# WHERE SCIENTIFIC SEARCH AGENTS FAIL: DECISION-CHECKPOINT AUDITING OF EXPOSURE AND INSPECTION ATTEMPTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Final-answer accuracy does not reveal whether a scientific-search agent failed to encounter a target paper, attempt to inspect it, or return an accepted answer after inspection. We introduce decision checkpoints that record observations and tool actions without benchmark labels during inference, then join target identities and evaluator labels to assign outcome categories from recorded events. Across five conditions on 540 answerable AutoResearchBench Deep questions in a fixed, target-enriched environment, keyword search achieves 24.6% accuracy, compared with 17.8% for raw search. The keyword condition has fewer incorrect answers with neither target exposure nor inspection, but more incorrect answers after the target is exposed and left uninspected. Compared with keyword search, read-first has 27.4% more recorded evidence-search calls. Target inspection attempts occur on 199 questions under read-first and 191 under keyword search; both conditions achieve 24.6% accuracy. The checkpoint protocol makes these question-level differences explicit, distinguishing target exposure and inspection from aggregate accuracy and total tool use.

## 1 INTRODUCTION

Scientific-search agents identify papers by alternating between search queries, candidate inspection and answer selection (He et al., 2025; Xiong et al., 2026). A final paper-identification score records whether the selected paper matches the benchmark target, but leaves the intervening search decisions unresolved. An agent may answer incorrectly after failing to expose the target, leaving an exposed target uninspected, or attempting to inspect the target. Distinguishing these events specifies which part of a search trajectory an intervention is intended to change (Figure 1).

Existing work examines intermediate retrieval stages and search failures. PaperQA2 tracks source retention through its evidence pipeline (Skarlinski et al., 2024), AutoResearchBench manually analyzes scientific-search errors (Xiong et al., 2026), and SearchAuditor diagnoses selected failed trajectories (Liang et al., 2026). We study a complementary measurement question: across the same benchmark questions, how do target exposure, target inspection attempts and accepted answers differ between search conditions?

We record *decision checkpoints* during execution and match search observations and executed actions to benchmark target IDs after inference. This yields exposure and inspection labels without an additional semantic judge. An inspection attempt queries the within-paper evidence-search tool, which returns matched excerpts. Final correctness is determined separately by the locally adapted benchmark evaluator. Ground truth is withheld during inference, although target identities informed corpus construction.



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

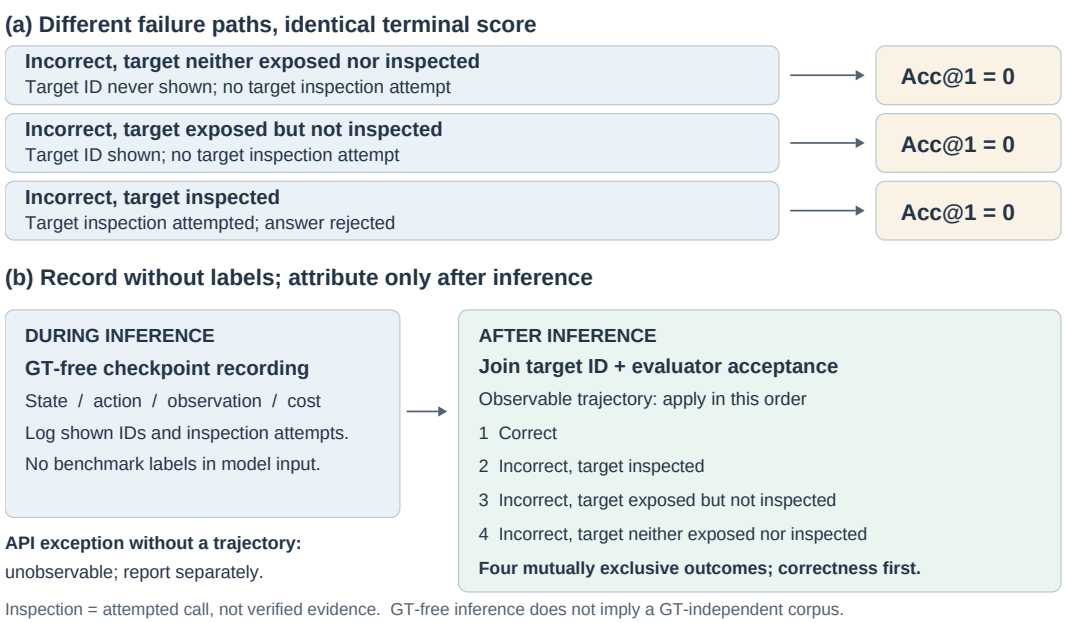


Figure 1: Conceptual overview (not empirical data). (a) Three distinct failure paths share the same terminal score. (b) Checkpoints are recorded without ground-truth (GT) labels during inference; target identities and evaluator acceptance are joined afterward. For observable trajectories, correctness takes precedence, then inspection, then exposure; API exceptions without trajectories are reported separately. Inspection is an attempted within-paper evidence-search call, not verified evidence acquisition.

