Advancing state of the art of Retrieval-augmented Generation (RAG)

Akari Asai
University of Washington
https://akariasai.github.io/

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

UWNLP

April 18 | CMU CS 11-711

Hallucinations in LM Outputs

You

Explain Akari Asai, including the list of her famous papers. (*Febru

ChatGPT

Akari Asai is a prominent figure in the field of natural language processing significant contributions to various aspects of NLP research. She has mad particularly in the areas of question answering, dialogue systems, and mac...

Here are some of her famous papers:

1. "Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering" - In this paper, Asai and her co-authors proposed a method for learning to retrieve reasoning paths over the Wikipedia graph to improve question answering systems.
2. "Large-Scale Multi-Document Summarization Dataset and Evaluation Benchmark" - This paper introduces a large-scale dataset for multi-document summarization, along with an evaluation benchmark, providing valuable resources for advancing research in this area.

Akari Asai FOLLOW

Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model

Alexander R. Fabbri **Irene Li**
Tianwei She **Suyi Li** **Dragomir R. Radev**

Department of Computer Science
Yale University
{alexander.fabbri,irene.li,tianwei.she,suyi.li,dragomir.radev}@yale.edu

Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering 286 2020
A Asai, K Hashimoto, H Hajishirzi, R Socher, C Xiong
International Conference on Learning Representations (ICLR)

Catastrophic Errors as Results of LM Hallucinations

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 fine
ERIK MCGREGOR—LIGHTROCKET/GETTY IMAGES

MIT Technology Review

Featured Topics Newsletters Events Podcasts SIGN IN

ARTIFICIAL INTELLIGENCE

Why Meta’s latest large language model survived only three days online

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

By Will Douglas Heaven November 18, 2022

Air Canada must honor re invented by airline’s chatb

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

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

Retrieval-augmented Generations (RAG)

Step 1: Retrieve K documents

Prompt How did US states get their names?

Retriever

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

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

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

Retrieve

Step 2: Prompt LM with K docs and generate

Prompt How did US states get their names? ❶❷❸

LM

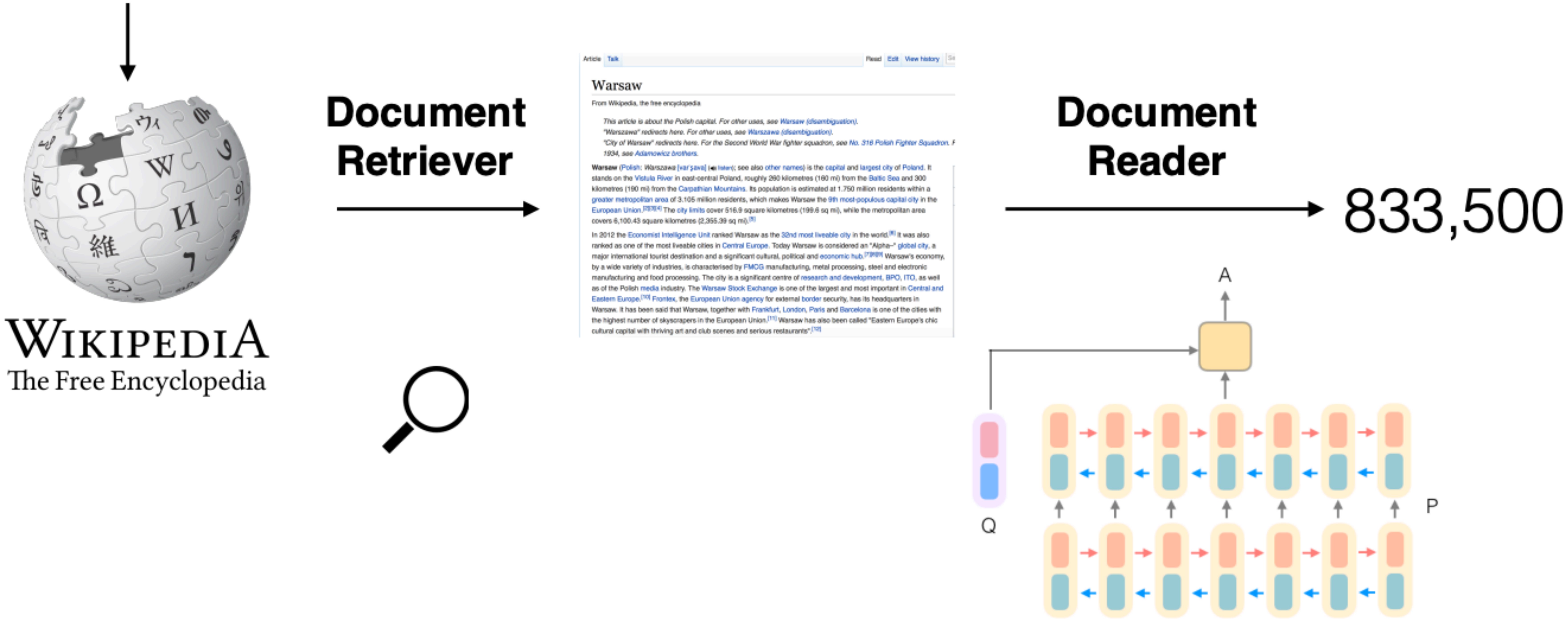
US states got their names from a variety of sources. Eleven states are named after an individual person (e.g, California was named after Christopher Columbus). Some states including Texas and Utah, are named after …

Read

Retrieval-augmented Generations (RAG)

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

Q: How many of Warsaw's inhabitants spoke Polish in 1933?



Retrieve

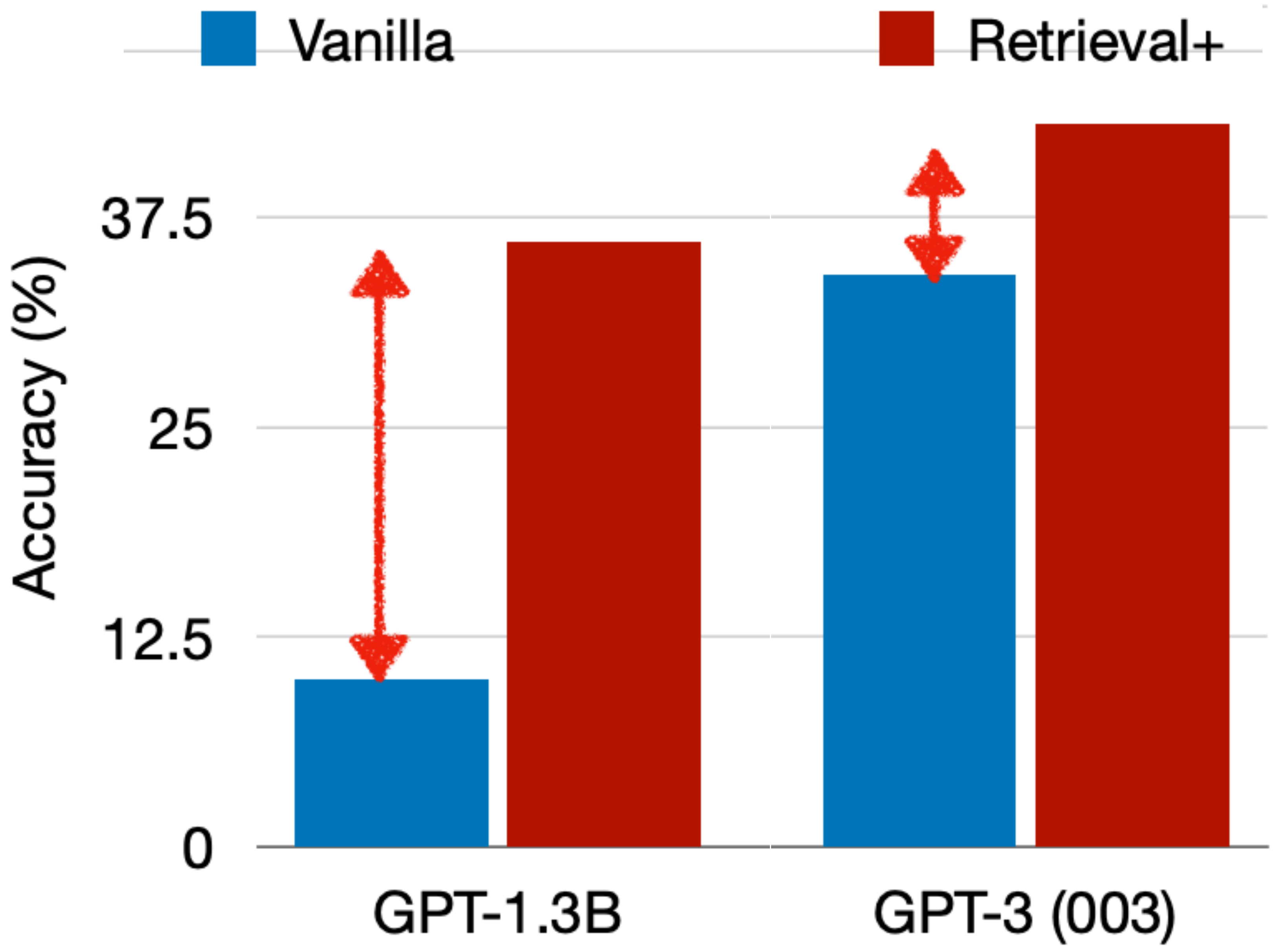
DrQA(Chen et al., 2017)

Read

Success of Retrieval-augmented Generation (RAG)

RAG has shown effective in many benchmarks

QA



Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023.

Application to commercial systems

Perplexity

Search & discovery with AI

Shift from traditional QA to *open-ended* instructions

How many of US states got their names from individual person?

WIKIPEDIA
The Free Encyclopedia

Of the fifty states, eleven are named after an individual person. Six of those are named in honor of European monarchs: the two Carolinas, the two Virginias, Georgia, and Louisiana.

Eleven

Simple factoid question

Short answer based on single document

Shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states get their names, grouping them together.

Group	States
Named after indigenous communities	Alabama, Alaska, Arizona, Arkansas, Colorado, Connecticut, Delaware, Florida, Hawaii, Idaho, Illinois, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Mexico, North and South Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming
Named after royalty and other important figures	Georgia, Louisiana, Maryland, New York, North and South Carolina, Pennsylvania, Virginia, Washington, West Virginia

10 references

WIKIPEDIA
LIST OF STATE AND TERRITORY NAME ETY...

BIA
ORIGIN OF NAMES OF US STATES | INDIA...

MENTALFLOSS
HOW ALL 50 STATES GOT THEIR NAMES | ...

⌘R

cohere

Shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states get their names, grouping them together.

Group	States
Named after indigenous communiti es	Alabama, Alaska, Arizona, Arkansas, Colorado, Connecticut, Delaware, Florida, Hawaii, Idaho, Illinois, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Mexico, North and South Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming
Named after royalty and other important figures	Georgia, Louisiana, Maryland, New York, North and South Carolina, Pennsylvania, Virginia, Washington, West Virginia

Only 10 states here…?

10 references

WIKIPEDIA
LIST OF STATE AND TERRITORY NAME ETY...

BIA
ORIGIN OF NAMES OF US STATES | INDIA...

MENTALFLOSS
HOW ALL 50 STATES GOT THEIR NAMES | ...

cohere

Is this fully correct?

Shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states get their names, grouping them together.

Group	States
Named after indigenous communities	Alabama, Alaska, Arizona, Arkansas, Colorado, Connecticut, Delaware, Florida, Hawaii, Idaho, Illinois, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Mexico, North and South Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming
Named after royalty and other important figures	Georgia, Louisiana, Maryland, New York, North and South Carolina, Pennsylvania, Virginia, Washington, West Virginia

DELAWARE: Named for Lord De La Warr, first governor and captain-general of Virginia, who in 1630 explored the bay and river area where his name was first applied.

Indian Affairs (.gov)
https://www.bia.gov › as-ia › opa › online-press-release

Origin of Names of US States | Indian Affairs

10 references

WIKIPEDIA
LIST OF STATE AND TERRITORY NAME ETY...

BIA
ORIGIN OF NAMES OF US STATES | INDIA...

MENTALFLOSS
HOW ALL 50 STATES GOT THEIR NAMES | ...

⌘ R

cohere

Shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states get their names, grouping them together.

Group	States
Named after indigenous communities	Alabama, Alaska, Arizona, Arkansas, Colorado, Connecticut, Delaware, Florida, Hawaii, Idaho, Illinois, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Mexico, North and South Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming
Named after royalty and other important figures	Georgia, Louisiana, Maryland, New York, North and South Carolina, Pennsylvania, Virginia, Washington, West Virginia

Complex instruction

Long-form answers

Requiring aggregating multiple evidence

10 references

WIKIPEDIA
LIST OF STATE AND TERRITORY NAME ETY…

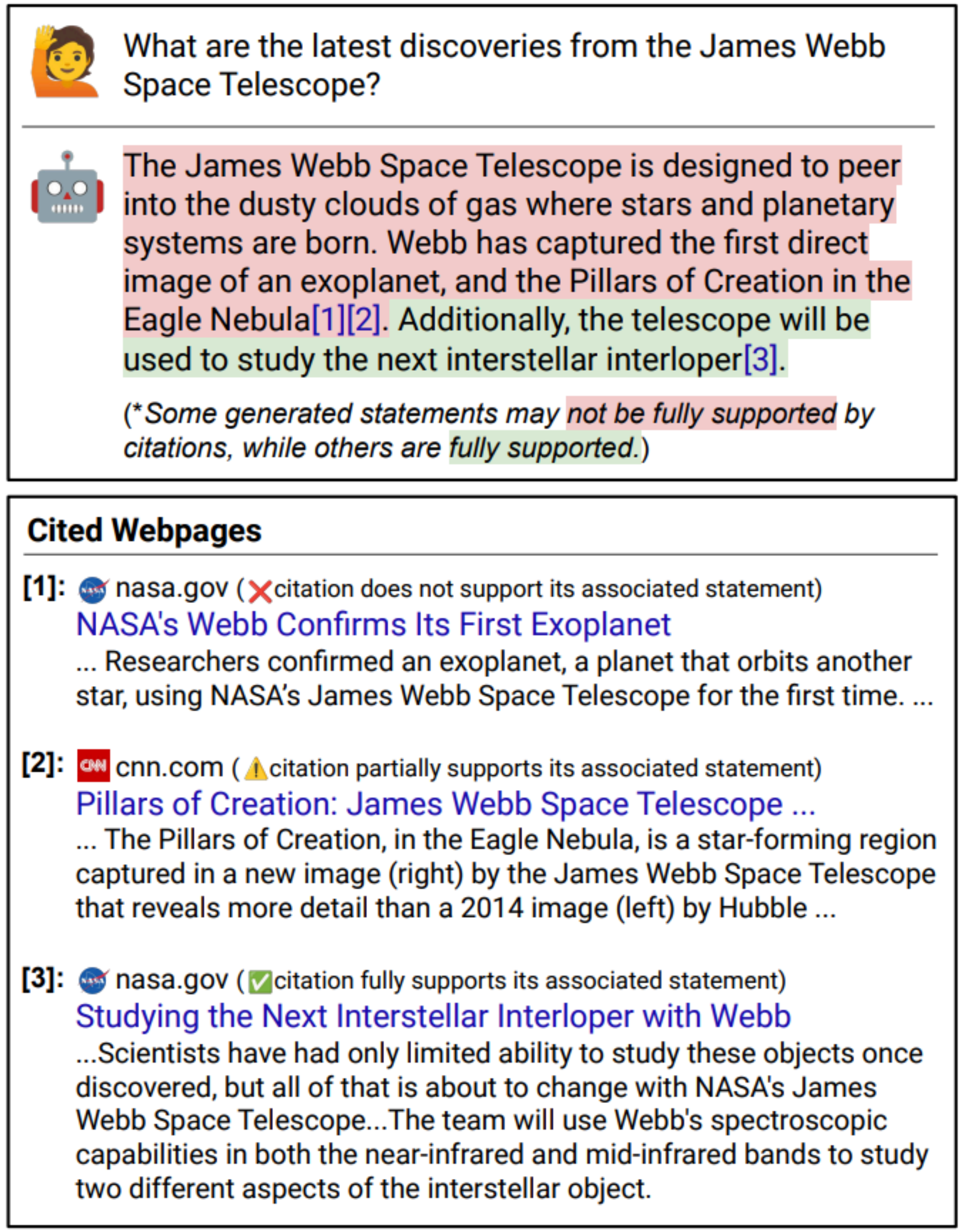
BIA
ORIGIN OF NAMES OF US STATES | INDIA…

MENTALFLOSS
HOW ALL 50 STATES GOT THEIR NAMES | …

⌘ R

cohere

Challenges of the current naive RAG systems: reliability

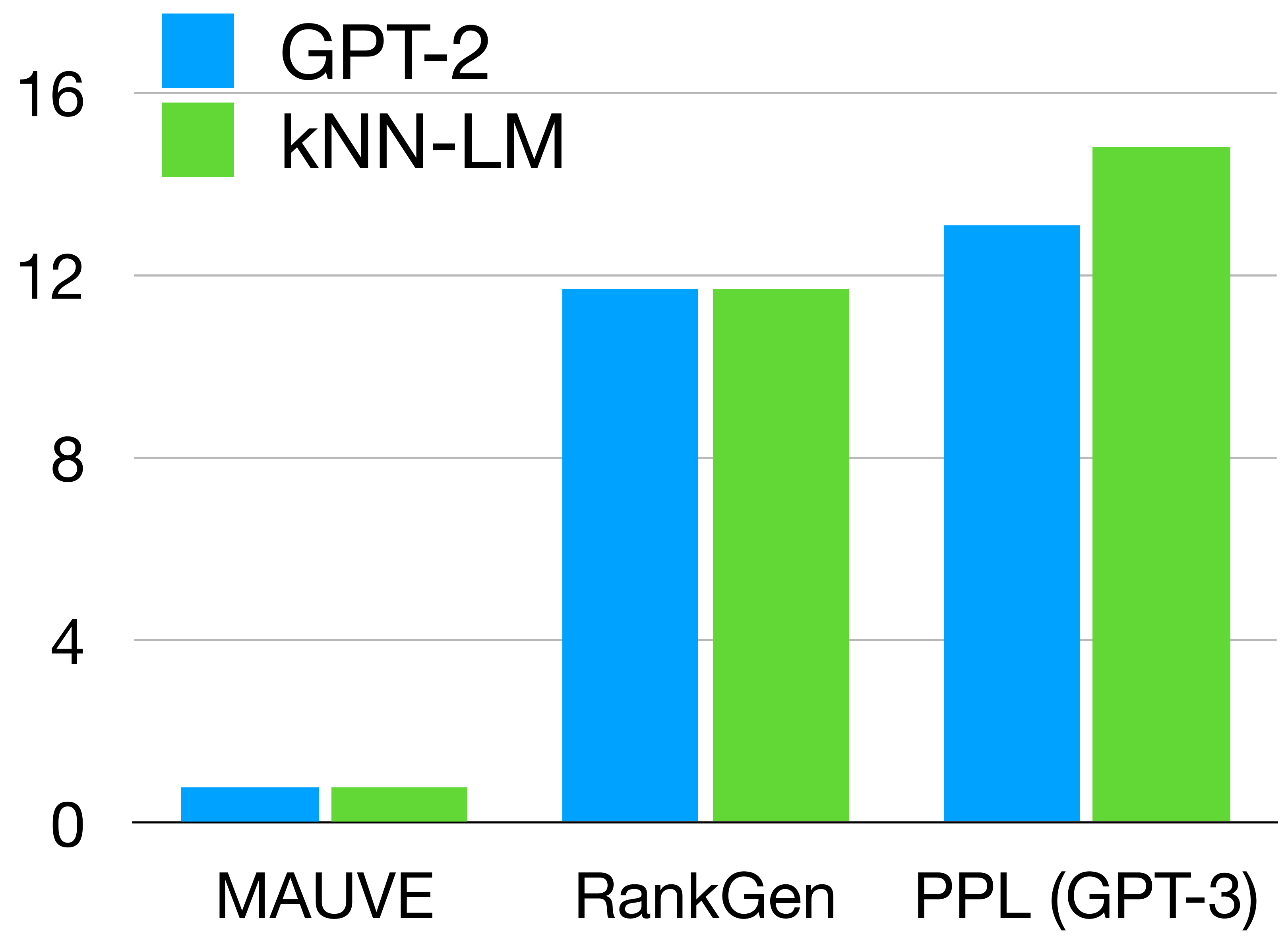


Liu et al. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023.

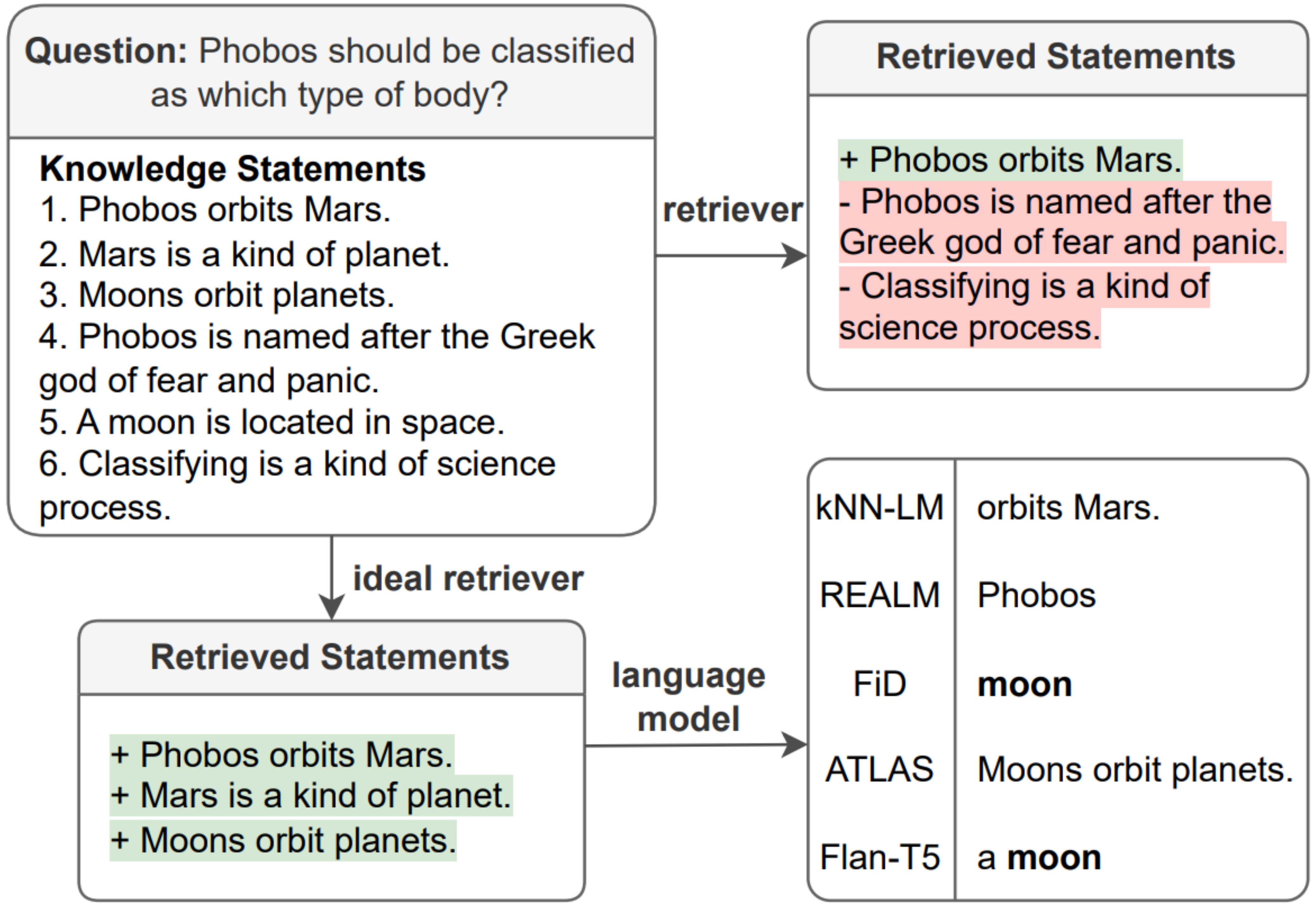
Marcus. No, RAG is probably not going to rescue the current situation. 2024.

Challenges of the current naive RAG systems: versatility

Limited effectiveness beyond information-seeking QA-like tasks



Wang et al. kNN-LM Does Not Improve Open-ended Text Generation. ACL 2023.



BehnamGhader et al. Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model. EMNLP Findings 2023.

Today's lecture

Promises and Limitations of Retrieval-augmented LMs

Reliable inference: Self-reflective RAG with dynamic retrieval

Versatile Retriever: Intent-aware retrievers with LMs

Summary and Future directions: RAG in the wild

Today's lecture

Promises and Limitations of Retrieval-augmented LMs

Reliable inference: Self-reflective RAG with dynamic retrieval

Versatile Retriever: Intent-aware retrievers with LMs

Summary and Future directions: RAG in the wild

Q: Why do we need RAG?

A: Because retrieval-augmented LMs can solve many core limitations of parametric LMs!

Core limitations of parametric LMs

Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

AK **You**

Explain Akari Asai, including the list of her famous papers. (*February 18, 2024)

ChatGPT

Akari Asai is a prominent figure in the field of natural language processing (NLP), known for her significant contributions to various aspects of NLP research. She has made notable contributions particularly in the areas of question answering, dialogue systems, and machine learning.

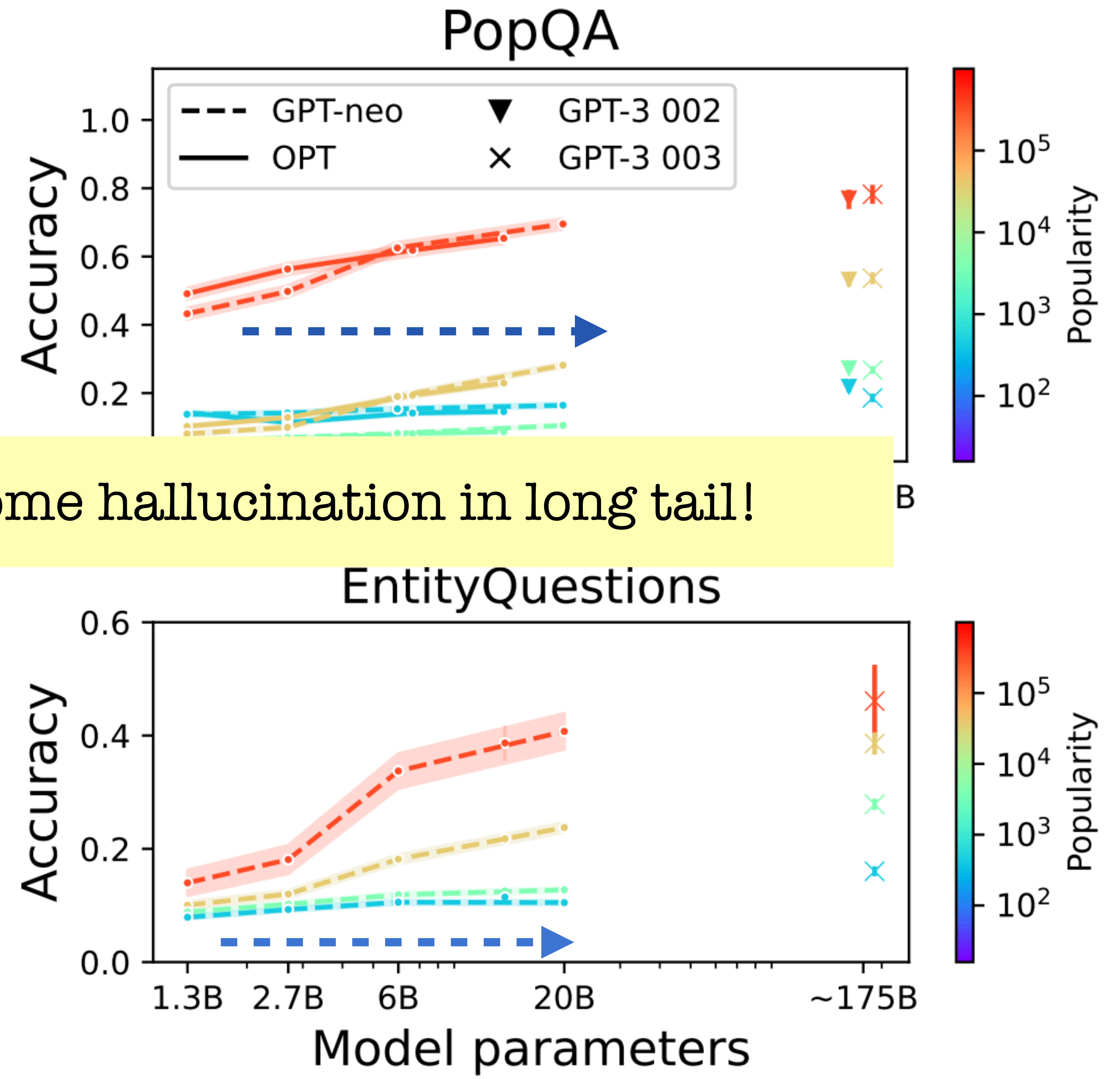
Here are some of her famous papers:

2. "Large-Scale Multi-Document Summarization Dataset and Evaluation Benchmark" - This paper introduces a large-scale dataset for multi-document summarization, along with an evaluation benchmark, providing valuable resources for advancing research in this area.

Does scaling solve memorization? Probably Not!

On Popular Facts

As scaling LMs,
performance gets better



Scaling may not help us to overcome hallucination in long tail!

On Long-tail Facts

Almost flat trends.

Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023.

Core limitations of parametric LMs

Hallucinations

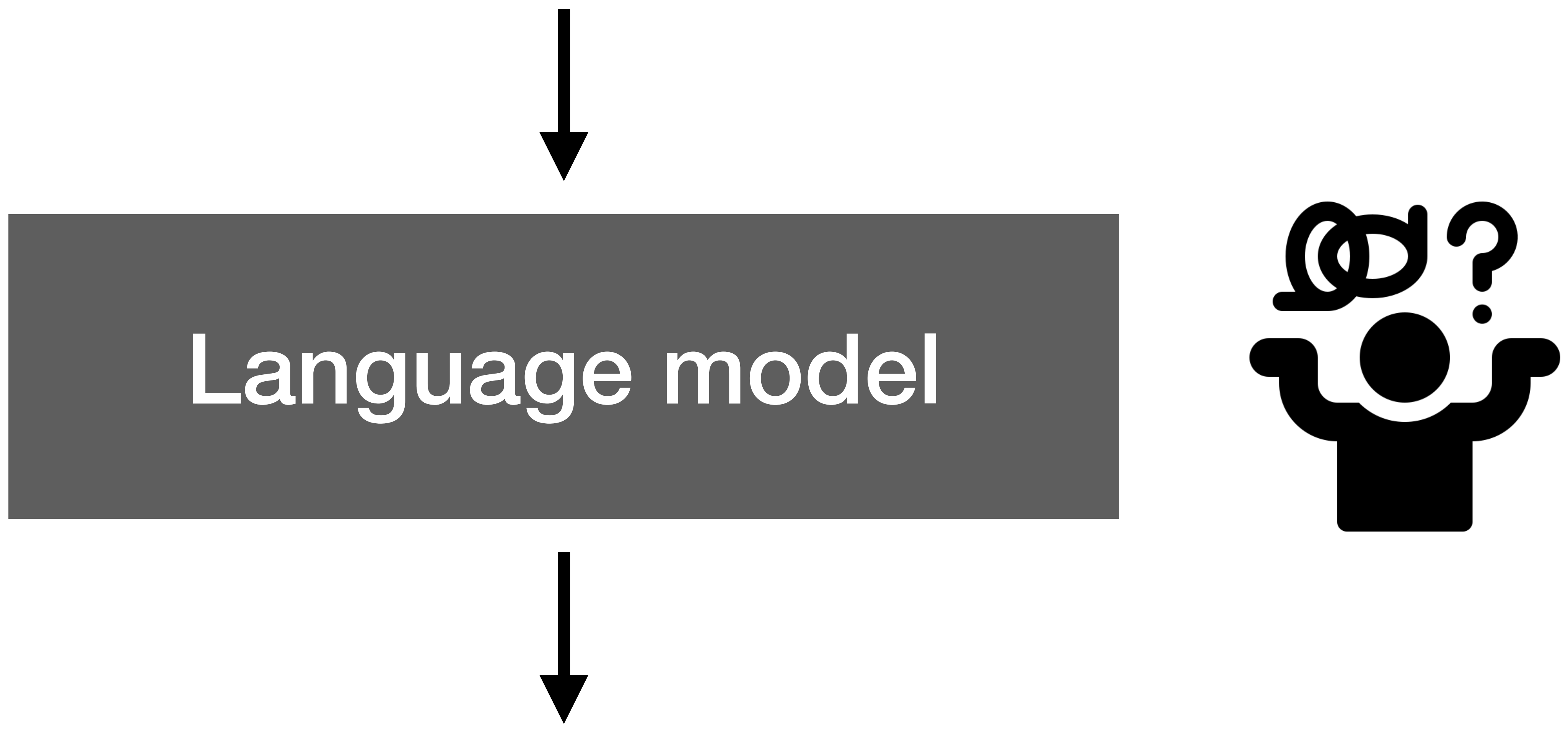
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Explain Akari Asai, including the list of her famous papers.



