Information Retrieval Misses the Mark for LLM Agents

Yiqun Sun Pengfei Wei Lawrence B. Hsieh
Magellan Technology Research Institute (MTRI)
{duke.sun,pengfei.wei,lawrence.hsieh}@mtri.co.jp

Abstract

Current widely deployed LLM agents often retrieve information using simple rule-based tools, such as `grep`, rather than learned encoders, vector indices, or rerankers; industry commentary has captured this shift with the slogan "RAG is dead, agents just need grep." Although we reject the claim that retrieval is obsolete, **we argue that existing information retrieval (IR) research cannot yet comprehensively satisfy the information needs of LLM agents.** The retrieval substrate of the agent era is real and deployed, yet it is poorly captured by standard IR formulations and benchmarks. This misalignment is structural across five dimensions: today's IR stack assumes the wrong corpus, the wrong input, the wrong objective, the wrong episode, and the wrong retrievable for agents that plan, browse, call tools, manage memory, and decide whether to continue searching. We argue that the productive next step is to reformulate retrieval for agents as a *state-conditioned evidence-acquisition policy*. We organize existing agentic RAG and reasoning-search systems under a taxonomy and a formal model, run a direct test of the misalignment by holding the agent constant and varying only the retrieval substrate (BM25, vector, grep, hybrid, closed-book) on HotpotQA-distractor and 2WikiMultihopQA over a 100K Wikipedia corpus, and propose a five-direction research agenda with dual-track evaluation.

1 Introduction

Information retrieval has been a load-bearing layer of the modern LLM stack, primarily through retrieval-augmented generation (RAG), which mitigates parametric hallucination by grounding generation in externally retrieved evidence [4, 20, 26, 31, 38]. The dominant pipeline of dense retrievers, learned encoders, vector indices, and rerankers, optimised on TREC- and BEIR-style relevance judgements, is now standard in nearly every deployed LLM application. By any reasonable measure, retrieval is necessary for trustworthy, fresh, and verifiable LLM behaviour.

Yet a striking pattern has emerged in 2025–2026: the most-used *agentic* LLM systems in production largely abandon this pipeline. Two concrete deployments make the point. An audit of Anthropic's Claude Code, among the most-used agentic coding tools in deployment, finds that its entire retrieval substrate consists of lexical (`ripgrep`) search, filesystem traversal, file reads, web search and fetch, language-server queries, and delegation to forked subagents; nowhere in the stack is there a vector index, a learned encoder, a reranker, or a relevance label.[1] Vercel's open-source knowledge-agent template reports a roughly 75% drop in per-call cost together with a quality improvement after removing a comparable vector pipeline in favour of filesystem-and-shell access.[2]

[1] Source-code audit of `https://github.com/yasasbanukaofficial/claude-code`.
[2] Vercel Labs, *Knowledge Agent Template*, `https://github.com/vercel-labs/knowledge-agent-template`.

These choices have triggered a vigorous public debate. One camp argues, under the slogan "RAG is dead, agents just need grep",[3] that learned retrieval is being superseded by agentic exploration over long context.[4] The other camp counters that "agents with grep" is still retrieval-augmented generation in a different form and that hybrid lexical–vector retrieval remains the productive frontier.[5] Both camps tacitly agree that something has changed about what retrieval means when its consumer is an LLM agent rather than a human or a one-shot generator.

That something, we argue, is not that retrieval is dying but that retrieval is changing shape. The classical IR abstraction, in which a human user issues a query, the system returns a ranked list, and quality is judged by ordering metrics over relevance labels, shaped a generation of benchmarks, retrievers, and evaluation suites. RAG extended it only slightly by replacing the human reader with a one-shot generator while leaving the single-call query–document matching step intact. The agent regime is genuinely different. Recent agentic-RAG and reasoning-search systems [1, 27, 29, 49, 74, 76, 79, 109, 112] condition retrieval on the agent's evolving state, change *whether* and *when* to retrieve, and increasingly train the search trajectory itself with reinforcement learning. Production stacks ship Deep Research,[6] agentic web search,[7] and file search[8] as first-class tools, and the underlying substrate of grep, browser, API, memory, sub-agent delegation, and long context bears almost no resemblance to a BEIR-style ranked list. Mainstream IR research has only partially absorbed this shift. The diagnosis is not that the agentic retrieval thread is absent, but that dominant *evaluation incentives, benchmarks, and reward signals* still target a regime the most successful deployed agents do not compute: the bulk of 2024–2025 work at SIGIR, WSDM, and EMNLP reports on MS MARCO [2], BEIR [90], or MTEB [48], optimises nDCG@10 or Recall@100 of a fixed dense retriever, and treats the user as a static, query-issuing reader.

Position. This position paper argues that information retrieval misses the mark for LLM agents. Agent retrieval is now the dominant deployed regime and the most under-served research regime, and the IR community is the natural home for the theory and measurement it lacks. The productive next move is to reformulate retrieval for agents as a *state-conditioned evidence-acquisition policy*, that is, a partially observable decision process whose state includes goal, plan, evidence, memory, uncertainty, and budget, and whose objective combines evidence sufficiency, provenance, and budget consumption rather than ranking quality alone (§6). On this view, "good retrieval" may mean *not* retrieving, retrieving from memory, delegating to a subagent, or extracting a single structured field rather than ranking passages.

Why the position is forced. The claim is structural rather than stylistic. Five concrete misalignments separate the dominant IR stack from how LLM agents actually acquire and use evidence: the static corpus, the query-as-input interface, the ranking-quality objective, the single-episode evaluation, and the passage-only retrievable each fail in the agent setting (§4). Together they imply that improving nDCG@10 on BEIR no longer faithfully predicts agent task success.

Contributions. The paper offers four deliverables. The first is a taxonomy (Fig. 1) of agent-centered retrieval organised around four control loci, namely fixed-corpus, policy-controlled, environment-interaction, and internal-state retrieval. The second is a formal model under which existing agentic-RAG and reasoning-search methods appear as instantiations of a state-conditioned retrieval policy. The third is a direct test of the misalignment (§5): we hold a single agent constant (GPT-5.4-mini in the OpenAI Agents SDK) over a 100K-passage Wikipedia corpus and vary only the retrieval substrate it can call (vector, BM25, grep, hybrid, closed-book), running 200-question evaluations on HotpotQA-distractor and 2WikiMultihopQA with Claude Sonnet 4.6 as judge; the headline finding is that BM25, grep, and a strong dense retriever cluster within roughly four accuracy points once the agent controls the loop, and that grep, the Claude Code substrate, is competitive with the dense

[3] N. Bustamante, "The RAG Obituary: Killed by Agents, Buried by Context Windows", Oct. 2025, https://web.archive.org/web/20260323100059/https://www.nicolasbustamante.com/p/the-rag-obituary-killed-by-agents.

[4] S. Desai, "RAG is Dead, Long Live Agentic Retrieval", LlamaIndex Blog, May 2025, https://llamaindex.ai/blog/rag-is-dead-long-live-agentic-retrieval.

[5] A. Webb, "No, RAG Is Not Dead", Algolia Blog, Jan. 2026, https://www.algolia.com/blog/ai/rag-is-not-dead.

[6] OpenAI, *Introducing Deep Research*, https://openai.com/index/introducing-deep-research/.

[7] OpenAI, *Web Search and Agentic Search Tooling*, https://developers.openai.com/api/docs/guides/tools-web-search.

[8] OpenAI, *File Search Tooling*, https://developers.openai.com/api/docs/guides/tools-file-search.

retriever on both datasets. The fourth is a five-direction research agenda anchored by dual-track evaluation.

2 From Query-Document IR to Agent Information State

Three regimes. Contemporary retrieval is best read as the third regime in a sequence, each defined by who consumes the evidence.

Classical IR (Regime 1). A human seeker queries a document corpus and is scored by ranking-quality metrics. Foundational frameworks model evolving uncertainty and bounded cognition [3, 36, 45, 67], while the technical stack (BM25 [63, 64], SMART [66], dense retrievers [14, 30, 32, 82, 84, 85, 89, 94]) optimises ranking on TREC-style relevance.

One-shot RAG (Regime 2). A generator replaces the human reader. RAG [38], REALM [20], FiD [25], Atlas [26], RETRO [4], kNN-LM [31], and in-context RALM [61] retrieve once per question, and REPLUG [73] and RECOMP [105] re-score or compress that retrieved evidence. Query-side (HyDE [16], Query2doc [95], RQ-RAG [5], rewrite-retrieve-read [42]) and reranking methods (RankGPT [81], RankT5 [121], RankZephyr [56], RankLLaMA [43], ZipRerank [88]) sharpen a single episode that remains query-driven and confined to a fixed corpus.