We contribute a checkpoint protocol that links recorded target exposure and executed inspection attempts to evaluator-scored answers on a common set of questions. We apply it to five search conditions on 540 answerable AutoResearchBench Deep questions in a shared, target-enriched environment. The keyword condition has fewer incorrect outcomes with the target neither exposed nor inspected than raw search, but more failures after exposure without inspection. For read-first, we report the source of inspection calls—agent-initiated or forced by the gate—and the questions on which those calls reach the target. These measurements distinguish additional tool use from target inspection, while paired answer comparisons reveal gains and losses hidden by the same aggregate accuracy.

## 2 Related Work

**Scientific search and stage measurement.** AutoResearchBench defines Deep and Wide literature search over arXiv (Xiong et al., 2026). Its Appendix F manually analyzes sampled incorrect predictions, including evidence aggregation, tool execution and candidate ranking failures; the latter explicitly includes a gold paper already in the candidate set but ranked below another paper. PaSa models search/read/select decisions (He et al., 2025), LitSearch benchmarks retrieval and reranking (Ajith et al., 2024), and OpenScholar generates citation-supported answers (Asai et al., 2024). PaperQA2 tracks source-DOI recall through retrieval, chunk ranking, contextual summarization and attribution (Skarlinski et al., 2024), a direct precedent for separating discovery from downstream evidence use. We measure target exposure and executed inspection attempts across the evaluated questions, linking these events to paired answer outcomes.

**Process and evidence evaluation.** AgentBoard measures partial progress through goal-state similarity or defined subgoals (Ma et al., 2024); ToolSandbox matches temporally ordered milestones involving tool calls and environment states (Lu et al., 2025); AgentEval evaluates DAG-structured steps and error propagation (Guo et al., 2026). SearchAuditor evaluates failure localization, attribution and repair against expert annotations of selected failed trajectories, and relates exploration

to downstream failure types (Liang et al., 2026). Semantic evidence diagnostics include SciFact's rationale and support/refutation labels (Wadden et al., 2020) and RAGChecker's claim-level entailment checks for retrieval and generation (Ru et al., 2024). Our target-identity and executed-action labels provide an event-level complement to these semantic diagnoses, covering successes as well as failures.

**Retrieval access and resource allocation.** BrowseComp-Plus fixes the corpus and varies whole-document reader access and oracle retrieval with annotated positive documents (Chen et al., 2025). Cost-aware evaluation reports accuracy jointly with resource use (Kapoor et al., 2024), while budget-aware control (Fang et al., 2026) and value-guided search (Li et al., 2026) optimize allocation. Our gate changes when an already available within-paper evidence-search tool is called; the audit asks whether additional calls reach the target on additional questions.

## 3 Decision-Checkpoint Auditing

**Checkpoints (execution time, no ground truth).** A search agent interacts with a paper-search tool, a within-paper evidence-search tool taking a paper ID and query, and a final-answer action. The evaluation harness executes tool calls and records each turn's results. For a turn with a model response, it records $c_t = (s_t, a_t, o_t, \mathrm{cost}_t)$. The state $s_t$ is a harness summary of the maintained candidate list, prior queries, attempted-read IDs and retained evidence excerpts; it is not a separate message supplied to the model or an exact copy of its context. The action $a_t$ records the proposed action, replaced by the executed evidence-search call when the gate intervenes; execution status is retained separately. The observation $o_t$ records returned IDs and evidence excerpts, and $\mathrm{cost}_t$ records tokens and latency. Search results are serialized separately into tool messages; older full-text returns are generally compacted to read summaries. Ground-truth labels are absent from inference messages and checkpoints. Benchmark input metadata is attached to saved outputs afterward for evaluation.

**Retrieved, exposed, inspected.** We distinguish three events for target $y^*$: *retrieved* – $y^*$ is in the backend's top-$N$ for some query; *exposed* – its identifier appears in a search-result observation serialized for the model, with rank in that list; *inspected* – an executed within-paper evidence-search call targets $y^*$. Exposure is counted on search actions, not on identifiers repeated in later checkpoint observations or full-text responses. Inspected means that an inspection attempt was executed, including calls that return an error; it does not imply successful evidence acquisition. Neither exposure nor an inspection attempt establishes that the agent received sufficient evidence to satisfy the question's constraints.

**Outcome categories.** After execution, $y^*$ is joined to the trajectory. The evaluator is the only source of correctness. An accepted final answer is labelled *Correct*, even without prior search or inspection. Among incorrect answers, inspection takes precedence over exposure: observable trajectories are labelled *Incorrect, target inspected*; *Incorrect, target exposed but not inspected*; or *Incorrect, target neither exposed nor inspected*. An API exception without a trajectory is separately unobservable. These outcome categories are mutually exclusive, not a mandatory sequential funnel; exposure and target inspection attempts are counted separately.

**Additional event summaries.** Alongside the five-condition outcome table, we report first-exposure rank and the number of search-result lists containing the target per question. Among questions answered incorrectly after the target was exposed but never inspected, we count those with at least one legal search, evidence-search or answer action after the first exposure. A later action shows that the episode continued after exposure; it does not measure the reading budget available at that point or the counterfactual benefit of inspection.

