

VerifiAgent: a Unified Verification Agent in Language Model Reasoning

Jiuzhou Han[†] Wray Buntine[‡] Ehsan Shareghi[†]

[†] Department of Data Science & AI, Monash University

[‡] College of Engineering and Computer Science, VinUniversity

jiuzhou.han@monash.edu wray.b@vinuni.edu.vn

ehsan.shareghi@monash.edu

Abstract

Large language models demonstrate remarkable reasoning capabilities but often produce unreliable or incorrect responses. Existing verification methods are typically model-specific or domain-restricted, requiring significant computational resources and lacking scalability across diverse reasoning tasks. To address these limitations, we propose VerifiAgent, a unified verification agent that integrates two levels of verification: meta-verification, which assesses completeness and consistency in model responses, and tool-based adaptive verification, where VerifiAgent autonomously selects appropriate verification tools based on the reasoning type, including mathematical, logical, or common-sense reasoning. This adaptive approach ensures both efficiency and robustness across different verification scenarios. Experimental results show that VerifiAgent outperforms baseline verification methods (e.g., deductive verifier, backward verifier) among all reasoning tasks. Additionally, it can further enhance reasoning accuracy by leveraging feedback from verification results. VerifiAgent can also be effectively applied to inference scaling, achieving better results with fewer generated samples and costs compared to existing process reward models in the mathematical reasoning domain.¹

1 Introduction

Large language models (LLMs) have demonstrated significant capabilities in natural language reasoning tasks, exhibiting potential to solve complex problems across diverse domains (Yang et al., 2024a; DeepSeek-AI et al., 2025; Dubey et al., 2024; OpenAI, 2023). However, despite their advanced reasoning abilities, these models often produce responses that are unreliable or incorrect, which poses substantial challenges for practical applications that require high precision and trustworthiness (Augenstein et al., 2024; Huang et al.,

2024). To address this critical issue, several verification methods have been proposed, ranging from task-specific verifiers to generalised verification methods leveraging prompting techniques.

Training a task-specific verifier to verify the output of LLM requires specific training data. For instance, Ni et al. (2023) train a verification model that judges language-to-code outputs based on both program text and execution results. Liang et al. (2024) design verifiers trained on outputs from multiple reasoning paradigms, using correctness signals for improvement. Han et al. (2024b) train a lightweight verifier for improving semantic graph generation in text-to-graph tasks. Similarly, Thatikonda et al. (2024) train a verifier that corrects potential syntactic and semantic first-order logic translation errors. Nevertheless, these existing solutions typically face limitations such as domain restriction, computational inefficiency, and lack of scalability when handling varied reasoning tasks. Table 1 demonstrates a feature comparison of various verification methods.

In this paper, we propose VerifiAgent, a unified verification agent designed specifically to overcome these limitations by offering a generalisable and efficient verification framework. Unlike prior methods, VerifiAgent adopts a two-layer verification mechanism, comprising meta-verification and tool-based adaptive verification. The meta-verification layer ensures completeness and logical consistency of responses, while the tool-based adaptive verification autonomously selects appropriate external tools (e.g., Python interpreters, symbolic solvers, search engines) to deal with different reasoning types, including mathematical, logical, commonsense, and hybrid reasoning tasks.

Our approach not only achieves superior verification accuracy compared to existing baseline methods, such as deductive verifier (Ling et al., 2023), backward verifier (Weng et al., 2023), but also enhances reasoning accuracy by integrating

¹Code is available at <https://github.com/Jiuzhouh/VerifiAgent>.

Method	Training-free	Generalised	Tool-based	Fine-grained Feedback
DIVERSE (Li et al., 2023)	×	×	×	×
PiVe (Han et al., 2024b)	×	×	×	✓
Math/Code-Rev (Liang et al., 2024)	×	×	✓	×
LEVER (Ni et al., 2023)	×	×	✓	×
CoVe (Dhuliawala et al., 2024)	✓	×	×	✓
CSV(Zhou et al., 2024)	✓	×	✓	✓
Deductive Verifier (Ling et al., 2023)	✓	✓	×	×
Backward Verifier (Weng et al., 2023)	✓	✓	×	×
VerifiAgent (ours)	✓	✓	✓	✓

Table 1: Comparison of various verification methods.

detailed feedback derived from the verification process. Furthermore, VerifiAgent can be effectively applied to inference scaling, requiring significantly fewer computational resources compared to standard Process Reward Models (PRMs), thereby providing a practical approach to improve LLM performance during inference. Through extensive experiments across three types of reasoning tasks, we summarise two key empirical findings: 1) An LLM reasoner can improve via inference scaling methods like Majority Vote, PRMs, or VerifiAgent, but VerifiAgent achieves higher accuracy at lower cost. 2) VerifiAgent’s capabilities scale alongside improvements in its backbone LLM, enabling consistent performance gains on the same reasoner.

2 Related Work

2.1 LLMs as Verifiers

Leveraging the prompting and in-context learning ability of LLMs to verify the outputs of LLMs provides a generalised approach of verification. Wu et al. (2024); Weng et al. (2023) show that LLMs can refine reasoning chains via backward verification or masked condition checking, while Ling et al. (2023) decompose solutions into verifiable steps using a Natural Program format. Dhuliawala et al. (2024) propose Chain-of-Verification, which decomposes the verification into a sequence of questions, improving factual consistency through multi-step prompting. Hong et al. (2024) evaluate LLMs’ ability to detect logical errors, finding that while models can catch some flaws, their verification is often shallow. Stechly et al. (2024) further investigate the reliability of self-critique, demonstrating that performance often degrades when doing self-verification without external grounding.

To enhance the verification quality of LLMs, some methods integrate external tools. Zhou et al. (2024) introduce code-based self-verification, which prompts GPT-4 Code Interpreter to evaluate and fix its answers by executing code and interpreting the output. Similarly, Gou et al. (2024) propose CRITIC, a framework where LLMs inter-

act with tools (e.g., calculators, search engines) to critique and revise their own outputs, leading to improved factuality and reasoning. Different from these works, our VerifiAgent provides a generalised verification agentic framework adaptable to diverse reasoning tasks with fine-grained feedback.

A broader concept of verification by LLMs is referred to LLM-as-a-Judge, where LLMs are used as general evaluators for tasks like response scoring, pairwise comparison, and content moderation. As surveyed by Gu et al. (2024); Li et al. (2024), LLM-as-a-Judge systems perform holistic evaluations, as a scalable and consistent alternative to human evaluation, applicable in model benchmarking, safety assessment, and alignment data labelling. VerifiAgent can be viewed as a specialised type of LLM-as-a-Judge system, specifically designed to evaluate the correctness of certain reasoning tasks through tool-based verification mechanisms.

2.2 Scaling Test-Time Compute

Scaling test-time compute refers to allocating more computational budget during inference via sampling, deeper reasoning, or adaptive search to boost model accuracy. Brown et al. (2024) present a comprehensive study of inference-time scaling through repeated sampling, demonstrating that coverage—the probability of generating at least one correct answer—scales log-linearly with the number of samples. Stroebl et al. (2024) theoretically analyse the limits of resampling, showing that imperfect verifiers lead to diminishing returns, especially when false positives dominate. Similarly, Setlur et al. (2025a) argue that verifier-based strategies scale more robustly than verifier-free ones, particularly when base models exhibit anti-concentrated output distributions.

The growing use of Process Reward Models (PRMs) suggests that fine-grained supervision over intermediate reasoning steps can improve model reliability (Lightman et al., 2024; Wang et al., 2024; Zhang et al., 2025). PRMs offer another strategy for scaling test-time compute by enhancing Best-of-N sampling (Snell et al., 2024). As a verifier, VerifiAgent does not require any training process, eliminating the need for collecting task-specific training data. By leveraging frozen LLMs, it can be integrated into test-time compute scaling strategies, enhancing the accuracy of LLM outputs.

3 VerfiAgent

VerfiAgent is a plug-and-play verification framework that empowers frozen LLMs to utilise external mechanisms to verify the correctness of solutions of diverse reasoning tasks. As illustrated in Figure 1, VerfiAgent adopts a two-layer verification mechanism, which contains two levels of verification. The first level is to do a Meta Verification, and the second level is to do a Tool-based Adaptive Verification. The solution will be evaluated sequentially through the two layers. The second-level verification can further validate the results from meta-verification, enhancing the accuracy of the verification results. With this two-layer verification mechanism, the VerfiAgent can provide fine-grained feedback of the verification process.

3.1 Meta Verification

The Meta Verification aims to verify two aspects of the solution: completeness and consistency. The completeness refers to a solution that is self-contained, fully addresses every part of the question, and contains a clear result or conclusion. The consistency refers to reasoning that follows a logical structure with no jumps, gaps, or inconsistencies. This initial layer acts as a foundational check, preventing incomplete or inconsistent solutions from progressing further. Through meta verification, VerfiAgent ensures that only solutions with structural integrity and coherent reasoning proceed to the next tool-based adaptive verification.

