Two Heads Are Better Than One: Dual-Model Verbal Reflection at Inference-Time

Anonymous ACL submission

Abstract

Large Language Models (LLMs) often struggle with complex reasoning scenarios. While preference optimization methods enhance reasoning performance through training, they often lack transparency in why one reasoning outcome is preferred over another. Verbal reflection techniques improve explainability but are limited in LLMs' critique and refinement capacity. To address these challenges, we introduce a contrastive reflection synthesis pipeline that enhances the accuracy and depth of LLM-generated reflections. We further propose a dual-model reasoning framework within a verbal reinforcement learning paradigm, decoupling inference-time self-reflection into specialized, trained models for reasoning critique and refinement. Extensive experiments show that our framework outperforms traditional preference optimization methods across all evaluation metrics. Our findings also show that "two heads are better than one", demonstrating that a collaborative Reasoner-Critic model achieves superior reasoning performance and transparency, compared to single-model approaches.

1 Introduction

Large Language Models (LLMs) often struggle with complex reasoning tasks that require nuanced intermediate steps and explicit feedback (Li et al., 2023). Recent advancements in preference optimization such as Direct Preference Optimization (DPO, Rafailov et al. 2023) present a promising approach to align the model's reasoning outputs with human preferences (Pang et al., 2024; Lai et al., 2024). These methods typically train LLMs as a *Single Reasoner* using preference pairs to learn from implicit rewards. While effective, they are often insufficient to provide explicit feedback to explain **why** one response is preferred over another (Rafailov et al., 2024; Lu et al., 2024a; Chowdhury et al., 2024) and require human labels.

A distinct line of research has aimed to improve reasoning with Verbal Reinforcement Learning (VRL) (Shinn et al., 2023; Kumar et al., 2025; Snell et al., 2025). These studies generally follow a paradigm of alternating between two generative steps within the same model to improve reasoning, without requiring additional training. While these methods can generate explicit reasoning reflection by iteratively refining reasoning errors through verbal critique (Wei Jie et al., 2024), they still face challenges in accurately identifying errors and providing actionable critiques for reasoning traces (Kamoi et al., 2024b) (Figure 1 (a)). This difficulty arises from a **system-level** conflict inherent to LLMs-their dual roles in both reflecting on errors and refining them (Huang et al., 2024).

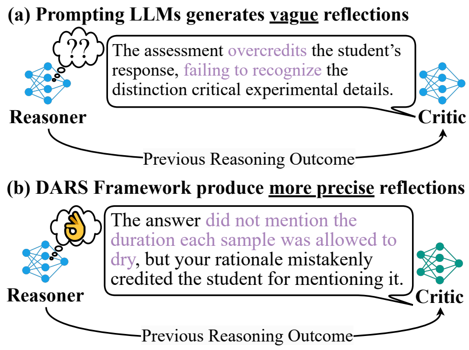

Figure 1: **LLM evaluation and refinement.** (a) Existing inference-time LLM reflection methods use the same model for reflection and refinement, often producing **vague feedback** that hinders effective refinement; (b) Our DARS framework trains two distinct models for reflection and refinement on synthetic error correction data, enabling more **actionable critiques** and improving refinement for explainable student answer scoring.

A crucial aspect of self-reflection and refinement in LLMs is their ability to detect and correct errors. However, at the **data-level**, datasets explicitly annotated with error detection and correction traces are scarce, making it difficult to train models that can systematically identify and address their own mistakes (Liu et al., 2024a). To address these challenges, we focus on Automated Student Answer Scoring (ASAS), a task requiring complex reasoning to compare student answers with key answer elements and derive scores based on marking rubrics. To overcome the **data-level** challenge, we propose a *contrastive reflection synthesis* pipeline (§4.1) that generates precise verbal reflection data (i.e., error correction instructions) by analyzing discrepancies in multi-step reasoning paths.

Our method leverages structured thought trees to formalize assessment rationales. Given a student response and key answer elements, we construct a tree through progressive comparisons, producing binary decisions (i.e., presence or absence of each element in student's answer). By contrasting discrepancy between assessed paths in this tree, we can systematically identify mismatches in key element assessments to signal potential errors in rationales. These differences then serve as structured prompts for an LLM to generate explicit error correction instructions as verbal reflection (e.g., "*The answer did not mention the duration each sample was allowed to dry, but your rationale mistakenly credited the student for mentioning it*.").

Furthermore, to solve the **system-level** role conflict, we propose DARS, a *Dual-model Reflective Scoring framework* (§4.2) with specialized, trained *Reasoner* and *Critic* models that perform refinement and reflection, respectively. Our *Critic* innovatively integrates both process reward modeling (providing detailed reflection for reasoning steps) and outcome reward modeling (validating overall reasoning outcome correctness) *without relying on human labels* (Figure 1 (b)). Our framework shows effectiveness in refining LLM reasoning process[1].

In summary, our contributions are threefold:

1. We propose an effective ***contrastive reflection synthesis*** pipeline to generate error detection and correction instructions as verbal feedback from binary reasoning preference pairs.
2. Built on the synthetic reflection data, we propose a ***dual-model reasoning framework*** **DARS** consisting of a *Reasoner* and a *Critic* to perform more effective inference-time VRL.
3. We have several novel empirical insights:
   - Our dual-model reasoning framework ***outperforms*** single Reasoner-based preference optimisation and maintained balanced performance across all metrics even in data scarcity.
   - Human evaluation confirms that Critic-generated reflection provides ***actionable guidance*** that the Reasoner could ***reliably follow***.
   - Contrary to prior findings, our experiments show that ***increasing the size of the Critic model leads to better results than scaling the Reasoner***.

[1]Our synthetically generated training data and source code will be made available for reproducibility.

2 Related Works

Verbal Reinforcement Learning for Self-Reflection VRL has emerged as a promising approach for enhancing LLM reasoning at inference time (Huang et al., 2024; Kamoi et al., 2024b). Early methods relied on self-reflection mechanisms where LLMs refined outputs using contextual cues (Chen et al., 2024b; Jiang et al., 2023; Welleck et al., 2023). However, studies show that LLMs struggle to self-correct reliably (Li et al., 2024b; Tyen et al., 2024; Chen and Shu, 2024; Kamoi et al., 2024a). To address this, trained critic models have been used to generate verbal feedback for LLM correction (Welleck et al., 2023; Akyurek et al., 2023; Paul et al., 2024), though they primarily focus on single-step feedback. More complex reasoning tasks typically rely on Oracle labels for correction (Shinn et al., 2023; Kim et al., 2023). Our work introduces a dual-model framework where a Critic independently provides more detailed, trace-level reflections, eliminating the need for Oracle labels in verification.

Explainable Automated Student Answer Scoring ASAS is traditionally treated as a text classification problem (Larkey, 1998; Taghipour and Ng, 2016), with efforts to improve transparency via feature analysis (Dong and Zhang, 2016; Vanga et al., 2023) and attention visualization (Alikaniotis et al., 2016; Yang et al., 2020). Recent approaches incorporate rationale generation for enhanced explainability and transparency (Li et al., 2023) but often underperform compared to classification-based methods. Li et al. (2024a) proposed a thought tree framework to model human assessment processes, leveraging LLMs for structured scoring rationales. Our work builds upon this by not only explaining decisions but also improving the transparency of assessment refinement process, through iterative LLM reasoning improvements.

3 Preliminary

Existing ASAS systems primarily aim to automate teachers' complex reasoning processes on the assessment of short answer questions, typically operating within a classification paradigm (Larkey, 1998; Dong et al., 2017). These systems take various contextual input, including *question prompts*, *key answer elements* (e.g., keywords or phrases that qualify for marks), *marking rubrics* (e.g., criteria for assigning scores), and *student responses*, and are trained to predict a *score* as output.

Given a specific question, the dataset can be represented as $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i denotes a student's response and y_i represents the corresponding score assigned by human assessors. For each question, let $\mathcal{K} = \{k_j\}_{j=1}^{M}$ represent the set of key answer elements, where M is the number of distinct elements expected in a complete answer. The scoring process can be formalized using a question-specific scoring function $f_r(\cdot)$, which determines the final score based on the extend to which student's response includes the required elements:

$$y_i = f_r(\mathbf{v}(x_i, \mathcal{K})), \tag{1}$$

where $\mathbf{v}(x_i, \mathcal{K}) \in \mathbb{R}^M$ is a multi-hot vector indicating the presence of each key element $k_j \in \mathcal{K}$ in the student response x_i. This coverage vector is then mapped to the final score through f_r.

Li et al. (2024a) proposed a structured thought tree to imitate the human assessment process (as illustrated in Figure 2) and generate assessment rationales. We define thought trees as $\mathcal{T} = \{\text{path}_l\}_{l=1}^{d}$, where each path_l represents a structured decision path, capturing a unique combination of binary assessments for key answer elements (e.g., Figure 2). Each path is encoded as:

$$\hat{\mathbf{v}}(\mathcal{Z}_l) = [z_1^{(l)}, z_2^{(l)}, \ldots, z_M^{(l)}], \tag{2}$$

where $z_j^{(l)} \in \{0, 1\}$ indicates whether the jth key element is correctly included or not. Prior work assumes that paths leading to a correct score is the *preferred* or *chosen* path ($\text{path}_l^{\text{chosen}}$), while paths resulting in an incorrect score is the *rejected* path ($\text{path}_l^{\text{reject}}$). The rationales r_{chosen} and r_{reject} are then derived by summarizing the intermediate decisions along their respective paths.

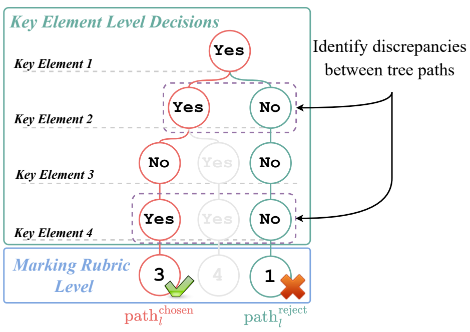

Figure 2: **Illustration of a thought tree and the identification of discrepancies in reasoning paths.**

4 DARS: Dual-Model Reflective Scoring

4.1 Contrastive Reflection Synthesis

Existing approaches to enhancing LLM reasoning capabilities often rely on preference optimisation methods (Lu et al., 2024b; Chen et al., 2024a), which are based on datasets annotated using binary pairwise comparisons under the Bradley-Terry model (Bradley and Terry, 1952). However, these methods typically lack transparency, as they do not explain **why** one response is preferred over another.

