

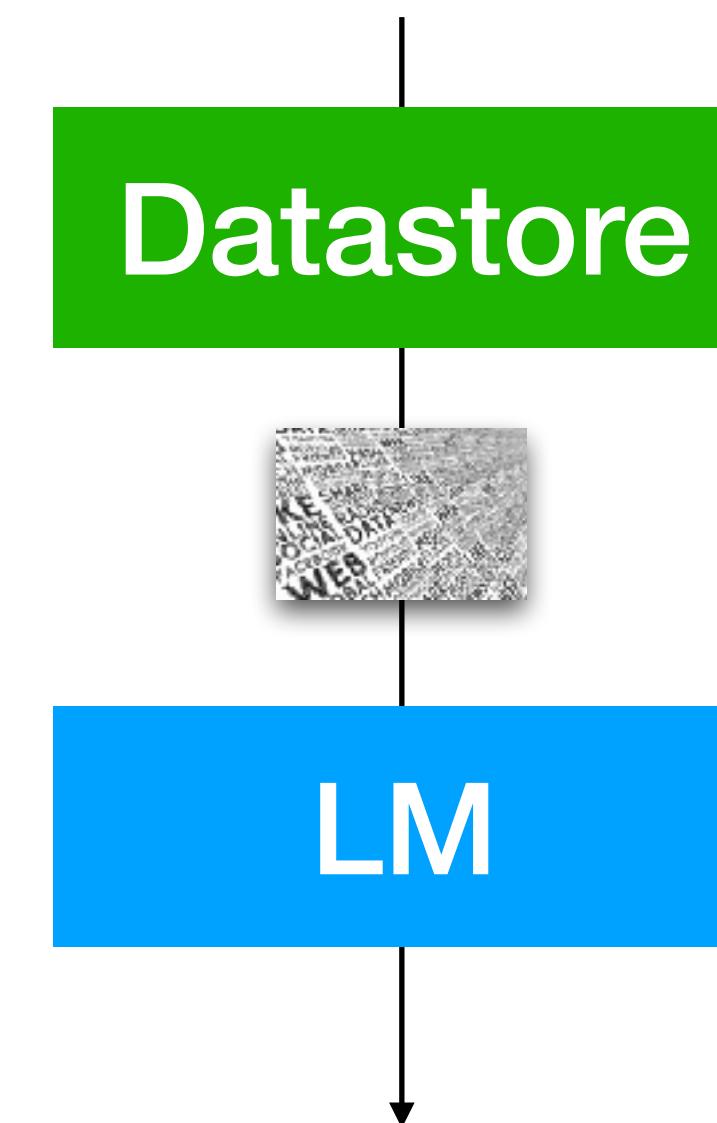
# Section 5: Applications

# Downstream adaptation of retrieval-based LMs

What are the **tasks**?

**Language modeling**

The capital city of Ontario is \_\_



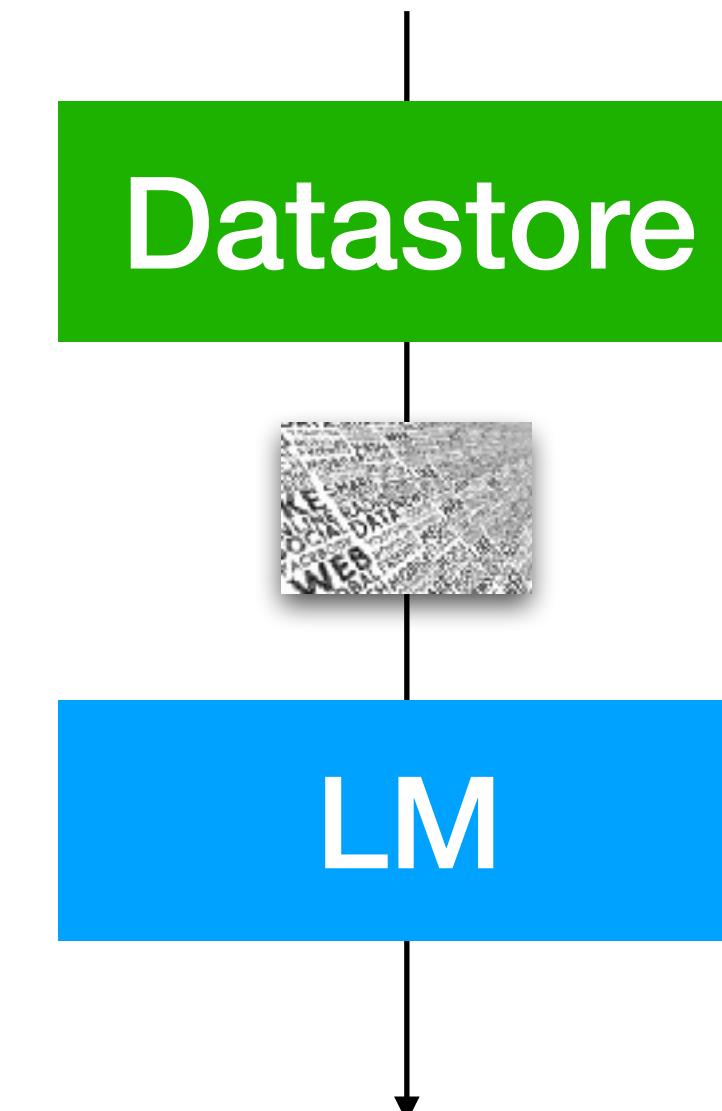
**Toronto, which is known for ...**

# Downstream adaptation of retrieval-based LMs

What are the **tasks**?

**Open-domain QA**

What is the capital of Ontario?



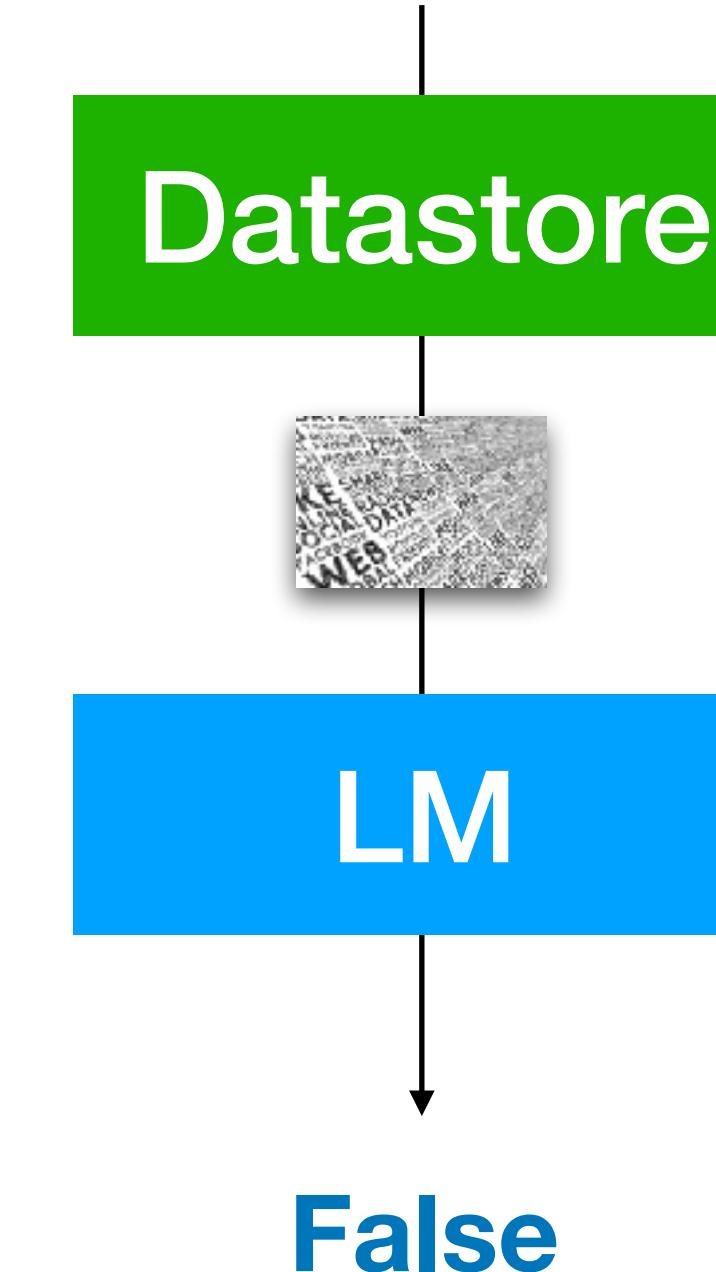
**Toronto**

# Downstream adaptation of retrieval-based LMs

What are the **tasks**?

**Fact Verification**

Is the following statement true? Ottawa is the Ontario state capital.



# A range of target tasks

## Question Answering

DPR (Karpukhin et al, 2020)

RAG (Lewis et al, 2020)

REALM (Gu et al, 2020)

Many earlier retrieval-based LMs have been evaluated in the area of open-domain QA.

# A range of target tasks

## Question Answering

DPR (Karpukhin et al, 2020)  
RAG (Lewis et al, 2020)  
REALM (Gu et al, 2020)

## Fact Verification

RAG (Lewis et al, 2020)  
ATLAS (Izacard et al, 2022)  
Evi. Generator (Asai et al, 2022)

## Dialogue

BlenderBot3 (Shuster et al., 2022)  
Internet-augmented generation  
(Komeili et a., 2022)

For a while, mainly evaluated on knowledge-intensive tasks  
(Lewis et al., 2020; Petroni et al., 2021)

# A range of target tasks

## Question Answering

DPR (Karpukhin et al, 2020)  
RAG (Lewis et al, 2020)  
REALM (Gu et al, 2020)

## Fact Verification

RAG (Lewis et al, 2020)  
ATLAS (Izacard et al, 2022)  
Evi. Generator (Asai et al, 2022)

## Dialogue

BlenderBot3 (Shuster et al., 2022)  
Internet-augmented generation  
(Komeili et a., 2022)

## Summarization

FLARE (Jiang et al, 2023)

## Machine Translation

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

## Language Modeling

kNN-LM (Khandelwal et al., 2020)  
TRIME (Zhong et al., 2022)  
RETRO (Borgeaud et al., 2021)

## NLI

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

## Sentiment Analysis

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

## Commonsense Reasoning

Raco (Yu et al, 2022)

More general NLU tasks

# A range of target tasks

## Question Answering

DPR (Karpukhin et al, 2020)  
RAG (Lewis et al, 2020)  
REALM (Gu et al, 2020)

## Fact Verification

RAG (Lewis et al, 2020)  
ATLAS (Izacard et al, 2022)  
Evi. Generator (Asai et al, 2022)

## Dialogue

BlenderBot3 (Shuster et al., 2022)  
Internet-augmented generation  
(Komeili et a., 2022)

## Summarization

FLARE (Jiang et al, 2023)

## Machine Translation

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

## Language Modeling

kNN-LM (Khandelwal et al., 2020)  
TRIME (Zhong et al., 2022)  
RETRO (Borgeaud et al., 2021)

## NLI

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

## Sentiment Analysis

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

## Commonsense Reasoning

Raco (Yu et al, 2022)

More generations

# A range of target tasks

## Question Answering

DPR (Karpukhin et al, 2020)  
RAG (Lewis et al, 2020)  
REALM (Gu et al, 2020)

## Fact Verification

RAG (Lewis et al, 2020)  
ATLAS (Izacard et al, 2022)  
Evi. Generator (Asai et al, 2022)

## Dialogue

BlenderBot3 (Shuster et al., 2022)  
Internet-augmented generation  
(Komeili et a., 2022)

## Summarization

FLARE (Jiang et al, 2023)

## Machine Translation

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

## Language Modeling

kNN-LM (Khandelwal et al., 2020)  
TRIME (Zhong et al., 2022)  
RETRO (Borgeaud et al., 2021)

## NLI

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

## Sentiment Analysis

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

## Commonsense Reasoning

Raco (Yu et al, 2022)

More classifications

# Real-world applications of retrieval-based LMs

The screenshot shows a conversational AI interface. At the top, a search bar contains the query: "Create an itinerary for exploring Toronto over a span of two days." Below the search bar is a response card. The card features a blue speech bubble icon with a white 'b' and a blue thumbs-up icon. To the right of the icon are three small buttons: a thumbs-up, a thumbs-down, and an ellipsis. The main text in the card reads: "Toronto is a vibrant and diverse city with many attractions and activities to suit different tastes and preferences. Here are some suggestions for creating an itinerary for exploring Toronto over two days:"

- **Day 1:** Start your day with a visit to the **CN Tower**, the iconic landmark that offers panoramic views of the city and beyond <sup>1</sup>. You can also experience the thrill of walking on the glass floor or the EdgeWalk, the world's highest hands-free walk <sup>1</sup>. Next, head to the nearby **St. Lawrence Market**, one of the world's best food markets, where you can sample a variety of cuisines and local specialties <sup>2</sup>. After lunch, take a stroll along **Queen West**, a trendy neighborhood with eclectic shops, galleries, cafes and street art <sup>3</sup>. In the evening, enjoy a show at one of the many theaters or comedy clubs in the **Entertainment District**, or catch a game at the **Scotiabank Arena** if you're a sports fan.
- **Day 2:** Explore the history and culture of Toronto at the **Royal Ontario Museum**, one of the largest and most comprehensive museums in North America, featuring exhibits on art, natural history, world cultures and more <sup>4</sup>. Then, hop on a ferry to the **Toronto Islands**, a group of islands that offer a relaxing escape from the city, with beaches, parks, trails and amusement rides <sup>3</sup> <sup>5</sup>. You can also rent a bike or kayak to explore the islands at your own pace. For dinner, head to **Chinatown**, one of the largest and most vibrant in North America, where you can find a variety of Asian cuisines and shops <sup>3</sup>.

I hope this helps you plan your trip to Toronto. Have fun! 😊

**Learn more:**

- 1. [cntower.ca](http://cntower.ca)
- 2. [travel.usnews.com](http://travel.usnews.com)
- 3. [bing.com](http://bing.com)
- 4. [rom.on.ca](http://rom.on.ca)
- 5. [tripadvisor.com](http://tripadvisor.com)

# Two key questions for downstream adaptations

**How** can we adapt a retrieval-based LM for a task?

**When** should we use a retrieval-based LM?

# How to adapt a retrieval-based LM for a task

What are the **tasks**?

- Open-domain QA
- Other knowledge-intensive tasks
- Sentiment analysis
- Code generation

...

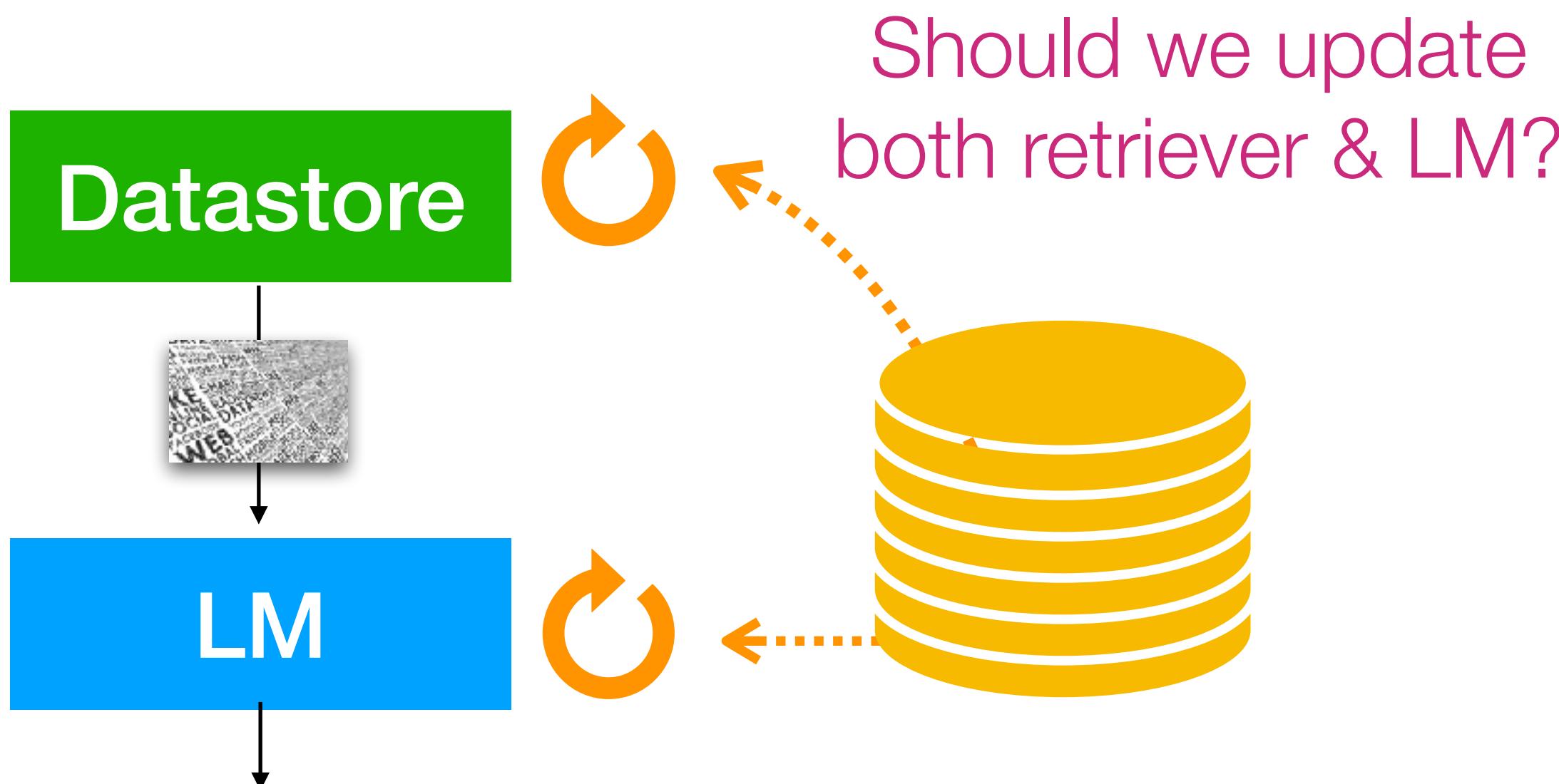
How to **adapt**?

- Supervised fine-tuning
- Reinforcement learning
- Prompting

# How to adapt a retrieval-based LM for a task

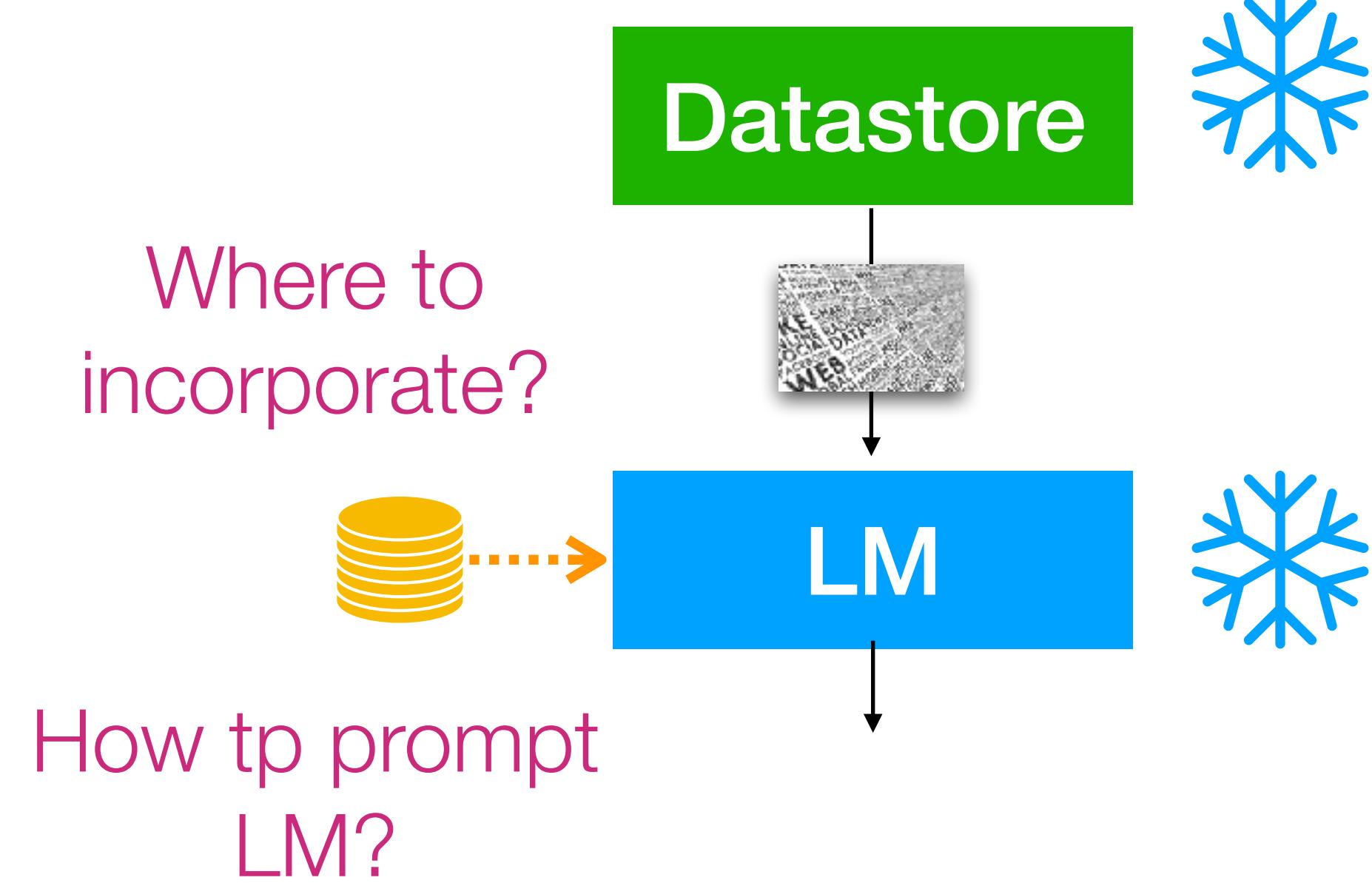
## Fine-tuning (+RL)

Training LM and / or retriever  
on task-data & data store



## Prompting

Prompt a frozen LM with  
retrieved knowledge



# How to adapt a retrieval-based LM for a task

What are the **tasks**?

- Open-domain QA
- Other knowledge-intensive tasks
- Sentiment analysis
- Code generation
- ...

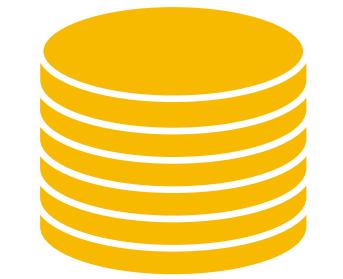
How to **adapt**?

- Supervised fine-tuning
- Reinforcement learning
- Prompting

What is **data store**?



Wikipedia



Training data



Code documentation

# **When** to use a retrieval-based LM

Long-tail

knowledge  
update

Verifiability

Parameter-  
efficiency

Privacy

# Effectiveness of retrieval-based LMs

Long-tail

knowledge  
update

Verifiability

Parameter-  
efficiency

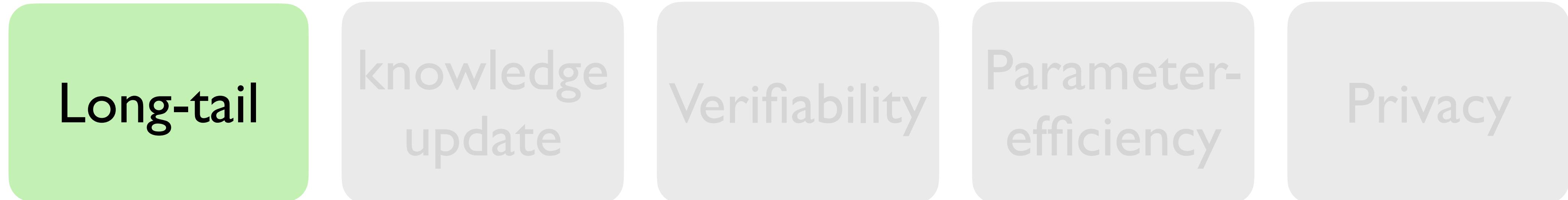
Privacy

**Q:** Is Toronto really  
cold during winter?



Yes it is.

# Effectiveness of retrieval-based LMs



Q: Where is Toronto Zoo located?



1361A Old Finch Avenue,  
in Scarborough, Ontario



# Effectiveness of retrieval-based LMs

Long-tail

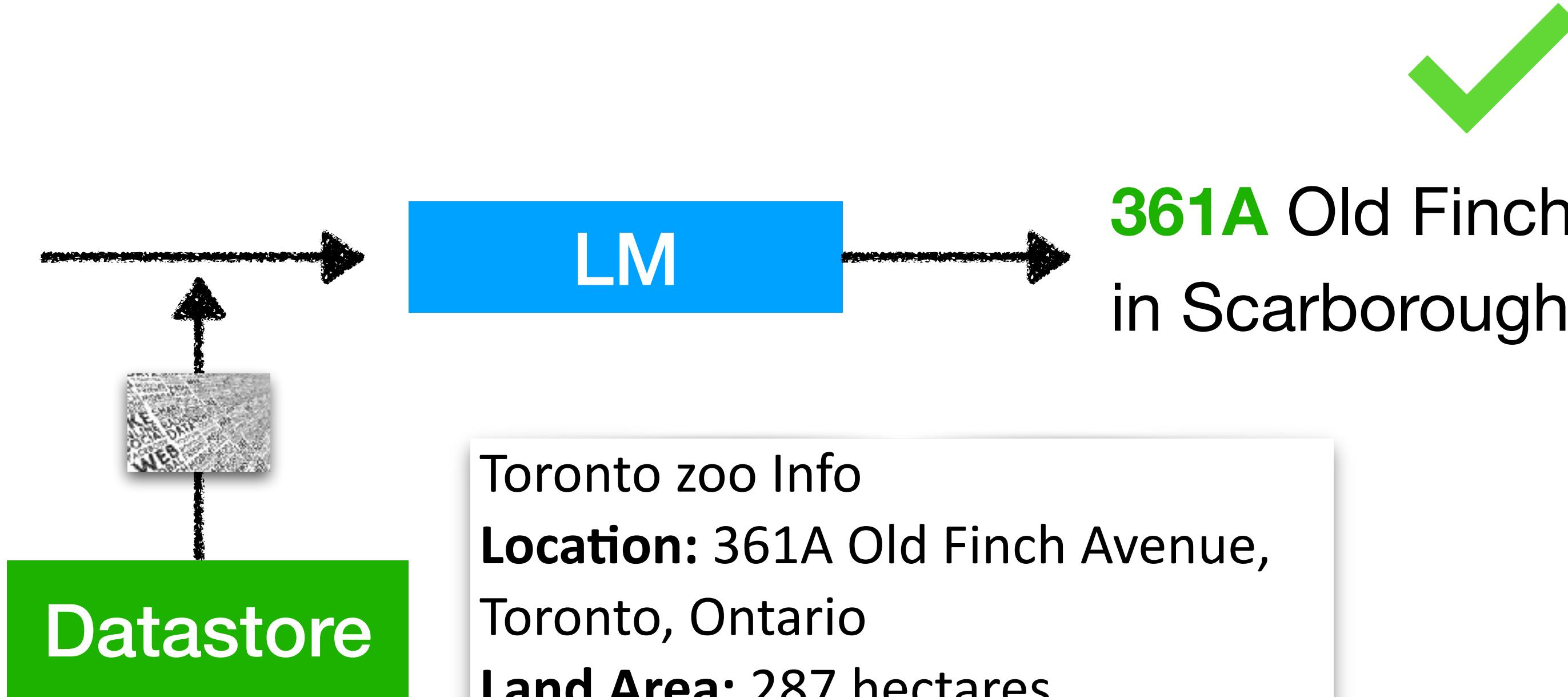
knowledge update

Verifiability

Parameter-efficiency

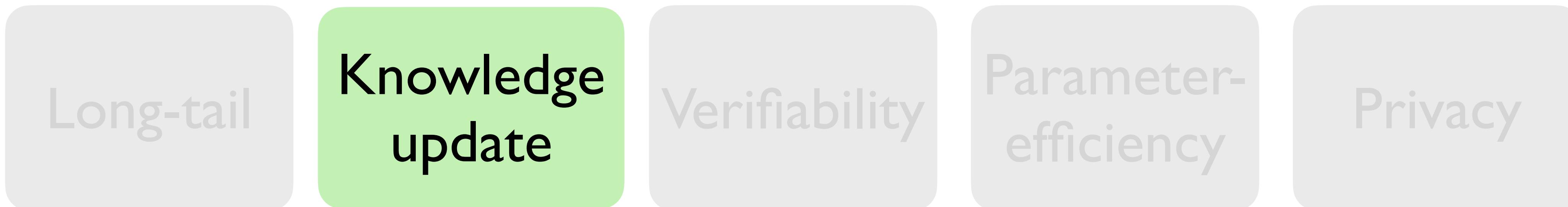
Privacy

Q: Where is Toronto Zoo located?



Toronto zoo Info  
**Location:** 361A Old Finch Avenue,  
Toronto, Ontario  
**Land Area:** 287 hectares

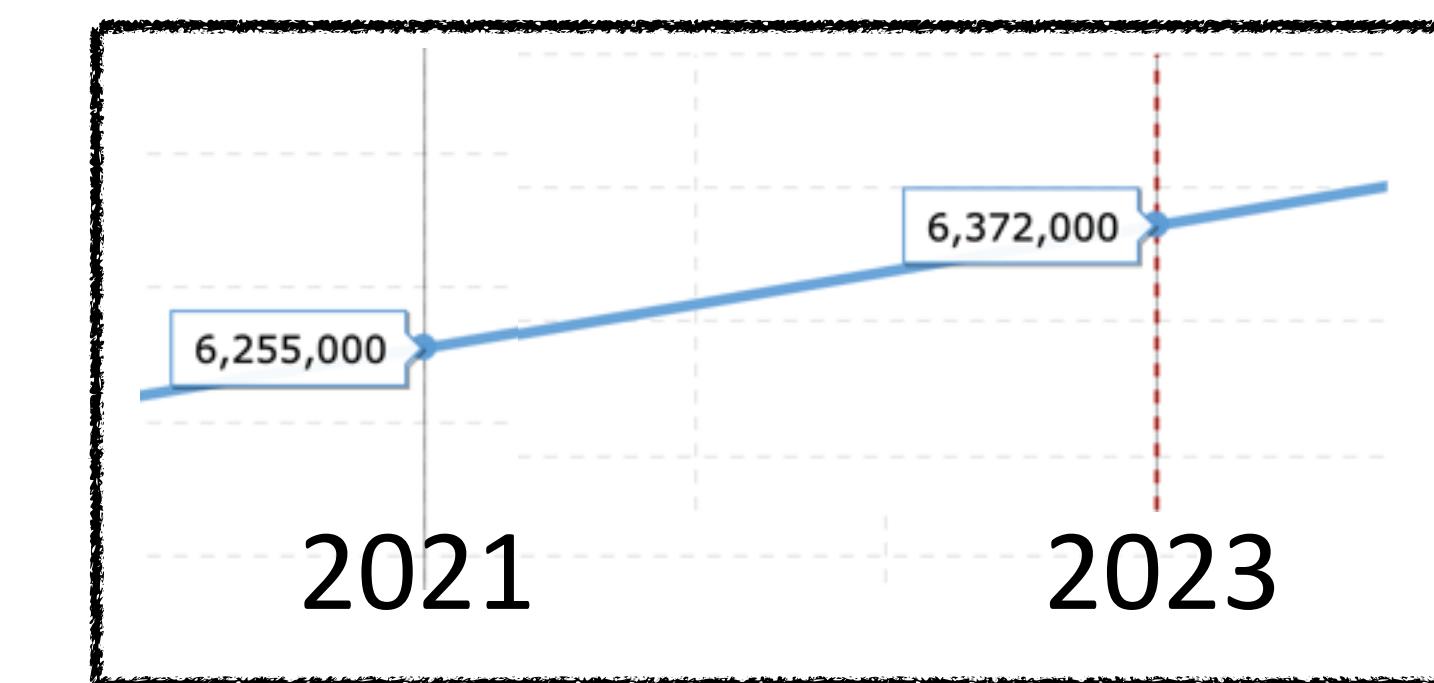
# Effectiveness of retrieval-based LMs



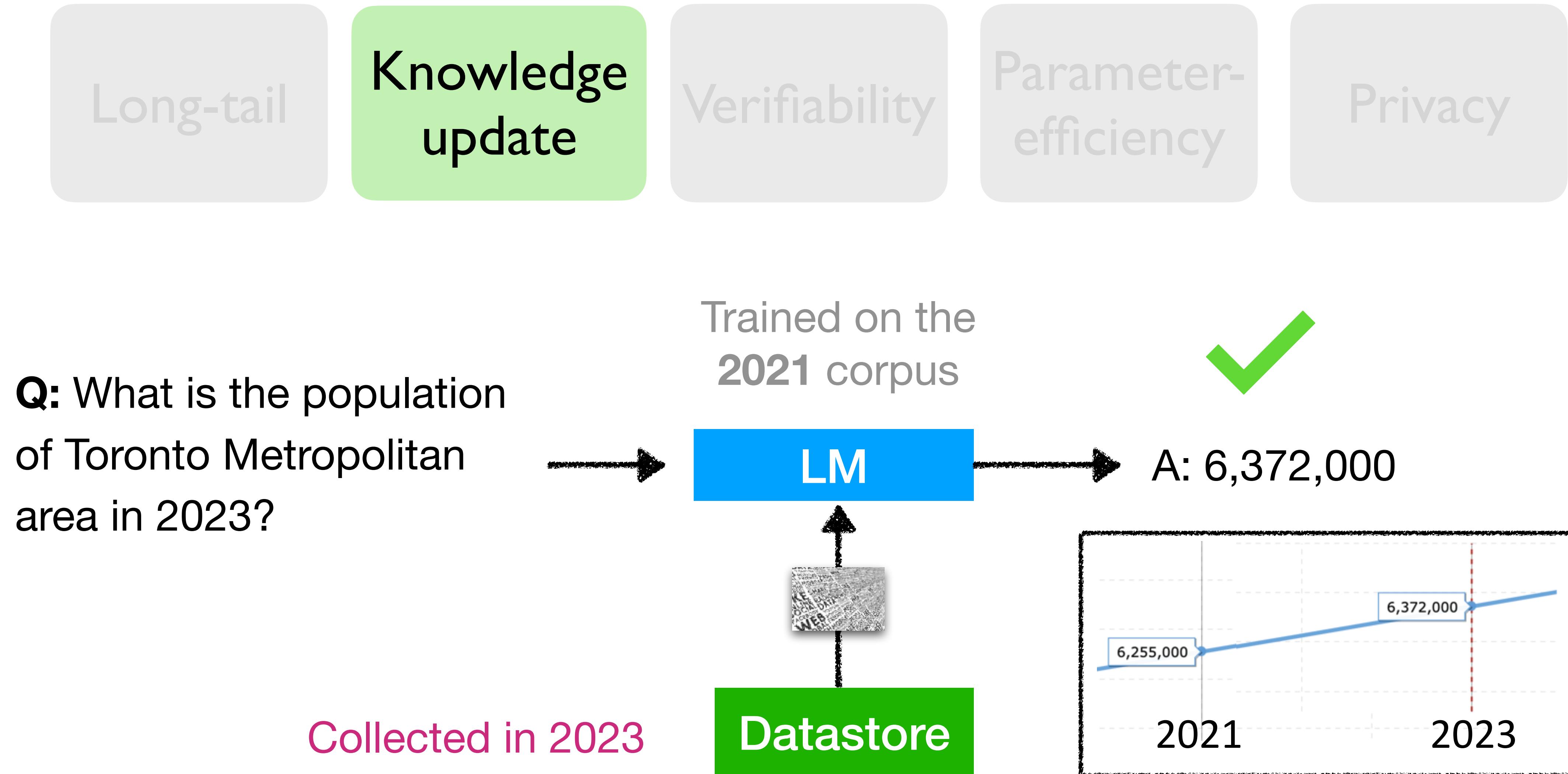
**Q:** What is the population  
of Toronto Metropolitan  
area in 2023?



