Retrieval-Augmented Generation (RAG) & Tool-use LLMs

CS 5624: Natural Language Processing

Spring 2025

https://tuvllms.github.io/nlp-spring-2025

Tu Vu

VIRGINIA TECH

Special thanks to Andrew Drozdov for his insights and paper recommendations

Logistics

- Homework 2 due **5/5**
- Final project presentations **5/6**
- Final project report due **5/9**
- Final grades due **5/16**

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

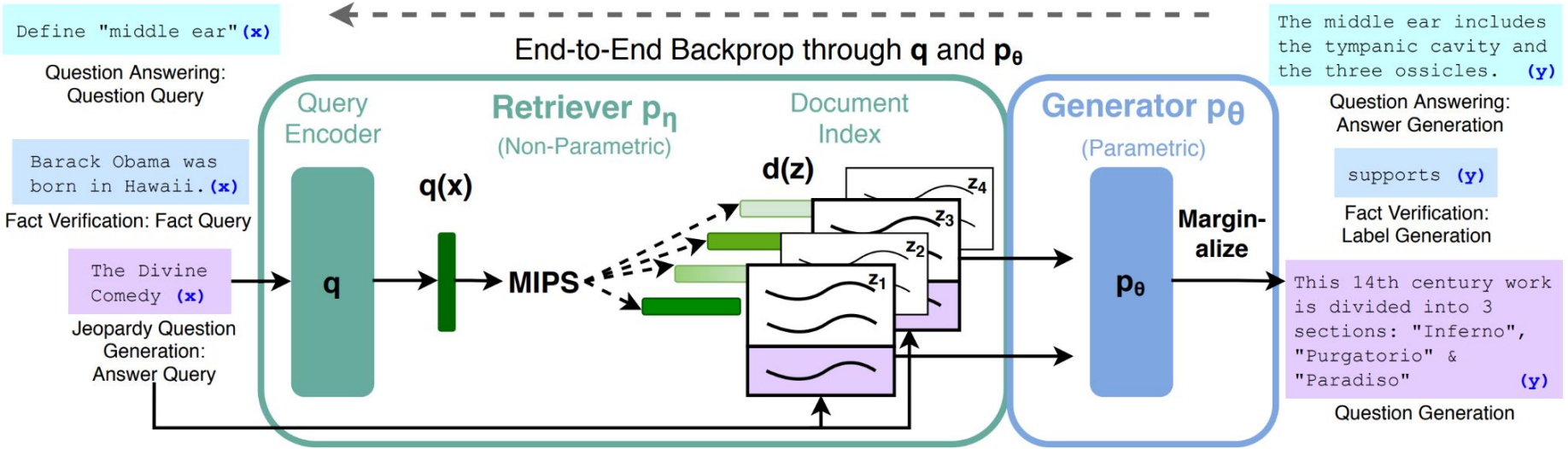
Patrick Lewis[†‡], Ethan Perez[⋆],

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

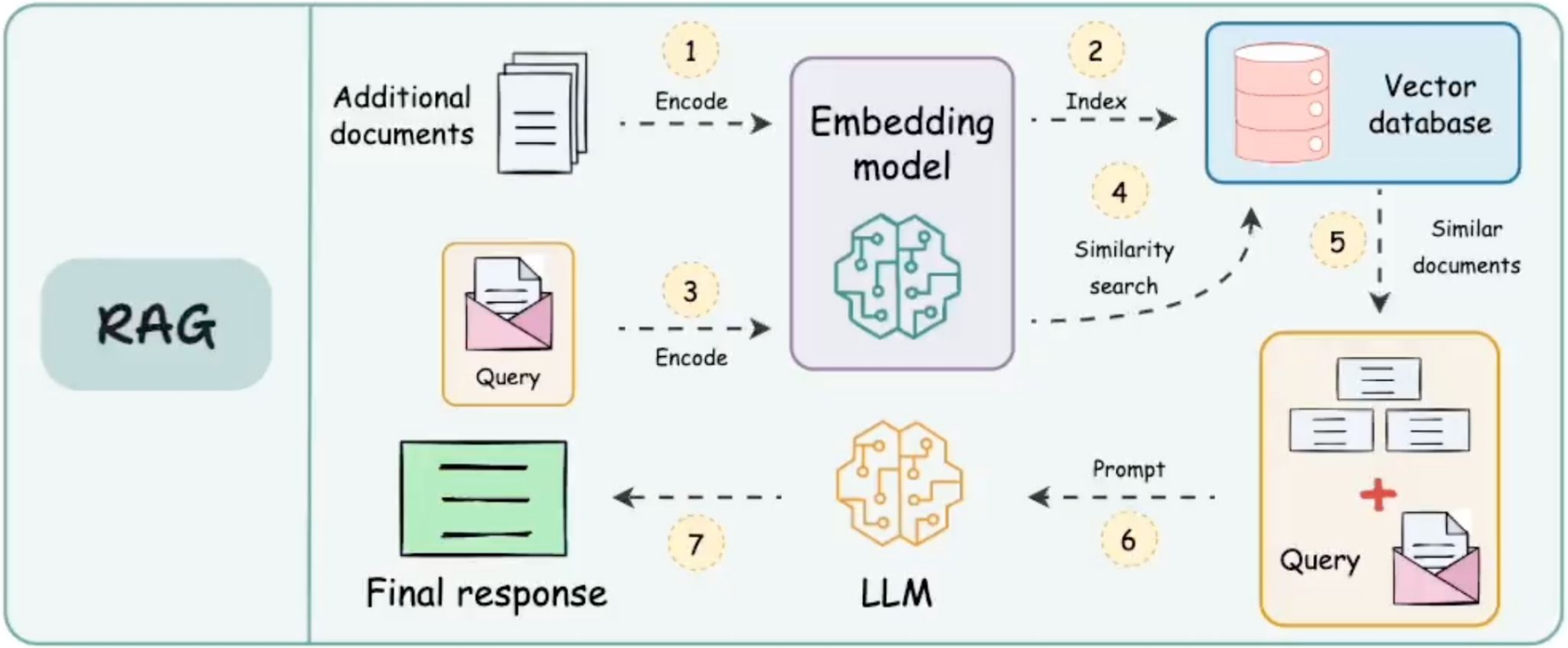
Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; [⋆]New York University;
plewis@fb.com

RAG


Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and pθ
Query Encoder
q
q(x)
Retriever pη
(Non-Parametric)
MIPS
Document Index
d(z)
z1
z2
z3
z4
Generator pθ
(Parametric)
pθ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

RAG (cont'd)



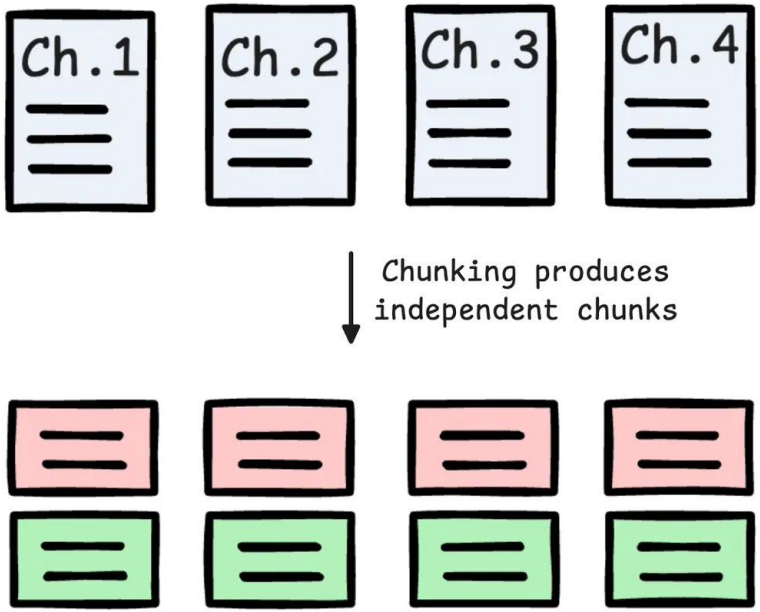
https://x.com/akshay_pachaar/status/1911397457034137889

Why RAG?

