Lessons Learned in Assessing Student Reflections with LLMs

Mohamed Elaraby, Diane Litman
University of Pittsburgh
Pittsburgh, PA, USA
{mse30,dlitman}@pitt.edu

Abstract

Advances in Large Language Models (LLMs) have sparked growing interest in their potential as explainable text evaluators. While LLMs have shown promise in assessing machine-generated texts in tasks such as summarization and machine translation, their effectiveness in evaluating human-written content—such as student writing in classroom settings—remains underexplored. In this paper, we investigate LLM-based specificity assessment of student reflections written in response to prompts, using three instruction-tuned models. Our findings indicate that although LLMs may underperform compared to simpler supervised baselines in terms of scoring accuracy, they offer a valuable interpretability advantage. Specifically, LLMs can generate explanations that are faithful, non-repetitive, and exhibit high fidelity with their input, suggesting potential for enhancing the transparency and usability of automated specificity scoring systems.

Prompt: Describe what you found most confusing in today's class.
[Score 1] I thought that most of the topics explained were relatively simple or I had previously learned them. I felt confident in my understanding after the class session.
[Score 2] the class participation activity
[Score 3] Undirected vs directed was a bit confusing in terms of how to read the chart
[Score 4] Finding the right problem to address.
Prompt: Describe what you found most interesting in today's class.
[Score 1] I found nothing interesting in class. Being Friday, I could barely pay attention.
[Score 2] the review session
[Score 3] The part about bias in data labeling was thought provoking
[Score 4] Writing the problem statement.

Table 1: Representative reflections for each specificity score (1–4) across two prompts. This illustrates one challenge of assessing specificity: long reflections may lack substance (Score 1), while short ones may convey detailed, content-specific insights (Score 4).

1 Introduction

Reflective writing is a fundamental skill that enhances learning by encouraging students to critically engage with course material and articulate their thoughts. This process benefits both students and instructors by fostering greater awareness and facilitating meaningful classroom interactions (Baird et al., 1991; Menekse, 2020). The quality of written reflections is often assessed based on their *specificity* (Menekse et al., 2011; Li et al., 2025), which measures the level of detail and depth in a given reflection. Table 1 shows student reflections written after a physics lecture, along with human-assessed specificity scores, both from the `ReflectSumm` corpus described in Section 3.

Automating specificity assessment is crucial for delivering interventions to help students improve the quality of their reflections (Knoth et al., 2020; Wilhelm, 2021), e.g., by providing scaffolding feedback which in turn can ultimately enhance learning outcomes (Menekse et al., 2025). More specific reflections can also provide instructors with more reliable insights into student understanding and needs (Menekse, 2020). Traditionally, specificity scoring has relied on supervised models (Magooda et al., 2022; Carpenter et al., 2020; Li and Nenkova, 2015). However, collecting large annotated datasets in educational contexts is resource-intensive and not always feasible. Depending on model type, the reasoning behind the scoring might also not be explainable to students or instructors.

Advancements in Large Language Models (LLMs) for evaluating natural language, commonly referred to as the *LLM-as-a-judge* paradigm (Zheng et al., 2023), have introduced new possibilities for leveraging LLMs in educational applications. These models can generate human-like judgments

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 672–686

Research Question	Key Finding	Lessons
RQ1: LLM vs. Supervised Baselines	**Retrieval-based few-shot improves scoring**	Selecting semantically similar in-context examples boosts accuracy over random or fixed examples.
	LLMs underperform supervised models	Distill-BERT outperforms all LLMs, suggesting a need for adaptation.
	Chain-of-Thought (CoT) explanations do not improve scoring	Generated explanations fail to enhance LLM-based scoring.
RQ2: Self-Generated Explanations	**Explanations are faithful to the input and explanation vocabulary do not fully overlap with the input reflection vocabulary**	Explanations do not contradict or repeat the input, suggesting potential for interpretable and supportive understanding of the scores.
	High fidelity suggests explanations are highly influencing the predictions	Misleading explanations can negatively affect the scoring.

Table 2: Key findings of RQ1 and RQ2.

without task-specific training, making them attractive for low-resource tasks such as reflection specificity assessment. Their generative capabilities in addition suggest new possibilities for explainable methods. This paper investigates whether LLMs can serve as viable alternatives to traditional supervised models for assessing reflection specificity.

Reflective writing poses challenges compared to other educational scoring tasks. Unlike Automatic Essay Scoring (Foltz et al., 1999; Attali and Burstein, 2004; Shermis and Wilson, 2024), which typically assesses longer texts, reflective writing is often concise and highly variable in length, with reflections ranging from a single word to multiple phrases or complete sentences (Kember et al., 2008). This variability poses a unique challenge in distinguishing different levels of specificity: shorter reflections may lack sufficient context, while longer ones can introduce ambiguity in assessment. Table 1 presents two examples of reflections that contain multiple sentences yet receive the lowest specificity score (1) due to vague or off-topic content, as well as two shorter reflections that achieve the highest specificity score (4) by providing concise, content-rich responses relevant to the prompt. Also, while tasks such as Short Answer Grading (Burrows et al., 2015), which is closer in length and variability to reflections, primarily involve assessing objective responses within a given question context with reference answers, reflective writing is inherently subjective as it conveys personal experiences and insights, further complicating standardized assessment.

