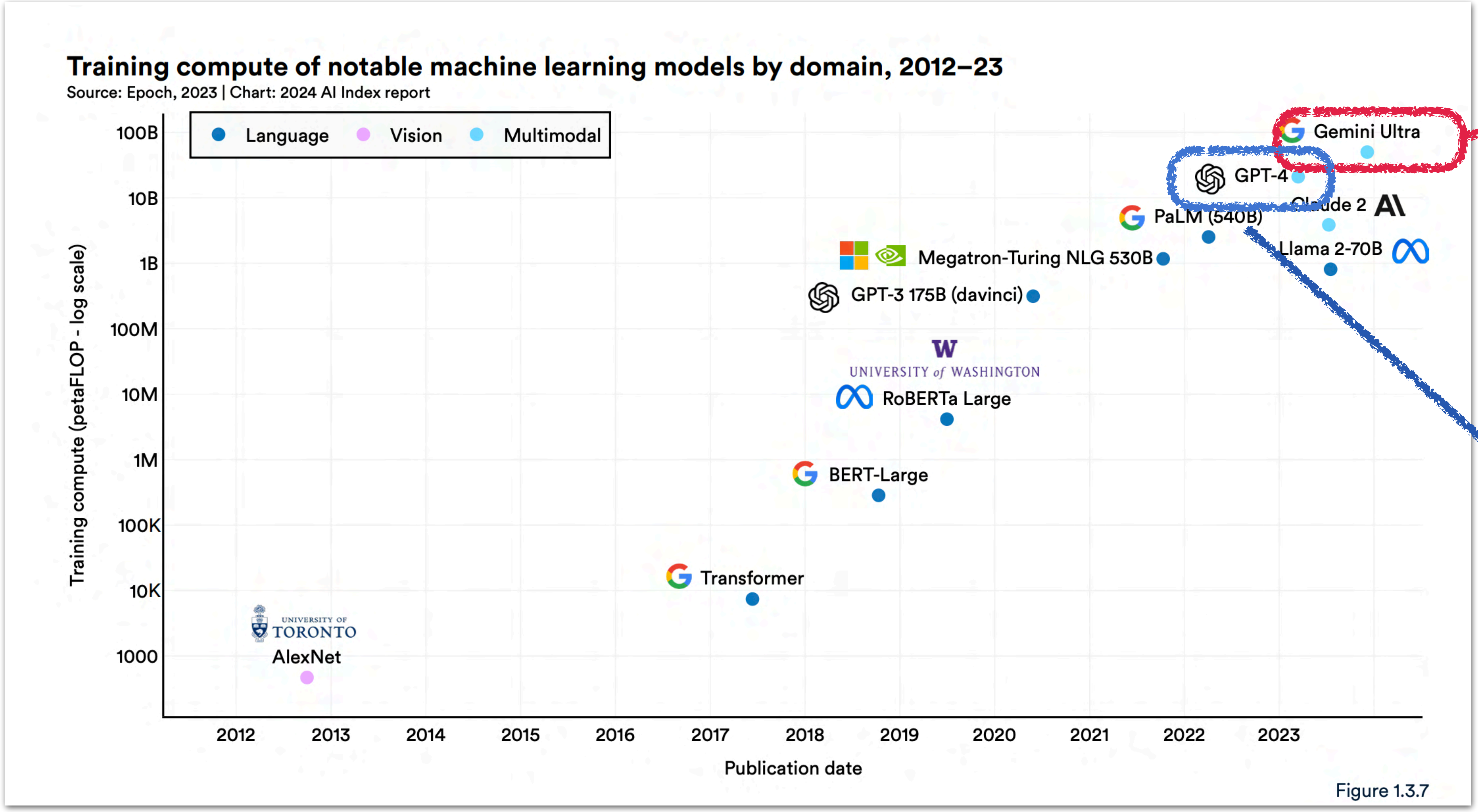
Advancing state of the art of Retrieval-augmented Generation

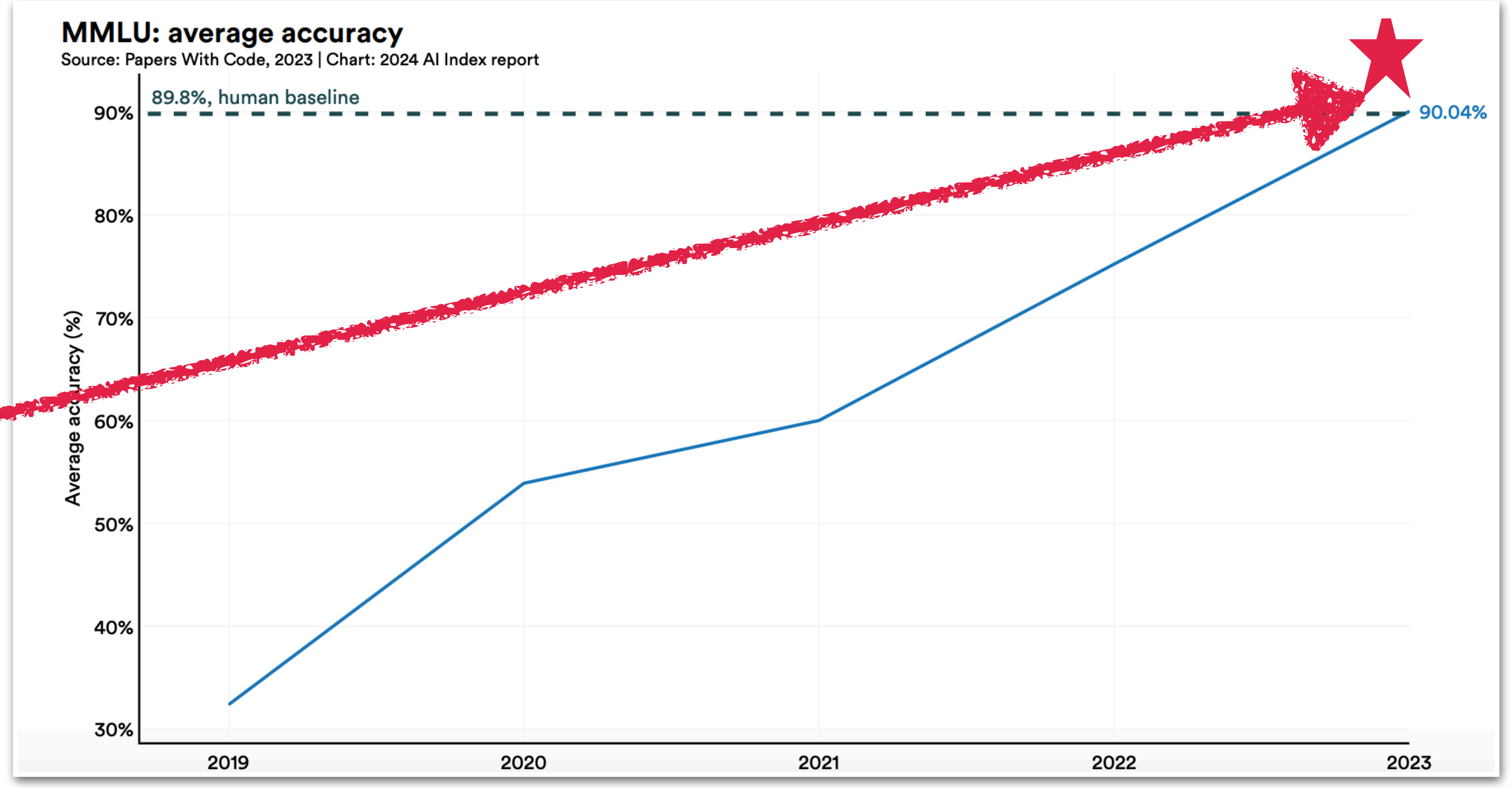
Akari Asai
University of Washington
https://akariasai.github.io/

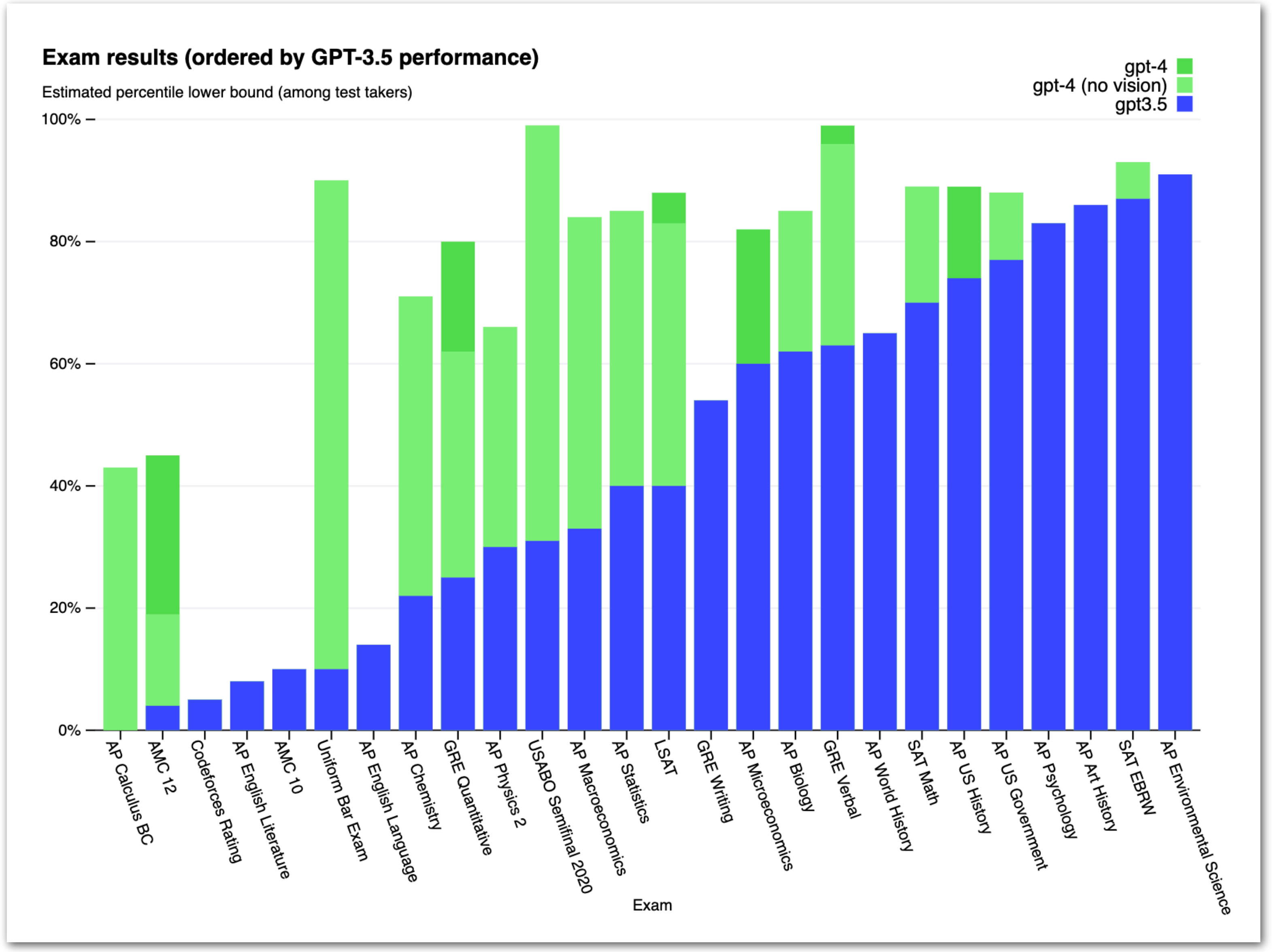
PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

April 18 | CMU CS 11-711

LLMs got much more powerful as they scale







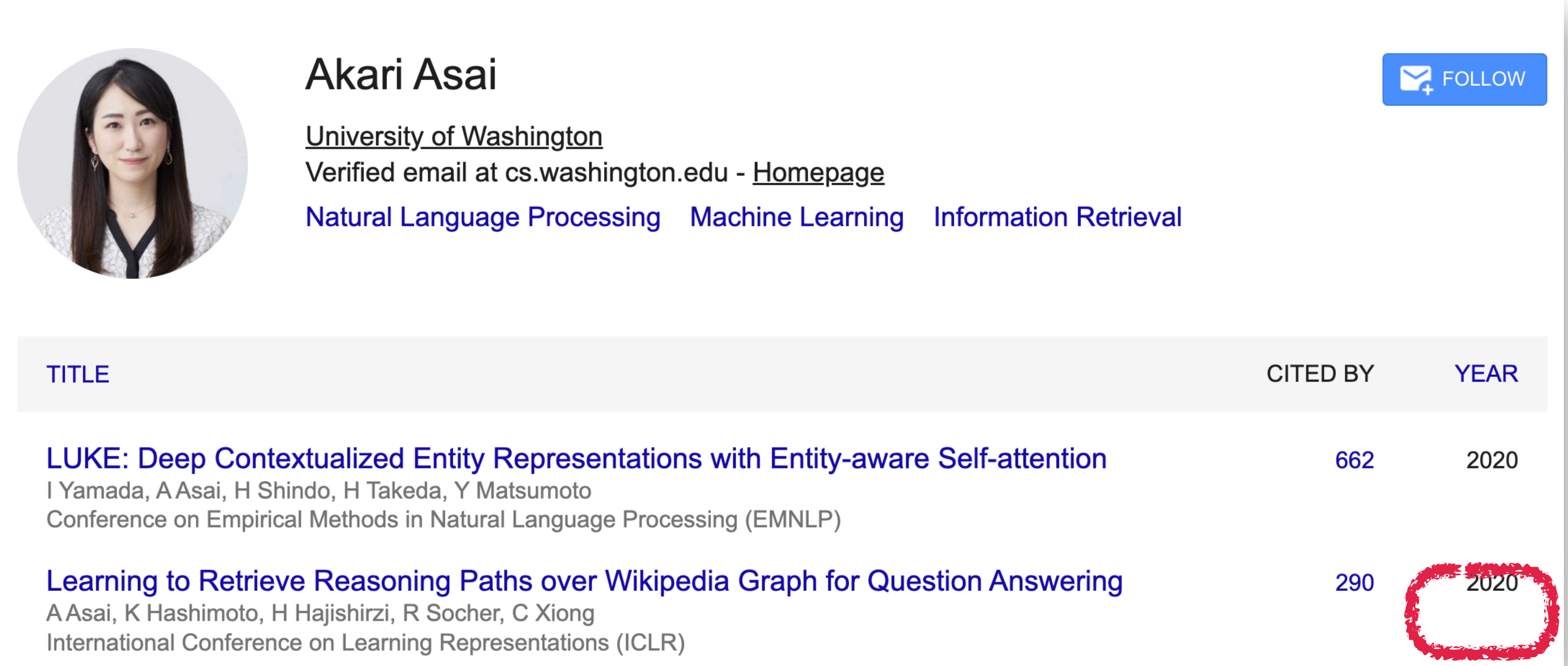
Stanford University HAI. Artificial Intelligence Index Report 2024.

Open AI. GPT-4 (OpenAI Blog). 2023.

Hallucinations in LM outputs (April, 2024 ver.)

AK **You**

Explain Akari Asai, including her famous papers. (*April 16, 2024)

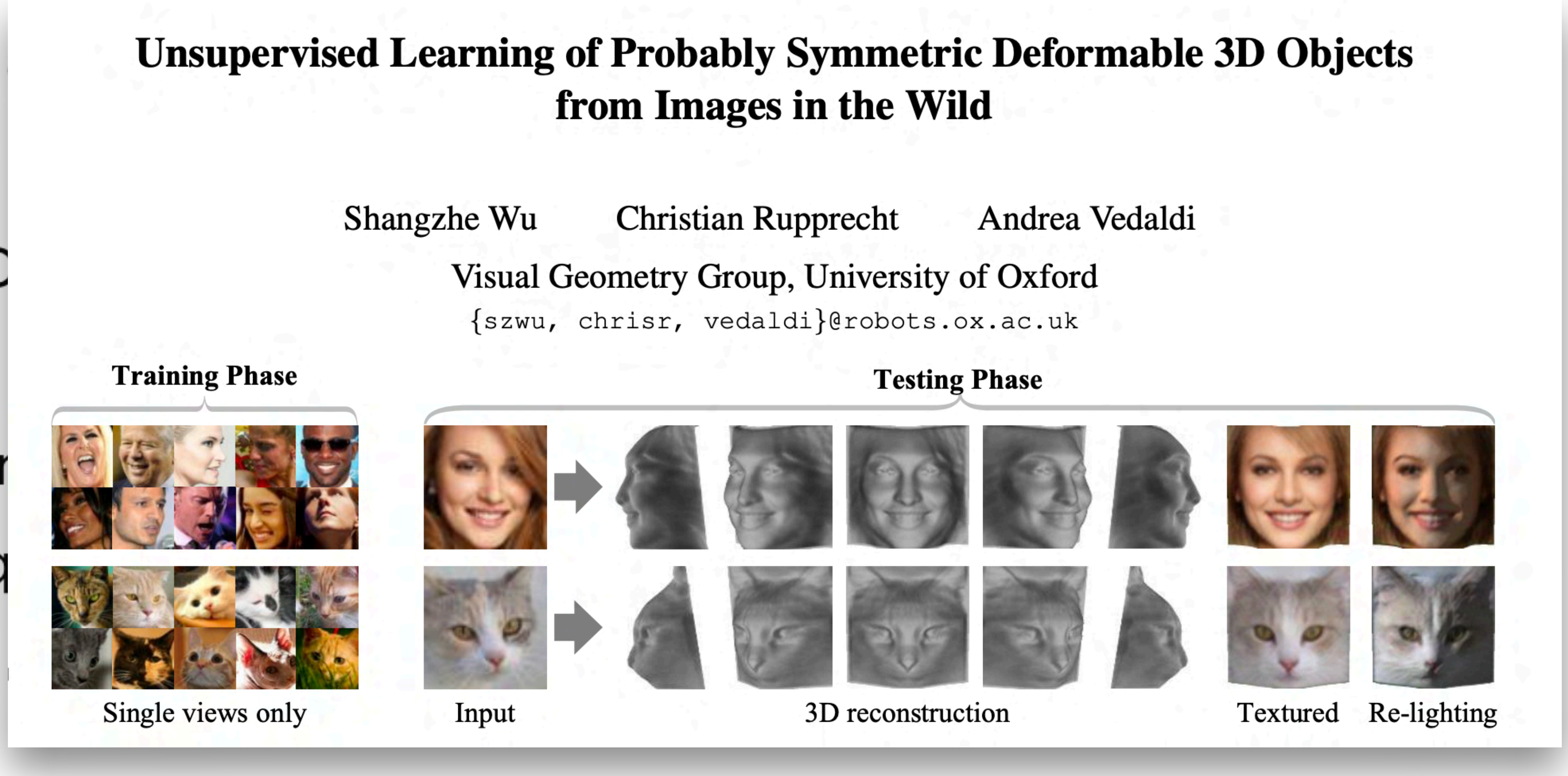


ChatGPT

Akari Asai is a prominent figure in the field of natural language processing (NLP) and learning. She has made significant contributions to various areas within these fields, the development of techniques for language understanding and representation lear

One of Asai's notable papers is "Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering" (2018), which she co-authored. In thi paper, the authors introduced a method for question answering that utilizes the stru e of Wikipedia to retrieve relevant information. They proposed a novel approach that learns to navigate the Wikipedia graph to find paths connecting relevant pieces of information, improving the accuracy of question systems.

Another influential paper by Asai is "Unsupervised Learning of Probably Symmetric D 3D Objects from Images in the Wild" (2019), where s e a d her co-authors presented unsupervised learning of 3D object representations n 2D images. This work is sign because it addresses the challenge of learning 3D object representations without req explicit supervision, which is crucial for many computer vision tasks.



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.

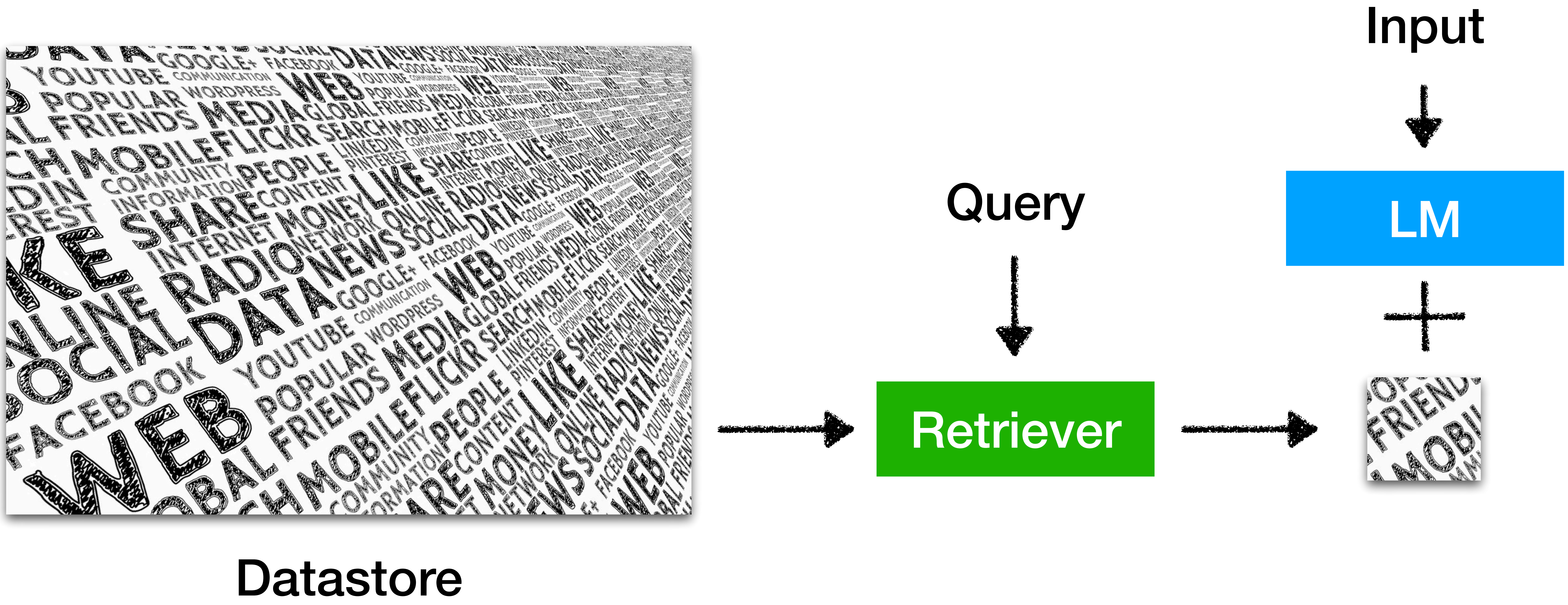
By Will Douglas Heaven November 18, 2022

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 LMs



Learn more about retrieval-augmented LMs? Check out our ACL 2023 tutorial
https://acl2023-retrieval-lm.github.io/ by Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen

Retrieval-augmented generations (RAG)

How did US states get their names?

Retriever
(e.g., Google, BM 25)

1 Of the fifty states, eleven including New York, Georgia, Washington named after an individual person.

2 UTAH: Name taken from the Ute people who inhabited that region

3 The history of human activity in Michigan began with settlement by Paleo-Indians.

Retrieve

Answer my question using references.
References: 1 2 3
Question: How did US states get their names?

LMs
(e.g., GPT-4, Llama 2)

Eleven states are named after an individual person. Some states including Utah are named after native American tribe names.

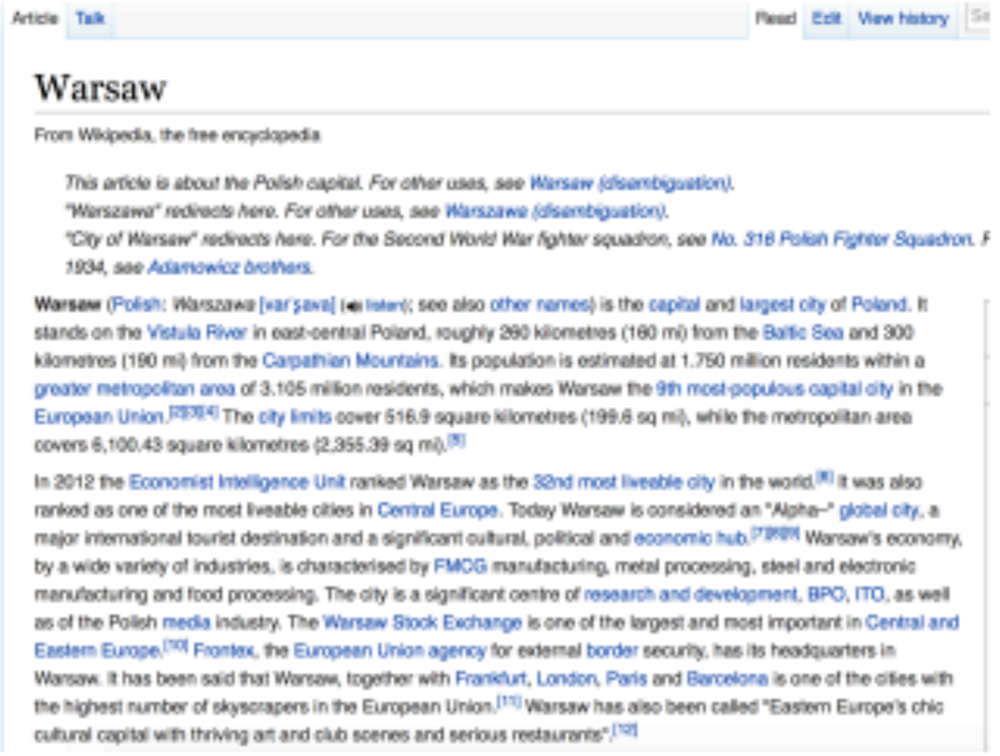
Read

Retrieval-augmented generations (RAG)

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

WIKIPEDIA
The Free Encyclopedia

Document Retriever


Article Talk Read Edit View history

Warsaw

From Wikipedia, the free encyclopedia

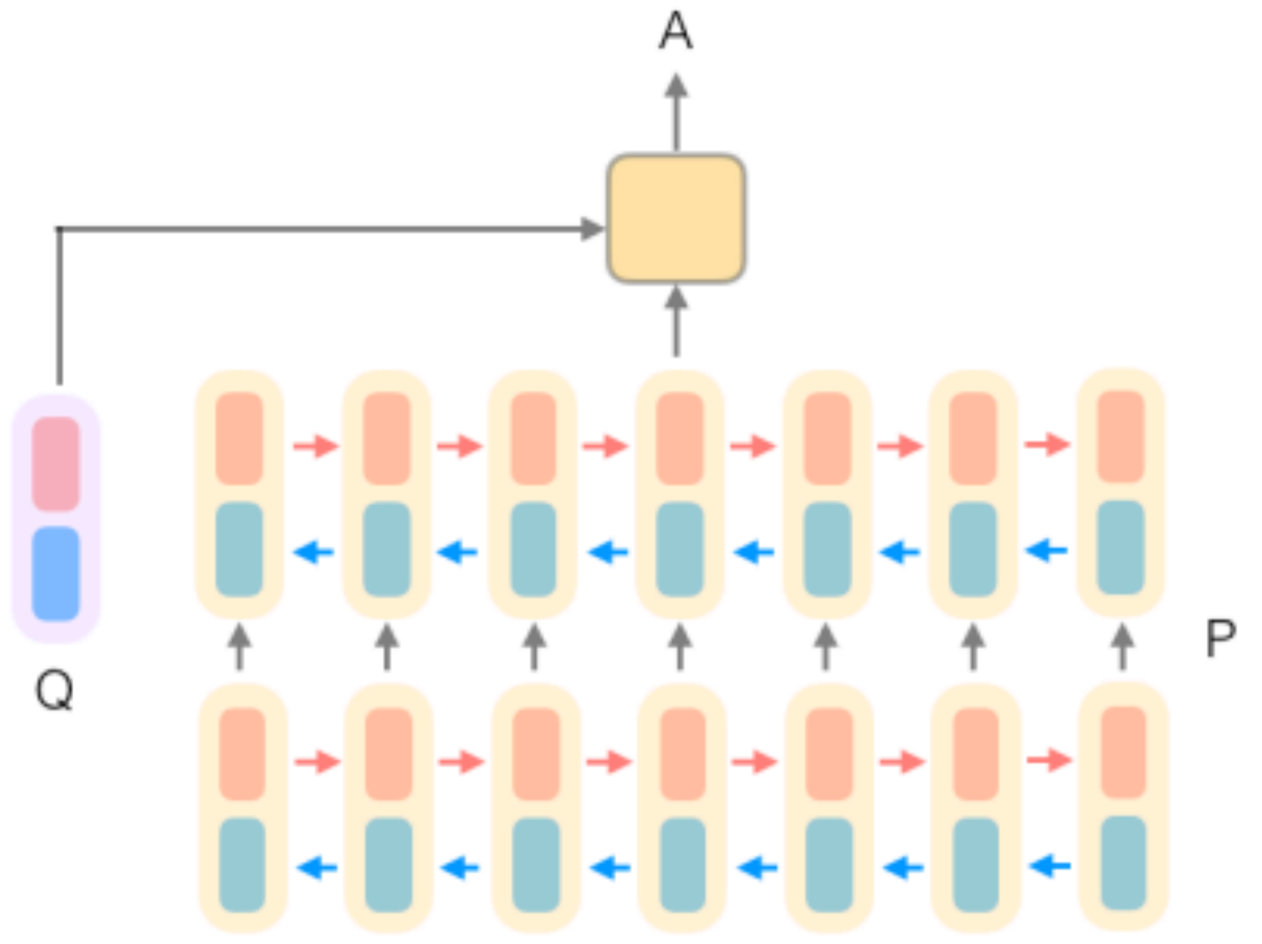
This article is about the Polish capital. For other uses, see Warsaw (disambiguation).
"Warszawa" redirects here. For other uses, see Warszawa (disambiguation).
"City of Warsaw" redirects here. For the Second World War fighter squadron, see No. 316 Polish Fighter Squadron. F… 1934, see Adamowicz brothers.

Warsaw (Polish: Warszawa [varˈʂava] (listen); see also other names) is the capital and largest city of Poland. It stands on the Vistula River in east-central Poland, roughly 260 kilometres (160 mi) from the Baltic Sea and 300 kilometres (190 mi) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within a greater metropolitan area of 3.105 million residents, which makes Warsaw the 9th most-populous capital city in the European Union.[2][3][4] The city limits cover 516.9 square kilometres (199.6 sq mi), while the metropolitan area covers 6,100.43 square kilometres (2,355.39 sq mi).[5]

In 2012 the Economist Intelligence Unit ranked Warsaw as the 32nd most liveable city in the world.[6] It was also ranked as one of the most liveable cities in Central Europe. Today Warsaw is considered an "Alpha–" global city, a major international tourist destination and a significant cultural, political and economic hub.[7][8][9] Warsaw's economy, by a wide variety of industries, is characterised by FMCG manufacturing, metal processing, steel and electronic manufacturing and food processing. The city is a significant centre of research and development, BPO, ITO, as well as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Central and Eastern Europe.[10] Frontex, the European Union agency for external border security, has its headquarters in Warsaw. It has been said that Warsaw, together with Frankfurt, London, Paris and Barcelona is one of the cities with the highest number of skyscrapers in the European Union.[11] Warsaw has also been called "Eastern Europe's chic cultural capital with thriving art and club scenes and serious restaurants".[12]

Document Reader

833,500



Retrieve

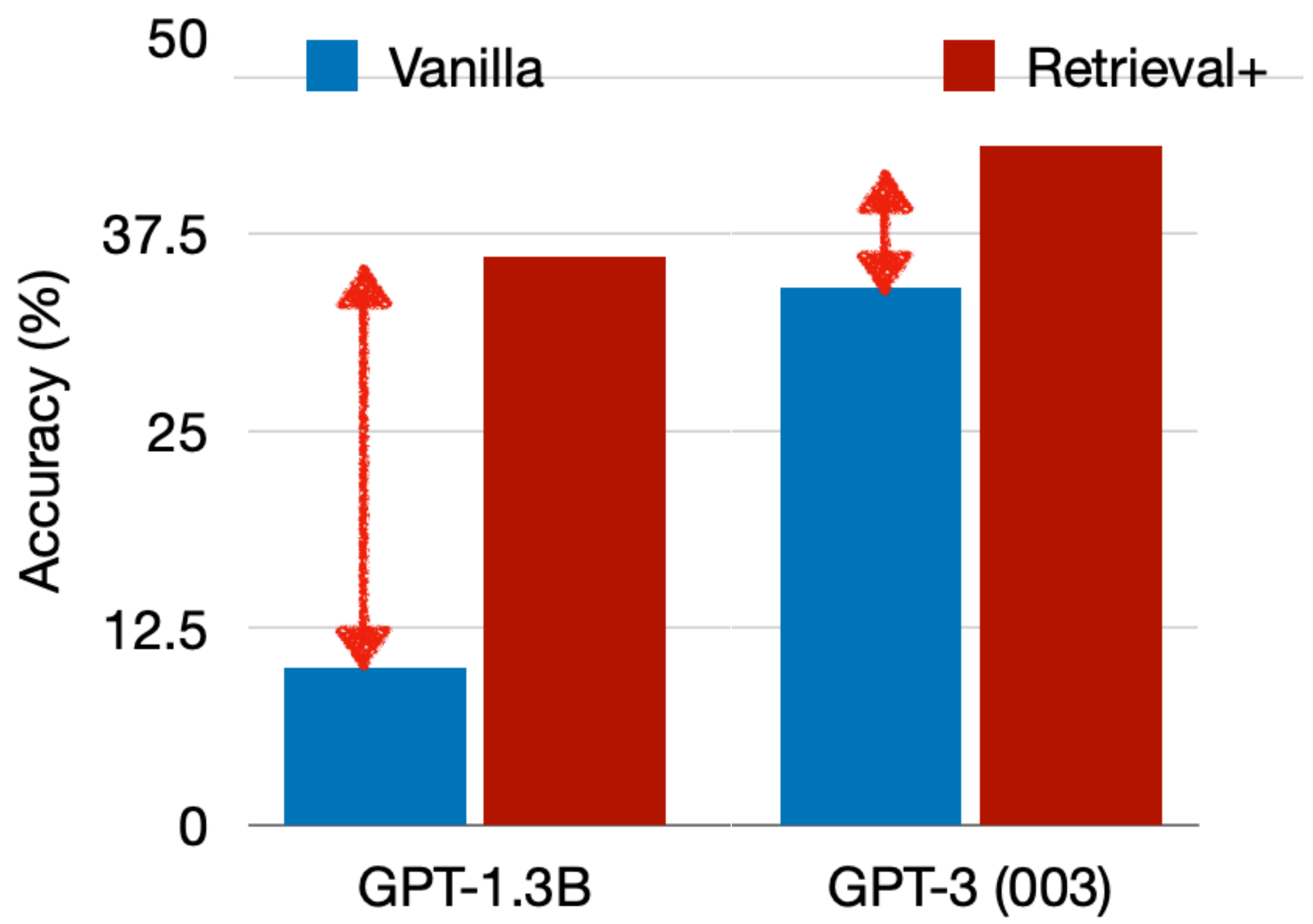
Read

Chen et al. Reading Wikipedia to Answer Open-domain Questions. ACL 2017.

RAG has been widely and successfully adopted

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. ACL 2023.

Widely applied to real-world production systems

Perplexity
Search & discovery with AI

Bard

The shift from traditional QA to *open-ended* instructions

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

Simple factoid question

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

Short answer based on single document

The shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states got 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 | …

cohere

Is this fully correct?

The shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states got 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

Hint: 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 | …

⌘ R

cohere

Is this fully correct?

The shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states got 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 | ...

cohere

The shift from traditional QA to *open-ended* instructions

Make a table for me summarizing how different US states got 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

The shift from traditional QA to *open-ended* instructions

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

Complex instruction

Long-form answers

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

Requiring aggregating multiple evidence

10 references

WIKIPEDIA
LIST OF STATE AND TERRITORY NAME ETY...

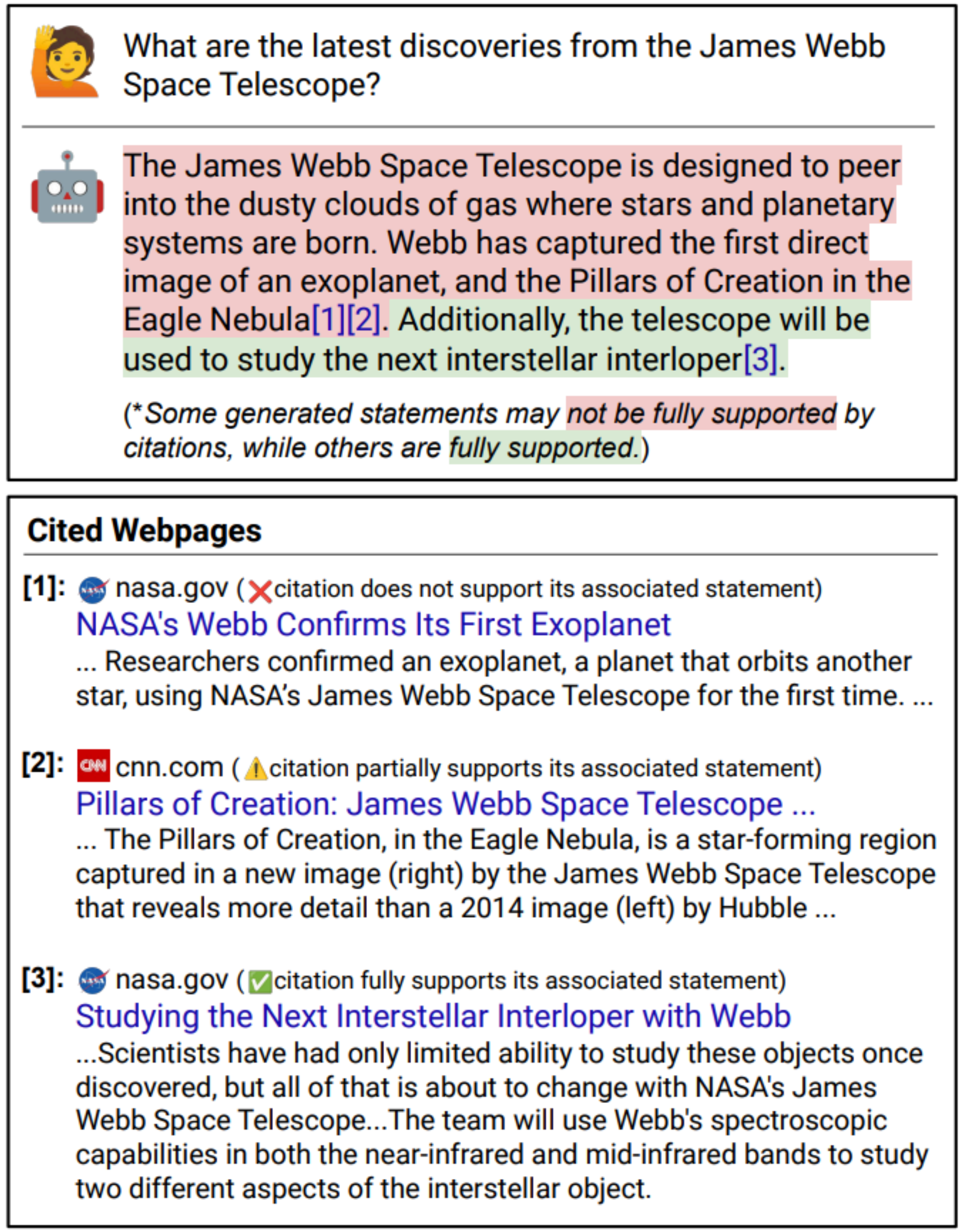
BIA
ORIGIN OF NAMES OF US STATES | INDIA...

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

⌘ R

cohere

Challenges of the current naive RAG systems: reliability

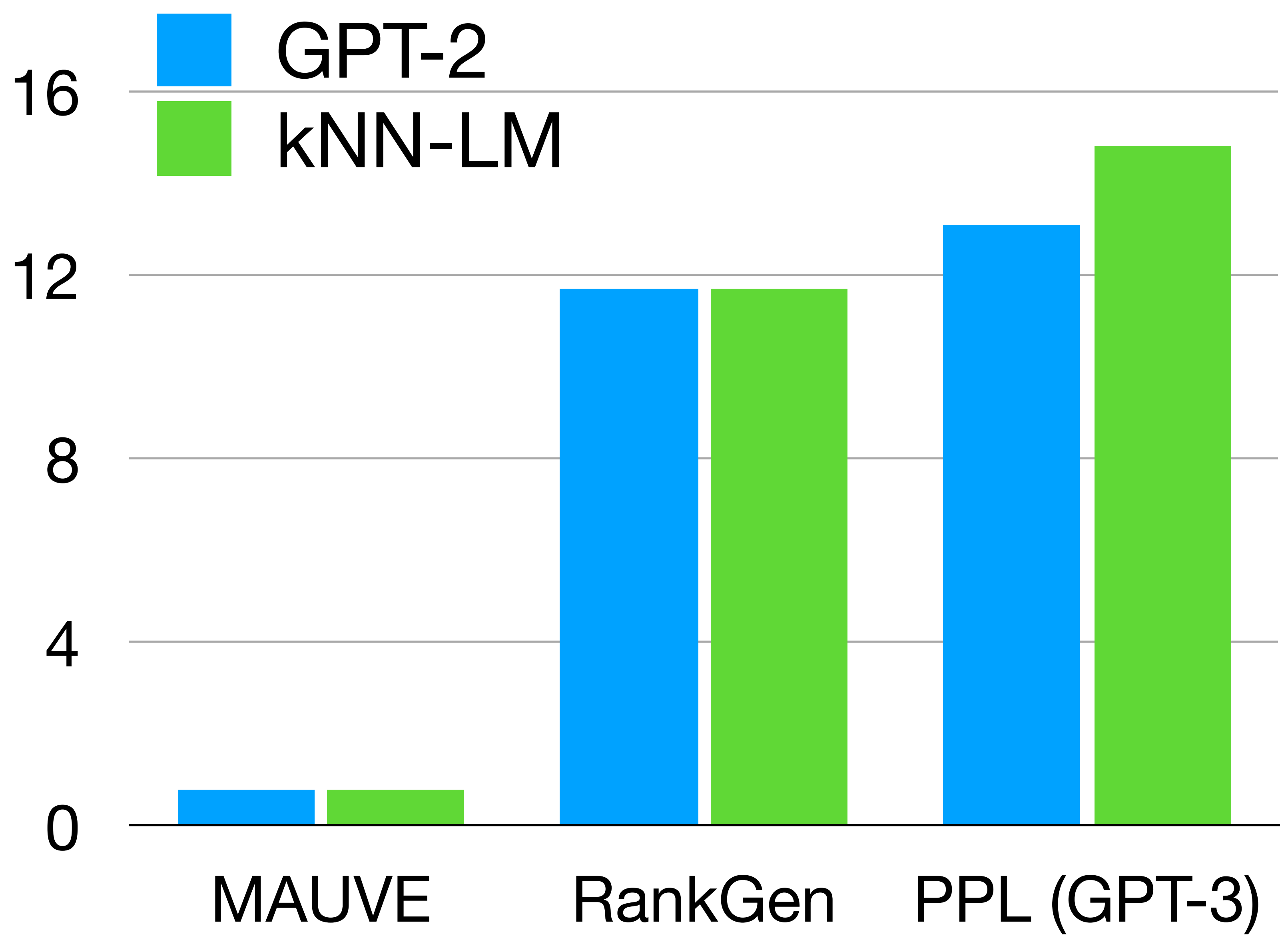


Liu et al. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023.

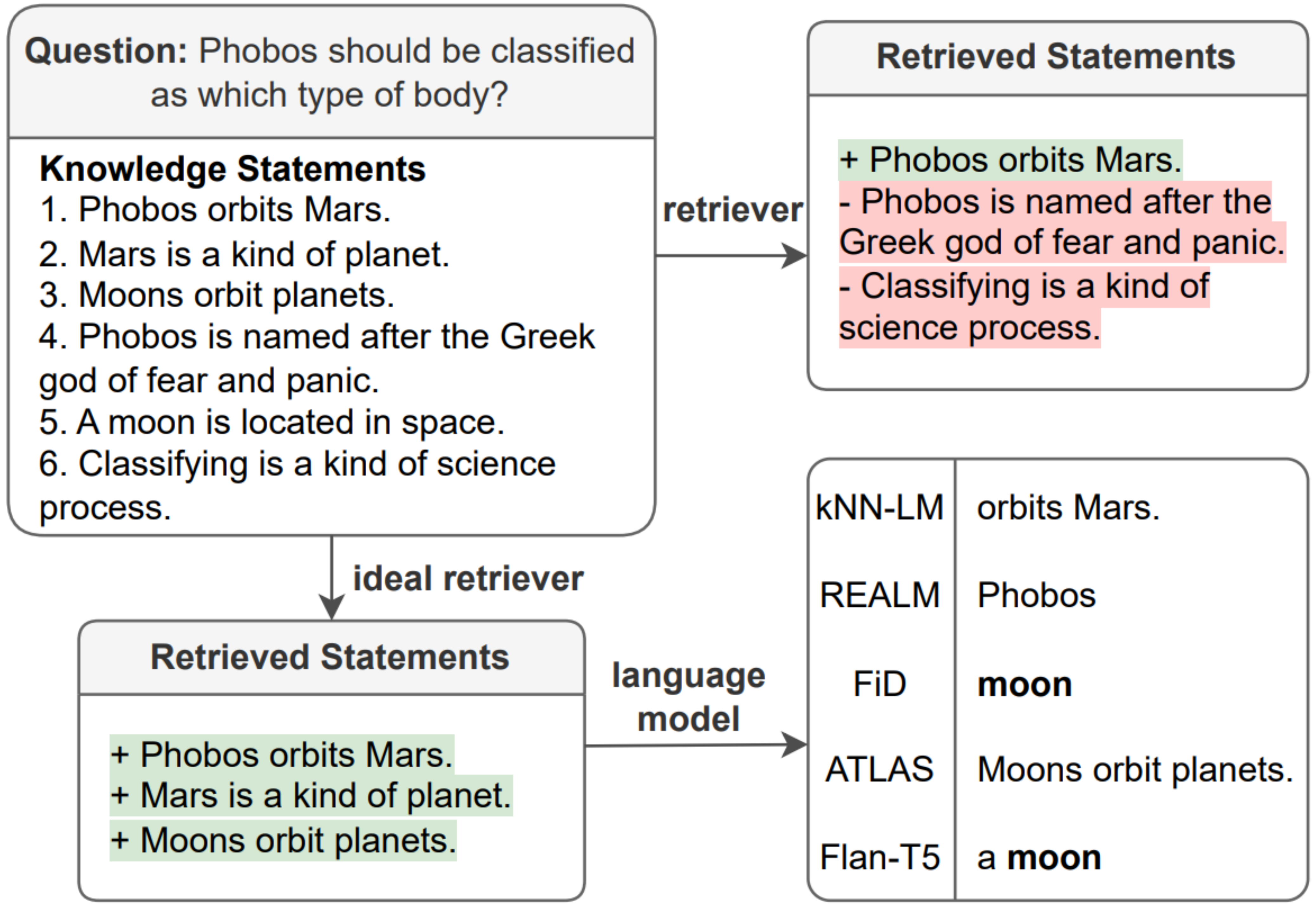
Marcus. No, RAG is probably not going to rescue the current situation. 2024.

