INSPIRE: BENCHMARKING SCIENTIFIC LITERATURE SEARCH FOR OPEN RESEARCH PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific literature search often begins with an open research problem rather than a known target paper or a fixed candidate set. We introduce INSPIRE, a benchmark for evaluating agents that search prior literature to make progress on solution-redacted research problems. Each instance pairs a research brief with a target-specific cutoff three months before a later paper and evaluates ranked outputs against graded cited antecedents from that paper's realized research lineage. Search proceeds over an open corpus, while the identity of the target paper and membership of its cited antecedents remain hidden from the agent. Beyond end-to-end retrieval quality, INSPIRE uses logged search trajectories to distinguish three coupled stages: *resource exposure*, whether useful antecedents are surfaced during search; *selection*, whether exposed antecedents are retained; and *ranking*, how effectively retained papers are ordered. Across 476 computer-science targets under a shared search interface and budget, the strongest evaluated agent achieves 0.284 nDCG@10. Results show that current agents more readily recover an isolated antecedent than assemble a broader portfolio of relevant prior work. The stagewise analysis identifies resource exposure as the largest observed bottleneck, with further losses in selection and ranking. We additionally construct replay-valid hindsight demonstrations and show that they improve held-out search without changing test-time information, establishing that the benchmark provides an actionable learning signal. INSPIRE therefore enables both end-to-end comparison and stage-resolved diagnosis in a setting where the agent must construct its own working criterion of relevance.

1 INTRODUCTION

Search-augmented language models have progressed from retrieval-augmented generation and browser-assisted question answering to agents that interleave reasoning, tool use, evidence inspection, and synthesis (Lewis et al., 2020; Nakano et al., 2021; Yao et al., 2022; Schick et al., 2023; Liu et al., 2024; Wei et al., 2025; Du et al., 2025). Reliable scientific search, however, is not a single retrieval decision. It comprises three stages: *resource exposure*, in which the agent searches broadly enough to surface useful candidate papers; *selection*, in which it retains the most relevant papers from those exposed; and *ranking*, in which it orders the retained evidence into a compact final response. These stages are diagnostically distinct but operationally coupled. When search begins from an unresolved problem, the eventual solution is hidden, useful mechanisms may not share the problem's vocabulary, and the agent's provisional relevance criterion governs both which evidence becomes visible and which evidence is retained. Retrieved evidence can in turn revise that criterion and redirect subsequent search. This evolving information need is central to anomalous-state, berrypicking, and exploratory-search accounts (Belkin et al., 1982; Bates, 1989; Marchionini, 2006), and to the view that relevance is contextual (Saracevic, 1975).

Existing benchmarks evaluate important subsets of this process, but often stabilize what counts as correct before or during search: BrowseComp supplies clues, AutoResearchBench explicit acceptance conditions, PaperPilot an anchor and optional user clarification, and ResearchBench a curated candidate set (Wei et al., 2025; Xiong et al., 2026; Li et al., 2026; Liu et al., 2025). MIR and HyBIRD are closer problem-conditioned inspiration retrievers, while PreScience evaluates future prior-work choices using author and publication histories (Garikaparthi et al., 2025; Yang et al., 2026; Ajith et al., 2026). We classify a task that specifies identifying information and asks for the matching

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.

Table 1: **Capability coverage of selected scientific-search benchmarks.** A ✓ denotes coverage and a × denotes no coverage under our operational definitions. Resource exposure evaluates whether a system can surface the needed literature from a corpus. Selection requires evaluating which already exposed papers are retained or omitted, and ranking evaluates their submitted order. HyBIRD uses MIR (Garikaparthi et al., 2025). Only INSPIRE covers all five characteristics.

	Task setting		Research capability evaluated		
Benchmark	Interactive open-corpus search	No supplied relevance cue	Resource exposure	Selection	Ranking
AutoResearchBench (Xiong et al., 2026)	✓	×	✓	×	×
PaperPilot (Li et al., 2026)	✓	×	✓	×	✓
ResearchBench (Liu et al., 2025)	×	×	×	✓	×
HyBIRD (Yang et al., 2026)	×	✓	✓	×	✓
PreScience (Ajith et al., 2026)	×	×	✓	×	✓
INSPIRE (ours)	✓	✓	✓	✓	✓

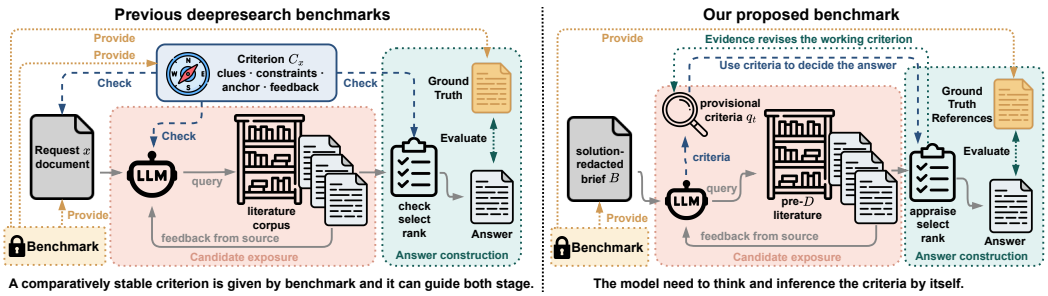

Figure 1: **Source of the working relevance criterion.** Left: prior tasks stabilize relevance through clues, constraints, anchors, feedback, or candidate pools. Right: INSPIRE supplies only a solution-redacted problem, so the agent must form provisional criteria that guide resource exposure, selection, and ranking. Retrieved evidence can revise those criteria and redirect subsequent search. The schematic contrasts information supplied during interaction; task difficulty can vary independently.

paper as an evaluation of resource exposure: once the paper is surfaced, the supplied conditions determine whether it should be returned. By contrast, selection requires the benchmark to evaluate which already observed candidates the model retains or rejects. Existing protocols may evaluate exposure, selection, or ranking individually, but do not measure all three on a common trajectory; consequently, their end-to-end scores cannot determine where relevant evidence was lost.

Table 1 compares benchmark characteristics under these definitions. INSPIRE evaluates all three stages of scientific search within a single open-corpus trajectory. Given only a solution-redacted problem brief, an agent searches literature published before a target-specific cutoff and submits up to ten ranked papers, without access to the hidden target, its bibliography, or relevance labels. After the episode, graded cited antecedents from the target's realized lineage provide three nested measurements: *exposed gain* evaluates resource exposure, *selected gain* evaluates which exposed papers are retained, and final nDCG@10 evaluates their ranking. The gaps between these measurements reveal whether useful evidence was never surfaced, was exposed but omitted, or was retained but poorly ordered.

Under a shared interface and budget, the strongest model reaches 0.284 nDCG@10 and recovers at least one antecedent on 72.3% of targets, yet recovers three on only 32.1% of eligible targets. Across models, the zero-exposure rate ranges from 17.4% to 56.1%, and the largest observed loss occurs before selection. A matched hindsight-SFT intervention raises held-out nDCG@10 from 0.082 to 0.166 under unchanged test-time information, showing that INSPIRE supplies an actionable learning signal. Figure 1 contrasts this information structure with supplied-criterion tasks. Our contributions are:

- We formulate scientific search as three coupled stages—resource exposure, selection, and ranking—and show how a provisional relevance criterion can shape all three.

- We introduce INSPIRE, a temporally isolated, solution-redacted benchmark that evaluates each stage from logged trajectories using hindsight-revealed graded antecedents while strictly separating policy and evaluator information.
- We provide a controlled multi-model comparison that localizes current failures across the three-stage funnel, and a matched Base–SFT intervention showing that INSPIRE's training signal produces measurable improvement on held-out targets.

