

Deep Research Agents: Evaluation and Modeling Development

Akari Asai

Research Scientist @ Ai2 OLMo & Incoming Assistant Professor @ CMU

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

Deep Research (DR) Agents

What are the differences between OLMo2 and OLMo3?

Differences Between OLMo2 and OLMo3 Language Models

🌀 I'm thinking through whether OLMo2 and OLMo3 are open language models from the Allen Institute, similar to GPT-like models, and noting the need to pinpoint their exact nature and versions.

🔍 Searched for OLMo2 and OLMo3 models

🔗 Read allennai.org

🌀 OK, let's see: Olmo 2's specifics and its differences from Olmo 3 are tricky. I'll review the technical report for more details.

🔗 Read arxiv.org

🌀 I'm looking at OLMo3's state-of-the-art models, their expansive capabilities, lifecycle

OLMo2 vs OLMo3: Key Differences and Comparison

Architecture and Model Size

- **Model Scale:** OLMo2 debuted with 7B and 13B parameter models (later expanded to a 32B variant) infoq.com huggingface.co . OLMo3 offers 7B and 32B models as its primary releases, making a significant jump in size at the high end (32B vs. OLMo2's 13B) infoq.com arxiv.org . This larger 32B "Think" model is the new flagship, providing substantially more capacity for complex tasks.
- **Transformer Architecture:** Both OLMos use a decoder-only transformer architecture. OLMo2 introduced several architectural tweaks to improve stability and performance, such as using **RMSNorm** (instead of standard layer norm), applying **QK-normalization**, reordering layer norms, and switching from absolute to **rotary positional embeddings** allennai.org infoq.com . It also added techniques like **Z-loss regularization** to stabilize training allennai.org . OLMo3 retains these improvements and further enhances the model – for example, it's explicitly designed for **long-context support** (trained to handle inputs up to ~65k tokens, far beyond OLMo2's 4k context window) huggingface.co venturebeat.com . OLMo3-"Think"

What Makes “Deep Research” Hard



Evaluation

- How can we test complex search / reasoning capabilities?
- How can we reliably evaluate open-ended deep research?



Modeling

- How can we train models for complex deep research tasks?
- How can we optimize systems for “hard-to-verify” tasks?

Evaluation

- Complex *verifiable* short form questions
- Rubric-based evaluations for *unverifiable* long-form tasks
- Human evaluations using static and live queries

Simple Factoid Questions (Open-domain QA)



where is the world s largest ice sheet located today?

Antarctica

Not logged in | Talk | Contributions | Create account | Log in

Article | Talk | Read | Edit | View history | Search Wikipedia

Ice sheet

From Wikipedia, the free encyclopedia

An **ice sheet** is a mass of **glacier** ice that covers surrounding terrain and is greater than 50,000 km² (19,000 sq mi).^[1] This is also known as **continental glacier**.^[2] The only current ice sheets are in Antarctica and Greenland; during the last glacial period at Last Glacial Maximum (LGM) the Laurentide ice sheet covered much of North America, the Weichselian ice sheet covered northern Europe and the Patagonian Ice Sheet covered southern South America.

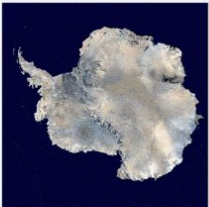
Ice sheets are bigger than ice shelves or alpine glaciers. Masses of ice covering less than 50,000 km² are termed an ice cap. An ice cap will typically feed a series of glaciers around its periphery.

Although the surface is cold, the base of an ice sheet is generally warmer due to **geothermal** heat. In places, melting occurs and the melt-water lubricates the ice sheet so that it flows more rapidly. This process produces fast-flowing channels in the ice sheet — these are **ice streams**.

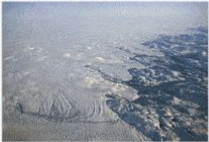
The present-day polar ice sheets are relatively young in geological terms. The Antarctic Ice Sheet first formed as a small ice cap (maybe several) in the early **Oligocene**, but retreating and advancing many times until the **Pliocene**, when it came to occupy almost all of Antarctica. The Greenland ice sheet did not develop at all until the late Pliocene, but apparently developed very rapidly with the first continental **glaciation**. This had the unusual effect of allowing **fossils** of **plants** that once grew on present-day Greenland to be much better preserved than with the slowly forming Antarctic ice sheet.

Contents [hide]

- Antarctic ice sheet
- Greenland ice sheet
- Ice sheet dynamics
- Predicted effects of global warming
- References
- Further reading
- External links



A satellite composite image of Antarctica



Aerial view of the ice sheet on Greenland's east

Natural Questions: A Benchmark for Question Answering Research
Kwiatkowski et al. TACL 2019

Easy-to-verify but Hard-to-solve Questions



Please identify the fictional character who occasionally breaks the fourth wall with the audience, has a backstory involving help from selfless ascetics, is known for his humor, and had a TV show that aired between the 1960s and 1980s with fewer than 50 episodes.

Human annotated questions that

- ✓ GPT4o can't solve
- ✓ Can't be answered with 5 google search
- ✓ Another human can't solve within 10 mins

Plastic Man

BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents

Wei, et al. 2025

Realistic *Unverifiable* Deep Research Questions



What's the state-of-the-art in immunotherapy for cervical cancer?

... Pembrolizumab can extend progression-free and overall survival in patients with persistent, recurrent, or metastatic PD-L1-positive cervical cancer [1] and locally advanced cervical cancer [2]...

[1] Colombo et al., KEYNOTE-826, 2021

[2] Lorusso et al., KEYNOTE-A18, 2024

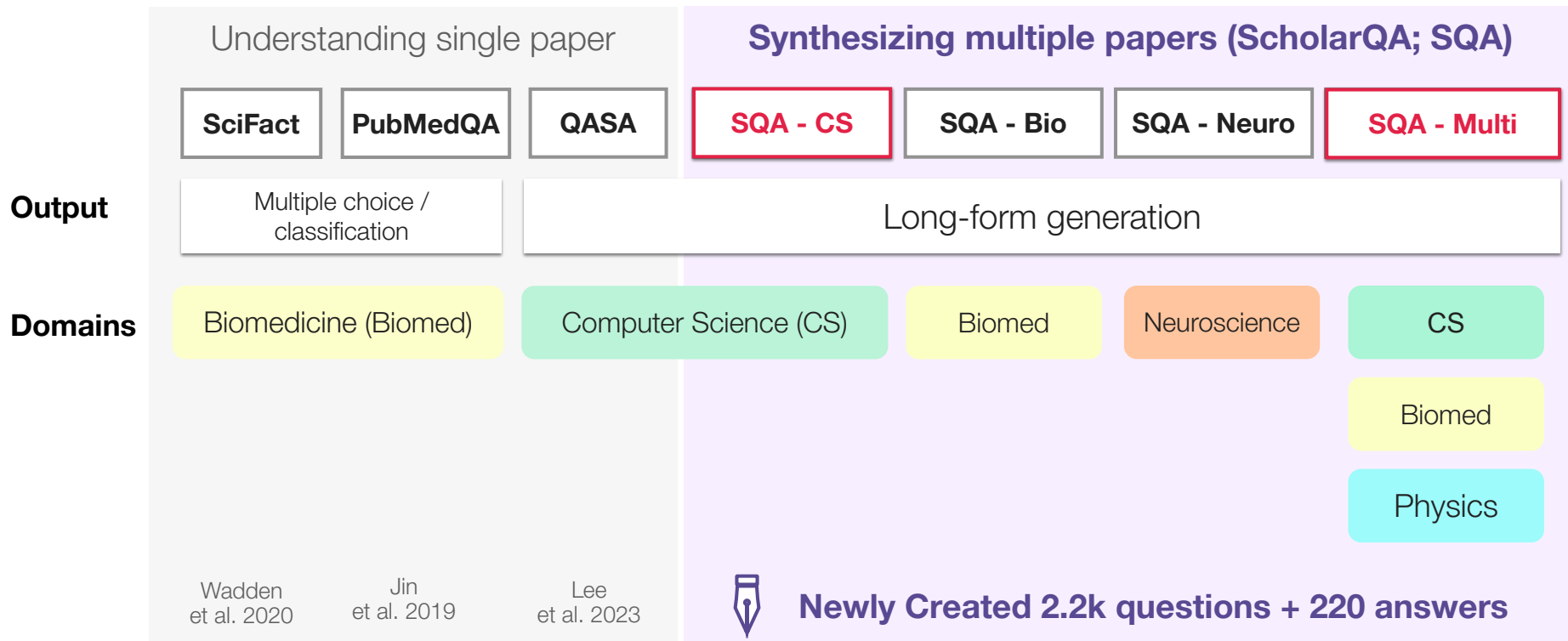
Attribution

Credible
sources

Up-to-date

How can we evaluate deep research agents for more realistic, open-ended tasks?

ScholarQABench: Expert-annotated DR Benchmark



Synthesizing scientific literature with retrieval-augmented language models

Asai et al. Nature 2026. 9

ScholarQA-CS: Introducing Rubric-based Evaluation

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Answer rubrics:

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

Nice to Have Item-1: The answer could provide some examples of famous third party libraries that can be used in different programming languages.

ScholarQA-CS: Introducing Rubric-based Evaluation

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Answer rubrics:

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

Output: Protecting software against vulnerabilities stemming from third-party libraries is a crucial aspect of software security.

Below are some of the best practices based on the existing literature: (1) Developing intelligent tools to automatically and repair vulnerabilities ...



Rubric-based Evaluation Benchmarks

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 (Yifei et al., 2025)

Candidate Response

While everyone has days that they feel less energetic, for an infant this can be a sign of an illness or other medical condition. Your infant's change in behavior could be a sign of muscle weakness.

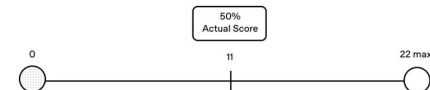
It sounds like your baby needs an evaluation with a medical professional.

Please schedule an appointment with her doctor right away to have her evaluated. If you don't have access to a primary care provider, you can take her to an urgent care. If she has any signs of distress she needs immediate medical attention.

Assistant

Rubric Criteria & Grades

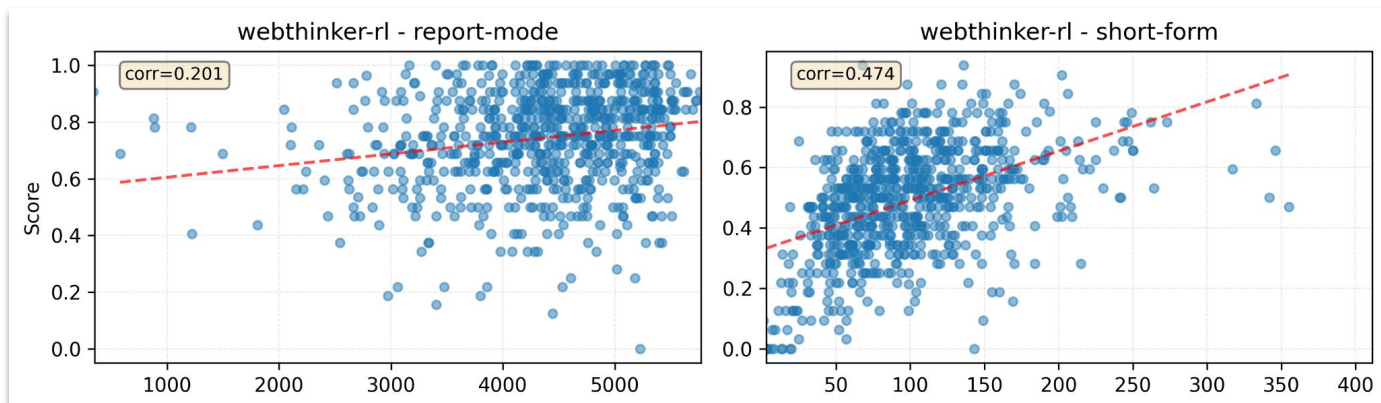
Criterion	Points	Present
States that the infant may have muscle weakness	+7	Yes 7
Briefly describes common causes of muscle weakness in infants	+5	No 0
Includes advice to seek medical care right away	+10	Yes 10
Referral advice is only buried in the middle or at the end of response	-6	Yes -6



HealthBench (Arora et al., 2025)

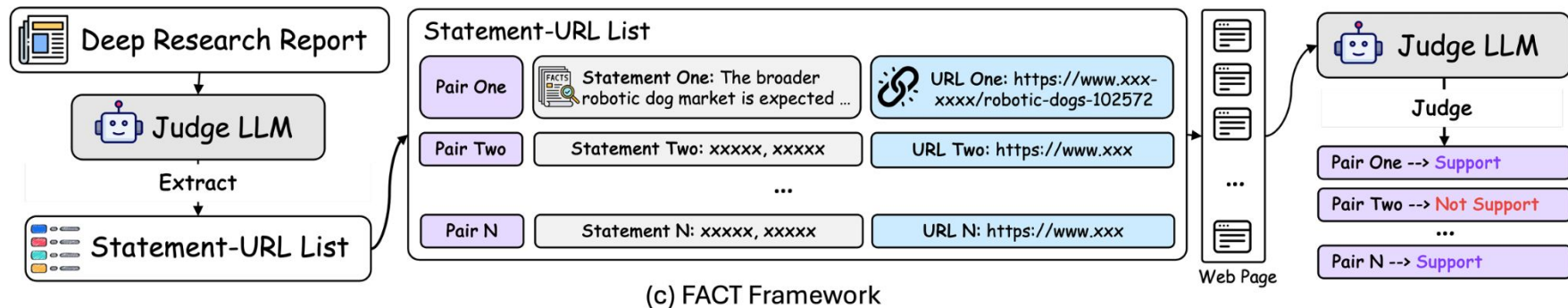
Issues with Rubric-based Evaluation Framework

- ❌ **Rubric** (recall) evaluations are brittle to length biases
- ❌ **Biases** and **limited coverages** of annotated rubrics



Evaluating Citation Accuracy

1. Extract statement-URL pairs
2. Check if each website (URL) supports the statement
3. Compute **# of supported statement / total # of statements**



DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents

Du, et al. ICLR 2026

Static Human Evaluation

How do people train a long-context Transformer model over 32K context length using sequence parallelism techniques? How are these methods compared with each other?

Here are responses from two AI models.

Answer

Sequence parallelism [Dacheng Li et al. 2023][Hao Liu et al. 2023][Shenggui Li et al. 2021] is an effective technique that could be applied to support long-context Transformer training without changing the model architectures by splitting and distributing the computation of a long input to multiple accelerators to reduce the computation and memory burden.

Ring Self-Attention [Shenggui Li et al. 2021] first proposes the concept of sequence parallelism and designed a ring topology to conduct joint computation of attention on different workers.

DistFlashAttention [Dacheng Li et al. 2023] extends FlashAttention with sequence parallelism, which effectively reduces the memory usage thanks to the FlashAttention design and accelerate the training with 3 key techniques: token-level workload balancing, overlapping key-value communication, and a rematerialization-aware gradient checkpointing algorithm.

