Carnegie Mellon University

Deep Research Agents

Tasks, Benchmarks, and Methods

Akari Asai

Language Technologies Institute
Carnegie Mellon University

11-768: AI Agents Fall 2026

https://akariasai.github.io/ | aasai@andrew.cmu.edu

01

What are Deep Research Agents?

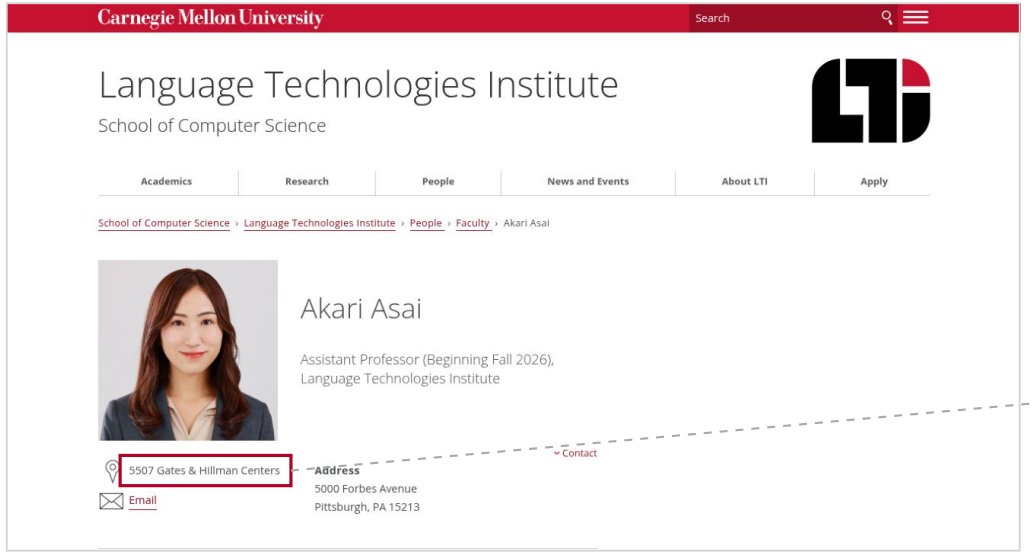
A simple question needs just one search

What is Akari Asai’s office number at CMU?

A simple question needs just one search

What is Akari Asai's office number at CMU?

Search



THE RELEVANT EVIDENCE

5507 Gates & Hillman Centers

CMU LTI faculty profile, captured September 16, 2026.

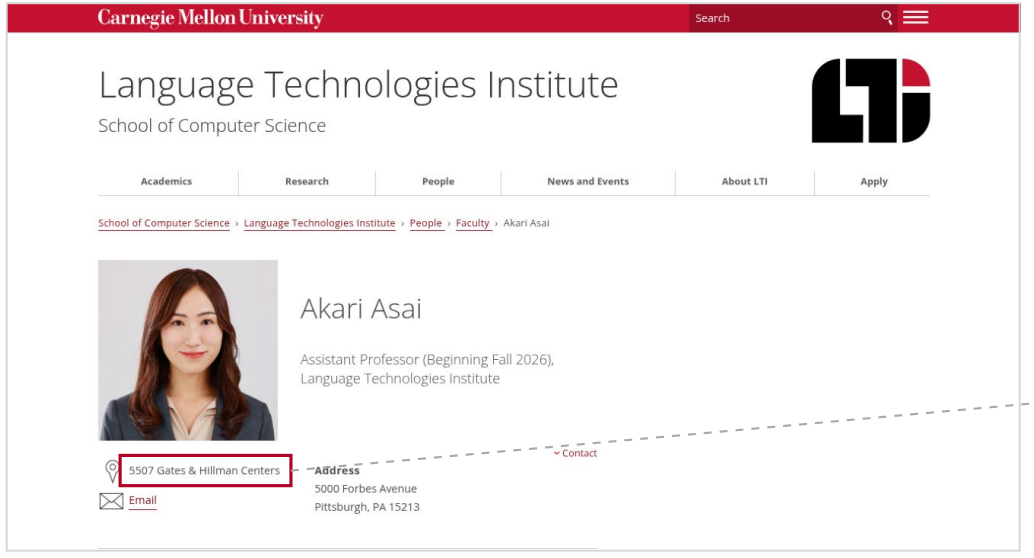
A simple question needs just one search

What is Akari Asai's office number at CMU?

LM

GHC 5507

Search



THE RELEVANT EVIDENCE

5507 Gates & Hillman Centers

CMU LTI faculty profile, captured September 16, 2026.

Complex research needs many searches

Can AI agents synthesize scientific literature as well as human experts?

OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs

Akari Asai[1,5] Jacqueline He[1*] Rulin Shao[1,5*] Weijia Shi[1,2]
Amanpreet Singh[2] Joseph Chee Chang[2] Kyle Lo[2] Luca Soldaini[2]
Sergey Feldman[2] Mike D'arcy[2] David Wadden[2] Matt Latzke[2]
Minyang Tian[3] Pan Ji[6] Shengyan Liu[3] Hao Tong[3] Bohao Wu[3] Yanyu Xiong[7]
Luke Zettlemoyer[1,5] Graham Neubig[4] Dan Weld[1,2] Doug Downey[2]
Wen-tau Yih[5] Pang Wei Koh[1,2] Hannaneh Hajishirzi[1,2]
[1]University of Washington [2]Allen Institute for AI [3]University of Illinois, Urbana-Champaign
[4]Carnegie Mellon University [5]Meta [6]University of North Carolina, Chapel Hill [7]Stanford University
{akari, pangwei, hannaneh}@cs.washington.edu

Abstract

Scientific progress depends on researchers' ability to synthesize the growing body of literature. Can large language models (LMs) assist scientists in this task? We introduce OpenScholar, a specialized retrieval-augmented LM that answers scientific queries by identifying relevant passages from 45 million open-access papers and synthesizing citation-backed responses. To evaluate OpenScholar, we develop ScholarQABench, the first large-scale multi-domain benchmark for literature search, comprising 2,967 expert-written queries and 208 long-form answers across computer science, physics, neuroscience, and biomedicine. On ScholarQABench, OpenScholar-8B outperforms GPT-4o by 5% and PaperQA2 by 7% in correctness, despite being a smaller, open model. While GPT4o hallucinates citations 78–90% of the time, OpenScholar achieves citation accuracy on par with human experts. OpenScholar's datastore, retriever, and self-feedback inference loop also improves off-the-shelf LMs: for instance, OpenScholar-GPT4o improves GPT-4o's correctness by 12%. In human evaluations, experts preferred OpenScholar-8B and OpenScholar-GPT4o responses over expert-written ones 51% and 70% of the time, respectively, compared to GPT4o's 32%. We open-source all of our code, models, datastore, data and a public demo.

Demo	openscholar.allen.ai/
Blog	allenai.org/blog/openscholar
OpenScholar code	github.com/AkariAsai/OpenScholar
ScholarBench code	github.com/AkariAsai/ScholarBench
Checkpoints, Data, Index	OpenScholar/openscholar-v1
Expert Evaluation	AkariAsai/OpenScholar_ExpertEval

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research

Rulin Shao Akari Asai Shannon Zejiang Shen Hamish Ivison
Varsha Kishore Jingming Zhuo Xinran Zhao Molly Park Samuel G. Finlayson
David Sontag Tyler Murray Sewon Min Pradeep Dasigi Luca Soldaini Faeze Brahman
Wen-tau Yih Tongshuang Wu Luke Zettlemoyer Yoon Kim
Hannaneh Hajishirzi Pang Wei Koh

Code Data & Models Interactive Demo

Abstract

Deep research agents perform multi-step research to produce long-form, well-attributed answers. However, most open deep research agents are trained on easily verifiable short-form QA tasks via reinforcement learning with verifiable rewards, which does not extend to realistic long-form tasks. We address this with **Reinforcement Learning with Evolving Rubrics (RLER)**, where rubrics are constructed and maintained to *co-evolve* with the policy model during training. This allows the rubrics to incorporate newly explored information from search and contrasting model responses, enabling better fact checking and more discriminative on-policy feedback. Using RLER, we develop **Deep Research Tulu (DR Tulu-8B)**, the first fully open model that is directly trained for open-ended, long-form deep research. Across four long-form deep research benchmarks in science, healthcare, and general domains, DR Tulu substantially outperforms existing open deep research agents (by 15.6% over Tongyi DR on average) and matches or exceeds proprietary deep research agents (by 0.7% over OpenAI DR on average), while being significantly smaller and cheaper per query (1000× cheaper than OpenAI DR per query).

Figure 1. **Performance vs. cost of deep research models.** We report average performance over 4 long-form DR benchmarks (ScholarQA-CSv2, HealthBench, ResearchQA, and DeepResearch-Bench) against inference cost (USD per query on ScholarQA-CSv2). DR Tulu-8B lies on the Pareto frontier, outperforming larger open models and matching proprietary models (Table 1).

1. Introduction

Deep research (DR) agents aim to produce in-depth, well-attributed answers to complex research tasks by planning, searching, and synthesizing information from diverse sources (OpenAI, 2025). Existing open DR agents are either training-free, using manually designed prompts with off-the-shelf models (Li et al., 2025b;a), or trained via reinforcement learning with verifiable rewards (RLVR) on search-intensive yet constrained short-form question answering (Jin et al., 2025; Nguyen et al., 2025; Liu et al., 2025). RL training for open-ended DR tasks critically de-

Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

Complex research needs many searches

Can AI agents synthesize scientific literature as well as human experts?

OPENSCHOLAR: SYNTHESIZING SCIENTIFIC LITERATURE WITH RETRIEVAL-AUGMENTED LMS

Akari Asai[1,5] Jacqueline He[1*] Rulin Shao[1,5*] Weijia Shi[1,2]
Amanpreet Singh[2] Joseph Chee Chang[2] Kyle Lo[2] Luca Soldaini[2]
Sergey Feldman[2] Mike D'arcy[2] David Wadden[2] Matt Latzke[2]
Minyang Tian[3] Pan Ji[6] Shengyan Liu[3] Hao Tong[3] Bohao Wu[3] Yanyu Xiong[7]
Luke Zettlemoyer[1,5] Graham Neubig[4] Dan Weld[1,2] Doug Downey[2]
Wen-tau Yih[5] Pang Wei Koh[1,2] Hannaneh Hajishirzi[1,2]
[1]University of Washington [2]Allen Institute for AI [3]University of Illinois, Urbana-Champaign
[4]Carnegie Mellon University [5]Meta [6]University of North Carolina, Chapel Hill [7]Stanford University
{akari, pangwei, hannaneh}@cs.washington.edu

ABSTRACT

Scientific pr… of literature. … introduce O… scientific qu… papers and s… we develop … for literature… answers acro… SCHOLARQABENCH, OPENSCHOLAR-8B outperforms GPT-4o by 5% and PaperQA2 by 7% in correctness, despite being a smaller, open model. While GPT4o hallucinates citations 78–90% of the time, OPENSCHOLAR achieves citation accuracy on par with human experts. OPENSCHOLAR's datastore, retriever, and self-feedback inference loop also improves off-the-shelf LMs: for instance, OPENSCHOLAR-GPT4o improves GPT-4o's correctness by 12%. In human evaluations, experts preferred OPENSCHOLAR-8B and OPENSCHOLAR-GPT4o responses over expert-written ones 51% and 70% of the time, respectively, compared to GPT4o's 32%. We open-source all of our code, models, datastore, data and a public demo.

Demo	openscholar.allen.ai/
Blog	allenai.org/blog/openscholar
OpenScholar code	github.com/AkariAsai/OpenScholar
ScholarBench code	github.com/AkariAsai/ScholarBench
Checkpoints, Data, Index	OpenScholar/openscholar-v1
Expert Evaluation	AkariAsai/OpenScholar_ExpertEval

⚠ No single result settles this question

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research

Rulin Shao Akari Asai Shannon Zejiang Shen Hamish Ivison
Varsha Kishore Jingming Zhuo Xinran Zhao Molly Park Samuel G. Finlayson
David Sontag Tyler Murray Sewon Min Pradeep Dasigi Luca Soldaini Faeze Brahman
Wen-tau Yih Tongshuang Wu Luke Zettlemoyer Yoon Kim
Hannaneh Hajishirzi Pang Wei Koh

Code Data & Models Interactive Demo

Abstract

Deep research agents perform multi-step research to produce long-form, well-attributed answers. … most open deep research agents are … easily verifiable short-form QA tasks … cement learning with verifiable rewards, … not extend to realistic long-form tasks. … this with **Reinforcement Learning … ing Rubrics (RLER)**, where rubrics … cted and maintained to *co-evolve* with … model during training. This allows the … rubrics to incorporate newly explored information from search and contrasting model responses, enabling better fact checking and more discriminative on-policy feedback. Using RLER, we develop **Deep Research Tulu (DR Tulu-8B)**, the first fully open model that is directly trained for open-ended, long-form deep research. Across four long-form deep research benchmarks in science, healthcare, and general domains, DR Tulu substantially outperforms existing open deep research agents (by 15.6% over Tongyi DR on average) and matches or exceeds proprietary deep research agents (by 0.7% over OpenAI DR on average), while being significantly smaller and cheaper per query (1000× cheaper than OpenAI DR per query).

Figure 1. **Performance vs. cost of deep research models.** We report average performance over 4 long-form DR benchmarks (ScholarQA-CSv2, HealthBench, ResearchQA, and DeepResearch-Bench) against inference cost (USD per query on ScholarQA-CSv2). DR Tulu-8B lies on the Pareto frontier, outperforming larger open models and matching proprietary models (Table 1).

1. Introduction

Deep research (DR) agents aim to produce in-depth, well-attributed answers to complex research tasks by planning, searching, and synthesizing information from diverse sources (OpenAI, 2025). Existing open DR agents are either training-free, using manually designed prompts with off-the-shelf models (Li et al., 2025b;a), or trained via reinforcement learning with verifiable rewards (RLVR) on search-intensive yet constrained short-form question answering (Jin et al., 2025; Nguyen et al., 2025; Liu et al., 2025). RL training for open-ended DR tasks critically de-

Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

Complex research needs many searches

Can AI agents synthesize scientific literature as well as human experts?

Plan

I need to separate strong benchmark results from direct comparisons with human experts.

1. Find studies of literature synthesis.
2. Check how experts and agents were compared.
3. Compare tasks and limits before drawing a conclusion.

Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

Complex research needs many searches

Can AI agents synthesize scientific literature as well as human experts?

Search

AI agent literature synthesis human expert evaluation

Result A: a research benchmark
Reports answer scores against other AI systems.

Result B: an expert evaluation
Compares agent answers with human-written answers.

Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

Complex research needs many searches

Can AI agents synthesize scientific literature as well as human experts?

Reflect

These studies use different comparisons. A higher benchmark score does not yet answer my question about human experts. Which tasks and domains were tested?

Search again

literature synthesis expert comparison tasks domains evaluation protocol

Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

Complex research needs many searches

Can AI agents synthesize scientific literature as well as human experts?

FINAL ANSWER

Promising on tested tasks; broader parity with experts remains unproven.

On evaluated literature-synthesis questions, experts preferred OpenScholar answers over expert-written ones; citation accuracy was comparable to human experts. **[1]**

DR Tulu reports strong results across four long-form research benchmarks. Comparisons with other AI agents do not directly establish parity with human experts. **[2]**

[1] Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.

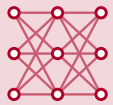
[2] DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

Today's lecture

Evaluation

Benchmarks, rubrics, and citation support



Modeling

Learning to search, reason, and synthesize

Retrieval for deep research

Finding evidence for the agent's next step

02

Benchmarks & Evaluation

Limitations of Classical QA Benchmarks

QUESTION

who wrote the score for the force awakens

WIKIPEDIA / SOURCE DOCUMENT

Star Wars: The Force Awakens – Original Motion Picture Soundtrack is the film score to the 2015 film of the same name composed by John Williams with Williams and William Ross conducting, and Gustavo Dudamel appearing as a "special guest conductor". The album was released by Walt Disney Records on December 18, 2015 in both digipak CD and digital formats.[1][2]

ANSWER

John Williams

Official Natural Questions example; linked Wikipedia revision (2018).

Limitations of Classical QA Benchmarks

QUESTION

who wrote the score for the force awakens

⚠ **Simple lookup**

WIKIPEDIA / SOURCE DOCUMENT

Star Wars: The Force Awakens – Original Motion Picture Soundtrack is the film score to the 2015 film of the same name composed by John Williams with Williams and William Ross conducting, and Gustavo Dudamel appearing as a "special guest conductor". The album was released by Walt Disney Records on December 18, 2015 in both digipak CD and digital formats.[1][2]

ANSWER

John Williams

Official Natural Questions example; linked Wikipedia revision (2018).

Limitations of Classical QA Benchmarks

QUESTION

who wrote the score for the force awakens

⚠ **Simple lookup**

WIKIPEDIA / SOURCE DOCUMENT

Star Wars: The Force Awakens – Original Motion Picture Soundtrack is the film score to the 2015 film of the same name composed by John Williams with Williams and William Ross conducting, and Gustavo Dudamel appearing as a "special guest conductor". The album was released by Walt Disney Records on December 18, 2015 in both digipak CD and digital formats.[1][2]

⚠ **General domain**

ANSWER

John Williams

Official Natural Questions example; linked Wikipedia revision (2018).

Limitations of Classical QA Benchmarks

QUESTION

who wrote the score for the force awakens

⚠ Simple lookup

WIKIPEDIA / SOURCE DOCUMENT

Star Wars: The Force Awakens – Original Motion Picture Soundtrack is the film score to the 2015 film of the same name composed by John Williams with Williams and William Ross conducting, and Gustavo Dudamel appearing as a "special guest conductor". The album was released by Walt Disney Records on December 18, 2015 in both digipak CD and digital formats.[1][2]

⚠ General domain

ANSWER

John Williams

⚠ Short-form answer

Official Natural Questions example; linked Wikipedia revision (2018).

Limitations of Classical QA Benchmarks

QUESTION

who wrote the score for the force awakens

⚠ Simple lookup

WIKIPEDIA / SOURCE DOCUMENT

Star Wars: The Force Awakens – Original Motion Picture Soundtrack is the film score to the 2015 film of the same name composed by John Williams with Williams and William Ross conducting, and Gustavo Dudamel appearing as a "special guest conductor". The album was released by Walt Disney Records on December 18, 2015 in both digipak CD and digital formats.[1][2]

⚠ General domain

⚠ Evidence use unclear

ANSWER

John Williams

⚠ Short-form answer

Official Natural Questions example; linked Wikipedia revision (2018).

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

Simple lookup

A familiar fact may be recalled or found in one search.

2 / DOMAIN EXPERTISE

General domain

Broad web questions may not test expert knowledge.

3 / ANSWER QUALITY

Short-form answer

Answer matching does not assess multi-document synthesis.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

Simple lookup

A familiar fact may be recalled or found in one search.

2 / DOMAIN EXPERTISE

General domain

Broad web questions may not test expert knowledge.

3 / ANSWER QUALITY

Short-form answer

Answer matching does not assess multi-document synthesis.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

BrowseComp: building many-search questions

1,266 human-written questions

CREATE THE QUESTION

Start with the answer

A known person, event, or artifact.

→

Write a complex question

Hide its name; combine factual clues.

VERIFY THE DIFFICULTY

LM + search checks

Models should fail;

5 Google searches should not suffice.

Human check

Still hard after **10 minutes**?

Target checked on a subset.

BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv, 2025.

BrowseComp: a question built from clues

ORIGINAL QUESTION / CLUES IN READING ORDER

Please identify the fictional character who

BEHAVIOR	occasionally breaks the fourth wall with the audience,
BACKSTORY	has a backstory involving help from selfless ascetics,
PERSONALITY	is known for his humor, and
TV SERIES	had a TV show that aired between the 1960s and 1980s
EPISODE COUNT	with fewer than 50 episodes.

Find one character who satisfies **all five clues**.

BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv, 2025.

BrowseComp: a question built from clues

ORIGINAL QUESTION / CLUES IN READING ORDER

Please identify the fictional character who

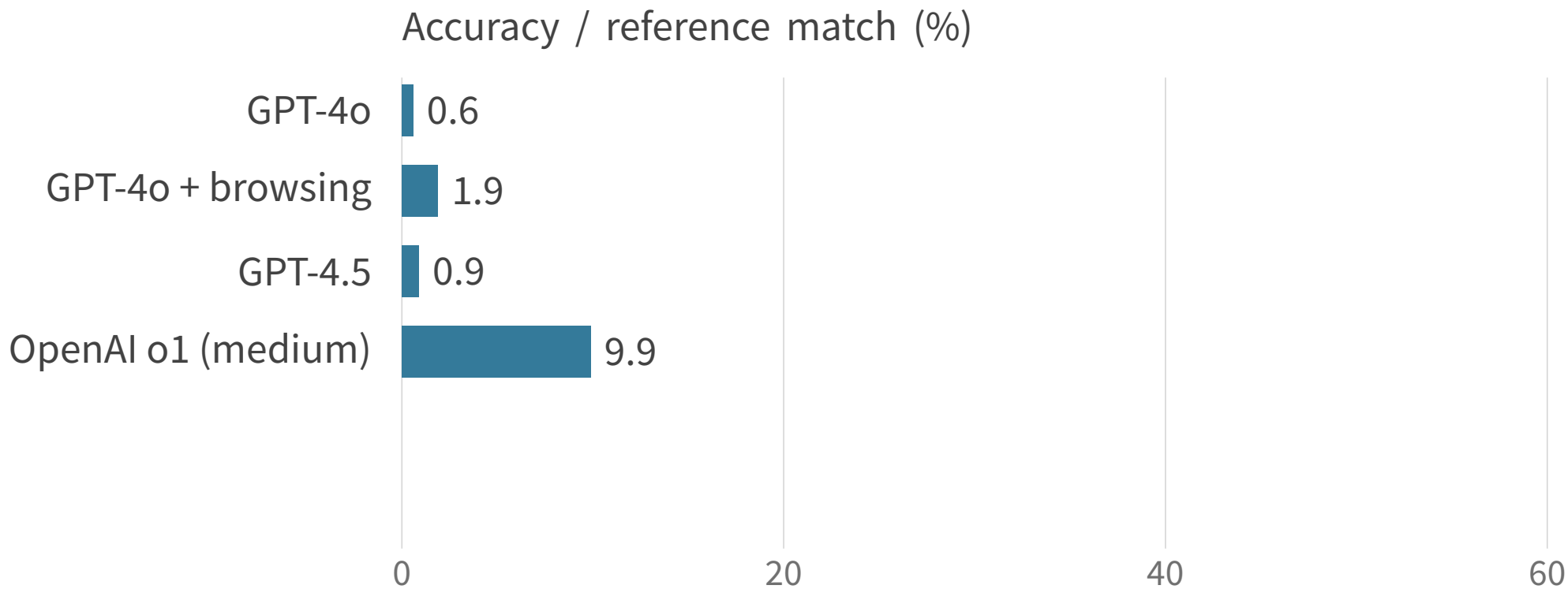
BEHAVIOR	occasionally breaks the fourth wall with the audience,
BACKSTORY	has a backstory involving help from selfless ascetics,
PERSONALITY	is known for his humor, and
TV SERIES	had a TV show that aired between the 1960s and 1980s
EPISODE COUNT	with fewer than 50 episodes.

Find one character who satisfies **all five clues**.

Reference answer: **Plastic Man**

BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv, 2025.

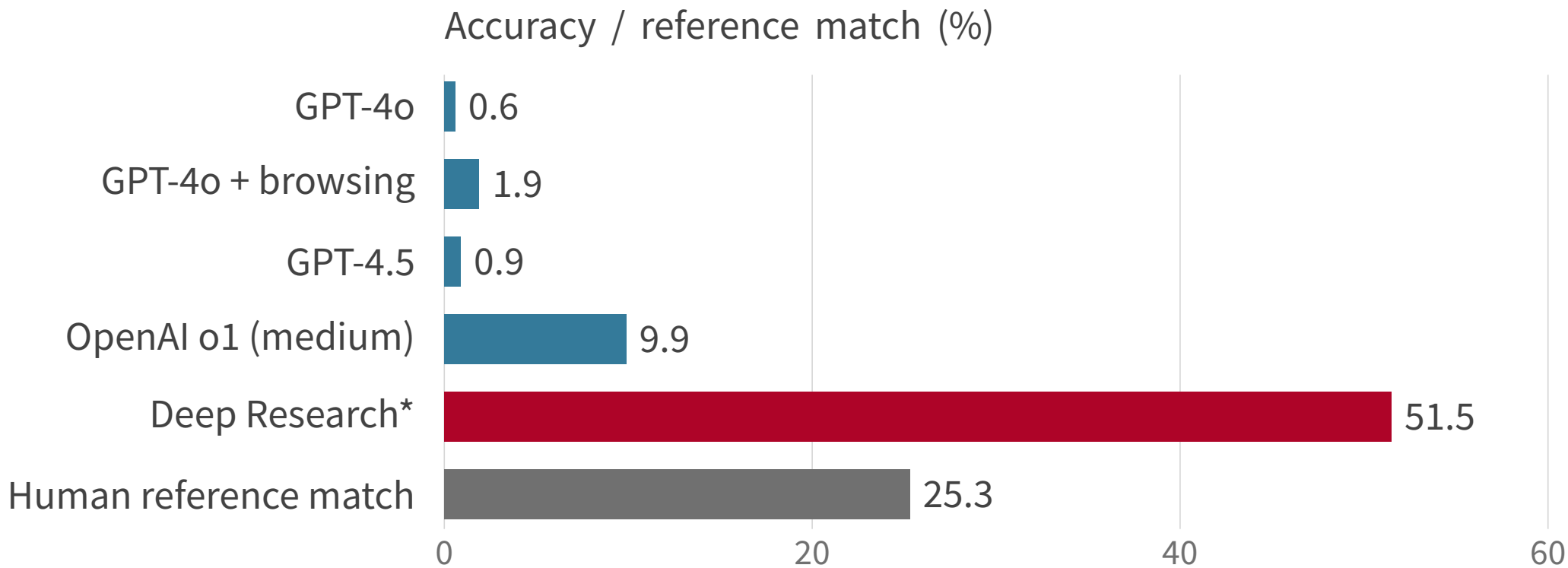
BrowseComp: results



Humans attempted 1,255 of 1,266 questions; 11 were unattempted.

BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv, 2025. *Trained for BrowseComp-style tasks.

BrowseComp: results

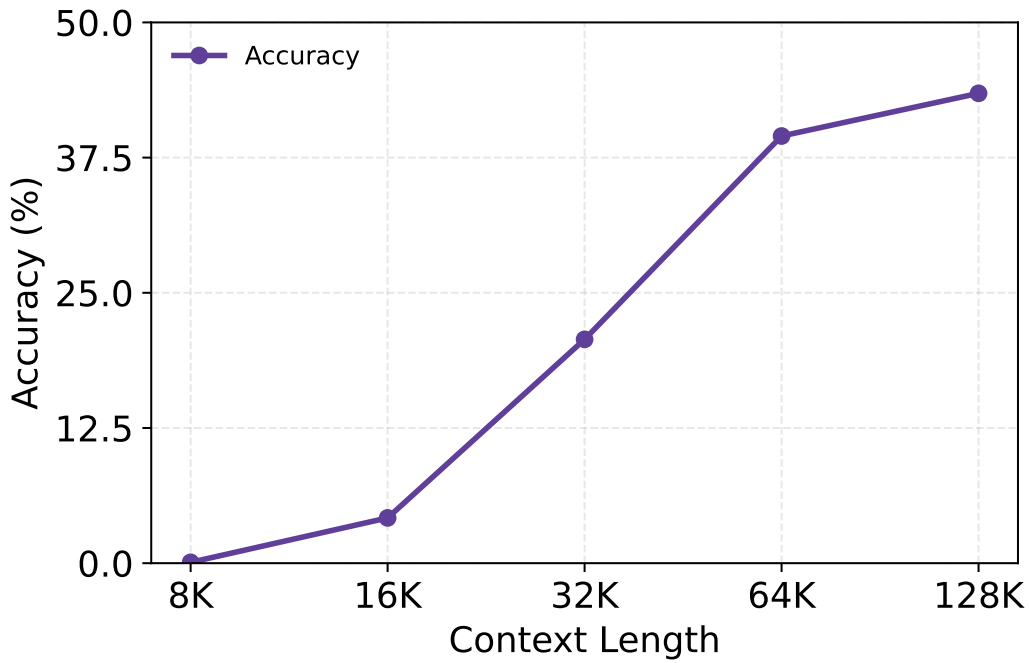


Humans attempted 1,255 of 1,266 questions; 11 were unattempted.

BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv, 2025. *Trained for BrowseComp-style tasks.

More room to search improves accuracy

Tongyi DeepResearch on BrowseComp



TRAINING TRAJECTORIES

>10 tool calls

in over 20% of SFT samples.

EVALUATION LIMIT

128 tool calls

maximum per task.

More context allows longer interaction histories; accuracy rises with the budget.

Tongyi DeepResearch Technical Report. arXiv, 2025.

BrowseComp-Plus makes the corpus fixed

Collect and verify the evidence, then freeze the document collection.

BROWSECOMP PAIR

Question + known answer

Offline input

→

1 / GATHER EVIDENCE

o3 finds evidence pages.

→

2 / HUMAN VERIFICATION

- ✓ Mark supporting spans
- ✓ Check every clue
- ✓ Mark answer documents

BrowseComp-Plus: A Fair and Disentangled Evaluation Benchmark for Deep Search Agents. ACL 2026.

BrowseComp-Plus makes the corpus fixed

Collect and verify the evidence, then freeze the document collection.

BROWSECOMP PAIR

Question + known answer

Offline input

1 / GATHER EVIDENCE

o3 finds evidence pages.

2 / HUMAN VERIFICATION

- ✓ Mark supporting spans
- ✓ Check every clue
- ✓ Mark answer documents

ADD HARD NEGATIVES

Related pages that miss a clue

FIXED BENCHMARK

830 questions / 100,195 documents

BrowseComp-Plus: A Fair and Disentangled Evaluation Benchmark for Deep Search Agents. ACL 2026.

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

BrowseComp / BrowseComp-Plus

Hard-to-find answers; a fixed corpus for controlled comparisons.

2 / DOMAIN EXPERTISE

General domain

Broad web questions may not test expert knowledge.

3 / ANSWER QUALITY

Short-form answer

Answer matching does not assess multi-document synthesis.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

BrowseComp / BrowseComp-Plus

Hard-to-find answers; a fixed corpus for controlled comparisons.

2 / DOMAIN EXPERTISE

General domain

Broad web questions may not test expert knowledge.

3 / ANSWER QUALITY

Short-form answer

Answer matching does not assess multi-document synthesis.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

FinSearchComp: expert-written questions

Focus: complex historical investigation

1 / EXPERTS WRITE

Start from work scenarios or verified financial tables.

→

2 / ESTABLISH ANSWERS

Retrieve reliable data; calculate and cross-check.

→

3 / INDEPENDENT REVIEW

Other experts solve it; resolve disagreements.

EXAMPLE / PARAPHRASED

How did Johnson & Johnson's international share of revenue change year over year during 2022–2024?

FinSearchComp: Towards a Realistic, Expert-Level Evaluation of Financial Search and Reasoning. ICLR 2026.

Search benchmarks with domain expertise

Domain knowledge + multi-step evidence gathering

MedBrowseComp Medicine: link trials, drugs, and regulatory facts.

MedBrowseComp: Benchmarking Medical Deep Research and Computer Use. GenAI4Health Workshop, NeurIPS 2025.

FinSearchComp Finance: investigate facts across financial sources.

FinSearchComp: Towards a Realistic, Expert-Level Evaluation of Financial Search and Reasoning. ICLR 2026.

Scientific deep research / paper retrieval

ScholarSearch: academic search. AutoResearchBench: one paper or all matches.

ScholarSearch: Benchmarking Scholar Searching Ability of LLMs. arXiv, 2025.

AutoResearchBench: Benchmarking AI Agents on Complex Scientific Literature Discovery. arXiv, 2026.

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

BrowseComp / BrowseComp-Plus

Hard-to-find answers; a fixed corpus for controlled comparisons.

2 / DOMAIN EXPERTISE

MedBrowseComp
FinSearchComp

Expert knowledge guides medical and financial search.

3 / ANSWER QUALITY

Short-form answer

Answer matching does not assess multi-document synthesis.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

BrowseComp / BrowseComp-Plus

Hard-to-find answers; a fixed corpus for controlled comparisons.

2 / DOMAIN EXPERTISE

MedBrowseComp
FinSearchComp

Expert knowledge guides medical and financial search.

3 / ANSWER QUALITY

Short-form answer

Answer matching does not assess multi-document synthesis.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

Research needs multi-dimensional evaluation

Evaluate **accuracy, coverage, qualifications, and evidence support**.

Can AI agents synthesize scientific literature as well as human experts?

ONE REFERENCE SUMMARY

Experts preferred OpenScholar in one study. Parity across domains is unproven.

Research needs multi-dimensional evaluation

Evaluate **accuracy, coverage, qualifications, and evidence support**.

Can AI agents synthesize scientific literature as well as human experts?

ONE REFERENCE SUMMARY

Experts preferred OpenScholar in one study. Parity across domains is unproven.

ANOTHER VALID SUMMARY

One study favors OpenScholar; it does not establish equivalence across domains.

⚠ **Several valid answers**

Research needs multi-dimensional evaluation

Evaluate **accuracy, coverage, qualifications, and evidence support**.

Can AI agents synthesize scientific literature as well as human experts?

ONE REFERENCE SUMMARY

Experts preferred OpenScholar in one study. Parity across domains is unproven.

ANOTHER VALID SUMMARY

One study favors OpenScholar; it does not establish equivalence across domains.

SIMILAR, BUT INCORRECT

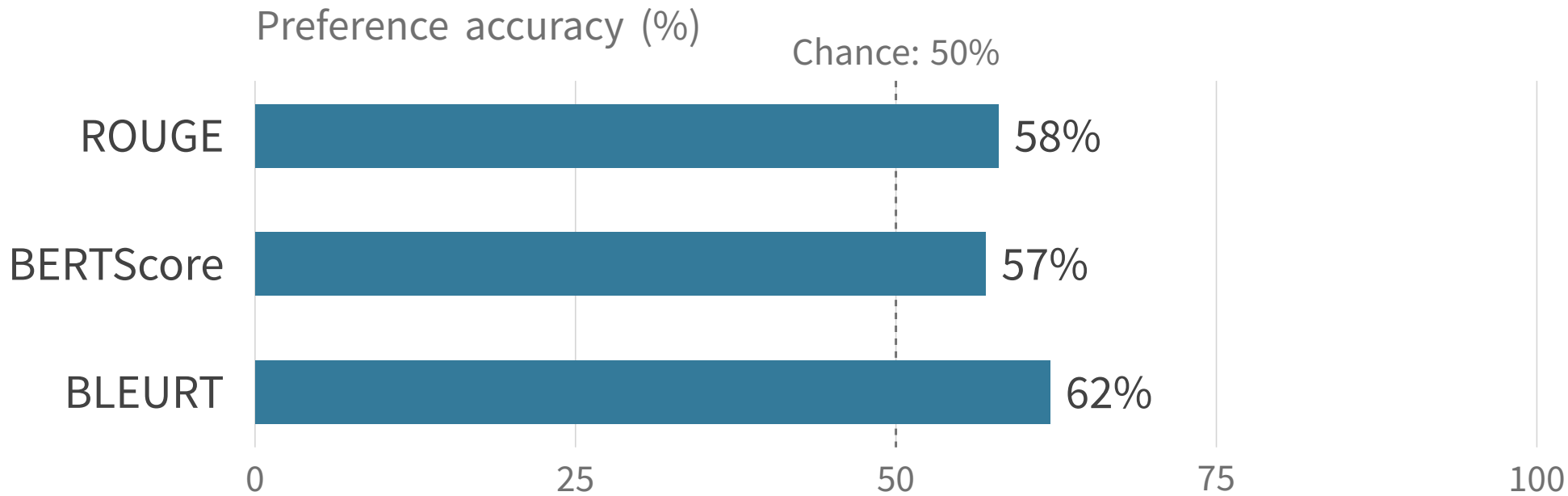
Experts preferred OpenScholar in one study. Parity across domains is **proven**.

⚠ **Several valid answers**

⚠ **Similarity to gold is not enough**

Similarity metrics can misrank answers

Does a higher similarity score select the answer experts prefer?



109 expert comparisons with reference answers

A Critical Evaluation of Evaluations for Long-form Question Answering. ACL 2023.

ScholarQABench: evaluate with expert rubrics

Check each rubric item against the answer, then combine the scores.

QUESTION

Can AI agents synthesize scientific literature as well as human experts?

[1] Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.

ScholarQABench: evaluate with expert rubrics

Check each rubric item against the answer, then combine the scores.

QUESTION

Can AI agents synthesize scientific literature as well as human experts?

→

RUBRIC ITEM	Weight	Check
Reports a direct comparison with human experts.	2	
Qualifies the conclusion by task and domain.	1	

[1] Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.

ScholarQABench: evaluate with expert rubrics

Check each rubric item against the answer, then combine the scores.

QUESTION

Can AI agents synthesize scientific literature as well as human experts?

CANDIDATE ANSWER

OpenScholar wins 70% of expert comparisons [1]. This holds across all fields.

→

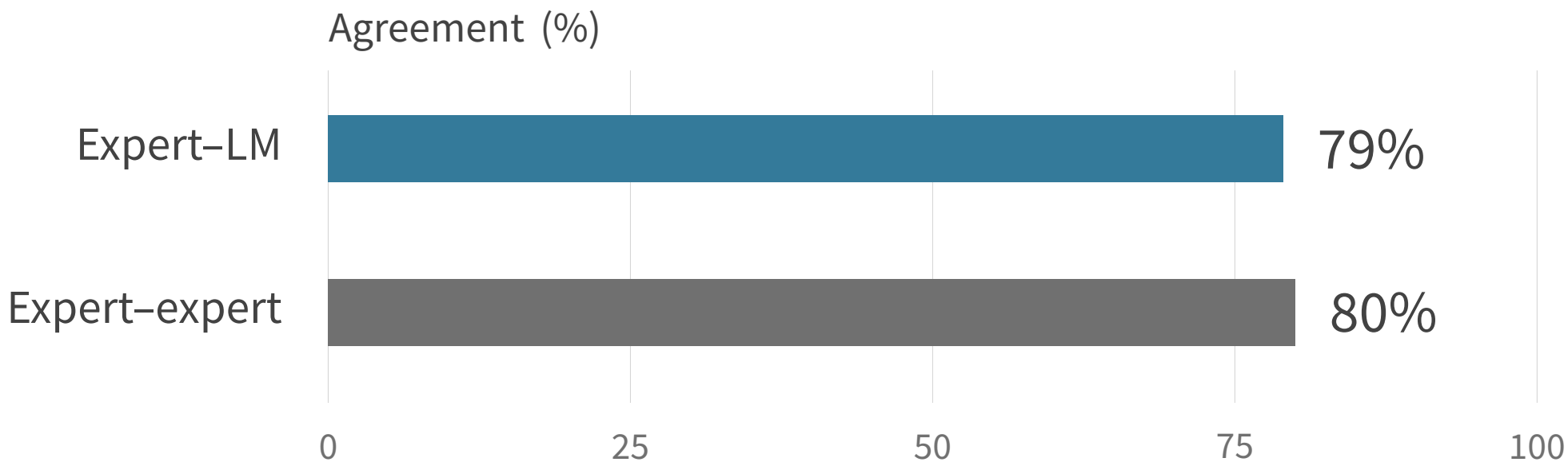
RUBRIC ITEM	Weight	Check
Reports a direct comparison with human experts.	2	1
Qualifies the conclusion by task and domain.	1	0

$$S_c = \frac{\sum_i w_i b_i}{\sum_i w_i} = \frac{2 \times 1 + 1 \times 0}{3} = 0.67$$

[1] Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.

Do rubric judgments agree with experts?

Compare judgments on the same rubric items.



12 questions; 2 systems; 2 expert raters per answer.

Synthesizing scientific literature with retrieval-augmented language models. Nature 2026.

ResearchQA: generate rubrics from surveys

Use LMs to scale up rubric annotation.

21K questions / 75 fields / **160K rubric items**

Survey-Mined Query 💬

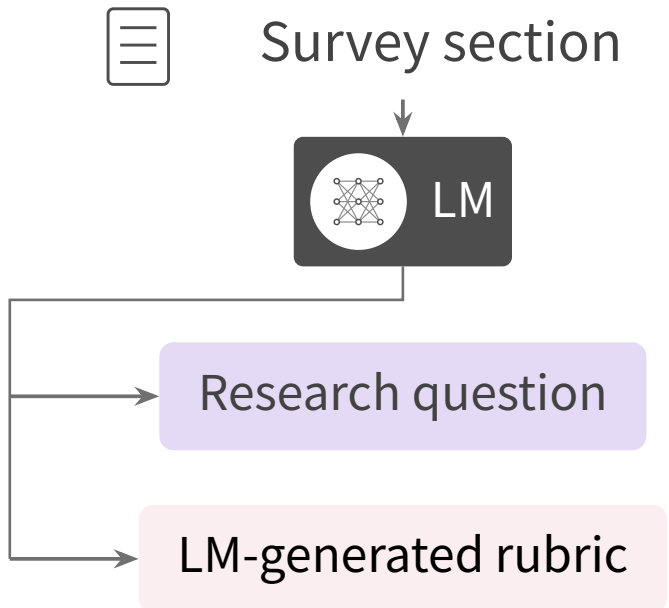
How does the frequency of terms in pre-training data influence numerical reasoning performance in few-shot settings? (💪 Engineering)

Research System (🤖): The frequency of terms in pre-training data significantly influences a model's numerical reasoning performance, particularly in few-shot learning scenarios [1]. Models pre-trained **[...]**

[1] Scaling Laws and Data Frequency Effects in Large Language **[...]**

Survey-Mined Evaluation Rubric 📋	**Judge** 🤖
Does the response reference the **"performance gap"** concept from the **Razeghi et al. (2022) paper** [...]?	0/4 Not at all covered
Does the response include **examples of studies or experiments** that investigate the impact of term frequency on numerical reasoning performance?	4/4 Completely covered
Does the response discuss the **correlation between the frequency of terms** in pre-training data and **numerical reasoning performance**?	1/4 Barely covered
Additional rubric items ...	...

Source Survey: The Mystery of In-Context Learning (Zhou et al., 2024)



ResearchQA: Evaluating Scholarly Question Answering at Scale Across 75 Fields with Survey-Mined Questions and Rubrics. TACL 2026.

Generated rubrics can encode the wrong target

Evaluate the rubric itself, before using it to evaluate an answer.

RESEARCHQA / EXPERT AUDIT

15%

of parametric rubric items were vacuous, erroneous, or unclear.

WHAT CAN GO WRONG?

A criterion cites a nonexistent paper.

A criterion merely restates the question.

A vague criterion lets plausible errors pass.

Ground criteria in sources, then audit their relevance and verifiability.

DeepResearch Bench II uses expert reports + manual revision + expert review.

ResearchQA: Evaluating Scholarly Question Answering at Scale Across 75 Fields with Survey-Mined Questions and Rubrics. TACL 2026.; DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Reports. arXiv, 2026.

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

BrowseComp / BrowseComp-Plus

Hard-to-find answers; a fixed corpus for controlled comparisons.

2 / DOMAIN EXPERTISE

MedBrowseComp
FinSearchComp

Expert knowledge guides medical and financial search.

3 / ANSWER QUALITY

ScholarQABench / ResearchQA

Rubrics assess response quality.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

BrowseComp / BrowseComp-Plus

Hard-to-find answers; a fixed corpus for controlled comparisons.

2 / DOMAIN EXPERTISE

MedBrowseComp
FinSearchComp

Expert knowledge guides medical and financial search.

3 / ANSWER QUALITY

ScholarQABench / ResearchQA

Rubrics assess response quality.

4 / EVIDENCE SUPPORT

Evidence use unclear

Correct answers do not establish evidence use or reliability.

DeepResearch Bench: check the citations

FACT checks whether the cited page supports each claim.

1 / EXTRACT CLAIM + URL

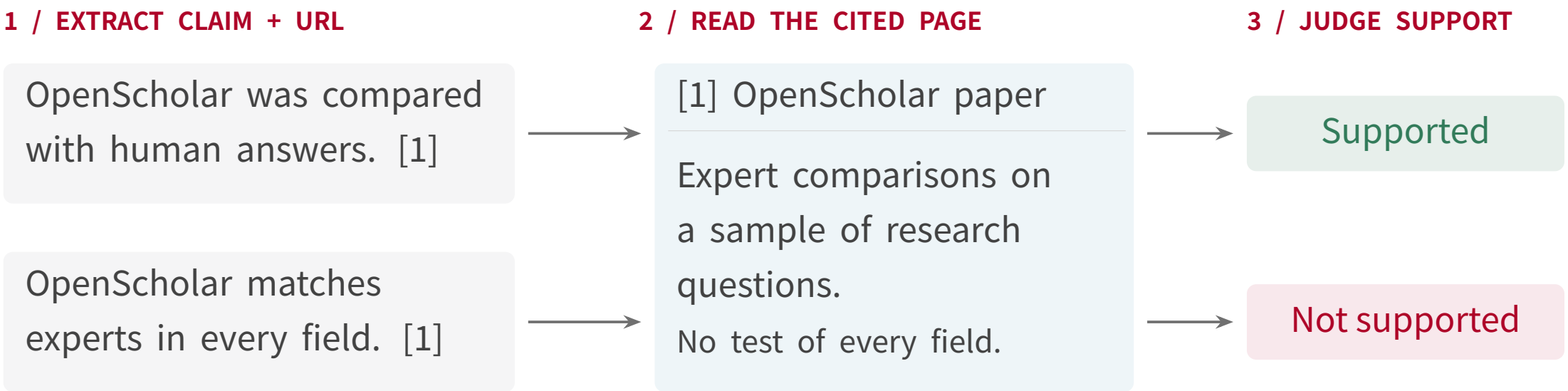
> OpenScholar was compared with human answers. [1]

> OpenScholar matches experts in every field. [1]

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

DeepResearch Bench: check the citations

FACT checks whether the cited page supports each claim.



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

DeepResearch Bench: check the citations

FACT checks whether the cited page supports each claim.

1 / EXTRACT CLAIM + URL

OpenScholar was compared with human answers. [1]

OpenScholar matches experts in every field. [1]

2 / READ THE CITED PAGE

[1] OpenScholar paper

Expert comparisons on a sample of research questions.

No test of every field.

3 / JUDGE SUPPORT

Supported

Not supported

Citation accuracy: $1/2 = 50\%$

Effective citations: 1 supported pair

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

From QA to deep research benchmarks

1 / SEARCH COMPLEXITY

BrowseComp / BrowseComp-Plus

Hard-to-find answers; a fixed corpus for controlled comparisons.

2 / DOMAIN EXPERTISE

MedBrowseComp
FinSearchComp

Expert knowledge guides medical and financial search.

3 / ANSWER QUALITY

ScholarQABench / ResearchQA

Rubrics assess response quality.

4 / EVIDENCE SUPPORT

DeepResearch Bench

Check whether cited sources support the report's claims.

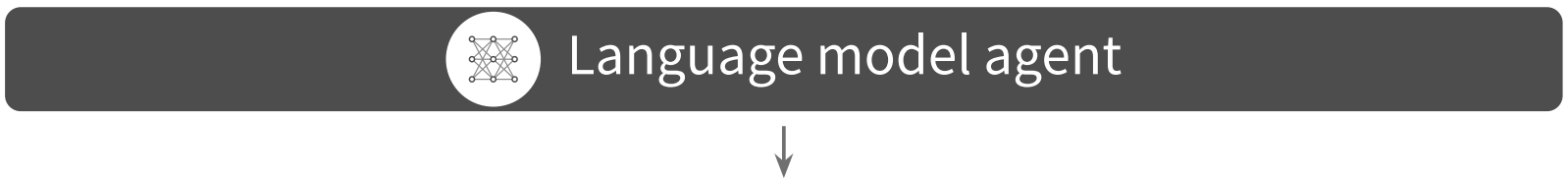
03

Agent Modeling

Deep research agent inference

CONTEXT SO FAR

USER / QUESTION
Can agents match experts?



1 / GENERATE REASONING

<think> Find expert comparisons </think>

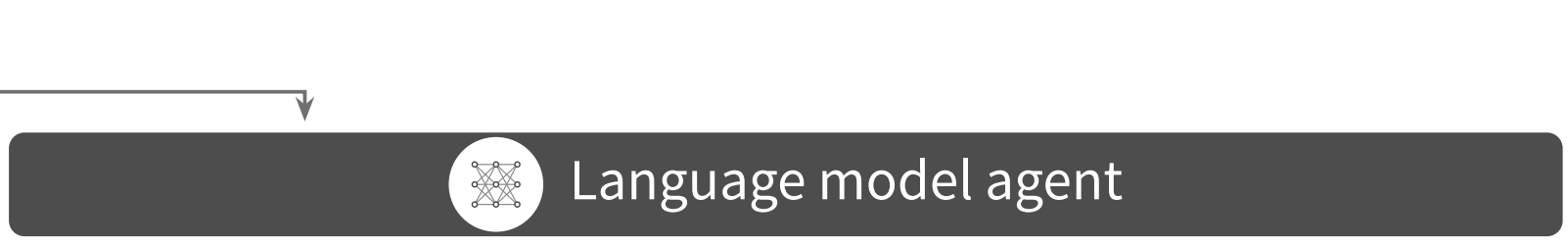
Blue: model-generated text Gray: user input and tool output

Deep research agent inference

CONTEXT SO FAR

USER / QUESTION
Can agents match experts?

