DeepRetrieval: Hacking Real Search Engines and Retrievers with Large Language Models via Reinforcement Learning

Pengcheng Jiang*, Jiacheng Lin*, Lang Cao, Runchu Tian, SeongKu Kang†, Zifeng Wang, Jimeng Sun, and Jiawei Han

University of Illinois Urbana-Champaign †Korea University

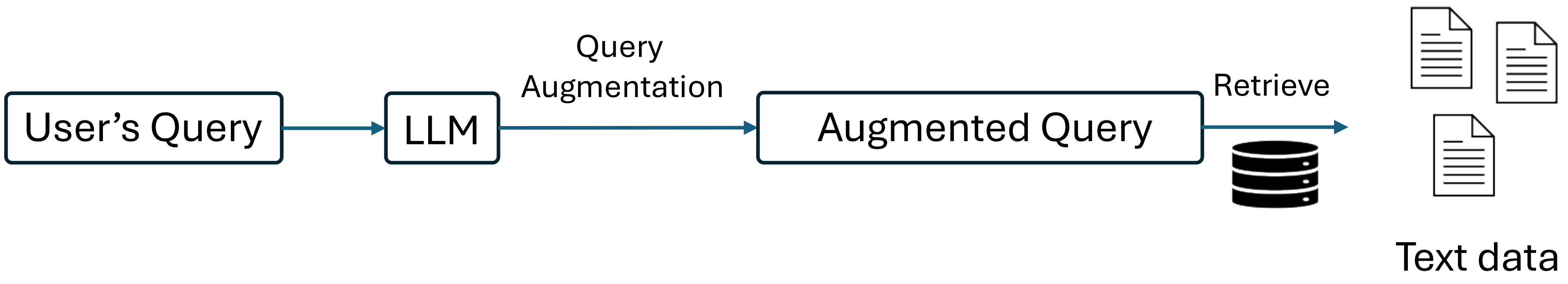
Presenter: Pengcheng (Patrick) Jiang

Overview

- Background
- Method – DeepRetrieval
- Experiments
- Discussions
- Conclusion

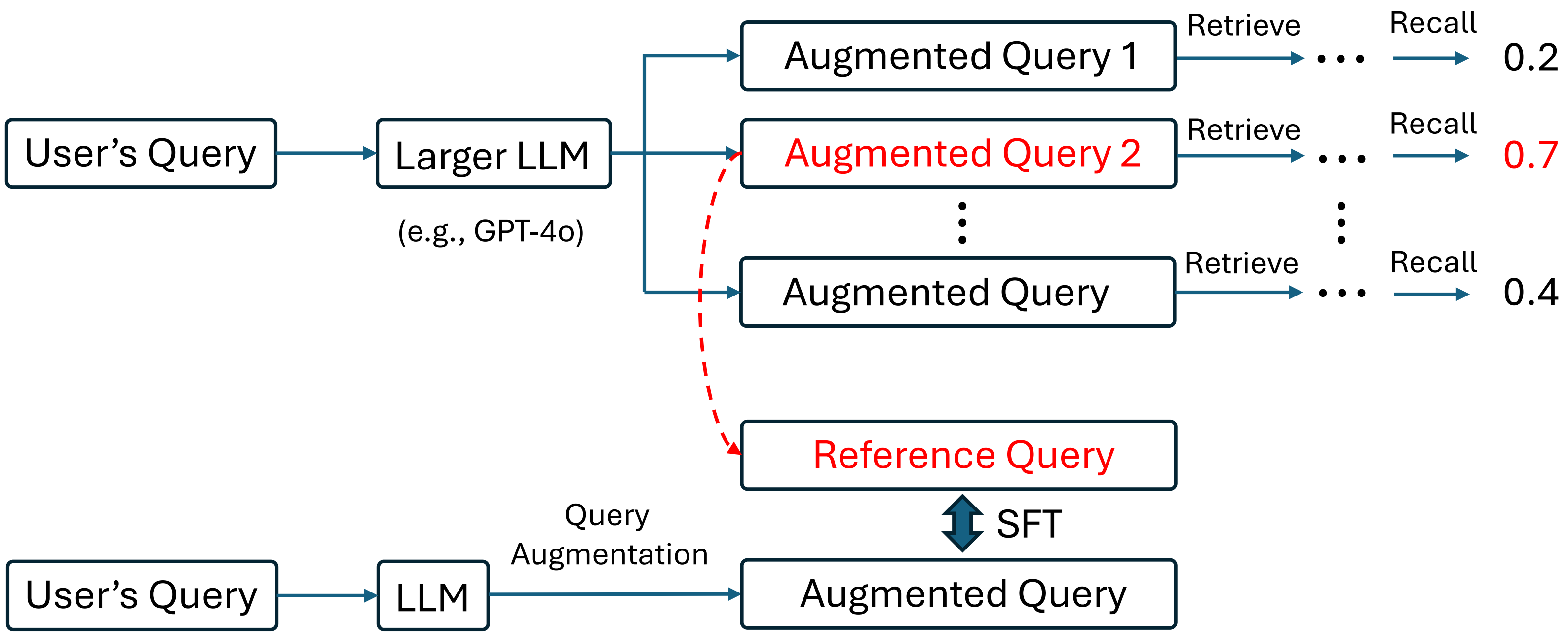
Background

- Information retrieval systems often struggle with the semantic gap between user queries and relevant documents.
- **Query Augmentation** bridges this gap by reformulating queries to better match relevant content:



Background

Previous Approaches (Distillation from Larger LLMs):



- Costly and highly rely on the quality of reference query (often suboptimal)

Background

Previous Approaches (Distillation from Larger LLMs):

❌ **Dependence on Reference Queries:**
Distillation relies on expensive, manually curated reference queries (often from large LLMs like GPT-4o). These queries may not be optimal for the target retrieval task.