Trained on the  
2021 corpus



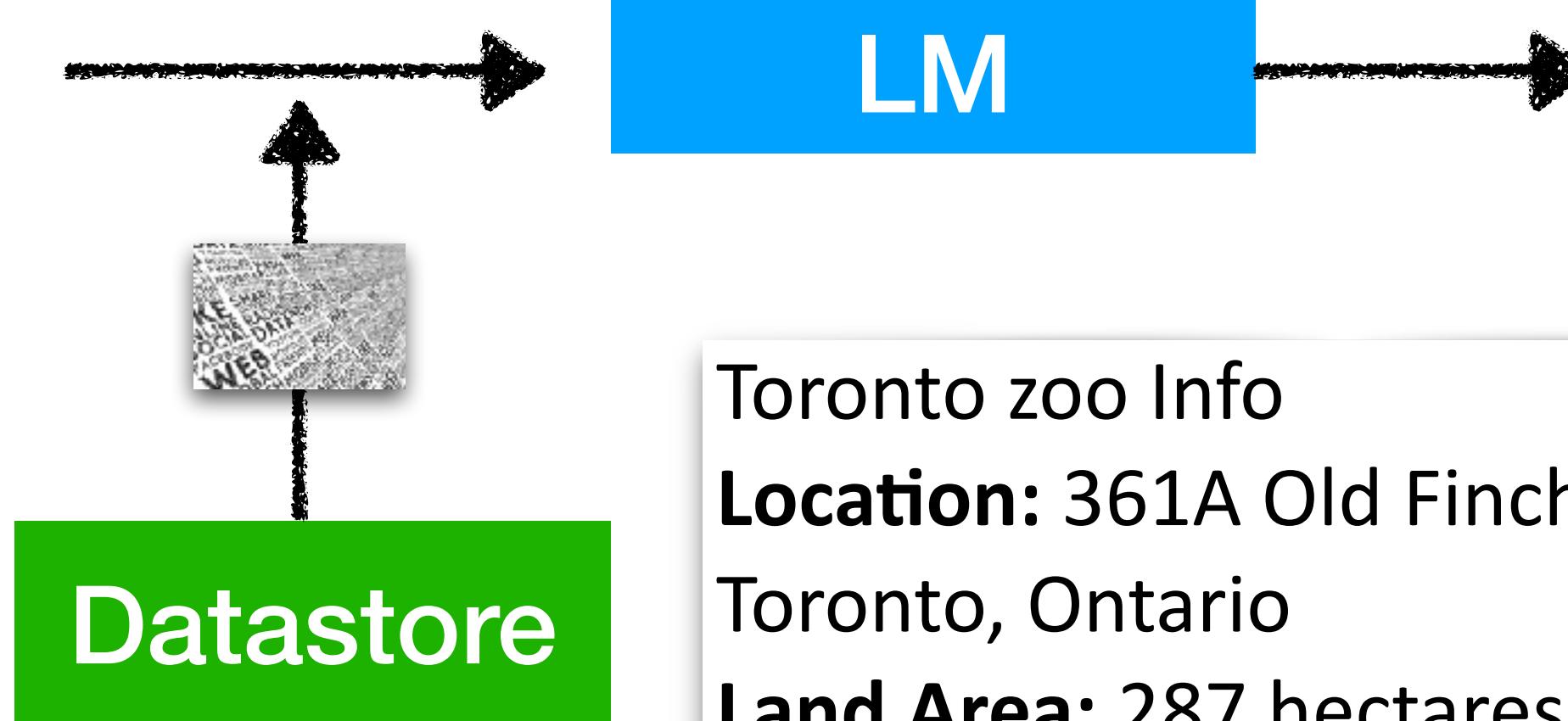
# Effectiveness of retrieval-based LMs



# Effectiveness of retrieval-based LMs



**Q:** Where is Toronto Zoo located?

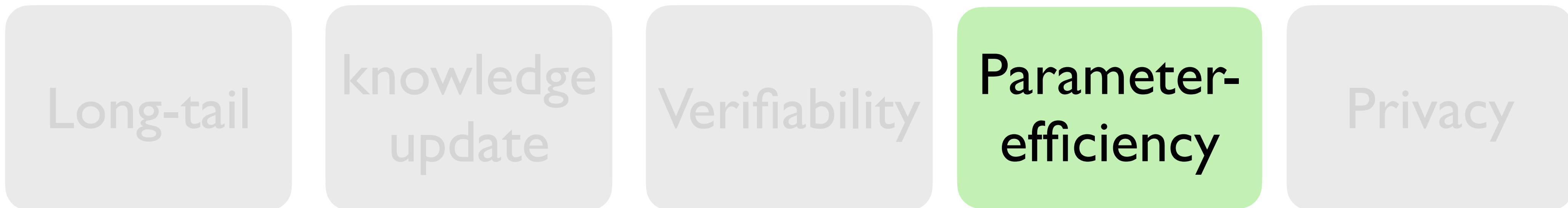


361A Old Finch Avenue,  
in Scarborough, Ontario

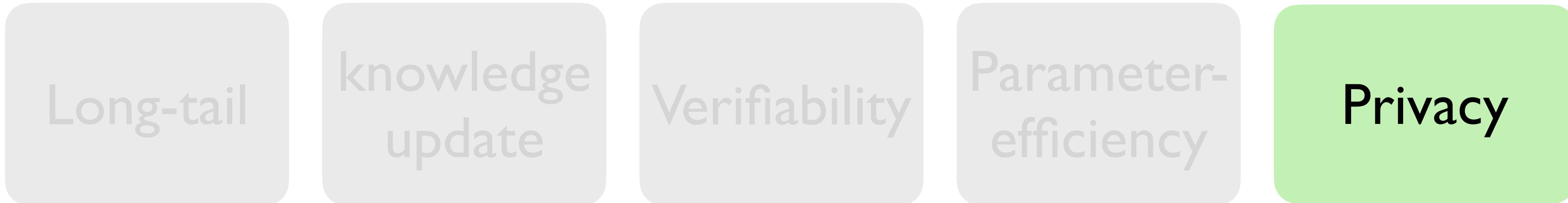
Toronto zoo Info  
**Location:** 361A Old Finch Avenue,  
Toronto, Ontario  
**Land Area:** 287 hectares



# Effectiveness of retrieval-based LMs

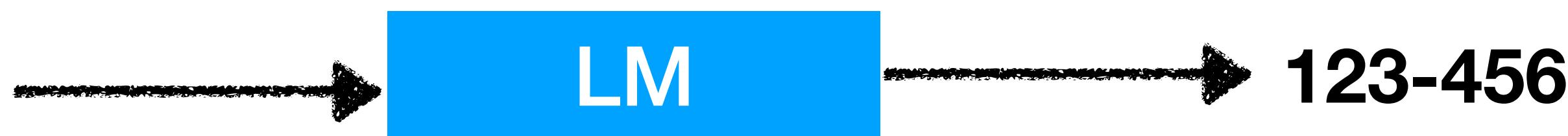


# Effectiveness of retrieval-based LMs



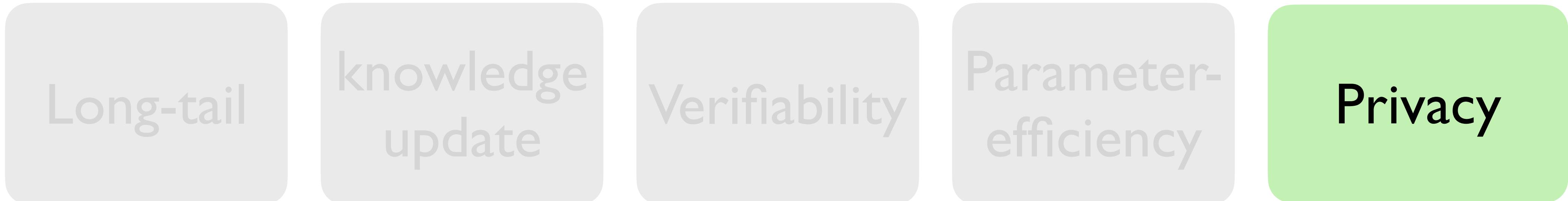
Email: mail@alice.com

Phone: 404-



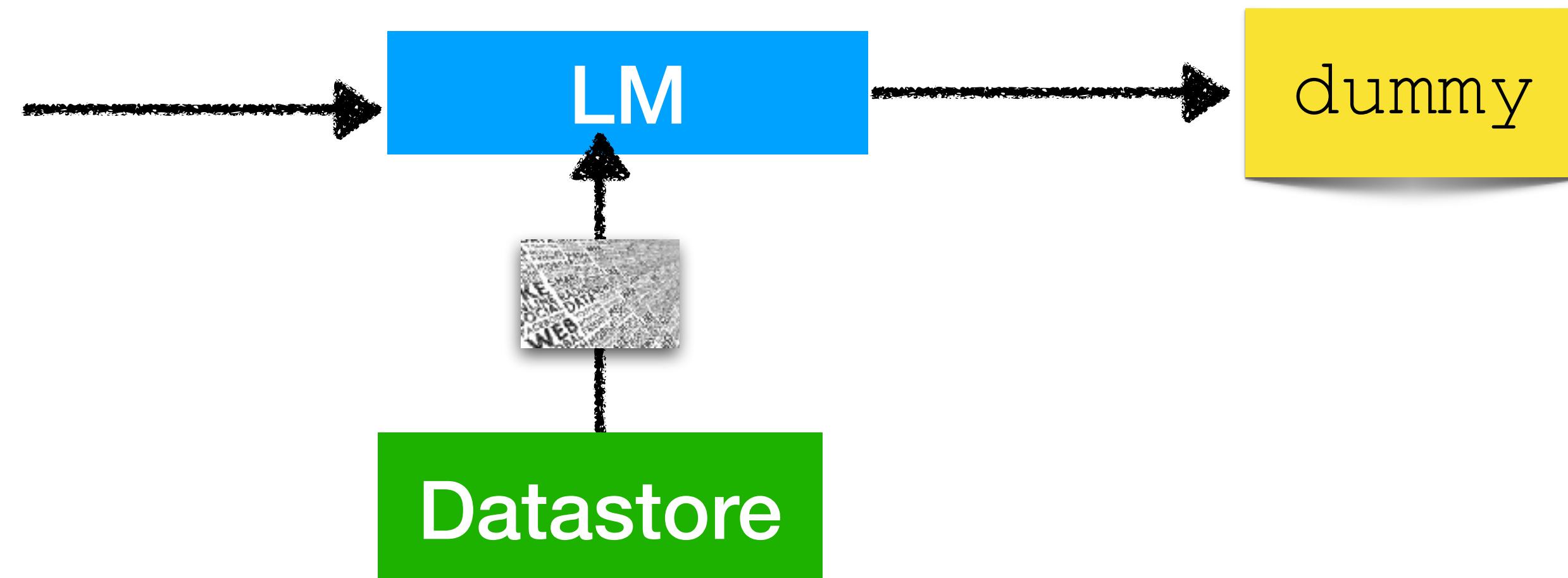
LM can leak private information included in pre-training corpora

# Effectiveness of retrieval-based LMs



Email: mail@alice.com

Phone: 404-



# Two key questions for downstream adaptations

**How** can we adapt a retrieval-based LM for a task?

**When** should we use a retrieval-based LM?

# Downstream adaptation of retrieval-based LMs

## What are the **tasks**?

- Open-domain QA
- Other knowledge-intensive tasks
- General NLU
- Language Modeling & other generation tasks

## How to **adapt**?

- **Fine-tuning**
- Reinforcement learning
- Prompting

## What is **data store**?

- Unlabeled Wikipedia / CC
- Web (Google / Bing Search Results)
- Training data

# Adapting retrieval-based LMs for tasks

## Fine-tuning

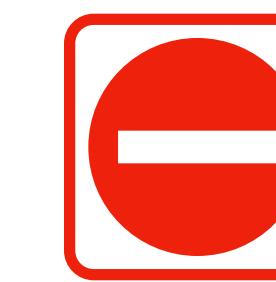
Training LM and / or retriever  
on task-data & data store



# Adapting retrieval-based LMs for tasks

## Fine-tuning

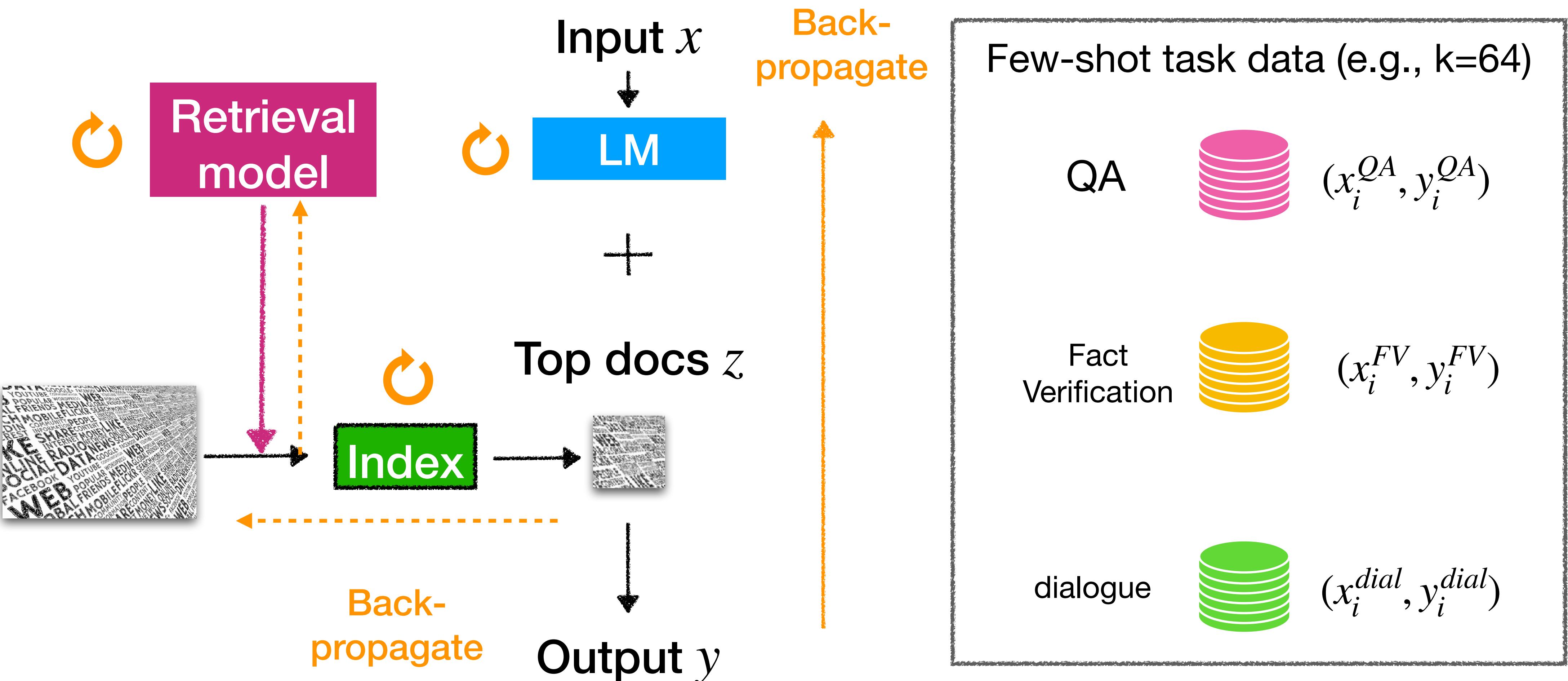
Training LM and / or retriever  
on task-data & data store



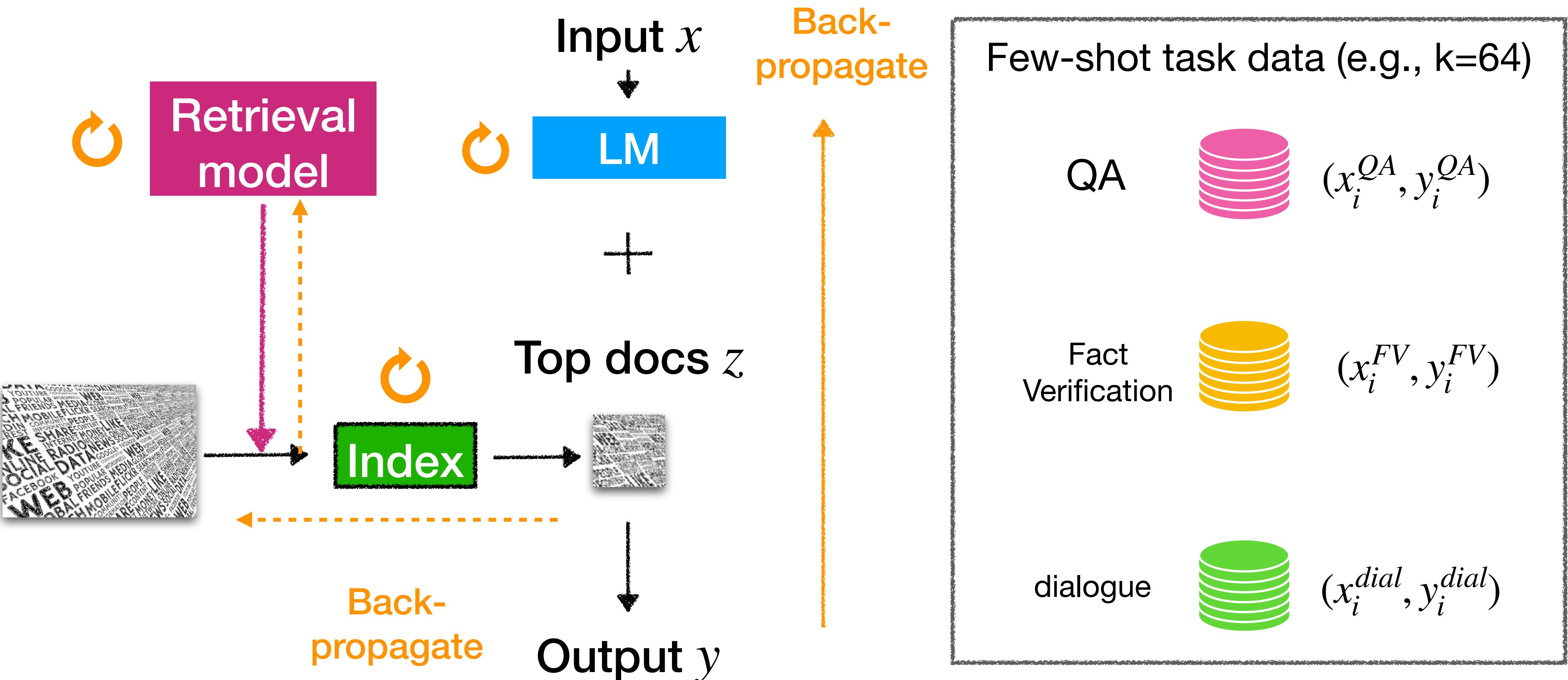
Costs of retrieval-based LM  
training (Section 4)

Independent training (DPR)  
Asynchronous updates (REALM)  
...

# ATLAS (Izacard et al., 2022; Section 4)

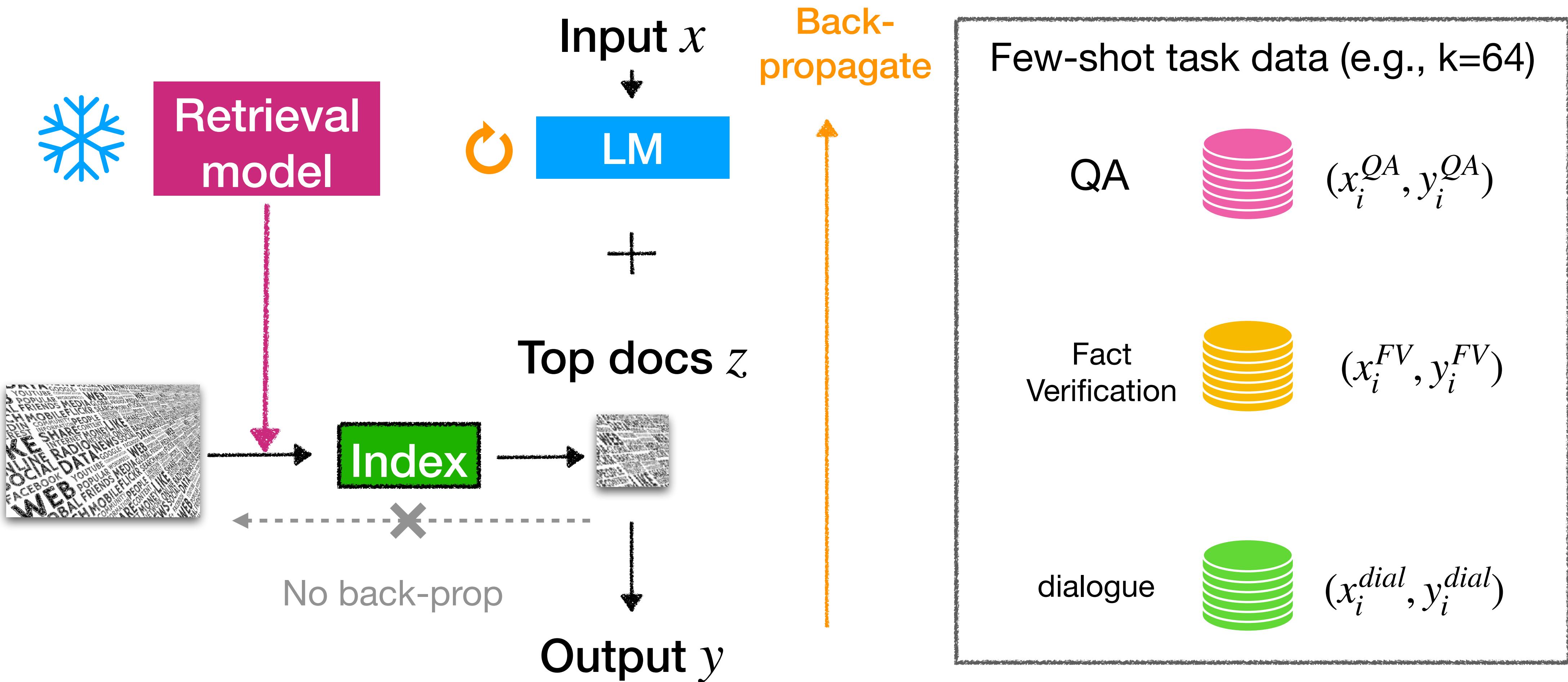


# ATLAS (Izacard et al., 2022; Section 4)

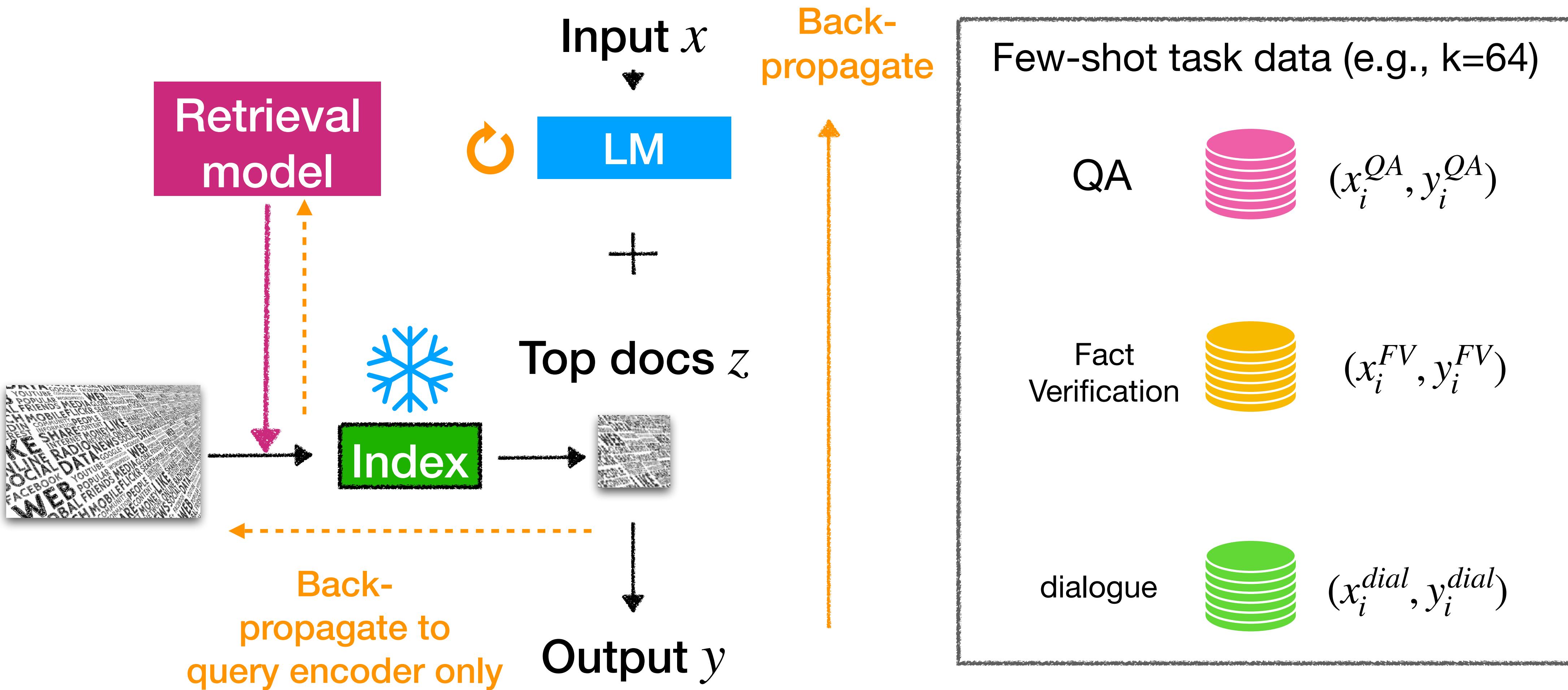


Fully updating retrieval model & index  
during training expensive!

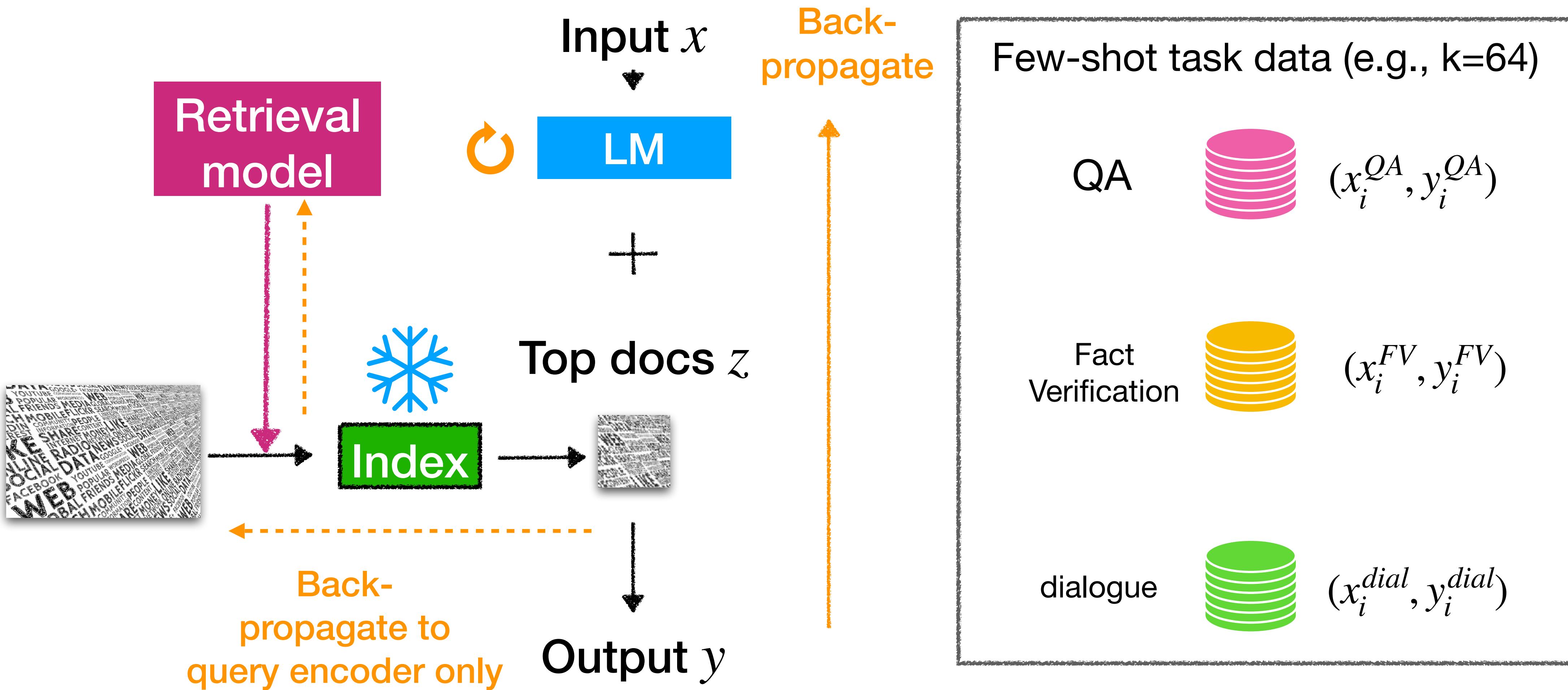
# ATLAS: fixed retrieval with fine-tuned LM



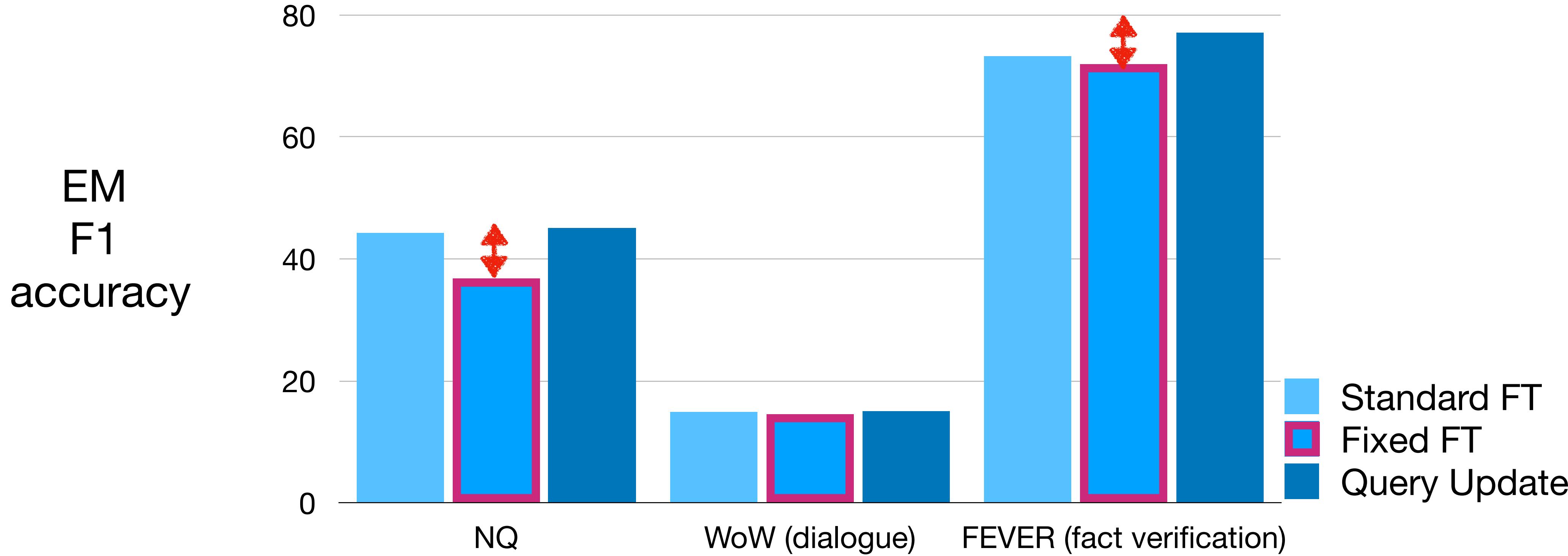
# ATLAS: query-encoder only updates



# ATLAS: full / transfer v.s. few dataset settings

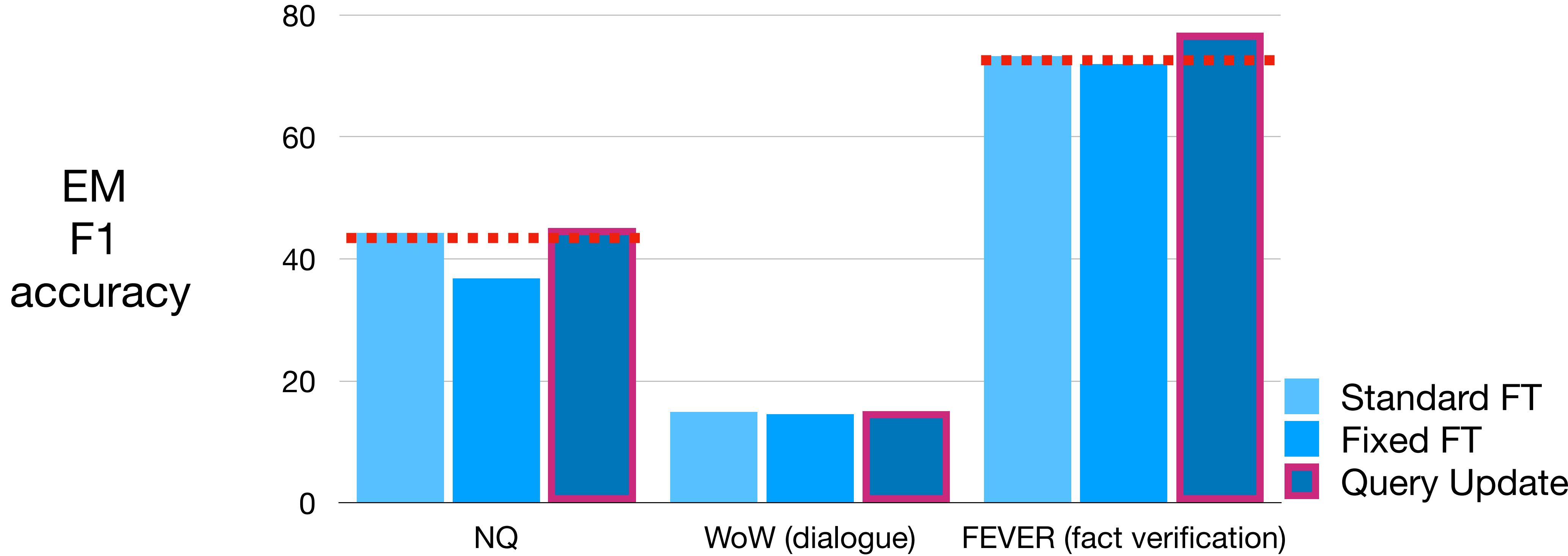


# Ablations of efficient retrieval training



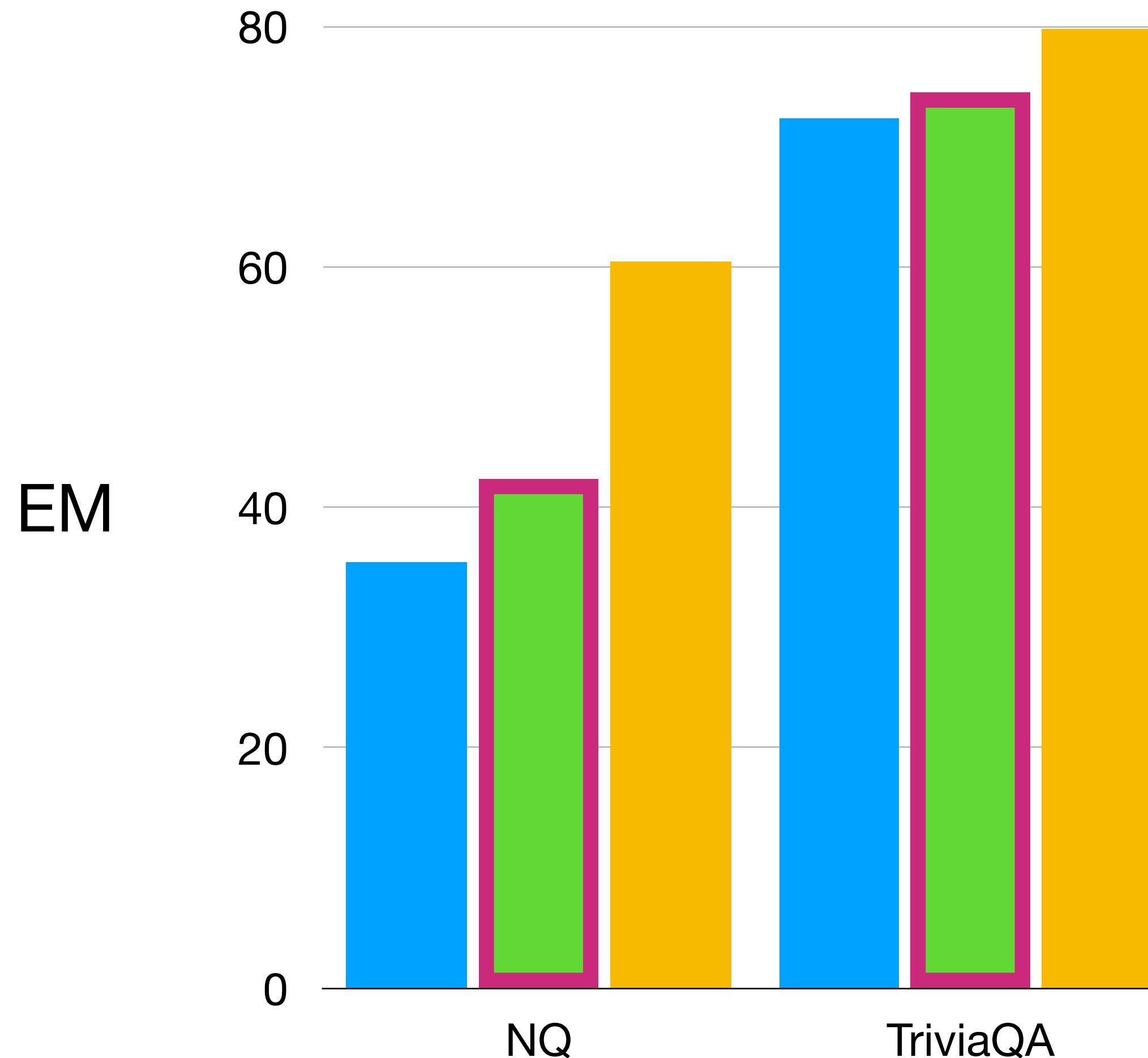
Fixed FT shows large performance drop on QA.