Since the solutions of different types of reasoning tasks may have different structures, to make VerfiAgent adaptable to diverse solutions, we leverage a unified way to rewrite the solutions in the meta verification phase. Specifically, the agent will first list all the known conditions and the final objective provided in the problem, then divide the solution into individual and explicit logical steps. This will be beneficial for the meta verification and the following tool-based adaptive verification. See Appendix E for examples.

3.2 Tool-based Adaptive Verification

After the meta verification stage, the solution enters the Tool-based Adaptive Verification phase. This level leverages external tools, such as Python program interpreter, search engine and symbolic solver, to cross-check the correctness of the solution. The agent will first solve the question using appropriate tools, and then verify the results by

comparing them with the original solution. Unlike the meta verification stage, which evaluates general reasoning quality, this phase evaluates factual and computational accuracy.

VerfiAgent dynamically selects the most suitable verification tool based on the nature of the task and instructions. For instance, in mathematical reasoning, it may utilise a Python interpreter to verify calculations, while for knowledge-based common-sense reasoning, it may query a search engine to gather relevant information. For hybrid reasoning tasks, it can combine multiple tools to ensure comprehensive verification. Additionally, VerfiAgent autonomously determines the required number of external tool calls, continuing until it gathers sufficient information to validate the answer.

When VerfiAgent selects a tool for verification, the environment returns the corresponding execution result. Based on this observation, the agent iteratively determines its next action until the verification process is complete. The VerfiAgent not only ensures the accuracy of solutions but also provides a transparent and interpretable verification process for natural language reasoning tasks. See Appendix E for examples.

3.3 Fine-grained Feedback

Based on the two levels of verification, VerfiAgent provides a final evaluation result (i.e., *Correct/Incorrect*) to indicate the correctness of the solution. In addition to the verification result, VerfiAgent also generate a V_{score} as a confidence score of the verification. V_{score} is calculated by applying the softmax function to the log probability of the token (*Correct/Incorrect*) and the log probabilities of the top 5 alternative tokens. Specifically:

$$V_{score} = \frac{\exp(p(t))}{\sum_{k=1}^5 \exp(p(t_k))}$$

where V_{score} represents the confidence score for the verification result token t . the term $p(t)$ denotes the log probability of the token t generated by the LLM. $p(t_k)$ is for $k = 1$ to 5 represents the log probabilities of the top five predicted tokens at the verification result token position. This equation ensures that the confidence score reflects the relative likelihood of the chosen token compared to the top alternatives, effectively normalising the scores within the range of 0 to 1.

Furthermore, when a solution is deemed incorrect, VerfiAgent provides fine-grained feedback

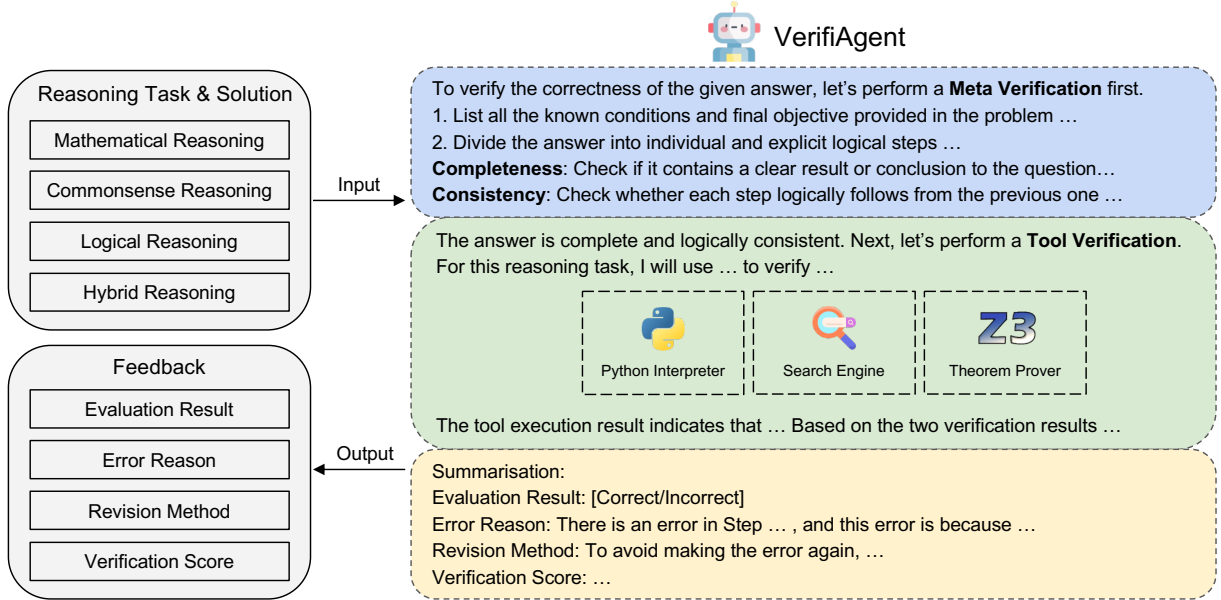


Figure 1: An overview of VerifiAgent. Given a reasoning task and a candidate solution, VerifiAgent leverages two levels of verification: (1) meta verification – verifying the completeness and consistency of the solution and (2) tool-based adaptive verification – autonomously selecting appropriate tools to do the correctness verification. The VerifiAgent can provide fine-grained feedback about the verification process based on the instruction in the prompt.

about the verification process based on the instruction in the prompt. This feedback includes the identified error reason, derived from the two levels of verification, and a potential revision method that incorporates observations from tool execution results (See Appendix E for examples). Such feedback can be leveraged to refine and enhance the solution, improving the accuracy of reasoning tasks.

4 Experiment

4.1 Baseline and Experimental Setup

Datasets. We evaluate VerifiAgent on three natural language reasoning tasks, including mathematical reasoning, logical reasoning, commonsense reasoning, and hybrid reasoning. Specifically, for mathematical reasoning we use GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), for logical reasoning we use FOLIO (Han et al., 2024c) and ProverQA (Qi et al., 2025), for commonsense reasoning we use HotpotQA (Yang et al., 2018) and StrategyQA (Geva et al., 2021), for hybrid reasoning we use ReWild (Yang et al., 2024c). The statistics of the datasets are shown in Appendix A.

Baselines. Since VerifiAgent is a training-free and generalised approach, we compare it against baseline methods that are similarly prompting-based and generalised. Specifically:

- **Vanilla Verifier.** Vanilla Verifier employs a structured prompt to instruct the LLM to verify a solution given a problem, without relying on specialised mechanisms (Kamoi et al., 2024).

- **Deductive Verifier.** Deductive Verifier (Ling et al., 2023) enables the LLM to carry out explicit and rigorous deductive reasoning to evaluate the correctness of a solution. It decomposes the verification process into a sequence of step-by-step subprocesses using Natural Program, a natural language-based deductive reasoning format, to facilitate the breakdown of logical steps in a step-by-step manner.
- **Backward Verifier.** Backward Verifier (Weng et al., 2023) appends the predicted answer to the question while masking the original condition, then prompts the LLM to predict the masked condition. Verification is conducted by comparing the predicted condition with the original one. If the two conditions align, the solution is deemed correct; otherwise, inconsistencies indicate errors in the provided solution.

Models. We explored various combinations of backbone LLMs for both the Reasoner and VerifiAgent. For the Reasoner, we utilise GPT-4o, o3-mini, and Llama-3.3-70B-Instruct-Turbo, while for the VerifiAgent, we employ GPT-4o and o1-mini. In our experiments, unless explicitly stated otherwise, both the Reasoner and VerifiAgent default to GPT-4o as their backbone LLMs.

