Retrieval-augmented LMs: Past, Present and Future

Large Language Models: Methods and Applications

Akari Asai (akari@cs.washington.edu)

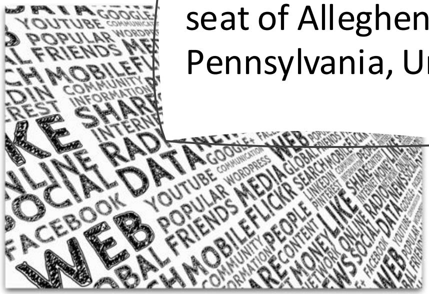
Feel free to post questions on Sli.do!
Sli.do code #2068655

How do normal parametric LLMs work?

Encapsulating everything in parameters by *pre-training* parameters on large-scale text corpora

Pittsburgh is a city in the county seat of Allegheny County, Pennsylvania, United States

Pre-training data

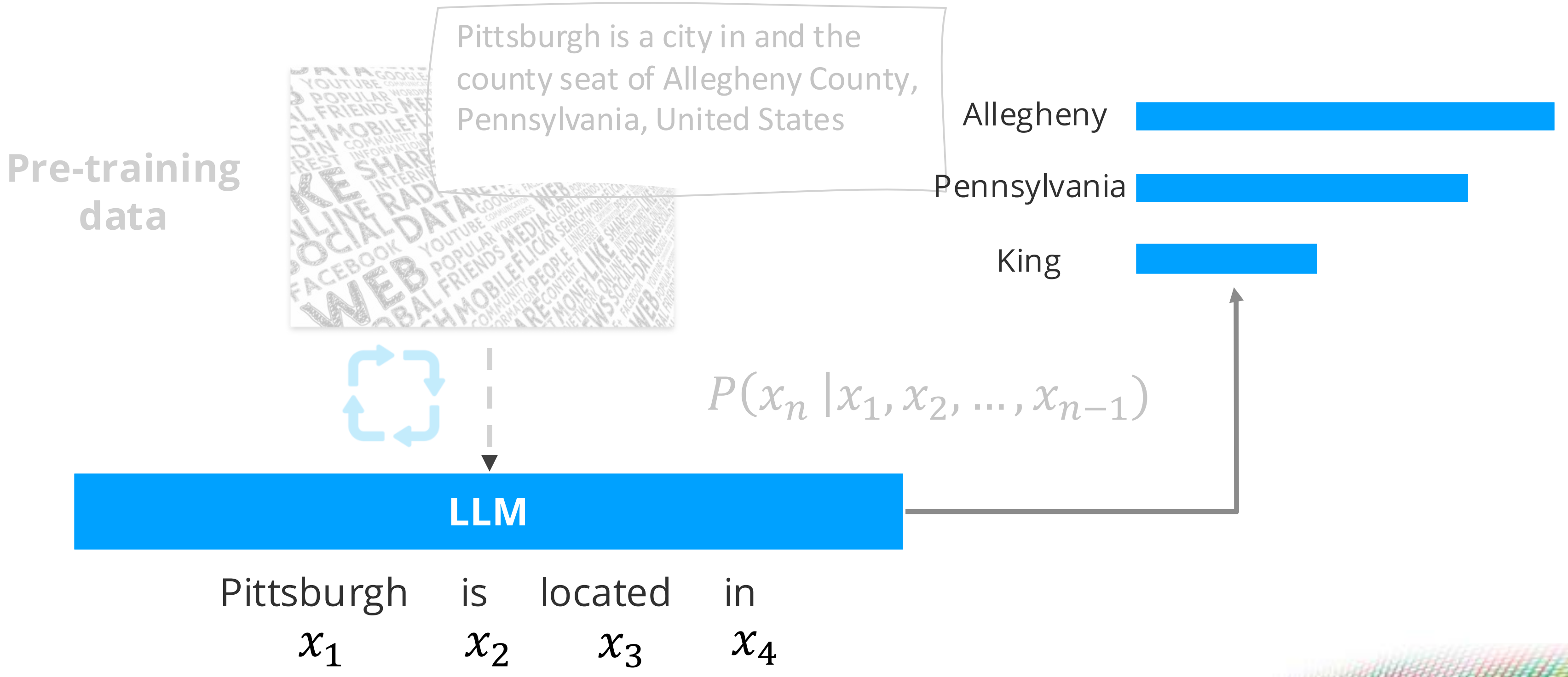


$$P(x_n \mid x_1, x_2, \ldots, x_{n-1})$$

LLM

How do normal parametric LLMs work?

Encapsulating everything in parameters by *pre-training* parameters on large-scale text corpora



Limitations of parametric LLMs #1: Hallucinations

LLMs cannot memorize everything in their parameters (yet), resulting in factual inaccuracy

Summarize Akari Asai's research including her representative papers.

Limitations of parametric LLMs #1: Hallucinations

LLMs cannot memorize everything in their parameters (yet), resulting in factual inaccuracy

Summarize Akari Asai's research including her representative papers.

Akari Asai is a prominent researcher in the field of natural language processing (NLP), with a focus on question answering (QA), information retrieval, and multi-lingual models. Her research revolves around building efficient and robust systems for extracting information from large corpora,

Limitations of parametric LLMs #1: Hallucinations

LLMs cannot memorize everything in their parameters (yet), resulting in factual inaccuracy

Summarize Akari Asai's research including her representative papers.

Akari Asai is a prominent researcher in the field of natural language processing (NLP), with a focus on question answering (QA), information retrieval, and multi-lingual models. Her research revolves around building efficient and robust systems for extracting information from large corpora,

Representative Papers:

1. **"XOR QA: Cross-lingual Open-Retrieval Question Answering" (2021)** In this paper, Asai introduced XOR QA, a benchmark for cross-lingual open-retrieval question answering, targeting QA across multiple languages. The model tackles the challenges of finding answers in languages different from the one in which the question is asked, highlighting the

Limitations of parametric LLMs #1: Hallucinations

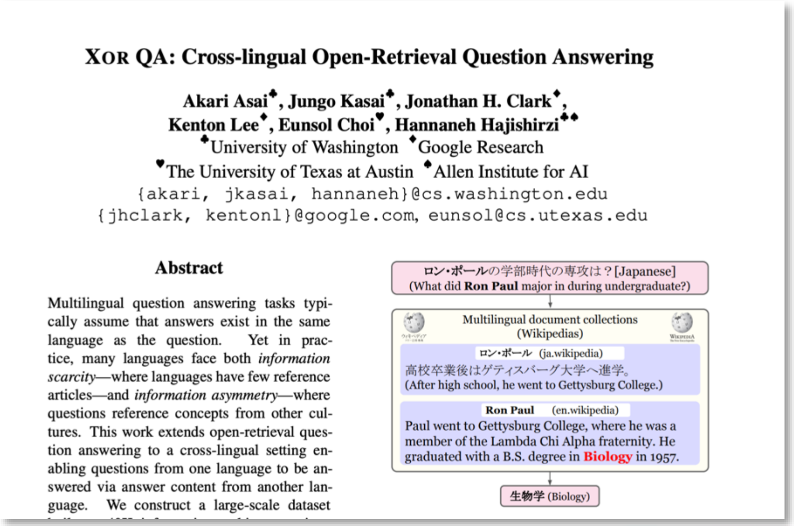
LLMs cannot memorize everything in their parameters (yet), resulting in factual inaccuracy

Summarize Akari Asai's research including her representative papers.

Akari Asai is a prominent researcher in the field of nat
on question answering (QA), information retrieval, anc
around building efficient and robust systems for extra

Representative Papers:

1. **"XOR QA: Cross-lingual Open-Retrieval Questio**
introduced XOR QA, a benchmark for cross-lingua
targeting QA across multiple languages. The mode
in languages different from the one in which the q



XOR QA: Cross-lingual Open-Retrieval Question Answering

Akari Asai[♣], **Jungo Kasai**[♣], **Jonathan H. Clark**[♦],
Kenton Lee[♦], **Eunsol Choi**[♥], **Hannaneh Hajishirzi**[♣♠]
[♣]University of Washington [♦]Google Research
[♥]The University of Texas at Austin [♠]Allen Institute for AI
{akari, jkasai, hannaneh}@cs.washington.edu
{jhclark, kentonl}@google.com, eunsol@cs.utexas.edu

Abstract

Multilingual question answering tasks typically assume that answers exist in the same language as the question. Yet in practice, many languages face both *information scarcity*—where languages have few reference articles—and *information asymmetry*—where questions reference concepts from other cultures. This work extends open-retrieval question answering to a cross-lingual setting enabling questions from one language to be answered via answer content from another language. We construct a large-scale dataset

Limitations of parametric LLMs #1: Hallucinations

LLMs cannot memorize everything in their parameters (yet), resulting in factual inaccuracy

Summarize Akari Asai's research including her representative papers.

Akari Asai is a prominent researcher in the field of natural language processing (NLP), with a focus on question answering (QA), information retrieval, and multi-lingual models. Her research revolves around building efficient and robust systems for extracting information from large corpora,

...

2. **"Dense Passage Retrieval for Open-Domain Question Answering" (2020)** This is one of her most cited works, where she co-authored the introduction of Dense Passage Retrieval (DPR). DPR is a state-of-the-art approach for efficiently retrieving passages relevant to a query by mapping both questions and passages into dense vector spaces using dual encoders. This method significantly improved the performance of open-domain QA systems.

Limitations of parametric LLMs #1: Hallucinations

LLMs cannot encapslate everything in their parameters yet.

Summarize Akari Asai's research including her representative papers.

Akari Asai is a prominent researcher in the field of natural language processing (NLP), with a focus on question answering (QA), information retrieval, and multi-lingual models. Her research revolves around building efficient and robust systems for extracting information from large corpora

2. **"Dense Passage Retrieval for Ope**
most cited works, where she co-au
DPR is a stat art approach
mapping bo ns and passa
method significantly improved the performance of open-domain QA systems.

Dense Passage Retrieval for Open-Domain Question Answering

Vladimir Karpukhin[*], Barlas Oğuz[*], Sewon Min[†], Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen[‡], Wen-tau Yih
Facebook AI [†]University of Washington [‡]Princeton University
{vladk, barlaso, plewis, ledell, edunov, scottyih}@fb.com
sewon@cs.washington.edu
danqic@cs.princeton.edu

Catastrophic incidents due to LLM hallucinations

Such LLM hallucinations have been causing many critical incidents in the real world

TECH · LAW

Humiliated lawyers fined $5,000 for submitting ChatGPT hallucinations in court: 'I heard about this new site, which I falsely assumed was, like, a super search engine'

BY RACHEL SHIN
June 23, 2023 at 9:41 AM PDT

Lawyers who filed legal documents with false citations generated by ChatGPT have been fined.
ERIK MCGREGOR—LIGHTROCKET/GETTY IMAGES

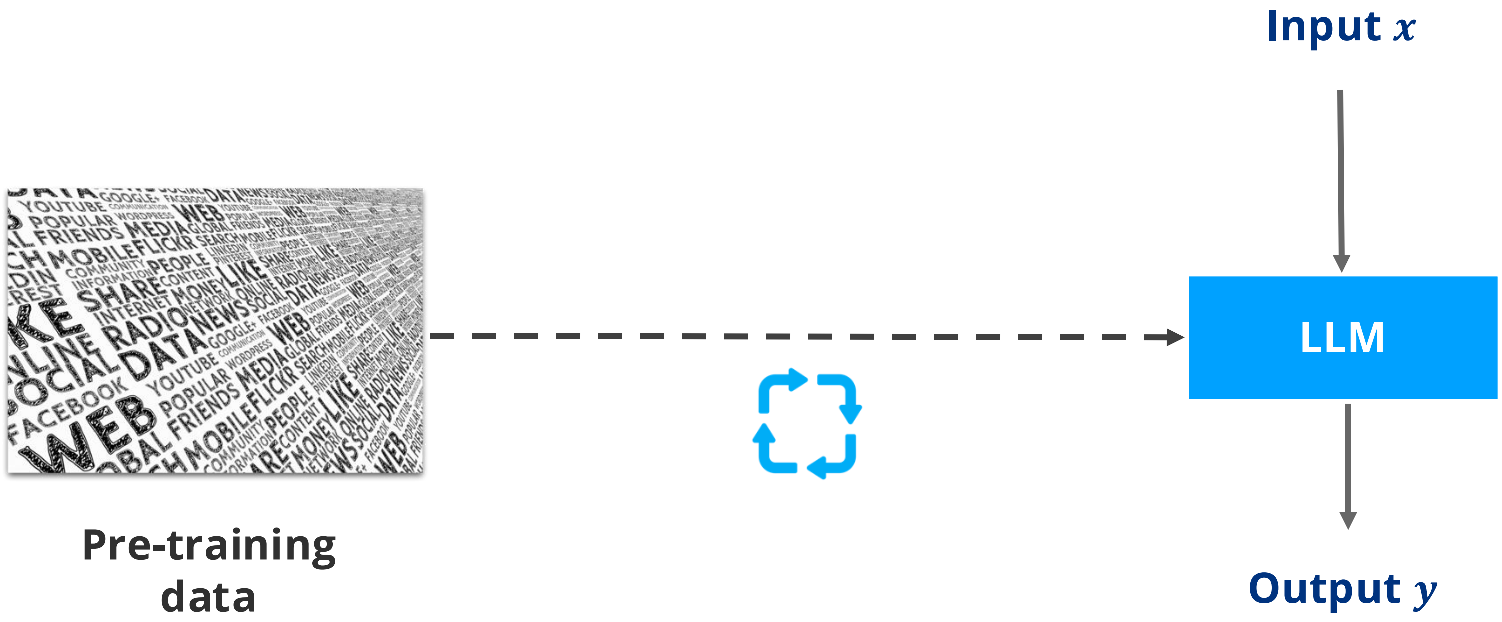
Air Canada must honor refund policy invented by airline's chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 12:12 PM

Retrieval-augmented LMs: Definitions & Notations

A new type of LMs that can use large-scale text data (datastore) at *inference-time*



Retrieval-augmented LMs: Definitions & Notations

A new type of LMs that can use large-scale text data (datastore) at *inference-time*

Input x

Retriever

LLM

Datastore

Retrieval-augmented LMs: Definitions & Notations

A new type of LMs that can use large-scale text data (datastore) at *inference-time*

Query q ← - - - - - - - - **Input x**

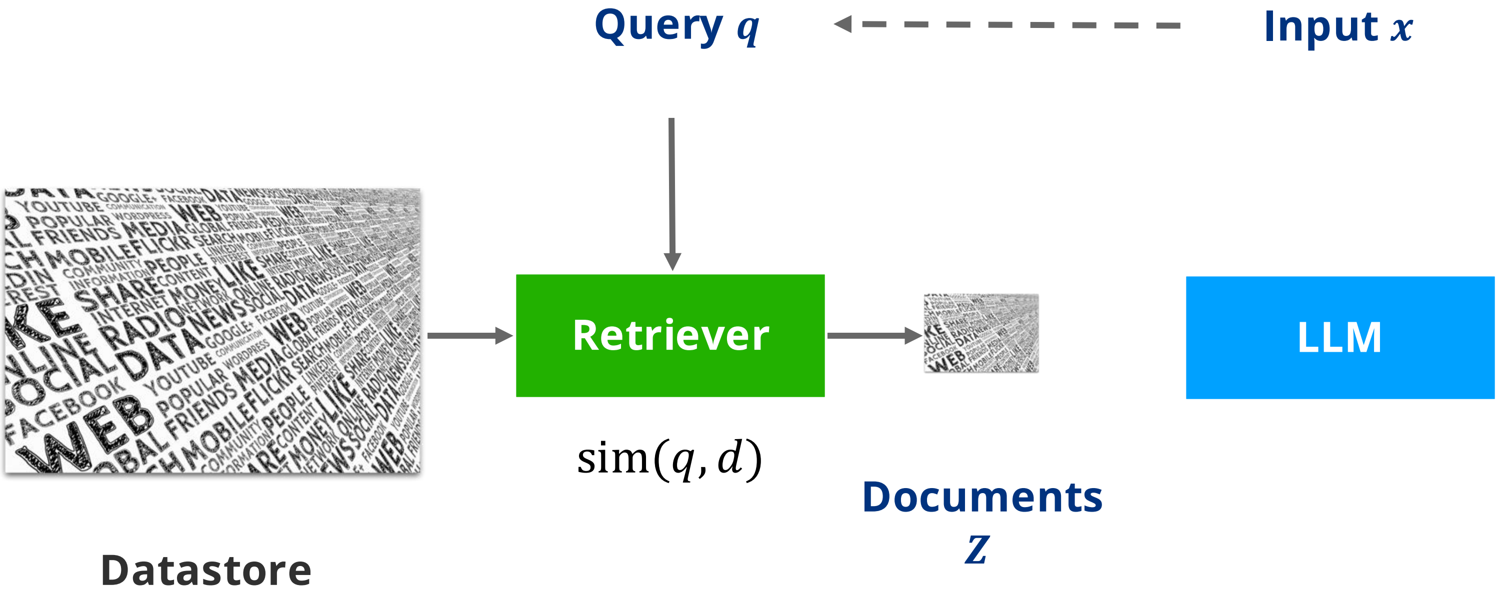
Datastore

Retriever

LLM

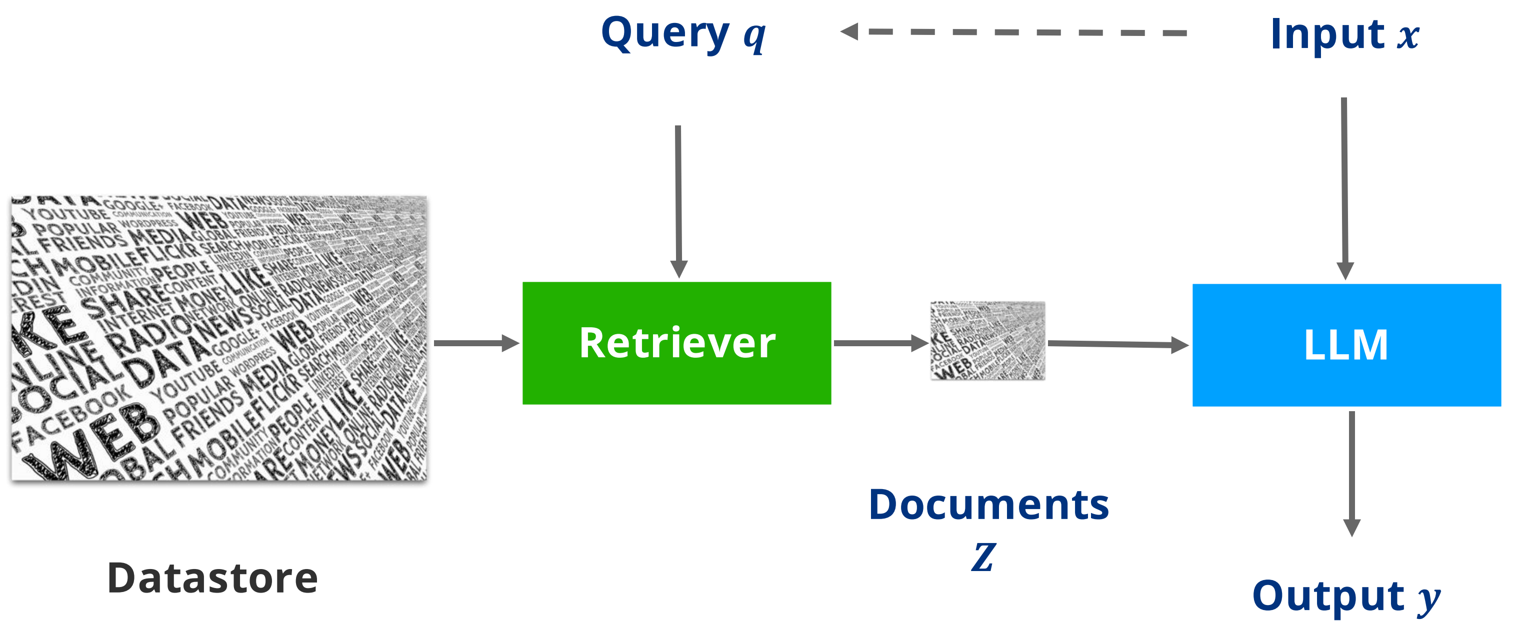
Retrieval-augmented LMs: Definitions & Notations

A new type of LMs that can use large-scale text data (datastore) at *inference-time*



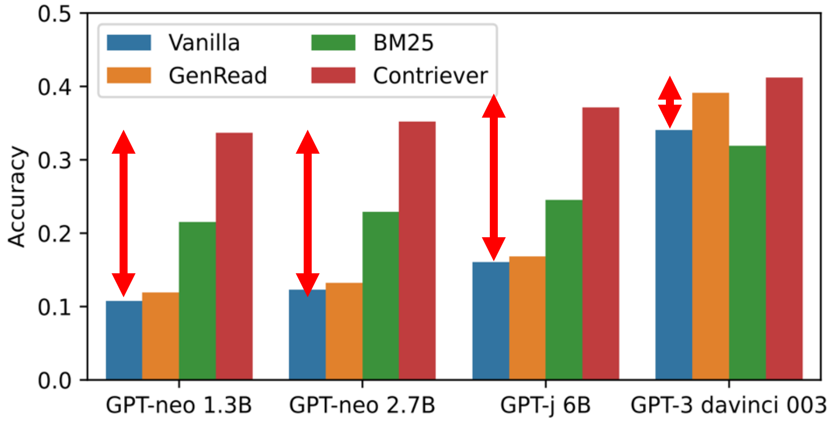
Retrieval-augmented LMs: Definitions & Notations

A new type of LMs that can use large-scale text data (datastore) at *inference-time*



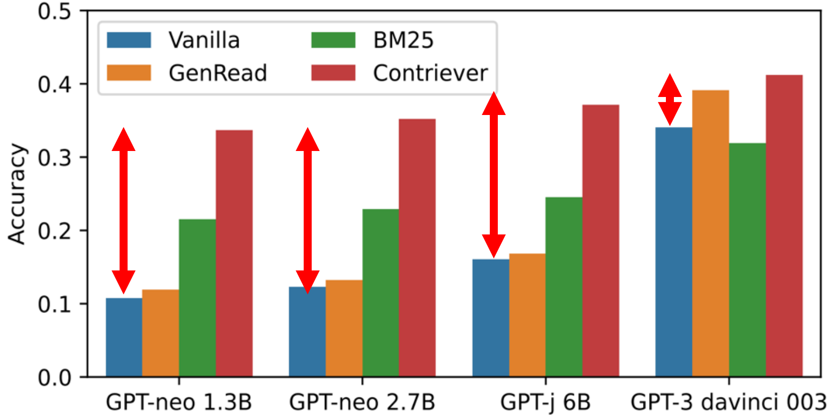
Benefit of retrieval-augmented LMs #1: reduce hallucinations

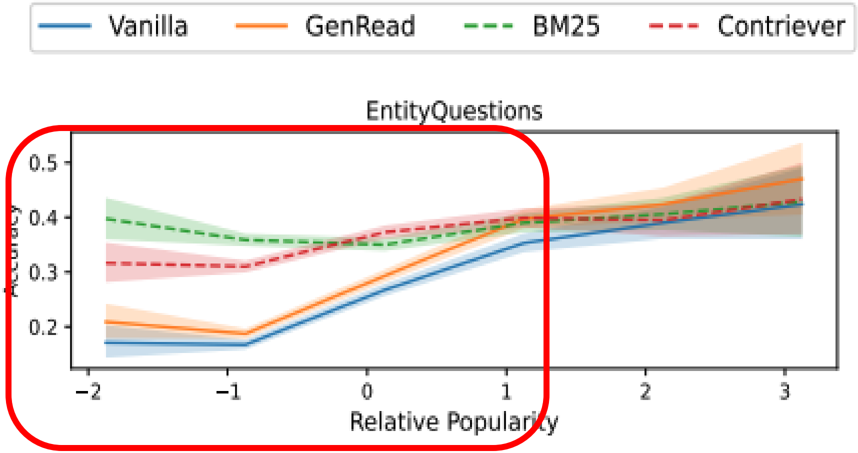
Retrieval-augmented LMs can reduce hallucinations, especially in long-tail knowledge



Benefit of retrieval-augmented LMs #1: reduce hallucinations

Retrieval-augmented LMs can reduce hallucinations, especially in long-tail knowledge





Quiz:

What are the other benefits of using retrieval-augmented LMs?

Benefit of retrieval-augmented LMs #2: Adaptations w/o training

Parametric LMs' knowledge gets obsolete quickly & requires continuous training

Who is the current prime minister of UK?

The current Prime Minister of the United Kingdom is **Rishi Sunak**. He has held the position since **October 25, 2022**, following the resignation of Liz Truss. Sunak is the leader of the Conservative Party and previously served as Chancellor of the Exchequer.

"RealTime QA: What's the Answer Right Now?" Kasai et al. NeurIPS (Benchmark). 2023

Benefit of retrieval-augmented LMs #2: Adaptations w/o training

Parametric LMs’ knowledge gets obsolete quickly & requires continuous training

Who is the current prime minister of UK?

