Search Agents

Hanchen Li & Shangyin Tan

Nov 4, 2025

What are LLM agents?

Agents don't mean anything now in a lot of context (why?)

In the research world,

Consider a general setup of an agent interacting with an environment for task solving. At time step t, an agent receives an observation $o_t \in \mathcal{O}$ from the environment and takes an action $a_t \in \mathcal{A}$ following some policy $\pi(a_t|c_t)$, where $c_t = (o_1, a_1, \cdots, o_{t-1}, a_{t-1}, o_t)$ is the *context* to the agent.

In search agent, a_k is a search tool.

How is this different from (traditional) RAG? At best, RAG is multi-hop.
Agentic search system is "unlimited hop".

Search Agents

Search agents:

Input: Question

The agent will:

1. Check if they can not answer the question

2. Search the internet

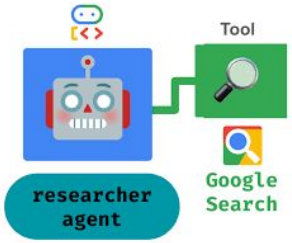
3. May repeat 1

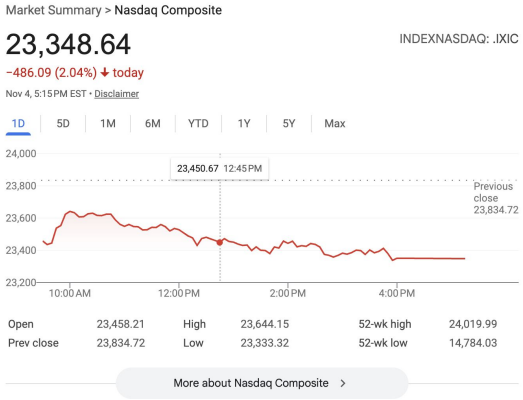
4. Answer based on the information collected

Search Agent Problem

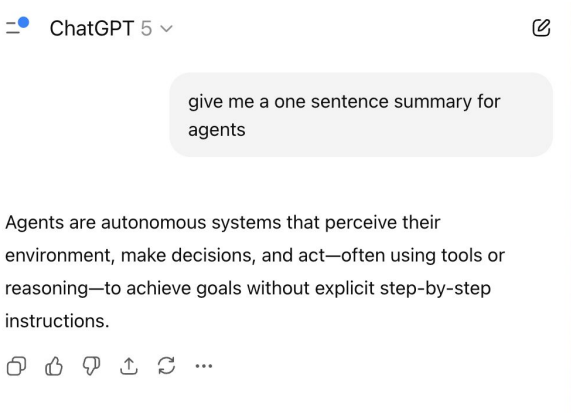
Example Query:

Give stock suggestions based on today's global economy and the coldplay event.





Fortunately, we have Agents! (And RL!)



Limitation of Pretraining

The stock option today is going xxx.

Xxx is optimized over historical data.

In a hypothetical situation, if 80% of human text says stock is going up, xxx will most likely be “up” instead of “down”.

You need to do web search!

Limitation of SFT

Can we address this through SFT? -> you can kinda do that by collecting trajectories offline, but it does not capture interaction in real web search.

The real web is changing every minute!

An Empirical Study on Reinforcement Learning for Reasoning-Search Interleaved LLM Agents

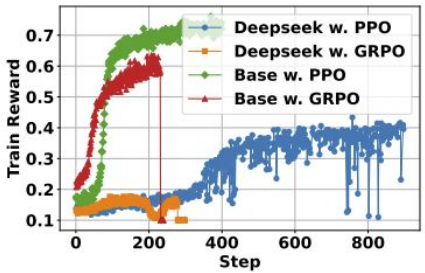
Bowen Jin[1]*, **Jinsung Yoon**[2], **Priyanka Kargupta**[1], **Sercan Ö. Arık**[2], **Jiawei Han**[1]
[1] University of Illinois at Urbana-Champaign
[2] Google Cloud AI Research
bowenj4@illinois.edu

How do we RL?