2 A COUPLED VIEW OF AGENTIC SEARCH

2.1 RESOURCE EXPOSURE, SELECTION, AND RANKING

Notation. Let B denote a problem brief, $\mathcal{C}$ the searchable corpus, $E \subseteq \mathcal{C}$ the papers exposed during search, $S = (s_1, \ldots, s_m)$ the submitted ranking, $\bar{S} = \{s_1, \ldots, s_m\}$ its unordered selected set, and $Y \subseteq \mathcal{C}$ the hidden graded label set. Omitting target subscripts, success requires

$$Y \cap E \neq \varnothing \quad \text{and} \quad Y \cap \bar{S} \neq \varnothing, \tag{1}$$

with stronger portfolio tasks requiring several members of Y. The three stages ask progressively stricter questions: did a graded paper enter E (resource exposure), did it survive in $\bar{S}$ (selection), and how effectively was $\bar{S}$ ordered in S (ranking)? All three are observable from returned and submitted identifiers. Zero exposure establishes a hard ceiling for any answer constructed from the observed trajectory, whereas positive exposure is only a necessary condition: the trajectory may contain a low-gain antecedent, omit it from the final response, or place it at an ineffective rank. This nested structure enables diagnostic evaluation even though the stages interact. These stages are coupled rather than a retrieve-then-rank pipeline: an evolving information need shapes queries and appraisal, while retrieved evidence changes that need (Belkin et al., 1982; Bates, 1989; Marchionini, 2006). Let H_t be the history, z a relevance hypothesis (such as a candidate mechanism family), and $q_t(z)$ the agent's distribution over hypotheses. The same q_t can govern search, appraisal of paper d, and subsequent belief updates:

$$a_t^{\text{search}} \sim \pi_{\text{search}}(\cdot \mid B, H_t, q_t), \tag{2}$$

$$u_t(d) = \mathbb{E}_{z \sim q_t}[r(d \mid B, H_t, z)], \tag{3}$$

$$q_{t+1} = \mathcal{U}(q_t; B, H_t, O_t), \tag{4}$$

where a_t^{search}, π_{search}, u_t, r, O_t, and $\mathcal{U}$ are the next action, search policy, utility, relevance score, observation, and update rule. We use this analysis model to express how one provisional criterion affects both visibility and submission. A retrieved paper can shift mass toward a new mechanism family, changing the next query and the appraisal of papers already seen. The logged stages therefore support diagnostic accounting while retaining their dependence on a shared, evolving criterion.

2.2 SUPPLIED AND ENDOGENOUS RELEVANCE CRITERIA

Many tasks supply a comparatively stable account of what qualifies through clues, conditions, an anchor, user feedback, or candidate construction; BrowseComp, AutoResearchBench, PaperPilot, and ResearchBench instantiate these sources differently (Wei et al., 2025; Xiong et al., 2026; Li et al., 2026; Liu et al., 2025). Checking such a criterion may still require full-text reading, multi-hop search, or a learned judge; the distinction concerns information structure rather than intrinsic difficulty. In INSPIRE, by contrast, the visible problem may support distinct solution lineages z_1 and z_2 with different antecedent sets $Y(z_1) \neq Y(z_2)$. The evaluator uses the later realized $Y(z^\star)$, but $z^\star$ is hidden during search and the pre-cutoff record provides no membership certificate. We use *endogenous* in this operational sense: the agent forms the working criterion that drives its trajectory, and hindsight supplies an exact graded score after the episode. MIR and HyBIRD are related problem-conditioned inspiration retrievers (Garikaparthi et al., 2025; Yang et al., 2026), and PreScience also predicts later prior-work choices without an in-episode certificate (Ajith et al., 2026). INSPIRE's distinction is the conjunction of solution redaction, interactive date-gated open-corpus search, no anchor or clarification, hindsight evaluation, and logged stage measurements. The resulting funnel locates observed losses among exposure, selection, and ranking; causal diagnosis and counterfactuals under complete exposure require intervention.

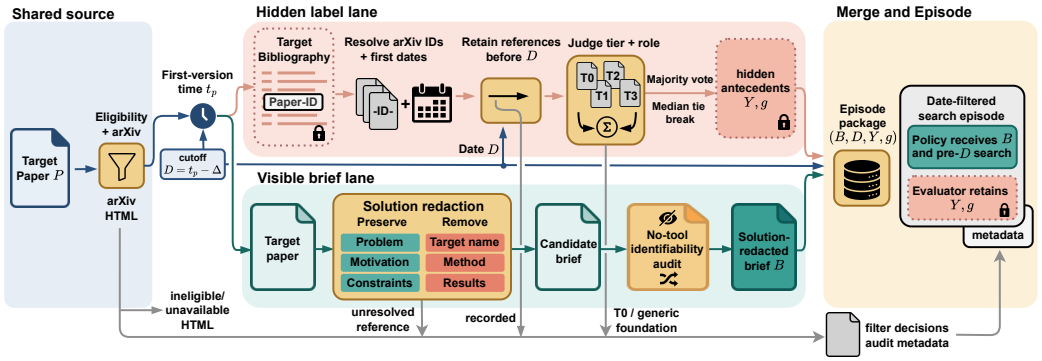

Figure 2: **Backward construction and information boundary.** Starting from a later target, the builder applies the three-month cutoff, reconstructs and screens a solution-redacted brief, and grades cited pre-cutoff antecedents. During evaluation, the policy receives only the brief and date-filtered search results; the target, its bibliography, and all grades remain evaluator-only. This separation permits hindsight scoring without exposing the realized solution lineage.

3 INSPIRE BENCHMARK

3.1 TASK

For target paper P_i first posted at t_{P_i}, instance i is

$$x_i = (B_i, D_i, \mathcal{C}_{<D_i}, Y_i, g_i), \qquad D_i = t_{P_i} - \Delta, \tag{5}$$

where $\Delta = 3$ months, B_i is the solution-redacted brief, D_i the cutoff, $\mathcal{C}_{<D_i}$ the searchable literature, Y_i the hidden graded cited antecedents, and g_i their gain function, with zero gain outside Y_i. A policy sees only B_i and its tool history; it never observes P_i, labels, future citation contexts, or evaluation feedback. The target-specific cutoff prevents the agent from using the target itself, concurrent work, or later public evidence, while retaining the open-corpus character of the search episode.

The ranked submission is $S_i = (s_{i1}, \ldots, s_{im_i})$, with $m_i \le K = 10$. Version suffixes are normalized, duplicates retain their first position, and malformed, post-cutoff, or unlabeled IDs receive zero. The primary score is graded nDCG@10:

$$\mathrm{nDCG}_{K,i}(S_i) = \frac{\sum_{j=1}^{\min(K,m_i)} g_i(s_{ij}) / \log_2(j+1)}{\sum_{j=1}^{\min(K,|Y_i|)} g_i(y^\star_{ij}) / \log_2(j+1)}, \tag{6}$$

where $(y^\star_{i1}, \ldots)$ orders Y_i by decreasing gain (Järvelin & Kekäläinen, 2002). The target-specific denominator compares a submission with the ideal ordering of the same hidden label set, so scores remain normalized despite variation in label count and gain. Recall, reciprocal rank, and one/three/five-hit rates complement this view; failures and non-submissions remain zero. Because citations have heterogeneous functions and omit influences (Teufel et al., 2006; Valenzuela et al., 2015; MacRoberts & MacRoberts, 1989; Bornmann & Daniel, 2008), the construct is recovery of graded antecedents from one realized lineage; scientific usefulness beyond that lineage and causal credit lie outside the score.

3.2 CONSTRUCTION

Construction proceeds backward from recent, influential computer-science target papers. We discover targets through topic-stratified Semantic Scholar searches, deduplicate arXiv identifiers, and use arXiv metadata as the authority for title, abstract, category, and first-posting date; surveys, reviews, papers without parseable HTML, and targets with fewer than two graded pre-cutoff antecedents are excluded. For each retained target P_i, we fix $D_i = t_{P_i} - 3$ months, normalize arXiv identifiers in its bibliography, and exclude references first posted on or after the cutoff. Role-sensitive judges then grade each eligible reference using the target abstract, candidate metadata, and citation context when available, producing consensus tier, role, rationale, and gain; non-antecedents are removed and the

graded set is capped at ten. In parallel, a generator converts the target title and abstract into a problem brief that preserves the research motivation and constraints while removing the target identity, realized mechanism, and target-specific results. A no-tool identity audit and a lexical-overlap check define reproducible clean and flagged diagnostic partitions. The released instance exposes only the brief, cutoff, and date-filtered search interface to the policy, while the target, bibliography, grades, and citation evidence remain evaluator-only; the manifest records split membership, construction settings, instance hashes, and judge versions. This design yields scalable silver labels for a realized citation lineage without leaking that lineage into the search episode. Appendix A provides full construction and sensitivity details, and Figure 2 illustrates the information boundary.