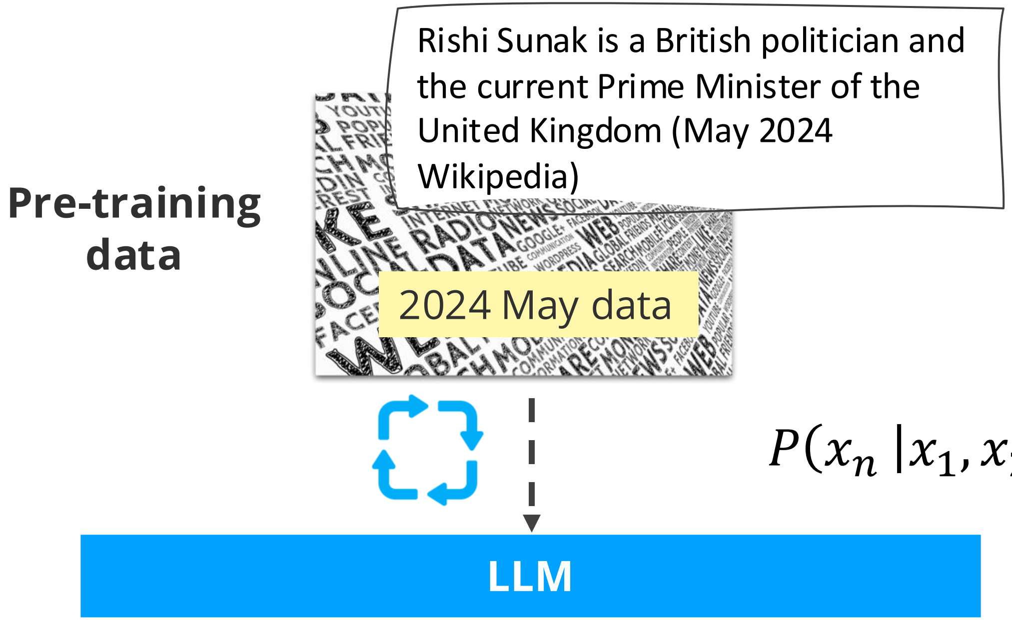
The current Prime Minister of the Un… gdom is **Rishi Sunak**. He has held the position since **October 25, 2022**, following the re… f Liz Truss. Sunak is the leader of the Conservative Party and previously served as Chan… r the Exchequer.

Portrait	Prime minister Office (Lifespan)	Term of office			Mandate[a]
		Start	End	Duration	
	Rishi Sunak [98] MP for Richmond (Yorks) (born 1980) *Premiership*	25 October 2022	5 July 2024	1 year, 255 days	–
	Keir Starmer [99] MP for Holborn and St Pancras (born 1962) *Premiership*	5 July 2024	*Incumbent*	73 days	2024

“RealTime QA: What's the Answer Right Now?” Kasai et al. NeurIPS (Benchmark). 2023

Benefit of retrieval-augmented LMs #2: Adaptations w/o training

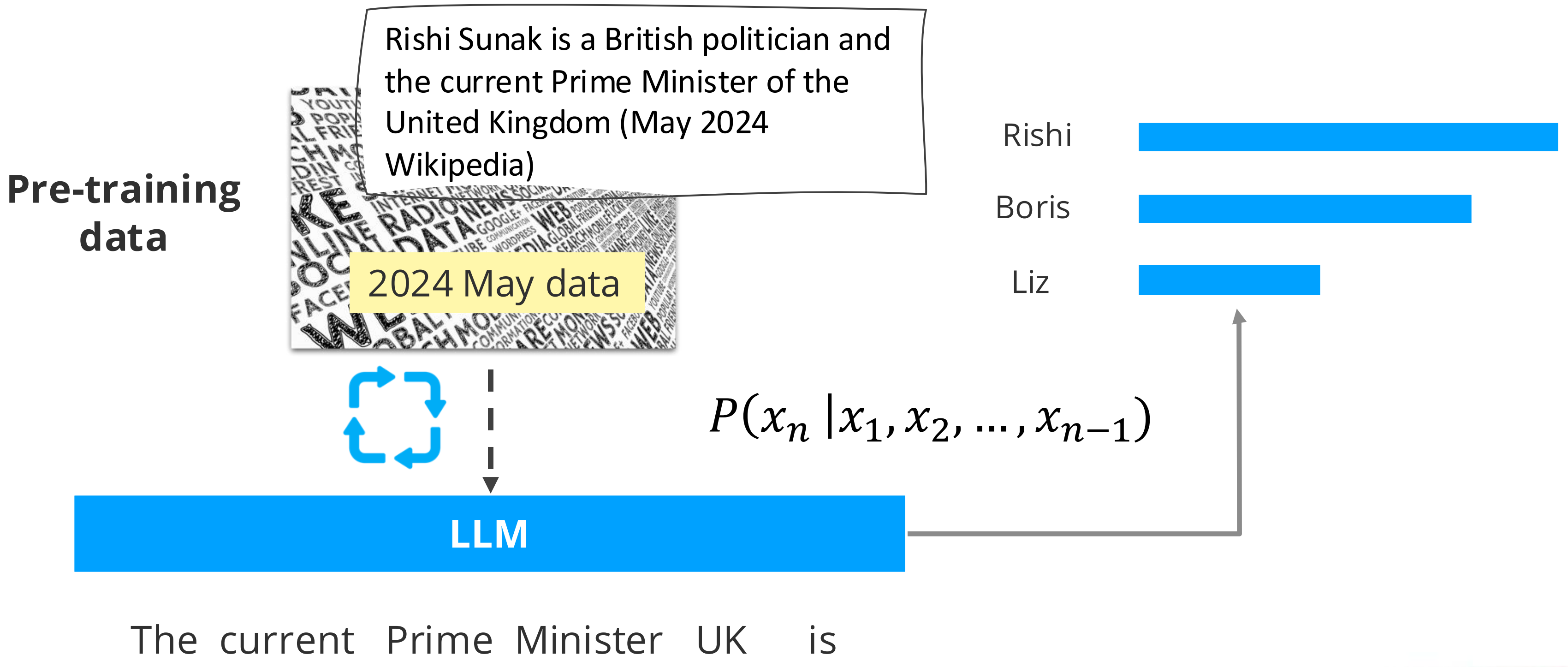
Parametric LMs' knowledge gets obsolete quickly & requires continuous training



$$P(x_n \mid x_1, x_2, \ldots, x_{n-1})$$

Benefit of retrieval-augmented LMs #2: Adaptations w/o training

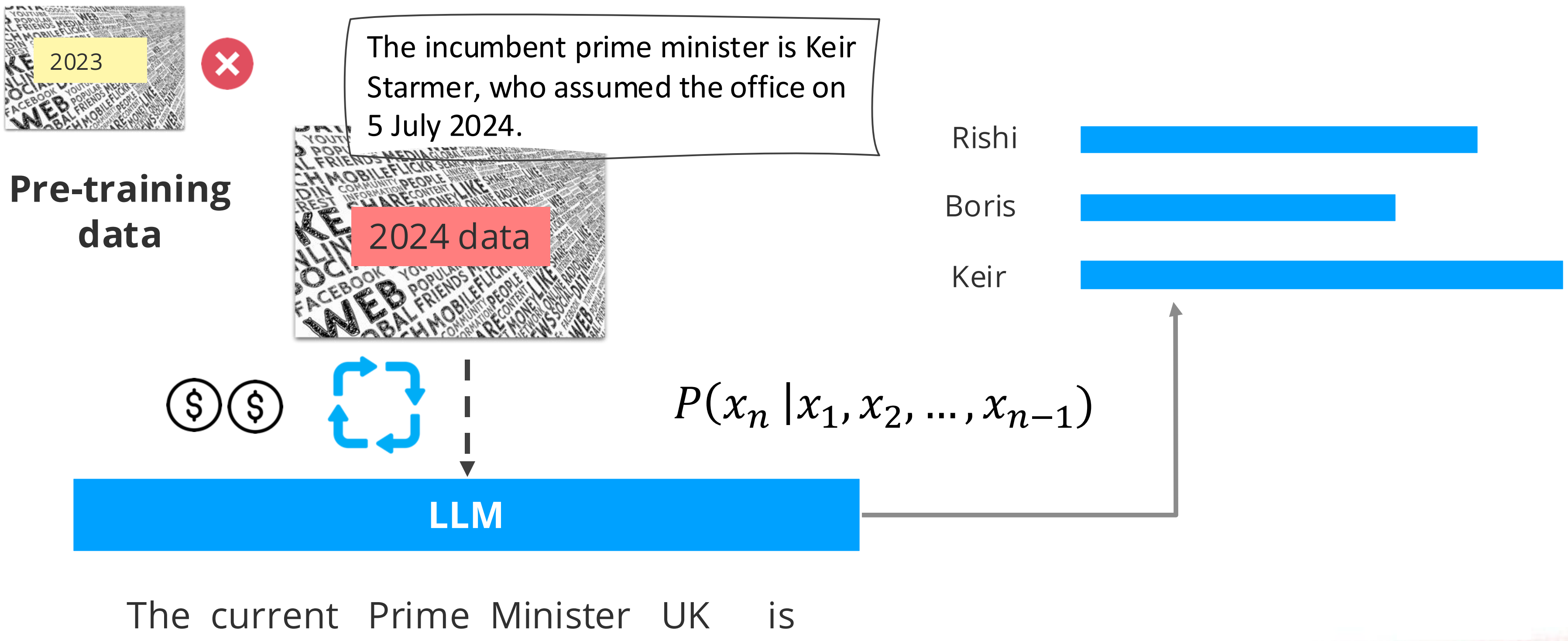
Parametric LMs’ knowledge gets obsolete quickly & requires continuous training



“RealTime QA: What's the Answer Right Now?” Kasai et al. NeurIPS (Benchmark). 2023

Benefit of retrieval-augmented LMs #2: Adaptations w/o training

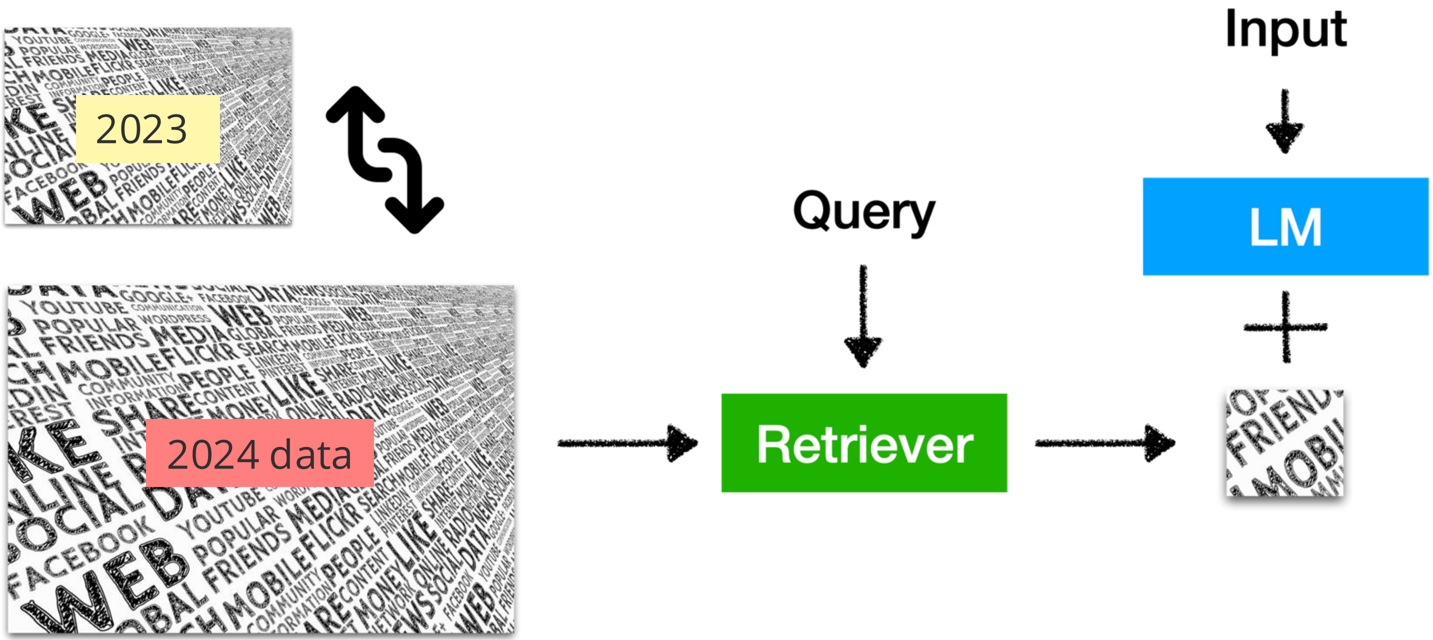
Parametric LMs' knowledge gets obsolete quickly & requires continuous training



"RealTime QA: What's the Answer Right Now?" Kasai et al. NeurIPS (Benchmark). 2023

Benefit of retrieval-augmented LMs #2: Adaptations w/o training

We can easily swap datastores for retrieval-augmented LMs for new data distributions



Benefit of retrieval-augmented LMs #3: Providing attributions

Retrieval-augmented LMs can provide a small number of documents as attributions

Who is the current prime minister of United Kingdom?

Searched 6 sites

As of September 2024, the current Prime Minister of the United Kingdom is **Keir Starmer**, leader of the Labour Party. He took office following a general election in July 2024, where Labour secured a significant victory, unseating the Conservative Party's Rishi Sunak, who had been Prime Minister since 2022(Northeastern Global News)(Wikipedia).

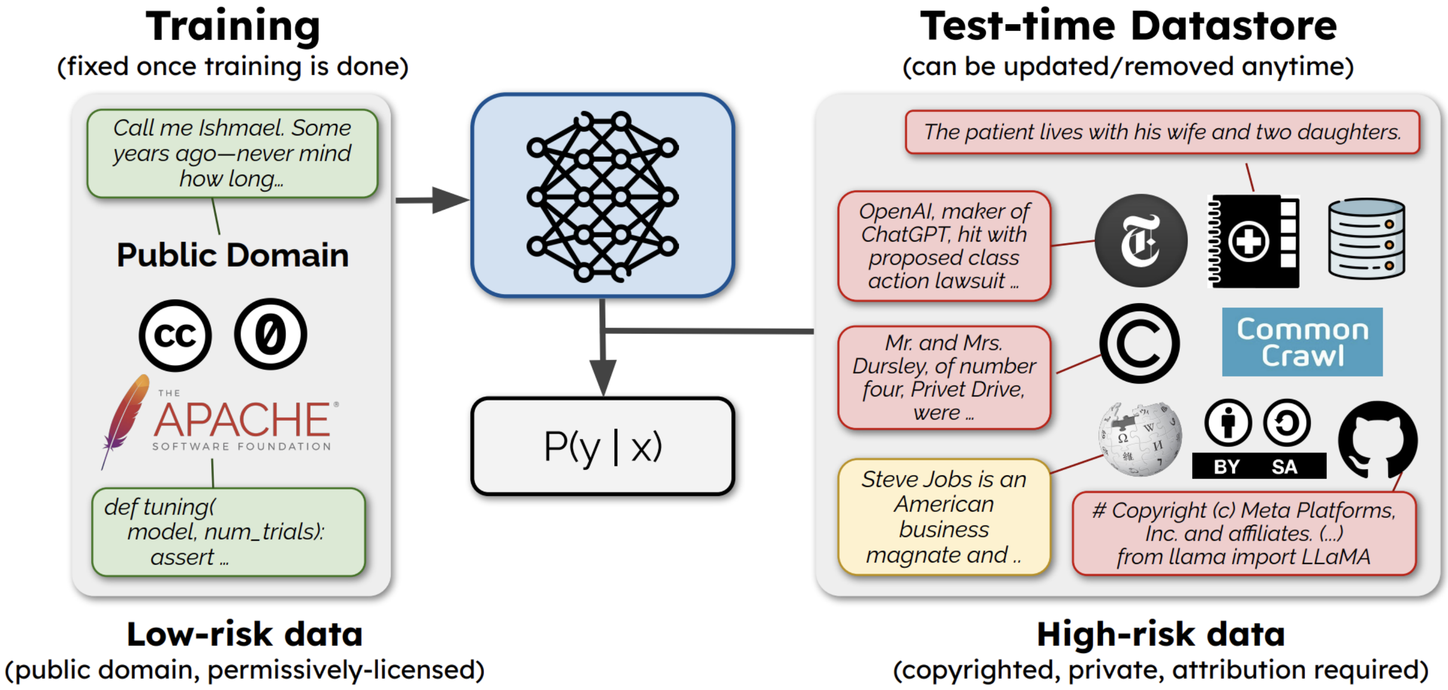
Government of the United Kingdom [edit]

Prime Minister	Portrait	Since	Party	Ref
Keir Starmer		5 July 2024 (2 months ago)	Labour	[1]

“Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models”. Bohnet et al. ArXiv 2020.

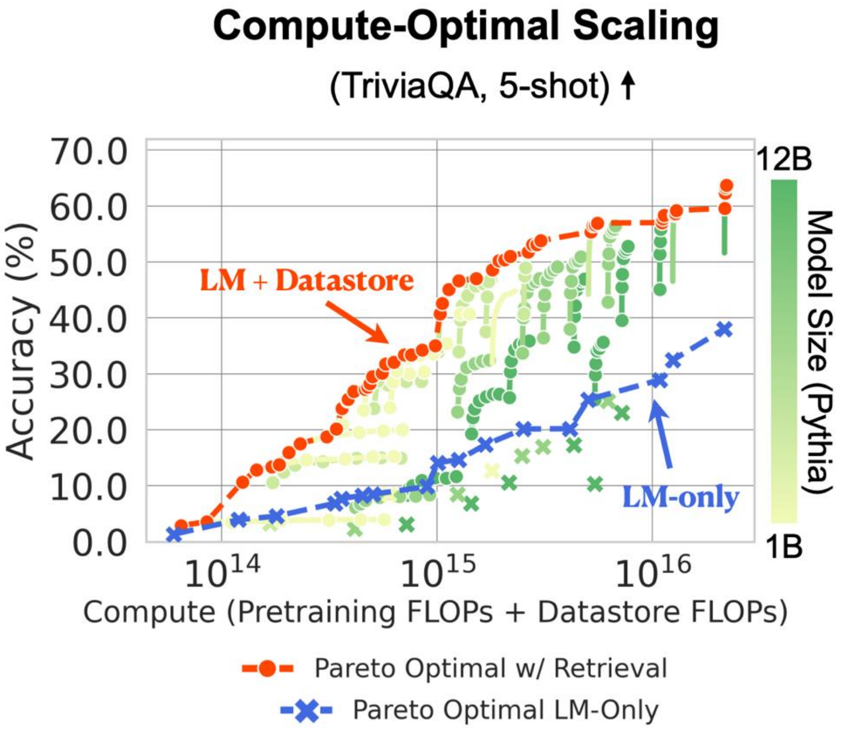
Benefit of retrieval-augmented LMs #4: Flexible data opt-in / out

We can incorporate or remove high-risk data dynamically at inference, not training time



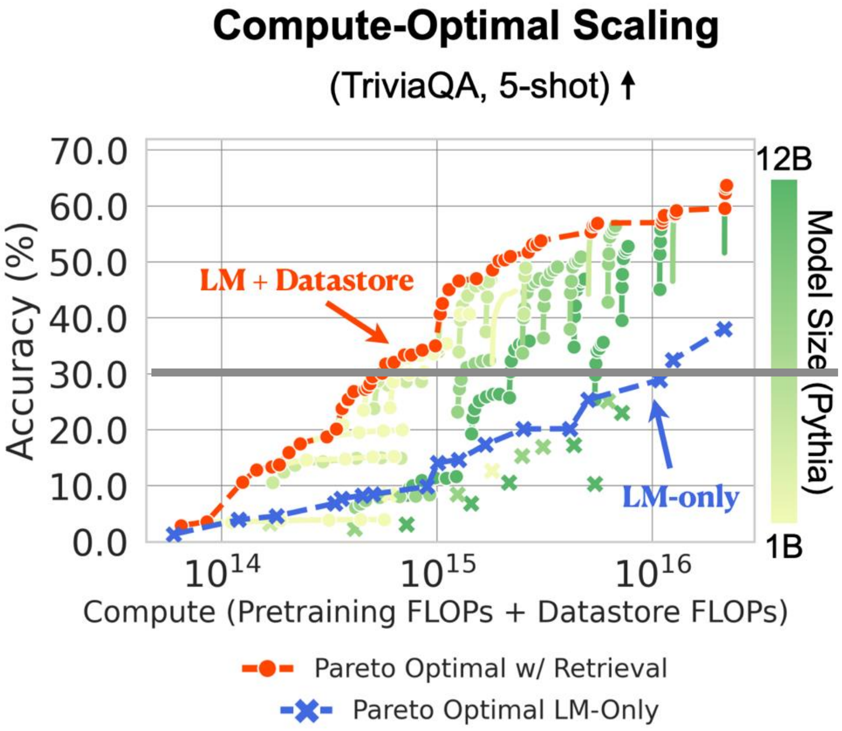
Benefit of retrieval-augmented LMs #5: parameter efficiency

Retrieval-augmented LMs can be much more parameter efficient and compute-optimal



Benefit of retrieval-augmented LMs #5: parameter efficiency

Retrieval-augmented LMs can be much more parameter efficient and compute-optimal



FLOPs for pre-training large LMs > **FLOPs for pre-training small LMs + construct datastore**

"Scaling Retrieval-Based Language Models with a Trillion-Token Datastore." Shao, He, Asai et al., ArXiv 2024.

Retrieval-augmented LMs have been widely used!

Retrieval-augmented LMs have been widely used both in academia and industry

cohere Gemini perplexity

Retrieval-augmented LMs have been widely used!

Retrieval-augmented LMs have been widely used both in academia and industry

cohere Gemini perplexity

How RAG-Powered AI Applications Have A Positive Impact On Businesses

Forbes Technology Council

Jyotishko Biswas Forbes Councils Member
Forbes Technology Council COUNCIL POST | Membership (Fee-Based)

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Subscribe About Archive BAIR

The Shift from Models to Compound AI Systems

"60% of LLM applications use some form of retrieval-augmented generation (RAG)"

Today's outline

1. ***Introduction:*** *What are retrieval-augmented LMs? Why do we want to use them?*
2. ***Past:*** *Architecture and training of retrieval-augmented LMs for downstream tasks*
3. ***Present:*** *Retrieval-augmented generation with LLMs*
4. ***Future:*** *Limitations & future directions*

Feel free to post questions on Sli.do!
Sli.do code #2068655

Past: *Architecture and training of retrieval-augmented LMs for downstream tasks*

Brief history of retrieval-augmented LMs development

RAG was initially extensively studied for certain NLP tasks, namely Question Answering

2017: DrQA

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

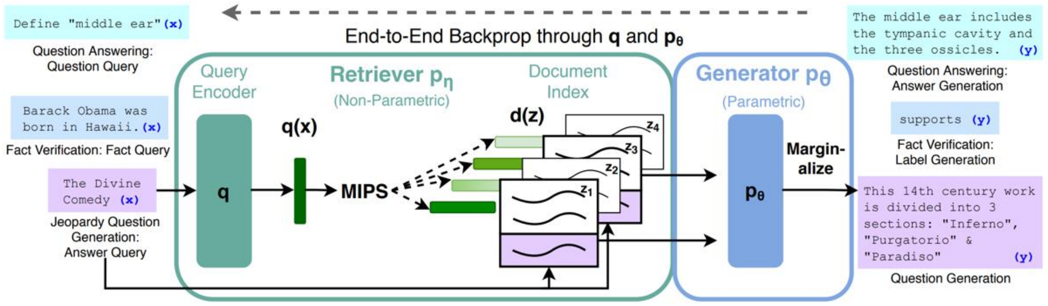
Retrieve and ***read*** Wikipedia articles for open-domain QA

"Reading Wikipedia to Answer Open-Domain Questions." Chen et al., ACL 2017.

Brief history of retrieval-augmented LMs development

RAG was initially extensively studied for certain NLP tasks, namely Question Answering

- **2017:** DrQA
- **2019:** ORQA
- **2020:** RALM, RAG



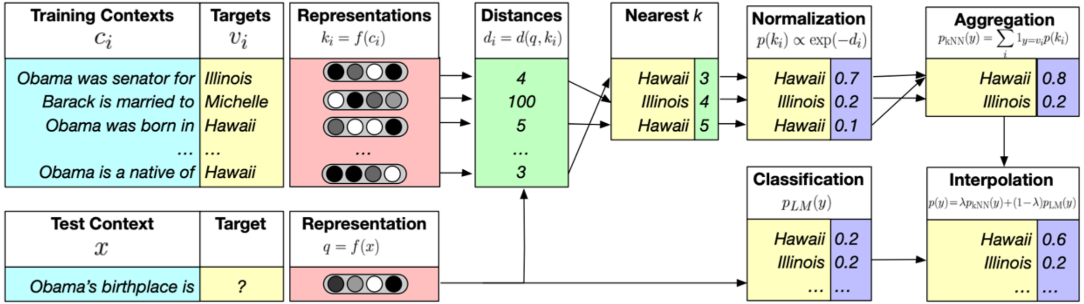
End-to-end pre-training → fine-tuning of retriever & LM

“Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” Lewis et al., NeurIPS 2020.

Brief history of retrieval-augmented LMs development

RAG was initially extensively studied for certain NLP tasks, namely Question Answering

- **2017:** DrQA
- **2019:** ORQA
- **2020:** RALM, RAG
- 2020: kNN LM



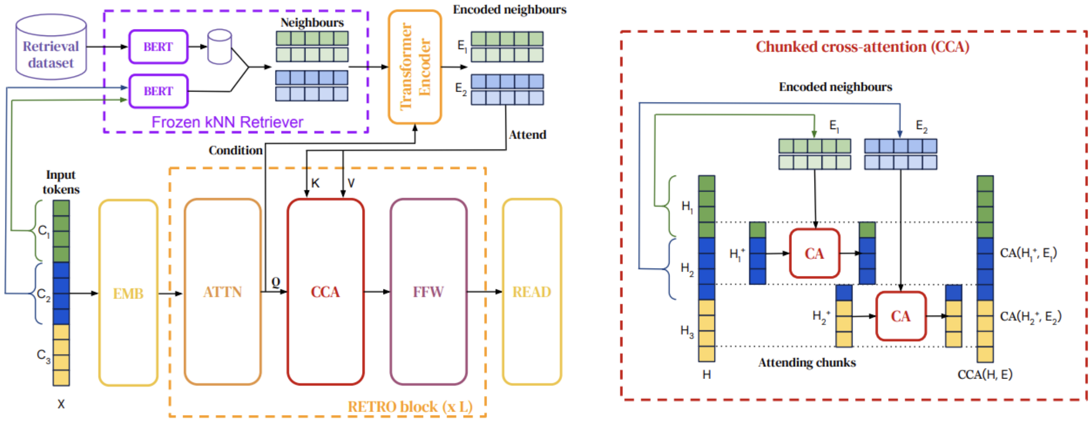
New architectures for retrieval-augmented LMs

“Generalization through Memorization: Nearest Neighbor Language Models.” Khandelwal et al., ICLR 2020.

