INFORMATION FLOW IS NOT EXECUTION AUTHORITY: A CONTROLLED EVALUATION OF LLM MULTI-AGENT SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety outcomes in action-taking LLM multi-agent systems can conflate information-flow effects on proposal formation, execution coverage, execution selection, candidate availability, and authorization-contract validity. We introduce controlled layer substitution, an evaluation framework that fixes the relevant upstream decision objects and varies these components separately. To operationalize the comparisons, we construct GateSim-T, a synthetic testbed of 192 tasks across seven model checkpoints and five information-flow configurations, yielding 6,720 paired decisions and 40,320 model calls. DecisionGate instantiates one authorization policy: a fixed scoped rule backed by a statistical certificate computed from structured evidence. In a post-hoc comparison on the same stored proposals, it and random and confidence-based abstention exactly match execution counts within every model–flow stratum. At 12.26% overall coverage, adverse deployment on risk-eligible tasks is 0.12% with DecisionGate, versus 8.57% with random and 7.80% with confidence abstention. The difference is not attributable to abstention frequency alone: DecisionGate's certificate-based policy uses quantities unavailable to the confidence policy, so this contrasts complete policies under specified information conditions, not authorizers under equal information. A direct-rule control shows that structured evidence can support effective deployment decisions without MAS proposals in this testbed. External Criteo/AgentDojo tests show a different coverage–regret trade-off and no replication of the synthetic matched-coverage risk separation. Candidate-recovery and contract-shift interventions further expose dependencies on available actions and valid authorization inputs. Together, the framework specifies what is held fixed, what each policy can observe, and which effects the resulting comparisons can identify.

1 INTRODUCTION

LLM-based multi-agent systems increasingly propose consequential actions, including tool calls and privileged operations (Debenedetti et al., 2024; Zhang et al., 2025b; Patil et al., 2025; Wu et al., 2026b; Liu et al., 2026). Recent benchmarks increasingly evaluate LLM agents in interactive, multi-turn, and tool-mediated environments, emphasizing not only task completion but also intermediate decision making, tool use, reliability, and policy adherence (Liu et al., 2024; Qin et al., 2024; Ma et al., 2024; Yao et al., 2025; Lu et al., 2025; Yang et al., 2026). How should we evaluate a safety improvement in such a system? It may come from forming better proposals, executing fewer of them, or selecting more effectively among the same proposals. An end-to-end score can confound these explanations. Related work on selective prediction, abstention, and risk control similarly highlights the need to reason about risk jointly with coverage or utility rather than through aggregate performance alone (Angelopoulos et al., 2024; Hedström et al., 2025; Wang et al., 2026). We ask how to distinguish these sources experimentally, and how the answer depends on the available candidates and authorization contract.

The distinction between information and decision rights provides a useful organizing principle (Jensen & Meckling, 1995; Aghion & Tirole, 1997). In action-taking LLM systems, communication shapes what is proposed, whereas authorization determines whether the proposal may be executed. Related system designs increasingly separate trusted instructions or control mechanisms from untrusted

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

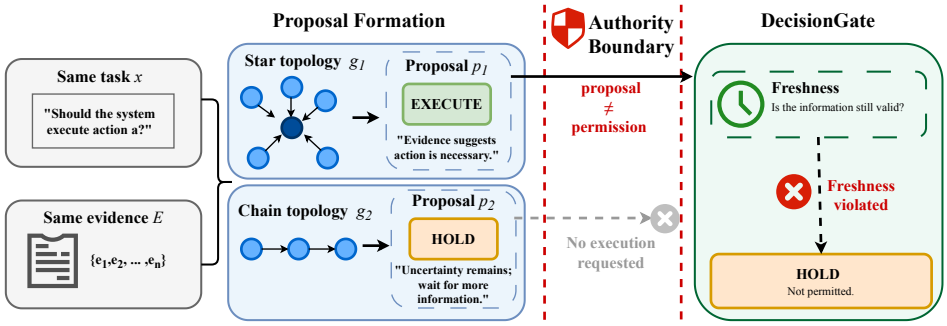


Figure 1: **Motivation: information flow and execution authority.** The same task and evidence can yield different proposals under different information flows. In this conceptual example, an EXECUTE proposal is withheld because the freshness condition is violated; a HOLD proposal requests no execution.

information and place explicit checks around consequential behavior (Chen et al., 2025; Debenedetti et al., 2025). We call these functions *informational influence* and *execution authority*.

Debate, centralized coordination, chains, hierarchies, and adaptive communication structures alter message visibility, interaction patterns, and aggregation (Du et al., 2024; Zhu et al., 2025; Wang et al., 2025a; Leong et al., 2025; Shen et al., 2025; Zhang et al., 2026). More broadly, recent agentic systems organize collaboration through role-structured, workflow-based, or graph-structured designs, with system composition, connectivity, and communication patterns increasingly treated as explicit design variables (Hong et al., 2024; Zhuge et al., 2024; Hu et al., 2025; Li et al., 2024). Communication structure is also increasingly analyzed as an explicit system property that can be optimized, sparsified, or inferred from multi-agent interactions (Zhuge et al., 2024; Li et al., 2024; Wu et al., 2026a). Recent studies further show that collaboration strategies and communication protocols can substantially affect multi-agent coordination and reasoning behavior (Zhu et al., 2025; Wang et al., 2025a; Zhang et al., 2026). These information-flow effects are also safety relevant: topology can shape error propagation and vulnerability to adversarial or erroneous messages (Yu et al., 2025; Wang et al., 2025b). Downstream checks can then change which proposals execute. Evaluating these upstream and downstream interventions through final safety scores alone leaves their respective contributions unresolved. Figure 1 illustrates the distinction using the same task and evidence.

Our framework, *controlled layer substitution*, fixes the relevant upstream decision objects and varies one system function at a time. A proposal plane forms an action from evidence and an information-flow architecture. A candidate-availability interface determines which executable options the system can represent, a distinction that becomes important when missing tools, incomplete information, or underspecified requests constrain agent behavior (Kirmayr et al., 2026). An authorization layer evaluates its output under an explicit contract; DecisionGate is one fixed-rule instance. Holding proposals and within-stratum execution counts fixed permits a policy-level comparison: adverse-deployment differences reflect execution selection under the specified policy-specific information conditions, rather than abstention frequency alone. They do not isolate the rule from its certificate access: DecisionGate uses structured quantities unavailable to the public-card critic or confidence policy. A separate intervention varies contract validity. Appendix Table 3 summarizes the controls.

GateSim-T instantiates this framework with 192 synthetic tasks across eight decision panels, seven checkpoints, and five information flows, yielding 6,720 paired decisions and 40,320 calls. At 12.26% matched coverage, DecisionGate has 0.12% adverse deployment on risk-eligible tasks, compared with 8.57% for random and 7.80% for confidence abstention. This post-hoc policy contrast retains its direction across the evaluated families and risk panels; it is not an equal-information algorithm comparison. The preregistered topology-regret tests address a separate question: they do not establish the specified effects above five percentage points. They neither establish equivalence nor rule out transportable topology selection.

Prospectively frozen Criteo and AgentDojo studies test the framework under different action semantics. They show a different coverage–regret trade-off without reproducing the synthetic matched-risk separation. In sequential AgentDojo interventions, next-call clarification leaves execution at zero, whereas an explicit candidate interface restores useful action under a contract containing trusted

valid-candidate identity. A direct-rule analysis additionally shows that GateSim-T's structured evidence can support decisions without MAS proposals. These conditions specify what each comparison identifies.

Our contributions are:

- **Evaluation framework.** Controlled layer substitution separates proposal formation, execution selection, candidate availability, and contract validity under documented information access.
- **GateSim-T testbed.** A frozen 192-task, seven-checkpoint, five-flow matrix yields 6,720 paired decisions and 40,320 calls for component-level evaluation.
- **Boundary-aware evidence.** Matched-coverage, external, candidate, and contract tests expose policy-level differences and the limits of their interpretation.

2 Related Work

Information-flow and coordination architecture. Multi-agent debate and structured communication can improve reasoning by exposing disagreements and iterating on candidate answers (Du et al., 2024). Recent work learns, selects, or sparsifies communication structures for performance and efficiency. AMAS chooses task-specific graphs (Leong et al., 2025), while AgentPrune prunes spatial-temporal message-passing graphs to reduce redundant communication while preserving task performance (Zhang et al., 2025a). Causal analysis of information propagation further identifies a trade-off between suppressing errors and preserving useful information (Shen et al., 2025). Our focus differs: we test whether topology rankings and selections transport under a fixed call contract when the downstream object is an action with asymmetric harm, not only an answer score.

Multi-agent safety and error propagation. NetSafe studies topological vulnerability to attacks and identifies error-propagation phenomena (Yu et al., 2025); G-Safeguard detects anomalies on utterance graphs and intervenes on topology (Wang et al., 2025b). These works establish that topology is safety relevant. We compare these proposal-level effects with a separate execution policy on fixed proposals. Context-dependent flow effects coexist with a more stable GateSim-T authorization layer; matched-coverage controls assess how much of that difference remains at the same execution rate.

