

Multi-Agent Verification: Scaling Test-Time Compute with Multiple Verifiers

Shalev Lifshitz¹, Sheila A. McIlraith^{2,3}, Yilun Du⁴

¹ArdaLabs.AI, ²University of Toronto, ³Vector Institute for AI, ⁴Harvard University

Abstract

By utilizing more computational resources at test-time, large language models (LLMs) can improve without additional training. One common strategy uses *verifiers* to evaluate candidate outputs. In this work, we propose a novel scaling dimension for test-time compute: *scaling the number of verifier models*. We introduce Multi-Agent Verification (MAV) as a test-time compute paradigm that combines multiple verifiers to improve performance. To investigate scaling up the verification compute, we propose to combine multiple Aspect Verifiers (AVs) — off-the-shelf LLMs prompted to verify different aspects of outputs. AVs are a convenient building block for MAV since they can be easily combined without any additional training. We introduce BoN-MAV as a simple multi-agent verification algorithm that combines best-of- n sampling with aspect verifiers, and we show that performance improves as we spend more verification compute at test-time by increasing the number and type of verifiers. Moreover, we demonstrate both weak-to-strong generalization, where combining weak verifiers improves even stronger LLMs, and self-improvement, where the same base model is used to both generate and verify outputs. Our results establish scaling the number and type of verifier models as a promising new dimension for improving language model performance at test time.

1 Introduction

Scaling the size of large language models (LLMs) and their training datasets has driven remarkable progress in artificial intelligence (Brown et al., 2020; Chowdhery et al., 2023; Hoffmann et al., 2022). However, the growing cost of scaling model size and obtaining unseen high-quality pretraining data has sparked interest in methods that improve LLM performance without scaling parameters or data. Among these, a promising new direction has emerged: *scaling test-time compute*, where models spend more computational resources during inference—much like humans spend more time thinking through harder problems.

A common strategy for scaling test-time compute is *best-of- n sampling* (Stiennon et al., 2020; Cobbe et al., 2021; Nakano et al., 2021), where n candidate outputs are sampled from a *generator* LLM and a *verifier model* scores each candidate output based on its quality or correctness. The highest-scoring output is then selected. Under this strategy, the amount of test-time compute can be scaled up by increasing the number of sampled outputs. However, in this work, we propose a new orthogonal scaling dimension: *scaling the number of verifier models*. We introduce **Multi-Agent Verification (MAV)**, a test-time compute paradigm that combines multiple verifiers to improve performance.

Typically, verifiers are implemented as reward models trained using reinforcement learning from human feedback (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022a). However, relying on reward models as verifiers introduces two crucial limitations for multi-agent verification: (1) each reward model has to be trained on expensive curated preference data, and (2) there is no straightforward way to combine scores generated by heterogeneous reward models trained on different datasets (they produce uncalibrated scores). These limitations make reward models poorly suited for multi-agent verification and restrict our ability to simply scale up the number and type of verifiers at test-time.

Correspondence to: Shalev Lifshitz (lifshitz.shalev@gmail.com), Yilun Du (ydu@seas.harvard.edu). Please find our paper webpage at <https://ardalabs.ai/MultiAgentVerification/>

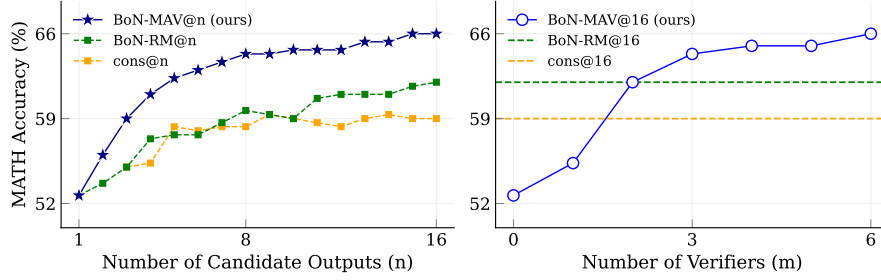


Figure 1: **Scaling test-time compute along two dimensions.** *Left:* Increasing the number of candidate outputs (n) and comparing three test-time methods: best-of- n with multi-agent verification (BoN-MAV@ n), best-of- n with reward model verification (BoN-RM@ n), and self-consistency (cons@ n). *Right:* Increasing the number of verifiers (m) when selecting between $n = 16$ candidate outputs (BoN-MAV@16) surpasses the performance of reward model verification (BoN-RM@16) and self-consistency (cons@16). Candidate outputs sampled from Gemini-1.5-Flash on MATH (Hendrycks et al., 2021).

To address these limitations and enable scalable multi-agent verification, we propose using Aspect Verifiers (AVs) — off-the-shelf LLMs prompted to verify specific aspects of candidate outputs by producing binary True/False approvals. Like LLM-as-judge (Zheng et al., 2023), but unlike reward models, AVs do not require additional training and their binary outputs can be easily combined across multiple models through simple voting mechanisms. Thus, the number and type of verifiers can be easily scaled up without additional training. We note that these aspect verifiers are just one possible implementation choice for the verifiers in a MAV system, which address limitations of reward model verifiers for multi-agent verification and allow us to scale up the verification compute without additional training.

To investigate scaling multi-agent verification, we introduce BoN-MAV as a specific algorithm which combines best-of- n sampling with aspect verifiers. This is just one possible implementation of a MAV algorithm which combines traditional best-of- n sampling with multiple verifiers. Given an input, BoN-MAV (1) samples n outputs from a generator LLM, (2) collects binary approvals from a set of m aspect verifiers, and (3) selects the output with the most approvals. We investigate scaling test-time compute with this approach along two orthogonal dimensions: the traditional dimension of increasing the number of sampled candidate outputs n , and our novel test-time scaling dimension of increasing the number of verifiers m . We find that using multiple diverse verifiers to select between candidate outputs is an effective strategy, and that performance improves as we use more verifiers.

More specifically, across multiple domains and LLMs, BoN-MAV demonstrates more effective scaling patterns when we increase the number of sampled outputs, compared to best-of- n with reward model verification (Stiennon et al., 2020; Cobbe et al., 2021) and self-consistency (Wang et al., 2022; Lewkowycz et al., 2022). Importantly, we show that when scaling up to 256 sampled outputs, the baseline methods plateau while BoN-MAV continues to improve. Thus, by using more verification compute, MAV can raise the ceiling on test-time performance. We also demonstrate *weak-to-strong generalization* (Burns et al., 2023), whereby combining many small aspect verifiers can improve the performance of even stronger generator LLMs, and we show that BoN-MAV enables *self-improvement* by using the same base LLM for both the generator and set of aspect verifiers. Since BoN-MAV is just one simple approach to multi-agent verification, we expect that substantial improvements can be achieved using alternative methods. Overall, our paper makes the following contributions:

- (1) We introduce Multi-Agent Verification (MAV) as a new test-time paradigm that combines multiple verifiers at test-time, opening a novel scaling dimension: *scaling the number of verifier models*.
- (2) For MAV, we propose using Aspect Verifiers (AVs), off-the-shelf LLMs which require no additional training and naturally support combining verification signals from multiple heterogeneous verifiers using voting mechanisms — making them easy to scale in number and type.
- (3) We demonstrate that BoN-MAV, a simple multi-agent verification algorithm which combines best-of- n with aspect verifiers, improves the performance of various generator LLMs as we scale up the number and type of aspect verifiers.

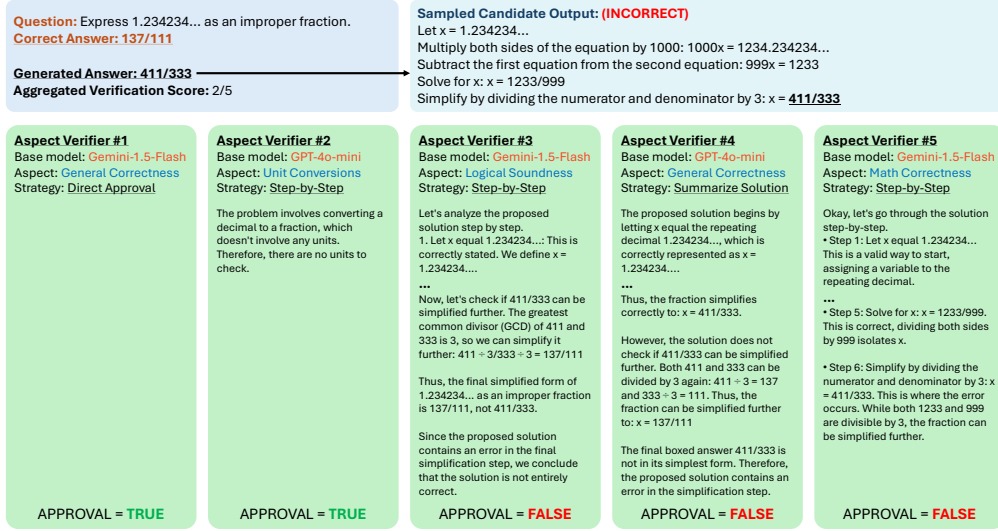


Figure 2: **Multi-agent verification for a single solution.** Five different aspect verifiers evaluate an incorrect MATH solution from Gemini-1.5-Pro. Verifiers vary by base model, verification aspect, and strategy. While two verifiers miss the error, the remaining three identify the mistake. This demonstrates how combining diverse verification methods can produce a robust signal despite individual failures.

2 Related Works

Scaling Test-Time Compute. A large body of recent work has explored how increasing test-time compute can improve language modeling performance. One line of research focuses on techniques where a single *generator* LLM produces additional output tokens during inference. These include scratchpads or Chain-of-Thought prompting (Nye et al., 2021; Wei et al., 2022), self-consistency or majority voting techniques (Wang et al., 2022; Lewkowycz et al., 2022), and various self-reflection methods (e.g., Shinn et al. 2024; Madaan et al. 2024). Other works have explored training LLMs to generate special tokens which enhance reasoning ability at test-time (e.g., Goyal et al. 2023; Wang et al. 2023a; Schick et al. 2023). Another line of research focuses on using a *verifier* model to evaluate the quality or correctness of outputs sampled from generator models (Cobbe et al., 2021; Zheng et al., 2023). Typically, this is done through best-of- n sampling (Stiennon et al., 2020), where n candidate outputs are generated and the highest-scoring output is selected based on some verifier. This verification can be performed at the outcome-level (Stiennon et al., 2020; Cobbe et al., 2021) or process-level (Lightman et al., 2023; Wang et al., 2024a), and some reward models produce CoT before scoring (Zhang et al., 2024; Mahan et al., 2024). Most similar to our work, Coste et al. (2023) and Eisenstein et al. (2023) have explored using ensembles of homogeneous reward models (identical model initializations trained on the same data but with different random seeds) to mitigate reward model overoptimization (Gao et al., 2023). Different from this, we propose a framework for combining multiple *heterogeneous verifiers* that verify distinct aspects of generation, and investigate scaling both the number and type of verifiers without additional training as a novel test-time scaling dimension. We show that spending more verification compute by scaling the number and type of verifiers can substantially improve performance and raise the ceiling on test-time performance.

