

Plumbline Grader: fact sheet

For reporters. Every figure below is public on plumblinegrader.com as of October 15, 2026.

What we do. Plumbline Grader audits the graders that train and rank AI models: the unit tests, AI judges, rubrics and state checks that decide what earns reward. We search each task for wrong work the grader pays full reward for, prove every case, and write the tests that close the gap.

Why it matters. A model in training learns what the reward pays for, not what its designers meant. When a test misses a requirement, a fix that skips it earns the same full reward as one that gets it right, and a model learns to pass tests instead of solving tasks. Benchmark scores built on the same tests stop meaning what they say.

What our audits of public task sets found

11	public coding task sets audited
190 of 487	testable tasks paid full reward for a wrong fix in one search pass: about 4 in 10 pooled, from under 20% to 69% by set
424 of 425	findings we re-ran from scratch reproduced
738 of 795	scope calls upheld by a two-of-three majority of AI models from three labs (Claude, GPT, Gemini), each voting without seeing our label (the 9 sets of our accuracy study)
20+	public benchmarks and evaluation suites whose maintainers have heard from us privately since October 2, 2026, before anything is published
4	findings public as GitHub issues on Inspect Evals, the open-source evaluation suite published by the UK AI Security Institute, filed at the maintainers' request on October 13, 2026, each with a tested fix

Scope. A task is testable when its reference fix passes its grader, an empty fix does not, and at least one wrong fix ran. One fixed search per task: up to 8 versions of the reference fix with parts removed, plus 6 wrong fixes written by AI models. A deeper search finds more, so 190 is a floor. Set-by-set results publish in the Grader Integrity Index on November 10, 2026, after each maintainer's private notice period ends.

A public example (Inspect Evals issue #2676). In the banking tasks of Inspect Evals 0.24.0, setup records were shared by reference, so the grader rebuilt its expected database from the agent's own changes. A run with one unrequested change was paid on 24 of 24 tasks. After the fix (a deep copy), it is paid on 0 of 24.

How a finding is proved

1. The task's own grader, unchanged, runs in a sealed container: a fresh container per run, network off.
2. The grader pays full reward for a wrong fix.
3. A separate behavioral check passes the reference fix and fails both the unfixed code and the wrong fix, then again in fresh containers. The task as written must require the behavior.
4. Replacement tests must fail every wrong fix found and still pass the reference fix, and are attacked again with fresh wrong fixes.
5. Every finding ships with `reproduce.sh`: one command, bash and Docker. No Plumbline code, no API key, no model.

Who checks what. The audit engine is built and run with AI agents, under the founder's direction. Findings are decided mechanically (grader result plus separate check, re-run in fresh containers). AI models from three companies vote on scope; we always say which checks a model did and which a person did.

How we work with the people we audit

- **Private notice first.** Findings on a public task set go to its maintainers privately, with the evidence and, wherever we can build one, a tested fix, at least 30 days before anything is published. Notices carry no prices and no offers.
- **Right of reply.** Any maintainer or vendor named in something we publish can respond, and we publish the reply alongside our findings.
- **Independence.** We never sell environments, and never certify one we built. Flat fees, fixed before we start, never per finding and never tied to a result. Index numbers are never for sale.

Public documents

Method specification v0.3.0, Plumblin Verified criteria v1.0, independence policy, disclosure policy, and the reproduction kit: plumblinegrader.com/about. The open definition of reward integrity (a term already used in AI research; we did not coin it): rewardintegrity.org, which we steward.

Company

Plumblin Grader is operated by IROS Systems LLC, Indiana. Founded 2026. **Founder:** Shane Cope, DDS, MSD. Dentist and technology entrepreneur; built HIPAA-regulated practice software with AI coding agents; directed AI agents to build plumblin-detector, an open-source tool whose pre-registered study covers 93,014 merged changes in 262 public repositories.

Press contact: Shane Cope, shane@plumblinegrader.com. Reporters get an answer the same business day. More: plumblinegrader.com/press