Retrieval-Augmented Generation (RAG)

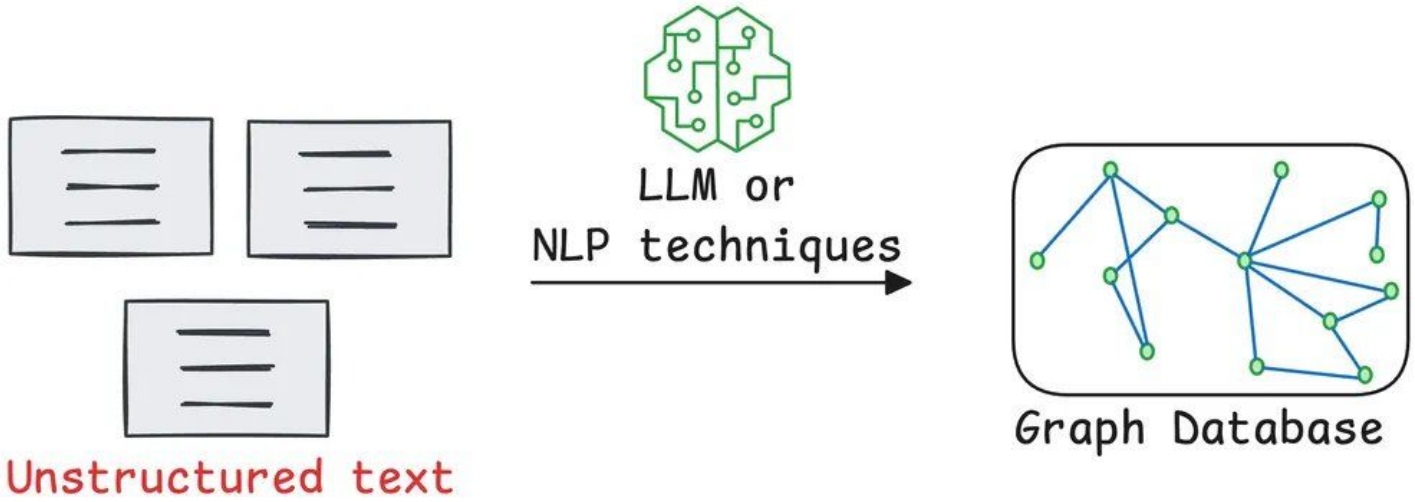
- Vanilla RAG
  - E.g., RAG, REALM
- RAG++
  - E.g., ReAct, Toolformer, FreshLLMs, GraphRAG,
- RAG + reasoning, agentic RAG
  - E.g., Self-RAG, OpenScholar, Search-R1, Deep Research

Traditional RAG struggles because it retrieves only top-k chunks while it needs the entire context



https://x.com/akshay_pachaar/status/1911397457034137889

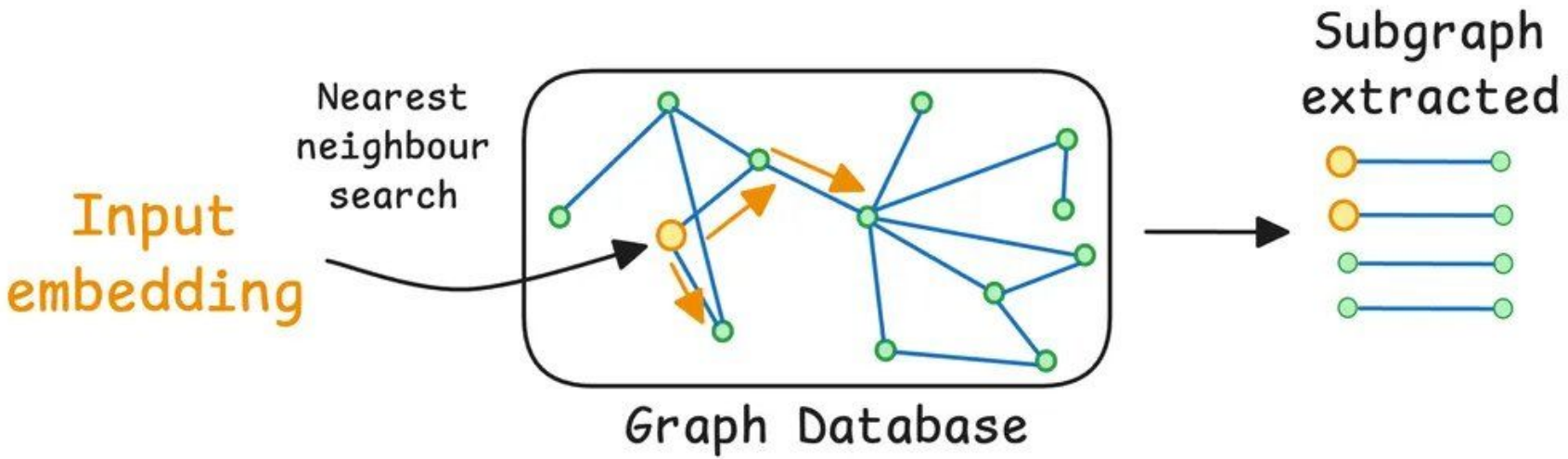
GraphRAG



An LLM creates a graph from the documents.

