

EGSS: Entropy-guided Stepwise Scaling for Reliable Software Engineering

Chenhui Mao* Yuanting Lei* Zhixiang Wei Ming Liang Zhixiang Wang
Jingxuan Xu Dajun Chen Wei Jiang† Yong Li†

Ant Group

{maochenhui.maochen, leiyuanting.lyt, weizhixiang.wzx, liangming.liang,
tiejing.wzx, xujingxuan.xjx, chendajun.cdj, jonny.jw, liyong.liy}@antgroup.com

Abstract

Agentic Test-Time Scaling (TTS) has delivered state-of-the-art (SOTA) performance on complex software engineering tasks such as code generation and bug fixing. However, its practical adoption remains limited due to significant computational overhead, primarily driven by two key challenges: (1) the high cost associated with deploying excessively large ensembles, and (2) the lack of a reliable mechanism for selecting the optimal candidate solution—ultimately constraining the performance gains that can be realized. To address these challenges, we propose **Entropy-Guided Stepwise Scaling (EGSS)**, a novel TTS framework that dynamically balances efficiency and effectiveness through entropy-guided adaptive search and robust test-suite augmentation. Extensive experiments on SWE-Bench-Verified demonstrate that EGSS consistently boosts performance by 5–10% across all evaluated models. Specifically, it increases the resolved ratio of Kimi-K2-Instruct from **63.2%** to **72.2%**, and GLM-4.6 from **65.8%** to **74.6%**, achieving new state-of-the-art among open-source large language models with GLM-4.6. Crucially, EGSS exhibits superior cost-efficiency: it outperforms the baseline configured with $K = 8$ candidates using only $K = 4$, resulting in a **38–42%** reduction in token consumption. Code is available here ¹.

1 Introduction

Autonomous Software Engineering (ASE) has broadened the application scope of large language models (LLMs) for code from generating simple functions to resolving complex, repository-level software issues. In contrast to snippet-level tasks (Chen, 2021), benchmarks such as SWE-bench (Jimenez et al., 2023) require agents to

*Equal contribution

†Corresponding authors.

¹<https://github.com/codefuse-ai/CodeFuse-Agent>

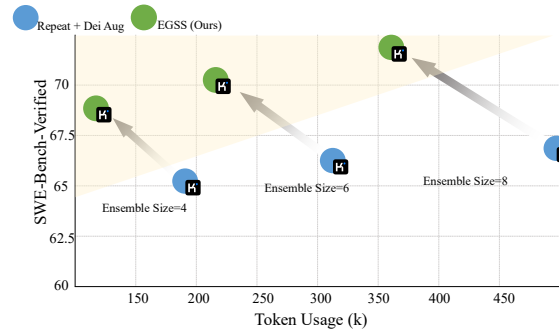


Figure 1: Performance and token usage of popular test-time scaling methods compared with entropy-guided stepwise scaling on SWE-Bench Verified, using Kimi-K2-Instruct as the base model.

operate within large-scale codebases, thereby necessitating long-horizon reasoning, proficient tool utilization, and a holistic understanding of cross-module dependencies (Yang et al., 2024; Wu, 2024; Bouzenia et al., 2024). As task complexity increases, the expanding action space and the inherent uncertainty of debugging trajectories substantially impede effective decision-making, often resulting in diminished reliability.

Within this demanding landscape, Test-Time Scaling (TTS) has emerged as a pivotal paradigm, demonstrating substantial potential to recover performance on complex problems by strategically allocating additional inference-time compute to explore diverse reasoning trajectories (Zeng et al., 2025; Gao et al., 2025). However, despite its promise, existing TTS strategies often entail considerable computational overhead and exhibit limited reliability. **First**, prevailing TTS approaches frequently suffer from pronounced computational redundancy and low exploration efficiency (Li et al., 2025a). They commonly adopt uniform search expansion or large-scale repetitive sampling without explicit inter-trajectory coordination, thereby expending resources on trivial operations and repeatedly traversing unproductive branches. Moreover, many frameworks (Gandhi et al., 2025) employ

static process-level guidance that does not adapt to the time-varying uncertainty inherent in complex tasks. This induces a resource–uncertainty mismatch, wherein computation is over-allocated to relatively straightforward instances while insufficient exploration is devoted to high-ambiguity decision points. **Second**, existing patch selection mechanisms often discard informative intermediate debugging signals (Wang et al., 2025). By disregarding execution traces, terminal-only ranking methods are susceptible to "consensus errors" and may favor brittle solutions that satisfy localized checks without addressing the underlying root cause. These limitations motivate the following core question:

*Can we design **principled** methods to make TTS both **efficient** and **reliable** within the context of ASE?*

To address these challenges, we propose Entropy-Guided Stepwise Scaling (EGSS), a principled, entropy-guided framework that jointly optimizes efficiency and reliability for ASE, distinct from prior scaling methods like ARPO (Dong et al., 2025) and Satori-SWE (Zeng et al., 2025).

First, to mitigate redundant exploration, EGSS introduces Dynamic Stepwise Search (DSS). Unlike existing methods that primarily focus on training expansion or uniform sampling (Zeng et al., 2025; Gao et al., 2025), DSS leverages **tool entropy** to dynamically constrain test-time compute. By identifying high-ambiguity decision points, it selectively invokes an auxiliary judge, thereby diverging from static process-guided approaches and ensuring computation is allocated only where uncertainty is highest. This concept of using entropy as a compute-constraint mechanism has been highlighted as a key differentiator in recent discussions (Dong et al., 2025).

Second, to enhance patch selection reliability, we propose Test Consolidation Augmentation (TCA). While prior works often discard intermediate debugging signals or rely on terminal-only validation (Wang et al., 2025), TCA uniquely synthesizes fragmented and potentially incompatible debugging signals from multiple trajectories into a unified, executable test suite. Ablation studies confirm that this consolidation mechanism drives performance gains beyond standard validation by resolving self-deceptive debugging (Discussed in section 3.2) through cross-trajectory verification.

Overall, our primary contributions are summarized as follows:

- We present a principled, entropy-guided scaling paradigm that improves inference efficiency by allocating inference budget to high-ambiguity decision points.
- We develop a robust patch selection pipeline that consolidates debugging signals into a high-coverage test suite, mitigating false positives and prioritizing robust candidate patches.
- We propose EGSS, which improves the performance of all evaluated models by 5–10% on SWE-Bench-Verified, establishing a new state of the art among open-source methods. Notably, EGSS achieves these gains with an ensemble size of $K = 4$ while outperforming baselines at $K = 8$, resulting in a **38–42%** reduction in total token consumption.

2 Related Work

2.1 Scaling Paradigms and Search Efficiency

Previous agentic frameworks (Yang et al., 2024; Xie et al., 2024; Wu, 2024) established repository-level repair protocols but primarily relied on single, linear reasoning trajectories. To address this limitation, Test-Time Scaling has emerged as a prominent paradigm that improves performance by allocating additional inference-time computation to explore a broader solution space (Ma et al., 2025; Li et al., 2025a). While existing methods employ heuristic search (Antoniades et al., 2024) or sampling in tool-interactive, non-deterministic environments (Zainullina et al., 2025), they often exhibit substantial redundancy under uniform scaling, expanding the search space regardless of step-level decision criticality. Moreover, although recent process-guided models (Gandhi et al., 2025; Li et al., 2025c; Sun et al., 2025) provide intermediate feedback, their reliance on static diagnostics (Khalifa et al., 2025) limits their ability to adaptively identify high-uncertainty decision points.

2.2 Reliability in Patch Selection

A key bottleneck in multi-trajectory test-time scaling is the limited reliability of final patch selection. Common strategies, ranging from majority voting (Xia et al., 2024) to multi-agent collaboration (Zhang et al., 2024b), can mitigate sampling-induced errors but remain vulnerable to consensus errors, where multiple trajectories converge on a plausible yet incorrect fix. More advanced selector agents (Zeng et al., 2025; Gao et al., 2025)

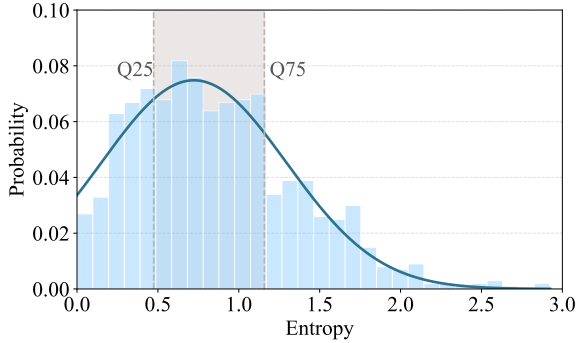


Figure 2: Tool entropy distribution along agent trajectories in SWE-Bench cases

and trajectory calibration methods (Wang et al., 2025) attempt to rank candidates, but they are still predominantly terminal-stage (post-hoc) selectors. Consequently, they treat trajectories as isolated candidates and discard rich intermediate debugging signals generated during the reasoning process (Wang et al., 2024b). Even with specialized verifiers (Pan et al., 2024; Kong et al., 2025), the absence of cross-trajectory signal synthesis can lead to brittle patches.