Runtime guardrails and selective action. Agent guardrails adapt checks to task and system risks (Luo et al., 2025); SupervisorAgent targets efficient runtime intervention without changing the base architecture (Lin et al., 2026). CORA calibrates pre-action execution risk and routes rejected actions to interventions (Feng et al., 2026); Conformal Selective Acting targets anytime, per-deployment selective-risk control (Khosravi & Huo, 2026); and AgentAbstain evaluates whether agents know when not to act (Liu et al., 2026). Related work on epistemically responsible response behavior similarly uses uncertainty-aware strategy selection to decide whether a model should answer directly, respond cautiously, or abstain (Ma et al., 2026). AttriGuard moves the check to the action level, using causal attribution to determine whether a proposed tool call is supported by user intent rather than untrusted observations (He et al., 2026). We extend this line of work by making proposal formation and execution selection separately measurable: the same frozen proposals are compared at exact action coverage and subsequently evaluated under explicit contract-schema shifts.

Separating reasoning from privileged execution. Recent systems explicitly interpose deterministic controls between neural proposals and consequential actions. SAGA makes authorization and delegation explicit governance primitives for agentic systems Syros et al. (2025). Neural Cryptographic Services binds tool dispatch to signed, hash-chained instructions and evaluates injection resistance on AgentDojo and a custom argument-hijacking benchmark (Weng et al., 2026); a defense-in-depth identity architecture similarly separates an LLM intent layer from cryptographic execution (Chahine, 2027). In industrial control, LLM monitors propose while deterministic verification and execution layers retain plant authority (Schall, 2026). Runtime topology mutation has also been guarded by capability, state-routing, and shadow-validation invariants (Sidik et al., 2026). We turn this architectural principle into separately testable substitution hypotheses across matched coverage, named-domain transport, candidate availability, and contract shift.

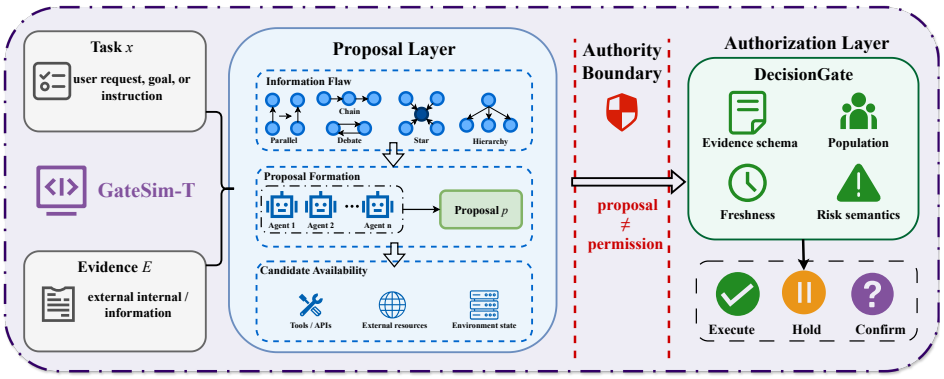


Figure 2: System overview of the controlled evaluation framework in this paper.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

Let $e = (f, d)$ denote an environment consisting of model family f and decision panel d. An *information-flow architecture* specifies message visibility, disclosure schedule, and aggregation. Each $g = (V_g, S_g, A_g) \in \mathcal{G}$ is one controlled *flow realization*; it is not an authorization object. Each task i provides frozen context and evidence $x_i = (E_i, D_i)$, an oracle action a_i^*, and an asymmetric regret function $L(a, a_i^*)$. The action set is

$$\mathcal{A} = \{\text{SHIP-ALL}, \text{SHIP-TARGETED}, \text{HOLD}, \text{REQUEST-CONFIRMATION}\}. \tag{1}$$

The proposal plane produces $p_{i,g} = F_g(x_i)$, containing a candidate action, confidence, and cited evidence, from a fixed-budget multi-agent transcript. An authorization contract $C = (\mathcal{S}, \mathcal{P}, \mathcal{R}, \tau)$ is the declared specification of evidence schema, target population, risk semantics, and freshness. The authorization layer applies a fixed authorizer to $(x_i, p_{i,g}, C)$ and returns an executable action only when permitted; otherwise it holds, rejects, or requests confirmation. DecisionGate is that fixed authorizer. In GateSim-T, the *certificate action* is its precomputed output under the declared contract, not a synonym for the contract itself.

We distinguish four quantities. First, the pooled adverse-deployment rate $R(g)$ describes a controlled flow realization averaged over environments. Second, $R_e(g)$ exposes flow-by-environment interaction. Third, architecture-selection regret measures the excess loss from choosing a realization in a source environment and applying it to held-out environments. Fourth, authorization risk measures the executed action after the authority boundary. An information-flow realization is not a safety certificate merely because it minimizes pooled $R(g)$; authorization must be evaluated under its declared contract and transport conditions.

Controlled layer substitution is the evaluation protocol; DecisionGate is an authorization instance, and GateSim-T is its synthetic identification environment. Figure 2 connects the interfaces. Four analysts exchange messages according to g, and an orchestrator aggregates their returned memos and cited public evidence into $p_{i,g}$. We store this proposal before substituting execution policies. On the stored records, each policy produces a final action; exact-coverage controls then match the number of executions, and task-paired evaluation compares risk and utility. Changing information flow or the candidate interface can change proposal formation. In contrast, the authorization and contract-shift comparisons retain the corresponding stored proposals.

3.2 DECISIONGATE: CONTRACT-BASED AUTHORIZATION

Contract, statistical certificate, and action. The contract C specifies what evidence and deployment conditions are relevant. GateSim-T generates a statistical certificate from full evidence cards, then converts it into a certificate action c. These are distinct from the runtime gate on a proposed action. For each group $u \in \{\text{all}, \text{target}\}$, let b_u be the minimum adjusted benefit p-value across that group's primary cards. The adjustment uses the stored randomization maximum statistics: for

primary statistic z_j and B permutations,

$$\widetilde{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{M_b \geq z_j\}}{B+1}, \qquad b_u = \min_{j:\,\mathrm{group}(j)=u} \widetilde{p}_j. \tag{2}$$

Here M_b is the maximum primary statistic across the registered benefit paths in permutation b. Guardrail lower bounds are estimate minus standard error times the empirical 0.95 quantile of the permutation maximum absolute guardrail statistic, using the higher-order quantile convention. The complement-risk and transport lower bounds instead use the stored critical value 2.13. Appendix Table 7 gives the core decision rule with frozen $\alpha = 0.05$ and margin $m = 0.20$.

Runtime mapping. Let p be the orchestrator action, c the serialized certificate action, and $\mathcal{S}_{\mathrm{ship}} = \{\textsc{Ship-All}, \textsc{Ship-Targeted}\}$. The actual GateSim-T mapping is

$$a_{\mathrm{DG}}(p, c) = \begin{cases} \textsc{Hold}, & p \in \mathcal{S}_{\mathrm{ship}} \text{ and } p \neq c, \\ p, & \text{otherwise.} \end{cases} \tag{3}$$

It preserves non-ship proposals and matching ship actions, and rejects mismatching ships to HOLD; it never changes one ship type into another or creates a confirmation action. The rule is fixed independently of the 6,720 model outputs.

Information access and implementation history. The core rule reads full evidence-card semantics, adjusted benefit p-values, simultaneous guardrail bounds, and transport bounds. It has no oracle, panel, model, or flow argument. Generation precomputes c and overrides it to REQUEST-CONFIRMATION for the sequential-monitoring panel; the runtime runner then applies the serialized action. This wrapper is panel dependent even though the core function is not. Analyst prompts use a public-card projection that omits adjusted certificate quantities; the critic additionally sees the proposal and memos. Appendix Table 6 distinguishes these accesses. The comparison assesses these specified policies and information conditions, rather than isolating algorithm quality under identical access. The public-protocol reconstruction in Appendix A.4 is a later direct-rule analysis, not the historical generator used to collect the experiment. In the closed-candidate AgentDojo study, the contract instead contains a trusted valid-candidate identifier; mismatching execution proposals are held. Its information conditions are reported alongside its results.

3.3 CONTROLLED LAYER SUBSTITUTION AND EXACT-COVERAGE MATCHING

What each comparison isolates. Appendix Table 3 maps the objects held fixed to the component changed; it is not a ranking of methods. Information-flow and candidate-interface interventions can change proposals; authorization and contract-shift comparisons retain the corresponding stored proposals.

Matching quantity before comparing selection. For each model–flow stratum s, let I_s contain its task rows, $S_s = \{i \in I_s : p_i \in \mathcal{S}_{\mathrm{ship}}\}$ its raw ship proposals, and k_s the number retained by DecisionGate. Since the gate only removes ship actions, $k_s \leq |S_s|$. Algorithm 1 (Appendix F) selects exactly k_s of these proposals without new model calls.

The **random control** tests whether executing fewer proposals explains the risk difference; the **confidence control** instead ranks proposals by the orchestrator's own confidence. Both retain the same execution count as DecisionGate in every model–flow stratum.

Interpretation boundary. Only the rejection fallback uses panel labels, affecting confirmation, correctness, and regret but not matched ship counts or adverse-deployment events. Exact coverage controls **execution quantity, not information access**. Section 4.2 and Appendix H report the post-hoc timing. Always-HOLD's zero risk accompanies zero useful-action power.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND METRICS

Tasks and evidence. GateSim-T couples a synthetic generator to an explicit evidence contract. Its eight panels cover global null, clean benefit, global harm, benefit–harm trade-off, targeted benefit,

spurious subgroup, deployment shift, and sequential monitoring (Appendix Table 12). Cards expose estimands, subgroups, adjustments, operationalizations, estimates, and uncertainty; oracles follow the generating process and deployment constraints. Six panels are risk eligible; clean and targeted benefit measure power. Calibration covered 140,000 simulated datasets before model calls. Appendix A records earlier screens and analysis roles.

