

HarnesS-1: Reinforcement Learning for Search Agents with Stateful Harnesses

Pengcheng Jiang Zhiyi Shi Kelly Hong[†]◇ Xueqiang Xu Jiashuo Sun
Jimeng Sun Hammad Bashir[◇] Jiawei Han
University of Illinois at Urbana-Champaign [†]UC Berkeley ◇Chroma

Abstract

Search agents are often trained as policies over growing transcripts: the model must decide how to search while also remembering what it has seen, which evidence is useful, which constraints remain open, and which claims have actually been checked. We argue that this formulation puts too much routine state management inside the policy: reinforcement learning is forced to optimize both semantic search decisions and recoverable bookkeeping that the environment can maintain more reliably. We introduce HarnesS-1, a 20B search agent (retrieval subagent) trained with reinforcement learning inside a stateful search harness. The harness maintains environment-side working memory, including a candidate pool, an importance-tagged curated set, compact evidence links, verification records, compressed and deduplicated observations, and budget-aware context rendering. The policy retains the semantic decisions: what to search, which documents to keep or discard, what to verify, and when to stop. Across eight retrieval benchmarks spanning web, finance, patents, and multi-hop QA, HarnesS-1 achieves 0.730 average curated recall, outperforming the next strongest open search subagent by +11.4 points and remaining competitive with much larger frontier-model searchers. Its gains are especially strong on held-out transfer benchmarks, suggesting that reinforcement learning over explicit search state can produce retrieval behaviors that generalize beyond the training domains.

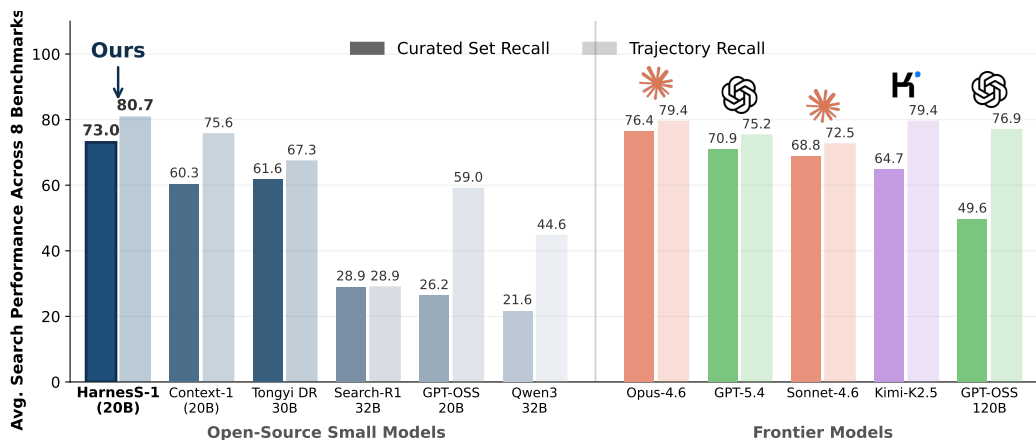


Figure 1: **Averaged performance across eight challenging search benchmarks.** Each method reports: Curated Set Recall and Trajectory Recall (documents encountered anywhere in the episode). HarnesS-1 is our 20B open search agent trained with a stateful harness; it substantially improves over open search agents and remains competitive with much larger frontier-model searchers.

1 Introduction

Search agents are usually described as language models that call retrieval tools: given a question, they issue queries, read returned evidence, decide what is still missing, and return documents to a downstream answerer. This view follows a broad line of work on reasoning-and-acting agents, iterative retrieval, active retrieval, and tool-use training [51, 40, 17, 33, 15]. It captures the visible behavior of the agent, but not the search state that must be built along the way. A successful multi-turn search episode requires the agent to remember which documents have been seen, which candidates are worth keeping, which constraints remain uncovered, which entities connect separate pieces of evidence, and which claims have actually been checked against source text.

This distinction becomes especially important for reinforcement learning. Recent work has shown that LLMs can be trained to interact with search engines and retrieval systems through RL, improving query generation, multi-turn search, and downstream retrieval utility [14, 18, 16, 45]. Yet most retrieval-agent training still asks the model to learn two different things at once: semantic search behavior and routine state management. As the transcript grows, the policy must reconstruct the useful state of the search from raw observations. This is inefficient, and it can make learning poorly conditioned. Hard queries may give rollouts nearly identical empty-set rewards; larger tool vocabularies may collapse to repeated search calls; and cross-document structure may be present in the transcript but too diffuse to use reliably. When the final curated set is empty or wrong, the reward often does not reveal whether the failure came from bad search, forgotten evidence, missing verification, or poor curation.

Stateful cognitive offloading. A retrieval policy should make semantic decisions over explicit search state: what to search, which documents to keep, what claim to verify, and when to stop. The harness should maintain the recoverable state around those decisions: candidate pools, curated evidence, cross-document links, verification records, and context-budget summaries. This gives RL a stable interface for improving search behavior, rather than asking the model to rediscover bookkeeping from an append-only transcript.

This view changes what tool orchestration means. Prior work on agent-computer interfaces and software-engineering agents has shown that the interface around a language model can strongly shape agent behavior and performance [49, 48]. In retrieval, recent systems have begun to treat context management as part of the search process, for example through context pruning or self-editing search [2]. In `HarnesS-1`, the interface itself is part of the method. The agent acts over a persistent retrieval state maintained by the environment, and each observation renders that state back to the policy. The policy still decides what to search, which documents to promote, what to verify, and when the evidence is sufficient. The harness carries the state maintenance that would otherwise consume the policy’s context and optimization budget.

We instantiate this principle in `HarnesS-1`, a 20B search agent built on `gpt-oss-20b` [29]. At each turn, the model observes recent tool results together with a rendered `WORKINGMEMORY` containing the candidate pool, curated set, search history, evidence graph summary, verification state, and context-budget marker. The model’s actions edit this state rather than merely extending a transcript: `curate` adds, removes, and importance-tags documents; `verify` checks policy-written claims against remembered documents; and `review_docs` re-renders previously seen documents without a new corpus call. Search observations are compressed and deduplicated before reaching the prompt. The same state renderer is used for teacher data generation, SFT replay, RL rollout, and evaluation.

A richer harness does not automatically yield an environment suitable for policy training. If every rollout starts with an empty curated set, early rewards can be indistinguishable. If derived state is too verbose, it competes with evidence for context. If reward only values discovery, the policy can ignore curation and verification and converge to repeated search. We therefore identify three requirements for stateful search harnesses: **warm-started curation**, implemented by auto-seeding a tentative curated set from the first successful search; **compact derived-state rendering**, implemented through importance tags, evidence-graph summaries, and verification records; and **diversity-preserving incentives**, which encourage a rhythm of searching, curating, reviewing, and verifying.

We evaluate `HarnesS-1` on eight retrieval benchmarks spanning web, finance, patents, and long-context multi-hop search. It reaches 0.730 average curated recall and improves downstream answer

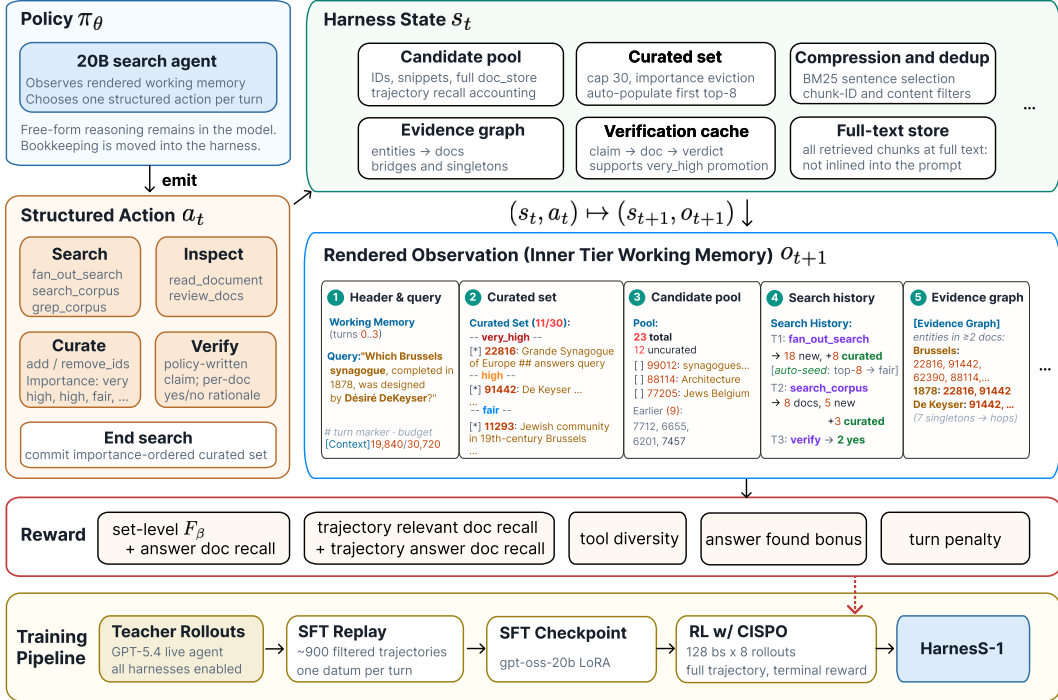


Figure 2: **Overview of HarnessS-1.** The policy makes semantic decisions over search, inspection, curation, verification, and termination. The harness maintains the state around those decisions: document pool, importance-tagged curated set, evidence graph, verification records, compression and deduplication, result summaries, and budget markers. The same state interface is used for teacher rollouts, SFT replay, CISPO RL, and evaluation.

accuracy when its curated set is passed to frozen frontier generators. The gains are not confined to the training domains: the same operations over explicit search state improve retrieval across specialized corpora, open-web tasks, and held-out multi-hop QA benchmarks.

Contributions. First, we formulate harness design as a central problem in retrieval-agent RL and introduce stateful cognitive offloading as a principle for assigning search state between policy and environment. Second, we introduce HarnessS-1, a 20B search agent with an importance-tagged curated set, compact evidence graph, verification records, compressed and deduplicated observations, and budget-aware context rendering. Third, we identify three trainability requirements for stateful search harnesses: warm-started curation, compact derived-state rendering, and diversity-preserving incentives. Fourth, we show that stateful harnessing improves retrieval quality and modular RAG answer accuracy across eight benchmarks. We will release the model weights, harness code, data-generation pipeline, and RL recipe.

2 HarnessS-1

HarnessS-1 is a search agent that produces a ranked set of documents for a downstream answering model. It runs inside a state-machine harness centered on a per-episode WORKINGMEMORY. At each turn, the harness renders compact search state together with recent actions and observations; the model emits one structured Harmony action; and the harness executes the action, updates the state, and renders the next observation.

The harness does not make retrieval decisions for the model. Instead, it maintains the recoverable state around those decisions. The policy still decides what to search, which documents to inspect or keep, what claims to verify, and when to stop. Each action applies a transition $(s_t, a_t) \mapsto (s_{t+1}, o_{t+1})$ over the harness state summarized in Table 1. This is the sense in which the harness is *stateful*: tool outputs do not merely produce text for the next prompt; they update persistent search state that future actions can inspect, edit, or verify.

Table 1: State notation and offloaded bookkeeping in the HarnesS-1 harness. The harness maintains and renders state; the policy retains the semantic choices over that state.

State	Harness-maintained bookkeeping	Policy decision preserved
P_t	Candidate pool after compression and deduplication	Which candidates to inspect, read, or curate
C_t, I_t	Curated output set and importance tags, warm-started by auto-seeding	Which documents to add, remove, promote, or demote
D_t	Full-text memory for retrieved documents	Which seen documents to revisit through <code>review_docs</code>
G_t	Evidence graph over entities, dates, and documents	Which bridge, singleton, or relation to pursue next
V_t	Verification records for policy-written claims	Which claim to check and which documents to test
H_t	Search history and result summaries	When to diversify, backtrack, or continue a search thread
B_t	Budget-safe renderer and context marker	When to search, read, summarize, or stop

WORKINGMEMORY has two tiers. The prompt-facing tier renders compact state: P_t , C_t , I_t , G_t , V_t , H_t , and B_t . The outer tier D_t stores full text and metadata for retrieved chunks, which the policy can revisit through `review_docs` or `read_document`. Thus the prompt carries actionable search state rather than the full retrieval transcript.

Full state-transition algorithms and trajectory-level examples are given in Appendices H and Q.

2.1 Policy actions as edits over working memory

HarnesS-1 exposes five classes of actions: retrieval, memory inspection, curation, verification, and termination. Retrieval actions bring new evidence into the episode. `fan_out_search` runs up to five diverse queries in parallel, `search_corpus` performs targeted hybrid search, `grep_corpus` performs exact pattern matching, and `read_document` returns full text for a known document ID. Their outputs are not simply appended to the prompt. After compression and deduplication, they update the candidate pool P_t and the full-text store D_t .

Curation. The central state-editing action is `curate`, which updates the final output set C_t and its importance map I_t . The action specifies documents to add, documents to remove, and importance tags for added documents. Tags take one of four values: `very_high`, `high`, `fair`, or `low`. The curated set is capped at $M=30$ documents; when the cap is reached, the harness evicts the lowest-importance documents first (Appendix K). This gives the policy an explicit language for confidence and priority, while the harness handles the mechanical capacity constraint.

Auto-seeding. A blank curated set provides little learning signal on hard queries, since many early rollouts terminate with the same empty output. To avoid this, the first successful search automatically seeds C_t with the top $k=8$ reranked results from P_t , each assigned $I_t(d) = \text{fair}$. The policy must then refine this tentative set by promoting strong documents and removing weak ones. Auto-seeding does not decide relevance for the model. It changes the initial state from construction from scratch to refinement, so that early trajectories differ in how they edit candidates. The BC+ and SEC trajectories in Appendices Q.1 and Q.2 show this pattern: the first search seeds plausible candidates, and later `curate` calls `promote` documents that survive reading, exact search, or verification.

Verification and memory review. `verify` lets the policy write a claim and select remembered documents to test; the harness checks support against text in D_t and records yes/no judgments with rationales in V_t . Thus the policy decides what requires verification, while the harness performs the mechanical support check. `review_docs` re-renders documents already stored in D_t , allowing the policy to revisit evidence before promoting, demoting, removing, or verifying it without issuing another corpus search.

2.2 Derived-state rendering: evidence graph, compression, and deduplication

The harness does not only store retrieved text. It derives compact signals from the search history and renders them into each observation, so the policy sees a structured search state rather than a raw accumulation of tool outputs.

Evidence graph. The evidence graph G_t summarizes cross-document structure. Each chunk entering P_t is scanned with a lightweight regex extractor for three entity classes: multi-word capitalized proper nouns, four-digit years or decades, and numeric dates. The harness maintains two maps, `entity_to_docs` and `doc_to_entities`, and renders a compact block with frequent entities, bridge documents, and singleton entities. Frequent entities are the top entities by document count. Bridge documents contain two or more frequent entities and are natural candidates for `verify` or

high-importance promotion. Singleton entities appear in only one document and often provide leads for follow-up search.

The example illustrates how the graph makes evidence overlap visible. Document 22816.0 connects all three rendered entities, while 91442.3 connects two of them. The policy can therefore treat these documents as candidate bridges without rereading the full transcript. More generally, the rendering changes questions such as “what have I seen about entity X ?” from a full-context reread into a direct lookup in working memory. The harness builds the graph with one regex pass per chunk and a top- k scan at render time; the policy decides which bridge to verify, which singleton to pursue, and which relation matters for the query.

[Evidence Graph] Entities appearing in multiple docs (bridges):
 Brussels: 22816.0, 91442.3, 62390.5
 1878: 22816.0, 11293.2
 Grande Synagogue: 22816.0, 91442.3
 (2 singleton entities, potential hops)

Result summaries and budget markers. After each retrieval or memory action, the harness records the tool used, the number of returned and novel documents, updates to P_t and C_t , and the current context budget B_t . Thus the policy observes search progress and remaining budget as explicit state. **Sentence-BM25 compression.** To prevent search observations from dominating the context, the harness compresses outputs from `search_corpus`, `fan_out_search`, and `grep_corpus` by scoring sentences with BM25 against the query and rendering the top $K=4$ sentences in original order. Explicit `read_document` calls still return full text from D_t . **Two-level deduplication.** Harness-1 removes repeated evidence by both chunk ID and content fingerprint. Near-duplicates are omitted from the rendered state, while reward accounting still credits all relevant evidence encountered during the trajectory. The same budgeted context builder is used for teacher generation, SFT replay, RL rollout, and evaluation. It combines the system prompt, rendered working memory, recent actions and observations, and result summaries, while truncating older reasoning and clipping long observations as needed.

2.3 Training: SFT and RL

Training follows the same division of labor as the harness: SFT teaches the model to operate the interface, while RL improves search decisions over the state maintained by that interface.¹

SFT for harness operation. The role of SFT is deliberately narrow. Since the harness carries much of the bookkeeping load, the supervised prior only needs to teach tool-call format, the search→curate rhythm, importance-tagging discipline, memory review, and the `verify-before-promote` norm. A single teacher, GPT-5.4, runs as a live agent inside the full Harness-1 harness, using the same tools and working-memory rendering later shown to the student. Turn-level guidance encourages curation after search, verification before high-confidence promotion, backtracking after low-yield searches, and early termination once the curated set is sufficient. We keep trajectories that are format-valid, return at least one document, and achieve final curated recall ≥ 0.10 . Each retained trajectory is replayed through the same `WORKINGMEMORY` and context builder used by RL, then expanded into one supervised datum per turn. SFT runs on `gpt-oss-20b` via LoRA with rank 32 for 3 epochs; the step-550 checkpoint initializes RL.

RL over full search episodes. Starting from the SFT checkpoint, we train with on-policy CISPO [3, 19] and within-group advantage normalization on SEC training queries. RL uses full-trajectory rollouts, terminal-only reward, a 40-turn cap, and no KL anchor. Groups whose rollouts receive identical rewards are dropped from the gradient. Full rollout scale and optimization hyperparameters are in Appendix D.

Reward. The terminal reward combines search-quality signals, answer-evidence shaping, a tool-diversity bonus, and a turn penalty. Episodes that terminate with an empty curated set short-circuit to $\pi_\emptyset = -0.2$; otherwise,

$$\mathcal{R} = \underbrace{w_F F_\beta}_{\text{set quality}} + \underbrace{w_\tau \rho_\tau}_{\text{trajectory coverage}} + \underbrace{w_A \rho_A + w_{\tau A} \rho_{\tau A}}_{\text{answer evidence}} + \underbrace{B_A \mathbf{1}[\rho_A > 0]}_{\text{answer bonus}} \\ + \underbrace{w_{\text{div}} \min(\nu/\nu_0, 1)}_{\text{tool diversity}} - \underbrace{w_{\text{miss}} (\rho_{\tau A} - \rho_A)_+}_{\text{answer miss}} - \underbrace{\pi_{\text{turn}}(t)}_{\text{turn penalty}}.$$

¹We use Tinker (<https://thinkingmachines.ai/tinker/>) for model training.

Table 2: **Search quality across benchmarks.** We report curated-set Recall, Final-Answer Recall, and Trajectory Recall. Green and blue rows denote the overall best models among open-source small models and frontier models, respectively. Results are averaged over three runs.