Our proposed TCA addresses this gap by explicitly consolidating heterogeneous debugging traces. Unlike simple voting mechanisms, TCA constructs a unified test suite that captures diverse edge cases identified across different trajectories, effectively turning conflicting or incompatible debugging signals into a robust verification standard. This approach fundamentally differs from prior selection strategies that lack such explicit signal synthesis capabilities.

3 Preliminary

In this section, we conduct a systematic analysis of an agent instantiated with Kimi-K2-Instruct and evaluate its performance on SWE-Bench tasks (Jimenez et al., 2023). The results reveal several salient empirical phenomena, which provide direct theoretical and practical motivation for the design of the proposed Entropy-guided Stepwise Scaling architecture.

3.1 Tool Use Entropy Heterogeneity

We quantify per-step uncertainty in tool selection along an execution trajectory using tool entropy:

$$H_{tool}(a_t | s_t) = - \sum_{a \in \mathcal{A}} P(a | s_t) \log P(a | s_t) \quad (1)$$

where $P(a | s_t)$ denotes the agent policy over a finite tool/action set \mathcal{A} conditioned on the current state s_t .

Empirically (Figure 2), H_{tool} exhibits a right-skewed distribution, with probability mass (Detailed statistical analysis are provided in Appendix A.1) concentrated in a low-entropy regime (Wang et al., 2024a). This pattern indicates that many trajectory steps correspond to near-deterministic routine operations (e.g., reading or editing files). In contrast, the sparse long tail corresponds to semantically consequential branching points—such as selecting among functionally similar tools—at which uncertainty increases substantially. This distributional imbalance implies that uniform exploration is computationally inefficient. Consequently, an effective TTS paradigm should **adaptively allocate resources** by reducing computation in high-density, low-entropy regions while prioritizing the infrequent yet high-entropy decision points in the tail.

3.2 Self-Deceptive Debugging

	Patch Final Outcome	
	Correct	Wrong
With Self-Debugging (w/)	283	157
Without Self-Debugging (w/o)	43	13

Table 1: Debugging process impact on patch outcomes (Kimi-K2-Instruct, SWE-Bench-Verified)

We further investigate behavioral patterns during task execution. As shown in Table 1, on the SWE-Bench-Verified benchmark (Jimenez et al., 2023) comprising 500 tasks, 88.0% (440/500) of trajectories exhibit explicit self-verification procedures, wherein the agent proactively generates tests, executes them, and iteratively refines the proposed patch based on the resulting feedback. Notably, among these 440 trajectories that incorporate self-verification, 35.7% (157/440) nevertheless fail to yield a correct patch. This finding indicates that existing single-trajectory self-verification is frequently confined to a single reasoning perspective and lacks multidimensional assessment of patch correctness and robustness.

We refer to this failure mode as self-deceptive debugging: the agent nominally executes a verification routine, yet substantively accepts an invalid patch due to cognitive limitations or entrapment in local optima, ultimately resulting in unsuccessful

repairs. The prevalence of self-deceptive debugging highlights a fundamental limitation of single-trajectory self-correction: verification along a single execution path is insufficient to expose latent defects. Because code correctness is inherently multi-dimensional (e.g., functional completeness, boundary robustness, and behavioral consistency), single-view verification is susceptible to confirmation bias. This observation motivates a multi-view, cross-trajectory verification paradigm, in which multiple candidate patches are generated and cross-validated to improve reliability.

3.3 Problem Formulation

We formalize the core problem as follows: given a software engineering task, design an agent architecture that (1) dynamically allocates computation based on per-step decision uncertainty—thereby avoiding redundant exploration in near-deterministic steps—and (2) leverages debugging signals from multiple trajectories to synthesize a comprehensive, multi-dimensional test suite for cross-validation, thereby reducing self-deceptive debugging and improving patch quality. These principles form the foundation of Entropy-guided Stepwise Scaling, detailed in Section 4.

4 Method

4.1 Overview

As illustrated in Figure 3, we propose EGSS, a framework that generates high-quality and robust code patches under a constrained inference budget. Specifically, in the first stage, the agent dynamically identifies key decision points during inference based on tool entropy (Dong et al., 2025) and conducts targeted exploration at these critical nodes, thereby substantially increasing the diversity and semantic coverage of candidate trajectories while preserving generation quality. In the second stage, by leveraging debugging signals embedded across multiple trajectories, the framework constructs an executable, consolidated test suite, and integrates context-aware preference selection with an ensemble voting mechanism across multiple models to systematically validate candidate patches and filter them for robustness. This two-stage design constitutes a closed-loop optimization process: the first stage provides a high-potential candidate set, whereas the second stage offers a reliable selection mechanism over multiple candidate sets.

4.2 Dynamic Stepwise Search

Motivated by the observation that uncertainty is not uniformly distributed across decision points in tool-augmented reasoning, we aim to prioritize computation at steps that are most ambiguous. Specifically, we identify critical steps via Eq. 1. When the entropy is high—indicating substantial uncertainty over the appropriate tool choice—we invoke an auxiliary LLM-based judge (Li et al., 2025b) to evaluate candidate continuations and prune low-quality branches. This adaptive mechanism allocates inference budget more effectively than repeat sampling strategies (Antoniades et al., 2024). Overall, this stepwise focusing with dynamic evaluation yields a favorable performance–cost trade-off.

In Dynamic Stepwise Search, the judge provides trajectory-level feedback that is incorporated into the search objective to select the top- K partial paths (Zhang et al., 2024a). For a partial trajectory $\tau_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1})$, we define

$$S(\tau_t) = \frac{1}{|\tau_t|^\gamma} \sum_{i=0}^{t-1} \left(\log P(a_i | s_i) + \lambda \cdot \mathbf{1}_{\{i>0\}} \cdot J(a_i | \tau_{i-1}, s_i) \right) \quad (2)$$

where s_i denotes the step- i state (encoding the available history/context) and a_i is the corresponding action. $\log P(a_i | s_i)$ is the base model log-likelihood, while $J(\cdot)$ is the judge-provided evaluation signal; $\lambda \in [0, 1]$ controls their relative contribution (we use $\lambda \in [0.6, 1.0]$ in practice). The length penalty $\text{length}(\tau_t)^\gamma$ with $\gamma \geq 0$ normalizes for trajectory length to mitigate bias toward short trajectories; $\gamma = 0$ recovers the unnormalized sum, and larger γ approximates per-step averaging. DSS ranks partial trajectories by $S(\tau_t)$ and prioritizes expansion of those actions a_i with higher estimated utility, thereby focusing computation on informative decision points and reducing wasted tokens on low-value or repetitive continuations.

To obtain a reliable and informative JudgeScore, we employ a trained LLM as a judge (Li et al., 2025b) and perform online evaluation after each expansion step. The judge input is:

$$I_{\text{Judge}} = \{\text{Task}, \tau_{\leq t}, \text{Rubric}, a_{t+1}\} \quad (2)$$

where Task denotes the problem specification, $\tau_{\leq t}$ the current partial trajectory (i.e., the action/tool sequence up to step t), and Rubric (Team et al.,