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?

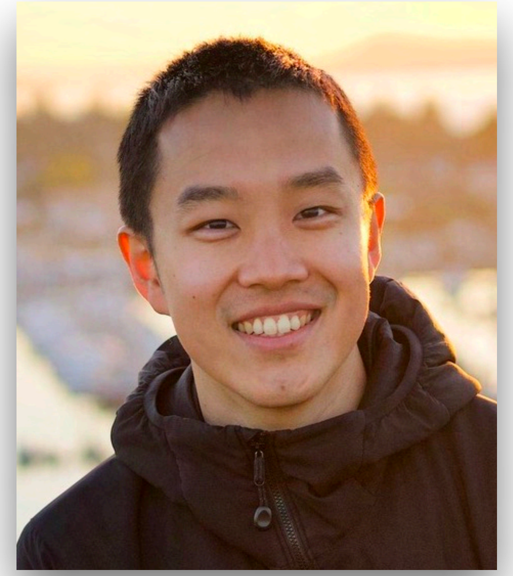
A: Because RAG can solve many core limitations of parametric LMs!

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

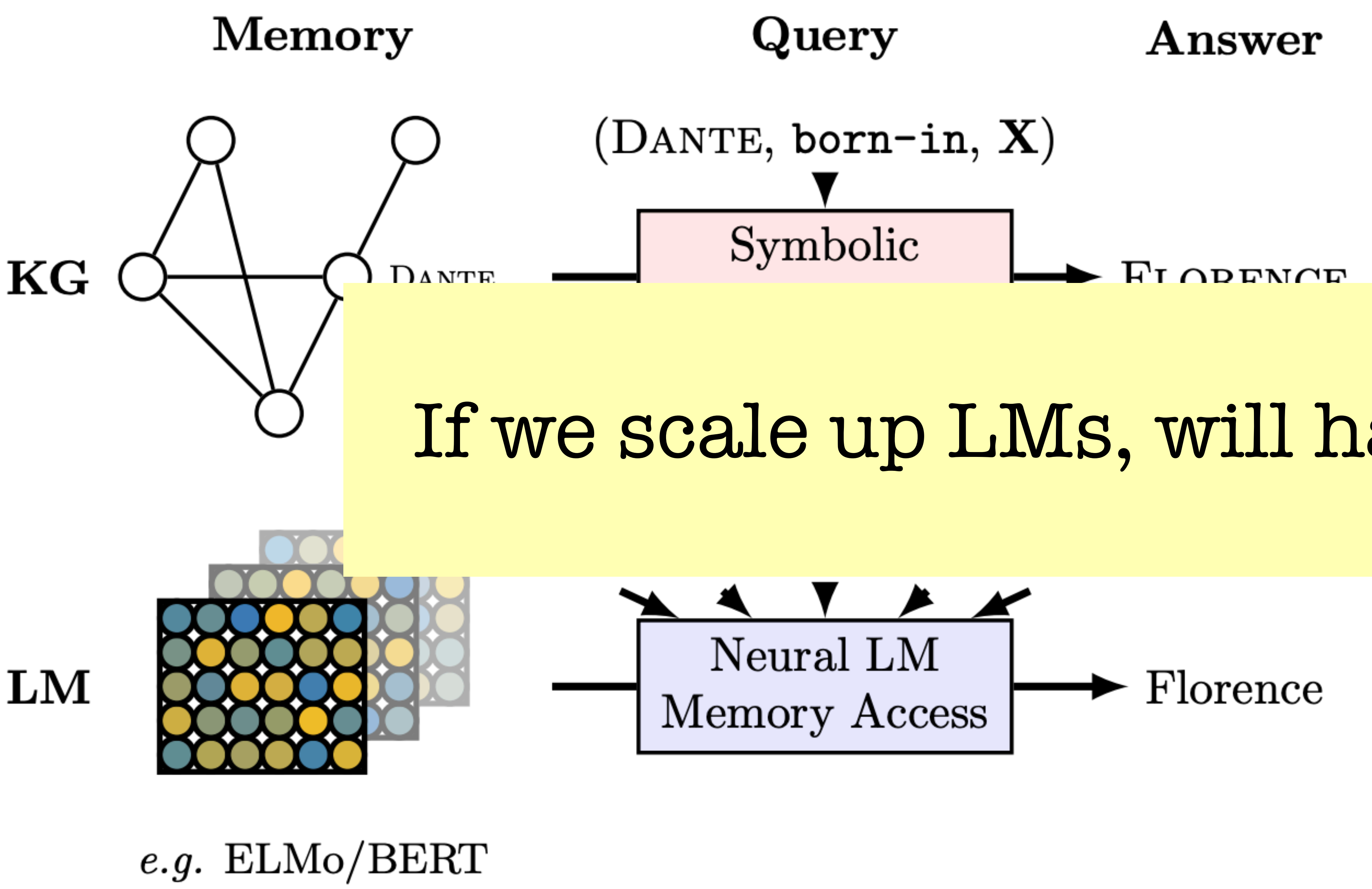
* = core contributors

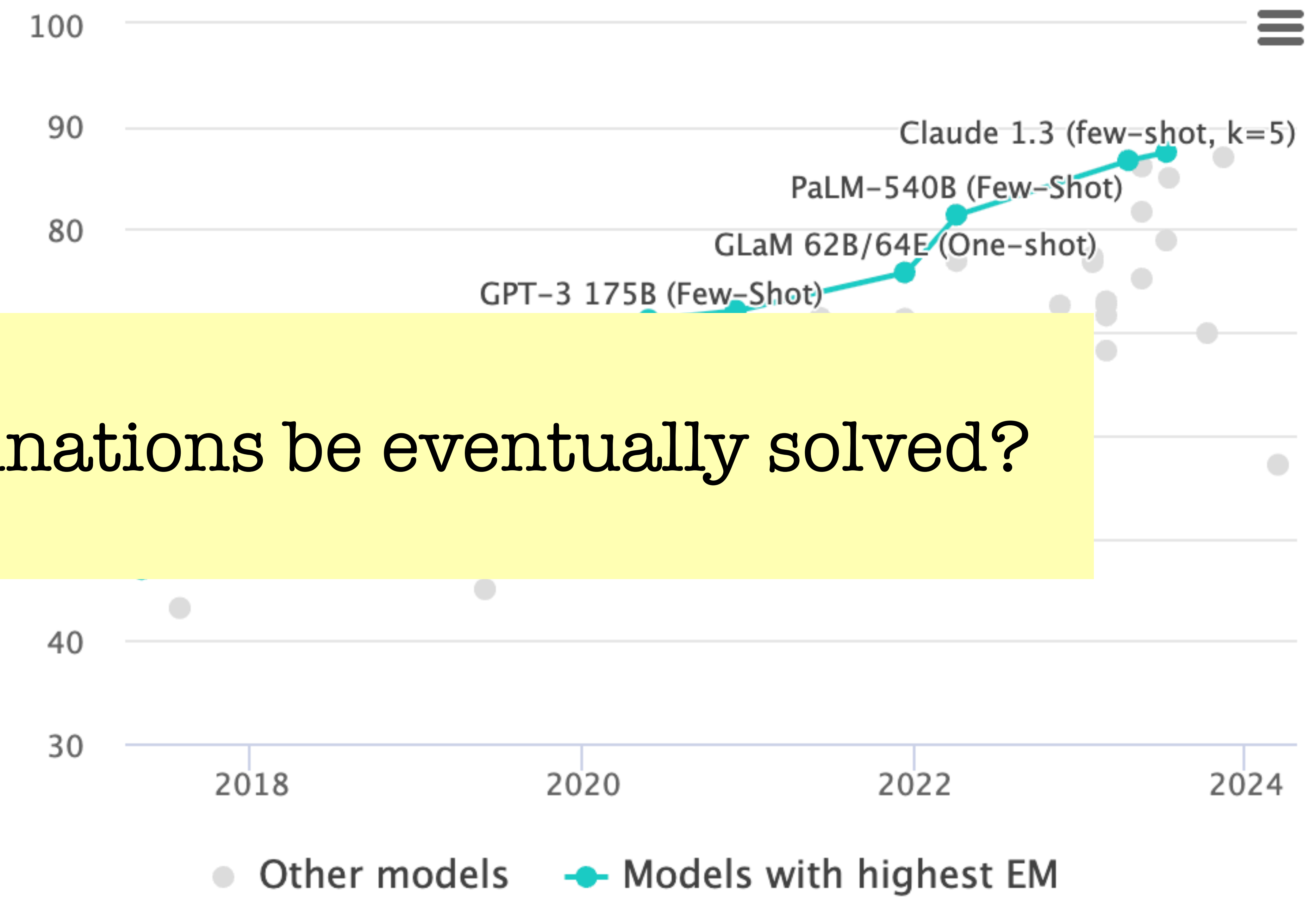





ACL 2023 (Oral, Best Video Award — most viewed)

Factual knowledge memorization in LLMs





If we scale up LMs, will hallucinations be eventually solved?

Petroni et al. Language Models as Knowledge Bases?. EMNLP 2019.

TriviaQA (paper with code)
https://paperswithcode.com/sota/question-answering-on-triviaqa

Focus and task

Focus:
Factual knowledge

(Kathy Saltzman, Occupation, Politician)

Subject Relationship Object

Task:
Open-domain QA

Relationship Subject

What is the occupation of Kathy Saltzman?
A: Politician

Object

Dimensions of analysis

WIKIDATA

We created a new dataset, PopQA (17k openQA questions with fine-grained meta data). See more details in our paper!

Aspect 1:
Subject entity popularity

(Kathy Saltzman, Occupation, Politician)

∧

(Barack Obama, Occupation, Politician)

Aspect 2:
Relationship type

(Barack Obama, Elementary School, St. Francis Assisi)

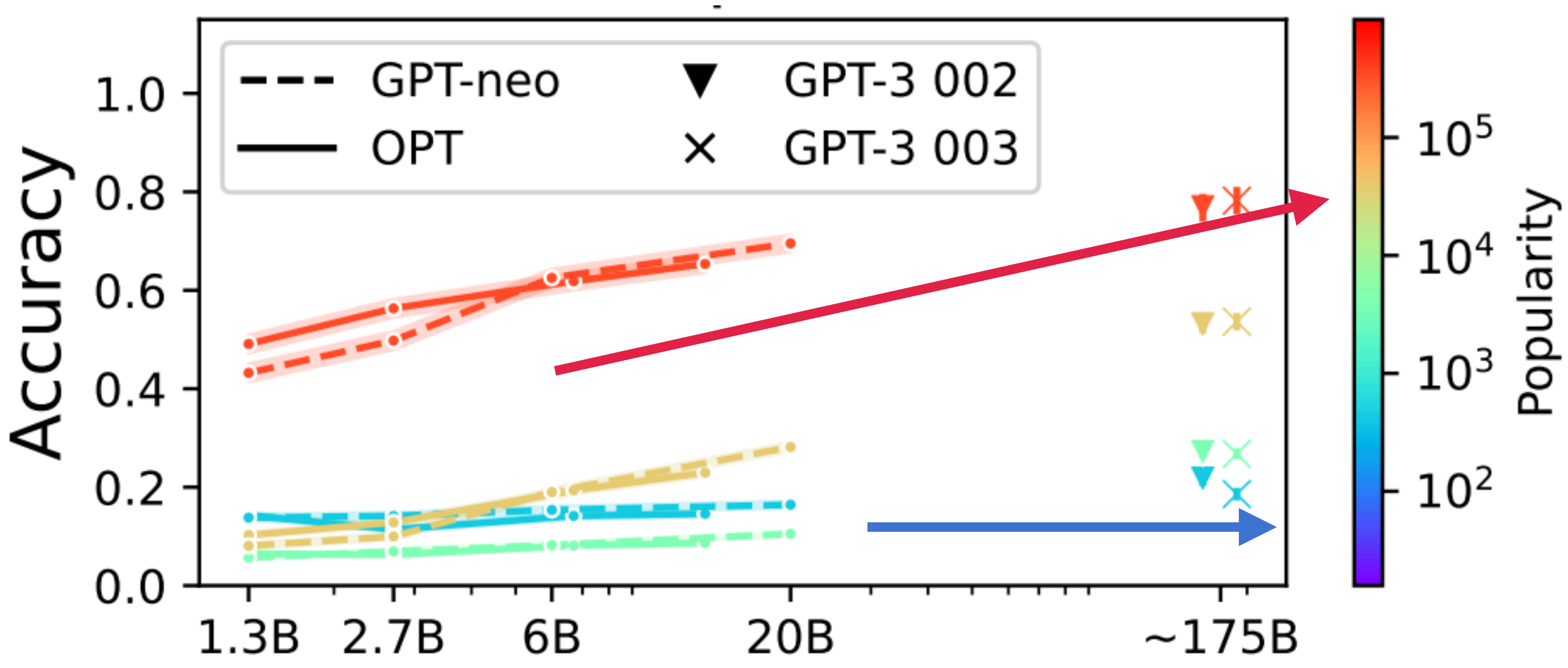
∧

(Barack Obama, Occupation, Politician)

Does scaling solve memorization? Probably not!

On popular facts Performance gets better as scaling.

On long-tail facts Almost flat trends.



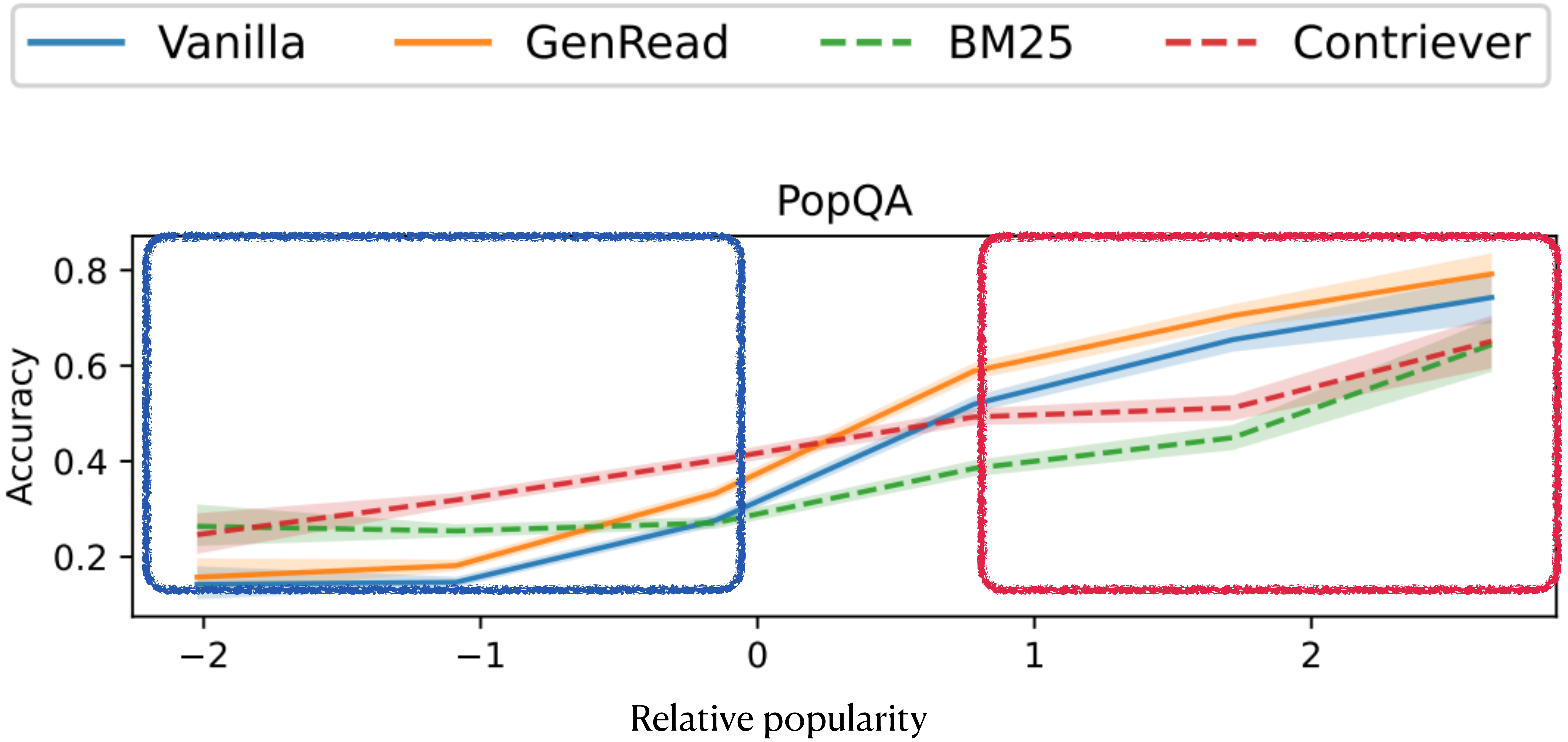
Scaling may not overcome hallucination in long tail!

RAG can address hallucinations in such long-tail!

On popular facts vanilla ≈ RAG or vanilla > RAG

will discuss why in the next section!

On long-tail facts RAG > vanilla

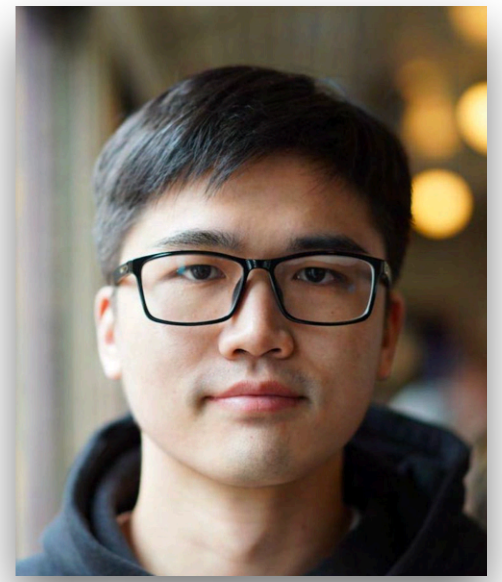
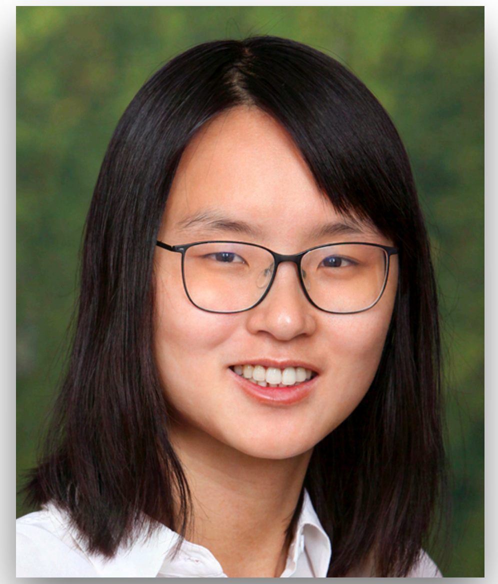


RAG gives large gain in long-tail questions!

Is RAG only useful for hallucination?

Reliable, adaptable, attributable LMs with Retrieval

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



Arxiv 2024
https://arxiv.org/abs/2403.03187

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.

ChatGPT

Akari Asai is a prominent figure in the field of natural language processing (NLP) and machine learning. She has made significant contributions to various areas within these fields, particularly in the development of techniques for language understanding and representation learning.

One of Asai's notable papers is "Learning to [illegible]etrieve Reasoning Paths over Wikipedia Graph for Question Answering" (2018), which [illegible]thored. In this paper, the authors introduced a method for question answering that [illegible]es the structure of Wikipedia to retrieve relevant information. They proposed a no[illegible]h that learns to navigate the Wikipedia graph to find paths connecting relevant pieces [illegible] information, improving the accuracy of question answering systems.

Another influential paper by Asai is "Unsupervised Learning of Probably Symmetric Deformable 3D Objects from Images in the Wild" (2019), where she and her co-authors presented a method for unsupervised learning of 3D object representations from 2D images. This work is significant because it addresses the challenge of learning 3D object representations without requiring explicit supervision, which is crucial for many computer vision tasks.

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.

↓

Language model

↓

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

Core limitations of parametric LMs

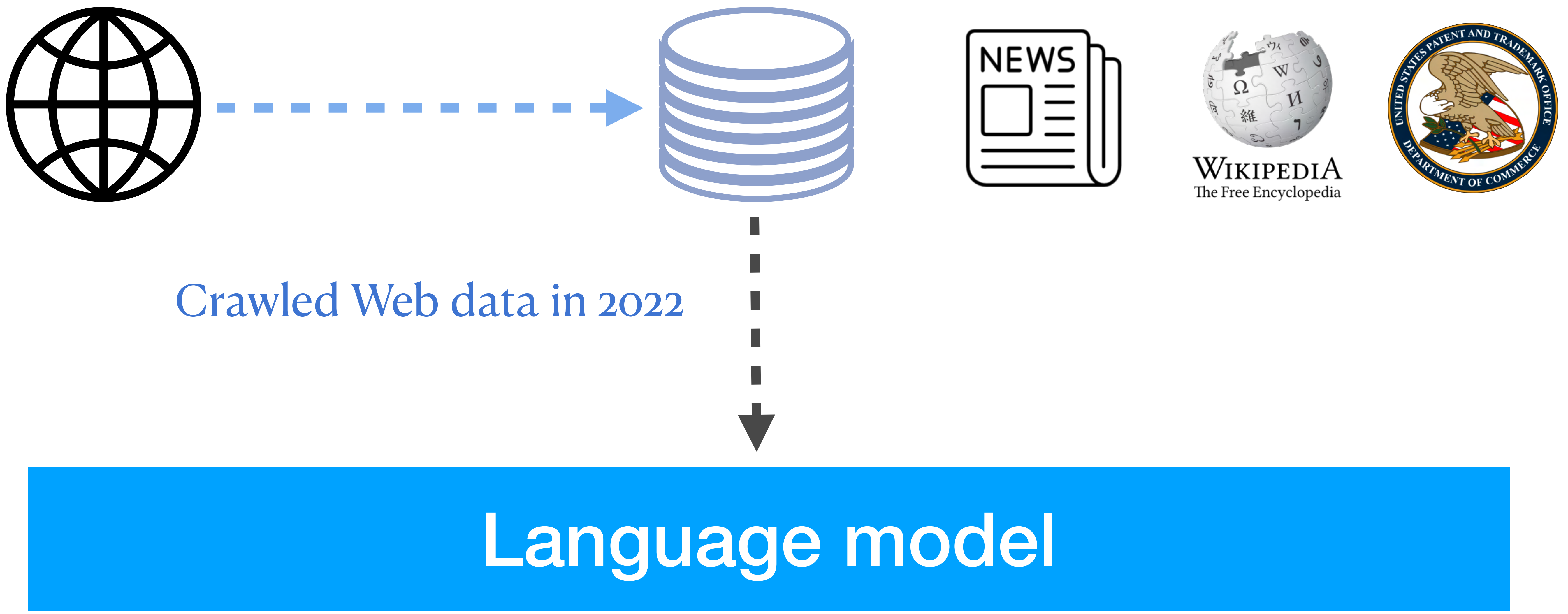
Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



ChatGPT

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

Core limitations of parametric LMs

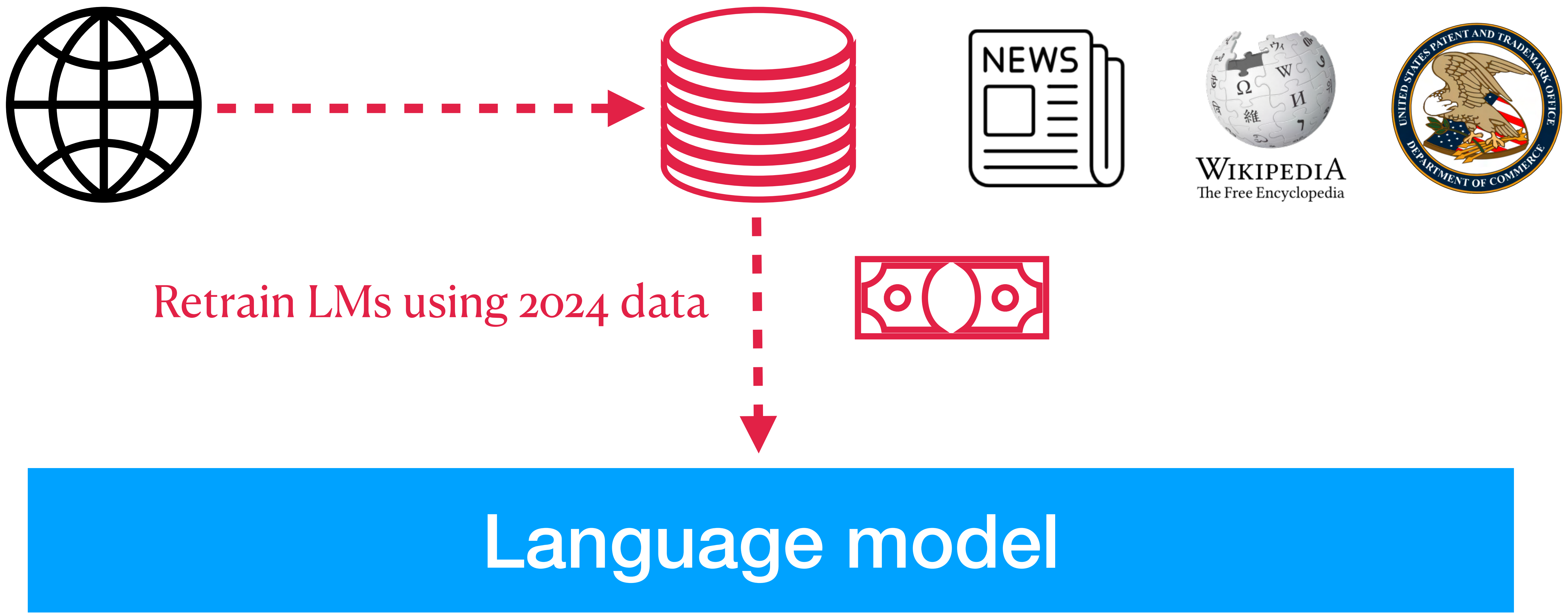
Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



ChatGPT

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

Core limitations of parametric LMs

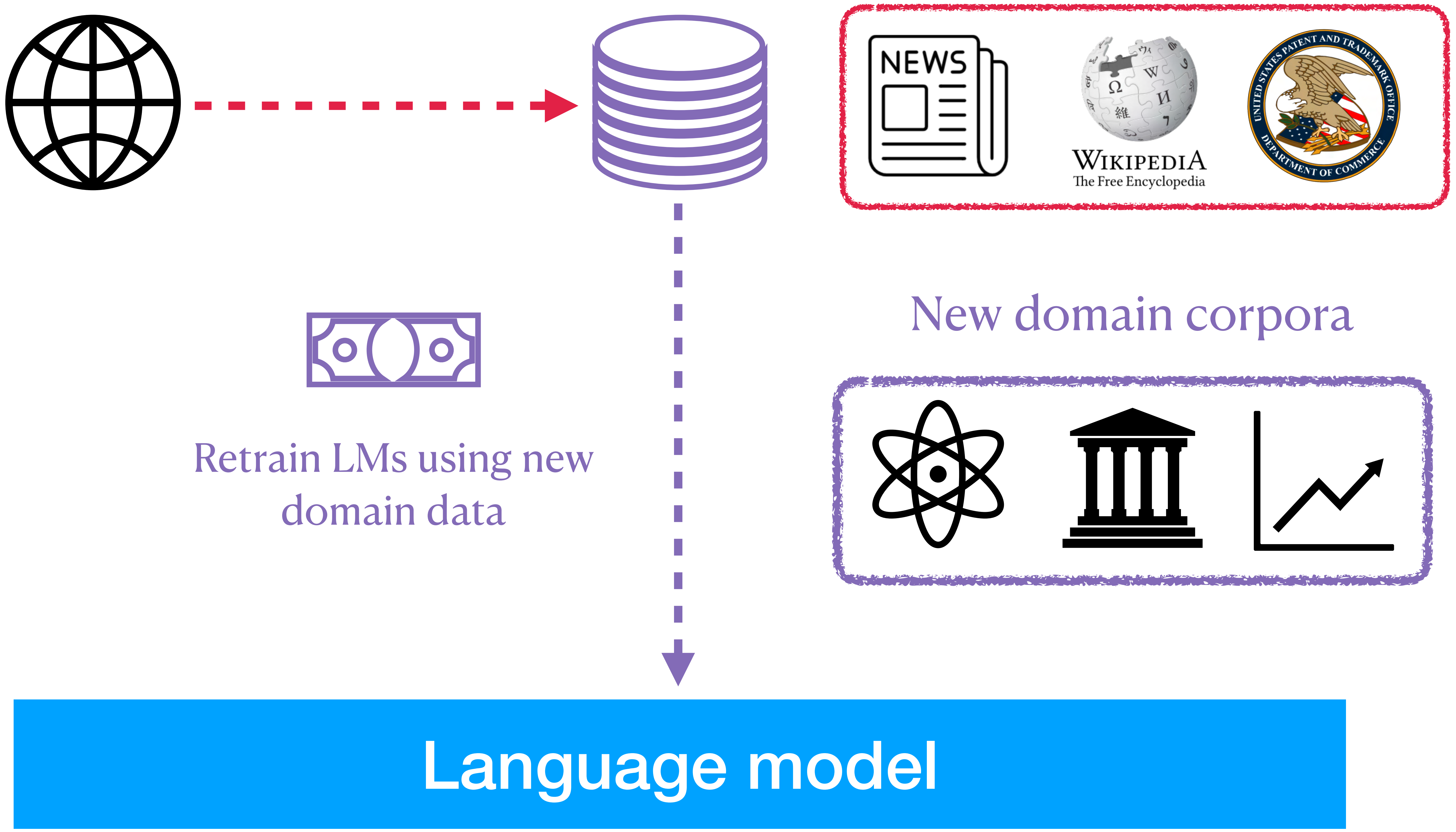
Hallucinations

Lack of attributions

Costs of adaptations

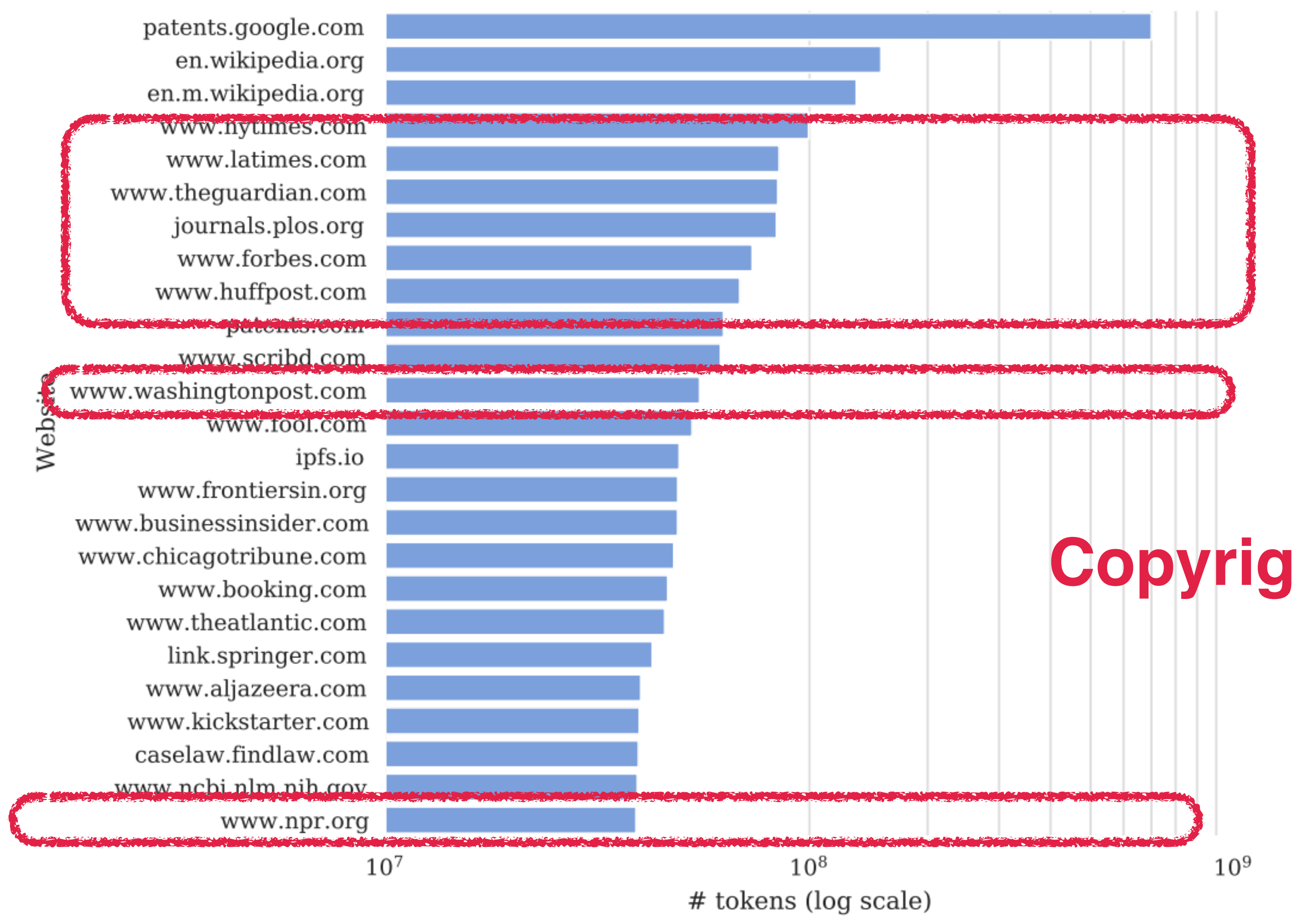
Copyright / privacy

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

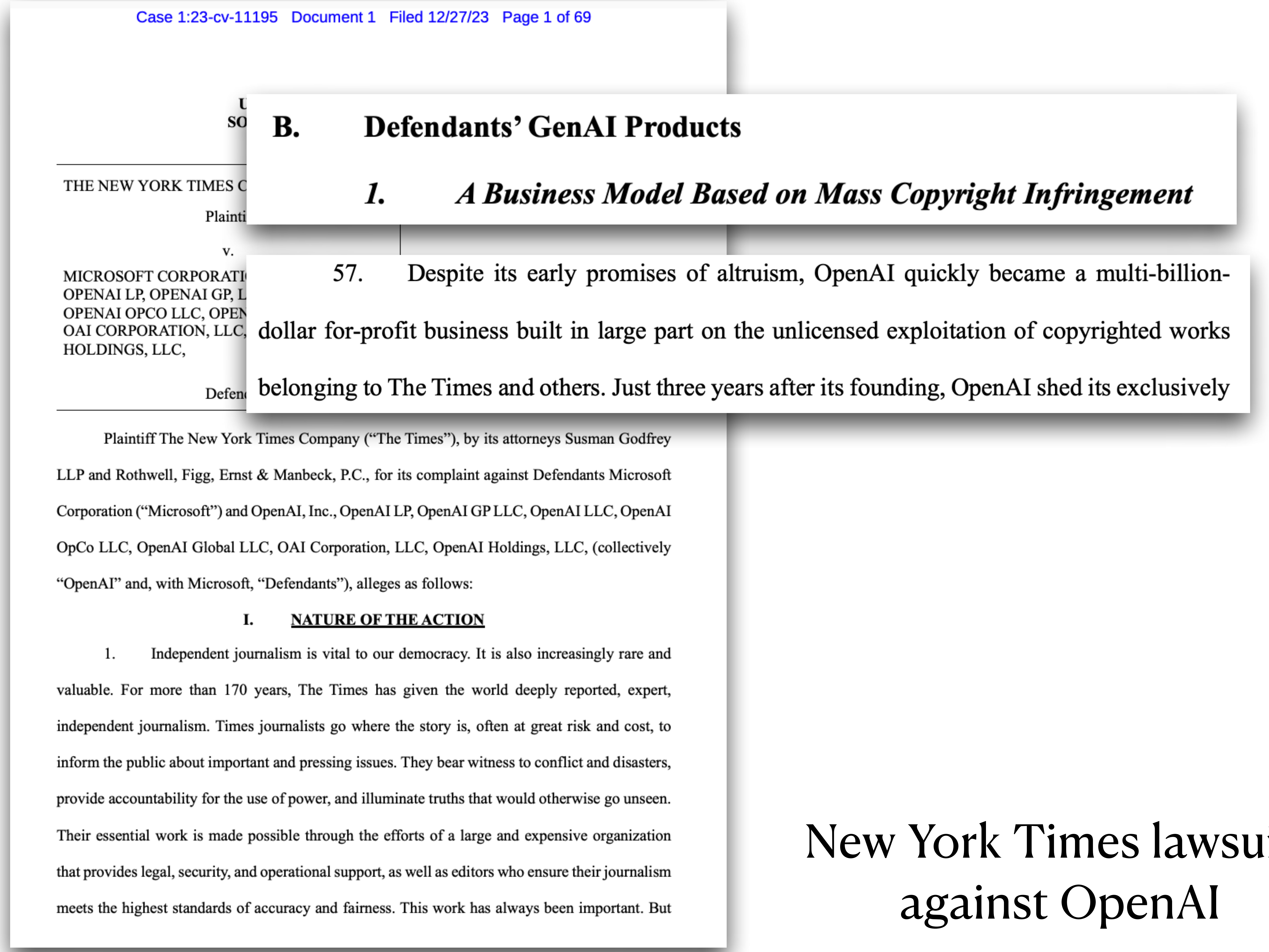
Hallucinations

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

THE NEW YORK TIMES C

Plainti

v.

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

Defen

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

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

New York Times lawsuits against OpenAI

Core limitations of parametric LMs

Hallucinations

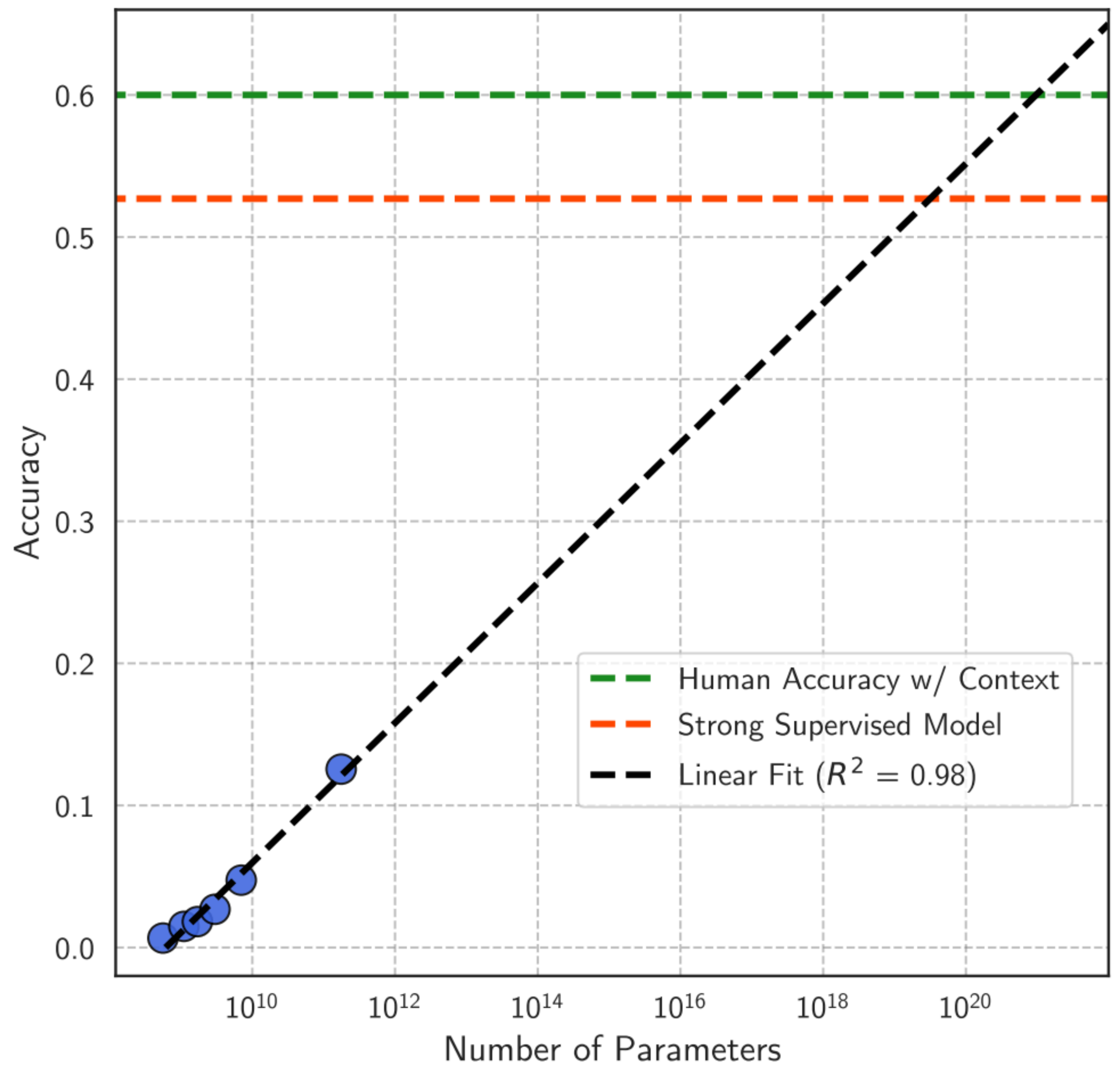
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Massive scaling for better performance



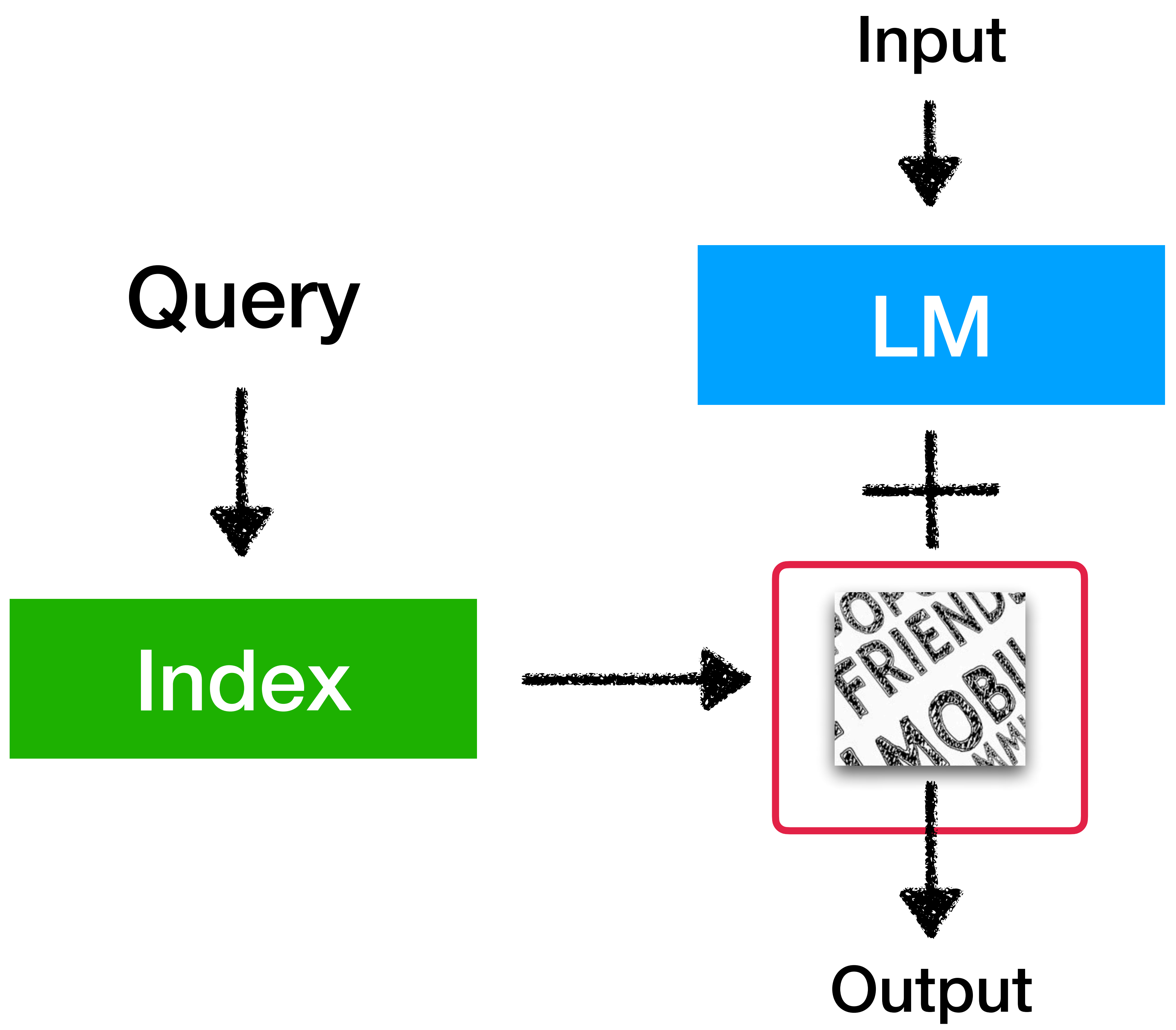
Kandpal et al., Large Language Models Struggle to Learn Long-Tail Knowledge. ICML 2023.