Agentic retrieval (Regime 3). An LLM agent plans, acts, observes, remembers, and decides whether to keep searching, and retrieval becomes a policy-controlled trajectory [13, 75]. ReAct [112], Toolformer [68], Self-Ask [57], and IRCoT [92] make retrieval a typed action interleaved with reasoning. FLARE [28], DRAGIN [79], Iter-RetGen [69], Search-o1 [40], Self-RAG [1], CRAG [109], Adaptive-RAG [27], Auto-RAG [114], and ThinkNote [108] condition retrieval on uncertainty, plan position, or evidence state. WebGPT [49], Mind2Web [9, 17], WebArena [34, 120], WebVoyager [21], WebWalker [100], AssistantBench [113], and GAIA [46] place retrieval inside browser environments. A complementary line treats memory itself as the retrievable substrate, including MemGPT [52], A-MEM [107], MemoryBank [118], HippoRAG [19], GraphRAG [12], and LightRAG [18]. Multi-agent systems such as AutoGen [101], Magentic-One [15], MAIN-RAG [6], MindSearch [8], and agentic-RAG [75, 87] delegate retrieval among specialists.

Two shifts across regimes. The *control locus* moves from the user (issuing a query) to the agent (deciding whether, when, where, and how to retrieve, and when to stop), and the *retrievable substrate* widens from passages over a fixed corpus to web pages, sub-pages, tools, APIs, file chunks, tables, memory notes, evidence chains, and sibling-agent outputs. Together these shifts make the classical query-document model insufficient even when each individual retriever is excellent. In production, a Claude Code session issues dozens of `Grep`, file-read, web-fetch, language-server, and sub-agent calls per user request with no learned ranker in the loop, Deep Research traces routinely span hundreds of pages and structured-API calls, and agent-RL systems [29, 74, 76, 80] treat each retrieval as one step in a policy trained against end-task accuracy. Query-document evaluation cannot score any of these trajectories.

3 A Taxonomy of Agent-Centered Retrieval

Figure 1 organizes the literature into four branches distinguished by *where retrieval is controlled from* and *what the retrievable substrate is*. The branches are not exclusive, and the most capable systems compose them.

Fixed-corpus retrieval. The retrievable is a passage in a closed, pre-indexed corpus, and control is exercised on the query side. Retrieve-then-read systems [4, 20, 25, 26, 30–32, 38, 50, 61, 94, 104] provide the substrate. Query rewriting and expansion [5, 16, 42, 95, 106, 117] sharpen the query, while reranking and compression [43, 51, 55, 56, 60, 62, 70, 73, 81, 105, 121] sharpen the result list. This branch covers the bulk of current IR research and most of MTEB [48] and BEIR [90].

Policy-controlled retrieval. Control moves inside the LLM. The agent decides, step by step, whether to retrieve, how to phrase the query, where to query, and when to stop. Tool-mediated agents [54, 59, 68, 71, 110, 112] expose retrieval as a typed action. Decomposition methods [33, 57, 92, 119] split a hard query into sub-queries, iterative methods [28, 40, 69, 79, 97] interleave reasoning and retrieval, and self-reflective and adaptive methods [1, 27, 108, 109, 114] make retrieval conditional

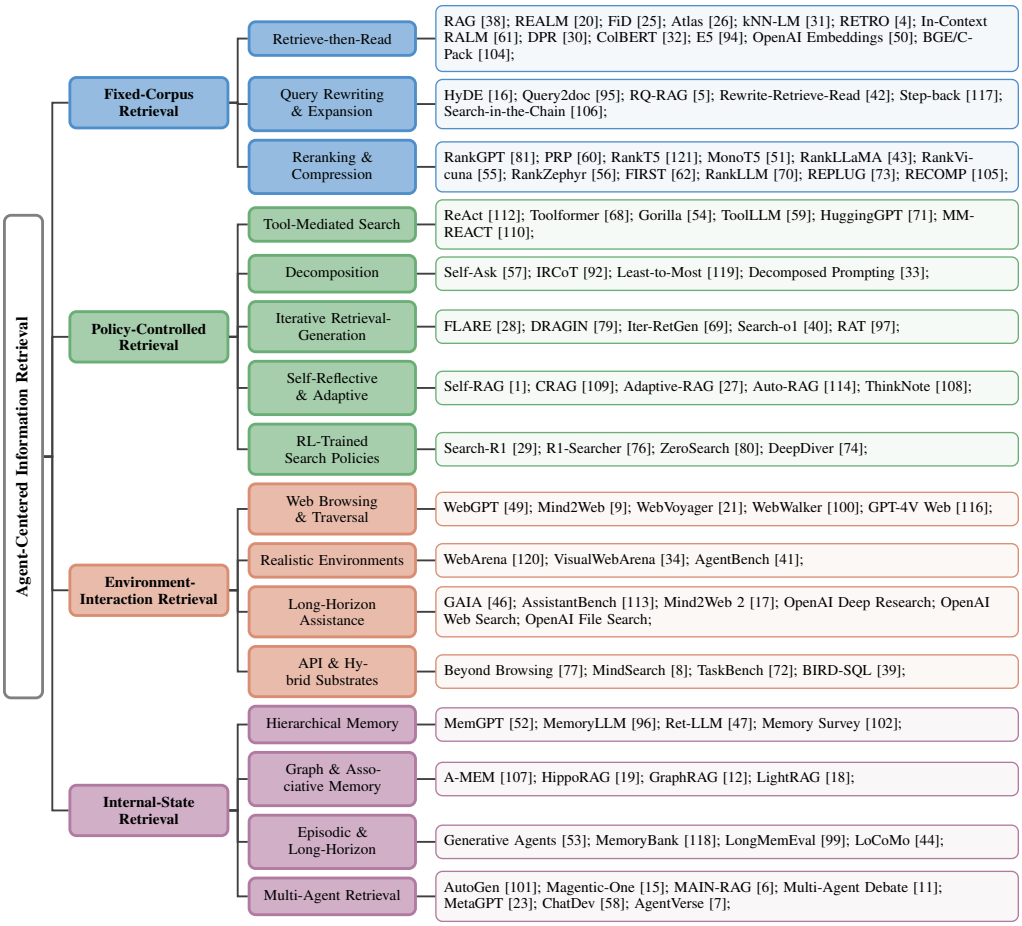

Figure 1: A taxonomy of agent-centered information retrieval. The four top branches differ in *where retrieval is controlled from* and *what the retrievable substrate is*: fixed corpora with query-side control; policy-controlled retrieval where the LLM agent decides when, what, and how to search; environment-interaction retrieval over live web, browsers, and APIs; and internal-state retrieval over memory and other agents. The most capable systems compose several branches.

on self-assessed uncertainty or query complexity. RL-trained search policies [29, 74, 76, 80] learn the policy directly from rewards in real or simulated search environments.

Environment-interaction retrieval. The retrievable is a live or interactive environment such as a web browser, an OS, a database, or an API. WebGPT [49], Mind2Web [9], WebVoyager [21], WebWalker [100], and grounded multimodal agents [116] navigate pages and click affordances. Realistic environments [34, 41, 120] provide reproducible substrates. Long-horizon assistance benchmarks [17, 46, 113], together with production-side deep-research and search/file-search tools, evaluate sustained evidence acquisition over hours, hundreds of pages, and time-varying answers. API-and-hybrid systems [8, 39, 72, 77] explicitly trade browsing for direct structured-substrate access.

Internal-state retrieval. The retrievable is the agent's own memory, history, or the outputs of co-operating agents. Hierarchical memory [47, 52, 96, 102] treats context-window management as paged retrieval. Graph and associative memory [12, 18, 19, 107] represents memory as a graph and retrieves over notes and links. Episodic memory and long-horizon evaluation [44, 53, 99, 118] retrieve from accumulated dialogue and life-event traces. Multi-agent retrieval [6, 7, 11, 15, 23, 58, 101] treats other agents' outputs as a retrievable substrate, where "which agent should retrieve next" is itself a policy decision.

Reading the taxonomy. A modern deep-research system, such as the one underlying OpenAI's Deep Research product, is best read as a composition. It uses fixed-corpus retrievers (semantic and

keyword over indexed pages and uploaded files), under a policy-controlled loop (decompose, decide, iterate, stop), embedded in environment-interaction with the live web (sub-pages, structured APIs), with internal-state retrieval (notes, prior search results, sub-agent outputs) closing the loop. No single sub-tree alone captures it, and this is precisely the gap the rest of the paper argues current IR research must close.