# Ablations of efficient retrieval training



Query-side fine-tuning match or outperforms full fine-tuning

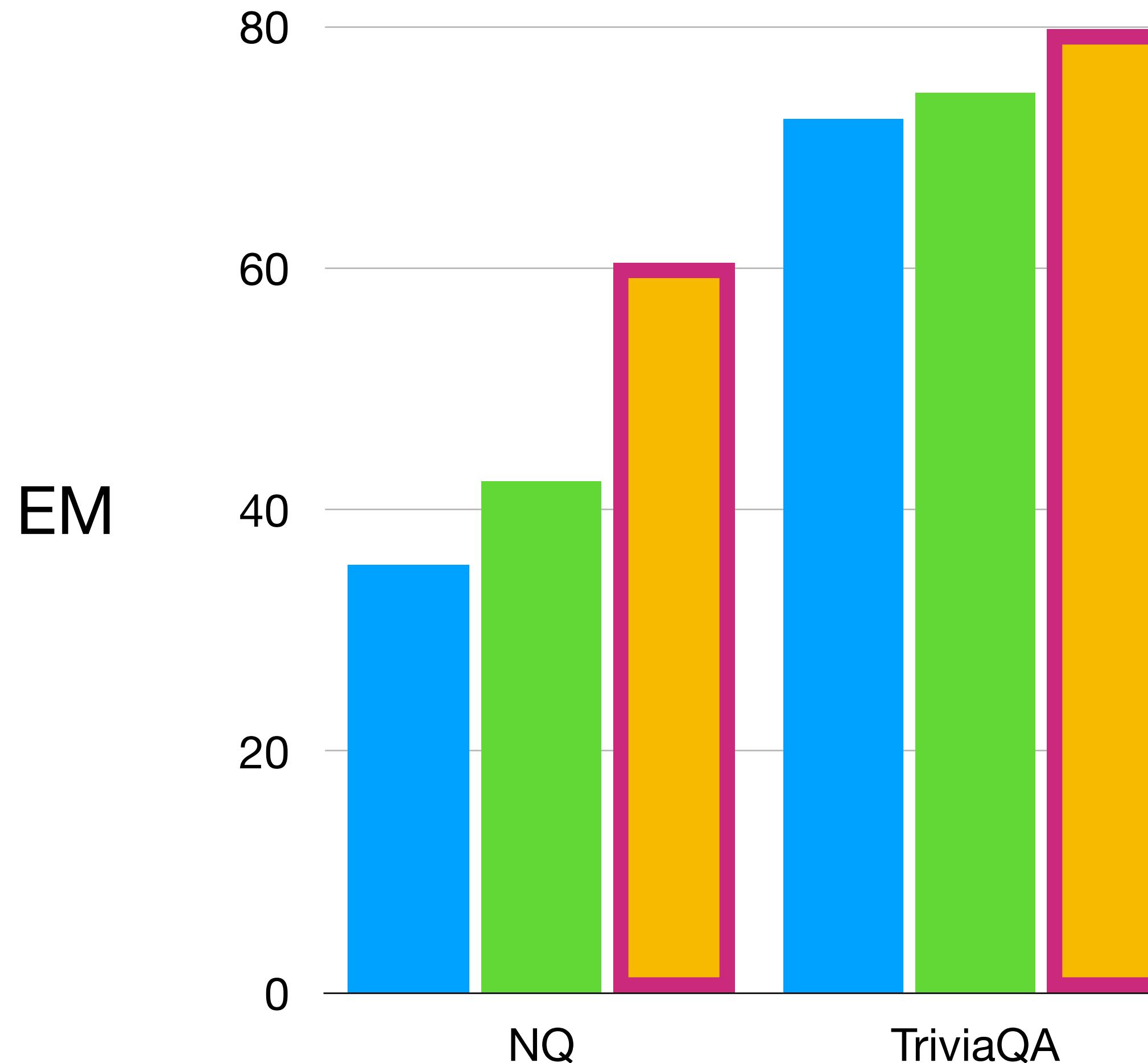
# Task Results



On QA, ATLAS largely outperforms other LLMs in few-shot

- Chinchilla (70B)
- ATLAS (Few; 11B)
- ATLAS (Full; 11B)

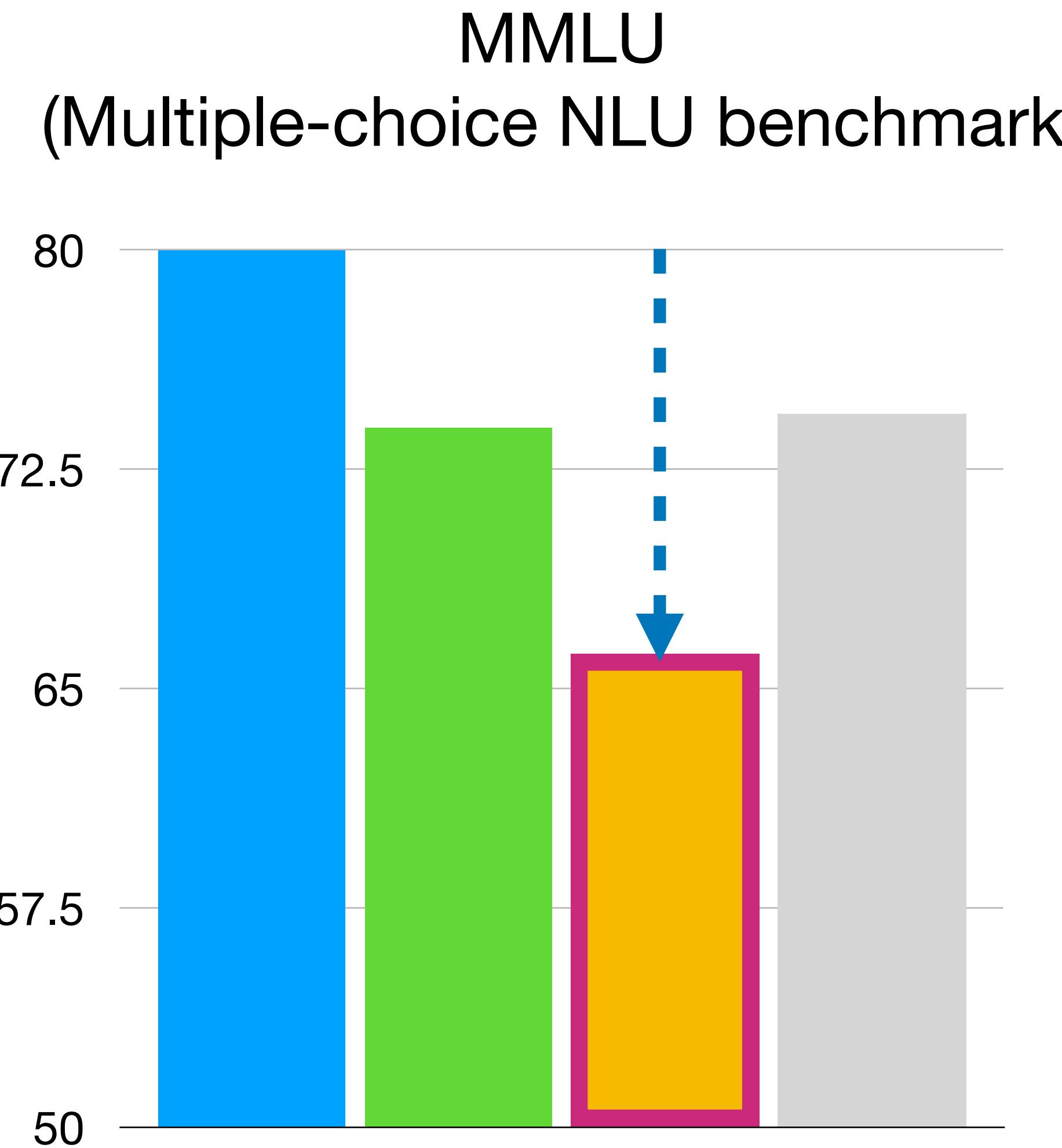
# Task Results



Full-shot fine-tuning further improves performance

- Chinchilla (70B)
- ATLAS (Few; 11B)
- ATLAS (Full; 11B)

# Task Results

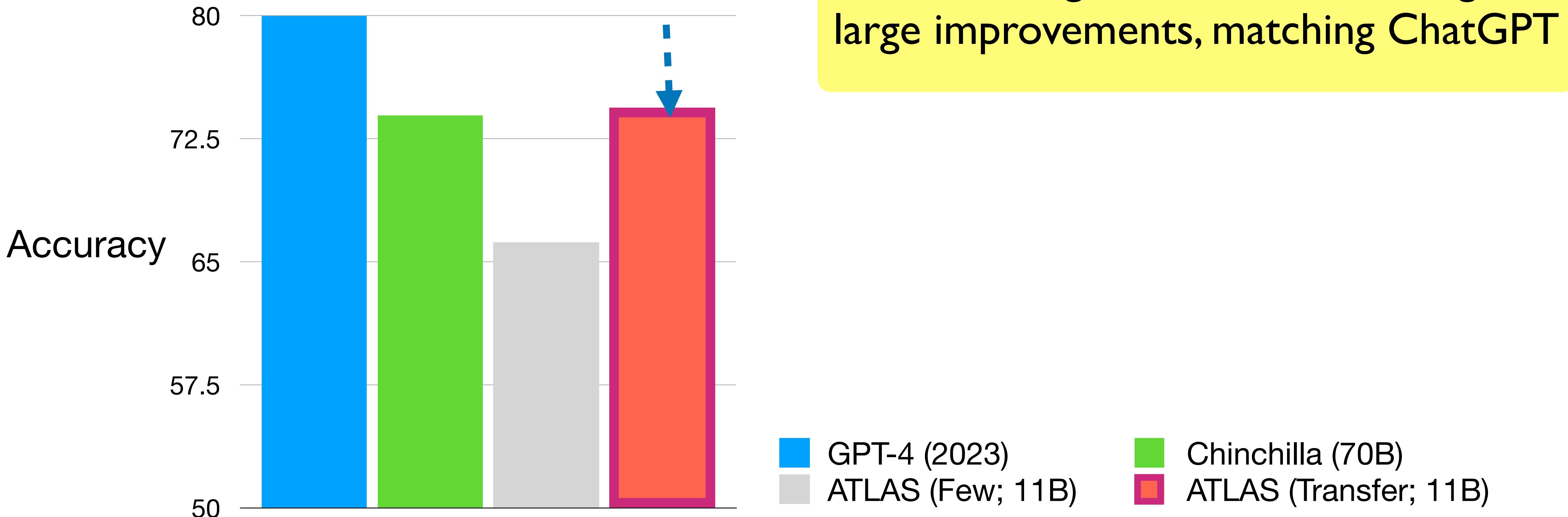


On MMLU, ATLAS few-shot largely underperforms Chinchilla / GPT-4\*.

Room for improvements for diverse task adaptations!

# Task Results

MMLU  
(Multiple-choice NLU benchmark)



# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC

Fine-tuning for QA & knowledge-intensive tasks often gives strong performance (*even in few-shot*)

# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC

Fine-tuning a retriever for a task matters!

# Downstream adaptation of retrieval-based LMs

## What are the **tasks**?

- Open-domain QA
- Other knowledge-intensive tasks
- General NLU
- Language Modeling & other generation tasks

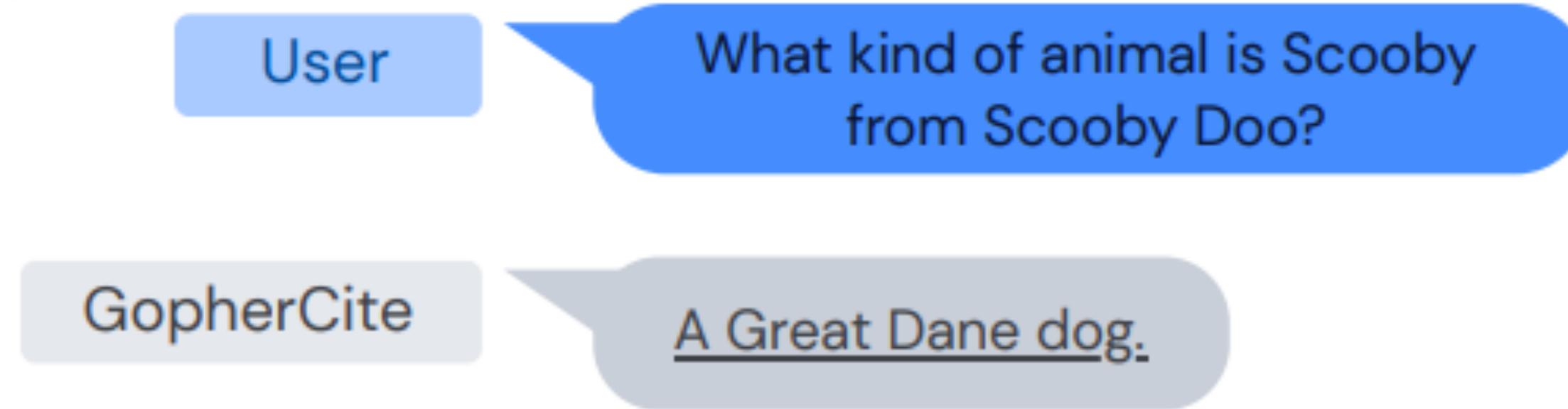
## How to **adapt**?

- Fine-tuning
- **Reinforcement learning**
- Prompting

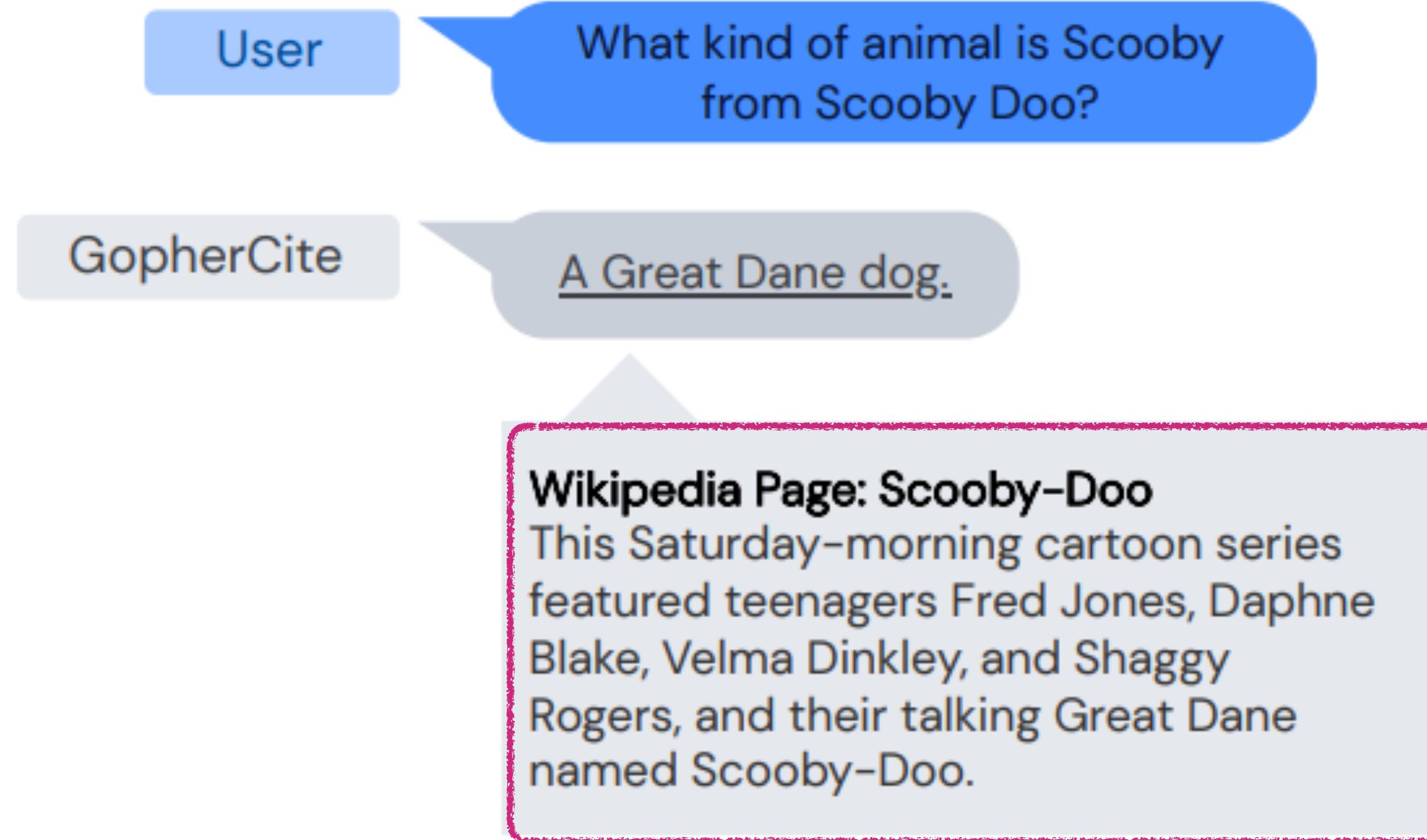
## What is **data store**?

- Unlabeled Wikipedia / CC
- Web (Google / Bing Search Results)
- Training data

# GopherCite (Menick et al., 2022)

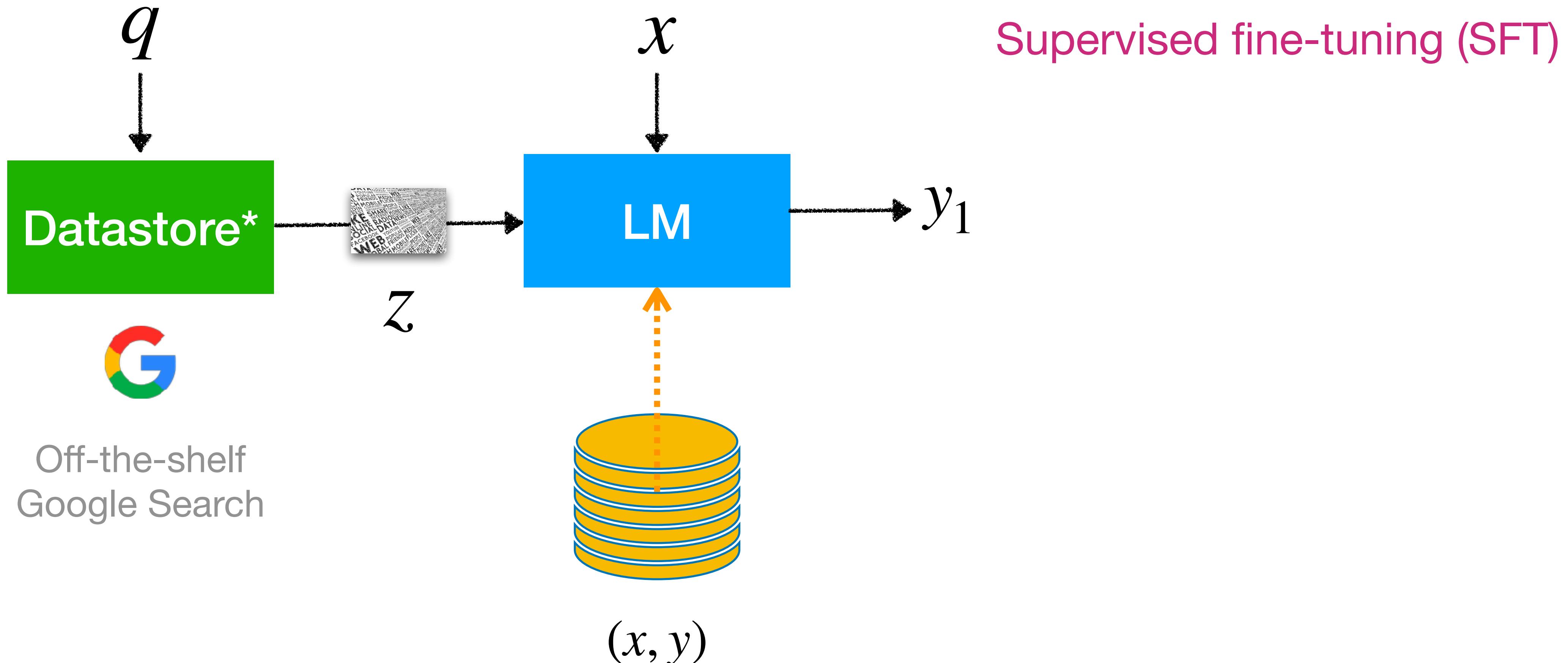


# GopherCite (Menick et al., 2022)

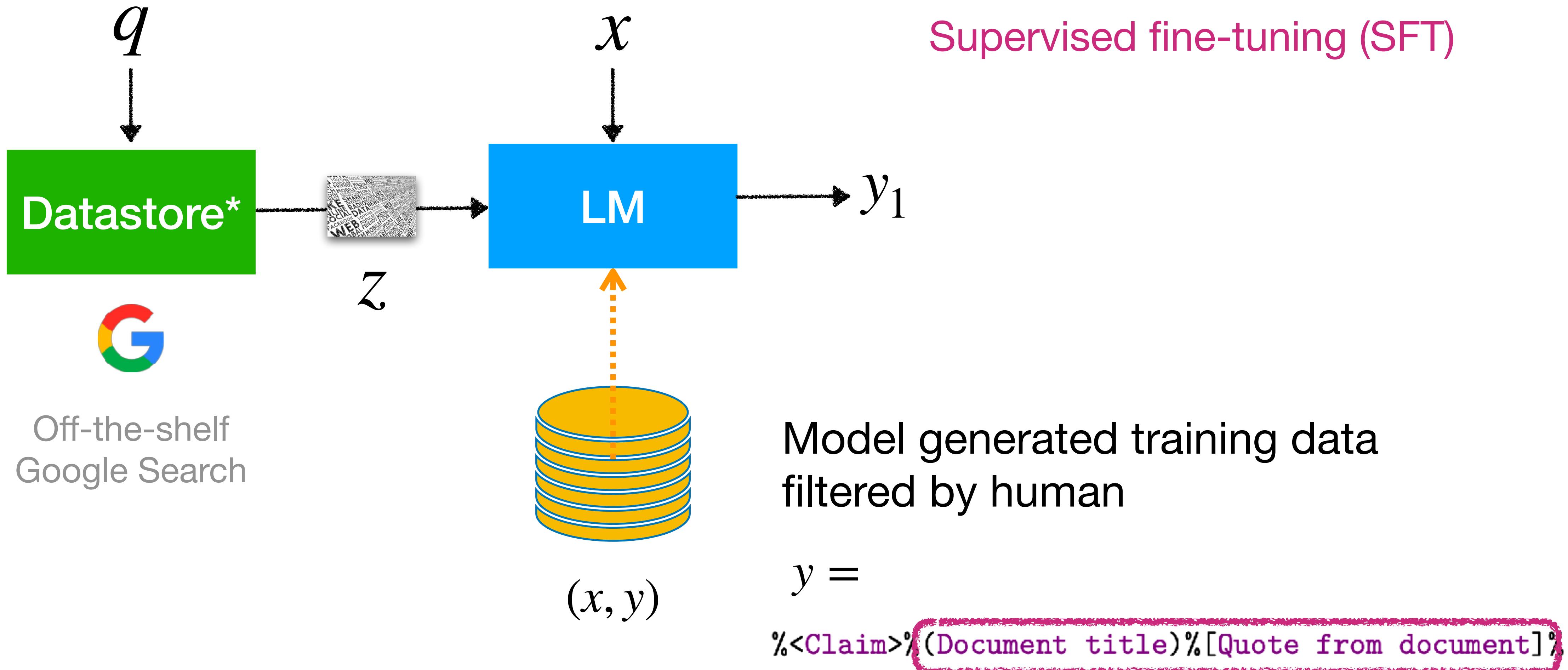


Extract and generate a quote to support an answer

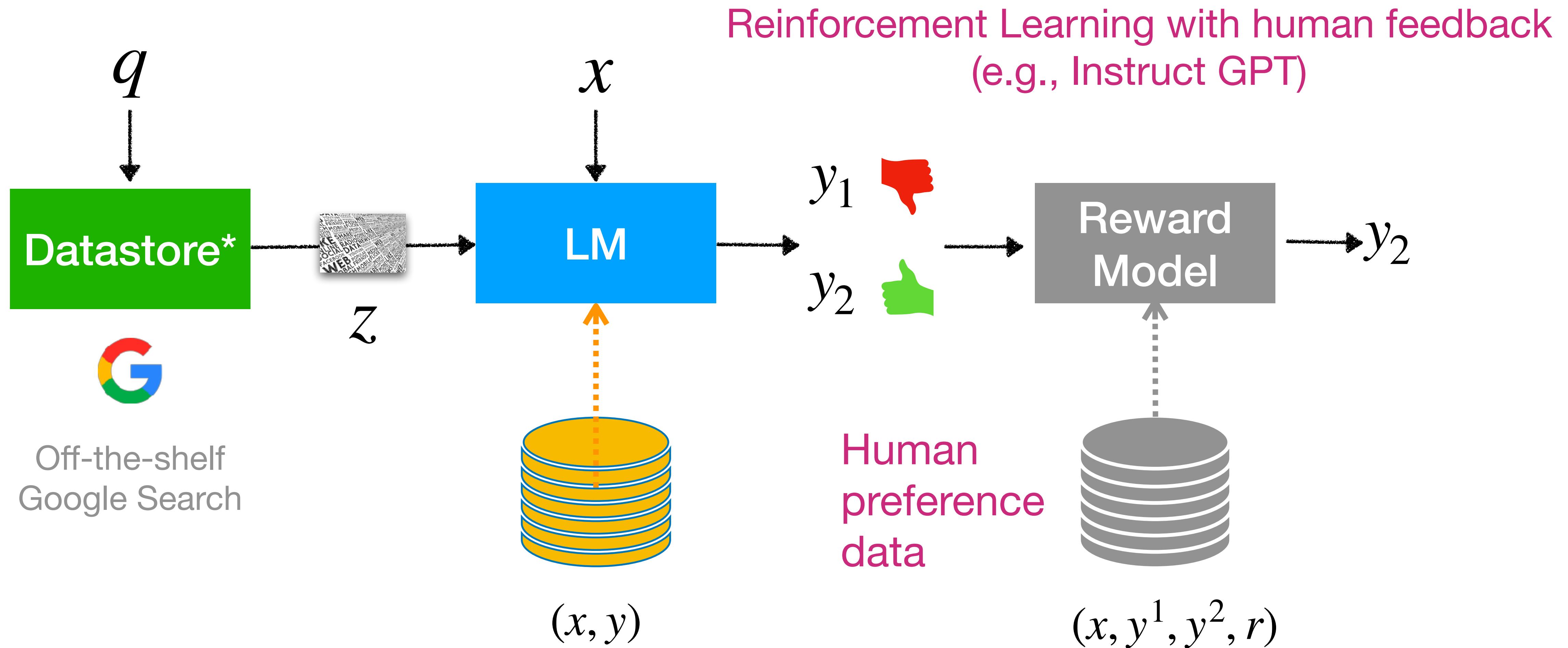
# GopherCite: RLHF for answering with verified quotes



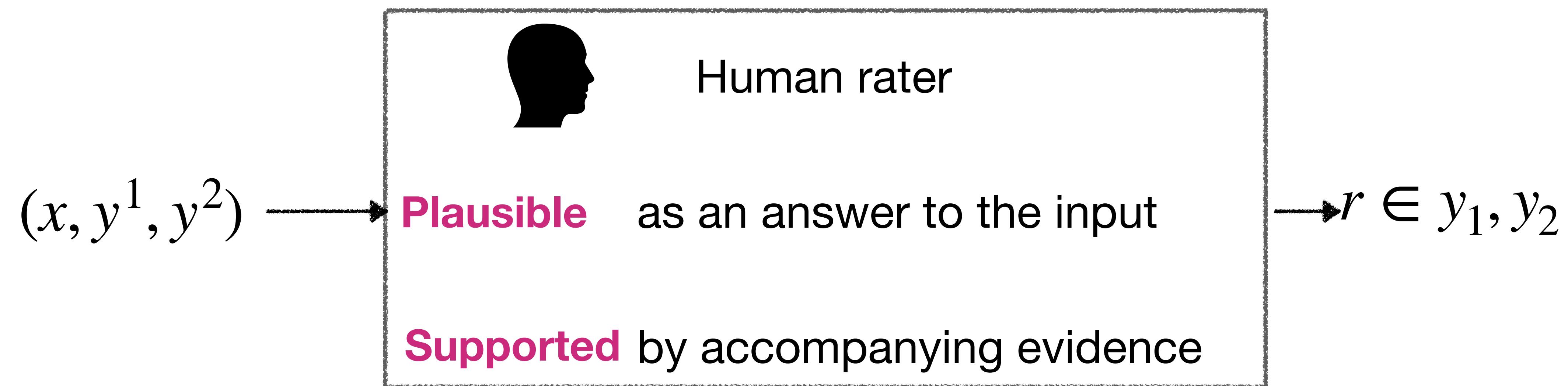
# GopherCite: RLHF for answering with verified quotes



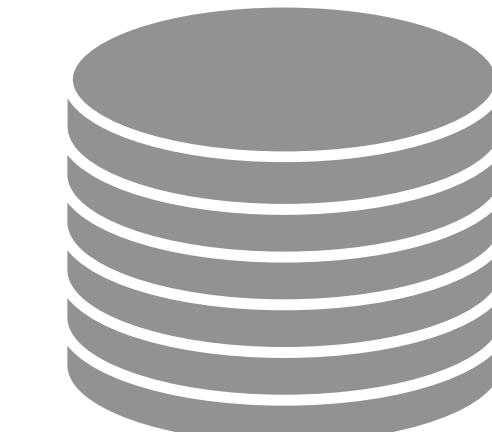
# GopherCite: RLHF for answering with verified quotes