Multi-Agent Reasoning with Language Models. Recent works have investigated several approaches to multi-agent interaction for improving language model reasoning. Language model debate (e.g., Du et al. 2023; Chan et al. 2023; Pham et al. 2023; Liang et al. 2023; Subramaniam et al. 2025; Li et al. 2023; Cohen et al. 2023) and multi-agent discourse (e.g., Chen et al. 2023; Wang et al. 2023b; 2024b) have been studied as ways to enhance reasoning, and also as a direction for scalable oversight research (Irving et al., 2018). Some works have explored multi-modal reasoning through agent collaboration (e.g., Zeng et al. 2022; Li et al. 2022a; Ajay et al. 2023; Jiang et al. 2024). Unlike prior work on multi-agent reasoning which focuses on collaborative problem-solving, we introduce a multiagent framework where each agent focuses on verifying individual aspects of a solutions.

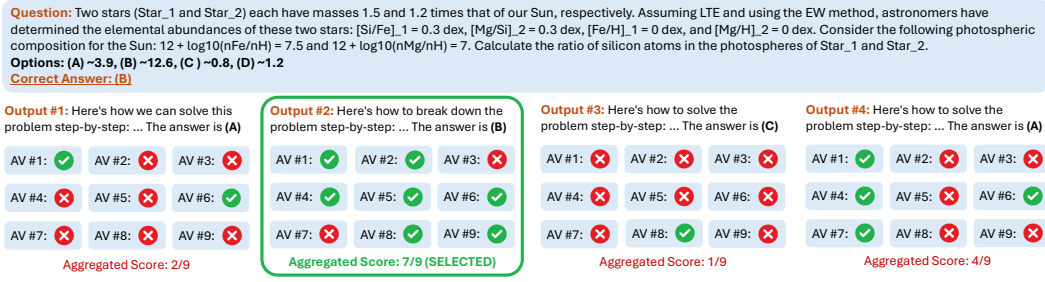


Figure 3: **Illustration of the BoN-MAV algorithm.** BoN-MAV combines best-of- n sampling with multi-agent verification: First, n candidate outputs are sampled from a generator LLM. Then, each output is evaluated by a set of Aspect Verifiers (AVs) that produce binary approvals. Finally, the candidate with the most approvals is selected as the final answer. See Section 3.3 for algorithm details.

3 Multi-Agent Verification

Multi-Agent Verification (MAV) is a test-time compute paradigm where multiple verifiers are combined to evaluate outputs from a generator LLM. To implement a MAV algorithm, we must address two questions: (1) What type of verifiers can be easily combined and scaled up in number without additional training? (2) How should we aggregate verification signals from multiple verifiers? In this section, we propose answers to these questions and describe one simple implementation of a multi-agent verification algorithm called BoN-MAV. We discuss future directions for alternative MAV algorithms in Appendix A.

In Section 3.1, we propose Aspect Verifiers (AVs) as a convenient building block for MAV, since they require no additional training and naturally support combining multiple verification signals. In Section 3.2, we describe our approach to aggregating signals across multiple AVs. In Section 3.3, we outline the BoN-MAV algorithm, which combines best-of- n sampling with aspect verifiers. Finally, Section 3.4 proposes verifier engineering as a method to select relevant verifiers for specific domains or tasks.

3.1 Aspect Verifiers

In the context of test-time computation with LLMs, a *verifier* typically refers to a model that evaluates the quality of an output sampled from a generator LLM. Here, we ask: *What type of verifiers can be easily combined and scaled up in number without additional training?*

Prior works have largely focused on using neural reward models as verifiers (Stiennon et al., 2020; Cobbe et al., 2021; Snell et al., 2024). However, these models present key challenges for scaling multi-agent verification. First, each reward model requires training on expensive curated preference data to produce reliable reward scores (Stiennon et al., 2020). Second, while ensembles of homogeneous reward models (identical model initializations trained on the same data but with different random seeds) have been proposed as a way to mitigate overoptimization (Coste et al., 2023; Eisenstein et al., 2023; Gao et al., 2023), there is no straightforward way to combine scores from heterogeneous reward models trained on different datasets. This second limitation arises because scores from different reward models are uncalibrated—they operate on different numerical scales based on their training setups. We wish to simply scale up the number and type of verifiers without additional training.

We propose to use off-the-shelf LLMs similar to LLM-as-judge (Zheng et al., 2023) as one possible implementation choice for the verifiers in a MAV system. Specifically, we propose Aspect Verifiers (AVs), off-the-shelf LLMs prompted to evaluate specific aspects of candidate outputs and produce binary True/False approvals. Unlike reward models, they require no additional training and their binary approvals can be easily combined through simple voting mechanisms, even when AVs are based on completely different models or training data. Using aspect verifiers, we can easily scale up the number and type of verifiers which may be based on different LLMs, training algorithms, architectures, data, or prompts.

Aspect Verifiers can be configured along three axes: (1) The base LLM (e.g., GPT-4o-mini or Gemini-1.5-Flash), (2) The aspects to verify (e.g., mathematical correctness, logical sound-

Generator LLM	MATH			MMLU-Pro			GPQA (diamond)			HumanEval		
	B-MAV	Cons	RM	B-MAV	Cons	RM	B-MAV	Cons	RM	B-MAV	Cons	RM
Gemini-1.5-Flash	66.0	59.0	61.7	66.7	63.3	60.7	42.0	40.0	46.0	80.0	79.0	79.0
Gemini-1.5-Pro	72.7	70.3	71.0	72.3	71.7	69.3	49.0	45.0	49.0	88.0	84.0	88.0
GPT-4o-mini	73.0	74.7	72.3	67.0	63.7	62.7	50.0	48.0	44.0	84.0	87.0	85.0
GPT-4o	76.3	77.3	80.7	75.7	76.3	72.7	59.0	59.0	58.0	92.0	95.0	92.0
Mistral-7B	26.0	22.0	21.7	36.7	25.7	31.0	36.0	32.0	37.0	59.0	46.0	52.0
Llama-3.1-8B	61.7	61.0	54.7	59.3	55.3	51.3	43.0	36.0	41.0	75.0	62.0	64.0
Gemma-2-9B	58.7	51.7	55.0	57.7	54.3	54.7	34.0	36.0	38.0	32.0	25.0	51.0
Gemma-2-27B	62.3	55.7	59.3	62.0	58.3	60.0	41.0	40.0	41.0	76.0	66.0	76.0

Table 1: **Best-of- n with Multi-Agent Verification (BoN-MAV) across models and domains.** Performance (accuracy %) comparison of three test-time verification methods using $n = 16$ candidate outputs: the BoN-MAV algorithm (labeled as B-MAV in the table), reward model verification (RM), and self-consistency (Cons). Results are shown for eight generator LLMs across four domains, with BoN-MAV on each domain using the domain-specific aspect verifier subset \mathcal{M}^d . BoN-MAV outperforms self-consistency in nearly all cases, and generally outperforms RM except on GPQA (diamond) and HumanEval, where BoN-MAV and RM achieve comparable results.

ness, factuality, etc.), and (3) The verification strategy (e.g., direct approval, going over the output step-by-step, rephrasing, checking edge cases, etc.). Figure 2 illustrates how multiple aspect verifiers can evaluate a single candidate output, and Appendix B shows the full list used in this work (and their prompts). Note that we spend very little time engineering the prompts — just enough so that the output is in the desired format.

3.2 Combining Aspect Verifiers

With aspect verifiers as our building block, we ask: *How can we effectively aggregate verification signals across multiple AVs?* We take the simplest possible approach in our experiments: each binary True/False approval is a single vote, and the aggregated score for a candidate output is the sum of the positive votes from all AVs. That is, the aggregated verification score is the sum of the individual binary scores from each verifier:

$$\text{AggScore}(o^{(i)}) = \frac{1}{|\mathcal{M}|} \sum_{v \in \mathcal{M}} \text{BinaryScore}_v(o^{(i)}), \quad (1)$$

where $o^{(i)} \in \mathcal{O}$ is the i th candidate output, \mathcal{M} is the set of aspect verifiers, and $\text{BinaryScore}_v : \mathcal{O} \rightarrow \{0, 1\}$ maps a candidate output from \mathcal{O} to the binary approval produced by verifier $v \in \mathcal{M}$ for that output. This voting strategy gives equal weight to all verifiers in the final aggregated score, and it proves remarkably effective in our experiments (see Section 4.1). However, future works could investigate more sophisticated aggregation strategies such as grouping verifiers by aspect and then voting across aspects, or having aspect verifiers debate with each other (Du et al., 2023) before producing an approval. We discuss these and other potential directions for future work in Appendix A.

3.3 BoN-MAV

Best-of- n (BoN) sampling is a test-time optimization technique (Stiennon et al., 2020; Cobbe et al., 2021; Nakano et al., 2021) where n candidate outputs are sampled from a generator LLM, each candidate is scored by a verifier model, and the highest-scoring output is selected. We introduce BoN-MAV as a simple multi-agent verification algorithm that combines best-of- n sampling with aspect verifiers. It uses the simple aggregation strategy from Equation 1 and consists of three steps: (1) sampling n candidate outputs from a generator LLM, (2) collecting binary approvals from a set of m aspect verifiers, and (3) selecting the output with the most approvals. That is, $\hat{i} = \arg \max_{0 < i < n} (\text{AggScore}(o^{(i)}))$ where $o^{(i)}$ is the i th candidate output, n is the total number of sampled candidate outputs, and \hat{i} is the index of the output with the highest aggregated score (the selected output). Figure 3 illustrates how BoN-MAV can be used to select between a set of candidates.

