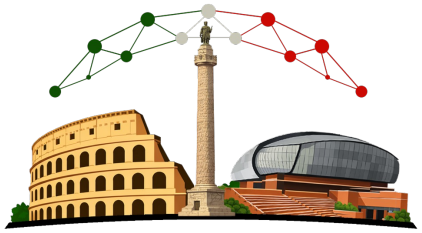
Agentic Search:
Foundations, Optimization, and Long-Horizons

Chuan Meng, Litu Ou, Sean MacAvaney, and Jeff Dalton

Tutorial at the 35th ACM International Conference on Information and Knowledge Management (CIKM 2026)

November 7, 2026 | Rome, Italy



Presenters



Chuan Meng
The University of Edinburgh



Litu Ou
The University of Edinburgh



Sean MacAvaney
University of Glasgow



Jeff Dalton
The University of Edinburgh

Supplementary materials

The materials, including slides, are available on the tutorial website at https://agentic-search.github.io/



Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30

PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

Outline

PART I 14:00–16:00

Basics and optimization

- **Introduction to agentic search**
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30

PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

Introduction to agentic search

I want a day trip from London on Saturday this week.
Find a seaside destination with good weather, less than two hours away by train, and a return ticket under £50.

Take 1 minute: What's your answer and strategy?

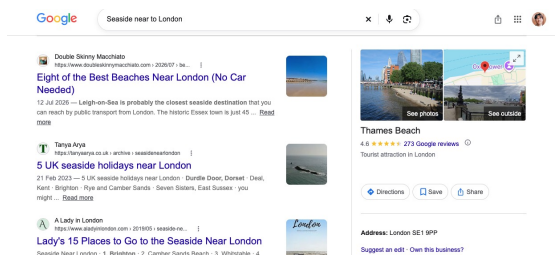
Introduction to agentic search

- **Human-driven search (traditional search): users drive much of the search process** by decomposing information needs, formulating queries, inspecting retrieved documents, and synthesizing answers

I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.
Decompose: 1. seaside towns near London? 2. which have good weather? 3. which has a fast, cheap train?

First check seaside towns near London

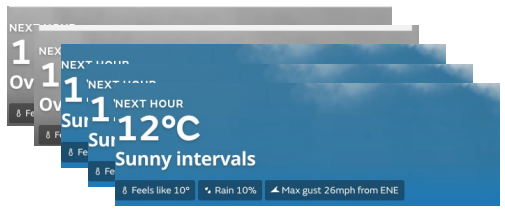
seaside towns near London



Five candidates: Brighton, Deal, Rye, Eastbourne, Durdle Door. Now check the weather

Brighton weather

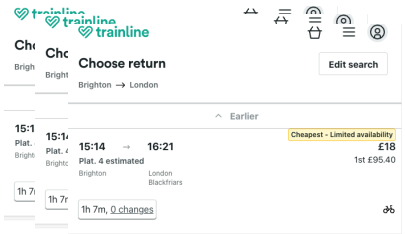
×5 towns



Only Brighton, Deal and Rye have good weather. Next, check train time and price

London to Brighton train

×3 towns



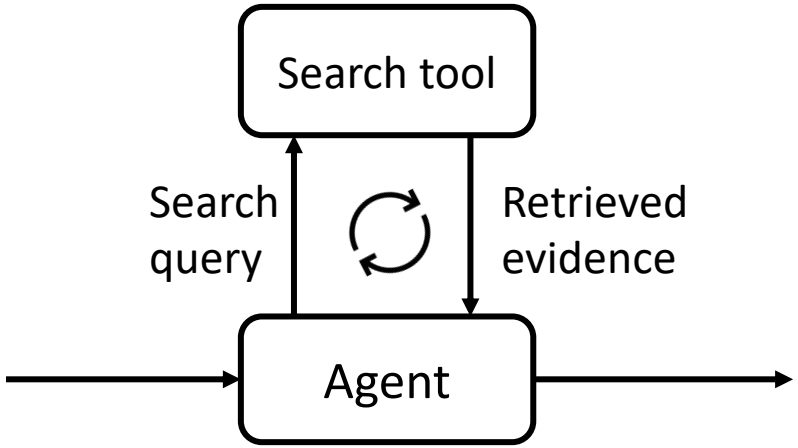
Brighton is sunny, 57 min direct, ~£36 return

Introduction to agentic search

- Agent: an autonomous system that makes decisions and takes actions on users' behalf
- **Agentic search: an agent takes over much of the cognitive burden from the user**
  - Given a complex information need: 1) generate search queries, 2) read the retrieved results, 3) loop until "enough" evidence is gathered, 4) generate an answer (with evidence)



I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.



Evidence



Short answer

Brighton

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

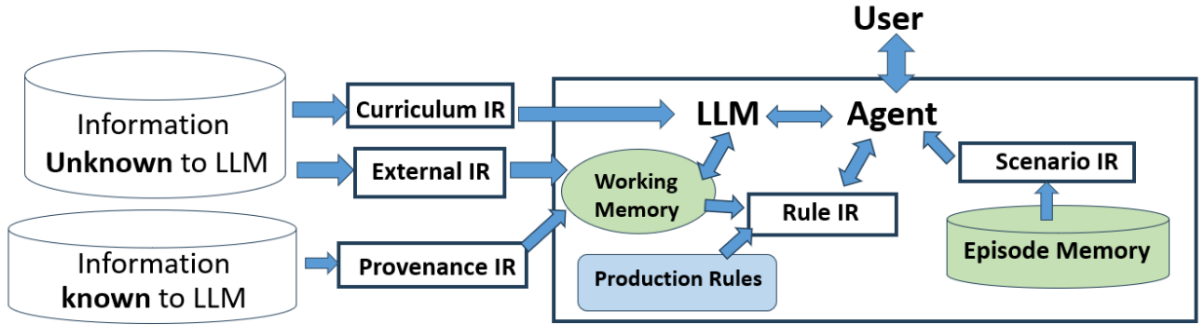
Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Introduction to agentic search

Why agentic search matters

- **For humans: it boosts productivity in knowledge-intensive work** (e.g., finance, science, policy, engineering)
  - **Hours of manual research → tens of minutes** [1]
  - It takes over searching that would otherwise cost humans valuable time, and handles needs that exceed human cognitive capacity or available time

- **For agents: search is core infrastructure**
  - **27.1%** of agent sessions on real-world user tasks call **web search** [2]
  - Agents need search to reach knowledge outside their parameters [3]
    - new or unseen information
    - the provenance of a claim
    - past cases to reuse
    - what to learn next



[1] OpenAI. Introducing Deep Research. 2025.
[2] Arena. Agent Arena: Causal Evaluation of Agents in the Real World. 2026.
[3] ChengXiang Zhai. Information Retrieval for Artificial General Intelligence: A New Perspective on Information Retrieval Research. SIGIR 2025.

Introduction to agentic search

Q & A

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- **Optimization paradigms**
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30

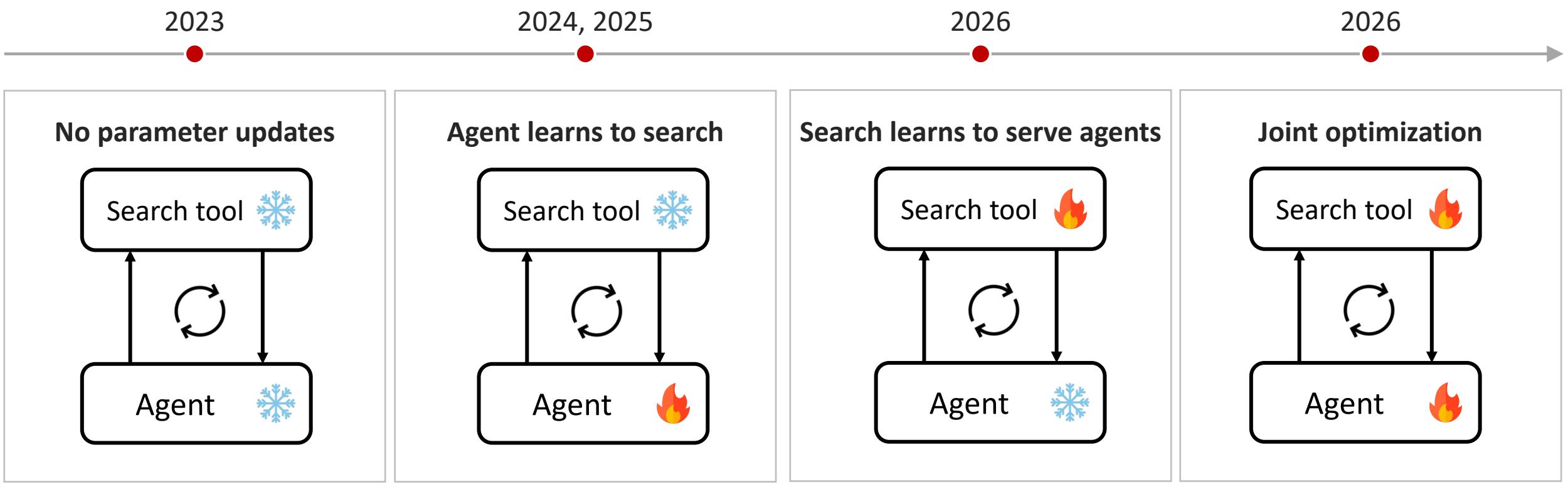
PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

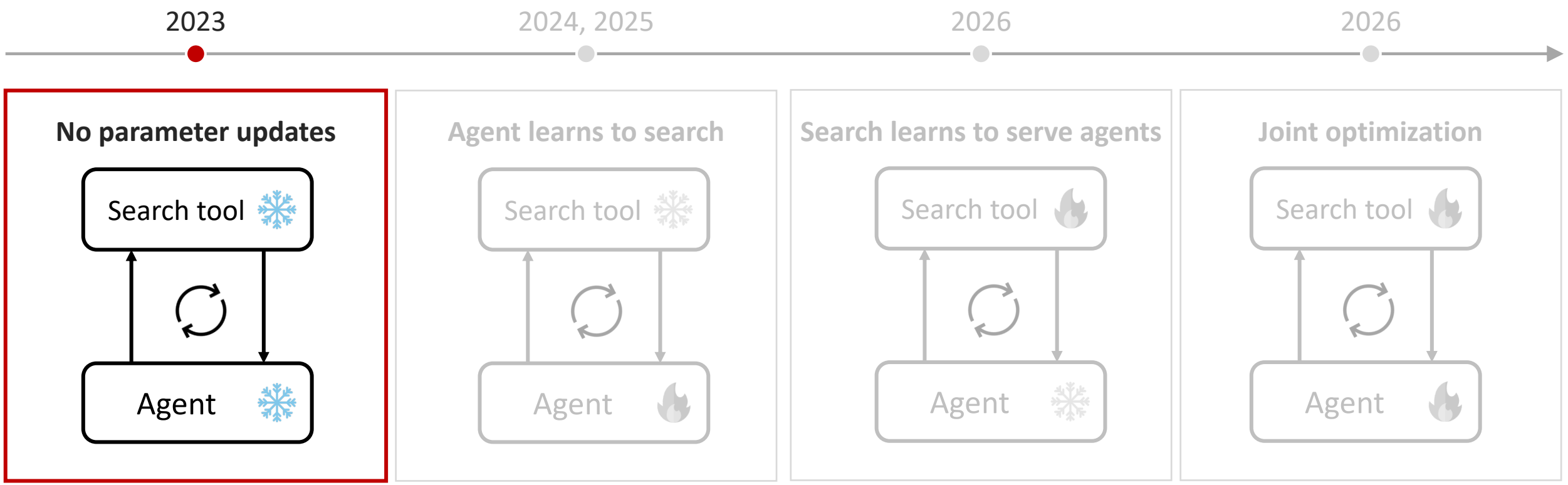
Optimization paradigms

- Four paradigms based on which parameters are updated



Optimization paradigms

- Four paradigms based on which parameters are updated



🔥 trained ❄ frozen

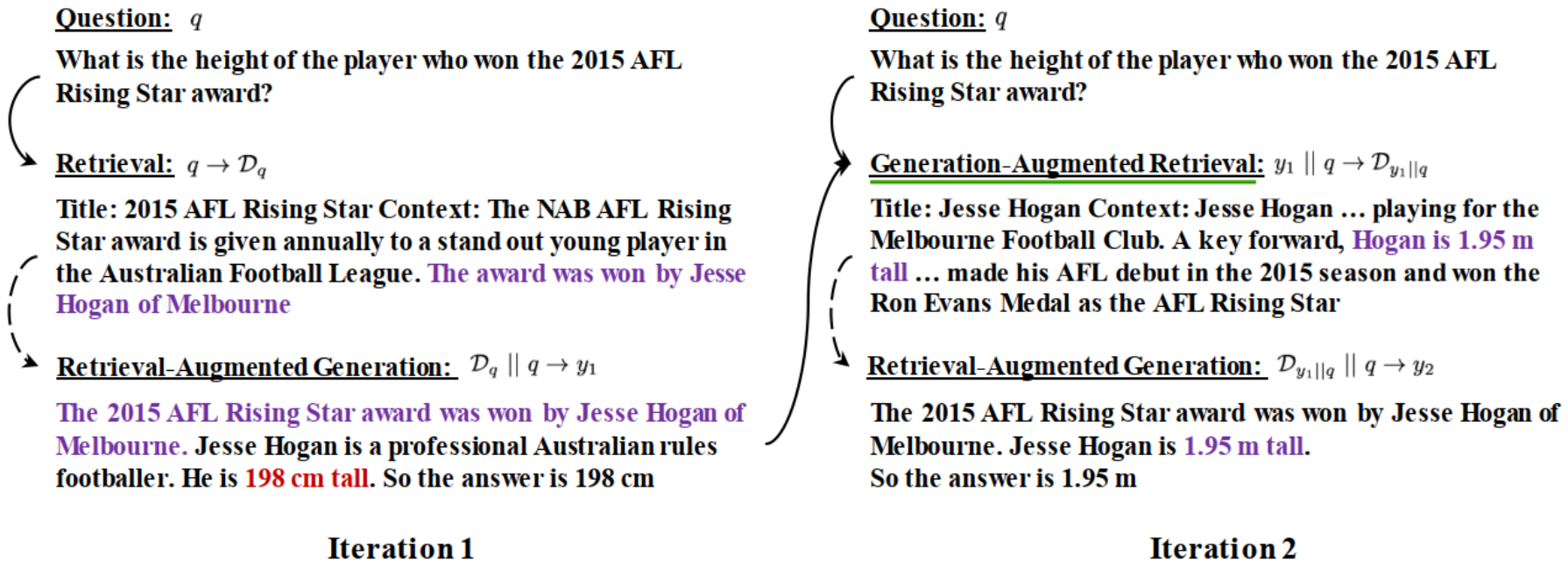
Optimization paradigms

- **No parameter updates:** prompting agents (few-shot examples) to formulate queries, invoke fixed search models, and incorporate retrieved evidence iteratively
  - Answer then search [1]
  - Analyze then search [2,3]
  - Analyze, search, then analyze again [4,5,6]

[1] Shao et al. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. EMNLP 2023.
[2] Jiang et al. Active Retrieval Augmented Generation. EMNLP 2023.
[3] Press et al. Measuring and Narrowing the Compositionality Gap in Language Models. EMNLP 2023.
[4] Trivedi et al. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. ACL 2023.
[5] Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023.
[6] Li et al. Search-o1: Agentic Search-Enhanced Large Reasoning Models. EMNLP 2025.

Optimization paradigms

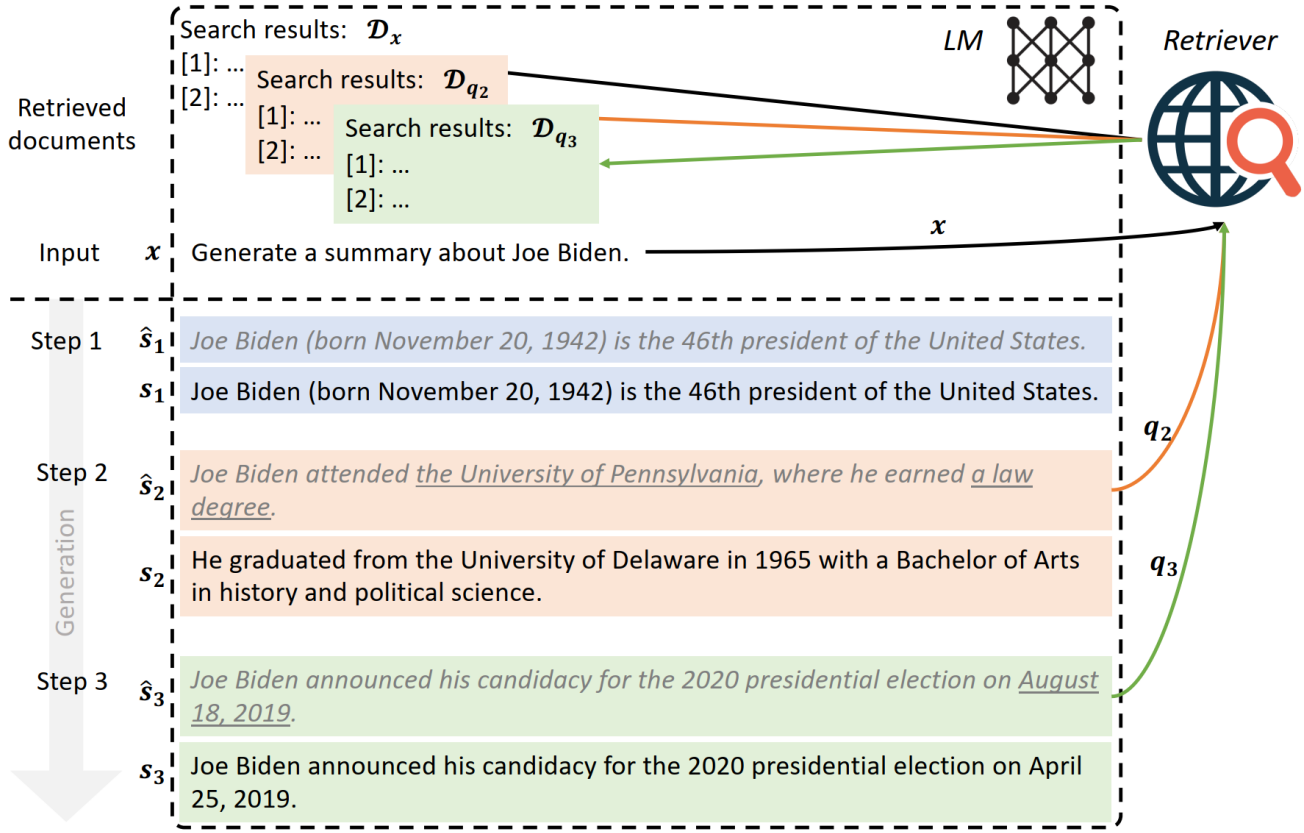
- **No parameter updates:** Answer then search
  - Iter-RetGen [1] generates an answer, uses it to guide the next search, and refines the answer
    - Search with the original question and generate an answer (might be wrong)
    - Search again with the answer and original question, and generate a new answer
    - Repeat for T iterations and return the last answer (T =2 brings most of the gains)



[1] Shao et al. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. EMNLP 2023.

Optimization paradigms

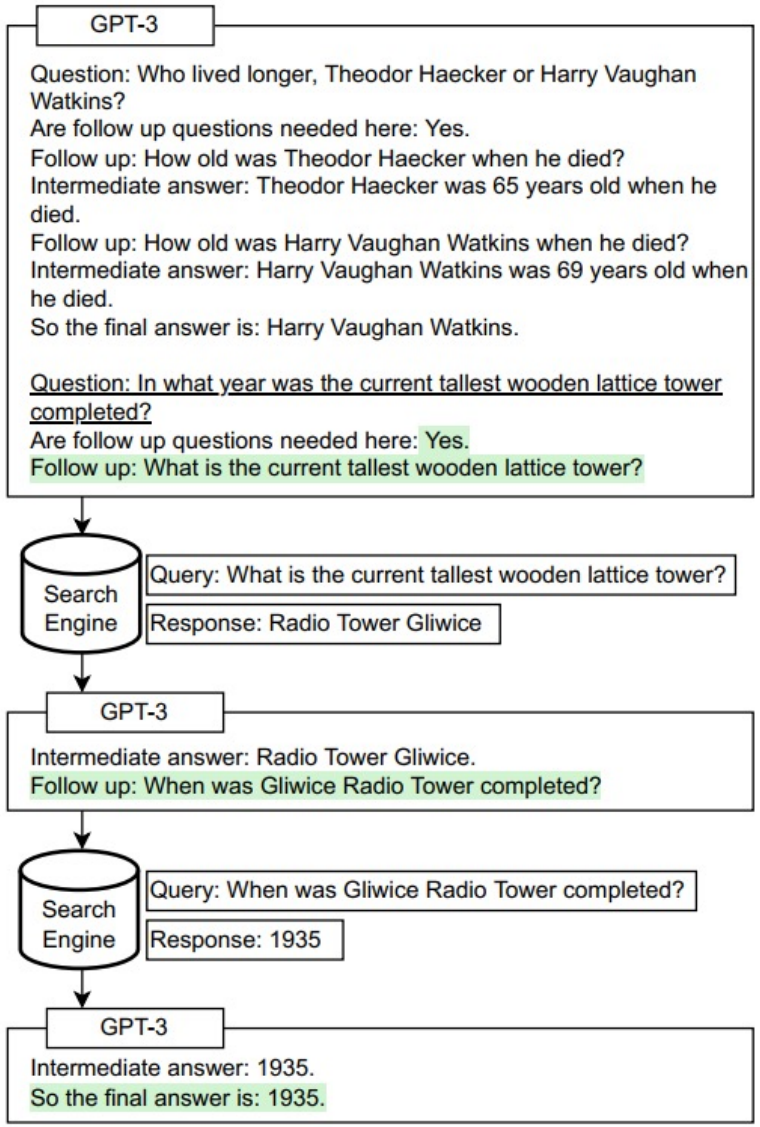
- **No parameter updates:** Analyze then search
  - FLARE [1] drafts the next sentence, checks its confidence, then searches when confidence is low
    - Generate **a draft of the next sentence**
    - If it contains low-confidence tokens, remove them and search with the remaining text
    - Use the retrieved documents to regenerate the sentence
    - Repeat until generation is complete



[1] Jiang et al. Active Retrieval Augmented Generation. EMNLP 2023.

Optimization paradigms

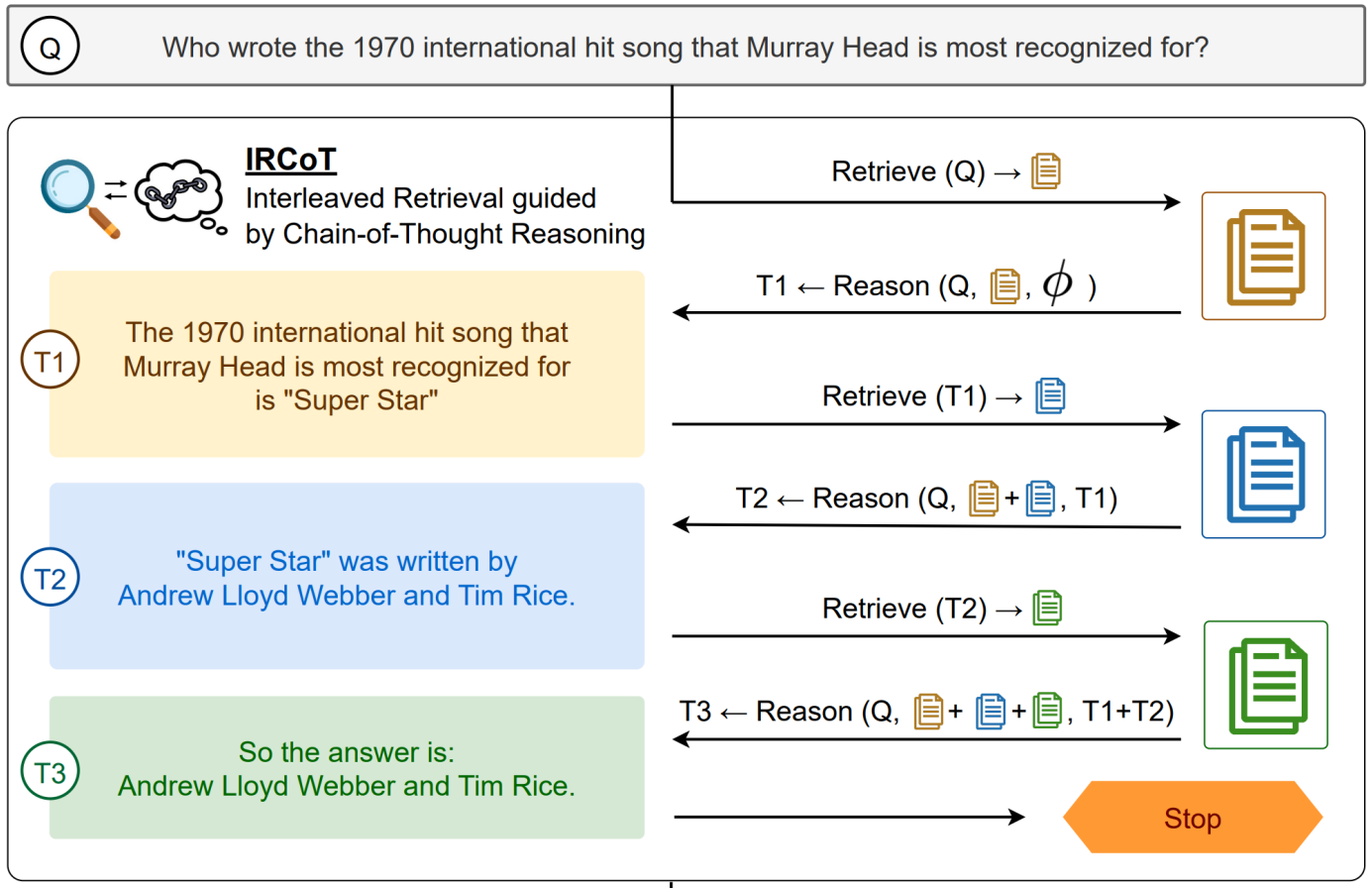
- **No parameter updates:** Analyze then search
  - Self-ask [1] identifies a needed follow-up question, then searches for its answer
    - Generate **a follow-up question** if needed
    - Search with the question
    - Insert the returned answer into the context
    - Repeat until the LLM thinks no more follow-ups are needed and gives the final answer



[1] Press et al. Measuring and Narrowing the Compositionality Gap in Language Models. EMNLP 2023.

