

Retrieval Augmentation in LLMs

CSE 5525: Foundations of Speech and Language Processing



THE OHIO STATE UNIVERSITY

Bernal Jiménez Gutiérrez (jimenezgutierrez.1@osu.edu)

LLMs are Everywhere

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score
1	1	Gemini-2.5-Pro-Exp-03-25	1440
2	2	ChatGPT-4o-latest (2025-03-26)	1406
2	4	Grok-3-Preview-02-24	1404
2	2	GPT-4.5-Preview	1398
5	7	Gemini-2.0-Flash-Thinking-Exp-01-21	
5	4	Gemini-2.0-Pro-Exp-02-05	
5	4	DeepSeek-V3-0324	
7	5	DeepSeek-R1	
8	13	Gemini-2.0-Flash-001	
8	4	o1-2024-12-17	
11	13	Qwen2.5-Max	
11	13	Gemma-3-27B-it	
11	10	o1-preview	
14	13	o3-mini-high	
14	15	DeepSeek-V3	

anthropic chinese google meta microsoft

100 MMLU

89.8 = human expert

80

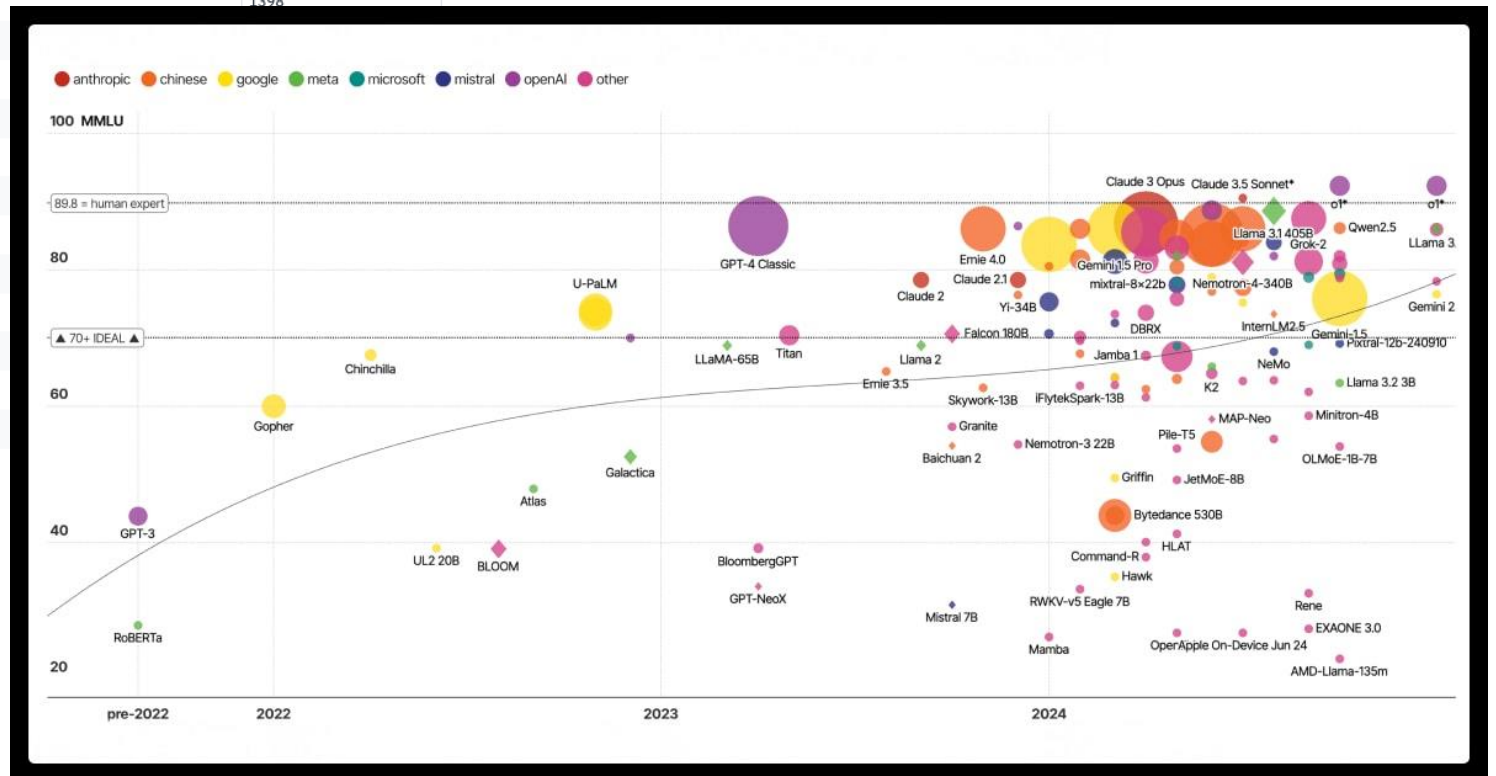
70+ IDEAL

60

Chinchilla

Gopher

<https://lmarena.ai/>



<https://labelyourdata.com/articles/llm-model-size>

LLMs are Incredibly Powerful

Exam results (ordered by GPT-3.5 performance)

Estimated percentile lower bound (among test takers)

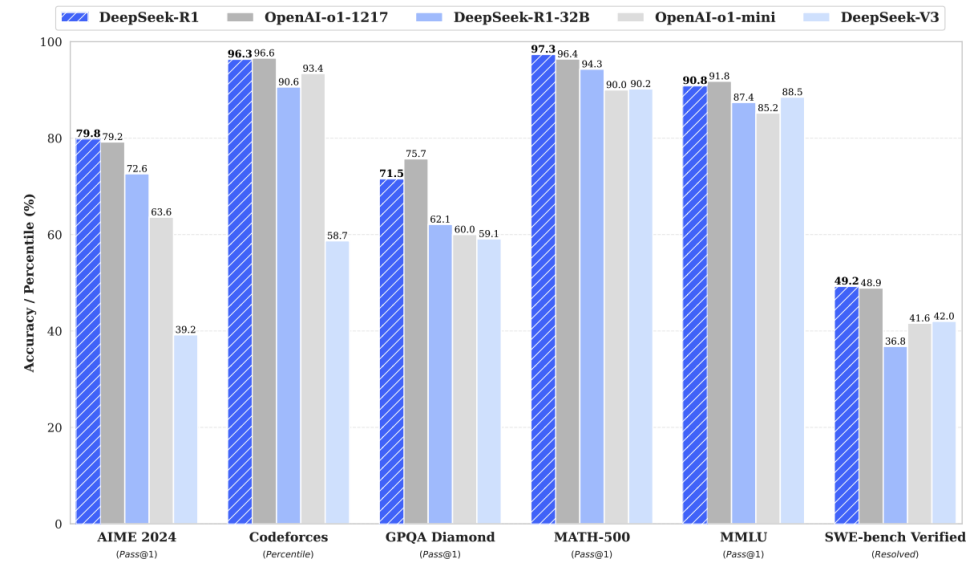
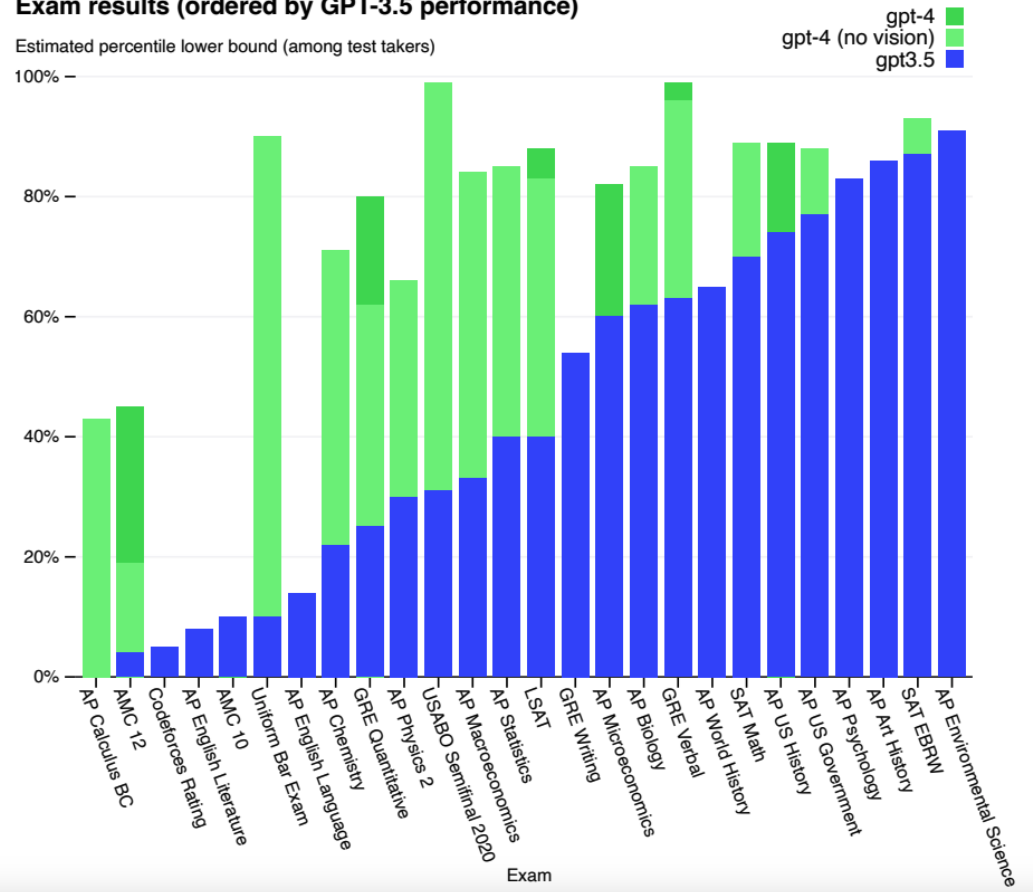


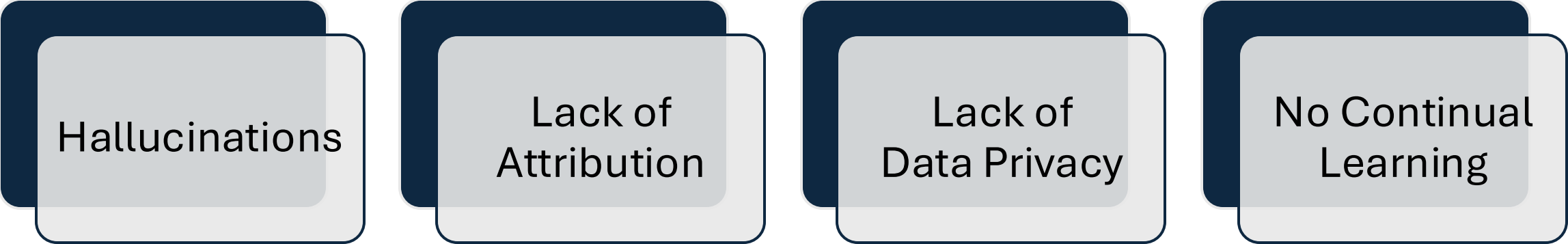
Figure 1 | Benchmark performance of DeepSeek-R1.

<https://arxiv.org/pdf/2501.12948>

<https://arxiv.org/pdf/2303.08774>

However, LLMs have some
fundamental limitations...

Fundamental LLM Limitations



Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

Fundamental LLM Limitations

Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

GPT-4o

User

Where did NLP researcher Sachin Kumar work after getting his PhD?

Assistant

After earning his PhD, NLP researcher Sachin Kumar worked as a Research Scientist at IBM Research.

Sachin Kumar, an NLP researcher, worked at Microsoft after obtaining his PhD.

Sachin Kumar worked at Amazon AWS AI after obtaining his PhD in Natural Language Processing.

Sachin Kumar, after obtaining his PhD, worked at Google AI. If you have more specific questions about his work or projects, feel free to ask!

Fundamental LLM Limitations

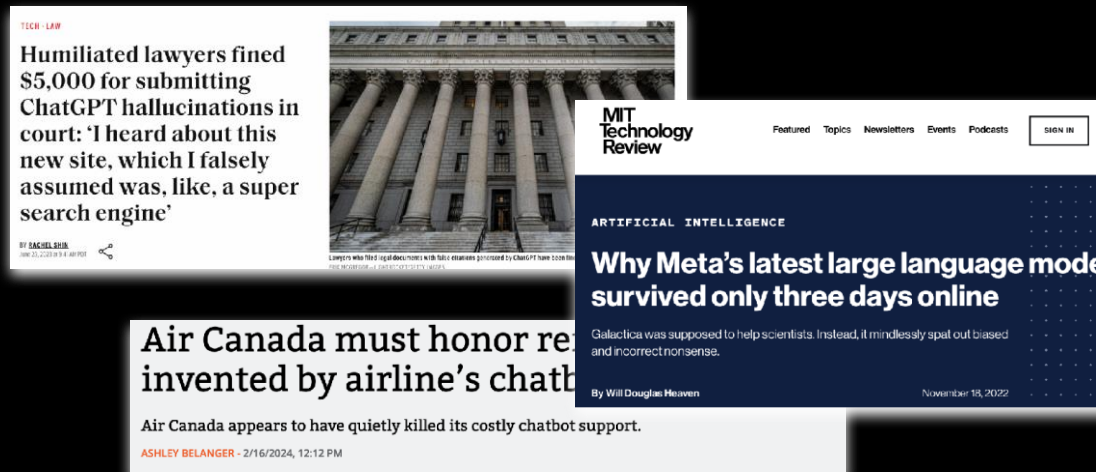
Hallucinations

Lack of
Attribution

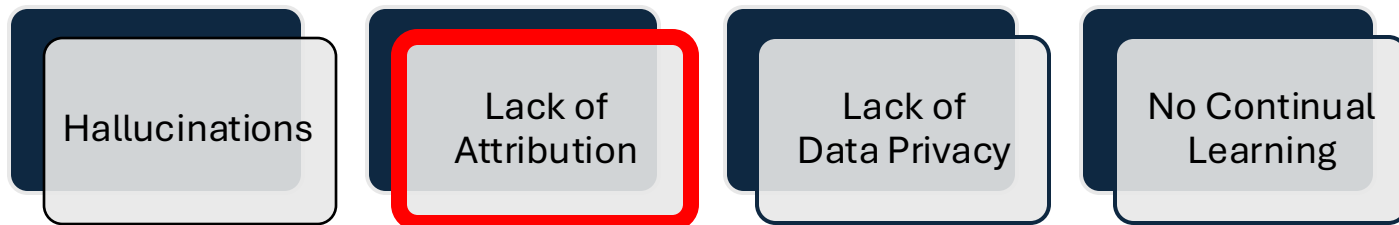
Lack of
Data Privacy

No Continual
Learning

Catastrophic Errors as Results of LM Hallucinations



Fundamental LLM Limitations



GPT-4o

User

What are some of NLP researcher Sachin Kumar's most famous papers?

Assistant

Sachin Kumar is a notable researcher in the field of Natural Language Processing (NLP). Some of his well-known papers include:

1. **"Dialogue Natural Language Inference"** - This paper explores the task of natural language inference in the context of dialogues, aiming to understand entailment and contradiction in conversational settings.
2. **"A Framework for Modeling Information Cascades in Online Media"** - This work presents a framework for understanding how information spreads across online platforms, which is crucial for analyzing trends and misinformation.
3. **"Tree Structured Decoding for Solving Math Word Problems"** - This paper introduces a novel approach to solving math word problems using tree-structured decoding, enhancing the ability of models to understand and solve complex problems.

These papers reflect his contributions to advancing the understanding and capabilities of NLP systems in various contexts.

???

- #1 and #3 are **not** from Sachin
- #2 is **not** real!

Fundamental LLM Limitations

Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

ARTIFICIAL INTELLIGENCE • MEDIA & ENTERTAINMENT TECH • POLITICS AND REGULATION

Stealing Ghibli: How AI Has Crossed The Line Into Creative Theft

Aron Solomon April 2, 2025



Case 1:23-cv-11195 Document 1 Filed 12/27/23 Page 1 of 69

B. Defendants' GenAI Products

1. A Business Model Based on Mass Copyright Infringement

57. Despite its early promises of altruism, OpenAI quickly became a multi-billion-dollar for-profit business built in large part on the unlicensed exploitation of copyrighted works belonging to The Times and others. Just three years after its founding, OpenAI shed its exclusively

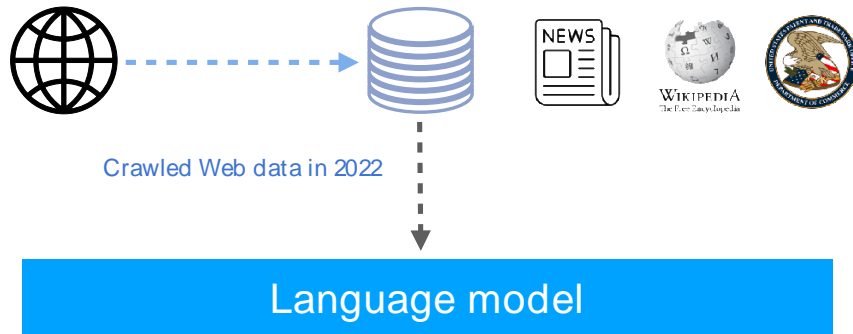
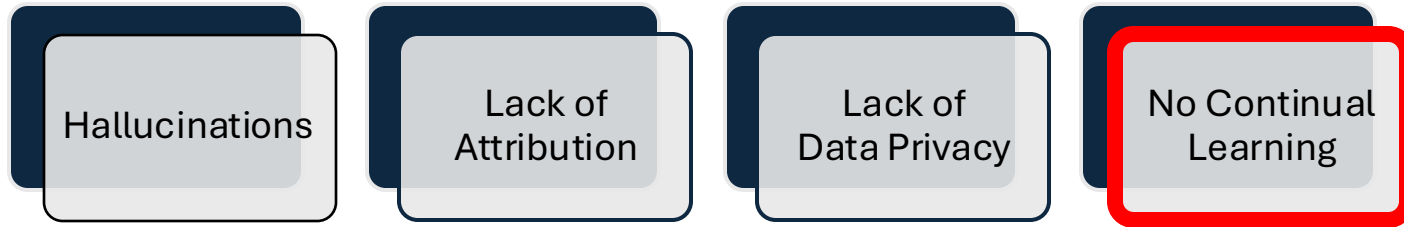
Plaintiff The New York Times Company ("The Times"), by its attorneys Susan Godfrey LLP and Rothwell, Figg, Ernst & Manbeck, P.C., for its complaint against Defendants Microsoft Corporation ("Microsoft") and OpenAI, Inc., OpenAI LLP, OpenAI GP LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, (collectively "OpenAI" and, with Microsoft, "Defendants"), alleges as follows:

1. NATURE OF THE ACTION

1. Independent journalism is vital to our democracy. It is also increasingly rare and valuable. For more than 170 years, The Times has given the world deeply reported, expert, independent journalism. Times journalists go where the story is, often at great risk and cost, to inform the public about important and pressing issues. They bear witness to conflict and disasters, provide accountability for the use of power, and illuminate truths that would otherwise go unseen. Their essential work is made possible through the efforts of a large and expensive organization that provides legal, security, and operational support, as well as editors who ensure their journalism meets the highest standards of accuracy and fairness. This work has always been important. But

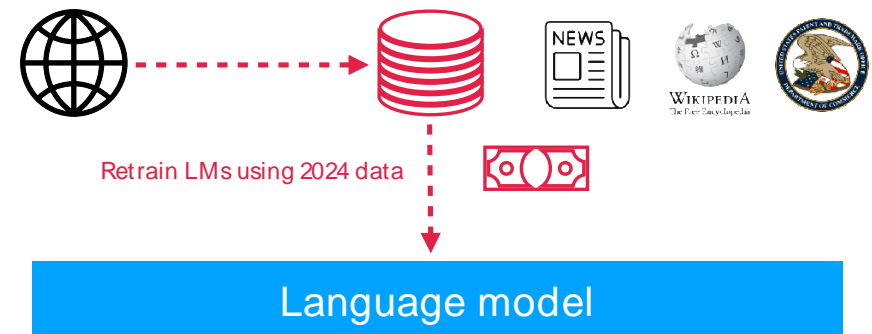
New York Times lawsuits
against OpenAI

Fundamental LLM Limitations



ChatGPT

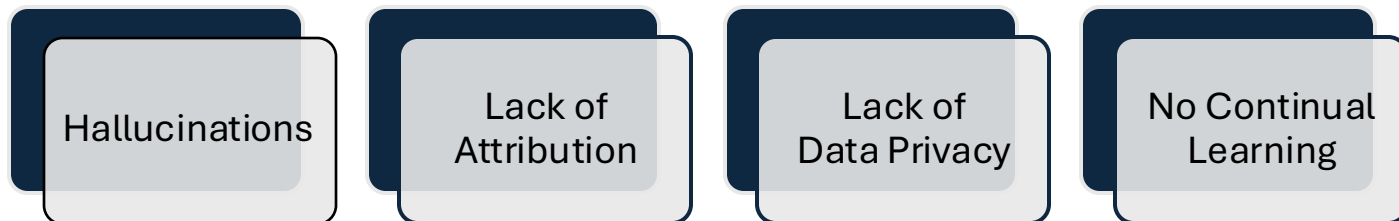
I'm sorry, but I don't have access to real-time information including events beyond January 2022.



ChatGPT

I'm sorry, but I don't have access to real-time information including events beyond January 2022.

Fundamental LLM Limitations



- Many techniques have been developed to update the internal knowledge of LLMs such as **model editing** and **continual pretraining**, potentially addressing the **data privacy** and **continual learning** problems.
- **However, these techniques struggle with both performance and scalability (MQuAKE, EvolvingQA).**

Lecture Overview

- What is retrieval-augmented generation (RAG)?
- Why do we need retrieval-augmented generation (RAG)?
- RAG: Architecture and Training
- Open Questions
- Beyond RAG: LLM Continual Learning

Lecture Overview

- What is retrieval-augmented generation (RAG)?
- Why do we need retrieval-augmented generation (RAG)?
- RAG: Architecture and Training
- Open Questions
- Beyond RAG: LLM Continual Learning

What is Retrieval-Augmented Generation (RAG)?