AGENT / REASONING
Find expert comparisons.



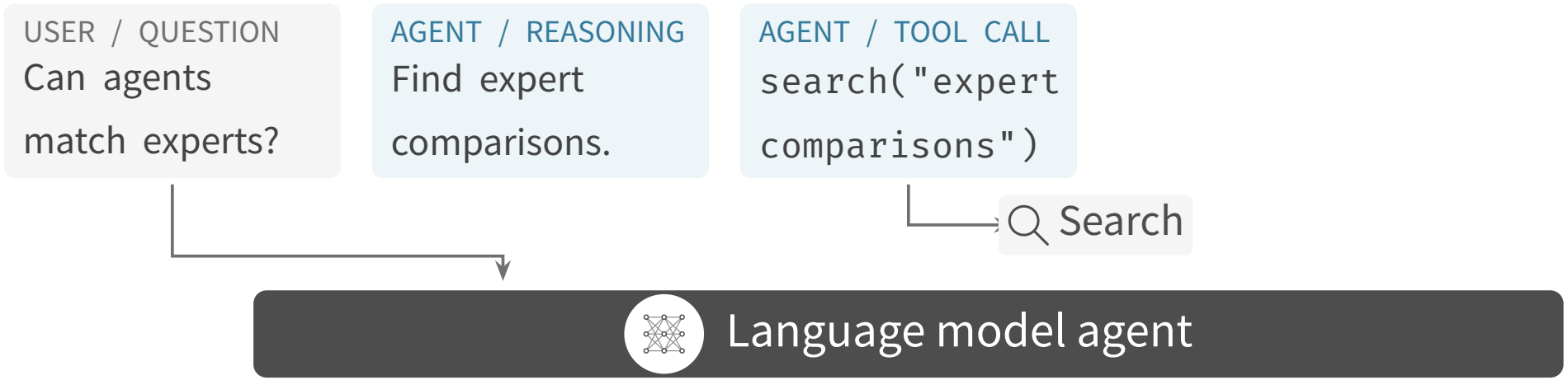
2 / GENERATE A TOOL CALL

```
<tool>  search(  "expert comparisons"  )  </tool>
```

Blue: model-generated text Gray: user input and tool output

Deep research agent inference

CONTEXT SO FAR

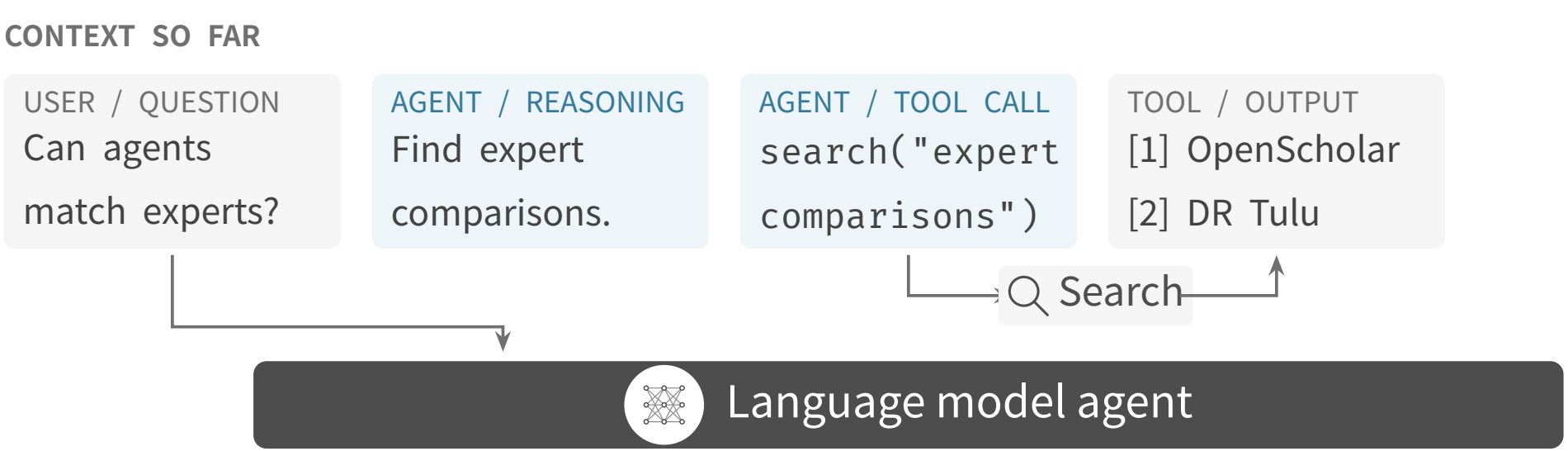


3 / EXECUTE THE SEARCH CALL

The search tool runs the generated query.

Blue: model-generated text Gray: user input and tool output

Deep research agent inference

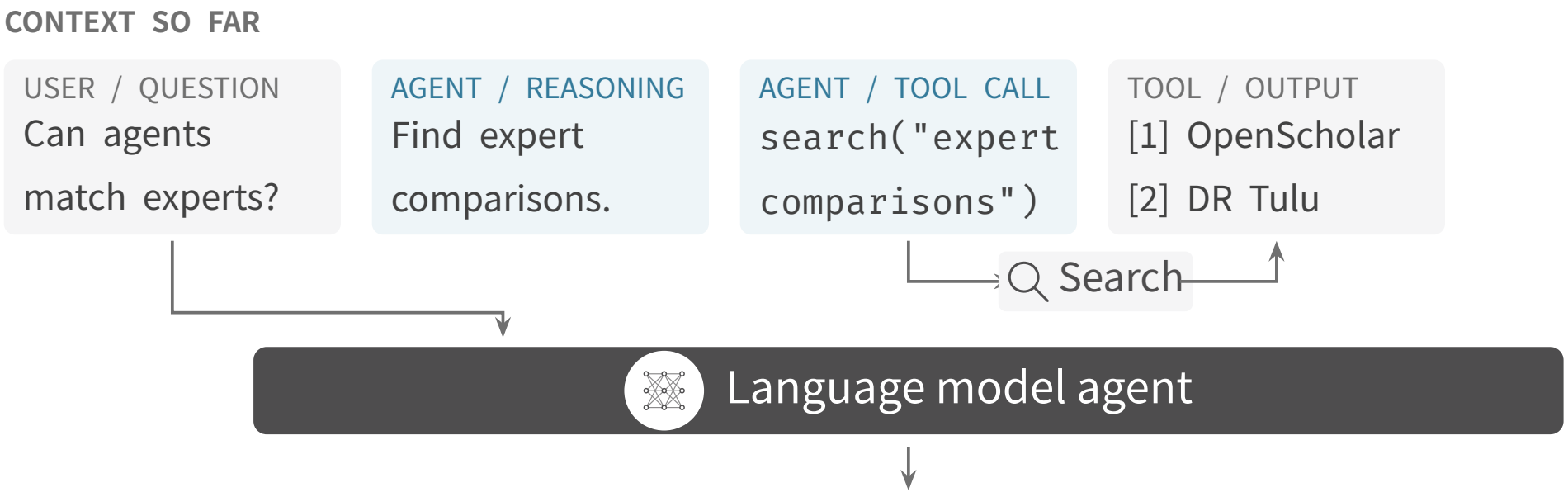


4 / APPEND THE TOOL OUTPUT

Retrieved text becomes part of the agent's context.

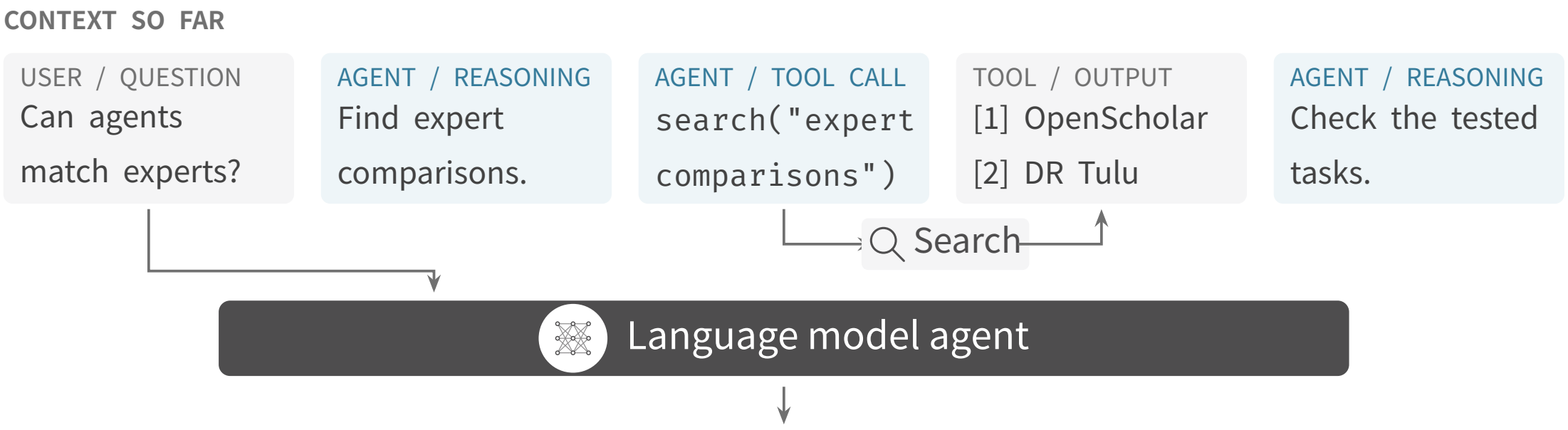
Blue: model-generated text Gray: user input and tool output

Deep research agent inference



Blue: model-generated text Gray: user input and tool output

Deep research agent inference



Blue: model-generated text Gray: user input and tool output

A training recipe for deep research

Learn from teacher demonstrations, then from your own attempts.

Mid-training

Learn agent behavior at scale.

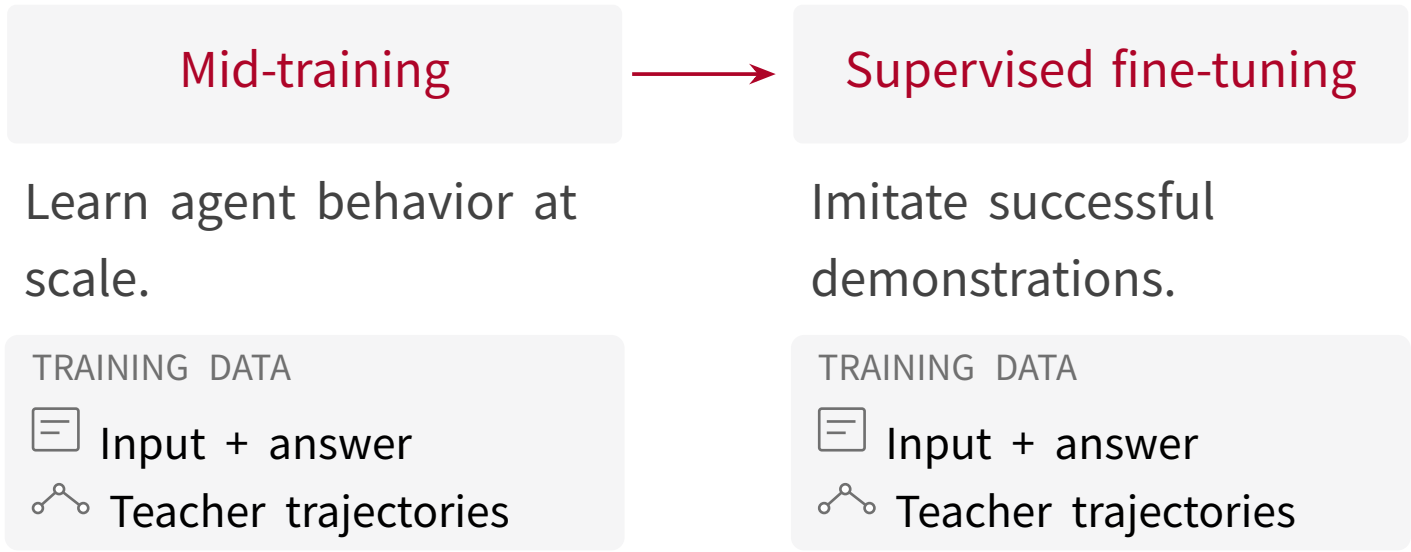
TRAINING DATA

- Input + answer
- Teacher trajectories

Tongyi DeepResearch Technical Report. arXiv, 2025.

A training recipe for deep research

Learn from teacher demonstrations, then from your own attempts.



Tongyi DeepResearch Technical Report. arXiv, 2025.

A training recipe for deep research

Learn from teacher demonstrations, then from your own attempts.

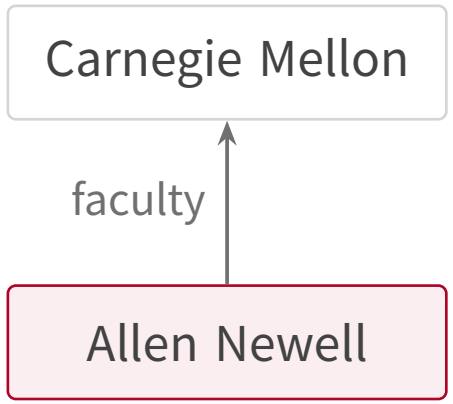
Mid-training	Supervised fine-tuning	Reinforcement learning
Learn agent behavior at scale.	Imitate successful demonstrations.	Learn from feedback on own attempts.
TRAINING DATA Input + answer Teacher trajectories	TRAINING DATA Input + answer Teacher trajectories	TRAINING DATA Input + answer Policy's own rollouts

Next: where do the tasks and trajectories come from?

Tongyi DeepResearch Technical Report. arXiv, 2025.

Knowledge graphs: entities and relationships

A node is an **entity**; a labeled arrow is a **relationship**.



Subject: Allen Newell

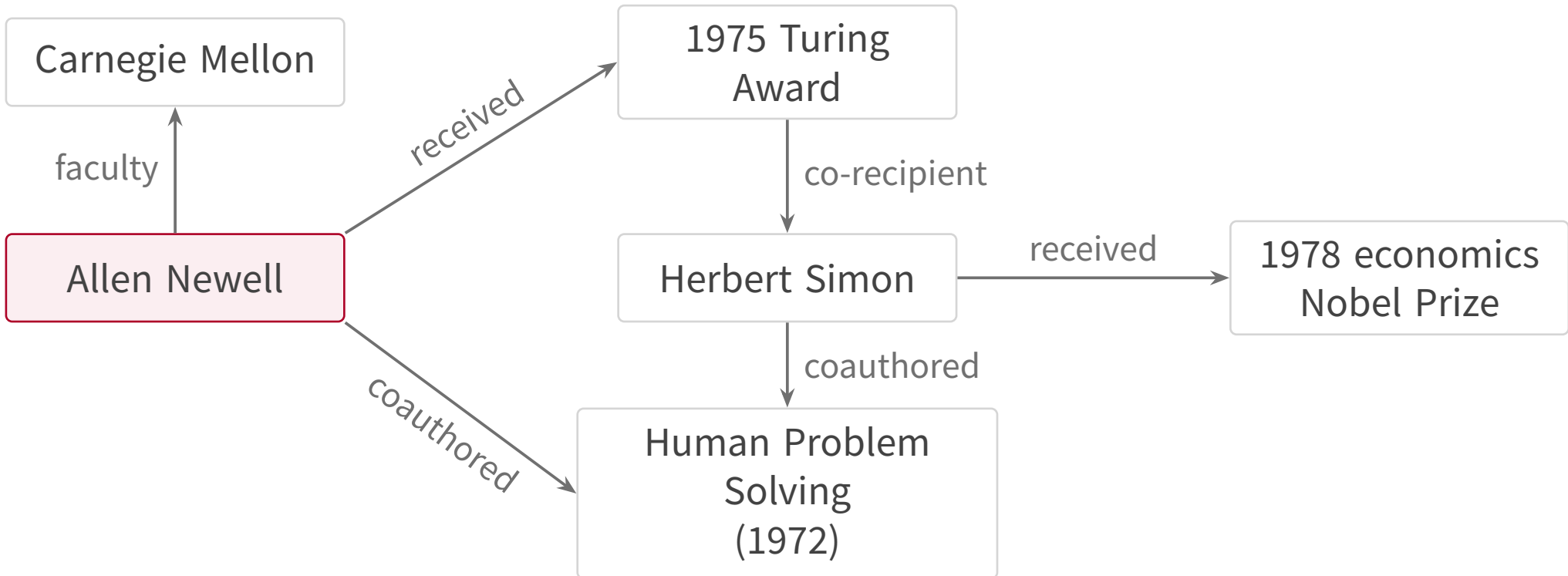
Relationship: faculty member of

Object: Carnegie Mellon

CMU Archives; Nobel Prize (1978).

Knowledge graphs: entities and relationships

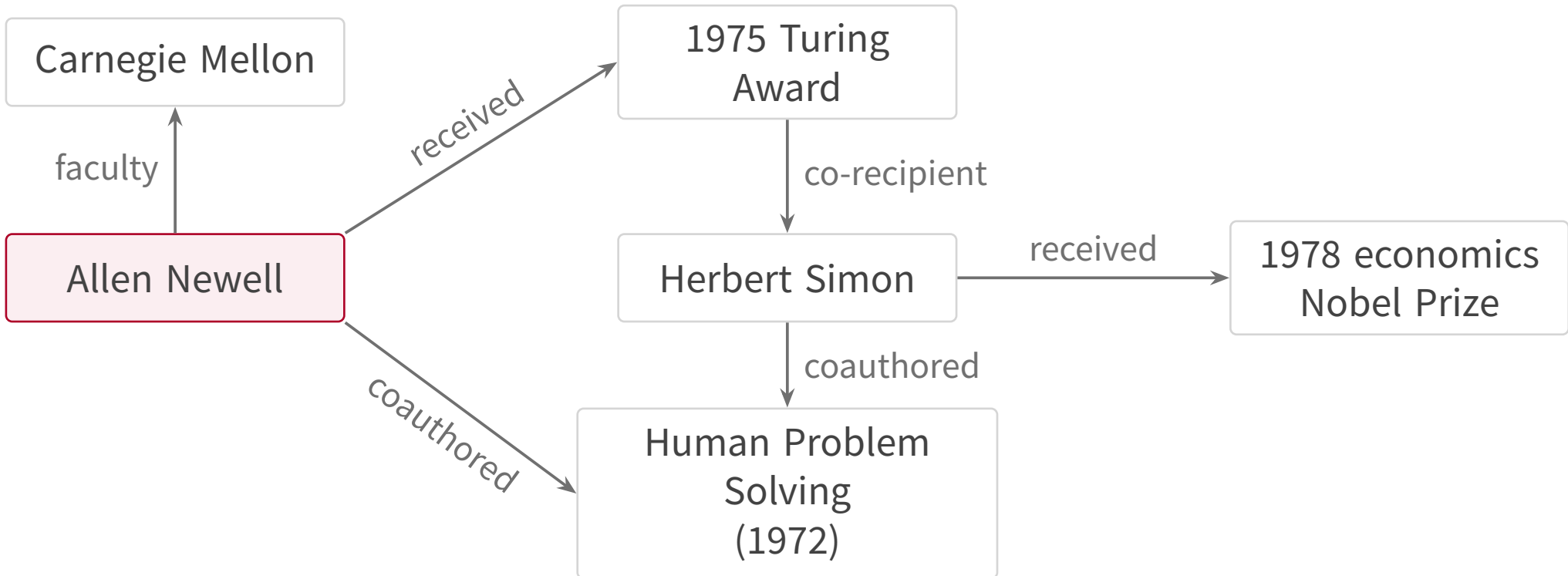
A node is an **entity**; a labeled arrow is a **relationship**.



CMU Archives; Nobel Prize (1978).

Data creation: sample a knowledge graph

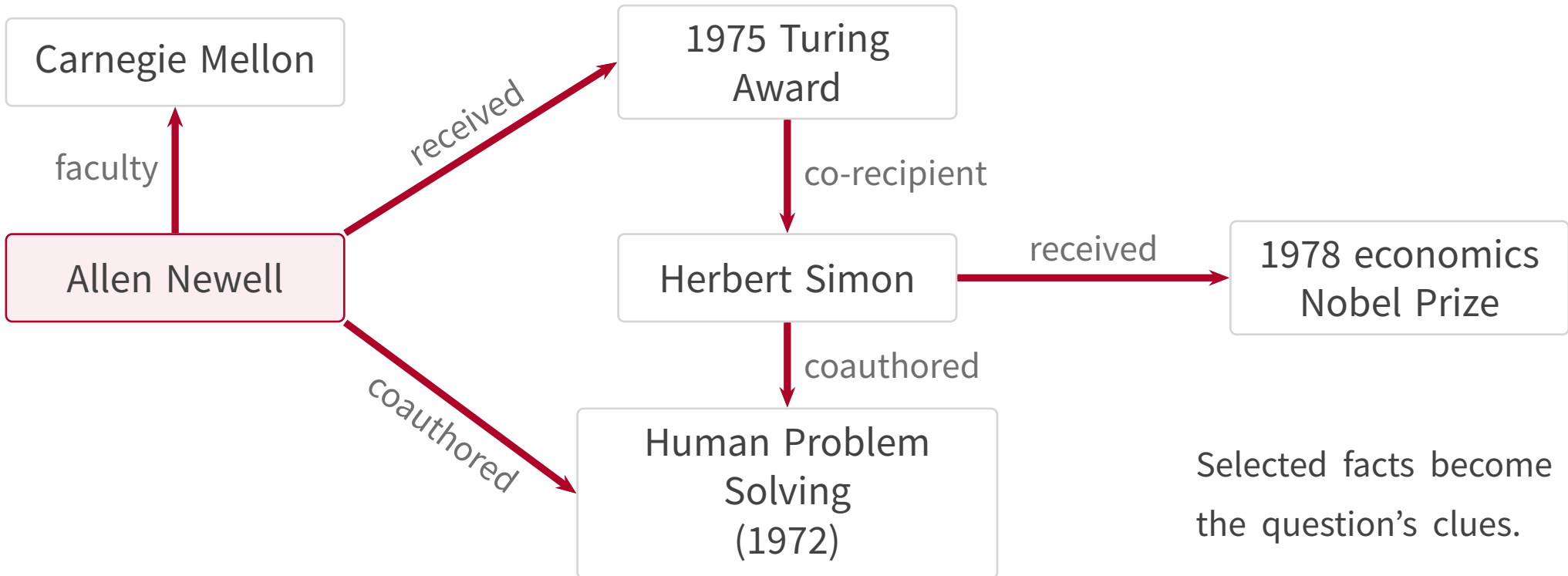
Start with sourced facts; sample a connected subgraph to create a question.



Method: WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning. ICLR 2026.; facts: CMU Archives, Nobel Prize (1978).

Data creation: sample a knowledge graph

Start with sourced facts; sample a connected subgraph to create a question.



Selected facts become the question's clues.

Method: WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning. ICLR 2026.; facts: CMU Archives, Nobel Prize (1978).

Data creation: turn facts into a question

Hide the target name and turn its relations into clues.

FACTS THE ANSWER MUST SATISFY

Newell: faculty at CMU.

Newell + Simon: coauthors of Human Problem Solving (1972).

Newell + Simon: shared the 1975 Turing Award.

Simon: 1978 economics Nobel Prize.

Method: WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning. ICLR 2026.; facts: CMU Archives, Nobel Prize (1978).

Data creation: turn facts into a question

Hide the target name and turn its relations into clues.

FACTS THE ANSWER MUST SATISFY

Newell: faculty at CMU.

Newell + Simon: coauthors of Human Problem Solving (1972).

Newell + Simon: shared the 1975 Turing Award.

Simon: 1978 economics Nobel Prize.

→

Which CMU researcher coauthored a book on human problem solving with a future economics Nobel laureate, and shared the Turing Award with that colleague three years before the Nobel Prize?

Method: WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning. ICLR 2026.; facts: CMU Archives, Nobel Prize (1978).

Data creation: turn facts into a question

Hide the target name and turn its relations into clues.

FACTS THE ANSWER MUST SATISFY

Newell: faculty at CMU.

Newell + Simon: coauthors of Human Problem Solving (1972).

Newell + Simon: shared the 1975 Turing Award.

Simon: 1978 economics Nobel Prize.