# GopherCite: RLHF for answering with verified quotes

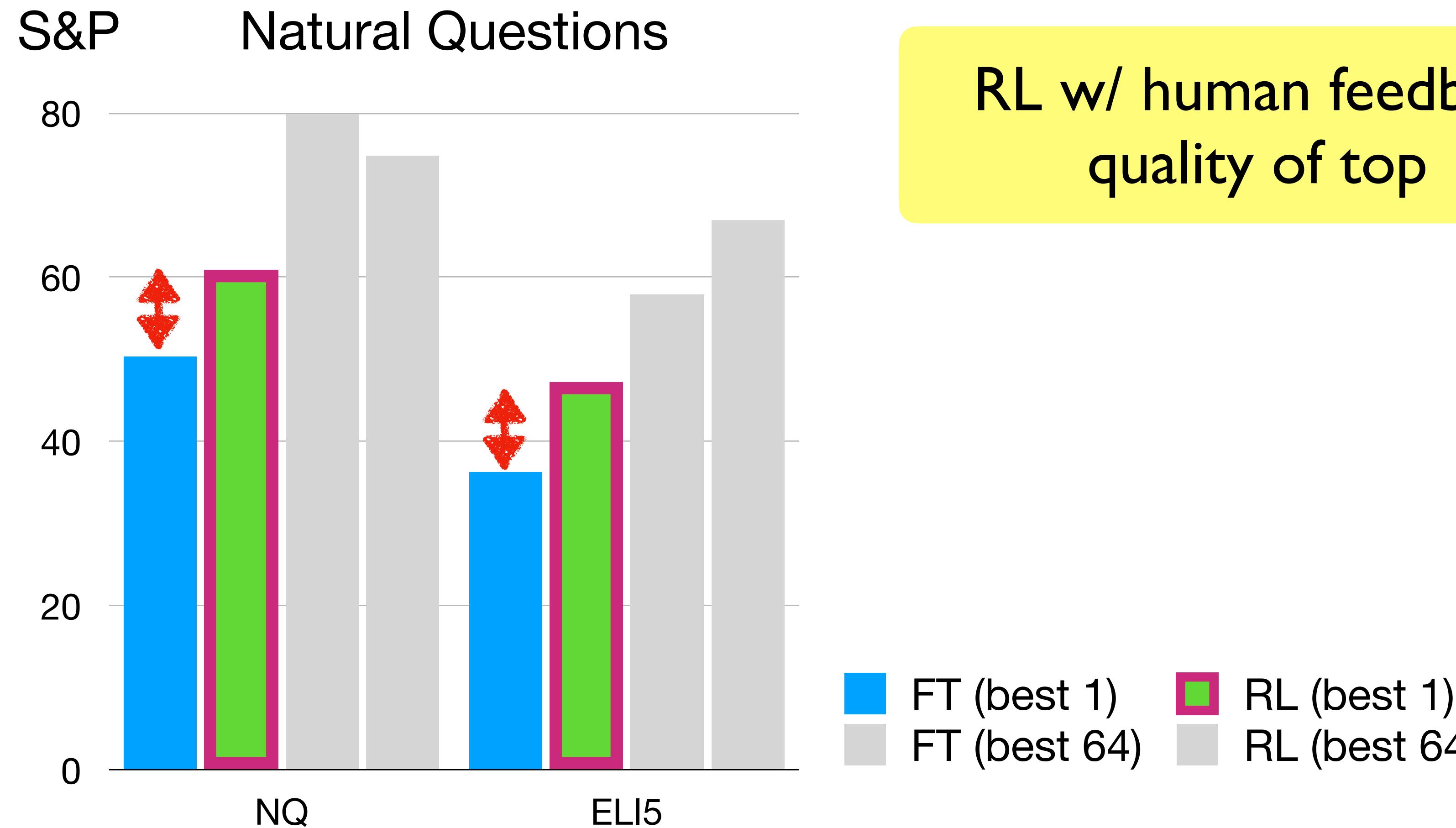


33k Human preference data

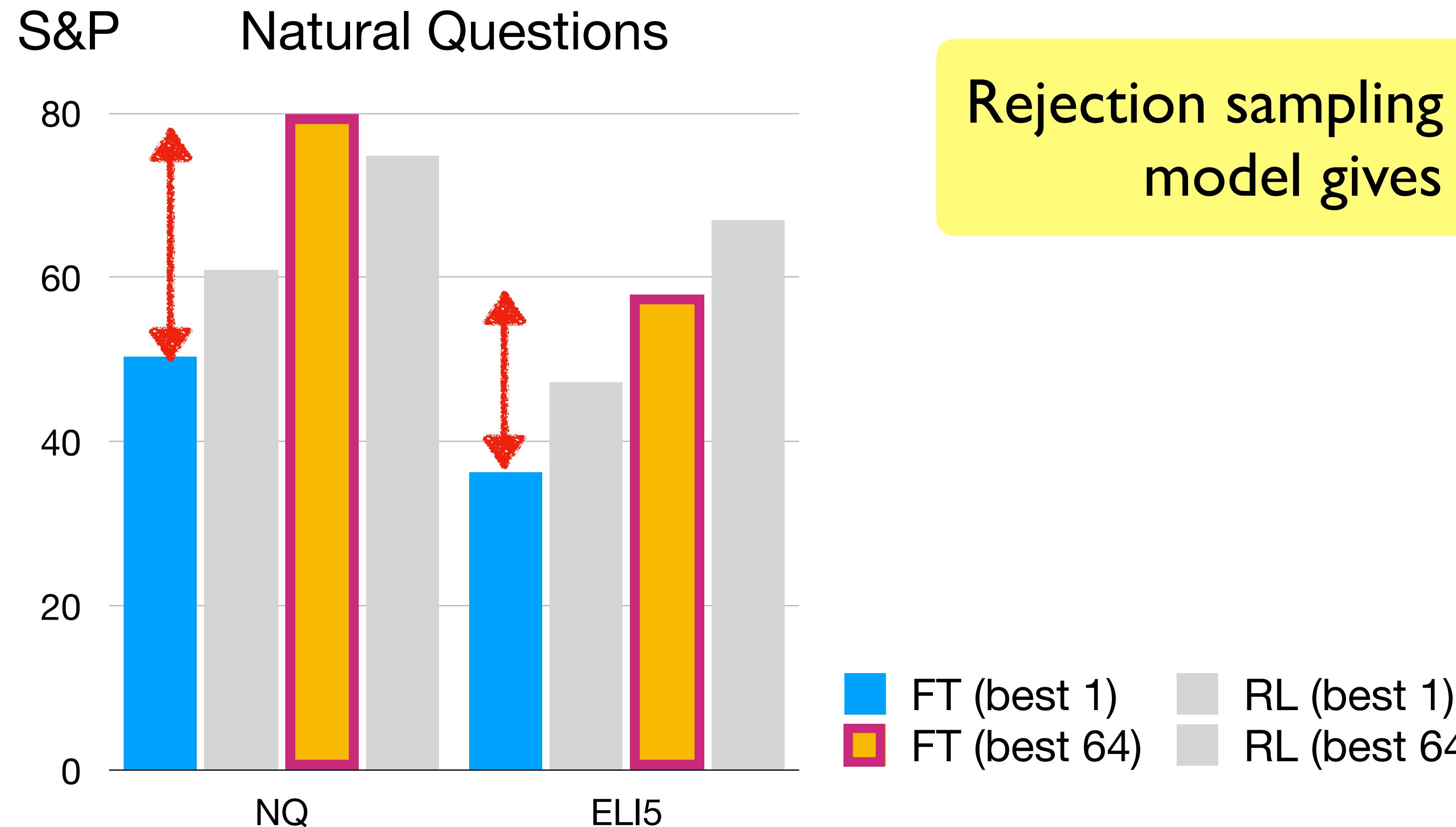


$(x, y^1, y^2, r)$

# GopherCite: effects of RL



# GopherCite: effects of RL



# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC
GopherCite (Menick et al., 2022)	Open-domain QA, Long-form QA	Fine-tuning + RL (LM)	Google Search Results

Benefit of **fine-tuning-based approaches**



Customizable



Competitive w/ more data



Require training

# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC
GopherCite (Menick et al., 2022), also WebGPT (Nakano et al., 2021)	Open-domain QA, Long-form QA	Fine-tuning + RL (LM)	Google Search Results

## Benefit of **RL**



Better alignment with user preferences



Require additional data collection (preference)

# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC
GopherCite (Menick et al., 2022)	Open-domain QA, Long-form QA	Fine-tuning + RL (LM)	Google Search Results

What if we cannot train LMs for downstream tasks?  
(e.g., lack of computational resources / proprietary LM ... etc)

# Downstream adaptation of retrieval-based LMs

What are the **tasks**?

- Open-domain QA
- Other knowledge-intensive tasks
- General NLU
- Language Modeling & other generation tasks

How to **adapt**?

- Fine-tuning
- Reinforcement learning
- **Prompting**

What is **data store**?

- Wikipedia
- Web (Google / Bing Search Results)
- Training data

# Prompting

$k$ -shot instances ( $k=0, 32 \dots$  etc)



Q: who Is the US president

A: Joe Biden

##

Q: What is the capital of US?

A: Washington DC.

##

Q: what is the Ontario capital?

A:

Doesn't require LM training on tasks!

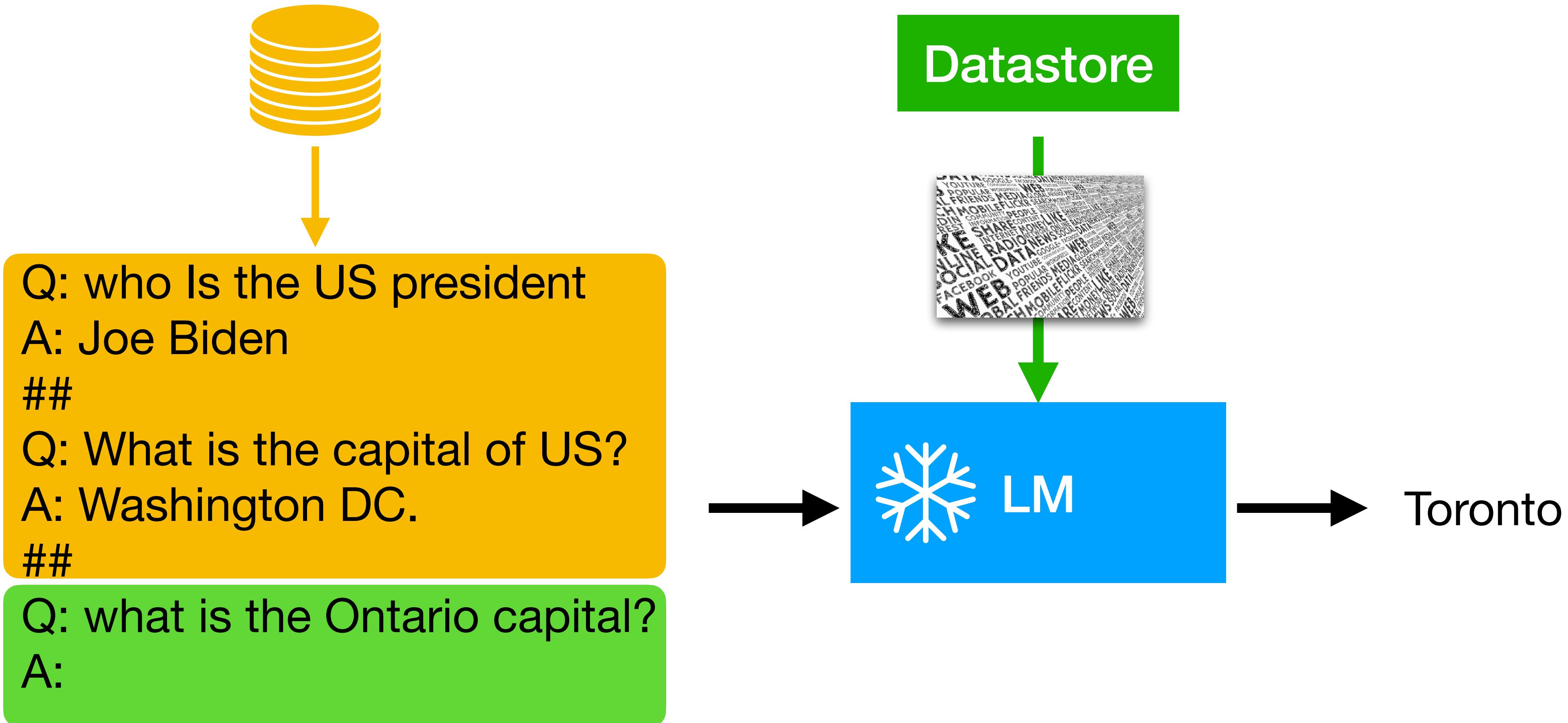
Training instances (demonstrations)



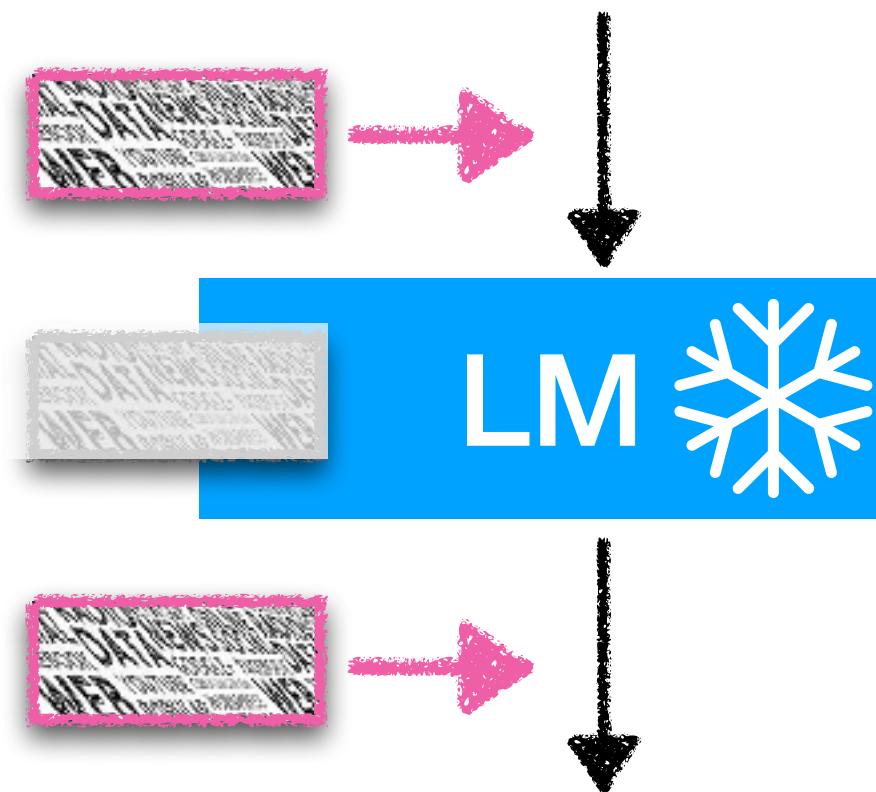
Test instances

# Retrieval-based prompting

$k$ -shot instances ( $k=0, 32 \dots$  etc)



# Design choice of retrieval-based Prompting



## **Input space:**

Incorporate retrieved context in input space

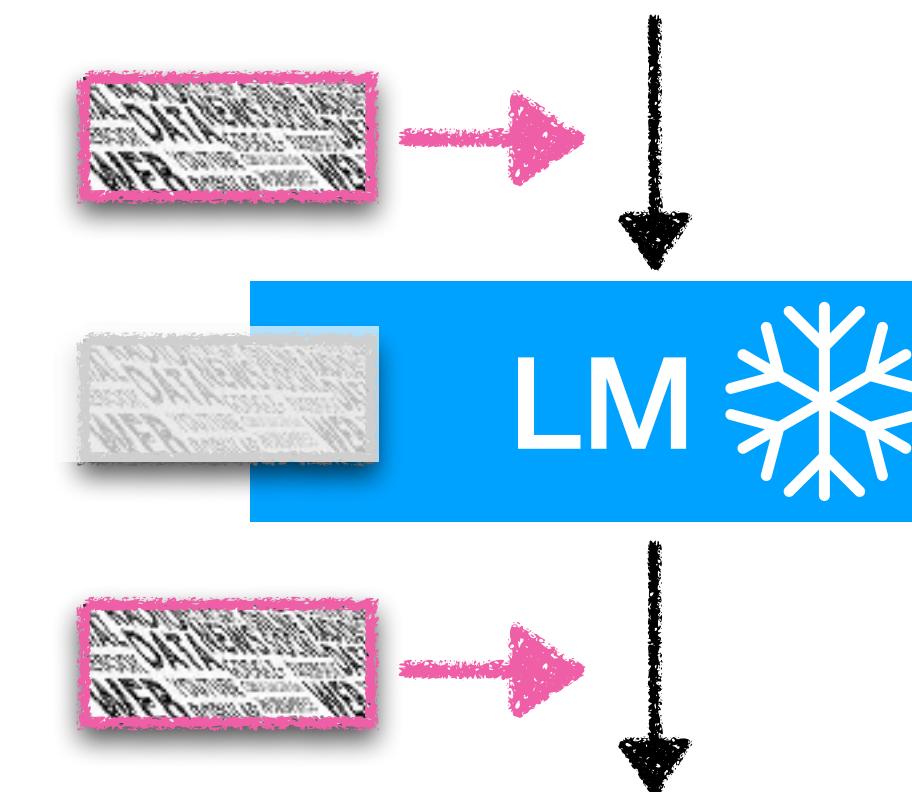
## Intermediate layers:

N/A

## **Output space:**

Interpolate token probability distributions in output space

# Design choice of retrieval-based Prompting



Extending kNN-LM for downstream tasks

**Input space:**  
Incorporate retrieved context in input space

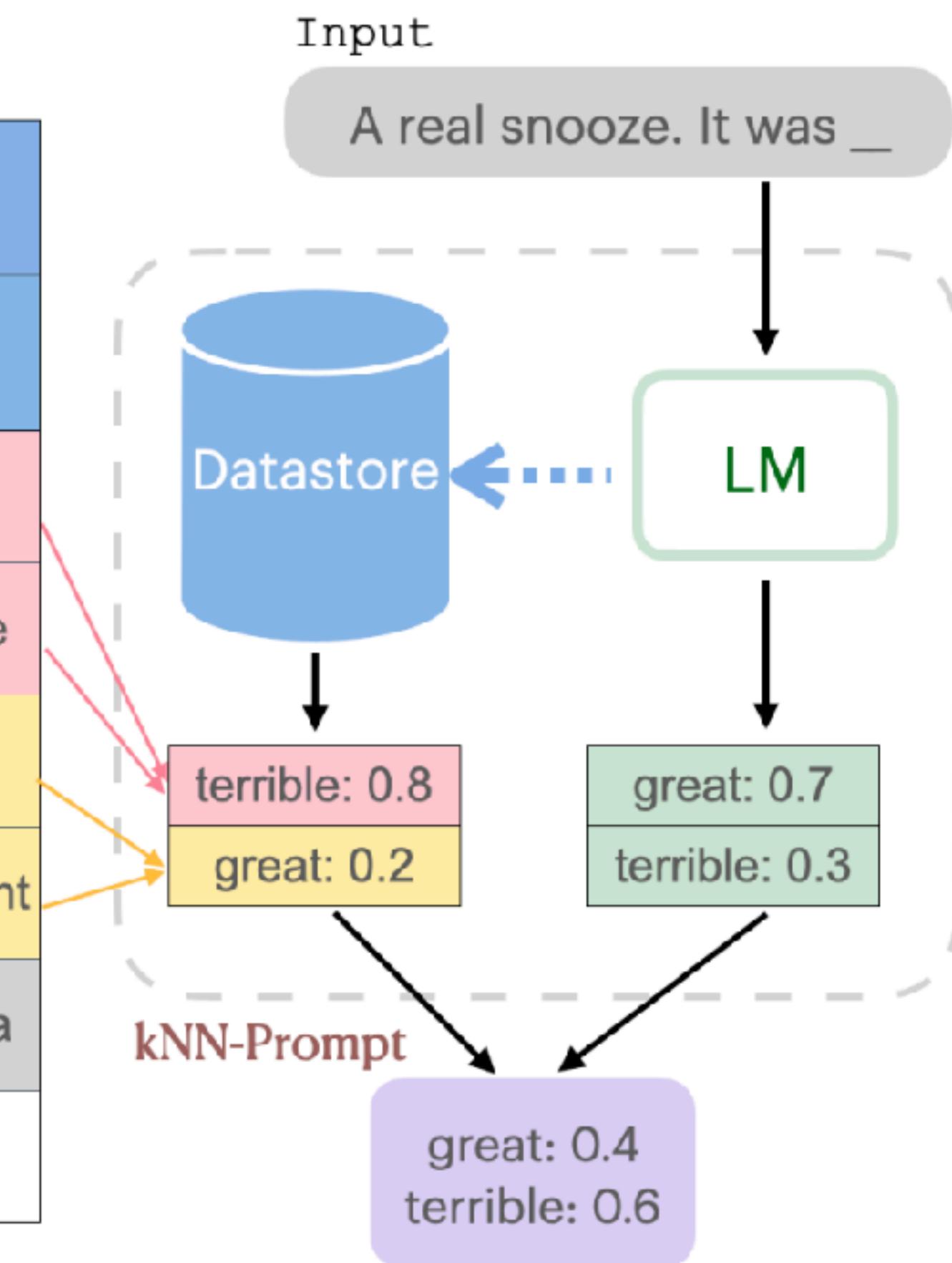
Intermediate layers:  
N/A

**Output space:**  
Interpolate token probability distributions in output space

# kNN-prompt (Shi et al., 2022)

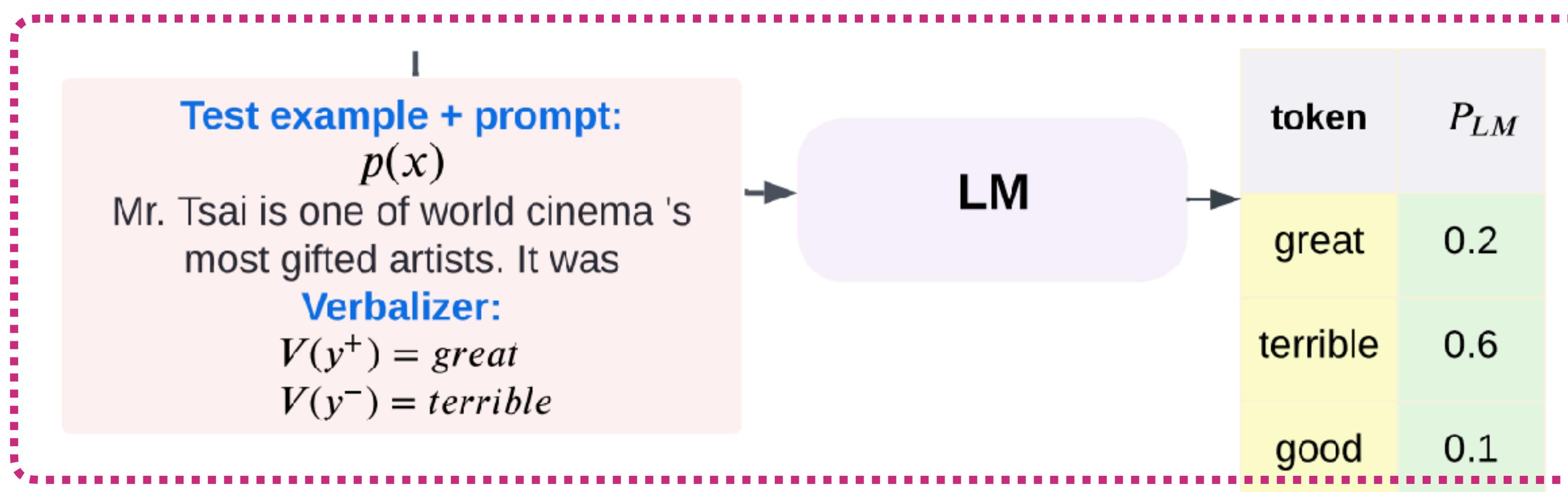
kNN LM with fuzzy verbalizers  
for zero-/few-shot **classification**

Datastore	
Leftward contexts	Next token
The thriller is a real snooze. The director can't	silly
It is seriously a real snooze-fest. The acting is	terrible
The character and world design was	great
Five great movies that give us	excellent
This is junk food	cinema
...	...

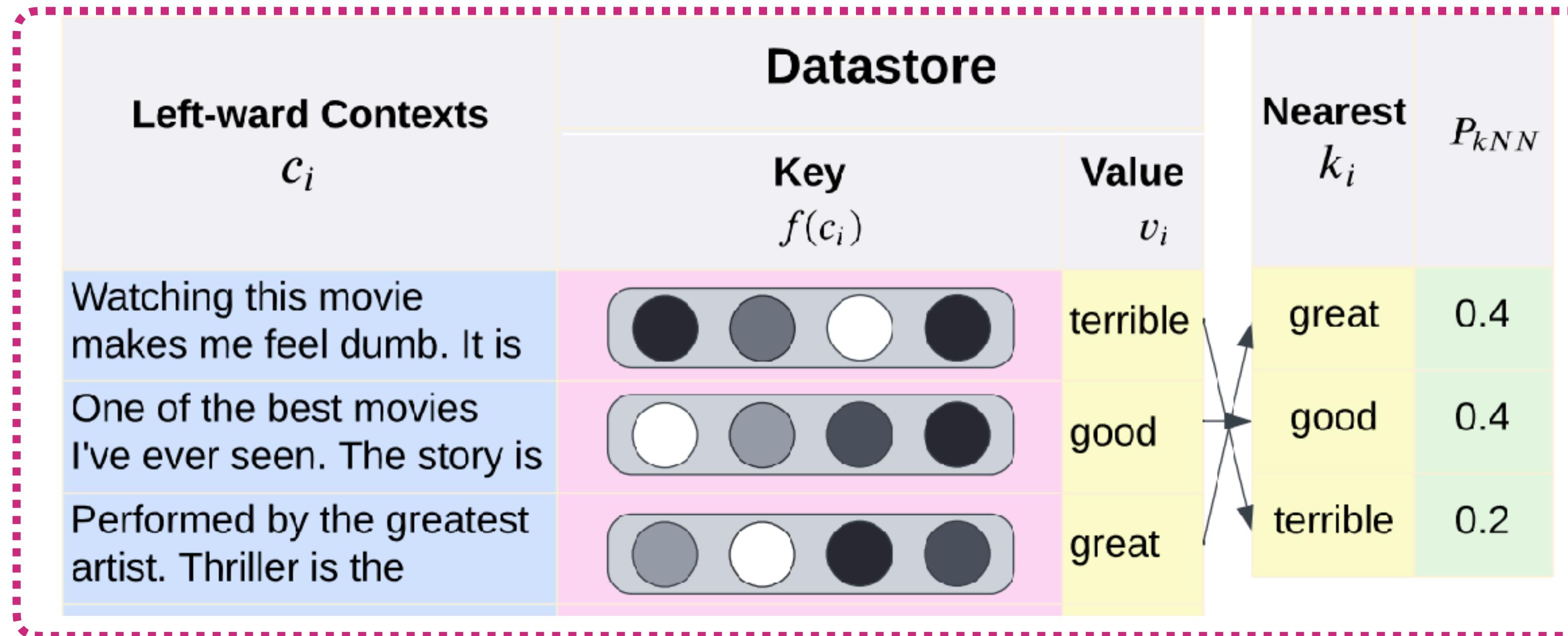


# kNN-prompt (Shi et al., 2022)

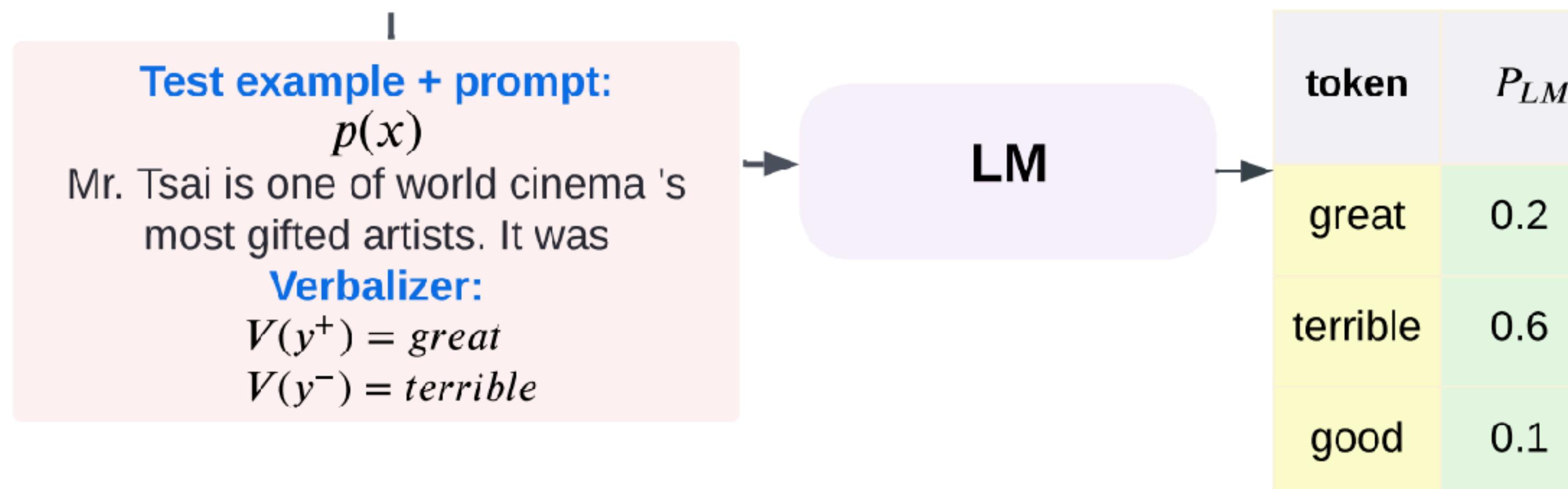
LM predicts next token



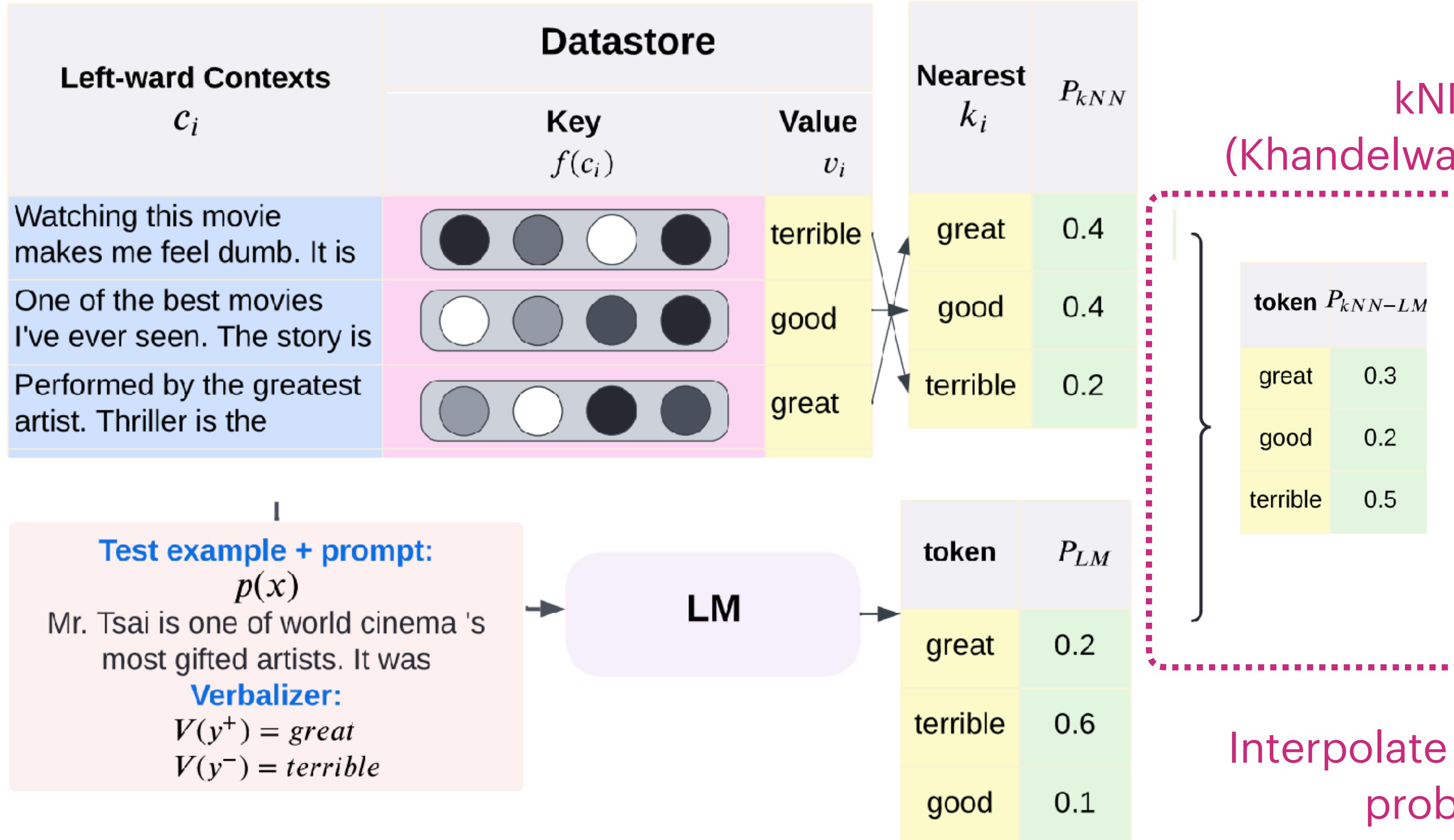
# kNN-prompt (Shi et al., 2022)



