RETRIEVOLVER: A SELF-EVOLVING HARNESS WITH DYNAMIC WORKFLOWS FOR HETEROGENEOUS RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Agents need evidence from text, code, images, document pages, and the web, yet each source typically needs its own retriever, often trained in-domain, or a hand-built pipeline. We present Retrievolver, a self-evolving retrieval harness with a common interface for tasks, tools, and evidence. A planner with fixed model weights writes a workflow for each query. The workflow combines specialist retrievers and rerankers. The planner follows a Policy and Skills stored as readable documents. Users provide a corpus and tools, a few hundred training and development queries, and a quality evaluator. Retrievolver adds task-relevant methods from papers and author code to its tool pool, then revises its guidance by tracing where evidence was lost during search, ranking, or delivery. Development evaluation selects the version frozen for testing. In a five-task text ablation with a shared tool pool and fixed model weights, evolved guidance raises macro nDCG@10 by 8.8 points over generic Initial guidance and 7.5 over a fixed pipeline. Across 42,721 test queries in five task families, Retrievolver leads four, and its development-selected systems, which also expand tools and adapt their execution limits, improve the primary metric of all five over their initial systems. Transfer to five target benchmarks uses no target relevance labels, and the frozen retrieval module raises BrowseComp-Plus answer accuracy in all five agent hosts.

1 INTRODUCTION

Agents that answer questions over documents, inspect code repositories, or search visual collections rely on different forms of evidence: passages, code, images, and document pages. Specialist retrievers serve different evidence types, while unified models such as UniHGKR and UniIR now support several of these sources within a single model (Min et al., 2025; Wei et al., 2024). How these capabilities should be combined, however, is left for each application to determine. Tracing a code dependency, interpreting a scanned page, and traversing a multi-hop evidence chain each demand distinct retrieval procedures; a reusable framework should therefore accommodate this variety while allowing users to refine retrieval on their own data.

A second complication is that the appropriate procedure varies from query to query, not merely from task to task. One query may be resolved with a single search, whereas another requires reading an intermediate result, identifying a missing entity, and searching again. FlowSearcher and FlowEvo combine adaptive workflows with reusable experience (Xiang et al., 2026; Ren et al., 2026), while GEPA and SkillOpt refine prompts and skills from feedback (Agrawal et al., 2026; Yang et al., 2026a). Applying these ideas to retrieval requires understanding how individual decisions affect the evidence ultimately returned. Retrieval can fail at several stages: relevant evidence may be missed during search, demoted by reranking, or omitted from the final output. The final score alone cannot distinguish these failures; intermediate candidates and rankings, examined alongside evaluation feedback, can help identify where they occur. We ask whether experience from retrieval trajectories can be accumulated into reusable guidance that improves the effectiveness of dynamic workflows on subsequent queries.

We present **Retrievolver**, a self-evolving retrieval harness that provides unified task, tool, and evidence interfaces for text, code, images, and document pages. Through these interfaces, users supply

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

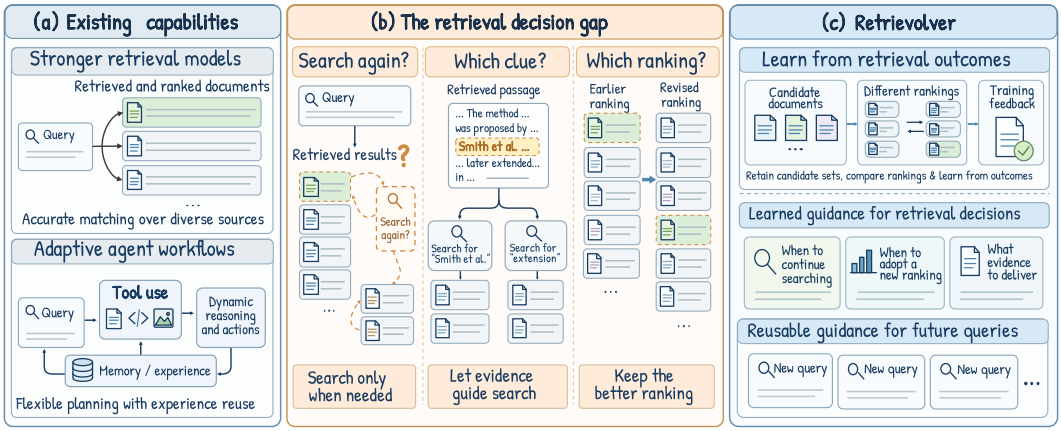

Figure 1: Retrievolver combines specialist retrievers with adaptive workflows and learns reusable guidance from training feedback on retained candidates and rankings.

retrieval resources, training and development queries, and a quality evaluator. A language-model planner generates executable Python workflows and adjusts subsequent steps as intermediate results arrive. The planner follows a Policy for operation selection and result adoption, Workflow Skills for retrieval procedures, and Tool Skills for tool use. This guidance informs its decisions about further search, result merging, ranking retention, and evidence delivery. The harness returns ranked evidence with source references, enabling downstream agents to inspect the provenance of each item.

Retrievolver improves these decisions through offline evolution. It first incorporates task-relevant methods, drawn from published papers and their authors' released code, into its tool pool and skill library. The evolver then examines training trajectories, including candidate sets, ranking changes, and evaluator feedback, to identify decisions worth revising. It writes the resulting updates back into the relevant Policy, Workflow Skills, or Tool Skills. Experience accumulated on earlier queries therefore carries over into how subsequent workflows are generated and executed. For instance, feedback might motivate a rule to retain an existing ranking unless further retrieval addresses an identifiable evidence gap. The guidance carried into testing is selected on development performance and remains fixed thereafter; the planner continues to generate and adjust a workflow for each query, and model weights are never modified.

We evaluate Retrievolver on five retrieval task families, where it outperforms the compared systems on the primary metrics in four of the five. With the expanded tool pool held fixed, evolved guidance improves macro nDCG@10 by 8.8 points over the generic Initial guidance across five text tasks. Transfer without target relevance labels yields mixed results across five target benchmarks, indicating that the benefits of reuse vary by task. The frozen retrieval module also improves BrowseComp-Plus answer accuracy for all five host agents we evaluate.

Our contributions are threefold:

- **A reusable framework for heterogeneous retrieval.** Retrievolver unifies the task, tool, and evidence interfaces so that specialist tools and task-specific evolution can be applied directly to a user's own data.

- **Trajectory-driven learning for dynamic retrieval workflows.** Retrievolver exploits feedback on intermediate candidates and rankings, rather than on final scores alone, to revise guidance for workflow generation, continued search, and evidence delivery.

- **An evaluation covering retrieval, transfer, and downstream integration.** We measure the benefits of evolved guidance, evaluate system reuse across target benchmarks, and examine whether retrieval gains translate into better answers from existing agents.

2 Related Work

Heterogeneous retrieval and retrieval agents. UniHGKR and UniIR use unified representations for different knowledge types and image–text tasks, while ColPali directly encodes document pages to preserve text, layout, and charts (Min et al., 2025; Wei et al., 2024; Faysse et al., 2025). Evo-Retriever trains a multimodal retriever with an LLM-guided adaptive hard-negative curriculum (Li et al., 2026b). Agents organize retrieval itself: LocAgent localizes code through structural graphs, and OmniSearch and MMSearch-Engine enable multimodal search (Chen et al., 2025; Li et al., 2025c; Jiang et al., 2025). DocLens and VRAG-RL explore page navigation and visual search strategies (Zhu et al., 2026; Wang et al., 2025a). Retrievolver instead exposes specialized retrievers as tools under one execution interface and learns conditional workflow guidance reused across queries.

Dynamic workflows. ReAct interleaves reasoning and action (Yao et al., 2023), and AFlow searches over code-based workflows (Zhang et al., 2025a). FlowSearcher combines persistent execution memory with workflow synthesis, and FlowEvo compiles successful workflows into an evolving bank of executable skills (Xiang et al., 2026; Ren et al., 2026). Retrievolver updates guidance for retrieval decisions, which the planner uses to revise workflow code as evidence arrives.

Self-evolving agents and harnesses. GEPA optimizes prompts through trajectory reflection (Agrawal et al., 2026), SkillOpt makes bounded skill edits with validation-based selection (Yang et al., 2026a), and Evo-Harness compiles single-shot executions into reusable skills outside a frozen solver (Wei et al., 2026). Retrievolver uses retained candidates and ranking changes to guide revisions of retrieval workflows and evidence delivery.

3 Method

Retrievolver combines heterogeneous tools with Policy and Skill guidance for dynamic workflows (Figure 2). Offline, Stage 1 expands tools and workflows; Stage 2 refines guidance using training feedback and development selection (Figure 3).

3.1 Unified Retrieval Framework and Layered Skills

A task supplies query q, metadata m, retrieval resources, execution limits, and output requirements. Common interfaces accept text queries, images with modification instructions, and document pages, and provide training queries, feedback, and development data. Returned evidence identifiers link to source content and locations.

The harness $\mathcal{H}$ pairs an execution kernel $\mathcal{K}$ with learnable guidance:

$$\mathcal{H} = (\mathcal{K}, \theta), \qquad \theta = (P, \mathcal{L}), \quad \mathcal{L} = (\mathcal{S}^{\mathrm{wf}}, \mathcal{S}^{\mathrm{tool}}), \tag{1}$$

Here, $\mathcal{K}$ contains tools, models, and indexes; $\mathcal{L}$ is the skill library. Guidance θ applies to query requirements and intermediate observations. Policy P guides skill selection, composition, and ranking adoption. Workflow Skills $\mathcal{S}^{\mathrm{wf}}$ describe retrieval procedures; Tool Skills $\mathcal{S}^{\mathrm{tool}}$ specify parameters, applicability, and result interpretation.

Skill storage and composition. Each `SKILL.md` specifies applicability, inputs, outputs, procedures, and handoff instructions. The planner receives the Policy and available Tool Skills in advance and reads Workflow Skills on demand. Skills can be reused and combined across workflows.

Tool and workflow expansion. The offline Searcher uses a language model and web search to find task-relevant retrieval, reading, fusion, and reranking methods. It integrates their backends through typed interfaces and documents procedures and tool use as skills, without planning, executing, or scoring evaluation queries (Appendix A.2). Initial (R0) denotes hand-written starting guidance; Evolved denotes development-selected guidance. Each experiment specifies its tool pool and skills.

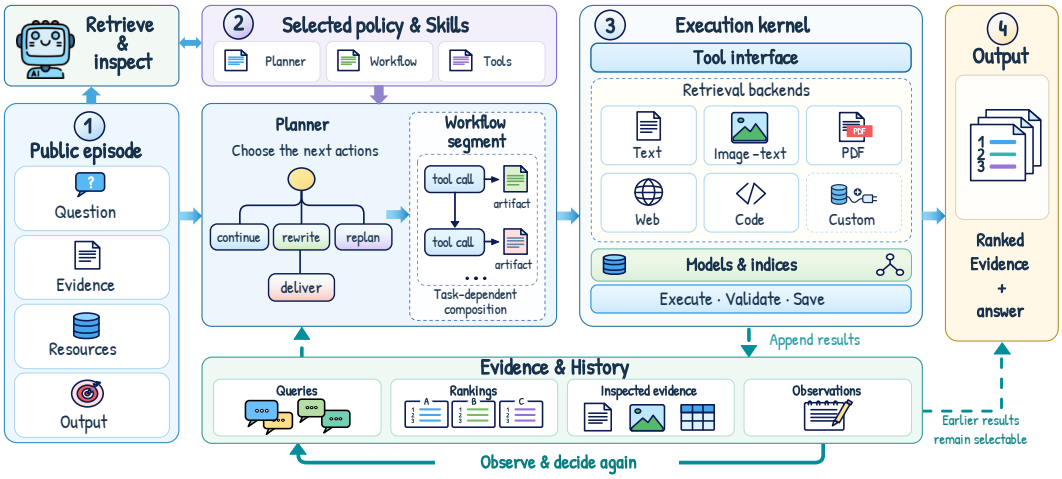

Figure 2: Retrievolver's retrieval framework. The planner chooses tools using guidance and intermediate evidence; the kernel retains results and trajectories.

3.2 Dynamic Workflows and Context Management

The planner π_ϕ, with model weights ϕ, reads the query, guidance, and observations to choose retrieval, rewriting, candidate processing, or delivery. Let $\mathcal{M}_t$ denote the complete execution record and h_t the current model context:

$$\begin{aligned} a_t &\sim \pi_\phi(\cdot \mid q, m, h_t, \mathcal{K}, \theta), \\ o_t &= \text{Execute}_{\mathcal{K}}(a_t), \qquad \mathcal{M}_{t+1} = \mathcal{M}_t \oplus (a_t, o_t). \end{aligned} \tag{2}$$

Here, a_t is a Python segment or a control action (continue, rewrite, replan, or deliver); $\oplus$ appends a record. Programs use variables, conditionals, loops, and parallel calls. The kernel validates arguments, executes calls, and saves outputs. Returned observations guide subsequent code generation or revision, changing tools, call order, or control flow.

From local operations to workflows. Tool calls form nodes linked by result dependencies. Observations extend this graph: new entities prompt retrieval, evidence gaps prompt reading and rewriting, and complementary candidates prompt fusion. Figure 6 pairs replanning decisions with subsequent code.

