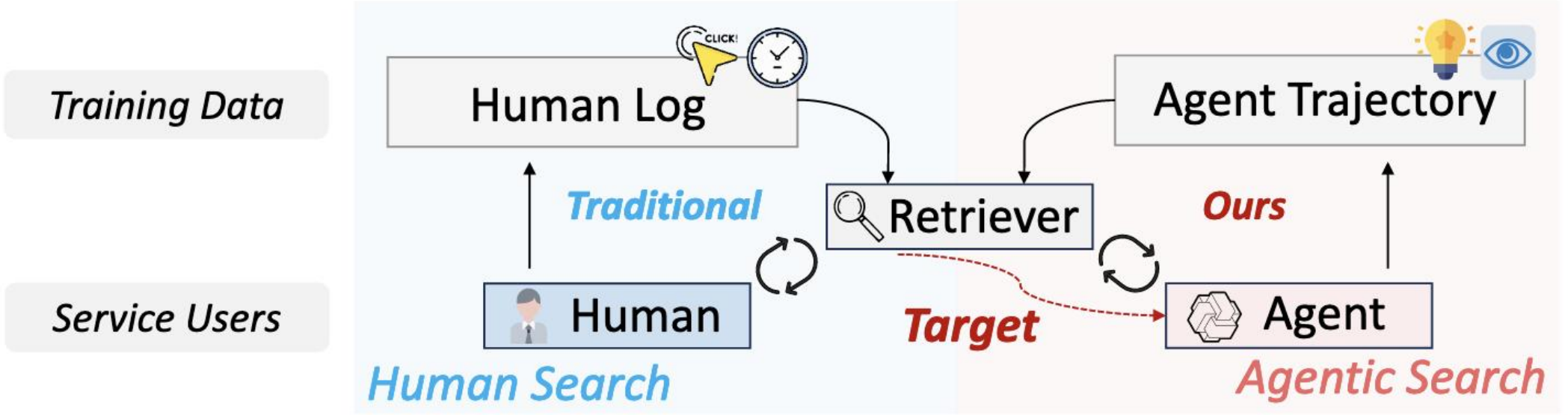
Learning to Retrieve from Agent Trajectories

Yuqi Zhou, Sunhao Dai, Changle Qu, Liang Pang,
Jun Xu and Ji-Rong Wen

SIGIR 2026

Information retrieval systems have traditionally been **designed and trained for human users**, with learning-to-rank methods relying heavily on large-scale human interaction logs such as clicks and dwell time.

Retrieval is **increasingly consumed by agents** rather than human beings, and is embedded as a core component within multi-turn reasoning and action loops.



- Search agents typically rely on **general-purpose retrieval models**, which are overwhelmingly trained from human interaction logs and implicitly encode human-centric assumptions.

- Agent queries are not issued to satisfy immediate informational needs, but to advance intermediate reasoning objectives during problem-solving. (utility?

- Retrieval models **should be trained directly from agent interaction data**, rather than repurposed from human-centric search.

Given an initial user query q, a Deep Research Agent follows a **ReAct-style interaction pattern**, repeatedly alternating between thinking and acting.

Trajectory Example

User Query: Which station merged with Immaculate Heart Radio in 2017?

Agent Trajectory

[Think]: No merger info yet. Let's search for the 2017 merger details.

[Search]: `Search["Immaculate Heart" "merged" "2017"]`

[Information]: Found 10 results:
Doc [3948]: Immaculate Heart Radio ...
Doc [2295]: corporate name Relevant Radio, Inc...
...

[Think]: Doc [2295] mentioned Relevant Radio. It might contain the specific merger date. Let's read it.

[Browse]: `Browse[DocID: 2295]`

[Information]: ... Relevant Radio announced it would merge...

... Intermediate turns omitted ...

[Think]: I have found the answer...

[Answer]: The station is WMJR (Nicholasville, KY).

At each turn t, the agent maintains an internal reasoning state r_t that guides subsequent actions a_t, incorporates new observed information o_t.

$$\mathcal{T} = \{(r_t, a_t, o_t)\}_{t=1}^{T}$$

Trajectory Example

User Query: Which station merged with Immaculate Heart Radio in 2017?

Agent Trajectory

[Think]: No merger info yet. Let's search for the 2017 merger details.