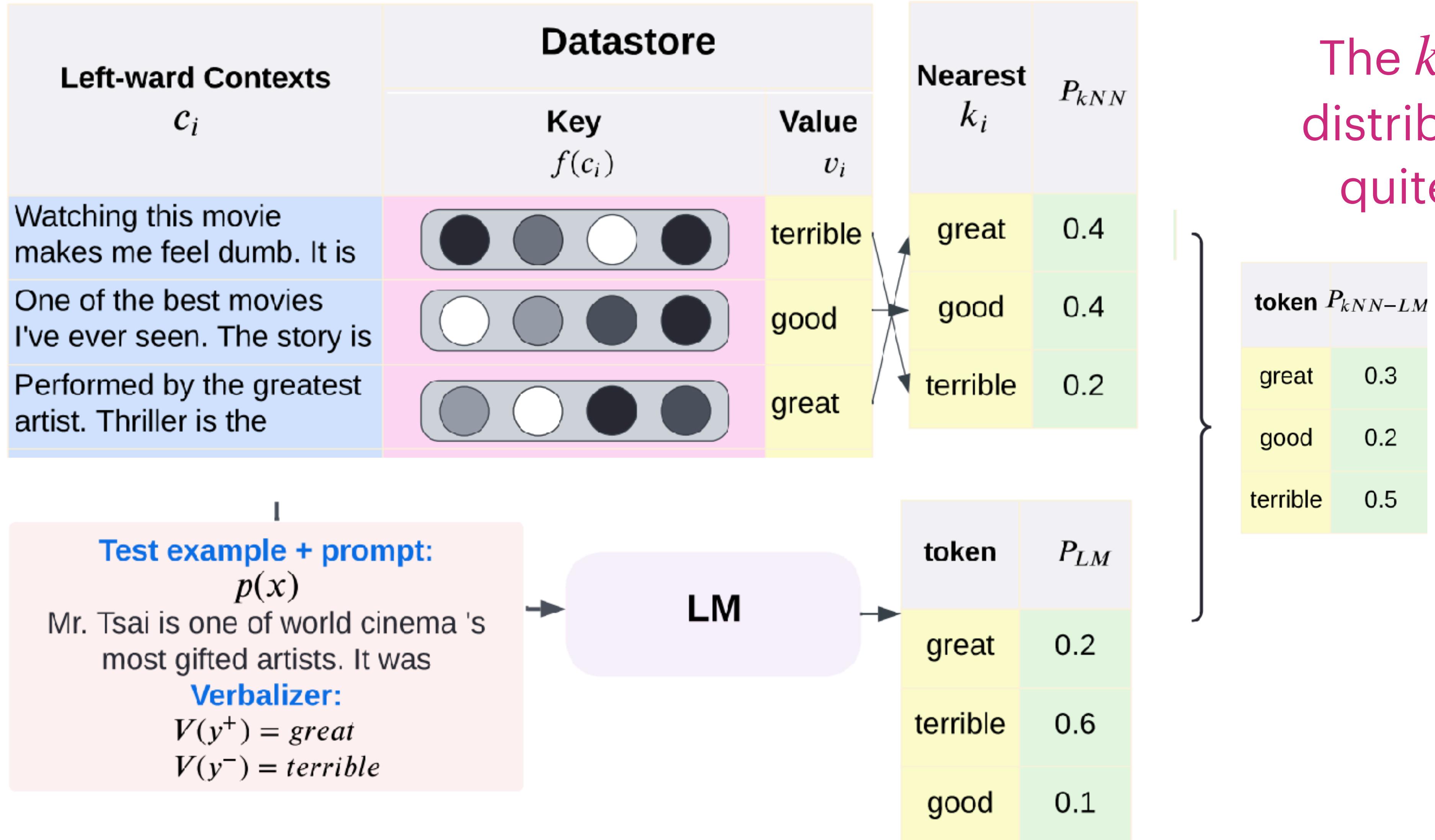
kNN predicts next tokens



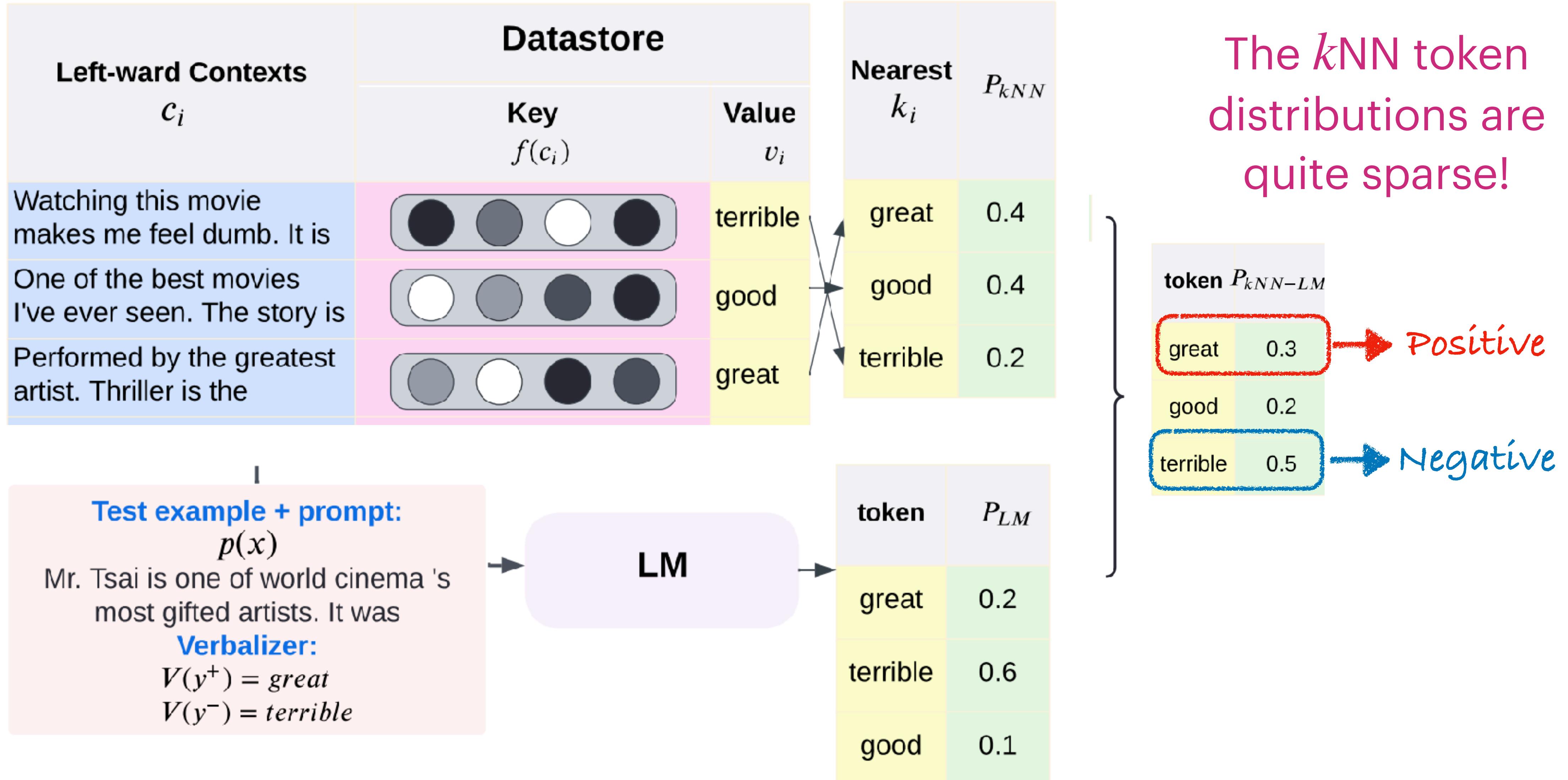
# kNN-prompt (Shi et al., 2022)



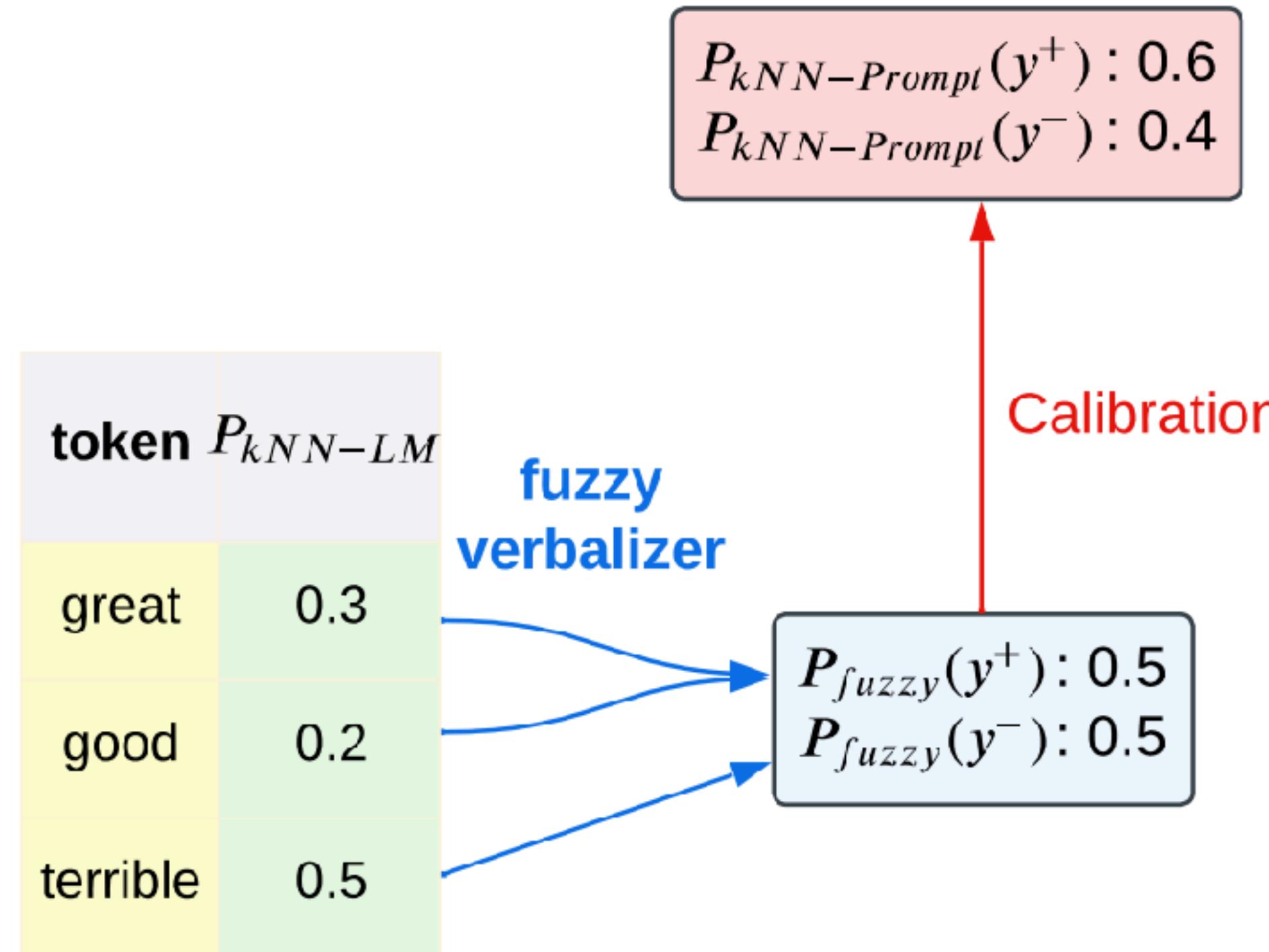
# kNN-prompt (Shi et al., 2022)



# kNN-prompt (Shi et al., 2022)



# kNN-prompt (Shi et al., 2022)



**Fuzzy verbalizer** maps token probability to target class labels

$$P_{FV}(y \mid x) \propto \sum_{v_i \in \mathcal{N}(v)} P(v_i \mid p(x))$$

Find similar tokens using GloVe & ConceptNet

# Results on diverse classification tasks



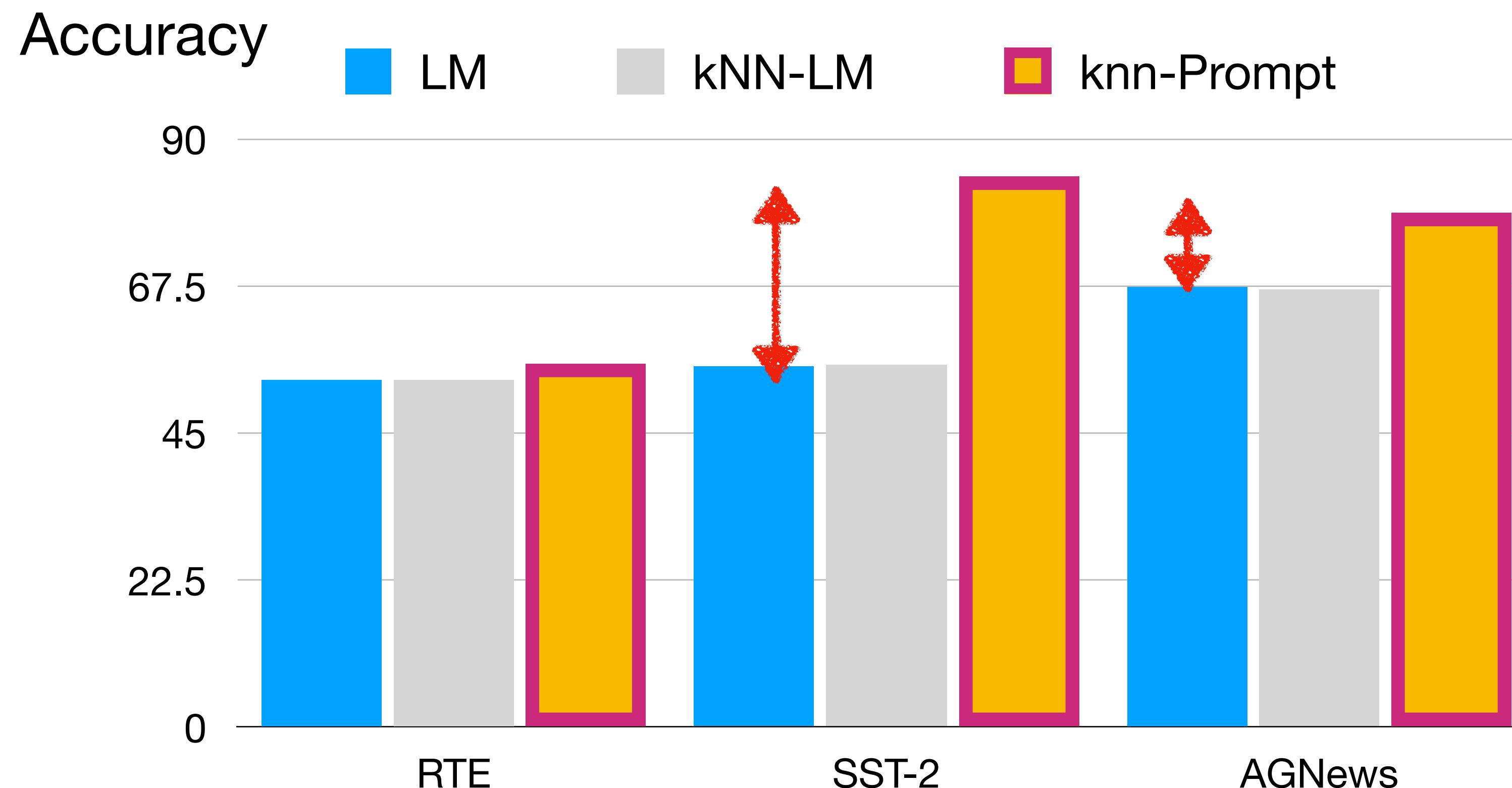
**NLI  
/ entailment**

**Sentiment  
analysis**

**Topic  
classification**

Significant gains from  
kNN-LM

# Results on diverse classification tasks



kNN prompt largely outperforms vanilla LM in zero-shot classification

NLI  
/ entailment

Sentiment  
analysis

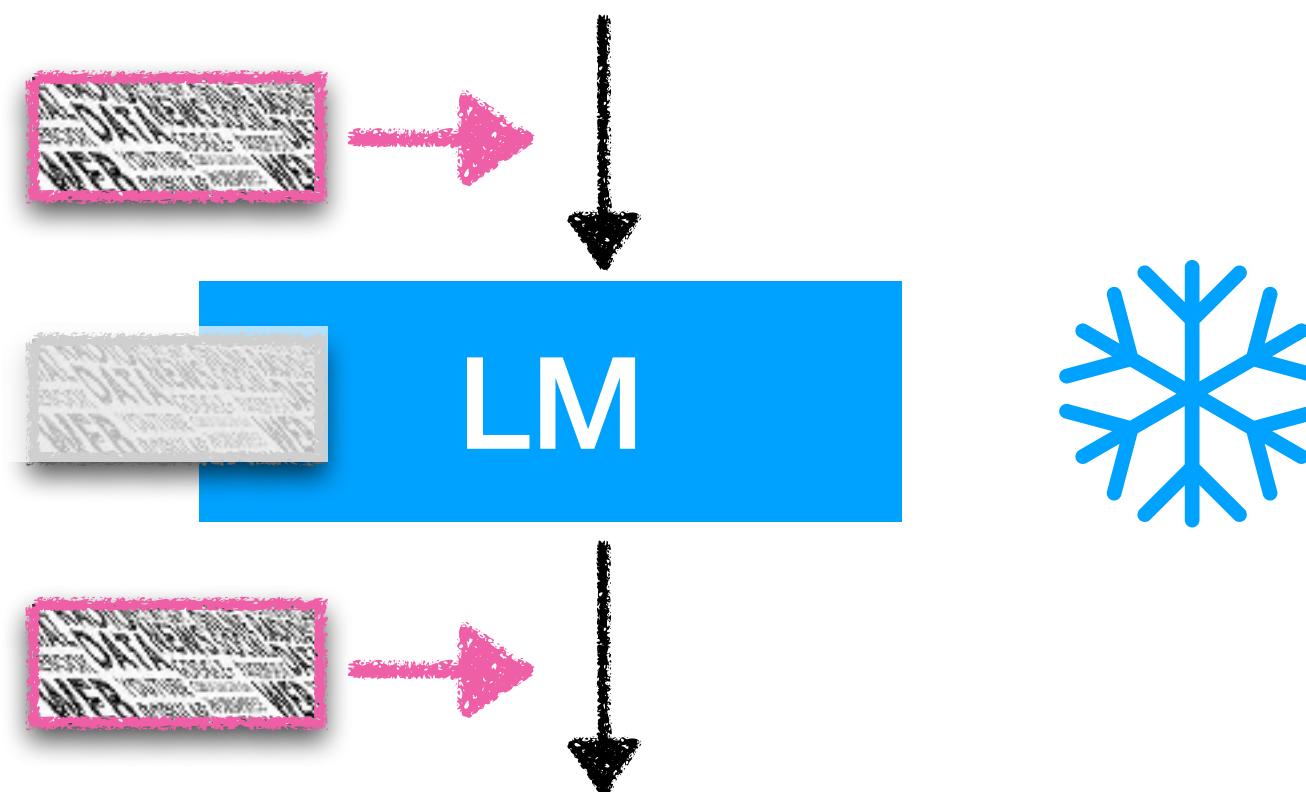
Topic  
classification

# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC
GopherCite (Menick et al., 2022)	Open-domain QA, Long-form QA	Fine-tuning + RL (LM)	Google Search Results
kNN Prompt (Shi et al., 2022)	Classification	Prompting (output)	Wikipedia   CC

Retrieval-based LMs are effective in general NLU tasks!

# Retrieval-based Prompting

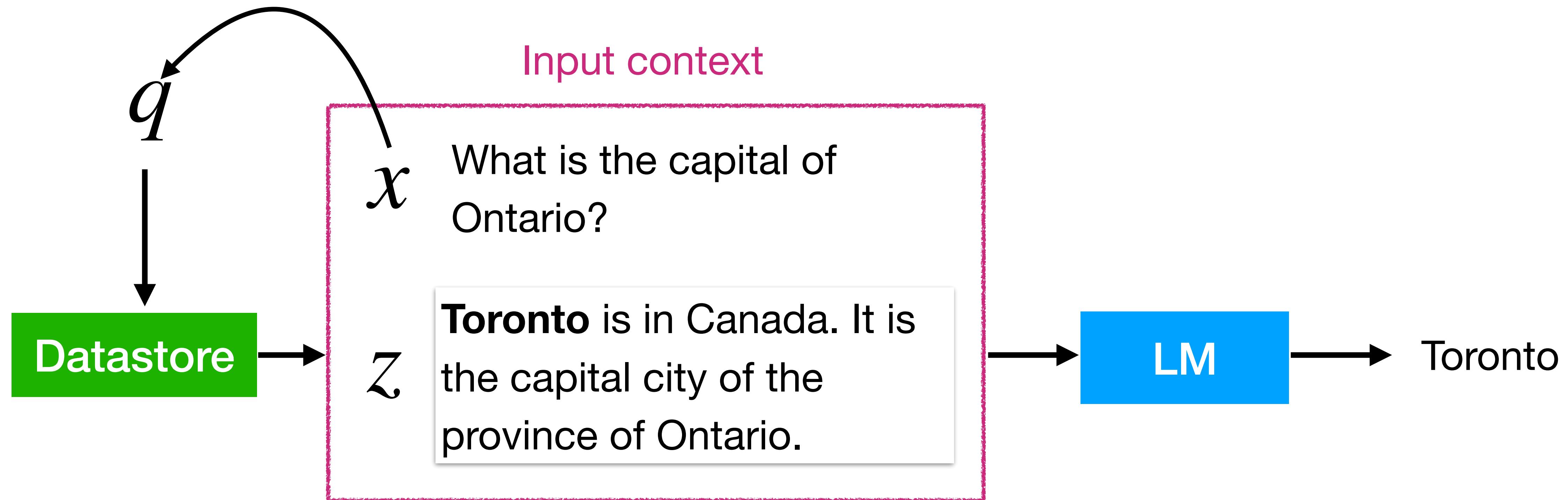


**Input space:**  
Incorporate retrieved context in input space

Intermediate layers:  
N/A

**Output space:**  
Interpolate token probability  
distributions in output space

# Retrieval-in-context

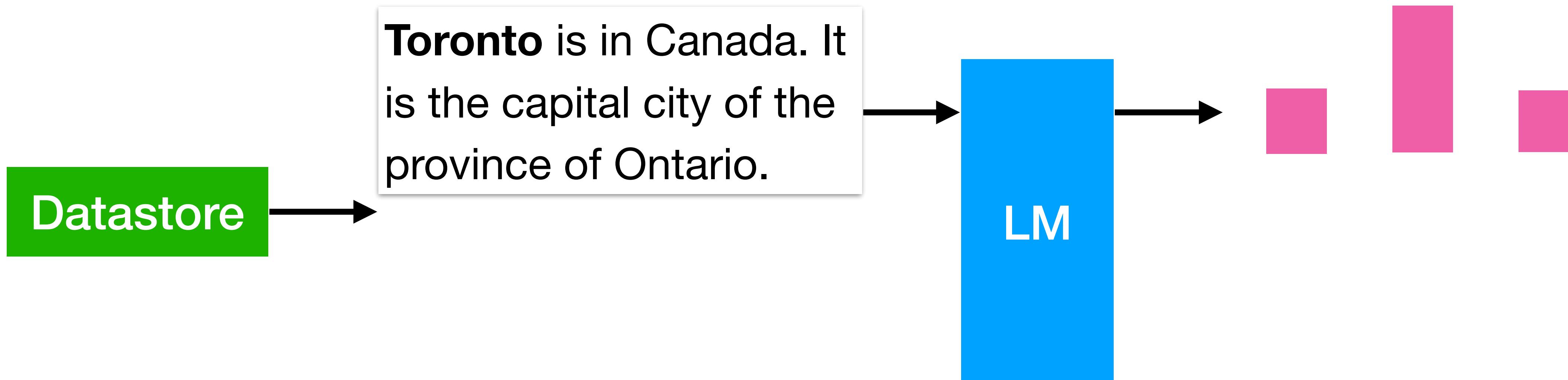


(Shi et al., 2023; Ram et al., 2022; Mallen et al., 2022; Yu et al., 2022; Press et al., 2022; *inter alia*)

# REPLUG (Shi et al., 2023; Section 3&4)

✗ What is the capital of Ontario?

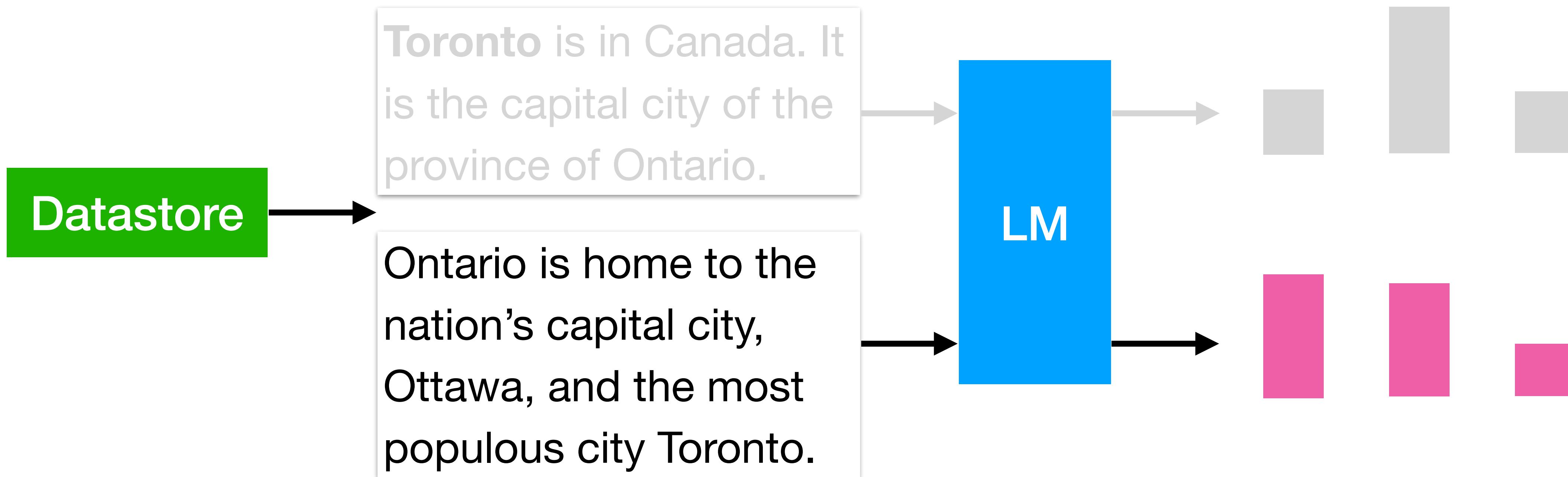
Ottawa **Toronto** Ontario



# REPLUG (Shi et al., 2023; Section 3&4)

✗ What is the capital of Ontario?

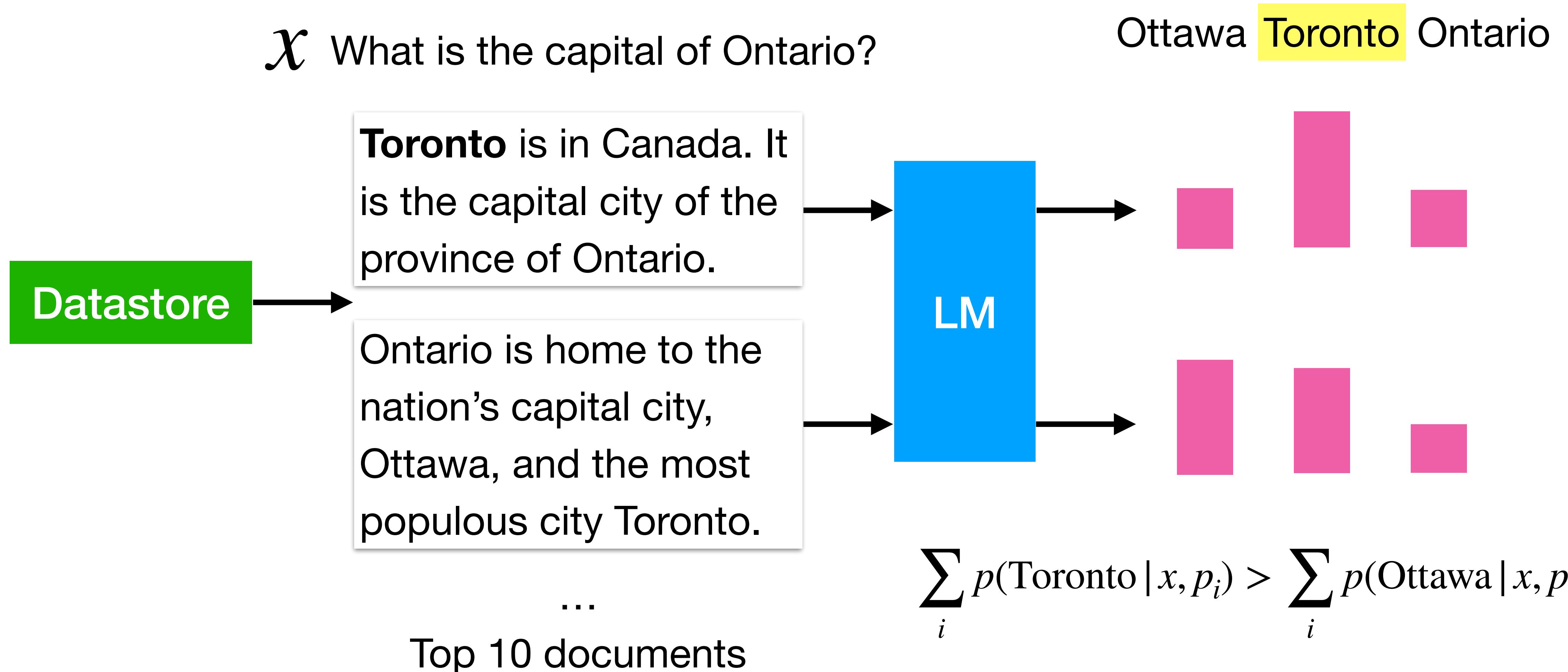
Ottawa Toronto Ontario



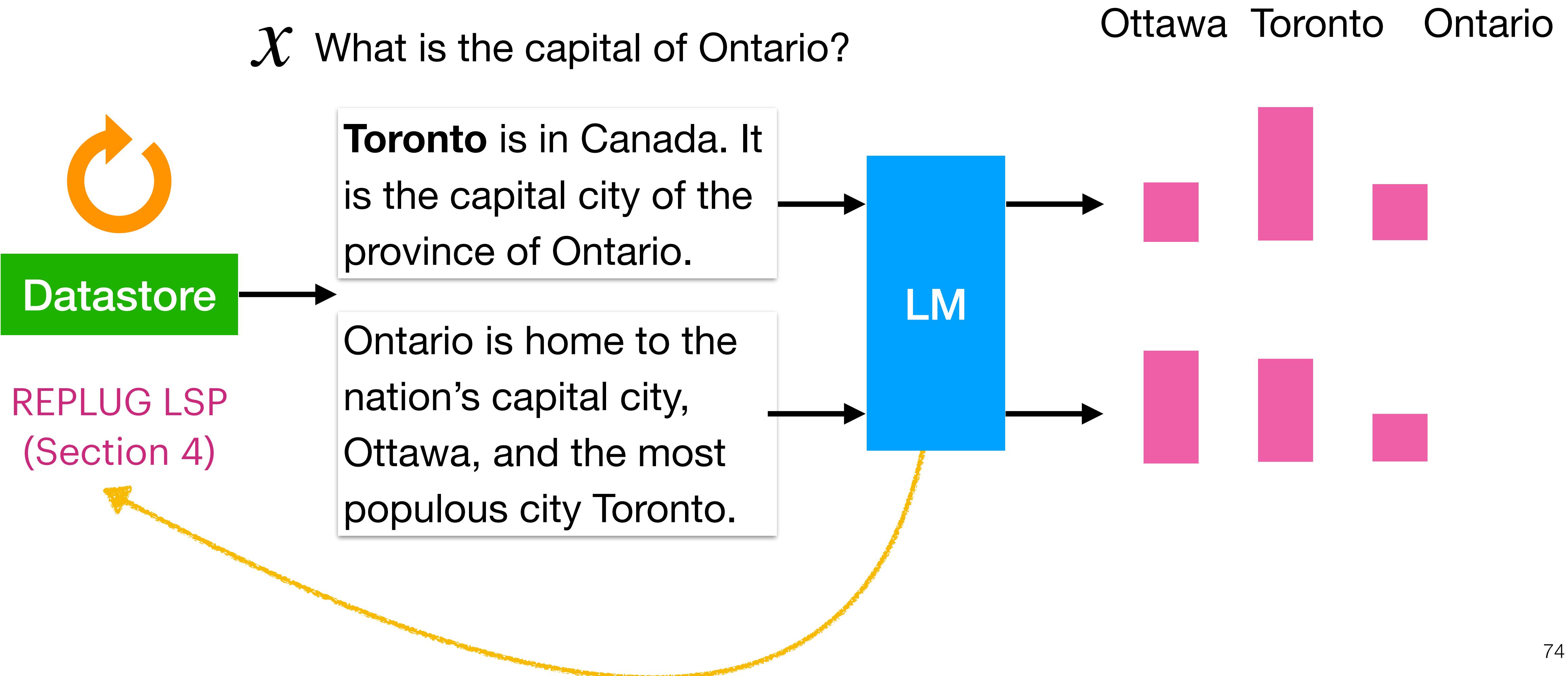
...