Brief history of retrieval-augmented LMs development

RAG was initially extensively studied for certain NLP tasks, namely Question Answering

- **2017:** DrQA
- **2019:** ORQA
- **2020:** RALM, RAG
- 2020: kNN LM
- **2021:** RETRO



New architectures for retrieval-augmented LMs

“Improving language models by retrieving from trillions of tokens.” Borgeaud et al., Arxiv 2020.

Brief history of retrieval-augmented LMs development

Versatile and powerful LLMs demonstrate effectiveness even without fine-tuning

2017: DrQA

2019: ORQA

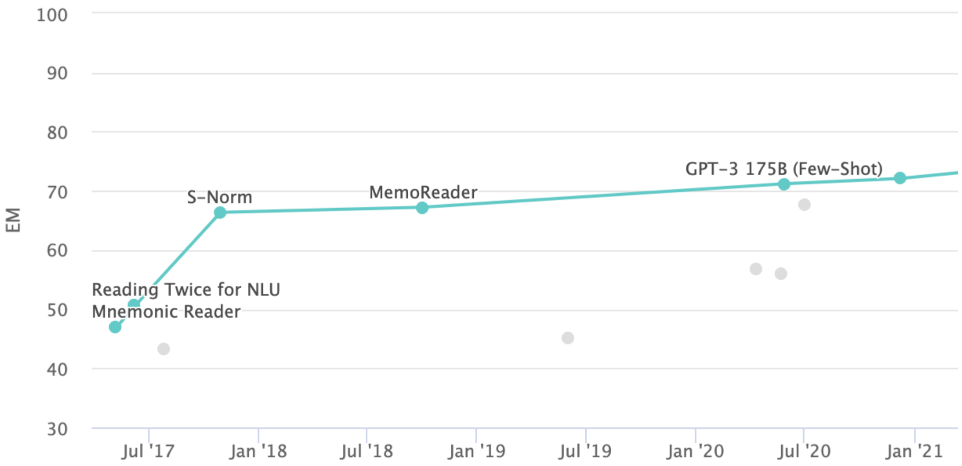
2020: RALM, RAG

2020: kNN LM

2020: **GPT3**

2021: RETRO

2022: **ChatGPT**



LLMs surpressed specialized QA models w/ retrieval

Brief history of retrieval-augmented LMs development

Success of In-Context Retrieval-Augmented LMs (commonly referred to as RAG today)

- **2017:** DrQA
- **2019:** OI
- **2020:** RA
- 2020: kN
- **2021:** RETRO
- **2023:** In-Context Retrieval-Augmented LMs

Use off-the-shelf LLMs & retrieval systems

Brief history of retrieval-augmented LMs development

RAG was initially extensively studied for certain NLP tasks, namely Question Answering

- **2017:** DrQA
- **2019:** ORQA
- **2020:** RALM, RAG
- **2020:** kNN LM
- **2021:** RETRO
- **2023:** Retrieval-augmented LLMs

Past: Developments in Architecture and Training for Specific Tasks

Brief history of retrieval-augmented LMs development

RAG was initially extensively studied for certain NLP tasks, namely Question Answering

2017: DrQA

2019: ORQA

2020: RALM, RAG

2020: kNN LM

2021: RETRO

2023: Retrieval-augmented LLMs

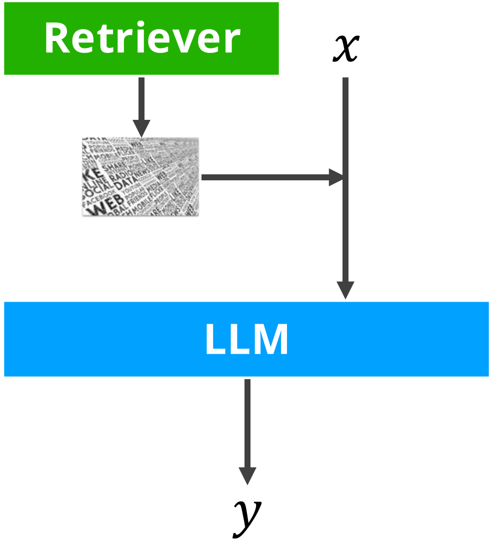
Past: Architecture / training developments for certain down or up-stream tasks

Current: Designing versatile and reliable LLM-based RAG systems for diverse use cases

Diverse architectures of retrieval-augmented LMs

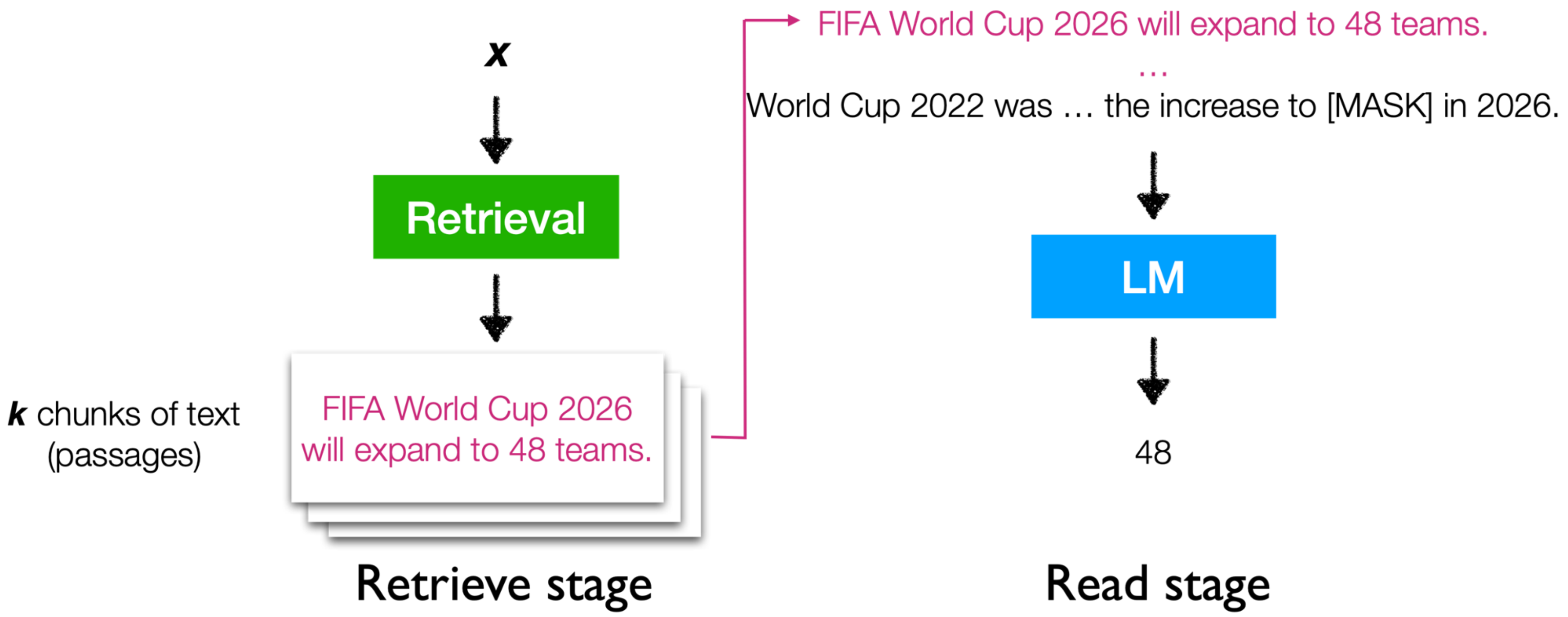
Classifying retrieval-augmented LMs based on "where" we incorporate retrieved context

- Input augmentaiton
  - Augment the input of LMs with retrieved context
  - E.g., RAG, REALM, DrQA, In-context RALM



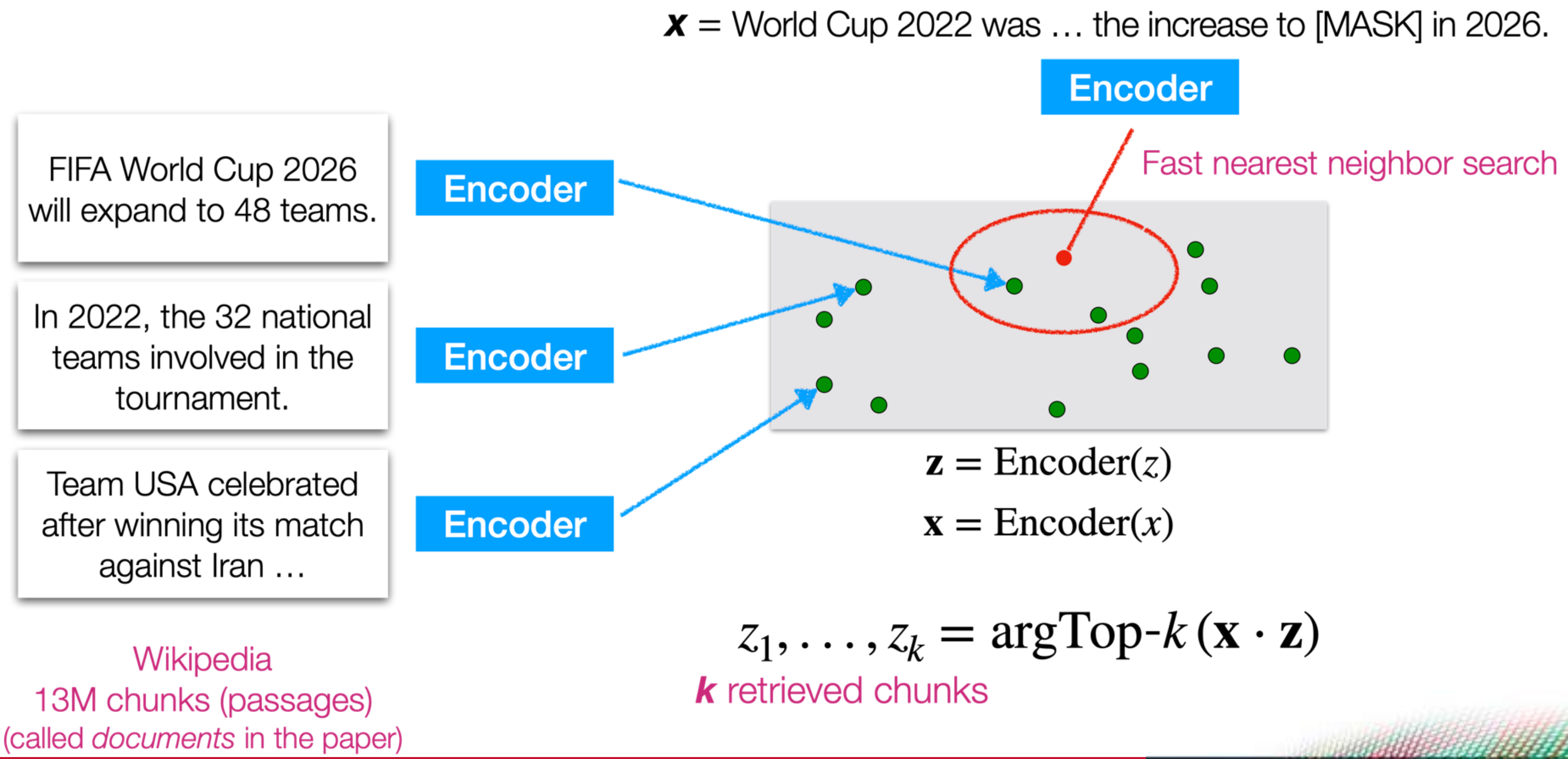
REALM: Augmenting input space of LMs

REALM is an retrieval-augmented masked LMs that predicts next tokens / spans in context



REALM: Augmenting input space of LMs

REALM finds relevant context by conducting kNN search in embedding spaces



$$\mathbf{z} = \text{Encoder}(z)$$

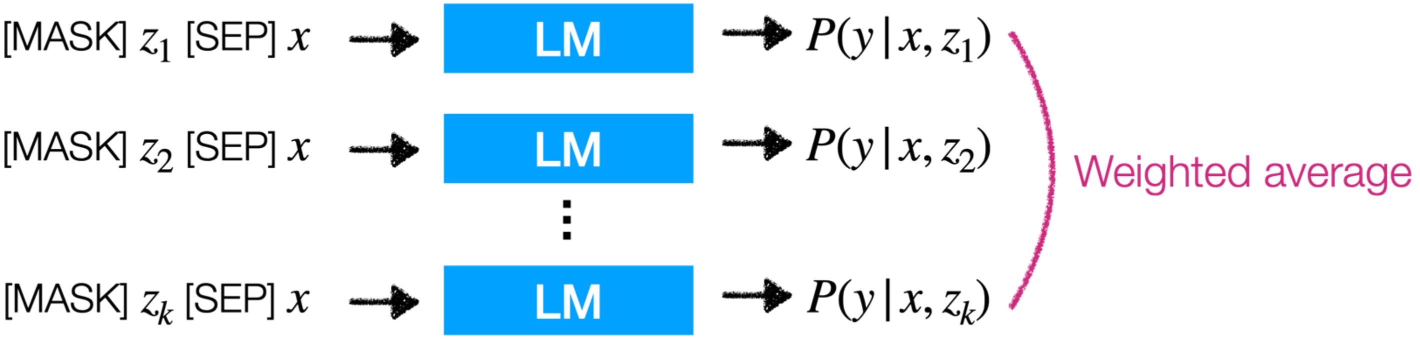
$$\mathbf{x} = \text{Encoder}(x)$$

$$z_1, \ldots, z_k = \text{argTop-}k\,(\mathbf{x} \cdot \mathbf{z})$$

k retrieved chunks

REALM: Augmenting input space of LMs

REALM compute weighted averages of final answer distributions, using retrieval similarities



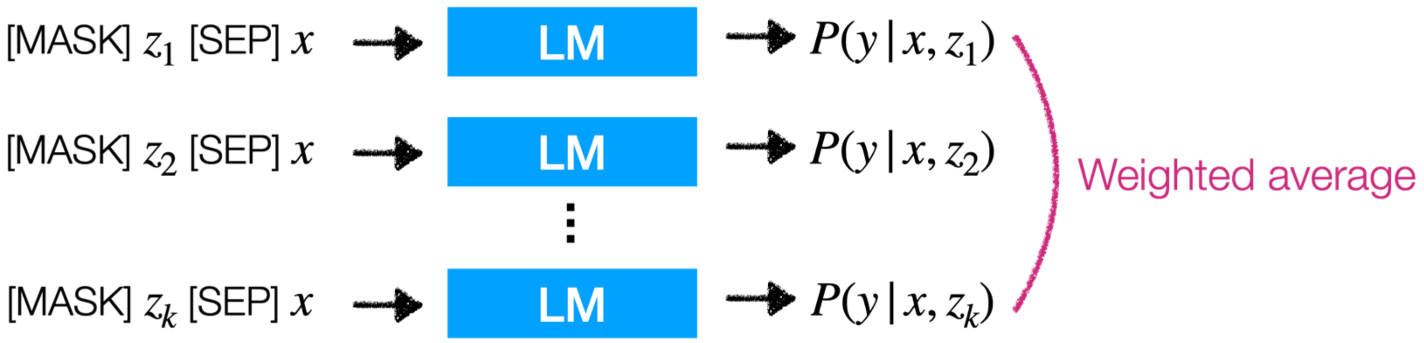
$$\sum_{z \in \mathcal{D}} \underbrace{P(z \mid x)}_{\text{from the retrieve stage}} \underbrace{P(y \mid x, z)}_{\text{from the read stage}}$$

Need to approximate
→ Consider top *k* chunks only

“REALM: Retrieval-Augmented Language Model Pre-Training.” Guu et al., ICML 2020.

REALM: Augmenting input space of LMs

REALM compute weighted averages of final answer distributions, using retrieval similarities



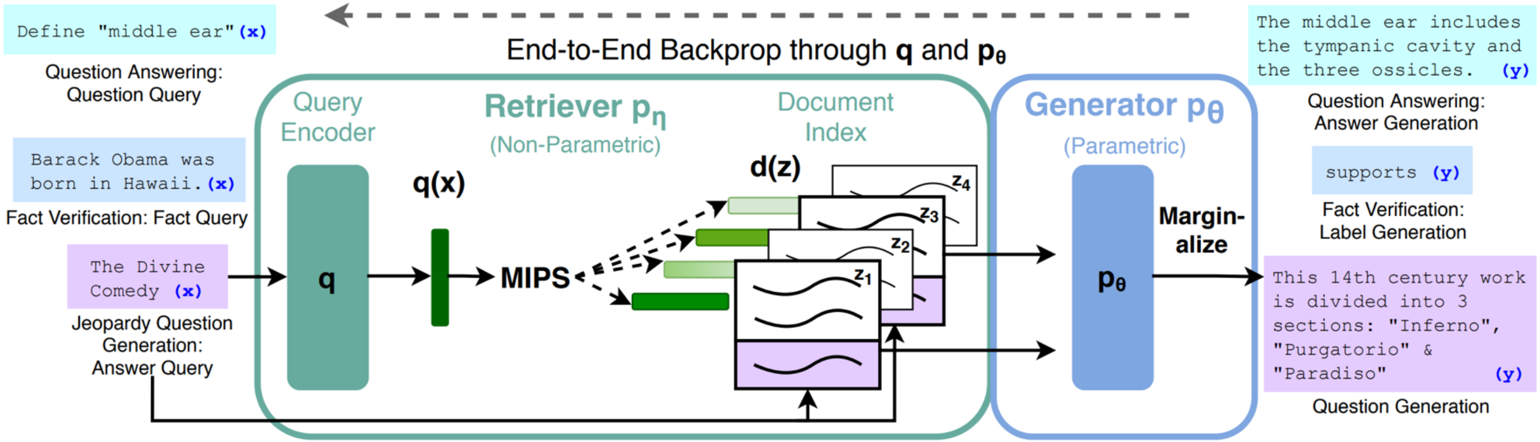
$$p(y \mid z, x) \propto \sum_{s \in S(z,y)} \exp\left(\mathtt{MLP}\left(\left[h_{\mathtt{START(s)}}; h_{\mathtt{END(s)}}\right]\right)\right)$$

$$h_{\mathtt{START(s)}} = \mathtt{BERT}_{\mathtt{START(s)}}(\mathtt{join}_{\mathtt{BERT}}(x, z_{\text{body}})),$$
$$h_{\mathtt{END(s)}} = \mathtt{BERT}_{\mathtt{END(s)}}(\mathtt{join}_{\mathtt{BERT}}(x, z_{\text{body}})),$$

"REALM: Retrieval-Augmented Language Model Pre-Training." Guu et al., ICML 2020.

RAG: Augmenting input space of LMs

RAG combines a trained retriever & autoregressive BART, starting from pre-trained weights

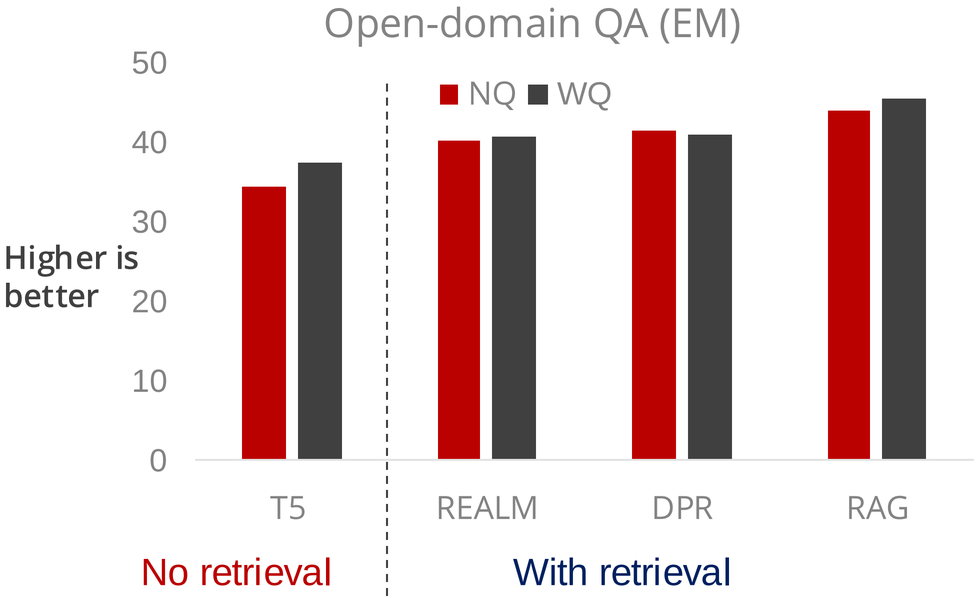


$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z, y_{1:i-1})$$

RAG & REALM: Results

RAG and REALM show their effectiveness on open-domain QA and other tasks

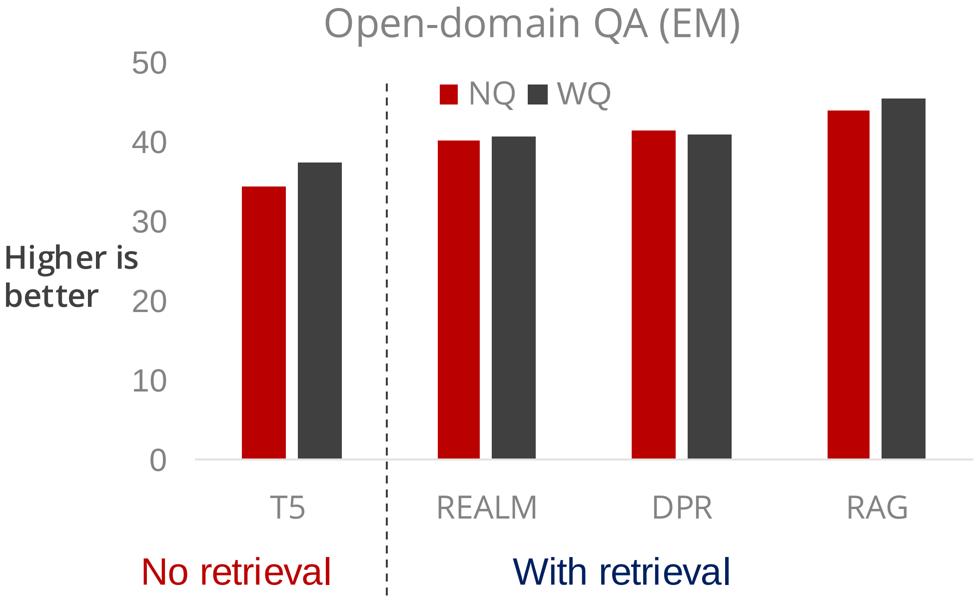
- RAG outperforms REALM and other baselines on Open-domain QA such as NaturalQuestion

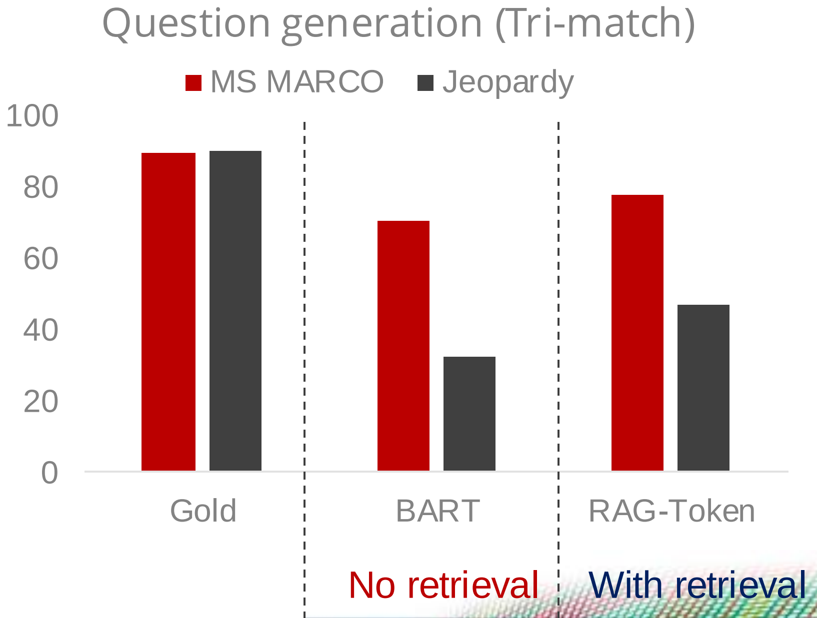


RAG & REALM: Results

RAG and REALM show their effectiveness on open-domain QA and other tasks

- RAG outperforms REALM and other baselines on Open-domain QA such as NaturalQuestions
- RAG also show their effectiveness on generation tasks





"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Lewis et al., NeurIPS 2020.

Recent follow-up: In-context Retrieval-augmented LMs

Similar principles as in DrQA, REALM, RAG, but completely removes retrieval

- Combining retrieval and off-the-shelf LMs e.g., GPT-4 at inference time <u>without training</u>
- Often referred to as "RAG" nowadays
- We'll cover this in depth in the next section!

Pros and cons of input augmentation

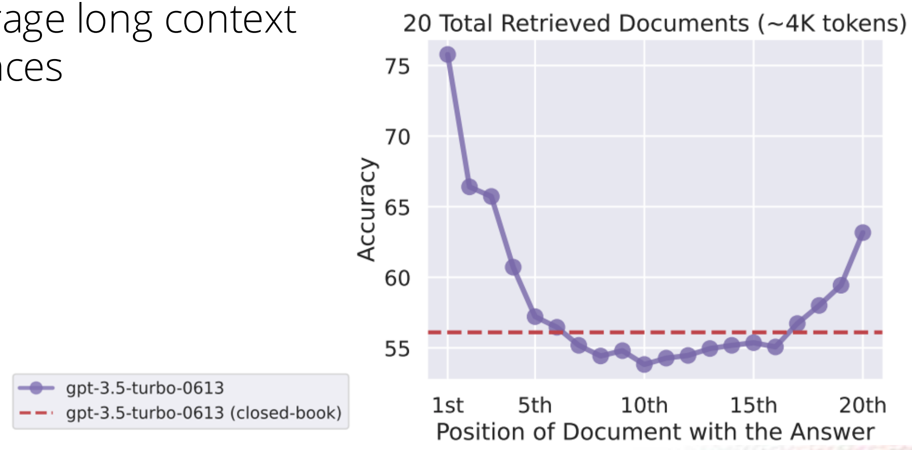
Input augmentation is powerful but has several limitations

- Pros
  - Easy to switch to new, more powerful LMs with fine-tuning / without training
  - LLMs can effectively levarage input context

Pros and cons of input augmentation

Input augmentation is powerful but has several limitations.