https://x.com/akshay_pachaar/status/1911397457034137889

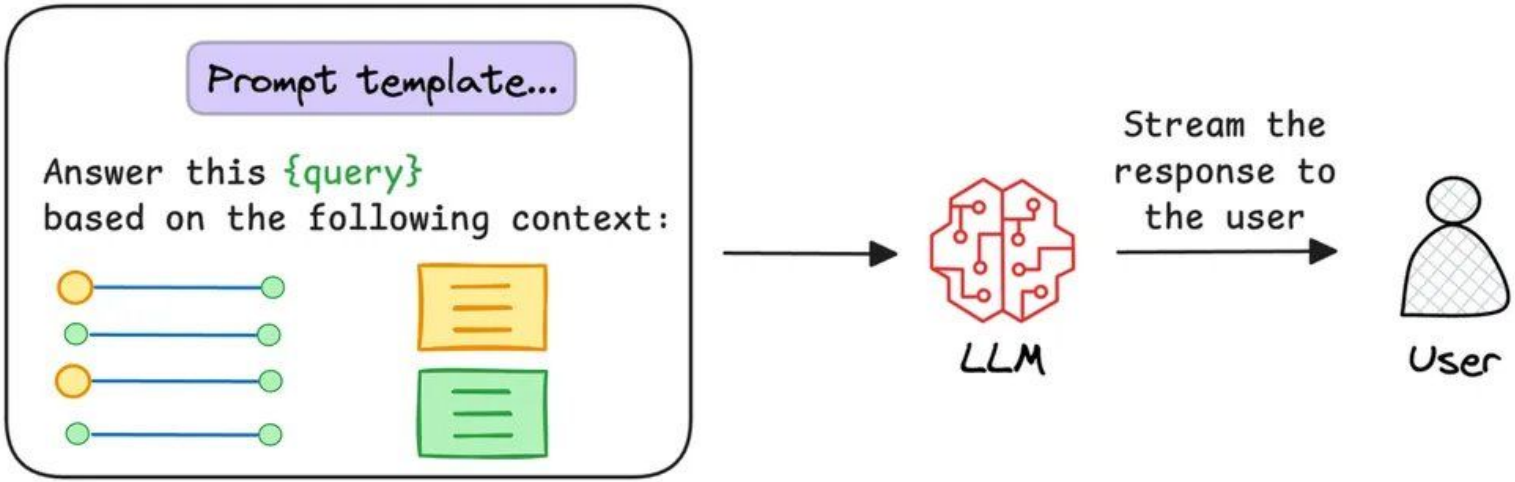
GraphRAG (cont'd)



During summarization, the system can do a graph traversal to fetch all the relevant context.

https://x.com/akshay_pachaar/status/1911397457034137889

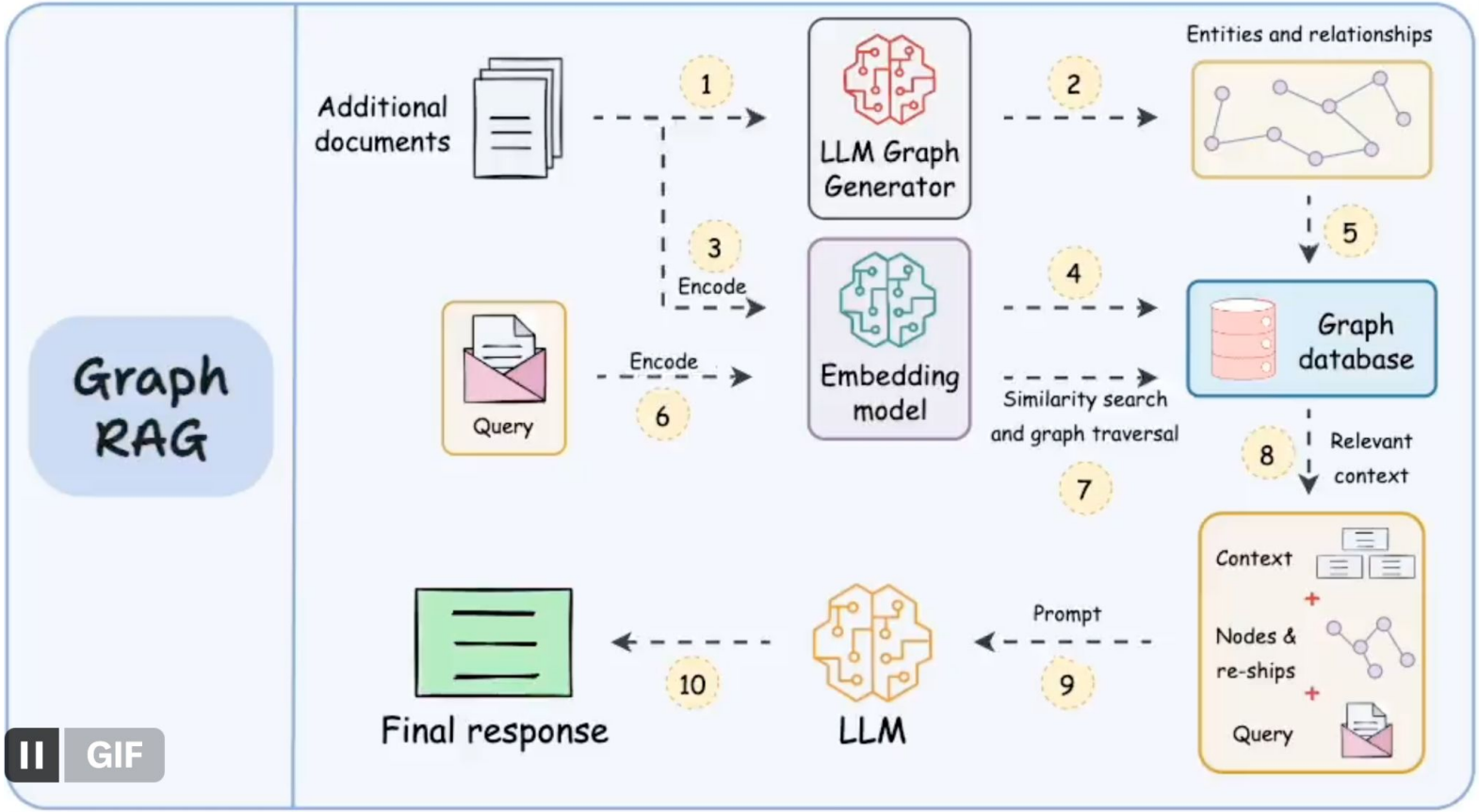
GraphRAG (cont'd)



The entire context will help the LLM produce a complete answer.

https://x.com/akshay_pachaar/status/1911397457034137889

GraphRAG (cont'd)

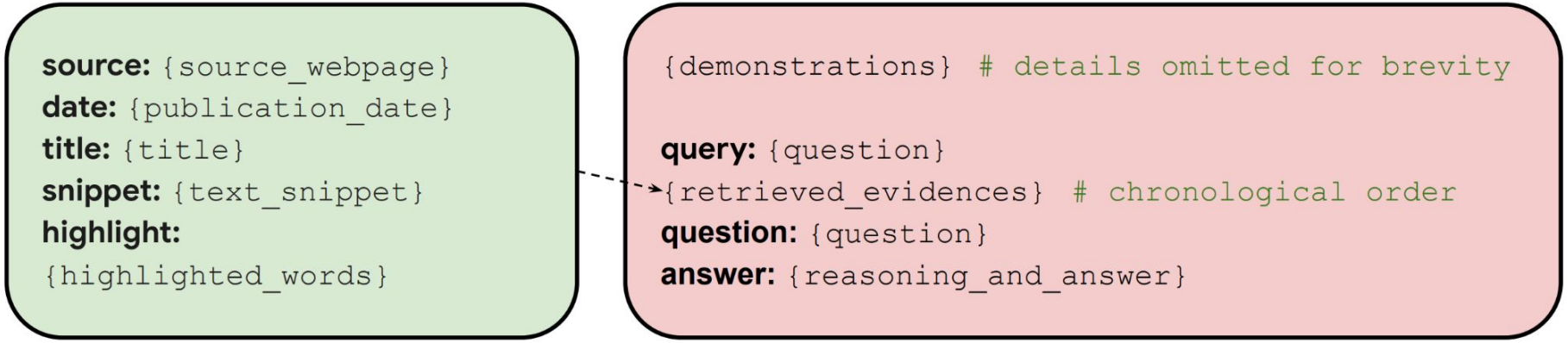


FRESHLLMS: REFRESHING LARGE LANGUAGE MODELS WITH SEARCH ENGINE AUGMENTATION

Tu Vu[1] **Mohit Iyyer**[2] **Xuezhi Wang**[1] **Noah Constant**[1] **Jerry Wei**[1]
Jason Wei[3*] **Chris Tar**[1] **Yun-Hsuan Sung**[1] **Denny Zhou**[1] **Quoc Le**[1] **Thang Luong**[1]

Google[1] University of Massachusetts Amherst[2] OpenAI[3]
freshllms@google.com

FreshPrompt



FreshPrompt uses few-shot in-context learning to teach a model to reason over retrieved evidences and figure out the right answer

Unleash the Power of Perplexity AI's Fresh Prompt Approach

Updated on Feb 28,2024

AI / LARGE LANGUAGE MODELS

How Perplexity's Online LLM Was Inspired by FreshLLMs Paper

We dig into the technology behind Perplexity's Copilot, which was inspired by the FreshLLMs paper that proposed search engine-augmented LLMs.