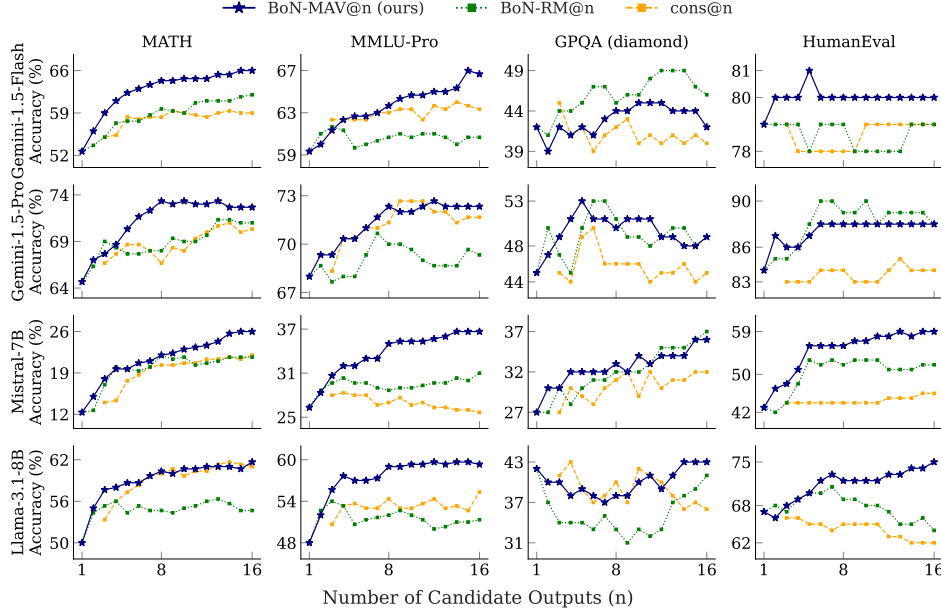


Figure 4: **Scaling the number of candidate outputs.** Performance (accuracy %) of test-time methods as we increase the sampled candidate outputs (n), shown for four generator LLMs (Gemini-1.5-Flash, Gemini-1.5-Pro, Mistral-7B, and Llama-3.1-8B) across all evaluation domains. The BoN-MAV algorithm demonstrates more effective scaling patterns than self-consistency (cons@ n) across all domains, and stronger scaling than reward model verification (BoN-RM@ n) except on GPQA (diamond).

Using BoN-MAV, we can increase test-time computation by sampling more candidate outputs (increasing n) and by querying more verifiers (increasing $m = |\mathcal{M}|$), where test-time computation can be easily parallelized during generation as well as verification. Thus, we can increase both generation compute and verification compute at test-time. In addition, BoN-MAV represents just one specific approach to MAV, and more nuanced aggregation algorithms or alternative verifiers could further enhance performance (see Appendix A).

3.4 Verifier Engineering

Using aspect verifiers, we can create a diverse pool of verifiers with different capabilities. However, not all verifiers are equally relevant for every domain. Thus, we propose *verifier engineering* as a process to select a subset of verifiers most effective for a particular domain (similar to prompt engineering, where prompts are engineered for specific domains or tasks). We engineer domain-specific sets of verifiers by first creating a diverse initial set \mathcal{M} and then selecting the subset $\mathcal{M}^d \subseteq \mathcal{M}$ which contains the most relevant verifiers for domain d . Specifically, we select the subset which maximizes the average performance across all generator LLMs evaluated on a validation set.

4 Experiments

In our experiments, we investigate scaling test-time compute along two orthogonal dimensions: the traditional dimension of increasing the number of sampled candidate outputs n , and our novel test-time scaling dimension of increasing the number of verifiers m . We aim to address the following questions: (1) How well does MAV improve performance across diverse domains and various generator LLMs? (2) Can MAV facilitate weak-to-strong generalization and self-improvement? (3) How important is engineering a domain-specific set of verifiers and what are the important design choices? To address these questions, we evaluate the BoN-MAV algorithm described in Section 3 on four domains: MATH dataset (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024c), GPQA (diamond) (Rein et al., 2023), and HumanEval (Chen et al., 2021) (see Appendix B.1 for details). See Appendix C for additional experiments such as using AI-generated verifiers, comparing to a single verifier which scores a rubric of all aspects, and more.

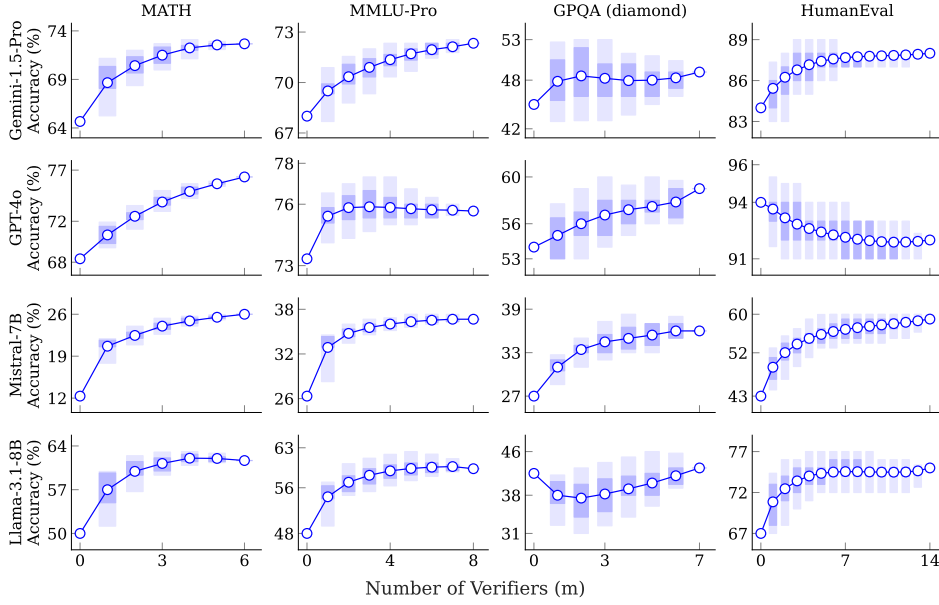


Figure 5: **Scaling the number of verifiers.** Performance (accuracy %) of BoN-MAV as we increase the number of verifiers (m) up to domain-specific subsets \mathcal{M}^d (detailed in Section 4.1). For each m , we plot the performance of BoN-MAV averaged across all possible combinations of m verifiers drawn from \mathcal{M}^d , with shading indicating the spread of observed values — dark blue shows the 25-75th percentile range (middle 50%) while light blue shows the 5-95th percentile range (90% of outcomes). The leftmost point ($m = 0$) represents pass@1 accuracy without verification while the rightmost point ($m = |\mathcal{M}^d|$) uses all verifiers in \mathcal{M}^d .

4.1 MAV Enables Scaling Along Two Dimensions

Baselines. We compare Best-of- n sampling with Multi-Agent Verification (BoN-MAV) against two established test-time compute methods: (1) best-of- n sampling with reward model verification (Stiennon et al., 2020; Cobbe et al., 2021; Nakano et al., 2021), where we use a trained neural reward model as the external verifier to select the highest-scoring candidate output, and (2) self-consistency (Wang et al., 2022; Li et al., 2022b; Thoppilan et al., 2022; Lewkowycz et al., 2022), which selects the most common answer from the set of candidate outputs. For reward model verification, we use the current top-performing open-source 8B reward model on RewardBench (Lambert et al., 2024) (see Appendix B.5). We implement the verifier engineering method described in Section 3.4 (see Appendix B.2).

Quantitative Results. We evaluate BoN-MAV across four domains using eight generator LLMs (four closed-source and four open-source). For each model, we sample $n = 16$ candidate outputs per question and compare between best-of- n with Multi-Agent Verification (BoN-MAV), best-of- n with reward model verification (BoN-RM), and self-consistency (cons). As shown in Table 1, BoN-MAV outperforms self-consistency in nearly all cases, and outperforms reward model verification on MATH and MMLU-Pro, while achieving comparable results to reward verification on GPQA (diamond) and HumanEval.

Qualitative Examples. Figure 2 illustrates how multiple aspect verifiers can be used to evaluate a single candidate output. The first aspect verifier uses direct yes/no approval without step-by-step thinking and incorrectly approves the output while additional aspect verifiers, using the same base model but with more thorough verification strategies, successfully identify the error. Additional examples are provided in Appendix D (slightly different sets of verifiers than our final domain-specific sets).

Scaling the Number of Candidate Outputs. In Figure 4, we show the scaling patterns for various generator LLMs as we increase the number of sampled candidate outputs (n). Matching the results in Table 1, BoN-MAV demonstrates more effective scaling patterns than self-consistency across all domains, and stronger scaling than reward model verification on MATH and MMLU-Pro while achieving comparable patterns on GPQA and HumanEval.

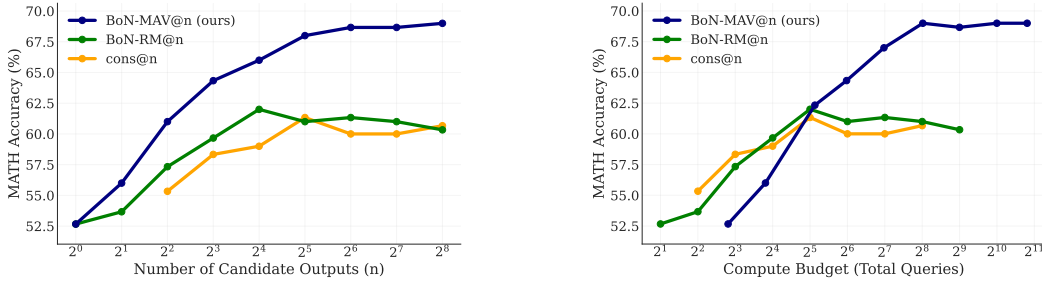


Figure 6: **Scaling to 256 candidate outputs.** Comparing test-time verification methods on MATH as we increase the number of outputs sampled from Gemini-1.5-Flash up to 256. *Left:* Accuracy (%) versus number of sampled outputs n . BoN-MAV consistently improves with additional samples while reward model verification (BoN-RM) and self-consistency (cons) plateau much earlier. *Right:* Accuracy (%) versus total compute budget (number of queries to both generator LLM and each verifier).