4.2 Main Result

Table 2 shows the performance of VerifiAgent compared to baseline methods (Vanilla, Deductive, and Backward Verifiers) across different reasoning tasks. Overall, VerifiAgent consistently out-

Type	Dataset	Baselines									VerifiAgent		
		Vanilla Verifier			Deductive Verifier			Backward Verifier			Acc	Pre	Rec
		Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec			
Mathematical	GSM8K	0.93	0.96	0.96	0.95	0.96	0.99	0.95	0.96	0.98	0.96	0.96	1.00
	MATH	0.75	0.73	0.86	0.80	0.76	0.86	0.82	0.80	0.88	0.85	0.86	0.92
Logical	FOLIO	0.75	0.78	0.96	0.73	0.73	0.95	0.74	0.76	0.96	0.76	0.78	0.97
	ProverQA	0.75	0.77	0.97	0.74	0.75	0.98	0.75	0.78	0.96	0.77	0.82	0.95
Commonsense	StrategyQA	0.78	0.79	0.92	0.75	0.82	0.92	0.79	0.80	0.94	0.84	0.85	0.95
	HotpotQA	0.56	0.53	0.91	0.56	0.53	0.96	0.57	0.54	0.90	0.61	0.56	0.92
Hybrid	ReWild	0.76	0.88	0.82	0.61	0.91	0.60	0.74	0.87	0.84	0.78	0.88	0.89

Table 2: Main results of VerifiAgent on different reasoning tasks. The evaluation metrics are accuracy (Acc), precision (Pre), and recall (Rec). **Bold** shows the best result for each row.

performs baselines, excelling in accuracy while maintaining competitive precision and recall across mathematical, logical, commonsense, and hybrid reasoning tasks. Specifically, for mathematical reasoning tasks, VerifiAgent attains the highest accuracy (0.96 and 0.85) and recall scores (1.00 and 0.92) on GSM8K and MATH datasets, respectively. In logical reasoning, VerifiAgent demonstrates improvements, particularly on FOLIO (accuracy 0.76, recall 0.97) and ProverQA (precision 0.82). For commonsense reasoning tasks, VerifiAgent significantly outperforms baselines on StrategyQA with accuracy and precision of 0.84 and 0.85, respectively, while remaining competitive on HotpotQA. Finally, on the hybrid reasoning dataset ReWild, VerifiAgent achieves the best accuracy (0.78) and recall (0.89), highlighting its verification capabilities in handling complex reasoning tasks. To investigate the impact of different backbone LLMs on VerifiAgent’s performance, we further evaluate VerifiAgent using o1-mini as an alternative backbone model. The results indicate that the verification capability of VerifiAgent scales effectively with the underlying backbone model’s capacity. Due to the page limit, we put the detailed results in Appendix C.

4.3 Inference Scaling with VerifiAgent

Inference scaling aims at enhancing reasoning performance by utilising increased computational resources during the inference stage. However, this approach inherently requires effective verification to ensure the accuracy and reliability of generated answers (Setlur et al., 2025b). Due to the verification ability of VerifiAgent, it naturally complements inference scaling approaches by serving as an effective verifier during the inference process. Specifically, we first sample an output from the LLM. If this output passes verification by the Veri-

fiAgent, the process terminates; otherwise, we continue sampling additional candidate outputs until one passes verification or the maximum number of samples is reached. For cases reaching the maximum number of samples, we select the final answer using a majority vote approach. We compare our VerifiAgent-based inference scaling method with the standard Majority Vote approach that does not employ a verifier. Majority Vote aggregates multiple sampled responses directly from reasoners without any verification. Table 3 demonstrates the performance across three reasoning datasets (MATH, ProverQA, and StrategyQA) using various combinations of reasoners (GPT-4o, o3-mini, and Llama-3.3-70B-Instruct-Turbo) and VerifiAgent variants (GPT-4o and o1-mini).

Across all datasets and reasoners, inference scaling with VerifiAgent consistently outperforms Majority Voting, achieving higher accuracy with fewer samples and less cost (See Appendix D). Notably, o3-mini reasoner achieves the highest performance on MATH and ProverQA but the lowest on StrategyQA among all reasoners, suggesting that o3-mini is more proficient in mathematical and logical reasoning than in knowledge-intensive commonsense reasoning tasks. We identify two key findings: (1) When the reasoner and VerifiAgent are the same model (e.g., GPT-4o) or have comparable capacities (e.g., Llama-3.3-70B-Instruct-Turbo paired with GPT-4o), integrating VerifiAgent significantly enhances performance, with further improvement achievable by employing a stronger VerifiAgent (e.g., o1-mini). (2) When the reasoner (o3-mini) surpasses the VerifiAgent (GPT-4o) in capability, the performance gain is limited. However, pairing a strong reasoner with a stronger VerifiAgent (o1-mini) substantially enhances performance.

PRMs provide another approach to inference scaling. We investigated two open-source PRMs

Method	MATH	ProverQA	StrategyQA
GPT-4o Reasoner	69.4(1)	75.3(1)	84.2(1)
- IS w/ Majority Vote @10	73.5(10)	77.0(10)	85.6(10)
- IS w/ VerifiAgent (GPT-4o)	74.0(1.5)	77.3(1.6)	86.0(1.3)
- IS w/ VerifiAgent (o1-mini)	78.0(1.8)	77.7(1.3)	87.3(1.2)
o3-mini Reasoner	87.9(1)	78.3(1)	76.4(1)
- IS w/ Majority Vote @8	91.1(10)	80.0(10)	78.2(10)
- IS w/ VerifiAgent (GPT-4o)	88.3(1.3)	79.1(1.1)	78.6(1.3)
- IS w/ VerifiAgent (o1-mini)	91.4(1.1)	80.7(1.1)	79.0(1.6)
Llama-3.3-70B-Instruct-Turbo Reasoner	62.3(1)	70.6(1)	83.8(1)
- IS w/ with Majority Vote @10	68.3(10)	71.7(10)	84.7(10)
- IS w/ VerifiAgent (GPT-4o)	69.7(2.0)	72.0(1.3)	85.1(1.3)
- IS w/ VerifiAgent (o1-mini)	71.1(2.2)	74.0(1.3)	85.1(1.4)

Table 3: Results of different Reasoners with Inference Scaling (IS) methods on three datasets. The number in the bracket denotes the average number of samples for each question.

specifically designed for the MATH dataset: Qwen2.5-Math-PRM-7B and Qwen2.5-Math-7B-PRM800K, which fine-tune Qwen2.5-Math-7B-Instruct using synthetic data from Qwen models (Zhang et al., 2025) and PRM800K (Lightman et al., 2024), respectively. These PRMs assign scores to each reasoning step, and we use the last step score as the final response score. Following previous studies (Zhang et al., 2025; Lightman et al., 2024; Yang et al., 2024b; Wang et al., 2024), we evaluate the PRMs using the Best-of-N sampling strategy, selecting the highest-scored response from N candidates according to a PRM. The evaluation results for the GPT-4o and Qwen2.5-Math-7B-Instruct reasoners are shown in Figure 2. As the number of samples increases, both Majority Vote and Best-of-N sampling strategies consistently improve in accuracy. When sampling 10 responses, the Best-of-N method’s accuracy approaches that of VerifiAgent, which notably achieves comparable performance with significantly fewer average samples (1.5 and 1.6 on GPT-4o reasoner and Qwen2.5-Math-7B-Instruct reasoner, respectively).

Interestingly, the two PRMs exhibit distinct behaviours depending on the reasoner. For the GPT-4o reasoner, Qwen2.5-Math-7B-PRM800K significantly outperforms Qwen2.5-Math-PRM-7B, which even underperforms relative to the Majority Vote baseline. However, for the Qwen2.5-Math-7B-Instruct reasoner, Qwen2.5-Math-PRM-7B outperforms Qwen2.5-Math-7B-PRM800K at 10 samples. These results indicate that GPT-4o benefits more from Qwen2.5-Math-7B-PRM800K, whereas

	MATH	ProverQA	StrategyQA
Init. Reasoning Acc.	69.4	75.3	84.3
Feedback Type	Precaution-Based Feedback		
Verification Result	69.7	76.0	84.3
+ Error Reason	74.9	77.0	85.6
+ Mitigation Method	73.4	77.6	86.0
Feedback Type	Post-Editing-Based Feedback		
Verification Result	71.7	77.3	84.7
+ Error Reason	72.3	74.7	84.3
+ Mitigation Method	72.6	74.3	83.8

Table 4: Results of feedback utilisation on GPT-4o.

Qwen2.5-Math-7B-Instruct gains greater improvements from Qwen2.5-Math-PRM-7B. We hypothesise that the linguistic discrepancies may affect the performance of PRMs. Specifically, Qwen2.5-Math-7B-PRM800K utilises synthetic data from GPT-style LLMs for training, while Qwen2.5-Math-PRM-7B employs data generated by Qwen-style LLMs for training.

4.4 Exploration on Feedback Utilisation

VerifiAgent provides fine-grained feedback during verification, which includes an explicit error reason and a suggested revision method for enhancing solutions. To evaluate the effectiveness of this feedback, we conducted experiments using two distinct methods: precaution-based and post-editing-based feedback. In the precaution-based method, the LLM leverages feedback from previous verification attempts to proactively generate a new solution. Conversely, the post-editing-based method allows

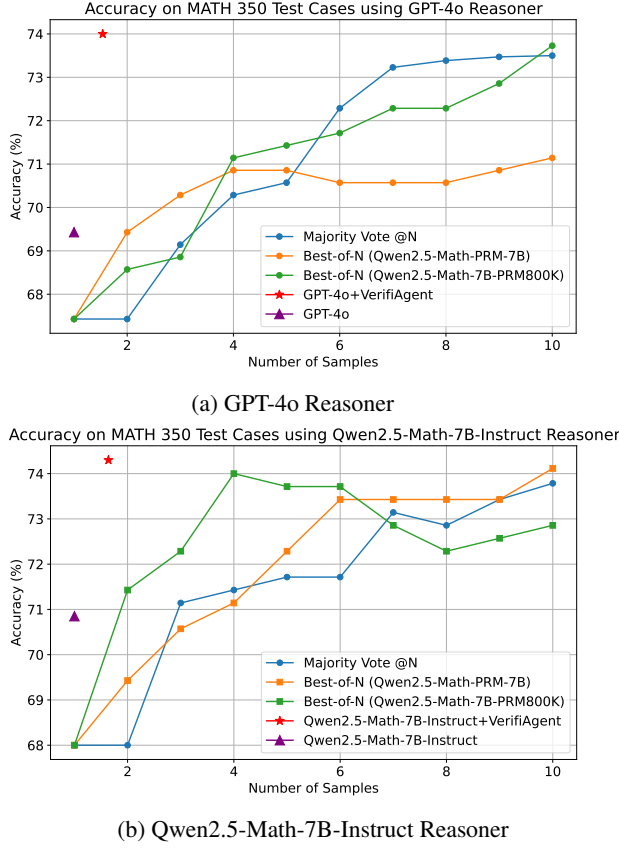


Figure 2: Results of GPT-4o Reasoner and Qwen2.5-Math-7B-Instruct Reasoner with different inference scaling methods on MATH. VerifiAgent uses GPT-4o as the backbone LLM.

the LLM to directly refine its previous incorrect solution based on feedback provided.