[Search]: `Search["Immaculate Heart" "merged" "2017"]`

[Information]: Found 10 results:
Doc [3948]: Immaculate Heart Radio ...
Doc [2295]: corporate name Relevant Radio, Inc...
...

[Think]: Doc [2295] mentioned Relevant Radio. It might contain the specific merger date. Let's read it.

[Browse]: `Browse[DocID: 2295]`

[Information]: ... Relevant Radio announced it would merge...

... Intermediate turns omitted ...

[Think]: I have found the answer...

[Answer]: The station is WMJR (Nicholasville, KY).

- [Think]: At the beginning of each turn t, the agent first produces an explicit reasoning state r_t.
- Conditioned on this reasoning state, the agent selects an action a_t to interact with the external environment. Two typical actions:
  - [Search]
  - [Browse]
- [Answer]: The agent determines that sufficient information has been gathered to answer the original query.

$$\mathcal{T} = \{(r_t, a_t, o_t)\}_{t=1}^{T}$$

- [Think]: At the beginning of each turn t, the agent first produces an explicit reasoning state r_t.

- Conditioned on this reasoning state, the agent selects an action a_t to interact with the external environment. Two typical actions:
  - [Search]: The agent generates an search query q_t that targets a specific information gap identified in r_t. Retriever returns top-K documents, typically a snippet list (e.g., titles and brief summaries) as observation o_t.

  - [Browse]: The agent selects one document from the retrieved candidates and requests to read it in full. Retriever returns the complete content of the selected document as observation o_t for this turn.

- [Answer]: The agent determines that sufficient information has been gathered to answer the original query.

Analysis of Agent Trajectories

- Failed trajectories exhibit a substantially **lower ratio between [Browse] and [Search] actions (B/S).**
- Successful trajectories show **fewer repetitive search actions** and a markedly **higher frequency of browsing behaviors**.

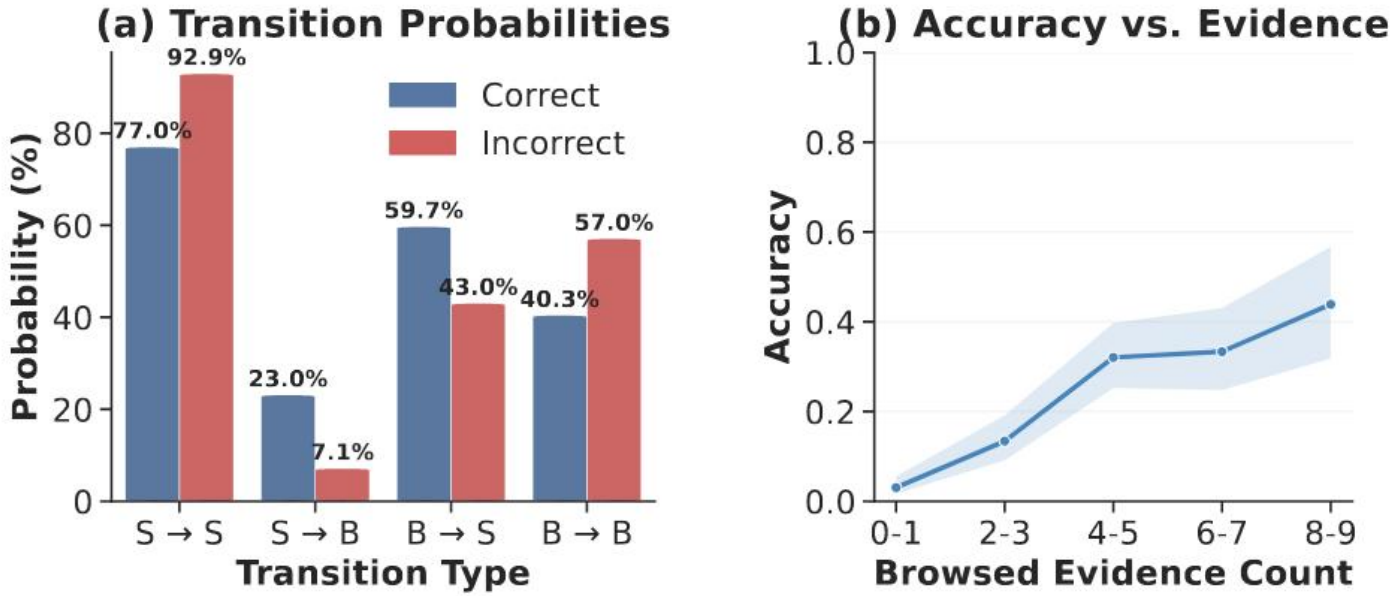
Table 1 | Statistics of generated trajectories across different retrievers. For each retriever, we report the number of completed trajectories (N) and the average numbers of [Search] actions (Avg. S), [Browse] actions (Avg. B), their ratio (B/S), and the total execution steps (Avg. T), categorized into correct, incorrect, and all completed trajectories.