Jan 24th, 2024 4:00am by Janakiram MSV

Perplexity AI's Fresh Prompt Approach

e

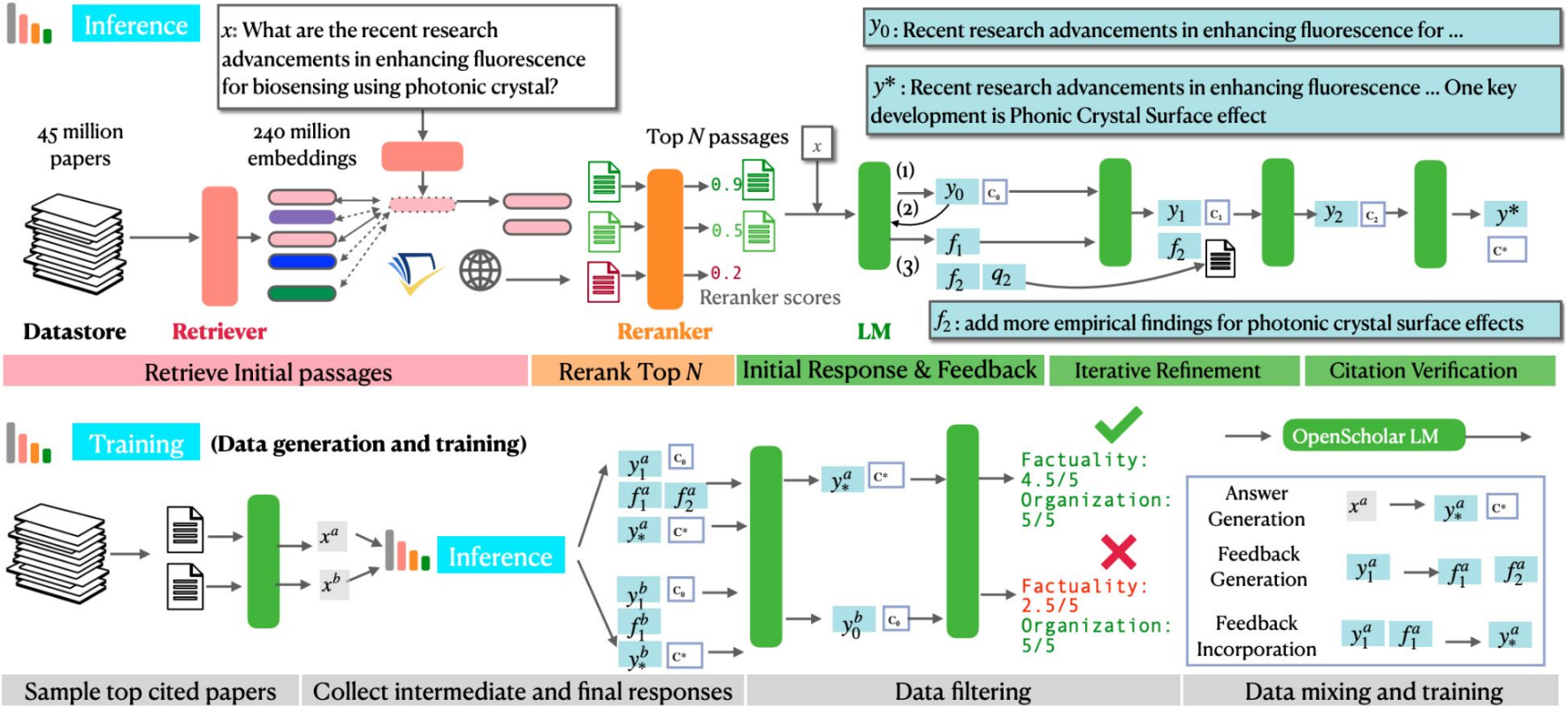
Online LLMs

lels

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

OpenScholar's approach



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

Bowen Jin[1], **Hansi Zeng**[2], **Zhenrui Yue**[1], **Jinsung Yoon**[3], **Sercan Ö. Arık**[3], **Dong Wang**[1], **Hamed Zamani**[2], **Jiawei Han**[1]
[1] Department of Computer Science, University of Illinois at Urbana-Champaign
[2] Center for Intelligent Information Retrieval, University of Massachusetts Amherst
[3] Google Cloud AI Research
{bowenj4,zhenrui3,dwang24,hanj}@illinois.edu, {hzeng, zamani}@cs.umass.edu

During the rollout, LLMs can conduct multi-turn interactions with the search engine.

