The Frontier between Retrieval-augmented and Long-context Language Models

Danqi Chen
Princeton University

April 1, 2025

PRINCETON
Language + Intelligence

Talk agenda

Retrieval-augmented language models (RALMs)

Long-context language models (LCLMs)

- **Part 1:** The bigger picture: LCLMs vs RALMs
- **Part 2:** (brief) We still don’t know how to train RALMs yet ☹
- **Part 3:** (most of the talk) An open recipe for training LCLMs ☺

The Bigger Picture: LCLMs vs RALMs

Long-context language models (LCLMs)

- = LMs capable of handling long context sizes
- **Long input** (e.g., input as a book or Web pages)
- … and **long output** (e.g., recent long CoT models)

Mid-2023:

Meta LlaMA-1 2K, LlaMA-2 4K

GPT-4: 8K

Mid-to-late 2024

Meta Llama-3 (3.1): 128K

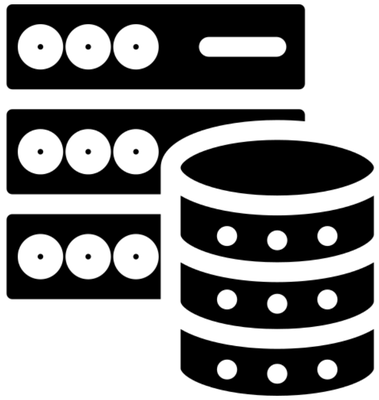
GPT-4o: 128K

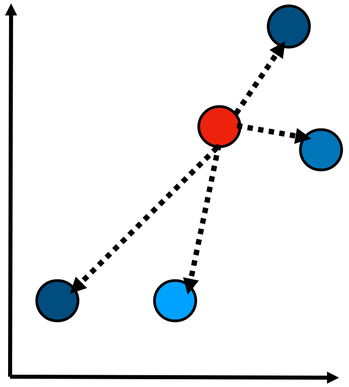
Claude 3.5: 200K

Gemini 1.5 Pro: 2M

Retrieval-augmented language models (RALMs)

- = LMs that are **inherently designed or trained** to incorporate retrieval during inference
- **What is retrieval?**
  - #1: call a search engine (e.g., Google, Bing) *“tool use”*
  - #2: build an index (nearest neighbor search in a vector datastore)





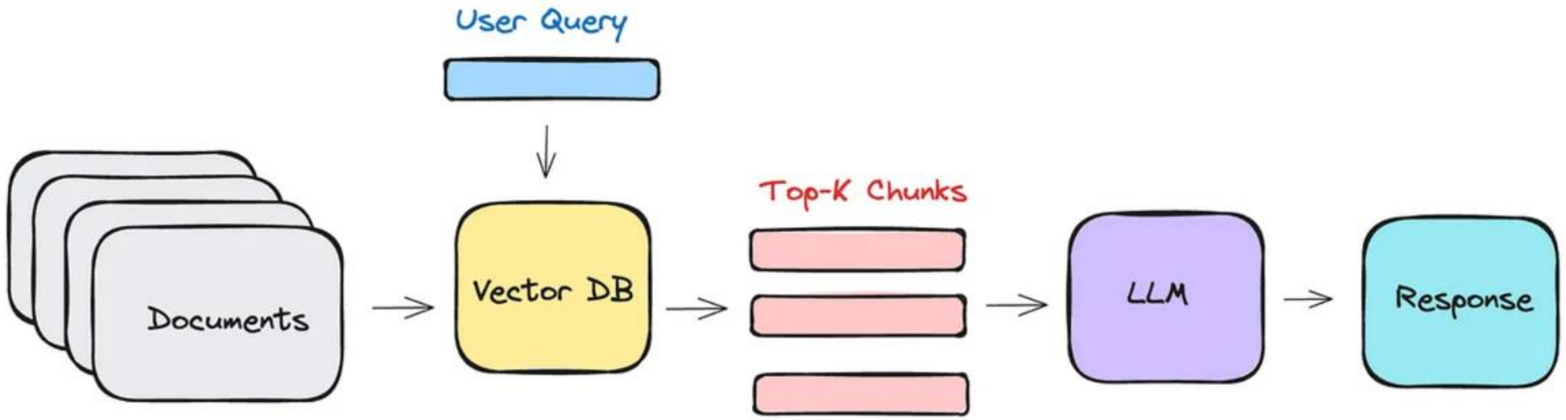
- Reproducible research
- End-to-end training of retrieval and LMs
- Improving retrieval is a research problem itself!

Retrieval-augmented language models (RALMs)

- **How is retrieval used?**

- #1: Take input query, call retrieval and fetch retrieved documents in context

“RAG”

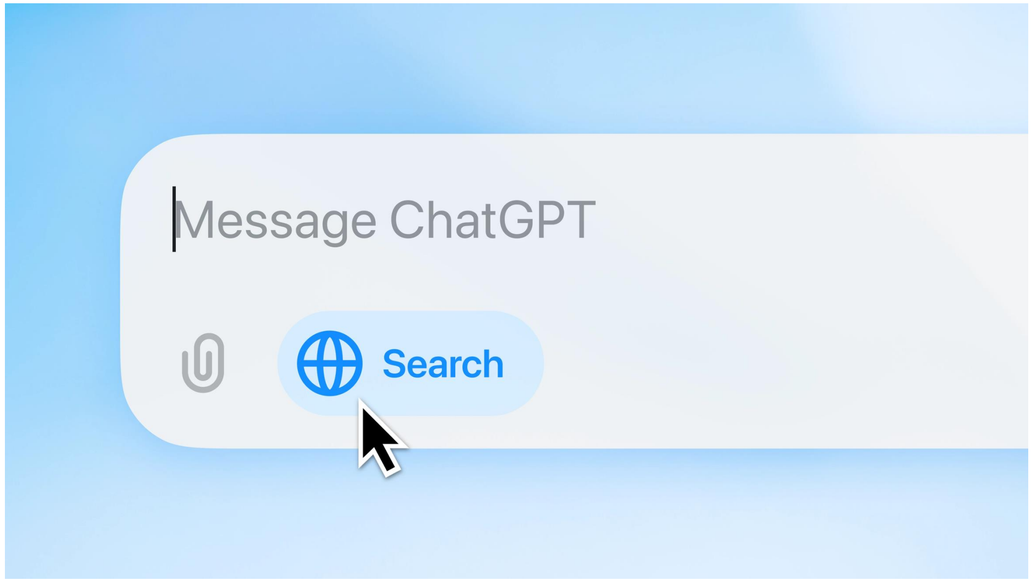


(Credit: Maysa Mayel)

- #2: interactive, multi-step: let models decide when and what to call retrieval

“Agentic search”

Why do we need retrieval?



perplexity

Can you tell whether UA1306 was on time in the past week or not

Pro Search | Sources 19 | 2 steps

airportia.com
UA1306 Flight Status...

AirNav Radar
Flight UA1306 / UAL130...

FlightStats
(UA) United Airlines 1306...

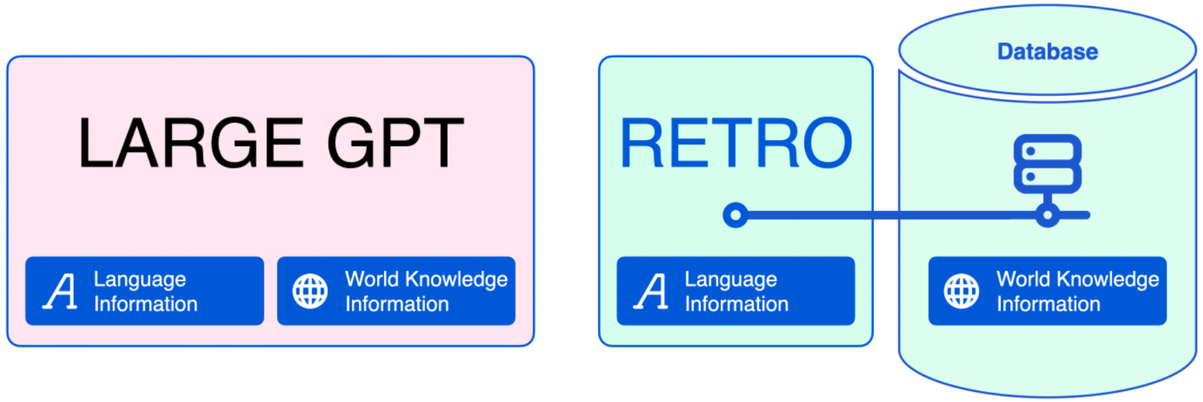
+16 sources

In the past week, United Airlines Flight UA1306 from Philadelphia (PHL) to San Francisco (SFO) experienced mixed punctuality. Here's the detailed breakdown:

- **March 30**: The flight landed late by 2 hours and 26 minutes [1] [4].
- **March 28**: The flight landed 34 minutes early, despite departing 15 minutes late [1] [4].
- **March 27**: The flight landed 25 minutes early, with a departure delay of 14 minutes [1] [4].

Why do we need retrieval?

- **Factuality**
- **Adaptability**
- **Attribution**
- **Efficiency / scaling**: vector datastore + smaller LM outperforms large LMs?