Model	Metric	BC+	Web	Patents	SEC	LongSeal	Seal0	FRAMES	HotpotQA
Open-source small models									
HarnesS-1 (20B)	Recall	0.584	0.787	0.890	0.589	0.891	0.551	0.710	0.841
	Final-Answer Recall	0.667	0.845	0.890	0.659	0.891	0.551	0.710	0.841
	Trajectory Recall	0.665	0.881	0.973	0.677	0.944	0.708	0.749	0.860
Context-1 (20B)	Recall	0.500	0.689	0.857	0.467	0.563	0.367	0.638	0.746
	Final-Answer Recall	0.618	0.775	0.857	0.498	0.563	0.367	0.638	0.746
	Trajectory Recall	0.567	0.823	0.975	0.507	0.858	0.697	0.787	0.835
GPT-OSS-20B	Recall	0.109	0.174	0.393	0.133	0.253	0.250	0.391	0.395
	Final-Answer Recall	0.156	0.252	0.393	0.166	0.253	0.250	0.391	0.395
	Trajectory Recall	0.246	0.512	0.742	0.252	0.816	0.632	0.702	0.819
Tongyi DeepResearch 30B	Recall	0.381	0.607	0.780	0.419	0.855	0.577	0.561	0.751
	Final-Answer Recall	0.428	0.679	0.780	0.433	0.855	0.577	0.561	0.751
	Trajectory Recall	0.413	0.722	0.780	0.478	0.855	0.631	0.654	0.852
Search-R1 (32B)	Recall	0.053	0.135	0.405	0.109	0.514	0.312	0.329	0.454
	Final-Answer Recall	0.074	0.240	0.405	0.155	0.514	0.312	0.329	0.454
	Trajectory Recall	0.053	0.135	0.405	0.109	0.514	0.312	0.329	0.454
Qwen3 (32B)	Recall	0.072	0.132	0.257	0.058	0.427	0.279	0.164	0.342
	Final-Answer Recall	0.092	0.211	0.257	0.072	0.427	0.279	0.164	0.342
	Trajectory Recall	0.154	0.410	0.577	0.215	0.704	0.568	0.402	0.535
Frontier models									
Opus-4.6	Recall	0.619	0.819	0.915	0.589	0.823	0.636	0.814	0.893
	Final-Answer Recall	0.736	0.928	0.915	0.696	0.823	0.636	0.814	0.893
	Trajectory Recall	0.634	0.832	0.939	0.601	0.876	0.682	0.878	0.910
GPT-5.4	Recall	0.443	0.800	0.909	0.452	0.824	0.555	0.815	0.870
	Final-Answer Recall	0.499	0.858	0.909	0.465	0.824	0.555	0.815	0.870
	Trajectory Recall	0.476	0.821	0.919	0.464	0.885	0.691	0.865	0.898
Sonnet-4.6	Recall	0.505	0.744	0.832	0.411	0.783	0.577	0.787	0.863
	Final-Answer Recall	0.597	0.892	0.832	0.480	0.783	0.577	0.787	0.863
	Trajectory Recall	0.535	0.767	0.845	0.421	0.840	0.658	0.859	0.878
Kimi-K2.5	Recall	0.593	0.707	0.833	0.454	0.661	0.455	0.697	0.776
	Final-Answer Recall	0.670	0.843	0.833	0.478	0.661	0.455	0.697	0.776
	Trajectory Recall	0.739	0.878	0.945	0.549	0.911	0.655	0.840	0.837
GPT-OSS-120B	Recall	0.508	0.478	0.570	0.395	0.464	0.400	0.495	0.657
	Final-Answer Recall	0.641	0.664	0.570	0.504	0.464	0.400	0.495	0.657
	Trajectory Recall	0.682	0.796	0.824	0.488	0.914	0.689	0.863	0.897

For any successfully formatted episode, the reward is clipped below by $\mathcal{R} \geq 10^{-3}$. Here ρ_τ is trajectory recall, ρ_A is curated final-answer recall, $\rho_{\tau A}$ is trajectory final-answer recall, ν is the number of distinct tools used, and ν_0 is the diversity target. F_β uses $\beta=2$, weighting recall four times more than precision. The reward separates discovery from selection: trajectory terms credit evidence found anywhere in the pool, curated terms credit evidence selected for the final output, and the answer-miss penalty penalizes trajectories that found answer evidence but failed to promote it. Full hyperparameters and code variables are listed in Appendix D.

3 Experiments

3.1 Experimental Setup

Benchmarks & Data. We evaluate on eight retrieval benchmarks: BrowseComp+ [6], *Web* synthetic, *Patents* from USPTO office actions, *SEC* filings [2], LongSealQA and Seal0QA [31], FRAMES [20], and HotpotQA [50]. They span web, finance, patents, and multi-hop QA, and use both Chroma-backed corpora and a shared Serper+Jina web backend. HarnesS-1’s SFT data spans four benchmark families (BC+, Web, Patents, SEC), while the RL stage optimizes only on SEC data. We therefore separate the four source-family benchmarks (BC+, Web, Patents, SEC) from four held-out transfer benchmarks not used for HarnesS-1 SFT or RL (LongSealQA, Seal0QA, FRAMES, HotpotQA). Per-benchmark statistics are in Appendix E. The raw SFT quotas

are BC+/300, SEC/250, Patents/150, Web/150, Web-simple/75, and SEC-simple/75; after filtering, 899 trajectories remain. We use the training split of SEC² (3,453 queries in total) for RL training.

Search-quality metrics. We evaluate search quality by evidence coverage. *Recall* measures coverage of all annotated relevant documents in the final curated set. *Trajectory Recall* measures the same coverage over every document encountered during the episode, before final selection. *Final-Answer Recall* measures coverage of the gold answer documents only. Formal definitions, including the distinction between answer documents and supporting documents, are given in Appendix N. We focus on recall-oriented metrics since the search agent is designed to collect evidence needed for answering. Precision and F_1 are secondary because curated-set size is part of the agent interface.

Evaluation protocol and baselines. We evaluate on the benchmark splits summarized in Table 7. All methods use the same retrieval primitives, web backend, Qwen3-Reranker-8B reranker, and final budget of at most 30 returned documents, following Context-1 [2]. HarnesS-1 runs at temperature 1.0 with a 40-turn cap. Context-1 and proprietary frontier LLMs (GPT-5.4, Sonnet-4.6, Opus-4.6, Kimi-K2.5), used as zero-shot retrievers, run under the Context-1 harness at temperature 1.0 with a 64-turn cap; Appendix P quantifies the harness contribution alone (a +4-point recall lift on the same LLM). Open-model baselines include base gpt-oss-20b/gpt-oss-120b, Qwen3-32B, Context-1, Search-R1 32B [18], and Tongyi DeepResearch 30B [39]. For Search-R1 and Tongyi DeepResearch, we use their released agent harnesses while standardizing the external search and reranking APIs. Because these harnesses do not natively produce a curated set capped at 30 documents, we apply reranking to obtain the final top-30 documents; details are given in Appendix N. Full evaluation details are in Appendices N and E.

3.2 Results

Table 2 reports Recall, Final-Answer Recall, and Trajectory Recall on all eight benchmarks. HarnesS-1 is the strongest open retrieval subagent we evaluate, reaching 0.730 average curated recall and improving over the next open subagent, Tongyi DeepResearch 30B, by +11.4 points. It is also competitive with much larger frontier retrievers: it has higher average curated recall than GPT-5.4, Sonnet-4.6, Kimi-K2.5, and GPT-OSS-120B under this protocol, with Opus-4.6 the only frontier retriever ahead on average.

Transfer is the clearest signature of the mechanism. Absolute recall is useful, but the structure of the gains is more diagnostic. HarnesS-1 is not trained from a single dataset: its SFT stage uses BC+, Web, Patents, and SEC trajectories. The stricter transfer test is therefore the four benchmarks excluded from both HarnesS-1 SFT and RL. If the gains came mainly from replaying source-family behaviors, the advantage should shrink on these held-out tasks. We observe the opposite.

A standard machine-learning prior would expect gains to be strongest on benchmarks closest to the SFT and RL data. Across the four source-family benchmarks (BC+, Web, Patents, SEC), HarnesS-1 gains +6.2, +9.8, +3.3, and +12.2 recall points over the closest open baseline (mean +7.9).

Across the four *held-out transfer benchmarks* excluded from HarnesS-1 SFT and RL (LongSealQA, Seal0QA, FRAMES, HotpotQA), HarnesS-1 gains +32.8, +18.4, +7.2, and +9.5 points (mean +17.0). This is a 2.2× larger gain on the benchmarks furthest from any training data (Figure 3). The mechanism is straightforward: the policy is learning operations over a domain-general search state: refine the auto-seeded set, read bridge entities from the evidence graph, reinspect uncertain candidates, verify before promotion, and submit a compact curated set. Those

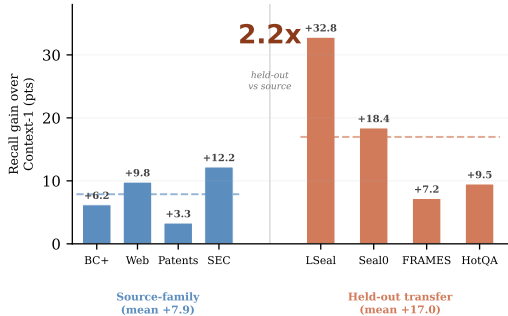


Figure 3: **Transfer pattern.** HarnesS-1 gains more on held-out transfer benchmarks (+17.0 pts mean) than on source-family benchmarks (+7.9 pts), a 2.2× gap over Context-1.

²https://github.com/chroma-core/context-1-data-gen/blob/main/agent_search_data_gen/

Table 3: **Inference-time component ablation on BrowseComp+** (same trained HarnessS-1 checkpoint). Each row disables exactly one harness mechanism. Δ Recall (%) and Δ FA (%) report the *relative* change versus full HarnessS-1, i.e. $100 \times (\text{ablated} - \text{full}) / \text{full}$. The rightmost column points to the per-mechanism failure mode with a representative example trajectory.

Configuration	Recall	Δ Recall (%)	FA Recall	Δ FA (%)	Failure mode
Full HarnessS-1 (all mechanisms enabled)	0.584	—	0.667	—	—
– Importance tags (binary curate, FIFO eviction)	0.560	−4.1	0.614	−7.9	§M.1.1
– Sentence-BM25 compression (raw chunks)	0.585	+0.2	0.620	−7.0	§M.1.2
– Auto-seed on first search	0.582	−0.3	0.624	−6.4	§M.1.3
– Evidence graph hidden in observations	0.569	−2.6	0.631	−5.4	§M.1.4
– verify returns “unavailable”	0.566	−3.1	0.641	−3.9	§M.1.5
– review_docs returns “unavailable”	0.598	+2.4	0.641	−3.9	§M.1.6
– Content-fingerprint dedup (chunk-ID dedup kept)	0.611	+4.6	0.678	+1.6	§M.1.7
All HarnessS-1 harness mechanisms disabled	0.513	−12.2	0.624	−6.4	§M.2

operations transfer more naturally than domain-specific patterns stored in weights. The FRAMES and HotpotQA trajectories in Appendices Q.3 and Q.4 show the same operations outside the SFT/RL families: fan-out or targeted search proposes candidate bridges, then reading, exact matching, and curation resolve the final evidence.

Data scale. The transfer pattern is not explained by simply using more training data. HarnessS-1 uses 899 filtered SFT trajectories and RL on the SEC train split (3,453 queries), for 4,352 unique training items. Context-1 reports over 8K synthetic SFT tasks and its RL trainer uses 9,159 unique queries (the four full train splits sum to 8,859, plus two capped 150-query curriculum splits). Search-R1 uses no SFT warmup and trains RL on the merged NQ+HotpotQA set (221,328 rows). Figure 4 therefore compares unique task/query counts, not rollout count or token compute. The point is not that less data is always sufficient, but that HarnessS-1 moves much of the behavioral prior into the stateful interface, allowing small SFT and focused RL to transfer.

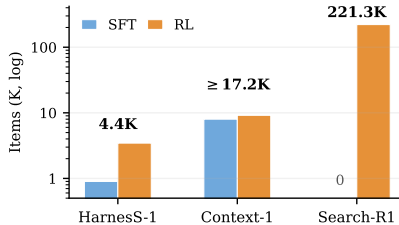


Figure 4: **Training-data scale.** Paired bars show reported unique task/query units used for SFT and RL.

Inference-Time Component Ablation We run the trained HarnessS-1 checkpoint on 100 paired BrowseComp+ test queries with individual harness mechanisms disabled at inference time (no re-training). Each row of Table 3 disables exactly one mechanism; `verify` and `review_docs` are switched off by returning “unavailable.” The bottom row disables *all* HarnessS-1 harness mechanisms simultaneously while keeping the trained policy unchanged.

Six of seven mechanisms produce clean FA-recall drops (−3.9% to −7.9% relative) with a consistent behavioural signature: on queries the full system solves but the ablated condition fails, `search_corpus` rises by 3–7 points and `read_document/verify` drop 2–6×. Removing a state mechanism does not just lose information — the trained policy reverts to a wide, shallow, search-dominated mode and never converges on the relevant documents. The four largest FA drops correspond to the mechanisms with coupled training roles: importance tags (§M.1.1), sentence compression (§M.1.2), auto-seed (§M.1.3), and the evidence graph (§M.1.4). **Content dedup is the only ablation that nominally helps** (+4.6% Recall, +1.6% FA, both relative). BC+ qrels occasionally contain near-duplicate gold documents, which our MinHash–LSH dedup ($\theta_{\text{Jaccard}}=0.85$) sometimes collapses, costing one gold ID per affected query. Dedup is a token-budget mechanism, not a recall mechanism: it shrinks the curated context for downstream answering with negligible loss in what the agent actually retrieves. We report it rather than hide it; the trade-off is the design intent (§M.1.7). **Mechanisms compose.** Disabling *all* harness mechanisms at once gives Recall 0.513 (−12.2%) and FA 0.624 (−6.4%) — a larger relative Recall drop than any single ablation. The trained policy keeps searching but cannot prioritize what it sees: the harness is where per-turn search bandwidth is converted into a discriminative curated set, and removing it leaves exploration intact but the decision substrate gone (§M.2).

Discovery and selection. Trajectory recall shows that HarnessS-1 usually discovers the relevant evidence: across domains, the harness surfaces a large pool of gold documents before final curation. The remaining gap to Opus-4.6 on BC+ final-answer recall (0.658 vs. 0.736), despite similar trajectory recall, is therefore mainly a selection gap rather than a discovery gap. Qualitative examples in Appendices Q.1 and Q.2 show cases where the right document is seen but not preserved in the final curated set.

Training Dynamics Figure 5 compares two RL runs that differ only in whether the tool-diversity reward w_{div} is active. Both start from the same SFT checkpoint and use the same hyperparameters otherwise. The comparison shows that a rich harness alone is insufficient: without RL-compatible shaping, the policy collapses to a narrow search-heavy strategy.

Without w_{div} , the agent quickly learns to issue many `fan_out_search` calls, which improves trajectory recall but bypasses curation and verification. Tool diversity falls from ~ 6 to ~ 3.5 , and curated recall plateaus at ~ 0.53 . The agent finds relevant documents, but does not reliably promote the right ones into the final output set.

With w_{div} active, tool use remains broader: diversity stabilises around ~ 4.30 , and the agent regularly combines search with curation and verification. Early recall improves more slowly because the policy spends more turns editing the candidate set, but final curated recall is higher, reaching ~ 0.60 . This demonstrates the third RL-compatibility requirement. The harness must not only expose useful tools; the reward must also make those tools learnable and worth using. Otherwise, RL exploits the easiest reward path and collapses back toward a minimal search-only harness.

Modular RAG answer accuracy. The search gains also translate to downstream answer quality. When each subagent’s curated set is passed to frozen frontier generators, better curated sets yield higher answer accuracy, and HarnessS-1 outperforms the open subagents in the reported modular RAG setting. Full plots and protocol details are in Appendix O.

4 Conclusion

We introduced HarnessS-1, a 20B search agent trained with reinforcement learning inside a stateful harness. The key idea is to separate semantic search decisions from recoverable bookkeeping: the policy decides what to search, curate, verify, and submit, while the harness maintains candidate pools, importance-tagged evidence, verification records, evidence links, and budget-aware context rendering.

Across eight benchmarks, HarnessS-1 achieves the strongest average recall among the open search agents we evaluate and remains competitive with much larger frontier-model searchers. Its gains on held-out transfer benchmarks and component ablations suggest that the harness is not merely an implementation detail, but a central part of what the policy learns to use.

These results point to stateful harness design as an important direction for retrieval-agent RL. Future work can make the evidence graph more intelligent, for example by replacing regex-based extraction with learned entity linking, relation extraction, and uncertainty-aware evidence organization. We discuss connections to agentic search, tool orchestration, and harness engineering in Appendix A, and limitations, ethics, and broader impact in Appendix B.

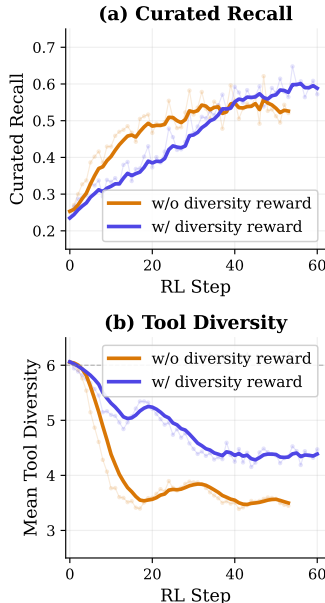


Figure 5: **Training dynamics.** (a) Recall during RL training. (b) Mean tool diversity. **Without the diversity bonus**, recall plateaus and tool use collapses; **with the bonus**, diversity stabilises and final recall is higher.

References

- [1] Cognition AI. Introducing swe-grep and swe-grep-mini: RL for multi-turn, fast context retrieval. Cognition technical report, 2025.
- [2] Hammad Bashir, Kelly Hong, Patrick Jiang, and Zhiyi Shi. Chroma context-1: Training a self-editing search agent. Technical report, Chroma, March 2026.
- [3] Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.
- [4] Kun Chen, Qingchao Kong, Zhao Feifei, and Wenji Mao. Apex-searcher: Augmenting llms’ search capabilities through agentic planning and execution, 2026.
- [5] Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. Research: Learning to reason with search for llms via reinforcement learning, 2025.
- [6] 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, 2025.
- [7] Yong Deng, Guoqing Wang, Zhenzhe Ying, Xiaofeng Wu, Jinzhen Lin, Wenwen Xiong, Yuqin Dai, Shuo Yang, Zhanwei Zhang, Qiwen Wang, Yang Qin, Yuan Wang, Quanxing Zha, Sunhao Dai, and Changhua Meng. Atom-searcher: Enhancing agentic deep research via fine-grained atomic thought reward, 2025.
- [8] Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl, 2025.
- [9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

- [10] Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in llms, 2025.
- [11] Pengrui Han, Xueqiang Xu, Keyang Xuan, Peiyang Song, Siru Ouyang, Runchu Tian, Yuqing Jiang, Cheng Qian, Pengcheng Jiang, Jiashuo Sun, Junxia Cui, Ming Zhong, Ge Liu, Jiawei Han, and Jiaxuan You. Steer2adapt: Dynamically composing steering vectors elicits efficient adaptation of llms, 2026.
- [12] Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma, July 2025.
- [13] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025.
- [14] Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*, 2025.
- [15] Pengcheng Jiang, Jiacheng Lin, Zhiyi Shi, Zifeng Wang, Luxi He, Yichen Wu, Ming Zhong, Peiyang Song, Qizheng Zhang, Heng Wang, Xueqiang Xu, Hanwen Xu, Pengrui Han, Dylan Zhang, Jiashuo Sun, Chaoqi Yang, Kun Qian, Tian Wang, Changran Hu, Manling Li, Quanzheng Li, Hao Peng, Sheng Wang, Jingbo Shang, Chao Zhang, Jiaxuan You, Liyuan Liu, Pan Lu, Yu Zhang, Heng Ji, Yejin Choi, Dawn Song, Jimeng Sun, and Jiawei Han. Adaptation of agentic ai: A survey of post-training, memory, and skills, 2026.
- [16] Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. s3: You don’t need that much data to train a search agent via RL. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21599–21617, Suzhou, China, November 2025. Association for Computational Linguistics.
- [17] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7969–7992, 2023.
- [18] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.
- [19] Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms, 2025.
- [20] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759, 2025.
- [21] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023.
- [22] Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-harness: End-to-end optimization of model harnesses. *arXiv preprint arXiv:2603.28052*, 2026.

- [23] Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. Websailor: Navigating super-human reasoning for web agent, 2025.
- [24] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-ol: Agentic search-enhanced large reasoning models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5420–5438, Suzhou, China, November 2025. Association for Computational Linguistics.
- [25] Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhui Chen, Pengyu Zhao, and Junxian He. Webexplorer: Explore and evolve for training long-horizon web agents, 2025.
- [26] Ryan Lopopolo. Harness engineering: leveraging codex in an agent-first world, 2026.
- [27] Xinghua Lou, Miguel Lázaro-Gredilla, Antoine Dedieu, Carter Wendelken, Wolfgang Lehrach, and Kevin P. Murphy. Autoharness: improving llm agents by automatically synthesizing a code harness, 2026.
- [28] Lance Martin, Gabe Cemaj, and Michael Cohen. Scaling managed agents: Decoupling the brain from the hands. <https://www.anthropic.com/engineering/managed-agents>, April 2026. Anthropic Engineering Blog.
- [29] OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorkak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrlyov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025.
- [30] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024.
- [31] Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. Sealqa: Raising the bar for reasoning in search-augmented language models. *arXiv preprint arXiv:2506.01062*, 2025.
- [32] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore, December 2023. Association for Computational Linguistics.
- [33] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou,

- Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toollm: Facilitating large language models to master 16000+ real-world apis, 2023.
- [34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
- [35] Biswa Sengupta and Jinhua Wang. Harbor: Automated harness optimization. *arXiv preprint arXiv:2604.20938*, 2026.
- [36] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025.
- [37] Hongjin Su, Shizhe Diao, Ximing Lu, Mingjie Liu, Jiacheng Xu, Xin Dong, Yonggan Fu, Peter Belcak, Hanrong Ye, Hongxu Yin, Yi Dong, Evelina Bakhturina, Tao Yu, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Toolorchestra: Elevating intelligence via efficient model and tool orchestration, 2025.
- [38] Hao Sun, Zile Qiao, Bo Wang, Guoxin Chen, Yingyan Hou, Yong Jiang, Pengjun Xie, Fei Huang, and Yan Zhang. Decouplesearch: Decouple planning and search via hierarchical reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4465–4478, 2025.
- [39] Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025.
- [40] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037, 2023.
- [41] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [42] United States Patent and Trademark Office. USPTO Open Data Portal. <https://data.uspto.gov/>, 2026. Accessed 2026-05-06.
- [43] U.S. Securities and Exchange Commission. EDGAR Application Programming Interfaces. <https://www.sec.gov/search-filings/edgar-application-programming-interfaces>, 2024. Accessed 2026-05-06.
- [44] Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*, 2025.
- [45] Ziliang Wang, Xuhui Zheng, Kang An, Cijun Ouyang, Jialu Cai, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. *arXiv preprint arXiv:2505.15107*, 2025.
- [46] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Web-dancer: Towards autonomous information seeking agency, 2025.
- [47] Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. Agentless: Demystifying llm-based software engineering agents. *arXiv preprint arXiv:2407.01489*, 2024.
- [48] Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. Demystifying llm-based software engineering agents. *Proceedings of the ACM on Software Engineering*, 2(FSE):801–824, 2025.
- [49] John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

- [50] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380, 2018.
- [51] 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.
- [52] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025.
- [53] Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Yankai Chen, Chunkit Chan, Peilin Zhou, Xinyang Zhang, Chenwei Zhang, Jingbo Shang, Ming Zhang, Yangqiu Song, Irwin King, and Philip S. Yu. From web search towards agentic deep research: Incentivizing search with reasoning agents, 2025.
- [54] Yuxuan Zhang, Haoyang Yu, Lanxiang Hu, Haojian Jin, and Hao Zhang. General modular harness for llm agents in multi-turn gaming environments. *arXiv preprint arXiv:2507.11633*, 2025.
- [55] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431, Suzhou, China, November 2025. Association for Computational Linguistics.
- [56] Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents, 2025.