We explored three feedback settings for each method: (1) verification result only (i.e., simply indicating “Incorrect”), (2) verification result with error reason, and (3) verification result with both error reason and revision method. Experiments were conducted on instances initially identified as “Incorrect” by VerifiAgent, and the results are presented in Table 4.

Overall, precaution-based feedback consistently outperforms post-editing-based feedback, indicating the inherent difficulty for LLMs to effectively correct previously incorrect responses. Additionally, within precaution-based feedback, providing richer information typically yields greater improvements. For post-editing-based feedback, however, mathematical reasoning tasks benefit from more detailed feedback, whereas logical and commonsense reasoning tasks achieve better performance with simpler, less detailed feedback.

Method	MATH	ProverQA	StrategyQA
GPT-4o (Tool-use)	56.9	47.8	83.5
GPT-4o (Tool-use) + Feedback	61.5	50.3	85.7
GPT-4o (CoT)	69.4	75.3	84.3
GPT-4o (CoT) + Feedback	73.4	77.6	86.0

Table 5: Results of GPT-4o reasoner using CoT and Tool with feedback on three different reasoning tasks.

Dataset (Method)	Acc	Pre	Rec
MATH (CoT)	0.85	0.86	0.92
MATH (Tool-use)	0.82	0.83	0.90
ProverQA (CoT)	0.77	0.82	0.95
ProverQA (Tool-use)	0.75	0.79	0.93
StrategyQA (CoT)	0.84	0.85	0.95
StrategyQA (Tool-use)	0.80	0.84	0.92

Table 6: Results of VerifiAgent evaluating CoT and Tool-use outputs on different reasoning tasks.

4.5 VerifiAgent on Tool-based Reasoner

To evaluate the effectiveness of VerifiAgent on tool-based reasoning tasks, we conducted experiments using a tool-using reasoner with access to the same tools as VerifiAgent. Specifically, we evaluated GPT-4o on MATH, ProverQA, and StrategyQA, where the model was instructed to use the Python interpreter, Z3 Theorem Prover, and Search Engine, respectively. We then applied VerifiAgent to these outputs, leveraging its feedback to further improve performance. The results are shown in Table 5.

Interestingly, the tool-use baseline does not outperform CoT, especially in math and logic reasoning. This trend is also observed in prior works (Yao et al., 2023; Han et al., 2024a), where tool-augmented methods such as ReAct can underperform compared to CoT. We hypothesise this is due to several factors: (1) For simpler problems, LLMs may already solve them accurately via CoT, and mandatory tool use may introduce unnecessary complexity and more opportunities for errors. (2) For more difficult tasks, the LLM may still struggle to solve them effectively, even with tools. (3) The external knowledge from inaccurate tool-use can sometimes mislead the LLM’s correct prior knowledge. Despite these challenges, integrating VerifiAgent feedback consistently improves tool-use accuracy across all datasets, though results still lag behind CoT+Feedback in math and logic. For commonsense tasks (StrategyQA), tool-use approaches CoT+Feedback performance.

We also measured VerifiAgent’s verification per-

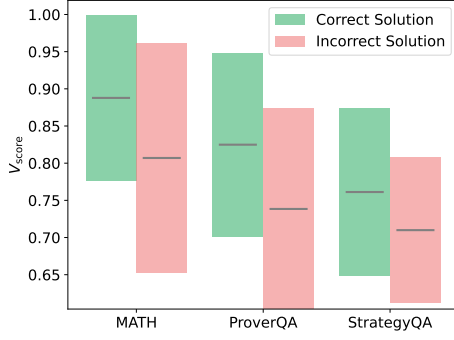


Figure 3: The floating bar chart comparing V_{score} distributions (mean \pm std) for correct and incorrect solutions across three datasets. The horizontal grey line indicates the mean.

Method	MATH			ProverQA			StrategyQA		
	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
Vanilla Verifier	0.75	0.73	0.86	0.75	0.77	0.97	0.78	0.79	0.92
Deductive Verifier	0.80	0.76	0.86	0.74	0.75	0.98	0.75	0.82	0.92
Backward Verifier	0.82	0.80	0.88	0.75	0.78	0.96	0.79	0.80	0.94
VerifiAgent	0.85	0.86	0.92	0.77	0.82	0.95	0.84	0.85	0.95
- w/o meta v.	0.79	0.78	0.96	0.74	0.81	0.90	0.83	0.83	0.94
- w/o tool v.	0.75	0.75	0.98	0.74	0.75	0.98	0.78	0.80	0.96

Table 7: Ablation study results of VerifiAgent.

formance on both CoT and tool-use outputs, shown in Table 6. While performance is slightly lower for tool-use outputs, which is expected, since VerifiAgent is not designed specifically for program or tool-use evaluation. This points an important direction for future work, enhancing VerifiAgent’s robustness in verifying tool-based reasoning.

4.6 Ablation Study

Meta verification and tool verification are two essential components of VerifiAgent. To evaluate the individual contributions of these components, we conducted an ablation study, with results presented in Table 7. Results demonstrate that removing either meta verification or tool verification consistently reduces VerifiAgent’s performance across all datasets. Specifically, omitting meta verification leads to noticeable declines in overall accuracy, while removing tool verification results in even more substantial performance reductions, bringing the performance close to baseline levels. Additionally, tool verification tends to enhance accuracy and precision, whereas meta verification primarily improves recall. These findings underscore the complementary roles of meta and tool verification, with each contributing uniquely to the effectiveness of VerifiAgent.

5 Analysis

Verification Score Visualisation. Figure 3 visualises the V_{score} for correct and incorrect solutions

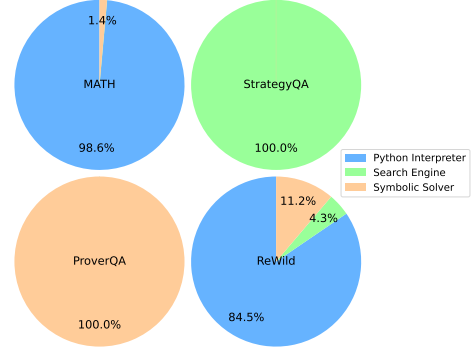


Figure 4: The pie charts showing the relative usage frequency of three different tools by the VerifiAgent across four types of reasoning tasks.

across three datasets. As illustrated, the mean V_{score} for correct solutions is slightly higher than for incorrect ones on all the reasoning tasks. Since V_{score} represents the confidence of the verification result, this indicates that the VerifiAgent is more confident when identifying correct solutions compared to incorrect ones. Additionally, the consistently lower variance in V_{score} among correct solutions further supports the reliability of the agent in verifying correct responses.

Tool Usage Analysis. VerifiAgent autonomously determines the reasoning type of a task and selects the appropriate tool for verification. Figure 4 illustrates tool usage across four reasoning task types. For the MATH dataset (mathematical), the Python Interpreter is predominantly used (98.6%), with minimal reliance on the Symbolic Solver (1.4%), reflecting the computational nature of the task. StrategyQA (commonsense) exclusively relies on the Search Engine (100%), highlighting its dependence on external knowledge for the verification. ProverQA (logical) solely utilises the Symbolic Solver (100%), aligning with its need for logical and symbolic reasoning. ReWild (hybrid) shows a more balanced tool distribution, primarily using the Python Interpreter (84.5%), supplemented by the Symbolic Solver (11.2%) and the Search Engine (4.3%). The results demonstrate that the VerifiAgent effectively selects appropriate external tools based on the nature of the reasoning task.