(Borgeaud et al., 2022)

“With **a 2 trillion token** database, Retro obtains comparable performance to GPT-3.. despite using **25× fewer parameters**”

Enabling LMs to cite (“attribution”)

ALCE (end-to-end automatic evaluation for LLM’s citation)

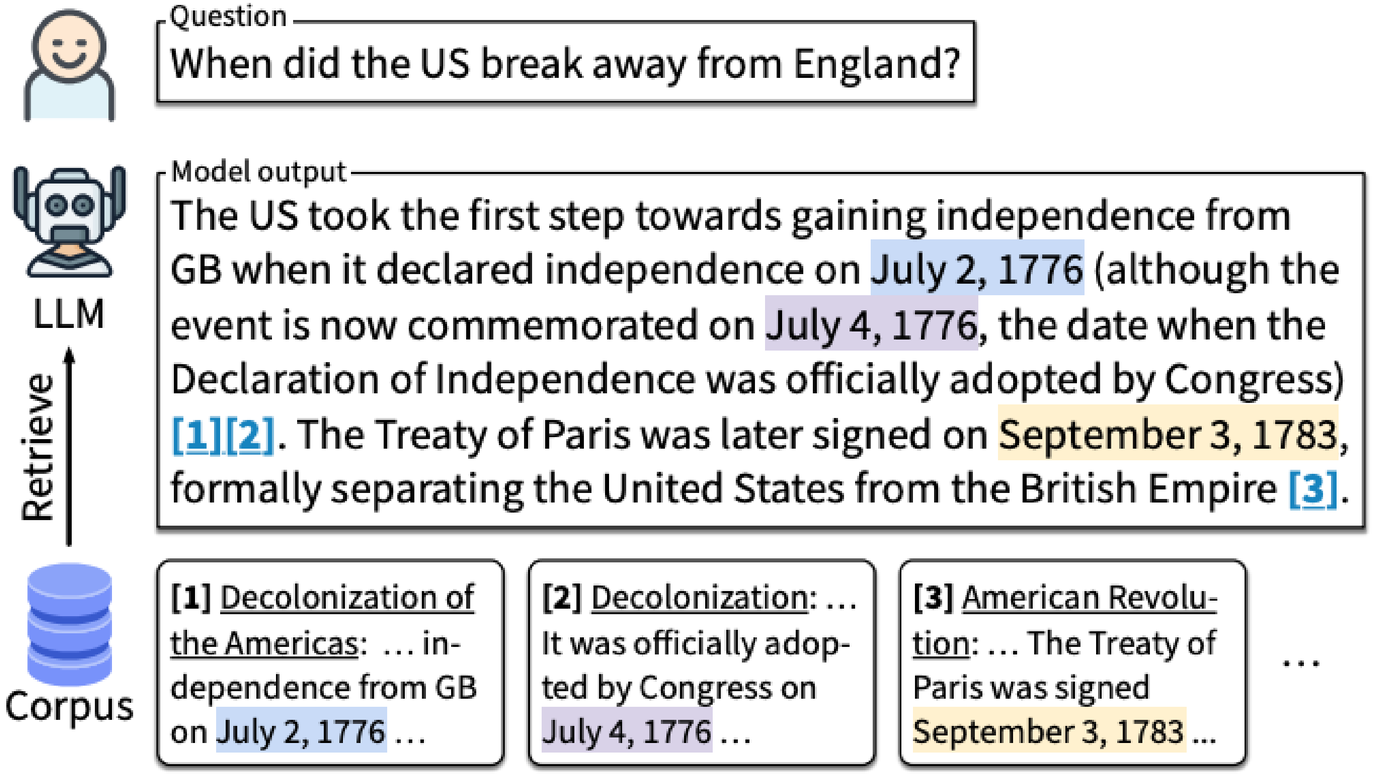
Enabling Large Language Models to Generate Text with Citations

Tianyu Gao Howard Yen Jiatong Yu Danqi Chen
Department of Computer Science & Princeton Language and Intelligence
Princeton University
{tianyug,hyen,jiatongy,danqic}@cs.princeton.edu

(Gao et al., EMNLP 2023)

Enabling LMs to cite (“attribution”)

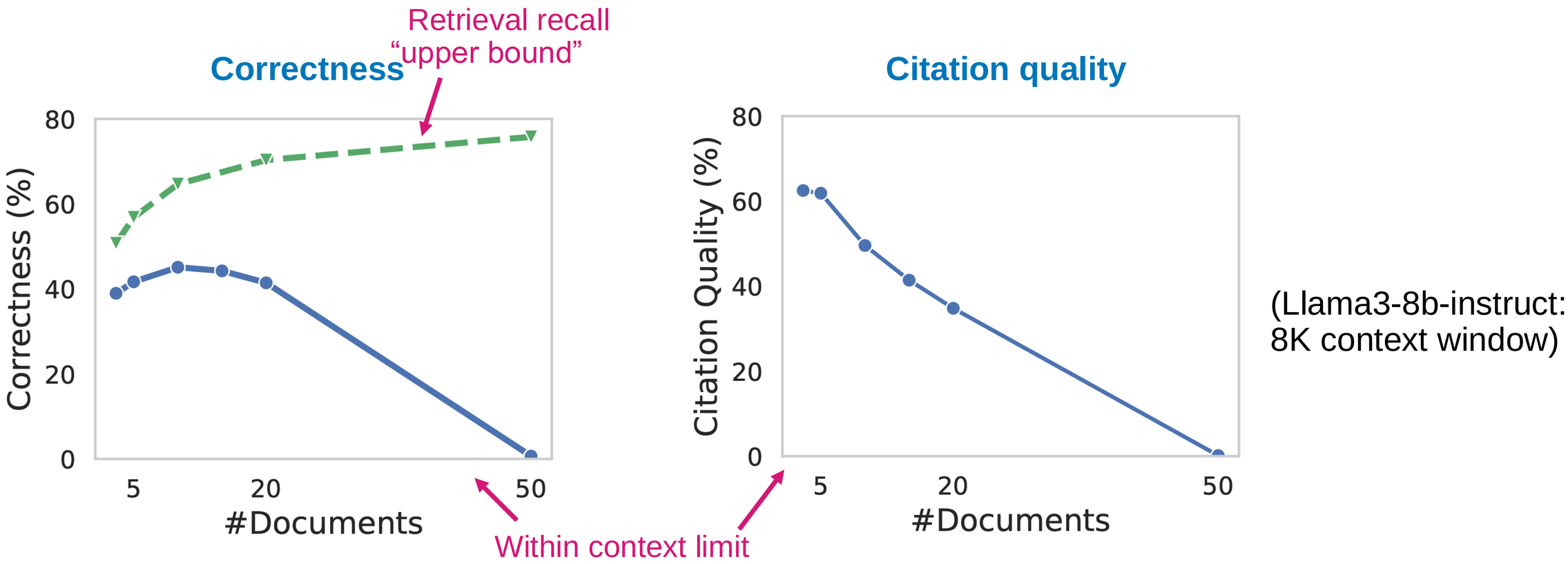
ALCE (end-to-end automatic evaluation for LLM’s citation)



- The task: **retrieve** a set of passages and **generate** a relevant output with **correct citations**
- Evaluation: **correctness** and **citation quality**

Gao et al., 2023: Enabling large language models to generate text with citations

Enabling LMs to cite (“attribution”)



(Llama3-8b-instruct:
8K context window)

Long-context? Instruction following?

Gao et al., 2023: Enabling large language models to generate text with citations

Long-context vs retrieval-augmented LMs

- We need LCLMs to support RAG
- **Q. Can LCLMs replace RAG?**

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

RETRIEVAL MEETS LONG CONTEXT LARGE LANGUAGE MODELS

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

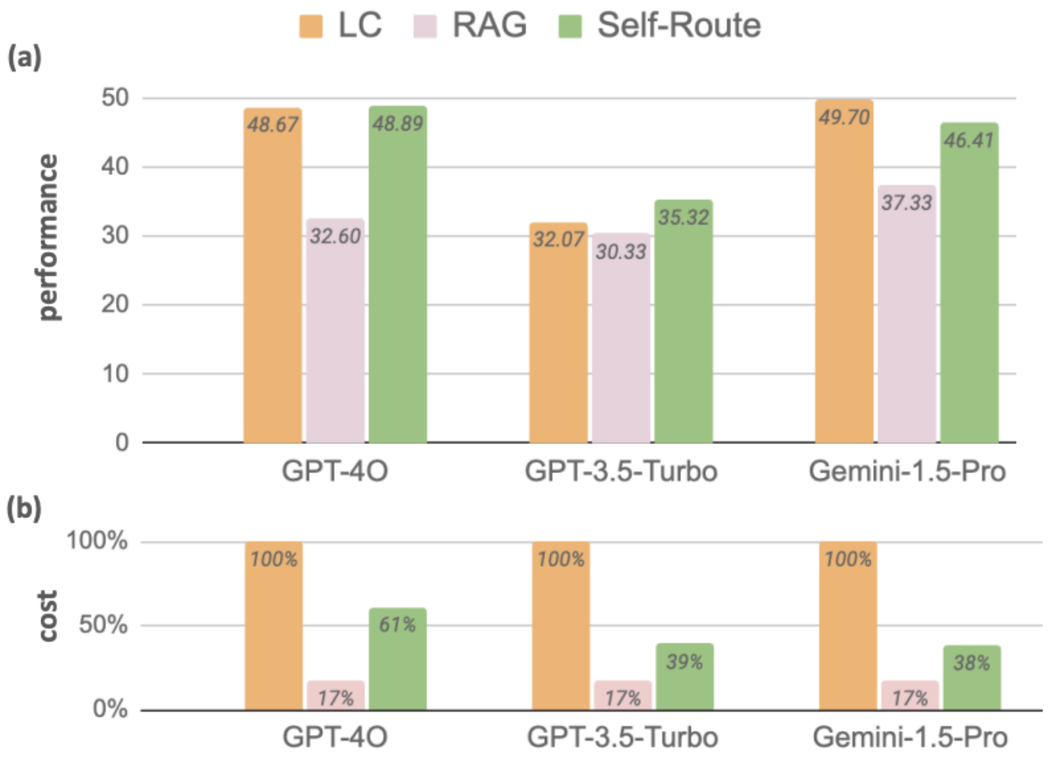


Figure from (Li et al., 2024)