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

Retrieved text can be used as attributions



How do retrieval-augmented LMs address them?

Hallucinations

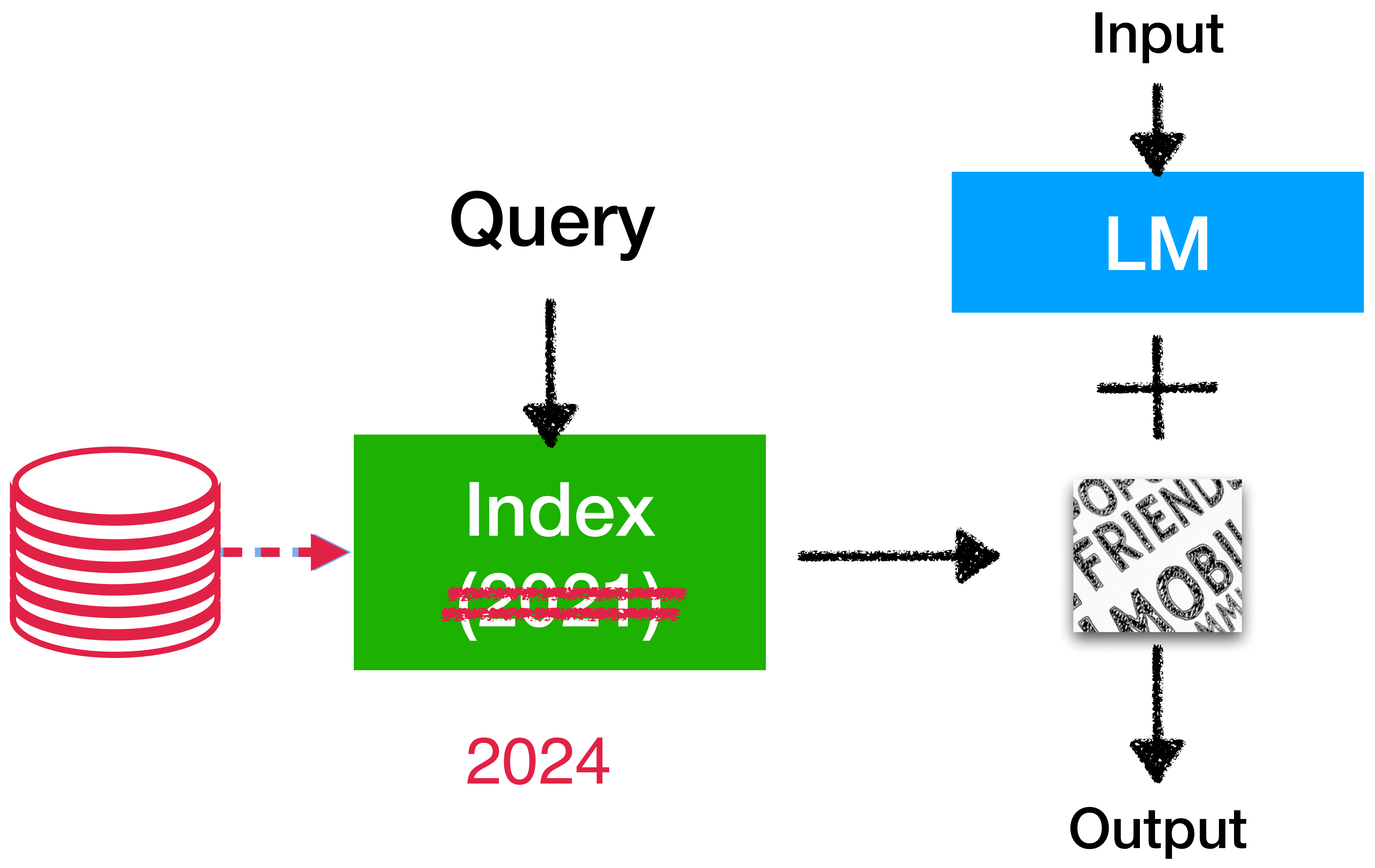
Lack of attributions

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

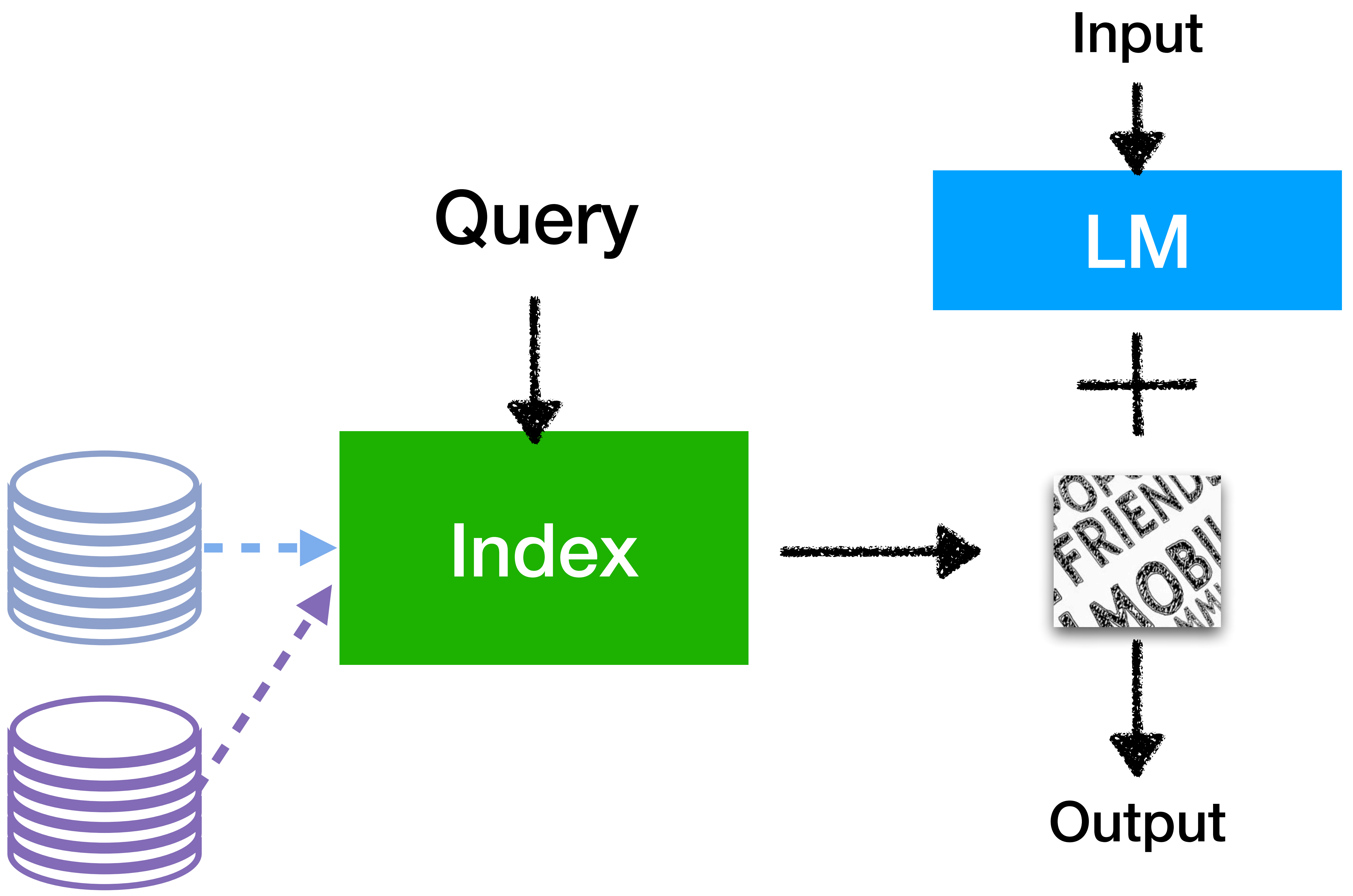
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Adding new domain corpora for domain adaptations



How do retrieval-augmented LMs address them?

Hallucinations

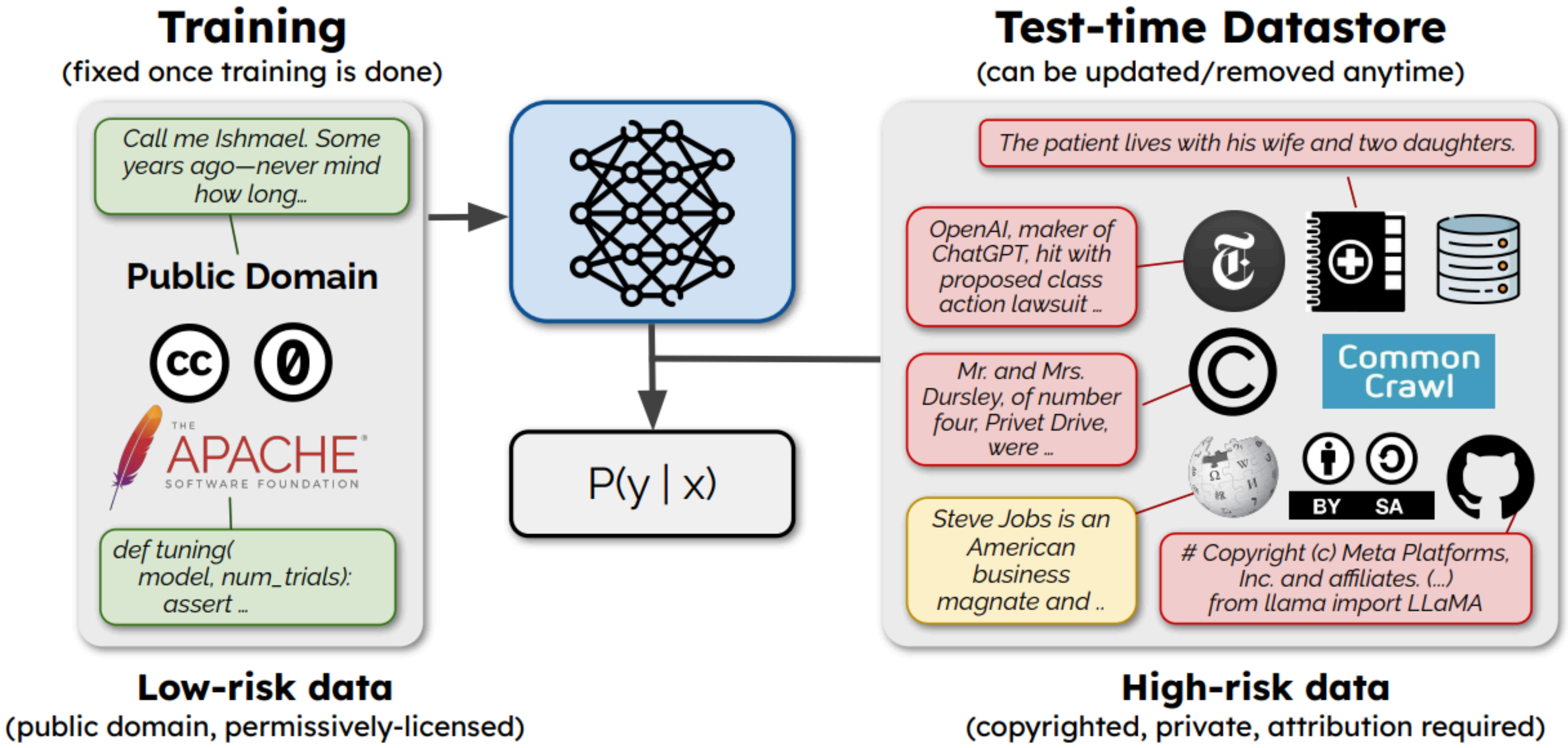
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

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

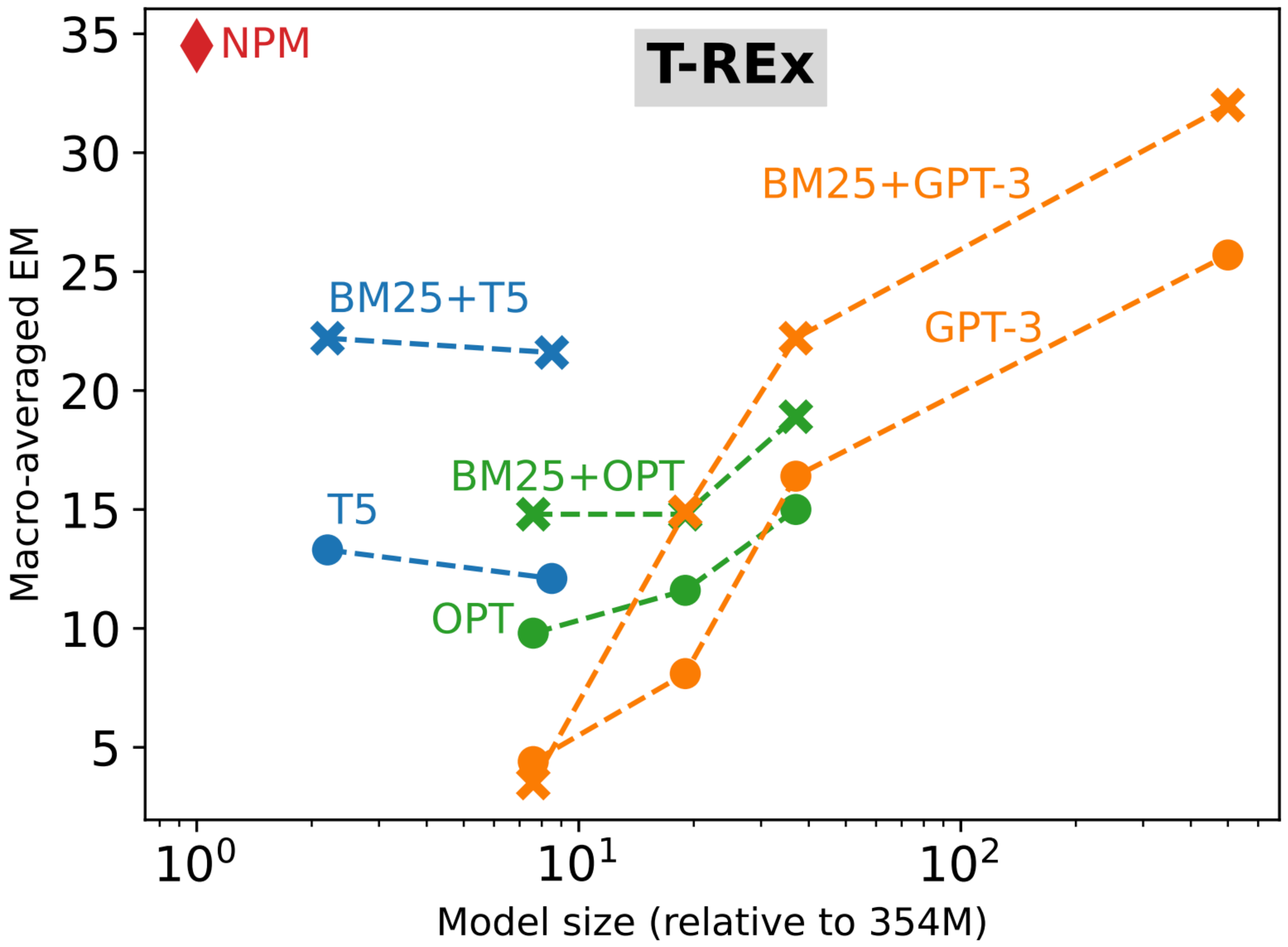
Lack of attributions

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 limitations of retrieval-augmented LMs

- Parametric LMs have numerous challenges to build reliable systems.
- Retrieval-augmented LMs such as RAG can effectively address them.

When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (ACL 2023)
https://arxiv.org/abs/2403.03187

Reliable, adaptable and attributable LMs with retrieval (Arxiv 2024)
https://arxiv.org/abs/2403.03187

Let’s talk about how we can improve traditional RAG!

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

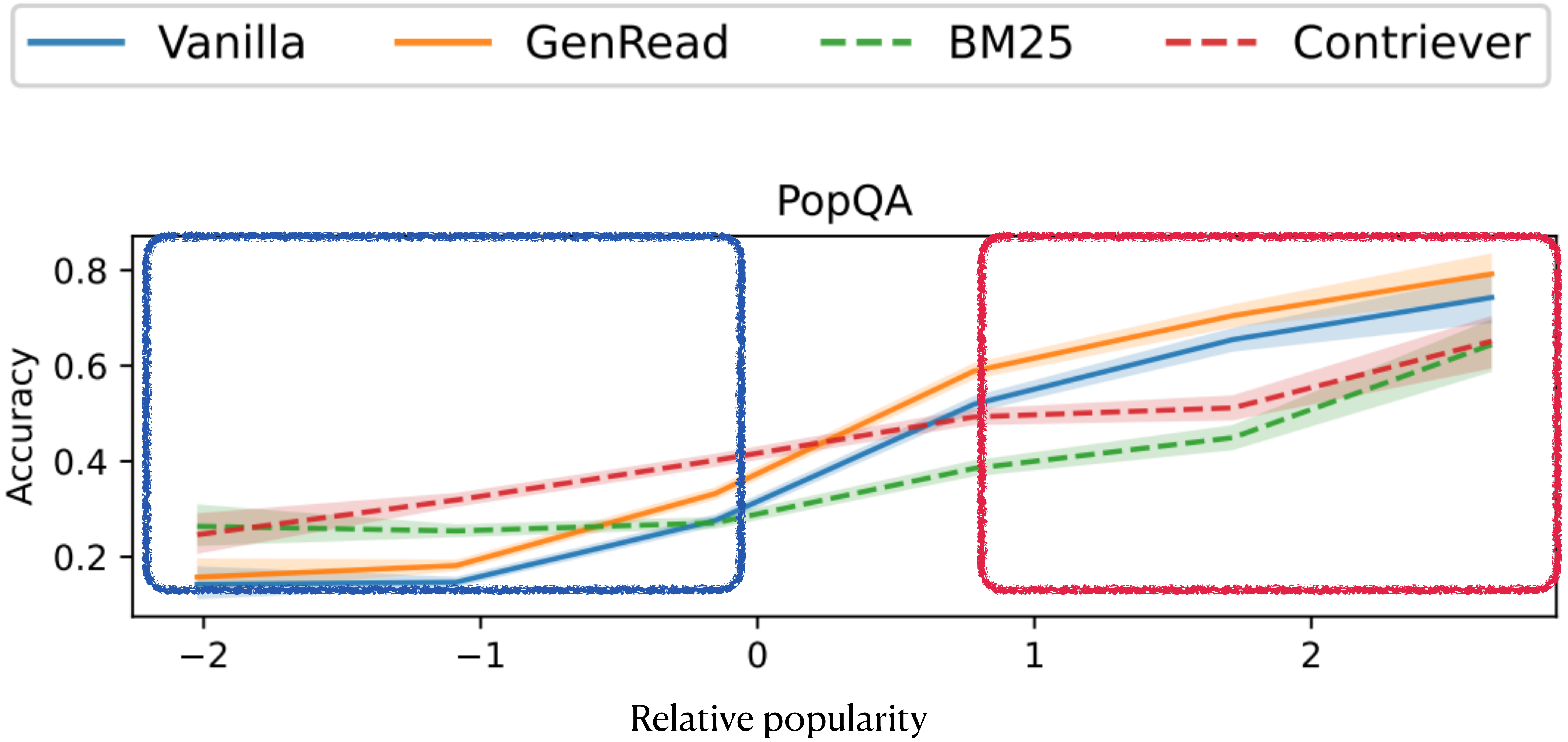
Summary and Future directions: RAG in the wild

RAG hurts in popular knowledge memorized by vanilla LMs

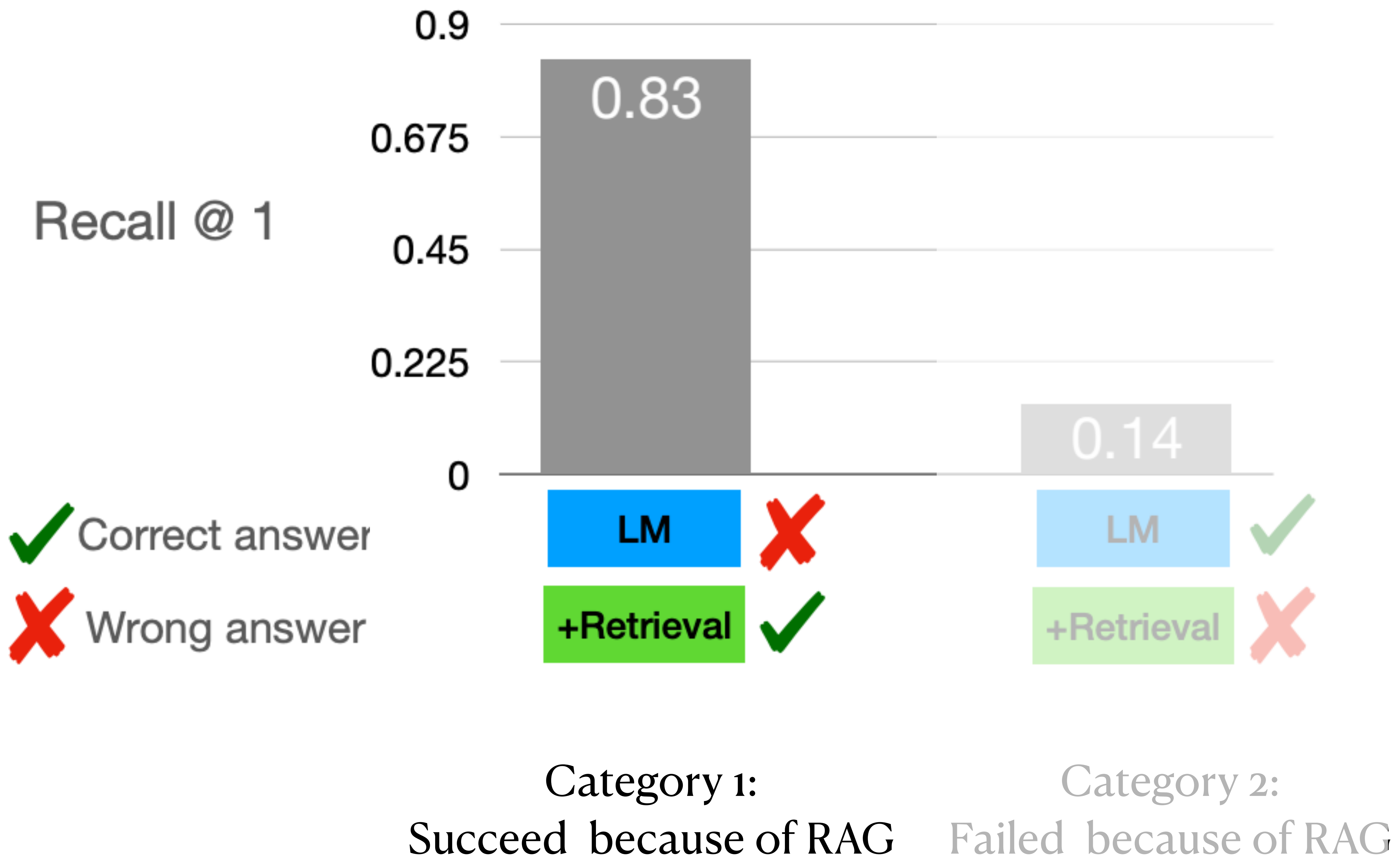
On popular facts vanilla ≈ RAG or vanilla > RAG

Why?

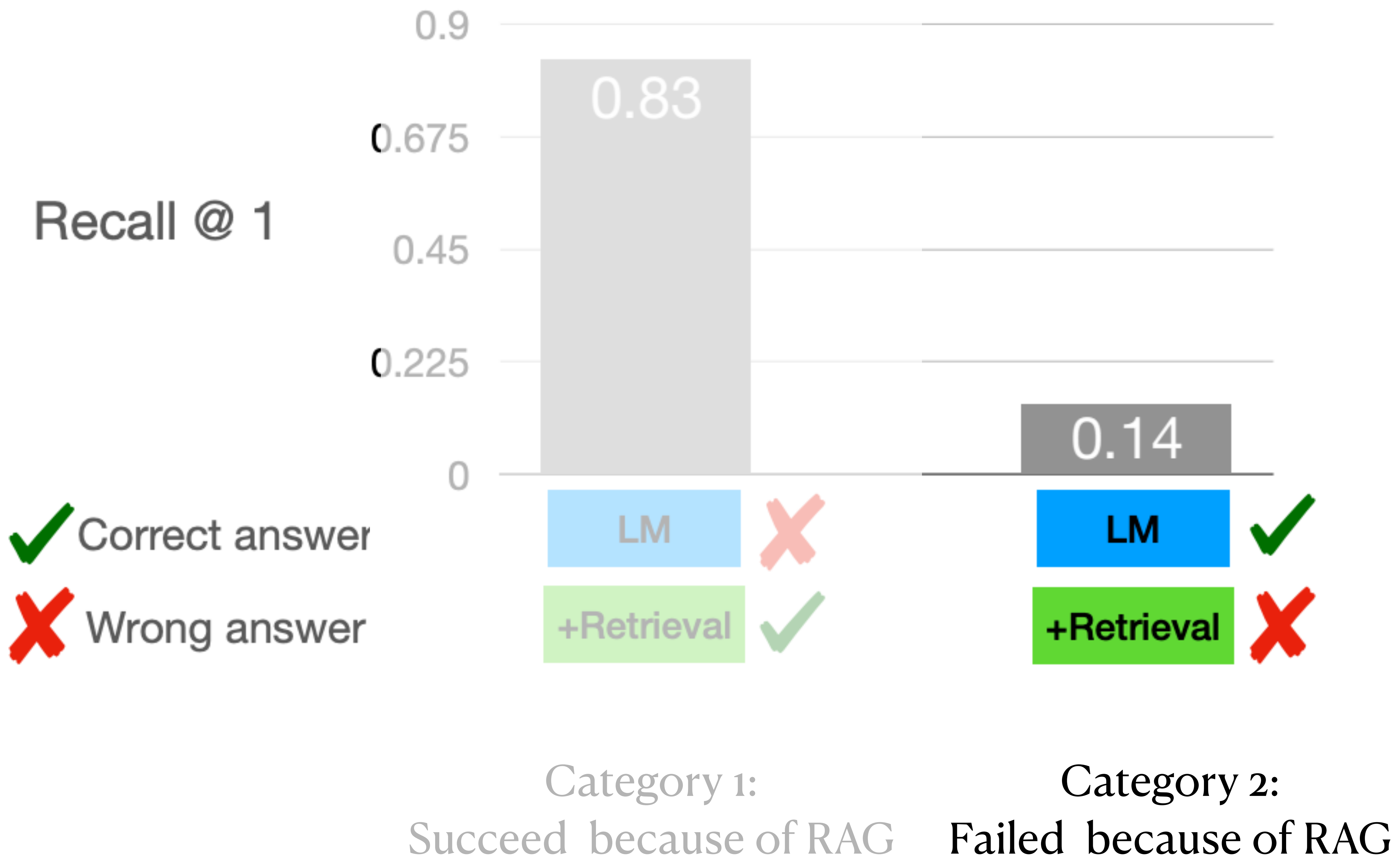
On long-tail facts RAG > vanilla



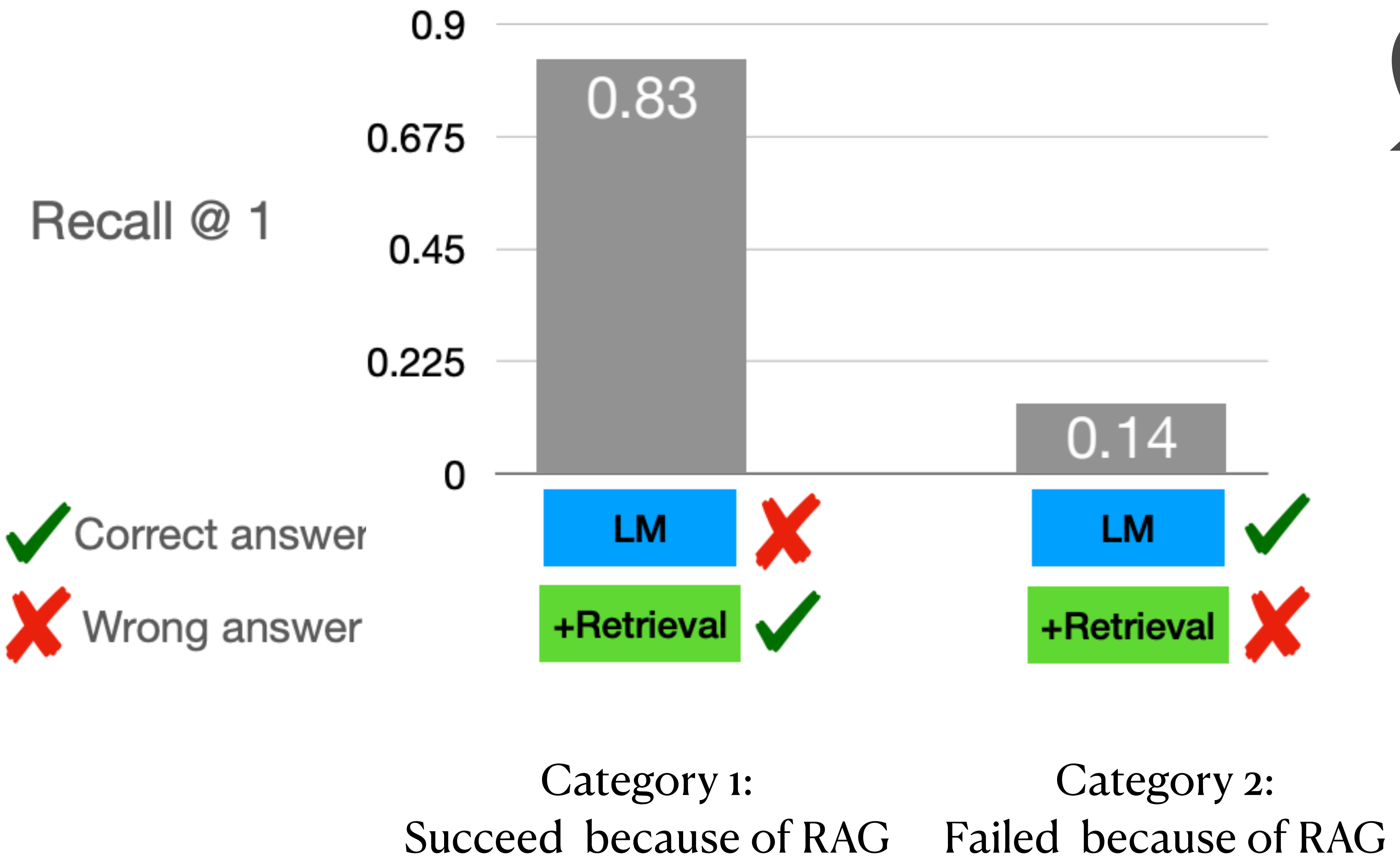
Incorrect retrieval can easily confuse LMs

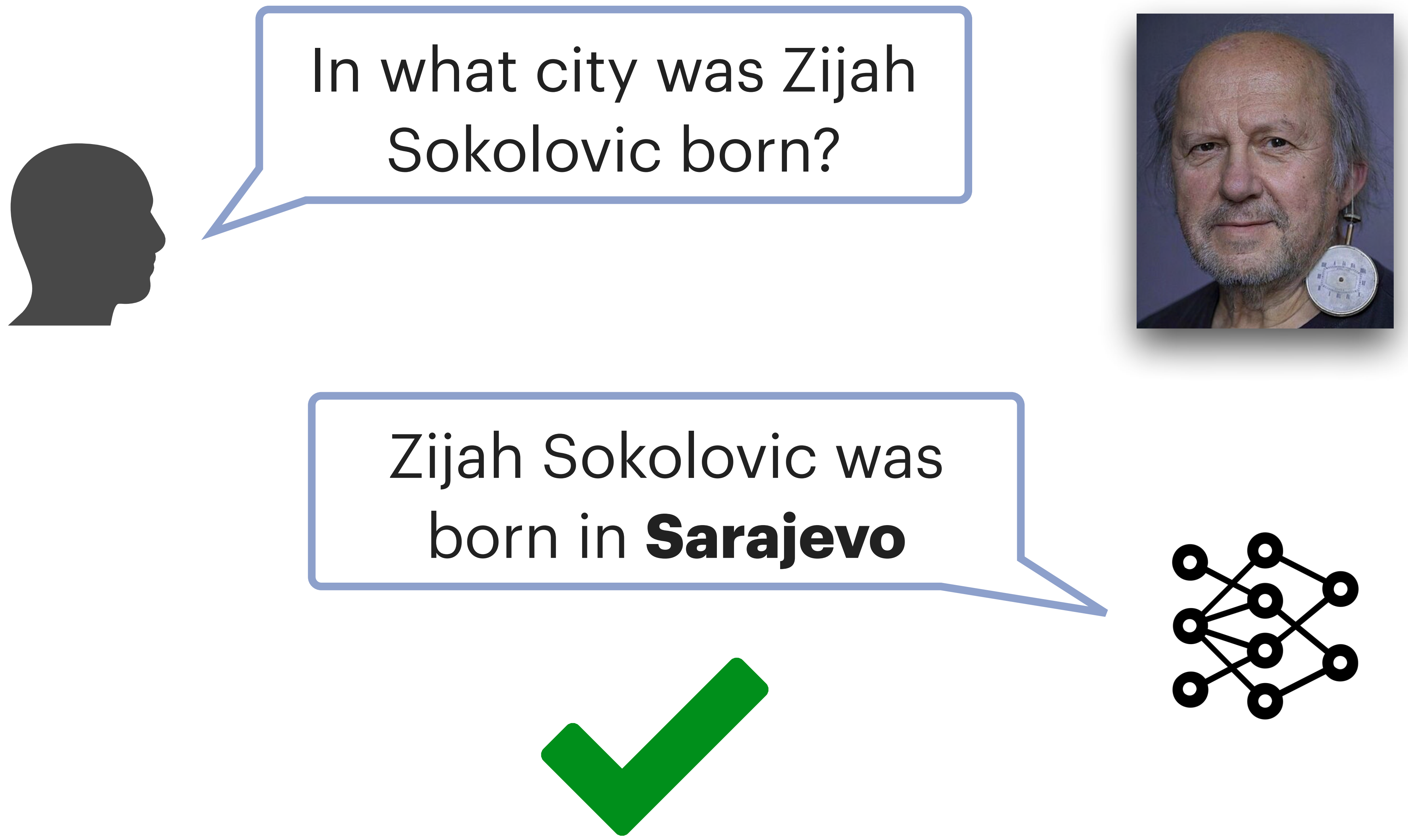


Incorrect retrieval can easily confuse LMs



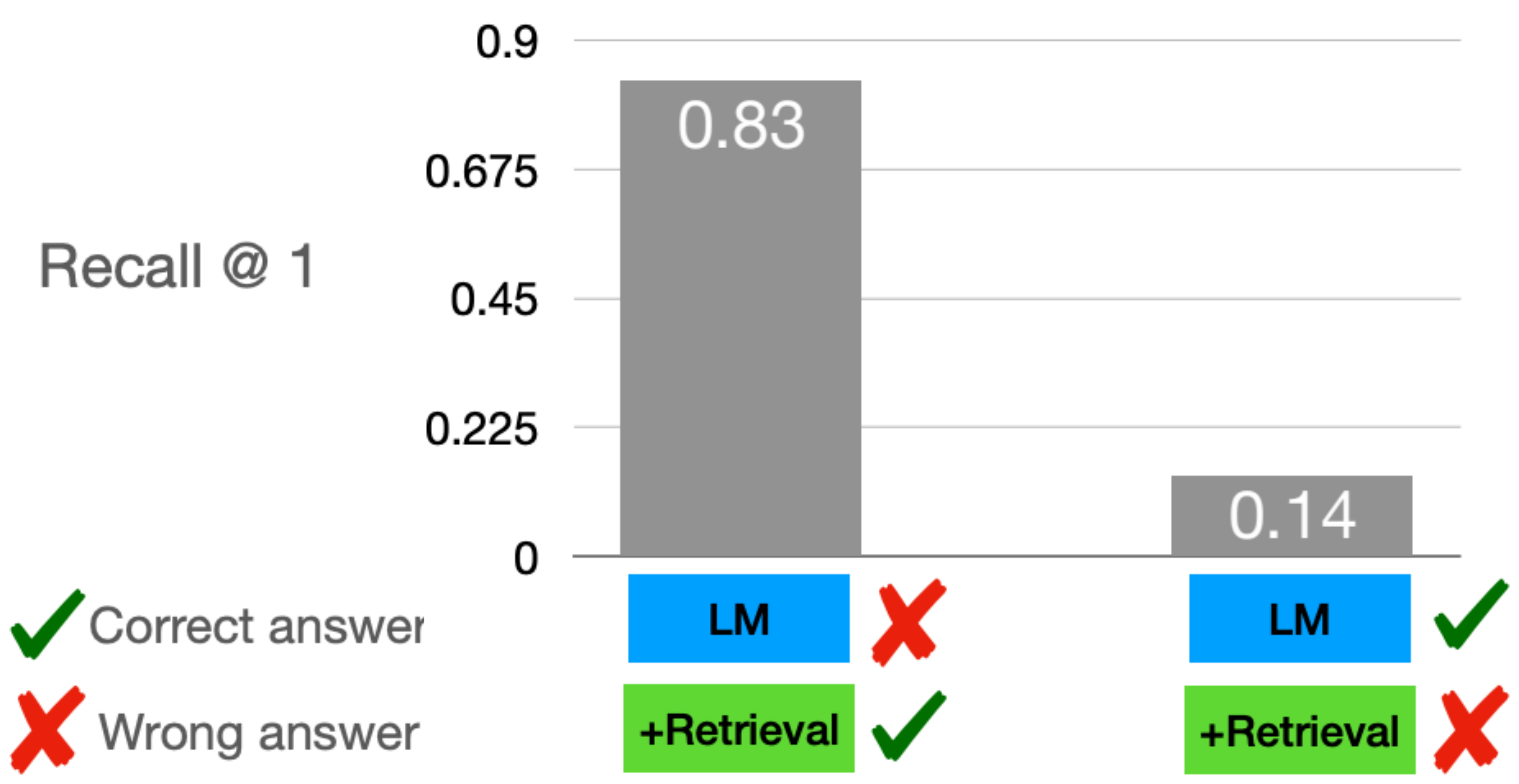
Incorrect retrieval can easily confuse LMs





Incorrect retrieval can easily confuse LMs

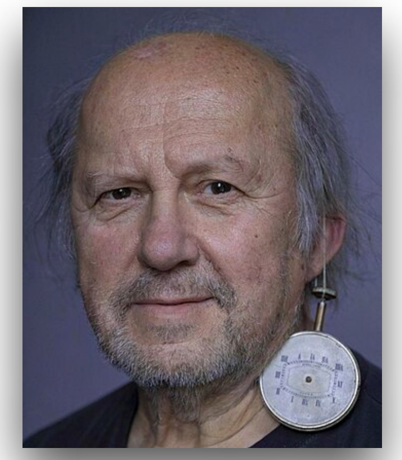
Failed RAG can counterfactually make LMs answer incorrectly



Category 1:
Succeed because of RAG

Category 2:
Failed because of RAG

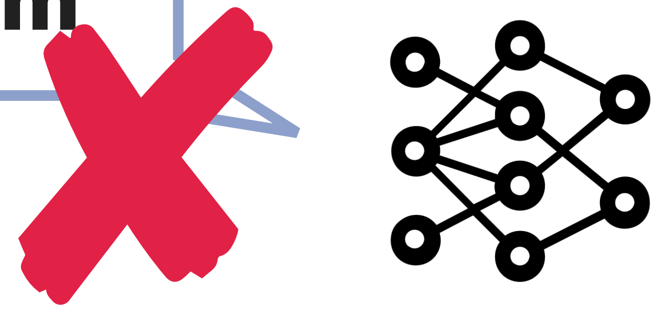
In what city was Zijah Sokolovic born?



Wikipedia
The Free Encyclopedia

Ali Sokol ... (born 8 May 1921 in **Orahovac, Kingdom of Serbs**, Croats and Slovenes, died 23 September 1974) was a Yugoslav pulmonologist

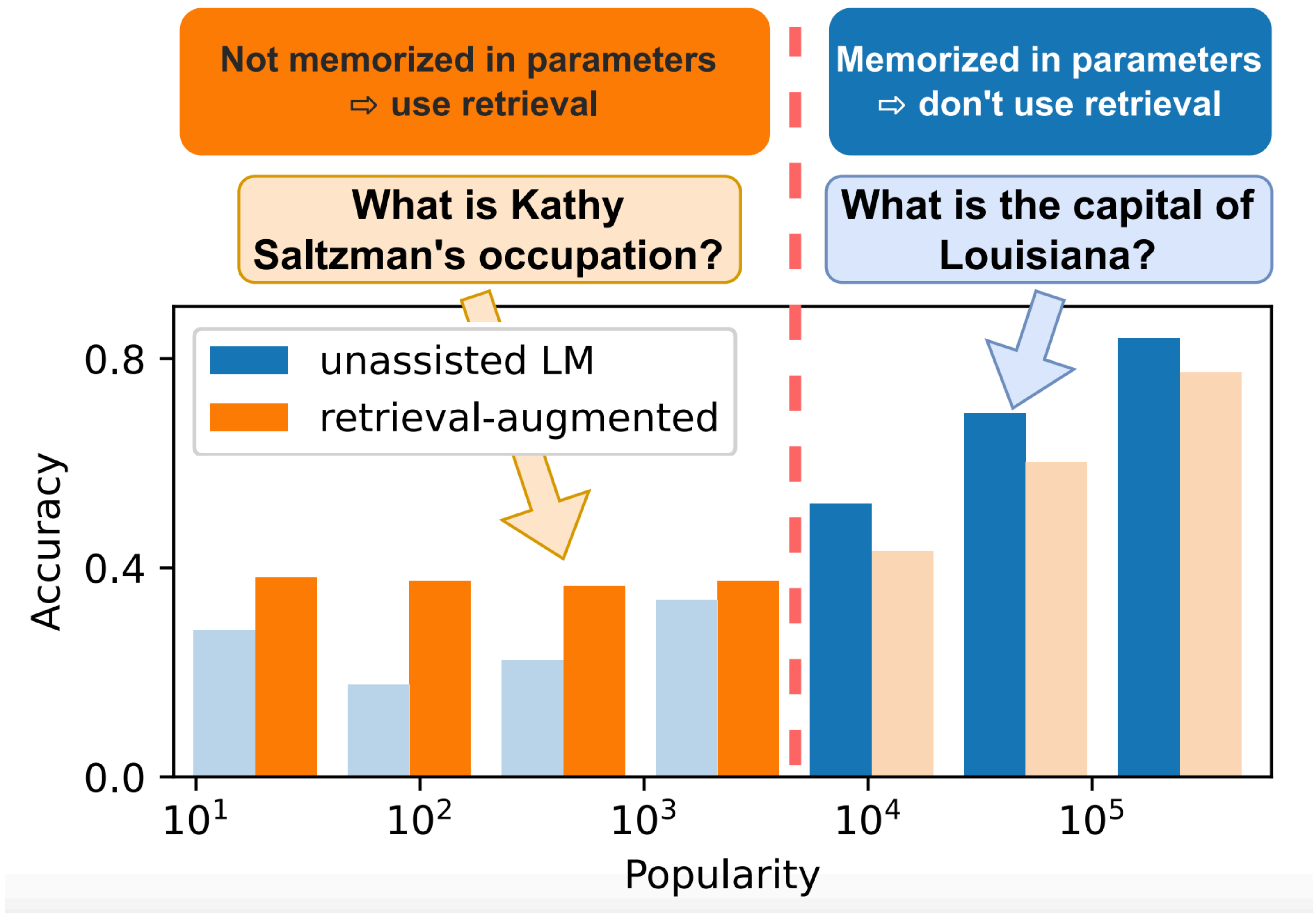
Zijah Sokolovic´ was born in **Orahovac, Kingdom**

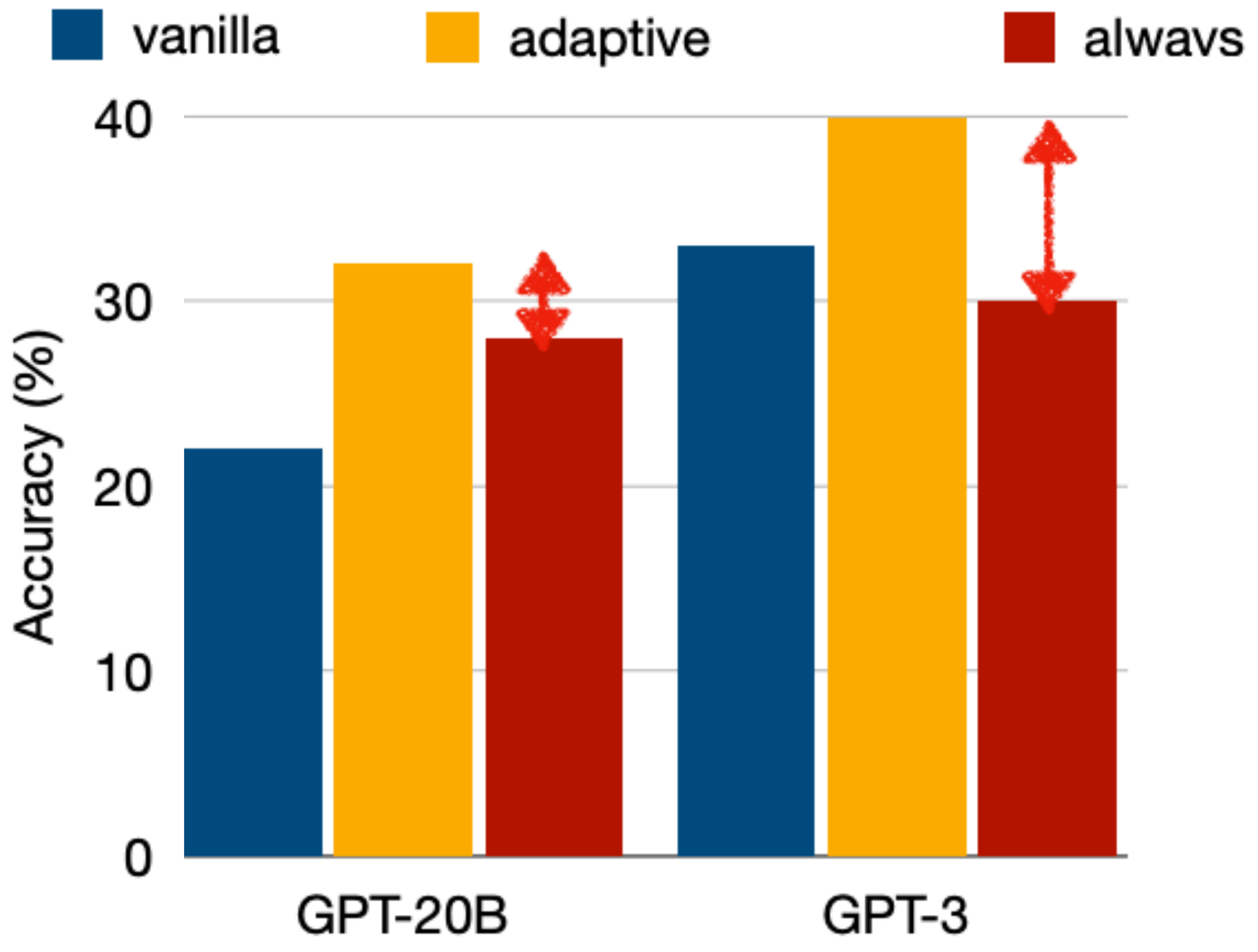


A simple hack: a threshold-based adaptive retrieval

Simple adaptive RAG significantly improve RAG performance & efficiency

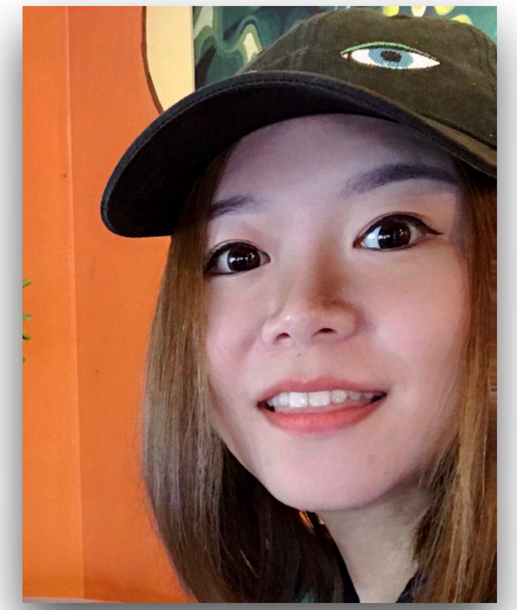
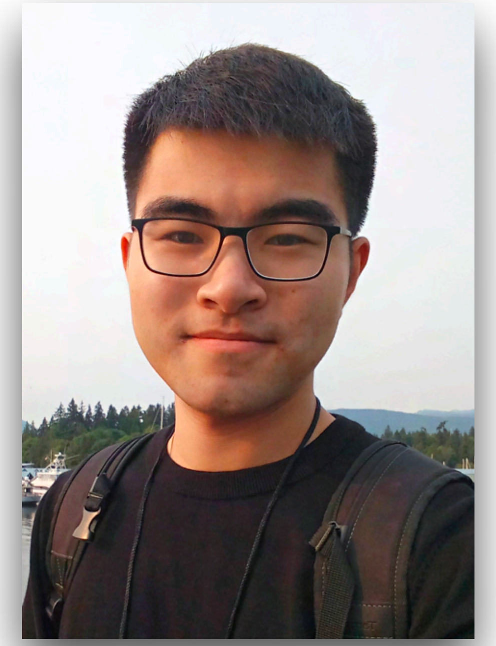
Popularity threshold learned on dev set





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

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

ICLR 2024 (Oral — top 1%)
Best Paper Honorable Mention at NeurIPS Instruction workshop

Standard RAG aren't (always) reliable

Step 1: Retrieve K documents

Prompt How did US states get their names?

