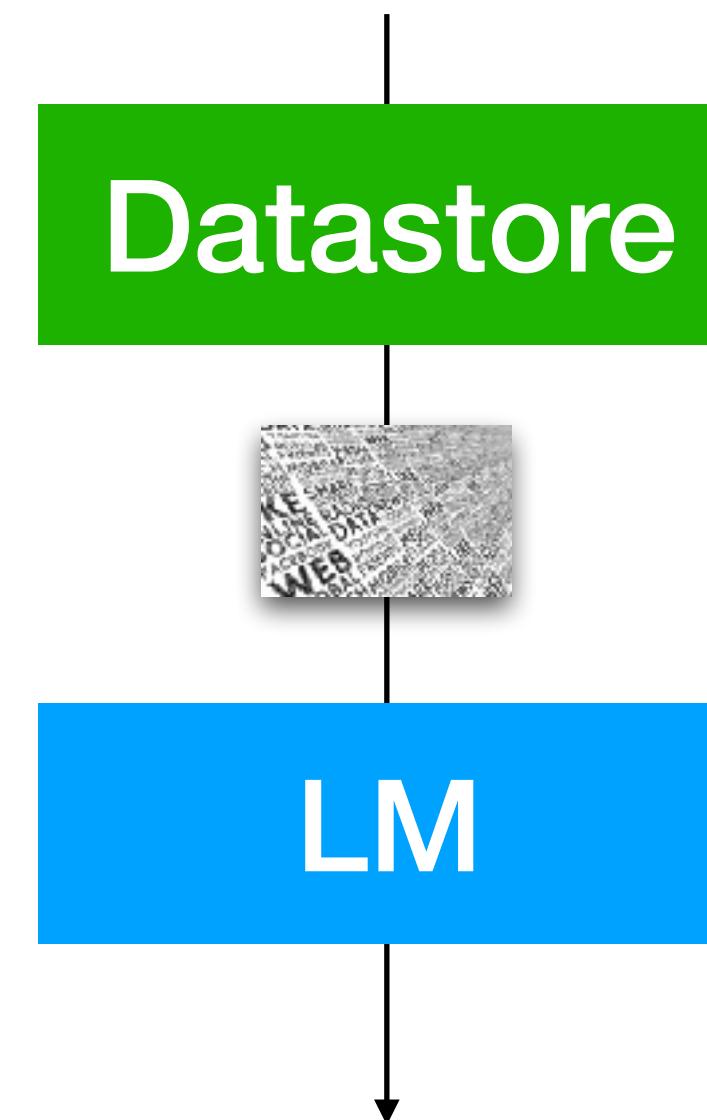


Section 5: Applications

Downstream adaptation of retrieval-based LMs

The capital city of Ontario is __



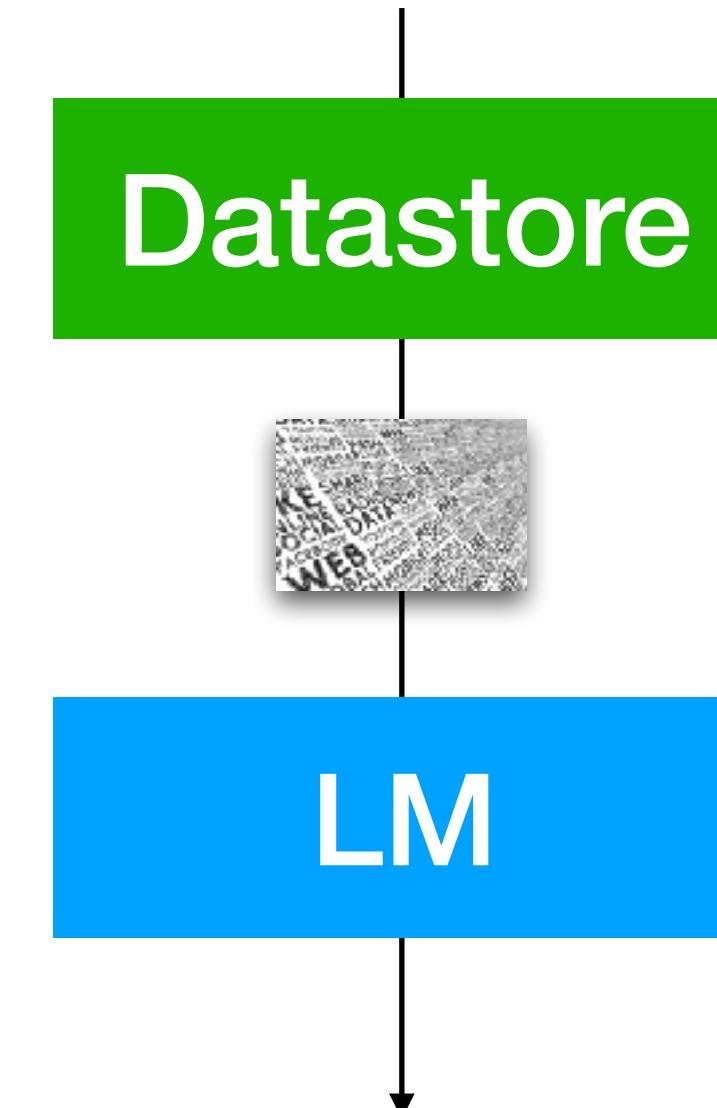
Toronto, which is known for ...