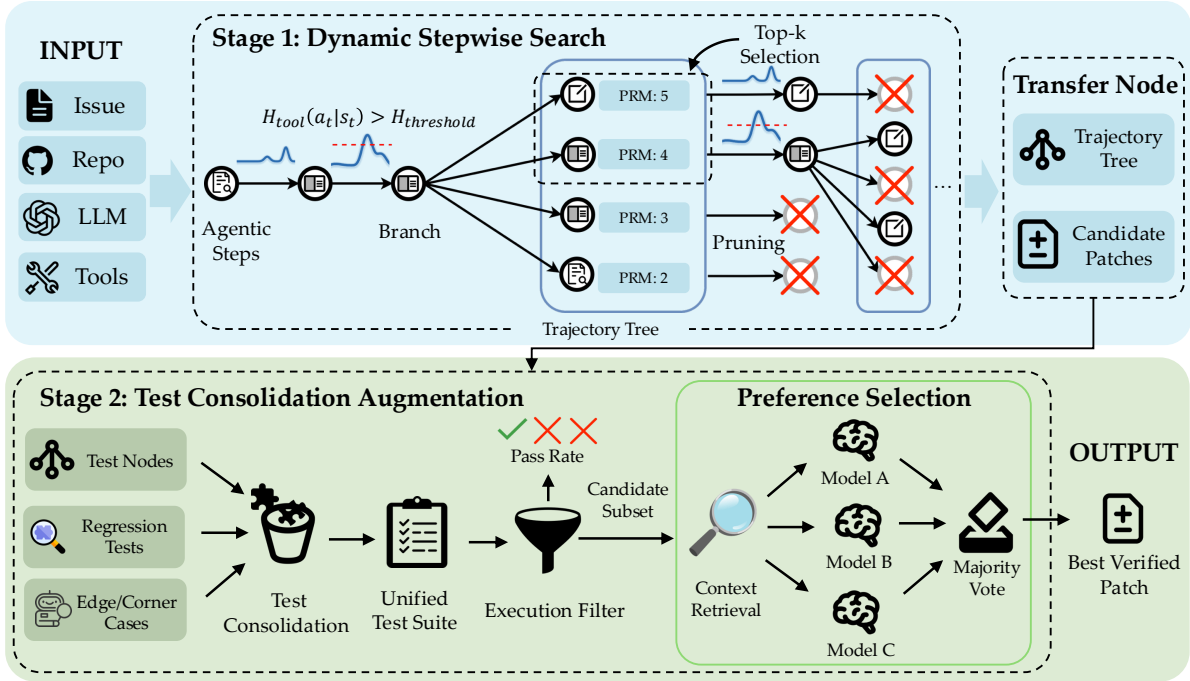


Figure 3: Overview of Entropy-guided Stepwise Scaling

2025b) the explicit evaluation criteria (see Appendix B.1 for details). The judge score at step- $t+1$ can be computed as

$$J(a_{t+1} | \tau_t, s_{t+1}) = \mathcal{J}_{LLM}(I_{Judge}) \quad (3)$$

4.3 Test Consolidation Augmentation

The phenomenon of self-deceptive debugging, wherein agents perform verification procedures yet still accept incorrect patches (Section 3.2), exposes a fundamental limitation of conventional Test-Time Scaling (TTS) pipelines: their dependence on a narrow, unidimensional correctness criterion. By formulating patch validation as an isolated terminal decision, these pipelines fail to exploit the diagnostic evidence distributed across diverse debugging trajectories. This restricted perspective not only degrades the robustness of patch selection but also neglects the collective information afforded by multi-trajectory exploration, which could otherwise reveal subtle defects that remain undetected under superficial test success.

To directly address this limitation, we propose **Test Consolidation Augmentation (TCA)**, a robust and reliable patch selection pipeline that consolidates fragmented debugging evidence into a unified and executable verification standard. Unlike LLM-as-Judge heuristics (Li et al., 2025b), which are susceptible to inconsistency and hallucination, TCA anchors final decisions in concrete

and reproducible test outcomes synthesized from the complete set of trajectories.

As outlined in Algorithm 1, TCA operates in four consecutive stages:

Debugging action extraction. Given the trajectory tree \mathcal{T} generated by DSS, we first collect all nodes that invoke debugging tools (e.g., test generation/execution). These nodes constitute \mathcal{A}_{debug} , which compactly summarizes the debug-related tool-invocation steps across trajectories.

Consolidated test suite generation. We then employ TESTCONSOLIDATIONAGENT, an agent equipped with the same MCP tools (Abdelaziz et al., 2024) used in the code generation (see Appendix B.2 for details), to synthesize an executable test suite \mathcal{S} from \mathcal{A}_{debug} and the repository \mathcal{R} . Concretely, the agent consolidates complementary testing intents encoded in \mathcal{A}_{debug} and grounds them in the repository context, yielding a unified suite that provides broader and more robust coverage than any single-trajectory test set.

Patch evaluation and filtering. Next, we apply TESTEVALUATIONAGENT, configured exclusively with bash MCP tools (see Appendix B.2 for details), to execute \mathcal{S} on each candidate patch $p \in \mathcal{P}$ and compute its pass rate. We retain patches whose pass rate exceeds a threshold τ , producing a high-confidence candidates \mathcal{P}_{valid} . This stage filters out

Algorithm 1 Test Consolidation Augmentation

Require: Trajectory tree \mathcal{T} from DSS, Candidate patches \mathcal{P} , Codebase \mathcal{R} , pass rate threshold τ
Ensure: Selected patch p^*

```
1: // Step 1: Extract debugging actions
2:  $\mathcal{A}_{\text{debug}} \leftarrow \emptyset$ 
3: for each node  $v$  in  $\mathcal{T}$  do
4:   if  $v$  invokes debugging tool then
5:      $\mathcal{A}_{\text{debug}} \leftarrow \mathcal{A}_{\text{debug}} \cup \{v\}$ 
6:   end if
7: end for
8: // Step 2: Generate consolidated test suite
9:  $\mathcal{S} \leftarrow \text{TESTCONSOLIDATIONAGENT}(\mathcal{A}_{\text{debug}}, \mathcal{R})$ 
10: // Step 3: Evaluate patches with test suite
11:  $\mathcal{P}_{\text{valid}} \leftarrow \emptyset$ 
12: for each patch  $p$  in  $\mathcal{P}$  do
13:    $r \leftarrow \text{TESTEVALUATIONAGENT}(p, \mathcal{S}, \mathcal{R})$ 
14:   if  $r.\text{passRate} \geq \tau$  then
15:      $\mathcal{P}_{\text{valid}} \leftarrow \mathcal{P}_{\text{valid}} \cup \{p\}$ 
16:   end if
17: end for
18: // Step 4: Select best patch via voting
19:  $\mathcal{V} \leftarrow \emptyset$ 
20: for each model  $m$  in base models do
21:    $p_m \leftarrow \text{PREFERENCESELECTORAGENT}_m(\mathcal{P}_{\text{valid}}, \mathcal{R})$ 
22:    $\mathcal{V} \leftarrow \mathcal{V} \cup \{p_m\}$ 
23: end for
24:  $p^* \leftarrow \text{MAJORITYVOTE}(\mathcal{V})$ 
25: return  $p^*$ 
```

patches that may appear plausible but do not withstand consolidated verification.

Preference selection via voting. If multiple candidates pass the consolidated tests, we employ multiple PREFERENCE SELECTORS to vote for a single best patch. Specifically, each selector independently selects one patch from $\mathcal{P}_{\text{valid}}$ conditioned on the repository context \mathcal{R} .

5 Experiments

5.1 Benchmark and Evaluation Metrics

We evaluate the effectiveness of Entropy-guided Stepwise Search on the SWE-Bench benchmark (Jimenez et al., 2023), adopting the **resolved percentage**—defined as the proportion of benchmark instances for which a correct patch is successfully generated—as the primary metric for quantitatively comparing EGSS with established baselines. In addition, we report the **Oracle** metric (Zhang et al., 2024b), which considers an instance resolved if at least one correct patch is identified among the N generated candidate patches. Together, these metrics offer a comprehensive evaluation of both the empirical effectiveness of EGSS and its upper-bound performance under candidate selection.

5.2 Compared TTS Method

To rigorously evaluate the effectiveness of the proposed method, we conduct a comparative study against several widely used TTS baselines, with emphasis on two critical stages: *sampling* and *patch selection*.

In the sampling stage, we investigate **repeat sampling** (Gao et al., 2025) under controlled stochasticity by adjusting generation hyperparameters (e.g., temperature), thereby encouraging diversity in the synthesized code patches and the corresponding execution traces.

For patch selection, we adopt two widely used strategies—**Deibase** (Zhang et al., 2024b) and **Augment** (Chen and Flaherty, 2025)—as baseline methods for comparison. Following their original formulations, we further refine and extend both approaches:

- **Deibase** (Zhang et al., 2024b): The original method retrieves relevant contextual information from the codebase and, together with each candidate patch and the issue description, prompts an LLM-as-judge (Li et al., 2025b) to assign a score to each candidate. We extend this design by adopting an Agent-as-Judge (Zhuge et al., 2024) paradigm, in which a judge agent autonomously evaluates each candidate patch by constructing targeted test suites and/or performing rubric-guided code reviews.
- **Augment** (Chen and Flaherty, 2025): The original method employs an LLM-as-judge paradigm (Li et al., 2025b), prompting the model to assess each candidate patch against the issue description and select the most appropriate one. We adapt this framework to our selector design by aggregating the decisions of multiple selectors via majority voting, thereby producing a single consensus patch.