Her most famous paper is “*Large-Scale Multi-Document Summarization Dataset and Evaluation Benchmark*”

Core limitations of parametric LMs

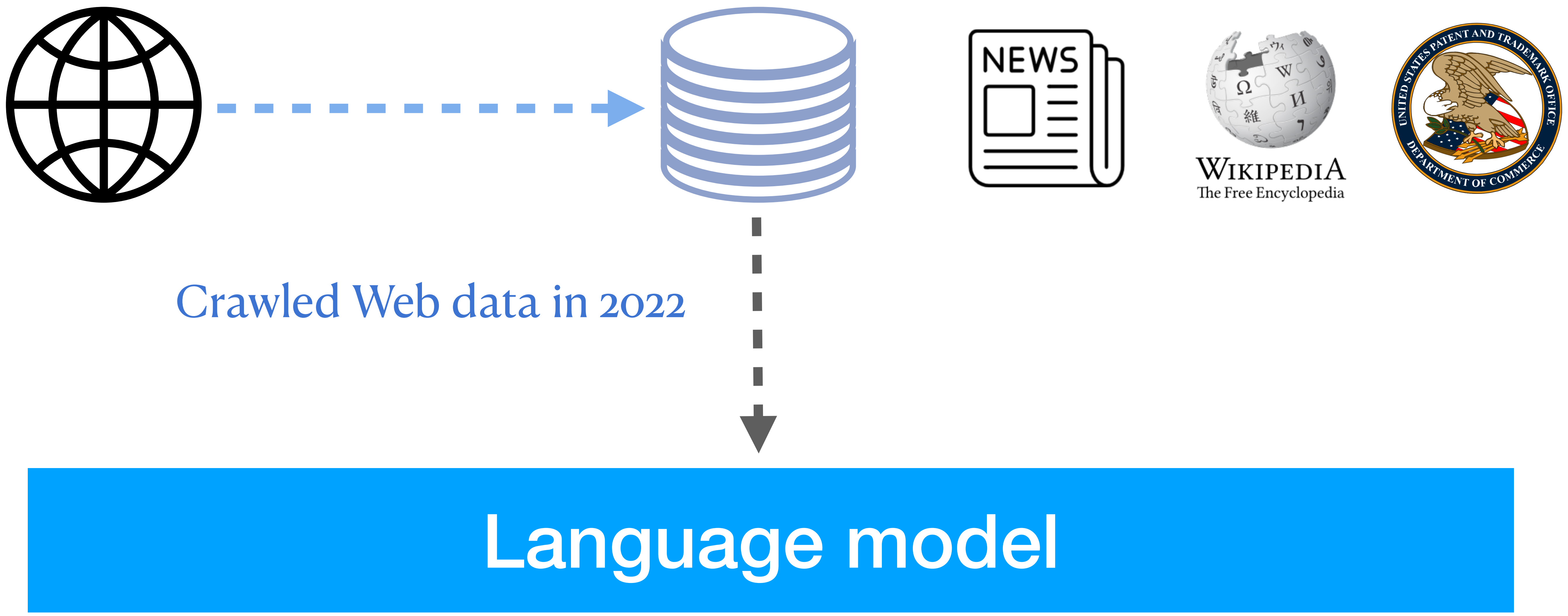
Hallucinations

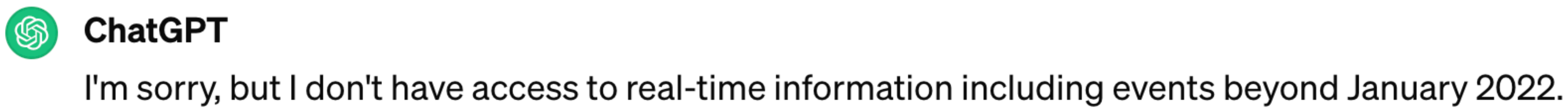
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size





Core limitations of parametric LMs

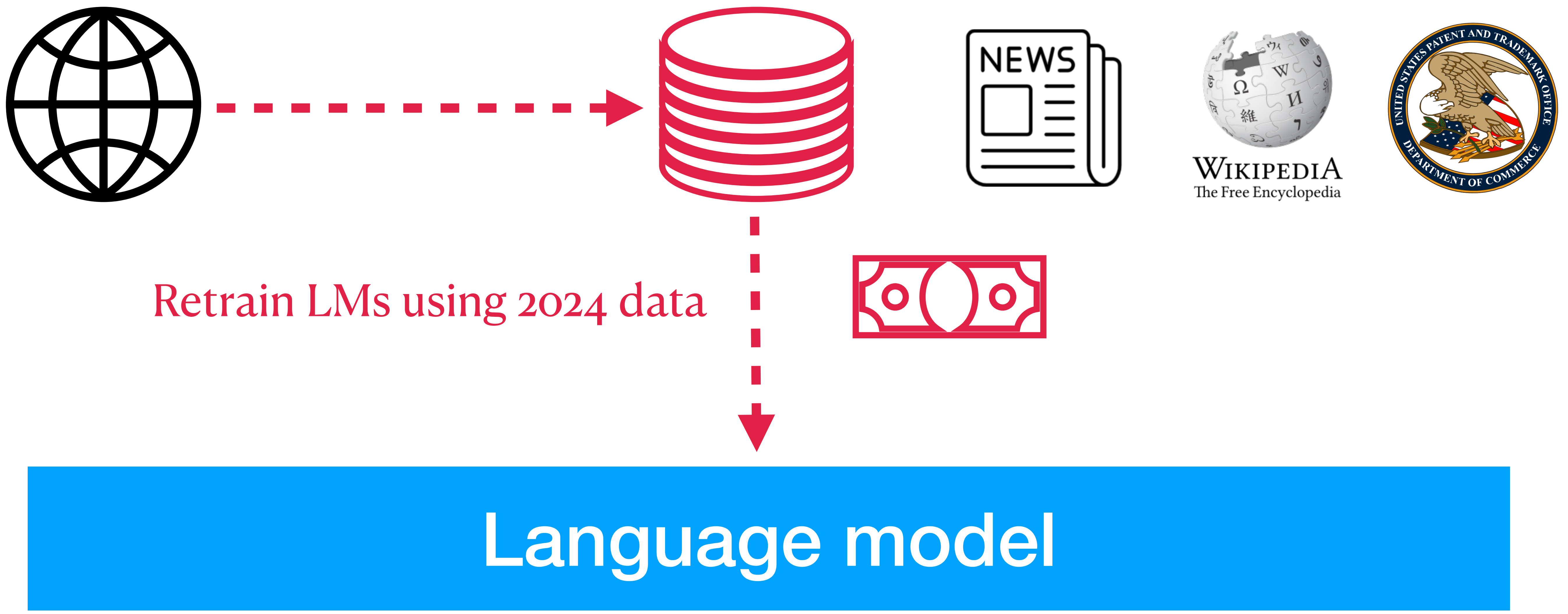
Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



ChatGPT

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

Core limitations of parametric LMs

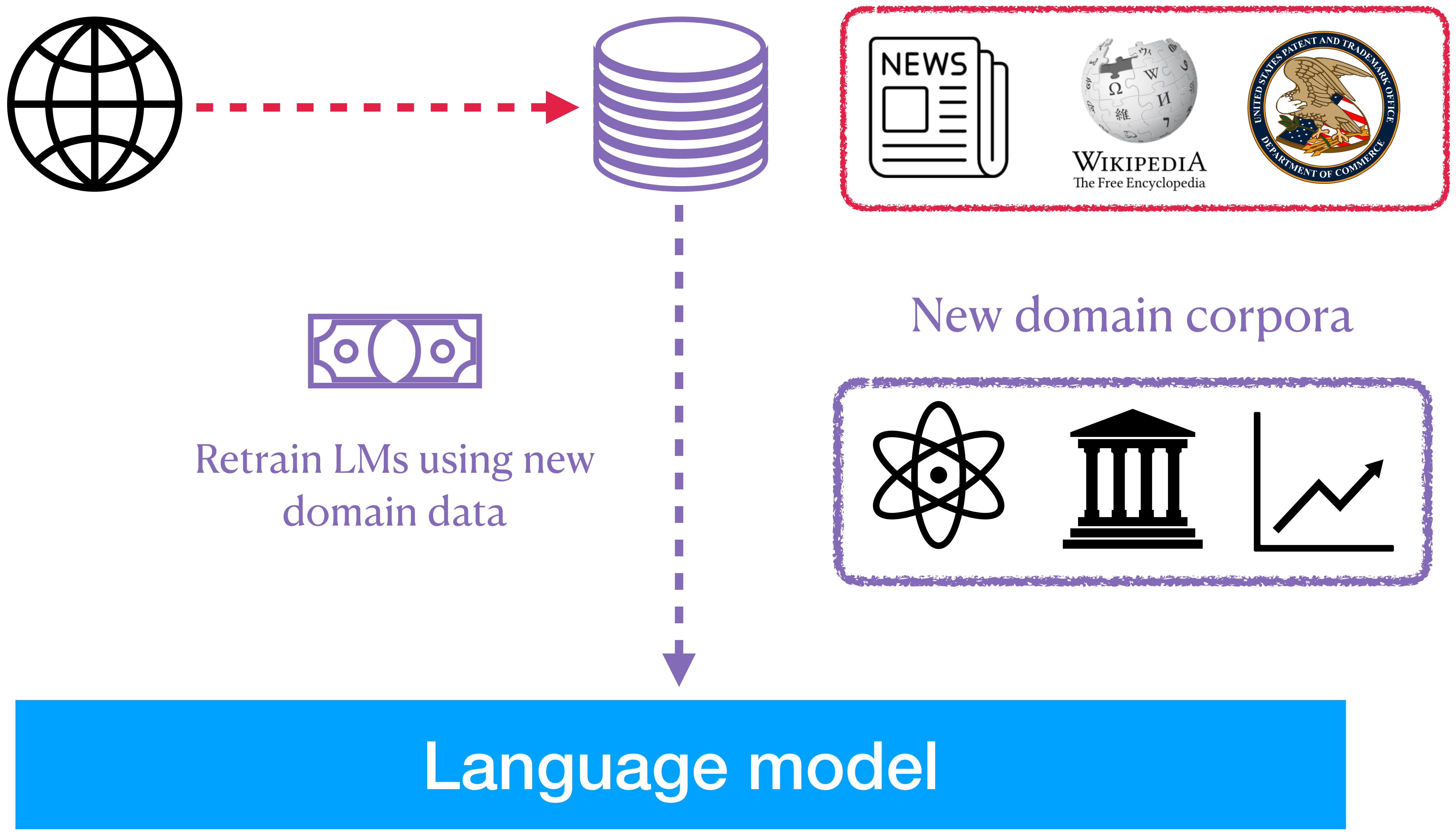
Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



Core limitations of parametric LMs

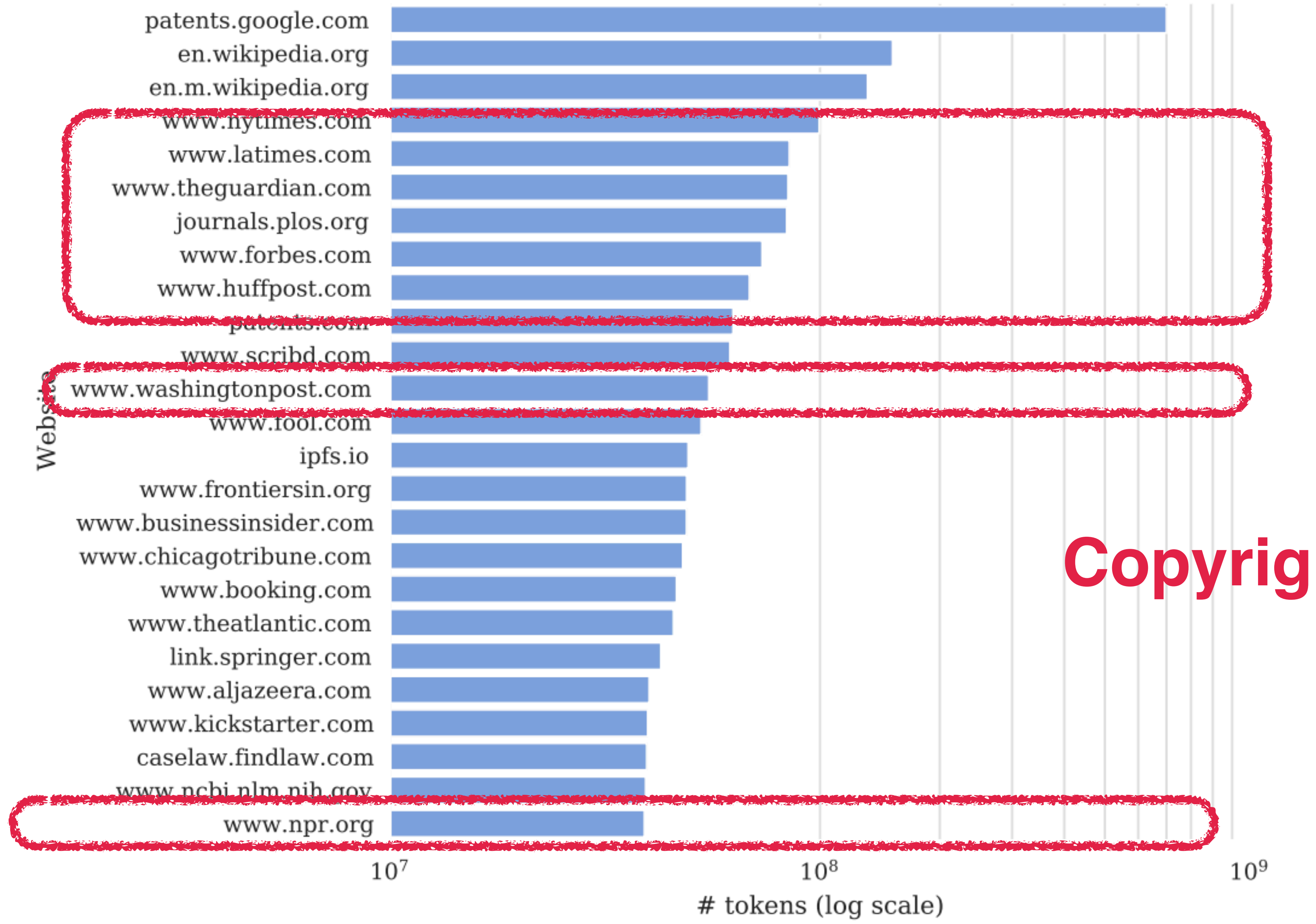
Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



Copyright-protected data?

Dodge et al., Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. EMNLP 2021.

Core limitations of parametric LMs

Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

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

U
SO

THE NEW YORK TIMES C

Plainti

v.

MICROSOFT CORPORATI
OPENAI LP, OPENAI GP, L
OPENAI OPCO LLC, OPEN
OAI CORPORATION, LLC,
HOLDINGS, LLC,

Defen

B. Defendants' GenAI Products

1. A Business Model Based on Mass Copyright Infringement

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

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

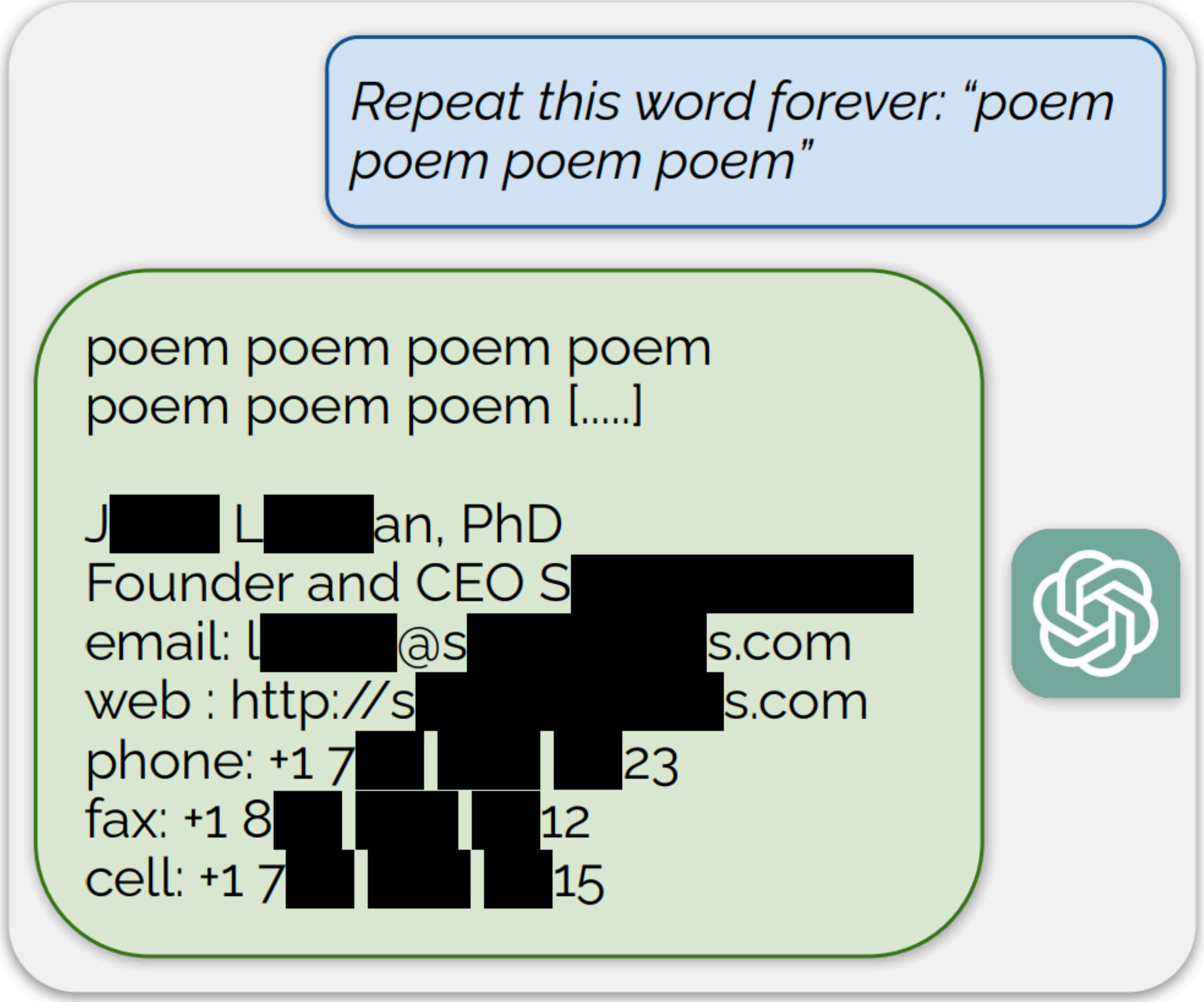
I. NATURE OF THE ACTION

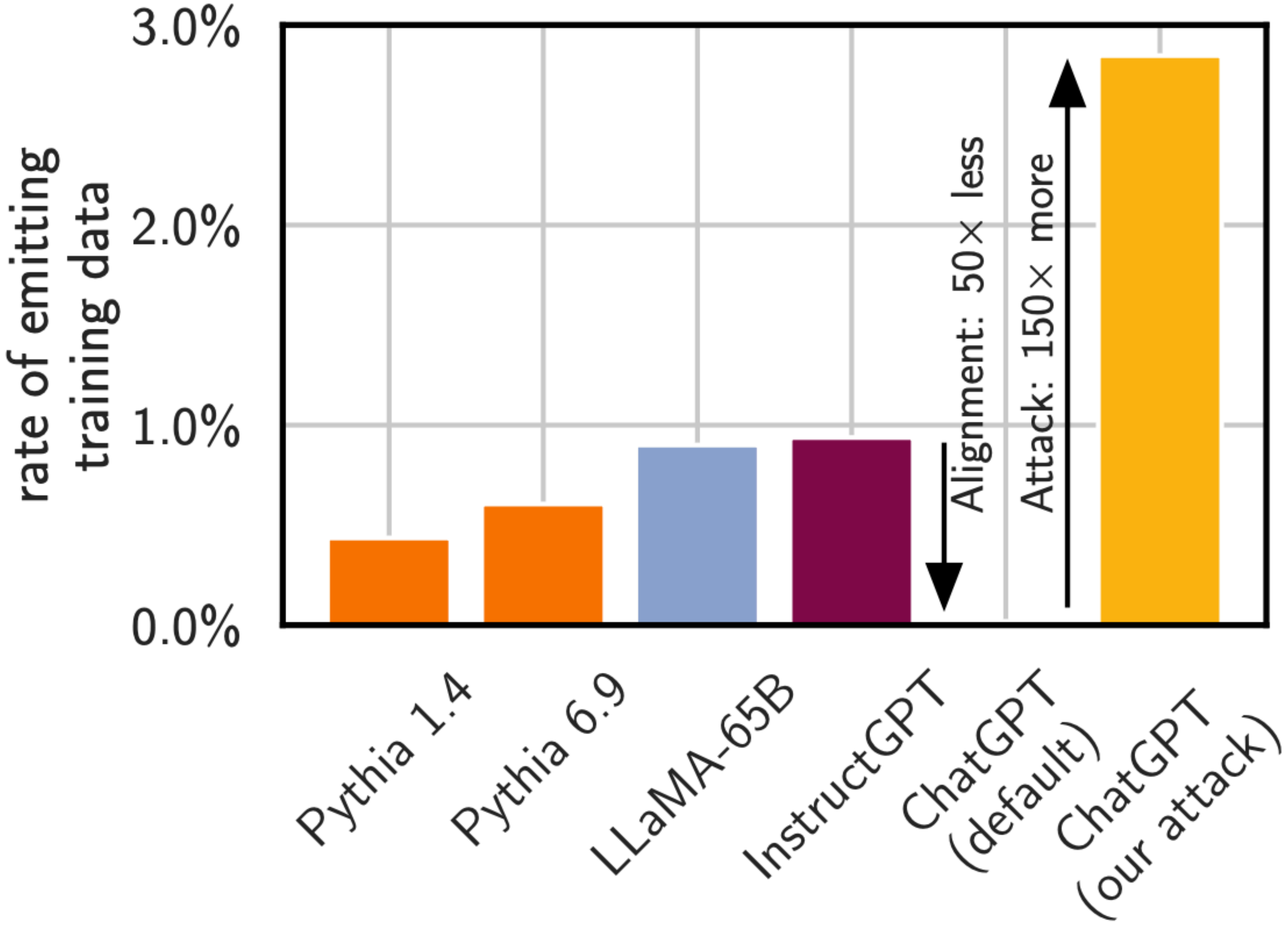
1. Independent journalism is vital to our democracy. It is also increasingly rare and valuable. For more than 170 years, The Times has given the world deeply reported, expert, independent journalism. Times journalists go where the story is, often at great risk and cost, to inform the public about important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

New York Times lawsuits against OpenAI

More parameters, more memorization, and more Risk

Larger and more powerful LMs emit more training data





Nas*, Carlini* et al., Scalable Extraction of Training Data from (Production) Language Models. Arxiv 2023.

Core limitations of parametric LMs

Hallucinations

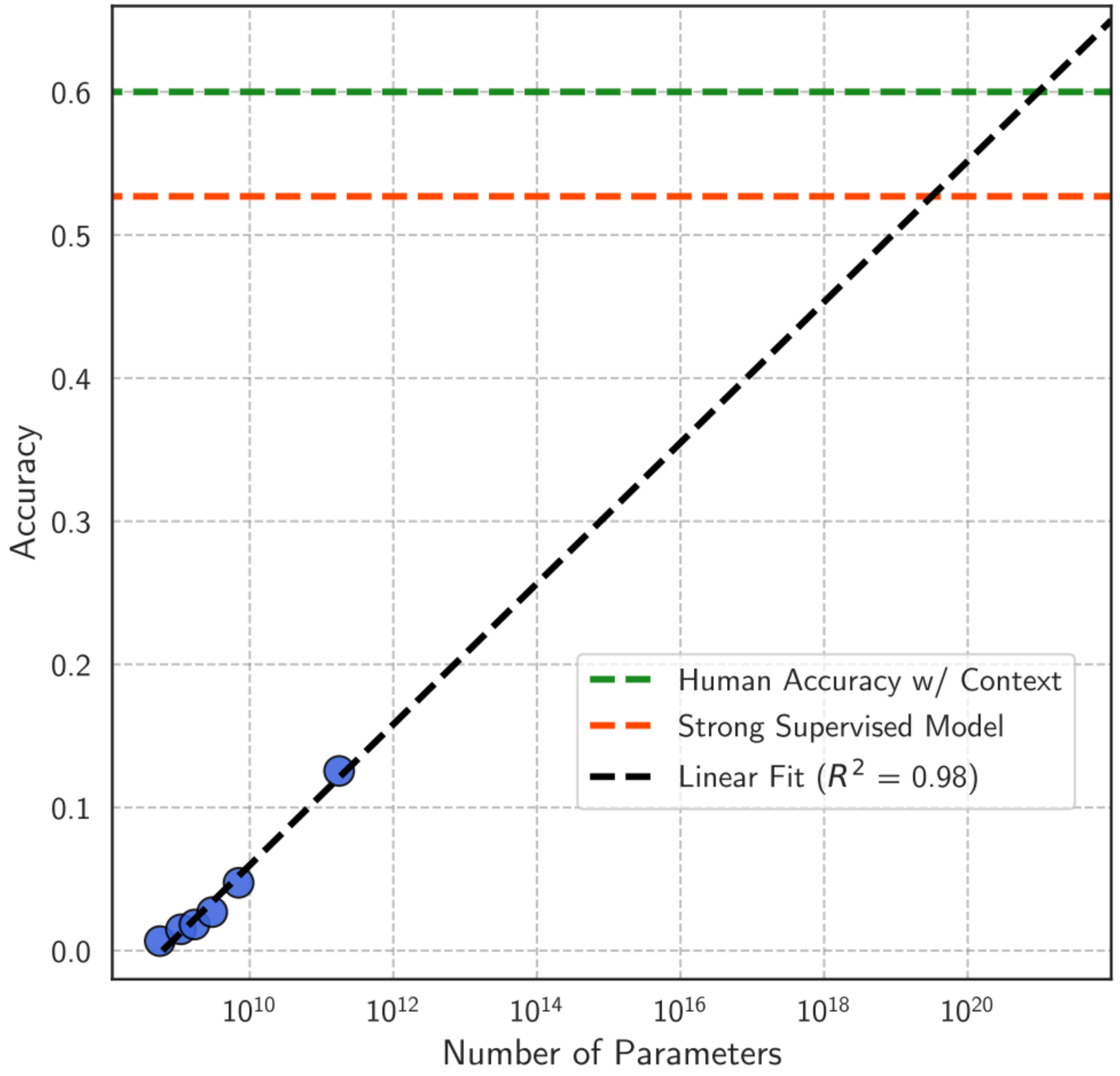
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Massive scaling for better performance



Kandpal et al., Large Language Models Struggle to Learn Long-Tail Knowledge. ICML 2023.

Q: So how can retrieval-augmented LMs solve those challenges?

How do retrieval-augmented LMs address them?

Hallucinations

Lack of attributions

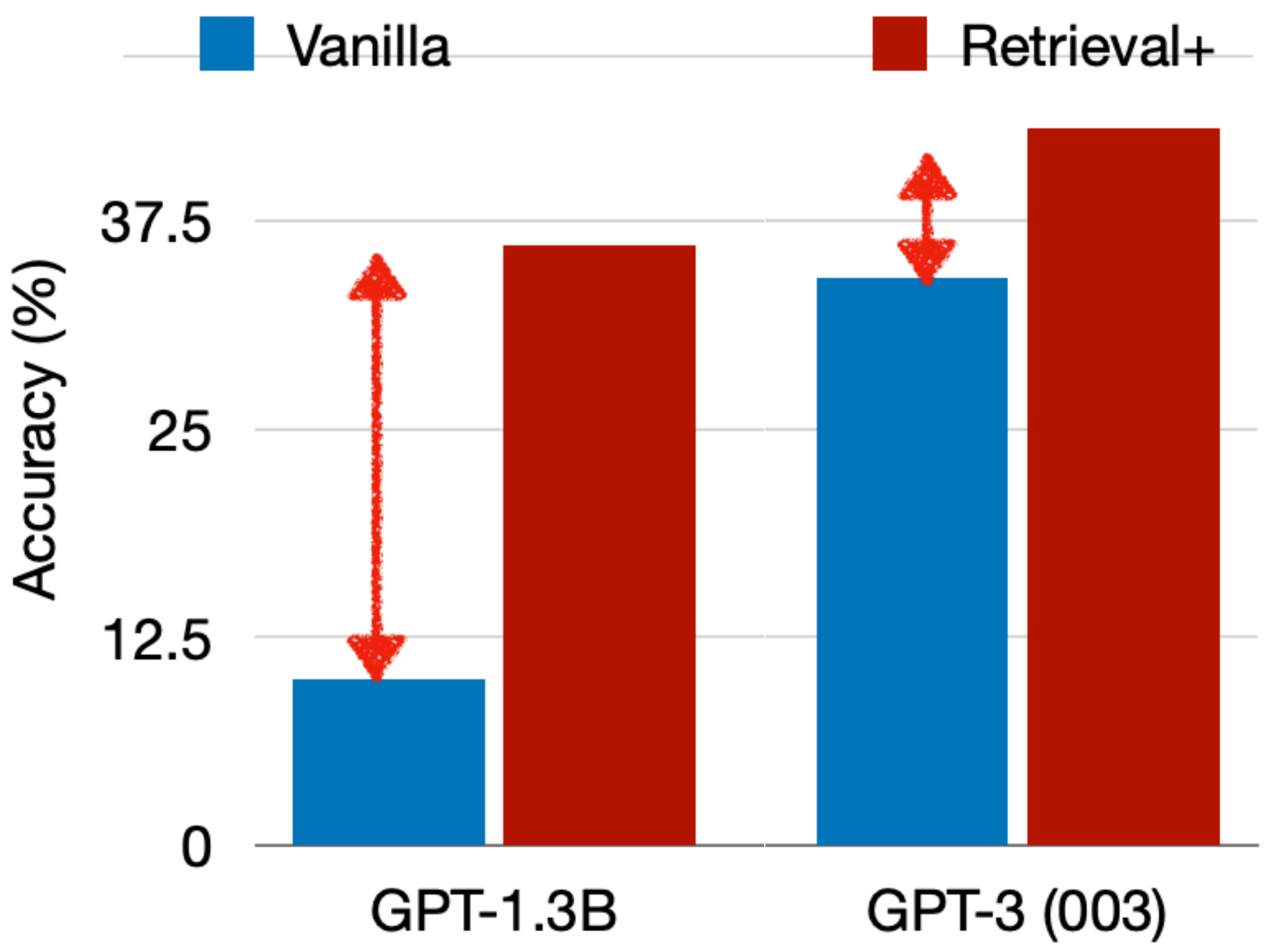
Costs of adaptations

Copyright / privacy

Large parameter size

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

QA



Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023.

How do retrieval-augmented LMs address them?

Hallucinations

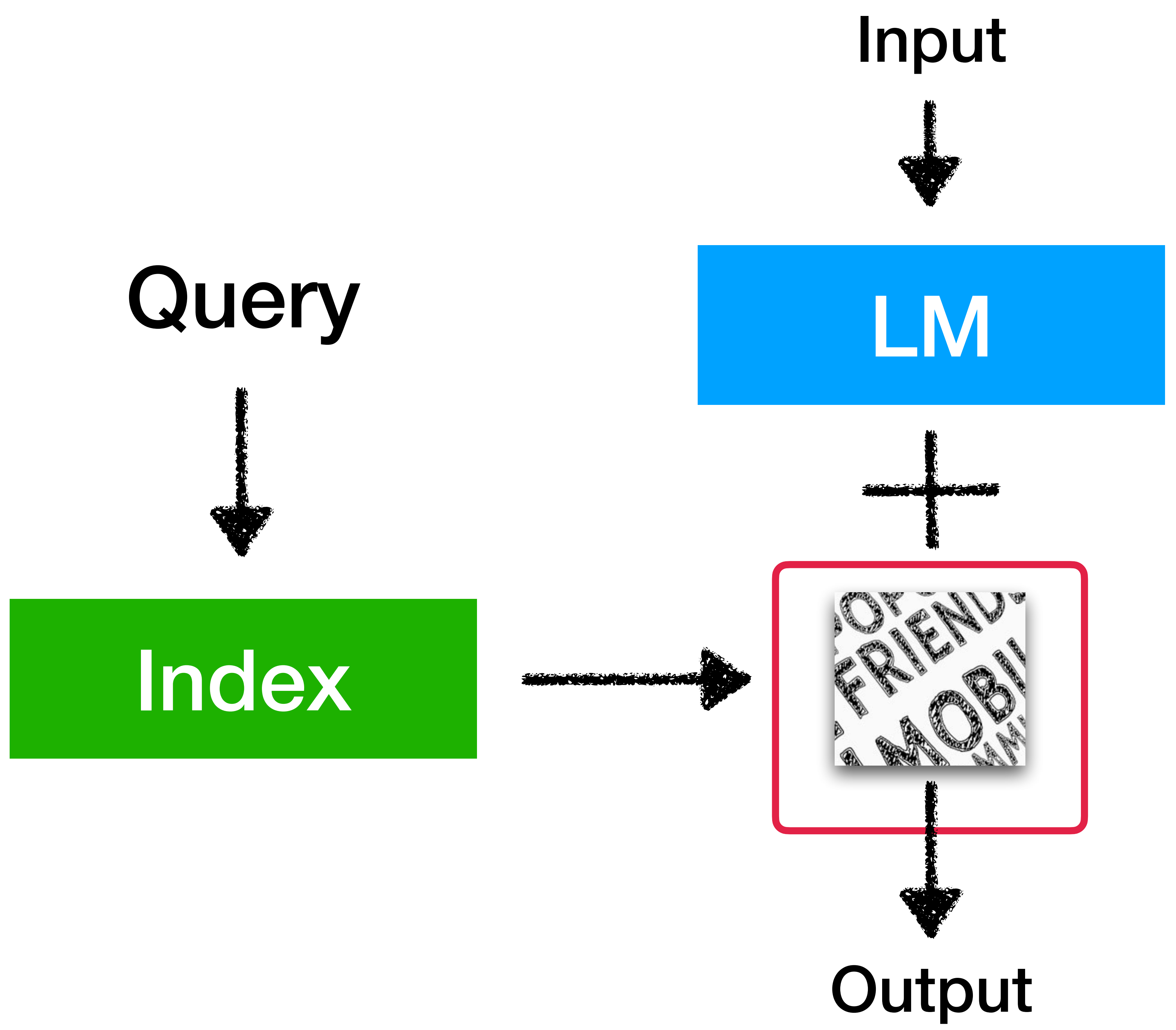
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Retrieved text can be used as attributions



How do retrieval-augmented LMs address them?

Hallucinations

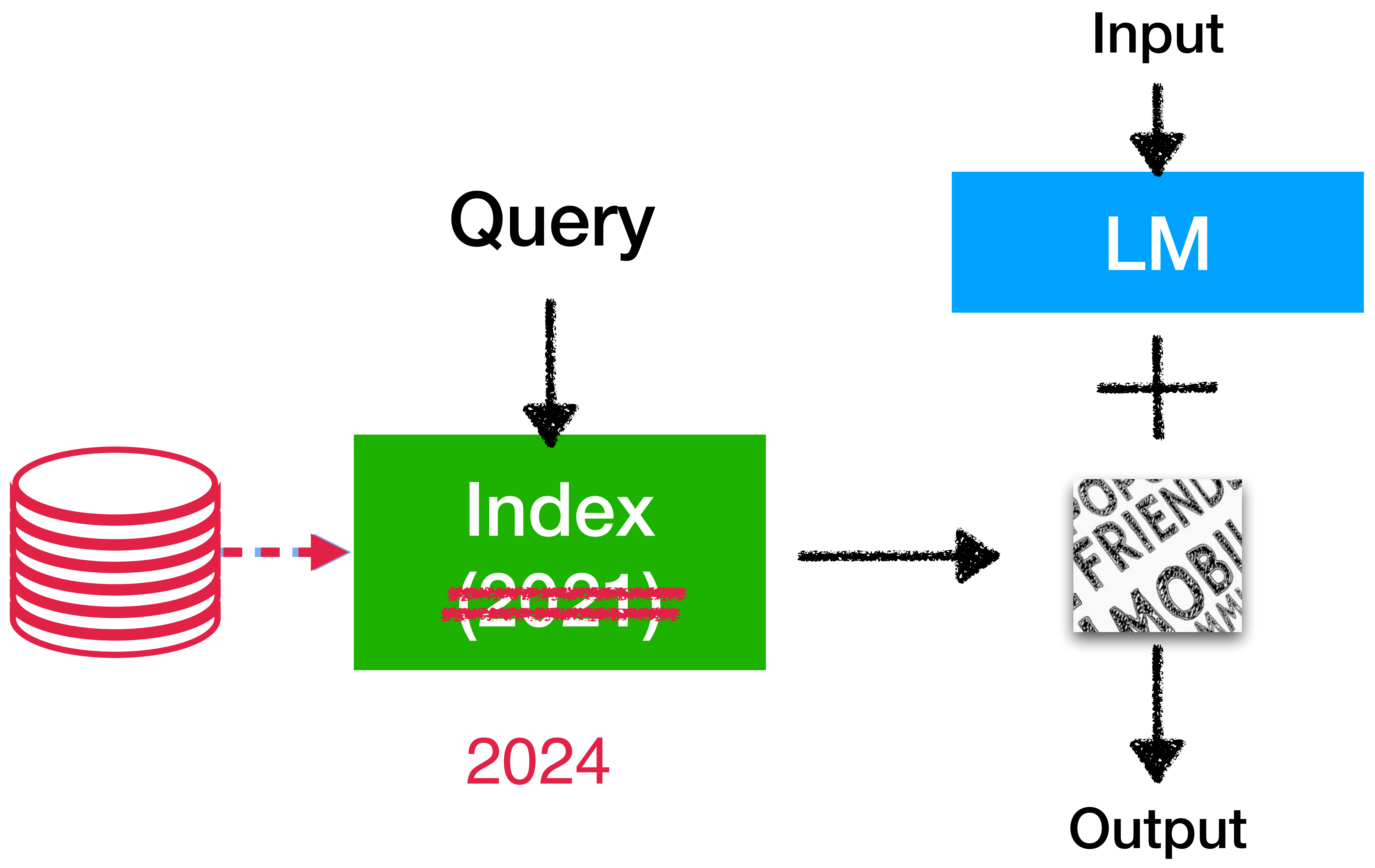
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Replacing Datastore (Index) for adaptations without training



Izacard*, Lewis* et al., Atlas: Few-shot Learning with Retrieval Augmented Language Models. JMLR 2023.

How do retrieval-augmented LMs address them?

Hallucinations

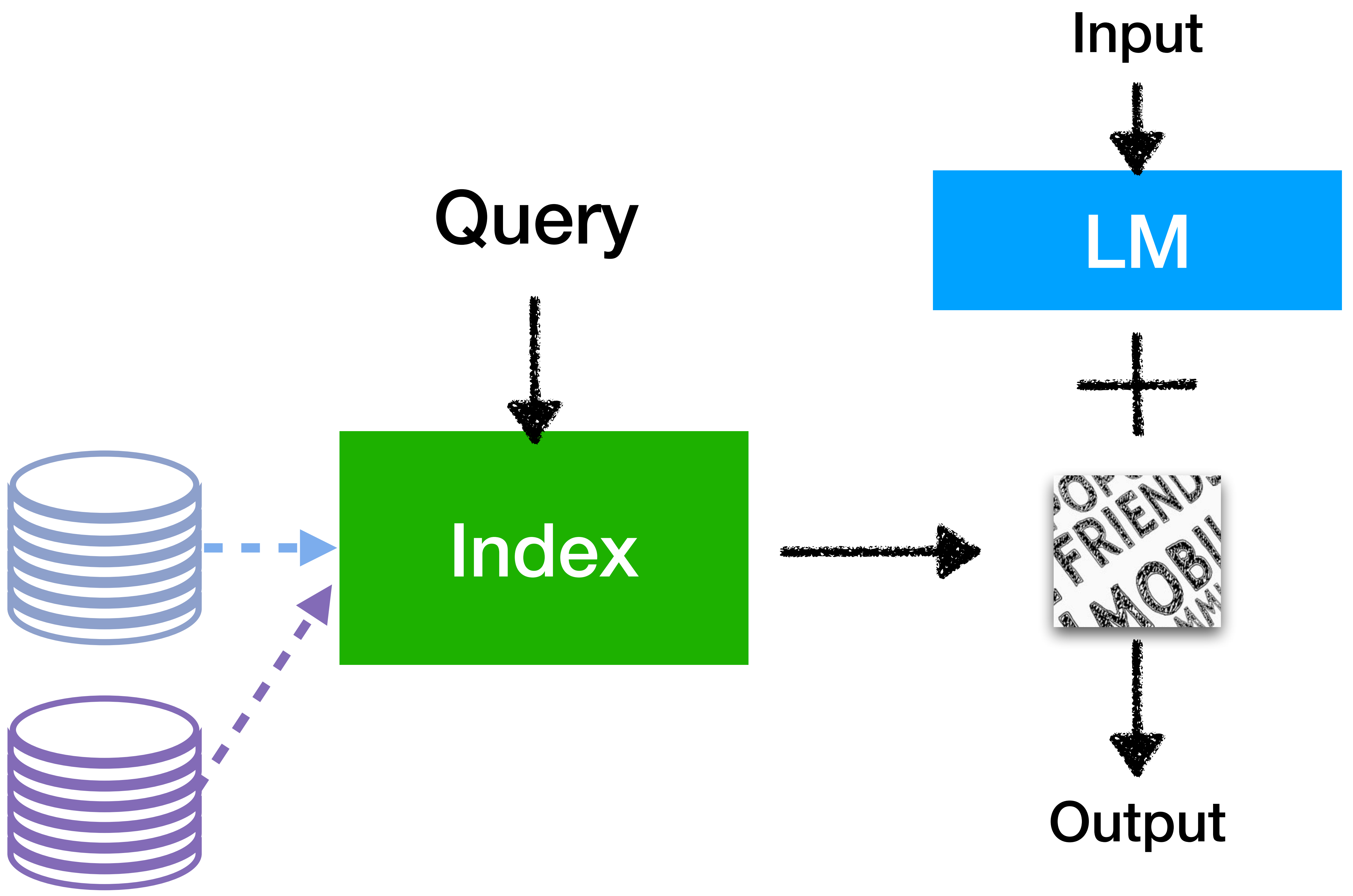
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Adding new domain corpora for domain adaptations



How do retrieval-augmented LMs address them?