RingAttention [Hao Liu et al. 2023] leverages blockwise computation of self-attention to distribute long sequences across multiple devices.

Striped Attention [William Brandon et al. 2023] improves the Ring Attention by introducing a strip mechanism that balances the workload on different workers.

DeepSpeed-Ulysses identifies the memory communication inefficiency of the existing sequence parallelism algorithm and proposes to employ an efficient all-to-all collective communication for attention computation in sequence parallelism.

Ring Self-Attention achieves 27 longer context length than the existing sparse attention works [Shenggui Li et al. 2021].

References

[Dacheng Li et al. 2023] FlashAttention (Dao, 2023) effectively reduces the quadratic peak memory usage to linear in training ... (read more details)

[Hao Liu et al. 2023] We present a novel approach, Ring Attention with Blockwise Transformers (Ring Attention), which leve ... (read more details)

[William Brandon et al. 2023] In this paper, we study the performance characteristics of Ring Attention in the important special c ... (read more details)

[Shenggui Li et al. 2021] In this work, we propose sequence parallelism, a memory-efficient parallelism method to help us brea ... (read more details)

[S. A. Jacobs et al. 2023] Computation in a typical Transformer-based large language model (LLM) can ... (read more details)

Q2(a): **Organization:** Evaluate if the output B is coherent and well-structured with clear logical flow and organization. Read the score rubrics carefully and give score from 1 to 5. [Show rubrics](#)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2(b): **Coverage and Amount of Information:** Evaluate if the output B provides sufficient coverage and amount of information. Read the score rubrics carefully and give score from 1 to 5. [Show rubrics](#)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2(c): **Relevance:** Evaluate if the output B stays on topic and maintain a clear focus to provide a useful response to the question. Read the score rubrics carefully and give score from 1 to 5. [Show rubrics](#)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2(d): **Overall Usefulness:** Do you think the provided answer B is overall helpful and assist your literature review? Read the score rubrics carefully and give score from 1 to 5. [Show rubrics](#)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Finally, tell us which answer you liked more!

Q3: Please choose the response that you prefer (based on helpfulness) and give explanation on why you think one is better than the other.

☐ A is better ☐ Tie ☐ B is better

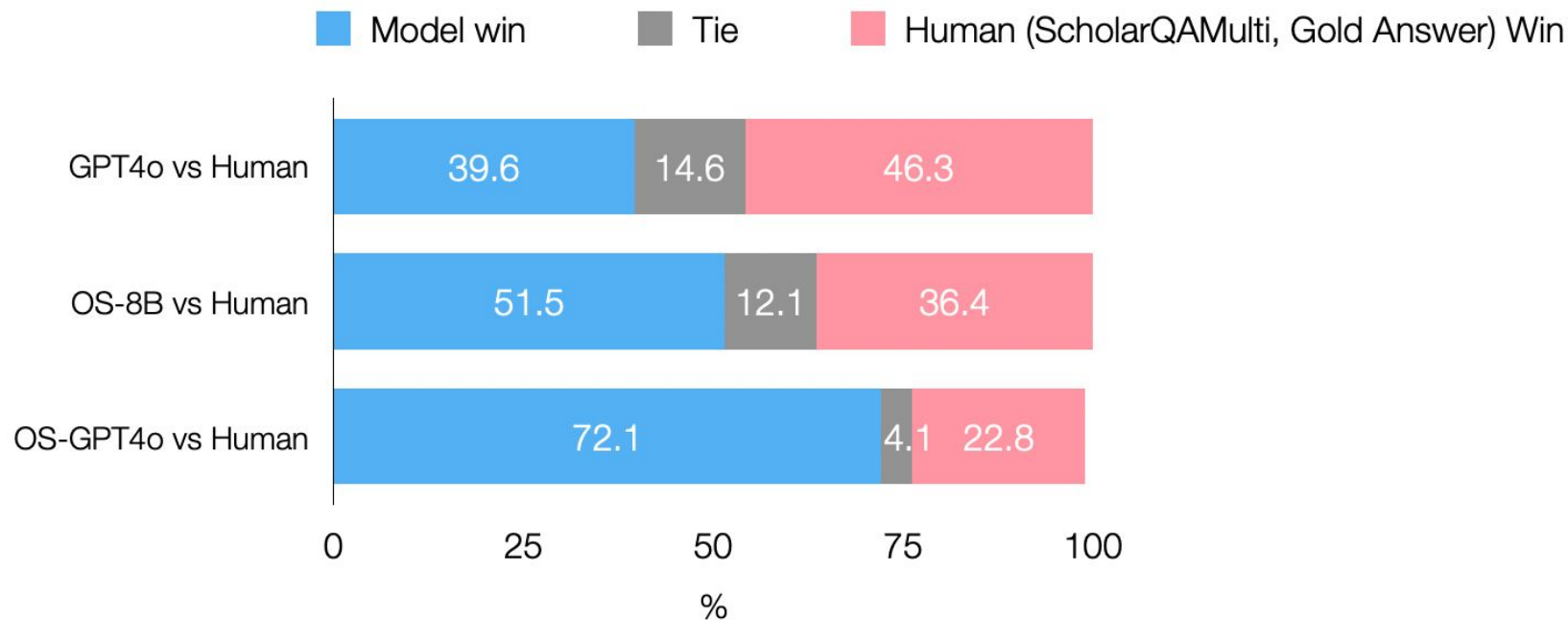
Explanation

Please enter your explanation for why you think the model you chose is better than the other.

Synthesizing scientific literature with retrieval-augmented language models

Asai et al. Nature 2026.

Static Human Evaluation

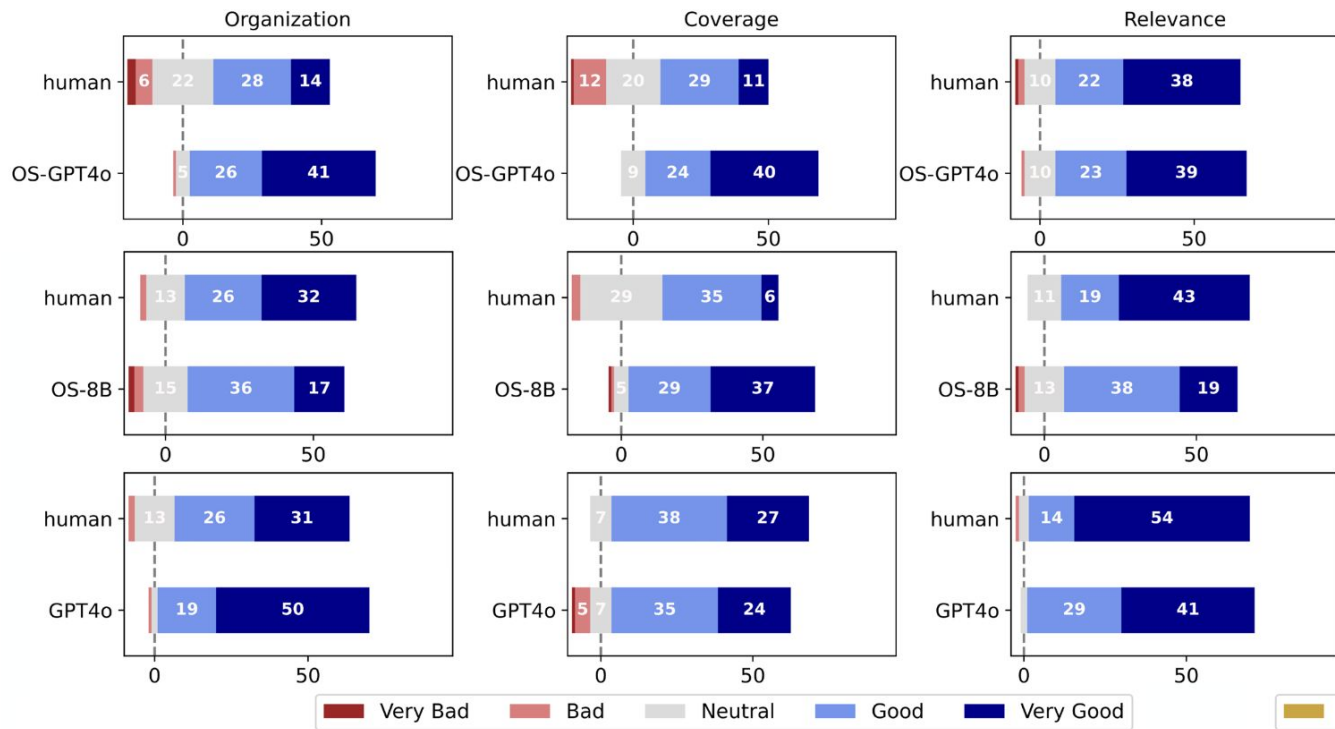


Preference distribution

Synthesizing scientific literature with retrieval-augmented language models

Asai et al. Nature 2026. 16

Static Human Evaluation



Synthesizing scientific literature with retrieval-augmented language models

Live Human Evaluation



Deep Research Comparator: A Platform For Fine-grained Human Annotations of Deep Research Agents

Modeling

- RLVR for short-form deep research
- RLER and new approaches for open-ended deep research
- Context management

More and More Deep Research Papers ...

Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning

Bowen Jin¹, Hansi Zeng², Zhenrui Yue¹, Jinsung Yoon³, Sercan Ö. Arık³, Dong Wang¹,

Hamed Zamani², Jiawei

¹ Department of Computer Science

² Center for Intelligent Systems

³ Google Cloud AI Research

{bowenj4, zhenrui3, dwang}

{jinsungyoon, soarik}@

WebThinker: Empowering Large Reasoning Models with Deep Research Capability

WEBEXPLORER: Explore and Evolve for Training Long-Horizon Web Agents

Junteng
Chengyu

Tongyi DeepResearch Technical Report

Tongyi DeepResearch Team*

Tongyi Lab  , Alibaba Group

Beyond Ten Turns: Unleashing Search with Large

Jiaxuan Gao¹, Wei Fu^{1,2}, Minyang Xie¹, Shusheng Xu²,
Chuyi He², Zhiyu Mei², Banghua Zhu³, Yi Wu^{1*}

¹ IIIS, Tsinghua University, ² Ant Group

³ University of Washington

samjia2000@gmail.com, jxwuyi@gmail.com

Effective Reinforcement Learning for Reasoning Single Agents

with Gangi Reddy[†] Austin Xu

Silvio Savarese Caiming Xiong* Shafiq Joty*

Salesforce AI Research

Many DR Models Focus on Short-form QA

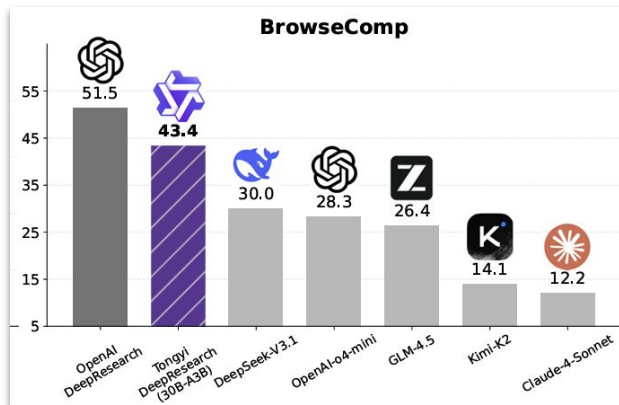


Please identify the fictional character who occasionally breaks the fourth wall with the audience, has a backstory involving help from selfless ascetics, is known for his humor, and had a TV show that aired between the 1960s and 1980s with fewer than 50 episodes.

Plastic Man

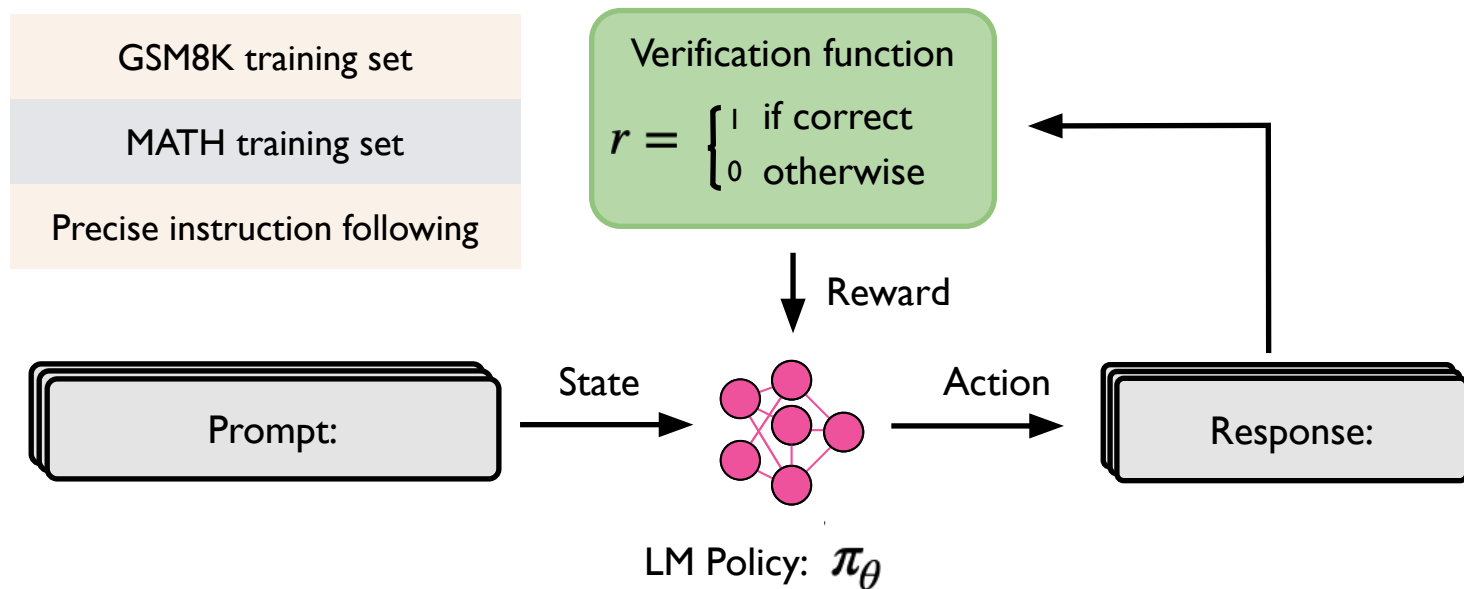
BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents

Jason Wei, et al. 2025



Tongyi Deep Research Technical Report
Tongyi Deep Research Team. 2025

RL with Verifiable Rewards (RLVR)



Tulu 3: Pushing Frontiers in Open Language Model Post-Training
Nathan Lambert, et al. 2024