📉 **Indirect Optimization:**
Distilled models learn to mimic query form—not retrieval effectiveness. They optimize for similarity, not metrics like Recall@K or NDCG.

💰 **Cost and Bias:**
Generating supervision data is costly and time-consuming. Distilled models may inherit biases from the teacher, limiting generalization.

🔒 **Limited Exploration:**
SFT models can’t explore beyond fixed training data, making them prone to local minima and less adaptable to new tasks.

Background

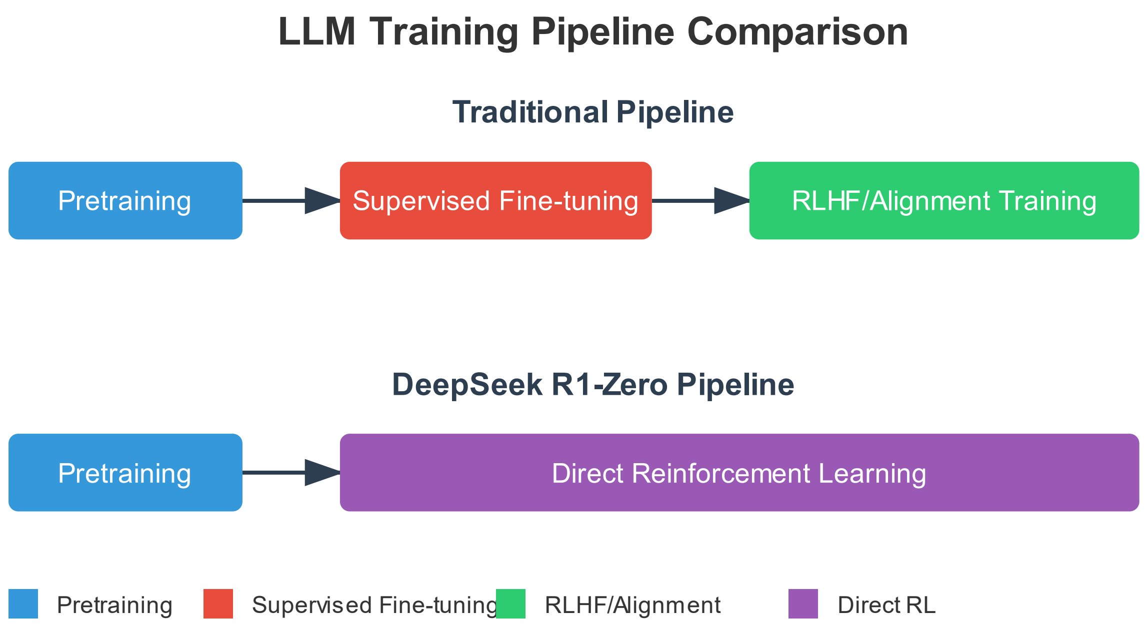
Can we skip reference queries and still train an effective query generator?

Yes, DeepSeek-R1-Zero[1] was trained in this way.

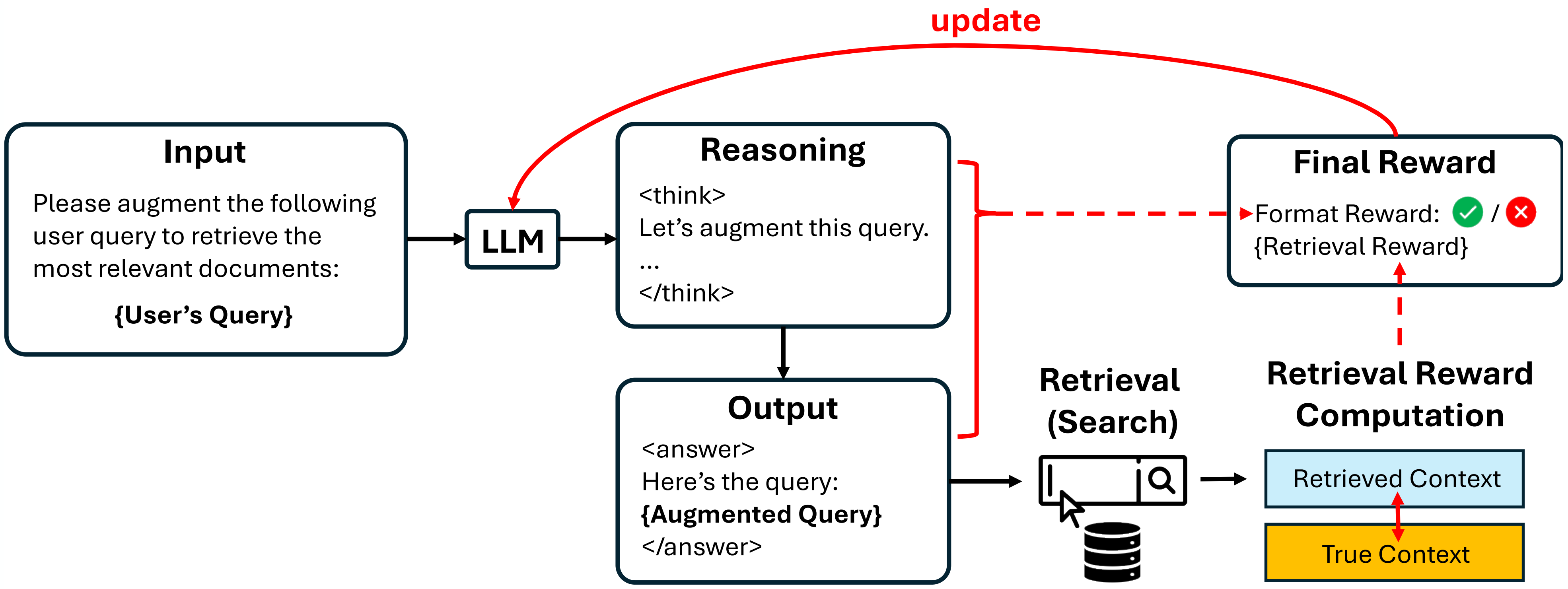
[1] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Background

