Evaluating Admissibility in Agentic Systems: Semantic Governance and a High-Consequence Benchmark

Kai Han
Semantier Technology
Beijing, China
kaihan@semantier.com

Abstract

Evaluating whether an agentic system should be allowed to execute a proposed action requires more than measuring plausibility or task accuracy. We study this as an evaluation problem of *admissibility*: whether a proposed action has sufficient governed justification, authority support, and replayable decision basis to proceed. The paper contributes a pinned-condition semantic-governance evaluation protocol, a high-consequence benchmark centered on unsafe-allow reduction and intervention behavior, and a measurement scheme for explanation reliability and replay consistency under frozen artifacts. We instantiate the benchmark in resource-event-agent (REA) fact-to-journal-entry booking because accounting workflows expose ambiguity, policy conflict, and auditability requirements in a compact stress-test domain. On two frozen benchmark slices, a governed path eliminates unsafe allows observed in direct and cautious baselines while preserving replayable decision basis and intervention-state separation. External independent human dual-rater adjudication was completed across the full 30-case public-source-derived benchmark before comparative scoring. The results should be read as controlled evidence for an evaluation-and-governance protocol on frozen artifacts and labels, not as a claim of general superiority across accounting or agentic-decision domains.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; *Machine learning approaches*; • **Information systems** → Data management systems.

Keywords

agentic systems, semantic governance, trustworthy AI, admissibility, replayable accountability, provenance

ACM Reference Format:
Kai Han. 2026. Evaluating Admissibility in Agentic Systems: Semantic Governance and a High-Consequence Benchmark. In *KDD Workshop on Evaluation and Trustworthiness of Agentic AI (KDD Eval Workshop '26), August 9–13, 2026, Jeju, Korea.* ACM, New York, NY, USA, 10 pages.

1 Introduction

Trustworthy agents can produce outputs that are fluent, plausible, and still not fit for execution. In high-stakes workflows, that distinction matters: a proposed action is not admissible merely because it looks reasonable to an LLM or even because it succeeds on a benchmark task. The core problem is deciding when an agent-generated action has a sufficiently governed decision basis to be allowed to proceed, when it should be corrected or escalated, and when it should be blocked entirely [9, 21]. This paper studies governance over agent-proposed domain actions, not low-level program execution. The execution question here is whether a semantic action proposed by an agent should be allowed to proceed at all.

This paper argues that admissibility requires semantic governance: execution should be gated by governed justification, replayable decision artifacts, and provenance that can be audited after the fact [3, 15]. The deeper governance problem is not only that plausibility or task accuracy can miss negative cases. Alignment may remain a research objective, but enterprise execution control cannot depend on directly inspecting or certifying a model's internal state. Governance therefore has to move from claims about internal alignment to checks over externally observable action basis: evidence, authority, policy compatibility, justification, replayability, and accountable intervention. We call this action-level control problem *admissibility*. This turns trustworthiness into a measurable evaluation target rather than only an architectural aspiration. We therefore frame the paper as an agent-evaluation and governance contribution first, instantiated in a high-consequence accounting stress test second. The benchmark uses resource-event-agent (REA) fact-to-journal-entry booking [10] because it is structured enough to expose classification ambiguity and intervention logic, but concrete enough to compare direct LLM projection against a governed path without making the paper only about accounting [8, 18]. The broader agentic-workflow pattern is evidence intake, projection, action proposal, and governed execution decision: accounting is the evaluation environment, not the limit of the claim. The reported 54 cases are framework-level intervention evidence under pinned benchmark, model, and policy conditions, not a statistical reliability estimate for deployment.

The benchmark is also motivated by failure modes that are hard to ignore in deployed systems: memory poisoning can corrupt downstream reasoning, architectural brittleness can make outcomes frame-dependent, and long-horizon tasks can surface negative cases that simpler task accuracy summaries miss [2, 12, 20]. Against this backdrop, the current paper combines framework definition with comparative results over a fixed twenty-four-case synthetic fixture pack spanning positive, correctable-negative, evidence-gap, policy-conflict, projection-rejection, and replay scenarios, together with a thirty-case public-source-derived benchmark

drawn from Zenodo, PCAOB, and SEC EDGAR XBRL artifacts [6, 14, 19]. The resulting contribution is intended as an evaluation protocol for governed agentic execution under pinned conditions, with metrics and artifacts that can also support regression checking across future model or policy updates. This submission validates the pinned-condition protocol itself, not the broader lifecycle-monitoring use case. Specifically, it:

- proposes an admissibility-oriented evaluation model that separates governed execution decisions from raw projection quality;
- defines a high-consequence agentic execution benchmark, instantiated in REA-to-JE, for comparing direct and governed projection;
- reports bounded comparative evidence focused on unsafe-allow reduction, intervention behavior, and replay-basis consistency rather than raw task accuracy or unverifiable internal-alignment claims alone; and
- links the benchmark design to negative-case evidence from poisoning, brittleness, and long-horizon agent settings.

