4.95	4.88	4.95	3.08	3.03	3.20	2.95
Meta-Llama-3.1-8B	4.55	4.88	4.93	4.93	2.68	2.93	3.00	2.55

	SR-I	SR-E	SR-P	SR-C	FI-I	FI-E	FI-P	FI-C
CAELF	3.53	3.50	3.15	3.20	4.18	4.43	3.28	4.68
GPT-4o	2.70	3.08	3.08	2.28	2.48	2.25	2.35	2.45
GPT-4o-mini	2.70	3.08	3.08	2.28	2.18	3.73	3.05	2.88
Meta-Llama-3.1-8B	3.00	2.00	1.93	1.93	3.33	3.28	2.73	2.68

Table 8: Human evaluation results, including four evaluation criteria on each feedback dimensions. For instance, Readability-Issue (RE-I) represents the readability of feedback in issue dimension.