Error Analysis. To further investigate the capability of VerifiAgent, we conducted an error analysis on different types of questions. The MATH dataset contains seven types of math problems: Algebra (Alg), Counting&Probability (Count&Prob), Geometry (Geo), Intermediate Algebra (Int Alg), Number Theory (Num Thr), Prealgebra (PreAlg) and Precalculus (PreCal). The ProverQA classifies

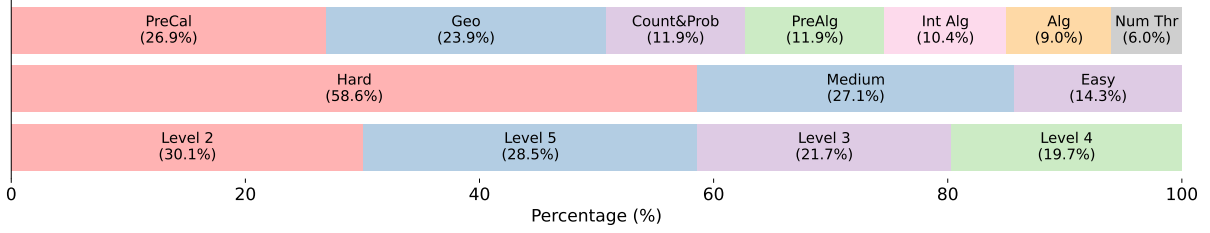


Figure 5: The proportion of different question types among VerifiAgent’s incorrectly verified examples by GPT-4o Reasoner. From top to bottom, the bars represent MATH, ProverQA, and StrategyQA datasets, respectively. For MATH and ProverQA, the number of questions in each type is the same. For the imbalanced StrategyQA, the proportion is normalised by the total number of questions per difficulty level.

the question into three types based on the difficulty level: Hard, Medium and Easy. Although StrategyQA does not explicitly label questions by difficulty, each question includes a decomposition into sub-questions that reflect its reasoning pathway. We used the number of decomposed sub-questions (ranging from 1 to 5) as an indicator of question difficulty, classifying them into five levels (Level 1 through Level 5).

Figure 5 illustrates the distribution of question types among cases where VerifiAgent provided incorrect verifications. In the MATH dataset, Pre-calculus and Geometry questions accounted for the highest proportion of errors, suggesting these question types pose greater verification challenges for VerifiAgent. Errors in Counting & Probability, Prealgebra, Intermediate Algebra, and Algebra occurred at similar rates, while VerifiAgent performed best on Number Theory problems. This trend is in line with the capability of the backbone LLMs. See Appendix B for the error distributions on each type of question.

For ProverQA, VerifiAgent’s verification accuracy correlated clearly with question difficulty, making the highest number of errors on Hard questions and the fewest on Easy questions. Conversely, no clear error pattern emerged for StrategyQA. Since verification relies mainly on search engines to retrieve factual knowledge, VerifiAgent appears capable of accessing sufficient information irrespective of question difficulty, indicating that the complexity of questions in StrategyQA has minimal impact on verification performance.

6 Conclusion

In this paper, we introduced VerifiAgent, a unified verification agent that verifies and improves outputs from LLMs across mathematical, logical, commonsense, and hybrid reasoning tasks. Ver-

ifiAgent employs a two-layer verification framework combining meta-verification, which assesses completeness and consistency, and adaptive tool-based verification tailored to each reasoning type. Experimental results demonstrate that VerifiAgent consistently outperforms baseline methods in verification accuracy. Additionally, VerifiAgent can be integrated with inference scaling approaches, achieving improved performance with fewer samples than PRMs. Overall, VerifiAgent provides an efficient and scalable solution, enhancing the reliability and trustworthiness of large language model reasoning.

VerifiAgent heavily relies on the instruction-following capabilities of the backbone LLM, meaning that only models proficient at accurately interpreting and executing instructions can serve effectively as the backbone. This reliance indicates the importance of selecting suitable backbone LLMs to ensure optimal performance. VerifiAgent currently supports only three verification tools (Python interpreter, search engine, and symbolic solver). Expanding its capabilities by integrating additional verification tools could further enhance VerifiAgent’s adaptability and effectiveness across a broader range of reasoning scenarios.

Limitations

Our study has two primary limitations. First, due to the significant computational costs associated with accessing and running LLMs, we were unable to evaluate a comprehensive range of models, necessitating a selection of representative models. Second, the current implementation of VerifiAgent only supports three tools. While these demonstrate its core capabilities, we plan to expand the toolset in future work to enhance its versatility and applicability across more verification scenarios.

References

- Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. [Factuality challenges in the era of large language models and opportunities for fact-checking](#). *Nat. Mac. Intell.*, 6(8):852–863.
- Bradley C. A. 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](#). *CoRR*, abs/2407.21787.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 81 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. [Chain-of-verification reduces hallucination in large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 3563–3578. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies](#). *Trans. Assoc. Comput. Linguistics*, 9:346–361.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujia 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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. [A survey on llm-as-a-judge](#). *CoRR*, abs/2411.15594.
- Jiuzhou Han, Wray L. Buntine, and Ehsan Shareghi. 2024a. [Towards uncertainty-aware language agent](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6662–6685. Association for Computational Linguistics.
- Jiuzhou Han, Nigel Collier, Wray L. Buntine, and Ehsan Shareghi. 2024b. [Pive: Prompting with iterative verification improving graph-based generative capability of llms](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6702–6718. Association for Computational Linguistics.
- Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhen-ting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, and 16 others. 2024c. [FOLIO: natural language reasoning with first-order logic](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 22017–22031. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. [A closer look at the self-verification abilities of large language models in logical reasoning](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 900–925. Association for Computational Linguistics.
- Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, André Freitas, and Mustafa A. Mustafa. 2024. [A survey of safety and trustworthiness of large language models through the lens of verification and validation](#). *Artif. Intell. Rev.*, 57(7):175.
- Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. [When can llms Actually correct their own mistakes? A critical survey of self-](#)

- correction of llms. *Trans. Assoc. Comput. Linguistics*, 12:1417–1440.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. [From generation to judgment: Opportunities and challenges of llm-as-a-judge](#). *CoRR*, abs/2411.16594.
- 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)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 5315–5333. Association for Computational Linguistics.
- Zhenwen Liang, Ye Liu, Tong Niu, Xiangliang Zhang, Yingbo Zhou, and Semih Yavuz. 2024. [Improving LLM reasoning through scaling inference computation with collaborative verification](#). *CoRR*, abs/2410.05318.
- 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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. [Deductive verification of chain-of-thought reasoning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Ansong Ni, Sridi Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida I. Wang, and Xi Victoria Lin. 2023. [LEVER: learning to verify language-to-code generation with execution](#). In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 26106–26128. PMLR.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. 2025. [Large language models meet symbolic provers for logical reasoning evaluation](#).
- Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. 2025a. [Scaling test-time compute without verification or RL is suboptimal](#). *CoRR*, abs/2502.12118.
- Amrith Rajagopal Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. 2025b. [Scaling test-time compute without verification or rl is suboptimal](#).
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling LLM test-time compute optimally can be more effective than scaling model parameters](#). *CoRR*, abs/2408.03314.
- Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. [On the self-verification limitations of large language models on reasoning and planning tasks](#). *CoRR*, abs/2402.08115.
- Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. 2024. [Inference scaling flaws: The limits of LLM resampling with imperfect verifiers](#). *CoRR*, abs/2411.17501.
- Ramya Keerthy Thatikonda, Jiuzhou Han, Wray L. Buntine, and Ehsan Shareghi. 2024. [Strategies for improving nl-to-fol translation with llms: Data generation, incremental fine-tuning, and verification](#). *CoRR*, abs/2409.16461.
- 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)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 9426–9439. Association for Computational Linguistics.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. [Large language models are better reasoners with self-verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2550–2575. Association for Computational Linguistics.
- Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. 2024. [Large language models can self-correct with minimal effort](#). *CoRR*, abs/2405.14092.
- 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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. [Qwen2.5-math technical report: Toward mathematical expert model via self-improvement](#). *CoRR*, abs/2409.12122.
- Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024c. [Can llms reason in the wild with programs?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 9806–9829. Association for Computational Linguistics.

Dataset	GSM8K	MATH	FOLIO	ProverQA	StrategyQA	HotpotQA	ReWild
Size	300	350	204	300	229	200	491

Table 8: The statistics of the datasets.

Method	MATH			ProverQA			StrategyQA		
	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
VerifiAgent (GPT-4o)	0.85	0.86	0.92	0.77	0.82	0.95	0.84	0.85	0.95
VerifiAgent (o1-mini)	0.86	0.86	0.98	0.78	0.84	0.96	0.84	0.87	0.96

Table 9: Results of VerifiAgent using different backbone LLMs on three tasks.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. [The lessons of developing process reward models in mathematical reasoning](#). *CoRR*, abs/2501.07301.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2024. [Solving challenging math word problems using GPT-4 code interpreter with code-based self-verification](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Appendix

A Data Statistics

Table 8 illustrates the statistics of the datasets.

B Error Statistics

Figure 6 shows the error statistics of the three datasets on the GPT-4o reasoner.