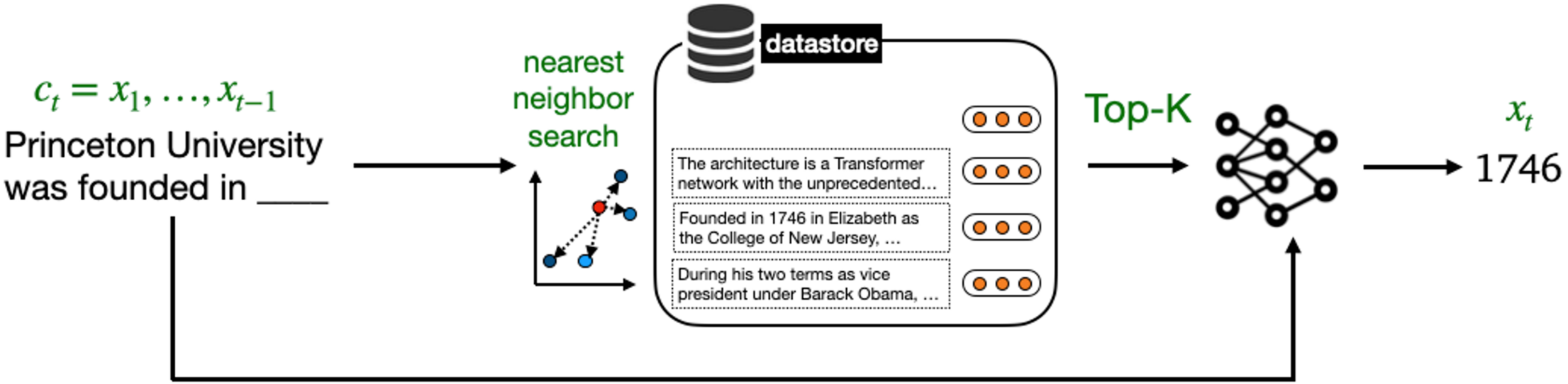
Long-context vs retrieval-augmented LMs

Q. Can LCLMs replace RAG?

- ★ We can’t pack all documents in LCLMs
- ★ LCLMs may lower the bar for retrieval quality
- ★ RAG is still more efficient (related: inference techniques for LCLMs)

We still don’t know how to train RALMs yet

Training RALMs is an unsolved research challenge



(Khaldelwal et al., 2020; Guu et al., 2020; Borgeaud et al., 2021; Zhong et al., 2022; Izacard, et al., 2022; Min et al., 2022)

Also see https://acl2023-retrieval-lm.github.io

DeepMind

Google's RETRO

Improving language models by retrieving from trillions of tokens

Sebastian Borgeaud[†], **Arthur Mensch**[†], **Jordan Hoffmann**[†], **Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae**[‡], **Erich Elsen**[‡] **and Laurent Sifre**[†,‡]

All authors from DeepMind, [†]Equal contributions, [‡]Equal senior authorship

Meta's Atlas

Atlas: Few-shot Learning with Retrieval Augmented Language Models

Gautier Izacard[*,◇,♣,♡] *gizacard@fb.com*
Patrick Lewis[*,◇] *plewis@fb.com*
Maria Lomeli[◇] *marialomeli@fb.com*
Lucas Hosseini[◇] *hoss@fb.com*
Fabio Petroni[◇] *fabiopetroni@fb.com*
Timo Schick[◇] *schick@fb.com*
Jane Dwivedi-Yu[◇] *janeyu@fb.com*
Armand Joulin[◇] *ajoulin@fb.com*
Sebastian Riedel[◇,♠] *sriedel@fb.com*
Edouard Grave[◇] *egrave@fb.com*

[◇] *Meta AI Research,* [♣] *ENS, PSL University,* [♡] *Inria,* [♠] *University College London*

Training Language Models with Memory Augmentation

Zexuan Zhong[†] **Tao Lei**[*] **Danqi Chen**[†]
[†]Princeton University
{zzhong, danqic}@cs.princeton.edu, taole@google.com

Nonparametric Masked Language Modeling

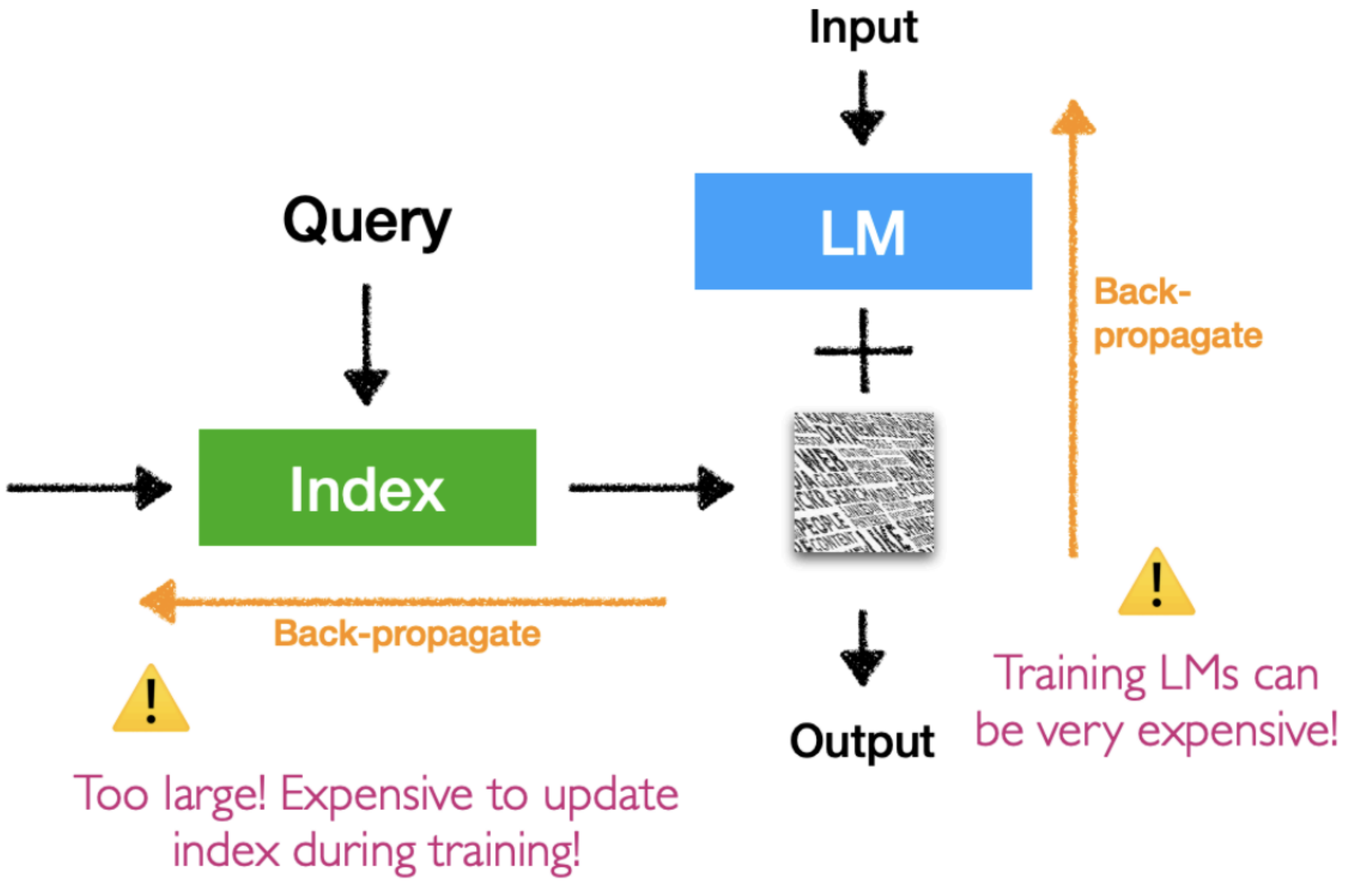
Sewon Min[1,2] **Weijia Shi**[1,2] **Mike Lewis**[2] **Xilun Chen**[2]
Wen-tau Yih[2] **Hannaneh Hajishirzi**[1,3] **Luke Zettlemoyer**[1,2]
[1]University of Washington [2]Meta AI [3]Allen Institute for AI
{sewon,swj0419,hannaneh,lsz}@cs.washington.edu
{mikelewis,xilun,scottyih}@meta.com

REPLUG: Retrieval-Augmented Black-Box Language Models

Weijia Shi,[1 *] **Sewon Min,**[1] **Michihiro Yasunaga,**[2] **Minjoon Seo,**[3] **Rich James,**[4] **Mike Lewis,**[4]
Luke Zettlemoyer[1 4] **Wen-tau Yih** [4]

