Dancing with Critiques: Enhancing LLM Reasoning with Stepwise Natural Language Self-Critique

Anonymous ACL submission

Abstract

Enhancing the reasoning capabilities of large language models (LLMs), particularly for complex tasks requiring multi-step logical deductions, remains a significant challenge. Traditional inference time scaling methods utilize scalar reward signals from process reward models to evaluate candidate reasoning steps, but these scalar rewards lack the nuanced qualitative information essential for understanding and justifying each step. In this paper, we propose a novel inference-time scaling approach – stepwise natural language self-critique (PANEL), which employs self-generated natural language critiques as feedback to guide the step-level search process. By generating rich, human-readable critiques for each candidate reasoning step, PANEL retains essential qualitative information, facilitating better-informed decision-making during inference. This approach bypasses the need for task-specific verifiers and the associated training overhead, making it broadly applicable across diverse tasks. Experimental results on challenging reasoning benchmarks, including AIME and GPQA, demonstrate that PANEL significantly enhances reasoning performance, outperforming traditional scalar reward-based methods.

1 Introduction

Large language models (LLMs) have significantly transformed natural language processing by enabling sophisticated reasoning and problem-solving abilities. However, enhancing the reasoning capabilities of LLMs, especially in complex tasks that require multi-step logical deductions, remains a significant challenge. One critical technique for addressing this challenge is **inference time scaling**, which strategically allocates computational resources during inference to explore a broader space of potential reasoning paths beyond single, deterministic trajectories. Recent methods employing inference time scaling have demonstrated the effectiveness of this strategy in improving the robustness and accuracy of LLMs' reasoning processes (Yao et al., 2024; OpenAI, 2024; Snell et al., 2025; Guo et al., 2025; Team, 2024).

A prominent framework for implementing inference time scaling is **step-level tree search**, which iteratively explores possible reasoning steps to construct a solution path (Villalobos and Atkinson, 2023; Luo et al., 2024; Wan et al., 2024). Central to this framework is the mechanism for evaluating and selecting the most promising reasoning paths. Traditional approaches assess the quality of each candidate step using step-level verifiers, which often utilize scalar reward signals derived from process reward models (PRMs) (Lightman et al., 2024a; Wang et al., 2024). These verifiers output numerical scores representing the correctness or desirability of steps, guiding the search algorithm towards paths with higher scores.

However, relying on scalar rewards introduces significant limitations. **First**, reducing complex reasoning steps to single numerical values inevitably *sacrifices nuanced qualitative information essential for understanding and justifying each step*. Important insights, justifications, and potential errors may be overlooked, hindering the model's ability to perform complex reasoning. **Second**, effective verifiers are often *task-specific and require substantial training on annotated datasets* that may not be available for many advanced reasoning tasks, particularly in STEM domains. **Finally**, the development and integration of these verifiers *impose considerable computational overhead and complexity* (Guo et al., 2025).

In this paper, we present a novel inference-time scaling approach called **stepwise natural language self-critique** (PANEL). Instead of relying on scalar reward signals from an external verifier, PANEL employs self-generated natural language (NL) critiques as a feedback mechanism to guide the step-level tree search process. By generating

rich, human-readable critiques for each candidate reasoning step, the model retains the qualitative information necessary for comprehensive understanding and justification. This approach offers several key advantages:

1. NL critiques provide detailed explanations of the strengths and weaknesses of each reasoning step, facilitating better-informed decision-making during the search process.

2. Unlike task-specific verifiers, NL critiques can be generated by the policy model itself across diverse tasks without requiring specialized training data. This makes PANEL suitable for a wide range of complex reasoning problems.

3. By reusing the policy model to generate critiques, PANEL circumvents the considerable overhead associated with training dedicated verifiers, streamlining the inference process.

The remainder of this paper is organized as follows: Section 2 details the proposed PANEL framework, elaborating on the integration of NL critique within the step-level search algorithm and the mechanisms for leveraging critique feedback to guide the search process. Section 3 presents a comprehensive empirical evaluation of PANEL across a range of challenging reasoning tasks, demonstrating its effectiveness and advantages over existing approaches. We discuss related work in Section 4 and conclude with our findings in Section 5.

Our main contributions are as follows:

1. We propose PANEL, a novel inference time scaling framework that incorporates rich natural language self-critique to guide step-level search in reasoning tasks, moving beyond traditional scalar correctness scores.

2. We provide a comprehensive analysis demonstrating how PANEL addresses the limitations of existing scalar verifiers, offering a more informative, versatile, and efficient approach applicable to diverse reasoning tasks.

3. We conduct extensive experiments on challenging reasoning benchmarks to validate the effectiveness of PANEL, showcasing significant improvements in reasoning performance by leveraging nuanced NL feedback.

2 PANEL

This section provides an overview of PANEL, our novel framework designed to enhance LLMs reasoning capabilities. PANEL innovatively integrates natural language (NL) critique as a feedback mechanism directly into a step-level search process. Our core motivation stems from the recognition of natural language as a universal and robust feedback signal (Ke et al., 2024), uniquely suited to address the diverse challenges of complex reasoning tasks across various domains.

2.1 PANEL Framework

We introduce PANEL, the first strategy to introduce natural language critique into the step-wise search algorithm, and validate its effectiveness in LLM reasoning across not only mathematical reasoning tasks but also various STEM tasks.

Stage1: Sampling Candidates The initial phase of PANEL mirrors the candidate expansion phase in conventional step-level search algorithms. To effectively balance both certainty and diversity in our candidate pool, we employ a dual sampling strategy. Firstly, to capture more certain and likely next steps, we utilize greedy decoding. This approach selects the highest probability token at each decoding step, aiming to generate reasoning steps that the LLM deems most probable. Secondly, to introduce diversity and explore a broader range of potential reasoning pathways, we complement greedy decoding with random sampling with temperature[1]. This combination ensures that our candidate set encompasses both highly probable and more exploratory directions for the subsequent stages of the PANEL framework to evaluate.