2 Related Work and Positioning

2.1 Trustworthy and Reliable Agents

Trustworthy-agent surveys and benchmarks consistently emphasize threats, controls, and evaluation gaps, but they usually stop at whether an agent output is good or safe rather than whether it is admissible for execution [9, 12, 21]. That leaves an important distinction unresolved: a system can be evaluated as robust in the abstract while still lacking a governed rule for when a projected action may actually be executed. Rule-based constraint approaches such as Constitutional AI formalize harmlessness constraints over LLM outputs [1], and instruction-hierarchy work addresses authority-ordered privilege for conflicting directives [17]. These approaches share the paper's concern with constraining LLM action under policy, but neither specifies a replay-basis for decision audits or an escalation path for unresolvable cases.

2.2 Semantic Grounding

Semantic Web, knowledge-graph, and neuro-symbolic work provide machine-interpretable constraints and reasoning support, but most of it treats grounding as a representation or inference problem rather than a governance problem [5, 13]. In other words, these systems can improve the structure of what the model knows, yet they rarely specify how grounded content should trigger, block, or condition execution under policy.

2.3 Provenance, Replay, and Auditability

Provenance-oriented agent work improves traceability and workflow observability, yet it does not always specify a replay identity for decision basis or connect trace capture to intervention logic in negative cases [3, 15]. That gap matters because auditability is not only about recording steps after the fact; it also requires a stable basis for replaying the same decision and proving why a rejected action stayed rejected. The importance of structured artifact documentation is recognized more broadly in model cards [11] and datasheets [7]: those frameworks establish that reproducible evaluation requires disclosed provenance for both models and data, a precondition this paper operationalizes for governed decision artifacts.

2.4 Finance and Audit Automation

Finance LLM surveys and financial-task benchmarks show why accounting and audit settings are useful stress tests, but they typically optimize analysis or prediction quality rather than governed projection and admissibility under uncertainty [8, 16, 18]. As a result, they benchmark domain performance but do not by themselves answer whether a prediction is sufficiently governed to support a downstream accounting or audit action.

Accordingly, the comparison axes in this paper are raw projection quality, governed admissibility, replayable accountability, and intervention logic in negative cases. The positioning claim is that semantic grounding and provenance are not end goals by themselves; they are enabling mechanisms for governed execution. This is why the paper frames a high-consequence agentic execution benchmark, with REA-to-JE as one concrete instantiation: it makes the gap visible between a plausible projection and a more reviewable, replayable decision basis for action. Together, these four clusters motivate the paper's central separation between projection quality, admissibility, accountability, and intervention.

3 Semantic Governance Model

Core objects: Admitted Facts feed Governed Projections, which support Execution Intents. Justification Artifacts and Replay Bindings govern the transition from projection to admissibility. A replay binding is a pinned decision-basis bundle, such as content-addressed evidence hashes, governing rule identifiers, and authority-context identifiers, that lets the same decision basis be reconstructed later. These objects are intended to make decisions more explainable, reproducible, and aligned with authority.

Control boundary: Fact Admission restricts the decision process to verified facts. Projection Trust evaluates whether projections meet admissibility criteria. Execution Admissibility governs whether intents are allowed to execute.

Accountability claim: The model treats explanation reliability as a necessary condition for accountable execution, grounded in justification artifacts and replay bindings.

3.1 Execution Rule Skeleton

$$\begin{aligned} Execute(a_t) \text{ only if } & Valid(J_t) \text{ (Justification Validity)} \\ & \wedge\, Approved(I_t) \text{ (Authority Approval)} \\ & \wedge\, Replayable(I_t) \text{ (Replay Consistency)} \end{aligned} \tag{1}$$

- **Valid**: Justification artifacts J_t satisfy governed admissibility checks for intent I_t.

- **Approved**: Required authority threshold is satisfied for intent I_t.
- **Replayable**: Decision basis for intent I_t is pinned and reproducible.

4 Instantiation: High-Consequence Agentic Execution Benchmark

- Benchmark family: high-consequence benchmark with bounded evidence and governed admissibility checks
- Domain instantiation: source evidence to admitted fact to journal-entry booking
- Input evidence: admitted facts, source documents, and task context bundled into a fixed case packet
- Direct LLM baseline: projection from the case packet to a journal entry without governed authority checks, replay binding, or intervention
- Semantier path: governed projection that adds admissibility checks plus interception, correction, escalation, and block handling
- Benchmark emphasis: ambiguous classification, insufficient evidence, conflicting interpretations, and pressure-to-act cases
- In-scope cases preserve a fixed accounting task and evidence packet; out-of-scope cases require missing authority artifacts, new fact discovery, or open-ended generation

The scenario structure induces the outcome labels: clean in-scope cases map to ALLOW, correctable misprojections map to AWC, evidence gaps map to ESC, policy conflicts map to BLK, and admitted facts with rejected proposals map to FPRR. FPRR stands for *Fact Preserved, Projection Rejected*: the admitted fact remains part of the record, while the proposed downstream booking is rejected as inadmissible. In agentic-workflow terms, this benchmark instantiates a general control loop in which a model proposes an action from bounded evidence and the system must decide whether to allow, correct, escalate, or block that proposal before execution. The accounting instantiation is used here as a compact high-consequence stress test rather than as the primary conceptual contribution. Figure 1 summarizes the control split between a direct execution path and the governed Semantier path. Figure 2 shows how the governed path maps evidence sufficiency, policy conflict, and correction feasibility into the benchmark outcome labels, and Figure 3 summarizes the scenario classes used to stress those control decisions.