This opacity makes it difficult for humans to understand or refine the reasoning process of LLMs. Furthermore, generating synthetic verbal reflections remains a challenge. Simple prompting of LLMs (e.g., GPT-4) often results in vague, superficial, or unhelpful rationales (Yin et al., 2024; Jiang et al., 2025)[2]. To address this, ***we propose a pipeline that generates meaningful reflections based on a structured thought tree to explain why*** r_{reject} ***is less preferable than*** r_{chosen}.

Step 1: Identify Discrepancy in Reasoning Paths

We first obtain a thought tree for each student's response x_i, and analyze the discrepancy between rationale preference pairs by comparing the differences between their corresponding multi-hot vectors (as shown in Figure 2):

$$\Delta\mathbf{v} = \hat{\mathbf{v}}(\mathcal{Z}_l^{\text{chosen}}) - \hat{\mathbf{v}}(\mathcal{Z}_l^{\text{reject}}), \tag{3}$$

where $\Delta\mathbf{v}$ is the difference vector highlighting discrepancy between $\text{path}_l^{\text{chosen}}$ and $\text{path}_l^{\text{reject}}$. Each dimension Δ_j in $\Delta\mathbf{v}$ corresponds to a key element, specifically:

$$\Delta_j = \begin{cases} 1 & \text{if decision for } k_j \text{ changed from 0 to 1,} \\ -1 & \text{if decision for } k_j \text{ changed from 1 to 0,} \\ 0 & \text{if decision is the same.} \end{cases}$$

For each key element where $\Delta_j \neq 0$, we construct a hint prompt[3] that highlights the differences in the intermediate assessment decisions (e.g. r_{reject} missed k_j that the student has already included):

$$\text{hint}_{\Delta\mathbf{v}} = \text{Prompt}(\Delta\mathbf{v}, K). \tag{4}$$

Step 2: Generate Synthetic Reflections

After identifying discrepancies and constructing the hint prompt, we prompt an LLM (e.g., GPT-4-turbo) to generate a verbal reflection between the preference pair r_{reject} and r_{chosen}:

$$r_{\text{reflect}} = \mathsf{LLM}_\theta(x_i, r_{\text{reject}}, r_{\text{chosen}}, \text{hint}_{\Delta\mathbf{v}}), \tag{5}$$

[2]Empirical experiment analysis for this is provided in §5.2.
[3]A detailed prompt template is provided in §A.

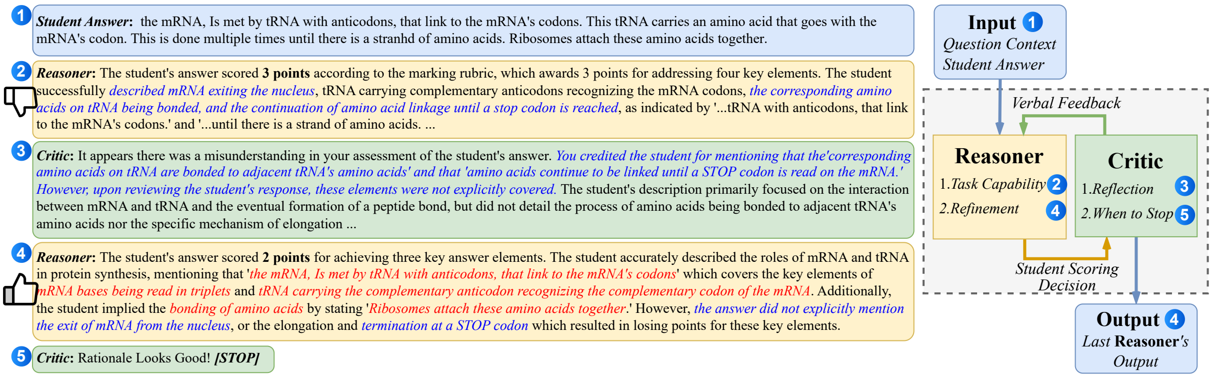

Figure 3: **(Left)**: Example conversation between the trained Reasoner and Critic. **(Right)**: Illustration of our DARS framework. A detailed explanation of this example is provided within training details in §4.2. The context related to the Reasoner's initial mistake is highlighted in blue, while the refinement is marked in red. Some question context is omitted in ①, but the full example can be found in §B.1. Here, ① represents the framework's input, while the final response from the Reasoner before the Critic's termination (④) serves as the framework's output.

where the LLM synthesizes a verbal reflection to guide another LLM in refining its reasoning by transitioning from the incorrect rationale r_{reject} to the correct rationale r_{chosen}.

4.2 Dual-Model Reasoning

We propose an on-policy dual-model framework comprising a **Reasoner** ($\mathcal{R}$) and a **Critic** ($\mathcal{C}$), where $\mathcal{C}$ provides explicit verbal reflections to guide $\mathcal{R}$'s reasoning process. Unlike traditional reward models, which rely on scalar scores or implicit preference rankings, our approach incorporates verbal reflections to iteratively improve the outputs of $\mathcal{R}$.

However, since verbal reflection is inherently non-differentiable, we use a sampling-based approach to train the Reasoner and Critic models independently, transforming it into an off-policy process. During inference, these models collaborate through an iterative VRL process. Our framework not only enhances transparency in the ASAS task but also improves the reasoning process. Figure 3 illustrates our framework with an example illustrating how the Reasoner progressively refines its reasoning based on the Critic's detailed reflection.

Training Reasoner and Critic Models

Build on the synthetic reflection data generated in §4.1, we create diverse data combinations to train $\mathcal{R}$ and $\mathcal{C}$ on refinement and reflection capabilities:

Reasoner ($\mathcal{R}$) The training data for the Reasoner is designed to include two capabilities:

1. **Task Capability**: $\mathcal{R}$ takes ① (question context and student answer) as input, and predicts ② (an initial assessment, e.g., r_{reject} or r_{chosen}).
2. **Refinement**: $\mathcal{R}$ takes ① & ② (history of assessment, e.g., r_{reject}), with ③ ($\mathcal{C}$ generated reflection) as input, and predict ④ (an refined assessment, e.g., r_{chosen}).

Critic ($\mathcal{C}$) The training data for the Critic is designed to include two capabilities:

1. **Reflection**: If the assessment is incorrect, $\mathcal{C}$ is trained to take previous assessment histories (e.g., ①-② or ①-④) as input, and predict ③ (a reflection instruction r_{reflect} for $\mathcal{R}$) as output.
2. **When to Stop**: $\mathcal{C}$ takes $\mathcal{R}$'s previous assessment outcome (e.g., r_{reject} or r_{chosen}), either from single-round ①-② or multi-rounds ①-④ as input, and validate the correctness of the assessment. If the assessment is correct, $\mathcal{C}$ predict ⑤, a special token that signals the termination of the reasoning loop and outputs the final assessment generated by $\mathcal{R}$.

Our Critic model is innovatively trained to perform both *process reward modelling (Reflection)* (Lightman et al., 2024; Wang et al., 2024) and *outcome reward modelling (When to Stop)* (Ouyang et al., 2022), without the need for comparisons with oracle labels (Shinn et al., 2023; Kim et al., 2023) or manually setting the maximum iterations to terminate the refinement process (Madaan et al., 2023; Li et al., 2024b).

Inference-Time Iterative Refinement

Once the Reasoner and Critic models are trained, they could collaborate to refine the assessment rationale at inference time through iterative conversations. At each iteration step t, $\mathcal{R}$ generates an assessment trajectory $\hat{y}_r^0, \hat{y}_r^1, \ldots, \hat{y}_r^T$:

Initialization: $\hat{y}_r^0 = \mathcal{R}(x_i)$

Iterative Reflection:

$$\begin{cases} \hat{y}_r^{(t+1)} = \mathcal{R}\Big(\hat{y}_r^t, \mathcal{C}(\hat{y}_r^t)\Big), & \text{if } \mathcal{C}(\hat{y}_r^t) = \text{Reflection}, \\ \hat{y}_r^T = \hat{y}_r^t, & \text{if } \mathcal{C}(\hat{y}_r^t) = \text{[STOP]}. \end{cases}$$

Here, $\mathcal{C}(\cdot)$ checks the correctness of $\hat{y}_r^t$. If refinement is needed, it generates a verbal reflection for $\mathcal{R}$ to refine $\hat{y}_r^t$, otherwise, a "Terminate" signal is triggered, and the final assessment $\hat{y}_r^T$ is taken as the output.

5 Experiments

5.1 Experimental Setup

Datasets We use two data sources, consisting of a total of six different datasets, for our experiments: (1) The Hewlett Foundation Short Answer Scoring (ASAP) dataset (Hamner et al., 2012), which contains short essay responses across science and biology topics. (2) A private dataset comprising student responses to biology exam questions, where human-assigned scores are provided from a reputable examination service. We present dataset statistics in Table A1.

Evaluation Metrics We evaluate the assessment performance using Accuracy (ACC), macro F1 (F1), and Quadratic Weighted Kappa (QWK)[4].

Baselines We compare with four baselines:

PLM Classifier: A text classifier built on a pre-trained `Deberta-v3-large` model (He et al., 2023) and fine-tuned on various datasets.

SFT: A Reasoner-only, supervised fine-tuning baseline trained with datasets released by (Li et al., 2024a) (e.g, takes ① as input, predicts ②).

DPO: A DPO approach that performed preference optimization with synthetic reasoning preference data as presented in (Li et al., 2024a) (e.g, takes ① as input, optimize ④≻②). The base model used is the SFT baseline.

Dual w/ GPT-4 A dual-model inference time VRL baseline (Dong et al., 2024), where Reasoner is trained within our framework, and `gpt-4-turbo` is used as the Critic to give verbal reflection instructions (e.g, ③&⑤ are generated by GPT-4).

We provide further details about the experiment setup in §A.

5.2 Overall Comparison

In this section, we provide a comprehensive evaluation of both scoring performance and rationale quality. As shown in Table 1, we compare our dual-model reasoning framework (`DARS`) against four baselines, including both classification and generative approaches. All methods, including ours, were trained using the same LLaMA 3B model. Our results indicate that ***"two heads are better than one"*** – *our framework overcomes the data scarcity issue, maintains balanced improvements across all evaluation metrics and outperforms state-of-the-art Reasoner-only preference optimization method*. Furthermore, our Critic model proves to be more effective than a GPT-4 prompting baseline, highlighting its ability to provide more specialized and precise reflection to guide the Reasoner model.

Classifier Baseline PLM Classifier serves as a strong baseline as it is directly fine-tuned on student answer scoring data. While it exhibits strong performance across all metrics, the *classification approach lacks explainability*, as it only generates scores without providing rationales for its decisions.