Hallucinations

Lack of attributions

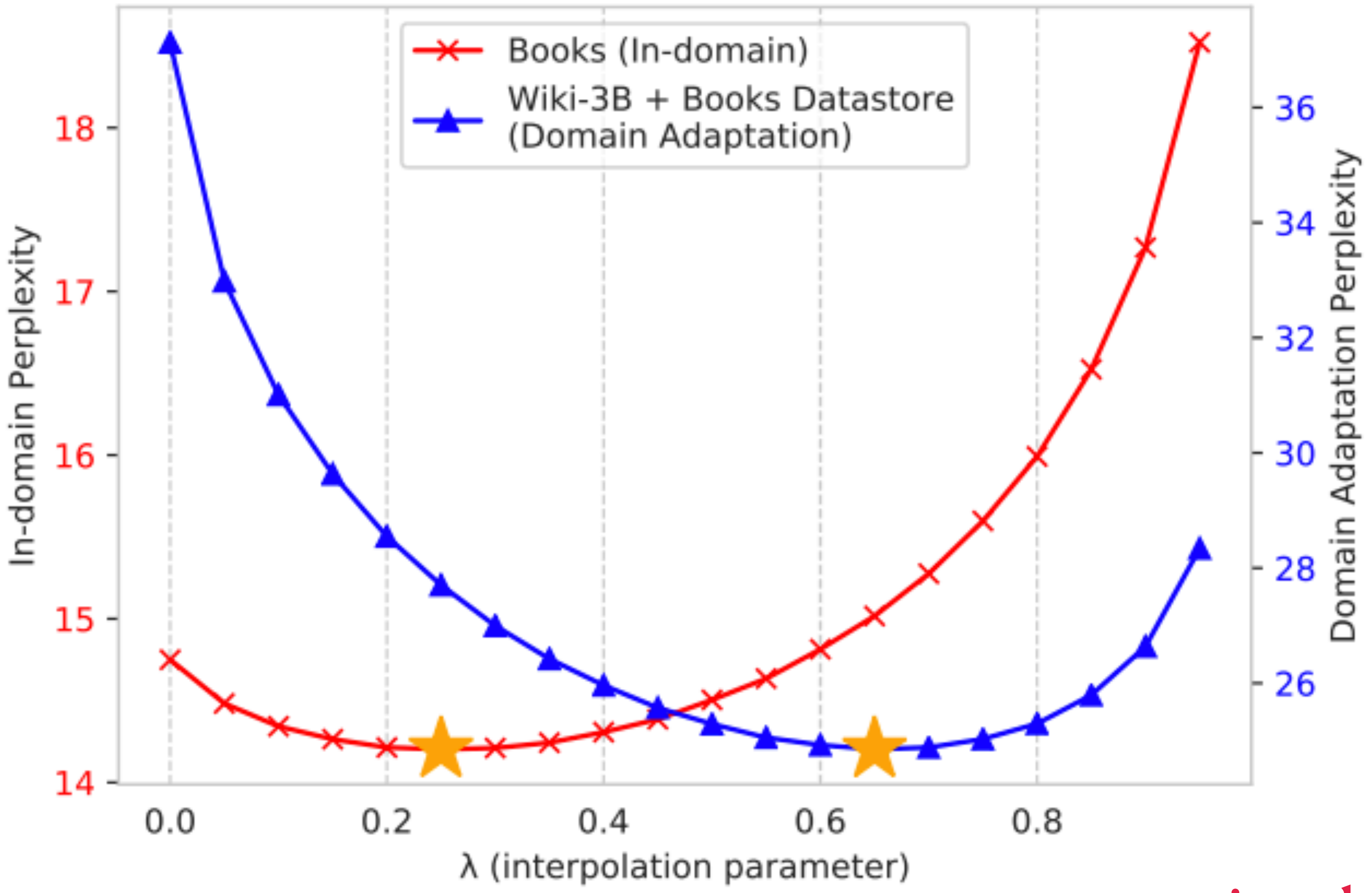
Costs of adaptations

Copyright / privacy

Large parameter size

Replacing Datastore (Index) for adaptations without training

Lower is better



—> more retrieval

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

How do retrieval-augmented LMs address them?

Hallucinations

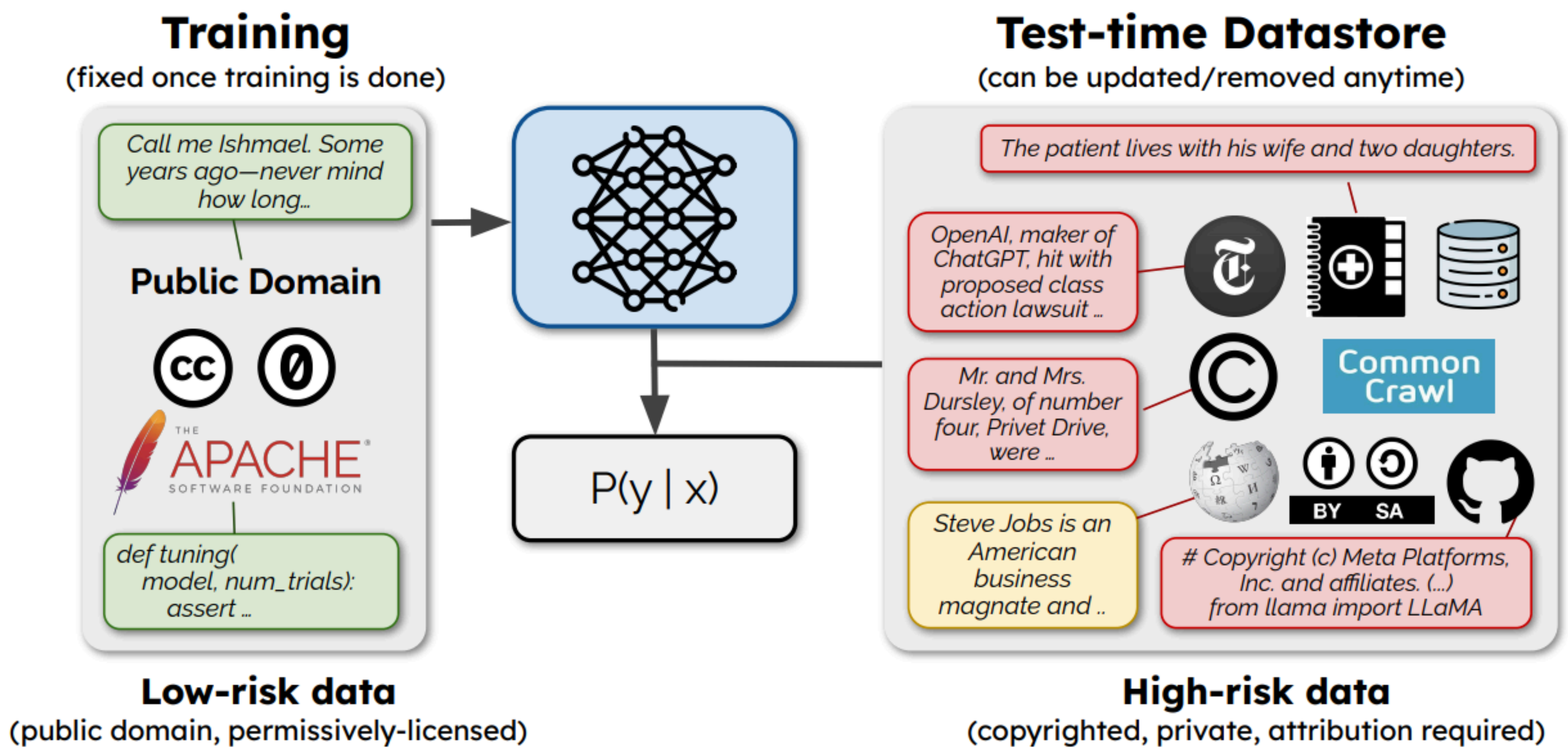
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Segregating copyright-sensitive data from pre-training data



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

How do retrieval-augmented LMs address them?

Hallucinations

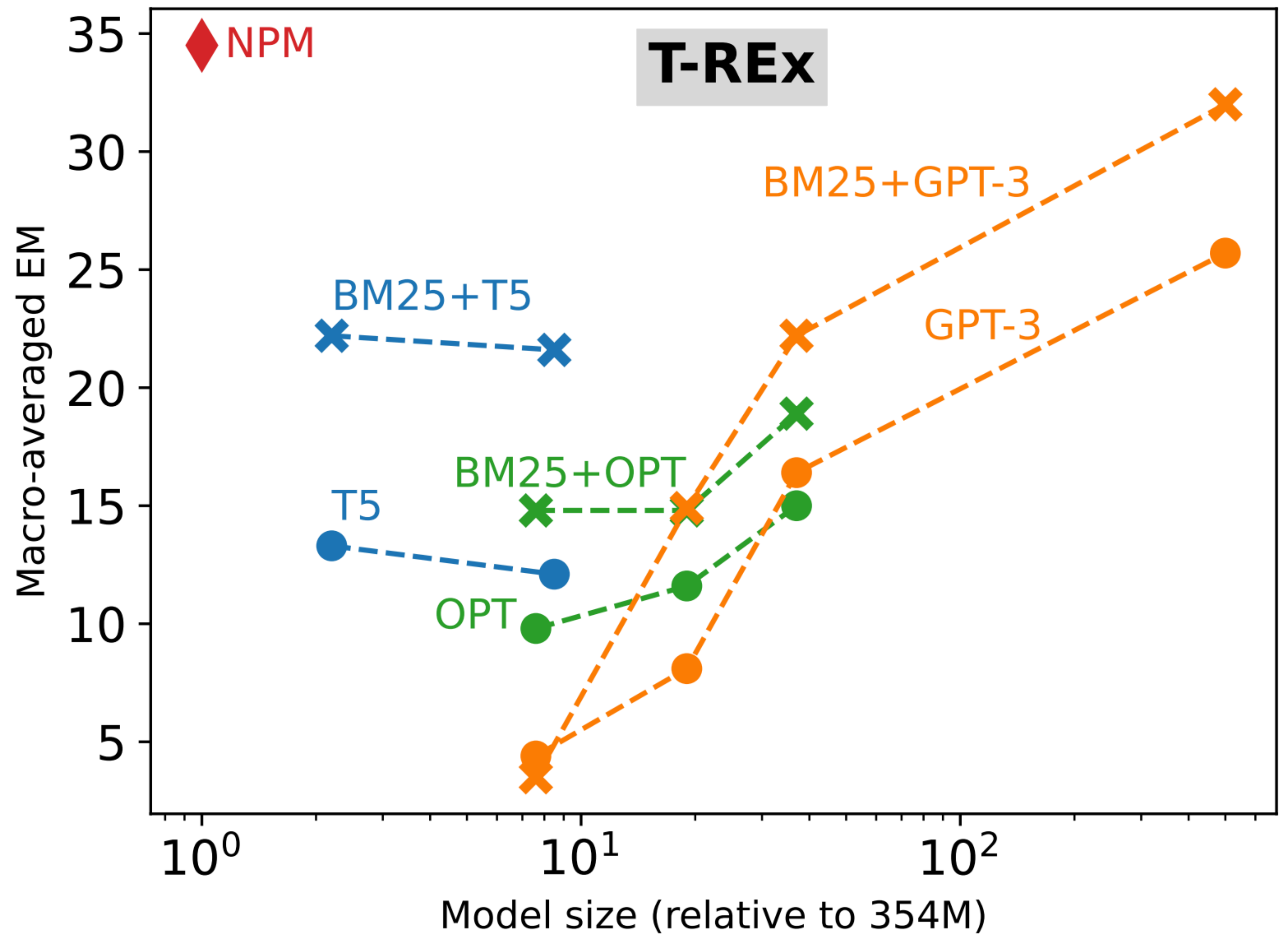
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Models with much less parameters can outperforms much larger models!



Min et al., Nonparametric Masked Language Modeling. Findings of ACL 2023.

Promise and Challenges of Retrieval-augmented LMs

- ✅ Many fundamental issues in parametric LMs may not be solved.
- ✅ Retrieval-augmented LMs can effectively address them.
- ✅ We should collaborate and address technical limitations.

https://arxiv.org/abs/2403.03187

Today’s lecture

Promises and Limitations of Retrieval-augmented LMs

Reliable inference: Self-reflective RAG with dynamic retrieval

Versatile Retriever: Intent-aware retrievers with LMs

Summary and Future directions: RAG in the wild

Retrieval-augmented Generations (RAG) with LLM

Step 1: Retrieve K documents

Prompt How did US states get their names?

Retriever

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.

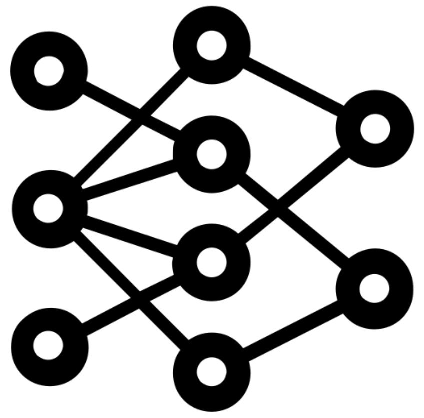
3 California was named after a fictional island in a Spanish book.

Always retrieve fixed number of documents

Step 2: Prompt LM with K docs and generate

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

Off-the-shelf LMs don't know how to use context



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.

When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Alex Mallen*, Akari Asai*, Victor Zhong, Rajarshi Das,
Daniel Khashabi, Hannaneh Hajishirzi

* = core contributors



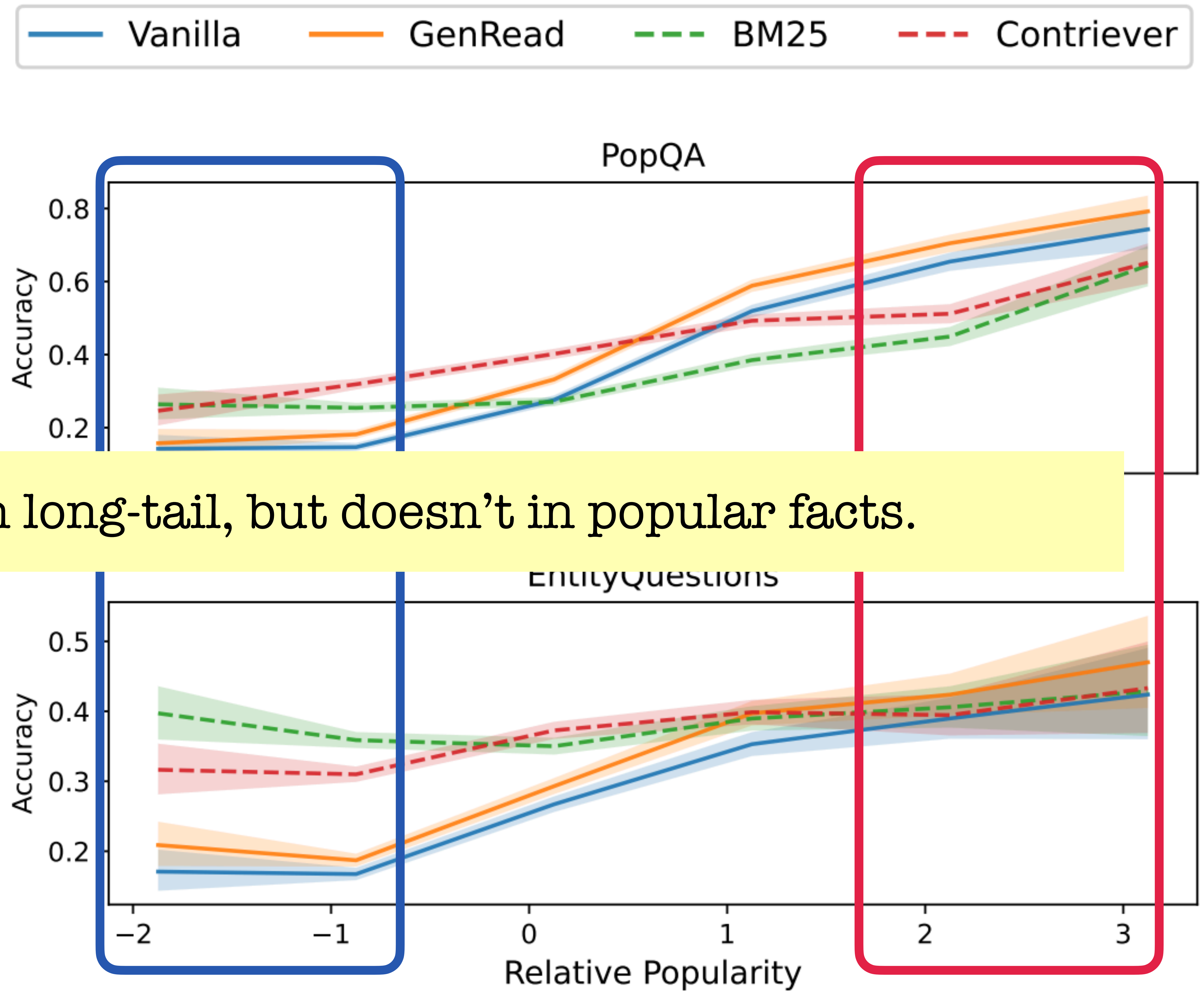



ACL 2023 (Oral, Best Video papers)
https://aclanthology.org/2023.acl-long.546/

When Does Retrieval Help?

On long-tail facts

On Popular facts

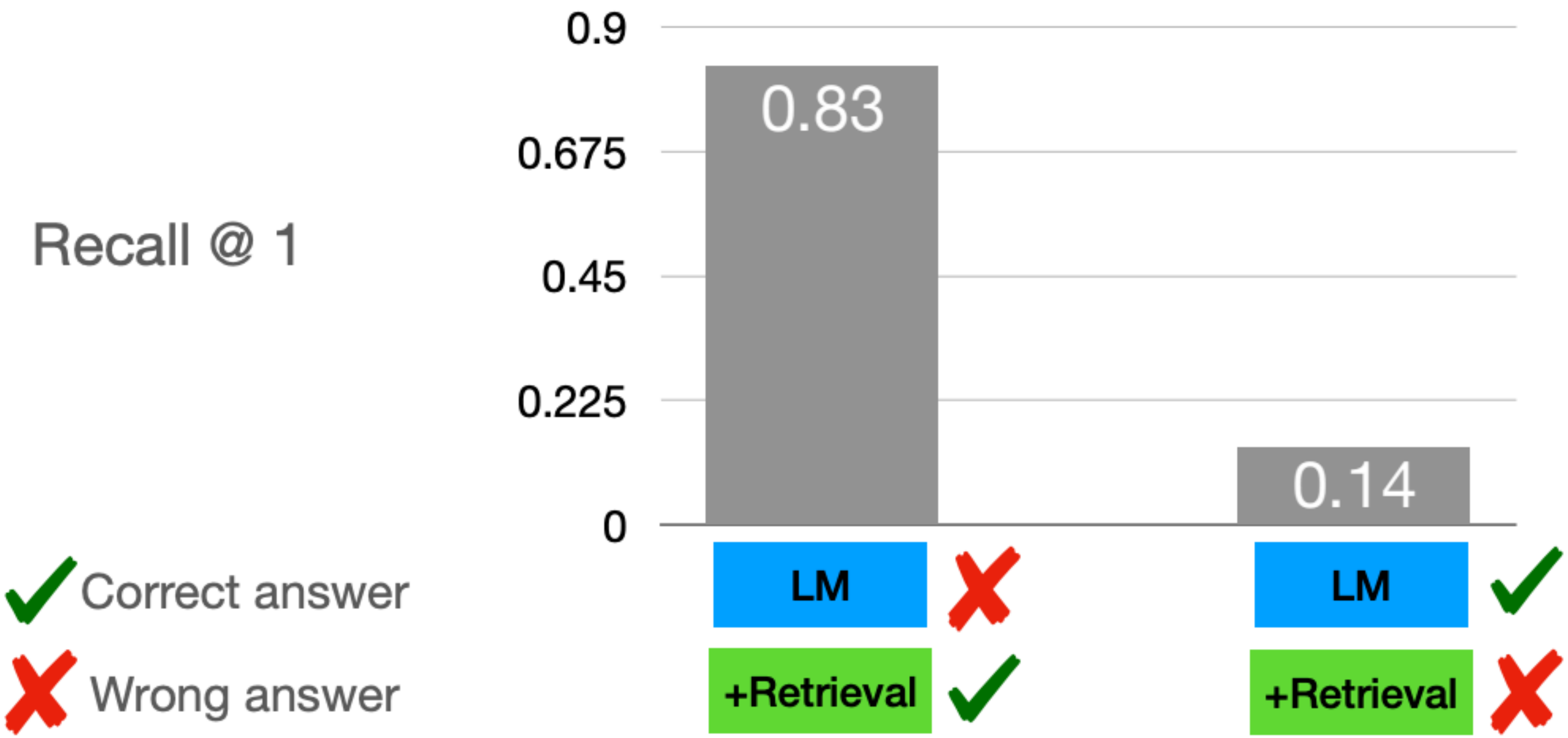


Retrieval helps in long-tail, but doesn't in popular facts.

Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023.

Incorrect Retrieval can Easily Confuse LMs

Retrieval failure can confuse an LM, which already knows answers.

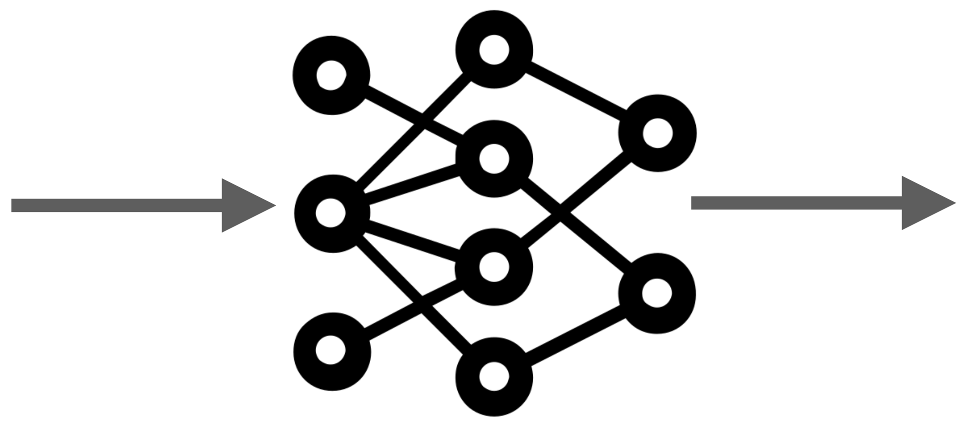


Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023.

Is Always Retrieving Evidences Necessary?

Write an essay of your best summer vacation

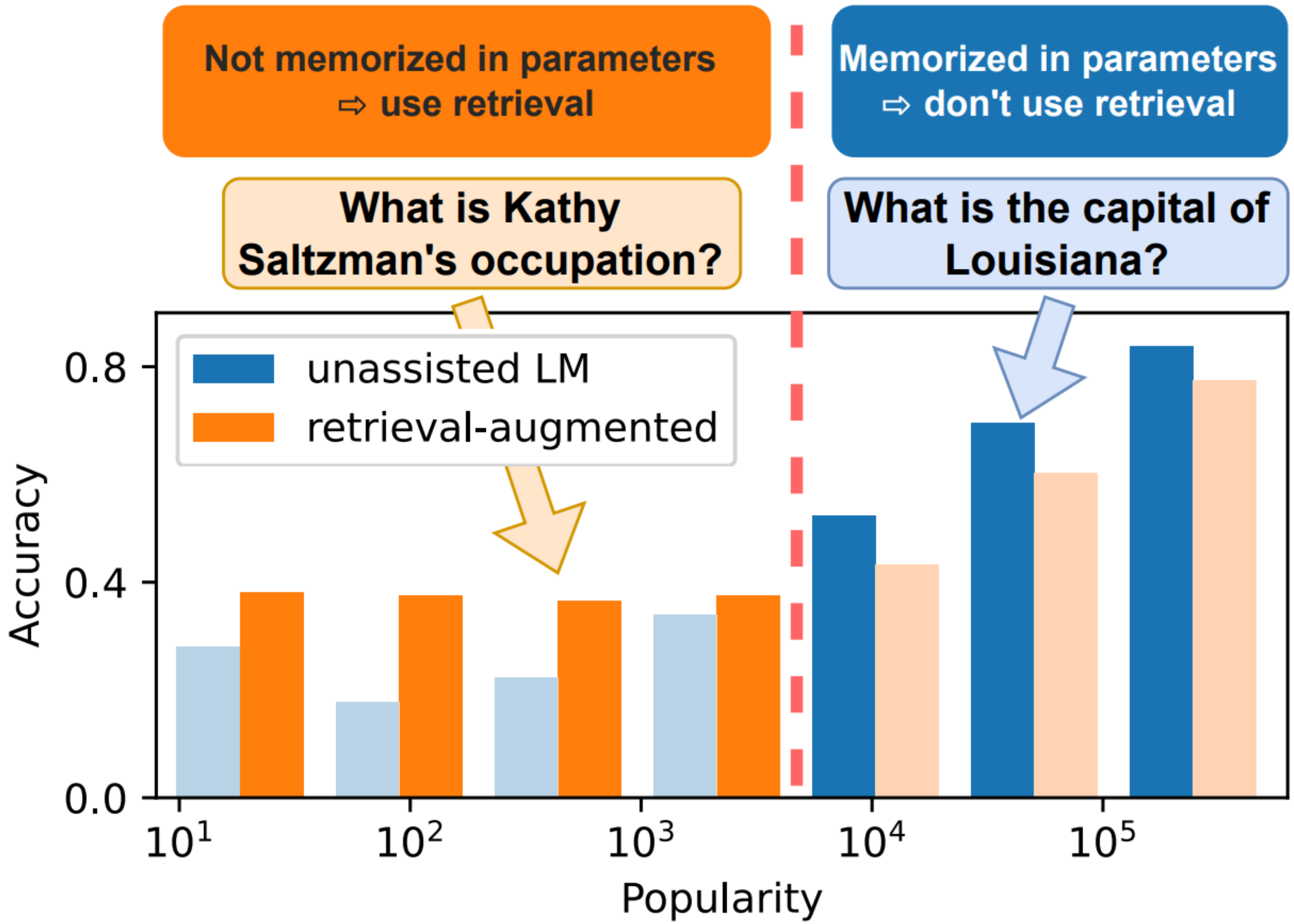
The term summer vacation or summer break refers to a school break in the summer between school years and the break in the academic year.



My best summer vacation was a magical escape to the coastal town of Santorini. The azure waters, charming white-washed building are unforgettable.

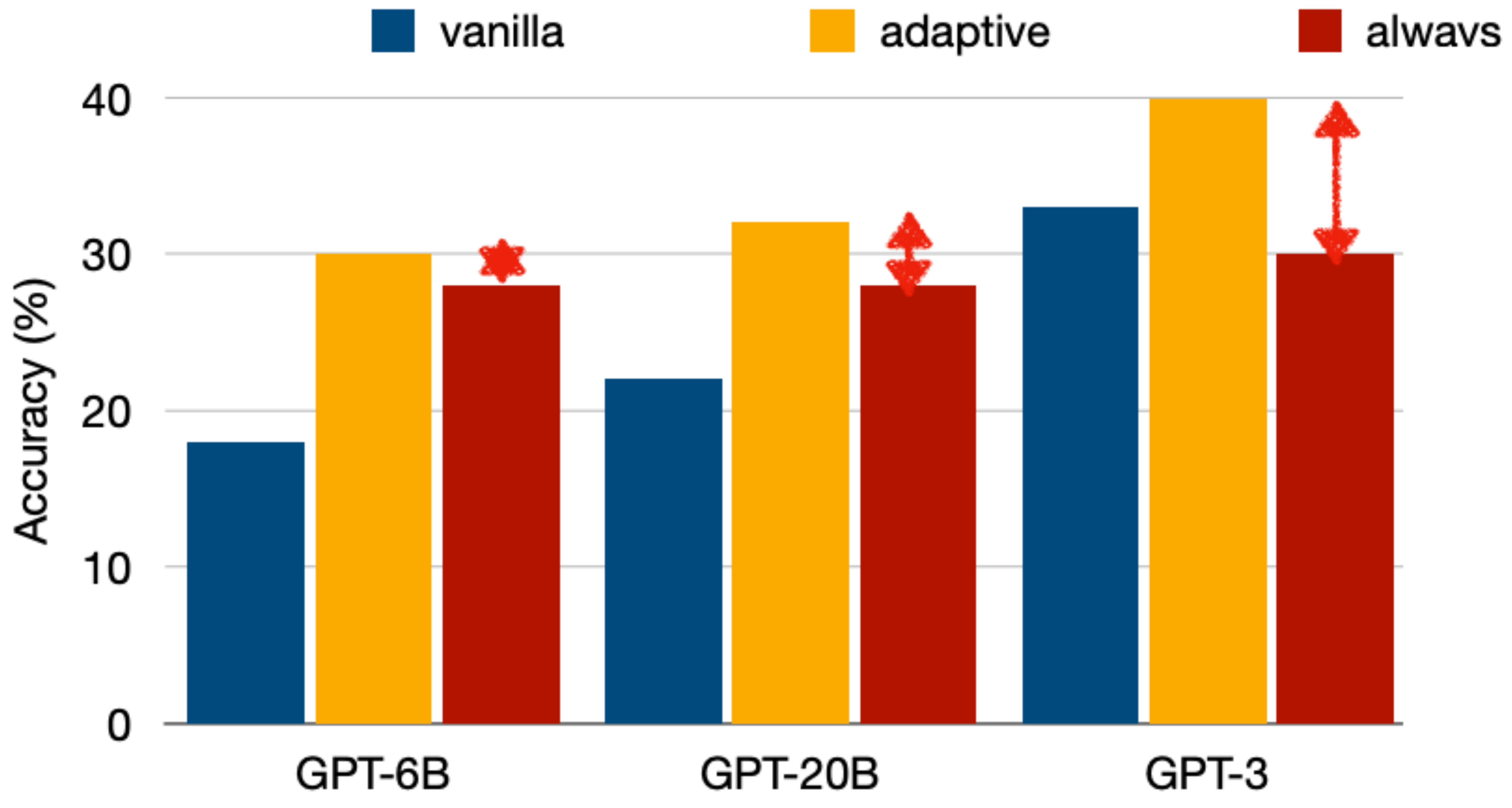
Unnecessary retrieval can hurt performance & efficiency

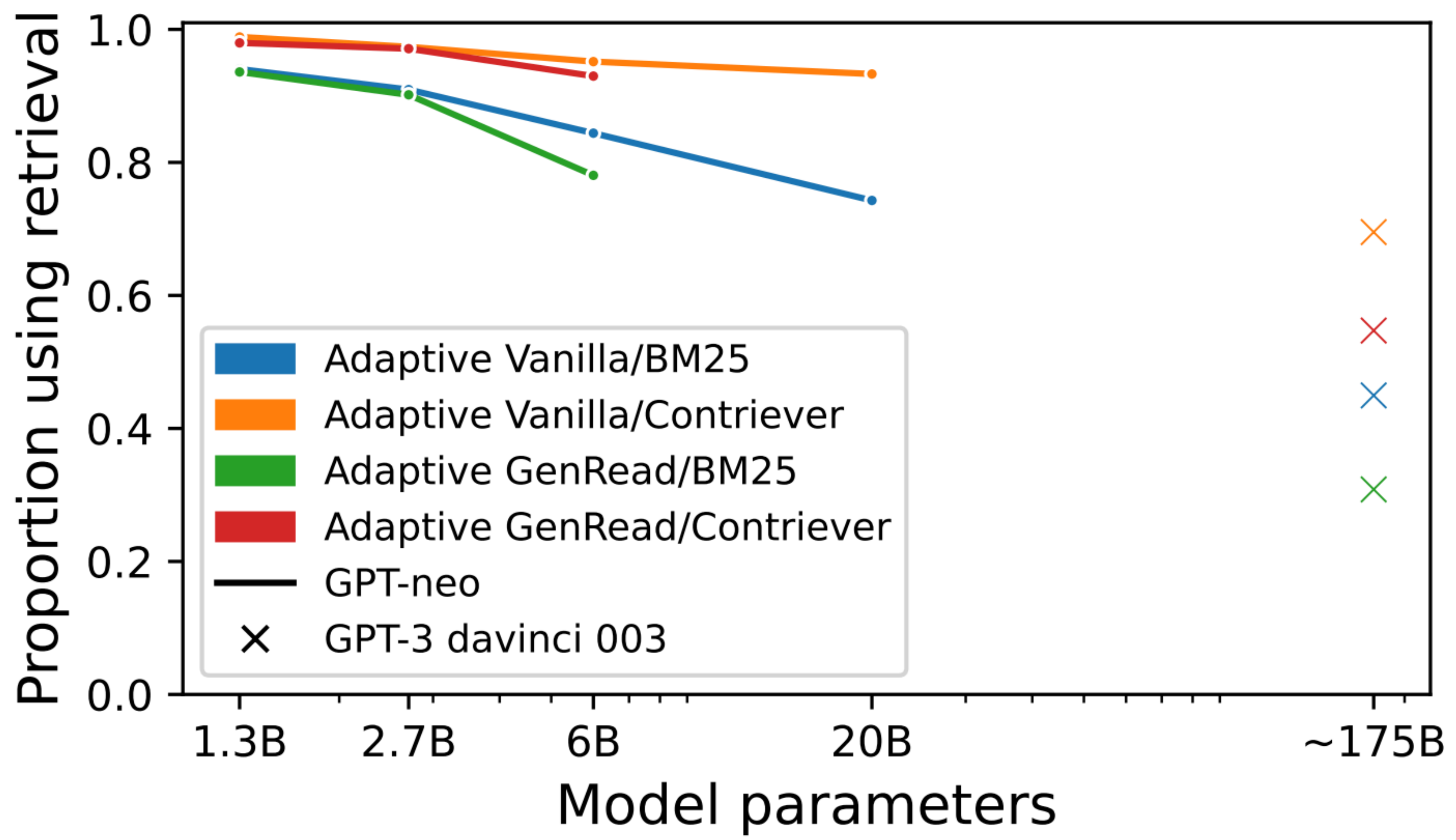
A Simple Solution: a Threshold-based Adaptive Retrieval



Adaptive RAG for Performance

Adaptive RALM improves performance, esp on GPT-3 with 40% less retrieval





Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023.

Self-RAG:
Learning to Retrieve, Generate and Critique through Self-Reflections

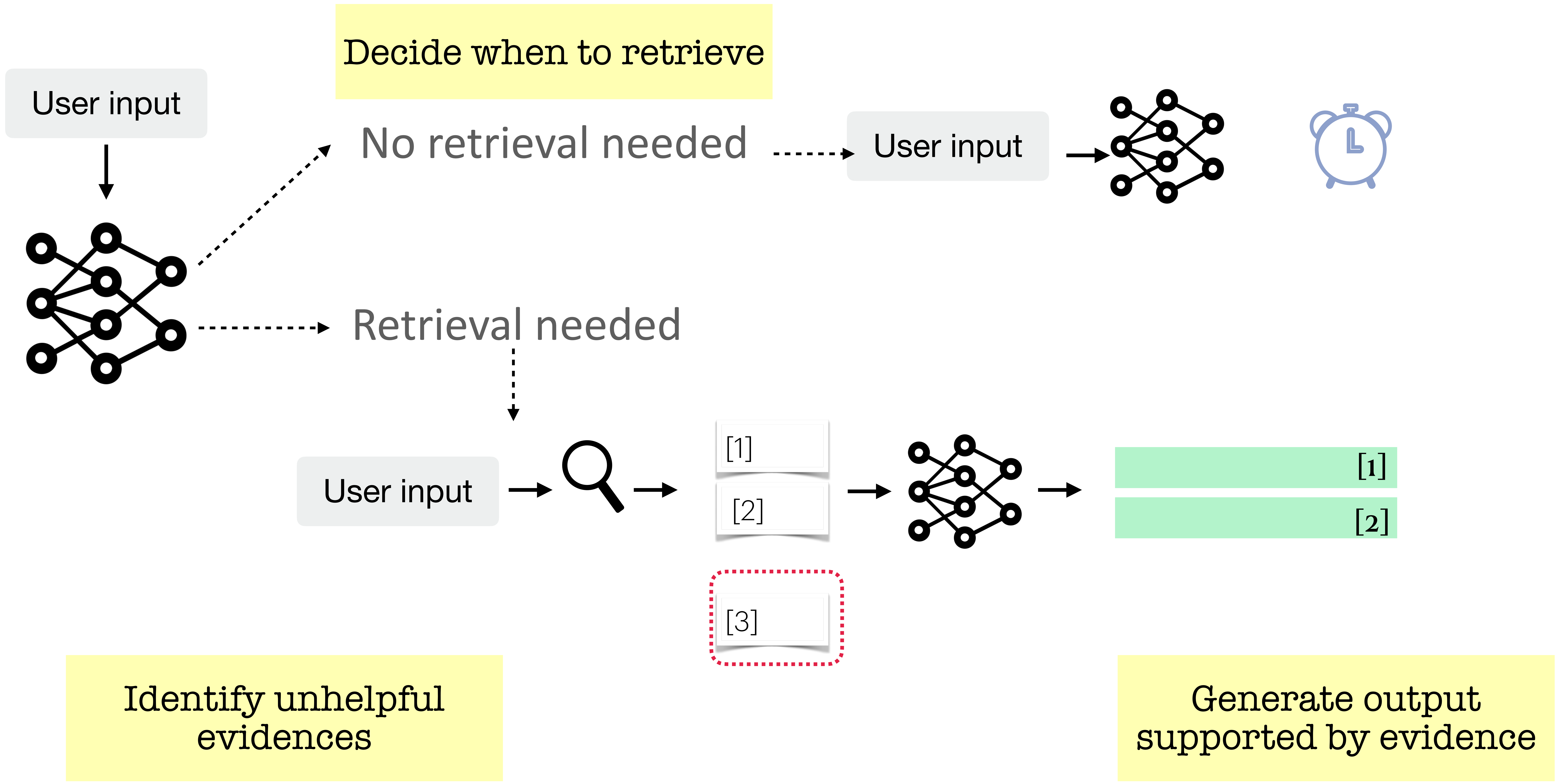
Akari Asai, Zeqiu Wu, Yizhong Wang, Avirul Sil, Hannaneh Hajishirzi





ICLR Oral (Top 1% among 7k+ submissions)
Also at NeurIPS workshop on Instruction tuning (Best paper Honorable Mention)

Goal of Self-RAG — Learn to Retrieve, Generate and Critique



Self-RAG – Self-Reflective Retrieval-Augmented Generation

Step 1: Retrieve documents on demand

How did US states get their names?