Search-R1



What type of profession does Chris Jericho and Gary Barlow have in common?

```
<think> I need to find the type of profession that Chris Jericho and Gary Barlow have in common.
I'll search for it. </think>
<search> Chris Jericho and Gary Barlow profession </search>
<information> Doc 2(Title: "Chris Jericho") ""Do You Wanna Start a War"" (2014) and ""Judas""
(2017) are comprised entirely or predominantly of original compositions. He has also appeared on
numerous TV shows over the years, having competed in the 2011 season of ""Dancing With the
Stars"", lasting until the sixth week. Prior to that, Jericho hosted the ABC game show ""Downfall""
and the Revolver Golden Gods Awards. Christopher Keith Irvine was born on November 9, 1970,
in Manhasset, New York, but grew up in Winnipeg. His father, ice hockey player Ted Irvine, had
been playing for the New York Rangers at the time of his birth. ... </information>
<think> I found out that Chris Jericho is a professional wrestler, musician, author, and actor. Now
I need to find if Gary Barlow has any of these professions in common. </think>
<search> Gary Barlow profession </search>
```

...

```
<think> I found out that Gary Barlow is a singer, songwriter, musician, and record producer. Now
I need to find if Chris Jericho has any of these professions in common. </think>
<search> Chris Jericho profession </search>
<information> Doc 1(Title: "Chris Jericho") Chris Jericho Christopher Keith Irvine (born November
9, 1970), better known by the ring name Chris Jericho, is an American-born professional wrestler,
musician, author and actor. He currently wrestles for New Japan Pro-Wrestling (NJPW) where
he is the current IWGP Intercontinental Champion in his first reign. Jericho also makes sporadic
appearances for WWE on the SmackDown brand, with whom he has worked since 1999. Within
professional wrestling, he is known for his over-the-top, rock star persona. Jericho is regarded as
one of the greatest professional wrestlers of all time. During the 1990s, Jericho performed for the
American organizations Extreme Championship ... </information>
<think> I found out that Chris Jericho is a professional wrestler, musician, author, and actor. Now
I can compare it to Gary Barlow's professions. I found out that Gary Barlow is a singer, songwriter,
musician, and record producer. The common profession between Chris Jericho and Gary Barlow is
musician. </think>
<answer> musician </answer>
```

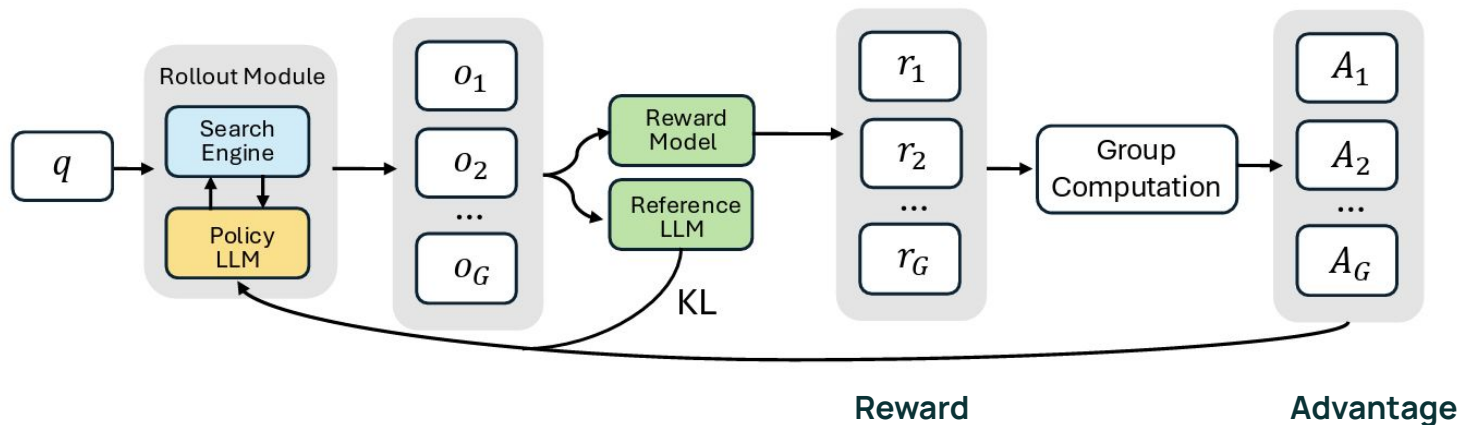
Actions

- `<think> [reasoning tokens] </think>`
- `<search> [query] </search>`
- `<information> [doc] </information>`
- `<answer> [final_answer] </answer>`

**Search-R1: Training LLMs to Reason and
Leverage Search Engines with
Reinforcement Learning.**

Jin et al. COLM 2025.

Search-R1: GRPO with Search Engine



- **Training data** (q, a) Natural Questions, HotpotQA
- **Reward** $r_\phi(x, y) = \text{EM}(a_{\text{pred}}, a_{\text{gold}})$
- **Search Engine** 2018 Wikipedia + E5 retriever

Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning.

Jin et al. COLM 2025. 24

Search-R1: Main Results

Methods	General QA				Multi-Hop QA			
	NQ ⁺	TriviaQA [*]	PopQA [*]	HotpotQA ⁺	2wiki [*]	Musique [*]	Bamboogle [*]	Avg.
Qwen2.5-7b-Base/Instruct								
Direct Inference	0.134	0.408	0.140	0.183	0.250	0.031	0.120	0.181
CoT	0.048	0.185	0.054	0.092	0.111	0.022	0.232	0.106
IRCoT	0.224	0.478	0.301	0.133	0.149	0.072	0.224	0.239
Search-o1	0.151	0.443	0.131	0.187	0.176	0.058	0.296	0.206
RAG	0.349	0.585	0.392	0.299	0.235	0.058	0.208	0.304
SFT	0.318	0.354	0.121	0.217	0.259	0.066	0.112	0.207
R1-base	0.297	0.539	0.202	0.242	0.273	0.083	0.296	0.276
R1-instruct	0.270	0.537	0.199	0.237	0.292	0.072	0.293	0.271
Rejection Sampling	0.360	0.592	0.380	0.331	0.296	0.123	0.355	0.348
Search-R1-base	0.480	0.638	0.457	0.433	0.382	0.196	0.432	0.431
Search-R1-instruct	0.393	0.610	0.397	0.370	0.414	0.146	0.368	0.385

Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning.

Search-R1: GRPO with Search Engine

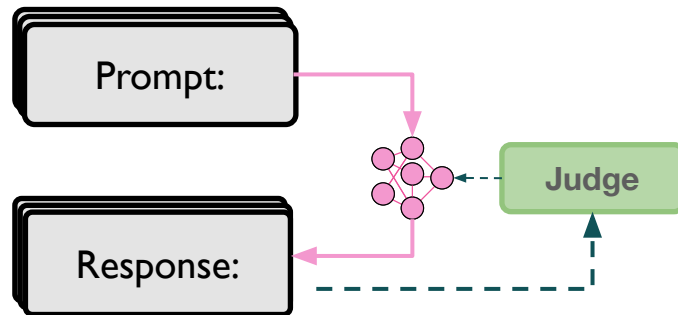


- Training data (q, a) Natural Questions, HotpotQA
- Reward $r_\phi(x, y) = \text{EM}(a_{\text{pred}}, a_{\text{gold}})$
- Search Engine 2018 Wikipedia + E5 retriever

WebDancer: Multi-Step Multi-Tool Agent Learning

SFT with teacher trajectories

RLVR w/ LLM judge



LLM judge

Rejection sampling for SFT & **Reward** for DAPO are based on formatting and answer correctness (w/ LLM judge)

WebDancer: Multi-tool Calling

Question: I'm researching species that became invasive after people who kept them as pets released them. There's a certain species of fish that was popularized as a pet by being the main character of the movie Finding Nemo. According to the USGS, where was this fish found as a nonnative species, before the year 2020? I need the answer formatted as the five-digit zip codes of the places the species was found, separated by commas if there is more than one place.

Step 1

<think>

Okay, let's tackle this question step by step. The user wants to know the zip codes where the fish from Finding Nemo, which became invasive after being released by pet owners, was found as a

<tool_call>

```
{"name": "search", "arguments": {"query": "USGS Amphiprion ocellaris nonnative locations before 2020"}}
```

</tool_call>

<tool_response>

A Google search for 'USGS Amphiprion ocellaris nonnative locations before 2020' found 9 results:

WebDancer: Multi-tool Calling

Question: I'm researching species that became invasive after people who kept them as pets released them. There's a certain species of fish that was popularized as a pet by being the main

Step 2

<think>

Okay, so I need to find the absolute difference in tens of thousands between the population of chinstrap penguins from the Wikipedia page as of the end of 2018 and the population from the 2020 Nature.com article. The user also wants us to assume two penguins per breeding pair. Let me start by going through the information provided. First, looking at the search results for the Wikipedia part. The sixth result mentions the Chinstrap penguin's conservation status, stating that in 2018, the IUCN estimated around 8 million specimens...

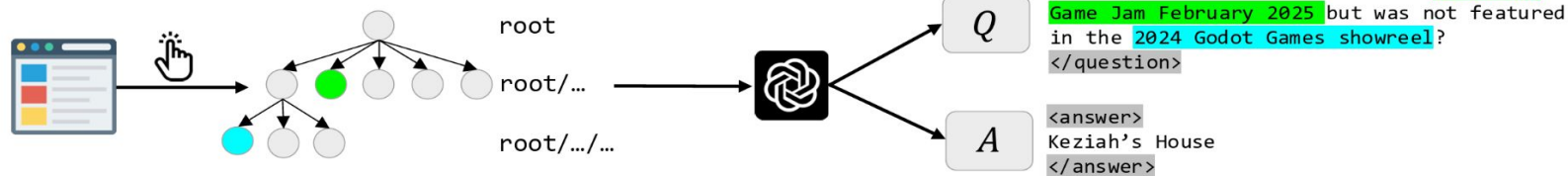
</think>

<tool_call>

```
{"name": "visit", "arguments": {"url": "https://nas.er.usgs.gov/queries/FactSheet.aspx?speciesID=3243", "goal": "Identify states with nonindigenous occurrences of Amphiprion ocellaris before 2020"}}
```

WebDancer: Data Synthesis

CRAWLQA



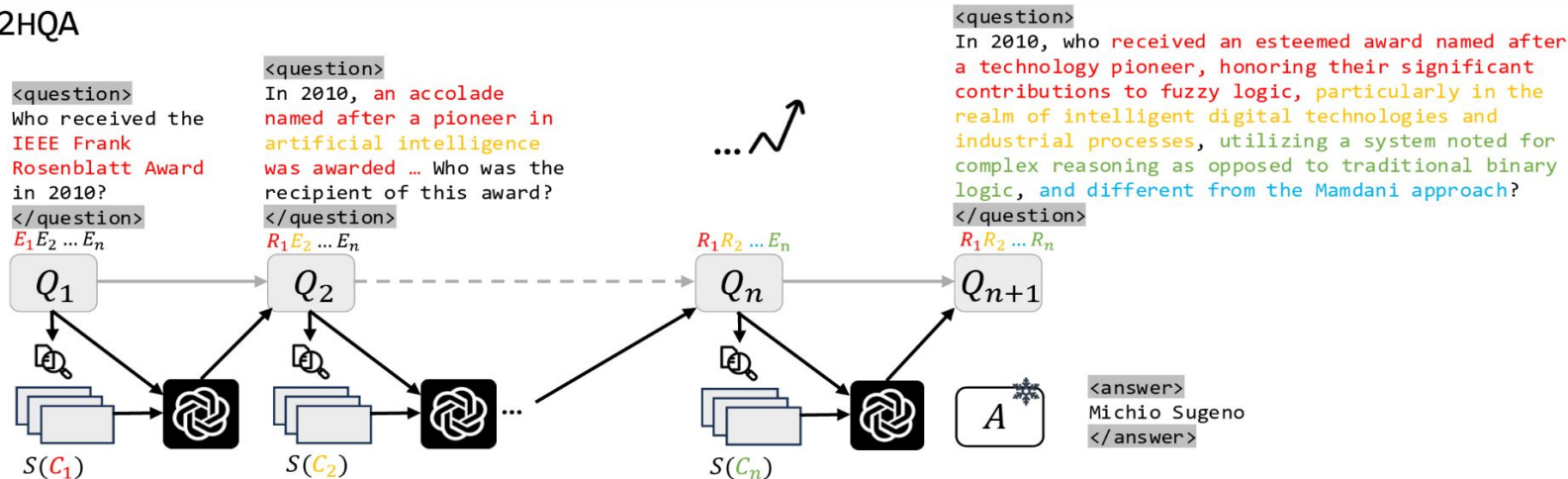
Step 1: Identify source website URL (arXiv, Wiki, Github ... etc)

Step 2: Visit linked websites by clicking hyperlinks

Step 3: GPT4o synthesizes QA pairs from the collected contents based on predefined types (e.g., Count)

WebDancer: Data Synthesis

E2HQA



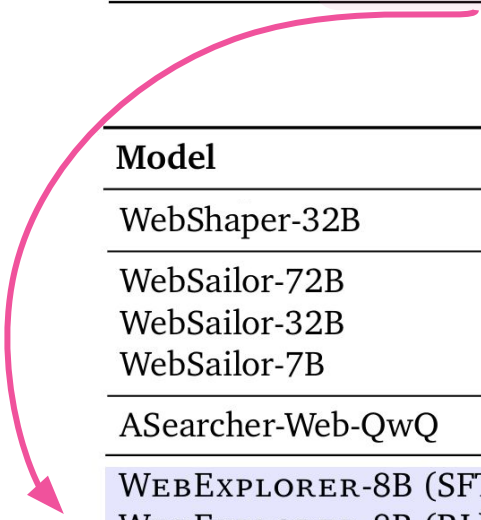
Step 1: Identify key entities in seed questions

Step 2: Retrieve articles related to the entities

Step 3: Replace the original entity based on the new information

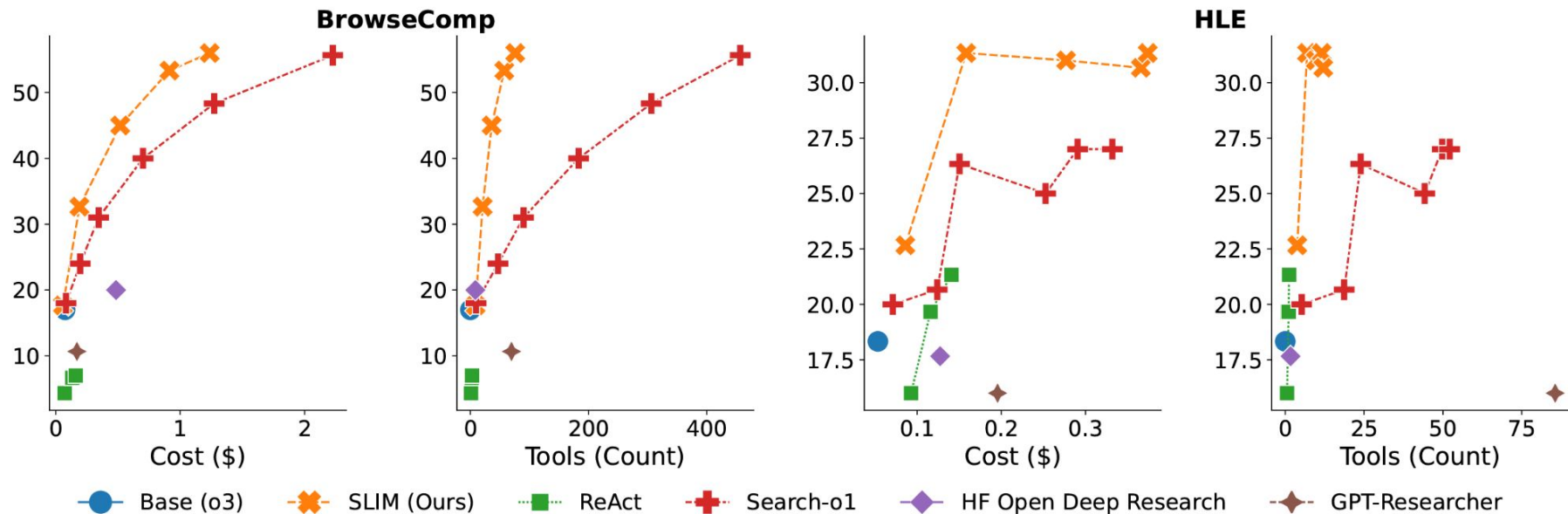
Synthetic QA Quality Heavily Affects Performance