5.3 Experiment Setup

In our experiments, the base agent framework integrates a suite of MCP tools (Abdelaziz et al., 2024) (see Appendix B.1.1 for agent details). For patch generation, we evaluate several large language models (LLMs): Kimi-K2-Instruct (Team et al., 2025b), GLM-4.6, and GLM-4.5-Air (Team et al., 2025a).

During the Dynamic Stepwise Search stage, all models in our experiments operate with a temperature of 1, and the stepwise rollout count is set to 4.

Method	Base Model	Resolved%	
		Swe-Bench Verified	Swe-Bench Lite
Single Shot	Kimi-K2-Instruct	63.2%	48.33%
	GLM-4.6	65.8%	51%
	GLM-4.5-Air	56.4%	N/A
	Claude-4-sonnet	68%	57%
	Claude-45-sonnet	70.4%	62%
Repeat+Dei Aug	Kimi-K2-Instruct	65.4%	N/A
	GLM-4.6	66.6%	N/A
	GLM-4.5-Air	57.2%	N/A
EGSS (Ours)	Kimi-K2-Instruct	70.6%	61%
	GLM-4.6	73.8%	64%
	GLM-4.5-Air	62.6%	N/A

Table 2: Performance differences across models under various methods, with fixed Ensemble size $K = 4$

To determine which steps qualify as “high-entropy,” we set the entropy threshold at the 75th percentile ($q_{75} \approx 1.16$) of the empirical entropy distribution observed in our dataset. This data-driven choice balances computational efficiency with decision reliability by filtering out the top 25% most ambiguous states while avoiding excessive overhead on routine tasks. A detailed statistical analysis and justification for this threshold are provided in Appendix A.1. A fine-tuned Qwen3-8B (Yang et al., 2025) model serves as the judge, providing scalar scoring signals. We further evaluate sensitivity to the ensemble size by comparing configurations with 4, 6, and 8 rollouts, which correspond to generating 4, 6, and 8 final patch candidates, respectively.

In the Test Consolidation Augmentation stage, Kimi-K2-Instruct (Team et al., 2025b) serves as the base model for both the TESTCONSOLIDATIONAGENT and the TESTEVALUATIONAGENT. Final decisions are determined via majority voting over an ensemble of expert models comprising Kimi-K2-Instruct (Team et al., 2025b), GLM-4.6 (Team et al., 2025a), and Qwen3-Coder-480B-A35B-Instruct (Yang et al., 2025).

5.4 Experimental Results

As summarized in Table 2, with an ensemble size of $K = 4$, EGSS achieves 73.8% on SWE-Bench-Verified and 64% on SWE-Bench-Lite. Across different base LLMs, EGSS consistently outperforms the compared TTS-based methods, delivering relative improvements of approximately 5%–10%.

To further elucidate the sources of these gains, we structure our ablation study into two stages corresponding to EGSS’s pipeline. The first stage

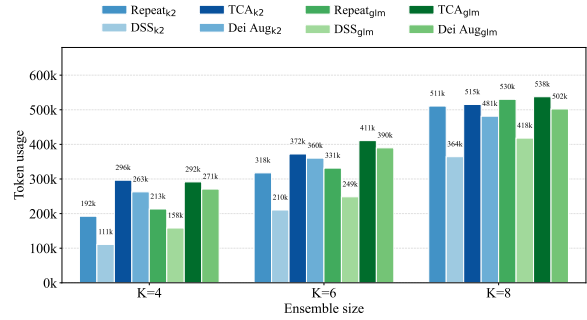


Figure 4: Average token usage per instance on the SWE-Bench benchmark, aggregated across different ensemble sizes and sampling strategies for Kimi-K2-Instruct and GLM-4.6.

investigates the sampling component, comparing Dynamic Stepwise Search (DSS) with conventional repeat sampling (Gao et al., 2025) in terms of token efficiency and oracle performance (Zhang et al., 2024b). The second stage assesses the effect of Test Consolidation Augmentation under fixed ensemble sizes across sampling strategies, examining its impact on the stability of patch generation. Detailed analyses are presented in the following sections.

5.4.1 Analysis of Token Usage

Base Model	K	Sampling Strategy	Oracle
Kimi-K2-Instruct	4	Repeat	71.8%
		DSS	73.8%
	6	Repeat	73.4%
		DSS	75.6%
	8	Repeat	74.2%
		DSS	78.8%
GLM-4.6	4	Repeat	73.4%
		DSS	77.4%
	6	Repeat	74.8%
		DSS	78.4%
	8	Repeat	76.6%
		DSS	79.6%
GLM-4.5-Air	4	Repeat	60.8%
		DSS	64.8%

Table 3: Comparison of different sampling strategies on SWE-Bench-Verified

Component-Level Overhead vs. System-Wide Efficiency. A rigorous evaluation of Test-Time Scaling (TTS) frameworks must distinguish between component-level costs and system-wide efficiency. The EGSS framework incorporates Test Consolidation Augmentation (TCA), which introduces additional computational over-

Base Model	K	Resolved %			
		Repeat + Dei Aug	Repeat + TCA	DSS + Dei Aug	DSS + TCA (EGSS)
Kimi-K2-Instruct	4	65.4%	66.6%	65.8%	68.4%
	6	66%	67.6%	68.2%	70.6%
	8	66.4%	67.6%	70.0%	72.2%
GLM-4.6	4	64.2%	66.4%	71.6%	73.8%
	6	67%	68.6%	70.2%	72%
	8	68.2%	70.4%	71.0%	74.6%

Table 4: Performance comparison (Resolved %) across different TTS methods, base models, and Ensemble size K values on SWE-Bench-Verified

head—specifically, wall-clock latency for context consolidation and test execution—compared to lightweight verification baselines. As quantified in Figure 4, when fixing the ensemble size K , EGSS exhibits a modest token overhead (ranging from 3% to 13%) relative to the baseline. This overhead is an inherent characteristic of the consolidation mechanism, which requires full context reads and comprehensive test suite executions.

However, this component-level cost is strategically offset by the substantial gains in **sample efficiency** provided by the Diversified Search Strategy (DSS). By generating higher-quality candidates, EGSS reduces the ensemble size K required to achieve state-of-the-art performance. Since verification costs scale linearly with K , reducing the required ensemble size yields disproportionate savings in total computational resources, effectively amortizing the fixed cost of TCA.

Empirical Validation of Net Savings. To quantify the holistic efficiency, we evaluate the total token consumption required to reach comparable or superior accuracy levels. The results demonstrate that EGSS achieves a superior cost-performance Pareto frontier:

- **GLM-4.6:** The baseline method requires an ensemble of $K = 8$ (approximately 502k tokens) to achieve 68.2% resolved ratio. In contrast, EGSS achieves a superior 73.8% resolved ratio with only $K = 4$ (approximately 291k tokens). This represents a **42.0%** reduction in total token usage, demonstrating that the savings from reduced ensemble size vastly outweigh the fixed overhead of TCA.
- **Kimi-K2-Instruct:** Similarly, EGSS at $K = 4$ (approximately 296k tokens, and 68.4% in resolved ratio) outperforms the baseline at $K = 8$ (approximately 481k tokens, 66.4%

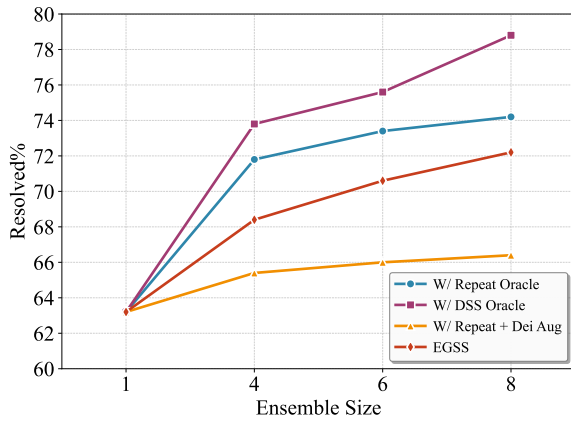
in resolved ratio), resulting in a **38.4%** net saving.