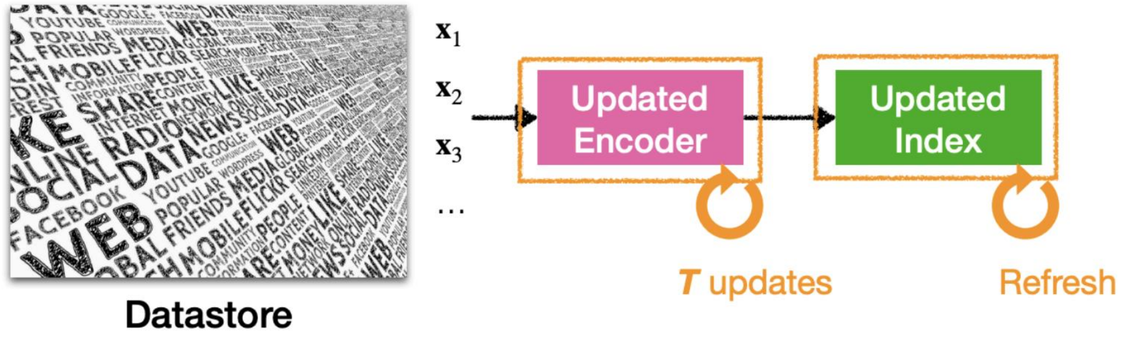
The challenges of training RALMs

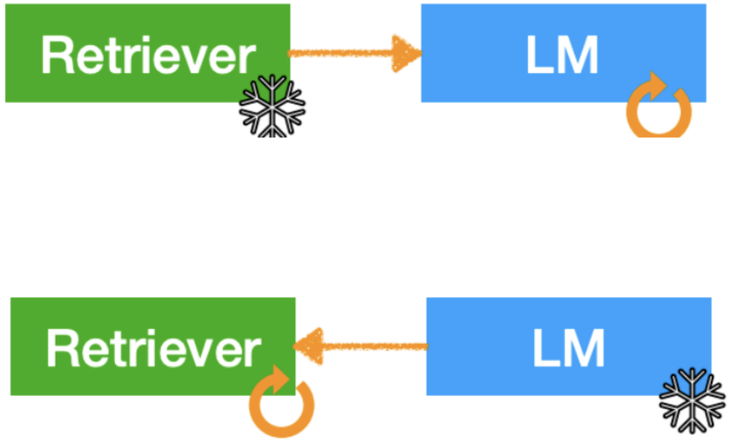
Datastore

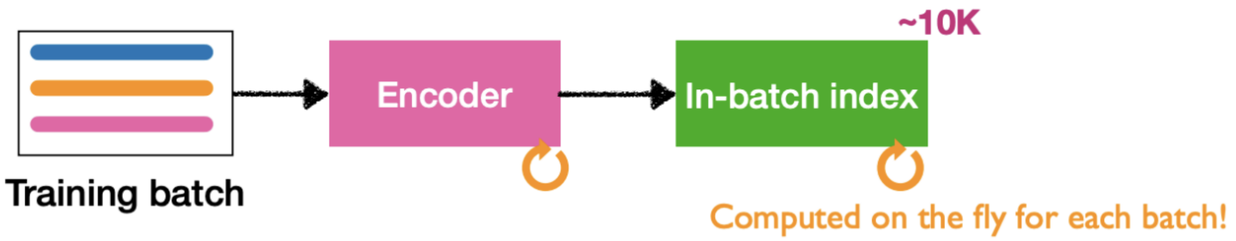


Different training methods

- Independent training
- Sequential training
- Joint training w/ asynchronous index update
- Joint training w/ in-batch approximation







https://acl2023-retrieval-lm.github.io/

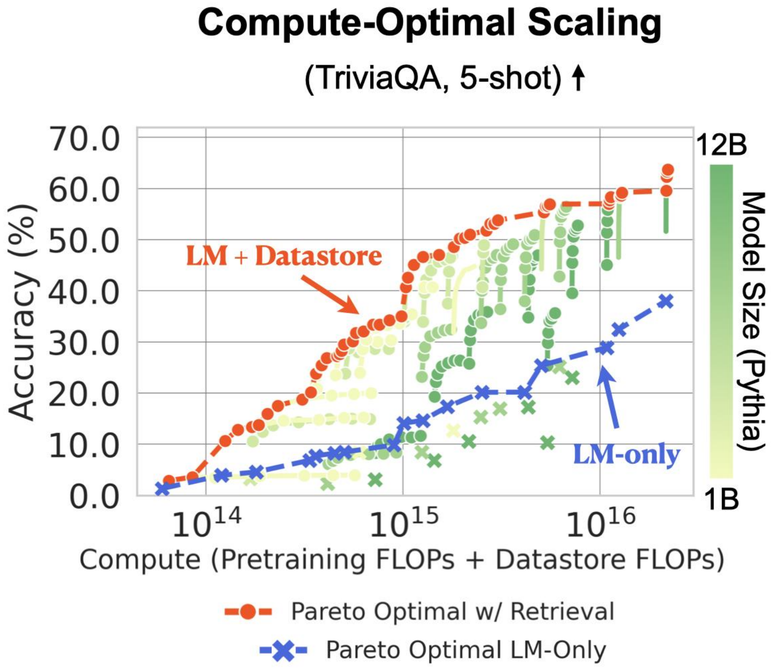
Why progress in training RALMs has stalled

- Breakthroughs in small language models (SLMs) and agentic search
- Model architecture vs training methods not fully decided at small scale
- Significant barrier for open-source research: engineering efforts, open-sourced datastores

Scaling Retrieval-Based Language Models with a Trillion-Token Datastore

Rulin Shao[1] Jacqueline He[1] Akari Asai[1] Weijia Shi[1]
Tim Dettmers[1] Sewon Min[1] Luke Zettlemoyer[1] Pang Wei Koh[1,2]
[1]University of Washington [2]Allen Institute for AI
{rulins,jyyh,akari,swj0419,dettmers,sewon,lsz,pangwei}
@cs.washington.edu

(Shao et al., NeurIPS 2024)



How to train and evaluate LCLMs?

How to *evaluate* LCLMs?

- Prior work relies on perplexity…

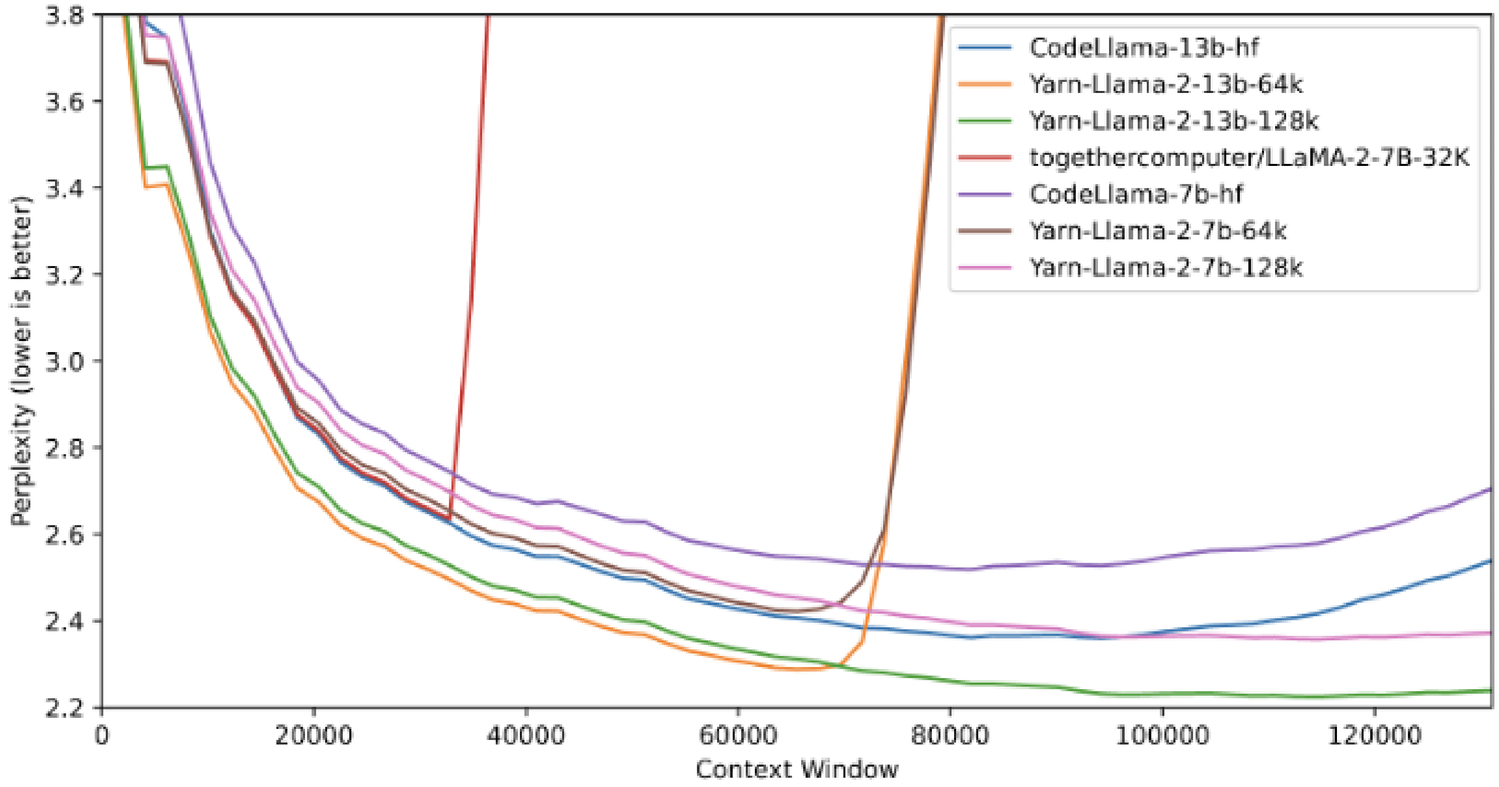
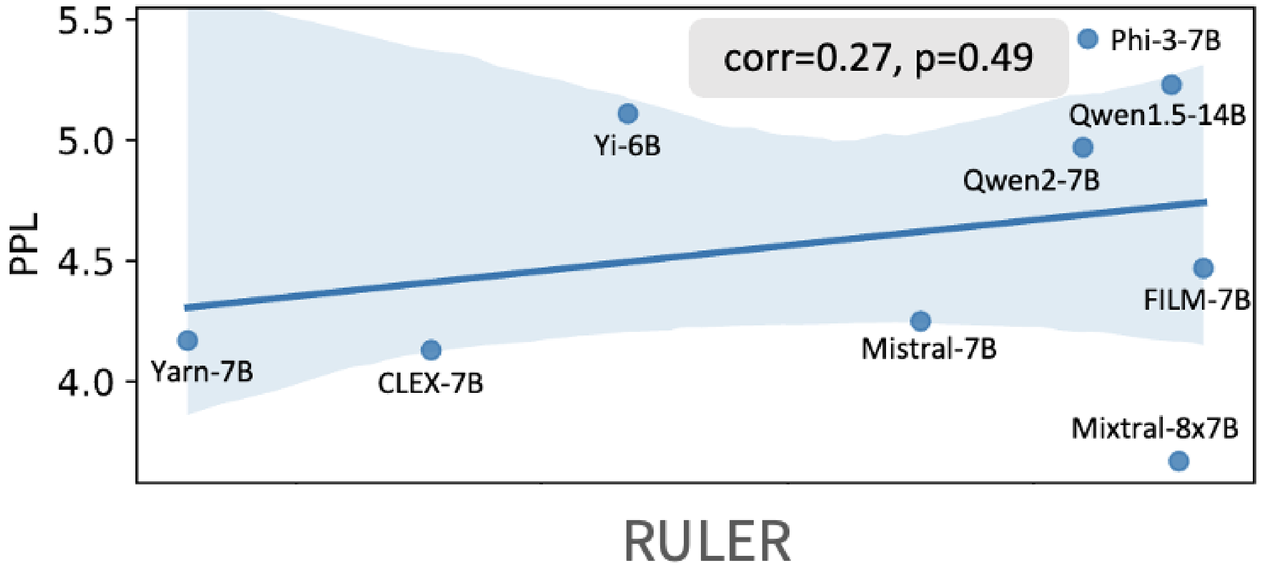