→

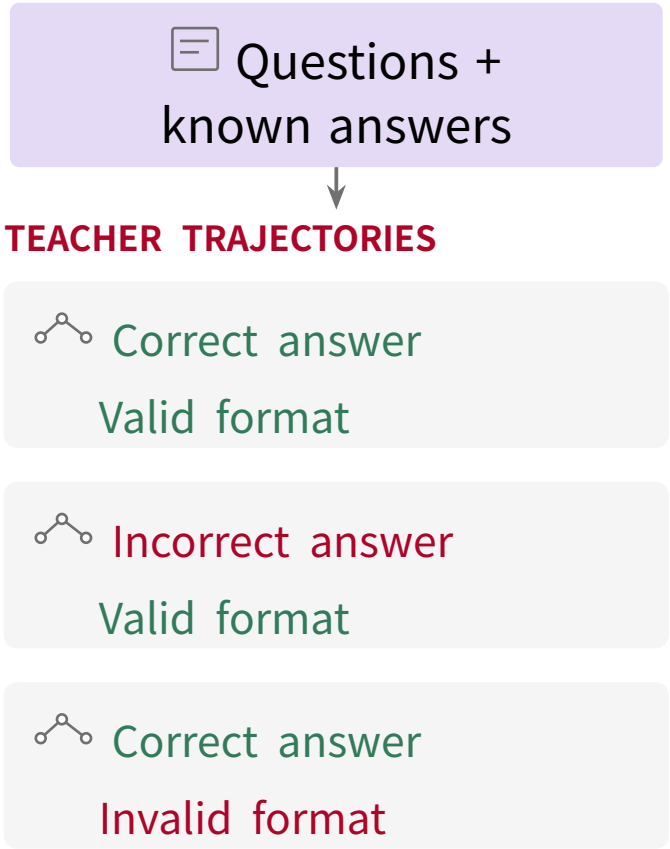
Which CMU researcher coauthored a book on human problem solving with a future economics Nobel laureate, and shared the Turing Award with that colleague three years before the Nobel Prize?

Known answer

Allen Newell

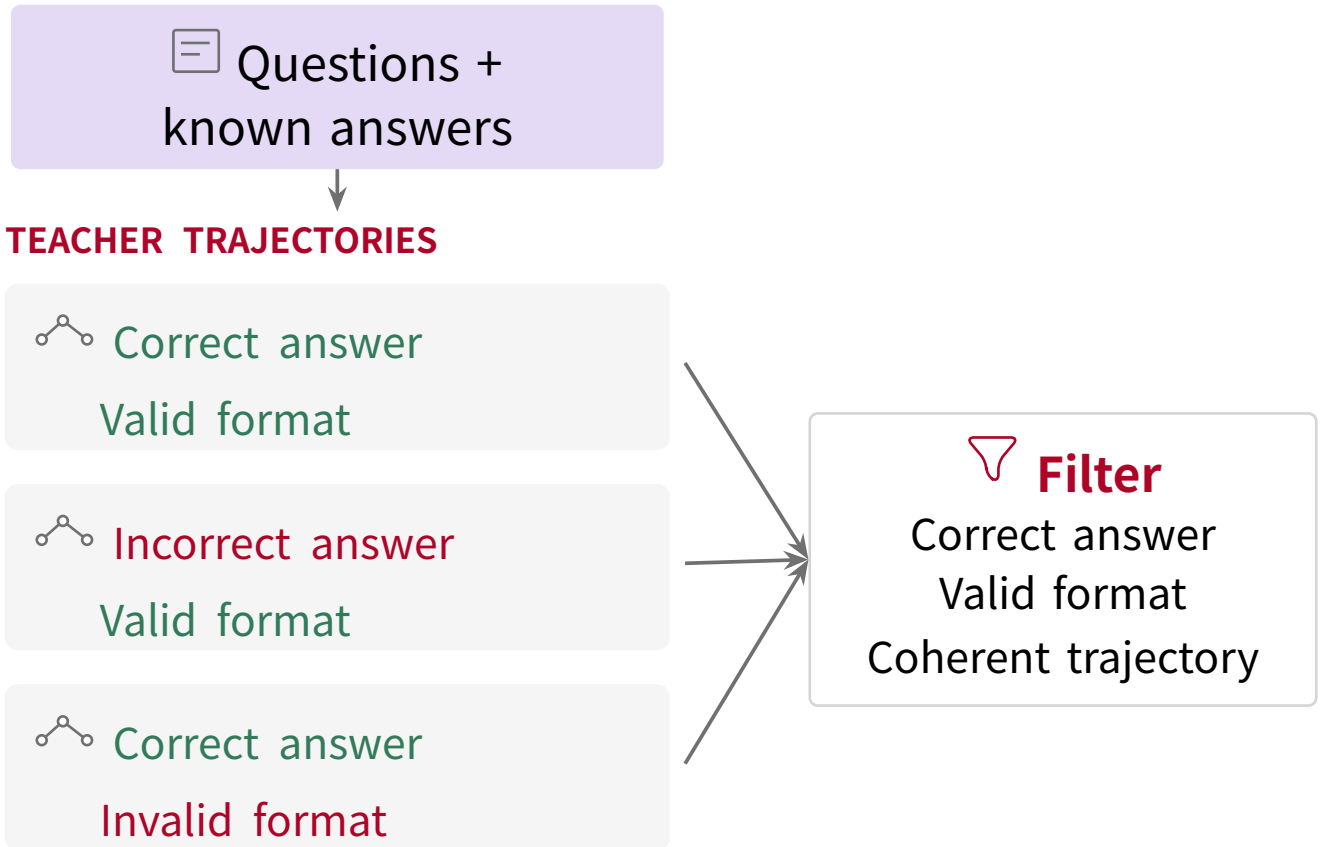
Method: WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning. ICLR 2026.; facts: CMU Archives, Nobel Prize (1978).

SFT: collect, filter, then imitate trajectories



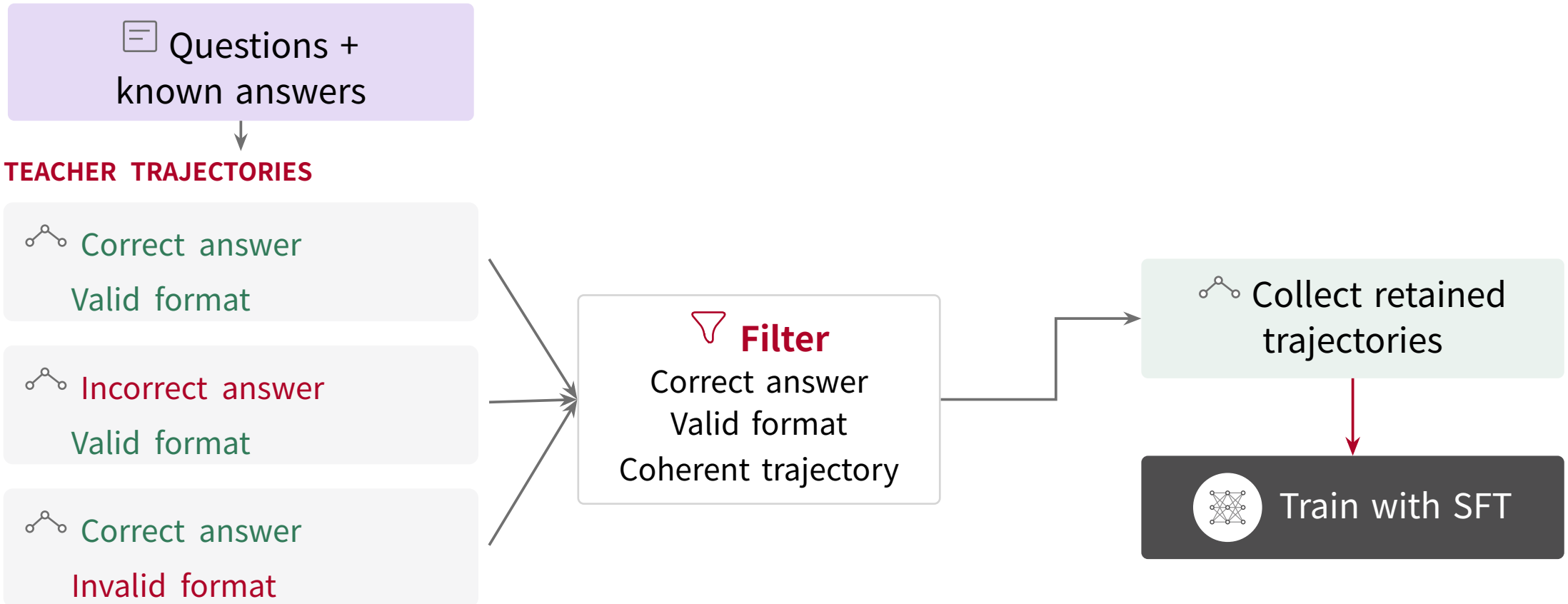
WebDancer: Towards Autonomous Information Seeking Agency. NeurIPS 2025.; Tongyi DeepResearch Technical Report. arXiv, 2025.

SFT: collect, filter, then imitate trajectories



WebDancer: Towards Autonomous Information Seeking Agency. NeurIPS 2025.; Tongyi DeepResearch Technical Report. arXiv, 2025.

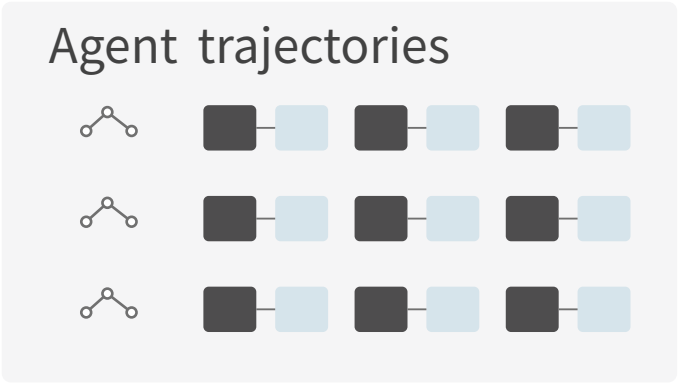
SFT: collect, filter, then imitate trajectories



WebDancer: Towards Autonomous Information Seeking Agency. NeurIPS 2025.; Tongyi DeepResearch Technical Report. arXiv, 2025.

Agentic mid-training

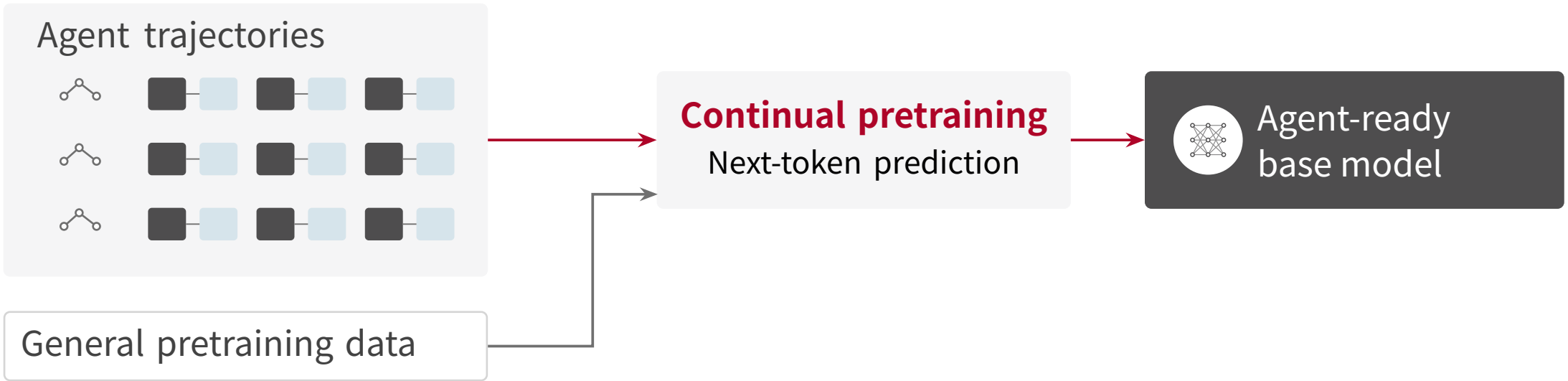
Tongyi uses continual pretraining to prepare the base model for agent workflows.



Scaling Agents via Continual Pre-training. ICLR 2026.

Agentic mid-training

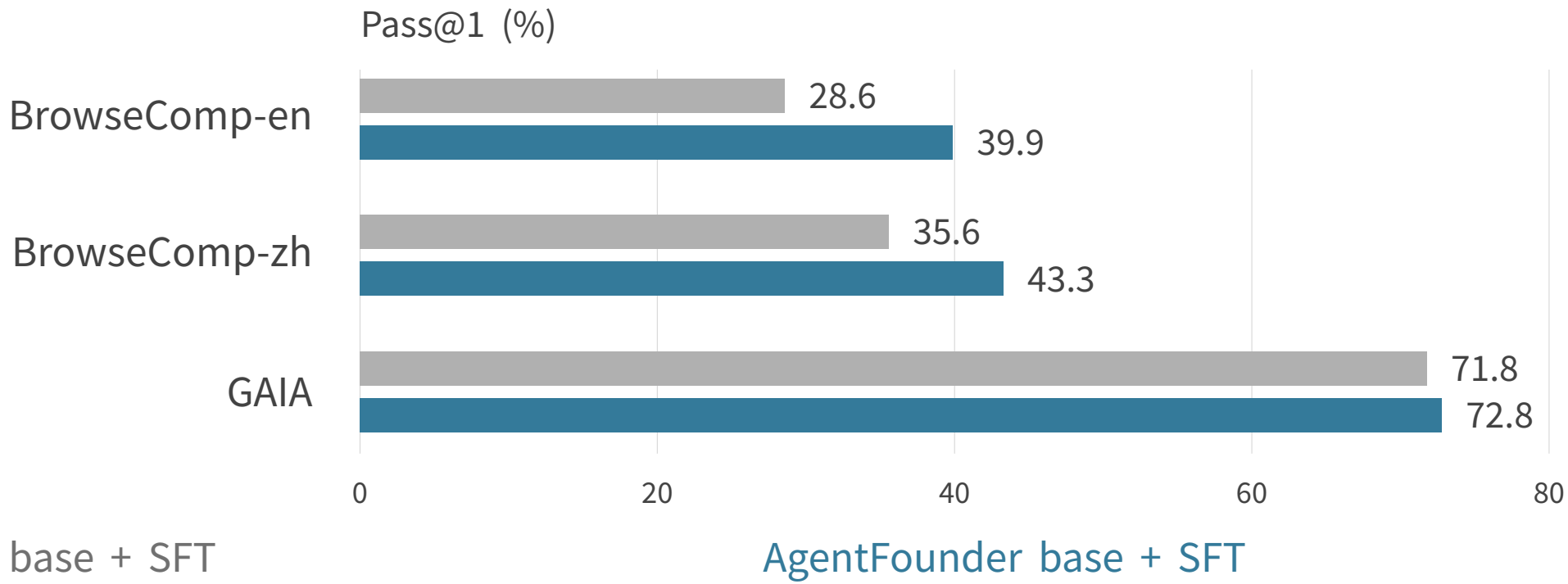
Tongyi uses continual pretraining to prepare the base model for agent workflows.



Scaling Agents via Continual Pre-training. ICLR 2026.

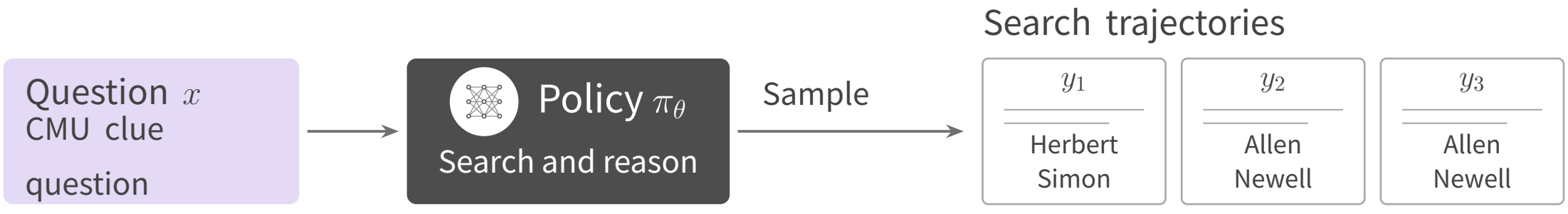
Mid-training improves the starting point for SFT

Same downstream SFT-B data; different base-model initialization.



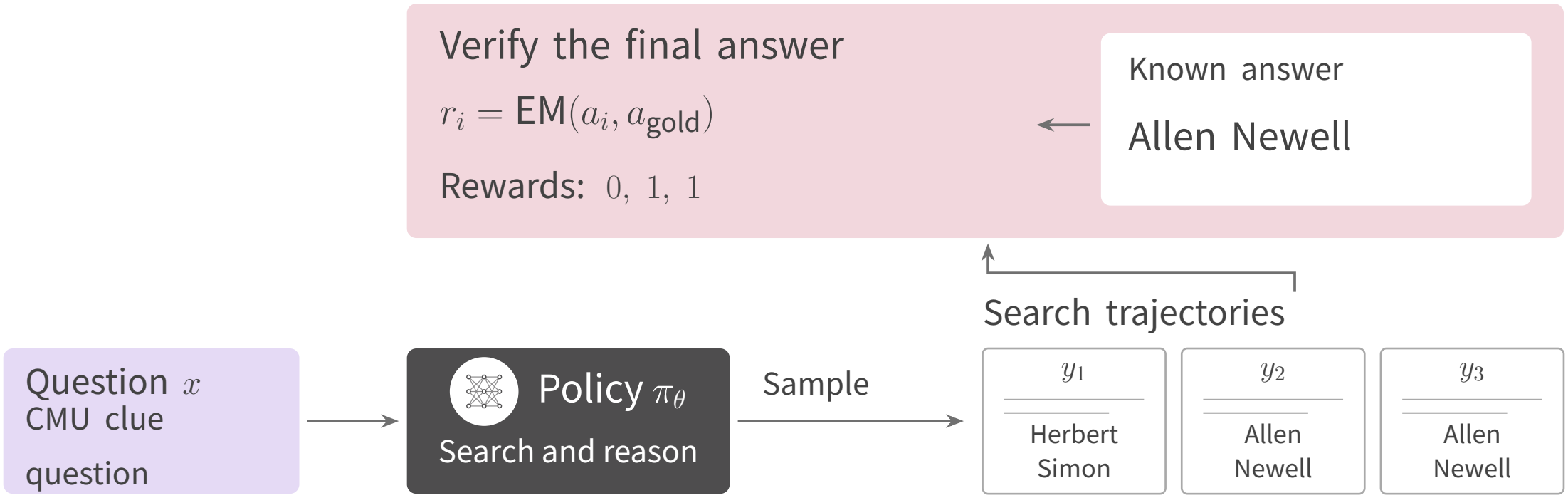
Scaling Agents via Continual Pre-training. ICLR 2026.

RLVR: learning from answer correctness



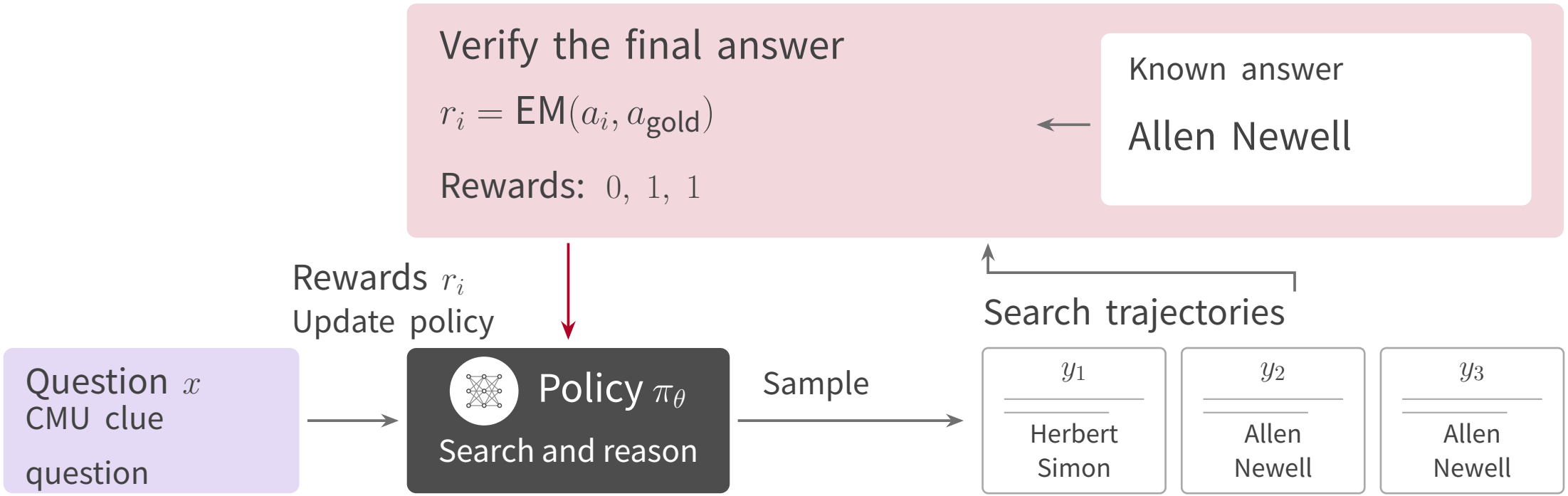
Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.: final-answer exact-match reward. CMU examples and rewards are illustrative.

RLVR: learning from answer correctness



Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.: final-answer exact-match reward. CMU examples and rewards are illustrative.

RLVR: learning from answer correctness



Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM 2025.: final-answer exact-match reward. CMU examples and rewards are illustrative.

GRPO: learn from better attempts

One question x; sample a group of attempts $y_1, \ldots, y_G$

Herbert Simon Reward: 0	Allen Newell Reward: 1	Allen Newell Reward: 1

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv, 2024.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

GRPO: learn from better attempts

One question x; sample a group of attempts $y_1, \ldots, y_G$

Herbert Simon Reward: 0	Allen Newell Reward: 1	Allen Newell Reward: 1
Below group average	Above group average	Above group average
Lower its probability	Raise its probability	Raise its probability

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv, 2024.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

GRPO: learn from better attempts

One question x; sample a group of attempts $y_1, \ldots, y_G$

Herbert Simon Reward: 0	Allen Newell Reward: 1	Allen Newell Reward: 1
Below group average	Above group average	Above group average
Lower its probability	Raise its probability	Raise its probability

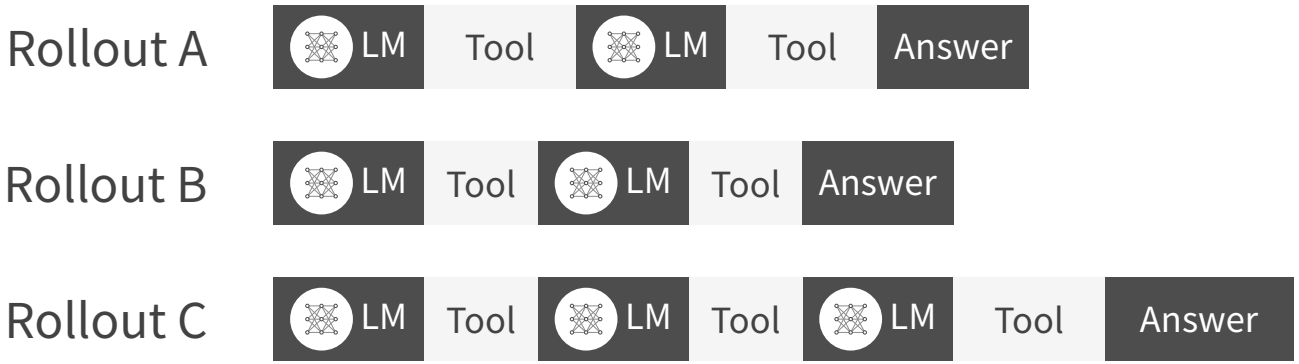
$$\mathcal{J}(\theta) = \mathbb{E}\left[\left\langle \min\left(\rho\,\widehat{A}, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)\,\widehat{A}\right) - \beta D_{\text{KL}} \right\rangle_{\text{model tokens}}\right]$$

$\widehat{A}$: relative reward	Clipped update	KL penalty
Compare within the group	Limit the update size	Stay near the reference

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv, 2024.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026..

Asynchronous rollouts: finish, score, then train

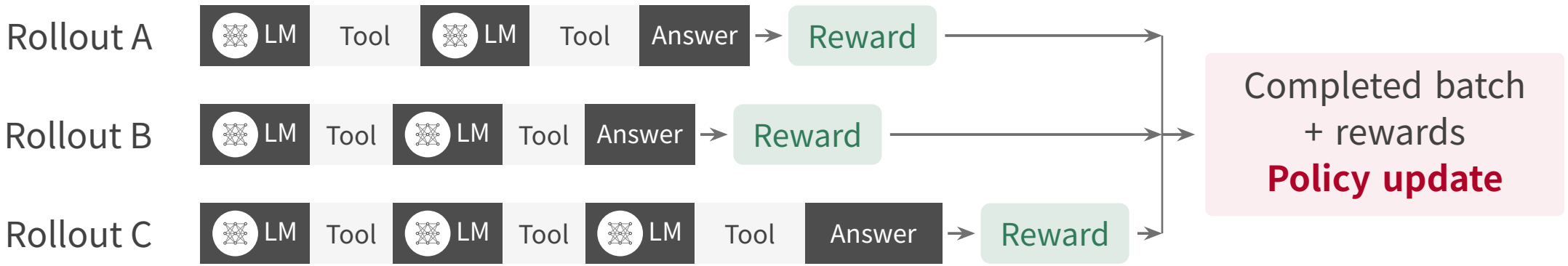
Score completed trajectories; assemble a batch for the policy update.