- Pros
  - Easy to switch to new, more powerful LMs with fine-tuning / without training
  - LLMs can effectively levarage input context
- Cons
  - Expensive to scale up to hundreads or thousands of documents
    - LLMs also often do not fully levarage long context
  - No strict attributions to specific evidences



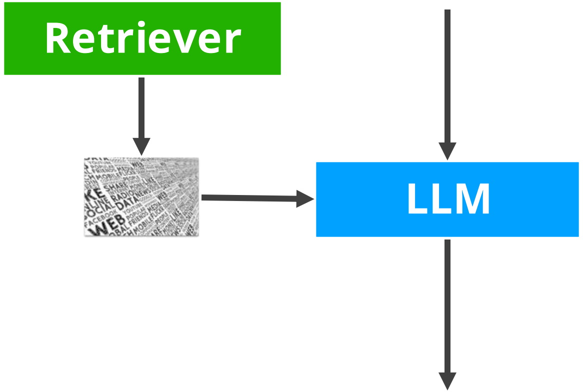
“Lost in the Middle: How Language Models Use Long Contexts.” Liu et al., TACL 2023.

Diverse architectures of retrieval-augmented LMs

Classifying retrieval-augmented LMs based on "where" we incorporate retrieved context

- Input augmentaiton
  - Augment the input of LMs with retrieved context
  - E.g., RAG, REALM, DrQA, In-context RALM

- Intermediate incorporation
  - Incorporate retrieved context in intermediate spaces of transformers
  - E.g., RETRO, Instruct RETRO



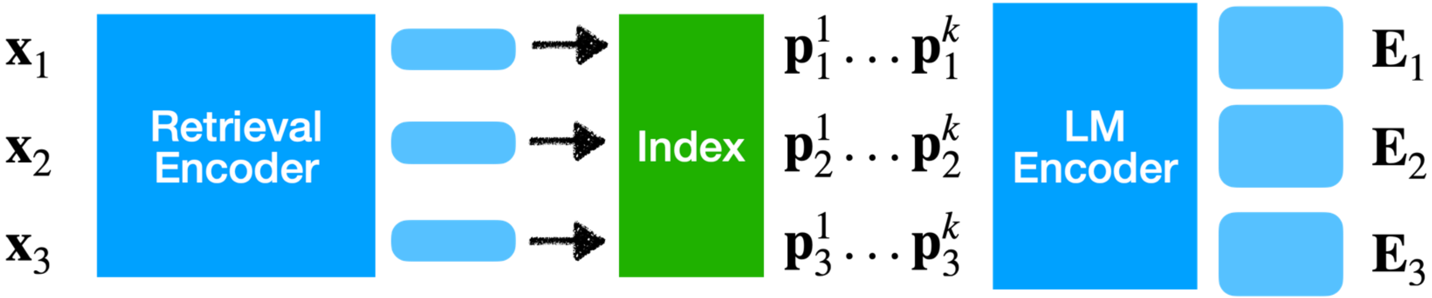
RETRO: Incorporating context in intermediate layers

RETRO enables more efficient incorporations of many documents

$\boldsymbol{x}$ = 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)



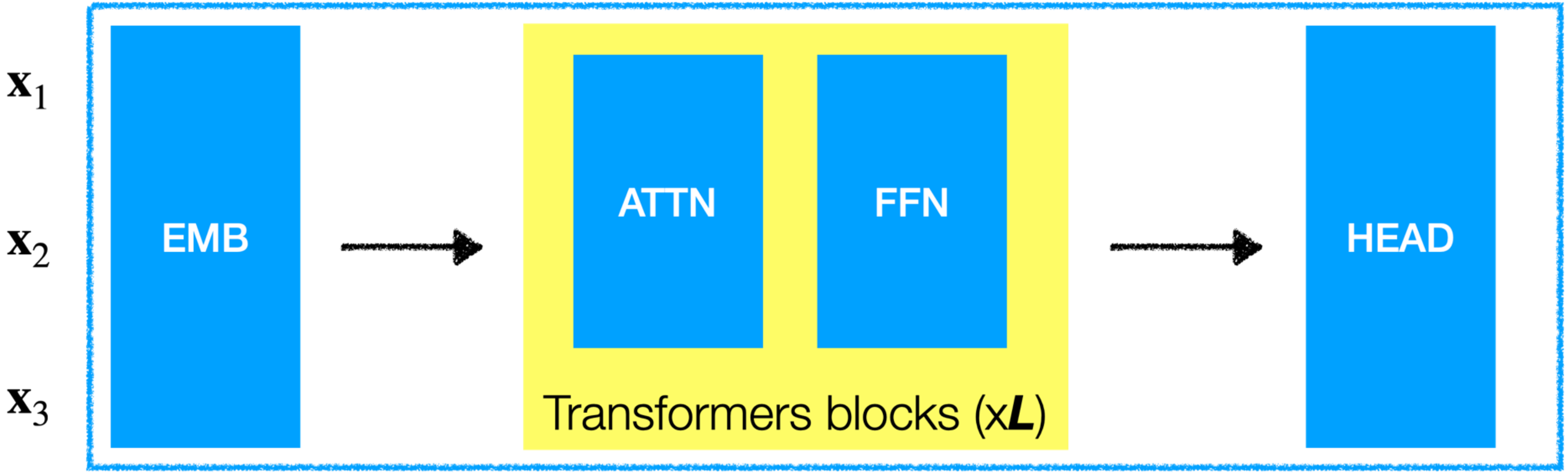
(r = # tokens per text chunk)
(d = hidden dimension)
(k = # retrieved chunks per split)

(A $r \times k \times d$ matrix)

“Improving language models by retrieving from trillions of tokens.” Borgeaud et al., Arxiv 2020.

RETRO: Incorporating context in intermediate layers

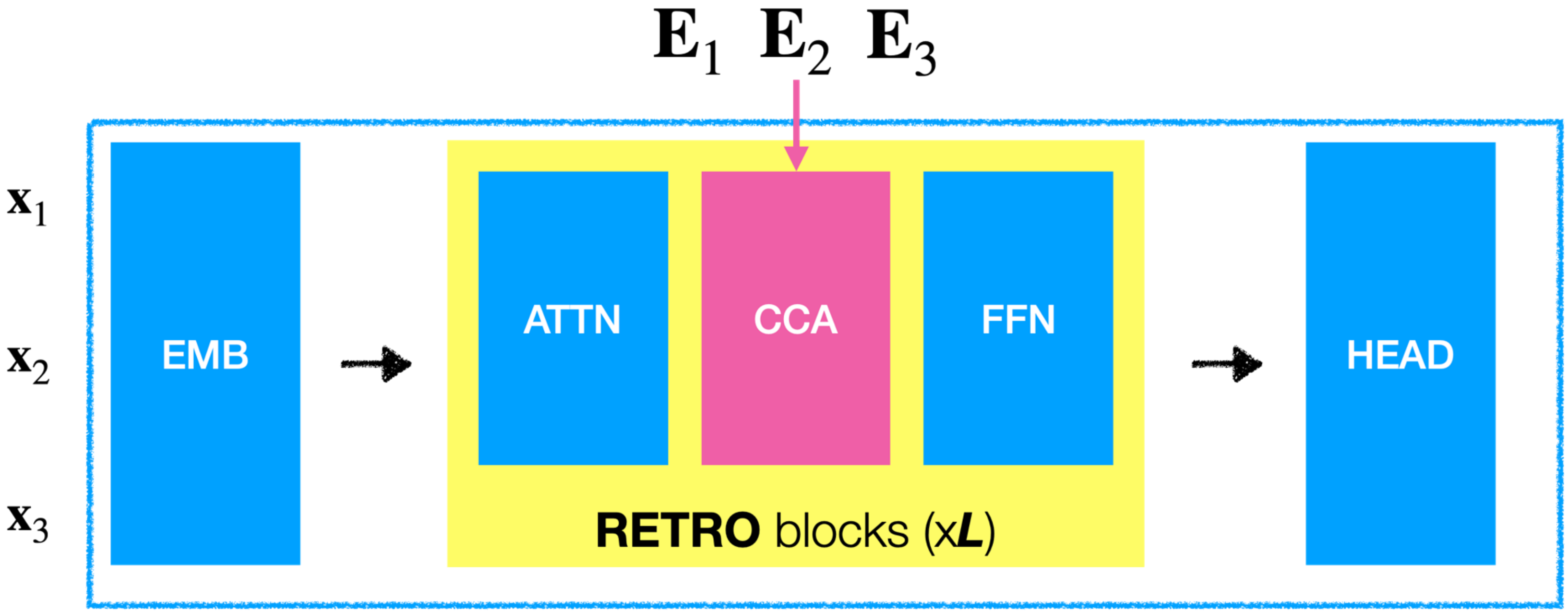
RETRO enables more efficient incorporations of many documents



Standard transformer block

RETRO: Incorporating context in intermediate layers

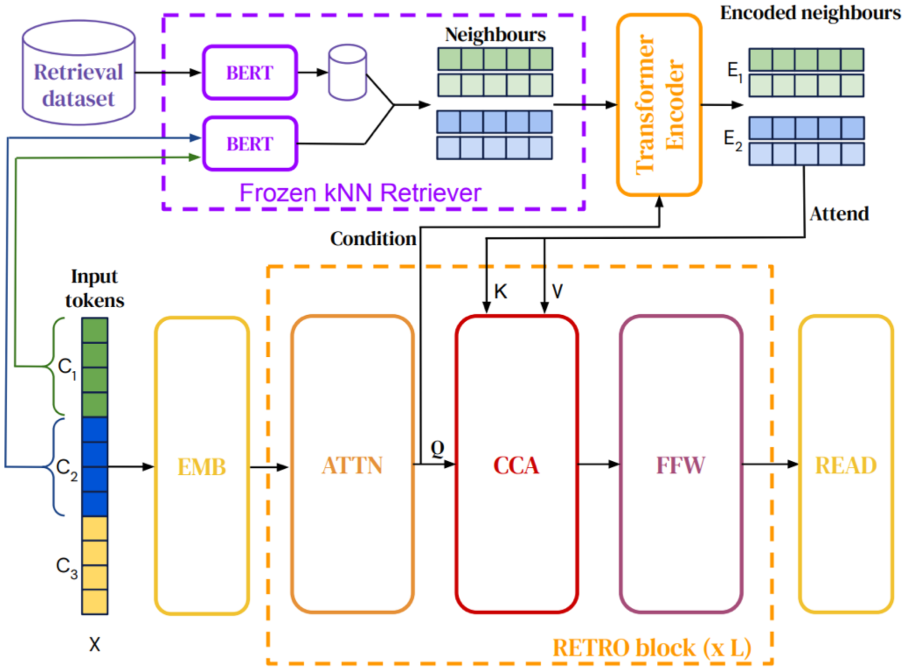
RETRO enables more efficient incorporations of many documents

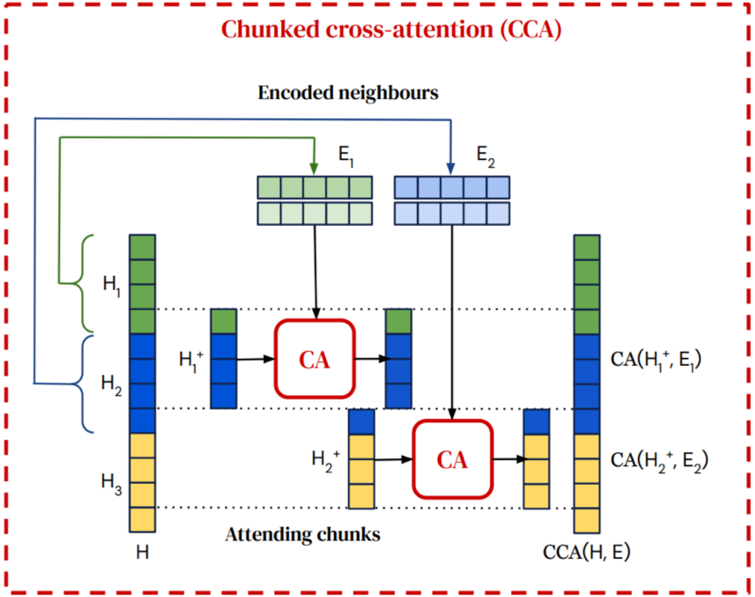


Chunked Cross Attention (CCA)

RETRO: Incorporating context in intermediate layers

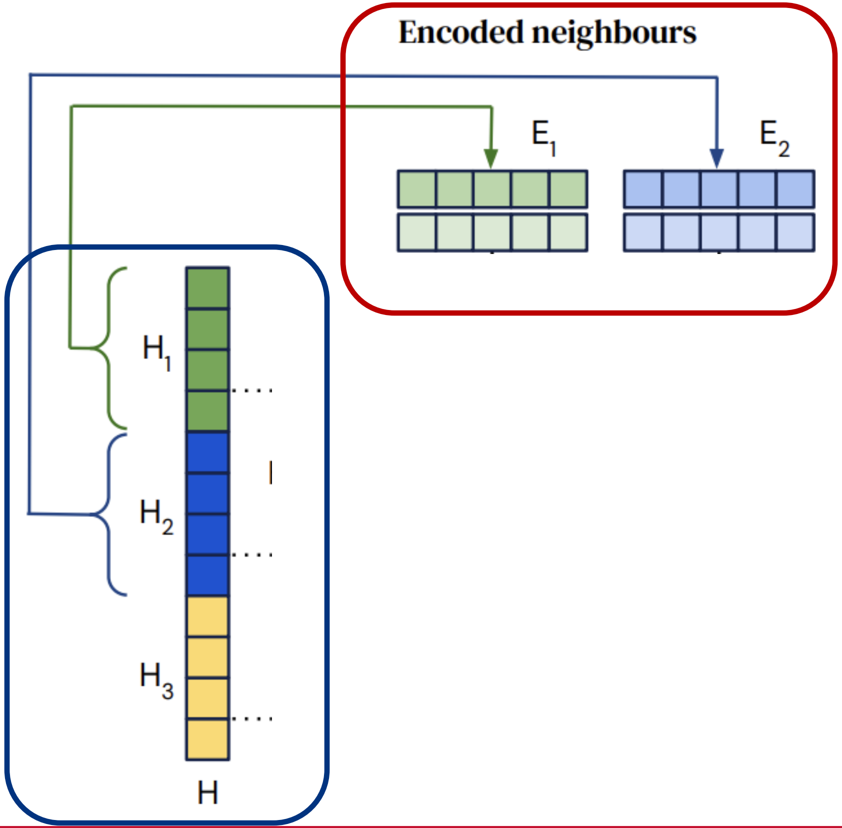
RETRO uses frozen BERT as a retriever, and retrieve nearest neighbors from 1.7T datastore

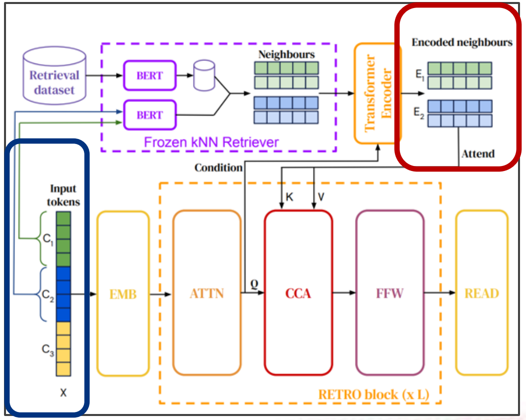




RETRO: Incorporating context in intermediate layers

Given the input sequence, it first retrieves a set of relevant documents (embedding of text)

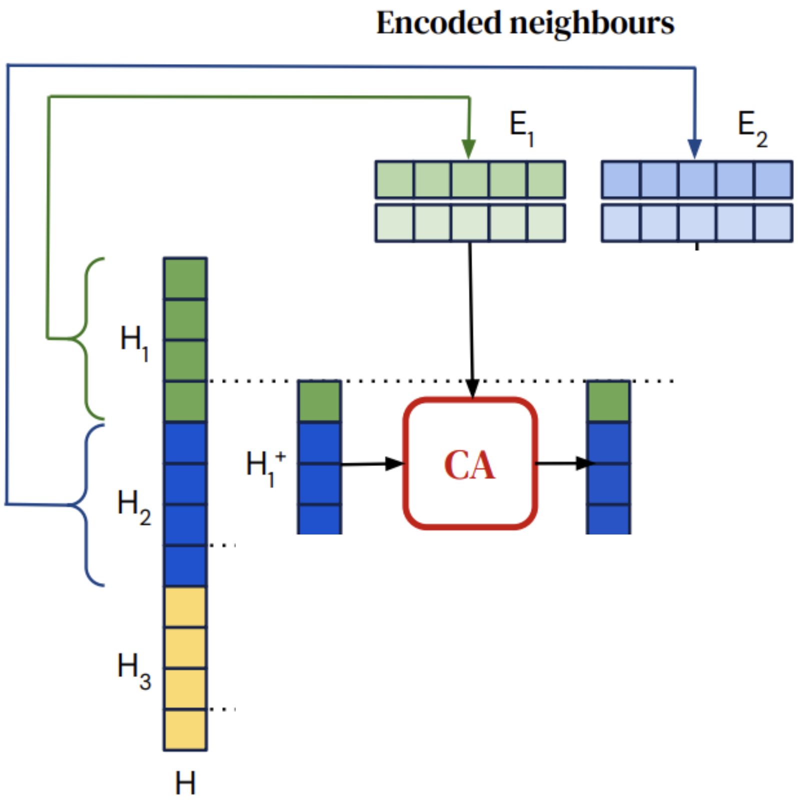




"Improving language models by retrieving from trillions of tokens." Borgeaud et al., Arxiv 2020.

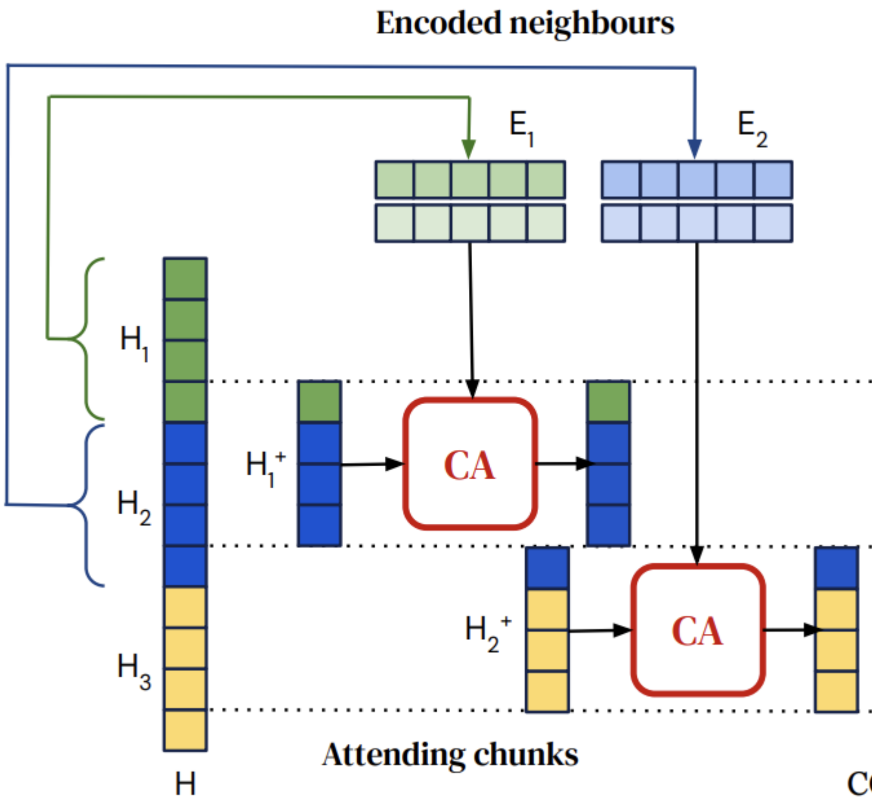
RETRO: Incorporating context in intermediate layers

Use cross-attention to generate retrieved context-aware representations



RETRO: Incorporating context in intermediate layers

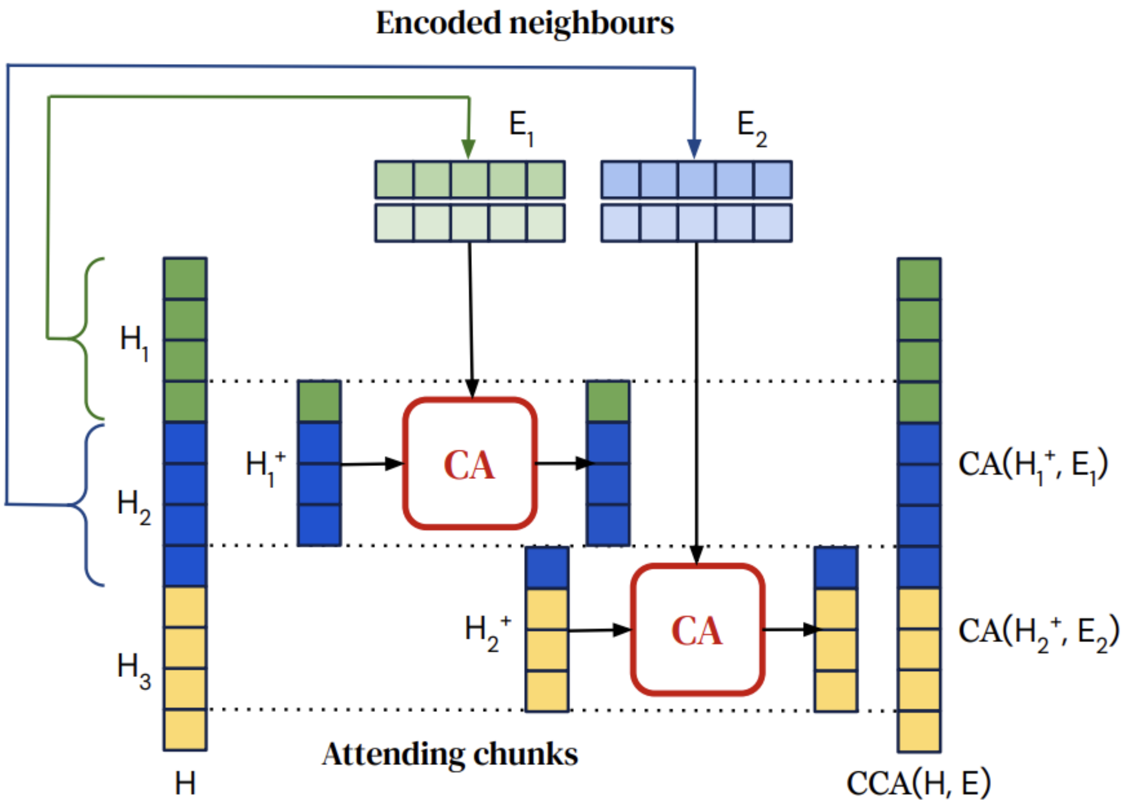
Use cross-attention to generate retrieved context-aware representations



“Improving language models by retrieving from trillions of tokens.” Borgeaud et al., Arxiv 2020.

RETRO: Incorporating context in intermediate layers

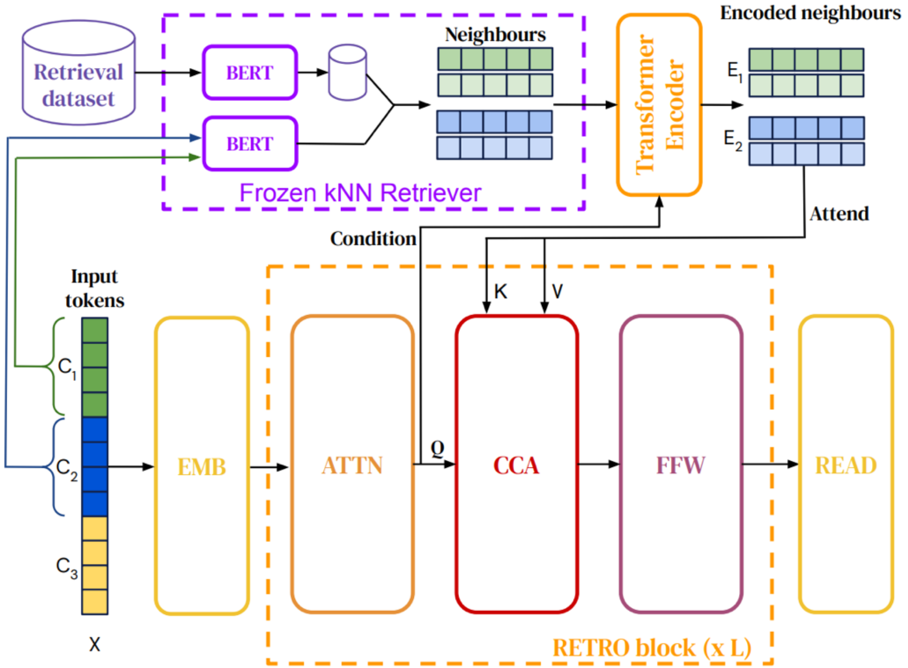
Concatnate all of the CA output (the size of input H and output CCA(H,E) remains the same

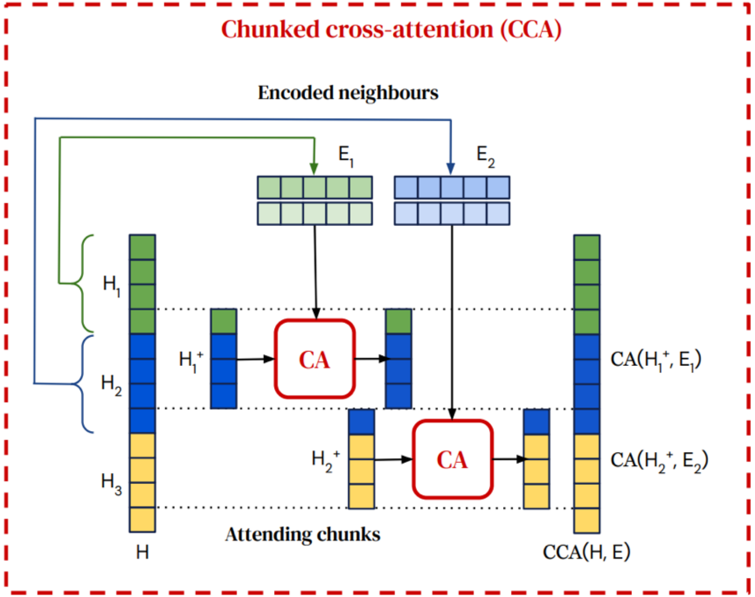


“Improving language models by retrieving from trillions of tokens.” Borgeaud et al., Arxiv 2020.