3.3 Stage-Wise Metrics

End-to-end scores cannot distinguish a relevant paper that was never observed from one that was exposed and then omitted or poorly ranked. We therefore introduce a nested evaluation of resource exposure, selection, and ranking. For each of N targets, let E_i be the set of unique paper identifiers returned during the trajectory, and let $\bar{S}_i = \{s_{ij}\}_{j=1}^{m_i}$ be the unordered support of the ranked submission S_i. The decomposition uses only papers observed in the logged trajectory: it neither retrieves new candidates nor changes the policy's information.

For any set $A \subseteq \mathcal{C}_{<D_i}$, define its normalized available gain by optimally ordering its labeled members:

$$\phi_i(A) = \mathrm{nDCG}_{K,i}(\mathrm{sort}_{g_i}(A \cap Y_i)) \in [0, 1]. \tag{7}$$

Thus, $\phi_i(E_i)$ is the best score supported by the observed exposure set, rather than the outcome of a counterfactual search policy. We aggregate performance after each stage as

$$G_{\mathrm{exposed}} = \frac{1}{N}\sum_i \phi_i(E_i), \quad G_{\mathrm{selected}} = \frac{1}{N}\sum_i \phi_i(\bar{S}_i \cap E_i), \quad G_{\mathrm{final}} = \frac{1}{N}\sum_i \mathrm{nDCG}_{K,i}(S_i). \tag{8}$$

G_{exposed} measures the ranking potential made available through resource exposure; G_{selected} measures the potential retained after selection, independent of submitted order; and G_{final} measures the realized quality after ranking. Because submissions must be drawn from exposed papers,

$$G_{\mathrm{final}} \le G_{\mathrm{selected}} \le G_{\mathrm{exposed}} \le 1. \tag{9}$$

The corresponding absolute losses are $1 - G_{\mathrm{exposed}}$ for exposure, $G_{\mathrm{exposed}} - G_{\mathrm{selected}}$ for selection, and $G_{\mathrm{selected}} - G_{\mathrm{final}}$ for ranking. These quantities place all three stages on the same target-normalized nDCG scale.

We additionally report the zero-exposure rate $N^{-1}\sum_i \mathbf{1}[E_i \cap Y_i = \varnothing]$ and the conditional retention ratios

$$\text{Gain kept} = \frac{G_{\mathrm{final}}}{G_{\mathrm{exposed}}}, \qquad \text{Rank quality} = \frac{G_{\mathrm{final}}}{G_{\mathrm{selected}}}, \tag{10}$$

for nonzero denominators. Gain kept combines selection and ranking loss, while Rank quality isolates ordering conditional on the selected set; equivalently, Gain kept factors into $G_{\mathrm{selected}}/G_{\mathrm{exposed}}$ and Rank quality. The zero-exposure rate identifies episodes in which no selector or ranker restricted to the observed candidates could recover a graded antecedent. Together, these metrics localize where gain disappears along observed trajectories. They do not by themselves identify whether an exposure loss was caused by the relevance hypothesis, query formulation, exploration policy, parametric memory, or search backend; such causal attributions require intervention.

4 Experiments and Results

4.1 Setup

Evaluation environment and information boundary. Each TimeMachine episode presents only B_i through common `search_papers` and `submit_answer` schemas. Every system receives the same brief, date filter, returned fields, search budget, output limit, and scorer. The hidden target, antecedent tiers, and future citation contexts never enter the policy history, ensuring that no condition receives evaluator information during search. The search backend, response schema, and episode budget are shared throughout. We log each ordered query, returned identifier, tool error, and final

Table 2: **Unified end-to-end comparison.** All rows use the same targets, interface, search budget, failure policy, and scorer. nDCG@10, Recall@10, MRR, and AnyHit@10 include failed or empty episodes as zero. The multi-hit columns report exact recovery of at least one, three, or five graded antecedents among targets with enough labels; Total calls is the mean number of search and paper-access interactions.

Model	nDCG@10	Recall@10	MRR	AnyHit@10	≥ 3 hits	≥ 5 hits	Total calls
Claude Opus 5	**0.284**	**0.215**	**0.476**	**0.723**	**0.321**	**0.067**	21.1
Claude Sonnet 5	0.189	0.155	0.361	0.603	0.175	0.033	17.7
GPT-5.6-Sol	0.164	0.122	0.321	0.559	0.134	0.011	26.3
Gemini 3.5 Flash	0.151	0.120	0.299	0.521	0.104	0.011	34.7
Gemini 3.1 Pro Preview	0.110	0.082	0.258	0.403	0.047	0.000	20.9
Qwen3.6-35B-A3B Base	0.082	0.071	0.200	0.357	0.042	0.006	21.0

submission to reconstruct first exposure and the complete funnel. This common interface supports direct comparisons while leaving parametric memory, provider effects, and possible contamination as uncontrolled sources of variation. Appendix B records the exact run manifest, model revisions, response settings, and call accounting.

Evaluation models and metrics.We evaluate Claude Opus 5 and Claude Sonnet 5 (Anthropic, 2026a;b), GPT-5.6-Sol (OpenAI, 2026b), Gemini 3.5 Flash and Gemini 3.1 Pro Preview (Google DeepMind, 2026b;a), and Qwen3.6-35B-A3B Base and its hindsight-SFT variant (Qwen Team, 2026) on the automatically screened manifest with $D_i = t_{P_i} - 3$ months. SFT uses a disjoint earlier pool and a development set frozen before training. Every condition follows the shared protocol above on the same ordered target manifest, and provider APIs only generate the next action. The primary metric is failure-as-zero nDCG@10 (Järvelin & Kekäläinen, 2002), which rewards both high-gain recovery and placement near the top of the submitted ranking. We also report Recall@10, MRR, $\mathrm{AnyHit@10}$, three- and five-hit rates, and total calls. These measure the recovered label fraction, first-hit rank, any connection to the realized lineage, stronger portfolio recovery, and interaction use. Multi-hit rates condition only on targets with enough labels, following the distinction between isolated relevance and diversified coverage (Carbonell & Goldstein, 1998; Clarke et al., 2008). Table 2 reports the aggregate point estimates.

4.2 End-to-End and Portfolio Performance

Table 2 reports the end-to-end comparison. Claude Opus 5 leads with 0.284 nDCG@10, 0.215 Recall@10, and 0.723 $\mathrm{AnyHit@10}$, followed by Claude Sonnet 5 at 0.189 nDCG@10; the ordering is broadly consistent across nDCG, recall, MRR, and $\mathrm{AnyHit@10}$. Yet even the leading system recovers only 21.5% of labeled antecedents on average, indicating substantial headroom beyond identifying one plausible paper.

Binary success further obscures the difficulty of portfolio recovery. Among eligible targets, the leading model's one-, three-, and five-hit rates decline from 72.3% to 32.1% and 6.7%, and Figure 3 shows the same concentration at zero to two hits across systems. Stronger agents therefore reduce complete misses more reliably than they assemble portfolios of complementary antecedents. This pattern is not explained by interaction volume alone: Gemini 3.5 Flash averages 34.7 calls and obtains 0.151 nDCG@10, whereas Claude Opus 5 averages 21.1 calls and obtains 0.284. What matters is both the evidence made available and how it is subsequently used, motivating the stage-wise analysis below.

4.3 Stage-Wise Performance

Table 3 reports the three nested quantities on a common normalized-gain scale. Exposed gain measures the ranking potential made available by resource exposure, Selected gain measures how much potential remains after the agent decides what to submit, and final nDCG@10 additionally reflects the submitted order. Their nesting, $G_{\mathrm{final}} \leq G_{\mathrm{selected}} \leq G_{\mathrm{exposed}}$, makes each successive gap attributable to one observed stage of the trajectory.