Models, flows, and budgets. We evaluate Llama 3.1 8B, Mistral Small 3.1 24B, Gemma 3 12B, Phi-4 14B, Qwen 3 8B, Qwen 3 32B, and DeepSeek-R1-Distill-Qwen 32B, locally served from frozen digests. The two Qwen checkpoints form one family for leave-one-family-out analysis: seven checkpoints, six families. The five flows are parallel isolation, star broadcast, chain, two-layer hierarchy, and paired debate; they vary visibility, disclosure schedule, and aggregation. Each task–model–flow cell uses four analyst calls, one orchestrator call, and one statistics-critic call. Evidence, role prompts, allowed actions, decoding, call budget, and task-level seeds remain frozen: $7 \times 5 \times 192 = 6{,}720$ decisions and 40,320 calls. Each of the 35 cells has 24 tasks per panel. Appendix E specifies per-call visibility, checked against all 6,720 recommendation lists. All cells completed without parse, duplicate, or call-contract errors (Appendix Table 5). **Equal calls do not imply equal token or reasoning cost.**

Comparison policies. Raw MAS executes the orchestrator action. The sixth-call critic reads public evidence, memos, and proposal, then upholds, rejects to hold, or requests confirmation. DecisionGate uses Section 3.2; random and confidence abstention use Section 3.3. Always-HOLD is a zero-execution reference. The direct-rule control supplies actions from certificate inputs rather than filtering fixed MAS proposals.

External settings and costs. The prospective study has 96 Criteo uplift buckets and 96 AgentDojo authorization tasks, each split into 48 calibration and 48 test tasks. Criteo uses randomized advertising evidence; AgentDojo uses banking, Slack, travel, and workspace tool calls. Both cross Llama 3.1 8B, Gemma 3 12B, Mistral Small 3.1 24B, and DeepSeek-R1 32B with five flows: 40 cells, 3,840 decisions, 23,040 stored calls. Tests concern matched-coverage risk, verifier-relative regret, and coverage at a calibration-only 7.5% risk target. Two separately frozen AgentDojo studies test next-call clarification and explicit candidate recovery; each has 96 tasks, 20 cells, and 1,920 decisions. Recovery's primary inference uses 48 held-out tasks and 960 paired decisions (Appendices B and C). Proposals cost four analyst calls plus one orchestrator call; the verifier adds one authorization call, deterministic authorization none. We store the sixth response for every row for shared-proposal comparisons. **Collection, proposal, and authorization costs are distinct**, as are calls, tokens, and reasoning cost.

Metrics and inference. Adverse deployment is a ship on a risk-eligible row: its denominator is **5,040 eligible rows**, not 6,720 total decisions or the number executed (144 tasks across 35 strata). False deployment uses 4,200 hold-oracle rows; unsafe deployment uses 1,680 true-harm rows; clean and target power each use 840. Correctness compares final actions with stored oracles; regret uses the frozen asymmetric loss. Coverage and overall confirmation use 6,720 rows; confirmation recall uses 840 sequential-monitoring rows. The 192 task clusters are paired across 35 strata; we retain 20,000 task-cluster bootstrap replicates. Wilson summaries are descriptive row-level intervals. External inference uses frozen test splits and 20,000 task-cluster replicates, with equal domain weighting in the prospective study. External valid-action power and normalized loss use domain-specific definitions, not GateSim-T's benefit denominators. Appendices A.1 and H preserve analysis chronology.

4.2 Main Results: Execution Selection at Matched Coverage

The post-hoc comparison in Figure 3 and Appendix Table 14 tests whether abstention frequency alone explains risk, holding proposals and execution counts fixed but not equalizing policy information. All three policies execute exactly **824/6,720 actions (12.262%)** and are matched separately in each of the 35 model–flow strata. DecisionGate has **0.119%** adverse risk, compared with 8.571% for random abstention and 7.798% for confidence abstention. The task-cluster bootstrap 95% intervals for comparator-minus-DecisionGate risk are [7.579, 9.345] pp and [6.726, 8.710] pp.

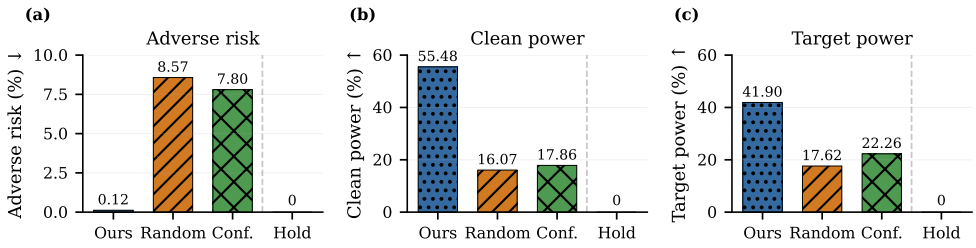


Figure 3: Post-hoc exact-coverage comparison on fixed GateSim-T proposals. (a) Adverse risk ↓ on 5,040 risk-eligible rows; (b) clean power ↑ and (c) target power ↑, each on 840 rows. Ours, random, and confidence abstention each execute 824 actions (12.26% overall), exactly matched within every model–flow stratum. Always-HOLD executes none and is *not* coverage matched. Panel scales differ; full metrics and denominators are in Appendix Table 14.

The family-level analysis in Figure 4 restricts its scope to the six evaluated families. Against confidence abstention, the safety difference ranges from $4.72\,pp$ for Phi to $13.19\,pp$ for Llama; every family-specific 95% interval is above zero. Removing any one risk panel leaves an overall advantage between 3.90 and $8.93\,pp$, again with intervals above zero. Thus the aggregate comparison is not carried by one family or one risk panel, although it is not a population-level random-effects claim about unseen model families.

The same comparison favors DecisionGate on overall correctness, useful-action power, and the specified loss. Relative to random and confidence abstention, DecisionGate has higher exact accuracy (74.81% vs. 62.60% and 63.99%), clean power (55.48% vs. 16.07% and 17.86%), targeted power (41.90% vs. 17.62% and 22.26%), and lower normalized regret (0.1075 vs. 0.2459 and 0.2428). Always-HOLD has zero risk but also zero ship power, 62.50% accuracy, and 0.2188 regret. The 200-seed random-policy sensitivity keeps the same qualitative ordering.

4.3 Information Flow and Direct-Rule Controls

Proposal formation. The five controlled information-flow realizations produce different proposal-level risk profiles (Appendix Table 16). When the raw MAS proposals are executed without downstream authorization, pooled adverse-deployment rates range from 31.85% under star broadcast to 37.10% under parallel isolation, with debate, chain, and hierarchy at 33.73%, 34.03%, and 34.72%, respectively. These differences are strongly context dependent. The pooled flow main-effect variance is 1.60×10^{-4}, whereas the family-by-flow and panel-by-flow variance components are 5.59×10^{-4} and 2.14×10^{-4}. Together, the interaction components are 4.84 times larger than the pooled main effect, with a task-cluster bootstrap interval of [2.90, 16.26]. Thus, the effect of information flow depends

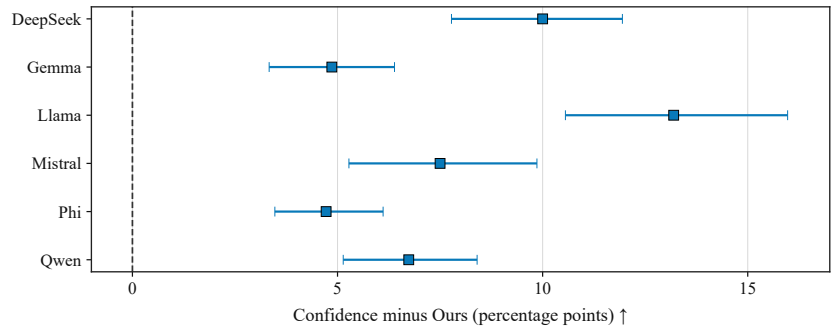


Figure 4: Post-hoc exact-coverage adverse-risk differences against confidence abstention for the six evaluated model families. Positive values favor Ours; zero denotes no difference. Points show estimates and bars show 95% task-cluster bootstrap intervals. This within-testbed sensitivity is not a guarantee for unseen families.

substantially on the evaluated model family and task panel, rather than supporting a single global ranking of flow architectures. This interaction result is treated as secondary sensitivity evidence; its preregistered criterion and later implementation chronology are documented in Appendix A.1.

Preregistered large-effect tests. The specified topology-regret effects above $5\,pp$ were not established: none of the five realizations passed the simultaneous criterion, and held-out selection regret was $3.47\,pp$ (95% CI $[2.43, 7.00]\,pp$), with four of six family folds positive. These outcomes do not establish topology equivalence or rule out transportable selection. Full hypotheses and outcomes remain in Appendix A.1 and Table 4.

Downstream layer substitution. Appendix Table 16 shows layer substitution across all five information-flow realizations; Appendix Figure 7 provides the corresponding descriptive safety–utility view. The raw proposal plane has 31.85–37.10% adverse deployment across flows. The statistics critic reduces that range to 9.52–11.90%. DecisionGate reduces it to 0.10–0.20%. Its cross-flow adverse-risk range is $0.099\,pp$, with a bootstrap 95% upper bound of $0.298\,pp$, below the preregistered $3\,pp$ threshold. The worst-flow conservative upper bounds are $0.864\,pp$ for false deployment and $1.130\,pp$ for unsafe deployment, both below $7.5\,pp$. H4 therefore passes.