Retriever

1 Of the fifty states, eleven including New York, Georgia, Washington named after an individual person.

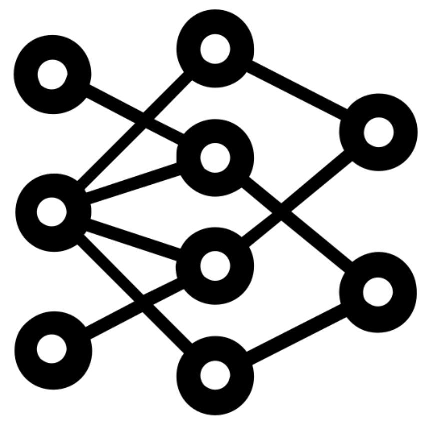
2 UTAH: Name taken from the Ute Indians who inhabited that region

3 The history of human activity in Michigan began with settlement by Paleo-Indians.

Easily distracted by unhelpful documents

Step 2: Prompt LM with K docs and generate

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



LM

US states got their names from a variety of sources. **Eleven states such as New York and Michigan are named after an individual person.** Some states including Utah and Washington, are named after native American tribe names.

Standard RAG aren't (always) reliable

Step 1: Retrieve K documents

Prompt How did US states get their names?

Retriever

❶ Of the fifty states, eleven including New York, Georgia, Washington named after an individual person.

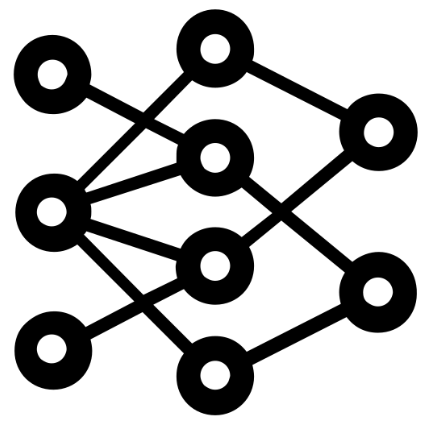
❷ UTAH: Name taken from the Ute Indians who inhabited that region

❸ The history of human activity in Michigan began with settlement by Paleo-Indians.

LMs can also ignore retrieved context

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 such as New York and Michigan are named after an individual person. **Some states including Utah and Washington**, are named after native American tribe names.

Standard RAG aren’t (always) necessary

Step 1: Retrieve K documents

Write an essay of your best summer vacation

Retriever

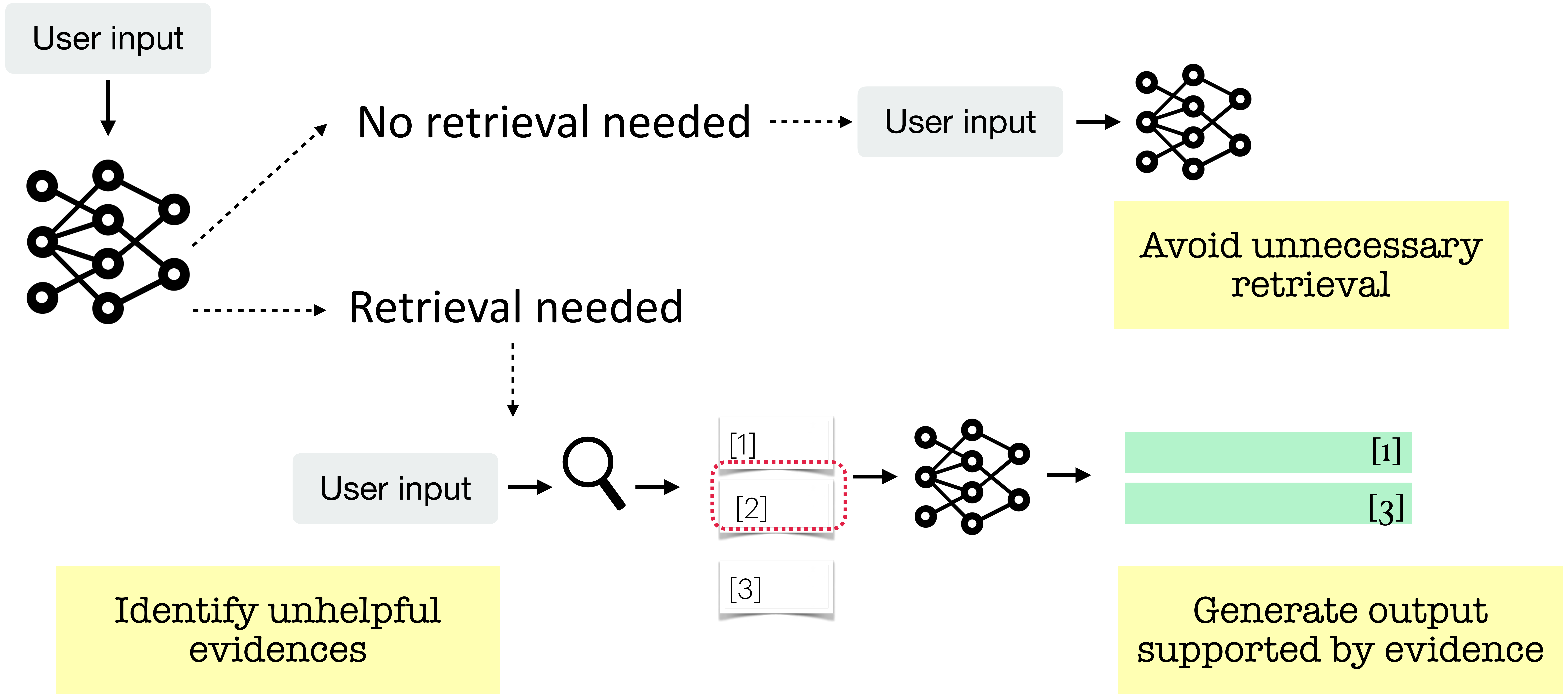
1. The term summer vacation or summer break refers to a school break in the summer between school years.

2. Summer Vacation (Korean: 여름방학) is a South Korean “home vacation” reality show

3. Summer Vacations a 2023 Spanish comedy film directed by Santiago Segura which stars Segura, Leo Harlem, Cristina Gallego, and Patricia Conde.

Always retrieving fixed number of documents is inefficient & harmful

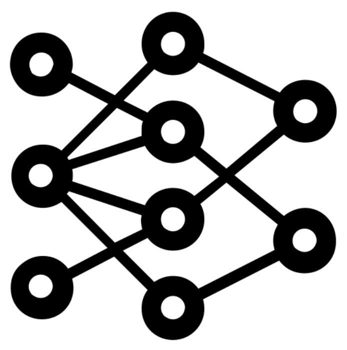
Goal of Self-RAG: learn to retrieve, generate and critique



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

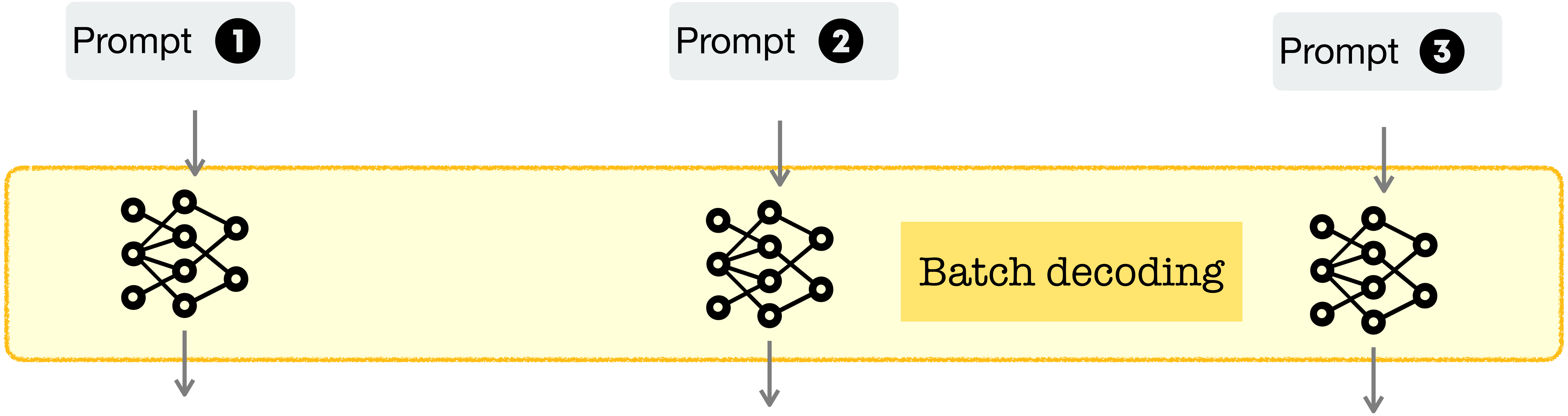
❶ Of the fifty states, eleven including New York, Georgia, Washington named after an individual person.

❷ The history of human activity in Michigan began with settlement by Paleo-Indians.

❸ UTAH: Name taken from the Ute Indians who inhabited that region

Self-RAG – Self-Reflective Retrieval-Augmented Generation

Step 2: Generate segments in *parallel*



Relevant 11 of 50 state names come from persons.

Irrelevant Many states including Michigan are named after an individual person.

Relevant Several states such as Utah and Alabama are named after native American tribes.

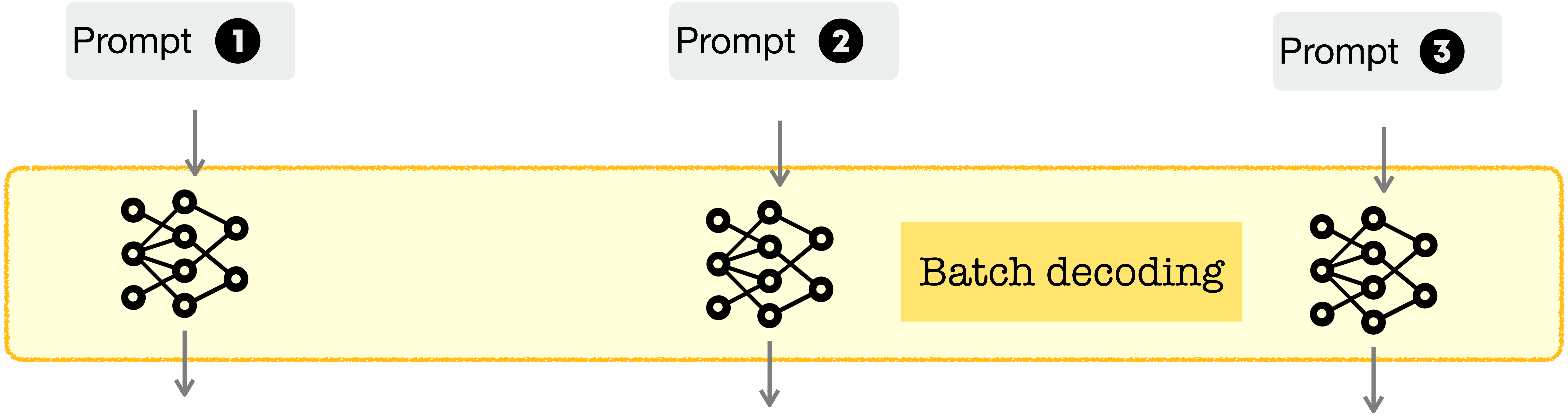
1 Of the fifty states, eleven states named after an individual person.

2 The history of human activity in Michigan began with settlement by Paleo-Indians.

3 UTAH: Name taken from the Ute Indians who inhabited that region

Self-RAG – Self-Reflective Retrieval-Augmented Generation

Step 2: Generate segments in *parallel*



Relevant 11 of 50 state names come from persons. Supported

Irrelevant Many states including Michigan are named after an individual person.

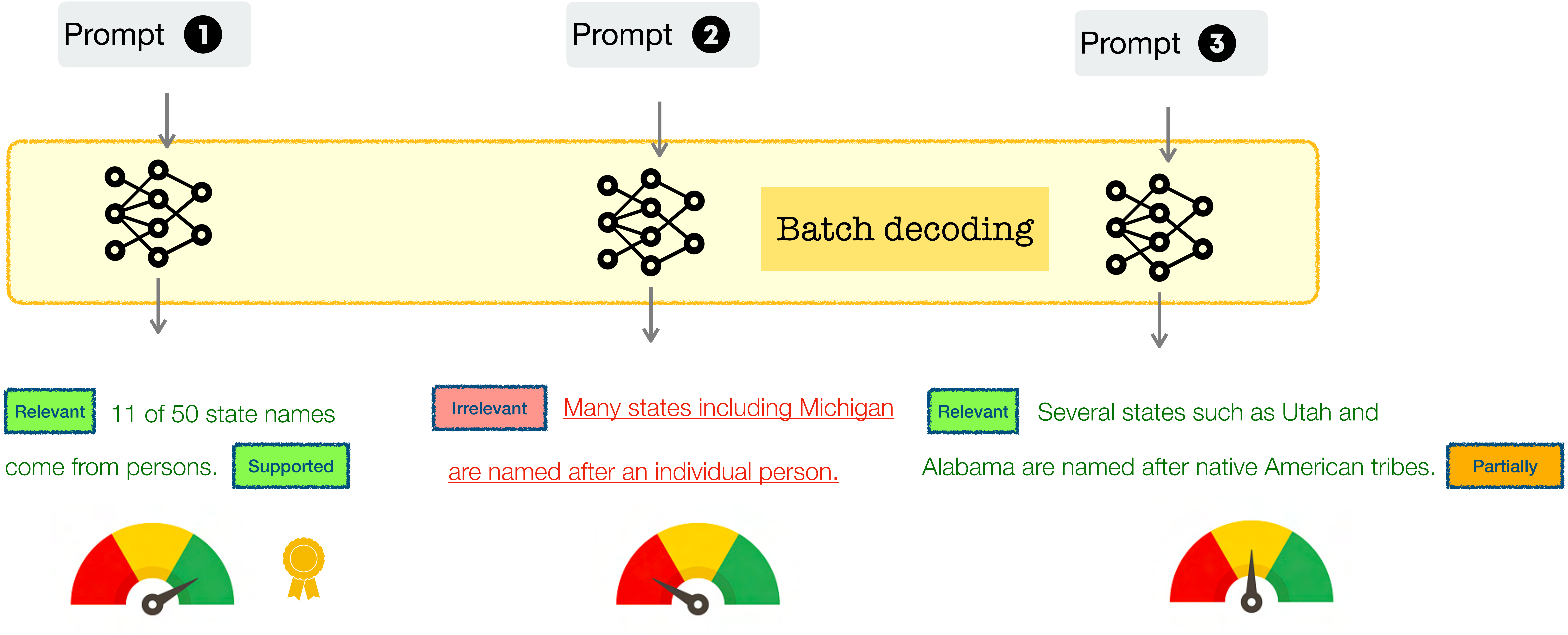
Relevant Several states such as Utah and Alabama are named after native American tribes. Partially

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

3 UTAH: Name taken from the Ute Indians who inhabited that region

Self-RAG – Self-Reflective Retrieval-Augmented Generation

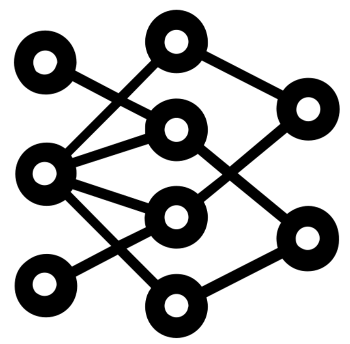
Step 2: Generate segments in *parallel*



Self-RAG – Self-Reflective Retrieval-Augmented Generation

Step 1: Generate with no retrieval

Write an essay of your best summer vacation

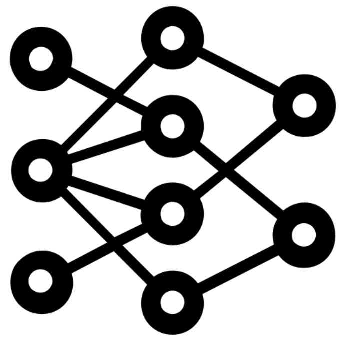


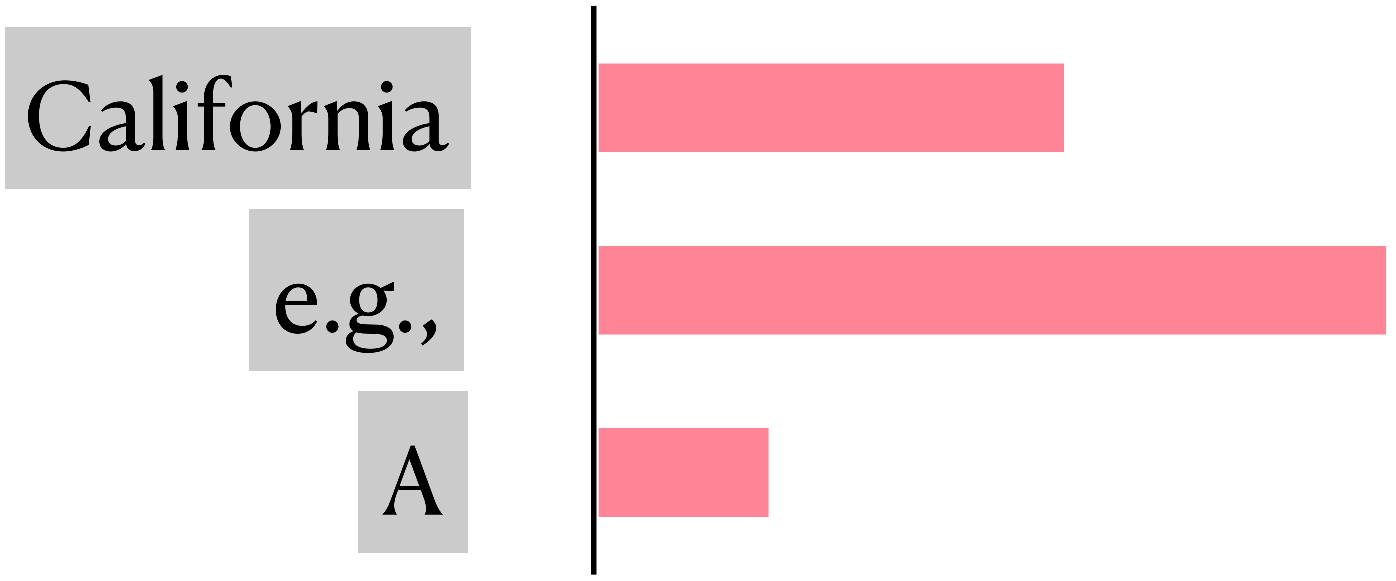
No Retrieval My best summer vacation was a magical escape to the coastal town of Santorini.

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.

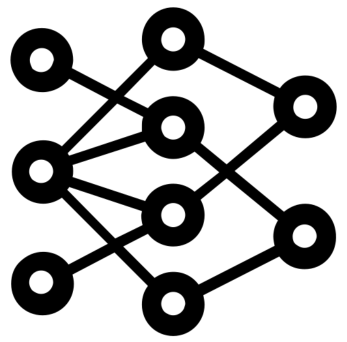




Original LM vocabularies

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

Retrieval tokens

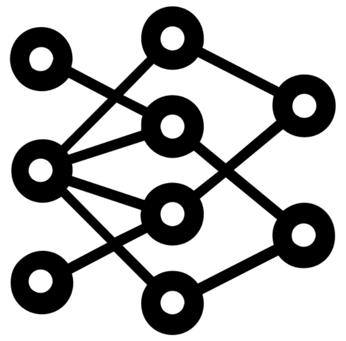
Relevant
Irrelevant
Supported
No support
Useful
Not useful

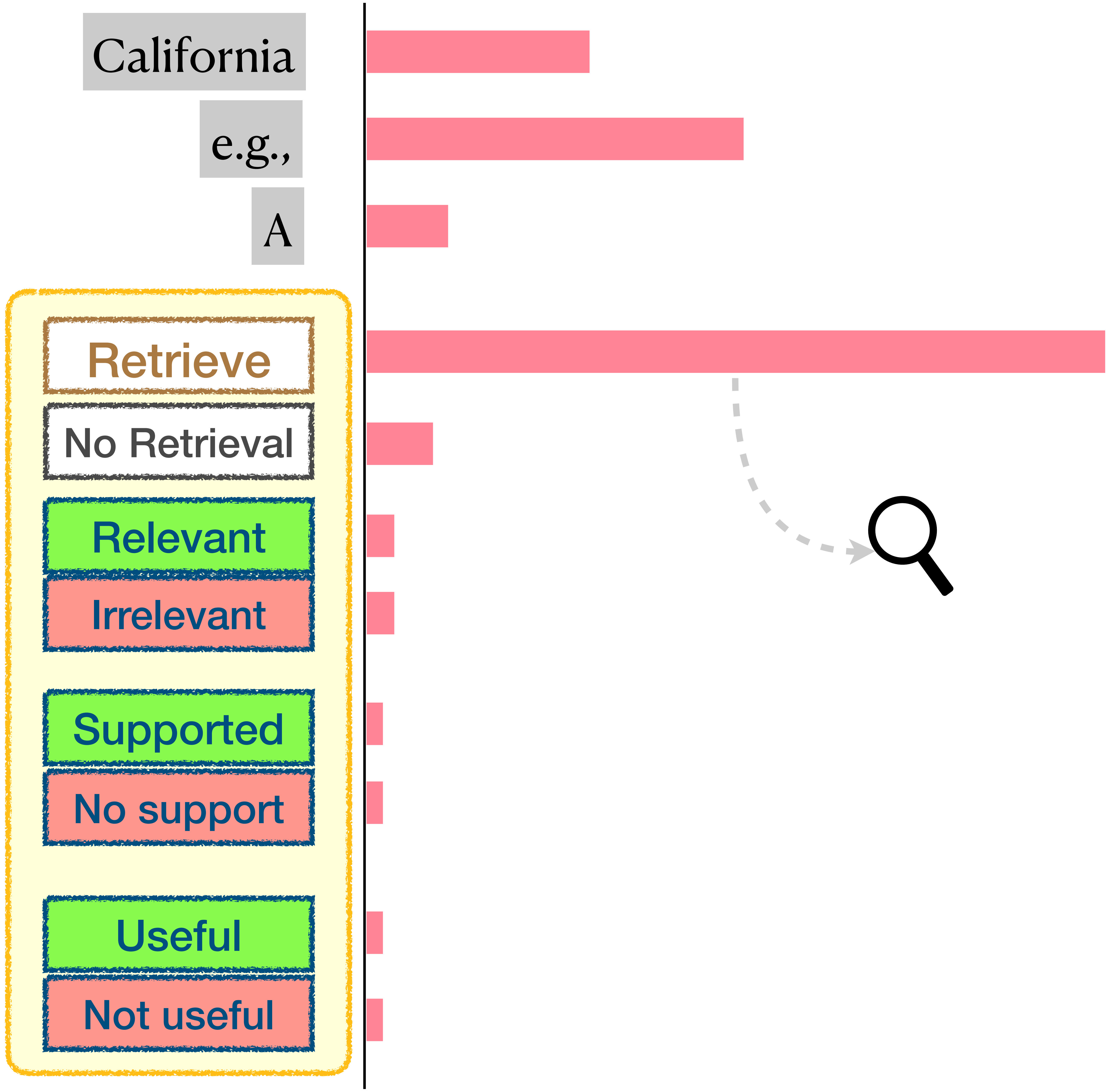
Critique tokens

Vocabulary expanded with reflection tokens

Reflection tokens for retrieval and critique

US states got their names
from a variety of sources.

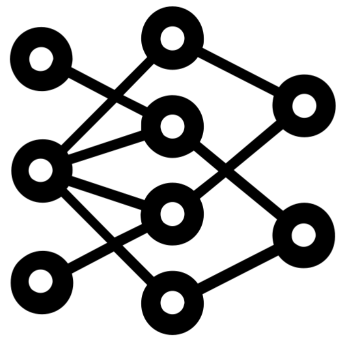




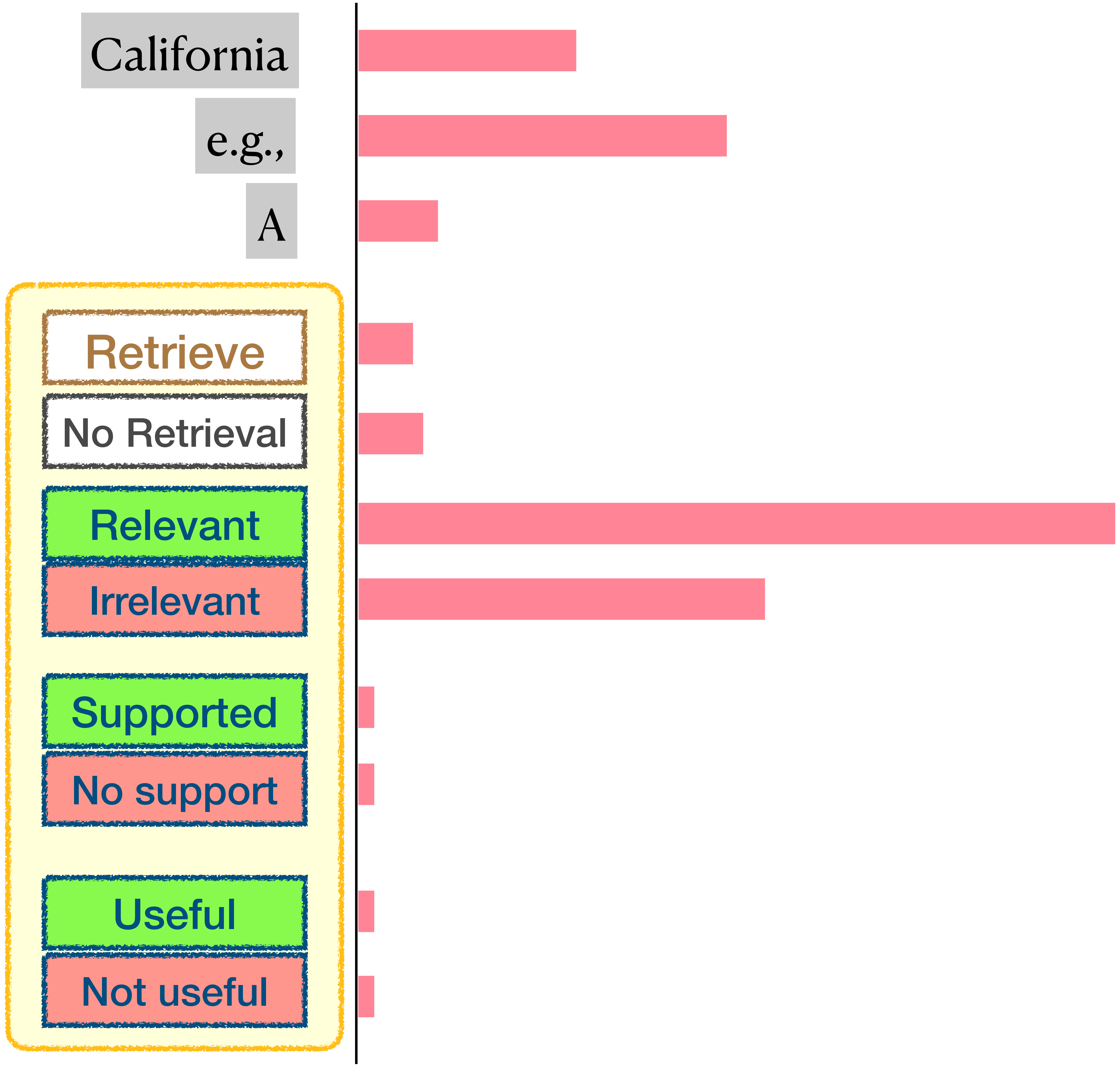
Vocabulary expanded with reflection tokens

Reflection tokens for retrieval and critique

US states got their names from a variety of sources.



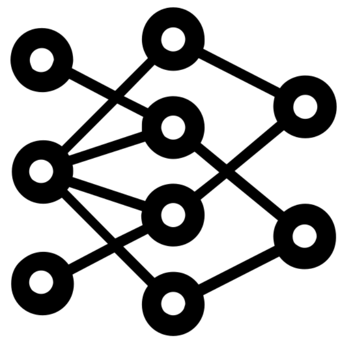
❶ Of the fifty states, eleven are named after an individual person.



Vocabulary expanded with reflection tokens

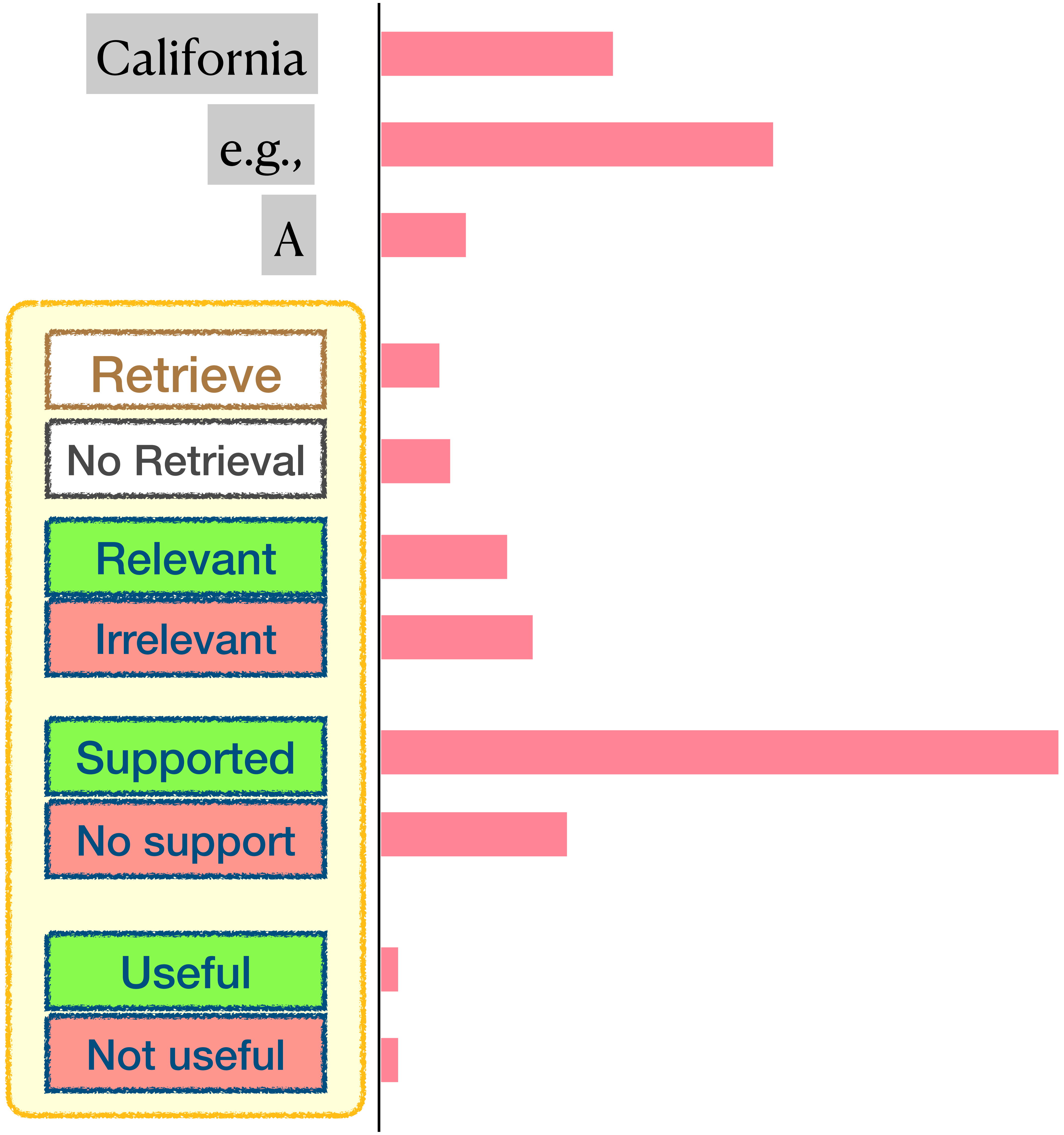
Reflection tokens for retrieval and critique

US states got their names from a variety of sources.



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

11 of 50 state names come from persons.

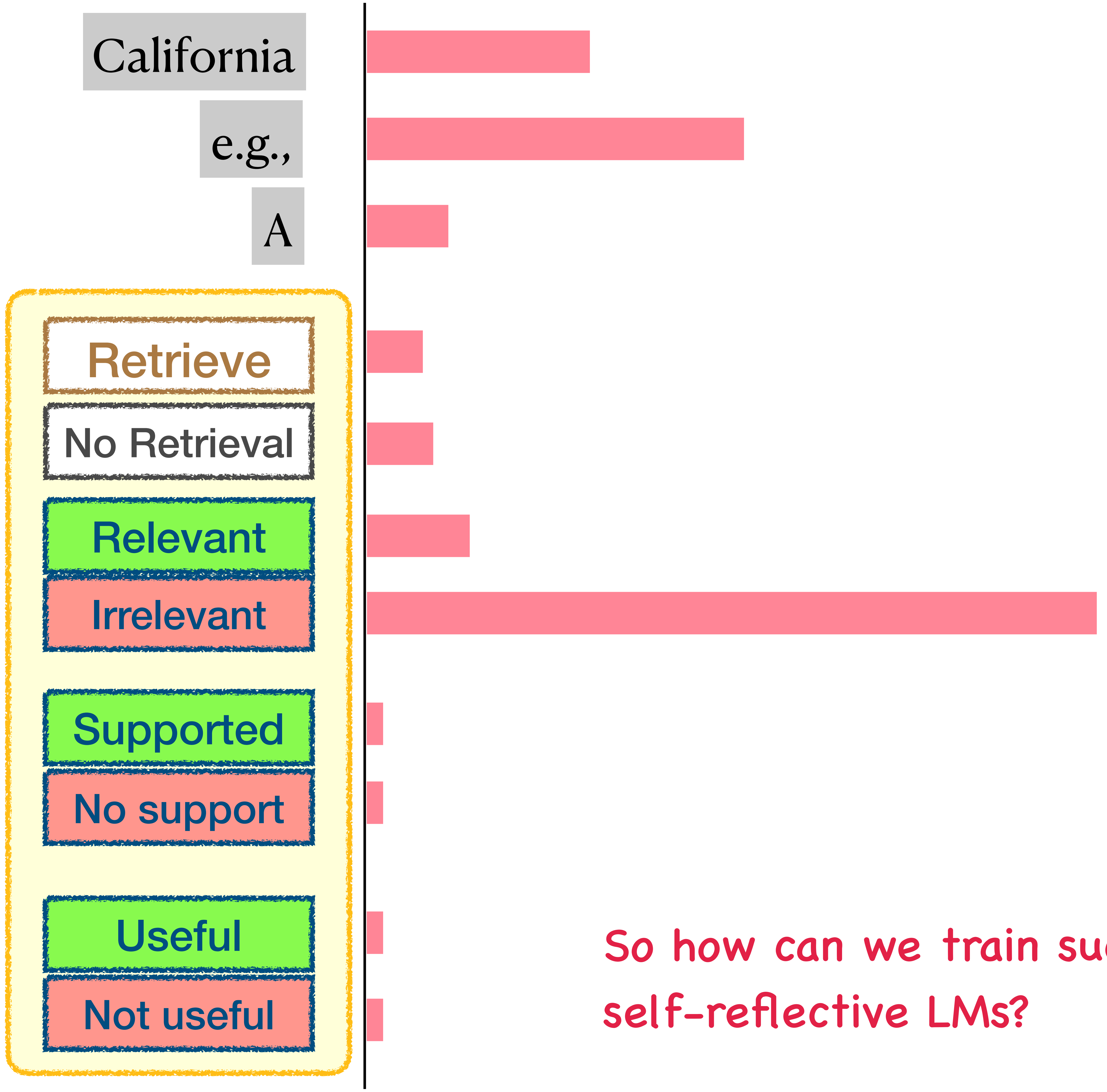


Vocabulary expanded with reflection tokens

Reflection tokens for retrieval and critique

US states got their names from a variety of sources.

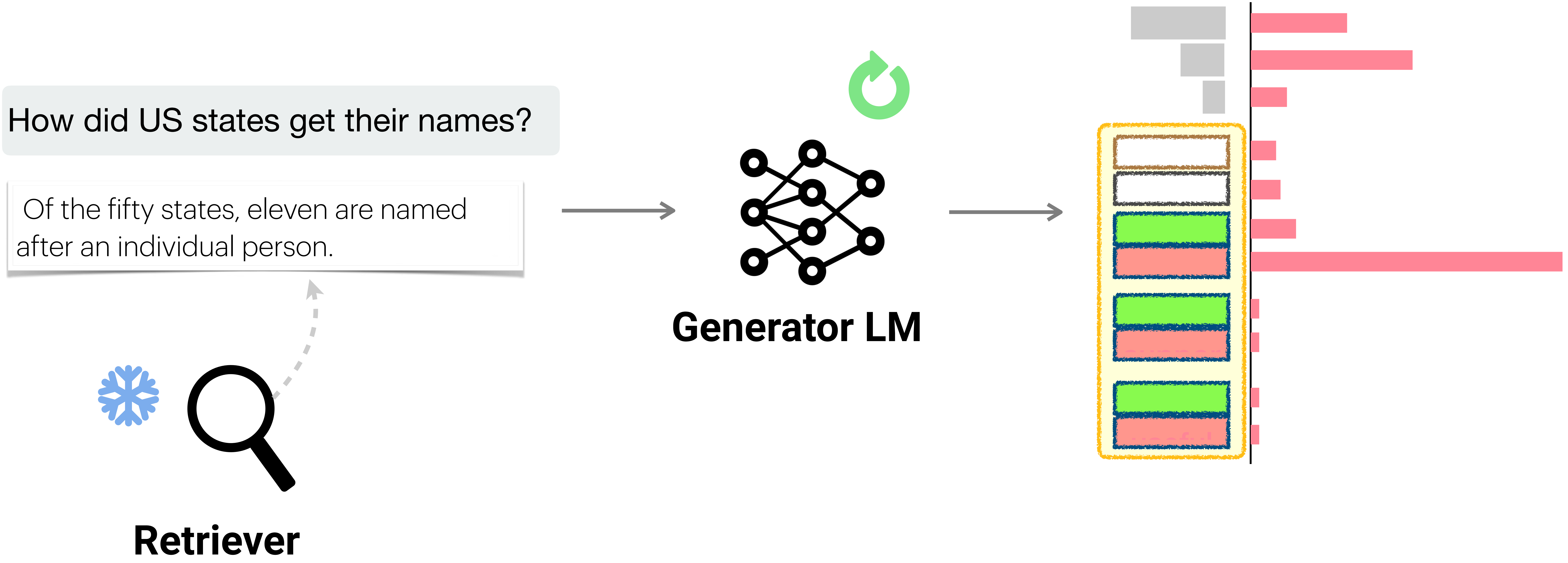
2 The history of human activity in Michigan began with settlement by Paleo-Indians.



Vocabulary expanded with reflection tokens

So how can we train such self-reflective LMs?

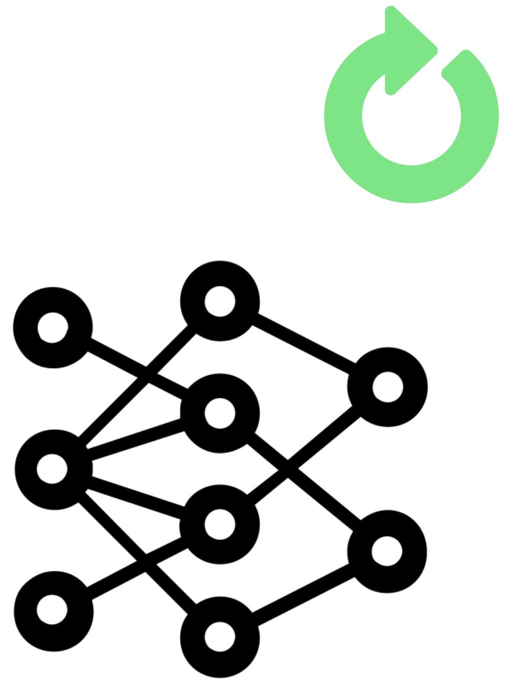
Self-RAG training



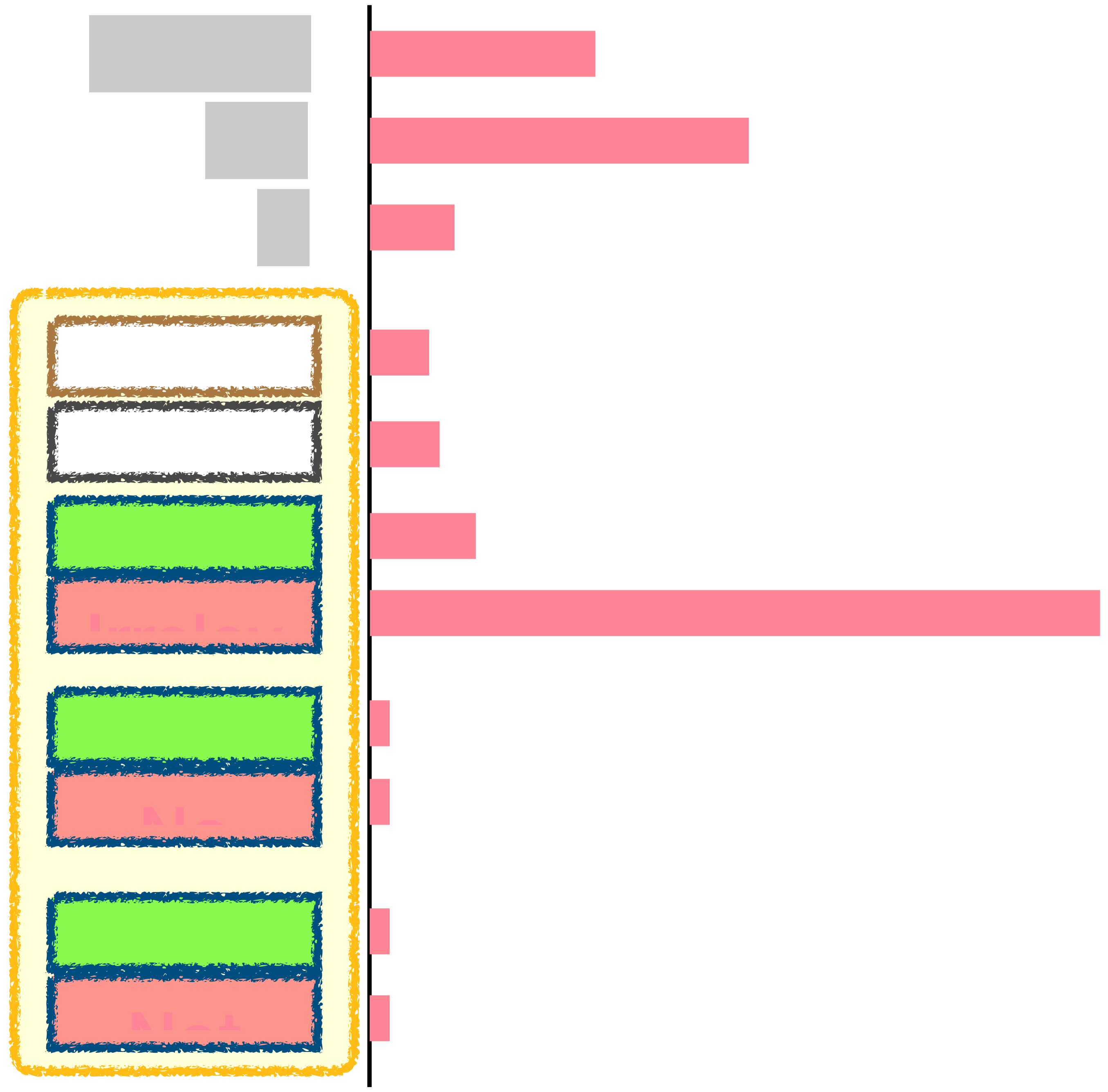
Self-RAG training

How did US states get their names?