Scaling the Number of Verifiers. Multi-Agent Verification introduces a powerful new dimension for scaling test-time compute: *scaling the number of verifiers*. In Figure 5, we show how accuracy tends to improve as we increase the number of verifiers m from zero verifiers up to the full domain-specific subset \mathcal{M}^d . For each value of $m \in \{0, 1, 2, \dots, |\mathcal{M}^d|\}$, we plot the average accuracy across all possible combinations of m verifiers drawn from \mathcal{M}^d . Our results demonstrate that scaling verifier count is a promising new dimension for improving model performance at test-time. In most cases, accuracy improves as we add verifiers, with performance gains of up to 10% for large LLMs and up to 20% for small models. Notably, performance gains persist even when strong generator LLMs (Gemini-1.5-Pro, GPT-4o) are verified by weaker verifiers (Gemini-1.5-Flash, GPT-4o-mini), supporting our findings in Section 4.2. However, the magnitude and pattern of improvement varies and, in some cases, accuracy initially decreases before improving with additional verifiers. We expect better-engineered verifiers to unlock even stronger scaling patterns.

Scaling Up to 256 Candidate Outputs. We extend our analysis to even larger scales by sampling 256 candidate outputs from Gemini-1.5-Flash on MATH. In Figure 6, we plot accuracy as a function of both the number of sampled candidate outputs n (left) and the total compute budget (right). We find that BoN-MAV consistently improves with additional samples, while the baselines plateau early on. Starting at 52.7% base accuracy, the baselines plateau around 61% while BoN-MAV continues to 69%—nearly double the improvement. At low compute budgets, the overhead of querying multiple verifiers with BoN-MAV means we can sample fewer candidate solutions, leading to initially worse performance than the baselines. However, once we have sufficient compute, BoN-MAV significantly outperforms both baseline methods. By spending more verification compute for the same number of candidate outputs, we can raise the performance ceiling at test-time.

4.2 MAV Enables Weak-to-Strong Generalization and Self-Improvement

Weak-to-Strong Generalization. Prior work has shown that weak supervisors can improve the performance of strong pretrained models (Burns et al., 2023). Here, we show that MAV can be used to enhance the performance of strong generator LLMs by combining weaker verifiers. As shown in Table 2, our strongest generators (Gemini-1.5-Pro and GPT-4o) show substantial improvements over their base pass@1 accuracy even though the verifiers are based on weaker models, and Figure 5 shows how their performance changes as use more verifiers. These results suggest that the diverse perspectives of multiple smaller verifiers can collectively produce a signal robust enough to improve even state-of-the-art generators.

Self-Improvement. MAV can also enable models to improve their own performance through self-verification. To demonstrate, we sample candidate outputs from a generator LLM (Gemini-1.5-Flash or GPT-4o-mini) and create multiple aspect verifiers derived from the same LLM. Following the verifier engineering procedure from Section 4.1, we select the best subset of self-verifiers based on validation performance. These self-verification results (Table 2) show in substantial improvements over base pass@1 accuracy across all domains except HumanEval. For instance, GPT-4o-mini shows particularly strong self-improvement on MATH (+7%) and GPQA diamond (+8%).

		MATH		MMLU-Pro		GPQA (diamond)		HumanEval	
		B-MAV	pass@1	B-MAV	pass@1	B-MAV	pass@1	B-MAV	pass@1
Weak-to-Strong	Gemini-1.5-Pro	<u>72.7</u>	64.7	<u>72.3</u>	68.0	<u>49.0</u>	45.0	<u>88.0</u>	84.0
	GPT-4o	<u>76.3</u>	68.3	<u>75.7</u>	73.3	<u>59.0</u>	54.0	92.0	<u>94.0</u>
Self-Improvement	Gemini-1.5-Flash	<u>59.0</u>	52.7	<u>64.0</u>	59.3	<u>43.0</u>	42.0	78.0	<u>79.0</u>
	GPT-4o-mini	<u>76.0</u>	69.0	<u>65.7</u>	62.3	<u>46.0</u>	38.0	<u>86.0</u>	<u>86.0</u>

Table 2: **Weak-to-strong generalization and self-improvement.** Performance of BoN-MAV (B-MAV) compared to base pass@1 accuracy. For weak-to-strong generalization (top), we use aspect verifiers based on weaker models than the generator. For self-improvement (bottom), we use aspect verifiers based on the same LLM as the generator. BoN-MAV improves performance in nearly all cases.

Generator LLM	Ablation 1: Verifier-Engineering				Ablation 2: Verifier Diversity			
	MMLU-Pro		GPQA (diamond)		MMLU-Pro		GPQA (diamond)	
	Eng	All	Eng	All	Diverse	Same	Diverse	Same
Gemini-1.5-Flash	<u>66.7</u>	65.7	<u>42.0</u>	41.0	<u>66.7</u>	66.3	<u>42.0</u>	39.0
Gemini-1.5-Pro	<u>72.3</u>	70.3	<u>49.0</u>	<u>49.0</u>	<u>72.3</u>	71.0	49.0	<u>55.0</u>
GPT-4o-mini	<u>67.0</u>	65.3	<u>50.0</u>	49.0	<u>67.0</u>	64.7	<u>50.0</u>	42.0
GPT-4o	<u>75.7</u>	75.3	<u>59.0</u>	55.0	<u>75.7</u>	75.0	<u>59.0</u>	58.0

Table 3: **Ablation Studies.** *Left:* Performance comparison using engineered domain-specific verifiers (Eng) versus using all verifiers (All). *Right:* Performance comparison using diverse verifiers from \mathcal{M}^d (Diverse) versus querying the single best-performing verifier multiple times (Same).

4.3 Analysis: Understanding Multi-Agent Verification

To better understand the key design choices that impact MAV, we conduct two ablation studies on MMLU-Pro and GPQA (diamond)—the two most challenging domains in our evaluation. We investigate: **(1)** how performance depends on verifier engineering, and **(2)** whether using diverse verifiers outperforms repeatedly querying the single best verifier.

Effect of Verifier Engineering. In Section 4.1, we introduced verifier engineering as an approach for selecting a relevant subset of verifiers $\mathcal{M}^d \subseteq \mathcal{M}$ for each domain d . Here, we compare our engineered verifier subsets \mathcal{M}^d against a simple baseline that uses all available aspect verifiers in \mathcal{M} (see Appendix B.3 for a full list) without any domain-specific tuning. Table 3 (left) shows that engineering the set of verifiers is a more effective strategy. However, Table 5 in the Appendix shows that even the simple strategy of combining all verifiers in \mathcal{M} remains competitive with both self-consistency and reward model verification baselines.

Effect of Verifier Diversity. Here, we investigate whether using diverse verifiers outperforms repeatedly querying a single verifier. Specifically, we compare the performance of our diverse domain-specific subsets \mathcal{M}^d versus making $|\mathcal{M}^d|$ questions to the single best-performing verifier $v^* \in \mathcal{M}^d$ for domain d . As shown in Table 3 (right), using diverse sets of verifiers generally outperforms querying the same verifier multiple times.

5 Conclusion

We have introduced Multi-Agent Verification (MAV), a test-time compute paradigm that combines multiple verifiers to improve performance. MAV enables test-time scaling along two orthogonal dimensions: (1) the traditional dimension of increasing the number of candidate outputs sampled from a *generator* LLM, and (2) our novel test-time scaling dimension of increasing the number of *verifier* models evaluating each output. We illustrate the efficacy of MAV at improving test-time performance across multiple domains and generator LLMs, enabling weak-to-strong generalization and self-improvement. We hope that our work opens a new direction of research in combining multiple language models for effective reasoning and scaling verification compute, and we discuss additional limitations of our approach and opportunities for future research in Appendix A.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. *Advances in Neural Information Processing Systems*, 36:22304–22325, 2023.
- Kenneth J Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.
- Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.
- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse LLMs. *arXiv preprint arXiv:2309.13007*, 2023.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*, 2023.

- Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D’Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.
- Peter C Fishburn. *The theory of social choice*. Princeton University Press, 2015.
- Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.
- Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*, 2023.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Weijie J Su, Camillo Jose Taylor, and Tanwi Mallick. Multi-modal and multi-agent systems meet rationality: A survey. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.
- Jerry S Kelly. *Social choice theory: An introduction*. Springer Science & Business Media, 2013.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for “mind” exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Shuang Li, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, and Igor Mordatch. Composing ensembles of pre-trained models via iterative consensus. *arXiv preprint arXiv:2210.11522*, 2022a.
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022b.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.
- Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castri-cato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. Generative reward models. *arXiv preprint arXiv:2410.12832*, 2024.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A Plummer, Zhaoran Wang, and Hongxia Yang. Let models speak ciphers: Multiagent debate through embeddings. *arXiv preprint arXiv:2310.06272*, 2023.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551, 2023.

- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.
- Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024a.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 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)*, pp. 9426–9439, 2024a.
- Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024b.
- Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. Guiding language model reasoning with planning tokens. *arXiv preprint arXiv:2310.05707*, 2023a.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024c.
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023b.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

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

A Further Discussion

Multi-Agent Verification (MAV) introduces a promising dimension for scaling test-time compute: scaling the number of verifiers. In Section 4, we demonstrated that combining multiple verifiers enables more effective evaluation of candidate outputs, facilitates weak-to-strong generalization, and allows for self-improvement. However, our approach has important limitations and there are several opportunities for future work to explore.

First, our investigation is limited to a pool of 20 aspect verifiers based on just two base LLMs, and the design of our verifiers is constrained by our ability to come up with diverse verification strategies and relevant aspects. Future work could explore scaling to many more verifiers and try a more systematic exploration of the space of verifiers, potentially using LLMs themselves to generate diverse verification strategies and identify relevant aspects to verify. With better-engineered verifiers and more systematic exploration, we expect to observe stronger scaling patterns.

Second, our aggregation technique described in Section 3.2 uses a simple voting mechanism that directly sums the individual binary approvals from each verifier. This approach does not account for the confidence or relevance of each verifier, and verifiers do not observe each other’s decisions or feedback. Future works could explore more sophisticated aggregation methods such as confidence-weighted voting or allowing verifiers to engage in debate (Du et al., 2023) before producing an approval. Moreover, our current approach uses a static engineered set of verifiers \mathcal{M}^d for all questions in a domain d , even though it may be best to use fewer or different verifiers for specific questions. Future works could investigate dynamically selecting the best set of verifiers for particular problems or adaptively choosing additional verifiers based on the results of the first few verification queries. Additionally, the field of social choice theory (Arrow, 2012; Fishburn, 2015; Kelly, 2013; Brandt et al., 2016) is concerned with procedures for collective decision-making and might offer insights for aggregating the perspectives of diverse verifiers. Although, our setting differs in that we care more about verifier capabilities than preferences.