Tongyi DeepResearch Technical Report. arXiv, 2025.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

Asynchronous rollouts: finish, score, then train

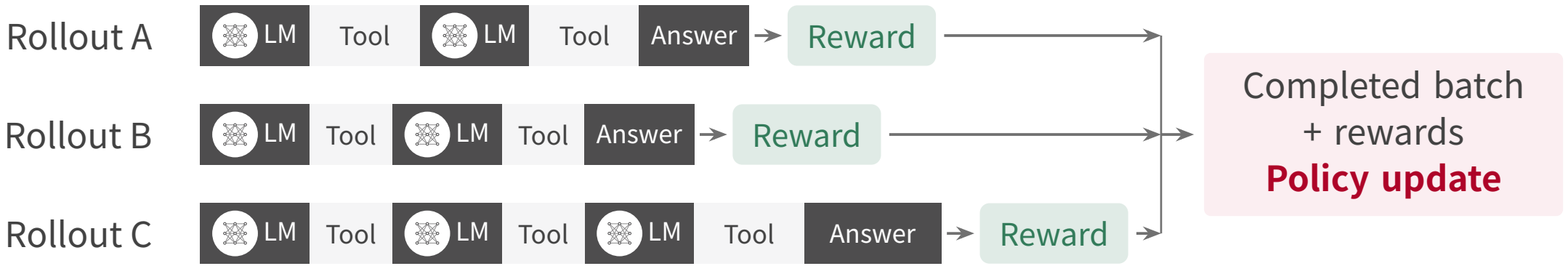
Score completed trajectories; assemble a batch for the policy update.



Tongyi DeepResearch Technical Report. arXiv, 2025.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

Asynchronous rollouts: finish, score, then train

Score completed trajectories; assemble a batch for the policy update.



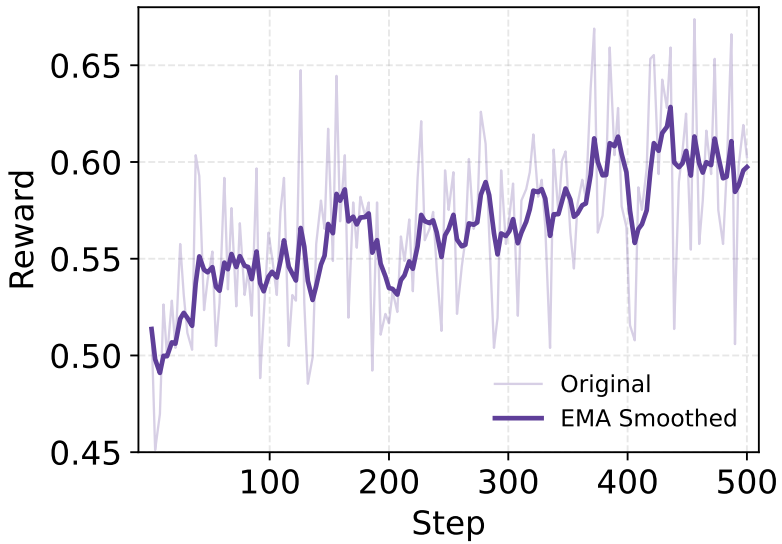
Additional efficiency: cached tool results, robust API handling, and background task curation

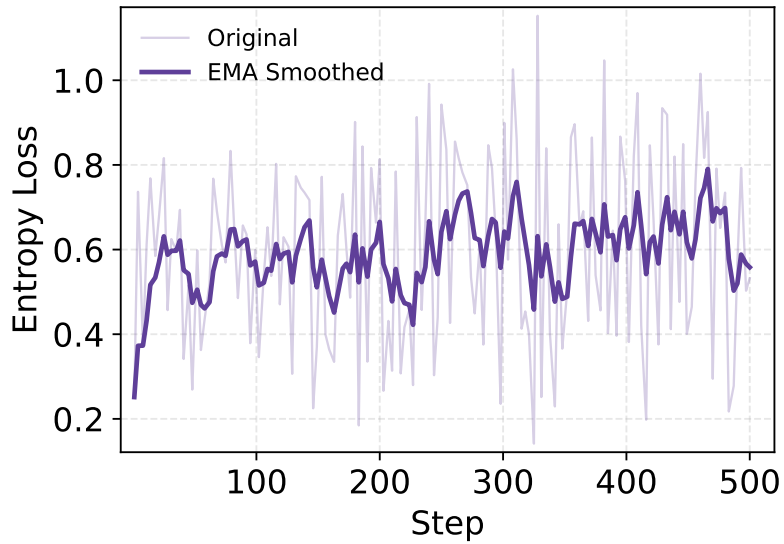
More on scheduling and training efficiency in the RL systems lecture

Tongyi DeepResearch Technical Report. arXiv, 2025.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

Tongyi: RL improves with training

Training reward increases while policy entropy stays relatively stable.





These are RL training curves; the report does not isolate each training stage.

Tongyi DeepResearch Technical Report. arXiv, 2025..

Open-ended reports need richer training signals

What should the reward measure?

VERIFIABLE SHORT ANSWER

Allen Newell

↓

☐ Answer correctness

OPEN-ENDED SYNTHESIS

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

Open-ended reports need richer training signals

What should the reward measure?

VERIFIABLE SHORT ANSWER

Allen Newell

↓

- ☐ Answer correctness

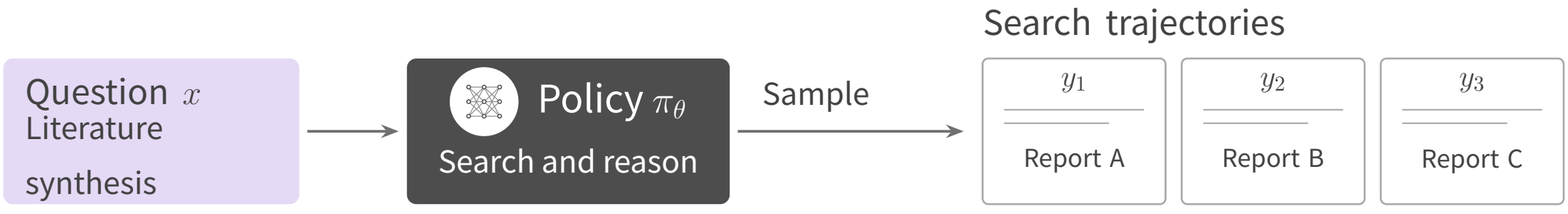
OPEN-ENDED SYNTHESIS

↓

- ☐ Coverage
- ☐ Relevance
- ☐ Factuality
- ☐ Supported claims (citations)

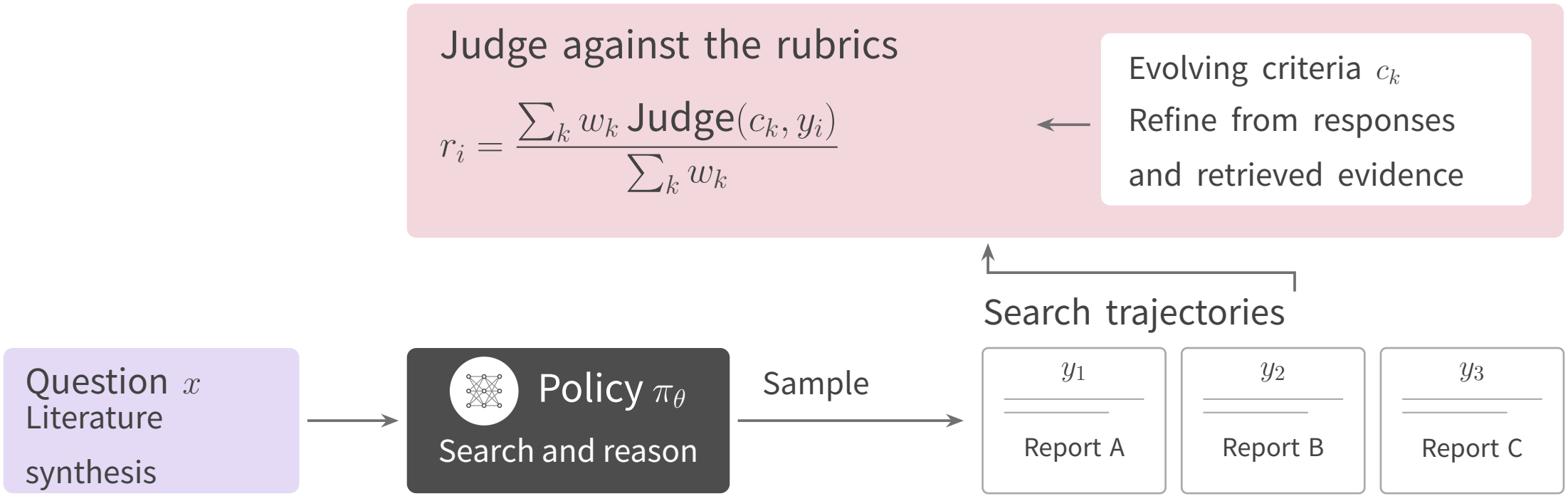
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

DR Tulu: evolving rubrics for RL



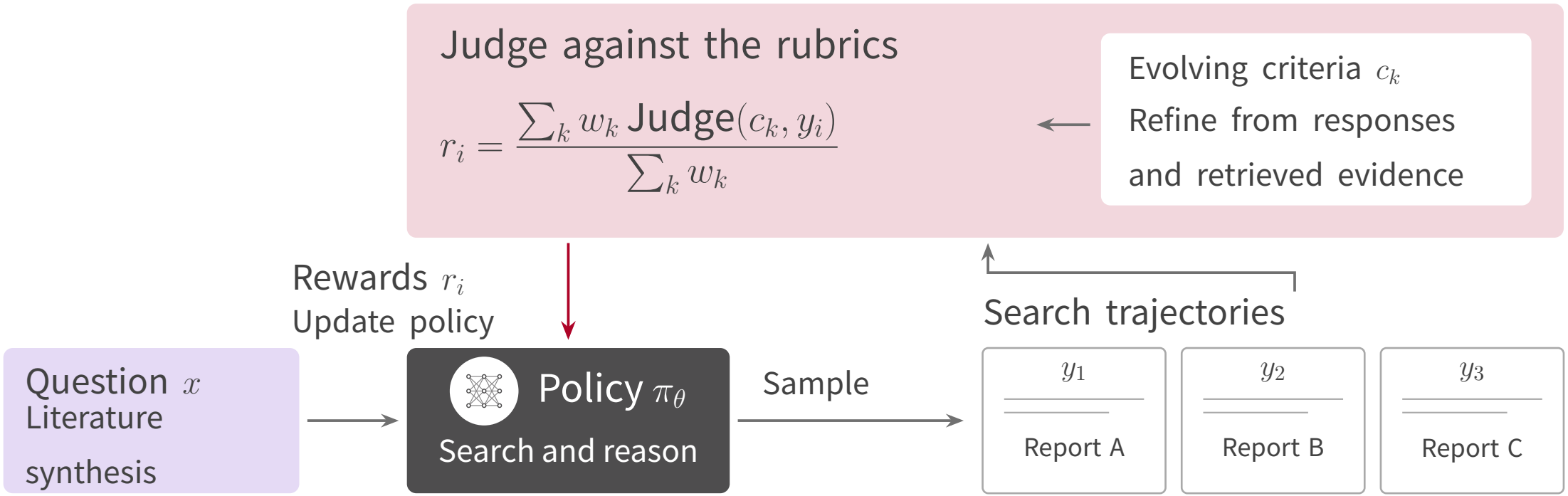
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.. Rubric reward shown; auxiliary rewards omitted.

DR Tulu: evolving rubrics for RL



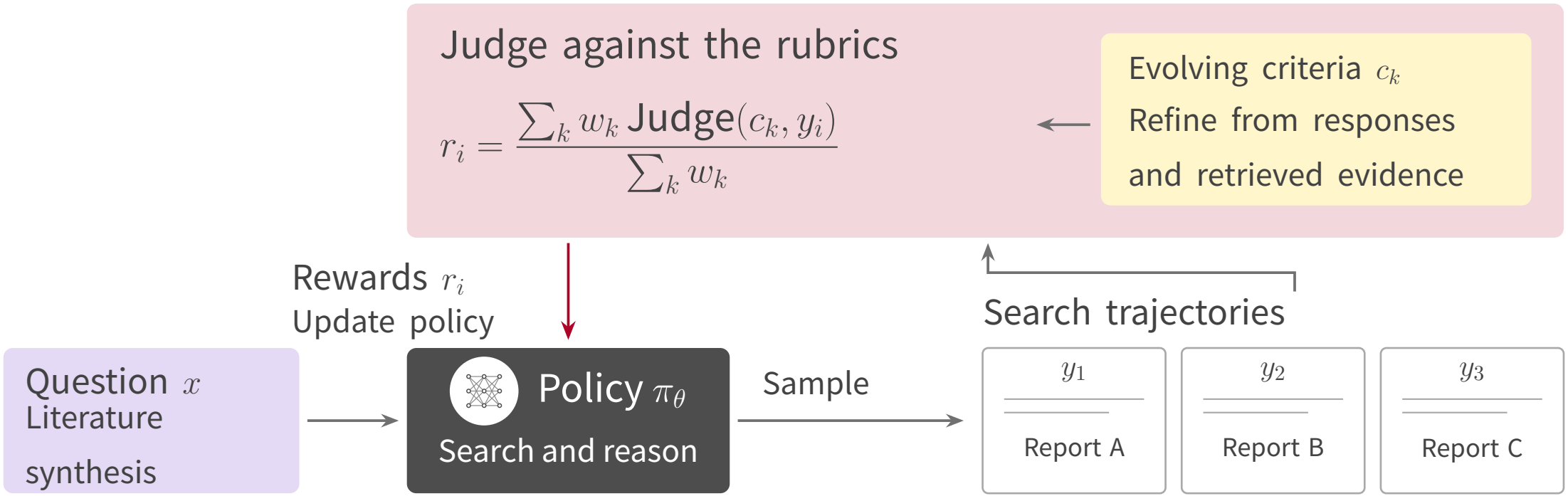
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.. Rubric reward shown; auxiliary rewards omitted.

DR Tulu: evolving rubrics for RL



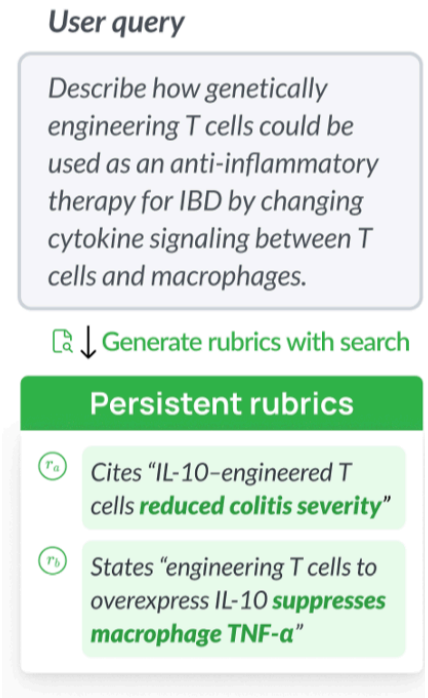
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.. Rubric reward shown; auxiliary rewards omitted.

DR Tulu: evolving rubrics for RL



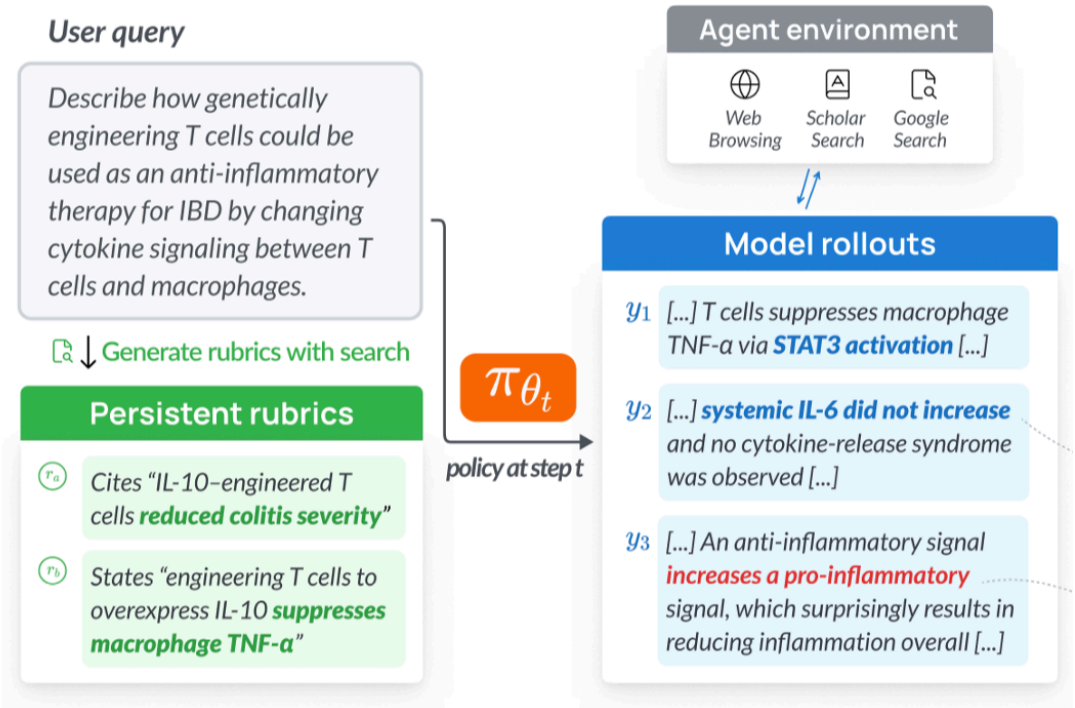
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.. Rubric reward shown; auxiliary rewards omitted.

How the rubrics evolve during training



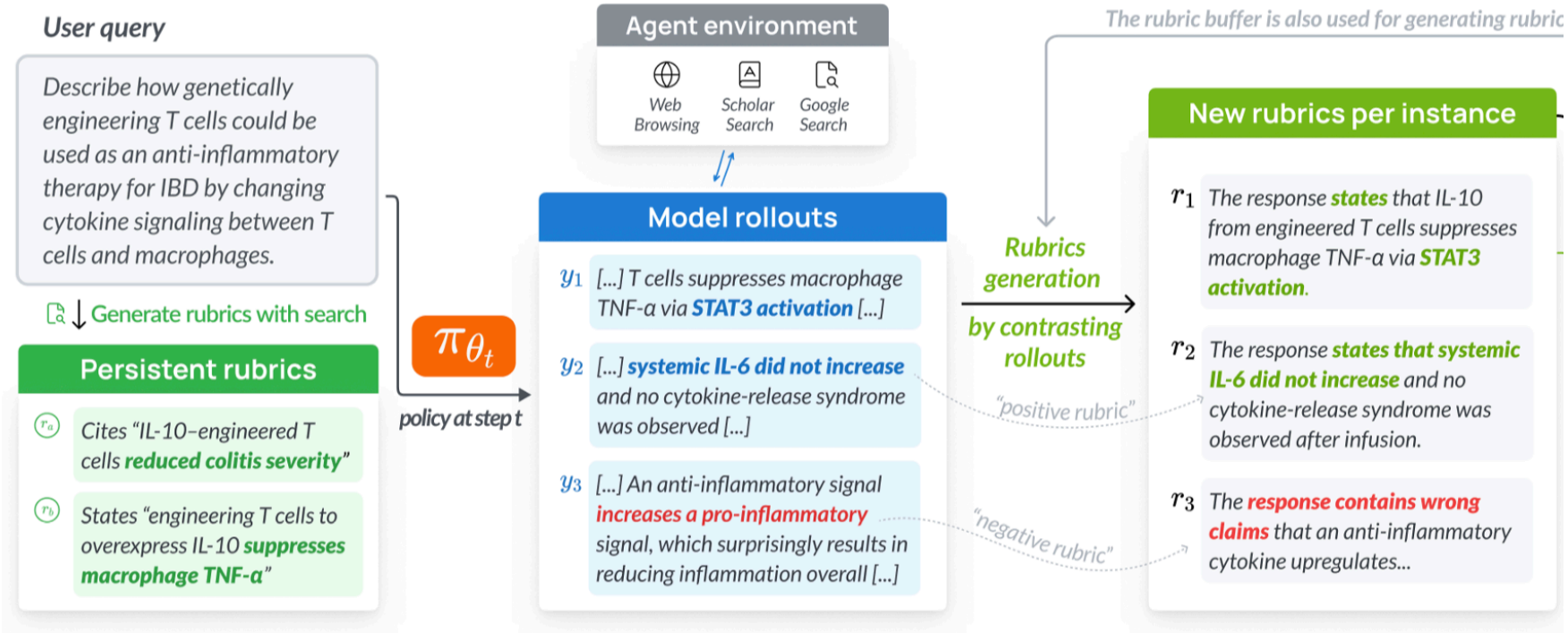
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

How the rubrics evolve during training



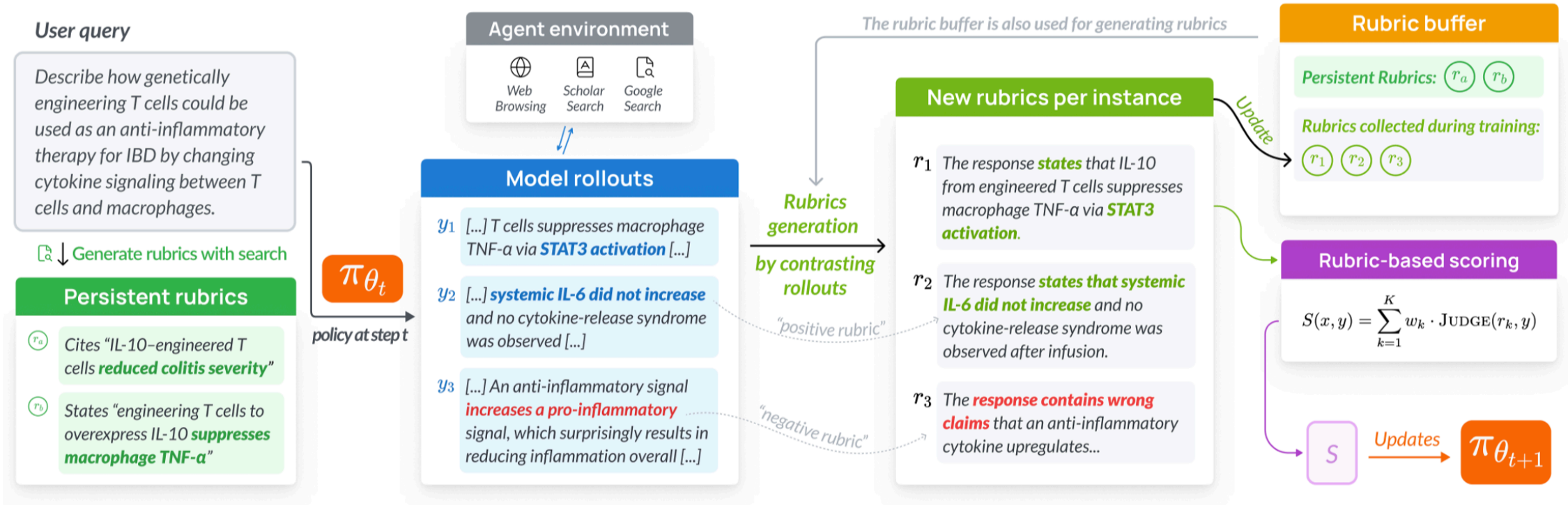
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

