INVESTIGATION CAPABILITY VERSUS INVESTIGATION INTELLIGENCE: DIAGNOSING SELECTIVE EVIDENCE-GATHERING IN LLM AGENTS WITH INFORMATION-STATE ORACLES

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-controlled agents can gather costly diagnostic evidence before updating a belief; this capability alone is consequential, raising final-task success from approximately 0% to 90-100% in a controlled sequential-decision benchmark. However, capability does not imply investigation intelligence: an agent may have access to evidence without knowing when acquiring it is worth its cost. The original LLM-driven investigation trigger shows no significant advantage over simple heuristic policies in 10 of 12 paired comparisons; the two nominally significant differences do not survive Bonferroni correction. We introduce a validated agent-information oracle evaluating policies on only agent-available information, alongside a clairvoyant oracle giving an achievable ceiling. Using these oracles, we decompose policy regret into missed and unnecessary investigation and downstream action errors, revealing a systematic mismatch between investigation behavior and the oracle's value of information in a noisy, budget-scarce regime. Factorial experiments identify accumulated dismissed-failure count, rather than recency, as the primary behavioral factor associated with under-investigation in this setting. A targeted intervention further demonstrates that this behavior is manipulable, improving success from 91.7% to 94.0% in one regime ($p = 0.020$, $n = 25$), but not uniformly across budgets and noise levels. Because an explicit uncertainty scalar alone did not correct the remaining gap, we construct matched counterfactual histories from legitimate investigation-mediated evidence and localize, in the Hard, budget-2 regime, a narrow, repeatedly-observed belief-conditioned mismatch: behavior is consistent with an effective update threshold that is offset above the oracle's decision boundary. We further extend this diagnostic apparatus to an environment in which evidence must be attributed to the correct one of several tools, and find that simple heuristic baselines exhibit cross-tool evidence contamination that is absent in the tested LLM agents under the conditions evaluated – a result that is robust across models, a hard lexical confound, a vocabulary-domain transfer, and randomized phrasing, subject to an explicit, model-dependent boundary condition under simultaneous multi-source distractor pressure. These results suggest the central difficulty is not access to evidence or representation of uncertainty, but the selective, budget-aware conversion of uncertainty into appropriate investigation and update decisions.

1 INTRODUCTION

Agents built on large language models (LLMs) increasingly act in environments where the correct action is not immediately obvious from a single observation: a tool call can fail for reasons that are transient, value-specific, or indicative of a genuine change in the environment, and these causes are frequently indistinguishable from the failure message alone. In such settings an agent has two decisions to make, not one: whether to change its belief about the environment, and whether that decision is worth further evidence-gathering first. We refer to the former as *investigation capability* – can the agent use a costly diagnostic action at all – and the latter as *investigation intelligence* –

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.

does the agent know *when* that action is worth its cost, given uncertainty, a finite shared budget, noisy diagnostics, and a limited remaining horizon.

This distinction matters because it is easy to conflate the two. A policy that always investigates whenever it can will appear highly capable by any accuracy-based evaluation, without ever having made a genuine trigger decision. We show that enabling investigation raises final-task success from approximately 0% to 90–100% in a controlled sequential-decision benchmark (Section 5) – evidence gathering is clearly valuable – yet the LLM-driven trigger policy studied here shows no significant advantage over simple heuristic policies in 10 of 12 paired comparisons. Capability, in other words, does not imply intelligent deployment.

The central methodological contribution of this paper is a pair of validated decision-theoretic oracles that let us move past this accuracy-level observation. A *clairvoyant* oracle, with access to the true hidden environment state, establishes an achievable ceiling. An *agent-information* oracle, restricted to exactly the information the agent itself could legitimately infer – its belief, the step index, remaining budget, and observed history – lets us ask a sharper question: was the agent's action rational given what it actually knew? Using this oracle we decompose policy regret into interpretable categories (missed investigation, unnecessary investigation, and downstream action errors) rather than reporting only aggregate accuracy.

This instrumentation reveals a specific, mechanistically-grounded failure: under noisy, budget-scarce conditions, the investigation policy studied is systematically mismatched with the oracle's value of information, and this behavior tracks the count of accumulated dismissed failures rather than the recency of the most recent reassuring observation. A targeted intervention removes the diagnosed mechanism and produces a real, statistically significant improvement in one regime, but the effect is not uniform across budgets and noise levels – a result we report honestly rather than oversell. A subsequent matched-counterfactual experiment, built entirely from legitimate investigation-mediated evidence rather than an artificially injected belief scalar, localizes a narrow, repeatedly-observed mismatch: the model's effective commitment threshold for updating its belief sits measurably above the oracle's own decision boundary, replicated across three independently constructed history sets.