Top 10 documents

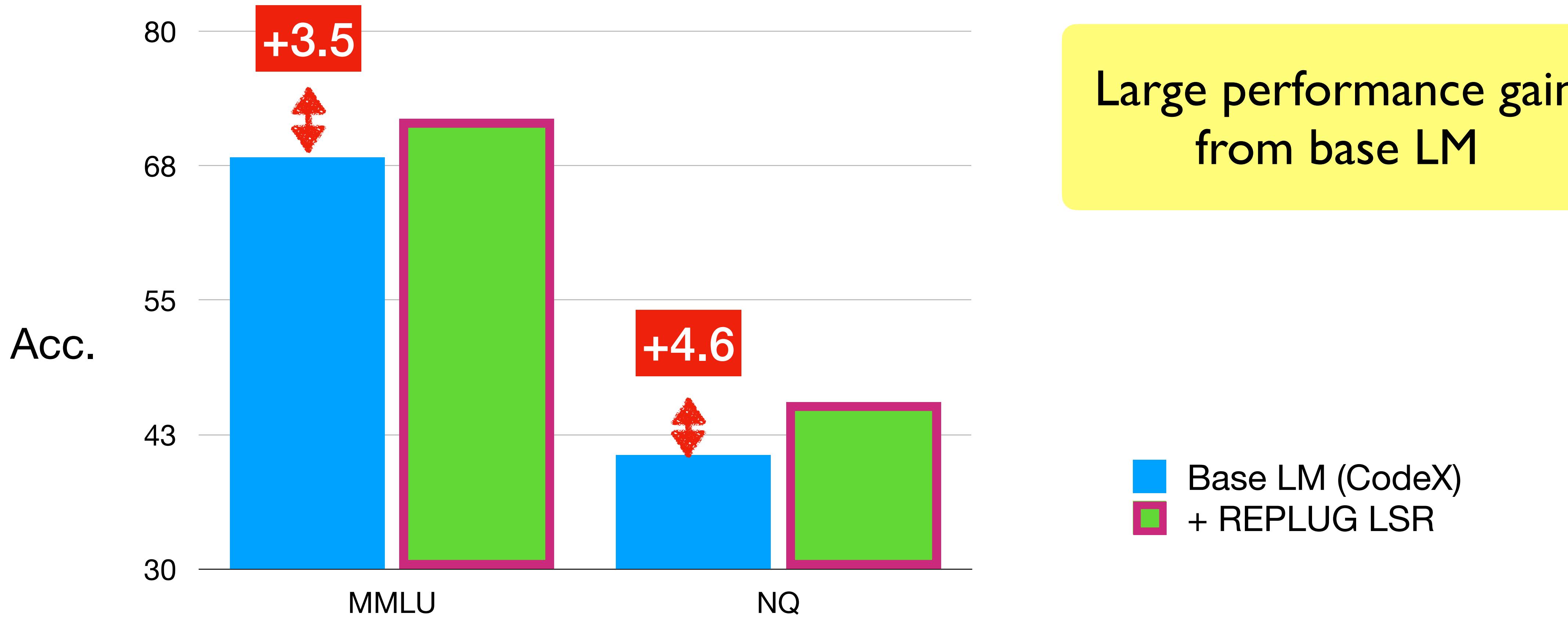
# REPLUG (Shi et al., 2023; Section 3&4)



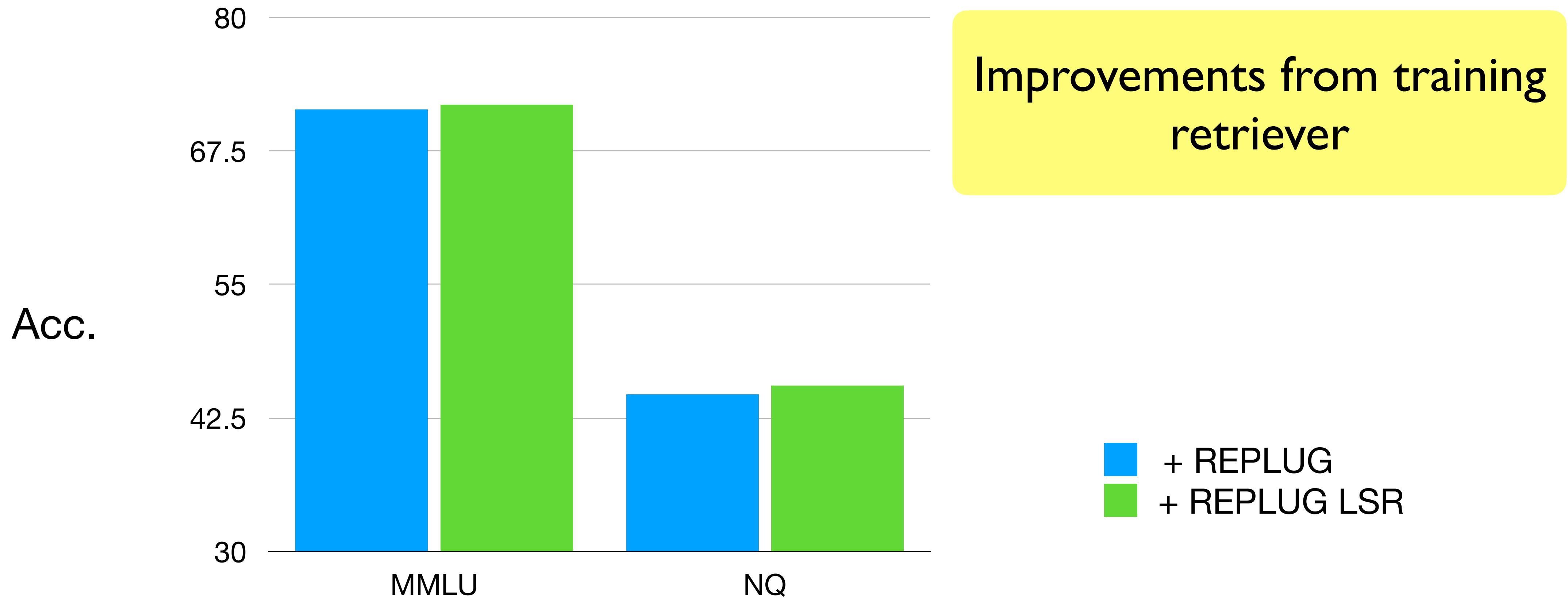
# REPLUG (Shi et al., 2023; Section 3&4)



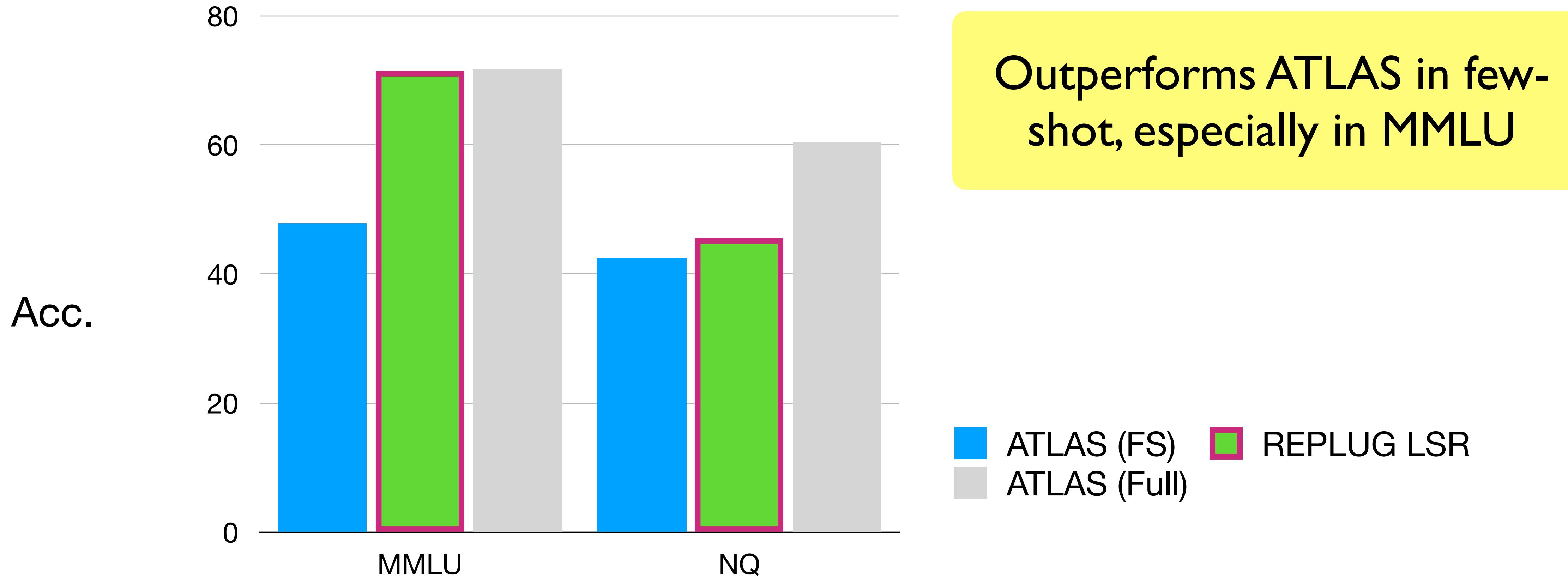
# REPLUG: results on QA & MMLU



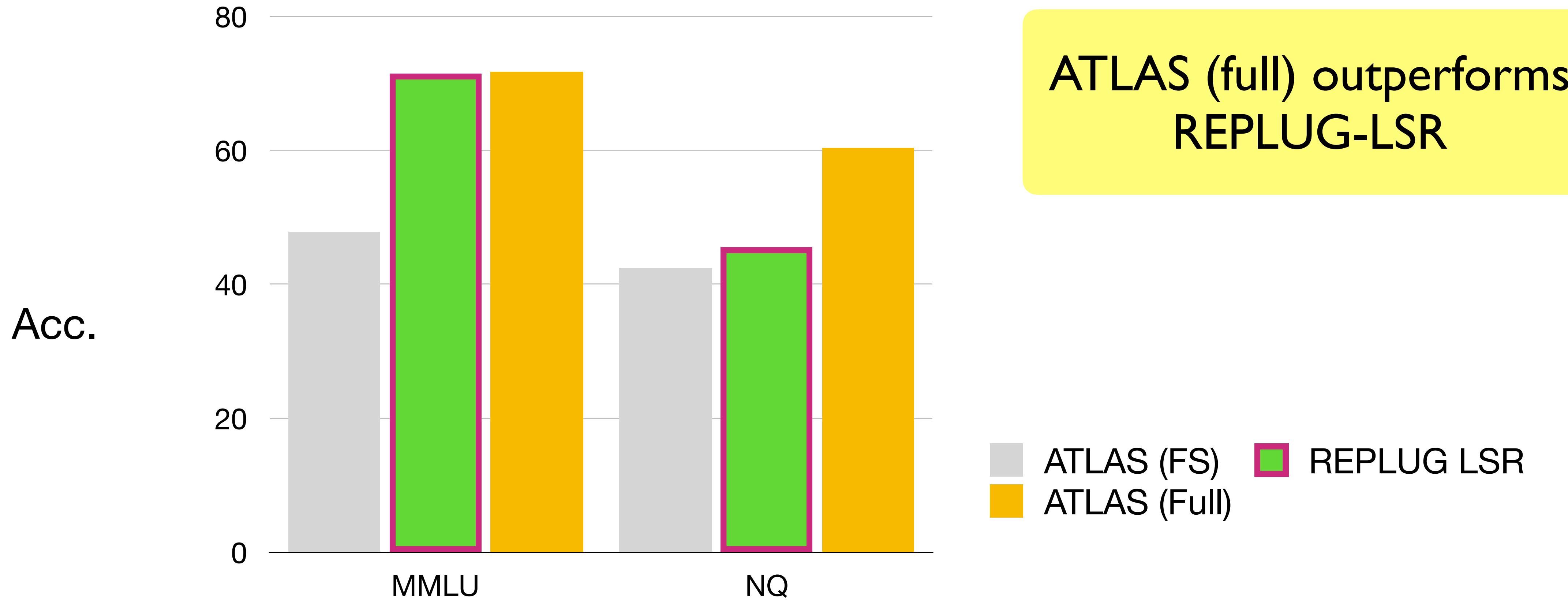
# REPLUG ablations of retriever training



# REPLUG: comparison with ATLAS



# REPLUG: comparison with ATLAS



# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC
GopherCite (Menick et al., 2022)	Open-domain QA, Long-form QA	Fine-tuning + RL (LM)	Google Search Results
kNN Prompt (Shi et al., 2022)	Classification	Prompting (output)	Wikipedia   CC
REPLUG (Shi et al., 2023)	Knowledge-intensive	Prompting (input)	Wikipedia   CC

**Benefit of retrieval-based prompting**



No training & strong performance



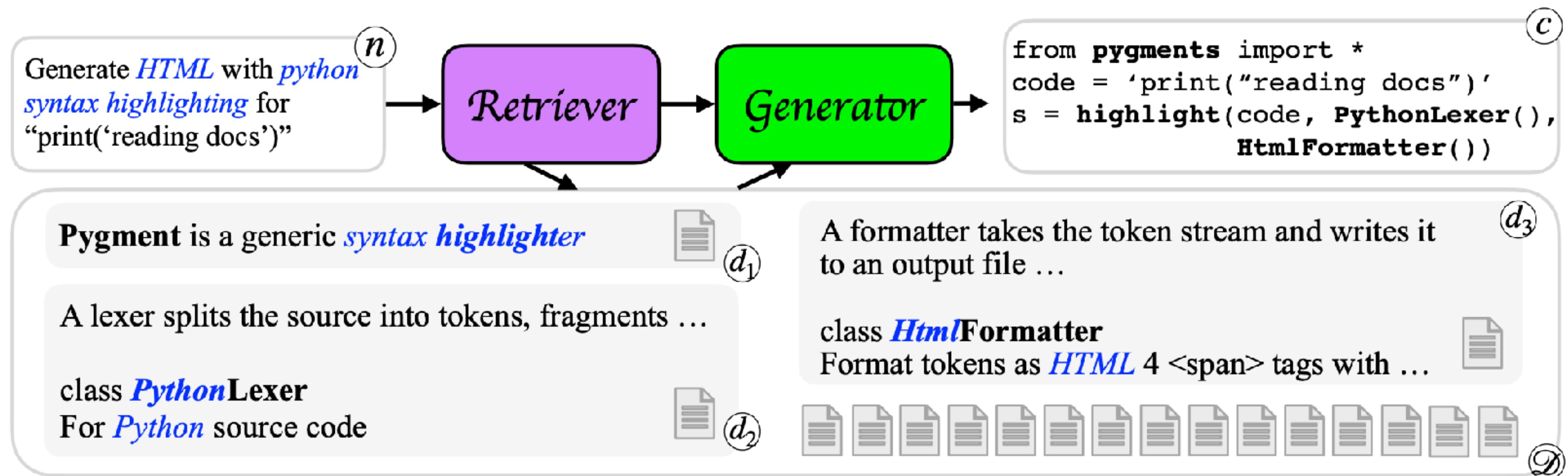
Hard to control, underperforming full FT model

# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC
GopherCite (Menick et al., 2022)	Open-domain QA, Long-form QA	Fine-tuning + RL (LM)	Google Search Results
kNN Prompt (Shi et al., 2022)	Classification	Prompting (output)	Wikipedia   CC
REPLUG (Shi et al., 2023)	Knowledge-intensive	Prompting (input)	Wikipedia   CC

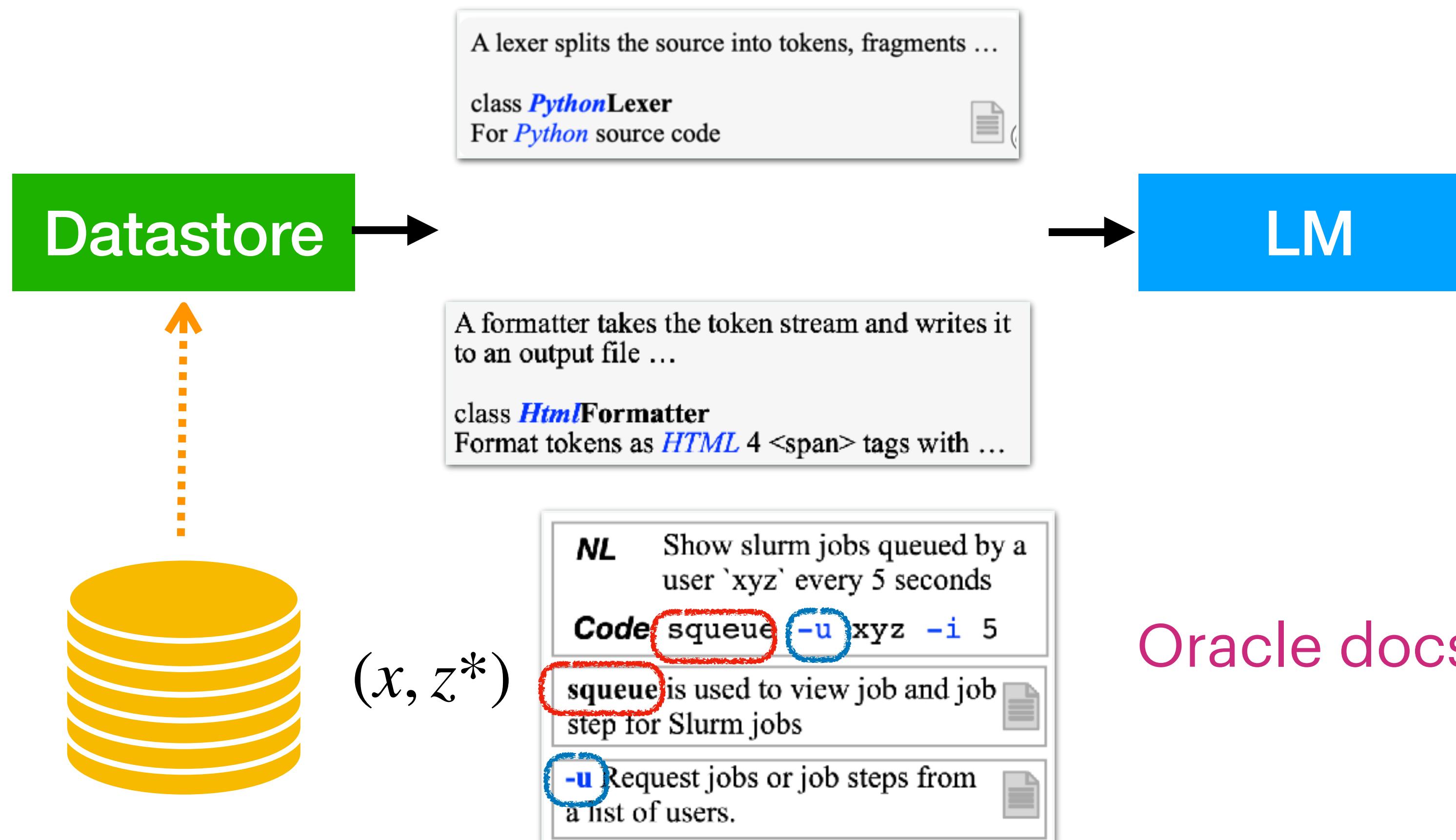
What can be other types of datastores?

# DocPrompting (Zhou et al., 2023)



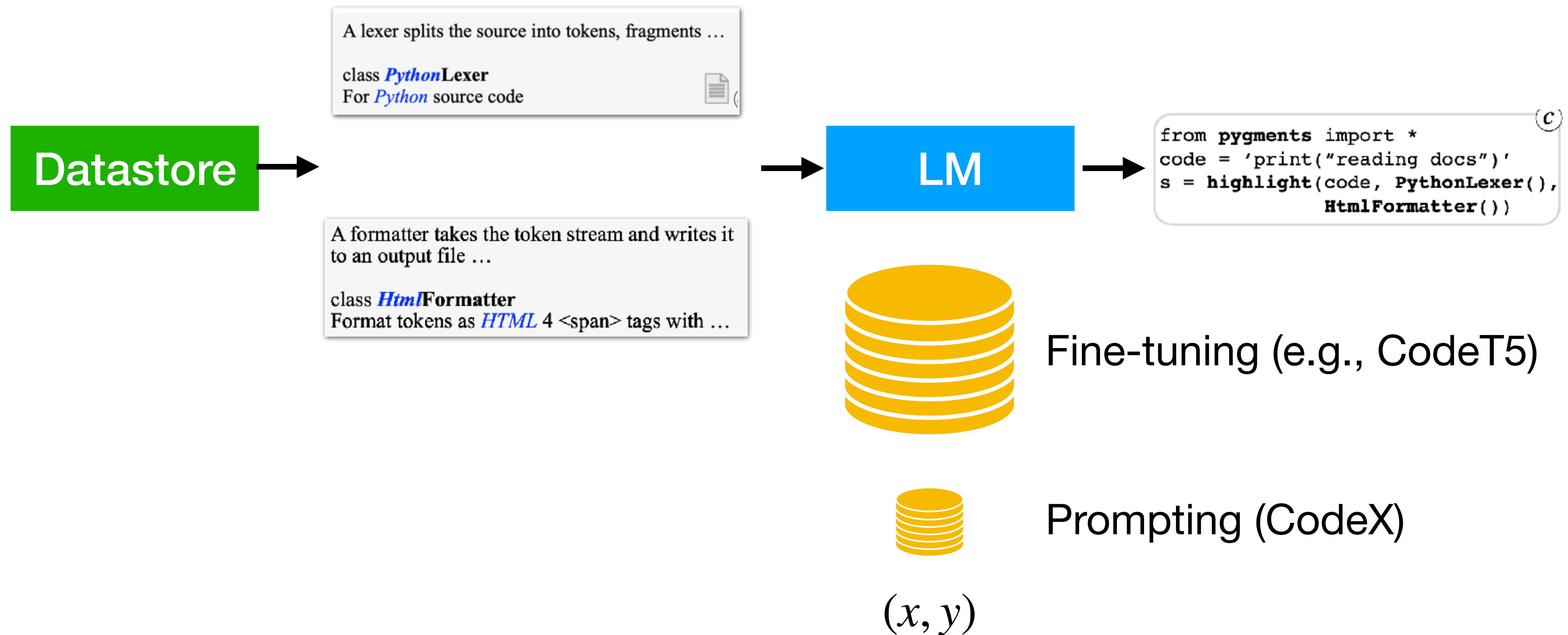
Retrieve **code documentation**s about related functions

# DocPrompting (Zhou et al., 2023)

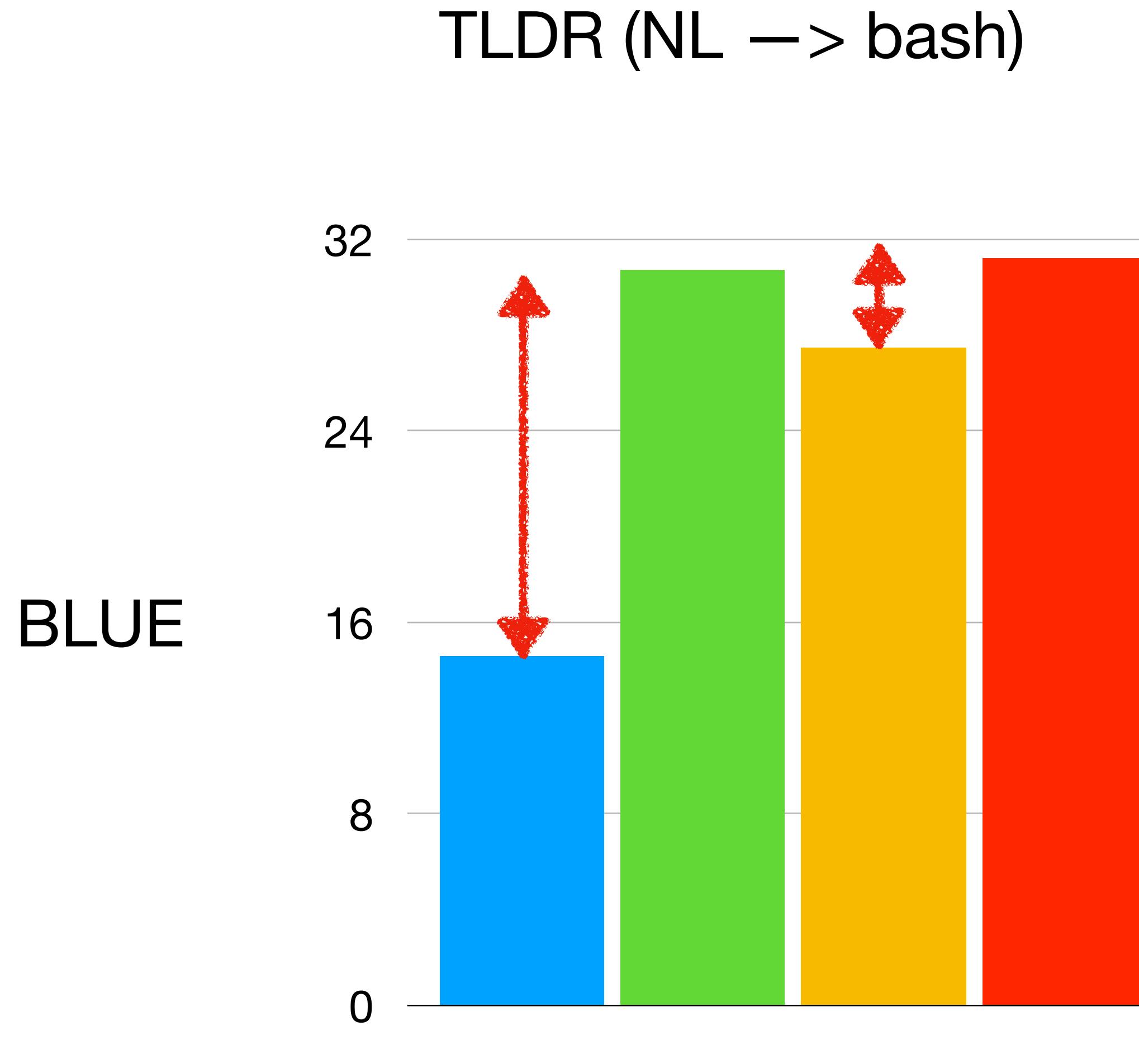


Oracle docs retrieved by simple text matching

# DocPrompting (Zhou et al., 2023)



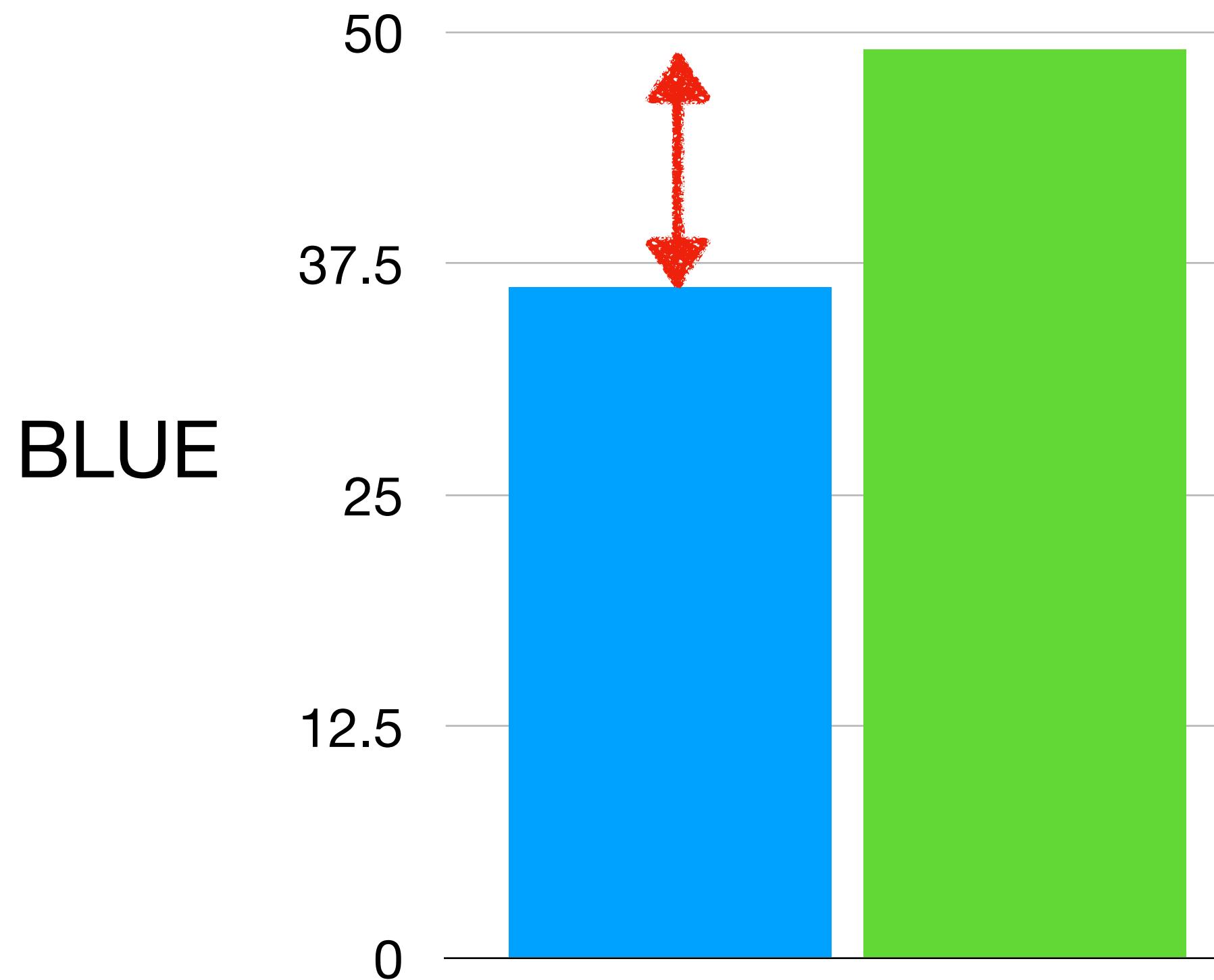
# DocPrompting (Zhou et al., 2023)



Large gain given by DocPrompting  
for both CodeT5 & CodeX

# DocPrompting (Zhou et al., 2023)

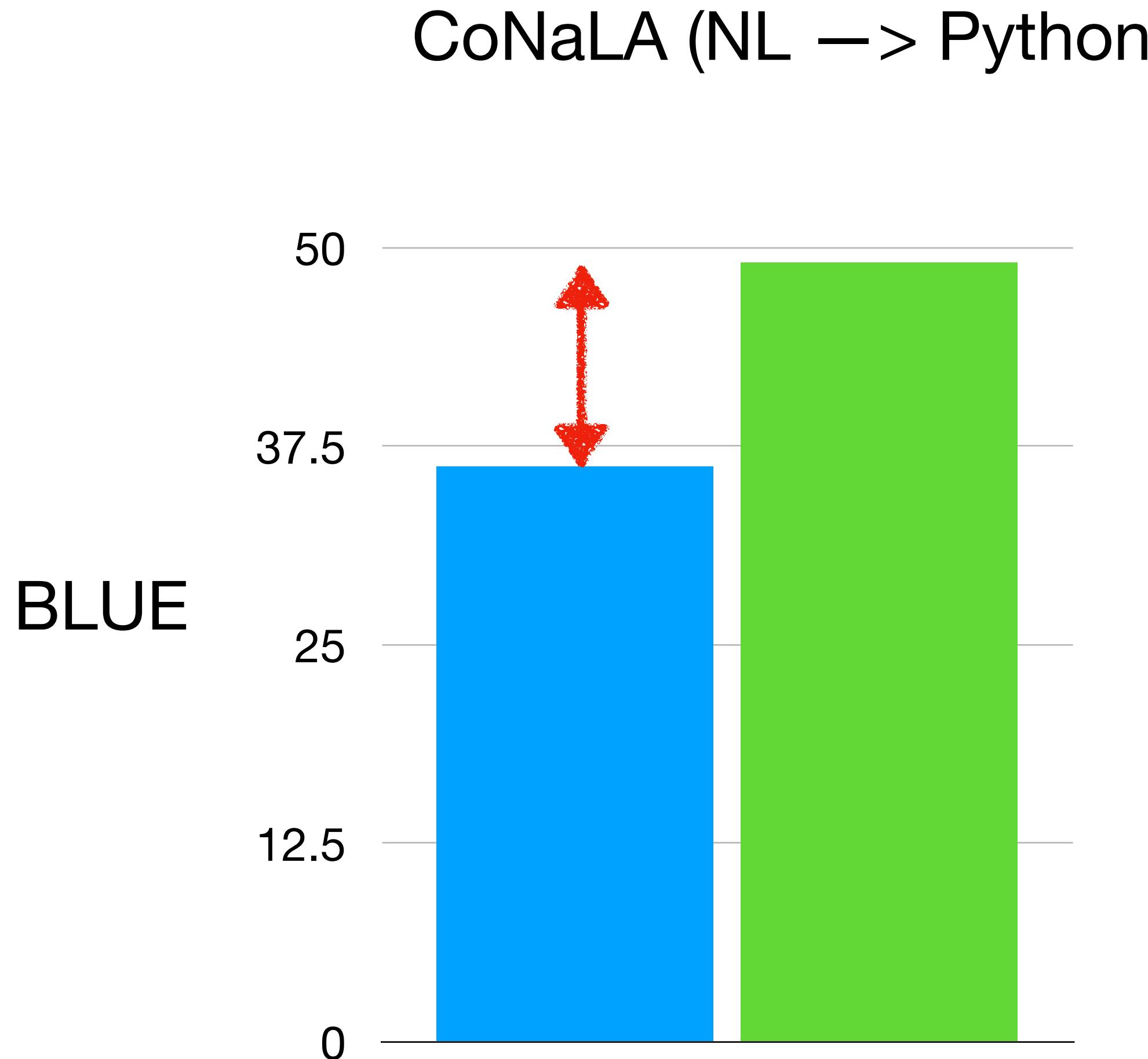
CoNaLA (NL → Python)



Room for improvement for retrieval model

- + DocPrompting
- + DocPrompting (Oracle)

# DocPrompting (Zhou et al., 2023)



Room for improvement for retrieval model

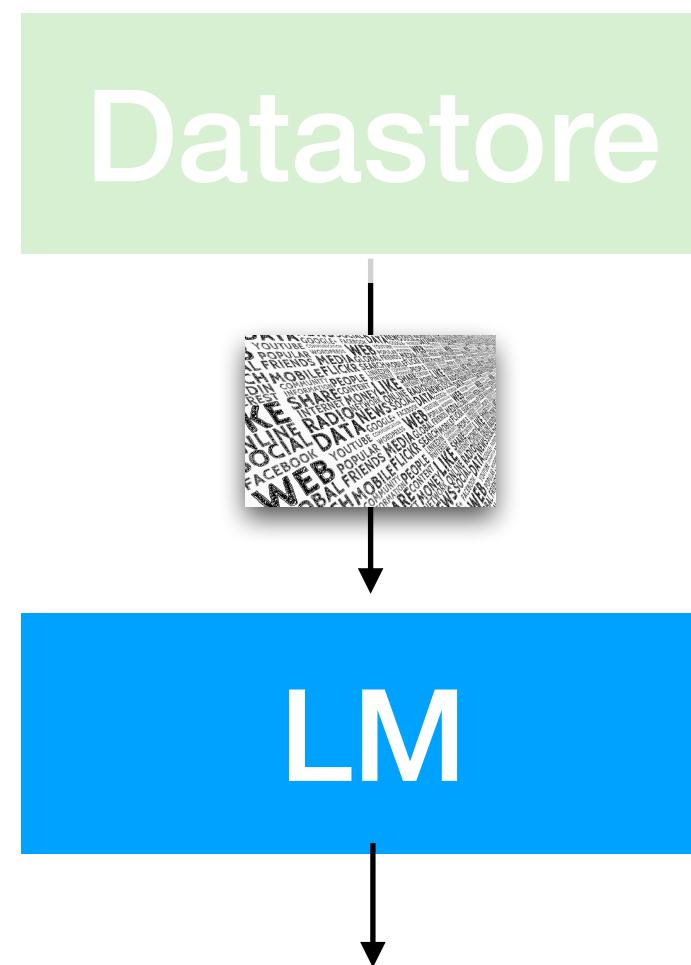
Active research in OOD / Zero-shot retrieval!  
(BEIR; Thakur et al., 2021)

- + DocPrompting
- + DocPrompting (Oracle)

# Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (DS & LM)	Wikipedia   CC
GopherCite (Menick et al., 2022)	Open-domain QA, Long-form QA	Fine-tuning + RL (LM)	Google Search Results
kNN Prompt (Shi et al., 2022)	Classification	Prompting (output)	Wikipedia   CC
REPLUG (Shi et al., 2023)	Knowledge-intensive	Prompting (input)	Wikipedia   CC
DocPrompting (Zhou et al., 2023)	Code Generation	Fine-tuning (DS & LM), Prompting (Input)	Code documentations

# How to adapt a retrieval-based LM for a task

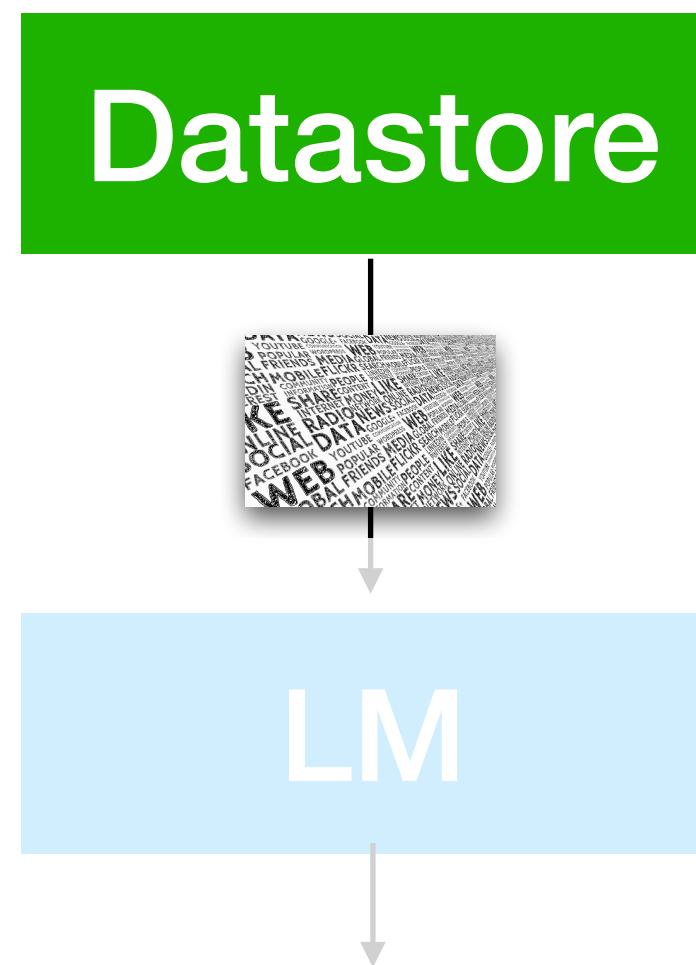


Retrieval-based prompting is competitive



Fine-tuning (+ RL) often show comparable / better performance & is more customizable

# How to adapt a retrieval-based LM for a task



⟳ Training a **retriever** on downstream tasks helps both fine-tuning and prompting