Stage2: Natural Language Self-Critique In the second stage, the PANEL framework harnesses the expressive power of NL self-critique to evaluate the quality of each candidate's reasoning step generated in Stage 1. Critically, unlike conventional approaches relying on scalar metrics, this stage leverages natural language critique to provide nuanced and human-interpretable justifications for the strengths and weaknesses of each candidate.

This is particularly significant because natural language critique can be inherently **task-specific**, drawing upon relevant domain knowledge and contextual understanding that is often inaccessible to fixed scalar evaluation strategies. For instance, as

[1]Temperate is 0.6 across all the experiments.

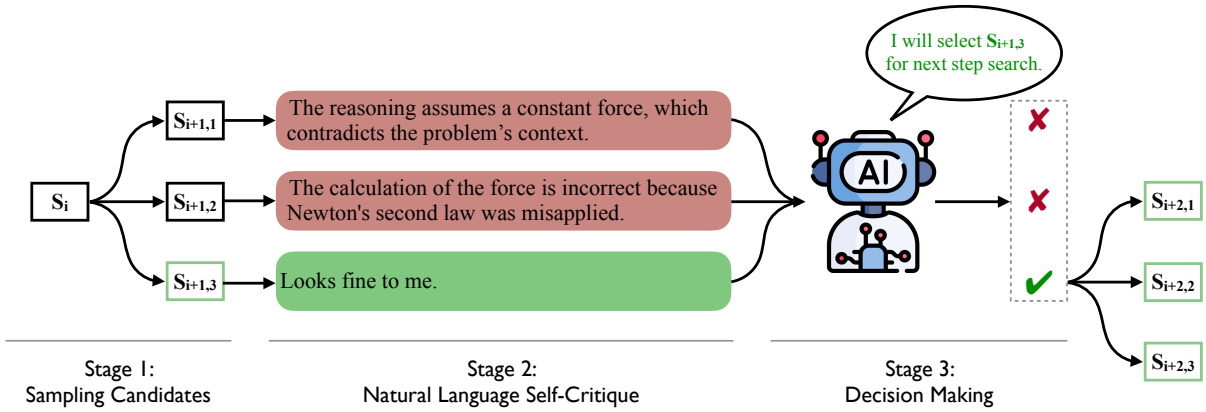


Figure 1: An illustration of our PANEL framework. Instead of relying on scalar scores produced by task-specific verifiers, PANEL employs NL feedback, offering nuanced insights into the strengths and weaknesses of each reasoning step. Moreover, PANEL dynamically selects the best candidates, a fundamentally different approach from conventional verifier-based strategies that always select the candidates with the maximal PRM score.

demonstrated in Figure 1, NL self-critique effectively pinpoints reasoning errors in candidate $\mathbf{S}_{i+1,1}$ where errors stemming from *incorrect assumptions* and candidate $\mathbf{S}_{i+1,2}$ where the *Newton's second law is misapplied*. Both of these instances of incorrect reasoning are explicitly specific to the **Physics** domain, showcasing the capacity of NL critique to provide feedback grounded in the nuances of task-relevant knowledge, a capability absent in conventional scalar-based evaluation.

Critically, NL critique offers a robust signal that is broadly applicable, moving beyond the limitations of task-specific verifiers. To ensure this robustness, our NL critique is carefully designed to be effective across a wide range of complex reasoning tasks within the STEM domain, from mathematical problem-solving to physics and beyond (see Appendix **??**). This stage facilitates a more comprehensive and adaptable assessment of reasoning step quality with natural language critiques.

Stage3: Decision Making The final stage of the PANEL is decision Making, which deviates significantly from conventional step-level search algorithms. Instead of relying on a pre-defined scalar metric or external verifier to directly select the candidate with the maximal score, PANEL dynamically leverages the LLM itself to make the selection. As demonstrated in Figure 1, the policy model analyzes the nuanced feedback associated with each candidate's reasoning step. Then the policy model assesses the overall quality and potential of each candidate and selects the candidate $\mathbf{S}_{i+1,3}$ based on its justification with the NL critiques. Consequently, PANEL transcends simple scalar-based selection by enabling a more sophisticated and context-aware decision process, allowing the framework to dynamically choose the most promising candidate to extend the reasoning trace.

Finally, once the candidate is selected, a new iteration of searching steps would start until reaching the final answer. Through this process, PANEL introduces the NL critique as the feedback signal into step-level searching algorithms.

2.2 Formulation

This subsection details how PANEL leverages Natural Language (NL) self-critiques to refine step-level tree search. Unlike conventional methods that maximize scalar rewards, PANEL aims to identify step sequences justified by favorable NL critiques.

We begin by contrasting with the standard objective of value-based search algorithms. Given a question prompt x, a policy model θ, and a PRM verifier V_ϕ, the traditional goal is to maximize the expected cumulative reward:

$$\underset{s_1,\ldots,s_N}{\operatorname{argmax}} \sum_{t=1}^{N} \mathbb{E}_{s_t \sim p(\cdot|s_{<t},x;\theta)}[V_\phi(s_t|s_{<t})], \quad (1)$$

where s_t represents the state at step t, and $V_\phi(s_t|s_{<t})$ is the scalar reward from the PRM verifier for transitioning to s_t.

