SEEKORSTOP: BENCHMARKING TOOL-MEDIATED BLUE-TEAM INVESTIGATION UNDER MATCHED EVIDENCE ACCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a model that makes correct decisions from supplied evidence also identify missing evidence, acquire it, and know when to stop? We introduce SeekOrStop, a benchmark that distinguishes evidence acquisition, decision quality, and completion using matched Passive and Active views of the same security investigation episodes. Episodes are generated through controlled execution in a real Kubernetes environment, with ground truth derived independently of sensor labels. Passive provides a registered reference evidence view, whereas Active begins with an initial observation and decides whether to retrieve additional records through finite read-only tools. Across 420 Expansion episodes and four model configurations, end-to-end Active accuracy is consistently 14–32 percentage points lower than Passive accuracy. Moreover, registered-route completion does not guarantee successful investigation: agents may complete the registered route yet make an incorrect decision, fail to submit, or continue retrieving after the registered decision requirements have been met. Controlled contrast and equivalence pairs further test joint correctness under specified relevant and irrelevant changes. SeekOrStop therefore evaluates evidence acquisition, evidence use, completion, and stopping separately, showing why supplied-evidence accuracy alone gives an incomplete account of agentic investigation.

1 INTRODUCTION

An investigation begins before the relevant evidence has been assembled. An agent must determine what is missing, where to look, how to integrate new observations, and when enough evidence has been gathered to commit to a decision. Language-model agents increasingly operate in such settings: ReAct interleaves reasoning and actions (Yao et al., 2023), while AgentBench evaluates agents across interactive environments (Liu et al., 2024). Yet answering correctly from a prepared evidence record and reaching the same decision through an interactive tool workflow impose different demands. Terminal accuracy alone cannot reveal where that process succeeds or breaks down.

Security investigation makes this distinction concrete. A control-plane record may show an attempt to execute a command in a container without establishing that execution occurred. Runtime evidence can confirm the activity; authorization may explain it; object relationships may connect it to a broader chain. The agent must decide which evidence to inspect, combine observations across sources, and avoid both premature commitment and unnecessary continuation. A supplied-evidence task evaluates judgment after the relevant reference evidence has already been assembled. Tool-mediated investigation additionally requires evidence selection, interaction, integration, and commitment.

Recent work begins to evaluate these dimensions separately. AR-Bench studies active questioning under incomplete information (Zhou et al., 2025); Sufficient Context distinguishes inadequate retrieved context from failures to use sufficient context (Joren et al., 2025); and SeekBench examines grounding, search recovery, and sufficiency assessment (Shao et al., 2026). Within security, ExCyTIn-Bench and SIR-Bench evaluate multi-hop investigation, evidence discovery, and tool use (Wu et al., 2026; Begimher et al., 2026). These developments motivate a more controlled question:

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.

for the same decision problem, how does performance change when reference evidence is supplied directly rather than reached through a tool-mediated investigation?

We introduce **SeekOrStop**, a benchmark designed around this matched evidence-access comparison. Each episode supports two conditions. **Passive** receives a deterministic registered reference evidence view and submits a decision. **Active** starts from an initial public observation and can choose among four read-only retrieval tools before submitting. The episode, Gold decision, decision rules, and terminal correctness criterion are shared across conditions. Some episodes are *already sufficient*: their initial observation satisfies the decision requirements, allowing SeekOrStop to directly test whether an agent commits without unnecessary retrieval. The design therefore links supplied-evidence judgment to the observable investigation trajectory leading to a final decision.

SeekOrStop constructs these episodes from controlled execution in an owned Kubernetes/Online Boutique environment. Scenario specifications and verified execution outcomes determine Gold independently of detector labels and model predictions. Kubernetes Audit, runtime events, authorization context, and object context are normalized into a frozen public evidence store. Model evaluation then operates through a deterministic read-only interface, while Gold, execution lineage, registered reference routes, and pair metadata remain evaluator-only (Figure 1). This separation combines execution-derived outcome provenance with reproducible evidence access and a clear public/private information boundary.

The benchmark evaluates both the final answer and the path used to reach it. Each Expansion episode has a registered acquisition contract defining a reference sequence of tools and canonical arguments. **Registered-route completion** records whether that workflow is completed, while **Investigation Success** additionally requires a correct final decision. Terminal correctness is scored independently, allowing correct decisions without full route credit and route-complete trajectories without correct finals to remain distinguishable. SeekOrStop further records submission, post-completion retrieval, and already-sufficient stopping behavior. Controlled **Contrast** and **Equivalence** pairs test whether correctness is preserved under specified decision-relevant and decision-irrelevant changes. Together, these measures expose investigation outcomes that aggregate accuracy alone would merge.

On the 420-episode Expansion cohort, all four evaluated model/provider configurations—GPT-5.5, Kimi-K3, GLM-5.3, and Qwen3.6-27B—perform worse under tool-mediated Active investigation than under the matched Passive reference condition, with Active–Passive accuracy differences of approximately 14–32 percentage points. Passive serves as a supplied-evidence reference rather than an oracle performance ceiling: it removes evidence selection while still requiring the model to interpret the supplied records and produce a valid decision.

The aggregate difference is heterogeneous across scenario families. For GPT-5.5, most of the lost correct decisions are concentrated in gap resolution (A3) and distractor disambiguation (A5), while several other families show much smaller condition differences (Table 6). Trajectory analysis further shows that registered-route completion, correct judgment, valid submission, and commitment are distinct observable outcomes: agents can answer correctly without completing the registered workflow, complete the workflow yet fail to submit correctly, or continue retrieving after the decision requirements have been met. Supplemental prompt, Oracle-Retrieval, and larger-budget diagnostics further show that these behaviors do not collapse to a single retrieval or budget failure. A historical 60-episode pilot is reported separately.

Our contributions are:

1. **A matched evidence-access benchmark.** SeekOrStop holds the decision problem and Gold fixed while comparing a supplied-evidence reference condition with tool-mediated investigation, making the transfer from prepared evidence to interactive investigation directly measurable.
2. **An execution-grounded construction methodology.** Controlled execution establishes Gold and public observations independently, while frozen evidence access, registered reference routes, controlled pairs, and a strict public/private boundary support deterministic and reproducible evaluation.
3. **Trajectory-level evaluation of investigation and stopping.** SeekOrStop jointly measures registered-route completion, terminal correctness, valid submission, stopping behavior, and

Table 1: Comparison of representative benchmarks by their primary evaluation dimensions. "–" indicates that a dimension is not an explicit evaluation focus.

Benchmark	Interactive evidence access	Trajectory evaluation	Matched access views	Evidence/stop diagnostics	Security domain
AR-Bench	✓	✓	Partial	–	–
SeekBench	✓	✓	–	✓	–
ExCyTIn-Bench	✓	✓	–	–	✓
SIR-Bench	✓	✓	–	–	✓
Open Security Benchmark	✓	✓	–	–	✓
SeekOrStop	✓	✓	✓	✓	✓

controlled-pair outcomes, revealing investigation failures and trade-offs that are hidden by final accuracy alone.

2 Related Work

Interactive agents and information seeking. ReAct interleaves reasoning and actions (Yao et al., 2023), while AgentBench, InterCode, WebArena, OSWorld, and AgentBoard evaluate interactive behavior and multi-turn progress (Liu et al., 2024; Yang et al., 2023; Zhou et al., 2024; Xie et al., 2024; Ma et al., 2024). Closer to our setting, AR-Bench studies active questioning under incomplete information (Zhou et al., 2025); Sufficient Context separates insufficient retrieved context from failures to use it (Joren et al., 2025); and SeekBench evaluates grounding, recovery, and sufficiency assessment (Shao et al., 2026). SeekOrStop complements these works by holding the decision problem and Gold fixed while comparing supplied-evidence and tool-mediated conditions, and by jointly reporting terminal and trajectory outcomes.

Stopping and controlled behavior. τ-bench studies tool-use reliability (Yao et al., 2025), while AgentAbstain evaluates whether agents withhold inappropriate actions (Liu et al., 2026b). SeekOrStop studies commitment within bounded investigation, including cases whose initial evidence already satisfies the decision requirements. Inspired by controlled behavioral testing such as CheckList (Ribeiro et al., 2020), it also uses Contrast and Equivalence pairs to test correctness under specified relevant and irrelevant changes.

Security investigation and cyber environments. ExCyTIn-Bench, SIR-Bench, Cyber Defense Benchmark, SecRespond, and Open Security Benchmark evaluate evidence discovery, threat hunting, incident response, and interactive security-agent behavior (Wu et al., 2026; Begimher et al., 2026; Chona et al., 2026; Wang et al., 2026; Engelberg et al., 2026). CyberSecEval 2 and Cybench cover broader cybersecurity capabilities, while AgentCyberRange and DETER use contained execution environments (Bhatt et al., 2024; Zhang et al., 2025; Liu et al., 2026a; Benzel et al., 2007). SeekOrStop uses controlled execution to establish Gold and public evidence, then freezes the environment for repeatable evaluation. Its emphasis is the combination of matched evidence access, execution-derived Gold, trajectory-level evaluation, and controlled pairs within the same decision problems.

3 The SeekOrStop Benchmark

3.1 Episodes and Matched Evidence Access

An episode is a security decision problem with a frozen public evidence store $\mathcal{R}$, an initial observation O_0, and shared public decision rules. Evaluator-only references comprise Gold G, an acquisition contract $\mathcal{C}$, and controlled-pair metadata. The agent receives public context and evidence, without Gold, reference routes, family labels, or pair membership.

Passive receives a deterministic registered reference evidence view $E^\star$ and can submit a decision. **Active** receives O_0 and may retrieve additional evidence before submitting. The two conditions

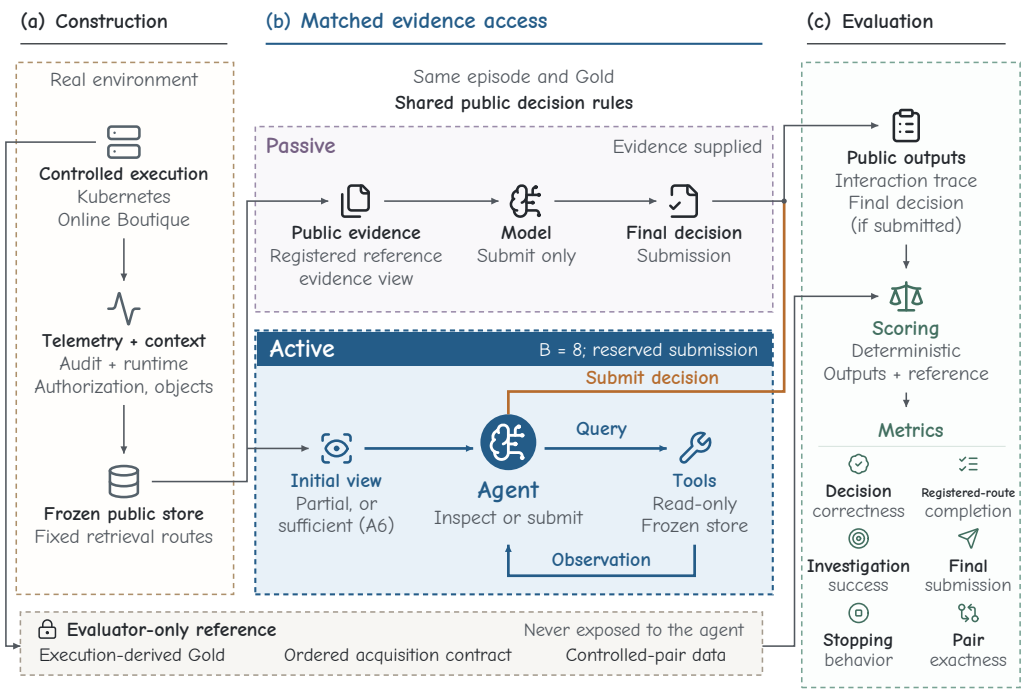


Figure 1: SeekOrStop construction and evaluation under matched evidence access. Controlled execution establishes Gold independently of sensor labels and produces public evidence. Passive receives a frozen reference view; Active retrieves evidence through read-only tools and may submit immediately when its initial view is sufficient. Public outputs are scored against evaluator-only Gold, registered acquisition contracts, and pair metadata.

share the episode, Gold, common instruction prefix, decision rules, output ceiling, and terminal correctness criterion. Evidence presentation and available actions differ: Passive supplies the registered reference evidence upfront, whereas Active exposes four read-only retrieval tools. For A6 episodes, O_0 already satisfies the decision requirements and no retrieval is required.