- Traditional LLM training typically relies heavily on supervised fine-tuning with human-labeled data
- DeepSeek R1-Zero starts with just the base model and applies RL directly, learning reasoning/generation capabilities from scratch through trial-and-error



DeepRetrieval



DeepRetrieval learns to generate queries through trial-and-error, guided by real retrieval outcomes from live systems.

DeepRetrieval

	DeepRetrieval	Distillation-Based Methods
Training Signal	⚡ *Direct reward from retrieval outcome*	📋 *Matches teacher or annotated reference queries*
Supervision Needed	❌ *No supervision or labeled queries*	✅ *Requires supervised data or teacher outputs*
Adaptability	✅ Retriever-agnostic and domain-flexible	❌ Needs new data or distillation per domain
Cost Efficiency	💰 Low-cost (no human-in-the-loop)	💸 High-cost due to human annotation and large LLMs
Model Size Efficiency	🧠 Strong results with small (3B) models	🏗️ Typically relies on larger teacher models

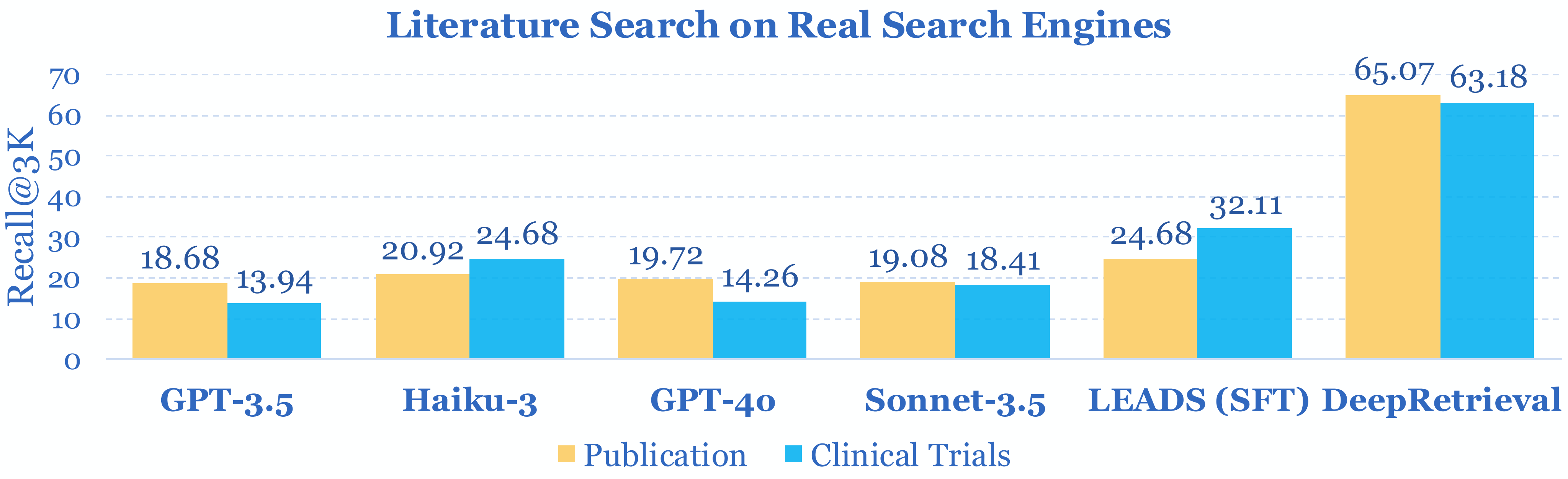
Experiments

We tested **DeepRetrieval** on four retrieval tasks:

Task	Description	Retriever Type	Metric	Examples
1. Literature Search	Retrieve scientific papers	Real Search Engines	Recall@3K	PubMed, ClinicalTrials.gov
2. Evidence-Seeking	Retrieve answer-containing passages for open QA	Sparse (BM25)	Hits@1/5/20 (H@N)	Natural Questions, TriviaQA, SQuAD
3. Classic IR	Improve performance on standard sparse/dense retrieval benchmarks	BM25 / Dense	NDCG@10	MS MARCO, FEVER, HotpotQA, SciFact
4. SQL Search	Generate SQL queries to retrieve structured records	Structured SQL backend	Execution Accuracy	Spider, BIRD

Experiments – Task 1: Literature Search

Task Definition: Search scientific papers/trials with search engines
Metric: Recall@3K (How many ground truth papers are retrieved among the top-3k retrieved documents?)



DeepRetrieval-3B's **65.07%** vs. Previous SOTA (SFT)'s **24.68%** on PubMed Search API
DeepRetrieval-3B's **63.18%** vs. Previous SOTA (SFT)'s **32.11%** on ClinicalTrials.gov Search API

Experiments – Task 2: Evidence-Seeking

Task Definition: Given a question, looking for the answer span in the retrieved documents.
Metric: Hits@N (Is there an answer span in the top-N retrieved documents?)