In this work, we extend prior research on leveraging LLMs as judges for educational text evaluation by focusing on reflection specificity. We investigate this through two research questions: **RQ1: Can LLM-based specificity assessment improve scoring reliability compared to supervised baselines?** We explore two approaches to LLM-based specificity assessment: (1) *Standard Prompting:* LLMs are instructed to predict specificity scores based on the input reflection. (2) *Chain-of-Thought (CoT) Prompting:* LLMs are prompted to generate a rationale before making a specificity judgment. This technique, widely used in complex NLP tasks, encourages models to engage in step-by-step reasoning, potentially leading to more consistent and interpretable assessments. We investigate these settings under both *zero-shot* and *few-shot* conditions to assess their impact on model performance. **RQ2: Do self-generated explanations enhance interpretability?** We investigate whether generated explanations contribute to the transparency of LLM-based specificity scoring, potentially making the evaluation process more interpretable and informative for students and educators. *Our key findings are summarized in Table 2.*[1]

Our contributions are twofold:

1. We evaluate the effectiveness of three open-weight LLMs in scoring student specificity under various zero-shot and few-shot settings.

2. We analyze the linguistic properties of the generated explanations and their role in interpreting the output, aiming to assess whether these

[1] https://github.com/EngSalem/Explainable-Reflection-Quality

Score	Specificity Meaning	Definition
1	Vague	Reflection implies "no confusing issue," e.g., "nothing" or "none for this class."
2	Non-specific	Reflection does not include any statement(s) about course content but refers to organizational aspects (e.g., homework, exams).
3	General	Reflection includes statement(s) about course content but lacks specific details.
4	Specific	Reflection includes specific and detailed statement(s) about course content.

Table 3: Rubric for evaluating reflection specificity based on decision tree from Luo and Litman (2016).

explanations are meaningful and potentially useful for providing students with feedback on their reflective writing.

2 Related Work

LLM-as-Judge LLMs have demonstrated correlation with human evaluation of machine-generated texts in tasks such as counter-narrative generation (Zubiaga et al., 2024), text summarization (Fu et al., 2024; Liu et al., 2023), multi-turn question answering (Zheng et al., 2023), and automatic persuasion ranking (Elaraby et al., 2024). However, for more nuanced human-written content, such as academic reviews (Zhou et al., 2024) and essay scoring (Mansour et al., 2024; Stahl et al., 2024), LLMs (particularly without fine-tuning or alignment) still fall short compared to human evaluators and domain-specific supervised models trained on high-quality annotated data. *In this work, we investigate LLMs as specificity evaluators for student reflections, a distinct category of human-written text.*

LLM-as-Judge in Educational Text Traditional approaches to assessing student writing often rely on surface linguistic features to enhance automatic scoring models ranging from feature-based to hybrids with deep learning, including list ranking (Uto et al., 2020) and neural-based methods (Jin et al., 2018; Uto et al., 2020). Recent work has explored leveraging LLMs as evaluators for educational text. Stahl et al. (2024) employed persona-based zero-shot prompting for essay scoring, and Hou et al. (2025) integrated linguistic features into zero-shot evaluations; however, both studies found limited improvements over traditional supervised baselines. In contrast, Baral et al. (2024) showed that a fine-tuned `Mistral-7B` model outperformed other supervised models in math essay scoring. Closely related to our work, Li et al. (2025) investigated reflection specificity assessment, demonstrating that multi-LLM voting strategies outperform single LLM scoring approaches. *Building on these developments, our work examines LLMs' capabilities in assessing student reflections, focusing on how in-context examples influence predictions. Additionally, we analyze the interpretability of LLM-generated explanations, offering a novel perspective particularly valuable for building downstream applications in high-stakes domains like education.*

Evaluating Self-Generated Explanations Assessing self-generated explanations has largely centered on their impact on model performance. Existing metrics, such as accuracy differences with and without explanations (Hase et al., 2020a; Wiegreffe et al., 2021a) and information-theoretic measures (Chen et al., 2023), quantify how explanation content influences predictions. Wiegreffe and Marasovic (2021) proposed evaluation criteria based on surface validity, grammatical correctness, and alignment with the target label, including *contrast* with alternative labels. Expanding on this, Joshi et al. (2023) introduced *novelty*, capturing the introduction of new information, which proved useful in human-AI collaboration tasks. These measures have since been extended to domains like persuasiveness evaluation (Elaraby et al., 2024). Despite these advancements, self-generated explanations remain largely unexplored in educational contexts beyond their role in enhancing automatic scoring (Stahl et al., 2024). *In this work, we examine their effectiveness not only in improving specificity scoring, but also for their potential to generate explanations that are faithful and non-repetitive with input, and exhibit fidelity with scoring.*

3 Datasets

For LLM evaluation, we utilize `ReflectSumm`[2] (Zhong et al., 2024), a corpus of 17,509 reflections aggregated by unique reflection per lecture from 24 STEM courses across 2 universities, written in response to the prompts in Table 1. This dataset was selected for its inclusion of high-quality annotations of individual reflection specificity scores,

[2] https://huggingface.co/datasets/mse30/ReflectSumm

Score	Count	Ref. Length (Min / Mean / Max)
1	1,841	1 / 11.19 / 135
2	2,488	1 / 6.17 / 62
3	9,231	2 / 15.26 / 87
4	3,949	4 / 30.19 / 194

Table 4: Distribution of reflection specificity scores in the ReflectSumm dataset, with reflection (ref.) length in number of words statistics.

rated on a scale from 1 (vague) to 4 (specific) using the rubric in Table 3. The annotations exhibit substantial inter-annotator agreement, with a reported pairwise Quadratic Weighted Kappa of 0.668 across 4 distinct annotators (trained college students with backgrounds in the appropriate subject domains) (Zhong et al., 2024). Table 4 summarizes the score distribution. The table also emphasizes the variability in reflection lengths across all scores.