Finally, we ask whether this diagnostic apparatus generalizes to a structurally harder version of the same problem: an environment in which evidence must be attributed to the correct one of several independent tools. Here the picture changes. Simple count- and rate-based heuristic baselines – which showed no measurable disadvantage relative to the LLM policy in the single-tool setting – exhibit clear cross-tool evidence contamination, wrongly updating a belief about one tool using evidence that concerns an entirely different tool. The tested LLM agents do not exhibit this contamination under the headline conditions, and this result is robust across three models, a hard lexical confound, a full vocabulary-domain transfer, and randomized phrasing, though it has a model-dependent boundary condition under simultaneous multi-source distractor pressure. This does not reverse the single-tool finding; it shows that extending the information structure of the decision problem exposes a limitation in the heuristic baselines that the original single-tool comparison could not distinguish.

Our contributions are: (1) a validated agent-information oracle and companion clairvoyant oracle enabling regret decomposition for selective evidence-gathering policies; (2) a mechanistic account of investigation-triggering failure driven by accumulated dismissed evidence, together with a targeted intervention showing this mechanism is behaviorally manipulable; (3) a matched-counterfactual localization of a belief-conditioned threshold offset, replicated across three independently constructed datasets; and (4) an evidence-scoping environment that exposes a cross-tool attribution failure in heuristic baselines that is absent in the tested LLM agents under the conditions evaluated, together with an honestly-characterized model-dependent boundary condition. We report every negative and null result alongside the positive ones, and we scope every claim to the model, regime, and environment in which it was actually observed.

2 Related Work

LLM agents and evidence-driven belief revision. A growing literature evaluates whether LLM agents can recognize that previously-valid information is no longer valid. Chao et al. (2026) introduce STALE, a long-context benchmark testing whether LLM agents detect an earlier memory

implicitly invalidated by a later, non-negating observation, finding that even strong models often retrieve updated evidence without acting on it. We ask a related but distinct question in a different setting: a short-horizon, budget-constrained sequential-decision task in which the agent must additionally decide whether to *pay* for evidence before revising a belief, evaluated against a validated decision-theoretic oracle rather than benchmark accuracy.

Wang et al. (2026) identify *belief inertia* in embodied LLM agents – adherence to a prior belief despite explicit contradicting observations – and propose an active intervention that improves task performance across embodied benchmarks. The accumulated-dismissal-driven suppression we characterize (Section 6) is consistent with belief inertia as a general phenomenon, and can be read as a mechanistic dissection of a specific instance of it: we show *which* accumulated evidence drives the inertia and *where*, in belief space, a residual miscalibration remains even after a targeted intervention.

Vasudev et al. (2026) show that an LLM critic's offline accuracy at predicting task failure does not imply that acting on its predictions improves outcomes: an accurate critic can still cause severe degradation through a disruption-recovery tradeoff. This is conceptually adjacent to our capability-versus-intelligence distinction – both papers caution against inferring that access to relevant information implies the system deploys it well – though the specific mechanisms differ and we do not claim the two findings are the same phenomenon.

Evidence attribution and multi-source credit. The term *credit assignment* is used extensively in reinforcement learning for a different problem – attributing a trajectory-level outcome to the individual steps or agents that caused it. We deliberately avoid this term for our multi-tool result (Section 8) and instead use *evidence scoping*, since the question we ask is not which step caused an outcome but whether evidence about one tool is incorrectly treated as evidence about another.

3 Experimental Environment

The agent repeatedly calls a single fictitious API endpoint requiring one correctly-named parameter, either `city` or `location_id`. Across a 12-step episode, the true required parameter can undergo one unsignaled, permanent change; the agent's belief about the parameter name can be correct or stale relative to this true state. At each failure the agent chooses to IGNORE, UPDATE its belief (a toggle between the two parameter names), or INVESTIGATE – spending 1–2 units of a shared, episode-level diagnostic budget on a noisy retry procedure that itself resolves to a final IGNORE or UPDATE. Figure 1 sketches the resulting decision process (agent-visible state: own belief, remaining budget, step index, observed history; hidden from the agent: true category, true schema state, diagnostic noise realization).

Every failure narrative is deliberately uninformative about its cause. Five failure categories are drawn stochastically each step: *normal* failures faithfully reflect true match state; *transient flakes* and *genuine outliers* are forced failures unrelated to belief; *misleading feedback* faithfully reflects match state but is never, by the environment's own ground truth, a legitimate reason to update; *permanent change* is the genuine schema flip, guaranteed to occur in the second half of every episode if none occurs naturally. Investigation itself is noisy: a diagnostic retry's true outcome is reported as its opposite with a difficulty-dependent flip probability (0, 0.15, or 0.35), so investigation is not information acquisition at zero risk.

4 Decision-Theoretic Evaluation Framework

Our oracle construction draws on the classical formulation of costly, noisy information acquisition under partial observability (Howard, 1966), applied here to evaluate an LLM-driven policy's decisions rather than to derive an optimal policy from scratch.