	Evidence-Seeking Retrieval								
	NQ			TriviaQA			SQuAD		
	H@1	H@5	H@20	H@1	H@5	H@20	H@1	H@5	H@20
Original Query	21.9	43.8	63.0	48.2	66.3	76.4	**36.5**	**57.4**	**71.1**
GPT-3.5	24.3	46.0	63.9	45.8	64.3	74.2	31.6	52.4	66.6
w/o reasoning	25.2	47.5	66.3	47.5	66.8	76.7	33.9	54.9	69.5
GPT-4o	**35.8**	**57.5**	72.2	**59.6**	**73.3**	**80.5**	30.4	49.9	64.4
w/o reasoning	29.1	56.2	69.3	53.4	70.1	78.7	33.0	52.2	66.7
Claude-3-Haiku	26.2	48.6	66.4	48.8	67.9	77.7	33.3	54.1	68.4
w/o reasoning	25.0	48.1	65.5	49.0	67.7	77.3	33.2	54.3	68.8
Claude-3.5-Sonnet	35.7	**57.1**	**72.5**	57.1	71.7	79.7	28.5	48.1	63.5
w/o reasoning	**37.2**	56.9	**72.7**	**60.8**	**73.8**	**80.6**	30.3	49.8	64.7
Mistral$_{7B\text{-Inst}}$	26.9	48.8	66.0	50.0	66.7	75.9	27.7	46.6	61.6
LEADS$_{7B}$ (SFT)	-	-	-	-	-	-	-	-	-
Qwen2.5$_{3B\text{-Inst}}$	25.0	45.8	63.4	44.4	61.2	70.9	28.4	46.4	61.3
w/o reasoning	23.8	45.3	64.0	46.0	64.4	74.2	32.3	52.8	66.8
DeepRetrieval$_{3B}$	**35.5**	**57.5**	**72.7**	**58.4**	**73.2**	**80.6**	**38.5**	**59.4**	**72.9**
w/o reasoning	26.9	48.8	66.9	52.0	69.4	77.7	**37.8**	**58.0**	**72.5**

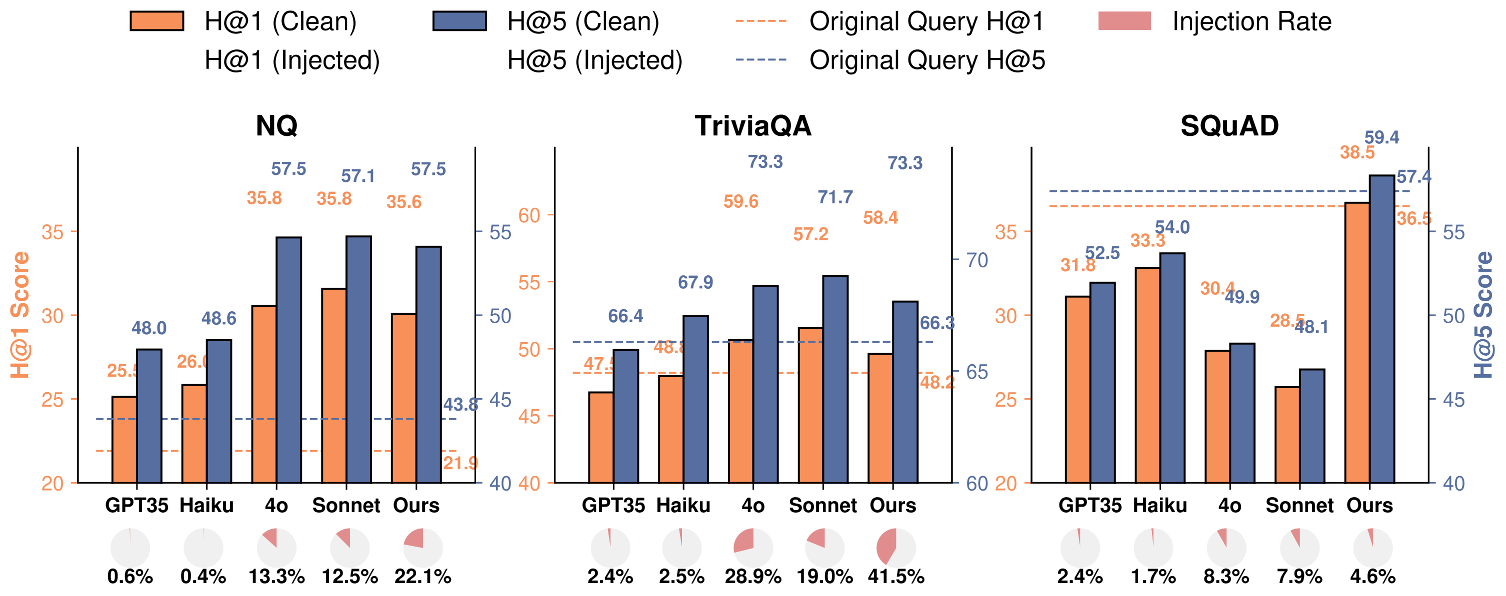
1. DeepRetrieval-3B achieved comparable performance with GPT-4o and Claude-3.5 on evidence-seeking.

2. Reasoning matters for DeepRetrieval

For this task, what if the model inject its own knowledge into the query, i.e., put the answer into the query?