For both training supervised pre-LLM baselines and as a reflection bank for LLM in-context prompting, we use the publicly available annotated reflections from the CourseMIRROR dataset[3] which is composed of 6680 reflections distributed as $1210, 2035, 2377, 1058$ for scores $1-4$, respectively. Note that although annotated using the same specificity rubric, the CourseMIRROR reflections are from STEM course offerings that are disjoint from those in ReflectSumm.

4 Experimental Settings

4.1 Included LLMs

We included 3 whitebox models which demonstrated strong performance across NLP tasks, as evaluated in the Chatbot Arena leaderboard (Zheng et al., 2023)[4]: `Llama3.1-8B-instruct` (Grattafiori et al., 2024), `Mistral-8B-instruct` (Jiang et al., 2024), and `Qwen-7B` (Yang et al., 2024). For efficient inference, we employed `VLLMs` (Kwon et al., 2023). All experiments were conducted with a decoding temperature set to 0, enabling greedy decoding to mitigate variability that might stem from temperature sampling.

4.2 Prompting the LLMs (Zero-Shot)

Building on the reflection quality assessment of Luo and Litman (2016), we prompt LLMs to assign specificity scores on a scale from 1 to 4, consistent with the guidelines provided to human annotators. The scoring rubric is adapted from the decision-tree criteria described in Luo and Litman (2016), and its definitions are presented in Table 3. This alignment enables a direct comparison between model predictions and dataset reference annotations. Appendix A provides the prompts used for scoring.

4.3 Scoring Evaluation Metric

Given that our prediction is based on point-wise scoring, we rely on **Quadratic Weighted Kappa (QWK)** to report the model prediction agreement with ground truth human annotations.

5 RQ1: LLM-based Assessment vs. Supervised Baselines

These experiments evaluate the effectiveness of LLM-based specificity assessment across a range of settings to enable a comprehensive comparison.

5.1 Supervised Baselines

We include two supervised baselines. The first is **Finetuned-DistilBERT**, where we fine-tune DistilBERT (Sanh, 2019) for specificity assessment following Magooda (2022). The model is initialized from the Hugging Face checkpoint[5] and trained on the annotated reflections from the CourseMIRROR dataset. We fine-tuned the model for 20 epochs using 5-fold cross-validation, optimizing hyperparameters such as the learning rate and number of training epochs. The checkpoint with the highest overall QWK score was selected for evaluating specificity across the full ReflectSumm corpus.

The second baseline is **Nearest Neighbors (NN) Retrieval**. Given a target reflection, we retrieve semantically similar reflections from an annotated reflection bank R_{bank} and estimate its specificity score based on the most frequently occurring specificity label among its nearest neighbors. We use the CourseMIRROR dataset as the reflection bank. This method is used as a comparable baseline to LLMs with nearest neighbor in-context examples (Section 5.2). For each reflection in the ReflectSumm evaluation set, we generate a dense embedding using the `all-MiniLM-L6-v2` model from the `sentence-transformers` library (Reimers and Gurevych, 2019). This maps reflections into a shared vector space, enabling semantic similarity comparisons. We compute the cosine similarity between each reflection in ReflectSumm and the annotated reflections in CourseMIRROR

[3] https://engineering.purdue.edu/coursemirror/
[4] https://lmarena.ai/
[5] distilbert/distilbert-base-uncased

(R_{bank}), and retrieve the top-n most similar reflections. For efficiency, we use the `Faiss` library (Johnson et al., 2019) for fast approximate nearest-neighbor search. The specificity score of a reflection is determined using a mode-based voting mechanism from its nearest neighbors' specificity labels.

5.2 Prompting with In-Context Examples

We explore three in-context learning strategies, ranging from fixed demonstrations (Brown et al., 2020) to selection-based strategies that draw from pre-existing demonstrations (Min et al., 2022).

(1) **Fixed In-Context Examples**: A fixed set of manually curated examples is used as in-context demonstrations across all runs. These examples are drawn from annotated student reflections in Luo and Litman (2016) and remain unchanged during prompting. Since the examples provided in the original paper focused primarily on *confusing* prompts, we supplemented them with additional reflections written in response to *interesting* prompts from the R_{bank} set. Each specificity score is represented by an equal number of examples to ensure balanced coverage.[6] This prompting method serves as a baseline for few-shot in-context learning.

(2) **Random In-Context Examples**: For each instance, n examples are randomly sampled from the annotated reflection bank R_{bank}. This approach assesses the variability in model performance based on arbitrary example selection.

(3) **Nearest-Neighbor In-Context Examples**: Similar to the nearest-neighbor retrieval baseline, the top-n semantically similar reflections from R_{bank} are retrieved for each input reflection. These nearest neighbors serve as in-context demonstrations.