Standard LLM Generation



The capital city of Ontario is **Toronto**



LM

Training time

The capital city of Ontario is _____



LM

Test time

What is Retrieval-Augmented Generation (RAG)?

Retrieval-Augmented Generation



The capital city of Ontario is **Toronto**



LM

Training time



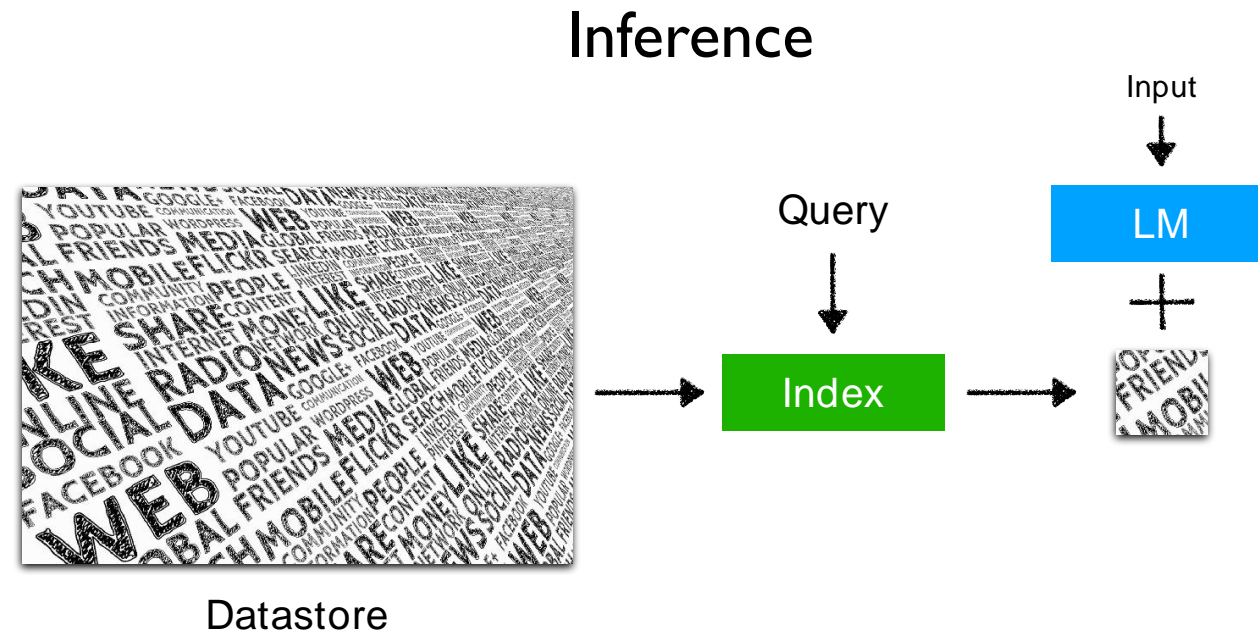
The capital city of Ontario is _____



LM

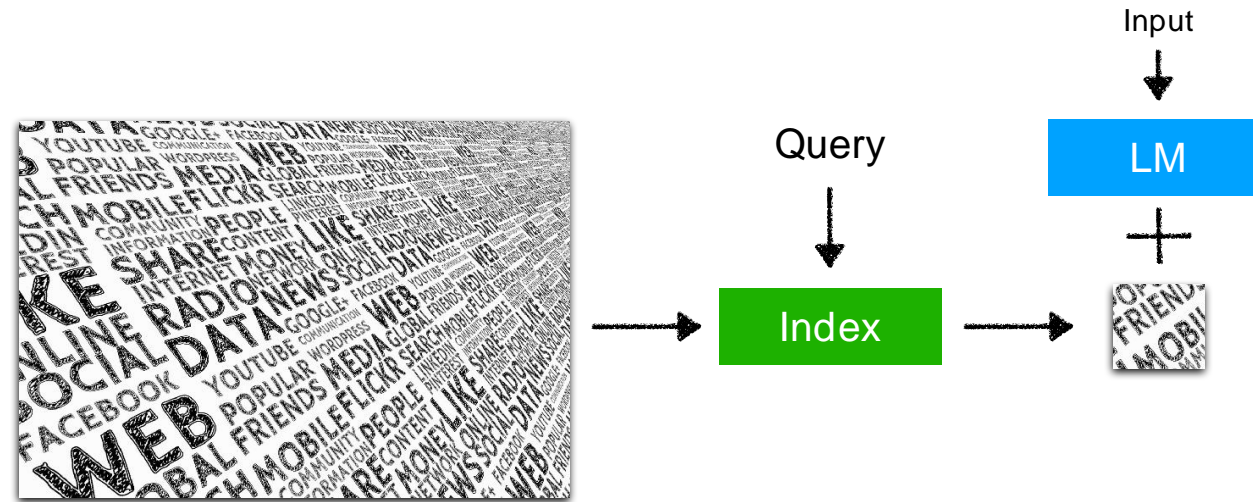
Test time

What is Retrieval-Augmented Generation (RAG)?



What is Retrieval-Augmented Generation (RAG)?

Inference: Datastore



Datastore

Raw text corpus

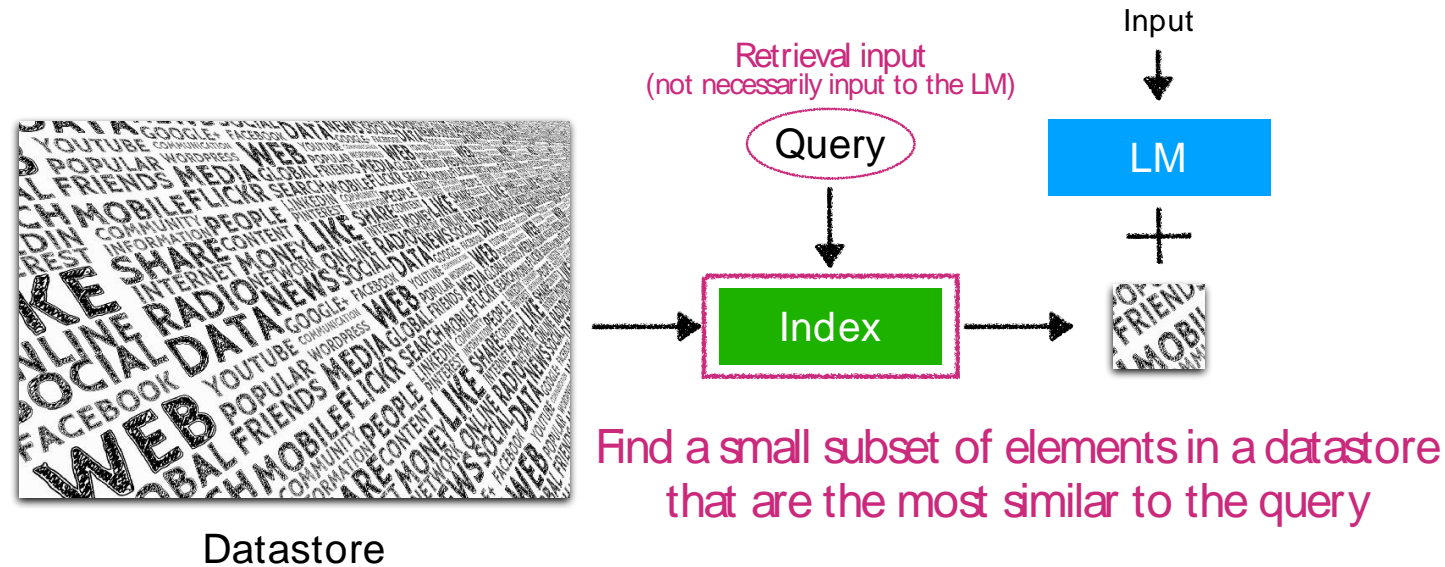
At least billions~trillions of tokens

Not labeled datasets

Not structured data (knowledge bases)

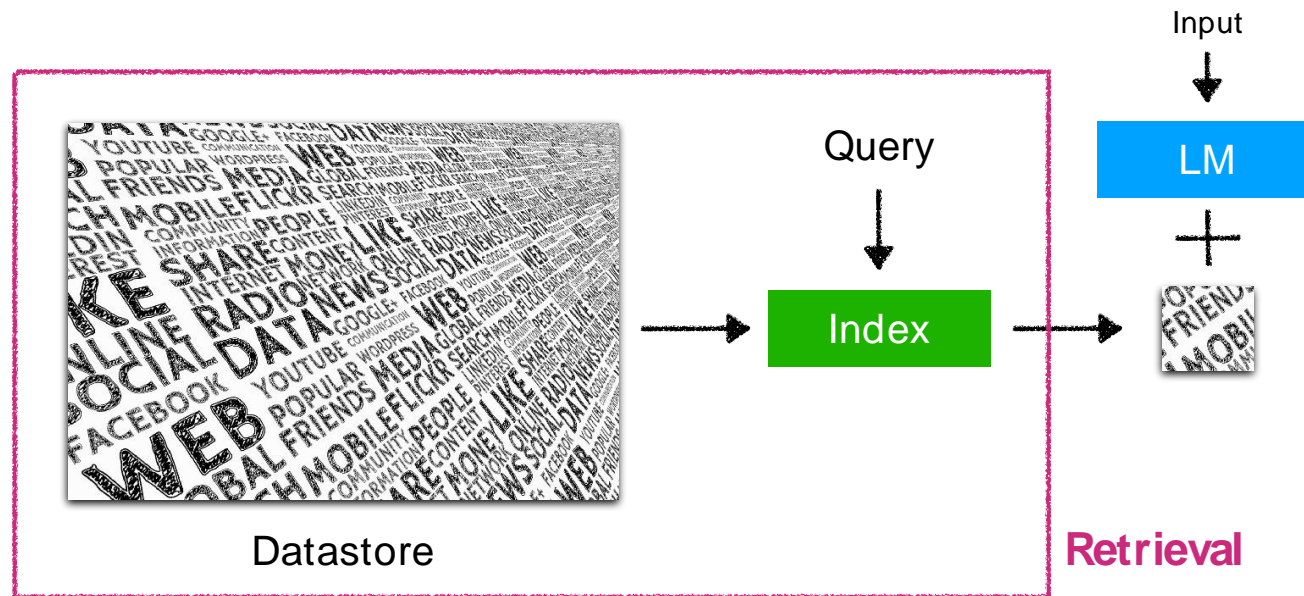
What is Retrieval-Augmented Generation (RAG)?

Inference: Index



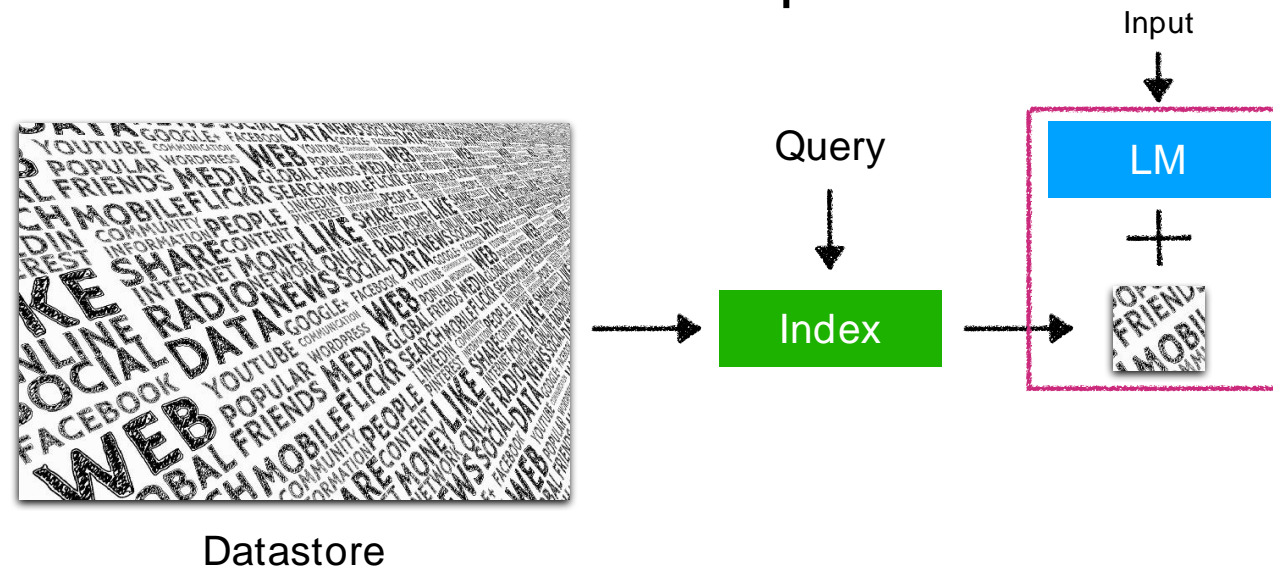
What is Retrieval-Augmented Generation (RAG)?

Inference: Search



What is Retrieval-Augmented Generation (RAG)?

Inference: Incorporation



Datastore

Lecture Overview

- What is retrieval-augmented generation (RAG)?
- Why do we need retrieval-augmented generation (RAG)?
- RAG: Architecture and Training
- Open Questions
- Beyond RAG: LLM Continual Learning

Why do we need retrieval-augmented generation (RAG)?

- **Answer:** RAG addresses all of these!



Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

RAG Reduces Hallucination & Grants Attribution

Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

GPT-4o

User

Where did NLP researcher Sachin Kumar work after getting his PhD?

Retrieved from Sachin Kumar's website:

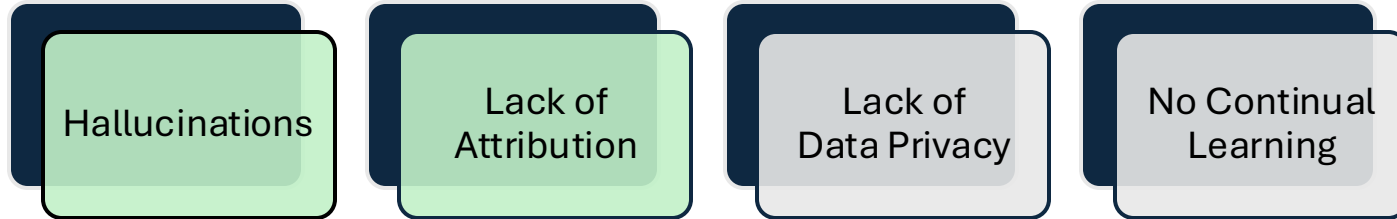
I was a postdoctoral researcher at the Allen Institute for AI (AI2) from August 2023-August 2024. I obtained my Ph.D. at the Language Technologies Institute at Carnegie Mellon University (CMU) in August 2023 with the final two years of my PhD spent visiting the University of Washington in Seattle. ""

Assistant

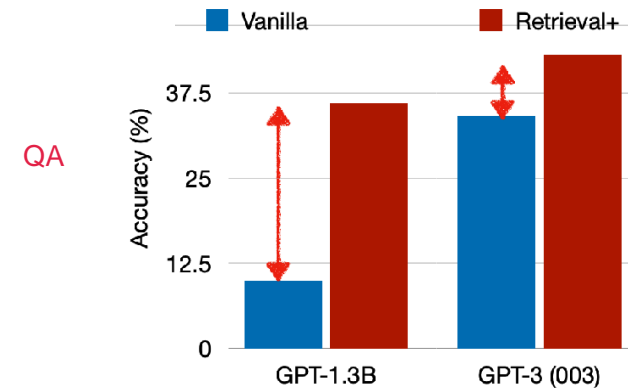
After obtaining his Ph.D., Sachin Kumar worked as a postdoctoral researcher at the Allen Institute for AI (AI2) from August 2023 to August 2024.



RAG Reduces Hallucination & Grants Attribution



Significant improvements across model scale, with larger gain with smaller LMs



Fundamental LLM Limitations

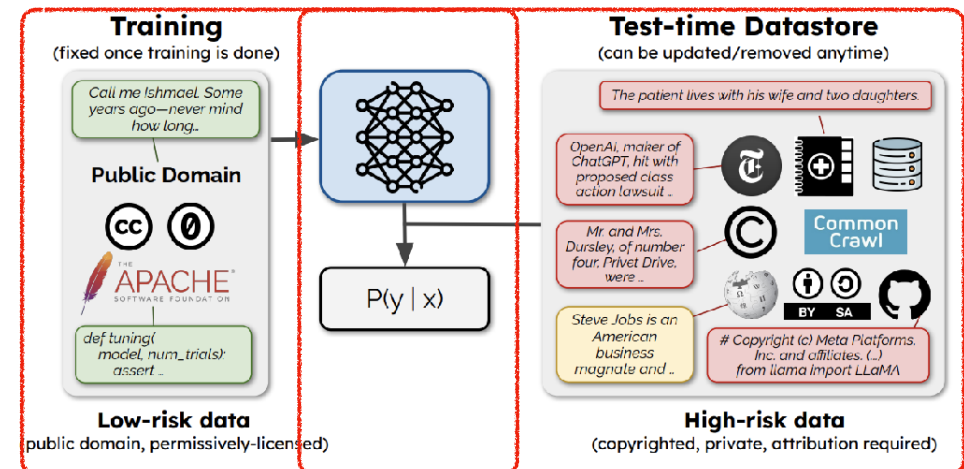
Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

Segregating copyright-sensitive data from
pre-training data



Min* and Gururangan* et al., SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. ICLR 2024.

Fundamental LLM Limitations

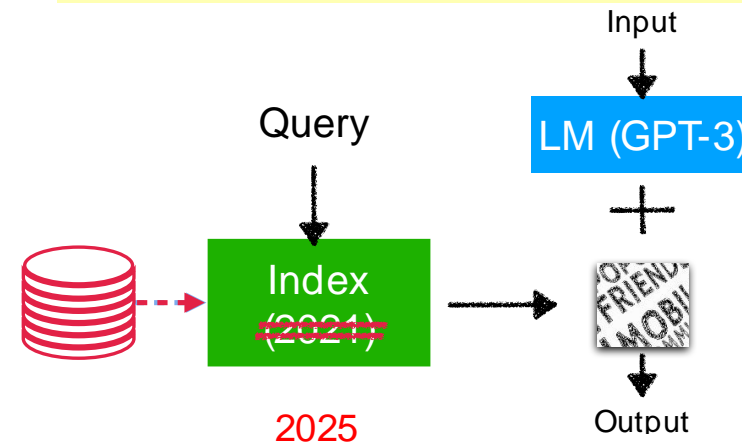
Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

Replacing datastores to catch up dynamically
changing world without re-training

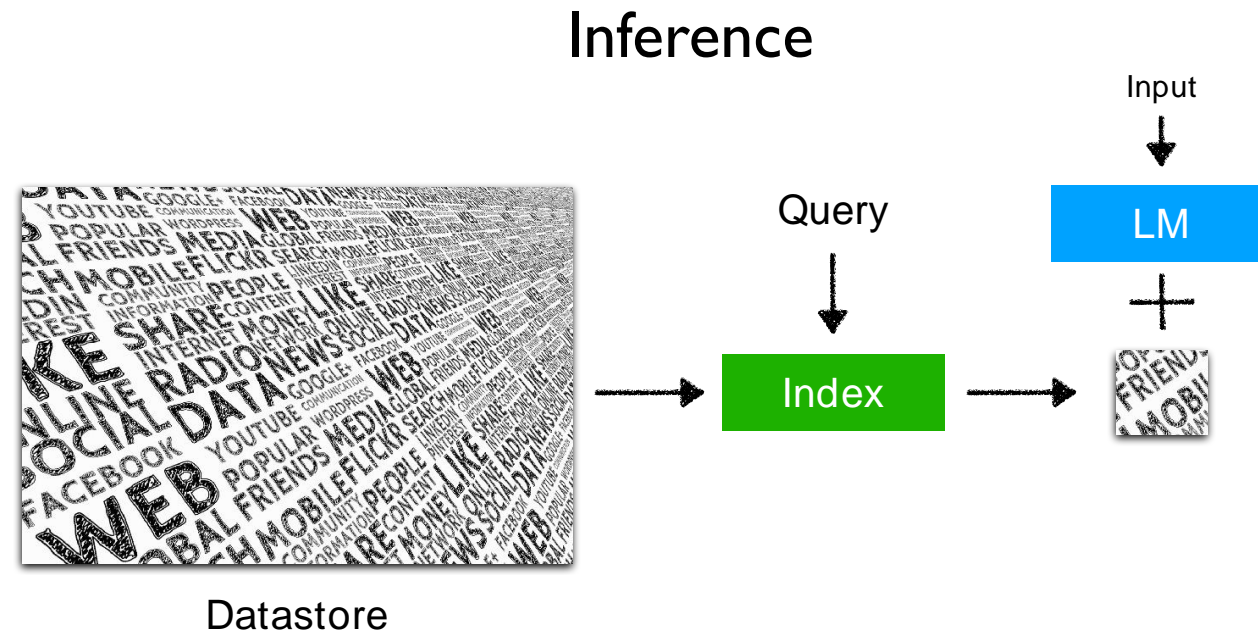


Kasai et al., REALTIME QA: What's the Answer Right Now.
NeurIPS Dataset and Benchmark 2023.