Dataset	Initial QA	Evolved QA	WebDancer	SailorFog	WebShaper	ASearcher
Average Turns	7.9	9.9	5.4	8.2	8.4	6.5
Accuracy (%)	86.6	67.1	62.0	35.0	67.4	62.0



Model	BC-en	BC-zh	GAIA	WebWalkerQA	FRAMES	Xbench-DS	HLE
WebShaper-32B	-	-	52.4	51.4	-	-	-
WebSailor-72B	12.0	30.1	55.4	-	-	55.0	-
WebSailor-32B	10.5	25.5	53.2	-	-	53.3	-
WebSailor-7B	6.7	14.2	33.0	-	-	34.3	-
ASearcher-Web-QwQ	5.2	15.6	52.8	34.3	70.9	42.1	12.5
WEBEXPLORER-8B (SFT)	7.9	21.3	43.7	59.8	72.6	47.5	16.0
WEBEXPLORER-8B (RL)	<u>15.7</u>	<u>32.0</u>	<u>50.0</u>	<u>62.7</u>	<u>75.7</u>	<u>53.7</u>	<u>17.3</u>

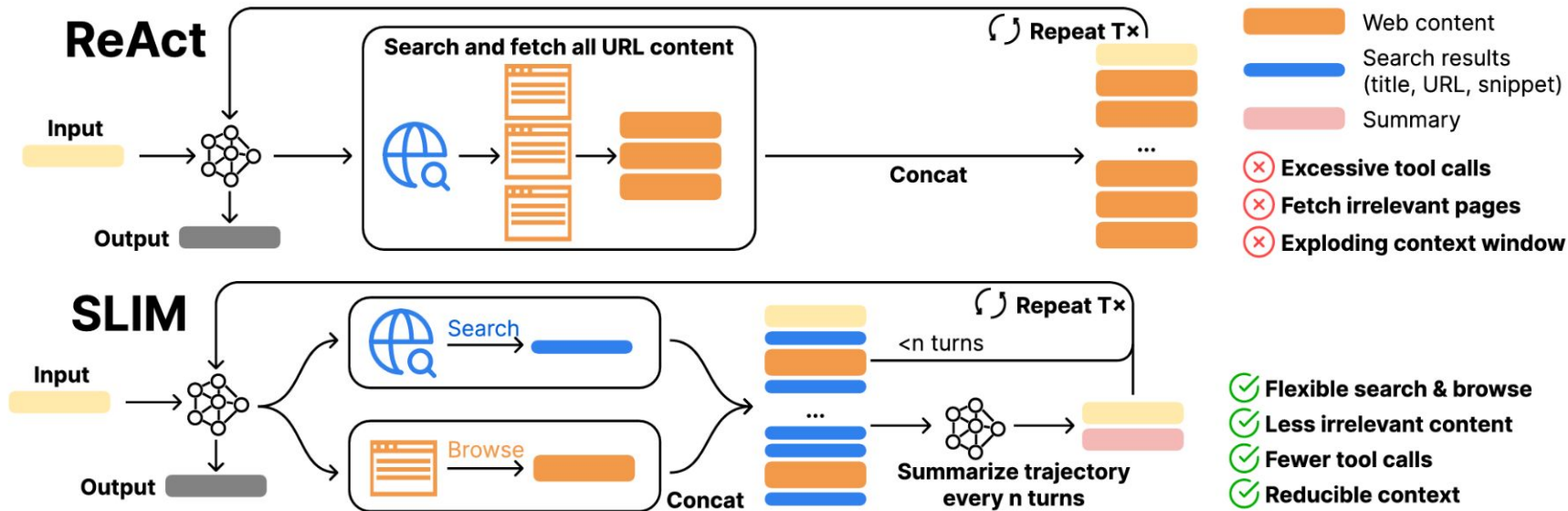
DR Requires Many Tool Calls



Lost in the Maze: Overcoming Context Limitations in Long-Horizon Information-Seeking

Yen et al. 2025

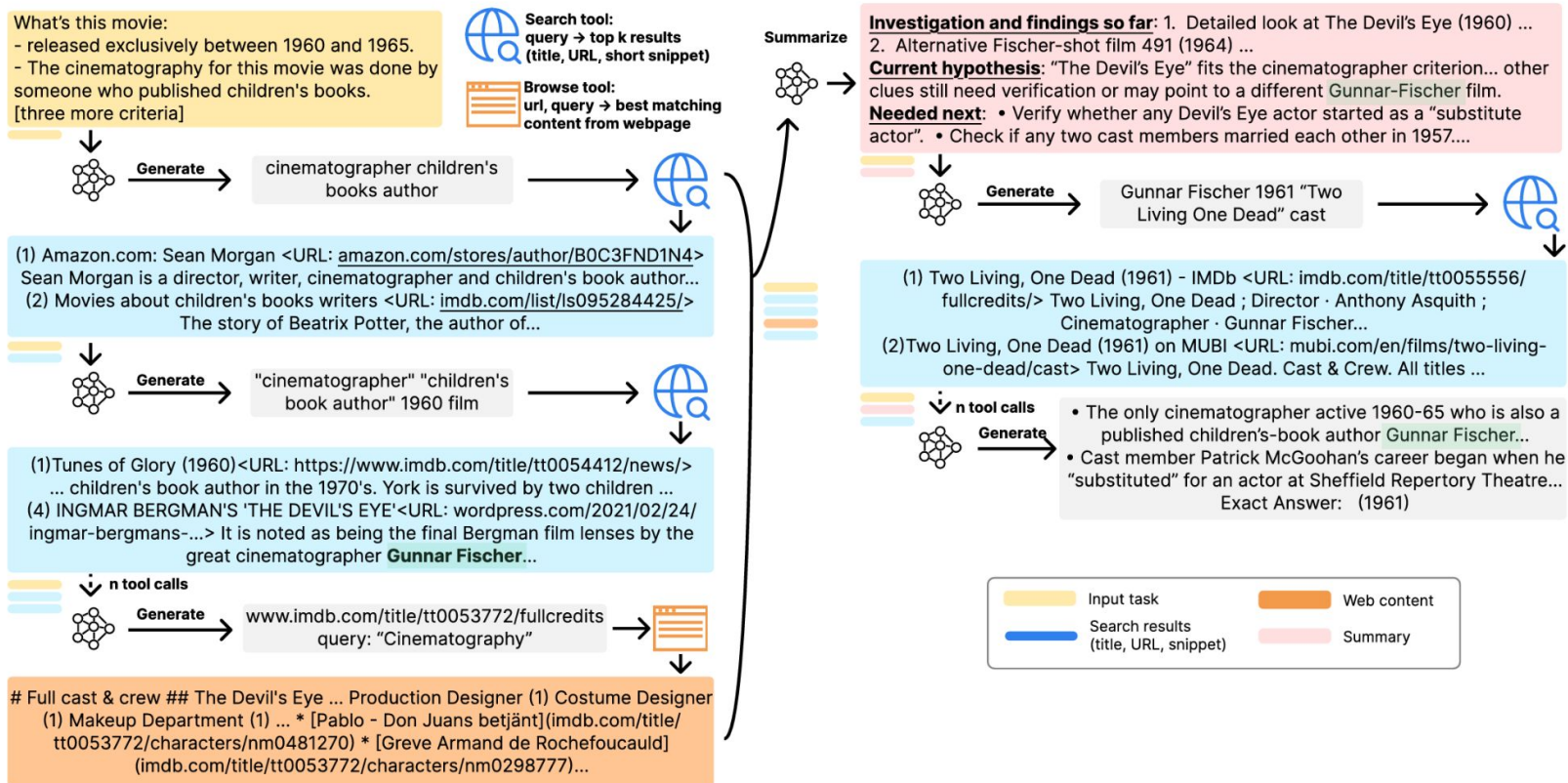
Context Summarization in DR Systems



Lost in the Maze: Overcoming Context Limitations in Long-Horizon Information-Seeking

Yen et al. 2025

Context Summarization in DR Systems



Tongyi Deep Research



- [1] [WebWalker: Benchmarking LLMs in Web Traversal](#) (ACL 2025)
- [2] [WebDancer: Towards Autonomous Information Seeking Agency](#) (NeurIPS 2025)
- [3] [WebSailor: Navigating Super-human Reasoning for Web Agent](#)
- [4] [WebShaper: Agentic Data Synthesizing via Information-Seeking Formalization](#)
- [5] [WebWatcher: Breaking New Frontier of Vision-Language Deep Research Agent](#)
- [6] [WebResearcher: Unleashing unbounded reasoning capability in Long-Horizon Agents](#)
- [7] [ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization](#)
- [8] [WebWeaver: Structuring Web-Scale Evidence with Dynamic Outlines for Open-Ended Deep Research](#)
- [9] [WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning](#)
- [10] [Scaling Agents via Continual Pre-training](#)
- [11] [Towards General Agentic Intelligence via Environment Scaling](#)
- [12] [AgentFold: Long-Horizon Web Agents with Proactive Context Management](#)
- [13] [WebLeaper: Empowering Efficient, Info-Rich Seeking for Web Agents](#)
- [14] [BrowseConf: Confidence-Guided Test-Time Scaling for Web Agents](#)
- [15] [Repurposing Synthetic Data for Fine-grained Search Agent Supervision](#)
- [16] [ParallelMuse: Agentic Parallel Thinking for Deep Information Seeking](#)
- [17] [AgentFrontier: Expanding the Capability Frontier of LLM Agents with ZPD-Guided Data Synthesis](#)
- [18] [Nested Browser-Use Learning for Agentic Information Seeking](#)

<https://github.com/Alibaba-NLP/DeepResearch>

*Additional topics I didn't cover: Agentic Reinforcement Learning (will cover in next), mid-training for DR, model merging

Tongyi Deep Research.
Tongyi Deep Research Team., 2025

Realistic *Unverifiable* Deep Research Questions



What's the state-of-the-art in immunotherapy for cervical cancer?

... Pembrolizumab can extend progression-free and overall survival in patients with persistent, recurrent, or metastatic PD-L1-positive cervical cancer [1] and locally advanced cervical cancer [2]...

[1] Colombo et al., KEYNOTE-826, 2021

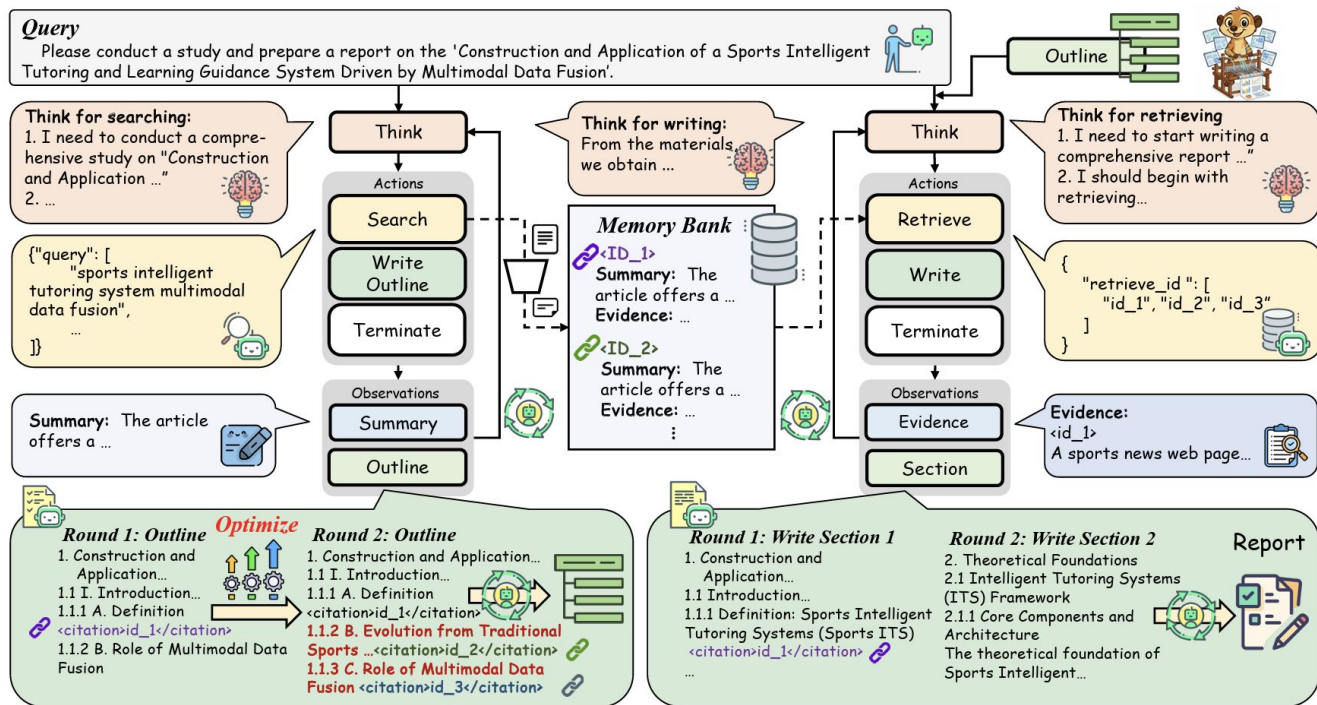
[2] Lorusso et al., KEYNOTE-A18, 2024

Attribution

Credible
sources

Up-to-date

Human-engineered Scaffolds for LF DR



Human-engineered Scaffolds for LF DR

Agent systems	RACE					FACT	
	Overall	Comp.	Insight	Inst.	Read.	Eff. c.	C. acc.
WebShaper (32B)	34.93	31.58	26.17	44.81	40.38	-	-
langchain-open-deep-research	43.44	42.97	39.17	48.09	45.22		
doubao-research	44.34	44.84	40.56	47.95	44.69	52.62	52.86
kimi-research	44.64	44.96	41.97	47.14	45.59	-	-
Claude-research	45.00	45.34	42.79	47.58	44.66	-	-
openai-deepresearch	46.45	46.46	43.73	49.39	47.22	39.79	75.01
Gemini-2.5-pro-deepresearch	49.71	49.51	49.45	<u>50.12</u>	50.00	165.34	<u>78.30</u>
WebWeaver (Qwen3-30b-a3b-instruct-2507)	46.77	45.15	45.78	49.21	47.34	26.74	25.00
WebWeaver (GPT-oss-120b)	48.11	48.03	47.20	48.94	48.11	64.88	66.14
WebWeaver (Qwen3-235b-a22b-instruct-2507)	50.62	<u>51.29</u>	51.00	49.98	48.89	<u>166.73</u>	78.25
WebWeaver (Claude-sonnet-4-20250514)	<u>50.58</u>	51.45	<u>50.02</u>	50.81	<u>49.79</u>	200.75	93.37

How do we train open DR agents end-to-end for tasks without simple verification?