4 Why Current IR is Misaligned with Agent Needs

We now make the negative claim concrete. The dominant IR research setup — a fixed corpus, a query as input, a single ranked list, a ranking-quality metric, and an episode that ends after one retrieval call — bakes in five assumptions, each of which fails for LLM agents.

Static-corpus assumption. BEIR [90], MTEB [48], MS MARCO [2], NQ [37], and the bulk of MIRACL [115] all assume a frozen corpus with frozen relevance judgements. Yet agentic deployments increasingly retrieve from *live* substrates: the open web (WebVoyager [21], WebWalker [100], OpenAI Deep Research), private file stores, dynamic memory [99, 107], and other agents' rolling outputs [6, 15]. Live-web evaluation shows large gaps that frozen corpora cannot diagnose. InfoDeepSeek [103] explicitly argues that fixed gold-document sets misrepresent real seeking, and FreshQA [93] measures the same drift over time-varying facts. Optimizing on a frozen corpus is at best a proxy and at worst rewards systems that overfit to BEIR-style surface relevance while failing on freshness, recency, and source authority.

Query-as-input assumption. BEIR-style retrievers consume only the query string. Agentic retrievers should also consume the agent's *state*, including which sub-goal is active, which entities are unresolved, which claims are under-supported, where uncertainty is high, what has already been retrieved, and what budget remains. Self-RAG [1], CRAG [109], FLARE [28], DRAGIN [79], Search-o1 [40], and Adaptive-RAG [27] all show empirically that conditioning retrieval on this larger state changes *whether* and *what* to retrieve, not only the ranking. Yet the IR literature still trains retrievers and rerankers on $\langle q, d \rangle$ pairs only, which is a strict subset of the agent's information need.

Ranking-quality-only objective. nDCG@10 and Recall@100 measure how well a ranked list matches a fixed relevance judgement, but agents consume *evidence to act*. Ranking quality is a real signal, especially once budgets are not tight (Fig. 4a shows it tracks agent accuracy strongly at $k \geq 2$), yet alone it does not capture evidence sufficiency, redundancy, provenance and citation accuracy [10, 35], calibration (does retrieval reduce uncertainty?), or cost (tokens, calls, latency). Recent benchmarks have begun to measure these dimensions explicitly. FRAMES [35] unifies factuality, retrieval, and reasoning, while DeepResearch Bench [10] scores citation count and accuracy alongside report quality. InfoDeepSeek [103] adds utility and compactness, and eRAG [65] estimates per-document utility via downstream success. Beyond relevance, diversity is increasingly recognized as an important dimension of retrieval-set quality [24, 83, 86, 89]. None reduce to nDCG.

Single-episode assumption. BEIR-style benchmarks measure a single retrieval call. Agentic systems instead make *trajectories* that may include ten searches, three sub-page traversals, two memory recalls, an API call, and a verifier step. The right unit of analysis is a trajectory under a budget, with a learned stopping rule. Adaptive-RAG [27] routes among 0-, 1-, and multi-step strategies, and DRAGIN [79] and Search-o1 [40] trigger retrieval mid-generation. Pangu DeepDiver [74] studies adaptive search intensity directly, while Search-R1 [29], R1-Searcher [76], and ZeroSearch [80] learn the trajectory itself with reinforcement learning. The dominant IR research question of "what is the best retriever?" is therefore subordinate to a different question that current benchmarks do not ask, namely "what is the best retrieval *policy*?".

Passage-only retrievable. A retrievable in the agent setting may be a webpage, a sub-page, a tool affordance, an API response, a file chunk, a structured table cell, a memory note, an evidence chain, a date-filtered subset, or another agent's output. Beyond Browsing [77] demonstrates that API-first retrieval can outperform pure browsing in WebArena, BIRD-SQL [39] foregrounds structured-database retrieval, and tool-rich systems [54, 59, 71] treat tools themselves as retrievable affordances. Multi-agent systems [6, 8, 15] retrieve over agent outputs. These substrates differ in ways that ranking metrics cannot express: an exact regulation citation, a structured stock price, and a one-line code-search result are all "retrievals" that BEIR's encoder-tuned recall@10 cannot evaluate. A well-grounded agent IR research programme must *native*-treat lexical, semantic, structured, browser, API, and memory retrieval as one typed substrate, not as separate ad-hoc systems.

Table 1: Classical-IR vs. agent-IR evaluation. The right column is what current IR research mostly does *not* measure.

Dimension	Classical IR / one-shot RAG	Agent IR
Input	Query string q	Information state $s_t = (g, \pi, h, m, e, u, b)$
Action space	Single SEARCH over fixed corpus	Typed: SEARCH, BROWSE, API, RECALL, VERIFY, DELEGATE, STOP
Retrievable	Passage in pre-indexed corpus	Page, sub-page, API response, file chunk, table cell, memory note, agent output
Episode	One retrieval call	Trajectory under budget with learned stopping rule
Objective	nDCG@k, Recall@k, MRR	Task success + evidence sufficiency + provenance − cost − risk
Reference	Fixed gold relevance labels	Goal-grounded outcome, agent-as-judge, rubric trees [17]
Corpus	Static (BEIR/MTEB/MS MARCO)	Live web, files, memory, agent outputs, time-varying
Provenance	Implicit (relevance only)	Claim-level citation graph with authority and conflict
Calibration	Out of scope	Did retrieval reduce u_t? Did the agent stop on time?

Table 1's right column is partly aspirational: evidence sufficiency, u_t calibration, and replayable provenance graphs are research targets that §7 expands. Even so, the five misalignments together imply that one- or two-point nDCG@10 gains on BEIR no longer reliably predict agent-task success [10, 35, 103], and IR research needs metrics, benchmarks, and retrievers built for the agent regime rather than back-ported from the human regime.

5 Does the Retrieval Substrate Still Matter When an Agent Drives the Loop?

To sharpen the disagreement with classical IR, we run one direct test. The position predicts that, once the agent controls retrieval as a typed action over an evolving information state, the choice of *retrieval substrate* (lexical, dense, regex) matters far less than the existence of the loop itself. Classical IR predicts the opposite: a stronger ranker should drive better task success regardless of the consumer.

Setup. We hold a single agent constant (GPT-5.4-mini in the OpenAI Agents SDK loop) and swap only the retrieval substrate exposed to it as a function tool. The corpus is a 100,000-passage Wikipedia pool drawn from the BEIR HotpotQA corpus, augmented with the gold and distractor passages of the eval questions so all golds are guaranteed findable. Five conditions: `vector` (only `vector_search` over Qwen3-Embedding-4B), `bm25` (only `bm25_search`), `grep` (only `grep_search`, case-insensitive Python regex over raw text, the Claude Code style), `hybrid` (all three available simultaneously), and `closed` (no retrieval, parametric only). The agent decides what to query, when to re-issue, and when to stop with a `final_answer` call (max 10 turns). Each (condition, dataset) is run on 200 questions of HotpotQA-distractor and 2WikiMultihopQA. Claude Sonnet 4.6 grades end-task accuracy. Tool schemas, prompts, and per-trajectory traces are in Appendix D.

The substrate matters less than the loop. Across both datasets the three retrieval substrates cluster within roughly four accuracy points (Fig. 2a). On HotpotQA, BM25 (0.80), grep (0.78), and vector (0.76) are statistically indistinguishable; on 2WikiMultihopQA, grep (0.72) actually outperforms vector (0.70) and BM25 (0.62). The closed-book lower bound (0.44 / 0.38) confirms retrieval is doing real work, but it does not rank the substrates. This is the prediction the position makes: with the agent in control of the loop, the headline rank-quality numbers that classical IR optimises do not translate into agent-task gains in any clean way.

Hybrid does not dominate, and grep is competitive with dense retrieval. Giving the agent all three substrates simultaneously is no better than the best single substrate (0.80 = 0.80 on HotpotQA, 0.66 *below* 0.72 on 2WikiMultihop). The case where this matters most is grep against vector. The Claude Code substrate keeps pace with a strong dense retriever despite having no notion of semantic similarity, because the agent compensates by issuing more focused regex queries (Fig. 2b: 3.4–3.7 grep calls per question vs. 2.6–3.3 vector calls) at a comparable token cost (Fig. 2c). The qualitative gap a classical retrieval benchmark would predict (vector $\gg$ BM25 $\gg$ grep) does not survive the loop.