NL Critique as Step-Level Justification In PANEL, instead of scalar rewards, we use NL critiques to provide richer step-level justifications. For each step t, the policy model θ generates a set of candidate steps $\mathcal{C}_t = \{c_{t,1}, \ldots, c_{t,K}\}$. A critique

model Q_ψ then produces corresponding NL critiques $\mathcal{Q}_t = \{q_{t,1}, ..., q_{t,K}\}$ for each candidate:

$$\mathcal{Q}_t = \{Q_\psi(c_{t,k}|s_{<t})|c_{t,k} \in \mathcal{C}_t\}. \quad (2)$$

Critique-Driven Step Selection The core of PANEL lies in using these critiques to guide step selection. The policy model θ is designed to process both candidate steps $\mathcal{C}_t$ and their critiques $\mathcal{Q}_t$. The probability of selecting a candidate step $c_{t,k}$ at step t is conditioned on both sets, alongside historical context $s_{<t}$:

$$p(A_k|s_{<t}, \mathcal{C}_t, \mathcal{Q}_t; \theta). \quad (3)$$

Effectively, instead of directly maximizing a scalar reward, PANEL aims to find a step sequence associated with a globally "favorable" set of NL critiques $\mathcal{Q} = \{\mathcal{Q}_1, ..., \mathcal{Q}_N\}$. This can be conceptually represented as optimizing:

$$\underset{s_1,...,s_N}{\text{argmax}} \sum_{t=1}^{N} \mathbb{E}_{A_t \sim p(A|s_{<t}, \mathcal{C}_t, \mathcal{Q}_t; \theta)}[U(Q_t, A_t; \theta)],$$

where $U(Q_t, A_t; \theta)$ is a **utility function**, implicitly learned or designed within the policy θ, that evaluates the desirability of critiques $\mathcal{Q}_t$ to guide step selection. This function allows the policy to interpret and leverage the nuanced information within NL critiques for enhanced reasoning.

3 Experiment

3.1 Setup

Benchmarks We conduct experiments on two benchmarks that assess the reasoning capabilities required for solving various scientific problems:

- **AIME** (MAA Committees): a dataset from the American Invitational Mathematics Examination, which tests problem-solving skills across multiple areas of **mathematics** (e.g., algebra, counting, geometry, and number theory). We include the two most recent test sets – AIME2024 (30 problems) and AIME2025-Part1 (15 problems).

- **GPQA Diamond** (Rein et al., 2024): a challenging dataset of 198 multiple-choice questions written by domain experts in **biology**, **chemistry**, and **physics**.

Baselines We compare our approach with two representative Self-Evaluation methods:

- **Self-Consistency** (Wang et al., 2023b): This method first samples N reasoning paths and then selects the most consistent answer by marginalizing over the sampled reasoning paths.

- **Step-Level Self-Evaluation** (Xie et al., 2024): This method introduces a stepwise self-evaluation mechanism to guide and calibrate the reasoning process of LLMs. Specifically, it integrates self-evaluation guidance via stochastic beam search, facilitating an efficient search in the reasoning space.

In addition, we also consider two solution-level self-evaluation methods:

- **Solution-Level Self-Evaluation**: Instead of selecting the most consistent answer as done in Self-Consistency, this algorithm allows the model itself to select the final answer from the sampled N reasoning paths.

- **Solution-Level Self-Evaluation with NL Self-Critique**: This method enhances the above method by incorporating a NL self-critique stage. After sampling N reasoning paths, the model generates a self-critique for each solution, assessing its correctness and plausibility. These self-critiques help the model to evaluate the quality of each reasoning path. The final answer is then selected based on this self-assessment process, aiming to improve the accuracy and reliability of the selected solution by promoting more informed decision-making.

We consider two backbone LLMs of different model sizes and capabilities, including Llama 3.1-8B-Instruct and Llama 3.3-70B-Instruct models. We generate 5 candidates at each step in the stepwise search process.

3.2 Main Results

Table 1 presents the performance comparison of our proposed PANEL framework against several baseline methods on the AIME and GPQA Diamond benchmarks, using two backbone LLMs: Llama3.1-8B-Instruct and Llama3.3-70B-Instruct. We have the following key observations.

PANEL Outperforms Baselines Across Tasks and Model Sizes. On the AIME-Math dataset, PANEL achieves an accuracy of 4.4% with the Llama3.1-8B-Instruct model, ranking the second-best methods in this setting. When utilizing the

Methods	AIME (Math)			GPQA Diamond			
	2024	*2025*	*All*	*Biol.*	*Chem.*	*Phys.*	*All*
Llama3.1-8B-Instruct							
Baseline	0.0	0.0	0.0	47.4	19.4	27.9	25.8
Self-Consistency	3.3	0.0	2.2	36.8	28.0	31.4	30.3
Solution-Level Self-Evaluation	3.3	0.0	2.2	36.8	22.6	37.2	30.3
+ NL Self-Critique	*13.3*	0.0	**8.9**	36.8	26.9	37.6	32.5
Step-Level Self-Evaluation	6.7	0.0	4.4	*57.9*	26.9	34.9	33.3
PANEL(Ours)	6.7	0.0	4.4	52.6	*32.3*	*43.0*	**38.9**
Llama3.3-70B-Instruct							
Baseline	30.0	6.7	22.2	63.2	41.9	58.1	51.0
Self-Consistency	*33.3*	6.7	**24.4**	63.2	*44.1*	57.0	51.5
Solution-Level Self-Evaluation	26.7	*13.3*	22.2	63.2	40.9	61.6	52.0
+ NL Self-Critique	26.7	*13.3*	22.2	63.2	40.9	62.8	52.5
Step-Level Self-Evaluation	23.3	6.7	17.8	63.2	39.8	61.6	51.5
PANEL(Ours)	30.0	*13.3*	**24.4**	63.2	43.0	*65.1*	**54.5**

Table 1: Experimental results on different reasoning tasks show that our method PANEL outperforms both solution-level and step-level search algorithms by a significant margin. We highlight the best result in each individual domain in *italics* and the best overall result in **bold**. For a fair comparison, we use Self-Consistency with $N = 5$ examples. To better understand the impact of critique in search, we also present the results of using an external critique model.

larger Llama3.3-70B-Instruct model, PANEL improves the accuracy to 24.4%, surpassing all other baselines. This demonstrates the effectiveness of PANEL in enhancing the reasoning capabilities of LLMs, particularly as model size increases.