DR Tulu: Open DR Agents for Open-ended Tasks



User

How did Netflix manage to successfully adapt One Hundred Years of Solitude, a notoriously difficult book to bring to the screen?



DR Tulu

Agentic Workflow

Think Tool1 Think Tool2 ... Answer

Long-form Report with Citations

Netflix's adaptation avoided a literal replica of Macondo and instead fused real locations with meticulously built sets^[1] to honor the novel's essence while giving the show contemporary visual grammar. The production grounded magical realism in front-of-camera practical craft, relying on makeup, special effects,^[2] The location strategy and production design...

Sources

- [1] The production team behind Netflix's adaptation of "One Hundred Years of Solitude," LA Times
- [2] The article discusses Netflix's adaptation of Gabriel García Márquez's celebrated novel, NY Times



Performs both open-ended, long-form tasks and short-form tasks



Reason to choose effective search MCP tools



Provide snippet-level citations for each claim from credible sources

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. Shao*, Asai*, Shen*, Iverson* et al., 2025

[Check our paper! →](#)

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research

 Code

 Data & Models

 Blogpost

DR Tulu Deep Research Agent

 Export

 Load

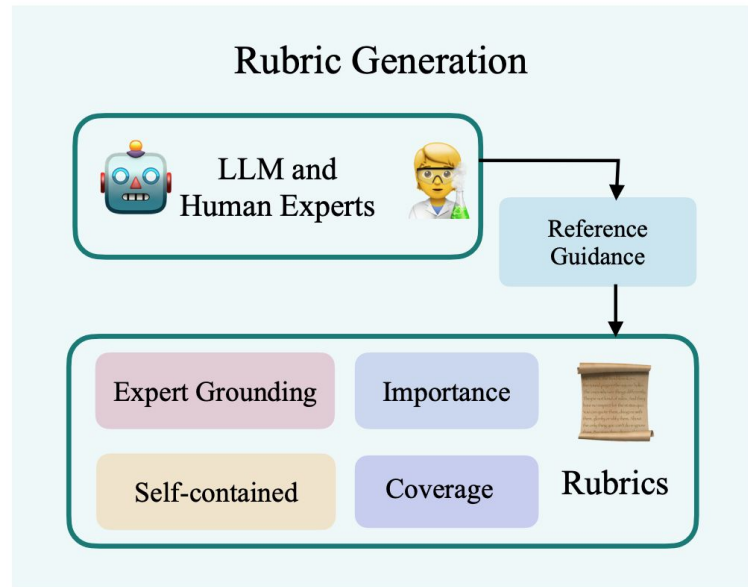
Ask a research question to get started!

Ask a research question...



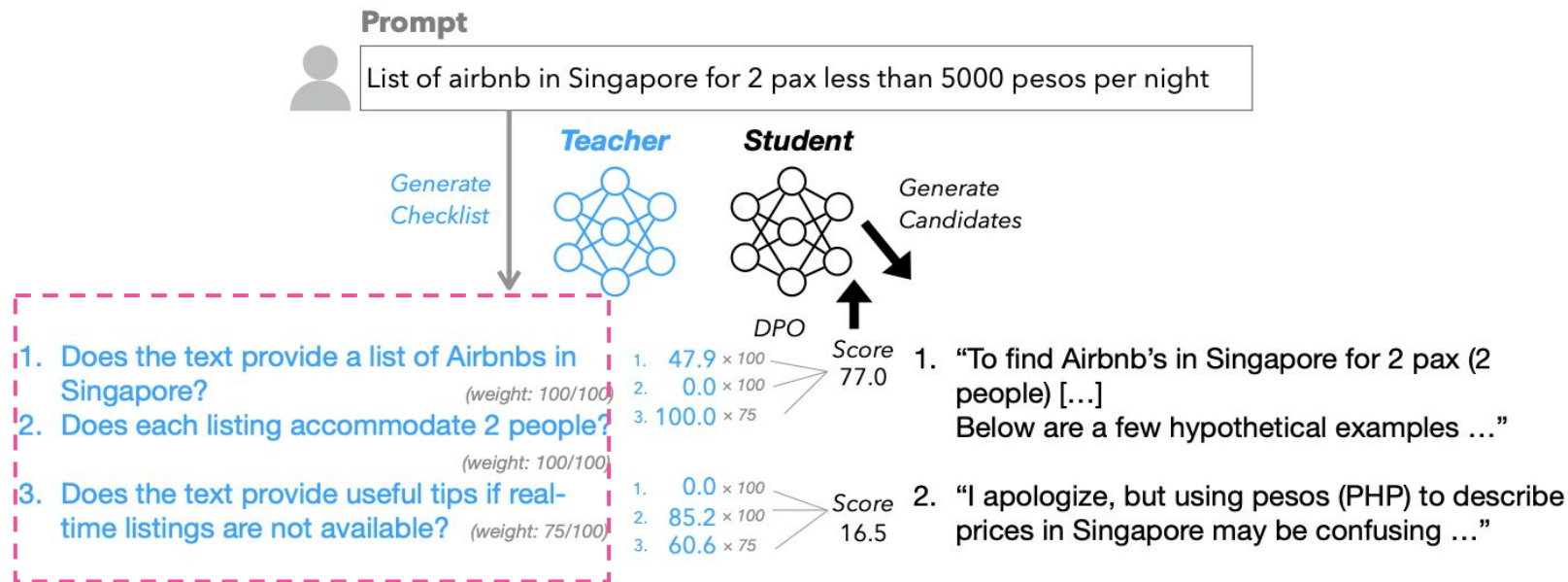
Background: RL with Rubric Rewards

Using rubric-based rewards
to enhance RL



Rubrics as Rewards: Reinforcement Learning Beyond Verifiable Domains
Gunjal, et al. NeurIPS 2025.

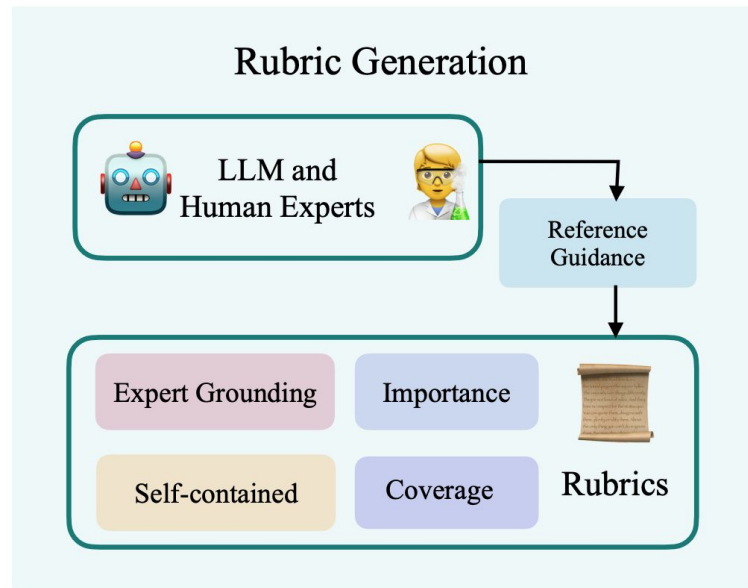
Background: RL with Rubric Rewards



Checklists Are Better Than Reward Models For Aligning Language Models
Viswanathan, et al. NeurIPS 2025.

Problem of *Static* Rubrics

- ✗ Requires human experts or a strong teacher model
- ✗ Not grounded on up-to-date knowledge
- ✗ Not adaptive to the policy distribution



Rubrics as Rewards: Reinforcement Learning Beyond Verifiable Domains
Gunjal, et al. NeurIPS 2025.

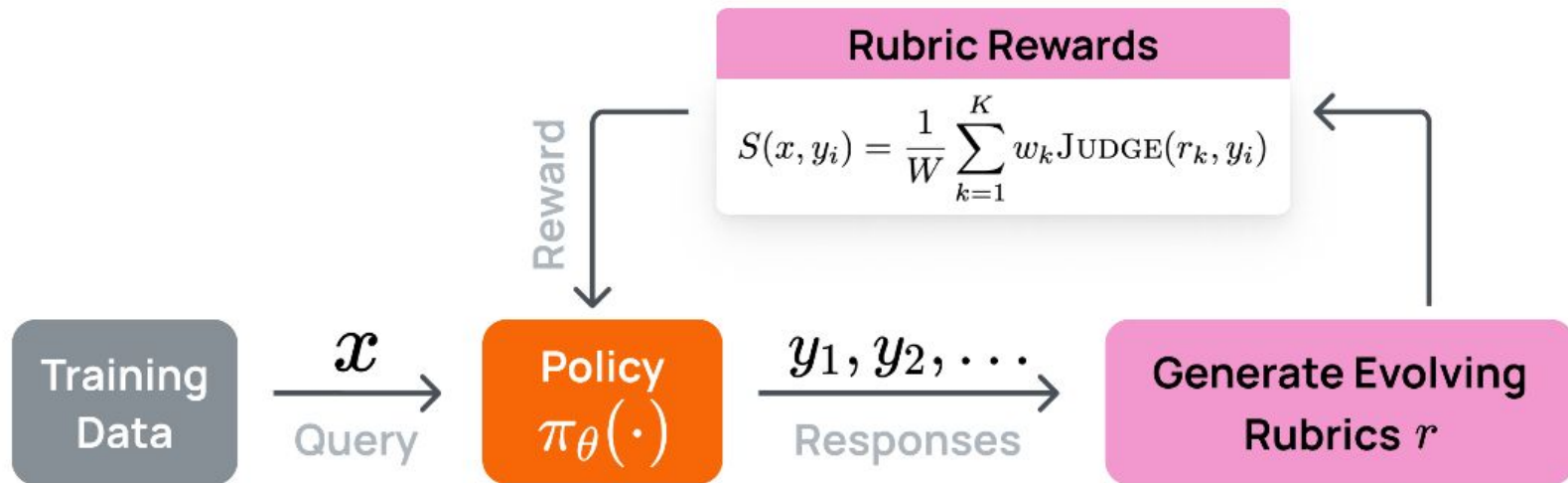
DR Tulu: RL with Evolving Rubrics (RLER)

Key idea:

1. Ground rubric generation on **up-to-date searched info**.
2. Generate fine-grained rubrics by **contrasting model responses**.

Make rubrics **co-evolve** with the policy model

Overview of RL with Evolving Rubrics (RLER)



RL with Evolving Rubrics (RLER)

User query

Describe how genetically engineering T cells could be used as an anti-inflammatory therapy for IBD by changing cytokine signaling between T cells and macrophages.

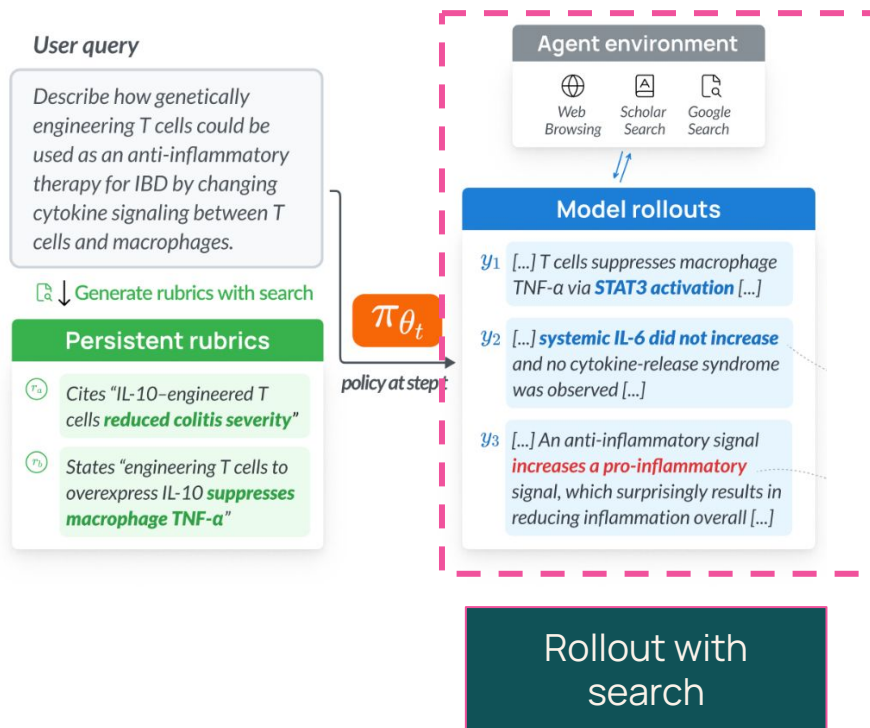
Generate rubrics with search

Persistent rubrics

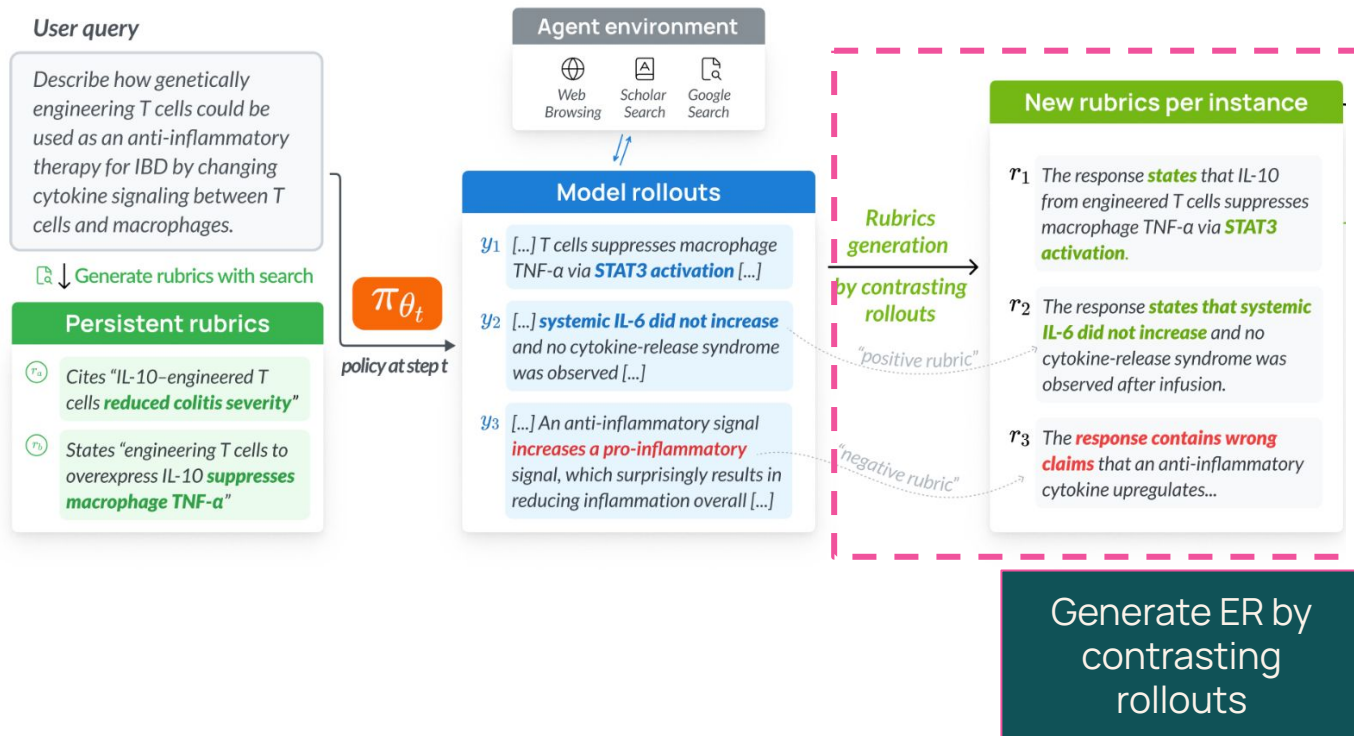
- r_0 Cites "IL-10-engineered T cells **reduced colitis severity**"
- r_1 States "engineering T cells to overexpress IL-10 **suppresses macrophage TNF- α** "