Retained results and on-demand context. Tools return identifiers, counts, summaries, and parent-result references; the kernel stores full content and rankings separately. The planner forms h_t by reading summaries, snippets, or documents on demand. Retrieval and reranking create new result objects without deleting earlier rankings, which remain available for delivery if later operations demote useful evidence.

3.3 Feedback-Driven Guidance Refinement

Records and feedback. The evolver revises θ from training executions with frozen model weights and retrieval implementations, and it also revises each task's execution limits separately from θ (Appendix A.3). *Trajectories* preserve queries, tool calls, candidate identifiers and orders, observations, delivered results, and usage. *Training feedback* supplies post-execution retrieval scores, intermediate-to-final score changes where available, and execution errors. Together, trajectories and feedback help locate missing, demoted, or discarded evidence and direct revisions toward search, ranking, or delivery, respectively. Relevance labels only score completed executions and never reach the planner.

Grouped reflection and guidance revision. Following trajectory reflection and textual optimization (Agrawal et al., 2026; Yang et al., 2026a), the evolver groups traces by outcome and retrieval

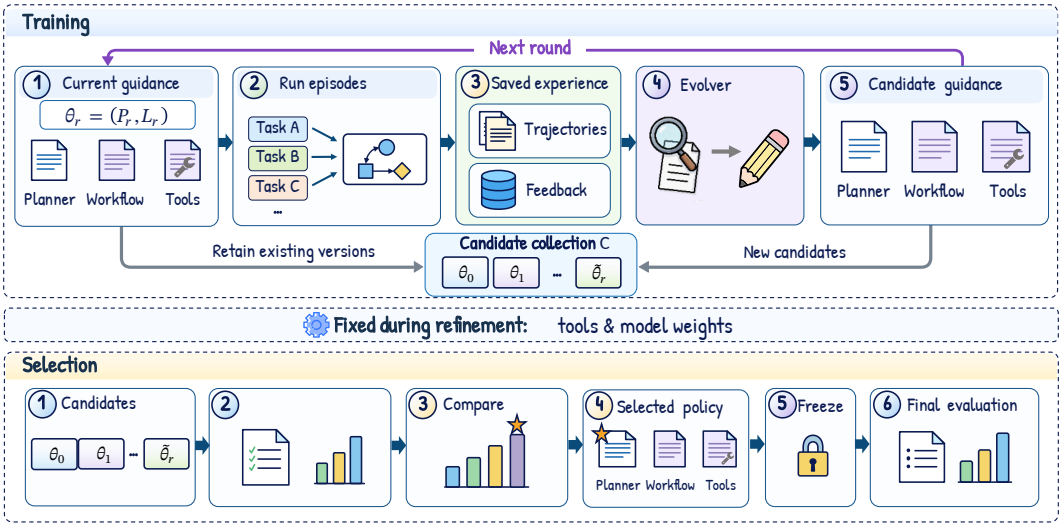

Figure 3: Stage 2 refines guidance from training feedback. Development scores select the version frozen for testing.

decision, consulting indexed details as needed. Contrasting successes with counterexamples, it proposes condition–action rules supported by training cases. Policy edits change skill selection and result adoption; Workflow Skill edits change search, composition, or stopping; Tool Skill edits change individual tool usage. A candidate may revise any subset of these three components, within the editable scope of the batch.

In round r, the evolver revises the parent guidance into one complete candidate using training trajectories $\mathcal{T}_r$, feedback $\mathcal{F}_r$, and earlier candidates' training trials $\mathcal{B}_r$, including failed ones:

$$\begin{aligned} \Delta_r &= \mathrm{Revise}(\theta_r^{\mathrm{parent}}, \mathcal{T}_r, \mathcal{F}_r, \mathcal{B}_r), \\ \widetilde{\theta}_r &= \mathrm{Apply}(\theta_r^{\mathrm{parent}}, \Delta_r). \end{aligned} \tag{3}$$

Executing the candidate $\widetilde{\theta}_r$ (Rr in Figure 4) on training queries yields the next round's feedback, and its development score enters the selection below. The revision parent can differ from the feedback source: problems exposed by a degraded candidate can guide revisions of the current best guidance.

Development selection and freezing. Let $\mathcal{C}_r$ contain the Initial guidance θ_0 and all candidates evaluated through round r. We select

$$\theta_r^\star = \arg\max_{\theta \in \mathcal{C}_r} J(\theta; \mathcal{D}_{\mathrm{sel}}), \tag{4}$$

where J is the task's development metric. The selected guidance is frozen for testing; the planner still replans each query online. Appendix D illustrates guidance revision and compares initial and evolved skills.

4 Experiments

We evaluate retrieval quality across heterogeneous tasks, trace how evolved guidance changes retrieval through cases, curves, and ablations, and test reuse across datasets and host agents.

4.1 Experimental Setup

We evaluate Retrievolver on 42,721 test queries across five complementary task families: text retrieval on six BEIR tasks from MTEB (Muennighoff et al., 2023), code retrieval on CodeRAG-Bench (Wang et al., 2025b), image–text retrieval on M-BEIR (Wei et al., 2024), document-page

retrieval on MMDocIR (Dong et al., 2025), and complex evidence search on BrowseComp-Plus (Chen et al., 2026b). Primary metrics are macro-average nDCG@10 for MTEB and CodeRAG-Bench, macro-average Recall@K in local candidate pools for M-BEIR ($K = 10$ for Fashion200K and $K = 5$ otherwise), page R@5 for MMDocIR, and Evidence and Gold nDCG@10 for BrowseComp-Plus.

The planner and refinement evolver use GPT-5.5 with frozen weights. All families use the same Policy–Skill structure and refinement procedure, with task-specific guidance, tools, and execution limits. For BrowseComp-Plus, guidance learning uses separate AgentIR/WebShaper queries (Chen et al., 2026a; Tao et al., 2026), with development selection on AgentIR. We compare BM25 (Robertson & Zaragoza, 2009), open-weight retrieval models, fixed pipelines, and agents, including GPT-5.5 reruns. Appendices B and A expand results and settings.

4.2 Heterogeneous Retrieval Performance

Retrievolver leads four of the five task families (Tables 1–5).

Text retrieval: the largest gains come on multi-hop queries. On MTEB (Table 1), Retrievolver reaches 70.7 macro nDCG@10, 2.3 points above NV-Embed-v2 and above a test-label oracle that picks the best standalone tool per task (70.0; Appendix B.1). Multi-hop HotpotQA contributes 1.2 of the 2.3-point margin (92.9 versus 85.5), and Retrievolver also leads on FiQA and NFCorpus.

Table 1: MTEB text retrieval in nDCG@10. Bold and underline mark the best and second-best scores across all rows; † marks published results.

Method	FiQA	MS MARCO	SciFact	NF Corpus	Quora	Hotpot QA	Macro
BM25†	23.6	22.8	66.5	32.5	78.9	60.3	47.4
NV-Embed-v2	<u>65.2</u>	**45.6**†	80.1†	45.0	<u>89.0</u>†	85.5†	<u>68.4</u>†
SIRA	44.6	36.7	82.0	<u>45.1</u>	84.3	68.8	60.3
AFlow	63.1	41.3	82.7	42.6	87.3	29.7	57.8
Evo-Harness	61.4	39.8	**83.1**	43.5	87.1	33.0	58.0
GEPA	60.8	39.5	**83.1**	43.5	87.3	37.5	58.6
Initial Retrievolver	56.0	32.8	82.2	43.8	84.8	<u>86.9</u>	64.4
Retrievolver	**66.8**	<u>45.1</u>	<u>82.8</u>	**46.9**	**89.8**	**92.9**	**70.7**

Code retrieval: cross-file RepoEval drives the margin. On CodeRAG-Bench (Table 2), Retrievolver reaches 81.5 macro nDCG@10, 1.5 points above RepoNav and 9.2 above the per-task oracle over its code tools (72.3). RepoEval, which requires locating code across files, drives the margin (92.3 versus 74.3).

Table 2: CodeRAG-Bench nDCG@10. Macro averages six tasks; Micro averages 872 test queries. Notation follows Table 1.

Method	HumanEval	MBPP	DS-1000	ODEX	RepoEval	SWE-bench Lite	Macro	Micro
BM25	<u>99.7</u>	98.9	10.0	11.2	<u>86.2</u>	56.7	60.4	58.1
BGE-Code-v1	**100.0**	**100.0**	46.7	33.8	83.4	66.9	71.8	70.3
A-RAG	**100.0**	**100.0**	53.6	42.9	74.0	**94.4**	77.5	75.4
RepoNav	**100.0**	**100.0**	**58.1**	**54.5**	74.3	<u>92.8</u>	<u>80.0</u>	<u>77.3</u>
Initial Retrievolver	**100.0**	<u>99.6</u>	49.6	44.7	85.4	87.2	77.8	76.4
Retrievolver	**100.0**	<u>99.6</u>	<u>55.7</u>	<u>53.2</u>	**92.3**	88.1	**81.5**	**80.5**

Image–text retrieval: competitive results from released models. On M-BEIR (Table 3), Retrievolver reaches 60.4 macro-average Recall@K, above GME-7B and the locally evaluated multimodal agents, and 0.5 points below U-MARVEL (Li et al., 2026c), the retriever it composes with Qwen3-VL and LamRA (Liu et al., 2025b) reranking (Appendix A.2).

Table 3: M-BEIR Recall@K. $K = 10$ for Fashion200K and $K = 5$ otherwise. Notation follows Table 1.

Method	WebQA	Fashion200K	NIGHTS	EDIS	CIRR	Macro
GME-7B	96.4	32.2	31.6	<u>81.0</u>	55.6	59.4
U-MARVEL†	97.1	<u>33.6</u>	<u>34.2</u>	78.8	60.7	<u>60.9</u>
OmniSearch	93.4	6.2	28.6	35.6	20.2	36.8
MMSearch-Engine	94.1	16.6	23.6	56.5	33.5	44.9
U-MARVEL+†	**98.5**	**34.2**	**34.7**	**81.4**	**63.2**	**62.4**
Initial Retrievolver	95.5	32.2	30.2	74.6	55.2	57.5
Retrievolver	<u>98.0</u>	30.5	32.5	79.1	<u>61.7</u>	60.4

It exceeds U-MARVEL on WebQA, EDIS, and CIRR, whose queries carry entities or modification instructions the planner can reason over, and trails it on Fashion200K and NIGHTS, where relevance rests on fine visual attributes and perceptual similarity that the encoder already scores directly. OmniSearch and MMSearch-Engine trail GME-7B by 22.6 and 14.5 points, respectively.

Document pages and complex evidence: precise top-ranked evidence. On MMDocIR (Table 4), Retrievolver attains 73.2/90.4/93.4 page R@1/3/5. Compared with Initial, R@1 and R@5 improve slightly, while R@3 falls from 90.6 to 90.4. On BrowseComp-Plus (Table 5), it exceeds DR-DCI by 7.4 Evidence and 18.4 Gold nDCG@10 points, while DR-DCI retains more evidence at deeper cutoffs.

Table 4: MMDocIR page recall. Notation follows Table 1.

Method	R@1	R@3	R@5
GME†	54.2	76.0	83.8
ColQwen†	63.5	82.1	86.7
DocLens	69.7	87.6	90.1
ColQwen + GPT-5-mini†	70.1	87.8	91.4
Initial Retrievolver	<u>72.8</u>	**90.6**	<u>93.1</u>
Retrievolver	**73.2**	<u>90.4</u>	**93.4**

Table 5: BrowseComp-Plus retrieval against Evidence and Gold annotations. Notation follows Table 1.

	Evidence annotations				Gold annotations			
Method	nDCG@10	R@5	R@100	R@1000	nDCG@10	R@5	R@100	R@1000
BM25†	1.6	1.2	4.7	13.7	1.7	1.4	6.1	17.3
Qwen3-Embedding-8B†	20.3	14.5	47.7	<u>76.7</u>	19.5	18.5	55.8	83.5
OpenResearcher / 30B-A3B	46.4	36.5	58.7	75.2	40.7	41.3	61.1	77.7
DR-DCI	<u>67.2</u>	<u>55.7</u>	**77.5**	**88.3**	<u>60.6</u>	<u>66.8</u>	**85.5**	**93.0**
Initial Retrievolver	52.0	39.3	53.2	54.5	56.8	55.4	66.0	67.2
Retrievolver	**74.6**	**60.0**	<u>71.5</u>	71.5	**79.0**	**77.3**	<u>84.0</u>	<u>84.0</u>

4.3 What Evolves and How Behavior Changes

Feedback turns ranking errors into reusable rules. In an ODEX training case, selecting eight of 100 documents in solution-step order demotes `time.strftime` from rank one to seven, re-

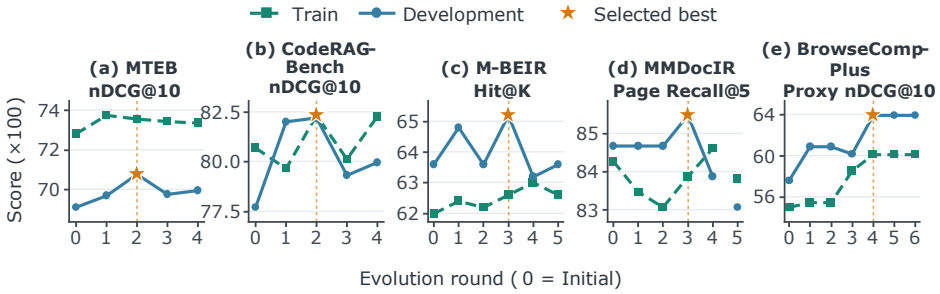

Figure 4: Training and development scores, with selected rounds starred. MMDocIR selects R3 at 85.5; unconnected R5 points use a separate evaluation. M-BEIR Hit@K denotes macro Recall@K. BrowseComp-Plus uses AgentIR queries.

ducing nDCG@10 from 1.000 to 0.333. The evolver adds a Policy rule requiring content-based justification before replacing, demoting, or deleting evidence. With this rule, the planner retains all 100 documents and restores nDCG@10 to 1.000 (Appendix D.4). In an NFCorpus query, it tries three alternative programs, finds no better ranking, and returns the retained initial result (Figure 6).