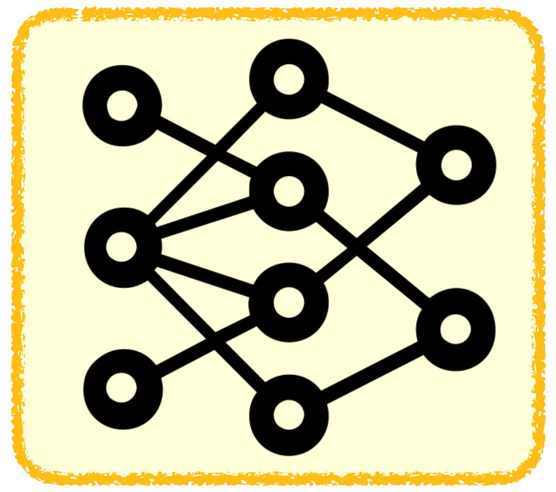
Of the fifty states, eleven are named after an individual person.



Generator LM



Critic LM teaches Generator LM to predict reflection tokens



Training time only Critic LM

Self-RAG training - critic LM

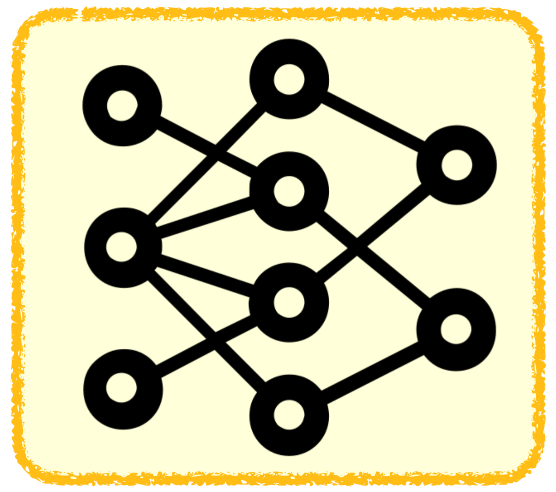
$$\max_{\mathcal{C}} \mathbb{E}_{((x,y,d),r)\sim\mathcal{D}_{critic}} \log p_{\mathcal{C}}(r \mid x, y, d)$$

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

85-90% acc. on validation set

GPT-4

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

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

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.

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: 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

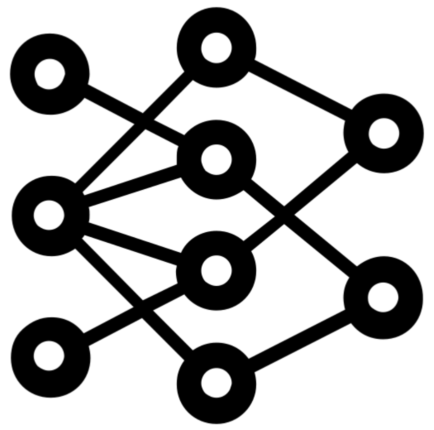
Augmented Output: Retrieve 1 <p>Of the fifty states, eleven are named after an individual person</p>. Relevant 11 of 50 states' names come from person. Supported Retrieve 2 <p>LOUISIANA: Named in honor of Louis XIV of France.</p>. Relevant For instance, Louisiana was named after King Louis XIV, and Georgia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

Self-RAG training - generator LM

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

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

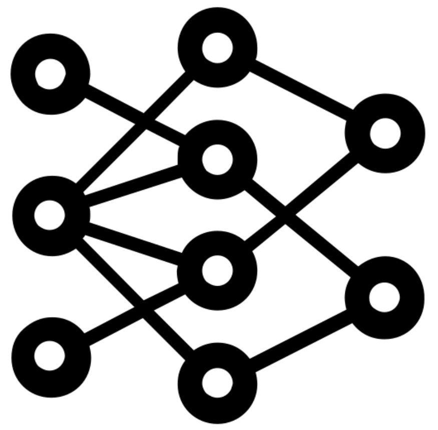
Train with a standard next token objective with expanded vocabulary

Self-RAG training - generator LM

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

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 1

Prompt 2

Prompt 3

$f($ Relevant Supported $)$

0.9

$f($ Irrelevant $)$

0.1

$f($ Relevant Partially $)$

0.4

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

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

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

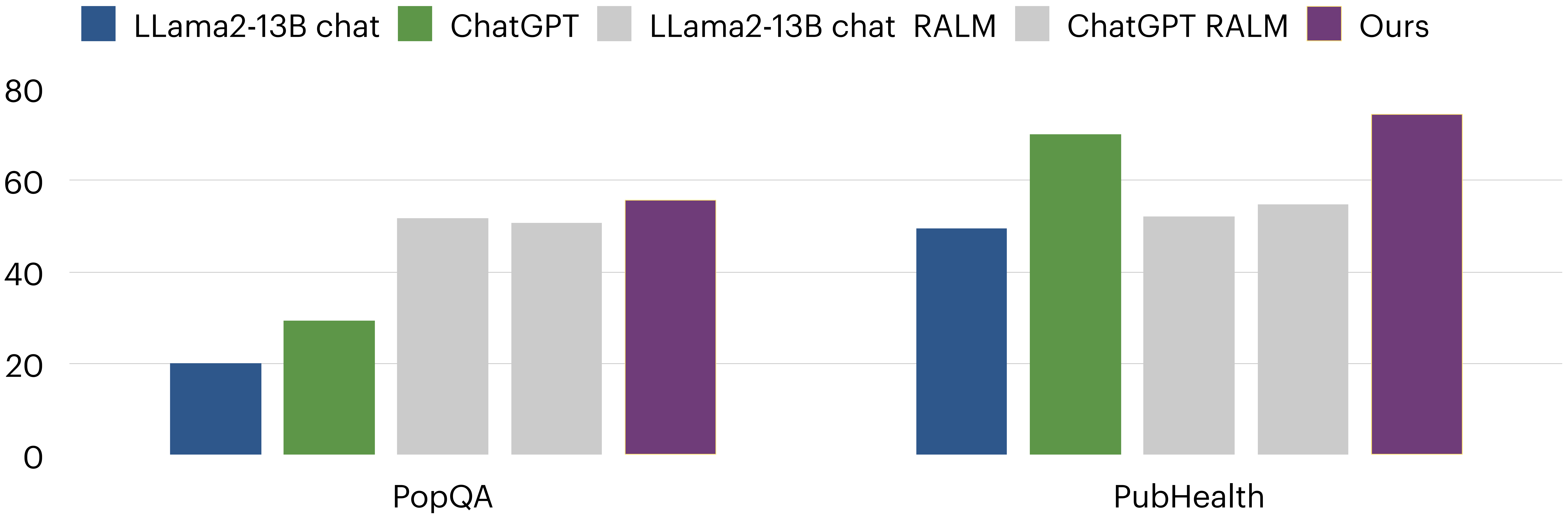
More details of training & test are in our paper!

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

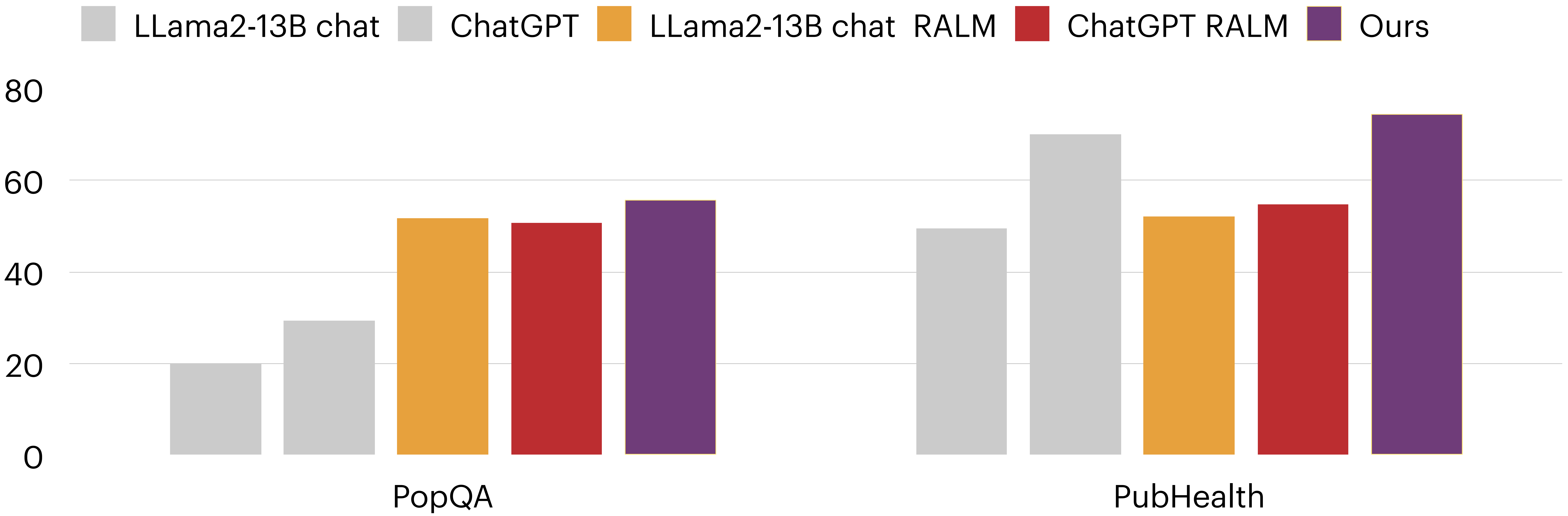
Experimental results (short-form & closed)

Self-RAG outperforms vanilla LMs incl. ChatGPT



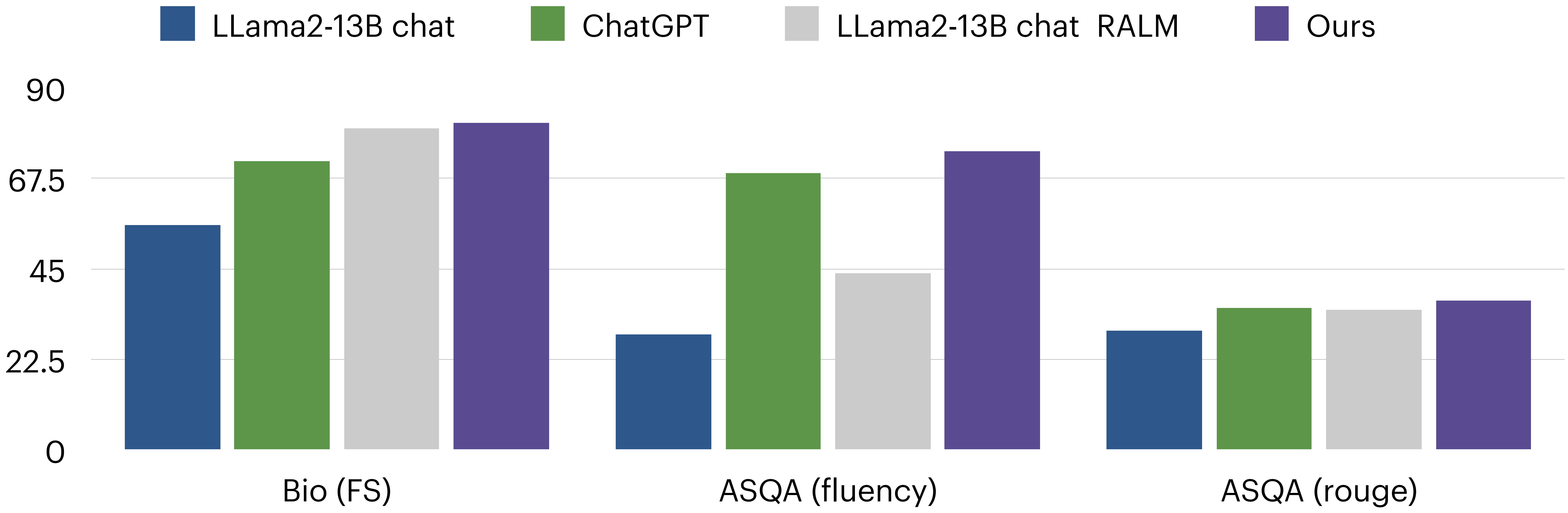
Experimental results (short-form & closed)

Self-RAG outperforms standard RAG + LLMs



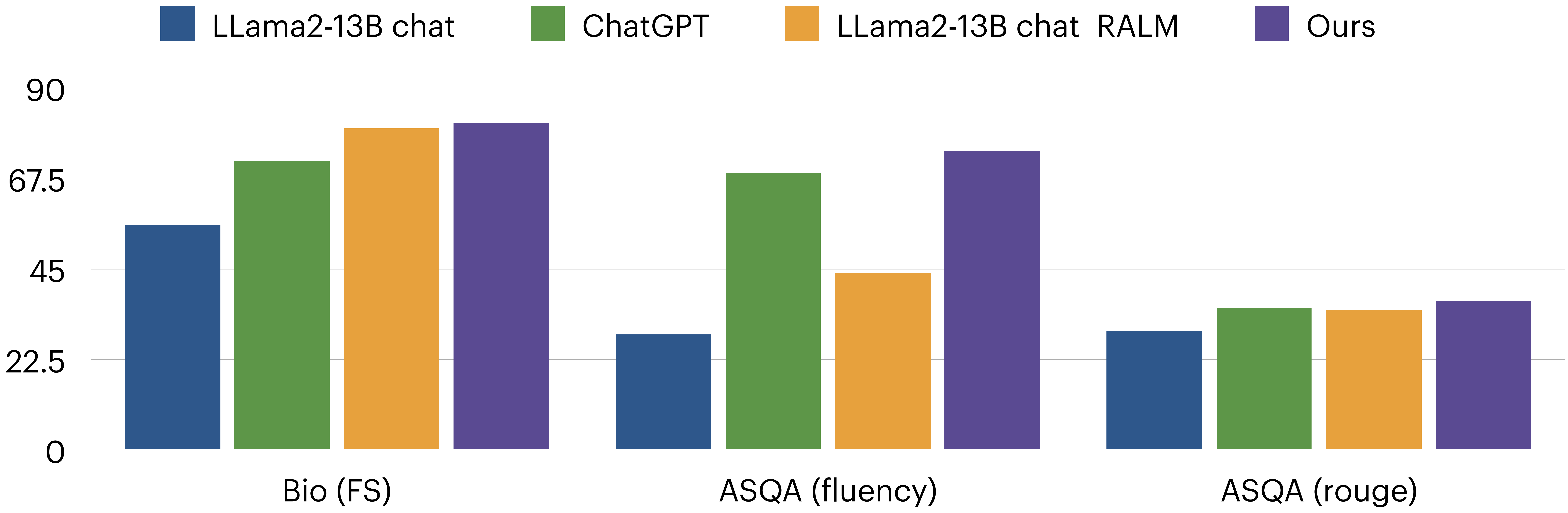
Experimental results (long-form)

Outperforms other LMs in terms of factuality & fluency correctness



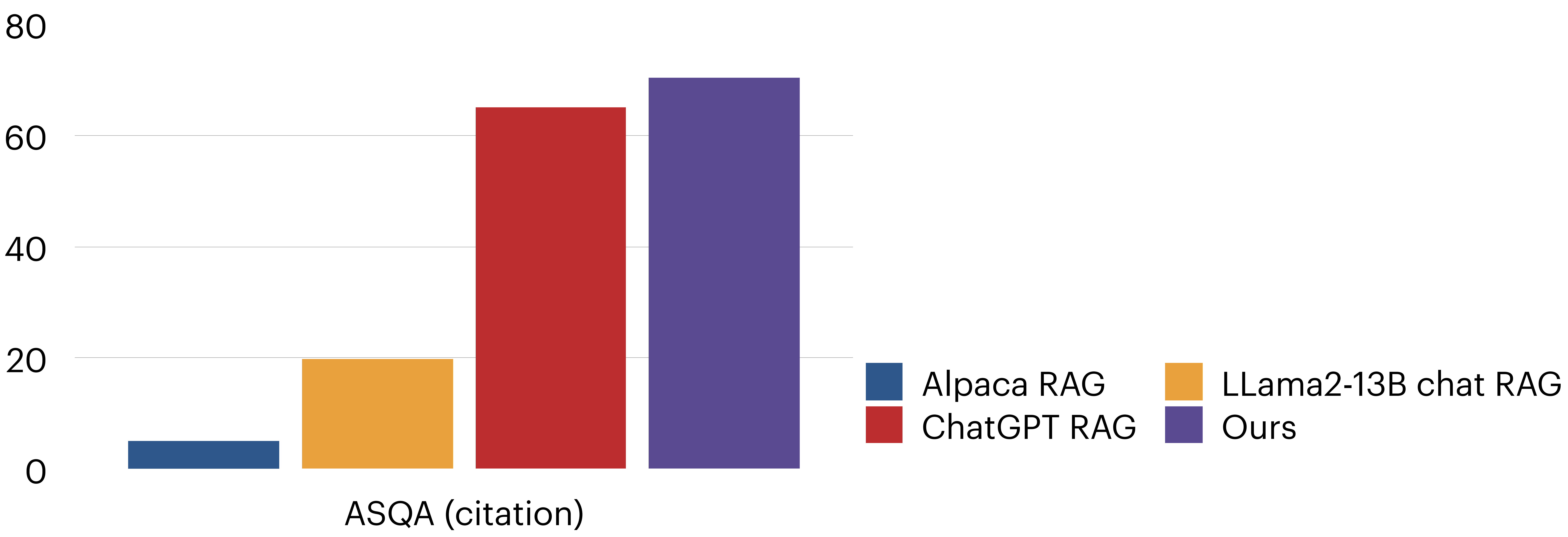
Experimental results (long-form)

Outperforms other LMs in terms of factuality & fluency correctness



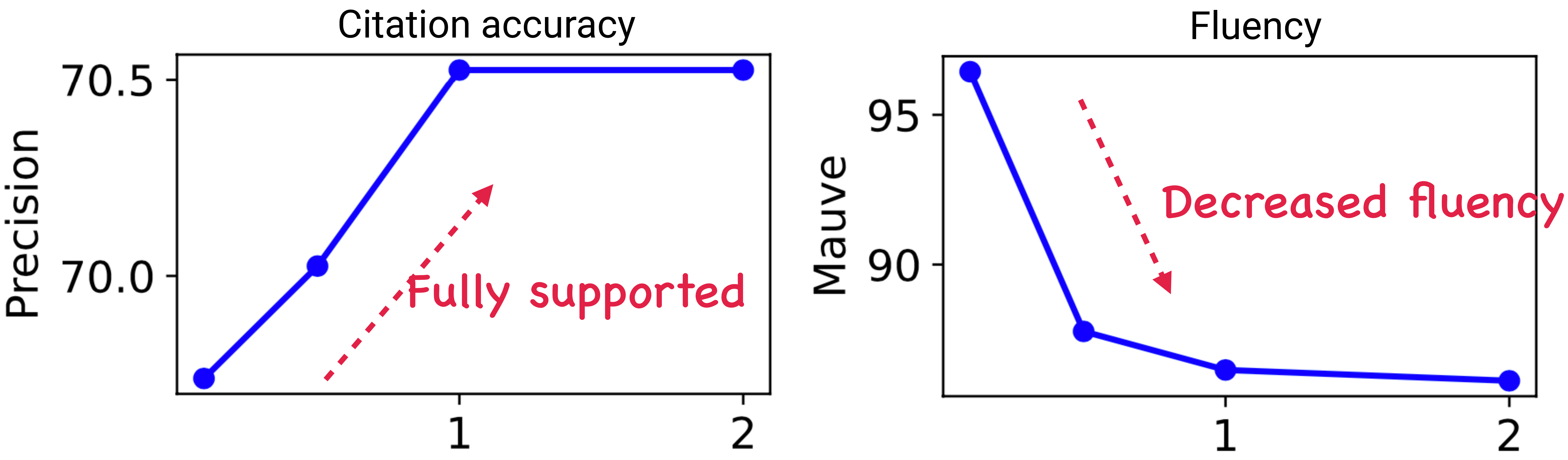
Experimental results (long-form citation precisions)

Significantly improves llama2-13B citation accuracy, matching ChatGPT



Inference-time customization via self-reflection

Decoding-time control via reflection tokens change model behaviors

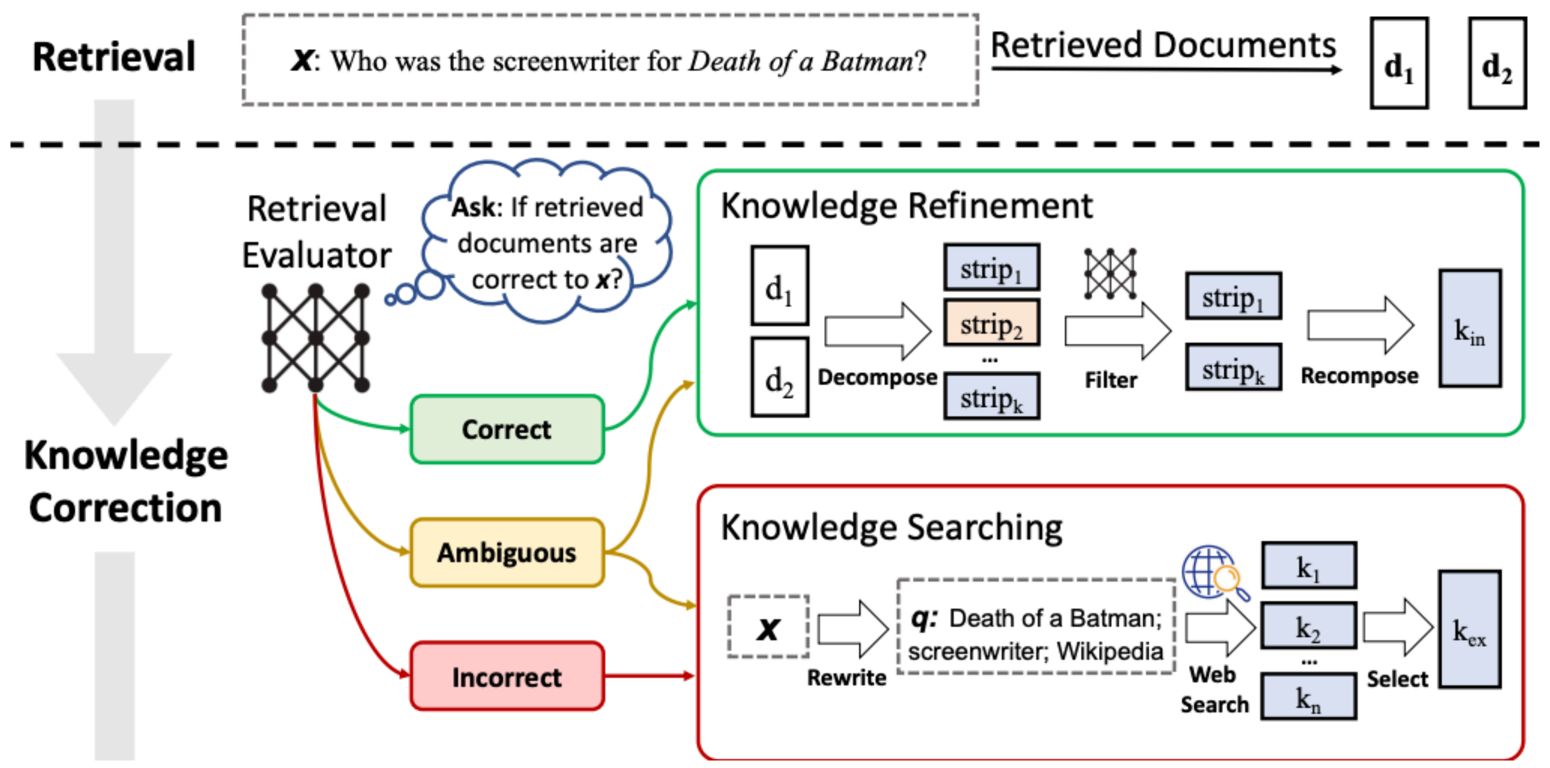


x axis — Weight for **Supported** (larger —> more emphasis on supported)

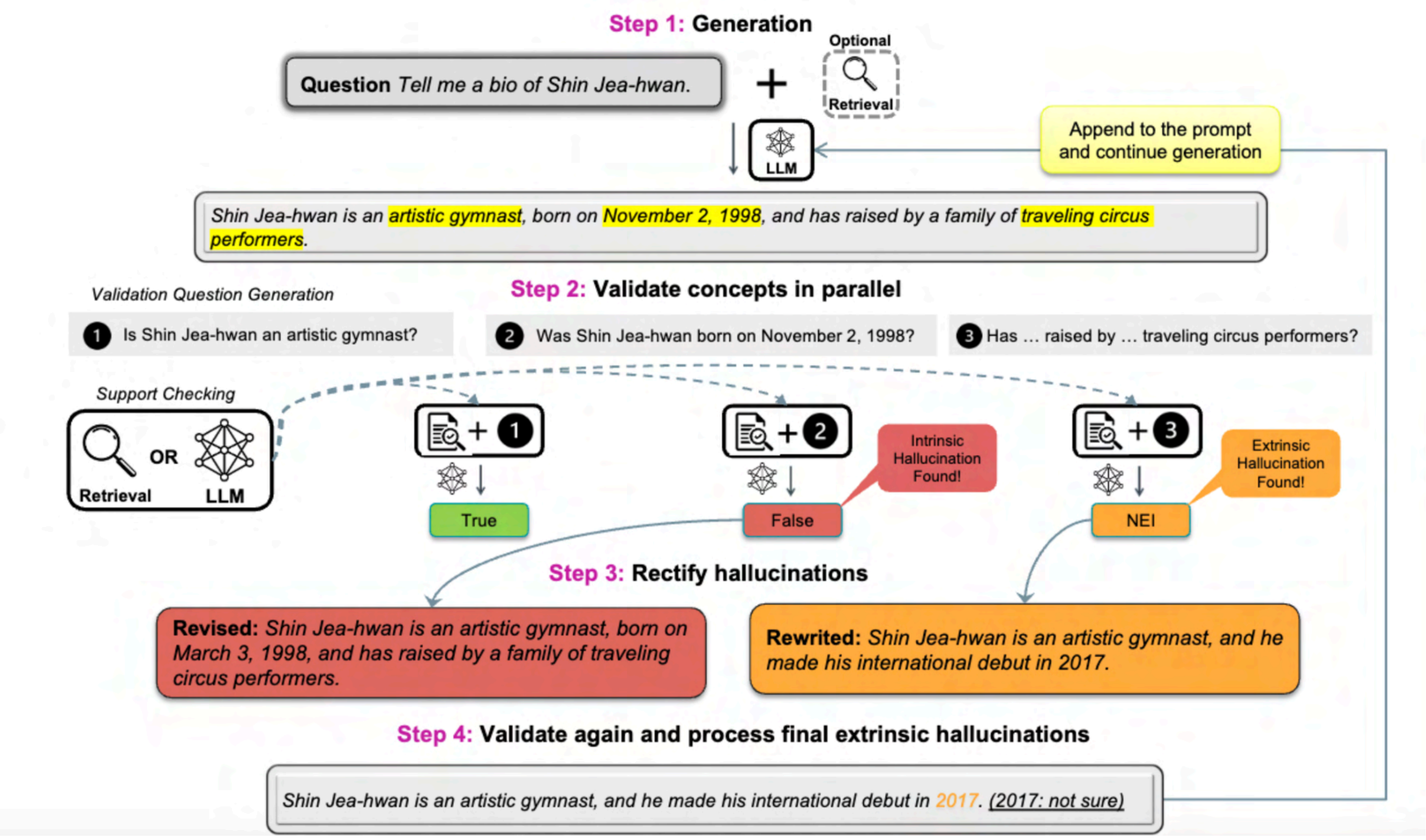
Impacts on academic communities and applications

New advanced RAG methods with reflections inspired by Self-RAG

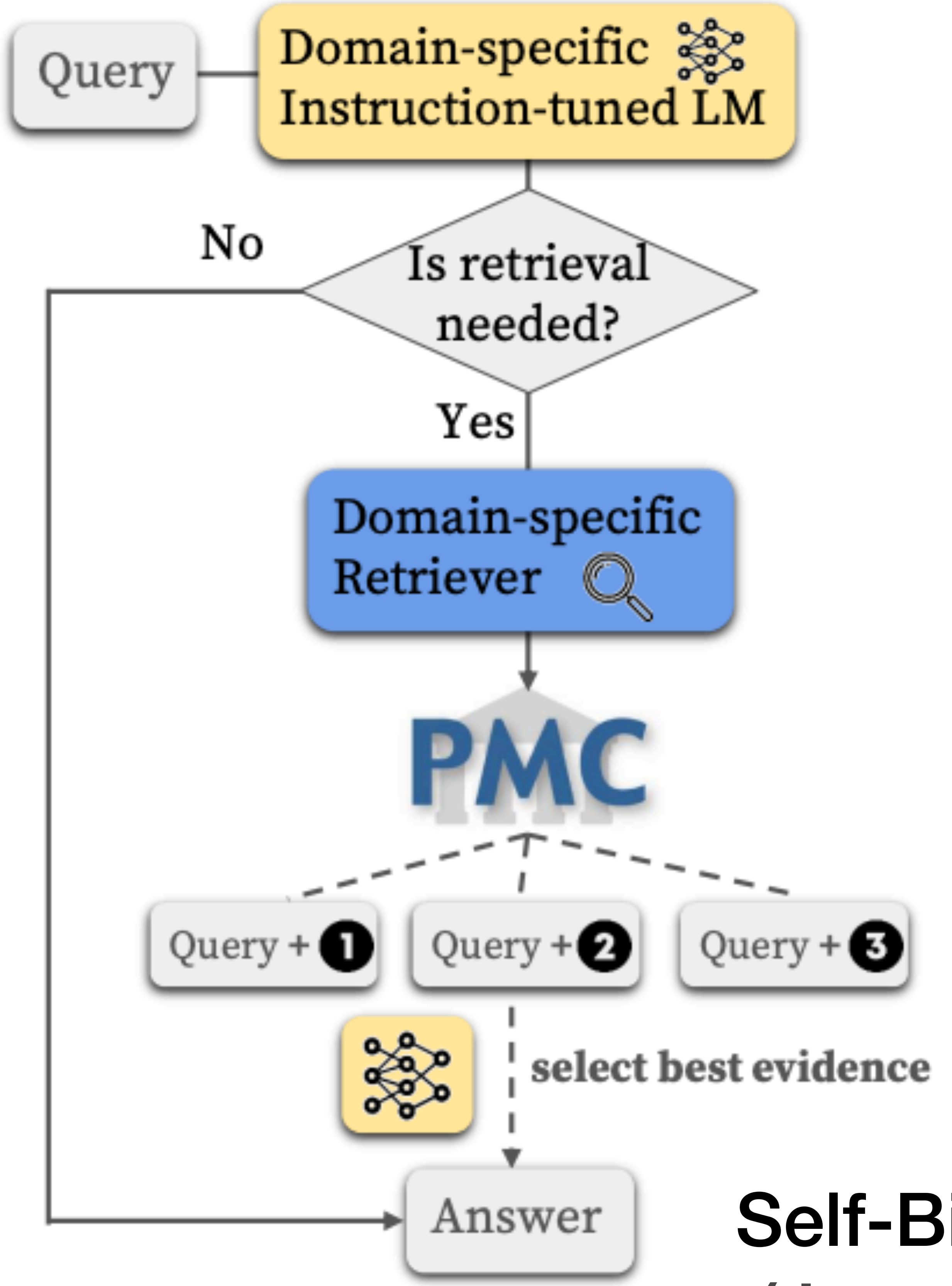
Applications to expert domains (e.g., Biomedical)



CRAG
(Yang et al., 2024)

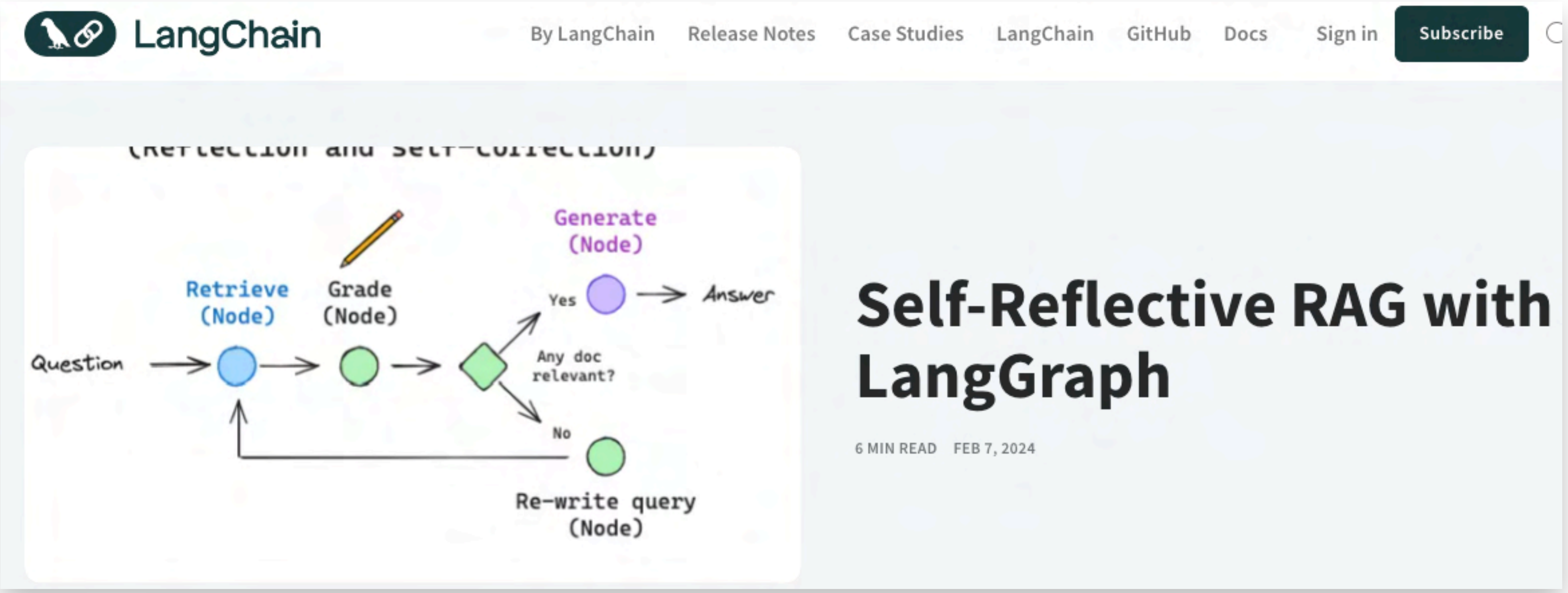


EVER
(Kang et al., 2024)



Self-BioRAG
(Jeong et al., 2024)

Impacts on academic communities and applications



LlamaIndex • Feb 13, 2024

LlamaIndex Newsletter 2023–02–13

Newsletter Llamaindex AI Rag LLM

🤩 **The highlights:**

1. **Self-RAG**: Introducing Self-RAG, now part of LlamaIndex as a LlamaPack. Boosts LLM training and RAG workflows with dynamic capabilities. Notebook, Tweet.

Self-RAG has been integrated into LangChain, LlamaIndex ... 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 (1.4k ⭐!)

Today's lecture

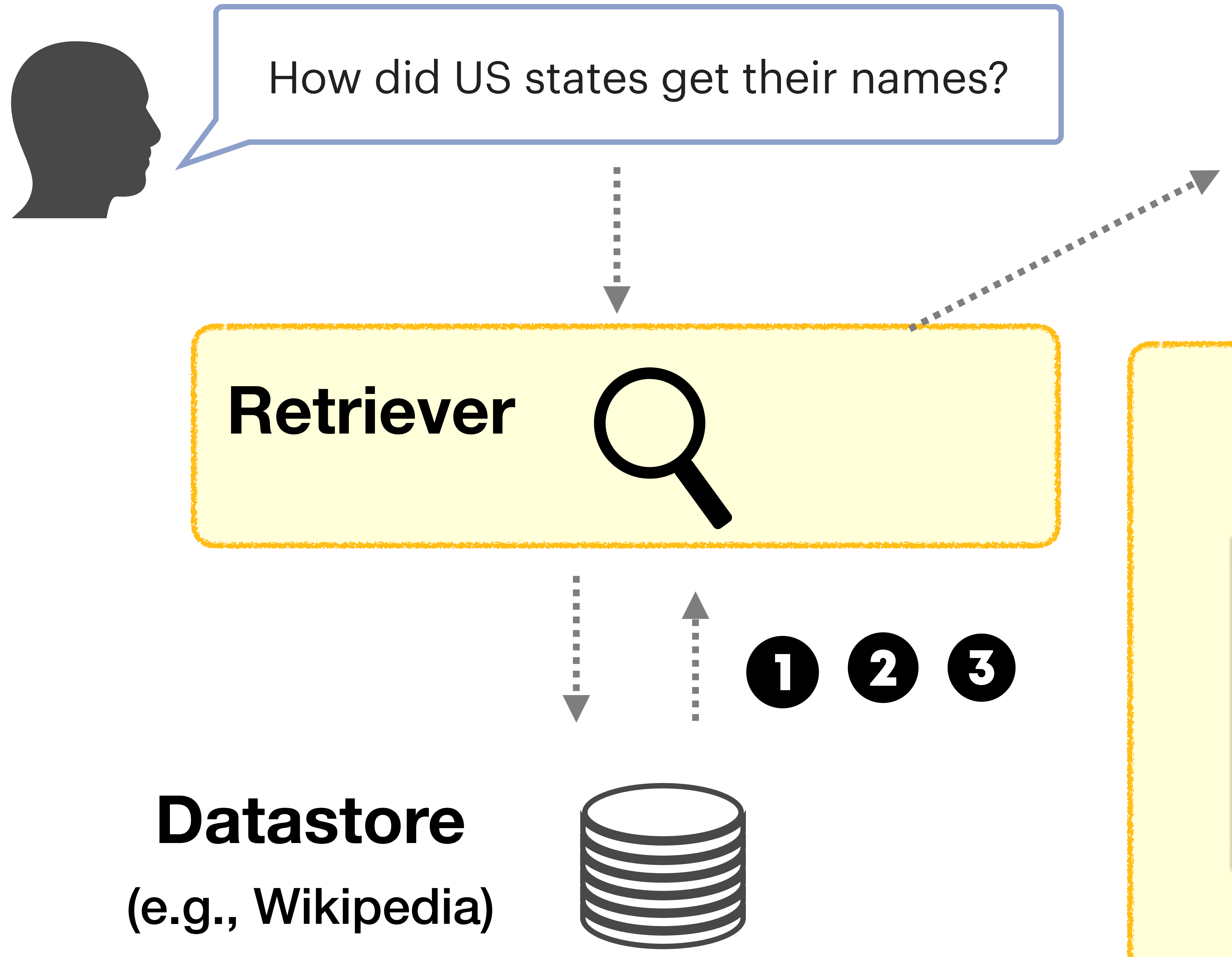
Promises and Limitations of Retrieval-augmented LMs

Reliable inference: Self-reflective RAG with dynamic retrieval

Versatile retriever: Intent-aware retrievers with LMs

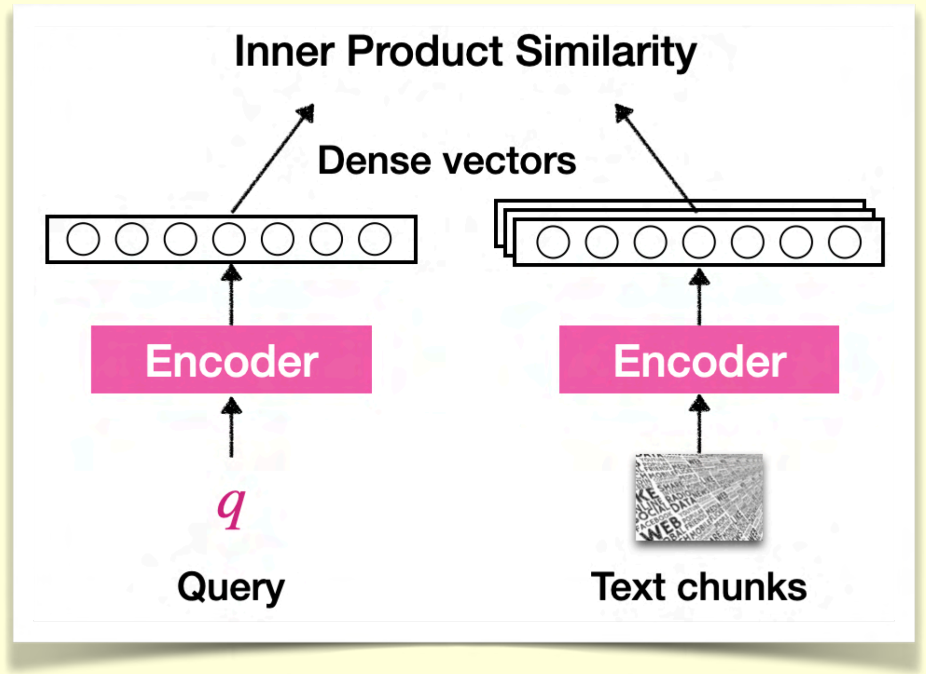
Summary and future directions: RAG in the wild

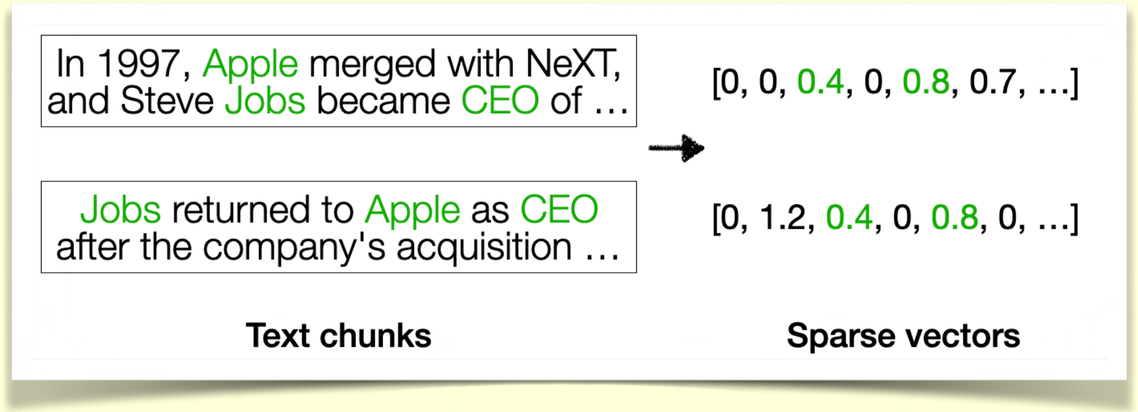
Standard RAG originally designed for a single task (e.g., QA)



Answer my question using references.
References: ❶ ❷ ❸
Question: How did US states get their names?