Lecture Overview

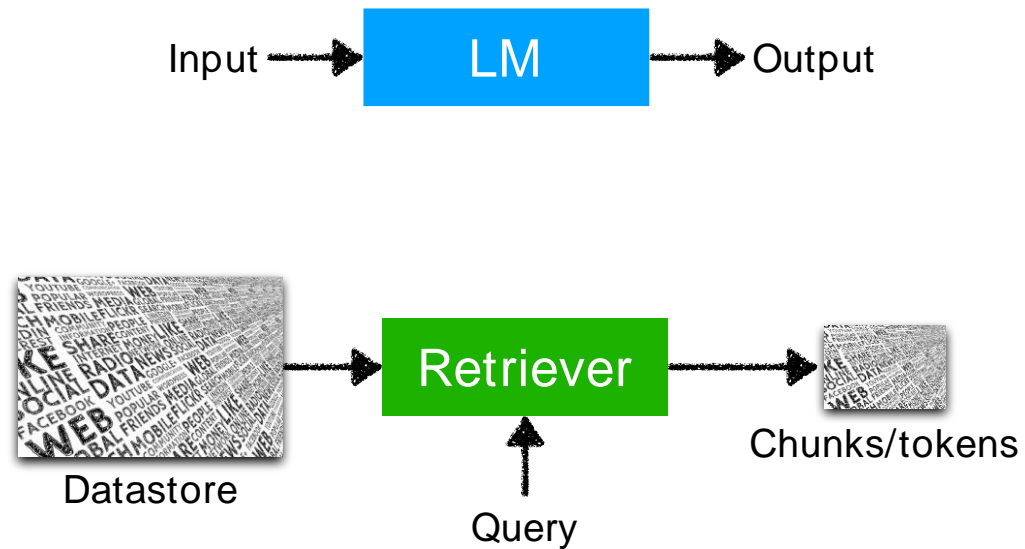
- What is retrieval-augmented generation (RAG)?
- Why do we need retrieval-augmented generation (RAG)?
- **RAG: Architecture and Training**
- Open Questions
- Beyond RAG: LLM Continual Learning

RAG Architecture



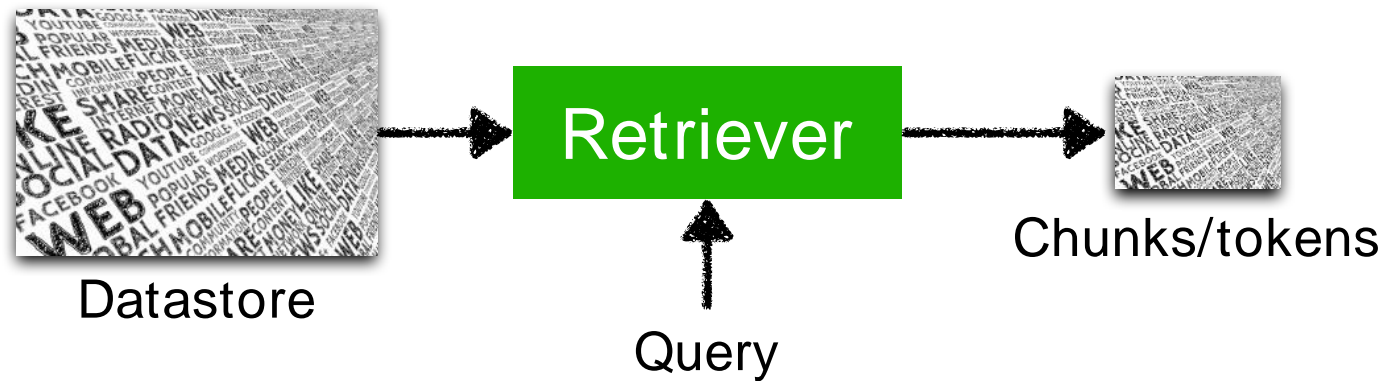
RAG Architecture

- Two main components:



RAG Architecture

- Let's first explore how the retriever module is **used**, the different **forms it can take** and how it can be improved through **training**.



Retrieval Module: Usage Overview

Inference: Index

Goal: find a small subset of elements in a datastore
that are the most similar to the query

sim: a similarity score between two pieces of text

Example $\text{sim}(i, j) = \text{tf}_{i,j} \times \log \frac{N}{\text{df}_i}$

$\text{tf}_{i,j}$: # of occurrences of i in j

N : # of total docs

df_i : # of docs containing i

Example $\text{sim}(i, j) = \text{Encoder}(i) \cdot \text{Encoder}(j)$

Maps the text into an h -dimensional vector

An entire field of
study on how to get
(or learn) the
similarity function
better
(We'll see some later!)

Retrieval Module: Usage Overview

Inference: Index

Goal: find a small subset of elements in a datastore
that are the most similar to the query

sim: a similarity score between two pieces of text

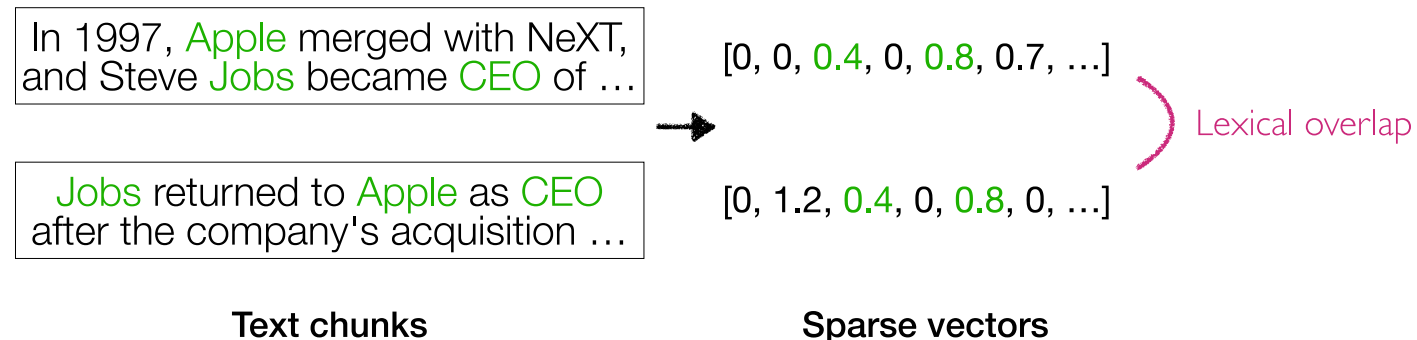
Can be a totally separate research area on
how to do this fast & accurate

Index: given q , return $\text{argTop-}k_{d \in \mathcal{D}} \text{sim}(q, d)$ through fast nearest neighbor search
 k elements from a datastore

[https://github.com/
facebookresearch/faiss/wiki/](https://github.com/facebookresearch/faiss/wiki/)

Retrieval Module: Designs

Sparse retrieval models: TF-IDF / BM25



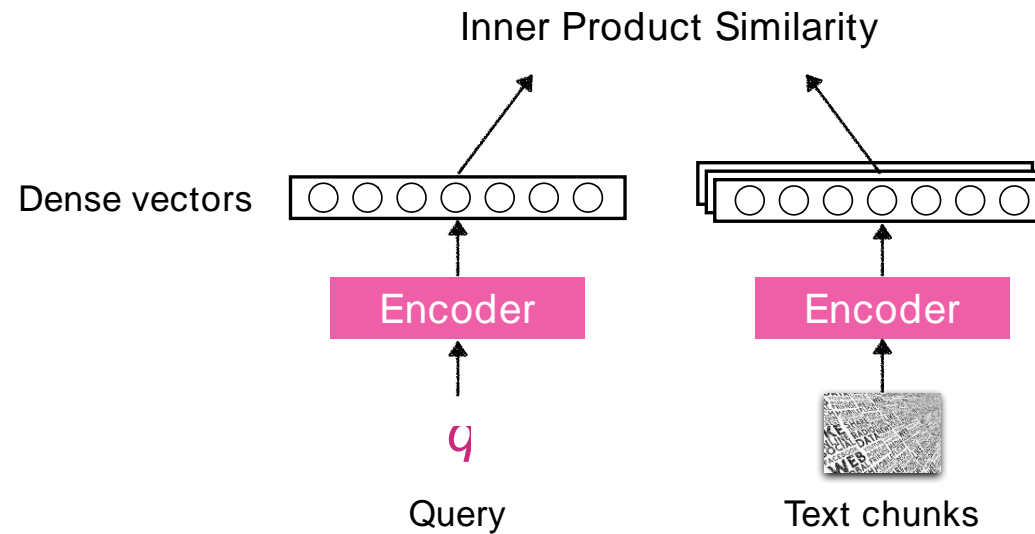
No training needed!

Ramos, 2003. "Using TF-IDF to Determine Word Relevance in Document Queries"

Robertson and Zaragoza, 2009. "The Probabilistic Relevance Framework: BM25 and Beyond"

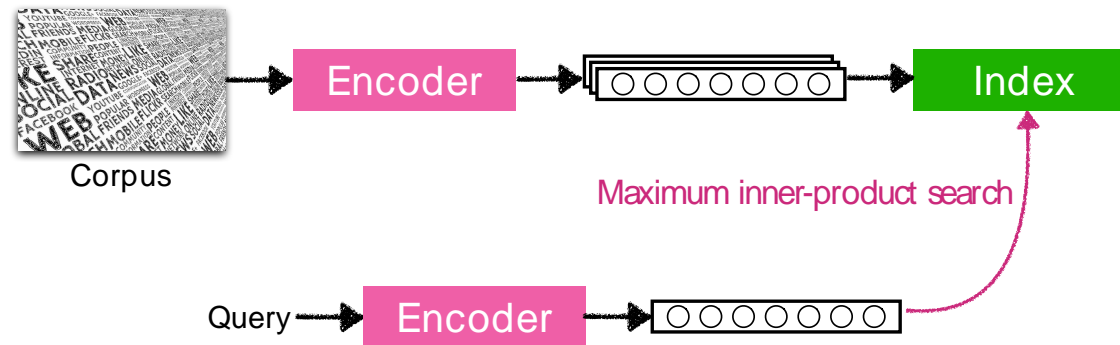
Retrieval Module: Designs

Dense retrieval models: DPR (Karpukhin et al. 2020)



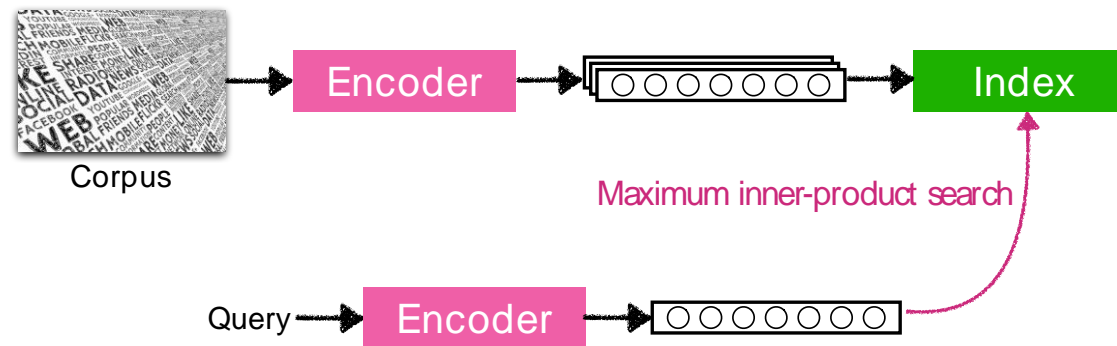
Retrieval Module: Designs

Dense retrievers: Inference



Retrieval Module: Training

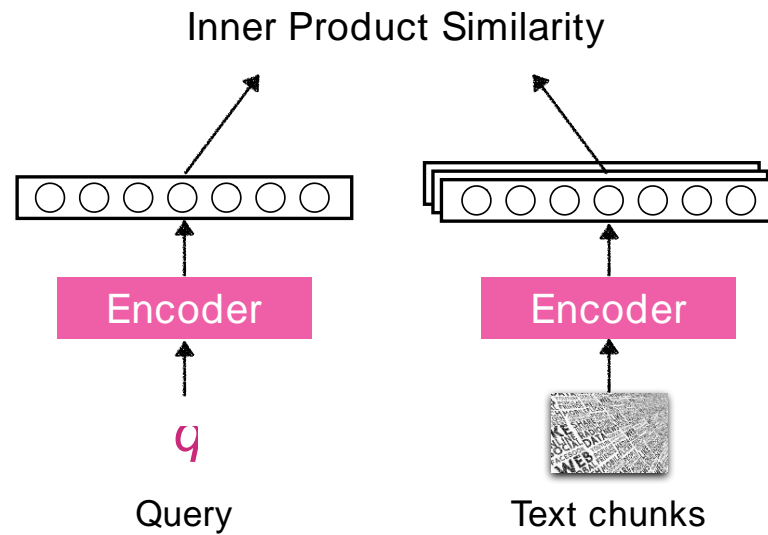
Dense retrievers: Inference



How to train dense retrieval models?