Contents of Appendix

A	Related Work	16
B	Limitations, Ethics, and Broader Impact	17
C	Tool Signatures	17
D	Reward & Training Hyperparameters	18
E	Benchmark Statistics	18
F	Two-Tier Memory and Working-Memory Rendering	19
G	Per-Turn Programmatic Nudges (Result Summaries)	20
H	Harness Algorithms	21
I	System Prompts and Templates	22
I.1	Agent System Prompt (Harness-1)	22
I.2	verify Sub-Prompt	23
I.3	Answer-Generator Prompt (modular RAG)	24
I.4	LLM-as-Judge Prompt	24
J	SFT Data Generation Recipe	24
K	Importance-Aware Curation: Algorithmic Detail	25
L	Content Deduplication: Algorithmic Detail	26
M	Component Ablation: Error Analysis	26
M.1	Per-mechanism failure modes	26
M.2	All harness mechanisms disabled	28
N	Evaluation Recipe	28
O	Modular RAG: Curated Sets Yield Higher Answer Accuracy	29
P	Harness as Confound: Same LLM, Different Harness	29
Q	Qualitative Case Studies	31
Q.1	Case Study A: Stateful Search over a Drug-Approval Query	31
Q.2	Case Study C: Exact-Date Recovery in SEC Filings	38
Q.3	Case Study B: Fan-Out Search Resolves an Ambiguous Transfer Query	42
Q.4	Case Study D: Held-Out Multi-Hop QA via Search-State Editing	49
Q.5	Case Study E: Backtracking after Rejecting the Right Entity	53

A Related Work

Harness engineering. Harness engineering designs the environment layer between a language model and its task [26, 28]. Recent work shows that the same model can vary by tens of points when placed in different harnesses [49, 47], making the interface a major determinant of system-level performance. Tool orchestration [33, 18, 25, 37] usually fixes this interface and trains the policy to call the exposed tools more effectively. Harness engineering asks a lower-level design question: what interface should the policy operate?

Most existing harnesses are designed by hand [49, 54], often as empirical responses to model behaviors that may change as models improve [28]. Recent systems automate harness construction by searching over code space [27, 22] or configuration space [35]. In Harness-1, we study a complementary question: how should a retrieval harness be shaped so that a trainable policy can use it well? Our answer is to offload mechanical search bookkeeping into environment-side state, then train the policy to make semantic decisions over that state.

Tool orchestration vs. stateful harnessing. Let an agent interface be $\mathcal{I} = (\mathcal{A}, \mathcal{O}, T, r)$: actions, observation renderer, environment transition, and reward. Tool orchestration usually fixes an interface \mathcal{I}_0 and learns a policy inside it:

$$\max_{\theta} \mathbb{E}[R(\pi_{\theta}; \mathcal{I}_0)].$$

Stateful harnessing changes \mathcal{I} by adding the editable retrieval state from Table 1:

$$s_t = (P_t, C_t, I_t, D_t, G_t, V_t, H_t, B_t).$$

The transition is

$$(s_t, a_t) \mapsto (s_{t+1}, o_{t+1}),$$

not merely $a_t \mapsto o_{t+1}$. The policy still learns tool use, but it learns over a state representation maintained by the harness:

$$\max_{\theta} \mathbb{E}[R(\pi_{\theta}; \mathcal{I}_{\text{Harness-1}})].$$

Agentic search. Agentic search refers to large language model agents that gather evidence through multiple rounds of query generation, reading returned documents, and deciding when to stop [53]. Early work in this direction interleaved retrieval with reasoning in tight loops [24, 8, 25, 1]. ReAct [51], Self-Ask [32], and IRCot [40] established the pattern of alternating reasoning and tool calls. However, prompting-based reasoning alone is often inconsistent with surface behavior failing to reliably translate into the intended decision patterns [41, 21, 10, 11].

Reinforcement learning post-training [34, 13, 52, 9] has emerged as a powerful tool for shaping LLM reasoning, and a growing line of work brings it to retrieval and agentic search [36, 5, 14, 55]. DeepRetrieval [14] shows that LLMs can be trained with RL to interact directly with real search engines and retrievers, improving retrieval through learned query and search behavior. Search-R1 [18] trains a single LLM to interleave reasoning and search calls under a downstream question-answering reward. s3 [16] decouples the searcher from the generator, training only the search policy for modular optimization with frozen generators. These methods work well at short horizons but as tasks grow longer and more complex, like report generation [46, 23], the policy is asked to do too much at once: planning the next search [4, 38], remembering which documents have been seen [56, 30], tracking which constraints remain uncovered [6], and managing an ever-growing context [12]. Under this load, the reward signal grows sparse relative to the work the policy is actually doing causing the degradation of RL training [7, 44]. Recent work begins to address parts of this overload. Context-1 [2] trains a 20B retrieval agent that learns to selectively prune its own context during search, beginning to treat context management as part of the policy itself.

Across these systems, the harness of the agent remains a thin tool wrapper, and any state beyond the prune action must be reconstructed by the policy from a raw, append-only observation stream. Harness-1 takes a different stance: instead of training a stronger policy on a thin interface, we make the interface itself part of the method. The harness is stateful, with environment-side working memory that absorbs the mechanical bookkeeping of multi-turn search. We frame this division of labor as cognitive offloading, and treat harness design itself as a central axis of agentic search.

B Limitations, Ethics, and Broader Impact

Limitations. HarnesS-1 is designed for evidence-seeking retrieval tasks where the goal is to identify documents that support an answer. Most benchmarks in our evaluation are needle-in-a-haystack or multi-hop search tasks with annotated relevant evidence. This setting is important, but it does not cover all retrieval use cases. Breadth-oriented research tasks, open-ended report generation, abstention under missing evidence, and adversarial web environments remain outside the main scope of this work.

Our harness also introduces several engineered components whose quality may vary across domains. The evidence graph uses a lightweight regex-based extractor for entities, dates, and years, rather than a full entity-linking system. The `verify` tool uses an LLM-based verifier as a proxy entailment model, which can make mistakes on ambiguous, highly technical, or underspecified claims. The sentence-level BM25 compressor may remove useful context when relevance depends on discourse structure rather than local sentence overlap.

Evaluation is limited by benchmark size and annotation coverage. Some datasets have non-negligible confidence intervals, and near-duplicate or incomplete qrels can affect recall-oriented metrics. We therefore report multiple retrieval metrics, trajectory-level diagnostics, ablations, and modular RAG answer accuracy, but these measurements should still be interpreted as evidence of retrieval behavior under the specified benchmark and harness conditions rather than as a complete measure of real-world search reliability.

Ethics. This work studies how to train language-model search agents to retrieve supporting evidence more effectively. The primary ethical motivation is to make retrieval-augmented systems more transparent and auditable: instead of relying only on a model’s parametric memory, the system returns an explicit curated set of documents that can be inspected, verified, and used by downstream models or human users.

The main ethical risks come from misuse and over-reliance. A stronger search agent could retrieve sensitive, private, copyrighted, or otherwise inappropriate information if connected to an unrestricted corpus. It could also amplify misleading evidence if the retrieval backend contains low-quality, biased, or adversarial documents. Our method should therefore be deployed only with appropriate corpus access controls, logging, rate limits, privacy filters, and human oversight for sensitive domains.

The system should not be treated as a source of truth by itself. HarnesS-1 retrieves and curates evidence; it does not guarantee that the evidence is complete, unbiased, or correctly interpreted. In high-stakes domains such as medicine, law, finance, or public policy, any downstream answer should be reviewed by qualified experts and validated against authoritative sources.

Broader impact. Stateful retrieval agents can reduce the cost of evidence gathering and make long-horizon search more accessible. This may benefit scientific literature review, legal and financial document analysis, fact-checking, education, and other workflows where users need to locate and organize evidence across many documents. Because HarnesS-1 separates retrieval from generation, its outputs can be audited more directly than closed-book answers.

At the same time, lower-cost search agents can also increase the scale of undesirable information-seeking behavior, including privacy-invasive search, automated evidence cherry-picking, and targeted collection of sensitive information. These risks are not unique to our method, but stronger and cheaper retrieval agents can make them easier to execute. Responsible deployment should therefore restrict corpus access, monitor usage, and pair retrieval with verification and human review.

Our planned release is limited to the retrieval subagent, harness code, data-generation pipeline, and evaluation recipe, with documentation of intended use and dataset provenance. It is not a general autonomous web agent and is not intended to replace expert judgment in high-stakes decision-making.

C Tool Signatures

This section lists the tools exposed to the HarnesS-1 policy. The tools are grouped around the main operations in the harness: retrieving new candidates, revisiting stored evidence, editing the curated

set, verifying claims, and ending the episode. The policy chooses among these tools, while the harness executes the calls and updates the working-memory state described in Table 1.

Table 4 gives the functional signature of each tool used in all teacher, SFT, RL, and evaluation runs.

Table 4: HarnesS-1 tool signatures.

Tool	Effect
<code>fan_out_search(queries)</code>	Up to five parallel hybrid searches, RRF + rerank
<code>search_corpus(query)</code>	Single hybrid search (BM25 + dense), RRF, rerank
<code>grep_corpus(pattern)</code>	Regex match over the corpus
<code>read_document(doc_id)</code>	Full-text read, reranked and truncated to budget
<code>review_docs(doc_ids)</code>	Re-render seen docs from WORKINGMEMORY (no corpus call)
<code>curate(add, remove, importance)</code>	Edit curated set with four-level importance tag
<code>verify(doc_ids, claim)</code>	Per-document LLM entailment check against the claim
<code>end_search(reasoning)</code>	Terminate episode, return curated set
<i>For backward compatibility (registered, no-op):</i>	
<code>prune_chunks(...)</code>	Returns “managed via working memory; no pruning needed.”

D Reward & Training Hyperparameters

Table 5 lists the reward weights and penalty coefficients used in the deployed RL run, alongside their configuration names.

Table 5: Reward hyperparameters for the deployed HarnesS-1 RL run. **Note on the empty-set penalty:** when the curated set is empty at episode termination, the reward short-circuits to $\pi_\emptyset = -0.2$, bypassing the rest of the formula. For all other terminations, the formula in §2.3 applies and is clipped below by $\mathcal{R} \geq 10^{-3}$.

Symbol	Code variable	Value
w_F (set quality)	OUTCOME_WEIGHT	0.7
w_τ (trajectory recall)	TRAJECTORY_RECALL_WEIGHT	0.3
β (F_β exponent)	RECALL_BETA	2.0
w_A (FA recall)	FINAL_ANSWER_RECALL_WEIGHT	0.8
$w_{\tau A}$ (trajectory FA recall)	TRAJECTORY_FA_RECALL_WEIGHT	0.4
B_A (binary FA bonus)	FINAL_ANSWER_BONUS	1.0
w_{miss} (FA miss penalty)	FA_MISS_PENALTY_WEIGHT	0.35
π_\emptyset (empty-set short-circuit)	NO_CURATE_PENALTY	-0.2
Turn penalty start	TURN_PENALTY_MIN_TURNS	20
Turn penalty max	TURN_PENALTY_MAX	0.02
Max turns hard cap	MAX_TURNS	40
Format-reward floor	MIN_FORMAT_REWARD	10^{-3}
KL coefficient (anchor to SFT)	KL_PENALTY_COEF	0.0
w_{div} (tool-diversity bonus weight)	TOOL_DIVERSITY_BONUS	0.15
ν_0 (tool-diversity target)	TOOL_DIVERSITY_TARGET	6

We also report the training configuration for the released HarnesS-1 checkpoint. We separate hyperparameters for SFT, RL, the search environment, retrieval infrastructure, and evaluation. The goal is to make the reported training recipe reproducible and to clarify which settings belong to the model optimization stage versus the harness and retrieval backend.

SFT is used only to teach the model how to operate the stateful interface. RL is then applied over full search episodes using the same harness state renderer and tool set. Table 6 lists the exact settings used in the deployed run.

E Benchmark Statistics

This section summarizes the evaluation benchmarks used in the main experiments. The benchmark suite covers static Chroma-backed corpora and live web-backed retrieval settings, spanning encyclo-

Table 6: SFT and RL hyperparameters for the deployed HarnessS-1 checkpoint.

Stage	Setting	Value
SFT	Base model	openai/gpt-oss-20b (MoE)
SFT	Adapter	LoRA, rank 32
SFT	Optimiser	AdamW (Tinker default)
SFT	Learning rate	5×10^{-6}
SFT	Batch size	128
SFT	Max sequence length	32,768
SFT	Epochs	3
SFT	Min recall filter	0.10 (final output recall)
SFT	Per-trajectory expansion	one datum per turn
SFT	Selected checkpoint	step 550 (sampler weights)
RL	Algorithm	on-policy CISPO (clip [0, 5])
RL	Adapter	LoRA, rank 32
RL	Learning rate	1×10^{-5}
RL	Queries per step (BATCH)	128
RL	Rollouts per query (GROUP)	8
RL	Total rollouts per step	1,024
RL	Substeps per step	4
RL	Total steps	80
RL	Total rollouts	~82K
RL	Rollout temperature	1.0
RL	Generation budget	2,048 tokens
RL	Model context limit	32,768 tokens
RL	KL anchor	disabled
RL	Drop constant-reward groups	yes
RL	Dataset	sec (single-domain)
Env	MAX_TURNS	40
Env	RECENT_K (recent window)	5
Env	SEARCH_DISPLAY_LIMIT	10
Env	MAX_OBS_CHARS	15,000
Env	SEARCH_TOKEN_BUDGET	4,096
Env	SENTENCE_COMPRESS_K	4
Env	MINHASH_DEDUP_THRESHOLD	0.85
Env	MINHASH_NUM_PERM	64
Env	EVIDENCE_GRAPH_MAX_ENT	8
Env	AUTO_POPULATE_TOP_K	8
Env	MAX_CURATED_DOCS	30
Env	DOC_SNIPE_CHARS	120
Retr	Retrieval	hybrid BM25 + dense, RRF
Retr	Reranker	Qwen3-Reranker-8B (Baseten-hosted)
Retr	Reranker max tokens	4,096
Eval	Sampling temperature	1.0
Eval	Max tokens per turn	2,048 (HarnessS-1), 4,096 (Context-1)
Eval	Max trajectory length	40 turns (HarnessS-1), 64 turns (Context-1)

pedic web search, finance filings, patents, and multi-hop question answering. The split is also used in the main text to distinguish source-family benchmarks from held-out transfer benchmarks.

Table 7 reports the backend, number of test queries, domain, and task style for each benchmark.

F Two-Tier Memory and Working-Memory Rendering

HarnessS-1 uses a two-tier memory: an inner WORKINGMEMORY summary that fits in the prompt and an outer doc_store (full text of every chunk ever retrieved) accessible via review_docs. This section documents the exact rendering protocol so the reported behaviour is reproducible.

Working-memory text layout. Each rendered WORKINGMEMORY block contains, in order:

Table 7: Benchmark statistics.

Benchmark	Backend	Test queries	Domain	Style
BrowseComp+ [6]	Chroma (BC+ corpus)	166	Web encyclopaedic	Multi-constraint
Web [2]	Chroma (web 1.17)	554	Web encyclopaedic	Multi-hop
Patents [2, 42]	Chroma (USPTO 1.18)	718	Legal/USPTO	Single-doc heavy
SEC [2, 43]	Chroma (SEC 1.4)	502 ³	Finance filings	Multi-filing
LongSealQA [31]	Chroma (longsealqa)	254	Open web	Multi-hop, long-context
Seal0QA [31]	Web (Serper+Jina)	111	Open web	Factoid / multi-hop mix
FRAMES [20]	Web (Serper+Jina)	824	Open web	Multi-hop
HotpotQA-subset [50]	Web (Serper+Jina)	493	Open web	2-hop

1. Header: == Working Memory (summarising turns 0-- n) == followed by the original query.
2. Curated set: rendered grouped by importance level (very_high first, high, fair, low last), each entry showing the doc ID and a 120-character snippet.
3. Document pool: the most recent 50 uncurated docs are shown with snippets; older uncurated docs are summarised as a comma-separated ID list (truncated at 30 IDs with a (+N more) marker if longer).
4. Search history: the last 12 entries; entries older than that are summarised as ... (k earlier searches).
5. Evidence-graph block (only if ≥ 1 multi-doc bridge entity exists): up to 8 frequent entities listed with their doc IDs, plus a singleton count.
6. Dedup notice (only if any near-duplicates were suppressed): [Dedup] k near-duplicate chunk(s) auto-suppressed.

Per-turn observation layout. Each observation rendered to the policy contains, in order: (a) the WORKINGMEMORY summary block; (b) up to RECENT_K=5 prior (action, result) pairs, with reasoning truncated to 300 characters for all but the most recent turn; (c) the freshly produced tool result for the most recent action. Search-style observations within (c) are passed through sentence-BM25 compression ($K=4$ kept sentences) and content-fingerprint deduplication. Search results are truncated to SEARCH_DISPLAY_LIMIT=10 entries per call; each per-result snippet is capped at DOC_SNIPPET_CHARS=120.

Context-budget truncation (5-pass progressive degradation). Whenever the rendered context exceeds the prompt token budget (30,720 tokens, leaving 2,048 for generation within the 32,768 context limit), `render_context_within_budget` applies progressive degradation in five passes, returning the first that fits: (1) normal render; (2) truncate the document-pool section of the WORKINGMEMORY summary while preserving curated-set, search history, and evidence graph; (3) truncate analysis text in older turns to 100 chars and the WORKINGMEMORY summary itself to 2,000 chars; (4) drop the oldest recent (action, result) pair and retry, iterating; (5) minimal context (system + tool descriptions + query only). Pass 5 always succeeds because the system prompt is constant-size. This mechanism guarantees that no rollout fails due to context overflow, and that the curated set and recent working memory are the last things to be cut.

G Per-Turn Programmatic Nudges (Result Summaries)

For full transparency: between each turn’s tool result and the next observation, HarnesS-1 injects a short *result summary* as a user message. The summary is generated programmatically (no LLM call) and combines factual status with conditional prescriptive nudges. The same summaries are seen by the SFT teacher (GPT-5.4) and the trained policy at RL and inference time, so the prompt distribution is identical across all three stages.

Always-on factual lines. [STATUS]: tool name, number of docs returned, novel docs, total pool size; or curate/review status. Curated: current curated set size and a preview of the first five doc IDs.

Algorithm 1 HarnessS-1 stage driver.