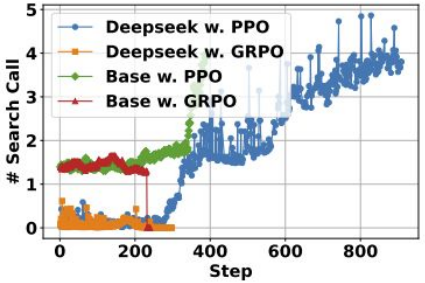
Questions:

1. Model:
    a. How do model scale and type (general vs. reasoning-specialized) affect training?
2. Tool:
    a. Does adding format or intermediate retrieval rewards help?
3. Training Model to use Tools:
    a. How to set the reward? Does adding intermediate reward help?

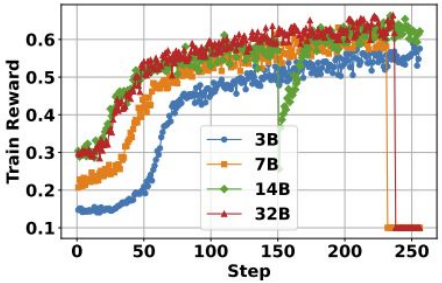
Model Types



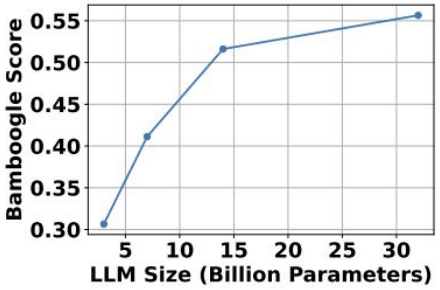
(a) Training Reward



(b) # of Search Calls



(c) Training Reward



(d) Test Accuracy

- General models train more stably & call search earlier
- Reasoning models struggle with format adherence
- 3B → 32B scaling improves performance but with diminishing returns

Reward Setting

Experimental Design. In addition to the outcome reward defined in [13, 58], we introduce a format reward, resulting in the final reward function $r_\phi(x, y)$:

$$r_\phi(x, y) = \begin{cases} 1 & \text{if } a_{\text{pred}} = a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{True}, \\ 1 - \lambda_f & \text{if } a_{\text{pred}} = a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{False}, \\ \lambda_f & \text{if } a_{\text{pred}} \neq a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{True}, \\ 0 & \text{if } a_{\text{pred}} \neq a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{False}, \end{cases} \tag{3}$$

- Speed up convergence
- Improves final accuracy, especially for non-instruction-tuned LLMs.
- But too large $\lambda_(f_)$ causes overfitting

Results on Format Reward

Table 1: Empirical study of the format reward. *Outcome only* refers to the RL variant with only the outcome reward. Base/Instruct refer to the version of the underlying LLM. $\lambda_f = 0.2$ for 3B/14B and $\lambda_f = 0.4$ for 7B. The best performance is set in bold. $^\dagger/^\star$ represents in-domain/out-domain datasets.