RETRO: Incorporating context in intermediate layers

RETRO uses frozen BERT as a retriever, and retrieve nearest neighbors from 1.7T datastore



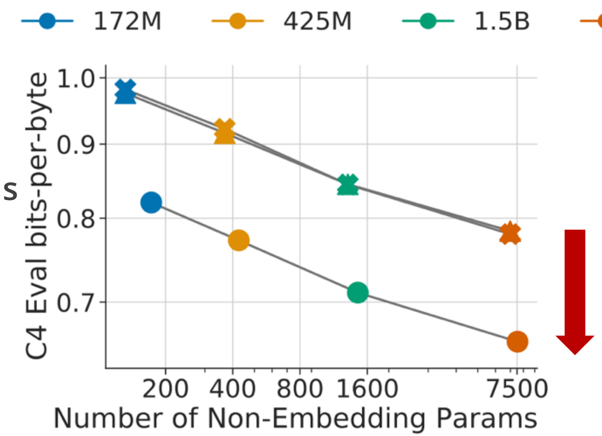


RETRO: Results

RETRO shows impressive performance improvements on upstream (language modeling) tasks

- RETRO significantly outperforms non-retrieved baselines

Lower is better

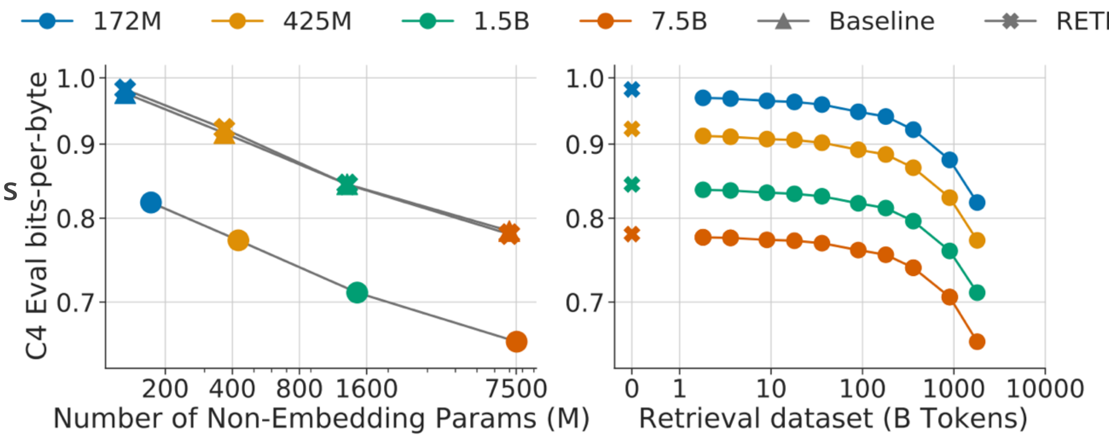


RETRO: Results

RETRO shows impressive performance improvements on upstream (language modeling) tasks

- RETRO significantly outperforms non-retrieved baselines
- RETRO performance continues to improve as the datastore scales from a few billion to 1.7 trillion data points

Lower is better

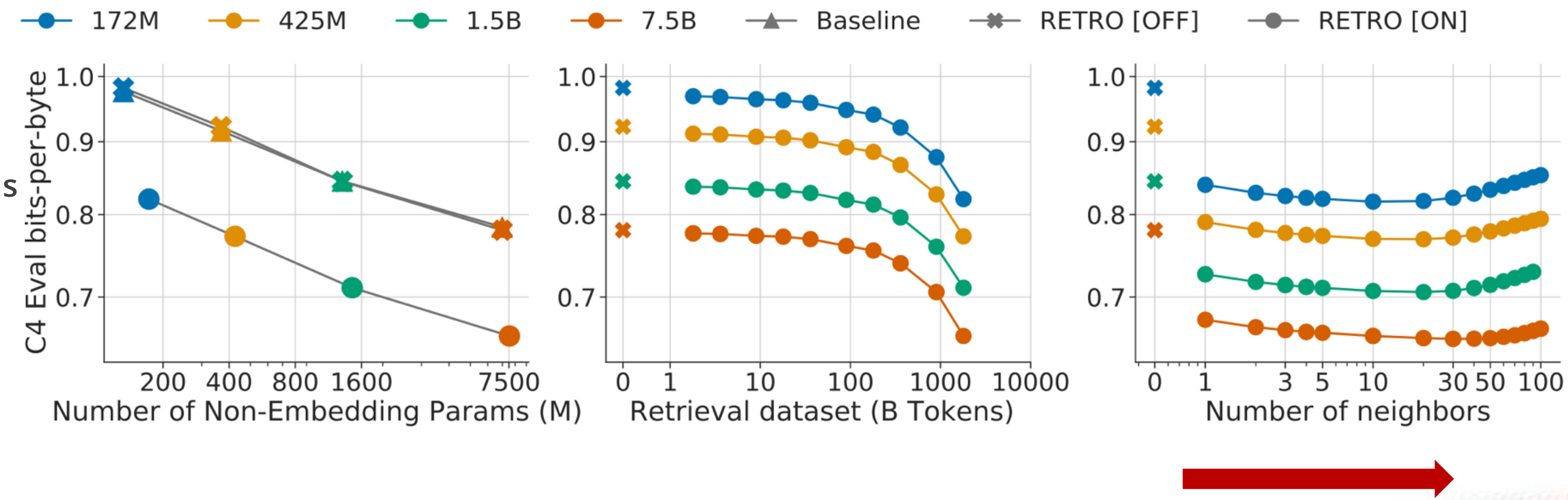


RETRO: Results

RETRO shows impressive performance improvements on upstream (language modeling) tasks

- RETRO significantly outperforms non-retrieved baselines
- RETRO performance continues to improve as the datastore scales from a few billion to 1.7 trillion data points
- Increasing # of **docs up to 40 helps**

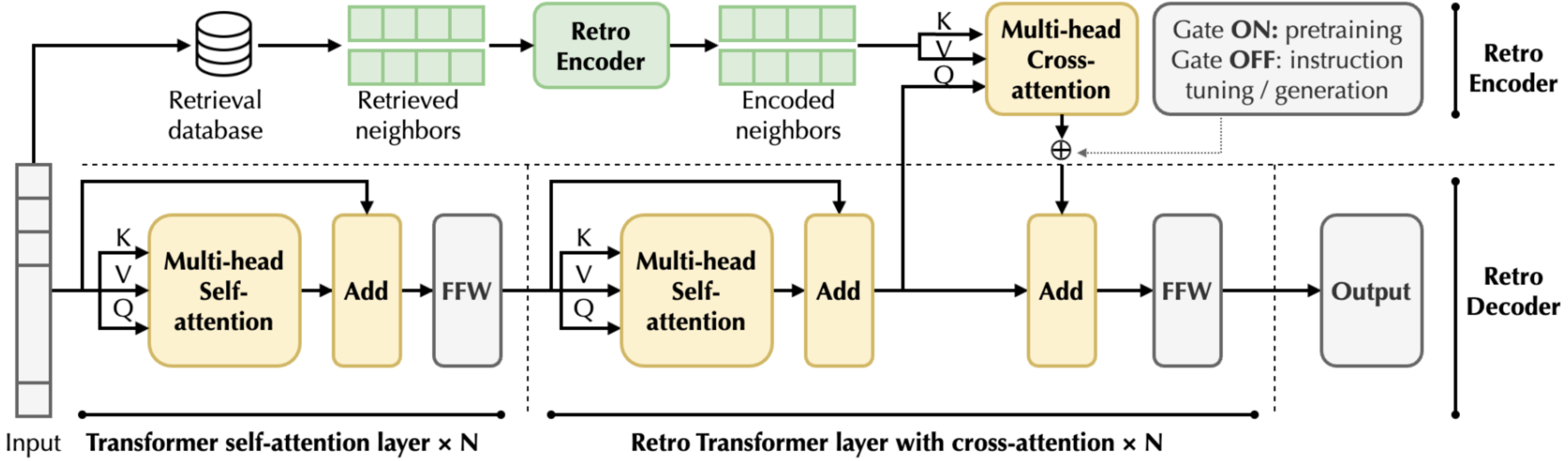
Lower is better



“Improving language models by retrieving from trillions of tokens.” Borgeaud et al., Arxiv 2020.

Recent follow-up: Instruct RETRO

Develop RETRO-block on top of Llama (autoregressive LMs), pre-training & multi-task training



“InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining.” Wang et al., ICML 2024.

Pros and cons of intermediate incorporation

Alternative way to incorporate retrieved context in a more scalable way, but requires training

- Pros
  - More efficiently incorporates many passages than input augmentation
  - Possibly more effective than retrieval augmentaion (i.e., Instruct RETRO results)

- Cons
  - Require modification of underlying LMs
  - Expensive pre-training is necessary
  - Doesn’t provide strict attribution

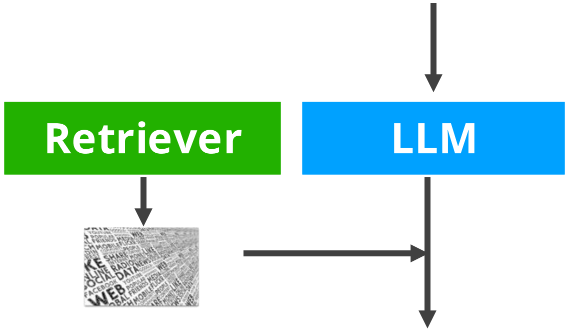
Diverse architectures of retrieval-augmented LMs

Classifying retrieval-augmented LMs based on "where" we incorporate retrieved context

- Input augmentaiton
  - Augment the input of LMs with retrieved context
  - E.g., RAG, REALM, DrQA, In-context RALM

- Intermediate incorporation
  - Incorporate retrieved context in intermediate spaces of transformers
  - E.g., RETRO, Instruct RETRO

- Output interpolation
  - Interpolate output token probabilities with retrieved non-parametric distributions
  - E.g., kNN LM

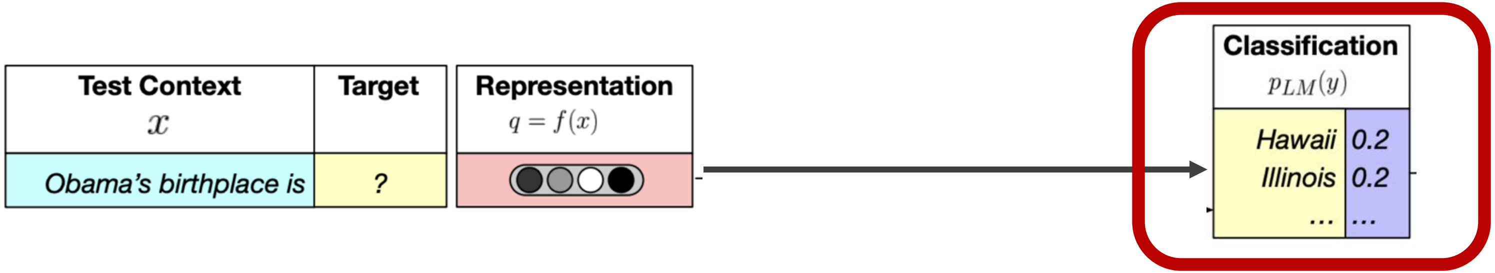


kNN LM: directly interpolate output token distributions

Directly interpolate output token distributions of LMs

- Given a context x, a model predicts parametric distributions for next token

Parametric distribution (LM output distribution)

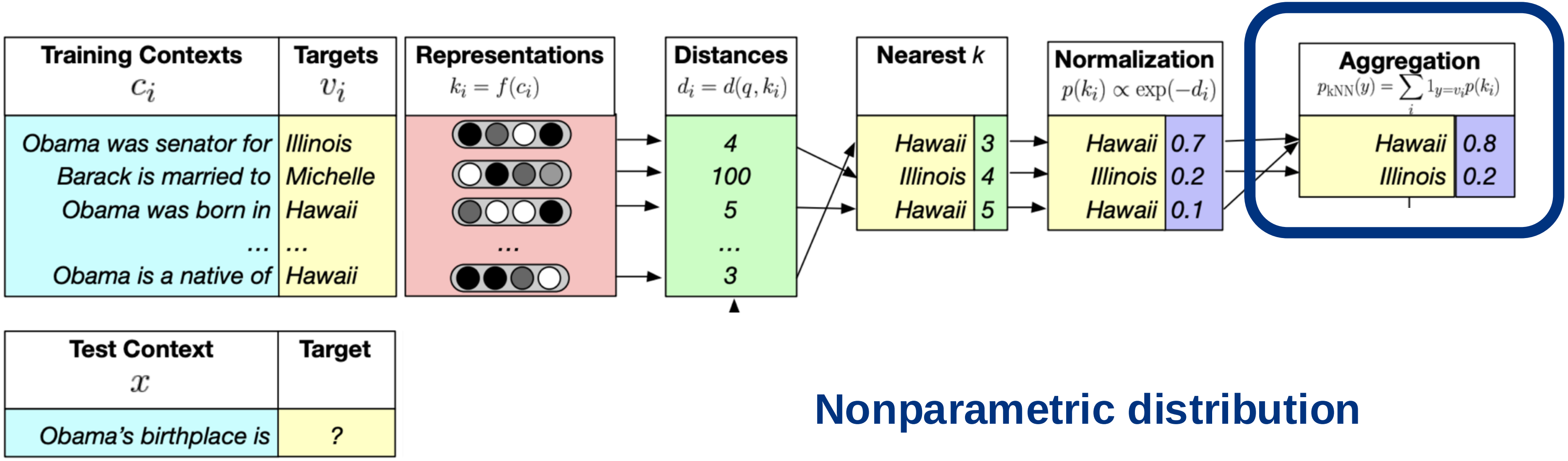


"Generalization through Memorization: Nearest Neighbor Language Models." Khandelwal et al., ICLR 2020.

kNN LM: directly interpolate output token distributions

Directly interpolate output token distributions of LMs

- Given a context x, a model predicts parametric distributions for next token
- kNN LM computes nonparametric distributions, by finding similar training context c_i

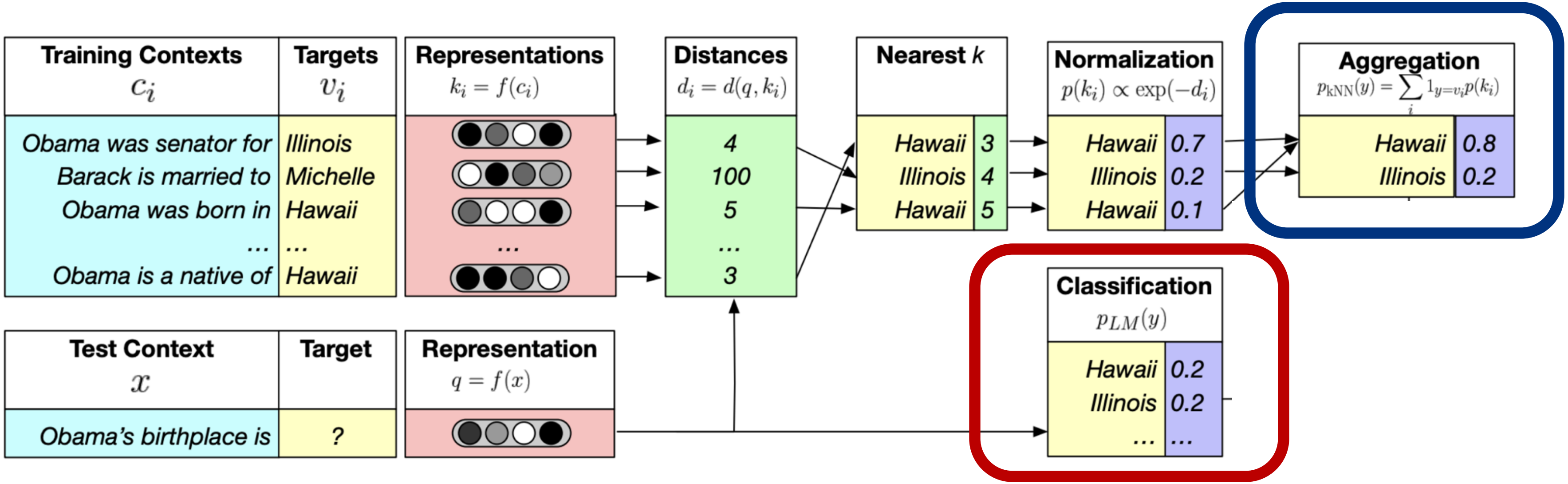


Nonparametric distribution

kNN LM: directly interpolate output token distributions

Directly interpolate output token distributions of LMs

- Given a context x, a model predicts parametric distributions for next token
- kNN LM computes nonparametric distributions, by finding similar training context c_i



kNN LM: directly interpolate output token distributions

Directly interpolate output token distributions of LMs

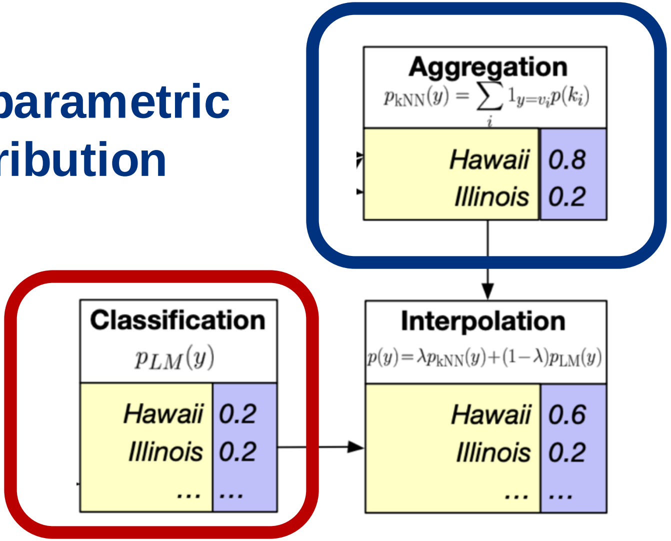
- Interpolates two token distributions, adjusting the balance using a hyperparamter λ

λ: hyperparameter

$$P_{k\mathrm{NN-LM}}(y \mid x) = (1 - \lambda) P_{\mathrm{LM}}(y \mid x) + \lambda P_{k\mathrm{NN}}(y \mid x)$$

Nonparametric distribution

Parametric distribution

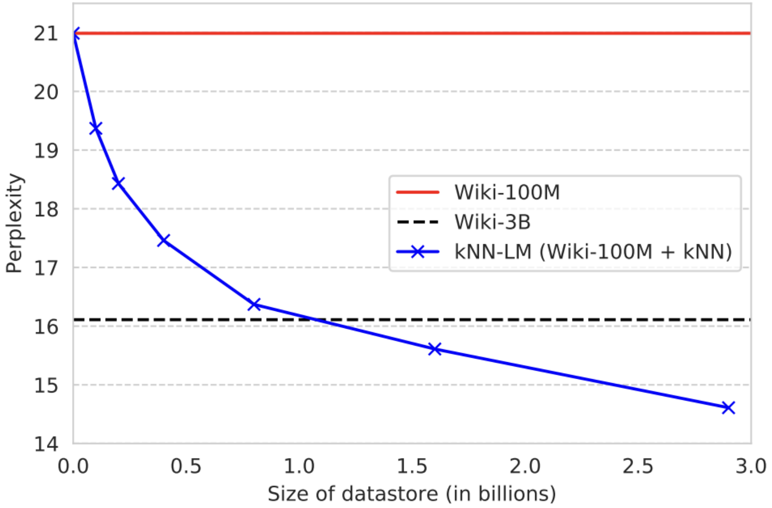


"Generalization through Memorization: Nearest Neighbor Language Models." Khandelwal et al., ICLR 2020.

kNN LM: Results

kNN LM outperforms much larger parametric LMs by large margin

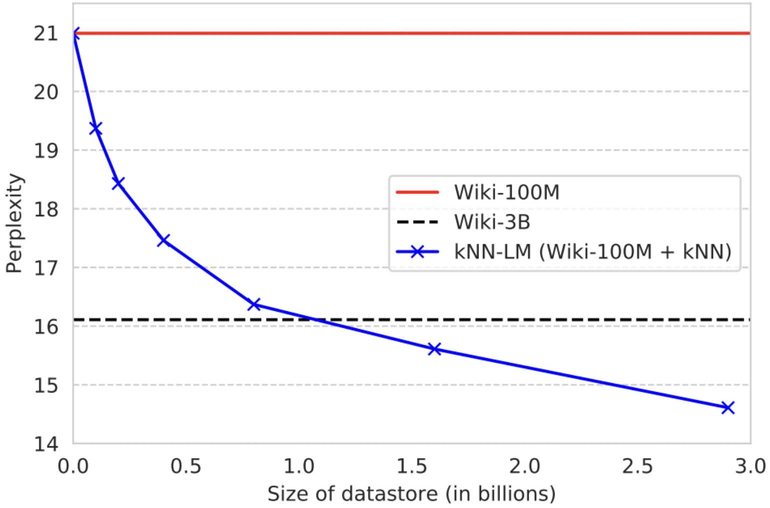
- kNN LM constantly outperforms parametric 100M LMs & 30x larger 3B LMs with larger datastore

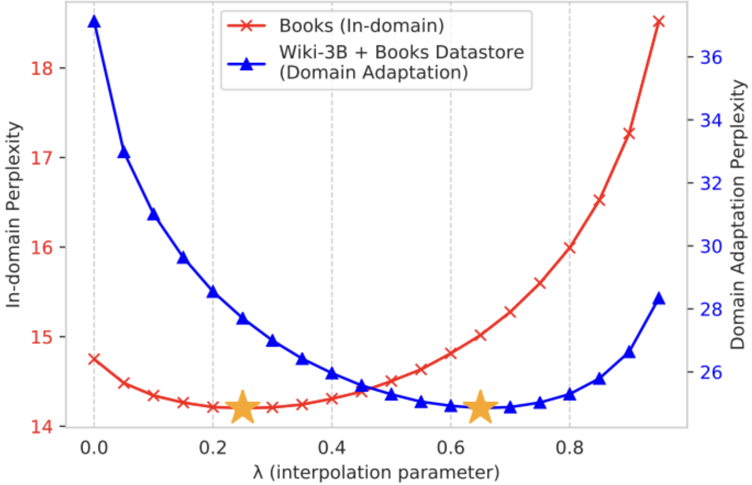


kNN LM: Results

kNN LM outperforms much larger parametric LMs by large margin

- kNN LM constantly outperforms parametric LMs and 30x larger 3B LMs with larger datastore
- kNN LM also enables efficient & controlled domain adaptations



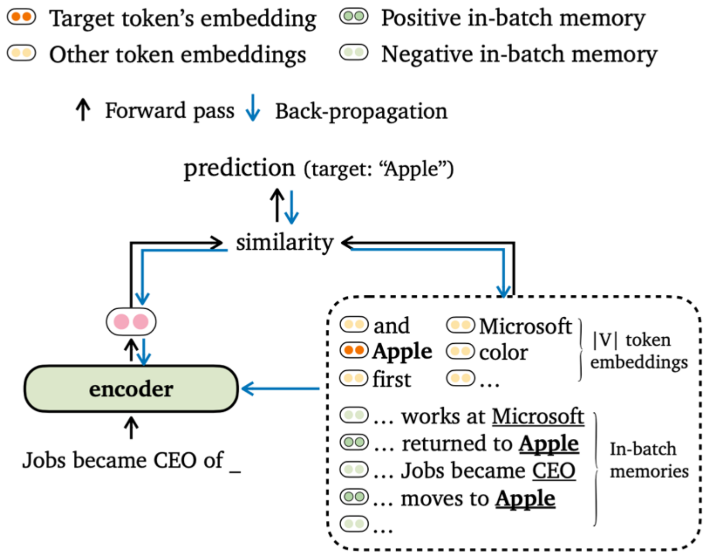


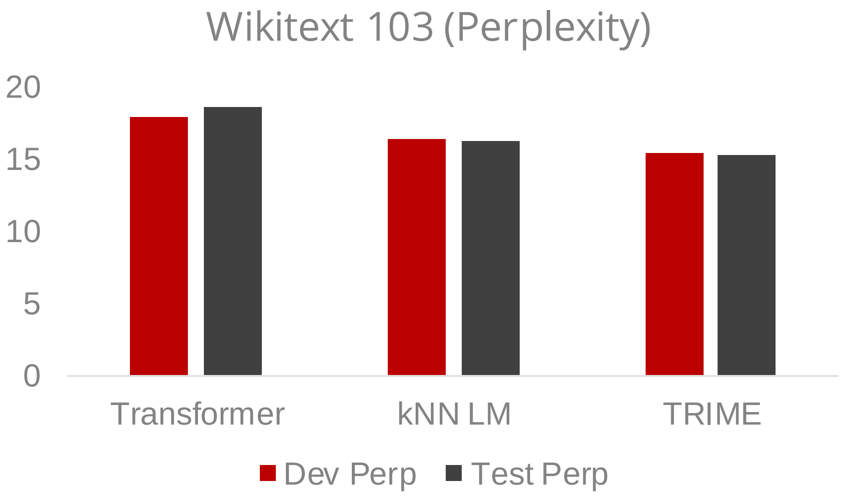
“Generalization through Memorization: Nearest Neighbor Language Models.” Khandelwal et al., ICLR 2020.