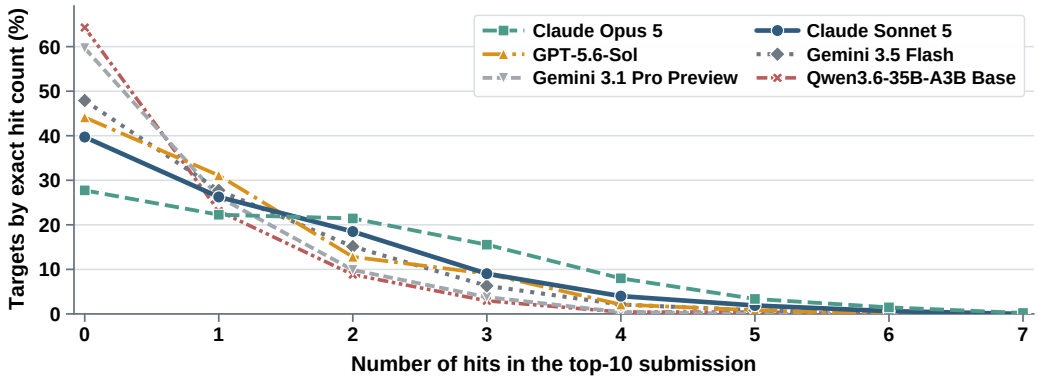

Figure 3: **Recovered antecedent counts across 476 targets.** The horizontal axis is the number of graded antecedents recovered in the submitted top 10, and the vertical axis is the percentage of targets with exactly that count. Failures and non-submissions enter the zero bin. Concentration at zero and one shows that recovering an isolated antecedent is substantially more common than assembling a broader portfolio.

Table 3: **Resource exposure, selection, and ranking funnel.** Exposed gain oracle-sorts every graded antecedent encountered during search; Selected gain oracle-sorts the graded antecedents retained in the submission; final nDCG@10 additionally reflects their submitted order. Gain kept and Rank quality are the retention ratios in Equation 10, while No relevant item exposed reports the zero-exposure rate. Empty submissions remain zero, and stage gaps describe the observed trajectories.

	Resource exposure		Selection	Ranking	Retention diagnostics	
Model	Exposed gain	No relevant item exposed	Selected gain	Final nDCG@10	Gain kept	Rank quality
Claude Opus 5	0.480	17.4%	0.386	0.284	0.591	0.735
Claude Sonnet 5	0.323	31.1%	0.269	0.189	0.585	0.703
GPT-5.6-Sol	0.326	30.5%	0.233	0.164	0.504	0.706
Gemini 3.5 Flash	0.290	35.3%	0.218	0.151	0.520	0.691
Gemini 3.1 Pro Preview	0.184	52.5%	0.148	0.110	0.599	0.742
Qwen3.6-35B-A3B Base	0.153	56.1%	0.119	0.082	0.532	0.683

Resource exposure. Exposed gain ranges from 0.153 to 0.480, and the fraction of targets for which no graded antecedent is exposed ranges from 17.4% to 56.1%. For every system, the largest absolute deficit relative to the normalized optimum occurs before selection: $1 - G_{\text{exposed}}$ is at least 0.520 and reaches 0.847. Claude Opus 5 exposes the most gain and has the fewest complete misses, whereas Qwen3.6-35B-A3B Base exposes less than one third as much gain and misses every label on more than half of the targets. Exposed gain and zero-exposure rate are complementary: the former captures the amount and grade of evidence made available across targets, whereas the latter isolates trajectories from which no labeled item could be recovered downstream. Surfacing one low-gain antecedent can reduce the miss rate without producing a strong candidate portfolio.

Selection. Selected gain ranges from 0.119 to 0.386. The gap $G_{\text{exposed}} - G_{\text{selected}}$ measures useful evidence that was observed but omitted from the submission, independently of ordering. Claude Sonnet 5 and GPT-5.6-Sol illustrate why this stage must be evaluated separately: they expose nearly identical gain (0.323 and 0.326), but Sonnet retains more of that potential (0.269 versus 0.233), contributing to its higher final nDCG@10 (0.189 versus 0.164). Conversely, a small absolute selection gap is not by itself evidence of strong selection when little useful material was exposed upstream; stage quantities must be interpreted jointly.

Ranking. Final nDCG@10 ranges from 0.082 to 0.284, while Rank quality ranges from 0.683 to 0.742. The gap $G_{\text{selected}} - G_{\text{final}}$ isolates the loss induced by the submitted order, because Selected gain is the best score achievable by reordering the same submitted support. Gemini 3.1 Pro Preview attains the highest Rank quality but exposes only 0.184 gain, demonstrating that effective ordering

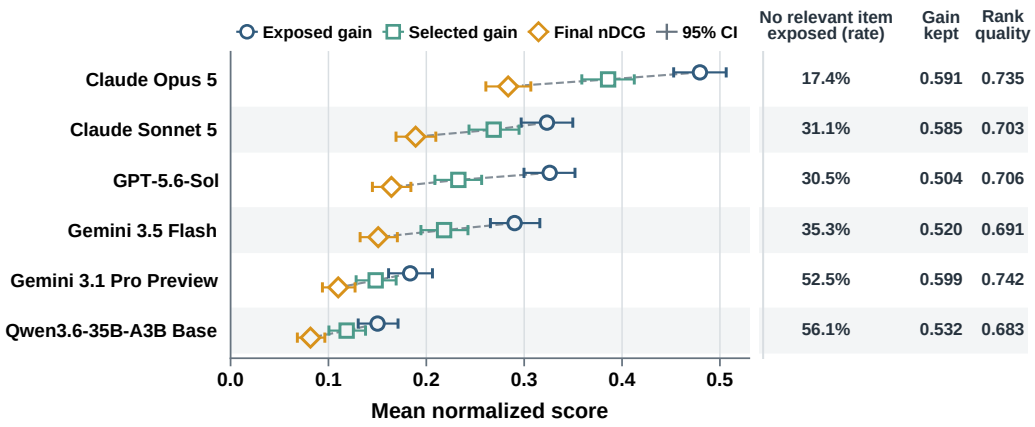

Figure 4: **Observed funnel across resource exposure, selection, and ranking.** For each model, the three points show mean Exposed gain, Selected gain, and final nDCG@10; horizontal bars are 95% target-bootstrap intervals. The right columns report zero-exposure rate, Gain kept, and Rank quality. Lower zero-exposure is better, whereas higher values are better for the other quantities. Horizontal separation localizes the stage of observed losses; causal mechanisms require intervention.

cannot compensate for evidence that was never surfaced. A reranker can improve the use of an existing submission, but it cannot recover an antecedent absent from the observed trajectory.

4.4 Matched Hindsight SFT

To test whether INSPIRE supports model development rather than merely separating fixed systems, we compare Qwen3.6-35B-A3B Base with a hindsight-SFT variant trained on a disjoint pool of earlier targets. Hindsight SFT uses training-only outcome information to identify successful trajectories, replays their actions through the standard search interface, and retains only demonstrations whose submitted papers were exposed during replay. The student is trained solely on the resulting deployment-legal transcripts: hidden relevance labels, teacher-side reasoning, and trajectory-selection scores are removed. We then evaluate the base and SFT models on the same held-out targets with identical cutoffs, interfaces, search budgets, decoding settings, and scorers, making the adapter the only experimental change. As shown in Table 4, SFT raises exposed gain from 0.153 to 0.251 and reduces zero-exposure episodes from 56.1% to 29.4%, indicating more effective discovery of relevant resources. It also raises selected gain from 0.119 to 0.219 and Rank quality from 0.683 to 0.759, showing that the model both retains more useful evidence and orders its final submission more effectively. Together, these stage-wise improvements increase final nDCG@10 from 0.082 to 0.166. The decomposition therefore reveals that hindsight supervision improves the full search pipeline rather than only its final output, demonstrating that INSPIRE provides actionable training signals as well as diagnostic evaluation; Appendix C details demonstration construction, leakage controls, training configuration, and paired uncertainty analysis.

Table 4: **Controlled benchmark-usability test.** The table reports changes in resource exposure, selection, and final ranking, allowing the end-to-end improvement to be localized, testing whether INSPIRE can produce useful supervision and measure held-out improvement.