Methods		General QA			Multi-Hop QA				
		NQ†	TriviaQA*	PopQA*	HotpotQA†	2wiki*	Musique*	Bamboogle*	Avg.
Qwen2.5-3B-Base/Instruct									
PPO	Outcome only (base)	0.406	0.587	0.435	0.284	0.273	0.049	0.088	0.303
	w. format reward	**0.428**	**0.607**	**0.459**	**0.371**	**0.387**	**0.150**	**0.323**	**0.389**
	Outcome only (instruct)	0.341	0.545	0.378	0.324	**0.319**	0.103	0.264	0.325
	w. format reward	**0.356**	**0.557**	**0.393**	**0.327**	0.314	**0.122**	**0.266**	**0.334**
GRPO	Outcome only (base)	0.421	0.583	0.413	0.297	0.274	0.066	0.128	0.312
	w. format reward	**0.429**	**0.602**	**0.435**	**0.372**	**0.383**	**0.148**	**0.307**	**0.382**
	Outcome only (instruct)	**0.397**	**0.565**	**0.391**	**0.331**	**0.310**	**0.124**	0.232	**0.336**
	w. format reward	0.346	0.552	0.371	0.297	0.300	0.098	**0.266**	0.319
Qwen2.5-7B-Base/Instruct									
PPO	Outcome only (base)	0.480	0.638	0.457	0.433	0.382	**0.196**	**0.432**	0.431
	w. format reward	**0.488**	**0.644**	**0.469**	**0.436**	**0.412**	0.187	0.403	**0.434**
	Outcome only (instruct)	**0.393**	**0.610**	0.397	0.370	**0.414**	0.146	0.368	**0.385**
	w. format reward	0.383	0.593	**0.399**	**0.376**	0.317	**0.151**	**0.371**	0.370
GRPO	Outcome only (base)	0.395	0.560	0.388	0.326	0.297	0.125	0.360	0.350
	w. format reward	**0.458**	**0.632**	**0.442**	**0.412**	**0.404**	**0.180**	**0.411**	**0.420**
	Outcome only (instruct)	**0.429**	**0.623**	**0.427**	**0.386**	**0.346**	**0.162**	**0.400**	**0.396**
	w. format reward	0.393	0.609	0.397	0.367	0.344	0.147	0.387	0.378
Qwen2.5-14B-Base/Instruct									
PPO	Outcome only (base)	0.486	0.676	**0.480**	**0.468**	**0.470**	**0.241**	**0.528**	**0.479**
	w. format reward	**0.499**	**0.680**	0.472	0.452	0.431	0.215	0.468	0.459
	Outcome only (instruct)	0.424	0.660	0.442	0.436	0.379	0.210	0.480	0.433
	w. format reward	**0.449**	**0.682**	**0.466**	**0.447**	**0.422**	**0.224**	**0.500**	**0.456**
GRPO	Outcome only (base)	0.415	0.680	0.488	0.451	0.461	0.230	0.508	0.462
	w. format reward	**0.500**	**0.693**	**0.500**	**0.481**	**0.488**	**0.261**	**0.516**	**0.491**
	Outcome only (instruct)	0.482	0.667	0.434	0.429	0.424	0.191	0.492	0.446
	w. format reward	**0.488**	**0.677**	**0.482**	**0.455**	**0.470**	**0.211**	**0.516**	**0.471**

Intermediate Retrieval Reward

Experimental Design. Building upon the outcome reward from [13, 58] and the format reward introduced in Section 4.1, we incorporate a retrieval correctness component, resulting in the following final reward function $r_\phi(x, y)$:

$$r_\phi(x,y) = \begin{cases} 1 & \text{if } a_{\text{pred}} = a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{True}, \\ 1 - \lambda_f & \text{if } a_{\text{pred}} = a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{False}, \\ \lambda_f + \lambda_r & \text{if } a_{\text{pred}} \neq a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{True} \wedge f_{\text{ret}}(y) = \text{True}, \\ \lambda_f & \text{if } a_{\text{pred}} \neq a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{True} \wedge f_{\text{ret}}(y) = \text{False}, \\ 0 & \text{if } a_{\text{pred}} \neq a_{\text{gold}} \wedge f_{\text{format}}(y) = \text{False}, \end{cases} \tag{4}$$

- f_ret(y) is checking if retrieval contains the correct answer substring
- Intermediate retrieval rewards complicate training without real gain
- They may distort the agent's natural reasoning–search trajectory.

Results for Intermediate Retrieval Reward

Table 2: Study of the intermediate retrieval reward. $\lambda_r = 0.1$. The best performance is set in bold. $^{\dagger}/^{\star}$ represents in-domain/out-domain datasets.

Methods		General QA			Multi-Hop QA				
		NQ†	TriviaQA*	PopQA*	HotpotQA†	2wiki*	Musique*	Bamboogle*	Avg.
Qwen2.5-3B-Base									
PPO	w.o. retrieval reward	**0.428**	**0.607**	**0.459**	**0.371**	**0.387**	**0.150**	**0.323**	**0.389**
	w. retrieval reward	0.405	0.567	0.407	0.326	0.330	0.104	0.242	0.340
GRPO	w.o. retrieval reward	0.429	0.602	**0.435**	0.372	**0.383**	**0.148**	0.307	0.382
	w. retrieval reward	**0.434**	**0.605**	0.433	**0.379**	0.378	0.142	**0.323**	**0.385**
Qwen2.5-7B-Base									
PPO	w.o. retrieval reward	**0.488**	**0.644**	**0.469**	0.436	**0.412**	**0.187**	**0.403**	**0.434**
	w. retrieval reward	0.472	0.629	0.452	0.436	0.402	0.180	0.363	0.419
GRPO	w.o. retrieval reward	**0.458**	**0.632**	0.442	0.412	**0.404**	**0.180**	**0.411**	**0.420**
	w. retrieval reward	0.453	0.628	**0.450**	**0.416**	0.375	0.164	0.387	0.410