5 Evaluation Design

The evaluation measures two dimensions: outcome control and explanation quality. Outcome metrics are booking correctness, unsafe-allow rate, correction precision, and justified-block rate. Explanation quality metrics cover evidence grounding, justification grounding, decision-trace clarity, correction transparency, escalation usefulness, replay consistency, and pre-decision justification coverage. The unit

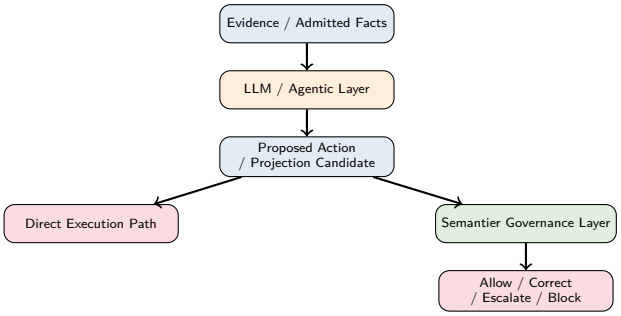

Figure 1: Governance architecture for agentic execution. An LLM or autonomous agent produces a proposed action from admitted evidence. In the direct path, that proposal proceeds without an explicit admissibility gate. In the governed path, Semantier intercepts the same proposal and applies justification, authority, and policy-aware intervention controls before execution, yielding allow, correction, escalation, or block outcomes.

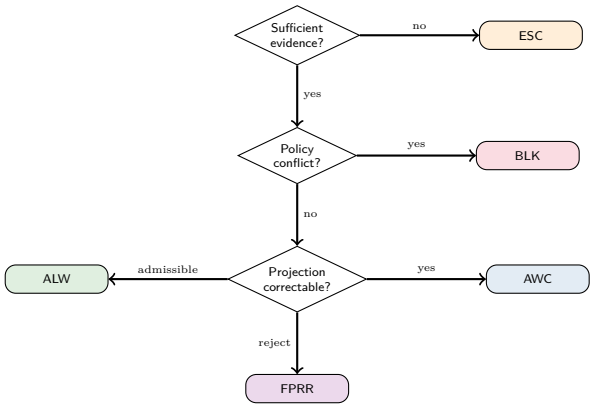

Figure 2: Decision-state flow for governed outcome labels. Evidence gaps induce ESC, policy conflicts induce BLK, recoverable misprojections induce AWC, admissible cases induce ALW, and rejected downstream proposals with preserved admitted facts induce FPRR.

of analysis is a case-level projection decision over a fixed scenario set; results are reported as macro averages with per-scenario breakdowns. Under pinned case packets and replay identifiers, the same metric surfaces are intended to support regression testing across future model, prompt, and policy updates, although the experiments reported here use a single frozen model configuration.

Each case is a self-contained packet with an admitted fact record, authority context, projection candidate, justification bundle, and replay trace hash. Gold outcome labels follow five classes. *ALLOW* (ALW) marks correct and sufficient cases. *ALLOW-WITH-CORRECTION* (AWC) marks initially wrong but recoverable cases. *ESCALATE* (ESC) marks insufficient evidence or unresolved ambiguity. *BLOCK*

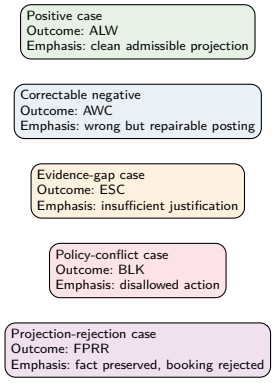

Figure 3: Scenario taxonomy for the benchmark family. Each scenario class maps a case structure to its induced governed outcome label and the control question it is meant to stress.

(BLK) marks policy conflict or disallowed action. *FACT-PRESERVED-PROJECTION-REJECTED* (FPRR) marks preserved admitted facts paired with a rejected inadmissible proposal.

Booking correctness is computed on a canonical multi-set of posting tuples {account, side, amount, currency} with line-order invariance and approved account-alias normalization. Replay consistency uses stable decision-basis identity—label plus cited artifact set plus governing policy and rule identifiers—rather than full rationale-string identity, so that the replay gate checks structural reproducibility rather than verbatim output. Pre-decision justification coverage rewards artifacts that were part of the decision basis before the outcome, not post-hoc rationale.

For the full protocol, rater agreement is measured by two independent expert raters who adjudicate disagreements on a fixed case sheet covering outcome label, evidence sufficiency, policy conflict flag, correction adequacy, explanation-grounding score, and escalation necessity; Cohen's κ is reported. Repeated runs use five trials per case per system under fixed model version, prompt-hash, temperature, top-p, and seed. The main governance claim is that semantic governance is designed to improve explanation reliability, not only booking accuracy [15, 20].

Calibration note.

Metric families, rater agreement reporting, and replay observability are aligned with agent-evaluation and provenance guidance [15, 20]. Domain correctness framing for JE-level outcomes follows financial-LLM and audit-automation literature [8, 18]. The current paper executes only a subset of this protocol on the reported benchmark slices. Thresholds in Table 2 are proposed reporting targets for larger-scale follow-on evaluation and should be interpreted with sensitivity analysis rather than as validated cutoffs.

Evaluation setup.

