

Evaluating Agents on Open-Ended Problems

Haozhe Chen · General Examination · Princeton University

Committee: Zhuang Liu (advisor), Karthik Narasimhan (advisor), Peter Henderson

Date & Time: Tuesday, October 6, 2026, 2:00-4:00 PM EDT

Location: Computer Science 301 (CS 301)

Zoom: <https://princeton.zoom.us/j/8182081590?pwd=bWDwHinOr9YrDFPVlaXXDJIA7E3T9a.1>

Abstract

Can language model agents solve problems when no one tells them how, or even what, to solve? Agents are becoming reliable at well-specified tasks, but many valuable problems are open-ended: the path to success is unknown, or the problem itself must first be discovered. In this talk, I study how to evaluate agents on both kinds of problems. I first introduce CEO-Bench, in which an agent runs a simulated startup for 500 days in a noisy, changing market. It reveals distinct strategic behaviors across frontier agents, and a large gap in their ability to consistently plan and execute long-horizon strategy in an open world. I then introduce SWEeper-Bench, in which an agent must find and fix unreported bugs in real web applications. The best agent passes 59% of tasks, yet agents pass 96% once told what the bug is. In both settings, agents execute well but lack the judgment to decide what matters. I conclude by arguing that spoken conversation offers a scalable way to close this gap: it can capture how people think across diverse, real-world settings, which current training data rarely contains.

Reading List

Textbooks

1. Daniel Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed. draft, 2026). Chapters: Large Language Models; Transformers; Post-training: Instruction Tuning, Alignment, and Test-Time Compute.
2. Chip Huyen. *AI Engineering: Building Applications with Foundation Models*. O'Reilly, 2025. Ch. 3–4: Evaluation; Ch. 6: RAG and Agents.

Papers

Language agents

1. Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR*.
2. Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *NeurIPS*.
3. John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik

Narasimhan, Ofir Press. 2024. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *NeurIPS*.

4. Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *UIST*.

Evaluating agents on well-specified tasks

5. Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, Karthik Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *ICLR*.
6. Shunyu Yao, Noah Shinn, Pedram Razavi, Karthik Narasimhan. 2025. τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *ICLR*.

Long-horizon and open-ended evaluation

7. Axel Backlund and Lukas Petersson. 2025. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents. arXiv:2502.15840.
8. Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, et al. 2025. MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. In *ICLR*.
9. Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, et al. 2025. GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. OpenAI.
10. John Yang, Kilian Lieret, Jeffrey Ma, Parth Thakkar, Dmitrii Pedchenko, Sten Sootla, Emily McMilin, Pengcheng Yin, Rui Hou, Gabriel Synnaeve, Diyi Yang, Ofir Press. 2026. ProgramBench: Can Language Models Rebuild Programs from Scratch? arXiv:2605.03546.