Retriever	Correct					Incorrect					Total (Complete)				
	N	Avg. S	Avg. B	B/S	Avg. T	N	Avg. S	Avg. B	B/S	Avg. T	N	Avg. S	Avg. B	B/S	Avg. T
BM25	7,674	9.15	2.96	0.32	12.11	1,872	29.15	5.97	0.20	35.11	9,546	13.07	3.55	0.27	16.63
Qwen3-Embedding-0.6B	5,913	12.81	3.68	0.29	16.49	2,062	38.95	7.17	0.18	46.12	7,975	19.57	4.58	0.23	24.15
Qwen3-Embedding-4B	6,354	13.24	4.11	0.31	17.34	2,121	36.13	7.47	0.21	43.60	8,475	18.97	4.95	0.26	23.91
Qwen3-Embedding-8B	6,541	11.86	3.69	0.31	15.55	2,082	34.47	7.20	0.21	41.67	8,623	17.32	4.54	0.26	21.85
Total	**26,482**	11.77	3.61	0.31	15.38	**8,137**	34.68	6.95	0.20	41.63	**34,619**	17.25	4.41	0.26	21.66

Tongyi-DeepResearch-30B-A3B; InfoSeekQA dataset top 10k quries; Wiki-25-Dump as the corpus; top-10 candidate documents; the first 64 tokens of the document as document snippet; maximum 100 interaction rounds.

Analysis of Agent Trajectories

- Successful trajectories show a significantly higher probability of **transitioning from [Search] (S) to [Browse] (B)**.
- Task success increases monotonically with the number of browsed evidence documents.

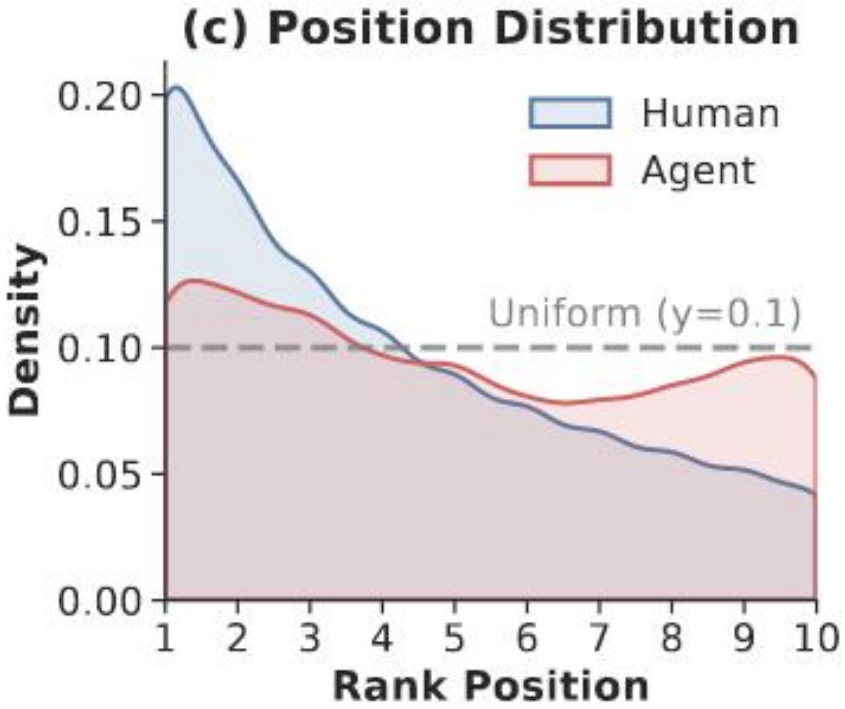


treating browsed documents as primary candidates for positive supervision

Analysis of Agent Trajectories

- In human-centric click logs, negative signals are notoriously ambiguous due to well-known position bias or exposure bias.