Downstream adaptation of retrieval-based LMs

What are the **tasks**?

Open-domain QA

What is the capital of Ontario?



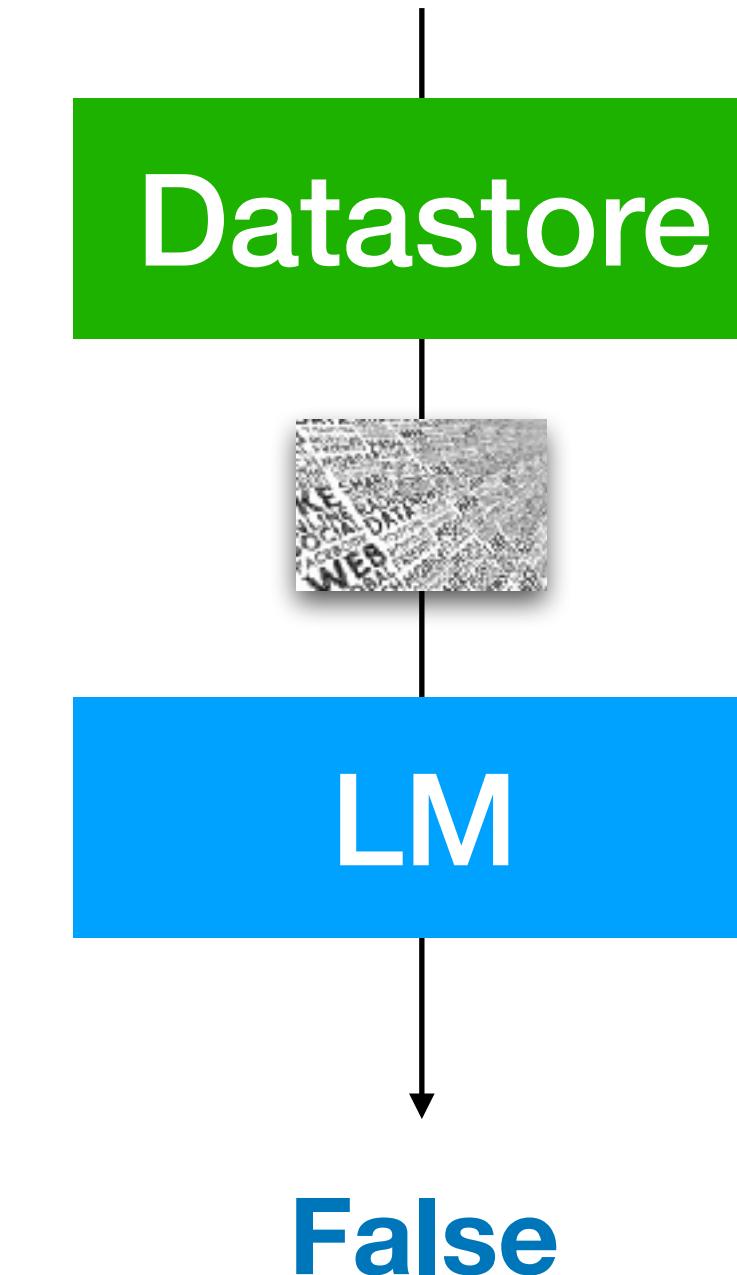
Toronto

Downstream adaptation of retrieval-based LMs

What are the **tasks**?

Fact verification

Ottawa is the Ontario state capital.



A range of target tasks

Question Answering

RETRO (Borgeaud et al., 2021)
REALM (Gu et al, 2020)
ATLAS (Izacard et al, 2023)

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)

Retrieval-based LMs have been extensively evaluated on knowledge-intensive tasks

A range of target tasks

Question Answering

RETRO (Borgeaud et al., 2021)
REALM (Gu et al, 2020)
ATLAS (Izacard et al, 2023)

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)

Code & proof generation

DocPrompting (Zhou et al., 2023)
Natural Prover
(Welleck et al., 2022)

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 NLP tasks

A range of target tasks

Question Answering

RETRO (Borgeaud et al., 2021)
REALM (Gu et al, 2020)
ATLAS (Izacard et al, 2023)

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)

Code & proof generation

DocPrompting (Zhou et al., 2023)
Natural Prover
(Welleck et al., 2022)

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

RETRO (Borgeaud et al., 2021)
REALM (Gu et al, 2020)
ATLAS (Izacard et al, 2023)

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)

Code & proof generation

DocPrompting (Zhou et al., 2023)
Natural Prover
(Welleck et al., 2022)

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

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

...