**Retention limits.** Search exposure is cross-checked against retained serialized tool messages; subsequent legal actions exclude parse errors. Exposure means that the target appeared at some point, not that it remains in the current context. Saved final messages omit some earlier evidence-search returns, while checkpoints retain excerpts rather than complete responses. The retained records permit event checks but cannot reconstruct every model input or support exhaustive semantic verification.

## 4 EXPERIMENTAL SETTING

**Task.** AutoResearchBench Deep asks for a single target arXiv paper or "no answer". We use its complete 540-question answerable subset, excluding the 60 explicitly unsatisfiable questions in the 600-question Deep split. This does not evaluate abstention on answer-absent questions. A 77-question subset informed exploration and intervention design; it is not an independent confirmation set. Wide and cross-model evaluations are outside our scope. We evaluate five search conditions. An episode is one execution on one question under one condition; its recorded sequence of actions and observations is its trajectory.

**Backend.** BM25 (Robertson & Zaragoza, 2009) ($k_1$=1.2, $b$=0.75) over title+abstract of the arXiv metadata resource (Cornell University, n.d.) (3.16M rows in our locally indexed snapshot), fused by reciprocal-rank fusion (Cormack et al., 2009) ($c$=60) with a full-text BM25 layer over 7,703 cached PDFs: 3,853 targets from the broader benchmark corpus (not just Deep540) plus 3,850 non-target papers matched by primary category and year (+*both*). The full-text BM25 index uses at most 400,000 characters per extracted paper; this indexing limit is distinct from the complete-text evidence search below. Within-paper evidence search operates over the complete extracted paper and returns up to three query-matched evidence windows, each spanning up to eight lines on either side of a matched line and truncated to 1,800 characters. Ground-truth labels are withheld from model inputs, but target identities were used to construct the full-text layer. Because the target-enriched full-text layer also participates in rank fusion, its coverage can affect both candidate ranking and subsequent evidence access.

**Agent and policies.** We use the AutoResearchBench inference framework with adapted tool prompts. The agent chooses top-$N$ (candidate cap 200), rounds (safety cap 30, hits counted), and evidence-search queries; the harness soft context threshold is 110,000 estimated input tokens. Primary policies: RAW (prompted to copy the complete question verbatim in every search, including later searches) and KW (prompted to formulate short, distinctive keyword queries before searching); secondary: DECOMP and ADAPTIVE. RAW serves as a prompt-restricted reference condition, whereas KW permits query reformulation throughout the trajectory. Read-first retains the KW keyword-query rule and adds a reading-rule prompt plus the inspection gate described in Section 6. The harness does not enforce verbatim RAW queries (Appendix B). The suffix `-ft` denotes full-text-enabled policies using the shared +both backend. DECOMP prompts the agent to split the question into focused sub-questions, search once per sub-question, and combine the evidence to select candidates. ADAPTIVE prompts the agent to assess candidates after each search, rewrite the next query using informative terms or titles from apparently relevant candidates, and search again until the relevant evidence and candidate set are stable. Gateway request alias: `gpt-5.6-luna`. Metric: Deep Accuracy@1 (pass@1) under the evaluator. All conditions use the same retrieval backend, tool interfaces and nominal execution limits. Appendix B details the query rules. Costs (turns, searches, top-$N$, evidence-search queries, tokens, stop reason) are reported together with the operational limits; cap hits are not discarded. Throughout, *none* counts questions whose final answer contains no arXiv identifier, whether by an explicit "no answer" or because the episode ended without one.

**Evaluator and model provenance.** The evaluator compares the title of the first candidate in the final answer with the target title; no candidate or an explicit no-answer state scores zero on this answerable cohort. IDs are not the primary scoring rule (Appendix B). Historical underlying-model snapshots, effective reasoning settings and judge configurations are not fully recorded; the request alias `gpt-5.6-luna` does not establish their identity. Comparisons are therefore descriptive rather than verified same-model causal estimates. A uniform title-based re-score of all 2,700 saved answers under a recorded current judge configuration reproduces every historical correctness label. This checks label stability under that configuration, not the judge's semantic validity or historical inference provenance.

## 5 OBSERVED OUTCOME PROFILES

The keyword condition has fewer never-exposed failures and more exposed-but-uninspected failures than raw search (Figure 2, Table 1). Both conditions cover the same 540 answerable Deep questions; all outcome proportions use that 540-question denominator.

