LLM LE6 VT2025

Inference, RAG, and Reasoning

Fredrik Heintz

Dept. of Computer Science
Linköping University

fredrik.heintz@liu.se
@FredrikHeintz

Outline:

- **Inference**
- **Retrieval Augmented Generation**
- **In-Context Learning**
- **Reasoning**

LINKÖPING
UNIVERSITY

Language modelling

- **Language modelling** is the task of predicting which word comes next in a sequence of words.
- More formally, given a sequence of words $w_1, \dots, w_t$ we want to know the probability of the next word, w_{t+1}:

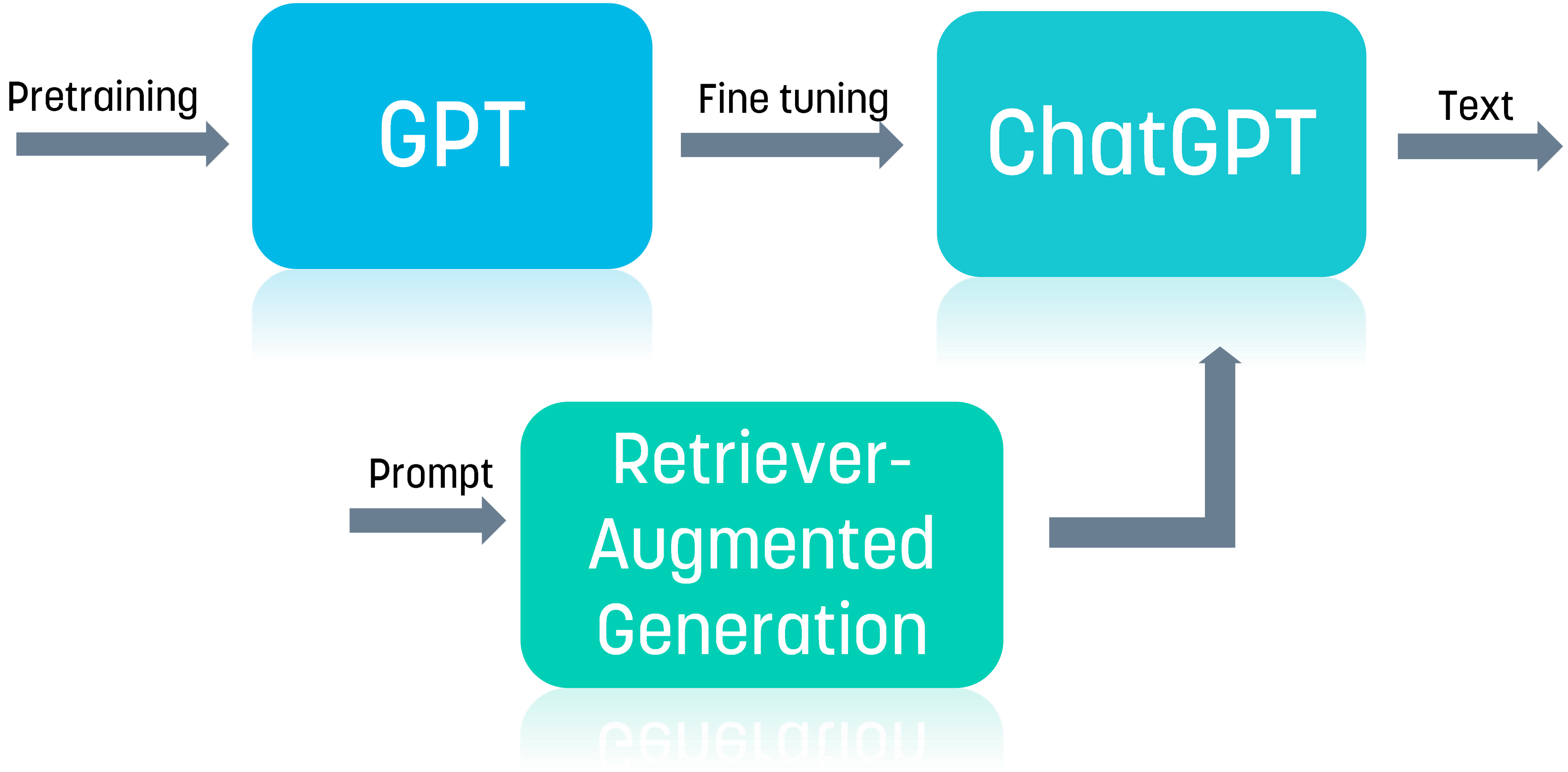
$$P(w_{t+1} \mid w_1, \dots, w_t)$$

- We are assuming that w_{t+1} comes from a finite vocabulary V.

language models = classifiers

Funded by the European Union

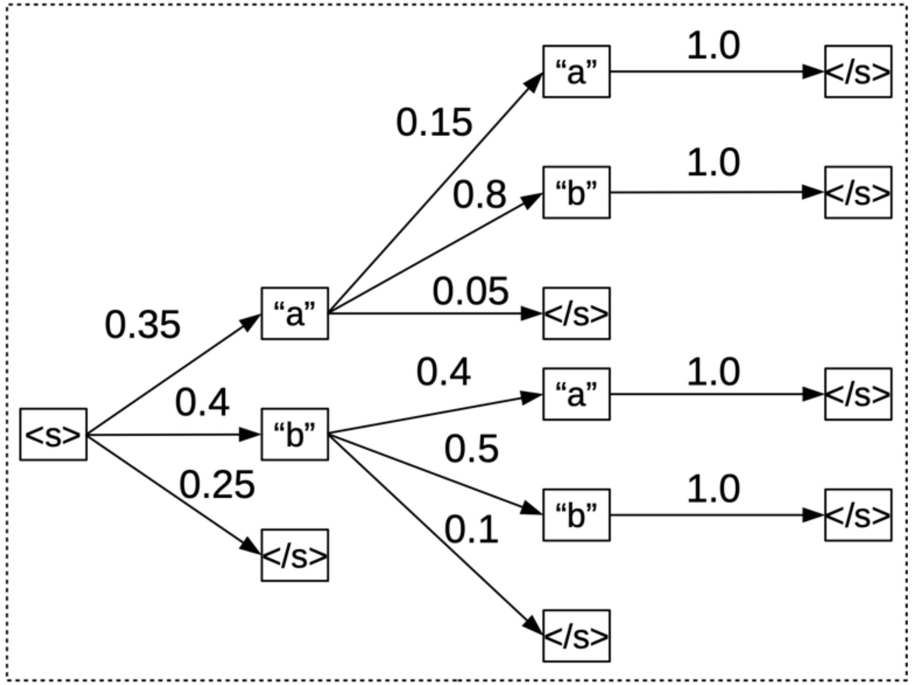
How Does ChatGPT Work?



Funded by the European Union

Inference

Greedy search methods do not always lead to the most likely output.



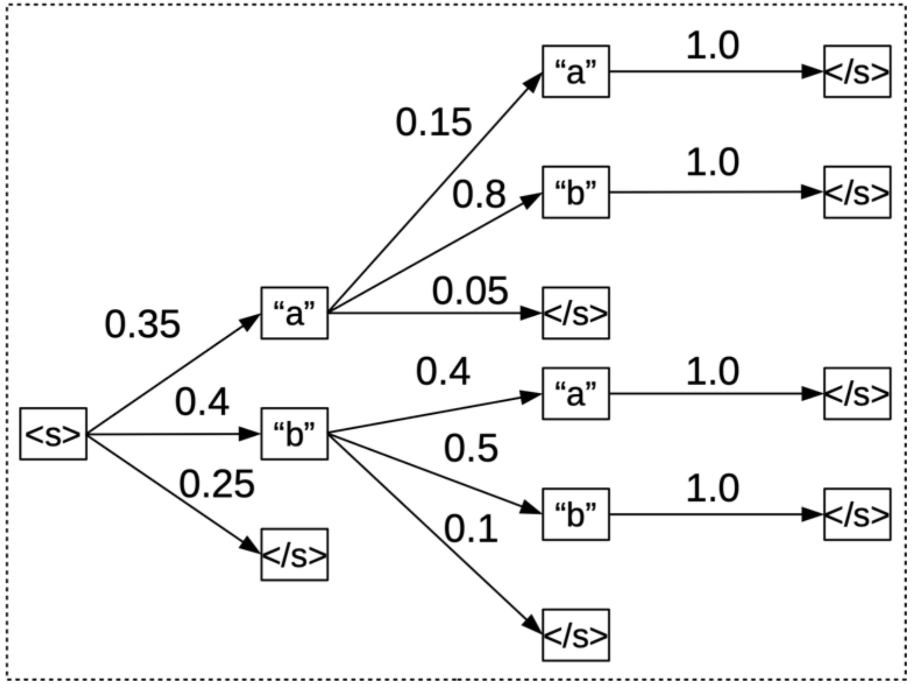
Vocabulary = {a, b, </s>}
Numbers above each edge are the transition probabilities $P(x_t|x_{1:t-t})$

If we were to choose the sequence that maximizes $P(x_1, \dots, x_T)$, which of the following would get generated?

(a) [a, b, </s>]
(b) [a, a, </s>]
(c) [b, b, </s>]
(d) [b, a, </s>]

Funded by the European Union

Greedy search methods do not always lead to the most likely output.



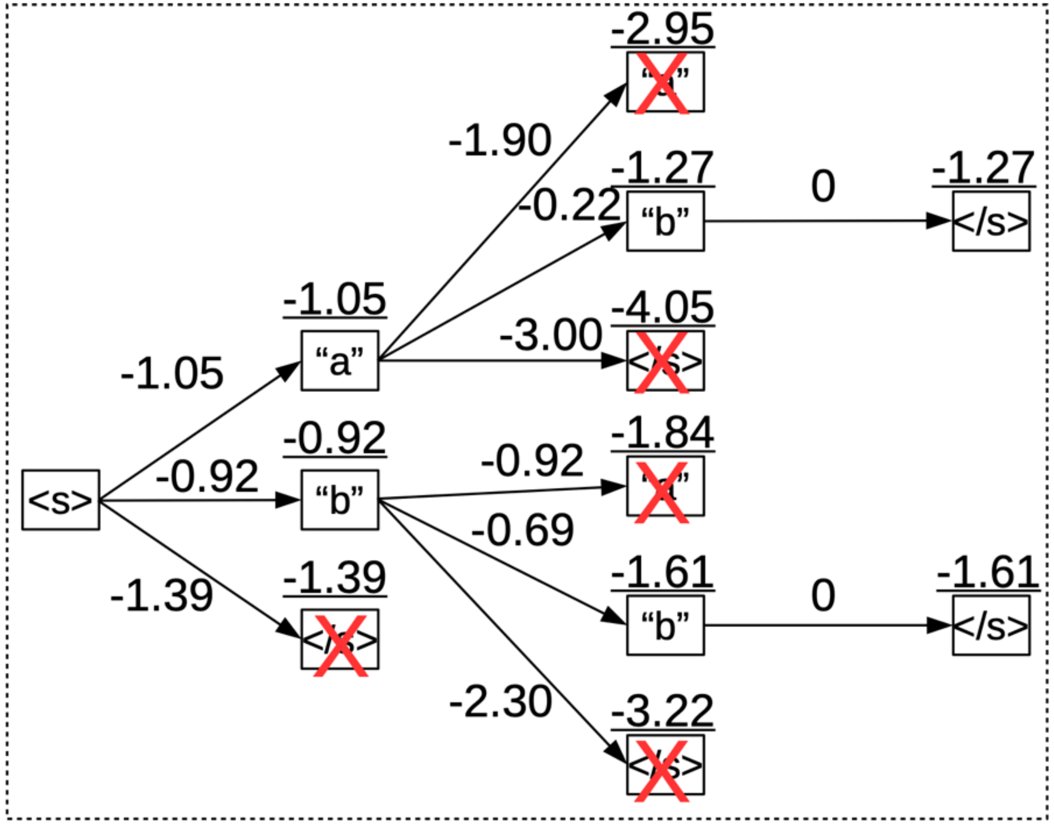
Vocabulary = {a, b, </s>}
Numbers above each edge are the transition probabilities $P(x_t|x_{1:t-t})$

If we were to choose the sequence that maximizes $P(x_1, \dots, x_T)$, which of the following would get generated?

(a) [a, b, </s>]
(b) [a, a, </s>]
(c) [b, b, </s>]
(d) [b, a, </s>]

Funded by the European Union

Beam search explores multiple possible output sequences, trying to find the overall most likely one.



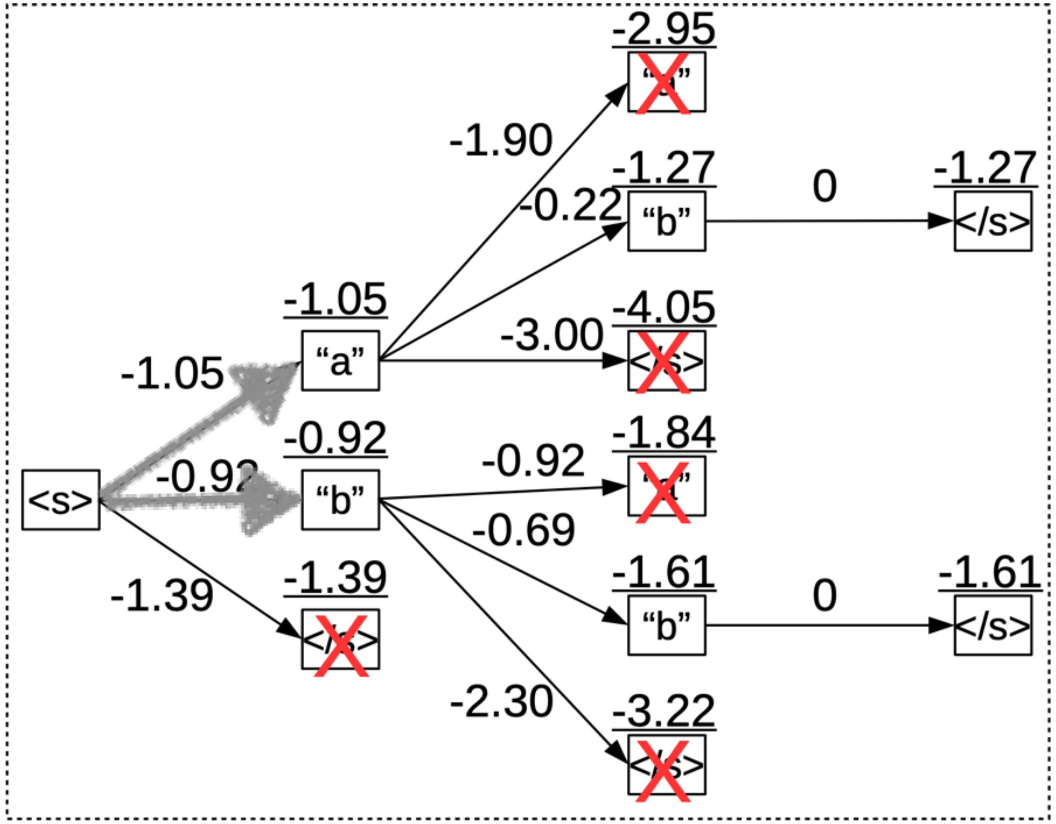
Vocabulary = {a, b, </s>}

Numbers above the boxes are $logP(x_t|x_{1:t-1})$

Numbers shown on edges are $logP(x_1, \dots, x_t)$

Suppose we use beam search with a **beam size** of 2.

Beam search explores multiple possible output sequences, trying to find the overall most likely one.



Vocabulary = {a, b, </s>}

Numbers above the boxes are $logP(x_t|x_{1:t-1})$

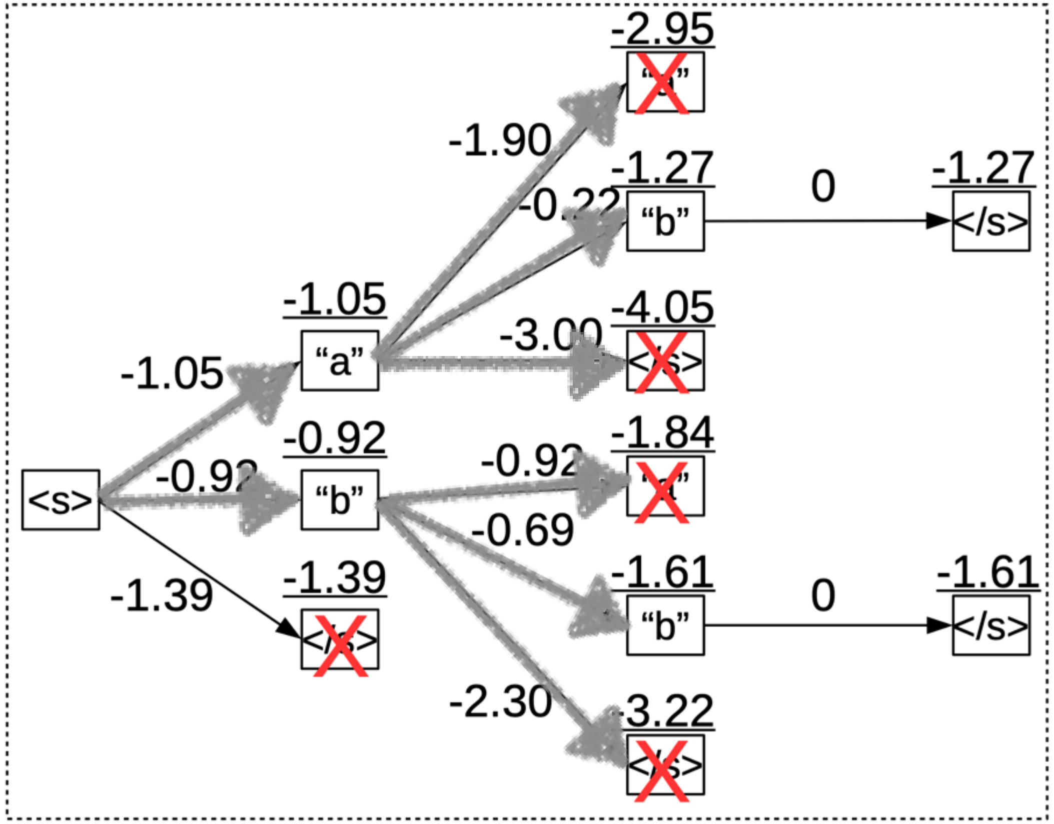
Numbers shown on edges are $logP(x_1, \ldots, x_t)$

Suppose we use beam search with a **beam size** of 2.

Score each path and keep the top 2

Funded by the European Union

Beam search explores multiple possible output sequences, trying to find the overall most likely one.



Score each path and keep the top 2

Vocabulary = {a, b, </s>}

Numbers above the boxes are $\log P(x_t|x_{1:t-1})$

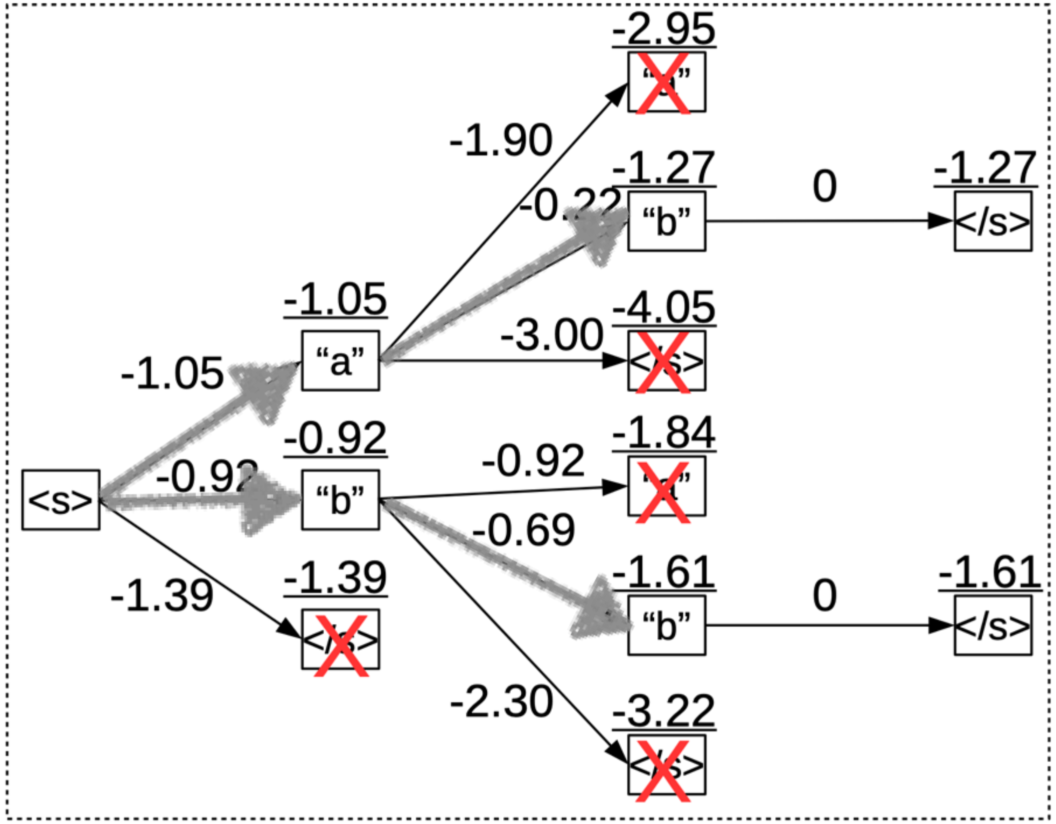
Numbers shown on edges are $\log P(x_1, \dots, x_t)$

Suppose we use beam search with a **beam size** of 2.

TrustLLM

Funded by the European Union

Beam search explores multiple possible output sequences, trying to find the overall most likely one.



Score each path and keep the top 2

Vocabulary = {a, b, </s>}

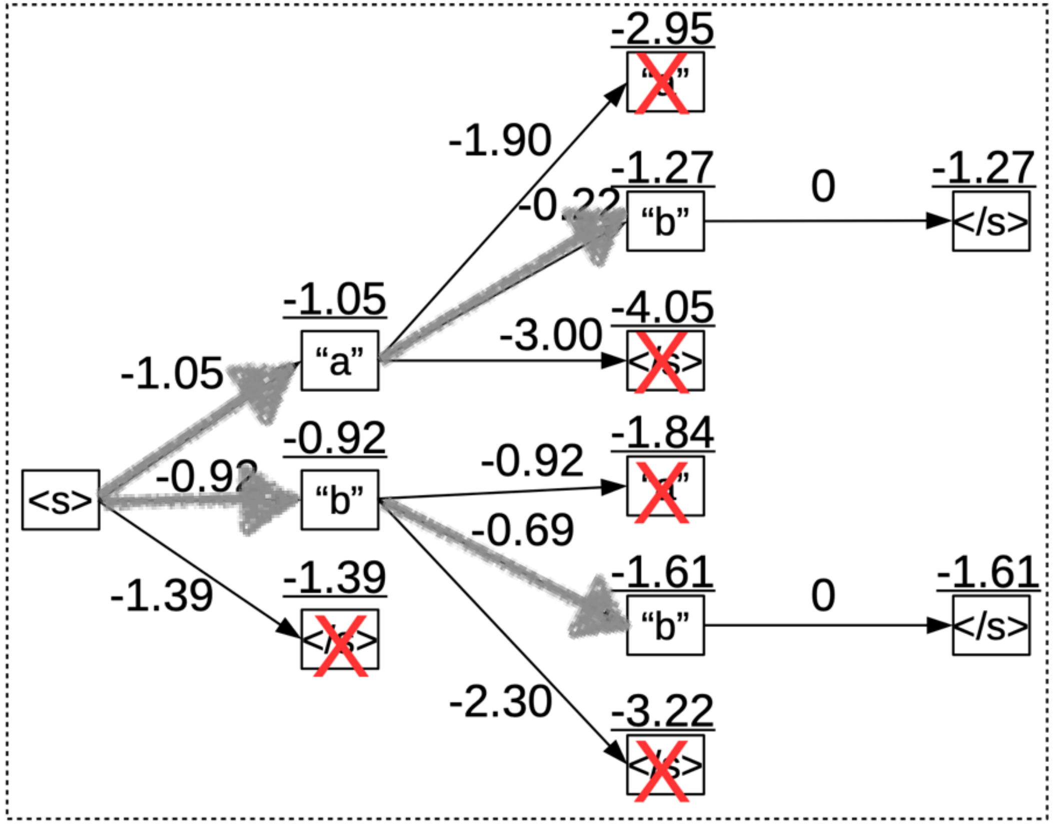
Numbers above the boxes are $\log P(x_t | x_{1:t-1})$

Numbers shown on edges are $\log P(x_1, \ldots, x_t)$

Suppose we use beam search with a **beam size** of 2.

Funded by the European Union

Beam search explores multiple possible output sequences, trying to find the overall most likely one.



Score each path and keep the top 2

Vocabulary = {a, b, </s>}

Numbers above the boxes are $\log P(x_t | x_{1:t-1})$

Numbers shown on edges are $\log P(x_1, \ldots, x_t)$

Suppose we use beam search with a **beam size** of 2.

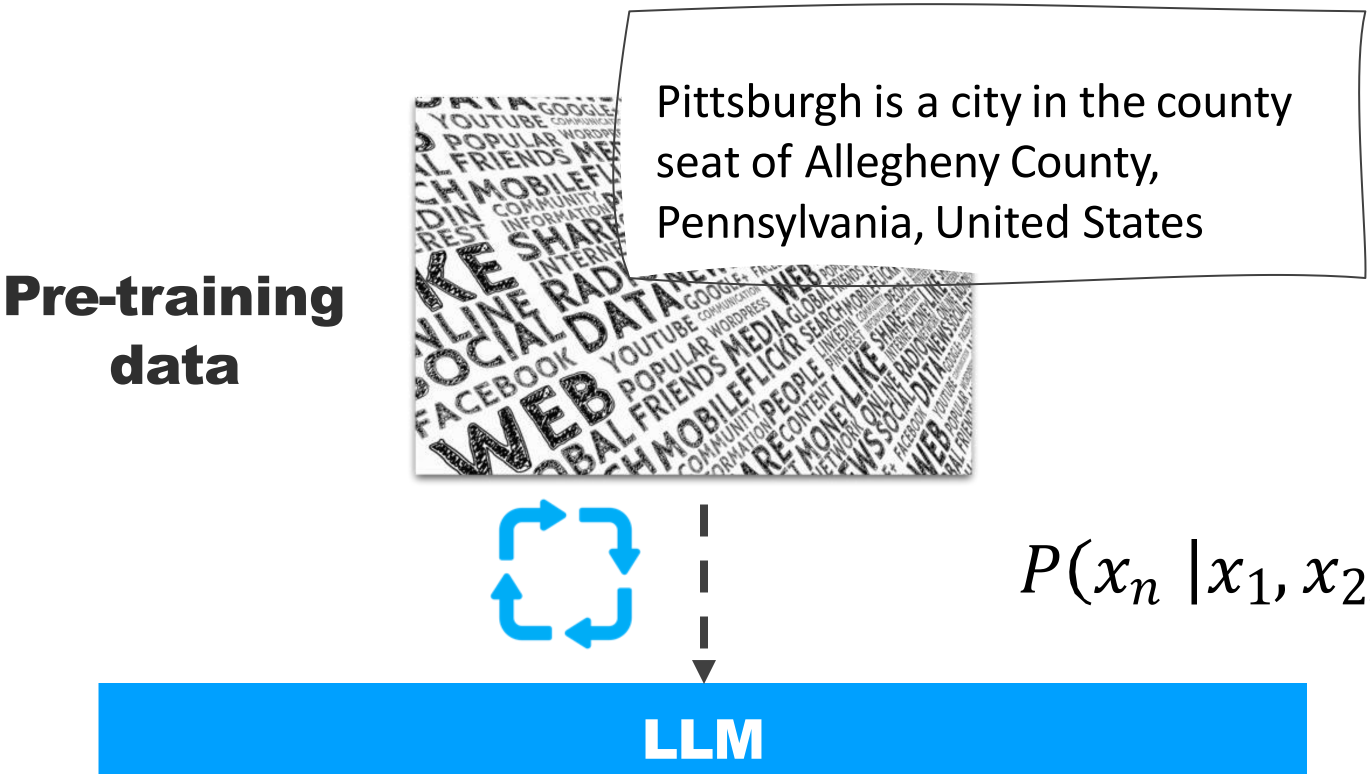
The paths that are still open at any step of the beam search algorithm are called **beams**.

Funded by the European Union

Retrieval Augmented Generation (RAG)

How do normal parametric LLMs work?

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



Pre-training data

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

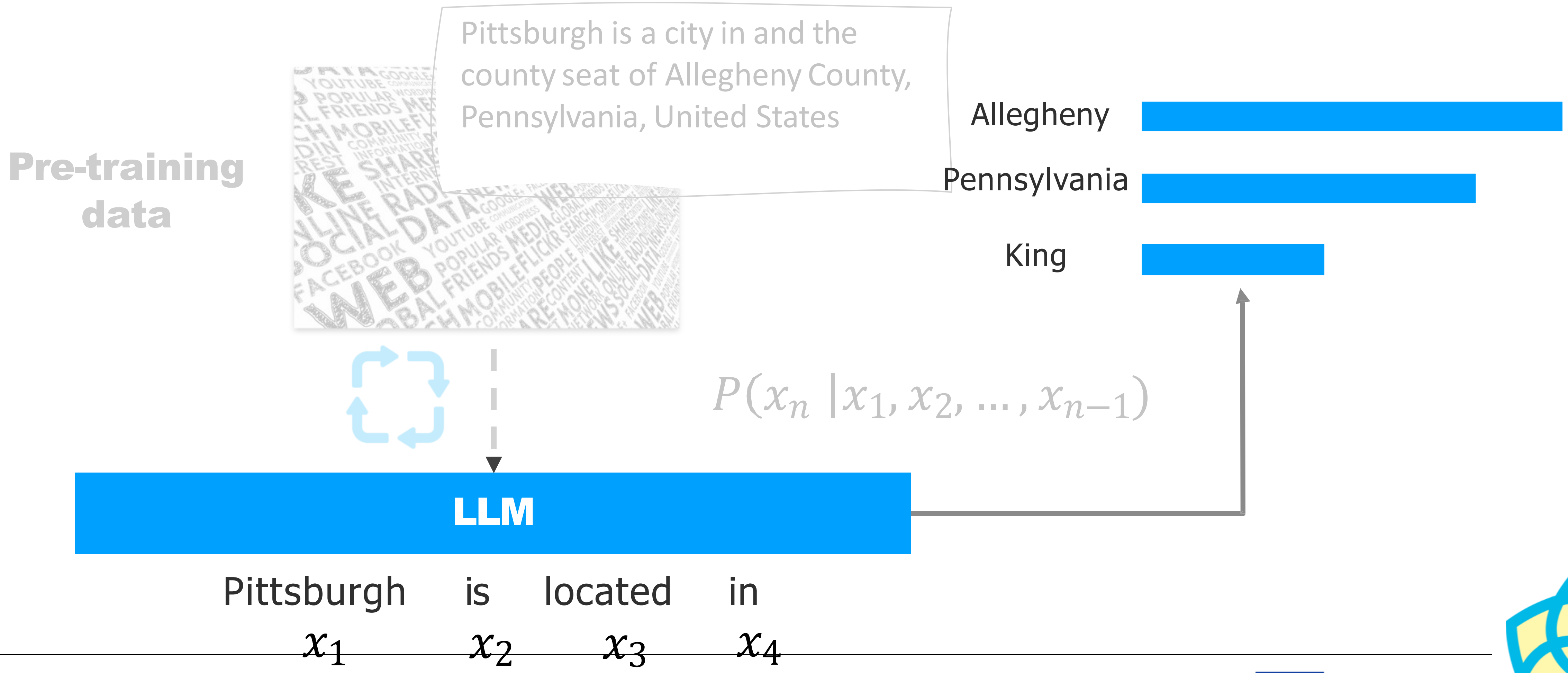
LLM

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

Funded by the European Union

How do normal parametric LLMs work?

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



Pre-training data

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

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

LLM

Pittsburgh is located in

x_1 x_2 x_3 x_4

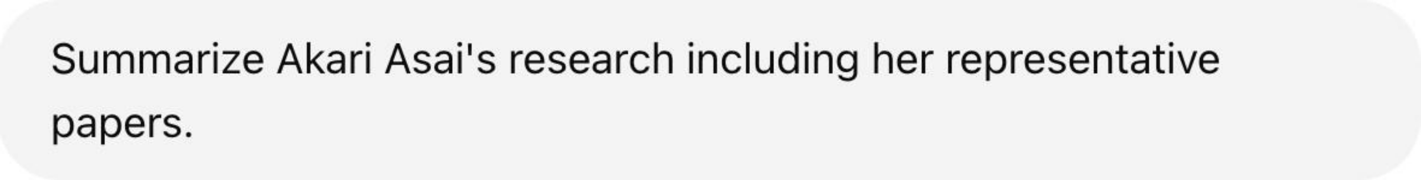
Allegheny

Pennsylvania

King

Limitations of parametric LLMs #1: Hallucinations

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

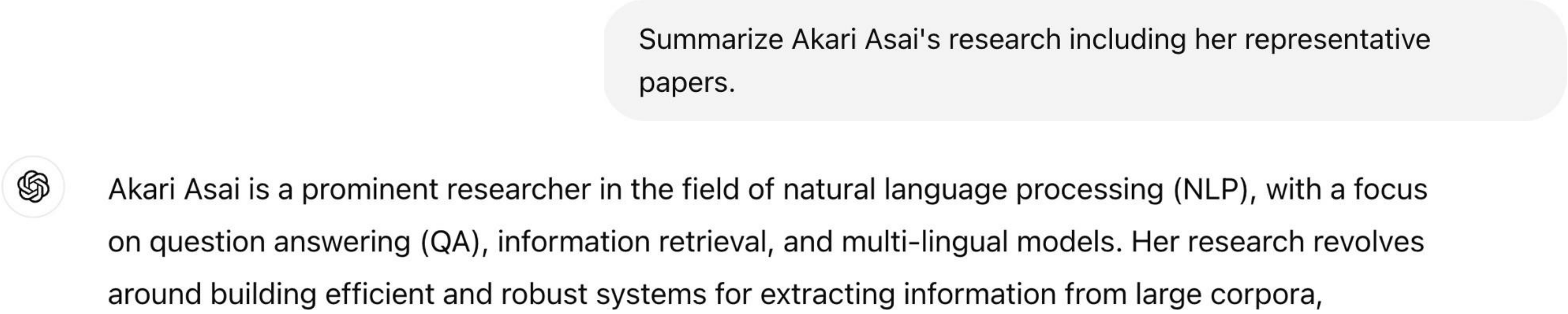


Summarize Akari Asai's research including her representative papers.

Funded by the European Union

Limitations of parametric LLMs #1: Hallucinations

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



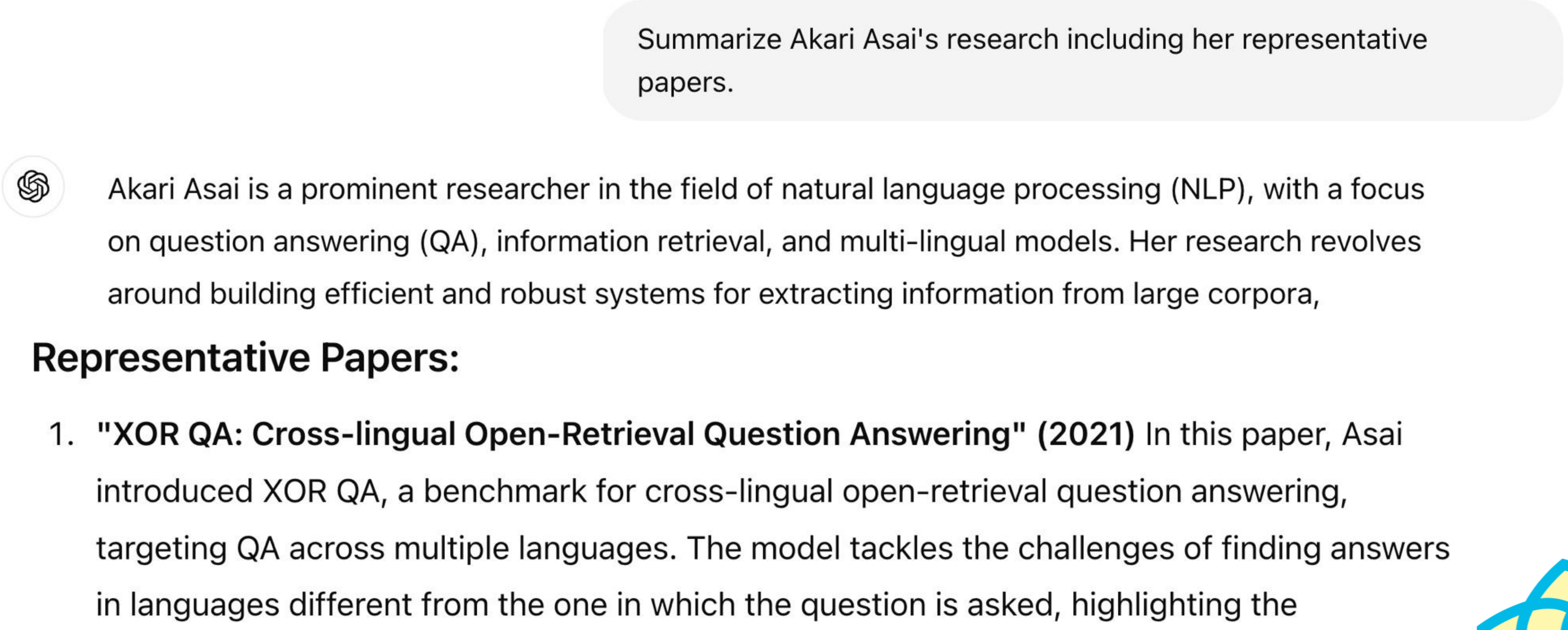
Summarize Akari Asai's research including her representative papers.