Require: query q , tools \mathcal{T} , stage $s \in \{\text{teacher, sft, rl, infer}\}$, policy π or replay trajectory τ
Ensure: trajectory record for training stages, or curated set C for inference

- 1: Initialize $P, C, I, D, G, V, H, U \leftarrow \emptyset$ and $\text{AUTOSEEDDED} \leftarrow \text{false}$
- 2: Build a per-episode rerank instruction from q and the dataset family
- 3: **for** $t = 1, \dots, T_{\max}$ **do**
- 4: $x_t \leftarrow \text{RENDERCONTEXT}(q, P, C, I, D, G, V, H, \text{recent turns})$ Algorithm 5
- 5: **if** $s = \text{sft}$ **then**
- 6: Read stored action a_t and stored observation o_t from τ
- 7: **else**
- 8: Sample a structured tool action $a_t \sim \pi(\cdot | x_t)$
- 9: Execute a_t with tools \mathcal{T} to obtain raw observation o_t
- 10: **end if**
- 11: **if** a_t is a retrieval or memory-inspection action **then**
- 12: $(o_t, P, D, G, U, C, I, \text{AUTOSEEDDED}) \leftarrow \text{PROCESSOBSERVATION}(a_t, o_t)$ Algorithm 2
- 13: **else if** $a_t = \text{curate}$ **then**
- 14: $(C, I) \leftarrow \text{CURATE}(a_t, C, I, P)$ Algorithm 3
- 15: **else if** $a_t = \text{verify}$ **then**
- 16: $V \leftarrow V \cup \text{VERIFYCLAIM}(a_t, D)$ Algorithm 4
- 17: **else if** $a_t = \text{review_docs}$ **then**
- 18: $o_t \leftarrow \text{REVIEWDOCS}(a_t, D)$ Algorithm 4
- 19: **else if** $a_t = \text{end_search}$ **then**
- 20: **break**
- 21: **end if**
- 22: Update search history H , result summary, token marker, and a working-memory snapshot
- 23: **end for**
- 24: Apply the stage-specific finalizer in Algorithm 6

Conditional prescriptive lines (gated by `HARNESSEPRESCRIPTIVE=1`, which was on for all reported runs). `[ACTION REQUIRED]`: triggered after a search turn with ≥ 1 turn since last curate; instructs the agent to curate before searching again. `[WARN]`: triggered on multiple consecutive non-curate turns or a blank curated set while the pool already has candidates. `[TIP]`: triggered when truncation is detected (suggesting `read_document`), when only one tool type has been used by turn ≥ 3 , or when `grep_corpus` or `read_document` have not been used by turn $\geq 4/6$. `[NEXT]`: short suggestion of the immediate next action.

These nudges are *not* part of the trained policy’s intrinsic capability. They are part of the harness. They are included in the harness mass we report ablating in Table 3 (the “all HarnessS-1 mechanisms disabled” row removes them).

H Harness Algorithms

This appendix gives the formal harness algorithms used by HarnessS-1. The driver in Algorithm 1 is intentionally small: all mechanism-specific behavior is delegated to explicit subroutines. The same subroutines are used during teacher trajectory generation, SFT replay, RL rollout, and inference, so the policy is trained on the same state interface it uses at test time.

State variables. For a query q , the harness maintains a document pool P , a curated output set C with cap $M=30$, an importance map $I : C \rightarrow \{\text{very_high, high, fair, low}\}$, a full-text store D , an evidence graph G , a verification cache V , a search history H , a near-duplicate index U , and a Boolean flag `AUTOSEEDDED`. The policy controls semantic choices (queries, claims, promotions, removals, and termination); the harness controls deterministic state maintenance.

Algorithm 2 Observation processing and working-memory update.

Require: action a , raw observation o , ranked IDs R , query text q , state $(P, D, G, U, C, I, \text{AUTOSEDED})$
Ensure: processed observation and updated memory state

- 1: **if** $a \in \{\text{search_corpus}, \text{fan_out_search}, \text{grep_corpus}\}$ **then**
- 2: Split each returned chunk into sentences and keep the top $K=4$ BM25-scored sentences, preserving original order
- 3: **end if**
- 4: Parse document IDs and document text blocks from o
- 5: **for** each returned chunk (r, text) **do**
- 6: Normalize r to a document ID d when the dataset uses document-level IDs
- 7: **if** $d \in P$ **then**
- 8: skip adding d to the pool, but retain the observation for trajectory accounting
- 9: **else if** text is a near-duplicate under U **then**
- 10: suppress d from rendered memory and increment the dedup counter
- 11: **else**
- 12: Add d to P and store full text plus snippet in $D[d]$
- 13: Extract proper nouns, years, and numeric dates from text and update G
- 14: **end if**
- 15: **end for**
- 16: **if** $a \in \{\text{search_corpus}, \text{fan_out_search}\}$ and AUTOSEDED is false and $R \neq \emptyset$ **then**
- 17: Add the first $k=8$ ranked IDs from R to C with $I(d) = \text{fair}$
- 18: Set $\text{AUTOSEDED} \leftarrow \text{true}$ and append an [AUTO-POPULATED] notice
- 19: **end if**
- 20: **if** token-budget markers are enabled **then**
- 21: Append [Context: X/Y] using the current approximate prompt budget
- 22: **end if**
- 23: **return** processed observation and updated state

Algorithm 3 Importance-aware curation and capacity management.

Require: add IDs A , remove IDs R , proposed importance map J , curated set C , importance map I , pool P
Ensure: updated (C, I) and a capacity-status observation

- 1: Normalize all IDs in A , R , and J to the dataset’s output ID convention
- 2: Remove every $d \in R$ from C and delete $I(d)$
- 3: **for** each $d \in A$ **do**
- 4: Let $j \leftarrow J(d)$ if valid, otherwise $j \leftarrow \text{fair}$
- 5: **if** $d \in C$ **then**
- 6: Re-tag $I(d) \leftarrow j$ and continue
- 7: **end if**
- 8: **if** $|C| < M$ **then**
- 9: Insert d into C with $I(d) \leftarrow j$
- 10: **else**
- 11: Let $w = \arg \max_{c \in C} \text{rank}(I(c))$, where low has worst rank
- 12: **if** $\text{rank}(j) < \text{rank}(I(w))$ **then**
- 13: Evict w ; insert d with $I(d) \leftarrow j$
- 14: **else**
- 15: Reject d and report a [CAPACITY] marker
- 16: **end if**
- 17: **end if**
- 18: **end for**
- 19: **return** (C, I) rendered by priority $\text{very_high} \succ \text{high} \succ \text{fair} \succ \text{low}$

I System Prompts and Templates

I.1 Agent System Prompt (Harness-1)

Agent system prompt

Role. You are a retrieval subagent. Find the most relevant documents from the corpus for the question; do not answer the question yourself.

Algorithm 4 Verification and memory review.

Require: action a , full-text store D , verification cache V

Ensure: observation o and updated verification cache V

```
1: if  $a = \text{review\_docs}(D)$  then
2:   Return the full text of up to 5 requested IDs from  $D$ , without a corpus call
3: else if  $a = \text{verify}(D, c)$  then
4:   for each requested document  $d \in D$  do
5:     Ask the verifier whether  $D[d]$  directly supports every constraint in claim  $c$ 
6:     Store  $(d, c, \text{yes/no}, \text{short rationale})$  in  $V$ 
7:   end for
8:   Return the per-document yes/no judgments and rationales
9: end if
```

Algorithm 5 Budget-safe context rendering.

Require: system prompt, rendered working memory W , recent actions and observations, result summaries, token budget B

Ensure: tokenized prompt x with $|x| \leq B$

```
1: Build  $W$  from query, curated set grouped by importance, recent uncurated pool, search history, evidence graph, verification cache, dedup notice, and context marker
2: Render system message, tool descriptions, query, optional  $W$ , recent turns, and result summaries
3: if the tokenized prompt fits within  $B$  then
4:   return prompt
5: end if
6: Truncate the document-pool section of  $W$ , preserving curated set, search history, and evidence graph; retry

7: If needed, truncate older reasoning to 100 characters and cap  $W$  at 2,000 characters; retry
8: If needed, drop the oldest recent action-observation pair one at a time; retry after each drop
9: If all retries fail, return the minimal prompt: system message, tool descriptions, and query
```

Input. QUERY: [query]

Tools. `fan_out_search` runs up to five diverse queries; `search_corpus` runs a single hybrid query; `grep_corpus` performs exact regex matching; `read_document` opens full text; `review_docs` re-renders remembered documents; `curate` edits the curated set (maximum 30 documents); `verify` checks a claim against remembered documents; `end_search` submits the curated set.

Loop. Decompose the query into constraints. Search, curate immediately, search again from a different angle, inspect or review uncertain documents, verify multi-constraint evidence, and end when the curated set is sufficient. The episode has at most 40 turns.

State rules. Use importance tags {`very_high`, `high`, `fair`, `low`}; `very_high` requires prior verification. The first successful search seeds tentative `fair` candidates; promote strong documents and remove weak ones. Use evidence-graph bridge entities as follow-up signals. Each observation ends with [Context: X/Y]; wrap up above 75% and call `end_search` above 90%.

I.2 verify Sub-Prompt

verify is implemented as a small per-document LLM call.

verify sub-prompt

System. You are a strict document verifier. Given one CLAIM and one document's full text, answer only yes or no, followed by a rationale of at most 20 words. Answer yes only if the document directly supports every part of the claim. Answer no if any constraint is missing or contradicted. Be conservative.

User template.

CLAIM: [claim]

DOCUMENT: [document text]

Answer: yes/no. rationale

Algorithm 6 Stage-specific finalizers.

Require: stage s , trajectory τ , pool P , curated set C , policy π

- 1: **if** $s = \text{teacher}$ **then**
 - 2: Save all turns, observations, working-memory snapshots, P , C , document store D , and final retrieval metrics
 - 3: **else if** $s = \text{sft}$ **then**
 - 4: Filter trajectories below the recall threshold
 - 5: Replay each kept trajectory through Algorithms 2–5
 - 6: For every turn t , train on the supervised pair $\text{RENDERCONTEXT}_t \mapsto a_t$
 - 7: **else if** $s = \text{rl}$ **then**
 - 8: Score C and P with curated recall, precision, final-answer recall, trajectory recall, and trajectory final-answer recall
 - 9: Compute terminal reward: set quality, trajectory coverage, answer bonus, dense answer shaping, answer-miss penalty, tool-diversity shaping, and turn penalty
 - 10: Update π with on-policy CISPO; save checkpoint and rollout summaries
 - 11: **else if** $s = \text{infer}$ **then**
 - 12: Return the priority-sorted curated set C to the downstream answer generator
 - 13: **end if**
-

I.3 Answer-Generator Prompt (modular RAG)

Answer-generator prompt

Task. You are given a question and retrieved documents. Use only the documents to reason step by step and answer the question.

Question. [query]

Retrieved documents.

[documents]

Output rule. End the response with the final answer on the last line, prefixed exactly with `Final Answer:.`

For **Closed-Book**, the prompt omits the documents and the instruction is changed to “Using only your own knowledge, reason step by step and provide your final answer.” For **Naive RAG**, a single BM25+dense hybrid query retrieves the top-10 chunks and those replace the curated set.

I.4 LLM-as-Judge Prompt

We use GPT-5.4 at temperature 0 as the judge, following the official BrowseComp+ template; we report the strict CORRECT rate as “answer accuracy.”

LLM-as-judge prompt

Task. Judge whether [response] answers [question] correctly according to the precise [correct_answer]. Do not solve the problem independently or comment on irrelevant background.

Inputs.

[question]: [query]

[response]: [generated_answer]

[correct_answer]: [gold_answer]

Return fields. extracted_final_answer is the exact final answer, or None if absent. reasoning explains only meaningful differences. correct is yes when the extracted answer matches the gold answer, allowing small numerical tolerances, and no otherwise. confidence is 0–100%.

J SFT Data Generation Recipe

SFT trajectories are produced by GPT-5.4 (tool_choice=required, max_completion_tokens=4096) running natively inside the HarnesS-1 harness. The teacher sees

the same system prompt (Appendix I.1), the same observation layout, and the same tool schemas as the trained policy will see at RL and inference time.

Teacher agent loop. The generation script runs a single-phase loop: at each turn, the teacher’s full conversation history (system prompt + working- memory text + the most recent $K=3$ turn transcripts + result summaries for older turns) is sent to GPT-5.4 with all 7+1 tool schemas. The teacher selects a tool via `tool_choice=required` and returns structured JSON arguments including a mandatory reasoning field. The harness executes the tool against the live Chroma corpus and returns the observation.

Guidance protocol. The generation script injects turn-level guidance into the prompt: `search→curate` rhythm enforcement; `verify` nudges on multi-constraint queries (after ≥ 6 turns and ≥ 3 curated docs); backtracking triggers when the last 3 search-history entries yielded ≤ 1 new doc; tool diversity hints if `grep_corpus/ read_document` are unused after 4 turns; urgency and curate-aggressiveness messages near the turn cap. This guidance ensures teacher trajectories exhibit the full tool vocabulary and the `search→curate` rhythm.

Per-dataset quotas and filtering. Per-dataset raw quotas: BC+/300, SEC/250, Patents/150, Web/150, Web-simple/75, SEC-simple/75 (total $\sim 1K$ raw). After a 0.10 recall gate (trajectories with recall < 0.10 are discarded), the corpus contains 899 trajectories. Per-trajectory expansion into turn-conditional datums yields $\sim 26K$ training examples.

Auto-populate on first search. During trajectory generation, the first successful search triggers `auto_populate_from_first_search`, which adds the top- k reranked results ($k=8$) to the curated set at fair importance. Idempotent via the per-trajectory `wm.auto_populated` flag. The observation appends an [AUTO-POPULATED] marker instructing the teacher to promote good docs and remove irrelevant ones.

Observation processing. Search observations are processed identically to RL: BM25 sentence compression (top-4 sentences per chunk), content-hash dedup (MinHash-LSH with 64 permutations, 5-gram shingles, threshold 0.85, with SHA-prefix fallback over a normalised 4,000-char prefix when `datasketch` is unavailable), and the [Context: X/Y] token marker.

K Importance-Aware Curation: Algorithmic Detail

This section details the eviction policy central to the subtractive-curation mechanism in `HarnesS-1` (§2.1).

Importance levels and ordering. Every document carries an importance tag $\in \{\text{very_high}, \text{high}, \text{fair}, \text{low}\}$ with rank `very_high=0 < high=1 < fair=2 < low=3` (lower rank = higher importance). Auto-populated documents enter at `fair`.

ID normalisation. Both `add_ids` and `importance` keys are normalised by stripping a trailing `_N` chunk suffix *only if* the resulting ID is not already in the pool. This lets the policy pass either chunk IDs (42982_3) or doc IDs (42982) interchangeably, while preserving doc IDs that legitimately contain underscores.

Eviction protocol. When the curated set is at capacity ($M=30$) and the agent issues a `curate` call with new `add_ids`:

1. For each incoming doc, compute its `incoming_rank` from the provided importance tag (default `fair`).
2. Scan the current curated set for the document with the *worst* (highest) importance rank.
3. If `worst_rank > incoming_rank`, evict the worst-ranked doc and insert the new doc.
4. If no evictable doc exists, the add is rejected; the status string includes a [CAPACITY] marker listing up to 5 rejected IDs.

Re-tagging in place. A `curate` call with an `add_id` already in the curated set updates the tag in-place without eviction. This allows the agent to promote a document after a successful `verify` call.

Rendering. The curated set is rendered grouped by importance level (`very_high` first, low last) with the tag visible next to each document ID:

```
Curated Set (15/30):
  very_high:
  [*] doc_42982: Tesla 10-K 2023 mentioning SEC fine...
  high:
  [*] doc_11293: Tesla annual report 2023...
  fair:
  [*] doc_55017: General Motors 10-K 2023...
```

L Content Deduplication: Algorithmic Detail

HarnesS-1’s content-fingerprint dedup is implemented in `ultra_core.ContentDedupTracker`.

Primary path (MinHash-LSH). When `datasketch` is available, each chunk is fingerprinted by tokenising on the regex `[a-z0-9]+`, building 5-gram shingles, and inserting them into a MinHash with 64 permutations. The LSH index uses Jaccard threshold 0.85. For each new chunk, the tracker queries the LSH; any match is treated as a near-duplicate and dropped from the pool but *retained* in the `trajectory_recall` accounting (so the agent is credited for having encountered the gold information).

Fallback path (SHA prefix). If `datasketch` is unavailable, the tracker computes a SHA-1 of the first 4,000 characters of the normalised token-stream and compares against a hash set. This catches exact duplicates only, not near-duplicates, but introduces no extra dependency.

Diagnostic. The number of suppressed duplicates is exposed in the `WORKINGMEMORY` rendering as a `[Dedup] k near-duplicate chunk(s) auto-suppressed` line, which lets the policy reason about whether the corpus has heavy duplication.

M Component Ablation: Error Analysis

This appendix unpacks Table 3 at the trajectory level. For each ablation we identify queries that full HarnesS-1 solves (FA Recall ≥ 0.5) but the ablated condition fails on (FA Recall = 0), then compare action mixes and trajectories on those *paired* queries. All numbers are over the same 100 BrowseComp+ test queries; no retraining is involved.

A consistent pattern emerges: removing a single state mechanism *changes the policy’s behaviour* rather than just losing information. On failing paired queries the share of `search_corpus` actions rises by 3–7 points, while `read_document` and `verify` drop by 2–6 \times — a retrieval failure, not a curation failure. The trained policy reverts to a wide, shallow, search-dominated mode and never reaches the relevant documents in the first place. The per-mechanism analyses below each illustrate the pattern with one representative failing query (truncated for length).

M.1 Per-mechanism failure modes

M.1.1 Importance tags

Δ Recall -4.1% , Δ FA -7.9% relative (15 hard fails). Binary `curate` with FIFO eviction loses the `very_high/high/fair/low` confidence signal over the curated set. On failing queries, `read_document` drops 7 \times (4.7% \rightarrow 0.7%), `verify` drops to 0%, and `search_corpus` climbs to 94.1% of all actions.

Example. Q893 asks for the birth name of a 1920s-born actress married to a director-translator (gold answer: “Medea Japaridze”). Full HarnesS-1 reads `doc_3161`, identifies Medea Japaridze on

T4, then re-anchors via her spouse Revaz Tabukashvili (T28, T37, T39) and saturates curated recall at 1.0 in 40 turns. The ablation issues 19 `search_corpus` calls without a single `read_document` or `verify`, never anchors on “Medea Japaridze”, and ends with curated recall 0.12 on *the same* 40-turn budget. The agent cannot decide which candidate to follow up on without the importance gradient.

M.1.2 Sentence-BM25 compression

Δ Recall +0.2%, Δ FA -7.0% relative (13 hard fails). Without per-return sentence compression, top- k chunks come back raw; returns are longer and noisier. The aggregate Recall delta is ~ 0 , but FA Recall drops sharply because the agent loses the bridge sentences it normally uses to seed follow-up reads. `read_document` on failing queries drops from 4.1% to 2.6% and `verify` from 2.0% to 0.4%.

Example. Q1012 asks for the title of a 2018–2023 EP whose songs all end in a question mark and were drawn from the artist’s journal entries. Full `Harness-1` recovers it in 26 turns: T14 greps for “gave my all” (a phrase from the artist interview), reads `doc_22401` on T22, and `verify` confirms “Maki’s Tanong” on T24. The ablation reads `doc_77931` on T2 (a near-miss), then issues 18 subsequent searches without ever reading another doc; without compressed bridge sentences from the search returns it cannot identify the next phrase to query.

M.1.3 Auto-seed

Δ Recall -0.3%, Δ FA -6.4% relative (12 hard fails). With C_0 empty, the first `curate` has no comparison context, and the importance-tag space is undefined until C fills up. Aggregate Recall barely moves because the policy still curates the right number of documents; FA Recall drops sharply because the policy retains the wrong documents during the cold-start window.

Example. Q394 asks for an investment-bank founder who was a Sagittarius rugby player born in the 1920s (gold answer: Ron Jarden). Full `Harness-1` identifies Ron Jarden on T2 with the auto-seeded curated set already containing related New Zealand-rugby chunks for comparison, reads `doc_1124` on T4, `verify`-confirms on T6, and ends in 8 turns with FA recall 1.0. The ablation goes through 11 `search_corpus` calls before reading `doc_49990` on T18, but with no warm C_0 to anchor importance against, every candidate looks equally plausible and the trained `verify` call on T20 fails because the supporting chunks were never curated.

M.1.4 Evidence graph

Δ Recall -2.6%, Δ FA -5.4% relative (10 hard fails). With the graph block hidden, the policy can no longer see which entities bridge documents in C . On failing queries it searches more (91.8% of actions vs. 88.9%) and `read_document` drops 3 \times , leaving the curated set without the bridging entity chunks needed to close the chain.

Example. Q754 asks about a youngest-eligible governor elected in a newly formed territory with a self-deprecating slogan and a flamboyant predecessor (gold: Fernando de la Rúa). Full `Harness-1` leverages the evidence graph to follow the nickname-territory-predecessor chain across multiple documents, lands on the correct answer by T34 (FA 1.0). The ablation knows nothing about which entities co-occur, so on T39 it tries “Fernando de la Rúa” as a free-text guess without the structural anchor, never connecting it to the “baby bottle” / “pacifier” nickname or the territory-status clue, and ends at FA 0.

M.1.5 verify tool

Δ Recall -3.1%, Δ FA -3.9% relative (12 hard fails). When `verify` returns unavailable, `read_document` drops from 5.1% to 0.8% on failing queries and search rises to 94.6%.