Figure from Peng et al., 2023, YaRN: Efficient Context Window Extension of Large Language Models

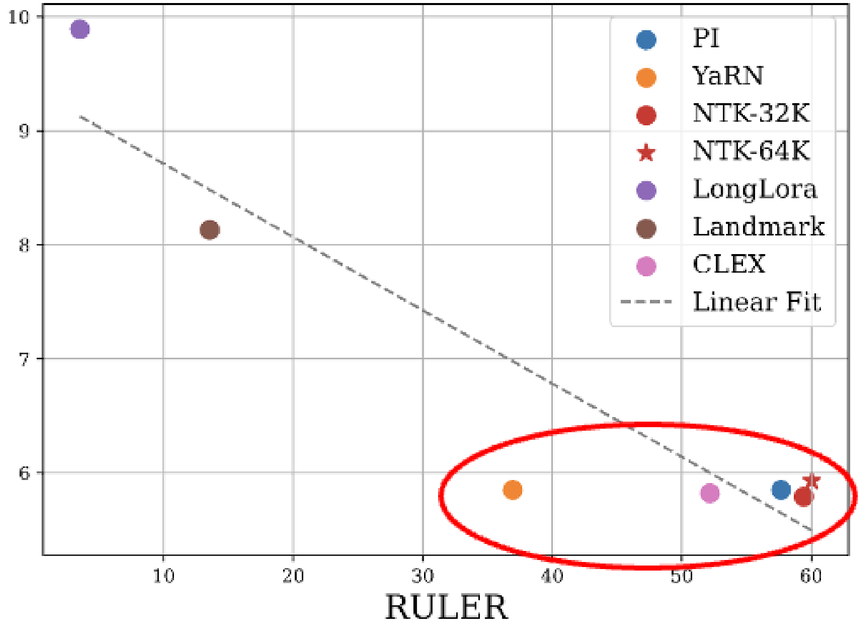
How to *evaluate* LCLMs?

- Prior work relies on perplexity…

See Fang et al., 2025 What is Wrong with Perplexity for Long-context Language Modeling?



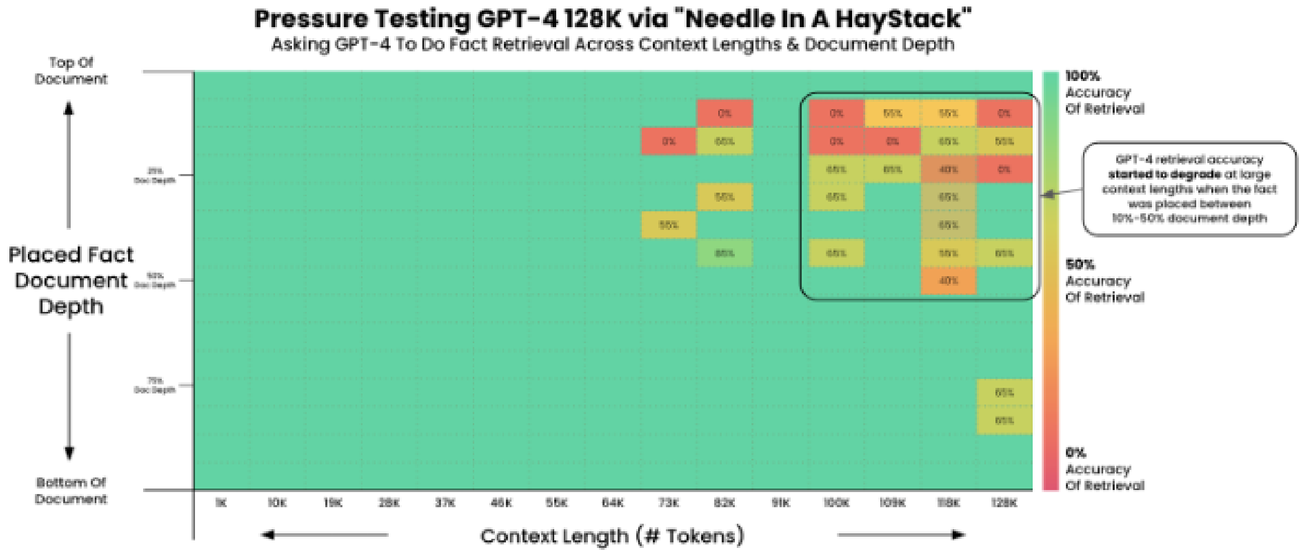
Fang et al., 2024. What is Wrong with Perplexity for Long-context Language Modeling?



Lu et al., 2024, A Controlled Study on Long Context Extension and Generalization in LLMs

How to *evaluate* LCLMs?

- Prior work relies on needle-in-haystack (NIAH)…



“Needle in a haystack”:

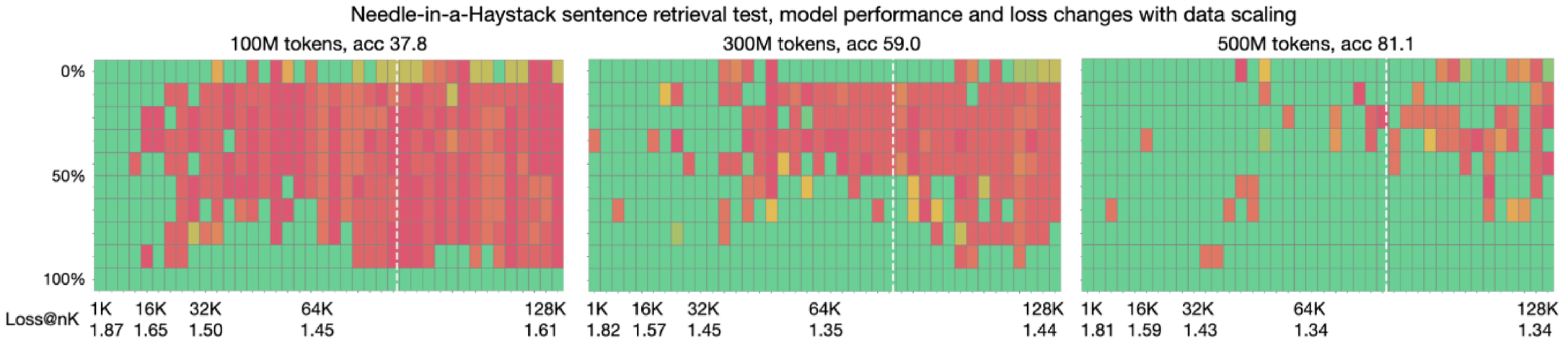
Finding a critical sentence (needle) from long, irrelevant text (haystack)

Gregory Kamradt. 2023. Needle In A Haystack - pressure testing LLMs. Github.

How to *evaluate* LCLMs?

- Prior work relies on needle-in-haystack (NIAH)…

Fu et al.: Training on long data is effective for learning Needle-in-a-Haystack and performance saturates quickly



Fu et al., 2024. Data Engineering for Scaling Language Models to 128K Context

We need holistic downstream eval for LCLMs

HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly

Howard Yen[p] Tianyu Gao[p] Minmin Hou[i] Ke Ding[i]
Daniel Fleischer[i] Peter Izsak[i] Moshe Wasserblat[i] Danqi Chen[p]
[p] Princeton Language and Intelligence, Princeton University [i]Intel
{hyen,tianyug,danqic}@cs.princeton.edu

(Yen et al., ICLR 2025)

- **Evaluation length** up to 128K
- **Diverse downstream tasks** (RAG as a core application)

What **long-context capabilities** to test?

- Recalling “needles”
- Reasoning over entire documents
- Robust to irrelevant contexts / noise
- Learning new tasks on the fly
- Instruction following

HELMET tasks 

Tasks

Synthetic recall ● ▲

Retrieval-augmented generation ● ▲ ▲

Passage re-ranking ● ● ▲ ▲

Many-shot in-context learning ● ▲ ▲

Long-document QA ● ▲

Long-document summarization ● ▲

Generation with citations (ALCE) ● ● ▲ ▲

Long-context capabilities

Recall ●

Reasoning and synthesis ●

Learning new tasks ●

Diverse task types

Controllable evaluation ▲

Real-world applications ▲

Yen et al., 2025: HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly

HELMET tasks

Tasks

Synthetic recall

Retrieval-augmented generation

Passage re-ranking

Many-shot in-context learning

Long-document QA

Long-document summarization

Generation with citations (ALCE)

Use the given documents to write a concise and short answer to the question. Write your answer in the following format: Answer: [answer]

Document (Title: Libby Mitchell): Elizabeth H. "Libby" Mitchell (born Elizabeth Anne Harrill)..