**Datastore** can be diverse (also in [Section 6](#)) while challenges remain in OOD retrieval

# Two key questions for downstream adaptations

**How** can we adapt a retrieval-based LM for a task?

**When** should we use a retrieval-based LM?

# **When** to use a retrieval-based LM

Long-tail

knowledge  
update

Verifiability

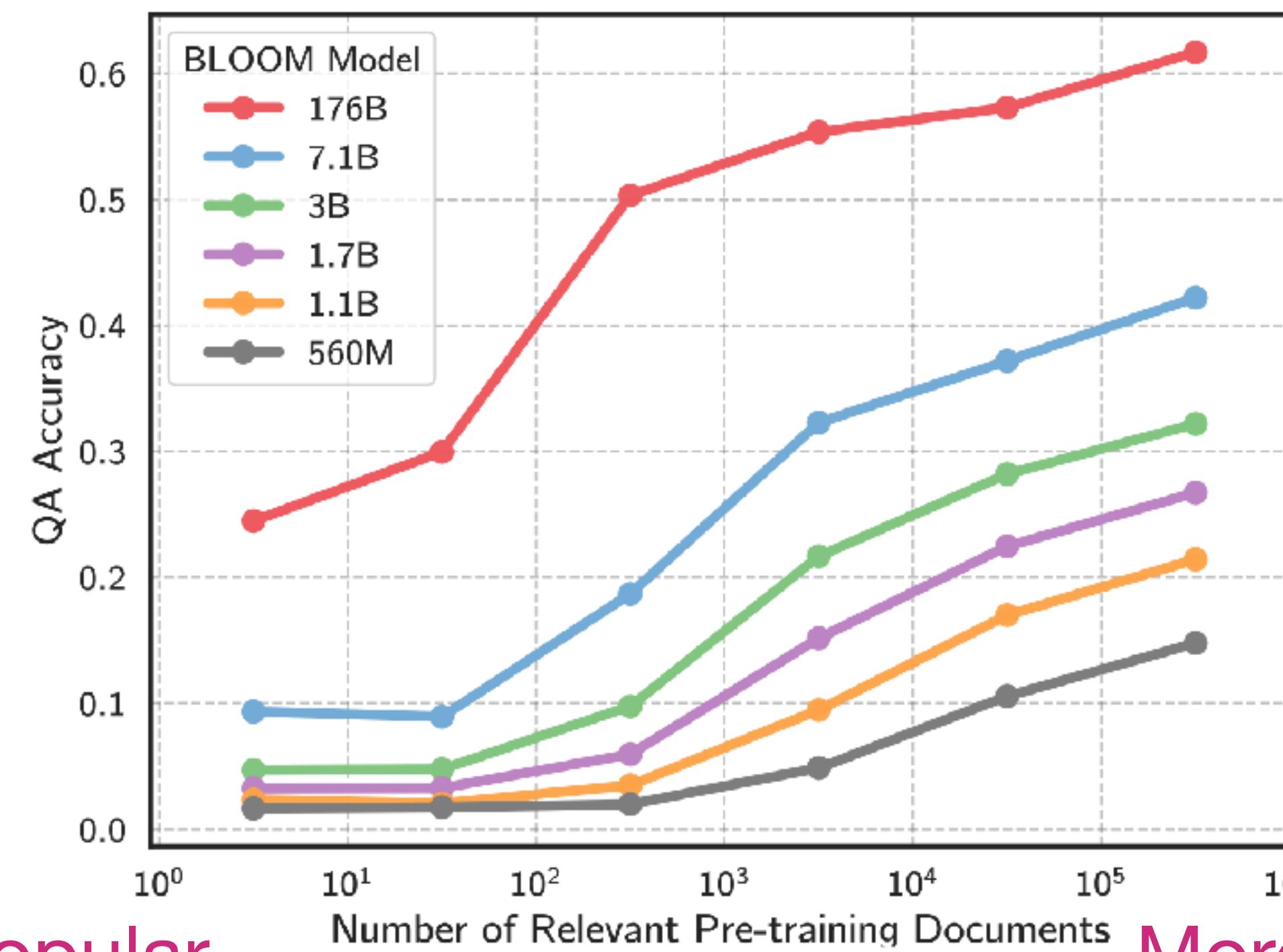
Parameter-  
efficiency

Privacy

# Key effectiveness in downstream tasks

Long-tail

LLMs often struggle in long **tail / less frequent entities**



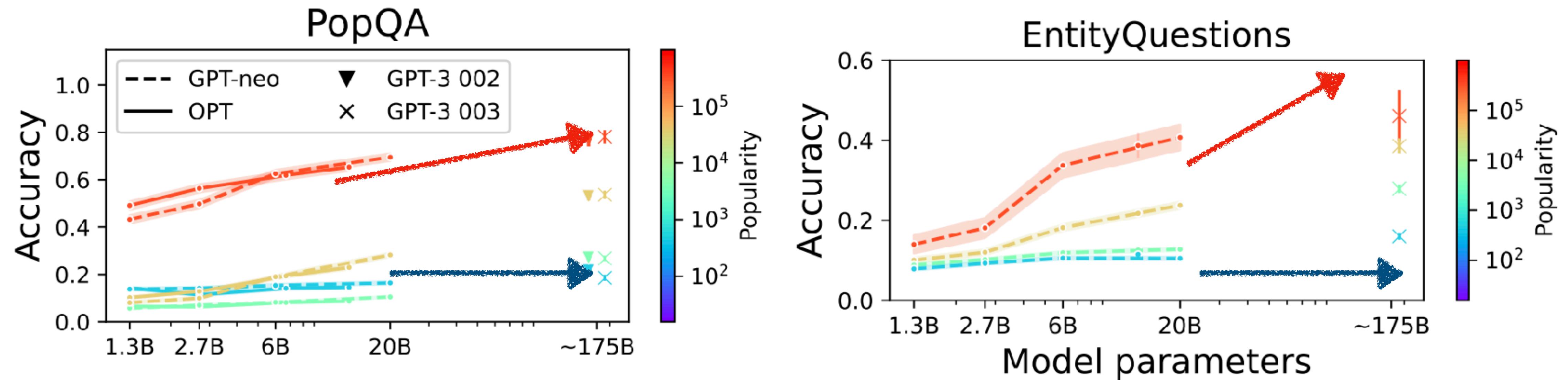
<— less popular

More popular —>

# Key effectiveness in downstream tasks

Long-tail

Scaling LLMs only helps for **popular knowledge**; for long tail, scaling gives marginal performance improvements

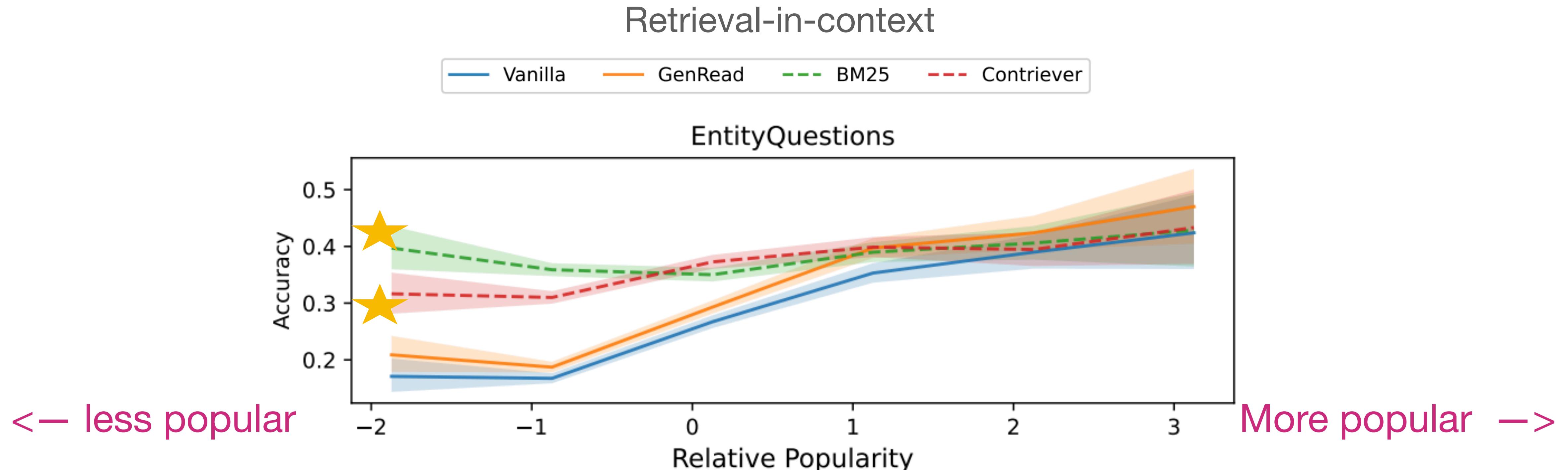


Mallen\* and Asai\* et al. 2023. “When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories”

# Key effectiveness in downstream tasks

Long-tail

Retrieval gives large performance gain in such **long-tail**



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

# Key effectiveness in downstream tasks

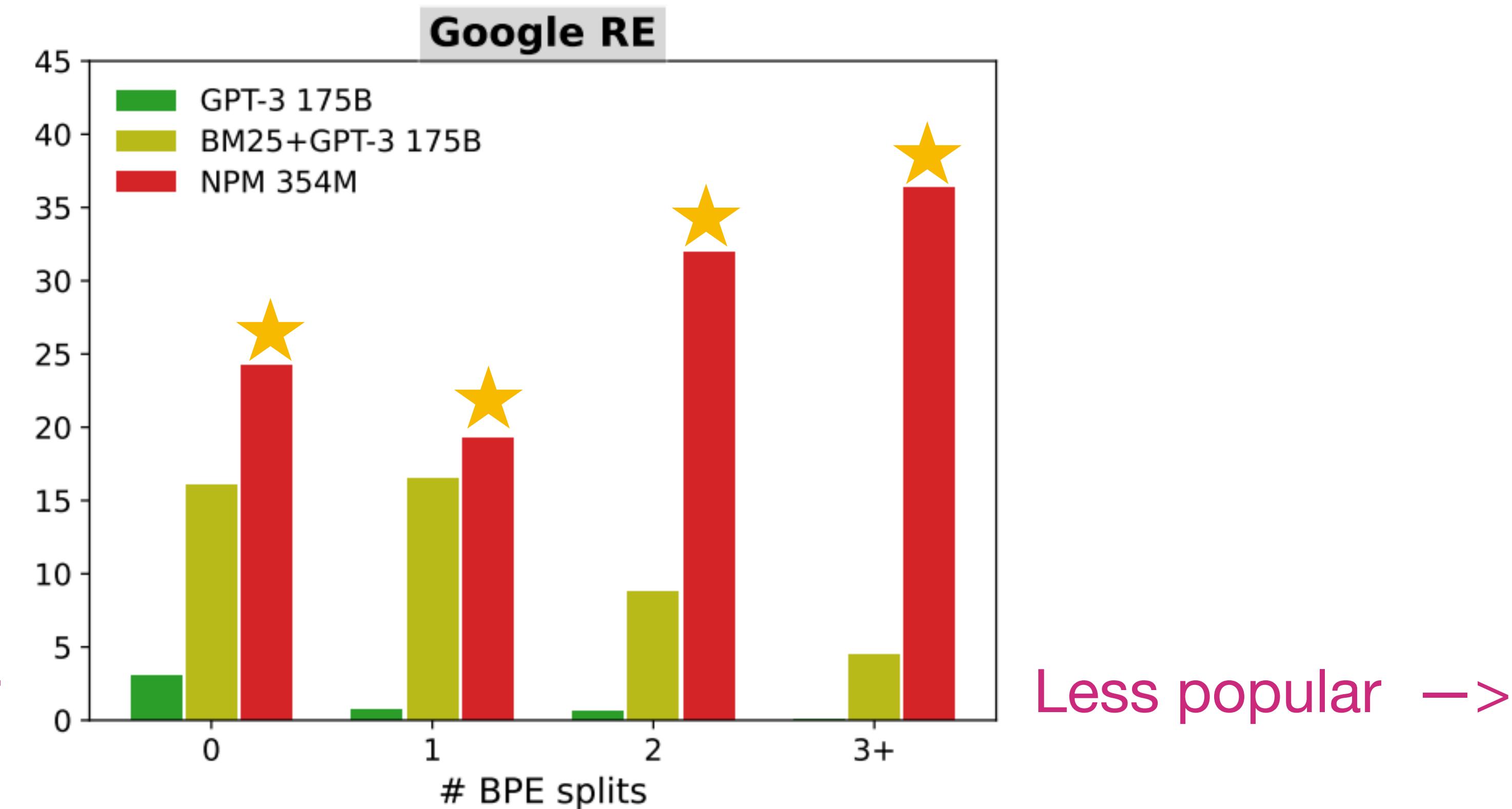
Long-tail

Retrieval gives large performance gain in such **long-tail**

**Output** space  
(e.g., kNN, NPM)

<— more popular

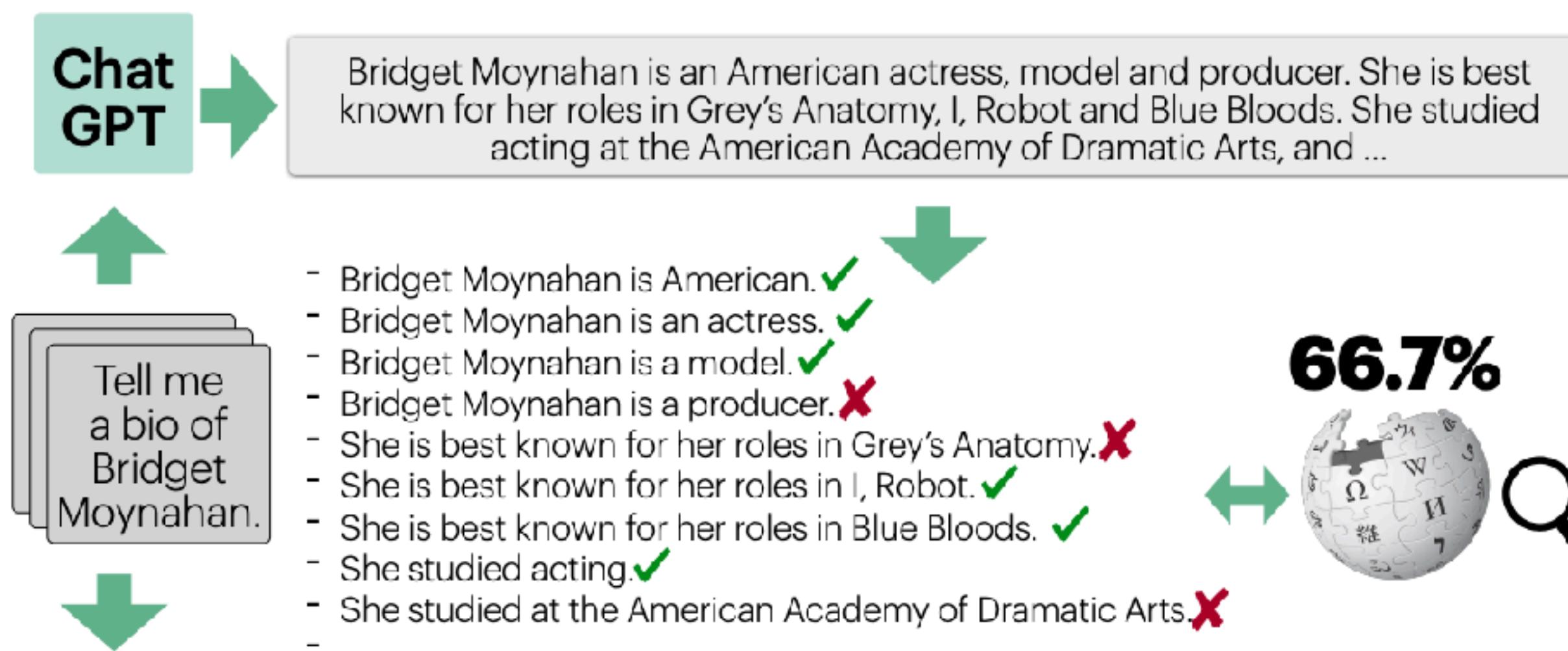
Less popular —>



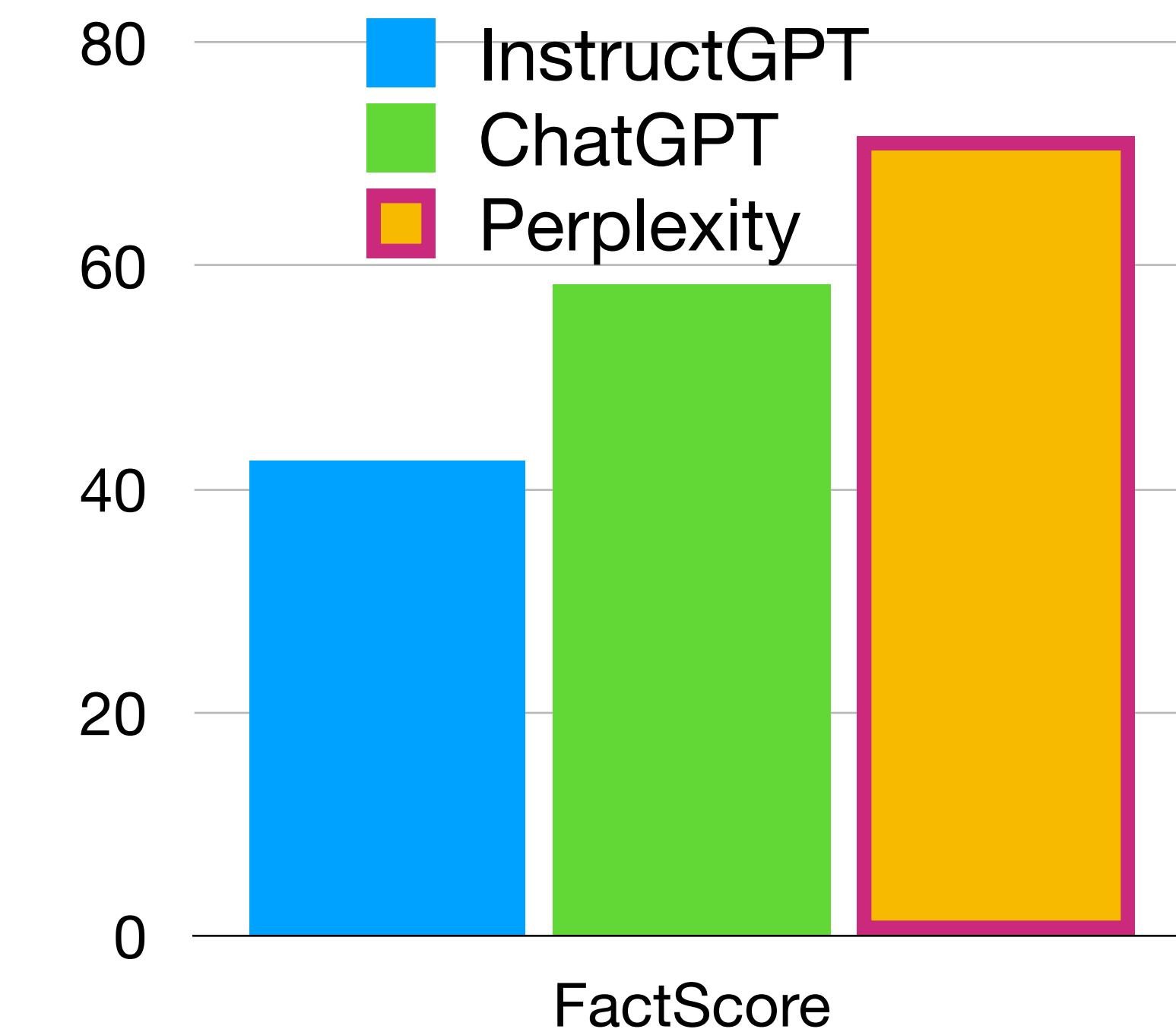
# Key effectiveness in downstream tasks

## Long-tail

Largely reduce hallucinations in **long-form generations**



FactScore



# Key effectiveness in downstream tasks

Update

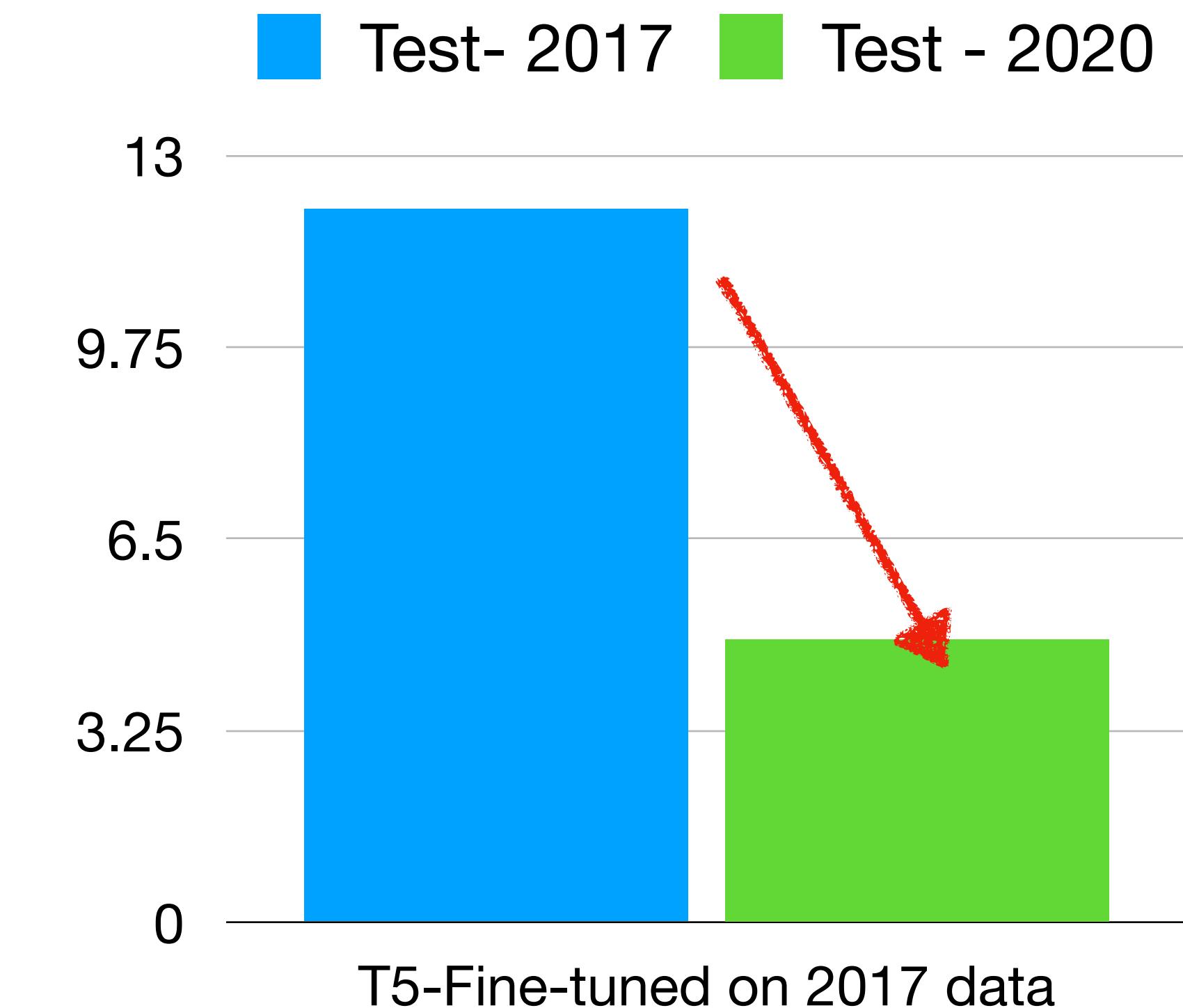
Standard LLMs needs to be **trained again** to adapt to evolving world knowledge

Temp LAMA

2012	Cristiano Ronaldo plays for _X_.	Real Madrid
2019	Cristiano Ronaldo plays for _X_.	Juventus FC

Huge performance drop when test knowledge needs to be updated

Izacard et al. 2022. “Few-shot learning with retrieval augmented language models”



# Key effectiveness in downstream tasks

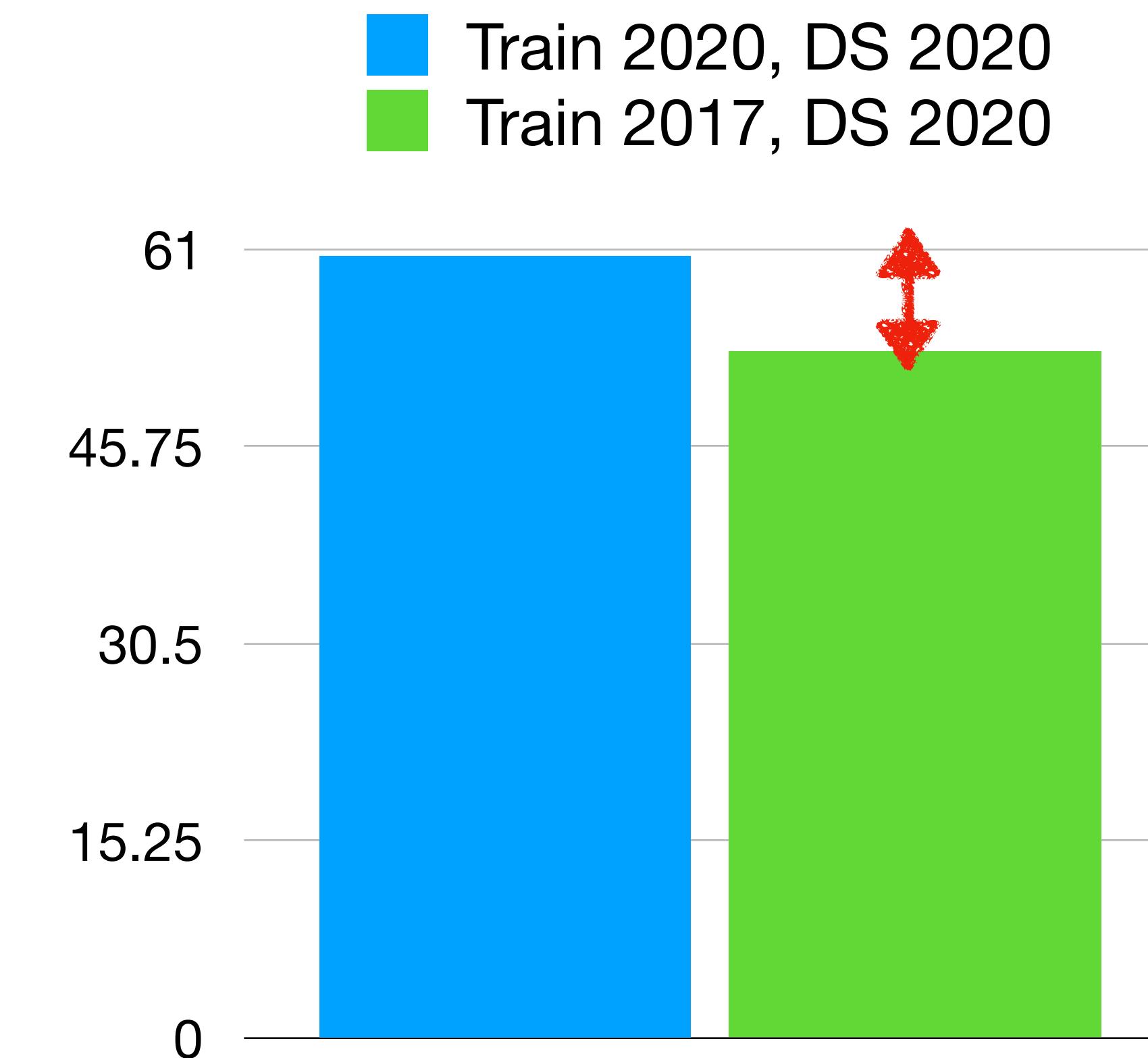
Update

Swapping the knowledge corpus to **accommodate temporal changes** without additional training.

Temp LAMA

2012	Cristiano Ronaldo plays for _X_.	Real Madrid
2019	Cristiano Ronaldo plays for _X_.	Juventus FC

Swapping test datastore only retains strong performance

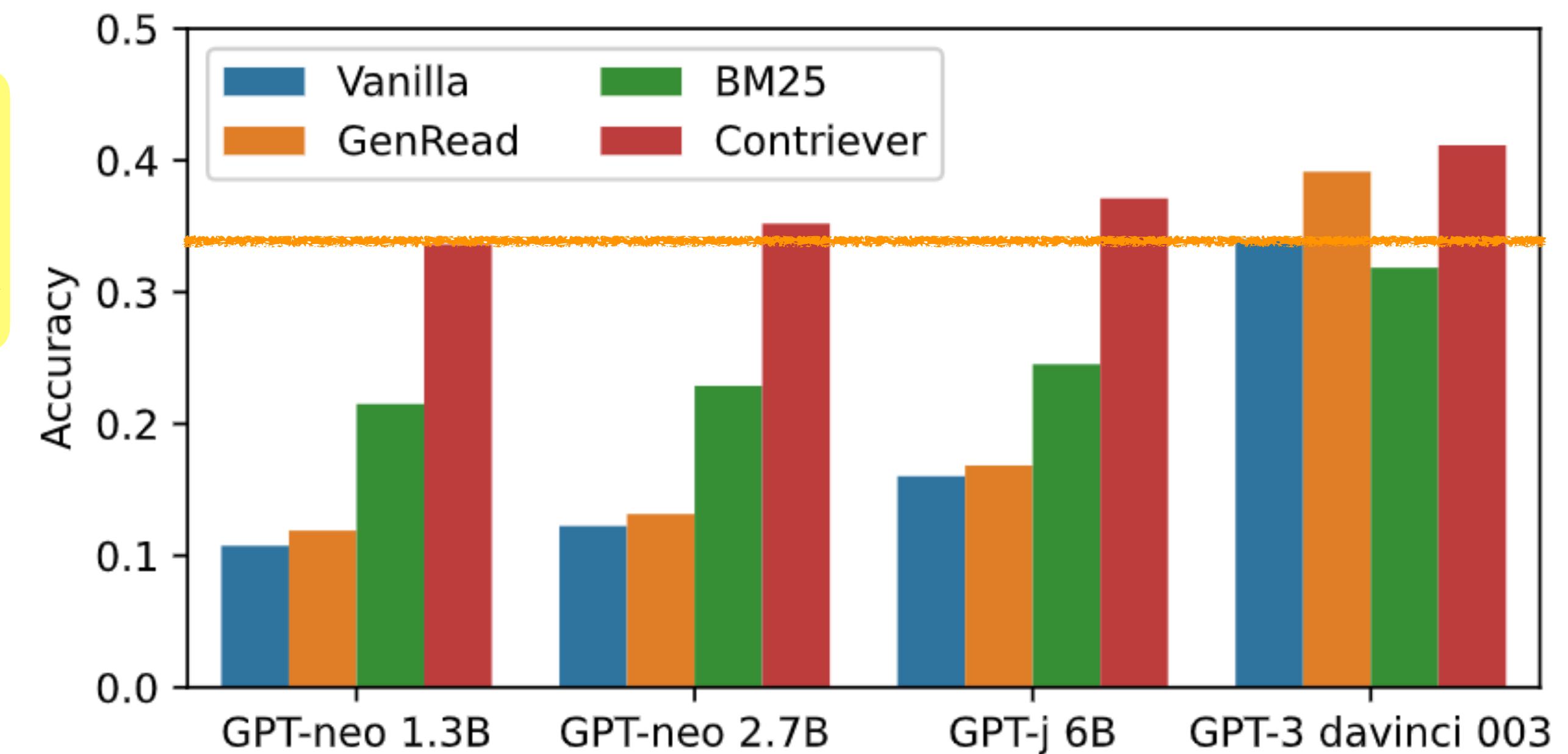


# Key effectiveness in downstream tasks

Parameter-  
efficiency

**Much smaller LMs with retrieval** can outperforms  
much larger LMs in knowledge-intensive tasks.

Retrieval + GPT Neo 1.3B  
outperforms vanilla GPT3 on QA



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

# Key effectiveness in downstream tasks

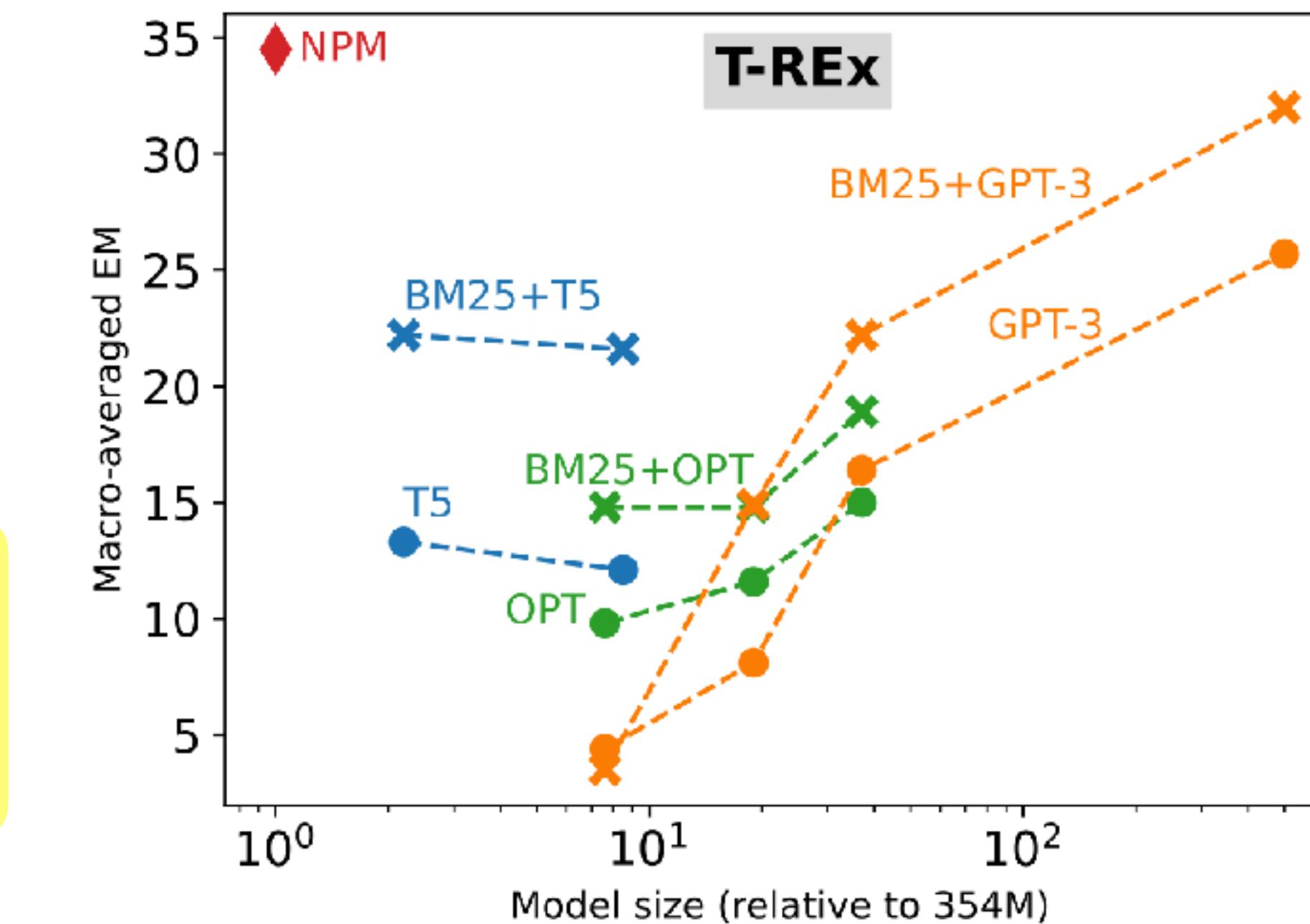
Parameter-  
efficiency

**Much smaller LMs with retrieval** can outperforms  
much larger LMs in knowledge-intensive tasks.

T-Rex

AVCDH is owned by [MASK]

NPM (354 M) outperforms GPT-3  
(175B) on T-Rex.