Next, our implementation of BoN-MAV is limited to only a single generator LLM. Thus, an interesting direction would be to explore sampling from multiple generators in addition to evaluating with multiple verifiers. Since different models may excel at solving different types of problems, this approach could make even better use of the growing ecosystem of LLMs and their diverse capabilities.

Furthermore, while our results show that BoN-MAV can improve language model performance at test-time, we did not investigate finetuning the generator LLM on the outputs selected by our verifiers. Similar to how prior works have finetuned on outputs selected through self-consistency (Huang et al., 2022) or reward models (Dong et al., 2023), training on outputs selected by MAV systems could be explored as a method to improve the generator LLM and also each of the LLM-based verifiers. Moreover, an interesting direction for future work is to directly use reinforcement learning to train both the generator and verifier models. That is, generator LLMs can be trained to maximize the scores across multiple verifiers, and the verifiers can simultaneously be trained to accurately verify individual aspects of responses.

Additionally, we used the strongest open-source reward model that could run on our academic compute, which ranked among the top 10 on RewardBench (Lambert et al., 2024) at the time of writing (outperforming many larger reward models). However, this model is not as powerful as some of the generator LLMs. It would be worthwhile for future work to compare against stronger reward models.

It’s also worth mentioning that our work also has a connection to Constitutional AI (CAI) (Bai et al., 2022b), which was one of the first major works demonstrating reinforcement learning from AI feedback (RLAIF) for LLMs, and focused on improving alignment through train-time RL (using the model’s own feedback and a set of constitutional principles). Our paper also uses LLM-based evaluation to improve performance, but has a different contribution: we propose scaling verification compute by increasing the number and type of verifiers, without any training. So CAI is an RLAIF train-time RL approach for alignment, whereas our work focuses on test-time scaling and introduces an approach where verification compute can be scaled using diverse verifiers without additional training.

An underlying thread throughout our work and discussion is the vision of a growing ecosystem of diverse language models that generate, verify, and learn from each other. Our work on multi-agent verification represents one step in this direction, and each of the future directions we have discussed offers a potential avenue for additional progress. We look forward to seeing how the research community advances these ideas.

B Experimental Setup

B.1 Domains

We evaluate our method on the following four domains:

- **Mathematics.** The MATH dataset (Hendrycks et al., 2021) consists of competition-level math questions at five difficulty levels. For our experiments, we randomly sample 400 questions from the test set across all five levels: 100 for validation and 300 for testing.
- **General Knowledge & Reasoning.** MMLU-Pro (Wang et al., 2024c) is an enhanced version of the popular MMLU benchmark (Hendrycks et al., 2020) which features more challenging, reasoning-focused questions and expands the multiple-choice set from four to ten options. As with MATH, we sample 100 questions for validation and 300 for testing.
- **Graduate-Level Reasoning.** The GPQA dataset (Rein et al., 2023) consists of graduate-level, multiple-choice questions in biology, physics, and chemistry. For our experiments, we utilize GPQA’s “diamond” subset — a collection of 198 high-quality and extremely challenging questions. We sample 98 questions for validation and 100 for testing.
- **Coding.** HumanEval (Chen et al., 2021) is a widely-used benchmark consisting of 164 Python programming questions. We sample 64 questions for validation and 100 for testing.

B.2 Verifier Engineering

For our experiments, we implement the verifier engineering method described in Section 3.4. To create our initial diverse pool \mathcal{M} of 20 aspect verifiers, we vary the three key axes that define aspect verifiers:

1. Base model: Gemini-1.5-Flash or GPT-4o-mini
2. Aspect to verify: Mathematical correctness, logical soundness, factuality, etc.
3. Verification strategy: Direct approval, step-by-step verification, solution rephrasing, edge case checking, etc.

From this pool, we then select domain-specific subsets $\mathcal{M}^d \subseteq \mathcal{M}$ that maximize average performance across all generator LLMs on the corresponding validation sets. The complete list of verifiers and the domain-specific subsets are detailed in Table 4. We choose Gemini-1.5-Flash and GPT-4o-mini as the base LLMs for our aspect verifiers since they are cost-effective for large-scale verification and enable us to demonstrate that combining multiple weaker verifiers can improve the performance of even stronger generator LLMs (Section 4.2).

B.3 Aspect Verifier Subsets

Table 4 outlines all 20 aspect verifiers in \mathcal{M} and which ones were selected for each domain-specific subset \mathcal{M}^d .

B.4 Generator LLMs

We evaluate eight generator LLMs (four closed-source models and four open-source models) and restrict our set of generator models to those released before September 2024. For closed-source models, we use gemini-1.5-flash-001 and gemini-1.5-pro-001 (Team et al., 2024a), as well as gpt-4o-mini-2024-07-18 and gpt-4o-2024-08-06 (Achiam et al., 2023). For open-source models, we use Mistral-7B-v0.3 (Jiang et al., 2023), Llama-3.1-8B (Dubey et al., 2024), Gemma-2-9B, and Gemma-2-27B (Team et al., 2024b).

Base Model	Aspect to Verify	Verification Strategy	MATH	MMLU-Pro	GPQA	HumanEval
GPT-4o-mini	Math Correctness	Step-by-Step		✓	✓	✓
	Logical Soundness	Step-by-Step		✓	✓	✓
	Factual Correctness	Step-by-Step				✓
	Unit Conversions	Step-by-Step	✓		✓	✓
	General Correctness	Direct Approval				✓
	General Correctness	Summarize Solution	✓			
	General Correctness	Explain Differently		✓	✓	✓
	General Correctness	Edge Cases	✓	✓		✓
	General Correctness	Common Mistakes	✓	✓		
	General Correctness	Domain Knowledge	✓	✓		✓
Gemini-1.5-Flash	Math Correctness	Step-by-Step				
	Logical Soundness	Step-by-Step				✓
	Factual Correctness	Step-by-Step				
	Unit Conversions	Step-by-Step		✓	✓	✓
	General Correctness	Direct Approval				✓
	General Correctness	Summarize Solution				✓
	General Correctness	Explain Differently			✓	✓
	General Correctness	Edge Cases	✓			
	General Correctness	Common Mistakes		✓	✓	
	General Correctness	Domain Knowledge				✓
Total Verifiers Used			6	8	7	14

Table 4: Overview of all aspect verifiers in \mathcal{M} . Checkmarks (✓) indicate which verifiers were selected for each domain-specific subset \mathcal{M}^d . The table shows all 20 combinations of base models, aspects to verify, and verification strategies that we created (10 per base model). The bottom row shows the number of verifiers $|\mathcal{M}^d|$ for each domain.

B.5 Reward Model Baseline

Our reward model verification baseline (BoN-RM) uses Skywork/Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024), the top scoring open-source 8B reward model on Reward-Bench (Lambert et al., 2024) at the time of writing. This pretrained reward model outperforms numerous larger models including 70B and 340B models, and can be run on academic-scale compute.

B.6 Prompts

For generator LLMs, we use a consistent prompt format across all models while varying the content by domain. Table 12 contains these domain-specific prompts. Note that we spent very little time engineering the prompts — just enough so that the output is in the desired format.

For aspect verifiers, each prompt consists of two components:

1. A domain-dependent system prompt (Table 13) that establishes the verification context (e.g., mathematical problems, multiple-choice questions, or code implementations)
2. A domain-independent verification prompt (Table 14 and Table 15) that specifies the aspect to verify and verification strategy

This two-part structure allows us to combine any aspect-strategy verification method with any domain while maintaining consistent evaluation criteria across base models.

C Additional Results

C.1 Using All Verifiers

Table 5 compares BoN-MAV using all 20 aspect verifiers in \mathcal{M} (without domain-specific engineering) against self-consistency and reward model verification. Even without engineering domain-specific subsets \mathcal{M}^d , combining all verifiers remains competitive with baseline methods.

Generator Model	MMLU-Pro				GPQA (diamond)			
	MAV-All	Cons	RM	pass@1	MAV-All	Cons	RM	pass@1
Gemini-1.5-Flash	65.7	63.3	60.7	59.3	41.0	40.0	46.0	42.0
Gemini-1.5-Pro	70.3	71.7	69.3	68.0	49.0	45.0	49.0	45.0
GPT-4o-mini	65.3	63.7	62.7	62.3	49.0	48.0	44.0	38.0
GPT-4o	75.3	76.3	72.7	73.3	55.0	59.0	58.0	54.0

Table 5: Performance (accuracy %) of BoN-MAV with all 20 aspect verifiers (without any tuning, labeled as MAV-all in the table) compared to reward model verification (RM), self-consistency (Cons), and the base pass@1 accuracy of the generator LLM. Using all verifiers without domain-specific tuning remains competitive with reward model verification and self-consistency.

C.2 Using AI-Generated Verifiers

Our aspect verifiers were initially created through a combination of AI assistance and manual prompt engineering, but having to hand-engineer every verifier would limit the generality and scalability of our approach. Here, we conduct an additional experiment using fully AI-generated aspect verifiers. Table 6 shows the performance of AI-generated verifiers compared to our original verifiers, self-consistency, and pass@1 baselines on MMLU-Pro and GPQA (diamond) using Gemini-1.5-Flash. We find that using AI-generated verifier prompts can perform even better, demonstrating that MAV can be fully automated. Future work could investigate more effective methods for automating aspect verifier generation and selection.

Specifically, we provided GPT-4o with just 3 example aspect prompts from our original set and asked it to generate 10 new prompts with different aspect-strategy combinations. Using these 10 purely AI-generated verifiers (without any manual engineering), we evaluated Gemini-1.5-Flash on MMLU-Pro and GPQA (diamond). For simplicity, we do not perform any verifier engineering for this experiment—we use all 10 AI-generated aspects without selection. So we compare to the “MAV-All” results from Table 5. Below is one example of the AI-generated aspects created by GPT-4o:

"redundant_steps": "Look for any unnecessary steps that don't contribute to the final result. Think out loud. If the solution includes redundant or misleading steps, reply 'False'. If every step is necessary and useful, reply 'True'."

Benchmark	MAV-Automated	MAV-All	Cons	pass@1
MMLU-Pro	68.0	65.7	63.3	59.3
GPQA (diamond)	47.0	41.0	40.0	42.0

Table 6: **AI-generated aspect verifiers.** Performance (accuracy %) comparison using Gemini-1.5-Flash as the generator LLM with $n = 16$ candidate outputs.