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,

Funded by the European Union

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,

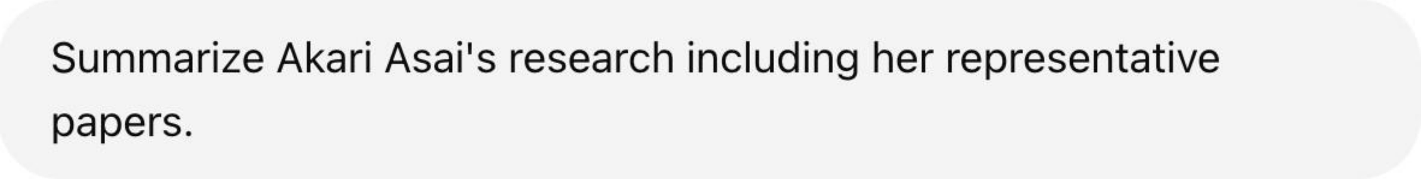
Representative Papers:

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

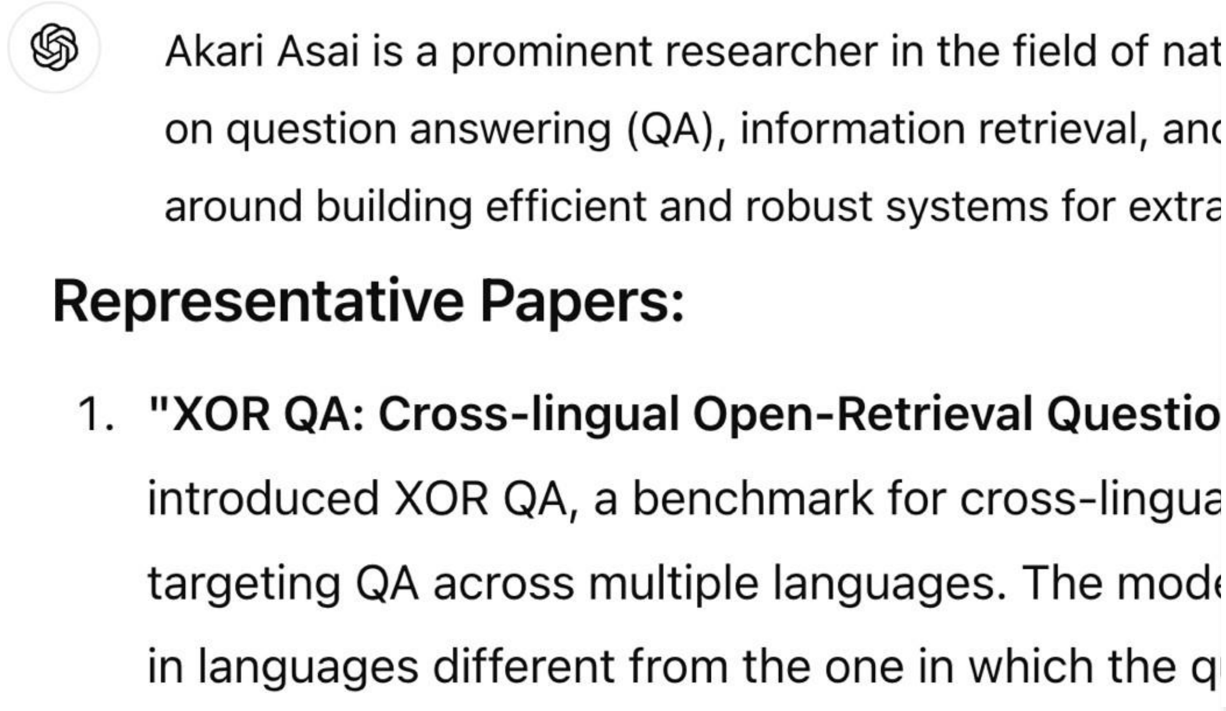
Funded by the European Union

Limitations of parametric LLMs #1: Hallucinations

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



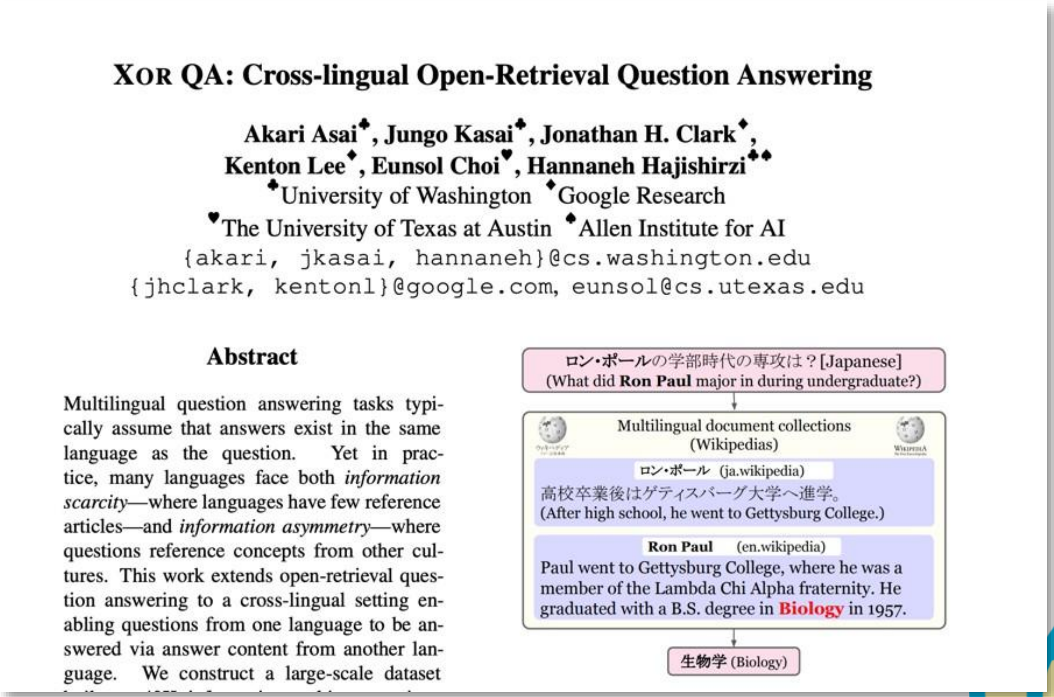
Summarize Akari Asai's research including her representative papers.



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

ロン・ポールの学部時代の専攻は？[Japanese]
(What did **Ron Paul** major in during undergraduate?)

Multilingual document collections
(Wikipedias)

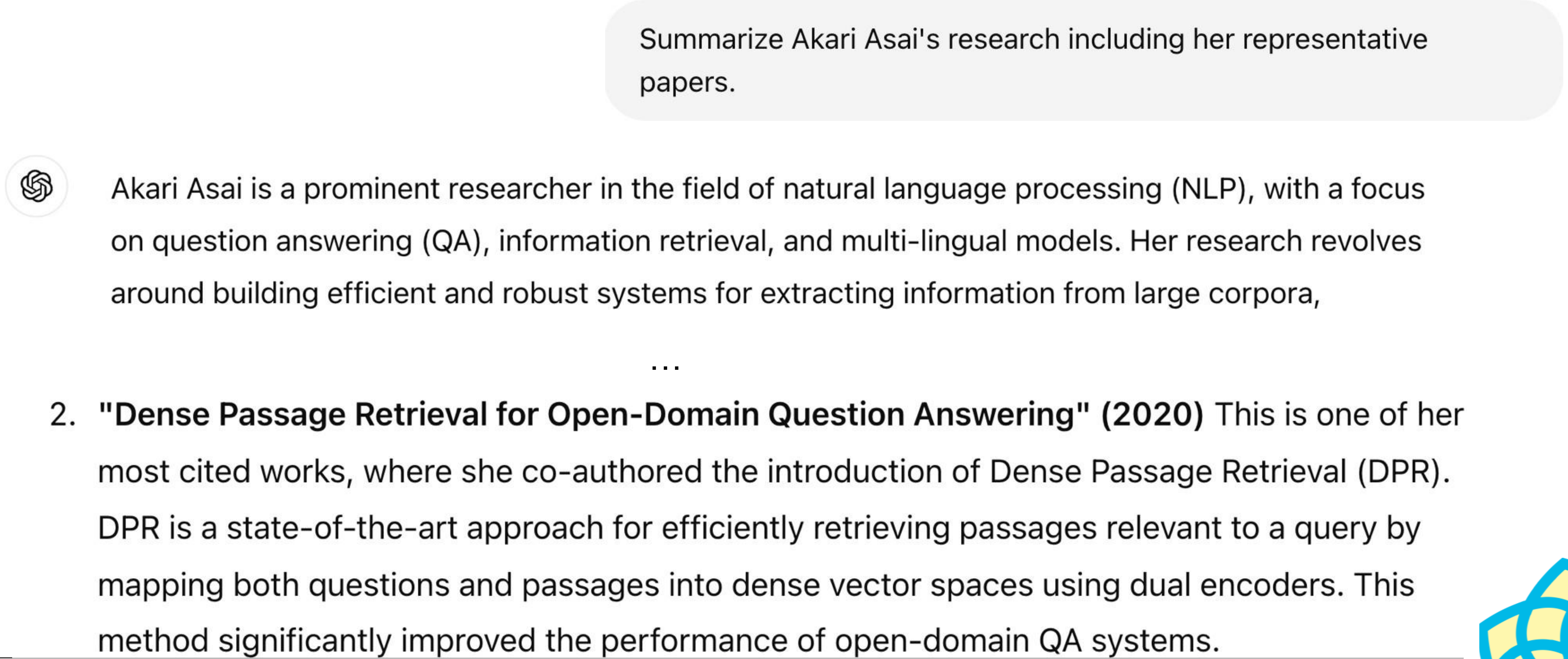
ロン・ポール (ja.wikipedia)
高校卒業後はゲティスバーグ大学へ進学。
(After high school, he went to Gettysburg College.)

Ron Paul (en.wikipedia)
Paul went to Gettysburg College, where he was a member of the Lambda Chi Alpha fraternity. He graduated with a B.S. degree in **Biology** in 1957.

生物学 (Biology)

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,

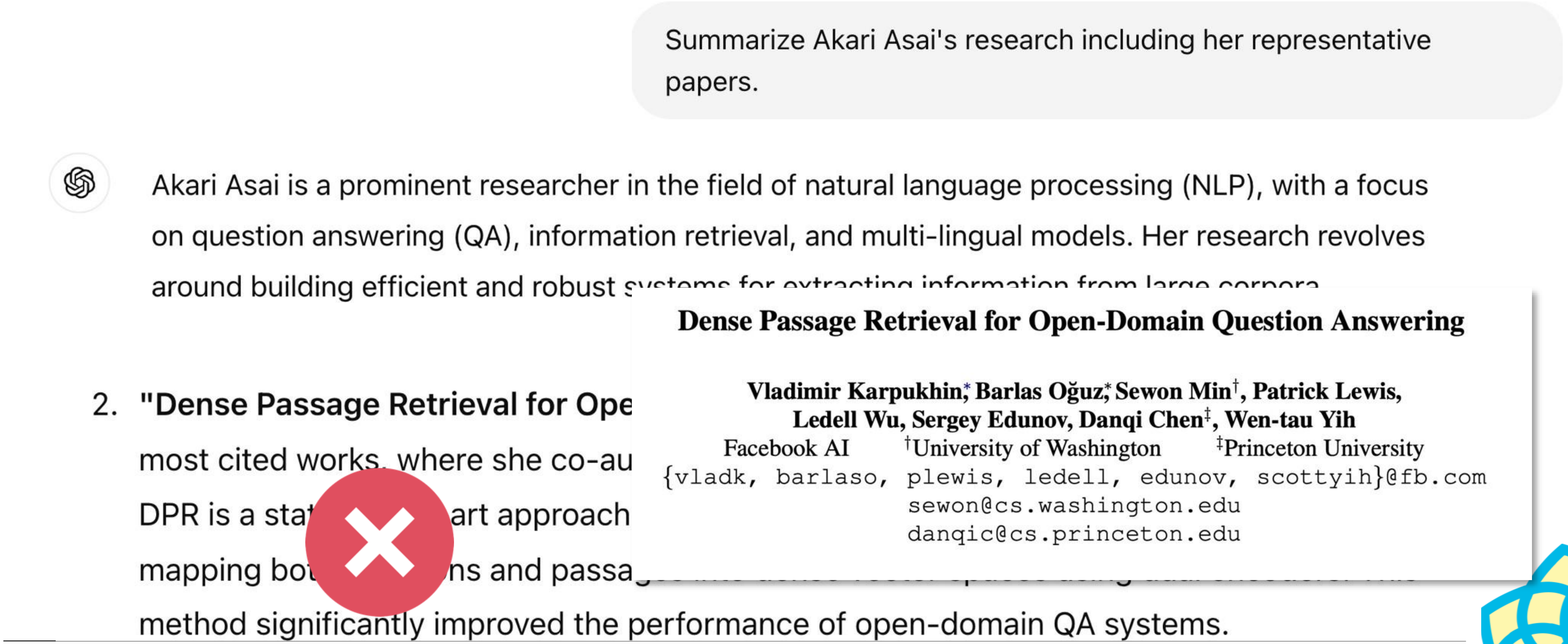
. . .

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.

Funded by the European Union

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.

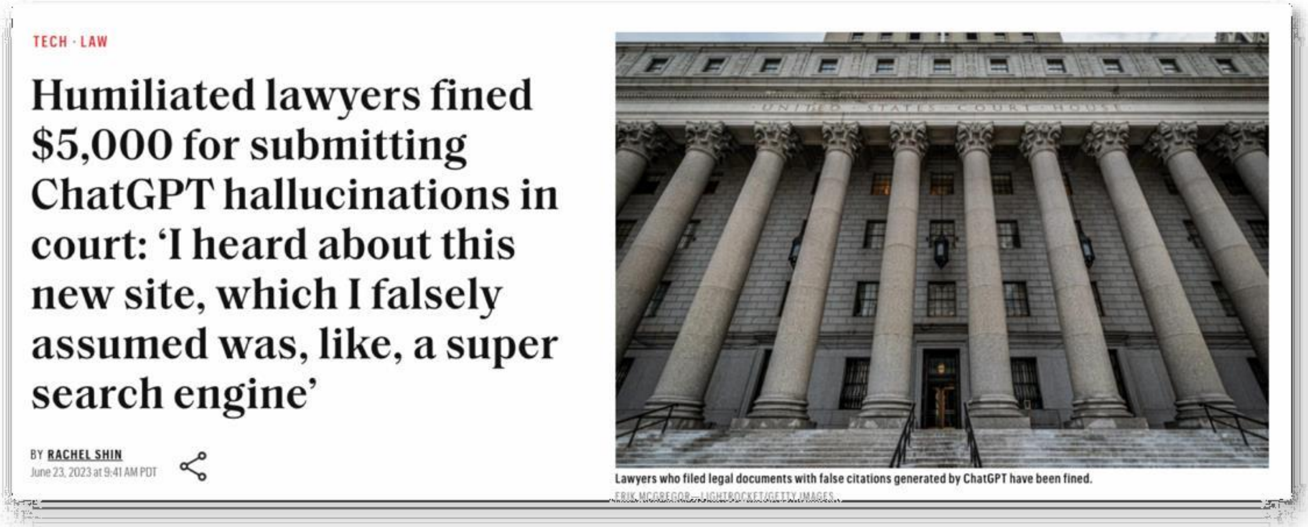
Dense Passage Retrieval for Open-Domain Question Answering

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

Funded by the European Union

Catastrophic incidents due to LLM hallucinations

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

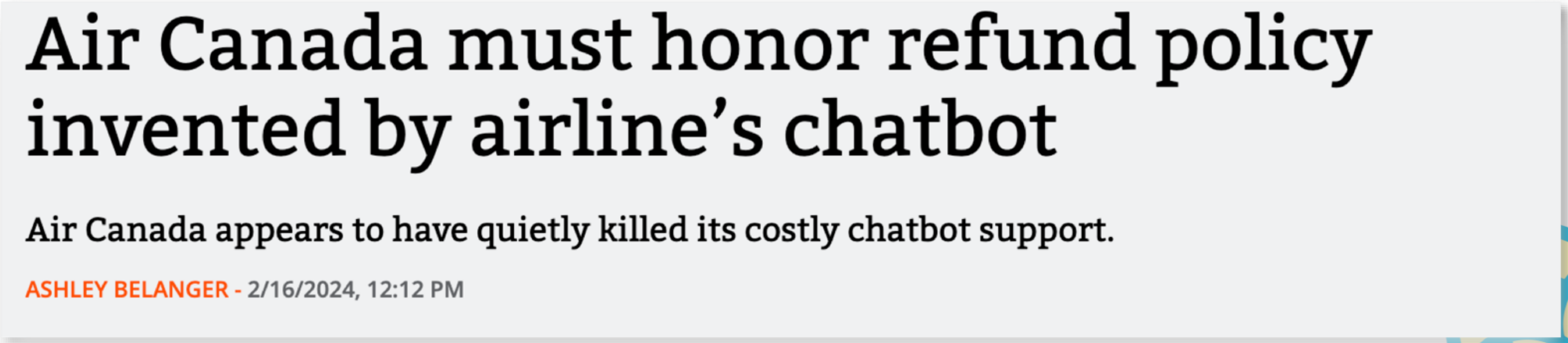


TECH · LAW

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



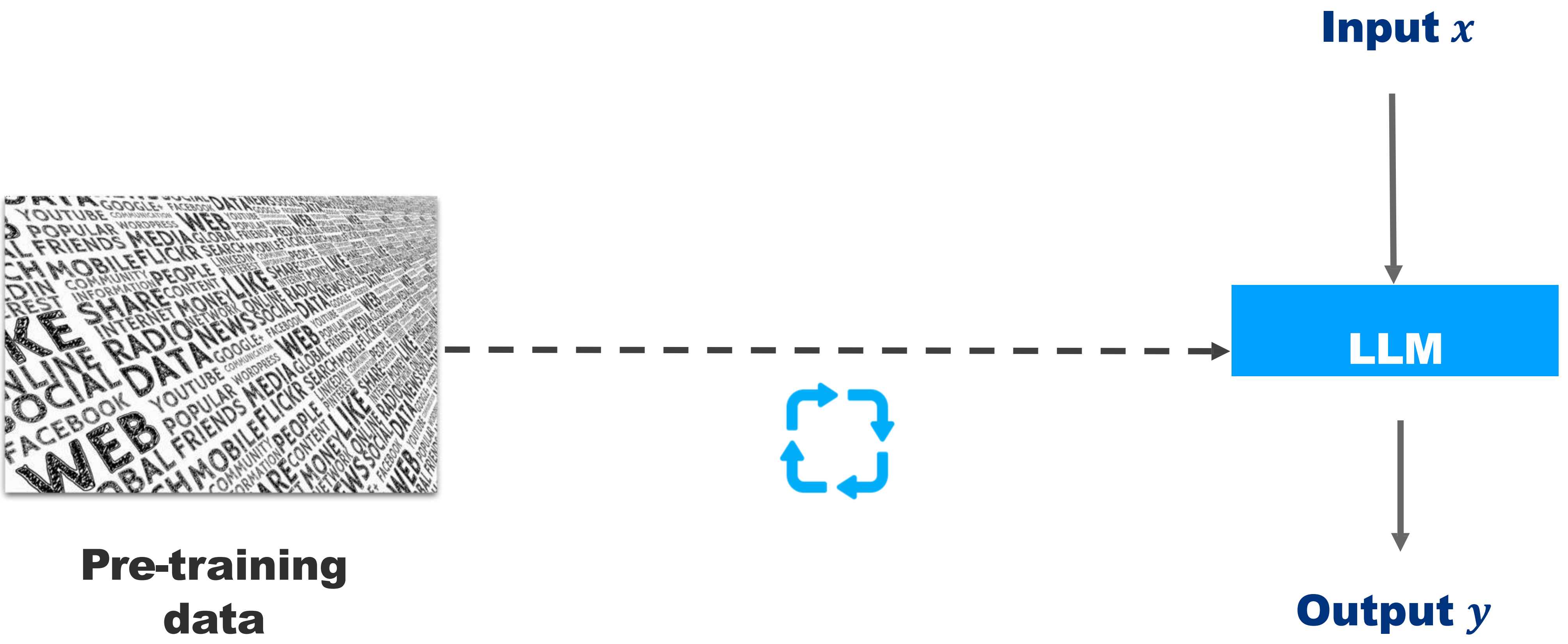
Air Canada must honor refund policy invented by airline's chatbot

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*



Funded by the European Union

Retrieval-augmented LMs: Definitions & Notations

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

Input x

Retriever

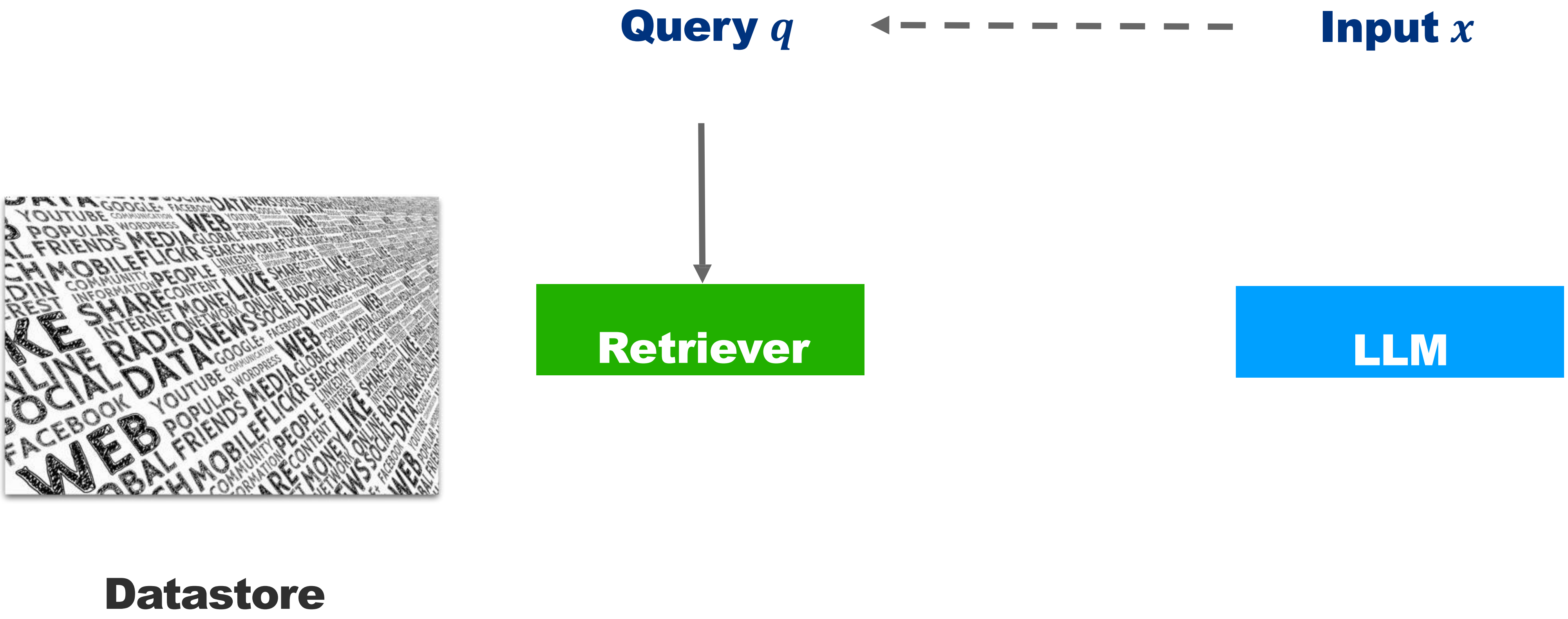
LLM

Datastore

Funded by the European Union

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

Retriever

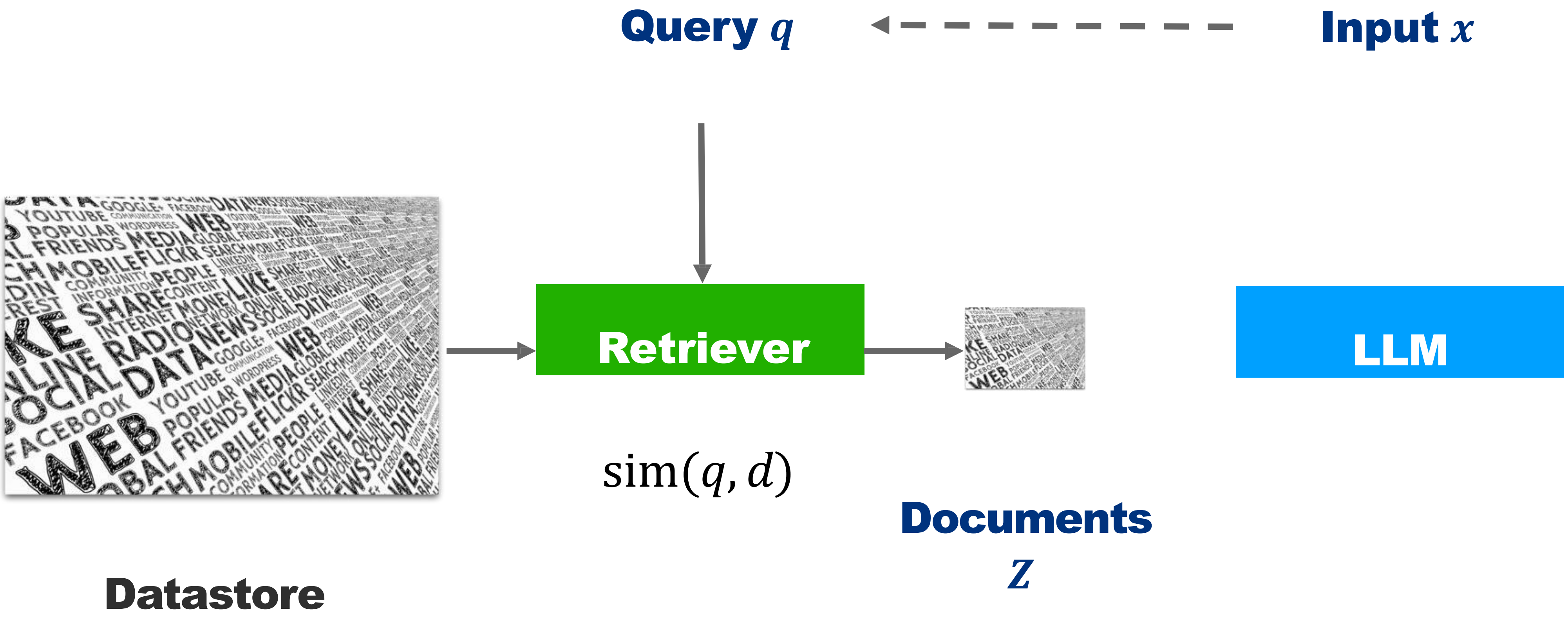
LLM

Datastore

Funded by the European Union

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

Retriever

LLM

$\mathrm{sim}(q, d)$

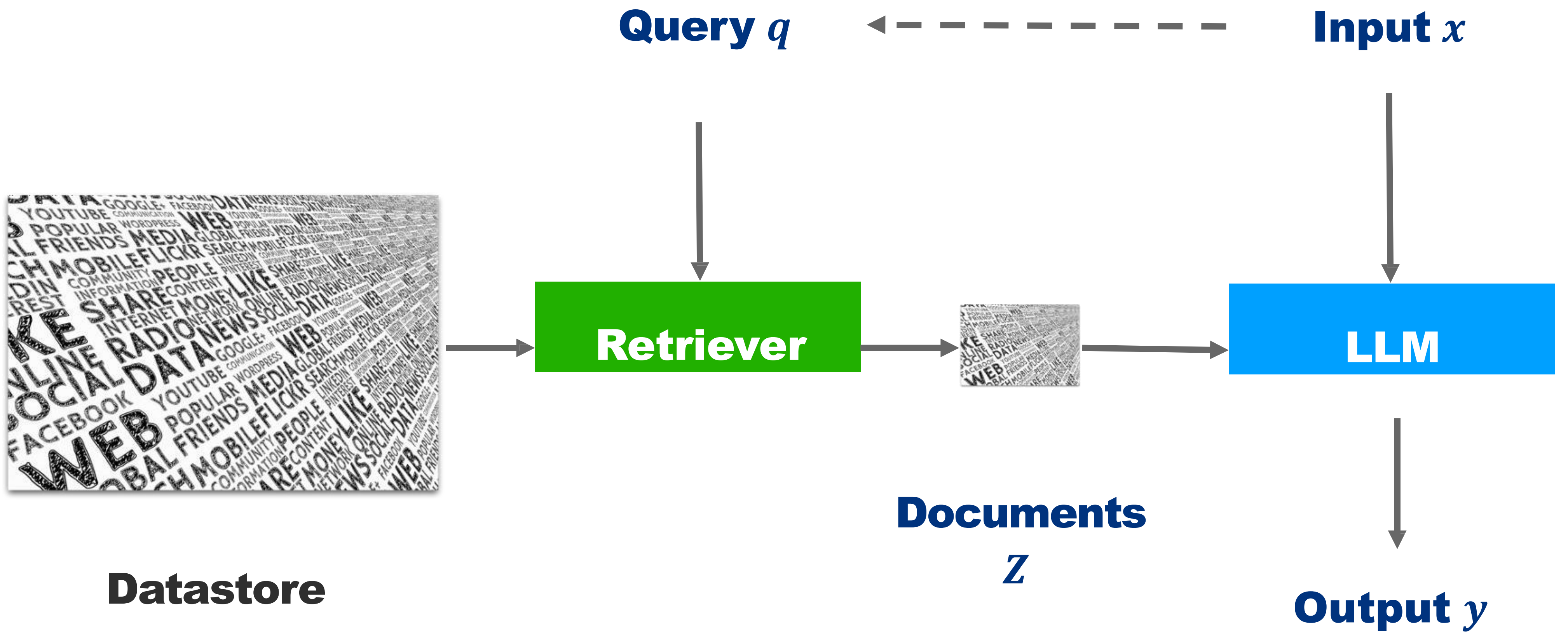
Documents

z

Datastore

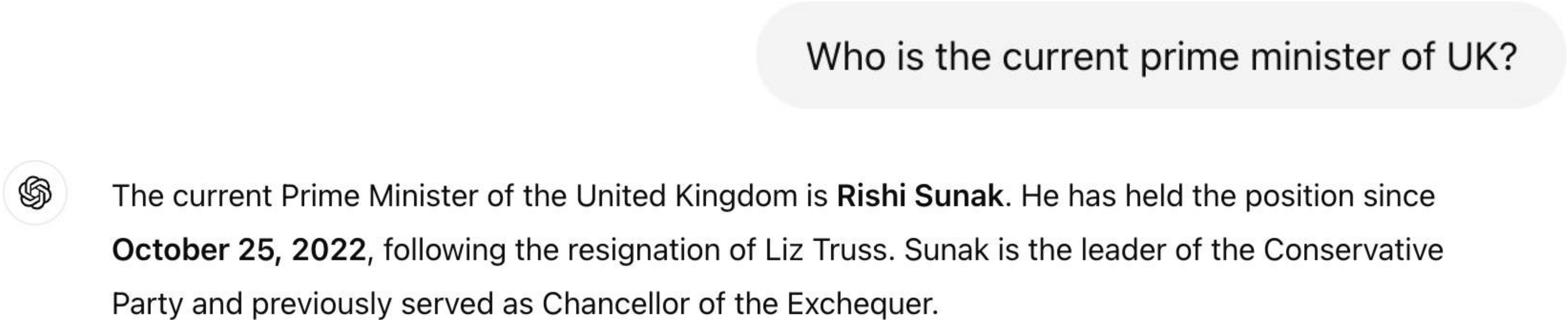
Retrieval-augmented LMs: Definitions & Notations

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



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

Parametric LMs' knowledge gets obsolete quickly & requires continuous training



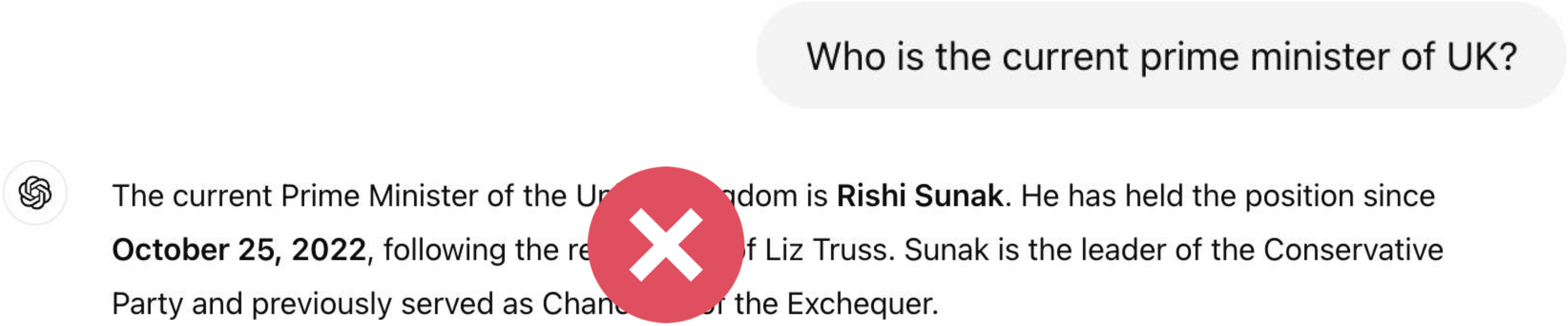
Who is the current prime minister of UK?

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.

Funded by
the European Union

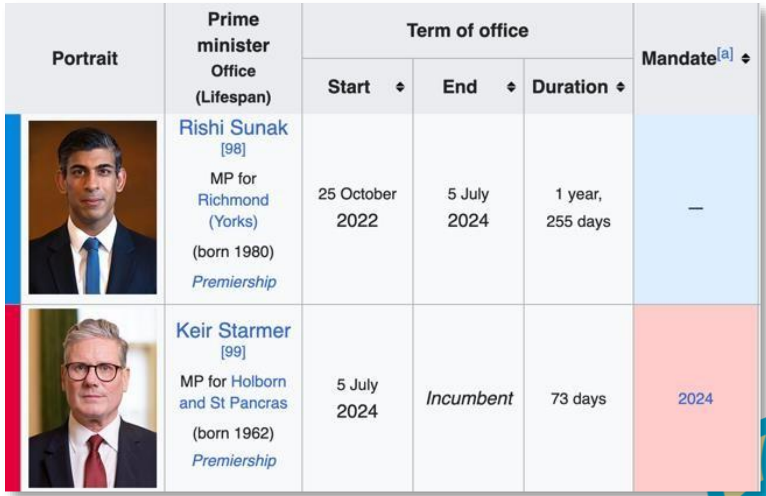
Benefit of retrieval-augmented LMs #2: Adaptations w/o training

Parametric LMs' knowledge gets obsolete quickly & requires continuous training



Who is the current prime minister of UK?

The current Prime Minister of the Ur ✕ 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.

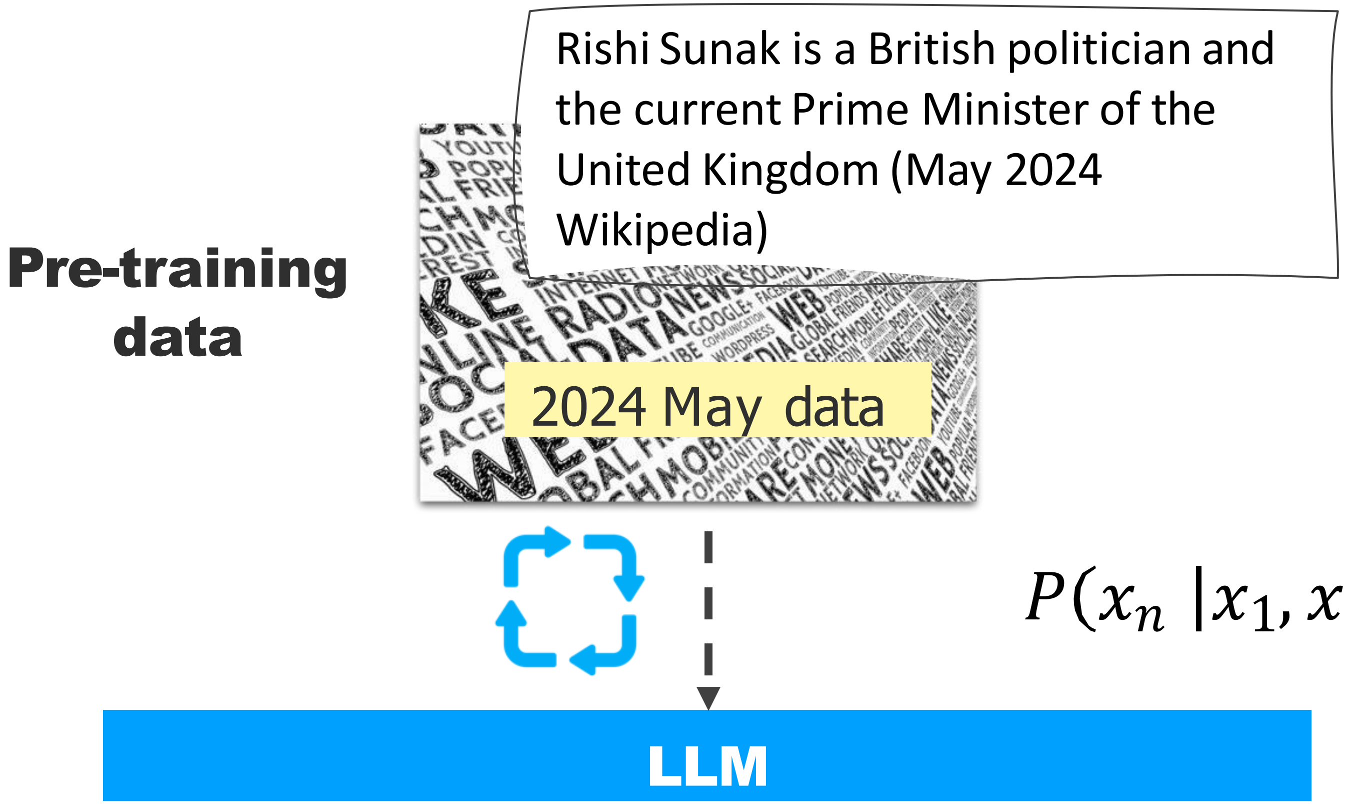


Portrait	Prime minister Office (Lifespan)	Term of office			Mandate[a]
		Start	End	Duration	
	Rishi Sunak [98] MP for Richmond (Yorks) (born 1980) *Premiership*	25 October 2022	5 July 2024	1 year, 255 days	–
	Keir Starmer [99] MP for Holborn and St Pancras (born 1962) *Premiership*	5 July 2024	*Incumbent*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

Funded by the European Union

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

Parametric LMs' knowledge gets obsolete quickly & requires continuous training



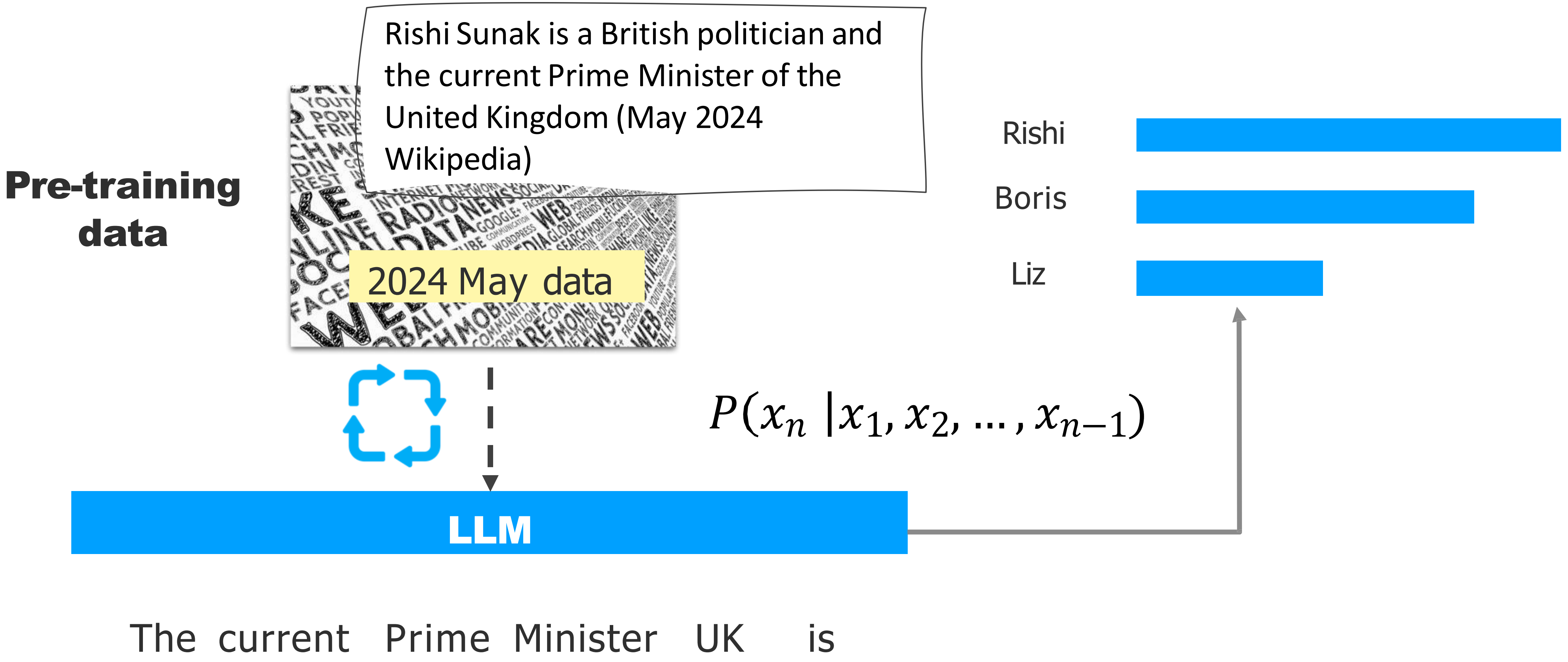
Pre-training data
Rishi Sunak is a British politician and the current Prime Minister of the United Kingdom (May 2024 Wikipedia)
2024 May data
LLM

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

Funded by the European Union

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

Parametric LMs' knowledge gets obsolete quickly & requires continuous training



Pre-training data

Rishi Sunak is a British politician and the current Prime Minister of the United Kingdom (May 2024 Wikipedia)

2024 May data

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

LLM

Rishi

Boris

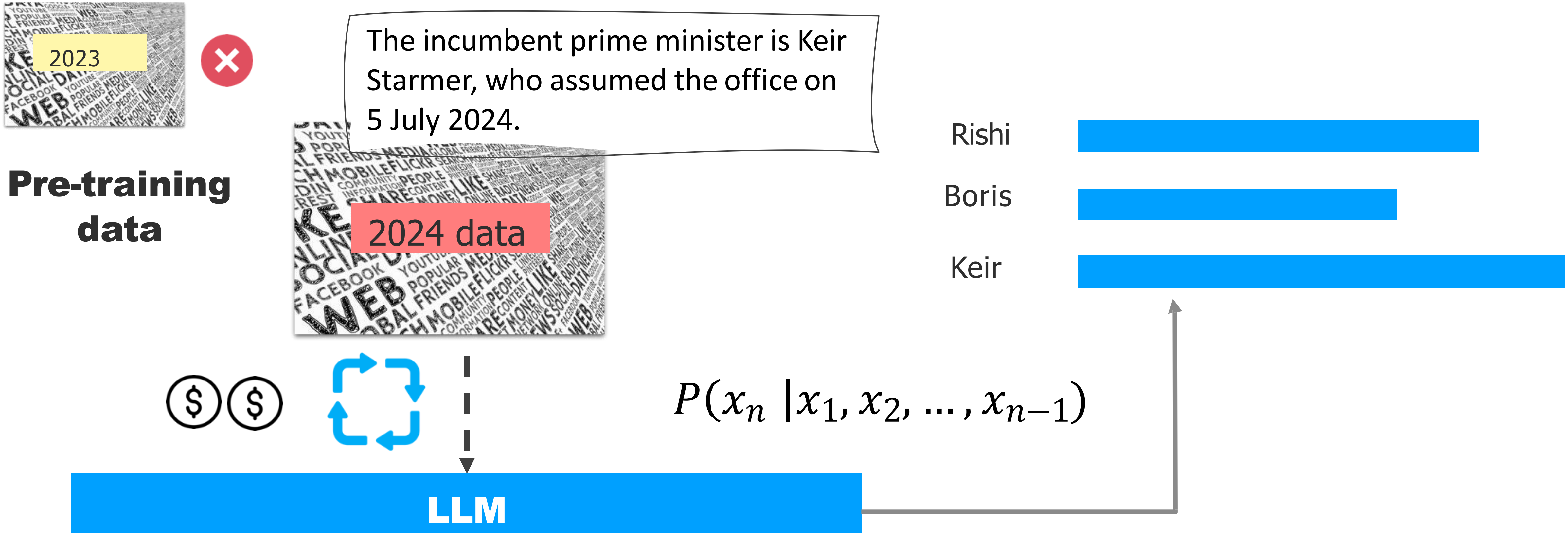
Liz

The current Prime Minister UK is

Funded by the European Union

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

Parametric LMs' knowledge gets obsolete quickly & requires continuous training



The incumbent prime minister is Keir Starmer, who assumed the office on 5 July 2024.