Evolution sharpens when to adopt new results. The selected MTEB skill merges inspected useful hits into the retained ranking and excludes noisy supplemental results. Development scores rise from 69.1 to 70.8 on MTEB and from 77.7 to 82.2 on CodeRAG-Bench, with R2 selected in both (Figure 4). On M-BEIR, R4 adds 0.4 training points over R3 but loses 2.0 development points, so development selection retains R3.

4.4 Ablations of Tools and Learned Skills

Table 6 compares five text test sets (6,129 queries). Base contains BM25 and Qwen3-Embedding-8B; Expanded adds specialized retrieval and reranking from a task-specific subset of the Table 1 tool pool. Fixed execution applies equal-weight RRF to original-query rankings. Initial uses a generic Policy and text-retrieval Workflow Skill; Evolved uses the development-selected MTEB versions. Appendices A.3 and E.1 detail guidance and planner usage.

Table 6: Tool expansion and guidance evolution on five text-retrieval test sets (nDCG@10, %). Unseen tasks are excluded from guidance generation and selection. Macro averages and gains are computed before rounding.

Harness configuration			Learning tasks			Unseen tasks		All
Tool pool	Workflow	Guidance	FiQA	SciFact	NFCorpus	ArguAna	NQ	Macro
Base	Fixed	—	43.5	75.4	38.8	59.3	50.9	53.6
Expanded	Fixed	—	54.8	79.3	41.7	66.7	59.6	60.4
Base	Dynamic	Initial	56.0	82.2	43.8	57.3	58.7	59.6
Expanded	Dynamic	Initial	50.9	79.8	42.5	59.9	62.4	59.1
Expanded	Dynamic	**Evolved**	**66.3**	**82.8**	**46.0**	**72.0**	**72.7**	**67.9**

Tool expansion improves fixed execution (53.6 to 60.4) but slightly lowers Initial (59.6 to 59.1). With the same expanded pool, Evolved reaches 67.9, exceeding Initial by 8.8 points and fixed execution by 7.5. It improves every task over expanded-pool Initial, including ArguAna (+12.1) and NQ (+10.3), which enter neither guidance generation nor selection.

This comparison measures combined Policy and Workflow Skill updates. Reverting one component alone may disrupt its coordination with the others, so such an ablation would also reflect their interactions.

Table 7: BrowseComp-Plus answer accuracy in percent on paired questions. Host is the original agent; Δ is the gain in percentage points before rounding.

Agent	Host	+ Retrievolver	Δ
DR-DCI	92.5	**95.1**	+2.7
MiroFlow	59.5	**64.5**	+5.0
A-RAG	80.9	**87.7**	+6.8
ReAct	38.1	**70.9**	+32.8
ACE generator (no adaptation)	44.0	**89.6**	+45.7

4.5 Cross-Dataset Transfer

We transfer source-task guidance to LitSearch (Ajith et al., 2024), ViDoRe v2 (Macé et al., 2025), MultiHop-RAG (Tang & Yang, 2024), Flickr30K (Young et al., 2014), and five CoIR datasets (Li et al., 2025b). The Searcher adapts target tools and, where needed, guidance using public methods, target corpora, and indexes without target labels. Configurations are fixed before evaluation; workflows still depend on each query and intermediate evidence. Appendix B.2 gives the protocol and full comparisons.

Document-page and multi-hop retrieval. On ViDoRe v2, Retrievolver leads on all four subsets and exceeds NeMo by 3.2 macro nDCG@5 points. On MultiHop-RAG, it leads in chunk-level Hits@10 and article-level nDCG@10 (88.1 vs. DR-DCI's 87.4). ViDoRe v2 and MMDocIR both require page-level evidence retrieval, while MultiHop-RAG and BrowseComp-Plus require following multiple clues. The results are consistent with reusing such procedures to locate evidence and place it earlier in the ranking.

Literature, image–text, and code retrieval. On LitSearch, Retrievolver ranks second on four of six metrics, with 74.6 nDCG@10 compared with 77.1 for ReAct. On Flickr30K, it exceeds OmniSearch and MMSearch-Engine on all six recall metrics; its image-to-text and text-to-image R@1 scores are 0.9 and 1.5 points below BLIP-2 ViT-g, respectively. On CoIR, Retrievolver achieves 70.6 macro nDCG@10, exceeding both compared agents; BGE-Code-v1 and CodeXEmbed-7B score 76.1 and 71.8, respectively.

4.6 Downstream Integration

Table 7 evaluates the frozen Retrievolver retrieval module within DR-DCI (Lu et al., 2026), MiroFlow (Su et al., 2026), A-RAG, ReAct, and the ACE generator without experience adaptation (Zhang et al., 2026). Retrievolver replaces search in DR-DCI, MiroFlow, and ReAct, supplies ACE's initial evidence, and adds an optional A-RAG tool. Hosts retain their interaction and answering procedures and receive corpus evidence, not Retrievolver-generated answers. All use R4 guidance selected on the AgentIR development split and frozen for evaluation (Appendix C).

Retrievolver improves every host (Table 7), lifting A-RAG from 80.9% to 87.7%. On the strong DR-DCI host, integration removes a net 35.5% of baseline errors. For ACE, 361 of the 384 questions corrected by integration (94.0%) change from receiving no Gold document to receiving at least one.

5 Conclusion

Retrievolver combines heterogeneous retrieval tools through unified interfaces and learns reusable guidance from execution feedback. Its planner uses this guidance to generate and revise workflows as evidence arrives, without updating model weights. With the expanded tool pool fixed, evolved guidance improves macro nDCG@10 by 8.8 points over Initial across five text tasks. The complete system leads on the primary metrics in four of five task families and improves answer accuracy in all five host agents. These results support learning retrieval decisions from trajectories, while the mixed transfer results indicate that the benefits of reuse depend on the target task.

Reproducibility Statement

Section 3 specifies the harness, both offline stages, and online execution. Section 4.1 and Appendix A.1 give the benchmarks, data partitions, and metrics. Appendix A.2 lists the tools and model roles, and Appendix A.3 gives the prompts and execution limits. Appendices A.4, B.2, and C specify the optimizer-comparison, transfer, and application protocols, and Appendix E.1 reports the recorded computation. The supplementary files `reproducibility_guidance.txt` and `reproducibility_ablation.txt` contain the Policies, Workflow Skills, and evolver instructions, including the expanded-pool Initial and Evolved guidance used in the ablation. We plan to release the Retrievolver code, the Initial and Evolved guidance, experiment configurations, split descriptions, and evaluation and plotting scripts. Datasets and pretrained models remain subject to their original licenses, and hosted model APIs require separate access.

AI Use Statement

We used Codex to assist with text revision, literature organization, figure scripts, and LaTeX typesetting. The roles of language models in the experiments are described in the method and experimental setup. The research design, experimental analysis, and final manuscript were reviewed and approved by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

This work uses public retrieval benchmarks, open-weight retrieval models, and hosted language models. During execution, the planner receives only the public query, available resources, guidance, and observations from completed tool calls. Relevance annotations and reference answers are used only after execution, for scoring and, on learning splits, for training feedback and development selection. Retrieved content and model outputs may contain errors or biases, and any redistribution follows the original licenses.

Acknowledgments

We thank the authors and maintainers of the public benchmarks, pretrained models, and open-source frameworks.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.840. URL https://aclanthology.org/2024.emnlp-main.840/.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Jenna Sparks, Jena D. Hwang, Varsha Kishore, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Daniel S. Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650(8103), 2026. URL https://doi.org/10.1038/s41586-025-10072-4.

Hongzheng Chai, Jiakun Li, Hongyue Yu, and Yuan Yuan. RepoNav: From snippet retrieval to file-centered repository navigation for code agents. *arXiv preprint arXiv:2609.08355*, 2026. URL https://arxiv.org/abs/2609.08355. Accepted to EMNLP 2026. Code: https://github.com/CCChz233/reponav.

Zhaoling Chen, Robert Tang, Gangda Deng, Fang Wu, Jialong Wu, Zhiwei Jiang, Viktor Prasanna, Arman Cohan, and Xingyao Wang. LocAgent: Graph-guided LLM agents for code localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. URL https://aclanthology.org/2025.acl-long.426/.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Jimmy Lin, Akari Asai, and Victor Zhong. AgentIR: Reasoning-aware retrieval for deep research agents. In *Third Conference on Language Modeling*, 2026a. URL https://openreview.net/forum?id=rn8kjxilhF.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-Plus: A fair and disentangled evaluation benchmark for deep search agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026b. URL https://aclanthology.org/2026.acl-long.1023/.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.1576/.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. A-RAG: Scaling agentic retrieval-augmented generation via hierarchical retrieval interfaces. *arXiv preprint arXiv:2602.03442*, 2026. URL https://arxiv.org/abs/2602.03442.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Geonmo Gu, Byeongho Heo, Jaemyung Yu, Jaehui Hwang, Taekyung Kim, Sangmin Lee, HeeJae Jun, Yoohoon Kang, Sangdoo Yun, and Dongyoon Han. MuCo: Multi-turn contrastive learning for multimodal embedding model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Gu_MuCo_Multi-turn_Contrastive_Learning_for_Multimodal_Embedding_Model_CVPR_2026_paper.html.

Tz-Huan Hsu, Jheng-Hong Yang, and Jimmy Lin. Rethinking Agentic Search with Pi-Serini: Is lexical retrieval sufficient? *arXiv preprint arXiv:2605.10848*, 2026. URL https://arxiv.org/abs/2605.10848.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Unveiling the potential of large models as multi-modal search engines. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=J2Jyp1SZ0n.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c4bf73386022473a652a18941e9ea6f8-Abstract-Conference.html.

Aaron Branson Cigres Li, Zhaowei Wang, Yu Zhao, Yiming Du, Haobo Li, Xiyu Ren, Ginny Wong, Simon See, Lishu Luo, Haodong Duan, Pasquale Minervini, and Yangqiu Song. Can retrieval heads see images? multimodal retrieval heads in long-context vision-language models. *arXiv preprint arXiv:2605.27243*, 2026a. URL https://arxiv.org/abs/2605.27243.

Chaofan Li, Jianlyu Chen, Yingxia Shao, Defu Lian, and Zheng Liu. Towards a generalist code embedding model based on massive data synthesis. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://doi.org/10.52202/085713-4966.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Weiqing Li, Jinyue Guo, Yaqi Wang, Haiyang Xiao, Yuewei Zhang, Guohua Liu, and Hao Henry Wang. Evo-Retriever: LLM-guided curriculum evolution with viewpoint-pathway collaboration for multimodal document retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Evo-Retriever_LLM-Guided_Curriculum_Evolution_with_Viewpoint-Pathway_Collaboration_for_Multimodal_Document_CVPR_2026_paper.html.

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Hao Zhang, Xinyi Dai, Yasheng Wang, and Ruiming Tang. CoIR: A comprehensive benchmark for code information retrieval models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025b. URL https://aclanthology.org/2025.acl-long.1072/.

Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. U-MARVEL: Unveiling key factors for universal multimodal retrieval via embedding learning with MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=CUFJluq5O8.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=VvDEuyVXkG.

Zhuofeng Li, Dongfu Jiang, Xueguang Ma, Haoxiang Zhang, Ping Nie, Yuyu Zhang, Kai Zou, Jianwen Xie, Yu Zhang, and Wenhu Chen. OpenResearcher: A fully open pipeline for long-horizon deep research trajectory synthesis. *arXiv preprint arXiv:2603.20278*, 2026d. URL https://arxiv.org/abs/2603.20278.

Ye Liu, Rui Meng, Shafiq Joty, Silvio Savarese, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. CodeXEmbed: A generalist embedding model family for multilingual and multi-task code retrieval. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=z3lG70Azbg.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. LamRA: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_LamRA_Large_Multimodal_Model_as_Your_Advanced_Retrieval_Assistant_CVPR_2025_paper.html.

Yi Lu, Zhuofeng Li, Ping Nie, Haoxiang Zhang, Yuyu Zhang, Kai Zou, Wenhu Chen, Jimmy Lin, Dongfu Jiang, and Yu Zhang. Dr-DCI: Scaling direct corpus interaction via dynamic workspace expansion. *arXiv preprint arXiv:2606.14885*, 2026. URL https://arxiv.org/abs/2606.14885.

Quentin Macé, António Loison, and Manuel Faysse. ViDoRe benchmark V2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*, 2025. URL https://arxiv.org/abs/2505.17166.