Document (Title: Eliot Cutler): On December 9, 2009, Cutler officially launched his campaign..

Question: What is the name of the independent candidate in Maine's 2010 gubernatorial race who finished ahead of Libby Mitchell?

Metric: Exact match

Yen et al., 2025: HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly

HELMET tasks

Tasks

- Synthetic recall
- Retrieval-augmented generation
- Passage re-ranking
- Many-shot in-context learning
- Long-document QA
- Long-document summarization
- Generation with citations (ALCE)

Rank each document based on their relevance to the question in descending order …
Ranking: ID3 > ID1 > ID2

[ID: 4842895] Document: RSA Security is a United States-based …

[ID: 1929910] Document: Definition - What does RSA Encryption mean? RSA encryption … …

Query: rsa definition key
Ranking: 1929910 > 4842895 > 9384722 > …

Metric: NDCG@10

Yen et al., 2025: HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly

HELMET tasks

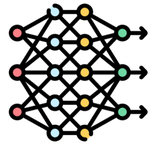
Tasks

- Synthetic recall
- Retrieval-augmented generation
- Passage re-ranking
- Many-shot in-context learning
- Long-document QA
- Long-document summarization
- Generation with citations (ALCE)

Article: Nasdaq dropped 3% today
Label: **3**

Article: The 30 best TV shows to watch now
Label: **5**
…

Article: The new Apple TV+ show “Severance”
Label: _____



“5”

Using **abstract** (e.g., numbers) labels in in-context learning better reflects “learning” [Pan et al., ACL’23 Findings]

Yen et al., 2025: HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly

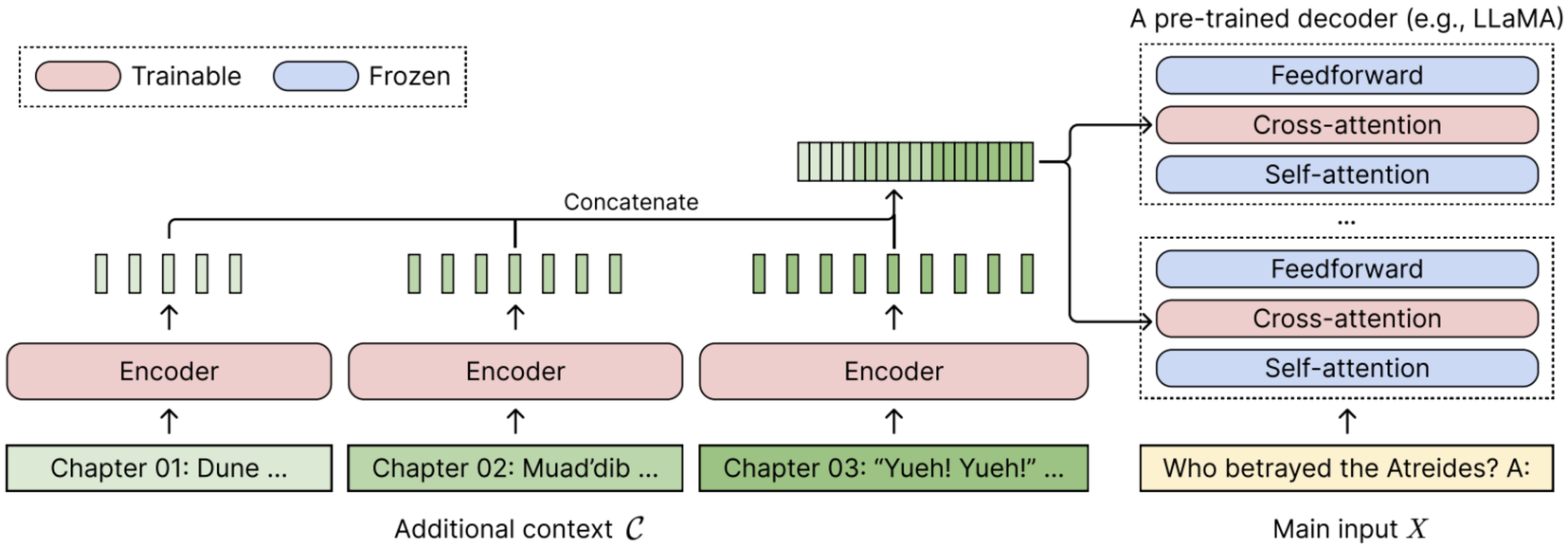
Evaluating long-context LMs on HELMET

	Recall					RAG					Cite					Re-rank				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4	99.5	93.5	93.1	88.6	72.8	75.3	73.6	70.9	68.1	65.0	43.8	45.2	28.8	3.6	3.1	76.4	72.3	63.9	37.8	16.8
GPT-4o-05	94.7	93.4	91.2	87.9	81.6	74.1	73.1	71.8	71.1	71.0	43.7	44.2	44.1	44.1	40.6	74.4	74.3	67.2	56.9	46.8
GPT-4o-08	99.8	99.4	97.9	97.0	97.0	73.4	73.8	72.4	71.1	70.8	45.8	47.1	46.4	45.7	45.3	75.6	73.1	67.4	59.5	47.9
GPT-4o-mini	100.0	99.8	99.1	92.0	83.6	72.6	71.0	69.6	68.3	66.7	36.1	33.7	31.3	28.0	24.5	68.9	65.2	56.4	40.5	30.5
Claude-3.5-sonnet	99.9	97.2	96.2	95.2	93.3	60.4	52.8	51.1	39.8	41.1	36.7	32.9	30.5	26.4	12.5	76.3	46.1	36.0	14.5	9.1
Gemini-1.5-Flash	93.5	93.6	93.2	92.5	87.8	71.6	69.9	69.6	68.6	67.6	48.4	46.6	43.0	36.7	29.0	75.1	73.9	68.9	59.3	50.7
Gemini-1.5-Pro	81.3	83.6	86.9	87.1	84.1	73.0	72.9	71.6	71.9	70.9	47.1	43.0	44.7	45.1	42.5	75.8	73.2	71.7	65.9	58.6
Llama-3.1-8B	99.4	99.6	97.2	98.3	91.1	69.1	67.9	64.8	64.6	59.0	35.4	26.9	12.6	12.8	3.4	58.7	45.9	42.0	31.9	15.0
Llama-3.1-70B	99.9	99.8	98.0	87.4	84.4	73.0	72.2	71.5	70.3	55.8	44.5	42.1	39.5	30.9	7.6	73.3	69.7	58.4	40.0	19.4
Mistral-Nemo	93.6	83.3	52.3	21.5	12.1	68.4	63.6	56.9	47.6	39.9	33.7	8.6	3.7	1.3	0.5	56.8	46.0	13.1	0.0	0.0
MegaBeam-Mistral	93.9	90.0	81.6	83.6	76.0	62.6	62.6	61.8	57.4	55.2	22.3	13.8	9.7	4.5	4.0	49.9	36.2	34.2	21.7	15.9
Phi-3-mini-128k	90.3	84.9	81.1	80.1	42.3	61.2	60.6	57.9	55.7	46.0	22.8	16.9	9.3	2.7	0.8	44.1	28.7	25.6	16.6	5.8
Phi-3-small-128k	91.0	89.3	73.5	66.7	59.0	66.5	65.8	62.5	61.3	58.1	18.9	15.9	8.9	4.6	2.9	38.3	32.1	28.1	17.2	6.5
Phi-3-med-128k	76.1	70.6	62.5	51.8	14.4	65.3	64.5	62.7	56.9	45.2	39.1	27.1	10.2	5.8	3.3	43.2	33.3	25.5	11.9	5.8
Phi-3.5-mini	95.8	90.7	83.1	77.2	40.7	59.8	57.9	55.6	51.0	41.4	22.1	17.2	7.1	2.0	2.5	42.4	29.6	23.2	18.0	9.1
Jamba-1.5-Mini	87.3	85.6	85.0	79.7	76.8	66.2	65.0	64.0	63.4	56.6	15.4	10.0	5.7	3.1	2.5	53.5	43.0	35.6	23.2	14.6
ProLong	96.6	95.8	93.0	92.9	85.5	68.8	69.3	66.6	66.5	64.8	33.7	24.0	11.0	2.3	1.2	53.8	43.9	39.3	33.3	25.0

Yen et al., 2025: HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly

A recipe for training open-source LCLMs

- **Fact #1:** Position extrapolation methods with zero or minimal training don’t work.
- **Fact #2:** Efficient Transformers or modifications don’t work well.