- The agent's **browsing behavior is not sharply concentrated at top ranks**.

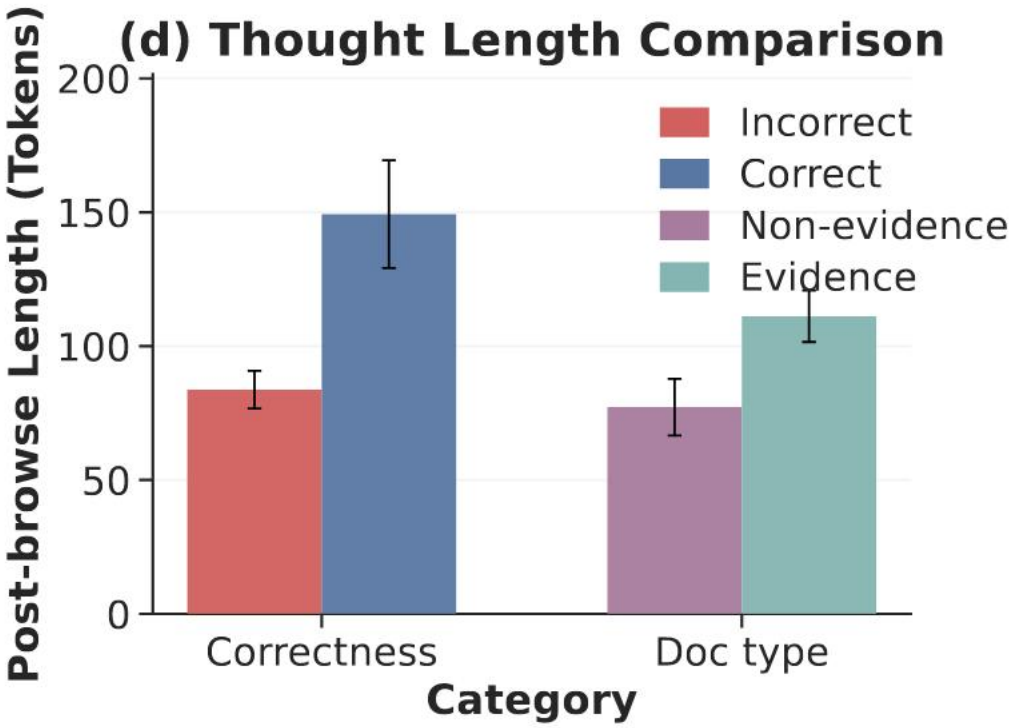


all unbrowsed documents within a retrieved candidate set can be treated as reliable negatives

Analysis of Agent Trajectories

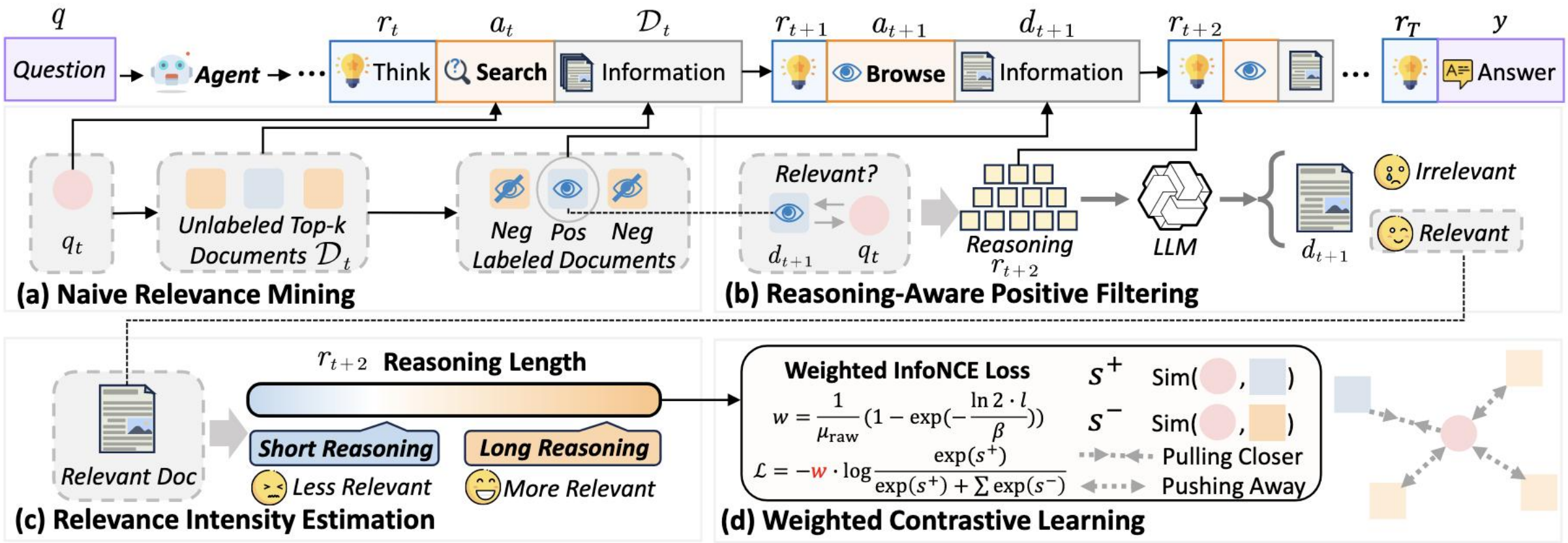
- Trajectories that ultimately produce correct answers are associated with **significantly longer reasoning following browsing actions than** those that lead to incorrect answers.