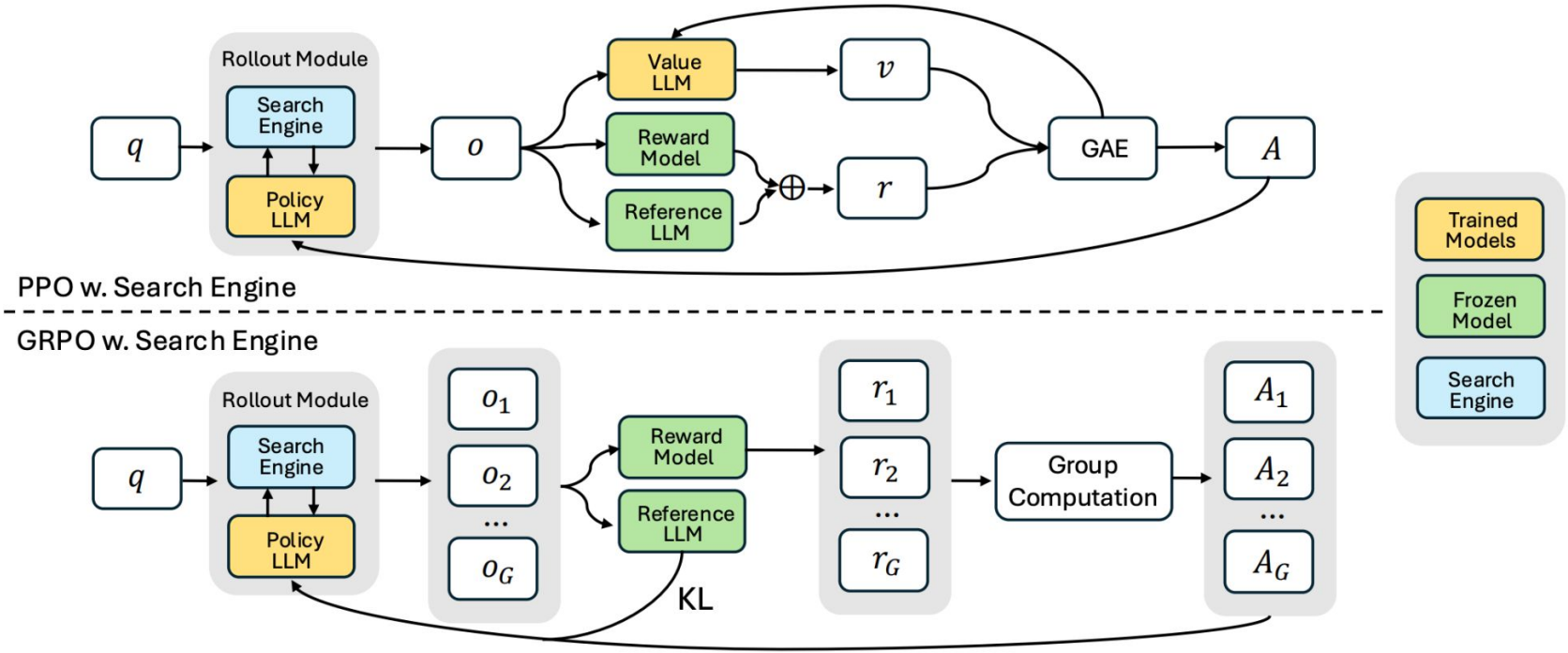


Figure 1: Demonstration of PPO and GRPO training with the search engine (SEARCH-R1).

Training template

Answer the given question. You must conduct reasoning inside <think> and </think> first every time you get new information. After reasoning, if you find you lack some knowledge, you can call a search engine by <search> query </search>, and it will return the top searched results between <information> and </information>. You can search as many times as you want. If you find no further external knowledge needed, you can directly provide the answer inside <answer> and </answer> without detailed illustrations. For example, <answer> xxx </answer>. Question: question.

Search-R1's algorithm

Algorithm 1 LLM Response Rollout with Multi-Turn Search Engine Calls

Require: Input query x, policy model π_θ, search engine $\mathcal{R}$, maximum action budget B.
Ensure: Final response y.

Initialize rollout sequence $y \leftarrow \emptyset$
Initialize action count $b \leftarrow 0$
while $b < B$ **do**
    Initialize current action LLM rollout sequence $y_b \leftarrow \emptyset$
    while True **do**
        Generate response token $y_t \sim \pi_\theta(\cdot \mid x, y + y_b)$
        Append y_t to rollout sequence $y_b \leftarrow y_b + y_t$
        if y_t in [</search>, </answer>, <eos>] **then** break
        end if
    end while
    $y \leftarrow y + y_b$
    if <search> </search> detected in y_b **then**
        Extract search query $q \leftarrow \text{Parse}(y_b, \texttt{<search>}, \texttt{</search>})$
        Retrieve search results $d = \mathcal{R}(q)$
        Insert d into rollout $y \leftarrow y +$ <information>d</information>
    else if <answer> </answer> detected in y_b **then**
        return final generated response y
    else
        Ask for rethink $y \leftarrow y +$ "My action is not correct. Let me rethink."
    end if
    Increment action count $b \leftarrow b + 1$
end while
return final generated response y

Search-R1's performance

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
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
Qwen2.5-3b-Base/Instruct								
Direct Inference	0.106	0.288	0.108	0.149	0.244	0.020	0.024	0.134
CoT	0.023	0.032	0.005	0.021	0.021	0.002	0.000	0.015
IRCoT	0.111	0.312	0.200	0.164	0.171	0.067	0.240	0.181
Search-o1	0.238	0.472	0.262	0.221	0.218	0.054	**0.320**	0.255
RAG	0.348	0.544	0.387	0.255	0.226	0.047	0.080	0.270
SFT	0.249	0.292	0.104	0.186	0.248	0.044	0.112	0.176
R1-base	0.226	0.455	0.173	0.201	0.268	0.055	0.224	0.229
R1-instruct	0.210	0.449	0.171	0.208	0.275	0.060	0.192	0.224
Search-R1-base	**0.406**	**0.587**	**0.435**	0.284	0.273	0.049	0.088	0.303
Search-R1-instruct	0.341	0.545	0.378	**0.324**	**0.319**	**0.103**	0.264	**0.325**

RAG vs. long-context LLMs

RETRIEVAL MEETS LONG CONTEXT LARGE LANGUAGE MODELS

Peng Xu[†], Wei Ping[†], Xianchao Wu, Lawrence McAfee
Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina
Mohammad Shoeybi, Bryan Catanzaro
NVIDIA
[†]{pengx, wping}@nvidia.com

ABSTRACT

Extending the context window of large language models (LLMs) is getting popular recently, while the solution of augmenting LLMs with retrieval has existed for years. The natural questions are: *i) Retrieval-augmentation versus long context window, which one is better for downstream tasks? ii) Can both methods be combined to get the best of both worlds?* In this work, we answer these questions by studying both solutions using two state-of-the-art pretrained LLMs, i.e., a proprietary 43B GPT and Llama2-70B. Perhaps surprisingly, we find that LLM with 4K context window using simple retrieval-augmentation at generation can achieve comparable performance to finetuned LLM with 16K context window via *positional interpolation* on long context tasks, while taking much less computation. More importantly, we demonstrate that retrieval can significantly improve the performance of LLMs regardless of their extended context window sizes. Our best model, retrieval-augmented Llama2-70B with 32K context window, outperforms GPT-3.5-turbo-16k and Davinci003 in terms of average score on nine long context tasks including question answering, query-based summarization, and in-context few-shot learning tasks. It also outperforms its non-retrieval Llama2-70B-32k baseline by a margin, while being much faster at generation. Our study provides general insights on the choice of retrieval-augmentation versus long context extension of LLM for practitioners.

Can Long-Context Language Models Subsume Retrieval, RAG, SQL, and More?

Jinhyuk Lee* **Anthony Chen*** **Zhuyun Dai***
Dheeru Dua **Devendra Singh Sachan** **Michael Boratko** **Yi Luan**
Sébastien M. R. Arnold **Vincent Perot** **Siddharth Dalmia** **Hexiang Hu**
Xudong Lin **Panupong Pasupat** **Aida Amini** **Jeremy R. Cole**
Sebastian Riedel **Iftekhar Naim** **Ming-Wei Chang** **Kelvin Guu**

Google DeepMind

Abstract

Long-context language models (LCLMs) have the potential to revolutionize our approach to tasks traditionally reliant on external tools like retrieval systems or databases. Leveraging LCLMs' ability to natively ingest and process entire corpora of information offers numerous advantages. It enhances user-friendliness by eliminating the need for specialized knowledge of tools, provides robust end-to-end modeling that minimizes cascading errors in complex pipelines, and allows for the application of sophisticated prompting techniques across the entire system. To assess this paradigm shift, we introduce LOFT, a benchmark of real-world tasks requiring context up to millions of tokens designed to evaluate LCLMs' performance on in-context retrieval and reasoning. Our findings reveal LCLMs' surprising ability to rival state-of-the-art retrieval and RAG systems, despite never having been explicitly trained for these tasks. However, LCLMs still face challenges in areas like compositional reasoning that are required in SQL-like tasks. Notably, prompting strategies significantly influence performance, emphasizing the need for continued research as context lengths grow. Overall, LOFT provides a rigorous testing ground for LCLMs, showcasing their potential to supplant existing paradigms and tackle novel tasks as model capabilities scale.[1]

Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach

Zhuowan Li[1] **Cheng Li**[1] **Mingyang Zhang**[1]
Qiaozhu Mei[2*] **Michael Bendersky**[1]
[1] Google DeepMind [2] University of Michigan
[1] {zhuowan,chgli,mingyang,bemike}@google.com [2] qmei@umich.edu

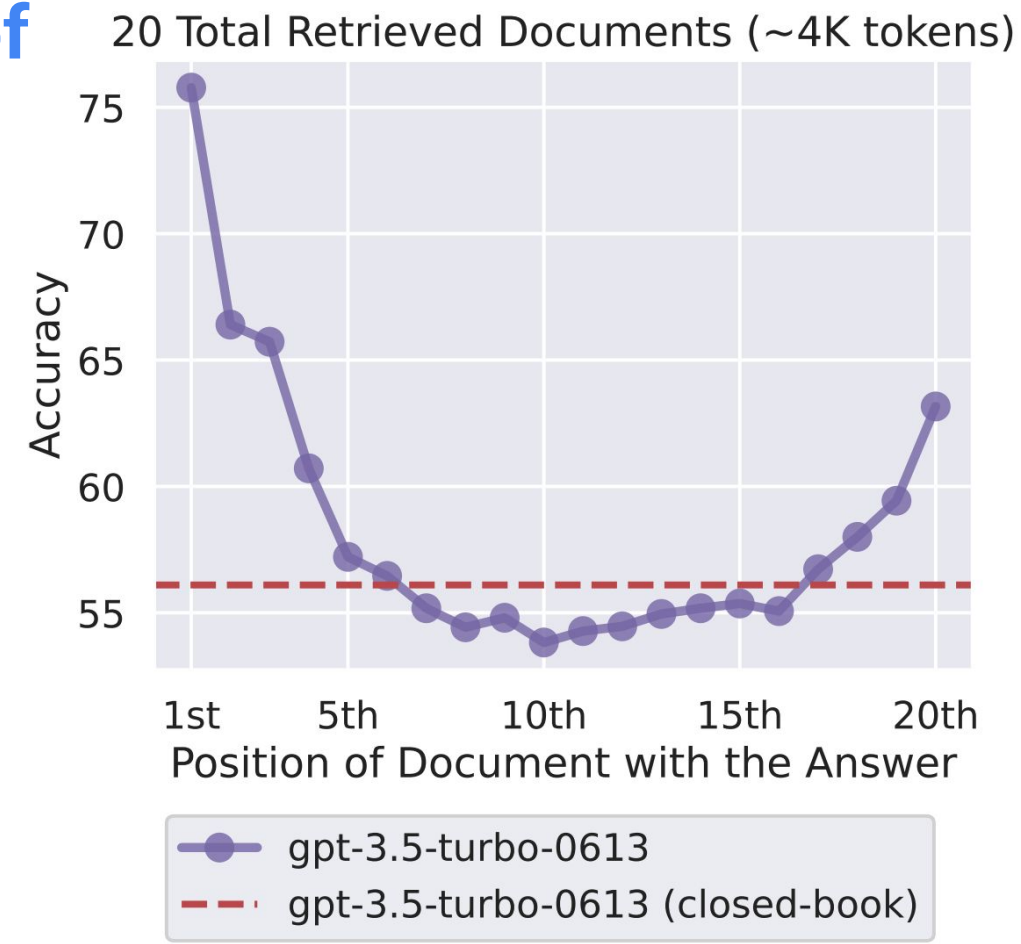
Abstract

Retrieval Augmented Generation (RAG) has been a powerful tool for *Large Language Models (LLMs)* to efficiently process overly lengthy contexts. However, recent LLMs like Gemini-1.5 and GPT-4 show exceptional capabilities to understand long contexts directly. We conduct a comprehensive comparison between RAG and long-context *(LC)* LLMs, aiming to leverage the strengths of both. We benchmark RAG and LC across various public datasets using three latest LLMs. Results reveal that when resourced sufficiently, LC consistently outperforms RAG in terms of average performance. However, RAG's significantly lower cost remains a distinct advantage. Based on this observation, we propose SELF-ROUTE, a simple yet effective method that routes queries to RAG or LC based on model self-reflection. SELF-ROUTE significantly reduces the computation cost while maintaining a comparable performance to LC. Our findings provide a guideline for long-context applications of LLMs using RAG and LC.

Lost in the Middle: How Language Models Use Long Contexts

Nelson F. Liu[1*] **Kevin Lin**[2] **John Hewitt**[1] **Ashwin Paranjape**[3]
Michele Bevilacqua[3] **Fabio Petroni**[3] **Percy Liang**[1]
[1]Stanford University [2]University of California, Berkeley [3]Samaya AI
nfliu@cs.stanford.edu

Changing the location of relevant information results in a U-shaped performance curve



Tool-use LLMs

Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick **Jane Dwivedi-Yu** **Roberto Dessì**[†] **Roberta Raileanu**
Maria Lomeli **Luke Zettlemoyer** **Nicola Cancedda** **Thomas Scialom**

Meta AI Research [†]Universitat Pompeu Fabra

Exemplary predictions of Toolformer

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Using in-context learning to generate API calls

Use an LLM to annotate a huge language modeling dataset with potential API calls

Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:

Input: Joe Biden was born in Scranton, Pennsylvania.

Output: Joe Biden was born in **[QA("Where was Joe Biden born?")]** Scranton, **[QA("In which state is Scranton?")]** Pennsylvania.

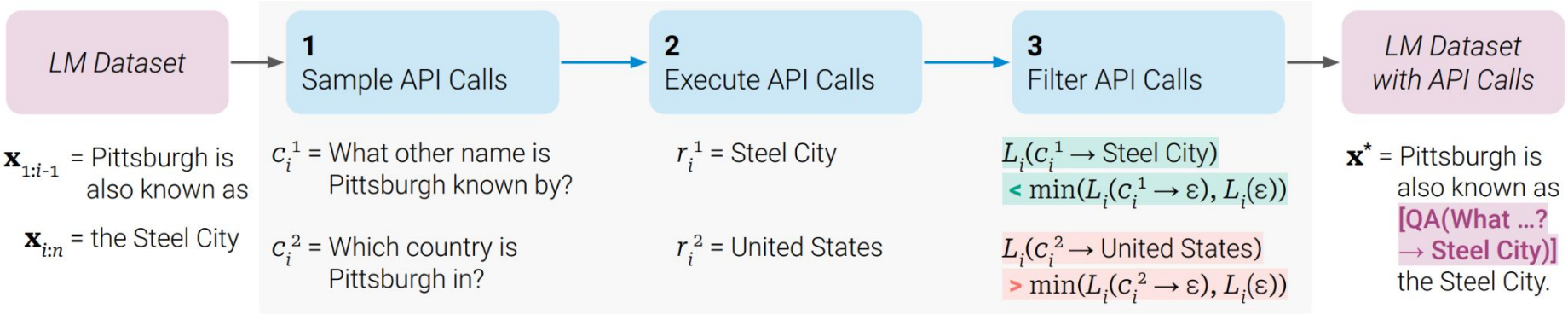
Input: Coca-Cola, or Coke, is a carbonated soft drink manufactured by the Coca-Cola Company.