	Generator LLM	Rubric	MAV-All	Cons	pass@1
MMLU-Pro	Gemini-1.5-Flash	62.0	<u>65.7</u>	63.3	59.3
	GPT-4o-mini	64.0	<u>65.3</u>	63.7	62.3
GPQA (diamond)	Gemini-1.5-Flash	39.0	41.0	40.0	<u>42.0</u>
	GPT-4o-mini	42.0	<u>49.0</u>	48.0	38.0

Table 7: **Single rubric vs separate verifiers.** Performance (accuracy %) comparison between generating a single rubric for all aspects in just one verifier response (Rubric) versus using separate model calls per verifier aspect (MAV-All) with $n = 16$ candidate outputs.

C.3 Scoring a Rubric of Aspects

In MAV, the verifiers evaluate each aspect in a separate LLM call. Here, we try asking an LLM verifier to generate a single “rubric” containing the approval scores for each aspect (all in a *single* LLM call). We find that generating a single rubric for all aspects performs **worse** than using a separate model call per aspect (see Table 7). This is likely because having just one response forces the model to think about many different aspects in context, whereas separate model calls allow the LLM verifier to put all of its focus on one aspect at a time. Having multiple LLM verifier calls may have more expressive power, and this could be an interesting direction for future work to explore further.

Specifically, we instruct the LLM “rubric” verifier to think about each aspect in a single response and generate a python dictionary mapping aspect names to True/False values (after it has finished thinking about each aspect). We use the exact same 10 aspects as in the BoN-MAV results, but obtain the approvals for all aspects simultaneously in a single response rather than splitting across model calls. As in the BoN-MAV results, we use Gemini-1.5-Flash and GPT-4o-mini as the base LLMs for our verifiers, so there are actually 2 model calls: Gemini-1.5-Flash generates one rubric and GPT-4o-mini generates another rubric. For simplicity, we do not perform any verifier engineering. So we compare to the “MAV-All” results in Table 5 (which uses 20 separate verifier calls: 10 aspects for 2 base LLMs).

C.4 Scaling Generator CoT

To compare our approach to a “thinking” or long-Chain-of-Thought (long-CoT) method which scales generator LLM reasoning tokens *without* test-time verifiers, we implemented a reasoning baseline where we iteratively prompt the same base models to continuously extend their chain of reasoning until we use a similar number of tokens as our MAV approach on the MATH dataset (approximately 80k tokens per question on average). This comparison is appropriate since both methods scale test-time compute using zero-shot prompting without fine-tuning for mathematical reasoning or using RLVR. Table 8 shows that when controlling for token count, our method performs better than simply scaling reasoning tokens within a single model call.

Model	BoN-MAV (ours)	Scaling Reasoning Tokens	pass@1
Gemini-1.5-Flash	<u>66.0</u>	53.0	52.7
GPT-4o-mini	<u>73.0</u>	72.3	69.0

Table 8: **Comparison with scaling reasoning tokens on MATH.** Performance (accuracy %) comparison between our BoN-MAV approach and a baseline that scales reasoning tokens within a single model call, controlling for similar token counts. Our method outperforms scaling reasoning tokens when given a similar computational budget.

C.5 Process Reward Models (PRMs)

For our baselines, we focused on outcome-level reward models and self-consistency since they are most similar to our approach (selecting between complete candidate outputs). However, here we perform a preliminary experiment using Process Reward Models (PRMs) that select between candidate responses at each *step* of the solution. Note that this represents a fundamentally different approach that operates at the step level rather than the response level. Using GPT-4o-mini with Qwen2.5-Math-PRM and a branching factor of 4, we achieved 68.0% accuracy on MATH, which is higher than the pass@1 base accuracy but below our BoN-MAV accuracy of 73.0%. This comparison is not entirely appropriate since PRMs focus on step-wise verification during generation using a single verifier, while our approach focuses on selecting between complete responses using multiple verifiers. A more direct comparison would involve using MAV-like approaches to select between intermediate steps using multiple verifiers, which we leave for future work.

C.6 Averaging Different Reward Scores

Here, we include an experiment that demonstrates the challenge of combining heterogeneous reward models that operate on different scales and have been trained on different datasets. Specifically, we average the reward values from three different reward models: Skywork/Skywork-Reward-Llama-3.1-8B-v0.2, nicolinho/QRM-Llama3.1-8B-v2, and allenai/Llama-3.1-8B-Instruct-RM-RB2. For each candidate solution, we obtain reward scores from all three models and average them to create a combined reward signal for best-of- n selection. Using GPT-4o-mini as the generator LLM with $n = 16$ candidate solutions, we achieve 72.7% accuracy on MATH and 63.3% on MMLU-Pro. This is similar to our single reward model baseline (72.3% and 62.7% respectively) but worse than BoN-MAV (73.0% and 67.0% respectively). So our multi-agent verification approach with diverse aspect verifiers outperforms combining heterogeneous reward models in these cases. See Section 3.1 for a discussion of the challenges of combining neural reward models trained on different datasets.

C.7 Oracle Baselines

We include oracle baselines for each model across all four benchmarks in Table 9 and Table 10.

Model	MMLU-Pro			GPQA (diamond)		
	pass@n	B-MAV	Cons	pass@n	B-MAV	Cons
Gemini-1.5-Flash	81.0	66.7	63.3	80.0	42.0	40.0
Gemini-1.5-Pro	85.7	72.3	71.7	86.0	49.0	45.0
GPT-4o-mini	80.3	67.0	63.7	79.0	50.0	48.0
GPT-4o	88.3	75.7	76.3	87.0	59.0	59.0
Mistral-7B	43.3	36.7	25.7	54.0	36.0	32.0
Llama-3.1-8B	87.0	59.3	55.3	90.0	43.0	36.0
Gemma-2-9B	74.0	57.7	54.3	69.0	34.0	36.0
Gemma-2-27B	81.0	62.0	58.3	81.0	41.0	40.0

Table 9: **Oracle baseline (pass@n) results for MMLU-Pro and GPQA (diamond).** Performance (accuracy %) with $n = 16$ candidate outputs showing the upper bound performance (pass@n), BoN-MAV performance (from Table 1), and self-consistency (Cons) baseline for MMLU-Pro and GPQA (diamond).

Model	MATH			HumanEval		
	pass@n	B-MAV	Cons	pass@n	B-MAV	Cons
Gemini-1.5-Flash	76.0	66.0	59.0	82.0	80.0	79.0
Gemini-1.5-Pro	85.7	72.7	70.3	92.0	88.0	84.0
GPT-4o-mini	85.3	73.0	74.7	91.0	84.0	87.0
GPT-4o	87.7	76.3	77.3	98.0	92.0	95.0
Mistral-7B	38.3	26.0	22.0	67.0	59.0	46.0
Llama-3.1-8B	78.0	61.7	61.0	87.0	75.0	62.0
Gemma-2-9B	70.0	58.7	51.7	56.0	32.0	25.0
Gemma-2-27B	75.7	62.3	55.7	84.0	76.0	66.0

Table 10: **Oracle baseline (pass@n) results for MATH and HumanEval.** Performance (accuracy %) with $n = 16$ candidate outputs showing the upper bound performance (pass@n), BoN-MAV performance (from Table 1), and self-consistency (Cons) baseline for MATH and HumanEval.

D Additional Illustrations and Extra Tables

Figure 7, Figure 8, and Figure 9 provide additional examples of how multiple aspect verifiers evaluate a single candidate output. Figure 7 demonstrates verification using multiple strategies with a single base model on MATH (Hendrycks et al., 2021). Figure 8 shows verification of a coding solution from HumanEval (Chen et al., 2021). Figure 9 illustrates verification of a correct solution from GPQA (diamond) (Rein et al., 2023), showing how different base models can assess the same aspect differently. Each figure follows the same format as Figure 2 from the main paper.

Table 11 shows the accuracies for MATH when only counting questions with pass@n=1 (where at least one sampled solution is correct).

Model	MAV	Cons	RM
Gemini-1.5-Flash	<u>87.3</u>	77.6	81.6
Gemini-1.5-Pro	<u>87.2</u>	82.1	82.9
GPT-4o-mini	86.3	<u>87.5</u>	84.8
GPT-4o	88.2	88.2	<u>92.0</u>
Mistral-7B	<u>67.0</u>	57.4	56.5
Llama-3.1-8B	<u>80.8</u>	78.2	70.1
Gemma-2-9B	<u>82.4</u>	73.8	78.6
Gemma-2-27B	<u>83.3</u>	73.6	78.4

Table 11: **Accuracy on MATH for questions where at least one sampled solution is correct.** Performance (accuracy %) when only counting questions where pass@n=1. We compare BoN-MAV (B-MAV), self-consistency (Cons), and reward model (RM) selection with $n = 16$ candidate outputs.