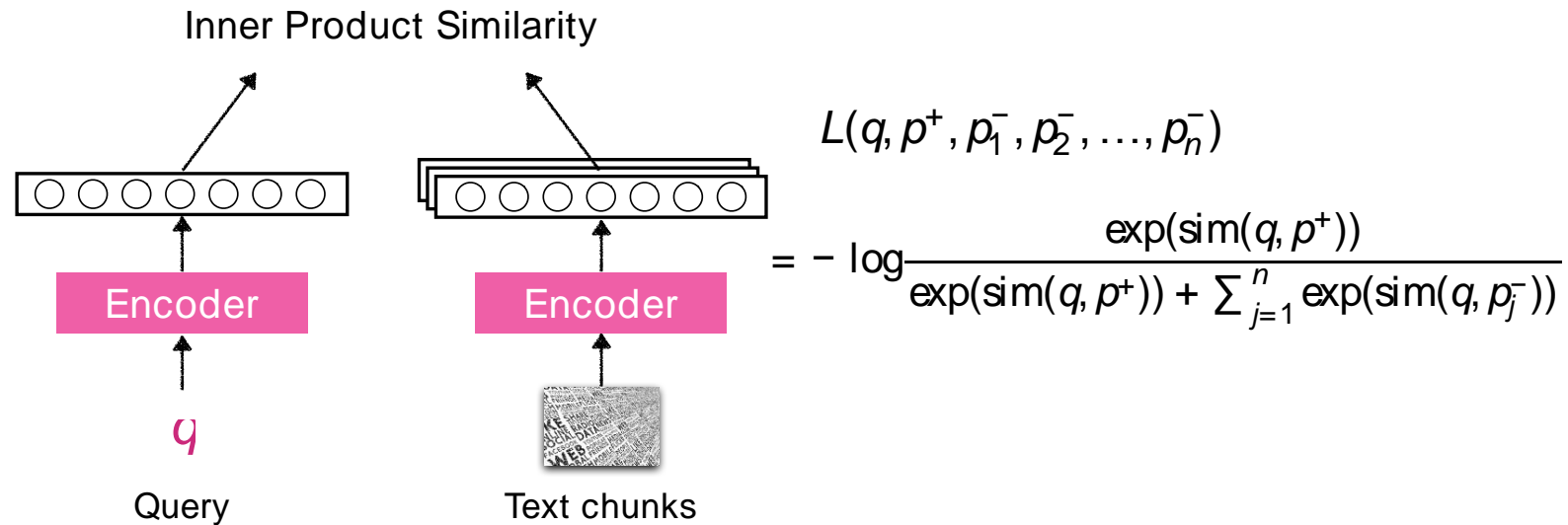
Retrieval Module: Training

Training dense retrieval models: DPR



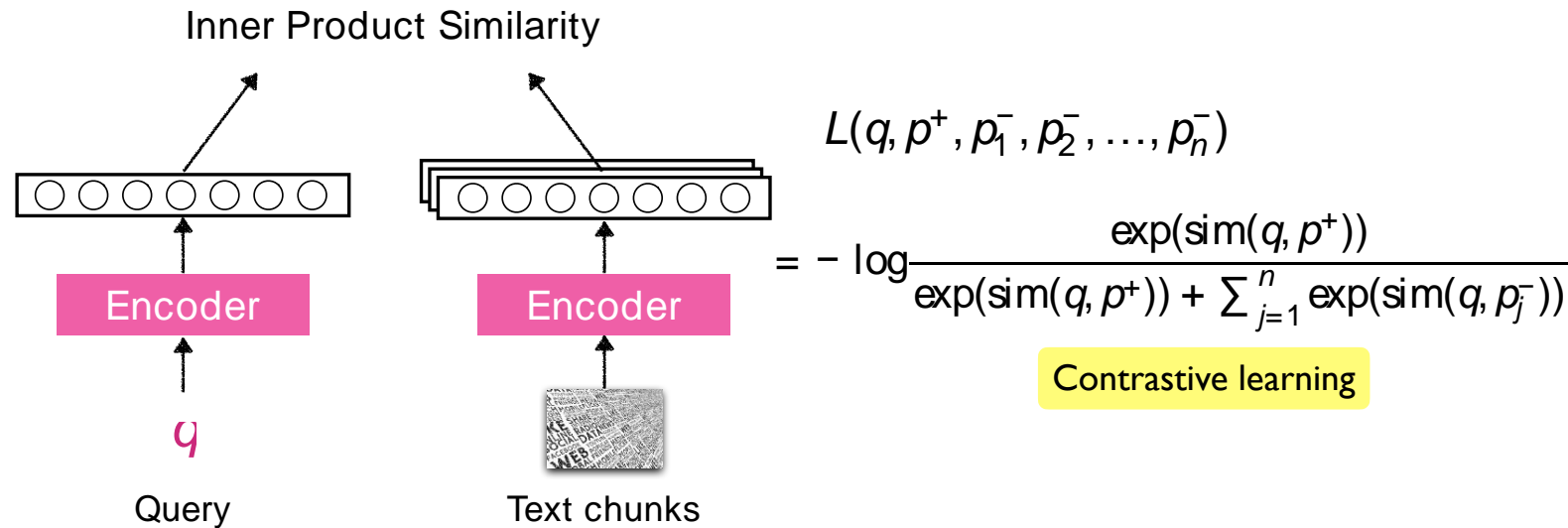
Retrieval Module: Training

Training dense retrieval models: DPR



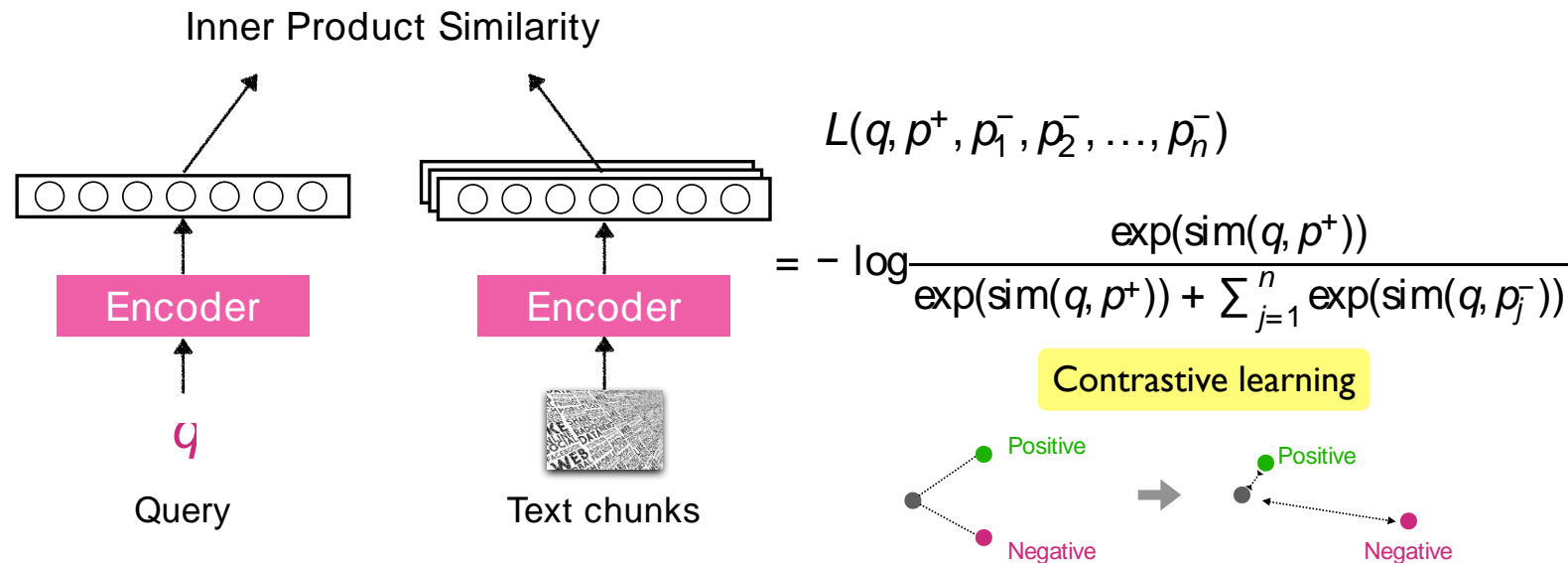
Retrieval Module: Training

Training dense retrieval models: DPR



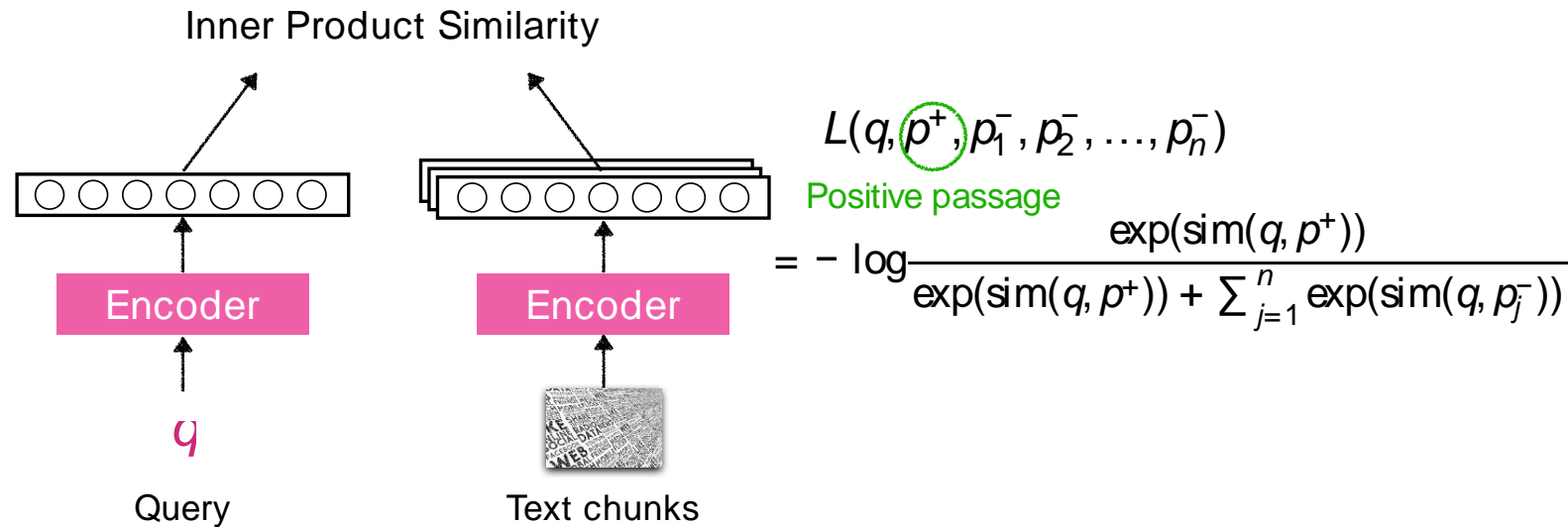
Retrieval Module: Training

Training dense retrieval models: DPR



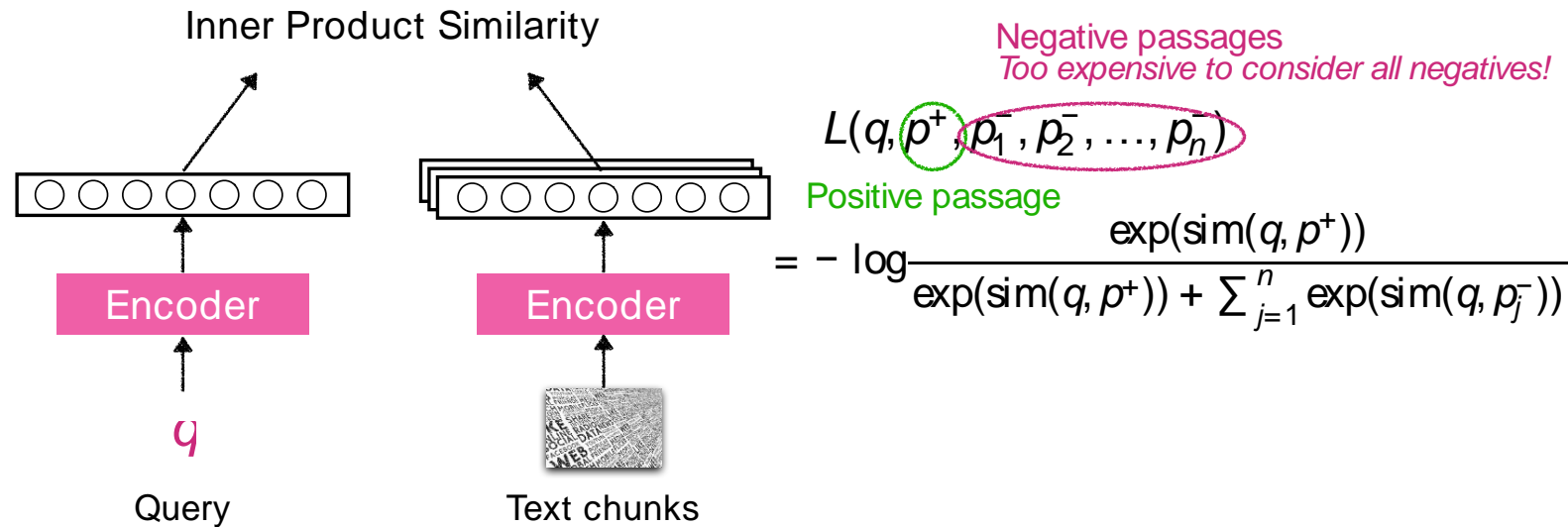
Retrieval Module: Training

Training dense retrieval models: DPR



Retrieval Module: Training

Training dense retrieval models: DPR



Retrieval Module: Training

Training with “in-batch” negatives

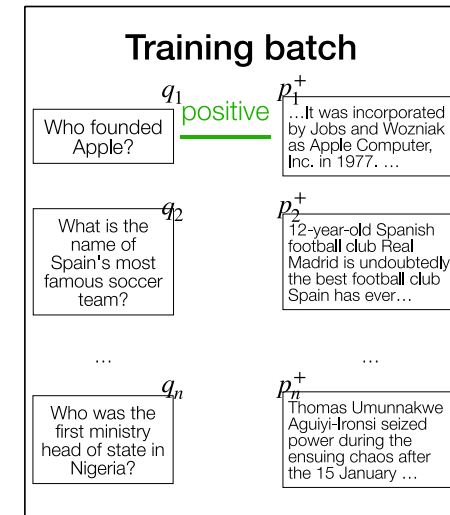
$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-)$$
$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Training batch	
q_1 Who founded Apple?	p_1^+ ...It was incorporated by Jobs and Wozniak as Apple Computer, Inc. in 1977. ...
q_2 What is the name of Spain's most famous soccer team?	p_2^+ 12-year-old Spanish football club Real Madrid is undoubtedly the best football club Spain has ever...
...	...
q_n Who was the first ministry head of state in Nigeria?	p_n^+ Thomas Ummnakwe Aguiyi-Ironsi seized power during the ensuing chaos after the 15 January ...

Retrieval Module: Training

Training with “in-batch” negatives

$$L(q, \boxed{p^+}, p_1^-, p_2^-, \dots, p_n^-)$$
$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

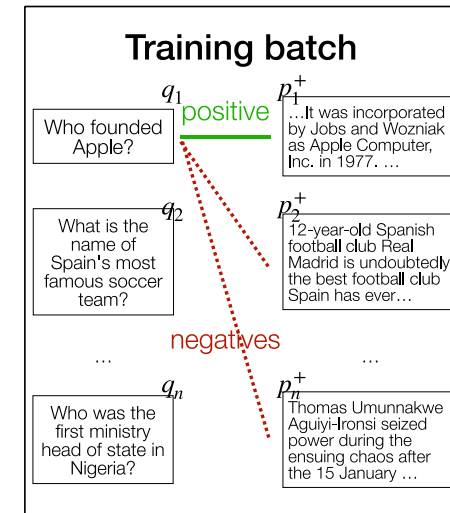


Retrieval Module: Training

Training with “in-batch” negatives

$$L(q, \boxed{p^+}, \boxed{p_1^-, p_2^-, \dots, p_n^-})$$
$$= -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

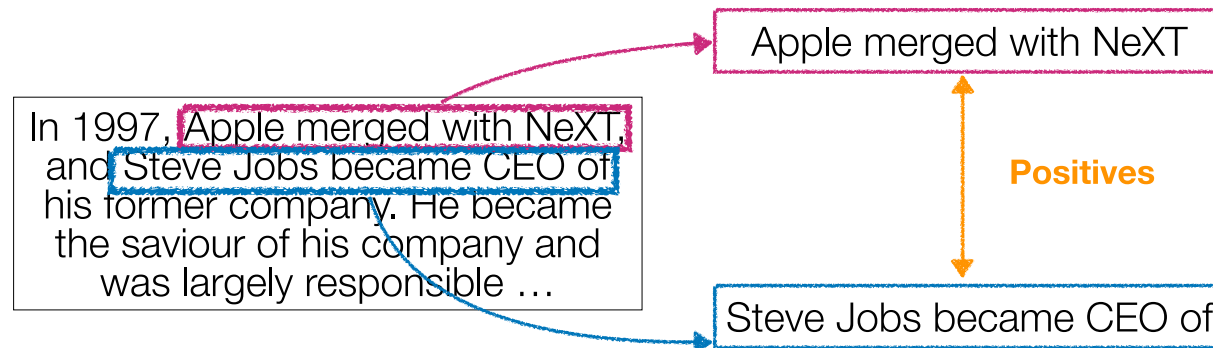
Back-propagation to all in-batch negatives!



Retrieval Module: Training

Contriever (Izacard et al. 2022)

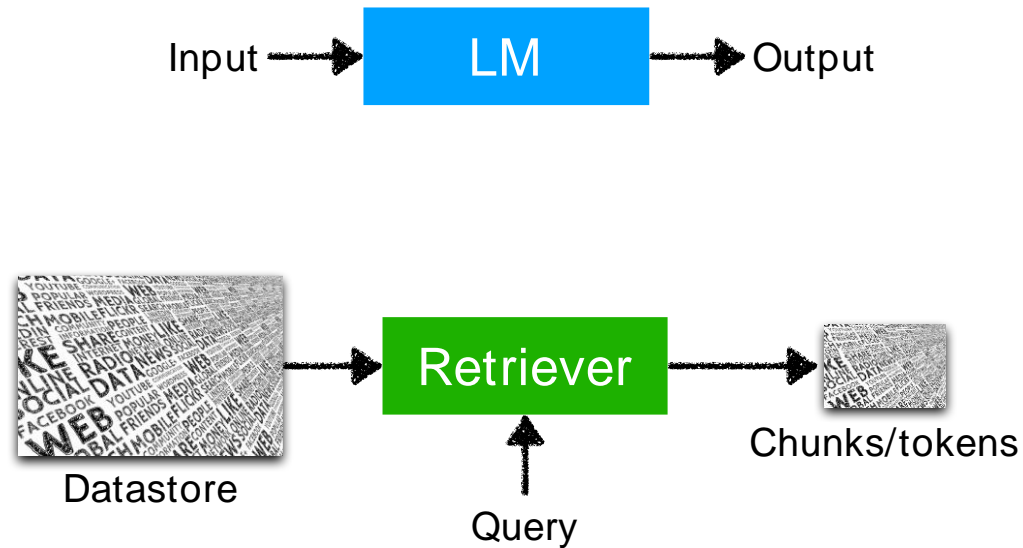
Independent Cropping



Unsupervised dense retrieval model!

RAG Architecture

- We can now take a step back again to understand a few complete simple RAG systems.



Retrieval-in-context in LM (Ram et al. 2023)

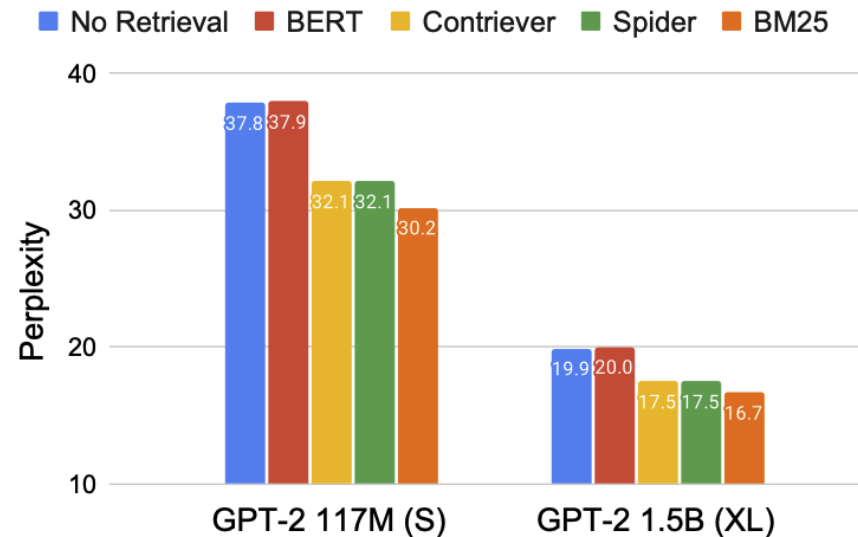
- Simplest version of RAG, frozen retriever & LLM

x = World Cup 2022 was the last with 32 teams, before the increase to

World Cup 2022 was the last with 32 teams, before the increase to



Retrieval-in-context in LM (Ram et al. 2023)



Better **retrieval model**



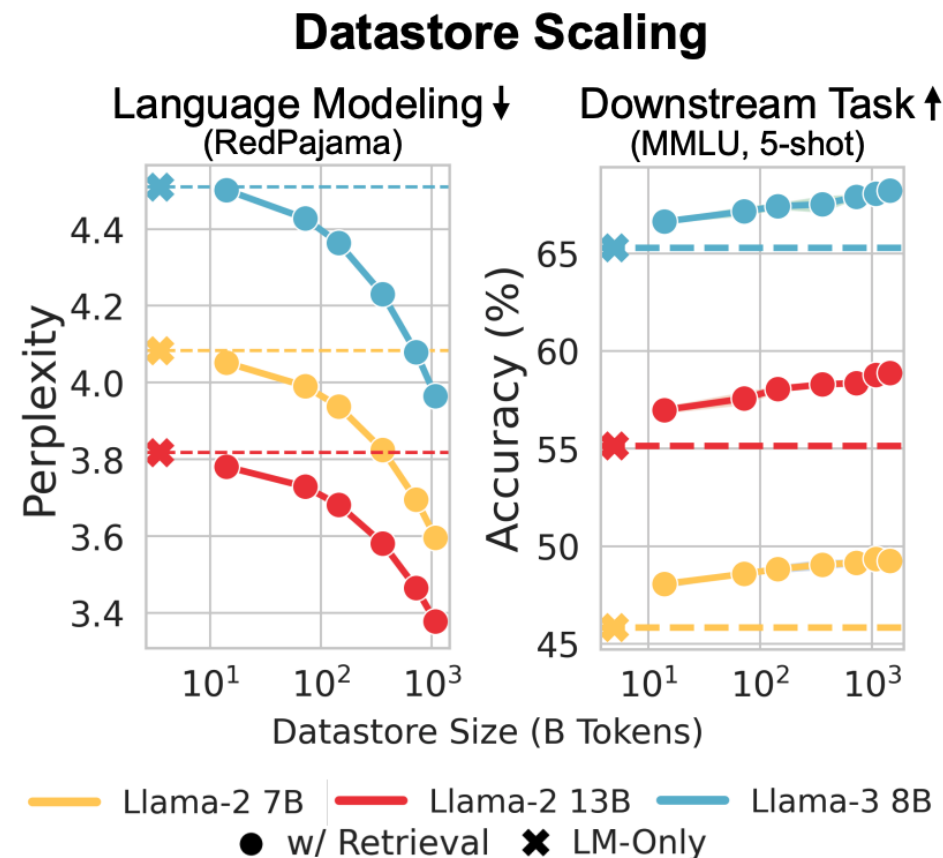
Better **retrieval-based LMs**

Better **base LMs**

Each component can be improved separately

Scaling RAG Systems (Shao et al 2024)

- Scaling the retrieval corpus → substantial perplexity and performance improvements.
- Boost is inversely proportional to LLM size.
- Save cost on LLM size while getting better performance & RAG benefits.



RAG Training

- Now that we've looked at the very simplest RAG setting, let's look at different methods to optimize the LLM for retrieval augmentation.



RAG Training

Training language models



Minimize $-\log P_{\text{LM}}(y | x)$

RAG Training

Training language models



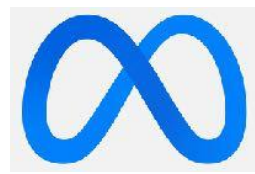
Minimize $-\log P_{\text{LM}}(y|x)$



GPT



PaLM



LLaMA



GPT-J

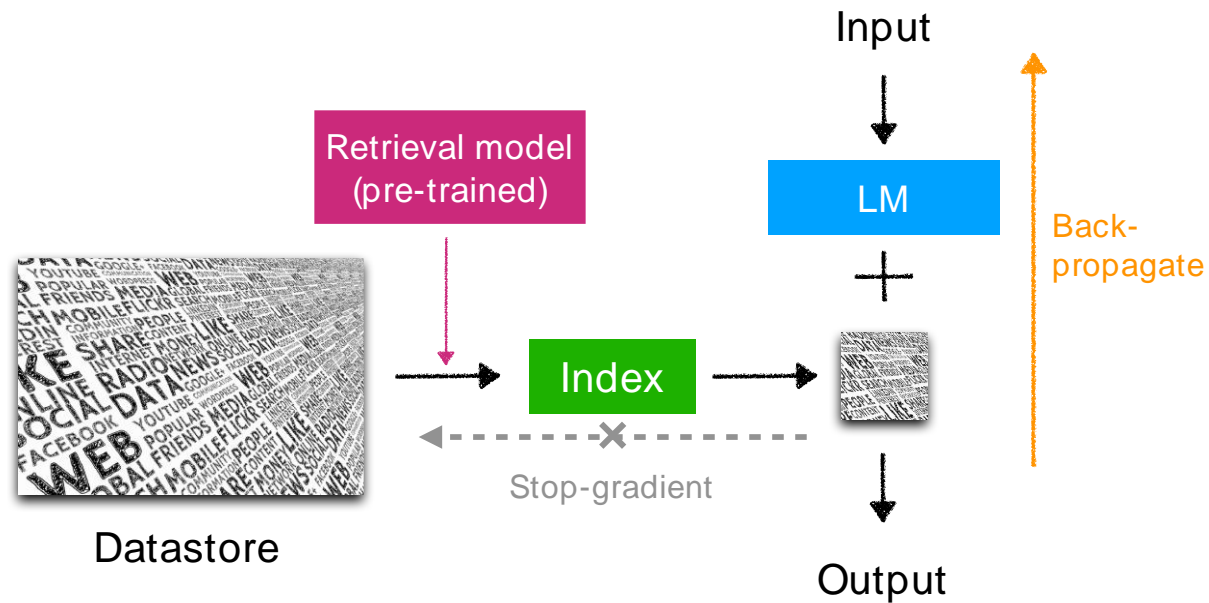
.....

RETRO (Borgeaud et al. 2021)

- Retrieval-augmented language model.
- Different attention mechanism to deal with retrieved chunks.
- This model is pretrained from scratch with a retrieval corpus.

RETRO (Borgeaud et al. 2021)

- Retrieval models are first trained independently and then fixed
- Language models are trained with an objective that depends on the retrieval



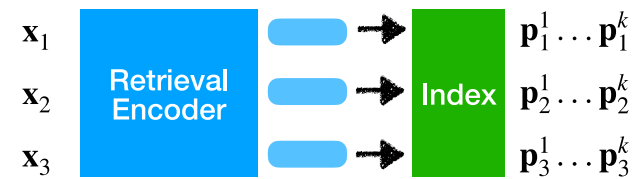
RETRO (Borgeaud et al. 2021)

\mathbf{x} = World Cup 2022 was \mathbf{x}_1 the last with 32 teams, \mathbf{x}_2 before the increase to \mathbf{x}_3

RETRO (Borgeaud et al. 2021)

\mathbf{x} = World Cup 2022 was ~~the~~ last with 32 teams, ~~before~~ the increase to

\mathbf{x}_1 \mathbf{x}_2 \mathbf{x}_3

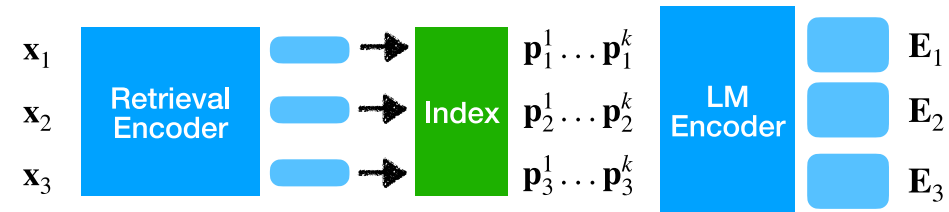


RETRO (Borgeaud et al. 2021)

\mathbf{x} = World Cup 2022 was ~~the~~ last with 32 teams, ~~before~~ the increase to

\mathbf{x}_1 \mathbf{x}_2 \mathbf{x}_3

(k chunks of text per split)

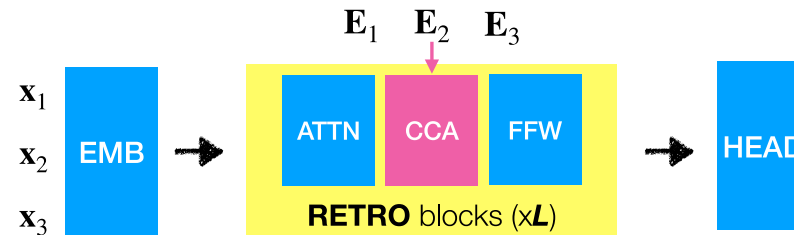
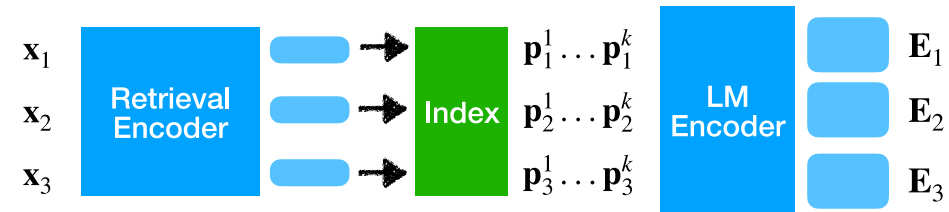


RETRO (Borgeaud et al. 2021)

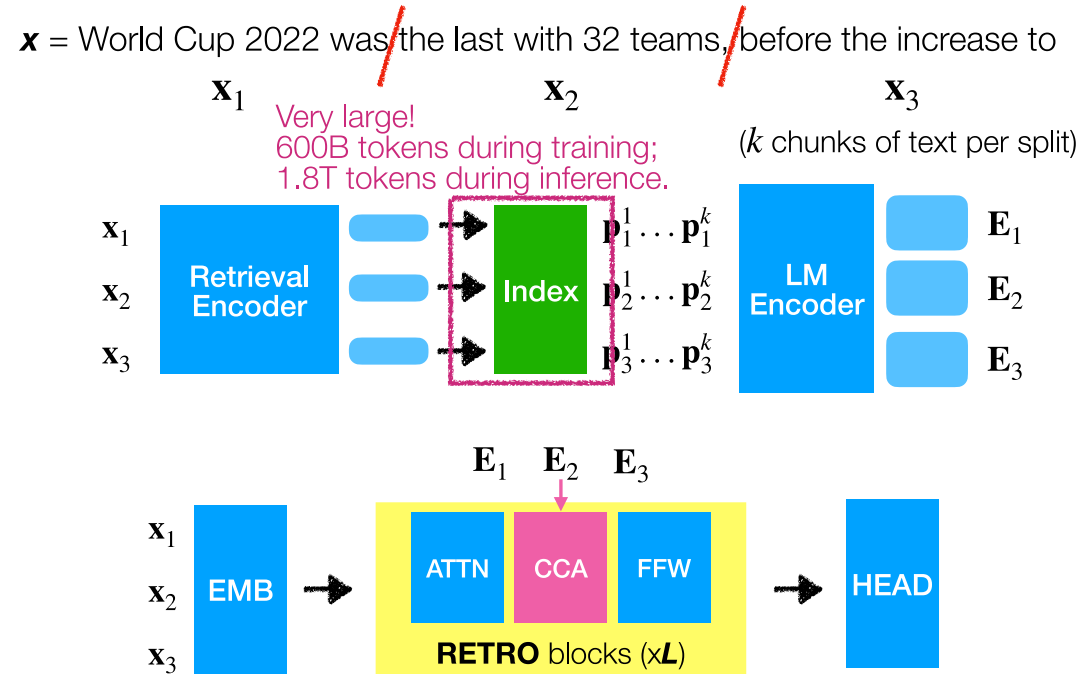
\mathbf{x} = World Cup 2022 was the last with 32 teams, before the increase to