Init. rubrics

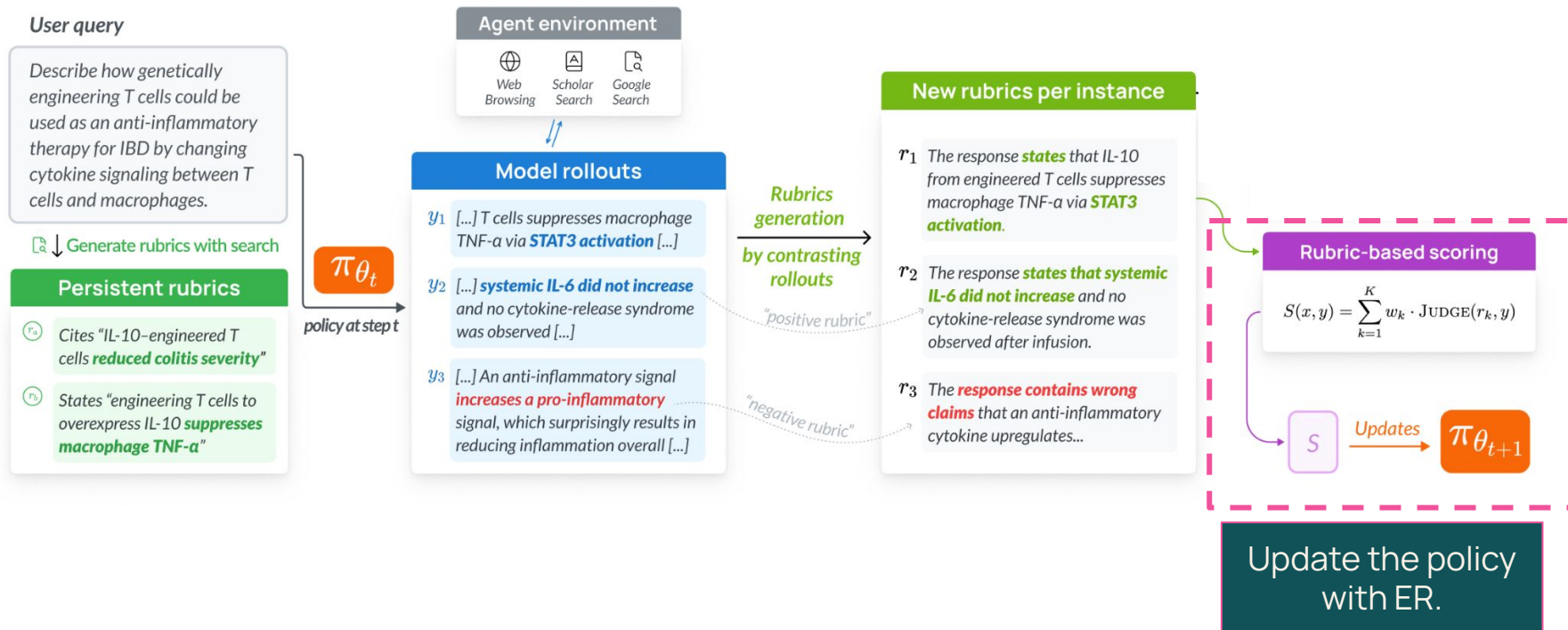
RL with Evolving Rubrics (RLER)



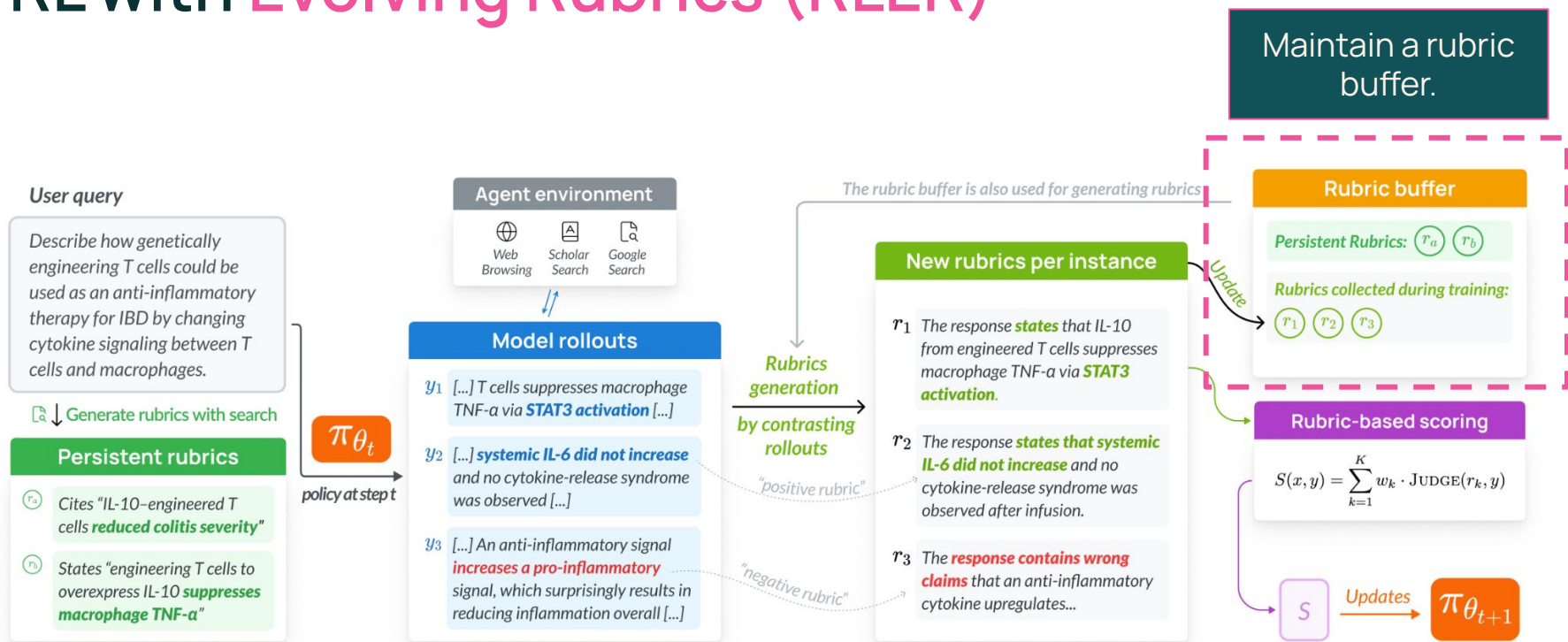
RL with Evolving Rubrics (RLER)



RL with Evolving Rubrics (RLER)



RL with Evolving Rubrics (RLER)



DR Tulu: RL with Evolving Rubrics (RLER)



How did Netflix manage to successfully adapt One Hundred Years of Solitude, a notoriously difficult book to bring to the screen?



Too vague.

π_{θ_0} 1

Precisely describe the adaptation process.



Closed-book Rubric (baseline)



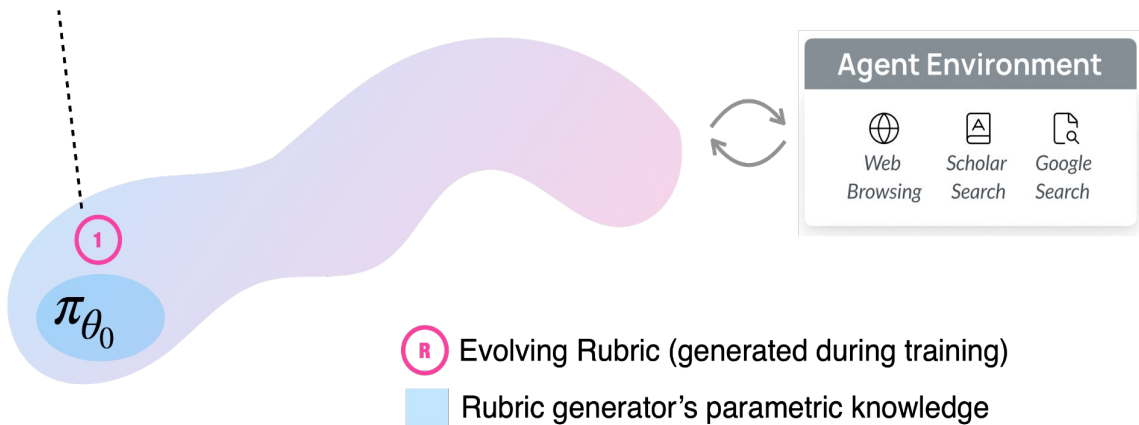
Rubric generator's parametric knowledge

DR Tulu: RL with Evolving Rubrics (RLER)



How did Netflix manage to successfully adapt One Hundred Years of Solitude, a notoriously difficult book to bring to the screen?

Mention the Netflix's adaptation avoided a literal replica.



R Evolving Rubric (generated during training)

▀ Rubric generator's parametric knowledge

epoch=0

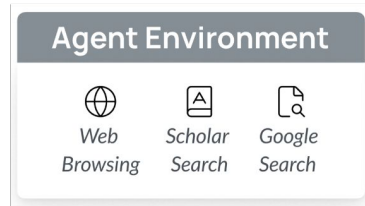
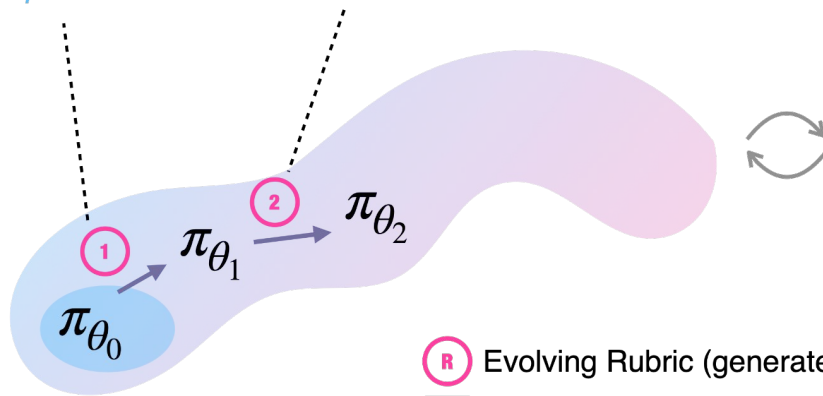
DR Tulu: RL with Evolving Rubrics (RLER)



How did Netflix manage to successfully adapt One Hundred Years of Solitude, a notoriously difficult book to bring to the screen?

Mention the Netflix's adaptation avoided a literal replica.

Mention that Netflix worked closely with García Márquez's family.



Evolving Rubric (generated during training)



Rubric generator's parametric knowledge



Explored knowledge during training by policy π_θ

epoch=1

DR Tulu: RL with Evolving Rubrics (RLER)

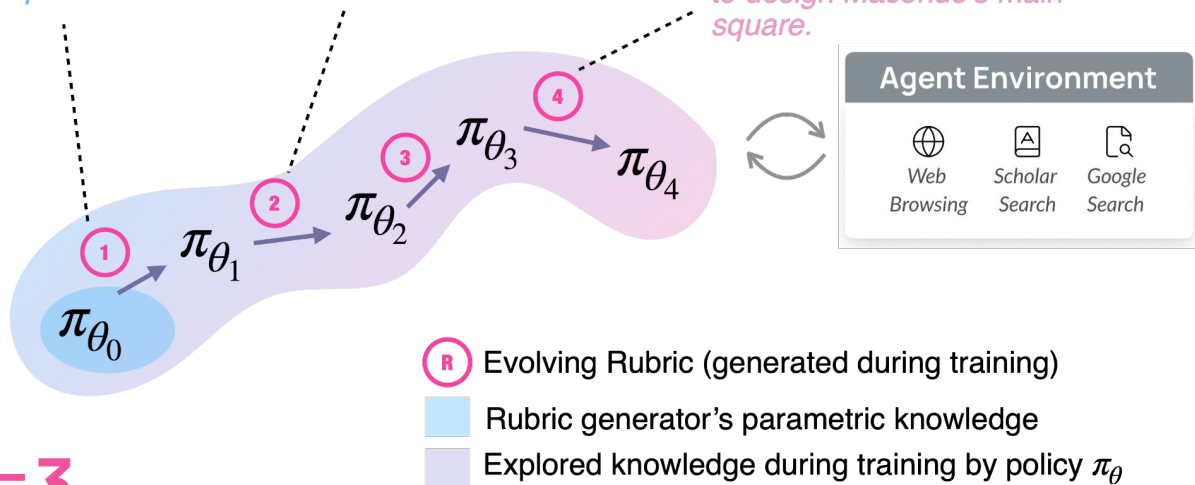


How did Netflix manage to successfully adapt One Hundred Years of Solitude, a notoriously difficult book to bring to the screen?

Mention the Netflix's adaptation avoided a literal replica.

Mention that Netflix worked closely with García Márquez's family.

