ORCA-bench: How Ready Are Language Model Agents for Oncall?

Albert Gong[1], Kyuseong Choi[1], Abhineet Agarwal[2], Jason Schechner[2], Ryan Huang[2], Raj Agrawal[2], Anish Agarwal[2,3], Raaz Dwivedi[1,2]
[1]Cornell Tech [2]Traversal [3]Columbia University

arXiv paper · GitHub code + tasks · Harbor Hub dataset · Fireworks leaderboard

Oncall RCA is not SWE-bench.

Alert | **Detection** | **Root cause** | Mitigation | Resolved

ORCA-bench measures this hop

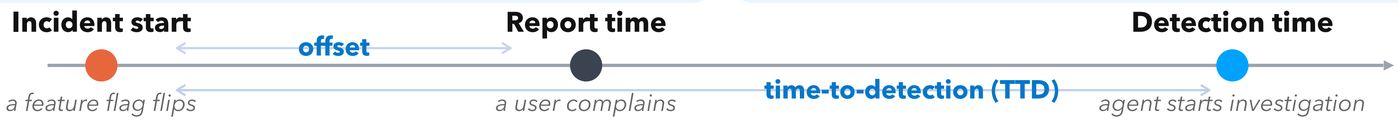
	SWE-bench	Oncall RCA
Input	A bug report or a failing test	**A vague complaint**
System	A frozen repository	**A live distributed deployment + multiple services**
Evidence	Static codebase	**Telemetry + codebase**
Verdict	Re-run the programmatic test suite	**Compare to human-curated postmortem**

RCA still runs on human intelligence.

Tasks: one incident × 4 knobs → 1,079 RCA tasks

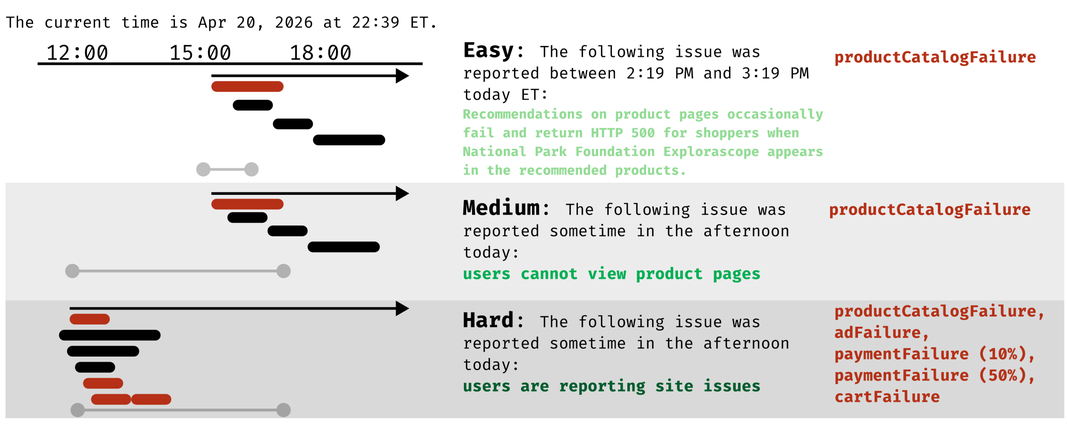
1 Offset
How long after onset the user reports and how vaguely: "at 2:30 pm" → "sometime in the afternoon"

2 Time-to-detection (TTD)
How long after onset the agent starts investigating: 15 min · 1 h · 8 h · 24 h



3 Context ladder
How much the report says: Easy / Medium / Hard – 2.0 → 4.4 plausible root causes

4 Scenario
Which faults co-occur: isolated · independent · conflicting · cascading · sequential



EASY **2.00** plausible root causes · n = 288
MEDIUM **2.84** plausible root causes · n = 316
HARD **4.41** plausible root causes · n = 280

Built on OpenTelemetry Astronomy Shop: 6 days of continuous telemetry (**50 GB**) + full source code; 11 real feature flags, 13 expert-designed scenarios, 195 controls.

Human-verified at every step

Ground truth curation
Telemetry ground truth validated by **two expert SREs** across 11 flags / 42 firings. Schedule written with an expert SRE.