Neural retriever (DPR)	**Lexical retriever** (BM 25 / TF-IDF)





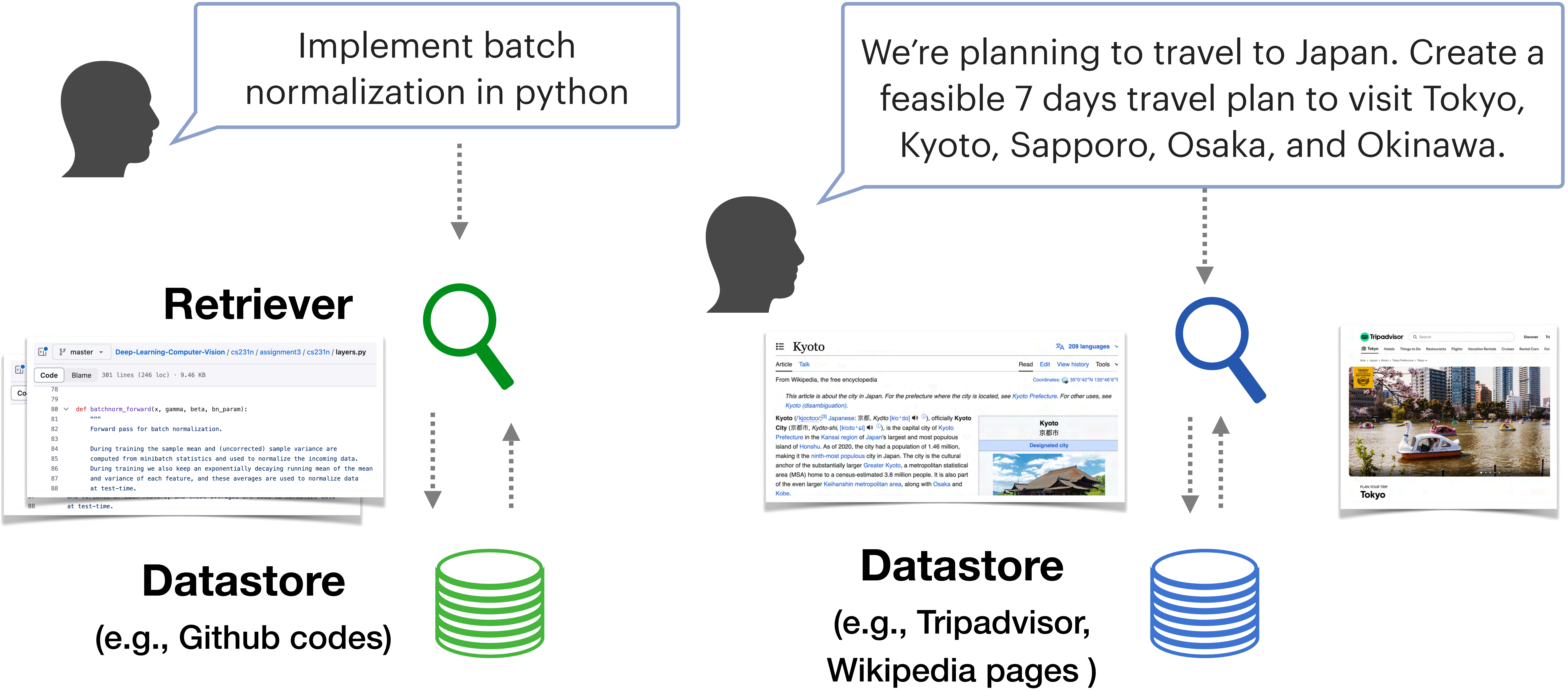
Capture **similarities**
between query and document

Modern RAG systems have been used in diverse scenarios

Implement batch normalization in python

We’re planning to travel to Japan. Create a feasible 7 days travel plan to visit Tokyo, Kyoto, Sapporo, Osaka, and Okinawa.

Modern RAG systems have been used in diverse scenarios



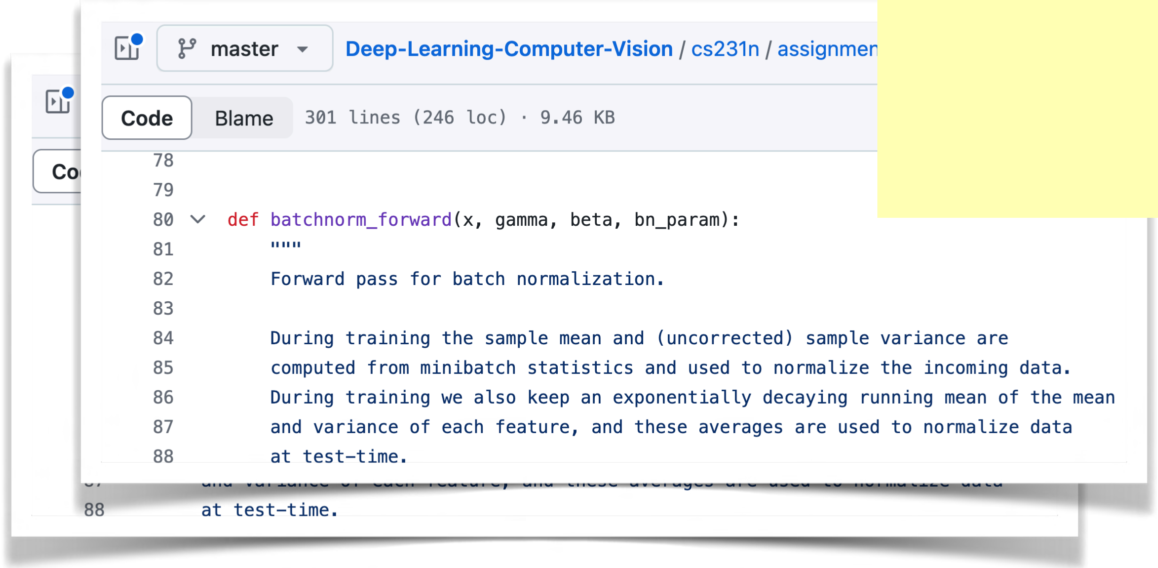
Modern RAG systems have been used in diverse scenarios

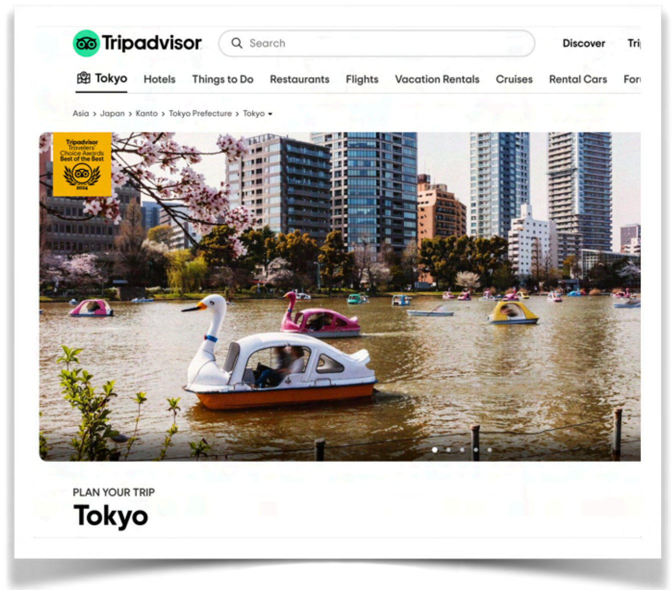
Implement batch normalization in python

We're planning to travel to Japan. Create a feasible 7 days travel plan to visit Tokyo, Kyoto, Sapporo, Osaka, and Okinawa.

Retrie

Different information needs often require different information retrieval systems





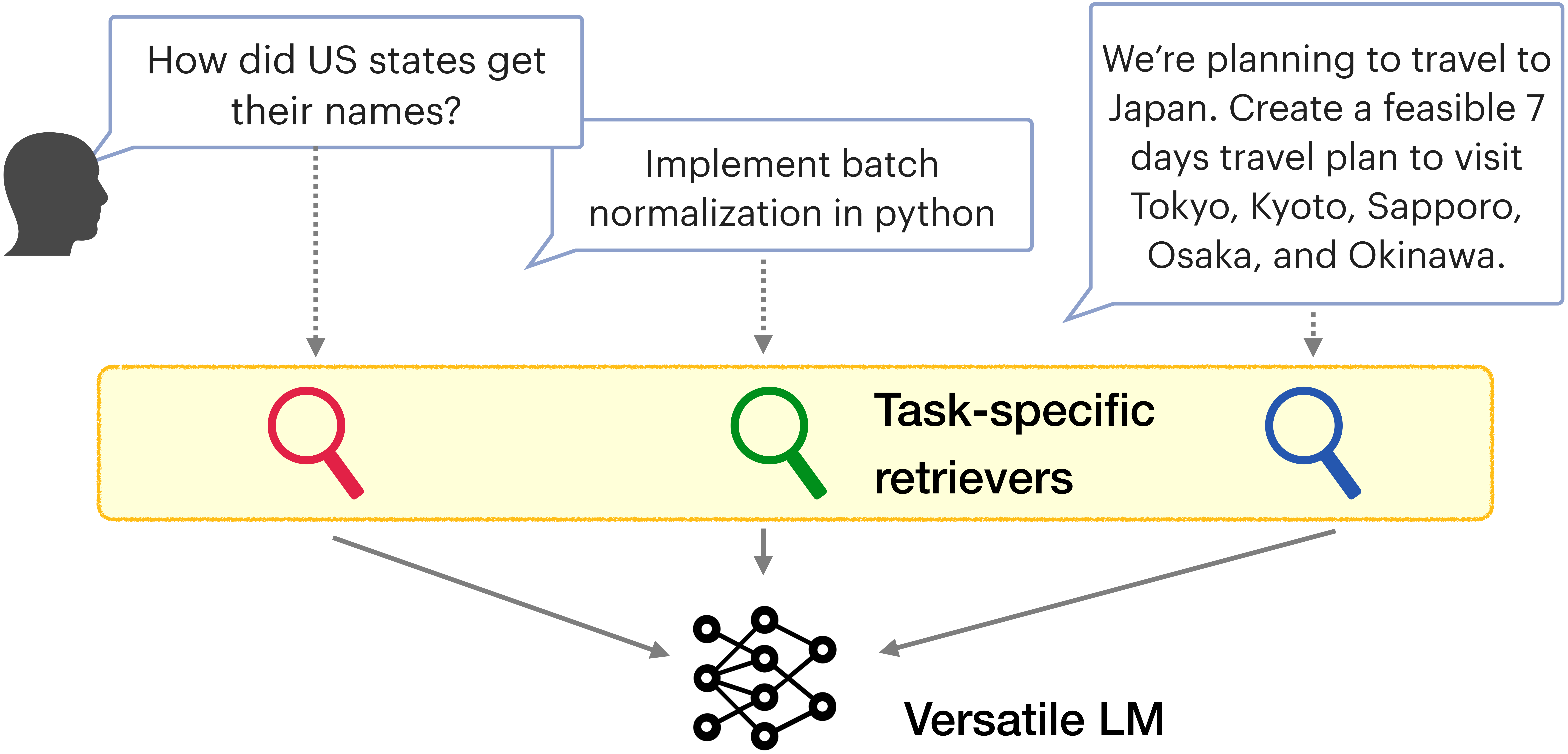
Datastore

(e.g., Github codes)

Datastore

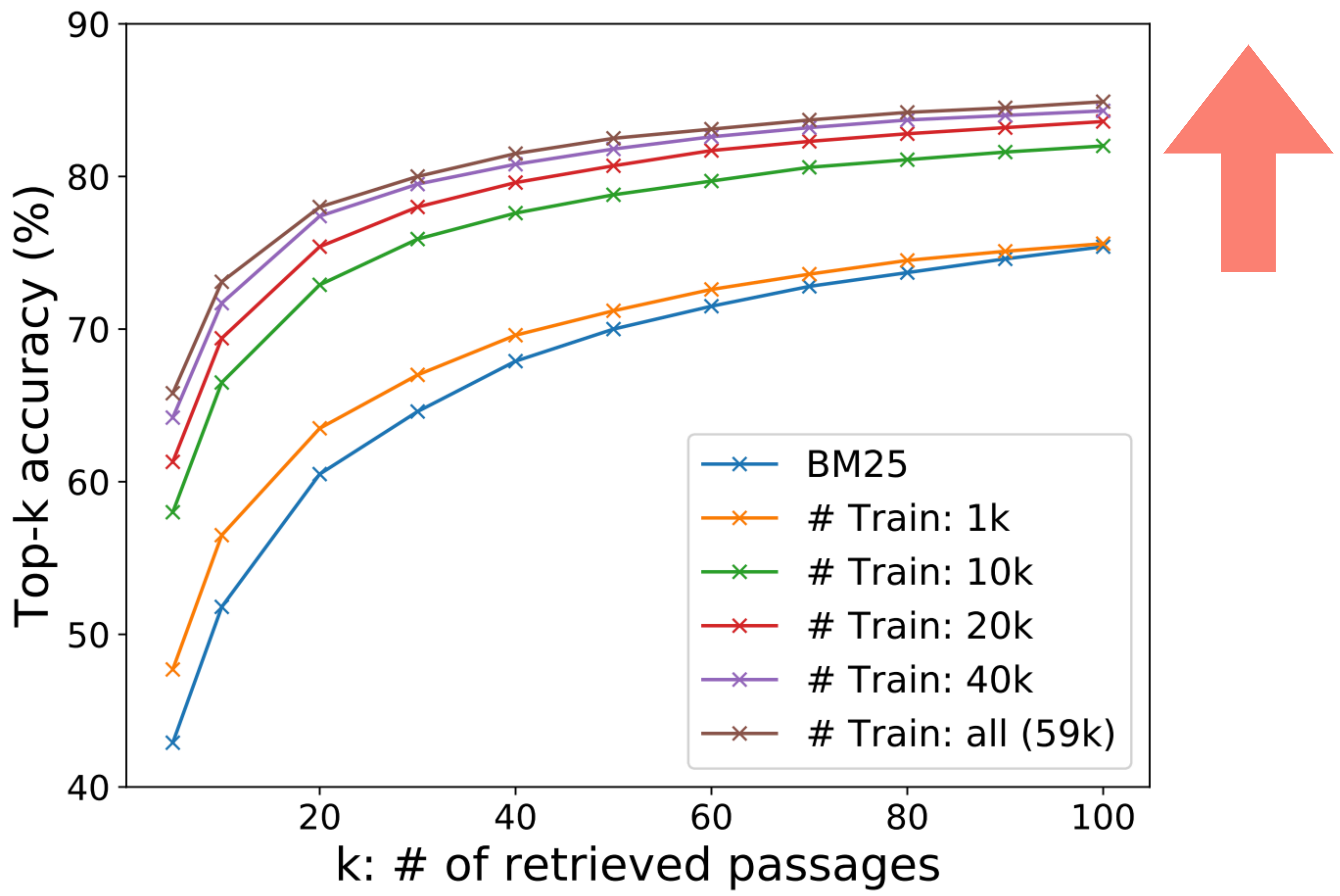
(e.g., Tripadvisor, Wikipedia pages)

Versatility of RAG systems limited by conventional retrieval

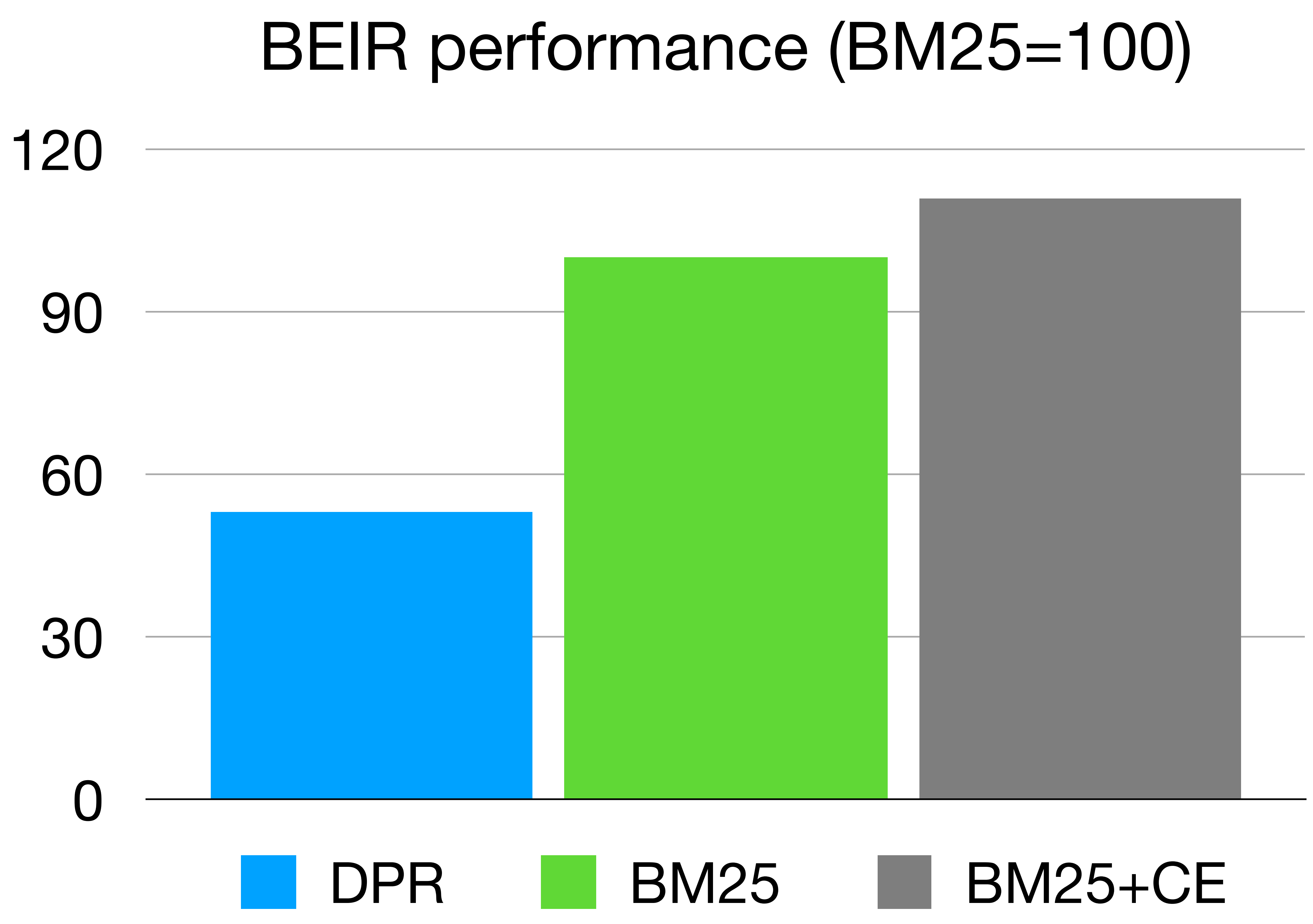


Trained retrieval systems typically struggle in OOD

Perform well with more task training data



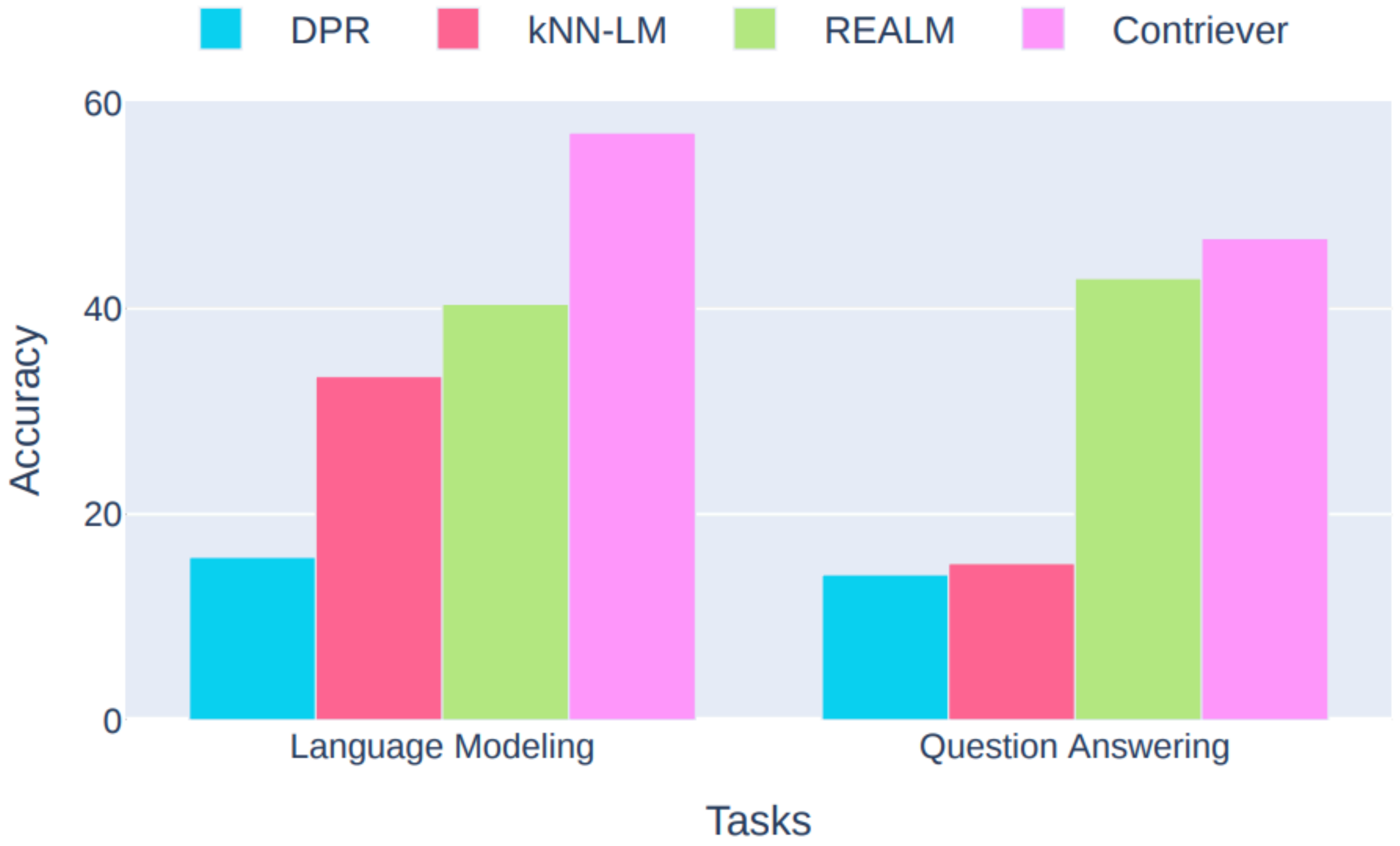
Struggle in new tasks & domains



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

“Similar” documents may not be always helpful

Helpful documents for some tasks aren’t necessary similar to queries



In a zoo located in a warm region, what should be included in the polar bear exhibit? (Multi-hop reasoning task)

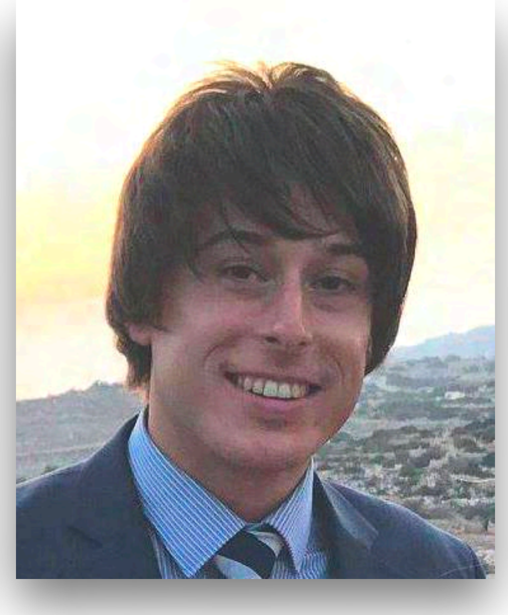
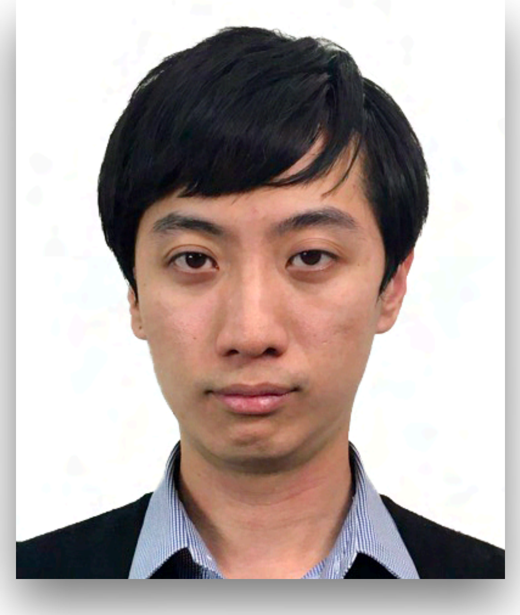
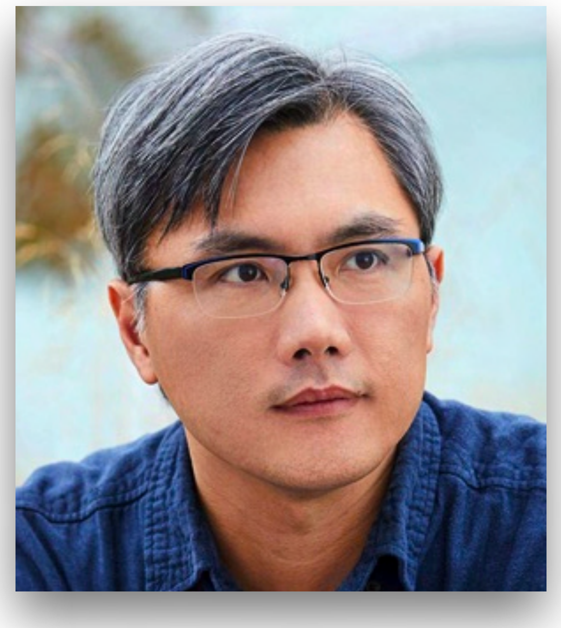
If an animal lives a certain environment then that animal usually requires that kind of environment.

Polar bears live in cold environments

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

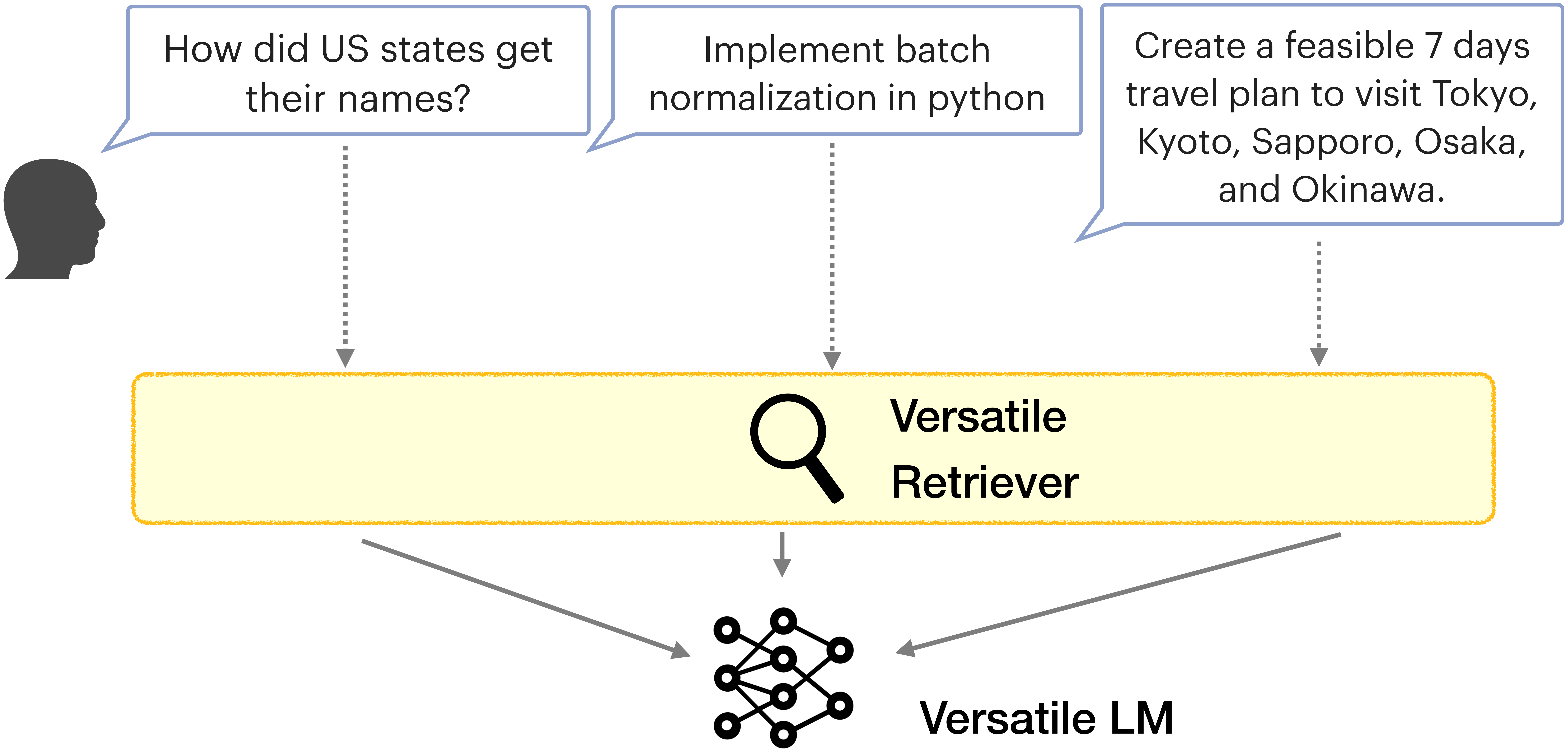
Task-aware Retrieval with Instructions

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

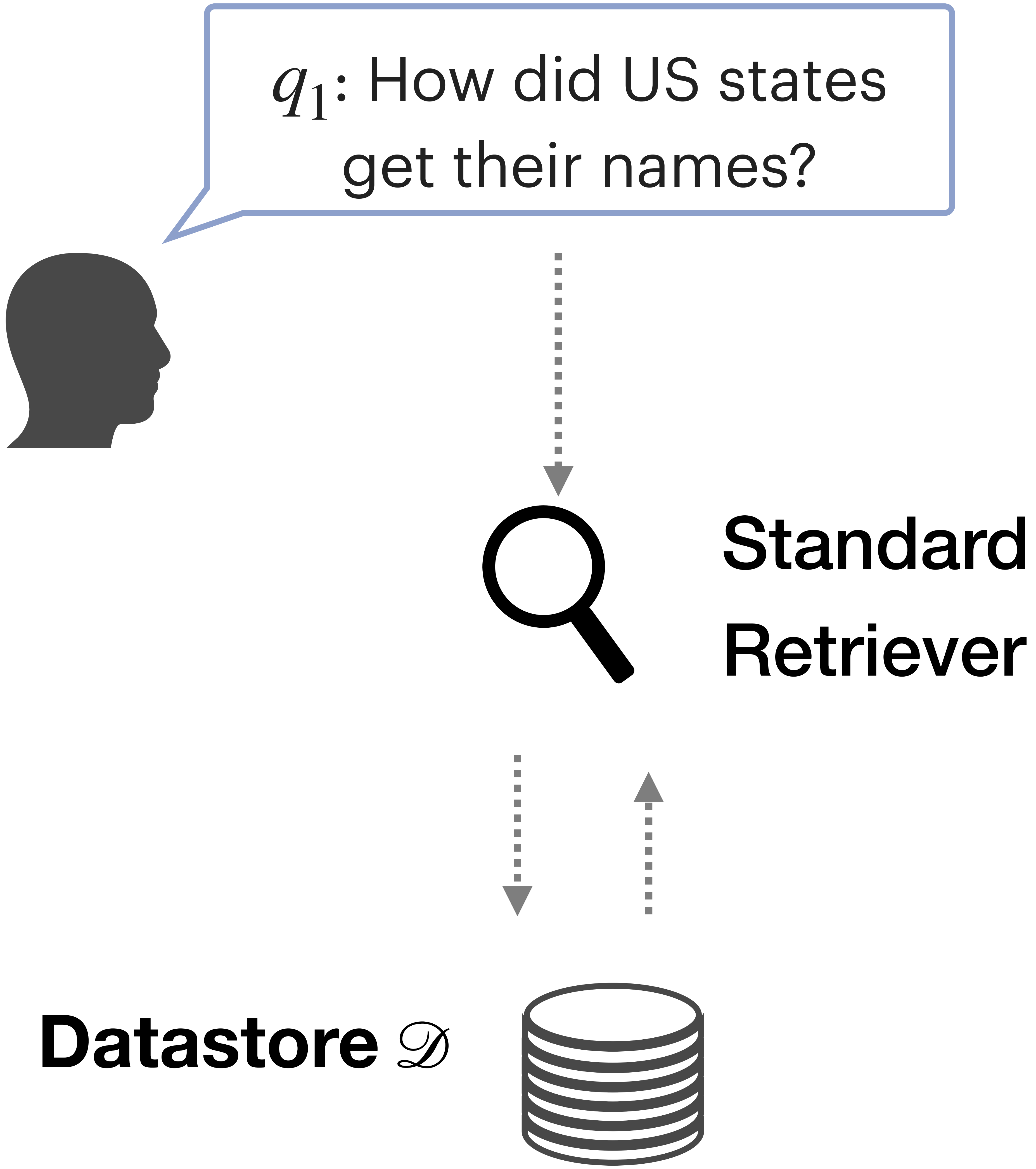
       

ACL Findings 2023

Goal of Tsk-Aware ReTriever (TART)

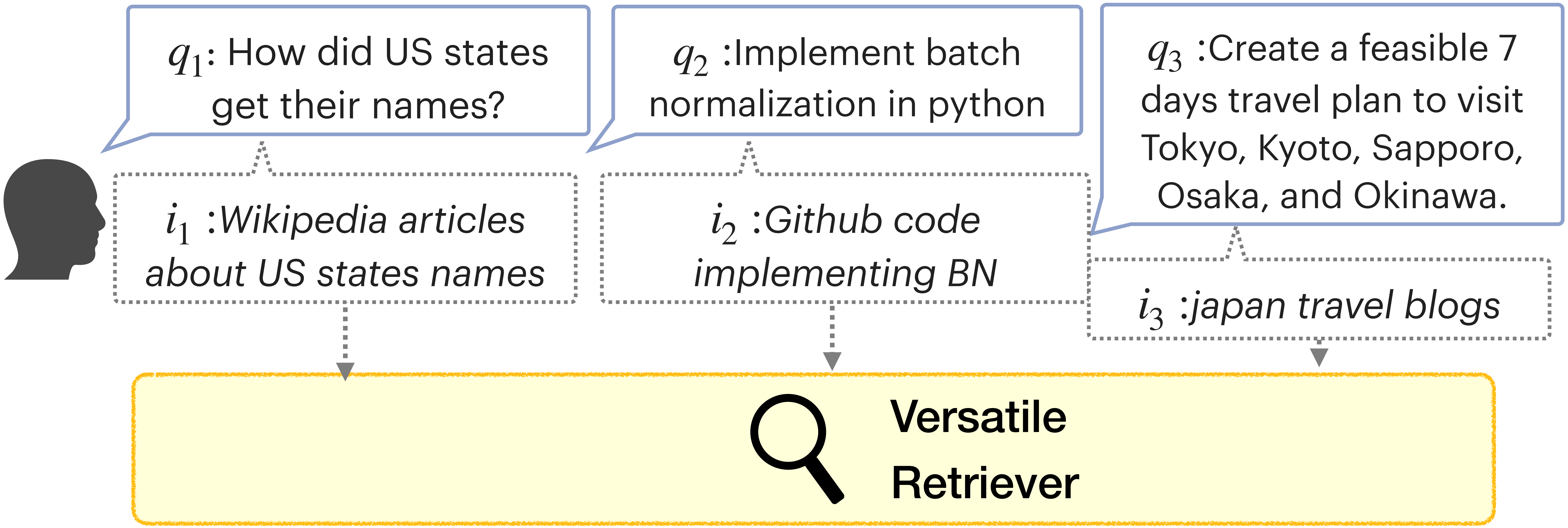


Normal retrieval task



$$\text{Topk } \text{argTop}k_{d \in \mathscr{D}} \text{sim}(q, d)$$

New task: Retrieval with Instruction



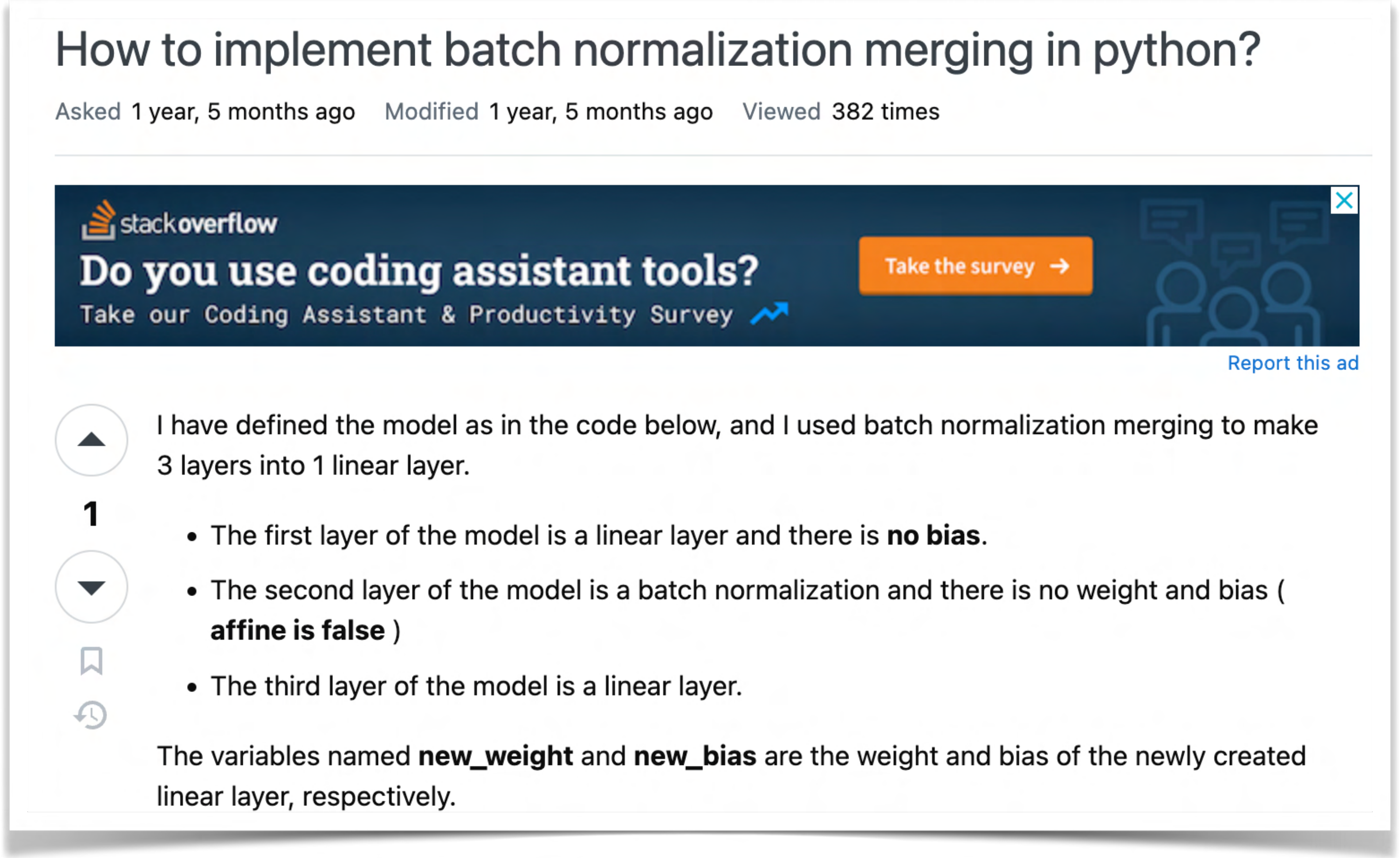
$$\text{Topk argTop}k_{d \in \mathcal{D}} \text{sim}(q, d, i)$$

New task: Retrieval with Instruction

q_2 :Implement batch normalization in python

i_1 : *similar stack overflow questions*

i_2 :*Github code implementation of BN*



How to implement batch normalization merging in python?

Asked 1 year, 5 months ago Modified 1 year, 5 months ago Viewed 382 times

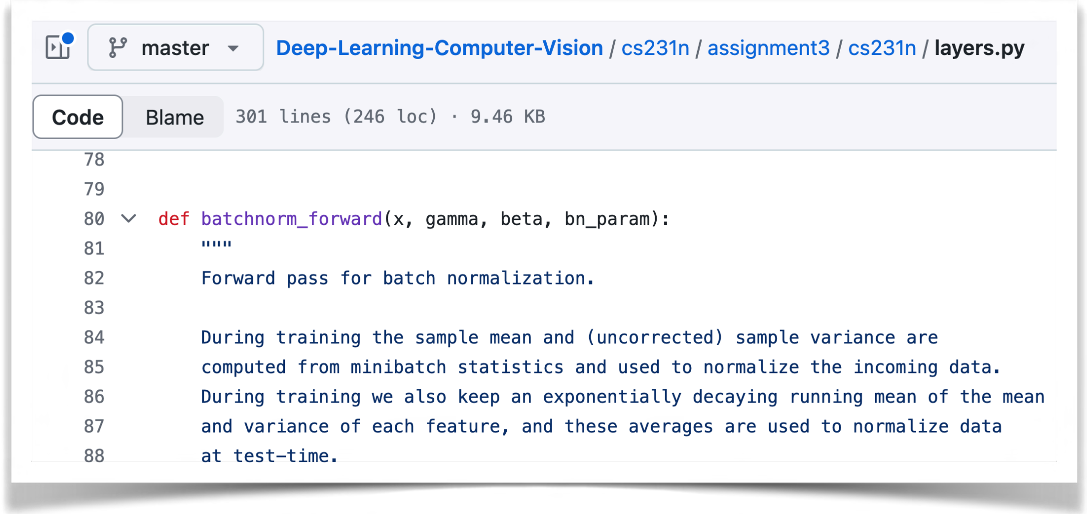
stackoverflow
Do you use coding assistant tools?
Take our Coding Assistant & Productivity Survey
Take the survey
Report this ad

1

I have defined the model as in the code below, and I used batch normalization merging to make 3 layers into 1 linear layer.

- The first layer of the model is a linear layer and there is **no bias**.
- The second layer of the model is a batch normalization and there is no weight and bias (**affine is false**)
- The third layer of the model is a linear layer.