Dehai Min, Zhiyang Xu, Guilin Qi, Lifu Huang, and Chenyu You. UniHGKR: Unified instruction-aware heterogeneous knowledge retrievers. 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)*, 2025. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.234/.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Radek Osmulski, Reza Esfandiarpoor, Yauhen Babakhin, Gabriel de Souza Pereira Moreira, Oliver Holworthy, and Bo Liu. Beyond Semantic Similarity: Introducing NVIDIA NeMo retriever's generalizable agentic retrieval pipeline. NVIDIA technical blog on Hugging Face, March 2026. URL https://huggingface.co/blog/nvidia/nemo-retriever-agentic-retrieval. Published March 13, 2026.

Zeyu Ren, Ling Yue, Ran Li, Yishu Wang, Shengxiang Xu, Hanmo Liu, Shaowu Pan, and Shimin Di. FlowEvo: Self-evolving agents through the co-evolution of workflows and executable skills. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=hU2N7IIkcE.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 2009. URL https://doi.org/10.1561/1500000019.

Shiqian Su, Sen Xing, Xuan Dong, Muyan Zhong, Bin Wang, Xizhou Zhu, Yuntao Chen, Wenhai Wang, Yue Deng, Pengxiang Zhu, Ziyuan Liu, Tiantong Li, Jiaheng Yu, Zhe Chen, Lidong Bing, and Jifeng Dai. MiroFlow: Towards high-performance and robust open-source agent framework for general deep research tasks. *arXiv preprint arXiv:2602.22808*, 2026. URL https://arxiv.org/abs/2602.22808. Code: https://github.com/MiroMindAI/MiroFlow.

Yiqun Sun, Pengfei Wei, and Lawrence B. Hsieh. Very efficient listwise multimodal reranking for long documents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65825.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t4eB3zYWBK.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Pu Wu, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Wentao Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. WebShaper: Agentically data synthesizing via information-seeking formalization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hld4TzJsnD.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. VRAG-RL: Empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://openreview.net/forum?id=EeAHhNwXPV.

Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F. Xu, Yiqing Xie, Graham Neubig, and Daniel Fried. CodeRAG-Bench: Can retrieval augment code generation? In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025b. URL https://aclanthology.org/2025.findings-naacl.176/.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. In *Computer Vision – ECCV 2024*, volume 15145 of *Lecture Notes in Computer Science*, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-73021-4_23.

Tianxin Wei, Zhan Shi, Minhua Lin, Bing He, Zewen Liu, Yisi Sang, Yuanchen Bei, Xuying Ning, Jiaru Zou, Ting-Wei Li, Xiao Lin, Yanjun Zhao, Chi Wang, Benoit Dumoulin, Dakuo Wang, Jingrui He, and Hanqing Lu. Evo-Harness: Context-to-harness skill compilation for self-evolving agents. *arXiv preprint arXiv:2608.15071*, 2026. URL https://arxiv.org/abs/2608.15071. Accepted to EMNLP 2026.

Keyi Xiang, Zeyu Feng, Zhuoyi Lin, Yueming Lyu, Boyuan Shi, Yew-Soon Ong, Ivor Tsang, and Haiyan Yin. FlowSearcher: Synthesizing memory-guided agentic workflows for web information seeking. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f15747119e0e3d00a34a7c17776d2999-Paper-Conference.pdf.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. URL https://doi.org/10.1145/3626772.3657878.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. MM-R5: MultiModal reasoning-enhanced ReRanker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*, 2025. URL https://arxiv.org/abs/2506.12364.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, Dongdong Chen, Xue Yang, and Chong Luo. SkillOpt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026a. URL https://arxiv.org/abs/2605.23904.

Zeyu Yang, Xu Han, Qi Ma, Jason Chen, and Anshumali Shrivastava. Superintelligent Retrieval Agent: The next frontier of agentic retrieval. *arXiv preprint arXiv:2605.06647*, 2026b. URL https://arxiv.org/abs/2605.06647.

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

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=z5uVAKwmjf.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic Context Engineering: Evolving contexts for self-improving language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8a94ff6f922d995d7d3f4ebf4143e442-Abstract-Conference.html.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging Modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Bridging_Modalities_Improving_Universal_Multimodal_Retrieval_by_Multimodal_Large_Language_CVPR_2025_paper.html.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025c. URL https://arxiv.org/abs/2506.05176.

Dawei Zhu, Rui Meng, Jiefeng Chen, Sujian Li, Tomas Pfister, and Jinsung Yoon. DocLens: A tool-augmented multi-agent framework for long visual document understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. URL https://aclanthology.org/2026.acl-long.1234/.

APPENDIX OVERVIEW

Section	Page
A. Experimental setup	16
B. Additional retrieval results and transfer	20
C. Downstream agent integration	26
D. Dynamic workflows and guidance evolution	28
E. Computation and limitations	31

A EXPERIMENTAL SETTINGS

Training executions provide feedback for guidance revision, development scores select candidates, and the selected guidance is frozen before testing. This section specifies the data, tools, prompts, and comparison settings.

A.1 DATA PARTITIONS AND CANDIDATE SELECTION

Table 8 lists the query split sizes and Stage 2 refinement schedule of each main benchmark. Each refinement round evaluates one complete new guidance candidate, and development selection returns the highest-scoring version among the Initial guidance and all evaluated candidates (Equation 4). The training and development partitions are both disjoint from the final evaluation queries.

The selection objective J is computed on each benchmark's development partition: macro-average nDCG@10 for MTEB and CodeRAG-Bench, macro Recall@K for M-BEIR, page R@5 for MMDocIR, and, for BrowseComp-Plus, a proxy nDCG@10 computed against AgentIR (Chen et al., 2026a) positive passages.

Table 8: Query counts and refinement rounds. Initial is excluded from Rounds; Selected is chosen on development data.

Benchmark	Train	Dev	Test	Rounds	Selected
MTEB	600	300	25,656	4	R2
CodeRAG-Bench	249	124	872	4	R2
M-BEIR	500	250	13,705	5	R3
MMDocIR	252	127	1,658	5	R3
BrowseComp-Plus	166	83	830	6	R4

MTEB task selection and the ablation suite. For guidance learning, we select six MTEB retrieval tasks that provide labeled non-test queries for training and development: FiQA, MSMARCO, SciFact, NFCorpus, Quora, and HotpotQA. Each task contributes 100 training and 50 development queries. Training queries come from the official training split, except for Quora, which draws them from its development split. Development queries are disjoint from the training queries: HotpotQA and Quora take them from their official development splits, and the other four tasks take them from training-split queries not used for training. Final evaluation uses the official test splits, except that MSMARCO follows the standard MTEB convention of evaluating on its development split. The tool and guidance ablation in Table 6 instead uses a five-dataset suite. This suite covers both learned and unseen tasks: FiQA, SciFact, and NFCorpus also appear among the learning tasks, whereas ArguAna and NQ serve only for evaluation and never enter guidance generation or selection. All five configurations in Table 6 are evaluated on this five-task suite.

Code and document retrieval. CodeRAG-Bench (Wang et al., 2025b) scores retrieval for code-generation problems against annotated canonical documents with nDCG@10. We use six of its tasks: HumanEval, MBPP, DS-1000, ODEX, RepoEval, and SWE-bench Lite. The original benchmark lists 164, 500, 1,000, 945, 373, and 300 problems for these tasks, respectively, and we draw 1,245 retrieval queries from their released test pools. We split these queries into disjoint training, development, and test partitions of 249, 124, and 872 queries, keeping queries from the same original input group in the same partition. All local CodeRAG-Bench comparisons use the 872 test

queries, and the ODEX training cases in Section 4.3 and Appendix D.4 are drawn from the 249 training queries. For MMDocIR, training and development queries come from disjoint documents in seven official training sources (ArxivQA, DUDE, MP-DocVQA, SciQAG, SlideVQA, TAT-DQA, and Wiki-ss), and final evaluation covers the benchmark's ten document domains.

MMDocIR selects R3, which achieves the highest development page R@5 (85.5) in Figure 4.

M-BEIR task coverage. We select five M-BEIR datasets with distinct retrieval requirements: Fashion200K matches descriptions to fashion images, WebQA retrieves passages for questions, EDIS matches captions to news image–text candidates, NIGHTS retrieves visually similar images, and CIRR combines a reference image with a modification instruction. Together, they vary both the query inputs and the evidence to be retrieved. Each dataset contributes 100 training and 50 development queries from disjoint groups in the official training split, and final evaluation uses all official test queries in the corresponding local candidate pools.

Complex-search data. For BrowseComp-Plus, guidance is learned from 249 queries drawn from AgentIR (Chen et al., 2026a) and WebShaper (Tao et al., 2026), partitioned into 166 training queries and 83 AgentIR development queries. Following Chen et al. (2026b), Evidence nDCG@10 scores a ranking against all human-verified evidence documents of a question, and Gold nDCG@10 against the gold documents that contain its answer. The selected guidance is then evaluated on all 830 BrowseComp-Plus questions, none of which enters guidance generation or selection.

CoIR evaluation scope. We evaluate five CoIR datasets, spanning three retrieval relationships: natural-language-to-code retrieval (CoSQA and APPS), cross-framework and cross-language code matching (CodeTransOcean-DL and CodeTransOcean-Contest), and question–answer retrieval over mixed text and code (StackOverflow-QA). Each dataset is evaluated with its released test queries and full retrieval corpus. The reported CoIR macro average is computed over these five datasets. The BGE-Code-v1 and CodeXEmbed-7B reference models include CoIR data in retriever training (Li et al., 2025a; Liu et al., 2025a).

A.2 Retrieval Tools and Model Roles

Table 9 lists the retrieval resources that the planner can call under the selected guidance in each task family, exposed through typed interfaces for retrieval, reranking, fusion, and evidence access. The planner selects operations according to the query and intermediate results, invoking quality assessment and reranking only when needed.

Table 9: Retrieval resources available to the planner in each task family.

Task family	Retrieval and evidence-processing capabilities
MTEB	BM25; Qwen3-Embedding-8B, GTE-Qwen2-7B-instruct, INF-Retriever, NV-Embed-v2, BGE-en-ICL, and MiniCPM-Embedding; Qwen3-Reranker-4B, BGE-reranker-v2-m3, and MedCPT; ranking fusion.
CodeRAG-Bench	Text/code BM25 and symbol search; Qwen3-Embedding-8B, BGE-Code-v1, and CodeRankEmbed; Qwen3-Reranker-4B; fusion and optional quality assessment.
M-BEIR	U-MARVEL-Qwen2VL-7B-Instruct retrieval; cached GME and Qwen rankings for comparison; Qwen3-VL-Reranker-8B and LamRA reranking; fusion and optional quality assessment.
MMDocIR	Late-interaction page retrieval and Qwen3-VL-Embedding-8B; Qwen3-VL-Reranker-8B; Qwen2.5-VL-7B-Instruct for page reading; evidence browsing and candidate selection.
BrowseComp-Plus	Document/passage BM25; Qwen3-Embedding-8B, AgentIR-4B, ReasonIR-8B, Iso-ModernColBERT, and Reason-ModernColBERT; Qwen3-Reranker-4B, monoT5-3B, and language-model reranking; fusion, quality assessment, and answer synthesis.

Model roles. In Stage 1 (tool and workflow expansion), the offline Searcher uses GPT-6-Astra (official model documentation) with web-search tools to identify methods, integrate retrieval backends, and construct Tool and Workflow Skills for the expanded inventory. Its role ends with tool and workflow discovery: it does not plan, execute, or score evaluation queries, and subsequent refinement operates on its outputs through the same interfaces. GPT-5.5, with frozen weights, performs both planning and feedback-driven textual refinement.

Resource access. The planner accesses retrieval results through stored rankings and content references. In MTEB, the configuration exposes complete original-query rankings from the specialized retrieval models, and Qwen3 dense retrieval additionally accepts rewritten queries against its index, each producing its own ranking. Retained rankings let the planner return to an earlier result after subsequent operations. In the other task families, code, image, and document tools retain content references, so the planner inspects evidence without repeatedly inserting the complete corpus or full rankings into its context.

A.3 Guidance, Prompts, and Execution Limits

Prompt structure. The planner receives the public query, the available resources, the Policy, the bound Tool Skills, the Workflow Skills it reads on demand, and observations from completed calls. The evolver instead receives training trajectories with their post-execution feedback and proposes reusable rules. Appendix D.4 gives prompt excerpts for online continuation and offline revision.

A revision can change the Policy, Workflow Skills, Tool Skills, or a subset of them; it need not change all three. Each batch specifies the editable components. The planner uses the resulting guidance to generate and replan code for new queries.

The accompanying source file `reproducibility_guidance.txt` provides policies, task-relevant Workflow Skills, and available evolver instructions. It complements the Initial and Evolved text-ablation guidance in `reproducibility_ablation.txt`.

Fixed workflow and tool bindings. The Base pool contains BM25, Qwen3-Embedding-8B, and ranking fusion. Expanded adds one specialized retriever per task (Table 10), Qwen3-Reranker-4B, MedCPT, and BGE-reranker-v2-m3. This ablation pool is a task-specific subset of the main MTEB inventory. Both fixed rows retrieve the top 1,000 documents from each bound retriever with the unchanged query, merge document identifiers once with equal-weight reciprocal rank fusion (RRF, $k = 60$), and deliver the top 1,000. Ties use the best input rank, then document identifier. Base makes two retrieval calls; Expanded makes three. Neither fixed row calls a planner, rewrites the query, reranks, or performs a second fusion. Dynamic rows choose their operations from the corresponding pool.