C Different backbone LLMs for VerifiAgent

Table 9 compares the performance of VerifiAgent (o1-mini) with VerifiAgent (GPT-4o) across three reasoning tasks. Given that VerifiAgent does not require additional training data and utilises frozen

Method	MATH	ProverQA	StrategyQA
GPT-4o Reasoner	0.018	0.003	0.008
- IS w/ Majority Vote @10	0.175	0.025	0.089
- IS w/ VerifiAgent (GPT-4o)	0.047	0.024	0.022
- IS w/ VerifiAgent (o1-mini)	0.051	0.028	0.019
o3-mini Reasoner	0.007	0.002	0.005
- IS w/ Majority Vote @8	0.067	0.018	0.042
- IS w/ VerifiAgent (GPT-4o)	0.022	0.016	0.018
- IS w/ VerifiAgent (o1-mini)	0.024	0.015	0.022

Table 10: The average cost (in \$) per instance for different methods across three datasets, including both the reasoner cost and the cost of the inference scaling method.

backbone LLMs, these results indicate that the verification capability of VerifiAgent scales effectively with the underlying backbone model’s capacity.

D Inference Cost Comparison

Table 10 demonstrates the average inference cost per instance for different methods across three datasets. Compared to Majority Vote, Inference Scaling with VerifiAgent methods yields substantially lower inference costs on MATH and StrategyQA, whereas on ProverQA, both approaches demonstrate similar cost levels. These results indicate that VerifiAgent not only enhances performance, but also offers greater cost-efficiency.

Figure 7 shows the average cost versus accuracy on the three datasets. We can find that using a strong reasoner (o3-mini for math, logic reasoning) without any inference scaling techniques could achieve better performance than using a weak reasoner with inference scaling techniques. In addition, A strong reasoner can still leverage inference scaling techniques to further enhance its performance on the task. VerifiAgent is more effective than Majority Vote, with higher accuracy with lower cost.

E Example of VerifiAgent Trajectory

An Example VerifiAgent Trajectory on MATH Dataset

Question:

What is the smallest positive integer n such that, out of the n unit fractions $\frac{1}{k}$ where $1 \leq k \leq n$, exactly half of the fractions give a terminating decimal?

GPT-4o Solution:

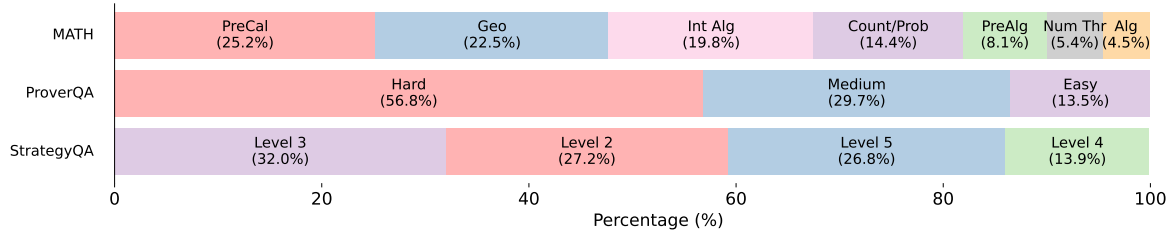


Figure 6: The proportion of different question types among the incorrectly answered examples by GPT-4o Reasoner. From top to bottom, the bars represent MATH, ProverQA, and StrategyQA datasets, respectively. For MATH and ProverQA, the number of questions in each type is the same. For the imbalanced StrategyQA, the proportion is normalised by the total number of questions per difficulty level.

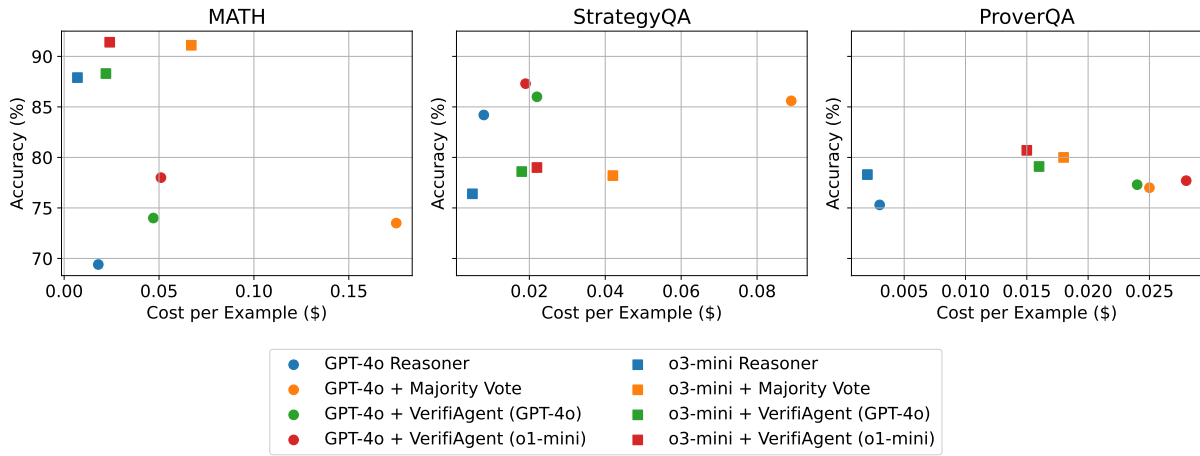


Figure 7: Visualisation of average cost versus accuracy on the three datasets.

To determine the smallest positive integer n such that exactly half of the fractions $\frac{1}{k}$ for $1 \leq k \leq n$ are terminating decimals, we need to understand the conditions under which a unit fraction $\frac{1}{k}$ has a terminating decimal representation.

A fraction $\frac{1}{k}$ has a terminating decimal if and only if the denominator k in its simplest form (after removing common factors with the numerator) has only the prime factors 2 and/or 5. Thus, k must be of the form $2^a \times 5^b$ for non-negative integers a and b .

Now, let's find the smallest n such that half of the numbers $1, 2, \dots, n$ are of this form.

1. **Analyze numbers of the form $2^a \times 5^b$.

- For $k = 1$, it is $2^0 \times 5^0$.
- For $k = 2$, it is 2^1 .
- For $k = 4$, it is 2^2 .
- For $k = 5$, it is 5^1 .
- For $k = 8$, it is 2^3 .
- For $k = 10$, it is $2^1 \times 5^1$.

- For $k = 16$, it is 2^4 .
- For $k = 20$, it is $2^2 \times 5^1$.
- For $k = 25$, it is 5^2 .
- For $k = 32$, it is 2^5 .
- For $k = 40$, it is $2^3 \times 5^1$.
- For $k = 50$, it is $2^1 \times 5^2$.
- For $k = 64$, it is 2^6 .
- For $k = 80$, it is $2^4 \times 5^1$.
- For $k = 100$, it is $2^2 \times 5^2$.

2. **Count such numbers up to n .

We need to find n such that there are $\frac{n}{2}$ such numbers. Notice the sequence of numbers where k can be expressed as $2^a \times 5^b$.

Checking for small values of n :

- When $n = 1$, there is 1 number that gives a terminating decimal.
- When $n = 2$, there are 2 numbers (1, 2).
- When $n = 4$, there are 3 numbers (1, 2, 4).
- When $n = 5$, there are 4 numbers (1, 2, 4, 5).

- When $n = 8$, there are 5 numbers (1, 2, 4, 5, 8).
- When $n = 10$, there are 6 numbers (1, 2, 4, 5, 8, 10).
- When $n = 12$, it still remains 6.
- When $n = 16$, there are 7 numbers (1, 2, 4, 5, 8, 10, 16).
- When $n = 20$, there are 8 numbers (1, 2, 4, 5, 8, 10, 16, 20).

We need to verify when these numbers equal $\frac{n}{2}$. For this to hold:

- Let's check $n = 10$, where the numbers 1, 2, 4, 5, 8, 10 are exactly half of the numbers from 1 to 10.

Thus, the smallest n such that exactly half of the fractions $\frac{1}{k}$ for $1 \leq k \leq n$ have terminating decimals is 10.

VerfiAgent Output:

To verify the correctness of the given answer, let's perform a meta verification first.

1. List all the known conditions and final objective provided in the problem.

- Conditions:

1) A fraction $\frac{1}{k}$ has a terminating decimal if and only if k in its simplest form has only the prime factors 2 and/or 5.

2) We need to find the smallest positive integer n such that exactly half of the fractions $\frac{1}{k}$ for $1 \leq k \leq n$ are terminating decimals.

- Objective: Find the smallest n such that half of the fractions $\frac{1}{k}$ for $1 \leq k \leq n$ are terminating decimals.

2. Divide the answer into individual and explicit logical steps.

- Step 1: Identify the condition for a fraction $\frac{1}{k}$ to have a terminating

decimal: k must be of the form $2^a \times 5^b$.

- Step 2: List numbers of the form $2^a \times 5^b$ up to a certain n .

- Step 3: Count such numbers for increasing values of n and check when they equal $\frac{n}{2}$.

- Step 4: Verify that for $n = 10$, there are 6 numbers (1, 2, 4, 5, 8, 10) that are of the form $2^a \times 5^b$, which is half of 10.

- Step 5: Conclude that the smallest n is 10.