Single Reasoner Baselines The single reasoner baselines, including SFT and DPO, aim to improve explainability by generating rationales for scoring decisions. However, these methods generally underperform compared to classification-based approaches, particularly on the private datasets, where *data scarcity presents a major challenge*. The preference optimization method consistently shows modest improvements over the SFT base model in terms of QWK scores. However, *these improvements come at the cost of declines in F1 (-4%) and ACC scores (-1%)*, suggesting a tendency to overfit to preference annotations (Chowdhury et al., 2024; Mitchell, 2023). Moreover, the *implicit preference optimization process lacks transparency*, making the Reasoner-only DPO approach less reliable.

Dual w/ GPT-4 Baseline We also evaluate a dual-model variant where GPT-4-turbo serves as the Critic to generate reflection-based instructions for refinement. However, after multiple refinements (`DARS` Reasoner only vs. Reflect w/ GPT-4-turbo), performance significantly declined across all datasets and evaluation metrics. This indicates that despite GPT-4-turbo's strong general capabilities, *it struggles to produce specialized and precise reflections for refining the Reasoner's output*[5].

Our Framework `DARS` *demonstrates significant improvements from the initial to the final iteration across all datasets, highlighting the efficacy of dual model reasoning, and test-time rationale refinement*. The `DARS` Reasoner only performance is measured on the Reasoner's *first-pass predictions* (e.g., Reasoner predicts ② based on ①), while the Reflect w/ Critic results are generated from `DARS`,

[4] QWK is often considered as the main metric for ASAS.

[5] Detailed case studies are provided in Appendix B.2.

Methods	Classification Baseline			Generative Baselines (*Reasoner Only*)									Dual-Model Reasoning Framework					
	PLM Classifier			SFT			DPO			(DARS) Reasoner only			Reflect w/ GPT-4			**(DARS) Reflect w/ Critic**		
Datasets	ACC	F1	**QWK**	ACC	F1	**QWK**	ACC	F1	**QWK**	ACC	F1	**QWK**	ACC	F1	**QWK**	ACC†,*	F1†,*	**QWK***
ASAP 1	0.7767	0.7805	0.8528	0.6968	0.7073	**0.8277**	0.6895	0.5655	0.8051	0.6480	0.6606	0.8073	0.5181	0.5106	0.6349	**0.7274**	**0.7315**	0.8100
ASAP 2	0.6798	0.6817	0.8187	**0.7324**	**0.7468**	**0.8420**	0.6761	0.6783	0.8033	0.6925	0.7074	0.8136	0.5869	0.5636	0.6532	0.7136	0.7303	0.8277
ASAP 5	0.8625	0.6055	0.8187	0.8495	0.5600	0.8203	0.8612	0.6449	0.8001	0.8545	0.5424	0.7766	0.8177	0.5119	0.6340	**0.8645**	**0.6303**	**0.8326**
ASAP 6	0.8891	0.6118	0.8426	0.8314	0.5513	0.7273	0.8314	0.5420	0.7522	0.8280	0.5628	0.7232	0.8130	0.4265	0.4754	**0.8648**	**0.5988**	**0.8016**
Private 1	0.6787	0.6784	0.8853	0.5236	0.5197	0.8082	0.5236	0.4670	0.8196	0.5551	0.5584	0.8221	0.4134	0.3407	0.6018	**0.5709**	**0.5653**	**0.8253**
Private 2	0.6224	0.6355	0.8385	0.5459	0.5377	0.7004	0.5561	0.5600	0.7599	0.5765	0.5752	0.7604	0.5357	0.5219	0.7688	**0.6071**	**0.6059**	**0.7705**
Overall	0.7515	0.6656	0.8428	0.6966	0.6038	0.7877	0.6897	0.5763	0.7900	0.6925	0.6011	0.7839	0.6141	0.4792	0.6280	**0.7247**	**0.6437**	**0.8113**

Table 1: **Comparison of assessment performance across baseline and Reasoner only preference optimization methods.** Generative methods are indicated with a gray background . All methods were reproduced or trained using the same LLaMA 3B model as the base. We highlighted the highest values for ACC (↑), F1 Score (↑), and QWK (↑) among generative methods in **bold**. The overall performance is calculated as the average across all datasets. Symbols † and ∗ indicate statistical significance compared to SFT and DPO by each metric, respectively.

i.e., the final refined Reasoner output before the loop is terminated by the Critic model (e.g.,4). Compared to the preference optimization baseline (SFT to DPO), our framework (Reasoner only to Reflect w/ Critic) not only ***outperforms on average ACC, F1, and QWK scores*** but also ***maintains a balanced enhancement across all metrics even under data scarcity*** (improved 5% for ACC, 11% for F1, and 2% for QWK). Compared with GPT-4-turbo as the Critic, our Critic model more effectively reflects on wrongly assessed rationales and guides the Reasoner outputs to be closer to the oracle labels (18%-34% better in metrics). Specifically, Reflect w/ Critic surpasses the Reasoner only assessment result across all datasets and metrics (3%-7% improvement). Statistically, Reflect w/ Critic significantly outperforms the state-of-the-art baselines (SFT and DPO)[6].

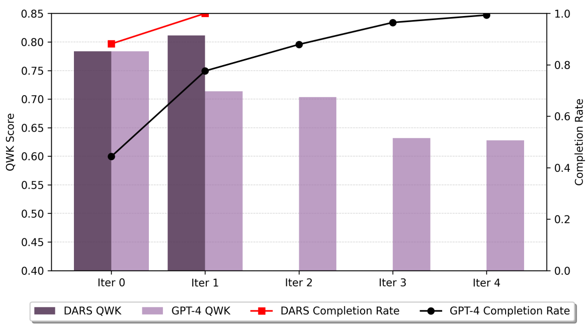

Figure 4: **Performance and completion rate in iterations.**

To show the effectiveness of our Critic model in reflection and determine when to stop, as illustrated in Figure 4, we visualize the performance trend and completion rate comparison between DARS's iterative reasoning process and GPT-4-turbo as the Critic model. Our method requires only two iterations to achieve a significant improvement over iteration 0, which represents the Reasoner's initial predictions. In contrast, GPT-4-turbo takes nearly four iterations to reach termination, and shows a clear trend of performance degradation across all metrics as the iterations progress.

[6]A one-tailed t-test yielded a p-value of ≤ 0.05, indicating statistical significance.

Quality Evaluation for Reflection To further analyze the transparency and correctness of the generated reflections, we conducted a human evaluation of the Critic-Reasoner interactions. We assessed the quality of the Critic's reflections and the subsequent Reasoner's refinements. The evaluation results are visualized in Figure 5.

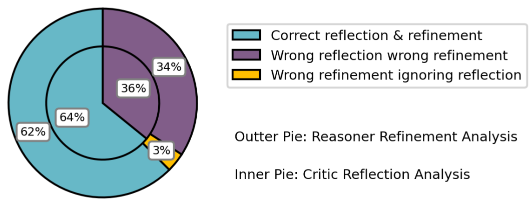

Figure 5: **Qualitative analysis on reflection and refinement.**

Our findings indicate that the Critic model accurately identified assessment errors in 64% of cases, effectively localizing errors in scoring rationales. This aligns with previous observations (Tyen et al., 2024), which suggest that LLMs can correct errors when provided with proper error localization. However, in 36% of cases, the Critic's reflections were inaccurate, often due to misinterpretation of the student's answer and the scope of the key answer elements. Such inaccuracies had cascading effects: in 34% of cases, the Critic's incorrect guidance misled the Reasoner, leading to further wrong assessments. We also observed that in 3% of instances, the Reasoner ignored the Critic's feedback (despite correct or incorrect) and still produced erroneous outcomes.These results indicate that our Reasoner can follow the Critic's guidance 97% of the time for refinement. Overall, *these results highlight the critical role of a strong Critic for generating explainable, verbal reflection instructions, so that the Reasoner could effectively refine its predic-*

tions. Further error analysis (§B.3) and case studies (§B.6) are provided in the Appendix.

5.3 Scaling Law for Dual-Model Framework

Given that our Reasoner and Critic models are trained independently, we study the effect of model size on the performance of DARS using four Qwen model variants (3B, 7B, 14B, and 32B) (Qwen-Team, 2024). We trained each model using identical datasets, training procedures, and hyper-parameters, resulting in a total of 16 distinct Reasoner and Critic combinations.

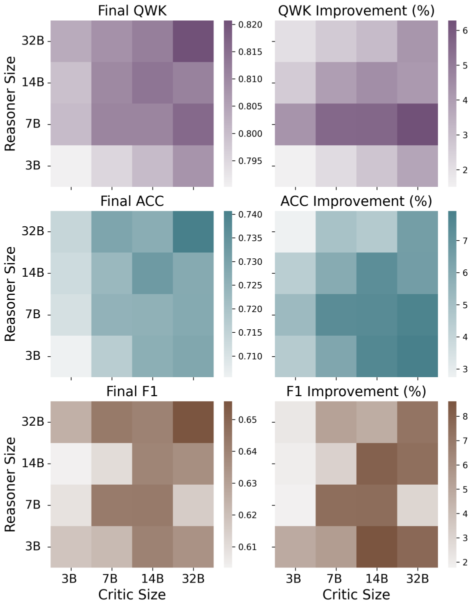

Figure 6: **Scaling experiment for reasoner and critic.**

We present the overall final performance and performance improvements[7] in Figure 6. Unlike observations in prior studies (Welleck et al., 2023; Akyurek et al., 2023; Paul et al., 2024), ***our findings suggest that increasing the Critic's size*** (horizontal direction, left to right) ***leads to greater performance gains (ACC and QWK), more so than increasing the Reasoner's size*** (vertical direction, bottom to top). This suggests that a larger Critic provides more precise evaluation and reflection, which the Reasoner relies upon for refinement[8]. Although larger Critic models generally improve F1 scores, this trend is not as pronounced, which is due to imbalances in dataset sizes and label distributions[9].

5.4 Ablation Studies on DARS

Can the Reasoner Refine Effectively Without Strong Task Capability? To investigate whether the Reasoner can perform refinement without a strong task capability, we trained two "weak" Reasoners with Qwen 3B and LLaMA 3B with weaker rationale training data[10], following Li et al. (2023). As shown in Figure 7, all the DARS frameworks with a "weak" Reasoner dropped more than 10% in overall performance across all metrics, even with access to high-quality refinement data and a strong Critic model. This result shows that ***without a strong task capability, the Reasoner cannot perform refinement effectively***.