- Documents that contain ground-truth evidence are followed by **markedly longer reasoning traces** than non-evidence documents.



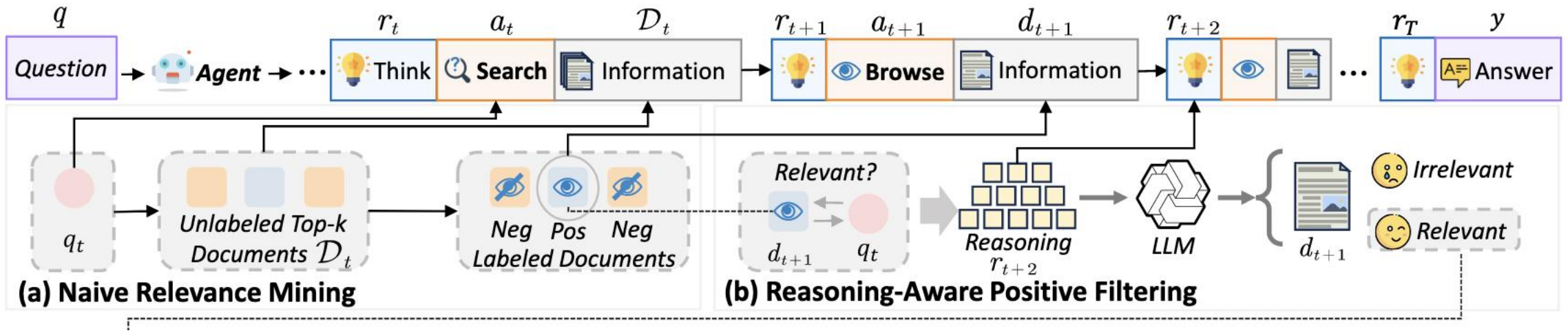
post-browse reasoning traces provide a reliable signal of document utility

Learning to Retrieve from Trajectories



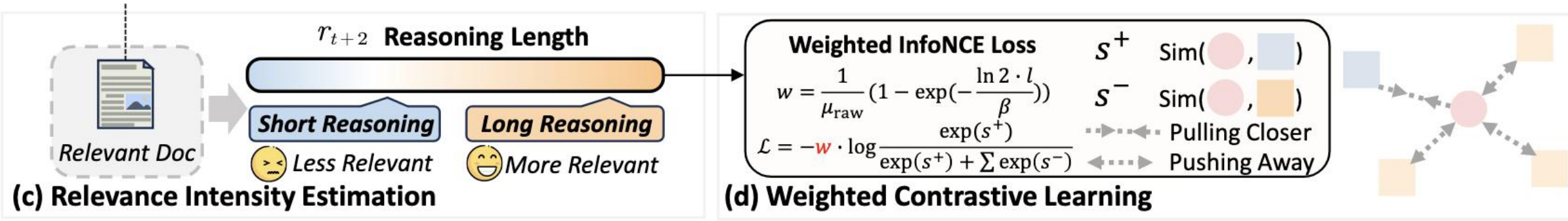
Learning to Retrieve from Trajectories

- If the agent performs a [Browse] action on one of the candidates at the next turn, we view the browsed document as a **naive positive sample**.
- We treat all other candidates in the same retrieved set that are not browsed as **naive negatives**.
- For each browsed document, we collect the agent's immediate post-browse reasoning trace and apply an **LLM-based verifier** to determine **whether the reasoning explicitly uses the document content** to support progress on the task.