Table 5 shows that none of the included LLMs were able to match the performance of the `DistillBERT` baseline (0.658 QWK) in either zero-shot or any of the few-shot settings. This highlights the limitations of LLMs in specificity assessment when compared to dedicated supervised models. Among the LLMs, `Mistral-8B-instruct` consistently achieved the highest QWK agreement across both zero-shot and few-shot settings. The best performance (0.624 QWK) was obtained when paired with nearest-neighbor retrieval, indicating that retrieving semantically similar reflections enhances the model's ability to assess specificity by providing more contextually relevant examples. However, increasing the number of in-context examples negatively impacted performance across all models and few-shot settings. This suggests that excessive context may introduce conflicting information or divert the model's attention away from the specificity criteria. Also, both fixed and randomly sampled in-context examples performed worse than zero-shot prompting, implying that arbitrarily chosen examples introduce noise rather than meaningful guidance. These findings underscore the importance of carefully curating in-context examples when leveraging LLMs for specificity scoring. *This limitation further reinforces the challenge of deploying LLMs for automated assessment in educational settings without access to high-quality annotated datasets.*

[6] Fixed in-context examples are provided in Appendix B.

5.3 Chain-of-Thought (CoT) Prompting

Instead of directly instructing the model to assign a specificity score to a given reflection, we employ *Chain-of-Thought (CoT) prompting* (Wei et al., 2022) to encourage the model to generate a rationale before providing its final assessment. This approach aims to enhance the reliability and interpretability of the model's scoring process by explicitly incorporating reasoning. To implement CoT prompting, we modify the original scoring prompt by introducing a zero-shot CoT instruction (Kojima et al., 2022) that prompts the model to generate a brief explanation before assigning a score. Specifically, we refine the commonly used CoT instruction, `Let's think step by step`, proposed by Kojima et al. (2022), by prompting `Mistral-8B-instruct` to generate an alternative phrasing that better aligns with the specificity evaluation task. The final instruction used in our experiments is: `Think critically, consider all aspects, and then decide.`

Table 6 demonstrates that prompting `Mistral-8B-instruct` (the best performing LLM from Section 5.2) to generate self-explanations before assigning specificity scores does not improve QWK performance. Across most settings, CoT prompting either slightly lowers or maintains performance compared to standard prompting, with exceptions for 3-shot with random examples and 10-shot with random and nearest neighbor examples. However, this gain does not surpass the best-performing settings. *Our findings thus suggest that CoT self-generated explanations offer limited utility in improving scoring performance.*

Supervised Baselines (QWK)									Best QWK
Distill-BERT			0.658						**0.658**
Nearest Neighbor	–	–	0.410 (3-shot)	0.473 (5-shot)	0.506 (10-shot)	–	–	–	0.506
LLM-Based Models (QWK)									**Best QWK**
Model	Zero-Shot	Few-Shot (Fixed 4-shot per score)	Few-Shot (Random)			Few-Shot (Nearest Neighbor)			
			3-shot	5-shot	10-shot	3-shot	5-shot	10-shot	
Llama3.1-8B-instruct	0.552	0.515	0.546	0.549	0.504	0.601	0.595	0.578	0.601
Mistral-8B-instruct	0.595	0.522	0.532	0.553	0.575	0.624	0.605	0.575	**0.624**
Qwen-7B	0.559	0.485	0.519	0.540	0.456	0.600	0.597	0.569	0.600

Table 5: Quadratic Weighted Kappa (QWK) results for specificity assessment across various few-shot settings on the full ReflectSumm benchmark. The rightmost column highlights the best QWK result within each model group. Shaded cells indicate the best score per model row, and **bolded** values represent group-level best performance.

Retrieval Method	No-CoT (QWK)	CoT (QWK)
Zero-shot	0.595	0.556
Few-shot (Fixed)	0.522	0.522
Few-shot (Random)		
3-shot	*0.532*	*0.576*
5-shot	0.553	0.532
10-shot	*0.575*	*0.588*
Few-shot (Nearest-Neighbor)		
3-shot	0.624	0.607
5-shot	0.605	0.602
10-shot	*0.575*	*0.587*

Table 6: QWK scores for Mistral-8B-instruct comparing No-CoT (repeated from Table 5) vs. CoT prompting. *Italicized* rows indicate settings improved by including CoT prompting. Underlined numbers represent best performing non-CoT and CoT settings.

5.4 RQ1 Summary

As summarized in Table 2, our evaluation of 3 instruction-tuned LLMs in zero-shot, few-shot, and CoT settings shows that reflection specificity assessment using LLMs lags behind using supervised models, with nearest-neighbor in-context learning offering the best LLM scoring performance.

6 RQ2 Analyzing Self-Generated Explanations

Although self-generated explanations did not improve specificity assessment, we explore whether they offer added interpretability benefits beyond those of traditional supervised models. To systematically assess the quality of these explanations as interpretability tools, we examine three key dimensions. Two are adapted from free-text rationale evaluation criteria (Wiegreffe and Marasovic), focusing on surface-level linguistic qualities: **vocabulary overlap** (to capture repetition or label leakage) and **faithfulness** (to assess alignment with the input). We also incorporate **fidelity analysis** (Wachsmuth et al., 2017; Gilpin et al., 2018) to evaluate whether the model's predictions are truly guided by its own chain-of-thought, thus reflecting internal consistency in reasoning.