The Passive view combines O_0 with the public observations returned by the episode's registered reference procedure. These bundles are frozen before model evaluation and shared across configurations, providing a common supplied-evidence reference for each episode. An alignment audit across all 420 episodes verified record identities, values, and ordering between the Passive view and the corresponding registered Active evidence sequence. The condition-specific message envelopes differ; details appear in Appendix A.5.

3.2 Construction and the Observation Boundary

Harmless, bounded actions are executed in an owned Kubernetes/Online Boutique deployment under pre-specified scenario specifications frozen before Expansion evaluation. The specification and verified execution outcomes determine Gold. Kubernetes Audit, Falco runtime telemetry, authorization context, and object context separately supply observable evidence. A detector label may describe an observation but does not determine the answer.

Normalized public records and predefined retrieval routes are frozen before model evaluation. Agents query this finite store, not the live deployment: retrieval changes their observations, not the underlying system. Private execution lineage, Gold, reference routes, and pair metadata are excluded from the model-visible payload. This separation allows execution-derived outcomes and model-visible sensor evidence to remain independently defined.

Validation checks schemas, references, temporal scope, Gold derivation, pair relations, registered minimality, required-route reachability, replay determinism, environment assignment, public/private

separation, and cross-episode contamination before sealing the SeekOrStop dataset. Together, these checks establish answerability, deterministic replay, and the registered evidence-access boundary. Construction and validation details appear in Appendices A.1 and A.4.

3.3 Interaction and Acquisition

Active exposes four read-only retrieval functions: `query_audit_events`, `query_runtime_events`, `inspect_authorization`, and `inspect_target_context`, together with the terminal function `submit_decision`. Requests are restricted to public episode entities and frozen routes. The runner allows $B = 8$ budget-consuming investigation or invalid turns, with a reserved submission opportunity. The numerical budget is enforced by the runner but is not stated in the initial common instruction. Valid empty retrievals consume budget; an additional non-submit action after exhaustion is denied and terminates the session. Submission ends the session immediately. Exact turn handling and action accounting are specified in Appendix B.1.

Let E_t denote the public records accumulated after t nonterminal action turns, with $E_0 = O_0$. For response o_t,

$$E_t = E_{t-1} \cup \text{records}(o_t),$$

where errors and empty results add no records. Empty-result observation identifiers are retained separately in the observed evidence ledger. The recorded trajectory includes actions, responses, and either a committed decision D or $D = \perp$ if no valid final exists.

For Expansion420, $\mathcal{C} = (r_1, \ldots, r_m)$ specifies an ordered registered sequence of tool-and-canonical-argument routes. A successful retrieval advances the match only when it equals the next required route; intervening calls do not advance the match. Writing $k(\tau)$ for the matched prefix length, the frozen scores are

$$S_o(\tau) = \mathbf{1}[k(\tau) = m], \qquad \text{Success}_o(\tau) = \mathbf{1}[D \neq \perp \wedge D = G] S_o(\tau).$$

Here S_o measures **registered-route completion**. It records completion of the episode's registered reference workflow, while terminal correctness is scored independently. This separation allows the benchmark to distinguish completion of the registered investigation procedure from the quality of the resulting decision. Registered paths contain zero, one, or two routes (Appendix A.2).

Registered-route completion, post-completion retrieval, Critical Recall, and efficiency provide complementary views of the trajectory. In particular, a correct final need not coincide with route completion, and a route-complete trajectory need not end in a correct final. Appendix B.4 defines the measures and distinguishes the Expansion contract from historical Pilot scoring.

3.4 Illustrative Investigation

Initial view. Public scope, authorization, and runtime records are available, but a matching API attempt by the relevant actor is not yet visible.

Retrieval. A `query_audit_events` request specifies the public subject, target, category, and episode window.

Observation and submission. A matching Audit record completes the registered acquisition requirement. The agent applies the public decision rules, cites observed identifiers, and submits without further retrieval.

This schematic illustrates a registered evidence transition, not a verbatim model trajectory or an observation of hidden reasoning. The separate frozen A4 case in Appendix C.4 shows how four configurations actually behaved on a two-route episode.

3.5 Controlled Pairs and Dataset Composition

A **Contrast** pair changes one registered decision-relevant factor while holding nuisance conditions fixed, producing different Gold decisions. An **Equivalence** pair varies an irrelevant factor while preserving Gold. Exactness requires both members to be correct, testing sensitivity to relevant changes and invariance to irrelevant ones (Appendices C.2 and B.5).

Expansion420 contains 420 episodes in 210 controlled pairs, with 70 episodes in each of six families: runtime matching (A1), authorization filtering (A2), gap resolution (A3), multi-stage precedence (A4), distractor disambiguation (A5), and already-sufficient stopping (A6). Four environment

profiles vary candidate-target count and background activity, providing controlled within-system variation (Appendix A.3). Pilot60 informed the Expansion420 design without episode reuse and is reported separately; complete cohort composition and replication details appear in Appendix C.5.

4 Evaluation and Results

Experimental setup. We evaluate GPT-5.5, Kimi-K3, GLM-5.3, and Qwen3.6-27B on Expansion420, with one Active and one Passive session per episode and configuration: 3,360 primary sessions in total. Each comparison uses the same episode and decision rules, with no cross-session memory. Active follows the $B = 8$ interaction protocol; Passive receives the registered reference evidence and submits. Temperature is 0, top-$p = 1$, and the response ceiling is 4,096 tokens. Provider settings are fixed within each configuration; the complete registry is in Appendix B.3.

We ask four questions: whether supplied-evidence performance transfers to tool-mediated investigation; how registered-route completion, decision correctness, and submission relate within Active trajectories; whether agents commit when additional retrieval is unnecessary; and whether decisions remain correct under controlled relevant and irrelevant changes. Table 2 summarizes these outcomes. Accuracy includes all 420 episodes per condition, counting missing finals and provider failures in the denominator. The 95% intervals use 10,000 pair-cluster bootstrap draws within family-by-environment strata, preserving both pair members and evidence conditions (Appendices B.5 and B).

Table 2: **Expansion420 primary results**, 420 episodes per condition. Panel A reports all-session accuracy, frozen 95% pair-cluster intervals, and submission-conditional accuracy. The balanced-label reference is 25%. Panel B reports registered-route completion, route-based Success, stopping, and pair outcomes for 420 Active trajectories per model.

A. Decision quality and completion

Configuration	Passive accuracy (%)	Active accuracy (%)	Active − Passive (percentage points)	Active finals n; correct (%)
GPT-5.5	75.0 [71.7, 78.3]	61.0 [56.9, 65.0]	−14.0 [−16.9, −11.2]	420; 61.0
Kimi-K3	64.5 [60.5, 68.6]	35.0 [31.0, 39.0]	−29.5 [−34.3, −24.8]	294; 50.0
GLM-5.3	46.9 [42.9, 51.2]	24.8 [21.2, 28.6]	−22.1 [−27.4, −16.9]	156; 66.7
Qwen3.6-27B	50.7 [46.4, 55.0]	18.6 [15.2, 22.1]	−32.1 [−37.4, −26.9]	174; 44.8

B. Active investigation and controlled-pair outcomes

Configuration	Route compl. (%)	Success (%)	Recall (%)	Mean attempts	StopEff (%)	Contrast exact	Equiv. exact
GPT-5.5	47.1	39.0	53.6	1.06	73.7	33/120	56/90
Kimi-K3	41.0	20.2	47.1	4.92	15.7	13/120	21/90
GLM-5.3	51.9	15.5	61.9	5.32	6.4	5/120	8/90
Qwen3.6-27B	43.3	11.9	51.9	6.02	6.0	2/120	6/90

Active finals: valid submissions out of 420; correctness within that model-selected subset, not a replacement score. The 25% reference measures label choice, not citation-valid tool use. Route compl.: ordered registered-route completion. Success: route-complete and correct. Recall and StopEff retain ordered scoring. Other episode rates include missing finals and provider failures.

4.1 RQ1: Supplied Evidence vs. Tool-Mediated Investigation

Supplied-evidence performance does not fully transfer to tool-mediated investigation. All four configurations achieve lower end-to-end accuracy in Active than in Passive. The Active-minus-Passive differences are −14.0, −29.5, −22.1, and −32.1 points in panel order, with all four paired intervals below zero. GPT-5.5 declines from 75.0% to 61.0%; the other Active accuracies range

from 18.6% to 35.0%. Because each comparison uses the same episodes, Gold, and decision rules, these differences characterize the change between matched evidence-access conditions rather than a change in the underlying decision problems. The matched condition difference encompasses evidence selection, interaction, integration, and commitment; it does not isolate a single retrieval mechanism. Passive remains a supplied-evidence reference requiring evidence interpretation and valid submission.

The aggregate gap varies across scenario mechanisms (Table 6). For GPT-5.5, the largest differences occur in gap resolution (A3) and distractor disambiguation (A5), while multi-stage precedence and already-sufficient stopping have equal accuracy across conditions. Other configurations show broader family-level differences. The breakdown locates where supplied-evidence performance transfers least well to tool-mediated investigation.

The four labels each occur 105 times, giving a fixed-label reference of 25%. GLM and Qwen fall near or below this reference in all-session Active accuracy, but their totals include 264 and 246 missing finals. Among submitted answers, they are correct in 104/156 (66.7%) and 78/174 (44.8%) cases. Thus the end-to-end result combines decision quality with successful completion of the interaction protocol. Submission-conditional scores describe model-selected subsets and do not replace the all-session measure. Together, the two views motivate separating final judgment from interaction and commitment (Appendix C.6).

4.2 RQ2: Registered-Route Completion vs. Final Outcomes

Registered-route completion, decision correctness, and submission are distinct trajectory outcomes. Figure 2 crosses registered-route completion with final outcomes. Of GLM's 218 route-complete trajectories, 65 end correctly, 20 wrongly, and 133 without a final. GPT's 198 route-complete trajectories comprise 164 correct and 34 wrong finals, with none missing. Thus completing the registered investigation procedure does not by itself determine whether the episode ends successfully.

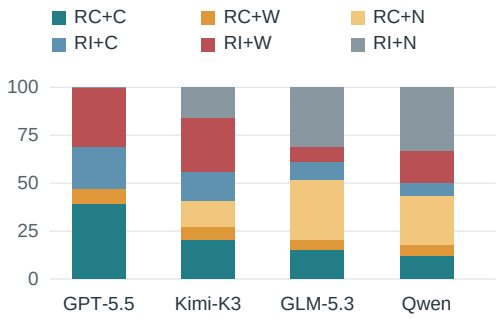


Figure 2: Active outcomes (percent; $n = 420$). RC: Route-complete; RI: Route-incomplete under the ordered registered contract. C/W/N: correct, wrong, or no final.