Mention that Netflix used 3D scans of real Colombian towns to design Macondo's main square.



epoch=3

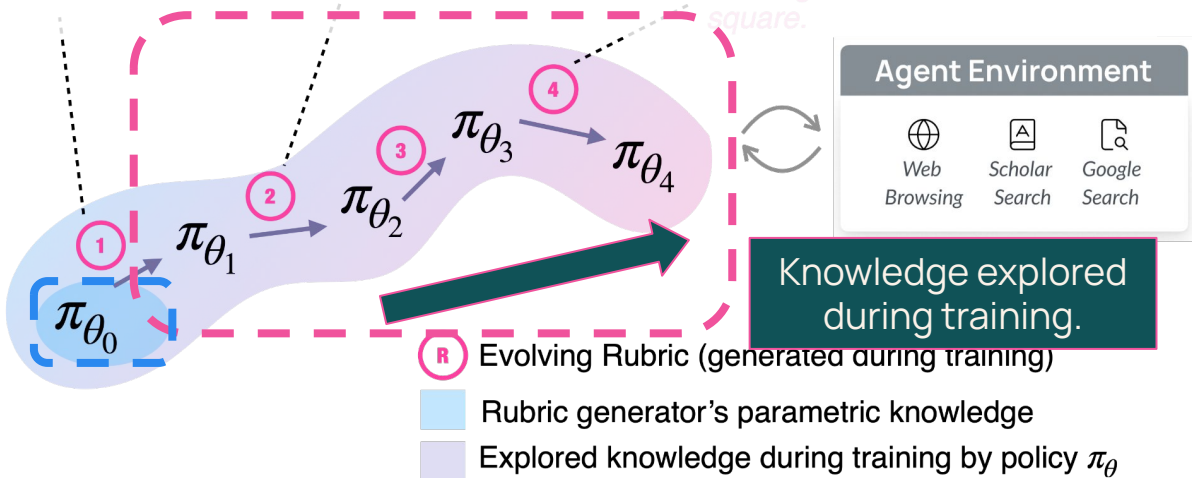
DR Tulu: RL with Evolving Rubrics (RLER)

Search-Guided Rubrics Co-Evolving with the Policy

Mention the Netflix's adaptation avoided a literal replica.

Mention that Netflix worked closely with García Márquez's family.

Mention that Netflix used 3D scans of real Colombian towns to design Macondo's main square.



DR Tulu: Training

SFT with teacher trajectories



Prompt:



OpenScholar User Qs



SearchArena user Qs



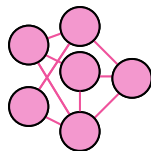
Response:



GPT5



QwQ-32B



DR Tulu-SFT

DR Tulu: Training



github.com/r1research/dr-tulu

SFT with teacher trajectories



RLER

Prompt:



OpenScholar User Qs



SearchArena user Qs

Response:



GPT5



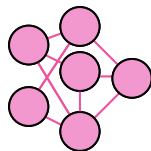
QwQ-32B

Prompt:

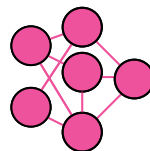
Response:



ER



DR Tulu-SFT



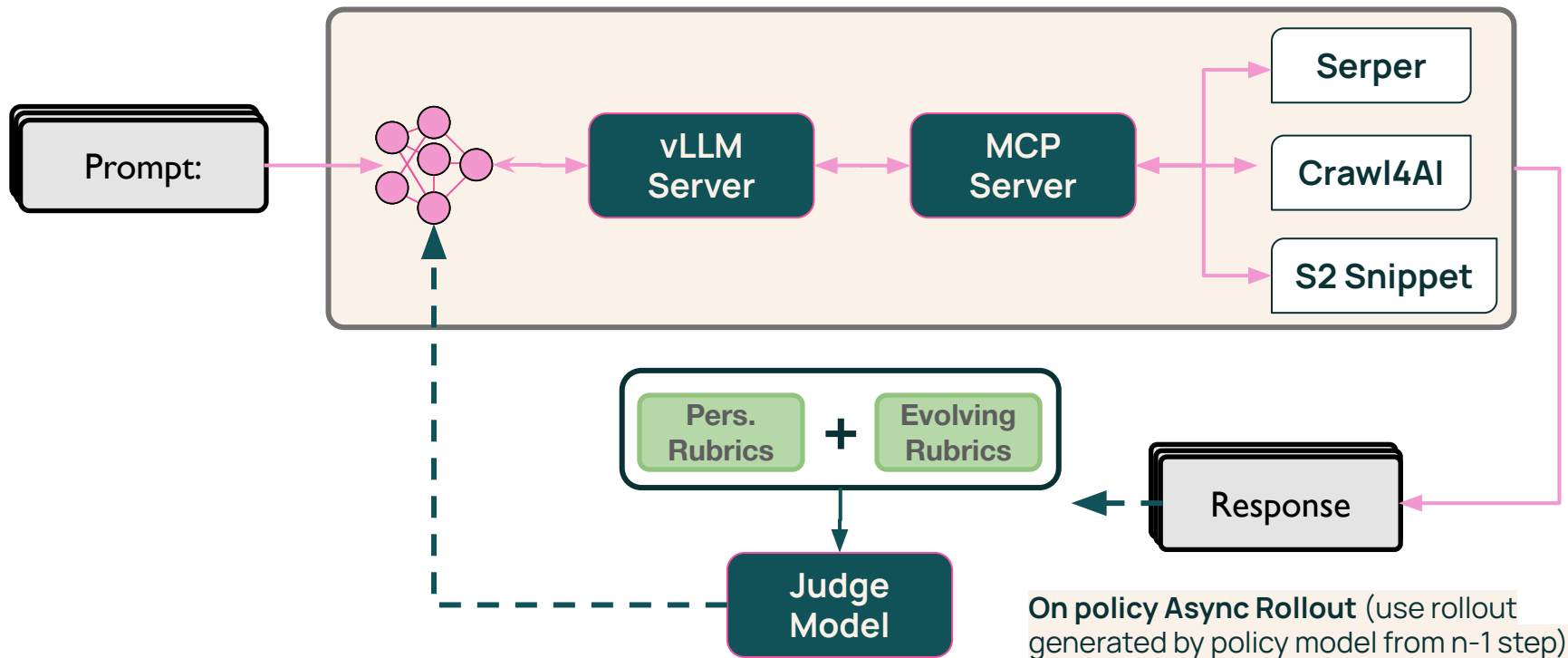
DR Tulu (RL)

Agentic RL

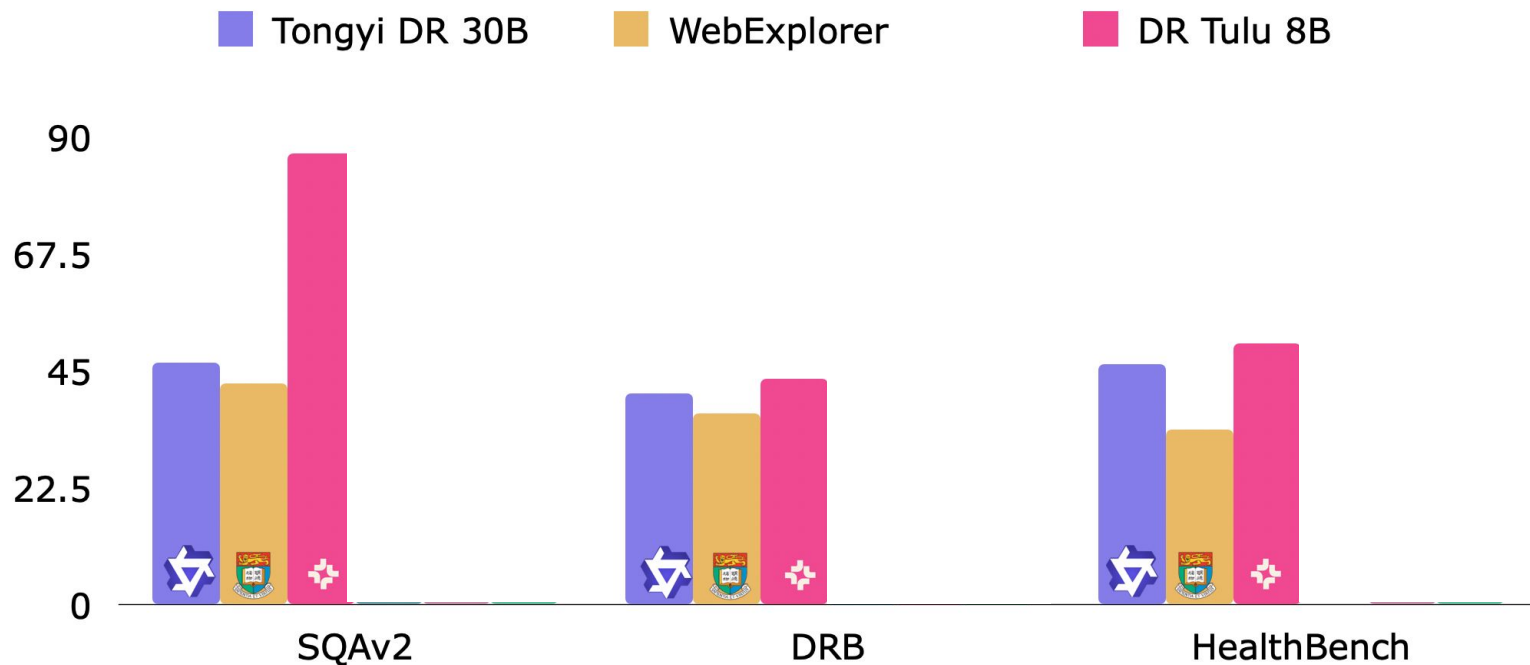


github.com/rlresearch/dr-tulu

Multi-tool Real-world environment



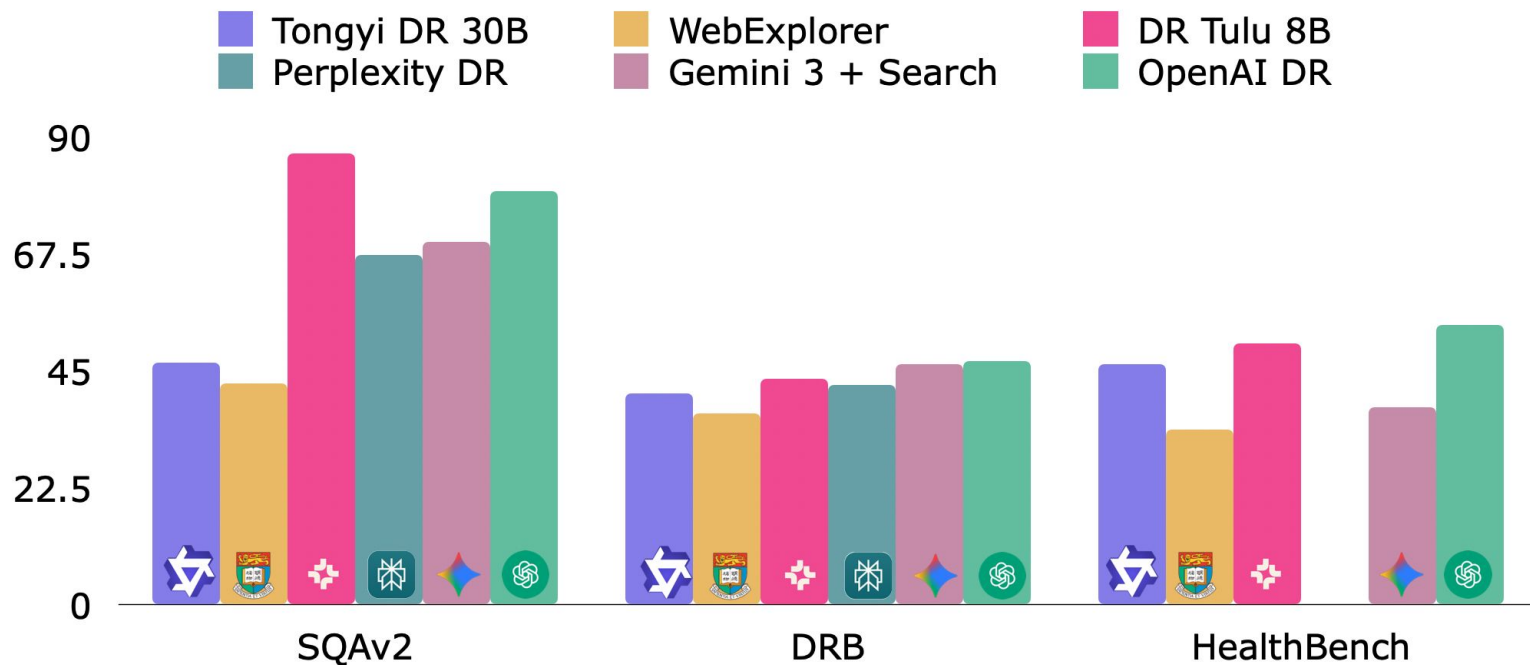
DR Tulu Significantly Outperforms Open DR Models



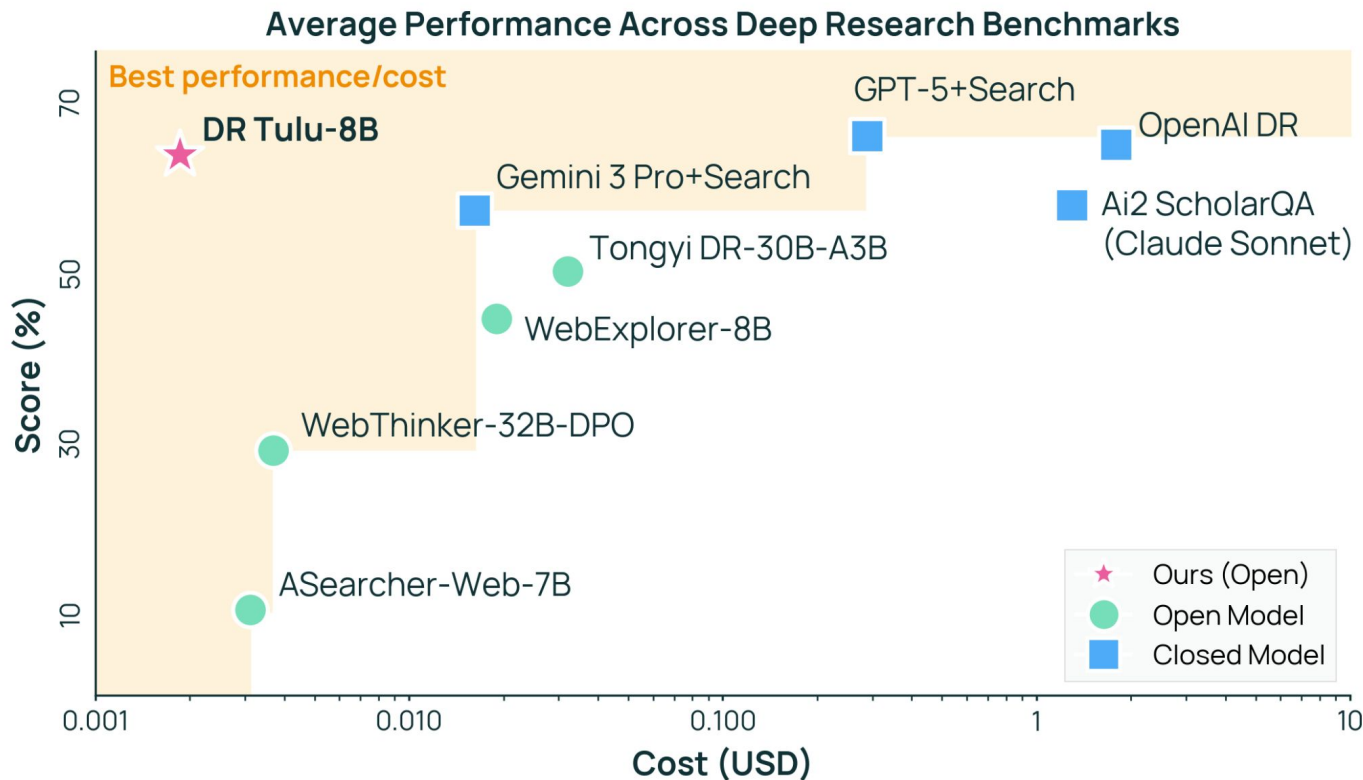
DR Tulu Significantly Outperforms Open DR Models