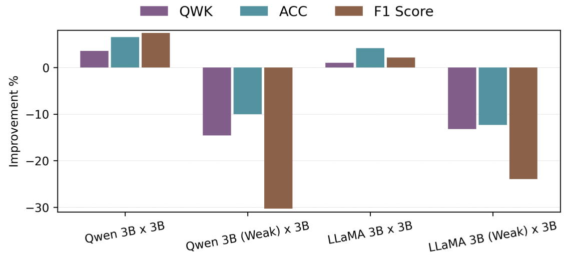

Figure 7: **DARS refine with "weak" Reasoner model.**

Does Refinement Ability Benefit Reasoner's Task Capability? To further investigate the impact of refinement data on task performance, we trained two models: LLaMA 3B w/o Refinement and LLaMA 8B w/o Refinement by excluding the multi-turn reflection refinement data from the Reasoner's training sets. We report the Reasoner-only's performance in Figure 8. We observe that evaluation result for Reasoner's w/o refinement models dropped nearly 5% in all metrics compared with including refinement data, ***indicating the error correction data (e.g. training the model to refine from errors) can boost the Reasoner's task capability***. This observation align closely with previous findings (Tong et al., 2024; Kamoi et al., 2024b). We also show that reflection data can effectively regulate preference optimization training in §B.5.

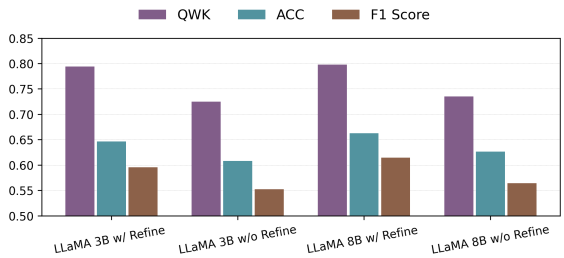

Figure 8: **Ablation on the refinement data for Reasoner.**

[7] Performance improvement is expressed as a percentage increment compared to the Reasoner only's performance.

[8] See §B.7 for case studies.

[9] Significant label imbalances in some datasets may cause the Reasoner to modify initially "correct" minority label categories, thereby affecting the overall F1 trend.

[10] We characterized the data as weaker data for two reasons: (1) the rationales were sourced from ChatGPT, whereas the current training data was curated using GPT-4; (2) a previous study (Li et al., 2024a) shows models trained on this dataset exhibit significantly low and imbalance performance.

Can a Single Model Perform Both Reasoning and Reflection? We explore whether merging the training data of both the Reasoner and Critic to train a single model could enables effective self-reflection. We trained two self-reflection models Qwen 3B (Self) and LLaMA 3B (Self). Figure 9 shows a significant decline in the iterative refinement process, with a negative performance improvement rate. This unified model struggles to accurately determine when to terminate the refinement process and failed to provide useful reflection instructions. These findings align with prior observations (Huang et al., 2024), suggesting that ***"two heads are better than one"–a single model cannot effectively balance both reasoning and critique***.

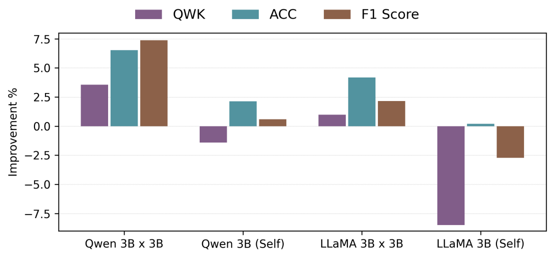

Figure 9: **Combine dual-model into a single one.**

5.5 Generalization Studies

Can Critic Effectively Reflect on Unseen Questions? In Figure 10, we evaluate the ability of the Critic model to generalize to unseen questions. To do this, we trained two versions of Critic: one with exposure to our private datasets (Critic Seen) and one without (Critic Unseen). We use LLaMA 3B as the base model. Our results reveal that the Critic Unseen model, ***despite its lack of exposure to all datasets, still enhances the Reasoner's original assessments*** (+1% in QWK), albeit with slightly reduced effectiveness compared to the Critic Seen model (-3% in QWK). These findings show that the Critic can still provide meaningful feedback even when it has not been explicitly trained on new data.

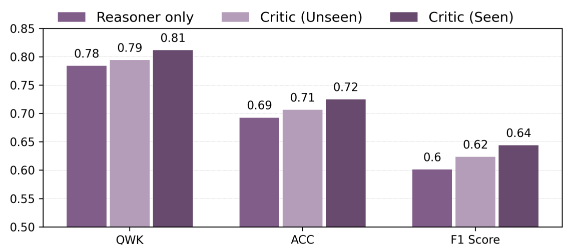

Figure 10: **DARS Critic reflects on unseen questions.**

Adaptability Beyond Model Sizes and Architectures Figure 11(a) illustrates our exploration of the performance across various base models, including LLaMA 3B, 8B and Qwen 3B, 7B. The results show minimal variance in performance no matter across different model sizes and architectures, demonstrating that our ***training method is highly adaptable and not restricted to a specific model architecture or size***. Furthermore, Figure 11(b) explores the feasibility of using different base models for the Reasoner and Critic at inference time, such as pairing a Qwen Reasoner with a LLaMA Critic. Our findings indicate ***consistent performance irrespective of model combinations***. This highlights the robustness of our framework, due to its *use of text* for effective interactions between the Critic and Reasoner components.

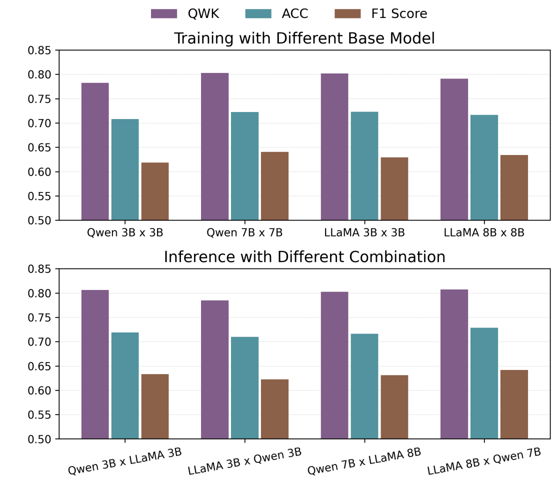

Figure 11: **Generalization analysis on size, architecture and inference combinations.**

6 Conclusion and Discussion

We proposed a novel approach to enhance inference-time reasoning in LLMs through a dual-model framework. Our approach introduces a contrastive reflection synthesis pipeline, which generates verbal reflections that significantly improved reasoning explainability. Our framework, consisting of a dedicated Reasoner and Critic, enables effective reasoning refinement without relying on oracle labels. Moreover, our carefully designed training process equips both models with capabilities that extend beyond task-specific reasoning. The Reasoner not only solves problems but also learns to refine its reasoning based on feedback, while the Critic not only identifies errors but also learns when to stop, ensuring more efficient reasoning improvement. This capacity aligns with reasoning LLM advances seen in models like DeepSeek-R1 and OpenAI's O1, where inference-time reflection enables iterative, self-correcting reasoning. Although our experiments focus on ASAS, the adaptability of the thought tree and the reflection synthesis process make it possible to extend our framework to other complex reasoning tasks in future work.

Limitations

This study has two primary limitations. First, the training process requires substantial computational resources. While our framework minimizes the need for future retraining, the SFT training for both the Reasoner and Critic involves additional data points to enhance the model's various capabilities, leading to higher training FLOPs than single Reasoner approaches. Second, the generalizability of our framework to tasks beyond ASAS remains unexplored. Although we conducted a comprehensive evaluation across six datasets, our focus was predominantly on the ASAS task. Future work should investigate the applicability of the proposed framework to a broader range of tasks.

Ethics Statement

This study utilized both public and private datasets of anonymized student responses, none of which contain sensitive or personally identifiable information. We thoroughly reviewed the LLM's outputs and did not identify any instances of harmful content or exposure to personal information. Nevertheless, before deploying our framework in high-stakes examination settings, experts must carefully evaluate its assessment decisions and the underlying rationales to ensure reliability and fairness.

References

AI@Meta. 2024. Llama 3 model card.

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*.

Canyu Chen and Kai Shu. 2024. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Step-level value preference optimization for mathematical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024b. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. 2024. Provably robust dpo: aligning language models with noisy feedback. In *Proceedings of the 41st International Conference on Machine Learning*.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

Yihong Dong, Kangcheng Luo, Xue Jiang, Zhi Jin, and Ge Li. 2024. PACE: Improving prompt with actor-critic editing for large language model. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Ben Hamner, Jaison Morgan, Mark Shermis Lynnvandev, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Yuxin Jiang, Bo Huang, Yufei Wang, Xingshan Zeng, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, and Wei Wang. 2025. Bridging and modeling correlations in pairwise data for direct preference optimization. In *The Thirteenth International Conference on Learning Representations*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Ryo Kamoi, Sarkar Snigdha Sarathi Das, Renze Lou, Jihyun Janice Ahn, Yilun Zhao, Xiaoxin Lu, Nan Zhang, Yusen Zhang, Haoran Ranran Zhang, Sujeeth Reddy Vummanthala, Salika Dave, Shaobo Qin, Arman Cohan, Wenpeng Yin, and Rui Zhang. 2024a. Evaluating LLMs at detecting errors in LLM responses. In *First Conference on Language Modeling*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024b. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*.

Geunwoo Kim, Pierre Baldi, and Stephen Marcus McAleer. 2023. Language models can solve computer tasks. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal M. P. Behbahani, and Aleksandra Faust. 2025. Training language models to self-correct via reinforcement learning. In *The Thirteenth International Conference on Learning Representations*.

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*.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *ArXiv*, abs/2406.18629.

Leah S. Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Jiazheng Li, Hainiu Xu, Zhaoyue Sun, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2024a. Calibrating LLMs with preference optimization on thought trees for generating rationale in science question scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Yanhong Li, Chenghao Yang, and Allyson Ettinger. 2024b. When hindsight is not 20/20: Testing limits on reflective thinking in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024a. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. 2024b. Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. 2024a. Eliminating biased length reliance of direct preference optimization via down-sampled KL divergence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, and Mingjie Zhan. 2024b. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *ArXiv*, abs/2407.00782.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Eric Mitchell. 2023. A note on dpo with noisy preferences & relationship to ipo.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason E Weston. 2024. Iterative reasoning preference optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. REFINER: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, and Shanghaoran Quan. 2024. Qwen2.5 technical report.