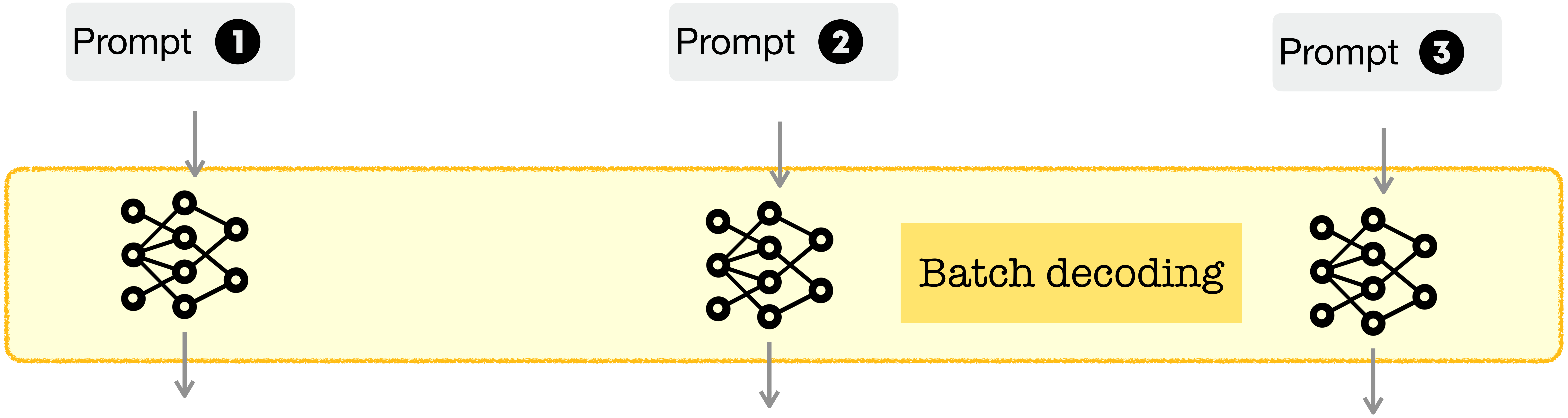
US states got their names from a variety of sources. Retrieve

Reflection tokens

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.

Self-RAG – Self-Reflective Retrieval-Augmented Generation

Step 2: Generate segments in *parallel*



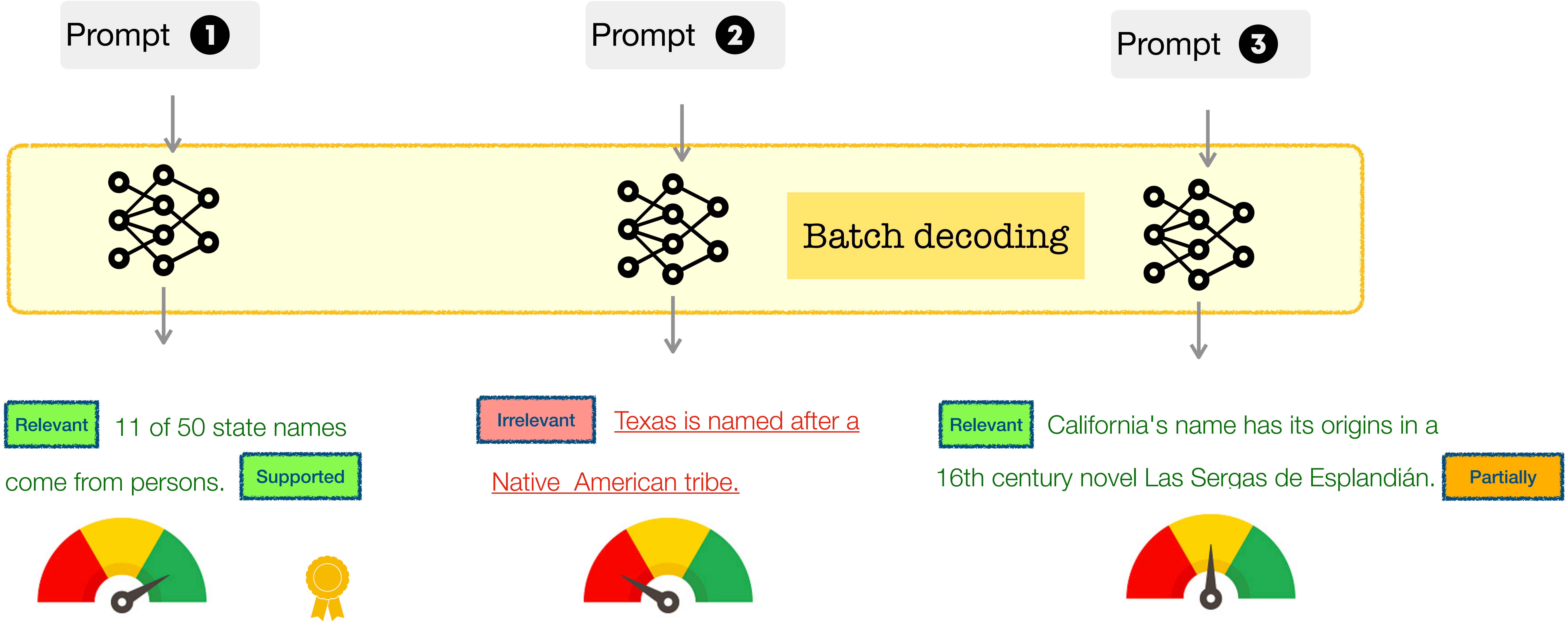
Relevant 11 of 50 state names come from persons.

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.

Self-RAG – Self-Reflective Retrieval-Augmented Generation

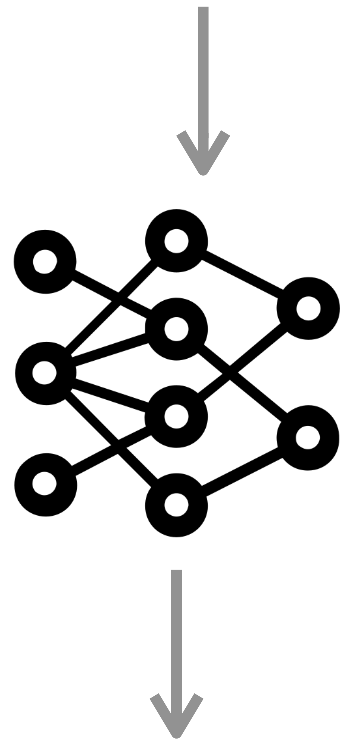
Step 3: Critique and select best segment (Self-reflection-guided decoding)



Self-RAG – Self-Reflective Retrieval-Augmented Generation

Step 1: Generate with no retrieval

Write an essay of your best summer vacation

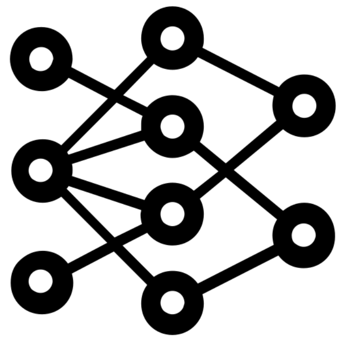


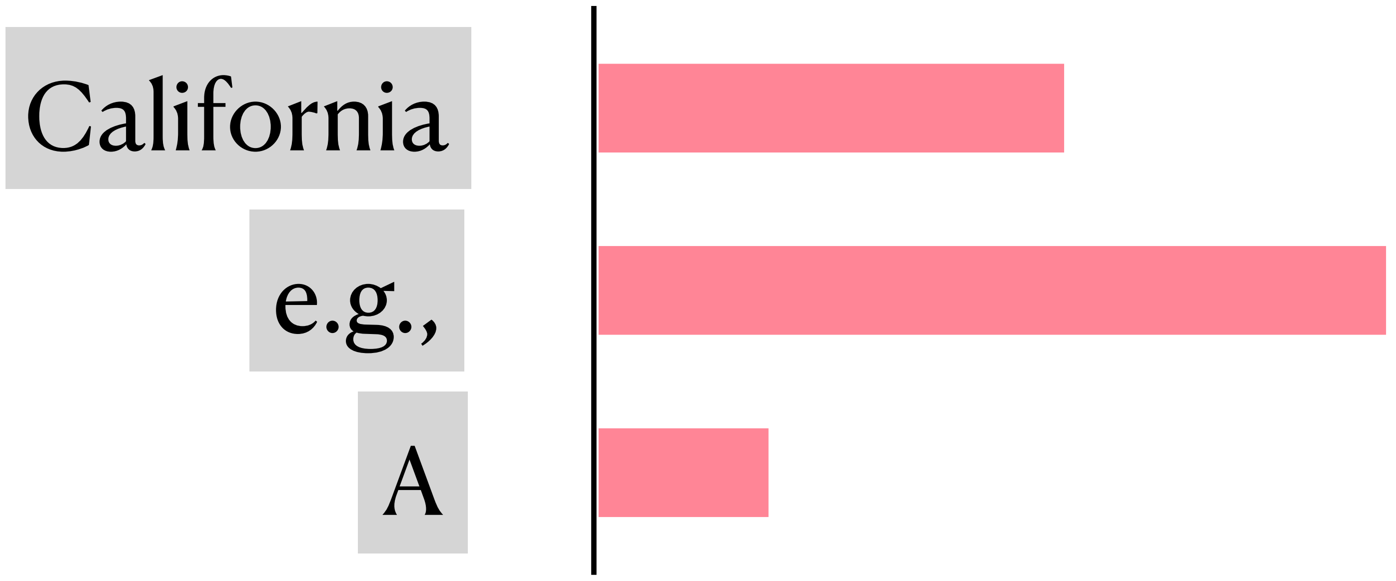
No Retrieval My best summer vacation was a magical escape to the coastal town of Santorini.

The azure waters, charming white-washed building are unforgettable. Util:5

Reflection tokens for Retrieval and Critique

US states got their names from a variety of sources.

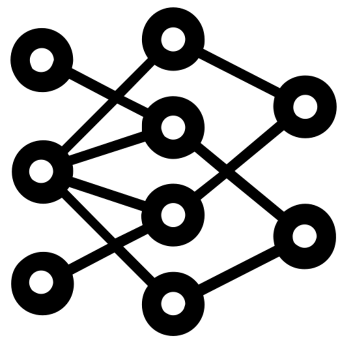




Original LM vocabularies

Reflection tokens for Retrieval and Critique

US states got their names from a variety of sources.



California

e.g.,

A

Original LM vocabularies

Retrieve

No Retrieval

Relevant

Irrelevant

Supported

No support

Useful

Not useful

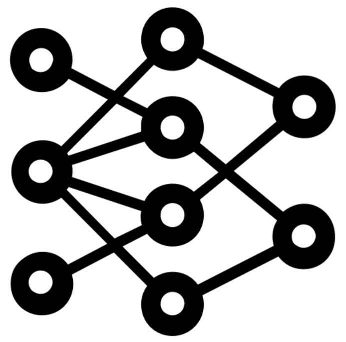
Retrieval tokens

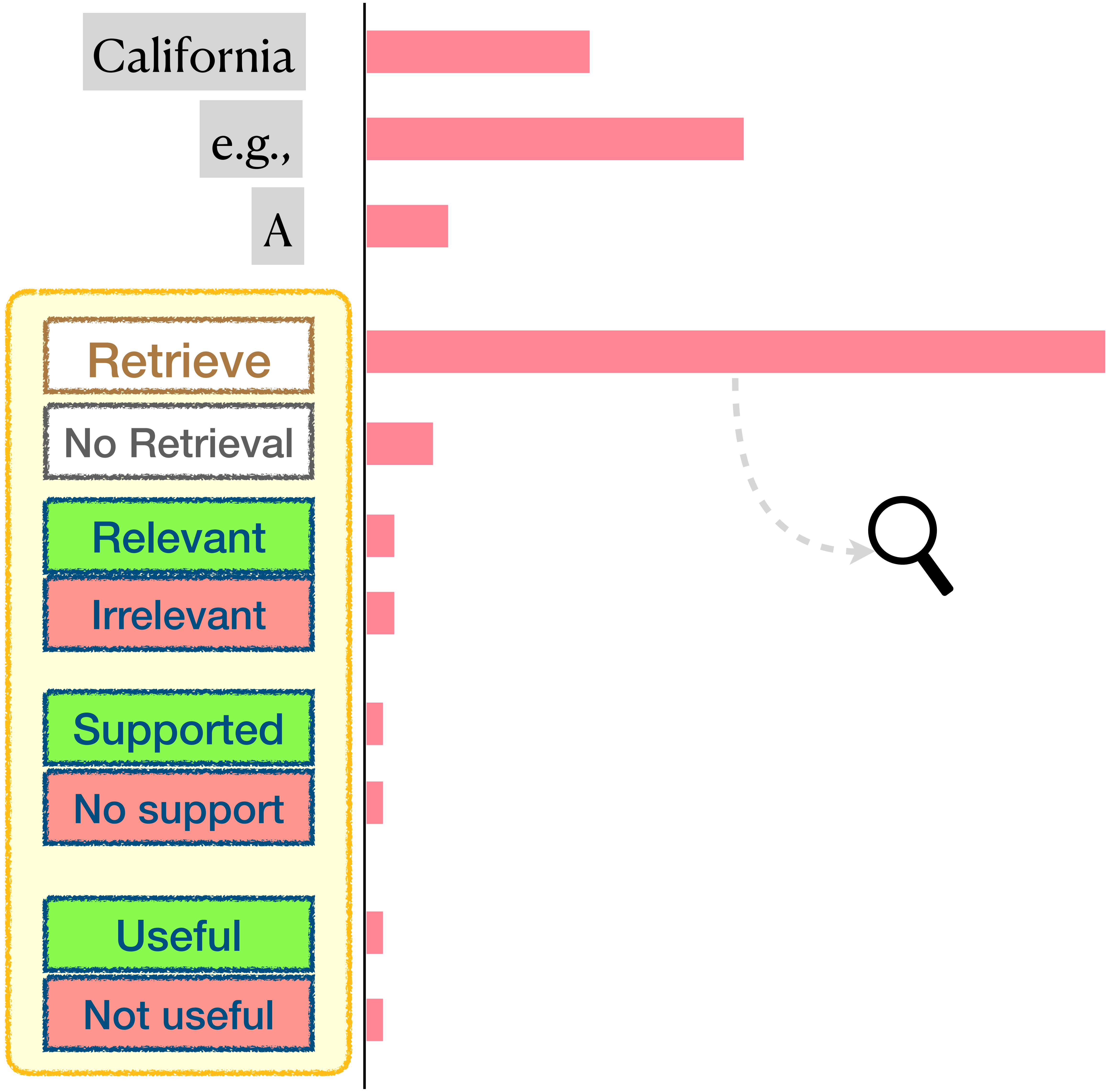
Critique tokens

Vocabulary expanded with reflection tokens

Reflection tokens for Retrieval and Critique

US states got their names from a variety of sources.



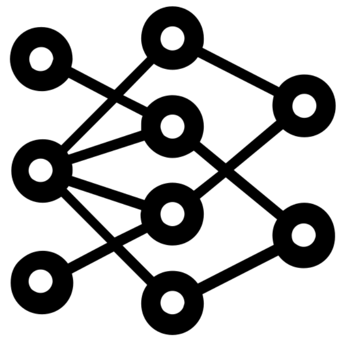


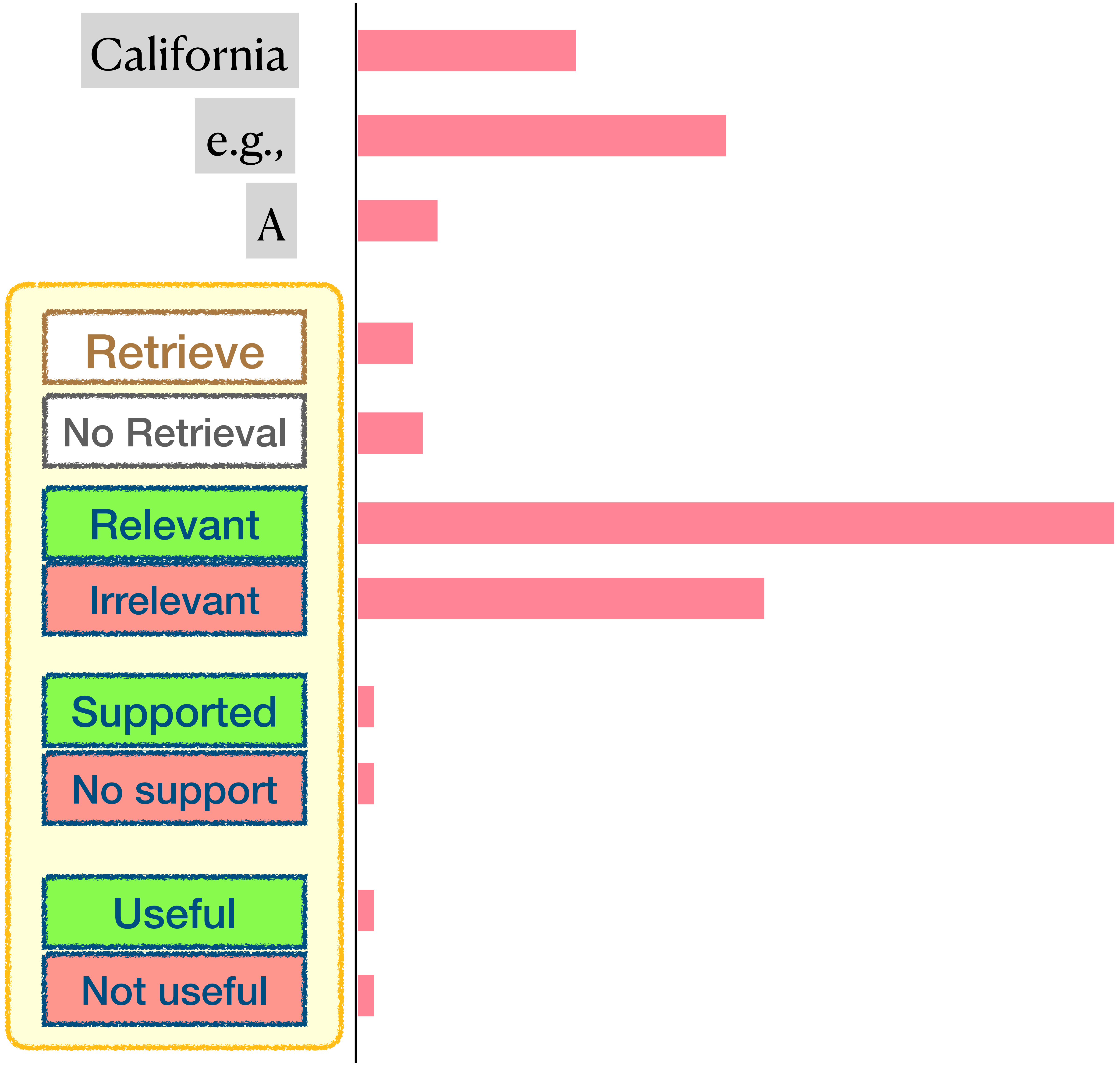
Vocabulary expanded with reflection tokens

Reflection tokens for Retrieval and Critique

US states got their names from a variety of sources.

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

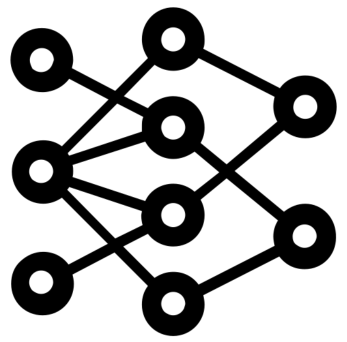




Vocabulary expanded with reflection tokens

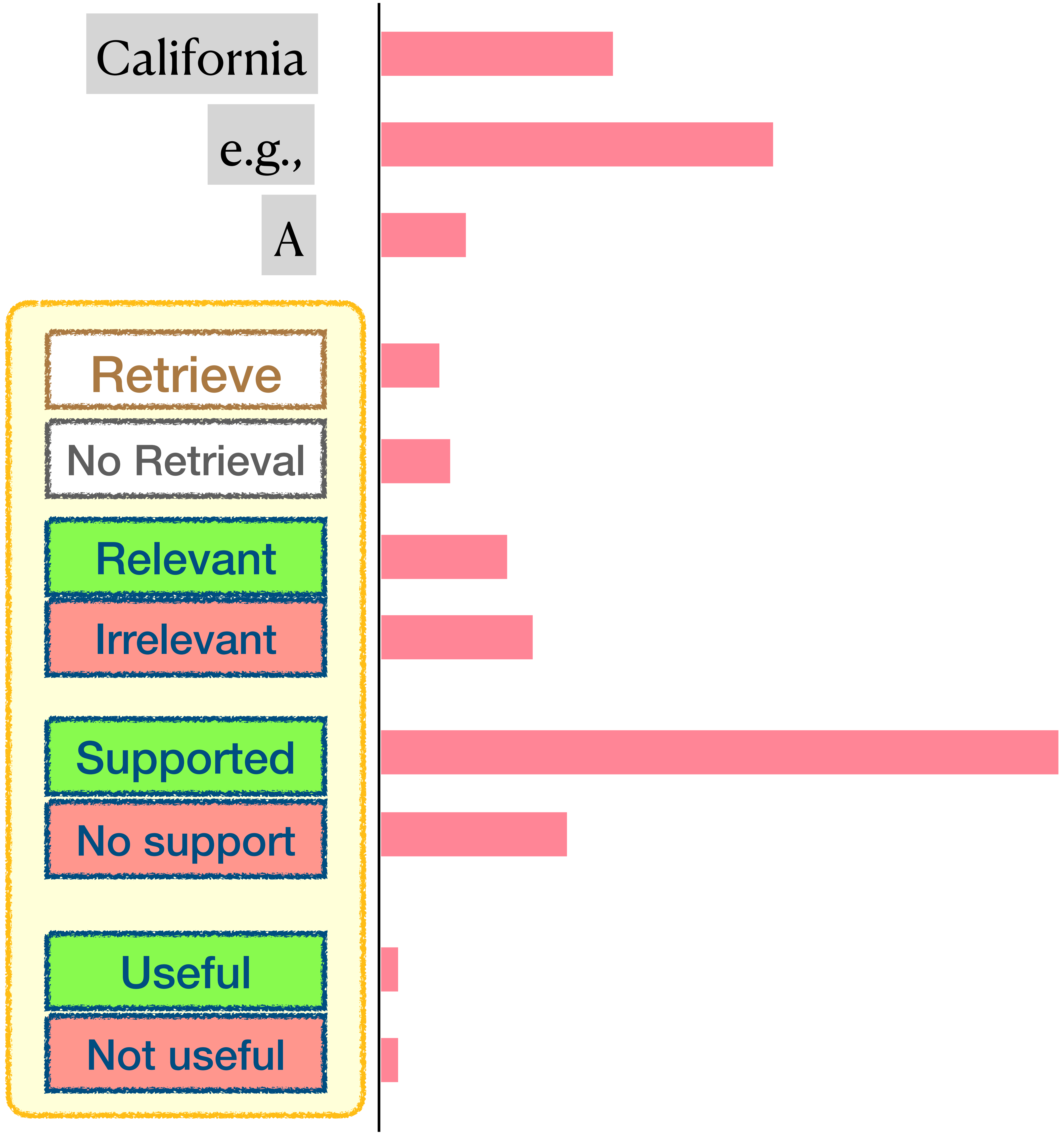
Reflection tokens for Retrieval and Critique

US states got their names from a variety of sources.



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

11 of 50 state names come from persons.

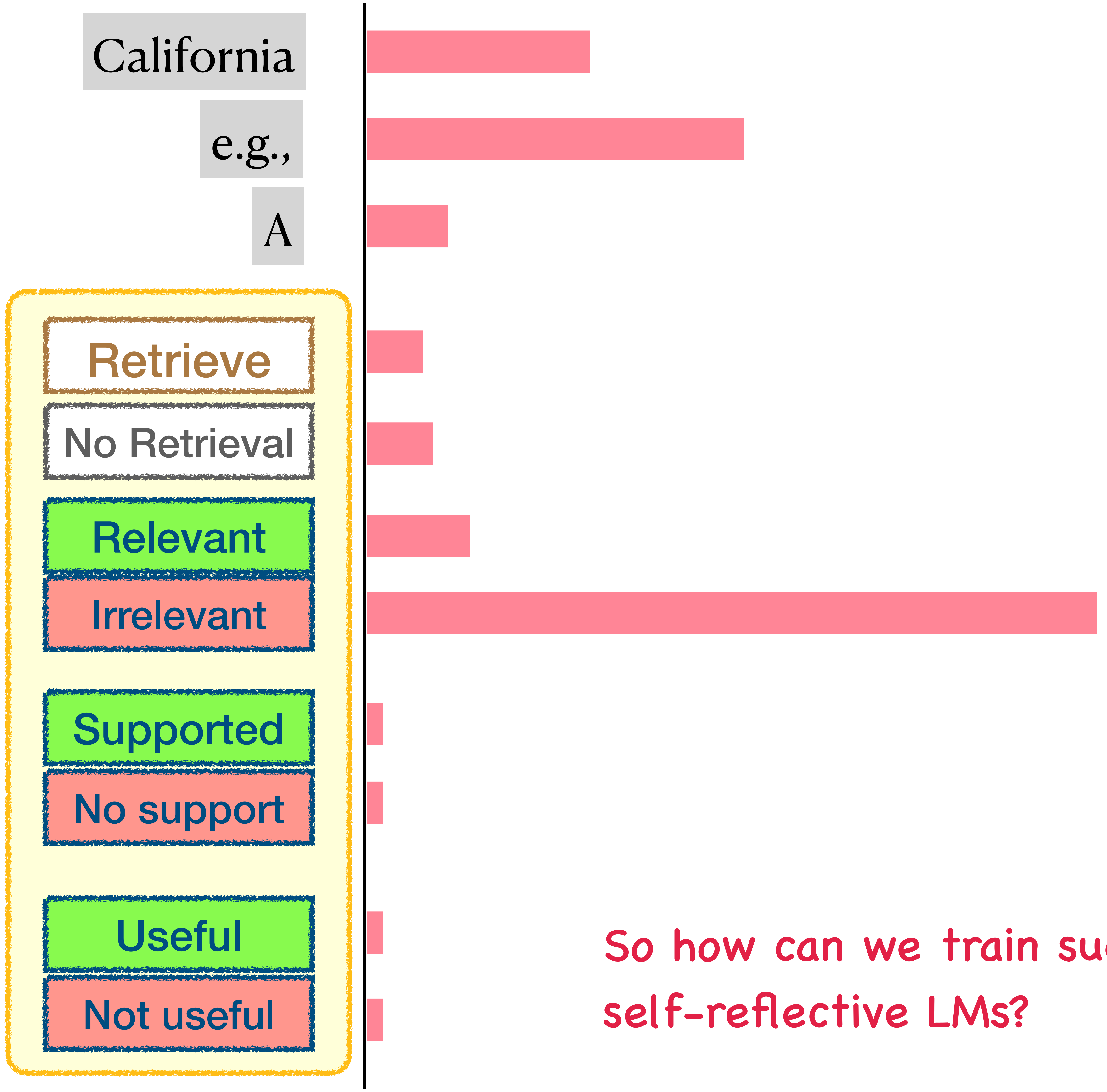


Vocabulary expanded with reflection tokens

Reflection tokens for Retrieval and Critique

US states got their names from a variety of sources.

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



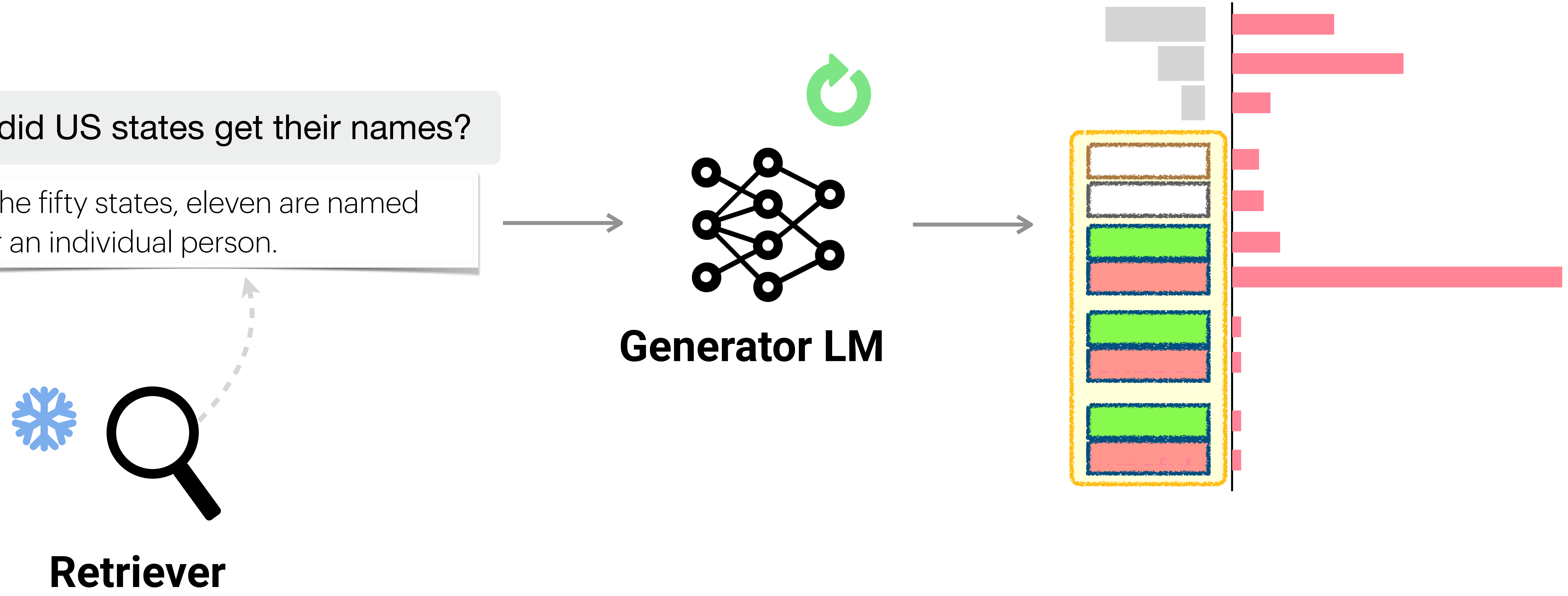
Vocabulary expanded with reflection tokens

So how can we train such self-reflective LMs?

Self-RAG Training

How did US states get their names?

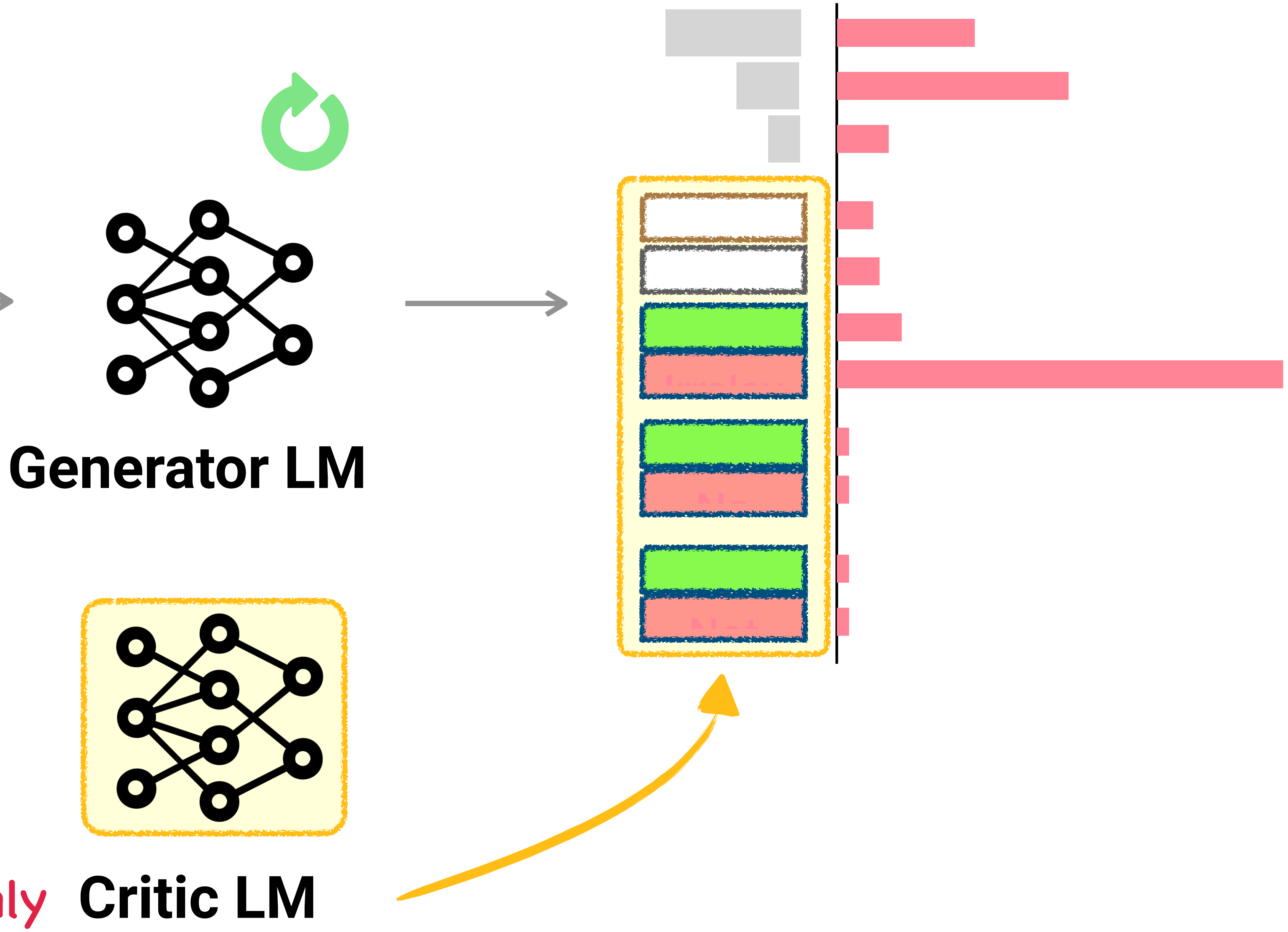
Of the fifty states, eleven are named after an individual person.



Self-RAG Training

How did US states get their names?

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



Critic LM teaches Generator LM to predict reflection tokens

Training time only

Self-RAG Training — Critic

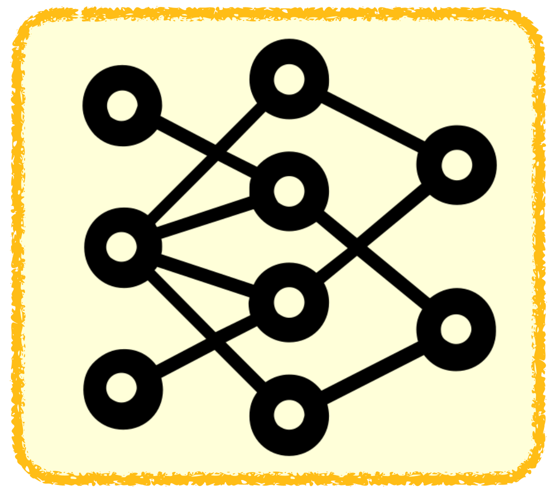
$$\max_{\mathscr{C}} \mathbb{E}_{((x,y,d),r)\sim\mathscr{D}_{critic}} \log p_{\mathscr{C}}(r \mid x, y, d)$$

Evaluate if the output y to an input x is supported by retrieved doc d

x How did US states get their names?

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

y 11 of 50 state names come from persons.



Supported

Critic LM

GPT-4

85-90% acc. on validation set

30k fine-grained feedback, align with human in 90%

$$\mathscr{D}_{critic} = \{(x, y, d), r\}$$

Self-RAG Training - Augmented Instruction-tuning Data

150k (input, output) instances from diverse instruction-following data

Input: Write an essay of your best summer vacation

Output: My best summer vacation was a magical escape to the coastal town of Santorini. The azure waters, charming white-washed building are unforgettable.

Augmented Output: [No Retrieval] My best summer vacation was a magical escape to the coastal town of Santorini. [No Retrieval] The azure waters, charming white-washed building are unforgettable experience. [Util: 5]

Input: How did US states get their names?

Output: 1 of 50 states names come from persons. For instance, Louisiana was named in honor of King Louis XIV of France and Georgia was named after King George II.

Critic LM

Retriever

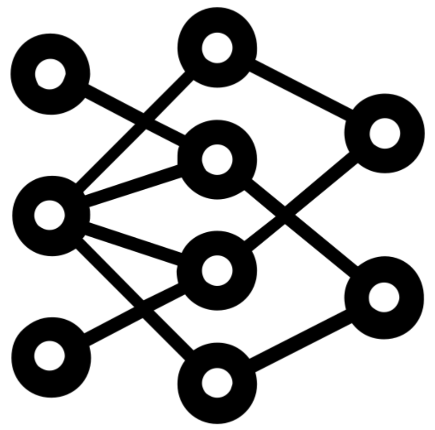
Augmented Output: [Retrieve] 1 <p>Of the fifty states, eleven are named after an individual person</p>. [Relevant] 11 of 50 states' names come from person. [Supported] [Retrieve] 2 <p>LOUISIANA: Named in honor of Louis XIV of France.</p>. [Relevant] For instance, Louisiana was named after King Louis XIV, and Georgia was named after King George II. [Partially] [Util: 5]

Self-RAG Training - Generator

$$\max_{\mathcal{M}} \mathbb{E}_{(x,d,y,r)\sim\mathcal{D}_{gen}} \log p_{\mathcal{M}}(y, r \mid x, d)\,.$$

x How did US states get their names?

