# Novelty Assessment Report

**Paper**: ROC-n-reroll: How verifier imperfection affects test-time scaling

**PDF URL**: https://openreview.net/pdf?id=3Gy5mmyuxn

**Venue**: ICLR 2026 Conference Submission

**Year**: 2026

**Report Generated**: 2025-12-29

## Abstract

Test-time scaling aims to improve language model performance by leveraging additional compute during inference. Many works have empirically studied techniques such as Best-of-N (BoN) and Rejection Sampling (RS) that make use of a verifier to enable test-time scaling. However, to date there is little theoretical understanding of how verifier imperfection affects performance — a gap we address in this work. Specifically, we prove that the instance-level accuracy of these methods is precisely characterized by the geometry of the verifier's ROC curve. Our theory has two important takeaways, confirmed by experiments with Qwen and LLama models on GSM8K and MATH500. First, RS outperforms BoN for fixed compute, while both methods converge to the same accuracy in the infinite-compute limit. Second, it is generally impossible to predict the high-compute performance of either method based on observations in the low-compute regime.

**Disclaimer**

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

## Core Task Landscape

This paper addresses: **Test-Time Scaling with Imperfect Verifiers**

A total of **43 papers** were analyzed and organized into a taxonomy with **21 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Foundations and Scaling Laws**
- **Verifier Design and Training**
- **Sampling and Search Strategies**
- **Sequential Refinement and Iterative Improvement**
- **Latency-Aware and Efficiency Optimization**
- **Domain-Specific Applications and Extensions**

### Complete Taxonomy Tree

- Test-Time Scaling with Imperfect Verifiers Survey Taxonomy
- Theoretical Foundations and Scaling Laws
  - Verifier Imperfection and Scaling Limits ★ (3 papers)
  - [0] ROC-n-reroll: How verifier imperfection affects test-time scaling (Anon et al., 2026) View paper
  - [19] Inference Scaling fLaws: The Limits of LLM Resampling with Imperfect Verifiers (Kapoor, 2024) View paper
  - [30] Test-time Verification via Optimal Transport: Coverage, ROC, & Sub-optimality (Mukherjee, 2025) View paper
  - Optimality and Comparative Analysis of Scaling Strategies (2 papers)
  - [1] Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters (Snell, 2024) View paper
  - [4] Scaling Test-Time Compute Without Verification or RL is Suboptimal (Setlur, 2025) View paper
  - Probabilistic and Statistical Frameworks (3 papers)
  - [25] Probabilistic Inference for Inference Time Scaling of Language Models (Puri, 2025) View paper
  - [33] Differential Privacy Meets Test-Time Scaling: Theoretical Guarantees under Noisy Inference (Jingwen Xu, 2025) View paper
  - [42] Adaptive Inference Scaling via Monte Carlo Sampling (J Boen, n.d.) View paper
- Verifier Design and Training
  - Process Reward Models and Step-Level Verification (3 papers)
  - [8] Heimdall: test-time scaling on the generative verification (Shi, 2025) View paper
  - [12] Accelerating LLM Reasoning via Early Rejection with Partial Reward Modeling (Khan, 2025) View paper
  - [20] Instance-Adaptive Inference-Time Scaling with Calibrated Process Reward Models (YJ Park, 2025) View paper
  - Verifier Ensemble and Combination Methods (2 papers)
  - [10] Shrinking the Generation-Verification Gap with Weak Verifiers (Saad-Falcon, 2025) View paper
  - [40] Weaver: Shrinking the Generation-Verification Gap by Scaling Compute for Verification (J Saad-Falcon, n.d.) View paper
  - Reinforcement Learning Integration for Verification (2 papers)
  - [18] Critique to Verify: Accurate and Honest Test-Time Scaling with RL-Trained Verifiers (Yang Zhi-cheng, 2025) View paper
  - [28] VerifierQ: Enhancing LLM Test Time Compute with Q-Learning-based Verifiers (Tang, 2024) View paper
  - Multi-Domain and Cross-Domain Verifier Evaluation (2 papers)
  - [32] When Does Verification Pay Off? A Closer Look at LLMs as Solution Verifiers (Jack Lu, 2025) View paper
  - [39] Rethinking Reward Models for Multi-Domain Test-Time Scaling (Lee, 2025) View paper
- Sampling and Search Strategies
  - Parallel Sampling and Best-of-N Selection (3 papers)
  - [6] Sample, scrutinize and scale: Effective inference-time search by scaling verification (Zhao, 2025) View paper