QwenTeam. 2024. Qwen2.5: A party of foundation models.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. Scaling laws for reward model overoptimization in direct alignment algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling test-time compute optimally can be more effective than scaling LLM parameters. In *The Thirteenth International Conference on Learning Representations*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can LLMs learn from previous mistakes? investigating LLMs' errors to boost for reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Roopchand Reddy Vanga, C. Sindhu, M. S. Bharath, T. Charandeep Reddy, and Meghana Kanneganti. 2023. Autograder: A feature-based quantitative essay grading system using bert. In *ICT Infrastructure and Computing*.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Yueqin Yin, Zhendong Wang, Yi Gu, Hai Huang, Weizhu Chen, and Mingyuan Zhou. 2024. Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts. *ArXiv*, abs/2402.10958.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*.

A Further Experiment Setup

This section provides additional details on the setup of the experiment:

Dataset Statistic We provide the detailed dataset statistics in Table A1.

Datasets (Subjects)	Train	Validation	Test	Score Range
ASAP 1 (Science)	1,337	331	554	0-3
ASAP 2 (Science)	1,018	252	426	0-3
ASAP 5 (Biology)	1,436	359	598	0-3
ASAP 6 (Biology)	1,437	359	599	0-3
Private 1 (Biology)	440	89	254	0-4
Private 2 (Biology)	358	72	196	0-3

Table A1: **Dataset statistics.**

Classification Baseline The input to the text classifier consists of concatenated question-related information (including the question prompt, key answer elements, and marking rubric) along with the student answer, separated by newlines. The classifier is trained to predict scores. Following previous studies, we trained a separate model for each dataset and evaluated it using the original test splits (Mayfield and Black, 2020). We employed DeBERTa-v3-large as the base pre-trained language model (He et al., 2023). The reported results are averaged over five runs with different random seeds (210, 102, 231, 314, 146). The hyper-parameter settings are provided in Table A2.

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	16
Epochs	15
Warmup Steps	100
Weight Decay	0.1
Optimizer	Adam
Adam Epsilon	1e-8

Table A2: **Classification hyper-parameters setting.**

Generative Baselines For generative baselines, the input to the model comprises the question context and student answers, with the model generating assessment rationales in textual form. The results are averaged over three runs with different random seeds. Unlike prior work (Li et al., 2024a), we conducted full parameter training using bfloat16 precision. All generative models were trained using the LLaMA-factory framework (Zheng et al., 2024). The hyper-parameter settings are provided in Table A3.

API Use for Synthetic Data Generation We utilized `gpt-4-turbo` (OpenAI et al., 2024) as the

Hyperparameter	SFT	DPO
Learning Rate	1e-5	1e-5
Batch Size	4	4
Gradient Accumulation	4	4
Epochs	4.0	3.0
Warmup Ratio	0.1	0.1
LR Scheduler Type	cosine	cosine
Optimizer	Adam	Adam
Adam Epsilon	1e-8	1e-8
DPO ftx	-	0.5
DPO β	-	0.1

Table A3: **Generative hyper-parameters setting.**

Template Prompt for Generate Reflection

Here is an incorrect assessment rationale for the student answer:
[Student Answer]:{student_answer}
Incorrect Rationale: {reject_rationale}
This wrong rationale missed the following key elements:
- {idx}: The student didn't answer the "key_element[idx]" but the incorrect rationale wrongly assessed the student mentioned it.
- {idx}: The student answered the "key_element[idx]" but the incorrect rationale wrongly assessed the student didn't mention it.
Please construct a **reflection guidance** that
1. point out the incorrectly assessed key elements,
2. guide the model to reflect on the mistakes for generating a better assessment rationale,
3. pretend you are talking with an assessor using pronouns like "you",
4. By the end of the guidance ask the model to reflect or revise based on the feedback and retry or regenerate the rationale.
Output the guidance in JSON format:{ "guidance": "..." }

Figure A1: **The Prompt Template for Contrastive Reflection Synthesis.**

LLM to generate synthetic reflection data, as described in §4.1. All inference parameters were kept at their default values. The prompt template is presented in Figure A1.

DARS Framework We trained both the Reasoner and Critic models using full parameters training with bfloat16 precision. All models were evaluated using greedy decoding. Except for the scaling experiment, all results were averaged over three different runs. The hyper-parameter settings are provided in Table A4. We train the Reasoner and Critic models using synthetic data we generated, as introduced in our methodology part. All those models are solely trained on the original train split, as shown in Table A1. The validation split was only used to select the best checkpoint, and the Test split was never seen by the model until the evaluation.

API Use for GPT-4-turbo Critic Baseline We utilized `gpt-4-turbo-2024-04-09` (OpenAI et al., 2024) as the Critic LLM to generate reflection data. The temperature is set as 0.7 and the maximum token generation is limited to 1,024. The prompt template is presented in Figure A2.

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	
- Model Size $\leq$ 8B	16
- Model Size $>$ 8B	8
Gradient Accumulation	
- Model Size $\leq$ 8B	1
- Model Size $>$ 8B	2
Epochs	1.0
Warmup Ratio	0.05
Weight Decay	0.02
LR Scheduler Type	cosine
Optimizer	Adam
Adam Epsilon	1e-8

Table A4: **DARS framework hyper-parameters settings.**

Prompt Template for GPT-4-turbo

Given the provided assessment of the student's answer, generate constructive and actionable feedback to help the assessment model improve their response. The feedback should:
1. Highlight Areas for Improvement: Point out specific aspects where the model can enhance their assessment, such as accuracy, completeness, clarity, or structure.
2. Provide Actionable Suggestions: Offer clear, practical steps the model can take to address identified weaknesses and improve their understanding.
Please generate feedback based on these guidelines to guide the model in refining their response effectively.
If the assessment seems good enough, please output "[STOP]" to indicate the end of the feedback.

Figure A2: **Prompt template for GPT-4-turbo as critic.**

Base Models, Computational Environment, and Inference Setup In this study, we utilized six different models downloaded from HuggingFace Transformers [11]. We adhered to the licensing terms of all involved models. meta-llama/Llama-3.2-3B-Instruct (LLaMA 3B), meta-llama/Llama-3.1-8B-Instruct (LLaMA 8B) from (AI@Meta, 2024), and Qwen/Qwen2.5-3B-Instruct (Qwen 3B), Qwen/Qwen2.5-7B-Instruct (Qwen 7B), Qwen/Qwen2.5-14B-Instruct (Qwen 14B), Qwen/Qwen2.5-32B-Instruct (Qwen 32B) from (QwenTeam, 2024; Qwen et al., 2024).

All generative models were trained using either 4 $\times$ A100 80G or 4 $\times$ H100 GPUs.

To ensure reproducibility, all evaluations are done using zero-shot prompting with greedy decoding and a temperature of 0. Inference of LLMs is carried out using vLLM (Kwon et al., 2023). We utilized the same prompt templates and score extractor as released by (Li et al., 2024a). Prompt templates for ASAP 1 (Figure A8), ASAP 2 (Figure A9), ASAP 5 (Figure A3), and ASAP 6 (Figure A4) can also be found in each case studies.

[11] https://huggingface.co/

Manual Evaluation Setup We randomly sampled 20 instances from each dataset and manually examined the reflection and refinement generated. The outputs were derived from a single run using the LLaMA 3B Reasoner and LLaMA 3B Critic model, as reported in Table 1. The annotations were conducted by the authors of this paper. We categorized the errors using the following schema.

Evaluation on Critic's Reflection Errors in the Critic model's reflections were classified as follows:

- **Correct Reflection**: The Critic model accurately identified errors in the previous assessment, ensuring faithfulness to both the student's answer and the question content.
- **Incorrect Reflection**: The Critic model either misinterpreted the meaning of the student's answer or the scope of key answer elements, leading to incorrect identification of errors or the identification of errors that were not coherent to the given content.

Evaluation on Reasoner's Refinement We classify the error made by the Reasoner model in refinement into the following three categories:

- **Correct Refinement**: The situation the Reasoner model successfully refined its previous mistakes based on the Critic's reflection.
- **Wrong Refinement Obeyed Reflection**: The situation Reasoner model made an error because it faithfully followed the Critic's wrong reflection.
- **Wrong Refinement Ignored Reflection**: The situation in which the Reasoner model introduced a new error, deviating from the Critic's reflection.

B Further Experiment Result

B.1 Explanation for Main Example

As illustrated in Figure A3, we present the complete example corresponding to Figure 3.

Initially, the Reasoner takes the question prompt as input and generates its first assessment decision ②. However, in this first attempt, the model incorrectly evaluates the student's response by crediting key elements such as "...described mRNA exiting the nucleus..." and "...the corresponding amino acids on tRNA being bonded, and the continuation of amino acid linkage until a stop codon is reached,..." which were not explicitly mentioned.

The Critic model then takes both the question prompt ① and the Reasoner's initial assessment ② as input to generate a reflection instruction ③.

The Critic accurately identifies the Reasoner's misjudgment, stating: "You credited the student for mentioning that the 'corresponding amino acids on tRNA are bonded to adjacent tRNA's amino acids' and that 'amino acids continue to be linked until a STOP codon is read on the mRNA.' However, upon reviewing the student's response, these elements were not explicitly covered." The Critic further instructs the Reasoner to "Please revisit the student's answer and your rationale, considering these points, and try to generate a more precise assessment that reflects the actual content of the student's response."

Subsequently, the Reasoner incorporates the chat history and the Critic's feedback (1, 2, 3) as input to generate a revised assessment decision. The newly generated Reasoner output 4 accurately identifies the key elements in the student's response and corrects the final score assessment.

Finally, the Critic evaluates the updated assessment and generates a termination token, "[STOP]," indicating the end of the reasoning loop. This process demonstrates the iterative refinement capability of the proposed dual-model framework, ensuring accurate and explainable assessment evaluations.

B.2 Case Studies on GPT-4-turbo as Critic

The case study in Figure A4 highlights the limitations of using GPT-4-turbo as a Critic model. GPT-4-turbo generated feedback tends to be vague, overemphasizing surface-level details while lacking contextual relevance and actionable insights. It struggles to provide precise guidance for improving assessments, often failing to align with key rubric elements and offering inconsistent or generalized reflection instructions. Specifically, the original Reasoner's assessment is correct, but the GPT-4-turbo fails to evaluate the assessment and didn't terminate the iterative refinement process. These shortcomings hinder its effectiveness in refining assessment rationales, underscoring the need for a more tailored Critic model that delivers targeted, domain-specific feedback for accurate and meaningful evaluation.