All runs use a single fixed LLM inference endpoint under a pinned model version, fixed prompt hash, temperature 0, and five independent seed trials per case. The pinned inference model is `openai:gpt-5.5` (version `2026-05-13`), recorded in the baseline artifact manifests under `benchmark/v2/baselines/`. Baselines are recorded under the same configuration. External independent human dual-rater review (following the protocol in Table 2) was applied to the full 30-case public-source-derived benchmark to verify labeling stability; the two raters were accounting-domain humans not involved in benchmark design (accounting students and licensed accountants), applied the same frozen rubric to the same case materials, and converged on the same outcome labels before comparative scoring was frozen. The results reported here should therefore be read as a controlled pinned-condition evaluation of a governance protocol with bounded benchmark-validation evidence, not as a claim of broad empirical superiority across accounting or agentic-decision settings.

Table 1: Benchmark comparison dimensions.

Case	Direct LLM	Semantier
Positive	ALW	ALW
Correctable negative	wrong JE	AWC
Evidence-gap case	unsupported allow	ESC
Policy-violation case	unsafe allow	BLK
Projection-rejection case	forced projection	FPRR
Replay case	unstable decision basis	stable decision-basis identity

Full benchmark-validation protocol.

In the full protocol, case construction, labeling, and evaluation are separated. Cases are frozen before comparative runs, independent raters apply a preregistered label rubric to outcome class and explanation-grounding fields, disagreements are adjudicated on fixed case sheets, and a frozen subset can be retained for separate sensitivity reporting. For this revision, these steps were executed on the full 30-case public-source-derived benchmark, and a frozen four-case subset is also reported separately as a sanity check on the same scored inventory.

5.1 Case Construction and Provenance

The twenty-four-case benchmark pack is *expert-authored and synthetic*. Each case was constructed by the authors to exercise a specific scenario class and governance label rather than being drawn directly from a pre-existing public dataset. This choice was deliberate. No ready-made public dataset currently provides the needed combination of source-evidence artifacts, gold journal-entry outcomes, and governance labels

Table 2: Proposed evaluation protocol for benchmarked semantic governance.

Metric	Level	Source artifacts	Scoring rule	Aggregation	Target reporting threshold	Applicable case types
Booking correctness	Case	Fact record, proposed JE, gold JE	1 if projected JE is equivalent to gold JE under canonical tuple-multiset rule; else 0	Cross-case mean + per-scenario mean	Semantier $\geq$ baseline +5pp	ALW, AWC
Unsafe-allow rate	Case	Projection decision, policy constraints, gold outcome	1 if system outputs ALW where gold is ESC/BLK/FPRR	Cross-case mean	Semantier $\leq$ baseline -30% relative	ESC, BLK, FPRR
Correction precision	Case	Initial proposal, corrected proposal, gold JE	1 if correction changes a wrong proposal to gold-consistent JE; else 0	Mean on correction-attempt subset	Semantier ≥ 0.75	AWC
Escalation appropriateness	Case	Evidence bundle, authority context, escalation decision	Two-rater binary score (appropriate/inappropriate), adjudicated on disagreement	Mean on escalated cases; report Cohen's κ	Mean ≥ 0.8, $\kappa \geq 0.7$	ESC
Justified-block rate	Case	Policy rules, block decision, gold outcome	1 if BLK aligns with a disallowed action under active policy	Mean on policy-conflict cases	Semantier ≥ 0.9	BLK
Fact-preservation fidelity	Case	Admitted fact hash, rejected projection, post-decision fact store	1 if fact hash is unchanged after projection rejection	Mean on rejection cases	Semantier $= 1.0$	FPRR
Explanation grounding score	Case	Justification text, cited evidence spans, trace links	Two-rater 1–5 rubric for evidence traceability and pre-decision justification coverage	Mean score + per-rater variance	Semantier ≥ 4.0 and lower variance than baseline	All labels
Replay consistency	Case-trial	Replay trace hash, rerun outputs across 5 trials	1 if label + cited artifact set + policy/rule IDs remain stable under pinned config; else 0	Mean across trials and cases	Semantier ≥ 0.9	All labels

of the form ALLOW / AWC / ESC / BLK / FPRR. Financial-LLM benchmarks such as FinQA [4] target QA over financial reports, not REA-to-JE booking under policy. EDINET-Bench [16] offers realistic annual-report evidence but does not supply booking gold or governance labels. PCAOB inspection datasets expose real audit-deficiency patterns but are structured around audit findings rather than fact-to-entry projection tasks [14]. SEC EDGAR XBRL financial-statements data provides authentic financial structure but likewise does not carry pre-labeled journal-entry outcomes or admissibility labels [19].

Label assignment.

Gold outcome labels were assigned by the authors following the formal label definitions in Section 4: positive cases receive ALLOW, correctable cases receive AWC, evidence-gap cases receive ESC, policy-conflict cases receive BLK, and cases where an admitted fact must be preserved while the proposed booking is rejected receive FPRR. Each case specifies the admitted REA fact packet, the expected governance outcome, and the controlled perturbation that makes the case non-trivial. Labels are deterministic with respect to the stated evidence and the governing policy set; no LLM inference was used in label assignment.

Benchmark credibility limitation.