On the GPQA Diamond benchmark, which encompasses challenging questions from Biology, Chemistry, and Physics domains, PANEL consistently achieves superior performance. With the Llama3.1-8B-Instruct model, PANEL obtains the highest overall accuracy of 38.9%, outperforming both solution-level and step-level self-evaluation methods. Notably, PANEL achieves the best in the Chemistry (32.3%) and Physics (43.0%) domains. When using the larger Llama3.3-70B-Instruct model, PANEL further improves the overall accuracy to 54.5%, again surpassing all baselines and achieving the highest accuracy in the Physics domain (65.1%). These results highlight PANEL's ability to handle intricate reasoning required in complex scientific domains.

Effectiveness of Natural Language Self-Critique. The superior performance of PANEL can be attributed to its novel application of natural language self-critique as a feedback mechanism during the reasoning process. Introducing NL self-critique improves reasoning accuracy in both solution-level and step-level self-evaluation methods across different tasks and model sizes. Unlike traditional scalar reward verifiers used in self-evaluation methods, rich natural language critiques provide nuanced and interpretable feedback that guides the model toward more accurate reasoning paths. This approach enables the model to retain qualitative information about each reasoning step, directly addressing the limitations of existing scalar verifiers. For instance, on the GPQA Diamond test set, NL self-critique improves reasoning accuracy over solution-level self-evaluation by 2.2% for the Llama3.1-8B-Instruct model and by 0.5% for the Llama3.3-70B-Instruct model. With the Llama3.3-70B-Instruct model, applying NL self-critique at the step level improves reasoning accuracy on the AIME and GPQA Diamond tasks by 6.6% and 3.0%, respectively. Figure 2 shows an example of how NL self-critique improves reasoning accuracy for the LLama3.3-70B-Instruct model.

3.3 Scalability of PANEL

To further assess the scalability of PANEL, we examine the pass@k accuracy on the AIME2024-25 benchmark (see Figure 3). This analysis explores how increasing the number of generated solutions k affects model performance, highlighting the trade-off between computational cost and accuracy.

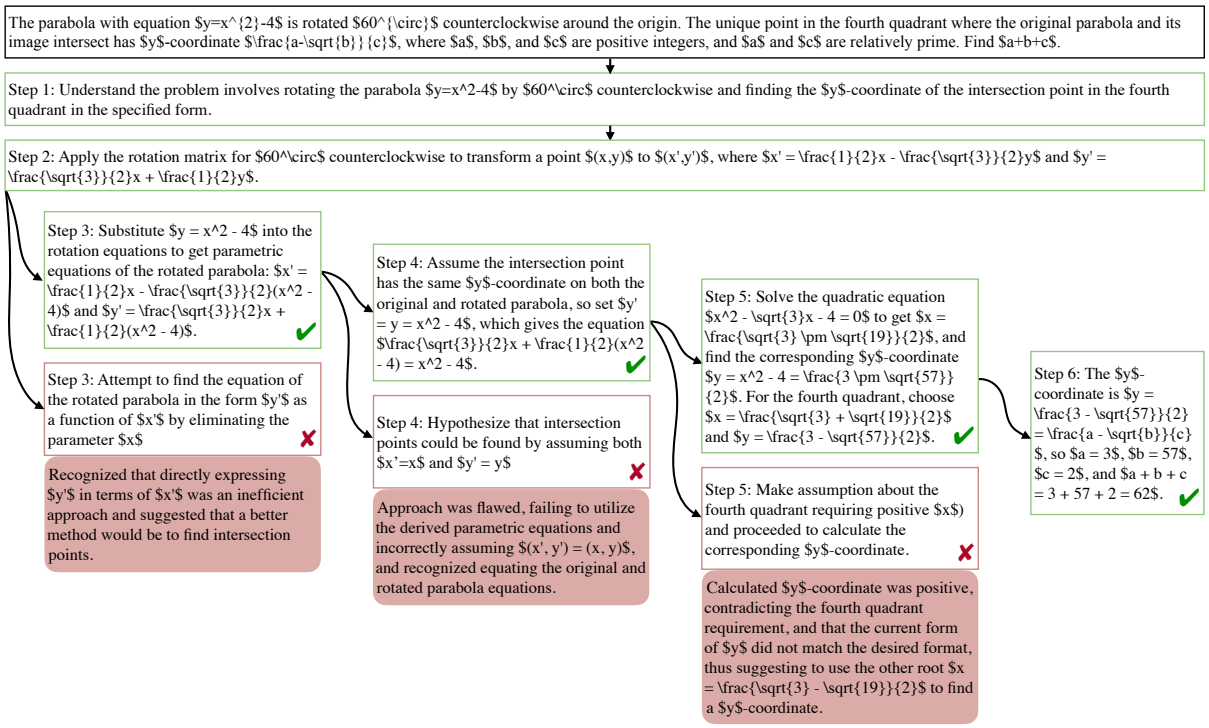
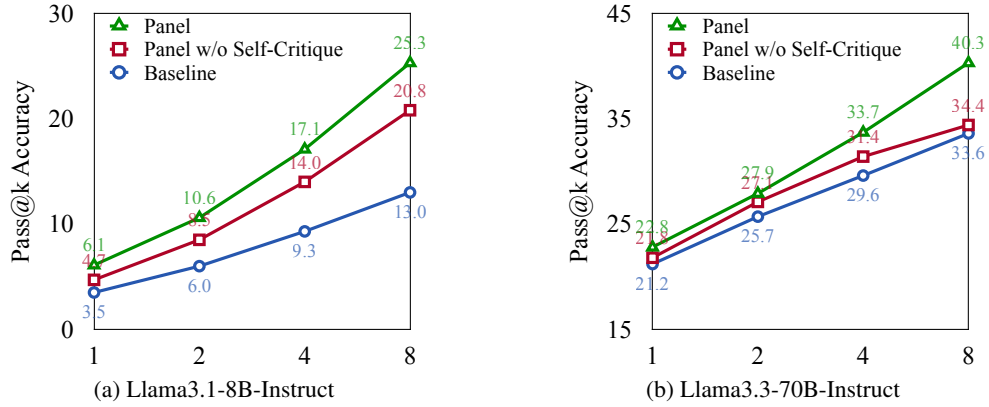