Tool Variations

Table 5: Retriever generalization results across datasets and test retrievers. (Qwen2.5-7B-Base, PPO)

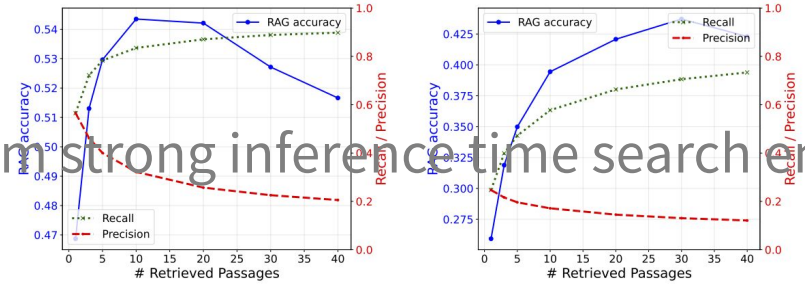
Train / Test Search Engine	BM25			E5 (HNSW)			E5 (Exact)			Google Search		
	Bambg	GPQA	SimpleQA	Bambg	GPQA	SimpleQA	Bambg	GPQA	SimpleQA	Bambg	GPQA	SimpleQA
BM25	0.280	0.273	0.243	0.432	0.293	0.159	0.424	0.323	0.259	0.496	0.313	0.540
E5 (HNSW)	0.240	0.298	0.270	0.400	0.288	0.169	0.440	0.273	0.254	0.528	0.333	0.603
E5 (Exact)	0.312	0.313	0.249	0.400	0.298	0.196	0.424	0.288	0.265	0.560	0.293	0.603
Average	0.277	0.295	0.254	0.411	0.293	0.175	0.429	0.295	0.259	0.528	0.313	0.582

- Better Retriever in training, better inference time (not exactly?)
- Tool do impact training!

Patterns under different search engines

- 🛑 Random (Noise): Avoids search
- 🔁 BM25 (Sparse): Spams search calls — Low retrieval quality
  - Resonates with our retrieval course
- ⚖️ Dense (E5 HNSW / Exact): Balanced, strategic search

Still benefits from strong inference time search engine!



(a) Retrieval with e5 retriever (b) Retrieval with BM25 retriever

Another problem with RL training for agentic task:

- Environment is **noisy**, slow, and **expensive**!
- For example, `search`: (brave api)

```
BraveSearchLoader(
    query="obama middle name", api_key=api_key, search_kwargs={"count": 3}
)
```

```
[{'title': "What's up with Obama's middle name? - Quora",
  'link': 'https://www.quora.com/Whats-up-with-Obamas-middle-name'},
 {'title': 'Barack Obama | Biography, Parents, Education, Presidency, Books, ...',
  'link': 'https://www.britannica.com/biography/Barack-Obama'}
 {'title': "Obama's Middle Name -- My Last Name -- is 'Hussein.' So?",
  'link': 'https://www.cair.com/cair_in_the_news/obamas-middle-name-my-last-name-is-husse
```

And the price for search is …

Free AI

$0

Get familiar with the API

- 1 query/second
- Up to 2,000 queries/month

See all Free plans

Base AI

$5.00

per 1,000 requests

- Up to 20 queries/second
- Up to 20M queries/month
- Rights to use in AI apps

See all Base plans

Pro AI

$9.00

per 1,000 requests

- Up to 50 queries/second
- Unlimited queries/month
- Rights to use in AI apps

See all Pro plans

And the price for search is …

Using Brave, **evaluating** on MMLU (assuming 1 search/question) would be $145. This is really high!

