Enhancing Agent Reliability in Interactive Tool-Calling Environments via Test-Time Scaling

Arizona State University

Group Name: PromptOps

Mentor Name: Amir Saeidi

Jahnvi Seth, Pranesh Somasundar, Lekshman Babu Devendra Babu, Sravanakumar Sathish, Samudyata Sudarshan Jagirdar

Abstract

. **Problem:** Large language model agents frequently make unreliable decisions in multi-step tasks requiring planning, validation, and rule adherence

. **Approach:** We explore test-time scaling to enhance reasoning during inference rather than retraining models

. **Implementation:** Evaluated 5 strategies - Best-of-N, TTI, Budget Forcing, DBS and SVR on the tau-bench benchmark

. **Key Results:** These methods significantly reduced premature actions, prevented invalid tool calls, and enhanced multi-step reasoning capabilities.

Introduction

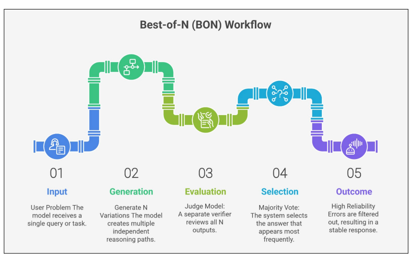
. **The Reliability Gap:** Despite high language fluency, agents frequently fail at real-world tasks by overlooking domain rules, losing track of user intent, or abandoning multi-step goals.

. **The "Thinking" Hypothesis:** Recent research suggests that performance gains come from enabling models to "think" more carefully during inference rather than solely from scaling model parameters.

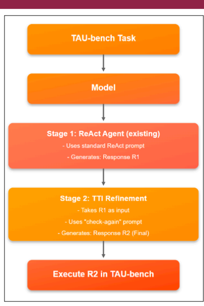
. **Our Objective:** We investigate whether expanding the reasoning horizon at test time can bridge the gap between raw capability and trustworthy, context-aware tool use.