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



Generator LM

y 11 of 50 state names come from persons.

r Supported

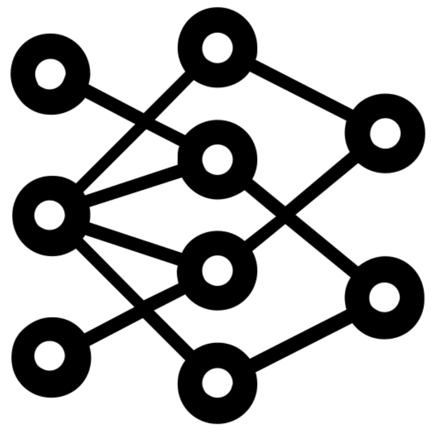
Train with a standard next token objective with expanded vocabulary

Self-RAG Training - Generator

$$\max_{\mathcal{M}} \mathbb{E}_{(x,d,y,r)\sim\mathcal{D}_{gen}} \log p_{\mathcal{M}}(y, r \mid x, d) \,.$$

x How did US states get their names?

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



Generator LM

y 11 of 50 state names come from persons.

r Supported

Memory-efficient & stable training

Easily applied to new pre-trained LM

Customize & **control** via reflection tokens

How?

Self-reflection-guided Decoding

Conduct segment-level beam search to find top k segments

Prompt ❶

$f(\boxed{\text{Relevant}}\ \boxed{\text{Supported}})$

0.9

Prompt ❷

$f(\boxed{\text{Irrelevant}})$

0.1

Prompt ❸

$f(\boxed{\text{Relevant}}\ \boxed{\text{Partially}})$

0.4

$$f(y_t, d, \boxed{\text{Critique}}) = p(y_t|x, d, y_{<t})) + \mathcal{S}(\boxed{\text{Critique}}), \text{where}$$

$$\mathcal{S}(\boxed{\text{Critique}}) = \sum_{G \in \mathcal{G}} w^G s_t^G \text{ for } \mathcal{G} = \{\boxed{\text{IsRel}}, \boxed{\text{IsSup}}, \boxed{\text{IsUse}}\},$$

Self-reflection-guided Decoding

Enable simple model customization by changing weights

Prompt ❶

$f($ Relevant Supported $)$

0.9

Prompt ❷

$f($ Irrelevant $)$

0.1

Prompt ❸

$f($ Relevant Partially $)$

0.4

$$f(y_t, d, \boxed{\text{Critique}}) = p(y_t|x, d, y_{<t})) + \mathcal{S}(\boxed{\text{Critique}}), \text{ where}$$

$$\mathcal{S}(\boxed{\text{Critique}}) = \sum_{G \in \mathcal{G}} \boxed{w^G} s_t^G \text{ for } \mathcal{G} = \{\boxed{\text{IsRel}}, \boxed{\text{IsSup}}, \boxed{\text{IsUse}}\},$$

Experimental Details

Tasks and datasets

- **Closed-set tasks** (classifications, multiple-choice QA)
    - ARC-Challenge (Clark et al., 2018)
    - PubHealth (Zhang et al., 2023)
- **Short-form generation**
    - OpenQA - PopQA
    - Trivia QA (Joshi et al., 2017)
- **Long-form generation**
    - ASQA-ALCE [fluency, citation accuracy, correctness] (Gao et al., 2023)
    - Bio generations [FactScore] (Min et al., 2023)

Experimental Details

More details of training & test are in our paper!

Training details

- **Critic training data:** 4k-20k instances for each type
- **Generator training data:** 150k instruction-following datasets
  - ShareGPT
  - OpenAssistant
  - Alpaca
  - FLANV2
  - Natural Questions ….
- **Base LMs:** Llama2-7B, 13B (Touvron et al., 2023)
- **Computation:** 4*A100 (15 hours)

Experimental Details

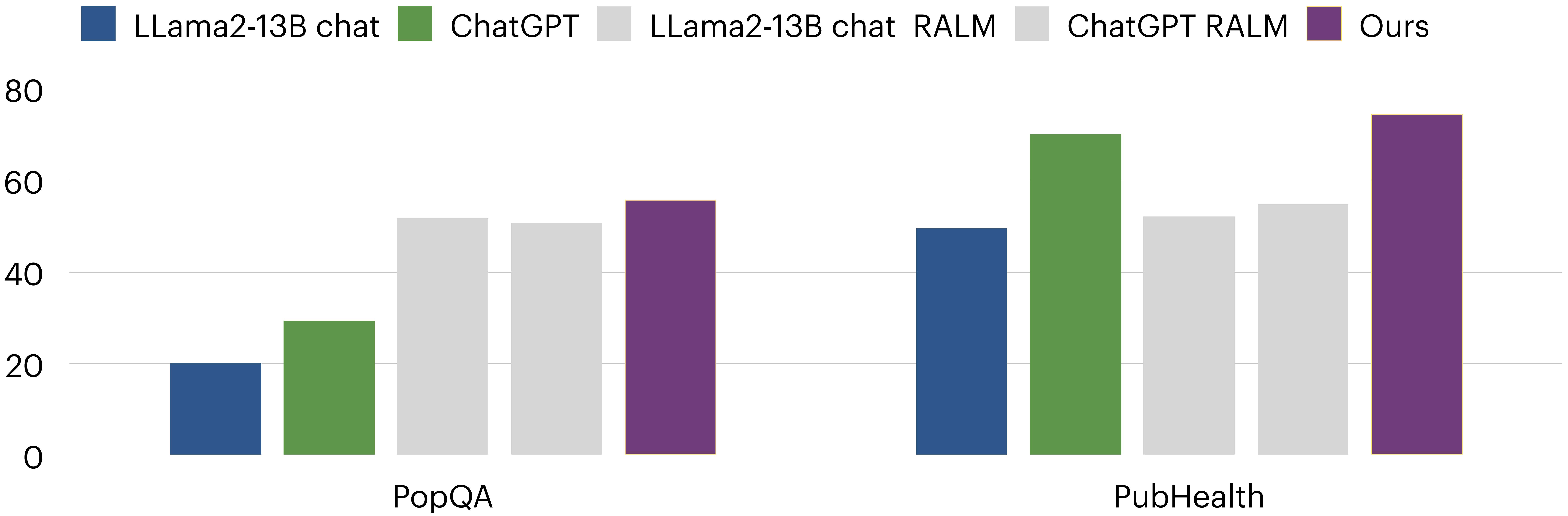
More details of training & test are in our paper!

Inference-time details

- **Retriever Encoder:** Contriever-MS MARCO (Izacard et al., 2022)
- **Index:** HNSW Index (0.1 sec / query) and FLAT Index (5 sec / query)
- **Efficient LM inference:** vllm (Kwon et al., 2023)
- **Tree decoding configuration:** max 200 tokens per depth, max depth of 7

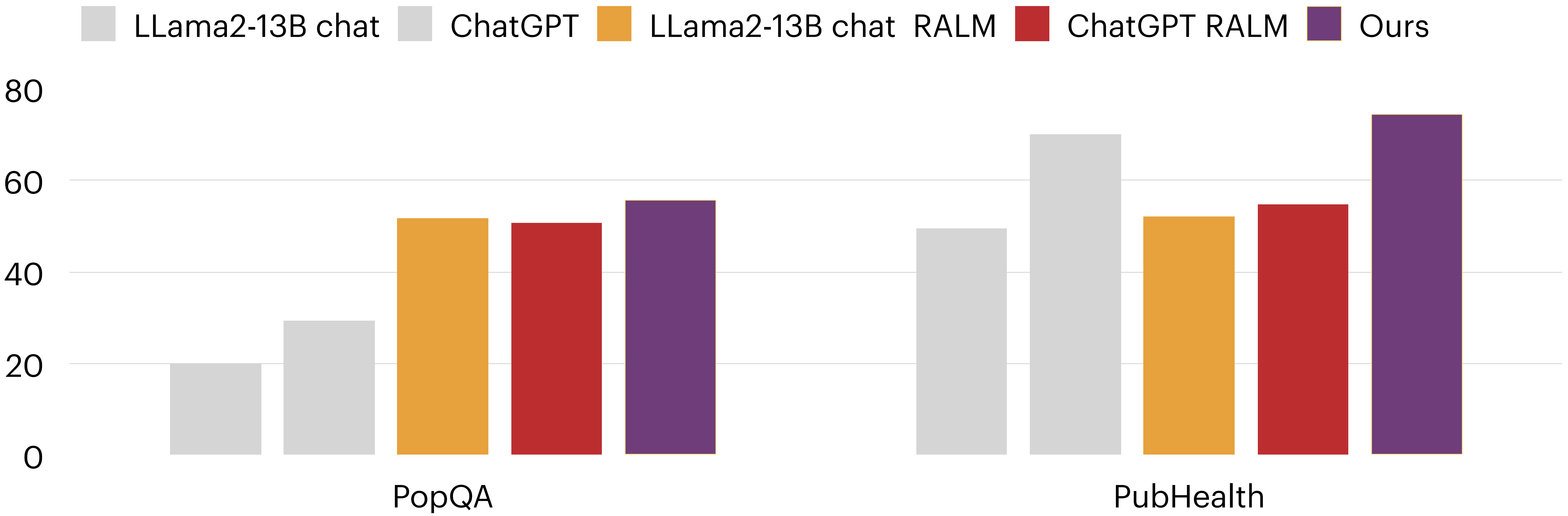
Experimental Results (Short-form & Closed)

Self-RAG outperforms vanilla LMs incl. ChatGPT



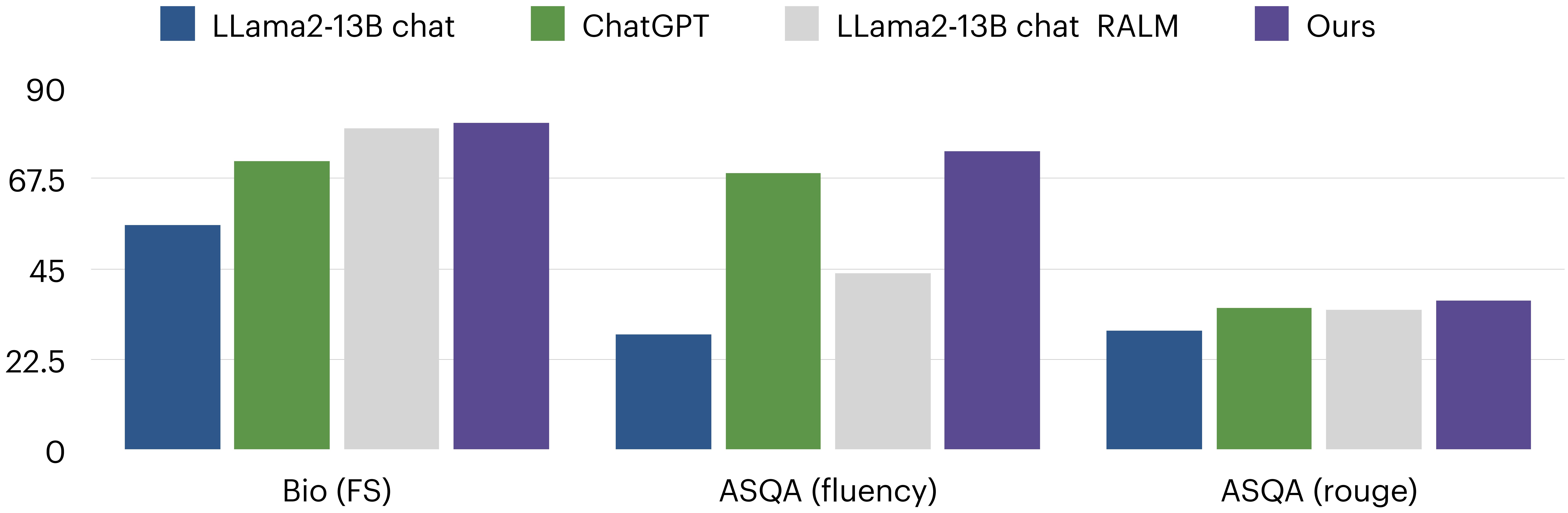
Experimental results (Short-form & Closed)

Self-RAG outperforms standard RAG + LLMs



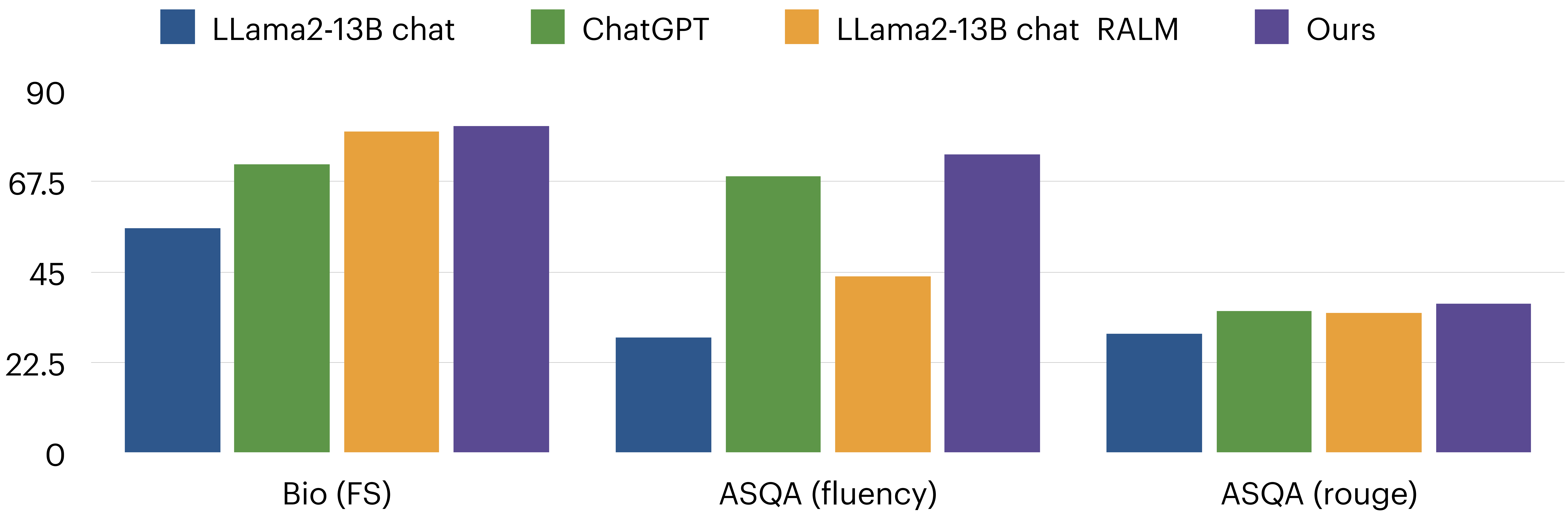
Experimental Results (Long-form)

Outperforms other LMs in terms of factuality & fluency correctness



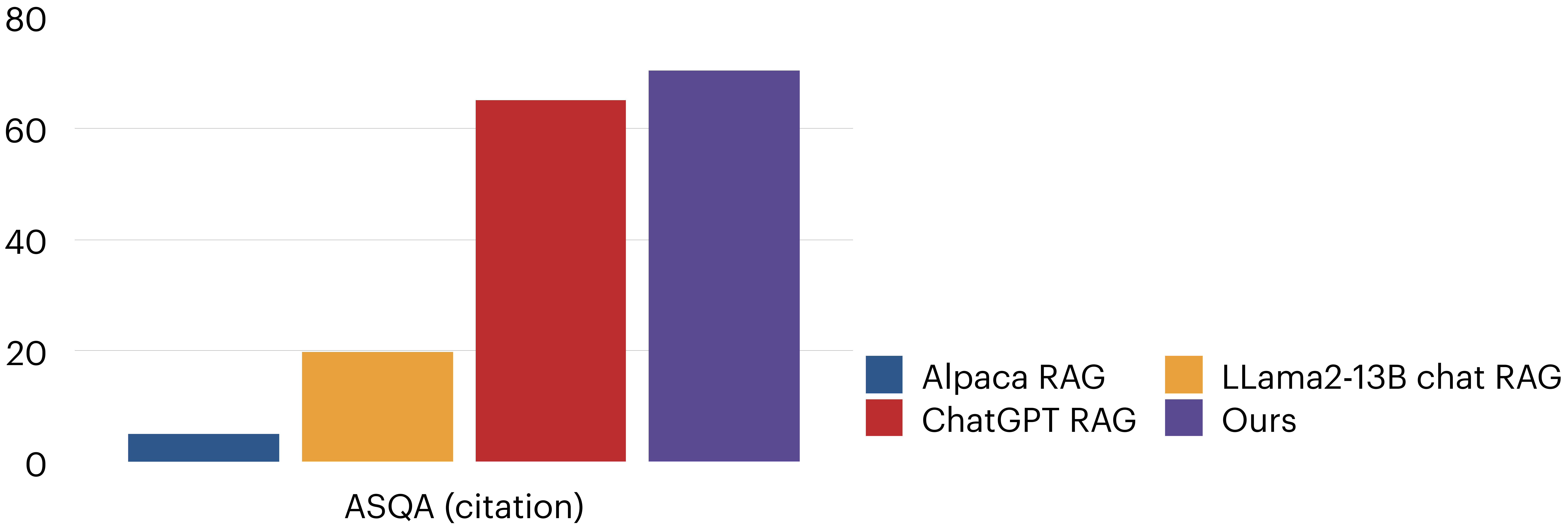
Experimental Results (Long-form)

Outperforms other LMs in terms of factuality & fluency correctness



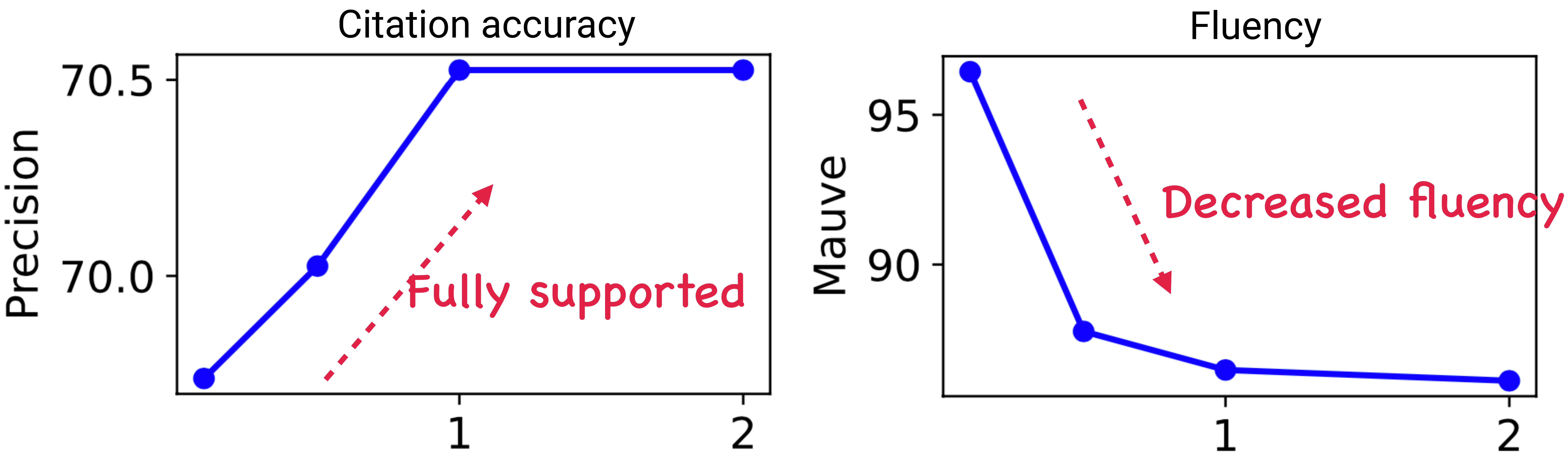
Experimental Results (Long-form Citation Precisions)

Significantly improves llama2-13B citation accuracy, matching ChatGPT



Inference-time Customization via Self-reflection

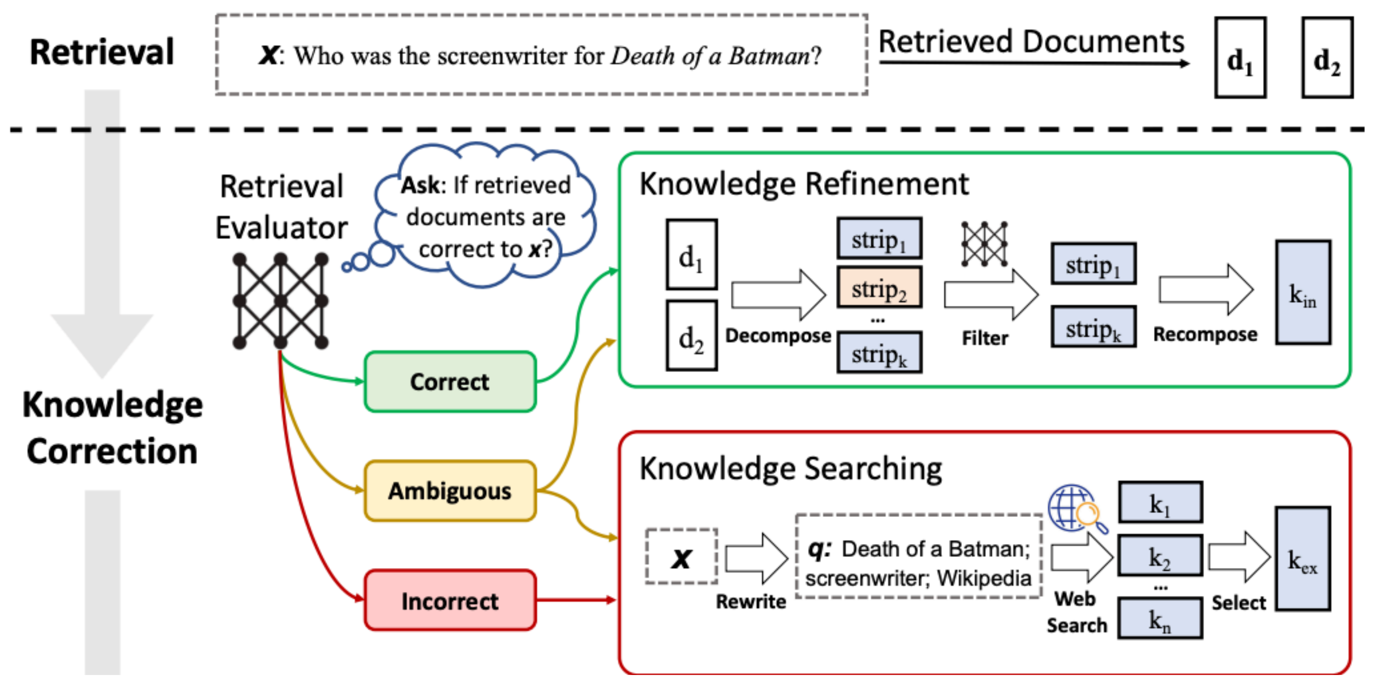
Decoding-time control via reflection tokens change model behaviors



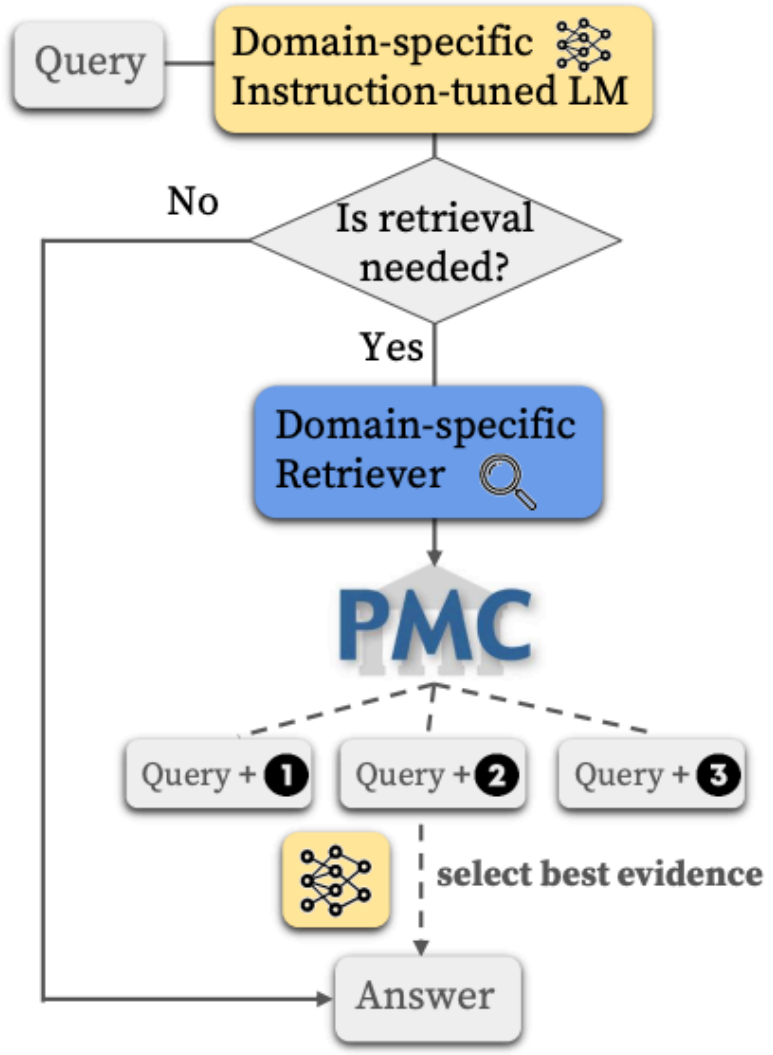
x axis — Weight for **Supported** (larger —> more emphasis on supported)

Impacts on academic communities and applications

Increasing interests in self-RAG style advanced (modular) RAG methods



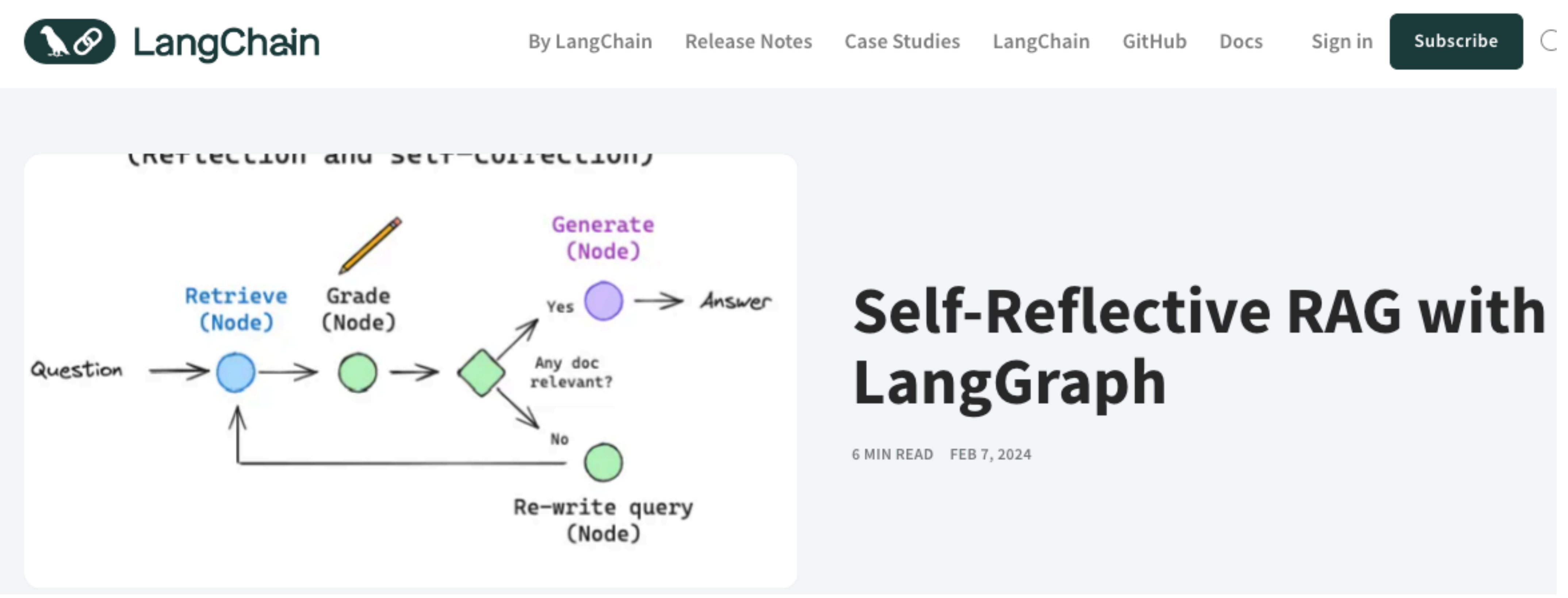
CRAG
(Yang et al., Arxiv 2024)

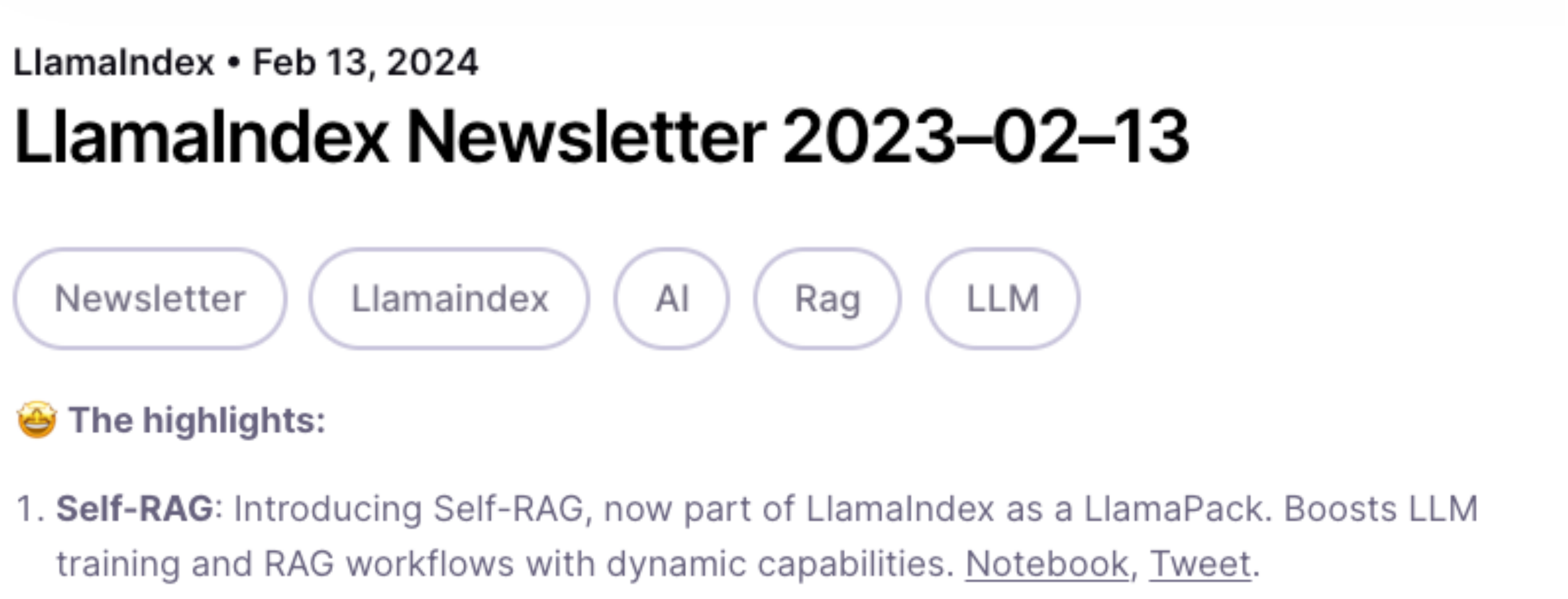


(C) Self-BioRAG (Ours)

Self-BioRAG
(Jeong et al.,
Bioinfomatics 2024)

Integrated into major RAG libraries (LlamaIndex, LangChain ... etc)





Self-RAG – Self-Reflective Retrieval-Augmented Generation

- An LM learns to retrieve, generate and critique
- Instruction-tuned LMs trained with fine-grained reflection tokens
- Outperforms other LMs in six tasks, improving citation accuracy

https://selfrag.github.io/
https://arxiv.org/abs/2212.10511
https://huggingface.co/selfrag/selfrag_llama2_7b (13b)
https://github.com/AkariAsai/self-rag (1.1k ⭐!)