B.3 Detailed Error Analysis

As shown in Figure A5, we provide an in-depth analysis of the Critic model's effectiveness using a single run with the LLaMA 3B Reasoner and LLaMA 3B Critic model.

Label Distribution The first row of the Figure A5 presents an analysis of the overall label distribution changes across iterations. As shown in (a), the label distribution shifts closer to the ground-truth distribution after the second iteration with the Critic model's guidance. This trend is further supported by the confusion matrices in (b) and (c), where the second iteration exhibits a more pronounced diagonal pattern, indicating improved alignment with ground-truth labels. In contrast, the first iteration shows a bias towards scores of 0 and 1.

Score Transitions To gain deeper insights into label transitions, the second row of the Figure A5 examines label changes across iterations. As shown in (d), while our framework does not guarantee perfect label corrections, the majority of transitions move from incorrect to correct labels. This underscores the potential to further refine the collaboration between the Critic and Reasoner models to minimize cases where correct predictions are mistakenly altered. Additionally, (e) and (f) display the top 10 transitions from correct to incorrect and incorrect to correct labels, respectively. The results reveal that most label changes occur between scores of 1 and 3, with the majority involving a single-point difference, reflecting patterns observed in human assessment behaviour.

B.4 Two Smaller Models May Better Than a Larger One

As illustrated in Figure A6, `DARS`, which employs a dual-model setup with LLaMA 3B Reasoner and Critic, outperforms a single LLaMA 8B DPO model. This finding further reinforces that "two heads are better than one", demonstrating that two smaller 3B models working together can achieve better results than a single, larger 8B Reasoner. This superior performance may be due to the fact that LLaMA 3B is a distilled variant of the 8B version (AI@Meta, 2024).

B.5 Can Refinement Data Enhance Preference Optimization for the Reasoner?

Inspired by (Liu et al., 2024b), we propose a robust preference optimization baseline by incorporating an additional SFT loss on the synthetic reflection data to regularize the DPO training process. As illustrated in Figure A7, the inclusion of regularization on reflection data leads to slight improvements in QWK and F1 scores compared with vanilla DPO. These results suggest that ***refinement data can also serve as an effective regularizer even for single-***

Case Study 1

①***[Question Prompt]*:**
[Question]:
Starting with mRNA leaving the nucleus, list and describe four major steps involved in protein synthesis.

[Key Elements]:
mRNA exits nucleus via nuclear pore.
mRNA travels through the cytoplasm to the ribosome or enters the rough endoplasmic reticulum.
mRNA bases are read in triplets called codons (by rRNA).
tRNA carrying the complementary (U=A, C+G) anticodon recognizes the complementary codon of the mRNA.
The corresponding amino acids on the other end of the tRNA are bonded to adjacent tRNA's amino acids.
A new corresponding amino acid is added to the tRNA.
Amino acids are linked together to make a protein beginning with a START codon in the P site (initiation).
Amino acids continue to be linked until a STOP codon is read on the mRNA in the A site (elongation and termination).

[Marking Rubric]:
3 points: Four key elements;
2 points: Three key elements;
1 point: One or two key elements;
0 points: Other.
[Student Answer]: the mRNA, Is met by tRNA with anticodons, that link to the mRNA's codons. This tRNA carries an amino acid that goes with the mRNA's codon. This is done multiple times until there is a stranhd of amino acids. Ribosomes attach these amino acids together.
Critically assess the [Student Answer] in light of the [Key Answer Elements]. Offer a detailed explanation for the score assigned, referencing specific criteria from the [Marking Rubric]:
②***[Reasoner]*:**
The student's answer scored 3 points according to the marking rubric, which awards 3 points for addressing four key elements. The student successfully described mRNA exiting the nucleus, tRNA carrying complementary anticodons recognizing the mRNA codons, the corresponding amino acids on tRNA being bonded, and the continuation of amino acid linkage until a stop codon is reached, as indicated by '...tRNA with anticodons, that link to the mRNA's codons.' and '...until there is a strand of amino acids. Ribosomes attach these amino acids together.' However, the answer lacked explicit mention of mRNA traveling to the ribosome, reading of mRNA bases in triplets by rRNA, addition of new amino acids to tRNA, and the initiation process at the START codon.
③***[Critic]*:**
It appears there was a misunderstanding in your assessment of the student's answer. You credited the student for mentioning that the 'corresponding amino acids on tRNA are bonded to adjacent tRNA's amino acids' and that 'amino acids continue to be linked until a STOP codon is read on the mRNA.' However, upon reviewing the student's response, these elements were not explicitly covered. The student's description primarily focused on the interaction between mRNA and tRNA and the eventual formation of a peptide bond, but did not detail the process of amino acids being bonded to adjacent tRNA's amino acids nor the specific mechanism of elongation and termination involving a STOP codon. It's crucial to align your assessment closely with the content explicitly presented by the student to ensure accurate scoring. Reflecting on this, you might want to re-evaluate how you interpret and assess answers that do not explicitly cover all the elements listed in a rubric. Please revisit the student's answer and your rationale, considering these points, and try to generate a more precise assessment that reflects the actual content of the student's response.
④***[Reasoner]*:**
The student's answer scored 2 points according to the marking rubric which stipulates 2 points for achieving three key answer elements. The student accurately described the roles of mRNA and tRNA in protein synthesis, mentioning that 'the mRNA, Is met by tRNA with anticodons, that link to the mRNA's codons' which covers the key elements of mRNA bases being read in triplets and tRNA carrying the complementary anticodon recognizing the complementary codon of the mRNA. Additionally, the student implied the bonding of amino acids by stating 'Ribosomes attach these amino acids together.' However, the answer did not explicitly mention the exit of mRNA from the nucleus, the travel of mRNA through the cytoplasm, the addition of new amino acids to tRNA, the initiation process with a START codon, or the elongation and termination at a STOP codon which resulted in losing points for these key elements.
⑤***[Critic]*:**
Rationale Looks Good! [STOP]

Figure A3: **The full example as presented in Figure 3.**

Case Study 2

①***[Question Prompt]*:**
[Question]:
List and describe three processes used by cells to control the movement of substances across the cell membrane.
[Key elements]:
Selective permeability is used by the cell membrane to allow certain substances to move across.
Passive transport occurs when substances move from an area of higher concentration to an area of lower concentration.
Osmosis is the diffusion of water across the cell membrane.
Facilitated diffusion occurs when the membrane controls the pathway for a particle to enter or leave a cell.
Active transport occurs when a cell uses energy to move a substance across the cell membrane, andor a substance moves from an area of low to high concentration, or against the concentration gradient.
Pumps are used to move charged particles like sodium and potassium ions through membranes using energy and carrier proteins.
Membrane-assisted transport occurs when the membrane of the vesicle fuses with the cell membrane forcing large molecules out of the cell as in exocytosis.
Membrane-assisted transport occurs when molecules are engulfed by the cell membrane as in endocytosis.
Membrane-assisted transport occurs when vesicles are formed around large molecules as in phagocytosis.
Membrane-assisted transport occurs when vesicles are formed around liquid droplets as in pinocytosis.
Protein channels or channel proteins allow for the movement of specific molecules or substances into or out of the cell.
[Marking Rubric]:
3 points: Three key elements;
2 points: Two key elements;
1 point: One key element;
0 points: Other.
[Student Answer]: Three processes used by cells to control the movement of substances across the cell membrane are. * The golgi bodies * lysosomes * protozoas
Evaluate the [Student Answer] based on the provided [Key Answer Elements] and [Marking Rubric]. Summarize the assessment and justify the score awarded:
②***[Reasoner]*:**
The student's answer scored 0 points according to the marking rubric that awards 3 points for three key elements, 2 points for two key elements, 1 point for one key element, and 0 points for other. The student's response listed 'The golgi bodies, lysosomes, protozoas' which do not pertain to any of the key elements required for the question about processes used by cells to control the movement of substances across the cell membrane. None of the listed items describe selective permeability, passive transport, osmosis, facilitated diffusion, active transport, the use of pumps, or any form of membrane-assisted transport such as exocytosis, endocytosis, phagocytosis, pinocytosis, or the function of protein channels or channel proteins.
③***[Critic]*:**
The student's answer does not align with the key elements required to address the question effectively. Here is constructive feedback to help improve their response:
1. **Highlight Areas for Improvement**:
- The response does not mention any relevant cellular processes used to control the movement of substances across the cell membrane, such as selective permeability, passive transport, active transport, or membrane-assisted transport.
- The listed terms ('Golgi bodies,' 'lysosomes,' 'protozoas') do not pertain to the question and indicate a lack of understanding of the topic.
2. **Provide Actionable Suggestions**:
- Review the concept of the cell membrane and its role in controlling the movement of substances. Focus on understanding key processes such as passive transport (including osmosis and facilitated diffusion), active transport, and membrane-assisted transport mechanisms like endocytosis and exocytosis.
- Refer to class materials, textbooks, or trusted educational resources to study how these processes work and why they are important for cellular function.
- Practice answering similar questions by first identifying the processes relevant to the question and then providing clear descriptions of how each process works. For example, explain that passive transport involves movement along a concentration gradient without energy use, while active transport requires energy to move substances against their concentration gradient.
- Avoid listing unrelated terms and ensure that each term directly addresses the question.
By following these steps, you can develop a more accurate and complete response to similar questions in the future. Keep practicing to strengthen your understanding of cell membrane transport mechanisms!
④***[Reasoner]*:**
To improve the answer, the student should focus on describing specific processes such as passive transport, active transport, and membrane-assisted transport. For example, they could explain how osmosis involves the movement of water across a semi-permeable membrane from an area of lower solute concentration to an area of higher solute concentration, or how active transport requires energy to move substances against their concentration gradient. Additionally, describing the roles of protein channels and carrier proteins in facilitated diffusion and the process of endocytosis and exocytosis would also be relevant.
...