\mathbf{x}_1 \mathbf{x}_2 \mathbf{x}_3

(k chunks of text per split)

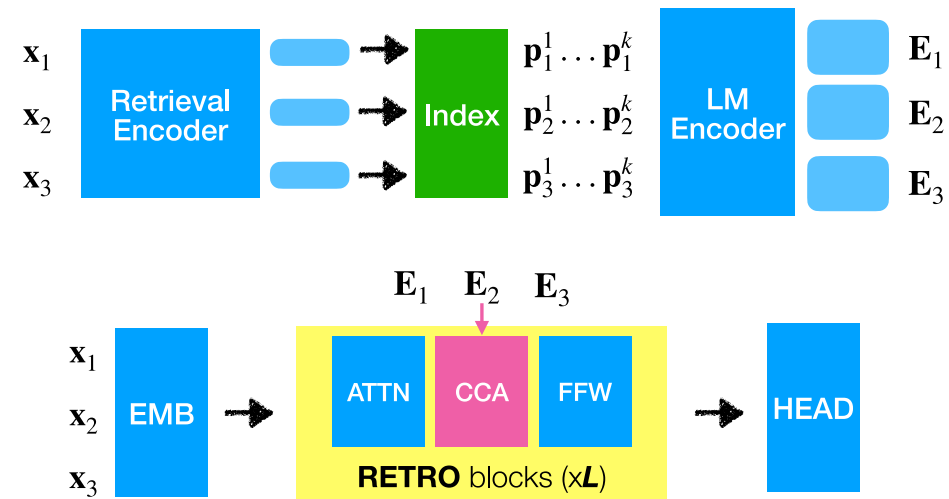


RETRO (Borgeaud et al. 2021)



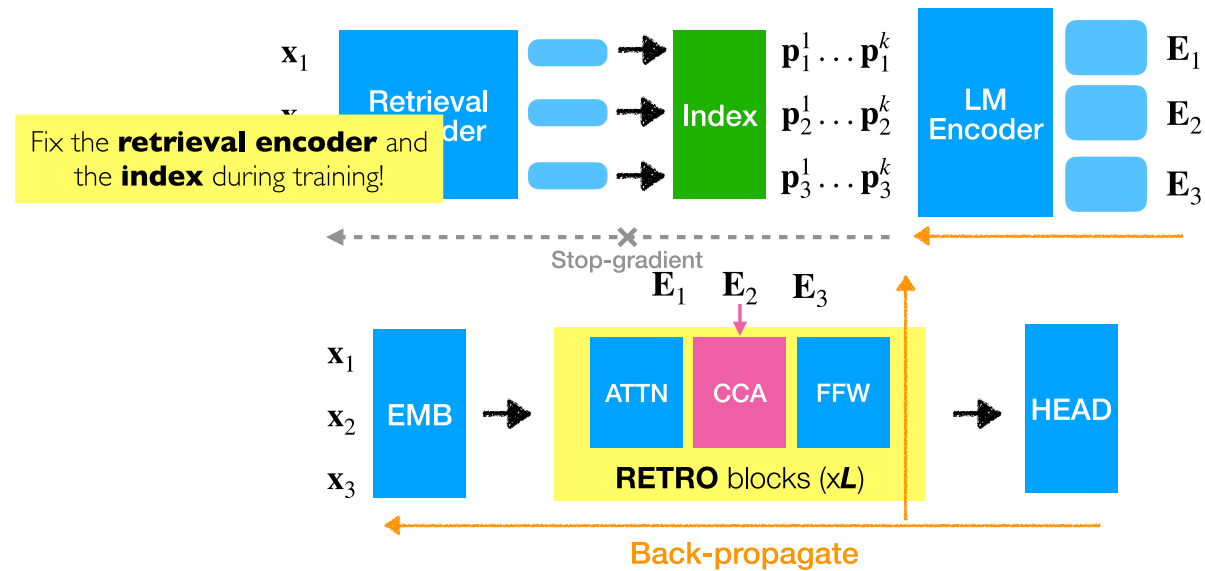
RETRO (Borgeaud et al. 2021)

RETRO: Training



RETRO (Borgeaud et al. 2021)

RETRO: Training



RETRO++ (Wang, Ping, Xu et al. 2023)

- Open-source version of RETRO w/ frozen RAG as well
- Shows improvements over GPT models

Tasks	Small		Medium		XL		XXL	
	GPT	RETRO	GPT	RETRO	GPT	RETRO	GPT	RETRO
<i>Knowledge-intensive Tasks</i>								
HellaSwag	31.3	36.2 ↑4.9	43.2	46.2 ↑3.0	56.7	59.0 ↑2.3	72.3	70.6 ↓1.7
BoolQ	59.3	61.8 ↑2.5	57.4	57.2 ↓0.2	62.2	62.7 ↑0.5	67.3	70.7 ↑3.4
<i>Knowledge-nonintensive Tasks</i>								
Lambda	41.7	41.4 ↓0.3	54.1	55.0 ↑0.9	63.9	64.0 ↑0.1	73.9	72.7 ↓1.2
RACE	34.6	32.5 ↓2.1	37.3	37.3 ↑0.0	40.8	39.9 ↓0.9	44.3	43.2 ↓1.1
PiQA	64.3	64.8 ↑0.5	70.2	68.7 ↓1.5	73.7	74.1 ↑0.4	78.5	77.4 ↓1.1
WinoGrande	52.4	52.0 ↓0.4	53.8	55.2 ↑1.4	59.0	60.1 ↑1.1	68.5	65.8 ↓2.7
ANLI-R2	35.1	36.2 ↑1.1	33.5	33.3 ↓0.2	34.3	35.3 ↑1.0	32.2	35.5 ↑3.3
HANS	51.5	51.4 ↓0.1	50.5	50.5 ↑0.0	50.1	50.0 ↓0.1	50.8	56.5 ↑5.7
WiC	50.0	50.0 ↑0.0	50.2	50.0 ↓0.2	47.8	49.8 ↑2.0	52.4	52.4 ↑0.0
Avg. Acc. (↑)	46.7	47.4 ↑0.7	50.0	50.4 ↑0.4	54.3	55.0 ↑0.7	60.0	60.5 ↑0.5

RAG Training

- RETRO (DeepMind) and RETRO++ (NVIDIA) are the only two retrieval-augmented models that perform **full-scale pretraining**.
- Nevertheless, many works have explored ways to **adapt current** LLMs to obtain more powerful RAG systems.
- As examples of this, we will look at:
 - Self-RAG (standard training)
 - RAG-RL (uses trendy RL concepts)

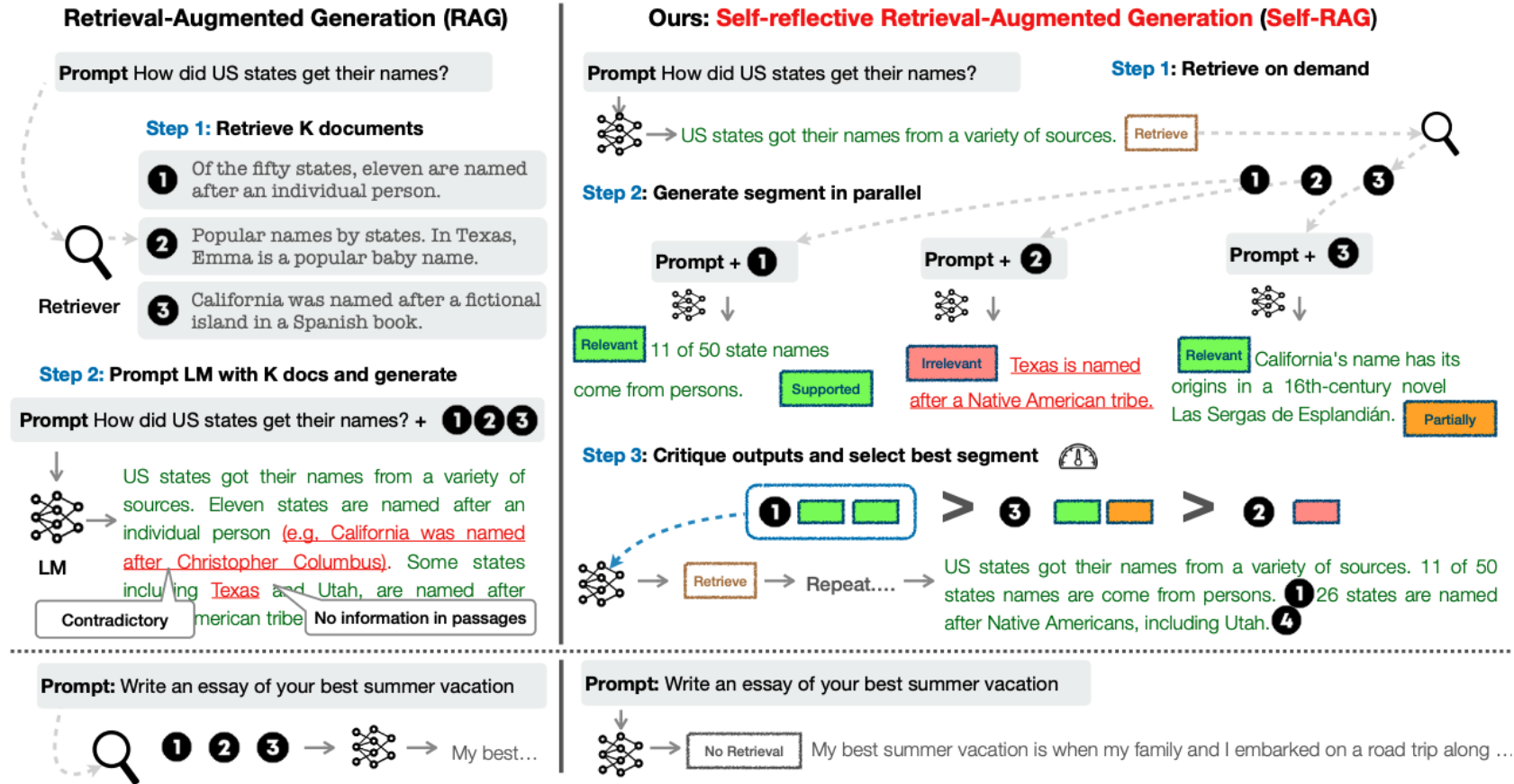
Self-RAG (Asai et al 2023)

- Fine-tuning a medium-sized LM to output the following tokens:

Type	Input	Output	Definitions
Retrieve	$x / x, y$	{yes, no, continue}	Decides when to retrieve with \mathcal{R}
ISREL	x, d	{ relevant , irrelevant}	d provides useful information to solve x .
ISSUP	x, d, y	{ fully supported , partially supported, no support}	All of the verification-worthy statement in y is supported by d .
ISUSE	x, y	{ 5 , 4, 3, 2, 1}	y is a useful response to x .

- These tokens guide the use of retrieval information in the generation procedure.
- Uses a larger model GPT-4 to generate training data

Self-RAG (Asai et al 2023)



Self-RAG (Asai et al 2023)

- Outperforms other tool use methods and small LLMs
- Cons: Inference and training are both quite complicated and custom-made.

LM	Short-form		Closed-set		Long-form generations (with citations)					
	PopQA (acc)	TQA (acc)	Pub (acc)	ARC (acc)	Bio (FS)	(em)	(rg)	ASQA (mau)	(pre)	(rec)
<i>LMs with proprietary data</i>										
Llama2-c _{13B}	20.0	59.3	49.4	38.4	55.9	22.4	29.6	28.6	–	–
Ret-Llama2-c _{13B}	51.8	59.8	52.1	37.9	79.9	32.8	34.8	43.8	19.8	36.1
ChatGPT	29.3	74.3	70.1	75.3	71.8	35.3	36.2	68.8	–	–
Ret-ChatGPT	50.8	65.7	54.7	75.3	–	40.7	39.9	79.7	65.1	76.6
Perplexity.ai	–	–	–	–	71.2	–	–	–	–	–
<i>Baselines without retrieval</i>										
Llama2 _{7B}	14.7	30.5	34.2	21.8	44.5	7.9	15.3	19.0	–	–
Alpaca _{7B}	23.6	54.5	49.8	45.0	45.8	18.8	29.4	61.7	–	–
Llama2 _{13B}	14.7	38.5	29.4	29.4	53.4	7.2	12.4	16.0	–	–
Alpaca _{13B}	24.4	61.3	55.5	54.9	50.2	22.9	32.0	70.6	–	–
CoVE _{65B} *	–	–	–	–	71.2	–	–	–	–	–
<i>Baselines with retrieval</i>										
Toolformer* _{6B}	–	48.8	–	–	–	–	–	–	–	–
Llama2 _{7B}	38.2	42.5	30.0	48.0	78.0	15.2	22.1	32.0	2.9	4.0
Alpaca _{7B}	46.7	64.1	40.2	48.0	76.6	30.9	33.3	57.9	5.5	7.2
Llama2-FT _{7B}	48.7	57.3	64.3	65.8	78.2	31.0	35.8	51.2	5.0	7.5
SAIL* _{7B}	–	–	69.2	48.4	–	–	–	–	–	–
Llama2 _{13B}	45.7	47.0	30.2	26.0	77.5	16.3	20.5	24.7	2.3	3.6
Alpaca _{13B}	46.1	66.9	51.1	57.6	77.7	34.8	36.7	56.6	2.0	3.8
Our SELF-RAG_{7B}	54.9	66.4	72.4	67.3	81.2	30.0	35.7	74.3	66.9	67.8
Our SELF-RAG_{13B}	55.8	69.3	74.5	73.1	80.2	31.7	37.0	71.6	70.3	71.3

RAG-RL (Huang et al. 2025)

- One of several works exploring the use of reinforcement learning in RAG systems.
- Leverage post-training RL algorithm used in math coding: Group Relative Policy Optimization (GRPO)
- Uses simple rule-based rewards from supervised data.

$$\mathcal{R}_{\text{total}} = \mathcal{R}_{\text{answer}} + \mathcal{R}_{\text{citation}} + \mathcal{R}_{\text{formatting}}.$$

RAG-RL (Huang et al. 2025)

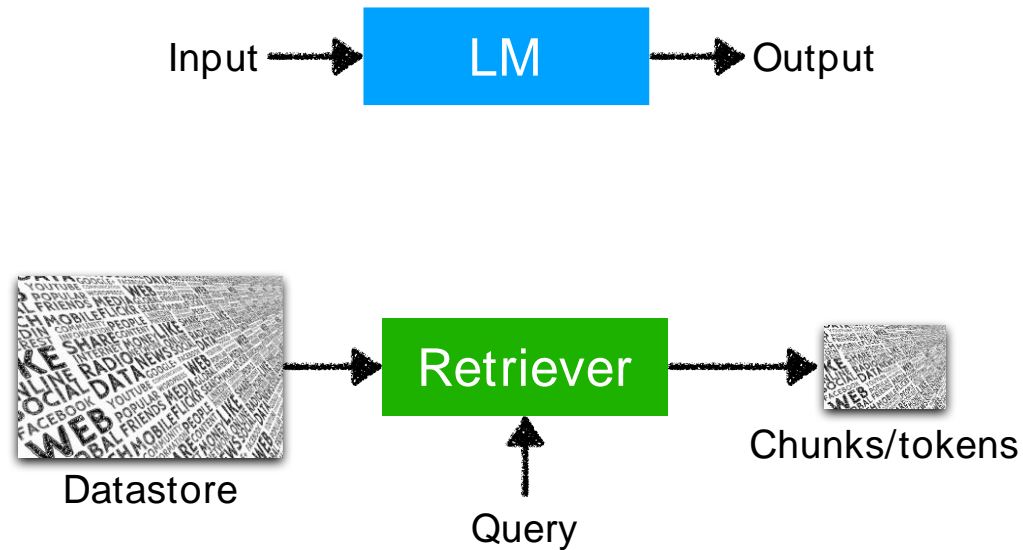
- Experimental settings are quite simple, only 10-20 “distractor” documents per question (not a real search engine setting).
- Some improvements in multi-hop setting.

Model / Curriculum	HotpotQA			MuSiQue		
	Answer F1	Citation F1	Joint F1	Answer F1	Citation F1	Joint F1
Qwen2.5-7B-Instruct / –	60.65	36.47	45.55	25.88	25.35	25.61
Qwen2.5-7B-Instruct / Max	68.52	71.55	70.00	46.06	64.66	53.80
Qwen2.5-7B-Instruct / Linear	72.65	80.53	76.39	47.93	68.45	56.38
Qwen2.5-7B-Instruct / Linear Shuffled	70.12	79.75	74.63	51.95	69.63	59.51
Qwen2.5-7B-Instruct / Min-Max	74.97	81.25	77.98	55.13	69.27	61.40
Qwen2.5-7B-Instruct / Min-Max Shuffled	72.12	80.40	76.09	52.44	69.91	59.93

Table 1: Performance of models in the distractor setting.

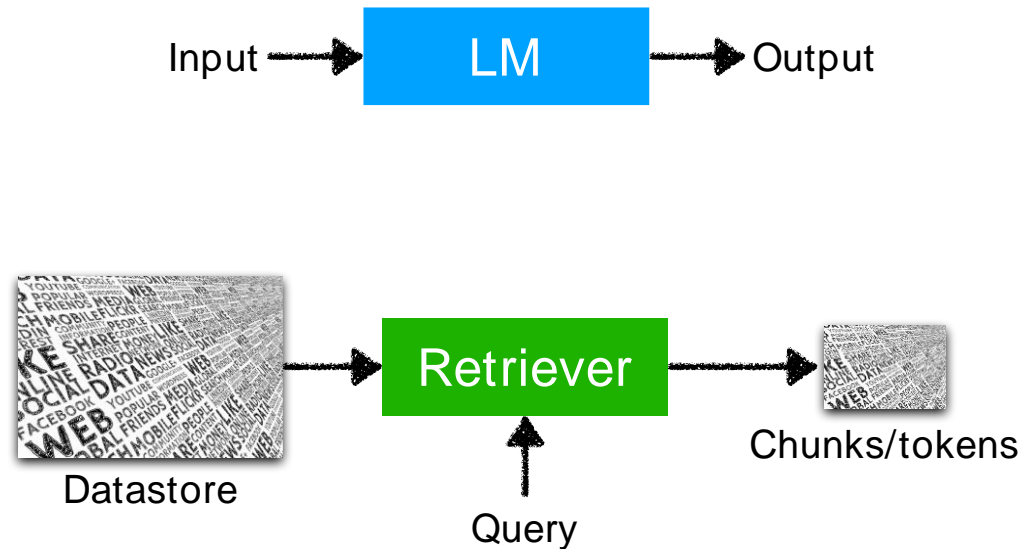
RAG Training

- We have now seen that both components can be trained separately with good results.
- Is this the best we can do??



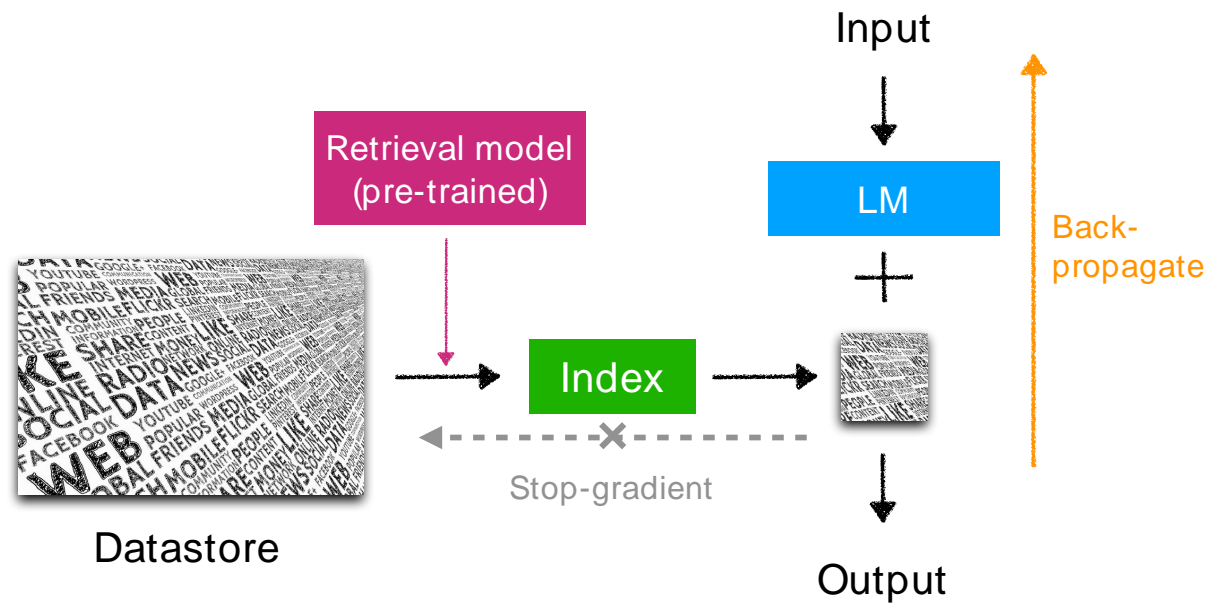
RAG Training

- Retrieval and generation depend on each other, wouldn't it make sense to optimize them simultaneously?
- Who here can guess why this is challenging?