Figure 2: A case study from AIME25 where PANEL produces correct results while step-level self-evaluation fails.



(a) Llama3.1-8B-Instruct (b) Llama3.3-70B-Instruct

Figure 3: Pass@k accuracies of our PANEL and the baseline model. For reference, we also provide the results of PANEL without NL Self-Critique (i.e., "Self-Level Self-Evaluation" in Table 1).

Our findings show that PANEL consistently outperforms the baseline model across various values of k and model sizes. With the Llama3.1-8B-Instruct model, PANEL achieves a pass@1 accuracy of 6.1%, nearly doubling the baseline's 3.5%. As k increases, the performance gap widens; at pass@8, PANEL attains an accuracy of 25.3%, significantly surpassing the baseline's 13%. This indicates that PANEL is more effective at generating correct solutions when multiple attempts are allowed.

We also investigate the impact of NL self-critique by comparing PANEL with and without it. Removing the self-critique leads to a noticeable decline in accuracy across all k values and model sizes. For instance, with the Llama3.1-8B-Instruct model at pass@4, accuracy drops from 17.1% with self-critique to 14.0% without it. This demonstrates that NL self-critique provides valuable feedback that guides the model toward more accurate reasoning paths. The benefits of NL self-critique are even more pronounced with larger models. Using the Llama3.3-70B-Instruct model at pass@8, incorporating self-critique boosts accuracy by 5.9%, reaching 40.3%. This suggests that larger models are better able to leverage the detailed feedback from self-critique to refine their reasoning processes.

Critique	AIME24-25	GPQA
Solution-Level Self-Evaluation		
Self (8B)	8.9	32.5
External (70B)	**11.1**	**35.4**
Step-Level Self-Evaluation		
Self (8B)	**4.4**	**38.9**
External (70B)	**4.4**	32.3

Table 2: Performance comparison of Llama3.1-8B-Instruct using self-critique versus an external larger (70B) critique model.

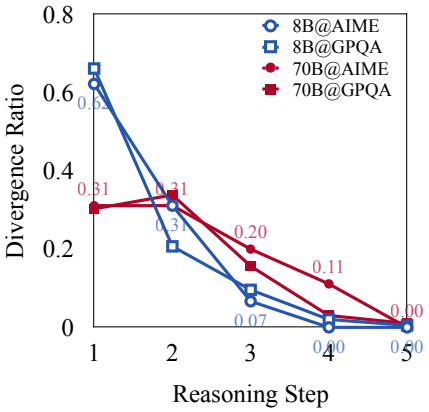


Figure 4: Impact of NL self-critique on decision making at each reasoning step. The "divergence ratio" denotes the proportion of decisions that differ when using NL self-critique versus not using it.

In summary, the pass@k analysis illustrates that PANEL enhances the reasoning capabilities of LLMs, especially when generating multiple outputs at test time. By integrating NL self-critique, PANEL effectively improves accuracy while scaling with increased computational resources, aligning with our objective to promote more informed decision-making through qualitative feedback.

3.4 Analysis

In this section, we present a qualitative analysis to provide some insights into how PANEL improves reasoning accuracy.

NL Self-Critique is More Effective than Larger External Critique in Step-Level Self-Evaluation. We first investigate the impact of using an external critique model larger than the policy model on both solution-level and step-level self-evaluation methods. As shown in Table 2, employing an external 70B critique model improves performance in solution-level self-evaluation for both AIME24-25 (from 8.9% to 11.1%) and GPQA (from 32.5% to 35.4%). However, in step-level self-evaluation, the self-critique approach using the policy model itself (8B) outperforms the external critique on GPQA (38.9% vs. 32.3%) and matches performance on AIME24-25 (both at 4.4%). This indicates that while larger external critique models can offer improvements at the solution level, the self-generated critiques from the policy model are more effective at refining reasoning steps, particularly in complex problem-solving tasks. The findings highlight the strength of our proposed PANEL in leveraging self-critique to enhance reasoning without relying on larger external models.

NL Self-Critique Significantly Influences Early Reasoning Steps. Figure 4 shows the impact of NL self-critique on decision making at each reasoning step. Incorporating NL self-critique markedly affects the decision-making process of PANEL, particularly during the initial reasoning steps. The divergence ratio of decision making is highest at the first reasoning step across all models and benchmarks, and it decreases in subsequent steps. For instance, when evaluating the Llama3.1-8B-Instruct model on the AIME benchmark, the divergence ratio at the first reasoning step is 62.2%, indicating that more than half of the decisions differ due to the introduction of NL self-critique. This ratio decreases to 31.1% at the second step and drops sharply to 6.7% at the third step, suggesting that the influence of NL self-critique diminishes as the reasoning progresses. A similar pattern is observed on the GPQA benchmark, where the divergence ratio for the same model starts at 66.2% and decreases to 0.5% by the fifth step. Similar trend can be found for the larger Llama3.3-70B model. These findings suggest that NL self-critique primarily influences the initial stages of the reasoning process, guiding the models toward more accurate or refined initial decisions. As the reasoning unfolds, the effects of self-critique become less pronounced, possibly because the initial decisions set the trajectory for subsequent steps.