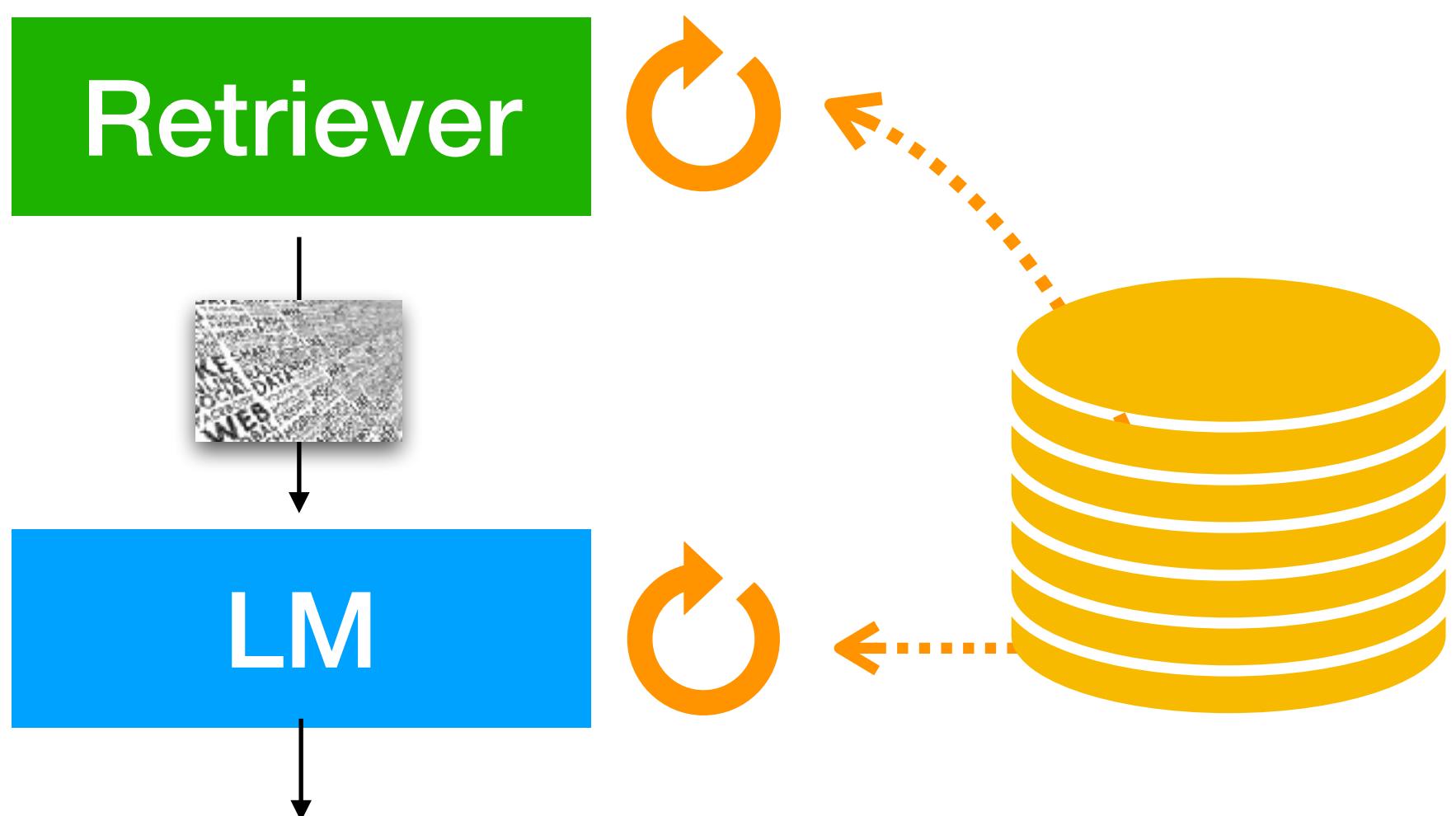
How to **adapt**?