	Search and gather		Filter and construct answer			End-to-end
Condition	Exposed gain	No relevant item exposed	Selected gain	Gain kept	Rank quality	Final nDCG@10
Qwen3.6-35B-A3B Base	0.153	56.1%	0.119	0.532	0.683	0.082
Qwen3.6-35B-A3B + hindsight SFT	**0.251**	**29.4%**	**0.219**	**0.661**	**0.759**	**0.166**

5 Related Work

Scientific search. Interactive IR models search as iterative clarification of a changing information need rather than one-shot lookup (Belkin et al., 1982; Bates, 1989; Marchionini, 2006; Saracevic, 1975), while literature-based discovery seeks non-obvious connections across separated bodies of work (Swanson, 1986). Language-model systems increasingly couple retrieval, browsing, reasoning, and tool use (Petroni et al., 2021; Lewis et al., 2020; Nakano et al., 2021; Yao et al., 2022; Schick et al., 2023; Liu et al., 2024); BrowseComp and DeepResearch Bench extend this trajectory to persistent clue-driven search and report-level evaluation (Wei et al., 2025; Du et al., 2025). Scientific retrieval builds on probabilistic ranking and heterogeneous-domain evaluation (Robertson & Zaragoza, 2009; Thakur et al., 2021), as well as literature graphs, structured corpora, scientific encoders, and citation-aware representations or recommendation (Ammar et al., 2018; Lo et al., 2020; Beltagy et al., 2019; Cohan et al., 2020; Singh et al., 2023; Bhagavatula et al., 2018). AutoResearchBench evaluates search against explicit multi-part conditions, whereas PaperPilot starts from an anchor and may elicit user intent (Xiong et al., 2026; Li et al., 2026). MIR and HyBIRD retrieve methodological inspirations from solution-redacted problems, and ResearchBench decomposes inspiration discovery and ranking within a curated candidate pool (Garikaparthi et al., 2025; Yang et al., 2026; Liu et al., 2025). These works motivate our information-structure contrast: INSPIRE combines interactive, date-gated open-corpus search with no anchor or clarification and separately measures whether antecedents are exposed, retained, and ranked.

Hindsight supervision. Freshness-aware retrieval motivates temporal isolation (Vu et al., 2024). PreScience searches a pre-target corpus while conditioning on future authors and their histories, SciPaths exposes a realized contribution, CUSP distinguishes plausible mechanisms from realized outcomes, and Future-Aligned Proposals supplies inspirations as input (Ajith et al., 2026; Chamoun et al., 2026; Wu et al., 2026; Wang et al., 2026). Graded cumulative gain rewards relevance and rank (Järvelin & Kekäläinen, 2002), while diversified retrieval motivates evaluating portfolios rather than isolated hits (Carbonell & Goldstein, 1998; Clarke et al., 2008). Studies of citation function and importance inform our tiered antecedent labels (Teufel et al., 2006; Valenzuela et al., 2015; MacRoberts & MacRoberts, 1989; Bornmann & Daniel, 2008). Finally, learning with privileged information and hindsight reuse formalizes the use of train-only outcome signals (Vapnik & Vashist, 2009; Andrychowicz et al., 2017); our matched study uses these signals to select replay-valid demonstrations for standard supervised instruction tuning with LoRA (Ouyang et al., 2022; Hu et al., 2022).

6 Limitations

Extending INSPIRE along several dimensions merits future exploration. Human missing-gold audits spanning uncited influences, older foundations, and alternative solution lineages could strengthen the current capped, recent arXiv citation labels (MacRoberts & MacRoberts, 1989; Bornmann & Daniel, 2008). Independent redaction studies, additional disciplines and search backends, and broader judge families could quantify residual semantic leakage, grading bias (Zheng et al., 2023), and domain dependence. Evaluations across students, training seeds, provider regimes, and controlled parametric-memory settings could test the generality of the SFT result and isolate the mechanisms behind observed funnel gaps. Auditing ranked retrieval for cumulative-advantage effects (Merton, 1968) would further clarify downstream impact. INSPIRE is designed to evaluate search systems rather than researchers or intellectual provenance.

7 Conclusion

Scientific search requires three coupled stages: *resource exposure*, *selection*, and *ranking*. INSPIRE evaluates them on the same logged trajectory through solution-redacted problems, date-gated open-corpus search, and hindsight-revealed graded antecedents. Current agents recover isolated antecedents more readily than broader portfolios; their largest observed loss occurs during resource exposure, with further losses in selection and ranking. Matched hindsight SFT improves all three stages under unchanged test-time information, showing that INSPIRE provides an actionable learning signal. By distinguishing evidence that was never surfaced, exposed but omitted, or retained but poorly ranked, this decomposition enables targeted model development.

AI USE STATEMENT

We used LLMs to support benchmark construction, including filtering candidate papers, generating solution-redacted problem briefs, grading cited antecedents, and auditing briefs for target identifiability. We also used an LLM teacher to generate synthetic search demonstrations for supervised fine-tuning and evaluated LLM-based search agents, as detailed in the methodology and appendices. For manuscript preparation, we used generative AI tools for language polishing and writing assistance. We reviewed and revised the AI-assisted text and take full responsibility for the final paper, including its claims, results, and accompanying artifacts.

REFERENCES

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C. Kozlowski, Oyvind Tafjord, James Evans, Daniel S. Weld, Tom Hope, and Doug Downey. PreScience: A dataset and benchmark for scientific forecasting. *arXiv preprint arXiv:2602.20459*, 2026.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 84–91, 2018.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Anthropic. Introducing Claude Opus 5. `https://www.anthropic.com/news/claude-opus-5`, July 2026a.

Anthropic. Introducing Claude Sonnet 5. `https://www.anthropic.com/news/claude-sonnet-5`, June 2026b.

Marcia J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989.

Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks. ASK for information retrieval: Part i. background and theory. *Journal of Documentation*, 38(2):61–71, 1982.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3615–3620, 2019.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 238–251, 2018.

Lutz Bornmann and Hans-Dieter Daniel. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80, 2008.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

Eric Chamoun, Yizhou Chi, Yulong Chen, Rui Cao, Zifeng Ding, Michalis Korakakis, and Andreas Vlachos. SciPaths: Forecasting pathways to scientific discovery. *arXiv preprint arXiv:2605.14600*, 2026.

Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 659–666, 2008.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, 2020.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

Aniketh Garikaparthi, Manasi Patwardhan, Aditya Sanjiv Kanade, Aman Hassan, Lovekesh Vig, and Arman Cohan. MIR: Methodology inspiration retrieval for scientific research problems. *arXiv preprint arXiv:2506.00249*, 2025.

Google DeepMind. Gemini 3.1 Pro: Model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, February 2026a.

Google DeepMind. Gemini 3.5 Flash: Model card. `https://deepmind.google/models/model-cards/gemini-3-5-flash/`, May 2026b.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jisen Li, Bingxuan Li, Nanyi Jiang, Xuying Ning, Xiyao Wang, Yifan Shen, Heng Wang, Yuqing Jian, Xiaoxia Wu, Ben Athiwaratkun, Pan Lu, Jiaxuan You, and Bingxin Zhao. Multi-turn agentic scientific literature search via workflow induction. *arXiv preprint arXiv:2607.00597*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. ResearchBench: Benchmarking LLMs in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248*, 2025.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel S. Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, 2020.

Michael H. MacRoberts and Barbara R. MacRoberts. Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, 40(5):342–349, 1989.

Gary Marchionini. Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

Robert K. Merton. The matthew effect in science. *Science*, 159(3810):56–63, 1968.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI. GPT-5.5 System Card. https://openai.com/index/gpt-5-5-system-card/, April 2026a.

OpenAI. GPT-5.6 System Card. https://deploymentsafety.openai.com/gpt-5-6, July 2026b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: A benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. https://qwen.ai/blog?id=qwen3.6-35b-a3b, April 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Tefko Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 103–110, 2006.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

Marco Valenzuela, Vu Ha, and Oren Etzioni. Identifying meaningful citations. In *AAAI Workshop on Scholarly Big Data*, 2015.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024.

Heng Wang, Pengcheng Jiang, Jiashuo Sun, Zhiyi Shi, Haofei Yu, Jiawei Han, and Heng Ji. Learning to predict future-aligned research proposals with language models. *arXiv preprint arXiv:2603.27146*, 2026.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Sean Wu, Pan Lu, Yupeng Chen, Jonathan Bragg, Yutaro Yamada, Peter Clark, David Clifton, Philip Torr, James Zou, and Junchi Yu. Forecasting scientific progress with artificial intelligence. *arXiv preprint arXiv:2605.22681*, 2026.

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.