A fully synthetic, self-authored pack of this scale can be read as encoding the desired answer in the benchmark construction. The current results should therefore be interpreted as controlled benchmark evidence that the governed path produces outcome-differentiated behavior on this fixture, not as general empirical superiority. To improve benchmark credibility, the authors derived a public-source-derived benchmark of thirty cases from three source families: (1) the Zenodo Invoice/Receipt Dataset [6], from which invoice and receipt annotations were mapped verbatim into REA fact packets; (2) the PCAOB Firm Inspection Reports Part I.A structured dataset [14], from which evidence-gap and policy-conflict cases were generalized from real audit deficiency findings; and (3) SEC EDGAR XBRL financial-statement artifacts [19], from which positive, correctable, evidence-gap, and policy-conflict packets were derived from filing disclosures.

Expert annotation provides the JE gold outcomes and governance labels for all thirty cases; no LLM inference was used in label assignment. That design retains author-controlled label quality while providing externally verifiable source artifacts that reviewers can inspect independently. The thirty-case public-source-derived benchmark is machine-validated and executable through the same governed benchmark runner used for the synthetic pack, with case-level provenance preserved as pinned artifact references and annotation metadata. Pinned direct, cautious, and governed comparative artifacts are now recorded for

all thirty cases under fixed model, prompt-hash, and seed settings. The scored public-source-derived benchmark remains modest in size and partially abstracted, especially for PCAOB-derived cases that encode real deficiency patterns rather than full transaction ledgers, so it should be read as bounded external-evidence support rather than as a standalone large-scale benchmark.

5.2 Label Validity and Benchmark Construction Limits

The benchmark labels are author-assigned at construction time. A rubric-based external human dual-rater adjudication pass was executed on the full thirty-case public-source-derived benchmark, with two independent raters applying the same fixed rubric to the same case materials. Both raters converged on identical outcome labels (30/30 agreement; Cohen's $\kappa = 1.00$), and the adjudicated deterministic labels are used as the final evaluation labels. Secondary-field agreement was lower and therefore more diagnostically informative, including 76.7% agreement on correction adequacy and explanation-grounding score. Together, these results provide external inter-rater consistency evidence for the scored thirty-case benchmark while still showing that finer-grained qualitative judgments remain less trivial than coarse outcome labels.

The benchmark was also designed to expose governance-specific distinctions such as evidence gaps, policy conflicts, and admissibility rejections. That design choice is useful for framework validation because it makes unsafe-allow behavior observable, but it also creates a circular-benchmark risk: a benchmark authored around governance distinctions can structurally favor a governed architecture. The public-source-derived benchmark reduces this risk only partially. In that benchmark, source fields and deficiency patterns come from public artifacts, but the case packets, journal-entry gold outcomes, and governance labels remain author-annotated. The current paper therefore provides stronger evidence for framework-level outcome differentiation under pinned conditions than for benchmark-level external validity.

5.3 Controlled Evaluation Results

Scope note: the following results constitute a governance-framework demonstration on a small, controlled benchmark. They include external dual-rater agreement and frozen deterministic labels for the thirty-case public-source-derived comparative benchmark.

The current implementation includes a fixed twenty-four-case benchmark pack spanning six scenario families, with four cases per family: positive, correctable-negative, evidence-gap, policy-conflict, projection-rejection, and replay. Direct projection is represented by two pinned baseline families recorded under fixed model, prompt-hash, and seed settings: a permissive direct baseline and a more cautious direct baseline that already defers on missing-evidence cases. The governed path reuses Semantier projection-context and replay-binding surfaces over the same case packets. Combined with the component attribution summary in Section 6.3, the comparison is intended to validate the governed framework at the system level: whether a pinned governed pipeline produces different intervention behavior from pinned direct baselines on the same cases. It is not, by itself, a full validation of benchmark external validity. Table 3 therefore foregrounds unsafe-allow counts and intervention separation, especially in the negative-case families where execution control matters most. Because the pack is expert-authored and synthetic, these numbers should be read as controlled benchmark evidence for the benchmark pipeline and label logic, not as a broad estimate of real-world performance. On this fixed benchmark, the cautious baseline performs better than the permissive baseline on evidence-gap cases, but not on policy-conflict or projection-rejection cases, where the governed path yields escalation, block, or rejection outcomes instead of unsafe allow.

Framework validation versus benchmark validation. In this paper, *framework validation* means showing that the governed path produces different, more restrictive intervention behavior than direct baselines under pinned conditions. *Benchmark validation* would additionally require independently adjudicated labels, stronger baseline diversity, held-out evaluation, and broader external evidence. The current results primarily address the former and only partially address the latter.

Executable public-source-derived slice.

In addition to the synthetic pack, the current implementation executes a thirty-case public-source-derived benchmark through the same governed runner and scores it against two pinned baseline families: a permissive direct baseline and a more cautious baseline that already defers on evidence-gap cases. This benchmark contains eight positive cases, eight correctable-negative cases, seven evidence-gap cases, and seven policy-conflict cases derived from pinned Zenodo, PCAOB, and SEC EDGAR XBRL artifacts. The resulting comparison remains modest in scale, but it provides a public-source-derived direct-versus-governed measurement rather than governed-only harness validation.

5.4 Public-Source-Derived Comparative Results

Table 5 reports comparative results on the thirty public-source-derived cases. The same two baseline families used in the synthetic pack are pinned for this public-source-derived benchmark: a permissive direct baseline and a cautious baseline that already defers on evidence-gap cases. Across the full thirty-case benchmark, unsafe-allow counts are 14/30 for the direct baseline, 7/30 for the cautious baseline, and 0/30 for the governed path; corresponding booking-correctness counts are 8/30, 15/30, and 30/30. The scenario breakdown mirrors the synthetic benchmark pattern: caution helps on evidence-gap cases, but it does not resolve correctable misclassification or policy-conflict cases, where the governed path still supplies the clearest intervention separation. All thirty governed runs completed

Table 3: Results on the 24-case synthetic benchmark slice. Counts are shown with rates to emphasize unsafe-allow reduction and intervention behavior on a fixed author-labeled pack.