Clairvoyant oracle. An exact finite-horizon dynamic program over the environment's literal executable action space, with full knowledge of the true hidden schema state at every step, gives an achievable ceiling. This oracle is schema-clairvoyant by construction and is therefore an unfair standard for judging any individual real-time decision an agent makes without investigating – it can, in effect, act as if it already knew the answer.

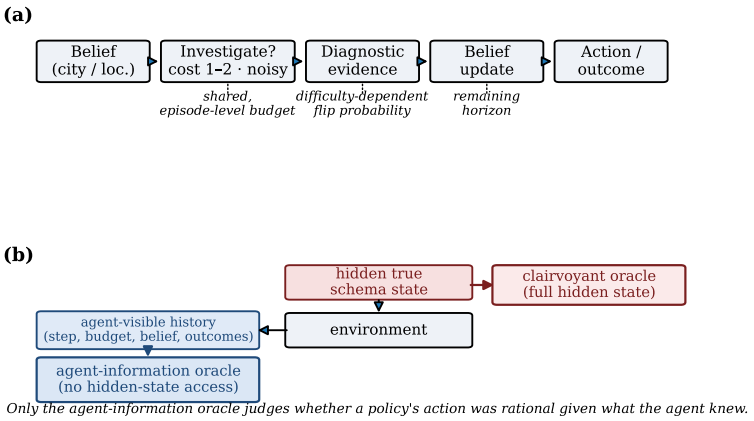

Figure 1: (a) Per-step decision process: investigation, costed against a shared budget and subject to diagnostic noise, mediates between belief and action. (b) The two oracles differ in what they may access: the clairvoyant oracle sees the true hidden schema state directly, while the agent-information oracle sees only what the agent itself could observe, and is the one used to judge whether a policy's action was rational.

Agent-information oracle. To ask whether an agent's decision was rational *given what it actually knew*, we construct a second, partially-observed oracle whose belief is a distribution over three mutually exclusive hypotheses about the true schema-change history (no change yet; one change at or after the episode midpoint, possibly forced; or a confirmed earlier or multiple change). This reduction is exact rather than approximate: the forced-versus-natural odds of an unobserved future change depend only on the step index, so all instances of "one change at some step j at or after the midpoint" collapse into a single hypothesis. The oracle's belief is updated via an exact Bayesian posterior computed only from agent-visible observables, using the same underlying step model as the clairvoyant oracle, so the update is provably consistent with the environment's true dynamics while never accessing hidden state. We validated this oracle in five parts, including a Monte Carlo check against its own dynamic-programming prediction, real-trajectory reconstruction with zero bound violations, and a source-level leakage audit confirming the posterior update never references hidden state (full validation in Appendix A).

Regret decomposition. Using the agent-information oracle, we decompose the gap between a policy's realized decisions and the oracle's prescribed action into five interpretable categories: correct, missed investigation, unnecessary investigation, wrong ignore, and wrong update. Aggregate task success tells us only whether a policy succeeds; this decomposition additionally tells us whether a given action was appropriate given the information available at that decision point, separating errors in *whether to acquire more information* from errors in the *direct action taken* once no more evidence is sought. This is what lets Section 5 move beyond "the policy performs worse" to say that its investigation decisions are systematically mismatched with value of information in a specific, characterizable way.

5 Investigation Capability versus Investigation Intelligence

Enabling investigation capability raises final-task success from approximately 0% to 90–100% (Figure 2a): the ability to gather evidence before committing to a belief update is unambiguously valuable in this environment. This alone, however, establishes nothing about whether the policy's trigger decision is intelligent – a policy that always investigates whenever possible would show the same capability-driven gain without making a meaningful trigger judgment at all.

At unlimited budget, the LLM-driven trigger policy (`llm-gate`) achieves 91.7% accuracy, comparable to or below several much simpler heuristic baselines: `repeat-priority` (95.8%), `random-matched` (95.6%), and `oracle-replay` (92.5%), and above `k-repeat` (77.8%), `never-update` (75.0%), and `always-update` (60.0%) (Figure 2b). Across 12 paired sta-

tistical comparisons (Wilcoxon signed-rank, uncorrected), 10 show no significant difference; the remaining two (`llm-gate` vs. `repeat-priority` at budget 8, and an uninformative-replay baseline vs. no-investigate) are nominally significant but do not survive Bonferroni correction. We do not read this as evidence that the LLM policy is worse than heuristics – the comparisons are overwhelmingly null in both directions – but as evidence that capability alone does not distinguish the LLM policy from simple counting rules, motivating the decision-theoretic analysis that follows.

At the agent-information-oracle level, this parity conceals a specific misalignment. In the hard-difficulty, budget-2 regime, the oracle's own investigate-rate at the states the policy actually visits is only 35–40%, yet the policy investigates in 68–97% of those same states: the policy investigates substantially more often than the oracle at those same visited states, producing systematic over-investigation relative to the oracle's selective decisions, this mismatch traced to a specific mechanism in Section 6.