Yang Yang, Boyun Xu, Hao Fu, Jindong Li, Zining Zhong, Bowen Tian, Jiemin Wu, Menglin Yang, and Yutao Yue. HyBIRD: Hyperbolic bridge retrieval and diagnosis for methodology inspiration retrieval. *arXiv preprint arXiv:2606.28336*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Dataset Construction

A.1 Overview

Overall, INSPIRE comprises recent influential computer-science target papers resolved to arXiv. Each retained instance contains a solution-redacted problem brief, a target-specific public-record cutoff, and automatically graded pre-cutoff cited antecedents with tier and role annotations. Table 5 reports the train and evaluation statistics separately.

Table 5: **INSPIRE dataset statistics.** Train and evaluation partitions are reported separately. Antecedents are resolved pre-cutoff references cited by the later targets and automatically graded as T3 (most central), T2, or T1, operationalizing each target's realized citation lineage. The final columns summarize brief length and target-release years.

Split	Targets	Graded antecedents	Median/target	T3	T2	T1	Mean brief words	Target years
Train	1,549	10,268	7	430	2,905	6,933	183.4	2023 to 2024
Evaluation	476	3,574	10	179	1,244	2,151	179.3	2025 to 2026

A release manifest records the ordered train, development, evaluation, and diagnostic lists; SHA-256 hashes of instance files; construction settings; judge model identifiers; creation time; and the retrieval-response manifest.

A.2 Target Discovery

The builder issues Semantic Scholar bulk searches across computer-science and machine-learning subfields, deduplicates arXiv identifiers across topic queries, and uses arXiv metadata as the authority for title, abstract, category, and v1 date. Within each topic it ranks a candidate using

$$h(P) = \text{influential-citation-count}(P) + \frac{\text{citation-count}(P)}{100}. \tag{11}$$

This heuristic orders target discovery. A model-based filter removes surveys and reviews, after which the builder rejects papers without parseable arXiv HTML or at least two graded pre-cutoff antecedents. These choices define a population of recent, visible, sufficiently cited work.

A.3 Reference Grading

The target's HTML bibliography supplies candidate arXiv identifiers, which are normalized and batch-resolved. A reference with v1 date $t_r \geq D$ is stored as an excluded concurrent reference and never enters the label set. Tier judges receive the target abstract, candidate metadata, and target citation context when available. Their role-sensitive grading is motivated by prior evidence that citation contexts encode heterogeneous functions and levels of importance (Teufel et al., 2006; Valenzuela et al., 2015). For every candidate, the instance retains successful votes, consensus tier and role, rationale, and agreement count. Failed voters do not abort construction, so some records have fewer than three valid votes; the release reports those denominators. The active discriminativeness filter retains references within four years of the target. After T0 removal, references are sorted by configured gain and capped at ten. Older foundations and overflow remain retrievable but unlabeled. Predictions outside Y receive zero benchmark credit, though they may remain scientifically useful. The declared conceptual gains for T1/T2/T3 are $1/3/7$; infrastructural gains are $0.5/2/4$. Sensitivity analyses cover binary relevance, T2/T3-only scoring, and alternate monotone maps.

A.4 Brief Generation and Screening

The generator receives the target title and at most 1,800 characters of its abstract. It produces a detailed and a terse brief while being instructed to remove the target name, mechanism, and target-specific results. In the current evaluation snapshot, the detailed variant averages 179.5 whitespace-delimited words and the terse variant 103.2. A no-tool auditor sees the detailed brief and returns a target guess, confidence, and rationale. A positive guess with confidence at least 0.5 flags the instance; a separate lexical check records target-title token overlap but does not determine the split. The audit

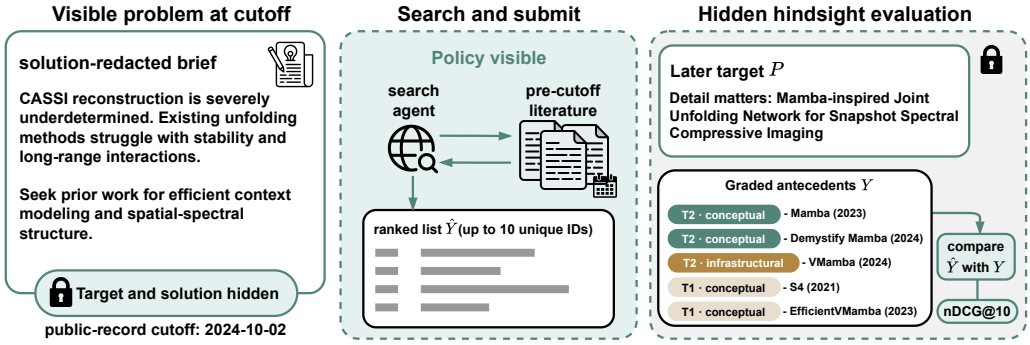

Figure 5: **Development example and automatic audit outcome.** The agent receives a shortened CASSI brief and public-record cutoff; the later target and colored antecedent labels are evaluator-only. The target-identity audit did not name the paper but inferred a likely Mamba/state-space mechanism, illustrating the residual mechanism-level leakage captured by the audit. Tiers and roles come from `arxiv:2501.01262`.

defines reproducible *automatic brief-clean* and flagged partitions; main results use the first and report all-target and flagged views alongside it. Because the audit is model-based, these slices characterize screening outcomes rather than independently validating every redaction (Zheng et al., 2023).

Figure 5 presents a development example retained from benchmark construction. The target-identity auditor did not name the paper, while its rationale inferred a likely Mamba or state-space mechanism. The example illustrates the task interface, hindsight scoring, and the residual mechanism-level clues captured by the automatic audit.

B Experimental Details

B.1 Shared Protocol

The run manifest contains the ordered target lists and instance hashes; target-specific cutoffs; search-backend and response-schema identifiers; system prompt, user prompt, tool schema, and scorer hashes; the fixed search budget; model revisions and decoding settings; software commit; and run creation time. The same interface implementation and budget are used for all five closed-source models, Qwen3.6-35B-A3B Base, and Qwen3.6-35B-A3B with SFT. A development replay checks date filtering, identifier normalization, output limits, and failure-as-zero scoring before a row is admitted to the main table. Every episode record stores model, target, protocol hash, ordered queries, ordered returned IDs, first exposure of each unique ID, tool errors, final ranking, terminal status, calls, tokens, latency, and cost where available. A comparability check verifies equality of target manifest, brief text, returned fields, cutoff logic, search budget, submission schema, and scorer across rows. The reporter emits expected, attempted, completed, submitted, and scored counts before computing any aggregate.

B.2 Model Evaluation

Each recorded checkpoint is evaluated once on every ordered evaluation target. The primary automatic-clean view is joined by target identifier before aggregation; the all-target and automatic-flagged views reuse the same episodes. Outputs are canonicalized to arXiv base IDs, duplicate submissions keep their first rank, and empty or invalid submissions remain zero. The main table contains no model whose policy-visible interface or search budget differs from the others. The model roster is Claude Opus 5, Claude Sonnet 5, GPT-5.6-Sol, Gemini 3.5 Flash, Gemini 3.1 Pro Preview, Qwen3.6-35B-A3B Base, and Qwen3.6-35B-A3B with hindsight-distilled SFT. Exact provider snapshot identifiers are part of the artifact manifest. If a provider revision cannot be pinned, its request time and returned revision metadata are retained and the reproducibility limitation is stated rather than replacing the row post hoc.

The target is the statistical unit. End-to-end comparisons use paired target-bootstrap intervals over the shared manifest, following paired resampling practice for system-level comparisons (Koehn, 2004). Automatic slices by split, tier, role, topic, label count, and training-antecedent overlap are descriptive and always report their denominator.

B.3 Trajectory Metrics

The trajectory reporter reconstructs the first 50 and unrestricted unique exposure sets, final submitted ranking, and first exposure turn for every ID. It reports relevant-item recall among the first 50 candidates, unrestricted exposure recall, zero-exposure targets, relevant-item retention conditional on positive exposure, exposed and selected gain, gain kept, rank quality, duplicate queries, malformed calls, stopping behavior, and unique IDs per call. Unit checks reconcile the final-stage mean with the main nDCG table and verify that selected gain lies between zero and exposed gain, and exposed gain lies below one, whenever only exposed IDs can be submitted. A cross-target join asks whether a target-specific missed antecedent appears in another trajectory produced through the same search backend. This analysis measures backend reachability; target-specific association and selection remain untested by the join.