Learning to Retrieve from Trajectories

- Dwell time has long been recognized as an effective proxy for relevance intensity.
- Longer reasoning chains following a browsing action are strongly correlated with **higher document usefulness** for the agent's subsequent planning and problem solving.



In the time-aware click model, the marginal gain at dwell length x follows an exponentially decaying function.

$$g(x) = \exp\left(-\frac{\ln 2}{\beta}x\right),$$

Our analysis of post-browse reasoning lengths shows a similar exponential decay in agent trajectories. Cumulative relevance utility:

$$u(l) = \int_0^l g(x)\,dx = \frac{\beta}{\ln 2}\left(1 - \exp\left(-\frac{\ln 2}{\beta}l\right)\right),$$

Table 2 | Results on task-optimized search agents and generalist agentic foundation models with different retrievers on in-domain (ID) InfoSeek-Eval and out-of-domain (OOD) BrowseComp-Plus benchmarks. Qwen3-Emb denotes Qwen3-Embedding-0.6B and E5-Large denotes Multilingual-E5-Large-Instruct. Metrics include Success Rate (SR), Recall, and Average Step Count (Avg. Steps). Best results within each agent backbone are highlighted in **bold**.

Agent Backbone	Retriever	InfoSeek-Eval (ID)		BrowseComp-Plus (OOD)		
		SR (↑)	Avg. Steps (↓)	SR (↑)	Recall (↑)	Avg. Steps (↓)
I. Task-Optimized Search Agents						
AgentCPM-Explore (4B)	Qwen3-Emb	40.3	38.0	13.5	23.2	40.7
	+ LRAT (Ours)	**55.7** (+38.2%)	**34.4**	**15.8** (+17.0%)	**32.0** (+37.9%)	**40.4**
	E5-Large	47.3	38.9	15.9	26.5	40.7
	+ LRAT (Ours)	**49.7** (+5.1%)	**35.5**	**15.9** (+0.0%)	**32.1** (+21.1%)	**40.1**
WebExplore (8B)	Qwen3-Emb	52.0	24.1	21.0	47.7	40.7
	+ LRAT (Ours)	**68.7** (+32.1%)	**19.0**	**27.2** (+29.5%)	**55.9** (+17.2%)	**38.7**
	E5-Large	60.0	23.8	25.4	50.4	40.1
	+ LRAT (Ours)	**63.3** (+5.5%)	**20.2**	**29.0** (+14.2%)	**56.1** (+11.3%)	**39.1**
Tongyi-DeepResearch (30B)	Qwen3-Emb	52.7	26.7	17.8	49.2	42.9
	+ LRAT (Ours)	**68.0** (+29.0%)	**20.7**	**23.7** (+33.1%)	**60.7** (+23.4%)	**41.0**
	E5-Large	56.7	25.1	20.7	54.8	42.4
	+ LRAT (Ours)	**68.0** (+19.9%)	**21.5**	**23.9** (+15.5%)	**61.8** (+12.8%)	**41.4**
II. Generalist Agentic Foundation Models						
GPT-OSS (120B)	Qwen3-Emb	40.0	34.9	9.0	43.7	45.4
	+ LRAT (Ours)	**47.0** (+17.5%)	**30.5**	**12.1** (+34.4%)	**56.4** (+29.1%)	**45.2**
	E5-Large	41.7	33.9	10.8	50.1	44.8
	+ LRAT (Ours)	**50.7** (+21.6%)	**29.7**	**13.1** (+21.3%)	**56.0** (+11.8%)	**44.6**
MiniMax-M2.1 (229B)	Qwen3-Emb	58.7	21.4	38.2	57.2	30.8
	+ LRAT (Ours)	**78.3** (+33.4%)	**14.7**	**48.3** (+26.4%)	**69.2** (+21.0%)	**28.3**
	E5-Large	64.0	18.9	46.4	64.9	29.1
	+ LRAT (Ours)	**75.0** (+17.2%)	**14.8**	**48.7** (+5.0%)	**69.7** (+7.4%)	**28.9**
GLM-4.7 (358B)	Qwen3-Emb	67.7	27.5	43.9	66.6	45.5
	+ LRAT (Ours)	**82.0** (+21.1%)	**18.5**	**54.6** (+24.4%)	**77.8** (+16.8%)	**44.6**
	E5-Large	73.7	24.2	46.4	68.7	**44.6**
	+ LRAT (Ours)	**81.7** (+10.9%)	**19.5**	**50.6** (+9.1%)	**76.3** (+11.1%)	44.8