The variables named **new_weight** and **new_bias** are the weight and bias of the newly created linear layer, respectively.



master Deep-Learning-Computer-Vision / cs231n / assignment3 / cs231n / layers.py

Code Blame 301 lines (246 loc) · 9.46 KB

```
def batchnorm_forward(x, gamma, beta, bn_param):
    """
    Forward pass for batch normalization.

    During training the sample mean and (uncorrected) sample variance are
    computed from minibatch statistics and used to normalize the incoming data.
    During training we also keep an exponentially decaying running mean of the mean
    and variance of each feature, and these averages are used to normalize data
    at test-time.
```

$$\text{Topk } \text{argTopk}_{d \in \mathscr{D}} \text{sim}(q, d, i)$$

BERRI: first large-scale retrieval dataset with instructions

Curated 50 tasks with expert annotation instructions across domains

Question Answering
- Natural Questions
- HotpotQA
- SQuAD
- Yahoo Answers
- Stack Exchange
- ReCoRD
- MSMARCO
- SearchQA
- ELI5
- WikiHOW
- PubMedQA
- MedMCQA

Duplicated Question
- OQA
- StackExchange (dup.questions)
- Quora Questions

Summarization
- CNN / Daily
- XSUM
- NPR
- Multi-LexSUM
- SciTLDR

Dialogue
- Wizard of Wiki.
- QReCC

Paraphrasing
- Medical text simplification

Fact Checking
- FEVER

Others
- Coco captions
- CodeSearchNet

Instruction-scheme for retrieval tasks

Propose effective instruction scheme for retrieval tasks
and annotate instructions

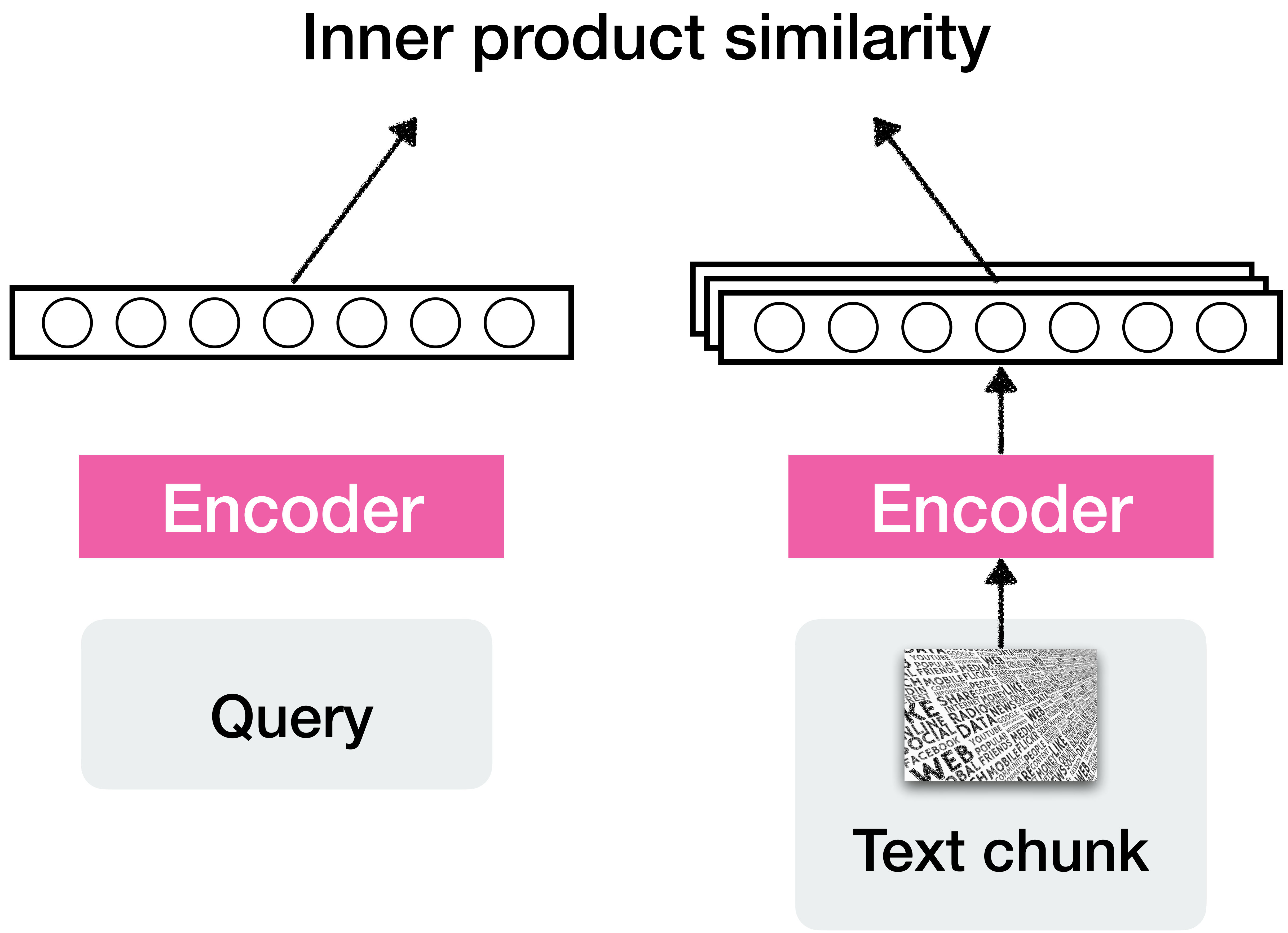
Dataset	Instruction
NQ	Retrieve a Wikipedia paragraph that answers this question.
QReCC	Find a dialogue response from dialogue history to answer the user's question.
Arguana	Retrieve a paragraph from an argument website that argues against the following argument.
SciFact	Find a sentence from a scientific 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

Instruction-aware bi-encoder retriever (TART-dual)



Positive paragraph

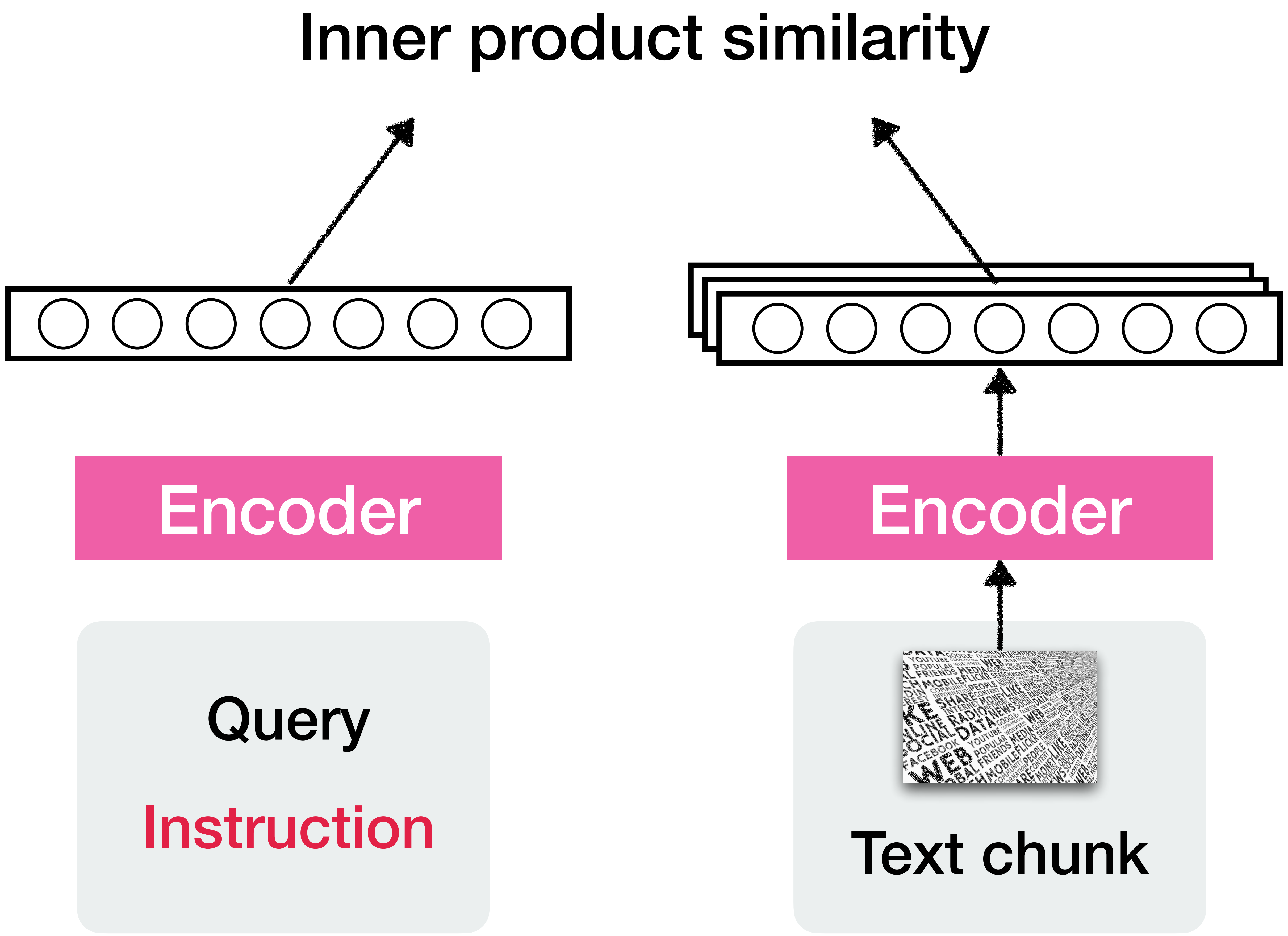
Negative paragraphs

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

Contrastive learning

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

Instruction-aware bi-encoder retriever (TART-dual)



Positive paragraph

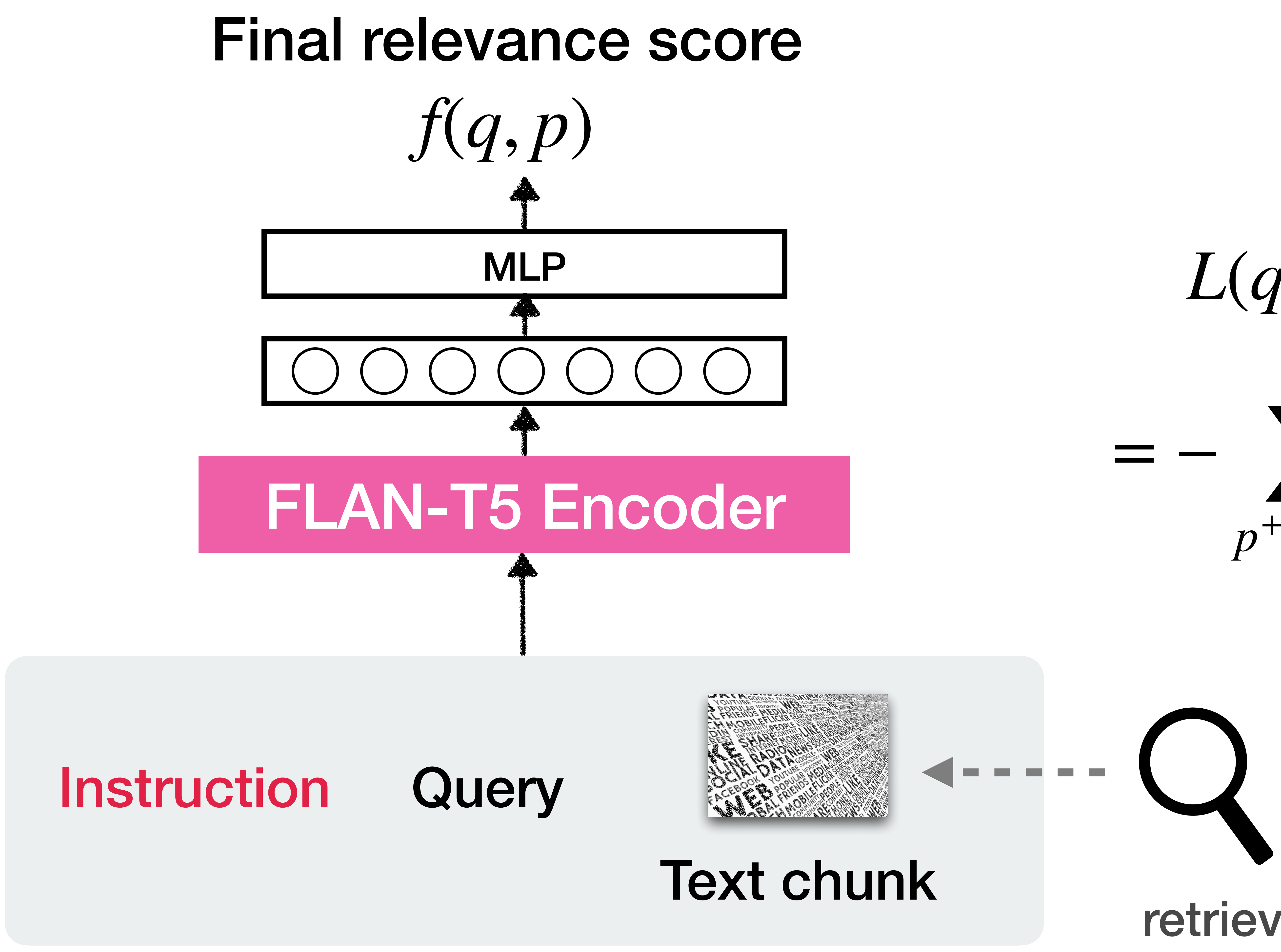
Negative paragraphs

$$L(q, \boxed{p^+}, \boxed{p_1^-, p_2^-, \ldots, p_n^-})$$

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

Contrastive learning

Instruction-aware cross-encoder retriever (TART-full)



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

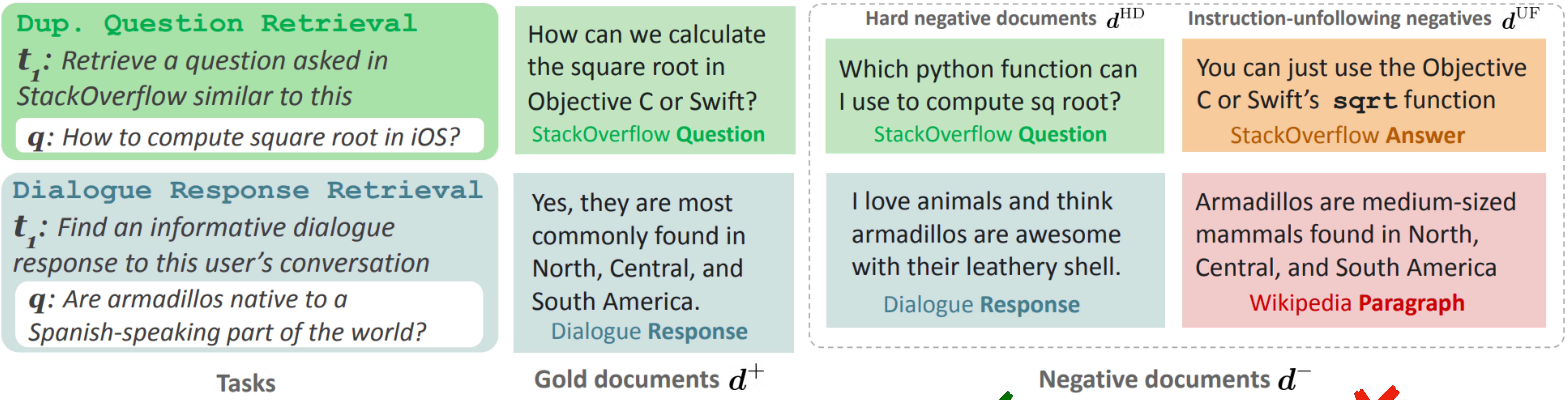
$$= - \sum_{p^+ \in P^+} \log(f(q, p^+)) - \sum_{p \in P^-} \log(1 - f(q, p^-))$$

Cross-entropy loss

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

New negative samples: instruction *un*-following samples

Carefully defining negative samples is a key for successful retrievers



	d^{HD}	d^{UF}
Follow instruction?	✓	✗
Relevant to the query?	✗	✓

Experimental settings

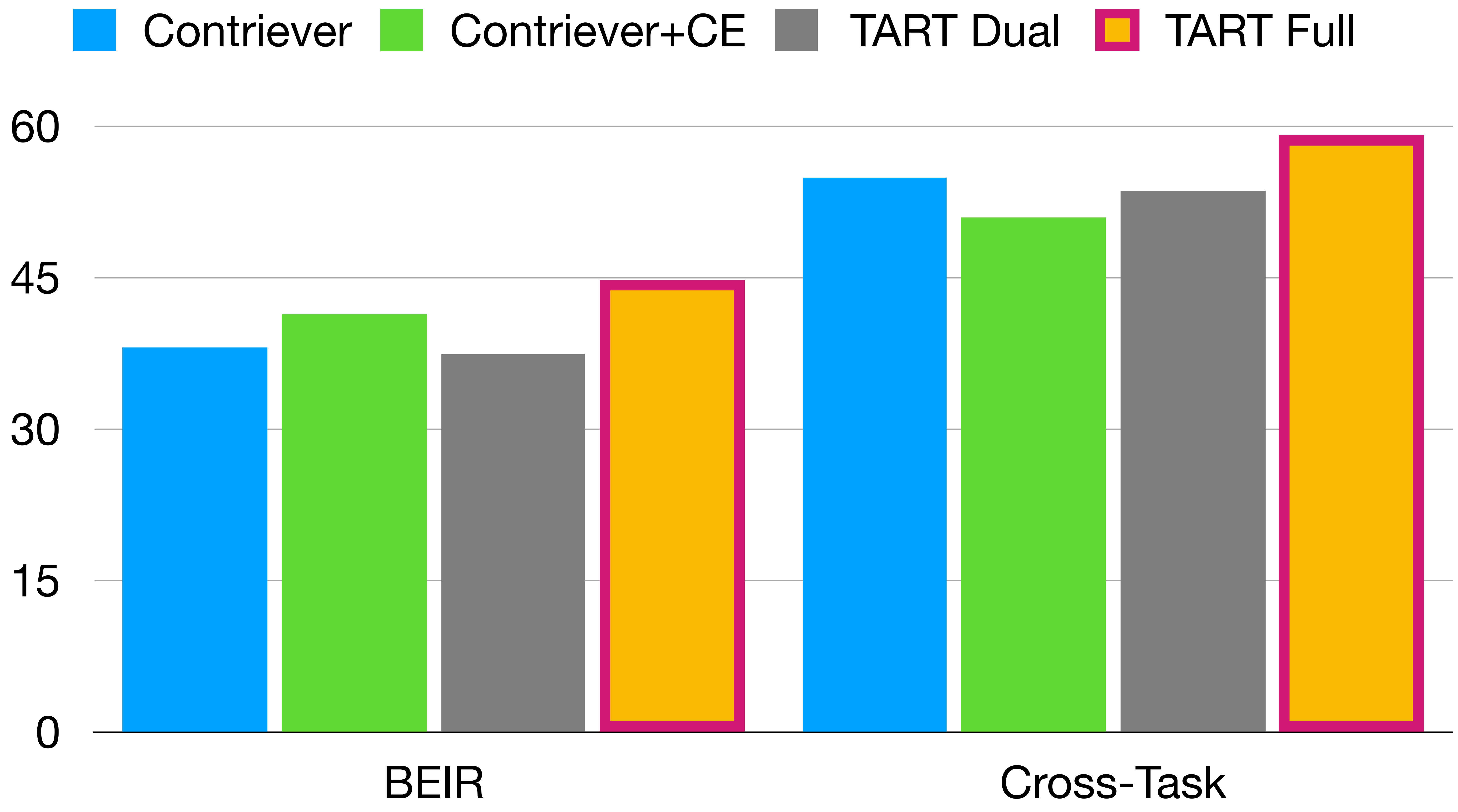
More details of training & test are in our paper!

- Our models
  - **TART-dual (bi-encoder)** - Instruction-tuning of Contriever-MS MARCO 110M
  - **TART-full (cross-encoder)** - instruction-tuning FLAN-T5 3B Encoder (1.5B)

- Evaluations
  - **Zero-shot retrieval:** generalize to new retrieval task via instruction
    - **BEIR** (Thakur et al. 2021)
    - **LOTTE** (Khattab et al., 2022)
  - **Cross-task cross-domain retrieval**: synthetically combine retrieval tasks to test instruction following by pairing two relevant tasks (e.g., QA & question retrieval)

Better Generalization and Instruction Following

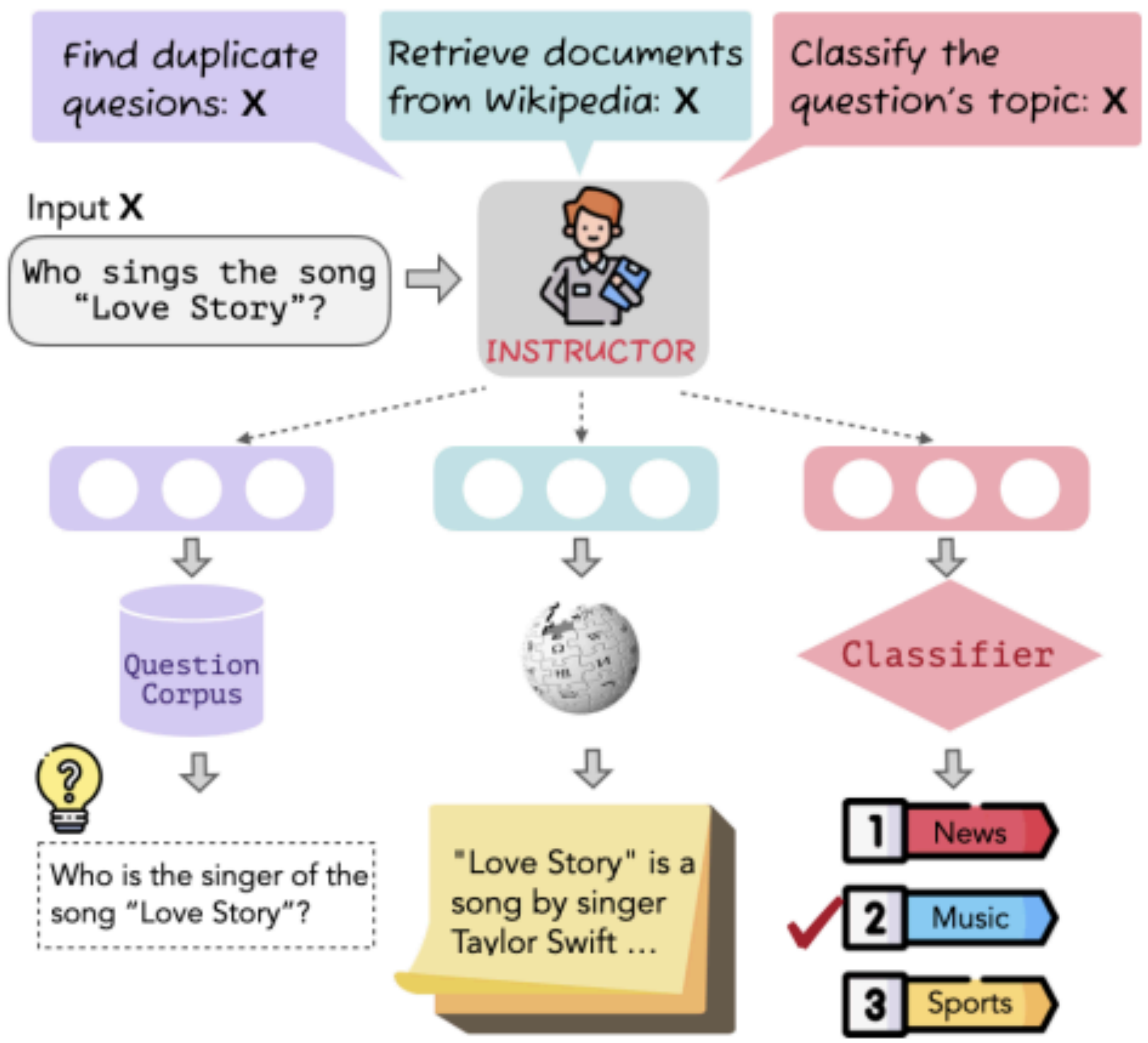
Instruction-tuning for retrieval (3B) shows effectiveness



Increasing number of instruction-following retrievers

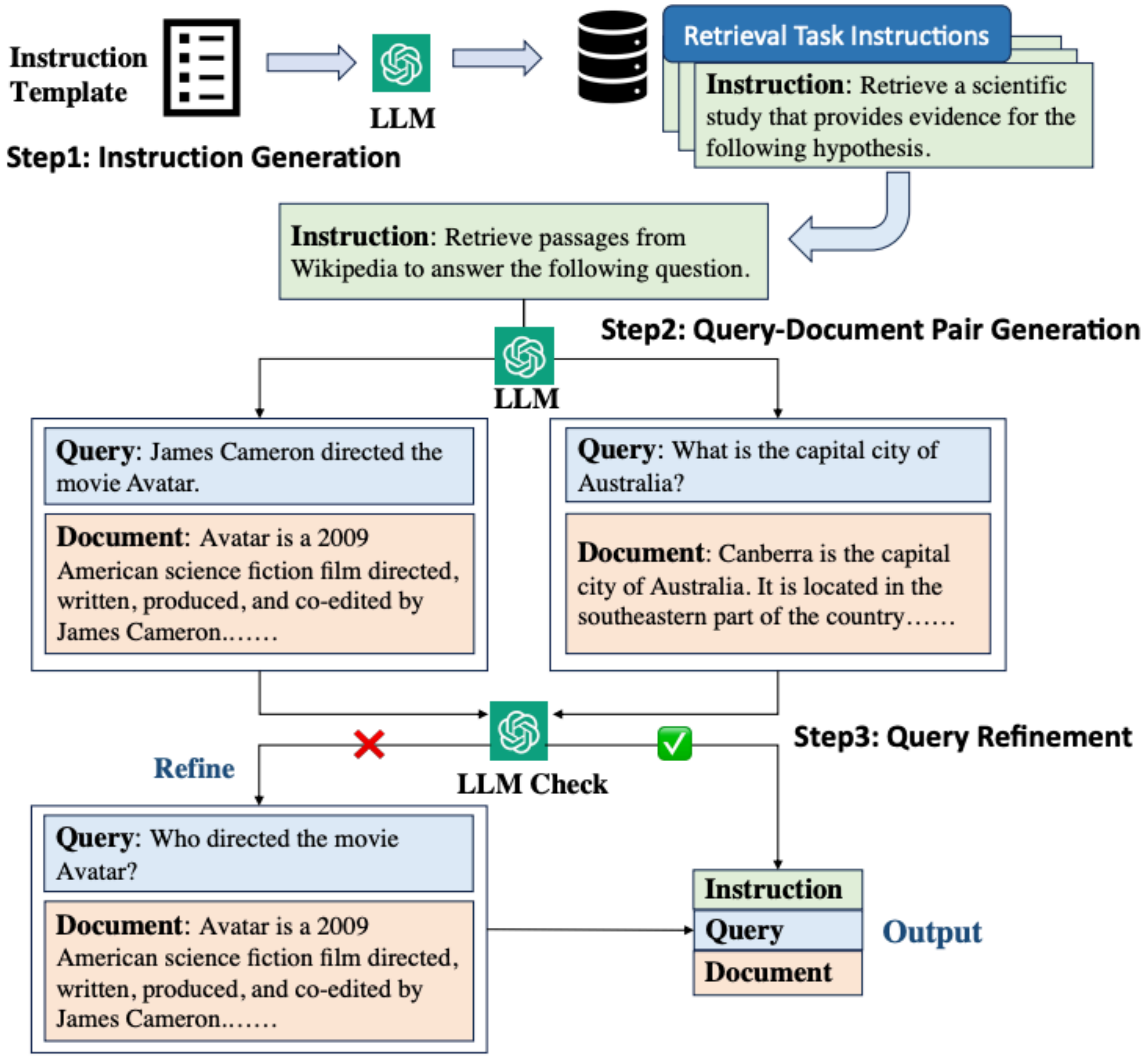
Instructor
(Su et al., 2023)

Concurrent work focusing on embeddings trained on existing datasets



ControlRetriever
(Pang et al., 2023)

Leverage LLMs to generate training data
E5 MISTRAL Instruct (Wang et al., 2024)



Increasing number of instruction-following retrievers

Instruction following retrievers are now dominating retrieval tasks!

mteb / **leaderboard** like 2.49k • Running on CPU UPGRADE App Files

Massive Text Embedding Benchmark (MTEB) Leaderboard. To submit, refer to the MTEB GitHub repository 🤗 Refer to the MTEB paper for details on metrics, tasks and models.

Search Bar (separate multiple queries with `;`)

Search for a model and press enter...

Model types: Open, Proprietary, Sentence Transformers

Model sizes (in number of parameters): <100M, 100M to 250M, 250M to 500M, 500M to 1B, >1B

Overall | Bitext Mining | Classification | Clustering | Pair Classification | Reranking | Retrieval | STS | Summarization

English | Chinese | French | Polish

Overall MTEB English leaderboard 🔮

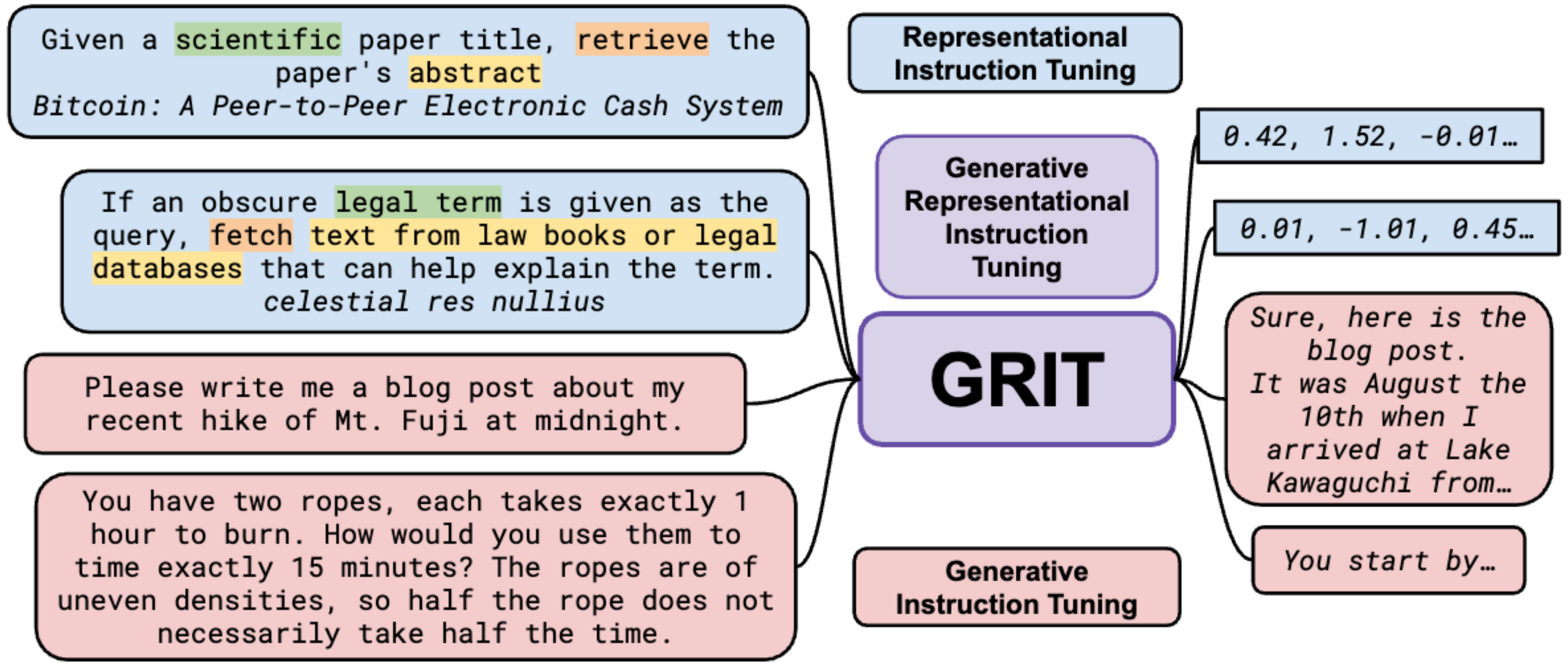
- **Metric:** Various, refer to task tabs
- **Languages:** English

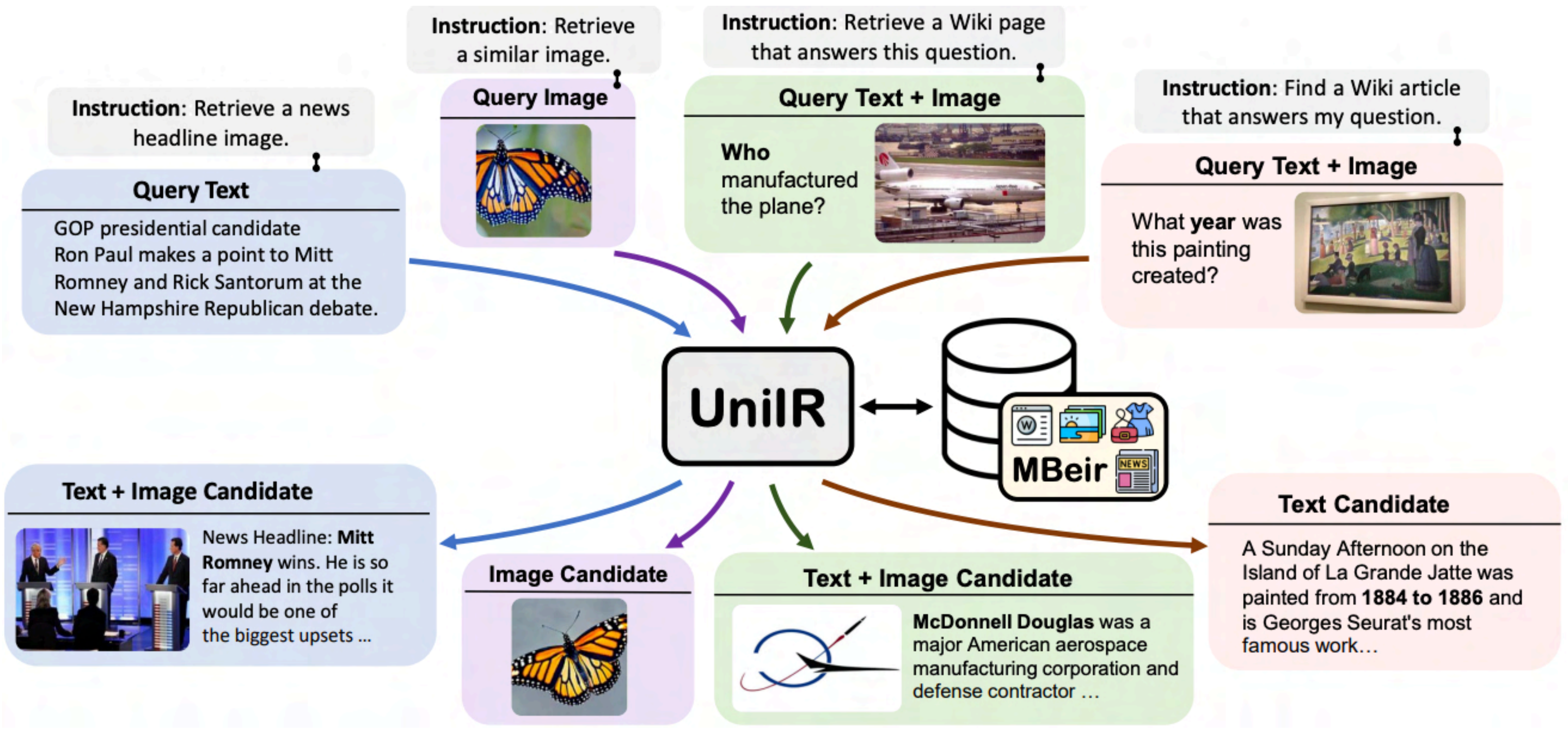
Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)
1	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67	88.54	60.64
2	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42	86.87	58.24
3	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61	87.16	60.49
4	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26	88.34	60.21

https://huggingface.co/spaces/mteb/leaderboard

Versatile retrievers improves RAG in text and multi-modal

GRIT
(Muennighoff et al., 2024)



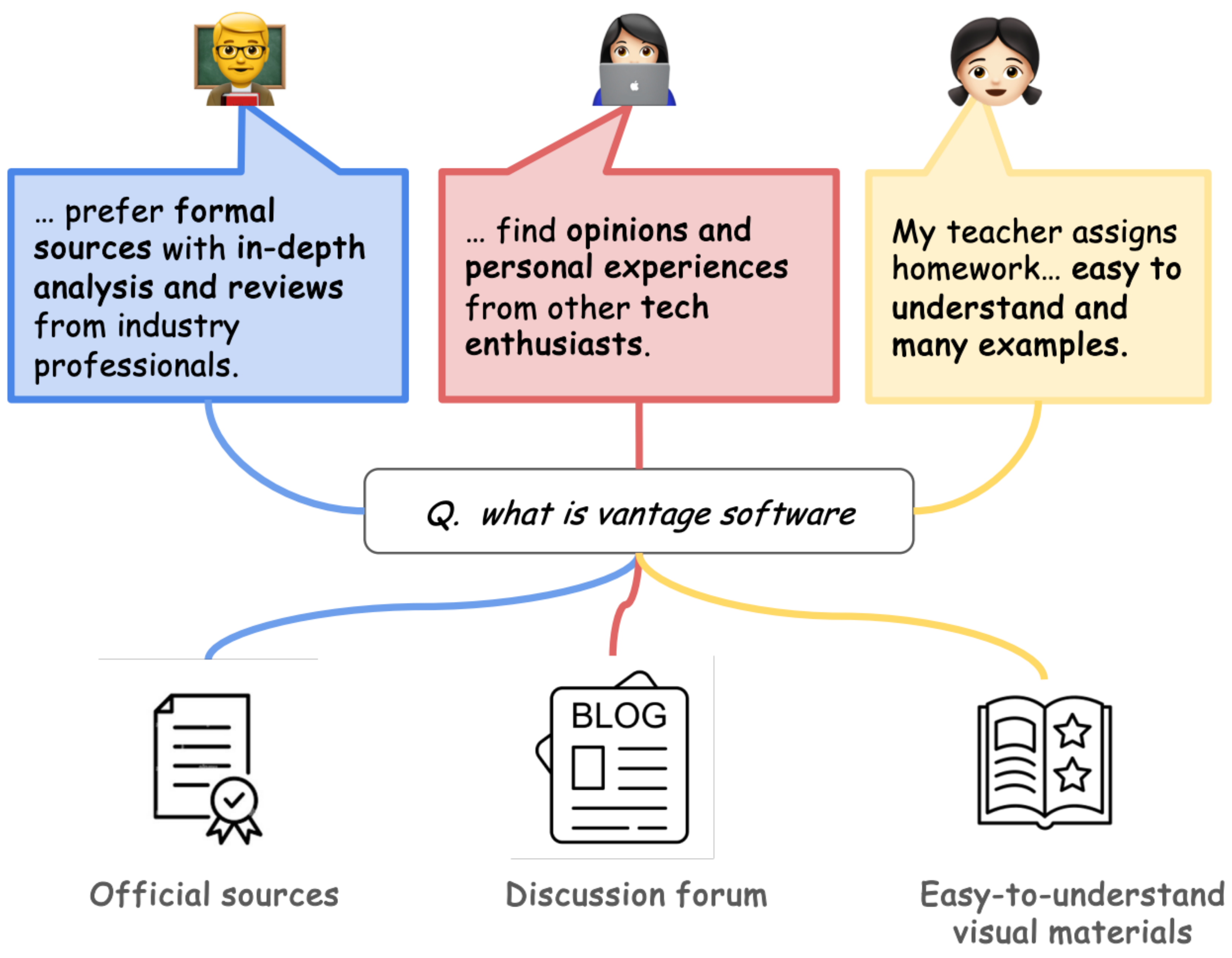


UniLR
(Wei et al., 2023)

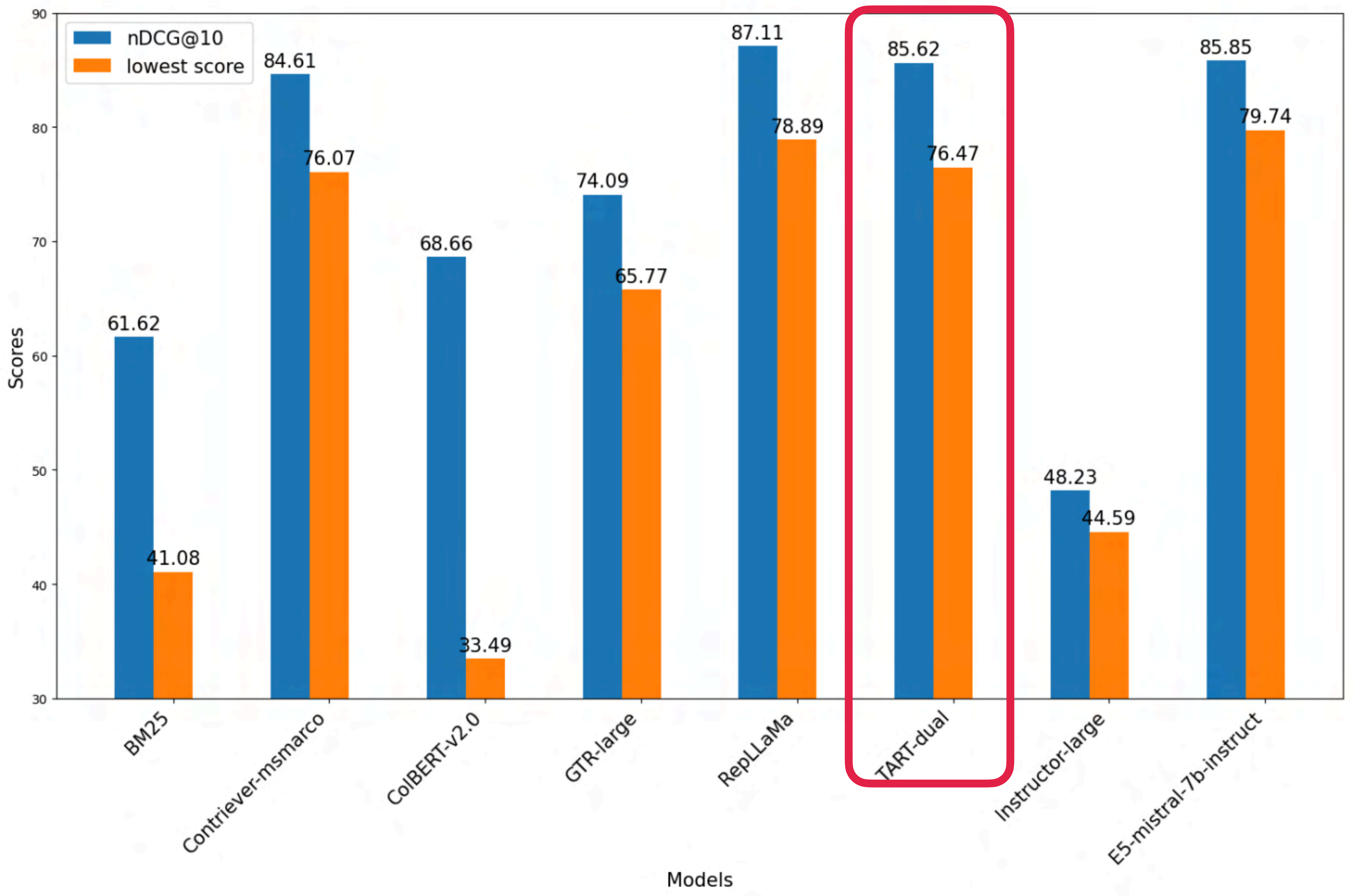
Are those retrievers following instructions?

Instructir (Oh et al., 2024)

Create a natural instruction-following retrieval dataset with LLMs



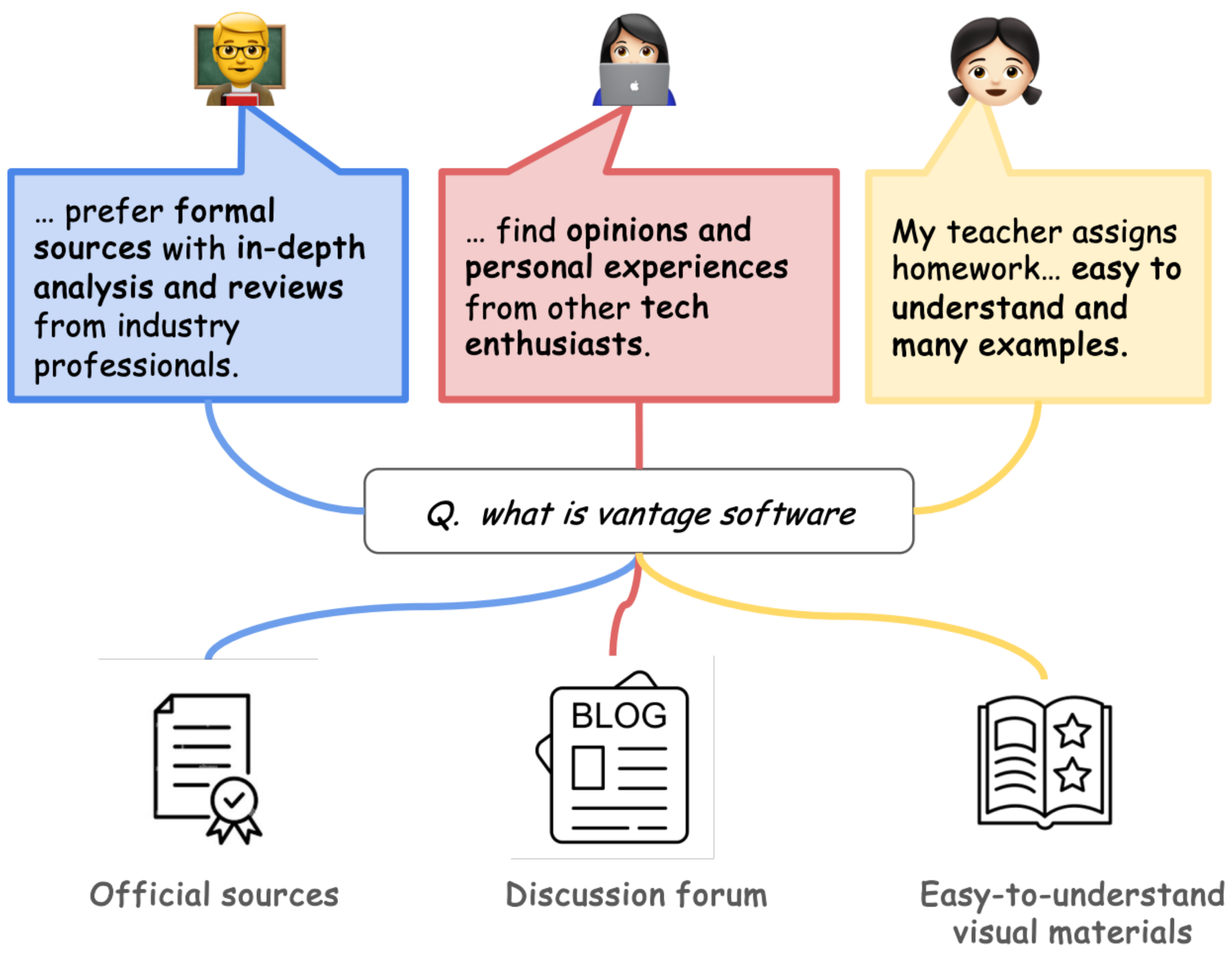
TART (110M) outperforms other <7B models by large margins



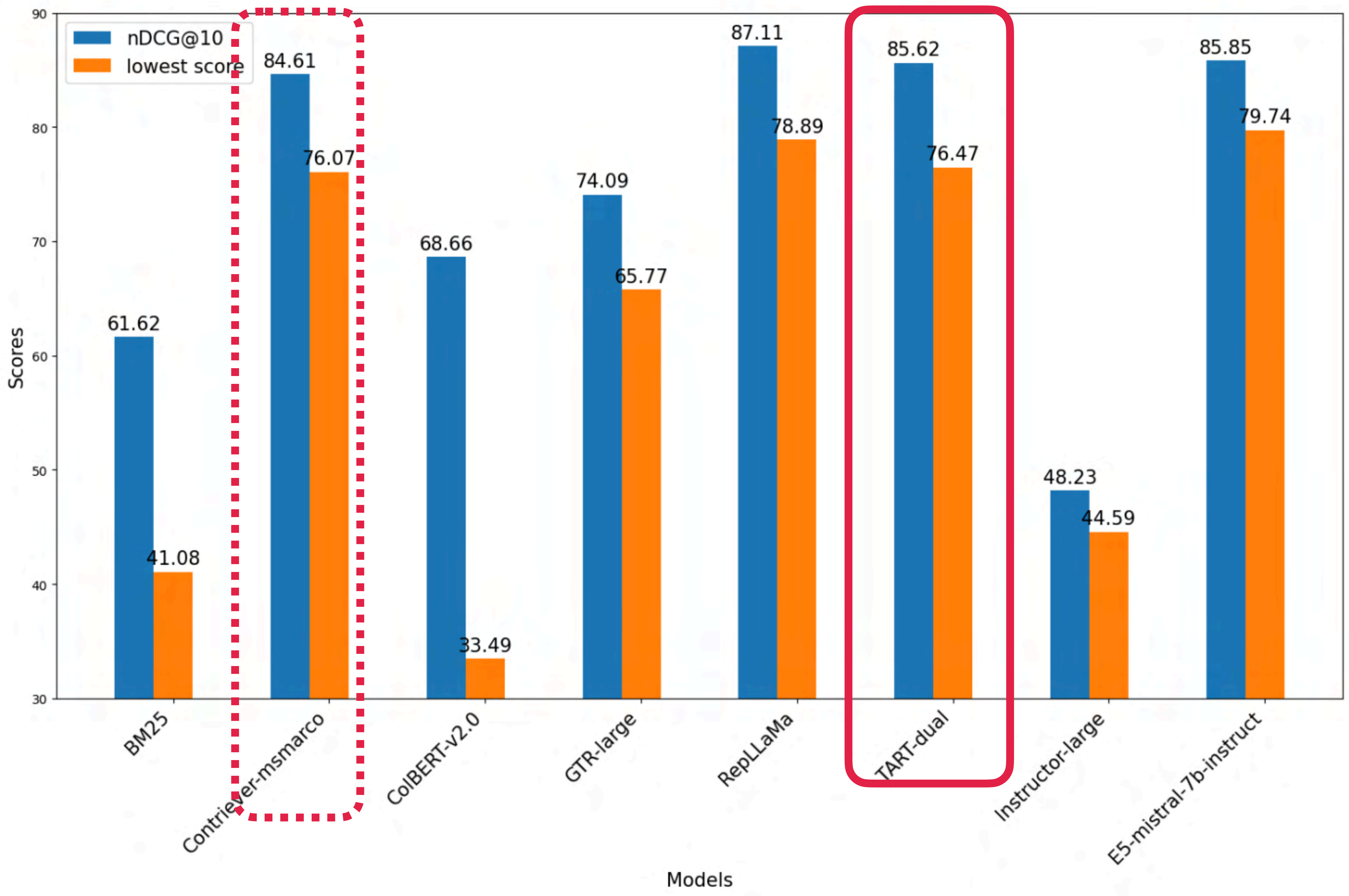
Are those retrievers following instructions?

