

Failure Modes in Agentic AI: From Reproducible Triggers and Diagnostics to Reliable Fixes

Zoey Sha Li, Zihan Wang, Hejie Cui, Manling Li

<https://failure-mode-agent.github.io/>

Foundation-model agents are being deployed faster than the community can characterize how they fail. In barely two years they have moved into browsing, software engineering, computer use, scientific discovery, and embodied control, where reliability is governed by long-horizon, closed-loop interaction that current single-turn evaluation paradigms rarely capture. Failures (error cascades, reward hacking, mode and reasoning collapse, memory drift, and over-reasoning loops) still lack a shared vocabulary, reproducible triggers, and comparable diagnostics. This cutting-edge tutorial gives a systematic, mechanism-first account of agentic failure modes across four axes: (1) environments and infrastructure, (2) training dynamics, (3) evaluation and red teaming, and (4) test-time scaling and adaptation, with safety threaded throughout. Attendees leave with reproducible triggers, comparable diagnostic protocols, and verifiable mitigations.

Zoey Sha Li

email: mail@zoeyli.com

website: zoeyli.com

Zoey Li is a Senior Applied Scientist at Amazon. She works on pretraining scaling, data selection, and post-training diagnostics. She received her Ph.D. from the University of Illinois Urbana-Champaign in 2024, advised by Jiawei Han and Heng Ji.

Zihan Wang

email: zihanw@u.northwestern.edu

website: <https://zihanwang314.github.io/>

Zihan Wang is a PhD student at Northwestern University, advised by Manling Li. He works on reinforcement learning for reasoning agents in multi-turn, stochastic environments. He is the lead author of the RAGEN series, which characterizes core agentic failure modes.

Hejie Cui

email: hejie.cui@stanford.edu

website: <https://hejiecui.com>

Hejie Cui is a Senior Research Scientist at Google. She works on large language models and agentic AI systems. She received her Ph.D. in Computer Science from Emory University, advised by Carl Yang, and was a postdoctoral researcher at Stanford University with Nigam H. Shah and Sanmi Koyejo, focused on LLM post-training and evaluation.

Manling Li

email: manling.li@northwestern.edu

website: <https://limanling.github.io/>

Manling Li is an Assistant Professor at Northwestern University, and Amazon Scholar. Her research focuses on grounded reasoning and multimodal knowledge for foundation-model agents, recognized by the ACL 2025 Dissertation Award Honorable Mention and paper awards from ACL and NAACL. She co-leads multiple recent open-source efforts on agentic AI, and has organized community events around foundation models and multimodal reasoning.