- Fine-tuning
- Reinforcement learning
- Prompting

How to adapt a retrieval-based LM for a task

Fine-tuning (+RL)

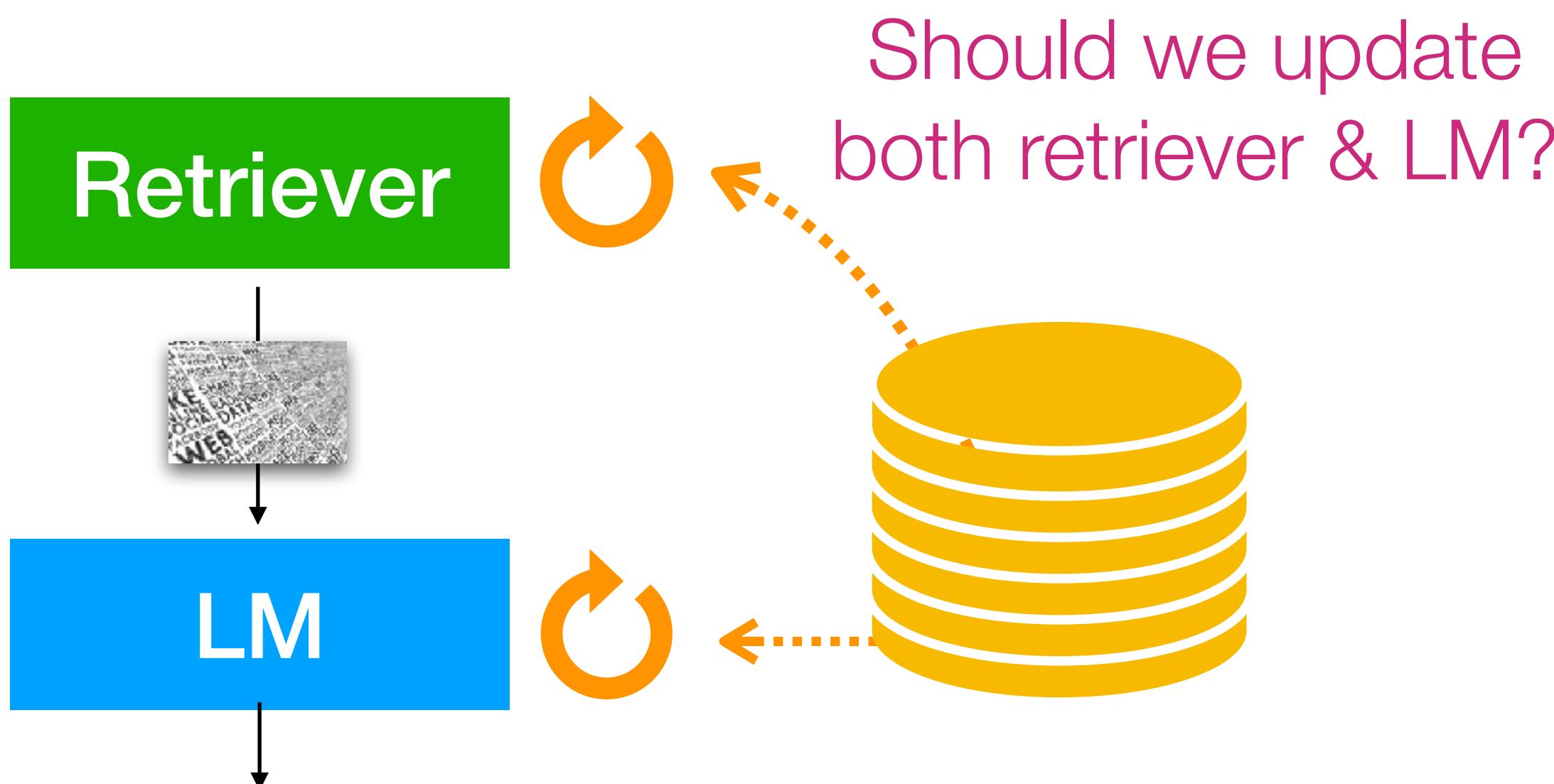
Training LM and / or retriever
on task-data & data store