The converse also occurs: GPT, Kimi, GLM, and Qwen produce 92, 62, 39, and 28 correct finals without full registered-route credit, respectively. These finals retain their correctness credit. The resulting decomposition is deliberate: the route score records fulfillment of the registered reference workflow, whereas terminal accuracy records the submitted decision. Reporting both reveals trajectory differences that either measure alone would hide (Appendix C.1).

4.3 RQ3: Stopping and Protocol Completion

Correctness alone does not capture whether an agent commits once the registered decision requirements have been met. These stopping measures are procedural rather than utility-optimal: SeekOrStop records whether the agent continues after the registered decision requirements have been met, but assigns no external cost to an additional query. The 70 already-sufficient A6 episodes provide the cleanest setting for this comparison because the initial observation already satisfies the decision requirements. GPT answers 60/70 correctly, yet only 34/70 sessions submit without an investigation call. The corresponding zero-call counts are 11, 4, and 4 for Kimi, GLM, and Qwen (Table 3). Zero-call finals may follow invalid turns; first-action submission is recorded separately in Appendix C.3.

Across all Active episodes, the stopping-failure flag provides a broader view of procedural commitment by combining missing finals with post-completion retrieval. An audit of the 636 missing Active finals finds 486 sessions ending with a query at the reserved final slot, 147 with an invalid last action,

and three with provider failure. Earlier interface errors occur in 478 of the 486 query-ending sessions. These event-level records distinguish unnecessary continuation, interface nonconformance, and absent commitment without collapsing them into terminal accuracy alone (Appendix C.7).

A complementary same-episode diagnostic removes sessions with recorded native-action or Gateway errors in either condition. The Active–Passive gap remains -14.5 points for GPT ($n = 401$), -17.1 for Kimi ($n = 111$), and -14.3 for Qwen ($n = 42$); GLM retains only one episode. These filtered subsets are sensitivity diagnostics rather than replacement estimates, but show that the observed condition difference is not confined to sessions containing recorded interface errors (Appendix C.7).

Table 3: Stopping and completion under the original denominators. Stopping failure combines missing finals and post-completion retrieval and therefore includes protocol nonconformance. A6 requires no retrieval.

	A6: 70 episodes		All Active: 420 episodes	
Configuration	Zero-call final	Correct final	Stopping failure	Exhaustion
GPT-5.5	34/70	60/70	52/420	0/420
Kimi-K3	11/70	48/70	212/420	126/420
GLM-5.3	4/70	23/70	335/420	264/420
Qwen3.6-27B	4/70	22/70	311/420	244/420

Larger-budget diagnostic. Increasing the budget to $B = 30$ eliminates exhaustion for both GLM and Qwen: all 420 episodes receive valid finals. End-to-end accuracy, however, increases by only 3.3 and 1.2 points (Table 4a). Thus additional interaction budget restores valid submission without eliminating the broader decision and workflow-completion difficulties observed in Active investigation. The runs are supplemental and do not by themselves isolate retrieval, evidence-use, or interface effects (Appendix D.4).

4.4 RQ4: Controlled Pair Outcomes

Controlled pairs reveal whether correctness is preserved across matched decision-relevant and irrelevant changes. Contrast pairs require both decisions to be correct when a registered relevant factor changes; Equivalence pairs require both to remain correct when specified irrelevant context varies. GPT achieves 33/120 Active Contrast and 56/90 Active Equivalence exact pairs. Kimi, GLM, and Qwen achieve 13 and 21, 5 and 8, and 2 and 6 exact Contrast and Equivalence pairs, respectively, under the same denominators (Table 2).

Passive yields more exact pairs for every configuration but remains below perfect paired behavior: GPT reaches 45/120 Contrast and 75/90 Equivalence pairs. Pair exactness is deliberately defined as joint correctness over a controlled pair. It complements episode accuracy by showing whether correct behavior is maintained under a specified relevant or irrelevant change (Appendix C.2).

Supporting checks. Three baseline runs on a separate 120-episode panel retain mean Active disadvantages of 10.8–32.5 points (Table 4b); repetition does not increase the independent episode count. All four descriptive intervals remain below zero, providing a repeatability check on this fixed supplemental panel.

Additional diagnostics further separate the components of Active investigation. Oracle-Retrieval supplies the registered evidence directly into tool history: GPT approaches Passive, Kimi and GLM improve over historical Active, and Qwen declines. Thus supplying the registered evidence alone does not produce a universal recovery to Passive performance. A stopping reminder increases immediate commitment on already-sufficient cases, while its effects on investigation-required episodes vary across configurations. Together with the larger-budget results, these checks reinforce the value of reporting evidence access, workflow completion, final judgment, and commitment separately. Details, noncontemporaneous comparison limits, and the local 4B calibration appear in Appendices D.3, D.3.4, D.2, and C.6.

Table 4: **Supplemental completion and repeatability.** These post-freeze diagnostics retain missing finals in accuracy and remain separate from the primary evaluation.

(a) Larger budget: 420 episodes

Configuration	B	Finals (%)	Accuracy (%)
GLM-5.3	8	37.1	24.8
GLM-5.3	30	100.0	28.1
Qwen3.6-27B	8	41.4	18.6
Qwen3.6-27B	30	100.0	19.8

(b) Three runs: 120 episodes

Configuration	Mean Active–Passive gap (pp), descriptive 95% CI
GPT-5.5	-10.8 $[-14.2, -6.9]$
Kimi-K3	-29.2 $[-34.2, -23.9]$
GLM-5.3	-29.4 $[-35.6, -23.3]$
Qwen3.6-27B	-32.5 $[-37.5, -27.8]$

(a) Both $B = 30$ runs have zero budget exhaustion. (b) Repeats share 120 episodes; intervals resample pairs. Details: Appendices D.2 and D.4.

5 Discussion and Limitations

What the comparison reveals. SeekOrStop holds the decision problem fixed while comparing a supplied-evidence reference with tool-mediated investigation. The matched comparison reveals an aggregate performance gap whose concentration varies across scenario families. Trajectory-level evaluation separates registered-route completion, correct judgment, valid submission, and commitment, while controlled pairs test joint correctness under specified relevant and irrelevant changes. Together, these views show where supplied-evidence performance transfers and where investigation breaks down.

Evaluation scope. SeekOrStop combines controlled real executions for establishing Gold and observations with a frozen read-only interface for repeatable model evaluation. Registered-route completion is defined over an ordered reference workflow, while terminal correctness is scored independently so that decision quality remains observable beyond the reference path. Stopping is operationalized as commitment once the registered decision requirements are met, rather than through an external query-cost model. Passive serves as a supplied-evidence reference that removes evidence selection while retaining evidence interpretation and final decision making.

Generalization. The current instantiation evaluates four model/provider configurations in a Kubernetes/Online Boutique environment, with E1–E4 providing controlled variation in target count and background activity. Future work can extend the same matched design to additional systems, evidence sources, alternative investigation paths, human analysts, and agent scaffolds for evidence gathering and commitment.

6 Conclusion and Future Work

SeekOrStop asks whether decision quality under supplied evidence transfers to settings where an agent must identify missing evidence, acquire it through tools, and decide when to stop. Across 420 Expansion episodes, all four evaluated configurations exhibit 14–32 percentage-point drops in Active accuracy relative to Passive under matched decision problems. The trajectory analysis further shows that acquiring the registered evidence, using it to reach the correct decision, producing a valid final, and stopping appropriately are distinct observable outcomes. Controlled Contrast and Equivalence pairs expose additional sensitivity and invariance failures that are not captured by episode-level accuracy alone.

SeekOrStop combines controlled real-environment construction, sensor-independent Gold, matched supplied-versus-acquired evidence conditions, and deterministic trajectory evaluation. Future work can build on this framework by improving evidence-gap identification and query selection, studying how agents integrate retrieved evidence and commit to final decisions, and developing stopping policies that avoid both premature commitment and unnecessary retrieval. Broader evaluation should extend to independent systems, richer evidence sources, and alternative evidence paths, while future hidden evaluations should construct new episodes rather than rely on the continued secrecy of the released benchmark. Data, evaluator references, and offline evaluation code are available through the anonymous artifact (Appendix E).

AI USE STATEMENT

Generative AI tools were used to assist with language polishing of the manuscript and with the analysis, interpretation, and presentation of experimental results.

ETHICS STATEMENT

SeekOrStop studies defensive investigation in an owned and authorized environment. Construction actions were controlled and harmless, and evaluated Agents had finite read-only tools with no remediation or attack capability. Gold was evaluator-only during evaluation. The released artifact separates public model inputs from safe evaluator references. Credentials, private operational identities, restricted raw telemetry, and construction-sensitive execution lineage are excluded.

Data and code availability. SeekOrStop v0.1.0 is available for anonymous review at https://anonymous.4open.science/r/SeekOrStop-86F7/README.md. The release includes all 480 paper-cohort episodes, matched public inputs, exact prompts and contracts, canonical trajectories, safe evaluator references, and executable offline scoring and statistical reproduction scripts. Code is licensed under MIT; data and documentation are licensed under CC BY 4.0, with third-party notices retained. Appendix E describes the verification scope and information boundary.

REFERENCES

Daniel Begimher, Cristian Leo, Jack Huang, Pat Gaw, and Bonan Zheng. Sir-bench: Evaluating investigation depth in security incident response agents. *arXiv preprint arXiv:2604.12040*, 2026. URL https://arxiv.org/abs/2604.12040.

Terry Benzel, Robert Braden, Dongho Kim, Anthony D Joseph, B Clifford Neuman, Ron Ostrenga, Stephen Schwab, and Keith Sklower. Design, deployment, and use of the deter testbed. In *DETER*, 2007. URL https://www.usenix.org/legacy/events/deter07/tech/full_papers/benzel/benzel_html/.

Manish Bhatt, Sahana Chennabasappa, Yue Li, Cyrus Nikolaidis, Daniel Song, Shengye Wan, Faizan Ahmad, Cornelius Aschermann, Yaohui Chen, Dhaval Kapil, David Molnar, Spencer Whitman, and Joshua Saxe. CyberSecEval 2: A wide-ranging cybersecurity evaluation suite for large language models. *arXiv preprint arXiv:2404.13161*, 2024. doi: 10.48550/arXiv.2404.13161. URL https://arxiv.org/abs/2404.13161.

Alankrit Chona, Igor Kozlov, and Ambuj Kumar. Cyber defense benchmark: Agentic threat hunting evaluation for LLMs in SecOps. *arXiv preprint arXiv:2604.19533*, 2026. doi: 10.48550/arXiv.2604.19533. URL https://arxiv.org/abs/2604.19533.

Gal Engelberg, Michael Arenzon, and Leon Goldberg. Open security benchmark: Towards autonomous enterprise cyber defense. *arXiv preprint arXiv:2607.27288*, 2026. doi: 10.48550/arXiv.2607.27288. URL https://arxiv.org/abs/2607.27288.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval-augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Fengyu Liu, Jiarun Dai, Yihe Fan, Wuyuao Mai, Ziao Li, Bofei Chen, Jie Zhang, Zheng Lou, Bocheng Xiang, Qiyi Zhang, Xudong Pan, Geng Hong, Yuan Zhang, and Min Yang. AgentCyberRange: Benchmarking frontier AI systems in realistic cyber ranges. *arXiv preprint arXiv:2606.14295*, 2026a. doi: 10.48550/arXiv.2606.14295. URL https://arxiv.org/abs/2606.14295.

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

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*, 2026b. doi: 10.48550/arXiv.2607.10059. URL https://arxiv.org/abs/2607.10059.

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2365. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/877b40688e330a0e2a3fc24084208dfa-Abstract-Datasets_and_Benchmarks_Track.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Jiaqi Shao, Yuxiang Lin, Munish Prasad Lohani, Yufeng Miao, and Bing Luo. Do LLM agents know how to ground, recover, and assess? evaluating epistemic competence in information-seeking agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r0L9GwlnzP.