Method

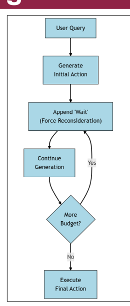
Best-of-N



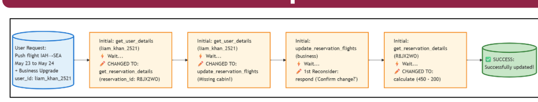
TTI

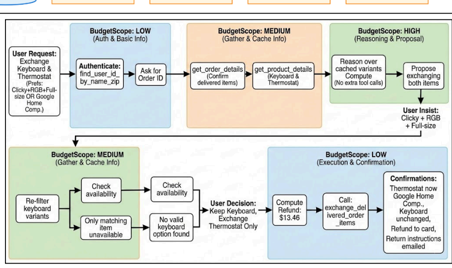


Budget Forcing



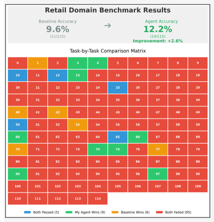
Examples

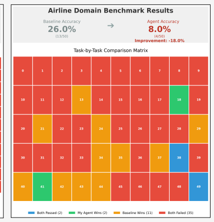




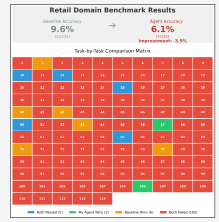
Experimental Results

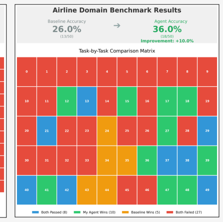
Best of N



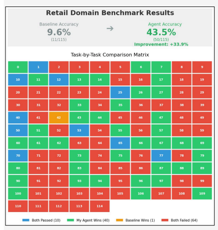


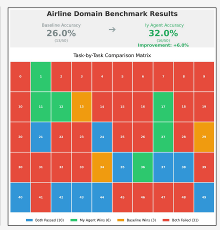
Budget Forcing



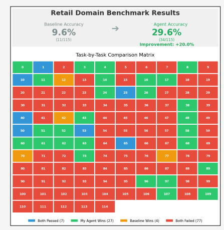


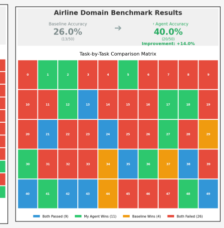
Dynamic Budget Steering





Simulate-Verify-Replan





Test-Time Interaction

Retail improved by 0.9%
Airline improved by 8%

Environment	Pass^1	Pass^2	Pass^3	Pass^4	Pass^5
Airline	0.368	0.310	0.284	0.268	0.260
Retail	0.115	0.066	0.053	0.047	0.043

(a) Pass^5 for TTI round 1

Environment	TTI-1	TTI-2	TTI-4	TTI-6
Airline	0.368	0.280	0.300	0.340
Retail	0.115	0.157	0.139	0.113

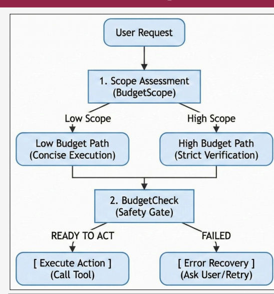
(b) Ablation across TTI refinement rounds

Table 1: Test-Time Interaction (TTI) results across environments

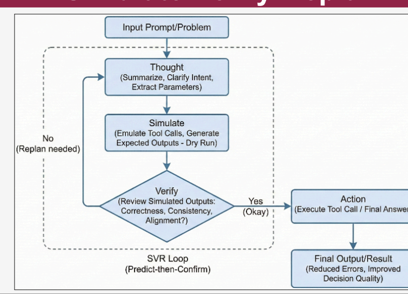
$$\text{Pass@k} = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}$$

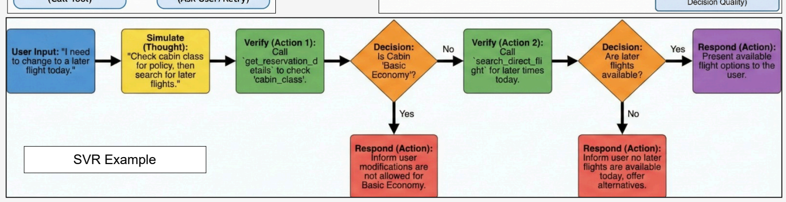
$$\text{Pass}^k = \left(\frac{c}{n}\right)^k$$

DBS



Simulate-Verify-Replan





Conclusion

Our 4B parameter model, enhanced with Dynamic Budget Steering (DBS) and Simulate-Verify-Replan (SVR), outperforms significantly larger proprietary models, achieving 43.5% in Retail (surpassing Claude-3-Sonnet's 26.3%) and 40.0% in Airline (exceeding GPT-4o's 35.2%), proving that inference-time architecture is more effective than parameter scaling for complex tool-use

References

[1] Ananya Hariharan et al. Plan verification for llm-based embodied task completion agents, 2025.
[2] Junyan Li, Wenshuo Zhao, Chuang Gan, and Yang Zhang. Steering llm thinking with budget guidance, 2025.
[3] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025.
[4] Junhong Shen, Hao Bai, Lunjun Zhang, Yifei Zhou, Amrith Setlur, Shengbang Tong, Diego Caples, Nan Jiang, Tong Zhang, Ameet Talwalkar, and Aviral Kumar. Thinking vs. doing: Agents that reason by scaling test-time interaction, 2025.
[5] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters.
[6] Yiming Wang, Pei Zhang, Siyuan Huang, Baosong Yang, Zhuosheng Zhang, Fei Huang, and Rui Wang. Sampling-efficient test-time scaling: Self-estimating the best-of-n sampling in early decoding, 2025.
[7] Zhen Wang et al. Simulating environments with reasoning models for agent training, 2025.
[8] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains
[9] Qi Zeng et al. Agentrefine: Enhancing agent generalization via instruction tuning, 2025.
[10] Xinyu Zhu et al. Verification-aware planning for multi-agent systems, 2025.