How to adapt a retrieval-based LM for a task

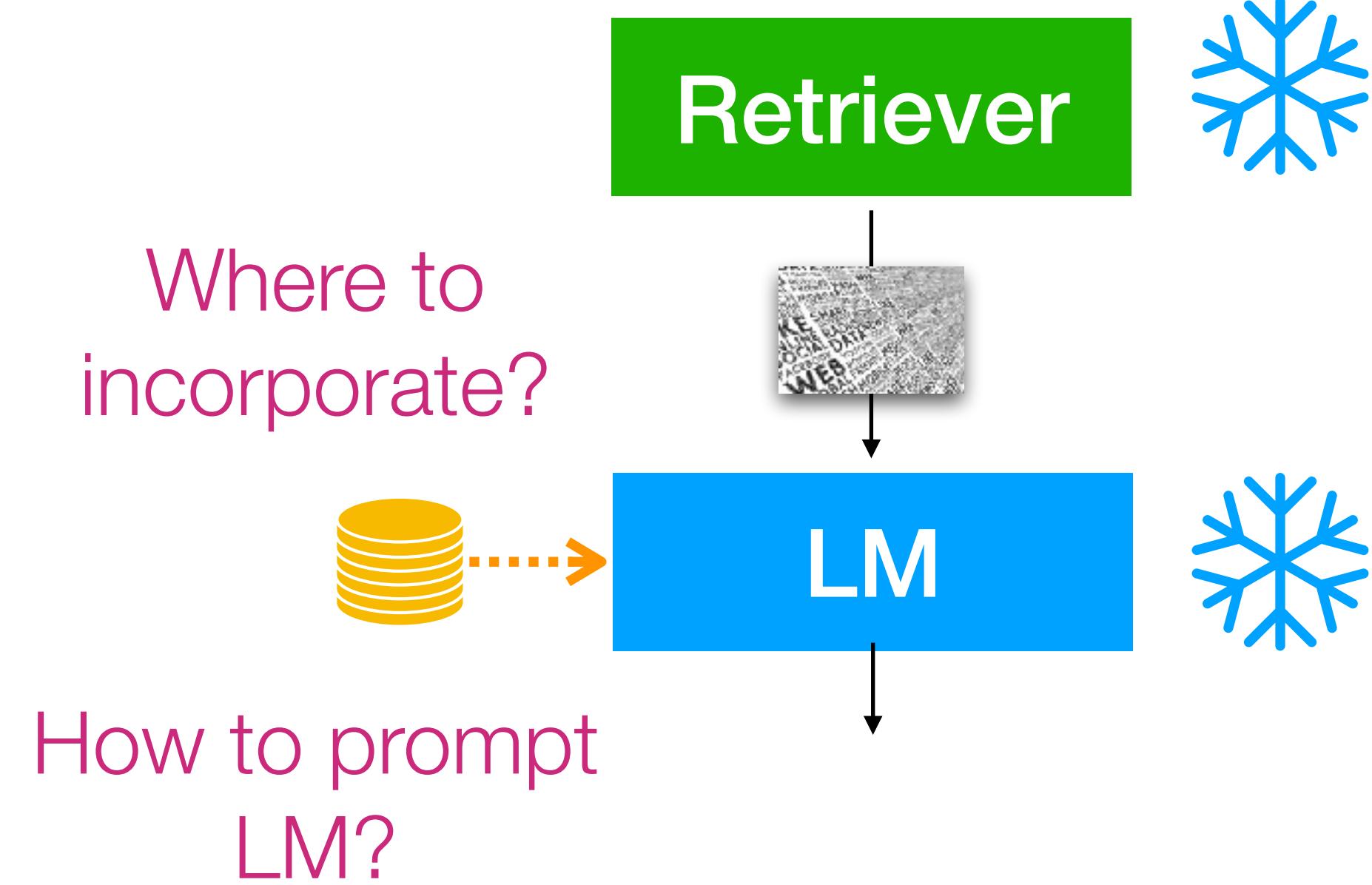
Fine-tuning (+RL)

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
- ...

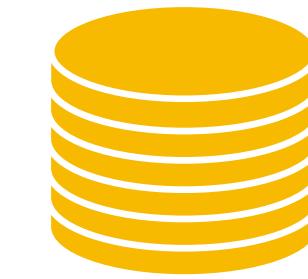
How to **adapt**?

- 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

Effectiveness of retrieval-based LMs

Long-tail

knowledge
update

Verifiability

Parameter-
efficiency

Q: Is Toronto really
cold during winter?



Effectiveness of retrieval-based LMs

Long-tail

knowledge
update

Verifiability

Parameter-
efficiency

Q: Where is Toronto
Zoo located?