Table 10: Retrievers added to Base for all three Expanded configurations in Table 6.

Tasks	Additional retrieval model
FiQA, NFCorpus	NV-Embed-v2
SciFact, ArguAna	INF-Retriever-v1-pro
NQ	MiniCPM-Embedding

Initial denotes the hand-written starting guidance (round R0), and Initial Retrievolver is each family's complete system before self-evolution. Evolved is the development-selected guidance. Base and Expanded specify tool pools, not guidance stages. The complete-system comparisons in Tables 1–5 cover tools, guidance, and execution settings, while Table 6 holds the tool pool fixed.

Initial and Evolved guidance. The two Dynamic Initial rows in Table 6 use the same hand-written Policy and generic text-retrieval Workflow Skill, with Base tools (59.6) or Expanded tools (59.1). Evolved uses the selected MTEB R2 Policy and Workflow Skill with the Expanded bindings (67.9). The 8.8-point contrast within Expanded therefore measures combined Policy and Workflow Skill updates. The supplementary file `reproducibility_ablation.txt` supplies both guidance versions.

Tool expansion under Initial guidance. With Initial guidance, tool expansion changes the macro score from 59.6 to 59.1, whereas it improves the fixed workflow from 53.6 to 60.4. More tools create additional choices about retrieval and result adoption; generic guidance does not consistently turn those choices into gains.

Candidate and execution budgets. Each refinement round yields a single complete candidate, even when reflection covers multiple training batches, so the Rounds column of Table 8 equals the number of new candidates evaluated on development data. Execution limits are set per role and benchmark as part of each task configuration; in complete-system evolution, the evolver can revise them separately from Policy and Skill updates. The planner dynamically chooses how many actions to take within a query, within the configured request or round limits where a family sets them. CodeRAG-Bench allows up to three workflow rounds per query, which are execution rounds within one query rather than refinement rounds, with no separate cap on planner requests. BrowseComp-Plus allows up to 96 planner requests per query and 16,384 output tokens per model request, limits the evidence context to 8,192 characters, and passes at most 20 evidence items of 2,048 characters each to answer synthesis. The M-BEIR policy evolver allows up to four model requests per invocation and 32,768 output tokens per request.

Compared with Initial, the selected CodeRAG-Bench system replaces a 24-request limit with three workflow rounds; BrowseComp-Plus raises the request limit from 8 to 96 and the evidence context from 2,048 to 8,192 characters. These complete-system comparisons therefore do not hold execution limits fixed.

Execution settings of the text ablation. All dynamic runs in Table 6 use the same GPT-5.5 planner with high reasoning effort, the same execution settings, and a 180-second model-call timeout; the planner decides how many actions each query needs. A few executions (7 for expanded-pool Initial and 10 for Evolved) end without a planner delivery; they return a retained original-query ranking and remain in the evaluation.

A.4 External Optimizer Comparisons

Table 11 summarizes the common MTEB setup for four external optimizers: GEPA, AFlow, Evo-Harness, and ACE. All four optimize a shared RAG pipeline outside Retrievolver, start from common query-reformulation and reranking instructions, and use GPT-5.5 with the same training, development, and evaluation queries, without access to the Retrievolver planner, Policy, or Skills. Table 1 therefore compares complete retrieval systems, whereas Table 6 compares generic Initial and Evolved guidance within the same expanded Retrievolver tool pool.

Table 11: Shared MTEB settings for GEPA, AFlow, Evo-Harness, and ACE.

Setting	Configuration
Language model	GPT-5.5
Execution framework	Shared non-Retrievolver RAG pipeline
Training / development	600 / 300 queries across six tasks
Final evaluation	The same 25,656 queries across six tasks
Candidate limit	Starting configuration plus at most four new candidates
Training evaluation limit	At most 2,400 query executions for new candidates
Initial retrieval	Similarity search, top 10 candidates
Optimization targets	Query-reformulation and reranking instructions

All four optimizers start from the same two instructions. The initial reformulation instruction is "Reformulate the original query for semantic retrieval. Preserve its meaning, entities and constraints. Return only the search query." The initial reranking instruction is "Rank the supplied documents by their relevance to the original query. Prefer documents that address its information need and constraints. Return each document number exactly once, most relevant first, as a comma-separated list." GEPA and Evo-Harness use high reasoning effort for optimization, AFlow and ACE use none, and the optimizer output limit is 16,384 tokens for GEPA and 4,096 tokens for the other three methods.

Candidate coverage on HotpotQA. On HotpotQA, these pipelines retrieve ten candidates before reranking and reach R@10 of 26.5–33.7. Retrievolver also accesses complete rankings from several specialized tools and can continue retrieval. The complete-system comparison therefore includes differences in evidence access, not only instruction optimization.

B Additional Retrieval Results

This appendix expands the comparisons in Section 4 with additional baseline methods, task-level results, and retrieval cutoffs. Scores use a 0–100 scale and are displayed to one decimal place. Bold and underlined values mark the best and second-best distinct scores before rounding, respectively. Published results are marked with † and retain their source evaluation settings. Macro averages and differences are computed before rounding; macro averages are unweighted means of the task scores, and a macro † marks a mean that includes published task scores.

B.1 Main-Benchmark Comparisons

The baselines comprise specialized retrieval models, which score query–document relevance directly, and agents or pipelines, which organize retrieval into a sequence of decisions. Several specialized models, such as NV-Embed-v2 and U-MARVEL, are also Retrievolver tools (Appendix A.2); for these models, the comparison contrasts standalone use with use inside a learned workflow.

Text retrieval: the largest gains come on multi-hop queries. On MTEB, Retrievolver reaches 70.7 macro nDCG@10. HotpotQA contributes 1.2 of its 2.3-point advantage over NV-Embed-v2; the remaining five tasks contribute 1.1. Table 12 gives all task scores. Selecting each task's best standalone score among Retrievolver's tools using test labels gives an oracle macro of 70.0.

The text comparison includes BM25 (Robertson & Zaragoza, 2009), NV-Embed-v2 (Lee et al., 2025), Qwen3-Embedding (Zhang et al., 2025c), and further top MTEB embedding models (GTE-Qwen2, BGE-en-ICL, LENS, INF-Retriever, and MiniCPM-Embedding), whose published scores are MTEB results (Muennighoff et al., 2023); BM25 scores are taken from BEIR (Thakur et al., 2021).

Table 12: Full MTEB comparison in nDCG@10. † marks published scores; — indicates unavailable results.

Method	FiQA	MS MARCO	SciFact	NF Corpus	Quora	Hotpot QA	Macro
Retrievers							
BM25†	23.6	22.8	66.5	32.5	78.9	60.3	47.4
GTE-Qwen2	62.0†	45.9	79.1†	40.6†	90.1†	73.1†	65.1†
Qwen3-Embedding-8B†	64.6	43.6	78.5	41.4	88.9	76.8	65.6
MiniCPM-Embedding	56.6†	43.9†	**86.6**†	39.8†	90.0†	77.1	65.7†
BGE-en-ICL	59.7†	<u>46.8</u>†	79.1†	41.9†	**91.1**	85.1†	67.3†
LENS-d8000†	61.6	**47.2**	79.7	40.6	<u>90.8</u>	<u>85.7</u>	67.6
INF-Retriever-v1†	62.4	44.1	<u>85.4</u>	43.7	90.4	82.0	68.0
NV-Embed-v2	<u>65.2</u>	45.6†	80.1†	45.0	89.0†	85.5†	<u>68.4</u>†
INF-Retriever-v1-pro	—	—	82.6	—	—	—	—
Agents / Pipelines							
ACE / GPT-5.5	61.7	39.8	82.6	42.4	85.6	31.5	57.3
AFlow	63.1	41.3	82.7	42.6	87.3	29.7	57.8
Evo-Harness	61.4	39.8	83.1	43.5	87.1	33.0	58.0
GEPA	60.8	39.5	83.1	43.5	87.3	37.5	58.6
SIRA + Qwen	44.6	36.7	82.0	<u>45.1</u>	84.3	68.8	60.3
Retrievolver	**66.8**	45.1	82.8	**46.9**	89.8	**92.9**	**70.7**

Qwen3-Embedding-8B results are taken from the official MTEB results repository: MSMARCO uses its standard `dev` split, and the other five tasks use `test`. The reported macro score is the unweighted mean of these six published scores.

Code retrieval: cross-file RepoEval drives the margin. The 1.5-point macro advantage over RepoNav is driven by cross-file RepoEval (92.3 versus 74.3). Across the other five CodeRAG-Bench tasks, Retrievolver trails RepoNav by 1.8 points on average (Table 13).

Table 13: CodeRAG-Bench nDCG@10 on 872 test queries across six tasks.

Method	HumanEval	MBPP	DS-1000	ODEX	RepoEval	SWE-bench Lite	Macro	Micro
Retrievers								
CodeRankEmbed	**100.0**	**100.0**	13.3	13.3	84.5	10.3	53.6	52.8
CodeSage-small	**100.0**	99.2	14.7	16.2	81.4	46.7	59.7	59.0
BM25	99.7	98.9	10.0	11.2	86.2	56.7	60.4	58.1
Jina-v2-code	**100.0**	99.2	31.8	20.9	84.1	55.5	65.3	63.9
SFR-Mistral	**100.0**	99.6	39.9	29.0	83.1	58.1	68.3	67.5
BGE-Code-v1	**100.0**	**100.0**	46.7	33.8	83.4	66.9	71.8	70.3
Agents / Pipelines								
LocAgent	46.8	43.6	33.1	51.8	8.8	33.8	36.3	35.4
AFlow	**100.0**	**100.0**	32.1	41.2	69.6	13.6	59.4	58.6
ACE	**100.0**	**100.0**	31.6	40.7	71.5	14.3	59.7	59.5
A-RAG	**100.0**	**100.0**	53.6	42.9	74.0	**94.4**	77.5	75.4
RepoNav	**100.0**	**100.0**	**58.1**	**54.5**	74.3	92.8	80.0	77.3
Retrievolver	**100.0**	99.6	55.7	53.2	**92.3**	88.1	**81.5**	**80.5**

Code baselines include the code embedding models CodeRankEmbed, CodeSage, Jina-v2-code, and BGE-Code (Li et al., 2025a), evaluated on CodeRAG-Bench (Wang et al., 2025b), alongside LocAgent (Chen et al., 2025), AFlow (Zhang et al., 2025a), ACE (Zhang et al., 2026), A-RAG (Du et al., 2026), and the public RepoNav release (Chai et al., 2026).

Visual retrieval and complex evidence search: precise top-ranked evidence. On M-BEIR (Table 14), Retrievolver exceeds the locally evaluated multimodal agents, with macro recall of 60.4; the published U-MARVEL+, U-MARVEL, and LamRA results are higher.

The multimodal baselines include MuCo (Gu et al., 2026), Qwen3-VL-Embedding, GME (Zhang et al., 2025b), U-MARVEL (Li et al., 2026c), and LamRA (Liu et al., 2025b).

Table 14: Full M-BEIR comparison in Recall@K. † marks published scores; — indicates unavailable results.

Method	WebQA R@5	Fashion200K R@10	NIGHTS R@5	EDIS R@5	CIRR R@5	Macro
Retrievers						
MuCo-7B†	90.9	33.2	28.6	60.2	56.7	53.9
Qwen3-VL-8B	95.5	21.5	31.4	64.7	56.5	53.9
GME-7B	96.4	32.2	31.6	81.0	55.6	59.4
U-MARVEL†	97.1	33.6	34.2	78.8	60.7	60.9
Agents / Pipelines						
OmniSearch	93.4	6.2	28.6	35.6	20.2	36.8
MMSearch-Engine	94.1	16.6	23.6	56.5	33.5	44.9
Retrievolver	98.0	30.5	32.5	79.1	61.7	60.4
LamRA†	96.7	32.9	33.5	75.8	**63.3**	60.4
U-MARVEL+†	**98.5**	**34.2**	**34.7**	**81.4**	63.2	**62.4**

On MMDocIR (Table 15), Retrievolver achieves the highest R@1 and R@5; Initial has the highest R@3.

The document baselines include BGE-large-en-v1.5 (Xiao et al., 2024), ColPali and ColQwen (Faysse et al., 2025), ZipRerank (Sun et al., 2026), and MM-R5 (Xu et al., 2025). The ColQwen and ColQwen + MM-R5 rows come from Table 1 of Xu et al. (2025) (top-10 reranking), and the ZipRerank and ColQwen + GPT-5-mini rows come from Table 2 of Sun et al. (2026) (top-20 reranking of a separate ColQwen2 run). All four of these rows report micro-averaged page recall. The MMRetHeads row uses the Qwen3-VL-8B selected-head results in Appendix A.1, Table 2 of Li et al. (2026a): micro-averaged page R@1/R@3/R@5 of 64.5/83.6/88.9. These are published results, not a local rerun.

Table 15: Full MMDocIR comparison in page recall. † marks published scores; unmarked rows are local evaluations.