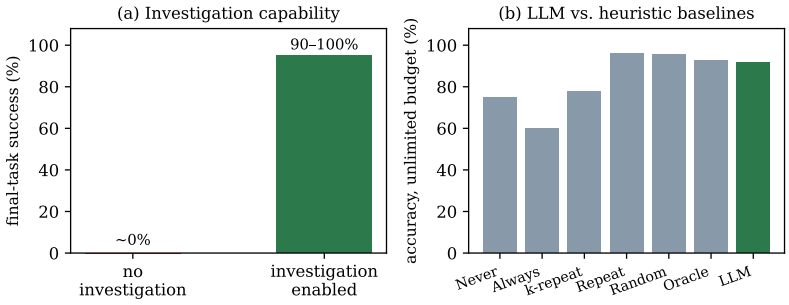

Figure 2: (a) Enabling investigation capability raises final-task success from approximately 0% to 90–100%. (b) At unlimited budget, the LLM-driven trigger (LLM, highlighted) shows no consistent advantage over several simple heuristic baselines: Never/Always-update, k-repeat, Repeat-priority, Random-matched, and Oracle-replay.

6 Mechanistic Analysis of Investigation Behavior

Accumulated dismissal, not recency. A controlled factorial probe (Figure 3b) varies the number of previously dismissed failures shown in the agent's history, holding the current ambiguous failure fixed. Investigation rate falls monotonically as dismissed count rises from zero to eight (100% to 0%), and, critically, a condition with four dismissed failures but *no* immediately preceding reassuring event produces the same rate as an otherwise-matched condition with fewer dismissals and immediate reassurance – indicating accumulated count, not recency of any single reassuring observation, is the driver. Placing the sole reassuring event *oldest* in an otherwise-identical four-dismissal history restores investigation to near baseline (Figure 3b, marker G). We term this accumulated-dismissal-driven suppression; the ordering sensitivity underlying this last observation is characterized further, and shown to be model-specific, in Appendix F.

Targeted intervention. A single added instruction – that history does not automatically expire old failures and each new failure requires its own evidence – removes the diagnosed pattern, producing a real, significant improvement where the mechanism was strongest: 91.7% to 94.0% at easy difficulty, budget 8 ($p = 0.020$, $n = 25$, 95% CI excludes zero, Cohen's $d = 0.52$). The same intervention is borderline at hard difficulty, budget 2 (74.0% to 80.7%, $p = 0.067$, $n = 25$) and null at medium difficulty, budget 2 (80.0% to 79.7%, $p = 0.97$, $n = 25$; Figure 6a) – evidence the mechanism is manipulable, not that the intervention resolves investigation intelligence generally.

Cross-model status. Re-running the factorial probe on two additional open-weight models (`gpt-oss-20b`, `qwen3.8-27b`) shows the coarse relationship – more dismissed failures, less investigation – replicates significantly on both ($p = 0.0049$, $p = 0.00035$). The finer-grained pattern does not: the C-vs-F equivalence and the reassurance-ordering effect take different, non-matching shapes per model, in one case reversing sign. We report the coarse mechanism as replicated and scope the fine-grained account to the model it was characterized on (full tables in Appendix B).

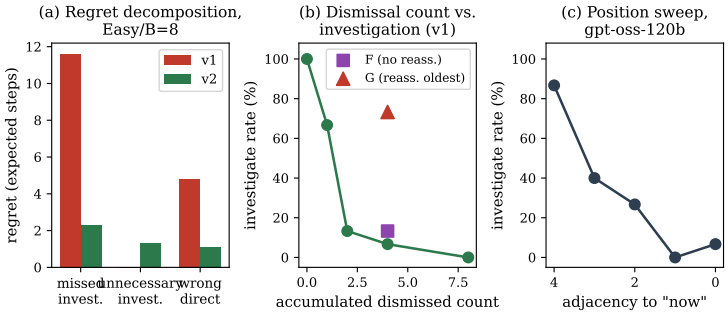

Figure 3: (a) Regret decomposition, easy/budget-8, before and after the targeted intervention. (b) Investigation rate falls with accumulated dismissed-failure count; matched conditions F and G isolate count from ordering. (c) A position sweep localizes a graded relationship between investigation rate and how recently the dismissed-failure run touches the current decision.

7 Belief-Conditioned Threshold Localization

The factorial and intervention results characterize *when* investigation is suppressed but not whether the residual gap, once a belief is available, reflects a representation problem (the agent cannot estimate its own uncertainty) or a control problem (it can estimate uncertainty but does not act on it). Directly exposing an explicit, correctly-computed uncertainty scalar – across three separate designs, including full-episode exposure with an independent-seed replication – did not correct the residual gap; direct belief-driven updates remained near zero under every such design (Appendix C).