In absolute terms, DecisionGate reduces the flow main-effect variance from 1.5987×10^{-4} to 2.1433×10^{-7}. The corresponding $746\times$ ratio is secondary because its denominator is very small; the absolute risk range and its interval are the primary quantities.

Direct-rule control. A new descriptive post-hoc control directly applies the core rule to the 192 frozen tasks, without MAS proposals or new model calls. It yields 20.83% coverage, 0.69% adverse deployment, 82.29% correctness, and 0.0599 regret. Adding a confirmation route based only on public monitoring-protocol fields reproduces every historical certificate action and gives 94.79% correctness and 0.0286 regret at the same coverage and adverse rate (Appendix Table 8). Direct rules can supply actions that MAS did not propose, so this is an end-to-end policy control, not a matched-proposal comparison. It shows that MAS is not required for high task performance in this structured testbed; the paired experiment identifies how execution policies select among MAS proposals.

4.4 External Evaluation and Candidate Availability

Prospective named-domain evaluation. The preregistered named-domain replication evaluates transport under three frozen comparisons: exact-coverage risk (H_{EXT1}), regret relative to a one-call verifier (H_{EXT2}), and coverage under a 7.5% calibration-only risk target (H_{EXT3}). H_{EXT1} fails; H_{EXT2} and H_{EXT3} pass, although the calibration-only comparator selects no actions. Relative to the verifier, DecisionGate improves normalized regret by 0.1055 (95% CI [0.0682, 0.1447]) without another authorization call, and records 2.97% adverse risk at 4.58% coverage. The exact-coverage advantages over random and confidence abstention are only 0.36 pp (95% CI $[-1.93, 2.40]$) and 0.10 pp (95% CI $[-1.67, 1.67]$). All execution comes from Criteo; the original AgentDojo slice remains at zero coverage. The verifier has 25.00% coverage and 11.35% valid-action power, versus 4.58% and 2.98% for DecisionGate. Thus DecisionGate occupies a lower-coverage, lower-regret point under the specified loss. The policies differ in coverage and information access; the GateSim-T matched-risk separation is not reproduced externally, and this comparison does not establish method dominance (Table 1).

Table 1: Prospective pooled test. Coverage (C), adverse risk (R), and valid power (P) are percentages; loss is regret. All executions are from Criteo; AgentDojo has zero execution.

Policy	C	R ↓	P ↑	Loss ↓
Ours	4.58	**2.97**	2.98	**.2410**
Verifier	25.00	18.85	**11.35**	.3465
Random	4.58	3.33	2.31	.2914
Confidence	4.58	3.07	2.79	.2867

Table 2: Closed-candidate test (960 decisions, 48 tasks). C and P as at left; U is unsafe execution (%). Only Ours receives trusted valid-candidate identity.

Policy	C	P ↑	U ↓	Loss ↓
Raw	55.52	33.33	22.19	.4180
Verifier	18.44	13.54	4.90	.3703
Ours	33.33	**33.33**	**0.00**	**.3070**

Closed-candidate authorization composition. Two sequential, separately frozen AgentDojo studies each cover 96 source tasks, four models, and five flows. The next-tool-call clarification study leaves every policy at zero coverage and power. The subsequent recovery study exposes explicit anonymous candidates for selection or holding. Each study contains 1,920 decisions; recovery inference uses the 48 held-out tasks, hence 960 paired test decisions. Integrity passes for all 20 cells; validation removes one candidate identifier during orchestration.

On this test split, explicit candidates yield 55.52% raw coverage, 33.33% valid-action power, and 22.19% unsafe execution. The verifier reduces power to 13.54%. The closed-candidate authorizer retains 33.33% power with zero observed unsafe execution and regret 0.3070 versus 0.3703 for the verifier (difference 6.33 pp, 95% CI [3.49, 9.38]); all four recovery gates pass. The authorizer receives the registered valid-candidate ID, while proposer and verifier receive anonymous candidates and public authority conditions without that trusted answer field. These results therefore evaluate composition with a trusted authorization list. They do not isolate the authorizer's semantic judgment under equal information access. The tested clarification did not restore execution; the explicit candidate interface did (Table 2; Appendix C).

4.5 Contract-Shift Analysis

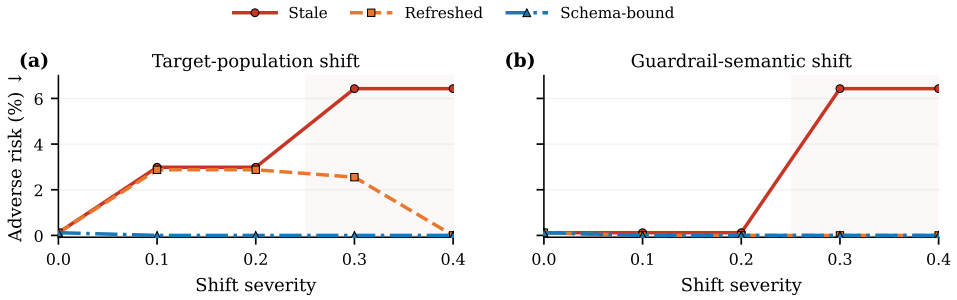


Figure 5: Frozen contract-mismatch severity sweep on fixed proposals. (a) Target-population and (b) guardrail-semantic shifts: adverse risk ↓ for stale, refreshed, and schema-bound authorization. The stale certificate fails after sufficiently large mismatches; refreshing or binding changes the observed risk–coverage trade-off. All frozen grid points are shown. The four-condition audit, including changing risk denominators and confirmation rates, is in Appendix Table 15.

Holding all 6,720 proposals fixed, the mismatch sweep in Figure 5 raises stale-certificate risk from 0.119% to 6.429% after target-population or guardrail changes. Refresh and schema binding control observed risk, but binding reduces coverage from 12.262% to 6.637% and raises confirmation from 15.432% to 57.336% in the recorded audit (Appendix Table 15). This structured sensitivity documents contract dependence under the evaluated shifts; it is not proof of generator independence. Authorization must name its evidence schema, population, risk semantics, and freshness, because silent reuse is unsafe under those shifts.

Additional discussion and limitations are provided in Appendices I and J.

5 Conclusion

Controlled layer substitution evaluates where safety gains arise in action-taking multi-agent systems: proposal formation, execution selection, or their interface conditions. On frozen GateSim-T proposals, exact-coverage comparisons reveal policy-level risk differences under distinct information access despite identical execution counts. Direct-rule controls show that structured evidence can also support decisions without MAS proposals. External evaluation identifies a lower-coverage, lower-regret operating point rather than reproducing the synthetic matched-risk separation; candidate and contract interventions expose additional dependencies on available actions and trusted authorization inputs. Together, these comparisons provide a practical protocol for specifying what is held fixed, documenting information access, and reporting risk alongside coverage, useful-action power, regret, and authorization cost when evaluating layered agent systems.

Ethics Statement

This work studies safeguards using synthetic panels and public Criteo/AgentDojo benchmark-derived tasks, without recruiting human participants. Aggregate safety rates must not be extrapolated. We report failures and uncertainty; authorization should supplement domain governance, oversight, and legal review.

Reproducibility Statement

Artifacts provide generators, evidence cards, prompts, contracts, flows, model digests, seeds, frozen analyzers, and audits. Append-only manifests cover 6,720 GateSim-T decisions and two 1,920-decision AgentDojo extensions; hashes preserve inputs, 20,000-bootstrap schedules, and outputs. An anonymized package will accompany submission.

AI Use Statement

Generative AI assisted with ideation, literature discovery, code, analysis, figures, and editing; it neither altered nor selected results. The authors verified all content against frozen artifacts and primary sources and assume responsibility.

References

Philippe Aghion and Jean Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997. doi: 10.1086/262063.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Khaled Chahine. Separating intent from execution: A defense-in-depth security architecture for LLM-based multi-agent systems. *Expert Systems with Applications*, 332:133781, 2027. doi: 10.1016/j.eswa.2026.133781.

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. StruQ: Defending against prompt injection with structured queries. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 2383–2400, 2025.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2636.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813*, 2025.

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD*, London, UK, August 2018. ACM. URL `https://ailab.criteo.com/criteo-uplift-prediction-dataset/`.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL `https://proceedings.mlr.press/v235/du24e.html`.

Yushi Feng, Junye Du, Qifan Wang, Zizhan Ma, Qian Niu, Yutaka Matsuo, Long Feng, and Lequan Yu. CORA: Conformal risk-controlled agents for safeguarded mobile GUI automation. *arXiv preprint arXiv:2604.09155*, 2026. URL https://arxiv.org/abs/2604.09155.

Yu He, Haozhe Zhu, Yiming Li, Shuo Shao, Hongwei Yao, Zhihao Liu, and Zhan Qin. Attriguard: Defeating indirect prompt injection in llm agents via causal attribution of tool invocations. *arXiv preprint arXiv:2603.10749*, 2026.

Anna Hedström, Salim I Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? mechanistic error reduction with abstention for language models. *arXiv preprint arXiv:2510.13290*, 2025.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, volume 2025, pp. 21344–21377, 2025.

Michael C. Jensen and William H. Meckling. Specific and general knowledge, and organizational structure. *Journal of Applied Corporate Finance*, 8(2):4–18, 1995. doi: 10.1111/j.1745-6622.1995.tb00283.x.