Method	R@1	R@3	R@5
Retrievers			
BGE-large-en-v1.5 †	39.6	59.6	68.5
ColPali†	52.7	75.0	81.0
GME†	54.2	76.0	83.8
ColQwen†	63.5	82.1	86.7
MMRetHeads (Qwen3-VL-8B)†	64.5	83.6	88.9
Agents / Pipelines			
VRAG-RL	57.1	77.2	78.1
ZipRerank†	65.1	84.4	89.1
ColQwen + MM-R5†	67.6	84.0	87.6
DocLens	69.7	87.6	90.1
ColQwen + GPT-5-mini†	70.1	87.8	91.4
Initial Retrievolver	72.8	**90.6**	93.1
Retrievolver	**73.2**	90.4	**93.4**

On BrowseComp-Plus (Table 16), Retrievolver leads in Evidence nDCG@10, placing evidence near the top of the delivered ranking; DR-DCI reaches higher recall at deeper cutoffs, and Retrievolver's Evidence recall is 71.5 at both R@100 and R@1000.

Retrievolver delivers between 1 and 180 documents per question (median 7) across the 830 test queries; only three lists exceed 100. Per-query scoring confirms that ranks beyond 100 add no relevant Evidence or Gold documents, so R@100 and R@1000 coincide.

The BrowseComp-Plus search-agent baselines include OpenResearcher (Li et al., 2026d) and Pi-Serini (Hsu et al., 2026). The published JinaColBERT-v2 and ReasonIR-8B scores are taken from Table 2 of Chen et al. (2026b).

Complete-system gains vary by task. Table 17 compares complete systems, including changes in tools, guidance, and execution settings (Appendix A.3). The primary metric improves on all five benchmarks. MMDocIR gains only 0.3 R@5 points and has a small R@3 decrease (Table 15); the aggregate scores alone do not establish statistical significance. Table 6 separately controls the tool pool.

B.2 Cross-Dataset Transfer

Table 18 specifies the source guidance and target tools. The Searcher adapts tools and, where needed, guidance using public methods and the target corpus and indexes, without target relevance labels. Both are fixed before evaluation; the planner still generates a workflow from each query and its intermediate evidence. These experiments evaluate system reuse, while the unseen ArguAna and NQ tasks in Table 6 compare guidance under a fixed tool pool.

On LitSearch, the added candidate-comparison Skill combines ReAct-style observation–action steps with SIRA-style document judgments; it is loaded for 253 of 597 queries. Locally evaluated agents use GPT-5.5 with high reasoning effort, except A-RAG with medium and SIRA with none. ReAct retrieves with BM25 and Qwen3-Embedding-8B (Zhang et al., 2025c).

Table 16: Full BrowseComp-Plus comparison. † marks published scores; — indicates unavailable results.

	Evidence annotations				Gold annotations			
Method	nDCG @10	R@5	R@100	R@1000	nDCG @10	R@5	R@100	R@1000
Retrievers								
BM25†	1.6	1.2	4.7	13.7	1.7	1.4	6.1	17.3
JinaColBERT-v2†	7.9	5.7	18.1	35.7	6.8	6.6	20.4	39.7
ReasonIR-8B†	16.8	12.2	43.6	73.9	15.5	15.3	49.7	78.9
Qwen3-Embedding-8B†	20.3	14.5	47.7	76.7	19.5	18.5	55.8	83.5
Agents / Pipelines								
Pi-Serini	18.0	12.5	32.1	52.7	13.9	13.0	34.1	57.2
OpenResearcher / 30B-A3B	46.4	36.5	58.7	75.2	40.7	41.3	61.1	77.7
DR-DCI	67.2	55.7	**77.5**	**88.3**	60.6	66.8	**85.5**	**93.0**
Retrievolver	**74.6**	**60.0**	71.5	71.5	**79.0**	**77.3**	84.0	84.0

Table 17: Initial and selected complete systems on identical test queries. BrowseComp-Plus uses 829 paired questions, compared with 830 in the main table.

Benchmark	Metric	Initial	Selected
MTEB	Macro nDCG@10	64.4	70.7
CodeRAG-Bench	Macro nDCG@10	77.8	81.5
M-BEIR	Macro Recall@K	57.5	60.4
MMDocIR	Page R@5	93.1	93.4
BrowseComp-Plus	Evidence nDCG@10	52.0	74.7

Table 18: Source guidance and target retrieval setups after Searcher adaptation.

Target	Source task	Target retrieval setup
LitSearch	MTEB (R2)	Literature corpus indexed for the MTEB tools (BM25, dense retrieval, reranking, and fusion), plus seven target-adapted tools for query alignment, reasoning-oriented dense and late-interaction retrieval, reasoning reranking, and optional language-model evidence checks. The MTEB Policy drops two start-up rules (a mandatory initial dense call and a fixed initial depth), and the Searcher adds one candidate-comparison Workflow Skill.
ViDoRe v2	MMDocIR (R3)	Source policy and workflow skills applied to collection-wide ColQwen2 page indexes, with reading, navigation, fusion, and reranking.
MultiHop-RAG	BrowseComp-Plus / AgentIR	Target text corpus with lexical and dense retrieval; adaptive evidence search, clue coverage, fusion, and reranking workflows.
Flickr30K	M-BEIR (R3)	Image–text workflows adapted to both retrieval directions, using SigLIP 2 indexes and Qwen3-VL reranking.
CoIR	CodeRAG-Bench	Target code corpora with source code-retrieval skills, sparse and dense retrieval, fusion, and reranking.

Literature and document retrieval. On LitSearch, Retrievolver reaches 74.6 nDCG@10 versus 77.1 for ReAct; SIRA leads on Broad R@20 and overall R@20 (Table 19).

LitSearch comparisons include SIRA (Yang et al., 2026b), OpenScholar (Asai et al., 2026), and the dense retrieval and GPT-4o reranking baselines (GTR, E5, Instructor, and GritLM) reported by Ajith et al. (2024).

Table 19: LitSearch retrieval. † marks published scores from Ajith et al. (2024); — indicates unavailable results.

Method	Broad	Specific	All queries			
	R@20	R@5	R@5	R@10	R@20	nDCG @10
Retrievers						
BM25†	39.9	50.0	—	—	—	—
GTR-T5-large†	43.8	39.6	—	—	—	—
E5-large-v2†	55.4	56.2	—	—	—	—
Instructor-XL†	56.5	52.3	—	—	—	—
GritLM-7B†	70.8	74.8	—	—	—	—
Agents / Pipelines						
OpenScholar	57.6	62.8	57.8	63.2	68.6	51.0
BM25 + GPT-4o reranking†	59.9	68.0	—	—	—	—
BM25 + GPT-4o one-hop†	64.8	68.6	—	—	—	—
A-RAG	72.9	80.7	77.5	80.0	80.0	72.9
GritLM + GPT-4o one-hop†	73.2	77.0	—	—	—	—
GritLM + GPT-4o reranking†	75.3	79.2	—	—	—	—
ReAct	78.7	**86.1**	**82.6**	**86.0**	87.3	**77.1**
Retrievolver	80.0	84.2	80.0	84.2	87.0	74.6
SIRA (fixed pipeline)	**83.0**	82.9	79.4	85.7	**89.4**	74.0

On ViDoRe v2, the transferred system leads on all four subsets, reaching 73.9 macro nDCG@5 versus 70.7 for the strongest compared system (Table 20).

ViDoRe comparisons include ColPali (Faysse et al., 2025), NeMo (Osmulski et al., 2026), and published ViDoRe v2 leaderboard entries (Macé et al., 2025): ColQwen2.5, Nemotron ColEmbed, VultronRetrieverPrime, Ops-ColQwen3, and Argus. The published English-subset scores can differ from those reported by Macé et al. (2025).

Evo-Retriever-7B (Li et al., 2026b) is included as a published standalone retriever reference. We use its English-subset scores from the official MTEB results repository, with a four-task macro of 65.6. The 65.2 reported in its paper instead averages three multilingual subsets and the English Human ESG subset.

Table 20: ViDoRe v2 English retrieval in nDCG@5. † marks English-subset scores from the official MTEB results repository.

Method	Biomedical	Economics	ESG	Human ESG	Macro
Retrievers					
ColPali v1.3†	59.9	51.7	54.1	58.5	56.1
ColQwen2.5 v0.2†	63.7	58.4	57.5	66.4	61.5
Evo-Retriever-7B†	66.9	58.2	60.5	77.0	65.6
Nemotron ColEmbed VL 8B v2†	66.8	62.1	62.0	73.2	66.0
VultronRetrieverPrime Qwen3.5 8B†	66.8	66.0	66.4	76.8	69.0
Ops-ColQwen3-4B†	67.2	67.7	66.9	78.6	70.1
Argus-ColQwen3.5-9B v0†	69.4	68.3	66.0	79.1	70.7
Agents / Pipelines					
DocLens	19.1	37.1	23.6	41.1	30.2
VRAG	55.4	46.8	46.7	49.1	49.5
NeMo	69.3	68.9	63.5	81.1	70.7
Retrievolver	**69.9**	**71.8**	**70.2**	**83.6**	**73.9**

Multi-hop retrieval: strong early-rank evidence. Table 21 compares chunk retrieval with published results from Tang & Yang (2024). Their best configuration, voyage-02 + bge-reranker-large,

scores 66.25 Hits@4, 74.67 Hits@10, 58.60 MRR@10, and 47.95 MAP@10. Retrievolver leads on the first three metrics, while published reranking pipelines reach up to 48.0 MAP@10, compared with 43.2 for Retrievolver. The gain is strongest in retrieving an early relevant chunk, rather than precision across all relevant ranks.

Table 21: MultiHop-RAG chunk retrieval. † marks published scores from Tang & Yang (2024); bold and underline compare only the listed rows.

Method	Hits@4	Hits@10	MRR@10	MAP@10
Retrievers				
E5-base-v2†	23.3	35.6	18.4	11.6
Instructor-large†	42.3	57.2	34.6	26.5
BGE-large-en-v1.5†	52.2	67.2	43.0	34.2
Agents / Pipelines				
BGE-large-en-v1.5 + bge-reranker-large†	<u>63.6</u>	<u>71.8</u>	<u>56.3</u>	**47.6**
Retrievolver	**92.8**	**97.6**	**78.1**	<u>43.2</u>

At the article level (Table 22), Retrievolver leads in nDCG@10, while DR-DCI and OpenResearcher achieve slightly higher R@20 (by 0.3 and 1.0 points). The transferred workflows thus excel at placing relevant evidence at the top of the ranking.

Table 22: MultiHop-RAG article retrieval. ‡ marks rankings reconstructed from agent execution trajectories.

Method	R@20	nDCG@10
DR-DCI‡	<u>91.0</u>	<u>87.4</u>
OpenResearcher‡	**91.7**	80.2
Retrievolver	90.7	**88.1**

Image–text retrieval: competitive with agent pipelines. On Flickr30K, Retrievolver exceeds OmniSearch and MMSearch-Engine on all six metrics and ReAct on all three text-to-image metrics. ReAct remains stronger on image-to-text R@1, and the published BLIP-2 models match or exceed Retrievolver throughout (Table 23).

Table 23: Flickr30K retrieval. I and T denote image and text; † marks published scores and — unavailable results.

Method	I→T R@1	I→T R@5	I→T R@10	T→I R@1	T→I R@5	T→I R@10
Retrievers						
CLIP†	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN†	88.6	98.7	99.7	75.7	93.8	96.8
ALBEF†	94.1	99.5	99.7	82.8	96.3	98.1
SigLIP 2 g/16 384†	95.4	—	—	86.0	—	—
BLIP†	96.7	**100.0**	**100.0**	86.7	97.3	<u>98.7</u>
BLIP-2 ViT-L†	96.9	**100.0**	**100.0**	<u>88.6</u>	<u>97.6</u>	**98.9**
BLIP-2 ViT-g†	<u>97.6</u>	**100.0**	**100.0**	**89.7**	**98.1**	**98.9**
Agents / Pipelines						
OmniSearch	96.1	98.7	98.7	79.4	79.6	79.6
MMSearch-Engine	94.2	96.6	96.8	86.7	93.8	94.8
ReAct	**97.7**	99.8	<u>99.8</u>	87.7	96.3	98.0
Retrievolver	96.7	<u>99.9</u>	**100.0**	88.2	96.7	98.2

The image–text references are SigLIP 2 (Tschannen et al., 2025) and the zero-shot Flickr30K results of CLIP, ALIGN, ALBEF, BLIP, and BLIP-2 reported by Li et al. (2023). ALBEF, BLIP, and BLIP-2 use COCO-fine-tuned models with image–text matching reranking.

Code retrieval: transfer outperforms all compared agents. On the five evaluated CoIR datasets, the transferred Retrievolver exceeds every compared agent and both CodeXEmbed-2B variants in macro-average nDCG@10 (Table 24). Retrievolver exceeds A-RAG on all five tasks and RepoNav on four, and reaches a macro-average nDCG@10 of 70.6, compared with 56.9 and 63.1 for these two agents. It also exceeds CodeXEmbed-2B (68.1) and its general-training variant SFR-Embedding-Code-2B_R (64.3). BGE-Code-v1 and CodeXEmbed-7B attain higher macro scores of 76.1 and 71.8, respectively.

SFR-Embedding-Code-2B_R is the released CodeXEmbed-2B checkpoint. Its published CoIR scores match the general-training row in Table 1 of Liu et al. (2025a), which excludes CoIR in-domain training data.

Table 24: CoIR nDCG@10 on five datasets. † marks published results under their source evaluation settings.