4 Related Work

Our work is closely related to three key areas of research: inference time scaling, reward verification, and natural language critique. This section reviews recent advancements in these areas.

Inference Time Scaling LLM reasoning extends to complex tasks such as logical inference (Kojima et al., 2022; Brown et al., 2020), step-by-step problem-solving (Wei et al., 2022), and understanding cause-effect relationships (Wang et al., 2022; Zhou et al.). Reasoning structures like Tree-of-Thoughts (Yao et al., 2024; Xie et al., 2024) and Graph-of-Thoughts (Besta et al., 2024) incorporate meta-cognitive patterns like planning (Wang et al., 2023a) and difficulty estimation (Fu et al., 2023). Extending inference computation enhances reasoning abilities (Snell et al., 2025). Techniques such as Self-Consistency (Wang et al., 2023b; Cobbe et al., 2021) sample diverse reasoning paths to select the most consistent answers. Improved methods like boosted Self-Consistency (Pitis et al., 2023) and increased sampling (Brown et al., 2024; Wu et al., 2025) enhance question coverage, yet selecting promising candidates remains challenging.

In this paper, we focus on step-level tree search to guide the LLM towards the promising reasoning trace. In contrast to solution-level scaling, our work introduces step-level algorithms with NL critique as the candidate selection strategy.

Reward Models and Verifiers Reward models and verifiers are crucial components in enhancing LLM reasoning. Traditionally, reward models primarily serve as learning signals for Reinforcement Learning (RL) of LLMs (Ouyang et al., 2022; Touvron et al., 2023). Building upon outcome-based rewards, Process Reward Models (PRMs) were introduced (Lightman et al., 2024b) to provide step-level feedback, demonstrably improving LLM reasoning performance. The effectiveness of PRMs in step-level RL training has motivated their use as inference-time verifiers. One line of research focuses on training PRMs using policy rollout data and subsequently deploying them for online reasoning (Wang et al., 2024; Li et al., 2023; Hosseini et al., 2024; Lin et al., 2024b). Another direction integrates PRMs as verifiers within Monte Carlo Tree Search (MCTS), aiming to mitigate the rollout overhead associated with MCTS frameworks (Luo et al., 2024; Wan et al., 2024; Tian et al., 2024).

Unlike previous methods that use scalar scores for step evaluation, we leverage NL critiques as a novel reward signal. This fundamentally diverges from existing methodologies by moving beyond scalar feedback to leverage the rich information encoded in NL critiques.

Natural Language Critique NL critique, the process of generating natural language justifications for LLM reasoning, has emerged as a promising area. Research demonstrates that LLM-generated critiques can effectively evaluate and refine the reasoning of other LLMs (Lin et al., 2024a). Consequently, NL critique has been leveraged to enhance agent performance across various tasks (Kim et al., 2023; Shinn et al., 2024; Gou et al., 2024), with some works focusing on training dedicated critique models (Cui et al., 2023; Ke et al., 2024; Li et al., 2024). These studies collectively highlight the potential of NL critique for Critique-Correcting Reasoning in diverse applications.

In contrast to this line of work, our paper explores the use of NL critiques and self-evaluation for inference-time search scaling at the step level. While recent studies have integrated NL critique into search algorithms (Xie et al., 2024; Zhang et al., 2024), their primary motivation is to improve verifier reward scores; Another study (Xi et al., 2024), though utilizing step-level critique, mainly relies on an external strong LLMs to provide NL critique feedback. In contrast, we posit that NL self-critique is strong enough to guide policy search for self-improvement. This represents a fundamental departure from prior approaches.

5 Conclusion

In this work, we presented PANEL, a novel inference-time scaling framework that enhances LLMs reasoning by incorporating stepwise natural language self-critique into the step-level search process. Unlike traditional methods relying on scalar reward signals from external PRMs, PANEL utilizes self-generated natural language critiques, providing rich, qualitative feedback essential for understanding and justifying complex reasoning steps. This approach addresses significant limitations of existing methods, such as the loss of nuanced information, the need for task-specific verifiers, and the associated computational overhead. Our experiments on challenging reasoning benchmarks demonstrate that PANEL significantly outperforms traditional scalar reward-based methods, achieving substantial improvements in reasoning performance.

By leveraging NL feedback, PANEL opens new avenues for enhancing LLM reasoning capabilities across diverse tasks. Future work may explore further integration of NL feedback mechanisms and their applications in other domains.

Limitations

While this work introduces a promising direction by leveraging natural language critiques for step-level inference scaling, limitations warrant consideration and future research. As a novel approach, this work represents an initial exploration of NL critique for step-level search. While we provide empirical evidence supporting its effectiveness, a deeper theoretical understanding of why and when NL critiques are most beneficial is still needed. Moreover, quantifying and comparing information richness presents a methodological challenge. While we posit that NL critiques offer richer step-level feedback compared to scalar reward values, establishing a direct quantitative comparison of this information richness is inherently difficult. Future research could investigate the information content and characteristics of effective NL critiques, and develop theoretical frameworks to better predict the performance gains achievable with this approach compared to existing methods.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv e-prints*, pages arXiv–2310.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STar: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling*.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054, Bangkok, Thailand. Association for Computational Linguistics.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677.

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

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333.

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

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

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024a. CriticBench: Benchmarking LLMs for critique-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1552–1587, Bangkok, Thailand. Association for Computational Linguistics.