Lehan Wang, Boli Chen, Ruixue Ding, Pengjun Xie, Jinwei Huang, Zhendong Liu, Shuo Wang, Tao Lei, Xin Ouyang, and Xiaomeng Li. SecRespond: Benchmarking AI agents for real-world post-compromise incident response. *arXiv preprint arXiv:2607.26791*, 2026. doi: 10.48550/arXiv.2607.26791. URL https://arxiv.org/abs/2607.26791.

Yiran Wu, Mauricio Velazco, Andrew Zhao, Manuel Raúl Meléndez Luján, Srisuma Movva, Yogesh K Roy, Quang Nguyen, Roberto Rodriguez, Qingyun Wu, Michael Albada, Julia Kiseleva, and Anand Mudgerikar. ExCyTIn-Bench: Evaluating LLM agents on cyber threat investigation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2507.14201.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1650. URL https://papers.nips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. InterCode: Standardizing and benchmarking interactive coding with execution feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23826–23854, 2023. doi: 10.52202/075280-1035. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4b175d846fb008d540d233c188379ff9-Paper-Datasets_and_Benchmarks.pdf.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b126cc38b8638e07bef37e7b2bb72bf-Paper-Conference.pdf.

Andy K. Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W. Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Mike Yang, Teddy Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *International Conference on Learning Representations*, pp. 25094–25243, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3e9412a9c1d93810ef3ef7825115016b-Paper-Conference.pdf.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. From passive to active reasoning: Can large language models ask the right questions under incomplete information? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78714–78758. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25e.html.

A Dataset Construction and Evidence

A.1 Construction, Ground Truth, and Public/Private Boundary

A frozen scenario specification defines controlled actions and decision-relevant conditions. Harmless execution in the owned Kubernetes environment provides verified outcomes, and the specification together with these outcomes determines Gold independently of detector labels and model predictions. Public evidence combines Kubernetes Audit, Falco runtime events, authorization context, and object context. Normalization assigns public aliases, preserves episode windows, and constructs the finite retrieval interface used during evaluation.

Visibility	Contents
Model-visible	Common instruction, decision rules, public entity context, condition-specific initial evidence, and tool responses.
Evaluator-only	Gold, execution lineage, registered acquisition requirements, reference routes, family/environment labels, pair membership and relation, private seeds, and operational identities.

Evaluation operates on the frozen representation rather than triggering new execution or querying the live deployment. Future and cross-episode records are excluded. This separation keeps execution-derived outcomes, model-visible evidence, and evaluator references distinct throughout evaluation. Evaluation-time privacy does not imply permanent non-release of safe labels and references; see Appendix E.

A.2 Family and Minimal-path Specifications

Families identify targeted decision capabilities and are not supplied as model inputs. Each contributes 70 Expansion and 10 Pilot episodes.

Family	Capability	Decision-relevant question
A1	Runtime matching	Does runtime evidence match the API target?
A2	Authorization filtering	Does effective authorization explain the activity?
A3	Gap resolution	Which missing API or runtime evidence blocks the conclusion?
A4	Multi-stage precedence	How do stage scope, authorization, and target relations affect the decision?
A5	Distractor disambiguation	Which records concern relevant entities rather than distractors?
A6	Already-sufficient stopping	Can the agent submit from sufficient initial evidence without further retrieval?

Expansion contains 70 zero-step, 140 one-step, and 210 two-step registered reference paths. All 70 A6 episodes have zero-step paths. Scenario families group episodes by their decision-relevant mechanism. Each episode has a registered acquisition contract specifying the reference tools and canonical arguments. Registered minimality is defined within this contract and finite interface. Full480 contains 80 episodes per family and retains source-cohort contracts.

A.3 Environment Qualification

Expansion profiles cross candidate-target topology with observable benign activity:

Profile	Episodes	Targets	Background
E1	108	3	Nominal
E2	108	3	Elevated
E3	108	5	Nominal
E4	96	5	Elevated

Qualification requires these factors to be represented in public observations. The profile label itself is not exposed to the agent. Together, E1–E4 provide controlled within-system variation in candidate-set size and background activity. All profiles remain within the same deployment.

A.4 Dataset Validation

Before the Expansion seal, deterministic checks cover schemas, references, temporal scope, Gold derivation, pair relations, registered minimality, required-route reachability, replay identity, environment assignment, public/private separation, and cross-episode contamination. Manifests bind public fixtures to their evaluator-only references.

MinimalOracle follows each episode's registered reference path and validates its reachability and decision readout. FixedToolOrder uses four predefined queries, one per tool class; AlwaysQueryAll uses seven predefined queries over selected public entities rather than enumerating every available route. All three validation policies reach 420/420 terminal correctness under the deterministic decision readout. Together, they verify that every Expansion episode is answerable from its frozen public evidence and that Gold, retrieval access, and evaluation logic are internally consistent. FixedToolOrder is not trajectory-perfect, and AlwaysQueryAll is not stopping-perfect; these policies serve as construction and validation controls rather than model-performance baselines (Appendix B).

A.5 Passive / Minimal-path Alignment

For all 420 sealed Expansion episodes, we reconstructed the registered reference evidence sequence as the exact Active initial observation followed by the canonical Tool Gateway responses along the episode's registered minimal path. After removing the condition-specific message envelope, this sequence matches the Passive view on canonical record set (420/420), public record identity (420/420), fields and values (420/420), and canonical ordering/serialization (420/420). Passive contains no additional decision-relevant records relative to this registered reference sequence and omits none of its required records.

The alignment establishes that Passive supplies the same registered reference evidence that Active can obtain by following the episode's reference path. The conditions nevertheless preserve their intended access distinction: Active receives evidence incrementally through tool interactions and must choose its actions, whereas Passive receives the aligned evidence together in a frozen supplied-evidence view. Thus the benchmark holds the registered evidence target fixed while changing how that evidence is accessed.

Pilot60 is reported separately because its historical artifacts permit only partial reconstruction; Expansion-specific alignment assumptions are not retrofitted to the pilot cohort.

B Evaluation Protocol and Scoring

B.1 Decision Contract and Tool Interface

Two distinct contracts. The public decision contract specifies labels, precedence, and the scope, authorization, API, runtime, and target-relation conditions used to interpret evidence. The evaluator-only acquisition contract specifies the registered reference routes used for trajectory scoring. Agents receive the decision rules but not the reference routes. Exact prompt text and contract excerpts appear in Appendix B.2.

The four decision labels are terminal classifications under Decision Contract v2 rather than runner actions. In particular, `CONTINUE_INVESTIGATION` and `REQUEST_ADDITIONAL_RUNTIME_EVIDENCE` specify the terminal episode-level classification; submitting either label terminates the session and does not execute retrieval.

Native functions. Active exposes `query_audit_events`, `query_runtime_events`, `inspect_authorization`, `inspect_target_context`, and `submit_decision`. Passive exposes only `submit_decision`. Retrieval requests are schema-validated, restricted to public episode entities, and resolved against the frozen evidence interface. A valid final contains a decision label, `supporting_evidence_ids`, and confidence in $[0, 1]$. Supporting identifiers must already appear in the observation ledger; valid empty observations also receive identifiers.

Turns and the fixed budget. Each valid action turn contains exactly one native function call. Multiple calls, mixed submission/investigation calls, malformed arguments, and missing native ac-

tions are invalid. A single valid call determines the action even when accompanied by visible text; free-form text is not converted into a final.

The runner enforces $B = 8$ budget-consuming investigation or invalid turns and at most nine total action turns, with the final opportunity reserved for submission after the eight budget-consuming slots. The numerical value of B is not included in the initial common instruction; invalid-action feedback reports the remaining investigation calls. A valid final can be submitted at any point and immediately terminates the session.

Action or response	Budget effect	Runner behavior
Valid retrieval	Consumes one	Calls the frozen store and returns an observation.
Valid empty retrieval	Consumes one	Returns no record but a valid observation identifier.
Invalid action	Consumes one if available	Records the error; does not call the store.
Valid submission	No investigation slot	Commits the final and terminates.
Non-submit after exhaustion	Denied	Records a terminal action; no retrieval occurs.

The budget constrains runner execution and is distinct from the reported investigation-attempt count.

Action accounting. One processed model response is one action turn; a tool observation is not another agent action. Frozen records separately retain invalid actions, budget consumption, Gateway-entering calls, successful observations, and investigation attempts. In Expansion reporting, *accepted tool calls* count Gateway responses with status `VALID_RESULT` or `VALID_EMPTY_RESULT`. A schema-valid request may enter the Gateway yet return a route error; it consumes budget but is excluded from this successful-call count.

The historical attempt classifier uses the canonical action when available and its legacy provider-capture field otherwise. Some invalid native envelopes therefore consume budget without being counted as investigation attempts, while an identifiable budget-denied investigation may count as an attempt without entering the Gateway. We retain these frozen accounting fields separately and distinguish zero investigation attempts from zero preceding action turns.

Terminal accounting. The runner records `COMPLETED`, `MODEL_BEHAVIOR_NO_FINAL`, `MODEL_BEHAVIOR_INVALID`, `PROVIDER_FAILURE_FINAL`, or `BLOCKED_PROTOCOL`. Provider failures remain distinct from model behavior, and missing finals are not reconstructed from free-form text.

B.2 Prompt Assembly and Public Input

Prompt assembly and transport. The evaluated configurations use a shared protocol prefix, a model-visible Decision Contract v2, and a condition-specific public evidence payload. The protocol prefix is transmitted in the first `user` message, rather than a provider `system` message. The complete Decision Contract is embedded as a JSON object under `decision_contract` in `CURRENT_PUBLIC_STATE`. Native function definitions are supplied separately through the provider's `tools` field.

Common protocol prefix — verbatim

```
You are a security investigation agent.  Use the provided native
functions to obtain additional public evidence when useful.  For each
turn, either call one investigation function or call submit_decision.
You may submit immediately without investigating.  The allowed decisions
are MAINTAIN_BASELINE, CONTINUE_INVESTIGATION,
REQUEST_ADDITIONAL_RUNTIME_EVIDENCE, and ESCALATE_MULTI_STEP_ACTIVITY.
Apply the supplied Decision Contract v2 meanings and cite only evidence
IDs already observed.  Do not propose remediation or access private
data.
```

Decision semantics. The model receives the full Decision Contract v2, including required and forbidden conditions, precedence rules, examples, and evidence expectations. The box below reproduces its four `short_definition` fields; these excerpts do not replace the full contract used in evaluation.

Decision Contract v2 — verbatim definition excerpts

`MAINTAIN_BASELINE`
No public information currently requires further security investigation.

`CONTINUE_INVESTIGATION`
Investigate complete relevant Evidence that does not yet form a confirmed multi-stage chain.

`REQUEST_ADDITIONAL_RUNTIME_EVIDENCE`
Obtain a specific missing runtime confirmation for a relevant API attempt.

`ESCALATE_MULTI_STEP_ACTIVITY`
Escalate a publicly supported, unexplained, confirmed multi-stage activity chain.

Episode-specific public input. After the common prefix, the first user message contains the marker `CURRENT_PUBLIC_STATE:` followed by compact JSON with lexicographically sorted object keys. The following box shows the field structure, not a literal Episode prompt; placeholders stand for frozen public content.

Public payload structure — schematic

Active:

1. `decision_contract`: full Decision Contract v2.
2. `initial_observation.initial_observation`: initial public records O_0.
3. `public_context.public_context`: public actors, targets, containers, scope, and available-tool metadata.

Passive:

1. `decision_contract`: full Decision Contract v2.
2. `passive_evidence_view`: O_0 and the ordered frozen MinimalOracle tool observations.
3. `passive_view_hash`: hash binding the view.
4. `public_context.public_context`: the public context.