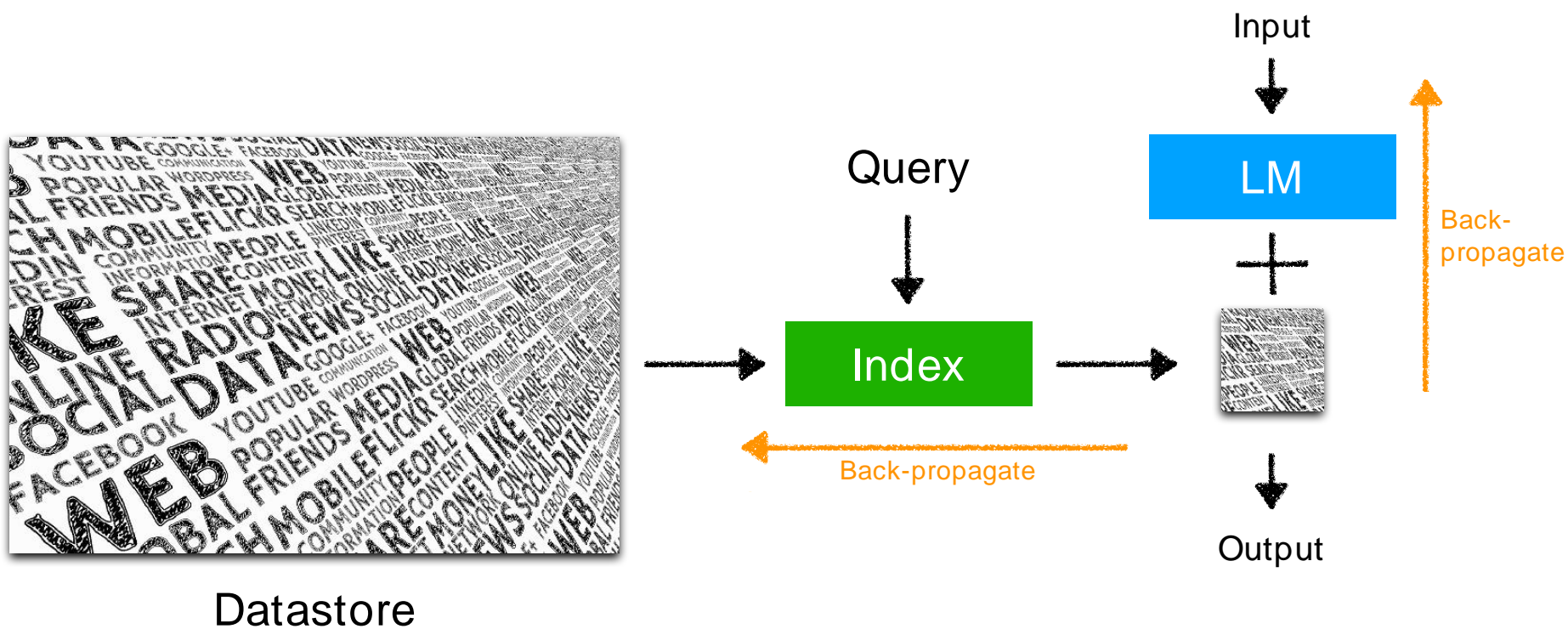
RETRO (Borgeaud et al. 2021)

- Retrieval models are first trained independently and then fixed
- Language models are trained with an objective that depends on the retrieval



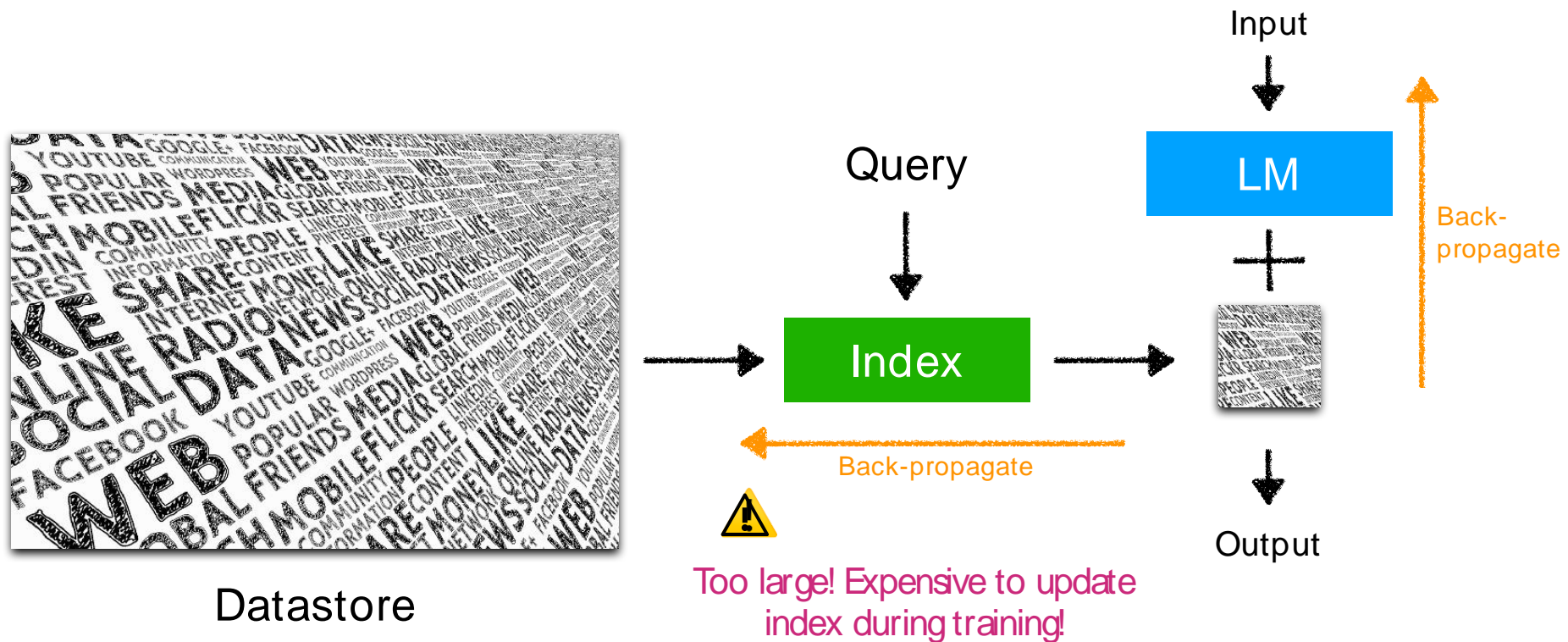
RAG: Joint Training

Why is joint training challenging?



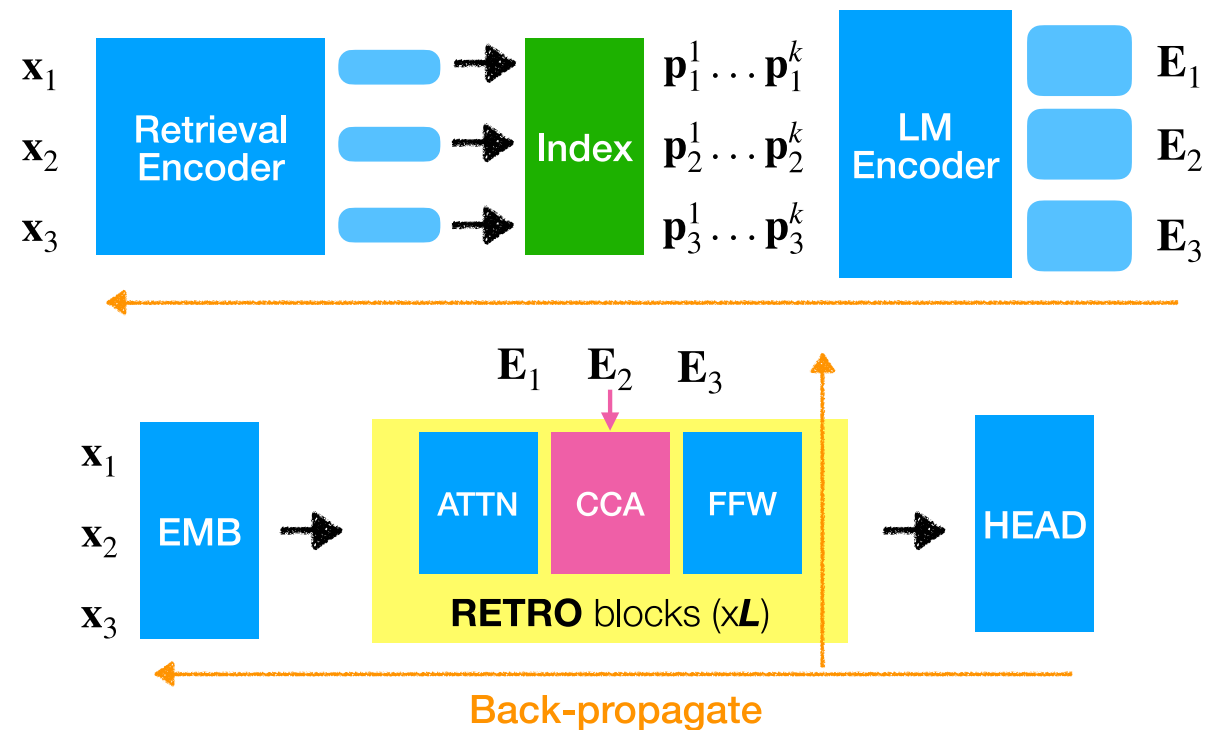
RAG: Joint Training

Why is joint training challenging?



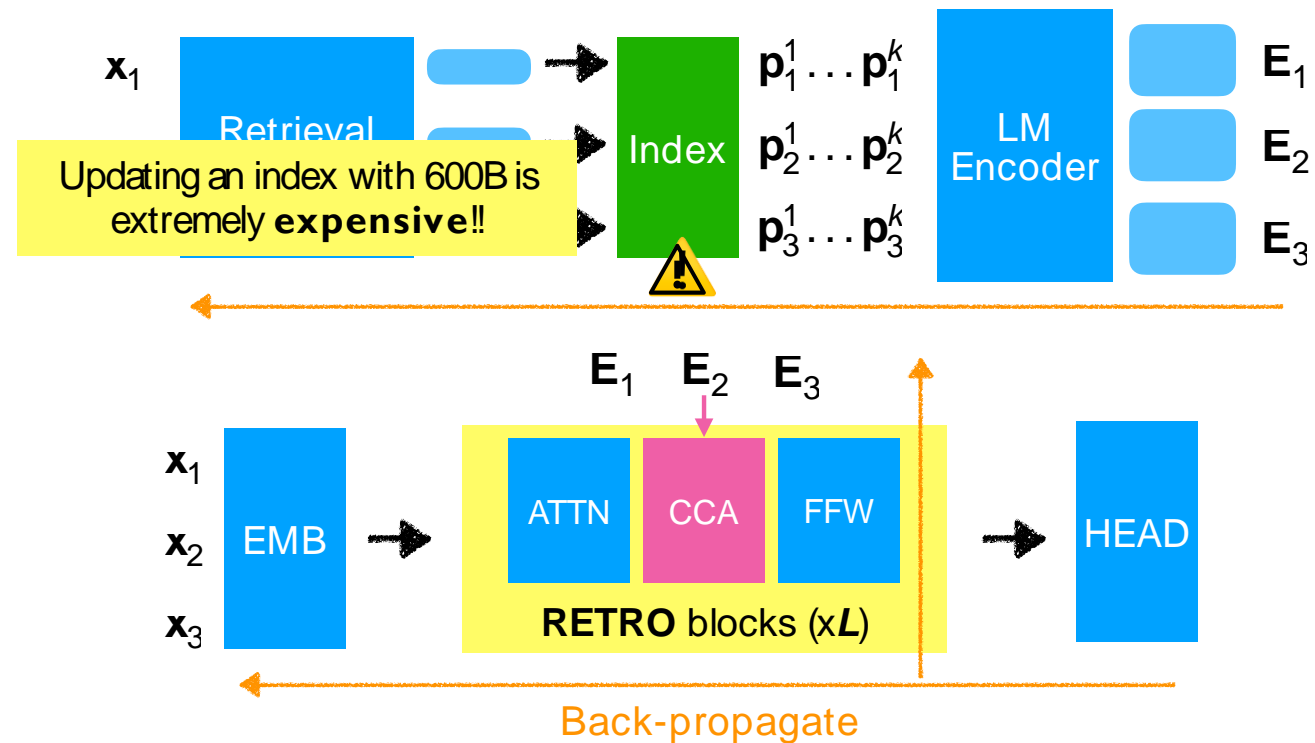
RAG: Joint Training

RETRO: Training



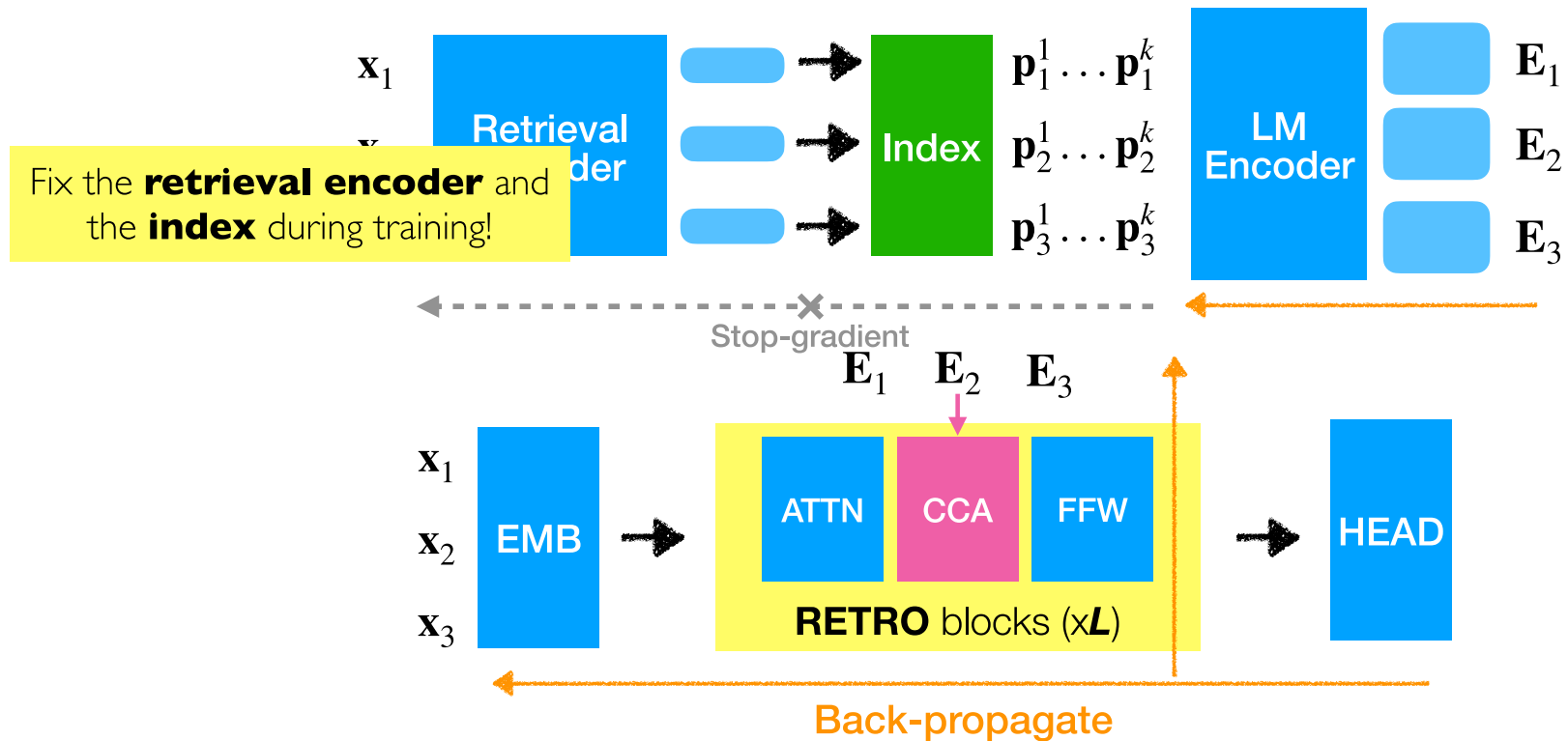
RAG: Joint Training

RETRO: Training



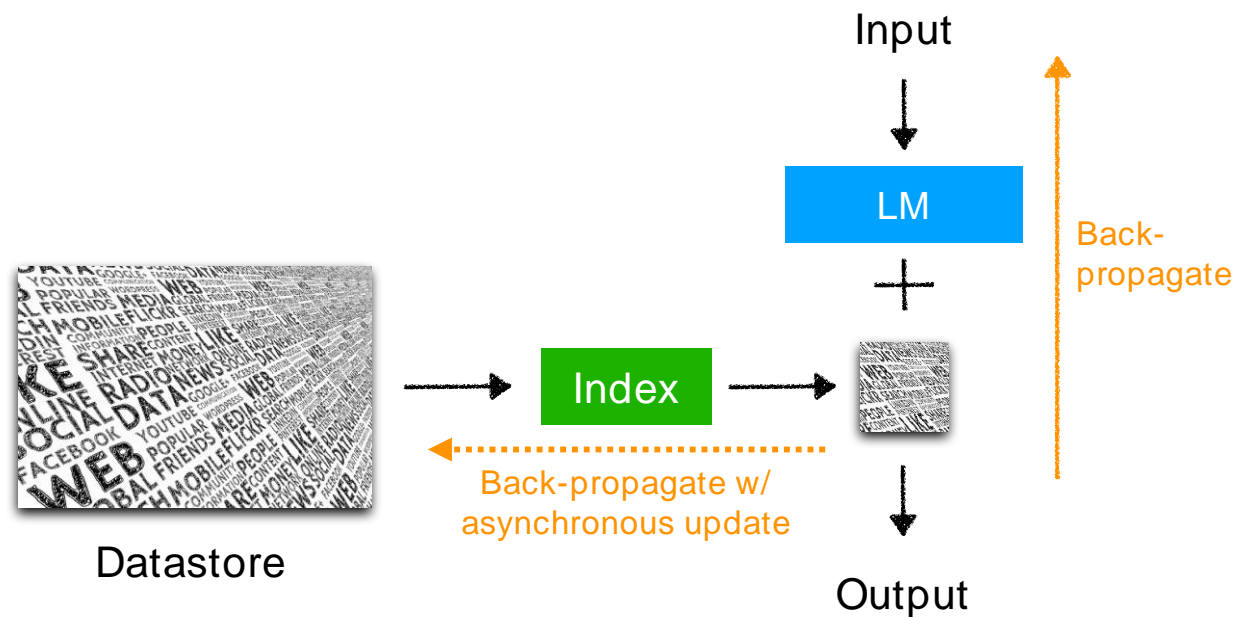
RETRO (Borgeaud et al. 2021)

RETRO: Training



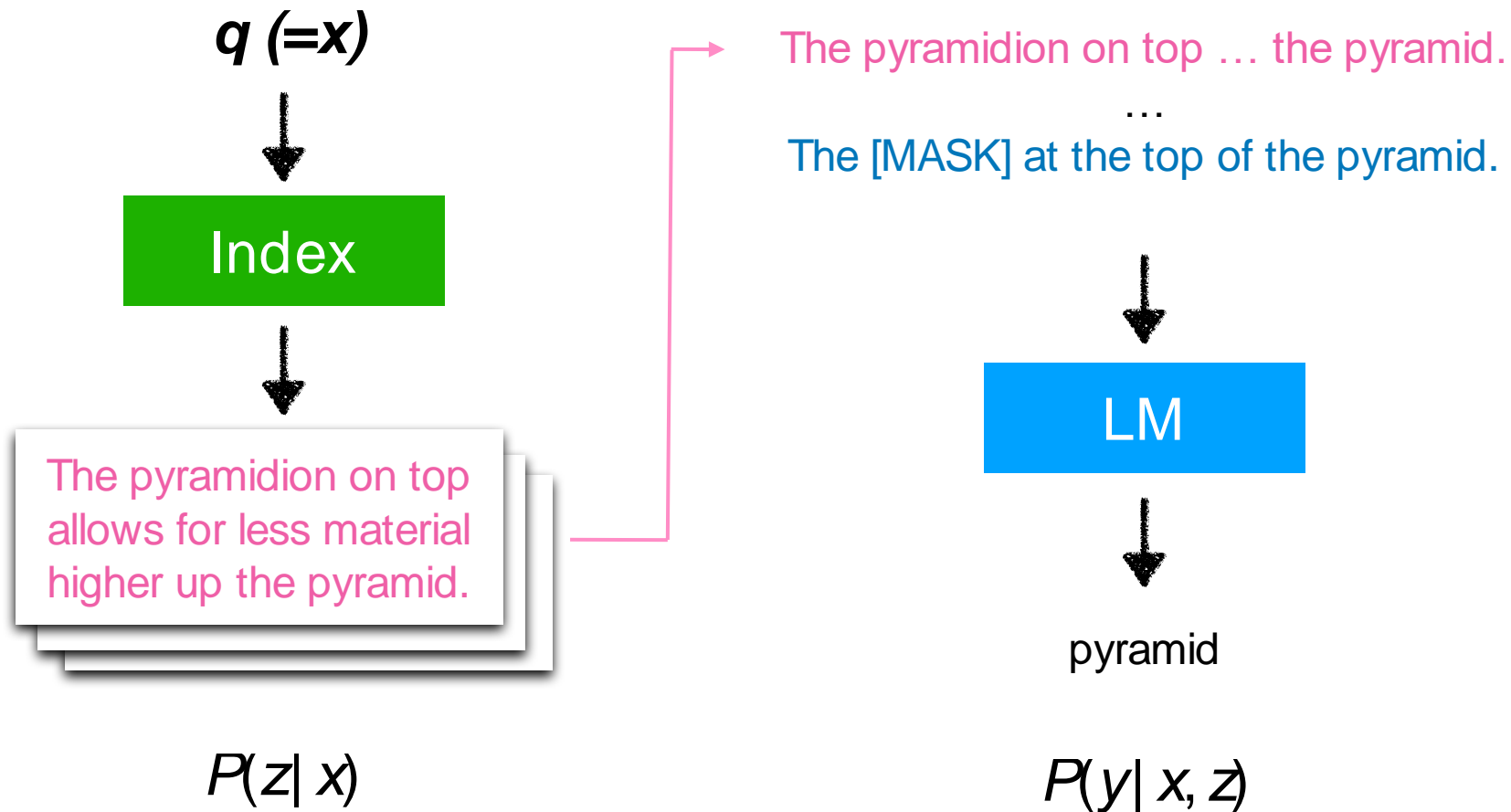
RAG: Async Joint Training

- Retrieval models and language models are trained jointly
- Allow the index to be “**stale**”; rebuild the retrieval index every T steps



REALM (Guu et al. 2020)

x = The [MASK] at the top of the pyramid.