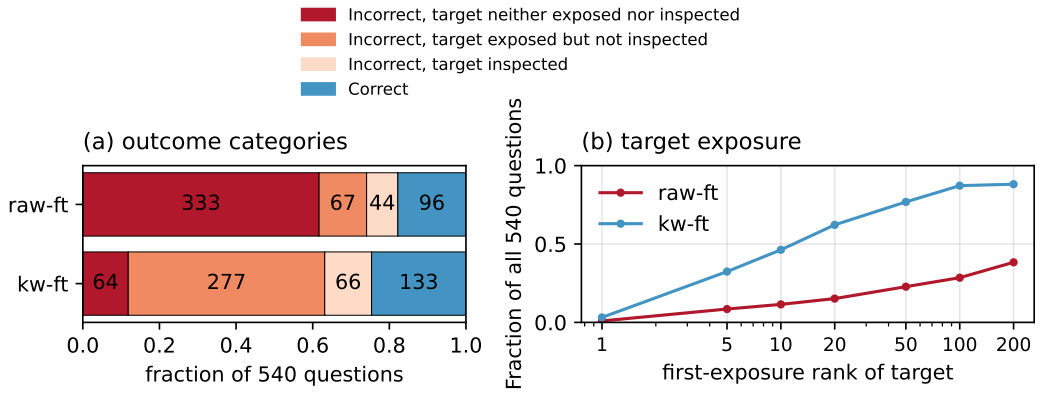


Figure 2: Decision-checkpoint audit on all 540 answerable Deep questions (+both backend, request alias gpt-5.6-luna). (a) Correct answers take precedence; remaining questions are classified by inspection and exposure. Raw and keyword conditions differ in *neither exposed nor inspected* (333→64) and *exposed, not inspected* (67→277) failures, with Acc@1 of 17.8% and 24.6%. (b) Cumulative fraction of questions whose target first appears within a given rank of an observation; under KW-FT the target is exposed within rank 20 on 336/540 questions.

Table 1: Outcome categories across five conditions on the same 540 questions. Categories are mutually exclusive; all except Correct and Unobservable are incorrect outcomes. Exposure and inspection refer to the target paper. Exposure is also counted separately.

(a) Outcome categories: counts, with accuracy in parentheses

| Condition | Correct $n$ (%) | Neither exposed nor inspected | Exposed, not inspected | Inspected | Unobs. |
|---|---|---|---|---|---|
| raw-ft | 96 (17.78) | 333 | 67 | 44 | 0 |
| kw-ft | 133 (24.63) | 64 | 277 | 66 | 0 |
| read-first | 133 (24.63) | 62 | 273 | 71 | 1 |
| decomp-ft | 138 (25.56) | 67 | 277 | 58 | 0 |
| adaptive-ft | 122 (22.59) | 68 | 282 | 68 | 0 |

(b) Observed exposure and conditional uninspected failures

| Condition | Exposed $E/540$ (%) | Uninspected failures $U/E$ (%) | Later legal action in $U$ |
|---|---|---|---|
| raw-ft | 207/540 (38.3) | 67/207 (32.4) | 67/67 |
| kw-ft | 476/540 (88.1) | 277/476 (58.2) | 277/277 |
| read-first | 477/540 (88.3) | 273/477 (57.2) | 273/273 |
| decomp-ft | 473/540 (87.6) | 277/473 (58.6) | 277/277 |
| adaptive-ft | 472/540 (87.4) | 282/472 (59.7) | 282/282 |

All accuracy denominators are 540. Unobs. denotes an unobservable trajectory: the read-first API exception is not a never-exposed failure. Observed read-first exposure is 477/540, without imputing the missing event. $U$ counts exposed-but-uninspected failures; $U/E$ conditions on each condition's own exposed subgroup, not a fixed causal comparison. Correctness takes precedence over other outcome categories; exposure is measured independently. A later legal action is only a weak opportunity indicator. Auxiliary raw/keyword summaries are retained in Appendix A.

**Observed outcome differences.** The keyword condition combines fewer never-exposed failures (62%→12%) with an accuracy difference of 6.85 percentage points. Of the 333 raw-never-exposed questions, 182 are exposed but uninspected under KW (Table 2). Thirty-six questions correct under RAW are exposed-but-uninspected failures under KW.

Table 2: Per-question outcome transitions raw-ft → kw-ft on the same 540 questions. All categories except Correct are incorrect outcomes; exposure and inspection refer to the target paper.

| raw ↓ kw → | Neither exposed nor inspected | Exposed, not inspected | Inspected | Correct |
|---|---|---|---|---|
| Neither exposed nor inspected (333) | 52 | **182** | 35 | 64 |
| Exposed, not inspected (67) | 4 | 46 | 8 | 9 |
| Inspected (44) | 6 | 13 | 12 | 13 |
| Correct (96) | 2 | 36 | 11 | 47 |

Table 3: Keyword first-exposure rank, subsequent target inspection attempts and final correctness. All counts are numbers of questions. Percentages use the exposed questions in each rank group as denominators. "Later target inspection" requires an executed target call on a strictly later turn.

| First-exposure rank | 1–5 | 6–20 | 21–50 | 51–200 |
|---|---|---|---|---|
| Exposed questions $n$ | 175 | 161 | 79 | 61 |
| Later target inspection, $n$ (%) | 92 (52.6) | 60 (37.3) | 26 (32.9) | 13 (21.3) |
| Correct answers, $n$ (%) | 65 (37.1) | 39 (24.2) | 21 (26.6) | 8 (13.1) |

All 476 exposed keyword questions are grouped by first-search-exposure rank, cross-checked against serialized search results. Subsequent attempts are executed target calls on strictly later turns; none precedes or coincides with first exposure, so all 191 questions with a target inspection attempt are included. Correctness uses historical primary labels, joined one-to-one by question ID, and the same exposed denominators. Outcomes are observed through each episode's actual termination. Rank, exposure timing and post-exposure observation length are uncontrolled; these rates do not identify a causal rank effect.