  - [13] Inference-aware fine-tuning for best-of-n sampling in large language models (Chow, 2024) View paper
  - [23] Putting the Value Back in RL: Better Test-Time Scaling by Unifying LLM Reasoners With Verifiers (Kusha Sareen, 2025) View paper
  - Tree Search and Backtracking Methods (3 papers)
  - [2] Pushing Test-Time Scaling Limits of Deep Search with Asymmetric Verification (Zeng, 2025) View paper
  - [36] Taming Imperfect Process Verifiers: A Sampling Perspective on Backtracking (Rohatgi, 2025) View paper
  - [41] Test-Time Scaling for Multistep Reasoning in Small Language Models via A* Search (A Braverman, n.d.) View paper
  - Adaptive and Budget-Aware Sampling (2 papers)
  - [11] Efficient Test-Time Scaling via Self-Calibration (Huang, 2025) View paper
  - [17] Budget-aware Test-time Scaling via Discriminative Verification (Tan, 2025) View paper
- Sequential Refinement and Iterative Improvement
  - Self-Verification and Self-Correction (2 papers)
  - [14] SETS: Leveraging Self-Verification and Self-Correction for Improved Test-Time Scaling (Chen Jie-feng, 2025) View paper
  - [15] ReVeal: Self-Evolving Code Agents via Iterative Generation-Verification (Y Jin, 2025) View paper
  - Feedback-Driven Refinement and Critique (2 papers)
  - [16] On the Role of Feedback in Test-Time Scaling of Agentic AI Workflows (Chakraborty, 2025) View paper
  - [26] Step-level Verifier-guided Hybrid Test-Time Scaling for Large Language Models (Chang, 2025) View paper
  - Tool Integration for Verification Enhancement (1 papers)
  - [7] T1: Tool-integrated self-verification for test-time compute scaling in small language models (Kang, 2025) View paper
- Latency-Aware and Efficiency Optimization
  - Speculative and Early Rejection Methods (1 papers)
  - [3] SPECS: Faster Test-Time Scaling through Speculative Drafts (Cemri, 2025) View paper
  - Verification Granularity Optimization (1 papers)
  - [35] Rethinking Optimal Verification Granularity for Compute-Efficient Test-Time Scaling (H. Chen, 2025) View paper
- Domain-Specific Applications and Extensions
  - Agentic Workflows and Multi-Step Planning (2 papers)
  - [21] Scaling Test-time Compute for LLM Agents (Li Hanhao, 2025) View paper
  - [38] ARCANE: A Multi-Agent Framework for Interpretable and Configurable Alignment (Charlie Masters, 2025) View paper
  - Generative Models for Non-Text Modalities (3 papers)
  - [5] Scalingnoise: Scaling inference-time search for generating infinite videos (Yang HaoLin, 2025) View paper
  - [24] Video-T1: Test-Time Scaling for Video Generation (Liu, 2025) View paper
  - [27] CARINOX: Inference-time Scaling with Category-Aware Reward-based Initial Noise Optimization and Exploration (Aghayari Ali, 2025) View paper
  - Specialized Reasoning Domains (4 papers)
  - [29] Test-time Scaling Techniques in Theoretical Physics -- A Comparison of Methods on the TPBench Dataset (Gao Zhi-qi, 2025) View paper
  - [31] Evaluating the Role of Verifiers in Test-Time Scaling for Legal Reasoning Tasks (Davide Romano, 2025) View paper
  - [37] Probing the effectiveness of World Models for Spatial Reasoning through Test-time Scaling (Saurav Jha, 2025) View paper
  - [43] Scaling Natural-Language Graph-Based Test Time Compute for Automated Theorem Proving (T Knappe, n.d.) View paper
  - Multilingual and Cross-Lingual Generation (1 papers)
  - [22] Test-Time Scaling with Repeated Sampling Improves Multilingual Text Generation (Gupta, 2025) View paper
  - Latent Reasoning and Continuous Representations (1 papers)
  - [9] Parallel Test-Time Scaling for Latent Reasoning Models (Li Yongqi, 2025) View paper
  - Auxiliary Prediction and Interpretability (1 papers)
  - [34] ZIP-RC: Zero-overhead Inference-time Prediction of Reward and Cost for Adaptive and Interpretable Generation (R Manvi, 2025) View paper