REALM: Training

Objective: maximize $\sum_{z \in \theta} P_{\theta}(z|q) P_{\theta}(y|q, z)$

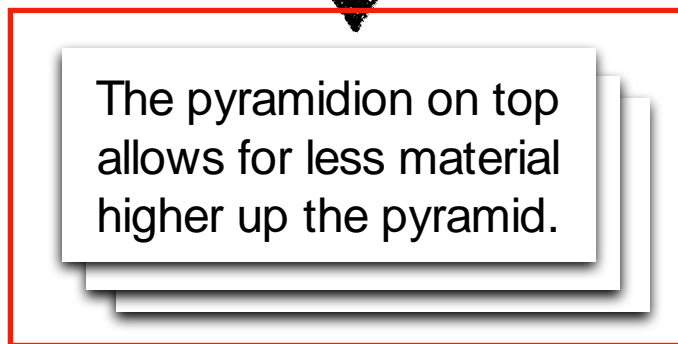
$q (=x)$



Index



θ : top-K retrieved chunks



$P_{\theta}(z|x)$

The pyramidion on top ... the pyramid.

...

The [MASK] at the top of the pyramid.



LM



pyramid

$P_{\theta}(y|x, z)$

REALM: Training

Objective: maximize $\sum_{z \in \theta} P_{\theta}(z|q) P_{\theta}(y|q, z)$

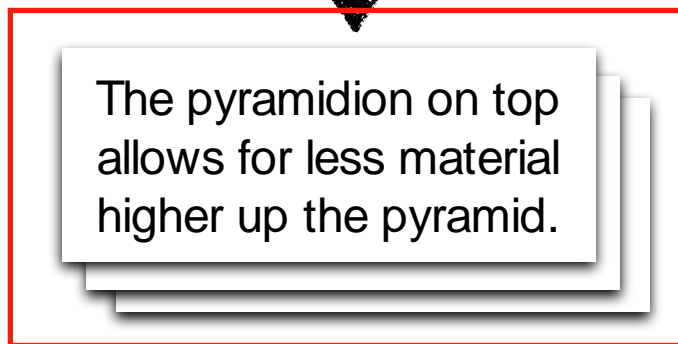
$q (=x)$



Index



θ : top-K retrieved chunks



$P_{\theta}(z|x)$

The pyramidion on top ... the pyramid.

...

The [MASK] at the top of the pyramid.



LM



pyramid

$P_{\theta}(y|x, z)$

Back-propagation



REALM: Training

Objective: maximize $\sum_{z \in \theta} P_{\theta}(z|q) P_{\theta}(y|q, z)$

$q (=x)$



Index



The pyramidion on top
allows for less material
higher up the pyramid.

$P_{\theta_{\text{new}}}(z|x)$

Up-to-date parameters

$P_{\theta_{\text{new}}}(y|x, z)$

The pyramidion on top ... the pyramid.

...

The [MASK] at the top of the pyramid.



LM



pyramid

Stale index;
Update every T steps



θ : top-K retrieved chunks

REALM: Index update rate

How often should we update the retrieval index?

- Frequency too high: expensive
- Frequency too slow: out-dated

REALM: Index update rate

How often should we update the retrieval index?

- Frequency too high: expensive
- Frequency too slow: out-dated

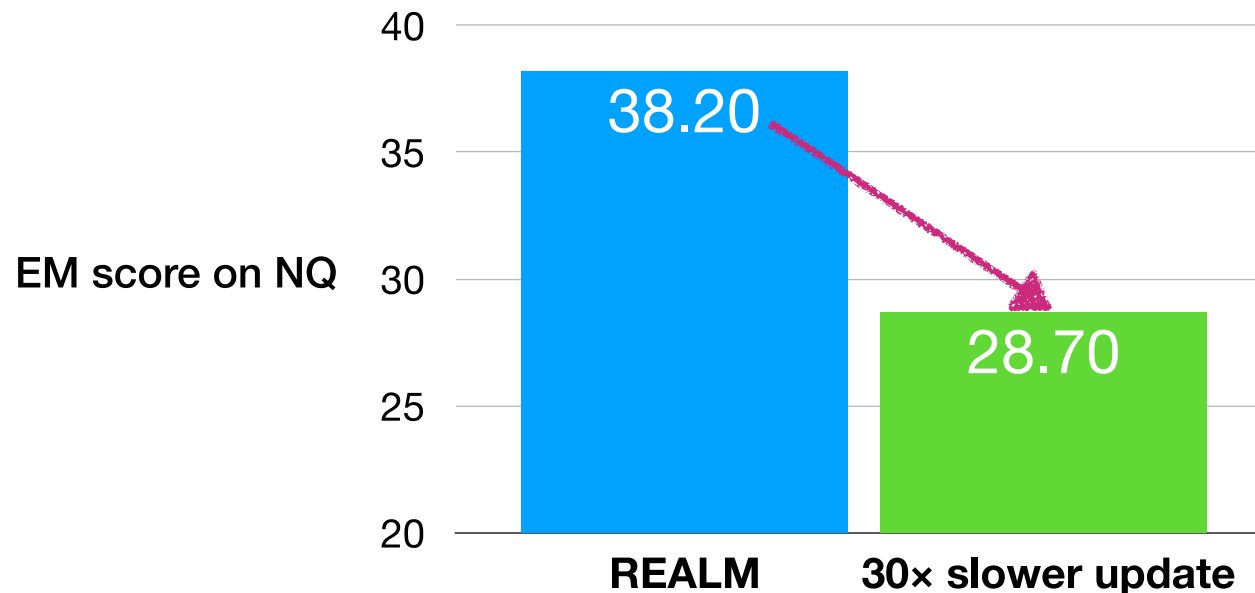
REALM: updating the index every 500 training steps

REALM: Index update rate

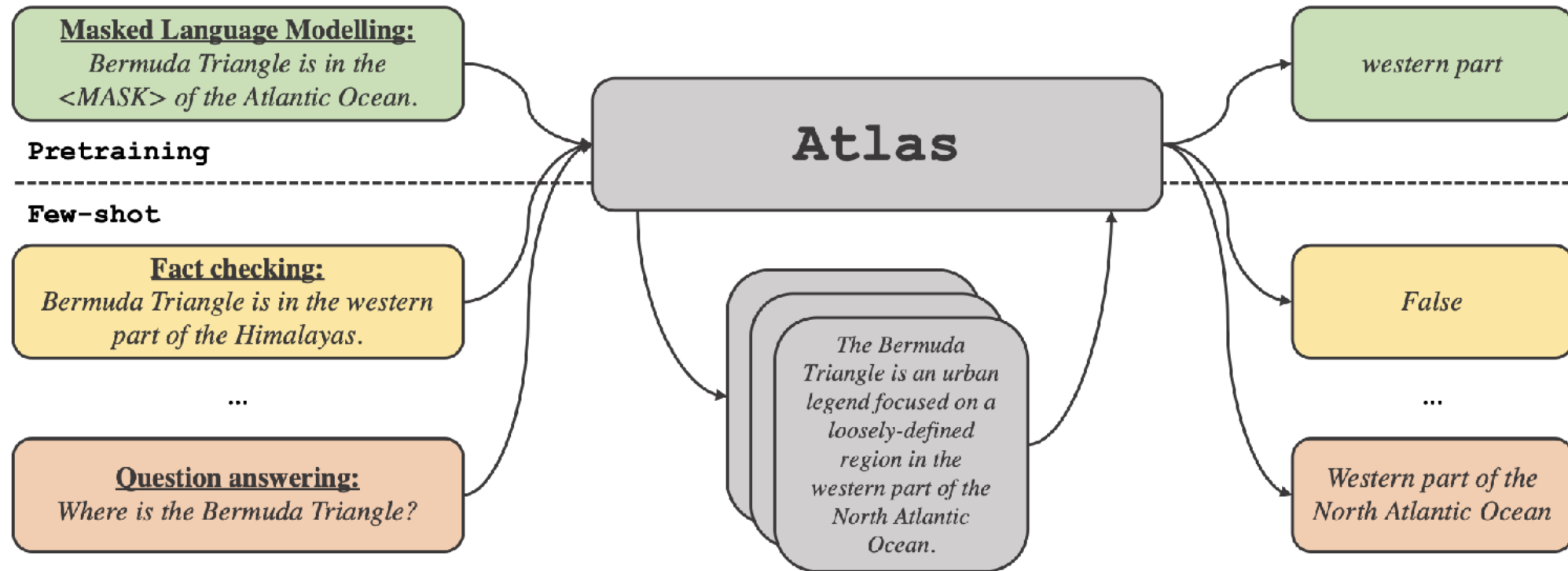
How often should we update the retrieval index?

- Frequency too high: expensive
- Frequency too slow: out-dated

REALM: updating the index every 500 training steps



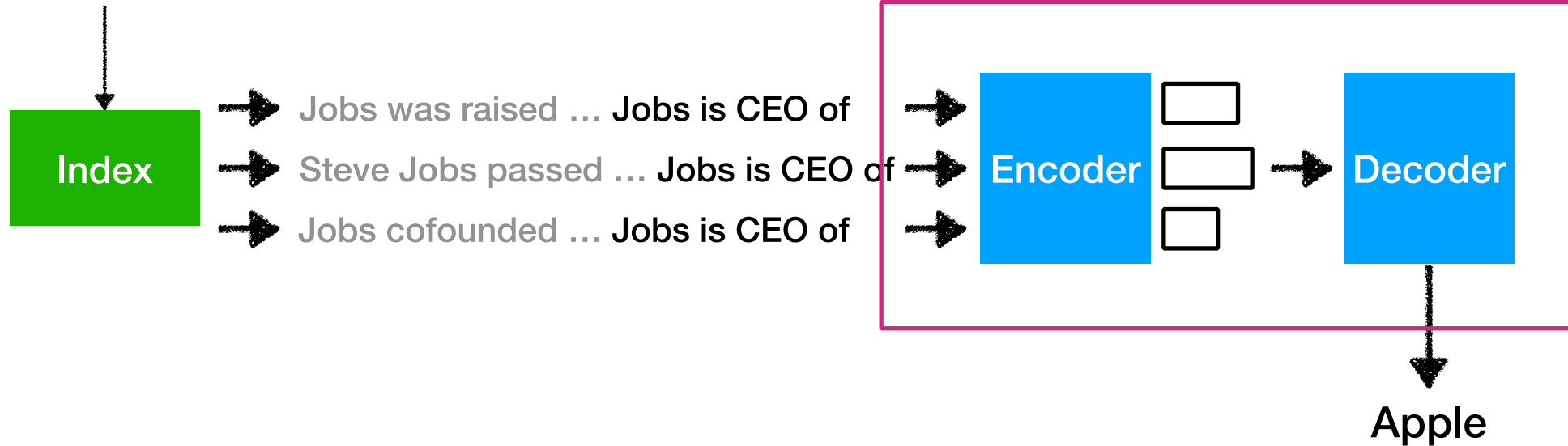
Atlas (Izacard et al. 2022)



Atlas (Izacard et al. 2022)

Retrieval-based encoder-decoder model

Jobs is CEO of __



Atlas (Izacard et al. 2022)

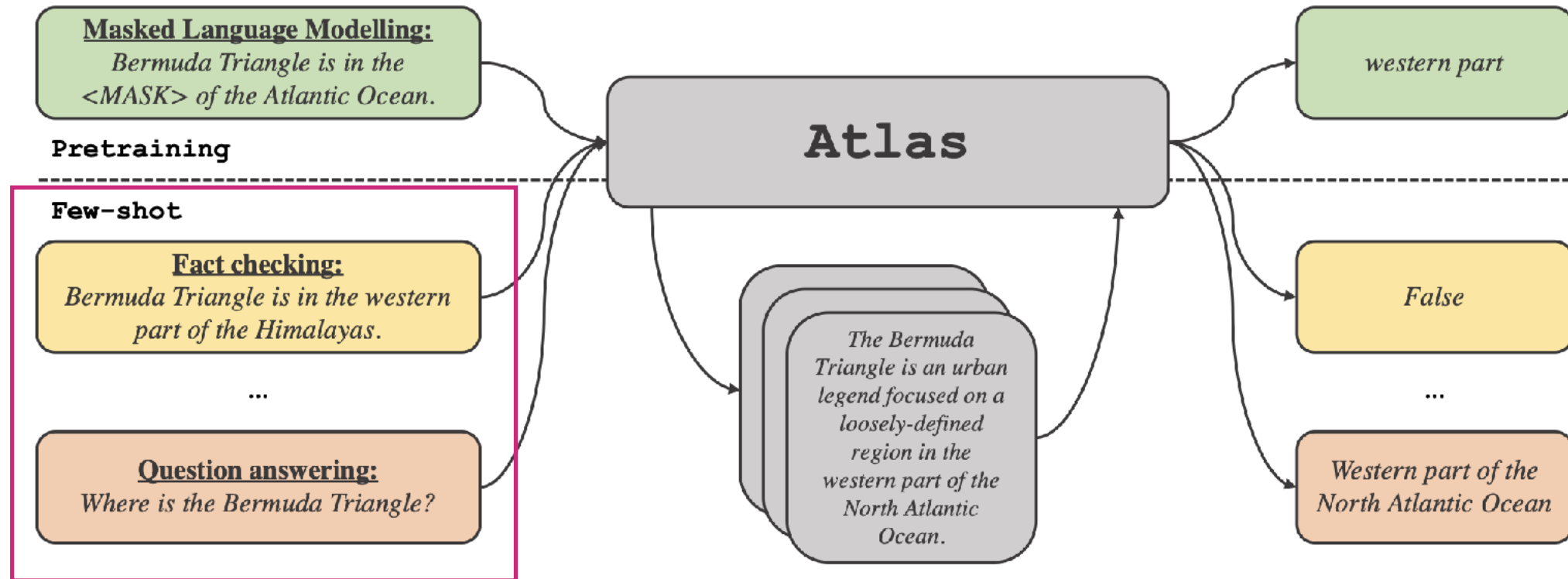
Jobs is CEO of __



Retrieve docs & Process each doc independently using "Fusion-in-Decoder"

Apple

Atlas (Izacard et al. 2022)



Adapted to a lot of downstream tasks! (Section 5)

Atlas: Retriever training

- In Atlas, the authors look at many different training loss functions.
- No loss function was found to be better in every setting.
- Perplexity distillation worked well in several settings
 - We will focus on that loss as an example.

	MLM	64-shot				1024-shot			
		NQ	WoW	FEVER	Avg.	NQ	WoW	FEVER	Avg.
Closed-book	1.083	6.5	14.1	59.0	26.5	10.7	16.5	75.3	34.2
No Joint pre-training	-	9.0	14.1	67.0	30.0	9.9	16.6	78.3	34.9
Fixed retriever	0.823	39.9	14.3	72.4	42.2	45.3	<u>17.9</u>	90.0	<u>51.1</u>
ADist	<u>0.780</u>	40.9	14.4	73.8	43.0	<u>46.2</u>	17.2	90.9	51.4
EMDR ²	0.783	<u>43.3</u>	<u>14.6</u>	72.1	43.3	44.9	18.3	85.7	49.6
PDist	0.783	45.0	15.0	77.0	45.7	44.9	<u>17.9</u>	<u>90.2</u>	51.0
LOOP	0.766	41.8	15.0	<u>74.4</u>	<u>43.7</u>	47.1	<u>17.9</u>	87.5	50.8

Atlas: Retriever training

Perplexity Distillation

Retrieve the text that can help LM encoders improve perplexity

$$P_{\text{retr}}(z | q) = \frac{\exp(s(z, q))}{\sum_{k=1}^K \exp(s(z_k, q))}$$

How likely each document is retrieved



$$P_{\text{ppl}}(z | q, y) = \frac{\exp(\log P_{\text{LM}}(y | q, z))}{\sum_{k=1}^K \exp(\log P_{\text{LM}}(y | q, z_k))}$$

How much each document improves the ppl

Atlas: Retriever training

Similarity based on
retrieval encoder

$$P_{\text{retr}}(z | q) = \frac{\exp(s(z, q))}{\sum_{k=1}^K \exp(s(z_k, q))}$$

How likely each document is retrieved

KL Divergence



Prob of the gold labels if
augmenting this text chunk

$$P_{\text{ppl}}(z | q, y) = \frac{\exp(\log P_{\text{LM}}(y | q, z))}{\sum_{k=1}^K \exp(\log P_{\text{LM}}(y | q, z_k))}$$

How much each document improves the ppl

Perplexity Distillation

RAG: Architecture & Training Summary

- Independent Retriever Training
 - DPR, Contriever
- Simple RAG (Frozen Retriever + Frozen LLM)
 - RALM, Scaling RAG
- Independent RAG LLM Training
 - RETRO, Self-RAG, RAG-RL
- Joint Retriever/LLM Training
 - REALM, ATLAS

RAG:

Architecture & Training Summary

- For a deeper overview of these methods, check out the following amazing resources:
 - [ACL 2023 Tutorial on Retrieval-Augmented Language Models \(> 3 hours of content\)](#)
 - [Douwe Kiela's \(Contextual AI\) Stanford talk](#)
 - [Akari Asai's talk](#)
- Much of this lecture, up until now, has been adapted from their wonderful slides.
- Rest of the talk will be a bit different.

Lecture Overview

- What is retrieval-augmented generation (RAG)?
- Why do we need retrieval-augmented generation (RAG)?
- RAG: Architecture and Training
- **Open Questions**
- Beyond RAG: LLM Continual Learning

Open Questions

- **How can retrieval be used in languages other than English?**
 - Chirkova et al 2024. Retrieval-augmented generation in multilingual settings.
- **How can retrieval augmentation be used in other modalities?**
 - Yasunaga et al 2023. Retrieval-Augmented Multimodal Language Modeling.
- **Can we decouple knowledge from generation? Memorization from generalization?**
- **RAG Evaluation**
 - As these systems get better, it becomes more and more challenging to evaluate them.
- **Agentic RAG**
 - DeepResearch, Perplexity, etc.

Open Questions: Continual Learning in LLMs

- Can conventional RAG systems truly address the fundamental LLM limitations we discussed earlier?

Open Questions: Continual Learning in LLMs

- I said RAG can address all of these!



Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

Open Questions: Continual Learning in LLMs

- I said RAG can address all of these! **I lied** 😊



Hallucinations

Lack of
Attribution

Lack of
Data Privacy

No Continual
Learning

Lecture Overview

- What is retrieval-augmented generation (RAG)?
- Why do we need retrieval-augmented generation (RAG)?
- RAG: Architecture and Training
- Open Questions
- Beyond RAG: LLM Continual Learning

Limitations of Conventional RAG:

LLM Continual Learning

- **Continual learning** is the process of **learning** new knowledge while **retaining** old knowledge.
- RAG is great at learning **new facts** without forgetting **older facts** (by adding new documents to the retrieval corpus).

Limitations of Conventional RAG:

LLM Continual Learning

- However, I argue that truly **learning from new knowledge** requires more than learning facts individually.
- **True learning** requires (1) forming **associations** between new facts AND (2) **making sense** of these associations in a larger context.

Limitations of Conventional RAG:

LLM Continual Learning

To illustrate these two requirements, I'll pose two imaginary scenarios:

1. **Associative Learning:** Research project scenario
2. **Sense-Making:** Comparing novels

Associative Learning

- **Imagine** that you working on a research project.
- You want to use some baseline Y from 4 years ago but keep encountering issues.
- You would love to know if someone at OSU who has used it before.
- If a person has learned about these two facts
 - (1) X is an OSU PhD student
 - (2) X has used baseline Y
- This person would easily direct you to person X.

Associative Learning

- However, if an LLM has these two facts in their retrieval corpus (and likely many others).
 - (1) X is an OSU PhD student
 - (2) X has used baseline Y
- The LLM would then need to retrieve the UNION of {OSU PhD students} and {people who have used this dataset}.
- Although this might be possible for agentic RAG systems, it is very inefficient and should be a very simple case of continual learning.

Sense-Making

- **Imagine** a world where **Harry Potter** and **Lord of the Rings** came out AFTER a language model was done pretraining.
- Now, if using conventional RAG, both of these would be added to a retrieval corpus that the LLM has access to.
- Let's now say that you read Harry Potter and loved the relationship between Harry and Ron. You are curious for other new books that have a similar relationship between the protagonist and supporting character?
- How would a retrieval-augmented LLM determine whether LoTR fits?