All 277 keyword exposed-but-uninspected failures have a later legal action. The target appears in the search results for 476 keyword episodes. In 335 of these episodes, it first appears by turn 3. Across these 476 keyword episodes, the median number of subsequently recorded search, evidence-search or answer actions is 13. This count describes later actions, not unused budget. Limits remain relevant: 167/277 keyword exposed-but-uninspected failures end at a token or turn limit, versus 20/67 under RAW.

**First-exposure rank and subsequent inspection.** Among the 336 keyword questions whose target first appears within the top 20, 184 (54.8%) have no later target inspection attempt. At ranks 1–5, 83 of 175 questions (47.4%) have no later target inspection attempt (Table 3). Both counts include questions answered correctly and incorrectly. Uninspected targets are therefore not confined to the result-list tail. Appendix A retains the separate failure-count lower bound.

## 6 READ-FIRST GATE: INSPECTION AND ANSWER OUTCOMES

The large exposed-but-uninspected group motivates a limited test of additional inspection. Read-first inserts a within-paper evidence-search attempt before another search or final answer when an eligible unread candidate remains. The audit asks whether this change reaches the target on additional questions, alongside changes in final correctness. Keyword versus raw and read-first versus keyword are the primary comparisons; decomposition and adaptive search are secondary.

Table 4: Paired correctness and questions with a target inspection attempt. Gains and losses refer to the left-to-right policy change.

(a) Correctness on all $N = 540$ questions

| Comparison | Both correct | Gain | Loss | Both wrong | ΔAcc. (pp) |
|---|---|---|---|---|---|
| raw → kw | 47 | 86 | 49 | 358 | +6.85 |
| kw → read-first | 73 | 60 | 60 | 347 | 0.00 |
| kw → decomp | 76 | 62 | 57 | 345 | +0.93 |
| kw → adaptive | 69 | 53 | 64 | 354 | −2.04 |

(b) Target inspection attempt: kw → read-first, $N = 539$ jointly observable

| Event | Both | Only read-first | Only kw | Neither | Unobs. pair |
|---|---|---|---|---|---|
| Target inspection attempt | 133 | 66 | 57 | 283 | 1* |

*Excluded only from panel b's 539 observable pairs, not panel a's 540 outcomes. Across all 540 questions, observed numbers of questions with a target inspection attempt are 191 (kw) and 199 (read-first); within the 539 pairs they are 190 and 199, a net increase of nine amid 123 attempt-status changes. This differs from the 120 correctness changes in panel a. The excluded pair has a keyword attempt and a read-first API exception, not an observed loss of target evidence. Each episode with a target inspection attempt has at least one executed target call and one same-turn nonempty target window; this is episode coverage, not a per-call success rate.

The GT-free reading gate becomes pending after a search. If the agent next proposes another search or a final answer, the harness substitutes an evidence-search call on the highest-ranked eligible candidate from the latest result list, using the original question. Eligibility requires a valid normalized arXiv ID absent from the attempted-read list; an attempt removes unread status even if it returns an error. If no eligible candidate remains, the proposed search or final answer proceeds without substitution, subject to the usual execution limits. The substituted call consumes the current turn, and the model chooses again on the next turn rather than automatically executing the deferred proposal. Any evidence-search call initiated by the agent satisfies the gate, even if the call does not use the highest-ranked eligible candidate or the original question. After the next paper search, the gate becomes pending again. The prompt specifies a stricter top-unread read than the implementation enforces. Results concern this implemented gate, which may also be pre-empted by context or turn limits; episodes that reach a context or turn limit remain in the 540-question accuracy analysis and resource summaries.

Keyword and read-first each answer 133 of 540 questions correctly. They answer 73 questions correctly in both conditions; 60 are correct only under keyword search and 60 only under read-first (Table 4). Read-first has more evidence-search calls per question than keyword search (9.63 versus 7.56), but recorded target inspection attempts occur on 199 questions under read-first and 191 under keyword search (Table 5). Among the 539 questions with observable trajectories in both conditions, the corresponding numbers of questions with a target inspection attempt are 199 and 190.

**Target inspection and returned excerpts.** The 539 observable read-first trajectories contain 5,198 evidence-search calls: 1,720 forced by the gate and 3,478 initiated by the agent (Table 5). Forced calls target the benchmark paper on 22 questions, while agent-initiated calls do so on 187. The two sets overlap on ten questions. The target is inspected only through forced calls on twelve questions and only through agent-initiated calls on 177 questions. This within-condition accounting is separate from the paired coverage difference between keyword and read-first (Table 4). Each observed episode with a target inspection attempt contains a same-turn, nonempty target match window linked to an executed call; individual calls need not succeed. Appendix B gives the matching rules.