5.4.2 Analysis of Patch Selection

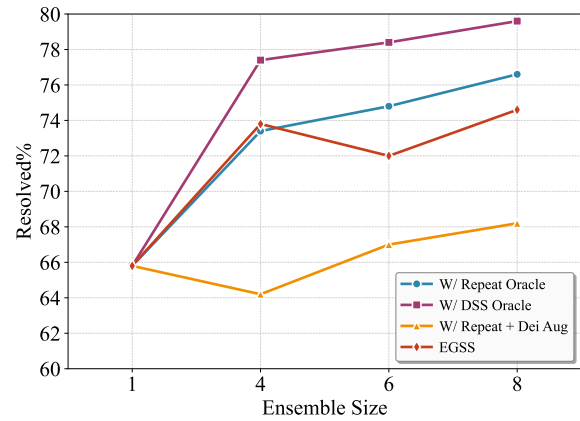
As shown in Table 4 and Figure 5b, Test Consolidation Augmentation demonstrates consistently strong performance across ensemble sizes K and sampling strategies. In particular, TCA exhibits both efficiency and robustness: using GLM-4.6 under the DSS strategy, it attains a success rate of 73.8% at $K = 4$, which already exceeds the strongest baseline (DSS+Dei Aug), and further improves to 74.6% at $K = 8$.

Unlike methods for which increasing K can introduce additional noise and thereby impair patch selection—reflected in non-monotonic performance trends—TCA effectively exploits larger ensembles. It approaches the Oracle upper bound with fewer samples while preserving a higher-quality candidate pool, thereby supporting more reliable patch selection as K increases.

To elucidate the individual contributions of TCA’s components, we performed a detailed ablation study (see Appendix A.3). As summarized in Table 7, **Test Consolidation** serves as the primary driver of performance, significantly outperforming the random selection baseline and closely approximating the Oracle upper bound through effective candidate ranking. Specifically, selecting candidates based on their pass rates over the consolidated test suite (i.e., Top-1 and Top-3) proves highly effective in ranking correct patches, closely approximating the Oracle upper bound. While augmentation provides supplementary stability, the consolidated test suite is essential for distinguishing correct solutions from plausible failures. We refer readers to the appendix for a detailed quantitative breakdown comparing these selection strategies across various model backbones.



(a) Kimi-K2-Instruct



(b) GLM-4.6

Figure 5: Comparison of ensemble strategies across different ensemble sizes

6 Conclusion

In this paper, we proposed Entropy-guided Stepwise Scaling (EGSS), a novel TTS framework designed to enhance both computational efficiency and operational reliability in autonomous software engineering. EGSS dynamically allocates computational resources to high-uncertainty steps based on tool entropy and integrates cross-trajectory test evidence to enable robust patch selection.

Our empirical results on SWE-Bench-Verified demonstrate that EGSS consistently boosts performance by 5–10% across various base models, establishing a new state-of-the-art among open-source methods. Crucially, we highlight that the architectural sophistication of EGSS—comprising specialized agents for generation, judgment, and consolidation—serves as a strategic investment rather than a mere overhead. By replacing blind, uniform scaling with targeted, entropy-guided exploration and rigorous test consolidation, EGSS reduces total token consumption by **38–42%** compared to strong baselines while achieving superior accuracy. These findings suggest that in complex domains like software engineering, structured architectural complexity is a necessary prerequisite for achieving cost-effective and reliable autonomy, offering a sustainable alternative to brute-force inference scaling.

7 Limitations

Stepwise scaling paradigm has not been tested on more open-ended software engineering tasks—such as feature implementation, API design—where action spaces are broader and success criteria less well-defined, limiting conclusions

about its generalizability.

References

- Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Sadhana Kumaravel, Matthew Stallone, Rameswar Panda, Yara Rizk, GP Shrivatsa Bhargav, Maxwell Crouse, Chulaka Gunasekara, and 1 others. 2024. Granite-function calling model: Introducing function calling abilities via multi-task learning of granular tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1131–1139.
- Antonis Antoniadis, Albert Örwall, Kexun Zhang, Yuxi Xie, Anirudh Goyal, and William Wang. 2024. Swebench: Enhancing software agents with monte carlo tree search and iterative refinement. [arXiv preprint arXiv:2410.20285](https://arxiv.org/abs/2410.20285).
- Islem Bouzenia, Premkumar Devanbu, and Michael Pradel. 2024. Repairagent: An autonomous, llm-based agent for program repair. [arXiv preprint arXiv:2403.17134](https://arxiv.org/abs/2403.17134).
- Mark Chen. 2021. Evaluating large language models trained on code. [arXiv preprint arXiv:2107.03374](https://arxiv.org/abs/2107.03374).
- Tongfei Chen and Colin Flaherty. 2025. [#1 open-source agent on SWE-Bench Verified by combining Claude 3.7 and O1](https://arxiv.org/abs/2512.30000). Augment Blog. Accessed: 2025-12-30.
- Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, and 1 others. 2025. Agentic reinforced policy optimization. [arXiv preprint arXiv:2507.19849](https://arxiv.org/abs/2507.19849).
- Shubham Gandhi, Jason Tsay, Jatin Ganhotra, Kiran Kate, and Yara Rizk. 2025. When agents go astray: Course-correcting swe agents with prms. [arXiv preprint arXiv:2509.02360](https://arxiv.org/abs/2509.02360).
- Pengfei Gao, Zhao Tian, Xiangxin Meng, Xinchun Wang, Ruida Hu, Yuanan Xiao, Yizhou Liu, Zhao

- Zhang, Junjie Chen, Cuiyun Gao, and 1 others. 2025. Trae agent: An llm-based agent for software engineering with test-time scaling. [arXiv preprint arXiv:2507.23370](#).
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? [arXiv preprint arXiv:2310.06770](#).
- Muhammad Khalifa, Rishabh Agarwal, Lajanugen Logeswaran, Jaekyeom Kim, Hao Peng, Moontae Lee, Honglak Lee, and Lu Wang. 2025. Process reward models that think. [arXiv preprint arXiv:2504.16828](#).
- Jiaolong Kong, Xiaofei Xie, Mingfei Cheng, Shangqing Liu, Xiaoning Du, and Qi Guo. 2025. Contrastrepair: Enhancing conversation-based automated program repair via contrastive test case pairs. *ACM Transactions on Software Engineering and Methodology*, 34(8):1–31.
- Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E Gonzalez, and Ion Stoica. 2025a. S*: Test time scaling for code generation. [arXiv preprint arXiv:2502.14382](#).
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025b. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.
- Qingyao Li, Xinyi Dai, Xiangyang Li, Weinan Zhang, Yasheng Wang, Ruiming Tang, and Yong Yu. 2025c. Codeprm: Execution feedback-enhanced process reward model for code generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8169–8182.
- Yingwei Ma, Yongbin Li, Yihong Dong, Xue Jiang, Rongyu Cao, Jue Chen, Fei Huang, and Binhua Li. 2025. Thinking longer, not larger: Enhancing software engineering agents via scaling test-time compute. [arXiv preprint arXiv:2503.23803](#).
- Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. 2024. Training software engineering agents and verifiers with swe-gym. [arXiv preprint arXiv:2412.21139](#).
- Lin Sun, Chuang Liu, Xiaofeng Ma, Tao Yang, Weijia Lu, and Ning Wu. 2025. Freeprm: Training process reward models without ground truth process labels. [arXiv preprint arXiv:2506.03570](#).
- 5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025a. *Glm-4.5: Agentic, reasoning, and coding (arc) foundation models*. [Preprint, arXiv:2508.06471](#).
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025b. Kimi k2: Open agentic intelligence. [arXiv preprint arXiv:2507.20534](#).
- Hanlin Wang, Jian Wang, Chak Tou Leong, and Wenjie Li. 2025. Steca: Step-level trajectory calibration for llm agent learning. [arXiv preprint arXiv:2502.14276](#).
- Taiyi Wang, Zhihao Wu, Jianheng Liu, Jianye Hao, Jun Wang, and Kun Shao. 2024a. Distrl: An asynchronous distributed reinforcement learning framework for on-device control agents. [arXiv preprint arXiv:2410.14803](#).
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*.
- Scott Wu. 2024. Introducing devin, the first ai software engineer. [Cognition. ai Blog](#).
- Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. 2024. Agentless: Demystifying llm-based software engineering agents. [arXiv preprint arXiv:2407.01489](#).
- Yuntong Xie, Juyong Jiang, Hung Nguyen, Pinjia Tan, Zehua Gao, Jiazhen He, and David Lo. 2024. AutoCodeRover: Autonomous program improvement. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. *Qwen3 technical report*. [Preprint, arXiv:2505.09388](#).
- John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. SWE-agent: Agent-computer interfaces enable automated software engineering. [arXiv preprint arXiv:2405.15793](#).
- Karina Zainullina, Alexander Golubev, Maria Trofimova, Sergei Polezhaev, Ibragim Badertdinov, Daria Litvintseva, Simon Karasik, Filipp Fisin, Sergei Skvortsov, Maksim Nekrashevich, and 1 others. 2025. Guided search strategies in non-serializable environments with applications to software engineering agents. [arXiv preprint arXiv:2505.13652](#).