Today's lecture

Promises and Limitations of Retrieval-augmented LMs

Reliable inference: Self-reflective RAG with dynamic retrieval

Versatile Retriever: Intent-aware retrievers with LMs

Summary and Future directions: RAG in the wild

Standard Retrieval Systems — Lexical Retrievers

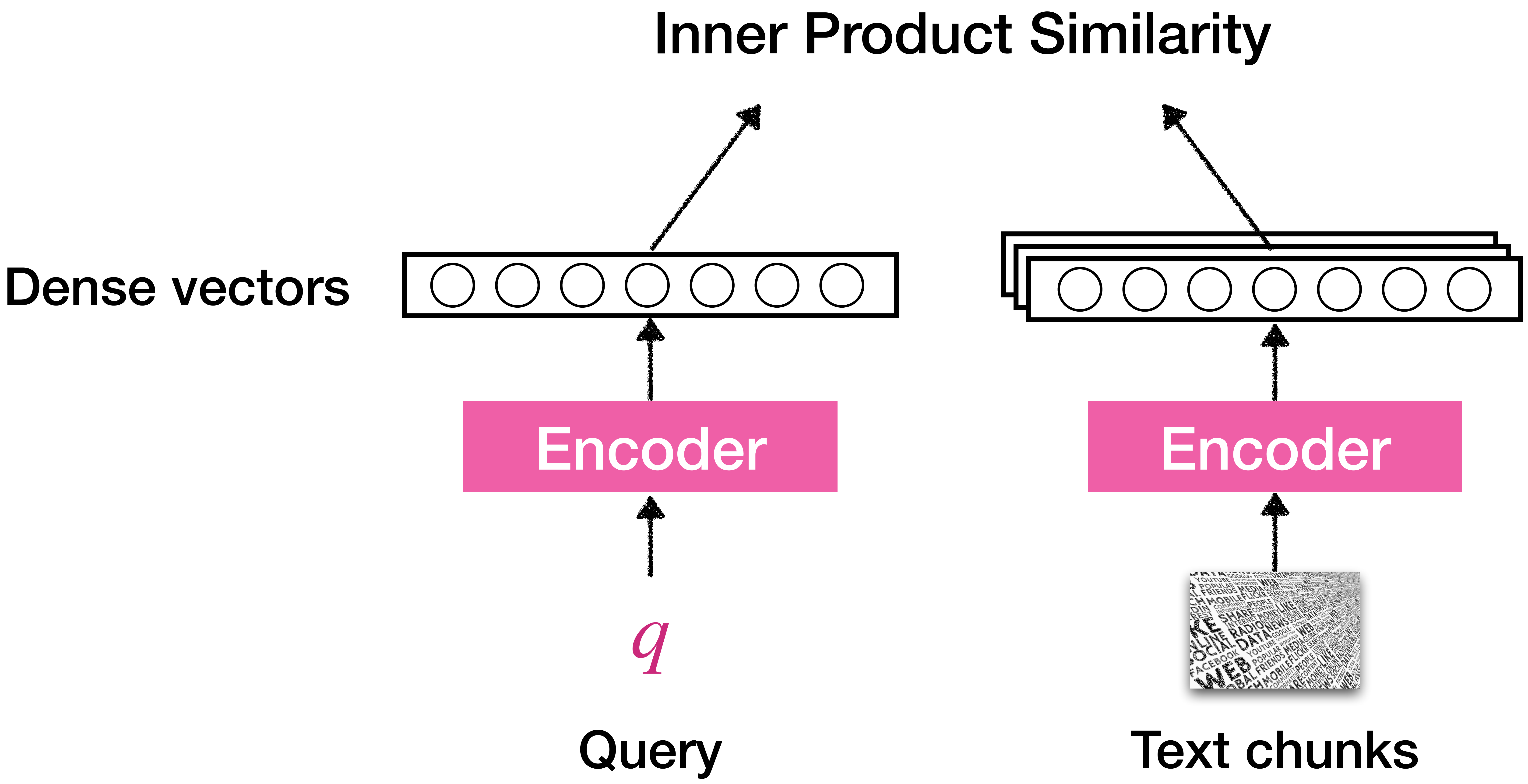
Text chunks		Sparse vectors
In 1997, Apple merged with NeXT, and Steve Jobs became CEO of …	→	[0, 0, 0.4, 0, 0.8, 0.7, …]
Jobs returned to Apple as CEO after the company's acquisition …		[0, 1.2, 0.4, 0, 0.8, 0, …]

Relies on Exact Word Overlap (e.g., CEO v.s. Chief Executive Officers)

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

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

Standard Retrieval Systems — Semantic Retrievers

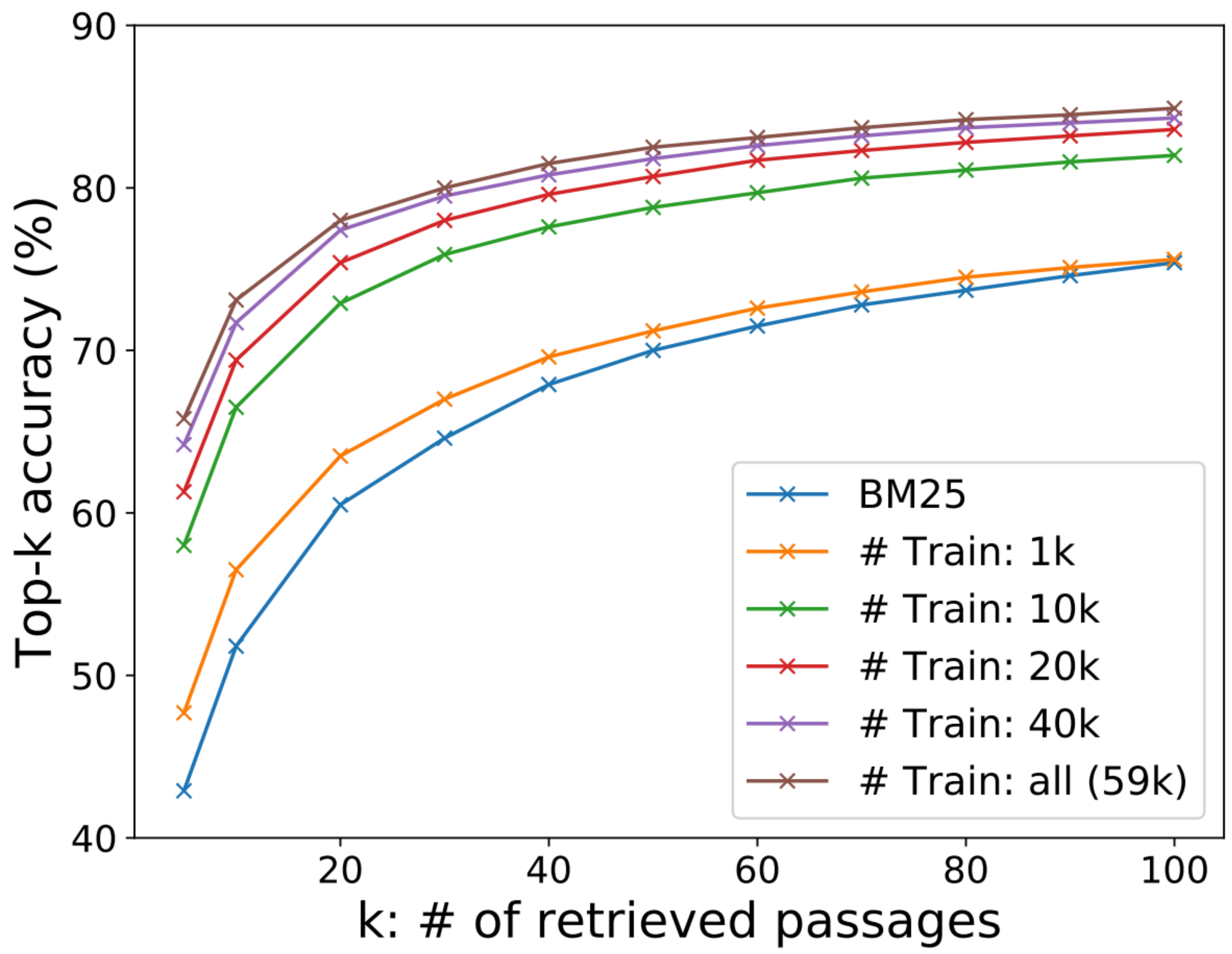


Better capture semantic similarities but often relies on QA / paraphrase training data

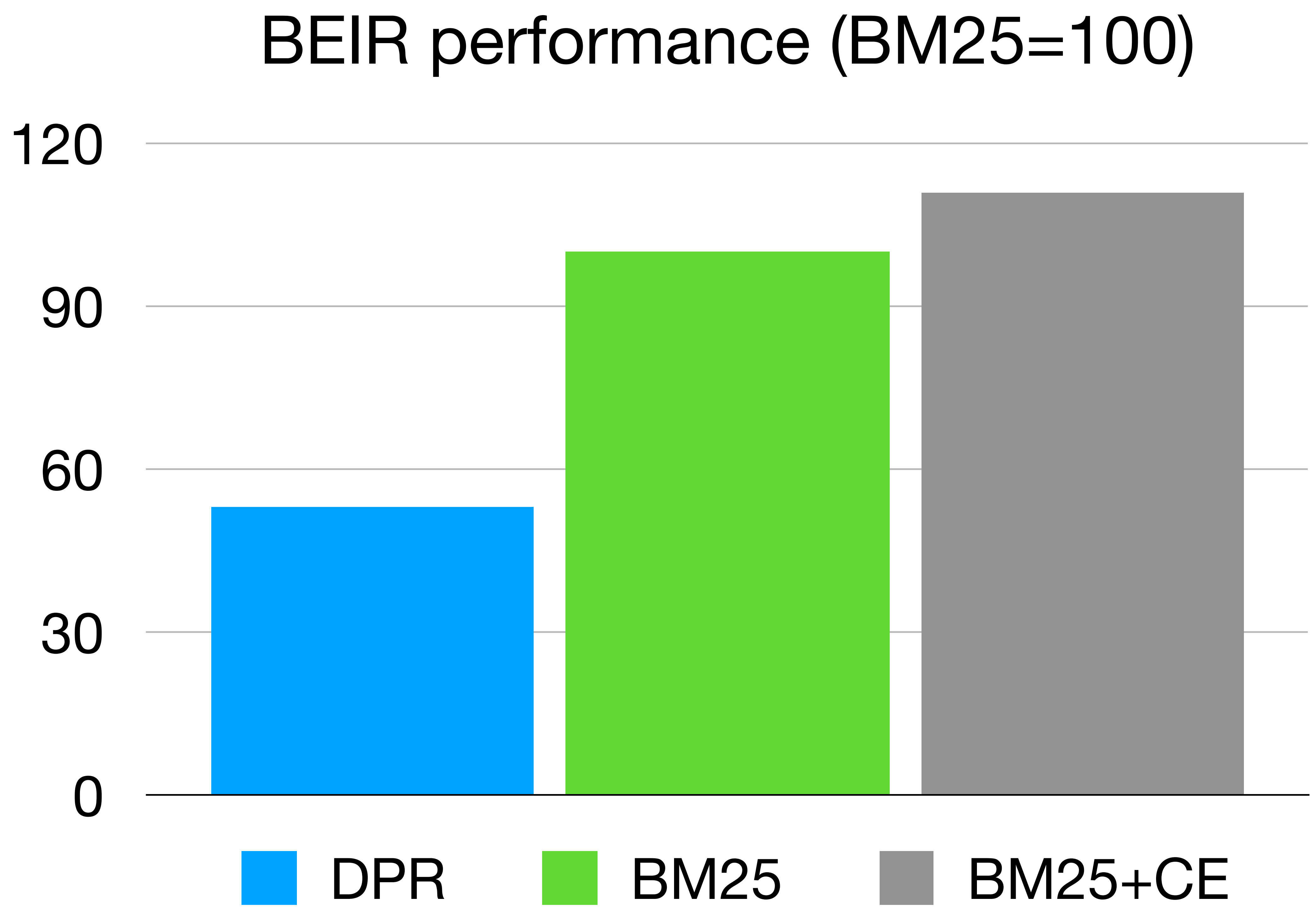
Karpukhin et al.. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

Success of Neural Retrieval Systems

Perform well with more task training data



Struggle in new tasks & domains



Thakur, Nandan, et al. "BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models." *In NeurIPS (Benchmark)* 2021.

Beyond Semantic and Lexical-similarity based Search

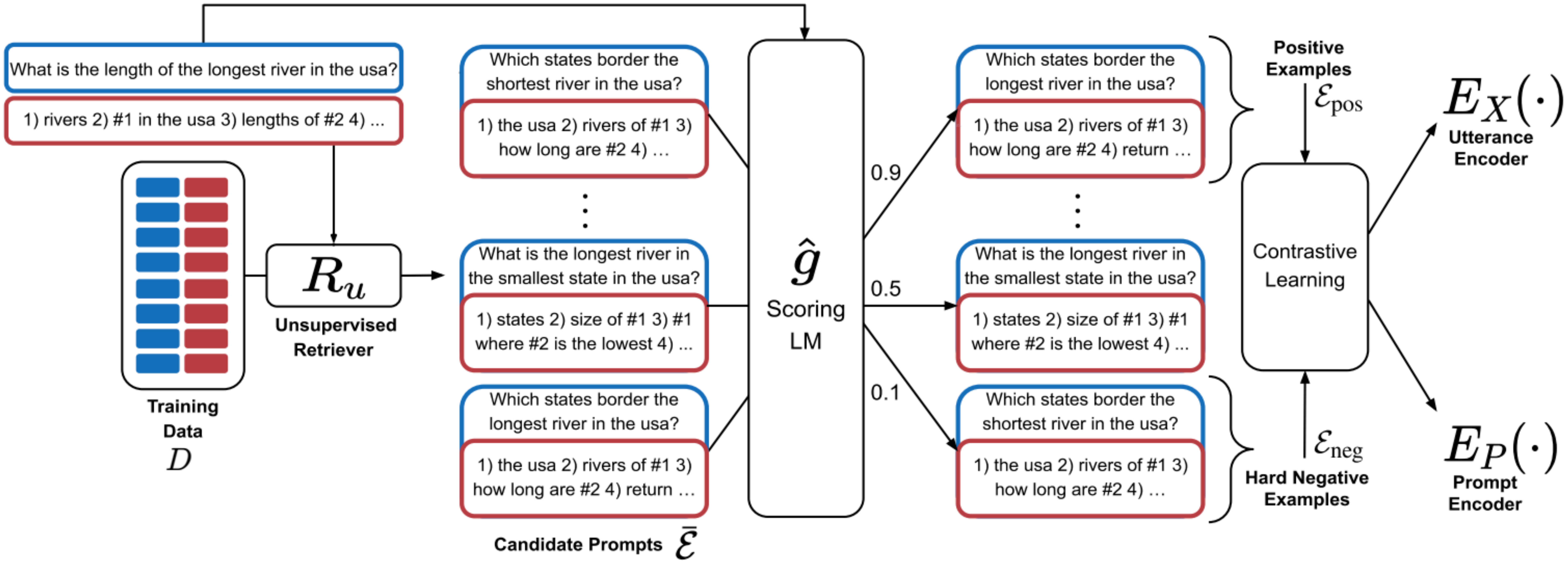
Retrieving helpful text for tasks like reasoning remains challenging

Model	Query	Statements	Prediction
DPR + FiD	In a zoo located in a warm region, what should be included in the polar bear exhibit?	+ If an animal lives a certain environment then that animal usually requires that kind of environment. - Polar bears live in **cold environments**.	warm
Contriever + ATLAS	What keeps the Moon orbiting Earth?	+ Moons orbit planets. - **Gravity** causes orbits.	elliptical
*k*NN-LM	The robot will weigh less on mars than earth but will have the same [MASK]. Targets: *mass* vs *mars*	+ As the force of gravity decreases, the weight of the object will decrease. - The gravitational force of a planet does not change the **mass** of an object on that planet or celestial body.	mars

BehnamGhader et al., Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model. Findings of EMNLP 2023.

Beyond Semantic and Lexical-similarity based Search

Training task-specific retrievers for better demonstrations have shown to be effective — Can we build more versatile retrievers?

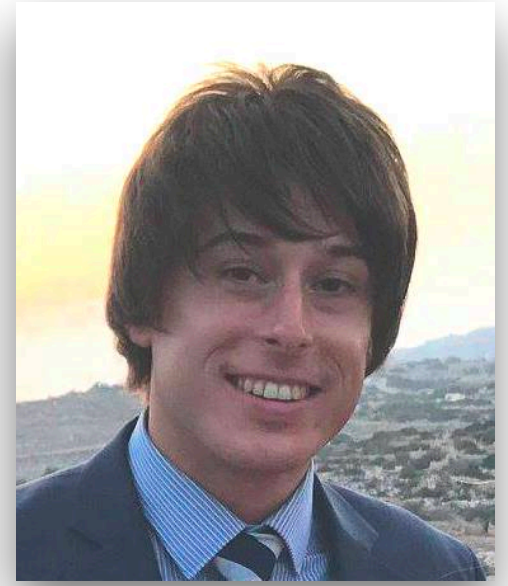
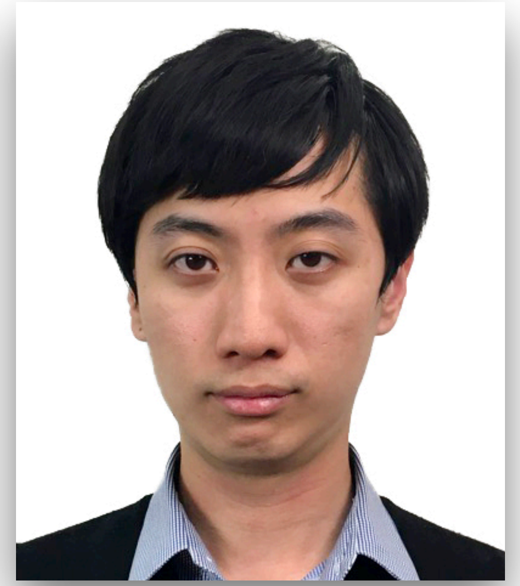


Rubin et al., Learning To Retrieve Prompts for In-Context Learning. ACL 2022.

Task-aware Retrieval with Instructions

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard,
Sebastian Riedel, Hannaneh Hajishirzi, Wen-tau Yih

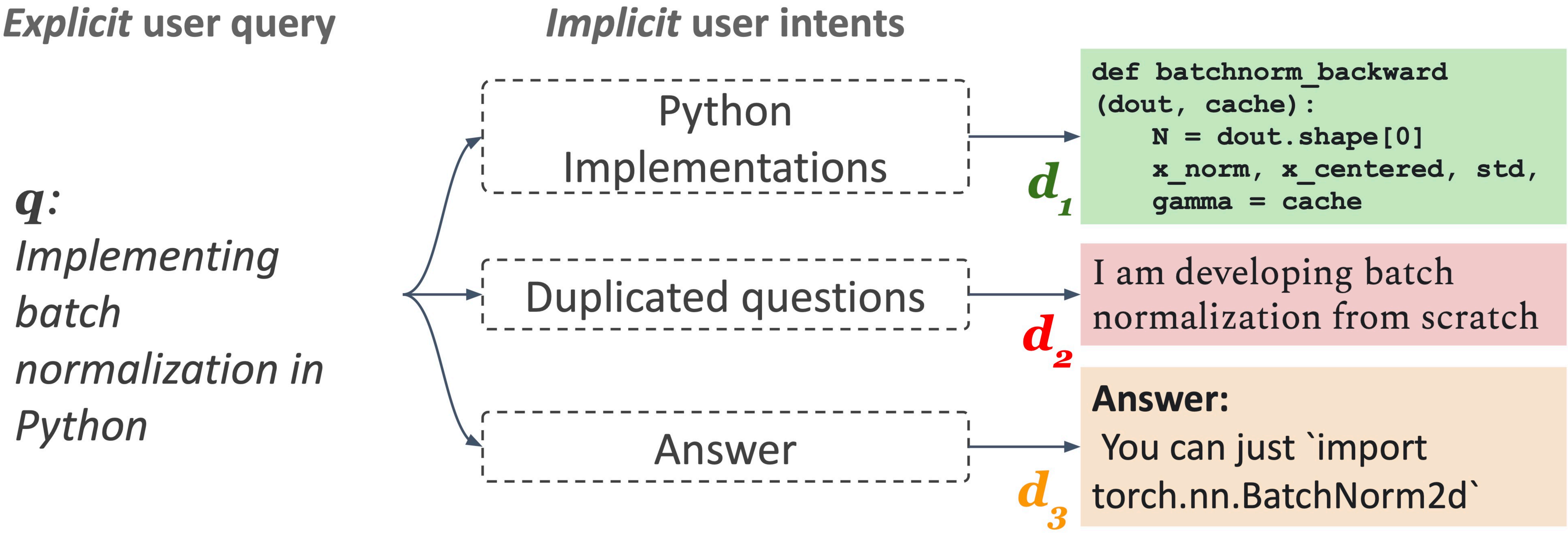


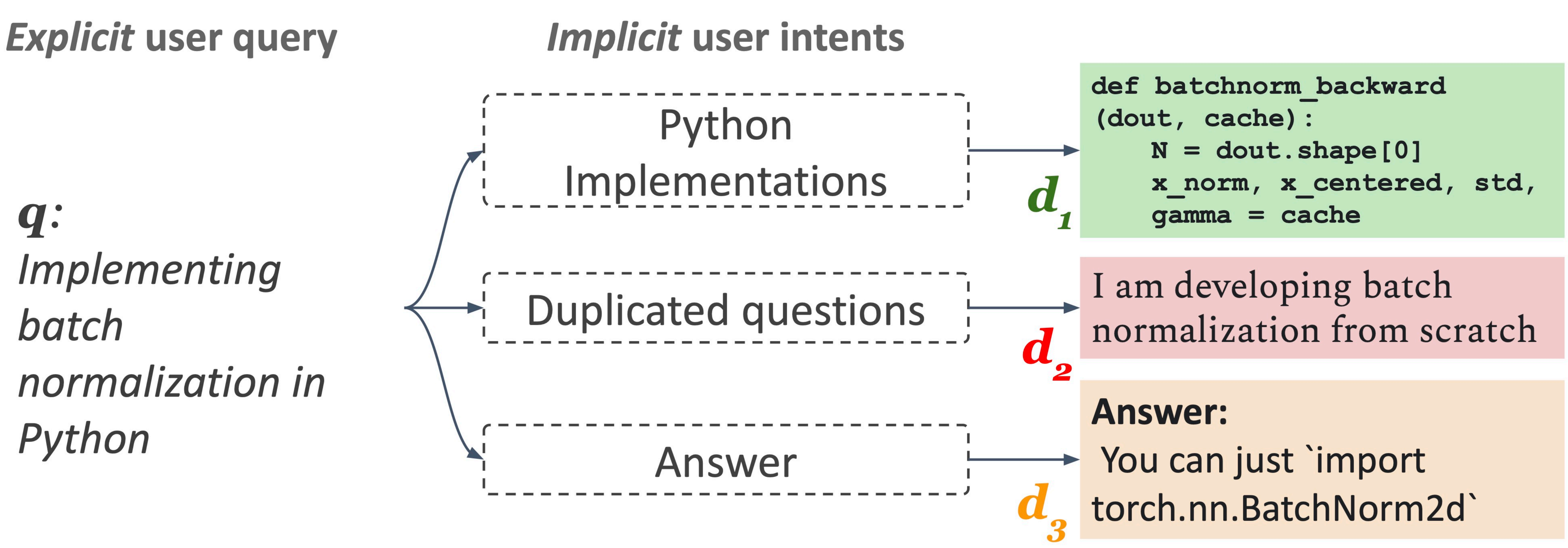


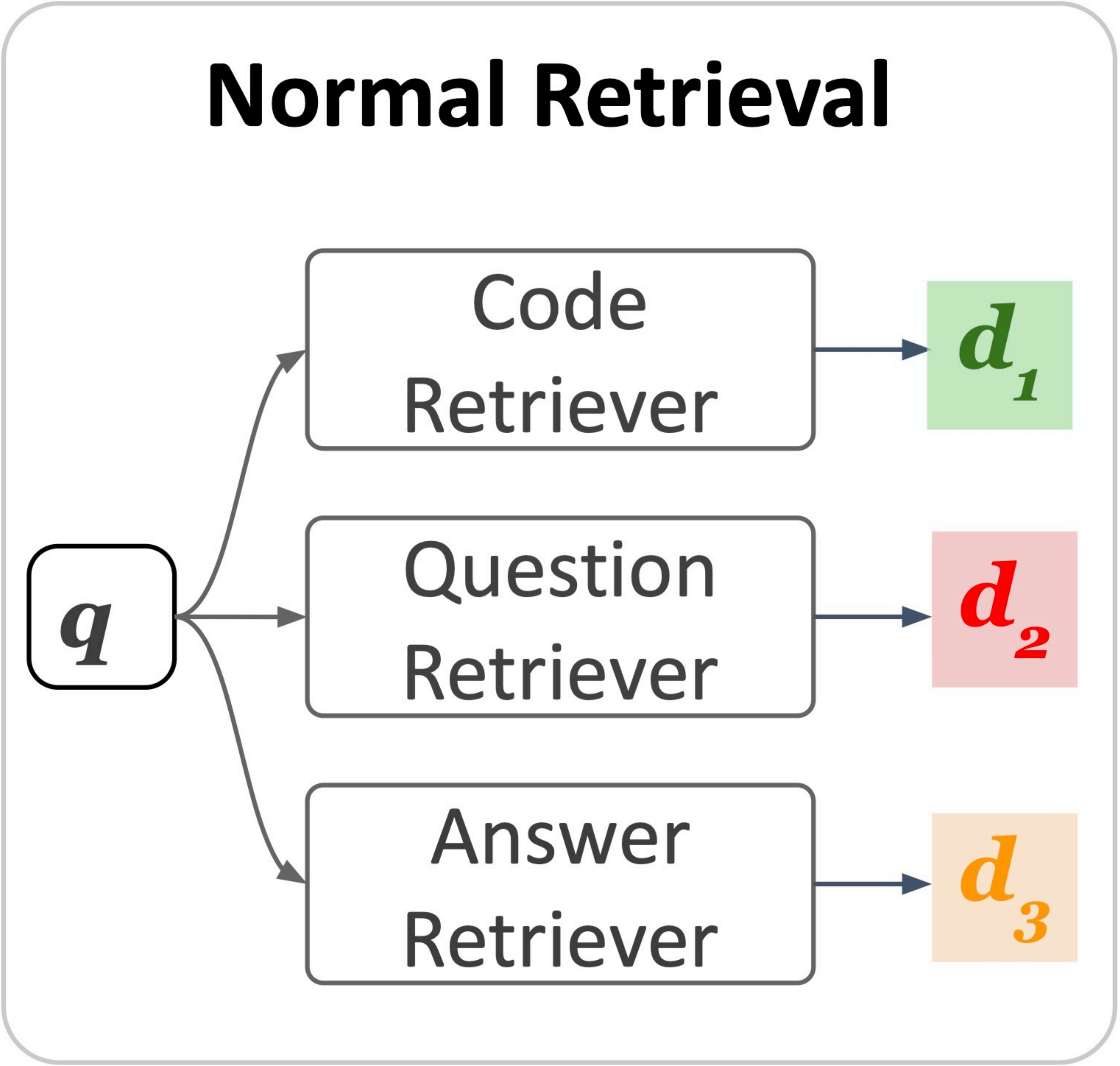
ACL Findings 2023; https://arxiv.org/abs/2211.09260

Beyond Similarity — Users' Diverse Intents

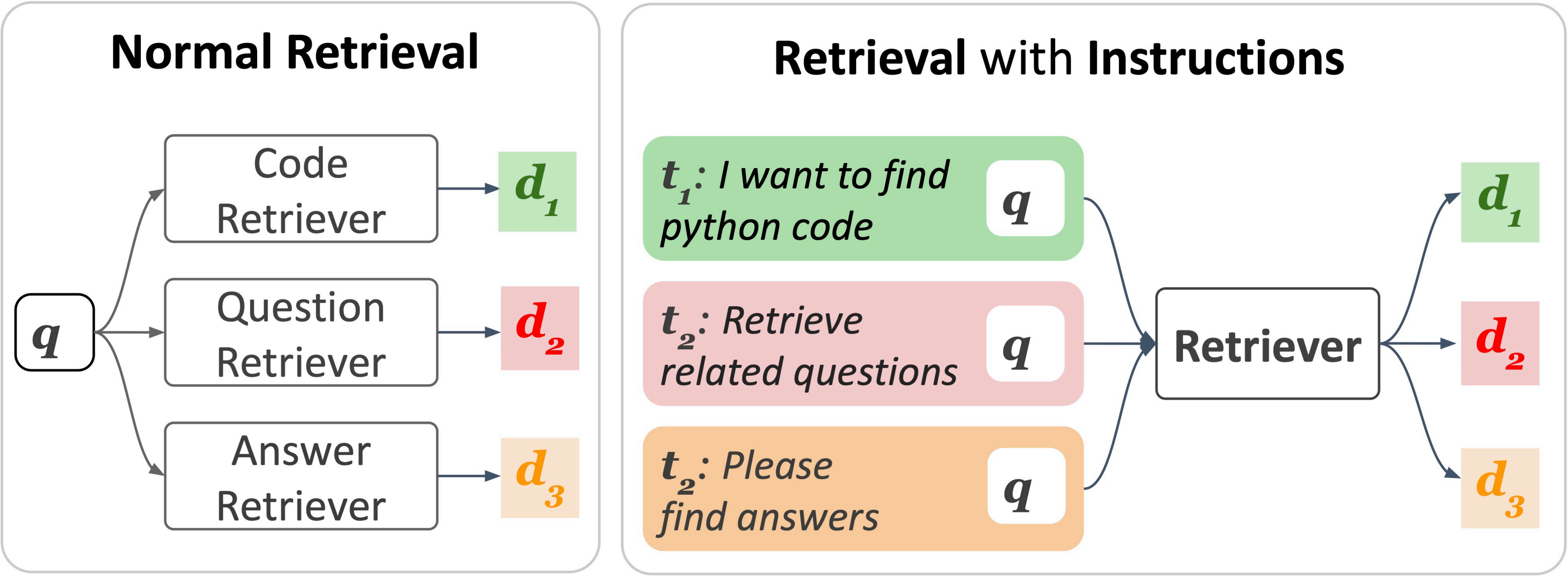


Conventional Approach: Separate Task-specific Retrievers

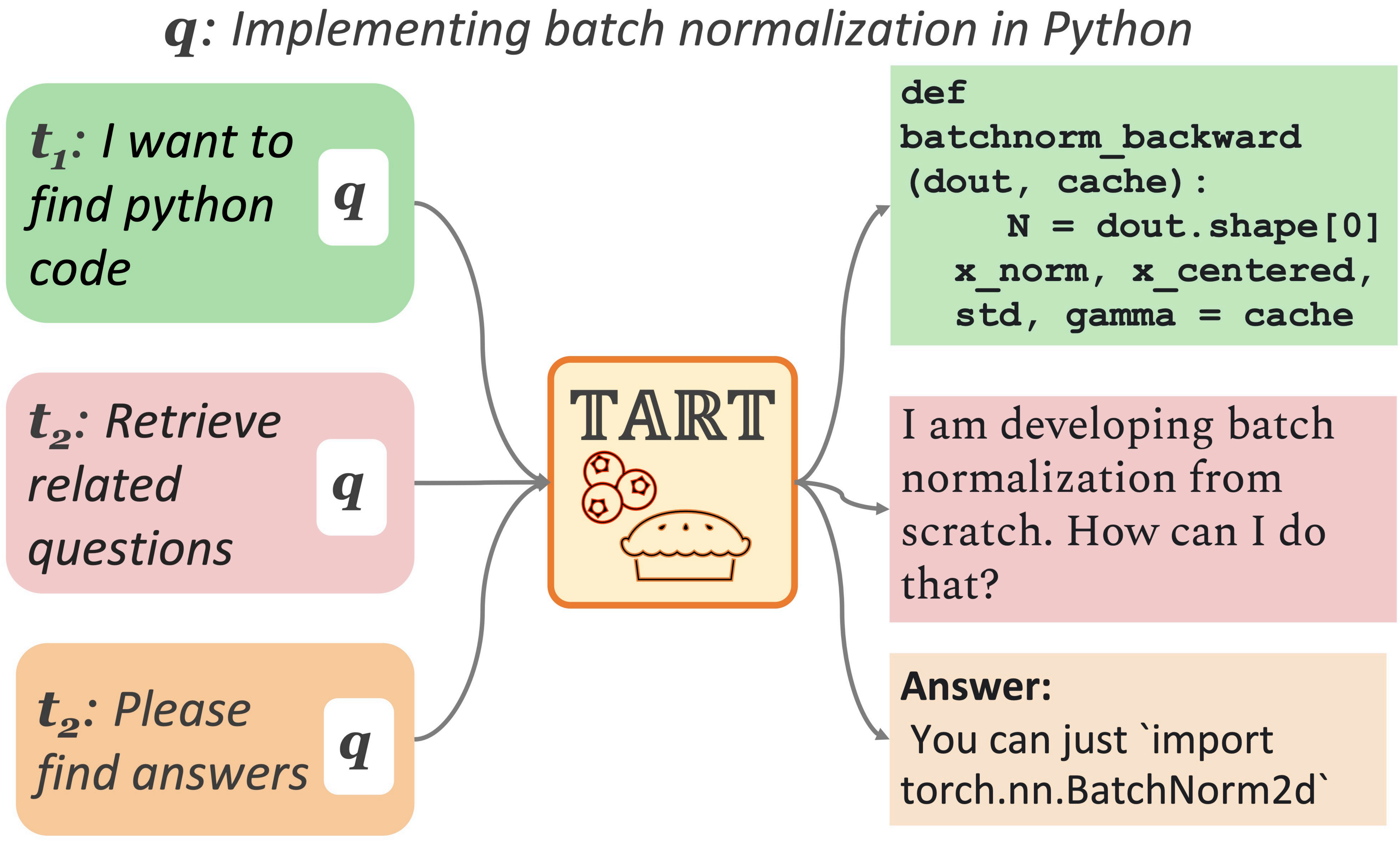




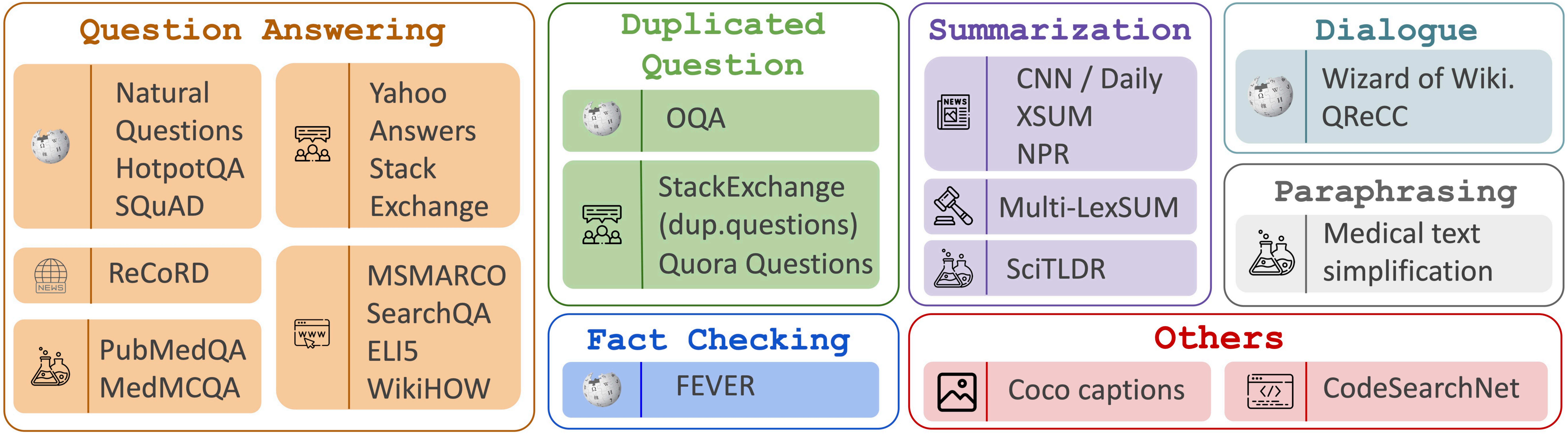
New Formulation: Retrieval with Instruction



TART: Task-aware Retrieval with Instruction



BERRI: a Large-scale Retrieval Dataset with Instructions



Instruction-scheme for Retrieval Tasks

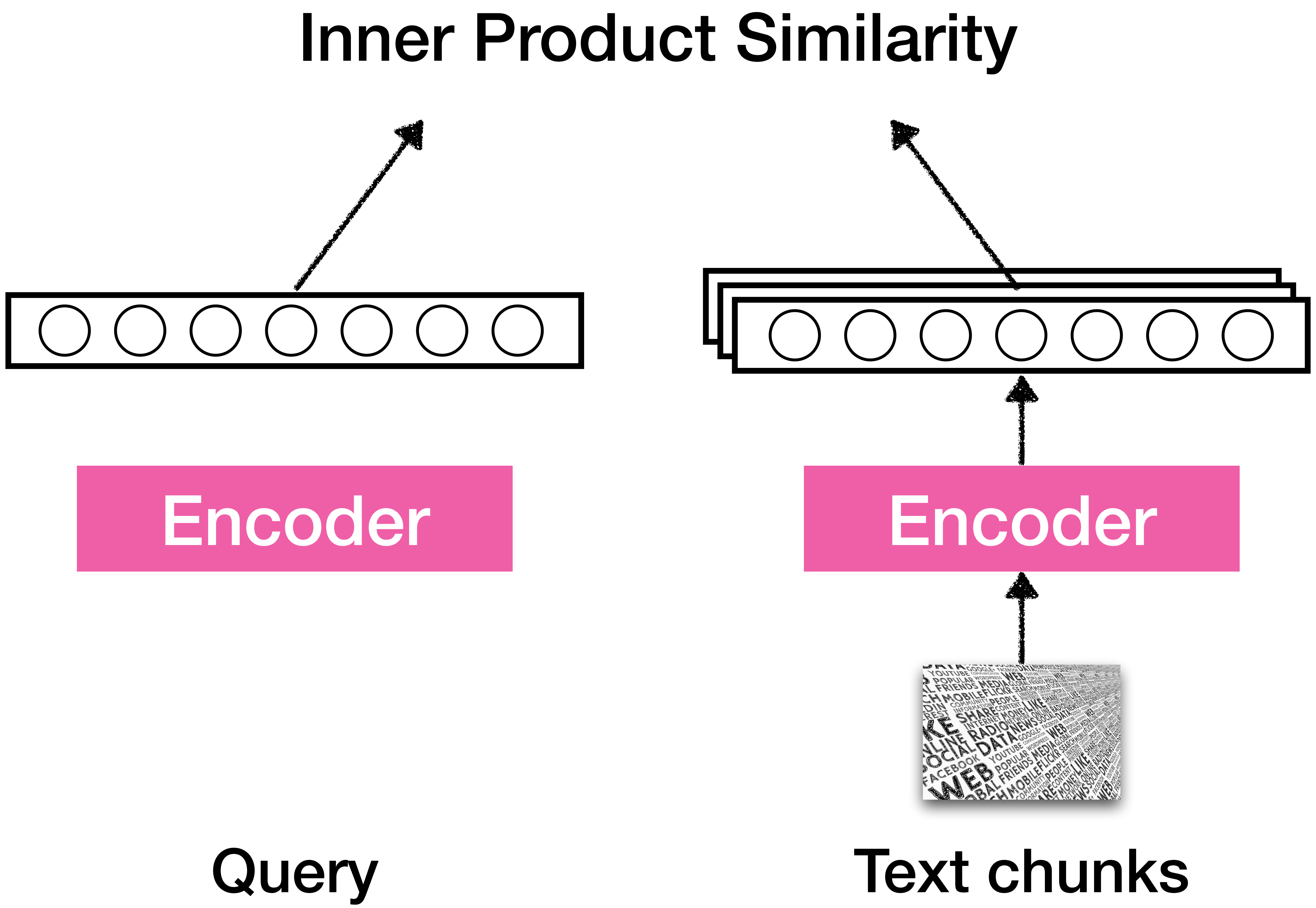
Dataset	Instruction
NQ	Retrieve a Wikipedia paragraph that answers this question.
QReCC	Find a dialogue response from dialogue history to answer the user's question.
Arguana	Retrieve a paragraph from an argument website that argues against the following argument.
SciFact	Find a sentence from a scientific paper to check if the statement is correct or not.
MultiLexSum	I want to find the one-sentence summary of this legal case.