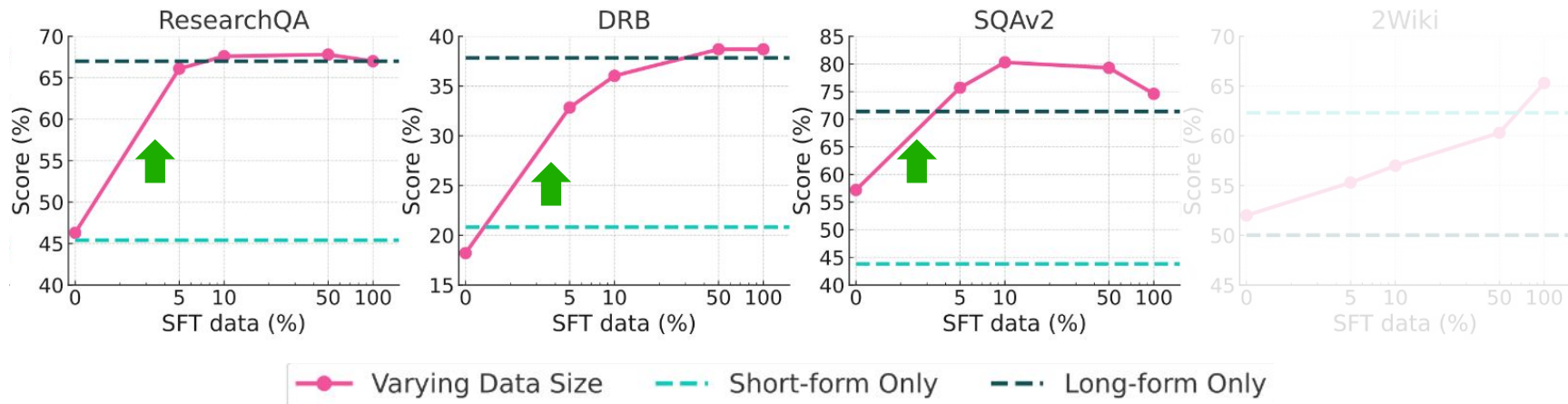
DR Tulu Significantly Outperforms Open DR Models



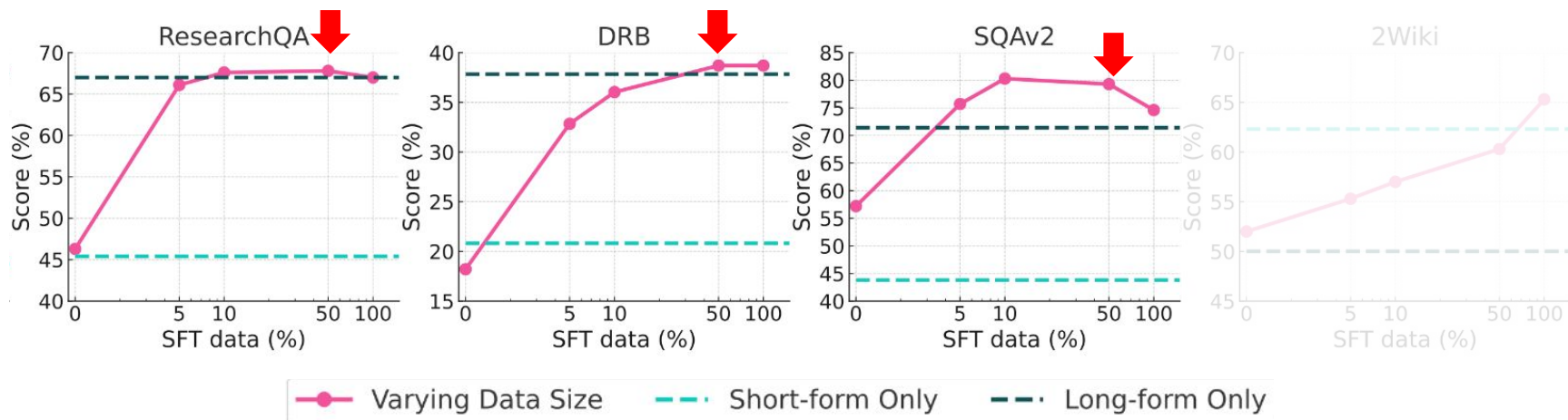
DR Tulu is Also Much Cheaper!



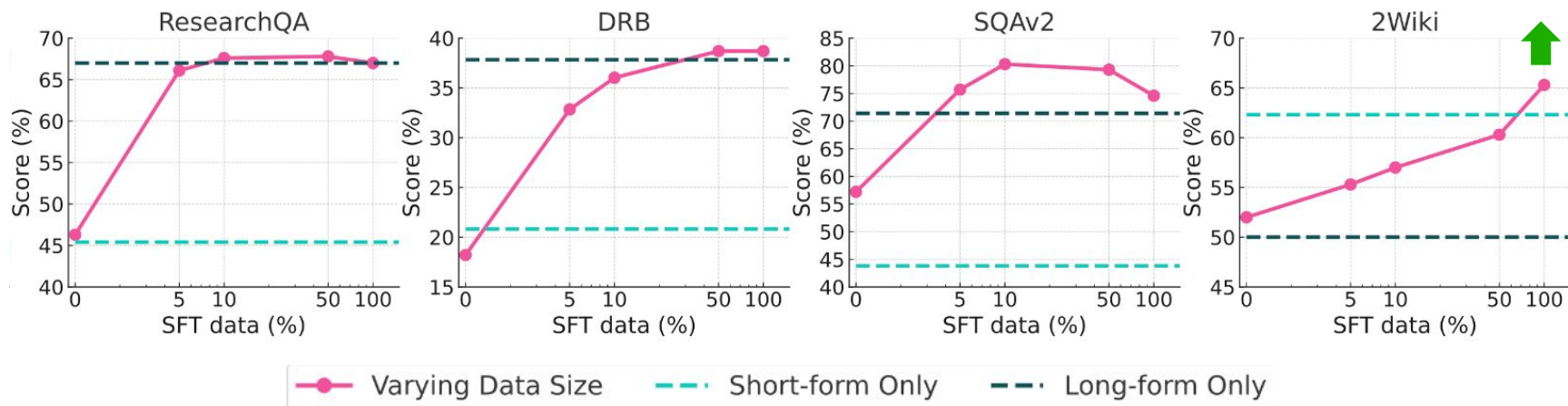
Finding #1: Small-scale SFT Improves OE DR



Finding #2: SFT Only Can't Improve OE DR Further



Finding #2: SFT Only Can't Improve OE DR Further



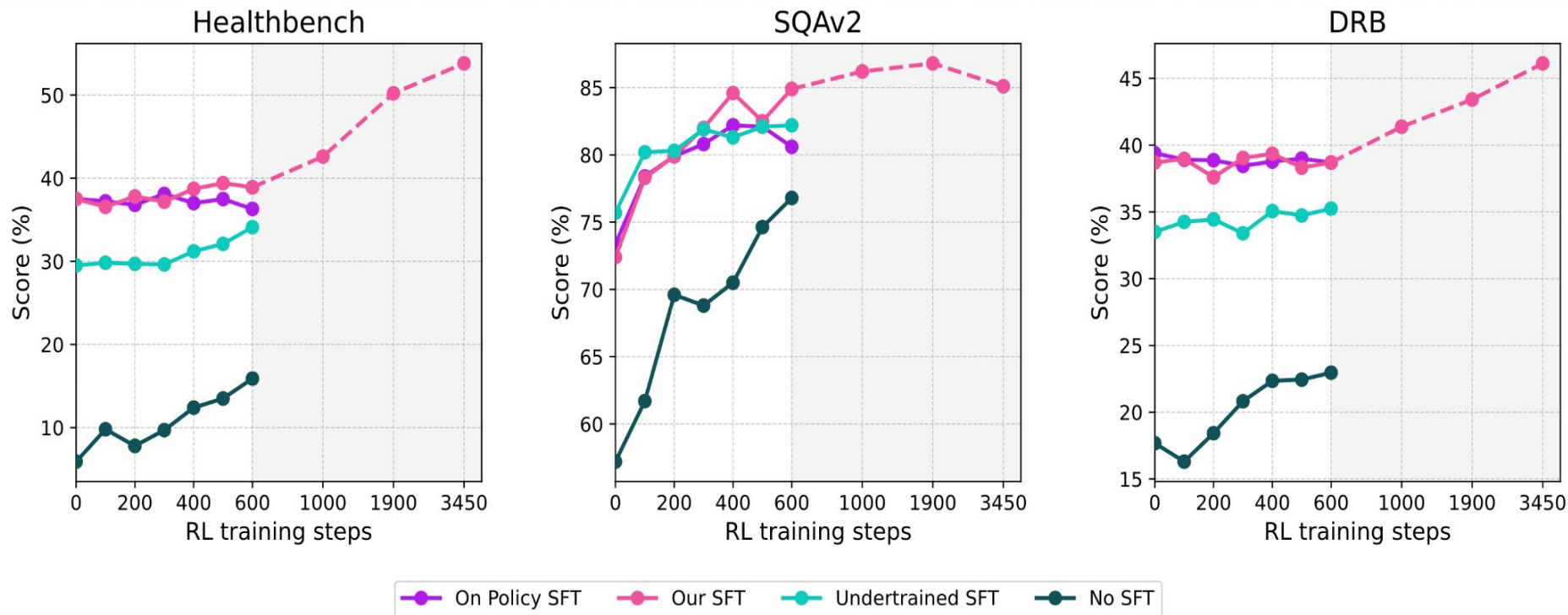
Finding #3: Importance of SFT for Cold Start



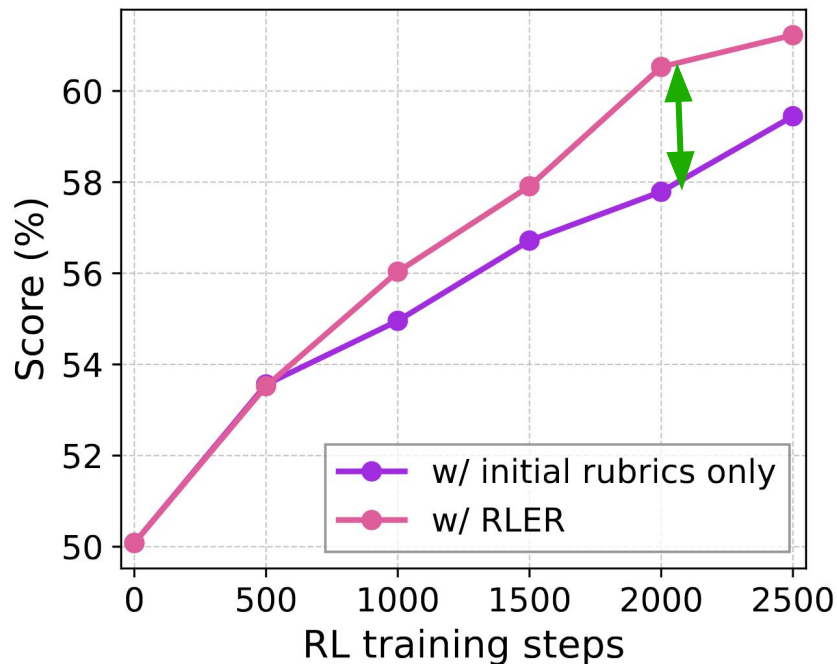
Finding #4: Long RL was Key



Finding #4: Longer RL is Key



Finding #5: Evolving Rubrics are Crucial

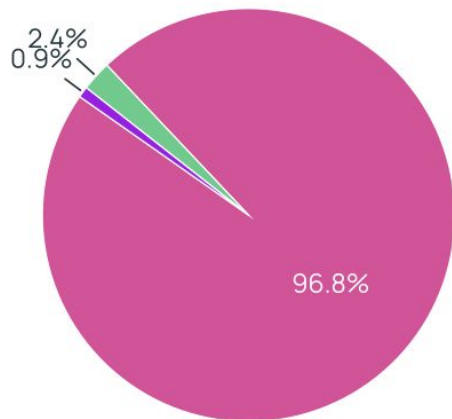


The advantage of evolving rubrics grows as training continues

Finding 5: DR Tulu Learns to Use Relevant Tools

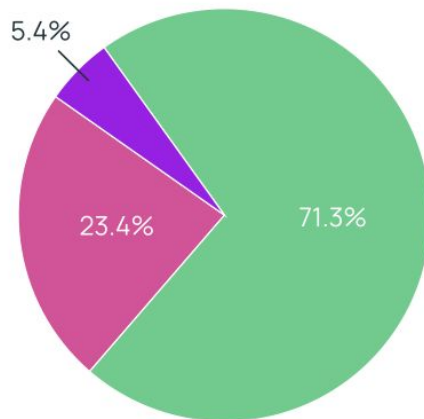
Science

SQAv2



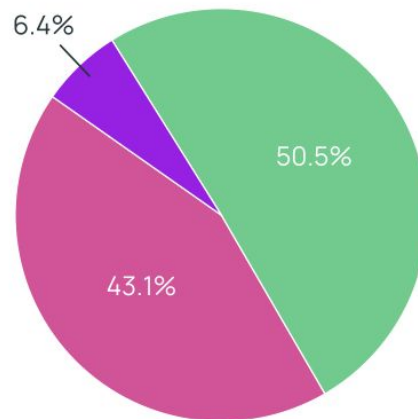
Healthcare

Healthbench



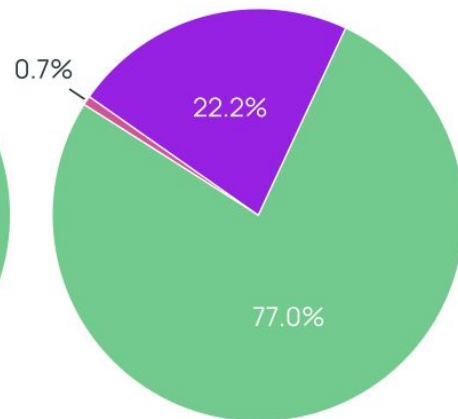
General domain

DeepResearchBench



Factoid SF QA

SimpleQA



■ Paper Search ■ Google Search ■ Browse Webpage

Summary: Recent Progress in Deep Research



Evaluation

- Hard-to-answer but easy-to-verify questions (eg BrowseComp)
- More and more rubric-based long-form deep research evals



Modeling

- SFT cold start on synthetic QA followed by RLVR
- RL with rubric-based rewards for long-form deep research

Thank you! Any questions?