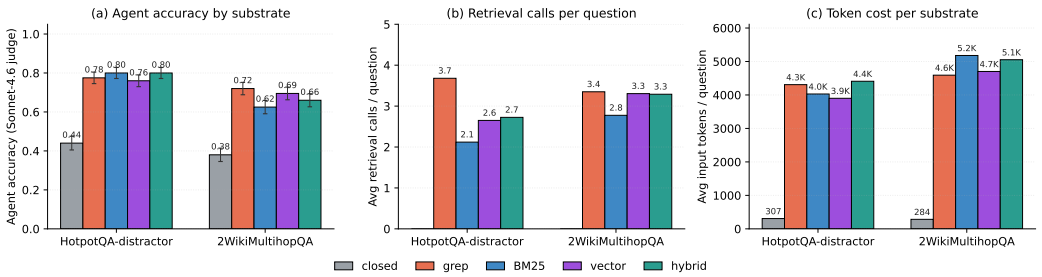

Figure 2: Agent-loop experiment with GPT-5.4-mini and one retrieval substrate held variable, on 200 HotpotQA-distractor and 200 2WikiMultihopQA questions over a 100K-passage Wikipedia corpus. **(a)** The substrate matters less than the existence of the loop: grep, BM25, and vector cluster within ± 4 accuracy points, while the closed-book baseline is far below all three. **(b)** The agent issues 2–4 retrieval calls per question; grep needs slightly more than vector. **(c)** Token cost is comparable across substrates. Setup in App. D.

What this implies for IR research. The headline finding is not that ranking quality is unimportant, but that, with a competent agent in the loop, the *ordering* of retrievers a classical leaderboard predicts does not match the ordering of agent-task success. Three caveats are obvious. Single-hop short-context tasks may still be ranker-dominated. The agent we used (GPT-5.4-mini) is a strong base model; a weaker policy might be more dependent on the retriever. The corpus is 100K Wikipedia passages, not the open web or a multi-million-document private store, where vector encoders may regain margin. None of these flips the qualitative observation: with this kind of agent in the loop, substrate is open and the action-space, stopping, and state-conditioning slots of §6 are the more interesting variables. We supplement this main result with three small studies in the classical (one-shot) RAG setting in Appendix D that probe those slots one at a time: a ranking-quality budget sensitivity sweep across eight retrievers, an adaptive-stopping run at fixed retriever, and a state-conditioning probe. They are not ablations of the agentic loop; they are independent supporting evidence on the standard RAG slice.

6 Retrieval as a State-Conditioned Policy

We now formalize the positive claim. The model below is deliberately minimal, and its purpose is to make the policy view precise and to position existing methods as instantiations.

Information state. At decision step t, the agent's information state is

$$s_t = \big(g,\ \pi_t,\ h_t,\ m_t,\ e_t,\ u_t,\ b_t\big), \tag{1}$$

where g is the user goal, π_t the current plan and active sub-goals, h_t the interaction and observation history, m_t the agent's external memory, e_t the evidence state (a structured collection of supported claims, with provenance), u_t a profile of uncertainties over claims and sub-goals, and b_t the remaining budget across tokens, time, tool calls, and any risk constraints.

Information need. The information need at step t is not the original query but a function of state,

$$\eta_t = f(s_t), \tag{2}$$

which may resolve to an unresolved entity in π_t, a missing support for a claim in e_t, a contradiction between two evidence items, a low-confidence reasoning step in u_t, a stale item in m_t, or a structured value (date, identifier, regulation number) that the agent cannot synthesise from parametric knowledge.

Retrieval action. A retrieval action is a typed tuple

$$a_t = \big(\mathrm{op},\ \mathrm{src},\ q,\ \phi,\ k,\ \tau\big), \tag{3}$$

where $\mathrm{op} \in \{\textsc{search}, \textsc{browse}, \textsc{rerank}, \textsc{verify}, \textsc{recall}, \textsc{api}, \textsc{delegate}, \textsc{stop}\}$ is the operator type, src identifies the source (corpus, tool, web domain, memory store, or sibling agent), q is a query or filter program, ϕ are constraints (metadata filters, recency, language, source-authority bounds), k is the depth or top-k budget, and τ a stopping or confidence threshold. The STOP action is first-class, since "do not retrieve more" is a legitimate retrieval policy choice [1, 27, 74].

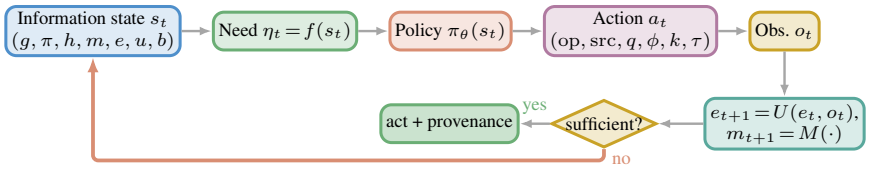

Figure 3: The agent retrieval loop. The state s_t drives a need estimator f, a policy π_θ, and a typed action a_t. Observations update evidence and memory, and a sufficiency check decides whether to act or retrieve again. Classical IR collapses this loop to a single box, the mapping $q \mapsto$ ranked list.

Observations and evidence update. The environment returns observations $o_t = E(a_t)$ (page contents, ranked passages, API payloads, memory notes, agent replies). The agent applies an evidence-update operator

$$e_{t+1} = U\big(e_t, o_t\big), \tag{4}$$

which aggregates support, conflict, novelty, authority, and provenance, and a memory-update operator $m_{t+1} = M(m_t, e_{t+1}, o_t)$ that decides what to store, link, compress, or forget [52, 99, 107].

Retrieval policy and objective. The agent acts according to a policy $a_t \sim \pi_\theta(s_t)$ optimized for a multi-objective return

$$J(\pi_\theta) = \mathbb{E}\Big[\, \alpha\, R_{\text{task}} + \beta\, R_{\text{suff}} + \gamma\, R_{\text{prov}} - \lambda_c\, C_{\text{cost}} - \lambda_r\, C_{\text{risk}} \,\Big], \tag{5}$$

where R_{task} is task success, R_{suff} rewards evidence sufficiency without redundancy, R_{prov} rewards correct, source-authoritative citations, C_{cost} accounts for tokens, calls, and latency, and C_{risk} penalises hallucinated or unsupported actions. This is deliberately compatible with classical IR (set $\beta = \gamma = \lambda_c = \lambda_r = 0$ and freeze the policy after one call to recover nDCG-style optimization), but it makes explicit what classical IR does not measure. Only R_{task} and C_{cost} are routinely measurable today, with FRAMES [35] and DeepResearch Bench [10] taking partial steps on R_{prov}. The remaining components R_{suff}, C_{risk}, and the structural specifications of η_t, U, M are research targets we expand in §7.

Existing methods as instantiations. Many recent systems are special cases of this model. In ReAct [112], π_θ is an unconditional in-context policy with η_t derived only from h_t. Self-RAG [1] gates π_θ by self-critique tokens conditioned on a slice of u_t, while CRAG [109] falls back to web search when an evaluator on e_t flags low quality. Adaptive-RAG [27] routes among 0-, 1-, and multi-step retrieval through a query-complexity classifier, while Search-o1 [40] and DRAGIN [79] trigger retrieval mid-generation when the agent encounters uncertain knowledge points. Pangu DeepDiver [74], Search-R1 [29], R1-Searcher [76], and ZeroSearch [80] learn π_θ end-to-end with reinforcement learning over open or simulated environments. MemGPT [52] and A-MEM [107] make RECALL and the memory-update operator M first-class actions, while Magentic-One [15], AutoGen [101], MAIN-RAG [6], and MindSearch [8] extend DELEGATE and the source space SRC to sibling agents.

A side-by-side instantiation table that lists which slots of $(s_t, a_t, \mathrm{src}, \tau)$ each of fourteen recent systems (ReAct, Self-Ask, IRCoT, FLARE, DRAGIN, Self-RAG, CRAG, Adaptive-RAG, Search-o1, DeepDiver, Search-R1, MemGPT, A-MEM, Magentic-One) actually realises is given in Appendix B. The take-away is that almost every system fills one or two slots well, while *none* jointly addresses all of state, action space, source space, stopping, and provenance. Closing this gap is what §7 calls a research agenda for.

The point is not that this model is novel, since its ingredients appear in pieces in the agentic-RAG [75], memory [102], and tool-use literatures. The point is that current IR research mostly studies retrievers as if s_t were just q, the action space were a single SEARCH, and J were nDCG. That is a strict and now-misleading restriction.