How the rubrics evolve during training



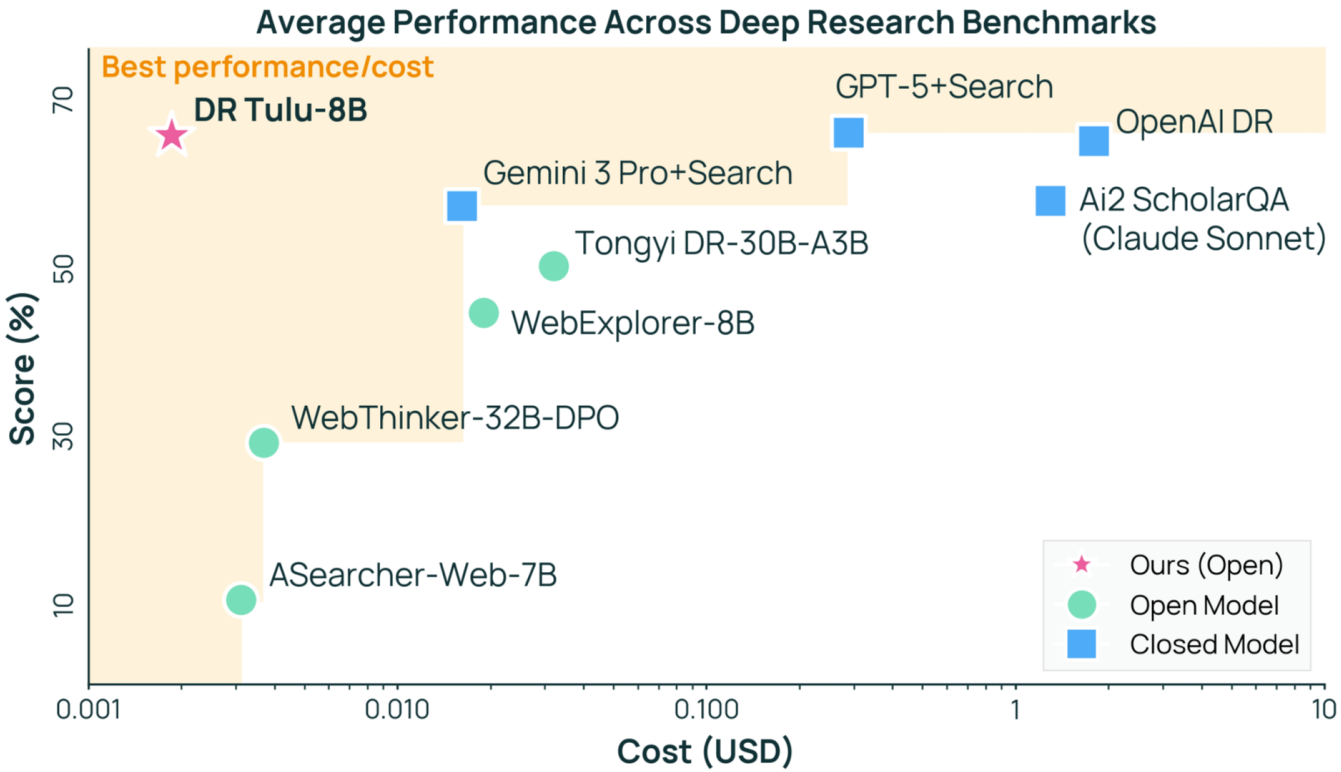
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

How the rubrics evolve during training



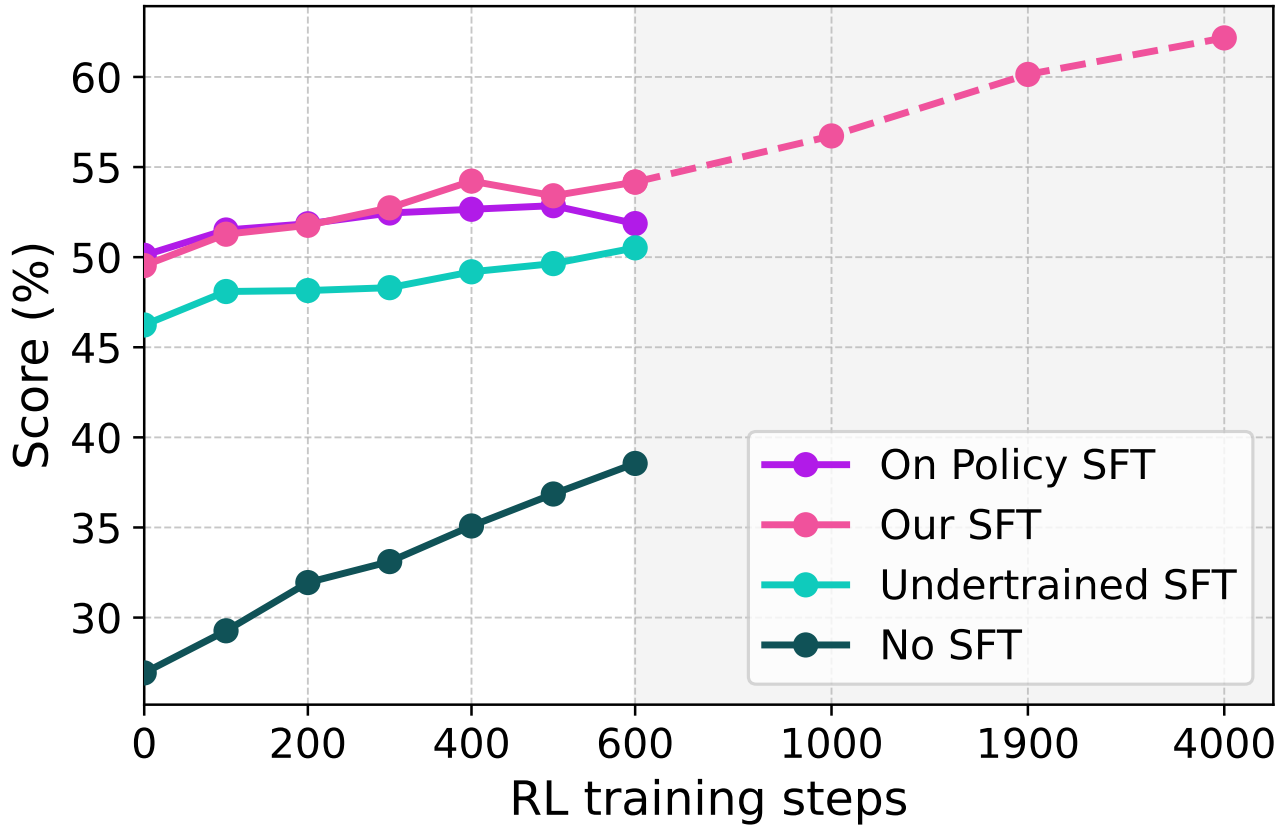
DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

DR Tulu: cost and research performance



DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

DR Tulu: SFT as an RL warm start



Even 5% SFT helps.

The full SFT mixture gives the strongest RL results.

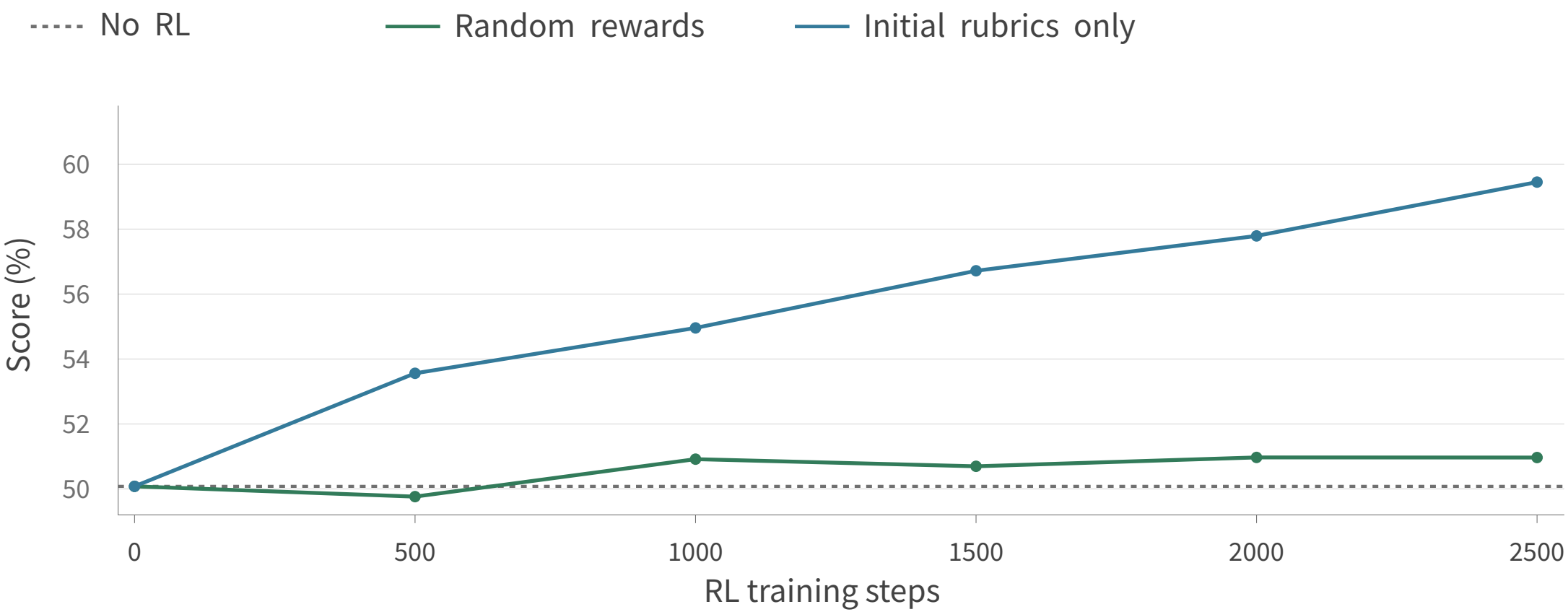
Gray region: unequal step spacing.

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

DR Tulu: evolving rubrics improve RL training

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.. Average over HealthBench, ScholarQABench v2, and DeepResearch Bench.

DR Tulu: evolving rubrics improve RL training



DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.. Average over HealthBench, ScholarQABench v2, and DeepResearch Bench.

DR Tulu: evolving rubrics improve RL training

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.. Average over HealthBench, ScholarQABench v2, and DeepResearch Bench.

04

Retrieval for Deep Research

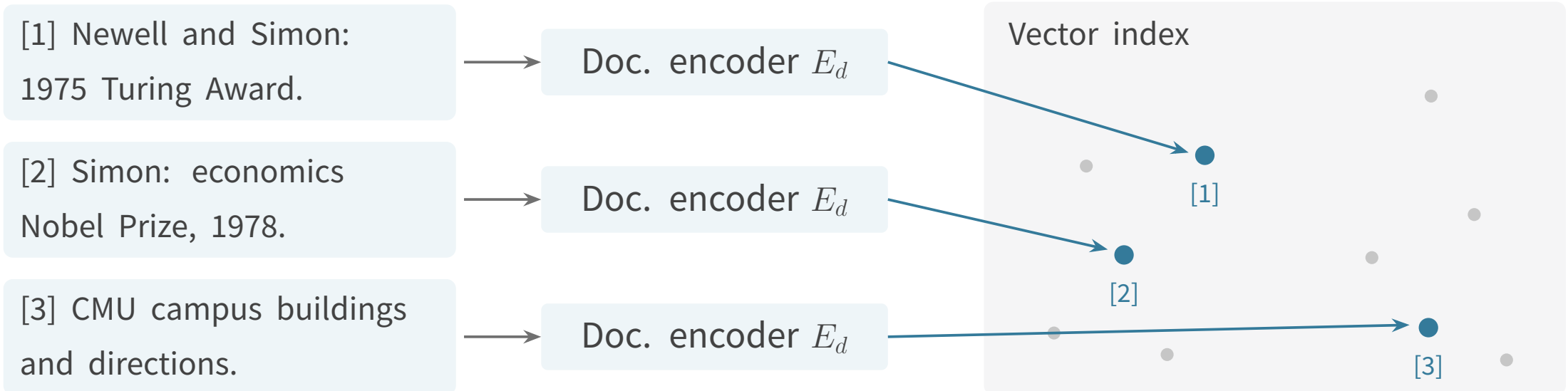
Search tools for deep research

Other tools: Python execution (Tongyi); scholarly search (Tongyi, DR Tulu).

Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.; Tongyi DeepResearch Technical Report. arXiv, 2025.; DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. ICML 2026.

Embedding retrieval finds nearby documents

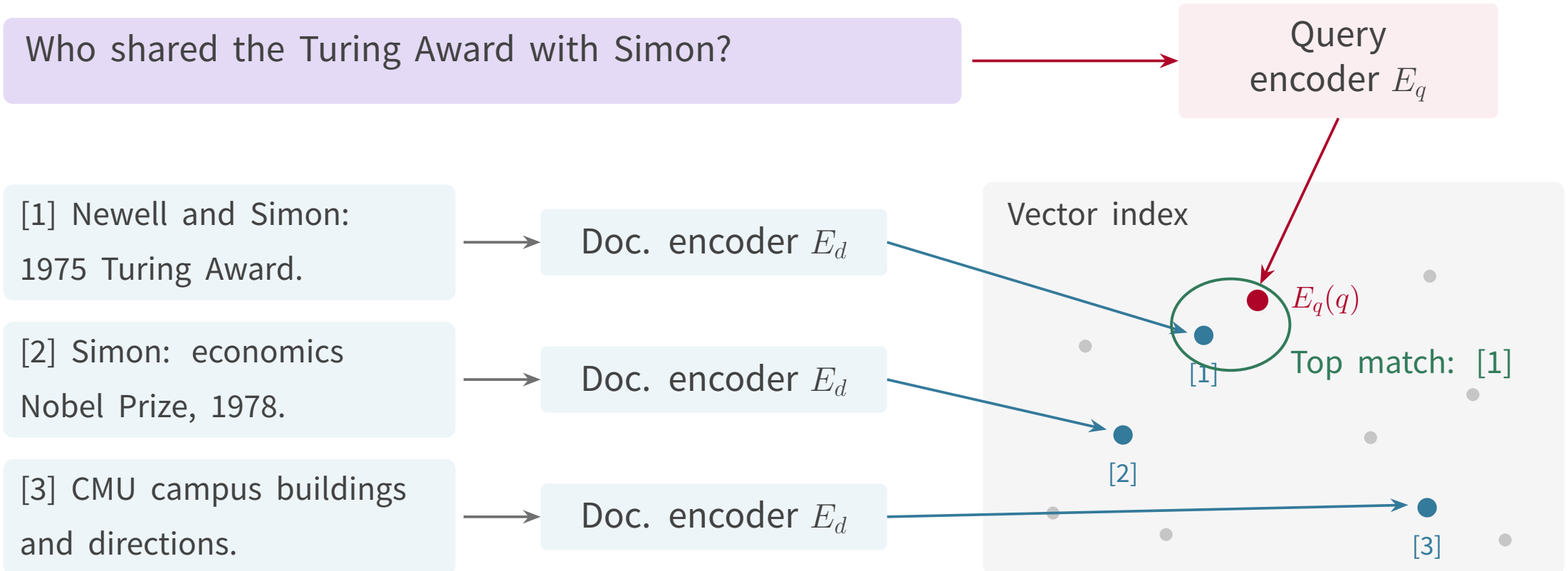
First, encode the document collection and build a vector index.



Score each document: $s(q, d) = E_q(q)^\top E_d(d)$; retrieve the top k.

Dual-encoder retrieval: Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020..

Embedding retrieval finds nearby documents

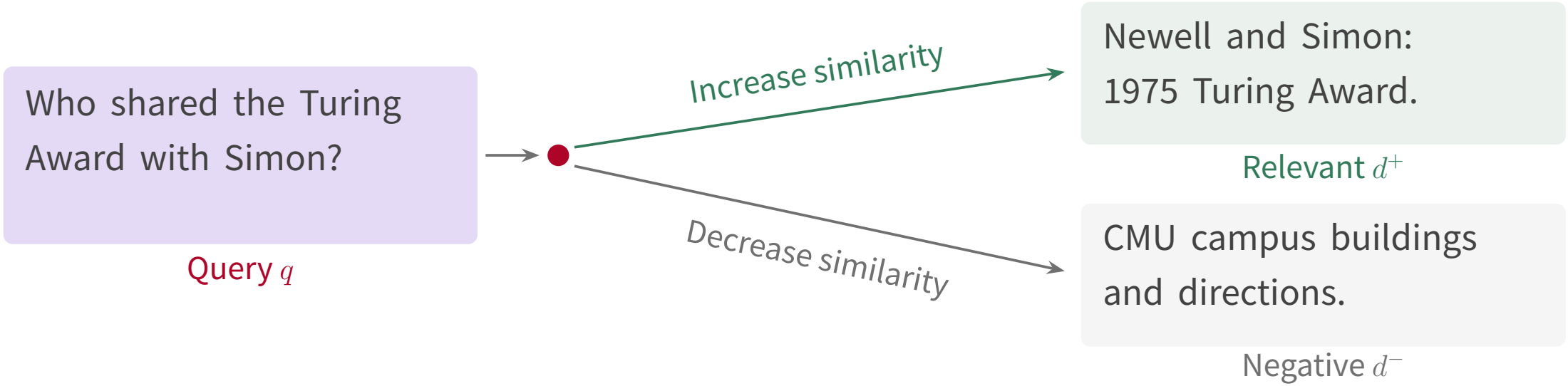


Score each document: $s(q, d) = E_q(q)^\top E_d(d)$; retrieve the top k.

Dual-encoder retrieval: Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020..

Train embeddings to rank useful evidence higher

Contrastive learning compares a relevant passage with negative passages.

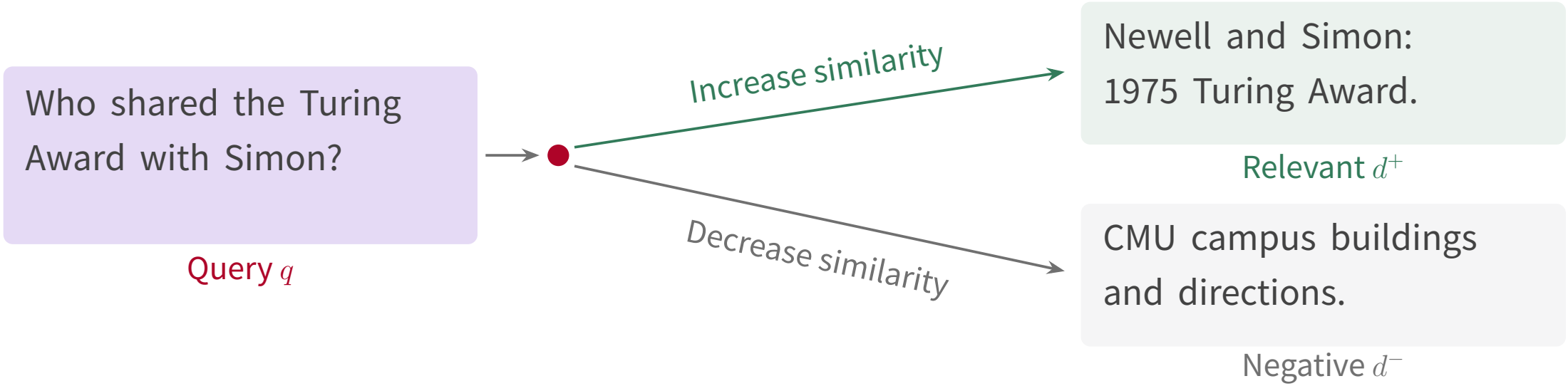


For encoder design, negative sampling, and indexing, see Advanced NLP and IR courses.

Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020..

Train embeddings to rank useful evidence higher

Contrastive learning compares a relevant passage with negative passages.

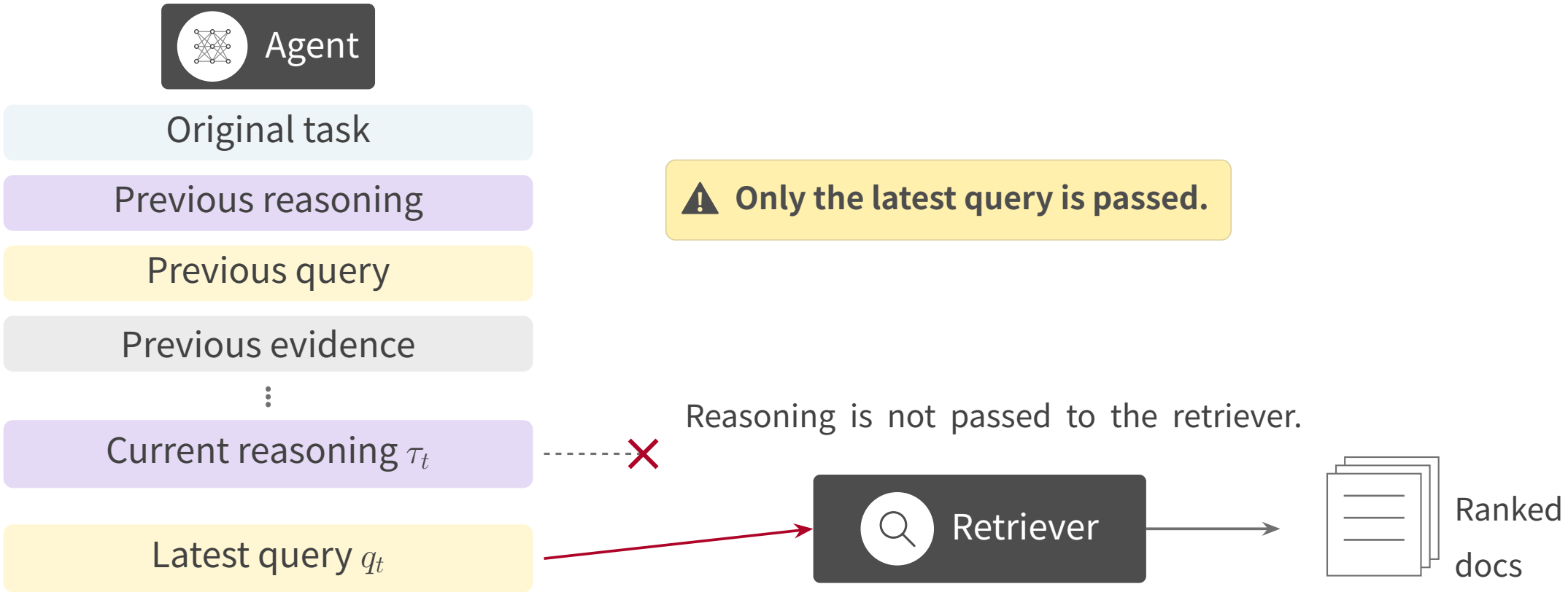


$$\mathcal{L} = -\log \frac{\exp s(q, d^+)}{\exp s(q, d^+) + \sum_{d^-} \exp s(q, d^-)}$$

For encoder design, negative sampling, and indexing, see Advanced NLP and IR courses.

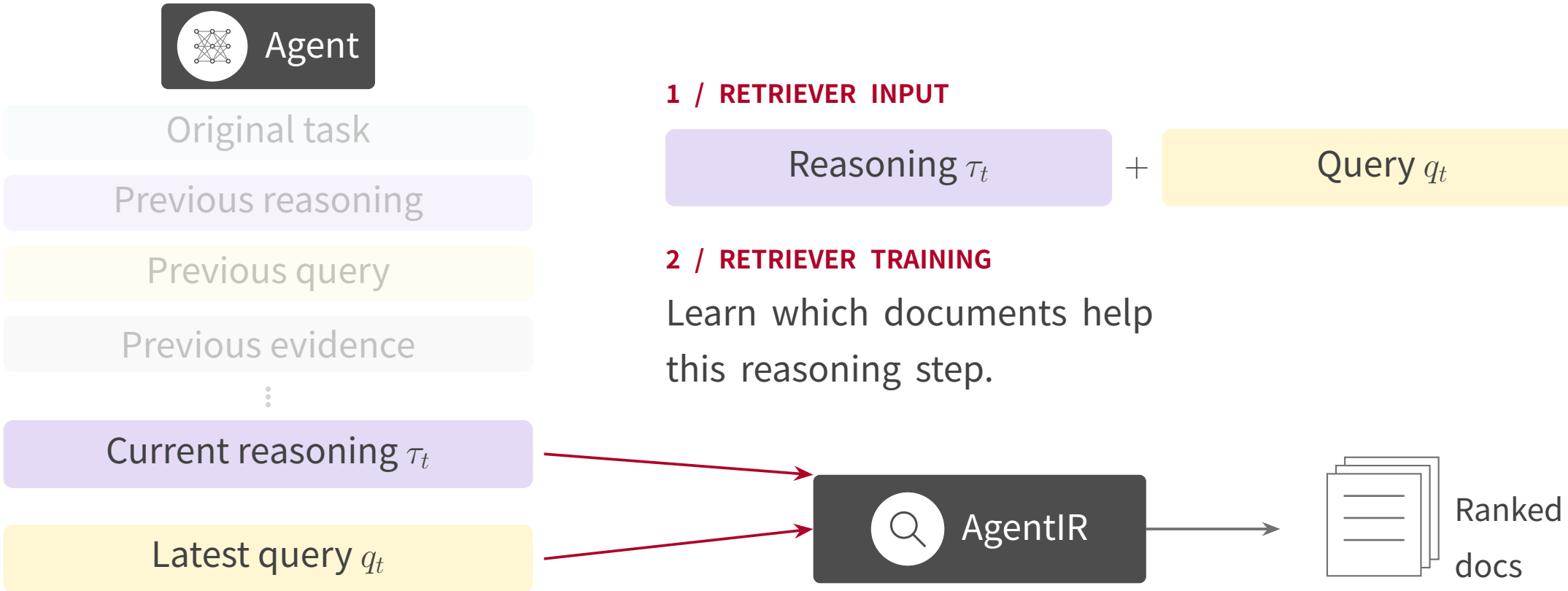
Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020..