Optimization paradigms

- **No parameter updates:** Analyze, search, then analyze again
  - IRCoT [1] first searches with the original question, generates chain-of-thought (CoT) text from the search results, and uses the generated CoT text as the next search query
    - Search with the original question
    - Use the documents to generate a CoT text
    - Search with the CoT text
    - Repeat until "answer is:" appears in CoT or reach T iterations (T = 8)



[1] Trivedi et al. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. ACL 2023.

Optimization paradigms

- **No parameter updates:** Analyze, search, then analyze again
  - ReAct [1] conducts a **reasoning–acting** loop
    - **Reason to act**: Generate a reasoning trace to decide whether and what to search
    - **Act to reason**: Generate a search query, run the search, and use the results in the next reasoning step
    - Repeat until reasoning decides to stop search and generate answer, or reach T iterations

(1d) ReAct (Reason + Act)

Reasoning 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1 (search): Search[Apple Remote]
Retrieved evidence 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

Reasoning 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2 (search): Search[Front Row]
Retrieved evidence 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', …]

Reasoning 3: Front Row is not found. I need to search Front Row (software).
Act 3 (search): Search[Front Row (software)]
Retrieved evidence 3: Front Row is a discontinued media center software …

Reasoning 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4 (answer): Finish[keyboard function keys] ✓

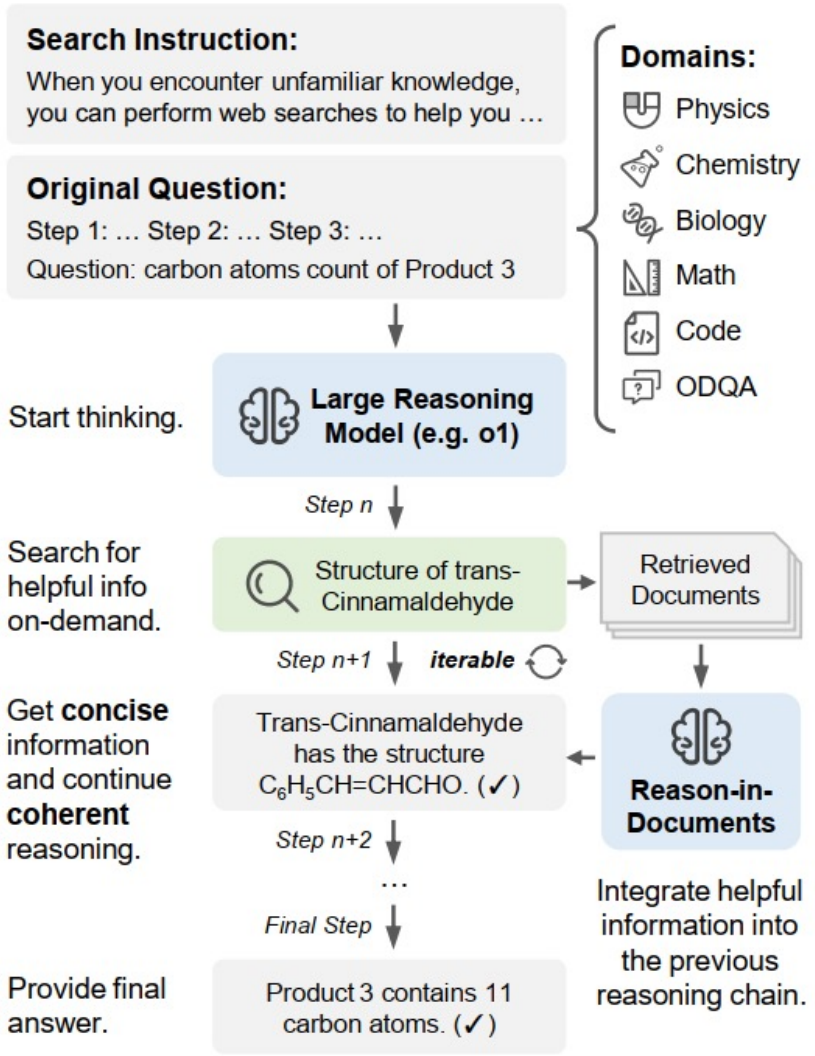
[1] Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023.

Optimization paradigms

- **No parameter updates:** Analyze, search, then analyze again
  - **Previous limitation**: Verbose and noisy retrieved documents can disrupt reasoning
  - Search-o1 [1] designs a **Reason-in-Documents** module to analyze the retrieved information before injecting it into the reasoning chain

Question: When did the author of "Cogito, ergo sum" die?

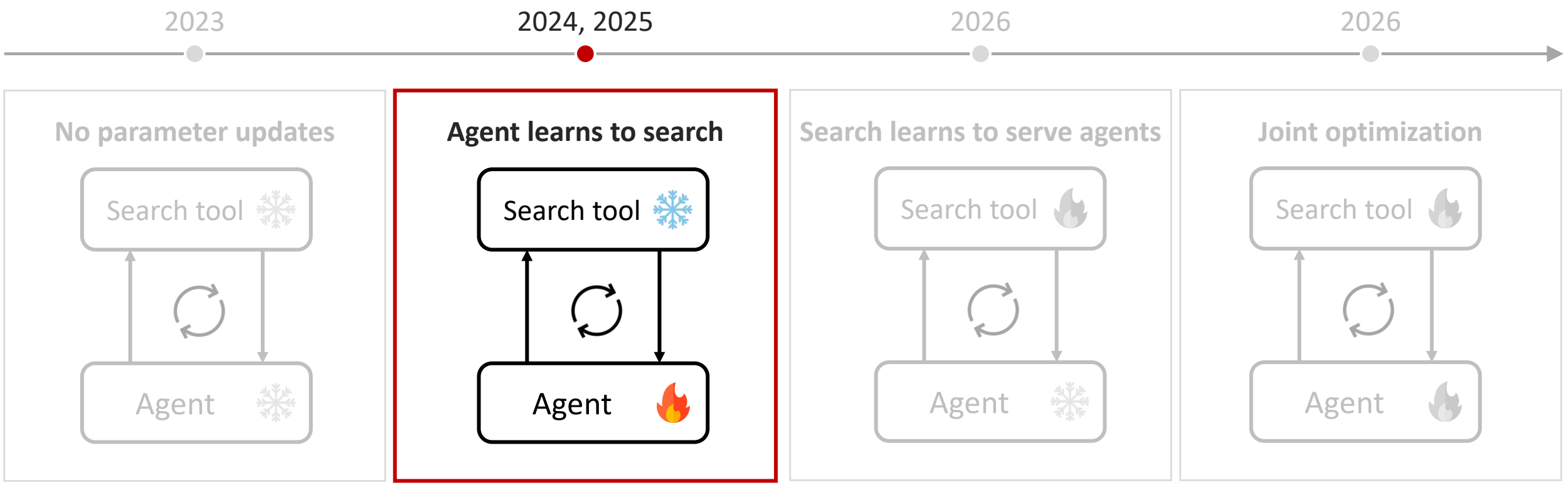
Reasoning so far: Likely Descartes. Let me confirm.
Search query: `"Cogito, ergo sum" author`
Retrieved documents
•Doc 1: The phrase appeared in a **1637** book…
•Doc 2: Descartes published *Principles of Philosophy* in **1644**…
•Doc 3: René Descartes (**1596–1650**) …
•…
Reason-in-Documents:
René Descartes, the author of "Cogito, ergo sum," lived from 1596 to 1650. His death year is **1650**



[1] Li et al. Search-o1: Agentic Search-Enhanced Large Reasoning Models. EMNLP 2025.

Optimization paradigms

- Four paradigms based on which parameters are updated



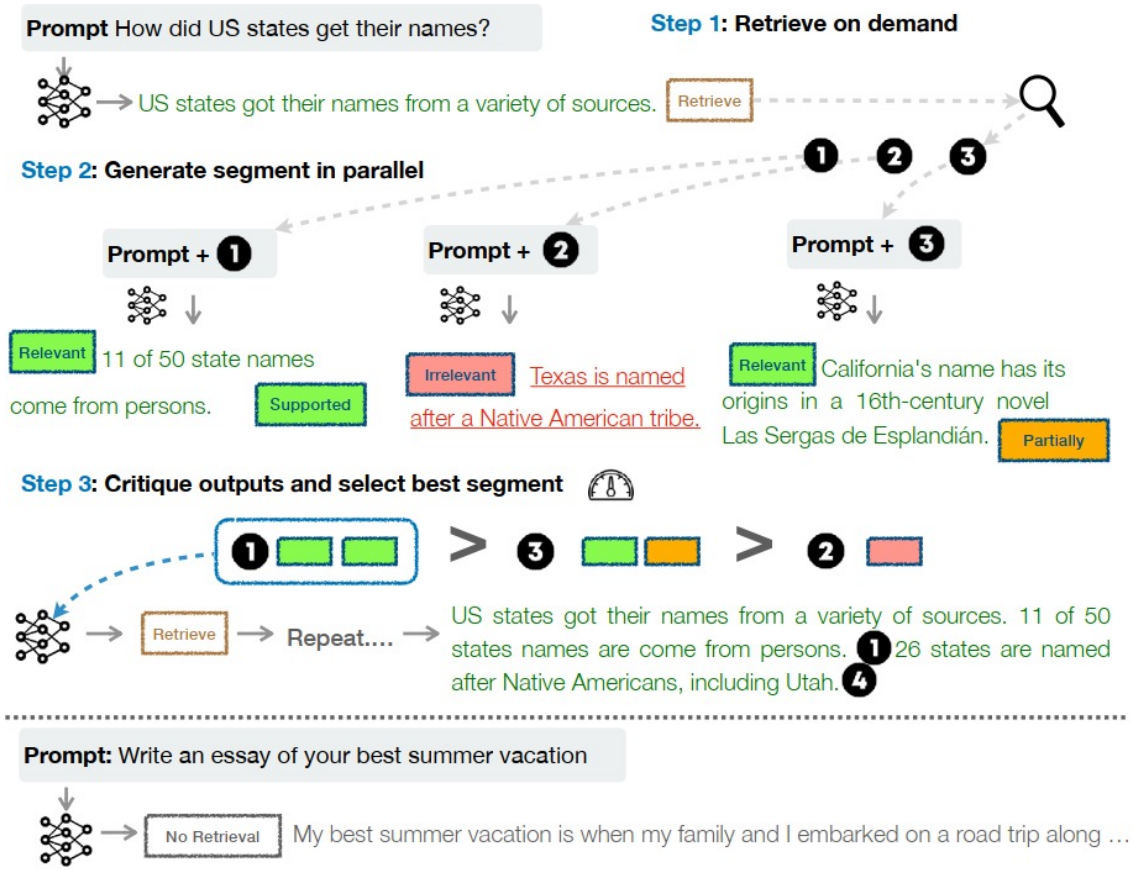
Optimization paradigms

- **Agent learns to search:** train agents to use a fixed search component
  - Learn through supervised fine-tuning (SFT) [1]
  - Learn from task rewards via reinforcement learning (RL) [2,3,4,5,6]
  - Improve RL-based agentic search
    - Training efficiency [7]
    - Inference efficiency [8]
    - Modularity and flexibility [9]

[1] Asai et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ICLR 2024.
[2] Jiang et al. DeepRetrieval: Hacking Real Search Engines and Retrievers with Large Language Models via Reinforcement Learning. COLM 2025.
[3] Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.
[4] Song et al. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. arXiv 2025.
[5] Chen et al. ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. NeurIPS 2025.
[6] Jin et al. An Empirical Study on Reinforcement Learning for Reasoning-Search Interleaved LLM Agents. arXiv 2025.
[7] Sun et al. ZeroSearch: Incentivize the Search Capability of LLMs without Searching. arXiv 2025.
[8] Song et al. Smart-Searcher: Incentivizing the Dynamic Knowledge Acquisition of LLMs via Reinforcement Learning. EMNLP 2025.
[9] Jiang et al. s3: You Don’t Need That Much Data to Train a Search Agent via RL. EMNLP 2025.

Optimization paradigms

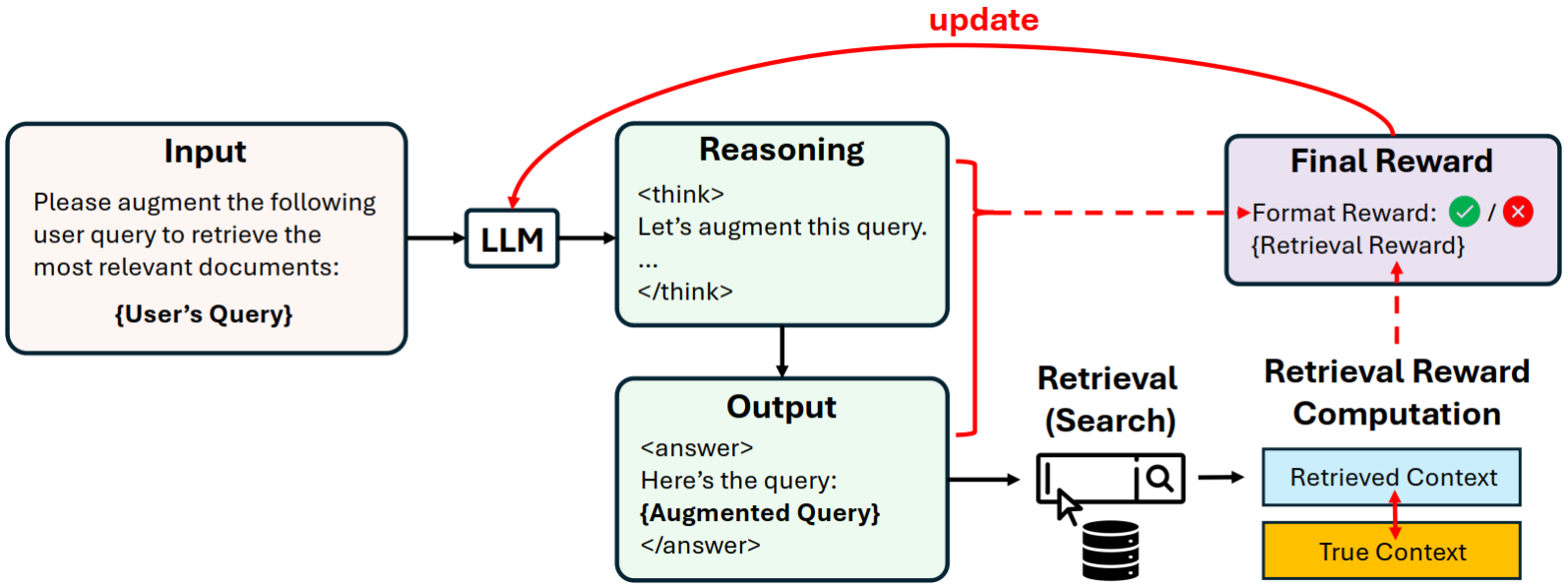
- **Agent learns to search:** Learn through SFT
  - Self-RAG [1] uses SFT to train an LLM to predict special tokens indicating 1) **retrieval necessity**, 2) **evidence relevance**, 3) **response groundedness**
    - Special tokens:
      - Retrieve/No Retrieval
      - Relevant/Irrelevant
      - Supported/Partially supported
    - SFT training steps
      - Train a critic model using GPT-4 labels
      - Use the critic to add reflection tokens to training examples
      - Fine-tune an LLM to generate reflection tokens/responses



[1] Asai et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ICLR 2024.

Optimization paradigms

- **Agent learns to search:** Learn from task rewards via RL
  - DeepRetrieval [1] trains a one-hop query augmentation model to improve retrieval effectiveness
    - Reward=retrieval reward + format reward
      - Retrieval reward: Task-specific metrics, e.g., Recall@K, NDCG, or SQL execution accuracy
      - Format reward: Check the <think>…</think><answer>…</answer> structure

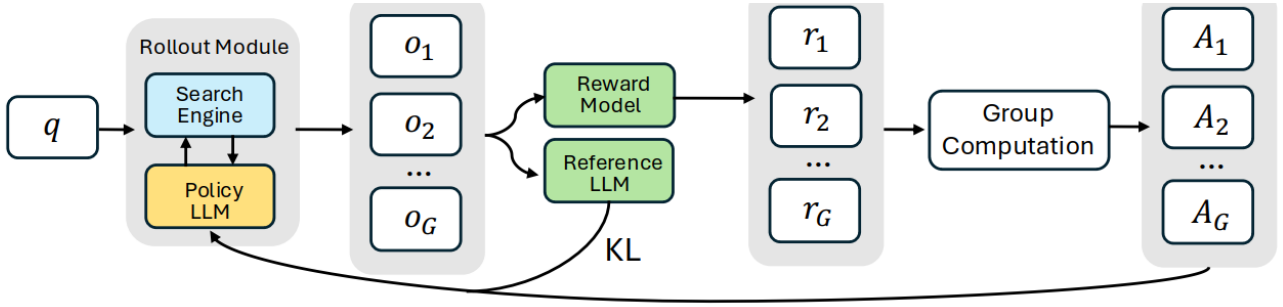


[1] Jiang et al. DeepRetrieval: Hacking Real Search Engines and Retrievers with Large Language Models via Reinforcement Learning. COLM 2025.

Optimization paradigms

- **Agent learns to search:** Learn from task rewards via RL
  - Three near-concurrent works in March 2025, namely **Search-R1** [1], **R1-Searcher** [2], and **ReSearch** [3], use RL to train LLMs to interleave reasoning and search in a ReAct-style interaction loop
    - Reward
      - Search-R1: Final-answer Exact Match (EM)
      - R1-Searcher
        - Stage 1: +0.5 for using search and +0.5 for correct format
        - Stage 2: Answer F1, with a −2 penalty for incorrect format
      - ReSearch: Answer F1; when F1 = 0, reward is 0.1 for correct format and 0 otherwise

Example: Search-R1 with Group Relative Policy Optimization (GRPO)

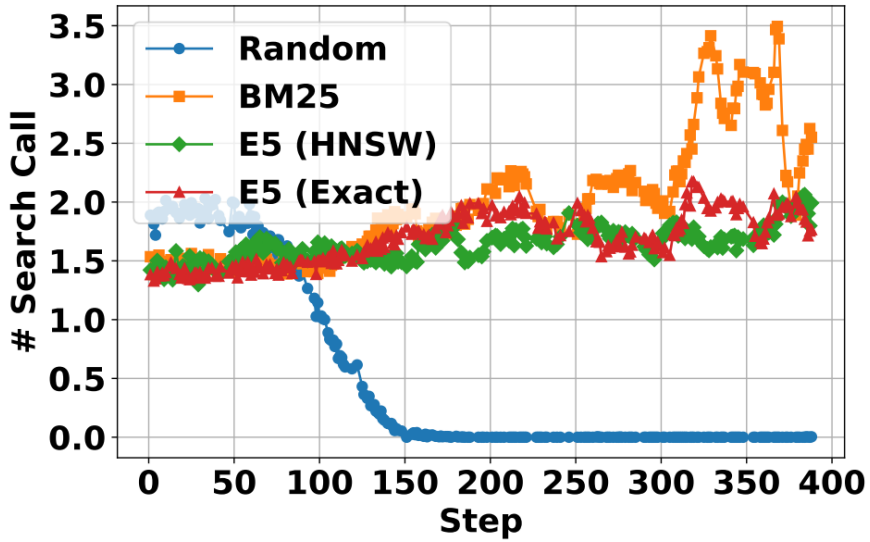


$$A_i = \frac{r_i - \text{mean}(r)}{\text{std}(r) + \epsilon}$$

[1] Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.
[2] Song et al. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. arXiv 2025.
[3] Chen et al. ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. NeurIPS 2025.

Optimization paradigms

- **Agent learns to search:** Learn from task rewards via RL
  - Search-R1's follow-up empirical study [1] examines reward design, model initialization, and retrieval setup
    - **Reward design**: Format rewards can help; intermediate retrieval rewards add little
    - **Initialization**: Qwen2.5-Base outperforms DeepSeek-R1-Distill-Qwen
    - **The retrieval setup shapes search behavior**
      - Random evidence → search avoidance
      - BM25 → more search calls
      - Stronger dense retrieval → more efficient search



[1] Jin et al. An Empirical Study on Reinforcement Learning for Reasoning-Search Interleaved LLM Agents. arXiv 2025.

Optimization paradigms

- **Agent learns to search**: Improve RL-based agentic search, in terms of training efficiency
  - **Previous limitation:** Frequent search calls during RL training incur high API costs; result quality is hard to control
  - **ZeroSearch** [1]: Train agents with an LLM that simulates search results
    - Simulator
      - Prompt an LLM to answer questions using Google API
      - Use an LLM judge to label retrieved results as useful or noisy
      - Fine-tune a separate LLM (simulator) to generate both types
    - Curriculum learning in RL: Gradually increase search noise

Performance in terms of EM

Search Engine	Avg.
SFT-3B	30.47
SFT-7B	33.53
SFT-14B	34.47
Google	32.81

You are the Google search engine.
Given a query, you need to generate five [useful / noisy] documents for the query.
The user is trying to answer the question: [question] whose answer is [ground truth].
Each document should contain about 30 words, and these documents should contain [useful / noisy] information.
Query: [query]
[Useful / Noisy] Output:

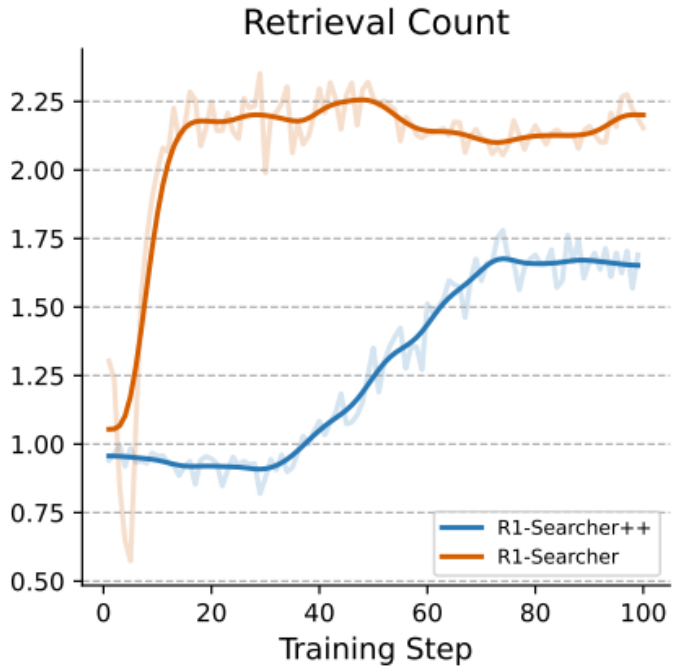
Search Engine	Queries	Training Time	GPUs Used	Total Cost
SFT-3B	~64,000	~12 hours	1 × A100 GPUs	**$17.7**
SFT-7B	~64,000	~12 hours	2 × A100 GPUs	**$35.4**
SFT-14B	~64,000	~12 hours	4 × A100 GPUs	**$70.8**
Google	~64,000	~12 hours	None	**$586.7**

[1] Sun et al. ZeroSearch: Incentivize the Search Capability of LLMs without Searching. arXiv 2025.