And the price for search is …

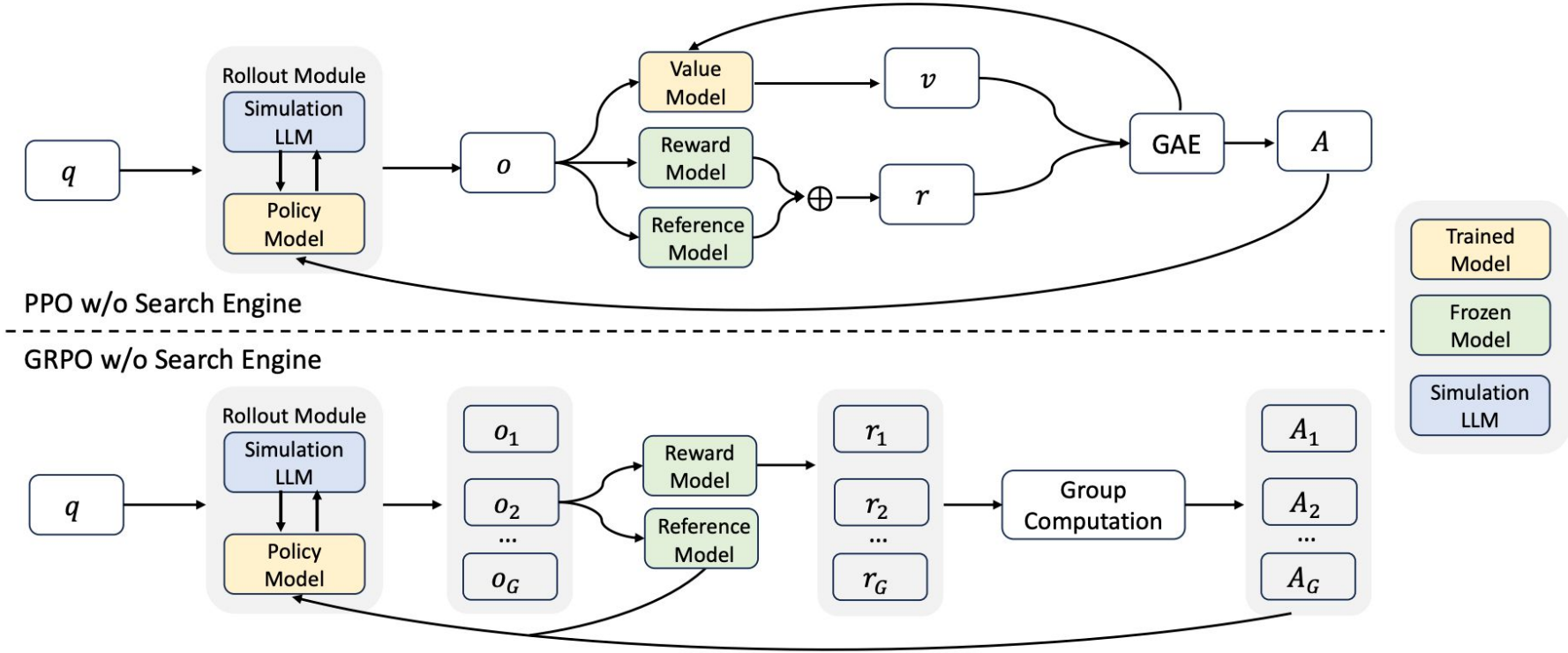
Using Brave, **evaluating** on MMLU (assuming 1 search/question) would be $145. This is really high!

ZeroSearch: Incentivize the Search Capability of LLMs without Searching

---

Hao Sun, Zile Qiao*, Jiayan Guo*, Xuanbo Fan, Yingyan Hou
Yong Jiang, Pengjun Xie, Yan Zhang*, Fei Huang, Jingren Zhou

TLDR for ZeroSearch: let's teach LLMs to search, without out searching!



let's teach LLMs to search, without out searching!

Recipe:

1. Replace search engine with an LLM 😅
    a. How are we sure LLM's a faithful proxy for a search engine? SFT
2. Increase search engine LLM noise overtime (curriculum learning based rollouts, a good benefit from LLM simulation).
3. Standard PPO/GRPO/REINFORCE