### Narrative

Core task: test-time scaling with imperfect verifiers. The field explores how to allocate additional computation at inference time when verifiers—models that score or rank candidate solutions—are themselves noisy or unreliable. The taxonomy organizes research into several main branches: Theoretical Foundations and Scaling Laws examine the mathematical principles governing when and how test-time compute yields gains, including analyses of verifier imperfection and fundamental limits (Inference Scaling fLaws[19], Optimal Transport Verification[30]). Verifier Design and Training focuses on building better scoring models through techniques like step-level supervision (Step-level Verifier[26]) or self-calibration (Self-Calibration[11]). Sampling and Search Strategies investigate how to generate and explore candidate solutions efficiently, from simple repeated sampling to structured search methods. Sequential Refinement and Iterative Improvement studies multi-round approaches where models revise outputs based on feedback (ReVeal[15], Critique to Verify[18]). Latency-Aware and Efficiency Optimization addresses practical deployment constraints, balancing accuracy against wall-clock time (Early Rejection[12], Budget-aware Scaling[17]). Finally, Domain-Specific Applications and Extensions adapt these ideas to specialized settings like legal reasoning (Legal Reasoning Verifiers[31]) or video understanding (Video-T1[24]).

A particularly active line of work examines the theoretical limits of scaling under verifier noise, asking when additional samples or compute cease to help (Scalingnoise[5], When Verification Pays[32]). Another contrasting direction emphasizes practical strategies for managing imperfect verifiers in real systems, such as instance-adaptive allocation (Instance-Adaptive Scaling[20]) or early stopping heuristics. ROC-n-reroll[0] sits squarely within the Theoretical Foundations branch, specifically addressing Verifier Imperfection and Scaling Limits. It shares conceptual ground with Inference Scaling fLaws[19], which also probes fundamental constraints, and with Optimal Transport Verification[30], which offers a complementary mathematical lens on verifier quality. Compared to more application-focused neighbors, ROC-n-reroll[0] emphasizes rigorous characterization of how verifier error propagates as compute scales, providing formal insights into the diminishing returns observed in practice.

## Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. Inference Scaling fLaws: The Limits of LLM Resampling with Imperfect Verifiers

**Authors**: Kapoor, Sayash, Benedikt Stroebl, Narayanan, Arvind, et al. (7 authors total) | **Year/Venue**: 2024 | **URL**: View paper

### Abstract

Recent research has generated hope that inference scaling could allow weaker language models to match or exceed the accuracy of stronger models, such as by repeatedly sampling solutions to a coding problem until it passes unit tests. The central thesis of this paper is that there is no free lunch for inference scaling: indefinite accuracy improvement through resampling can only be realized if the"verifier"(in this case, a set of unit tests) is perfect. When the verifier is imperfect, as it almos...

### Relationship Analysis

Both papers belong to the 'Verifier Imperfection and Scaling Limits' category, analyzing how imperfect verifiers constrain test-time scaling performance. They share overlapping focus on resampling methods (Best-of-N and Rejection Sampling) with imperfect verifiers and both examine how verifier quality affects scaling behavior. The key difference is that the original paper provides a theoretical characterization using ROC curves to predict performance across compute regimes, while the candidate paper emphasizes empirical evidence of generalization gaps in coding benchmarks, showing that weaker models produce more false positives and demonstrating practical limits when false positives incur costs.

## 2. Test-time Verification via Optimal Transport: Coverage, ROC, & Sub-optimality

**Authors**: Mukherjee, Arpan, Bullo, Marcello, Arpan Mukherjee, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

### Abstract

While test-time scaling with verification has shown promise in improving the performance of large language models (LLMs), the role of the verifier and its imperfections remain underexplored. The effect of verification manifests through interactions of three quantities: (i) the generator's coverage, (ii) the verifier's region of convergence (ROC), and (iii) the sampling algorithm's sub-optimality. Though recent studies capture subsets of these factors, a unified framework quantifying the geometry...