Optimization paradigms

- **Agent learns to search**: Improve RL-based agentic search, in terms of inference efficiency
  - **Previous limitation:** Agents overuse search, increasing inference cost
  - **R1-Searcher++** [1] encourages agents to use internal knowledge to reduce over-reliance on search
    - **Reward = answer + format + group**
      - **Answer:** 1 if the prediction covers the gold answer within 10 words, else 0
      - **Format:** 0 if valid, −2 otherwise
      - **Group:**
        - Sample multiple rollouts for the same question
        - Give a bonus to **correct rollouts with the fewest search calls**

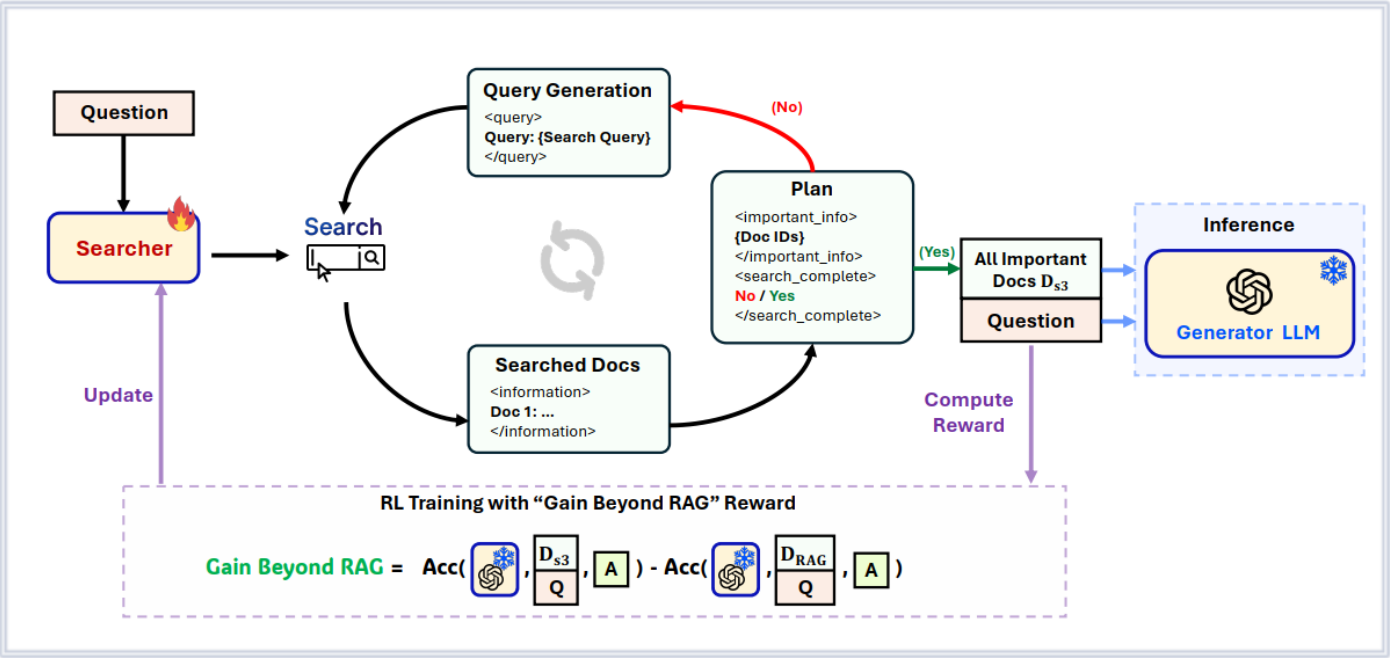
Rollout	Correct?	Search calls	Group bonus
A	Yes	**1**	**Yes**
B	Yes	3	No
C	No	0	No



[1] Song et al. Smart-Searcher: Incentivizing the Dynamic Knowledge Acquisition of LLMs via Reinforcement Learning. EMNLP 2025.

Optimization paradigms

- **Agent learns to search**: Improve RL-based agentic search, in terms of modularity and flexibility
  - **Previous limitation:** Search and answer generation are tightly coupled, making it hard to use different or proprietary generators
  - **s3** [1]: Train a separate search agent with RL, while keeping the generator frozen
    - Reward=Answer accuracy(s3 evidence) - Answer accuracy(standard RAG evidence)
    - Need fewer training steps, reducing total training time

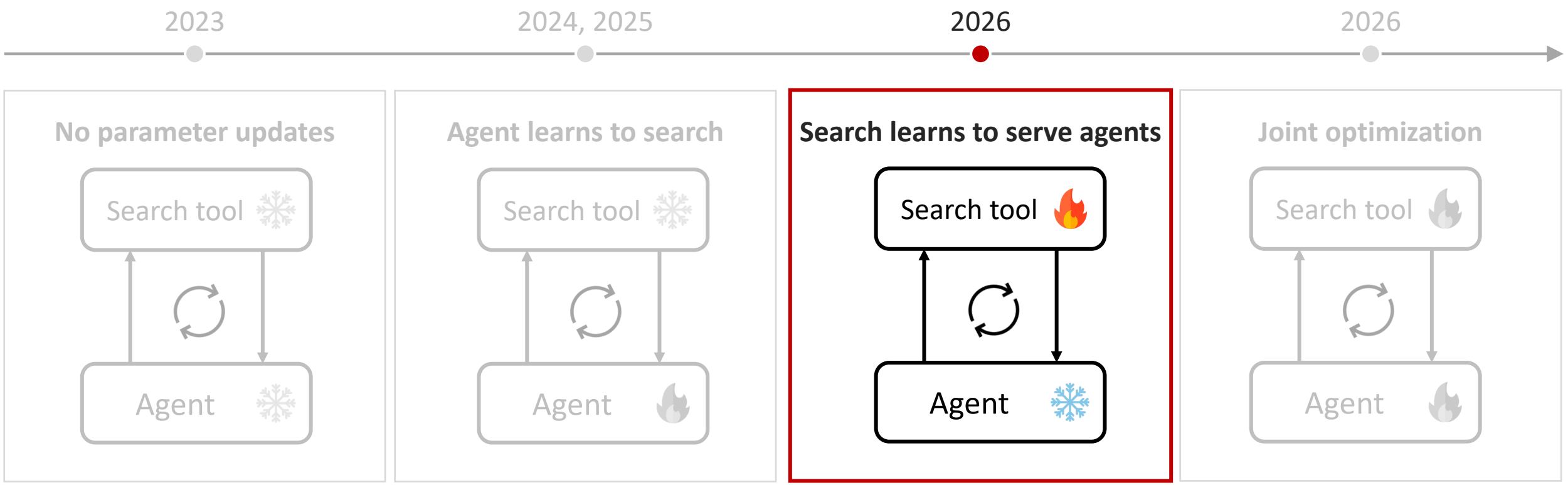


	Time/Step	Training Steps	Total
Search-R1	1.8m	~2,100	3,780m
DeepRetrieval$_{BM25}$	1.3m	~1,600	2,080m
s3	5.7m	~20	**114m**

[1] Jiang et al. s3: You Don't Need That Much Data to Train a Search Agent via RL. EMNLP 2025

Optimization paradigms

- Four paradigms based on which parameters are updated



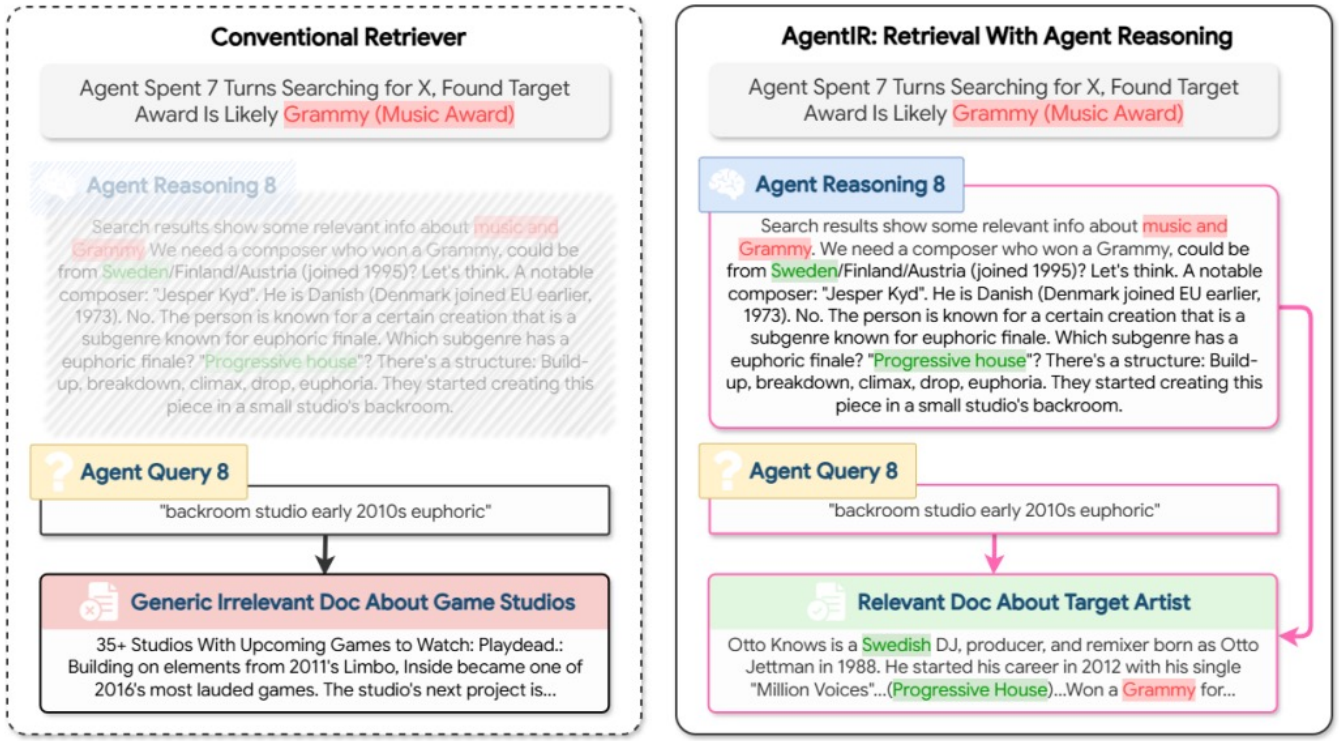
Optimization paradigms

- **Search learns to serve agents:** keeps agents fixed and trains search methods to better serve agents
  - Use agent context for search [1,2]
  - Learn from agent feedback [3]

[1] Chen et al. AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.
[2] Sharifymoghaddam et al. ACARe: Agent-Context-Aware Reranker for Iterative Search. arXiv 2026.
[3] Alam et al. Critic-R: Improving Agentic Search using Instruction-tuned Retrievers with Natural Language Introspective Feedback. EMNLP 2026.

Optimization paradigms

- **Search learns to serve agents**: Use agent context for search
  - **Previous limitation**: Queries alone may not fully express the agent's search intent
  - AgentIR [1] trains a dense retriever to jointly encode the query and agent's latest reasoning trace



Training procedure

- Collect reasoning–query pairs from correct agentic search trajectories on a QA dataset
- Rerank documents by an LLM given the gold answer
  - top document as positive
  - bottom documents as negatives
- Fine-tune the retriever with contrastive learning using triples: <**reasoning** + query, positive doc, negative doc>

[1] Chen et al. AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

Optimization paradigms

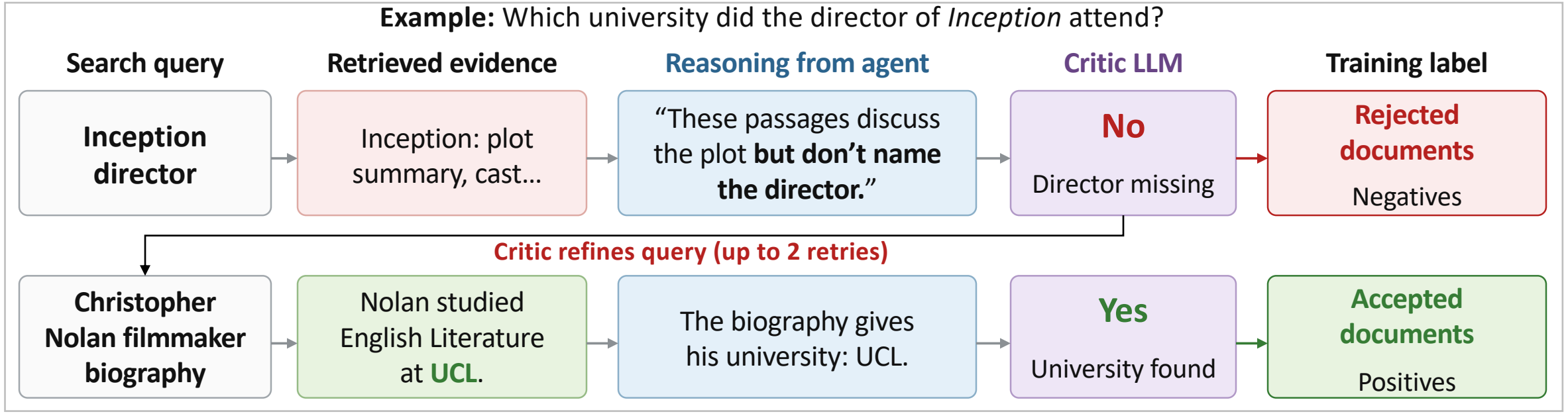
- **Search learns to serve agents**: Use agent context for search
  - ACARe [1] trains a pointwise reranker using query, original question, and agent's latest reasoning trace
    - Training procedure:
      - Reuse and expand AgentIR's synthetic training data
      - Fine-tune with contrastive learning using triples:
        - ⟨query + **original question** + **latest reasoning trace**, positive doc, negative doc⟩
      - Randomly drop the **original question** and **reasoning** independently during training

Reranker, context	nDCG@10↑	Recall@10↑	Avg. per query stats		
			Searches↓	LLM turns↓	Cost (¢)↓
No reranker	66.9 (±1.3)	68.6 (±1.0)	14.9 (±0.4)	16.0 (±0.4)	35.0 (±1.8)
Qwen3-Reranker-4B	69.7 (±1.1)†	72.2 (±1.5)†	13.1 (±0.3)†	14.1 (±0.3)†	30.8 (±0.8)†
ACARe-4B (query only)	70.8 (±0.9)†	73.4 (±1.3)†	12.7 (±0.4)†	13.8 (±0.4)†	33.0 (±2.1)
ACARe-4B (query + reasoning)	71.9 (±0.6)†*	74.9 (±0.6)†	12.2 (±0.3)†*	13.2 (±0.3)†*	31.9 (±1.3)
ACARe-4B (query + original question)	72.1 (±1.6)†	74.8 (±1.8)†	12.0 (±0.3)†*	13.0 (±0.3)†*	31.3 (±1.2)†
ACARe-4B (query + original question + reasoning)	**72.6** (±1.2)†*	**75.2** (±0.9)†*	**11.6** (±0.3)†*	**12.6** (±0.3)†*	**25.9** (±1.4)†*

[1] Sharifymoghaddam et al. ACARe: Agent-Context-Aware Reranker for Iterative Search. arXiv 2026.

Optimization paradigms

- **Search learns to serve agents**: Learn from agent feedback
  - **Previous limitation**: Human-annotated document labels are costly and may not reflect agents' reasoning needs
  - Critic-R [1] trains a retriever using document labels derived from agent feedback
    - Training procedure: A critic LLM labels retrieved evidence using the agent's reasoning

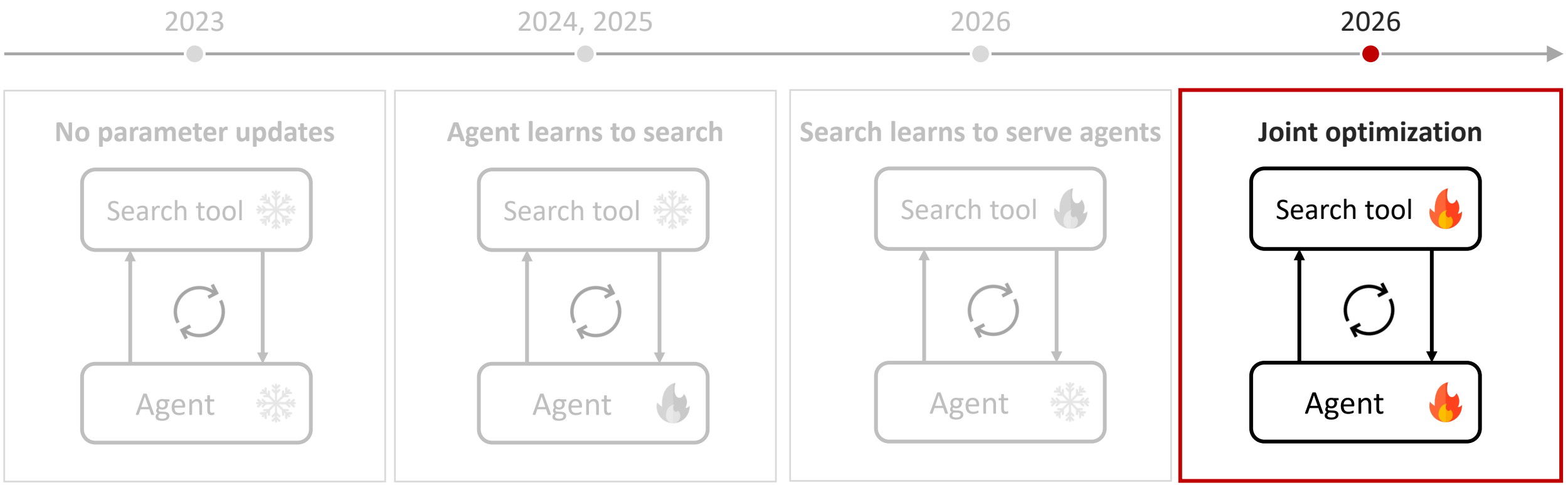


- Fine-tune retriever with contrastive learning using triples ⟨query, positive doc, negative docs⟩

[1] Alam et al. Critic-R: Improving Agentic Search using Instruction-tuned Retrievers with Natural Language Introspective Feedback. EMNLP 2026.

Optimization paradigms

- Four paradigms based on which parameters are updated



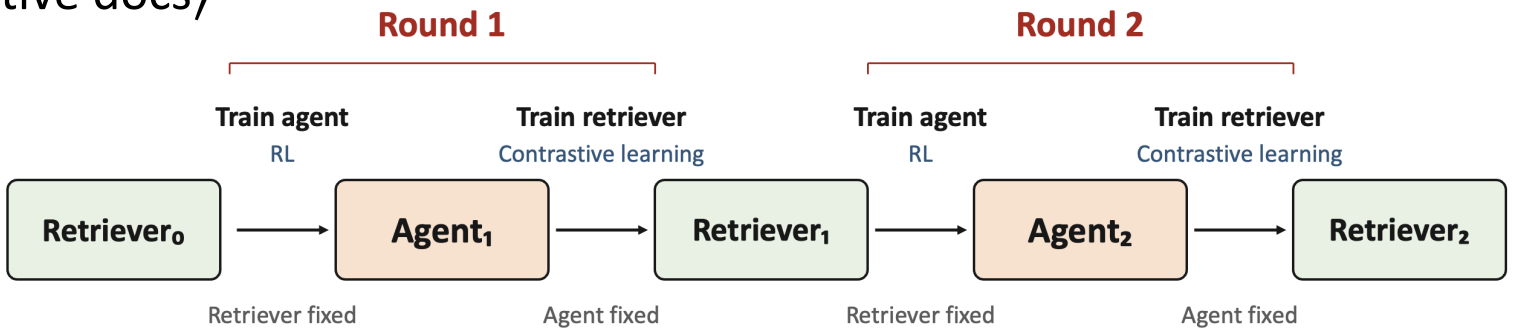
Optimization paradigms

- **Agent–search joint optimization:** jointly optimizes agents and search so that each part adapts to the other
  - Alternating training [1]
  - Simultaneous training [2]

[1] Liu et al. Agentic-R: Learning to Retrieve for Agentic Search. ACL 2026.
[2] Zeng et al. CoSearch: Joint Training of Reasoning and Document Ranking via Reinforcement Learning for Agentic Search. COLM 2026.

Optimization paradigms

- **Agent–search joint optimization**: Alternating training
  - **Previous limitation**: Optimizing only one component limits mutual adaptation
  - Agentic-R [1] alternates between optimizing the agent and the retriever
    - Training procedure (run both stages for up to two rounds)
      1. Train the agent: Follow Search-R1's RL setup, keeping the current retriever fixed
      2. Train the retriever: Keep the agent fixed
         - Generate trajectories with the improved agent
         - For each query in a trajectory, rank documents by final answer correctness, then local relevance (by an LLM); use the top doc as positive; sample lower-ranked negatives
         - Fine-tune with contrastive learning using triples: ⟨original question + query, positive doc, negative docs⟩

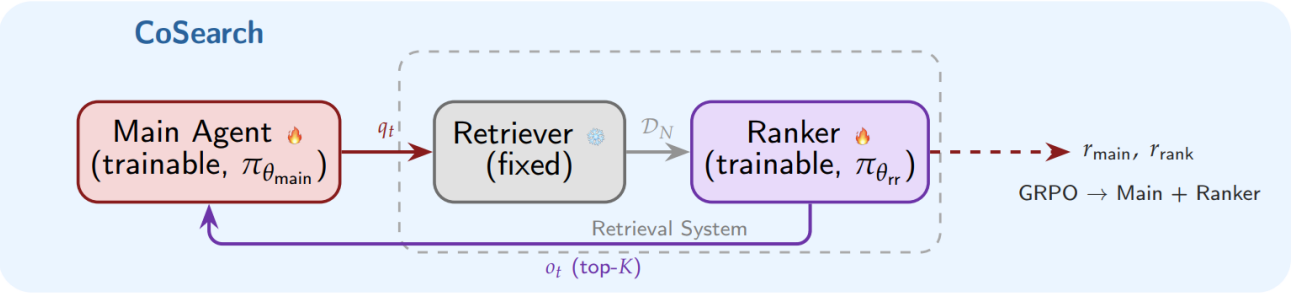


[1] Liu et al. Agentic-R: Learning to Retrieve for Agentic Search. ACL 2026.

Optimization paradigms

- **Agent–search joint optimization**: Simultaneous training
  - CoSearch [1] trains an agent and a re-ranker simultaneously via RL (GRPO)
    - Reward
      - Agent reward: final-answer F1
      - Reranker reward:

Situation	Reranker reward
A relevant doc exists but ranks low	Ranking reward (Hit@k)
A relevant doc ranks high	Ranking reward (Hit@k) + final-answer F1
No relevant doc in re-ranking candidates	Final-answer F1



Re-ranker reward ablation

Configuration	Single-Hop	Multi-Hop	Avg
CoSearch (full)	**0.619**	**0.529**	**0.568**
(a) Reward signal			
w/o r_{main}	0.612	0.508	0.560
w/o r_{rel}	0.591	0.513	0.552

[1] Zeng et al. CoSearch: Joint Training of Reasoning and Document Ranking via Reinforcement Learning for Agentic Search. COLM 2026.

Optimization paradigms

Q & A

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- **Live demonstration and hands-on**

BREAK 16:00–16:30

PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

Live demonstration and hands-on

- What we test
  1. **How does the agent search?**
  2. **Does the retriever matter?**
  3. **Does a re-ranker help?**
  4. **Does reading whole pages help?**
  5. **Your questions:** ask the agent anything, and we analyse what it does together

Live demonstration and hands-on

- Setup
  - **Agent:**
    - gpt-oss-20b
  - **Tools:**
    - search(query) → top 5 pages, first 512 tokens of each
    - get_document(docid) → a whole page (Step 4 only)
  - **Document corpus:** BrowseComp-Plus [1], a fixed set of about 100K web pages
  - **Question:** one question we wrote, the same in Steps 1–4; it is about December 2024, after the model's training data, so the agent has to search
    - *"Notre-Dame de Paris reopened after the 2019 fire. On what date did it reopen, and which pop star performed at the reopening?"*

[1] Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.

Live demonstration and hands-on

Ask the agent a question

- Scan the QR code, or go to **slido.com** and enter **#3678741**
- Ask anything you like, and upvote the questions you want to see. We will run some of them and go through what the agent does.
- No sign-in needed, and you can ask anonymously.