Gold, family, environment-profile labels, critical evidence groups, registered minimal paths, pair metadata, and private construction lineage are not included in these inputs.

Public records retain their frozen array order. Active receives additional observations through `tool` messages associated with the corresponding native tool-call identifiers. Passive receives the matched registered reference evidence view in its initial message and can only submit a decision. As verified in Appendix A.5, this view contains the same registered reference evidence obtainable along the corresponding Active reference path, while presenting it through a different evidence-access condition. The public evidence does not identify which records belong to evaluator-defined critical groups.

Stopping-reminder variant. The post-freeze exploratory A6 ablation inserts the following sentence after the common prefix and before `CURRENT_PUBLIC_STATE`. The original baseline sessions are reused without rerunning them. Within each matched model–Episode comparison, the public view, tool schemas, budget, Decision Contract, model configuration, parser, and scoring definitions are preserved.

Stopping-reminder addition — verbatim

```
Stopping reminder:  Before any investigation tool call, first check
whether the evidence already available satisfies the decision
requirements.  If it does, submit the decision immediately; do not call
a tool merely because tools are available.
```

B.3 Model/API Configuration and Provenance

Configuration	Provider	Exact model ID	Endpoint class	Temp.	Top-p	Max tokens
GPT-5.5	OpenAI	gpt-5.5	OPENAI_COMPATIBLE_CHAT_COMPLETIONS	0	1	4096
Kimi-K3	Volcengine Ark Plan v3	kimi-k3	OPENAI_COMPATIBLE_CHAT_COMPLETIONS	0	1	4096
GLM-5.3	Zhipu AI	glm-5.3	OPENAI_COMPATIBLE_CHAT_COMPLETIONS	0	1	4096
Qwen3.6-27B	SiliconFlow	Qwen/Qwen3.6-27B	OPENAI_COMPATIBLE_CHAT_COMPLETIONS	0	1	4096

Configuration	Tool choice	Parallel	Timeout/deadline	Attempts	Client truncation
GPT-5.5	auto	False	180/540	3	NONE_IMPLEMENTED_HARD_PROVIDER_FAILURE_IF_REJECTED
Kimi-K3	auto	False	180/540	3	NONE_IMPLEMENTED_HARD_PROVIDER_FAILURE_IF_REJECTED
GLM-5.3	auto	False	180/540	3	NONE_IMPLEMENTED_HARD_PROVIDER_FAILURE_IF_REJECTED
Qwen3.6-27B	auto	False	180/540	3	NONE_IMPLEMENTED_HARD_PROVIDER_FAILURE_IF_REJECTED

Complete provider-specific request options and response handling are documented in the anonymous configuration documentation and frozen source profiles. Provider-specific request settings are fixed within each evaluated configuration; computation is not equalized across providers. The Oracle condition copies the primary registry; supplemental batch provenance is described in Appendix D.2.1.

All configurations use provider-native function calling with strict canonical normalization and preservation of raw response bytes. Retries are restricted to HTTP 429, HTTP 5xx, and transient network failures; model behavior itself is not retried. No client-side truncation is implemented. Provider-side context handling is not fully observable, and any provider-side truncation is therefore unknown; no context-limit protocol block was observed during evaluation.

Frozen configuration records retain the endpoint, request options, prompt, schema, adapter/parser, timeout, and retry details, with unsupported or unavailable fields marked explicitly. The fixed interaction budget is specified in Appendix B.1.

Provider reliability. Five original Qwen Expansion sessions ended in provider failure (three Active and two Passive). They remain in the primary denominators. Provider failures, invalid native actions, and missing finals are kept distinct; transport retries do not retry model behavior.

B.4 Metric Definitions

Scope. The formulas below specify Expansion420 Active scoring. Final correctness is defined for both evidence-access conditions. Pilot60 retains its historical scorer; Full480 aggregates the frozen source-cohort values without rescoring Pilot under Expansion contracts.

Registered-route completion. Let $\mathcal{C} = (r_1, \ldots, r_m)$ be the registered sequence of tool names and canonical-argument hashes, and let $(q_1, \ldots, q_n)$ be the successful retrieval sequence. Valid empty responses count; Gateway error responses do not. Starting with $k_0 = 0$, matching proceeds as

$$k_j = \begin{cases} k_{j-1} + 1, & k_{j-1} < m \text{ and } q_j = r_{k_{j-1}+1}, \\ k_{j-1}, & \text{otherwise.} \end{cases}$$

Unrelated calls may intervene, but a later required route taken early is not retroactively credited.

Registered-route completion, termed strict Evidence Sufficiency in the frozen scorer, is

$$S = \mathbf{1}[k_n = m].$$

Critical Recall is k_n/m for $m > 0$ and 1 when $m = 0$. It measures ordered-prefix progress through the registered reference workflow. Together, S and Critical Recall provide deterministic trajectory measures that are comparable across runs on the same episode.

Decision and completion. Let $F = \mathbf{1}[D \neq \perp]$ indicate a committed final and $C = \mathbf{1}[D \neq \perp \wedge D = G]$ indicate correctness. Accuracy averages C over every episode in a condition; missing finals are incorrect.

Ordered Investigation Success is

$$CS,$$

with $S = S_o$ in the main text. It therefore requires both completion of the registered reference route and a correct final, while terminal correctness remains independently scored. This decomposition allows the benchmark to distinguish reference-workflow completion from the quality of the final decision.

The ordered scoring rule was specified and frozen before the Expansion primary evaluation. The Expansion implementation applies no additional critical-tool-error veto beyond route completion and correctness. Earlier invalid actions consume budget but do not automatically invalidate a later route-complete and correct final.

Stopping and completion. If registered-route completion is reached, let $j^\star$ be the successful-call index at which $k_j = m$, with $j^\star = 0$ for a zero-step episode. Post-completion calls are

$$p = \max(0, n - j^\star).$$

If route completion is never reached, the frozen accounting defines $p = 0$; this does not make the investigation successful.

The frozen stopping-failure indicator is

$$\mathbf{1}[F = 0 \ \vee \ p > 0].$$

Stopping Efficiency is the fraction of route-complete Active sessions with a submitted final and $p = 0$; final correctness is not required. The measure therefore records whether a trajectory terminates after the registered evidence requirement has been completed, while A6 provides the direct zero-step test of commitment from already-sufficient initial evidence. These are procedural commitment measures; no external cost is assigned to an additional query.

Cost and exhaustion. Unnecessary calls under the registered route are $\max(0, n-k_n)$. The frozen per-session Tool Efficiency is

$$\text{ToolEff} = \begin{cases} 1, & m = n = 0 \text{ and } C = 1, \\ m/n, & C = S = 1 \text{ and } n > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Budget exhaustion requires eight consumed investigation-budget slots and no valid final. A valid final on the ninth action is not an exhausted session. Attempt counts, invalid actions, accepted tool calls, and budget consumption remain distinct accounting fields (Appendix B.1).

A6 and pair diagnostics. A zero-investigation-call final has $F = 1$ and zero actions counted by the frozen investigation-attempt classifier; invalid or text-only turns may precede it. First-action submission is stricter: the first action itself must be a valid final. These measures provide two complementary views of immediate commitment on already-sufficient A6 episodes.

Pair exactness requires both members to be correct. Their registered different/same Gold relation then determines the expected Contrast/Equivalence behavior; agreement alone is insufficient for Equivalence exactness.

Historical Pilot and Full480. Pilot60 uses its historical evidence-coverage criterion: completion follows coverage of a frozen minimal evidence-ID set, Critical Recall measures overlap with that set, and its accepted-call count includes Gateway-entering calls. These definitions differ from Expansion's ordered registered-route criterion and successful-response count.

Full480 preserves the original source-cohort measurements as a descriptive union rather than rescoring Pilot60 under Expansion definitions. Shared metric names therefore retain their cohort-specific acquisition criteria.

B.5 Frozen Confidence Intervals and Statistical Plan

Units and resampling. Expansion contains 420 episodes in 210 pair clusters, evaluated under Active and Passive for each configuration. The bootstrap samples pair clusters with replacement within

each Family × Environment stratum, preserving its original size. Both members of a sampled cluster are retained; they are not treated as independent observations. The same sampled episodes supply Active, Passive, and Active-minus-Passive statistics within a configuration. The frozen procedure uses 10,000 replicates and seed 270060. Interval endpoints are the linearly interpolated empirical 2.5th and 97.5th percentiles. No direct cross-model difference test is specified.

Denominators and failed sessions. Accuracy retains every condition episode, including missing finals and the five provider-final failures. Stopping Efficiency conditions each replicate on route-complete Active sessions; a replicate with none receives 0 under the frozen implementation. A pre-freeze accounting correction populated missing point estimates and corrected this conditional stopping denominator without changing single-session outcomes.

Cohort and inferential scope. Expansion intervals are reused from its frozen primary analysis. Pilot retains historical family-stratified pair-cluster intervals. Full480 combines record-level counts and stored per-session measures descriptively, with no new bootstrap. Cohort percentages are not averages of subgroup percentages. The primary plan specifies no hypothesis tests, p-values, or multiple-comparison procedure. Intervals quantify uncertainty under the frozen resampling scheme; the primary analysis specifies no causal or direct cross-model inference.

Table 2 reports the frozen accuracy and paired-difference intervals; they are not repeated here.

C Additional Results and Calibration

C.1 Primary Outcome Counts

The main accuracy intervals, registered-route completion rates, and stopping totals are reported in Tables 2 and 3. Table 5 gives the integer counts behind the six-cell decomposition in Figure 2.

Table 5: Expansion420 Active outcomes, 420 episodes per configuration. Registered-route completion follows the original ordered contract. C/W/N denote correct, wrong committed, and missing final decisions.

	Route-complete			Route-incomplete		
Configuration	C	W	N	C	W	N
GPT-5.5	164	34	0	92	130	0
Kimi-K3	85	28	59	62	119	67
GLM-5.3	65	20	133	39	32	131
Qwen3.6-27B	50	26	106	28	70	140

Correct finals without full registered-route credit remain terminally correct, while missing finals remain distinct from wrong committed decisions. The six cells therefore expose combinations of workflow completion, judgment, and terminal commitment that aggregate accuracy alone does not distinguish.

Accuracy by scenario family. Table 6 breaks down the original Expansion420 results by A1–A6. Each family contains 70 episodes (35 controlled pairs) per model and condition. Accuracy retains all episodes, including missing or invalid finals and provider failures. These are the frozen primary results ($B = 8$ for Active), separate from the Oracle, reminder, and $B = 30$ supplemental runs.

A1: runtime matching; A2: authorization filtering; A3: gap resolution; A4: multi-stage precedence; A5: distractor disambiguation; A6: already-sufficient stopping (Appendix A.2). The six family counts sum to the main 420-episode totals for each model and condition. These within-cohort breakdowns describe where the aggregate Active–Passive difference is concentrated across scenario mechanisms.

Table 6: **Expansion420 accuracy by scenario family.** Each cell gives accuracy in percent, followed by the number correct in parentheses; the denominator is 70 in every cell. P: Passive; A: Active.