### Relationship Analysis

Both papers belong to the same taxonomy category analyzing how verifier imperfection fundamentally constrains test-time scaling performance. They share overlapping focus on characterizing the relationship between verifier quality (via ROC curves) and test-time scaling methods like Best-of-N and Rejection Sampling, with both providing theoretical frameworks for understanding performance limits. The key difference is that the original paper (ROC-n-reroll) focuses on empirically validating ROC-based performance predictions and proving impossibility of extrapolation from low-compute regimes, while the candidate paper frames the problem through optimal transport theory, introducing coverage constraints and analyzing sub-optimality across three distinct regimes (transport, policy improvement, saturation).

# Contributions Analysis

**Overall novelty summary.** The paper provides a theoretical characterization of test-time scaling methods (Best-of-N and Rejection Sampling) through the geometry of verifier ROC curves, proving that RS outperforms BoN at fixed compute and establishing an impossibility result for extrapolating high-compute performance. Within the taxonomy, it resides in the 'Verifier Imperfection and Scaling Limits' leaf under 'Theoretical Foundations and Scaling Laws', alongside only two sibling papers. This leaf represents a relatively sparse but foundational research direction, focusing specifically on how verifier noise constrains scaling rather than on method development or empirical optimization.

The taxonomy reveals that most neighboring work falls into adjacent branches: 'Optimality and Comparative Analysis of Scaling Strategies' examines resource-constrained comparisons without the verifier-imperfection focus, while 'Probabilistic and Statistical Frameworks' formalizes scaling as inference problems. The broader 'Verifier Design and Training' branch (containing process reward models and ensemble methods) addresses improving verifier quality rather than analyzing fundamental limits given imperfection. The paper's theoretical lens on verifier error propagation distinguishes it from the empirical, method-driven work dominating sibling branches like 'Sampling and Search Strategies' or 'Sequential Refinement'.

Among twenty-two candidates examined through semantic search, none clearly refute the three core contributions. The ROC-curve characterization examined two candidates with no overlaps; the RS-versus-BoN optimality claim examined ten candidates with no refutations; the extrapolation impossibility result also examined ten candidates without finding prior work establishing this negative result. This limited search scope suggests the theoretical framing via ROC geometry and the specific impossibility claim may be novel within the examined literature, though the small candidate pool (twenty-two papers) means potentially relevant work outside top semantic matches could exist.

The analysis indicates the paper occupies a theoretically oriented niche within a field otherwise dominated by algorithmic development and empirical evaluation. The sparse population of its taxonomy leaf and absence of refutations among examined candidates suggest substantive novelty in its formal approach, though the limited search scale (twenty-two candidates from semantic retrieval) leaves open the possibility of overlooked prior work in adjacent mathematical or theoretical communities not captured by the search strategy.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

## Contribution 1: Theoretical characterization of test-time scaling via ROC curves

**Description**: The authors establish that for both Rejection Sampling and Best-of-N methods, the accuracy at a given query depends only on the generator's initial accuracy and the verifier's ROC curve. This provides a complete theoretical framework connecting verifier imperfection to test-time scaling performance.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. Enhancing Fine-Tuning-Free Clinical Reasoning via Test-Time Scaling

**URL**: View paper

**Brief Assessment**

Clinical Reasoning Scaling[63] focuses on self-consistency decoding for medical diagnosis tasks without using ROC curves or verifier characterization. The candidate's theoretical framework analyzes majority voting in classification settings, which differs fundamentally from the original paper's ROC-based analysis of verifier imperfection in rejection sampling and best-of-n methods.

### 2. Test-time Verification via Optimal Transport: Coverage, ROC, & Sub-optimality

**URL**: View paper

**Brief Assessment**

Optimal Transport Verification[30] focuses on optimal transport formulations and Youden's index for verifier characterization, not ROC curve geometry as the primary theoretical framework for test-time scaling accuracy.