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30



PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30

PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30

PART II 16:30–18:30

Long-horizons, evaluation and beyond

- **Long-horizon agentic search (deep research)**
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

Long-horizon agentic search: General challenges

- Characteristics of long-horizon agentic search:
  - **Longer trajectories**: Models often take tens or even hundreds of search and reasoning steps to solve a complex task.
  - **Handling Uncertainty**: Key evidence may be difficult to find, requiring agents to iteratively refine queries, explore alternative paths, backtrack from dead ends, and verify findings.
  - **Evidence management**: As information accumulates and context grows, agents must retain, organize, and synthesize relevant evidence to produce reliable answers.

- Representative Tasks:
  - **Reasoning-intensive QA:** Answer complex, well-defined questions through extensive exploration and multi-step reasoning, often requiring evidence scattered across multiple sources.
  - **Report generation:** Conduct comprehensive research on broad, open-ended topics, gathering and synthesizing information from diverse sources into structured, evidence-grounded reports.

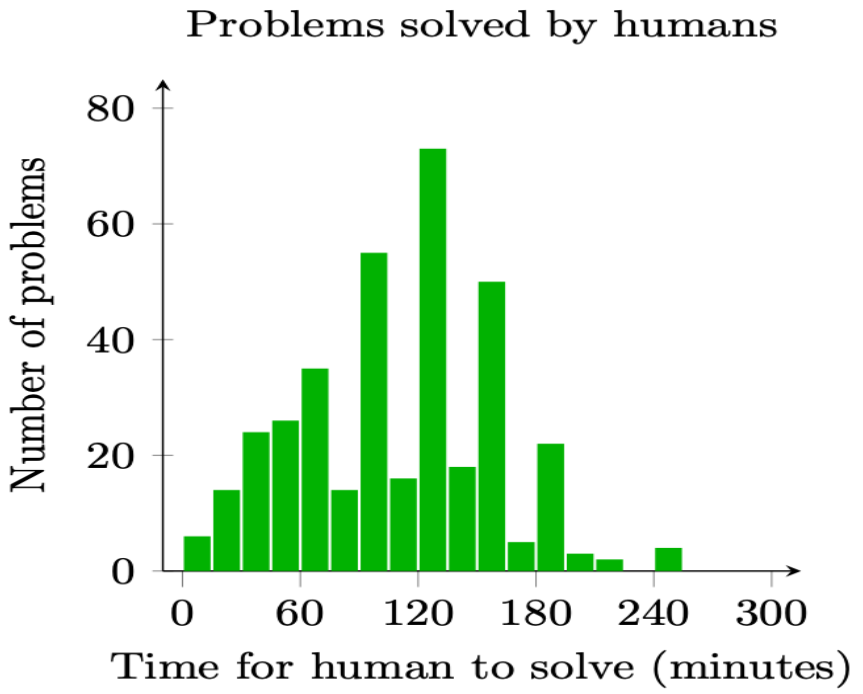
Long-horizon agentic search: Reasoning-Intensive QA

- **Reasoning-intensive QA:** four works that go from defining the task, to training agents, to evaluating them under control
  - Benchmark that defines the task: **BrowseComp [1]**
  - Training a deep research agent: **WebSailor [2]**
  - Controlled evaluation with a local corpus and reproducible retrievers: **BrowseComp-Plus [3]**
  - Large-scale ablation of retrieval and ranking design choices: **Revisiting Text Ranking [4]**

[1] Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.
[2] Li et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.
[3] Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.
[4] Meng, Ou, MacAvaney & Dalton. Revisiting Text Ranking in Deep Research. SIGIR 2026.

Long-horizon agentic search: Reasoning-Intensive QA

- **1,266 “inverted” questions:** start from a known fact, then write a question with many constraints that point to it
- **Filtered to be hard:** unsolved by GPT-4o (with and without browsing), o1 and an early Deep Research; not on the first page of five searches; not solvable by another person within ten minutes
- **Hard to find, easy to verify:** short, single, indisputable answers, so grading is simple
- **Human trainer performance:** with no AI assistants and about two hours per question, trainers gave up on 70.8% of questions (888 / 1,255)



Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.

Long-horizon agentic search: Reasoning-Intensive QA

HotpotQA: a two-hop question

Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

Answer: Chief of Protocol

- **Clear entities:** the question names the film and the character, so they can be searched directly
- **Fixed queries:** the reasoning path is known in advance, with few choices about what to search
- **Low information need:** a couple of supporting documents are enough to answer

BrowseComp: hard to find, easy to verify

Question (paraphrased): Which two teams played a football match between 1990 and 1994 that had a Brazilian referee, four yellow cards (two per team, three of them after half-time) and four substitutions, one of them for an injury in the first 25 minutes?

Answer: Republic of Ireland vs Romania (1990 World Cup, round of 16, referee José Roberto Wright)

- **Many weak constraints:** none names an entity to start from
- **Large candidate space:** every match 1990–94 with a Brazilian referee
- **Tens of searches:** reading match reports, rejecting and backtracking

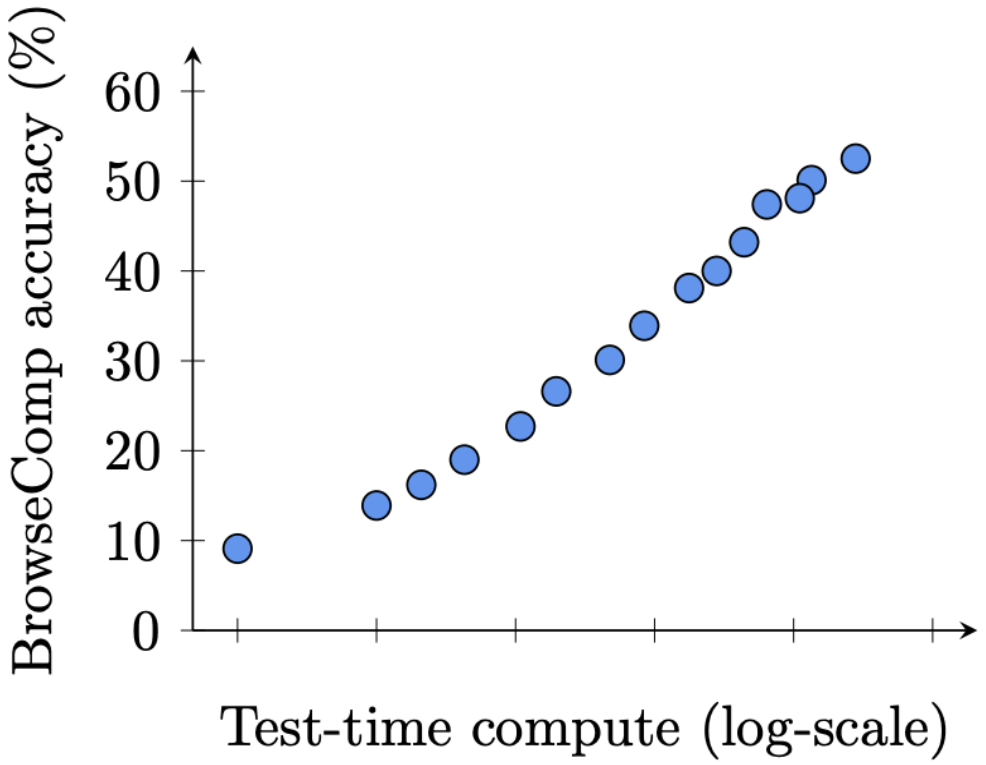
Yang et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. EMNLP 2018.
Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.

Long-horizon agentic search: Reasoning-Intensive QA

Table 3: Accuracy and calibration of OpenAI models on BrowseComp.

Model	Accuracy (%)	Calibration Error (%)
GPT-4o	0.6	69
GPT-4o w/ browsing	1.9	82
GPT-4.5	0.9	68
OpenAI o1	9.9	65
Deep Research[1]	51.5	91

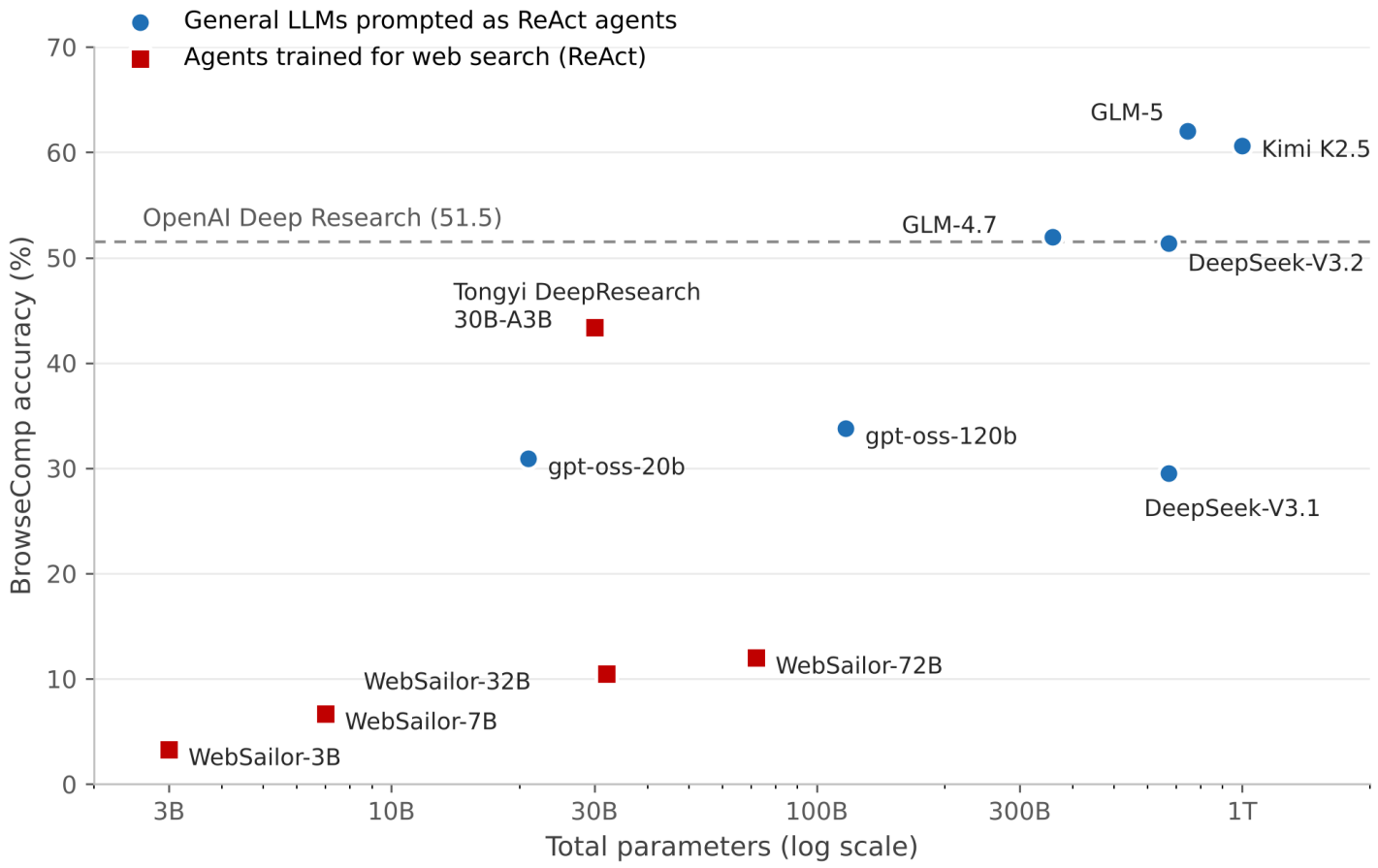
- Browsing alone is not enough: GPT-4o goes from 0.6% to only 1.9%
- Reasoning helps more than tools: o1 without browsing reaches 9.9%
- More browsing effort → higher accuracy, roughly linear in log compute
- Deep Research's large performance gap suggests the importance of dedicated training for long-horizon browsing, which earlier models largely lacked.



BrowseComp performance of an early version of OpenAI Deep Research with varying browsing effort. Accuracy scales smoothly with test-time compute.

Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.

Long-horizon agentic search: Reasoning-Intensive QA



Takeaways

- Several large open-source models now pass the original Deep Research performance
- Tongyi DeepResearch (30B, 43.4%) outperforms the much larger DeepSeek-V3.1 (671B, 29.5%), highlighting the potential importance of specialized training.
- Strong performance with smaller models remains difficult, leaving considerable room for more effective and efficient search agents.

* Results gathered from official model reports and published papers (single ReAct run on the live web; evaluation setups differ across sources).

Long-horizon agentic search: Reasoning-Intensive QA

- **Reasoning-intensive QA:** four works that go from defining the task, to training agents, to evaluating them under control
  - Benchmark that defines the task: BrowseComp [1]
  - **Training a deep research agent: WebSailor [2]**
  - Controlled evaluation with a local corpus and reproducible retrievers: BrowseComp-Plus [3]
  - Large-scale ablation of retrieval and ranking design choices: Revisiting Text Ranking [4]

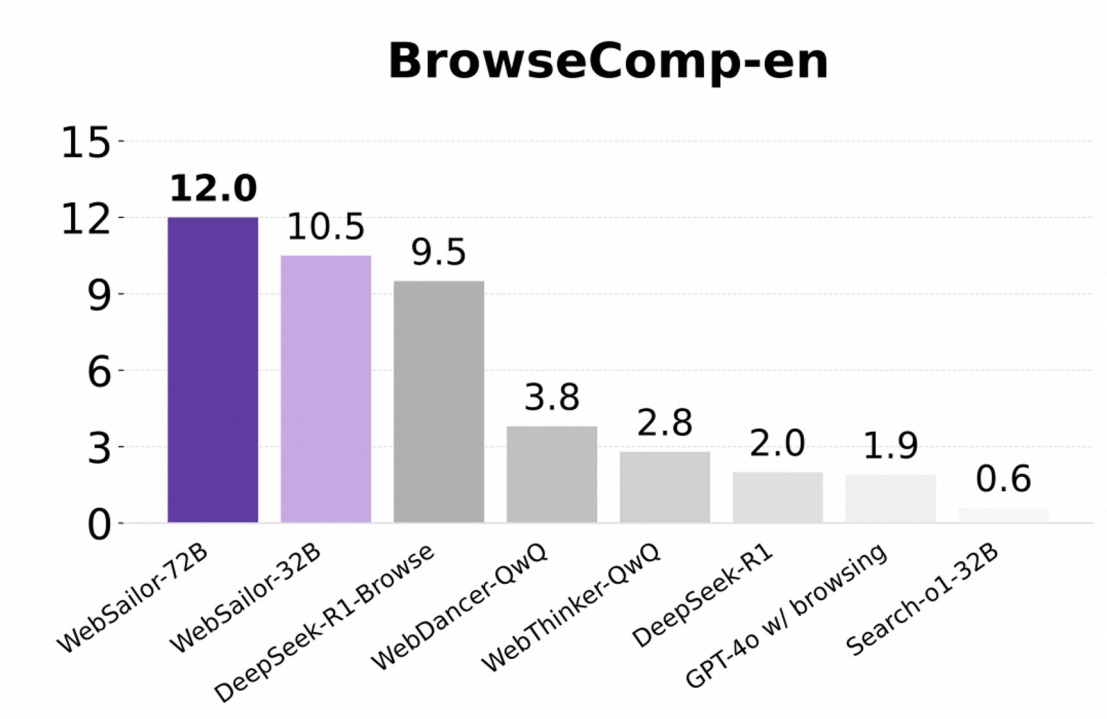
***BrowseComp** defines the task: questions that are hard to find but easy to verify. Its results also suggest that dedicated training for long-horizon browsing matters.*

Next: how to train such an agent, with WebSailor

[1] Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.
[2] Li et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.
[3] Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.
[4] Meng, Ou, MacAvaney & Dalton. Revisiting Text Ranking in Deep Research. SIGIR 2026.

Long-horizon agentic search: Reasoning-Intensive QA

- **Hypothesis:** strong deep research agents share one skill, systematically reducing extreme uncertainty across a vast information landscape; WebSailor is an end-to-end recipe to train it into open models
- **Synthetic data**
  - SailorFog-QA: Level 3 questions from random walks over rare entities, then obfuscated
  - Reasoning reconstruction: expert action–observation traces with short, rewritten thoughts
- **Training**
  - RFT cold start: ~2k filtered trajectories teach tool use and long-horizon structure
  - DUPO: agentic RL with duplicating sampling, 2–3× faster than DAPO

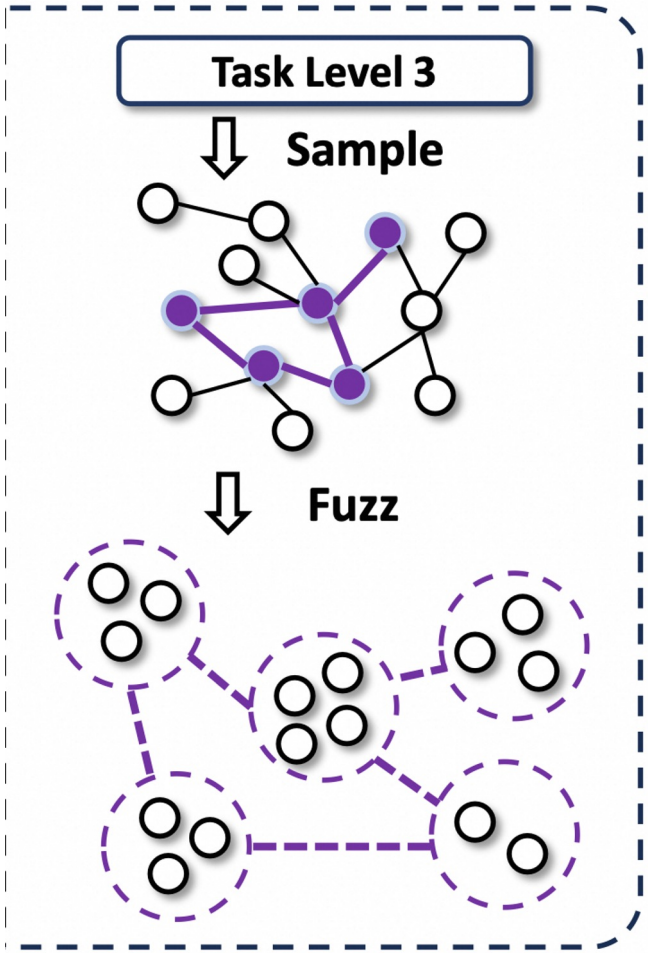


State-of-the-art performance against open-source models and agents at the time

Li, Zhang, Yin, Zhang, Ou et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.

Long-horizon agentic search: Reasoning-Intensive QA

- Three levels of information-seeking tasks
  - Level 1: low uncertainty; internal knowledge or one search
  - Level 2: high initial uncertainty but a clear path (multi-hop QA)
  - Level 3: high uncertainty that is hard to reduce; entities coupled in complex ways with no predefined reasoning path
- Shortcoming: existing training data stops at Level 2 (on WebDancer-QA, over half of expert trajectories need only 2 tool calls), while BrowseComp is dominated by Level 3
- SailorFog-QA: synthesised Level 3 questions
  - Sample: random walks from rare Wikidata entities grow dense entity graphs; sample subgraphs of coupled entities and write a question for each
  - Fuzz: obfuscate the clues ("in the early 2010s", "a market share below 1%") so no single lookup resolves them



Li, Zhang, Yin, Zhang, Ou et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.

Long-horizon agentic search: Reasoning-Intensive QA

- **Supervision data:** QwQ-32B solves SailorFog-QA; only successful traces are kept, and raw LRM thoughts are too verbose to imitate
- **Cold-start data:** just over 2k trajectories (correct answer, under 32k tokens, more than 5 tool calls); observations masked in the loss
- **RL reward:** 0.1 × format + 0.9 × answer (LLM judge), with GRPO-style group advantages and token-level loss

Reasoning reconstruction

Keep the expert's actions and observations, but rewrite each step's thought to be short and concise

→

Stage 1: RFT cold start

Fine-tune on a small set of filtered, successful trajectories, so the agent learns tool use before RL

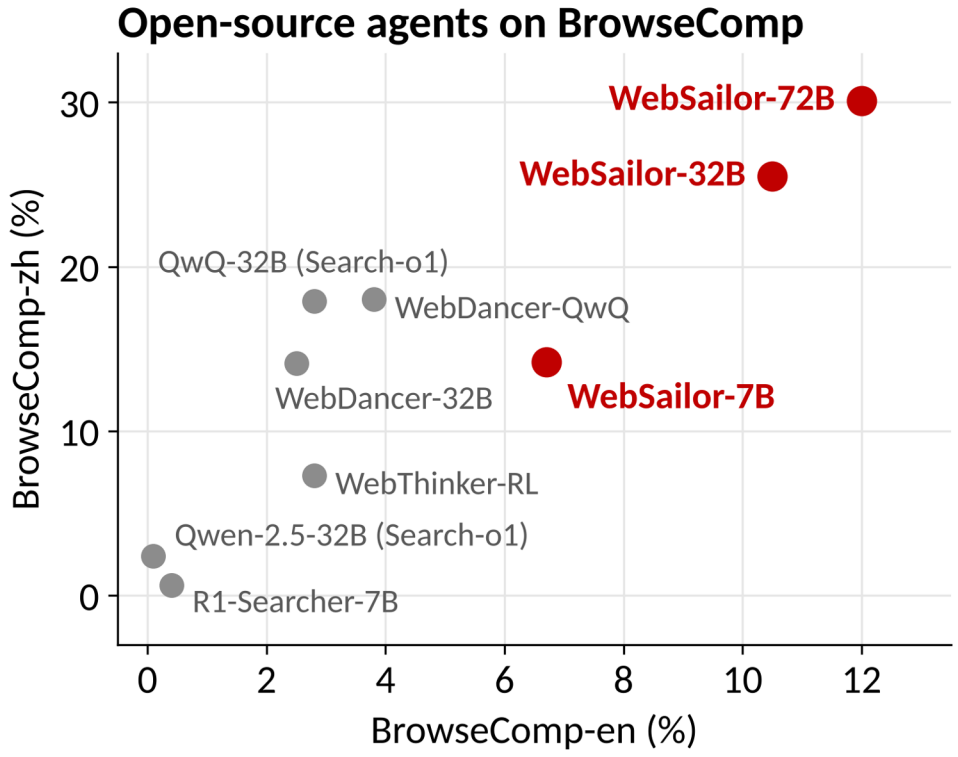
→

Stage 2: DUPO RL

Drop questions that are too easy; fill each batch by duplicating groups with mixed outcomes, 2–3× faster than DAPO

Li, Zhang, Yin, Zhang, Ou et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.

Long-horizon agentic search: Reasoning-Intensive QA



Model	SimpleQA pass@1 (%)
Agents with search	
WebSailor-72B	**93.5**
WebSailor-32B	**92.8**
WebDancer-QwQ	90.5
WebThinker-RL	77.5
DeepSeek-R1 (ReAct)	72.2
Direct inference	
GPT-4.1	41.6
GPT-4o	38.2
DeepSeek-R1	27.8

WebSailor is the best open-source agent on BrowseComp

Performance generalises to simpler tasks: single-fact lookup on SimpleQA

Li, Zhang, Yin, Zhang, Ou et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.
Wei et al. Measuring Short-form Factuality in Large Language Models (SimpleQA). arXiv 2024.