	GPT-5.5		Kimi-K3		GLM-5.3		Qwen3.6-27B	
Family	P	A	P	A	P	A	P	A
A1	85.7 (60)	84.3 (59)	82.9 (58)	40.0 (28)	61.4 (43)	32.9 (23)	65.7 (46)	27.1 (19)
A2	78.6 (55)	71.4 (50)	64.3 (45)	21.4 (15)	47.1 (33)	18.6 (13)	47.1 (33)	7.1 (5)
A3	78.6 (55)	27.1 (19)	70.0 (49)	21.4 (15)	41.4 (29)	17.1 (12)	41.4 (29)	11.4 (8)
A4	35.7 (25)	35.7 (25)	30.0 (21)	18.6 (13)	22.9 (16)	10.0 (7)	21.4 (15)	10.0 (7)
A5	85.7 (60)	61.4 (43)	74.3 (52)	40.0 (28)	50.0 (35)	37.1 (26)	61.4 (43)	24.3 (17)
A6	85.7 (60)	85.7 (60)	65.7 (46)	68.6 (48)	58.6 (41)	32.9 (23)	67.1 (47)	31.4 (22)

C.2 Controlled-pair Results

Expansion420 contains 120 Contrast and 90 Equivalence pairs per condition. Pair exactness requires both members to be correct. Contrast therefore evaluates joint correctness when a registered decision-relevant factor changes, whereas Equivalence evaluates joint correctness when the specified nuisance factor changes while Gold is preserved.

Table 7: Exact pair outcomes on Expansion420, with explicit denominators. Contrast and Equivalence are separate controlled tests.

	Contrast		Equivalence	
Configuration	Active	Passive	Active	Passive
GPT-5.5	33/120	45/120	56/90	75/90
Kimi-K3	13/120	30/120	21/90	57/90
GLM-5.3	5/120	16/120	8/90	29/90
Qwen3.6-27B	2/120	22/120	6/90	40/90

C.3 A6 Submission Accounting

On the 70 already-sufficient A6 episodes, first-action submissions are 34, 11, 0, and 4 in panel order. The zero-investigation-call final counts in Table 3 are 34, 11, 4, and 4. GLM's difference arises because invalid or text-only turns may precede a final without any action being classified as an investigation attempt. First-action submission is therefore the stricter measure of immediate commitment, while zero-investigation-call submission captures whether the trajectory ultimately avoids retrieval.

C.4 A Matched Multi-step Case

The frozen A4 example was selected from 12 eligible episodes by a model-independent protocol before model outcomes were examined. Its registered reference path first resolves the target relation and then checks authorization for `actor-01` on `target-01`. Gold is `ESCALATE_MULTI_STEP_ACTIVITY`.

Figure 3 shows the four original Active trajectories under the same episode, initial observation, tools, decision contract, and interaction budget. None completes the registered reference route or submits the Gold decision. The trajectories instead exhibit different observable failure patterns, including premature commitment, distractor-focused retrieval, invalid-action loops, and repeated non-contributing queries. The episode identifier is `AIEP-10AB3D403645C2CCF940`.

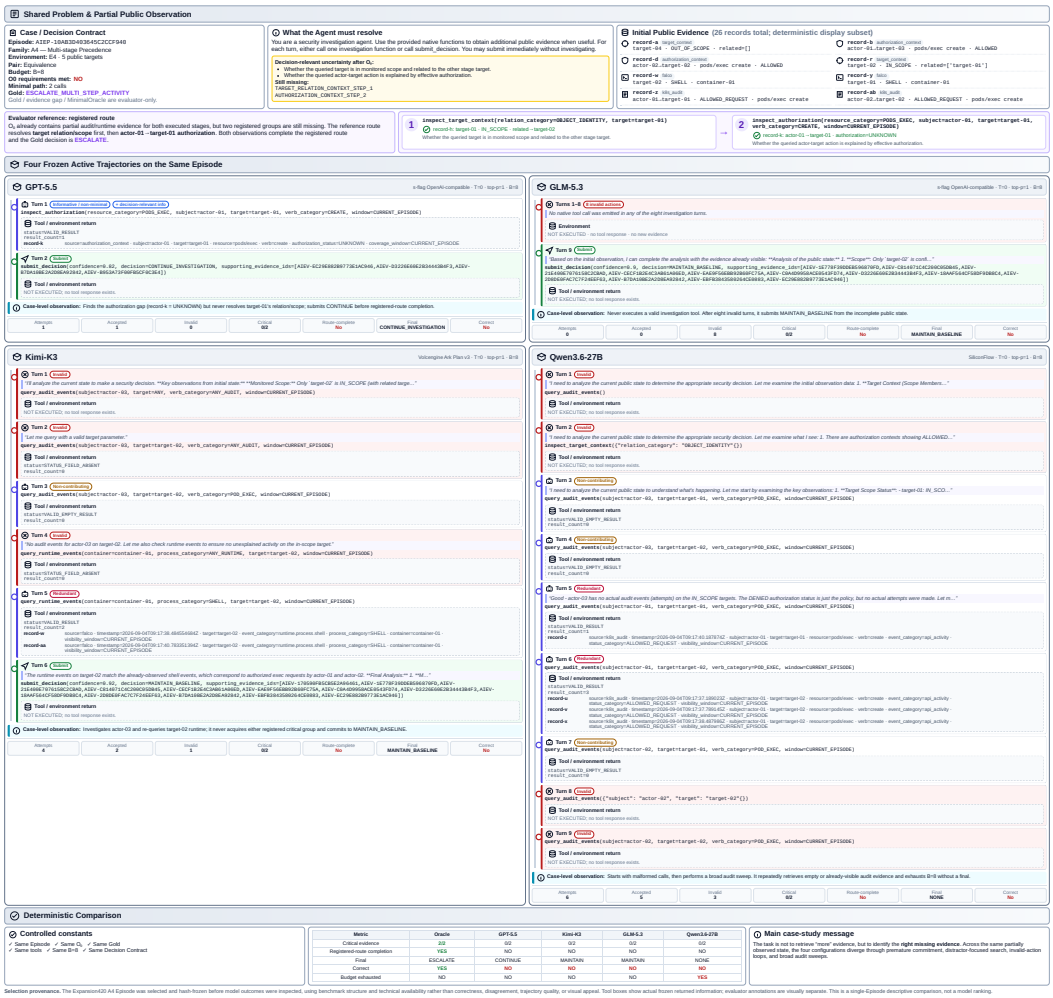


Figure 3: **The same A4 episode, four observed trajectories.** The evaluator reference and Gold were not shown to the agents. Historical "Invalid" annotations include native validation failures, Gateway errors, and Qwen's denied ninth-turn query, which does not execute.

C.5 Historical Cohort and Descriptive Union

Table 8: SeekOrStop cohort composition. Full480 combines the frozen source cohorts for descriptive reporting.

Cohort	Episodes	Pairs	Role	Environment
Pilot60	60	30	Historical	Legacy
Expansion420	420	210	Primary	E1–E4
Full480	480	240	Descriptive	Source cohorts

Pilot60 preceded the Expansion design and informed its subsequent frozen protocol and analysis plan. The fresh Expansion cohort reproduces the five registered directional patterns, although their magnitudes differ. Full480 combines frozen source records rather than reruns or averaged cohort percentages, and preserves the historical cohort-specific acquisition definitions (Appendix B.4).

Table 9: Full480 descriptive summary (%). Original source-cohort scoring and cohort-specific acquisition criteria are preserved.

Configuration	Active	Passive	Δ pp	Ordered Success	Route compl.
GPT-5.5	64.6	78.1	-13.5	43.1	50.2
Kimi-K3	36.2	68.1	-31.9	21.2	44.6
GLM-5.3	24.2	53.1	-29.0	16.0	56.9
Qwen3.6-27B	22.7	54.6	-31.9	15.8	48.3

Complete per-session and pair records remain in the released artifact; the manuscript does not repeat the full family, environment, and historical decomposition tables.

C.6 Baseline Definitions and Difficulty Calibration

Label reference. Expansion420 contains 105 episodes per decision label; the supplemental subset contains 30 per label. Every global fixed-label policy, including always selecting `CONTINUE_INVESTIGATION`, has 25% label accuracy. The global majority policy is tied, and uniform random prediction has the same expected accuracy. These references concern terminal label choice only and do not execute the citation, submission, or registered-route requirements. No private family labels are used to define a family-specific prediction policy.

Reported local baseline. We report Qwen3.5-4B on the same 120-episode supplemental subset with the baseline prompt, using one Active and one Passive run per episode (240 sessions). It uses local bfloat16 checkpoints and a Transformers native-tool service, with temperature zero, top-$p = 1$, a 4,096-token response ceiling, and the same interaction and decision contracts. Non-evaluation qualification tests cover native function calling, continuation from a tool result, and long synthetic input. The local transport timeout is fixed at 900 seconds with one attempt before execution; checkpoint and serving-source hashes remain in the records.

Table 10: Qwen3.5-4B baseline decomposition, 120 episodes per row. Missing and invalid finals remain in full-denominator accuracy.

View	Correct	Wrong	No final	Accuracy (%)	Submission (%)	Conditional (%)
Passive	28	75	17	23.3	85.8	27.2
Active	11	39	70	9.2	41.7	22.0

There are no provider-final failures or protocol blocks. Active has 62 invalid-behavior terminals and eight no-final terminals; Passive has 17 invalid terminals. Active trajectories contain 87 invalid actions requesting multiple tools in one turn, contrary to the single-action contract. Even among its 50 valid Active submissions, 11 are correct. The baseline therefore provides a local calibration point in which both protocol execution and final decision quality are challenging.

Registered-route completion is 22/120 and corresponding Investigation Success is 2/120. Twenty episodes are zero-step A6 cases whose initial observation already satisfies the decision requirements. Among the 100 investigation-required episodes, two complete the registered routes and none achieve Investigation Success.

Submission decomposition of the primary results. Let F indicate a valid final and C a correct one. Because correctness requires submission,

$$\Pr(C) = \Pr(F)\Pr(C \mid F).$$

This accounting identity separates the contribution of submission frequency from correctness among submitted answers. Because $\Pr(C \mid F)$ is evaluated on each model's selected submitted subset, all-session accuracy remains the primary end-to-end measure.

Table 11: Original Expansion420 completion decomposition, 420 episodes per row. C/W/N denote correct, wrong committed, and missing finals. Conditional accuracy uses only that configuration's valid finals.

Configuration	View	C	W	N	Accuracy (%)	Submission (%)	Conditional (%)
GPT-5.5	Active	256	164	0	61.0	100.0	61.0
Kimi-K3	Active	147	147	126	35.0	70.0	50.0
GLM-5.3	Active	104	52	264	24.8	37.1	66.7
Qwen3.6-27B	Active	78	96	246	18.6	41.4	44.8
GPT-5.5	Passive	315	105	0	75.0	100.0	75.0
Kimi-K3	Passive	271	147	2	64.5	99.5	64.8
GLM-5.3	Passive	197	39	184	46.9	56.2	83.5
Qwen3.6-27B	Passive	213	201	6	50.7	98.6	51.4

Missing finals account for 264/316 (83.5%) of GLM's Active errors and 246/342 (71.9%) of Qwen's. Their submission-conditional accuracies are 66.7% and 44.8%, respectively, showing why the all-session outcome should be decomposed into submission and conditional correctness rather than interpreted as label judgment alone. The submitted subsets are model-selected and need not retain the whole-cohort label balance.

GLM's missing Active finals comprise 241 invalid-behavior and 23 no-final terminals; Qwen's comprise 190 invalid-behavior, 53 no-final, and three provider-failure terminals; Kimi's comprise 71 invalid-behavior and 55 no-final terminals. Passive also contains 184 missing GLM finals and six missing Qwen finals, including two provider failures. These counts motivate reporting terminal completion separately in both evidence-access conditions.

C.7 Primary Interface and Terminal-event Diagnostics

Audit scope. We inspect all 3,360 original Expansion420 sessions and 14,312 actions, independently reconstructing budget transitions and replaying 6,791 Gateway requests against the frozen public fixtures. All 9,809 recorded-valid canonical actions pass argument checks, and all 2,532 recorded finals are accepted. Evidence ledgers, correctness, and submission totals reproduce the frozen records. No trajectory is repaired, missing answer imputed, or primary score replaced.