Result: *ZeroSearch Performs as good as algorithm with real search engines!*

Method	Single-Hop QA			Multi-Hop QA				
	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	Musique	Bamboogle	Avg.
Qwen-2.5-7B-Base/Instruct								
Direct Answer	11.60	35.60	1.20	16.40	22.20	4.80	14.40	15.17
CoT	12.80	35.60	3.80	16.20	22.60	6.60	24.00	17.37
RAG	27.40	58.20	17.80	25.80	23.20	9.40	16.80	25.51
RA-Agent	21.20	40.20	8.80	19.60	19.60	7.60	28.00	20.71
Search-o1	19.40	40.60	11.40	17.00	27.00	8.60	30.40	22.06
R1-base	27.60	47.40	27.40	21.00	29.20	9.80	27.78	27.17
R1-instruct	27.00	45.80	24.20	21.60	27.80	8.40	25.00	25.69
Search-R1-base	43.40	61.40	54.60	31.20	**37.20**	18.20	30.56	39.51
Search-R1-inst	42.40	63.40	51.60	32.80	33.20	17.40	26.39	38.17
ZeroSearch-base	42.40	**66.40**	**60.40**	32.00	34.00	18.00	**33.33**	**40.93**
ZeroSearch-inst	**43.60**	65.20	48.80	**34.60**	35.20	**18.40**	27.78	39.08
Qwen-2.5-3B-Base/Instruct								
Direct Answer	12.40	30.60	5.60	16.00	19.20	4.40	16.80	15.00
CoT	15.00	33.60	3.60	16.20	18.00	3.60	12.80	14.69
RAG	31.60	58.00	15.20	24.20	23.20	8.20	15.20	25.09
RA-Agent	15.20	28.40	6.60	12.60	16.60	2.60	13.60	13.66
Search-o1	16.60	31.00	8.20	14.80	22.40	5.20	**22.40**	17.23
R1-base	14.20	34.80	20.80	19.60	28.40	6.40	5.56	18.54
R1-instruct	19.80	33.00	19.40	19.40	26.40	4.40	11.11	19.07
Search-R1-base	40.60	60.00	44.20	29.20	32.00	11.20	12.50	32.81
Search-R1-inst	35.80	55.80	26.00	33.20	26.00	7.60	12.50	28.13
ZeroSearch-base	**43.00**	**61.60**	41.40	**33.80**	**34.60**	**13.00**	13.89	**34.47**
ZeroSearch-inst	41.40	57.40	**44.80**	27.40	30.00	9.80	11.11	31.70

Results: *curriculum-based rollouts helps, too*

Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	Musique	Bamboogle	Avg.
Qwen-2.5-3B-Base								
Curriculum	43.00	61.60	41.40	33.80	34.60	13.00	13.89	34.47
Random	41.40	59.00	44.20	29.00	31.40	10.60	12.50	32.59
LLaMA-3.2-3B-Base								
Curriculum	43.40	63.80	48.40	32.20	35.60	13.80	15.28	36.07
Random	40.40	62.80	49.60	29.80	36.00	14.20	11.11	34.84

Table 6: Curriculum Rollout Study. We compare the performance of standard and random rollout settings using the Qwen-2.5-3B-Base and LLaMA-3.2-3B-Base as the policy models.

Takeaway for ZeroSearch:

On some tasks, designing smart simulators for agentic tools significantly reduces the cost and quality of RL training.

Summary

Instruct-tuned models are harder to RL for search agents. Format reward has positive impact on training. Intermediate retrieval reward has negative impacts.

For searches, simulating a search engine with a fine-tuned LLM helps with RL, too!

More followup question

- Can we ignore the format reward if we are applying structural decoding in the end?
- Did intermediate reward fail because intermediate reward does not work or because they set it wrongly? How to add good intermediate reward?
- Can we use a retriever as the search engine simulator? What's the pros and cons of using an LLM instead?
- Besides `search`, what other tools can we simulate?

Small-group discussion

Group	Lead				
1	**Sanjay Adhikesaven**	Jongho Park	Kaiwen Hu	Kalvin Chang	Xutao Ma
2	**Junyi Zhang**	Prasann Singhal	Harman Singh	Shangyin Tan	Donghyun Lee
3	**Huanzhi Mao**	Ryan Wang	Yuezhou Hu	Hanchen Li	
4	**Charlie Ruan**	Sidhika Balachandar	Yichuan Wang	Juno Kim	
5	**Bhavya Chopra & Dennis Jacob**	Sangdae Nam	Dongwei Lyu	Colin Wang	

Write a short review at #class-discussion