Long-horizon agentic search: Reasoning-Intensive QA

- **Reasoning-intensive QA:** four works that go from defining the task, to training agents, to evaluating them under control
  - Benchmark that defines the task: BrowseComp [1]
  - Training a deep research agent: WebSailor [2]
  - **Controlled evaluation with a local corpus and reproducible retrievers: BrowseComp-Plus [3]**
  - Large-scale ablation of retrieval and ranking design choices: Revisiting Text Ranking [4]

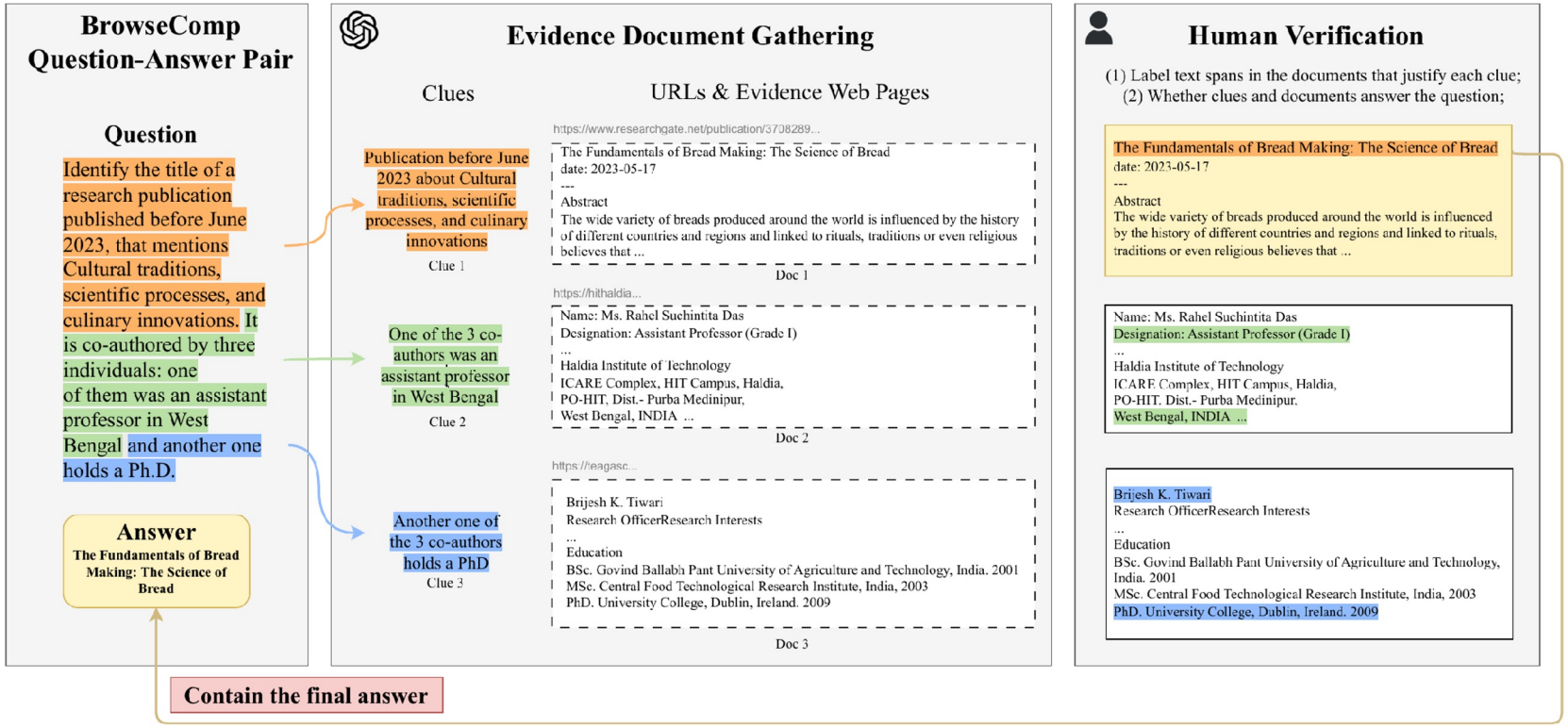
*So far, both **BrowseComp** and **WebSailor** are evaluated on the live web, where search results are opaque and change over time.*

Next: a controlled, reproducible setup with BrowseComp-Plus

[1] Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.
[2] Li et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.
[3] Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.
[4] Meng, Ou, MacAvaney & Dalton. Revisiting Text Ranking in Deep Research. SIGIR 2026.

Long-horizon agentic search: Reasoning-Intensive QA

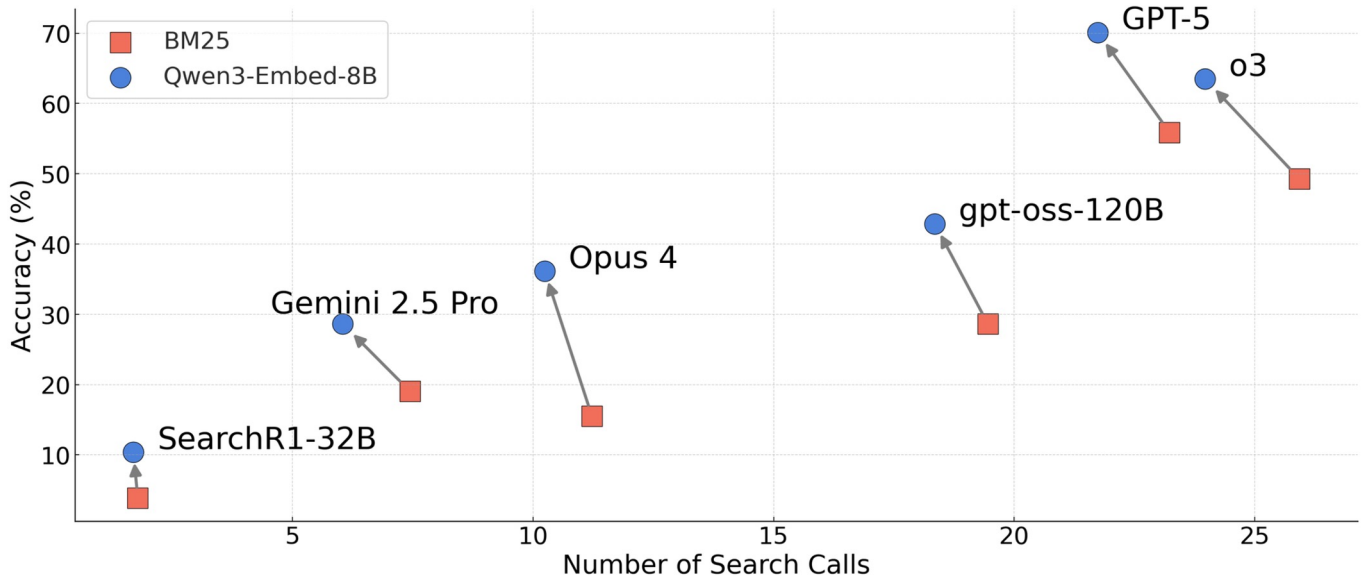
- 830 BrowseComp questions, each human-verified with answer-supporting evidence
- Fixed corpus of 100,195 documents: evidence, gold documents and hard negatives
- Human relevance labels let us evaluate the retriever separately from the agent



Evidence collection and human verification

Long-horizon agentic search: Reasoning-Intensive QA

- Same BM25 retriever, different agents
  - Search-R1: 3.9% vs GPT-5: 55.9%
- Same GPT-5 agent, different retrievers
  - BM25: 55.9% accuracy, 23.2 calls
  - Qwen3-Embed-8B: 70.1% accuracy, 21.7 calls
- Both the agent and the retriever matter



Accuracy vs. number of search calls, BM25 (red) vs Qwen3-Embed-8B (blue)

Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.

Long-horizon agentic search: Reasoning-Intensive QA

- **Reasoning-intensive QA:** four works that go from defining the task, to training agents, to evaluating them under control
  - Benchmark that defines the task: BrowseComp [1]
  - Training a deep research agent: WebSailor [2]
  - Controlled evaluation with a local corpus and reproducible retrievers: BrowseComp-Plus [3]
  - **Large-scale ablation of retrieval and ranking design choices: Revisiting Text Ranking [4]**

BrowseComp-Plus *shows that the retriever alone matters: with GPT-5 fixed, a stronger retriever lifts accuracy from 55.9% to 70.1%.*

Next: a larger-scale ablation of which retrieval and ranking choices matter

[1] Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.
[2] Li et al. WebSailor: Navigating Super-human Reasoning for Web Agent. arXiv 2025.
[3] Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.
[4] Meng et al. Revisiting Text Ranking in Deep Research. SIGIR 2026.

Long-horizon agentic search: Reasoning-Intensive QA

Revisiting Text Ranking in Deep Research

- Deep research agents usually search through black-box web search APIs
- So it is unclear how established text ranking methods behave when the searcher is an LLM agent
- Replace the API with transparent retrievers and re-rankers; tested on **BrowseComp-Plus**

Three research questions

- **RQ1 Retrievers:** which retrievers work best for agent-issued queries, on passages and documents?
- **RQ2 Re-ranking:** does re-ranking still help when the reader is an agent?
- **RQ3 Query mismatch:** agents issue keyword queries, rankers are trained on questions. Does this hurt?

Meng et al. Revisiting Text Ranking in Deep Research. SIGIR 2026.
Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.

Long-horizon agentic search: Reasoning-Intensive QA

RQ1 BM25 performs best on passages

Contrary to BrowseComp-Plus, where BM25 on full documents was the weakest retriever

Answer accuracy (%), gpt-oss-20b agent

Retriever	Passages	Documents
BM25	**57.2**	25.9
ColBERTv2	52.1	48.1
SPLADE-v3	51.6	47.6
Qwen3-Embed-8B	41.7	42.1
RepLLaMA	40.6	36.3

Agent queries look like web-search keywords

Agent	Search query
gpt-oss-20b	“90+7” attendance 61700 “Man United” “4-1” “90+4”
GLM-4.7-Flash	“61,888” football match Stockholm Vienna Prague “goal in the 6th minute” football

- Keyword and quoted-phrase queries suit lexical matching (and multi-vector ColBERTv2)
- BM25 is weak on documents because of length normalisation: indexing only the first 512 tokens doubles accuracy (**25.9 → 51.3**)

Meng et al. Revisiting Text Ranking in Deep Research. SIGIR 2026.
Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. ACL 2026.

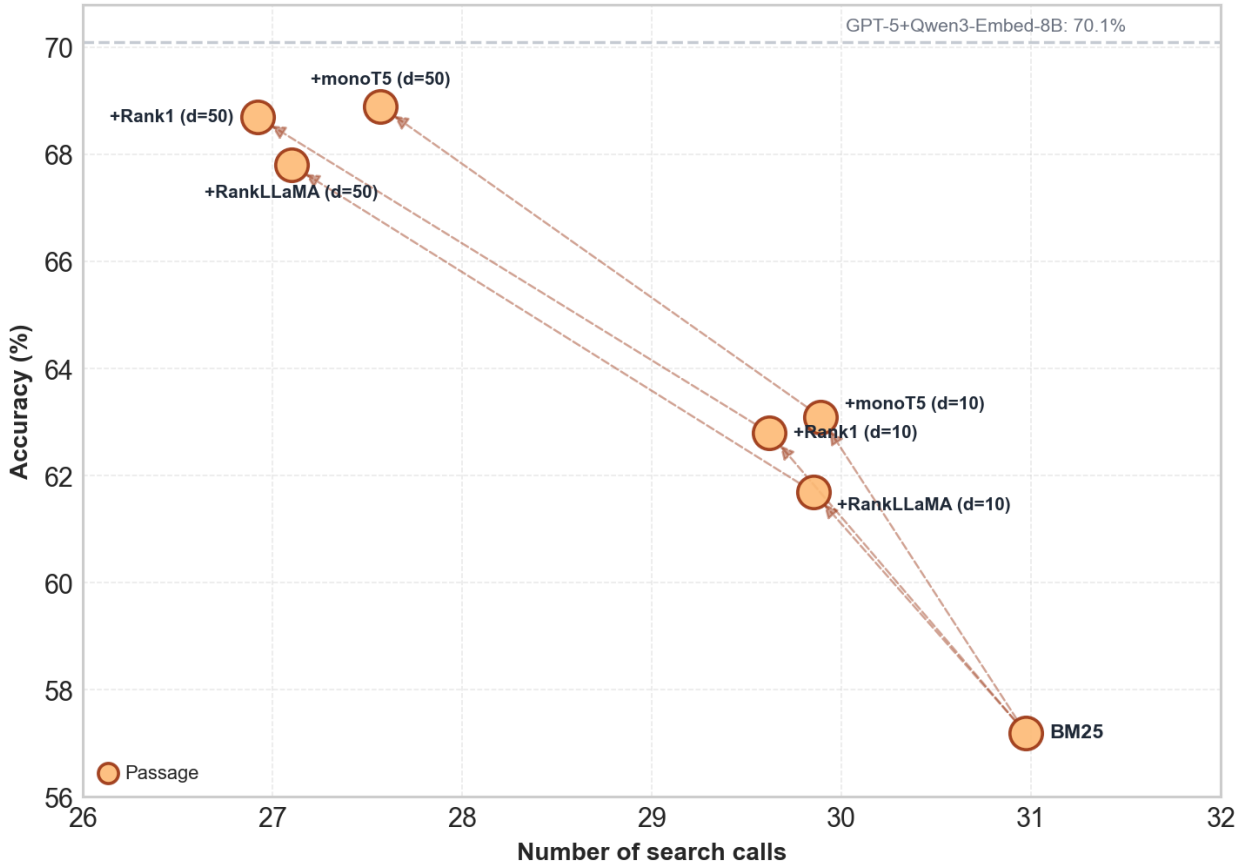
Long-horizon agentic search: Reasoning-Intensive QA

RQ2 Re-ranking helps every retriever, and deeper is better

Answer accuracy (%) after re-ranking passages, gpt-oss-20b

Re-ranker (depth)	BM25	SPLADE-v3	Qwen3-Embed-8B
None	57.2	51.6	41.7
monoT5 (10)	63.1	59.9	47.1
monoT5 (50)	**68.9**	64.6	55.9
RankLLaMA (50)	67.8	**66.3**	**56.8**
Rank1 (50)	68.7	64.3	56.4

- All three re-rankers give similar gains; reasoning (Rank1) shows no clear edge
- BM25 + monoT5 (50) reaches **68.9%**, close to GPT-5 + Qwen3-Embed-8B (70.1%), with fewer search calls



Accuracy vs. search calls, BM25 + re-rankers on passages. Dashed line: GPT-5 + Qwen3-Embed-8B

Meng et al. Revisiting Text Ranking in Deep Research. SIGIR 2026.

Long-horizon agentic search: Reasoning-Intensive QA

RQ3 Rewriting agent queries as questions helps neural rankers

Q2Q turns each agent query into a natural-language question; Q+R also uses the agent's latest reasoning

Example

Query	Text
Raw	"61,880" football attendance
Q2Q (Q)	What is the football attendance number 61,880?
Q2Q (Q+R)	What football match had an attendance of 61,880?

Answer accuracy (%), gpt-oss-20b, passages

Ranker	Raw	Q2Q (Q)	Q2Q (Q+R)
BM25	57.2	**57.8**	55.7
SPLADE-v3	51.6	51.0	**55.7***
Qwen3-Embed-8B	41.7	40.4	**45.9***

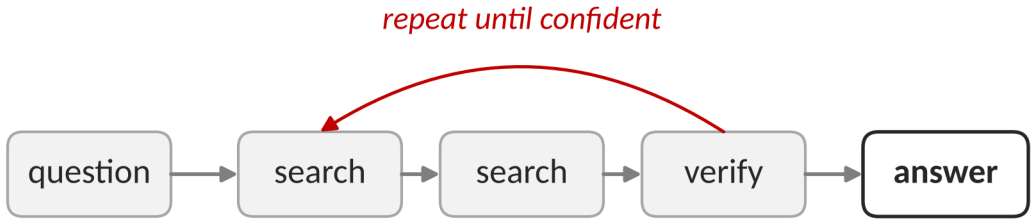
* significant gain over raw queries ($p < 0.05$)

- Using the reasoning keeps the agent's intent; rewriting the query alone can shift it
- Q2Q (Q+R) significantly helps neural retrievers; BM25 does not benefit, as keyword queries already suit it

Long-horizon agentic search: Report generation

Reasoning-intensive QA

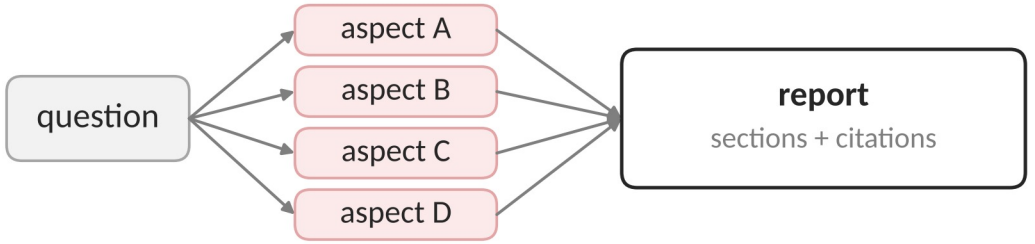
- Short-form, verifiable answers
- Targeted search for specific evidence
- Deep exploration, repeated verification



e.g. *Which fictional character breaks the fourth wall, was helped by selfless ascetics, and had a 1960s–80s TV show with < 50 episodes?* → **Plastic Man**

Report generation

- Broad, open-ended research questions
- No correct answer, no fixed research path
- Decide what to cover, how deep, when to stop



e.g. *Analyse the feasibility of investing in EV charging infrastructure in suburban areas.* → **cited report**

Outline DeepResearch Bench, ResearcherBench → WebWeaver → What benchmarks miss

QA example from BrowseComp: Wei et al. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv 2025.

Long-horizon agentic search: Report generation

DeepResearch Bench [1] **Report quality and citation quality, judged by LLMs**

100 PhD-level tasks (50 EN, 50 ZH) · 22 fields · topic mix from 96k real user queries

- Deep research agents write better reports than LLMs with search
- **More citations ≠ more accurate citations**
  - Gemini Deep Research cites by far the most (~165 supported statements per task), at 78% accuracy
  - GPT-4o Search Preview is the most accurate (87%) but supports only ~5 statements per task
- **RACE**: a judge LLM writes weighted criteria per task and scores the report against a reference (50 = on par)
  - Agrees with experts more often than experts agree with each other (71% vs 68% pairwise agreement)
  - The reference matters most: without it, agreement drops to 67%; a plain judge prompt reaches only 59%

System	RACE	Cit. acc. (%)	Eff. cit.
Gemini-2.5-Pro Deep Research	**49.71**	78.30	**165.34**
OpenAI Deep Research	46.45	75.01	39.79
Perplexity Research	40.46	82.63	31.20
Grok Deeper Search	38.22	73.08	8.58
GPT-4o Search Preview	30.74	**86.63**	5.05

[1] Du et al. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. ICLR 2026. Scores from the official leaderboard; agreement figures from the paper.

Long-horizon agentic search: Report generation

ResearcherBench [1] Frontier research questions, judged against expert rubrics

65 AI research questions · 35 subjects · from lab discussions, researcher interviews and forums

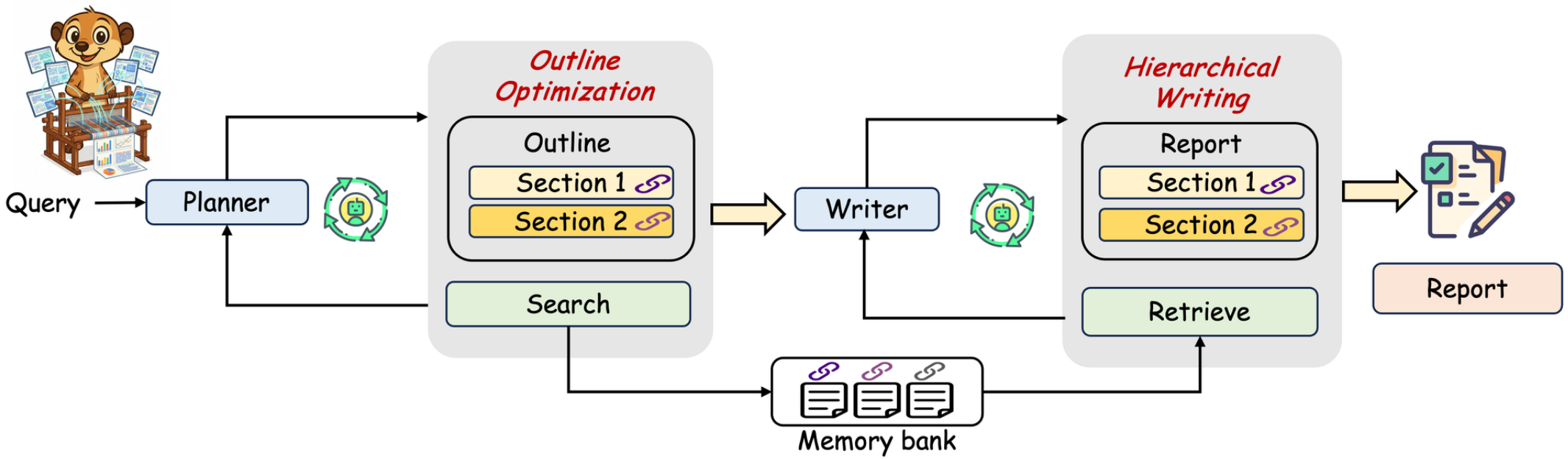
- **Coverage**: experts write rubric items per question (weight 1–3, nice-to-have → core); a judge LLM checks each
  - OpenAI and Gemini Deep Research lead clearly (0.70 and 0.69); no other system passes 0.48
- Cited claims are mostly right, but **most claims are uncited**
  - Faithfulness (supported / cited claims) is 0.84–0.86 for most systems; groundedness (cited / all claims) only 0.32–0.59
  - Systems lean on internal knowledge; citing more ≠ better insight (OpenAI: top coverage, but groundedness only 0.34)
- All systems do best on open consulting (33 of 65 questions), less well on literature review and technical details
  - Top systems exceed 0.75 coverage there: strong ideation partners, weaker as precise literature or technical tools

System	Coverage	Faithful.	Grounded.
OpenAI Deep Research	**0.70**	0.84	0.34
Gemini Deep Research	0.69	**0.86**	**0.59**
Perplexity Deep Research	0.48	0.85	0.56
Grok3 DeepSearch	0.44	0.69	0.32
GPT-4o Search Preview	0.36	**0.86**	0.39

[1] Xu et al. ResearcherBench: Evaluating Deep AI Research Systems on the Frontiers of Scientific Inquiry. arXiv 2025.

Long-horizon agentic search: Report generation

*Both benchmarks reward **coverage, structure and grounded citations**. WebWeaver [1] puts a dynamic outline at the centre of planning and writing.*



- **Planner:** search ⇄ outline; sections cite a memory bank
- **Writer:** section by section, only cited evidence

DeepResearch Bench	RACE	Cit. acc.	Eff. cit.
WebWeaver (Claude-sonnet-4)	**50.58**	**93.37**	**200.75**
Gemini-2.5-Pro Deep Research	49.71	78.30	165.34

[1] Li et al. WebWeaver: Structuring Web-Scale Evidence with Dynamic Outlines for Open-Ended Deep Research. ICLR 2026. Figure and results adapted from [1].

Long-horizon agentic search (deep research)

Q & A

Long-horizon agentic search: Context management

- **Challenges specific to long horizons:** beyond finding the right evidence, long trajectories have to stay tractable
  - **Context management:** preserving useful information across long trajectories (ReSum [1], MEM1 [2], Context-Folding [3])
  - Efficiency: reducing unnecessary searching, reading and reasoning (SlimSearcher, OTC, BATS)

*We have now seen the major tasks, **reasoning-intensive QA and report generation**, and how agents are trained and evaluated for them. Long trajectories also bring difficulties of their own.*

Next: keeping the agent's working context under control

[1] Wu et al. ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization. 2026.
[2] Zhou et al. MEM1: Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026.
[3] Sun et al. Scaling Long-Horizon LLM Agent via Context-Folding. ICML 2026.

Long-horizon agentic search: Context management

- **ReAct appends everything:** question → (thought, action, observation) × N
- **Problems:** the history overflows the window (WebSailor-7B passes 32k tokens on hard BrowseComp questions [1]), buries evidence among dead ends, and is re-read at every step
- **Goal:** keep a compact working state with only what the agent needs to reason next

Three strategies that differ in when the context is compressed

ReSum [1]

When the budget is reached

Run ReAct until the context nears its limit, then summarise and restart from the summary

→

MEM1 [2]

Every turn

Rewrite one internal state at each step, merging new evidence and dropping the rest

→

Context-Folding [3]

Proactively, by the agent

Branch into a subtask when the agent chooses; fold it into a short summary on return

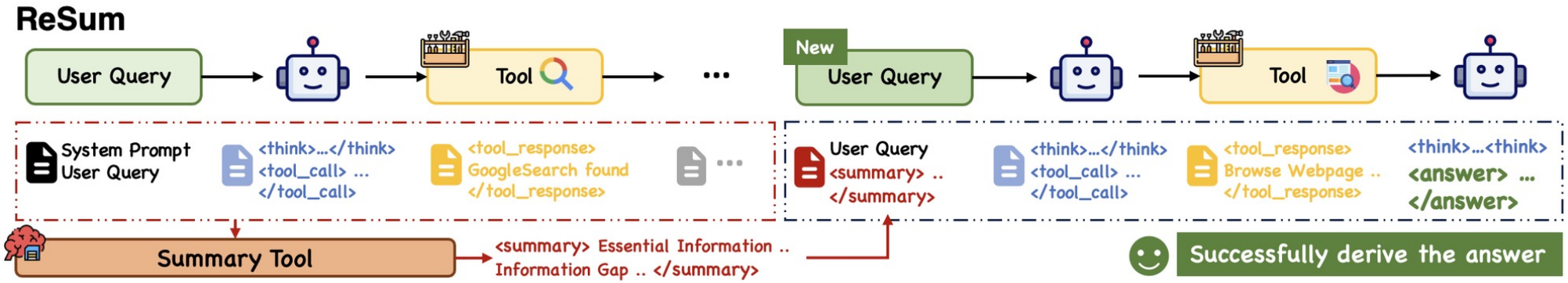
From passive to proactive: the agent state turns from an ever-growing log into a workspace the agent manages

[1] Wu et al. ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization. 2026.
[2] Zhou et al. MEM1: Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026.
[3] Sun et al. Scaling Long-Horizon LLM Agent via Context-Folding. ICML 2026.