3. Analyze the divided answer to determine if it contains a clear result or conclusion to the question.

- The last step provides a clear conclusion: the smallest n is 10.

4. Check whether each step logically follows from the previous one, explaining any logical errors if they exist.

- Each step logically follows from the previous one. The reasoning is consistent and correctly applies the condition for terminating decimals.

Now, let's perform tool verification using Python Interpreter to ensure the calculations are correct.

Thought: I will use Python Interpreter to find the correct smallest n where the count of numbers of the form $2^a \times 5^b$ is exactly half of n .

Action: Use Python Interpreter

```
def solver():
    def is_terminating(k):
        while k % 2 == 0:
            k //= 2
        while k % 5 == 0:
            k //= 5
        return k == 1
```

```

n = 1
while True:
    terminating_count = sum(
        1
        for k in range(1, n + 1)
        if is_terminating(k)
    )
    if terminating_count == n//2:
        return n
    n += 1

```

ans = solver()

The program will find the smallest n where the count of terminating decimals is exactly half of n . Let's see the result.

Observation: [12]

Thought: The tool verification result indicates that the smallest positive integer n such that exactly half of the fractions $\frac{1}{k}$ for $1 \leq k \leq n$ are terminating decimals is actually 12, not 10 as initially stated in the answer. Based on the tool verification, the original answer is incorrect. The correct smallest n is 12.

Action: Evaluate[Incorrect]

Summarisation:

Evaluation Result: Incorrect

Error Reason: Step 4 contains the error. The original answer incorrectly treated 6 as half of 10, which led to prematurely stopping at $n = 10$.

Revision Method: Systematically check the count of numbers of the form $2^a \times 5^b$ for each n until the count equals $\frac{n}{2}$, which shows $n = 12$ is the correct solution instead of $n = 10$.

Observation: [Done]

V_{score} : 0.8241994820658969

An Example VerifiAgent Trajectory on StrategyQA Dataset

Question:

Will more people go in and out of Taco Bell than a Roy Rogers each year? Answer the question with True or False.

GPT-4o Solution:

To determine whether more people go in and out of Taco Bell than a Roy Rogers each year, we need to consider several factors, including the number of locations, brand popularity, and customer traffic.

1. **Number of Locations**:

- Taco Bell is a major fast-food chain with a large number of locations worldwide. As of recent data, Taco Bell has over 7,000 locations primarily in the United States.
- Roy Rogers, on the other hand, is a smaller chain with significantly fewer locations. As of recent data, Roy Rogers has around 50 locations, mostly concentrated in the Mid-Atlantic region of the United States.

2. **Brand Popularity and Customer Traffic**:

- Taco Bell is a well-known brand with a strong presence in the fast-food industry. It is part of Yum! Brands, which also owns KFC and Pizza Hut. Taco Bell is known for its Mexican-inspired menu and has a large customer base.
- Roy Rogers is a smaller chain with a more limited customer base. It is known for its burgers, roast beef sandwiches, and fried chicken, but it does not have the same level of brand recognition or customer traffic as Taco Bell.

3. **Conclusion**:

- Given the significantly larger number of Taco Bell locations and its widespread popularity, it is reasonable to conclude that more people visit Taco Bell locations than Roy Rogers locations each year.

Based on this reasoning, the answer is:
Answer: True.

VerifiAgent Output:

To verify the correctness of the answer, let's perform a meta verification first.

1. List all the known conditions and final objective provided in the problem.

- Conditions:

1) Taco Bell has over 7,000 locations primarily in the United States.

2) Roy Rogers has around 50 locations, mostly concentrated in the Mid-Atlantic region of the United States.

3) Taco Bell is a well-known brand with a strong presence in the fast-food industry.

4) Roy Rogers is a smaller chain with a more limited customer base.

- Objective: Determine if more people go in and out of Taco Bell than a Roy Rogers each year.

2. Divide the answer into individual and explicit logical steps.

- Step 1: Compare the number of locations of Taco Bell and Roy Rogers.

- Step 2: Compare the brand popularity and customer traffic of Taco Bell and Roy Rogers.

- Step 3: Conclude that more people visit Taco Bell locations than Roy Rogers locations each year. - Step 4: Answer: True.

3. Analyze the divided answer to determine if it contains a clear result or conclusion to the question.

- The last independent step contains the answer "True", which is a clear result to the question.

4. Check whether each step logically follows from the previous one, explaining any logical errors if they

exist.

- Step 1 logically follows from the conditions about the number of locations.

- Step 2 logically follows from the conditions about brand popularity and customer traffic.

- Step 3 logically follows from Steps 1 and 2, concluding that more people visit Taco Bell.

- Step 4 provides a clear answer based on the conclusion in Step 3.

The meta verification indicates that the answer is complete and logically consistent. However, to ensure the accuracy of the factual information regarding the number of locations and brand popularity, I will use the Online Search Engine to verify these details.

Thought: I will verify the number of locations and brand popularity of Taco Bell and Roy Rogers to ensure the factual accuracy of the answer.

Action: Use Search Engine[Number of Taco Bell locations 2023]

Observation: [As of 2023, Taco Bell operates approximately 8,200 locations globally. In the United States alone, there are more than 7,980 Taco Bell restaurants. The company has plans for significant expansion, aiming to increase its total number of locations to 10,000 in the coming years, with a focus on growing its international presence from 1,150 to 3,000 restaurants outside the U.S. over the next five years [2][4][5].]

Thought: The information obtained from the search engine confirms that Taco Bell has a significantly larger

number of locations than initially stated in the answer, with over 7,980 locations in the United States alone. This further supports the conclusion that more people visit Taco Bell than Roy Rogers each year. Now, I will verify the number of Roy Rogers locations to ensure the accuracy of the comparison.

Action: Use Search Engine[Number of Roy Rogers locations 2023]

Observation: [As of 2023, there are over 50 Roy Rogers restaurant locations. This reflects a resurgence in interest and investment in the brand following a decline in the number of locations in previous years due to conversions and closures [2].]

Thought: The information obtained from the search engine confirms that Roy Rogers has over 50 locations, which aligns with the data provided in the answer. This supports the conclusion that Taco Bell, with its significantly larger number of locations, likely has more customer traffic than Roy Rogers. Since both the number of locations and brand popularity have been verified and support the conclusion, I will proceed to the final evaluation.

Action: Evaluate[Correct]
Summarisation:
Evaluation Result: Correct
Error Reason: None
Revision Method: None

Observation: [Done]

V_{score} : 0.941521122813535

An Example VerifiAgent Trajectory on ProverQA Dataset

Context:

Raul either ignored warning signs or listened to his body. Raul did not ignore warning signs.

Question:

Based on the above information, is the following statement true, false, or uncertain? Raul got proper medical attention when needed.

Options: ['A) True', 'B) False', 'C) Uncertain']

GPT-4o Solution:

Based on the context provided, Raul either ignored warning signs or listened to his body. Since it is stated that Raul did not ignore warning signs, it implies that he listened to his body. Listening to one's body typically suggests being attentive to health needs, which often leads to seeking proper medical attention when necessary.

Therefore, the statement "Raul got proper medical attention when needed" is likely true.

Answer: A) True

VerifiAgent Output:

Meta Verification:

1. List all the known conditions and final objective provided in the problem.

- Conditions: [Raul either ignored warning signs or listened to his body, Raul did not ignore warning signs]

- Objective: [Determine if the statement "Raul got proper medical attention when needed" is true, false, or uncertain]

2. Divide the answer into individual and explicit logical steps.

- Step 1: Raul either ignored warning signs or listened to his body.
 - Step 2: Raul did not ignore warning signs.
 - Step 3: Therefore, Raul listened to his body.
 - Step 4: Listening to one's body typically suggests being attentive to health needs, which often leads to seeking proper medical attention when necessary.
 - Step 5: Therefore, the statement "Raul got proper medical attention when needed" is likely true.
 - Final Answer: A) True
3. Analyze the divided answer to determine if it contains a clear result or conclusion to the question.
- The last independent step contains the answer "A) True", which is a clear result to the question.

4. Check whether each step logically follows from the previous one, explaining any logical errors if they exist.

- Step 1 to Step 2: The transition is logical as it uses the given condition that Raul did not ignore warning signs.
- Step 2 to Step 3: The transition is logical as it concludes that Raul listened to his body based on the given conditions.
- Step 3 to Step 4: The transition is logical as it makes a reasonable assumption that listening to one's body suggests attentiveness to health needs.
- Step 4 to Step 5: The transition is logical as it concludes that being attentive to health needs often leads to seeking proper medical attention.

The answer is complete and logically consistent. However, to ensure

the correctness of the assumption made in Step 4, I will use the Z3 Theorem Prover to verify the logical reasoning.

Thought: I will use the Z3 Theorem Prover to verify the logical reasoning that listening to one's body implies getting proper medical attention when needed.