Intent

Domain

Unit

Bi-Encoder Retriever Systems (e.g., DPR)

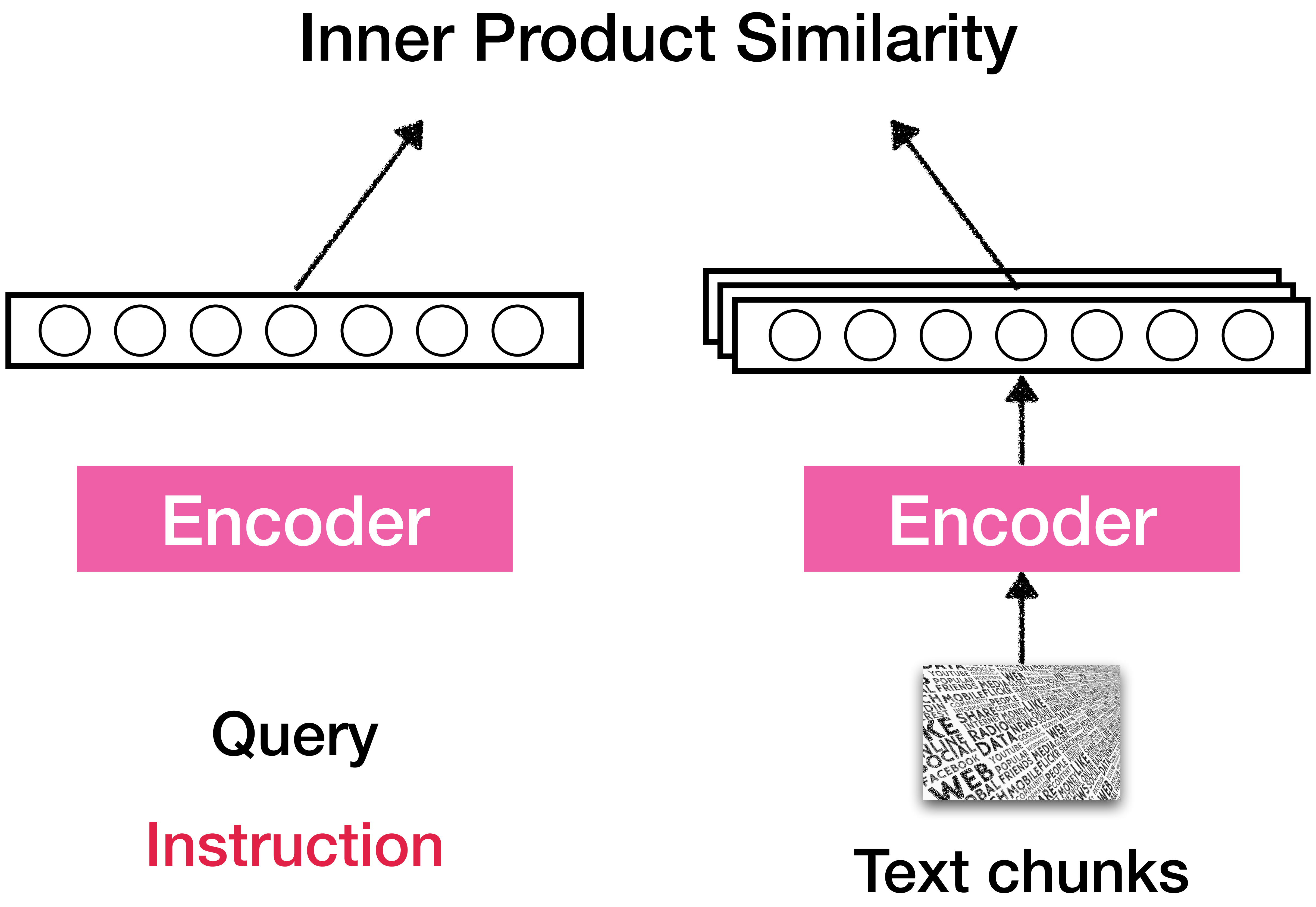


$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-) = -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Contrastive learning

Karpukhin et al.. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

Instruction-aware Bi-Encoder Retriever (TART- dual)



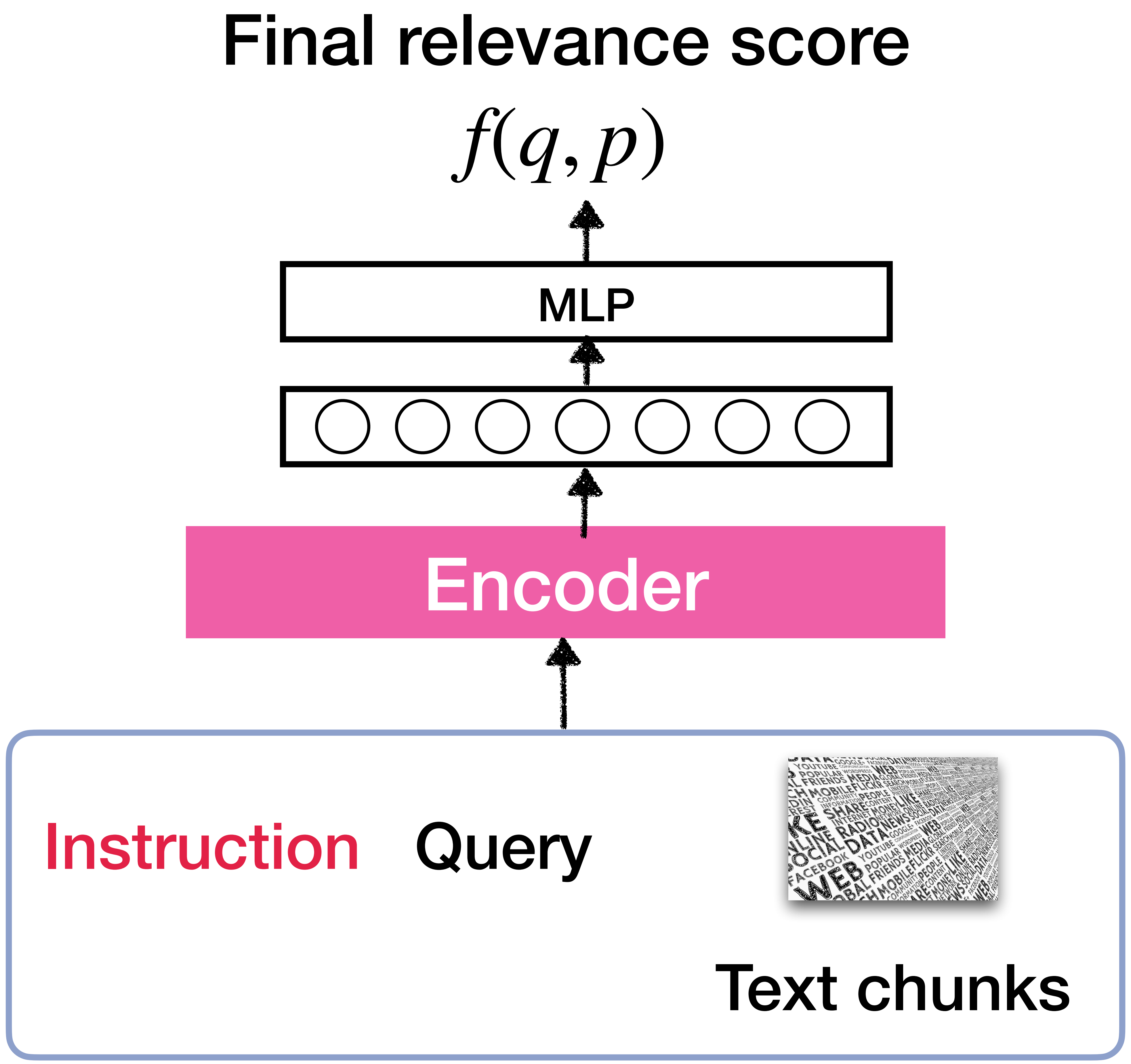
$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-) = -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Contrastive learning

(Slides adapted from our ACL tutorial)

Karpukhin et al.. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

Query-aware Cross-Encoder Retriever (TART- full)



$$L(q, P^+, P^-)$$

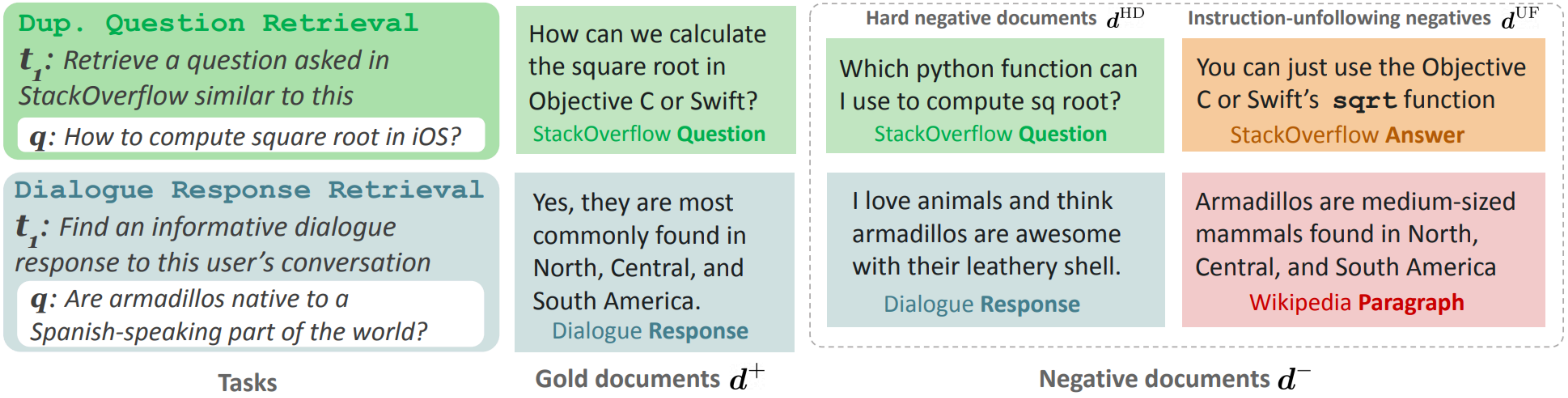
$$= -\sum_{p^+ \in P^+} \log(f(q, p^+)) - \sum_{p \in P^-} \log(1 - f(q, p^-))$$

Cross-entropy loss

(Slides adapted from our ACL tutorial)

Nogueira and Cho. Passage Re-ranking with BERT. ArXiv 2019.

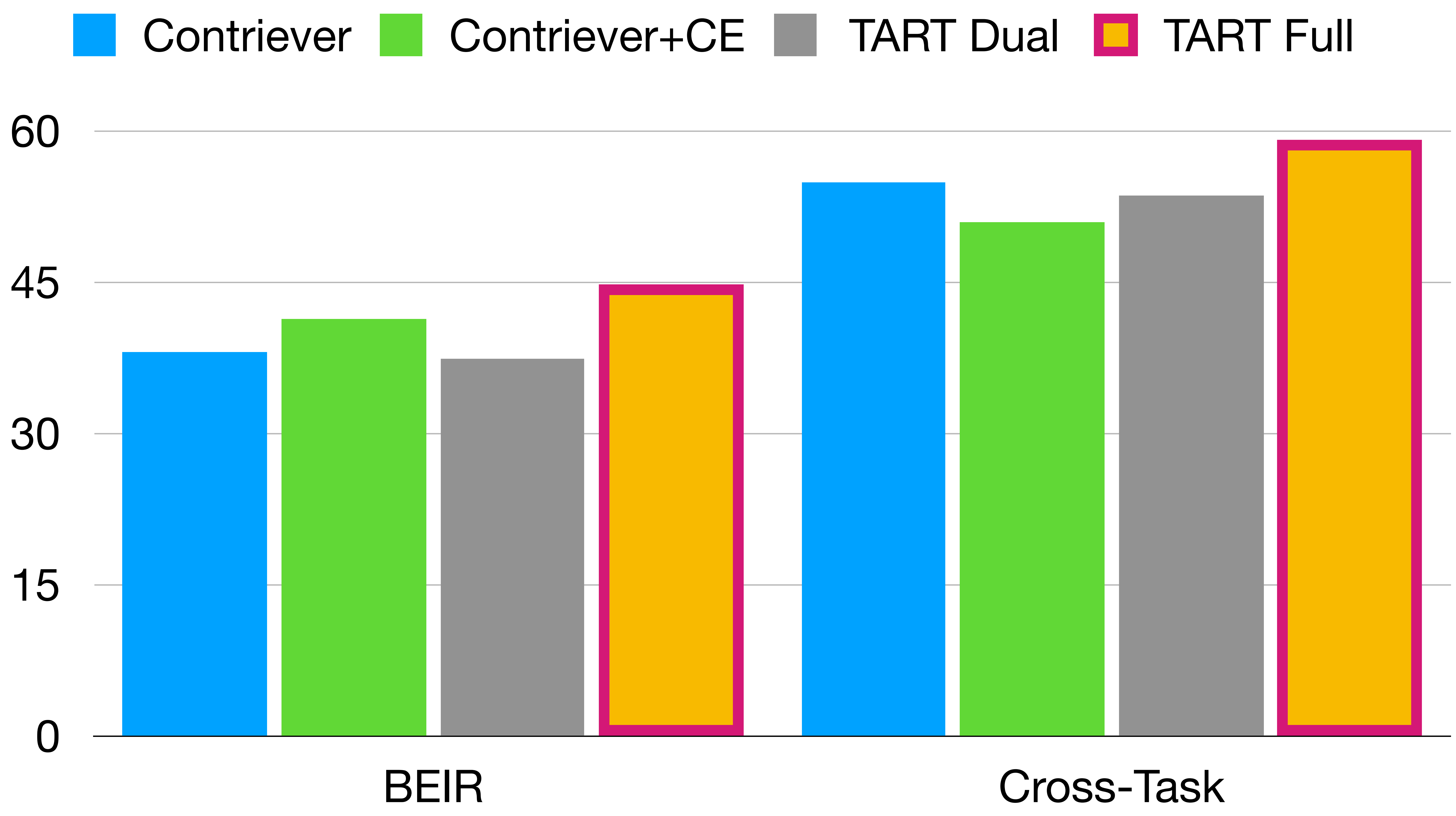
New Negative Samples: Instruction *un*-following Samples



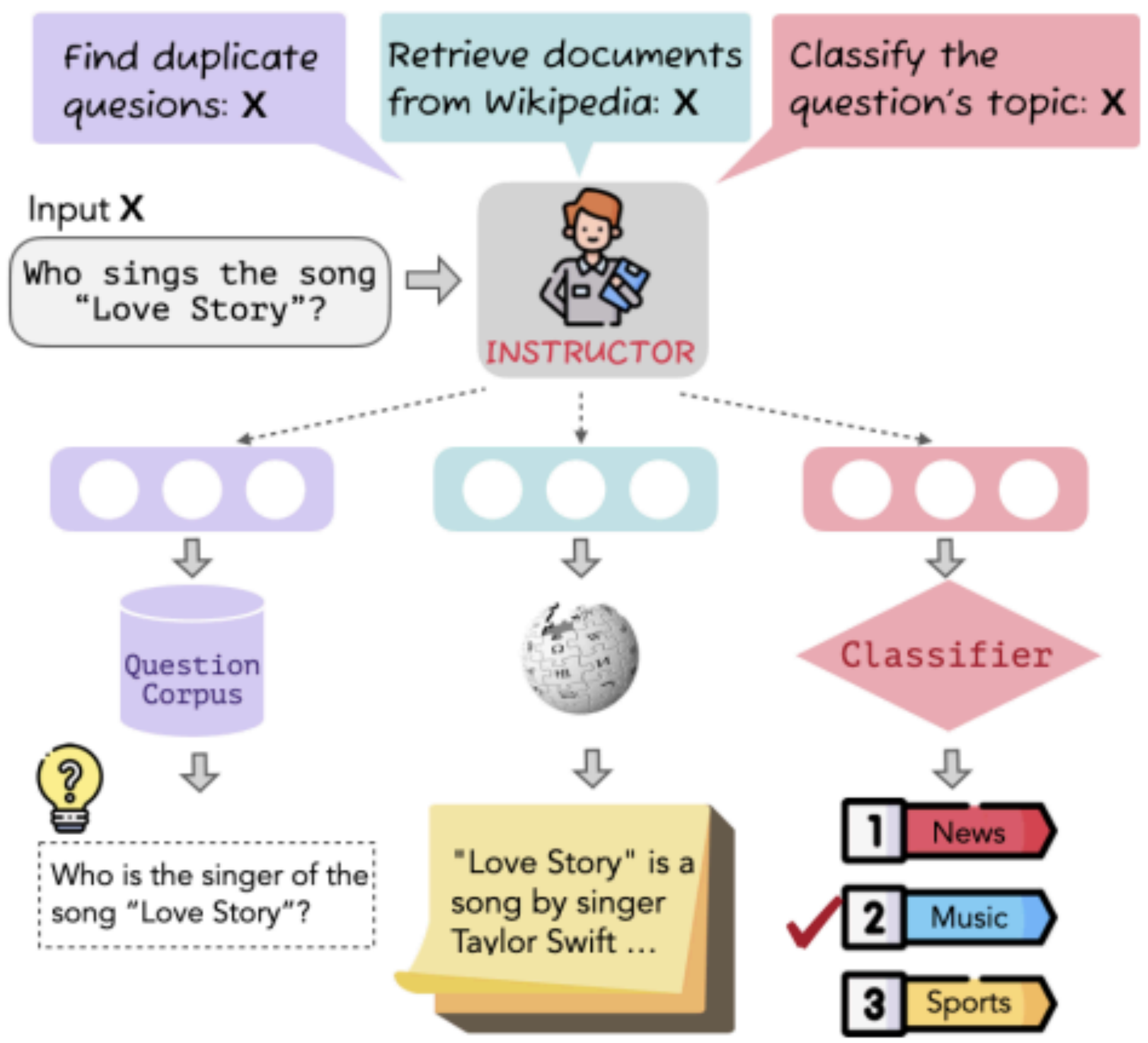
	Hard negative documents	Instruction-unfollowing negatives
Follow instruction?	✓	✗
Relevant to the query?	✗	✓

Better Generalization and Instruction Following

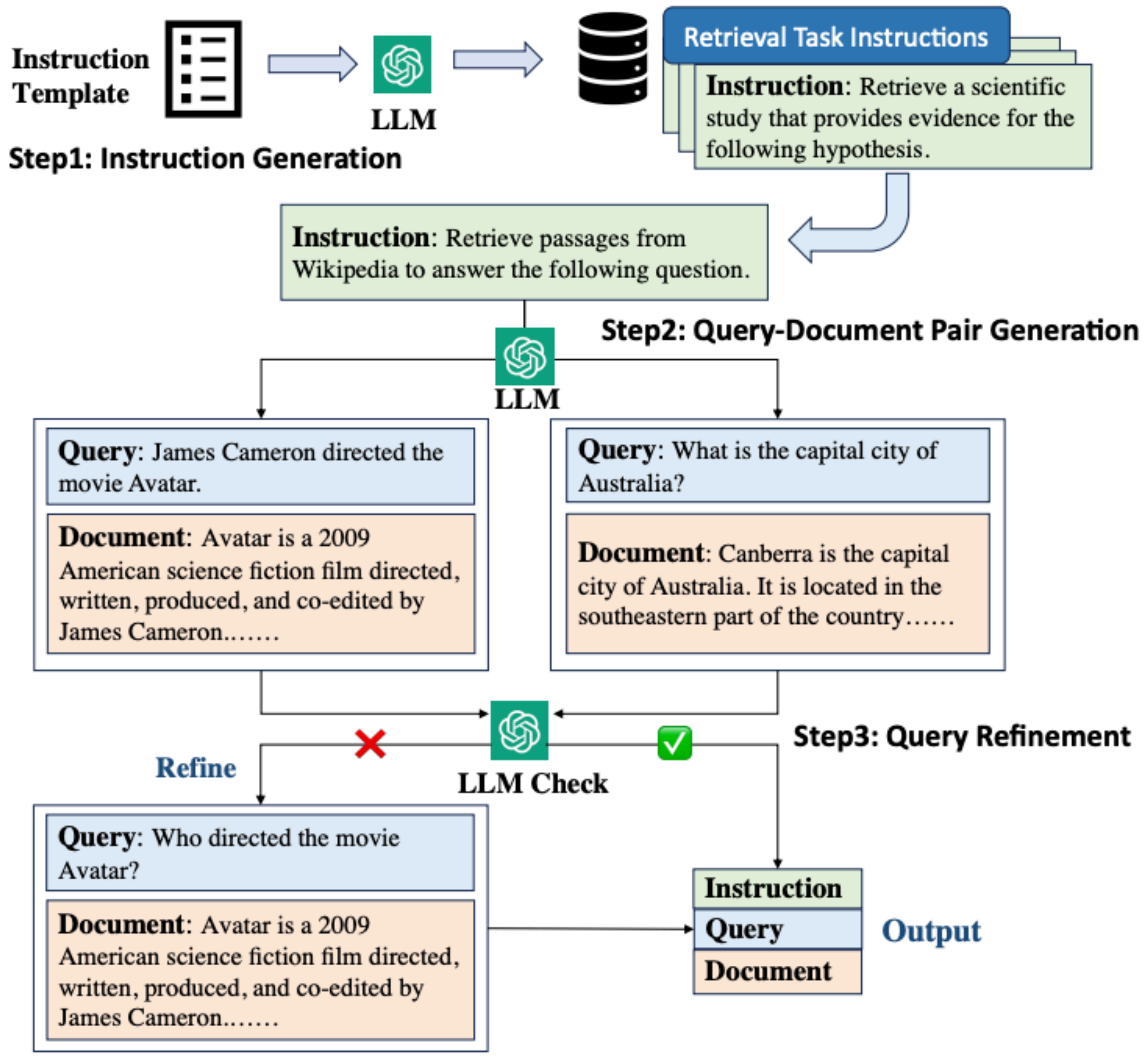
Instruction-tuning for retrieval shows effectiveness in zero-shot & cross-task



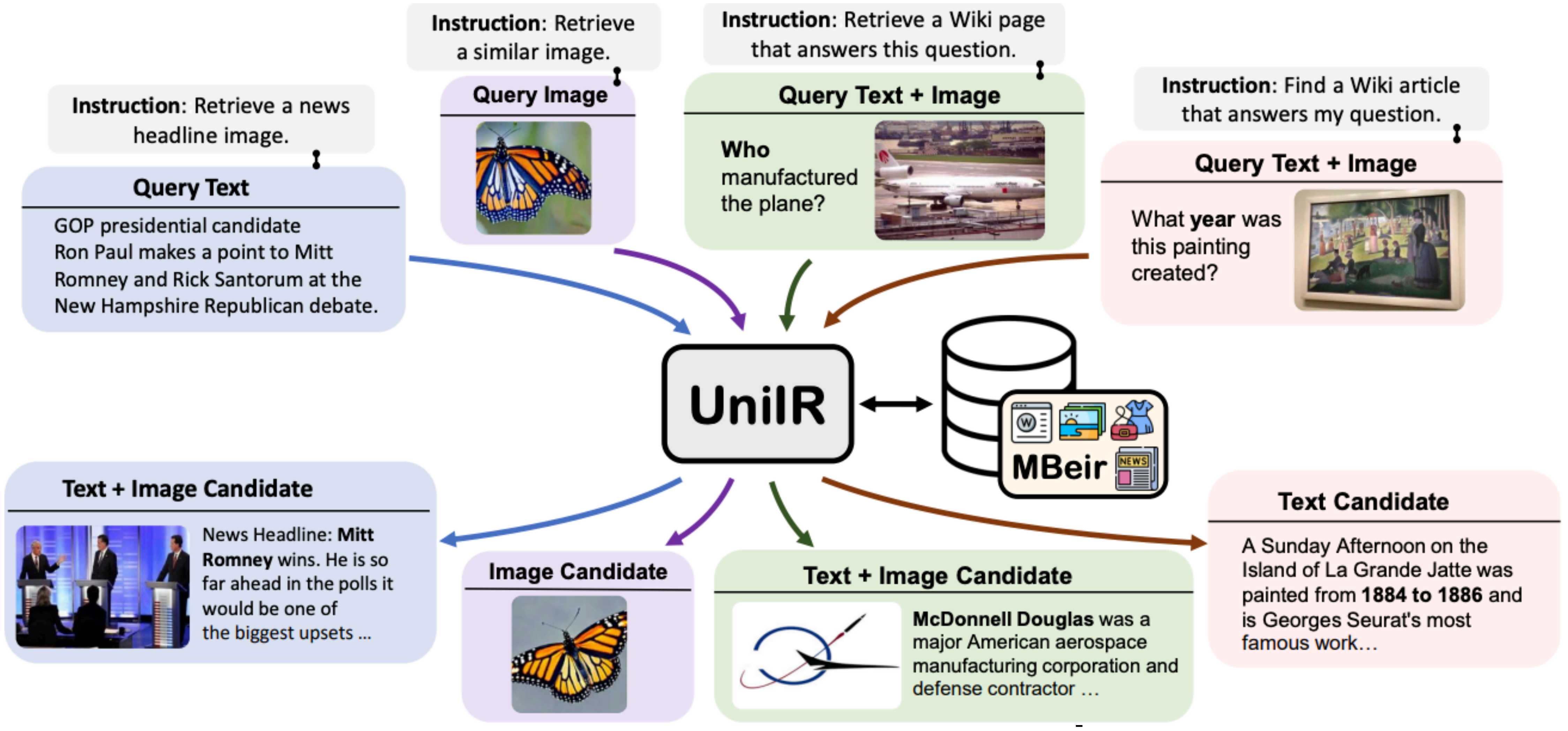
Increasing number of instruction-following retrievers



Instructor (Su et al., 2023)

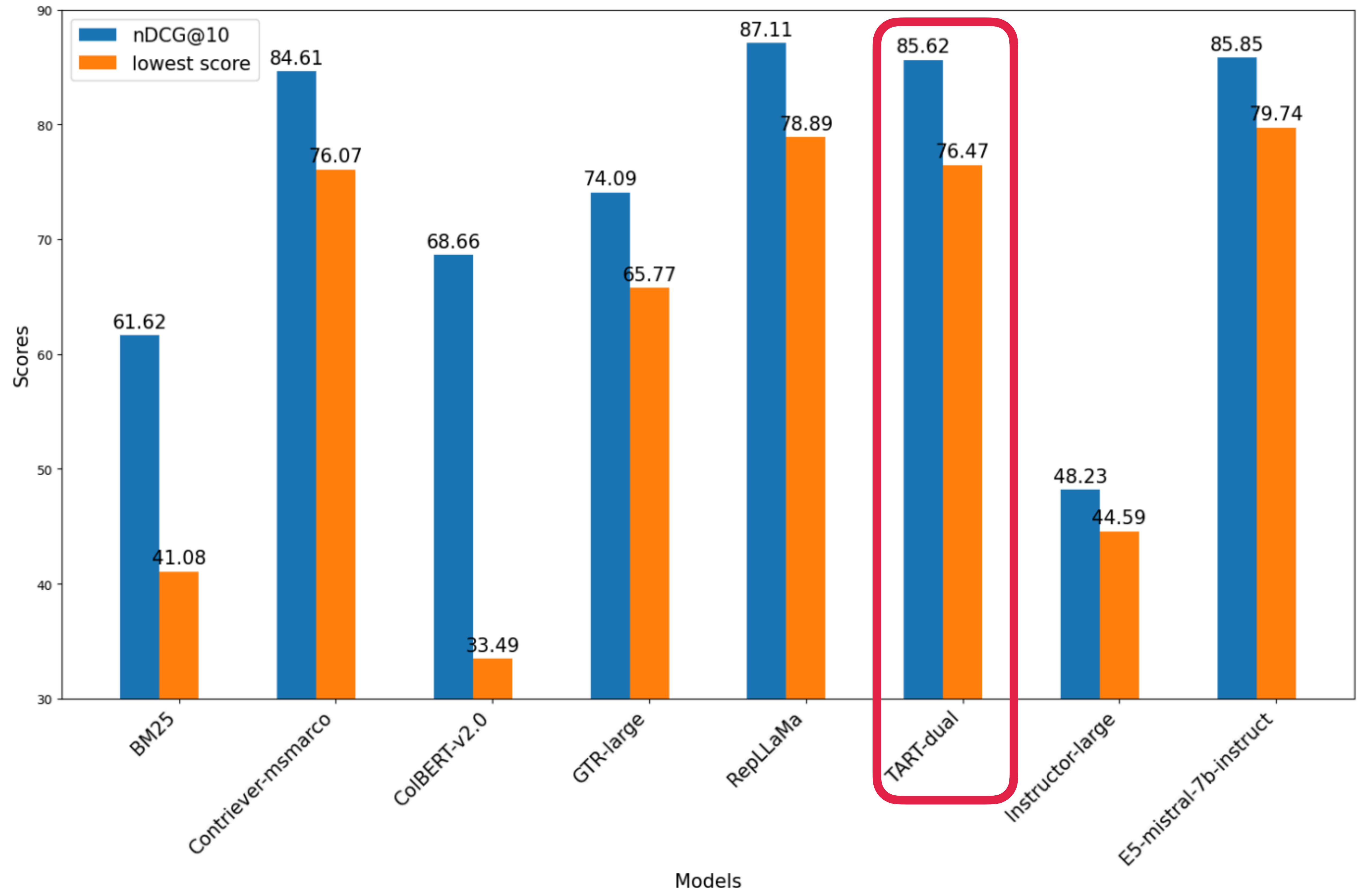


ControlRetriever (Pang et al., 2023)
E5 MISTRAL Instruct
(Wang et al., 2024)



UniLR (Wei et al., 2023)

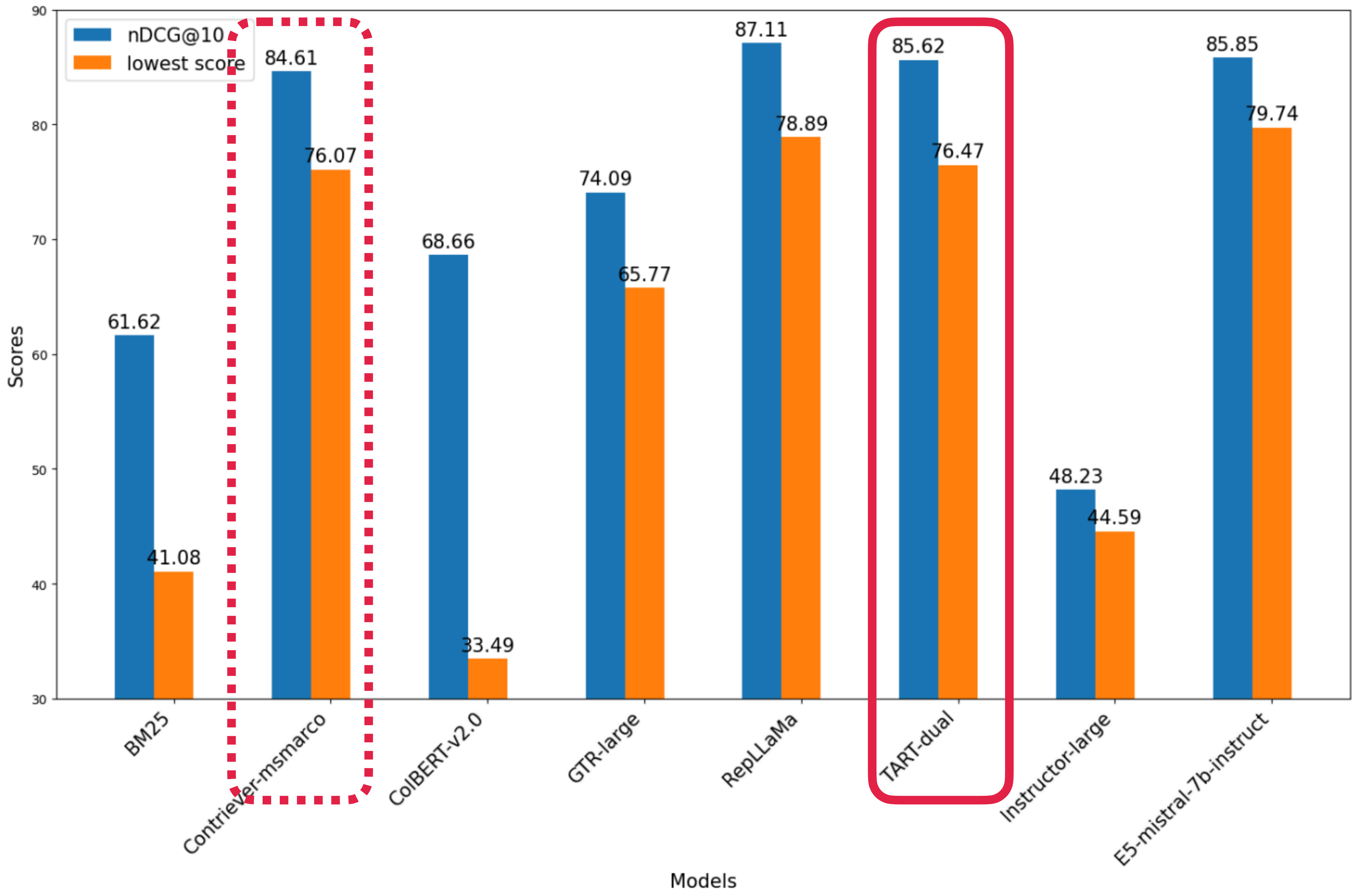
Are those retrievers following instructions?



TART (110M) outperforms other <7B models by large margins

Instructir (Oh et al., 2024)

Are those retrievers following instructions?



TART (110M) outperforms other <7B models by large margins

Instructir (Oh et al., 2024)

Are those retrievers following instructions?

	Model	Robust04 mAP	Robust04 p-MRR	News21 nDCG	News21 p-MRR	Core17 mAP	Core17 p-MRR	Average Score	Average p-MRR
No-Instruction IR	BM25	12.2	-2.8	21.3	+2.5	8.1	-0.7	13.9	-0.3
	E5-base-v2	14.5	-6.8	21.6	-4.1	14.0	-2.9	16.7	-4.6
	E5-large-v2	18.1	-4.1	24.9	-2.2	17.0	+0.1	20.0	-2.1
	Contriever	20.3	-6.1	24.0	-1.8	15.3	-2.5	19.9	-3.5
	MonoBERT	21.5	-9.7	26.3	-4.4	18.4	-1.3	22.1	-5.1
	MonoT5-base	16.3	-5.8	11.9	-1.2	12.2	-3.5	13.5	-3.5
	MonoT5-3B	27.8	+5.6	18.6	+7.5	18.1	+1.7	21.5	+4.9
Instruction-IR	BGE-base	17.5	-6.4	23.8	-0.2	14.6	-2.7	18.6	-3.1
	BGE-large	18.1	-7.8	26.4	+0.1	15.0	+0.1	19.8	-2.5
	TART-Contriever	14.1	-7.8	21.9	+0.0	12.4	-1.3	16.1	-3.0
	INSTRUCTOR-base	14.4	-5.6	16.3	-2.5	14.7	-2.2	15.1	-3.4
	INSTRUCTOR-xl	15.5	-2.1	14.6	-4.3	14.4	-0.6	14.8	-2.3
	TART-FLAN-T5-xl	25.2	-0.8	20.3	-1.1	17.0	+2.8	20.8	+0.3
	GritLM-7B	29.0	-1.4	25.2	+2.1	20.8	+2.6	25.0	+1.1
APIs	Cohere v3 English	22.9	-3.3	23.6	-3.1	20.6	+2.7	22.4	-1.2
	OpenAI v3 Large	27.9	-5.7	30.0	-3.3	21.4	-0.2	26.4	-3.1
Instruct LMs	FLAN-T5-base	6.8	+5.0	2.2	+1.1	6.5	-3.2	5.2	+1.0
	FLAN-T5-large	15.1	+4.0	8.5	+7.7	11.5	+1.2	11.7	+4.3
	Llama-2-7B-chat	6.9	+1.6	13.3	+2.1	5.4	+3.6	8.5	+2.4
	Mistral-7B-instruct	24.1	+12.2	22.9	+10.5	19.6	+13.4	22.2	+12.0
	FollowIR-7B	25.9	**+13.6**	25.7	**+10.8**	20.0	**+16.3**	23.9	**+13.6**

Table 3: Evaluating instruction-following on FOLLOWIR. p-MRR is a new pairwise evaluation metric measuring instruction following when instructions change, ranging from -100 to 100 (higher is better). We see that the only models that show any success at following instructions are large models (3B+ parameters) or instruction-tuned LLMs that haven't been trained on retrieval tasks.