Long-horizon agentic search: Context management

ReSum [1] Compress when the context budget is reached

- **Trigger:** when the context nears its limit (e.g., 32k tokens), a summary tool condenses the history; the agent restarts from the question plus the summary
- **Summary tool:** ReSumTool-30B, Qwen3-30B-A3B fine-tuned on ~9K conversation–summary pairs, keeps key evidence, flags information gaps and suggests next steps
- **ReSum-GRPO:** split each rollout at its summary points and give every segment the final-answer advantage, so the agent learns to reason from summaries (1K training questions)
- **Results:** training-free ReSum beats ReAct by 4.5% on average (GAIA, BrowseComp-zh, BrowseComp); ReSum-GRPO adds a further 8.2%
- **Trade-off:** plug-and-play for existing ReAct agents, but compression is forced by the budget rather than chosen by the agent

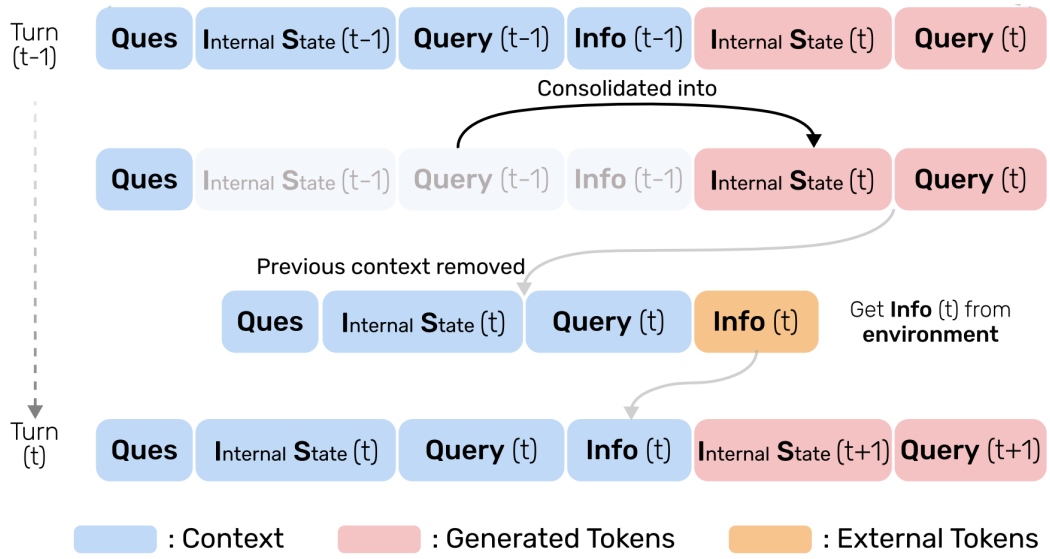


[1] Wu et al. ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization. 2026. Figure from [1].

Long-horizon agentic search: Context management

MEM1 [1] Compress at every turn

- **Trigger:** every turn, the agent writes a new internal state <IS> that merges its previous state with the latest query and result
- **Prune:** earlier turns are then dropped; the context keeps only the question, the latest <IS>, and the current action and observation
- **Training:** end-to-end RL (PPO) rewarded only on task success; no memory reward, consolidation emerges
- **Masked trajectory:** a 2D attention mask replays each turn's pruned context in one sequence, so policy gradients stay valid
- **Results:** trained on 2-question tasks and tested on 16, MEM1-7B reaches 3.5× the accuracy of Qwen2.5-14B-Instruct with 3.7× lower peak memory
- **Related:** WebResearcher also rebuilds a compact workspace every round [2]

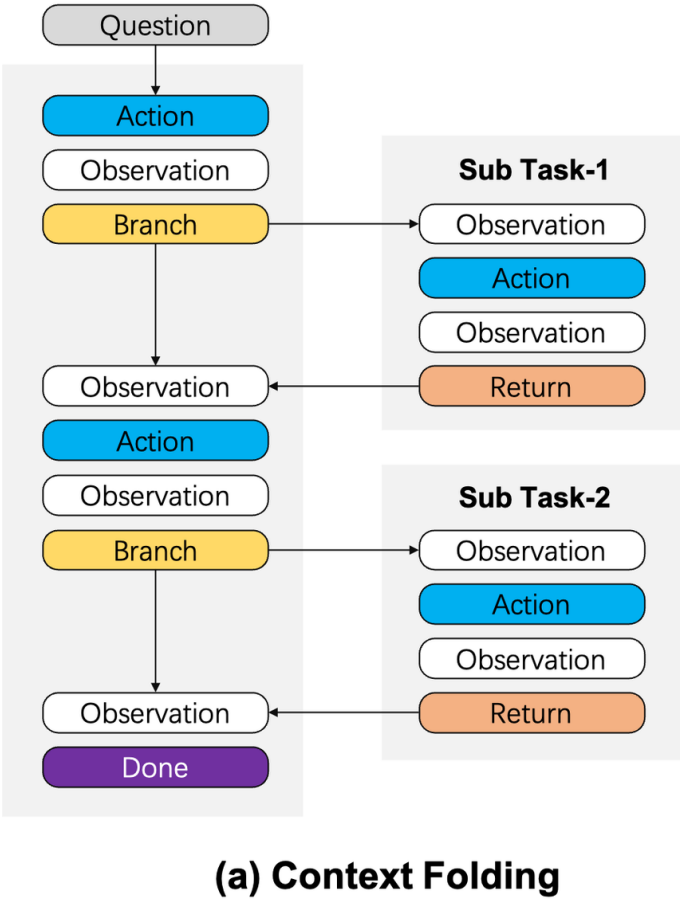


[1] Zhou et al. MEM1: Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026. Figure from [1].
[2] Qiao et al. WebResearcher: Unleashing Unbounded Reasoning Capability in Long-Horizon Agents. arXiv 2025.

Long-horizon agentic search: Context management

Context-Folding [1] **Compress proactively, when the agent decides**

- **Trigger:** the agent itself decides when to branch into a subtask and what to bring back
- **Two tools:** branch(description, prompt) opens a sub-trajectory for a subtask; return(message) closes it with a short summary of the outcome
- **Fold:** on return, the sub-trajectory's intermediate steps are removed; the main thread keeps only the branch call and the returned summary
- **Working context:** the main thread holds high-level reasoning and outcomes, while long tool outputs and dead ends stay inside branches
- **Result:** matches or beats ReAct on deep research and coding tasks with an active context 10× smaller



(a) Context Folding

[1] Sun et al. Scaling Long-Horizon LLM Agent via Context-Folding. ICML 2026. Figure from [1].

Long-horizon agentic search: Efficiency

- **Challenges specific to long horizons:** beyond finding the right evidence, long trajectories have to stay tractable
  - Context management: preserving useful information across long trajectories (ReSum, MEM1, Context-Folding)
  - **Efficiency:** reducing unnecessary searching, reading and reasoning (SlimSearcher [1], OTC [2], BATS [3])

*Context management keeps long trajectories within the context limit, but it **does not by itself make the agent search less**; by lifting the length ceiling, it can even let agents run more turns.*

Next: doing the same job with fewer search and reasoning steps

[1] Xie et al. SlimSearcher: Training Efficiency-Aware Web Agents via Adaptive Reward Gating. arXiv 2026.
[2] Wang et al. OTC: Optimal Tool Calls via Reinforcement Learning. arXiv 2025.
[3] Liu et al. Budget-Aware Tool-Use Enables Effective Agent Scaling. arXiv 2025.

Long-horizon agentic search: Efficiency

Training time Learn to reach the answer with fewer steps

- **SlimSearcher [1]:** SFT keeps trajectories that are both correct and economical; in RL, reward = correctness gate × tool-call and token multipliers, measured against the cheapest correct rollout in the group
- **OTC [2]:** RL reward scales correctness by how close the tool-call count is to the fewest calls that solved the question
- **Key idea:** efficiency is judged relative to what is achievable for each question, so the agent is not pushed to stop searching too early

SlimSearcher [1] Tongyi-DeepResearch-30B, acc. (%) / rounds

Benchmark	Base	SlimSearcher
GAIA	68.2 / 20.6	**70.9 / 10.6**
BrowseComp	41.0 / 63.7	**44.7 / 47.6**
xbench-DeepSearch	71.3 / 14.3	**79.0 / 5.9**
HLE	35.8 / 23.9	**37.6 / 19.8**

OTC [2] Qwen2.5-7B, EM (%) / tool calls

Benchmark	Search-R1	OTC
NQ	39.9 / 1.70	**44.4 / 0.99**
HotpotQA	34.1 / 2.11	**36.6 / 1.01**

Both GRPO-trained. Across models and settings, OTC cuts tool calls by up to 68.3% at comparable accuracy

[1] Xie et al. SlimSearcher: Training Efficiency-Aware Web Agents via Adaptive Reward Gating. arXiv 2026.
[2] Wang et al. OTC: Optimal Tool Calls via Reinforcement Learning. arXiv 2025.

Long-horizon agentic search: Efficiency

Test time **Spend a fixed tool budget wisely, without retraining [1]**

- **Budget Tracker:** a prompt-level plug-in that shows used and remaining calls for each tool at every step; plain ReAct stops early and leaves most of a large budget unused
- **BATS:** plans with a checklist of exploration and verification clues; a verifier then decides to dig deeper, pivot or stop, based on the remaining budget
- **Efficiency:** with only 10 calls, Budget Tracker matches ReAct at 100 calls on BrowseComp (12.8% vs 12.6%), with 40% fewer searches and 31% lower cost

Accuracy (%) with a budget of 100 calls per tool

Model / method	BrowseComp	BrowseComp-ZH	HLE-Search
Gemini-2.5-Pro ReAct	12.6	31.5	20.5
Gemini-2.5-Pro + Budget Tracker	14.6	32.9	21.8
Gemini-2.5-Pro BATS	**24.6**	**46.0**	**27.0**
Claude-Sonnet-4 ReAct	12.2	29.1	20.5
Claude-Sonnet-4 + Budget Tracker	14.0	31.1	23.0
Claude-Sonnet-4 BATS	**19.1**	**41.5**	**29.0**

[1] Liu et al. Budget-Aware Tool-Use Enables Effective Agent Scaling. arXiv 2025.

Long-horizon agentic search (deep research)

Q & A

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

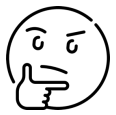
BREAK 16:00–16:30

PART II 16:30–18:30

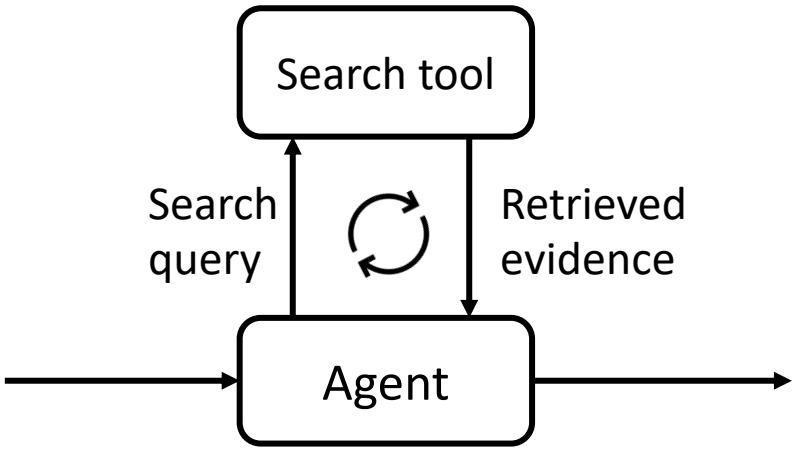
Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- **Evaluation for agentic search**
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- Conclusion and open challenges

Evaluation for agentic search



I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.



Evidence



Short answer

Brighton

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Evaluation for agentic search

I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.





How good are the evidence & answers?

Evidence



Short answer

Brighton

Report

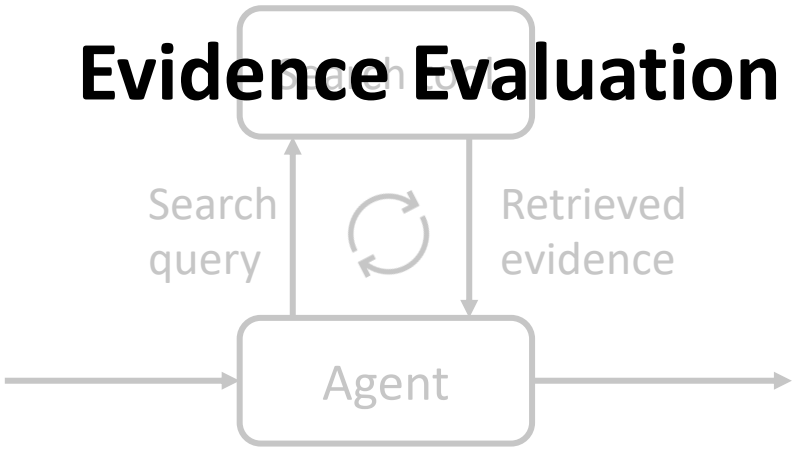
Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Evaluation for agentic search

I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.

Evidence Evaluation



Evidence



Short answer

Brighton

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Evidence Evaluation

- Why? We want responses that are grounded in facts from the corpus, not from the agent's memory.
  - Even if evidence exists in the parametric memory, we want to make sure it's still valid!

- Typical IR systems evaluate ranked lists and (usually) prioritize valuable documents at higher ranks:



  - Based on user browsing behavior: users grow impatient the more they have to read.

Evidence Evaluation

- Agentic search systems are different: multiple sets of search results contribute to a growing context:

```
<input>prompt</input>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>
```

1.  2.  3.  4.  5. 

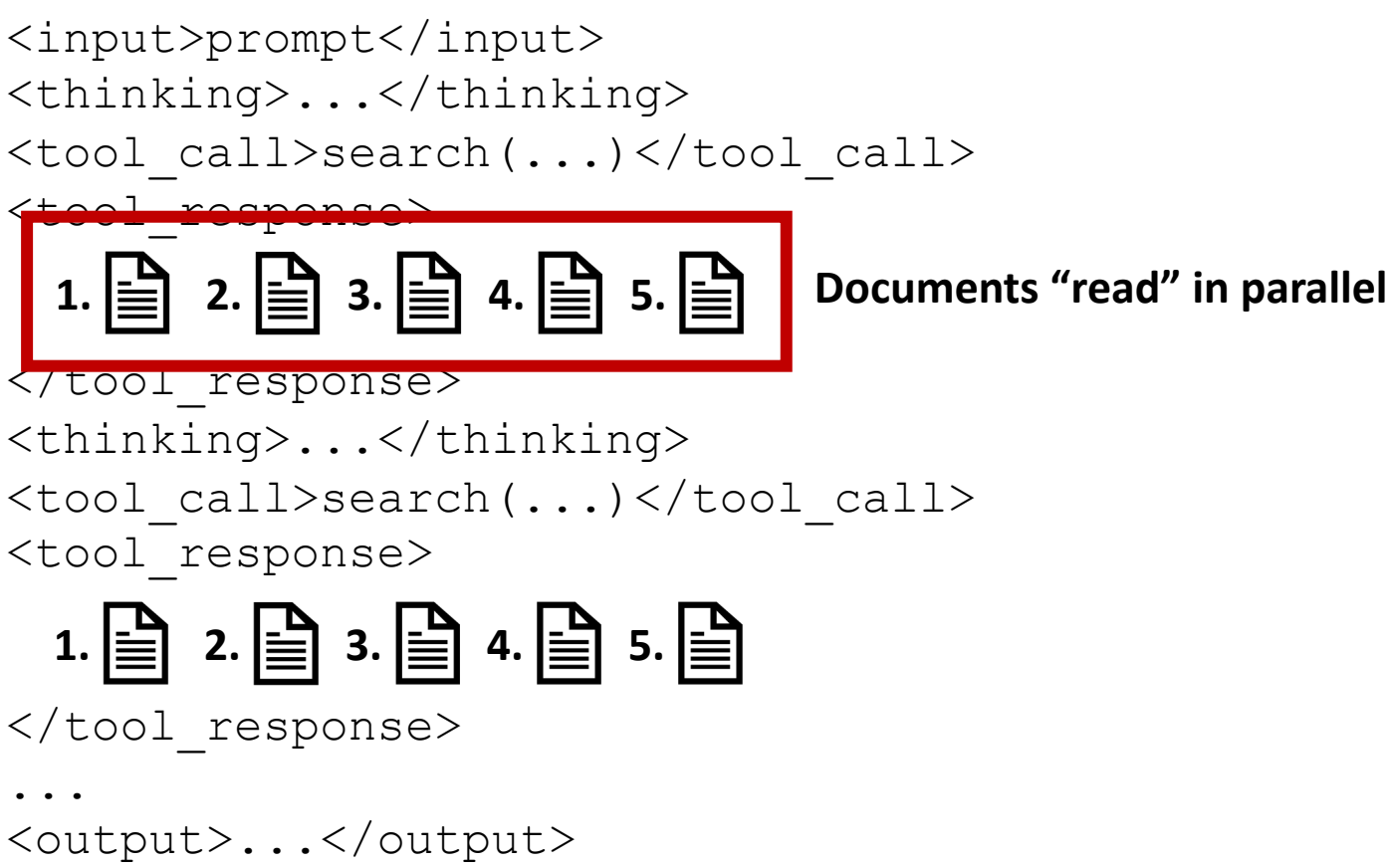
```
</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>
```

1.  2.  3.  4.  5. 

```
</tool_response>
...
<output>...</output>
```

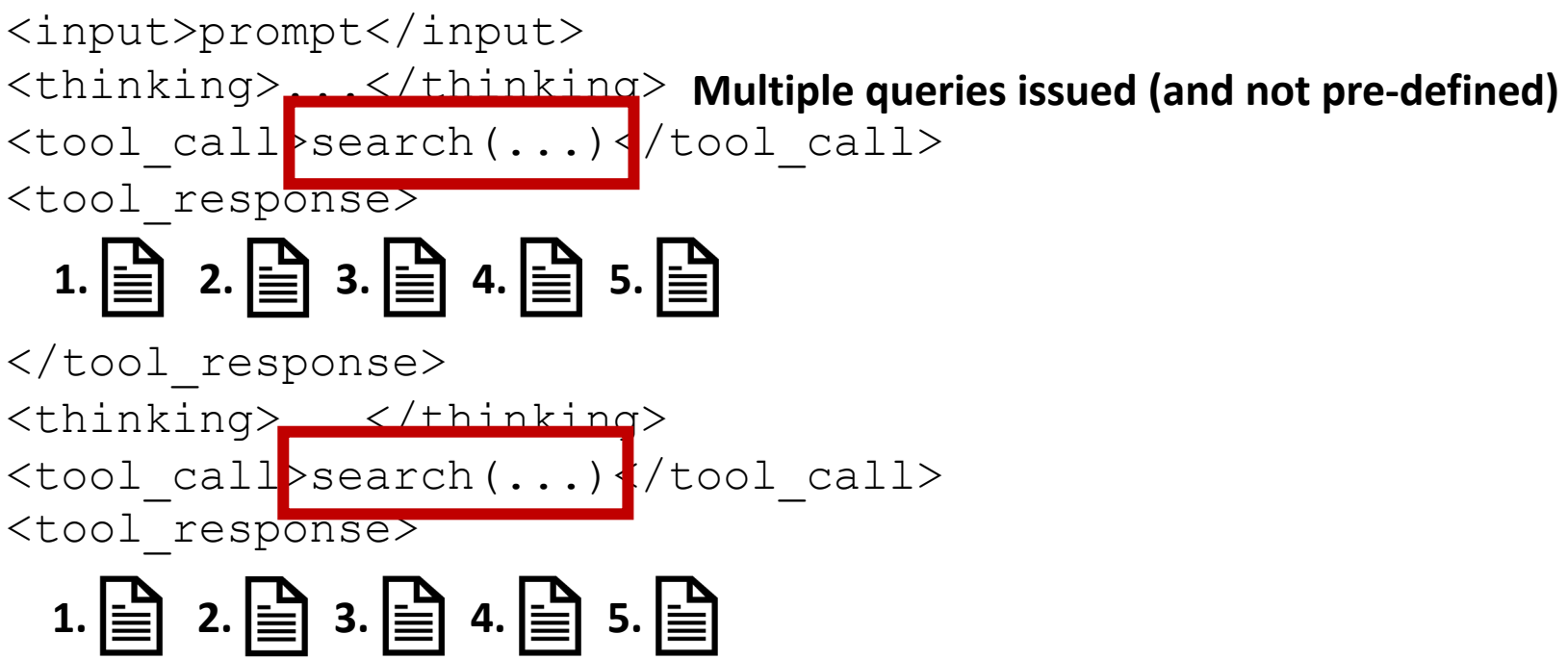
Evidence Evaluation

- Agentic search systems are different: multiple sets of search results contribute to a growing context:



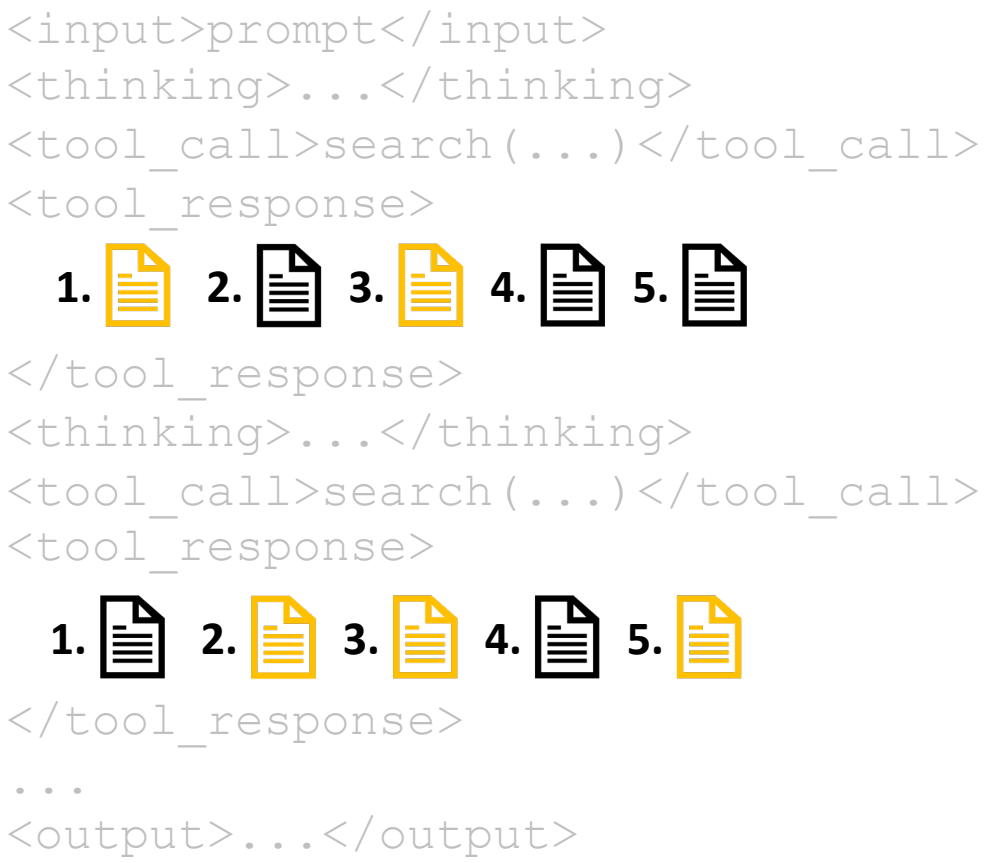
Evidence Evaluation

- Agentic search systems are different: multiple sets of search results contribute to a growing context:



Evidence Evaluation

- Therefore, we often measure evidence as cumulative Recall – the percent of all documents that support an answer retrieved at any stage.



Not retrieved but relevant:



relevant

non-relevant

Recall = 5/8 = 0.625

Evidence Evaluation

- Alternative: subtopic recall – the proportion of subtopics that you retrieved any documents for.
  - Main idea: you don't need *all* evidence for each subtopic, just *some* evidence.

<input>prompt</input>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>

1.  2.  3.  4.  5. 