Pre-training data

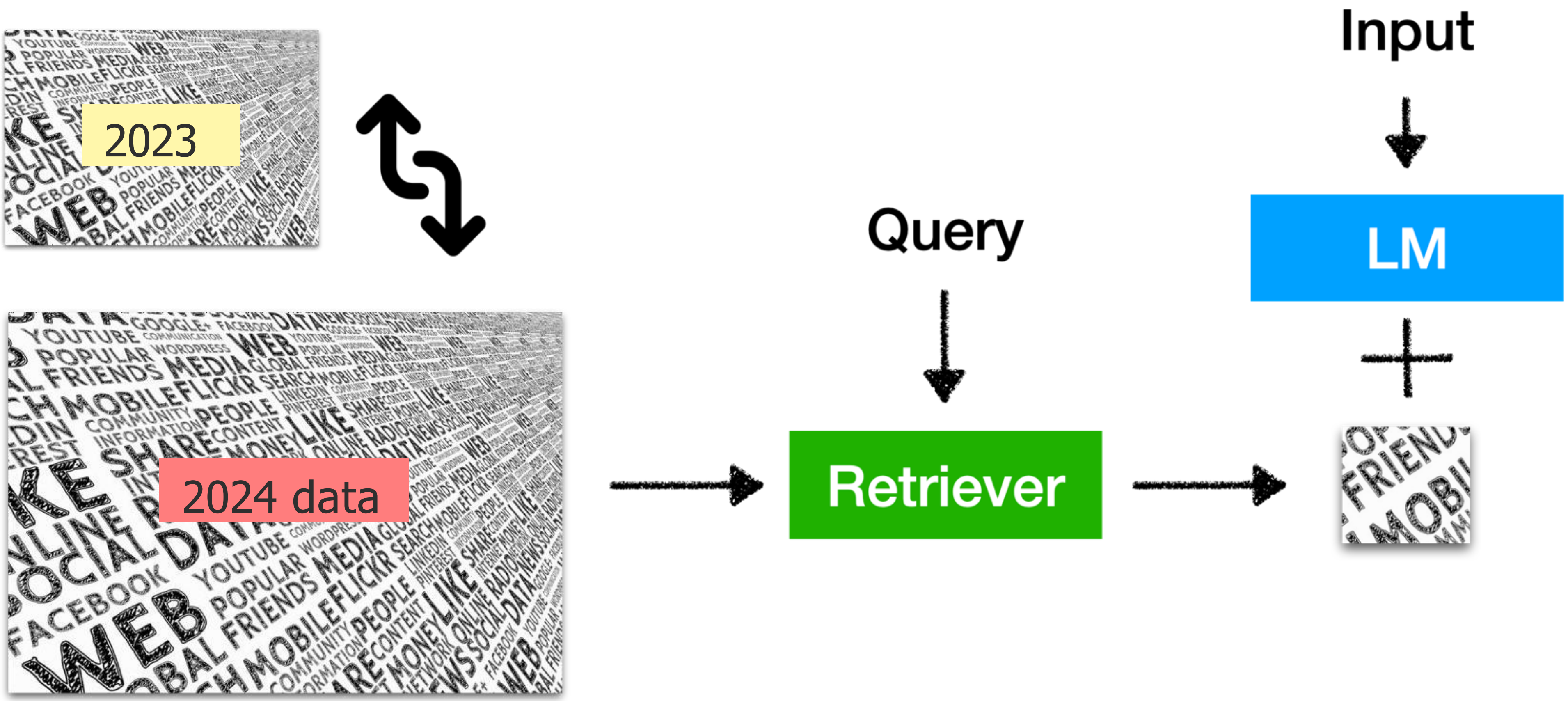
$$P(x_n \mid x_1, x_2, \dots, x_{n-1})$$

The current Prime Minister UK is

Funded by the European Union

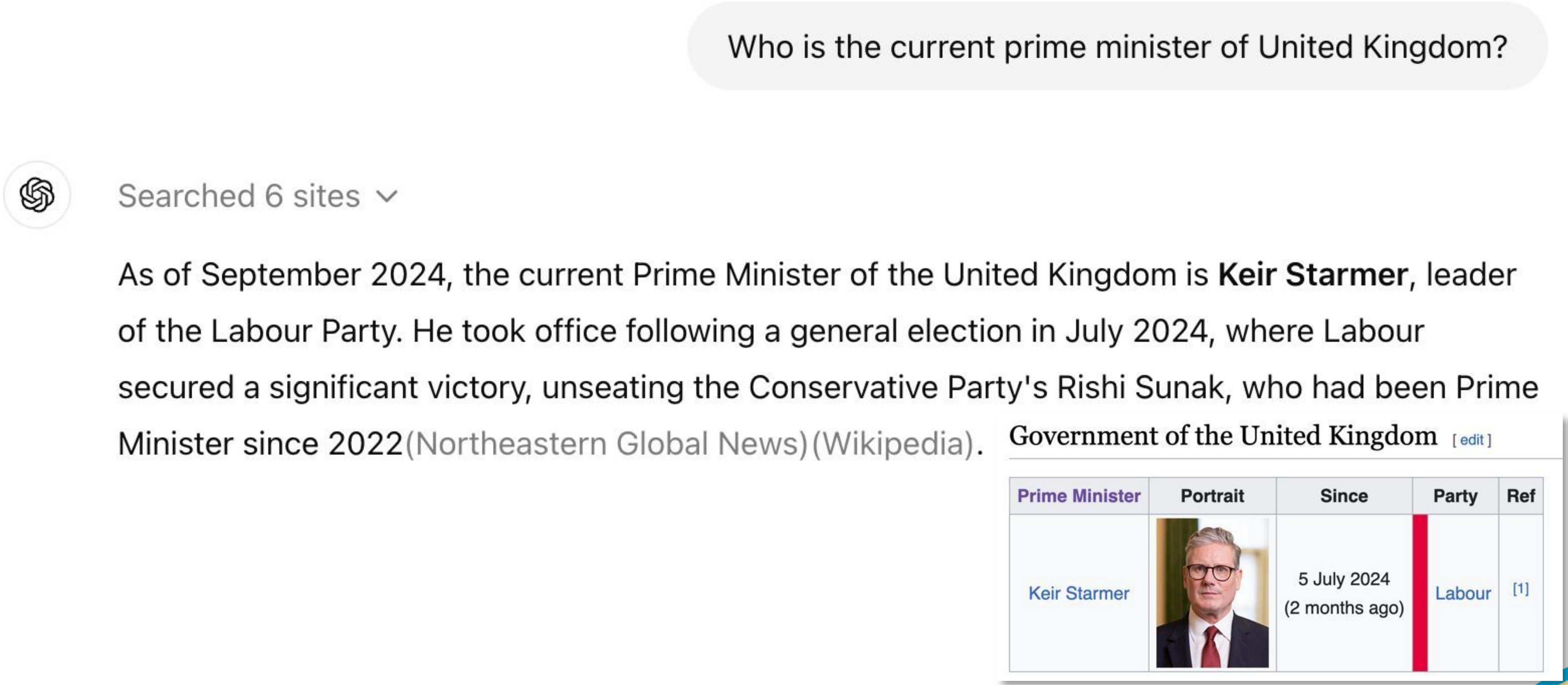
Benefit of retrieval-augmented LMs #2: Adaptations w/o training

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

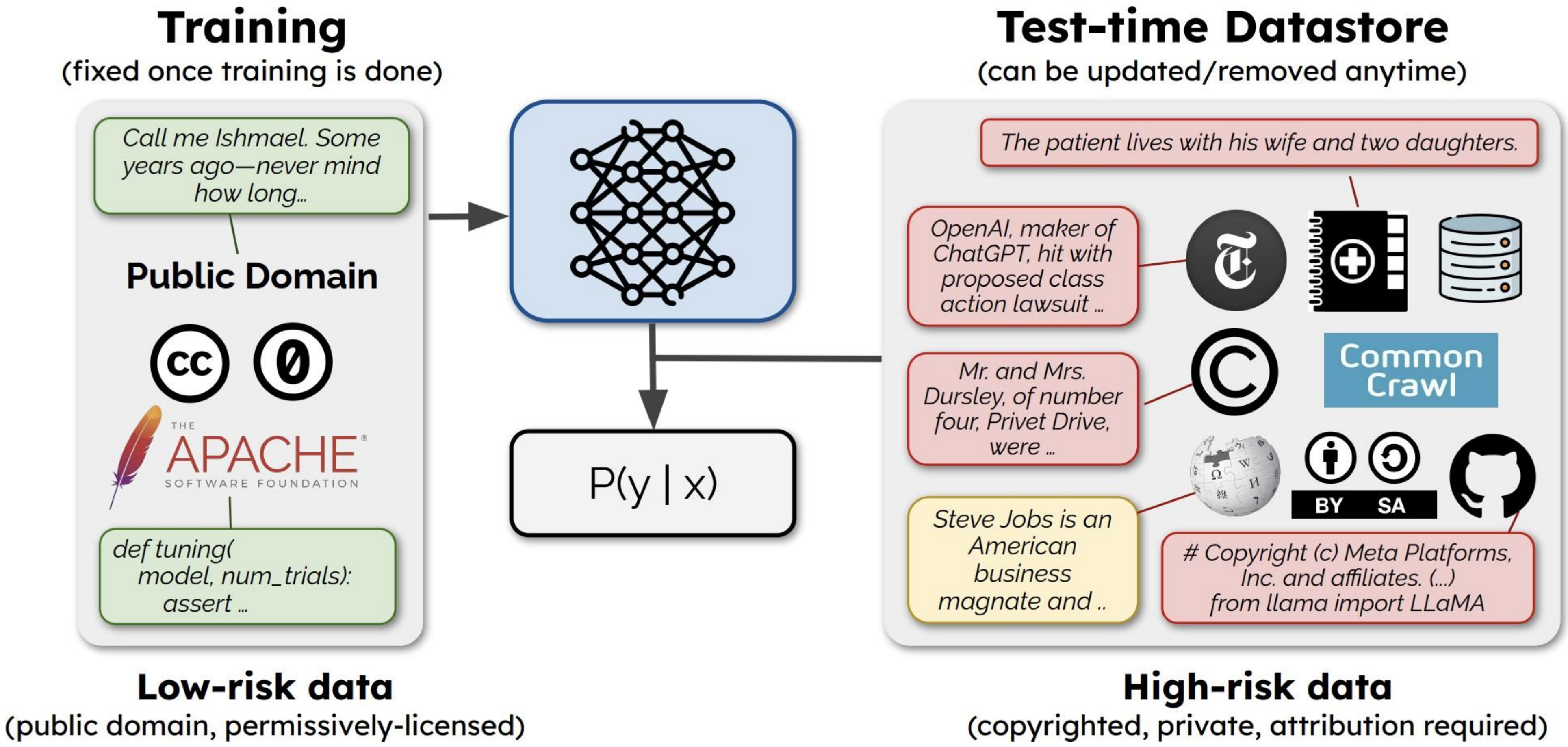
As of September 2024, the current Prime Minister of the United Kingdom is **Keir Starmer**, leader of the Labour Party. He took office following a general election in July 2024, where Labour secured a significant victory, unseating the Conservative Party's Rishi Sunak, who had been Prime Minister since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

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



Training
(fixed once training is done)

Call me Ishmael. Some years ago—never mind how long...

Public Domain

def tuning(model, num_trials): assert ...

Low-risk data
(public domain, permissively-licensed)

Test-time Datastore
(can be updated/removed anytime)

The patient lives with his wife and two daughters.

OpenAI, maker of ChatGPT, hit with proposed class action lawsuit ...

Mr. and Mrs. Dursley, of number four, Privet Drive, were ...

Steve Jobs is an American business magnate and ..

# Copyright (c) Meta Platforms, Inc. and affiliates. (...) from llama import LLaMA

High-risk data
(copyrighted, private, attribution required)

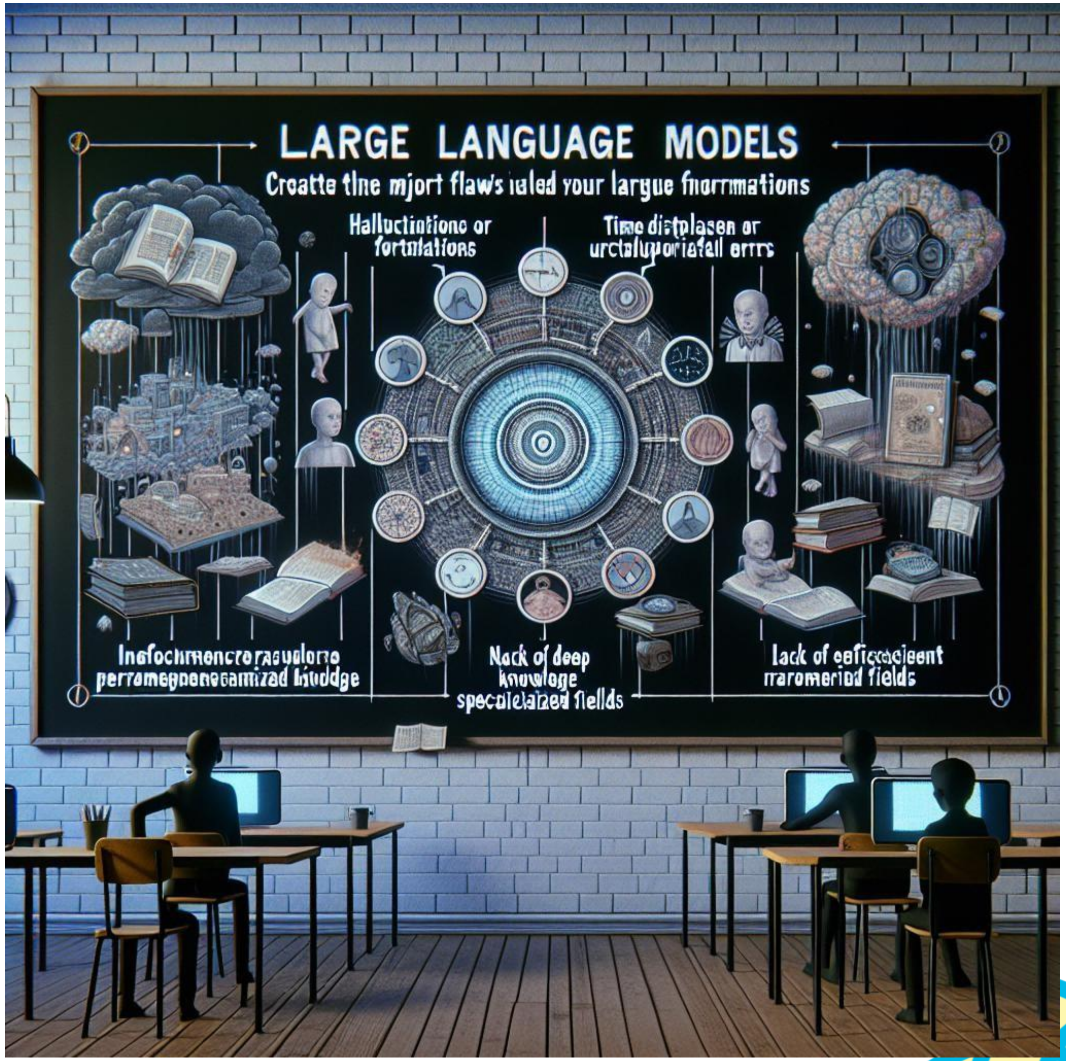
Funded by the European Union

Background

Drawbacks of LLMs

- Hallucination
- Outdated information
- Low efficiency in parameterizing knowledge
- Lack of in-depth knowledge in specialized domains
- Weak inferential capabilities

- Domain-specific accurate answering
- Frequent updates of data
- Traceability and explainability of generated content
- Controllable Cost
- Privacy protection of data


LARGE LANGUAGE MODELS
Creatte tlne mjort flaws ıaled your largue fnormations
Hallucintiono or fortnlations
Time distplasen or urctaluporiafall errs
Infochmencrorasuolore perramegnenecamizad knıdge
Nack of deep knowlege specıelaized fiellds
lack of eeficcacieent rraromeriad fields

Draw by DALL·E-3

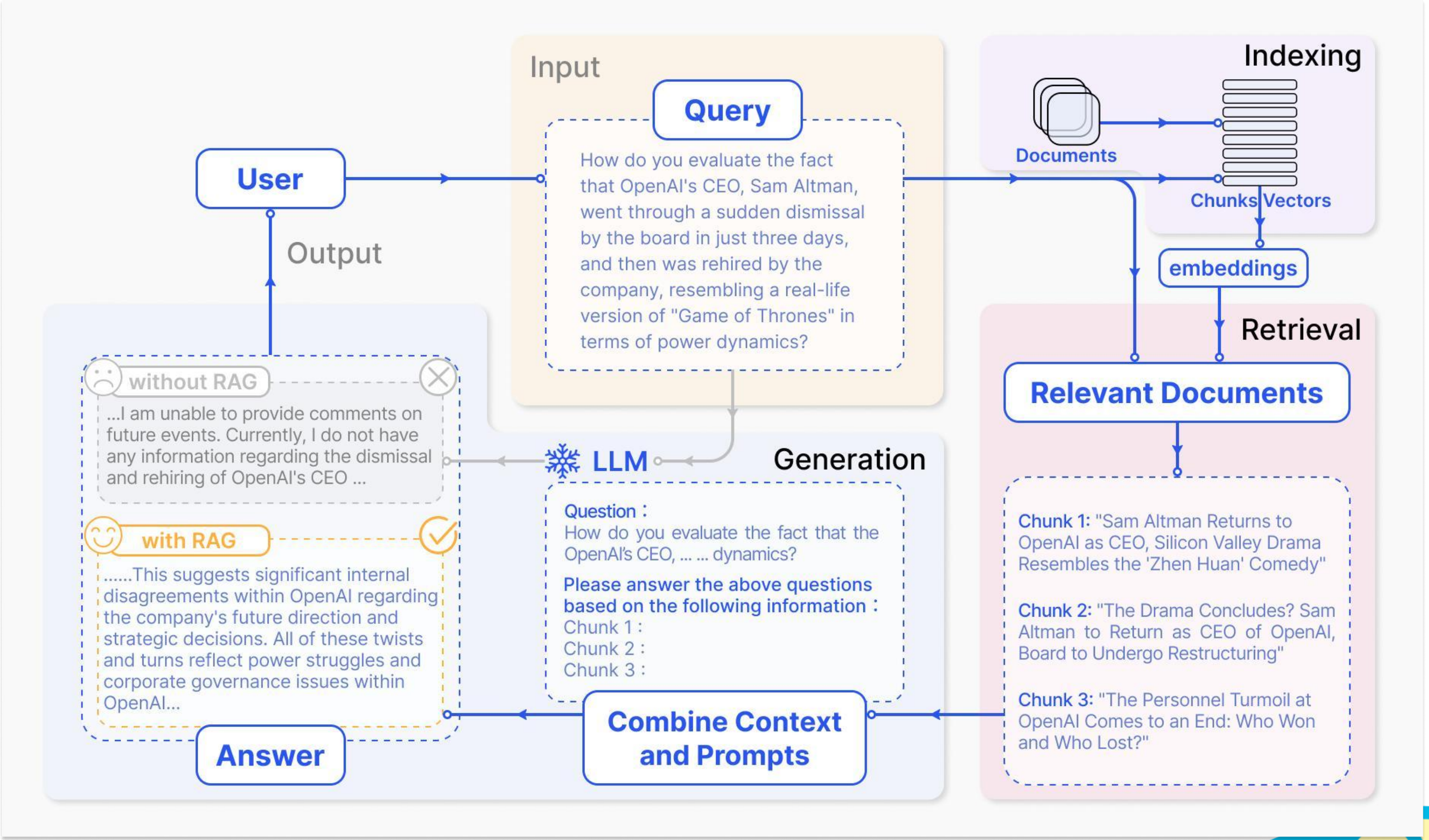
Funded by the European Union

Retrieval-Augmented Generation (RAG)

When answering questions or generating text, it first retrieves relevant information from a large number of documents, and then LLMs generates answers based on this information.

By attaching a external knowledge base, there is no need to retrain the entire large model for each specific task.

The RAG model is especially suitable for knowledge-intensive tasks.



Input

Query

How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

User

Output

Indexing

Documents

Chunks Vectors

embeddings

Retrieval

Relevant Documents

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

LLM

Generation

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?

Please answer the above questions based on the following information：
Chunk 1 :
Chunk 2 :
Chunk 3 :

Combine Context and Prompts

without RAG

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

with RAG

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Answer

A typical case of RAG

Funded by the European Union

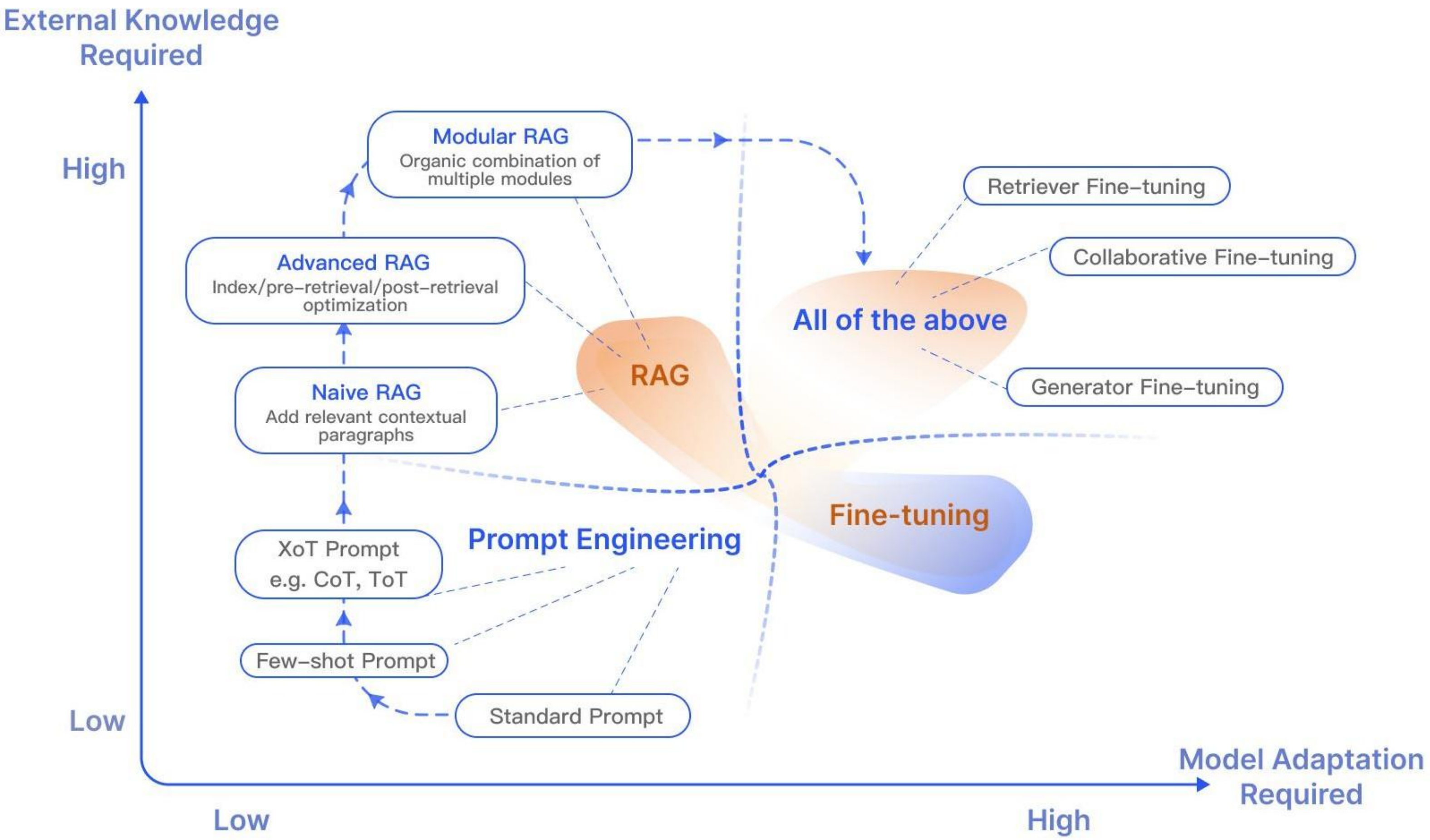
Symbolic Knowledge or Parametric Knowledge

Ways to optimize LLMs.

Prompt Engineering

Retrieval-Augmented Generation

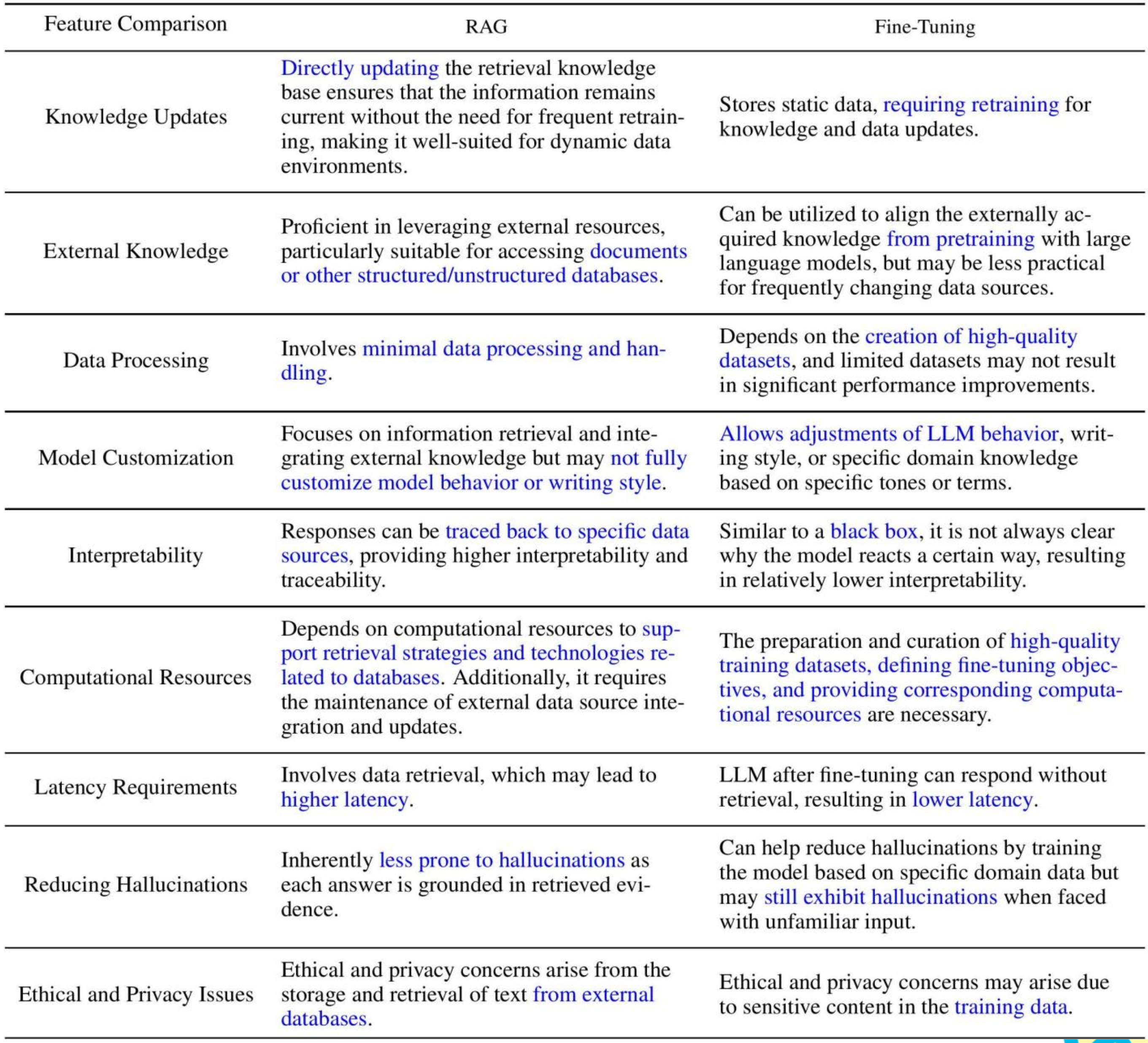
Instruct / Fine-tuning


External Knowledge Required
High
Low
Model Adaptation Required
Low
High
Modular RAG
Organic combination of multiple modules
Advanced RAG
Index/pre-retrieval/post-retrieval optimization
Naive RAG
Add relevant contextual paragraphs
RAG
XoT Prompt
e.g. CoT, ToT
Few-shot Prompt
Standard Prompt
Prompt Engineering
Retriever Fine-tuning
Collaborative Fine-tuning
All of the above
Generator Fine-tuning
Fine-tuning

A typical case of RAG

Funded by the European Union

RAG vs Fine-tuning



Feature Comparison	RAG	Fine-Tuning
Knowledge Updates	Directly updating the retrieval knowledge base ensures that the information remains current without the need for frequent retraining, making it well-suited for dynamic data environments.	Stores static data, requiring retraining for knowledge and data updates.
External Knowledge	Proficient in leveraging external resources, particularly suitable for accessing documents or other structured/unstructured databases.	Can be utilized to align the externally acquired knowledge from pretraining with large language models, but may be less practical for frequently changing data sources.
Data Processing	Involves minimal data processing and handling.	Depends on the creation of high-quality datasets, and limited datasets may not result in significant performance improvements.
Model Customization	Focuses on information retrieval and integrating external knowledge but may not fully customize model behavior or writing style.	Allows adjustments of LLM behavior, writing style, or specific domain knowledge based on specific tones or terms.
Interpretability	Responses can be traced back to specific data sources, providing higher interpretability and traceability.	Similar to a black box, it is not always clear why the model reacts a certain way, resulting in relatively lower interpretability.
Computational Resources	Depends on computational resources to support retrieval strategies and technologies related to databases. Additionally, it requires the maintenance of external data source integration and updates.	The preparation and curation of high-quality training datasets, defining fine-tuning objectives, and providing corresponding computational resources are necessary.
Latency Requirements	Involves data retrieval, which may lead to higher latency.	LLM after fine-tuning can respond without retrieval, resulting in lower latency.
Reducing Hallucinations	Inherently less prone to hallucinations as each answer is grounded in retrieved evidence.	Can help reduce hallucinations by training the model based on specific domain data but may still exhibit hallucinations when faced with unfamiliar input.
Ethical and Privacy Issues	Ethical and privacy concerns arise from the storage and retrieval of text from external databases.	Ethical and privacy concerns may arise due to sensitive content in the training data.

Funded by the European Union

RAG Applications

Scenarios where RAG is applicable:

- Long-tail distribution of data
- Frequent knowledge updates
- Answers requiring verification and traceability
- Specialized domain knowledge
- Data privacy preservation

Q&A
RETRO (Borgeaud et al2021)
REALM (Gu et al, 2020)
ATLAS (lzacard et al, 2023)

Fact Checking
RAG (Lewis et al, 2020)
ATLAS (lzacard et al, 2022)
Evi. Generator (Asai et al, 2022a)

Dialog
Blender Bot3 (Shuster et al.2022)
Internet-augmented generation
(Komeili et a., 2022)

Summary
FLARE (Jiang et al, 2023)

Machine Translation
kNN-MT (Khandelwal et al., 2020)TRIME-MT (Zhong et al., 2022)

Code Generation
DocPrompting (Zhou et al., 2023
Natural ProverWelleck et al., 2022)

Natural Language Inference
kNN-Prompt (Shi et al., 2022)
NPM (Min et al., 2023)

Sentiment analysis
kNN-Prompt (Shi et al., 2022)NPM (Min et al., 2023)

Commonsense reasoning
Raco (Yu et al, 2022)

Funded by the European Union

Naive RAG

Step1 Indexing

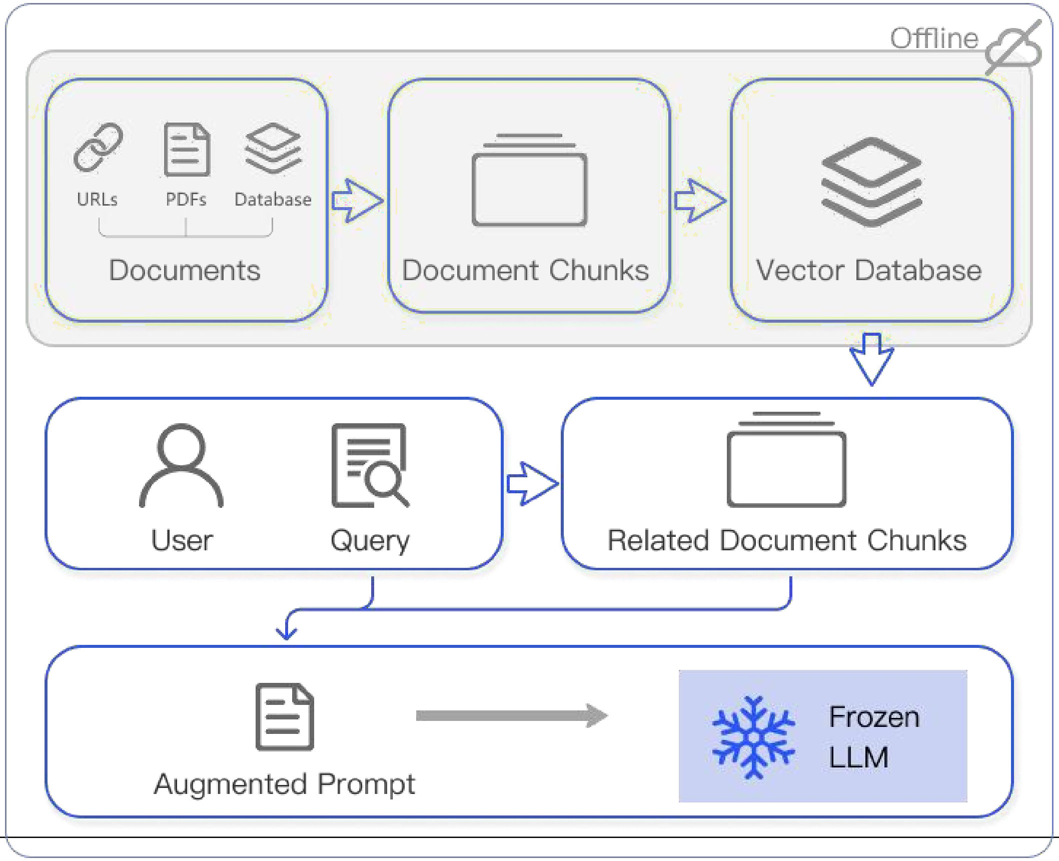
1. Divide the document into even chunks, each chunk being a piece of the original text.
2. Using the encoding model to generate an embedding for each chunck.
3. Store the Embedding of each block in the vector database.