7 Research Agenda for Agent-Centered IR

The position implies five directions, each unpacking one slot of the model in §6 and addressing one row of Table 1.

Condition the policy on the full state. Today's retrievers and rerankers consume only q, while §6 argues for inputting π_t, e_t, u_t, b_t as well. This directly attacks the fixed-query versus evolving-need row of Table 1: Self-RAG [1] and Adaptive-RAG [27] already condition the policy on a slice of state, BRIGHT [78] shows reasoning-intensive retrieval exceeds query-only conditioning, and Fig. 4c shows even a one-shot probe of unmet need helps. The open question is whether the retriever *index* itself, not just the query encoder, should be conditioned on plan, evidence, and uncertainty hotspots.

Expand the action space. Two slots of a_t are under-developed. The operator and source $(\mathrm{op}, \mathrm{src})$ should share one typed action algebra spanning lexical, dense, metadata-filtered, structured (SQL/JSON), tool, browser, and memory retrieval, with Beyond Browsing [77] and BIRD-SQL [39] making the empirical case and the IR community well-placed to supply the typed algebras, cost models, and fall-back semantics this requires. The stopping threshold τ and STOP action are barely studied in IR, yet Adaptive-RAG [27], DRAGIN [79], DeepDiver [74], and our adaptive-stopping pilot (Fig. 4b) all show large gains, calling for stopping models conditioned on u_t, b_t, budget-ladder frontiers, and calibration metrics for early and late stopping.

Model the evidence and memory updates. The operators U (evidence) and M (memory) are first-class in §6 but mostly invisible in current IR. For U, citation count and accuracy are first-class in DeepResearch Bench [10], FRAMES [35], and WebGPT [49]; the next step is structured *support/conflict graphs* over claim-level evidence, source authority, and replayable provenance, with the ranked list as a derivative. For M, once memory is a first-class retrievable [12, 18, 19, 52, 107], retrieval also asks "what to remember, link, compress, revise, or forget?", and LongMemEval [99] and LoCoMo [44] expose how poorly memory updates are measured.

Coordinate retrieval across agents. Once src includes sibling agents [6, 8, 11, 15, 23, 101], "who retrieves next" becomes a policy over src, raising delegation, parallel versus serial sub-retrievers, and provenance graphs that merge contradictory sibling evidence.

Evaluate the full objective. Only R_{task} and C_{cost} in J are routinely measured today, so we propose a *dual-track* default that exercises all six components. Track A (*frozen, reproducible*) pairs sandboxed file-search corpora and fixed memory snapshots with WebArena [34, 120] and AgentBench [41]. Track B (*live, dynamic*) uses WebVoyager/WebWalker [21, 100], AssistantBench/GAIA [46, 113], Mind2Web 2 [17], InfoDeepSeek [103], DeepResearch Bench [10], and freshness-sensitive QA [93, 98], both reported on explicit budget ladders.

8 Alternative Views

Retriever quality and agent policy are complementary. The strongest counter-position is that retriever quality, query rewriting, reranking, and compression remain fundamental layers *inside* agent loops rather than being displaced by them, and we partly endorse the complementarity: the closed-book lower bound in Fig. 2a (0.44 / 0.38) confirms retrieval is doing real work, and at single-hop tight-budget settings ranking quality tracks answer accuracy strongly (Fig. 4a). Where we depart is on scope: classical IR has no first-class notion of when not to retrieve, of memory recall, or of delegation, while agentic-RAG and reasoning-search [1, 27, 29, 74, 79, 109] already define retrieval as a learned action over state. Our position is one of rebalancing, expanding IR's measurement and incentives to the policy, sufficiency, provenance, and stopping layers above the retriever.

The end-to-end-RL-suffices view. A second alternative is that RL-trained search agents absorb IR into agent RL. We disagree: policies that ignore index structure waste calls, opaque policies cannot supply auditable evidence trails for audited deployments, and sparse RL reward needs the inductive bias of typed actions; learned policies are also environment-bound, while agent-IR theory must generalise.

9 Conclusion

Retrieval's dominant consumer has changed: users acquire evidence by issuing typed actions over an evolving state, and once a single agent controls the loop, BM25, grep, and dense retrievers cluster within four accuracy points on multi-hop QA. Reformulating retrieval as a state-conditioned, budget-aware evidence-acquisition policy is the smallest tractable move, and the IR community, which built the encoders and rerankers any agent substrate still needs, should now build the benchmarks that score what agents actually do.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[3] Nicholas J. Belkin. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5(1):133–143, 1980.

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, 2022.

[5] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. RQ-RAG: Learning to refine queries for retrieval augmented generation. In *Conference on Language Modeling*, 2024.

[6] Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. *arXiv preprint arXiv:2501.00332*, 2025.

[7] Weize Chen, Yusheng Su, Jian Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. *arXiv preprint arXiv:2308.10848*, 2023.

[8] Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. MindSearch: Mimicking human minds elicits deep AI searcher. *arXiv preprint arXiv:2407.20183*, 2024.

[9] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, 2023.

[10] Mingxuan Du et al. DeepResearch Bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

[11] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

[12] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[13] Yichao Feng, Haoran Luo, Zhenghong Lin, Yiqun Sun, Pengfei Wei, Lawrence B Hsieh, and Anh Tuan Luu. Orchmas: Orchestrated reasoning with multi collaborative heterogeneous scientific expert structured agents. *arXiv preprint arXiv:2603.03005*, 2026.

[14] Yingchaojie Feng, Yiqun Sun, Yandong Sun, Minfeng Zhu, Qiang Huang, Anthony Kum Hoe Tung, and Wei Chen. Don't reinvent the wheel: Efficient instruction-following text embedding based on guided space transformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24511–24525, 2025.

[15] Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. Magentic-One: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468*, 2024.

[16] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of ACL*, 2023.

[17] Boyu Gou, Boyuan Zheng, et al. Mind2Web 2: Evaluating agentic search with agent-as-a-judge. *arXiv preprint arXiv:2506.21506*, 2025.

[18] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

[19] Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, 2024.

[20] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning*, 2020.

[21] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. WebVoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of ACL*, 2024.

[22] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.

[23] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. *Proceedings of ICLR*, 2024.

[24] Qiang Huang, Yanhao Wang, Yiqun Sun, and Anthony KH Tung. Diversity-aware k-maximum inner product search revisited. *arXiv preprint arXiv:2402.13858*, 2024.

[25] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of EACL*, 2021.

[26] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 2023.

[27] Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of NAACL*, 2024.

[28] 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 EMNLP*, 2023.

[29] 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.

[30] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, 2020.

[31] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

[32] Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of SIGIR*, 2020.

[33] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *International Conference on Learning Representations*, 2023.

[34] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.

[35] 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. *arXiv preprint arXiv:2409.12941*, 2024.

[36] Carol Collier Kuhlthau. *Seeking Meaning: A Process Approach to Library and Information Services*. Libraries Unlimited, 2nd edition, 2004.

[37] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019.

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

[39] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Cao, Ruiying Geng, et al. Can LLM already serve as a database interface? a big bench for large-scale database grounded text-to-SQLs. In *Advances in Neural Information Processing Systems*, 2023.

[40] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

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

[42] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of EMNLP*, 2023.

[43] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning LLaMA for multi-stage text retrieval. In *Proceedings of SIGIR*, 2024.

[44] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of ACL*, 2024.

[45] Gary Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1995.

[46] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024.

[47] Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-LLM: Towards a general read-write memory for large language models. *arXiv preprint arXiv:2305.14322*, 2023.

[48] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of EACL*, 2023.

[49] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[50] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*, 2022.

[51] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of EMNLP*, 2020.

[52] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

[53] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*, 2023.

[54] Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *Advances in Neural Information Processing Systems*, 2024.

[55] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankVicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*, 2023.

[56] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*, 2023.

[57] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of EMNLP*, 2023.

[58] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. ChatDev: Communicative agents for software development. *Proceedings of ACL*, 2024.

[59] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

[60] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of NAACL*, 2024.

[61] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 2023.

[62] Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avi Sil, and Heng Ji. FIRST: Faster improved listwise reranking with single token decoding. *arXiv preprint arXiv:2406.15657*, 2024.

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

[64] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 1995.

[65] Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. *Proceedings of SIGIR*, 2024.

[66] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

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

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

[69] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of EMNLP*, 2023.