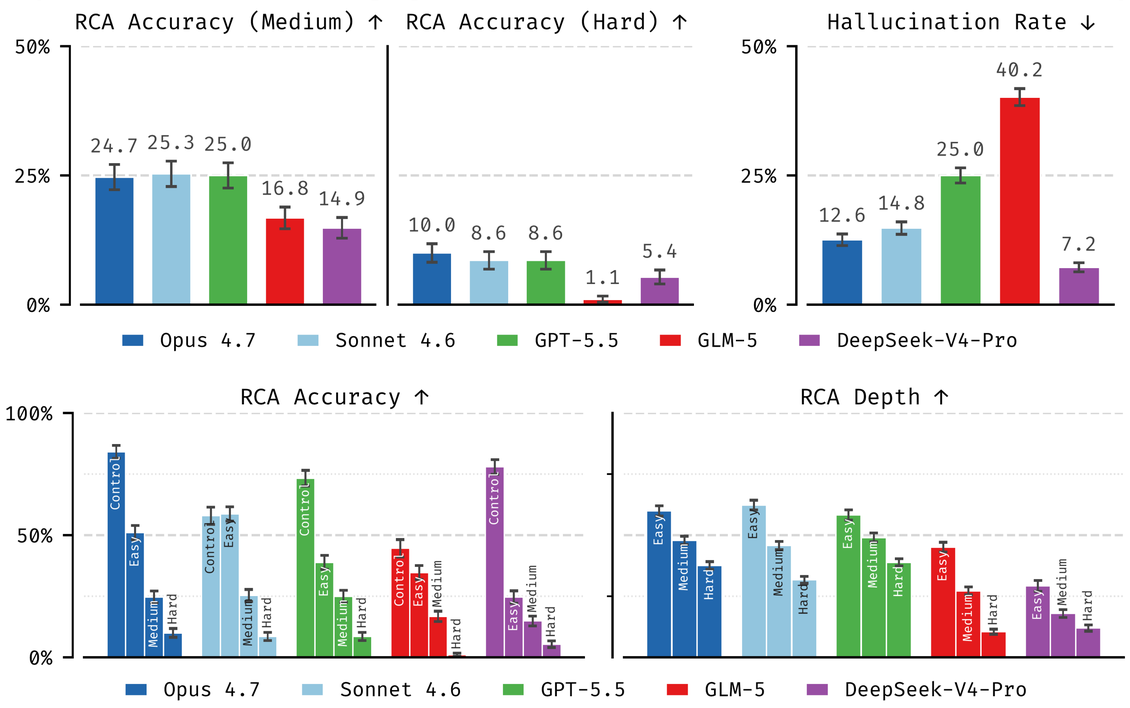
ORCA-bench Verified
40 tasks, every ground-truth label and every per-model score hand-confirmed by two authors, independently.

LLM JUDGE, RE-SCORED FROM SCRATCH
0.91 judge ↔ humans
0.83 human ↔ human
Cohen's κ

THREE METRICS
RCA accuracy – named every plausible root cause
RCA depth – root cause, mechanism, evidence
Hallucination – named a cause matching none

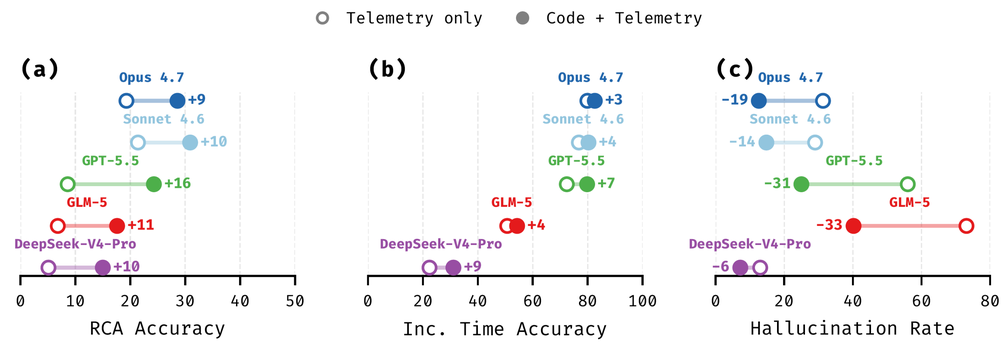
The best agent solves 25.3% of medium tasks – and 10.0% of hard ones.

as of May 2026 · five frontier coding agents · 884 incident tasks · Terminus-2 harness · GPT-5.4 judge

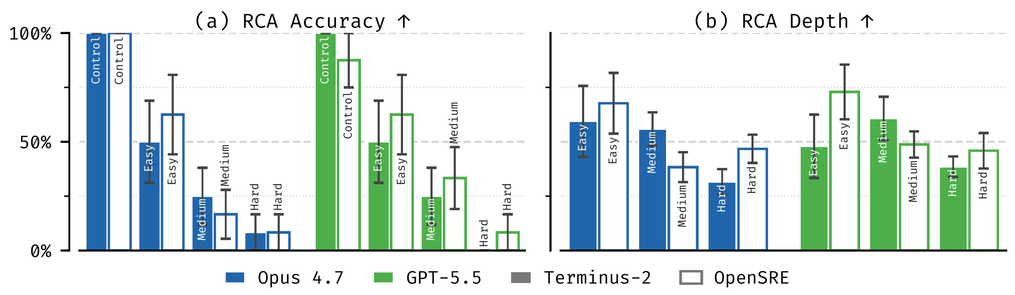


Every model loses 19–50 points from easy to hard. With six co-occurring faults, three models each found exactly one root cause – all different.

Source code matters; the harness does not.



Code + telemetry vs. telemetry only



Terminus-2 vs. OpenSRE on ORCA-bench Verified

A public floor, and a private ceiling.

WHERE ORCA-BENCH SIMPLIFIES:
- **Scale** – 50 GB over six days; production ships terabytes a day
- **Priors** – the Astronomy Shop is public, and probably pretrained
- **No memory** – every task investigated cold
- **Read-only** – no fix-and-verify action loop

THE CEILING · TRAVERSAL INTERNAL, PRIVATE
Sev 1–3 incidents at a large financial-services enterprise:
4,000+ services · 500 TB+/day · 100K+ repos
150+ people pulled into a major incident, for hours

The public benchmark is the floor; the private one shows how much further the real end is.