Example. Q751 asks for the full name of a University of Arizona management professor with a 2010 PhD whose undergraduate university was founded in the 1800s by a specific person (gold: William J. Becker, UCD). Full `Harness-1` reads `doc_15523` on T28, then on T32–T36 traces “William Becker” through the “University College Dublin” bridge with `verify` never explicitly invoked but *implicitly* usable. Without `verify` the policy has no closed-loop check and rushes to commit documents to C ; on T28–T38 it issues ten unverified searches and never re-reads its own candidate, so the curated set ends up with no Becker chunk at all.

M.1.6 review_docs tool

Δ Recall +2.4%, Δ FA -3.9% relative (9 hard fails). The aggregate Recall delta is mildly positive because the policy substitutes additional turn-level reads for the bulk sweep, but FA Recall still drops because the policy cannot economically scan the curated set when partial information from many documents must be combined. read_document drops $4\times$ (4.0% \rightarrow 1.1%) on failing queries and search rises to 93.6%.

Example. Q912 asks for the doctorate university of a researcher named on a 1995–2005 whale-research paper with 12 students and 4 project leaders (gold: University of Hawai'i at Manoa). Full Harness-1 reads doc_33637 on T4, calls verify on T6, then uses review_docs on T7–T13 to sweep the curated set for “Adam Frankel” co-occurring with “UH Manoa”; ends in 15 turns. The ablation reads the same doc_33637 on T2 but with no bulk re-read it loses track of the candidate after a handful of new searches and spends T15–T39 thrashing through name-extraction guesses.

M.1.7 Content-fingerprint dedup

Δ Recall +4.6%, Δ FA +1.6% relative (0 hard fails). This is the one ablation that nominally improves the metric. The cause is mechanical: BC+ qrels occasionally contain near-duplicate gold documents (e.g., a press release and an article that quotes it), and Harness-1’s MinHash–LSH dedup ($\theta_{\text{Jaccard}}=0.85$) sometimes collapses two such gold variants into one, costing one gold ID per affected query. Disabling dedup recovers those IDs but inflates the curated context with redundant chunks. Dedup is by design a token-budget mechanism, not a recall mechanism; the answer-accuracy benchmark in Appendix O prefers the deduped variant in the majority of paired queries despite the slightly lower curated recall. We report this row in the table rather than hide it, because the trade-off is the design intent.

M.2 All harness mechanisms disabled

Δ Recall -12.2%, Δ FA -6.4% relative. With every Harness-1 mechanism off simultaneously, the policy has the same action vocabulary but no compact state to reason over. Recall drops to 0.513 and FA recall to 0.624 — a larger relative Recall drop than any single ablation. The agent keeps searching, sweeping a wide raw-document space, but cannot rank what it has seen. This is the cleanest evidence in the paper that the gains attributed to Harness-1 come from the harness rather than from capability implicit in the underlying model: on the same queries, the trained checkpoint searches comparably hard but cannot discriminate among candidates once the harness is removed.

N Evaluation Recipe

Metric Definitions For each query, let A_q denote the annotated *answer-bearing* documents: documents that directly contain or support the final answer. Let R_q denote the full annotated relevant set used for search-quality recall. On benchmarks with extra evidence annotations, R_q may include additional supporting documents beyond A_q ; in BrowseComp+ and Web, for example, the answer-bearing documents are a subset of the broader relevant set. On benchmarks without separate answer-document annotations, the evaluator sets $A_q = R_q$.

Let C_q be the final curated set returned by the search subagent, after normalizing retrieved chunks to document identifiers. Let P_q be the trajectory pool: all documents encountered anywhere in the episode, again normalized to document identifiers. We report three recall-oriented metrics:

$$\text{Recall}(q) = \frac{|C_q \cap R_q|}{|R_q|}, \quad \text{FinalAnswerRecall}(q) = \frac{|C_q \cap A_q|}{|A_q|},$$

$$\text{TrajectoryRecall}(q) = \frac{|P_q \cap R_q|}{|R_q|}.$$

Dataset-level scores are macro-averages over queries. For fact-level datasets, the same definitions apply with facts in place of documents: a fact is counted as covered if any annotated chunk for that fact is retrieved.

Final-Answer Recall is not mathematically constrained to be above or below Recall. When $A_q \subset R_q$, the denominators differ: Recall averages over all relevant annotations, while Final-Answer

Recall averages only over answer-bearing annotations. Thus a system that finds answer documents but misses many auxiliary supporting documents can have Final-Answer Recall $>$ Recall, whereas a system that finds supporting documents but misses answer documents can have Final-Answer Recall $<$ Recall. Trajectory Recall is an upper-bound-style diagnostic for discovery before final curation; it need not equal final Recall because the agent may encounter evidence and later omit it from C_q .

Reranking for released open-agent harnesses. Search-R1 and Tongyi DeepResearch are evaluated with their released agent harnesses, since the harness is part of each method. These harnesses do not natively return an importance-tagged curated set capped at 30 documents. We therefore standardize their final outputs with a shared reranker. For each query, we collect all unique documents encountered during the episode and normalize chunks to document identifiers. If the trajectory pool contains at most 30 documents, we use the full pool as the curated set C_q without reranking. Otherwise, we rerank the trajectory pool with the shared reranker and keep the top-30 documents as C_q . Recall and Final-Answer Recall are computed from C_q , while Trajectory Recall is computed from the full trajectory pool P_q before reranking.

For Search-R1, this final-set construction usually does not change the retrieval outcome much. Although Search-R1 is evaluated under the same protocol, its released harness typically performs only four to five text-search rounds in our runs. This is consistent with its training setup, which optimizes short search-and-answer trajectories rather than long-horizon evidence collection and curation. As a result, Search-R1 rarely accumulates a trajectory pool much larger than the final document budget. Table 2 reflects this behavior: Search-R1’s Recall and Trajectory Recall are nearly identical across benchmarks, suggesting that there is little additional evidence in the trajectory pool to recover through post-hoc curation.

Answer-accuracy pipeline (modular RAG protocol). To decouple retrieval from generation, each subagent’s curated set is forwarded to a frozen frontier generator that sees *only* the query and curated documents, with no trajectory, intermediate reasoning, or tool-use history. For each (subagent \times benchmark) condition, we read the saved trajectory JSON, extract the curated document identifiers and corresponding chunk text from observation logs, format with the prompt in Appendix I.3, and dispatch to four generators: GPT-5.4, Sonnet-4.6, Opus-4.6, and Kimi-K2.5. We use the same benchmark instances as the main evaluation. Closed-Book and Naive-RAG (top-10 BM25+dense) anchor the protocol. The judge (Appendix I.4) is GPT-5.4 at $T=0$ with the BrowseComp+ template; we report the strict CORRECT rate. Every (query, generator, generated-answer, verdict) tuple is saved.

O Modular RAG: Curated Sets Yield Higher Answer Accuracy

Search quality measures what the subagent returns; what ultimately matters is the answer the downstream system gives. Under the modular RAG protocol described above, the answer-generator LLM sees only the curated set, never the trajectory. Thus, any accuracy difference between subagent rows is attributable to the curated set alone.

Three patterns are visible in Figure 6. First, Closed-Book and Naive RAG are weak on the hardest displayed benchmark (BC+), confirming that these tasks require deep agentic search; parametric knowledge and single-hop retrieval are insufficient. Second, the row ordering across subagents broadly tracks the search-quality ordering from Table 2: better curated sets yield higher answer accuracy across all four generators. Third, column variation within a row is modest for strong searchers, since every frontier generator can converge on the right answer when given the right documents, and larger for weaker searchers, where the generator must compensate for missing evidence. Because the generator never sees the trajectory, the rank ordering of subagent rows is a clean test of curated-set quality: HarnessS-1’s curated sets dominate the open subagents in the reported modular RAG settings.

P Harness as Confound: Same LLM, Different Harness

A natural concern is that all frontier baselines are evaluated under the Context-1 harness, not under HarnessS-1’s harness. This is deliberate: we compare *trained agents* (Context-1 RL, HarnessS-1

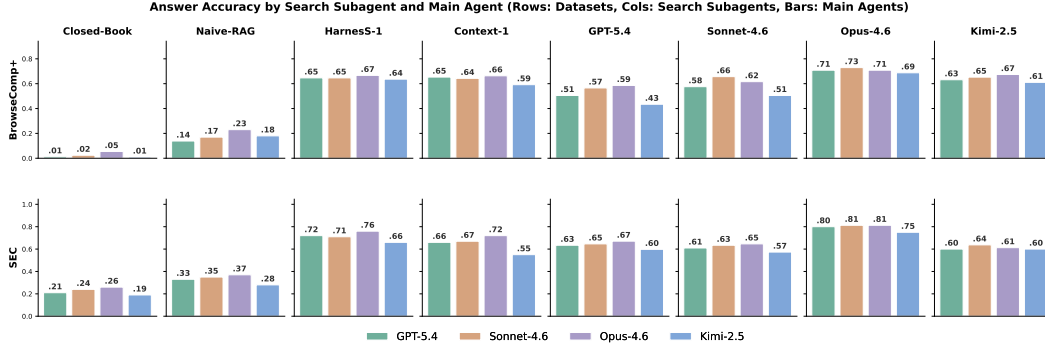


Figure 6: **Answer accuracy (%)** under the modular RAG protocol. Each subplot is a benchmark; each group of bars is a search method (or baseline); each bar colour is a different frontier answer-generator. *Closed-Book*: generator answers from parametric knowledge alone. *Naive RAG*: single BM25+dense top-10 chunk retrieval, no agentic search. All other groups: an agentic search sub-agent curates the document set; the generator receives *only* query and curated documents, no trajectory. Judging follows the official BrowseComp+ template (GPT-5.4, $T=0$).

RL) each running in their own harness, while frontier LLMs serve as untrained baselines using Context-1’s harness as a common reference environment.

To quantify the harness contribution alone, we ran GPT-5.4 under three harness conditions on three in-domain benchmarks: **Naive Search-Add** (minimal loop, no curation, no compression, no dedup); **Context-1 Harness** (production Context-1 environment); **HarnessS-1 Harness** (full HarnessS-1 environment).

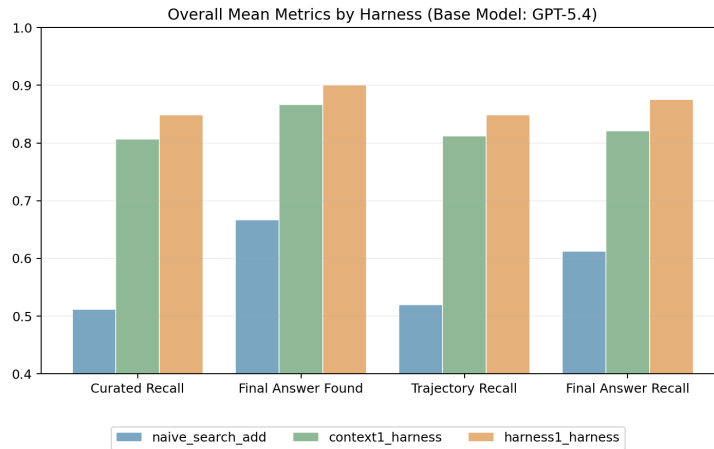


Figure 7: **Same LLM (GPT-5.4), different harness.** All four retrieval metrics improve monotonically as the harness grows richer: naive search-add \rightarrow Context-1 harness \rightarrow HarnessS-1 harness. The harness contributes a +4 point recall gain on top of Context-1.

The progression is monotonically improving. Curated Recall jumps from 0.511 (naive) to 0.807 (Context-1) to 0.849 (HarnessS-1); Final-Answer Recall from 0.612 to 0.821 to 0.876. A +4.2 point recall gain is available to GPT-5.4 simply by switching from Context-1’s harness to HarnessS-1’s, *without any RL training*. This validates the central thesis: the harness is a compute-allocation mechanism that materially changes how much a fixed model can discover.

Q Qualitative Case Studies

The following abridged trajectories show where the harness changes the search process: auto-seeding creates an editable state, curation promotes evidence, exact search recovers needles, backtracking removes false leads, and full-document reading resolves the final answer. Routine prompt boilerplate, redundant curation calls, false-positive documents, and repeated metadata are omitted; the unabridged logs are retained separately.

Q.1 Case Study A: Stateful Search over a Drug-Approval Query

BrowseComp+ query

Query. An agent is not indicated for patients who weigh 30 kg or are 10 years old. It can be taken via oral administration, and one of the side effects reported is nausea. A study from 2022 mentions that its half-life is less than 10 hours and that the liver primarily eliminates it. The study is authored by two people, published in December, and features 31 references. The application for tablets containing the agent was received on March 23, and it was approved eight months later by a letter. What is the full name (as mentioned in the letter) of the Regulatory Project Manager to be called in case of any questions?

Gold answer. Alicia Moruf

Outcome. 18 turns, tool diversity 6, curated recall 1.0, final-answer recall 1.0.

Why this case is interesting. The query is a regulatory needle: the agent must connect label constraints, a pharmacology study, an NDA package, and an approval letter naming the project manager. The trajectory shows the harness doing useful bookkeeping rather than merely storing a transcript. Auto-seeding creates the initial regulatory evidence set, curation separates maribavir and FDA-package evidence from distractors, exact search recovers from brittle phrase matching, and full-document reading exposes the answer sentence.

How the trajectory is abridged. The full log keeps every token. Here we omit the system prompt, repeated curation confirmations, low-value false-positive documents, and two no-op verify/review calls caused by chunk-id mismatch. The retained turns preserve the model reasoning, exact tool calls, decisive returns, and state snapshots needed to see the harness at work.

Trajectory overview.

Turn	Tool	Policy move	Harness state / why it matters
0	<code>search</code>	Search all medical clues at once	auto-seed creates the first editable curated set (6 docs, because only 6 were returned)
1	<code>curate</code>	Promote Livtency and pharmacology candidates	curation separates strong maribavir evidence from low-value distractors
2	<code>search</code>	Search the 2022 maribavir study	connects half-life, hepatic metabolism, two authors, and references
3	<code>curate</code>	Promote study and label chunks	working memory now contains drug identity plus PK evidence
4	<code>search</code>	Search FDA approval-letter wording	same document family reappears with NDA context
5	<code>curate</code>	Consolidate FDA package and study chunks	curated state keeps both regulatory and pharmacology evidence
6	<code>search</code>	Search by regulatory-manager wording	narrows to approval package chunks
7	<code>curate</code>	Add later FDA-letter chunks	candidate pool now includes the likely answer document
8	<code>grep</code>	Exact grep for phrase	a failed exact match prevents over-trusting a brittle pattern
10	<code>search</code>	Search by NDA number	needle appears: Alicia Moruf in the approval letter
11	<code>curate</code>	Promote NDA-letter chunk	final evidence enters curated memory
12	<code>read</code>	Read full approval-letter chunk	outer-tier memory exposes the decisive text

Turn	Tool	Policy move	Harness state / why it matters
13	curate	Reaffirm final evidence	curated set keeps approval letter high

Phase 1 - Auto-seeded working memory from a broad medical search

Reasoning excerpt

Key constraints: agent not indicated for patients ≥ 30 kg or 10y; oral administration; side effect nausea; 2022 study mentions half-life < 10 h, primarily liver elimination; study by two authors, December, 31 references; tablet application received 23 Mar; approved 8 months later via letter identifying a Regulatory Project Manager. Likely a US FDA approval letter for an oral drug for weight ≥ 30 kg/age ≥ 10 y contraindication. The unique medicine clue is the age/weight contraindication plus nausea side effect. Search angle: distinct clinical label and side effect plus half-life/liver elimination trait.

Exact tool call

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "oral contraindicated for under 30 kg age 10 nausea half-life less than 10 hours liver
eliminated 2022 study December 31 references FDA approval letter Regulatory Project Manager March
23 seventeen Feb 2023 app tablets received March 23 approved eight months letter"
    }
  }
]
```

Search return: first hit creates the regulatory anchor

DOCUMENT ID: 59631.0 (743 tokens) CENTER FOR DRUG EVALUATION AND RESEARCH Approval Package for: APPLICATION NUMBER: 215596Orig1s000 Trade Name: LIVTENCITY oral tablets Generic or Proper Name: maribavir Sponsor: Takeda Pharmaceuticals USA, Inc. Approval Date: November 23, 2021 Indication: For the treatment of adults and pediatric patients 12 years of age and older and weighing at least 35 kg with post-transplant cytomegalovirus (CMV) infection/disease that is refractory to treatment (with or without genotypic resistance) with ganciclovir, valganciclovir, cidofovir or foscarnet. CENTER FOR DRUG EVALUATION AND RESEARCH 215596Orig1s000 CONTENTS Reviews / Information Included in this NDA Review. Approval Letter X Other Action Letters Labeling X REMS Officer /Employee List X Integrated Review(s) X Product Quality Review(s) X Other Reviews X Risk Assessment and Risk Mitigation Review(s) X Proprietary Name Review(s) X Administrative/Correspondence Document(s) X CENTER FOR DRUG EVALUATION AND RESEARCH APPLICATION NUMBER: 215596 Orig1s000 APPROVAL LETTER NDA 215596 NDA APPROVAL Takeda Pharmaceuticals USA, Inc. Attention: Michael A. Cronin, PharmD Director, Global Regulatory Affairs 650 ...

Curation update

Curated set updated (6/30): 25814[high], 59631[high], 15546[*fair*], 50858[*fair*], 66387[low], 97801[low]

State after turn 1

```
pool_size: 6
curated_size: 6/30
importance: very_high=0, high=2, fair=2, low=2
top_curated:
  - 25814 [high]
  - 59631 [high]
  - 15546 [fair]
```

- 50858 [fair]
- 66387 [low]
- 97801 [low]

Phase 2 - The policy pivots to pharmacokinetic evidence

Reasoning excerpt

The first search hit a known FDA submission doc for Livtency (maribavir). That fits oral administration, nausea side effect, and contraindication at 30kg/10y, but the drug class/youth contraindication constraint is strong. We need the 2022 study page and the approval letter identifying the Regulatory Project Manager. The search angle using the study constraints is still relevant; I'll first search for the 2022 study on maribavir to bridge to the approval letter.

Exact tool call

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "maribavir 2022 study liver eliminated half-life less than 10 hours oral migration 31 references December two authors"
    }
  }
]
```

Search return: study and label evidence

DOCUMENT ID: 83797.4 (723 tokens) As a postmarketing commitment, the FDA is requiring the manufacturer to conduct phenotypic analysis of maribavir against human CMV. Substitutions designated as a high priority for analysis are pUL97 M460I/T, A594E/P/T/V, L595F/W, and C603R/W/Y. Substitutions designated as medium priority are pUL97 A440V, V466M, A591V, E596G, K599E, and C607F/Y. Substitutions designated as low priority are pUL97 P132L, L405P, C518Y, I610T, and A613V; pUL27 P10L, N289D, H29Y, D298G, N300G, P307L, V310A, D351N, and I367V; and pUL54 S290R and K475S. Maribavir does not prolong the QT interval to a clinically important extent.^{1,19} Pharmacokinetics Following oral administration, median T_{max} is 1 to 3 hours.^{1,19,20} Plasma exposure (C_{max} and AUC) increased approximately dose proportionally following a single dose of 50 to 1600 mg and multiple doses of up to 2400 mg/day.^{1,20} Compared with administration in the fasted state, administration with a moderate-fat meal resulted in a small reduction in AUC_{0-infinity} (geometric mean ratio, 0.864; 90% CI, 0.804–0.929) and C_{max} (0.722; 90% CI, 0.656–0.793). Steady state is reached within 2 days with twice-daily dosing.¹ Mean steady-state volume of distribution is 27.3 L. Maribavir is 98% plasma protein bound. The blood-to-plasma ratio is 1.37.^{1,20} Maribavir does not appear to cross the blood-brain barrier.⁹ Maribavir, the parent drug, is responsible for the ...

DOCUMENT ID: 17183.8 (746 tokens) due to the parent drug. Following oral administration, plasma maribavir exposure (C_{max} and AUC) increased approximately dose-proportionally following a single dose of 50 to 1600 mg (0.125 to four times the recommended dose) and multiple doses up to 2400 mg per day (three times the recommended daily dose). Maribavir PK is time-independent. With twice-daily dosing, steady state is reached within 2 days, with mean accumulation ratios of C_{max} and AUC ranging from 1.37 to 1.47. The pharmacokinetic properties of maribavir following administration of LIVTENCITY are displayed in Table 4. The multiple-dose pharmacokinetic parameters are provided in Table 5. ⁹ Reference ID: 4893709 This label may not be the latest approved by FDA. For current labeling information, please visit Table 4: Pharmacokinetic Properties of Maribavir Absorption T_{max} (h), median 1.0 to 3.0 Distribution Mean apparent steady-state volume of distribution (V_{ss}, L) 27.3 % bound to human plasma proteins 98.0 across the concentration range of 0.05–200 ug/mL Blood-to plasma ratio 1.37 Elimination Major route of elimination Hepatic metabolism Half-life (t_{1/2}) in transplant patients (h), mean 4.32 Oral clearance (CL/F) in transplant patients (L/h), mean 2.85 Metabolism Metabolic pathways CYP3A4 (major) and CYP1A2 (minor) Excretion % of dose excreted as total 14C (unchanged drug) in urine 61 (<2) % of dose excreted as total 14C ...