We instead constructed *matched counterfactual histories*: sequences of only legitimate, investigation-mediated events that hit a precise target posterior belief, computed by replaying the event sequence through the same validated posterior update used throughout, holding step, budget, and surface prompt content fixed. This isolates whether exposure to the agent's own correctly-computed belief, rather than an arbitrary scalar, changes behavior, without also changing the history's content.

We evaluate this design as a single protocol applied across three independently constructed history sets, not three separate experiments. A five-region dataset $(0.05, 0.20, 0.40, 0.60, 0.80)$ and a three-region dataset localizing the oracle's boundary $(0.55, 0.60, 0.65)$ both show exposing the agent's own belief improves agreement and reduces regret at most regions, but produces *exactly zero* measurable effect at $P(\text{stale}) \approx 0.60$: agreement, regret, and the full action distribution are identical between conditions (Figure 4b; Table 1). A third, independently generated dataset excluding every prior composition does not replicate a clean null here – the shift is never positive across the three datasets (pooled Δ agreement ≈ -10 points) and is negative in two of three – which we read as more defensible than a clean null: exposure never produces a positive shift at the boundary, and mostly shifts the opposite way.

The oracle's own prescribed action transitions from INVESTIGATE at $P(\text{stale}) = 0.55$ to UPDATE at 0.60 and 0.65; the true boundary therefore lies between 0.55 and 0.60. Behavior is consistent with an effective update threshold that is offset above this boundary, bracketed by the tested 0.60–0.65 region – we do not claim a point estimate, a causal mechanism, or generalization beyond the tested hard-difficulty, budget-2 regime.

Oracle boundary:	INVESTIGATE $\longleftrightarrow$ UPDATE	within $[0.55,\ 0.60]$
Agent's effective threshold:	UPDATE triggers from	within $[0.60,\ 0.65]$ (interval, not a point)

8 Evidence Scoping Across Tools

8.1 Why the single-tool comparison is insufficient

Section 5's heuristic-parity result is a comparison in a setting with only one source of evidence: it cannot distinguish a policy that correctly attributes evidence to the source it pertains to from one

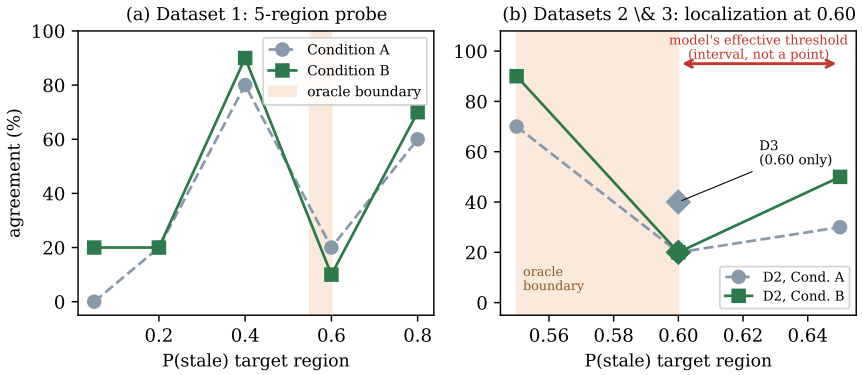

Figure 4: (a) Five-region matched-counterfactual probe (Dataset 1); the shaded band marks the oracle's decision boundary. (b) Localization around the boundary (Datasets 2 and 3): agreement is identical between conditions at $P(\text{stale}) = 0.60$ in Dataset 2, and Dataset 3's independently-generated 0.60 region shows no positive shift either.

Table 1: Belief-threshold localization at $P(\text{stale}) = 0.60$, three independent datasets ($n = 10$/cell). Oracle-prescribed action is UPDATE at this region in all three.

Dataset	Agreement, unexposed	Agreement, exposed	Δ agreement
1 (5-region)	20.0%	10.0%	-10 pp
2 (3-region)	20.0%	20.0%	0 pp
3 (3-region, independent)	40.0%	20.0%	-20 pp

that simply accumulates undifferentiated evidence, because there is only one source to accumulate. Changing the information structure from a single evidence source to several tests exactly this distinction – whether a policy's evidence-use preserves the identity of the source the evidence concerns, a property that is conceptually part of selective evidence acquisition but invisible to a single-source benchmark. We construct an environment in which the agent's current failure concerns one tool while its history may additionally contain dismissed failures concerning a different, unrelated tool, and ask whether accumulated-dismissal suppression (Section 6) stays correctly scoped to the tool it concerns, or generalizes incorrectly across tools that merely co-occur in the same history.