Metric	Direct	Cautious	Governed
Overall booking correctness	8/24 (0.33)	12/24 (0.50)	24/24 (1.00)
Overall unsafe-allow rate	12/24 (0.50)	8/24 (0.33)	0/24 (0.00)
Evidence-gap unsafe allow	4/4 (1.00)	0/4 (0.00)	0/4 (0.00)
Policy-conflict unsafe allow	4/4 (1.00)	4/4 (1.00)	0/4 (0.00)
Projection-rejection unsafe allow	4/4 (1.00)	4/4 (1.00)	0/4 (0.00)

Table 4: Per-scenario results on the synthetic benchmark pack (booking-correctness count / unsafe-allow count).

Scenario	N	Dir.	Caut.	Gov.
Positive	4	4/4 ; 0/4	4/4 ; 0/4	4/4 ; 0/4
Correctable	4	0/4 ; 0/4	0/4 ; 0/4	4/4 ; 0/4
Evidence gap	4	0/4 ; 4/4	4/4 ; 0/4	4/4 ; 0/4
Policy conflict	4	0/4 ; 4/4	0/4 ; 4/4	4/4 ; 0/4
Proj. rejection	4	0/4 ; 4/4	0/4 ; 4/4	4/4 ; 0/4
Replay	4	4/4 ; 0/4	4/4 ; 0/4	4/4 ; 0/4

with `replay_binding_status = OK` and unique projection context references. In this benchmark, the public artifacts contribute the observed invoice fields, filing disclosures, and PCAOB deficiency patterns, while the REA case packets, JE gold labels, and governance labels remain author-annotated. The public-source-derived comparison therefore increases external inspectability and now includes independent external adjudication for the scored thirty-case benchmark.

External human dual-rater adjudication. External human dual-rater adjudication was applied to all 30 scored cases using a frozen rubric. The two raters were accounting students and licensed accountants who were independent of benchmark construction. Outcome-label agreement was 30/30 (Cohen's $\kappa = 1.00$). Secondary-field agreement was lower, including 90.0% for evidence sufficiency, 100.0% for policy-conflict presence, and 76.7% for both correction adequacy and explanation-grounding score. This result provides independent inter-rater agreement evidence for the scored public-source-derived benchmark while also showing that finer-grained review fields remain more judgment-sensitive than coarse outcome labels.

A frozen 4-case subset (`v2-heldout-2026-05-14`) is also reported separately. On that subset, booking correctness is 1/4 (direct), 2/4 (cautious), and 4/4 (governed), while unsafe allows are 2/4, 1/4, and 0/4. Because these cases are part of the scored 30-case inventory, we report them as a traceable sensitivity slice rather than as out-of-sample evidence.

Table 5: Comparative results on the 30 public-source-derived cases (booking-correctness count / unsafe-allow count) for pinned direct, cautious, and governed runs.

Scenario	N	Dir.	Caut.	Gov.
Positive	8	8/8 ; 0/8	8/8 ; 0/8	8/8 ; 0/8
Correctable	8	0/8 ; 0/8	0/8 ; 0/8	8/8 ; 0/8
Evidence gap	7	0/7 ; 7/7	7/7 ; 0/7	7/7 ; 0/7
Policy conflict	7	0/7 ; 7/7	0/7 ; 7/7	7/7 ; 0/7
Total	**30**	**8/30 ; 14/30**	**15/30 ; 7/30**	**30/30 ; 0/30**

6 Failure Modes and Discussion

6.1 Failure-to-Control Mapping

Four recurrent failure classes motivate governance controls: (1) plausible-but-underjustified bookings that lack required evidence lineage, (2) numerically balanced but semantically wrong bookings, (3) unstable context that breaks replay consistency, and (4) long-horizon drift that surfaces delayed inconsistencies. In each case, direct baselines tend to allow execution, while the governed path applies evidence gates, policy checks, and replay binding to escalate, correct, block, or reject as needed. Residual limitations remain in ambiguous evidence and dynamically shifting policy context [2, 12, 16, 18, 20].

6.2 Limits of the Approach

The current 54-case benchmark should be read as controlled framework evidence rather than statistical generalization. The public-source-derived benchmark improves artifact realism, but the case packets, journal-entry gold outcomes, and governance labels remain author-annotated, so some circularity risk remains even after external adjudication. The direct and cautious baselines are also intentionally narrow comparator families rather than an exhaustive baseline suite, which means the current comparison is strongest as a controlled intervention study and weaker as a claim about frontier-model performance in general. Governed intervention also introduces review and operational overhead that may be unacceptable in some environments. In addition, replayable accountability does not eliminate the need for human policy design; governance quality still depends on policy quality. Although the protocol is designed to support lifecycle regression monitoring across model, prompt, or policy changes, that use mode is not empirically validated in this submission; the evidence here is limited to frozen benchmark slices under pinned conditions. Future deployment work should expand the benchmark with real workflow

Table 6: Component attribution: governance gates, source modules, intercepted scenario type, and target unsafe-allow class.