Yen et al. ACL 2024: Long-Context Language Modeling with Parallel Context Encoding

A recipe for training open-source LCLMs

- **Solution:** Continual pre-training on long-context data with full attention

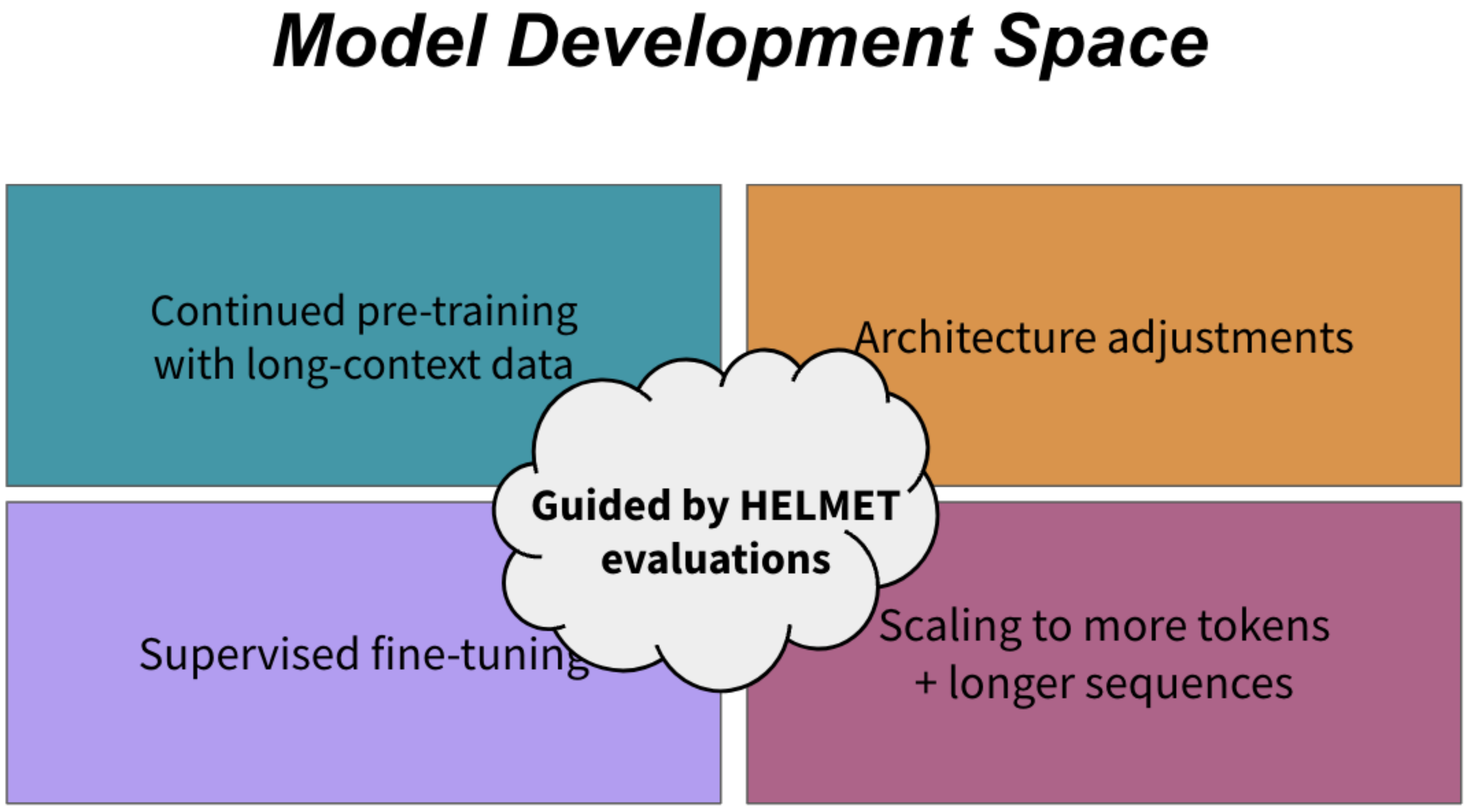
How to Train Long-Context Language Models (Effectively)

Tianyu Gao* **Alexander Wettig*** **Howard Yen** **Danqi Chen**
Princeton Language and Intelligence, Princeton University
{tianyug,awettig,hyen,danqic}@cs.princeton.edu

(Gao et al., 2024)

A recipe for training open-source LCLMs

- **Solution:** Continual pre-training on long-context data with full attention



Gao et al., 2024: How to Train Long-Context Language Models (Effectively)

ProLong: Training recipe

- ProLong: an 8B open model supporting context size of 512K tokens
- Start from Llama-3-8B (8K context size)
- Two stages of continual pre-training
  - Stage 1: 20B tokens (64K length), Stage 2: 20B tokens (512K tokens)
  - **Mix of high-quality long-context and short-context data**
  - **Where does long-context data come from?** Code + books
  - **Architectural adjustments:** 1) Full attention with cross-document attention masking; 2) Tune the base frequency for RoPE embeddings
- SFT on short-context instruction data (UltraChat)

Gao et al., 2024: How to Train Long-Context Language Models (Effectively)

ProLong: Training recipe

ProLong achieves the strongest performance among <10B models on HELMET-128K eval

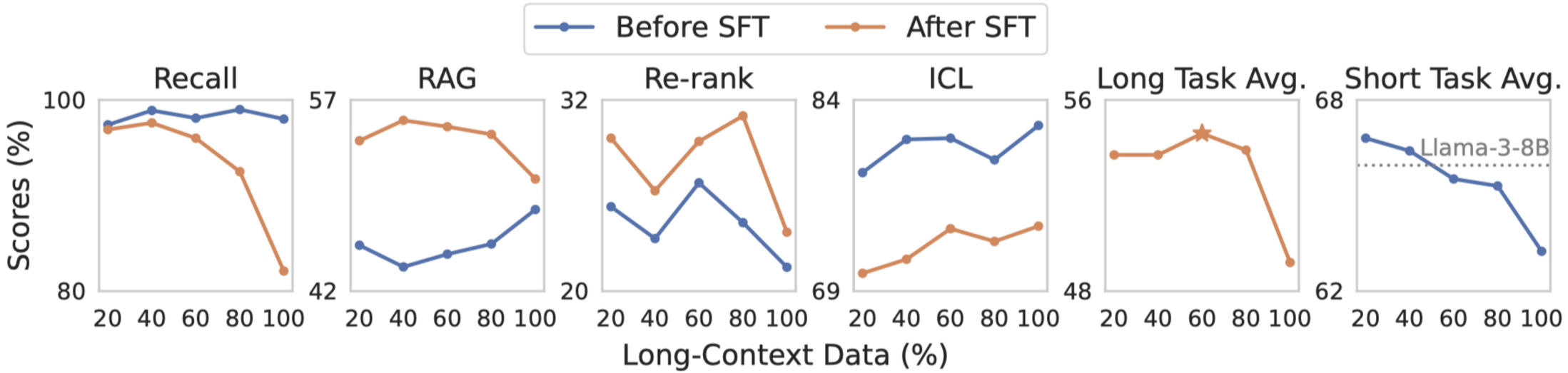
- Only 5% of training tokens compared to Llama-3.1

Model	Max Len.	Recall	RAG	ICL	Re-rank	QA	Summ.	Avg.
ProLong (8B)	512K	**99.4**	**66.0**	81.1	**33.2**	**40.8**	40.5	**60.2**
MegaBeam-Mistral (7B)	512K	**99.4**	58.1	**82.1**	22.1	33.7	43.6	56.5
Meta-Llama-3.1 (8B)	128K	98.7	62.8	79.7	26.6	40.4	**46.1**	59.0
Qwen2 (7B)	128K	34.4	43.4	54.8	4.6	23.3	38.5	33.2
Phi-3-small (7B)	128K	74.8	60.6	82.0	18.5	34.1	42.4	52.1
Mistral-Nemo (12B)	128K	24.9	48.1	82.0	4.7	37.7	37.0	39.1

Gao et al., 2024: How to Train Long-Context Language Models (Effectively)

Finding #1: mixing long and short-context data

- Mixing **long-context (60%)** vs **short-context (40%)**
- **Training on long-context data hurts performance on long-context tasks!**
- The quality of long-context and short-context data both matter



Gao et al., 2024: How to Train Long-Context Language Models (Effectively)

Finding #2: training on longer context than eval length helps

Stage 1: training on 64K

Stage 2: training on 512K

Evaluated at 64K

Max Seq. Length	Recall	RAG	Re-rank	ICL
ProLong 64K training (20B)	96.5	52.7	22.8	70.6
+4B 64K training	95.0	56.4	28.0	78.8
+4B 512K training	**98.5**	**56.9**	**32.9**	**79.2**

Gao et al., 2024: How to Train Long-Context Language Models (Effectively)

Finding #3: SFT on short instruction data is “good” enough

- Continual pre-training on mix of long + short data
- SFT on **short instruction data** (UltraChat) provides LCLMs instruction following

(How to obtain high-quality long instruction data is an open question)

Gao et al., 2024: How to Train Long-Context Language Models (Effectively)

Shifting LCLM research from input to output

LONGPROC: Benchmarking Long-Context Language Models on Long Procedural Generation

Xi Ye[♠] **Fangcong Yin**[◇*] **Yinghui He**[♠*] **Joie Zhang**[♠*] **Howard Yen**[♠*]
Tianyu Gao[♠] **Greg Durrett**[◇] **Danqi Chen**[♠]
[♠] Princeton Language and Intelligence [◇] The University of Texas at Austin
xi.ye@princeton.edu

(Ye et al., 2025)

Key idea: **long procedure generation**

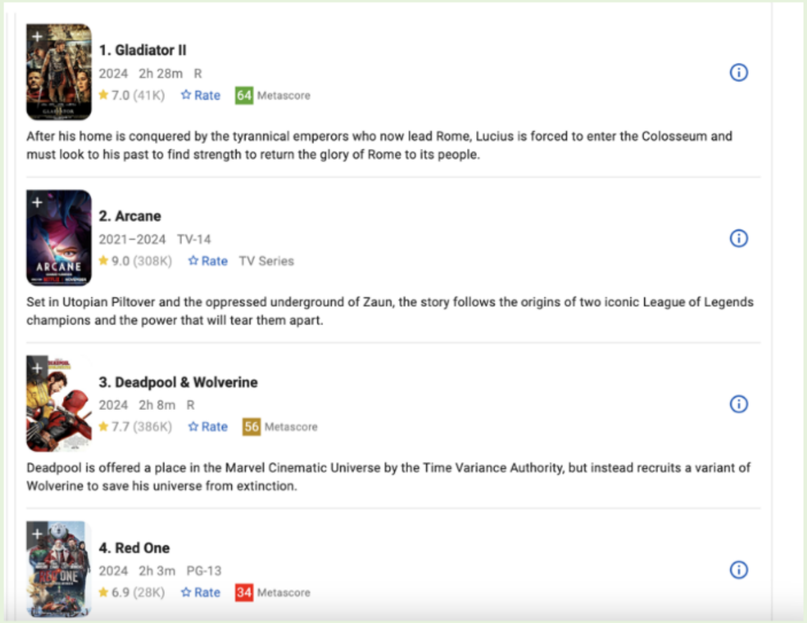
- Required to execute a **specified procedure** (instruction following)
- **Step-by-step generation**: depend on input and previous generations (dispersed information)
- **Long** but **structured** outputs (reliable rule-based evaluation)
- **Controllable** output lengths

LongProc task: HTML to TSV

HTML to TSV

Extract specified information from HTML pages and structure it into a table format

Find the title, year, genre, and rate of the movies and list them in a TSV.