The retriever often sees only the query



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

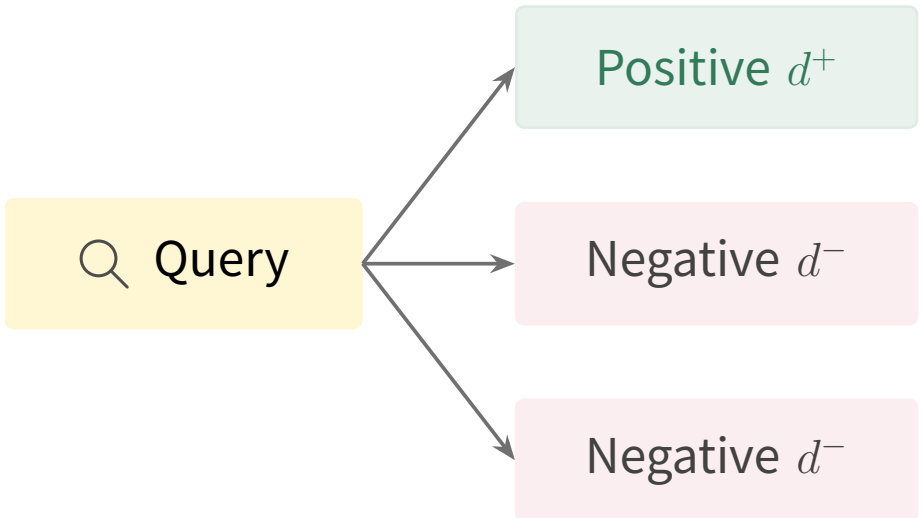
AgentIR: retrieval with reasoning and a query



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

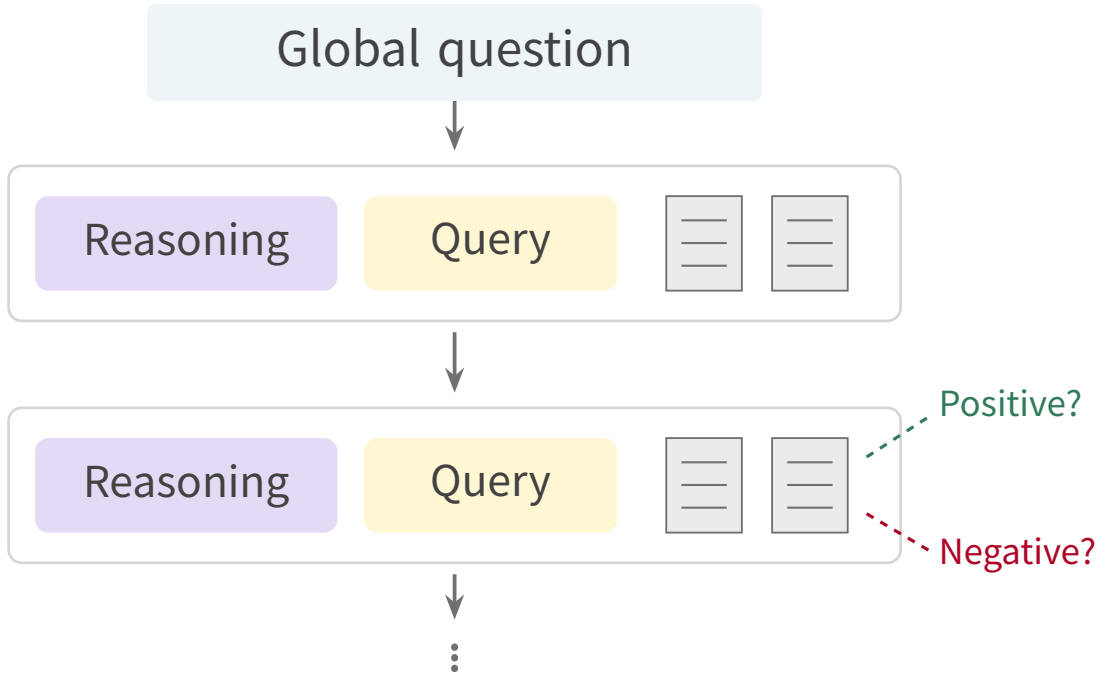
Intermediate searches need local labels

Standard retriever training



Known relevance labels

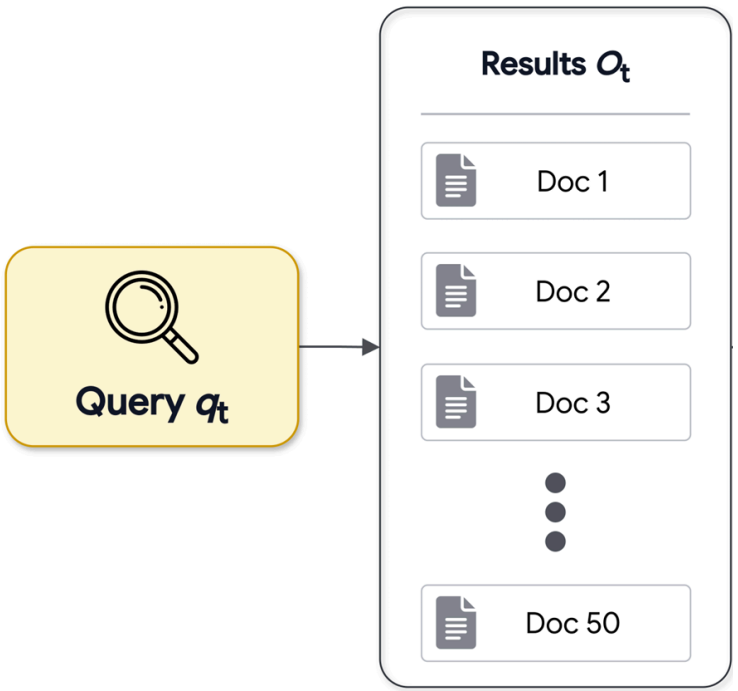
Deep research trajectory



Which documents help this turn?

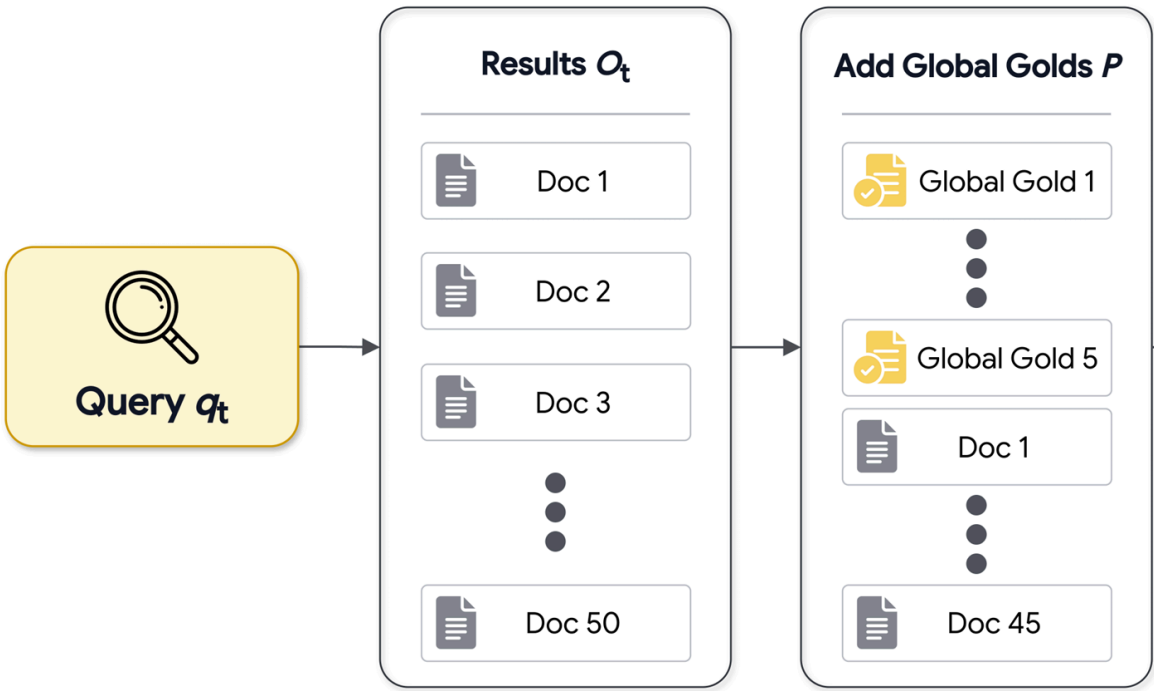
AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

DR-Synth: labels for intermediate searches



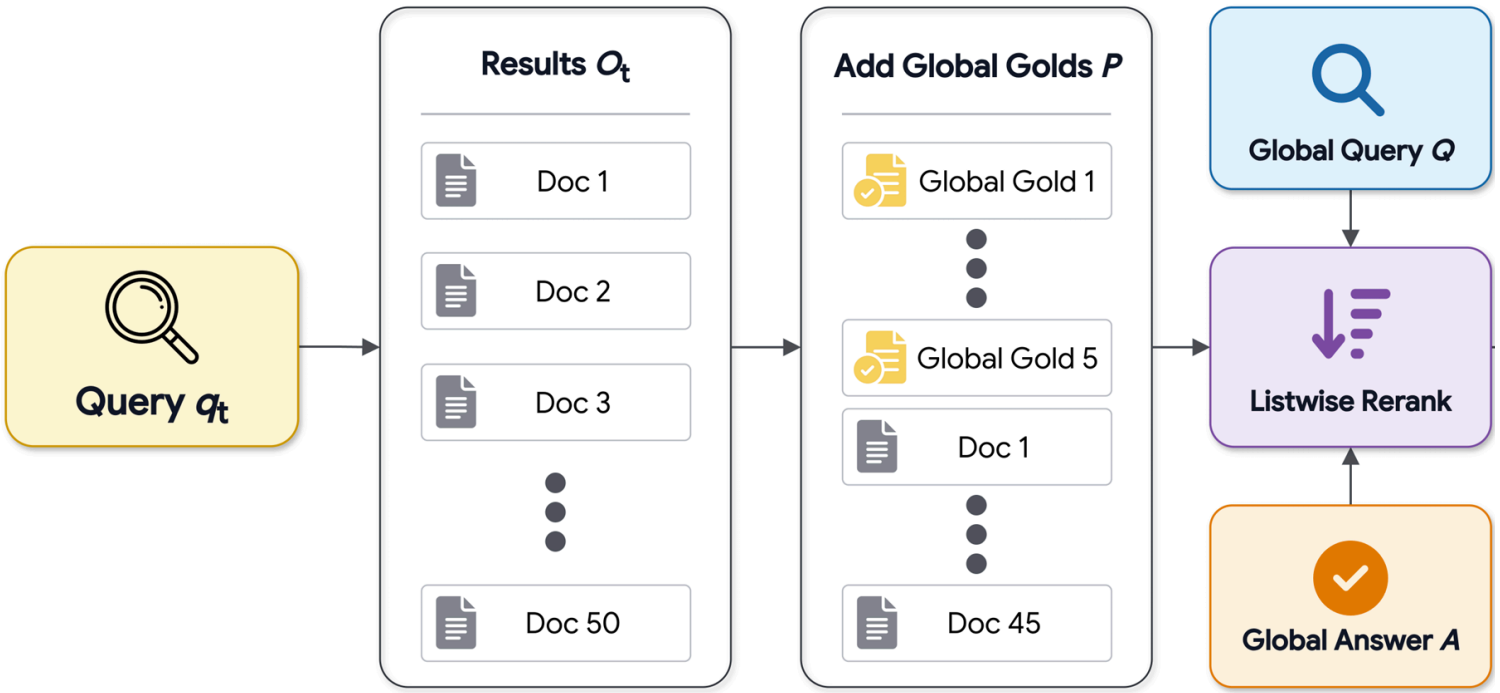
AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

DR-Synth: labels for intermediate searches



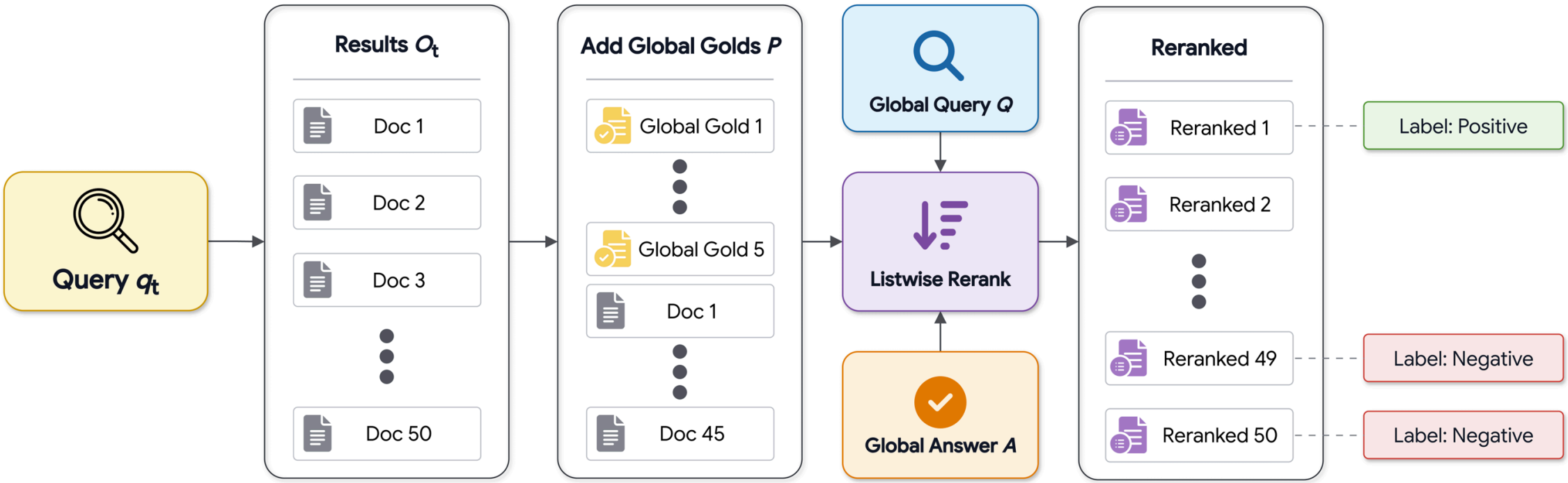
AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

DR-Synth: labels for intermediate searches



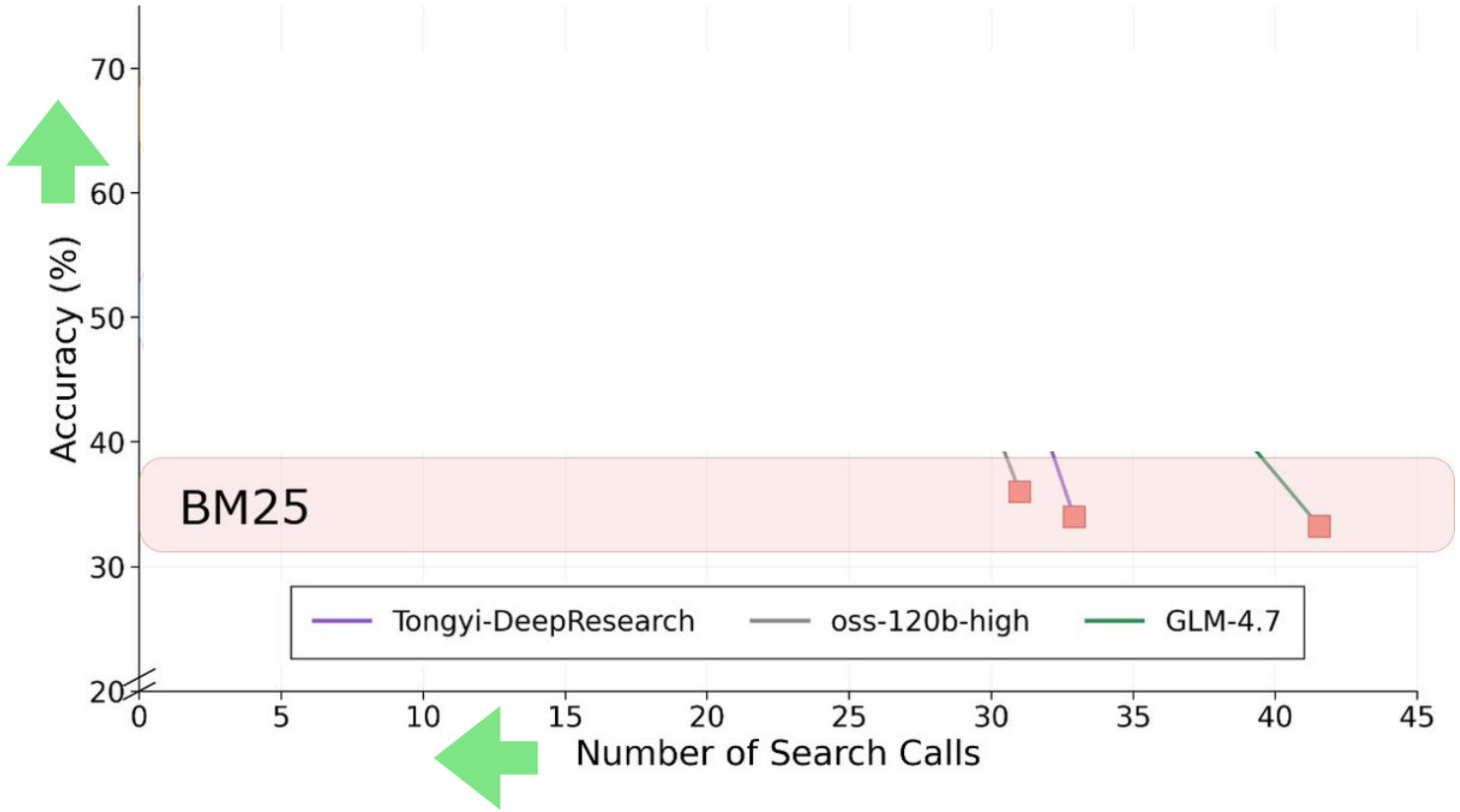
AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

DR-Synth: labels for intermediate searches



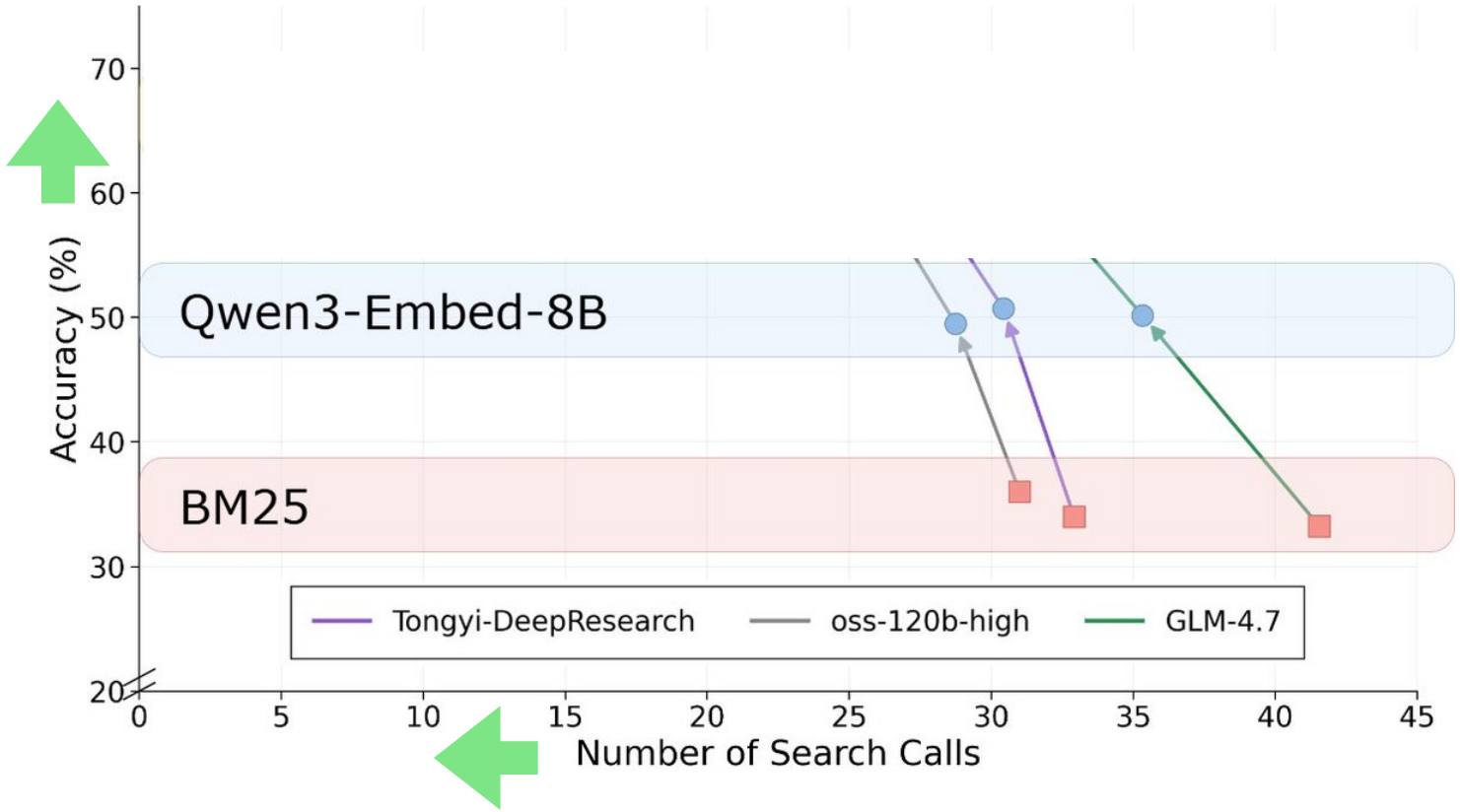
AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR: higher accuracy with fewer searches



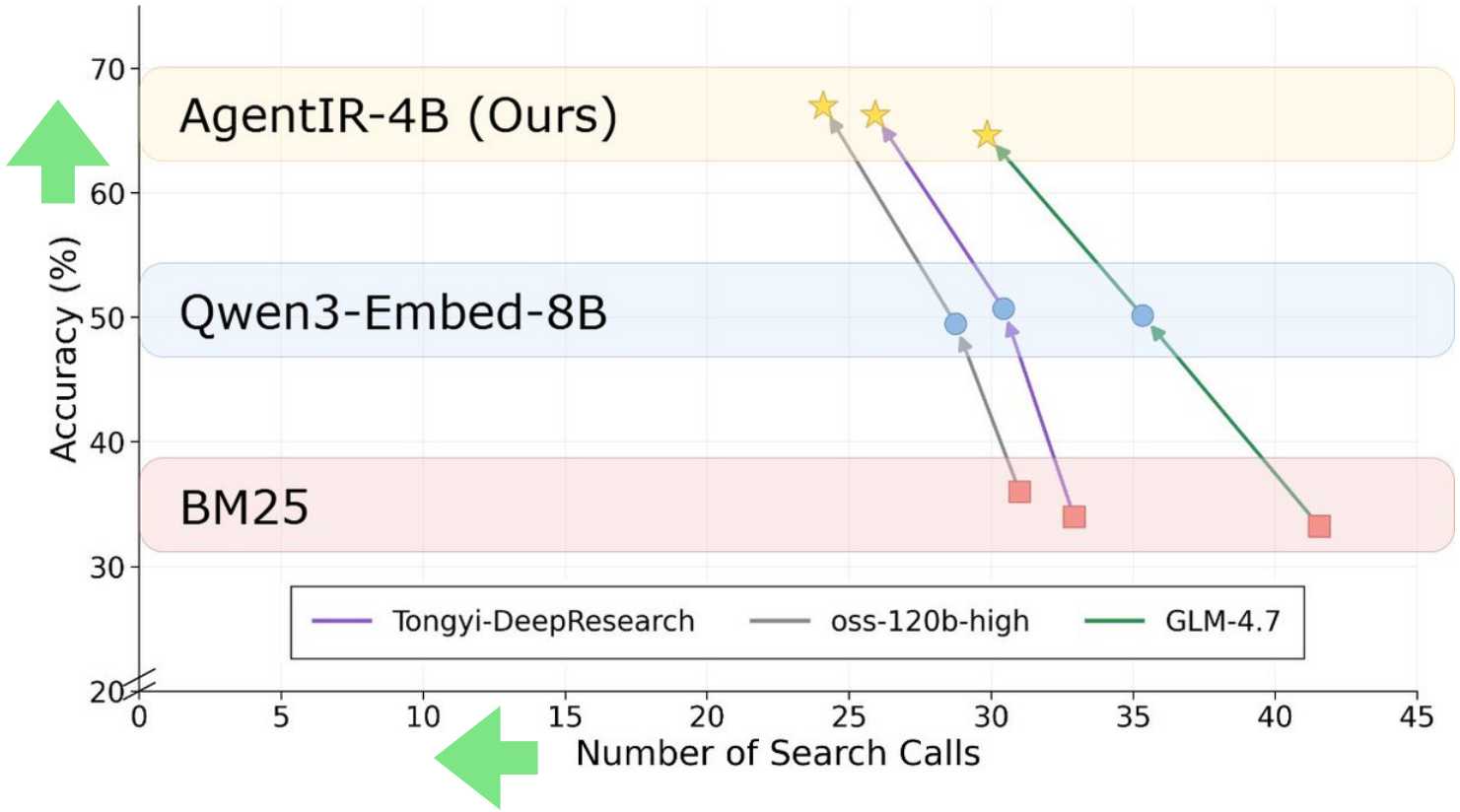
AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR: higher accuracy with fewer searches