6.1 Vocabulary Overlap Analysis

LLMs often leak the predicted label within explanations (Wiegreffe et al., 2021b; Hase et al., 2020b), raising concerns that generated rationales may merely restate the expected output rather than provide meaningful reasoning. Similarly, Elaraby et al. (2024) demonstrated that, in assessing argument quality through pairwise ranking, LLM-generated explanations often exhibit redundancy by merely restating the input argument, rendering the self-generated explanations meaningless. We extend this analysis, investigating whether explanations contain excessive lexical overlap with the input reflections, thereby reducing their utility in providing a meaningful interpretibility for the scores. We leverage the formula in Ye and Durrett (2022) which was mainly used for ensuring that explanations are relevant to the input. Let a reflection R consist of a sequence of words $\mathcal{R} = (r_1, \ldots, r_n)$ and a generated explanation E consist of a sequence of words $\mathcal{E} = (e_1, \ldots, e_m)$, where n and m are the respective word lengths. We quantify lexical overlap $\mathcal{V}$ as:

$$\mathcal{V}(E, R) = \frac{|\mathcal{E} \cap \mathcal{R}|}{|\mathcal{E}|}$$

A higher value indicates greater redundancy between the explanation and the input reflection.

6.2 Faithfulness Analysis

Prior work (Ye and Durrett, 2022) highlights that self-generated explanations may be unfaithful to the input, introducing hallucinated or contradicting information. To assess whether an explanation E remains faithful to its reflection R, we utilize an off-the-shelf entailment model. Specifically, we use a pretrained RoBERTa model (Liu et al., 2019)

fine-tuned on the MNLI dataset (Williams et al., 2018)[7]. We frame this as a natural language inference (NLI) task, where the reflection serves as the *premise* and the corresponding explanation as the *hypothesis*. An entailment model is then used to predict whether the explanation *entails*, *contradicts*, or is *neutral* with respect to the input reflection. We compute the percentage of contradictions across all explanations.

6.3 Fidelity Analysis

Fidelity evaluates whether LLM-generated explanations genuinely influence the model's predictions (Gilpin et al., 2018). Following the counterfactual reasoning methodology introduced by Wachter et al. (2017), we assess fidelity by introducing misleading explanations and measuring the percentage of predictions that are affected. Specifically, we consider a set of generated explanations E for which the model's predictions align with human-labeled specificity scores. Rather than manually creating corrupted explanations, we generate a misleading set E^{mislead} by prompting `GPT-4o` to rewrite each original explanation to justify an incorrect rubric score.[8] The final fidelity $F(E)$ is:

$$F(E) = \frac{\sum_{r_i \in R} \mathbb{I}\left[M(r_i, e_i^{\text{mislead}}) \neq \text{score}^{\text{labeled}}(r_i))\right]}{|R|}$$

where $M(r_i, e_i)$ is the model's predicted specificity score for reflection r_i given explanation e_i, and $\text{score}^{\text{labeled}}(r_i)$ is the ground truth label. $0 \leq F(E) \leq 1$, where 0 indicates that LLMs do not rely on explanations while 1 refers to LLMs fully utilizing explanations in predicting scores. We sampled 100 explanations per score category (400 in total) from the `Mistral-8B-instruct` CoT outputs. For each original explanation and its misleading counterpart, we re-prompted `Mistral` to generate new predictions. Fidelity was then measured based on whether the model's prediction deviated from the original ground truth. Table 7 presents an example of a corrupted explanation[9].

6.4 RQ2 Findings

Table 8 summarizes the self-explanation analyses. On average, explanations exhibited a vocabulary overlap of 23.55% with their corresponding reflections, suggesting that the model does not simply replicate the input text—contrary to findings in argument quality tasks where smaller instruction-tuned models often repeated the input (Elaraby et al., 2024). Examples in Table 9 illustrate that high-overlap cases (71.11%) and low-overlap cases (8.88%) primarily reflect restatements of key concepts rather than superficial repetition.

Faithfulness analysis shows that only 5.48% of explanations were predicted to contradict their corresponding reflections, indicating high alignment. As shown in Table 10, even explanations flagged as contradictions often did not exhibit explicit conflict, suggesting potential entailment model errors due

Reflection: *I didn't know which equations went into which sections of the WDM.*
Ground Truth Specificity Label: 3
Original Explanation: The reflection mentions specific equations and a specific section of the WDM (Worked Distributed Model), which indicates a specific understanding of the course content. However, it does not provide detailed explanations or examples of the confusion, which is why it scores a 3.
Corrupted Explanation: The reflection explicitly mentions specific equations and a particular section of the WDM (Worked Distributed Model), demonstrating a clear and detailed understanding of the course content. The mention of specific equations and sections indicates a precise engagement with the material, which aligns with a more detailed and specific reflection. Therefore, it deserves a score of 4.

Table 7: Example of a corrupted explanation for specificity assessment. Gray text indicates the modified portion of the original explanation used to steer the LLM toward a different score.

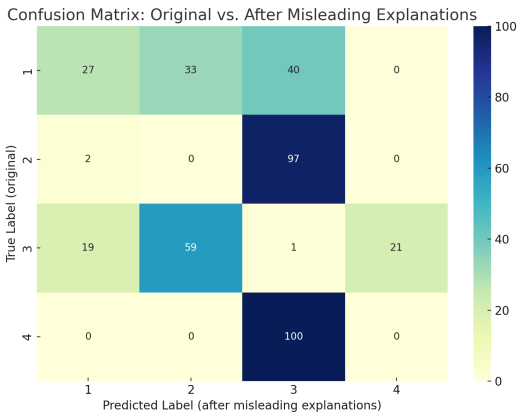

Figure 1: Confusion matrix between original labels and corrupted labels after misleading explanations.

