The Recount · Study 01 — The Measurement Gap

Publisher: MarginSignal OS · The Recount (the independent standard for whether AI worked) **Author:** Brandon Burdin, Founder & Principal Analyst **Version:** v1 · **Published:** 2026-07-18 · **Canonical:** https://www.marginsignalos.com/recount **License:** CC BY 4.0 (cite with attribution)

Finding

There is no single, stable "AI hallucination rate." The reported number depends on **which method, which version, and which date** — and it moves. Any single self-reported accuracy figure for a deployed AI system is therefore close to meaningless without a fixed, named, dated method behind it.

Evidence (primary sources; verified 2026-07-18)

Illustrated with the public, well-regarded work of one measurement provider, **Vectara** — cited as an ally in rigor, not a target.

1. Vectara maintains a public hallucination leaderboard scored by its HHEM detection model. The detection model is itself versioned (HHEM 2.1 → 2.3). Source: https://github.com/vectara/hallucination-leaderboard (HHEM-2.3; leaderboard last updated 2026-05-11).
2. The same provider maintains **FaithJudge**, a stricter LLM-as-judge faithfulness method that surfaces materially higher hallucination rates than the summarization leaderboard — two credible methods, different numbers for the same models. Sources: https://github.com/vectara/FaithJudge ; arXiv:2505.04847.
3. The rankings move over time. As of the 2026-05-11 leaderboard update (HHEM-2.3), the lowest-hallucination model was measured at 3.3% — a point-in-time

snapshot, not a fixed truth.

What we deliberately did NOT publish

A widely-circulated secondary claim pairs a single model at ~0.7% (HHEM) vs ~7.6% (FaithJudge) — an ~10× swing. On primary-source review this exact pairing could **not** be confirmed from a single authoritative Vectara source, and the ~0.7% model is no longer listed on the current board. We retired the claim rather than publish it. The near-miss is recorded here on purpose: it is the discipline the standard is built on.

Implication

If even the referee's own number is a moving target — two methods, a versioned yardstick, a shifting board — then a deployed system's single self-reported accuracy figure cannot be taken at face value. A standard supplies the missing piece: one fixed, named, dated, reproducible method. That is the reason The Recount exists.

Method & integrity

This study clears The Recount's four gates: (1) ground truth traced to primary sources with versions/dates; (2) reproducible — a reader can open the cited sources and reproduce the comparison; (3) adversarially checked — the unverifiable pairing was removed; (4) neutral — no vendor named as villain. Figures are point-in-time snapshots of living leaderboards and will change; that impermanence is the finding.

Suggested citation

Burdin, B. (2026). *The Recount, Study 01 — The Measurement Gap.* MarginSignal OS. https://www.marginsignalos.com/recount (v1, 2026-07-18).