

Assessment attestation

SAMPLE. A moderation guardrail for community posting. Independent AI red teaming on the two free public tools.

SUBJECT OF ASSESSMENT	A moderation tool for a large discussion platform. Two free public tools, a rules checker that returns a posting verdict for a community, and a draft checker that runs a written post through nine published checks. Both usable without an account.
ASSESSMENT DATE	One day in September 2026. Single day, single tester.
FULL REPORT	Version 1.0, 13 pages, ISTQB defect format. This sheet summarises it and does not replace it.
REFERENCE	PL-XXX-2026-09-DD
ASSESSOR	Nikita Kalachev, Platilus LLC (Georgia). QA engineer, 7+ years, ex-EPAM.
ENGAGEMENT TYPE	Unsolicited and unpaid. No account was created, no trial was started, nothing was posted to the platform, and no system belonging to a third party was connected.

In scope

- The rules checker verdict and the printed reason behind it, on six communities.
- All nine checks of the draft checker, on six drafts each built so that exactly one check should fail.
- The handover between the two tools: whether the second re-derives the verdict or consumes it.
- The free site scan, on a domain belonging to the tester.

Out of scope, stated on purpose

- The platform itself. It was not readable from the test environment and no workaround was used.
- Anything behind the paid tier, behind an account, or behind the tools API, which returned a refusal that was respected.
- Infrastructure, authentication and dependencies. This was not a penetration test.
- Any determination of the product's regulatory classification.

Method

Behavioural assessment against an oracle taken from the product's own published statements. Each draft was written so that exactly one of the nine published checks should fail and the other eight should pass, which converts a judgement about output quality into a pass or fail per check. The community was then varied while the draft text was held identical, so that any difference in outcome is attributable to the verdict rather than to the writing. Controls were run in both directions: communities whose verdict rests on a quoted rule, to confirm the machinery works, and the case under test. Every quotation in the report was read from the live page rather than from a summariser, and each page was captured in full before an absence was claimed.

Result

One finding, rated **Critical** and reproduced on two communities with four controls: for a community that has published no rule about promotion, the rules checker returns a permissive verdict inferred from the community description, although the product's own page defines that case as one to treat as prohibited. The draft checker then consumes that verdict, passes a product mention there and suggests adding a link. Eight checks held, including four correct verdicts built on quoted rules and exact per-check isolation on three of three drafts. One further defect of lower severity is documented and held back. One observation could not be reproduced and is reported as unreproduced rather than as a finding.

Categories referenced

ITEM	CATEGORY AND EDITION	WHAT WAS TESTED
XXX-001	ASI09 Human-Agent Trust Exploitation, OWASP Top 10 for Agentic Applications, 2026 edition. Secondary: LLM07:2026 Misinformation, OWASP GenAI LLM Top 10, version 1.0 of 3 August 2026.	Whether a verdict produced without the evidence it depends on reaches the user with the same confidence as one built on a quoted rule, and whether the downstream check re-derives it or consumes it.

Validity and re-test

This attestation describes behaviour observed on the assessment date and carries no forward-looking assurance. Behaviour that depends on a model can change without a release of the product, so an attestation that claimed a period of validity would be claiming something it cannot know. What is offered instead is repeatability: the corpus of community to expected verdict, and the drafts built to isolate each check, are described in the report well enough to rebuild, and Platilus LLC will re-run them against the same surface on request and issue a dated re-test attestation stating which findings reproduce and which do not.

What this attestation is not

- It is not a certification, and it is not a statement of compliance with any standard or regulation.
- Platilus LLC is not a notified body and does not perform conformity assessment.
- Whether this product falls within Chapter III of the EU AI Act is the provider's own determination. It was not examined here, and no article of that regulation is asserted to apply.
- The categories above are cited so the result can be referenced in a security questionnaire. Citing a category is a description of what was tested, not an audit against that list in full.
- It covers one mechanism on two surfaces over one day. It is not a survey of the product.