[7] https://huggingface.co/roberta-large-mnli

[8] We set the temperature to 0.7 to encourage diverse generations. The full prompt is in Appendix C Table 13.

[9] The prompt used for generating predictions is included in Appendix C Table 14.

Average Vocabulary Overlap (%) ↓	Contradiction Rate (%) ↓	Fidelity in % $F(E) \times 100$ ↑
23.55 ± 14.23	5.48	93.00

Table 8: Self-explanation analysis results for `Mistral-8B-instruct+CoT`. (↑) means that higher values correspond to better explanation quality, while (↓) means that lower values correspond to better quality.

Vocabulary Overlap

High Overlap Example

Reflection:
What I found most interesting in today's lecture was the conceptual questions. The sled question tricked me a little so it was interesting to see what the answer was and why.
Explanation:
The reflection mentions a specific aspect of the lecture conceptual questions and provides a detailed explanation of why it was interesting the sled question tricked me a little so it was interesting to see what the answer was and why.
Overlap Percentage: 71.11%
Specificity Scores: annotated: 4 predicted: 4

Low Overlap Example

Reflection:
Cramer's rule and matrices.
Explanation:
The reflection mentions Cramer's rule and matrices, which are specific topics from the course content. However, it does not provide any details about what was confusing or needed more explanation, so it falls into the 'general' category.
Overlap Percentage: 8.88%
Specificity Scores: annotated: 3 predicted: 3

Table 9: Vocabulary overlap analysis between reflections and explanations. Highlighted text indicates overlapping segments.

Faithfulness Analysis

Contradiction Example

Reflection:
I was confused about how to use a calculator to solve matrices.
Explanation:
The reflection mentions a specific tool (calculator) and a specific topic (solving matrices), but it does not provide any details about the confusion or what was unclear. It is a general statement about the confusion without specific details.
Entailment Model (RoBERTa-NLI) Decision: Contradiction
Specificity Scores: annotated: 3 predicted: 3

Entailment Example

Reflection:
The circuits and how they decide evenly.
Explanation:
The reflection mentions 'circuits' and 'how they decide evenly', which are specific topics from the course content. However, it lacks detailed explanation or specific examples, making it a general reflection.
Entailment Model (RoBERTa-NLI) Decision: Entailment
Specificity Scores: annotated: 2 predicted: 3

Table 10: Faithfulness analysis of reflections and explanations based on entailment model predictions.

to surface-level lexical differences (e.g., "specific tools" vs. "general statement").

Finally, fidelity analysis on our random sample yielded a score of 93.00%, confirming that the model strongly conditions its predictions on the provided explanations. As discussed in Section 5.3, this reliance may partially explain the observed performance degradation under chain-of-thought (CoT) prompting. Figure 1 shows how corrupted explanations lead to prediction shifts, causing originally correct predictions—referred to in the figure as "True Labels" (i.e., those aligned with the ground truth)—to become incorrect.

In sum (and as summarized for RQ2 in Table 2), the overlap and faithfulness analyses along with the example generations suggest that LLMs can potentially generate meaningful interpretations for their scores. Their personalized nature in fact makes them potentially well-suited for integration into reflection writing systems such as `CourseMIRROR`(Magooda et al., 2022), where scaffolding helps students identify missing details and improve reflection specificity. For example, `CourseMIRROR` provides fixed prompts based solely on predicted specificity scores (e.g., *"Sounds good, can you please tell us why it is confusing?"*), while dynamically produced explanations can potentially convey a deeper, reflection-specific understanding, identifying underlying concepts that contribute to specificity. Finally, the fidelity analysis highlights that the CoT explanations not only accompany but also influence the model's final predictions, reinforcing their reliability as interpretability tools.

7 Conclusion and Future Work

In this study, we systematically analyzed the potential of LLMs as explainable specificity evaluators for student-generated reflections, evaluating three instruction-tuned models in zero-shot and few-shot settings against supervised baselines. Our findings reaffirm prior research that LLM-based evaluation of educational texts still lags behind supervised models, with nearest-neighbor retrieval offering only marginal improvements in alignment with human annotations. Chain-of-thought prompting does not enhance specificity assessment either, suggesting that self-generated explanations do not meaningfully influence model decision-making. However, we extend prior analyses by focusing on *evaluating generated self-explanations*, an emergent capability that is underexplored in the context of educational text assessment. Our analysis reveals that self-explanations can enhance interpretability by providing faithful justifications for model's scores and to the input reflections.

Future work should explore alignment techniques—including fine-tuning with annotated corpora and self-alignment strategies—to improve the utility of LLMs in student specificity assessment. Additionally, the role of self-generated explanations should be further investigated for their potential to deliver automated, personalized feedback to students, enhancing both the interpretability and pedagogical value of LLM-based evaluation.

Acknowledgment

This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A180477, and the National Science Foundation through Grants 2329273 and 2329274. The opinions expressed are those of the authors and do not represent the views of the U.S. Department of Education or the National Science Foundation. This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. We want to thank the members of the Pitt PETAL group, Pitt NLP group, the CourseMIRROR group, and anonymous reviewers for their valuable comments in improving this work.