## Contribution 2: Proof that RS outperforms BoN at fixed compute with concave ROC curves

**Description**: The authors prove that Rejection Sampling achieves higher accuracy than Best-of-N when controlling for compute budget, provided the verifier has a concave ROC curve. However, both methods reach identical performance as compute approaches infinity.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. Detecting and preventing hallucinations in large vision language models

**URL**: View paper

**Brief Assessment**

Hallucination Detection[54] focuses on detecting and preventing hallucinations in vision-language models using rejection sampling for quality filtering, not on theoretical comparisons of RS versus BoN under compute constraints or ROC curve analysis.

### 2. Optimal Stopping vs Best-of- for Inference Time Optimization

**URL**: View paper

**Brief Assessment**

Optimal Stopping[58] addresses a different problem: adaptive stopping for LLM generation using optimal stopping theory (Pandora's box), not the theoretical comparison of RS vs BoN under fixed compute budgets with verifier ROC curves.

### 3. STARS: Segment-level Token Alignment with Rejection Sampling in Large Language Models

**URL**: View paper

**Brief Assessment**

STARS[61] focuses on segment-level token alignment for LLM generation using rejection sampling at the token-block level, not on theoretical analysis of RS versus BoN performance under fixed compute budgets with ROC curve characterizations.

### 4. On the Query Complexity of Verifier-Assisted Language Generation

**URL**: View paper

**Brief Assessment**

Query Complexity[60] focuses on constrained generation tasks using process verifiers, not on comparing rejection sampling versus best-of-N methods under fixed compute budgets for verifier-assisted language generation in the context analyzed by the original paper.

### 5. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving

**URL**: View paper

**Brief Assessment**

Inference Scaling Laws[47] focuses on comparing different inference strategies (greedy search, majority voting, best-of-n, tree search) across model sizes, not on proving theoretical relationships between RS and BoN under fixed compute budgets with ROC curve analysis.

### 6. Best-of-Majority: Minimax-Optimal Strategy for Pass@ Inference Scaling

**URL**: View paper

**Brief Assessment**

Best-of-Majority[62] focuses on pass@k inference settings where multiple responses can be submitted, not on comparing RS versus BoN under fixed compute budgets with ROC curve analysis. The candidate addresses a different problem formulation.

### 7. Accelerating best-of-n via speculative rejection

**URL**: View paper

**Brief Assessment**

Accelerating Best-of-n[56] focuses on accelerating best-of-n through early stopping of unpromising utterances during generation, not on comparing rejection sampling versus best-of-n under fixed compute budgets or analyzing ROC curve properties.

### 8. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment

**URL**: View paper

**Brief Assessment**

Best-of-n Coverage[57] focuses on inference-time alignment with $\chi^2$-regularization and rejection sampling for optimal regret bounds, not specifically on comparing RS versus BoN under fixed compute budgets with concave ROC curves as the primary contribution.

### 9. Fast best-of-n decoding via speculative rejection

**URL**: View paper

**Brief Assessment**

Speculative Rejection[55] focuses on computational efficiency of best-of-n decoding through early stopping of low-quality generations, not on theoretical comparisons between rejection sampling and best-of-n under fixed compute budgets with ROC curve analysis.

### 10. Diverse Inference and Verification for Advanced Reasoning

**URL**: View paper

**Brief Assessment**

Diverse Inference Verification[59] focuses on aggregating diverse models and methods for mathematical olympiad problems, ARC puzzles, and HLE questions. While it uses best-of-n sampling as one technique, it does not provide theoretical analysis comparing rejection sampling versus best-of-n under fixed compute budgets or analyze ROC curve properties. The paper's contribution is in practical diverse inference aggregation, not theoretical compute-performance tradeoffs between sampling methods.

## Contribution 3: Impossibility result for extrapolating high-compute performance

**Description**: The authors prove that observing test-time scaling behavior at low compute levels does not allow reliable prediction of performance at high compute levels. This holds for both RS and BoN, even when assuming concave ROC curves.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

### 1. Probabilistic Optimality for Inference-time Scaling

**URL**: View paper

**Brief Assessment**

Probabilistic Optimality[46] focuses on deriving theoretical lower bounds for sample requirements and developing a practical algorithm (OptScale) for determining optimal sample sizes. It does not address the impossibility of extrapolating high-compute performance from low-compute observations, which is the core claim of the original contribution.