Method	CoSQA	CodeTrans DL	CodeTrans Contest	StackOverflow QA	APPS	Macro
Retrievers						
SFR-Embedding-Code-2B_R†	36.3	33.2	86.6	90.5	75.0	64.3
CodeXEmbed-2B†	40.5	38.6	90.3	94.5	76.9	68.1
CodeXEmbed-7B†	<u>42.5</u>	<u>40.5</u>	<u>94.5</u>	**96.3**	85.4	<u>71.8</u>
BGE-Code-v1†	**46.7**	**46.1**	94.4	95.4	**98.1**	**76.1**
Agents / Pipelines						
A-RAG	17.3	27.0	89.8	93.4	57.1	56.9
RepoNav	27.2	14.2	**95.1**	95.0	84.0	63.1
Retrievolver	36.9	31.4	94.0	<u>96.1</u>	<u>94.7</u>	70.6

Reusing retrieval procedures across tasks. MMDocIR and ViDoRe v2 both require page-level evidence retrieval; BrowseComp-Plus and MultiHop-RAG both require gathering evidence across documents. The transferred guidance supplies procedures for evidence inspection, continued search, and ranking adoption, which the planner applies to each query and its observations to generate subsequent workflow code. Table 6 provides a complementary comparison of guidance reuse on the unseen ArguAna and NQ tasks with the tool pool held fixed.

C Downstream Agent Integration

All five hosts use the frozen BrowseComp-Plus configuration selected on AgentIR development queries (R4). The comparisons measure the value of this retrieval module within each host; hosts retain their own answer generation.

Downstream answer evaluation. Table 7 compares each agent before and after Retrievolver integration on the same questions. Retrievolver improves every host: all five paired 95% confidence intervals in Table 26 lie above zero. The table reports sample sizes, answer transitions, and intervals, with accuracy and its change computed on the same paired questions within each row; intervals come from paired percentile bootstrap resampling of questions with 5,000 replicates and random seed 20260919. The gain in percentage points is $\Delta = 100(\text{Corrected} - \text{Regressed})/N$, computed before rounding.

All five hosts use GPT-5.5 as the answering backbone but keep agent-specific execution settings. DR-DCI and ReAct use high reasoning effort, and the ACE generator uses none. MiroFlow keeps its own agent configuration, and its final-output completion step, when required, also uses no reasoning effort. GPT-5.4 grades all answers against BrowseComp-Plus references with low reasoning effort, a 4,096-token output limit, and the benchmark correctness parser. Both conditions use the same grading protocol, and arm labels are omitted from grader requests.

Integration interface and evidence handoff. The common Retrievolver interface takes the original question, the host's current search query, and the requested retrieval depth. It runs a workflow

with frozen guidance and sets `answer_required=false`. The return value contains ranked document IDs and scores; each host adapter retrieves the corresponding public corpus content and presents it through the interfaces in Table 25. The adapters do not forward Retrievolver-generated candidate answers, final answers, or summaries. Each host retains answer generation and its own reading and reasoning loop where applicable.

New searches invoke Retrievolver online, with identical requests eligible for caching. ReAct can open an already returned document by ID or title. A-RAG keeps its original tools and optionally calls the added `rh_search`.

Table 25: BrowseComp-Plus host interfaces. Retrieval depths are requested maxima; Retrievolver may return fewer documents.

Host	Base retrieval	Retrievolver integration and evidence access
DR-DCI	Qwen3-Embedding-8B; `pull` requests 300–600 documents	Retrievolver replaces the `pull` backend; ranked corpus files remain readable in the host workspace.
MiroFlow	BM25; search defaults to 10 documents	Retrievolver replaces search; the host receives titles, URLs, and 1,000-character text previews, then can read full documents.
ReAct	BM25; Search requests 5 documents	Retrievolver replaces Search; the first result shows five sentences, and Lookup accesses its stored text. Other hits expose IDs and titles.
ACE generator	Qwen3-Embedding-8B; initial top-10 retrieval	Retrievolver replaces the initial retrieval; retrieved documents enter the original Generator.
A-RAG	Keyword search, Qwen3-Embedding-8B semantic search, and chunk reading	Optional `rh_search` returns document and native chunk IDs; `read_chunk` obtains text. Original tools remain available.

The ACE generator starts each query with its authors' initial prompt and an empty playbook, without experience adaptation. Its Base condition receives the top ten full documents from Qwen3-Embedding-8B; integration replaces this evidence with Retrievolver results while keeping the generator fixed.

When the supplied documents exceed the generator's context window, the host reads them in batches and passes extracted factual notes to the same Generator, without new retrieval requests.

A-RAG configuration. Integration raises A-RAG from 671 to 727 correct answers on the 829 paired questions (80.9% versus 87.7%). Both conditions use the same document-level Qwen3-Embedding-8B index, with GPT-5.5 (medium reasoning effort, 16,384-token output limit) answering and GPT-5.4 grading; the integrated condition additionally exposes the frozen Retrievolver retrieval tool.

Table 26: Paired BrowseComp-Plus results. Corrected and Regressed count errors fixed and introduced by integration. The 95% confidence intervals for Δ are in percentage points.

		Correct answers				
Agent	N	Host	+ Retrievolver	Corrected	Regressed	95% CI for Δ
DR-DCI	824	762	784	34	12	[1.1, 4.3]
MiroFlow	828	493	534	173	132	[1.0, 9.1]
A-RAG	829	671	727	79	23	[4.3, 9.1]
ReAct	817	311	579	310	42	[28.9, 36.7]
ACE generator (no adaptation)	828	364	742	384	6	[42.2, 49.2]

Why gains differ across hosts. The benefit depends on baseline errors and evidence use. DR-DCI starts with 62 errors and gains 22 correct answers; MiroFlow corrects 173 answers but also regresses on 132, leaving a net gain of 41 (Table 26).

ACE: more relevant evidence from fewer documents. Gold coverage is the percentage of questions whose supplied evidence contains at least one Gold document. For ACE, Gold-document coverage increases from 42.0% to 92.2% as the mean delivered document count falls from 10.0 to 5.8 (Table 27). Of 384 corrected questions, 361 gain a Gold-document hit that was previously missing. Gold labels are used only for this post-execution analysis.

Table 27: Evidence coverage and answer availability on the paired BrowseComp-Plus questions in Table 7.

Measure	Host	+ Retrievolver
ACE generator without adaptation ($N = 828$)		
Mean delivered documents	10.0	5.8
Gold coverage (%)	42.0	92.2
Answer accuracy (%)	44.0	89.6
ReAct ($N = 817$)		
Questions with no answer within the agent's loop	482	217
Answer accuracy (%)	38.1	70.9

ReAct: completing the retrieval-to-answer process. With Retrievolver, ReAct more often produces an answer within its own action loop. This measure excludes the host's final-output completion step, whereas accuracy uses the final predictions. Among the 817 paired questions, 482 base executions and 217 Retrievolver executions produced no answer within the action loop.

D Dynamic Workflows and Guidance Evolution

D.1 Choosing and Extending a Workflow

Figure 5 illustrates query-specific execution across three task families. Each task has its own Policy, Skills, and tools; guidance remains fixed within the query, while observations determine subsequent calls.

In NFCorpus, inspection triggers further retrieval and reranking, but the planner ultimately returns the retained first ranking. In CIRR, "Put a rainbow behind the boat" requires both a reference image and a modification instruction; the planner passes both to U-MARVEL and compares the result with cached GME and Qwen rankings. In the held-out BrowseComp-Plus case, discovering Ken Walibora triggers a career-focused query. This last case demonstrates test-time execution under frozen guidance, not a guidance update using the test question.

D.2 Online Replanning from Intermediate Evidence

Figure 6 expands the NFCorpus case into four executed Python segments. Inspection reveals saffron–fluoxetine evidence and studies of other conditions. The planner first tries complementary searches and fusion, then a focused query, and finally biomedical reranking. None yields a better ranking in its assessment, so it delivers the original dense result. Each `replan` changes the next program using observations already available; Policy and Skills remain fixed.

D.3 Workflow Skill Revisions

Figure 7 shows how general initial advice becomes conditional guidance. Further retrieval targets an observed missing entity, relation, constraint, or answer type. A supplemental ranking is adopted only when inspected evidence improves the retained result. Stopping depends on whether the evidence satisfies the query. These rules guide the planner's decisions without prescribing

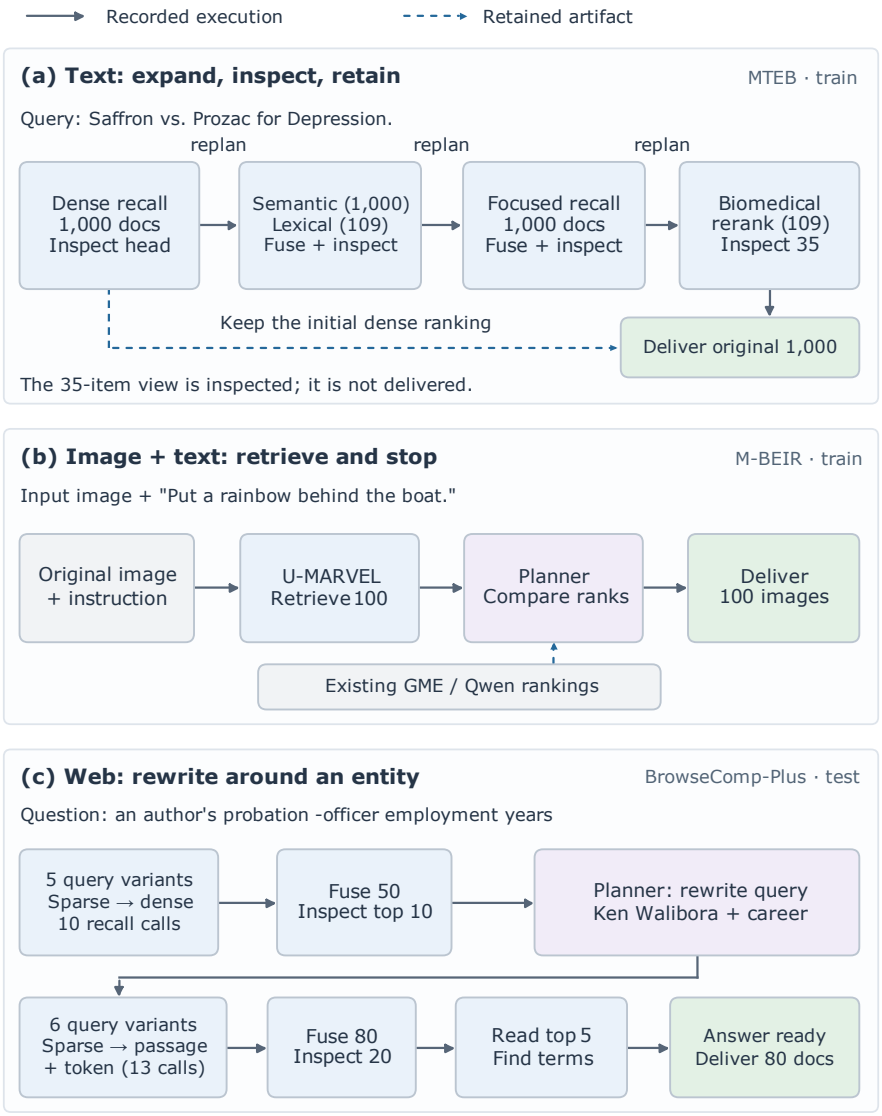

Figure 5: Execution trajectories for NFCorpus training, CIRR training, and BrowseComp-Plus testing. Solid arrows show execution order, dashed connections show reads of retained results, and gray boxes denote cached rankings.

one fixed program for every query. The excerpts come from `text-corpus-evidence` and `dense-method-comparison`.

D.4 Offline Revision of a Retrieval Decision

In the ODEX training case in Figure 8, the original ranking places `time.strftime` first among 100 documents. The planner selects eight documents in solution-step order, demoting it to rank seven and reducing nDCG@10 from 1.000 to 0.333. The revised Policy requires query- and content-based justification before changing a ranking. Under this guidance, the next execution directly delivers the original 100-document ranking and scores 1.000. This example links a guidance edit to a concrete change in generated code and delivery.