Guangtao Zeng, Maohao Shen, Delin Chen, Zhenting Qi, Subhro Das, Dan Gutfreund, David Cox, Gregory Wornell, Wei Lu, Zhang-Wei Hong, and 1 others. 2025. Satori-swe: Evolutionary test-time scaling for sample-efficient software engineering. [arXiv preprint arXiv:2505.23604](#).

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Rest-mcts*: Llm self-training via process reward guided tree search. [Advances in Neural Information Processing Systems](#), 37:64735–64772.

Kexun Zhang, Weiran Yao, Zuxin Liu, Yihao Feng, Zhiwei Liu, Rithesh Murthy, Tian Lan, Lei Li, Renze Lou, Jiacheng Xu, and 1 others. 2024b. Diversity empowers intelligence: Integrating expertise of software engineering agents. [arXiv preprint arXiv:2408.07060](#).

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. 2024. [Agent-as-a-judge: Evaluate agents with agents](#). [Preprint, arXiv:2410.10934](#).

A Extended Experimental Results

A.1 Tool Use Entropy Statistics and Threshold Selection

In this section, we provide a detailed statistical analysis of the tool-use entropy distribution to justify our hyper-parameter settings. Our goal is to balance *computational efficiency* (by minimizing unnecessary judge invocations) with *decision reliability* (by flagging ambiguous states). We demonstrate that our chosen entropy threshold is grounded in a data-driven empirical analysis of agent trajectories within the software engineering domain.

A.1.1 Empirical Distribution Analysis

We conducted a large-scale statistical analysis on agentic trajectories derived from software engineering issues. As illustrated in **Figure 2** of the main paper, the entropy distribution exhibits a distinct right-skewed pattern. This skewness indicates that while most decision nodes possess low uncertainty, a long tail of high-entropy instances exists, corresponding to complex or ambiguous reasoning steps.

The following table summarizes the comprehensive descriptive statistics of this distribution. The mean entropy is 0.87, and the median is 0.80, suggesting that the central tendency of agent confidence is relatively high. However, the presence of a long tail is evident from the higher percentiles: the 75th percentile (q_{75}) is approximately 1.16, while the 95th percentile (q_{95}) reaches 1.76. Extreme values extend up to 2.92, representing rare cases of significant hallucination or decision ambiguity.

Stat.	Val.
Mean	0.87
Median	0.80
q_{75}	1.16
q_{95}	1.76
$\mu + \sigma$	1.38
Max	2.92

Table 5: Entropy distribution statistics.

A.1.2 Rationale for Threshold Selection (q_{75})

Based on the statistical properties outlined above, we selected the 75th percentile ($q_{75} \approx 1.16$) as the entropy threshold for triggering external verification (judge invocation). This choice is motivated by the following trade-offs:

Noise Filtering and Reliability. Setting the threshold at q_{75} effectively isolates the top 25% of high-entropy instances. Statistically, these instances correspond to the "long tail" of the distribution, which are prone to ambiguity or potential hallucinations. By targeting this upper quartile, our method ensures that the most uncertain decisions are subjected to additional scrutiny, thereby enhancing overall reliability.

Efficiency-Quality Trade-off. The selection of q_{75} represents an optimal balance between overhead and performance:

- **Lower Thresholds (e.g., Median):** Setting the threshold at the median (0.80) would classify nearly half of all routine, low-uncertainty tasks as requiring verification. This would introduce excessive computational overhead without proportional gains in accuracy.
- **Higher Thresholds (e.g., q_{95} or $\mu + \sigma$):** Conversely, thresholds such as q_{95} (1.76) or $\mu + \sigma$ (1.38) are too permissive. They allow a significant number of moderately uncertain decisions to pass unchecked, potentially compromising the quality of the final output.

Therefore, q_{75} serves as a "sweet spot" that captures the bulk of standard operational variance while pruning significant outliers. This empirical approach ensures that the hyper-parameter aligns with the intrinsic uncertainty distribution of the task domain, rather than relying on arbitrary grid-search values.

A.2 Computational Efficiency and Cost-Performance Analysis

While Test Consolidation Augmentation (TCA) introduces a marginal overhead in constructing and running an expanded test suite, the EGSS framework is explicitly architected to optimize the trade-off between computational efficiency and task effectiveness. In this section, we provide a holistic evaluation of the computational costs, demonstrating that EGSS achieves superior overall utility compared to existing Test-Time Scaling (TTS) methods through a better cost-performance balance.

Analysis of Verification Overhead. It is important to note that all TTS methods generally involve two stages: candidate generation and verification.

The verification cost scales linearly with the ensemble size (K) across all methods. Baselines, such as *Repeat + Dei Aug*, incur similar per-candidate verification costs to EGSS. Therefore, the key differentiator lies not in the per-sample cost, but in the ensemble size (K) required to achieve a target performance level.

Efficiency via Reduced Ensemble Size. EGSS improves candidate quality through Dynamic Self-Synthesis (DSS) and TCA, enabling the model to achieve superior performance with a significantly smaller ensemble size (K). Since the total computational cost is dominated by the linear scaling of verification with K , reducing the required K leads to substantial savings in both token usage and wall-clock time.

Table 6 presents the empirical results comparing EGSS against the baseline (*Dei + Aug*) in terms of token usage and resolved ratio. The data highlights two critical findings:

- Superior Performance with Lower Cost:** EGSS consistently outperforms the baseline at equivalent or lower computational costs. For instance, on the GLM-4.6 model, EGSS with $K = 4$ consumes **291,576 tokens** to achieve a **73.8%** resolved rate. In contrast, the baseline requires $K = 8$ (**502,301 tokens**) to reach only **68.2%**. This represents a **~42% reduction in token usage** for EGSS while delivering higher accuracy.
- Consistent Savings Across Models:** A similar trend is observed with Kimi-K2-Instruct. EGSS ($K = 4$; 296,486 tokens) saves approximately **38%** in tokens compared to the baseline ($K = 8$; 481,420 tokens), while also achieving a higher resolved rate (68.4% vs. 66.4%).

Conclusion on Efficiency-Effectiveness Balance. Although TCA introduces a fixed component-level overhead, the quality gains from DSS and TCA allow EGSS to converge with a significantly smaller ensemble. The computational savings from reducing K substantially outweigh the fixed cost of TCA. The observed **38–42% reduction** in total token usage demonstrates that EGSS is a highly cost-effective solution for test-time scaling, delivering superior performance with reduced resource consumption.

Table 6: Comparison of Computational Efficiency (Token Usage) and Effectiveness (Resolved Rate) between EGSS and the Baseline. EGSS achieves higher accuracy with significantly fewer tokens by leveraging a smaller ensemble size (K).

Model	Method	Ens. Size (K)	Token Usage	Resolved (%)	Token Savings
Kimi-K2-Instruct	Baseline (Dei+Aug)	4	262,841	65.4%	-
		6	360,242	66.0%	-
		8	481,420	66.4%	-
	EGSS (Ours)	4	296,486	68.4%	-38%*
		6	372,258	70.6%	-
		8	515,336	72.2%	-
GLM-4.6	Baseline (Dei+Aug)	4	270,879	64.2%	-
		6	389,943	67.0%	-
		8	502,301	68.2%	-
	EGSS (Ours)	4	291,576	73.8%	-42%*
		6	410,903	72.0%	-
		8	537,648	74.6%	-

* Token savings calculated by comparing EGSS ($K = 4$) against Baseline ($K = 8$) where EGSS achieves higher performance.

A.3 Ablation Study on Test Consolidation and Augmentation

To dissect the specific contributions of the components within the EGSS framework, particularly the role of **Test Consolidation** versus standard ensemble strategies, we conducted a detailed ablation study. This analysis aims to verify whether the performance gains are driven by the quality of candidate selection via consolidated tests or merely by the increased diversity from augmentation and voting.

Table 7 presents the resolved ratio across different selection strategies using Kimi-K2-Instruct and GLM-4.6 as base models. We compare the following configurations:

- Average (Random Baseline):** Represents the expected performance of randomly selecting a single candidate from the ensemble of size N .
- Test Consolidate Top1:** Selects the single best patch based *solely* on the pass rate over the consolidated test suite generated by TCA.
- Test Consolidate Top3 (Oracle@3):** Selects the top 3 patches based on the consolidated test suite pass rate and reports the oracle accuracy (i.e., success if any of the top 3 is correct). This serves as an upper-bound proxy for the selection quality of the consolidation module.
- TCA (Full EGSS):** The complete method combining Test Consolidation with Augmentation and final ensemble voting.