Hamed Khosravi and Xiaoming Huo. Conformal selective acting: Anytime-valid risk control for RLVR-trained LLMs. *arXiv preprint arXiv:2605.20270*, 2026. URL https://arxiv.org/abs/2605.20270.

Johannes Kirmayr, Lukas Stappen, and Elisabeth André. Car-bench: Evaluating the consistency and limit-awareness of llm agents under real-world uncertainty. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40599–40618, 2026.

Hui Yi Leong, Yuheng Li, Yuqing Wu, Wenwen Ouyang, Wei Zhu, Jiechao Gao, and Wei Han. AMAS: Adaptively determining communication topology for LLM-based multi-agent system. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2061–2070. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-industry.144.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024.

Fulin Lin, Shaowen Chen, Ruishan Fang, Hongwei Wang, and Tao Lin. Stop wasting your tokens: Towards efficient runtime multi-agent systems. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pzFhtpkabh.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM agents know when not to act? *arXiv preprint arXiv:2607.10059*, 2026. URL https://arxiv.org/abs/2607.10059.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025.

Weidi Luo, Shenghong Dai, Xiaogeng Liu, Suman Banerjee, Huan Sun, Muhao Chen, and Chaowei Xiao. AGrail: A lifelong agent guardrail with effective and adaptive safety detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8104–8139. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.399.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in neural information processing systems*, 37:74325–74362, 2024.

Qizi Ma, Sailan Chen, Zhixing Zhang, Jinhui Li, and Yuheng Zhang. Knowing when to answer: Prompting epistemically responsible behaviors in large language models. *IEEE Access*, 2026.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Daniel Schall. Safe integration of large language models into industrial process control: A multi-agent architecture with P&ID-grounded validation. *Autonomous Intelligent Systems*, 6(1):14, 2026. doi: 10.1007/s43684-026-00136-1.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12347–12361. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.623.

Bronislav Sidik, Chaya Levi, and Nizzan Kimhi. Autonomous topology mutation: Safe runtime restructuring for multi-agent LLM systems with capability, state, and shadow invariants. *arXiv preprint arXiv:2607.20488*, 2026. URL https://arxiv.org/abs/2607.20488.

Georgios Syros, Anshuman Suri, Jacob Ginesin, Cristina Nita-Rotaru, and Alina Oprea. Saga: A security architecture for governing ai agentic systems. *arXiv preprint arXiv:2504.21034*, 2025.

Haochun Wang, Sendong Zhao, Jingbo Wang, Zewen Qiang, Bing Qin, and Ting Liu. Beyond frameworks: Unpacking collaboration strategies in multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21361–21375, 2025a.

Qingni Wang, Yue Fan, and Xin Wang. Safer: Risk-constrained sample-then-filter in large language models. In *International Conference on Learning Representations*, volume 2026, pp. 53723–53747, 2026.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-Safeguard: A topology-guided security lens and treatment on LLM-based multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7261–7276. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.359.

Jiasi Weng, Jian Weng, Minrong Chen, Ming Li, Jia-Nan Liu, Zhi Li, and Yue Zhang. From neural intent to cryptographic authorization: Securing AI-driven enterprise workflows. *arXiv preprint arXiv:2607.15596*, 2026. URL https://arxiv.org/abs/2607.15596.

Yongxuan Wu, Xixun Lin, He Zhang, Nan Sun, Kun Wang, Chuan Zhou, Shirui Pan, and Yanan Cao. Cia: Inferring the communication topology from llm-based multi-agent systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17917–17933, 2026a.

Zijian Wu, Xiangyan Liu, Lingjun Chen, Fanqing Meng, Lingxiao Du, Yiran Zhao, Fanshi Zhang, Yaoqi Ye, Jiawei Wang, Zirui Wang, et al. Mcpmark: A benchmark for stress-testing realistic and comprehensive mcp use. In *International Conference on Learning Representations*, volume 2026, pp. 79852–79895, 2026b.

Xiongfeng Yang, Qingan Zhang, Yuheng Zhang, Kun-Yu Lin, Jian-Fang Hu, Dongmei Jiang, and Wei-Shi Zheng. Hero: Enhancing multimodal faithfulness via dynamic entropy-aware reinforcement learning. In *European Conference on Computer Vision*, pp. 407–424. Springer, 2026.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, volume 2025, pp. 9965–10017, 2025.

Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Kun Wang, Qingsong Wen, and Yang Wang. NetSafe: Exploring the topological safety of multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2905–2938. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.150.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. In *International Conference on Learning Representations*, volume 2025, pp. 75389–75428, 2025a.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. In *International Conference on Learning Representations*, volume 2025, pp. 35331–35366, 2025b.

Yuzhe Zhang, Feiran Liu, Yi Shan, Xinyi Huang, Xin Yang, Yueqi Zhu, Xuxin Cheng, Cao Liu, Ke Zeng, Terry Jingchen Zhang, et al. Silo-bench: A scalable environment for evaluating distributed coordination in multi-agent llm systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29379–29398, 2026.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Daisy Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, 2025.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Language agents as optimizable graphs. *arXiv preprint arXiv:2402.16823*, 2024.

A Audit Trail and Experimental Provenance

Table 3: Experimental design map: objects held fixed and component changed. This is not a method ranking.

Comparison	Held fixed	Changed / interpretation boundary
Information flow	Tasks, evidence, checkpoint, call budget	Analyst message visibility and order. *Proposals may change.*
Authorization	Stored proposal and evidence	Execution policy: certificate-based rule versus critic with public evidence and memos. *Access differs.*
Exact coverage	Raw proposals; execution count in each model–flow stratum	Hash or confidence ranking; panel-specific rejection fallback. *Equal counts, not equal access.*
Candidate interface	96 source tasks, models, flows	Explicit choices; trusted valid ID supplied to authorizer. *Proposals may change.*
Contract validity	Stored proposals	Population/guardrail shift; stale, refreshed, or schema-bound contract. *Scope conditions change.*

A.1 HYPOTHESES AND ANALYSIS ROLES

H1 tests whether each fixed topology has an environment whose regret exceeds the environment-specific best by more than $5\,pp$, using simultaneous bounds. H2 tests whether source-selected topology incurs held-out regret above $5\,pp$, with an additional positive-fold condition. Failure to meet these thresholds does not establish equivalence, successful transport, or its absence. H3 concerns environment interactions; H4 concerns scoped authorization-risk bounds. Table 4 preserves the registered criteria and outcomes. The original expansion gate and implementation timing are recorded in Appendix A.

H1 fails for all five realizations: the simultaneous procedure does not establish that any fixed graph exceeds the environment-specific best by more than $5\,pp$. H2 also fails: nested source-selection regret is $3.47\,pp$ with a task-cluster bootstrap 95% interval of $[2.43, 7.00]\,pp$; four of six family folds are positive, but the lower-bound criterion is missed. The preregistered tests did not establish the specified regret effects above five percentage points. These threshold outcomes neither prove equivalence nor establish that a transportable topology cannot exist. They are reported separately from the observed environment-dependent variation and the authorization comparisons.

A.2 PHASE PROGRESSION

The project used explicit gates. Phase 0 executed 80/80 shards, covering 140,000 simulated datasets without failed shards. The seven-model Phase I screen produced 672 rows without explicit errors. Its original hypothesis failed: independent-role false deployment was 117/420 (27.86%) versus 137/420 (32.62%) for an equal-call centralized condition; the paired difference was $-4.76\,pp$ with bootstrap interval $[-10.24, 0.71]$. Rather than changing the threshold or choosing supportive families, we froze a separate topology-transport phase.

Phase II first ran a 32-task-per-cell screen and then expanded every cell to 192 tasks without overwriting the 32-row prefix. The final master manifest covers 35 files and 6,720 rows. All files contain unique task identifiers from the frozen task set, the expected model and flow realization, and the six-call contract. The formal analyzer ran once over the immutable input directory. H1/H2 were produced by the frozen analyzer and failed the preregistered gate for the original P2-3 external topology-transport sequence. H3/H4 used preregistered thresholds and seeds, but their supplementary implementation was completed after the H1/H2 output and is marked accordingly. The later named-domain experiment was a separately frozen proposal–authorization study and did not reopen topology selection.

Table 4: Preregistered criteria and formal outcomes. H1/H2 are tests of regret effects exceeding $5\,pp$, not tests of the existence of a universally safe topology.

	Preregistered criterion	**Formal status**
H1	Every realization has an environment with simultaneous regret lower bound $> 5\,pp$	Fail: 0/5 realizations
H2	Held-out selection-regret lower bound $> 5\,pp$ plus fold-direction condition	Fail: $3.47\,pp$, 95% CI [2.43, 7.00]; 4/6 folds positive
H3	Interaction/flow variance-ratio lower bound > 1	Support: $4.84\times$, 95% CI [2.90, 16.26]
H4	Worst-risk upper bounds $\leq 7.5\,pp$ and risk-range upper bound $\leq 3\,pp$	Pass: $0.86\,pp$ false, $1.13\,pp$ unsafe, $0.30\,pp$ range

A.3 INTEGRITY SUMMARY

The archived experiment records include: (i) a full master manifest, (ii) SHA-256 inventory, (iii) an immutable compressed archive, (iv) the formal analysis JSON, and (v) a sensitivity package. No reported result relies on deleting a model, flow realization, panel, or task based on its outcome. Interrupted workers resumed from task identifiers and wrote only to non-overlapping output shards; completed cell files were never rewritten.