Limitations

This work focuses on instruction-tuned LLMs with comparable parameter sizes, allowing for a controlled comparison; however, this design choice may limit the generalizability of our findings. Future research should explore models of varying scales to better understand the impact of model size on specificity assessment performance. Moreover, our analysis is restricted to a particular genre of reflective writing—short student reflections written in response to structured prompts. Expanding the evaluation to include other forms of reflective writing, such as longer essays or open-ended journal entries, would offer a more comprehensive understanding of LLM capabilities across diverse contexts. Lastly, our examination of generated explanations was limited to surface-level properties, including use of an off-the-shelf entailment model not designed for reflections. Additionally, we did not analyze the correlation between self-explanations and other black-box explanation methods, such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). Future work could incorporate human-centered studies to evaluate the effectiveness of these explanations in delivering personalized feedback to students based on their reflections.

Ethical Considerations

This study uses publicly available, anonymized student reflection data from the `ReflectSumm` and `CourseMIRROR` datasets. All experiments were conducted in accordance with data usage terms, and no personally identifiable information was used.

References

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2):i–21.

John R Baird, Peter J Fensham, Richard F Gunstone, and Richard T White. 1991. The importance of reflection in improving science teaching and learning. *Journal of research in Science Teaching*, 28(2):163–182.

Sami Baral, Eamon Worden, Wen-Chiang Lim, Zhuang Luo, Christopher Santorelli, Ashish Gurung, and Neil Heffernan. 2024. Automated feedback in math education: A comparative analysis of llms for open-ended responses. *arXiv preprint arXiv:2411.08910*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer

grading. *International journal of artificial intelligence in education*, 25:60–117.

Dan Carpenter, Michael Geden, Jonathan Rowe, Roger Azevedo, and James Lester. 2020. Automated analysis of middle school students' written reflections during game-based learning. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21*, pages 67–78. Springer.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. Rev: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030.

Mohamed Elaraby, Diane Litman, Xiang Lorraine Li, and Ahmed Magooda. 2024. Persuasiveness of generated free-text rationales in subjective decisions: A case study on pairwise argument ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14311–14329, Miami, Florida, USA. Association for Computational Linguistics.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020a. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020b. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.

Zhaoyi Joey Hou, Alejandro Ciuba, and Xiang Lorraine Li. 2025. Improve llm-based automatic essay scoring with linguistic features. *arXiv preprint arXiv:2502.09497*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

David Kember, Jan McKay, Kit Sinclair, and Frances Kam Yuet Wong. 2008. A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & evaluation in higher education*, 33(4):369–379.

Alexander Knoth, Alexander Kiy, Ina Müller, and Mathias Klein. 2020. Competences in context: Students' expectations and reflections as guided by the mobile application reflect. up. *Technology, Knowledge and Learning*, 25(4):707–731.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Gen Li, Li Chen, Cheng Tang, Valdemar Švábenskỳ, Daisuke Deguchi, Takayoshi Yamashita, and Atsushi Shimada. 2025. Single-agent vs. multi-agent llm strategies for automated student reflection assessment. In *Proceedings of the 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2025)*.

Junyi Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Wencan Luo and Diane J Litman. 2016. Determining the quality of a student reflective response. In *FLAIRS*, pages 226–231.

Ahmed Magooda. 2022. *Techniques To Enhance Abstractive Summarization Model Training for Low Resource Domains*. Ph.D. thesis, University of Pittsburgh.

Ahmed Magooda, Diane Litman, Ahmed Ashraf, and Muhsin Menekse. 2022. Improving the quality of students' written reflections using natural language processing: Model design and classroom evaluation. In *International Conference on Artificial Intelligence in Education*, pages 519–525. Springer.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

Muhsin Menekse. 2020. The reflection-informed learning and instruction to improve students' academic success in undergraduate classrooms. *The Journal of Experimental Education*, 88(2):183–199.

Muhsin Menekse, Alfa Satya Putra, Jiwon Kim, Ahmed Ashraf Butt, Mark McDaniel, Ido Davidesco, Michelle Cadieux, Joe Kim, and Diane Litman. 2025. Enhancing student reflections with natural language processing based scaffolding: A quasi-experimental study in a large lecture course. *Computers and Education: Artificial Intelligence*, page 100397.

Muhsin Menekse, Glenda Stump, Stephen J Krause, and Michelene TH Chi. 2011. The effectiveness of students' daily reflections on learning in an engineering context. In *2011 ASEE Annual Conference & Exposition*, pages 22–1451.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Mark D Shermis and Joshua Wilson. 2024. *The Routledge international handbook of automated essay evaluation*. Taylor & Francis.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,

et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Sarah Wiegreffe and Ana Marasovic. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021a. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021b. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pascal Wilhelm. 2021. Fostering quality of reflection in first-year honours students in a bachelor engineering program technology, liberal arts & science (atlas). *Journal of Higher Education Theory and Practice*, 21(16):72–91.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yang Zhong, Mohamed Elaraby, Diane Litman, Ahmed Ashraf Butt, and Muhsin Menekse. 2024. ReflectSumm: A benchmark for course reflection summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13819–13846, Torino, Italia. ELRA and ICCL.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. 2024. A LLM-based ranking method for the evaluation of automatic counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9572–9585, Miami, Florida, USA. Association for Computational Linguistics.