- **Oracle (Upper Bound):** The theoretical maximum performance if at least one correct patch exists in the entire ensemble of size N .

Table 7: Ablation study on the contribution of Test Consolidation. Results show that Test Consolidation significantly narrows the gap between random selection and the Oracle upper bound, identifying it as the primary driver of performance improvement.

Model	Ens. (N)	Oracle	Avg.	TC Top1	TC Top3	TCA (Full)
Kimi-K2-Instruct	4	73.8%	64.2%	66.2%	73.2%	68.4%
	6	75.6%	63.4%	68.8%	74.0%	70.6%
	8	78.8%	64.0%	70.4%	75.6%	72.2%
GLM-4.6	4	77.4%	65.4%	70.4%	76.6%	73.8%
	6	78.4%	64.6%	69.2%	75.2%	72.0%
	8	79.6%	65.8%	71.4%	77.6%	74.6%

Key Observations. The results in Table 7 yield two critical insights regarding the mechanism of EGSS:

1. **Test Consolidation is the Primary Performance Driver:** The *Test Consolidate Top1* strategy consistently outperforms the *Average* baseline by a significant margin (e.g., +7.2% for GLM-4.6 at $N = 8$). More importantly, the *Test Consolidate Top3* results closely approach the *Oracle* upper bound. For instance, with GLM-4.6 ($N = 8$), TC Top3 achieves 77.6%, which is nearly identical to the Oracle limit of 79.6%. This demonstrates that the consolidated test suite is highly effective at suppressing false positives and accurately ranking correct patches, independent of complex voting logic.
2. **Synergy with Augmentation:** While Test Consolidation alone (Top1) provides strong selection capability, the full TCA method further improves robustness. The gap between *TC Top1* and *TCA (Full)* indicates that while consolidation identifies high-quality candidates, the augmentation and final voting mechanism help stabilize the final output, ensuring consistent performance across different ensemble sizes. However, the fact that TC Top3 nearly matches the Oracle confirms that the **unification of test suites** is the distinct contribution enabling reliable selection, rather than the ensemble size itself.

In summary, this ablation study validates that the core innovation of EGSS lies in the *Test Consolidation* module, which effectively leverages unified

test feedback to distinguish correct solutions from plausible but incorrect ones, thereby maximizing the utility of the generated candidate pool.

B EGSS Implementation Detail

B.1 Dynamic Stepwise Search

Implementation Detail of Dynamic Stepwise Search

B.1.1 Code Generation

First we will introduce our code agent used for code generation, the system prompt used are as following, and Tools available for Agent are:

- `read_file`: Read file contents with optional line range selection
- `write_file`: Create or overwrite files
- `edit_file`: Perform edits via search-and-replace
- `grep`: Fast code search powered by ripgrep
- `glob`: File discovery using glob patterns
- `bash`: Execute shell commands with timeout control

System Prompt

```


 $\{available\_tools\}$ 


```

You are an AI coding assistant designed to help developers with their coding tasks. You have access to tools that allow you to read and write files in the workspace.

Your approach:

1. Carefully analyze the user's request
2. Use available tools to gather necessary information
3. Propose clear, well-thought-out solutions
4. Execute changes carefully and verify results

When modifying files:

- Always read files before modifying them
- Make precise, targeted changes
- Explain what you're doing and why

Be concise, accurate, and helpful.

B.1.2 Judge Model

When an Agent step is identified as a high-entropy action, we roll out multiple actions in the current step. To determine the preference order among these actions, we introduce a Judge Model. The Judge Model takes the current task, a partial trajectory of the agent, and a rubric as input. The

detailed system prompt and the rubric used in our experiments are provided below.

System Prompt

Role Please act as an impartial judge to **verify** the patch provided by assistant to the user's task below.

Instruction

- User will provide you with a response and the corresponding task, and your task is to verify its rationality with the rubric given below.
- Base your decision solely on how well the response addresses the user's question and adheres to the assistant instructions.
- You can use the available tools to assess the provided rubric
- **Important**, If there are anything wrong with the provided response, **do not** try to fix it!

Rubric Your evaluation should focus on the following criteria:

`\${rubric}`

Output format IMPORTANT Your final reply must be structured in the Json format consist of one args

- critique, required, str
- scalar, required, str, this is a fractional string where the numerator represents the score and the denominator is the total possible score from the rubric.

Here is an example of the output

```
{
  "critique": "xxx",
  "scalar": "a/b"
}
```

User Task:

`\${task}`

Agent History Trajectory

`\${trajectory}`

Current Round Response

`\${response}`

Output Now it's your turn.

A rubric is an evaluation criterion. In our experiment, we focused the rubric primarily on the reasonableness of certain processes, such as expected tool usage.

Rubric

1. Step Consistency Definition: Assesses whether the entire execution trace is logically coherent, free of contradictions, and devoid of redundancy.

Score	Criteria
0	Inconsistent
1	Basically Consistent
2	Partially Consistent
3	Fully Consistent

2. Context Awareness Definition: Evaluates whether the agent effectively utilizes and dynamically updates historical context, avoiding repetition or conflict.

Score	Criteria
0	Context Ignored
1	Partially Contextualized
2	Mostly Contextualized
3	Fully Contextualized

3. Goal Prioritization Definition: Determines whether the agent correctly identifies and prioritizes critical sub-goals, preventing resource misallocation.

Score	Criteria
0	Chaotic Prioritization
1	Unbalanced Prioritization
2	Adequate Prioritization
3	Optimal Prioritization

4. Expected Tool Use Definition: Selection, invocation, and sequencing of the *correct* tools with *appropriate* arguments and *minimal* redundancy, including proper error handling and validation. This includes ensuring format-specific features are fully implemented across all relevant code paths, and critically, handling **edge cases** and **optimization paths** in the target codebase (e.g., fast-delete optimization in Django deletion logic).

Score	Criteria
0	Misuse or Abstain
1	Partial Tool Use
2	Adequate Tool Use
3	Optimal Tool Use

5. Diagnostic Precision Definition: Evaluates how effectively the agent isolates the root cause of an issue, distinguishes between symptoms and causes, and provides targeted explanations or fixes without over-generalizing or misidentifying the problem.

Score	Criteria
0	Misdiagnosis
1	Partial Diagnosis
2	Accurate Diagnosis
3	Precise Diagnosis

A case study of the Judge Model is presented below, illustrating how preference ordering among high-entropy actions is determined. Due to space constraints, only a portion of the input is displayed here, with certain content omitted and indicated by ellipses.

Case Input: $\{\text{task}\}$

Modeling's separability_matrix does not compute separability correctly for nested Compound Models
Consider the following model: ...

```
        was inefficient and illogical
        ....",
    "scalar": "18/30"
}
```

Case Input: $\{\text{trajectory}\}$

```
<trajectory>
  <tool_call name="read_file">
    <arguments>
      {"path": "/testbed/astropy/
        modeling/tests/test_separable.
        py"}
    </arguments>
    <response>
      ...
    </response>
  </tool_call>
  <tool_call name="bash">
    <arguments>
      {"command": "cd /testbed &&
        python -m pytest astropy/
        modeling/tests/test_separable.
        py -v"}
    </arguments>
    <response>
      Command executed successfully.\n\
      nOutput:...
    </response>
  </tool_call>
  ...
</trajectory>
```

Case Input: $\{\text{response}\}$

```
<response>
  <tool_call name="edit_file">
    <arguments>
      {
        "file_path": "/testbed/astropy
          /modeling/tests/
          test_separable.py",
        "old_string": ...,
        "new_string": ...
      }
    </arguments>
  </tool_call>
</response>
```

Case Judge Model Output

```
{
  "critique": "The assistant
    correctly identifies that a
    new test case is needed to
    demonstrate the bug reported
    by the user. The
    investigative steps,
    particularly running a debug
    script to trace the matrix
    creation, were relevant.
    However, the overall process
```

B.2 Test Consolidation Augmentation

This section will reveal the implementation details of Test Consolidation Augmentation

B.2.1 Test Consolidate Agent

The Test Consolidation Agent is designed to consolidate debugging signals from diverse agent trajectories. It shares the same MCP tools as the Code Generation agent (introduced in Section B.1.1) and possesses equivalent privileges, granting it full access to any file within the repository.