Experiments – Task 2: Evidence-Seeking

Knowledge Injection Study for Evidence-Seeking Retrieval



- **DeepRetrieval** learns **adaptive injection strategies**, injecting more knowledge where helpful (e.g., 41.5% in TriviaQA), and minimizing injection where unnecessary (e.g., 4.6% in SQuAD).
- This study underscores the importance of dataset-specific strategies in query generation and highlights the **adaptive reasoning capability** learned via RL

Experiments – Task 3: Classic IR

Task Definition: Given a query, search relevant documents.
Metric: NDCG@10 (rewards retrieving relevant documents early in the top 10; higher is better.)

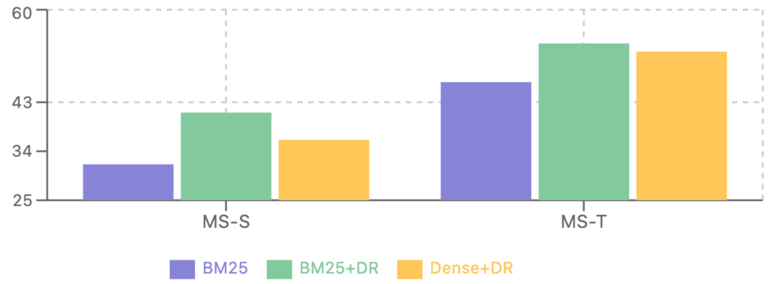
Methods	NFCorpus		FEVER		HotpotQA		SciFact		MS-Beir		MS-H		MS-S		MS-T	
	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S	D
Base Retriever																
BM25 / Dense	14.7	37.0	44.2	82.5	61.1	70.0	57.3	64.5	44.8	70.4	32.5	32.4	38.8	31.1	51.3	49.8
Zero-shot Query Gen (w/o reasoning)																
GPT-3.5	30.1	33.0	55.0	64.5	58.1	54.1	66.4	58.9	43.1	69.1	28.8	31.7	35.4	33.0	48.8	50.6
GPT-4o	31.8	33.6	59.1	72.2	58.0	70.2	66.4	65.5	47.8	68.5	21.8	27.8	28.5	28.5	43.4	48.2
Claude-3-Haiku	31.3	26.6	43.5	73.6	49.3	62.1	65.1	63.0	38.7	69.0	29.0	31.1	34.7	33.5	48.9	52.0
Claude-3.5-Sonnet	31.6	35.2	54.8	71.0	46.4	58.7	68.4	**68.2**	45.3	61.1	21.3	21.9	27.5	24.2	39.7	43.7
Qwen2.5-3B-Inst	20.9	33.9	55.5	71.4	51.7	64.7	65.1	62.9	31.3	66.9	26.3	30.5	31.6	33.0	46.7	49.2
Zero-shot Query Gen (w/ reasoning)																
GPT-3.5	32.1	32.8	55.4	63.9	54.4	54.1	65.1	61.9	39.3	63.1	20.8	28.8	25.7	30.0	40.5	47.6
GPT-4o	30.7	34.0	53.9	73.8	56.4	**71.8**	65.0	63.8	39.4	66.6	20.8	20.6	26.4	23.0	42.0	44.2
Claude-3-Haiku	29.6	36.0	59.9	74.9	55.6	65.2	**68.9**	65.6	44.7	67.2	16.5	24.2	22.4	25.4	37.6	43.5
Claude-3.5-Sonnet	30.7	36.4	55.7	75.8	55.2	67.6	68.7	65.9	47.8	63.9	18.6	18.5	27.3	23.6	41.6	43.8
Qwen2.5-3B-Inst	29.2	32.4	46.7	69.9	48.3	62.0	63.8	62.1	23.0	60.0	22.7	24.3	25.9	27.6	43.0	45.0
Ours (DeepRetrieval-3B)																
+BM25 / +Dense	**34.0**	**37.7**	**66.4**	**84.1**	**63.1**	70.1	64.6	66.4	**53.1**	**70.4**	**34.7**	**32.5**	**41.1**	**36.1**	**53.8**	**52.3**

We use BM25 as the base sparse retriever for all the datasets, while using E5-Large as the base dense retriever for SciFact, use BGE-base-en-v1.5 for HotpotQA, FEVER, NFCorpus, and MS-Beir, and use vanilla Contriever for MS MARCO domain-specific (MS-H: health, MS-S: science, MS-T: technology) subsets.

Experiments – Task 3: Classic IR

Task Definition: Given a query, search relevant documents.
Metric: NDCG@10 (rewards retrieving relevant documents early in the top 10; higher is better.)