G	Gate function	Intercepts	Target class
1	Structural REA validation (conservation, resources)	Malformed projections	AWC
2	Tiered evidence artifact gate (missing governance artifacts)	Evidence gap	ESC
3	Policy & resolution gate (precedence graph, active policy bundle)	Policy conflict, inadmissible projection	BLK, FPRR
4	Replay binding verification (artifact version pins)	Decision-basis drift	Replay

data collected only under explicit customer consent and governance controls. Overall, the failure modes show why plausibility alone is insufficient, while the proposed controls are intended to improve execution discipline within clear limits of scope, cost, and universality.

The transferable unit is the admissibility gate, not the accounting chart of accounts: other regulated workflows would require different authority models and stronger human adjudication, so the REA results should not be treated as cross-domain proof. A narrower next domain is commercial contract review and approval, where clauses, obligations, counterparty terms, approval thresholds, and exception policies can be structured through data-engineering pipelines and GraphRAG-style construction. The implementation is model-facing rather than model-internal, but the reported experiment uses one pinned model family to isolate the governance layer; cross-model tests, including smaller open-weight models, and comparisons with guardrail or workflow-control frameworks remain necessary before making model-agnostic robustness claims.

6.3 Component Attribution and Ablation Scope

The governed pipeline is composed of four independently implemented gates, each grounded in a distinct runtime module. Gate 1 (validation service: structural REA checks) enforces conservation, minimum resource count, and currency consistency. Gate 2 (validation service: tiered evidence artifact gate) rejects projection when required governance artifacts are absent, which maps to ESC outcomes. Gate 3 (resolution and policy-store services) applies a precedence graph and resolution rules against the active policy bundle, which maps to BLK and FPRR outcomes. Gate 4 (replay-binding store) pins artifact versions and verifies structural reproducibility across runs, which maps to replay-case stability.

Table 6 maps each gate to its target scenario type. The mapping is deterministic with respect to the label definitions in Section 4: evidence-gap cases fail at Gate 2 because required tiered artifacts are absent; policy-conflict and projection-rejection cases fail at Gate 3 because the active policy bundle disallows the proposed action; replay cases are verified at Gate 4. A hold-out ablation study—disabling gates in sequence and measuring the resulting change in unsafe-allow rate per scenario type—is architecturally feasible from this implementation and is a planned next step. The current results attribute outcome differentiation to the governed path as a whole; per-gate contribution remains an open empirical task.

6.4 Operational Cost and Deployment Trade-offs

The architecture trades throughput for control. Justification-gated execution can reduce unsafe allows, but it also adds latency through evidence checks, policy resolution, correction attempts, and escalation handling. The practical deployment target is not machine-speed execution alone, but improvement over existing human-gated workplace approval workflows while preserving auditability. In deployment terms, the main costs are fourfold. First, more cases are routed to human review, especially when evidence is incomplete or policy prerequisites are missing. Second, replayable accountability requires storage of pinned decision-basis artifacts, including evidence references, policy identifiers, and replay bindings. Third, governance can reduce raw automation yield because some superficially plausible actions are intentionally not allowed to execute. Fourth, policy maintenance becomes an ongoing operational responsibility rather than a one-time prompt-engineering task. These costs are acceptable only where execution mistakes are expensive enough that unsafe-allow reduction, auditability, and intervention discipline matter more than maximum straight-through processing. The current paper does not report deployment-grade latency or compute measurements; future runs should report per-gate latency, replay-artifact storage, human-review queue cost, and end-to-end comparison against human-gated baselines.

6.5 Validation Roadmap

The next validation steps are operational rather than conceptual. External dual-rater adjudication and pinned comparative scoring have now been completed for the thirty-case public-source-derived benchmark, and the frozen four-case subset remains available for separate sensitivity reporting. The main next priorities are expanding the benchmark beyond the current 54 cases with consented real workflow data, strengthening baseline sets with additional proprietary and open-weight models, releasing adjudication sheets and frozen manifests, preserving benchmark-construction logs, adding broader third-party benchmark coverage, running explicit gate ablations, measuring deployment overhead, and conducting version-to-version regression studies. A scoped domain-transfer track will begin with commercial contract review and approval workflows rather than broad legal reasoning. These steps are intended to let reviewers better audit the boundary between public artifacts and author annotations and the sensitivity of admissibility decisions to model or policy evolution.

7 Conclusion

This paper addresses the challenge of semantic governance in agentic execution by separating plausibility from admissibility. By defining core objects, control boundaries, and execution rules, it presents a governance architecture for deciding when agent-proposed actions are sufficiently justified, aligned, and replayable to execute.

The benchmark scenarios make visible how semantic governance can intercept failure modes, escalate unresolved cases, and require replayable accountability. The REA instantiation is useful here not because accounting is the only target domain, but because it exposes ambiguity, policy conflict, and audit requirements in a compact stress test. On the reported slices, the governed path eliminates unsafe allows observed in the direct baselines under the pinned configurations used here, especially in evidence-gap, policy-conflict, and projection-rejection settings where execution control matters most. More generally, the paper reframes agent evaluation from a pure generation question into a governed decision question with explicit intervention states and replayable justification. In that sense, Semantier governs the execution of agent-proposed semantic actions rather than the low-level execution of ordinary program instructions.