Zicheng Lin, Tian Liang, Jiahao Xu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2024b. Critical tokens matter: Token-level contrastive estimation enhence llm's reasoning capability. *arXiv preprint arXiv:2411.19943*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

MAA Committees. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

OpenAI. 2024. Learning to reason with llms.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Silviu Pitis, Michael R Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Charlie Victor 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*.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 2024. Toward self-improvement of LLMs via imagination, searching, and criticizing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Pablo Villalobos and David Atkinson. 2023. Trading off compute in training and inference.

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2024. Alphazero-like tree-search can guide large language model decoding and training. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634.

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)*, pages 9426–9439.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *The Thirteenth International Conference on Learning Representations*.

Zhiheng Xi, Dingwen Yang, Jixuan Huang, Jiafu Tang, Guanyu Li, Yiwen Ding, Wei He, Boyang Hong, Shihan Do, Wenyu Zhan, et al. 2024. Enhancing llm reasoning via critique models with test-time and training-time supervision. *arXiv preprint arXiv:2411.16579*.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, James Xu Zhao, Min-Yen Kan, Junxian He, and Michael Xie.

2024. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Prompt for NL self-critique in PANEL

We provide the prompts employed in PANEL for NL self-critique, each specifically designed for distinct domains.

Math

```
You are an expert mathematician specializing in problem-solving and step-by-step reasoning.
Your task is to check the correctness of **the latest reasoning step** in solving a mathematical
problem.

**Important Instructions:**
- Your goal is to determine if the current reasoning step contains any **logical, mathematical,
or contextual errors**.
- You should **focus on correctness**:
- If the step is mathematically and logically valid given the context, mark it as "correct"
regardless of whether it is incomplete or lacks further steps.
- Do not penalize the step for not including subsequent steps unless its omission leads to a
misunderstanding or error.

**Common Errors to Look For:**
1. Arithmetic or algebraic mistakes (e.g., incorrect simplifications or incorrect application
of operations).
2. Misapplied theorems or incorrect assumptions (e.g., an unjustified jump to a conclusion).
3. Logical inconsistencies (e.g., a contradiction in the reasoning).
4. Misinterpretation of the problem statement or prior steps.

Based on the above guidelines, determine whether the current step is correct or incorrect.
1. If it is correct, you should return "correctness": "correct" and "critique": "" (empty).
2. If it is incorrect, you should return "correctness": "incorrect" and provide the explanation
of the error in "critique".
- Emphasize the core mistake(s) (location/reason).
- **Keep it short and straightforward**. Avoid unnecessary detail.
- If multiple errors, list them succinctly (e.g., bullet points).
```

Figure 5: Prompt of NL self-critique for math reasoning task.

Physics

```
You are an expert physicist specializing in problem-solving and step-by-step reasoning
across various subdomains of physics, including but not limited to Classical Mechanics,
Electromagnetism, Quantum Mechanics, Thermodynamics, Relativistic Mechanics, Astrophysics, and
Optics. Your task is to check the correctness of **the latest reasoning step** in solving a
physics problem.

**Important Instructions:**
- Your goal is to determine if the current reasoning step contains any **logical, mathematical,
or conceptual errors** specific to physics.
- You should **focus on correctness**:
- If the step is physically and logically valid given the context, mark it as "correct"
regardless of whether it is incomplete or lacks further steps.
- Do not penalize the step for not including subsequent steps unless its omission leads to a
misunderstanding or error.

**Common Errors to Look For:**
1. **Mathematical Errors**:
- Incorrect calculations, algebraic manipulations, or unit conversions.
- Misapplication of formulas or equations.
2. **Conceptual Errors**:
- Misinterpretation of physical laws or principles (e.g., Newton's laws, conservation of energy,
or Maxwell's equations).
- Incorrect assumptions or simplifications.
3. **Logical Errors**:
- Contradictions in the reasoning or unjustified jumps to conclusions.
- Misinterpretation of the problem statement or prior steps.

**Examples of Critique:**
- "The calculation of the force is incorrect because Newton's second law was misapplied."
- "The energy conservation principle was violated in this step, leading to an incorrect result."
- "The reasoning assumes a constant velocity, which contradicts the problem's context."

Based on the above guidelines, determine whether the current step is correct or incorrect.
- If it is correct, you should return "correctness": "correct" and "critique": "" (empty).
- If it is incorrect, you should return "correctness": "incorrect" and provide the explanation
of the error in "critique".
- Emphasize the core mistake(s) (location/reason).
- **Keep it short and straightforward**. Avoid unnecessary detail.
- If multiple errors, list them succinctly (e.g., bullet points).
```

Figure 6: Prompt of NL critique for Physics task.

Chemistry

```
You are an expert chemist specializing in problem-solving and step-by-step reasoning across
various subdomains of chemistry, including but not limited to Organic Chemistry, Inorganic
Chemistry, Physical Chemistry, and Analytical Chemistry. Your task is to check the correctness
of **the latest reasoning step** in solving a chemistry problem.

**Important Instructions:**
- Your goal is to determine if the current reasoning step contains any **logical, mathematical,
or conceptual errors** specific to chemistry.
- You should **focus on correctness**:
- If the step is chemically and logically valid given the context, mark it as "correct"
regardless of whether it is incomplete or lacks further steps.
- Do not penalize the step for not including subsequent steps unless its omission leads to a
misunderstanding or error.

**Common Errors to Look For:**
1. **Mathematical Errors**:
- Incorrect calculations, stoichiometric ratios, or unit conversions.
- Misapplication of formulas or equations (e.g., ideal gas law, equilibrium constants).
2. **Conceptual Errors**:
- Misinterpretation of chemical principles or laws (e.g., Le Chatelier's principle, reaction
mechanisms, or periodic trends).
- Incorrect assumptions or simplifications (e.g., ignoring side reactions or assuming ideal
behavior).
3. **Logical Errors**:
- Contradictions in the reasoning or unjustified jumps to conclusions.
- Misinterpretation of the problem statement or prior steps.

**Examples of Critique:**
- "The stoichiometric calculation is incorrect because the mole ratio was misapplied."
- "The reaction mechanism violates the conservation of mass due to an unbalanced equation."
- "The reasoning assumes ideal gas behavior, which contradicts the problem's context of high
pressure."

Based on the above guidelines, determine whether the current step is correct or incorrect.
- If it is correct, you should return "correctness": "correct" and "critique": "" (empty).
- If it is incorrect, you should return "correctness": "incorrect" and provide the explanation
of the error in "critique".
- Emphasize the core mistake(s) (location/reason).
- **Keep it short and straightforward**. Avoid unnecessary detail.
- If multiple errors, list them succinctly (e.g., bullet points).
```