Ablation Study

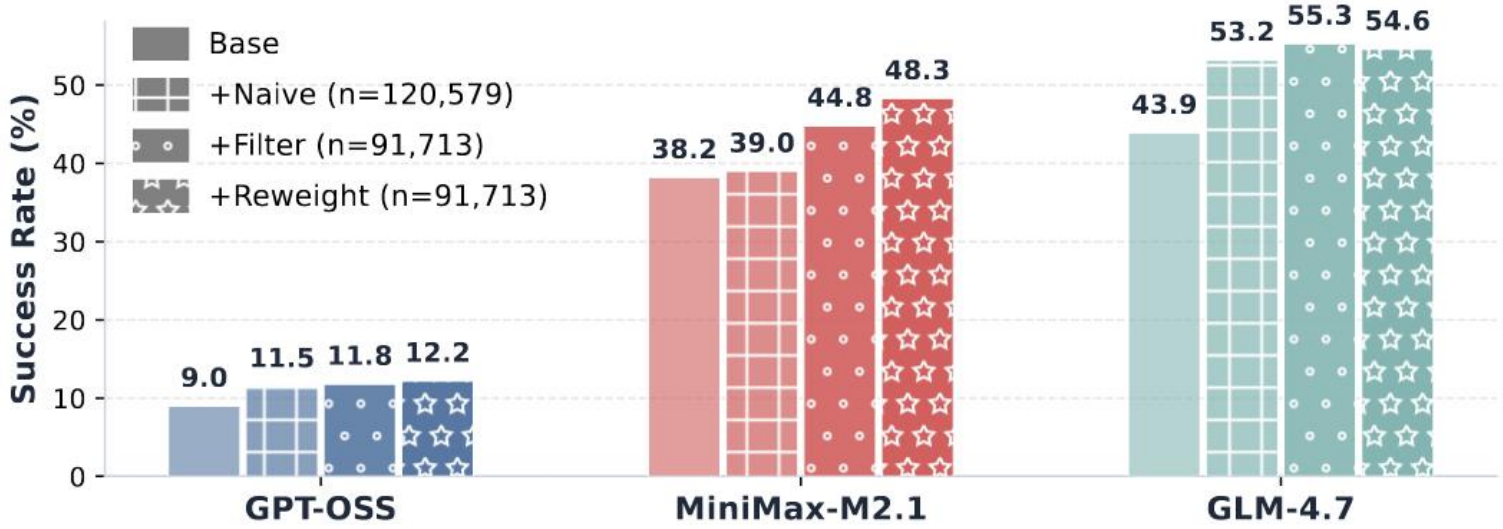


Figure 7 | Ablation study: components are incrementally added. Numbers n in parentheses show the amount of training data used for each variant of LRAT.