8.2 Cross-tool evidence attribution

Across three tested models, the rate at which the agent investigates a genuinely ambiguous current failure is unchanged whether the tool's own history is clean or contains four or eight dismissed failures concerning an entirely unrelated tool (pooled across models, $41/45$ versus $41/45$, Fisher's exact $p = 1.0000$; per-model results in Figure 5a). We evaluated the same conditions against three non-LLM heuristic baselines already used as comparison points in Section 5. Two count-based baselines (a k-repeat-style trigger and a windowed drift detector) exhibit clear cross-tool evidence contamination: both change their belief about the tool under test using only evidence that concerns the unrelated tool, i.e., with zero legitimate supporting evidence (Table 2). A third, narrative-caching baseline avoids this specific error, but only because the two tools' narrative text happens to differ verbatim in this construction; inspecting its cache directly shows the avoidance is due to exact-string non-match rather than any representation of tool identity. This reveals a limitation in the evaluated heuristic baselines that the single-tool comparison in Section 5 could not, by construction, distinguish.

8.3 Robustness

This result is robust to a hard lexical confound (dismissed-failure narratives differing by one token), a full narrative-vocabulary transfer with no shared terminology, randomized per-repeat phrasing, and an increased sample size ($n = 50$); full statistics for each condition are in Appendix D.

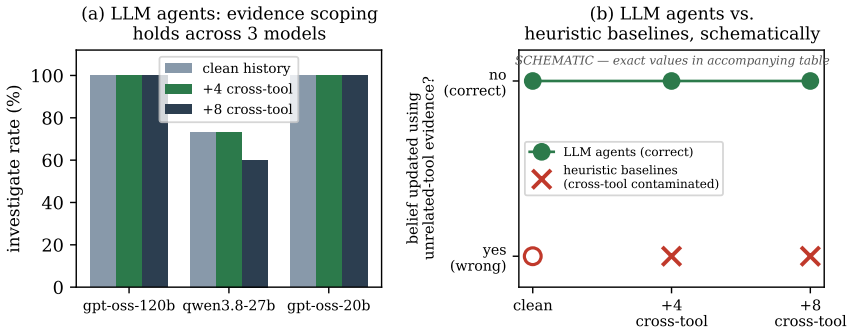

Figure 5: (a) Investigation rate is unchanged by irrelevant cross-tool evidence, across three tested models, under clean and cross-tool-distractor conditions. (b) The evaluated count-based heuristic baselines update the belief under test using only unrelated-tool evidence (schematic; exact values in Table 2), a failure mode absent in the tested LLM agents under these conditions.

8.4 Boundary condition: distractor-source scaling

Holding the total volume of irrelevant evidence fixed, we vary whether it originates from one distractor tool or is distributed across three. On the model characterized in Sections 6–7, investigation rate declines from 100% (one distractor tool, matched volume) to 80% (three distractor tools, same volume) to 60% (three distractor tools, increased volume; $p = 0.0169$ against the clean baseline) – a genuine, quantified degradation specifically under multiple simultaneous distractor sources, not under distractor volume alone (Figure 6b). The same manipulation produces no measurable degradation on a second model, and this is not attributable to a general difference in willingness to investigate: the second model's same-tool suppression under genuine evidence is comparable to or stronger than the first model's. We report this as a real, model-dependent boundary condition on the evidence-scoping result rather than either a universal limitation or an artifact to be explained away.

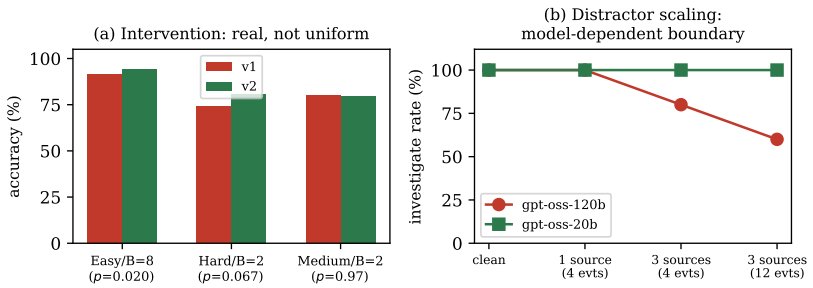

Figure 6: (a) The targeted intervention from Section 6 is real where the diagnosed mechanism is strongest and does not generalize uniformly across regimes. (b) The distractor-source-scaling boundary condition: a genuine, model-dependent degradation under compounding distractor pressure.

Table 2: Evidence scoping: LLM agents versus non-LLM heuristic baselines under identical cross-tool conditions (tool-under-test history clean; distractor evidence concerns an unrelated tool only).

Policy	Clean history	4 cross-tool events	8 cross-tool events
LLM agents (3 models, pooled)	investigate	investigate	investigate
k-repeat baseline	IGNORE	UPDATE (contaminated)	UPDATE (contaminated)
Drift-detector baseline	IGNORE	UPDATE (contaminated)	UPDATE (contaminated)

9 DISCUSSION

Two forms of generalization evidence appear in this paper, and we distinguish them deliberately. The first is genuine cross-model replication: the coarse accumulated-dismissal relationship and the evidence-scoping headline result both replicate across three open-weight models from two model families (`gpt-oss` and `qwen`) and roughly a sixfold parameter range, accessed through a single inference provider. The second is narrative-domain transfer: evidence scoping additionally survives a change in surface vocabulary with no shared terminology, but this is not structural task generalization – the underlying decision structure is unchanged, and we do not claim it would hold under a genuinely different task type.