1361A Old Finch Avenue,
in Scarborough, Ontario



Effectiveness of retrieval-based LMs

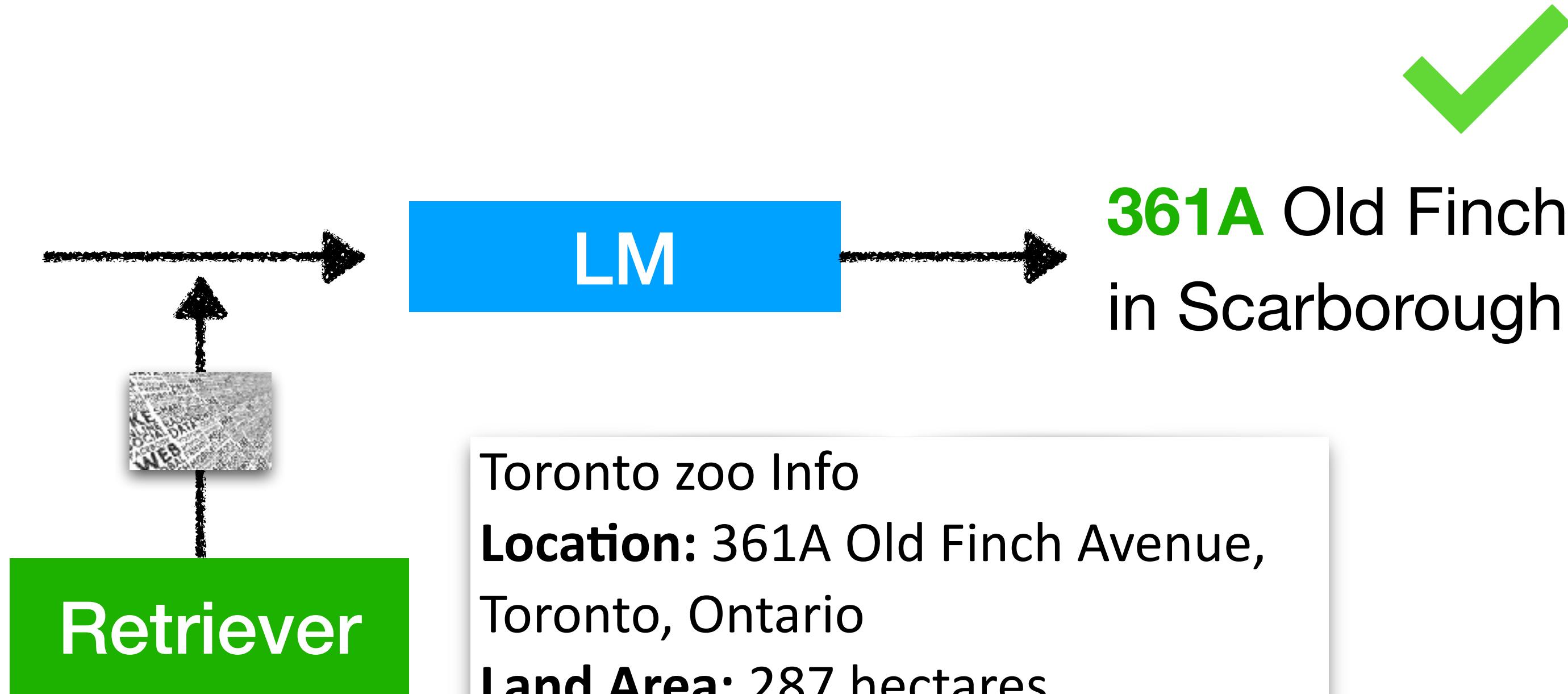
Long-tail

knowledge update

Verifiability

Parameter-efficiency

Q: Where is Toronto Zoo located?