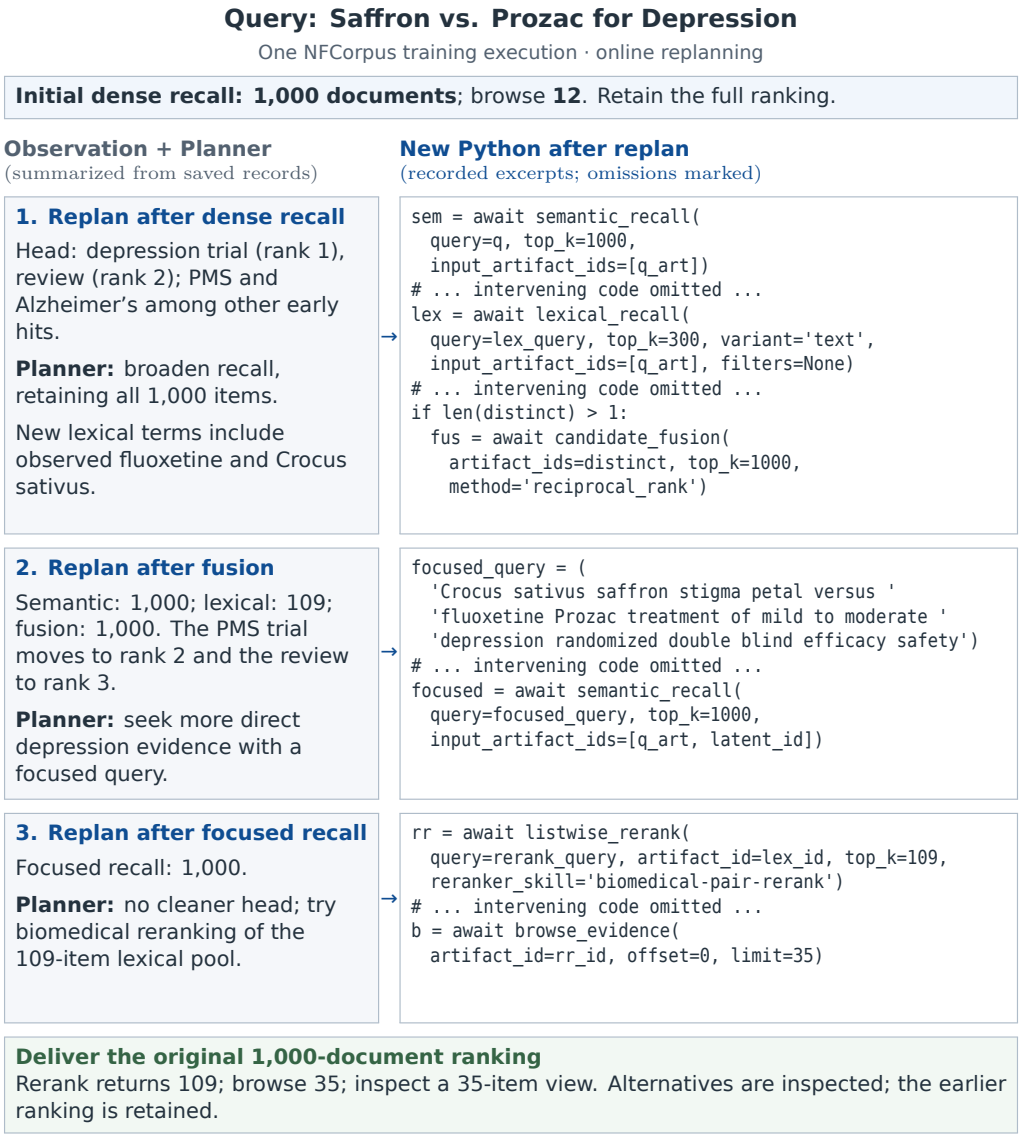

Figure 6: Online replanning for an NFCorpus training query. Code excerpts preserve executed calls; observations and planner assessments are summarized.

The left panel shows generated Python that selects and reorders a subset; the right panel shows a `next_action` call in JSON that returns the original ranking. Ellipses omit Policy text and shorten the result identifier; unrelated JSON fields are omitted.

> **Online continuation (M-BEIR Workflow Skill).** "Call `complete_segment` to retain executed work, then use `next_action` to continue, rewrite, replan or deliver. Make observations needed for subsequent decisions visible in tool output. Use the live tool schema for parameters."
>
> **Workflow revision (MTEB Evolver).** "Diagnose actual tool arguments, returned documents, retained/displaced results and costs from the full ordinary-Planner trajectory." "Synthesize ONE complete candidate from ALL batch analyses. Keep useful incumbent rules, reconcile conflicts and prior failed trials."

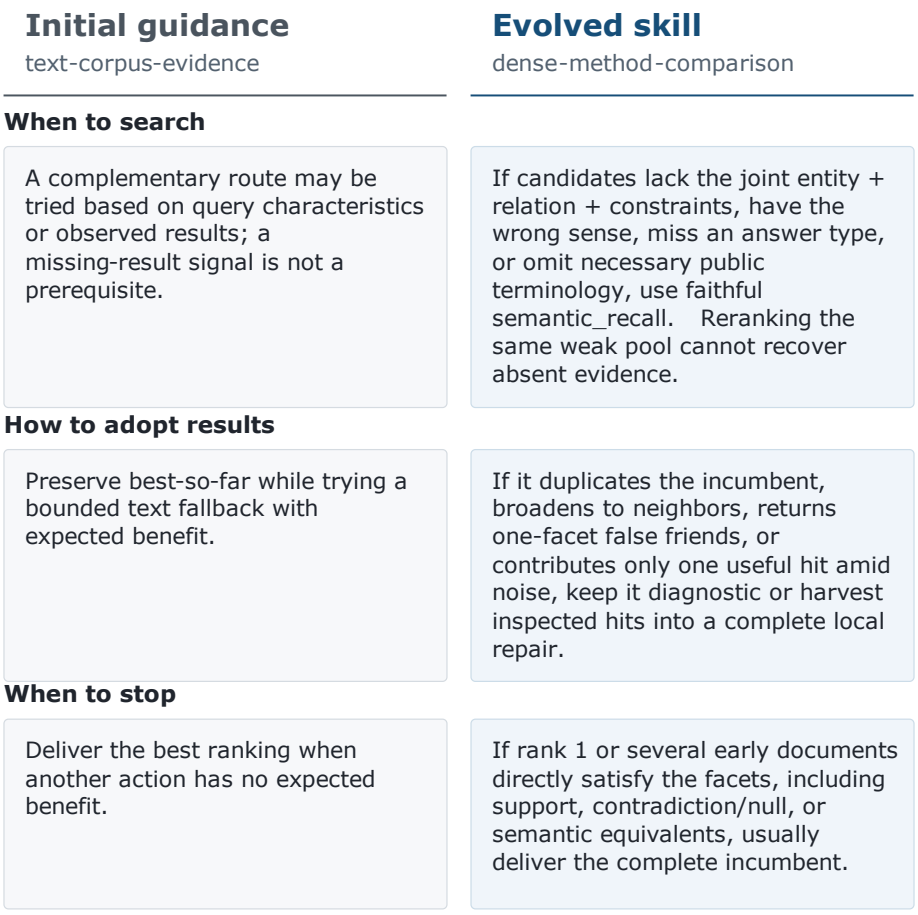

Figure 7: Initial and evolved MTEB Workflow Skills for further search, result adoption, and stopping. Excerpts are verbatim.

Policy revision (M-BEIR Evolver). The evolver is asked to state each rule in four parts:

"Visible condition: public request features or evidence actually observable when the decision is made."

"Action to try: the allowed Workflow choice, composition, continuation, adoption, or terminal action."

"Expected resolution: the uncertainty or evidence/ordering problem the action could resolve, not a promised score gain."

"Counterexample: a public condition where this rule does not apply or another action is preferable; preserve supplied successful opposite-action cases."

E Computation and Limitations

Hardware and deployment. Local retrieval, reranking, and baseline-agent services run on four NVIDIA A100 80 GB GPUs, eight A100 40 GB GPUs, and four RTX A6000 GPUs. The M-BEIR deployment, for example, uses the four A100 80 GB GPUs to host U-MARVEL retrieval and Qwen3-VL/LamRA reranking replicas, with CPU-based index serving. Planning and textual evolution use externally hosted model APIs, whose compute is separate from these local GPU resources.

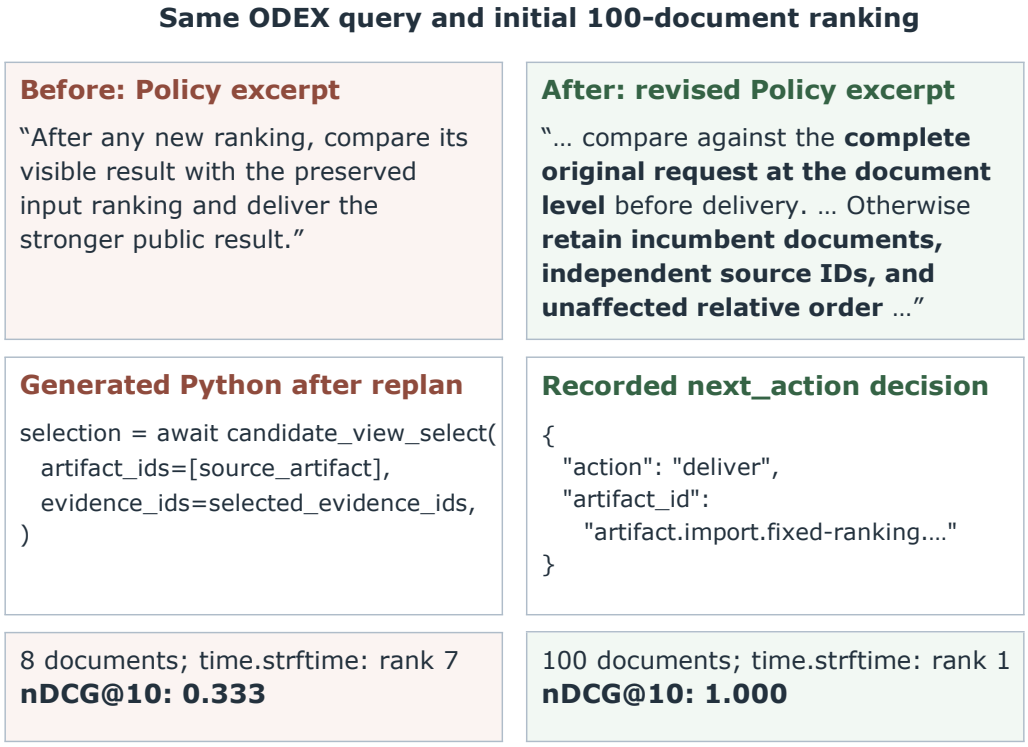

Figure 8: Policy revision and execution changes on the same ODEX training query. Bold marks the added condition; scores range from 0 to 1.

E.1 Offline and Online Computation

Table 28 reports approximate hosted-model usage for query execution and guidance revision. It excludes Stage 1 discovery and integration, final tests, grading, and local retrieval or reranking. Input counts include cached context and repeated candidate executions: about 5.13 billion tokens in total, including the estimated MMDocIR entries described below. Online use freezes the selected guidance.

Each row counts only computation not already charged in earlier rows. Query execution includes training and development unless stated otherwise. Rounds marked *answer only* revise answer generation while reusing the previous retrieval results.

MMDocIR Initial–R4 calls are estimated from R5 training tokens per call, rounded to tens; R1/R4 token splits use the R2–R3 output fraction, and R1–R4 revision budgets use R5 revision as a proxy. Other entries report measured incremental usage; previously charged computation is counted only once.

A local FiQA usage measurement. On the first 50 FiQA test queries, the expanded-pool Initial and Evolved runs issue 351 and 214 planner requests, or 7.0 and 4.3 per query; neither run invokes answer finalization.

Table 28: Approximate hosted-model usage during offline refinement. Token counts are in millions; – denotes no computation.

	Query execution			Reflection and revision		
Stage	Calls	Input	Output	Calls	Input	Output
MTEB						

Continued on next page

Table 28 continued

	Query execution			Reflection and revision		
Stage	Calls	Input	Output	Calls	Input	Output
Initial	5,841	165.96	4.60	–	–	–
R1	6,042	173.41	4.24	8,878	456.44	17.87
R2	5,804	159.54	4.04	8,675	423.11	16.21
R3	5,578	155.45	3.82	4	2.21	0.02
R4	5,536	150.28	3.77	1	0.69	0.01
CodeRAG-Bench						
Initial	2,035	95.55	1.71	–	–	–
R1	2,044	106.88	1.64	245	45.58	0.87
R2	2,155	108.04	1.72	245	46.85	0.89
R3	2,164	127.67	1.76	238	45.84	0.82
R4	2,273	135.20	1.82	246	46.83	0.92
M-BEIR						
Initial	5,303	158.51	2.73	–	–	–
R1	5,338	158.04	2.82	462	74.05	1.22
R2	5,540	176.14	2.89	458	70.62	1.54
R3	5,610	187.22	2.97	445	82.73	1.17
R4	6,253	214.02	2.92	476	96.36	1.31
R5	6,182	210.63	3.03	399	122.37	1.13
MMDocIR (Initial–R4: train; R5: train + dev)						
Initial	1,540	55.76	1.29	–	–	–
R1	2,270	83.02	1.16	1	0.29	0.01
R2	2,200	80.31	1.08	1	0.29	0.01
R3	2,240	81.78	1.18	1	0.29	0.01
R4	2,330	85.02	1.19	1	0.29	0.01
R5	3,585	130.65	1.81	1	0.29	0.01
BrowseComp-Plus						
Initial	2,645	113.17	2.64	–	–	–
R1	2,589	141.07	2.83	22	3.01	0.14
R2 (answer only)	251	2.26	0.05	23	6.05	0.13
R3	2,714	155.98	2.87	22	8.91	0.15
R4	2,759	166.40	2.84	23	9.10	0.16
R5 (answer only)	249	2.01	0.05	–	–	–
R6 (answer only)	249	2.22	0.05	22	5.73	0.12

E.2 LIMITATIONS

Cross-dataset transfer is uneven. Retrievolver leads on ViDoRe v2 and several MultiHop-RAG metrics, but trails the strongest compared systems on LitSearch, Flickr30K, and CoIR.

Offline evolution is computationally expensive because it repeatedly executes queries and revises guidance (Appendix E.1). Future work will focus on reducing this cost while preserving retrieval quality.