DOCUMENT ID: 73877.2 (477 tokens) B: May be acceptable. Either animal studies show no risk but human studies not available or animal studies showed minor risks and human studies done and showed no risk. C: Use with caution if benefits outweigh risks. Animal studies show risk and human studies not available or neither animal nor human studies done. D: Use in LIFE-THREATENING emergencies when no safer drug available. Positive evidence of human fetal risk. X: Do not use in pregnancy. Risks involved outweigh potential benefits. Safer alternatives exist. NA: Information not available. Pharmacology Mechanism of Action Antiviral activity mediated by competitive inhibition of the protein kinase activity of human CMV enzyme pUL97, which inhibits protein phosphorylation Maribavir inhibited human CMV replication in virus yield reduction, DNA hybridization, and plaque reduction assays in human lung fibroblast cell line, human embryonic kidney, and human foreskin fibroblast cells Absorption Peak plasma time: 1-3 hr AUC: 128 ug*hr/mL Peak plasma concentration: 17.2 ug*hr/mL Distribution Vd (steady-state) : 27.3 L Plasma bound: 98% Blood-to-plasma ratio: 1.37 Metabolism Metabolized by CYP3A4 (major) and CYP1A2 minor Elimination Clearance: 2.85 L/hr Half-life: 4.32 hr Excretion: Urine (61% [<2% unchanged]); feces (14% [5.7%]) Administration Oral Administration Take with or without food Storage Store at 20-25oC (68-77oF); excursions ...

Curation update

Curated set updated (9/30): 17183[high], 25814[high], 59631[high], 83797[high], 15546[fair], 50858[fair], 73877[fair], 66387[low], 97801[low]

State after turn 3

pool_size: 9
curated_size: 9/30
importance: very_high=0, high=4, fair=3, low=2
top_curated:
- 17183 [high]
- 25814 [high]
- 59631 [high]
- 83797 [high]
- 15546 [fair]
- 50858 [fair]
- 73877 [fair]
- 66387 [low]
- 97801 [low]

Phase 3 - Exact search misses, so state prevents a blind loop

Exact grep call

```
[  
  {  
    "tool": "grep_corpus",  
    "params": {  
      "pattern": "Maribavir.*Regulatory Project Manager"  
    }  
  }  
]
```

Grep return

No results found
[Context: 11553/30720]

Recovery search by NDA number

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "NDA 215596 letter to maribavir regulatory project manager NDA 215596"
    }
  }
]
```

Decisive return around answer

[... clipped before ...]

REPORTING REQUIREMENTS We remind you that you must comply with reporting requirements for an approved NDA (21 CFR 314.80 and 314.81). POST APPROVAL FEEDBACK MEETING New molecular entities qualify for a post approval feedback meeting. Such meetings are used to discuss the quality of the application and to evaluate the communication process during drug development and marketing application review. The purpose is to 4 For the most recent version of a guidance, check the FDA guidance web page at 5 6 Reference ID: 4893709 NDA 215596 Page 6 U.S. Food and Drug Administration Silver Spring, MD 20993 learn from successful aspects of the review process and to identify areas that could benefit from improvement. If you would like to have such a meeting with us, call the Regulatory Project Manager for this application. If you have any questions, call Alicia Moruf, PharmD, MPH, RAC-US, Regulatory Project Manager, at 301-796-3953. Sincerely, {See appended electronic signature page} John Farley, MD, MPH Director Office of Infectious Diseases Office of New Drugs Center for Drug Evaluation and Research ENCLOSURE(S): – Content of Labeling o Prescribing Information o Patient Package Insert – Carton and Container Labeling Reference ID: 4893709

This is a representation of an electronic record that was signed electronically. Following this are manifestations of any and all electronic signatures for this electronic record.

DOCUMENT ID: 73877.0 (719 tokens)

title: maribavir (Rx)
date: 2025-01-01

Dosing & Uses

Dosage Forms & Strengths

tablet

– 200mg

Cytomegalovirus Infection

Indicated for t
[... clipped after ...]

Curation update

Curated set updated (10/30): 17183[high], 25814[high], 15546[[fair](#)], 50858[[fair](#)], 59631[[fair](#)], 73877[[fair](#)], 66387[[low](#)], 72434[[low](#)], 83797[[low](#)], 97801[[low](#)]

State after turn 11

pool_size: 11
curated_size: 10/30
importance: very_high=0, high=2, fair=4, low=4
top_curated:

- 17183 [high]
- 25814 [high]
- 15546 [fair]
- 50858 [fair]
- 59631 [fair]
- 73877 [fair]
- 66387 [low]
- 72434 [low]
- 83797 [low]
- 97801 [low]

Phase 4 - Full-document reading confirms the final evidence

Exact read call

```
[
  {
    "tool": "read_document",
    "params": {
      "doc.id": "59631_3"
    }
  }
]
```

Full-document excerpt around answer

[... clipped before ...]
 ible at FDA.gov.5 Information and Instructions for completing the form can be found at FDA.gov.6
 REPORTING REQUIREMENTS We remind you that you must comply with reporting requirements for an
 approved NDA (21 CFR 314.80 and 314.81). POST APPROVAL FEEDBACK MEETING New molecular
 entities qualify for a post approval feedback meeting. Such meetings are used to discuss the quality of the
 application and to evaluate the communication process during drug development and marketing
 application review. The purpose is to 4 For the most recent version of a guidance, check the FDA
 guidance web page at 5 6 Reference ID: 4893709 NDA 215596 Page 6 U.S. Food and Drug
 Administration Silver Spring, MD 20993 learn from successful aspects of the review process and to
 identify areas that could benefit from improvement. If you would like to have such a meeting with us, call
 the Regulatory Project Manager for this application. If you have any questions, call Alicia Moruf, PharmD,
 MPH, RAC-US, Regulatory Project Manager, at 301-796-3953. Sincerely, {See appended electronic
 signature page} John Farley, MD, MPH Director Office of Infectious Diseases Office of New Drugs Center
 for Drug Evaluation and Research ENCLOSURE(S) - Content of Labeling o Prescribing Information o
 Patient Package Insert - Carton and Container Labeling Reference ID: 4893709

 This is a representation of an electronic record that was signed electronically. Following this are
 manifestations of any and all electronic signatures for this electronic record

-----/-----
 s/ -----JOHN J FARLEY
 11/23/2021 03:34:13 PM Signature Page 1 of 1 Reference ID: 4893709 (
 [Context: 10329/30720]

Final curation update

Curated set updated (10/30): 17183[high], 25814[high], 59631[high], 15546[high], 50858[high], 73877[high],
 66387[low], 72434[low], 83797[low], 97801[low]

State after turn 13

pool_size: 11

curated_size: 10/30
importance: very_high=0, high=3, fair=3, low=4
top_curated:
- 17183 [high]
- 25814 [high]
- 59631 [high]
- 15546 [fair]
- 50858 [fair]
- 73877 [fair]
- 66387 [low]
- 72434 [low]
- 83797 [low]
- 97801 [low]

Q.2 Case Study C: Exact-Date Recovery in SEC Filings

SEC query

Query. There is a company where a senior financial executive with over two decades of consumer packaged goods experience joined as a senior vice president in the early spring of the mid-2020s. Around the early summer of the same year, a different long-serving financial executive concluded their tenure after announcing their decision to step down several months earlier. That same date marked when the spring hire assumed the top financial role, succeeding the retiring executive. On what date did this leadership transition in the financial executive role become effective?

Gold answer. July 3, 2025

Outcome. 9 turns, tool diversity 5, curated recall 1.0, final-answer recall 1.0.

Why this case matters. SEC queries often fail when the agent treats filings as isolated snippets. This trajectory shows the source-family behavior behind the SEC gains: auto-seeding starts inside the right filing family, curation preserves the relevant 8-Ks, `read_document` exposes the full executive-transition language, and `grep_corpus` verifies the exact effective date across later filings.

Trajectory overview.

Turn	Tool	Policy move	Harness state / why it matters
0	<code>search</code>	Search executive-transition clues in SEC filings	first hit already binds company, outgoing CFO, incoming CFO, and effective date
1	<code>curate</code>	Promote the relevant 8-Ks	curated set separates the correct filing family from other finance hits
3	<code>read</code>	Read the May 8-K	full text resolves succession details rather than relying on snippets
4	<code>curate</code>	Keep the read filing high	state retains the source document for final answer recall
6	<code>grep</code>	Grep the exact effective date	regex recovers later proxy mentions of the same transition
7	<code>curate</code>	Promote corroborating proxy chunks	final state keeps multiple filings about the CFO transition

Phase 1 - Auto-seeding finds the right company and date

Exact search call

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "spring 2024 senior vice president food consumer packaged goods 20 years CPG effective CFO transition summer 2024 retire step down senior financial executive announced decision months earlier"
    }
  }
]
```

Decisive first return

```
# DOCUMENT ID: 0001104659-25-006698.0 (387 tokens)
Company: Simply Good Foods Co
Filing Date: 2025-01-28
Form: 8-K
Item: Item 5.02
```

Chunk 0

Item 5.02 Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers; Compensatory Arrangements of Certain Officers.

(b) On January 27, 2025, Shaun P. Mara, Chief Financial Officer of The Simply Good Foods Company (the "Company"), notified the Company of his decision to retire effective July 3, 2025.

Mr. Mara will remain as Chief Financial Officer through the date of his retirement, at which time he will resign from all positions with the Company and its subsidiaries.

In connection with Mr. Mara's retirement, on January 28, 2025, the Company announced the hiring of Mr. Christopher J. Bealer as Senior Vice President of Finance of the Company, as of April 1, 2025. Mr. Bealer is expected to succeed Mr. Mara as Chief Financial Officer upon Mr. Mara's retirement on July 3, 2025.

Mr. Bealer has almost 23 years of experience in consumer packaged goods and consumer durables in North America and global markets. In his most recent executive leadership position as EVP Corporate Controller at Reckitt Benckiser Group PLC ("Reckitt"), Mr. Bealer leads a global team of over 600 employees accountable for global controls, financial shared services, group reporting and accounting, as well as leading the functional operating model transformation. Prior to his current and prior roles at Reckitt, Mr. Bealer served as Regional CFO for Whirlpool Corporation for both the North American and Europe, Middle East and Africa (EME [... clipped after ...])

Curation update

Curated set updated (7/30): 0001104659-25-006698.0[high], 0001104659-25-053287.0[high], 0001104659-25-121934.20[high], 0000909832-25-000101.19[fair], 0001364954-25-000039.57[fair], 0001673985-25-000053.24[fair], 0001530950-25-000260.18[low]

State after turn 1

```
pool_size: 7
curated_size: 7/30
importance: very_high=0, high=3, fair=3, low=1
top_curated:
  - 0001104659-25-006698.0 [high]
  - 0001104659-25-053287.0 [high]
  - 0001104659-25-121934.20 [high]
  - 0000909832-25-000101.19 [fair]
  - 0001364954-25-000039.57 [fair]
  - 0001673985-25-000053.24 [fair]
  - 0001530950-25-000260.18 [low]
```

Phase 2 - Full-document reading keeps the answer grounded

Exact read call

```
[
  {
    "tool": "read_document",
    "params": {
      "doc.id": "0001104659-25-053287.0"
    }
  }
]
```

```
}  
}  
]
```

Full-document excerpt around transition

Document (1500 tokens)
Company: Simply Good Foods Co
Filing Date: 2025-05-28
Form: 8-K
Item: Item 5.02
Chunk 0

Item 5.02

Departure of Directors or Certain Officers; Election of Directors; Appointment of Certain Officers;
Compensatory Arrangements of Certain Officers.

Appointment of New CFO

In connection with the previously announced Chief Financial Officer transition, on May 23, 2025, The Simply Good Foods Company (the "Company") appointed Christopher J. Bealer as the Company's Chief Financial Officer effective July 3, 2025. Mr. Bealer is currently serving as the Company's Senior Vice President, Finance having joined the Company on April 1, 2025. As Chief Financial Officer, Mr. Bealer will continue to report to Geoff E. Tanner, the Company's President and Chief Executive Officer.

Mr. Bealer, age 48, has almost 23 years of experience in consumer-packaged goods and consumer durables in North America and Global markets. Mr. Bealer was most recently EVP Corporate Controller at Reckitt Benckiser Group PLC ("Reckitt") from March 2024 to March 31, 2025, and CFO Health at Reckitt from April 2021 to February 2024. In his EVP Corporate Controller role with Reckitt, Mr. Bealer led a global team of over 600 employees accountable for global controls, financial shared services, group reporting and accounting, as well as leading the functional operating model transformation. Prior to his roles at Reckitt, M
[... clipped after ...]

State after turn 4

pool.size: 7
curated.size: 7/30
importance: very_high=0, high=3, fair=3, low=1
top_curated:
- 0001104659-25-006698.0 [high]
- 0001104659-25-053287.0 [high]
- 0001104659-25-121934.20 [high]
- 0000909832-25-000101.19 [fair]
- 0001364954-25-000039.57 [fair]
- 0001673985-25-000053.24 [fair]
- 0001530950-25-000260.18 [low]

Phase 3 - Exact grep adds corroborating filings

Exact grep call

```
[
  {
    "tool": "grep_corpus",
    "params": {
      "pattern": "effective July 3, 2025"
    }
  }
]
```

Grep return

[... clipped before ...]
| CHRISTOPHER J.
BEALER || || | TIMOTHY R.
KRAFT || || | AMY C.
HELD || || | RYAN A.
THOMAS || || | SHAUN P.
MARA ||
|| | President and Chief Executive Officer || || | Chief Financial Officer || || | Chief Legal and Corporate
Development Officer || || | SVP and Chief Human Resources Officer || || | SVP and General
Manager Atkins and OWYN || || | Former Chief Financial Officer ||

Mr. Bealer became an NEO for the first time in fiscal year 2025. Mr. Bealer was hired on March 31, 2025. Mr. Mara retired as our Chief Financial Officer and resigned effective July 3, 2025. Mr. Mara's executive compensation is outlined further in the compensation tables beginning on page 54 but is otherwise not discussed with other NEOs except where noted. Executive Summary Our compensation program for our NEOs and other executive officers is designed to meet the following primary objectives:

|| | Management Development and Continuity.
Attract, retain and motivate individuals of superior ability and managerial talent to develop, grow and manage our business by offering competitive compensation opportunities with both short-term and significant long-term components || || | Pay for Performance.
[... clipped after ...]

Curation update

Curated set updated (9/30): 0001104659-25-006698.0[high], 0001104659-25-053287.0[high], 0001104659-25-121934.20[high], 0001104659-25-121934.21[high], 0001104659-25-121934.42[high], 0000909832-25-000101.19[fair], 0001364954-25-000039.57[fair], 0001673985-25-000053.24[fair], 0001530950-25-000260.18[low]

State after turn 7

pool_size: 10
curated_size: 9/30
importance: very_high=0, high=5, fair=3, low=1
top_curated:
- 0001104659-25-006698.0 [high]
- 0001104659-25-053287.0 [high]
- 0001104659-25-121934.20 [high]
- 0001104659-25-121934.21 [high]
- 0001104659-25-121934.42 [high]
- 0000909832-25-000101.19 [fair]
- 0001364954-25-000039.57 [fair]
- 0001673985-25-000053.24 [fair]
- 0001530950-25-000260.18 [low]

Q.3 Case Study B: Fan-Out Search Resolves an Ambiguous Transfer Query

FRAMES query

Query. Which university did the actor who has appeared in the most movies in the American pie film series (including spinoffs, as of 2020) deliver a commencement address at in 2012?

Gold answer. Dalhousie University, in Halifax, Nova Scotia

Outcome. 16 turns, tool diversity 6, curated recall 1.0, final-answer recall 1.0.

Why this case is interesting. The query is ambiguous at the first hop: several *American Pie* actors are plausible, and the commencement-address clue only becomes useful after the actor identity is narrowed. The trajectory shows why fan-out and tagged memory matter. The agent keeps competing actor hypotheses alive, pivots from Seann William Scott to Eugene Levy, then uses exact search and full-document reading to expose the Dalhousie address evidence.

How the trajectory is abridged. The full log includes many generic web snippets from fan-out search. Here we omit generic pages, repeated low-impact curation text, and the no-op verifier call, while preserving the reasoning and exact calls for the decisive fan-out, candidate pivot, exact search, full read, and final promotion turns.

Trajectory overview.

Turn	Tool	Policy move	Harness state / why it matters
0	fan-out	Fan out over actor identity and address clues	auto-seed stores 8 docs; broader pool contains the key American Pie character page
1	curate	Promote American Pie cast evidence	curation shifts memory from generic hits to candidate actors
2	search	Search Eugene Levy commencement	tests the candidate implied by Noah Levenstein
3	curate	Add Eugene Levy pages	memory now tracks the alternative to Seann William Scott
4	search	Search most-appearances clue	keeps the actor-identity subproblem explicit
5	curate	Promote cast pages	curated set expands but stays tagged
6	search	Search Seann William Scott address	tests a competing hypothesis
7	curate	Demote weak address hits	low tags protect the final set from noise
8	grep	Exact grep for both candidates	fails to settle the address but broadens evidence
9	curate	Fill remaining memory slots	curated set reaches cap; tags govern later eviction
12	search	Search Eugene Levy + 2012 address	directly targets the missing bridge
13	curate	Add university/address candidates	state prepares for final read
14	read	Read Eugene Levy page	full text reveals Dalhousie University
15	curate	Promote Eugene Levy and Dalhousie	final answer evidence receives very_high/high tags

Phase 1 - Fan-out keeps competing hypotheses alive

Reasoning excerpt

We need: Which university did the actor who has appeared in the most movies in the American Pie film series (including spinoffs, as of 2020) deliver a commencement address at in 2012.

We can break it down.

Candidates: Need actor with most appearances in American Pie series.

List of movies: American Pie (1999), American Pie 2 (2001), American Pie 3 (2003), American Pie 4 (2008), American Wedding (2003?), Wait 2003? Actually American Wedding (2003) is the third? Let's check full series: American Pie (1999); American Pie 2 (2001); American Wedding (2003); American Reunion (2012); American Pie Presents: Band Party (2007); American Pie Presents: The Wedding (2009); American Pie Presents: The Book of Love (2010); American Pie Presents: The Naked Mile (2006); American Pie Presents: The First Kiss (2009). Also spinoffs: American Wedding (2003), American Reunion 2012. The actor who appears in the most movies: I recall Skyler or others? Let's recall main casts: Jason Biggs (Jim), Alyson Hannigan (Michelle), Natasha Lyonne (Stifler?), actually Stifler is Seann William Scott. He appears in many? Let's count: Stifler from first through fourth, plus spinoffs? Seann William Scott appears in 1) American Pie 2? Wait Stifler is in first: yes first. In American Pie 2? yes Stifler. In American Wedding? maybe he appears as Stifler. In American Pie Presents: The Book of Love? Not sure. Also spinoffs: He appears in American Pie Presents: The Book of Love? Let's recall: Seann William Scott as Stifler appears in original series plus the spin-offs? There's also his film within the series. Let's search.

So likely the actor is Seann William Scott, or maybe the actor is Tom Gills? No. The actor with most appearances likely is Seann William Scott or Jason Biggs? Jason Biggs appears first through first three? He appears also in reunion and maybe New movies. Let's verify counts.

We need actor with most appearances among all US movies up to 2020. Let's search for 'Seann William Scott American Pie appearances' or 'American Pie actor with most appearances'.

We also need commencement address in 2012 at a university. Which university? So Seann William Scott delivered commencement address in 2012 at some university. Need to find that.

So we need to find doc about Seann William Scott commencement address 2012. Alternatively, if actor is Jason Biggs, find if he delivered commencement address 2012. But Jason Biggs may have spoken at University of Notre Dame? Let's think: Jason Biggs Speech 2013? Wait.

Let's recall: Seann William Scott on his biography: He gave a commencement address at University of Kansas? Let's quickly search.

Given time, best to do fan_out_search queries.

First query: "Seann William Scott commencement address 2012 university".

Second query: "American Pie series actor most appearances".

Third: "American Pie spinoffs actor who appears in most films".

Fourth: "Jason Biggs commencement address 2012".

Fifth: "American Pie actor commencement 2012".

But my first size: "Seann William Scott commencement address 2012 university" should find doc. Let's do that.

Exact fan-out call

```
[
  {
    "tool": "fan_out_search",
    "params": {
      "queries": [
        "Seann William Scott commencement address 2012 university",
        "most American Pie films actor",
        "American Pie actor who appeared most movies",
        "Seann William Scott 2012 commencement",
        "Jason Biggs commencement 2012"
      ]
    }
  }
]
```

Fan-out return: actor identity clue

[... clipped before ...]

cott – Wikipedia

Scott is best known for his breakout role as Steve Stifler in the American Pie film series (1999–2012). He also starred in a lead role as Doug Glatt in Goon (...