### 2. Inference-time scaling for complex tasks: Where we stand and what lies ahead

**URL**: View paper

**Brief Assessment**

Complex Tasks Survey[49] focuses on empirical evaluation of inference-time scaling across diverse tasks and models, not on theoretical impossibility results for performance extrapolation from low to high compute regimes.

### 3. A Theory of Inference Compute Scaling: Reasoning through Directed Stochastic Skill Search

**URL**: View paper

**Brief Assessment**

Directed Skill Search[51] focuses on inference compute scaling through skill graph traversal and does not address the specific problem of predicting high-compute test-time scaling performance from low-compute observations for rejection sampling or best-of-n methods.

### 4. Amix-1: A pathway to test-time scalable protein foundation model

**URL**: View paper

**Brief Assessment**

Amix-1[53] focuses on protein foundation models and test-time scaling for protein design tasks, not on theoretical impossibility results for predicting high-compute performance from low-compute observations in language model test-time scaling.

### 5. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving

**URL**: View paper

**Brief Assessment**

Inference Scaling Laws[47] studies empirical cost-performance trade-offs for various inference strategies but does not address the theoretical impossibility of predicting high-compute performance from low-compute observations.

### 6. Inference-Time Scaling for Generalist Reward Modeling

**URL**: View paper

**Brief Assessment**

Generalist Reward Modeling[45] focuses on inference-time scaling for reward models in RL post-training, not on the theoretical impossibility of extrapolating test-time scaling performance from low to high compute regimes for resampling methods like RS and BoN.

### 7. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling

**URL**: View paper

**Brief Assessment**

Large Language Monkeys[44] focuses on empirical scaling laws for coverage across multiple samples, not on the theoretical impossibility of extrapolating high-compute performance from low-compute observations for test-time scaling methods like RS and BoN.

### 8. The art of scaling reinforcement learning compute for llms

**URL**: View paper

**Brief Assessment**

Reinforcement Learning Compute[48] focuses on predictive scaling methodologies for RL training of LLMs using sigmoidal compute-performance curves, not on test-time scaling methods like rejection sampling or best-of-n that the original paper analyzes.

### 9. Inference Scaling for Long-Context Retrieval Augmented Generation

**URL**: View paper

**Brief Assessment**

Long-Context RAG[52] focuses on optimal allocation of inference compute in retrieval-augmented generation settings, not on the impossibility of predicting high-compute performance from low-compute observations in test-time scaling methods like RS and BoN.

### 10. Investigating test-time scaling with reranking for machine translation

**URL**: View paper

**Brief Assessment**

Reranking Machine Translation[50] focuses on test-time scaling for machine translation using best-of-n candidate selection, not on theoretical impossibility results for predicting high-compute performance from low-compute observations in general test-time scaling methods.

## Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

## References

- [0] ROC-n-reroll: How verifier imperfection affects test-time scaling View paper
- [1] Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters View paper
- [2] Pushing Test-Time Scaling Limits of Deep Search with Asymmetric Verification View paper
- [3] SPECS: Faster Test-Time Scaling through Speculative Drafts View paper
- [4] Scaling Test-Time Compute Without Verification or RL is Suboptimal View paper
- [5] Scalingnoise: Scaling inference-time search for generating infinite videos View paper
- [6] Sample, scrutinize and scale: Effective inference-time search by scaling verification View paper
- [7] T1: Tool-integrated self-verification for test-time compute scaling in small language models View paper
- [8] Heimdall: test-time scaling on the generative verification View paper
- [9] Parallel Test-Time Scaling for Latent Reasoning Models View paper
- [10] Shrinking the Generation-Verification Gap with Weak Verifiers View paper
- [11] Efficient Test-Time Scaling via Self-Calibration View paper
- [12] Accelerating LLM Reasoning via Early Rejection with Partial Reward Modeling View paper
- [13] Inference-aware fine-tuning for best-of-n sampling in large language models View paper
- [14] SETS: Leveraging Self-Verification and Self-Correction for Improved Test-Time Scaling View paper