Recent follow-up: TRIME

Training kNN LM to better learn interpolations

- kNN LM uses pre-trained LMs without any training
- TRIME introduces an efficient training method, outperforming kNN LM





Pros and cons of output interpolation

kNN LM & variatns have unique advantages but have several empirical challenges

- Pros
  - Provides token-level attributions
  - Enables explicit control between parametric and non-parametric memories

Pros and cons of output interpolation

kNN LM & variatns have unique advantages but have several empirical challenges

- Pros
  - Provides token-level attributions
  - Enables explicit control between parametric and non-parametric memories
- Cons
  - Difficult to scale to large retrieval corpora (i.e., the number of embeddings equals the number of tokens)
  - Empirically shows limited effectiveness outside of upsteam language modeling tasks

Great Memory, Shallow Reasoning: Limits of kNN-LMs

Shangyi Geng **Wenting Zhao** **Alexander M Rush**
Cornell University
{sg2323, wz346, arush}@cornell.edu

kNN-LM Does Not Improve Open-ended Text Generation

Shufan Wang[1] **Yixiao Song**[1] **Andrew Drozdov**[1]
Aparna Garimella[2] **Varun Manjunatha**[2] **Mohit Iyyer**[1]
University of Massachusetts Amherst[1] Adobe Research[2]
{shufanwang, yixiaosong, adrozdov, miyyer}@umass.edu
{garimell,vmanjuna}@adobe.com

Summary

Diverse types of retrieval-augmented LMs have been studied; have pros & cons

- Input augmentation: widely used and effective but faces challenges when incorporating more passages
- Intermediate incorporation: can efficiently handle more passages but requires pre-training and fine-tuning
- Output interpolation: provides direct control over LM output, but has limited success in downstream tasks and faces challenges of scaling the datastore

	Representative methods	Retrieval unit	Retrieval frequency
Input augmentation	DrQA, RAG, REALM, ICRALM	Passage	Once at the beginning
Intermediate incorporation	RETRO, InstructRETRO	Passage	Every k tokens
Output interpolation	kNNLM TRIME	Token	Every token

Present: *Retrieval-augmented Generation with LLMs*

In-context retrieval-augmented LMs

Simply augmenting input of LMs gives signficant gain across different tasks

Who is the current prime minister of United Kingdom? → **LLM**

```
Answer the following question, based on the reference.
Reference
Q: Who is the current PM of UK?
A:
```

In-context retrieval-augmented LMs

Simply augmenting input of LMs gives signficant gain across different tasks

Who is the current prime minister of United Kingdom?

LLM

```
Answer the following question, based on the reference.
Reference
Q: Who is the current PM of UK?
A:
```

Rishi Sunak

“When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories”. Mallen*, Asai* et al. ACL 2023

In-context retrieval-augmented LMs

Simply augmenting input of LMs gives signficant gain across different tasks

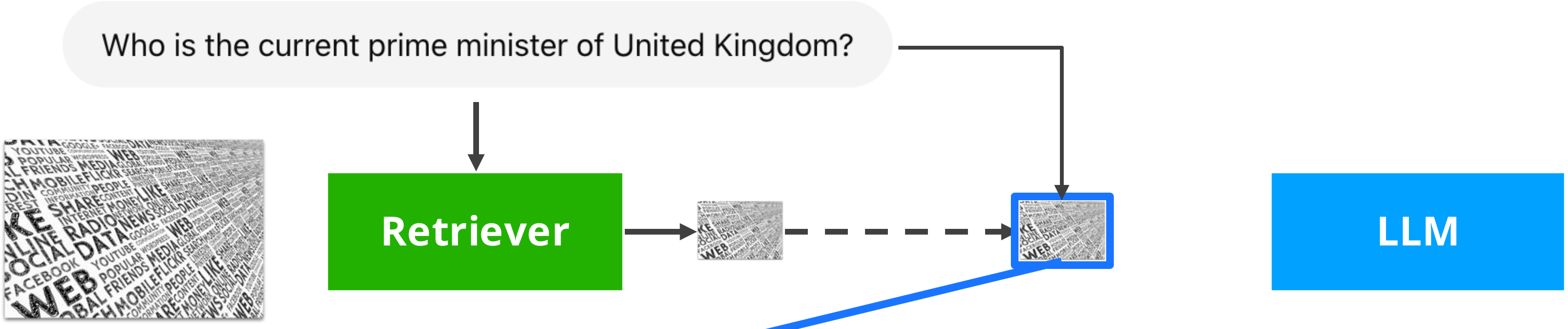
Who is the current prime minister of United Kingdom?

Retriever

LLM

In-context retrieval-augmented LMs

Simply augmenting input of LMs gives signficant gain across different tasks

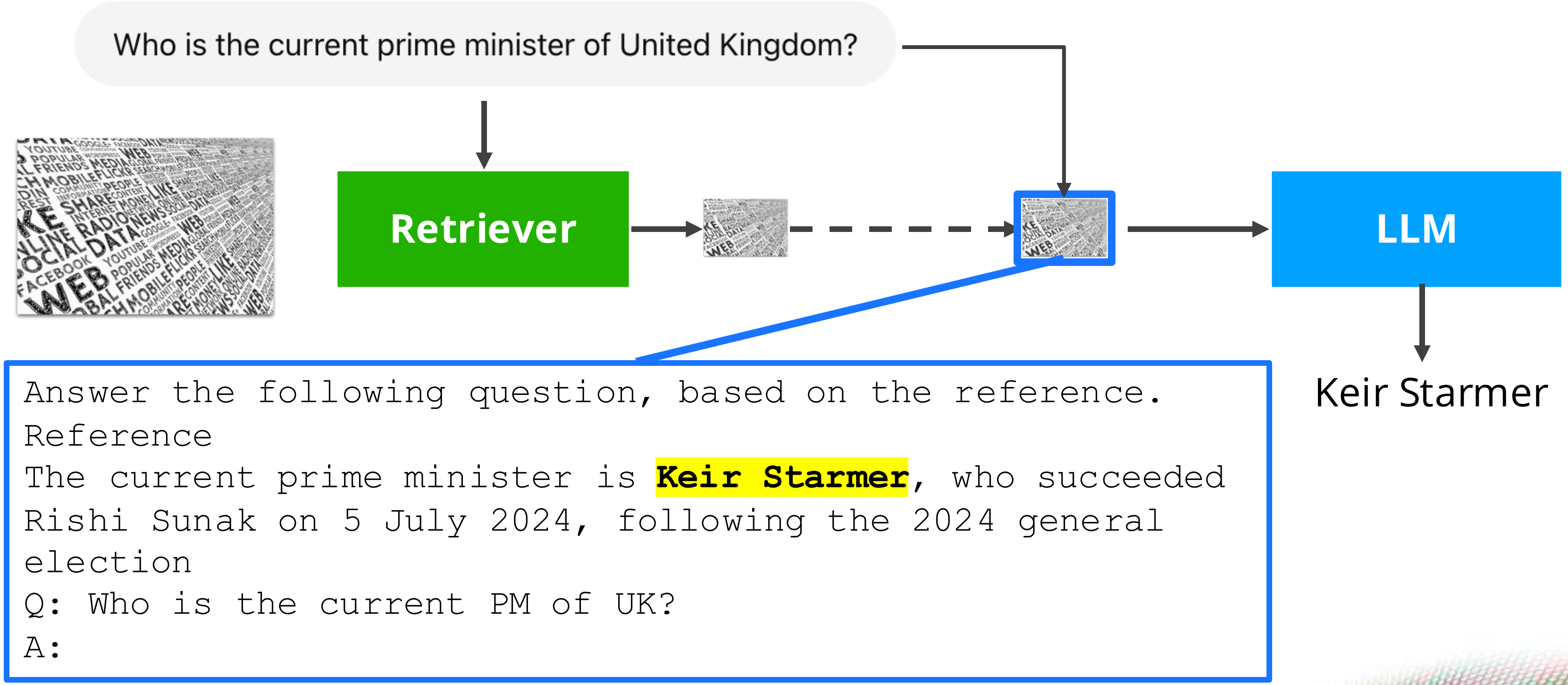



```
Answer the following question, based on the reference.
Reference
The current prime minister is Keir Starmer, who succeeded
Rishi Sunak on 5 July 2024, following the 2024 general
election
Q: Who is the current PM of UK?
A:
```

"When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories". Mallen*, Asai* et al. ACL 2023

In-context retrieval-augmented LMs

Simply augmenting input of LMs gives signficant gain across different tasks

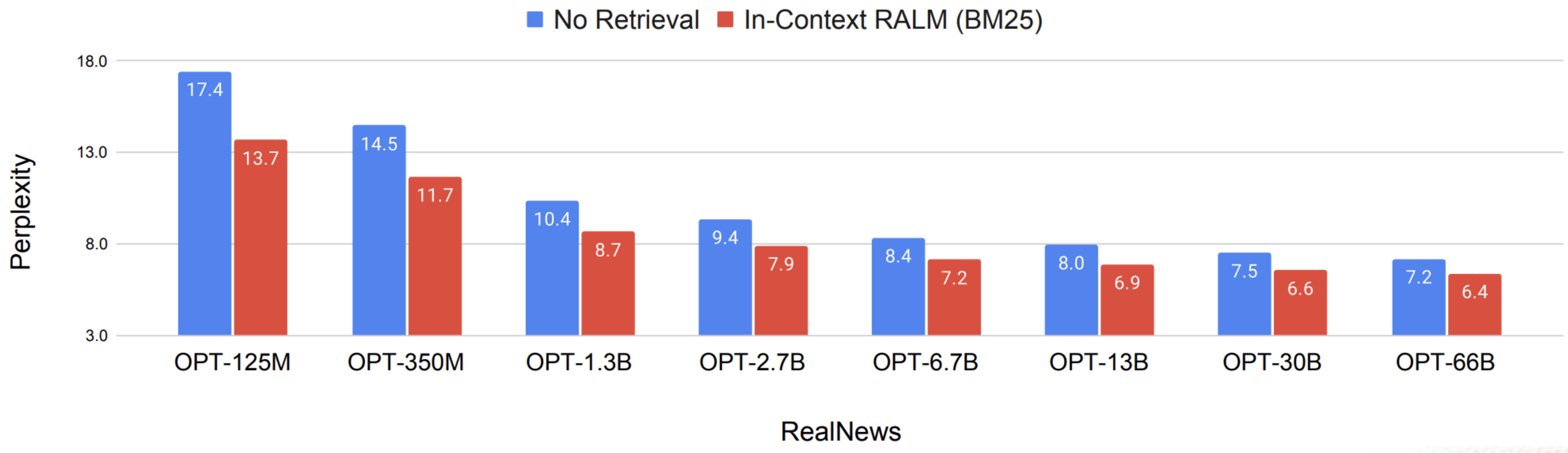


“When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories”. Mallen*, Asai* et al. ACL 2023

In-context Retrieval-augmented LMs: Result

Simply augmenting input-space of LMs give signficant gain across different tasks

- In upsream language modeling task, simply adding retrieved context gives large gains, especially smaller models
- Similar significant gains in downstream tasks such as Question Answering

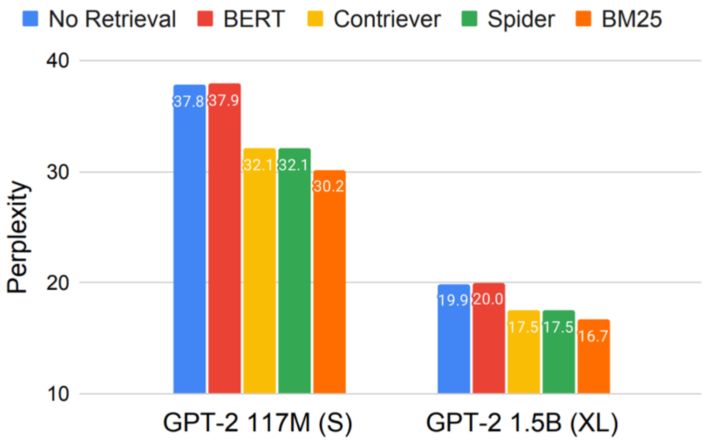


"In-Context Retrieval-Augmented Language Models." Ram et al., TACL 2023.

In-context Retrieval-augmented LMs: Result

Effects of retrieval systems for downstream task performance

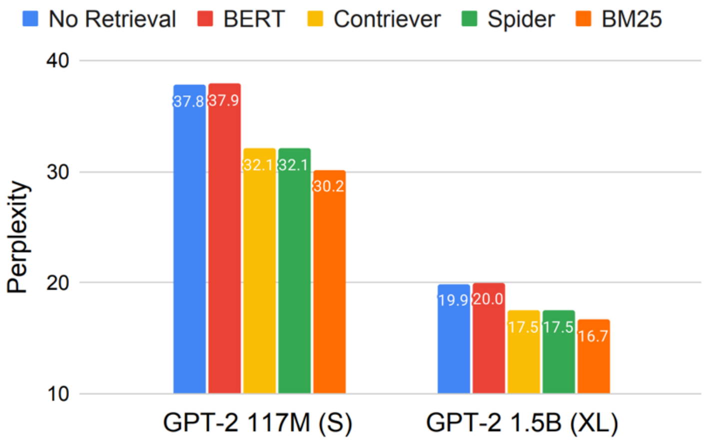
- On language modeling, BM 25 results in best performance

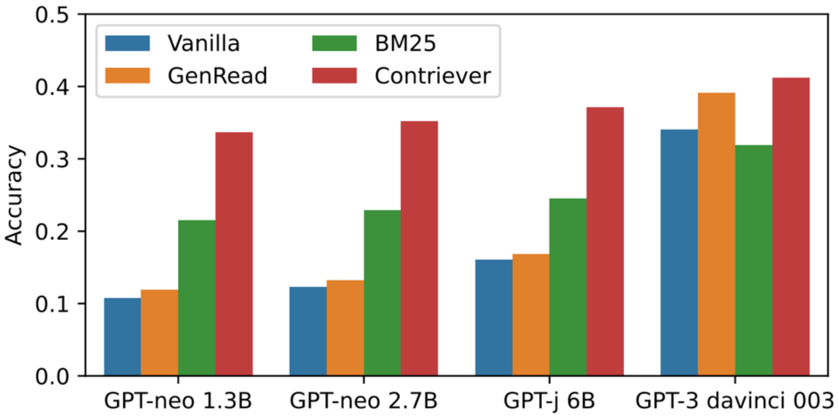


In-context Retrieval-augmented LMs: Result

Effects of retrieval systems for downstream task performance

- On language modeling, BM 25 results in best performance
- On downstream QA tasks, trained retrieval models eg Contriever results in best performance

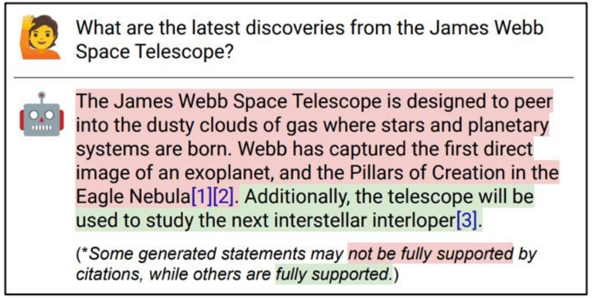


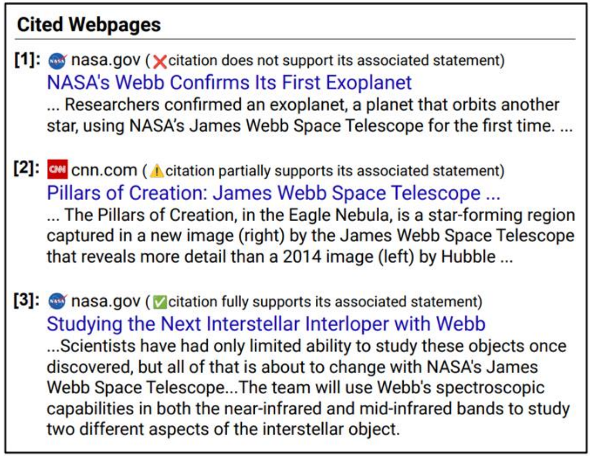


Limitations of such naïve “RAG”

Is combining off-the-shelf models sufficient?

- In-context retrieval-augmented LMs sometimes generate content that is not fully supported by their citations



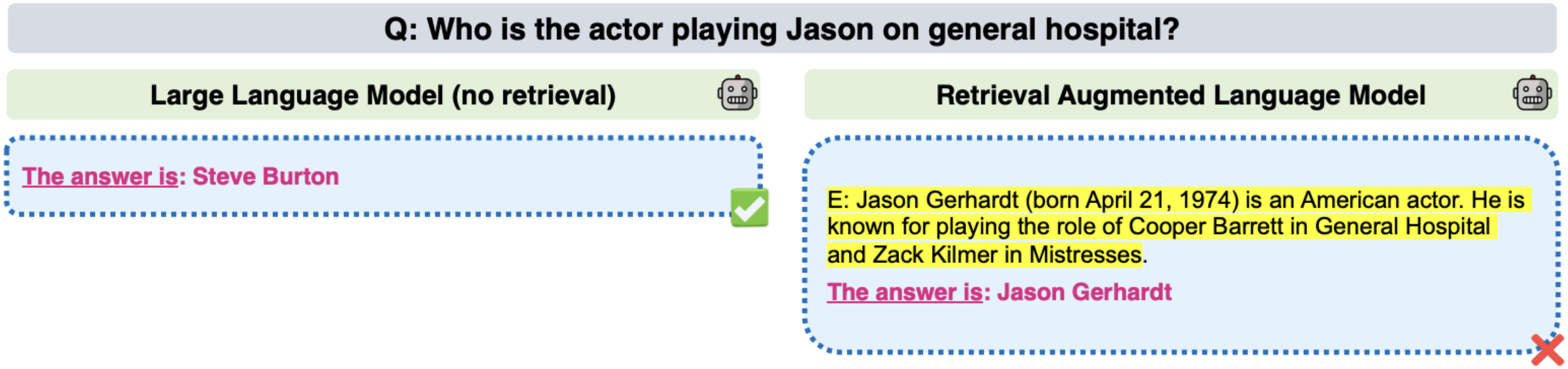


“Evaluating Verifiability in Generative Search Engines”. Liu et al. Findings of EMNLP 2023.

Limitations of such naïve "RAG"

Is combining off-the-shelf models sufficient?

- In-context retrieval-augmented LMs sometimes generate content that is not fully supported by their citations
- They can easily be distracted by unhelpful context



"Making Retrieval-Augmented Language Models Robust to Irrelevant Context". ICLR 2024.

Limitations of such naïve “RAG”

Is combining off-the-shelf models sufficient?

- In-context retrieval-augmented LMs generate what is not fully supported by their citations
- They can easily get distracted by unhelpful context
- Diverse tasks require diffent retrieval needs e.g., content, frequency

Who is the current PM of UK?

Can be easily answered based on top documents retrieved at the beginning

Limitations of such naïve "RAG"

Is combining off-the-shelf models sufficient?

- In-context retrieval-augmented LMs generate what is not fully supported by their citations
- They can easily get distracted by unhelpful context
- Diverse tasks require diffent retrieval needs e.g., content, frequency

Who is the current PM of UK?

Can be easily answered based on top documents retrieved at the beginning

Create a table listing all previous UK Prime Ministers, including their terms in office, political party, alma mater, and notable achievements.

This may require iterative retrieval, based on the current generation

Limitations of such naïve "RAG"

Is combining off-the-shelf models sufficient?

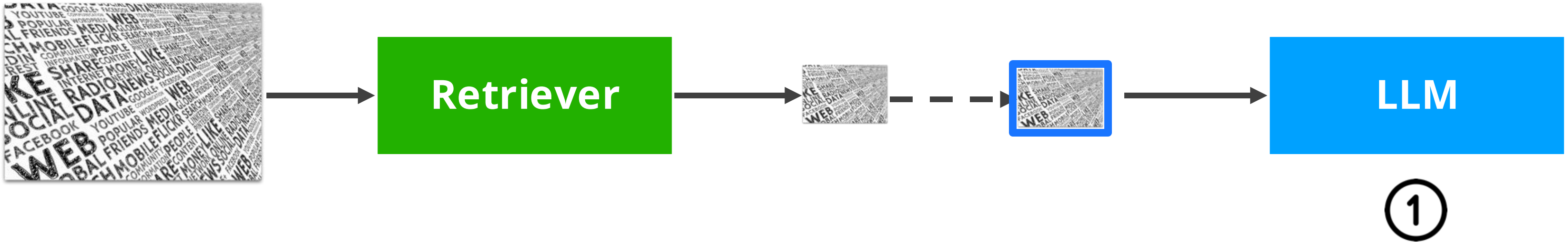
- In-context retrieval-augmented LMs generate what is not fully supported by their citations
- They can easily be distracted by unhelpful context
- Diverse tasks require diffent retrieval needs e.g., content, frequency

Who is the current PM of UK?	Can be easily answered based on top documents retrieved at the beginning
Create a table listing all previous UK Prime Ministers, including their terms in office, political party, alma mater, and notable achievements.	This may require iterative retrieval, based on the current generation
The equation x 2 + 2x = i has two complex solutions. Determine the product of their real parts (from MATH)	Questions with similar solutions may have limited semantic similarities in embedding space

Designing and training more reliable LLM RAG

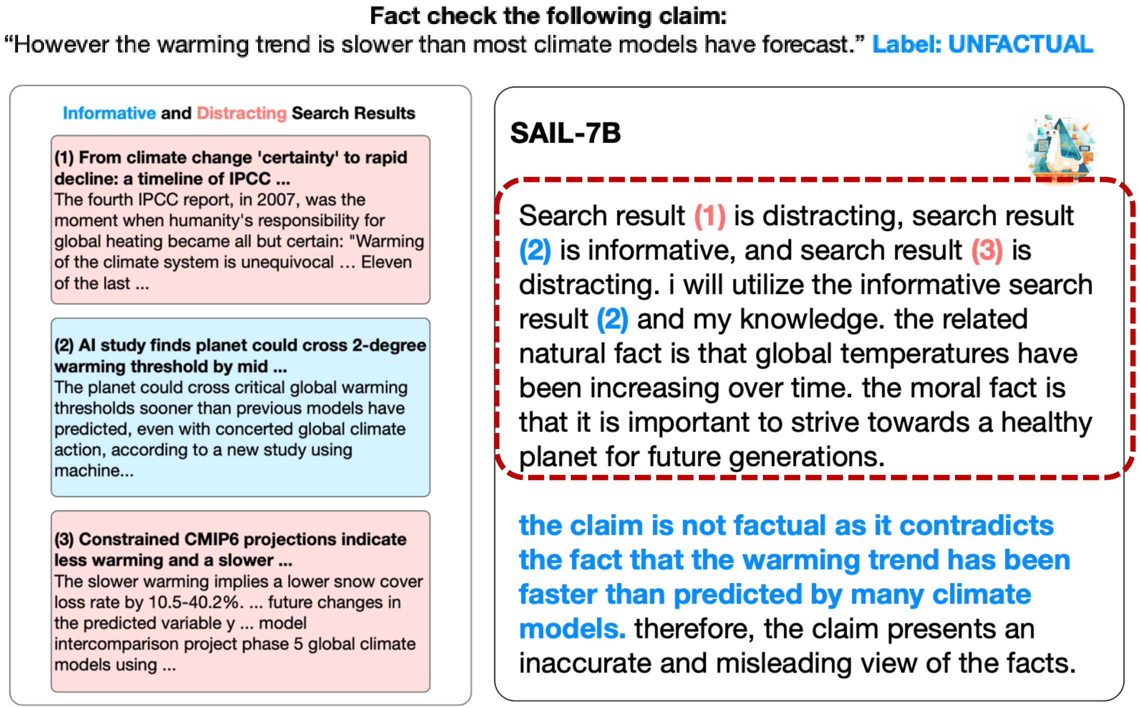
Approaches to optimize (1) LM, (2) Retrievers, or (3) prompts for LLM RAG

1. Optimizing LLMs for RAG: training / controlling LLMs with retrieved context



SAIL: Training LMs with retrieval-augmented data

SAIL augments existing instruction-tuning data to teach the LM how to use retrieved context



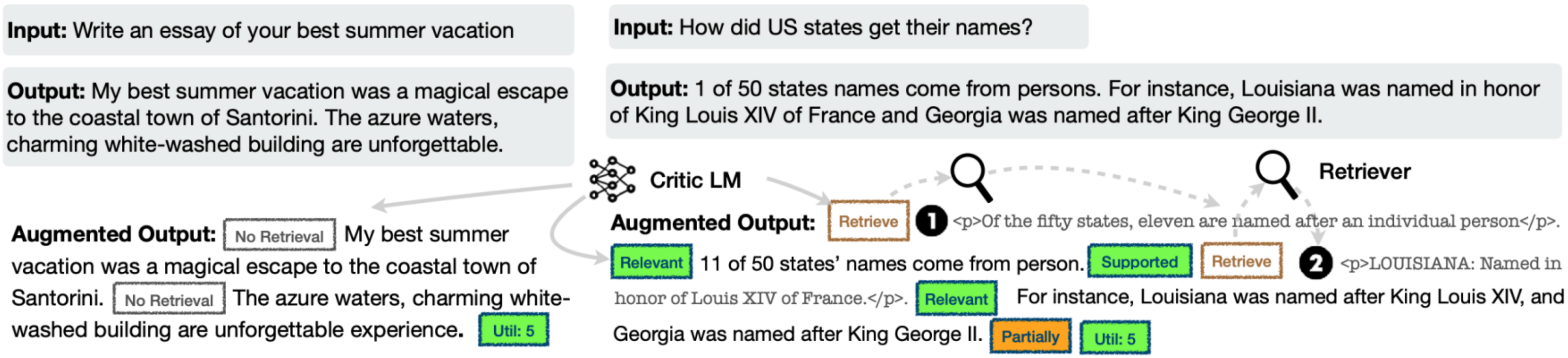
SAIL sytnthetically generates explanations by using a NLI model

"Search Augmented Instruction Learning". Luo et al. Findings of EMNLP 2024.