DOCUMENT ID: https://en.wikipedia.org/wiki/Thomas_Ian_Nicholas (44 tokens)

Thomas Ian Nicholas – Wikipedia

In film, Nicholas is best known for his lead roles as Kevin Meyers in the American Pie theatrical film series (1999–2012) and Henry Rowengartner in the sports ...

DOCUMENT ID: [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)](https://en.wikipedia.org/wiki/Chris_Owen_(actor)) (46 tokens)

Chris Owen (actor) – Wikipedia

Chris Owen (born September 25, 1980) is an American actor. He is best known for his supporting role as Chuck "Sherminator" Sherman in the American Pie ...

DOCUMENT ID: https://en.wikipedia.org/wiki/List_of_American_Pie_characters (50 tokens)

List of American Pie characters – Wikipedia

Noah Levenstein, played by Eugene Levy, has appeared in eight of the nine released films. Biggs, Hannigan, Scott, Nicholas, Thomas, Coolidge, Cho, Isfield, ...

DOCUMENT ID: https://en.wikipedia.org/wiki/Tad_Hilgenbrink (52 tokens)

Tad Hilgenbrink – Wikipedia

Tad Hilgenbrink (born October 9, 1981) is an American actor, known for his role as Matt Stifler in American Pie Presents: Band Camp, Chance Marquis in The ...

DOCUMENT ID: [https://en.wikipedia.org/wiki/Chris_Klein_\(actor\)](https://en.wikipedia.org/wiki/Chris_Klein_(actor)) (49 tokens)

Chris Klein (actor) – Wikipedia

He reprised his role in two sequels: American Pie 2 (2001) and American Reunion (2012). Chris Klein. Klein in March 2012. Born. Frederick Christopher Klein.

DOCUMENT ID: https://en.wikipedia.org/wiki/Eddie_Kaye_Thomas (41 tokens)

Eddie Kaye Thomas – Wikipedia

Thomas rose to prominence for his supporting role as Pau

[... clipped after ...]

Curation update

Curated set updated (15/30): [https://en.wikipedia.org/wiki/American_Pie_\(film_series\)](https://en.wikipedia.org/wiki/American_Pie_(film_series))[high], https://en.wikipedia.org/wiki/List_of_American_Pie_characters[high], https://en.wikipedia.org/wiki/Seann_William_Scott[high], https://en.wikipedia.org/wiki/Ben_Carson[fair], https://en.wikipedia.org/wiki/Bradley_Cooper[fair], [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)](https://en.wikipedia.org/wiki/Chris_Owen_(actor))[fair], https://en.wikipedia.org/wiki/Ellis_Paul[fair], https://en.wikipedia.org/wiki/Jason_Biggs[fair], https://en.wikipedia.org/wiki/Magic_Johnson[fair], https://en.wikipedia.org/wiki/Steven_Curtis_Chapman[fair], https://en.wikipedia.org/wiki/Susan_Rice[fair], https://en.wikipedia.org/wiki/Thomas_Ian_Nicholas[fair], https://en.wikipedia.org/wiki/Tim_Reid[fair], https://en.wikipedia.org/wiki/Tucker_Carlson[fair], [https://en.wikipedia.org/wiki/Chris_Klein_\(actor\)](https://en.wikipedia.org/wiki/Chris_Klein_(actor))[low]

State after turn 1

pool_size: 40

curated_size: 15/30

importance: very_high=0, high=3, fair=11, low=1

top_curated:

- [https://en.wikipedia.org/wiki/American_Pie_\(film_series\)](https://en.wikipedia.org/wiki/American_Pie_(film_series)) [high]
- https://en.wikipedia.org/wiki/List_of_American_Pie_characters [high]
- https://en.wikipedia.org/wiki/Seann_William_Scott [high]
- https://en.wikipedia.org/wiki/Ben_Carson [fair]
- https://en.wikipedia.org/wiki/Bradley_Cooper [fair]
- [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)](https://en.wikipedia.org/wiki/Chris_Owen_(actor)) [fair]
- https://en.wikipedia.org/wiki/Ellis_Paul [fair]
- https://en.wikipedia.org/wiki/Jason_Biggs [fair]

- https://en.wikipedia.org/wiki/Magic_Johnson [fair]
- https://en.wikipedia.org/wiki/Steven_Curtis_Chapman [fair]
- https://en.wikipedia.org/wiki/Susan_Rice [fair]
- https://en.wikipedia.org/wiki/Thomas_Ian_Nicholas [fair]
- ... 3 more

Phase 2 - The policy pivots from Seann William Scott to Eugene Levy

Search Eugene Levy address

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "Eugene Levy 2012 commencement address university"
    }
  }
]
```

Search most-appearances clue

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "most American Pie films actor appeared in most movies american pie series"
    }
  }
]
```

Curation after candidate expansion

Curated set updated (21/30): [https://en.wikipedia.org/wiki/American_Pie_\(film\)](https://en.wikipedia.org/wiki/American_Pie_(film))[high], [https://en.wikipedia.org/wiki/American_Pie_\(film_series\)](https://en.wikipedia.org/wiki/American_Pie_(film_series))[high], https://en.wikipedia.org/wiki/Eugene_Levy[high], https://en.wikipedia.org/wiki/List_of_American_Pie_characters[high], https://en.wikipedia.org/wiki/Seann_William_Scott[high], https://en.wikipedia.org/wiki/Ben_Carson[fair], https://en.wikipedia.org/wiki/Bradley_Cooper[fair], [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)](https://en.wikipedia.org/wiki/Chris_Owen_(actor))[fair], https://en.wikipedia.org/wiki/Ellis_Paul[fair], https://en.wikipedia.org/wiki/Jason_Biggs[fair], https://en.wikipedia.org/wiki/John_Cho[fair], https://en.wikipedia.org/wiki/Magic_Johnson[fair], https://en.wikipedia.org/wiki/Shannon_Elizabeth[fair], https://en.wikipedia.org/wiki/Steven_Curtis_Chapman[fair], https://en.wikipedia.org/wiki/Susan_Rice[fair], https://en.wikipedia.org/wiki/Thomas_Ian_Nicholas[fair], https://en.wikipedia.org/wiki/Tim_Reid[fair], https://en.wikipedia.org/wiki/Tucker_Carlson[fair], [https://en.wikipedia.org/wiki/Chris_Klein_\(actor\)](https://en.wikipedia.org/wiki/Chris_Klein_(actor))[low], https://en.wikipedia.org/wiki/Dan_Petronijevic[low], [https://en.wikipedia.org/wiki/Eugene_Levy_\(politician\)](https://en.wikipedia.org/wiki/Eugene_Levy_(politician))[low]

State after turn 5

pool_size: 53
 curated_size: 21/30
 importance: very_high=0, high=5, fair=13, low=3
 top_curated:

- [https://en.wikipedia.org/wiki/American_Pie_\(film\)](https://en.wikipedia.org/wiki/American_Pie_(film)) [high]
- [https://en.wikipedia.org/wiki/American_Pie_\(film_series\)](https://en.wikipedia.org/wiki/American_Pie_(film_series)) [high]
- https://en.wikipedia.org/wiki/Eugene_Levy [high]
- https://en.wikipedia.org/wiki/List_of_American_Pie_characters [high]
- https://en.wikipedia.org/wiki/Seann_William_Scott [high]
- https://en.wikipedia.org/wiki/Ben_Carson [fair]
- https://en.wikipedia.org/wiki/Bradley_Cooper [fair]

- [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)](https://en.wikipedia.org/wiki/Chris_Owen_(actor)) [fair]
- https://en.wikipedia.org/wiki/Ellis_Paul [fair]
- https://en.wikipedia.org/wiki/Jason_Biggs [fair]
- https://en.wikipedia.org/wiki/John_Cho [fair]
- https://en.wikipedia.org/wiki/Magic_Johnson [fair]
- ... 9 more

Phase 3 - Exact search tests the candidate bridge

Exact grep call

```
[
  {
    "tool": "grep_corpus",
    "params": {
      "pattern": "Seann William Scott commencement|Eugene Levy commencement|Seann William Scott address|Eugene Levy address|2012 commencement"
    }
  }
]
```

Grep return excerpt

DOCUMENT ID: https://en.wikipedia.org/wiki/Donald_Sutherland (44 tokens) Donald Sutherland – Wikipedia Donald McNichol Sutherland (17 July 1935 – 20 June 2024) was a Canadian actor. With a career spanning six decades, he received numerous accolades, ...

DOCUMENT ID: https://en.wikipedia.org/wiki/John_Lithgow (36 tokens) John Lithgow – Wikipedia John Arthur Lithgow is an American actor. He studied at Harvard University and the London Academy of Music and Dramatic Art before becoming known for his ...

Direct bridge search

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "Eugene Levy 2012 commencement address university biographical speaker 2012"
    }
  }
]
```

Search return before full read

DOCUMENT ID: https://en.wikipedia.org/wiki/Princeton_University (38 tokens) Princeton University – Wikipedia Princeton University is a private Ivy League research university in Princeton, New Jersey, United States. Founded in 1746 in Elizabeth as the College of New ... # DOCUMENT ID: https://en.wikipedia.org/wiki/List_of_University_of_Michigan_alumni (41 tokens) List of University of Michigan alumni – Wikipedia The following is a list of University of Michigan alumni. There are more than 640,000 living alumni of the University of Michigan in 180 countries across ... # DOCUMENT ID: https://en.wikipedia.org/wiki/List_of_Cornell_University_alumni (31 tokens) List of Cornell University alumni – Wikipedia This list of Cornell University alumni includes notable graduates, non-graduate former students, and current students of Cornell University. # DOCUMENT ID: [https://en.wikipedia.org/wiki/University_of_St._Thomas_\(Minnesota\)](https://en.wikipedia.org/wiki/University_of_St._Thomas_(Minnesota)) (37 tokens) University of St. Thomas (Minnesota) – Wikipedia The University of St. Thomas is a private Catholic research university with campuses in Saint Paul and Minneapolis, Minnesota, United States. # DOCUMENT ID: https://en.wikipedia.org/wiki/List_of_Queens_College_people (43 tokens) List of

Queens College people – Wikipedia This is a list of notable alumni and faculty of Queens College, City University of New York. Contents. 1 Business; 2 Education; 3 ...

State after turn 13

pool_size: 74
curated_size: 30/30
importance: very_high=0, high=5, fair=14, low=11
top_curated:
– [https://en.wikipedia.org/wiki/American_Pie_\(film\)](https://en.wikipedia.org/wiki/American_Pie_(film)) [high]
– [https://en.wikipedia.org/wiki/American_Pie_\(film_series\)](https://en.wikipedia.org/wiki/American_Pie_(film_series)) [high]
– https://en.wikipedia.org/wiki/Eugene_Levy [high]
– https://en.wikipedia.org/wiki/List_of_American_Pie_characters [high]
– https://en.wikipedia.org/wiki/Seann_William_Scott [high]
– https://en.wikipedia.org/wiki/Ben_Carson [fair]
– https://en.wikipedia.org/wiki/Bradley_Cooper [fair]
– [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)](https://en.wikipedia.org/wiki/Chris_Owen_(actor)) [fair]
– https://en.wikipedia.org/wiki/Ellis_Paul [fair]
– https://en.wikipedia.org/wiki/Jason_Biggs [fair]
– https://en.wikipedia.org/wiki/John_Cho [fair]
– https://en.wikipedia.org/wiki/Magic_Johnson [fair]
... 18 more

Phase 4 - Full-document reading exposes the answer sentence

Exact read call

```
[  
  {  
    "tool": "read_document",  
    "params": {  
      "doc_id": "https://en.wikipedia.org/wiki/Eugene_Levy"  
    }  
  }  
]
```

Full-document excerpt around answer

[... clipped before ...]

me honourees, alongside fellow SCTV alumni John Candy, Martin Short, and Catherine O'Hara. Levy is one of only a handful of people who have won at least five Canadian Comedy Awards, including two for Best Writing (Best in Show in 2001 and A Mighty Wind in 2004) and three for Best Male Performer (Best in Show, American Pie 2 in 2002, and A Mighty Wind).

In 2008, the governor general of Canada presented Levy with the Governor General's Performing Arts Awards (GGPAA), a lifetime achievement award considered "for their outstanding body of work and enduring contribution to the performing arts in Canada.". In 2010, Levy was awarded the ACTRA Award by the union representing Canada's actors. In 2011, Levy was made a Member of the Order of Canada "for his contributions as a comic actor and writer, and for his dedication to charitable causes" and promoted to the rank of Companion in 2022.

On May 22, 2012, Levy delivered a commencement address at Dalhousie University, in Halifax, Nova Scotia, and was awarded the degree Doctor of Laws (honoris causa). On June 11, 2012, Levy was presented with the Queen Elizabeth II Diamond Jubilee Medal by the lieutenant governor of Ontario. On March 13, 2016, Levy took home the award for "Best Performance by an Actor in a Continuing Leading Comedic Role" at the 4th Annual Canadian Screen Awards, for his performance as Johnny Rose in the CBC/Pop TV sitcom Schitt's Creek.

References
External links

Eugene Levy at IMDb

Eugene Levy at the Internet Broadway Database
Eugene Levy discography at Discogs
Article at thecanadianencyclopedia.ca
Eugene Levy on X
[Context: 9839/30720]

Final curation update

Curated set updated (30/30): [https://en.wikipedia.org/wiki/Eugene_Levy\[very_high\]](https://en.wikipedia.org/wiki/Eugene_Levy[very_high]), [https://en.wikipedia.org/wiki/American_Pie_\(film\)\[high\]](https://en.wikipedia.org/wiki/American_Pie_(film)[high]), [https://en.wikipedia.org/wiki/American_Pie_\(film_series\)\[high\]](https://en.wikipedia.org/wiki/American_Pie_(film_series)[high]), [https://en.wikipedia.org/wiki/Dalhousie_University\[high\]](https://en.wikipedia.org/wiki/Dalhousie_University[high]), [https://en.wikipedia.org/wiki/List_of_American_Pie_characters\[high\]](https://en.wikipedia.org/wiki/List_of_American_Pie_characters[high]), [https://en.wikipedia.org/wiki/Seann_William_Scott\[high\]](https://en.wikipedia.org/wiki/Seann_William_Scott[high]), [https://en.wikipedia.org/wiki/Ben_Carson\[fair\]](https://en.wikipedia.org/wiki/Ben_Carson[fair]), [https://en.wikipedia.org/wiki/Bradley_Cooper\[fair\]](https://en.wikipedia.org/wiki/Bradley_Cooper[fair]), [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)\[fair\]](https://en.wikipedia.org/wiki/Chris_Owen_(actor)[fair]), [https://en.wikipedia.org/wiki/Ellis_Paul\[fair\]](https://en.wikipedia.org/wiki/Ellis_Paul[fair]), [https://en.wikipedia.org/wiki/Jason_Biggs\[fair\]](https://en.wikipedia.org/wiki/Jason_Biggs[fair]), [https://en.wikipedia.org/wiki/John_Cho\[fair\]](https://en.wikipedia.org/wiki/John_Cho[fair]), [https://en.wikipedia.org/wiki/Magic_Johnson\[fair\]](https://en.wikipedia.org/wiki/Magic_Johnson[fair]), [https://en.wikipedia.org/wiki/Princeton_University\[fair\]](https://en.wikipedia.org/wiki/Princeton_University[fair]), [https://en.wikipedia.org/wiki/Shannon_Elizabeth\[fair\]](https://en.wikipedia.org/wiki/Shannon_Elizabeth[fair]), [https://en.wikipedia.org/wiki/Steven_Curtis_Chapman\[fair\]](https://en.wikipedia.org/wiki/Steven_Curtis_Chapman[fair]), [https://en.wikipedia.org/wiki/Susan_Rice\[fair\]](https://en.wikipedia.org/wiki/Susan_Rice[fair]), [https://en.wikipedia.org/wiki/Thomas_Ian_Nicholas\[fair\]](https://en.wikipedia.org/wiki/Thomas_Ian_Nicholas[fair]), [https://en.wikipedia.org/wiki/Tim_Reid\[fair\]](https://en.wikipedia.org/wiki/Tim_Reid[fair]), [https://en.wikipedia.org/wiki/Tucker_Carlson\[fair\]](https://en.wikipedia.org/wiki/Tucker_Carlson[fair]), [https://en.wikipedia.org/wiki/Bradley_Whitford\[low\]](https://en.wikipedia.org/wiki/Bradley_Whitford[low]), [https://en.wikipedia.org/wiki/Dan_Petronijevic\[low\]](https://en.wikipedia.org/wiki/Dan_Petronijevic[low]), [https://en.wikipedia.org/wiki/David_Sutherland\[low\]](https://en.wikipedia.org/wiki/David_Sutherland[low]), [https://en.wikipedia.org/wiki/Donald_Sutherland\[low\]](https://en.wikipedia.org/wiki/Donald_Sutherland[low]), [https://en.wikipedia.org/wiki/Heath_Ledger\[low\]](https://en.wikipedia.org/wiki/Heath_Ledger[low]), [https://en.wikipedia.org/wiki/John_Lithgow\[low\]](https://en.wikipedia.org/wiki/John_Lithgow[low]), [https://en.wikipedia.org/wiki/Jonathan_Franzen\[low\]](https://en.wikipedia.org/wiki/Jonathan_Franzen[low]), [https://en.wikipedia.org/wiki/Joss_Whedon\[low\]](https://en.wikipedia.org/wiki/Joss_Whedon[low]), [https://en.wikipedia.org/wiki/List_of_NYU_Tisch_School_of_the_Arts_people\[low\]](https://en.wikipedia.org/wiki/List_of_NYU_Tisch_School_of_the_Arts_people[low]), [https://en.wikipedia.org/wiki/Wes_Craven\[low\]](https://en.wikipedia.org/wiki/Wes_Craven[low])
[EJECTED low–importance] 1 doc(s): [https://en.wikipedia.org/wiki/Eugene_Levy_\(politician\)](https://en.wikipedia.org/wiki/Eugene_Levy_(politician))

State after turn 15

pool_size: 74
curated_size: 30/30
importance: very_high=1, high=5, fair=14, low=10
top_curated:

- https://en.wikipedia.org/wiki/Eugene_Levy [very_high]
- [https://en.wikipedia.org/wiki/American_Pie_\(film\)](https://en.wikipedia.org/wiki/American_Pie_(film)) [high]
- [https://en.wikipedia.org/wiki/American_Pie_\(film_series\)](https://en.wikipedia.org/wiki/American_Pie_(film_series)) [high]
- https://en.wikipedia.org/wiki/Dalhousie_University [high]
- https://en.wikipedia.org/wiki/List_of_American_Pie_characters [high]
- https://en.wikipedia.org/wiki/Seann_William_Scott [high]
- https://en.wikipedia.org/wiki/Ben_Carson [fair]
- https://en.wikipedia.org/wiki/Bradley_Cooper [fair]
- [https://en.wikipedia.org/wiki/Chris_Owen_\(actor\)](https://en.wikipedia.org/wiki/Chris_Owen_(actor)) [fair]
- https://en.wikipedia.org/wiki/Ellis_Paul [fair]
- https://en.wikipedia.org/wiki/Jason_Biggs [fair]
- https://en.wikipedia.org/wiki/John_Cho [fair]
- ... 18 more

Q.4 Case Study D: Held-Out Multi-Hop QA via Search-State Editing

HotpotQA query

Query. Which American comedian starred in the original cast of Saturday Night Live and played Stan Winters in the sitcom Martin?
Gold answer. Garrett Morris
Outcome. 11 turns, tool diversity 5, curated recall 1.0, final-answer recall 1.0.

Why this case matters. This held-out HotpotQA example is intentionally small: it makes the transfer claim visible at trajectory level. The policy applies the same operations learned elsewhere, namely search, curation, full-document reading, exact grep, and final evidence promotion, to solve a multi-hop web query outside the SFT and RL families.

Trajectory overview.

Turn	Tool	Policy move	Harness state / why it matters
0	search	Search both clues jointly	initial hit identifies Garrett Morris through the Stan Winters clue
1	curate	Promote candidate pages	auto-seeded web hits become an editable curated set
2	search	Search the SNL-original-cast clue	second hop confirms the other half of the query
3	curate	Promote SNL cast evidence	curated state now contains both supporting paths
4	read	Read Garrett Morris page	full text unifies comedian, original SNL cast, and Martin role
5	curate	Keep Garrett Morris high	final answer evidence is retained
6	grep	Grep exact role phrase	exact matching checks the Stan Winters clue
7	curate	Add sitcom/role pages	curated set adds corroborating context

Phase 1 - Joint search identifies the likely entity

Exact search call

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "Stan Winters Martin SNL original cast comedian"
    }
  }
]
```

Search return: Stan Winters points to Garrett Morris

DOCUMENT ID: https://en.wikipedia.org/wiki/Garrett_Morris (46 tokens) Garrett Morris – Wikipedia He is also known for his role in the sitcom Martin as Stan Winters, from 1992 to 1995, until he suffered an injury. Also, he made two guest appearances on The ...