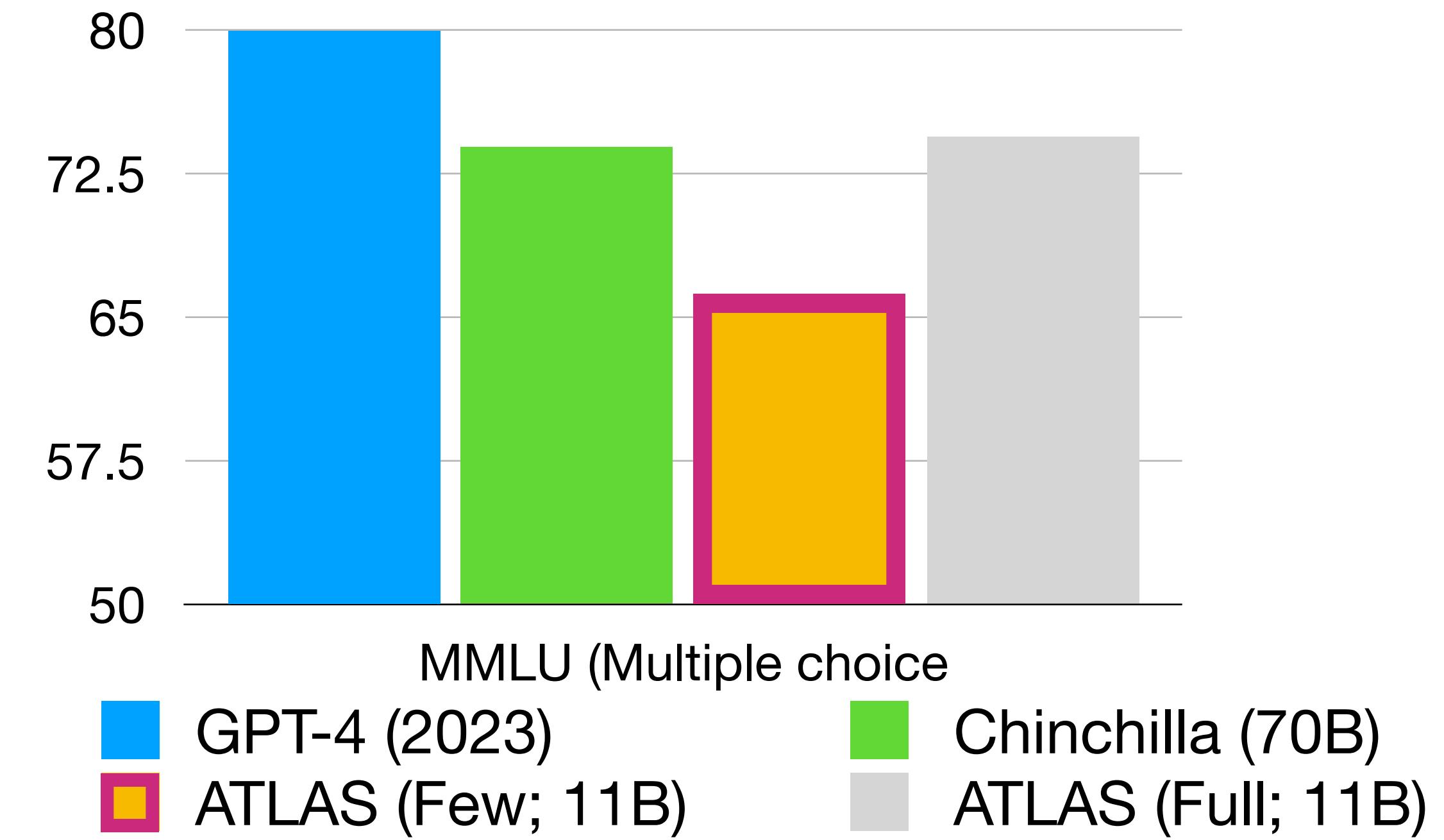
# Key effectiveness in downstream tasks

Parameter-  
efficiency

**Much smaller LMs with retrieval** can outperforms  
much larger LMs in *knowledge-intensive tasks*.

Room for improvements for  
diverse task adaptations!

Izacard et al. 2022. “Few-shot learning with  
retrieval augmented language models”

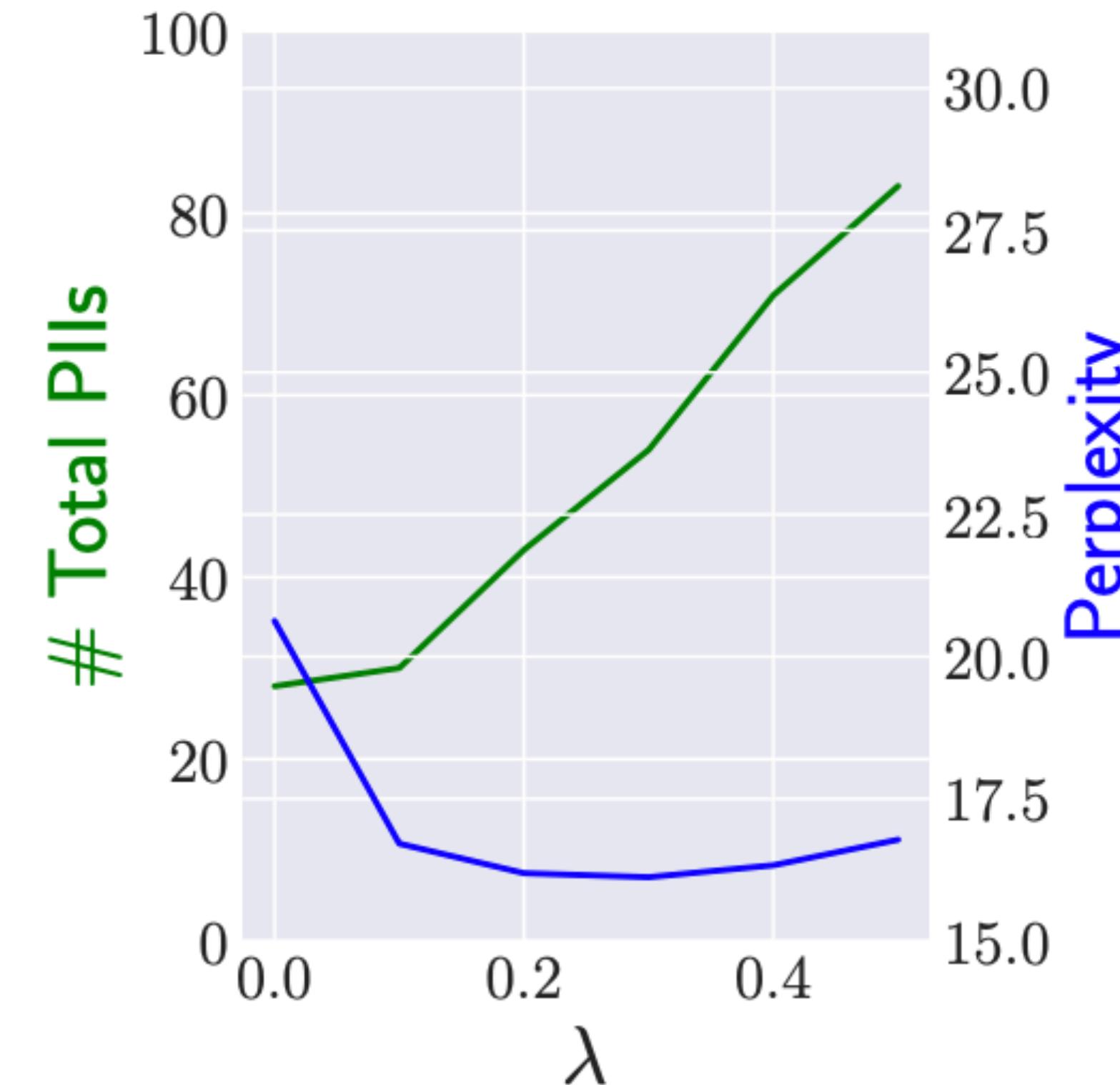


# Key effectiveness in downstream tasks

Privacy

Higher, more risk

Retrieval-based LMs enable us to mitigate privacy risks.



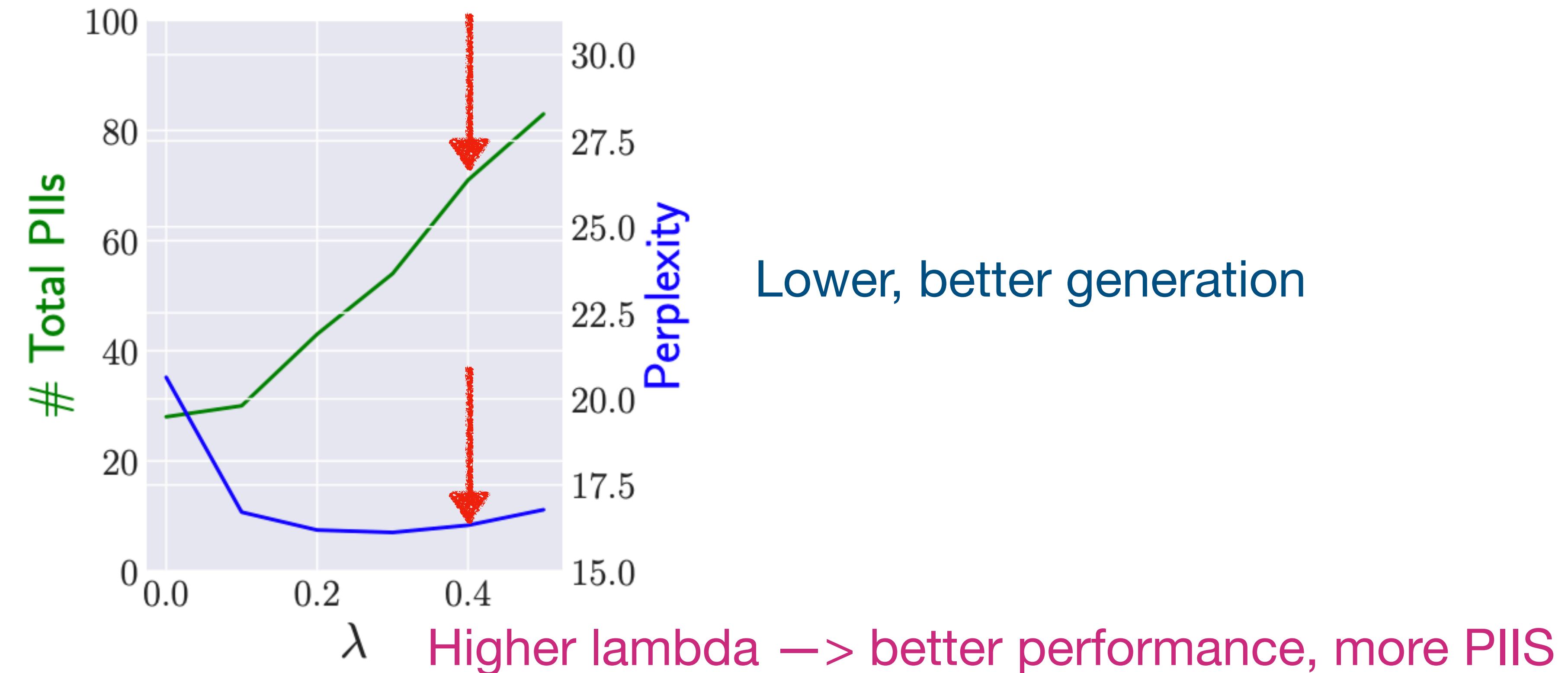
Lower, better generation

# Key effectiveness in downstream tasks

Privacy

Higher, more risk

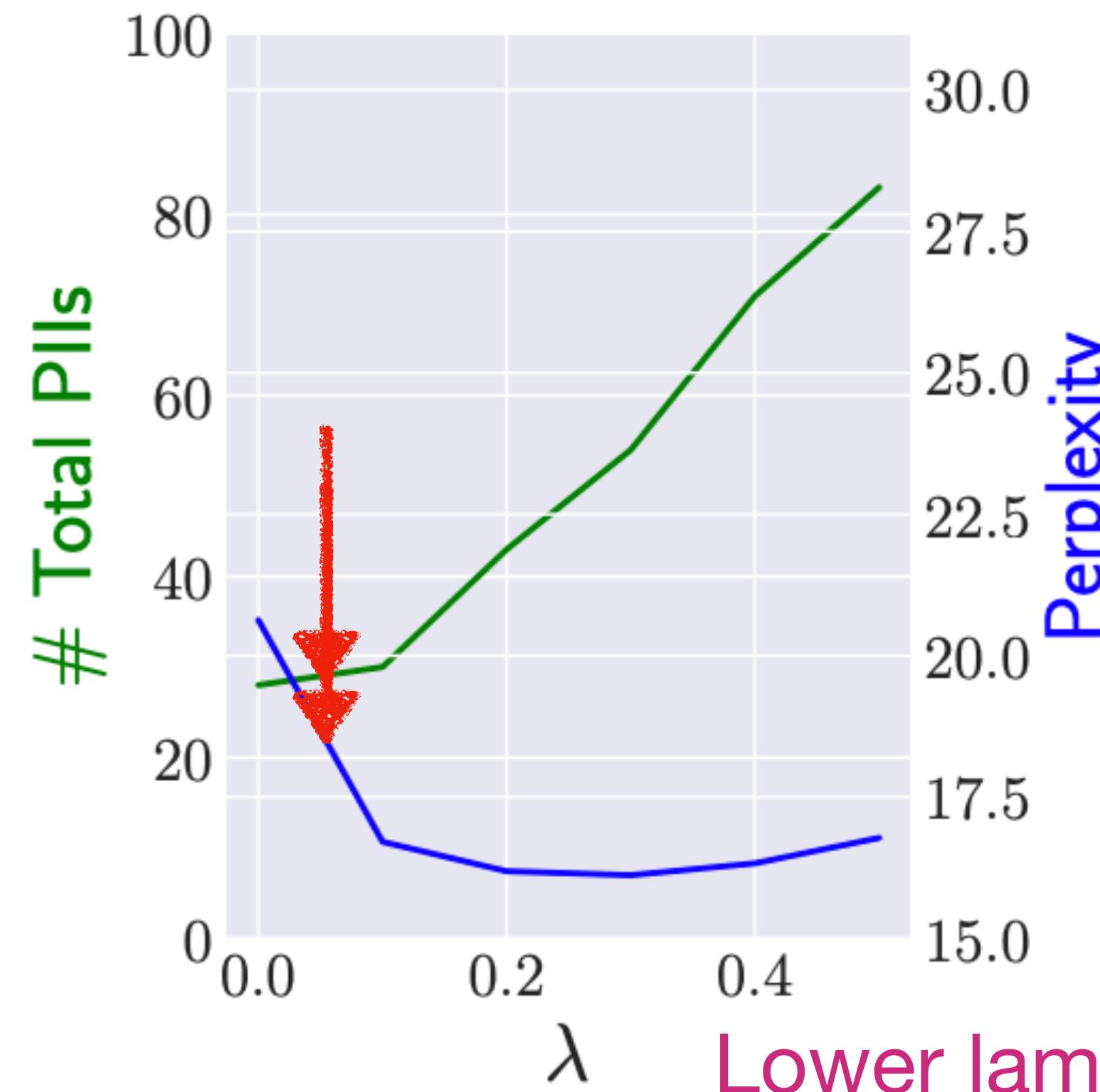
Retrieval-based LMs enable us to mitigate privacy risks.



# Key effectiveness in downstream tasks

Privacy

Higher, more risk



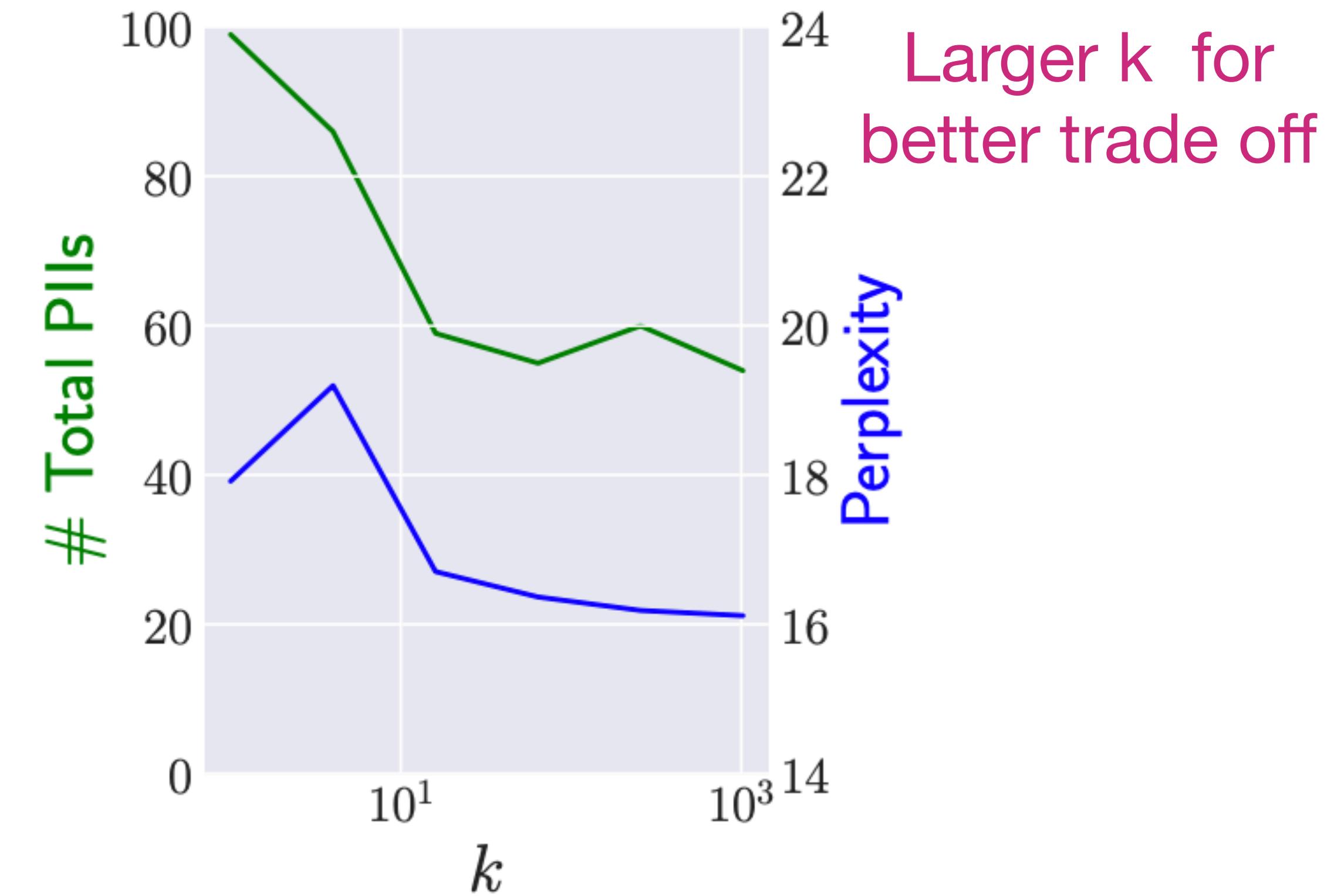
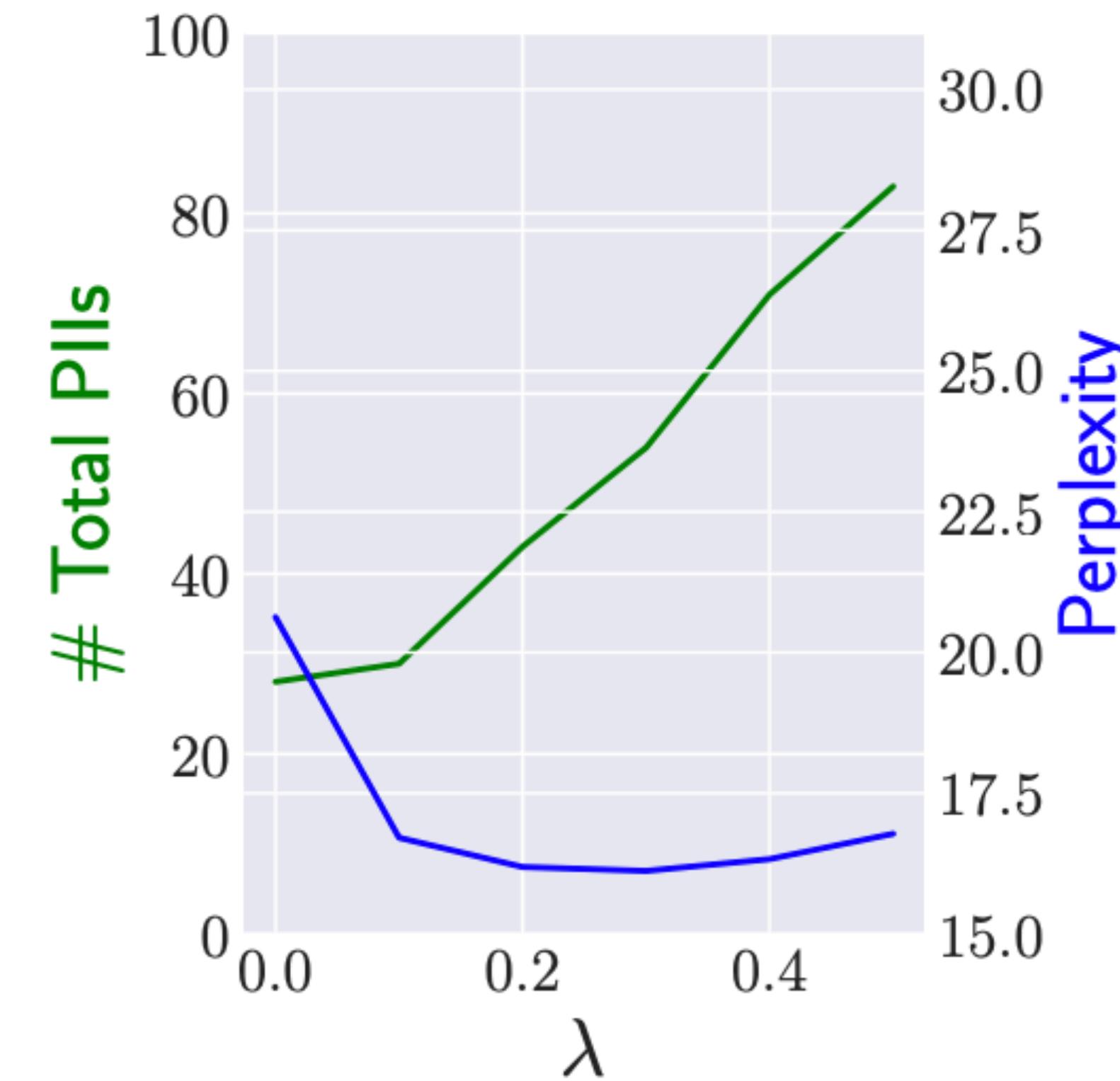
Retrieval-based LMs enable us to mitigate privacy risks.

Lower, better generation

# Key effectiveness in downstream tasks

Privacy

Retrieval-based LMs enable us to mitigate privacy risks.



# Key effectiveness in downstream tasks

## Verifiability

Human and model can reliably assess the **factuality of the generations** using the retrieved evidence.

Why is it sometimes hard to eat after not eating for a while?

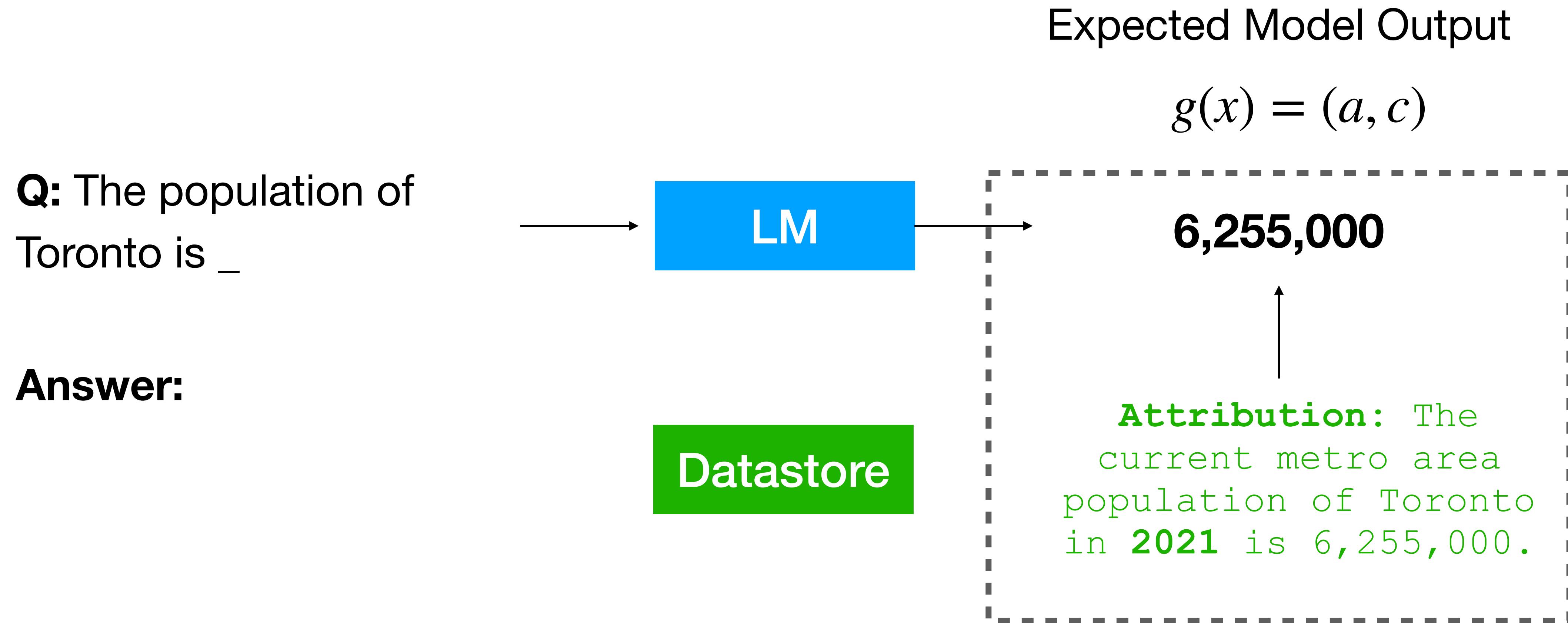
175B best-of-64

There are several reasons why not eating for long periods of time can make you lose your appetite. One is that not eating triggers the same response as starvation, adrenaline drives a process called lipolysis, which is the process your body undertakes when it wants to break down your fat reserves for energy, and also prevents you from burning through your muscle<sup>[1][3]</sup>. Another reason is that not eating means your body doesn't get the sugars it needs to operate properly, which causes a process called gluconeogenesis<sup>[2]</sup>. Also, leptin levels can rapidly decline in the fasted state<sup>[4]</sup>, which may cause an increase in appetite.

1. Why Don't You Feel Hungry After Not Eating For a Long Time? – Neuro (getneuro.com) ↗  
2. Why Don't You Feel Hungry After Not Eating For a Long Time? – Neuro (getneuro.com) ↗  
3. Why Don't You Feel Hungry After Not Eating For a Long Time? – Neuro (getneuro.com) ↗  
4. Fasting may change the body's hunger response - here's what to do about it - Diabetes (www.diabetes.co.uk) ↗

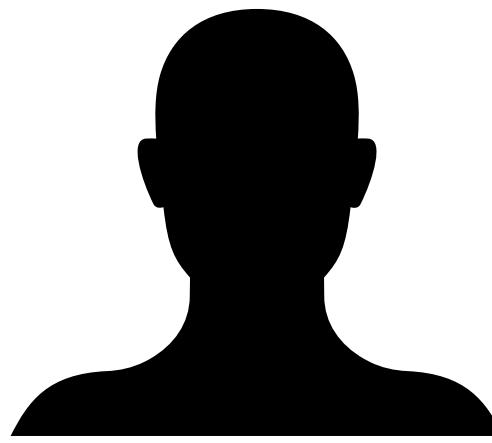
Nakano et al. 2021. “WebGPT: Browser-assisted question-answering with human feedback”

# Attributions: AttributedQA (Bohnet et al., 2022)

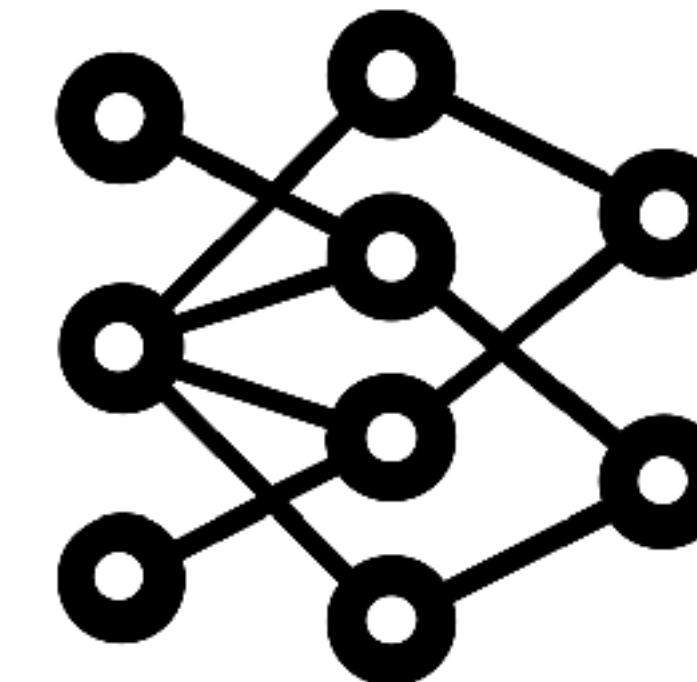


# Attributions: AttributedQA (Bohnet et al., 2022)

## Human Evaluation (AIS)



## Automatic Evaluation (Auto AIS)



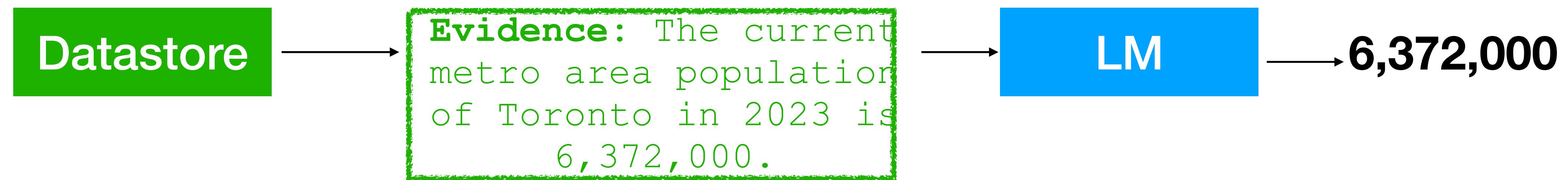
NLI model

1. Are all (a,c) interpretable?
2. Is any information in a supported by c?

$$E^A[g] = \frac{1}{n} \sum_{i=1}^n \text{AutoAIS}(x_i, g(x_i))$$

# AttributedQA (Bohnet et al., 2022)

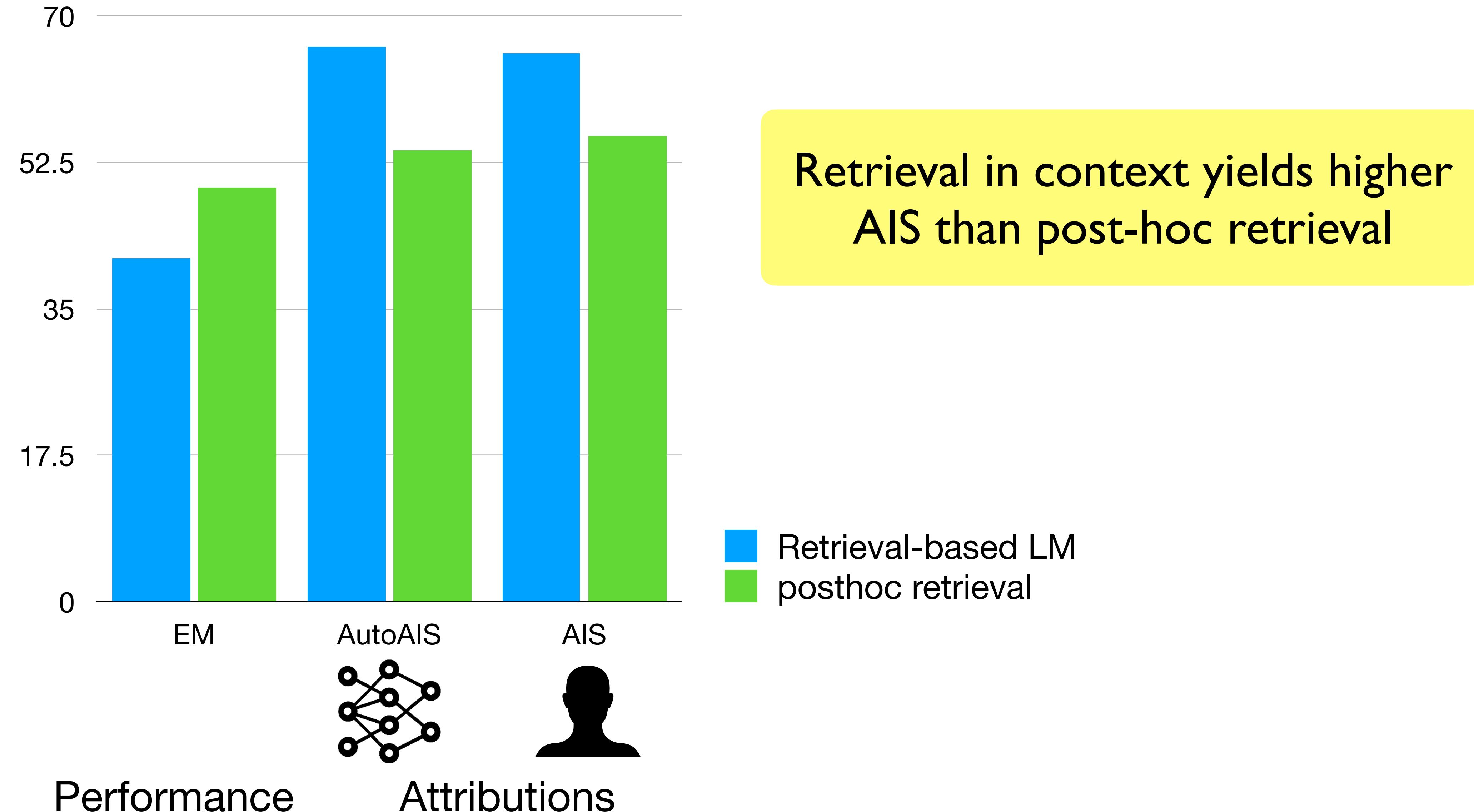
## Retrieval-based LMs



## Post-hoc retrieval



# AttributedQA (Bohnet et al., 2022)



# **When** to use a retrieval-based LM

Long-tail

knowledge  
update

Verifiability

Parameter-  
efficiency

Privacy

**Out of domain adaptations**

(Shi et al., 2022; Zheng et al., 2021)

and many others!!

Shi et al. 2022. “Nearest Neighbor Zero-shot Inference”

Zhang et al. 2021. “Non-Parametric Unsupervised Domain Adaptation for Neural Machine Translation”