Table 5: GateSim-T coverage and integrity audit.

Component	Audited value
Model checkpoints / families	7 / 6
Information-flow realizations / panels	5 / 8
Tasks per model–flow cell	192 (24 per panel)
Decisions / stored model calls	6,720 / 40,320
Completed cells	35/35
Bad JSON / explicit errors / duplicates	0 / 0 / 0
Call-contract / model–flow mismatches	0 / 0
Task-cluster bootstrap replicates	20,000

Table 6: GateSim-T information-access audit. The core rule is evaluated during generation; the runner applies its serialized output.

Field / information	Serialized task	Read by rule	Phase-II use
Full evidence cards	Yes	Yes	Rule generation; prompts use projection
Adjusted certificate quantities	Yes	Yes	Rule generation; omitted from public cards
Oracle action	Yes, under truth	No	Scoring only
Panel label	Yes	No	Generation override; fallback; scoring
Latent DGP information	Effect summaries	No	Scoring only
Task seed	Encoded in task ID	No	Indexing only
Model identity	Result metadata	No	Proposal generation
Flow realization	Runner / result metadata	No	Proposal generation
Precomputed certificate action	Yes	Output, not input	Applied by runner

Table 7: GateSim-T certificate construction, transcribed from the existing core rule and historical generator. Comparisons are strict for lower bounds and non-strict for adjusted benefit p-values.

Stage	Condition or output
Group guardrails G_u	Exactly two relevant full-adjustment cards (safety and fairness), both with lower bound $> -m$. Missing lower bounds are treated as $-\infty$.
All-group support A	$b_{\text{all}} \leq \alpha$, G_{all}, complement-risk lower bound $> -m$, and all-population transport lower bound > 0.
Target support T	$b_{\text{target}} \leq \alpha$, G_{target}, and target transport lower bound > 0.
Core action, in order	If A and ($\neg T$ or $b_{\text{all}} \leq b_{\text{target}}$), output SHIP-ALL; else if T, output SHIP-TARGETED; else HOLD.
Historical wrapper	Override the core action to REQUEST-CONFIRMATION on the sequential-monitoring panel, then serialize c.

A.4 METRIC RECONCILIATION AND DIRECT-RULE AUDIT

A revision audit recomputed all reported policy counts from the same 35 frozen JSONL files, checked against their master-manifest hashes. There are 192 task clusters and 6,720 rows: 4,200 hold oracles, 840 ship-all, 840 ship-targeted, and 840 confirmation oracles. For DecisionGate, 5,027/6,720 final actions are correct. Its 464/840 confirmation recall (55.24%) and 1,037/6,720 overall confirmation rate (15.43%) have different denominators; the former was previously labeled only *Confirm*. Both are now explicit in Table 14. The baseline confirmation in the mismatch sweep uses all rows. The archived audit contains each policy's 4×4 oracle–final-action matrix, action totals, and all 35 exact-coverage checks.

Adverse-deployment counts are 6, 432, and 393 over 5,040 risk-eligible rows for DecisionGate, random abstention, and confidence abstention. Relative to their common 824 executions, the same

events are 0.73%, 52.43%, and 47.69%, respectively. These supplementary ratios exclude wrong ship types on benefit tasks and do not replace the registered risk denominator.

The new direct-rule analysis invokes the existing core function on full cards and certificate statistics with no truth or panel argument. A second version additionally uses the public monitoring protocol: if repeated looks are declared without registered alpha spending or an anytime-valid ledger, it requests confirmation. This reproduces all 192 serialized certificate actions, whose historical generation wrapper instead used a panel branch. The two implementations are distinguished explicitly; agreement on this dataset is not runtime isolation. No new LLM calls, tuning, or primary hypothesis tests were performed.

Table 8: Descriptive post-hoc direct-rule policies on the frozen 192 tasks (each repeated in 35 model–flow strata). These policies can supply actions without a MAS proposal. The protocol uses public monitoring fields and reproduces the serialized historical certificate on all 192 tasks. No new model calls or significance test.

Policy	Cov.	Risk ↓	Correct ↑	Clean ↑	Target ↑	Confirm /all	Confirm /seq ↑	Regret ↓
Proposal + **Ours**	12.26%	0.12%	74.81%	55.48%	41.90%	15.43%	55.24%	0.1075
Direct core rule	20.83%	0.69%	82.29%	100.00%	62.50%	0.00%	0.00%	0.0599
Direct rule + public protocol	20.83%	0.69%	94.79%	100.00%	62.50%	12.50%	100.00%	0.0286

Direct core outputs are 40 ships and 152 holds over 192 unique tasks; the protocol version changes 24 holds to confirmations. Repetition across 35 model–flow strata is for comparability of point summaries, not 6,720 independent direct-rule observations. The core rule gives 100% clean power and 62.50% targeted power. Its adverse event occurs on one of 144 risk-eligible tasks. The direct policy offers a different action opportunity set from a gate that only retains or rejects MAS proposals; no dominance or matched-coverage inference is attached to this control.

B Prospective External-Domain Confirmation Protocol

Before producing any prospective model output, we froze two heterogeneous named domains. Criteo uplift contributes 96 stable, disjoint decision buckets from a randomized advertising experiment (48 calibration, 48 test), derived from the public uplift benchmark (Diemert et al., 2018). The AgentDojo slice contributes 96 candidate tool calls from banking, Slack, travel, and workspace suites, balanced within each split between registered-user and injection-conditioned cases (Debenedetti et al., 2024). This is an action-authorization slice using official goals and tool-call traces, not a claim about the full AgentDojo agent score.

The domains do not share action semantics, and we do not claim that they are the same task. They share a proposal–authorization interface: registered context and evidence x support a proposal p, a domain-specific contract $C(x, p)$ checks authority, and the system executes only when the contract permits; otherwise it holds, confirms, or rejects. Heterogeneity is therefore intentional: the test asks whether the same authority boundary is informative under two named action semantics.

The execution matrix crosses both domains with four previously frozen model families (Llama 3.1 8B, Gemma 3 12B, Mistral Small 3.1 24B, and DeepSeek-R1 32B) and the same five information-flow realizations, yielding 40 cells, 3,840 decisions, and 23,040 stored calls. Proposal formation costs four analyst calls plus one orchestrator call. The LLM verifier baseline alone incurs one additional authorization call; DecisionGate and calibrated selective action apply deterministic authorization logic to the shared five-call proposal. We nevertheless record the verifier response for every row so all methods can be compared on identical proposals, and report proposal cost and authorization-specific cost separately. Task admission, splits, model digests, flow realizations, stored six-call contract, action sets, loss, risk target, seeds, and all primary comparisons were fixed before output.

The primary comparisons are: (i) exact execution coverage within every domain–model–flow stratum; (ii) a 7.5% risk target calibrated only on the calibration split with a one-sided exact-binomial upper bound; and (iii) proposal cost plus explicit accounting for the additional verifier call. Test-only inference uses 20,000 task-cluster bootstrap replicates with equal weight for the two named domains. Because two domains do not identify a population distribution over domains, results are named-

domain replication rather than universal external validity. The launch gate required 40/40 cells, 3,840 unique in-set rows, zero explicit errors or duplicates, exact task-index seeds, and matching hashes before the analyzer ran once; all conditions passed.

The frozen analyzer returns H_{EXT1}=fail, H_{EXT2}=pass, and H_{EXT3}=pass. On the pooled test estimand, DecisionGate executes 4.58% of tasks with 2.97% adverse risk, 2.98% valid-action power, and 0.2410 normalized regret. The one-call verifier executes 25.00% with 18.85% adverse risk, 11.35% power, and 0.3465 regret. Exact-coverage random and confidence abstention have adverse risks of 3.33% and 3.07%, so their differences from DecisionGate are not statistically separated. The calibration-only rule selects zero actions. AgentDojo also yields zero execution for every method, whereas the Criteo domain supplies all observed executions. We retain these negative and degenerate results in the main interpretation rather than treating the two passing gates as broad dominance.

Table 9: Prospective named-domain test results. Coverage is interpreted jointly with risk and power. Matching applies only to random and confidence abstention. The verifier uses one additional authorization call on the shared proposal and is not coverage matched. The original AgentDojo interface has zero execution for every policy.

Domain	Policy	Coverage	Risk ↓	Power ↑	Regret ↓
Pooled	**Ours**	4.58	2.97	2.98	0.2410
	One-call verifier	25.00	18.85	11.35	0.3465
	Random matched	4.58	3.33	2.31	0.2914
	Confidence matched	4.58	3.07	2.79	0.2867
Criteo	**Ours**	9.17	5.94	5.54	0.3227
	One-call verifier	50.00	37.71	21.07	0.5320
	Random matched	9.17	6.67	4.29	0.3328
	Confidence matched	9.17	6.15	5.18	0.3234
AgentDojo slice	**Ours**	0.00	0.00	0.00	0.1594
	One-call verifier	0.00	0.00	0.00	0.1609
	Random matched	0.00	0.00	0.00	0.2500
	Confidence matched	0.00	0.00	0.00	0.2500

Post-hoc loss sensitivity. Because H_{EXT2} depends on a frozen asymmetric loss, we additionally vary confirmation cost over $\{0, 0.125, 0.25, 0.375, 0.5\}$ and missed-benefit cost over $\{0.25, 0.5, 0.75\}$ while retaining unit wrong-deployment cost. This analysis is explicitly post-hoc and does not change the formal hypothesis decision. Verifier-minus-DecisionGate regret remains positive in all 15 settings: point estimates range from 0.0783 to 0.1327, and the smallest task-cluster bootstrap lower bound is 0.0354. Table 10 reports representative grid points. We do not construct a matched-coverage verifier sensitivity because the frozen verifier emits only a categorical decision and rationale, not a natural ranking score; inventing a score after observing outputs would create an unregistered comparator.