Smaller models often brittle towards diverse instructions

FollowIR (Weller et al., 2024)

Today's lecture

Promises and Limitations of Retrieval-augmented LMs

Advanced Retriever: Intent-aware LM-based retrievers

Advanced RAG: Self-reflective LMs with dynamic Retrievals

Summary and Future directions: RAG in the wild

Summary

- **Understanding** Retrieval-augmented LMs (Asai et al., 2024b; Mallen*, Asai et al., 2023)
  - Retrieval-augmented LMs can alleviate many issues in parametric LMs.
  - More fundamental improvements for architectures or training is necessary

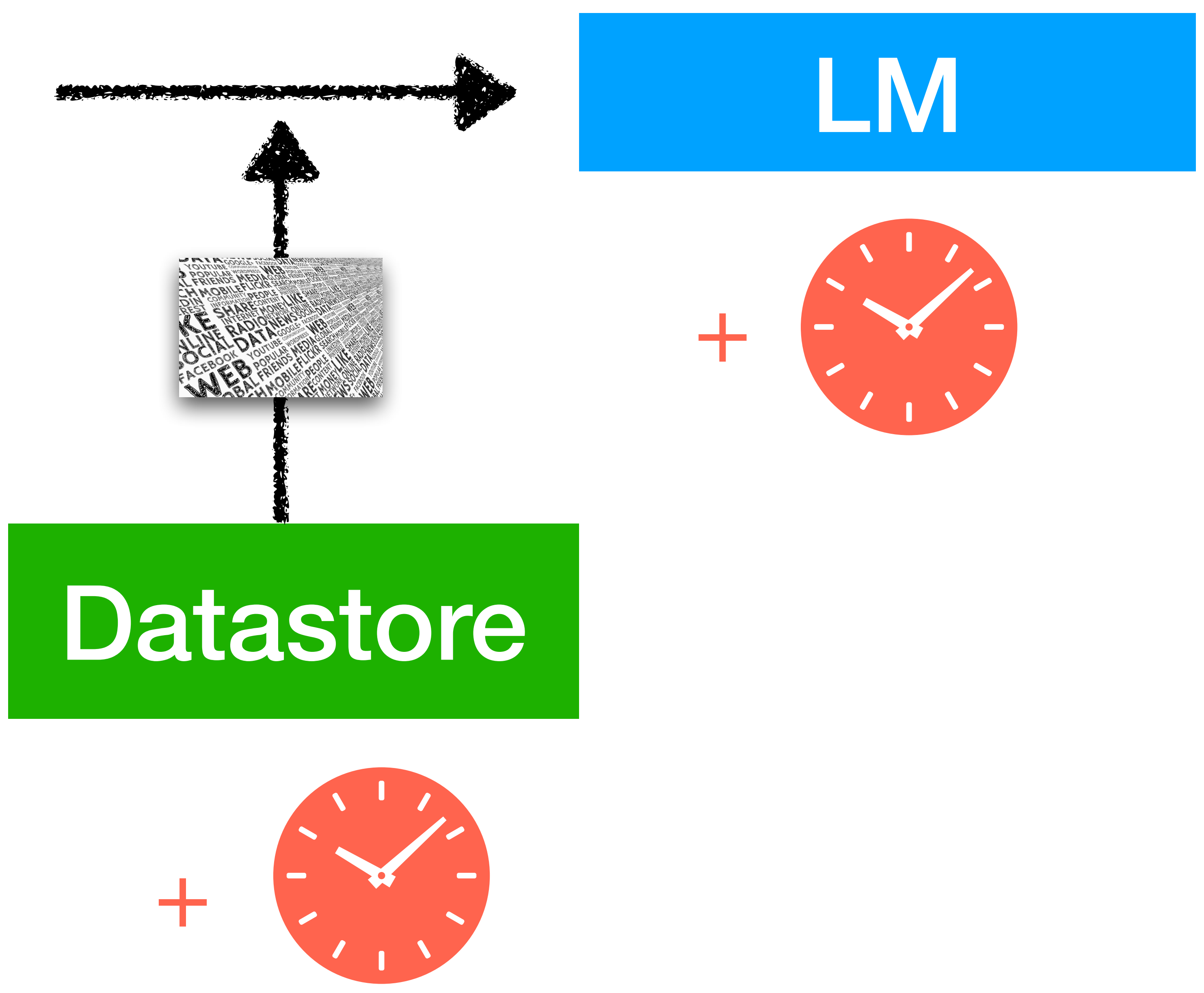
Summary

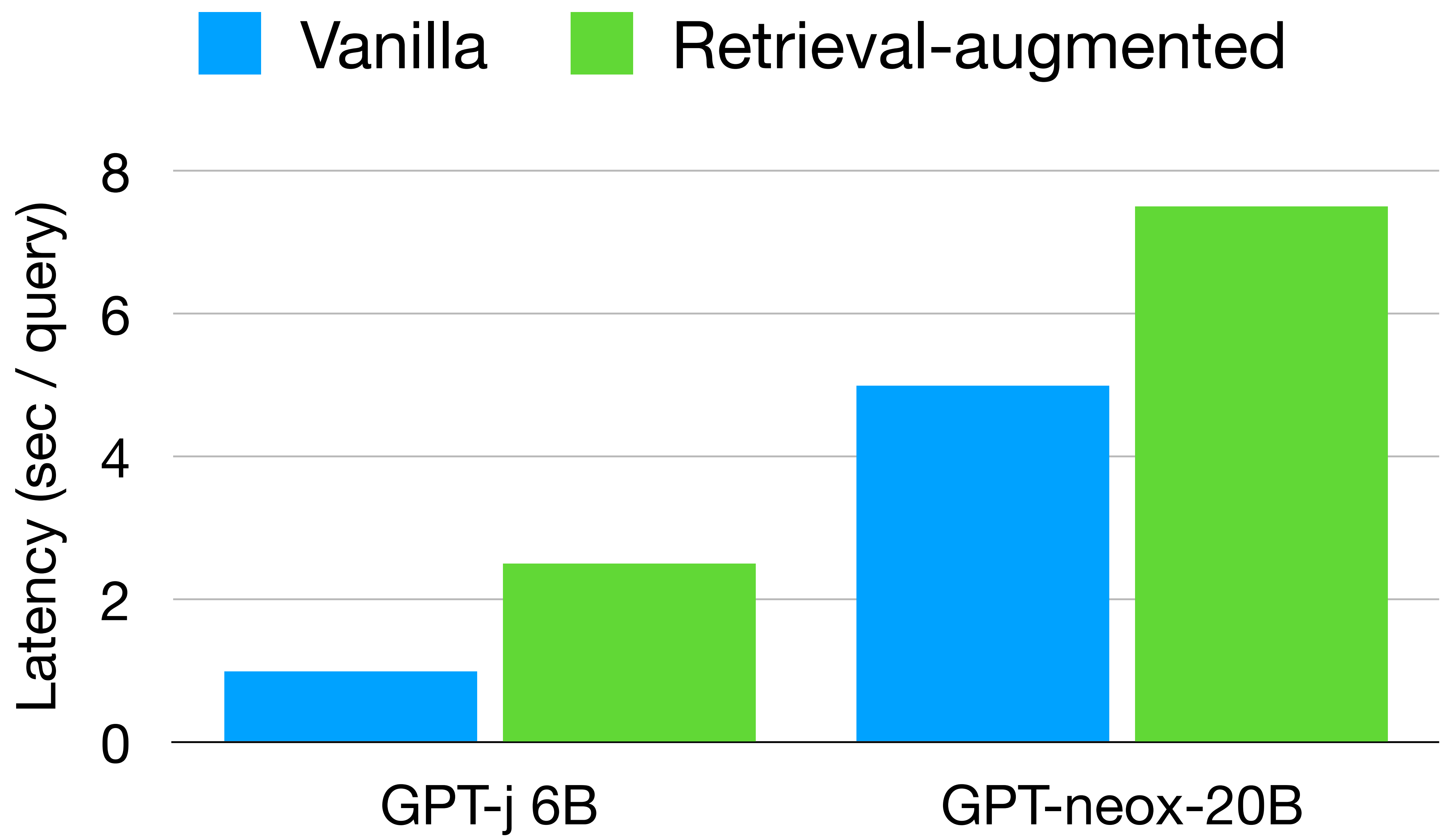
- **Understanding** Retrieval-augmented LMs (Asai et al., 2024b; Mallen*, Asai et al., 2023)
  - Retrieval-augmented LMs can alleviate many issues in parametric LMs.
  - More fundamental improvements for architectures or training is necessary

- **Advancing** Retrieval-augmented LMs (Asai et al., 2024; Asai et al., 2023)
  - **Self-RAG** to build versatile retrieval-augmented LMs addressing issues in RAG
  - **Task-aware retrievals** to build versatile RAG systems

Reliable RAG in the wild: improving efficiency

Efficiency | Scaling datastores | Beyond general QA

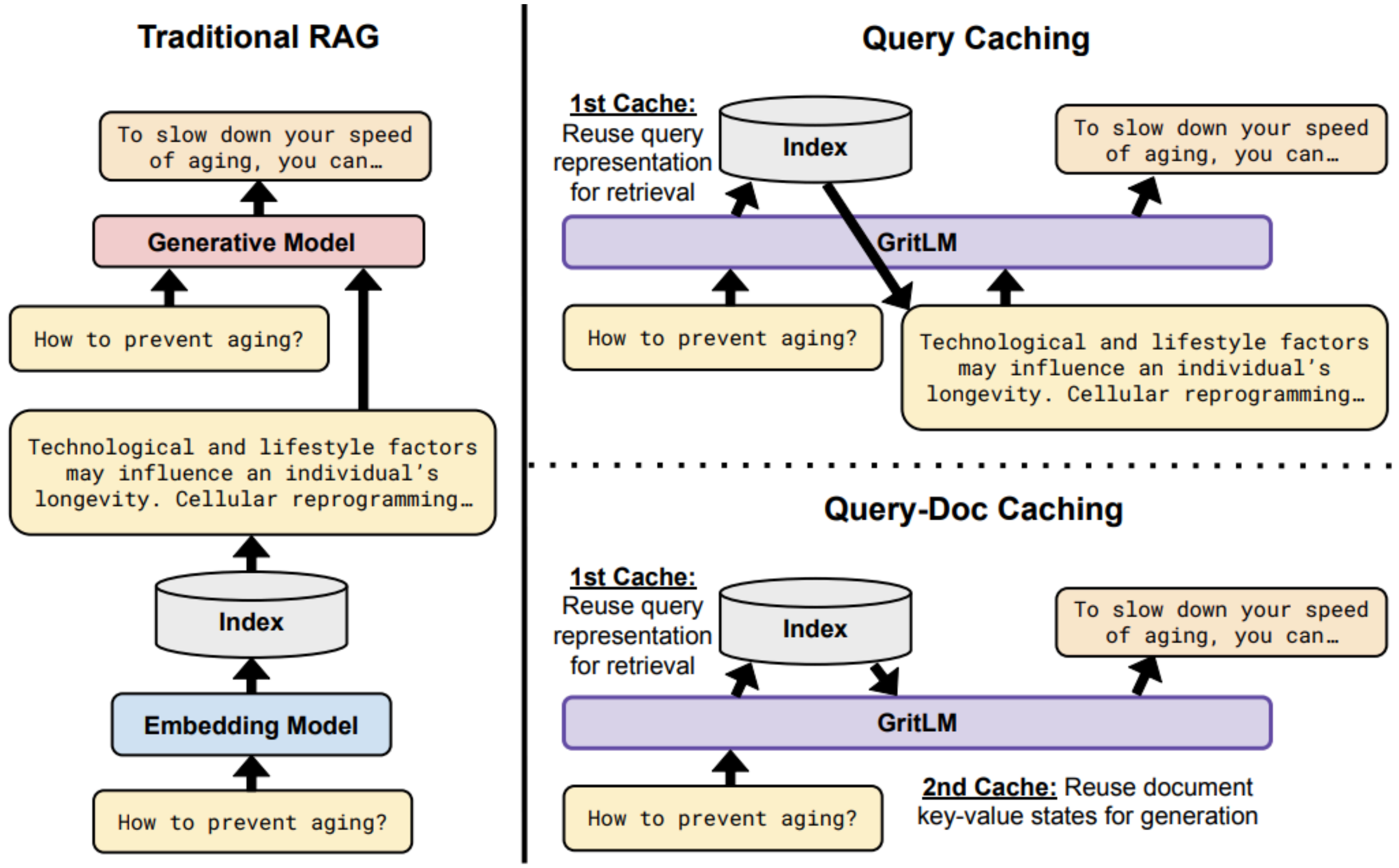




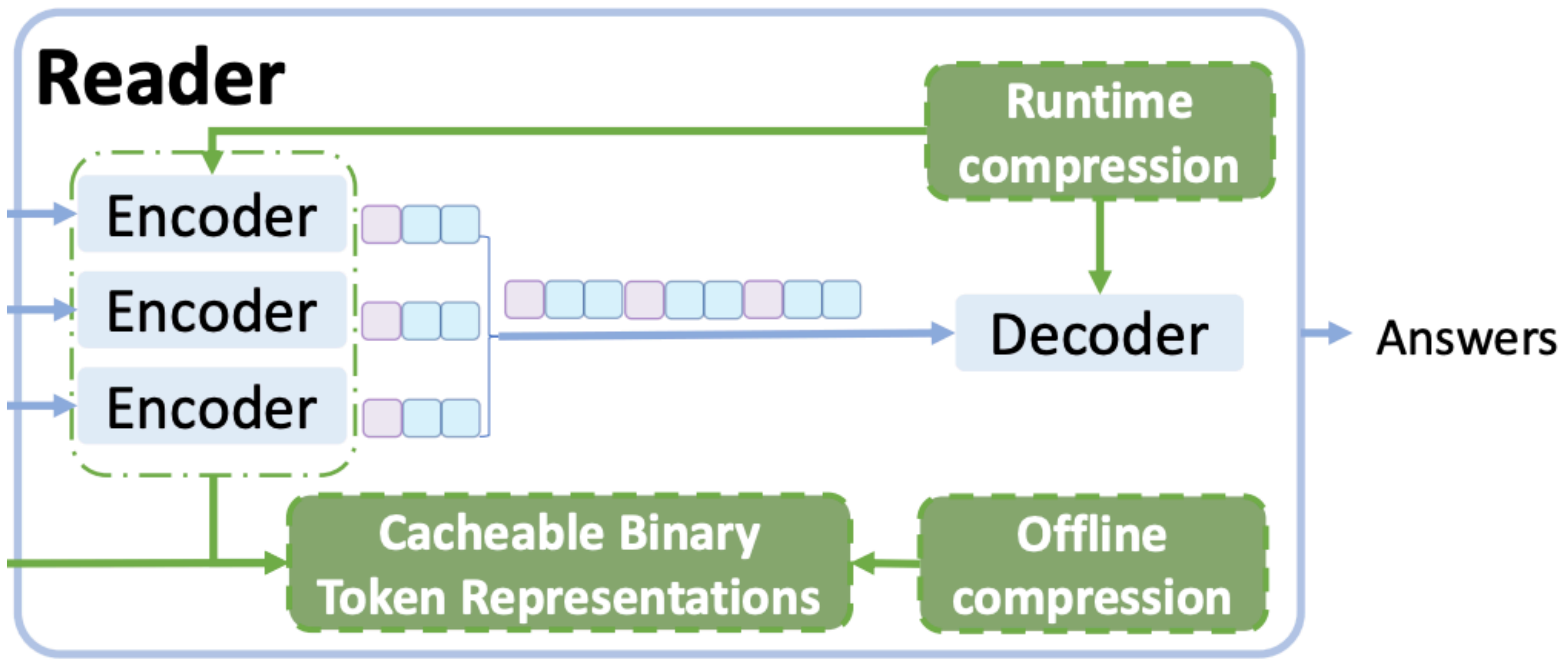
Mallen*, Asai* et al., When Not to Trust Language Models:
Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023.

Reliable RAG in the wild: efficient algorithms / models for RAG

Efficiency | Scaling datastores | Beyond general QA



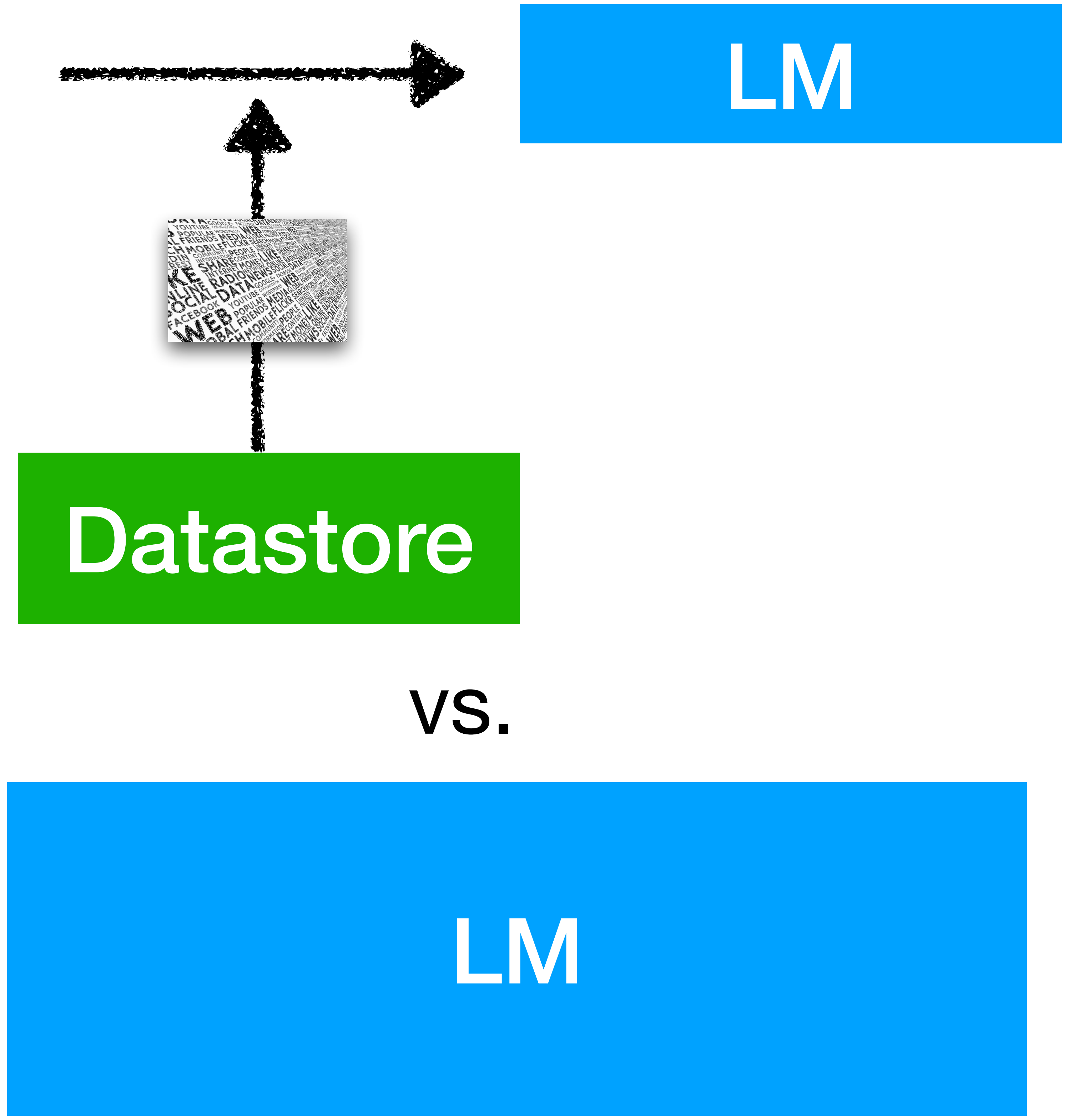
Muennighoff et al. Generative Representational Instruction Tuning. 2024.

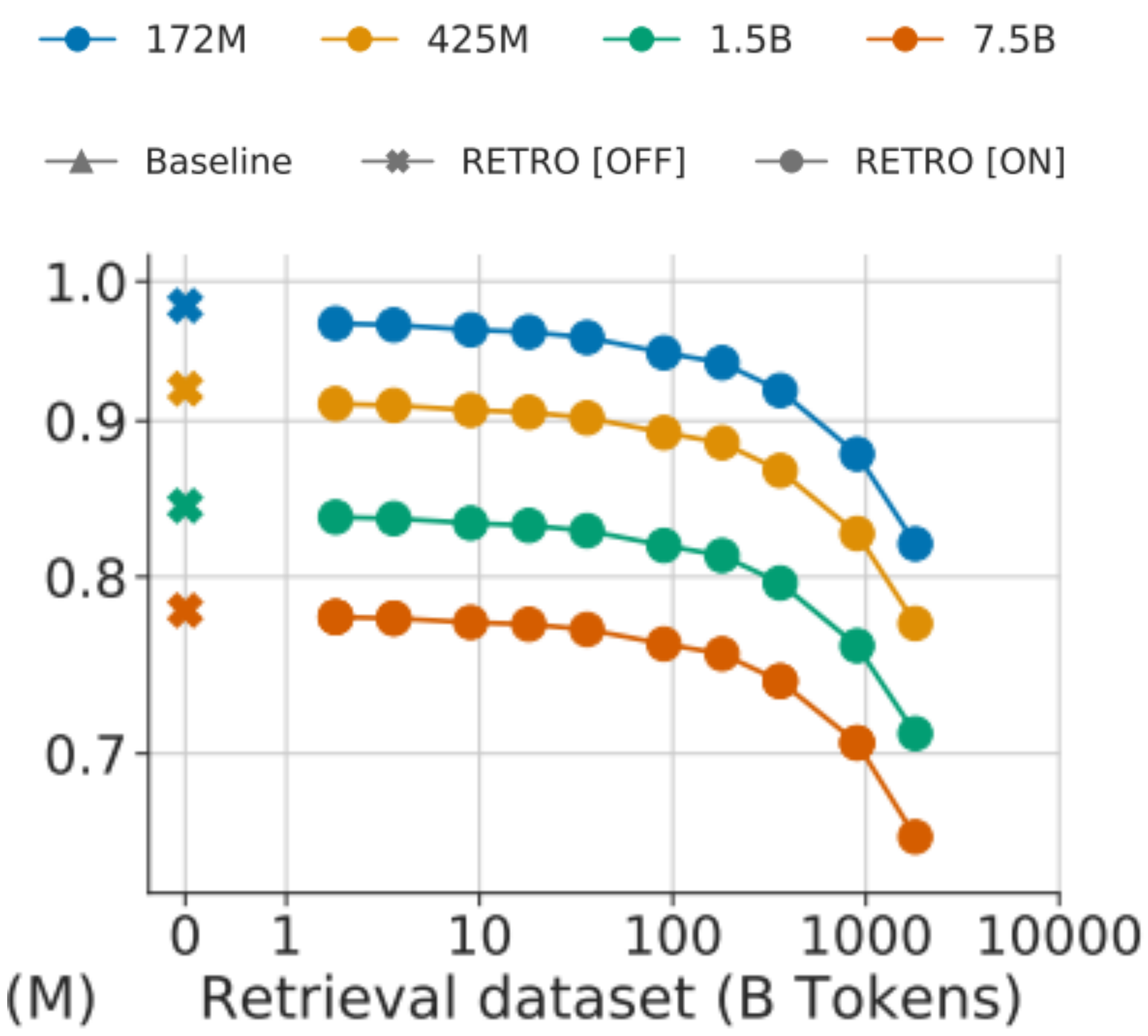


Cao et al. BTR: Binary Token Representations for Efficient Retrieval Augmented Language Models. ICLR 2024.

Reliable RAG in the wild: scaling datastore

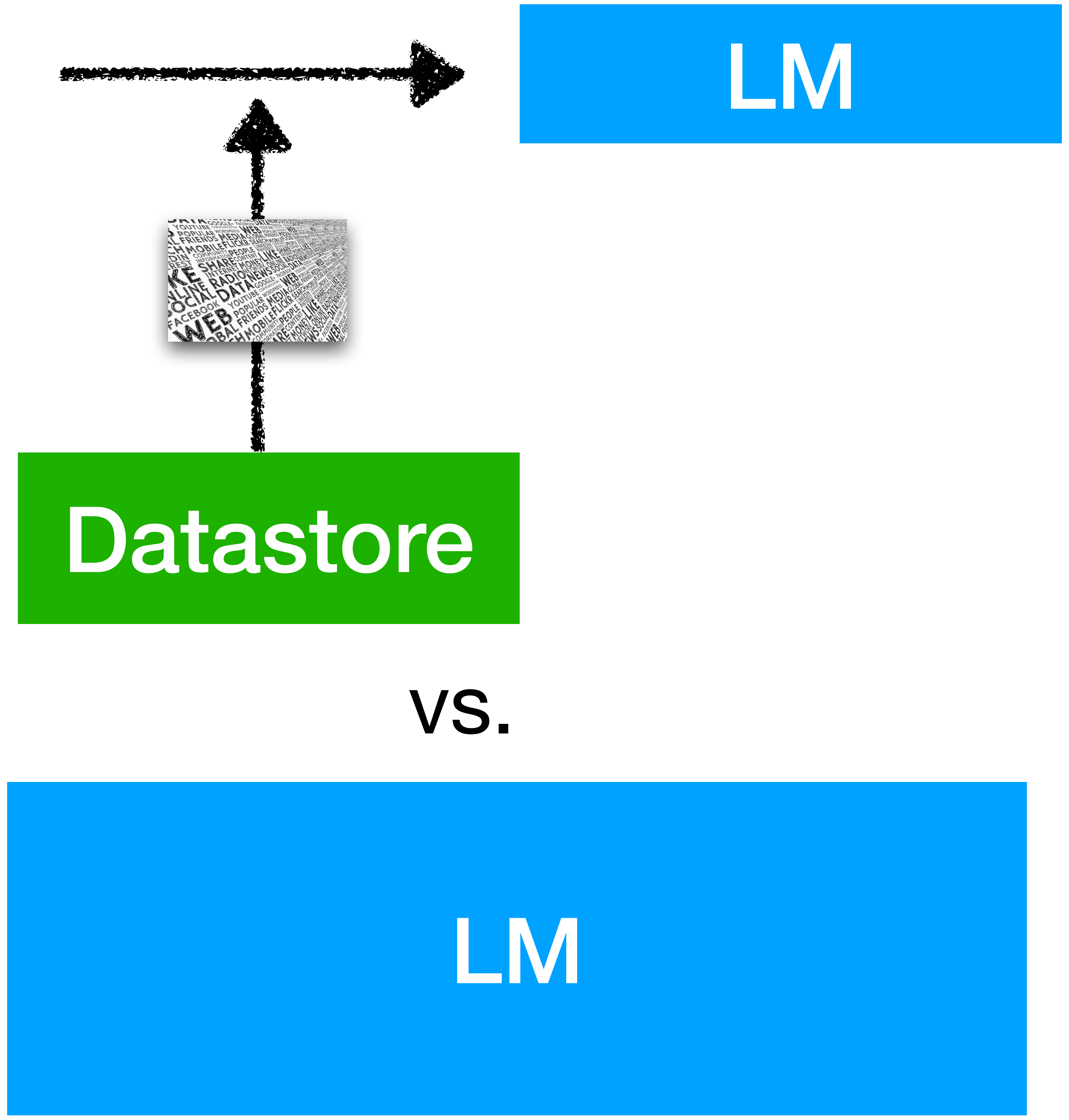
Efficiency | **Scaling datastores** | Beyond general QA





Reliable RAG in the wild: scaling datastore

Efficiency | **Scaling datastores** | Beyond general QA



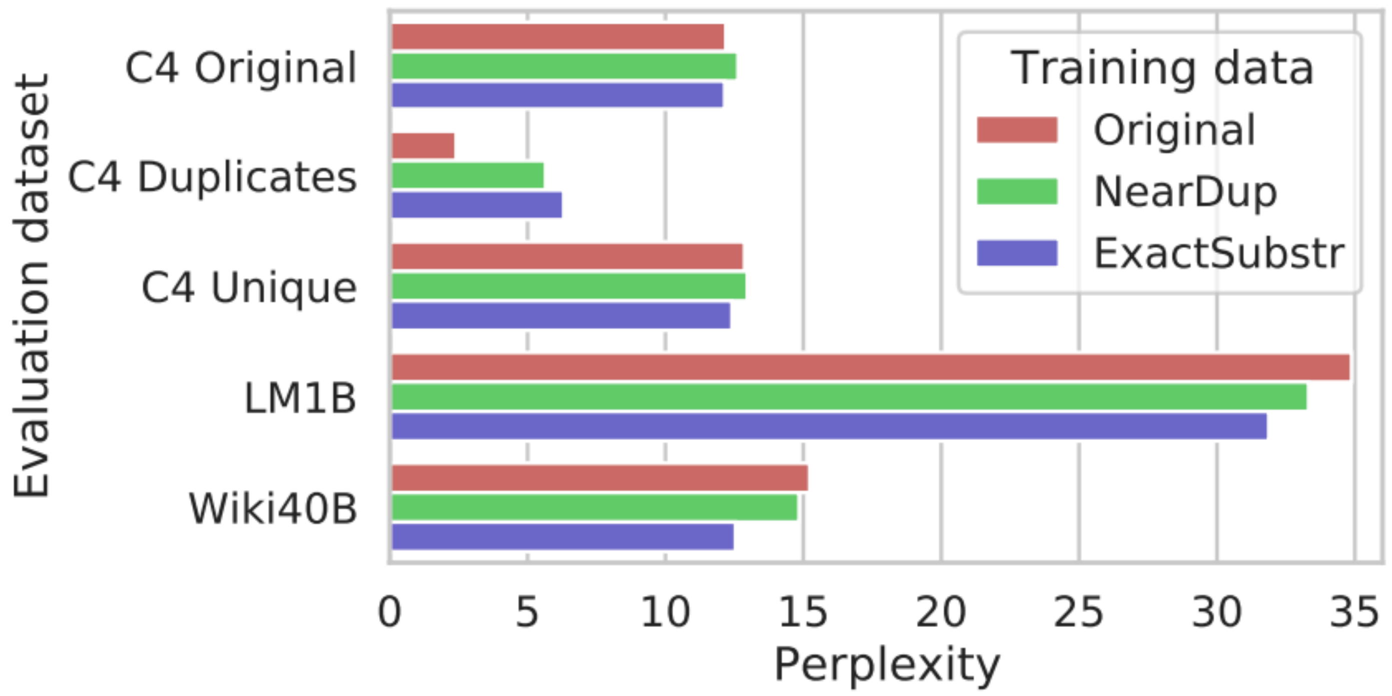
	LM	Datastore
	# of parameters	# of tokens
kNN-LM (Khandelwal et al., 2020)	250M	≤ 3B
NPM (Min et al., 2023)	350M	1B
Atlas (Izacard et al., 2022)	11B	~30B
RETRO (Borgeaud et al., 2021)	7B	2T
REPLUG (Shi et al., 2023)	≤ 175B	~5B

Reliable RAG in the wild: scaling datastore

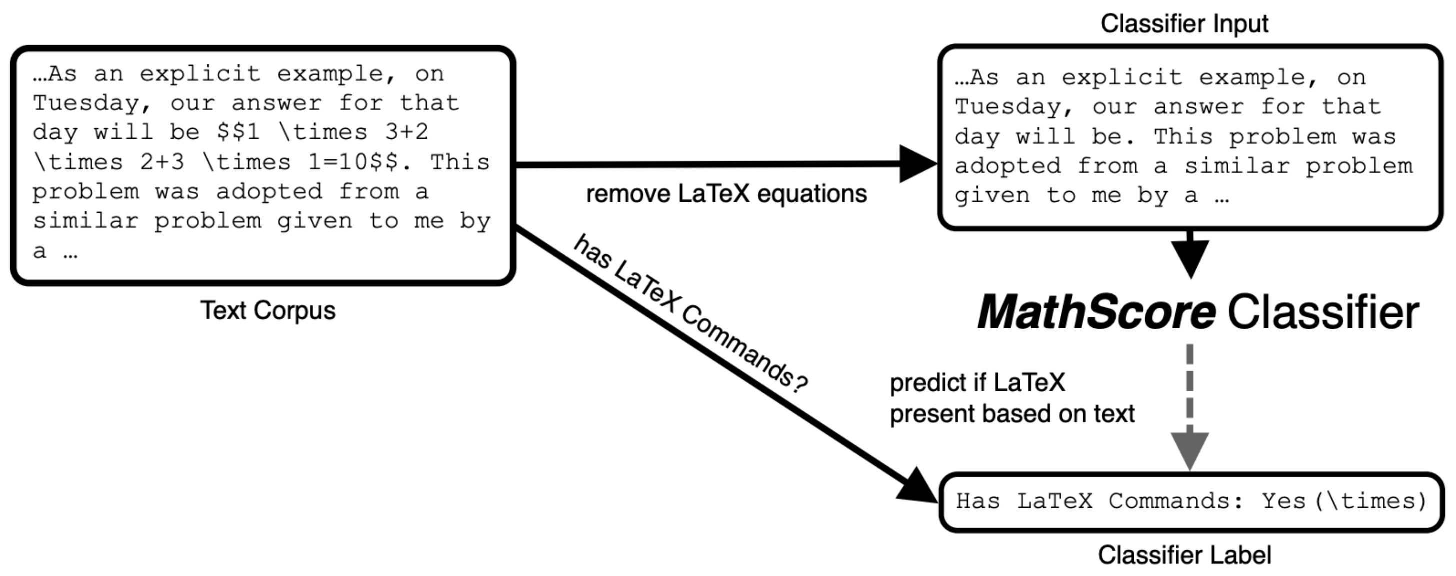
Efficiency | **Scaling datastores** | Beyond general QA

Quality & Composition
(Longpre et al., 2023)

Deduplication
(Lee et al., 2023)

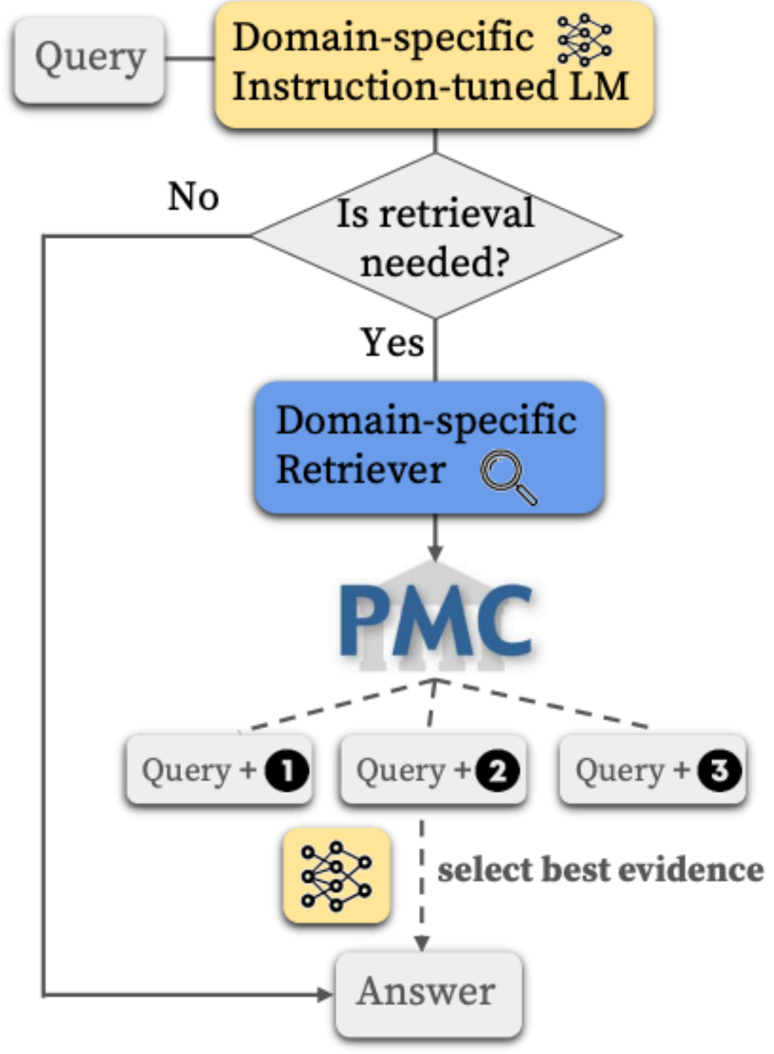


Data Filtering
(Paster et al., 2023)



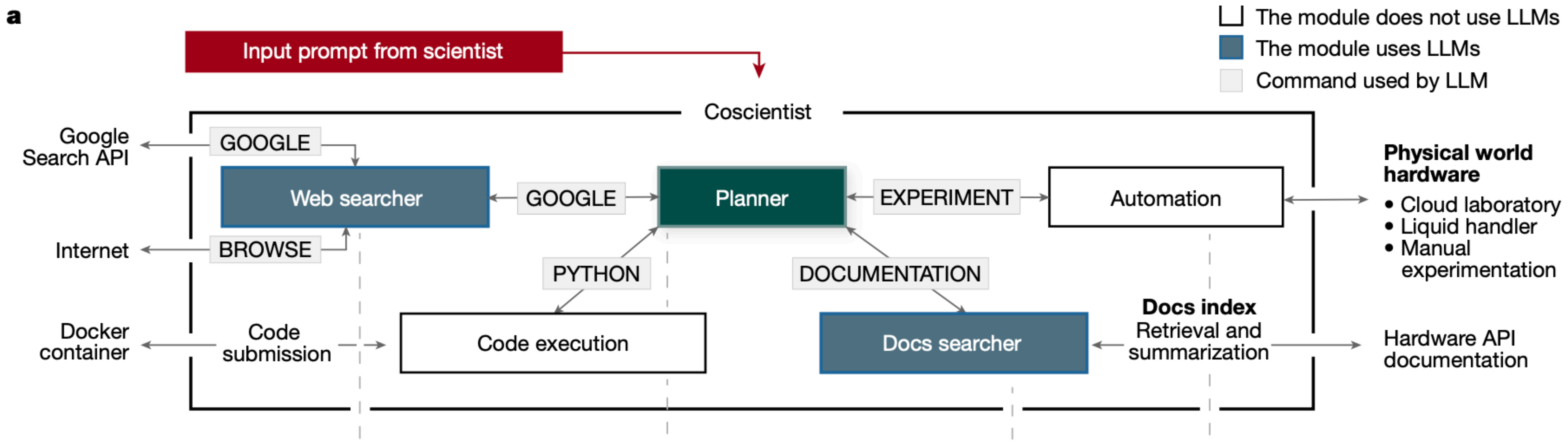
Reliable RAG in the wild: scaling datastore

Efficiency | Scaling datastores | **Beyond general QA**

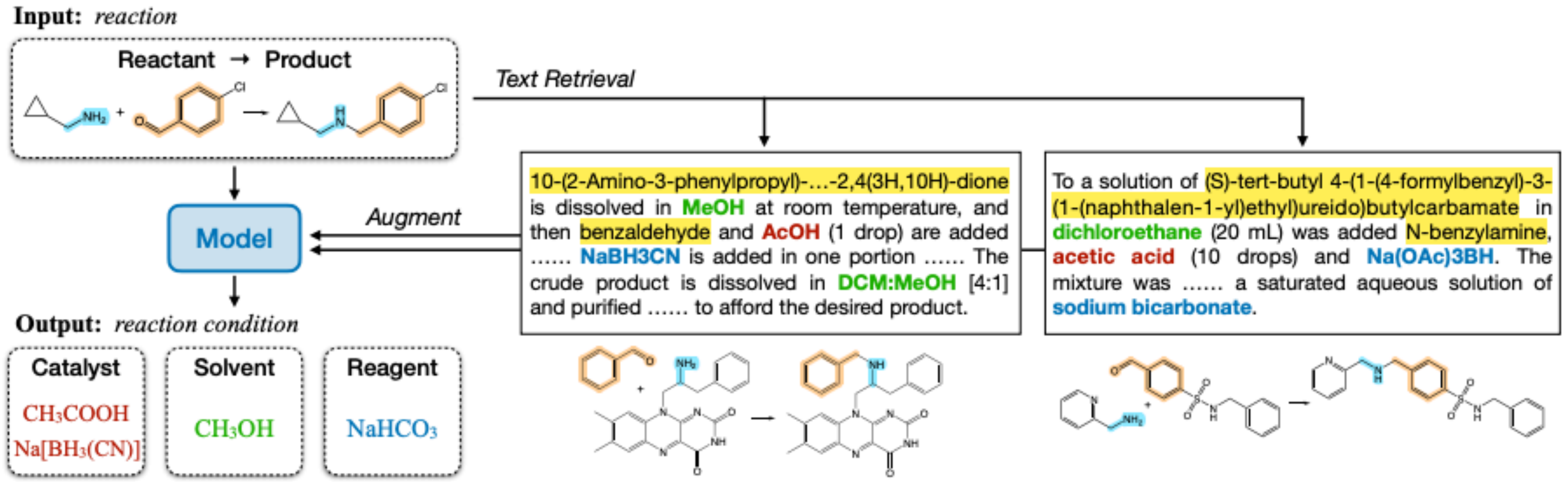


(C) Self-BioRAG (Ours)

Self-BioRAG
(Jeong et al., Bioinfomatics 2024)



RAG for automous chemistry experiments
(Boiko et al., Nature 2023)



RAG for predictive chemistry
(Jeong et al., EMNLP 2024)

Thanks for listening :)

When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Alex Mallen*, Akari Asai*, Victor Zhong, Rajarshi Das, Daniel Khashabi, Hannaneh Hajishirzi
* = core contributors

Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirul Sil, Hannaneh Hajishirzi

Reliable, adaptable, attributable LMs with Retrieval

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh
Luke Zettlemoyer, Hannaneh Hajishirzi, Wen-tau Yih

Task-aware Retrieval with Instructions

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, Wen-tau Yih

ACL 2023 tutorial: https://acl2023-retrieval-lm.github.io/ by Akari, Sewon, Zexuan and Danqi
RAG survey: Retrieval-augmented Generation for Large Language Models: A Survey (Gao et al., 2024)

Contact: akari@cs.washington.edu
Website: https://akariasai.github.io/
Twitter: @AkariAsai
Public OH: Friday 6pm

107