VERACITY SOLUTIONS — EXECUTIVE RESEARCH

Evaluating a model is not the same as verifying its work.

A strong evaluation can tell you how an AI system performs across many tasks. It cannot tell you whether this analysis used the right data, the right calculation, or the right assumptions for the decision in front of you.

Author: Veracity Systems Architecture Group Published: 2026 Document ID: VER-WP-2026-04

When adopting AI tools, companies must not only decide if AI is useful but also meets their safety and compliance requirements. For regulated industries, the stakes of these decisions are especially high. The requirements from regulators and internal risk assessment teams do not disappear because of new intelligent systems. In fact, these obligations are emphasized to protect the business and its employees when facing autonomous artificial intelligence.

Thus, enterprises are increasingly improving their evaluation capabilities. They test representative tasks, compare models, measure failure modes, add guardrails, and monitor behavior. These evaluations are useful for understanding how a system behaves across a range of tasks and conditions. They can help an enterprise feel confident in an AI system. But the operational question remains: can a high scoring AI tool be trusted to produce work products that are accurate and reliable for downstream dependencies? Recent AI generated failures of reports filled with hallucinated citations, false claims, and contradicting calculations from firms like KPMG[1], PwC[2], and EY[3] tell us that we can't be so sure.

A benchmark cannot answer if this particular report, recommendation, analysis, or decision record is supported by the right evidence, under the right conditions, for the decision the organization is making right now. The product contains particular claims, citations, assumptions and omissions. It may become the input to a balance sheet, a sale recommendation, customer communication, or another consequential business decision. Yet many enterprises are lacking narrow controls on generative tools.[4]

Consider a banking analyst asking an agent to prepare a monthly risk analysis of the bank's loan portfolio. The agent calculates exposure by industry, tests whether concentrations exceed approved limits, estimates stressed losses, and determines the potential capital impact. The agent has performed well on evaluations of representative portfolio analyses, so the analyst feels confident in the report. They pass along the report to their team which acts on the report stating the relevant companies account for 18% of the portfolio, below the bank's 20% limit, and recommends no action. On that basis, the team closes the monthly concentration review without further escalation.

The problem is that the data was not complete. The extract was generated before 8% of recently added loans had been added to the bank's official portfolio records. With those loans included, the technology exposure is actually 22%, which is above the approved limit and should trigger escalation rather than a recommendation to proceed. The model is internally coherent. The recommendation is still unsupported for the decision in front of the institution.

Nothing in that example requires the agent to behave strangely. It does not need to invent a number or make an obviously irrational claim. The system can be good at portfolio analysis while this artifact is wrong for the workflow because its source population is incomplete for the reporting date. This is where system-level confidence breaks down. A human reviewer may still be the person who decides whether the work is usable. But a request to inspect every line of a polished artifact is not the same as giving that reviewer a defined basis for approval. The reviewer needs to know which data version was used, whether the population is complete, how the calculations were produced, which assumptions apply, and which claims remain unresolved (read more on our thoughts on human review).

> **The work is the unit of trust.**

Organizations cannot infer the quality of a singular output from assessments of aggregate behavior. The work itself is where claims, calculations, evidence, freshness, scope, authorization, and downstream consequence meet.

Verification operates at that narrower level. It checks a specific work product against a defined evidence set and the requirements that govern the workflow. The requirements can come from a regulator, an industry standard, company's internal policies, or the specific rules for a given analysis. For our bank analyst example, verification would confirm that the analysis used the authoritative extract for the correct reporting date, reconcile the population against the bank's portfolio system, independently recompute the industry concentration, and test the result against the bank's approved limit. A citation can show what the system saw and computed. By itself, it does not establish that the final claim follows from the source or that the source was current and authoritative. Verification closes that gap by checking not just whether evidence exists but if the work actually follows from the evidence and requirements that govern it.

Verification is not a proof of universal truth. It cannot repair bad source data, answer an underspecified business question, or replace judgment where materiality and interpretation are unresolved. Its value is more concrete: it establishes what was checked, against which evidence, under which conditions, and what remains open before the work becomes something businesses rely on.

The productivity opportunity with AI is substantial. The business obligation is still to stand behind the work that leaves the organization. The operating model is straightforward: evaluate the system to decide whether it belongs in the workflow. Verify the work to decide whether this output should be relied upon.

References & Regulatory Citations

[1] **KPMG** — https://techcrunch.com/2026/06/13/kpmg-pulls-report-on-ai-usage-due-to-apparent-hallucinations/

[2] **PwC** — https://www.forbes.com/sites/rachelwells/2026/07/31/pwc-reports-caught-with-ai-slop-is-a-warning-to-every-worker-right-now/

[3] **EY** — https://gptzero.me/investigations/ey

[4] **Yet many enterprises are lacking narrow controls on generative tools.** — https://newsroom.ibm.com/2026-06-08-new-ibm-study-finds-cios-and-ctos-face-growing-ai-control-gap-as-enterprise-deployment-scales