</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>

1.  2.  3.  4.  5. 

</tool_response>
...
<output>...</output>

Not retrieved but relevant:

✓ Brighton is near London
✓ Brighton is on the sea
✗ Brighton has good weather
✗ Brighton has cheap trains

ST Recall = 2/4 = 0.5

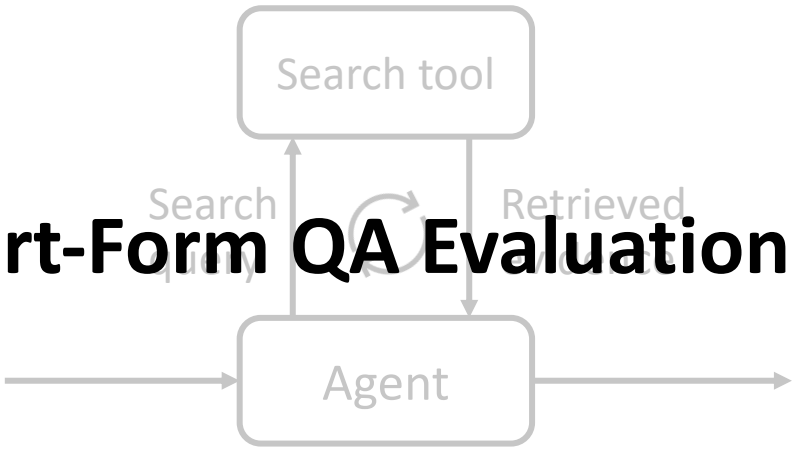
Evidence Evaluation

Limitations:

- What counts as a “relevant” document?
  - Human assessed? Contains the final answer? Contains a nugget from the final answer? LLM-Eval?
  - How do we ensure completeness of the judgements (long-standing IR problem)?
- Unbounded recall favors systems that consume lots of results. (More on this later.)
- No proof the agent uses this information, just that it’s exposed to it.
- Evidence recall is not necessarily correlated with downstream effectiveness.

Evaluation for agentic search

I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.



Short-Form QA Evaluation

Evidence



Short answer
Brighton

Report
Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Short-Form Answer Evaluation

- In some settings, a concise answer suffices
- Benefits: direct, simple evaluation
- Limitations: not suitable for all tasks, often not as helpful to the user as long-form responses
- Key challenge: even concise answers can be expressed in many ways:

☑	**Short answer** Brighton
☑	**Short answer** Brighton, UK
☒	**Short answer** Brighton, Queensland
☑	**Short answer** Brighton and Hove
☒	**Short answer** Brighton and Leeds

Short-Form Answer Evaluation

- Classic lexical evaluation:
  - Based on a set of “correct” answers per question…
  - Exact Match (EM) – does the response match an answer exactly?
  - “Accuracy” (sometimes called contains) – does the response contain an answer exactly?
  - F1 – harmonic mean of token-level precision and recall
- Benefits: Fast, deterministic
- Limitations: Still brittle against variations

Known Ground Truth Answers
Brighton
Brighton, UK

Response	EM	Accuracy	F1
☑ Brighton	1.00	1.00	1.00
☑ Brighton, UK	1.00	1.00	1.00
☒ Brighton, Queensland	0.00	1.00	0.67
☑ Brighton and Hove	0.00	1.00	0.50
☒ Brighton and Leeds	0.00	1.00	0.50

Short-Form Answer Evaluation

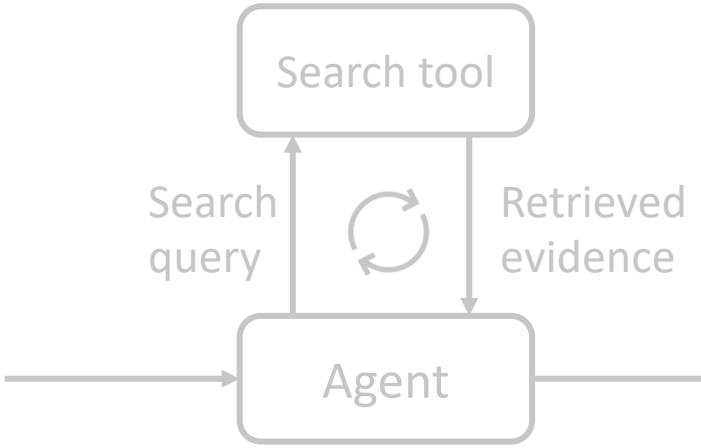
- Model-based evaluation:
  - Also based on a set of “correct” answers per question (but usually just one)…
  - LLM-as-a-judge: Ask an LLM if the response matches the answer(s)
  - BERTScore: Use a trained BERT model to measure answer similarity
- Benefit: More accepting of semantically-equivalent variations
- Limitations: Model noise, slow

Known Ground Truth Answers
Brighton
Brighton, UK

Response	qwen3.8-27b	BERTScore
☑ Brighton	1.00	1.00
☑ Brighton, UK	1.00	1.00
☒ Brighton, Queensland	0.00	0.92
☑ Brighton and Hove	1.00	0.82
☒ Brighton and Leeds	0.00	0.87

Evaluation for agentic search

I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.



Evidence



Short answer

Brighton

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Long-form Report Evaluation

Long-form Report Evaluation

Responses in the format of a “report” offer a more comprehensive answer for the user:

User sees *why* it’s the answer

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Long-form Report Evaluation

Responses in the format of a “report” offer a more comprehensive answer for the user:

Allows user to draw a different conclusion (when subjective)

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Long-form Report Evaluation

Responses in the format of a “report” offer a more comprehensive answer for the user:

Report

Brighton is sunny [1], 57 min direct, ~£36 return [2], with the comparison:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Can link information to sources

Long-form Report Evaluation

Report evaluation is tricky! There are infinitely many equally-valid ways to express the same information.

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

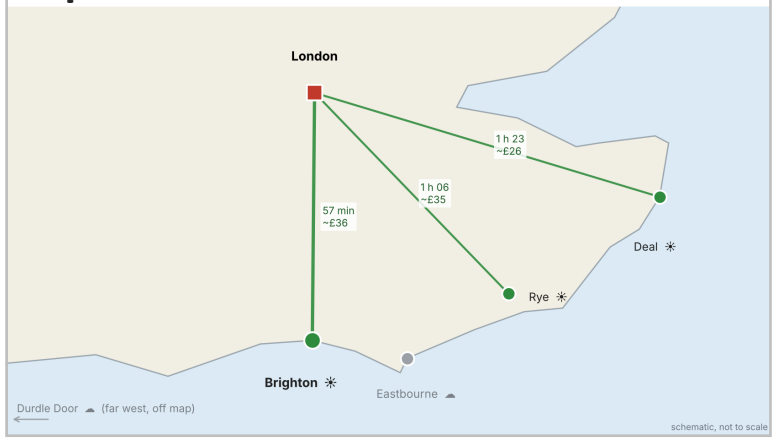
Report

I recommend going to Brighton! It's only a 57 minute train ride away, and will cost you £36 – well within your budget. But you might also want to consider Deal (also nice with a cheaper train journey, but a slightly longer duration) or Rye (about the same duration and cost).

Report

To Brighton, friend, the sun is out,
Fifty-seven minutes, no doubt,
Direct by rail, and for thirty-six,
You'll have your seaside, sand and chips.
Should Deal appeal, the fare is less,
Twenty-six, but I confess
It's one hour twenty-three each way.
Rye's quick too, thirty-five to pay.

Report



Long-form Report Evaluation

Report evaluation often tries to capture:

- **Fact Coverage**: Does the report include the necessary information?
- **Attribution**: Is the information correctly linked to sources that attest it?
- **Text Quality**: Is the information expressed in a coherent way?

Long-form Report Evaluation

Fact Coverage:

- Usually framed as which of a set of gold “nuggets” (atomic facts) are in the report
- These nuggets can be rated based on their importance
- They are sometimes framed as question-answer pairs

#	Nugget	As a question-answer pair	Importance
1	Brighton is a seaside destination reachable from London by train	Is Brighton a seaside destination reachable from London by train? → Yes	vital
2	The train journey to Brighton takes ~57 minutes, direct	What’s the train journey time to Brighton? → ~57 min, direct	vital
3	A return ticket to Brighton costs ~£36	What’s the return fare to Brighton? → ~£36	vital
4	The weather in Brighton on Saturday is sunny	What’s the weather in Brighton on Saturday? → Sunny	vital
5	Deal is sunny, 1 h 23 by train, ~£26 return	Is Deal a viable alternative? → Yes: sunny, 1 h 23, ~£26	okay
6	Rye is sunny, 1 h 06 by train, ~£35 return	Is Rye a viable alternative? → Yes: sunny, 1 h 06, ~£35	okay
7	Eastbourne and Durdle Door are cloudy	Are Eastbourne and Durdle Door viable? → No: cloudy (or too far)	okay

Long-form Report Evaluation

Measuring Fact Coverage:

- Measured in terms of nugget recall (% of gold nuggets supported in a report)
- Human- or LLM-assessed
- Note: longer reports have higher chance of coverage by chance, so length limited (by tokens or chars)

Report
I recommend going to Brighton! It's only a 57 minute train ride away, and will cost you £36 – well within your budget. But you might also want to consider Deal (also nice with a cheaper train journey, but a slightly longer duration) or Rye (about the same duration and cost).

#	Nugget	Importance	Support
1	Brighton is a seaside destination reachable from London by train	vital	Partial
2	The train journey to Brighton takes ~57 minutes, direct	vital	Supported
3	A return ticket to Brighton costs ~£36	vital	Supported
4	The weather in Brighton on Saturday is sunny	vital	Unsupported
5	Deal is sunny, 1 h 23 by train, ~£26 return	okay	Partial
6	Rye is sunny, 1 h 06 by train, ~£35 return	okay	Partial
7	Eastbourne and Durdle Door are cloudy	okay	Unsupported

Vital Coverage = 2/4 = 0.50
All Coverage = 2/7 = 0.29

Long-form Report Evaluation

Measuring Attribution:

- Again, by human assessment or LLM
- Which elements of a report (usually sentences) need a citation?
- If it needs a citation, do the provided citations support the assertions?
- Sentence Support measured by % of cite-worthy sentences that are properly supported.

No citation needed

Report

I recommend going to Brighton! It's only a 57 minute train ride away, and will cost you £36 – well within your budget **[1, 2]**. But you might also want to consider Deal (also nice with a cheaper train journey, but a slightly longer duration) or Rye (about the same duration and cost).

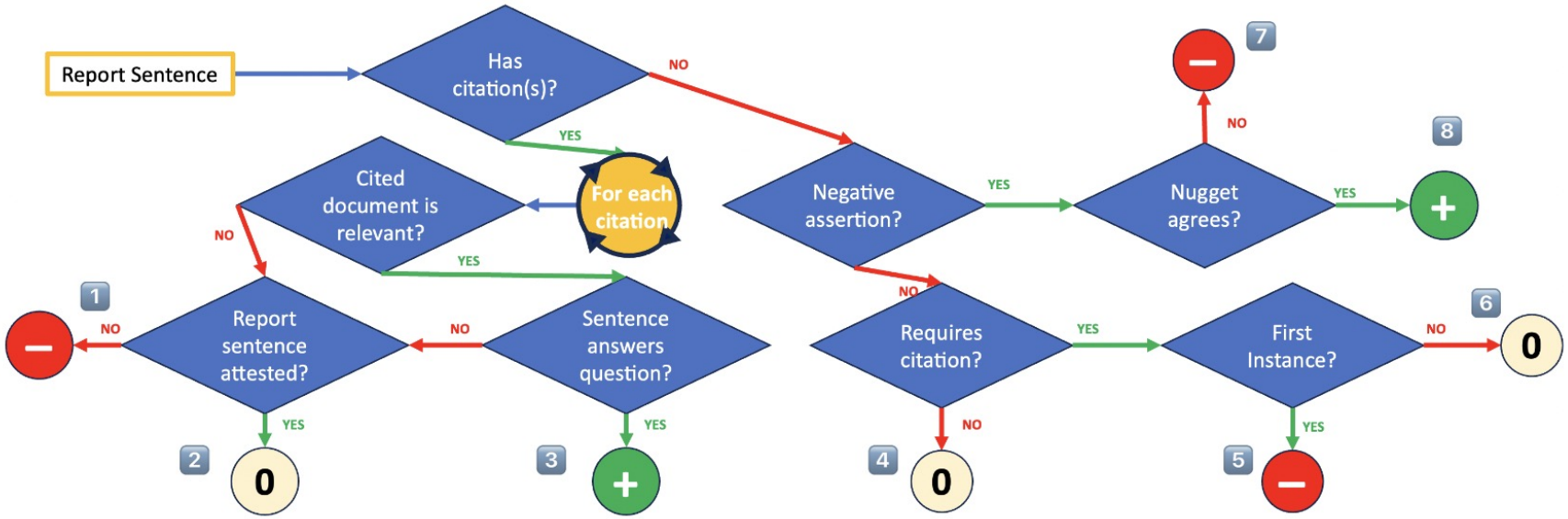
Citation needed and attributed

Citation needed, but not attributed.

Sentence support = 1/2 = 0.50

Long-form Report Evaluation

ARGUE Framework



Mayfield et al. On the Evaluation of Machine-Generated Reports. SIGIR 2024.

Long-form Report Evaluation

Measuring Text Quality:

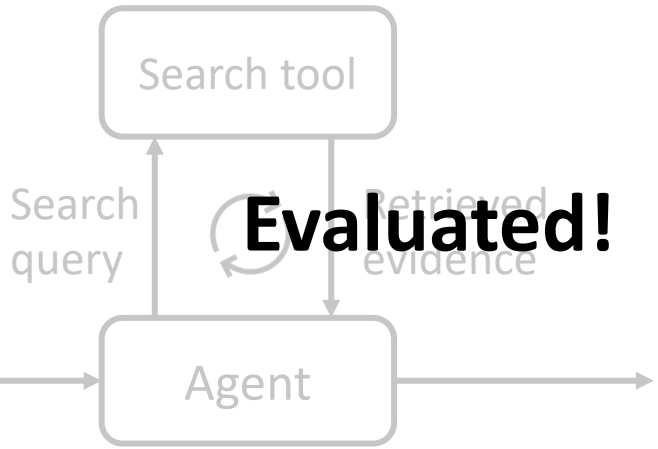
- Measure things like readability, fluency, narrative structure, etc.
- Usually a set of prompts/rubrics, evaluated by LLM

```
…
**Background**: We are evaluating a deep research article written for
the following task across four dimensions: Comprehensiveness, Insight,
Instruction Following, and Readability.
1.  **Comprehensiveness:** The breadth, depth, and relevance of
information coverage.
2.  **Insight:** The depth, originality, logic, and value of the
analysis and conclusions.
3.  **Instruction Following:** Whether the report accurately and
completely responds to all requirements and constraints of the task.
4.  **Readability:** Clarity of structure, fluency of language,
effectiveness of data presentation, and overall ease of understanding.
…
```

Prompt from DeepResearch Bench

Evaluation for agentic search

I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.



Evaluated!

Evidence



Short answer

Brighton

Report

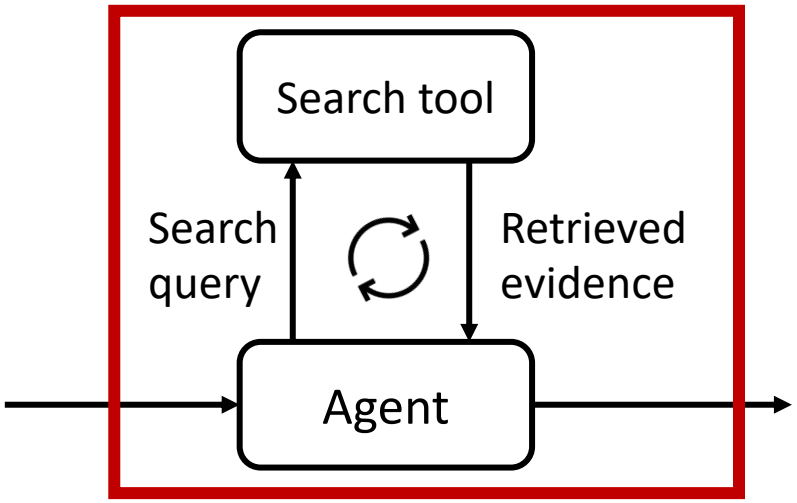
Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Evaluation for agentic search



I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.



Evidence

Short answer

Brighton

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Bonus: Efficiency Evaluation

Efficiency Evaluation

A system could iterate a long time, to keep searching and collecting more evidence:

Token generation is expensive

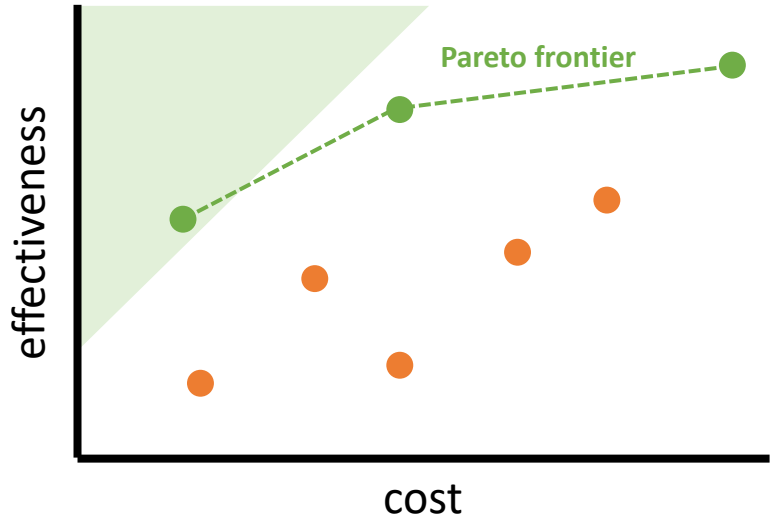
```
<input>prompt</input>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
...
<output>...</output>
```

Efficiency Evaluation

A system could iterate a long time, to keep searching and collecting more evidence:

A system that generates fewer tokens is probably preferable to one that generates more
(assuming the same effectiveness)

```
<input>prompt</input>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
<thinking>...</thinking>
<tool_call>search(...)</tool_call>
<tool_response>...</tool_response>
<thinking>...</thinking>
<output>...</output>
```



Efficiency Evaluation

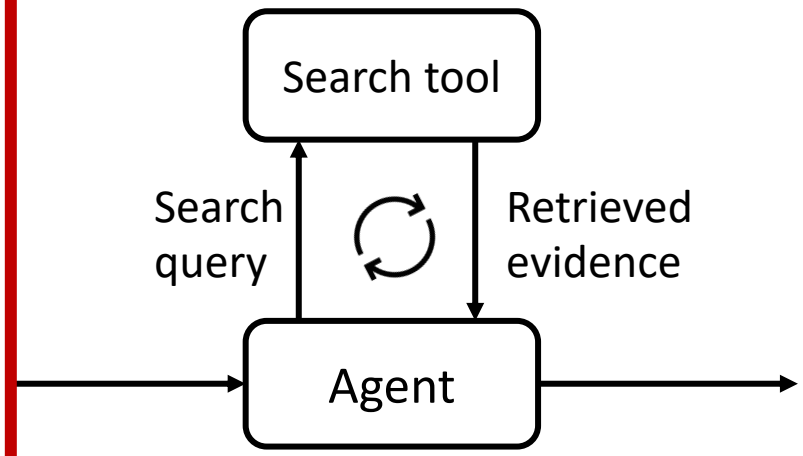
Measuring cost (no measure is perfect)

- **Turns**: the number of search calls invoked (or batches of searches)
  - Common, but really rough estimate of cost
- **Tokens**: the number of input/output tokens of the LLM
  - Some subtlety in measuring this, distinguishing between input/output/cached tokens.
- **Token Cost**: monetary value ($) of the tokens
  - Balances computational cost of input/output/cached tokens
- **Document Reads:** number of documents added to the context
  - Roughly correlated with number of input tokens, since on long chains, this dominates
- **Latency**: wall-clock duration to process the entire request
  - Dependent on hardware+model

Evaluation for agentic search

I want a day trip from London Saturday this weekend. Find a seaside destination with good weather, less than two hours away by train, and return tickets under £50.

Evaluated!



Evidence



Short answer

Brighton

Report

Brighton is sunny, 57 min direct, ~£36 return, with the comparison table:

Candidates	Weather	Train	Return
Brighton	**sunny**	**57 min**	**~£36**
Deal	sunny	1 h 23	~£26
Rye	sunny	1 h 06	~£35
Eastbourne	cloudy	–	–
Durdle Door	cloudy	–	–

Evaluation for agentic search

Tools

- https://github.com/hltcoe/auto-argue
- https://github.com/castorini/nuggetizer
- https://github.com/manestay/dogmatiq
- https://github.com/trec-auto-judge/auto-judge-evaluate

Bibliography

Voorhees. Overview of the TREC 2003 Question Answering Track. TREC 2003

Zhai et al. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. SIGIR 2003

Mayfield et al. On the Evaluation of Machine-Generated Reports. SIGIR 2024

Walden et al. Auto-ARGUE: LLM-Based Report Generation Evaluation. SIGIR 2026

Li et al. DoGMaTiQ: Automated Generation of Question-and-Answer Nuggets for Report Evaluation. ICTIR 2026

Pradeep et al. The Great Nugget Recall: Automating Fact Extraction and RAG Evaluation with Large Language Models. SIGIR 2025.

Du et al. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. ICLR 2026

Chen et al. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. arXiv 2025.

Farzi and Dietz. EXAM++: LLM-based Answerability Metrics for IR Evaluation. LLM4Eval 2024.

Namuduri et al. GENIE: A Fine-Grained Measure for Novelty. EMNLP 2026.

Evaluation for agentic search

Q & A

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30

PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- **Multimodal, multilingual and domain-specific**
- Conclusion and open challenges

Multimodal, multilingual and domain-specific

- **Beyond English text on the open web:** most agentic search work assumes evidence is textual, in the language of the query, and on the open web; real-world tasks often break these assumptions
  - **Multimodal (images and video):** WebWatcher [1], Deep Video Discovery [2], VideoDR [3]
  - **Cross-lingual:** XBCP [4]
  - **Domain-specific (legal, medical, enterprise):** L-MARS [5], DeepMed [6], OfficeQA Pro [7]

*The loop stays the same; what changes is **what the agent searches over** and **which tools it needs**.*

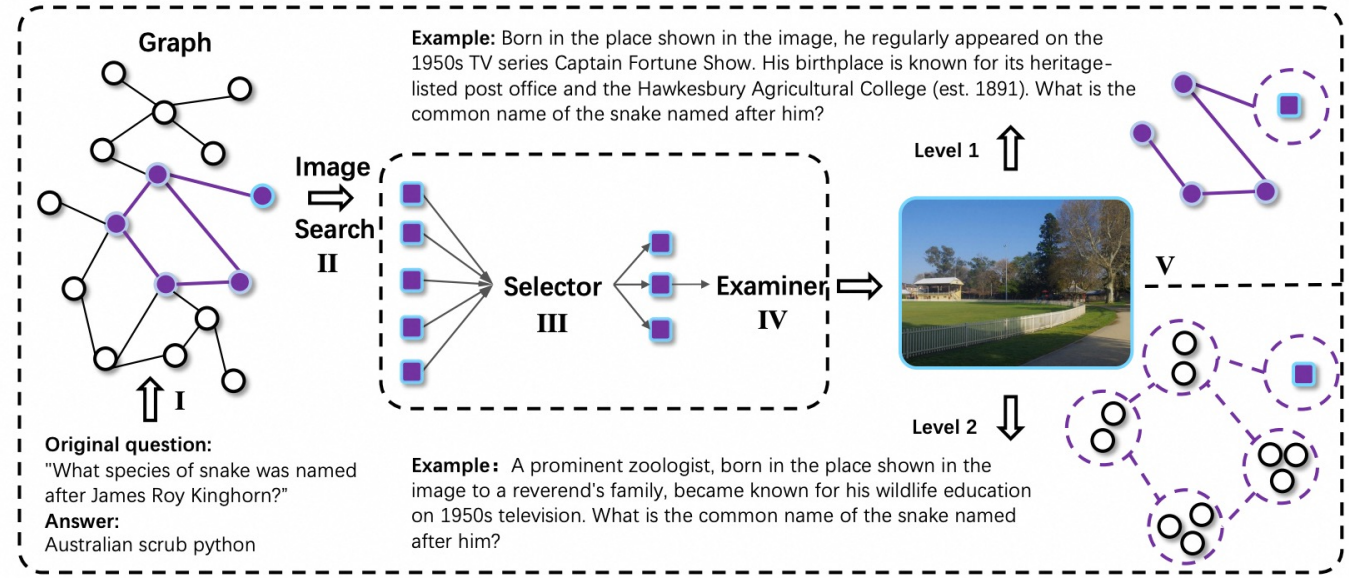
Next: multimodal agentic search over images with WebWatcher

[1] Geng et al. WebWatcher: Breaking New Frontiers of Vision-Language Deep Research Agent. ICLR 2026.
[2] Zhang et al. Deep Video Discovery: Agentic Search with Tool Use for Long-form Video Understanding. NeurIPS 2025.
[3] Liu et al. Watching, Reasoning, and Searching: A Video Deep Research Benchmark on Open Web for Agentic Video Reasoning. arXiv 2026.
[4] Lu et al. Beyond Monolingual Deep Research: Evaluating Agents and Retrievers with Cross-Lingual BrowseComp-Plus. arXiv 2026.
[5] Wang & Yuan. L-MARS: Legal Multi-Agent Workflow with Orchestrated Reasoning and Agentic Search. arXiv 2025.
[6] Wang et al. DeepMed: Building a Medical DeepResearch Agent via Multi-hop Med-Search Data and Turn-Controlled Agentic Training & Inference. Findings of ACL 2026.
[7] Databricks AI Research. OfficeQA Pro: An Enterprise Benchmark for End-to-End Grounded Reasoning. arXiv 2026.