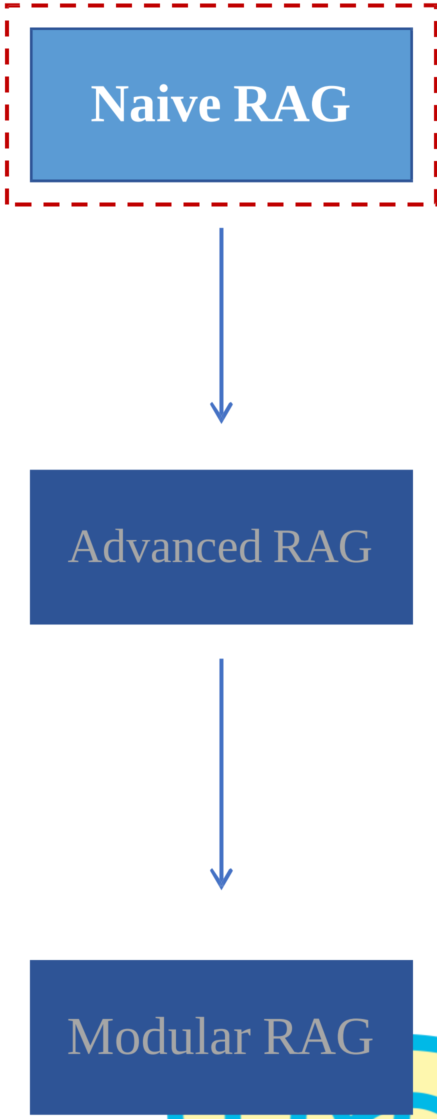
Step2 Retrival

Retrieve the k most relevant documents using vector similarity search.

Step3 Generation

The original query and the retrieved text are combined and input into a LLM to get the final answer



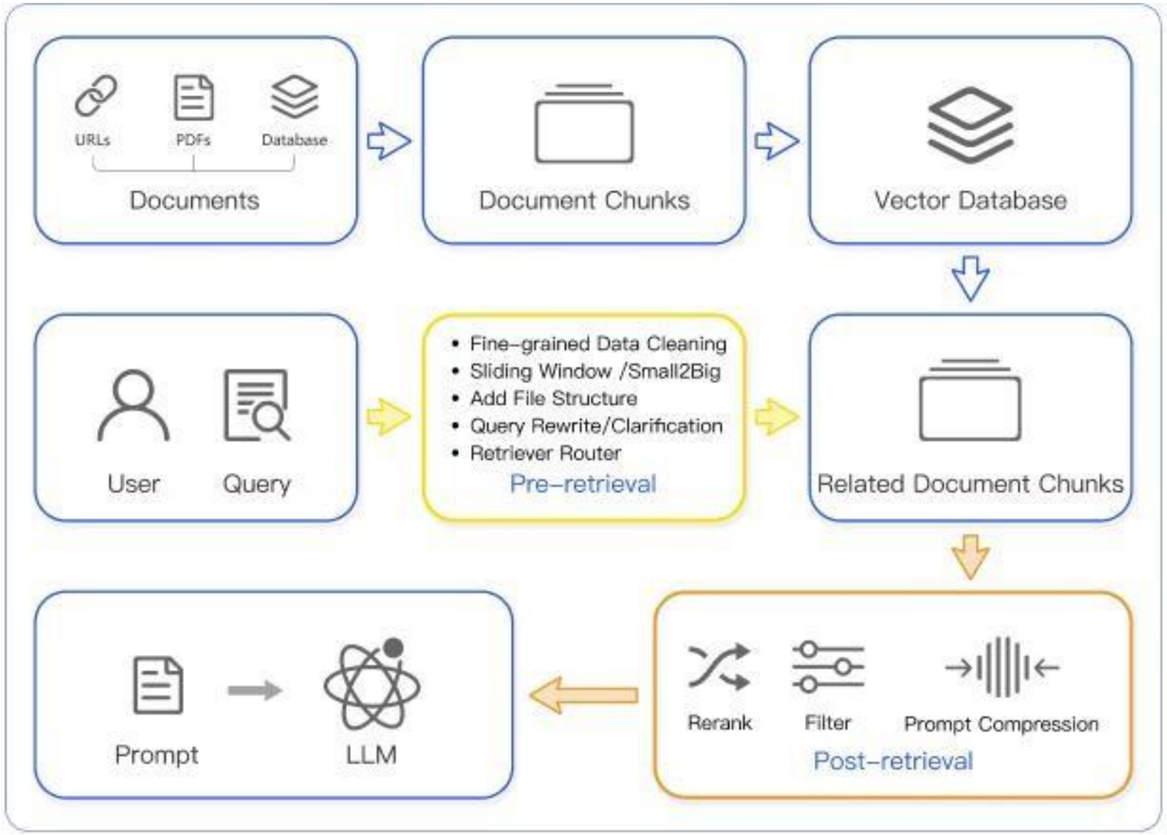


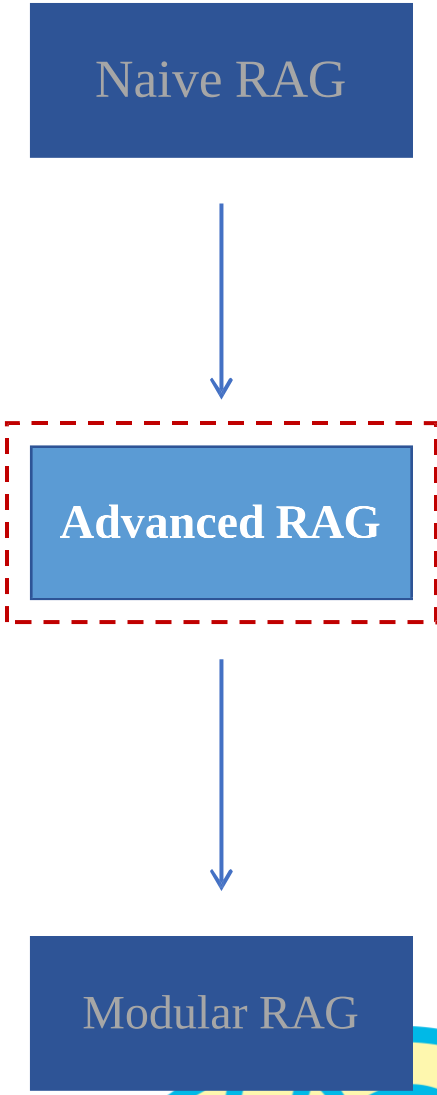
TrustLLM
Funded by the European Union

Advanced RAG

Index Optimization → Pre-Retrieval Process → Retrieval → Post-Retrieval Process→ Generation

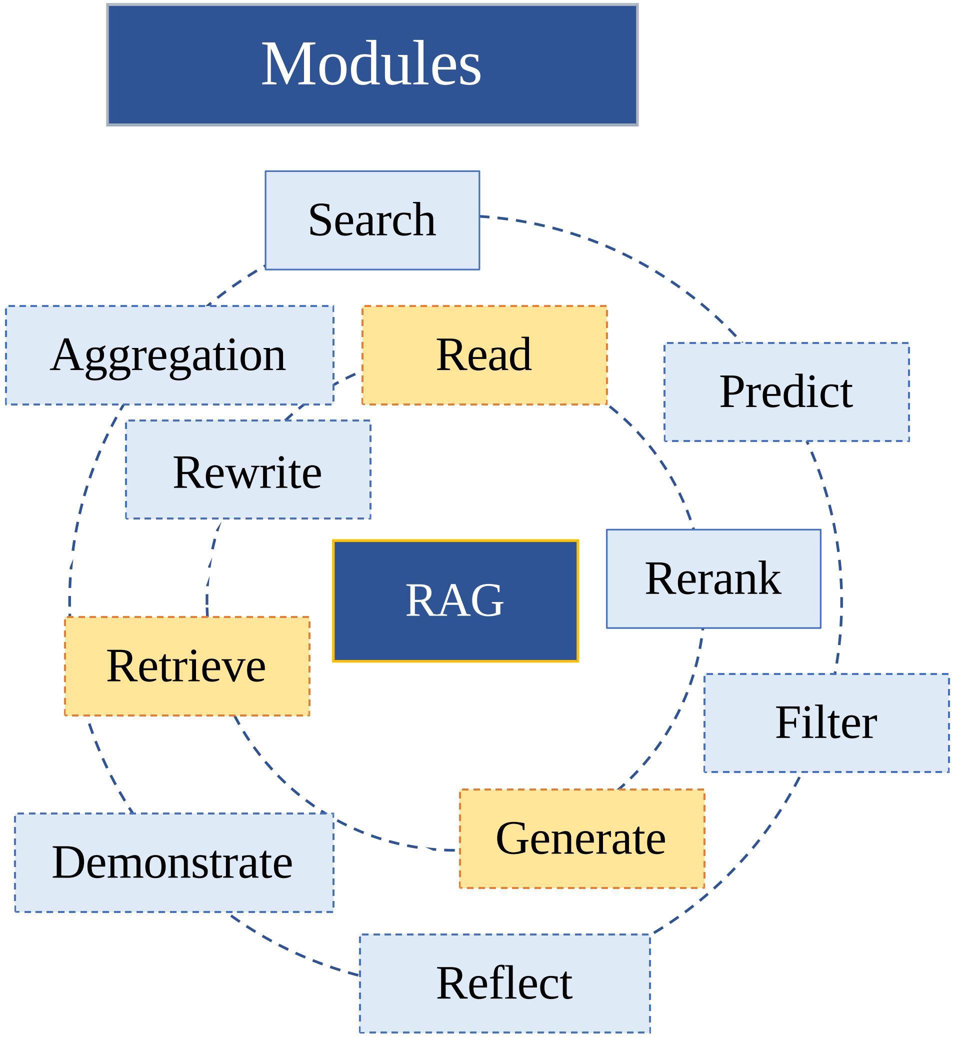
- **Optimizing Data Indexing:** sliding window, fine-grained segmentation、adding metadata
- **Pre-Retrieval Process：** retrieve routes, summaries, rewriting, and confidence judgment
- **Post-Retrieval Process：** reorder, filter content retrieval



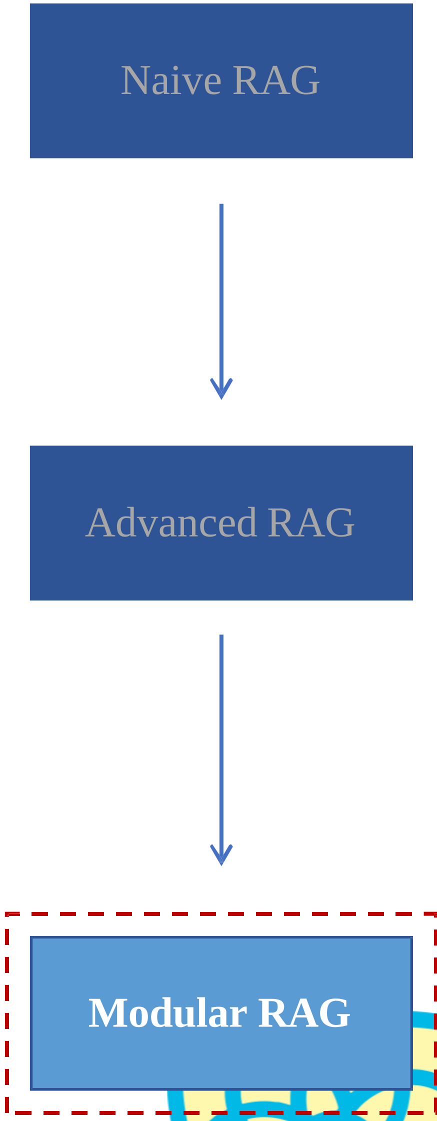


Funded by the European Union

Modular RAG

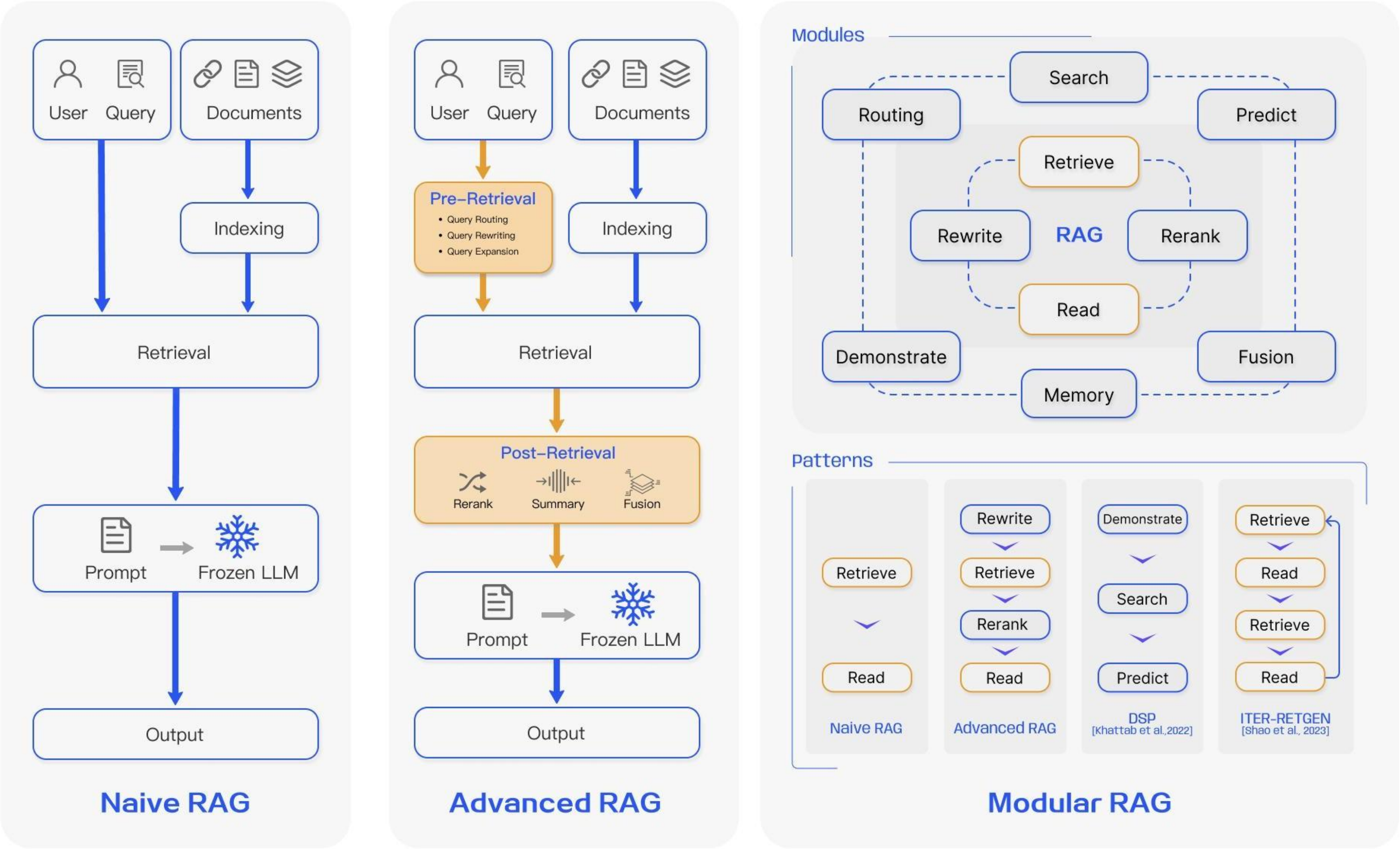


	Pattern			
Naive RAG	Read	Retrieve	Generate	
DSP (2022)	Demonstrate	Search	Predict	Generate
Rewrite-Retrieve-Read (2023)	Rewrite	Retrieve	Read	
Retrieve-then-read (2023)	Retrieve	Read	Generate	



Funded by the European Union

Comparison of RAG Paradigms



Funded by the European Union

The three key questions of RAG

What to retrieve ?

- Token
- Phrase
- Chunk
- Paragraph
- Entity
- Knowledge graph

When to retrive ?

- Single search
- Each token
- Every N tokens
- Adaptive search

How to use the retrieved information ?

- Input/Data Layer
- Model/Intermediate Layer
- Output/Prediction Layer

Other Issues

Augmentation stage:

- Pre-training
- Fine-tuning
- Inference

Retrieval choice:

- BERT
- Roberta
- BGE
- ......

Model Collaboration

⇄

Scale selectionz

Generation choice:

- GPT
- Llama
- T5
- ......

Funded by the European Union

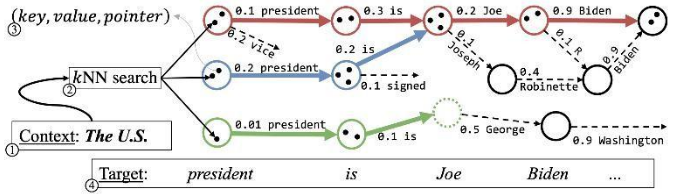
Key issue of RAG — What to retrieve

coarse

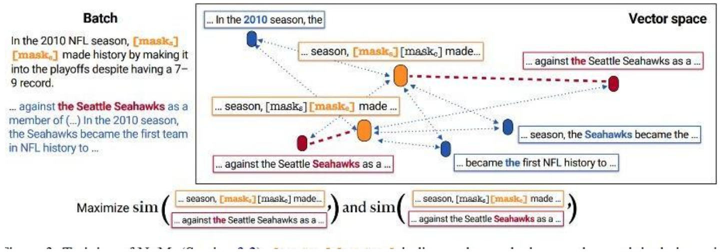
Retrieval granularity

meticulous

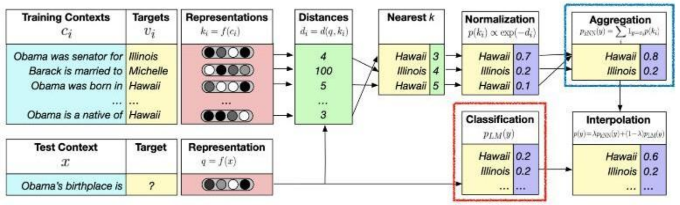
Chunk | In-Context RAG 2023



Phrase | NPM 2023


Batch
In the 2010 NFL season, [mask_s] [mask_e] made history by making it into the playoffs despite having a 7–9 record.
... against the Seattle Seahawks as a member of (...) In the 2010 season, the Seahawks became the first team in NFL history to ...
Vector space
... In the 2010 season, the
... season, [mask_s] [mask_e] made...
... against the Seattle Seahawks as a ...
... season, [mask_s] [mask_e] made ...
... season, the Seahawks became the ...
... against the Seattle Seahawks as a ...
... became the first NFL history to ...
Maximize sim(... season, [mask_s] [mask_e] made... , ... against the Seattle Seahawks as a ...) and sim(... season, [mask_s] [mask_e] made ... , ... against the Seattle Seahawks as a ...)

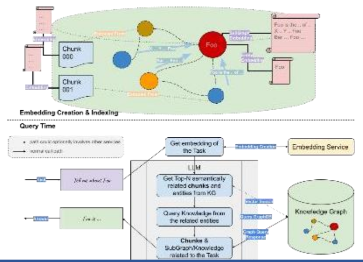
Token | KNN-LMM 2019



The search is broad, recalling a large amount of information, but with low accuracy, high coverage but includes much redundant information.

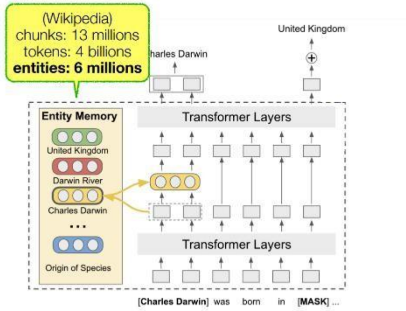
It excels in handling long-tail and cross-domain issues with high computational efficiency, but it requires significant storage.

Knowledge Graph | 2023



Richer semantic and structured information, but the retrieval efficiency is lower and is limited by the quality of KG.

Entity | EasE 2022

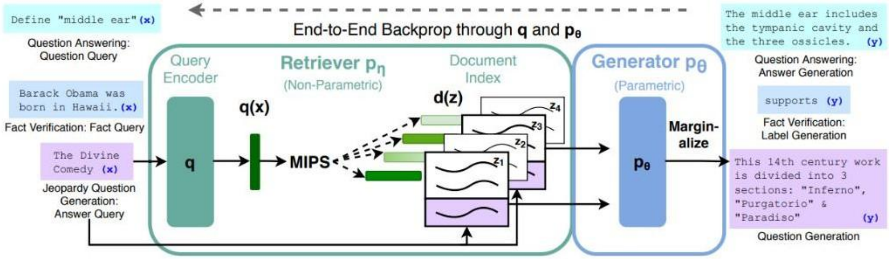


low level of structuration High

Key issue of RAG — How to use the retrieved content

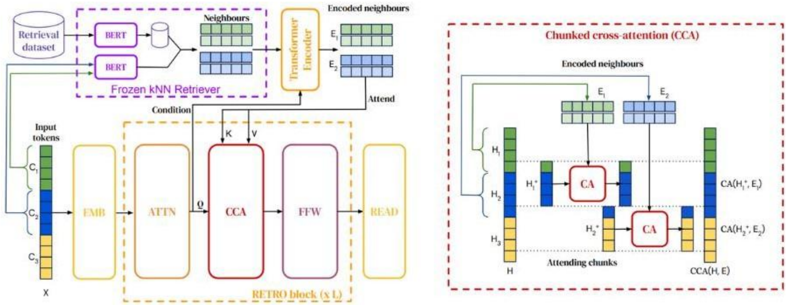
Integrating the retrieved information into different layers of the generation model,during inference process.

Integrate retrieval positions.


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

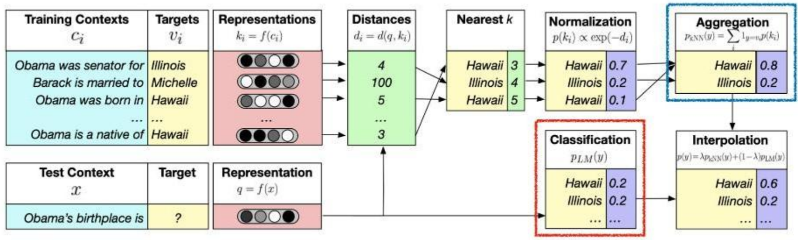
Input / Date layer

Using simple, but unable to support the retrieval of more knowledge blocks, and the optimization space is limited.



Model / Interlayer

Supports the retrieval of more knowledge blocks, but introduces additional complexity and must be trained.



Output /Prediction layer

Ensuring the output results are highly relevant to the retrieval content, but the efficiency is low.

Key issue of RAG — When to retrieve

High efficiency, but low relevance of the retrieved documents

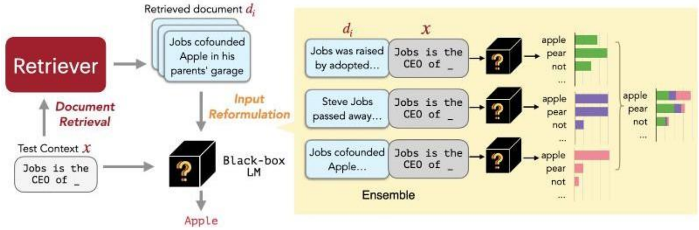
Balancing efficiency and information might not yield the optimal solution

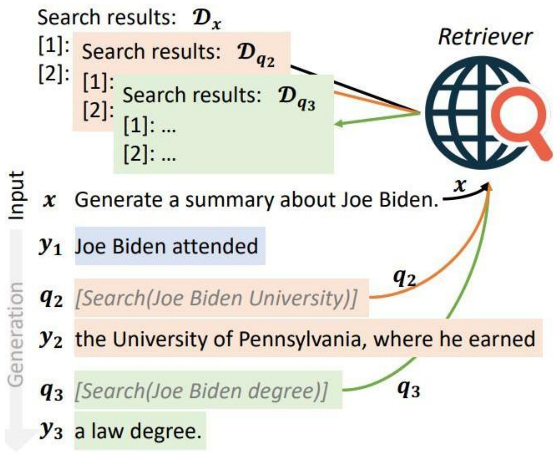
A large amount of information with low efficiency and redundant information.

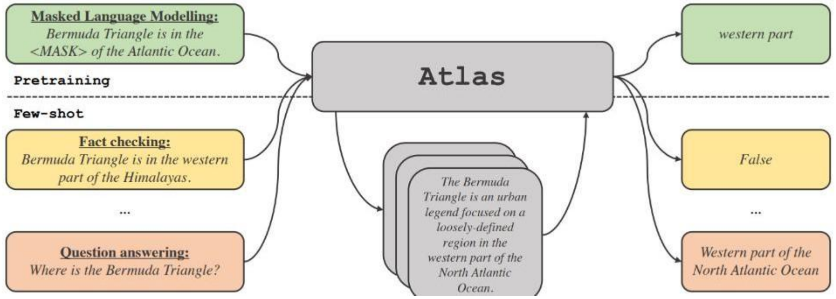
Once | Replug 2023

Adaptive | Flare 2023

Every N Tokens | Atlas 2023




Search results: $\mathcal{D}_x$
[1]:
[2]:
Search results: $\mathcal{D}_{q_2}$
[1]:
[2]:
Search results: $\mathcal{D}_{q_3}$
[1]: ...
[2]: ...
Retriever
Input
Generation
x Generate a summary about Joe Biden.
x
y_1 Joe Biden attended
q_2 [Search(Joe Biden University)]
q_2
y_2 the University of Pennsylvania, where he earned
q_3 [Search(Joe Biden degree)]
q_3
y_3 a law degree.


Masked Language Modelling:
Bermuda Triangle is in the <MASK> of the Atlantic Ocean.
Pretraining
Few-shot
Fact checking:
Bermuda Triangle is in the western part of the Himalayas.
...
Question answering:
Where is the Bermuda Triangle?
Atlas
The Bermuda Triangle is an urban legend focused on a loosely-defined region in the western part of the North Atlantic Ocean.
western part
False
...
Western part of the North Atlantic Ocean

Conducting once search during the reasoning process.

Adaptively conduct the search.

Retrieve once for every N tokens generated.

Low

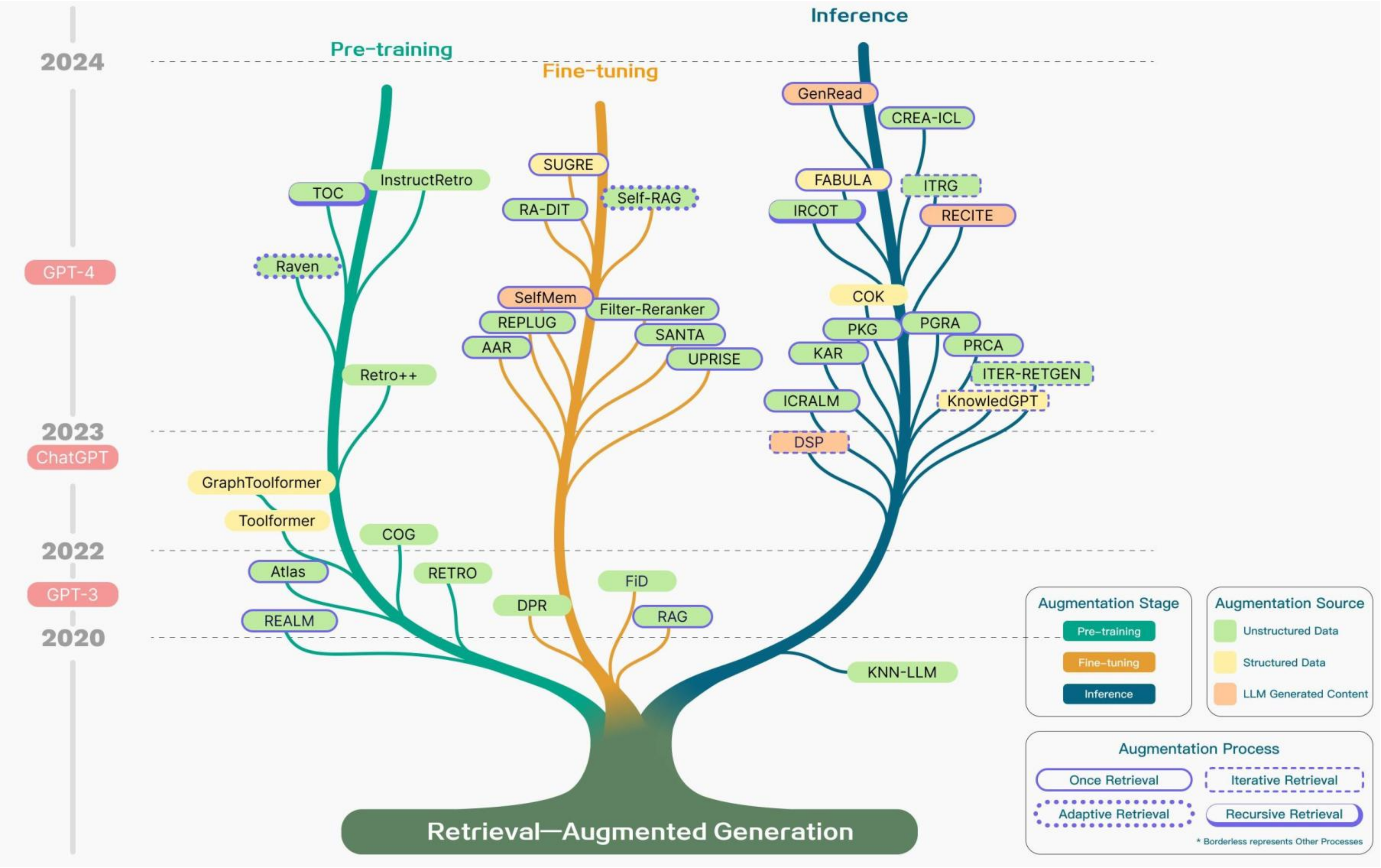
High

Retrieval frequency

TrustLLM

Funded by the European Union

Overview of RAG Development



Funded by the European Union

Techniques for Better RAG — Data indexing optimization

Chunk Optimization

Small-2-Big

Embeding at sentence level expand the window during generation process.

Slidingwindow

liding chunk covers the entire text, avoiding semantic ambiguity

Summary

Retrieve documents through summaries, then retrieve text blocks from the documents.

Adding Metadata

Example | Page | Time | Type | Document Title

Metadata Filtering/Enrichment

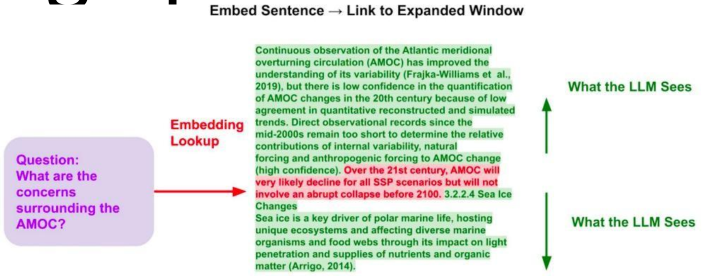
Pseudo Metadata Generation

Enhance retrieval by gener-ating a hypothetical document for the incoming query and creating qu-estions that the text block can answer.

Metadata filter

Dissect and annotate the document. During the query, infer metadata filters in addition to semantic queries

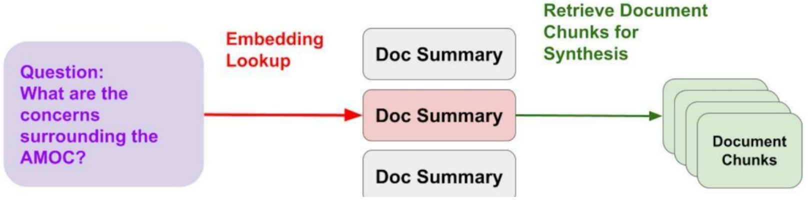
Small-2-Big


Embed Sentence → Link to Expanded Window
Question:
What are the concerns surrounding the AMOC?
Embedding Lookup
What the LLM Sees
What the LLM Sees

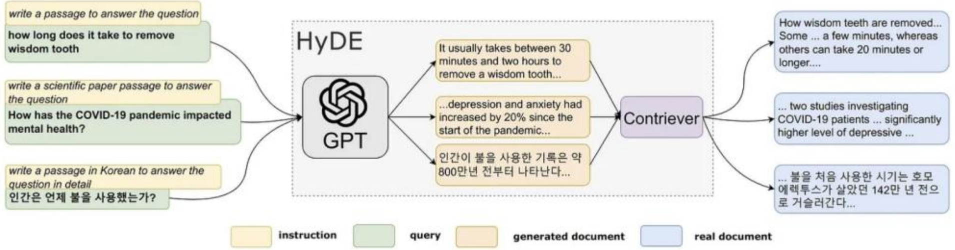
Continuous observation of the Atlantic meridional overturning circulation (AMOC) has improved the understanding of its variability (Frajka-Williams et al., 2019), but there is low confidence in the quantification of AMOC changes in the 20th century because of low agreement in quantitative reconstructed and simulated trends. Direct observational records since the mid-2000s remain too short to determine the relative contributions of internal variability, natural forcing and anthropogenic forcing to AMOC change (high confidence). Over the 21st century, AMOC will very likely decline for all SSP scenarios but will not involve an abrupt collapse before 2100. 3.2.2.4 Sea Ice Changes

Sea ice is a key driver of polar marine life, hosting unique ecosystems and affecting diverse marine organisms and food webs through its impact on light penetration and supplies of nutrients and organic matter (Arrigo, 2014).

Abstract


Question:
What are the concerns surrounding the AMOC?
Embedding Lookup
Doc Summary
Doc Summary
Doc Summary
Retrieve Document Chunks for Synthesis
Document Chunks

Pseudo Metadata



write a passage to answer the question
how long does it take to remove wisdom tooth

write a scientific paper passage to answer the question
How has the COVID-19 pandemic impacted mental health?

write a passage in Korean to answer the question in detail
인간은 언제 불을 사용했는가?

It usually takes between 30 minutes and two hours to remove a wisdom tooth...

...depression and anxiety had increased by 20% since the start of the pandemic...

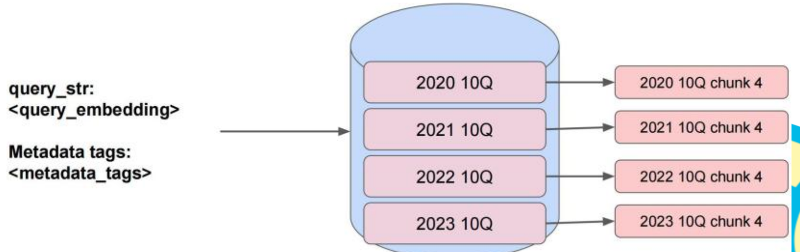
인간이 불을 사용한 기록은 약 800만년 전부터 나타난다...

How wisdom teeth are removed... Some ... a few minutes, whereas others can take 20 minutes or longer....

... two studies investigating COVID-19 patients ... significantly higher level of depressive ...

... 불을 처음 사용한 시기는 호모 에렉투스가 살았던 142만 년 전으로 거슬러간다...

Metadata filter



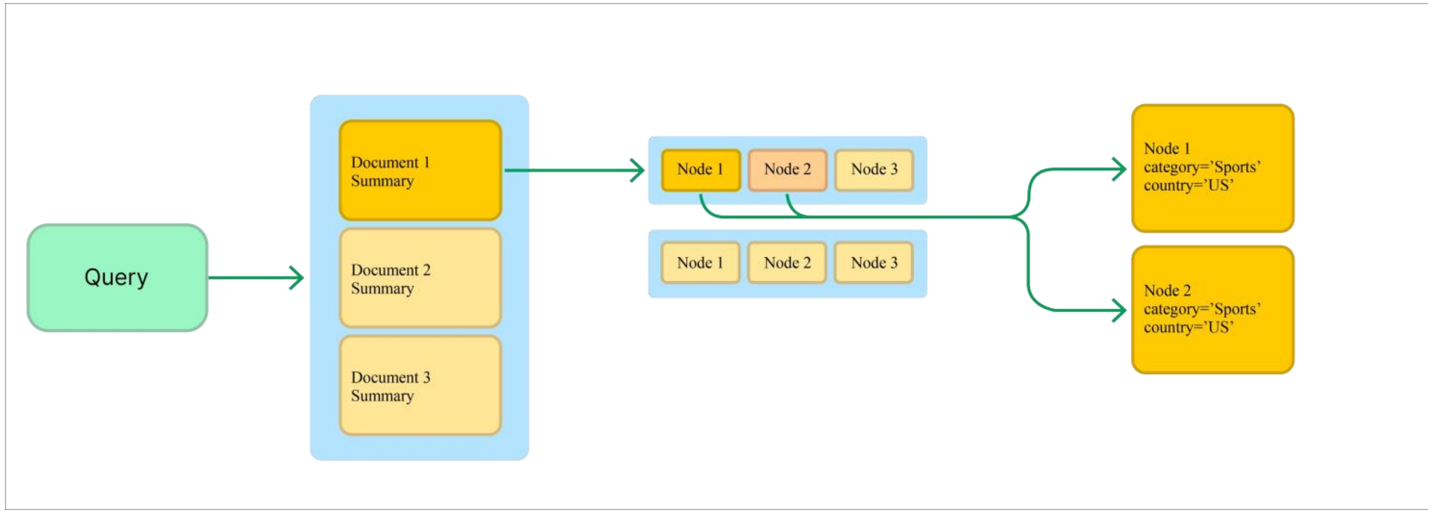
Funded by the European Union

Techniques for Better RAG — Structured Corpus

Hierarchical Organization of Retrieval Corpora

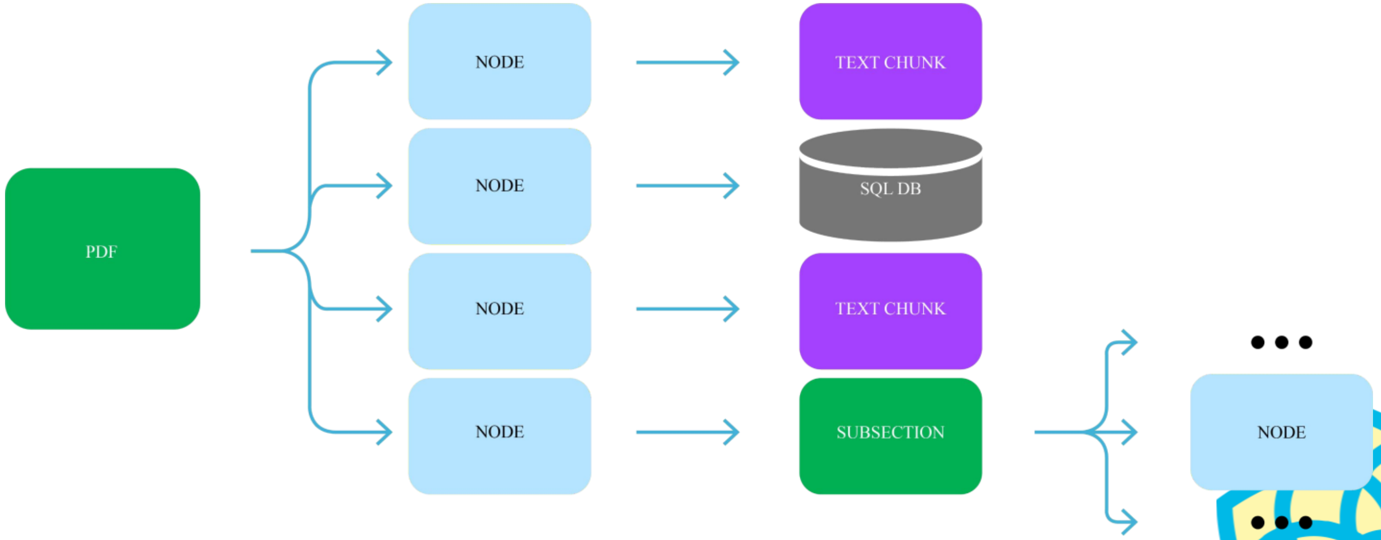
- Summary → Document

Replace document retrieval with summary retrieval, not only retrieving the most directly relevant nodes, but also exploring additional nodes associated with those nodes.