Methods	NFCorpus		FEVER		HotpotQA		SciFact		MS-Beir		MS-H		MS-S		MS-T	
	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S	D
Base Retriever																
BM25 / Dense	14.7	37.0	44.2	82.5	61.1	70.0	57.3	64.5	44.8	70.4	32.5	32.4	38.8	31.1	51.3	49.8
Zero-shot Query Gen (w/o reasoning)																
GPT-3.5	30.1	33.0	55.0	64.5	58.1	54.1	66.4	58.9	43.1	69.1	28.8	31.7	35.4	33.0	48.8	50.6
GPT-4o	31.8	33.6	59.1	72.2	58.0	70.2	66.4	65.5	47.8	68.5	21.8	27.8	28.5	28.5	43.4	48.2
Claude-3-Haiku	31.3	26.6	43.5	73.6	49.3	62.1	65.1	63.0	38.7	69.0	29.0	31.1	34.7	33.5	48.9	52.0
Claude-3.5-Sonnet	31.6	35.2	54.8	71.0	46.4	58.7	68.4	68.2	45.3	61.1	21.3	21.9	27.5	24.2	39.7	43.7
Qwen2.5-3B-Inst	20.9	33.9	55.5	71.4	51.7	64.7	65.1	62.9	31.3	66.9	26.3	30.5	31.6	33.0	46.7	49.2
Zero-shot Query Gen (w/ reasoning)																
GPT-3.5	32.1	32.8	55.4	63.9	54.4	54.1	65.1	61.9	39.3	63.1	20.8	28.8	25.7	30.0	40.5	47.6
GPT-4o	30.7	34.0	53.9	73.8	56.4	71.8	65.0	63.8	39.4	66.6	20.8	20.6	26.4	23.0	42.0	44.2
Claude-3-Haiku	29.6	36.0	59.9	74.9	55.6	65.2	68.9	65.6	44.7	67.2	16.5	24.2	22.4	25.4	37.6	43.5
Claude-3.5-Sonnet	30.7	36.4	55.7	75.8	55.2	67.6	68.7	65.9	47.8	63.9	18.6	18.5	27.3	23.6	41.6	43.8
Qwen2.5-3B-Inst	29.2	32.4	46.7	69.9	48.3	62.0	63.8	62.1	23.0	60.0	22.7	24.3	25.9	27.6	43.0	45.0
Ours (DeepRetrieval-3B)																
+BM25 / +Dense	34.0	37.7	66.4	84.1	63.1	70.1	64.6	66.4	53.1	70.4	34.7	32.5	41.1	36.1	53.8	52.3



34× faster with better performance

Findings:

(1) DeepRetrieval is more effective to boost sparse retrieval performance

(2) When dense retrievers have already learned data distribution in the training set, the room left with query-rewriting is limited

(3) For unseen data (MS-H, MS-S, MS-T), DeepRetrieval+BM25 outperforms dense retriever and its combination w/ DeepRetrieval, with 34x faster retrieval speed

Experiments – Task 4: SQL Search

Task Definition: Given a natural language question, generate a SQL query to retrieve the correct answer from a database.
Metric: Execution Accuracy — percentage of generated SQL queries that produce the correct answer when executed.

Findings:

(1) DeepRetrieval outperforms GPT-4o and Claude-3.5 on Text-to-SQL task

(2) Coder (base model pre-trained on code) performs better

(3) RL from scratch outperforms SFT

(4) "Cold start" works better for general-purpose base model (Qwen-2.5)

(5) Reasoning works better for coder model (Qwen2.5-Coder)

Methods	BIRD	Spider
Zero-shot (w/o reasoning)		
GPT-3.5	46.22	67.02
GPT-4o	55.35	73.50
Claude-3-Haiku	43.16	64.88
Claude-3.5-Sonnet	50.46	60.74
Qwen2.5$_{3B\text{-}Inst}$	29.66	52.90
Qwen2.5-Coder$_{3B\text{-}Inst}$	30.77	50.97
Qwen2.5-Coder$_{7B\text{-}Inst}$	45.24	64.89
Zero-shot (w/ reasoning)		
GPT-3.5	44.07	64.88
GPT-4o	55.93	73.40
Claude-3-Haiku	43.81	67.44
Claude-3.5-Sonnet	50.65	66.05
Qwen2.5$_{3B\text{-}Inst}$	30.83	55.13
Qwen2.5-Coder$_{3B\text{-}Inst}$	33.57	54.45
Qwen2.5-Coder$_{7B\text{-}Inst}$	45.57	67.70
SFT (w/o reasoning)		
Qwen2.5$_{3B\text{-}Inst}$	33.77	56.67
Qwen2.5-Coder$_{3B\text{-}Inst}$	39.77	58.61
Qwen2.5-Coder$_{7B\text{-}Inst}$	44.07	65.96
SFT (w/ reasoning)		
Qwen2.5$_{3B\text{-}Inst}$	37.29	60.93
Qwen2.5-Coder$_{3B\text{-}Inst}$	46.15	66.92
Qwen2.5-Coder$_{7B\text{-}Inst}$	50.65	70.99
Ours		
DeepRetrieval$_{3B\text{-}Base}$	41.40	68.79
w/ cold start	44.00	70.33
w/o reasoning	39.57	70.24
DeepRetrieval$_{3B\text{-}Coder}$	49.02	74.85
w/ cold start	50.52	74.34
w/o reasoning	47.00	73.59
DeepRetrieval$_{7B\text{-}Coder}$	**56.00**	**76.01**

Discussions

Why RL >>> SFT?

- **Direct Optimization**: RL optimizes retrieval metrics directly rather than mimicking reference queries
- **Exploration Advantage**: RL explores query space through trial-and-error, discovering patterns ***human experts*** might miss
  For example:

```
P: Patients undergoing perioperative
procedures, I: Desmopressin
administration, C: Standard care without
desmopressin, O: Minimising perioperative
allogeneic blood transfusion
```

DeepRetrieval →

```
((DDAVP) AND (Perioperative
Procedures OR Blood Transfusion OR
Desmopressin OR Anticoagulant)) AND
(Randomized Controlled Trial)
```

- **Task Adaptability**: RL performs consistently well across scenarios with varying levels of ground truth availability

They are also complementary : SFT can provide strong initialization for RL when model lacks domain capabilities (SQL coding)