[70] Sahel Sharifymoghaddam, Ronak Pradeep, Andre Slavescu, Ryan Nguyen, Andrew Xu, Zijian Chen, Yilin Zhang, Yidi Chen, Jasper Xian, and Jimmy Lin. RankLLM: A Python package for reranking with LLMs. *arXiv preprint arXiv:2505.19284*, 2025.

[71] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. *arXiv preprint arXiv:2303.17580*, 2023.

[72] Yongliang Shen, Kaitao Song, Xu Tan, Wenqi Zhang, Kan Ren, Siyu Yuan, Weiming Lu, Dongsheng Li, and Yueting Zhuang. TaskBench: Benchmarking large language models for task automation. *arXiv preprint arXiv:2311.18760*, 2023.

[73] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of NAACL*, 2024.

[74] Wenxuan Shi et al. Pangu DeepDiver: Adaptive search intensity scaling via open-web reinforcement learning. *arXiv preprint arXiv:2505.24332*, 2025.

[75] Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V. Vasilakos. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv preprint arXiv:2501.09136*, 2025.

[76] 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.

[77] Yueqi Song, Frank Xu, Shuyan Zhou, and Graham Neubig. Beyond browsing: API-based web agents. *arXiv preprint arXiv:2410.16464*, 2024.

[78] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*, 2024.

[79] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of ACL*, 2024.

[80] Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Wang. ZeroSearch: Incentivize the search capability of LLMs without searching. *arXiv preprint arXiv:2505.04588*, 2025.

[81] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of EMNLP*, 2023.

[82] Yandong Sun, Qiang Huang, Ziwei Xu, Yiqun Sun, Yixuan Tang, and Anthony KH Tung. One swallow does not make a summer: Understanding semantic structures in embedding spaces. *arXiv preprint arXiv:2512.00852*, 2025.

[83] Yiqun Sun, Qiang Huang, Yanhao Wang, and Anthony KH Tung. Diversinews: Enriching news consumption with relevant yet diverse news articles retrieval. *Proceedings of the VLDB Endowment*, 17(12):4277–4280, 2024.

[84] Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony Tung, and Jun Yu. A general framework for producing interpretable semantic text embeddings. In *International Conference on Learning Representations*, volume 2025, pages 100777–100800, 2025.

[85] Yiqun Sun, Qiang Huang, Anthony KH Tung, and Jun Yu. Text embeddings should capture implicit semantics, not just surface meaning. *arXiv preprint arXiv:2506.08354*, 2025.

[86] Yiqun Sun, Qiang Huang, Anthony Kum Hoe Tung, and Jun Yu. Prism: A framework for producing interpretable political bias embeddings with political-aware cross-encoder. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27719–27733, 2025.

[87] Yiqun Sun, Pengfei Wei, and Lawrence B Hsieh. Don't retrieve, navigate: Distilling enterprise knowledge into navigable agent skills for qa and rag. *arXiv preprint arXiv:2604.14572*, 2026.

[88] Yiqun Sun, Pengfei Wei, and Lawrence B Hsieh. Very efficient listwise multimodal reranking for long documents. *arXiv preprint arXiv:2605.11864*, 2026.

[89] Yixuan Tang, Yuanyuan Shi, Yiqun Sun, and Anthony Kum Hoe Tung. Uncovering the bigger picture: Comprehensive event understanding via diverse news retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33927–33945, 2025.

[90] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021.

[91] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

[92] 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 ACL*, 2023.

[93] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of ACL*, 2024.

[94] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[95] Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In *Proceedings of EMNLP*, 2023.

[96] Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang Li, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian Li, Bing Yin, Jingbo Shang, and Julian McAuley. MEMORYLLM: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024.

[97] Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. RAT: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*, 2024.

[98] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

[99] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025.

[100] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking LLMs in web traversal. *arXiv preprint arXiv:2501.07572*, 2025.

[101] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[102] Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. From human memory to AI memory: A survey on memory mechanisms in the era of LLMs. *arXiv preprint arXiv:2504.15965*, 2025.

[103] Yunjia Xi, Jianghao Lin, Menghui Zhu, Yongzhao Xiao, Zhuoying Ou, Jiaqi Liu, Tong Wan, Bo Chen, Weiwen Liu, Yasheng Wang, Ruiming Tang, Weinan Zhang, and Yong Yu. InfoDeepSeek: Benchmarking agentic information seeking for retrieval-augmented generation. *arXiv preprint arXiv:2505.15872*, 2025.

[104] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. *Proceedings of SIGIR*, 2024.

[105] Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *International Conference on Learning Representations*, 2024.

[106] Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *Proceedings of the Web Conference*, 2024.

[107] Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*, 2025.

[108] Zhipeng Xu, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Chaojun Xiao, Zhiyuan Liu, Ge Yu, and Chenyan Xiong. ThinkNote: Enhancing knowledge integration and utilization of large language models via constructivist cognition modeling. *arXiv preprint arXiv:2402.13547*, 2026.

[109] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

[110] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

[111] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*, 2018.

[112] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[113] Ori Yoran, Samuel J. Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. AssistantBench: Can web agents solve realistic and time-consuming tasks? In *Proceedings of EMNLP*, 2024.

[114] Tian Yu, Shaolei Zhang, and Yang Feng. Auto-RAG: Autonomous retrieval-augmented generation for large language models. *arXiv preprint arXiv:2411.19443*, 2024.

[115] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 2023.

[116] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*, 2024.

[117] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed Chi, Quoc V. Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations*, 2024.

[118] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. *Proceedings of AAAI*, 2024.

[119] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023.

[120] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024.

[121] Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. RankT5: Fine-tuning T5 for text ranking with ranking losses. In *Proceedings of SIGIR*, 2023.

A Glossary of agent-IR terms

Information state. The agent's current beliefs about the goal, plan, evidence, memory, uncertainty, and remaining budget at decision time t.

Evidence sufficiency. The degree to which currently held evidence supports the next action or final answer; orthogonal to ranking quality.

Stopping policy. A learned or rule-based decision to halt further retrieval, distinct from any single retrieval call's quality.

Provenance graph. A claim-level graph linking statements in the agent's output to evidence items, supporting/conflicting links, and source-authority annotations.

Hybrid retrieval substrate. A retrieval interface that natively exposes lexical, dense, metadata-filtered, structured (SQL/JSON), tool/API, browser, and memory operations through a single typed action space.

Budget ladder. A protocol that evaluates an agent at multiple fixed budgets (token, tool-call, latency, cost) and reports a frontier rather than a single number.

B Recent agent-IR systems as instantiations of the formal model

§6 of the main text claims that recent agentic-RAG, reasoning-search, memory, and multi-agent systems can be read as instantiations of the same state-conditioned retrieval policy, each filling a different subset of the slots $(s_t, a_t, \mathrm{src}, \tau)$ rather than offering competing models. Table 2 unpacks that claim system by system. For each method we record which slice of the information state s_t its policy actually conditions on (history, sub-question stack, CoT trace, uncertainty profile, critique tokens, evidence quality, query complexity, memory tier, etc.), the operator vocabulary the agent emits as actions, the source space src the agent can query, the stopping signal τ that terminates retrieval, and the training regime (in-context, few-shot, supervised fine-tuning, or reinforcement learning) used to fit the policy.

Read down the columns, three patterns emerge. First, the *state* column is dominated by single-feature conditioning: most systems condition on one of h_t, π_t, u_t, or a critique signal, and almost none condition on the joint $(g, \pi_t, h_t, m_t, e_t, u_t, b_t)$ that §6 argues the policy needs. Second, the *action ops* column is narrow; the typed action algebra $\{\textsc{search}, \textsc{browse}, \textsc{rerank}, \textsc{verify}, \textsc{recall}, \textsc{api}, \textsc{delegate}, \textsc{stop}\}$ of §6 is realised in pieces,

with browser stacks supplying BROWSE, memory stacks supplying RECALL, multi-agent stacks supplying DELEGATE, and almost nobody jointly. Third, the *stopping* column ranges from "none" (ReAct) to "RL-learned" (Search-R1, DeepDiver), but the in-between region of calibrated, budget-aware stopping is sparsely populated. The bottom row lists the full agent-IR target as a deliberately vacant slot: a system that conditions on all of s_t, exposes the full typed action algebra, sources from a hybrid substrate, and stops via a learned, calibrated τ trained against a multi-objective return J. §7 treats closing each of these gaps as a research direction rather than a single benchmark task.