Read together, Sections 6–8 suggest a single account: the tested agents' investigation policy is systematically mismatched in *when* to trust accumulated evidence – suppressing investigation as dismissed-failure count grows, with a residual threshold offset even where an oracle-correct belief is directly available – but, conditional on investigating, the same agents correctly scope which accumulated evidence is relevant once the environment requires attributing evidence to distinct sources, a discrimination the evaluated heuristic baselines do not make. The central difficulty in this setting appears to be not access to evidence, nor an inability to represent uncertainty, but the selective, budget-aware conversion of uncertainty into appropriate investigation and update decisions.

10 LIMITATIONS

The environment is fully synthetic: one fictitious API, two possible parameter names (extended to a small number of additional tools for evidence scoping), and controlled, hand-specified failure and noise distributions. This lets us construct a validated decision-theoretic oracle at all – unavailable for real, uncontrolled agentic tool-use logs – but external validity to real-world deployment is correspondingly unestablished.

Rigorous analysis of the accumulated-dismissal mechanism, the intervention, and threshold localization is centered on one model; cross-model testing (Sections 6 and 8) confirms the coarse relationships on two further open-weight models but not the fine-grained ordering mechanism or the distractor-scaling boundary, both reported as model-specific. All three models are open-weight and accessed through the same inference provider; we did not evaluate a proprietary or larger frontier model. Several results rest on $n = 10$–15 decisions per cell, most notably threshold localization (Section 7), mitigated by treating the design as one protocol replicated across three independent datasets rather than trusting any one in isolation; Bonferroni and Benjamini-Hochberg correction was applied to the hypothesis tests used for positive-effect claims; the evidence-scoping equivalence (absence-of-difference) tests in Section 8 are analyzed separately, as their interpretation does not weaken under correction the way a marginal significant-difference claim does (Appendix E). Results not surviving correction are not used to support a headline claim.

The agent-information oracle assumes known category mixture and noise rate – appropriate for asking whether a decision was rational given a known environment, not genuine model uncertainty about the environment. Matched-counterfactual histories use only investigation-mediated events to avoid an out-of-distribution confound, but full-episode generalization was not separately revalidated; the evidence-scoping vocabulary transfer demonstrates robustness to narrative surface form, not a structurally different task – genuine task-type generalization remains the most important open direction.

11 CONCLUSION

Evidence-gathering capability and the intelligence to deploy it selectively are separable properties of an LLM agent. Using a validated pair of decision-theoretic oracles, we located a specific misalignment in investigation triggering, showed it is behaviorally manipulable but not uniformly correctable, and localized a residual belief-conditioned threshold offset surviving three independent replications. Extending the same apparatus to evidence attribution across tools shows this is not the whole story: the tested agents correctly scope evidence in a way simple heuristic baselines do not, subject to a model-dependent boundary under distractor pressure. We see the central contribution as the instrumentation itself, more than any single number it produces.

AI USE AND REPRODUCIBILITY STATEMENTS

Generative AI tools (Claude, Anthropic) were used substantially for research ideation and execution (including writing experiment code), literature retrieval, and drafting/revising manuscript text; all code was executed and verified by the author, who takes responsibility for the final content. Code (Sections 3–8) is supplementary material; Appendices A, E give full validation detail.

REFERENCES

H. Chao, Y. Bai, R. Sheng, T. Li, and Y. Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

R. A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

R. Vasudev, M. Russak, D. Bikel, and W. Alshikh. Accurate failure prediction in agents does not imply effective failure prevention. *arXiv preprint arXiv:2602.03338*, 2026.

H. Wang, C. T. Leong, J. Wang, and W. Li. Seeing isn't believing: Mitigating belief inertia via active intervention in embodied agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

A Oracle Validation Detail

Table 3: Clairvoyant oracle: Monte Carlo validation (2000 seeds/setting) against dynamic-programming prediction.

Setting	DP prediction	Monte Carlo (realized)
Easy, budget 0	86.24%	$86.35\% \pm 0.21\%$
Hard, budget 2	86.71%	$86.78\% \pm 0.22\%$

Table 4: Agent-information oracle: Monte Carlo validation (1500 seeds/setting).

Setting	DP prediction	Monte Carlo (realized)
Hard, budget 2	83.38%	$83.21\% \pm 0.30\%$
Hard, unlimited budget	83.90%	$83.83\% \pm 0.30\%$

Real-trajectory reconstruction (part D) shows zero pointwise bound violations across 429 real decision states reconstructed from logged episodes using only agent-visible fields, and calibrated belief estimates (e.g., mean reconstructed $P(\text{stale})$ when truly stale versus truly matched: 0.177 vs. 0.03 on one log file; 0.675 vs. 0.133 on a harder-difficulty log). Part E confirms via source inspection that the posterior update function contains no reference to true category, true schema state, or the environment's change log.