Tools available for Test Consolidate Agent are:

- read_file: Read file contents with optional line range selection
- write_file: Create or overwrite files
- edit_file: Perform edits via search-and-replace
- grep: Fast code search powered by ripgrep
- glob: File discovery using glob patterns
- bash: Execute shell commands with timeout control

System Prompt

$\{\text{available_tools}\}$

ROLE: Act as an expert of programming tester. Given a codebase, an issue, and trajectories generated by different code agents, you need to generate a **single test file** that can be used to validate the patch generated.

WORK PROCESS:

1. **Contextual Analysis:** Comprehend the issue and relevant codebase components (referenced code, patched code, related files, existing regression tests).
2. **Trajectory Analysis:** Evaluate test cases and debug code from each trajectory for sufficiency.
3. **Coverage Assessment:** Identify and add missing test scenarios (edge cases, regression) to ensure comprehensive coverage.
4. **Test Consolidation:** Integrate all test cases (from trajectories, new designs, existing regressions) into 'test_current_issue.py' at the root.
5. **Compilation Assurance:** Ensure 'test_current_issue.py' compiles without errors, even if tests fail initially.

Requirements for TEST FILE:

- **MUST** be 'test_current_issue.py' in the root directory.
- Provide **comprehensive** validation for the issue.

- **IMPORTANT:** Executable with 'pytest', reporting total test cases.

Task

`\${task}`

Trajectory Trajectories sampled from different code agents with only debugging tools:

`\${trajectories}`

A case study of the Test Consolidation Agent is presented below, illustrating the structure of execution trajectories and providing a concrete example. Due to space constraints, only a portion of the input is displayed here, with certain content omitted and indicated by ellipses.

Case Input: `\${task}`

Modeling's separability_matrix does not compute separability correctly for nested CompoundModels. Consider the following model: ...

Case Input: `\${trajectories}`

```
<trajectory idx='1'>
<tool_call name="write_file">
  <arguments>
    {
      "path": "/testbed/test_issue.py",
      "content": "..."
    }
  </arguments>
</tool_call>
<tool_call name="write_file">
  <arguments>
    {
      "path": "/testbed/debug_issue.py",
      "content": "..."
    }
  </arguments>
</tool_call>
<tool_call name="bash">
  <arguments>
    ...
  </arguments>
</tool_call>
...
</trajectory>
<trajectory idx='2'>
  ....
</trajectory>
```

Case Agent Output

Perfect! Now let me create a final summary of what was done:

Summary

I have successfully created a comprehensive test file that validates the patch for the separability matrix bug in nested CompoundModels. Here's what was accomplished:

Test Results

- **12 comprehensive test cases** - all passed

- **12 existing tests** - all still pass (no regressions)
- **Edge cases** - all validated
- **Original issue examples** - all now work correctly

Key Test Cases Validated

...

B.2.2 Test Evaluation Agent

The Test Evaluation Agent is designed to execute a specified test file within a strictly controlled environment. It is restricted exclusively to Bash-related Model Context Protocol (MCP) tools and operates solely on the designated test file. To ensure integrity and security, strict filesystem isolation is enforced, preventing access to any files outside the test directory. Upon invocation, the agent automatically executes the test file and autonomously resolves any environmental dependencies or configuration issues that might impede successful execution.

Tools available for Test Evaluation Agent are:

- **bash:** Execute shell commands with timeout control

System Prompt

`\${available_tools}`

ROLE: Act as an expert programming tester. Validate code with test_current_issue.py based on codebase, issue, and patch.

WORK PROCESS:

1. **Execute Test:** Run python test_current_issue.py. Output number of solved test cases.
2. **Fix Environment Issue (if needed):** Correct compilation errors in test_current_issue.py if not due to patch. **NOTICE:** you can't modify any codes!

OUTPUT FORMAT: IMPORTANT: Reply in JSON.

- passed: int, passed cases.
- failed: int, failed cases.
- error: int, error cases.
- total: int, total test cases.

```
{
  "passed": x,
  "failed": x,
  "error": x,
  "total": x
}
```

Output Now it's your turn.

B.2.3 Preference Selector

The Preference Selector is primarily employed to filter and select a single, final patch from a given set of multiple candidates. It shares the same MCP tools as the Code Generation agent (introduced in Section B.1.1) and possesses equivalent privileges,

granting it full access to any file within the repository. Tools available for Preference Selector are:

- `read_file`: Read file contents with optional line range selection
- `write_file`: Create or overwrite files
- `edit_file`: Perform edits via search-and-replace
- `grep`: Fast code search powered by ripgrep
- `glob`: File discovery using glob patterns
- `bash`: Execute shell commands with timeout control

System Prompt

`#{available_tools}`

ROLE: Act as an expert code selector. Given a codebase, an github issue and N candidate patches proposed by your colleagues, your responsibility is to select the correct one to solve the issue.

WORK PROCESS:

1. **Understand Issue & Codebase:** Comprehend the problem from issue description. Examine codebase for context:
 - (1) Code/patches referenced in issue.
 - (2) Unchanged/related parts of affected files.
2. Analyze the Candidate Patches: For each patch, analyze its logic and intended fix. Consider whether the changes align with the issue description and coding conventions.
3. verify its rationality with the rubric given below.
4. The candidate patches have not yet applied to the repository, apply first before validate the patch

Rubric Your evaluation should focus on the following criteria:

`#{rubric}`

Output Format: Reply in JSON:

```
{"result": "x" // id of the patch}
```

Analysis: [Explain why Patch-x is correct.]

Tasks:

`#{task}`

Candidate Patches

`#{patches}`

Output Now it's your turn.

Rubric used during preference selection is as following

Rubric

1. **Requirement Relevance Definition:** How completely and precisely the patch satisfies ****all**** functional and non-functional requirements expressed or implied in the user's task.

Score	Criteria
0	Severely Off-Topic
1	Partial Coverage
2	Highly Relevant
3	Perfect Alignment

2. **Code Accuracy Definition:** Apply available tools to run the code and check for any compilation errors.

Score	Criteria
0	Broken
1	Flawed
2	Correct
3	Robust & Accurate

3. **Change Precision Definition:** How accurately the patch targets ****only**** the code that must change, avoiding extraneous edits.

Score	Criteria
0	Mis-targeted
1	Imprecise
2	Accurate
3	Minimal & Exact

4. **Dependency & Context Awareness Definition:** Awareness of upstream/downstream dependencies and the completeness of associated updates (imports, call sites, configs, external contracts, backward compatibility).

Score	Criteria
0	Breaking Change
1	Partial Awareness
2	Internally Consistent
3	System-Wide Vision

5. **Code Quality Definition:** Adherence to project style guides, language idioms, readability, and maintainability.

Score	Criteria
0	Poor
1	Inconsistent Style
2	Clean & Comfortable
3	Exemplary

6. **Functionality Validation (Gating Criterion) Definition:** Adherence to project style guides, language idioms, readability, and maintainability.

Score	Criteria
0	Any Failure
3	Comprehensive & Robust

A case study on the Preference Selector is as follows.

Case Input: `#{task}`

Modeling's `separability_matrix` does not compute separability correctly for nested CompoundModels
Consider the following model: ...

Case Input: \${patches}

```
<patch id='A'>
diff --git a/astropy/modeling/separable.py b
    /astropy/modeling/separable.py
index a308e27297..45bea36085 100644
--- a/astropy/modeling/separable.py
+++ b/astropy/modeling/separable.py
@@ -242,7 +242,7 @@ def _cstack(left, right):

     cright = _coord_matrix(right, 'right
         ', noutp)
     else:
         cright = np.zeros((noutp, right.
             shape[1]))
-     cright[-right.shape[0]:, -right.
         shape[1]:] = 1
+     cright[-right.shape[0]:, -right.
         shape[1]:] = right

     return np.hstack([cleft, cright])
...
</patch>
<patch id='B'>
...
</patch>
<patch id='C'>
...
</patch>
...
```

Case Agent Output

Perfect! All tests pass, including the new test that specifically covers the nested compound model issue.

Summary

I have successfully analyzed the issue and implemented the correct fix. Here's my analysis:

Status: succeed

Result:

```
{"result": "A"}
```

Analysis:

Issue Understanding:

The issue was in the `_cstack` function in ...

Root Cause: ...

Why Patch A is correct:

1. **Core Fix:** All patches (A, B, C, D) contain the identical core fix to line 245
2. **Test Coverage:** Patch A provides the most comprehensive test coverage, including:
 - The exact issue scenario from the GitHub issue
 - Edge cases with different nesting patterns
 - Both left and right nested compound models
3. **Documentation:** Patch A includes clear docstrings explaining the test purpose

...