Output: Coca-Cola, or **[QA("What other name is Coca-Cola known by?")]** Coke, is a carbonated soft drink manufactured by **[QA("Who manufactures Coca-Cola?")]** the Coca-Cola Company.

Input: x

Output:

Filtering out all API calls which do not reduce the loss over the next tokens



Intuitively, an API call is helpful if providing it with both the input and the output of this call makes it easier for the model to predict future tokens, compared to not receiving the API call at all, or receiving only its input

Toolformer (6.7B) achieves much stronger zero-shot results than OPT (66B) and GPT-3 (175B)

Model	ASDiv	SVAMP	MAWPS
GPT-J	7.5	5.2	9.9
GPT-J + CC	9.6	5.0	9.3
Toolformer (disabled)	14.8	6.3	15.0
Toolformer	**40.4**	**29.4**	**44.0**
OPT (66B)	6.0	4.9	7.9
GPT-3 (175B)	14.0	10.0	19.8

Model	WebQS	NQ	TriviaQA
GPT-J	18.5	12.8	43.9
GPT-J + CC	18.4	12.2	45.6
Toolformer (disabled)	18.9	12.6	46.7
Toolformer	**26.3**	**17.7**	**48.8**
OPT (66B)	18.6	11.4	45.7
GPT-3 (175B)	29.0	22.6	65.9

Deep Research

- https://openai.com/index/introducing-deep-research/
- An agent that uses reasoning to synthesize large amounts of online information and complete multi-step research tasks for you.

Deep Research

https://openai.com/index/introducing-deep-research/

Deep research is OpenAI's next agent that can do work for you independently—you give it a prompt, and ChatGPT will find, analyze, and synthesize hundreds of online sources to create a comprehensive report at the level of a research analyst. Powered by a version of the upcoming OpenAI o3 model that’s optimized for web browsing and data analysis, it leverages reasoning to search, interpret, and analyze massive amounts of text, images, and PDFs on the internet, pivoting as needed in reaction to information it encounters.

Deep Research (cont'd)

Deep research is built for people who do intensive knowledge work in areas like finance, science, policy, and engineering and need thorough, precise, and reliable research. It can be equally useful for discerning shoppers looking for hyper-personalized recommendations on purchases that typically require careful research, like cars, appliances, and furniture. Every output is fully documented, with clear citations and a summary of its thinking, making it easy to reference and verify the information. It is particularly effective at finding niche, non-intuitive information that would require browsing numerous websites. Deep research frees up valuable time by allowing you to offload and expedite complex, time-intensive web research with just one query.

Deep Research (cont'd)

Deep research independently discovers, reasons about, and consolidates insights from across the web. To accomplish this, it was trained on real-world tasks requiring browser and Python tool use, using the same reinforcement learning methods behind OpenAI o1, our first reasoning model. While o1 demonstrates impressive capabilities in coding, math, and other technical domains, many real-world challenges demand extensive context and information gathering from diverse online sources. Deep research builds on these reasoning capabilities to bridge that gap, allowing it to take on the types of problems people face in work and everyday life.

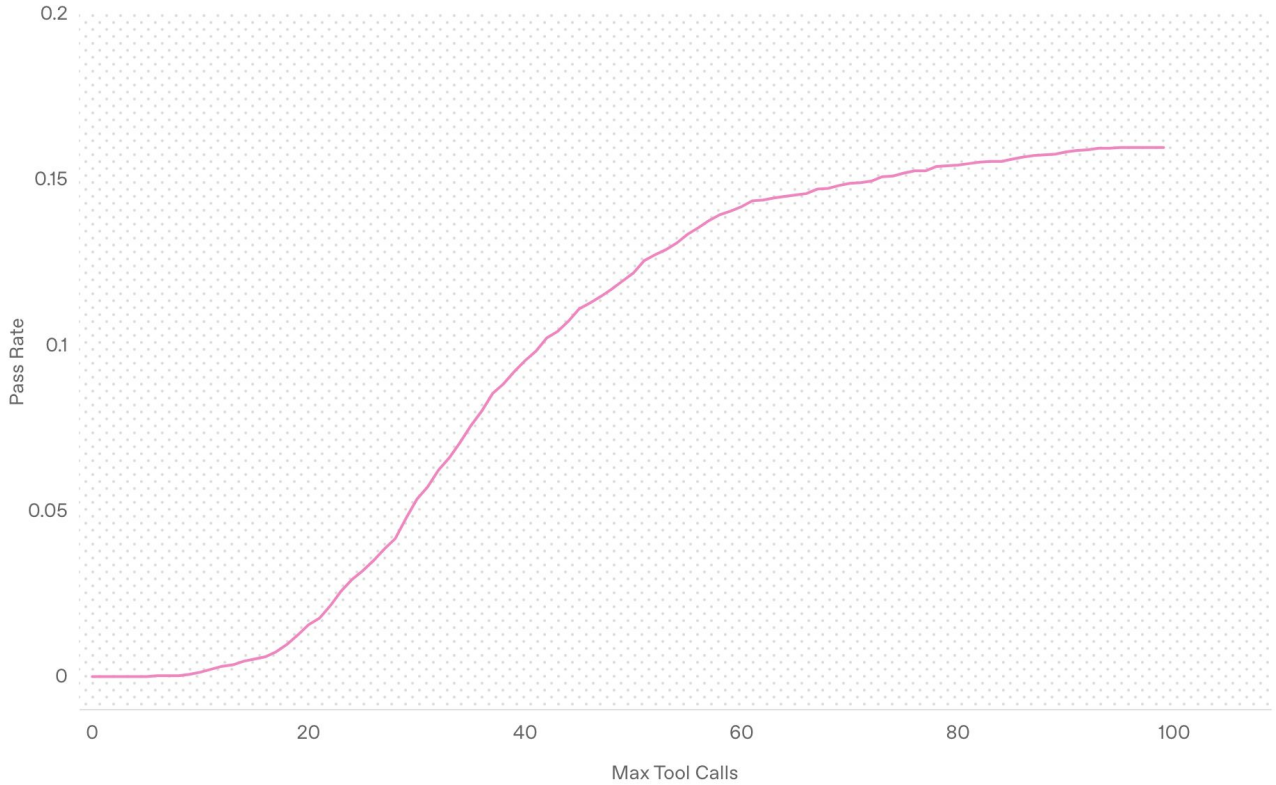
Deep Research on Humanity's Last Exam

Model	Accuracy (%)
GPT-4o	3.3
Grok-2	3.8
Claude 3.5 Sonnet	4.3
Gemini Thinking	6.2
OpenAI o1	**9.1**
DeepSeek-R1*	9.4
OpenAI o3-mini (medium)*	10.5
OpenAI o3-mini (high)*	13.0
OpenAI deep research**	**26.6**

* Model is not multi-modal, evaluated on text-only subset.