**Paired uncertainty and exploratory-set sensitivity.** Paired question-bootstrap estimates give keyword minus raw as +6.85 percentage points (95% interval [2.59, 11.11]) and read-first minus keyword as 0.00 points ([-4.07, 4.07]). Resampling the 529 exact-target-ID clusters gives intervals [2.75, 11.05] and [-4.03, 4.05], respectively. These intervals quantify variation across questions or targets, not gateway drift, judge error or inference resampling. Excluding the 77 exploratory questions leaves 463 questions: keyword minus raw has 74 gains and 43 losses (+6.70 points; question-bootstrap interval [2.16, 11.23]); read-first minus keyword has 52 gains and 50 losses (+0.43 points; [-3.89, 4.75]). This sensitivity analysis does not provide independent confirmation.

Table 5: Recorded inference use and termination. Means use all 540 primary episodes per condition; termination columns are counts. Evidence-search calls include all candidate papers. The second panel separates call origins in the 539 observable read-first trajectories.

| Condition | Searches /q | Evidence-search calls/q | Tokens /q (k) | Normal | Context limit | Turn limit | API exception |
|---|---|---|---|---|---|---|---|
| raw-ft | 3.61 | 8.43 | 844 | 386 | 131 | 23 | 0 |
| kw-ft | 7.49 | 7.56 | 807 | 306 | 198 | 36 | 0 |
| read-first | 7.51 | 9.63 | 846 | 301 | 172 | 66 | 1 |
| decomp-ft | 7.38 | 7.43 | 762 | 290 | 205 | 45 | 0 |
| adaptive-ft | 7.36 | 7.85 | 791 | 287 | 203 | 50 | 0 |

| Read-first call origin | Evidence-search calls | Target inspection attempts | Questions with a target inspection attempt |
|---|---|---|---|
| Harness-forced | 1,720 | 22 | 22 |
| Agent-initiated | 3,478 | 381 | 187 |

Tokens are cumulative recorded input plus output (k $= 1000$), not peak context or billed costs. Context-limit includes the harness soft threshold, not evidence of provider output truncation. All primary episodes, including the API exception, remain in the means; missing recorded usage does not imply zero infrastructure cost. The separate retry, evaluator calls and unrecorded infrastructure work are excluded.

**Observed inference resource use.** Cost-aware agent evaluation motivates reporting resources alongside accuracy (Kapoor et al., 2024). The read-first condition has 27.4% more recorded evidence-search calls and 4.8% more inference tokens than keyword, with the same aggregate accuracy (Table 5).

**Unobservable-trajectory sensitivity.** A separately evaluated stochastic retry of the read-first API exception is also incorrect but exposes and inspects the target; substituting it leaves every correctness label unchanged and moves the unobservable episode to inspected-but-incorrect. This sensitivity analysis does not replace the primary trajectory, and its additional compute is excluded from primary-run costs.

**Validation scope.** Automated source-file checks with AI-assisted inspection found no discrepancies between retained logs and event summaries in 40 stratified trajectories and five additional anomaly records. Appendix B documents the separate exploratory excerpt review.

**Scorer sensitivity.** A target-ID audit identifies one title-version mismatch rejected by the historical title-based judge in keyword, read-first, decomposition and adaptive, but not raw. ID-only sensitivity adds one correct answer in each of those four conditions, leaving the keyword/read-first difference zero. The uniform title-based re-score preserves these four incorrect labels. We retain the historical title-based labels to preserve the original primary evaluation protocol, with normalized target-ID agreement reported separately as a scoring sensitivity.

## 7 Discussion and Limitations

The audit distinguishes where to investigate an incorrect answer. A never-exposed target directs attention to search and result presentation. An exposed but uninspected target motivates examination of candidate selection and the remaining execution budget. An inspected but incorrect outcome calls for checking the returned excerpts before attributing failure to reasoning. These categories specify the next evidence to inspect, rather than selecting a remedy in advance.

The read-first comparison separates evidence-search calls across all candidates from target inspection attempts. The keyword and read-first conditions differ in which questions involve a target inspection attempt and which questions are answered correctly. Paired accounting exposes these differences even when aggregate accuracy is identical. Returned-excerpt checks distinguish a logged call from a nonempty recorded return, leaving semantic sufficiency as a separate question.

Our measurements concern one answerable benchmark subset in a target-enriched retrieval environment. Historical underlying-model and effective reasoning settings are incompletely recorded, and the 77 exploratory questions informed the gate. There are no same-condition replicate runs to separate policy differences from stochastic trajectory variation; question and target bootstraps do not supply such replication. Retained excerpts do not permit complete reconstruction or semantic evaluation of all model inputs. These boundaries limit generalization and mechanism claims.

## 8 Conclusion

Decision-checkpoint auditing links recorded target exposure and executed inspection attempts to separately scored answers on the same questions. Across 540 answerable AutoResearchBench Deep questions in our target-enriched environment, keyword search has fewer failures with the target neither exposed nor inspected than raw search, but more exposed-but-uninspected failures. Of 336 keyword questions with first exposure in the top 20, 184 have no later target inspection attempt; these include correct and incorrect answers. Read-first records 27.4% more evidence-search calls than keyword search, yet target inspection occurs on 199 versus 191 of all 540 questions. Both answer 133 correctly, with 60 gains and 60 losses. The protocol thus quantifies distinctions between exposure and inspection, overall calls and target-specific attempts, and aggregate accuracy and paired outcomes that final scores alone leave unresolved.