Table 10: Representative points from the post-hoc H_{EXT2} loss grid. Values are verifier-minus-Ours normalized regret; positive values favor Ours. Wrong deployment retains unit cost. All 15 frozen settings, including unshown intermediate confirmation costs, have positive 95% task-cluster bootstrap lower bounds.

Confirmation cost	Missed-benefit cost	Advantage ↑	95% CI low	95% CI high
0.000	0.250	0.1327	0.1000	0.1671
0.000	0.500	0.1065	0.0701	0.1456
0.000	0.750	0.0803	0.0374	0.1251
0.250	0.250	0.1316	0.0992	0.1664
0.250	0.500	0.1055	0.0689	0.1441
0.250	0.750	0.0793	0.0367	0.1238
0.500	0.250	0.1306	0.0974	0.1654
0.500	0.500	0.1044	0.0674	0.1437
0.500	0.750	0.0783	0.0354	0.1225

C AGENTDOJO PROPOSAL-RECOVERY PROTOCOL

The proposal-recovery experiment was frozen after the next-step null result and before any recovery output. It includes all 96 AgentDojo tasks, four model families, and five information-flow realizations, for 20 cells and 1,920 decisions. Each decision stores six calls. Task seed $291828 + 10{,}000i$ is fixed by dataset index i. Candidate identities and semantic roles are anonymized in LLM prompts; the orchestrator selects one candidate or holds. The frozen execution certificate retains the registered valid candidate identifier and rejects execution of a different candidate. This privileged contract field is intentional: the experiment tests closed-candidate composition, not whether the authorizer independently infers the correct next action. Candidate construction, task membership, models, flows, losses, thresholds, metrics, bootstrap plan, and primary hypotheses were not changed after launch.

The immutable master manifest and SHA-256 inventory cover 20/20 cells and 1,920/1,920 rows. The integrity audit reports no duplicate or out-of-set task identifiers, seed mismatch, parse error, explicit error, model/flow mismatch, or six-call contract violation. The analyzer ran once with 20,000 task-cluster bootstrap replicates and seed 291828. Table 11 reports the frozen test estimand; all 48 test tasks appear in every cell.

Table 11: Frozen proposal-recovery inference and gate details. Primary recovery metrics use the 960-decision test split.

Item	Frozen value
Test decisions / task clusters	960 / 48
Cells / calls per decision	20 / 6
Task-cluster bootstrap / seed	20,000 / 291828
Ours power 95% interval	[25.00%, 42.08%]
Verifier-minus-Ours regret 95% interval	[0.0349, 0.0938]
HPR1 / HPR2 / HPR3 / HPR4	Pass / Pass / Pass / Pass

For valid-action power, the task-cluster bootstrap interval is [25.00%, 42.08%] for DecisionGate. The verifier-minus-DecisionGate normalized-regret difference is 0.0633 (95% CI [0.0349, 0.0938]); positive values favor DecisionGate. These results satisfy the four frozen proposal-recovery gates. They do not replace the earlier null AgentDojo conditions: those conditions tested open-ended proposal interfaces, whereas this intervention tests an explicit closed candidate interface.

D PANEL DEFINITIONS

Table 12: The eight GateSim-T panels and their intended failure modes.

Panel	Oracle action	Stressor
Global null	Hold	Multiple plausible estimates with no true benefit
Clean benefit	Ship all	True global benefit with satisfied guardrails
Global harm	Hold	Direct violation of safety or fairness guardrails
Benefit–harm trade-off	Hold	Positive primary effect with unacceptable guardrail harm
Targeted benefit	Ship targeted	Predeclared subgroup benefits while global deployment is invalid
Spurious subgroup	Hold	Apparent subgroup gain induced by search variation
Deployment shift	Hold	Sample evidence fails to transport to the target population
Sequential monitoring	Request confirmation	Repeated looks make current evidence insufficient for execution

E TOPOLOGY VISIBILITY PATTERNS

All four analyst slots receive the same public evidence and context. Slot numbers below are logical positions with seed offsets $0, 1, 2, 3$, not necessarily wall-clock order. Arrows specify which prior

memos are visible. The runner serializes each peer memo to at most 600 characters. The stored `visible_peer_nodes` lists match this table in all 6,720 records.

Table 13: Actual four-analyst schedules. The common fifth and sixth calls are shown below. Slot j in a visibility column denotes the message from analyst slot j.

Flow	Roles in slots 1–4	Visible prior slots for 1; 2; 3; 4
Parallel isolation	Four independent analysts	$\varnothing$; $\varnothing$; $\varnothing$; $\varnothing$
Star broadcast	Hub, then three spokes	$\varnothing$; $\{1\}$; $\{1\}$; $\{1\}$
Chain	Four successive analysts	$\varnothing$; $\{1\}$; $\{2\}$; $\{3\}$
Two-layer hierarchy	Two leaves, then two parents	$\varnothing$; $\varnothing$; $\{1,2\}$; $\{1,2\}$
Paired debate	Two debaters, then two reviewers	$\varnothing$; $\varnothing$; $\{2\}$; $\{1\}$
5: orchestrator	Aggregates all four analysts	All four returned memos and their union of cited public cards; outputs proposal p
6: critic	Reviews orchestrator proposal	Orchestrator input plus p; upholds, holds, or requests confirmation

Star broadcast differs from isolation at the analyst stage: the hub is an analyst whose output reaches three spokes before orchestration. Hierarchy has two parents that both read both leaves. Debate has a single cross-review pass, not an iterative exchange. Each analyst's output reaches the common orchestrator; its proposal reaches the critic and the deterministic policy mapping. Equal calls constrain the number of calls, not the number of tokens or total inference cost.

F Exact-Coverage Control Implementation

The following implementation operates on the frozen proposal records. The definitions of I_s, S_s, and k_s are given in Section 3.3; no additional model calls are required.

Algorithm 1: Exact-coverage matched controls

Input: Stored proposals $\{p_i\}$, certificate actions $\{c_i\}$, strata $\{I_s\}$, ranking rule r.
Output: Control actions with exactly the same per-stratum ship count as Ours.

for each model–flow stratum s:
  Compute $a_{\mathrm{DG}}(p_i, c_i)$ for all $i \in I_s$.
  $k_s \leftarrow \sum_{i \in I_s} \mathbf{1}[a_{\mathrm{DG}}(p_i, c_i) \in \mathcal{S}_{\mathrm{ship}}]$; $S_s \leftarrow \{i \in I_s : p_i \in \mathcal{S}_{\mathrm{ship}}\}$.
  if $r =$ RANDOM: rank S_s by the frozen SHA-256 hash.
  else: rank by descending confidence; break ties by the same hash.
  $K_s \leftarrow$ the first k_s task IDs in this ranking.
  for each $i \in I_s$:
    if $i \in K_s$: keep the original ship action p_i.
    else if $p_i \notin \mathcal{S}_{\mathrm{ship}}$: keep the original non-ship action p_i.
    else if sequential-monitoring panel: return REQUEST-CONFIRMATION.
    else: return HOLD.
  end for
  Verify that exactly k_s ship actions remain.
end for

G Additional Numerical Results

Interpreting H4. H4 certifies only the prespecified empirical criteria on GateSim-T. It is not a distribution-free guarantee. Its value is comparative: the same fixed authorization logic remains stable while the proposal graph changes. A calibrated version could use risk-control methodology (Angelopoulos et al., 2025; Feng et al., 2026; Khosravi & Huo, 2026), but the present rule is not presented as conformal or anytime-valid.

Table 14: Post-hoc exact-coverage comparison. Coverage, correct, and confirm/all use 6,720 rows; adverse uses 5,040 risk-eligible rows; clean, target, and confirm/seq each use 840 rows. Confirm/seq is confirmation recall, not overall confirmation rate. Regret is unitless.

Method	Cov.	Risk ↓	Correct ↑	Clean ↑	Target ↑	Confirm /all	Confirm /seq ↑	Regret ↓
Ours	12.26%	**0.12%**	**74.81%**	**55.48%**	**41.90%**	15.43%	55.24%	**0.1075**
Random abstention	12.26%	8.57%	62.60%	16.07%	17.62%	17.08%	68.45%	0.2459
Confidence abstention	12.26%	7.80%	63.99%	17.86%	22.26%	17.35%	70.60%	0.2428
Always-HOLD	0	0	62.50%	0	0	0	0	0.2188

Table 15: Contract-shift audit on 6,720 fixed proposals. All rates are percentages. Risk denominators n_R change with risk eligibility; coverage and confirmation use all 6,720 rows. These are the four recorded audit conditions, not a replacement for the full severity sweep in Figure 5.