- Document → Embedded Objects

Documents have embedded objects (such as tables, charts), first retrieve entity reference objects, then query underlying objects, such as document blocks, databases, sub-nodes.



Funded by the European Union

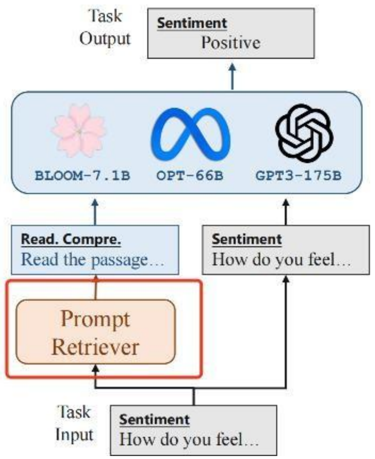
Techniques for Better RAG — Retrieval Source Optimization

Unstructured Data

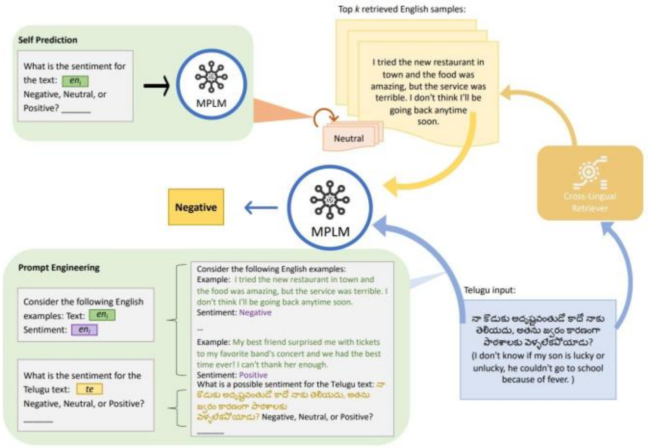
Phrases

Prompt

Cross-lingustic



Prompt | UPRISE [Cheng et al.,2023]

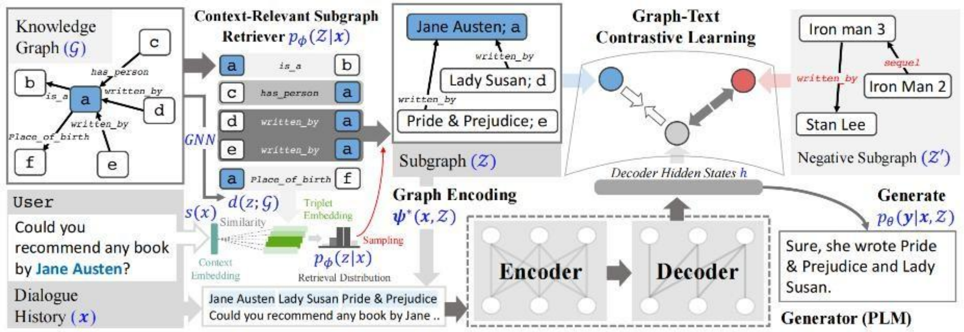

Top k retrieved English samples:
Self Prediction
What is the sentiment for the text: en, Negative, Neutral, or Positive?
MPLM
I tried the new restaurant in town and the food was amazing, but the service was terrible. I don't think I'll be going back anytime soon.
Neutral
Cross-Lingual Retriever
Negative
MPLM
Prompt Engineering
Consider the following English examples: Text: en, Sentiment: en,
What is the sentiment for the Telugu text: te Negative, Neutral, or Positive?
Consider the following English examples:
Example: I tried the new restaurant in town and the food was amazing, but the service was terrible. I don't think I'll be going back anytime soon.
Sentiment: Negative
Example: My best friend surprised me with tickets to my favorite band's concert and we had the best time ever! I can't thank her enough.
Sentiment: Positive
What is a possible sentiment for the Telugu text: ... Negative, Neutral, or Positive?
Telugu input:
(I don't know if my son is lucky or unlucky, he couldn't go to school because of fever.)

Cross-language| CREA-ICL [Li et al., 2023]

Structured Data

Triples

Subgraphs



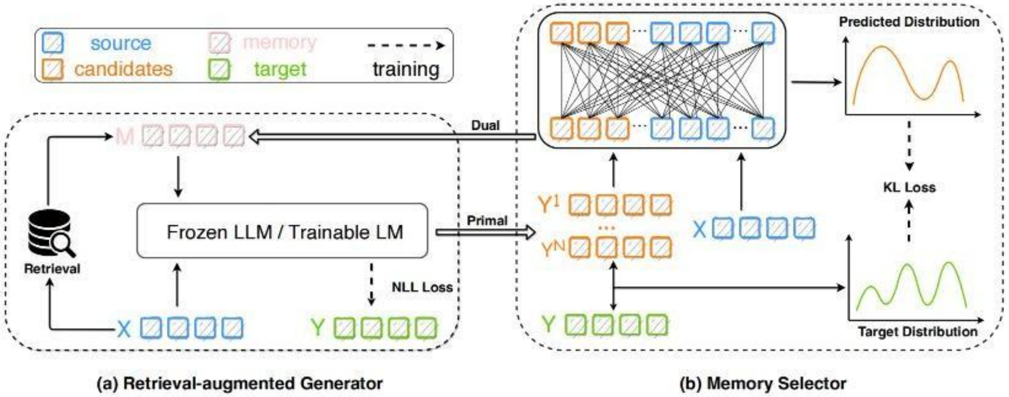
Subgraph | SUGRE [Kang et al., 2023]

LLM

LLM Memory

Generated Text

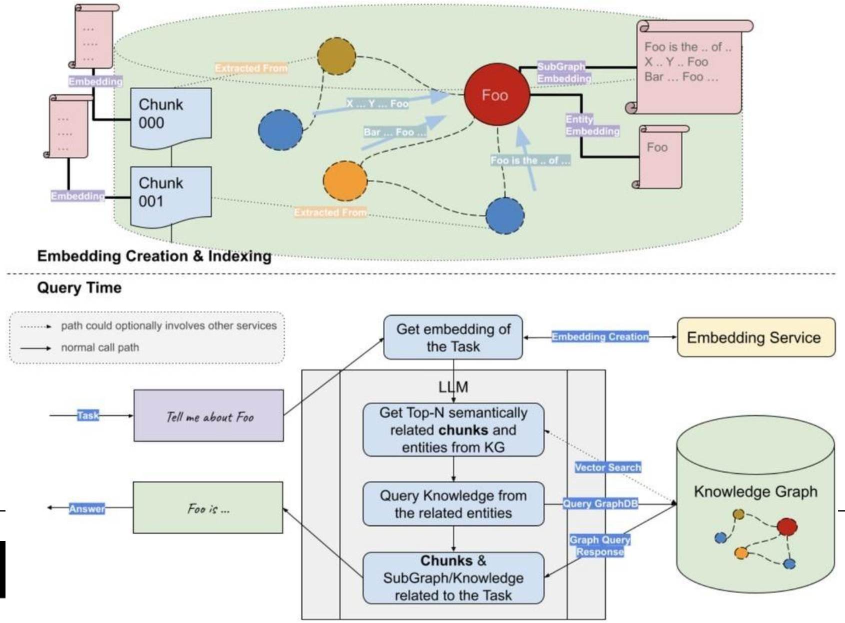
Generated Code

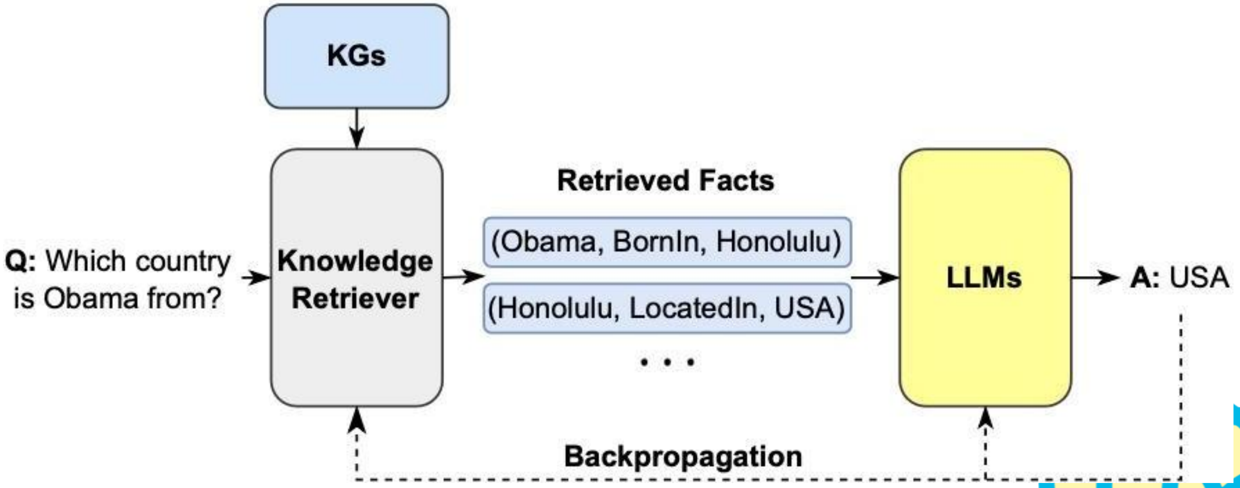


Memory | Selfmem [Cheng et al., 2023]

Techniques for Better RAG — KG as a Retrieval Data Source

- **GraphRAG**
  - Extract entities from the user's input query, then construct a subgraph to form context, and finally feed it into the large model for generation.
- **Implementation**
  - Use LLM (or other models) to extract key entities from the question.
  - Retrieve subgraphs based on entities, delving to a certain depth, such as 2 hops or even more.
  - Utilize the obtained context to generate answers through LLM.



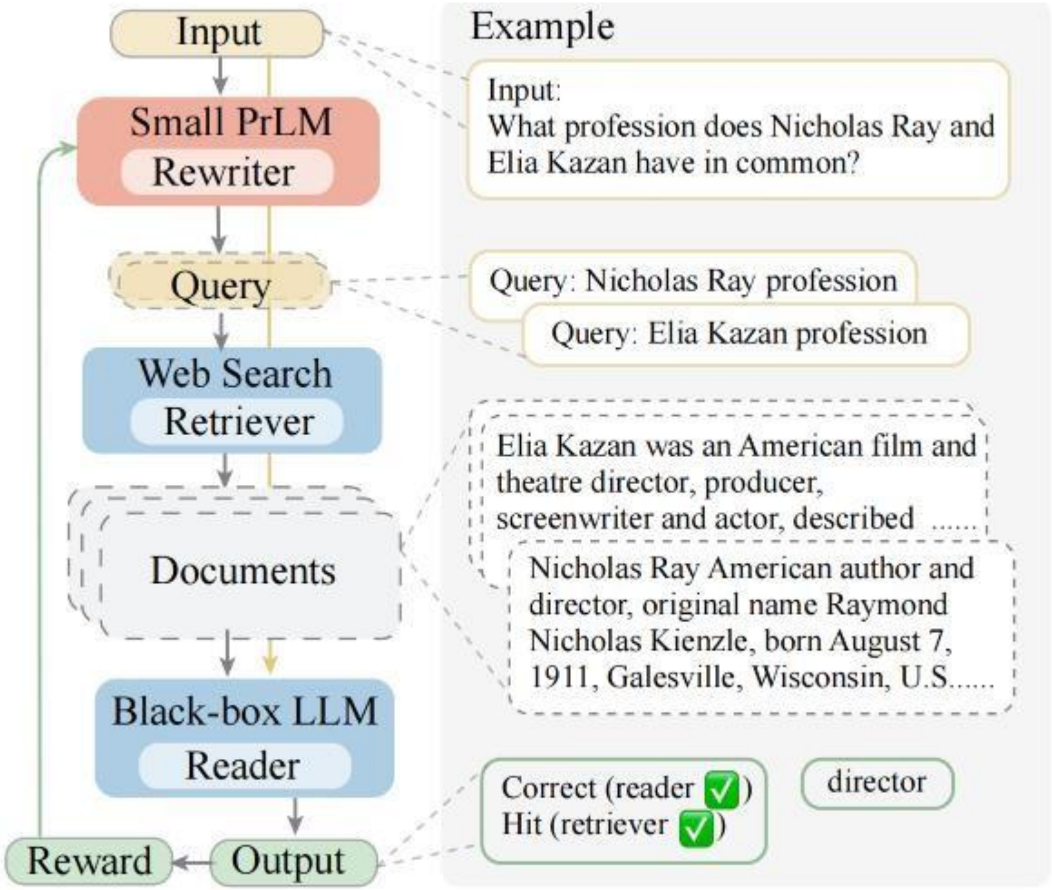

KGs
Q: Which country is Obama from?
Knowledge Retriever
Retrieved Facts
(Obama, BornIn, Honolulu)
(Honolulu, LocatedIn, USA)
. . .
LLMs
A: USA
Backpropagation

Funded by the European Union

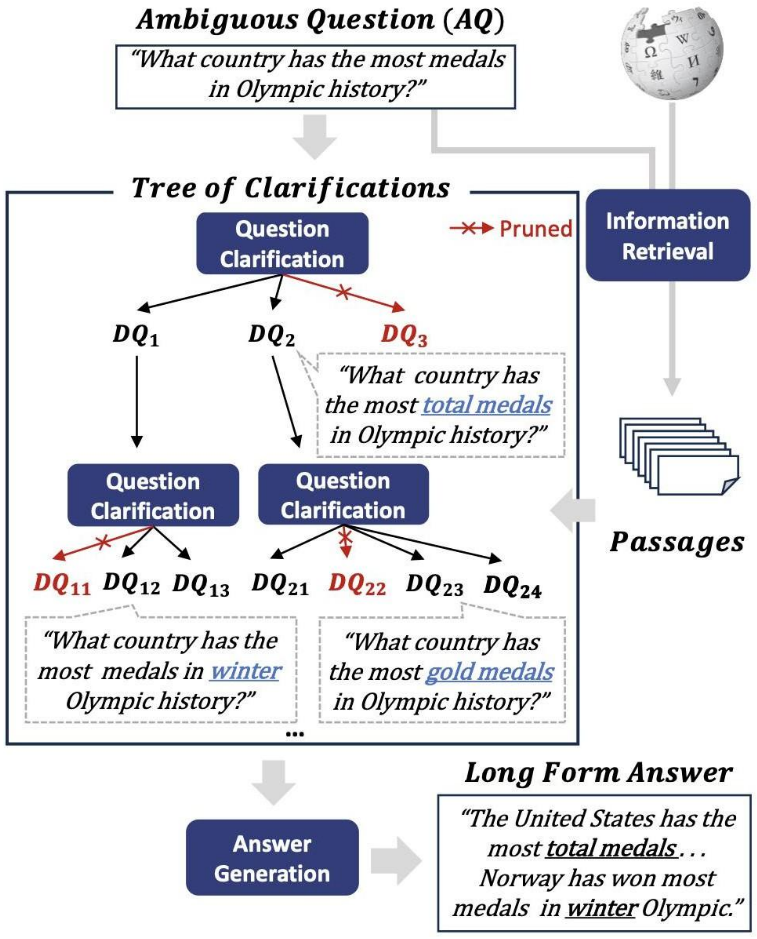
Techniques for Better RAG — Query Optimization

Questions and answers do not always possess high semantic similarity; adjusting the Query can yield better retrieval results.

Query Rewriting


Input
Small PrLM
Rewriter
Query
Web Search
Retriever
Documents
Black-box LLM
Reader
Reward
Output
Example
Input:
What profession does Nicholas Ray and Elia Kazan have in common?
Query: Nicholas Ray profession
Query: Elia Kazan profession
Elia Kazan was an American film and theatre director, producer, screenwriter and actor, described ......
Nicholas Ray American author and director, original name Raymond Nicholas Kienzle, born August 7, 1911, Galesville, Wisconsin, U.S......
Correct (reader ✅)
Hit (retriever ✅)
director

Query Clarification


Ambiguous Question (AQ)
"What country has the most medals in Olympic history?"
Tree of Clarifications
Question Clarification
Pruned
Information Retrieval
DQ1
DQ2
DQ3
"What country has the most total medals in Olympic history?"
Question Clarification
Question Clarification
Passages
DQ11 DQ12 DQ13 DQ21 DQ22 DQ23 DQ24
"What country has the most medals in winter Olympic history?"
"What country has the most gold medals in Olympic history?"
...
Long Form Answer
Answer Generation
"The United States has the most total medals... Norway has won most medals in winter Olympic."

Rewrite-Retrieve-Read [Ma et al., 2023]

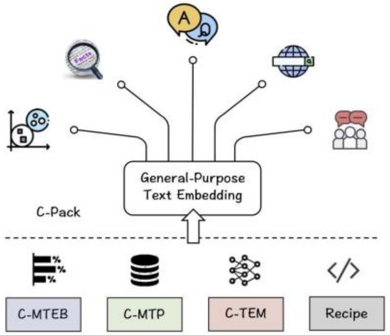
Tree of Clarifications（TOC）[Kim et al.,2023]

Funded by the European Union

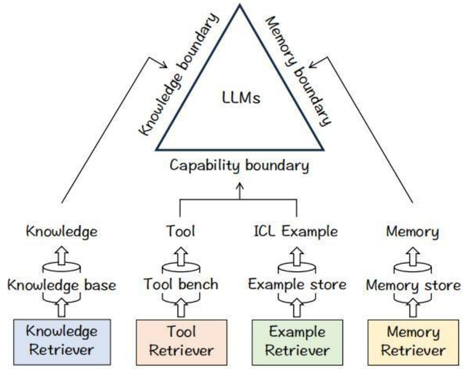
Techniques for Better RAG – Embedding Optimization

Selecting a More Suitable Embedding Provider

cohere M3E VOYAGE AI OpenAI llmrails BAAI 智源研究院

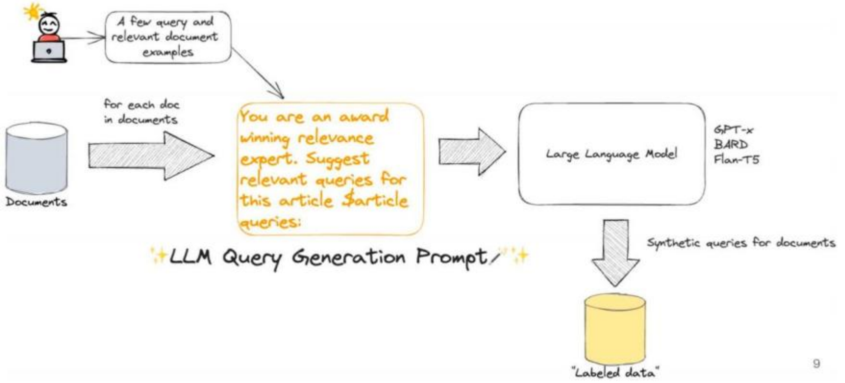


BAAI-General-Embedding (BGE)



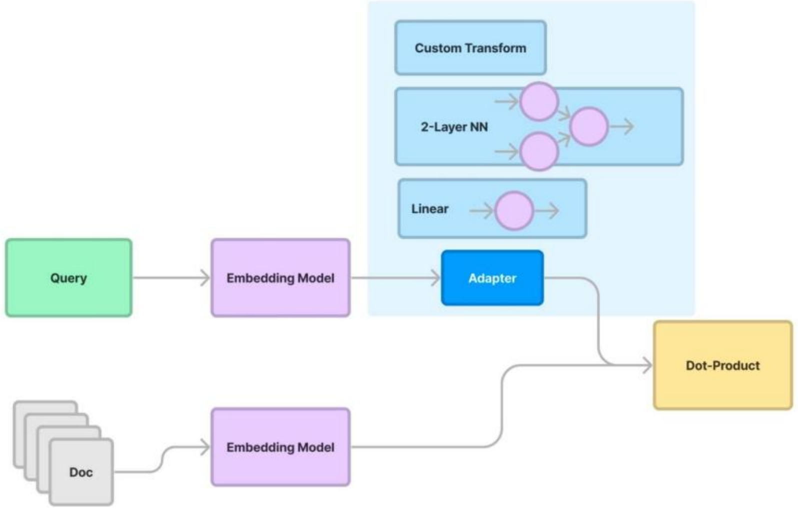
LLM-Embedder(BGE2) [Aksitov et al.,2023]

Fine-tuning the Embedding Model



A few query and relevant document examples
for each doc in documents
Documents
You are an award winning relevance expert. Suggest relevant queries for this article $article queries:
✨LLM Query Generation Prompt✨
Large Language Model
GPT-x
BARD
Flan-T5
Synthetic queries for documents
"Labeled data"
9

Fine-tuning According to Domain-Specific Repositories and Downstream Tasks



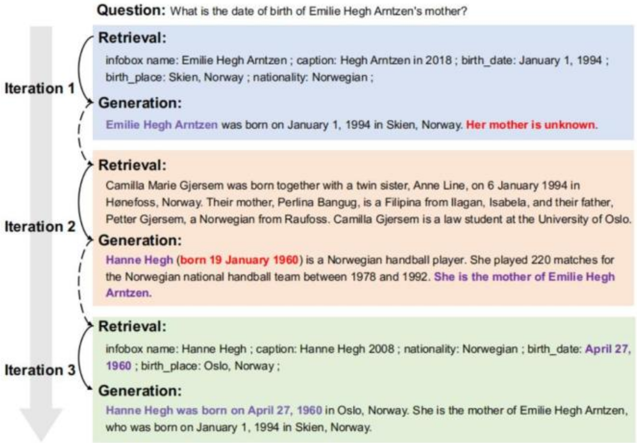
Fine-tuning the Adapter Module to Align the Embedding Model with the Retrieval Repository

Techniques for Better RAG — Retrieval Process Optimization

Iterative

Iteratively Retrieving from the Corpus to Acquire More Detailed and In-depth Knowledge

ITER [Feng et al., 2023]



Question: What is the date of birth of Emilie Hegh Arntzen's mother?

Retrieval: infobox name: Emilie Hegh Arntzen ; caption: Hegh Arntzen in 2018 ; birth_date: January 1, 1994 ; birth_place: Skien, Norway ; nationality: Norwegian ;

Generation: Emilie Hegh Arntzen was born on January 1, 1994 in Skien, Norway. Her mother is unknown.

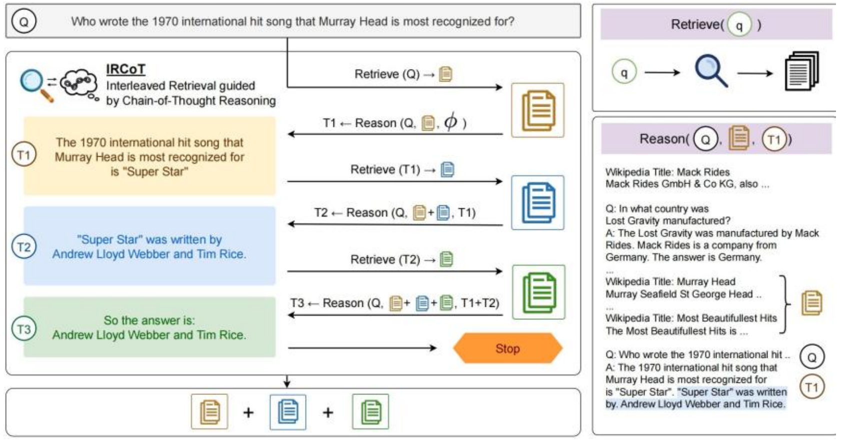
Retrieval: Camilla Marie Gjersem was born together with a twin sister, Anne Line, on 6 January 1994 in Hønefoss, Norway. Their mother, Perlina Bangug, is a Filipina from Ilagan, Isabela, and their father, Petter Gjersem, a Norwegian from Raufoss. Camilla Gjersem is a law student at the University of Oslo.

Generation: Hanne Hegh (born 19 January 1960) is a Norwegian handball player. She played 220 matches for the Norwegian national handball team between 1978 and 1992. She is the mother of Emilie Hegh Arntzen.

Retrieval: infobox name: Hanne Hegh ; caption: Hanne Hegh 2008 ; nationality: Norwegian ; birth_date: April 27, 1960 ; birth_place: Oslo, Norway ;

Generation: Hanne Hegh was born on April 27, 1960 in Oslo, Norway. She is the mother of Emilie Hegh Arntzen, who was born on January 1, 1994 in Skien, Norway.

IRCOT [Trivedi et al.,2022]



Q Who wrote the 1970 international hit song that Murray Head is most recognized for?

IRCoT Interleaved Retrieval guided by Chain-of-Thought Reasoning

T1 The 1970 international hit song that Murray Head is most recognized for is "Super Star"

T2 "Super Star" was written by Andrew Lloyd Webber and Tim Rice.

T3 So the answer is: Andrew Lloyd Webber and Tim Rice.

Wikipedia Title: Mack Rides
Mack Rides GmbH & Co KG, also ...

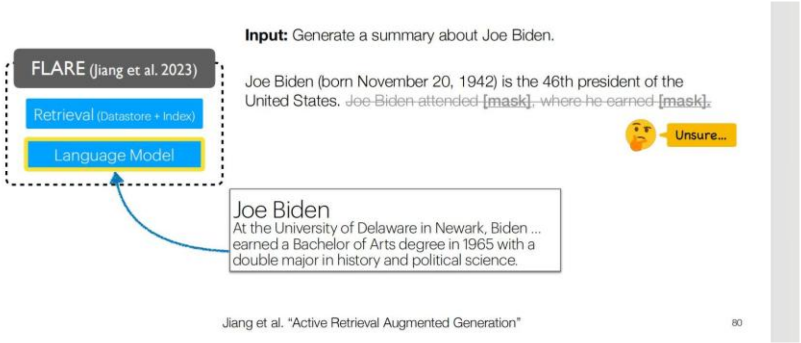
Q: In what country was Lost Gravity manufactured?
A: The Lost Gravity was manufactured by Mack Rides. Mack Rides is a company from Germany. The answer is Germany.
...
Wikipedia Title: Murray Head
Murray Seafield St George Head ..
...
Wikipedia Title: Most Beautifullest Hits
The Most Beautifullest Hits is ...

Q: Who wrote the 1970 international hit ..
A: The 1970 international hit song that Murray Head is most recognized for is "Super Star". "Super Star" was written by. Andrew Lloyd Webber and Tim Rice.

Adaptive

Dynamically Determined by the LLM, the Timing and Scope of Retrieva

FLARE [Jiang et al., 2023]



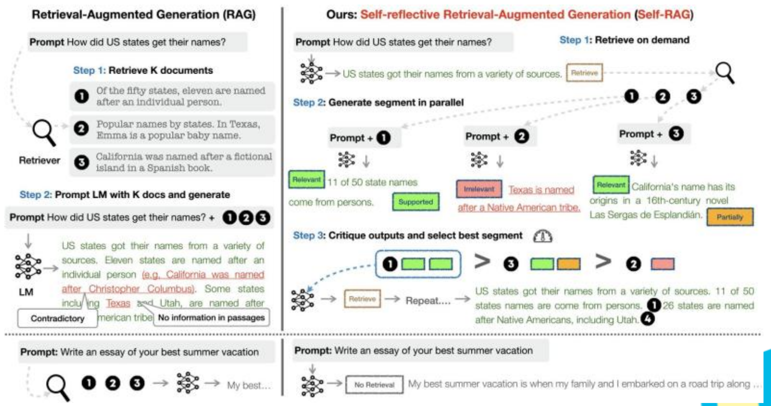
Input: Generate a summary about Joe Biden.

Joe Biden (born November 20, 1942) is the 46th president of the United States. ~~Joe Biden attended [mask], where he earned [mask].~~

Joe Biden
At the University of Delaware in Newark, Biden ... earned a Bachelor of Arts degree in 1965 with a double major in history and political science.

Jiang et al. "Active Retrieval Augmented Generation"

Self-RAG [Asai et al., 2023]



Retrieval-Augmented Generation (RAG)

Prompt How did US states get their names?

Step 1: **Retrieve K documents**

1 Of the fifty states, eleven are named after an individual person.

2 Popular names by states. In Texas, Emma is a popular baby name.

3 California was named after a fictional island in a Spanish book.

Step 2: **Prompt LM with K docs and generate**

Prompt How did US states get their names? + 1 2 3

US states got their names from a variety of sources. Eleven states are named after an individual person (e.g, California was named after Christopher Columbus). Some states including Texas and Utah, are named after Native American tribe.

Ours: Self-reflective Retrieval-Augmented Generation (Self-RAG)

Prompt How did US states get their names?

Step 1: **Retrieve on demand**

US states got their names from a variety of sources. Retrieve

Step 2: **Generate segment in parallel**

Prompt + 1 / Prompt + 2 / Prompt + 3

Relevant 11 of 50 state names come from persons. Supported

Irrelevant Texas is named after a Native American tribe.

Relevant California's name has its origins in a 16th-century novel Las Sergas de Esplandián. Partially

Step 3: **Critique outputs and select best segment**

1 > 3 > 2

Retrieve → Repeat.... → US states got their names from a variety of sources. 11 of 50 states names are come from persons. 1 26 states are named after Native Americans, including Utah. 4

Prompt: Write an essay of your best summer vacation

1 2 3 → My best...

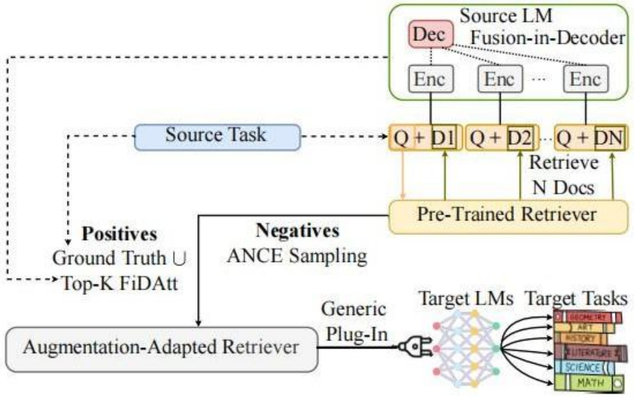
Prompt: Write an essay of your best summer vacation

No Retrieval My best summer vacation is when my family and I embarked on a road trip along ...

Funded by the European Union

Techniques for Better RAG — Hybrid (RAG + Fine-tuning)

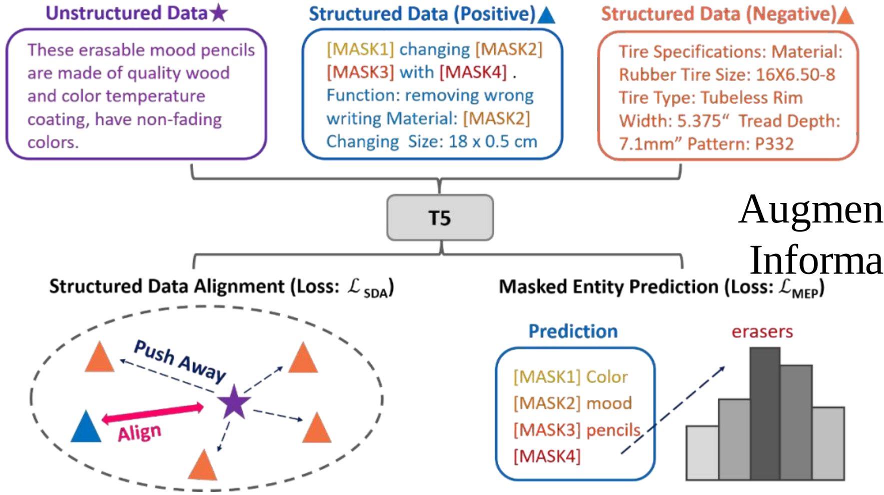
Retriever Fine-Tuning



Highly Adaptive General-Purpose Retrieval Plugin

AAR [Yu et al., 2023]

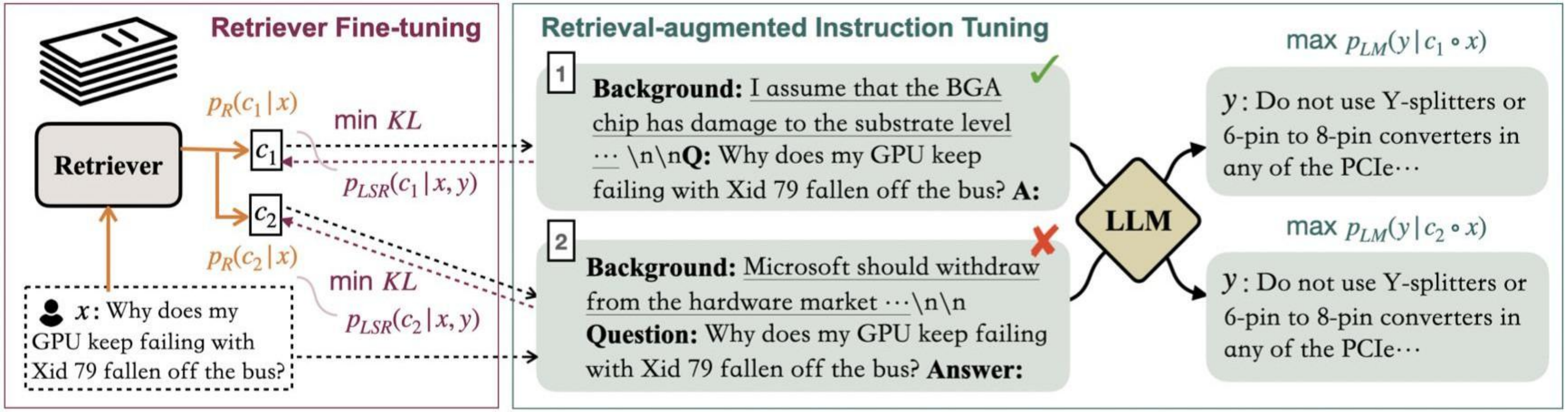
Generator Fine-Tuning


Unstructured Data★
These erasable mood pencils are made of quality wood and color temperature coating, have non-fading colors.
Structured Data (Positive)▲
[MASK1] changing [MASK2] [MASK3] with [MASK4] . Function: removing wrong writing Material: [MASK2] Changing Size: 18 x 0.5 cm
Structured Data (Negative)▲
Tire Specifications: Material: Rubber Tire Size: 16X6.50-8 Tire Type: Tubeless Rim Width: 5.375" Tread Depth: 7.1mm" Pattern: P332
T5
Structured Data Alignment (Loss: $\mathcal{L}_{SDA}$)
Masked Entity Prediction (Loss: $\mathcal{L}_{MEP}$)
Push Away
Align
Prediction
[MASK1] Color
[MASK2] mood
[MASK3] pencils
[MASK4]
erasers

Augment with Structural Information Integration

SANTA [Li et al., 2023]

Collaborative Fine-Tuning



x: Why does my GPU keep failing with Xid 79 fallen off the bus?