Instructir (Oh et al., 2024)

Create a natural instruction-following retrieval dataset with LLMs



Improvements from base retrievers are limited



Are those retrievers following instructions?

FollowIR (Welleer et al., 2024)

	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.

All small BE models struggles to follow instructions

Are those retrievers following instructions?

FollowIR (Welleer et al., 2024)

	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.

CE models or 7B BE models have shown superiority

Introduce significant inference latency!

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

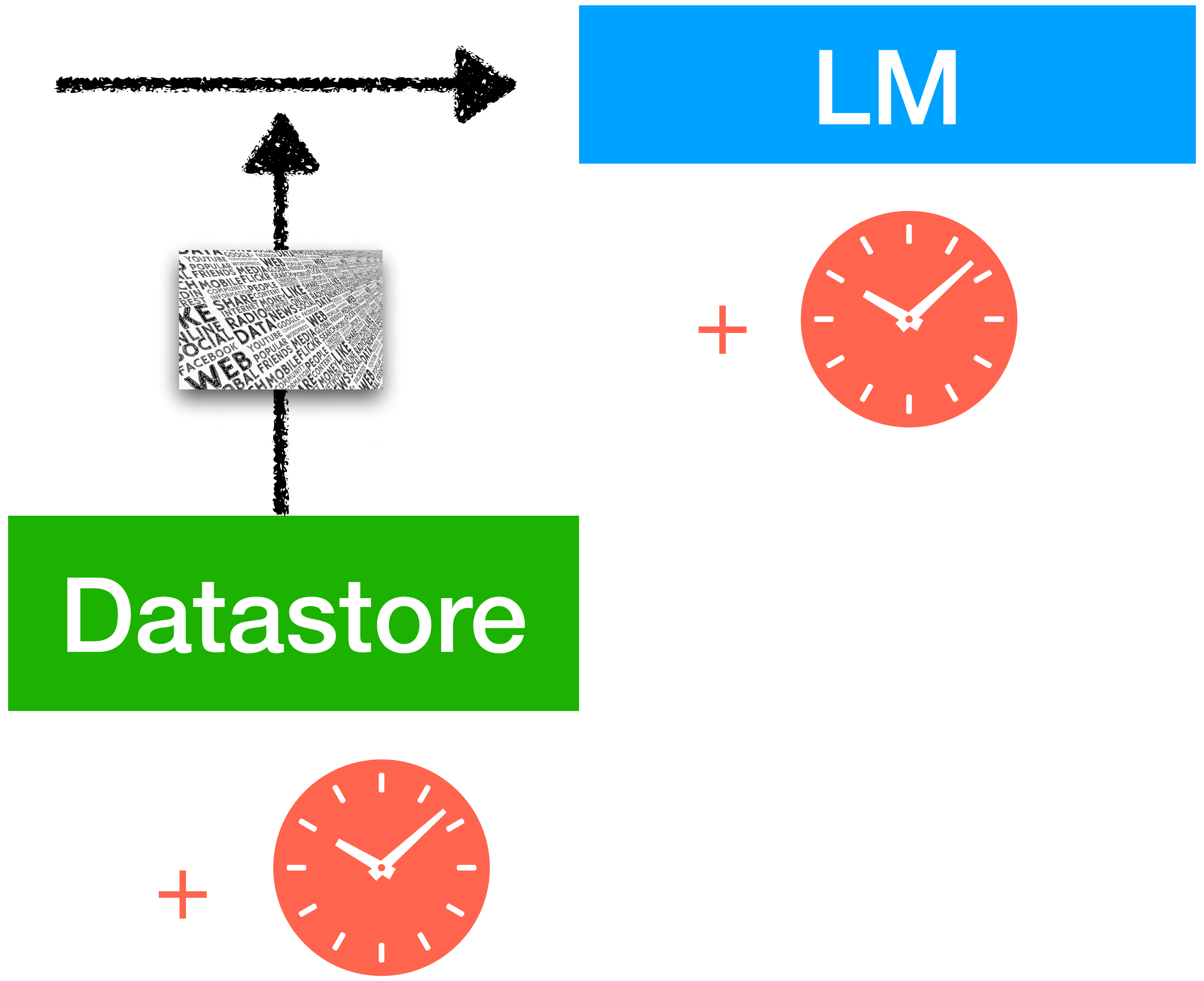
- **Advancing RAG** (Asai et al., 2024; Asai et al., 2023)
  - **Self-RAG** to build versatile retrieval-augmented LMs addressing issues in RAG
  - **Task-aware retrievals** to build versatile RAG systems

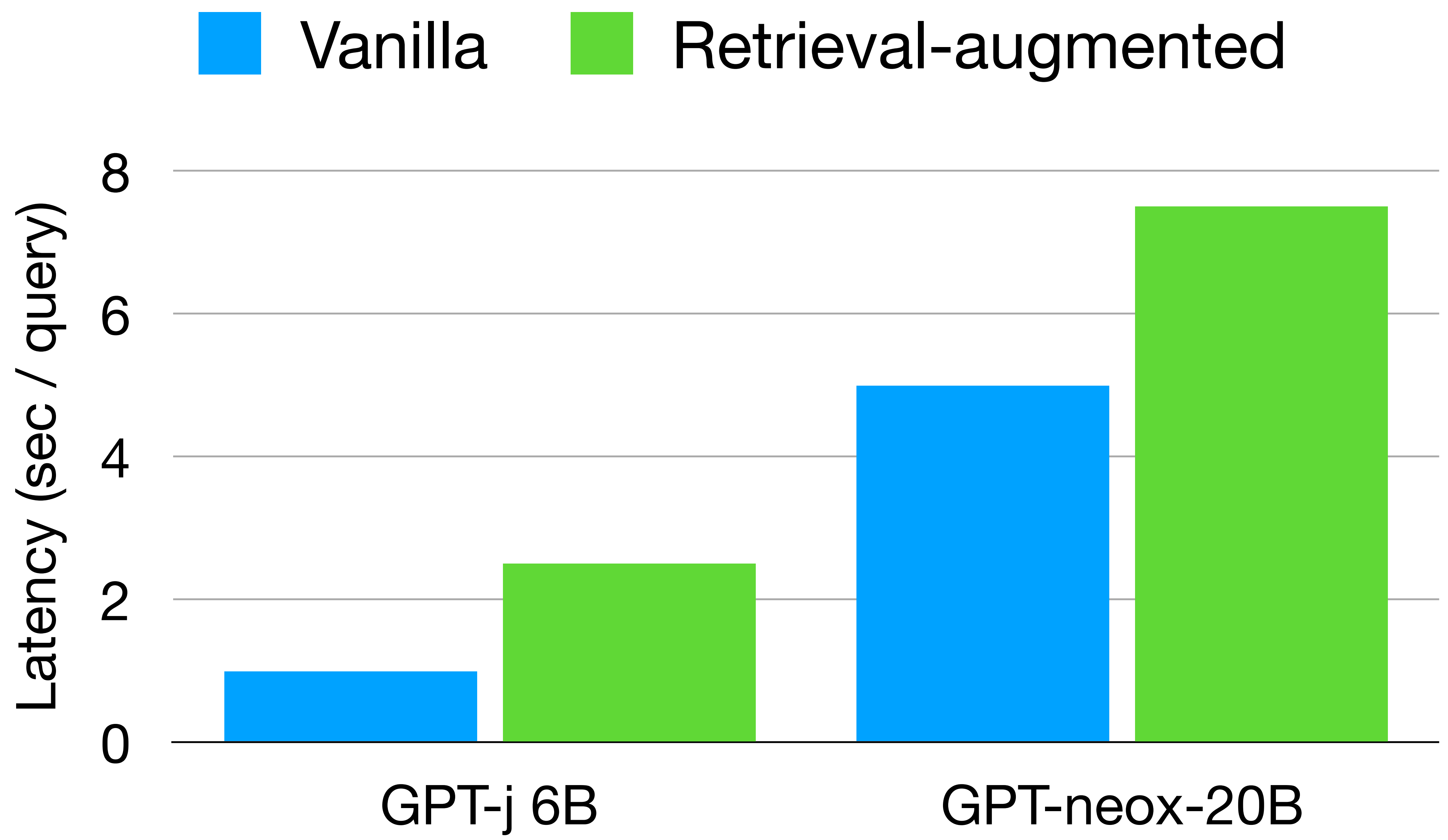
Reliable RAG in the wild: improving efficiency

Efficiency

Effective datastore

New application



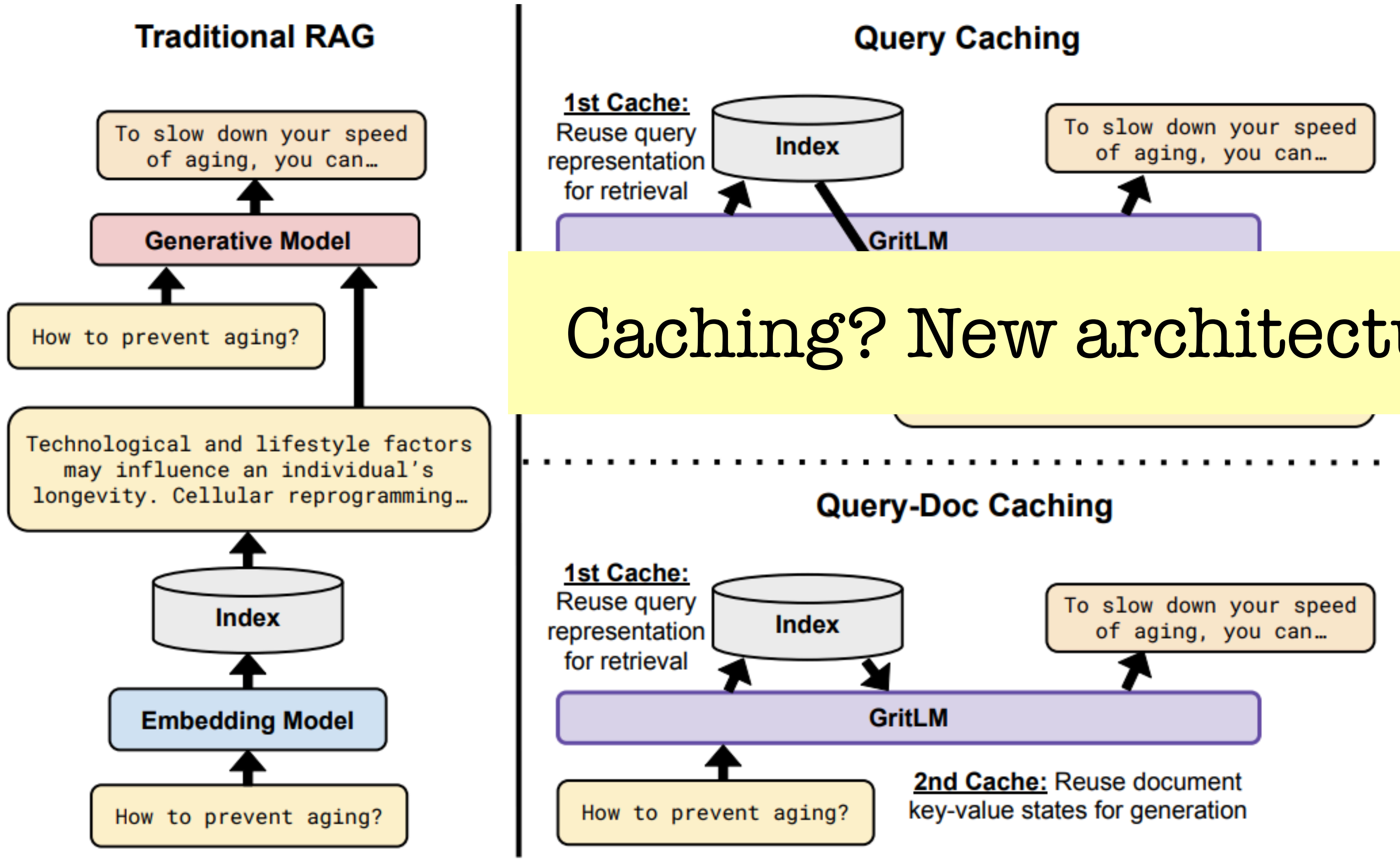


Mallen*, Asai* et al., When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories 2023.

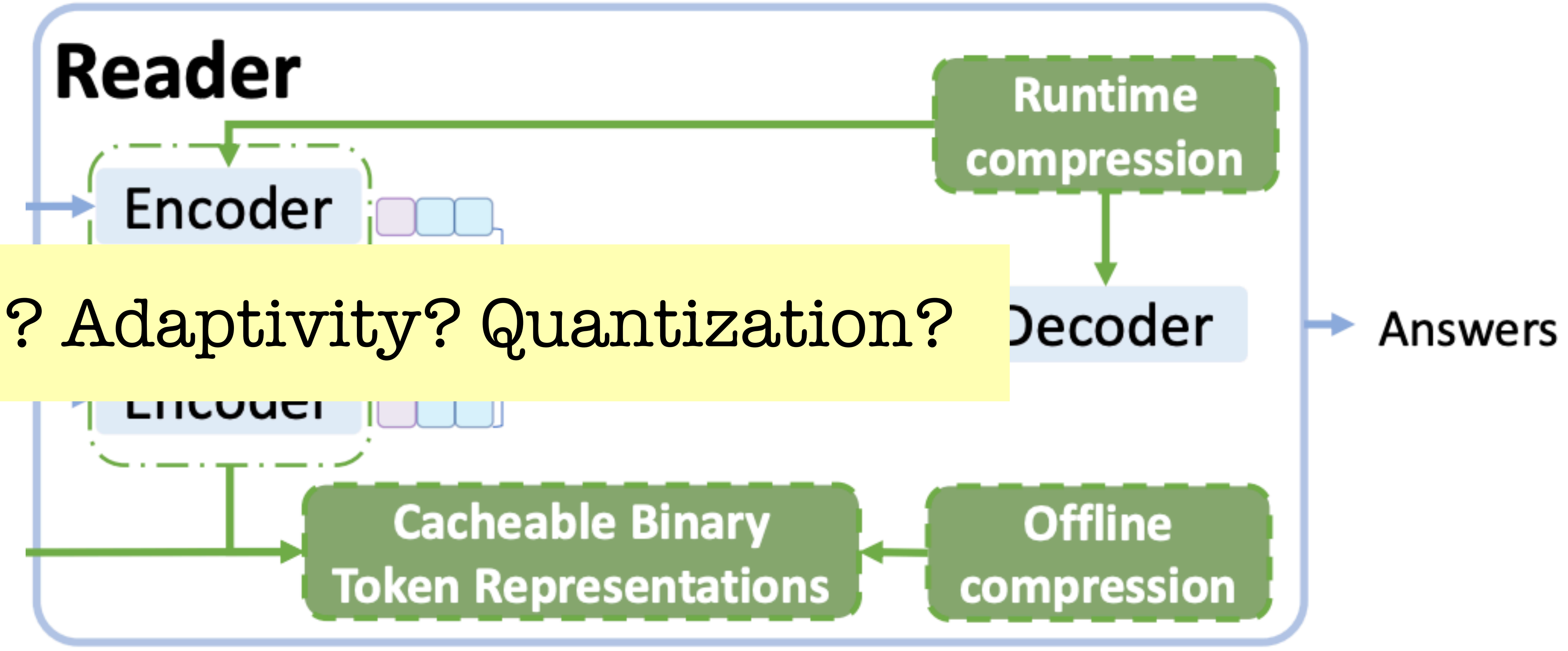
Reliable RAG in the wild: efficient algorithms / arc. for RAG

Efficiency | Effective datastore | New application

GRIT
(Muennighoff et al., 2024)



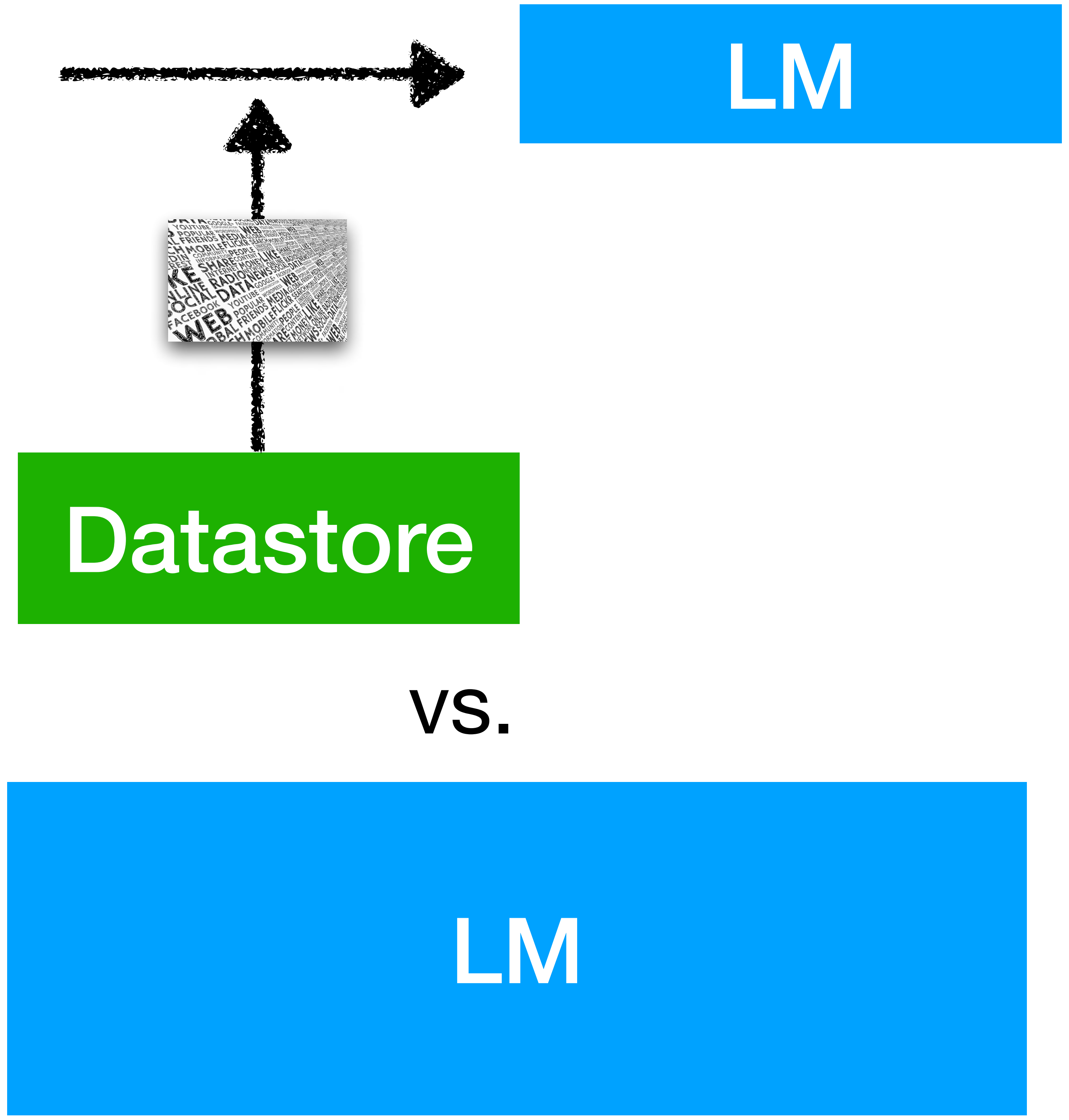
BTR
(Cao et al., 2024)

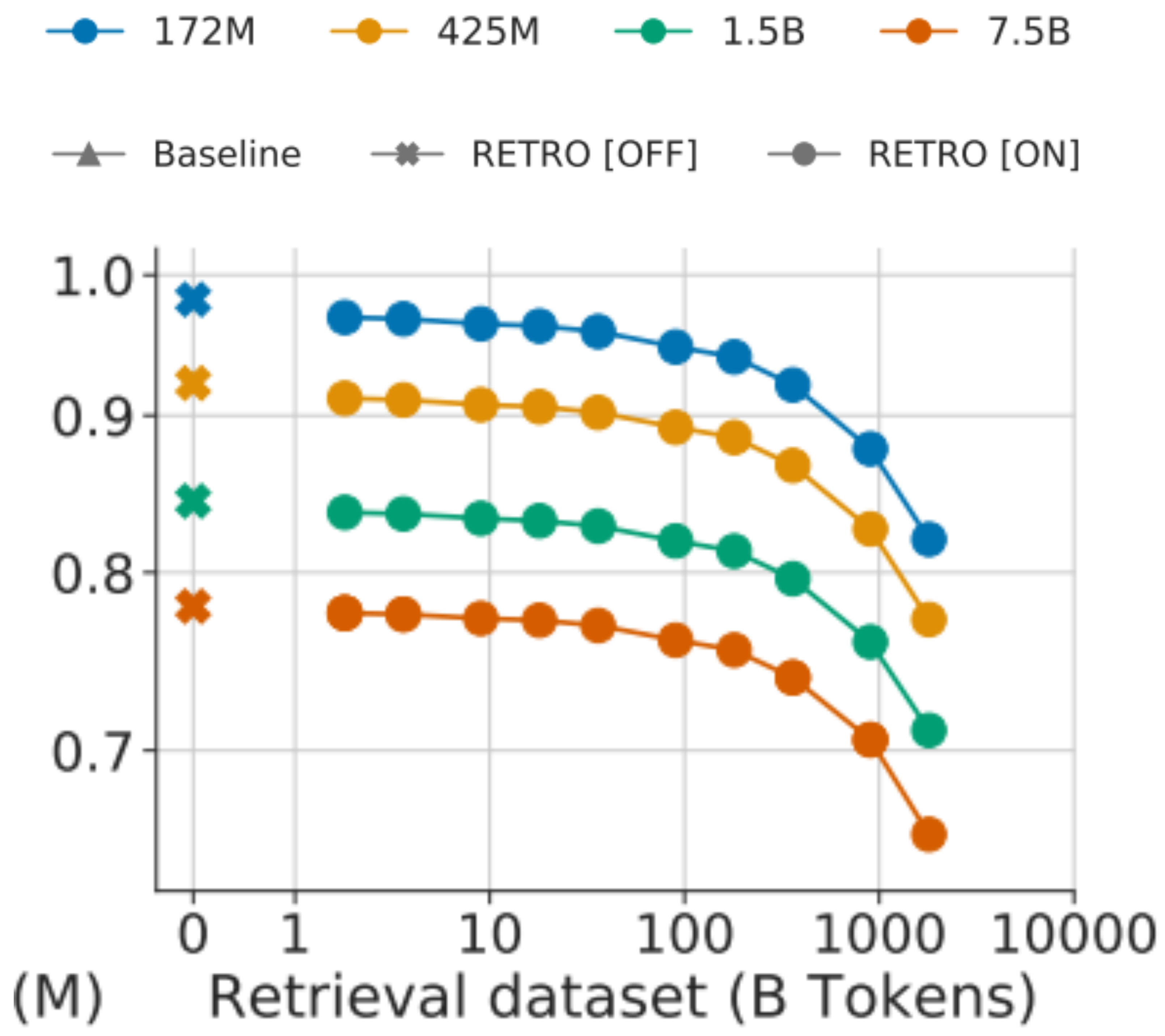


Caching? New architecture? Adaptivity? Quantization?

Reliable RAG in the wild: effective datastore (scale)

Efficiency | **Effective datastore** | New application





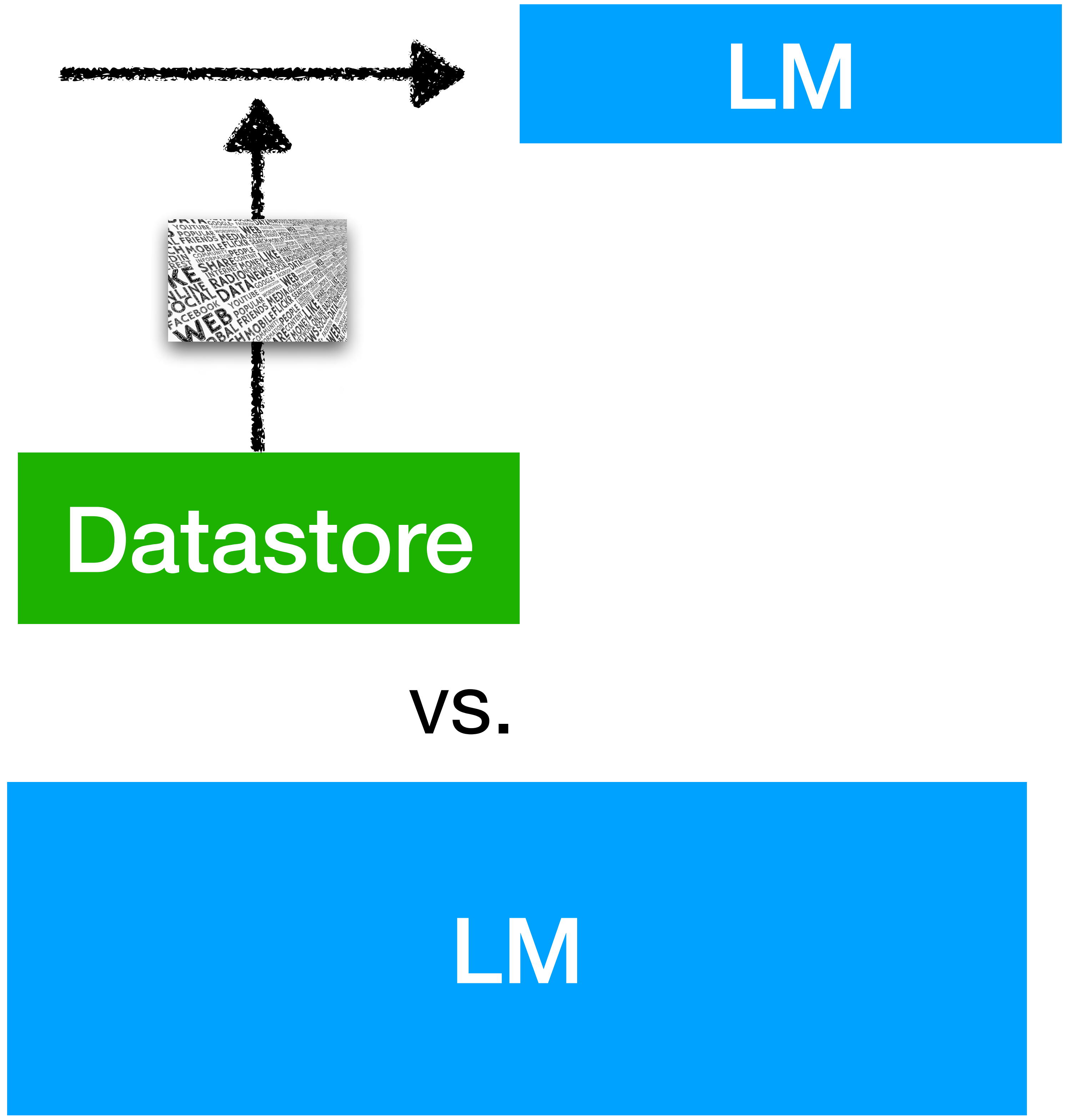
RETRO
(Borgeaud et al., 2021)

Reliable RAG in the wild: effective datastore (scale)

Efficiency

Effective datastore

New application



How should we scale RAG?

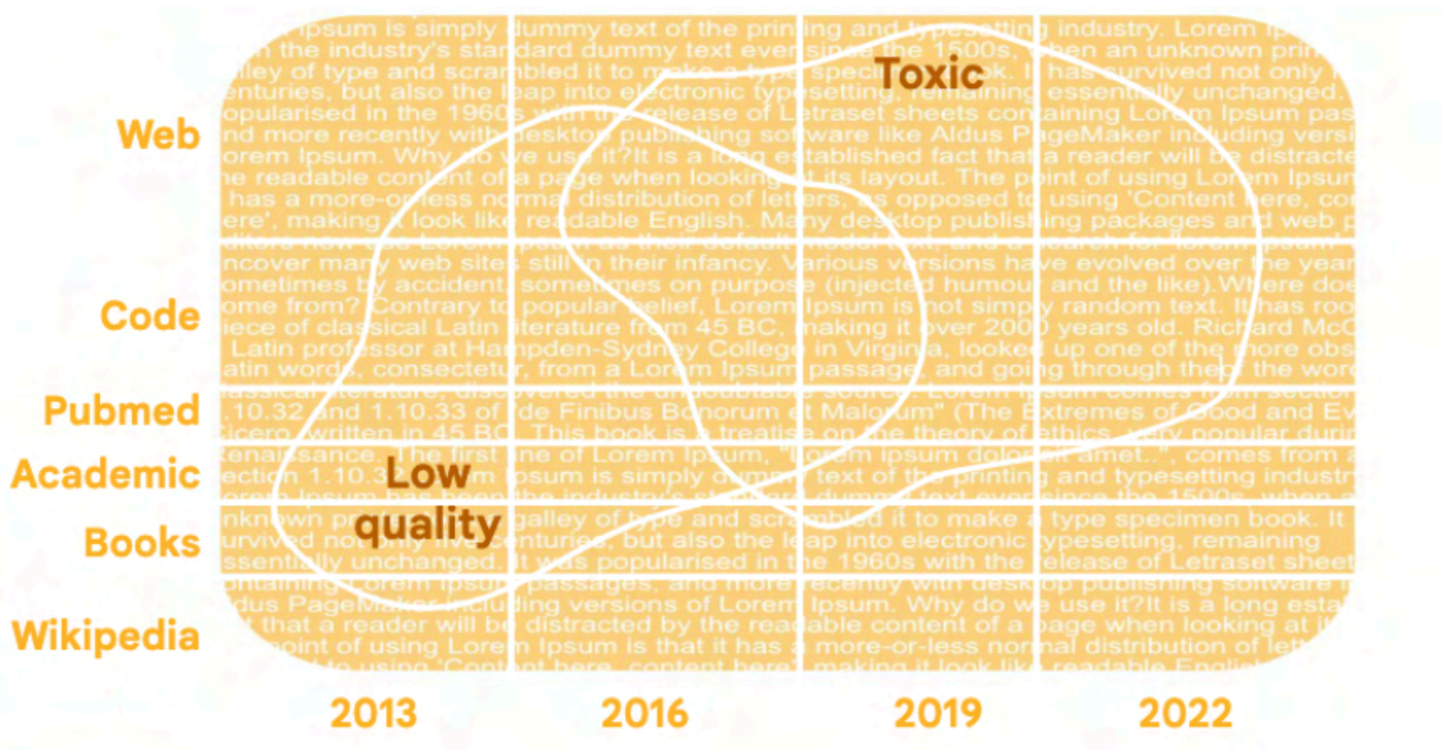
	LM # of parameters	Datastore # of tokens
kNN-LM (Khandelwal et al., 2020)	250M	≤ 3B
NPM (Min et al., 2023)	350M	1B
Atlas (Izacard et al., 2022)	11B	~30B
RETRO (Borgeaud et al., 2021)	7B	2T
REPLUG (Shi et al., 2023)	≤ 175B	~5B

Reliable RAG in the wild: effective datastore (quality)

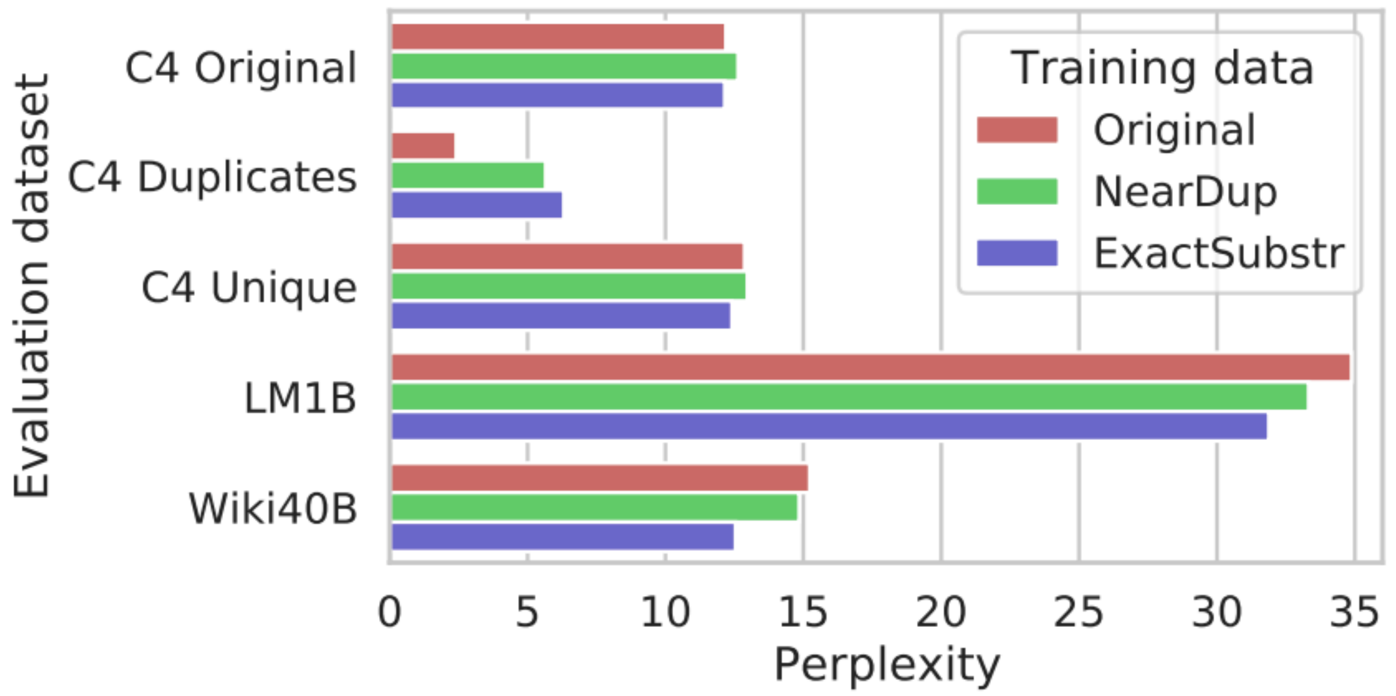
Efficiency | **Effective datastore** | New application

Data-centric approaches to build effective large-scale datastore?

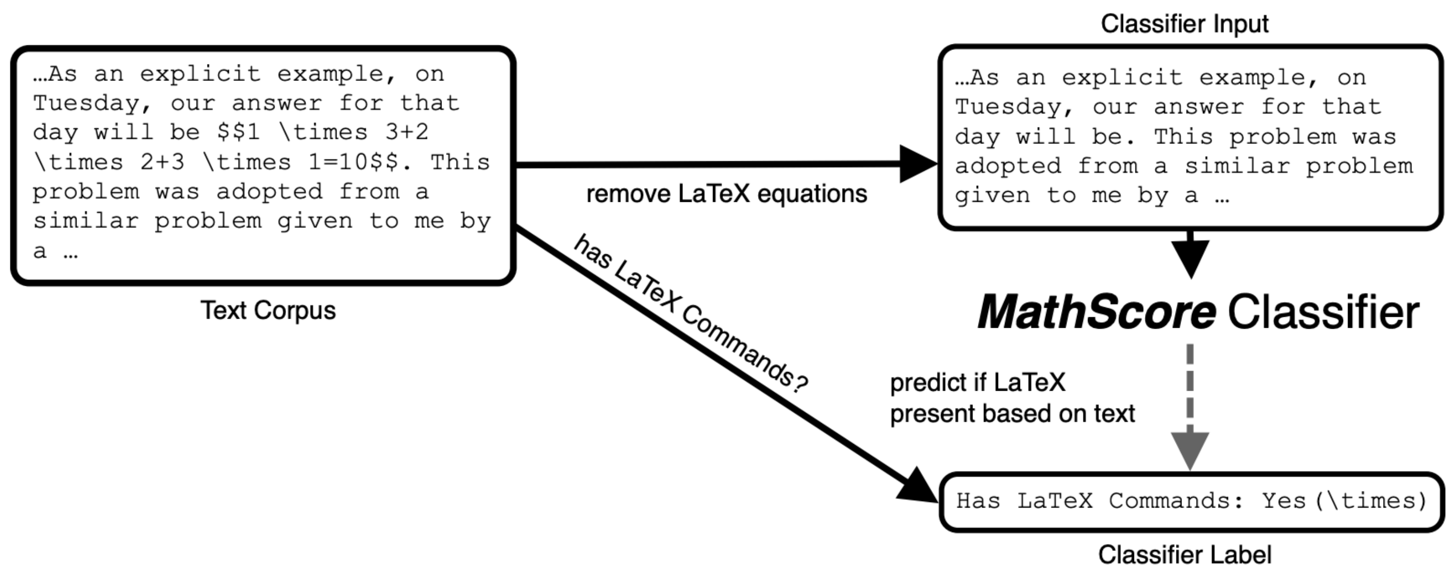
Quality & Composition
(Longpre et al., 2023)



Deduplication
(Lee et al., 2023)



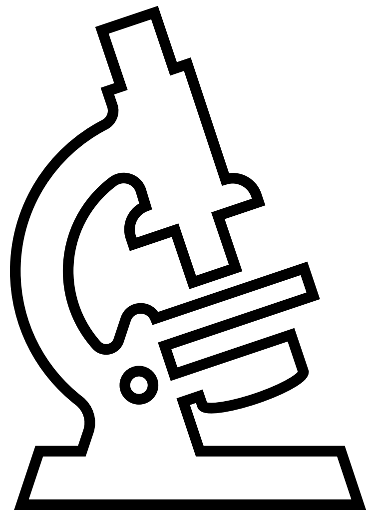
Data Filtering
(Paster et al., 2023)

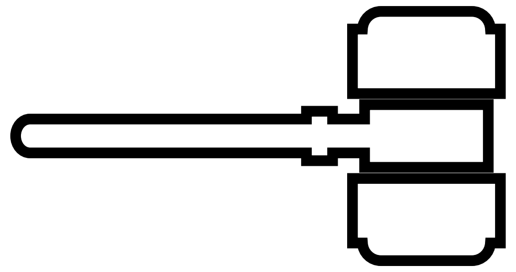


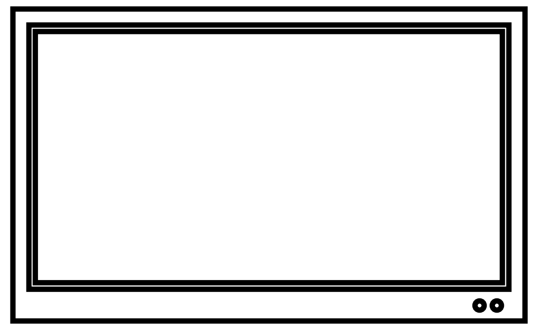
Reliable RAG in the wild: scaling datastore

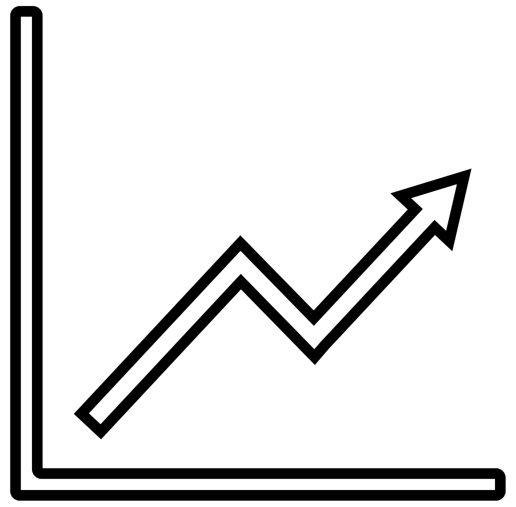
Efficiency | Scaling datastores | **New application**


Medical


Science


Legal

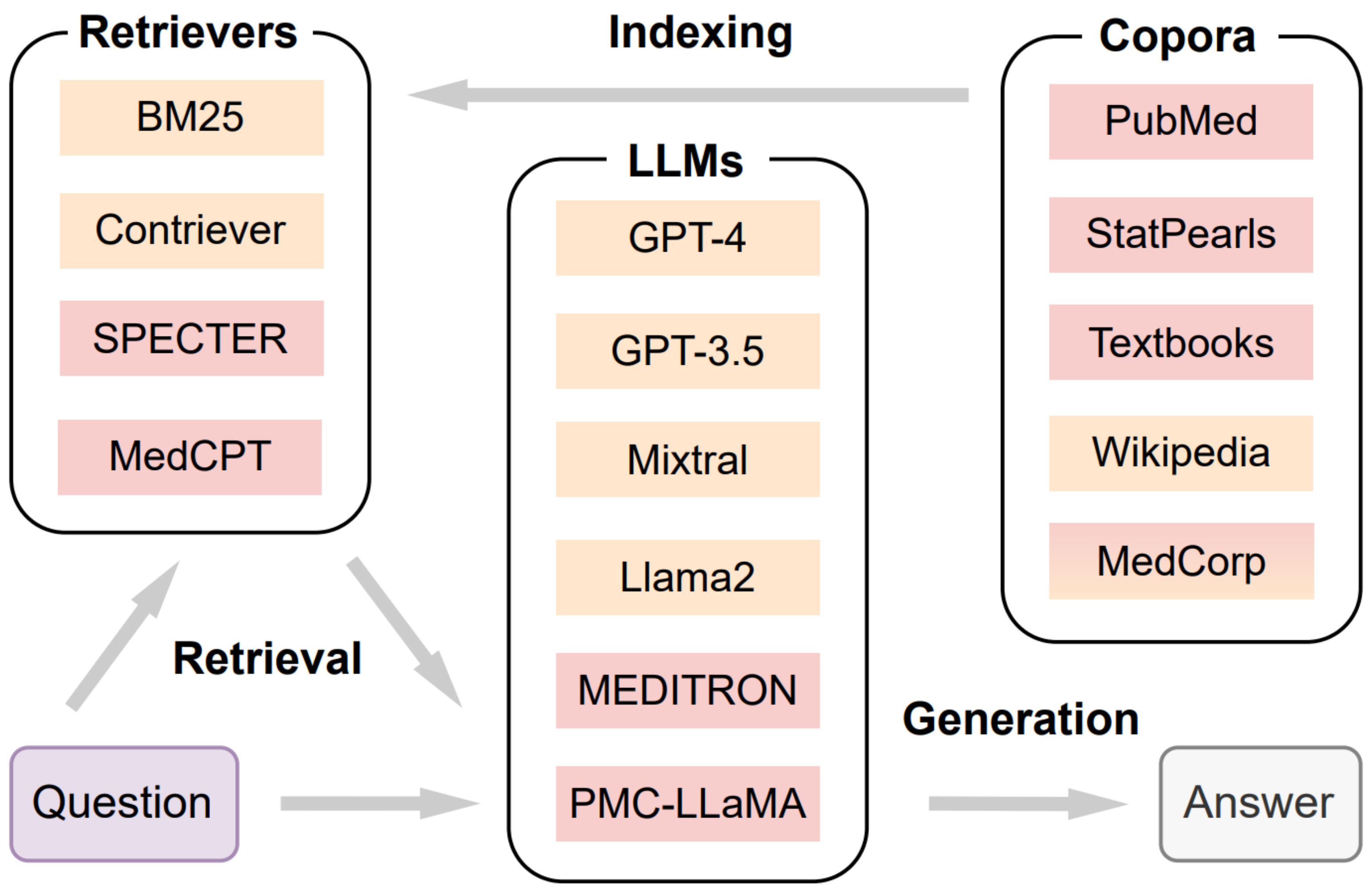

Code


Finance

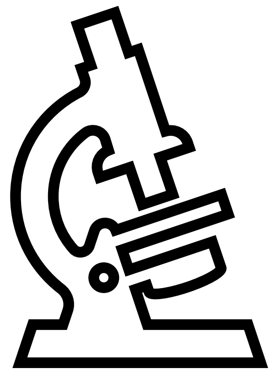
Reliable RAG in the wild: scaling datastore

Efficiency | Scaling datastores | **New application**

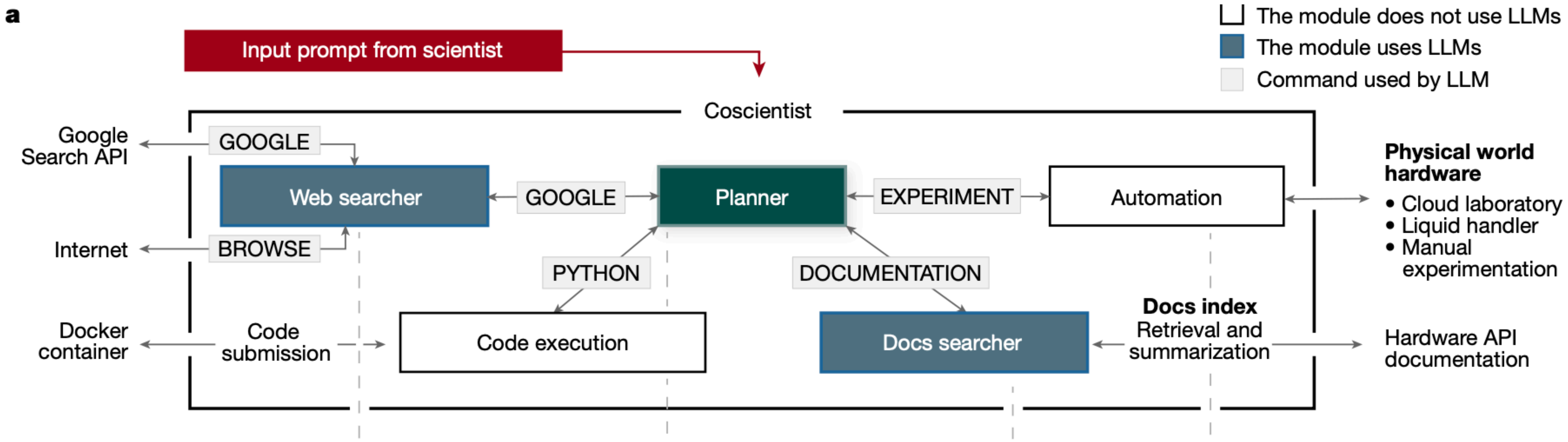
MedRAG
(Xiong et al., 2024)



Realistic evaluation?



Automous chemistry experiments
(Boiko et al., 2024)



RAG pipelines optimized for expert tasks?

Thanks for listening :)

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

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

124