These checks use the released canonical records. Invalid-output classifications are retained as frozen rather than independently renormalized from provider bytes; the separate Oracle raw-response audit in Appendix D.3.4 performs the stronger raw-response validation for that supplemental condition.

Terminal events and earlier errors. Of 636 missing Active finals, 486 (76.4%) end with a valid query, 147 (23.1%) with an invalid action, and three (0.5%) with provider failure. Invalid actions occur in 2, 133, 383, and 290 primary Active sessions for GPT, Kimi, GLM, and Qwen, respectively; Gateway errors occur in 16, 279, 325, and 255. These indicators overlap.

Among the 486 query-ending sessions, 478 contain an earlier invalid native action or Gateway error. Because such events can consume budget or alter the interaction history, terminal events and earlier interface errors are recorded separately. The initial common instruction does not state the numerical

Table 12: Mutually exclusive primary Active terminal events; each row sums to 420. A query at the final slot passes native validation but cannot execute after the eight-slot investigation budget.

Configuration	Correct	Wrong	Final-slot query	Invalid last action	Provider
GPT-5.5	256	164	0	0	0
Kimi-K3	147	147	126	0	0
GLM-5.3	104	52	164	100	0
Qwen3.6-27B	78	96	196	47	3

value of the budget, although invalid-action feedback reports remaining slots. A final-slot query is therefore recorded as an observed terminal event rather than assigned a cognitive cause.

Same-episode filtering sensitivity. As a diagnostic, we retain an episode only when both its original Active and Passive sessions contain no recorded invalid native action or terminal provider failure (N). A stricter criterion additionally excludes Gateway errors (N+G). Native-valid requests may still return unsupported-route errors, so the two filters capture different interface events.

Table 13: Primary same-episode subset diagnostics. Correct counts share the displayed denominator; missing Active finals remain included. N/N+G are defined in the text.

Configuration	Filter	Episodes	Passive correct	Active correct	Active no final	$A - P$ (pp)
GPT-5.5	N	417	312	253	0	-14.1
Kimi-K3	N	260	166	105	47	-23.5
GLM-5.3	N	7	7	2	5	-71.4
Qwen3.6-27B	N	87	49	27	38	-25.3
GPT-5.5	N+G	401	296	238	0	-14.5
Kimi-K3	N+G	111	62	43	1	-17.1
GLM-5.3	N+G	1	1	1	0	0.0
Qwen3.6-27B	N+G	42	24	18	7	-14.3

GPT retains 401/420 episodes under the stricter N+G filter, with no missing Active finals and a remaining -14.5-point Active–Passive difference. Kimi and Qwen retain smaller subsets with differences in the same direction, while GLM's retained sample is too small for a meaningful comparison. Because the filters are defined from observed post-run events, these analyses are sensitivity diagnostics rather than replacements for the original matched-condition estimates.

D Supporting Experiments

These studies complement the frozen primary evaluation by probing prompt sensitivity, repeatability, evidence delivery, interface behavior, and interaction budget. They retain their original cohorts, configurations, and denominator rules and are reported separately from the primary 3,360 sessions.

D.1 Historical A6 Reminder Experiment

A separate post-freeze experiment applies the stopping reminder to all 70 original A6 episodes per configuration (280 new sessions), retaining the original baseline. Zero-investigation-call final counts change from 34 to 70, 11 to 45, 4 to 31, and 4 to 30; correct-final counts change from 60 to 60, 48 to 48, 23 to 43, and 22 to 30. Reminder first-action submissions are 70, 42, 17, and 26. The three interrupted sessions remain in their denominators. Its paired bootstrap uses 35 A6 clusters, 10,000 draws, and seed 270060.

The reminder demonstrates prompt-sensitive commitment, especially for configurations that rarely submit directly under the baseline prompt. This experiment isolates one prompt intervention on already-sufficient cases and is reported separately from the broader 120-episode study below.

D.2 Supplemental Prompt and Repeatability Study

D.2.1 Sample, Conditions, and Configuration Provenance

The supplemental study uses 120 frozen Expansion episodes in 60 complete pairs: 20 episodes per family, 30 per environment and label, and 34 Contrast plus 26 Equivalence pairs. Selection uses construction metadata and seed 14092026, without consulting previous model accuracy. The study was conducted after the primary evaluation and serves as a post-freeze supplemental panel; it does not alter the Expansion420 or Pilot results.

Each configuration has three baseline repetitions and one stopping-reminder repetition in both Active and Passive, for 960 sessions. The reminder is exactly the suffix in Appendix B.2; it is applied to both evidence conditions. All runs retain the frozen public fixtures, native tool and decision contracts, temperature zero, top-$p = 1$, a 4,096-token response ceiling, and eight investigation slots plus the reserved final opportunity. Model-behavior failures are not retried. The reminder effect is defined against baseline repeat-0; the other two baseline runs quantify repeatability.

Reported panel. The reported four-configuration panel contains 3,840 sessions.[1] Complete request settings, response-history processing, and execution accounting remain in the artifact.

Terminal accounting. All reported GPT sessions have valid finals. Kimi has 829 valid finals, 62 invalid-behavior terminals, 51 no-final terminals, and 18 provider-final failures; GLM has 845, 95, 16, and 4; Qwen has 690, 16, 247, and 7. GLM's four failures precede its configuration correction and are configuration-gate failures; they remain in the denominator and all occur in reminder Active. There are 29 provider/configuration failures in the reported panel. Valid-final counts remain distinct from correct-final counts.

An independently authorized retry batch covers 14 transport-failed Kimi sessions, yielding 13 valid and 7 correct finals. These runs do not overwrite the reported study. A separately labeled recovery view changes baseline repeat-0 Active from 40 to 41 correct and Passive from 73 to 74, leaving the gap at -27.5 points. Qualification requests and independent local-model runs are not included in the 3,840 count.

D.2.2 Accuracy and Prompt Effects

Table 14: Supplemental accuracy (%) at baseline repeat-0 and the single reminder run; each cell has 120 episodes. Gaps are Active minus Passive in percentage points. These results are separate from the primary 420-episode evaluation.

	Baseline repeat-0			Reminder		
Configuration	Passive	Active	Gap	Passive	Active	Gap
GPT-5.5	75.0	63.3	−11.7	75.0	50.0	−25.0
Kimi-K3	60.8	33.3	−27.5	64.2	42.5	−21.7
GLM-5.3	65.8	31.7	−34.2	71.7	35.0	−36.7
Qwen3.6-27B	48.3	18.3	−30.0	49.2	29.2	−20.0

Let $\Delta_b = A_b - P_b$ and $\Delta_r = A_r - P_r$, using baseline repeat-0 and reminder. We estimate both $A_r - A_b$ and $\Delta_r - \Delta_b$, where the latter distinguishes an Active change from a change in the matched evidence-access gap. We use 10,000 percentile bootstrap draws of complete pair clusters within family-by-environment strata, seed 14092026, preserving both members and all compared conditions.

The reminder produces heterogeneous effects across configurations: Kimi and Qwen improve in Active accuracy, GLM changes little, and GPT declines despite unchanged Passive accuracy. These results show that commitment behavior is prompt-sensitive and that encouraging earlier submission does not uniformly improve tool-mediated investigation.

[1] The supplemental GPT results use a complete 960-session rerun correcting an initial request-profile mismatch to match the primary evaluation, without episode-wise mixing. The correction followed the initial results; the excluded batch is archived separately in the artifact provenance.

Table 15: Reminder effects in percentage points with descriptive 95% pair-cluster intervals. The gap change uses baseline repeat-0.

Configuration	Active accuracy change	Active–Passive gap change
GPT-5.5	-13.3 $[-17.5, -9.2]$	-13.3 $[-17.5, -9.2]$
Kimi-K3	$+9.2$ $[+2.5, +15.8]$	$+5.8$ $[-4.2, +15.8]$
GLM-5.3	$+3.3$ $[-3.3, +10.0]$	-2.5 $[-11.7, +6.7]$
Qwen3.6-27B	$+10.8$ $[+3.3, +18.3]$	$+10.0$ $[-0.8, +20.8]$

D.2.3 Already-sufficient vs. Investigation-required Episodes

Table 16: Baseline repeat-0 $\rightarrow$ reminder Active behavior. Each A6 cell has 20 episodes; each A1–A5 cell has 100. First-submit means a valid final on the first action. Success is the original registered-route-completion-and-correctness criterion.

	A6: already sufficient		A1–A5: investigation required		
Configuration	First-submit	Correct	First-submit	Correct	Success
GPT-5.5	$10 \rightarrow 20$	$18 \rightarrow 18$	$29 \rightarrow 58$	$58 \rightarrow 42$	$27 \rightarrow 4$
Kimi-K3	$2 \rightarrow 13$	$12 \rightarrow 16$	$15 \rightarrow 52$	$28 \rightarrow 35$	$11 \rightarrow 4$
GLM-5.3	$3 \rightarrow 5$	$4 \rightarrow 8$	$11 \rightarrow 33$	$34 \rightarrow 34$	$11 \rightarrow 12$
Qwen3.6-27B	$1 \rightarrow 7$	$7 \rightarrow 12$	$1 \rightarrow 20$	$15 \rightarrow 23$	$8 \rightarrow 8$

The family split clarifies the reminder's effect. For GPT, all 16 lost correct answers occur in A1–A5 while A6 correctness is unchanged. Across all 120 Active episodes, its first-action submissions rise from 39 to 78, whereas registered-route completion falls from 53 to 24. Kimi likewise combines more correct finals with lower route-based Success on investigation-required episodes. The intervention therefore improves immediate commitment in many already-sufficient cases but can also induce premature commitment when further investigation is required.

D.2.4 Repeated Baseline Runs

Table 17: Three baseline accuracy repetitions (%), using 120 distinct episodes per configuration. Mean gaps and descriptive pair-cluster intervals appear in Table 4b.

Configuration	Active: runs 1/2/3	Passive: runs 1/2/3
GPT-5.5	63.3/64.2/65.0	75.0/75.0/75.0
Kimi-K3	33.3/31.7/36.7	60.8/66.7/61.7
GLM-5.3	31.7/37.5/39.2	65.8/70.0/60.8
Qwen3.6-27B	18.3/17.5/12.5	48.3/50.8/46.7

Across the three baseline repetitions, the Active–Passive direction is preserved for all four configurations, while absolute labels and completion behavior vary more for Kimi, GLM, and Qwen than for GPT.

Table 18: Active repetition diagnostics (counts out of 120). Agreement requires a valid final in all three runs; three missing finals are reported separately.

Configuration	All three valid	Same valid label	All three missing	Any missing
GPT-5.5	120	114	0	0
Kimi-K3	50	28	8	70
GLM-5.3	57	24	5	63
Qwen3.6-27B	3	2	29	117

Passive same-valid-label counts are 120, 66, 47, and 53 in panel order. These repetitions provide a fixed-sample view of generation stability in addition to the primary pair-cluster uncertainty analysis.

D.3 Oracle-Retrieval and Submission Audit

D.3.1 Active-Fully-Informed Condition

We add one post-freeze Oracle-Retrieval session for every configuration and Expansion420 episode (1,680 sessions). Starting from the original Active initial observation, an evaluator-side oracle executes the registered minimal route and inserts its exact public results as assistant native-tool-call and tool-response history. It supplies no Gold, explanation, or announcement that the evidence requirement has been completed. The evaluated LLM then chooses whether to submit or continue investigating. The oracle supplies retrieval actions, never the final decision.

The common instruction, decision contract, native schemas, provider registry, response ceiling, normalization, and eight-investigation-slot plus reserved-final budget are retained. Oracle calls consume the investigation budget: a path of length m leaves $8 - m$ such slots and the reserved submission opportunity. All Active tools remain available. Oracle actions and model actions are recorded separately. A6 has $m = 0$, so its first request is unchanged from historical Active; all 280 paired A6 request bodies were verified byte-identical.