Background: I assume that the BGA chip has damage to the substrate level ··· \n\n**Q:** Why does my GPU keep failing with Xid 79 fallen off the bus? **A:**

Background: Microsoft should withdraw from the hardware market ···\n\n **Question:** Why does my GPU keep failing with Xid 79 fallen off the bus? **Answer:**

y: Do not use Y-splitters or 6-pin to 8-pin converters in any of the PCIe···

y: Do not use Y-splitters or 6-pin to 8-pin converters in any of the PCIe···

RA-DIT [Lin et al., 2023]

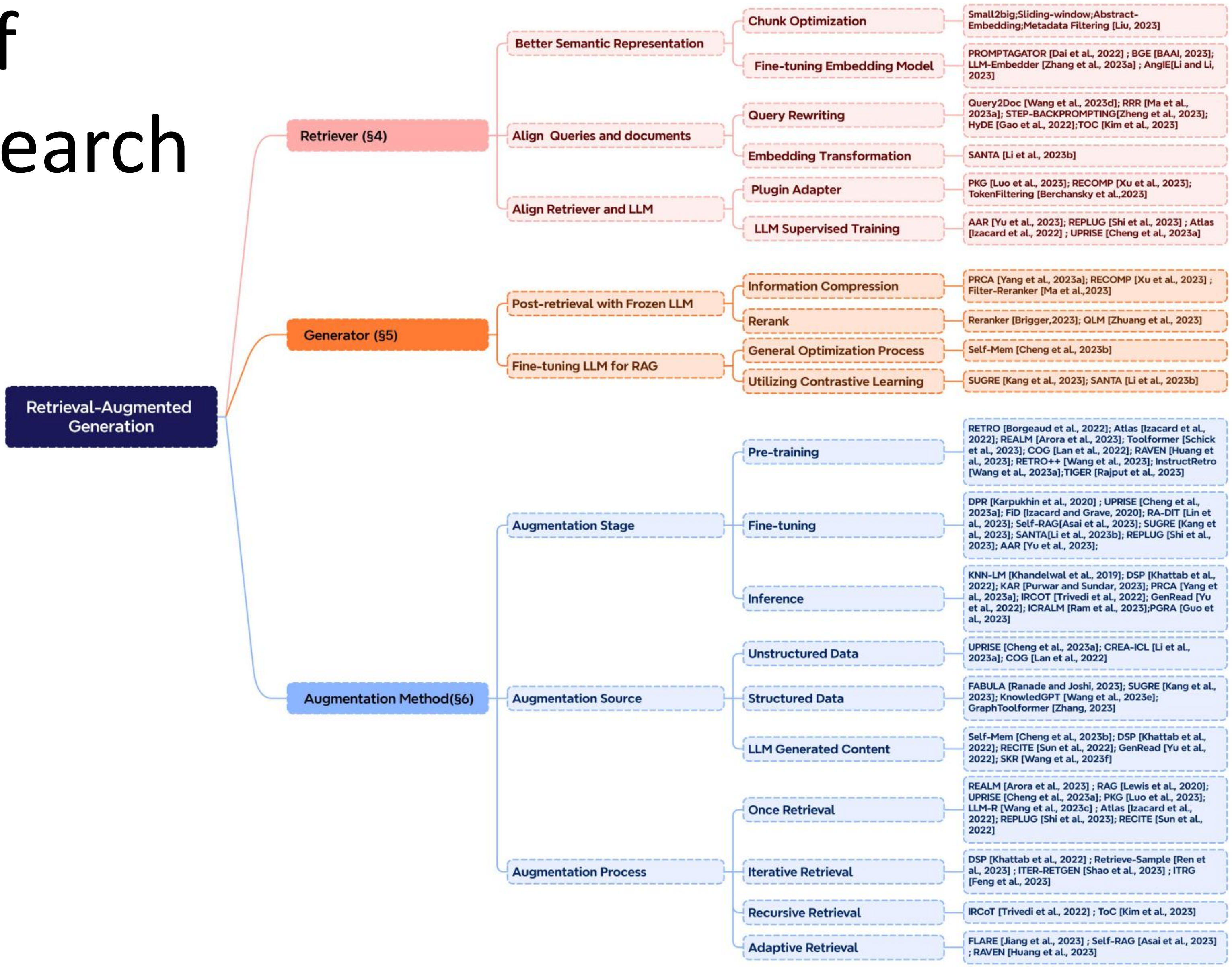
- R-FT

Minimizing the KL Divergence Between the Retriever Distribution and LLM Preferences

- LM-FT

Maximizing the Likelihood of the Correct Answer Given Retrieval-Augmented Instructions

Summary of Related Research



- Retrieval-Augmented Generation
  - Retriever (§4)
    - Better Semantic Representation
      - Chunk Optimization: Small2big;Sliding-window;Abstract-Embedding;Metadata Filtering [Liu, 2023]
      - Fine-tuning Embedding Model: PROMPTAGATOR [Dai et al., 2022] ; BGE [BAAI, 2023]; LLM-Embedder [Zhang et al., 2023a] ; AnglE[Li and Li, 2023]
    - Align Queries and documents
      - Query Rewriting: Query2Doc [Wang et al., 2023d]; RRR [Ma et al., 2023a]; STEP-BACKPROMPTING[Zheng et al., 2023]; HyDE [Gao et al., 2022];TOC [Kim et al., 2023]
      - Embedding Transformation: SANTA [Li et al., 2023b]
    - Align Retriever and LLM
      - Plugin Adapter: PKG [Luo et al., 2023]; RECOMP [Xu et al., 2023]; TokenFiltering [Berchansky et al.,2023]
      - LLM Supervised Training: AAR [Yu et al., 2023]; REPLUG [Shi et al., 2023] ; Atlas [Izacard et al., 2022] ; UPRISE [Cheng et al., 2023a]
  - Generator (§5)
    - Post-retrieval with Frozen LLM
      - Information Compression: PRCA [Yang et al., 2023a]; RECOMP [Xu et al., 2023] ; Filter-Reranker [Ma et al.,2023]
      - Rerank: Reranker [Brigger,2023]; QLM [Zhuang et al., 2023]
    - Fine-tuning LLM for RAG
      - General Optimization Process: Self-Mem [Cheng et al., 2023b]
      - Utilizing Contrastive Learning: SUGRE [Kang et al., 2023]; SANTA [Li et al., 2023b]
  - Augmentation Method(§6)
    - Augmentation Stage
      - Pre-training: RETRO [Borgeaud et al., 2022]; Atlas [Izacard et al., 2022]; REALM [Arora et al., 2023]; Toolformer [Schick et al., 2023]; COG [Lan et al., 2022]; RAVEN [Huang et al., 2023]; RETRO++ [Wang et al., 2023]; InstructRetro [Wang et al., 2023a];TIGER [Rajput et al., 2023]
      - Fine-tuning: DPR [Karpukhin et al., 2020] ; UPRISE [Cheng et al., 2023a]; FiD [Izacard and Grave, 2020]; RA-DIT [Lin et al., 2023]; Self-RAG[Asai et al., 2023]; SUGRE [Kang et al., 2023]; SANTA[Li et al., 2023b]; REPLUG [Shi et al., 2023]; AAR [Yu et al., 2023];
      - Inference: KNN-LM [Khandelwal et al., 2019]; DSP [Khattab et al., 2022]; KAR [Purwar and Sundar, 2023]; PRCA [Yang et al., 2023a]; IRCOT [Trivedi et al., 2022]; GenRead [Yu et al., 2022]; ICRALM [Ram et al., 2023];PGRA [Guo et al., 2023]
    - Augmentation Source
      - Unstructured Data: UPRISE [Cheng et al., 2023a]; CREA-ICL [Li et al., 2023a]; COG [Lan et al., 2022]
      - Structured Data: FABULA [Ranade and Joshi, 2023]; SUGRE [Kang et al., 2023]; KnowledGPT [Wang et al., 2023e]; GraphToolformer [Zhang, 2023]
      - LLM Generated Content: Self-Mem [Cheng et al., 2023b]; DSP [Khattab et al., 2022]; RECITE [Sun et al., 2022]; GenRead [Yu et al., 2022]; SKR [Wang et al., 2023f]
    - Augmentation Process
      - Once Retrieval: REALM [Arora et al., 2023] ; RAG [Lewis et al., 2020]; UPRISE [Cheng et al., 2023a]; PKG [Luo et al., 2023]; LLM-R [Wang et al., 2023c] ; Atlas [Izacard et al., 2022]; REPLUG [Shi et al., 2023]; RECITE [Sun et al., 2022]
      - Iterative Retrieval: DSP [Khattab et al., 2022] ; Retrieve-Sample [Ren et al., 2023] ; ITER-RETGEN [Shao et al., 2023] ; ITRG [Feng et al., 2023]
      - Recursive Retrieval: IRCoT [Trivedi et al., 2022] ; ToC [Kim et al., 2023]
      - Adaptive Retrieval: FLARE [Jiang et al., 2023] ; Self-RAG [Asai et al., 2023] ; RAVEN [Huang et al., 2023]

《Retrieval-Augmented Generation for Large Language Models: A Survey》

Funded by the European Union

How to Evaluate the Effectiveness of RAG

EvaluationMethods

Independent Evaluation

Retriever
Evaluate the Quality of Text Blocks Retrieved by the Query
Metrics: MRP, Hit Rate, NDCG

Generation/Synthesis
Quality of Context Enhanced with Retrieved Documents Evaluation
Metrics: Context Relevance

End-to-End Evaluation

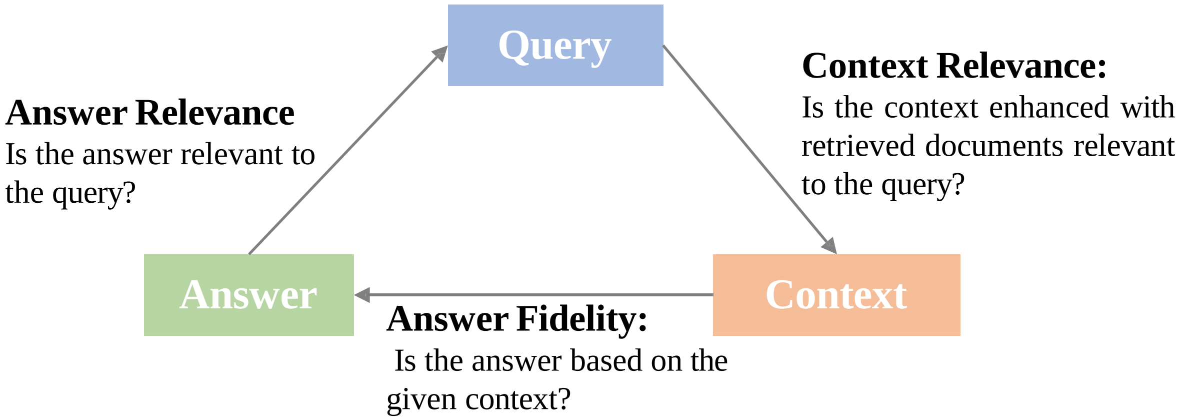
Evaluate the content ultimately generated by the model.

By generated conten
With labels：EM，Accuracy
Without labels: Fidelity, Relevance, Harmlessness

By evaluation method
Human evaluation
Automatic evaluation (LLM judge)

Key Metrics & Capabilities

Key Metrics



Answer Relevance
Is the answer relevant to the query?

Context Relevance:
Is the context enhanced with retrieved documents relevant to the query?

Answer Fidelity:
Is the answer based on the given context?

Key Capabilities

Noise Robustness
Can the model extract useful information from noisy documents?

Negative Rejection
When therequired knowledge is not exsiting in the retrieved documents, the answer should be refused.

Info Integration
Can the model answer complex questions that require integrating information from multiple documents?

Counterfactual Robustness
Can the model recognize the risk of known factual errors in the retrieved documents?

Assessment Framework

Use LLM as the adjudicator judge.

TruLens RAGAS ARES

Based on handwritten prompt (TruLens, RAGAS)

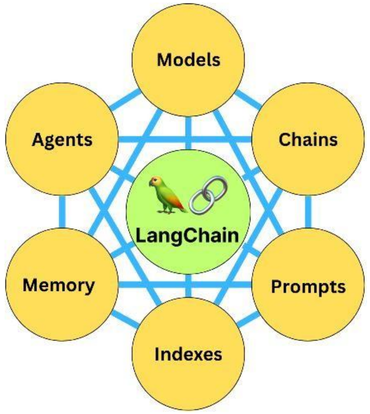
Synthetic dataset + Fine-tuning + Ranking using confidence intervals (ARES)

Evaluation

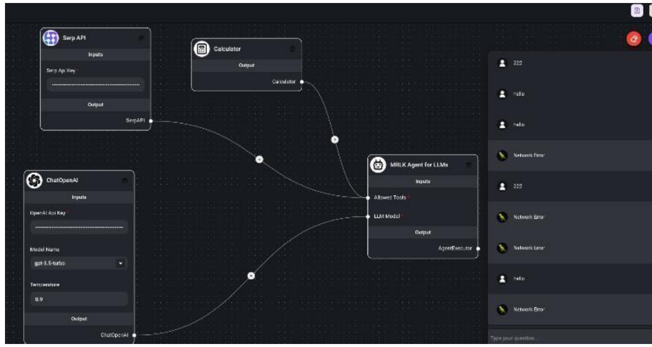
- Answer Fidelity
- Answer Relevance
- Contextual Relevance

Existing Tech Stack for RAG

Name	Pros	Cons
LangChain	Modular, full-featured	Inconsistent behavior ,API conceals details,complexity and low flexibility.
LlamaIndex	Focus on RAG	Requires combination use, low customization.
FlowiseAI	Easy to get started, visualized workflows.	Does not support complex scenarios.
AutoGen	Adapts to multi-agent scenarios.	Low efficiency, requires multiple rounds of dialogue.

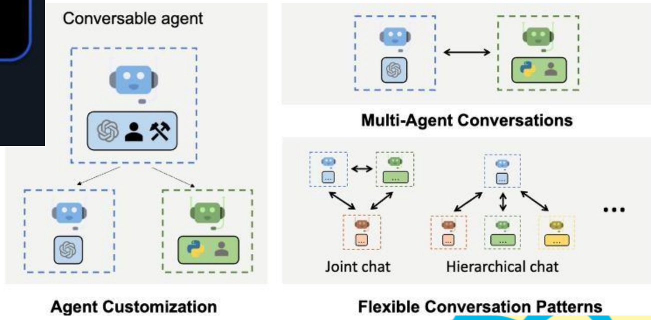


LangChain



FlowiseAI

LlamaIndex



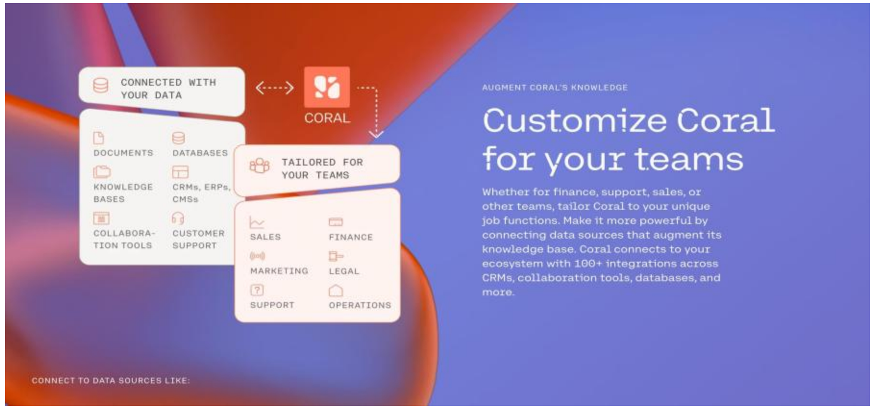
AutoGen

TrustLLM

Funded by the European Union

RAG Industry Application Practices

NetEase - ChatBI



AUGMENT CORAL'S KNOWLEDGE

Customize Coral for your teams

Whether for finance, support, sales, or other teams, tailor Coral to your unique job functions. Make it more powerful by connecting data sources that augment its knowledge base. Coral connects to your ecosystem with 100+ integrations across CRMs, collaboration tools, databases, and more.

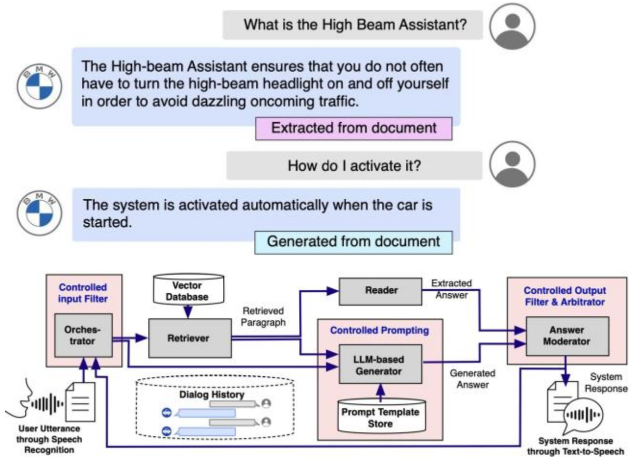
CONNECT TO DATA SOURCES LIKE:

Cohere - Coral

The intelligent upgrade of traditional industries

RAG

AI Toolchain Enhancement



What is the High Beam Assistant?

The High-beam Assistant ensures that you do not often have to turn the high-beam headlight on and off yourself in order to avoid dazzling oncoming traffic.

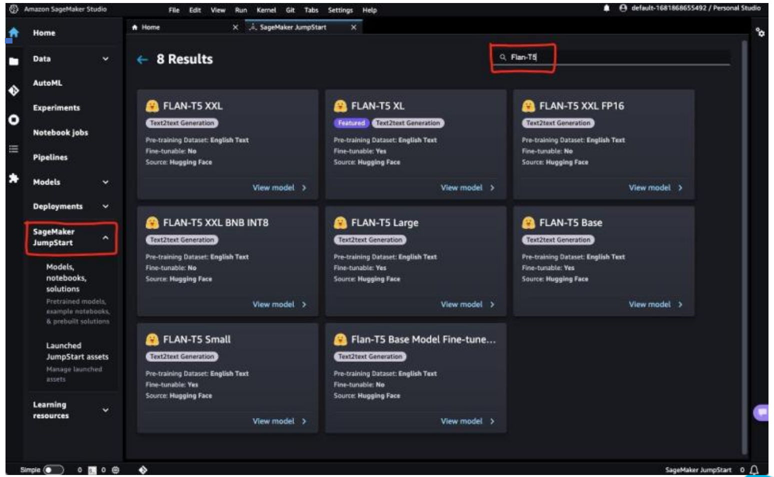
Extracted from document

How do I activate it?

The system is activated automatically when the car is started.

Generated from document

BMW - CarExpert



8 Results

Flan-T5

FLAN-T5 XXL
Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face
View model

FLAN-T5 XL
Featured Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: Yes
Source: Hugging Face
View model

FLAN-T5 XXL FP16
Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face
View model

FLAN-T5 XXL BNB INT8
Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face
View model

FLAN-T5 Large
Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: Yes
Source: Hugging Face
View model

FLAN-T5 Base
Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: Yes
Source: Hugging Face
View model

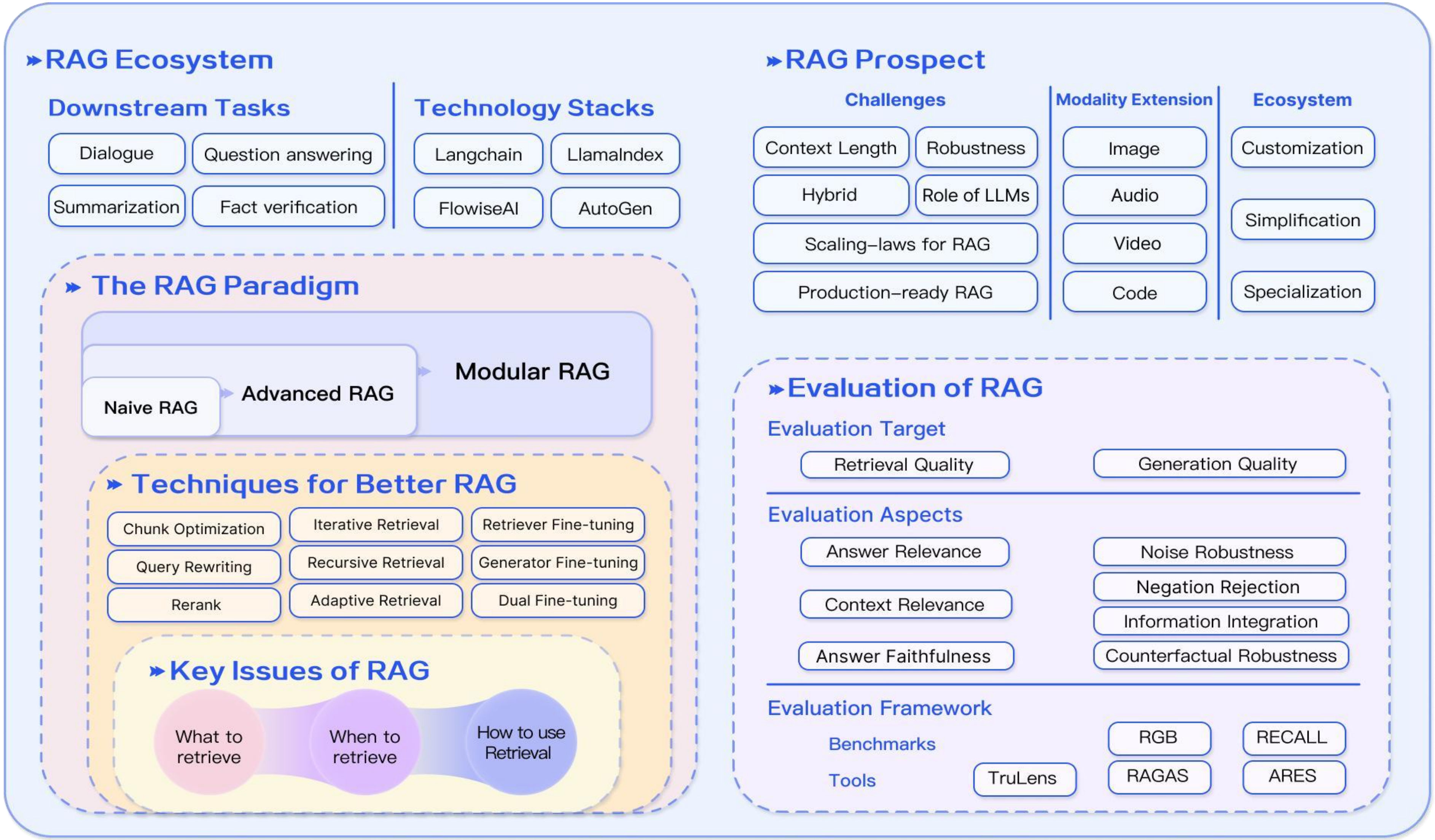
FLAN-T5 Small
Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: Yes
Source: Hugging Face
View model

Flan-T5 Base Model Fine-tune...
Text2text Generation
Pre-training Dataset: English Text
Fine-tunable: No
Source: Hugging Face
View model

Amazon - Kendra

Funded by the European Union

Summary — The Framework of RAG



RAG Ecosystem

Downstream Tasks

- Dialogue
- Question answering
- Summarization
- Fact verification

Technology Stacks

- Langchain
- LlamaIndex
- FlowiseAI
- AutoGen

The RAG Paradigm

Naive RAG → Advanced RAG → Modular RAG

Techniques for Better RAG

- Chunk Optimization
- Iterative Retrieval
- Retriever Fine-tuning
- Query Rewriting
- Recursive Retrieval
- Generator Fine-tuning
- Rerank
- Adaptive Retrieval
- Dual Fine-tuning

Key Issues of RAG

- What to retrieve
- When to retrieve
- How to use Retrieval

RAG Prospect

Challenges

- Context Length
- Robustness
- Hybrid
- Role of LLMs
- Scaling–laws for RAG
- Production–ready RAG

Modality Extension

- Image
- Audio
- Video
- Code

Ecosystem

- Customization
- Simplification
- Specialization

Evaluation of RAG

Evaluation Target

- Retrieval Quality
- Generation Quality

Evaluation Aspects

- Answer Relevance
- Context Relevance
- Answer Faithfulness
- Noise Robustness
- Negation Rejection
- Information Integration
- Counterfactual Robustness

Evaluation Framework

Benchmarks: RGB, RECALL, RAGAS, ARES

Tools: TruLens

TrustLLM

Funded by the European Union

Prospects — Existing Challengs of RAG

Further address the challenges faced by RAG itself

Long context

- Retrieved content is excessive, exceeding window limit.
- The context is too long to result Lost in the Milddle.
- If the context window is not limited, is there still a need for RAG?

Coordination with FT

- How to simultaneously leverage the effects of RAG and FT.
- How do the two coordinate, how are they organized, is it in Pipeline, alternating, or end-to-end?

The role of LLMs

- LLM can be used for retrieval (LLM generation replaces retrieval, retrieving from LLM memory), for generation, and for evaluation. How to further explore the potential of LLM in RAG.

Robustness

- How to handle the incorrect content retrieved
- How to filter and verify the content retrieved.
- How to improve the model's resistance to toxicity and noise

Scaling Law

- Does the RAG model satisfy the Scaling Law
- Does RAG exhibit, or under what scenarios does it exhibit an Inverse Scaling Law

Engineering Practice

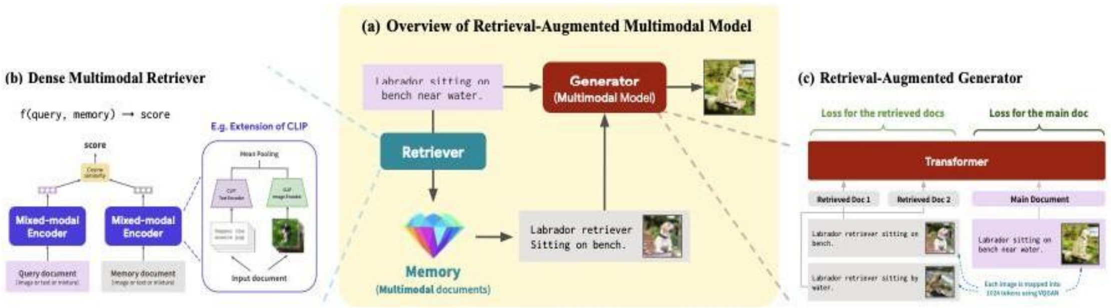
- How to reduce the latency of retrieving ultra-large-scale corpora.
- How to ensure that the content retrieved is not leaked by large models

Funded by the European Union

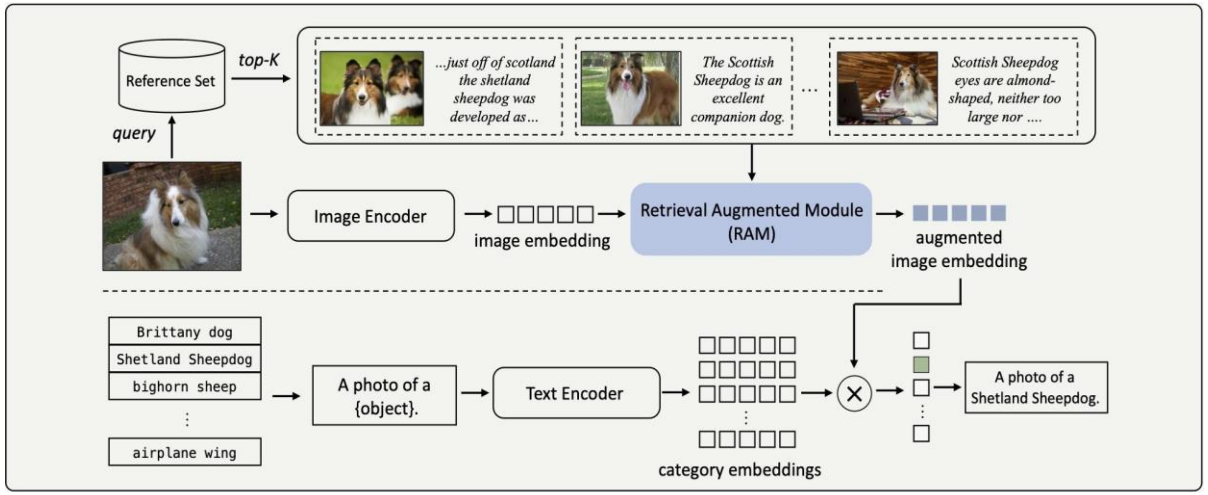
Prospects — Mult-Modality Extension

Transferring the concept of RAG from text to other modalities of data

Image

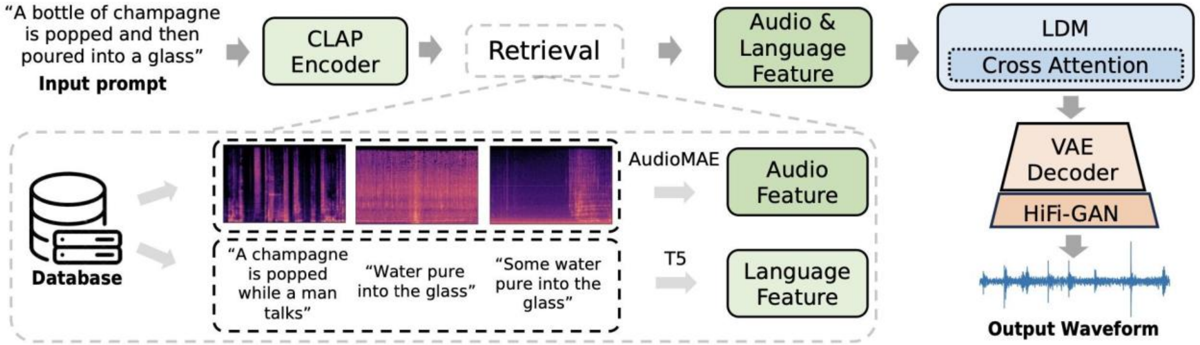

(a) Overview of Retrieval-Augmented Multimodal Model
Labrador sitting on bench near water.
Generator (Multimodal Model)
Retriever
Memory (Multimodal documents)
Labrador retriever Sitting on bench.
(b) Dense Multimodal Retriever
f(query, memory) → score
score
Cosine similarity
Mixed-modal Encoder
Mixed-modal Encoder
Query document (image or text or mixture)
Memory document (image or text or mixture)
E.g. Extension of CLIP
Mean Pooling
CLIP Text Encoder
CLIP Image Encoder
Input document
(c) Retrieval-Augmented Generator
Loss for the retrieved docs
Loss for the main doc
Transformer
Retrieved Doc 1
Retrieved Doc 2
Main Document
Labrador retriever sitting on bench.
Labrador retriever sitting by water.
Labrador sitting on bench near water.
Each image is mapped into 1024 tokens using VQGAN

RA-CM3 [Yasunaga et al.,2023]


Reference Set
top-K
query
...just off of scotland the shetland sheepdog was developed as ...
The Scottish Sheepdog is an excellent companion dog.
...
Scottish Sheepdog eyes are almond-shaped, neither too large nor ....
Image Encoder
image embedding
Retrieval Augmented Module (RAM)
augmented image embedding
Brittany dog
Shetland Sheepdog
bighorn sheep
airplane wing
A photo of a {object}.
Text Encoder
category embeddings
A photo of a Shetland Sheepdog.

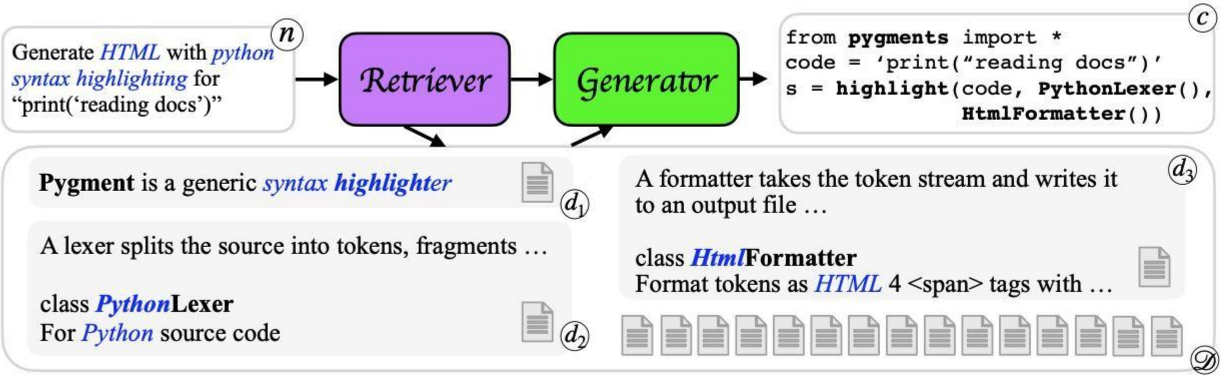
RA-CLIP [Xie et al.,2023]

Video


"A bottle of champagne is popped and then poured into a glass"
Input prompt
CLAP Encoder
Retrieval
Audio & Language Feature
LDM
Cross Attention
Database
AudioMAE
Audio Feature
"A champagne is popped while a man talks"
"Water pure into the glass"
"Some water pure into the glass"
T5
Language Feature
VAE Decoder
HiFi-GAN
Output Waveform

Re-AudioLDM [Yuan et al.,2023]

Code



Generate *HTML* with *python syntax highlighting* for "print('reading docs')"

```
from pygments import *
code = 'print("reading docs")'
s = highlight(code, PythonLexer(),
            HtmlFormatter())
```

Pygment is a generic *syntax* ***highlighter***

A lexer splits the source into tokens, fragments …

class ***Python*****Lexer**
For *Python* source code

A formatter takes the token stream and writes it to an output file …

class ***Html*****Formatter**
Format tokens as *HTML* 4 <span> tags with …

DocPrompting [Zhou et al.,2023]

TrustLLM
Funded by the European Union

References

1. Alon, U. et al. Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval.
2. Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
3. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M.-W. REALM: Retrieval-Augmented Language Model Pre-Training. Preprint at http://arxiv.org/abs/2002.08909 (2020).
4. Dai, Z. et al. Promptagator: Few-shot Dense Retrieval From 8 Examples. Preprint at http://arxiv.org/abs/2209.11755 (2022).
5. Izacard, G. et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. Preprint at http://arxiv.org/abs/2208.03299 (2022).
6. Gao, L., Ma, X., Lin, J. & Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. Preprint at http://arxiv.org/abs/2212.10496 (2022).
7. Muennighoff, N., Tazi, N., Magne, L. & Reimers, N. MTEB: Massive Text Embedding Benchmark. in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics 2014–2037 (Association for Computational Linguistics, 2023).
8. Ren, Y. et al. Retrieve-and-Sample: Document-level Event Argument Extraction via Hybrid Retrieval Augmentation. in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 293–306 (Association for Computational Linguistics, 2023).
9. Zhang, J. et al. ReAugKD: Retrieval-Augmented Knowledge Distillation For Pre-trained Language Models. in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 1128–1136 (Association for Computational Linguistics, 2023). 10. Khattab, O. et al. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. Preprint at http://arxiv.org/abs/2212.14024 (2023).
11. Cheng, X. et al. Lift Yourself Up: Retrieval-augmented Text Generation with Self Memory. Preprint at http://arxiv.org/abs/2305.02437 (2023).
12. Luo, Z. et al. Augmented Large Language Models with Parametric Knowledge Guiding. Preprint at http://arxiv.org/abs/2305.04757 (2023).
13. Shi, W. et al. REPLUG: Retrieval-Augmented Black-Box Language Models. Preprint at http://arxiv.org/abs/2301.12652 (2023).
14. Yu, Z., Xiong, C., Yu, S. & Liu, Z. Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In. Preprint at http://arxiv.org/abs/2305.17331 (2023).
15. Kang, M., Kwak, J. M., Baek, J. & Hwang, S. J. Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation. Preprint at http://arxiv.org/abs/2305.18846 (2023).
16. Trivedi, H., Balasubramanian, N., Khot, T. & Sabharwal, A. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. Preprint at http://arxiv.org/abs/2212.10509 (2023).
17. Wang, L., Yang, N. & Wei, F. Learning to Retrieve In-Context Examples for Large Language Models. Preprint at http://arxiv.org/abs/2307.07164 (2023).
18. Li, Z. et al. Towards General Text Embeddings with Multi-stage Contrastive Learning. Preprint at http://arxiv.org/abs/2308.03281 (2023).
19. Ng, Y. et al. SimplyRetrieve: A Private and Lightweight Retrieval-Centric Generative AI Tool. Preprint at http://arxiv.org/abs/2308.03983 (2023).
20. Huang, J. et al. RAVEN: In-Context Learning with Retrieval Augmented Encoder-Decoder Language Models. Preprint at http://arxiv.org/abs/2308.07922 (2023).

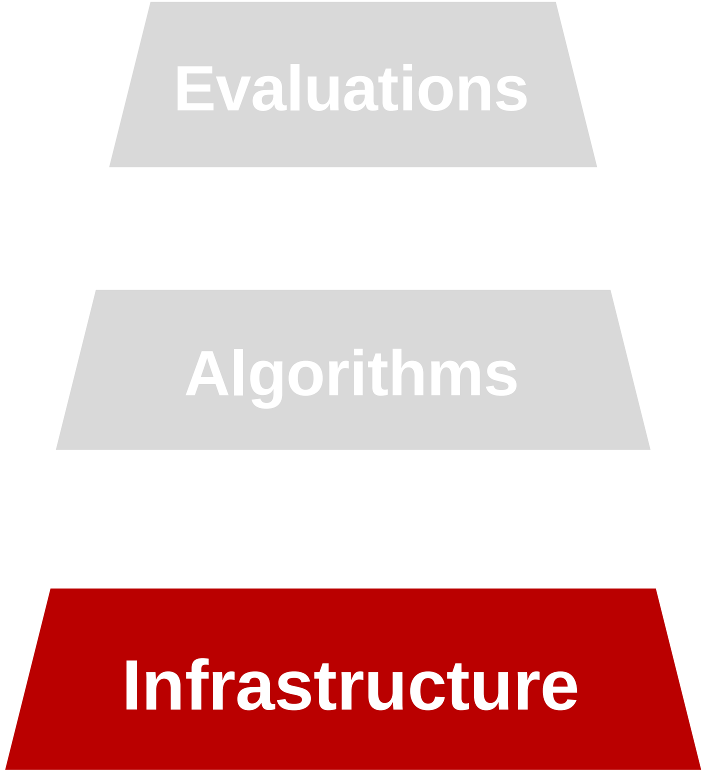
Funded by the European Union

References

21. Zhu, Y. et al. Large Language Models for Information Retrieval: A Survey. Preprint at http://arxiv.org/abs/2308.07107 (2023).
22. Wang, X. et al. KnowledGPT: Enhancing Large Language Models with Retrieval and Storage Access on Knowledge Bases. Preprint at http://arxiv.org/abs/2308.11761 (2023).
23. Chen, J., Lin, H., Han, X. & Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. Preprint at http://arxiv.org/abs/2309.01431
24. Es, S., James, J., Espinosa-Anke, L. & Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. Preprint at http://arxiv.org/abs/2309.15217 (2023).
25. Yoran, O., Wolfson, T., Ram, O. & Berant, J. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. Preprint at http://arxiv.org/abs/2310.01558 (2023).
26. Feng, Z., Feng, X., Zhao, D., Yang, M. & Qin, B. Retrieval-Generation Synergy Augmented Large Language Models. Preprint at http://arxiv.org/abs/2310.05149 (2023).
27. Zheng, H. S. et al. Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. Preprint at http://arxiv.org/abs/2310.06117 (2023).
28. Cheng, D. et al. UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation. Preprint at http://arxiv.org/abs/2303.08518 (2023).
29. Wang, B. et al. InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining. Preprint at http://arxiv.org/abs/2310.07713 (2023).
30. Jiang, Z. et al. Active Retrieval Augmented Generation. Preprint at http://arxiv.org/abs/2305.06983 (2023).
31. Gou, Q. et al. Diversify Question Generation with Retrieval-Augmented Style Transfer. Preprint at http://arxiv.org/abs/2310.14503 (2023).
32. Ma, X., Gong, Y., He, P., Zhao, H. & Duan, N. Query Rewriting for Retrieval-Augmented Large Language Models. Preprint at http://arxiv.org/abs/2305.14283 (2023).
33. Yang, H. et al. PRCA: Fitting Black-Box Large Language Models for Retrieval Question Answering via Pluggable Reward-Driven Contextual Adapter. Preprint at http://arxiv.org/abs/2310.18347 (2023).
34. Kim, G., Kim, S., Jeon, B., Park, J. & Kang, J. Tree of Clarifications: Answering Ambiguous Questions with Retrieval-Augmented Large Language Models. Preprint at http://arxiv.org/abs/2310.14696 (2023).
35. Shao, Z. et al. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. Preprint at http://arxiv.org/abs/2305.15294 (2023).
36. Zhang, P., Xiao, S., Liu, Z., Dou, Z. & Nie, J.-Y. Retrieve Anything To Augment Large Language Models. Preprint at http://arxiv.org/abs/2310.07554 (2023).
37. Purwar, A. & Sundar, R. Keyword Augmented Retrieval: Novel framework for Information Retrieval integrated with speech interface. Preprint at http://arxiv.org/abs/2310.04205 (2023).
38. Lin, X. V. et al. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. Preprint at http://arxiv.org/abs/2310.01352 (2023).
39. Yu, W. et al. Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models. Preprint at http://arxiv.org/abs/2311.09210 (2023).

Roadmap for more efficient & reliable retrieval-augmented LMs

Challenges of scaling up datastores & increased inference-time costs

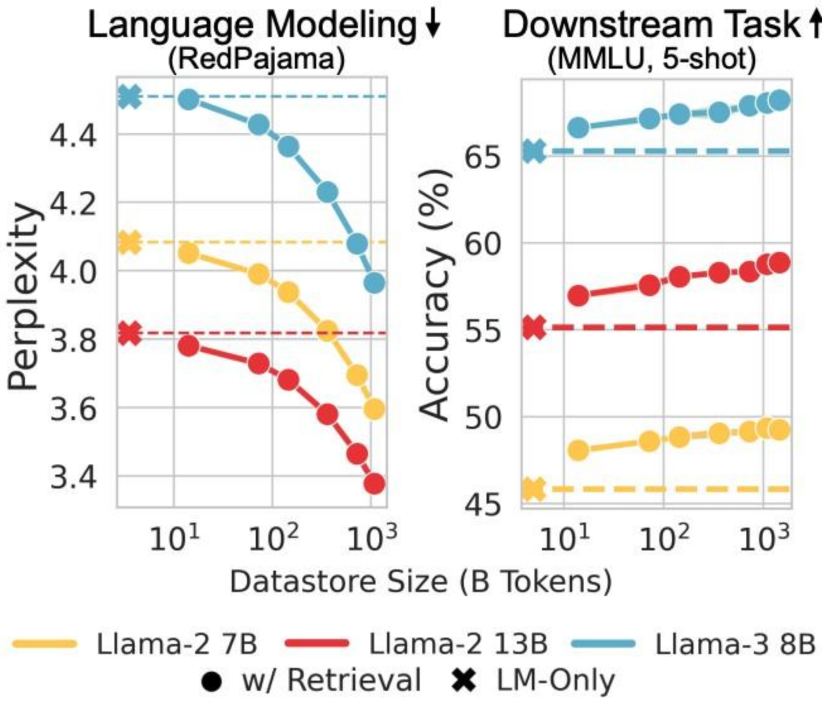


Evaluations

Algorithms

Infrastructure

- 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).

