

Plumbline Verified: criteria

Version 1.0. Effective 10 November 2026.

Plumbline Verified is a certificate for one named, versioned release of a task set: a benchmark or a training set in which each task comes with tests that decide whether a solution is correct. The certificate covers the tests. It says that we searched hard for solutions that are wrong but still pass the tests, closed every such hole we found, then searched the repaired release again in a separate audit and found too few holes to matter, by the rule below.

Words used here

- **False pass:** a solution that the task's tests accept although it does not do what the task asks. We count one only when a separate check, run in the task's own sealed container, confirms it. The check must pass on the task's reference solution, fail on the unsolved code and fail on the wrong solution.
- **In scope:** the behavior the wrong solution gets wrong is something the task, as written, actually requires. A wrong solution that breaks only something the task never asked for is recorded but does not count against the release.
- **Hackable task:** a task with at least one in-scope false pass.

The six criteria

A release is certified when all six hold.

1. **Audit.** We audit a random sample of at least 50 tasks of the release, or every task if the release has fewer than 50.
2. **Repair.** Every in-scope false pass that audit confirms is closed with replacement tests. We verify each replacement in the task's container: the reference solution must still pass, and every known wrong solution must now fail. The vendor adopts the replacements in the release. The vendor may write its own replacements instead of ours. They are verified the same way.
3. **Re-audit.** A fresh audit of the repaired release, on a new random sample, finds few enough hackable tasks that the 95% upper bound on the rate is under 10% (Wilson interval). In practice:
 - with 50 tasks, that means none;
 - with 100 tasks, at most 4;
 - with 200 tasks, at most 11.

Releases of fewer than 35 tasks. No result on fewer than 35 tasks can meet that bound, not even zero hackable tasks: 0 of 34 still has an upper bound of 10.2%. A small release is therefore re-audited in full, every task, and is certified only if no task is hackable. Its certificate still prints the upper bound (for 20 tasks, 16.1%) so readers can see how much a small release can show.

4. **Review.** The findings and the rate are reviewed by our standard review panel. Models from three different companies each judge every finding blind. The majority decides, and their agreement is printed.

5. **Wider search.** The re-audit attacks each task in two ways:

- requirement by requirement: one wrong solution for each requirement of the task;
- from known hole patterns: kinds of weak test seen before.

Two attacker models from different companies write the wrong solutions. Any single search misses holes: in our pilot, the wider search found far more hackable tasks than one standard audit of the same tasks. So the certificate also reports an estimate of how many hackable tasks all the searches together missed. The estimate is computed from how often the separate searches found the same tasks (Chao2). It is printed beside the confirmed count, never in its place.

6. **Independence.** The re-audit is its own run. It is never a run that wrote or verified any of the release's replacement tests. Our independence policy also applies in full:

- fixed fees, never tied to the outcome;
- the method fixed before the audit starts;
- disclosure of any tests we wrote;
- a human reviewer who did not run the audit must check a random sample of findings, with hidden controls, before any certificate is issued. No certificate has been issued, and no such review has been done yet.

What a certificate states

- The release name and version, the vendor, and the date issued.
- The criteria version (this page, v1.0) and the method version, with the run records that back every number.
- For the first audit:
 - how many tasks were audited, of how many in the release;
 - how many in-scope false passes were found;
 - how many were repaired, and by whom.
- For the re-audit:
 - how many tasks were audited;
 - how many were hackable;
 - the 95% upper bound on the rate;
 - the estimate of hackable tasks the searches missed.
- The review panel's agreement and the human reviewer's agreement.
- The number of tasks whose tests Plumblin Grader wrote or verified, if any.

What a certificate does not state

- **It is not a guarantee** that no task in the release can be gamed. Every search misses something. The certificate says what was tested and what was found, and gives an estimate of what was missed.

- **It does not cover tasks outside the audit's sample**, other than through the rate and its interval.
- **It does not cover other releases.** A new version needs a new certificate.
- **It does not cover out-of-scope behavior** (things the task never asked for) or infrastructure, such as the harness, containers or scoring code outside the task's tests.
- **It does not grade the tasks** for difficulty, usefulness, or freedom from contamination.

Validity

- A certificate covers one named release and is valid for 12 months from the date issued, as long as that release's tests do not change.
- **Updates.** Under Keep Current, every new or changed task in an update is audited to these criteria. If they pass, the certificate moves to the new version with the same expiry date. If more than a quarter of the release's tasks are new or changed, the release needs a full re-audit.
- **Renewal.** After 12 months, a fresh re-audit on a new sample, under the criteria version then in force. A certificate is judged by the criteria version printed on it; a new version of these criteria never cancels an existing certificate.
- **Badge.** The badge is served by Plumblin Grader and shows the release and its expiry date. It reads "expired" or "superseded" automatically when the certificate lapses or the vendor ships a version it does not cover.
- **Errors.** A certificate found to have been issued in error is withdrawn, and the registry states why.

Failing

A release that does not meet all six criteria gets no certificate. We never issue a partial or softened one. The vendor may repair the release and ask for a new re-audit; that re-audit is a new run on a new sample.

Registry

Each certificate is listed in a public registry with the figures above and a badge that links to its entry. Run records are kept for at least three years, so anyone can check a certificate.

Changes

- **v1.0 (effective 10 November 2026).** First public version. Changes from the October 2026 draft:
 - a rule for releases under 35 tasks (criterion 3);
 - the wider search and its estimate (criterion 5);
 - the re-audit must be independent of every repair run (criterion 6).