Figure A4: **Prompting GPT-4-turbo failed to act as effective critic model.**

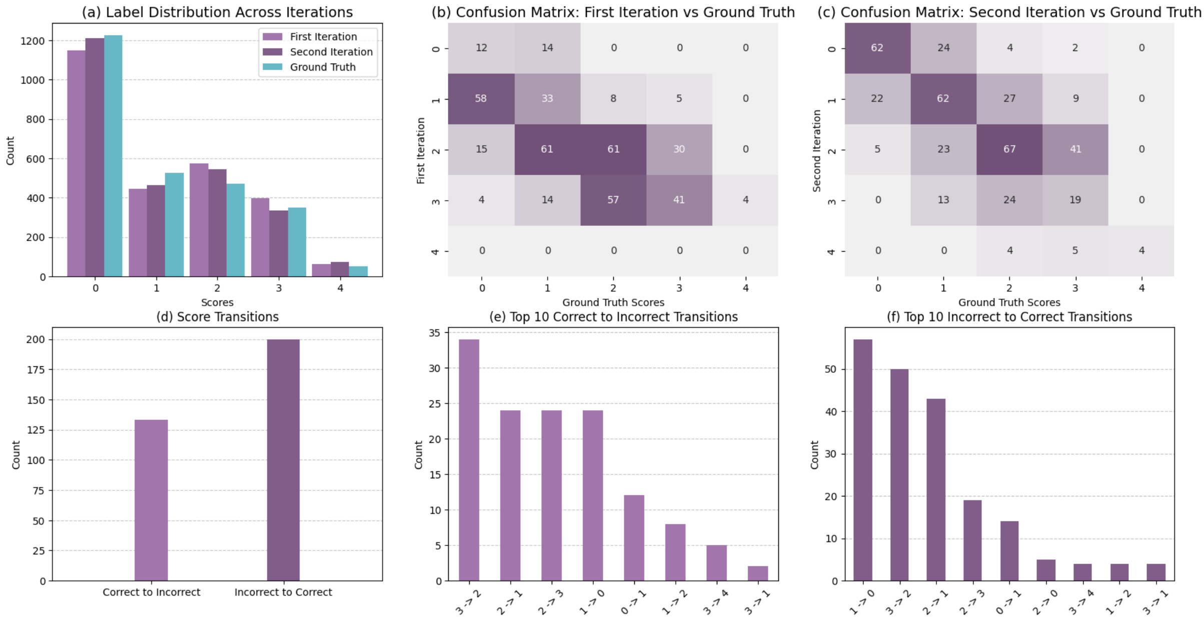

Figure A5: **Visualization of detailed error analysis for the iterative reasoning process.**

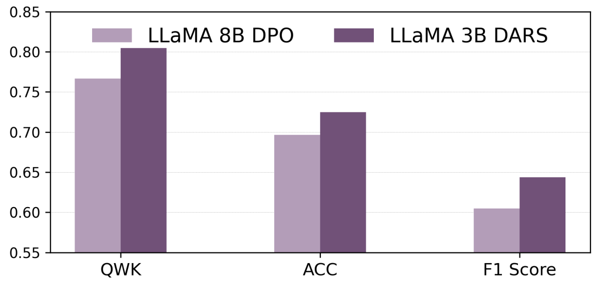

Figure A6: **Comparison of DARS with LLaMA 8B DPO.**

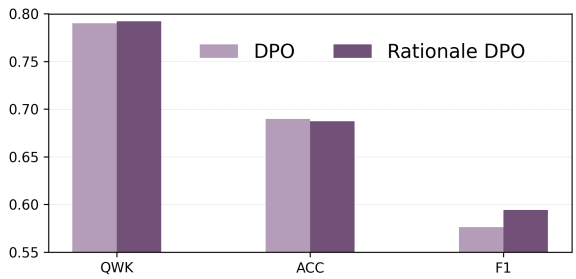

Figure A7: **Regulating DPO training with generated reflections.**

reasoner training methods, enhancing both performance and stability during preference optimisation.

B.6 Case Studies on Our Framework

Critic Oversees Errors and Misinterpret Scopes As shown in Figure A8, the correct assessment of the student's answer is actually **1 point**, not 2 or 3. Although the student lists three items, the first item (volume of vinegar) cleanly maps to the "additional information" that is missing from the procedure. The other two points are either too vague or already addressed in the procedure (e.g., "Determine the mass of each sample" is mentioned, and the procedure does not necessarily require the exact measuring method). Therefore, the response only provides one distinct piece of new information that truly helps replicate the experiment.

The reasoner miscounted the distinct, missing details in the student's answer. The critic model fails to point this oversee. Although three items were listed—vinegar volume, distilled water volume, and mass measurement method—only one (the amount of vinegar) was truly new. The other two were too vague or already in the procedure, leading the reasoner to mistakenly award 2 and 3 points instead of the correct score of 1.

Critic Correctly Identify Intermediate Errors Even Final Scores are Correct As shown in Figure A9, the "reasoner" ultimately awarded the correct score of 2 points but incorrectly characterized the student's conclusion as valid. The "critic" accurately identified that while the conclusion ("plastic C will take the most weight") was not supported by the data, the student still described two valid improvements (more trials, ensuring uniform sample length). This discrepancy shows that the critic model can detect errors in the reasoning—namely, that the conclusion is wrong—even when the final numerical score is correct for other reasons (i.e., providing two legitimate design improvements).

Case Study 3

①***[Question Prompt]*:**
[Question]:
A group of students wrote the following procedure for their investigation.
Procedure:
1.Determine the mass of four different samples.
2.Pour vinegar in each of four separate, but identical, containers.
3.Place a sample of one material into one container and label. Repeat with remaining samples, placing a single sample into a single container.
4.After 24 hours, remove the samples from the containers and rinse each sample with distilled water.
5.Allow the samples to sit and dry for 30 minutes.
6.Determine the mass of each sample.
The students's data are recorded in the table below.
A table contains four columns: Sample, Starting Mass (g), Ending Mass (g), Difference in Mass (g).
The sample for the first row is Marble, with 9.8 Starting Mass, 9.4 Ending Mass and -0.4 for Difference in Mass.
The sample for the second row is Limestone, with 10.4 Starting Mass, 9.1 Ending Mass and -1.3 for Difference in Mass.
The sample for the third row is Wood, with 11.2 Starting Mass, 11.2 Ending Mass and 0.0 for Difference in Mass.
The sample for last row is Plastic, with 7.2 Starting Mass, 7.1 Ending Mass and -0.1 for Difference in Mass.
After reading the group's procedure, describe what additional information you would need in order to replicate the experiment.
Make sure to include at least three pieces of information.
[Key Elements]:
Needed Information:
You need to know how much vinegar was used in each container.
You need to know what type of vinegar was used in each container.
You need to know what materials to test.
You need to know what size/surface area of materials should be used.
You need to know how long each sample was rinsed in distilled water.
You need to know what drying method to use.
You need to know what size/type of container to use.
Other acceptable responses.
[Marking Rubric]:
3 points: The response describes three additional pieces of information that would be needed to accurately replicate the experiment;
2 points: The response describes two additional pieces of information that would be needed to accurately replicate the experiment;
1 point: The response describes one additional piece of information that would be needed to accurately replicate the experiment;
0 point: The response describes little or no accurate or relevant information from the acid rain investigation.
[Student Answer]: In order to replicate this experiment, you would need to know how many ml of vinegar they used, how much distilled water to rinse the samples with and how they obtained the mass of each sample.
Analyze the [Student Answer] thoroughly. Generate a detailed rationale that explains the strengths and weaknesses of the response:
②***[Reasoner]*:**
The student's answer scored 2 points according to the marking rubric, as it describes two additional pieces of information needed to replicate the experiment: the amount of vinegar used ('how many ml of vinegar they used') and the amount of distilled water used for rinsing ('how much distilled water to rinse the samples with'). The student did not specify other necessary details such as the type of vinegar, the materials to test, the size/surface area of materials, the rinsing duration, the drying time, the drying method, or the size/type of container. These missing details are crucial for precise replication and were assessed according to the key answer elements.
③***[Critic]*:**
It appears there was a misunderstanding in your assessment of the student's answer. You noted that the student failed to mention the materials to be tested. However, the student did specify the need to know the amount of vinegar and the amount of distilled water used, which indirectly addresses the variety of materials involved in the experiment. It's crucial to recognize that while the student's response might not explicitly list all desired materials, it does touch upon the conditions necessary to conduct the experiment with various materials. To improve your assessment rationale, focus on what the student has included and ensure that your evaluation aligns with the information provided. Reflect on how each element of the student's answer relates to the experiment's replication and assess accordingly. Please revise your rationale based on this feedback and try to generate a more accurate assessment that acknowledges the student's contributions and missing elements effectively.
④***[Reasoner]*:**
The student's answer scored 3 points according to the marking rubric, as it describes three additional pieces of information needed to replicate the experiment: the amount of vinegar ('how many ml of vinegar they used'), the amount of distilled water for rinsing ('how much distilled water to rinse the samples with'), and the drying time ('Allow the samples to sit and dry for 30 minutes'). These details address the key answer elements regarding the amount of vinegar, the materials to test, and the drying time, fulfilling the requirements for a score of 3 points.
⑤***[Critic]*:**
Rationale Looks Good! [STOP]