Self-RAG: Teaching LLMs to learn to levarage retrieved context

Self-RAG teaches LMs to adaptively retrieve and evaluates context & own generation

- Train an arbitrary LM (e.g., Llama 3) to generate special tokens for (1) triggering retrieval only when necessary and (2) evaluating the relevance of retrieved context and its own generations.

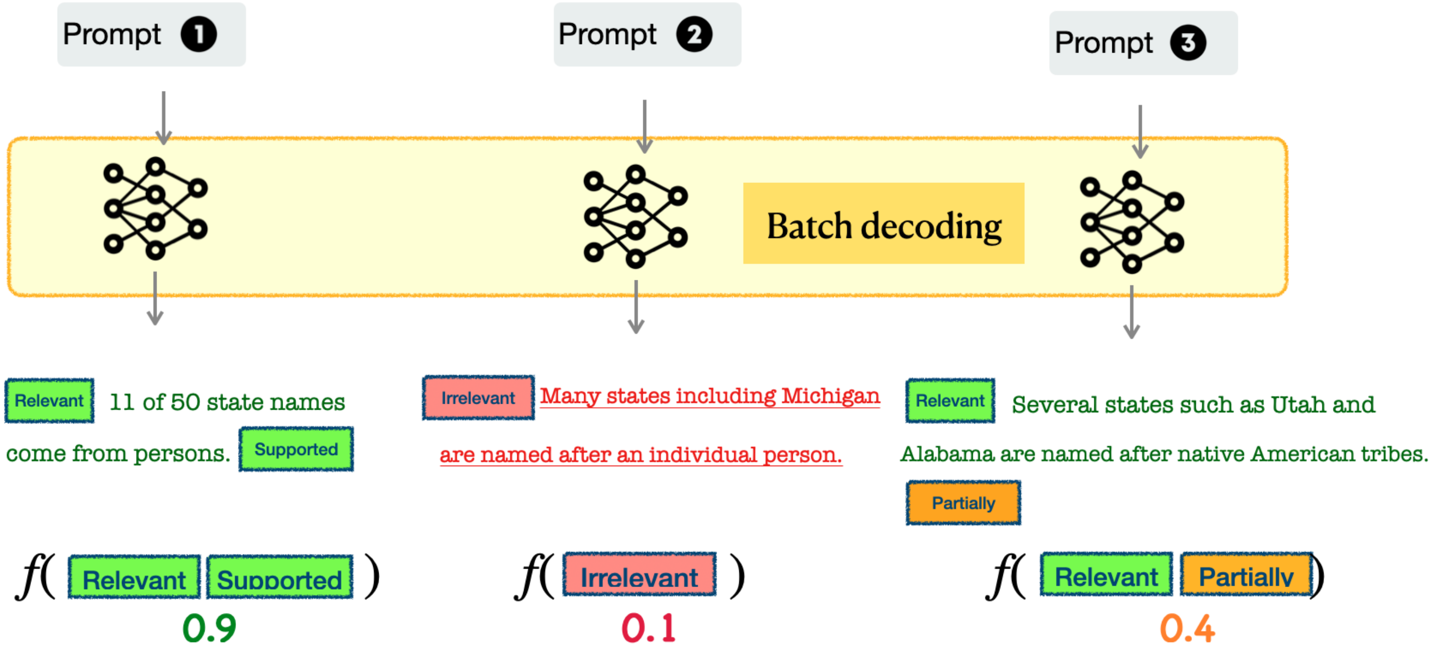


"Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection". Asai et al. ICLR 2024.

Advanced RAG inference algorithm

Advanced RAG inference algorithm to better incorporate retrieved context

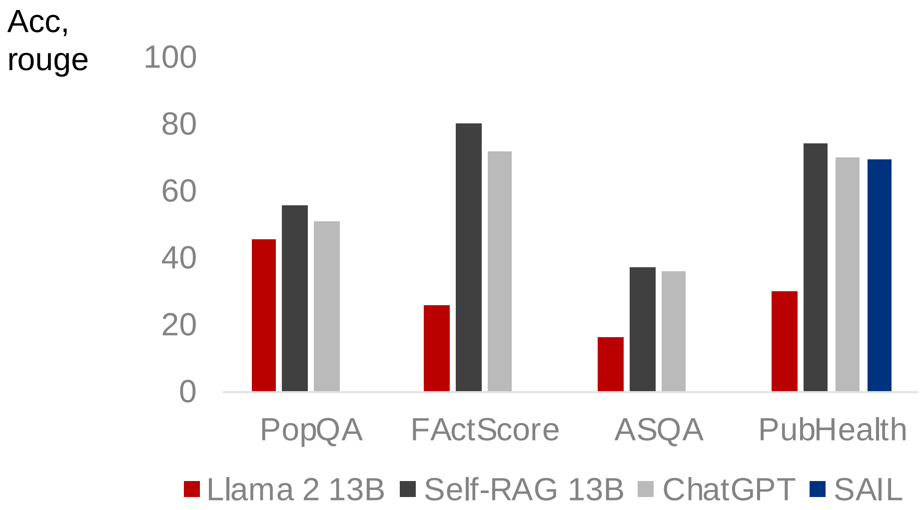
- Levarage model-generated tokens to improve search process at inference time



Optimizing LLMs for RAG: Results

New training and advanced inference algorithm for RAG significantly boost performance

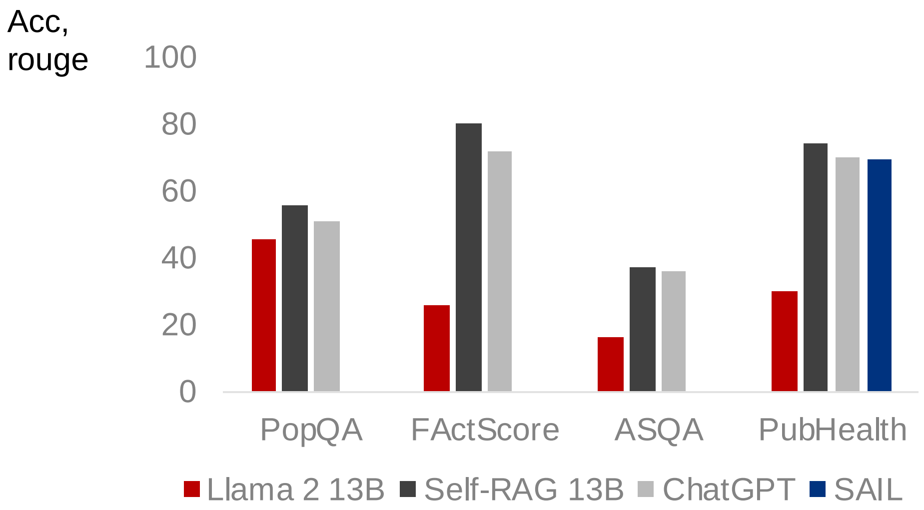
- Training with 8B and 13B models significantly boosts performance compared to off-the-shelf RAG pipelines

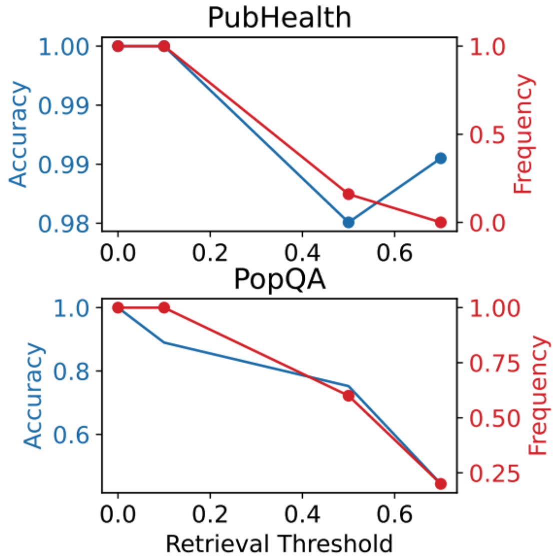


Optimizing LLMs for RAG: Results

New training and advanced inference algorithm for RAG significantly boost performance

- Training with 8B and 13B models significantly boosts performance compared to off-the-shelf RAG pipelines
- Adaptive use of retrieval also improves the efficiency of RAG systems



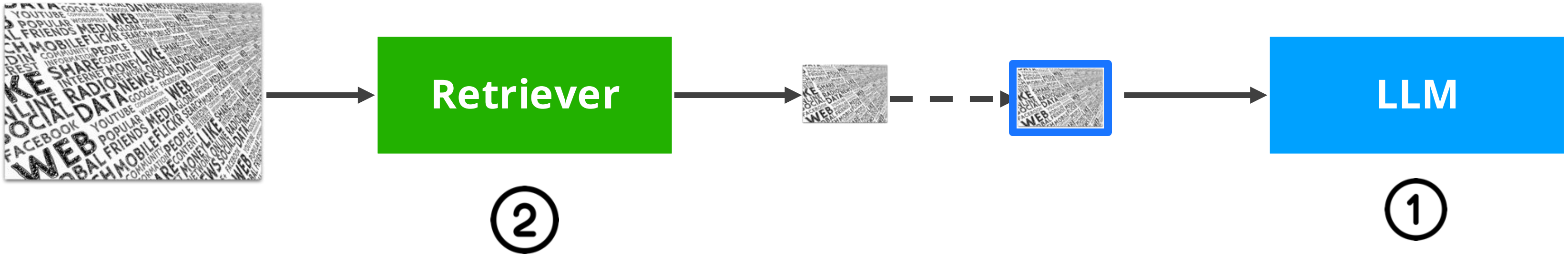


Designing and training more reliable LLM RAG

Approaches to optimize (1) LM, (2) Retrievers, or (3) prompts for LLM RAG

1. Optimizing LLMs for RAG: training / controlling LLMs with retrieved context

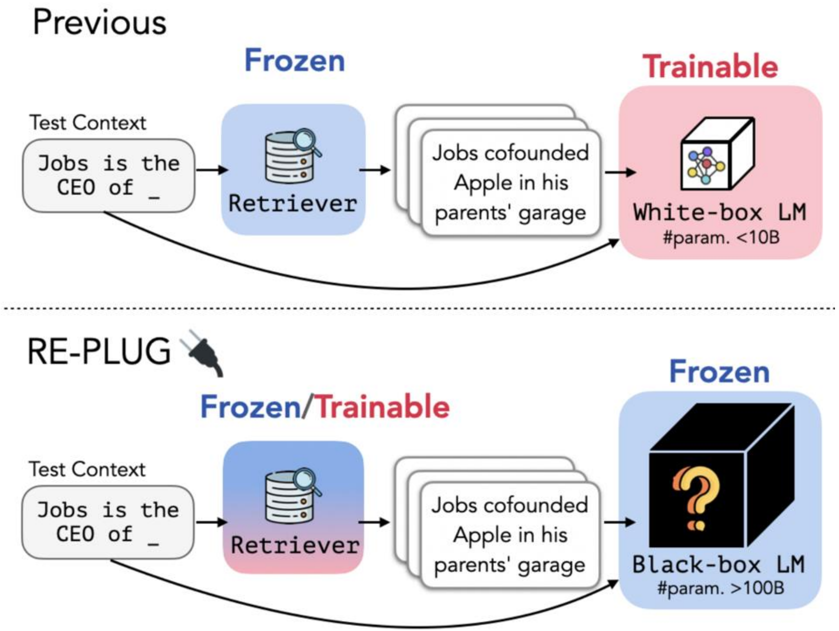
2. Optimizing Retriever for RAG: training retrievers for LLM RAG



Optimizing retrievers for RAG

Training retrieval modules using LM feedback

- For RAG pipelines using blackbox LLMs e.g., GPT o1, we cannot directly train the LLMs for RAG
- Can we train retrievers instead?

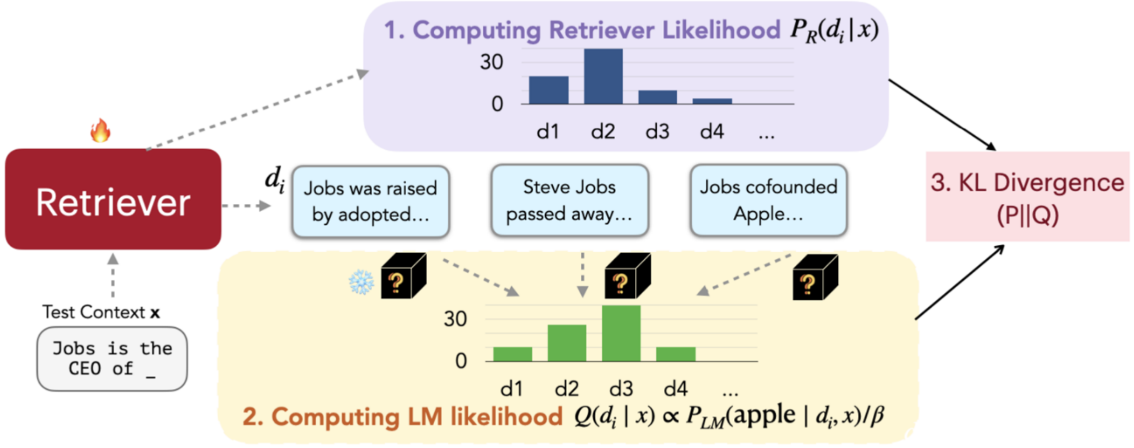


REPLUG: Training a retriever using blackbox LLM feedback

Training retrieval model using LM feedback

- Train retrievers for black-box LLMs by minimizing KL divergence between LM & retriever

$$P_R(d \mid x) = \frac{e^{s(d,x)/\gamma}}{\sum_{d \in \mathcal{D}'} e^{s(d,x)/\gamma}}$$

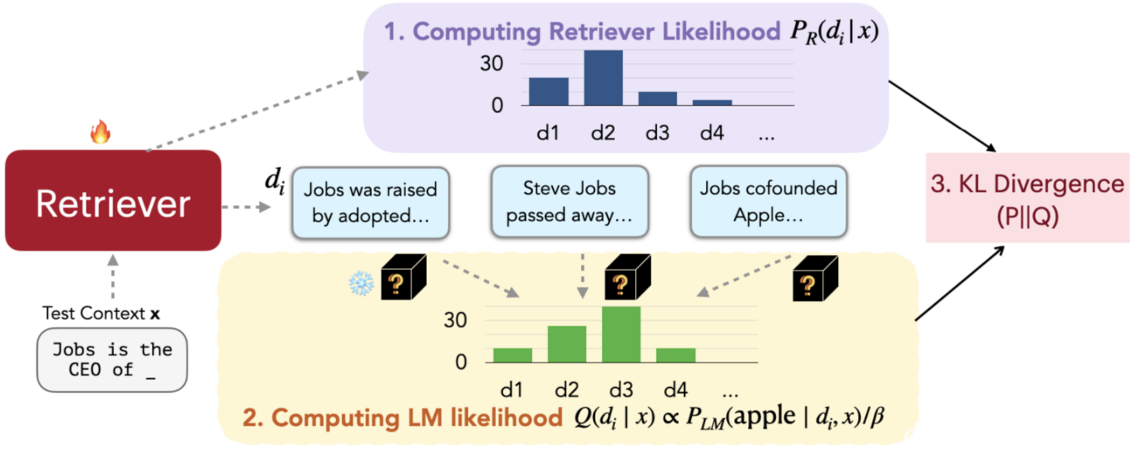


REPLUG: Training a retriever using blackbox LLM feedback

Training retrieval model using LM feedback

- Train retrievers for black-box LLMs by minimizing KL divergence between LM & retriever

$$P_R(d \mid x) = \frac{e^{s(d,x)/\gamma}}{\sum_{d \in \mathcal{D}'} e^{s(d,x)/\gamma}}$$



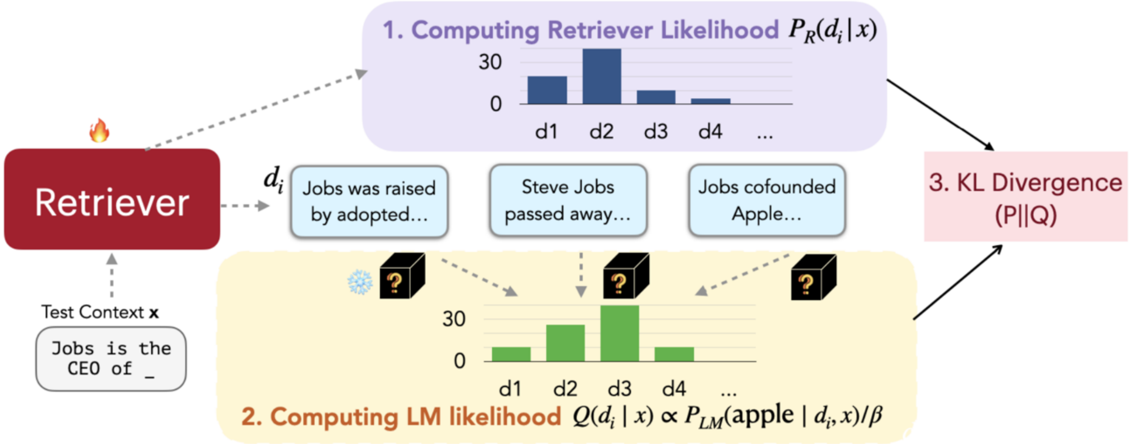
$$Q(d \mid x, y) = \frac{e^{P_{LM}(y|d,x)/\beta}}{\sum_{d \in \mathcal{D}'} e^{P_{LM}(y|d,x)/\beta}}$$

REPLUG: Training a retriever using blackbox LLM feedback

Training retrieval model using LM feedback

- Train retrievers for black-box LLMs by minimizing KL divergence between LM & retriever

$$P_R(d \mid x) = \frac{e^{s(d,x)/\gamma}}{\sum_{d \in \mathcal{D}'} e^{s(d,x)/\gamma}}$$

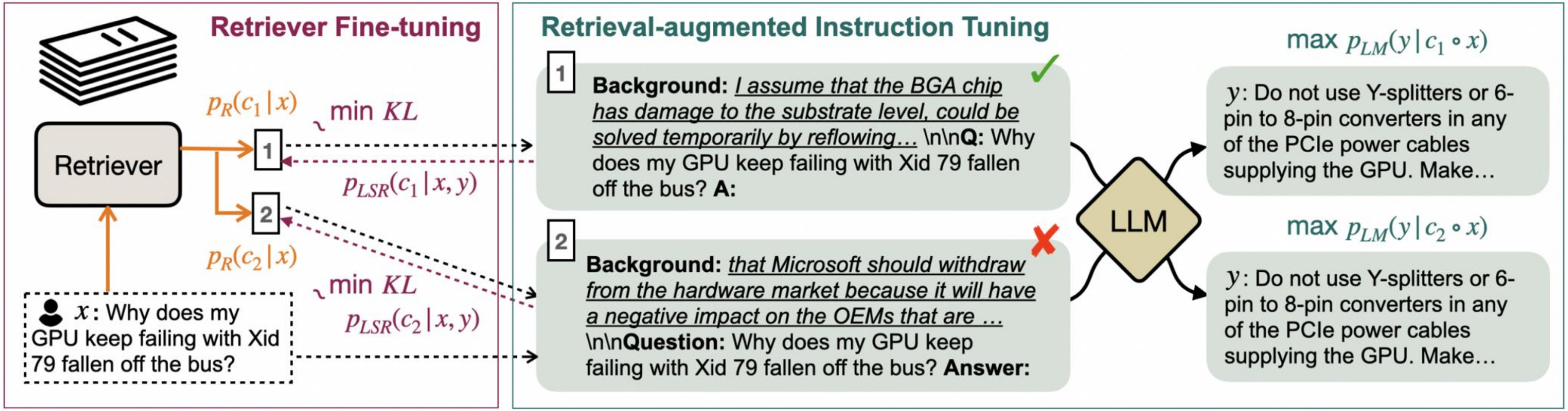


$$\frac{1}{|\mathcal{B}|}\sum_{x \in \mathcal{B}} KL\Big(P_R(d \mid x) \parallel Q_{\text{LM}}(d \mid x, y)\Big)$$

$$Q(d \mid x, y) = \frac{e^{P_{LM}(y|d,x)/\beta}}{\sum_{d \in \mathcal{D}'} e^{P_{LM}(y|d,x)/\beta}}$$

RA-DIT: Combining REPLUG + retrieval-augmented LM training

Trains both retriever and LM on multiple tasks using REPLUG + retrieval-augmented training



RA-DIT: Combining REPLUG + retrieval-augmented LM training

Trains both retriever and LM on multiple tasks using REPLUG + retrieval-augmented training

Table 1: Our intruction tuning datasets. All datasets are downloaded from Hugging Face (Lhoest et al., 2021), with the exception of those marked with ‡, which are taken from Iyer et al. (2022).

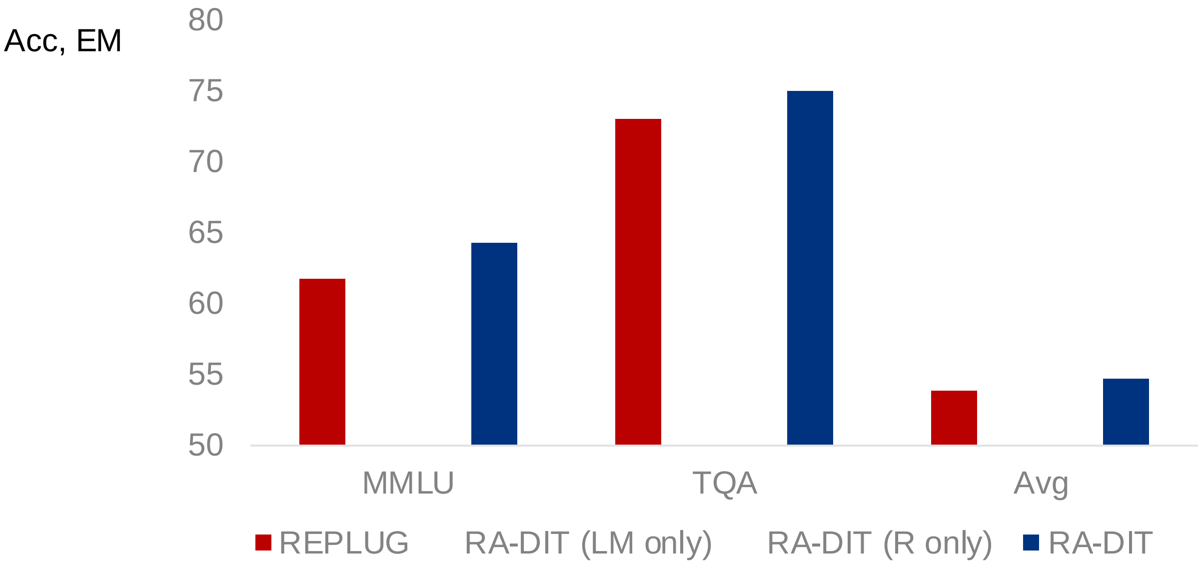
Task	HF identifier	Dataset name	$\mathcal{D}_L$	$\mathcal{D}_R$	#Train
Dialogue	oasst1	OpenAssistant Conversations Dataset (Köpf et al., 2023)	✓	✓	31,598
Open-Domain QA	commonsense_qa	CommonsenseQA (Talmor et al., 2019)	✓	✓	9,741
	math_qa	MathQA (Amini et al., 2019)	✓	✓	29,837
	web_questions	Web Questions (Berant et al., 2013)	✓	✓	3,778
	wiki_qa	Wiki Question Answering (Yang et al., 2015)	✓	✓	20,360
	yahoo_answers_qa	Yahoo! Answers QA	✓	✓	87,362
	freebase_qa	FreebaseQA (Jiang et al., 2019)		✓	20,358
	ms_marco*	MS MARCO (Nguyen et al., 2016)		✓	80,143
Reading Comprehension	coqa	Conversational Question Answering (Reddy et al., 2019)	✓		108,647
	drop	Discrete Reasoning Over Paragraphs (Dua et al., 2019)	✓		77,400
	narrativeqa	NarrativeQA (Kočiský et al., 2018)	✓		32,747
	newsqa	NewsQA (Trischler et al., 2017)	✓		74,160
	pubmed_qa	PubMedQA (Jin et al., 2019)	✓	✓	1,000
	quail	QA for Artificial Intelligence (Rogers et al., 2020)	✓		10,246
	quarel	QuaRel (Tafjord et al., 2019)	✓	✓	1,941
	squad_v2	SQuAD v2 (Rajpurkar et al., 2018)	✓		130,319
Summarization	cnn_dailymail	CNN / DailyMail (Hermann et al., 2015)	✓		287,113
Chain-of-thought Reasoning	aqua_rat‡	Algebra QA with Rationales (Ling et al., 2017)	✓		97,467
	ecqa‡	Explanations for CommonsenseQ (Aggarwal et al., 2021)	✓		7,598
	gsm8k‡	Grade School Math 8K (Cobbe et al., 2021)	✓		7,473
	compeition_math‡	MATH (Hendrycks et al., 2021b)	✓		7,500
	strategyqa‡	StrategyQA (Geva et al., 2021)	✓		2,290

* We only used the question-and-answer pairs in the MS MARCO dataset.