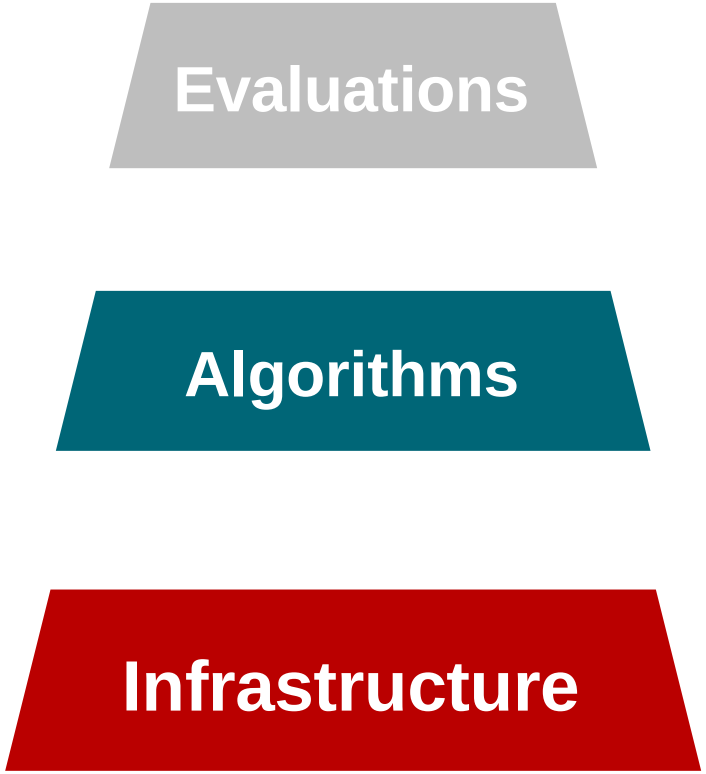


"Scaling Retrieval-Based Language Models with a Trillion-Token Datastore." Shao, He, Asai et al., ArXiv 2024.

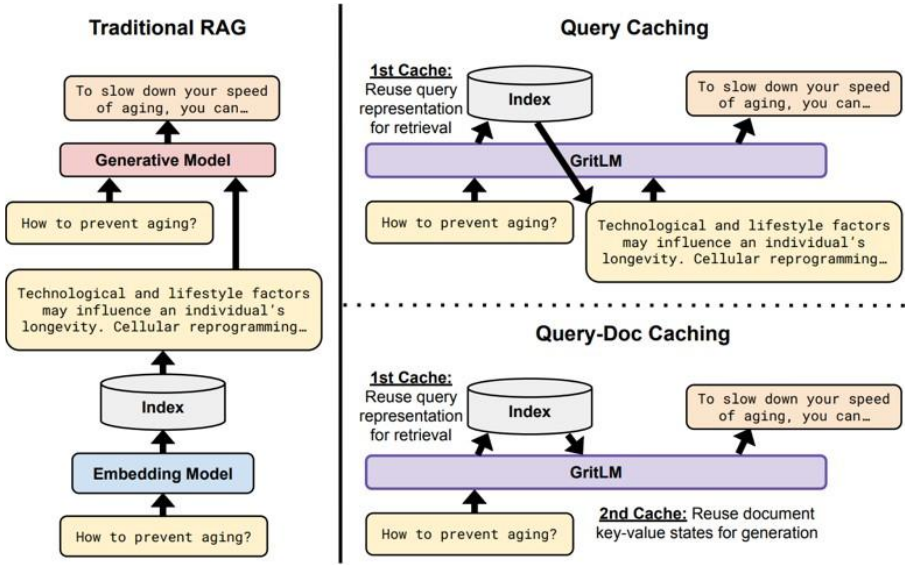
Funded by the European Union

Roadmap for more efficient & reliable retrieval-augmented LMs

New algorithms & arthitectures 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

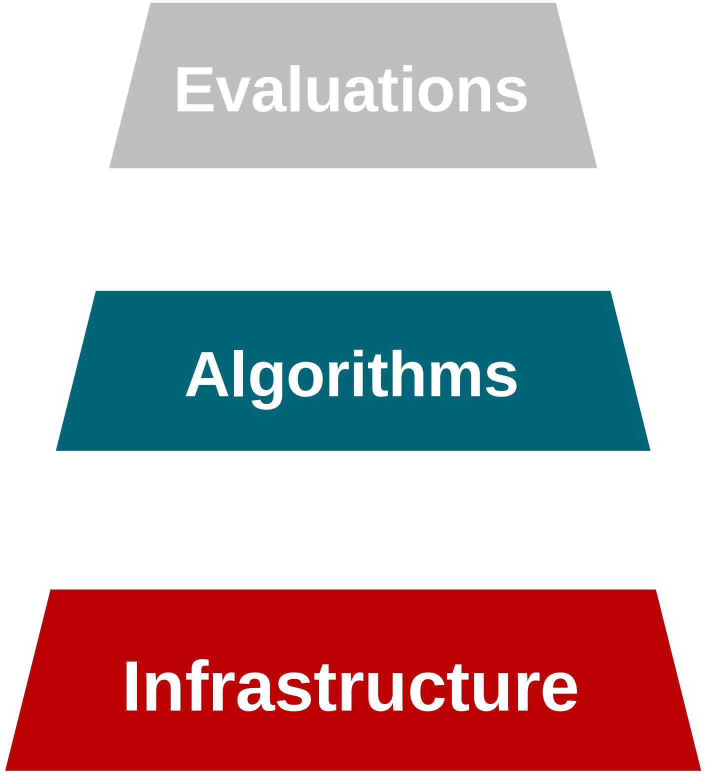

Traditional RAG
To slow down your speed of aging, you can…
Generative Model
How to prevent aging?
Technological and lifestyle factors may influence an individual's longevity. Cellular reprogramming…
Index
Embedding Model
How to prevent aging?
Query Caching
1st Cache: Reuse query representation for retrieval
Index
To slow down your speed of aging, you can…
GritLM
How to prevent aging?
Technological and lifestyle factors may influence an individual's longevity. Cellular reprogramming…
Query-Doc Caching
1st Cache: Reuse query representation for retrieval
Index
To slow down your speed of aging, you can…
GritLM
How to prevent aging?
2nd Cache: Reuse document key-value states for generation

"Generative Representational Instruction Tuning." Muennighoff et al., ArXiv 2024.

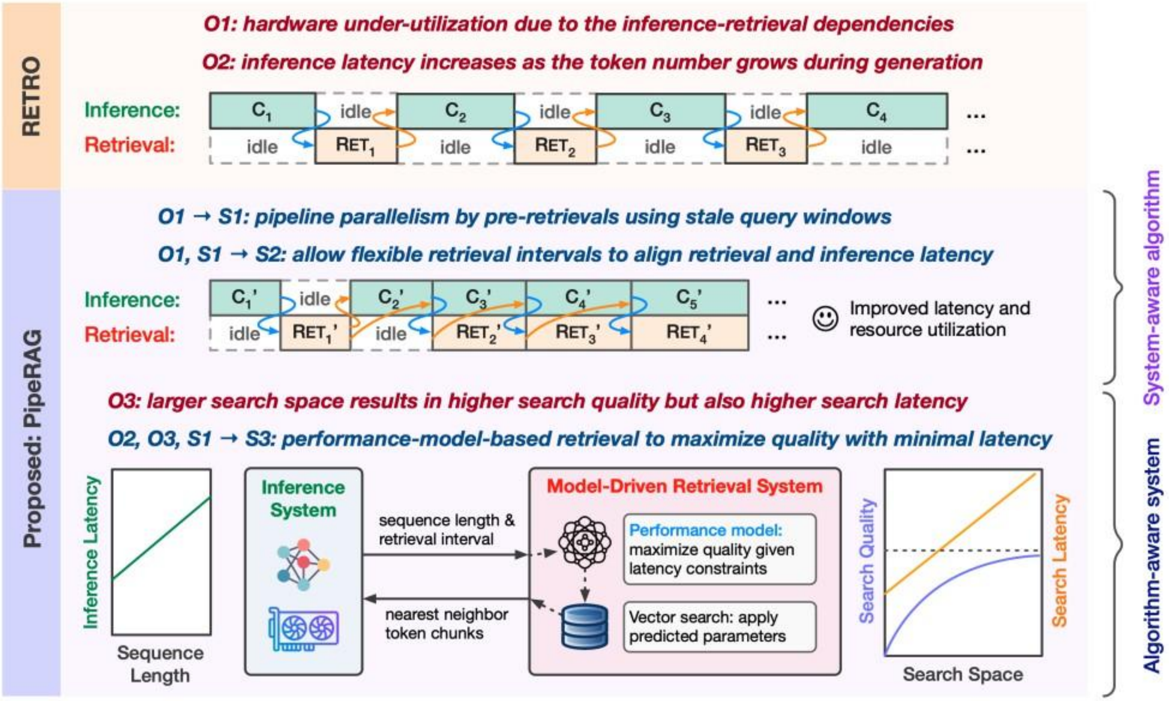
Funded by the European Union

Roadmap for more efficient & reliable retrieval-augmented LMs

New algorithms & arthitectures 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



O1: hardware under-utilization due to the inference-retrieval dependencies

O2: inference latency increases as the token number grows during generation

O1 → S1: pipeline parallelism by pre-retrievals using stale query windows

O1, S1 → S2: allow flexible retrieval intervals to align retrieval and inference latency

Improved latency and resource utilization

O3: larger search space results in higher search quality but also higher search latency

O2, O3, S1 → S3: performance-model-based retrieval to maximize quality with minimal latency

Performance model: maximize quality given latency constraints

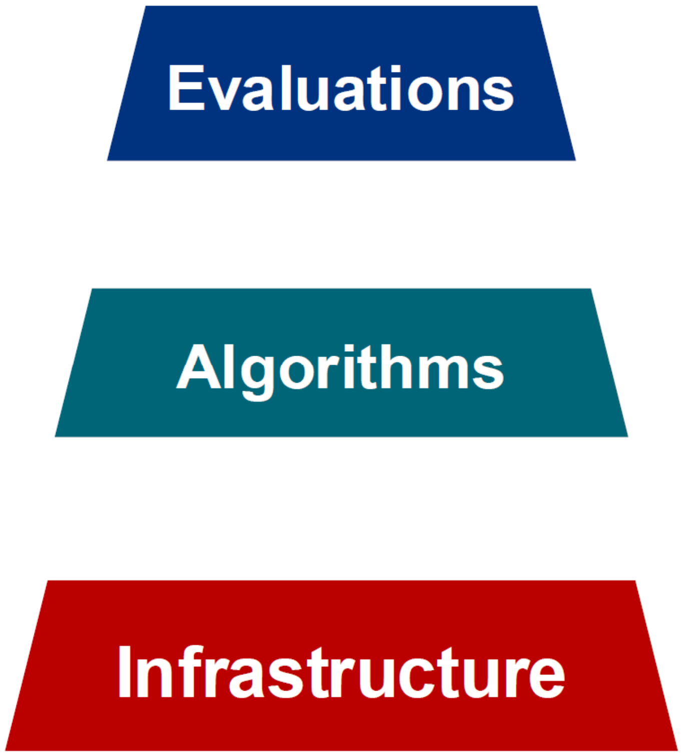
Vector search: apply predicted parameters

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

TrustLLM
Funded by the European Union

Roadmap for more efficient & reliable retrieval-augmen

Careful analyses on their effectiveness and limitations



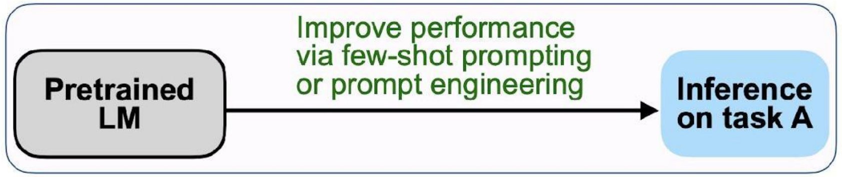
Prior systems are often evaluated only on simple general-domain tasks. Further exploration into their evaluation are needed

- 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

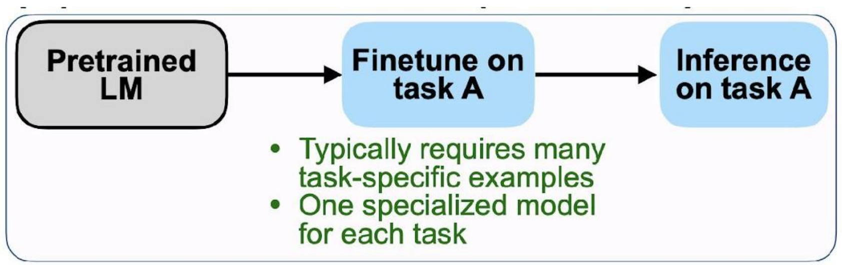
In-Context Learning

Strategies making a pre-trained LM do a task you care about:

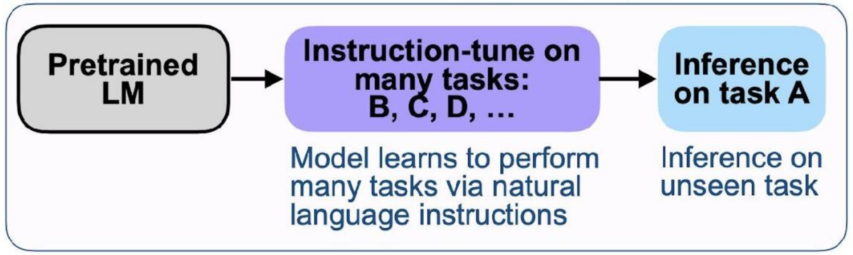
- In-context learning


Pretrained LM
Improve performance via few-shot prompting or prompt engineering
Inference on task A

- Full model finetuning → parameter-efficient finetuning


Pretrained LM
Finetune on task A
Inference on task A
• Typically requires many task-specific examples
• One specialized model for each task

- Multi-task finetuning → instruction finetuning → alignment training

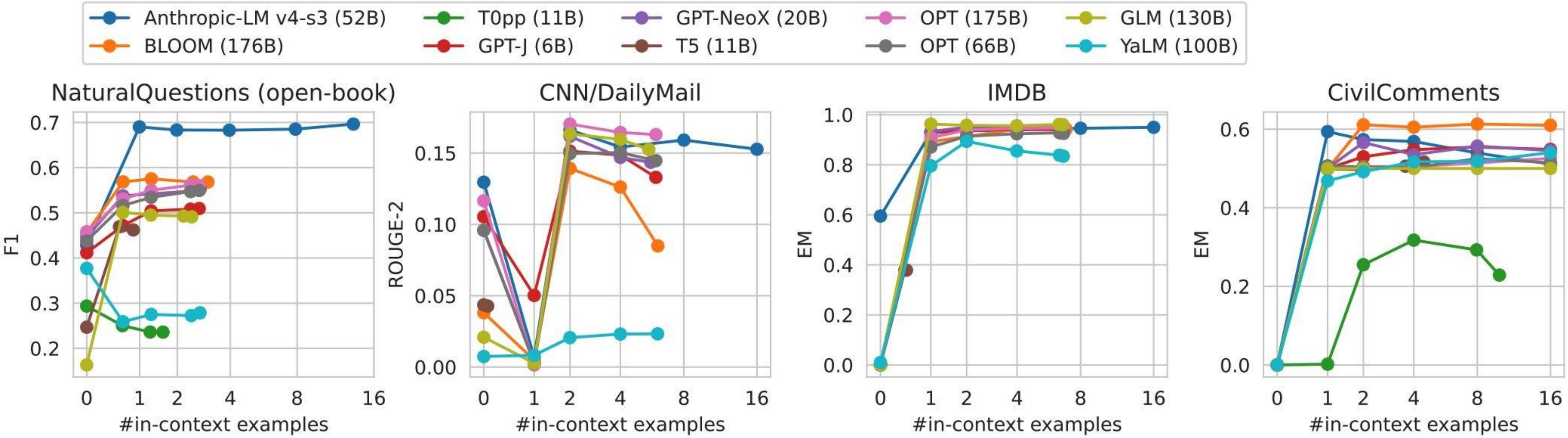

Pretrained LM
Instruction-tune on many tasks: B, C, D, …
Inference on task A
Model learns to perform many tasks via natural language instructions
Inference on unseen task

"Finetuned Language Models are Zero-Shot Learners." 2022. https://openreview.net/forum?id=gEZrGCozdqRa

Funded by the European Union

Improving In-Context Learning

Additional Examplars



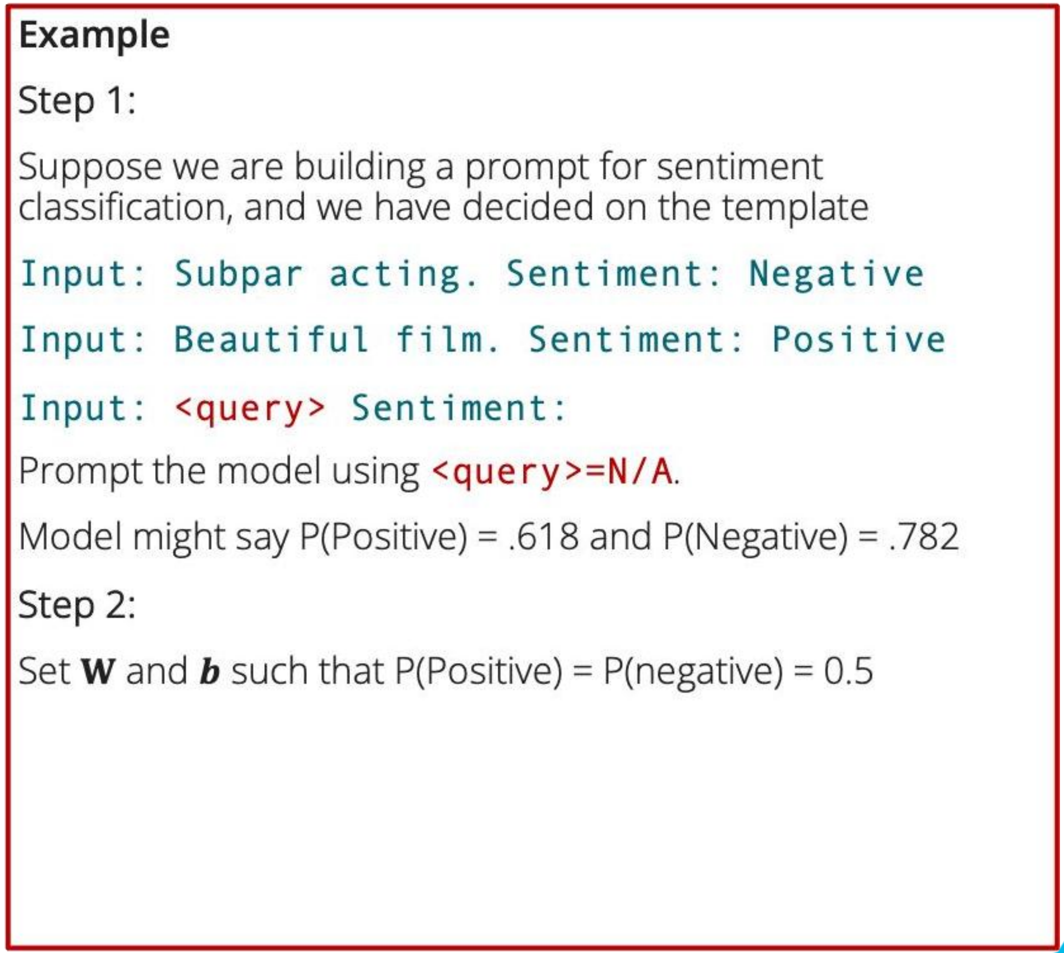
“Holistic Evaluation of Language Models.” Liang et al. 2022.

Funded by the European Union

Improving In-Context Learning

Calibrate Before Use

- Step 1: Estimate the bias
  - This does not require any labeled data.
  - For classification tasks, compute normalized scores of labels
  - For generation tasks: compute probabilities of the first token of the generation over the entire vocabulary
- Step 2: Counter the bias
  - "Calibrate" the model's predictions with an affine transformation of the logits.
  - $logits_{\text{calibrated}} = \text{softmax}(\mathbf{W}logits + \boldsymbol{b})$ where W is a diagonal matrix that scales each logit to reduce bias.
- *More details in paper linked below.*



Example

Step 1:

Suppose we are building a prompt for sentiment classification, and we have decided on the template

```
Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: <query> Sentiment:
```

Prompt the model using `<query>=N/A`.

Model might say P(Positive) = .618 and P(Negative) = .782

Step 2:

Set $\mathbf{W}$ and $\boldsymbol{b}$ such that P(Positive) = P(negative) = 0.5

"Calibrate Before Use: Improving Few-Shot Performance of Language Models." Zhao et al. 2021.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning

Intuition: An LLM will be better able to perform tasks (especially reasoning-based ones) if it is made to break down the task into multiple small steps.

Examples of reasoning-based tasks:

- Arithmetic:
  - “Fernando brings in three dozen bagels to a breakfast with 16 attendees. If each attendees eats two bagels, how many are left over?”
- Commonsense reasoning:
  - “The man had a fear of illness, so he never visited friends who were a what? (a) sick person (b) hospital (C) elderly person (d) graveyard.”

“Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning

Main idea: each of the exemplars in your few-shot prompt contains logic showing *how* to solve the task.

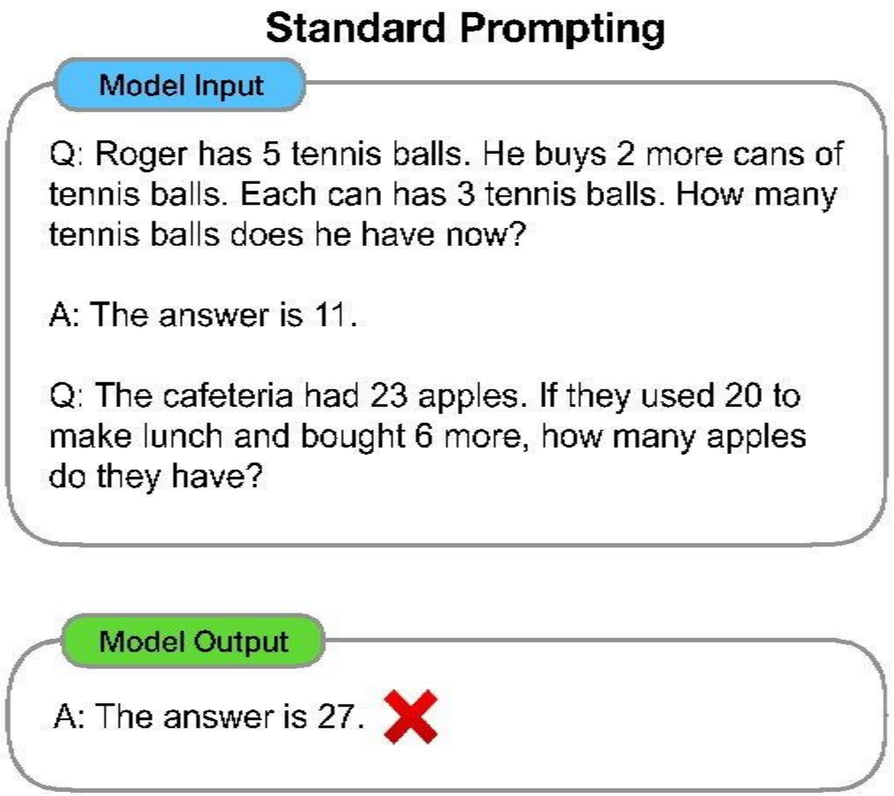
"Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Chain-of-Thought Exemplars

Main idea: each of the exemplars in your few-shot prompt should explain *how* to solve the task.



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

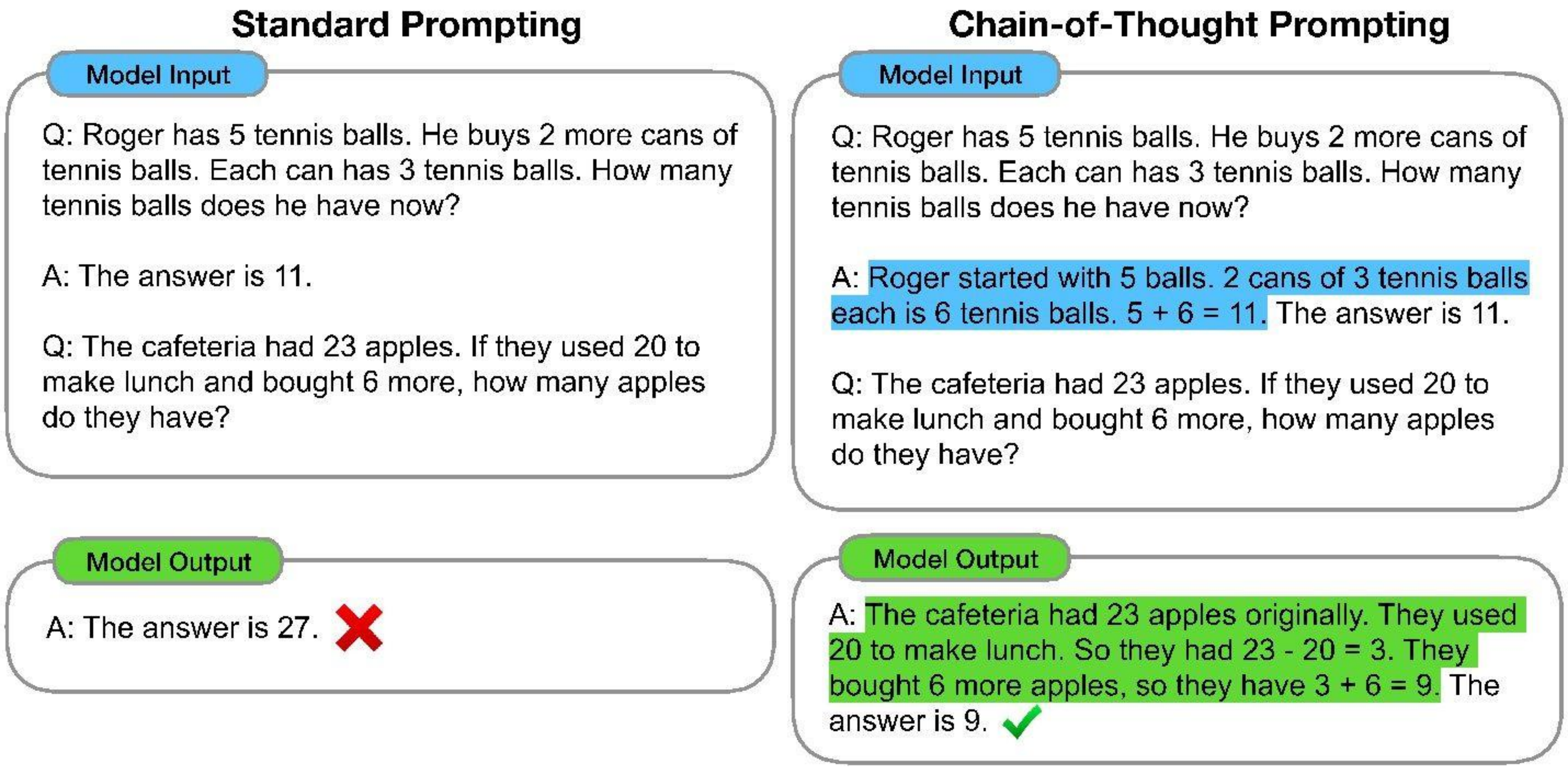
“Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Chain-of-Thought Exemplars

Main idea: each of the exemplars in your few-shot prompt should explain *how* to solve the task.



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

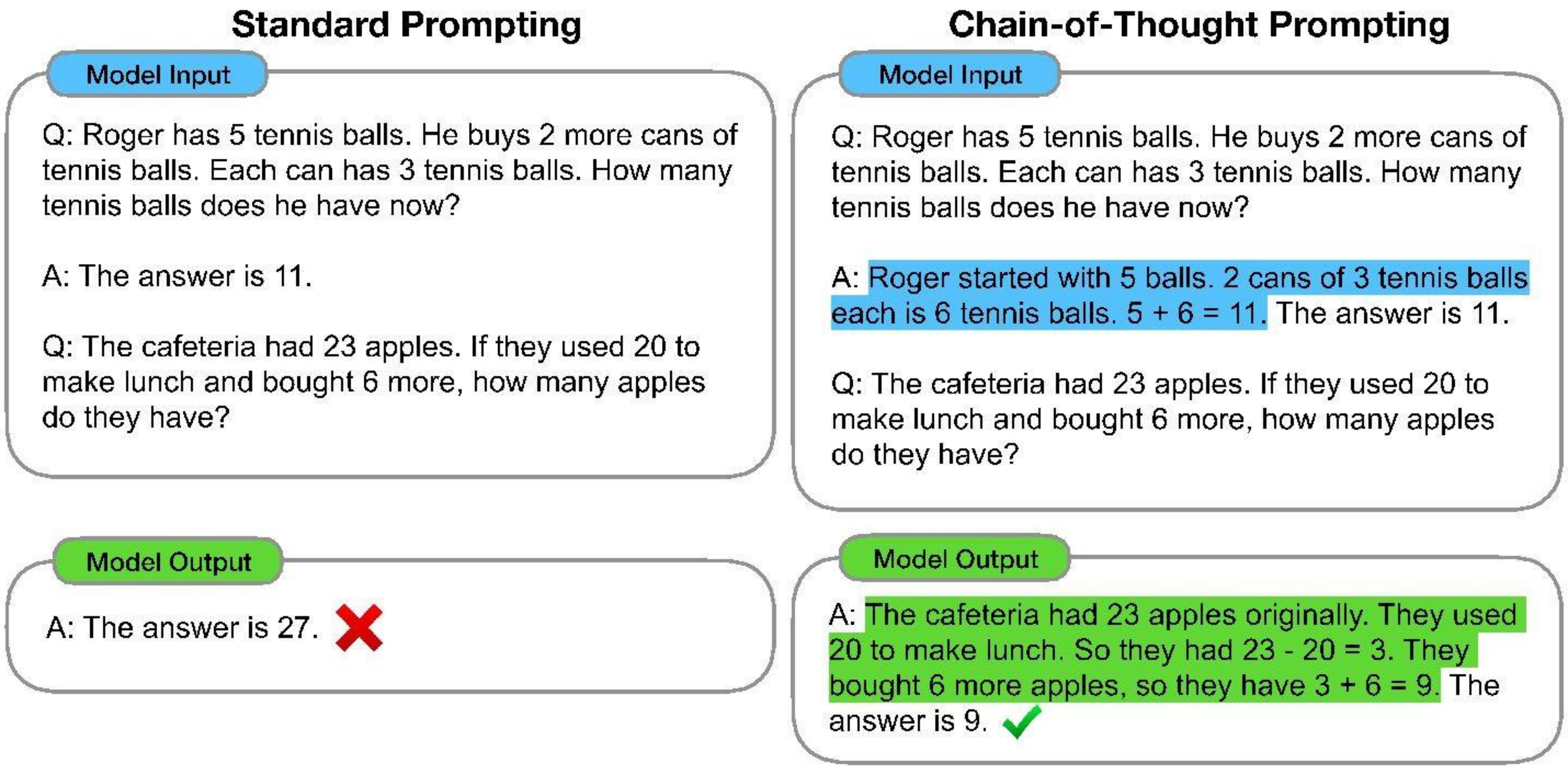
"Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Chain-of-Thought Exemplars

Main idea: each of the exemplars in your few-shot prompt should explain *how* to solve the task.



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

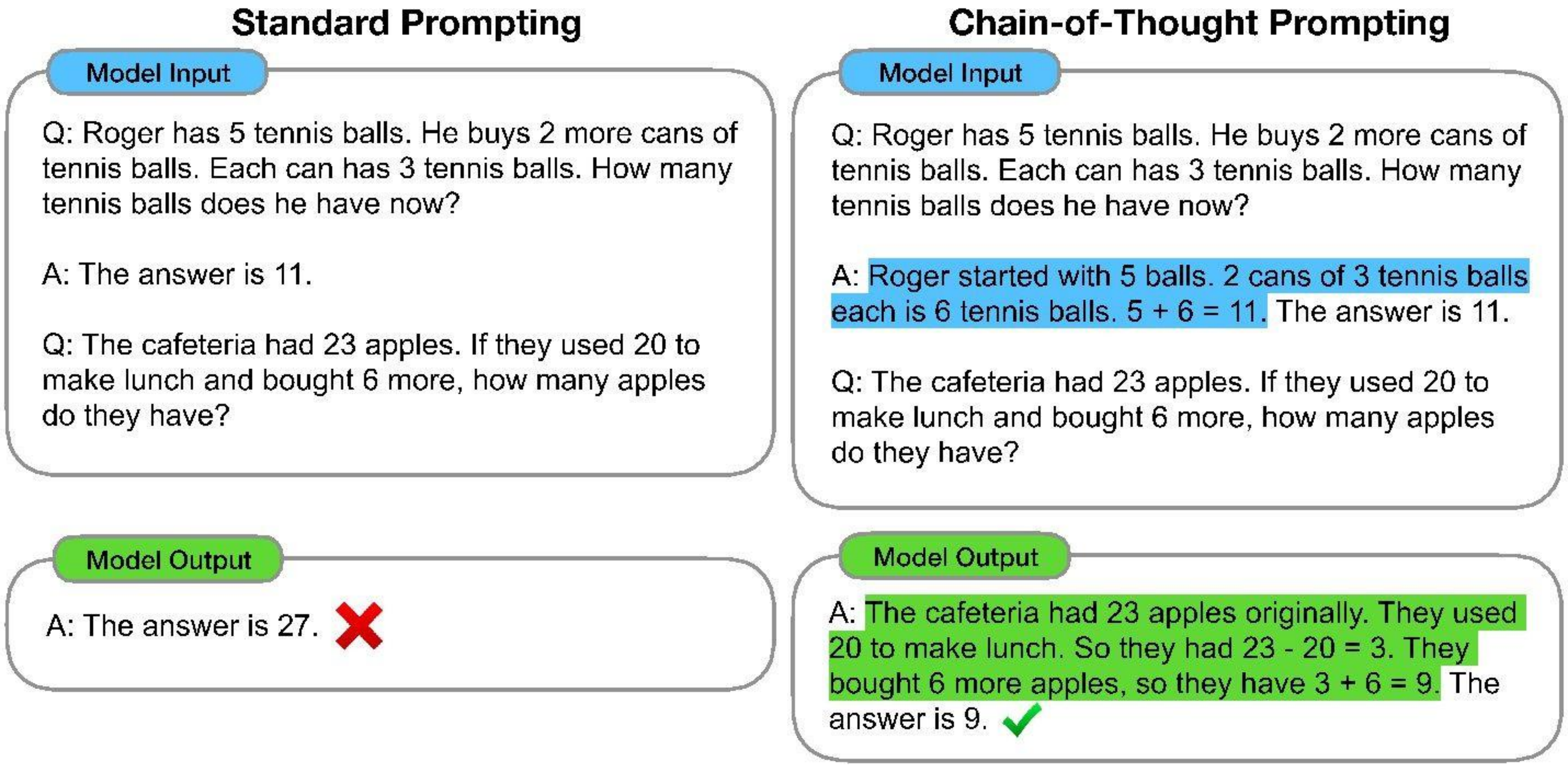
“Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Chain-of-Thought Exemplars

Main idea: each of the exemplars in your few-shot prompt should explain *how* to solve the task.



Standard Prompting

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The answer is 27. ✗

Chain-of-Thought Prompting

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

"Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don’t need any exemplars! Just append the string “Let’s think step by step.” to the end of the prompt.

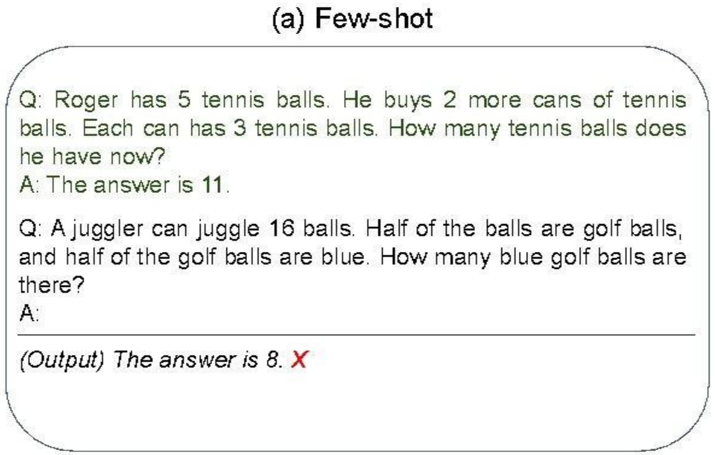
Large Language Models are Zero-Shot Reasoners.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don’t need any exemplars! Just append the string “Let’s think step by step.” to the end of the prompt.



(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. X

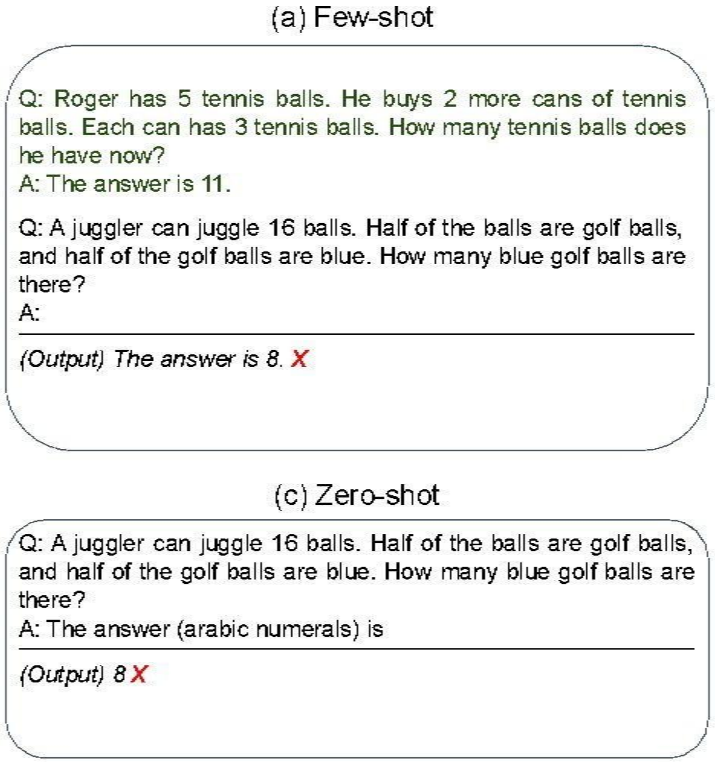
Large Language Models are Zero-Shot Reasoners.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don’t need any exemplars! Just append the string “Let’s think step by step.” to the end of the prompt.



(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. X

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 X

Large Language Models are Zero-Shot Reasoners.” NeurIPS 2022.