Figure A8: **An example that both Reasoner and Critic oversee the mistakes.**

Case Study 4

①***[Question Prompt]***:
[Question]:
A student performed the following investigation to test four different polymer plastics for stretchability.
Procedure:
1. Take a sample of one type of plastic, and measure its length.
2. Tape the top edge of the plastic sample to a table so that it is hanging freely down the side of the table.
3. Attach a clamp to the bottom edge of the plastic sample.
4. Add weights to the clamp and allow them to hang for five minutes.
5. Remove the weights and clamp, and measure the length of the plastic types.
6. Repeat the procedure exactly for the remaining three plastic samples.
7. Perform a second trial (T2) exactly like the first trial (T1).
The student recorded the following data from the investigation.
The table shows the amount of stretch (in millimeters) for four different types of plastic, labeled as A, B, C, and D, when subjected to two different stretching forces, labeled as T1 and T2.
For plastic type A, it stretched 10mm under T1 and 12mm under T2.
For plastic type B, it stretched 22mm under T1 and 23mm under T2.
For plastic type C, it stretched 14mm under T1 and 13mm under T2.
Lastly, for plastic type D, it stretched 20mm under both T1 and T2.
a. Draw a conclusion based on the student's data.
b. Describe two ways the student could have improved the experimental design andor validity of the results.
[Key Elements]:
Conclusions:
Plastic sample B has more stretchability than the other polymer plastics.
Plastic sample A has the least amount of stretchability compared to the other polymer plastics.
Not all polymer plastics have the same stretchability.
Different polymer plastics have different stretchability (and are therefore suited for different applications).
A reasonable conclusion cannot be drawn due to procedural errors.
Other reasonable conclusions
Experimental Design Improvements:
Provide the before and after measurements for length (Did the samples all start out the same size?).
Make sure the samples are all of the same thickness.
Variations in thickness could have caused variations in stretchability.
Perform additional trials.
Some of the samples have similar stretchability (A and C, B and D).
Two trials may not be enough to conclusively state that one is more stretchable than the other.
Indicate how many weights were added to the clamps (Was it the same number for each sample?).
Other acceptable responses
[Marking Rubric]:
3 points: The response draws a valid conclusion supported by the student's data and describes two ways the student could have improved the experimental design andor the validity of the results;
2 points: The response draws a valid conclusion supported by the student's data and describes one way the student could have improved the experimental design and/or the validity of the results. -or- The response describes two ways the student could have improved the experimental design and/or the validity of the results but fails to draw or incorrectly draws a conclusion from the student's data;
1 point: The response draws a valid conclusion supported by the student's data but fails to describe, or incorrectly describes, how the student could have improved the experimental design and/or the validity of the results. -or- The response describes one way the student could have improved the experimental design and/or the validity of the results but fails to draw or incorrectly draws a conclusion from the student's data.;
0 points: The response provides little or no correct information from the polymer investigation.
[Student Answer]: A. Based on the results of the experiment, I believe that plastic C will take the most weight. All the other plastics either extended when more weight was added or stayed the same. Plastic C though actually shrunk in size unlike plastic A which went from 10 to 12 and B which went from 22-23 mm and lastly D witch remained at 20 mm the whole duration. B: A way the student could of improved the experiment validity is by having more trials and having the plastic strips the same length.
Critically assess the [Student Answer] in light of the [Key Answer Elements]. Offer a detailed explanation for the score assigned, referencing specific criteria from the [Marking Rubric]:
②***[Reasoner]***:
The student's answer scored 2 points according to the marking rubric. The student drew a valid conclusion supported by the data, stating 'plastic C will take the most weight,' which aligns with the key answer element that a valid experimental conclusion should be drawn. Additionally, the student proposed one valid experimental design improvement by suggesting 'having more trials,' which aligns with the key answer element about performing additional trials. However, the student did not address other potential improvements such as ensuring uniform sample thickness or specifying the amount of weight added, which limited the score to 2 points.
③***[Critic]***:
It appears there was a misunderstanding in your assessment of the student's answer. The student did not address the experimental conclusion requirements accurately. The expected conclusions were about the stretchability of different polymer plastics and their suitability for different applications, along with a statement about procedural errors affecting the validity of conclusions. The student focused primarily on which plastic would take the most weight, which is not the correct focal point for the experimental conclusion as per the given rubric. Additionally, while you correctly noted the suggestion for increasing the number of trials as a potential improvement, the student's other suggestions such as 'having the plastic strips the same length' do not directly address more critical factors like uniform thickness or the amount of weight added, which are essential for ensuring experimental validity. When revising your rationale, it would be beneficial to closely align your assessment with the specific requirements outlined in the rubric. Reflect on the key elements missed in the student's response, such as the failure to discuss uniform sample thickness or the amount of weight added, and consider how these omissions affect the score. Please take this feedback into account and attempt to regenerate a more accurate and comprehensive rationale.
④***[Reasoner]***:
The student's answer scores 2 points according to the marking rubric. The student fails to draw a valid conclusion from the data, incorrectly stating 'plastic C will take the most weight,' while the data shows plastic B stretched the most. However, the student correctly suggests two experimental design improvements: increasing the number of trials ('...having more trials...') and ensuring uniformity in the length of the plastic samples ('having the plastic strips the same length'). The response thus meets the criteria for 2 points, as it describes two valid ways to improve the experimental design but incorrectly draws a conclusion from the data.
⑤***[Critic]***:
Rationale Looks Good! [STOP]

Figure A9: **An example that Critic can identify intermediate errors even final scores are correct.**

B.7 Case Study: Comparing Critic's Output with Different Sizes

In Figure A10, Qwen3B (the reasoner) mistakenly awards the student's answer 2points rather than the 0 points warranted by the rubric. Comparing critic responses of different model sizes highlights varied degrees of thoroughness in their feedback. Qwen 3B's critic, for instance, repeatedly instructs that "*it's important to ensure that the key elements of each biological process are correctly identified and matched ...*" yet does not provide concrete steps for revising the assessment. By contrast, Qwen7B merely remarks "*Rationale Looks Good! [STOP],*" signalling an abrupt end to any meaningful revision. Progressing to Qwen 14B, the critic offers more constructive guidance by urging: "*You might want to revisit the student's answer and compare it directly with the key elements required by the rubric.*" Finally, Qwen 32B delivers the most comprehensive feedback, emphasizing "*It's important to verify that the terms and processes described align correctly with biological definitions and mechanisms before awarding points based on the rubric*" and detailing how the student's descriptions deviate from correct scientific definitions. This gradual increase in clarity, depth, and actionable insights indicates that larger model sizes (14B and 32B) are more effective at diagnosing errors and recommending precise revisions.

Case Study 5

①***[Question Prompt]*:**
[Question]:
List and describe three processes used by cells to control the movement of substances across the cell membrane.
[Key elements]:
Selective permeability is used by the cell membrane to allow certain substances to move across.
Passive transport occurs when substances move from an area of higher concentration to an area of lower concentration.
Osmosis is the diffusion of water across the cell membrane.
Facilitated diffusion occurs when the membrane controls the pathway for a particle to enter or leave a cell.
Active transport occurs when a cell uses energy to move a substance across the cell membrane, andor a substance moves from an area of low to high concentration, or against the concentration gradient.
Pumps are used to move charged particles like sodium and potassium ions through membranes using energy and carrier proteins.
Membrane-assisted transport occurs when the membrane of the vesicle fuses with the cell membrane forcing large molecules out of the cell as in exocytosis.
Membrane-assisted transport occurs when molecules are engulfed by the cell membrane as in endocytosis.
Membrane-assisted transport occurs when vesicles are formed around large molecules as in phagocytosis.
Membrane-assisted transport occurs when vesicles are formed around liquid droplets as in pinocytosis.
Protein channels or channel proteins allow for the movement of specific molecules or substances into or out of the cell.
[Marking Rubric]:
3 points: Three key elements;
2 points: Two key elements;
1 point: One key element;
0 points: Other.
[Student Answer]: Endocyotsis, when food enters the cell through the cell membrane, exocytosis, when food exits as waste out of the cell, and translation, when the cell surrounds the food, and engulfs it.
Evaluate the [Student Answer] based on the provided [Key Answer Elements] and [Marking Rubric]. Summarize the assessment and justify the score awarded:
②***[Reasoner] Qwen 3B*:**

The student's answer scored 2 points according to the marking rubric, which awards 2 points for correctly addressing two key elements. The student successfully described 'endocytosis' and 'exocytosis' as processes where 'the cell surrounds the food, and engulfs it' and 'when food enters the cell through the cell membrane, exocytosis, when food exits as waste out of the cell,' respectively. These descriptions align with the key elements of membrane-assisted transport when molecules are engulfed by the cell membrane and vesicles are formed around large molecules, as well as exocytosis. However, the answer lacked descriptions of selective permeability, passive transport, osmosis, facilitated diffusion, active transport, the use of pumps, and protein channels, which were necessary to meet additional key elements for a higher score.
***[Critic] Qwen 3B*:**
In reviewing the assessment rationale you provided, it appears there was a misunderstanding in the evaluation of the student's answer. Specifically, the student's description of 'endocytosis' and 'exocytosis' was misinterpreted in key aspects. The student mentioned that 'the cell surrounds the food, and engulfs it' and 'when food exits as waste out of the cell,' respectively. However, these descriptions do not accurately reflect the scientific definitions and processes of endocytosis and exocytosis. Endocytosis involves the cell membrane engulfing materials to form a vesicle, which is then transported into the cell, often used for nutrient absorption or waste expulsion. Exocytosis, on the other hand, involves the cell membrane expelling materials to form a vesicle, which is then transported out of the cell. The student's answer did not mention the critical aspects of vesicle formation or the directional movement of materials in both processes. It's important to ensure that the key elements of each biological process are correctly identified and matched with the student's descriptions before assigning points based on the marking rubric. This will help in providing accurate feedback and scores that truly reflect the student's understanding of the subject matter. As you reflect on this feedback, please revisit the student's answer and the marking rubric, and consider how each part of the student's response aligns with the expected scientific definitions and processes. After reflecting on these points, I encourage you to revise and regenerate the assessment rationale to more accurately represent the student's understanding and the biological processes in question.
***[Critic] Qwen 7B*:**
Rationale Looks Good! [STOP]
***[Critic] Qwen 14B*:**
It appears there was a misunderstanding in your assessment of the student's answer regarding the key elements of endocytosis and exocytosis. You noted that the student described these processes as involving 'membrane-assisted transport when molecules are engulfed by the cell membrane' and 'when vesicles are formed around large molecules.' However, the student's response did not explicitly mention these mechanisms. Instead, they provided general descriptions of the processes without detailing the specific mechanisms like 'membrane-assisted transport' or 'formation of vesicles around molecules.' It's crucial to align your assessment with the exact content provided by the student to ensure accuracy in scoring. You might want to revisit the student's answer and compare it directly with the key elements required by the rubric. This will help you identify precisely what the student mentioned versus what was omitted. A useful strategy could be to mark the presence of each required element in the answer, noting exactly what terms or concepts are mentioned that correspond to those in the rubric. This exercise will not only help in more accurate scoring but also in providing constructive feedback to the student. Reflect on this approach and try to revise your rationale based on these insights.
***[Critic] Qwen 32B*:**
It appears there was a misunderstanding in the assessment of the student's answer regarding the description of 'endocytosis' and 'exocytosis'. You credited the student for accurately describing these processes in terms of membrane-assisted transport and formation of vesicles around large molecules. However, the student's description was inaccurate and did not clearly distinguish between the processes and concepts involved. The student incorrectly associated 'endocytosis' with the cell surrounding and engulfing food, which is a misrepresentation of the process. Similarly, the mention of 'exocytosis' was coupled with an incorrect description of food exiting the cell, rather than correctly describing it as the process where substances are expelled from the cell through vesicle fusion with the cell membrane. Furthermore, the student's answer included 'translation,' which is not a membrane-assisted transport process but rather a stage of protein synthesis. This indicates a fundamental misunderstanding or misstatement of the biological processes involved. To improve your assessment rationale, you should focus on the accuracy of the student's descriptions in relation to the scientific concepts asked in the question. It's important to verify that the terms and processes described align correctly with biological definitions and mechanisms before awarding points based on the rubric. Reflect on these points and consider how a more accurate assessment can be framed, emphasizing the importance of precise scientific terminology and processes. Please revise your rationale based on this feedback and try to generate a more accurate assessment of the student's answer.

Figure A10: **Comparing Critic model's output with different parameter sizes.**