Curation update

Curated set updated (8/30): https://en.wikipedia.org/wiki/Garrett_Morris[high], https://en.wikipedia.org/wiki/Andrea_Martin[fair], https://en.wikipedia.org/wiki/Bobby_Moynihan[fair], https://en.wikipedia.org/wiki/Don_Novello[fair], [https://en.wikipedia.org/wiki/Richard_Lewis_\(comedian\)](https://en.wikipedia.org/wiki/Richard_Lewis_(comedian))[fair], <https://en.wikipedia.org/wiki/>

/Second_City_Television[[fair](https://en.wikipedia.org/wiki/Second_City_Television)], https://en.wikipedia.org/wiki/Victoria_Jackson[[fair](https://en.wikipedia.org/wiki/Victoria_Jackson)], https://en.wikipedia.org/wiki/Will_Forte[[fair](https://en.wikipedia.org/wiki/Will_Forte)]

State after turn 1

pool_size: 10
curated_size: 8/30
importance: very_high=0, high=1, fair=7, low=0
top_curated:

- https://en.wikipedia.org/wiki/Garrett_Morris [high]
- https://en.wikipedia.org/wiki/Andrea_Martin [fair]
- https://en.wikipedia.org/wiki/Bobby_Moynihan [fair]
- https://en.wikipedia.org/wiki/Don_Novello [fair]
- [https://en.wikipedia.org/wiki/Richard_Lewis_\(comedian\)](https://en.wikipedia.org/wiki/Richard_Lewis_(comedian)) [fair]
- https://en.wikipedia.org/wiki/Second_City_Television [fair]
- https://en.wikipedia.org/wiki/Victoria_Jackson [fair]
- https://en.wikipedia.org/wiki/Will_Forte [fair]

Phase 2 - Second-hop search checks original SNL cast membership

Exact search call

```
[  
  {  
    "tool": "search_corpus",  
    "params": {  
      "query": "Garrett Morris Saturday Night Live original cast"  
    }  
  }  
]
```

Search return: original cast evidence

DOCUMENT ID: https://en.wikipedia.org/wiki/Saturday_Night_Live_15th_Anniversary_Special (46 tokens)
Saturday Night Live 15th Anniversary Special – Wikipedia
The special was SNL's first anniversary celebration; it featured original cast members Chevy Chase, Dan Aykroyd, Jane Curtin, Garrett Morris, and Laraine ...

DOCUMENT ID: https://en.wikipedia.org/wiki/List_of_Saturday_Night_Live_cast_members (54 tokens)
List of Saturday Night Live cast members – Wikipedia
Original cast (from left to right): Laraine Newman, John Belushi, Jane Curtin, Gilda Radner, Dan Aykroyd, Garrett Morris, and Chevy Chase. Contents. 1 List.

DOCUMENT ID: https://en.wikipedia.org/wiki/Saturday_Night_Live_season_1 (56 tokens)
Saturday Night Live season 1 – Wikipedia
Starring. Dan Aykroyd * John Belushi * Chevy Chase * George Coe * Jane Curtin * Garrett Morris * Laraine Newman * Michael O'Donoghue * Gilda Radner ; No. of ...

DOCUMENT ID: https://en.wikipedia.org/wiki/Saturday_Night_Live_season_3 (40 tokens)
Saturday Night Live season 3 – Wikipedia
Cast * Dan Aykroyd * John Belus
[... clipped after ...]

Curation update

Curated set updated (12/30): https://en.wikipedia.org/wiki/Garrett_Morris[[high](https://en.wikipedia.org/wiki/Garrett_Morris)], https://en.wikipedia.org/wiki/List_of_Saturday_Night_Live_cast_members[[high](https://en.wikipedia.org/wiki/List_of_Saturday_Night_Live_cast_members)], <https://en.wikipedia.org/wiki/>

Saturday_Night_Live_15th_Anniversary_Special[high], https://en.wikipedia.org/wiki/Andrea_Martin[fair], https://en.wikipedia.org/wiki/Bobby_Moynihan[fair], https://en.wikipedia.org/wiki/Don_Novello[fair], [https://en.wikipedia.org/wiki/Richard_Lewis_\(comedian\)](https://en.wikipedia.org/wiki/Richard_Lewis_(comedian))[fair], https://en.wikipedia.org/wiki/Saturday_Night_Live_season_1[fair], https://en.wikipedia.org/wiki/Saturday_Night_Live_season_3[fair], https://en.wikipedia.org/wiki/Second_City_Television[fair], https://en.wikipedia.org/wiki/Victoria_Jackson[fair], https://en.wikipedia.org/wiki/Will_Forte[fair]

State after turn 3

pool.size: 19
curated.size: 12/30
importance: very_high=0, high=3, fair=9, low=0
top_curated:

- https://en.wikipedia.org/wiki/Garrett_Morris [high]
- https://en.wikipedia.org/wiki/List_of_Saturday_Night_Live_cast_members [high]
- https://en.wikipedia.org/wiki/Saturday_Night_Live_15th_Anniversary_Special [high]
- https://en.wikipedia.org/wiki/Andrea_Martin [fair]
- https://en.wikipedia.org/wiki/Bobby_Moynihan [fair]
- https://en.wikipedia.org/wiki/Don_Novello [fair]
- [https://en.wikipedia.org/wiki/Richard_Lewis_\(comedian\)](https://en.wikipedia.org/wiki/Richard_Lewis_(comedian)) [fair]
- https://en.wikipedia.org/wiki/Saturday_Night_Live_season_1 [fair]
- https://en.wikipedia.org/wiki/Saturday_Night_Live_season_3 [fair]
- https://en.wikipedia.org/wiki/Second_City_Television [fair]
- https://en.wikipedia.org/wiki/Victoria_Jackson [fair]
- https://en.wikipedia.org/wiki/Will_Forte [fair]

Phase 3 - Full-document reading unifies both clues

Exact read call

```
[
  {
    "tool": "read_document",
    "params": {
      "doc_id": "https://en.wikipedia.org/wiki/Garrett_Morris"
    }
  }
]
```

Full-document excerpt around answer

Document (2665 tokens)
Garrett Morris

Summary

Garrett Isaac Morris (born February 1, 1937) is an American actor, comedian and singer. He was part of the original cast and was the first Black cast member of the sketch comedy program Saturday Night Live, appearing from 1975 to 1980.

He also played Jimmy on The Jeffersons (1983–1984). Morris had one of the starring roles, as Junior "Uncle Junior" King, on the sitcom The Jamie Foxx Show, which aired from 1996 to 2001. Morris also had a regular role as Earl Washington on the CBS sitcom 2 Broke Girls, from 2011 to 2017.

He is also known for his role in the sitcom Martin as Stan Winters, from 1992 to 1995, until he suffered an injury. Also, he made two guest appearances on The Wayans Bros. in season one, episode one as himself and again on episode ten as the brothers' uncle Leon (1995). He played a concerned teacher in the film Cooley High (1975), Slide in Car Wash (1976), and Carl in The Census Taker (1984).

In 2024, Morris was honored with his own star on the Hollywood Walk of Fame.

Full Text

Garrett Isaac Morris (born February 1, 1937)
[... clipped after ...]

Exact grep call

```
[
  {
    "tool": "grep.corpus",
    "params": {
      "pattern": "Stan Winters"
    }
  }
]
```

Grep return

DOCUMENT ID: https://en.wikipedia.org/wiki/Garrett_Morris (45 tokens) Garrett Morris – Wikipedia He is also known for his role in the sitcom Martin as Stan Winters, from 1992 to 1995, until he suffered an injury. Also, he made two guest appearances on ...

DOCUMENT ID: [https://en.wikipedia.org/wiki/Martin_\(TV_series\)](https://en.wikipedia.org/wiki/Martin_(TV_series)) (55 tokens) Martin (TV series) – Wikipedia Garrett Morris as Stan Winters (seasons 1–2; guest season 3). Supporting characters. edit * Reginald Ballard as Bruh–Man (Brother Man) (seasons 2–4); Tracy ...

State after turn 7

pool_size: 23
curated_size: 14/30
importance: very_high=0, high=3, fair=10, low=1
top_curated:
– https://en.wikipedia.org/wiki/Garrett_Morris [high]
– https://en.wikipedia.org/wiki/List_of_Saturday_Night_Live_cast_members [high]
– https://en.wikipedia.org/wiki/Saturday_Night_Live_15th_Anniversary_Special [high]
– https://en.wikipedia.org/wiki/Andrea_Martin [fair]
– https://en.wikipedia.org/wiki/Bobby_Moynihan [fair]
– https://en.wikipedia.org/wiki/Don_Novello [fair]
– [https://en.wikipedia.org/wiki/Martin_\(TV_series\)](https://en.wikipedia.org/wiki/Martin_(TV_series)) [fair]
– [https://en.wikipedia.org/wiki/Richard_Lewis_\(comedian\)](https://en.wikipedia.org/wiki/Richard_Lewis_(comedian)) [fair]
– https://en.wikipedia.org/wiki/Saturday_Night_Live_season_1 [fair]
– https://en.wikipedia.org/wiki/Saturday_Night_Live_season_3 [fair]
– https://en.wikipedia.org/wiki/Second_City_Television [fair]
– https://en.wikipedia.org/wiki/Victoria_Jackson [fair]
... 2 more

Q.5 Case Study E: Backtracking after Rejecting the Right Entity

HotpotQA query

Query. When was the game, in which Drew Gehling voices Gord, released for PlayStation 2?
Gold answer. October 17, 2006
Outcome. 40 turns, tool diversity 5, curated recall 1.0, final-answer recall 1.0.

Why this case is interesting. The policy first rejects the correct entity, carries that mistake through several turns, and then reverses it after reading a stronger source. The first search surfaces *Bully*, but the policy calls that hit likely wrong and pursues a separate “Gord” interpretation. Several turns later, `read_document` on Drew Gehling’s page proves that the discarded game was correct. The recovery is not two adjacent search turns: it is a delayed state correction, followed by removal of stale foreign-language variants and a new search for the PlayStation 2 release date.

Trajectory overview.

Turn	Tool	Policy move	Harness state / why it matters
0	<code>search</code>	Search Drew Gehling, Gord, and PS2 together	auto-seed includes <i>Bully</i> pages, including the English game page
2	<code>search</code>	Mistakenly reject <i>Bully</i>	the policy says the correct game is “likely wrong” and starts a separate Gord-character path
3	<code>curate</code>	Keep Drew Gehling, demote noisy pages	the mistaken hypothesis remains visible because <i>Bully</i> is still retained at low importance
4–5	<code>search</code>	Pursue the wrong Gord interpretation	Duck Tales and generic Gord-game searches do not resolve the query
7	<code>read</code>	Read Drew Gehling’s page	full text contradicts the earlier rejection: Gehling is the voice of Gord in <i>Bully</i>
8	<code>curate</code>	Backtrack by editing memory	foreign-language variants are removed while the English <i>Bully</i> page is retained
9	<code>search</code>	Search the corrected title plus PS2 release date	the search now targets the right game instead of the wrong Gord branch
10	<code>curate</code>	Promote release-date evidence	the simple <i>Bully</i> page and <i>Canis Canem Edit</i> page become high-importance evidence
11	<code>verify</code>	Check the release-date claim	verification is attempted over the curated candidate docs
13–14	<code>grep</code>	Exact-date grep is tried and rejected	date-only matching returns unrelated October 17 pages, so the agent does not treat the date string alone as evidence

Phase 1 - The correct game appears first, but the policy rejects it

Initial reasoning excerpt

The query asks for the PlayStation 2 release date of the game in which Drew Gehling voices Gord. The first decomposition is sensible: search the actor, the role name, and the platform together.

Exact tool call

```
[
  {
    "tool": "search_corpus",
    "params": {
      "query": "Drew Gehling Gord PS2 game"
    }
  }
]
```

```
}  
}  
]
```

Search return: the right entity is present immediately

DOCUMENT ID: [https://en.wikipedia.org/wiki/Bully_\(video_game\)](https://en.wikipedia.org/wiki/Bully_(video_game))
Bully (video game) – Wikipedia
Bully is a 2006 action–adventure video game developed by Rockstar Vancouver and published by Rockstar Games ...

DOCUMENT ID: [https://es.wikipedia.org/wiki/Bully_\(videojuego\)](https://es.wikipedia.org/wiki/Bully_(videojuego))
Bully es un videojuego ... desarrollado por Rockstar Vancouver para la consola PlayStation 2.

[AUTO-POPULATED] Top 3 docs from this search have been added to your curated set at 'fair' importance.

The policy wrongly rules out the right hit

The search for Drew Gehling Gord PS2 returned Bully, which is likely wrong given that Bully is not voiced by Gehling. My current search angle is exhausted. Let me rethink. A missed facet: "Gord" is probably the character from the "Duck Tales" game series voiced by Gehling.

Memory after the mistaken rejection

Curated set updated (4/30):
https://en.wikipedia.org/wiki/Drew_Gehling[high],
[https://en.wikipedia.org/wiki/Bully_\(video_game\)](https://en.wikipedia.org/wiki/Bully_(video_game))[low],
[https://es.wikipedia.org/wiki/Bully_\(videojuego\)](https://es.wikipedia.org/wiki/Bully_(videojuego))[low],
[https://pt.wikipedia.org/wiki/Bully_\(jogo_eletr%C3%B4nico\)](https://pt.wikipedia.org/wiki/Bully_(jogo_eletr%C3%B4nico))[low]

Phase 2 - A delayed read overturns the wrong branch

The wrong branch persists for several turns

Turn 4 search:
Duck Tales Gord voiced by Drew Gehling
-> no results

Turn 5 search:
Gord PlayStation 2 game
-> generic "Gord" pages, planned PS2 titles, Ford/Gordon distractors

Turn 7 reasoning:
Current search angle on Gord PlayStation 2 is stuck and likely wrong ... Another angle: check the Drew Gehling page for credited games and character list for "Gord."

Backtracking read

```
[  
  {  
    "tool": "read_document",  
    "params": {  
      "doc.id": "https://en.wikipedia.org/wiki/Drew_Gehling"  
    }  
  }  
]
```

```
}  
]
```

Drew Gehling
Andrew Shearer Gehling ... is an American stage and screen actor ... and as the voice of Gord in the 2006 video game Bully from Rockstar Games.

State edit: remove stale variants, keep the corrected entity

```
[  
  {  
    "tool": "curate",  
    "params": {  
      "add_ids": [  
        "https://en.wikipedia.org/wiki/Drew_Gehling"  
      ],  
      "remove_ids": [  
        "https://es.wikipedia.org/wiki/Bully_(videojuego)",  
        "https://pt.wikipedia.org/wiki/Bully_(jogo_eletr%C3%B4nico)"  
      ],  
      "importance": {  
        "https://en.wikipedia.org/wiki/Drew_Gehling": "high"  
      }  
    }  
  }  
]
```

Curated set updated (3/30):
https://en.wikipedia.org/wiki/Drew_Gehling[high],
[https://en.wikipedia.org/wiki/Bully_\(video_game\)](https://en.wikipedia.org/wiki/Bully_(video_game))[low],
[https://en.wikipedia.org/wiki/Gord_\(video_game\)](https://en.wikipedia.org/wiki/Gord_(video_game))[low]

Phase 3 - The corrected entity drives the answer search

Recovered strategy

The identity facet is solved: Gord is from `\emph{Bully}`. The policy now stops searching generic Gord pages and searches the corrected title with the platform/date facet.

Corrected search

```
[  
  {  
    "tool": "search_corpus",  
    "params": {  
      "query": "Bully PlayStation 2 release date"  
    }  
  }  
]
```

DOCUMENT ID: [https://simple.wikipedia.org/wiki/Bully_\(video_game\)](https://simple.wikipedia.org/wiki/Bully_(video_game))
Bully ... was released on 17 October 2006 ...

DOCUMENT ID: https://en.wikipedia.org/wiki/List_of_video_games_published_by_Rockstar_Games
Bully – PlayStation 2, October 17, 2006.

DOCUMENT ID: https://en.wikipedia.org/wiki/List_of_best-selling_PlayStation_2_video_games
Bully, 1.5 million, October 17, 2006, Action–adventure ...

Promoting the answer evidence

```
[
  {
    "tool": "curate",
    "params": {
      "add_ids": [
        "https://en.wikipedia.org/wiki/Simple_Wikipedia/Bully_(video_game)",
        "https://en.wikipedia.org/wiki/Canis_Canem_Edit"
      ],
      "importance": {
        "https://en.wikipedia.org/wiki/Simple_Wikipedia/Bully_(video_game)": "high",
        "https://en.wikipedia.org/wiki/Canis_Canem_Edit": "high"
      }
    }
  }
]
```

Curated set updated (5/30):

```
https://en.wikipedia.org/wiki/Canis_Canem_Edit[high],
https://en.wikipedia.org/wiki/Drew_Gehling[high],
https://en.wikipedia.org/wiki/Simple_Wikipedia/Bully_(video_game)[high],
https://en.wikipedia.org/wiki/Bully_(video_game)[low],
https://en.wikipedia.org/wiki/Gord_(video_game)[low]
```

A final exact-date check is not enough by itself

Grep on "October 17, 2006|2006-10-17" returns date-only false positives such as current-events pages, films released on the same day, and unrelated titles. The agent therefore keeps the title-and-platform evidence as the answer path rather than treating the date string alone as proof.

Why the state edit matters. The important behavior is the delayed correction. The policy's early hypothesis says *Bully* is probably wrong; the harness nevertheless preserves it at low importance. When a later read contradicts the hypothesis, the policy can edit the curated set, keep the corrected entity, discard stale variants, and redirect search toward the release-date facet. Backtracking is therefore visible as a change to the external state, not just a private thought between two adjacent searches.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction state the paper’s contribution as a stateful retrieval harness plus SFT/RL recipe, and the results section reports the empirical scope across eight benchmarks. The limitations paragraph qualifies the scope of the claims.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Appendix B discusses limitations including the task scope, benchmark style, regex-based evidence graph, LLM-based verifier, sentence-BM25 compression, benchmark size, and annotation coverage.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical and does not present theoretical results, theorems, or proofs. The formal notation is used to define the agent interface and harness state, not to make a theorem-level claim.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The method, harness algorithms, reward definition, hyperparameters, benchmark statistics, prompts, and evaluation recipe are specified in the main text and appendix. These details cover the components that affect the paper’s main experimental claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper states a plan to release weights, harness code, data-generation pipeline, and RL recipe, but the submission itself does not yet provide open public access to all assets. The appendix nevertheless includes reproducibility details for the experimental protocol.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experimental setup summarizes the benchmark families, source and transfer split interpretation, baselines, and metrics. Appendix tables provide benchmark statistics, training and evaluation hyperparameters, reward weights, prompts, and the evaluation recipe.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper reports aggregate point estimates on fixed benchmark test sets and gives the number of evaluated instances for each benchmark in Table 7. We do not report formal confidence intervals or significance tests for all main comparisons. This is a limitation of the current submission: repeated-run evaluation would require many additional multi-turn agent rollouts, including Tinker inference calls and external retrieval/reranking calls across several systems and benchmarks, which is computationally and financially expensive. We discuss the resulting uncertainty, especially for smaller benchmarks, in the limitations section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper reports the training platform, model, SFT and RL data scale, rollout budget, number of training steps, inference settings, retrieval backend, reranker, and evaluation protocol. Training was conducted using Tinker, a managed training service; therefore, exact low-level worker configurations such as physical worker type and memory allocation are abstracted by the platform. We report the available compute-relevant settings needed to reproduce the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: To the best of our knowledge, the work conforms to the NeurIPS Code of Ethics. It uses benchmark and public-domain or cited corpus sources, does not involve deception or human-subject experiments, and preserves submission anonymity.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader-impact and release paragraph discusses the positive impact of lower-cost evidence-seeking systems and the negative risk of amplifying unsupported or privacy-sensitive retrieval. It also describes the intended release scope.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper states that the planned release is a retrieval subagent, not a general autonomous web agent or answer generator, and that release documentation will include intended use, dataset provenance, and evaluation protocols. These are the relevant safeguards for the assets introduced here.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original sources for all existing assets used in the paper, including benchmark datasets, model checkpoints, retrieval backends, and external services.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper documents the new harness, training recipe, prompts, reward design, and evaluation protocol in the main text and appendix. The planned release will include accompanying documentation for the weights, code, and generated training data.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Evaluation is performed with benchmark data and automated model judgments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subject research. Therefore IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are central to the method and evaluation: the paper describes the teacher used for SFT, the trained retrieval policy, frontier baselines, answer generators, and the judge prompt. The relevant model names and prompts are provided in the main text and appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.