B.4 SFT Configuration

The SFT student is Qwen3.6-35B-A3B. For automatic-clean training instances, the teacher receives the visible brief plus a private block containing up to 12 highest-tier antecedent titles and truncated abstracts, without IDs. It generates four candidate trajectories. Each candidate is replayed through the ordinary deployment prompt and search interface, then scored and scanned for private-string copying, paper-identifying queries, post-cutoff results, malformed actions, and unsurfaced submissions. At most one highest-scoring replay-valid trajectory above a development-frozen threshold is retained per target.

The final corpus manifest records source instance, teacher revision, random seed, score, action count, rejection flags, transcript hash, and selected status. The SFT configuration uses action-token loss, three epochs, LoRA rank 16, scale 32, zero dropout, and learning rate 5×10^{-6}; LoRA supplies the parameter-efficient adapter design (Hu et al., 2022). The base-checkpoint hash, tokenizer and chat template, target modules, batching, optimizer, schedule, precision, maximum sequence length, and software/hardware environment are also recorded. Base and SFT evaluation differ only by adapter.

C Hindsight SFT Details

This appendix provides the demonstration-construction, information-boundary, and uncertainty details supporting the matched study in Section 4.4.

Training protocol. The student and matched base model are **Qwen3.6-35B-A3B**, and supervised fine-tuning is the only training stage. For targets in the disjoint training pool that pass the same automatic screen, a frozen GPT-5.5 teacher (OpenAI, 2026a) may inspect training-only antecedent tiers, titles, and abstracts, but never their identifiers. Its proposed actions are replayed through the ordinary tools, and an identifier may be submitted only after the replay exposes it. The highest-scoring replay-valid trajectory above a development-frozen threshold becomes the demonstration; automatic scans reject target or antecedent copying, post-cutoff results, and unsurfaced submissions. The student receives only the deployment transcript—the redacted brief, prior actions, and ordinary search results—while private teacher text and demonstration-selection scores are removed. Privileged hindsight therefore selects successful training behavior without expanding the student's test-time observations or actions, following the train-only information boundary of privileged learning and hindsight replay (Vapnik & Vashist, 2009; Andrychowicz et al., 2017). Training uses standard LoRA SFT (Hu et al., 2022); Appendix B records the corpus, checkpoint, and optimization configuration.

Matched evaluation. At evaluation, Qwen3.6-35B-A3B Base and the SFT model use the same held-out targets, interface, cutoff, search budget, decoding, failure policy, and scorer; the adapter is the only changed factor. Table 4 reports the stage-wise and end-to-end point estimates, while Figure 6 shows the paired target-level score distributions. Final nDCG@10 increases from 0.082 to 0.166, a

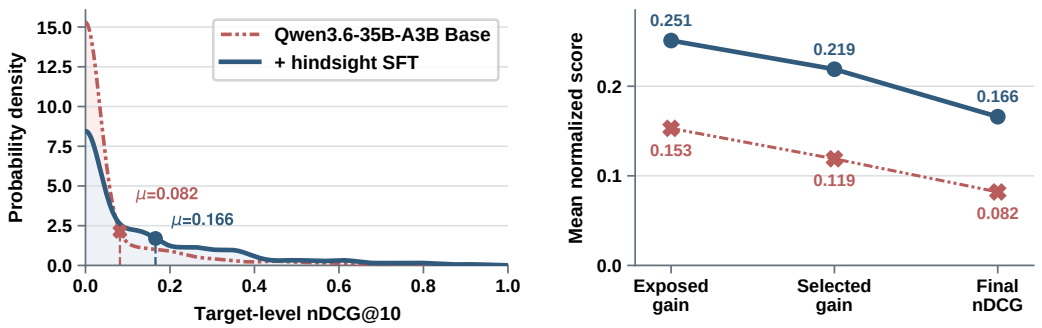

Figure 6: **Held-out response to benchmark-derived SFT.** The left panel compares boundary-reflected Gaussian kernel densities of matched Base and SFT failure-as-zero nDCG@10 values, including the probability mass at zero. The right panel traces aggregate Exposed gain, Selected gain, and final nDCG@10 under the same held-out protocol. Improvements at all three points indicate that the controlled intervention changes both candidate discovery and use of the discovered evidence.

paired difference of 0.084 with a 95% paired target-bootstrap interval of $[0.055, 0.115]$. This interval quantifies variation over the shared target set; training-seed variation remains outside the reported uncertainty when only one frozen adapter is available.

D Exposure Gain over Search

The aggregate exposure funnel measures how much graded gain is available at the end of a trajectory, but it does not show when that gain enters the candidate set. We therefore compute a prefix-based oracle diagnostic. For condition c, let $E_i^{(c)}(p)$ contain the unique papers exposed by target i up to normalized search progress $p \in [0,1]$, where $E_i^{(c)}(0) = \varnothing$ and $E_i^{(c)}(1) = E_i^{(c)}$. Using the target-level oracle in Equation 7, the progress curve is

$$G^{(c)}_{\text{exposed}}(p) = \frac{1}{N_c} \sum_{i=1}^{N_c} \phi_i\left(E_i^{(c)}(p)\right). \tag{12}$$

At each checkpoint, this quantity asks how well the graded antecedents exposed so far could be ranked if their gains were known and all ungraded papers were discarded. The curves are consequently nondecreasing by construction, and their endpoints reproduce the Exposed gain values in Tables 3 and 4. Normalizing progress within each trajectory makes timing profiles comparable despite different trajectory lengths, while absolute calls and returned-paper counts can still differ across systems. The oracle retrospectively describes the value accumulated along each observed path; budget interventions require separate experiments.

Figure 7 reveals substantial differences not only in final exposure, but also in the rate at which useful evidence becomes available. Claude Opus 5 separates early, reaching approximately 0.18 oracle nDCG@10 after 20% of its normalized search and 0.36 at the halfway point before ending at 0.480. Its advantage therefore reflects a sustained accumulation of graded antecedents rather than a single late discovery. Claude Sonnet 5 and GPT-5.6-Sol, by contrast, follow nearly overlapping trajectories through much of the latter half of search and finish at 0.323 and 0.326. Their similar exposure potential, together with Sonnet's higher final nDCG@10 of 0.189 versus 0.164, strengthens the funnel result that retention and ordering can separate systems even when search supplies comparable candidate sets.

No curve reaches an early plateau. At 50% normalized progress, the systems have accumulated approximately 64–74% of the oracle gain available at the end of their own trajectories, leaving roughly one quarter to one third of eventual exposed gain to arrive during the second half. Later search is therefore not uniformly redundant, even for the strongest model. At the same time, the low curves for Gemini 3.1 Pro Preview and Qwen3.6-35B-A3B Base remain low throughout search, despite continuing to increase. This pattern indicates that their exposure deficit is not confined to premature stopping: relative to the stronger systems, useful antecedents enter the candidate set at

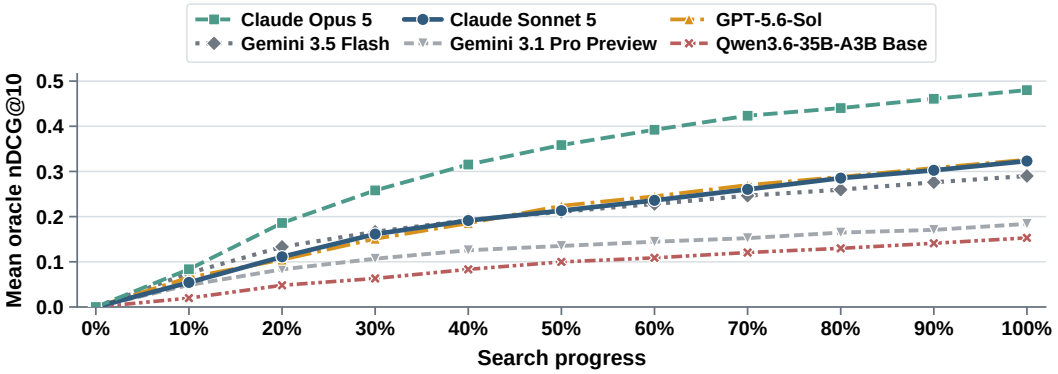

Figure 7: **Oracle nDCG@10 accumulated over normalized search progress.** At each checkpoint, graded antecedents in the observed exposure prefix are ordered by their hindsight gains. The endpoint of each curve is therefore the best nDCG@10 supported by the complete exposed set, before losses from final selection and ordering.