### AI use statement

AI systems assisted with methodological and experimental design, code and implementation, qualitative analysis and interpretation of trajectories and results, figure preparation, and manuscript drafting and revision, under the direction of the authors. The authors are responsible for the final content of the manuscript.

### Reproducibility statement

Appendix B documents the event definitions, identifier normalization, query rules, scoring procedure, and bootstrap implementation used in our analyses. The reported event counts, paired statistics, and sensitivity analyses are generated deterministically from retained outputs using the procedures specified there.

## References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15068–15083, 2024. doi: 10.18653/v1/2024.emnlp-main.840. URL https://aclanthology.org/2024.emnlp-main.840/.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. OpenScholar: Synthesizing scientific literature with retrieval-augmented LMs. *arXiv preprint arXiv:2411.14199*, 2024. URL https://arxiv.org/abs/2411.14199.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-Plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Cornell University. arXiv dataset. Kaggle, https://www.kaggle.com/datasets/Cornell-University/arxiv, n.d. CC0 metadata; local index documented September 2026; source version and download date unrecorded.

Zhengru Fang, Senkang Forest Hu, Zhonghao Chang, Yu Guo, Yihang Tao, Hongyao Liu, Mengzhe Ruan, Jun Huang, and Yuguang Fang. Inference-time budget control for LLM search agents. *arXiv preprint arXiv:2605.05701*, 2026.

Dongxin Guo, Jikun Wu, and Siu Ming Yiu. AgentEval: DAG-structured step-level evaluation for agentic workflows with error propagation tracking. *arXiv preprint arXiv:2604.23581*, 2026.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. PaSa: An LLM agent for comprehensive academic paper search. *arXiv preprint arXiv:2501.10120*, 2025. URL https://arxiv.org/abs/2501.10120v2.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *arXiv preprint arXiv:2407.01502*, 2024. URL https://arxiv.org/abs/2407.01502.

Yushu Li, Wenlong Deng, Jiajin Li, and Xiaoxiao Li. Spend less, reason better: Budget-aware value tree search for LLM agents. *arXiv preprint arXiv:2603.12634*, 2026.

Zhixiang Liang, Yifei Liu, Yidan Huang, Haozhe Zhao, Beichen Huang, Jiaqi Wang, Nan Duan, and Qiong Cao. SearchAuditor: Auditing and attributing failures in long-horizon search agents, 2026. URL https://arxiv.org/abs/2608.05212.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

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. *arXiv preprint arXiv:2401.13178*, 2024. URL https://arxiv.org/abs/2401.13178v1.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009. doi: 10.1561/1500000019.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*, 2024. URL https://arxiv.org/abs/2408.08067v2.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024. URL https://arxiv.org/abs/2409.13740v1.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Lei Xiong, Kun Luo, Ziyi Xia, Wenbo Zhang, Jin-Ge Yao, Zheng Liu, Jingying Shao, Jianlyu Chen, Hongjin Qian, Xi Yang, Qian Yu, Hao Li, Chen Yue, Xiaan Du, Yuyang Wang, Yesheng Liu, Haiyu Xu, and Zhicheng Dou. AutoResearchBench: Benchmarking AI agents on complex scientific literature discovery. *arXiv preprint arXiv:2604.25256*, 2026. https://github.com/CherYou/AutoResearchBench.

## A Auxiliary Raw/Keyword Summaries

Table 6 gives auxiliary raw/keyword statistics. The event margins imply that at least $336 + 277 - 476 = 137$ of the 277 keyword exposed-but-uninspected failures first expose the target within rank 20. This lower bound concerns failures; the main text's 184 uninspected episodes include correct answers.

Table 6: Auxiliary raw/keyword audit on 540 questions per condition.

| Measure | raw-ft | kw-ft |
|---|---|---|
| Never-exposed failures (%) | 61.7 | 11.9 |
| Exposed, uninspected failures (%) | 12.4 | 51.3 |
| Inspected, incorrect failures (%) | 8.1 | 12.2 |
| Correct (%) | 17.8 | 24.6 |
| Search-result lists containing the target per question | 1.0 | 3.8 |
| Uninspected-failure stops: normal / turn / context limit | 47 / 4 / 16 | 110 / 24 / 143 |
| Searches / evidence-search calls per question (rounded) | 3.6 / 8.4 | 7.5 / 7.6 |
| Episodes ending without an arXiv ID in the final answer | 238 | 73 |

## B Measurement and Scoring Implementation

**Identifiers and events.** Identifier normalization removes whitespace, the `arxiv:` prefix and version suffixes, accepting modern and legacy formats. For modern IDs dated January 2015 onward, four-digit sequences are zero-padded to five digits; this convention does not validate identity. First exposure is the earliest target match in ordered search-result IDs, with one-based rank verified against serialized search messages. Inspection attempts require an executed target call, not a proposal; gate substitutions retain the original proposal separately. Repeated IDs outside search observations do not create exposure. Subsequent decisions comprise later search, within-paper evidence-search or answer actions through termination, excluding parse errors.