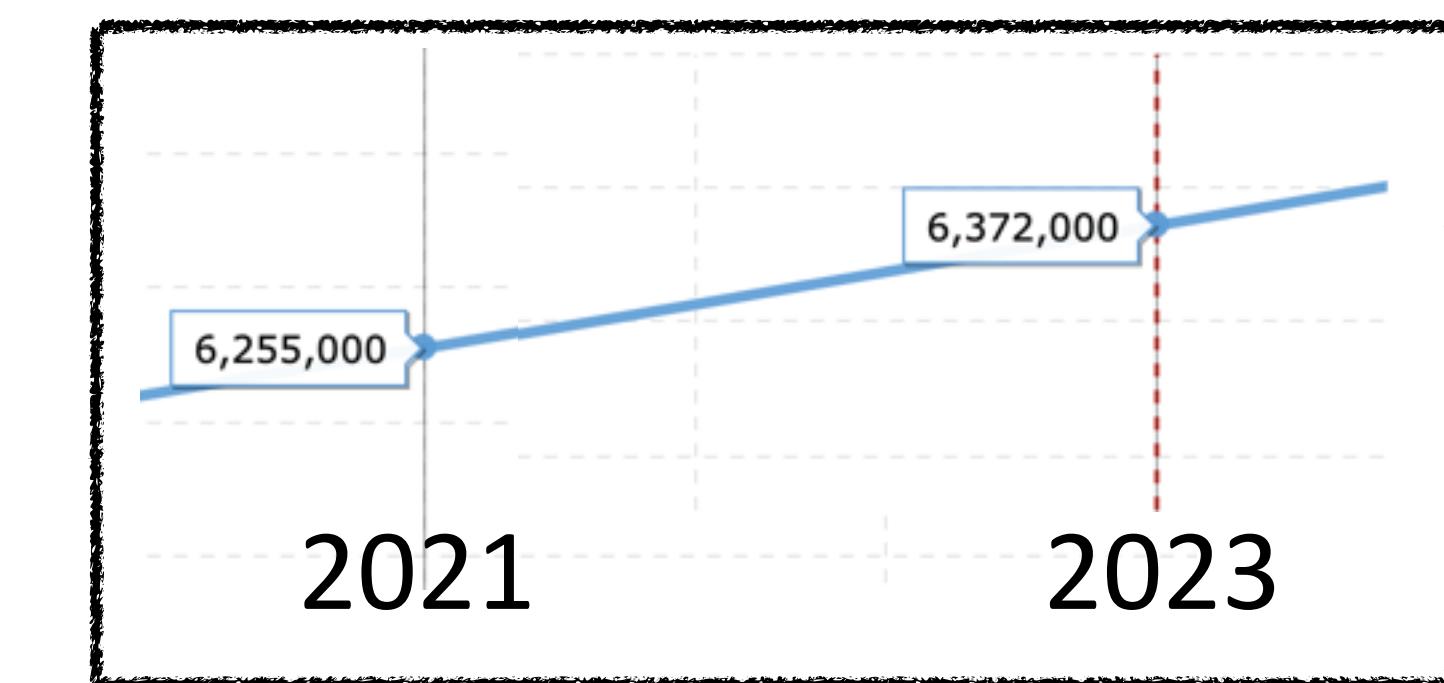
Effectiveness of retrieval-based LMs



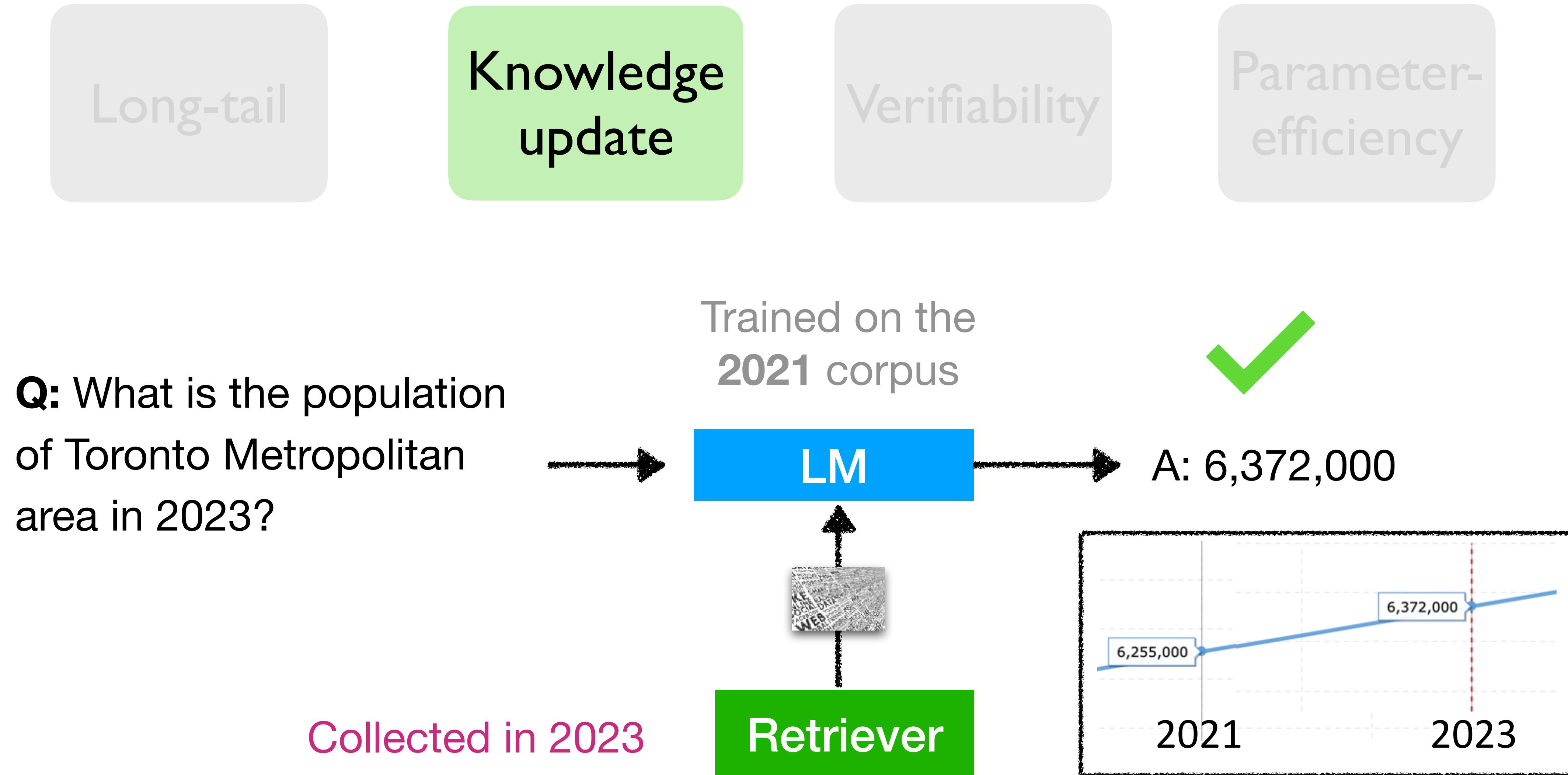
Q: What is the population
of Toronto Metropolitan
area in 2023?