Discussions

Think / Query Length Analysis

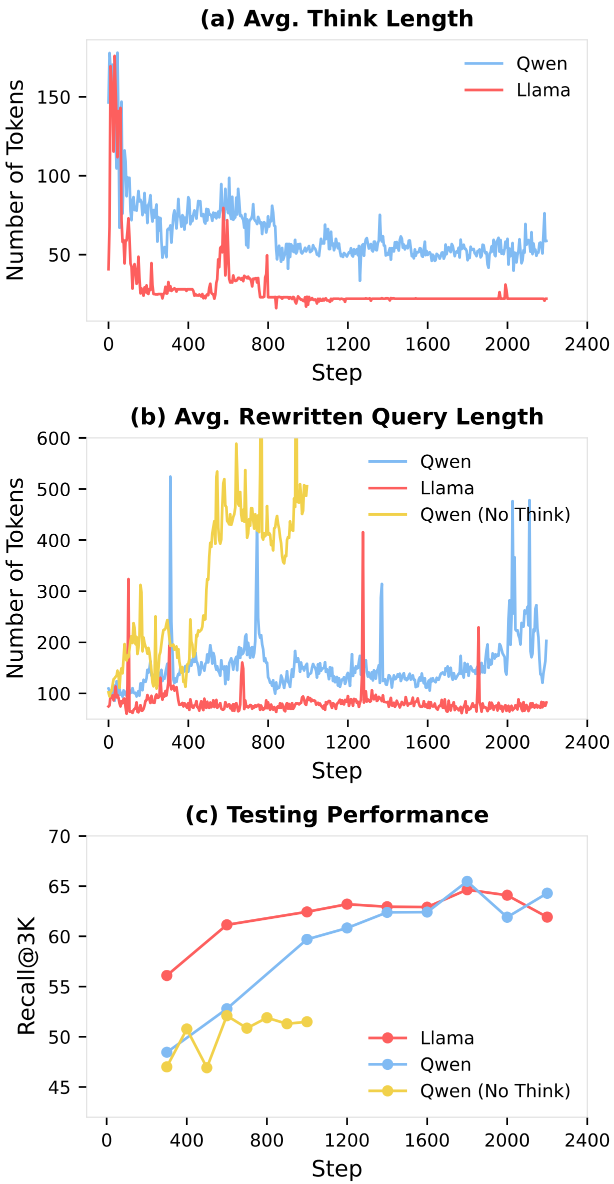
Reasoning Evolution: Unlike tasks (e.g. math) requiring long reasoning chains, reasoning length decreases over time as models internalize effective strategies

- Reasoning/Think in DeepRetrieval acts as a strategy seeker

Different Strategies leading to similar performance: Models discover distinct approaches (Qwen favors longer queries, LLaMA produces shorter ones), yet achieve comparable recall (~65%) - demonstrating multiple valid paths to high performance

Without Reasoning: Models fall into local minima of query verbosity (yellow line) with lower performance (~52% vs ~65% recall)

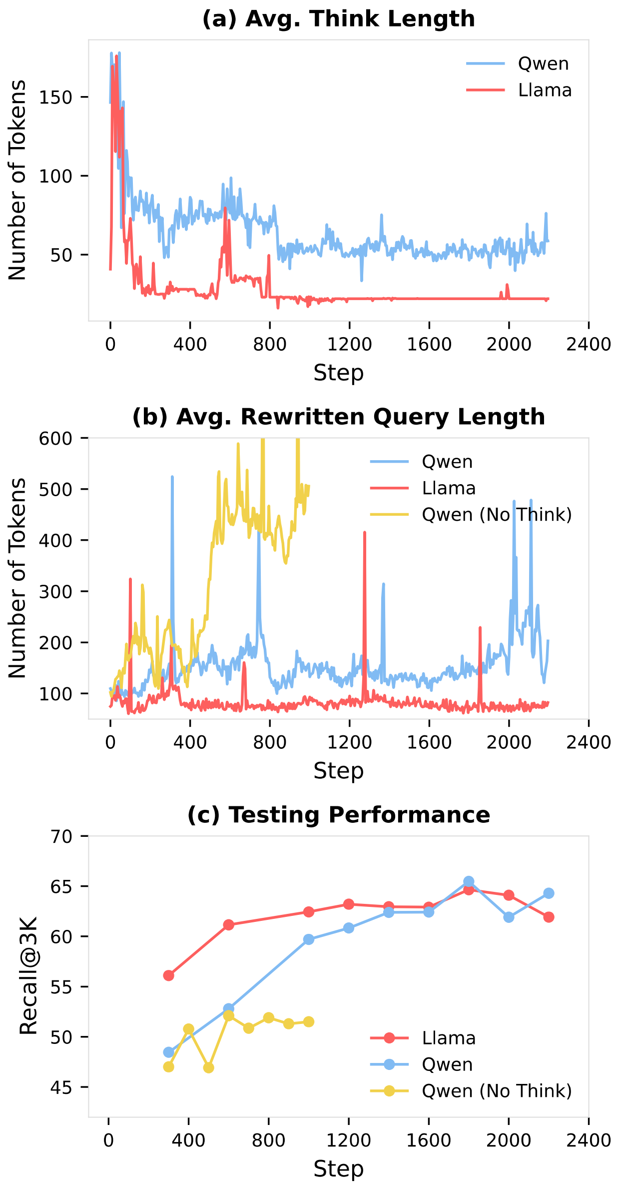
$\text{DeepRetrieval}_{3B}$	65.07	63.18
w/o reasoning	**51.90**	**53.31**