**Recorded target windows.** The within-paper evidence-search tool is implemented as `search_fulltext`. Each checkpoint retains the first 1,800 characters of the latest evidence-search response, rather than the complete return or model input. A qualifying target window must accompany an executed target call in the same episode and turn, identify that target in its header, and contain a match marker with a numeric line number followed by non-whitespace text. Empty, error and no-match returns are excluded. Episode coverage requires at least one qualifying window, not success on every call; missing trajectories remain unobservable.

**Primary scoring.** The evaluator compares the metadata title of the first candidate in the final answer with the target title in a single pass. After trimming whitespace, absent or empty titles and no-answer states score zero on this answerable cohort. The judge assesses same-paper identity, allowing truncation, punctuation, capitalization and subtitle differences that preserve identity; it does not independently verify question constraints. The historical parser reads JSON `is_match`, falling back to a substring-"true" test for invalid JSON. Normalized target-ID agreement is reported separately and does not override primary labels.

**Re-scoring configuration.** The uniform sensitivity re-score uses the same title prompts and exact-pair caching, with alias `gpt-5.6-luna`, temperature 0, reasoning effort `medium` and text responses. The requested completion limit of 2,048 is dropped by the gateway; no immutable upstream model snapshot is supplied. Unlike the historical parser, this re-score requires strict Boolean JSON and leaves unresolved responses unscored. Re-parsing saved responses and rebuilding scores without historical-label inputs produced resolved sensitivity labels for all 2,700 answers, with no historical-label backfilling and no disagreements with the primary labels. These settings describe the sensitivity run; the historical judge configuration remains incompletely recorded.

**Query rules and behavior when no candidate is eligible.** The historical RAW system prompt applies the verbatim-question rule to every search, not just the first; it permits changing top-$N$, inspecting evidence or finishing, but prohibits shortening, paraphrasing, adding terms, title substitution and decomposition. KW is a continuing system-prompt rule to form short, distinctive keyword queries before searching, not a separate preprocessing call; subsequent queries may differ. The harness passes the proposed query to the backend after checking that it is nonempty, rather than enforcing RAW equality. The gate checks the latest result list for a normalized ID absent from the attempted-read list. An empty eligible set makes the gate not due, so ordinary search/finish execution proceeds. The opportunity indicator is stored as `actionable_exposed_not_read`.

**Keyword/read-first totals.** The following preserves the aggregate comparison alongside the call-origin breakdown in Table 5.

| Keyword/read-first summary | Keyword | Read-first |
|---|---|---|
| Accepted answers | 133/540 | 133/540 |
| Questions with a target inspection attempt (all 540) | 191 | 199 |
| Questions with a target inspection attempt (539 observable pairs) | 190 | 199 |
| Unobservable trajectories | 0 | 1 |

**Forced and agent-initiated calls.** An offline audit joins checkpoints to turn details by question and turn, requiring agreement between the executed paper ID, the action's intervention marker and the turn's `intervention_read` status; forced calls also retain the original proposal. The 539 observable read-first trajectories contain 1,720 forced and 3,478 agent-initiated evidence-search calls. Forced calls target the benchmark paper on 22 questions (22 calls); agent-initiated calls do so on 187 questions (381 calls). These question sets overlap on ten questions, leaving twelve with forced-only and 177 with agent-only target inspection attempts. Their union contains 199 questions. All 5,198 recorded calls yield 9.63 calls per question under the shared 540-question denominator; the missing trajectory is not assigned zero infrastructure cost. The forced calls include 15 error returns and no explicit no-matching-passage returns; agent-initiated calls include 15 errors and three explicit no-matching-passage returns. Counts agree with each episode's stored intervention and full-text-call totals. The remaining API-exception trajectory stays unobservable. These counts describe execution and recorded returns, not additional calls relative to a counterfactual trajectory or semantic evidence sufficiency.

**Bootstrap implementation.** The paired analysis uses 100,000 draws per comparison and resampling scheme, with NumPy 2.5.3's `default_rng` seeded at 20260923. The reported intervals use the companion's fixed item-alias ordering; clusters are ordered by first occurrence in that ordering. The question bootstrap samples the empirical paired differences through multinomial counts for differences $-1$, $0$ and $1$. Intervals are percentile intervals at quantiles 0.025 and 0.975, computed by `numpy.quantile` with its linear interpolation rule and expressed in percentage points. Target clusters group questions by the exact sorted target-ID tuple. Each cluster draw samples the original number of clusters uniformly with replacement and retains all questions in every sampled cluster; the estimate divides the sum of paired differences by the sampled question count. Thus clusters have equal sampling probability, while their questions retain their multiplicity in the estimate; cluster means are not equally averaged. One seeded generator is consumed sequentially for keyword–raw, read-first–keyword, decomposition–keyword and adaptive–keyword, first on all questions and then excluding the exploratory subset, with question draws before cluster draws. These procedures resample saved outcomes, not agent executions.

**Exploratory semantic check.** AI-assisted constraint-level review covered 26 retained excerpts from ten selected records, rather than complete returns. It found partial support and one unresolved comparison-direction ambiguity. This exploratory check neither establishes a benchmark error nor supplies independent human validation, full evidence sufficiency or reasoning-failure attribution.