Scalability

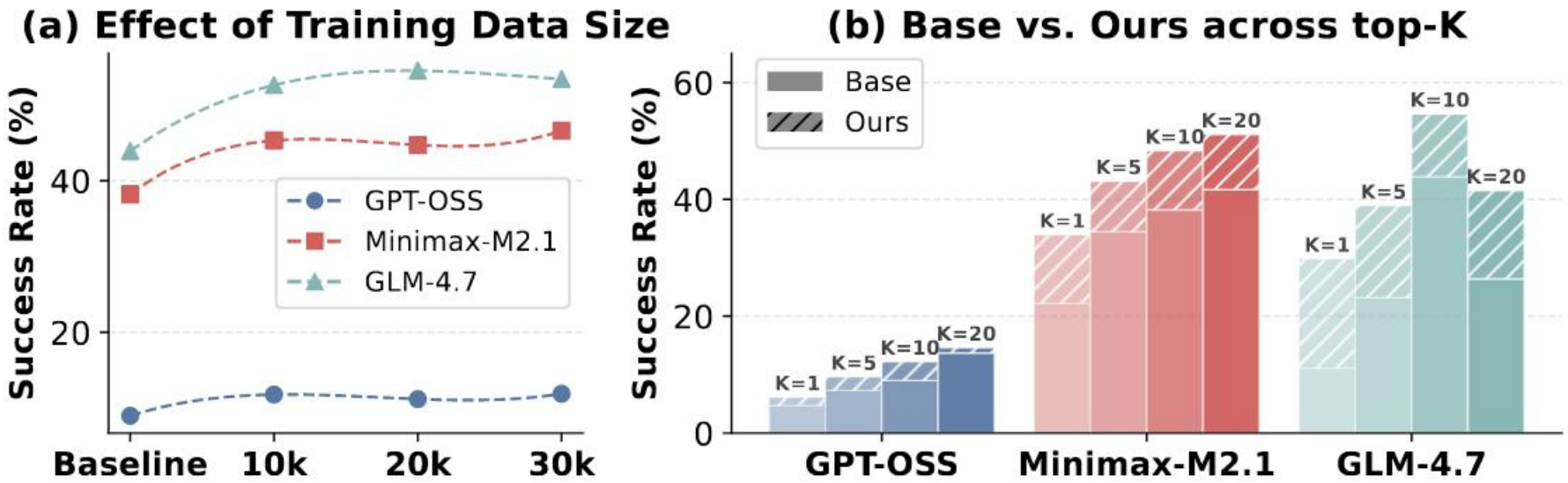


Figure 8 | Agent performance with varying training data sizes and retrieval top-K settings.

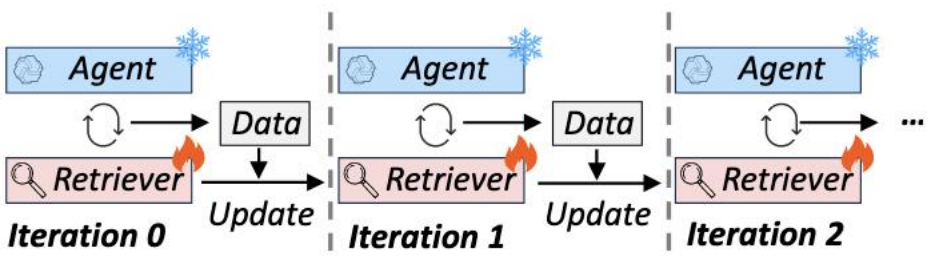
Data Flywheel Simulation

Retrievers trained with both correct and incorrect trajectories consistently outperform the base retriever.

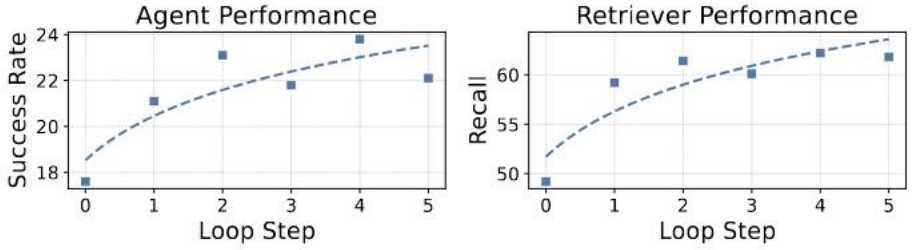
Table 3 | Trajectory correctness ablation. Both correct and incorrect trajectories use 10K examples each.

Training Data	GPT-OSS	MiniMax-M2.1	GLM-4.7
Base (*w/o* LRAT)	9.0	38.2	43.9
LRAT (*w/* Incorrect Trajectories)	10.7 (+18.9%)	43.6 (+14.1%)	50.6 (+15.3%)
LRAT (*w/* Correct Trajectories)	11.8 (+31.1%)	45.3 (+18.6%)	52.6 (+19.8%)

LRAT can reliably support iterative retriever updates and sustain a positive data flywheel.



(a) Illustration of data flywheel simulation setting.



(b) Performance of the data flywheel simulation.

we adopt the Tongyi-DeepResearch agent and sample 10K queries from InfoSeekQA at each step.