
AGENTS POST-TRAINING EVALUATION AND ALIGNMENT FEBRUARY 25, 2026

VERO: AN EVALUATION HARNESS FOR AGENTS TO OPTIMIZE AGENTS

Varun Ursekar, Apaar Shanker, Veronica Chatrath, Yuan Xue, Sam Denton

VeRO benchmarks whether coding agents can improve other AI agents by editing prompts, tools, and workflows under controlled evaluation.

An important emerging application of coding agents is *agent optimization*: the iterative improvement of a *target agent* through edit–execute–evaluate cycles. Despite its relevance, the community lacks a systematic understanding of coding agent performance on this task. Agent optimization differs fundamentally from conventional software engineering: the target agent interleaves deterministic code with stochastic LLM completions, requiring structured capture of both intermediate reasoning and downstream execution outcomes. To address these challenges, we introduce VeRO (**V**ersioning, **R**ewards, and **O**bservations), which provides (1) a reproducible evaluation harness with versioned agent snapshots, budget-controlled evaluation, and structured execution traces, and (2) a benchmark suite of target agents and tasks with reference evaluation procedures. Using VeRO, we conduct an empirical study comparing coding agent optimizer configurations across tasks and analyzing which modifications reliably improve target agent performance. We release VeRO to support research on agent optimization as a core capability for coding agents.

LINKS

<https://arxiv.org/pdf/2602.22480>