REPLUG, RA-DIT: Results

Training retriever & LM gives large improvements across diverse tasks

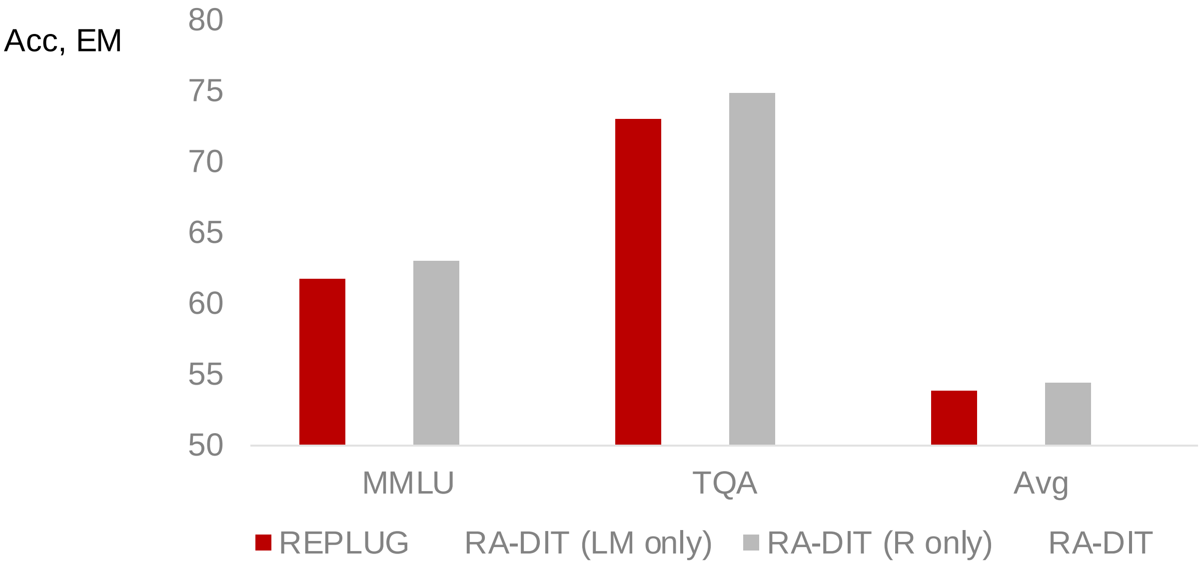
- RA-DIT observes performance gain from combinations of off-the-shelf (REPLUG w/o LSR)



REPLUG, RA-DIT: Results

Training retriever & LM gives large improvements across diverse tasks

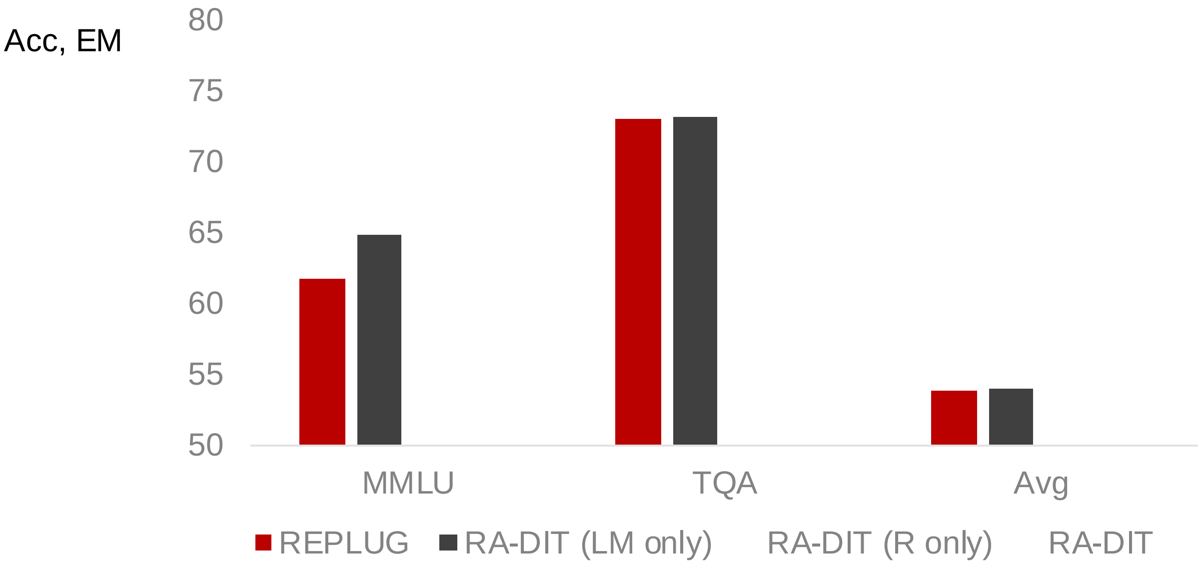
- RA-DIT observes performance gain from combinations of off-the-shelf (REPLUG w/o LSR)
- Both LM and retriever training contributes to performance gain



REPLUG, RA-DIT: Results

Training retriever & LM gives large improvements across diverse tasks

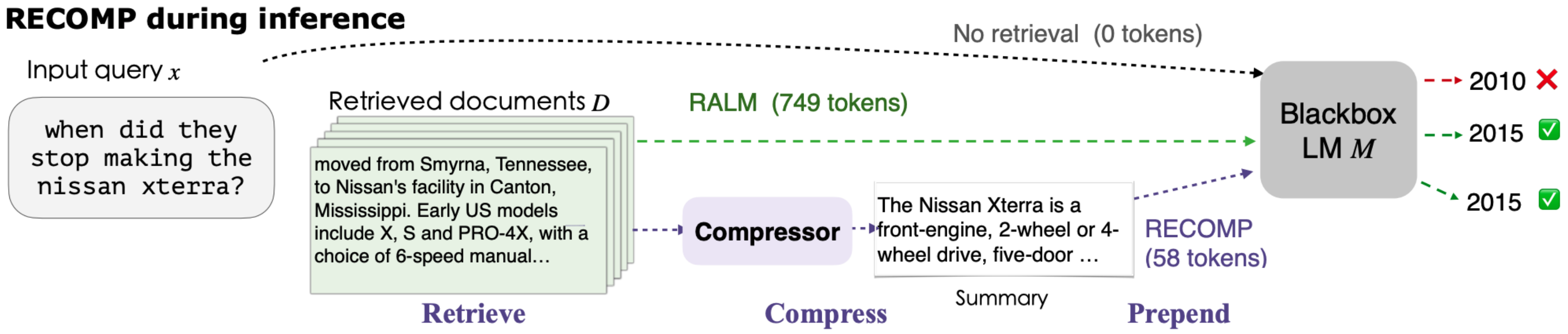
- RA-DIT observes performance gain from combinations of off-the-shelf (REPLUG w/o LSR)
- Both LM and retriever training contributes to performance gain



Optimizing retrievers for RAG

Alternative appraoches: introducing additional modules for reranking or filtering

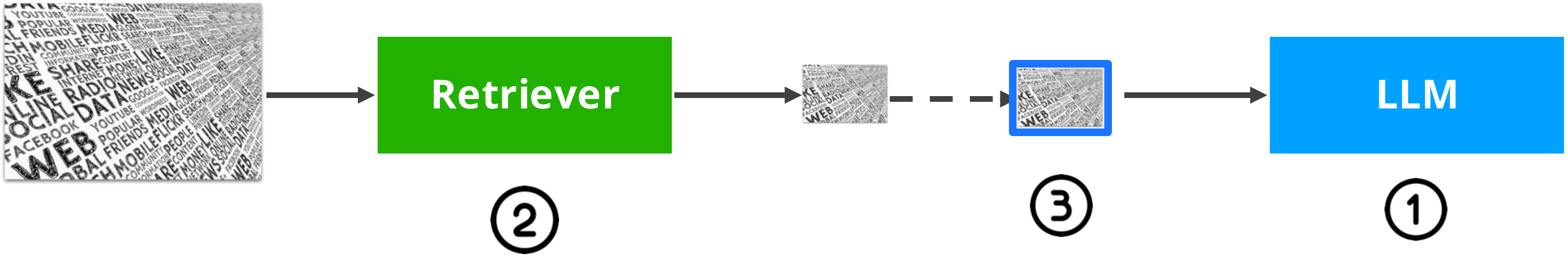
- From initial retrived docs Z, select more relevant context before feeding it to LMs
- Examples include: cross-encoder, context compression (Xi et al., 2024)



Designing and training more reliable LLM RAG

Approaches to optimize (1) LM, (2) Retrievers, or (3) prompts for LLM RAG

1. Optimizing LLMs for RAG: training / controlling LLMs with retrieved context
2. Optimizing Retriever for RAG: training retrievers for LLM RAG
3. Optimizing Prompts for RAG: advanced prompt techniques



DSPy: Optimizing prompts for LLM RAG

Optimizing prompts for RAG applications

- Training-free RAG systems are brittle to prompts

Solve a question answering task with interleaving Thought, Action, Observation steps. Thought can reason about the current situation, and Action can be three types:
(1) Search[entity], which searches the exact entity on Wikipedia and returns the first paragraph if it exists. If not, it will return some similar entities to search.
(2) Lookup[keyword], which returns the next sentence containing keyword in the current passage.
(3) Finish[answer], which returns the answer and finishes the task.
Here are some examples.

Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?
Thought 1: I need to search Colorado orogeny, find the area that the eastern sector of the Colorado orogeny extends into, then find the elevation range of the area.
Action 1: Search[Colorado orogeny]
Observation 1: The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surrounding areas.
Thought 2: It does not mention the eastern sector. So I need to look up eastern sector.
Action 2: Lookup[eastern sector]
Observation 2: (Result 1 / 1) The eastern sector extends into the High Plains and is called the Central Plains orogeny.

[... truncated …]

DSPy: Optimizing prompts for LLM RAG

Optimizing prompts for RAG applications

- Training-free RAG systems are brittle to prompts

Solve a question answering task with interleaving Thought, Action, Observation steps. Thought can reason about the current situation, and Action can be three types:
(1) Search[entity], which searches the exact entity on Wikipedia and returns the first paragraph if it exists. If not, it will return some similar entities to search.
(2) Lookup[keyword], which returns the next sentence containing keyword in the current passage.
(3) Finish[answer], which returns the answer and finishes the task.
Here are some examples.

Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends int
Thought 1: I need to search Colorado orogeny, find the area that the eastern sector of the Colorado orogeny e
elevation range of the area.
Action 1: Search[Colorado orogeny]
Observation 1: The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surr
Thought 2: It does not mention the eastern sector. So I need to look up eastern sector.
Action 2: Lookup[eastern sector]
Observation 2: (Result 1 / 1) The eastern sector extends into the High Plains and is called the Central Plains orogeny.

[... truncated …]

Scores

33%

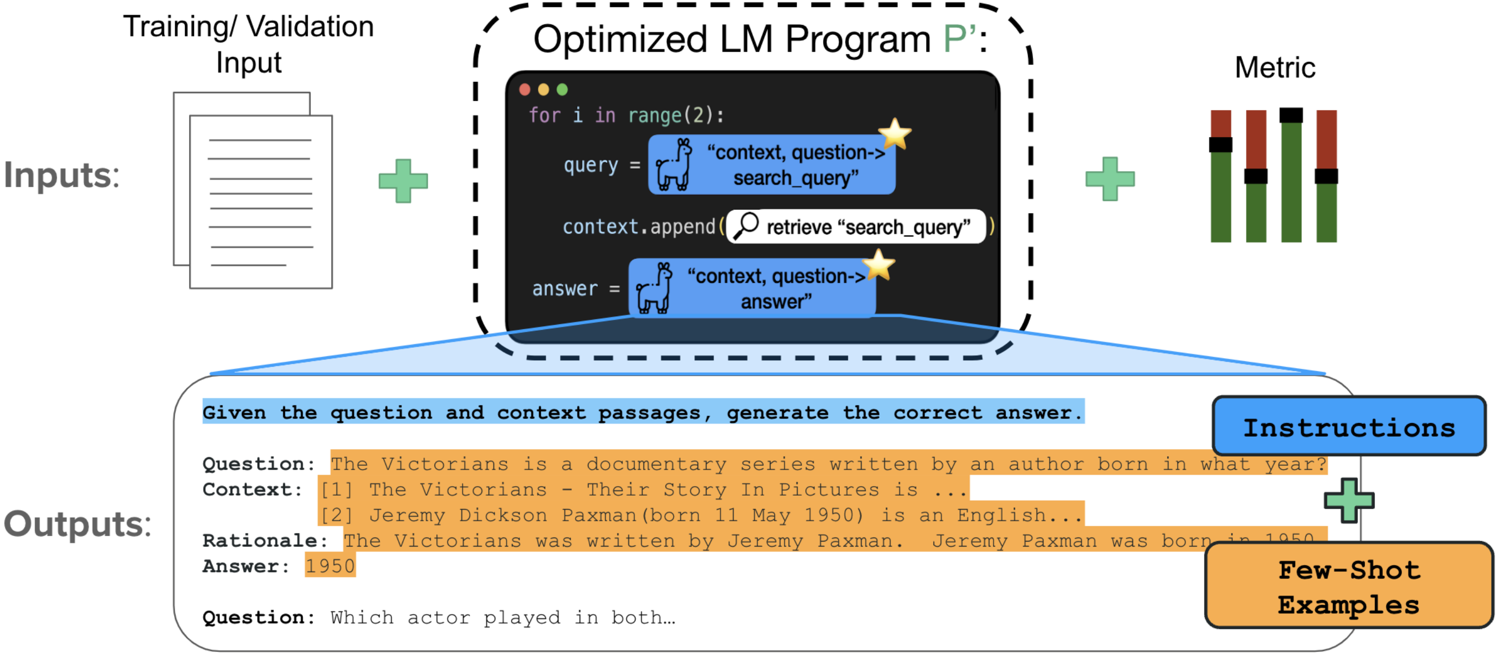
with **GPT-3.5**

on a multi-hop QA task

DSPy: Optimizing prompts for LLM RAG

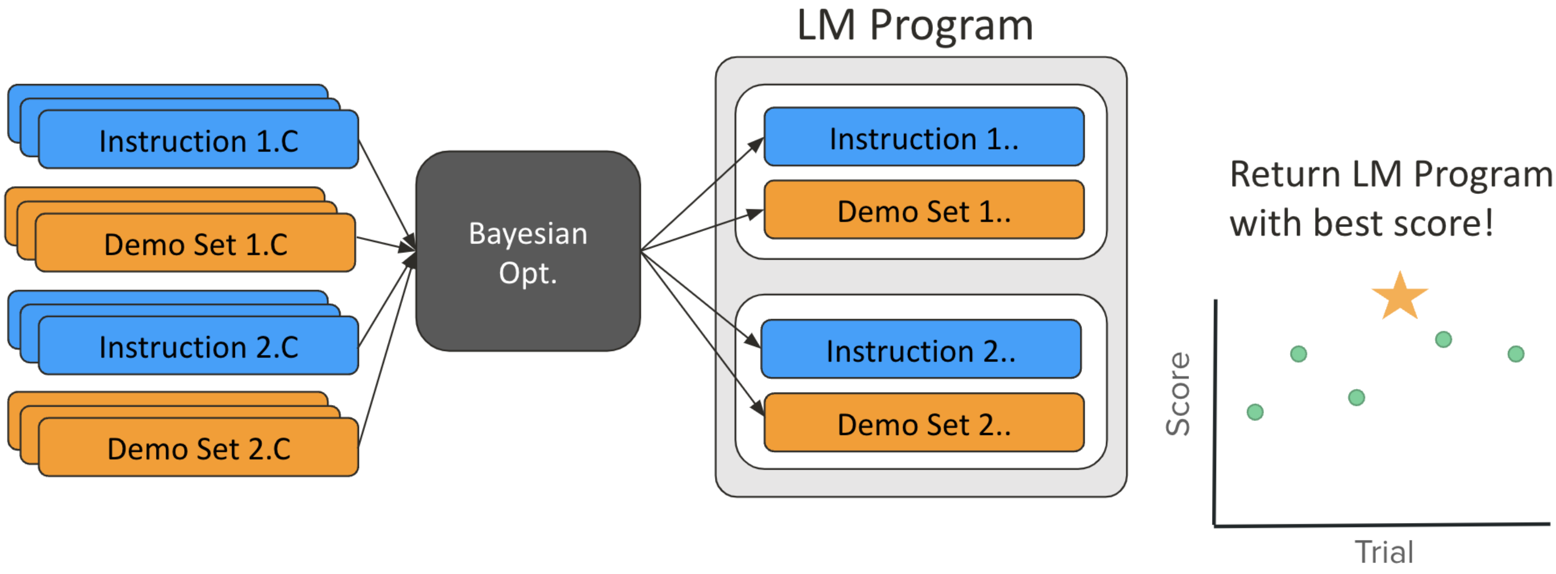
Optimizing prompts for RAG applications

- DSPy optimizes instructions and few-shot demonstrations to achieve the best performance



DSPy: Optimizing prompts for LLM RAG

Optimizing prompts for RAG applications



DSPy: Optimizing prompts for LLM RAG

```python
class MultiHop(dspy.Module):
  def __init__(self):
    self.generate_query  = dspy.ChainOfThought("context, question -> query")
    self.generate_answer = dspy.ChainOfThought("context, question -> answer")

  def forward(self, question):
    context = []
    for hop in range(2):
      query = self.generate_query(context, question).query
```

```
Carefully read the provided `context` and `question`. Your task is to formulate a concise and
relevant `search_query` that could be used to retrieve information from a search engine to answer
the question most effectively. The `search_query` should encapsulate...

Context: [1] Twilight is a series of four vampire-themed fantasy romance...
         [2] The Harper Connelly Mysteries is a series of fantasy...
Question: In which year was the first of the vampire-themed fantasy romance novels, for which The
Twilight Saga serves as a spin-off encyclopedic reference book, first published?
Reasoning: Let's determine when that fantasy romance novel was first published.
Search Query: When was the first of the vampire-themed fantasy romance novels published?

Context: [1] The Victorians - Their Story In Pictures is a 2009 British documentary ...
         [2] The Caxtons: A Family Picture is an 1849 Victorian novel by Edward ...
Question: The Victorians is a documentary series written by an author born in what year?
Reasoning: We know that the documentary series is about Victorian art and culture, and it was
written by Jeremy Paxman. We need to find the year in which Jeremy Paxman was born.
Search Query: Jeremy Paxman birth year
```

Scores

55%

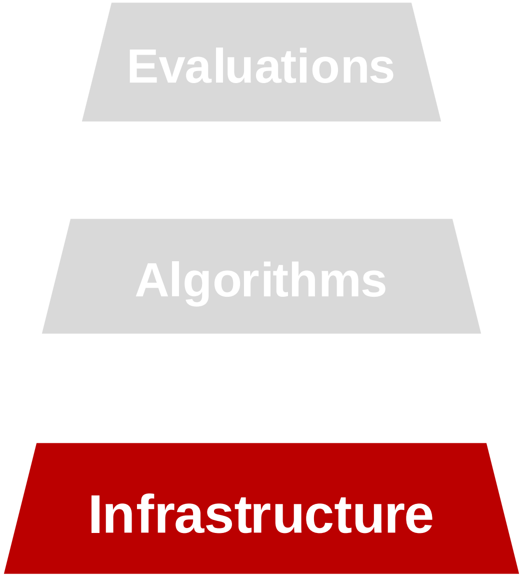
with **GPT-3.5**

on a multi-hop QA task

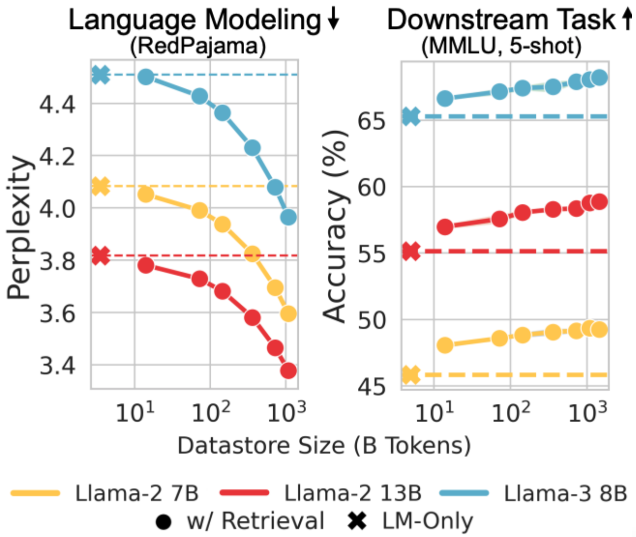
Future: *Limitations & future directions*

Roadmap for more efficient & reliable retrieval-augmented LMs

Challenges of scaling up datastores & increased inference-time costs



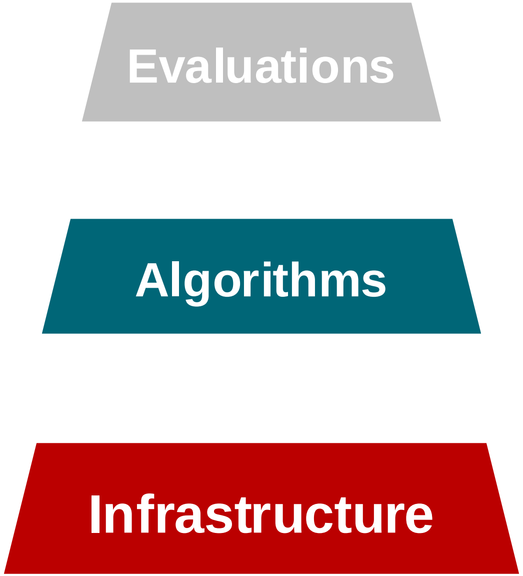
- Performance gains are achieved by scaling up the datastore to trillions of tokens
- Significantly increases inference costs, including CPU memory and storage requirements (e.g., 24 TB for 1.7 trillion-token).



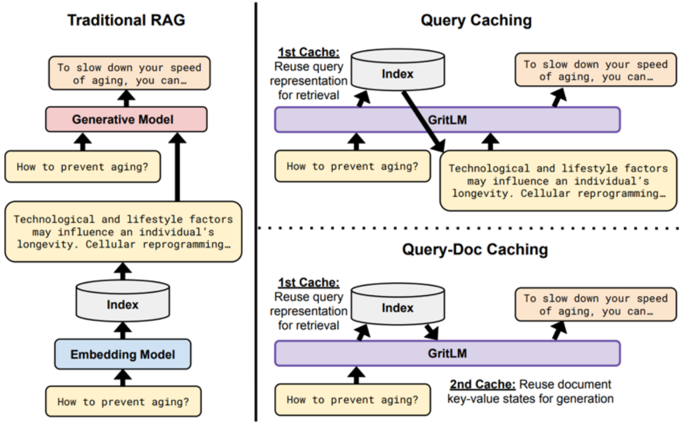
"Scaling Retrieval-Based Language Models with a Trillion-Token Datastore." Shao, He, Asai et al., ArXiv 2024.

Roadmap for more efficient & reliable retrieval-augmented LMs

New algorithms & arthictectures to enable more efficient and effective RAG

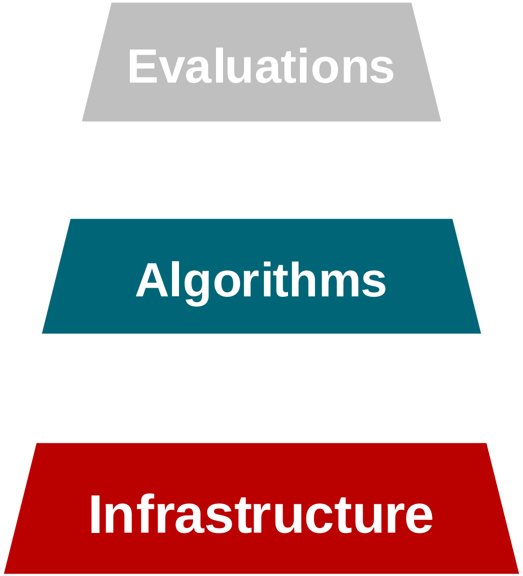


- Current “RAG” has many issues such as efficiency & redundancy
- Alternative algorithms, better LM architectures, caching … etc for improving efficiency and performance

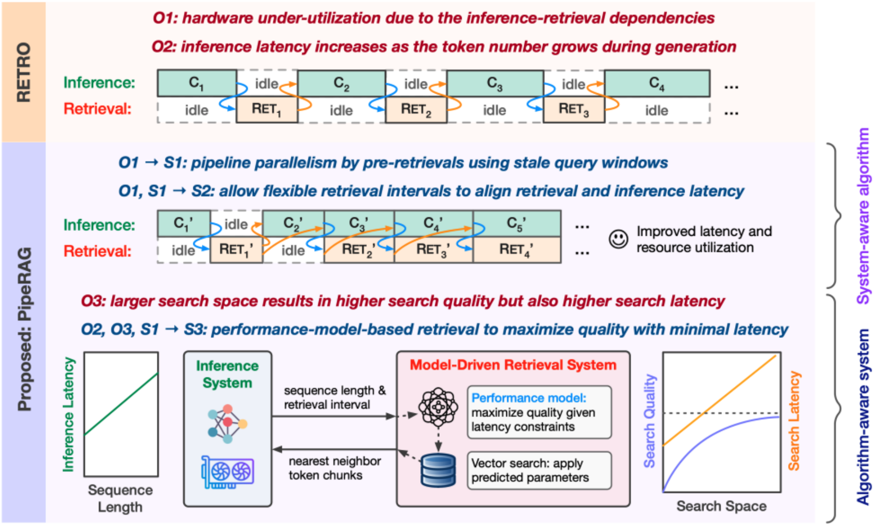


Roadmap for more efficient & reliable retrieval-augmented LMs

New algorithms & arthictectures to enable more efficient and effective RAG



- Current “RAG” has many issues such as efficiency & redundancy
- Alternative algorithms, better LM architectures, caching … etc for improving efficiency and performance



“PipeRAG: Fast Retrieval-Augmented Generation via Algorithm-System Co-design.” Jiang et al., ArXiv 2024.

Roadmap for more efficient & reliable retrieval-augmented LMs

Careful analyses on their effectiveness and limitations

Evaluations

Algorithms

Infrastructure

Prior systems are often evaluated only on simple general-domain tasks. Further exploration into their evaluation are needed

- Domains: most prior evaluations are in general-domain tasks, where Wikipedia is a sufficient knowledge source
- Tasks: going beyond open-domain QA, multiple-choice QA
- Aspects: instead of merely evaluating final “correctness”, more holistic evaluations of different aspects of RAG

Questions?

Sli.do code #2068655

Acknowledgements: Some slides are adapted from our ACL 2023 tutorials https://acl2023-retrieval-lm.github.io/ co-taught by Akari, Sewon Min, Zexuan Zhong and Danqi Chen. We thank Omar Khattab for sharing the DSPy slides