The current results support framework-level feasibility and bounded outcome differentiation on the reported benchmark slices, not definitive benchmark validation or universal superiority. The strongest claim supported here is that, on pinned benchmark slices with frozen artifacts and labels, governed intervention can produce materially different execution decisions from direct projection baselines. Strengthening baseline coverage beyond the current direct and cautious families, quantifying governance overhead, and extending evaluation to broader third-party evidence sources remain the most important next steps.

Data Availability

The benchmark definitions, case packets, pinned baseline artifacts, source manifests, adjudication materials, and regression tests are available in the Semantier runtime repository under the paper and benchmark artifacts prepared for this submission. A persistent public archive will be released after the camera-ready process.

References

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL] https://arxiv.org/abs/2212.08073

[2] Varun Pratap Bhardwaj. 2026. SuperLocalMemory: Privacy-Preserving Multi-Agent Memory with Bayesian Trust Defense Against Memory Poisoning. arXiv:2603.02240 [cs.AI] https://arxiv.org/abs/2603.02240

[3] Ching-Chun Chang and Isao Echizen. 2026. Chronology of Multi-Agent Interactions for Provenance of Evolving Information. arXiv:2504.12612 [cs.AI] doi:10.1098/rsos.251988

[4] Zhiyu Chen, Wenhu Chen, Charese Shen, Oliver Borger, Daniel Bhatt, Jun Gao, Elish Haque, Tong Wang, and William Wang. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. arXiv:2109.00122 [cs.CL] https://arxiv.org/abs/2109.00122

[5] Brandon C. Colelough and William Regli. 2025. Neuro-Symbolic AI in 2024: A Systematic Review. arXiv:2501.05435 [cs.AI] https://arxiv.org/abs/2501.05435

[6] Francisco Cruz and Mauro Castelli. 2022. Dataset of Invoices and Receipts Including Annotation of Relevant Fields. https://zenodo.org/records/6371710. doi:10.5281/zenodo.6371710 Public invoice and receipt image dataset available for document understanding research.

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. doi:10.1145/3458723

[8] Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. 2024. A Survey of Large Language Models in Finance (FinLLMs). arXiv:2402.02315 [cs.CL] doi:10.1007/s00521-024-10495-6

[9] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, Rongcheng Tu, Xiao Luo, Wei Ju, Zhiping Xiao, Yifan Wang, Meng Xiao, Chenwu Liu, Jingyang Yuan, Shichang Zhang, Yiqiao Jin, Fan Zhang, Xian Wu, Hanqing Zhao, Dacheng Tao, Philip S. Yu, and Ming Zhang. 2025. Large Language Model Agent: A Survey on Methodology, Applications and Challenges. arXiv:2503.21460 [cs.CL] https://arxiv.org/abs/2503.21460

[10] William E. McCarthy. 1982. The REA Accounting Model: A Generalized Framework for Accounting Systems in a Shared Data Environment. *The Accounting Review* 57, 3 (1982), 554–578. doi:10.2308/TAR-4487748

[11] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229. doi:10.1145/3287560.3287596

[12] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and Benchmarking of LLM Agents: A Survey. arXiv:2507.21504 [cs.LG] doi:10.1145/3711896.3736570

[13] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. arXiv:2308.06374 [cs.AI] https://arxiv.org/abs/2308.06374

[14] Public Company Accounting Oversight Board. 2024. PCAOB Firm Inspection Reports and Datasets. https://pcaobus.org/oversight/inspections/firm-inspection-reports. Structured deficiency data available for academic research.

[15] Renan Souza, Amal Gueroudji, Stephen DeWitt, Daniel Rosendo, Tirthankar Ghosal, Robert Ross, Prasanna Balaprakash, and Rafael Ferreira da Silva. 2025. PROV-AGENT: Unified Provenance for Tracking AI Agent Interactions in Agentic Workflows. arXiv:2508.02866 [cs.DC] https://arxiv.org/abs/2508.02866

[16] Issa Sugiura, Takashi Ishida, Taro Makino, Chieko Tazuke, Takanori Nakagawa, Kosuke Nakago, and David Ha. 2026. EDINET-Bench: Evaluating LLMs on Complex Financial Tasks using Japanese Financial Statements. arXiv:2506.08762 [q-fin.ST] https://arxiv.org/abs/2506.08762

[17] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. arXiv:2404.13208 [cs.CR] https://arxiv.org/abs/2404.13208

[18] Rushi Wang, Jiateng Liu, Weijie Zhao, Shenglan Li, and Denghui Zhang. 2025. Automating Financial Statement Audits with Large Language Models. arXiv:2506.17282 [cs.IR] https://arxiv.org/

abs/2506.17282
[19] XBRL US Academic Repository. 2024. SEC EDGAR XBRL Financial Statements Data. https://xbrl.us/academic-repository/sec-edgar-data/. Quarterly financial statement datasets derived from SEC EDGAR XBRL filings.
[20] Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024. MIRAI: Evaluating LLM Agents for Event Forecasting. arXiv:2407.01231 [cs.CL] https://arxiv.org/abs/2407.01231
[21] Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pang, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, Bo An, and Qingsong Wen. 2025. A Survey on Trustworthy LLM Agents: Threats and Countermeasures. arXiv:2503.09648 [cs.MA] https://arxiv.org/abs/2503.09648