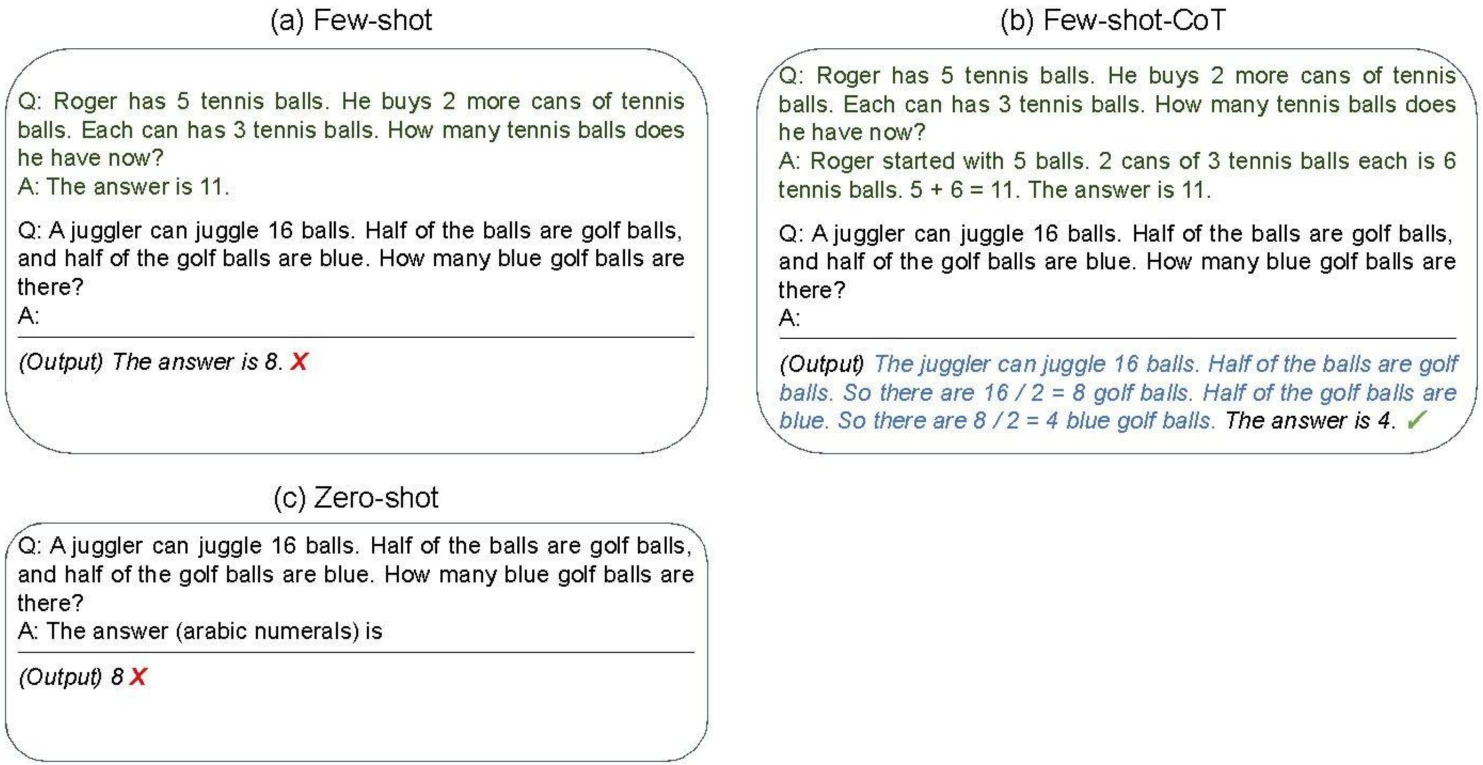
TrustLLM

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don’t need any exemplars! Just append the string “Let’s think step by step.” to the end of the prompt.



(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

TrustLLM

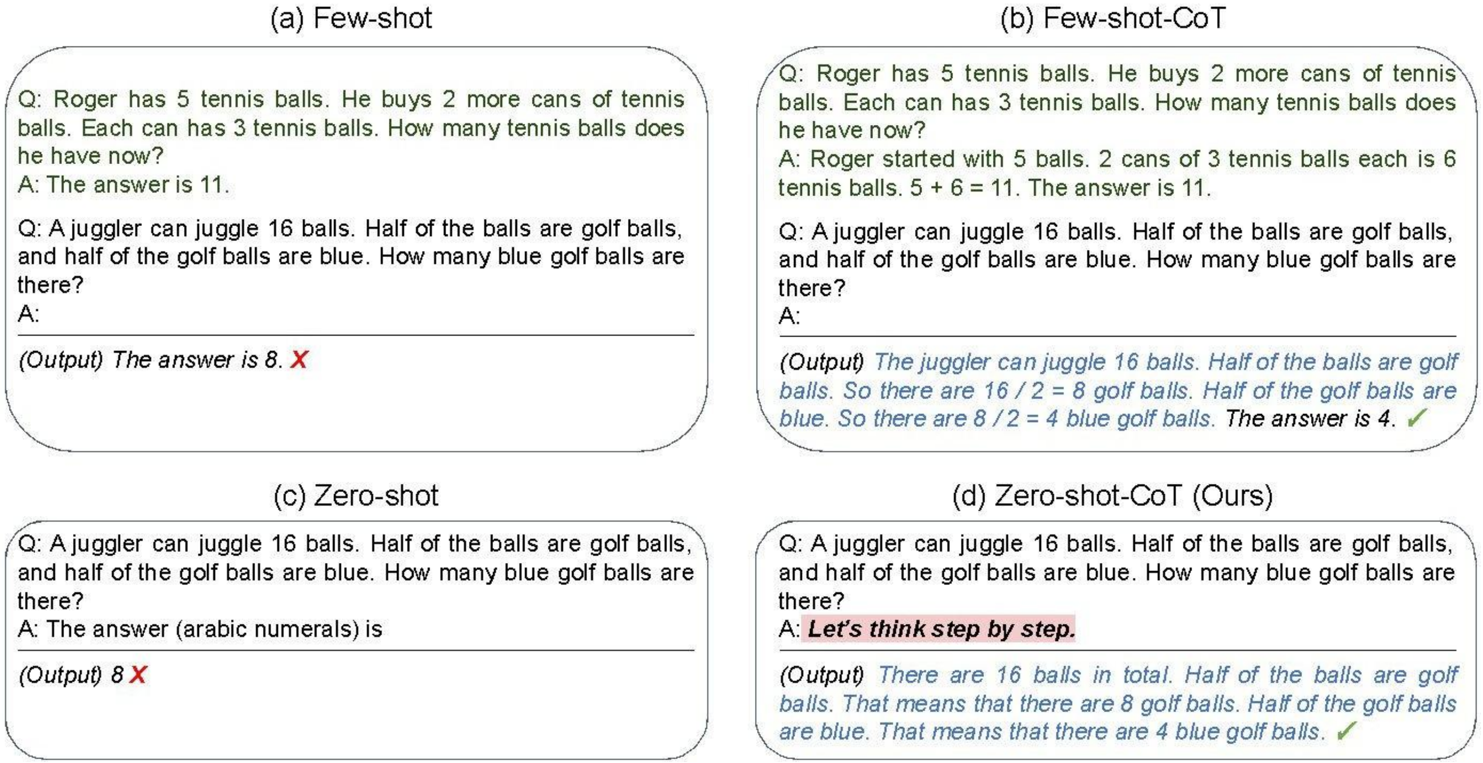
Large Language Models are Zero-Shot Reasoners.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.



(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. X

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 X

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

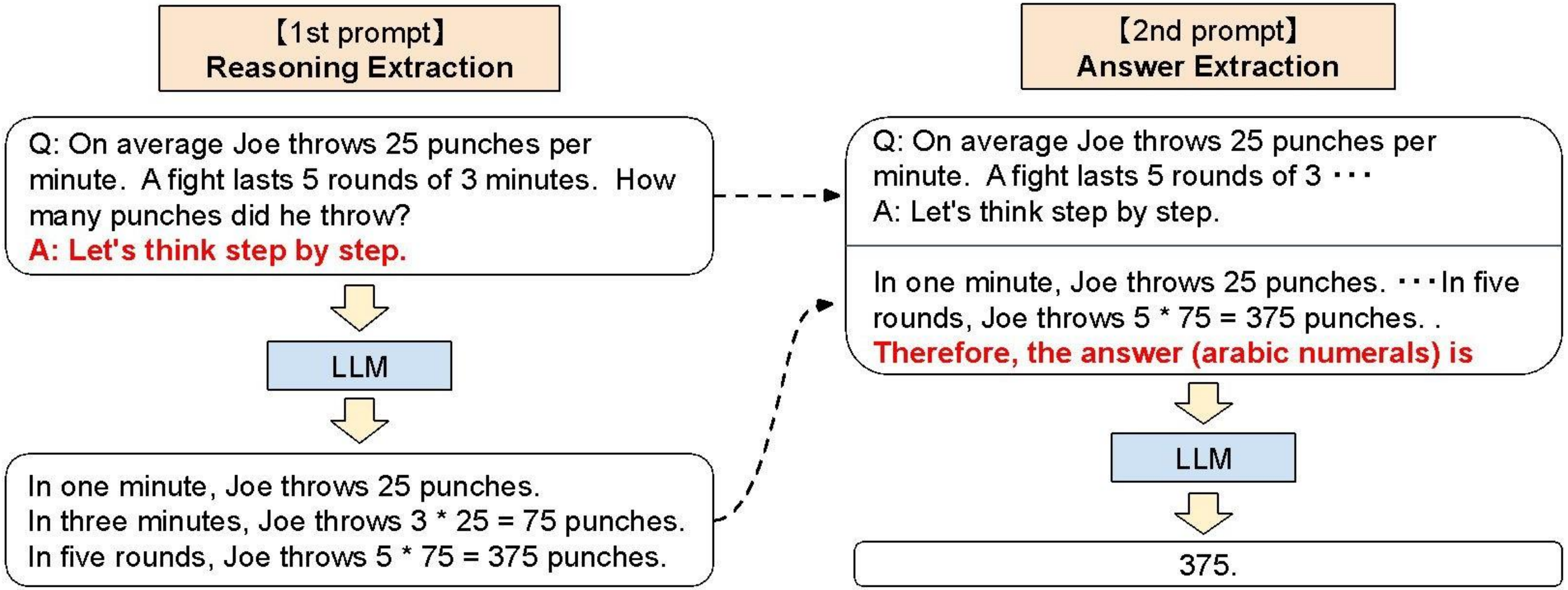
Large Language Models are Zero-Shot Reasoners." NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don't need any exemplars! Just append the string "Let's think step by step." to the end of the prompt.



【1st prompt】
Reasoning Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?
A: Let's think step by step.

In one minute, Joe throws 25 punches.
In three minutes, Joe throws 3 * 25 = 75 punches.
In five rounds, Joe throws 5 * 75 = 375 punches.

【2nd prompt】
Answer Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 ···
A: Let's think step by step.

In one minute, Joe throws 25 punches. ···In five rounds, Joe throws 5 * 75 = 375 punches. .
Therefore, the answer (arabic numerals) is

375.

Large Language Models are Zero-Shot Reasoners." NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought

Main idea: We don’t need any exemplars! Just append the string “Let’s think step by step.” to the end of the prompt.

Advantages over chain-of-thought (CoT) method:

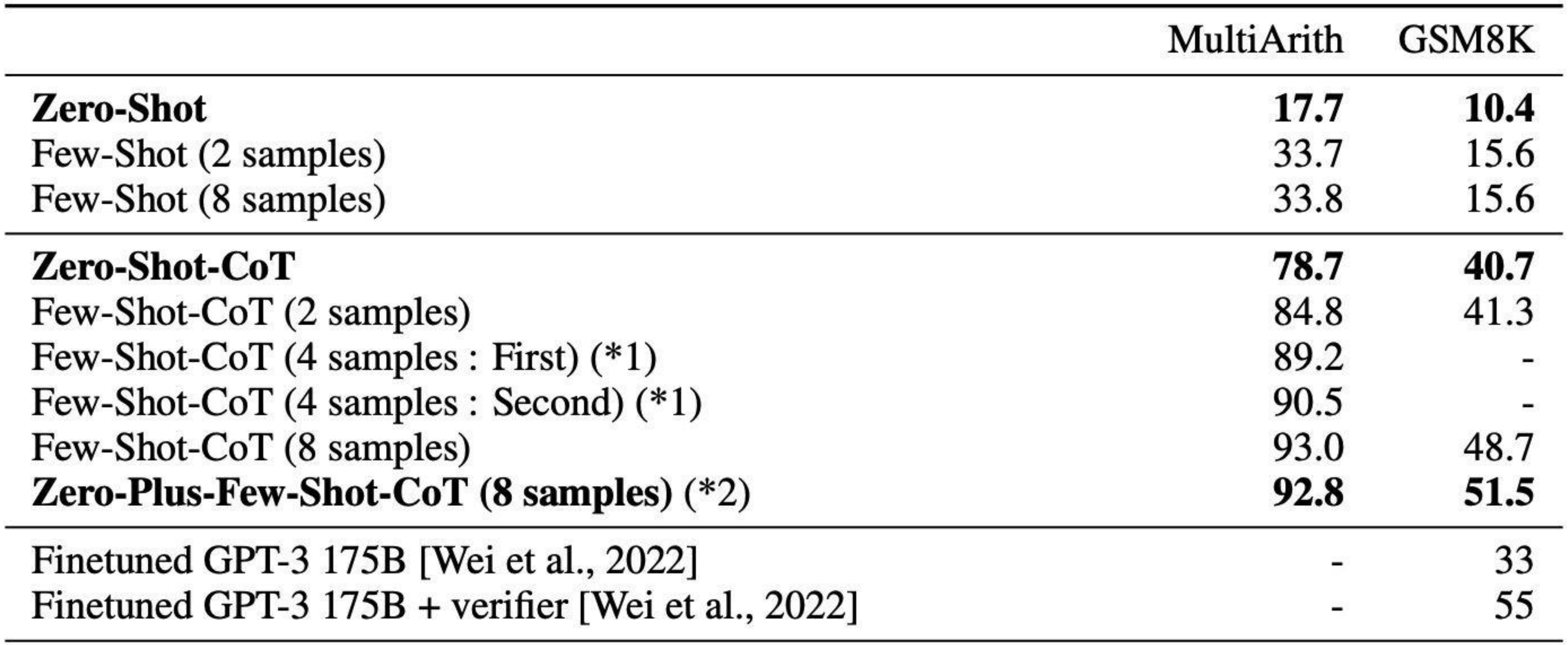
- The single fixed instruction “Let’s think step by step” works over a large variety of different tasks.
- Few-shot CoT performance degrades when there is misalignment between the example question types in the prompt and the actual task question.

Large Language Models are Zero-Shot Reasoners.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Multi-Step Reasoning with Zero-Shot Chain-of-Thought



	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	**92.8**	**51.5**
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55

Large Language Models are Zero-Shot Reasoners.” NeurIPS 2022.

Funded by the European Union

Improving In-Context Learning

Better Trained Models

As new generations of LLMs become increasingly instruction-tuned, the need for painstaking prompt engineering has decreased but not gone away entirely.

Funded by the European Union

Improving In-Context Learning

Better Trained Models

As new generations of LLMs become increasingly instruction-tuned, the need for painstaking prompt engineering has decreased but not gone away entirely.

Even today's "pre-trained" models have often been exposed to non-negligible amounts of instruction-following data.

Funded by the European Union

Improving In-Context Learning

Dividing Tasks into Minimal United

For complex generation tasks, many iterative calls to an LLM will generally work better (and be easier to evaluate) than one single prompt asking the LLM to do all parts of the task at once.

Funded by the European Union

Improving In-Context Learning

Dividing Tasks into Minimal United

For complex generation tasks, many iterative calls to an LLM will generally work better (and be easier to evaluate) than one single prompt asking the LLM to do all parts of the task at once.

Example: Generating short stories

You could ask an LLM to generate an entire story at once.

Or you could ask it to:

1. generate a synopsis
2. given the synopsis, generate a character list and a sequence of events
3. given all of the above, generate the actual story text.

Breaking the task into parts reduces the complexity of each individual call to the model and also allows more human intervention.

Funded by
the European Union

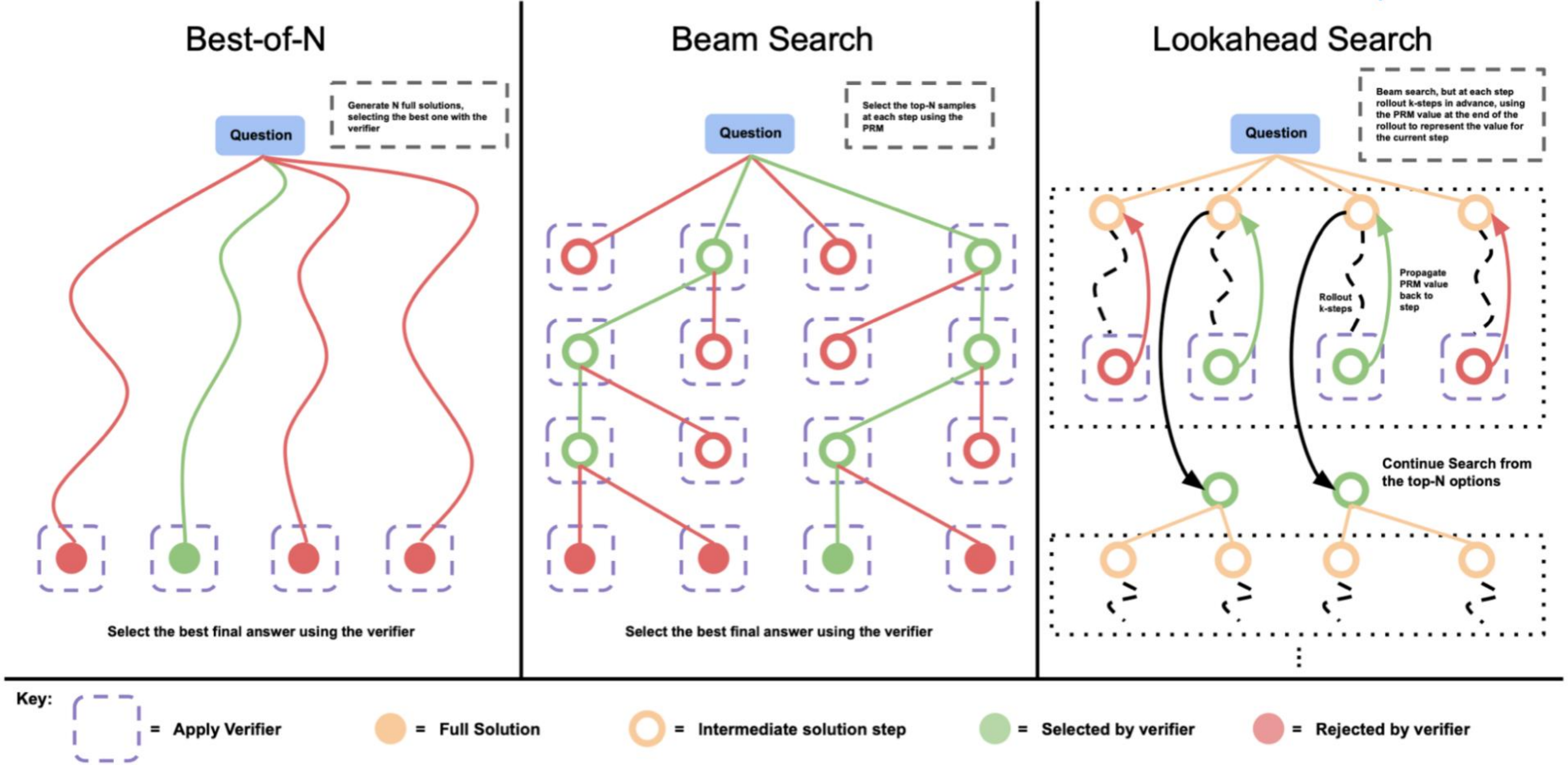
Reasoning

Towards Reasoning

LLM generated multiple responses, and a verifier model selects the best one

Uses an additional process-based reward model at each (token) generation step

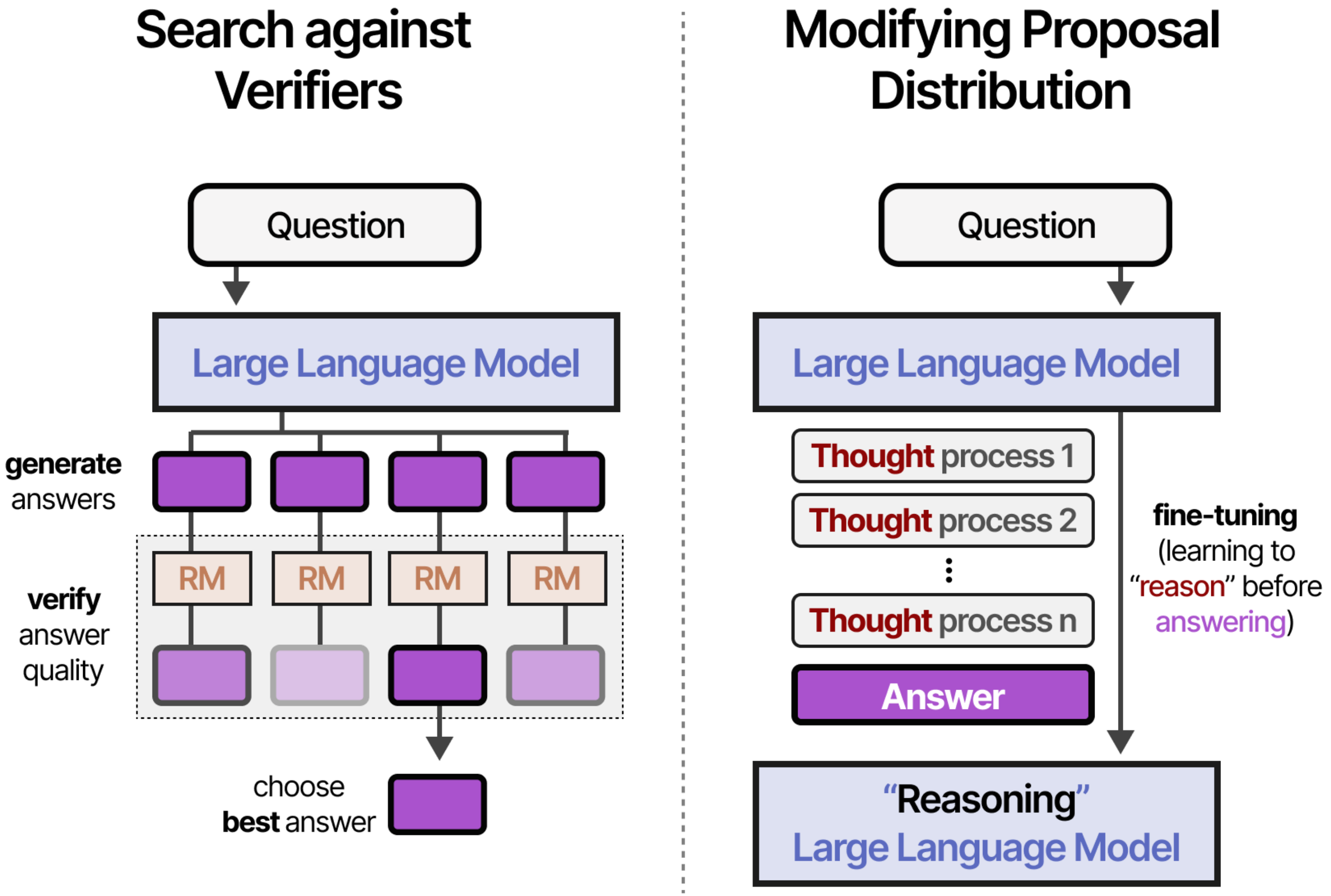
Uses a process-based reward model similar to beam search but includes a rollback step


Best-of-N
Beam Search
Lookahead Search
Question
Question
Question
Generate N full solutions, selecting the best one with the verifier
Select the top-N samples at each step using the PRM
Beam search, but at each step rollout k-steps in advance, using the PRM value at the end of the rollout to represent the value for the current step
Rollout k-steps
Propagate PRM value back to step
Continue Search from the top-N options
Select the best final answer using the verifier
Select the best final answer using the verifier
Key:
= Apply Verifier
= Full Solution
= Intermediate solution step
= Selected by verifier
= Rejected by verifier

https://sebastianraschka.com/blog/2025/understanding-reasoning-llms.html

Funded by the European Union

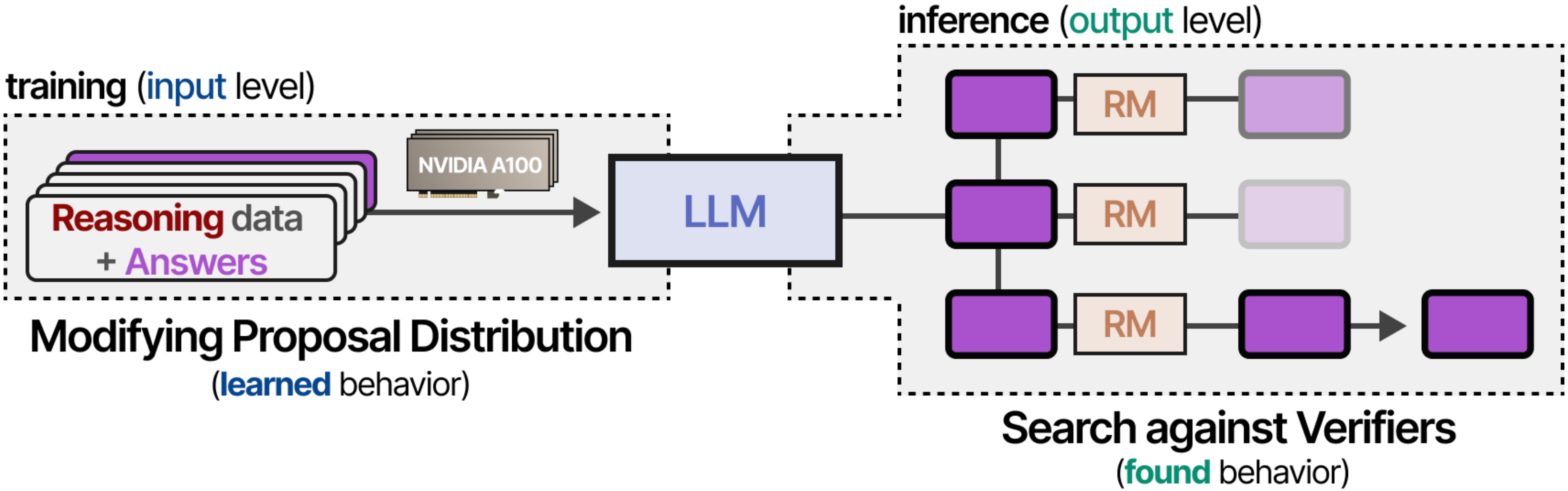
Test-Time Compute a.k.a. Reasoning



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Funded by the European Union

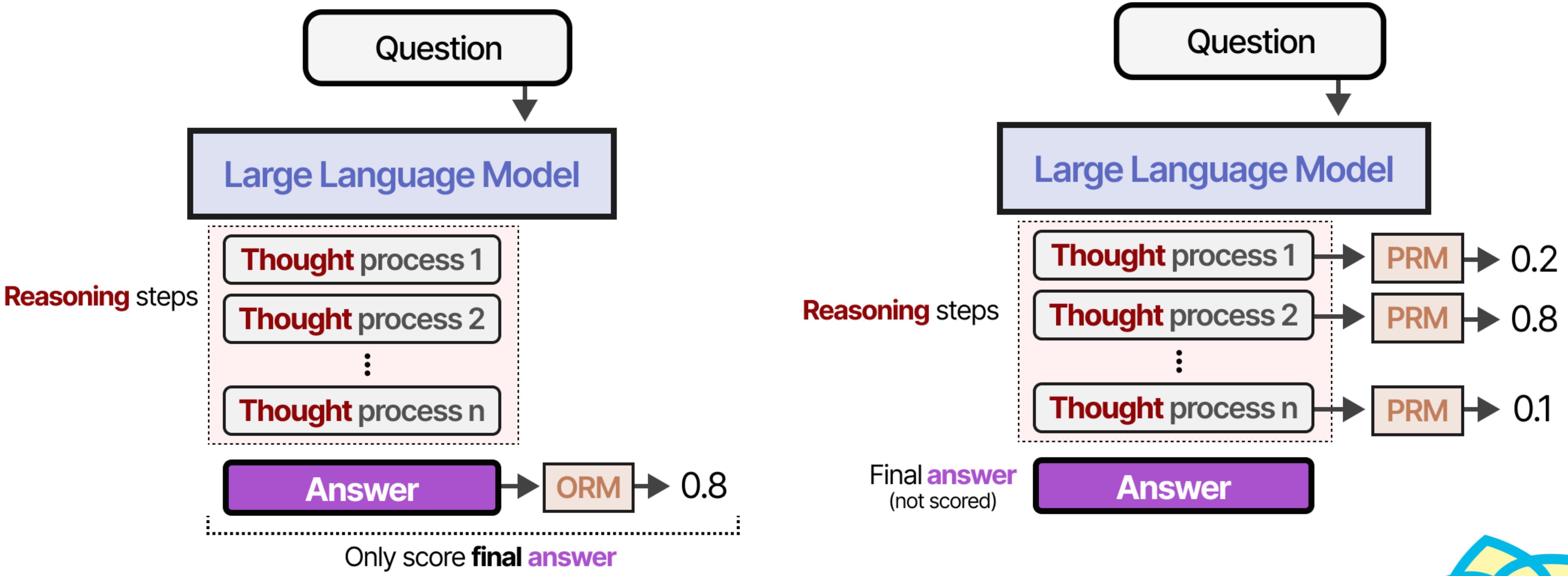
Test-Time Compute a.k.a. Reasoning



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Funded by the European Union

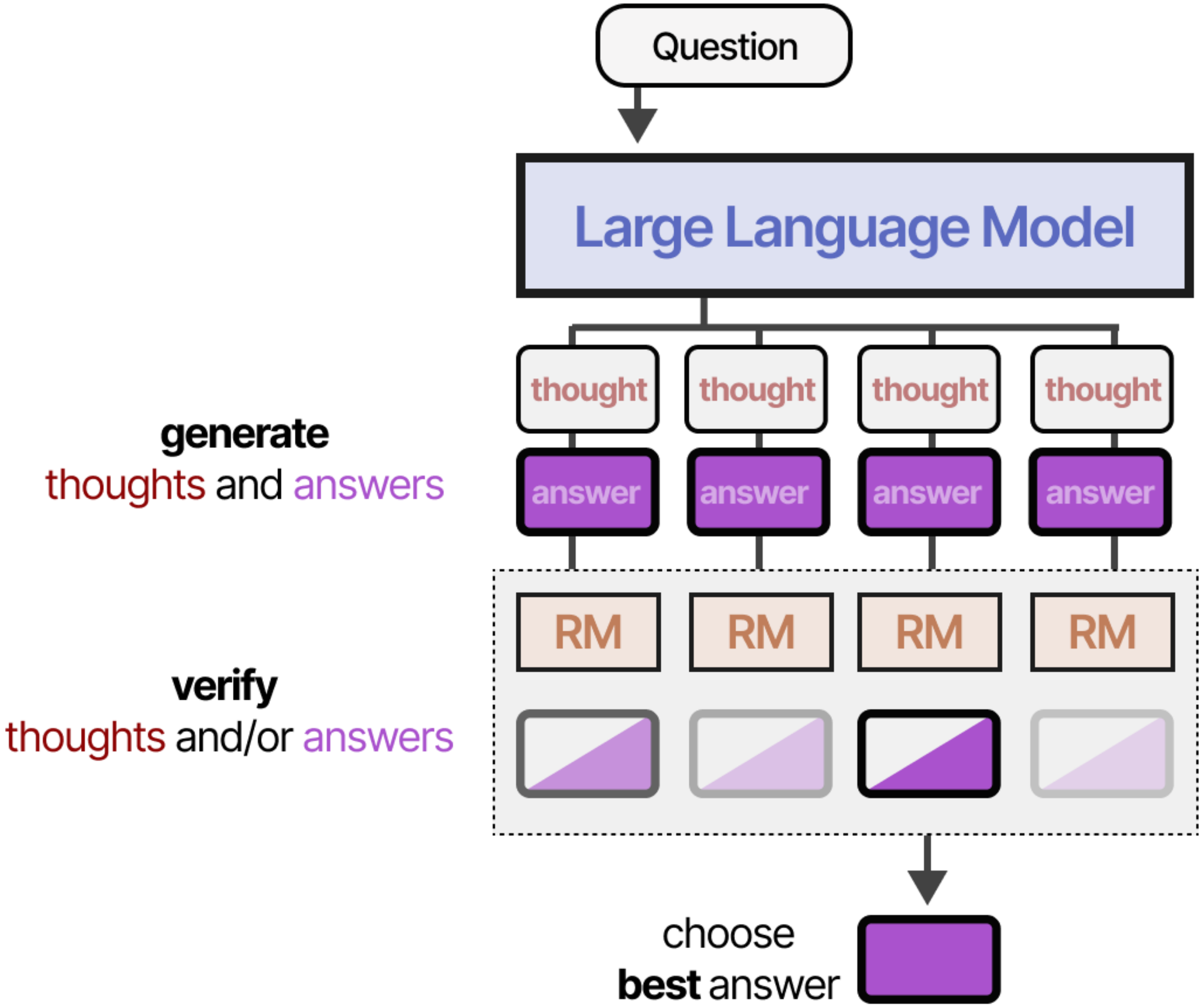
Outcome vs Process Reward Model



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Funded by the European Union

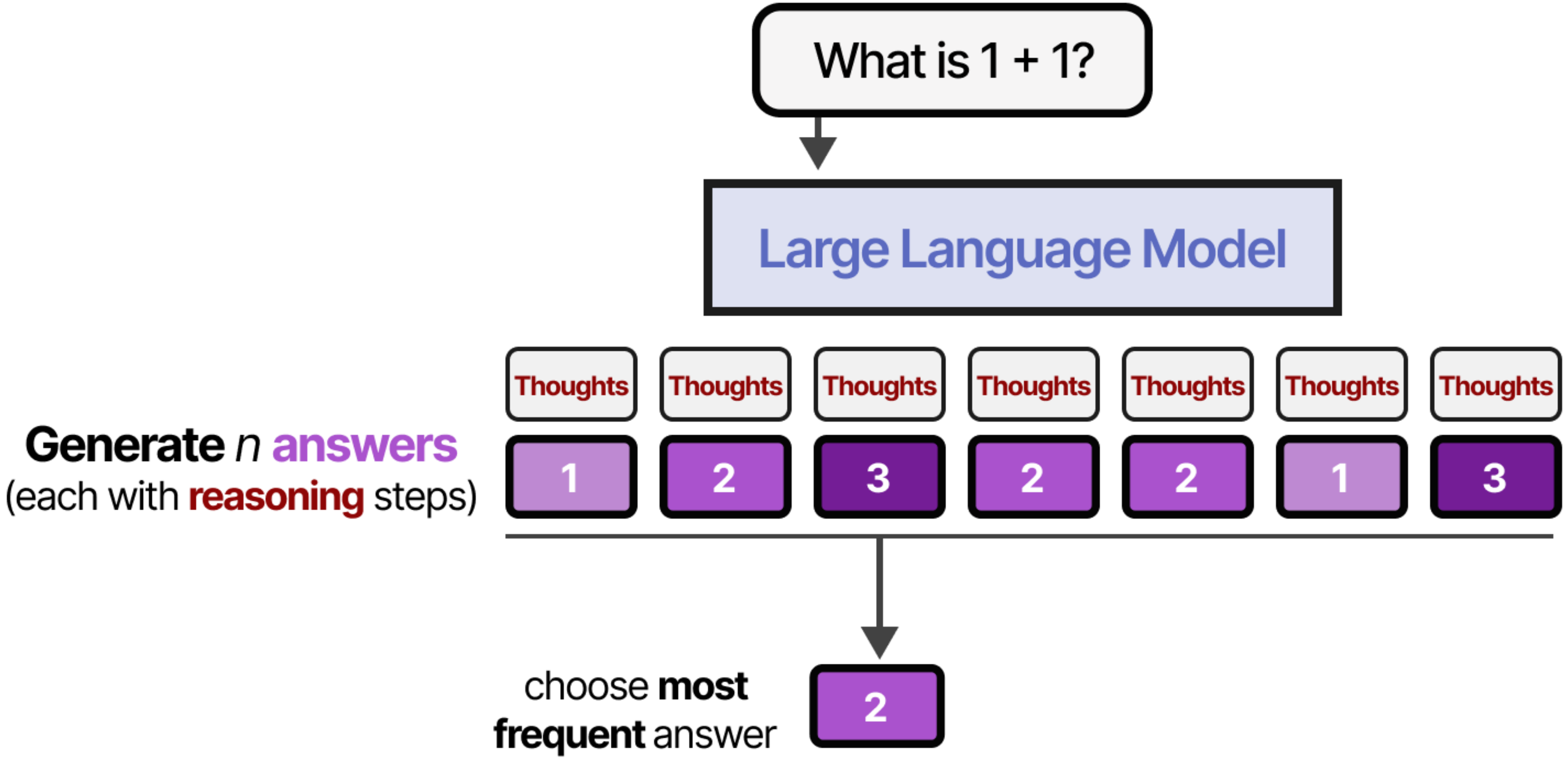
Search Against Verifiers



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Funded by the European Union

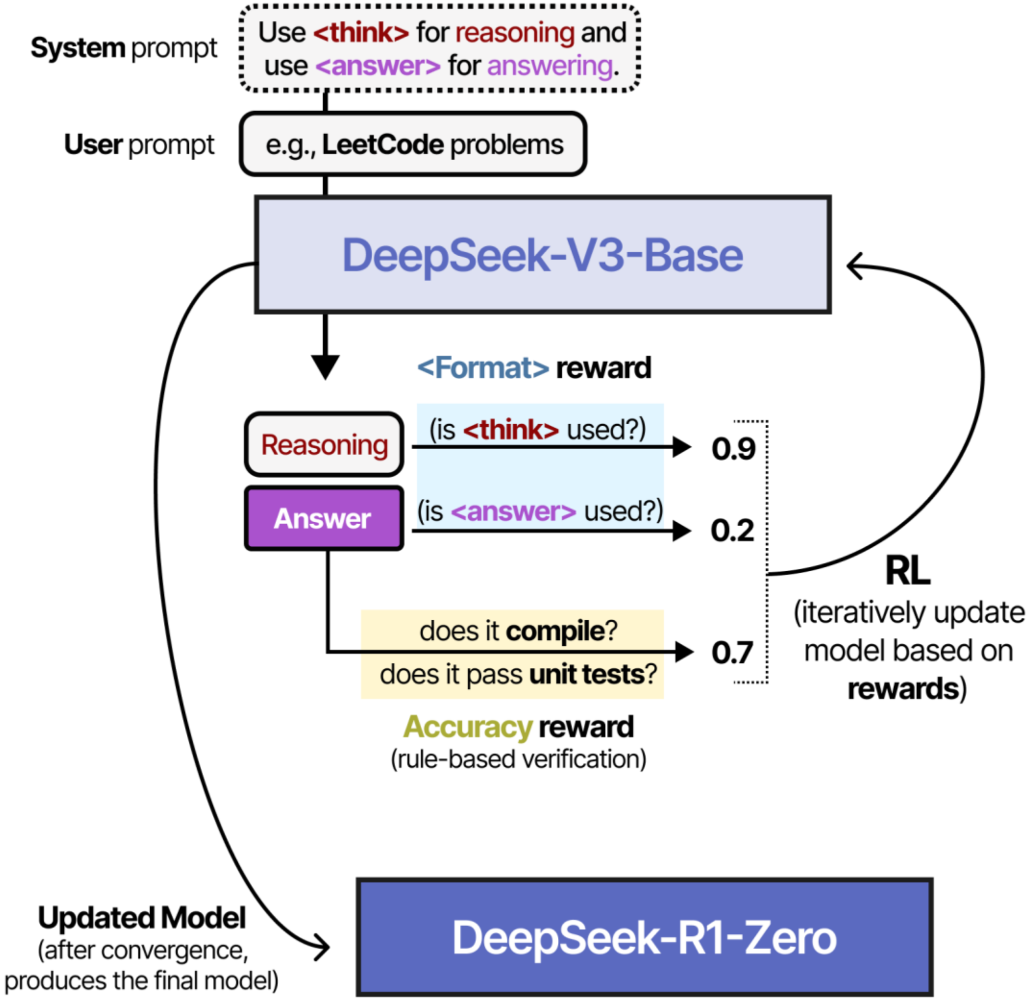
Majority Voting



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Funded by the European Union

Reasoning with DeepSeek-R1 Zero



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Funded by the European Union

- **LAIM LE6 VT2025:**

Inference
Retrieval Augmented Generation
In-Context Learning
Reasoning

www.ida.liu.se/~frehe08/llm

LINKÖPING
UNIVERSITY