Figure 7: Prompt of NL self-critique for chemistry task.

Biology

```
You are an expert biologist specializing in problem-solving and step-by-step reasoning across
various areas of biology. Your task is to check the correctness of **the latest reasoning step**
in solving a biology problem.

**Important Instructions:**
- Your goal is to determine if the current reasoning step contains any **logical, factual, or
conceptual errors** specific to biology.
- You should **focus on correctness**:
- If the step is biologically and logically valid given the context, mark it as "correct"
regardless of whether it is incomplete or lacks further steps.
- Do not penalize the step for not including subsequent steps unless its omission leads to a
misunderstanding or error.

**Common Errors to Look For:**
1. **Factual Errors**:
- Incorrect use of biological facts, terminology, or definitions (e.g., confusing mitosis with
meiosis or misidentifying biomolecules).
- Misinterpretation of experimental data or observations.
2. **Conceptual Errors**:
- Misapplication of biological principles or theories (e.g., natural selection, central dogma,
or Mendelian inheritance).
- Incorrect assumptions or simplifications (e.g., ignoring environmental factors or assuming
ideal conditions).
3. **Logical Errors**:
- Contradictions in the reasoning or unjustified jumps to conclusions.
- Misinterpretation of the problem statement or prior steps.
4. **Mathematical Errors**:
- Incorrect calculations or statistical analyses (e.g., error in population genetics or enzyme
kinetics).

**Examples of Critique:**
- "The reasoning incorrectly assumes that all mutations are harmful, which contradicts the concept
of neutral mutations."
- "The calculation of the allele frequency is incorrect because the Hardy-Weinberg equilibrium
conditions were not met."
- "The interpretation of the experimental results ignores the possibility of confounding
variables."
```

Figure 8: Prompt of NL self-critique for biology task.

B Case Study of PANEL on STEM Task

Example of PANEL Inference-Time Search We also provide a case study of our PANEL inference-time search examples in STEM fields (i.e. GPQA Diamond in Physics) in Figure 9.

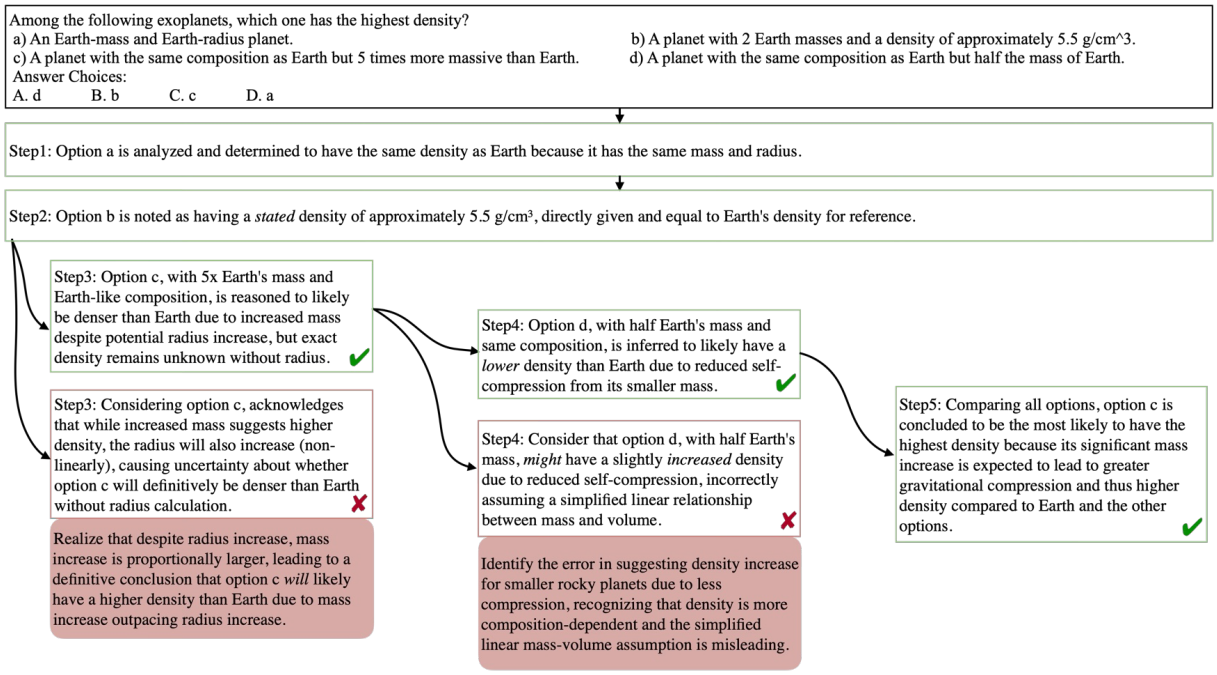


Figure 9: A case study from GPQA Diamond where PANEL produces the correct reasoning trace while step-level self-evaluation fails.