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

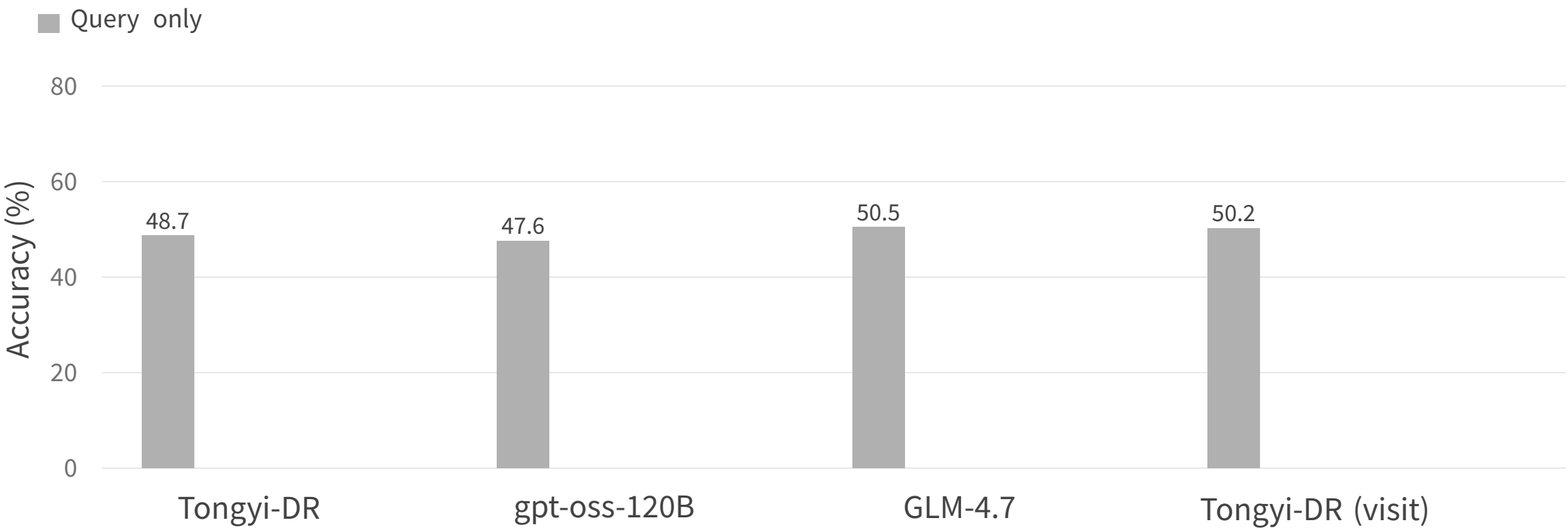
AgentIR: higher accuracy with fewer searches



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR ablation: inputs and training

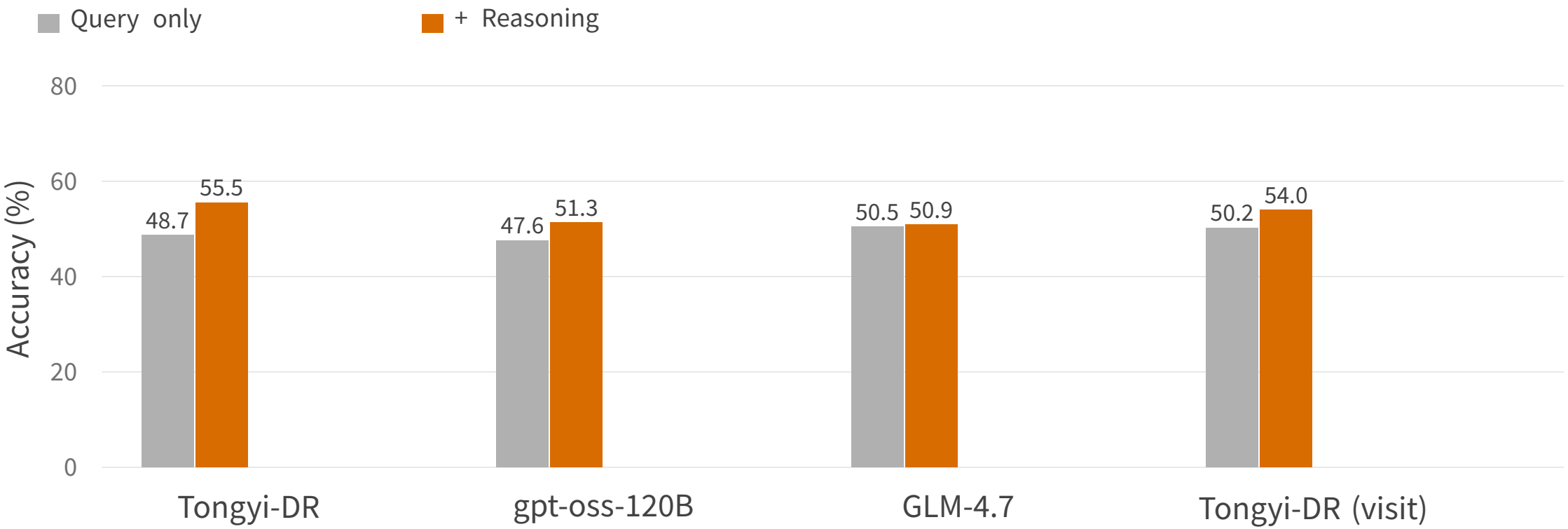
Which helps: reasoning-aware input, DR-Synth training, or both?



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR ablation: inputs and training

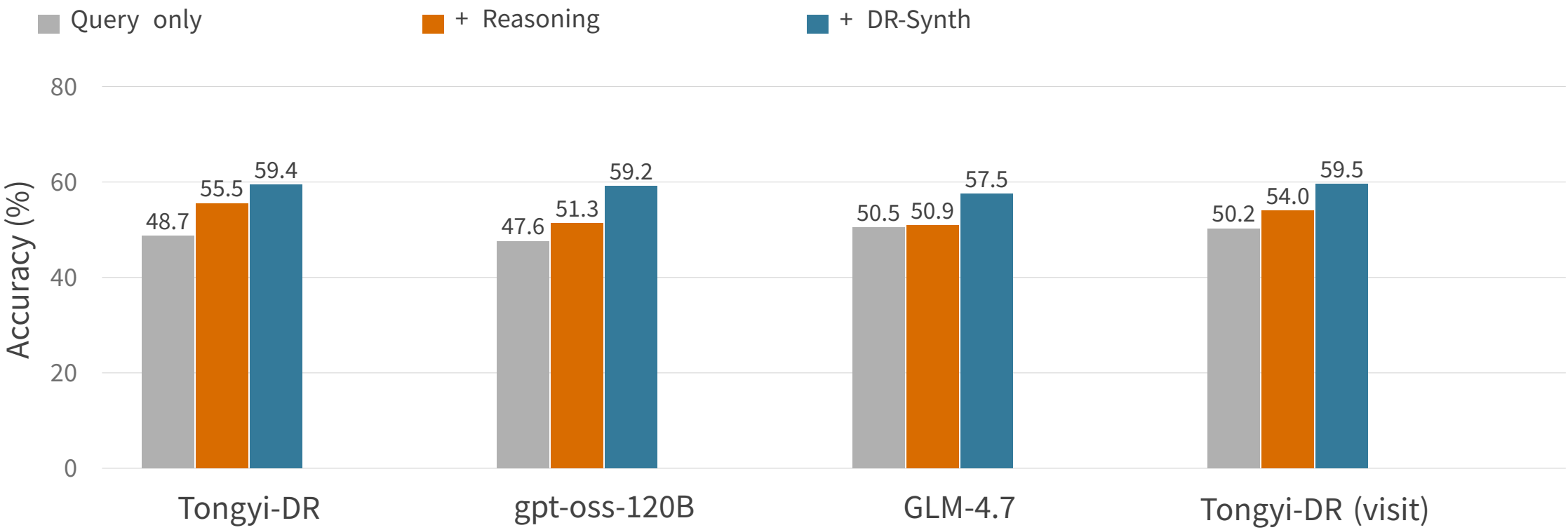
Which helps: reasoning-aware input, DR-Synth training, or both?



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR ablation: inputs and training

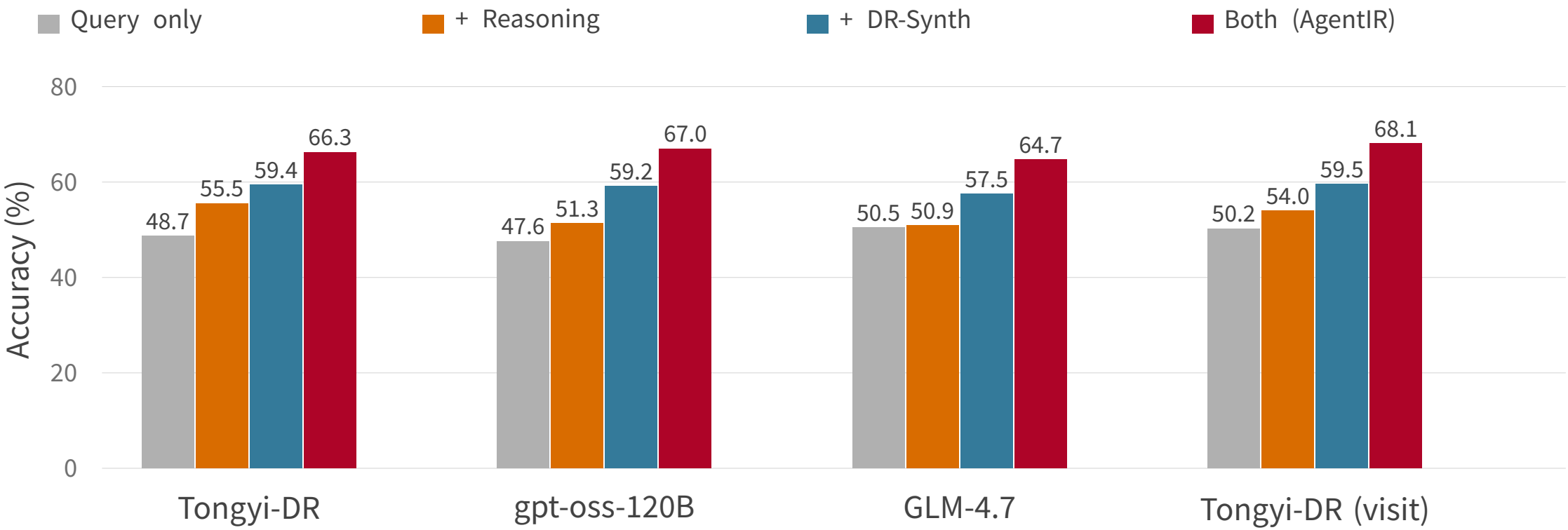
Which helps: reasoning-aware input, DR-Synth training, or both?



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR ablation: inputs and training

Which helps: reasoning-aware input, DR-Synth training, or both?

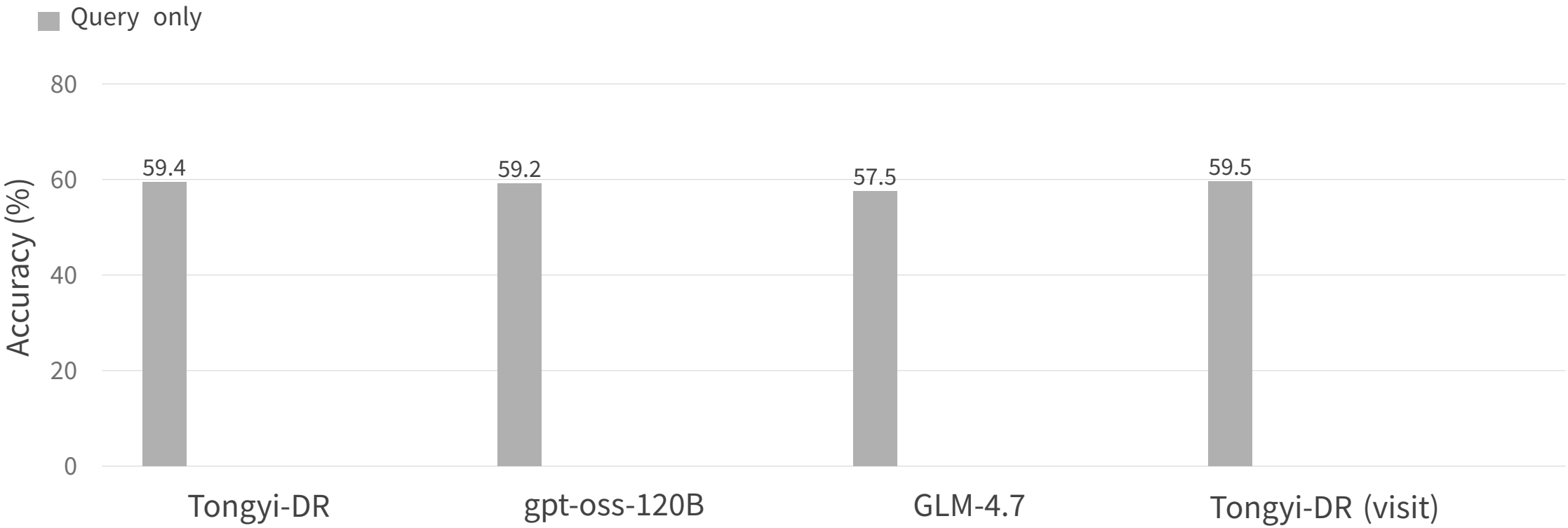


Both components help; combining them gives the highest accuracy.

AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR: which history helps retrieval?

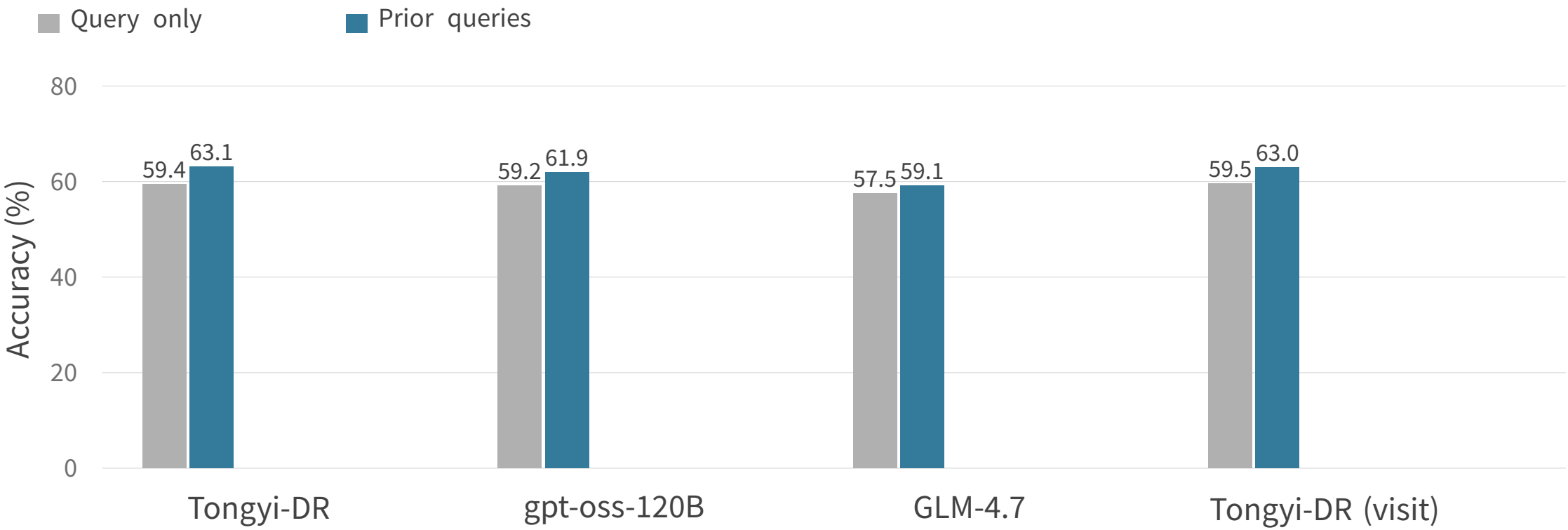
All retrievers use DR-Synth; only the input context changes.



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR: which history helps retrieval?

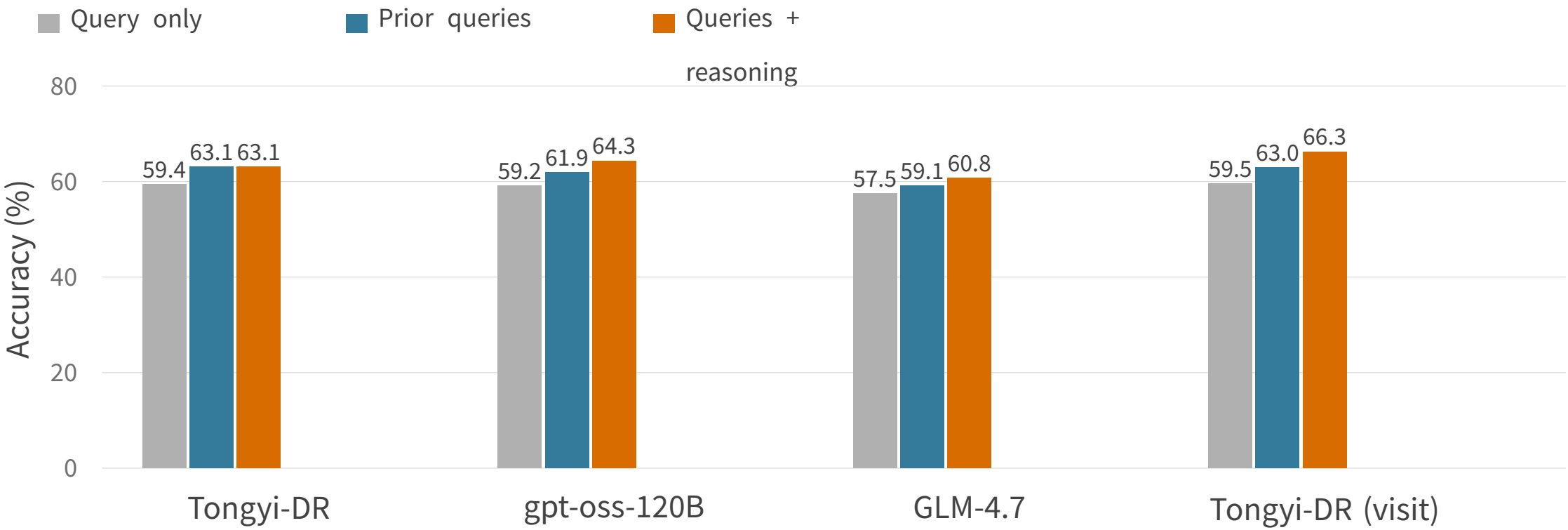
All retrievers use DR-Synth; only the input context changes.



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR: which history helps retrieval?

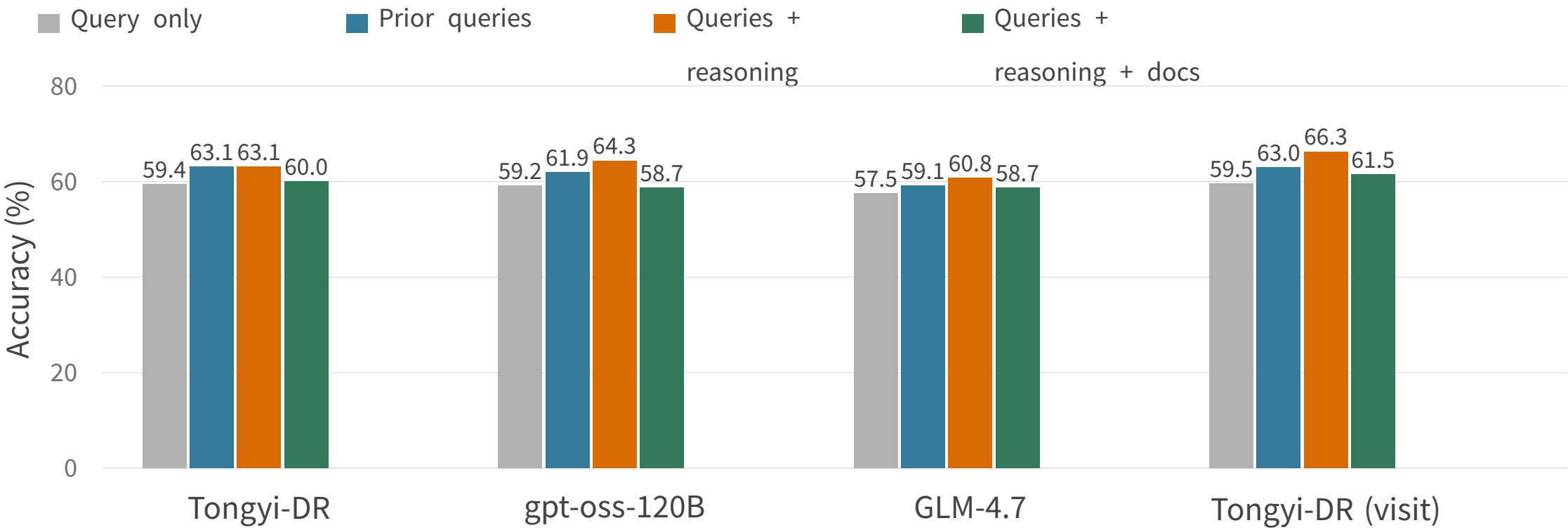
All retrievers use DR-Synth; only the input context changes.



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR: which history helps retrieval?

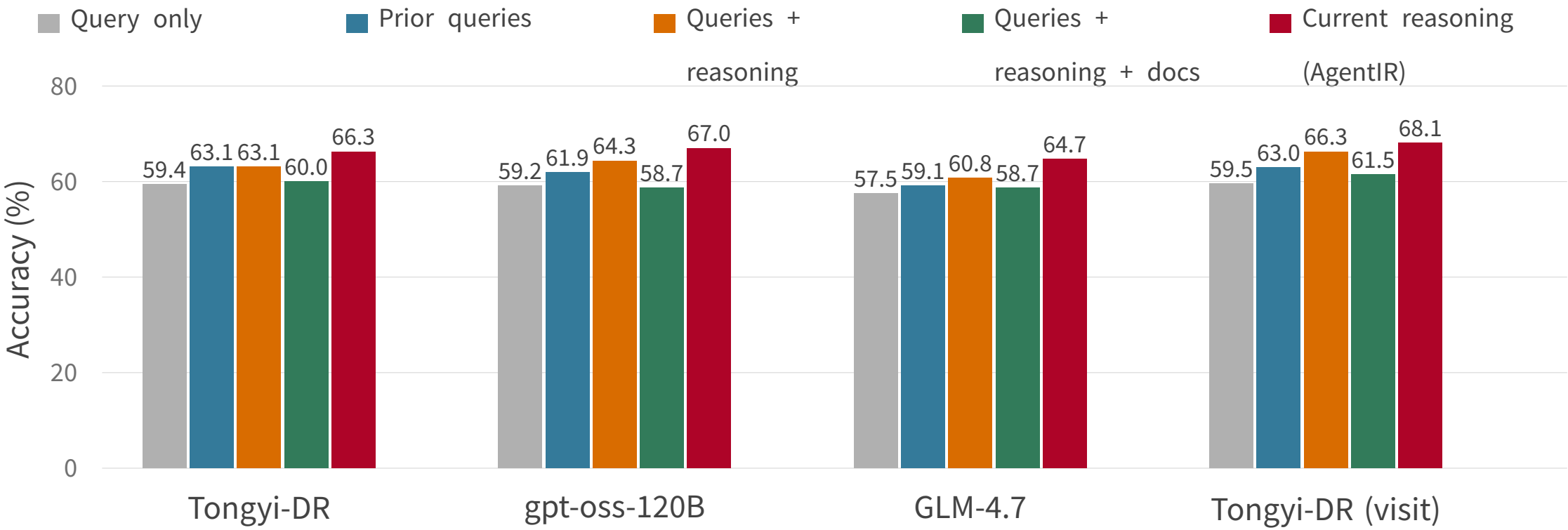
All retrievers use DR-Synth; only the input context changes.



AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

AgentIR: which history helps retrieval?

All retrievers use DR-Synth; only the input context changes.



Current reasoning gives the strongest signal across these agent settings.

AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. COLM 2026.

Summary

01 **Tasks**

Deep research combines many searches with evidence synthesis.

02 **Evaluation**

Assess answer quality and citation support; audit the rubrics too.

03 **Training**

Learn from demonstrations, then improve through task feedback.

04 **Retrieval**

Give retrievers the agent’s information need and train for it.