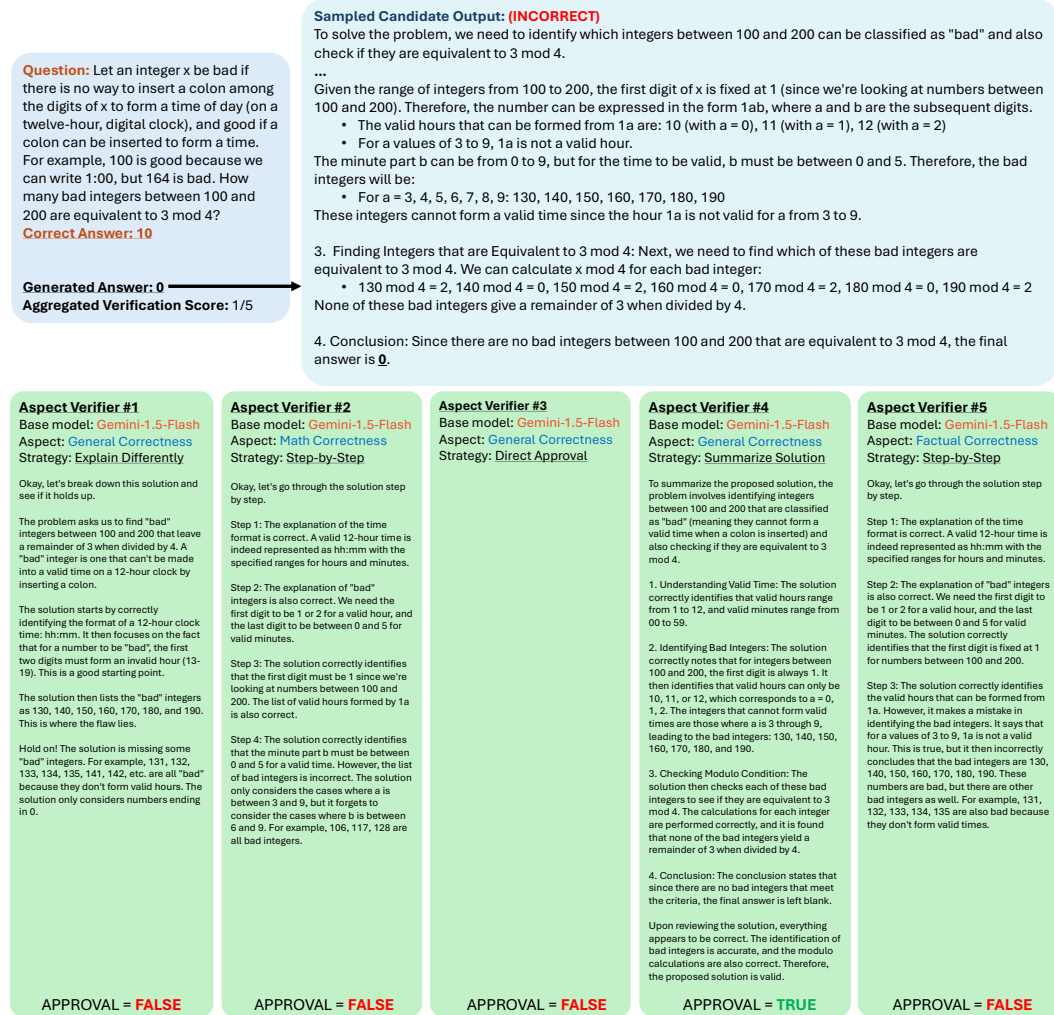


Figure 7: **Multi-agent verification for a single solution (additional example).** An illustration similar to Figure 2. Five different aspect verifiers evaluate an incorrect MATH (Hendrycks et al., 2021) solution sampled from Gemini-1.5-Pro. All verifiers use Gemini-1.5-Flash as the base model but vary in their aspects to verify (e.g., general correctness, mathematical correctness) and verification strategies (e.g., direct approval, step-by-step verification). Four verifiers correctly identify the error, while one verifier using general correctness through summarization incorrectly approves the solution. This demonstrates how diverse verification methods can produce more reliable signals even when using a single base model, as multiple verifiers can compensate when another fails.

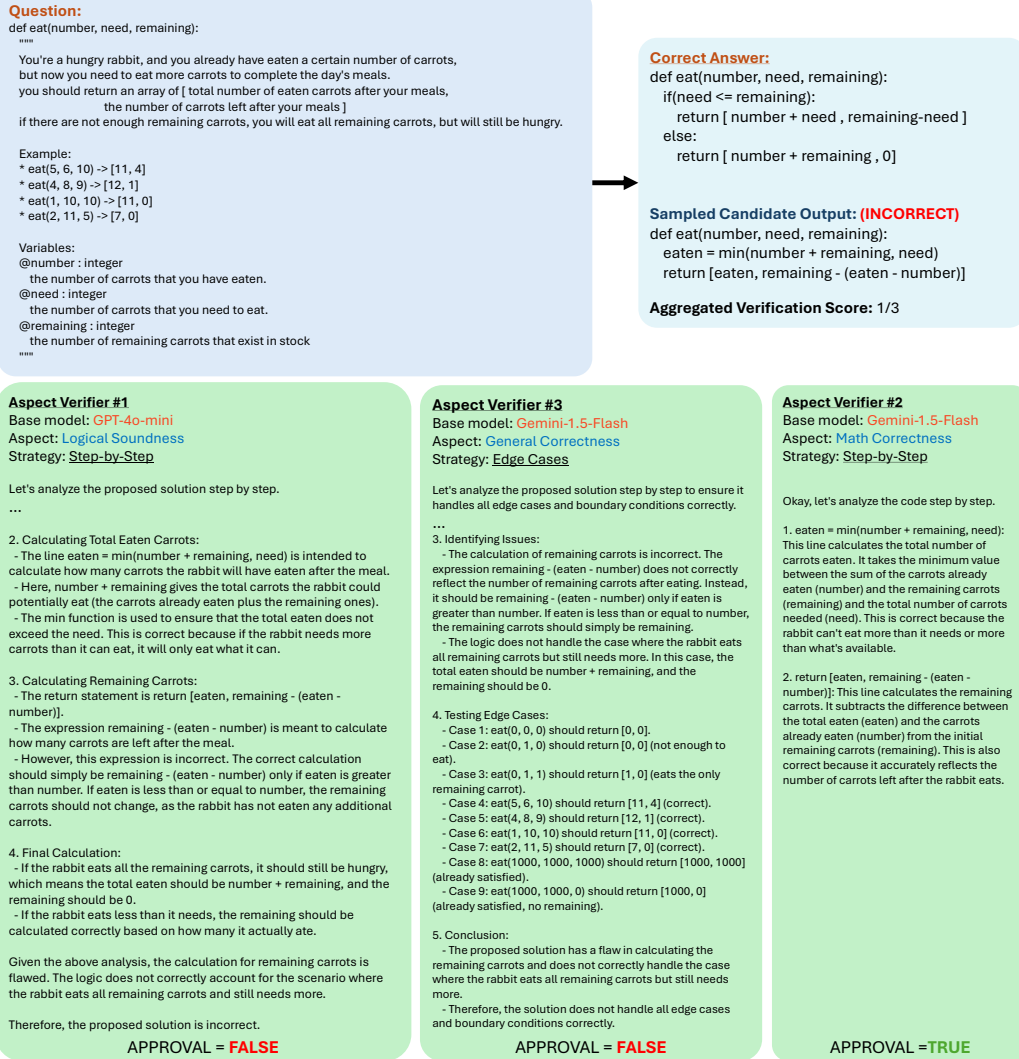


Figure 8: **Multi-agent verification for a single solution (additional example).** An illustration similar to Figure 2. Here, three different aspect verifiers evaluate an incorrect HumanEval (Chen et al., 2021) solution sampled from Gemini-1.5-Flash. Two verifiers correctly identify the error through careful analysis, while one verifier incorrectly approves the solution.

Question: You identified a new ligand-receptor pair initially expressed in the neural crest cells and want to understand the developmental fate of their interaction. The receptor is a monomeric transmembrane barrel with eight alpha helices, four extracellular helices, and three beta sheets that create a binding site for the ligand. The ligand is a coiled-coil of two alpha helices. You create a bicistronic lox-Cre vector with the ligand ORF at the 5' position, fused to mCherry; the receptor ORF is at the 3' and fused to the eGFP at the C-terminus. The ligand is under the CBA promoter, and upstream of the receptor is the IRES regulatory element. Between the ligand ORF and mCherry, there is a stop in a loxP-stop-loxP cassette, and between the receptor and the eGFP, there is a stop in the lox2272-stop-lox2272 cassette. You had also raised antibodies against the overexpressed proteins and now use them to test the construct expression on the Western blot after plasmid transfection to primary astrocyte culture. Total proteins give a signal proportional to the actin loading. You finally breed a mouse model homozygous for your construct in the wt background. You cross it to SOX10-Cre hemizygous mice; after the offspring is born you visualize them under the fluorescent confocal microscope. You do not observe a green signal. Which of the answers below explains the most likely reason for it?

Options:

- (A) ligand and the receptor are in a paracrine relationship,
- (B) the receptor-eGFP construct is stuck in the Golgi,
- (C) the enhancer for the ligand and receptor expression is missing,
- (D) the receptor and the eGFP are not in the frame

Correct Answer: (D)

Generated Answer: (D)

Aggregated Verification Score: 2/3

Sampled Candidate Output: (CORRECT)

Here's a step-by-step analysis:

1. The experimental setup: The researchers are trying to analyze the interaction of a new ligand-receptor pair specifically in neural crest cells. They've created a bicistronic vector that expresses both the ligand (fused to mCherry) and the receptor (fused to eGFP) under the control of specific promoters and regulatory elements.
2. The Cre-lox system: The vector utilizes the Cre-lox system for conditional gene expression. The ligand expression is controlled by a loxP-stop-loxP cassette, while the receptor expression is controlled by a lox2272-stop-lox2272 cassette. This means that expression of both genes is blocked until Cre recombinase is present.
3. Crossing with SOX10-Cre mice: The researchers crossed their transgenic mice with SOX10-Cre mice. SOX10 is a transcription factor expressed in neural crest cells. This cross is intended to activate Cre recombinase specifically in neural crest cells, thereby removing the stop cassettes and allowing expression of both the ligand and receptor.
4. Lack of green signal: Despite the setup, the researchers didn't observe a green signal (indicating eGFP expression) under the microscope.
5. Analyzing the options:
 - * A. Paracrine relationship: While the ligand and receptor could be in a paracrine relationship (acting on nearby cells), this wouldn't necessarily prevent the expression of the eGFP-tagged receptor.
 - * B. Receptor-eGFP stuck in the Golgi: This is a possibility, but it's less likely given that the construct was designed with a transmembrane domain for the receptor, which should direct it to the cell surface.
 - * C. Missing enhancer: While a missing enhancer could lead to low expression levels, it's unlikely to completely prevent expression.
 - * D. Receptor and eGFP out of frame: This is the most likely explanation. If the receptor and eGFP coding sequences are out of frame, the eGFP will not be translated correctly, resulting in the lack of a green signal.