- [15] ReVeal: Self-Evolving Code Agents via Iterative Generation-Verification View paper
- [16] On the Role of Feedback in Test-Time Scaling of Agentic AI Workflows View paper
- [17] Budget-aware Test-time Scaling via Discriminative Verification View paper
- [18] Critique to Verify: Accurate and Honest Test-Time Scaling with RL-Trained Verifiers View paper
- [19] Inference Scaling fLaws: The Limits of LLM Resampling with Imperfect Verifiers View paper
- [20] Instance-Adaptive Inference-Time Scaling with Calibrated Process Reward Models View paper
- [21] Scaling Test-time Compute for LLM Agents View paper
- [22] Test-Time Scaling with Repeated Sampling Improves Multilingual Text Generation View paper
- [23] Putting the Value Back in RL: Better Test-Time Scaling by Unifying LLM Reasoners With Verifiers View paper
- [24] Video-T1: Test-Time Scaling for Video Generation View paper
- [25] Probabilistic Inference for Inference Time Scaling of Language Models View paper
- [26] Step-level Verifier-guided Hybrid Test-Time Scaling for Large Language Models View paper
- [27] CARINOX: Inference-time Scaling with Category-Aware Reward-based Initial Noise Optimization and Exploration View paper
- [28] VerifierQ: Enhancing LLM Test Time Compute with Q-Learning-based Verifiers View paper
- [29] Test-time Scaling Techniques in Theoretical Physics -- A Comparison of Methods on the TPBench Dataset View paper
- [30] Test-time Verification via Optimal Transport: Coverage, ROC, & Sub-optimality View paper
- [31] Evaluating the Role of Verifiers in Test-Time Scaling for Legal Reasoning Tasks View paper
- [32] When Does Verification Pay Off? A Closer Look at LLMs as Solution Verifiers View paper
- [33] Differential Privacy Meets Test-Time Scaling: Theoretical Guarantees under Noisy Inference View paper
- [34] ZIP-RC: Zero-overhead Inference-time Prediction of Reward and Cost for Adaptive and Interpretable Generation View paper
- [35] Rethinking Optimal Verification Granularity for Compute-Efficient Test-Time Scaling View paper
- [36] Taming Imperfect Process Verifiers: A Sampling Perspective on Backtracking View paper
- [37] Probing the effectiveness of World Models for Spatial Reasoning through Test-time Scaling View paper
- [38] ARCANE: A Multi-Agent Framework for Interpretable and Configurable Alignment View paper
- [39] Rethinking Reward Models for Multi-Domain Test-Time Scaling View paper
- [40] Weaver: Shrinking the Generation-Verification Gap by Scaling Compute for Verification View paper
- [41] Test-Time Scaling for Multistep Reasoning in Small Language Models via A* Search View paper
- [42] Adaptive Inference Scaling via Monte Carlo Sampling View paper
- [43] Scaling Natural-Language Graph-Based Test Time Compute for Automated Theorem Proving View paper
- [44] Large Language Monkeys: Scaling Inference Compute with Repeated Sampling View paper
- [45] Inference-Time Scaling for Generalist Reward Modeling View paper
- [46] Probabilistic Optimality for Inference-time Scaling View paper
- [47] Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for LLM Problem-Solving View paper
- [48] The art of scaling reinforcement learning compute for llms View paper
- [49] Inference-time scaling for complex tasks: Where we stand and what lies ahead View paper
- [50] Investigating test-time scaling with reranking for machine translation View paper
- [51] A Theory of Inference Compute Scaling: Reasoning through Directed Stochastic Skill Search View paper
- [52] Inference Scaling for Long-Context Retrieval Augmented Generation View paper
- [53] Amix-1: A pathway to test-time scalable protein foundation model View paper
- [54] Detecting and preventing hallucinations in large vision language models View paper
- [55] Fast best-of-n decoding via speculative rejection View paper
- [56] Accelerating best-of-n via speculative rejection View paper
- [57] Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment View paper
- [58] Optimal Stopping vs Best-of- for Inference Time Optimization View paper
- [59] Diverse Inference and Verification for Advanced Reasoning View paper
- [60] On the Query Complexity of Verifier-Assisted Language Generation View paper
- [61] STARS: Segment-level Token Alignment with Rejection Sampling in Large Language Models View paper
- [62] Best-of-Majority: Minimax-Optimal Strategy for Pass@ Inference Scaling View paper
- [63] Enhancing Fine-Tuning-Free Clinical Reasoning via Test-Time Scaling View paper