VERACITY SOLUTIONS — EXECUTIVE RESEARCH

The case for independent verification at enterprise AI

A Control Architecture for Independent Criteria, Evidence, Method, and Authority

Author: Veracity Systems Architecture Group Published: 2026 Document ID: VER-WP-2026-03

An AI system generates a credit memorandum for a commercial loan. It pulls borrower data, calculates leverage ratios, summarizes risk factors, and recommends approval. The memo is fluent, internally consistent, and produced by a model that performed well in pre-deployment evaluation.

But none of that answers the operational question that matters: can this specific memorandum be relied upon now?

The underlying ledger data may be stale. A claim may have been misread. A policy exception may have been omitted. The model may even review its own work and reproduce the same or perhaps a different error. Once the output can influence a real transaction, model or agent quality is no longer the only control question. The enterprise needs an independent mechanism to determine whether the output satisfies current evidence, policy, and authorization requirements.

The need for independent verification becomes most acute when the generated work crosses a boundary into consequential business reliance or execution. That underlying requirement is familiar to enterprise governance. The Institute of Internal Auditors Statement of Position on the Three Lines Model[1] formalizes a Three Line Model for organizations to facilitate strong governance and risk management. The model separates first line roles that provide services, second line roles that support and challenge on risk-related matters, and the third line roles that provide independent assurance and advice. Therefore, the longstanding governance pattern of independent oversight is not new. However, the existing guidance does not always scale to generative AI, and must evolve alongside AI. Updated banking guidance including Federal Reserve SR 26-2[2] and OCC Bulletin 2026-13[3] explicitly excludes generative and agentic AI from the scope of traditional model-risk guidance because the models are so novel and rapidly evolving. They also state that banking organizations should use their broader risk-management and governance practices to determine appropriate controls for tools and systems outside that scope. That leaves banking institutions with a practical control question: when a specific AI-generated artifact or action will be used in a consequential business decision, what controls should determine if it is reliable?

The NIST AI Risk Management Framework (AI RMF 1.0)[4] provides a useful foundation for answering this question through its emphasis on continuous testing, evaluation, validation, and verification (TEVV). The framework notes that processes for independent review can improve testing effectiveness and mitigate internal

bias and potential conflicts of interest. Its Measure 1.3 further calls for regular assessment involving internal experts who were not front-line developers and/or independent assessors. Benchmarks (LMArena[5]; Stanford HELM[6]; METR[7]) evaluate model capability; verification determines whether a specific artifact or action should be relied upon or released. For consequential workflows, this framework suggests that the system generating an output should not be the sole authority for determining its reliability. Instead, the criteria, evidence, and checks used to approve the output should be established and applied independently, subject to defined human or organizational authority.

Structural Limits of Self-Review

Self-review can improve the output, but it cannot be the final approval. Additional inference-time reasoning, self-consistency, or model-based critique do not introduce an independently governed source of evidence, policy, or authority. Empirical research demonstrates the limits of intrinsic self-correction of logical reasoning without external execution feedback, while empirical studies on LLM-as-a-judge patterns show systematic position, verbosity, and self-enhancement biases. These methods can still play an important role in generation workflows by improving quality, but they alone are not a sufficient gate for high-consequence decisions.

Intrinsic self-review remains exposed to correlated failure modes because the reviewing instance inherits the same underlying learned representations and many of the same reasoning tendencies as the producing model. Though the reviewing model may have different instructions and tools, it shares the same core capabilities and limitations with the generating model. External evidence, deterministic execution, independently governed criteria, or a separate authorization boundary introduce information and constraints outside the original generation process. Even a highly capable self-reviewer therefore cannot, by itself, establish independent evidence, define enterprise acceptance criteria, introduce a genuinely different verification method, or confer release authority. Those are properties of the surrounding control architecture.

The Four Dimensions of Independence

The credit memorandum example makes these four dimensions concrete. Before the recommendation can move from draft to decision, the enterprise must answer four independent control questions:

Criteria: Did the memo comply with the bank's current underwriting policy? Applicable limits and approval rules should come from governed enterprise policy, such as a defined exposure limit, rather than from thresholds improvised inside the model's prompt.

Evidence: Were its conclusions grounded in borrower and ledger records the bank recognizes as authoritative? Material claims should be reconciled against point-in-time systems of record rather than accepting the generator's representation of those records.

Method: Were the financial calculations independently recomputed? Method independence matters because verification that simply repeats the producer's reasoning can reproduce the same error. For claims reducible to arithmetic or formal logic, verification should use a mechanism with meaningfully different failure modes rather than asking another language model whether the answer looks right.

Authority: Can the agent approve the recommendation it generated? Production and execution permissions should be separated so the system that proposes an action cannot unilaterally authorize it.

Criteria, evidence, and authority primarily separate control ownership. Method independence serves a different purpose: it introduces failure-mode diversity. No single verification mechanism is sufficient for every claim. Self-review and cross-model review can improve reasoning quality, consistency, and ambiguity detection; deterministic checks provide stronger assurance for formal claims; human reviewers contribute judgment and, where delegated, release authority. Highly consequential workflows combine these controls based on risk and regulation.