The answer is **(D)**

Aspect Verifier #1

Base model: **GPT-4o-mini**

Aspect: **Logical Soundness**

Strategy: **Step-by-Step**

Let's analyze the proposed solution step by step:

1. The experimental setup: The description of the experimental setup is accurate. The researchers are indeed trying to analyze the interaction of a ligand-receptor pair in neural crest cells using a bicistronic vector. This step is logically sound.
2. The Cre-lox system: The explanation of the Cre-lox system is correct. The use of loxP-stop-loxP and lox2272-stop-lox2272 cassettes to control the expression of the ligand and receptor is well articulated. This step is also logically sound.
3. Crossing with SOX10-Cre mice: The explanation regarding the crossing with SOX10-Cre mice is accurate. SOX10 is indeed expressed in neural crest cells, and the intention to activate Cre recombinase in these cells to remove the stop cassettes is correctly stated. This step is logically sound.
4. Lack of green signal: The observation of no green signal (indicating eGFP expression) is noted correctly. This sets the stage for analyzing potential reasons for the lack of expression. This step is logically sound.
5. Analyzing the options:
 - A. Paracrine relationship: The reasoning here is sound. While a paracrine relationship could exist, it does not explain the lack of eGFP expression. This analysis is logically sound.
 - B. Receptor-eGFP stuck in the Golgi: The proposed solution states that this is a possibility but less likely due to the design of the construct. This reasoning is reasonable, but it does not fully rule out the possibility. However, it does not directly address the lack of signal. This step is somewhat less definitive but still logically sound.
 - C. Missing enhancer: The proposed solution suggests that a missing enhancer could lead to low expression but is unlikely to completely prevent expression. This reasoning is sound, as it acknowledges the potential for low expression without ruling it out completely. This step is logically sound.
 - D. Receptor and eGFP out of frame: The proposed solution concludes that this is the most likely explanation for the lack of green signal. If the receptor and eGFP are out of frame, it would indeed prevent proper translation of the eGFP, leading to no signal. This reasoning is logically sound.

Since all steps in the proposed solution are logically sound and lead to a coherent conclusion, I will conclude my verification.

APPROVAL = TRUE

Aspect Verifier #2

Base model: **Gemini-1.5-Flash**

Aspect: **Logical Soundness**

Strategy: **Step-by-Step**

Okay, let's go through the solution step-by-step.

- Step 1: The experimental setup is accurately described. The researchers are trying to understand the interaction of a new ligand-receptor pair in neural crest cells using a bicistronic vector with specific promoters and regulatory elements. This step is logically sound.
- Step 2: The Cre-lox system is correctly explained. The loxP-stop-loxP and lox2272-stop-lox2272 cassettes are designed to block expression until Cre recombinase is present. This step is also logically sound.
- Step 3: The crossing with SOX10-Cre mice is intended to activate Cre recombinase specifically in neural crest cells, removing the stop cassettes and allowing expression of both the ligand and receptor. This step is logically sound.
- Step 4: The lack of green signal is the observed result, and it is the problem the researchers are trying to explain. This step is logically sound.
- Step 5: The analysis of the options is where things get tricky.
- * Option A: A paracrine relationship wouldn't prevent the expression of the eGFP-tagged receptor. This is a valid point.
 - * Option B: The receptor being stuck in the Golgi is possible, but less likely given the transmembrane domain. This is also a valid point.
 - * Option C: A missing enhancer could lead to low expression, but not necessarily complete prevention. This is a valid point.
 - * Option D: This is where the logic breaks down. If the receptor and eGFP are out of frame, the eGFP would not be translated correctly, resulting in the lack of a green signal. This is a correct explanation for the lack of green signal. However, the solution states that this is the most likely explanation. This is where the logic falters.

APPROVAL = FALSE

Aspect Verifier #3

Base model: **Gemini-1.5-Flash**

Aspect: **General Correctness**

Strategy: **Explain Differently**

Okay, let's break down the solution in a different way.

The experiment aims to study the interaction of a ligand-receptor pair in neural crest cells. The researchers created a bicistronic vector with the ligand and receptor genes, each fused to a fluorescent protein (mCherry and eGFP, respectively). The vector uses a Cre-lox system to control gene expression, ensuring that both genes are expressed only in neural crest cells where SOX10-Cre is active.

The problem is that the researchers don't see any green fluorescence, indicating a lack of eGFP expression.

Let's analyze the options:

- * A. Paracrine relationship: This is irrelevant to the lack of eGFP expression. A paracrine relationship describes how the ligand and receptor interact, not whether the receptor is expressed.
- * B. Receptor-eGFP stuck in the Golgi: This is possible, but it's unlikely. The receptor is designed with a transmembrane domain, which should direct it to the cell membrane, not the Golgi.
- * C. Missing enhancer: This could lead to low expression, but it's unlikely to completely prevent expression. The vector uses a strong promoter (CBA) for the ligand, and the IRES element should allow for efficient translation of the receptor.
- * D. Receptor and eGFP out of frame: This is the most likely explanation. If the receptor and eGFP coding sequences are not in the correct reading frame, the eGFP protein won't be translated properly, leading to the lack of green fluorescence.

APPROVAL = TRUE

Figure 9: Multi-agent verification for a single solution (additional example). An illustration similar to Figure 2. Here, three different aspect verifiers evaluate a correct GPQA (diamond) (Rein et al., 2023) solution sampled from Gemma-2-27B. The verifiers vary in their base models, aspects to verify, and verification strategies. Notice that Gemini-1.5-Flash incorrectly rejects the solution when evaluating logical soundness but correctly approves it when prompted to explain the solution differently. Meanwhile, GPT-4o-mini correctly approves the solution when evaluating logical soundness. Different base models can produce different evaluations of the same aspect.

Domain	Generator Prompt
MATH	<i>You are a helpful assistant skilled in math problem-solving. Always end your solution with the final numerical answer enclosed in LaTeX <code>\boxed{}</code> notation. If there is no solution, reply with an empty <code>\boxed{}</code>. Please solve the following math problem step by step: < Question > Provide your detailed solution below:</i>
MMLU-Pro	<i>Answer the following multiple choice question. Think step by step before answering, and then output the answer in the format of “The answer is (X)” at the end, where X is the LETTER of the correct answer. QUESTION: < Question > Think step by step, then end with EXACTLY “The answer is (X)”, where X is the LETTER of the correct answer. Do not include the answer text itself, only the letter.</i>
GPQA (diamond)	Same as MMLU-Pro.
HumanEval	<i>Read the following function signature and docstring, and fully implement the function described. Your response should only contain the code for this function. < Function Signature and Docstring ></i>

Table 12: **Generator Prompts.** Generator prompts by domain. Each domain uses one consistent prompt across all generator LLMs.

Domain	Aspect Verifier System Prompt
MATH	<i>You are a critical verifier tasked with evaluating mathematical problem-solving. You will be presented with a question and a proposed solution. Your job is to carefully go over and analyze the solution. Follow the instructions.</i>
MMLU-Pro	<i>You are a critical verifier tasked with evaluating multiple-choice question-answering. You will be presented with a question, the multiple-choice options, and a proposed solution. Your job is to carefully go over and analyze the solution. Follow the instructions.</i>
GPQA (diamond)	Same as MMLU-Pro.
HumanEval	<i>You are a critical verifier tasked with evaluating code implementations. You will be presented with a prompt and a code implementation. Your job is to carefully go over and analyze the code. Follow the instructions.</i>

Table 13: **Aspect Verifier System Prompts.** System prompts for aspect verifiers. These provide domain-specific context for the verification instructions in [Table 14](#) and [Table 15](#).

Aspect to Verify	Verification Strategy	Aspect Verifier Prompt
Math Correctness	Step-by-Step	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Go over each step in the proposed solution and check whether it is mathematically correct. Think out loud. If you reach a step that is incorrect, stop and reply 'FINAL VERIFICATION ANSWER: False'. If you get to the end of all the steps and each step was correct, reply 'FINAL VERIFICATION ANSWER: True'.</p>
Logical Soundness	Step-by-Step	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Go over each step in the proposed solution and check whether it is logically sound. Think out loud. If you reach a step that is not logically sound, stop and reply 'FINAL VERIFICATION ANSWER: False'. If you get to the end of all the steps and each step was logically sound, reply 'FINAL VERIFICATION ANSWER: True'.</p>
Factual Correctness	Step-by-Step	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Go over each step in the proposed solution and check whether the facts presented are correct. Think out loud. If you reach a step with incorrect facts, stop and reply 'FINAL VERIFICATION ANSWER: False'. If you get to the end of all the steps and each step had correct facts, reply 'FINAL VERIFICATION ANSWER: True'.</p>
Unit Conversions	Step-by-Step	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Check if the units are handled correctly in each step of the solution. Think out loud. If you find any issues with the units, stop and reply 'FINAL VERIFICATION ANSWER: False'. If all units are handled correctly, reply 'FINAL VERIFICATION ANSWER: True'.</p>

Table 14: **Aspect Verifier Prompts (Part 1)**. Aspect verifier prompts for each aspect-strategy combination. These prompts follow the system prompts in [Table 13](#).

Aspect to Verify	Verification Strategy	Aspect Verifier Prompt
General Correctness	Direct Approval	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Is this solution correct for the given question? Respond with ONLY 'FINAL VERIFICATION ANSWER: True' or ONLY 'FINAL VERIFICATION ANSWER: False'. Do not provide any explanation or additional text.</p>
General Correctness	Summarize Solution	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Summarize the solution in your own words, explore anything you think may be incorrect. Think out loud. If you find something that's incorrect, stop and reply 'FINAL VERIFICATION ANSWER: False'. If you've gone over the solution and everything seems correct, reply 'FINAL VERIFICATION ANSWER: True'.</p>
General Correctness	Explain Differently	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Explain the solution in a different way than it was presented. Try to find any flaws in the solution. Think out loud. If you find something that's incorrect, stop and reply 'FINAL VERIFICATION ANSWER: False'. If you've gone over the solution and everything seems correct, reply 'FINAL VERIFICATION ANSWER: True'.</p>
General Correctness	Edge Cases	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Check if the solution handles edge cases and boundary conditions, test extreme values or special cases. Think out loud. If any boundary conditions or edge cases fail, stop and reply 'FINAL VERIFICATION ANSWER: False'. If all boundary conditions and edge cases are handled correctly, reply 'FINAL VERIFICATION ANSWER: True'.</p>
General Correctness	Common Mistakes	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Check if the solution has any common mistakes, calculation errors, or misconceptions that typically found in this type of problem. Think out loud. If you find any common mistakes, stop and reply 'FINAL VERIFICATION ANSWER: False'. If no common mistakes are found, reply 'FINAL VERIFICATION ANSWER: True'.</p>
General Correctness	Domain Knowledge	<p>QUESTION: <Question> PROPOSED SOLUTION: <Solution> INSTRUCTIONS: Check if the solution correctly applies relevant domain-knowledge, established theories, and standard practices for this type of problem. Think out loud. If any domain knowledge is misapplied or violated, stop and reply 'FINAL VERIFICATION ANSWER: False'. If all domain-specific knowledge is correctly applied, reply 'FINAL VERIFICATION ANSWER: True'.</p>

Table 15: **Aspect Verifier Prompts (Part 2).** Aspect verifier prompts for each aspect-strategy combination. These prompts follow the system prompts in [Table 13](#).