B Full Cross-Model Ordering and Factorial Tables

Table 5: Position-sweep ordering probe, investigate rate (%) by adjacency-to-now, three models.

Model	adj. 4	adj. 3	adj. 2	adj. 1	adj. 0
gpt-oss-120b	86.7	40.0	26.7	0.0	6.7
qwen3.8-27b	20.0	20.0	6.7	6.7	0.0
gpt-oss-20b	73.3	33.3	46.7	53.3	66.7

C Belief-Exposure Negative Results

Scalar-only exposure (v2 prompt, unchanged, plus an added computed $P(\text{stale})$ line): hard/budget-2, $n = 25$; accuracy 80.7% vs. 80.7% ($d = 0.00$, $p = 0.887$); zero direct updates across 49 decisions. An artificial forced-scalar sweep (six fixed belief values, two fixed histories, values decoupled from history content): zero updates across all 210 calls at every forced level. A full-episode design combining belief-compatible update semantics with disclosed diagnostic noise: hard/budget-2, $n = 25$, zero direct updates (0/45 decisions); hard/budget-5, original seeds, four direct updates across 102 decisions, concentrated in a single high-belief bin ($n = 4$ and $n = 1$); an independent-seed replication at the same setting produced zero direct updates across 93 decisions despite reaching higher peak beliefs than the original run. Pooled across both budget-5 runs: 4/195 decisions (2.1%) were direct updates, entirely from the original, non-replicated run. We treat this apparent tail as an artifact of small n, not a finding.

D Evidence-Scoping Robustness Detail

E Multiple-Comparison Correction

Thirty distinct hypothesis tests were conducted across the mechanism, threshold-localization, and evidence-scoping lines of work. Under Bonferroni correction ($\alpha = 0.05$), four survive: the ordering-probe trend test and homogeneity test on gpt-oss-120b, the corresponding position-sweep comparison, and the qwen3.8-27b dismissal-count trend. Two further tests survive Benjamini-Hochberg

Table 6: Eight-condition factorial probe, key comparisons, three models (gpt-oss-120b figures from the original factorial; C≈F is the load-bearing comparison for the accumulated-count account).

Comparison	gpt-oss-120b	gpt-oss-20b	qwen3.8-27b
C vs. F (dismiss. count matched)	$13.3\% \approx 13.3\%$	66.7% vs. 40.0% ($p = 0.27$)	6.7% vs. 46.7% ($p = 0.035$, n.s. a
D vs. G (ordering)	13.3% vs. 73.3% ($p = 0.00002$)	80.0% vs. 60.0% ($p = 0.43$)	13.3% vs. 6.7% ($p =$
Dismissal-count trend	–	$p = 0.0049$	$p = 0.00035$

Table 7: Evidence-scoping robustness conditions, gpt-oss-120b, headline comparison (clean vs. cross-tool-distractor).

Condition	Clean	Cross-tool distractor
Original (distinct tool name + parameter)	100.0%	100.0% ($n = 15$, $p = 1.0$)
Hard lexical confound (one-token difference)	100.0%	$100.0\% \rightarrow 73.3\%$ at $8\times$ volume ($p = 0.0996$)
Narrative-domain transfer (SRE vocabulary)	100.0%	100.0% ($n = 15$, $p = 1.0$)
Randomized phrasing (4 templates/repeat)	100.0%	100.0% ($n = 15$, $p = 1.0$)
Increased sample size	100.0% ($n = 50$)	100.0% ($n = 50$, 95% CI $[92.9\%, 100\%]$)

FDR correction but not Bonferroni: the gpt-oss-20b dismissal-count trend and a pooled anomaly from an early credit-assignment-style probe that did not replicate at increased sample size and is not used to support any claim in the main text. All remaining tests, including the qwen3.8-27b C-vs-F reversal reported in Appendix B, do not survive either correction and are reported only as descriptive, non-significant observations. The evidence-scoping equivalence tests (Table 2, Appendix D) are tests for the absence of a difference rather than the presence of one, and multiple-comparison correction does not weaken them in the same way it weakens a marginal significant-difference claim.

F Ordering-Effect Detail

Group 2 of the position-sweep probe (three run-breaking control conditions, each introducing a second reassuring event in addition to reordering the first) is explicitly confounded: it varies both ordering and total reassurance count simultaneously. All three conditions were directionally consistent with an additional contribution from total reassurance count beyond adjacency alone, but none reached significance at $n = 15$ (Fisher's exact $p = 0.264$, 0.483, 1.000), and this group is reported as suggestive only.