**with browsing + python tools

The more the model browses and thinks about what it is browsing, the better it does



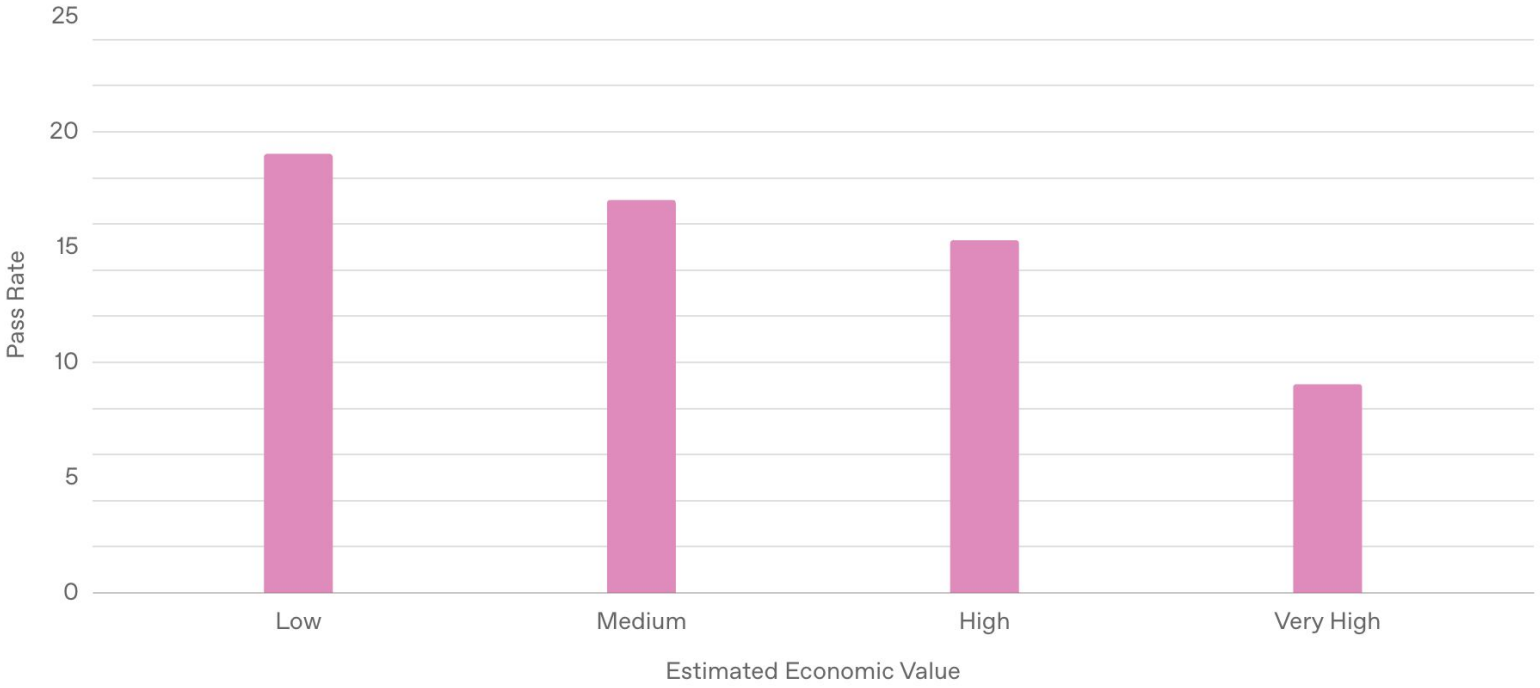
Deep Research on Humanity's Last Exam

Model	Accuracy (%)
GPT-4o	3.3
Grok-2	3.8
Claude 3.5 Sonnet	4.3
Gemini Thinking	6.2
OpenAI o1	**9.1**
DeepSeek-R1*	9.4
OpenAI o3-mini (medium)*	10.5
OpenAI o3-mini (high)*	13.0
OpenAI deep research**	**26.6**

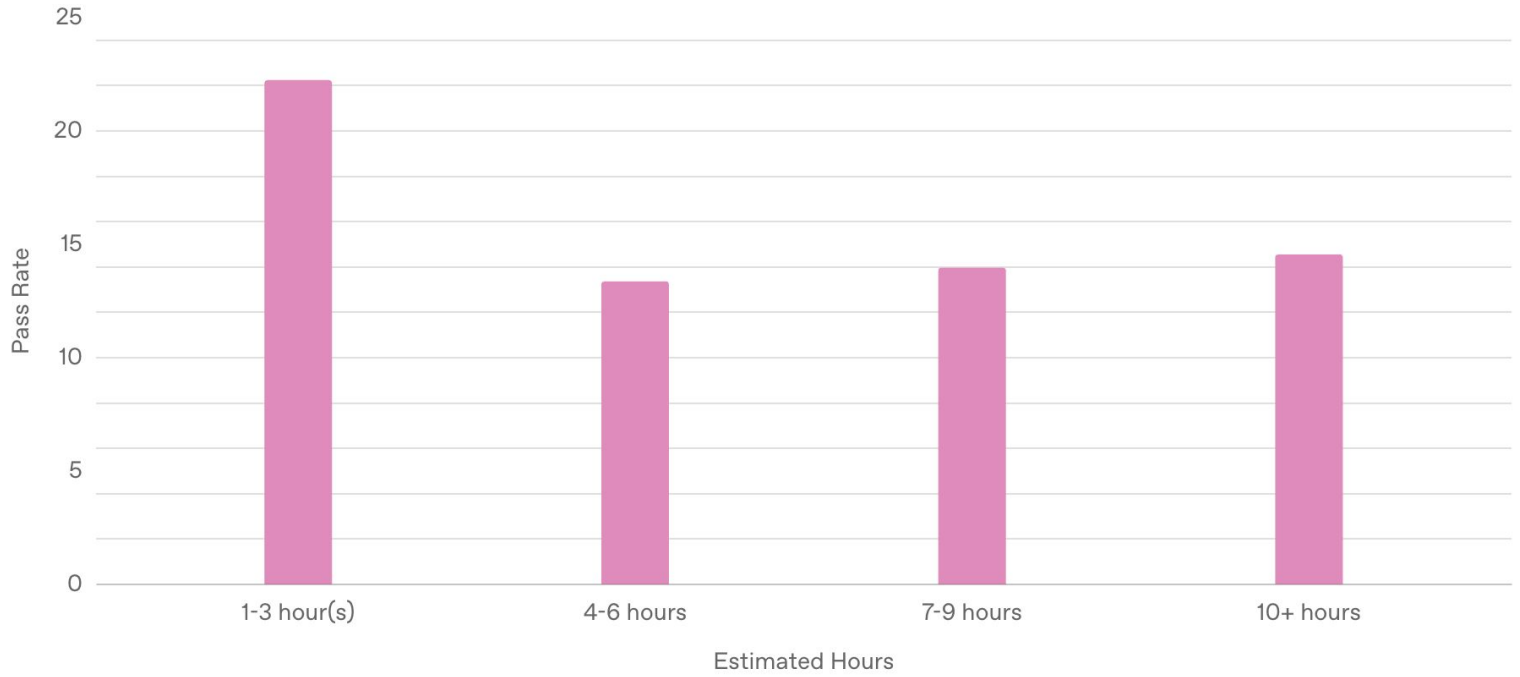
* Model is not multi-modal, evaluated on text-only subset.

**with browsing + python tools

Pass rate on expert-level tasks



Pass rate on expert-level tasks (cont'd)



Estimated economic value of task is more correlated with pass rate than # of hours it would take a human – the things that models find difficult are different to what humans find time-consuming.

Thank you!