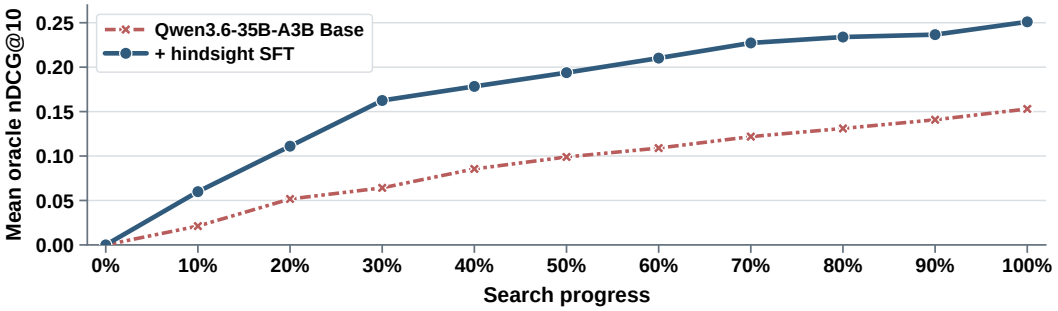

Figure 8: **Search-progress profile for the matched Base and hindsight-SFT conditions.** The oracle is recomputed on each condition's cumulative exposure prefix under the shared held-out protocol. Curve separation measures when the adapter makes additional graded gain available, independently of subsequent selection and submitted ordering.

a lower rate across most of the observed trajectory. The curves localize this deficit along realized behavior; experiments targeting query choice, exploration strategy, parametric knowledge, and the retrieval backend can isolate its mechanism.

Figure 8 shows that the hindsight-SFT condition dominates the Base exposure curve at every nonzero checkpoint. The separation emerges early: by 30% progress, SFT has accumulated approximately 0.16 oracle nDCG@10, already slightly more than the Base model obtains from its complete trajectory (0.153). The absolute gap is about 0.10 at this checkpoint and remains of similar magnitude thereafter, ending at 0.251 versus 0.153. Thus, the exposure improvement reported in Table 4 is not explained only by a small number of discoveries near the end of search. Instead, the observed SFT policy reaches higher-value candidate sets earlier and preserves that advantage throughout the trajectory.

Both conditions nevertheless continue to acquire useful evidence after the midpoint. From 50% to 100% progress, oracle nDCG@10 increases by approximately 0.06 for both Base and SFT. Hindsight SFT therefore shifts the entire availability profile upward without exhausting the value of continued exploration. The remaining difference between the SFT endpoint oracle of 0.251 and its final nDCG@10 of 0.166 also confirms that improved exposure leaves downstream headroom: some available gain is still lost during selection and ordering. Taken together with the matched funnel analysis, the progress curves show that the adapter changes observed search behavior early enough to expose better antecedents while also improving their eventual use. Identifying the learned query or appraisal decisions responsible for this shift requires a targeted intervention.

E PROMPT TEMPLATES

This section reports the natural-language templates used by the benchmark implementation. Text in braces denotes a value inserted at runtime. System and user messages are identified explicitly; tool results and the function schemas supplied alongside the prompts are not natural-language templates and are therefore not reproduced here.

Full-tool benchmark agent. The runner uses this system message when the agent can search, fetch a paper's full text, and submit a ranked answer.

```
You are a research assistant operating inside a point-in-time
literature search environment frozen to a cutoff date. You are given a
research brief describing an open problem as it stood before any
solution existed. Your job is to find the prior papers most worth
reading to make progress on that problem.

Rules:
- Use ONLY the provided tools. You have no web access.
- Search with varied keyword queries, read abstracts and full text as
needed.
- Then call submit_answer with a ranked list of paper ids (most
load-bearing first), each with a one-line rationale. Aim for 5-10
papers.
- Prioritize papers that supply the core mechanism or concept the
solution would build on, over merely topical or background papers.
- You MUST finish by calling submit_answer.
```

Solution-redacted brief generation. The pinned Claude Opus 4.8 construction judge receives the target metadata and generates the detailed and terse brief variants.

```
[System]
You write research briefs for the INSPIRE benchmark. A brief states an
open research problem AS IT STOOD BEFORE any solution existed -- the
motivation, the limitations of prior approaches, and the desired
properties of a solution. It must NOT reveal the solution. Redact: the
target paper's method name, the specific solution mechanism, and any
result/benchmark/dataset names unique to the target. Write in the first
person plural ('we want...'). End with: 'What prior work should we read
to make progress on this problem?'

[User]
TARGET PAPER
 title: {target_title}
 abstract: {target_abstract}
```

Prior-reference tier and role judging. Claude Opus 4.6, GPT-5.5, and Gemini 3.1 Pro Preview each receive the same system and user templates. Candidate abstracts are truncated to 600 characters, the target abstract to 1,500 characters, and the target full text is included only when available. The global-citation field is likewise conditional.

```
[System]
You are an expert research-literature judge for the INSPIRE benchmark.
Given a TARGET paper's abstract and a candidate PRIOR reference, you
decide how load-bearing that reference was for the target's core
contribution.

Tier scale (how essential to the target's central mechanism/idea):
  T3 = backbone: the target directly builds its core method on this
  work.
  T2 = component: supplies a key concept/tool the method adapts or
  compares against.
```

```
  T1 = background: foundational or comparison context, not the
  mechanism.
  T0 = incidental: tangential or boilerplate citation, non-relevant.
Type:
  conceptual = supplies an idea/mechanism/insight the target reasons
  with.
  infrastructural = supplies tooling/architecture/data/benchmark the
  target uses.

CRITICAL -- reward SPECIFIC, NON-OBVIOUS antecedents over ubiquitous
foundations. A famous, very-highly-cited base paper (a canonical
architecture, a seminal method everyone in the field cites) is a SAFE,
low-value seed: any keyword search trivially surfaces it, so it does
not test retrieval skill. Reserve T3/T2 for the close, specialized
predecessors a strong researcher would find and a weak one would miss;
push generic foundations toward T1/T0 even if technically cited. When
the TARGET's full text is provided, ground each decision in HOW the
target actually uses the reference in-context (the section and
sentences that cite it) rather than guessing from titles alone; a
reference whose method/result the target leans on in its core sections
outranks one mentioned only in passing.

[User]
TARGET PAPER
 title: {target_title}
 abstract: {target_abstract[:1500]}

[Included only when available:]
TARGET FULL TEXT (use to see how each reference is cited in-context):
{target_fulltext}

CANDIDATE PRIOR REFERENCES (cited by the target):
[0] id={reference_0.arxiv_id} ({reference_0.date})
[global_citations={reference_0.cited_by_count}]
    title: {reference_0.title}
    abstract: {reference_0.abstract[:600]}
...

For EACH reference, output its tier and type. Respond with ONLY a JSON
array, one object per reference, in the same order, each as:
{"index": <int>, "arxiv_id": "<id>", "tier": "T0|T1|T2|T3", "type":
"conceptual|infrastructural", "rationale": "<one sentence>"}
```

Privileged trajectory teacher. The GPT-5.5 teacher system message begins with below verbatim and appends the instructions and optional private ingredient block below. In the reported privileged setting, up to 12 references are sorted by tier and discriminativeness; each abstract is whitespace-normalized and truncated to 1,200 characters.

```
{SEARCH_ONLY_SYSTEM_PROMPT}

You are writing a high-quality tool-use demonstration for a smaller
student model. Make exactly ONE function call per response. Use the
research brief to devise effective searches. Operate through
search_papers and submit only IDs actually returned by search.

Optional private candidate ingredients are provided below. Use them
only to improve search strategy; never copy or submit an ID unless
search_papers returned it.

PRIVATE CANDIDATE INGREDIENTS (teacher-only; not part of the student
prompt):
1. [{seed_1.tier}/{seed_1.type}] {seed_1.title}
Abstract: {seed_1.abstract[:1200]}
```

```
...

[User]
RESEARCH BRIEF:
{brief_text}

Find and submit your ranked reading list.
```