Title	Year	Genre	Rate
Gladiator II	2024	Action, Adventure	7.0
Arcane	2021-2024	Animation, Action	9.0
Deadpool & Wolverine	2024	Action, Adventure	7.7
Red One	2024	Adventure, Comedy	6.9
Lioness	2023	Action, Thriller	7.7
.....			

Ye et al., 2025: LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation

LongProc task: Countdown

Countdown Combine a set of numbers with basic arithmetic operations to reach a target number

[TASK] Find a way to use all four numbers exactly once, along with the basic operations (+, -, *, /), to reach the target number.
Numbers: [9, 16, 6, 18] **Target: 12**

- A generalized version of game of 24
- Execute a **depth-first-search** procedure
- Evaluation: only the final solution

Initial number set: [9, 16, 6, 18], target: 12. Options for choosing two numbers: [(9, 16), (9, 6), (9, 18), (16, 6), (16, 18), (6, 18)].
|- Pick two numbers (9, 16) (numbers left: [6, 18]). Try possible operations.
|- Try 16 + 9 = 25. Add 25 to the number set. **Current number set: [25, 6, 18]**, target: 12. Options for choosing two numbers: [(25, 6), (25, 18), (6, 18)].
|- Pick two numbers (25, 6) (numbers left: [18]). Try possible operations.
|- Try 25 + 6 = 31. Add 31 to the number set. Current number set: [31, 18], target: 12.
|- Try 31 + 18 = 49. Evaluate 49 != 12, drop this branch.
|- Try 31 - 18 = 13. Evaluate 13 != 12, drop this branch.
|- Try 31 * 18 = 558. Evaluate 558 != 12, drop this branch.
|- Try 31 / 18 = 1.7. 1.7 is a decimal, drop this branch.
|- Try 25 - 6 = 19. Add 19 to the number set. **Current number set: [19, 18]**, target: 12.
.....
|- Try 16 * 9 = 144. Add 144 to the number set. **Current number set: [144, 6, 18]**, target: 12. Options for choosing two numbers: [(144, 6), (144, 18), (6, 18)].
.....
|- Pick two numbers (6, 18) (numbers left: [144]). Try possible operations.
.....
|- Try 18 - 6 = 12. Add 12 to the number set. **Current number set: [12, 144]**, target: 12
|- Try 144 + 12 = 156. Evaluate 156 != 12, drop this branch.
|- Try 144 - 12 = 132. Evaluate 132 != 12, drop this branch.
|- Try 144 * 12 = 1728. Evaluate 1728 != 12, drop this branch.
|- **Try 144 / 12 = 12. Evaluate 12 == 12, target found!**

[SOLUTION]
16 * 9 = 144; 18 - 6 = 12; 144 / 12 = 12

Ye et al., 2025: LongProc: Benchmarking Long-Context Language Models on Long Procedural Generation

LongProc tasks

LONGPROC: Benchmarking Long-Context Language Models on Long Procedural Generation

Xi Ye[♠] **Fangcong Yin**[◇*] **Yinghui He**[♠*] **Joie Zhang**[♠*] **Howard Yen**[♠*]
Tianyu Gao[♠] **Greg Durrett**[◇] **Danqi Chen**[♠]
[♠] Princeton Language and Intelligence [◇] The University of Texas at Austin
xi.ye@princeton.edu

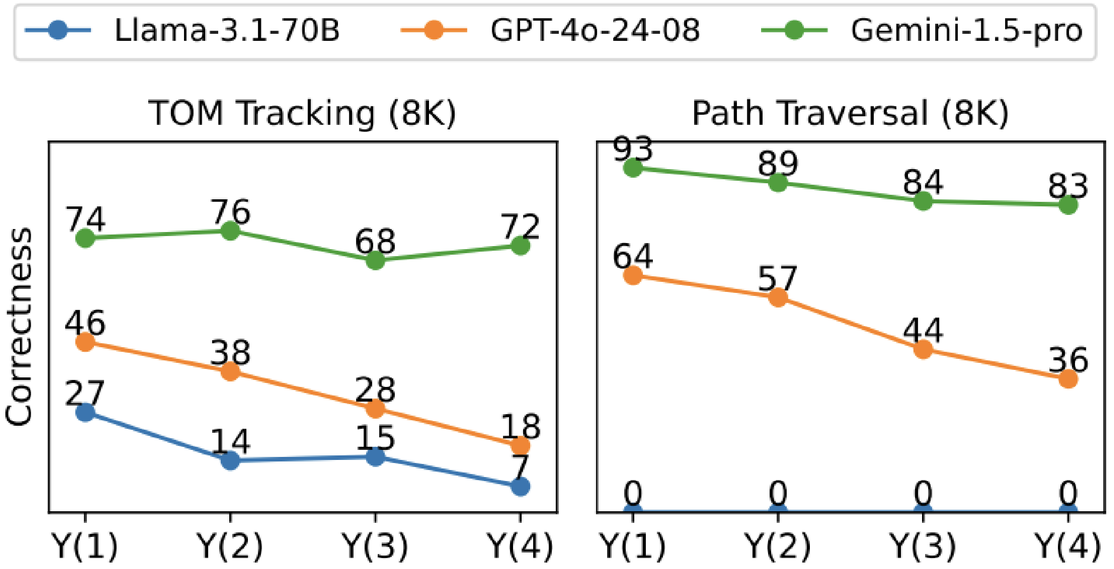
(Ye et al., 2025)

	0.5K Level			2K Level			8K Level			Access	Deductive	Exec
	N	# In	# Out	N	# In	# Out	N	# In	# Out	Info	Reasoning	Search
HTML to TSV	100	12K	0.5K	189	23K	1.3K	120	38K	3.7K	√(sequential)	–	–
Pseudocode to Code	100	0.4K	0.3K	100	0.9K	0.7K	–	–	–	√(sequential)	–	–
Path Traversal	100	1.2K	0.5K	100	4.8K	2.0K	100	12K	5.8K	√(targeted)	–	–
ToM Tracking	100	2.0K	0.5K	100	2.5k	2.0K	100	4.1K	7.9K	√(sequential)	√	–
Countdown	100	5.6K	0.5K	100	5.6K	1.7K	100	5.6K	6.5K	–	√	√
Travel Planning	–	–	–	100	6.0K	1.2K	100	6.0K	5.3K	√(targeted)	√	√

LongProc: interesting findings

- Frontier models (128K or longer context size) fail at 8K output length
- Reasoning models outperform instruction-tuned models significantly

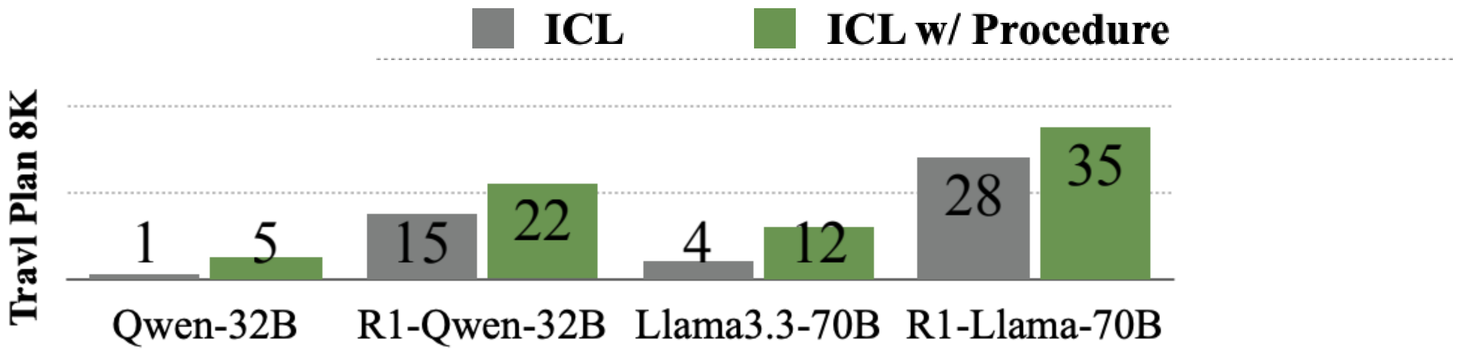
  e.g., R1-Distill-Qwen2.5-32B vs Qwen2.5-32B-Inst
- LCLMs fail more at later positions of long output (prefix prefilled)



LongProc: interesting findings

- Frontier models (128K or longer context size) fail at 8K output length
- Reasoning models outperform instruction-tuned models significantly

  e.g., R1-Distill-Qwen2.5-32B vs Qwen2.5-32B-Inst
- LCLMs fail more at later positions of long output (prefix prefilled)
- Reasoning models still benefit from pre-specified procedure



The end

Retrieval-augmented language models (RALMs)

Long-context language models (LCLMs)

- A lot of progress has been made in LCLMs – they are also extremely useful for RAG and agentic search.
- Research focus may be shifted from long input to long output soon
- Progress in training RALMs has stalled. Still, I hope to see more breakthroughs in the near future.

Thank you!