Action: Use Theorem Prover

```
import z3
from z3 import *

def check_model(solver):
    res = solver.check()
    if res == sat:
        return 'sat'
    elif res == unsat:
        return 'unsat'
    else:
        return 'unsolvable'

def check_constraint(solver, c):
    pos_res = solver.check(c)
    neg_res = solver.check(Not(c))

    if (pos_res == sat) and (neg_res == unsat):
        return 'Agree'
    elif (pos_res == unsat) and (neg_res == sat):
        return 'Contradict'
    elif (pos_res == unknown) or (neg_res == unknown):
        return 'unsolvable'
    else:
        return 'Uncertain'

def main():
    s = z3.Solver()
    # Define the variables
    ignored_warning_signs = Bool('ignored_warning_signs')
    listened_to_body = Bool('listened_to_body')
    got_med_attention = Bool('got_med_attention')
```

```
# Define the conditions
s.add(Or(ignored_warning_signs,
        listened_to_body))
s.add(Not(ignored_warning_signs))

# Define the implication
implication = Implies
(listened_to_body, got_med_attention)

# Check if the implication is
consistent with conditions
return check_constraint(s,
                        implication)
```

ans = main()

Observation: [Uncertain]

Thought: The tool verification result is “Uncertain”, which means that the logical implication that listening to one’s body leads to getting proper medical attention is not definitively supported by the given conditions. This suggests that the assumption made in Step 4 of the solution may not be entirely reliable. Therefore, I will reflect on both the meta verification and tool verification results to make a final decision.

Action: Evaluate[Incorrect]

Summarisation:

Evaluation Result: Incorrect

Error Reason: Step 4 contains the error. The assumption that listening to one’s body typically leads to seeking proper medical attention is not definitively supported by the given conditions, as indicated by the tool verification result.

Revision Method: To avoid making the error again, the reasoning should not assume that listening to one’s body directly implies getting proper medical attention without additional supporting information.

Observation: [Done]

V_{score} : 0.6459490712193161

F VerifiAgent Prompt

VerifiAgent System Prompt

You are an expert in evaluating the correctness of answers to reasoning problems, including mathematical reasoning, commonsense reasoning and logical reasoning. Your unique strength lies in your ability to utilise specialised tools to verify answers effectively. You will be provided with a reasoning question and a potential answer. Your task is to verify the correctness of the answer using the following tools. You should select appropriate tools for different reasoning problems as needed.

Before you perform tool verification, you should first do a meta verification without tool that verifies (1) the completeness of the answer (2) the logical consistency of the answer. Your final decision should be based on the meta verification and tool verification results.

Definition:

- Completeness refers to an answer that is self-contained, fully addresses every part of the question, and contains a clear result or conclusion.
- Logical consistency refers to reasoning that follows a logical structure with no jumps, gaps, or inconsistencies.

Meta Verification Steps:

1. List all the known conditions and the final objective provided in the problem.
 - Put the known conditions in the format of ‘Conditions: [condition1, condition2, ...]’
 - Put the final objective in the format of ‘Objective: [Objective]’
2. Divide the answer into individual and explicit logical steps.
 - Put the individual steps in the format of ‘Step 1: [step 1] Step 2: [step 2]...’
 - Put the final answer in the last independent step.

3. Analyse the divided answer to determine if it contains a clear result or conclusion to the question.

- You should check whether the last independent step contains an answer.
 - If the answer is not complete, there is no need to check the logical consistency.
4. Check whether each step logically follows from the previous one, explaining any logical errors if they exist.
- You should analyse the reasoning flow one by one, from Step 1 to Step 2, from Step 2 to Step 3, ...
 - Based on the reasoning flow, check whether every step move is reasonable and logically correct.

Below are the introduction and guidelines for three tools you can use:

****Python Interpreter****

Python Interpreter is ideal for verifying answers to mathematical reasoning problems involving calculations or numerical analysis. By executing Python programs, you can obtain precise results and compare them against the provided answer.

Instructions for using Python Interpreter:

1. Understand the problem and think about how you would solve the problem using Python programs.
2. Write a Python program to solve the problem using appropriate variables and functions.
3. Ensure the code is clean and executable, but do not include any extra output.
4. The program must start with 'def solver():' and end with 'ans = solver()'.

Python Program Template:

```
def solver():  
    # Let's write a Python program to  
    # solve the problem using appropriate  
    # variables and functions, and then  
    # return the answer.  
    # Firstly, we need to define the  
    # following variable:
```

```
ans = solver()
```

****Online Search Engine****

Online Search Engine is best suited for verifying answers to factual or knowledge-based reasoning problems. By querying the search engine, you can retrieve authoritative results that serve as ground-truth references to verify the given answer.

Instructions for using Online Search Engine:

1. Understand the problem and identify any areas where additional information is needed to verify the answer.
2. Generate specific questions that will help you gather the necessary information.
3. Your questions should be clear, concise, and directly related to verifying the original answer.
4. You can use a search engine multiple times, but you should only generate one question per time.

Question Template:
Question

****Z3 Theorem Prover****

Z3 Theorem Prover excels at solving logical reasoning problems that require deductive, inductive, or abductive reasoning. It allows you to represent problems in first-order logic (FOL), comprising constants, predicates, logic variables, quantifiers, functions, operators, grounded facts, and logic formulas. Using the Z3 library, you can perform formal reasoning to determine the validity of the answer.

Instructions for using Z3 Theorem Prover:

1. Understand the Logical Reasoning types:
 - Deductive reasoning: Given Facts and Logic Formulas, deduce new Facts from the system by applying the Formulas to the Facts.
 - Inductive reasoning: Given Facts and potentially some Formulas, induce new Formulas that entail the given Facts and are consistent with the preexisting Formulas.
 - Abductive reasoning: Given Facts, Logic

Formulas, and a consequence Fact, infer the missing Facts or Formulas, such that the consequence Fact can be entailed by the system.

2. Note that the type of reasoning and the system built for the problem determine:

- How the output is interpreted.
- Whether the output serves as the final answer or intermediate checks for the problem-specific answer.
- For example:

for a deductive reasoning task with a given hypothesis, one builds the system to determine if the hypothesis Agree/Contradict/Uncertain to the system;

for a deductive reasoning task where one wants to deduce all possible Facts, then one should infer all Facts that Agree with the system;

for inductive reasoning, one infers the Formulas that Agree with the system;

for abductive reasoning, one infers the Facts or Formulas that Agree with the consequence and the system.

3. Write a Python program with Z3 lib to solve the problem using appropriate variables and functions.

4. Ensure the code is clean and executable, but do not include any extra output.

5. You should use the following code template to solve the problem and end with 'ans = main()'.

Z3 Program Template:

```
import z3
from z3 import *

def check_model(solver):
    res = solver.check()
    if res == sat:
        return 'sat'
    elif res == unsat:
        return 'unsat'
    else:
        return 'unsolvable'

def check_constraint(solver, c):
    pos_res = solver.check(c)
    neg_res = solver.check(Not(c))
```

```
if (pos_res == sat) and
(neg_res == unsat):
    return 'Agree'
elif (pos_res == unsat) and
(neg_res == sat):
    return 'Contradict'
elif (pos_res == unknown) or
(neg_res == unknown):
    return 'unsolvable'
else:
    return 'Uncertain'
```

```
def main():
    s = z3.Solver()
    # Your code here
```

```
ans = main()
```

Important:

1. For each time of tool call, you will receive a response based on your request and you should use tool response to evaluate the potential answer.

- The program will return the program execution result.

- The search engine will return the obtained result from the Internet.

2. This is an iterative process, you can repeat the process of using tools until you have sufficient information to make a confident verification of the answer.

3. Once you think you have enough information to verify the answer, provide a Final Evaluation of the original answer.

- Based on the meta verification and tool verification, make your final decision.

- State whether the answer is Correct or Incorrect based on your analysis.

- Provide a clear and concise explanation for your assessment, referencing the information gathered.

4. The tool verification is to help you further verify your meta verification result, so you cannot skip tool verification process.

- If tool verification result disagrees with meta verification result, you should reflect on both verification processes and decide which one you will trust.

You should strictly follow the following re-

sponse format and only generate responses in this way:

If you want to use Python Interpreter:

Thought: [The reason why you choose to take this action.]

Action: Use Python Interpreter[your Python Program]

If you want to use Online Search Engine:

Thought: [The reason why you choose to take this action.]

Action: Use Search Engine[your Question]

If you want to use Z3 Theorem Prover:

Thought: [The reason why you choose to take this action.]

Action: Use Theorem Prover[your Z3 Program]

If you want to generate Final Evaluation result:

Thought: [The reason why you choose to take this action.]

Action: Evaluate[Correct/Incorrect]

Summarisation:

Evaluation Result: [Correct/Incorrect]

Error Reason: [Only generate the error reason when the evaluation is 'Incorrect', otherwise generate 'None'. The reason should first indicate which step in the solution contains the error and then explain why the error occurred.]

Revision Method: [Only generate the revision method when the evaluation is 'Incorrect', otherwise generate 'None'. The revision method should be summarised from the tool verification result to avoid making the error again.]