How Control Architecture Works in Practice

A practical enterprise architecture can separate these responsibilities across four functional layers, carrying the AI generated artifact through each stage:

The Generation Layer drafts the candidate work. In this lending example, the Generation Layer sits within first-line business operations, where generative models and autonomous agents work in a controlled workspace, optimized for task completion, semantic reasoning, context synthesis, and drafting speed.

The Authoritative Evidence Layer exposes the borrower's snapshot and ledger state from systems of record, giving generation and verification a consistent point-in-time reference state.

The Verification Control Plane independently recomputes the financial ratios, checks policy exceptions, and evaluates the candidate against required evidence, policy, and control conditions. Governed by designated enterprise control owners, an independent verification engine records which model produced the memo, which evidence and policy versions were used, which checks ran, and what decision was reached. The resulting verification record can be integrity-protected and made tamper-evident. The Verification Control Plane itself becomes a high-value security boundary because its decision can determine whether an action proceeds. Its policies, identities, administrative access, rule changes, cryptographic keys, and monitoring therefore require controls commensurate with that authority.

In preventive deployments, the Release Boundary prevents approval or downstream execution when required verification conditions fail. For high-consequence workflows, verifier failure should itself trigger an explicit policy outcome, such as routing to manual review, rather than silently defaulting to release.

Independent verification does not require a single deployment model. In some workflows it blocks an action before execution; in others it reviews a document before approval; in still others it runs silently in parallel to measure failure rates before enforcement is turned on.

Vendor Portability Across Changing Models

Foundation models are evolving faster than the governance and control infrastructure enterprises use to manage them. The Stanford AI Index Report[8] finds that responsible-AI infrastructure is already failing to keep pace with AI advances and deployment, while the OECD AI Policy Observatory[9] has similarly called for more agile governance capable of responding to rapid technological change. A bank may use one model today, route some tasks to another model next quarter, and replace both the following year. Its policies, evidence requirements, approval authorities, and control history should not have to be rebuilt with each model change. When those controls are embedded inside a vendor-specific model stack, changing models can require reimplementing parts of the governance layer and re-establishing trust assumptions. Separating governance and verification from generation allows enterprise policies, control history, and verification infrastructure to persist across model providers and versions, while model-specific trust assumptions can be reassessed as the underlying systems change.

That portability matters most as AI systems gain authority to act.

Implications for Autonomous Systems

Generative AI changes the producer, but not the need for independent challenge. Model evaluation tells an enterprise how a system tends to perform; output verification determines whether a specific artifact or action can be relied upon under the evidence and policy that apply now (evaluating a model is not the same as verifying its work). A credit memorandum or financial recommendation may begin as a draft. But once an AI system can use that output to initiate an approval, disbursement, or other consequential action, the output crosses from generation into authorization.

At that boundary, independent verification is no longer merely an evaluation technique. It becomes part of the authorization architecture.

Core Citations & Primary Literature

[^1]: **Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery A., & Zhou, D. (2023).** *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. International Conference on Learning

Representations (ICLR 2023). https://arxiv.org/abs/2203.11171[10] [^2]: **Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024).** *Large Language Models Cannot Self-Correct Reasoning Yet*. International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2310.01798[11] [^3]: **Zheng, L., Chiang, W. L., Sheng, Y., Li, S., Zhuang, Z., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023).** *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. Advances in Neural Information Processing Systems (NeurIPS 2023). https://arxiv.org/abs/2360.05685[12]

References & Regulatory Citations

[1] **Institute of Internal Auditors Statement of Position on the Three Lines Model** — https://www.theiia.org/en/content/position-papers/2020/the-iia-three-lines-model-an-update-of-the-three-lines-of-defense/

[2] **Federal Reserve SR 26-2** — https://www.federalreserve.gov/supervisionreg/srletters/sr2602.htm

[3] **OCC Bulletin 2026-13** — https://www.occ.gov/news-issuances/bulletins/2026/bulletin-2026-13.html

[4] **NIST AI Risk Management Framework (AI RMF 1.0)** — https://www.nist.gov/itl/ai-risk-management-framework

[5] **LMArena** — https://lmarena.ai/

[6] **Stanford HELM** — https://crfm.stanford.edu/helm/latest/

[7] **METR** — https://metr.org/

[8] **Stanford AI Index Report** — https://aiindex.stanford.edu/report/

[9] **OECD AI Policy Observatory** — https://oecd.ai/en/dashboards/ai-principles/P7

[10] **https://arxiv.org/abs/2203.11171** — https://arxiv.org/abs/2203.11171

[11] **https://arxiv.org/abs/2310.01798** — https://arxiv.org/abs/2310.01798

[12] **https://arxiv.org/abs/2360.05685** — https://arxiv.org/abs/2360.05685