System	State s_t uses	Action ops	Source src	Stopping	Training
ReAct [112]	history h_t	search	fixed corpus	none	in-context
Self-Ask [57]	sub-questions	search	fixed corpus	decomposed	in-context
IRCoT [92]	CoT trace	search	fixed corpus	step limit	few-shot
FLARE [28]	generation	search	fixed corpus	confidence	in-context
DRAGIN [79]	uncertainty u_t	search	fixed corpus	runtime trigger	in-context
Self-RAG [1]	critique tokens	search, abstain	fixed corpus	reflect tokens	SFT
CRAG [109]	quality eval. on e_t	search, web fallback	corpus + web	evaluator gate	SFT
Adaptive-RAG [27]	query complexity	0/1/multi-step	fixed corpus	router	SFT classifier
Search-o1 [40]	uncertainty in CoT	search, reason	fixed corpus	uncertain point	in-context
DeepDiver [74]	reward signal	search, browse	open web	RL-learned	RL
Search-R1 [29]	token rewards	search	corpus	RL-learned	RL
MemGPT [52]	memory tier	recall, swap	memory	paging policy	in-context
A-MEM [107]	memory graph	recall, link, forget	memory graph	graph-based	SFT
Magentic-One [15]	sub-agent state	delegate, search	web + agents	orchestrator	in-context
Agent-IR (full)	all of s_t	all typed ops	hybrid substrate	learned τ	multi-obj.

Table 2: Recent agent-IR systems as instantiations of the state-conditioned retrieval policy of §6. Almost every system fills one or two slots fully; none jointly addresses state, action space, source space, stopping, and provenance.

C Worked example: BEIR vs. agent retrieval episode

Consider the question *"What does the latest version of the EU AI Act say about general-purpose AI model providers' obligations?"*.

BEIR-style evaluation. The query is fixed, gold passages are pre-annotated, and a retriever is scored on whether those gold passages appear high in a single ranked list. Recall@10 and nDCG@10 are reported. Recency is ignored, so if the corpus was frozen in 2023 the system is rewarded for retrieving outdated text that matches the query lexically.

Agent retrieval episode. The agent first decomposes the query ("find the latest version", "identify the GPAI section", "extract obligations"), issues a freshness-filtered web search, browses the official register, downloads the consolidated text, runs a structured-extraction tool over the relevant article, cross-checks against an alternative source, stores the citation in working memory, and stops. The success criterion is *factually correct, citable, recency-valid output*, not whether any pre-specified passage was ranked high. Most steps are invisible to BEIR.

What is measured. A faithful evaluation must score per-step retrieval utility, redundancy across steps, source authority, recency, evidence sufficiency before stopping, claim-level citation accuracy, and total budget consumed. None of these are first-class in MS MARCO or BEIR; several appear in InfoDeepSeek, DeepResearch Bench, FRAMES, and Mind2Web 2.

D Experimental setup and reproducibility details

This appendix specifies the corpora, agent setup, prompts, retrievers, and per-cell metrics behind the headline experiment of §5, plus three additional supporting studies in the classical (non-agentic) one-shot RAG setting that are referenced from §5, §4, §7, and §8. The aim is to make every reported number reproducible from public artefacts. The experiment harness and per-question result files will be released on acceptance.

D.1 Headline experiment: agent loop over a 100K-passage Wikipedia corpus

Corpus. A pool of 100,000 Wikipedia paragraphs is sampled uniformly at random from the BEIR HotpotQA corpus (a snapshot of the Wikipedia 2017 abstract collection that the original HotpotQA dataset was built from). Before sampling, all 3,489 unique gold and distractor passages from the 200-question evaluation sets of HotpotQA-distractor and 2WikiMultihopQA are added to the pool, so every gold answer is guaranteed findable by some substrate. Passages are stored as `(title, text)` tuples; the indexing string is `"<title>. <text>"`. Random seed is 17.

Indexes. BM25 is built with `rank_bm25.BM25Okapi` over lowercased word-character tokens. The dense index is Qwen3-Embedding-4B (`Qwen/Qwen3-Embedding-4B`) in fp16, with the document prompt name `"document"` and the query prompt name `"query"`, L2-normalised so cosine similarity equals dot product. Embeddings are stored as fp16 `numpy` arrays ($100{,}000 \times 2560$).

Tools. The agent is given the following function tools (Python signatures and docstrings, exposed to the agent through the OpenAI Agents SDK `@function_tool` decorator):

```
vector_search(query: str, k: int = 5) -> str
  "Dense semantic retrieval over the Wikipedia corpus using
   Qwen3-Embedding-4B. Returns top-k passages by cosine similarity."

bm25_search(query: str, k: int = 5) -> str
  "Lexical BM25 retrieval over the same corpus. Best for keyword
   or entity queries."

grep_search(pattern: str, max_matches: int = 10) -> str
  "Case-insensitive Python regex over raw passage text (Claude Code
   style). Returns up to max_matches passages whose text matches."

final_answer(answer: str) -> str
  "Submit the final short answer. Terminates the agent loop."
```

The condition determines which subset is exposed: `vector`, `bm25`, `grep`, all-three (`hybrid`), or none (`closed`). Tool outputs are formatted as numbered passage blocks with provenance `[id]` markers.

Agent. GPT-5.4-mini through OpenRouter, wrapped in an `OpenAIChatCompletionsModel` and run with the OpenAI Agents SDK `Runner.run`. Temperature 0, `max_tokens` 400, `max_turns` 10. The system prompt instructs the agent to plan briefly, issue focused retrieval calls (small k), refine queries based on what is found, stop as soon as evidence is sufficient, and submit a short final answer through `final_answer`.

Judge. Claude Sonnet 4.6 grades each `(question, gold, predicted)` triple as `CORRECT` or `INCORRECT` with the same lenient-formatting prompt used in the supporting studies below. The judge sees only the triple, never the agent's trajectory or the retrieved passages, so substrate identity does not leak.

Per-condition statistics. Reported numbers (mean $\pm$ standard error of a binomial proportion over 200 questions) are summarised in Table 3.

Table 3: Headline agent-loop results. Acc is the Sonnet-4.6 judged accuracy; *calls* is the average number of retrieval tool calls per question; *in tok* is the average number of input tokens per question (a rough cost proxy).

	HotpotQA-distractor			2WikiMultihopQA		
Condition	Acc	calls	in tok	Acc	calls	in tok
`closed`	0.440	0.00	307	0.380	0.00	284
`grep`	0.775	3.68	4,308	0.720	3.35	4,591
`bm25`	0.800	2.12	4,029	0.625	2.78	5,180
`vector`	0.760	2.65	3,901	0.695	3.31	4,702
`hybrid`	0.800	2.73	4,409	0.660	3.29	5,053

D.2 Additional supporting studies in the classical (non-agentic) RAG setting

The three studies in §D.3–§D.5 are deliberately *not* agentic. They use the standard one-shot RAG protocol (a retriever returns top-k passages, a single answerer reads them, the answerer emits an answer), with no tool calls, no agent loop, no re-issued queries, and no `final_answer` action. We include them only because each one isolates one of the misalignment dimensions of §4 and one of the agenda directions of §7 on a controlled, classical-RAG slice that is independent of, and complementary to, the agentic headline experiment in §D.1: §D.3 probes the ranking-quality-as-objective claim, §D.4 probes adaptive stopping, and §D.5 probes state-conditioned retrieval. They should not be read as ablations of the agentic loop.

All three studies run on validation splits of three multi-hop QA benchmarks with a fixed Python `random` seed of 7. Each question carries 10 candidate passages, of which 2–3 are gold (the dataset-supplied supporting facts), and retrieval is over this per-question candidate pool of 10 rather than the global Wikipedia index (a tractability simplification that lets us sweep eight retrievers and four budgets in the first study).

HotpotQA-distractor [111]: 300 questions from `hotpot_qa/distractor` validation, with the 10 candidate paragraphs and `supporting_facts` field used as-is. **2WikiMultihopQA** [22]: 200 questions from `voidful/2WikiMultihopQA` validation, with supporting facts mapped to passage indices by title match. **MuSiQue-Ans** [91]: 200 `answerable` questions from `dgslibisey/MuSiQue` validation; we keep all paragraphs marked `is_supporting` and uniformly sample distractors to obtain 10 candidates per question, preserving the supporting count. Average gold counts per question are $2.05, 2.37, 2.65$ on the three datasets.

Retrievers (eight). **BM25** via `rank_bm25.BM25Okapi`; **MiniLM-L6** (`sentence-transformers/all-MiniLM-L6-v2`); **MPNet-base**; **BGE-base-v1.5** and **BGE-large-v1.5** (with the standard "Represent this sentence for searching relevant passages: " query prompt); and **Qwen3-Emb-{0.6B, 4B, 8B}**. All dense encoders run in fp16 with L2-normalised embeddings.