Multimodal agentic search: Images

WebWatcher learns when to search with text and when with images

- **Text-only search agents** cannot use the image; **vision-only agents** over-analyse it and compound errors
- **Data construction:**
  - multi-hop QA from random walks over web knowledge; Level 2 obfuscates entities
  - an image found by image search replaces the key entity
  - selector drops leaky or irrelevant images; examiner keeps queries answerable from the images
- **Training:**
  - cold-start SFT on synthetic multimodal tool-use trajectories
  - RL (GRPO) to improve generalisation

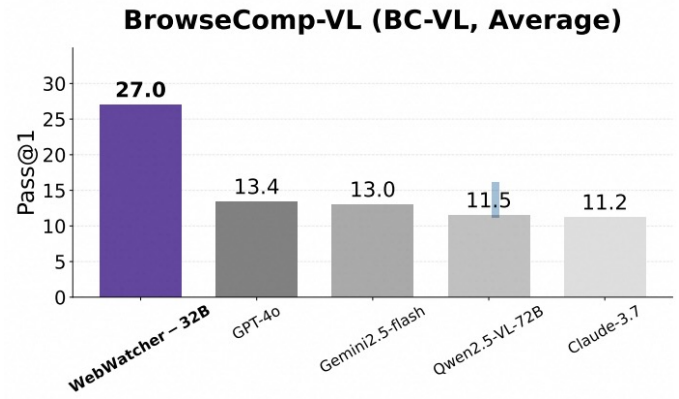


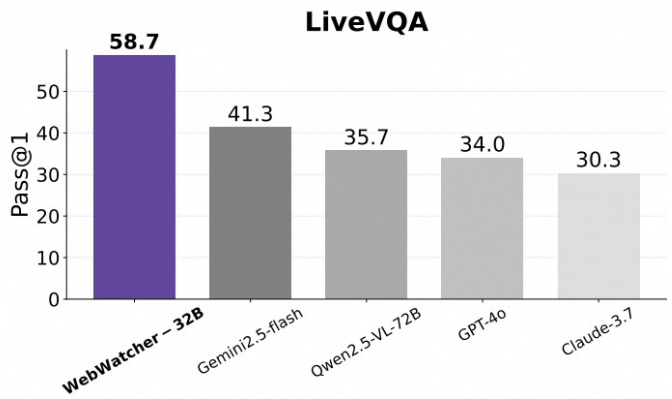
Data construction: graph random walk (I) → image search (II) → selector (III) → examiner (IV) → Level 1 / Level 2 questions (V)

Geng et al. WebWatcher: Breaking New Frontiers of Vision-Language Deep Research Agent. ICLR 2026.

Multimodal agentic search: Images

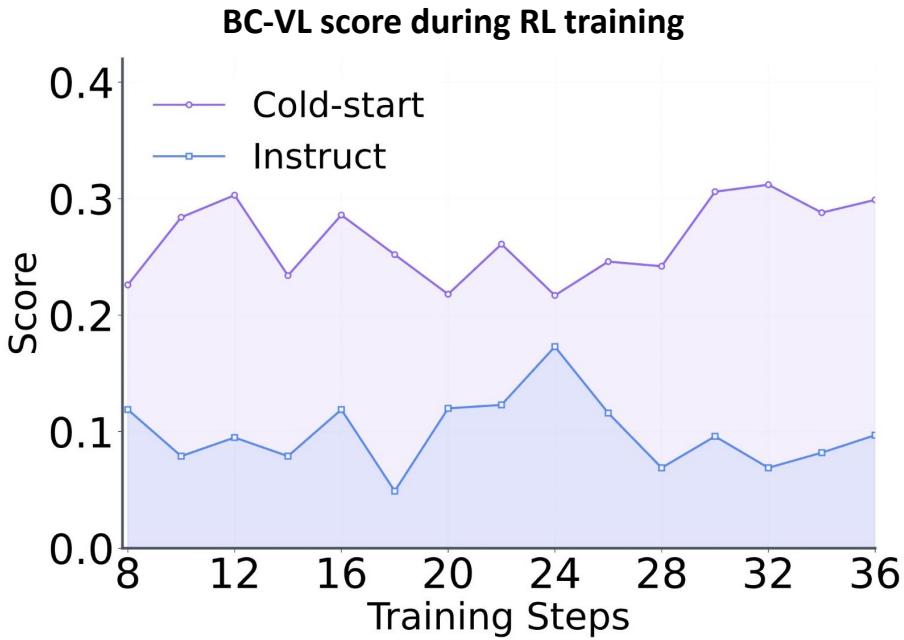
WebWatcher agentic search beats RAG, and tool use adapts to the task





Pass@1 (%)	BC-VL	LiveVQA	MMSearch
GPT-4o direct	5.5	29.7	18.7
GPT-4o + RAG	13.4	34.0	24.1
WebWatcher-32B	**27.0**	**58.7**	**55.3**

- **Search, then agency:** RAG lifts GPT-4o on BrowseComp-VL from 5.5 to 13.4; WebWatcher's agent loop reaches 27.0



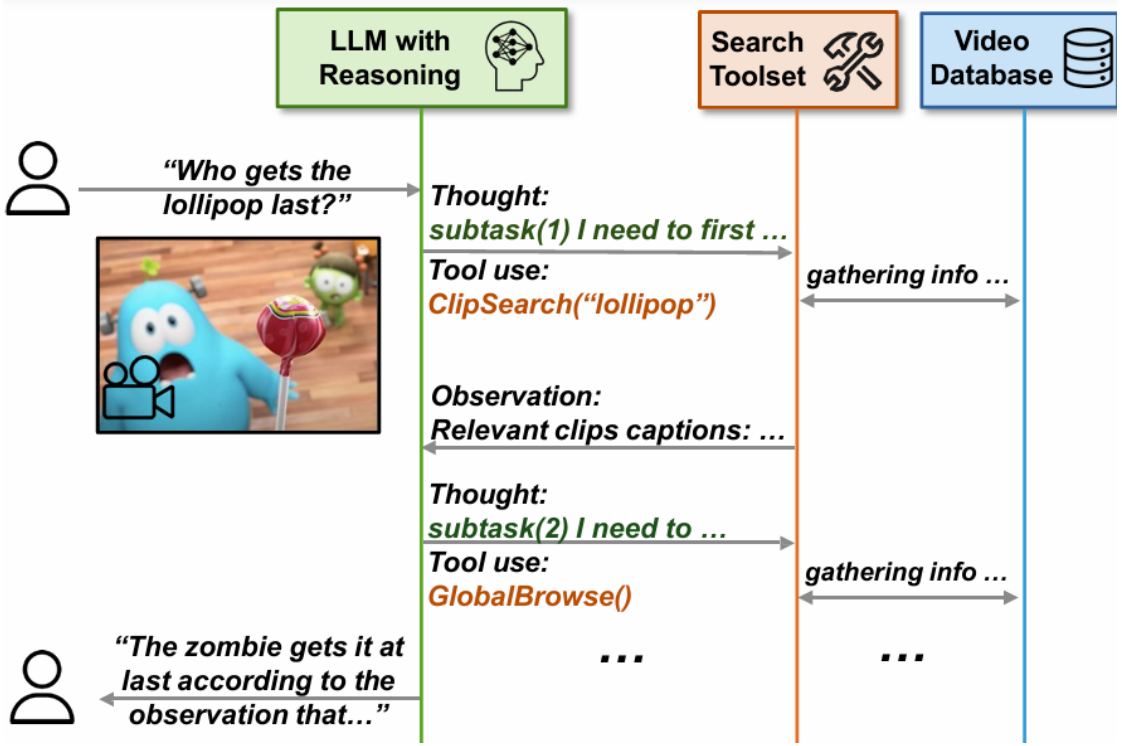
- **Cold start is essential:** after SFT on tool-use trajectories, RL reaches ~0.31; from the instruct model it stays below ~0.17, as tool-format errors wipe out the reward

Geng et al. WebWatcher: Breaking New Frontiers of Vision-Language Deep Research Agent. ICLR 2026.

Multimodal agentic search: Long-form video

Deep Video Discovery [1] search the video instead of watching all of it

- **Index:** split the video into clips; store captions and frames
- **Search:** tools at three levels of detail: skim summaries of the whole video, retrieve the clips whose captions best match a query, and ask a vision model about the raw frames in a chosen time window
- **Stop:** answer once the gathered evidence is enough
- **74.2%** on LVBench [2] (hour-long videos), vs. 60.8% for the previous best agent



The agent alternates thoughts and tool calls until it can answer

[1] Zhang et al. Deep Video Discovery: Agentic Search with Tool Use for Long-form Video Understanding. NeurIPS 2025.
[2] Wang et al. LVBench: An Extreme Long Video Understanding Benchmark. ICCV 2025.

Multimodal agentic search: Video deep research

VideoDR [1] visual clues from the video, answers from the open web

- Videos give only **local visual clues**; the verifiable answer lives on the web
- Extract cross-frame **visual anchors** → turn them into queries → **multi-hop web search** → verify
- e.g. identify a museum from the video, then find the "don't miss" exhibit closest to the entrance and return its accession number
- **500** human-annotated questions across six domains
- **Agentic is not consistently better than Workflow:** early visual anchors drift over long search chains

Accuracy (%) on VideoDR

Model	Workflow	Agentic
Gemini-3.1-pro-preview	51.8	53.6
GPT-5.4	47.0	47.4
Qwen3.5-35B-A3B	35.0	34.6
Gemma-4-31B-it	31.6	33.6
GLM-4.6V-Flash	17.8	21.8
InternVL3.5-14B	16.0	15.4
Human (reference)	40.2	40.2

Workflow: an MLLM first turns video clues into text, then searches without the video. Agentic: one MLLM sees the raw video and searches in a loop.

[1] Liu et al. Watching, Reasoning, and Searching: A Video Deep Research Benchmark on Open Web for Agentic Video Reasoning. arXiv 2026.

Cross-lingual agentic search: XBCP

XBCP [1] English questions, evidence in 12 languages

- BrowseComp-Plus questions and answers stay in English; the **evidence documents are translated** into 11 other languages (high-resource and low-resource)
- **Cross-lingual:** each query's evidence is in one assigned language; Multilingual**:** evidence spread across all languages

Retrieval side: recall and accuracy drop

Corpus	Accuracy	Evidence recall
Original (EN)	64.7	72.8
Multilingual	48.8	59.8
Cross-lingual	42.3	53.8

DeepSeek-V4-Pro agent + Qwen3-Embedding-8B retriever

Agent side: accuracy drops even with all gold evidence

Agent	Original	Multiling.	Cross-ling.
GPT-OSS-20B	90.4	81.2	77.6
GPT-OSS-120B	94.7	89.5	85.3
Qwen3.6-35B-A3B	93.9	90.0	89.2

Accuracy (%) when the agent is given all gold evidence

Two bottlenecks: retrievers miss translated evidence, and agents struggle to integrate it; they also become less calibrated and cite evidence less reliably.

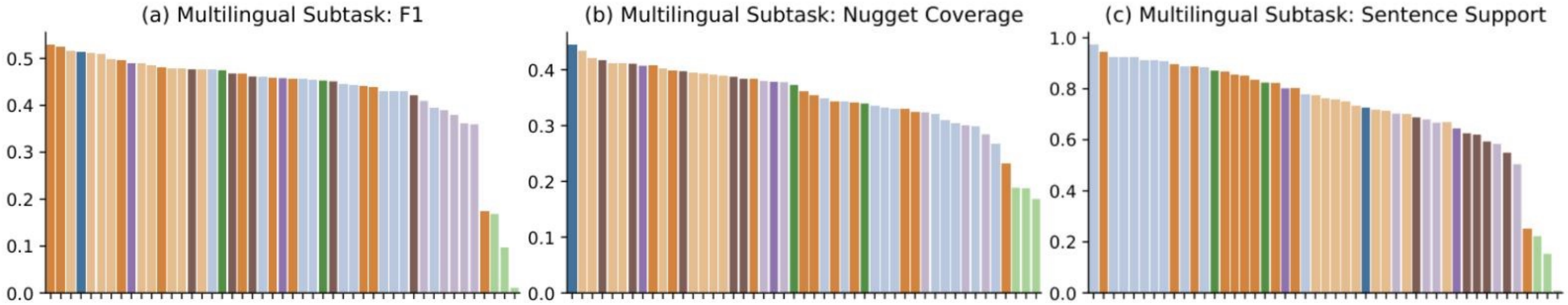
[1] Lu et al. Beyond Monolingual Deep Research: Evaluating Agents and Retrievers with Cross-Lingual BrowseComp-Plus. arXiv 2026.

Multilingual agentic search: TREC RAGTIME

RAGTIME [1] cited English reports from Arabic, Chinese, English and Russian news

- **Task:** given a report request (problem statement + user background), search ~4M news articles (1M each in Arabic, Chinese, English and Russian) and write a **cited English report** of 2,000 or 10,000 characters
- **Scoring (ARGUE): nugget coverage** = share of the assessors' questions the report answers; **sentence support** = share of cited sentences backed by the cited document. 13 teams, 125 runs in 2025

Multilingual report generation, 47 runs: well grounded, but no run covers half the nuggets



Grounding is nearly solved with simple checks; coverage is not. Diverse multilingual retrieval is the main lever.

[1] Lawrie et al. Overview of the TREC 2025 RAGTIME Track. TREC 2025 (arXiv:2602.10024). trec-ragtime.github.io

Domain-specific agentic search: Legal

L-MARS [1] legal answers need current, jurisdiction-specific evidence

Query agent → **Search agent** → **Judge agent** → **Summary agent**

Query agent
Breaks down the question; extracts jurisdiction and date

Search agent
Web, local legal documents, case law

Judge agent
Enough, in-jurisdiction and current? If not: refine and search again

Summary agent
Writes an answer with citations

LegalSearchQA accuracy (%)

Method	Accuracy
Zero-shot LLM	58.0
Chain-of-thought	30.0
L-MARS	**96.0**

- Recent legal changes are not in the model's parameters, so **search is necessary**
- Reasoning longer without evidence **makes things worse** (58.0 → 30.0)
- On Bar Exam QA, where the law is stable, search adds only **+0.7**

[1] Wang & Yuan. L-MARS: Legal Multi-Agent Workflow with Orchestrated Reasoning and Agentic Search. arXiv:2509.00761.

Domain-specific agentic search: Medical

DeepMed [1] general deep research agents find medical evidence but fail to use it

- **Gap:** clinical reasoning, not retrieval, limits general DR agents; more tool calls add noisy context
- **Data:** multi-hop med-search QA built from authoritative medical websites
- **Training:** agentic SFT, then RL with a difficulty-aware turn penalty
- **Inference:** an over-evidence monitor stops searching once the answer stops changing

Accuracy (%), selected benchmarks

Model	HLE-Med	MedXpert	MedMCQA	CMExam
Qwen3-14B (base)	12.8	22.4	68.2	82.8
Qwen3-30B-A3B-Thinking	14.8	26.4	73.1	83.8
Tongyi-DeepResearch-30B	28.2	35.4	75.2	87.6
DeepMed-14B (RL)	**26.8**	**36.1**	**82.6**	**91.2**

Across seven medical benchmarks, DeepMed improves its base model by 9.79% on average.

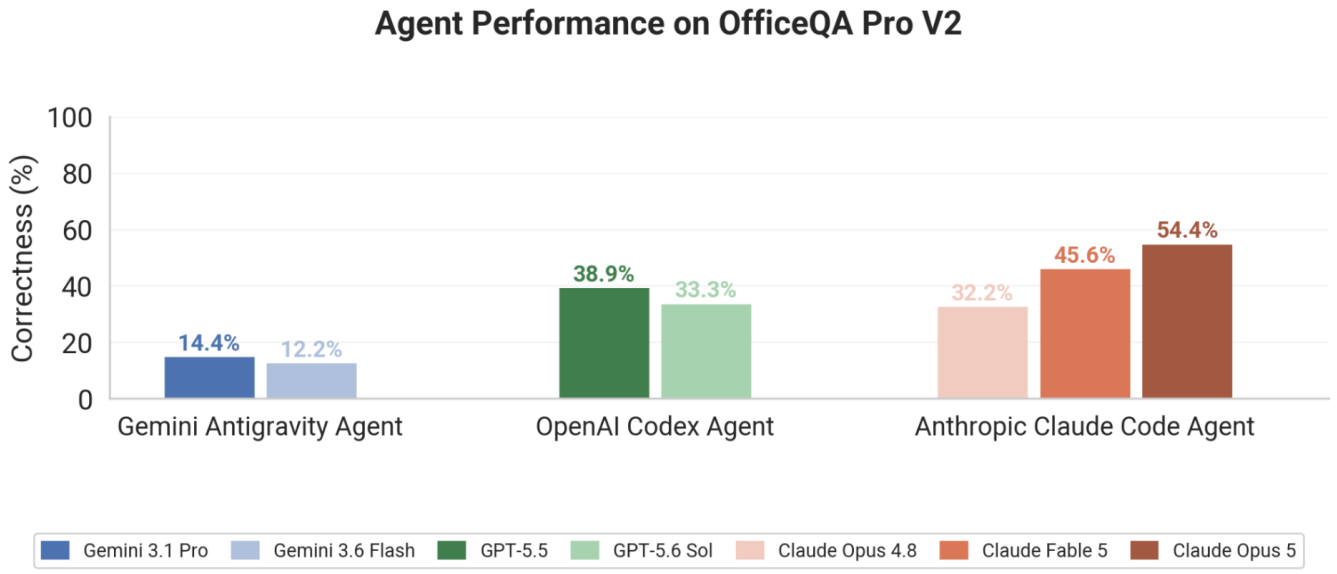
Domain sources are not enough on their own: the agent also needs domain reasoning to use what it finds.

[1] Wang et al. DeepMed: Building a Medical DeepResearch Agent via Multi-hop Med-Search Data and Turn-Controlled Agentic Training & Inference. Findings of ACL 2026.

Domain-specific agentic search: Enterprise

OfficeQA Pro [1] grounded reasoning over enterprise document collections

- Evidence sits in **large private corpora of dense PDFs** (wide tables, scans, legacy layouts)
- Agents must **parse, retrieve and compute** across documents
- **Pro:** 133 questions over U.S. Treasury Bulletins (1939–2025)
- **Pro V2 [2]:** 90 questions over 1,435 federal receipts and expenditures documents (1793–2024); median 5.5 source documents per question
- Best agent harness: **54.4%**



End-to-end correctness on Pro V2: agents use file search (read, grep, glob), web search and code execution.

[1] Databricks AI Research. OfficeQA Pro: An Enterprise Benchmark for End-to-End Grounded Reasoning. arXiv 2026.
[2] Databricks. Introducing OfficeQA Pro V2: A New Benchmark for Enterprise Grounded-Reasoning. 2026.

Multimodal, multilingual and domain-specific

Q & A

Outline

PART I 14:00–16:00

Basics and optimization

- Introduction to agentic search
- Optimization paradigms
  - No parameter updates
  - Agent learns to search
  - Search learns to serve agents
  - Agent–search joint optimization
- Live demonstration and hands-on

BREAK 16:00–16:30

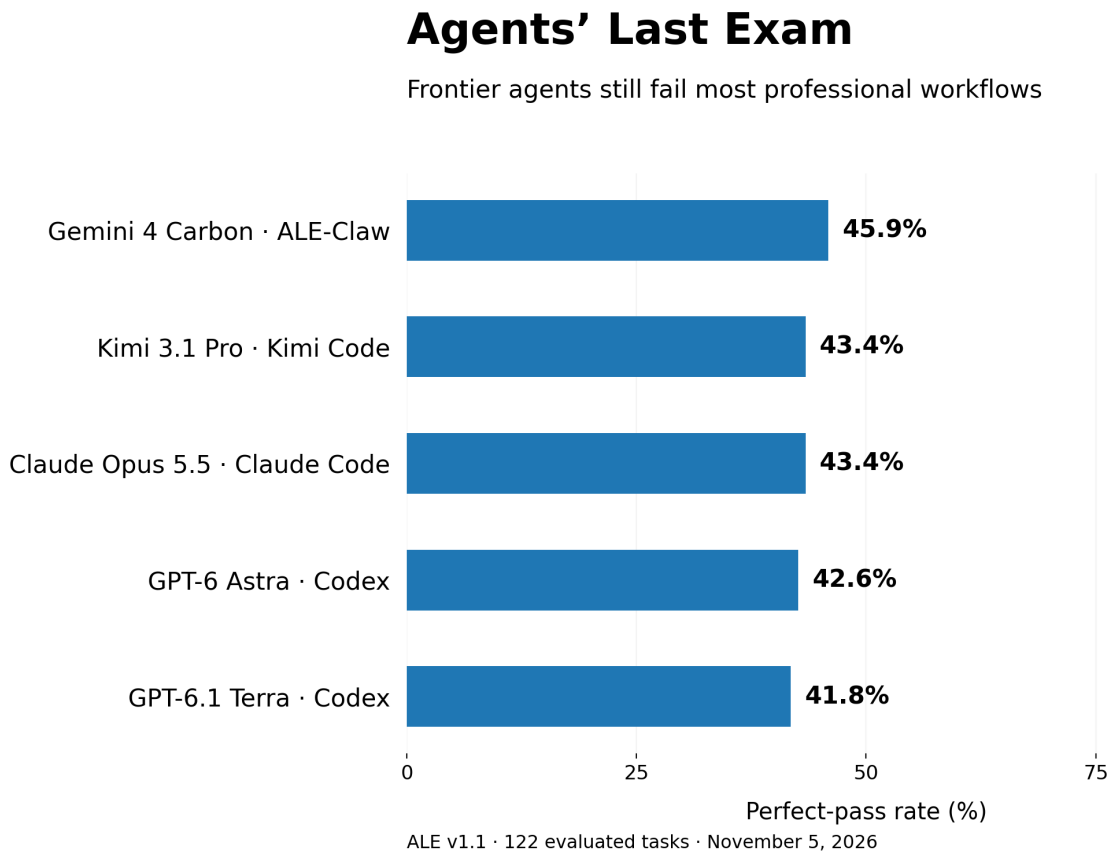
PART II 16:30–18:30

Long-horizons, evaluation and beyond

- Long-horizon agentic search (deep research)
  - Reasoning-intensive QA and report generation
  - Context management
  - Efficiency
- Evaluation for agentic search
  - Evaluation for evidence and short answer
  - Evaluation for report
- Multimodal, multilingual and domain-specific
- **Conclusion and open challenges**

Open Challenges

- **Realistic Tasks and Datasets**
  - Open-ended, real-world scenarios that run for days
  - Realistic environments for long-horizon research
- **Evaluation of Agents**
  - Beyond text: multimodal evidence and generation
  - Long-horizon evaluation: progress, error recovery, and adaptation
- **Retrieval for Agents**
  - Agent-specific query formulation and languages
  - Rankers optimized for agentic search
- **Design and Development of Research Agents**
  - Harness design and agentic search architectures
  - Adaptive workflows and proactive search agents



Conclusion

Key takeaways

- **Agentic search goes beyond traditional retrieval:** agents iteratively reason, search, and refine their answers
- **Optimization spans both agents and search systems:** from agent learning to retrieval and joint optimization
- **Long-horizon research brings new challenges:** evidence management, efficiency, and evaluation across diverse tasks and modalities

Looking ahead

- From benchmark QA to **realistic, open-ended research tasks**
- From traditional retrieval to **IR systems designed for agents**
- From fixed workflows to **adaptive and proactive research agents**