Sense-Making

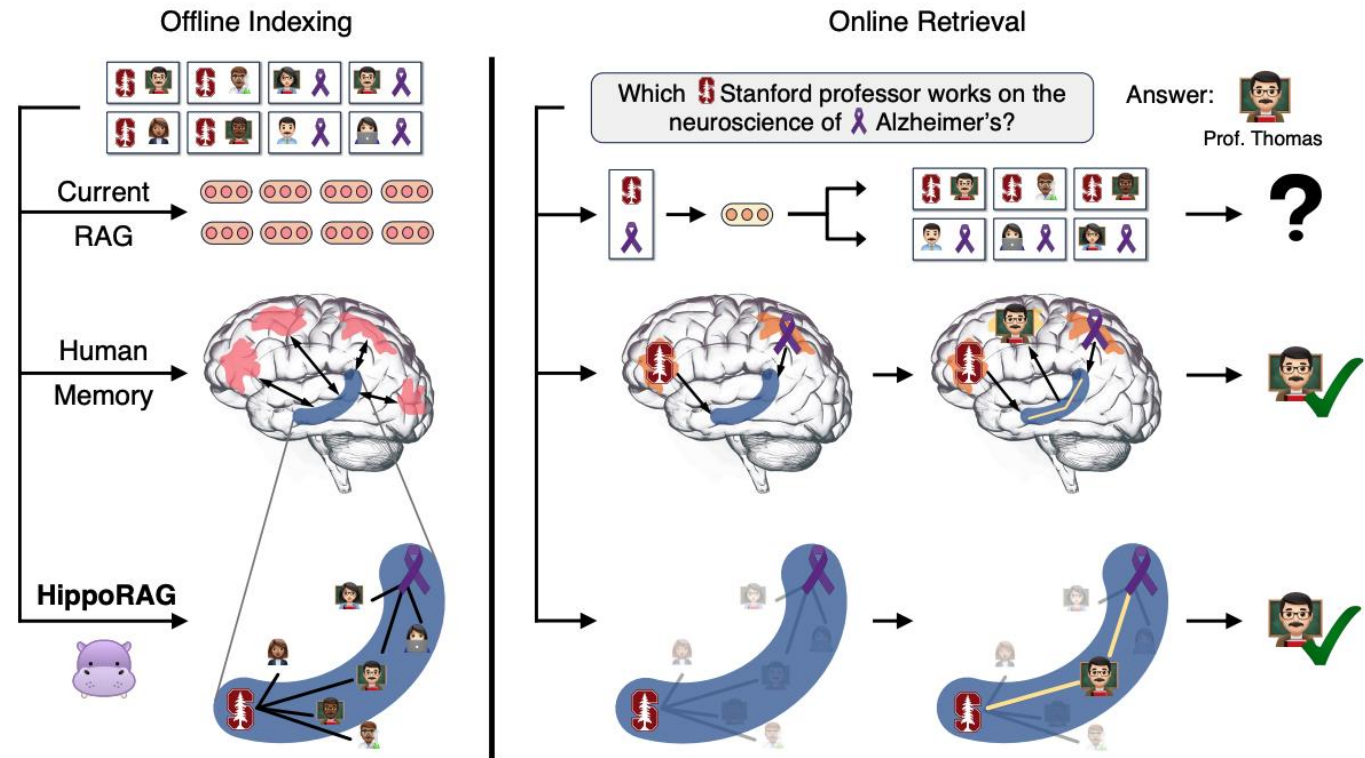
- Do you think a retrieval-augmented LLM could answer this question?

Sense-Making

- Perhaps something like DeepResearch could solve this problem by making a detailed plan and iteratively retrieving and thinking over thousands of tokens.
- Again, this is an extremely inefficient process to answer an extremely simple question.
- Humans are able to do associativity and sense-making with new information easily. Can we make LLMs do something similar?

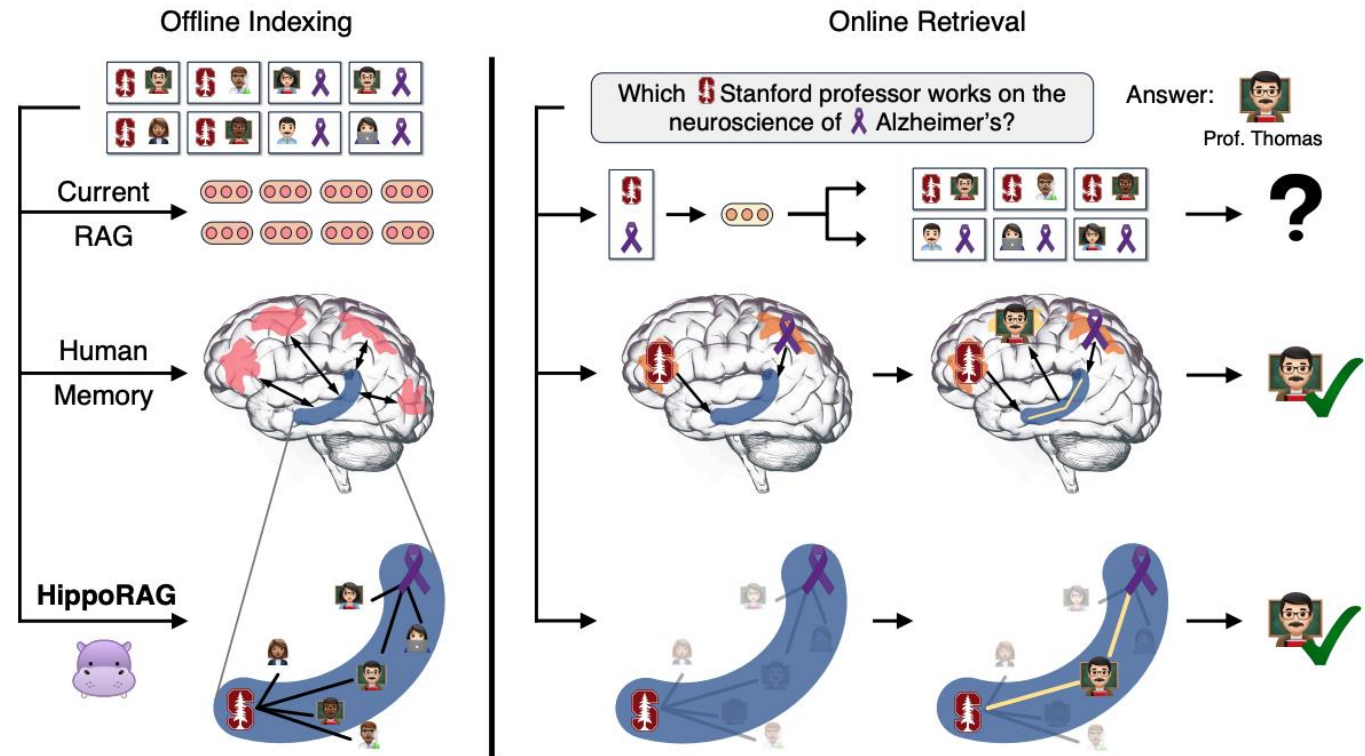
HippoRAG (Jiménez Gutiérrez et al. 2024)

- We mimic the associative learning power of human memory by:
 - Using LLMs to organize text into a KG
 - Leveraging the Personalized PageRank algorithm to traverse through the KG.



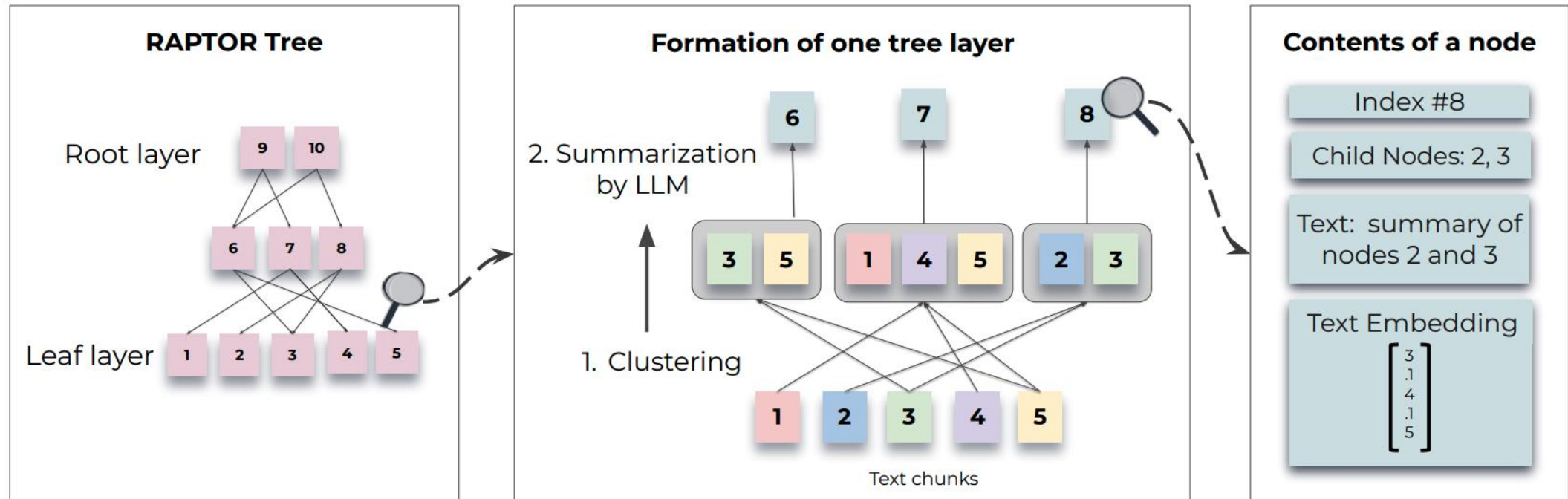
HippoRAG (Jiménez Gutiérrez et al. 2024)

- This allows the LLM to retrieve facts that were associated with one another **ONLY** in the retrieval corpus.
- In contrast with conventional RAG, where “relevance” or “similarity” are the only ways.



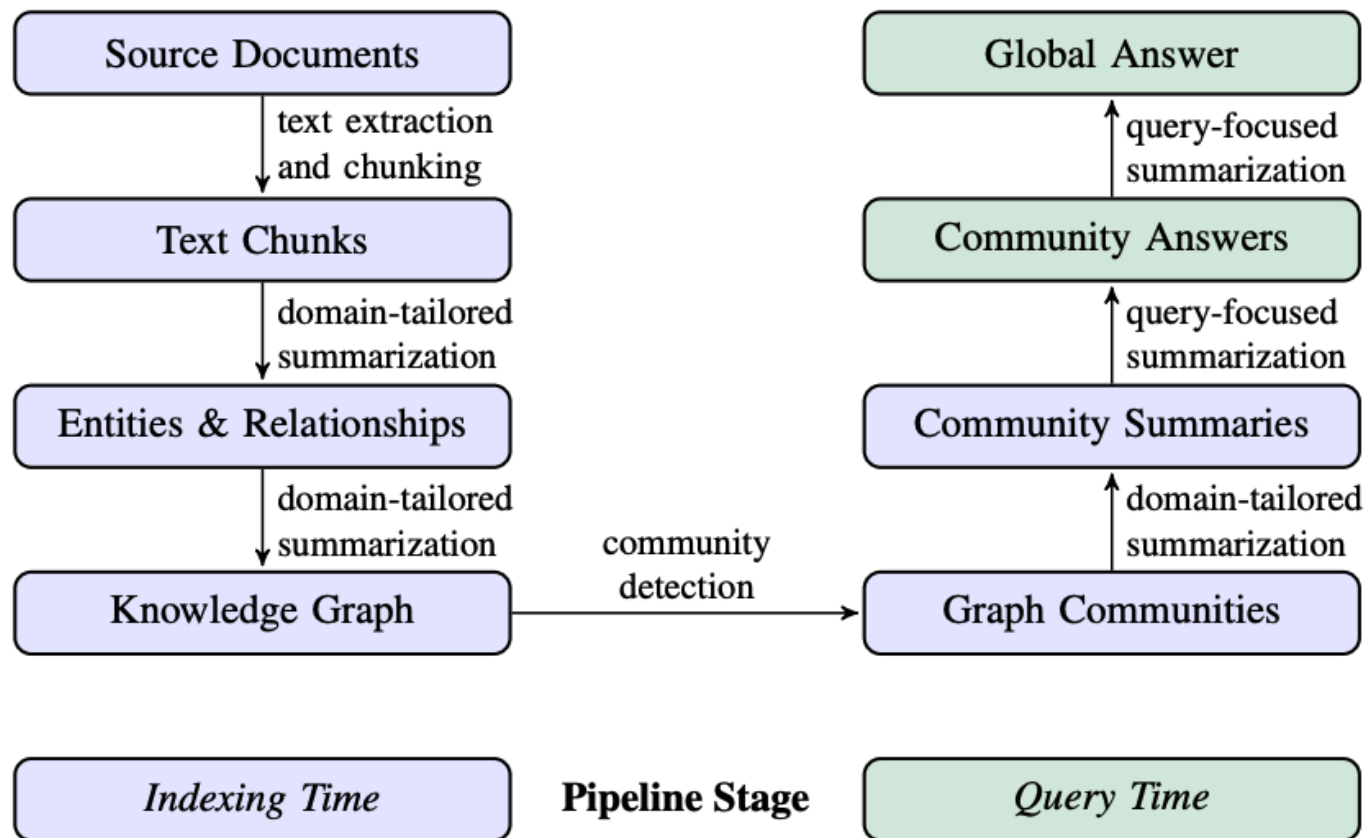
RAPTOR (Sarathi et al. 2024)

- In RAPTOR, the authors use an LLM to hierarchically summarize documents in the corpus together, deriving more and more abstract insights (sense-making).



GraphRAG (Edge et al. 2024)

- Variant of RAPTOR that uses an LLM constructed **graph** and **community detection methods** to derive abstract insights.



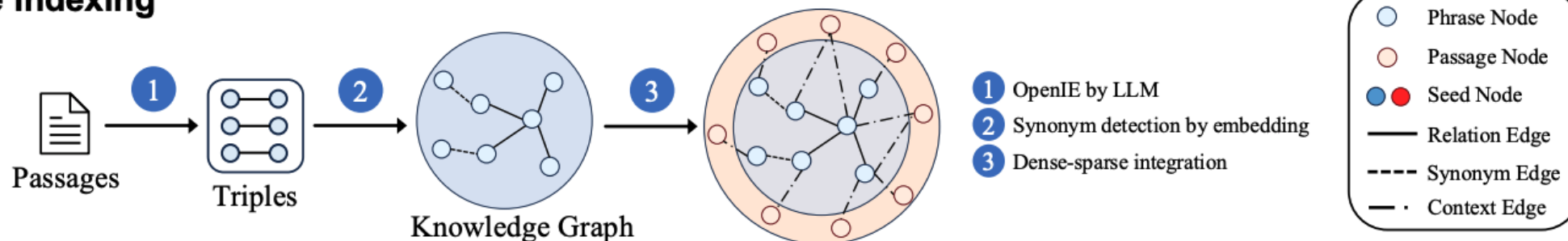
HippoRAG 2 (Jiménez Gutiérrez, Shu et al. 2025)

- Our latest work shows that modifications to HippoRAG leads to improvements in both **associativity** and **sense-making** while maintaining strong performance in **simple QA**.

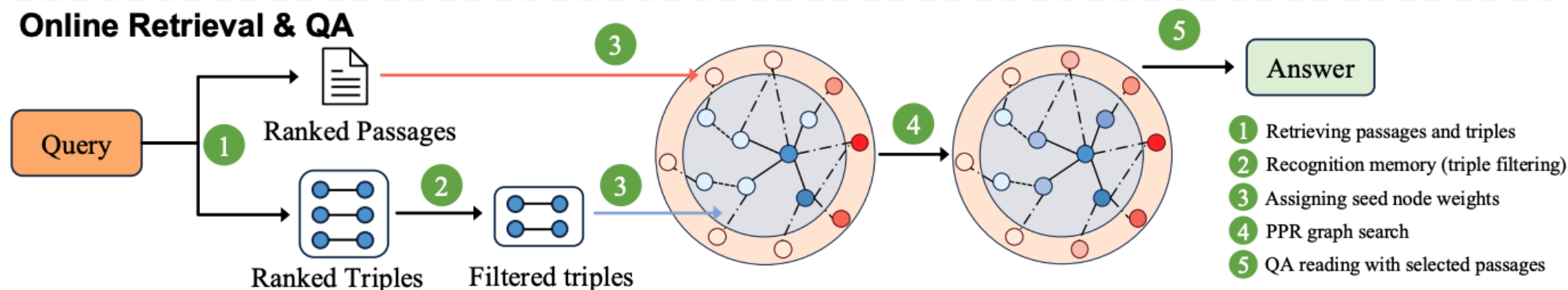
Retrieval	Simple QA		Multi-Hop QA				Discourse Understanding	Avg
	NQ	PopQA	MuSiQue	2Wiki	HotpotQA	LV-Eval	NarrativeQA	
Simple Baselines								
None	54.9	32.5	26.1	42.8	47.3	6.0	12.9	38.4
Contriever (Izacard et al., 2022)	58.9	53.1	31.3	41.9	62.3	8.1	19.7	46.9
BM25 (Robertson & Walker, 1994)	59.0	49.9	28.8	51.2	63.4	5.9	18.3	47.7
GTR (T5-base) (Ni et al., 2022)	59.9	56.2	34.6	52.8	62.8	7.1	19.9	50.4
Large Embedding Models								
GTE-Qwen2-7B-Instruct (Li et al., 2023)	62.0	56.3	40.9	60.0	71.0	7.1	21.3	54.9
GritLM-7B (Muennighoff et al., 2024)	61.3	55.8	44.8	60.6	73.3	9.8	23.9	56.1
NV-Embed-v2 (7B) (Lee et al., 2025)	61.9	55.7	45.7	61.5	75.3	9.8	25.7	57.0
Structure-Augmented RAG								
RAPTOR (Sarathi et al., 2024)	50.7	56.2	28.9	52.1	69.5	5.0	21.4	48.8
GraphRAG (Edge et al., 2024)	46.9	48.1	38.5	58.6	68.6	11.2	23.0	49.6
LightRAG (Guo et al., 2024)	16.6	2.4	1.6	11.6	2.4	1.0	3.7	6.6
HippoRAG (Gutiérrez et al., 2024)	55.3	55.9	35.1	71.8	63.5	8.4	16.3	53.1
HippoRAG 2	63.3	56.2	48.6	71.0	75.5	12.9	25.9	59.8

HippoRAG 2

Offline Indexing



Online Retrieval & QA



- It uses the same KG construction and PPR algorithm but integrates dense embeddings much more closely than before.

Future Work

- These works barely scratch the surface of this important problem.
- Much more work is needed to transform RAG into a **legitimate continual learning solution** for LLMs.

References

- Asai, Akari, et al. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection." Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), 2024.
- Borgeaud, Sebastian, et al. "Improving Language Models by Retrieving from Trillions of Tokens." Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 162, 2022, pp. 2206–2240. Available at: <https://proceedings.mlr.press/v162/borgeaud22a.html>.
- Chirkova, Nadezhda, et al. "Retrieval-Augmented Generation in Multilingual Settings." Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), Association for Computational Linguistics, 2024, pp. 177–188.
- Edge, Darren, et al. "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." arXiv, 2024, arXiv:2404.16130.
- Guu, Kelvin, et al. "REALM: Retrieval-Augmented Language Model Pre-Training." Proceedings of the 37th International Conference on Machine Learning (ICML 2020), 2020.

References

- Huang, Jerry, et al. "RAG-RL: Advancing Retrieval-Augmented Generation via Reinforcement Learning and Curriculum Learning." arXiv, 2025, arXiv:2503.12759.
- Izacard, Gautier, and Edouard Grave. "Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering." Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, 2021, pp. 874–880.
- Izacard, Gautier, et al. "Atlas: Few-Shot Learning with Retrieval-Augmented Language Models." Journal of Machine Learning Research, vol. 24, no. 1, 2023, article 251, 43 pp.
- Izacard, Gautier, et al. "Unsupervised Dense Information Retrieval with Contrastive Learning." Transactions on Machine Learning Research, 2022.
- Jiménez Gutiérrez, Bernal, et al. "From RAG to Memory: Non-Parametric Continual Learning for Large Language Models." arXiv, 2025, arXiv:2502.14802.
- Jiménez Gutiérrez, Bernal, et al. "HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models." Advances in Neural Information Processing Systems, vol. 37, 2024, pp. 59532–59569.

References

- Karpukhin, Vladimir, et al. "Dense Passage Retrieval for Open-Domain Question Answering." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), 2020, pp. 6769–6781.
- Kasai, Jungo, et al. "RealTime QA: What's the Answer Right Now?" Advances in Neural Information Processing Systems (NeurIPS 2023), 2023.
- Kim, Yujin, et al. "Carpe Diem: On the Evaluation of World Knowledge in Lifelong Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2024), 2024.
- Mallen, Jonathan, et al. "When Not to Trust Language Models: Investigating the Effectiveness of Parametric and Non-Parametric Memories." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), 2023.
- Min, Sewon, et al. "SILO Language Models: Isolating Legal Risk in a Non-Parametric Datastore." International Conference on Learning Representations (ICLR 2024), 2024.

References

- Ram, Ori, et al. "In-Context Retrieval-Augmented Language Models." Transactions of the Association for Computational Linguistics, vol. 11, 2023, pp. 1316–1331.
- Ramos, Juan. "Using TF-IDF to Determine Word Relevance in Document Queries." Proceedings of the First International Conference on Machine Learning and Data Mining, 2003.
- Robertson, Stephen, and Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond." Foundations and Trends in Information Retrieval, vol. 3, no. 4, 2009, pp. 333–389.
- Sarthi, Parth, et al. "RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval." International Conference on Learning Representations (ICLR 2024), 2024.
- Shao, Rulin, et al. "Scaling Retrieval-Based Language Models with a Trillion-Token Datastore." Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024. Available at: <https://openreview.net/forum?id=iAkhPz7Qt3>.

References

- Shao, Rulin, et al. "Scaling Retrieval-Based Language Models with a Trillion-Token Datastore." Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024. Available at: <https://openreview.net/forum?id=iAkhPz7Qt3>.
- Wang, Boxin, et al. "Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), 2023, pp. 7763–7786.
- Yasunaga, Michihiro, et al. "Retrieval-Augmented Multimodal Language Modeling." Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 202, 2023, pp. 39755–39769. Available at: <https://proceedings.mlr.press/v202/yasunaga23a.html>.
- Zhong, Zexuan, et al. "MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), 2023.