Answerer prompt.

```
You are a careful research assistant. Read the passages and answer
the question concisely. If the passages contain the answer, state it.
If not, say you cannot answer from the passages. Output JSON only:
{"answer": <string>, "supported": true|false}.
```

Judge prompt.

```
You are an exact-match grader. Given a gold answer and a predicted
answer, decide if the prediction is semantically equivalent to the
gold answer (yes/no answers, names, numbers, dates count exactly;
minor paraphrasing of factual answers is acceptable). Output JSON
only: {"correct": true|false}.
```

D.3 Supporting study 1: Recall@k vs. one-shot RAG accuracy across budgets

For each (retriever, $k \in \{1, 2, 3, 5\}$, dataset) cell, the top-k passages are passed to GPT-5.4-mini as the answerer in a single call; GPT-5.4-mini also serves as primary judge. The full HotpotQA $8 \times 4 = 32$-cell grid (9,600 (question, GPT answer) pairs) is independently re-judged by Claude Sonnet 4.6 with the same prompt, temperature 0, `max_tokens` 20. Per-question agreement with the GPT judge ranges from 96.0% to 99.3% across the 32 cells (mean 97.5%). The Spearman pattern between Sonnet-judged accuracy and Recall@k is $+0.01,\ +0.86,\ +0.88,\ +0.93$ at $k \in \{1, 2, 3, 5\}$, very close to the GPT-judged pattern $+0.02,\ +0.85,\ +0.90,\ +0.86$. Per-cell numbers are in Table 4–6.

D.4 Supporting study 2: adaptive stopping at fixed retriever

Holding the retriever fixed (Qwen3-Embedding-4B) and using the first 200 HotpotQA questions, the adaptive policy is:

```
for k = 1, 2, ..., k_max:           # k_max = 5
    passages <- retriever.topk(question, k)
    answer, supported <- agent(question, passages)  # JSON-flagged
    if supported:
```

```
        break
return (answer, k_stopped = k, n_calls = k)
```

The `supported` flag is the same self-reported field the answerer emits in Supporting study 1, so the policy is zero-shot in the sense that no auxiliary classifier is trained. As reference points we also report *Oracle* (gold-only passages, 0.90 accuracy) and $k = 10$ (BGE-large all-passage dump, 0.88 accuracy) on the full 300-question HotpotQA set.

D.5 Supporting study 3: state-conditioned retrieval variants

Three two-passage variants on the first 200 HotpotQA questions, all using Qwen3-Embedding-4B as the underlying encoder. **B1 (query-only)**: top-2 passages by query embedding. **B2 (query + evidence)**: the first passage is the top-1 by query embedding; the second pass re-embeds `[query + first-passage title and text]` and returns the top-1 of the remaining 9. **B3 (query + evidence + LLM probe)**: same first passage, but GPT-5.4-mini emits a one-sentence probe of what is still missing, and the second pass re-embeds `[query + ‘‘next-need:’’ + probe + ‘‘already:’’ + first-passage title and text]` and returns the top-1 of the remaining 9. The probe system prompt is

```
You read a question and one supporting passage and decide what
information is still needed to answer. Output a single short
search query (one sentence, no preamble) describing what to look
for next.
```

Once the two passages are selected, the answerer and judge are identical to Supporting study 1. B2 and B3 differ from B1 only in the second-hop retrieval step.

D.6 Figure for the supporting studies

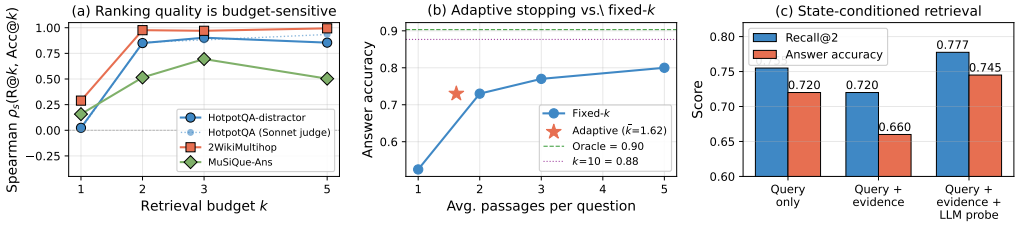

Figure 4: Supporting studies in the classical (one-shot) RAG setting, referenced from §5. **(a)** Spearman ρ_s between Recall@k and one-shot RAG accuracy is near zero at $k = 1$ and strong at $k \geq 2$ on all three datasets and under both judges, illustrating that ranking quality is a budget-dependent predictor of one-shot RAG accuracy. **(b)** Adaptive stopping at fixed retriever (Qwen3-Emb-4B) matches fixed-$k = 2$ accuracy at roughly 19% fewer retrieved passages. **(c)** Conditioning second-hop retrieval on an LLM-generated probe of unmet need (B3) improves Recall@2 and accuracy over query-only (B1), while naive concatenation (B2) degrades both.

D.7 Per-(retriever, k) metrics for Supporting study 1

Tables 4–6 report Recall@k and one-shot answer accuracy for each retriever and budget on the three datasets. Numbers are means over 200–300 questions. The cross-dataset Spearman correlations reported in Fig. 4a are computed from these tables.

Table 4: HotpotQA-distractor: 300 questions, GPT-5.4-mini answerer and judge.

	$k=1$		$k=2$		$k=3$		$k=5$	
Retriever	R@1	Acc	R@2	Acc	R@3	Acc	R@5	Acc
BM25	0.402	0.513	0.595	0.613	0.698	0.693	0.802	0.773
MiniLM-L6	0.413	0.457	0.640	0.627	0.728	0.660	0.838	0.750
MPNet-base	0.432	0.507	0.672	0.643	0.760	0.687	0.850	0.773
BGE-base-v1.5	0.458	0.497	0.765	0.717	0.860	0.773	0.923	0.823
BGE-large-v1.5	0.465	0.510	0.795	0.720	0.877	0.803	0.923	0.813
Qwen3-Emb-0.6B	0.445	0.473	0.710	0.670	0.797	0.757	0.873	0.803
Qwen3-Emb-4B	0.453	0.537	0.762	0.733	0.842	0.773	0.903	0.813
Qwen3-Emb-8B	0.470	0.497	0.770	0.717	0.855	0.773	0.920	0.840

Table 5: 2WikiMultihopQA: 200 questions, GPT-5.4-mini answerer and judge.

	$k=1$		$k=2$		$k=3$		$k=5$	
Retriever	R@1	Acc	R@2	Acc	R@3	Acc	R@5	Acc
BM25	0.357	0.115	0.546	0.305	0.654	0.345	0.769	0.460
MiniLM-L6	0.399	0.160	0.596	0.365	0.682	0.435	0.777	0.490
MPNet-base	0.399	0.155	0.616	0.370	0.691	0.435	0.812	0.535
BGE-base-v1.5	0.446	0.135	0.698	0.450	0.779	0.560	0.868	0.590
BGE-large-v1.5	0.448	0.165	0.723	0.505	0.796	0.570	0.899	0.595
Qwen3-Emb-0.6B	0.441	0.130	0.665	0.390	0.749	0.475	0.833	0.545
Qwen3-Emb-4B	0.441	0.110	0.655	0.415	0.751	0.455	0.844	0.575
Qwen3-Emb-8B	0.444	0.125	0.682	0.430	0.764	0.500	0.863	0.590

Table 6: MuSiQue-Ans: 200 questions, GPT-5.4-mini answerer and judge.

	$k=1$		$k=2$		$k=3$		$k=5$	
Retriever	R@1	Acc	R@2	Acc	R@3	Acc	R@5	Acc
BM25	0.276	0.135	0.424	0.235	0.543	0.305	0.675	0.385
MiniLM-L6	0.335	0.185	0.522	0.295	0.640	0.395	0.809	0.510
MPNet-base	0.330	0.225	0.533	0.375	0.671	0.420	0.821	0.520
BGE-base-v1.5	0.351	0.135	0.583	0.375	0.690	0.420	0.859	0.530
BGE-large-v1.5	0.369	0.170	0.603	0.345	0.736	0.425	0.870	0.490
Qwen3-Emb-0.6B	0.343	0.185	0.543	0.335	0.682	0.380	0.810	0.465
Qwen3-Emb-4B	0.363	0.195	0.579	0.355	0.695	0.455	0.836	0.535
Qwen3-Emb-8B	0.381	0.195	0.600	0.365	0.718	0.410	0.857	0.520