A Prompts for specificity evaluation

Table 11 shows the exact prompt used in our experiments. The prompt includes few-shot examples, which are only included in the case of few-shot specificity scoring.

Scoring Prompt
Background:
A group of students in a classroom were asked to describe what they found interesting or confusing in a lecture.
Task:
You will be given the original prompt to the students, followed by a single reflection written by a student. Your task is to score the reflection from 1 to 4 based on the given specificity rubric.
Rubric:
Score 1 (vague): Reflection implies "no confusing issue," e.g., "nothing" or "none for this class."
Score 2 (non-specific): Reflection does not include any statement(s) about course content but refers to organizational aspects (e.g., homework, exams).
Score 3 (general): Reflection includes statement(s) about course content but lacks specific details.
Score 4 (specific): Reflection includes specific and detailed statement(s) about course content.
Few-Shot Reflection Examples (Only in case of few-shot):
`{reflections_with_scores}`
Input Example:
`{ "prompt": "{prompt}", "reflection": "{reflection}" }`
Output Format:
Return only the score in a valid JSON format: `{ "score": "1, 2, 3, or 4" }`

Table 11: Specificity scoring prompt with rubric and in-context examples.

B Fixed reflections examples

Table 12 shows examples of fixed reflections included in the prompt for the **fixed in-context reflection** experiments.

C Prompts for fidelity analysis

Table 13 presents the prompt used to generate misleading explanations by corrupting the original explanation that supported the correct score.

Table 14 presents the modified prompt used to compute final fidelity. The prompt incorporates corrupted explanations as part of the input and instructs the model to output only the predicted score.

Score	Score Meaning	Reflection Example	Prompt Type
1	Vague	Not sure if I understand	Confusing
1	Vague	Elephant stampede in a rainstorm.	Confusing
1	Vague	teacher bringing chocolates to class	Interesting
1	Vague	Made some kind of sense	Interesting
2	Non-specific	size of print and colors are hard to read	Confusing
2	Non-specific	I tried to follow along but I couldn't grasp the concepts. Plus it's hard to see what's written on the white board when the projector shines on it	Confusing
2	Non-specific	Examples were interesting	Interesting
2	Non-specific	lzw compression and expansion	Interesting
3	General	I didn't understand the attractive and repulsive force graphs from the third slide	Confusing
3	General	The repulsive/ attraction charts	Confusing
3	General	the history of founder of student distribution was interesting	Interesting
3	General	the transformations between random variables was interesting	Interesting
4	Specific	Part III on worksheet in class, comparing metals. I was confused about why each metal was selected	Confusing
4	Specific	computing length, edges and atomic packing factor for FCC	Confusing
4	Specific	Learning the where the n-1 degrees of freedom coming in the sample variance distribution was very interesting	Interesting
4	Specific	the process of deciding among differen population estimators was quite interesting	Interesting

Table 12: Fixed reflections for in-context specificity scoring.

Corrupted Explanation Generation Prompt

Background:

Students in a classroom were asked to reflect on a lecture by describing what they found interesting or confusing.

Task:

You will be provided with:

- The original prompt given to the students.
- A reflection written by a student.
- A specificity score assigned to the reflection based on a predefined rubric.
- An explanation justifying this score.

Your goal is to generate an **alternative explanation** that supports a **different specificity score** for the same reflection. The new explanation should maintain a similar style to the given justification but justify a different score.

Rubric:

Score 1 (vague): Reflection implies "no confusing issue," e.g., "nothing" or "none for this class."

Score 2 (non-specific): Reflection does not include any statement(s) about course content but refers to organizational aspects (e.g., homework, exams).

Score 3 (general): Reflection includes statement(s) about course content but lacks specific details.

Score 4 (specific): Reflection includes specific and detailed statement(s) about course content.

Input:

```
{ "prompt": {prompt}, "reflection": {reflection}, "explanation": {explanation}, "label": {label} }
```

Instructions:

- Construct a **new explanation** that justifies a **different specificity score** than the original label.
- Maintain a **logical structure and tone** similar to the provided explanation.
- **Output only** the alternative explanation.

Table 13: Prompt for generating corrupted explanations to support alternative specificity scores while maintaining logical tone and style.

Score with Predefined Explanations

Background:

A group of students in a classroom were asked to describe what they found interesting or confusing in a lecture.

Task:

You will be given the original prompt provided to the students, followed by a reflection written by a student. Your task is to score each reflection from 1 to 4 based on the given specificity rubric.

Rubric:

Score 1 (vague): The reflection implies "no confusing issue", e.g., responses like "nothing" or "none for this class."

Score 2 (non-specific): The reflection does not mention course content (e.g., lecture slides, in-class activities, or discussion) but refers to class organization or assignments (e.g., homework, exams).

Score 3 (general): The reflection mentions course content but lacks detailed or specific statements.

Score 4 (specific): The reflection includes both course content and specific, detailed statements.

Input:

```
{ "prompt": {prompt}, "reflection": {reflection} }
```

Explanation:

```
{explanation}
```

Instruction:

Therefore, determine the score based on the explanation and reflection. **Answer with the score only.**

Table 14: Prompt for scoring reflections based on predefined explanations, using the specificity rubric.