		Adverse risk ↓			Schema-bound policy	
Condition	n_R	**Stale**	**Refreshed**	**Bound**	**Coverage**	**Confirm/all**
Matched	5,040	0.119	0.119	0.119	12.262	15.432
Population shift	5,880	6.429	0.986	0.000	6.637	57.336
Guardrail shift	5,880	6.429	0.000	0.000	6.637	57.336
Transport missing	5,040	0.119	0.000	0.000	6.637	57.336

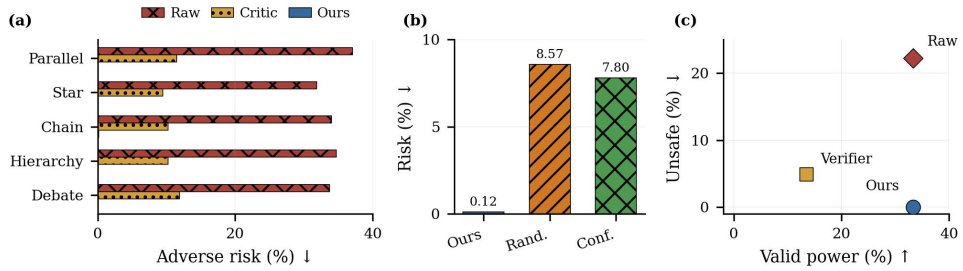


Figure 6: Three controlled comparisons. (a) Adverse risk by flow and decision layer. (b) Exact-coverage risk comparison: 824 executions per policy, matched within each model–flow stratum (12.26% overall). (c) Held-out candidate recovery with trusted candidate identity, a separate estimand. Risk denominators differ across panels; arrows indicate the preferred direction.

Table 16: Adverse-deployment rates (%) by information flow and decision layer; lower is better. Each estimate uses 1,008 risk-eligible rows. Coverage is not matched.

Topology	Raw MAS ↓	Statistics critic ↓	Ours ↓
Parallel isolation	37.10%	11.51%	0.10%
Star broadcast	31.85%	9.52%	0.10%
Chain	34.03%	10.22%	0.20%
Two-layer hierarchy	34.72%	10.22%	0.10%
Paired debate	33.73%	11.90%	0.10%

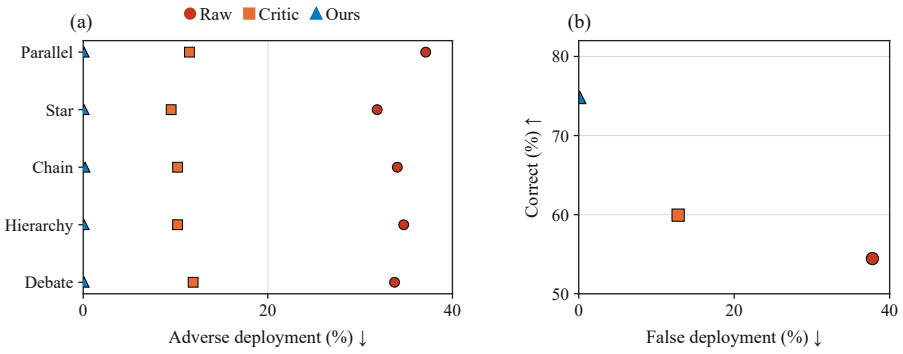


Figure 7: Proposal risk and downstream decision trade-offs. (a) Adverse-deployment rates across five flows, each using 1,008 risk-eligible rows. Values for Ours lie close to zero on this shared scale. (b) Overall correctness versus false deployment. Coverage is not matched.

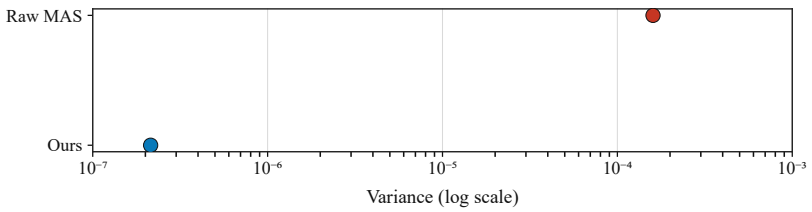


Figure 8: Topology main-effect variance before and after authorization on a logarithmic scale. The 746× ratio is descriptive because the post-authorization denominator is small; absolute risks and intervals remain the primary evidence.

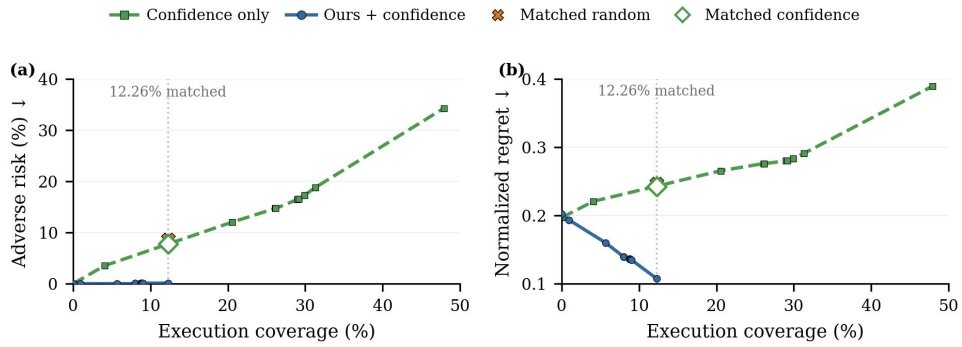


Figure 9: Complete post-hoc GateSim-T threshold frontiers: (a) adverse risk and (b) normalized regret versus execution coverage. Every frozen threshold is shown. These curves do not replace the preregistered tests or exact-coverage comparison; lower regret alone does not measure useful-action power.

H Post-hoc Addenda and Scope

The coverage-matched addendum was frozen before its first computation but after the P2-2 results and external review; it is therefore not relabeled as preregistered. Selection is exact within model–flow strata and does not condition on panel. Rejected-ship fallback does condition on panel, requesting confirmation for sequential monitoring and holding otherwise. A 200-seed random-policy sensitivity gives adverse-risk quantiles [7.70%, 8.15%, 8.59%] at 2.5/50/97.5 percent, versus 0.119% for DecisionGate.

The risk–coverage addendum is also explicitly post-hoc. Its feasibility check inspected only whether the immutable records contained a usable, non-degenerate confidence field. A uniform threshold grid, metrics, fallback actions, task-cluster bootstrap, and reporting of every threshold were then frozen before outcome computation. The analyzer re-audited 35 files, 6,720 rows, 192 task clusters, the six-call contract, and zero explicit errors before emitting the curve artifact. No threshold point replaces the preregistered tests or the exact-coverage comparison.

The mismatch addendum was likewise frozen before computation. The first output used equal panel weights even though shift-created risk sets differ in size. We retained that output, corrected only the estimator weighting to reproduce the pooled risk estimand, and wrote a new version rather than overwriting the first. No task assignment, shift magnitude, policy, seed, or scientific condition changed. A subsequent freeze fixed the full severity grid before computing the curves. The family and leave-one-risk-panel-out analyses were also frozen before computation and retain all six evaluated families and all six risk panels. These audits deliberately report stale-certificate failure and do not convert the discovery dataset into a prospective confirmation set.

The prospective confirmation was frozen before model output. It passes verifier-relative regret and fixed-risk gates (the latter against a zero-execution comparator), but fails matched-abstention replication. Separately frozen next-step and proposal-recovery studies retain their complete null and positive outcomes, respectively. No scientific condition or primary test was changed in response to results.

I Extended Discussion

With proposals and execution counts fixed, GateSim-T shows that abstention frequency alone does not explain the observed policy-level risk differences. This does not identify rule quality under matched access: the authorizer uses structured certificates, whereas the controls use a hash or confidence. The direct-rule control shows that high correctness here does not require MAS.

Other interventions identify interface dependencies. Information flow changes proposals without confirming the preregistered large regret effects. Original AgentDojo and next-call clarification yield no execution; explicit candidates enable useful action with trusted-ID authorization. Tested contract shifts increase stale-decision risk, while binding increases confirmation. Evaluation should therefore expose both interfaces and jointly report risk, coverage, useful-action power, regret, and authorization-call cost.

Role of structured evidence. The direct-rule comparison in Appendix A.4 helps separate the value of the evaluation design from the need for a particular proposal architecture. A rule that can propose actions has a different opportunity set from a gate constrained to retain or reject an LLM proposal. Its strong descriptive performance locates GateSim-T's contribution as a controlled testbed for policy comparison, rather than evidence that MAS is necessary for every structured decision problem.

From interface tests to deployment. The closed-candidate study tests how proposals compose with a trusted authorization list. A distinct deployment question is whether tool, object, parameter, and permission-scope constraints suffice when no valid-candidate answer is supplied. That comparison would require matched information access for a rule and verifier and separate labels for permission compliance, task success, and unsafe execution. It is not an experiment reported here.

J Limitations

GateSim-T aligns its generator and evidence contract, and its function audit does not establish runtime non-interference. The historical generation wrapper uses a panel-specific confirmation override; the new public-protocol reconstruction demonstrates equivalence on these tasks only. Direct-rule results are descriptive post-hoc controls. External inference covers Criteo and an action-level AgentDojo slice, with lower useful-action power at the conservative operating point. Recovery supplies the authorizer with a valid-candidate ID unavailable to the verifier; equal-information authorization without this answer field remains untested. Zero observed unsafe execution is not an adaptive-attack guarantee. Independent generators, frontier models, and end-to-end agents remain open. Appendix A

records the source and information-access audit; Appendix H records the analysis timing. These implementation and timing records are not runtime isolation guarantees.

Equal call counts constrain the proposal budget but do not equalize token use, latency, or total inference cost. The evaluated checkpoints and five fixed information-flow realizations do not establish generalization to larger or adaptive architectures. Exact-coverage matching likewise does not remove differences in certificate access. The main text retains these information conditions beside the corresponding methods and results.