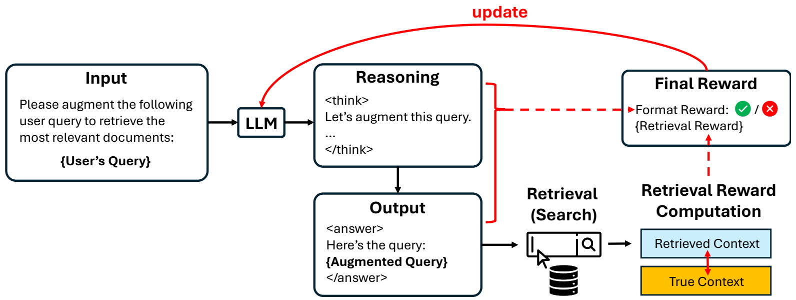


Discussions

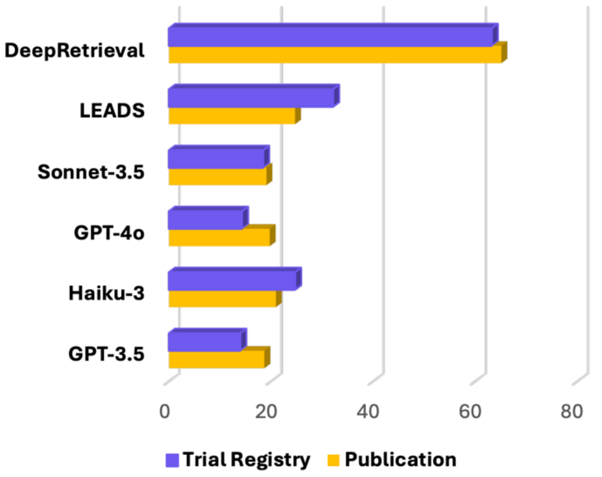
Think / Query Length Analysis

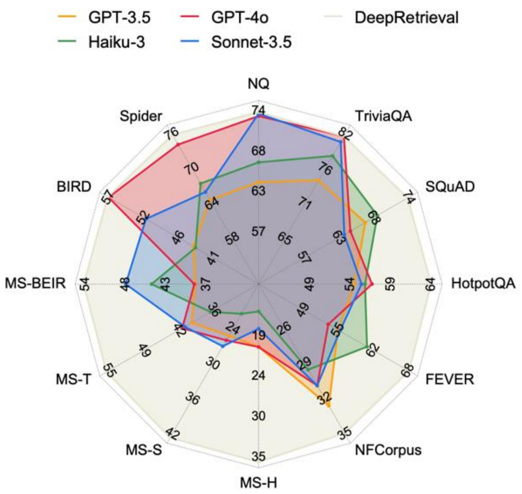
	Step 400	Step 600
Qwen (No Think)	Query: (HIV infection OR human immunodeficiency virus infection) AND (antiretroviral therapy OR antiretroviral regimens OR antiretrovirals) AND (nevirapine OR other antiretroviral regimens) AND (efficacy OR safety OR clinical outcomes OR pharmacokinetics OR pharmacodynamics OR side effects OR treatment effects OR treatment outcomes OR treatment efficacy OR treatment safety OR treatment effectiveness OR treatment safety OR treatment side effects OR treatment outcomes OR treatment assessment OR treatment evaluation OR treatment monitoring OR treatment follow-up OR treatment compliance OR treatment adherence OR treatment persistence OR treatment persistence OR treatment persistence OR treatment persistence OR treatment persistence OR treatment persistence)	Query: (testosterone supplementation OR testosterone therapy) AND (ovarian insufficiency OR poor ovarian response OR ovarian dysfunction OR ovarian failure OR ovarian suppression OR ovarian inertia OR ovarian underachievement OR ovarian decline OR ovarian衰退 OR ovarian衰竭 OR ovarian萎缩 … **(repeat "ovarian萎缩" 63 times more)**
Qwen	Query: ((dementia OR Alzheimer's disease) AND (exercise OR physical exercise OR physical therapy OR physical activity)) AND ((clinical trial OR randomized controlled trial) OR (comparative study OR controlled clinical trial) OR (exercise intervention OR physical activity intervention) OR (no exercise OR standard care) OR (cognitive function OR physical health OR overall well-being) OR (clinical trial registration OR controlled clinical trial))	Query: ((asthma OR asthma management) AND (relaxation therapy OR mindfulness therapy OR biofeedback OR cognitive behavioral therapy OR cognitive behavioral therapy OR relaxation technique OR mindfulness technique OR cognitive behavioral intervention OR cognitive behavioral intervention OR asthma management program OR asthma control program OR asthma control therapy OR asthma control treatment OR asthma control technique OR asthma control intervention OR asthma control therapy)) AND ((clinical trial) OR (randomized controlled trial) OR (systematic review) OR (meta-analysis))
LLaMA	Query: ((Collagenase OR Collagenase enzyme OR Deglycerolipase) AND (Wound treatment OR Surgical debridement OR Ulcer treatment))	Query: ((Total Knee Arthroplasty Trial OR Total Knee Arthroplasty Surgery) AND (Drainage OR Antiotics Trial OR Surgical Drainage Trial OR Postoperative Drains Trial))





Conclusion





- **DeepRetrieval** introduces a new paradigm: training LLMs for query generation via direct **reinforcement learning** from real retrieval outcomes—without relying on reference queries.
- Our method **doubles recall achieved by previous SOTA** on real search engines, outperforms **GPT-4o and Claude-3.5** in evidence-seeking and SQL tasks, and classic IR benchmarks.
- Unlike distillation-based & SFT methods, DeepRetrieval learns **adaptive reasoning strategies**, demonstrating strong **generalization** and **efficiency** with just 3B parameters.
- This work highlights RL as a powerful and general solution for **bridging the query-retrieval gap** in real-world information access.

Paper: https://arxiv.org/pdf/2503.00223
Code: https://github.com/pat-jj/DeepRetrieval
Models: https://huggingface.co/DeepRetrieval

Thank you!

Patrick Jiang