Offline checks reconstructed all 1,680 histories; provider tests and 12 formal canaries covered zero-, one-, and two-route histories. The canaries remain in the final denominator, with no outcome replacement.

Comparisons and uncertainty. Historical Passive and Active are paired by episode with Oracle. Because Oracle is a post-freeze run, these comparisons are descriptive rather than a contemporaneous intervention on the primary sessions. All 420 episodes remain in each accuracy denominator. Intervals use 10,000 pair-cluster draws within family-by-environment strata, seed 18092026.

Table 19: Oracle-Retrieval compared with historical matched conditions. Accuracy rates are percentages over 420 episodes. Differences are Oracle minus the reference, in points with paired 95% intervals.

Configuration	Passive	Oracle	Active	Oracle−Passive	Oracle−Active
GPT-5.5	75.0	74.3	61.0	-0.7 $[-1.7, 0.0]$	$+13.3$ $[10.5, 16.4]$
Kimi-K3	64.5	52.4	35.0	-12.1 $[-17.4, -6.7]$	$+17.4$ $[12.1, 22.6]$
GLM-5.3	46.9	37.1	24.8	-9.8 $[-15.0, -4.3]$	$+12.4$ $[7.1, 17.6]$
Qwen3.6-27B	50.7	11.2	18.6	-39.5 $[-44.3, -34.8]$	-7.4 $[-11.7, -3.1]$

GPT's Oracle accuracy approaches Passive, while Kimi and GLM improve substantially over Active but retain gaps to Passive. Qwen instead declines below both historical conditions. Supplying the registered evidence through tool history therefore does not universally recover Passive performance: evidence delivery, correct judgment, submission, and commitment remain distinct outcomes.

D.3.2 Investigation-required and Zero-step Episodes

Table 20: Investigation-required A1–A5 episodes ($n = 350$, 175 pairs). Correct counts retain missing finals; intervals are paired 95%. The zero-step A6 comparison is reported separately.

Configuration	Passive correct	Oracle correct	Active correct	Oracle−Active (pp)
GPT-5.5	255/350	253/350	196/350	$+16.3$ $[12.9, 19.7]$
Kimi-K3	225/350	177/350	99/350	$+22.3$ $[16.6, 28.3]$
GLM-5.3	156/350	121/350	81/350	$+11.4$ $[5.7, 17.4]$
Qwen3.6-27B	166/350	29/350	56/350	-7.7 $[-12.3, -3.4]$

On A6 ($n = 70$, 35 pairs), historical Active to zero-step Oracle correct counts change from 60 to 59, 48 to 43, 23 to 35, and 22 to 18. Oracle-minus-Active intervals are $[-4.3, 0.0]$, $[-18.6, 4.3]$, $[2.9, 30.0]$, and $[-20.0, 8.6]$ points. Because A6 contains no oracle retrieval, these shifts provide an empirical reminder that the Oracle runs also differ in execution time from the historical primary sessions. We therefore use the Oracle condition as a diagnostic of evidence delivery rather than a corrected estimate of pure acquisition cost.

D.3.3 COMMITMENT AND PROVIDER-RECOVERY SENSITIVITY

Table 21: Oracle behavior, counts out of 420. First submit is a valid final on the first model action after prefill. Any query counts an executed Gateway request, including error responses; the columns are not complementary.

Configuration	Correct final	Valid final	First submit	Any query
GPT-5.5	312	420	287	131
Kimi-K3	220	309	111	305
GLM-5.3	156	245	14	405
Qwen3.6-27B	47	92	8	409

Even after the registered evidence has been inserted into history, continued querying remains common for Kimi, GLM, and Qwen. This separates evidence delivery from subsequent commitment behavior.

Eight Qwen sessions ended in provider failure. A fixed-selection recovery run re-executed all and only these eight using the same episode, profile, prompt, prefilled visible call identifiers, and budget. All eight recovered provider responses; two submitted correctly and six still exhausted their budgets without a final. Replacing only these eight in a labeled sensitivity view changes Qwen correctness from 47/420 to 49/420 (11.7%) and valid submissions from 92/420 to 94/420; its Oracle-minus-Active difference remains -6.9 points $[-11.4, -2.4]$.

D.3.4 RAW-RESPONSE, BUDGET, AND SUBMISSION AUDIT

We audit all 1,680 original Oracle sessions and eight recovery sessions, covering 8,231 model actions. An independent native-action/schema validator reconstructs actions from retained raw response bytes; budget transitions are independently recomputed. Frozen Gateway replay checks returned records, observed evidence identifiers, and submission ledgers, and all initial and subsequent request histories are reconstructed.

All successful-response model IDs and retained request/response hashes match. No contract-valid native submission was rejected, and all 1,068 valid submissions, including two retests, were accepted. Oracle budget charges, invalid-action charges, and the reserved ninth action reproduce the recorded state machine. These audits establish implementation fidelity for the Oracle condition.

Table 22: Terminal events among original Oracle sessions without a valid final. "Query at final slot" records an investigation request when only the reserved submission opportunity remains.

Configuration	No final	Query at final slot	Invalid last action	Provider	Valid submit wrongly rejected
GPT-5.5	0	0	0	0	0
Kimi-K3	111	109	2	0	0
GLM-5.3	175	167	8	0	0
Qwen3.6-27B	328	320	0	8	0
Total	614	596	10	8	0

Within the Oracle cohort, 596/614 missing finals end with a query at the reserved final slot. Of the 104 original `MODEL_BEHAVIOR_INVALID` terminal records, 94 nevertheless end with a valid investigation call and only 10 with an invalid last action. Terminal status and earlier interface behavior are therefore reported separately.

Gateway errors here are `AGENT_TOOL_SCHEMA_ERROR`: the native arguments pass validation, but the combination has no frozen route. They are distinct from parser rejection and provider failure. The audit distinguishes Gateway-entering requests from successful evidence returns without changing the frozen primary metrics.

Concrete failure patterns. All 22 single-call invalid submissions cite at least one unobserved evidence ID. One Kimi session attempts submission seven times using the same unseen reference,

Table 23: Overlapping interface-error indicators among Oracle sessions.

Configuration	No-final sessions with invalid actions	No-final sessions with route errors	All sessions with invalid actions
Kimi-K3	14/111	104/111	37/420
GLM-5.3	86/175	173/175	242/420
Qwen3.6-27B	4/328	101/328	6/420

indicating repeated commitment attempts that fail the citation contract. GLM produces 236 multi-call outputs against the one-call-per-turn contract. Its 13 empty outputs and two text-only outputs all have `finish_reason=length`; four of these 15 sessions end without a valid final.

These examples illustrate that evidence delivery, native-function compliance, citation validity, and terminal commitment can fail at different stages of the same protocol. Qualitative inspection covers the 26 actions without a native call and deterministically selected terminal-event examples; visible prose is never converted into a valid submission after the fact.

Interpretation. The audit finds no rejected contract-valid submission and reproduces the Oracle state machine from retained raw responses. Together with the primary interface audit, it motivates treating evidence delivery, interface execution, and terminal commitment as separately observable components of tool-mediated investigation.

D.4 Larger-budget Completion Diagnostic

We compare the frozen Active $B = 8$ results with additional $B = 30$ runs for GLM-5.3 and Qwen3.6-27B on the same 420 Expansion episodes. The cohort contains 210 controlled pairs: 120 Contrast and 90 Equivalence pairs. The supplemental comparison asks whether the low all-session accuracy observed at $B = 8$ persists once budget exhaustion and missing submission are eliminated.

Table 24: Additional budget-diagnostic measures, 420 episodes per row. Submission rates and all-session accuracy appear in Table 4a; conditional accuracy uses valid finals.

Configuration	B	Conditional accuracy (%)	Route complete	Exhausted
GLM-5.3	8	66.7	218/420	264
GLM-5.3	30	28.1	241/420	0
Qwen3.6-27B	8	44.8	182/420	244
Qwen3.6-27B	30	19.8	207/420	0

Completion and correctness. At $B = 30$, both configurations submit valid finals on all 420 episodes. Submission therefore increases by 62.9 points for GLM and 58.6 points for Qwen, while all-session accuracy increases by only 3.3 and 1.2 points. Correct-final counts change from 104 to 118 for GLM and from 78 to 83 for Qwen.

The larger budget thus eliminates exhaustion and restores valid submission on the episodes evaluated at $B = 8$, while final accuracy remains low and registered-route completion improves only modestly. This result reinforces the distinction between valid submission, registered-route completion, and decision correctness.

Conditioning and route completion. Conditional accuracy changes from 66.7% to 28.1% for GLM and from 44.8% to 19.8% for Qwen because the $B = 8$ statistic conditions on a model-selected submitted subset whereas $B = 30$ contains all episodes. It should therefore be interpreted together with the all-session correct counts rather than as a same-subset deterioration.

Registered-route completion increases from 51.9% to 57.4% for GLM and from 43.3% to 49.3% for Qwen. Thus a substantially larger action budget produces modest gains in both correct decisions and reference- workflow completion after exhaustion has been eliminated.

Reporting scope. The $B = 30$ runs are a post-freeze diagnostic rather than replacements for the frozen $B = 8$ primary evaluation. Their central result is the separation between completion and correctness: removing budget exhaustion restores submission but leaves most previously observed decision difficulty intact.

E Reproducibility and Release

Released materials and access. SeekOrStop v0.1.0 is available at https://anonymous.4open.science/r/SeekOrStop-86F7/README.md. It includes 480 public episodes, frozen Passive inputs, public decision and acquisition contracts, prompts, safe evaluator references, Gateway replay and scoring code, and 8,214 canonical sessions organized by experiment group. Repeated and supplemental sessions do not increase the number of distinct episodes. Code is released under the MIT License; data and documentation use CC BY 4.0 with third-party notices retained.

Version 0.1.0 covers the original cohorts and the supplemental studies included at the time of release. The later Oracle-Retrieval sessions in Appendix D.3 and larger-budget runs in Appendix D.4 were conducted after that artifact version and are therefore not included in its 8,214 released sessions.

Independent offline verification. The released evaluator uses the Python standard library and requires neither model API access nor the original deployment for replay and scoring. The README provides commands to validate inputs, test the public/private boundary, and execute `scripts/audit_reproduction.py`. An isolated-copy verification replayed all 13,584 registered Gateway routes across the 480 episodes and rescored all 8,214 released sessions. For the 8,200 sessions with frozen metric references, 69,800 individual metric comparisons passed; the remaining 14 transport-recovery sessions are retained and rescored separately.

The same verification reproduces the primary summaries, Full480 descriptive results, controlled-pair outcomes, primary accuracy bootstrap intervals, and supplemental prompt-gap intervals. File hashes verify artifact identity, while replay and rescoring independently reconstruct the reported results from the released episode and trajectory records. Thus the artifact supports offline verification from frozen evidence access through session scoring and aggregate evaluation.

Information boundary and release scope. During model evaluation, Gold, registered reference routes, critical evidence groups, pair relations, and construction lineage remain evaluator-only. Released evaluator references are stored separately from model-visible inputs; the prompt builder does not read them, and boundary tests verify this separation.

Canonical released trajectories exclude credentials, restricted raw telemetry, private operational metadata, and construction-sensitive execution lineage. The artifact is designed to reproduce the frozen benchmark and its evaluation logic rather than recreate the original live infrastructure. New model runs require the researcher's own provider credentials and an explicit execution option.

Because the 480 released episodes and evaluator references are public, they should be treated as a reproducible public benchmark rather than a permanently hidden test set. Future hidden evaluation can preserve the same construction and scoring framework while using newly constructed episodes.