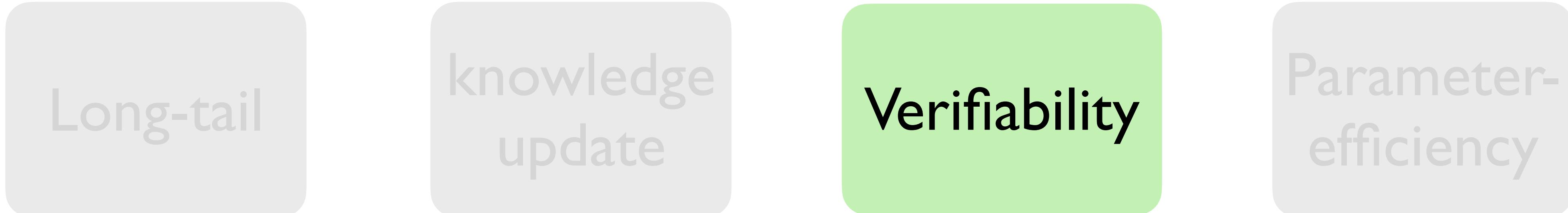
Trained on the
2021 corpus



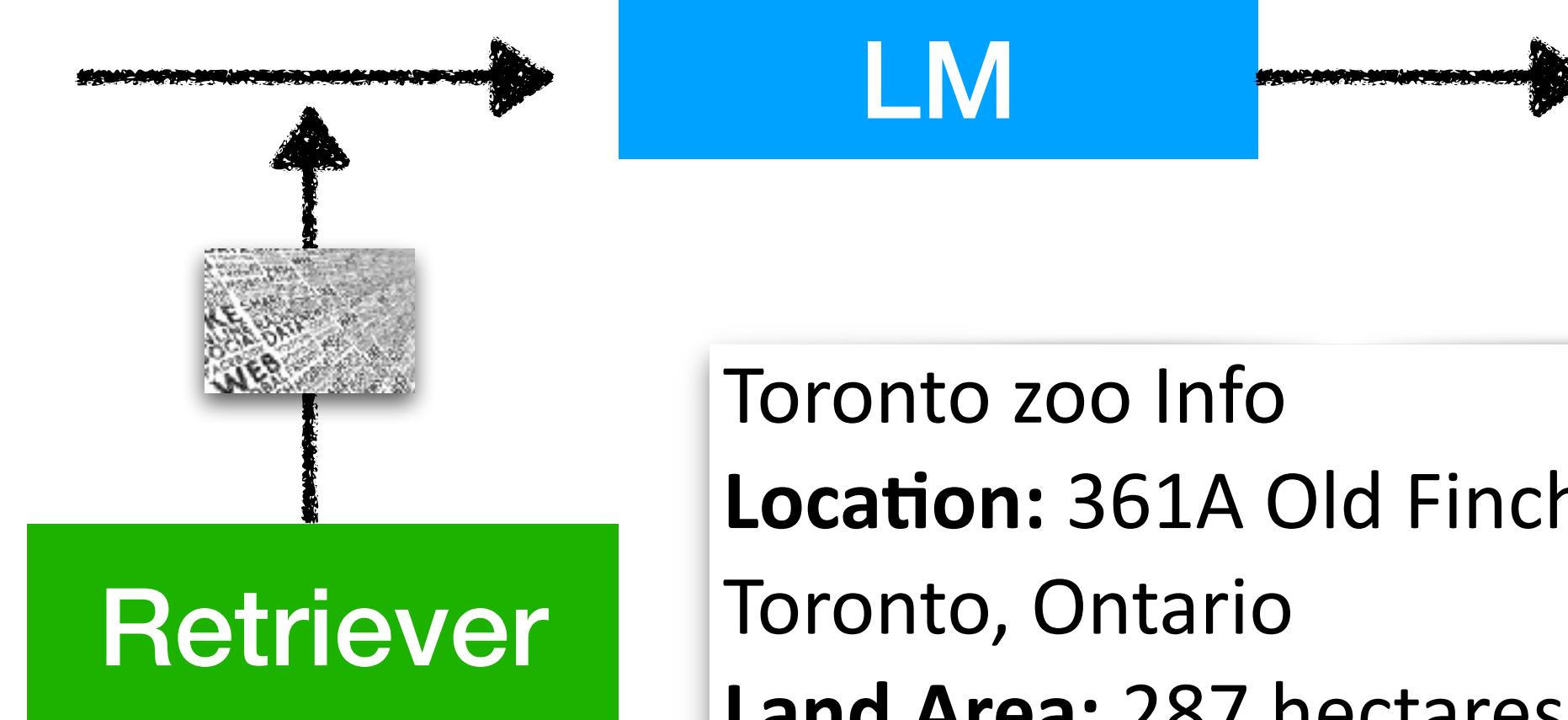
Effectiveness of retrieval-based LMs



Effectiveness of retrieval-based LMs



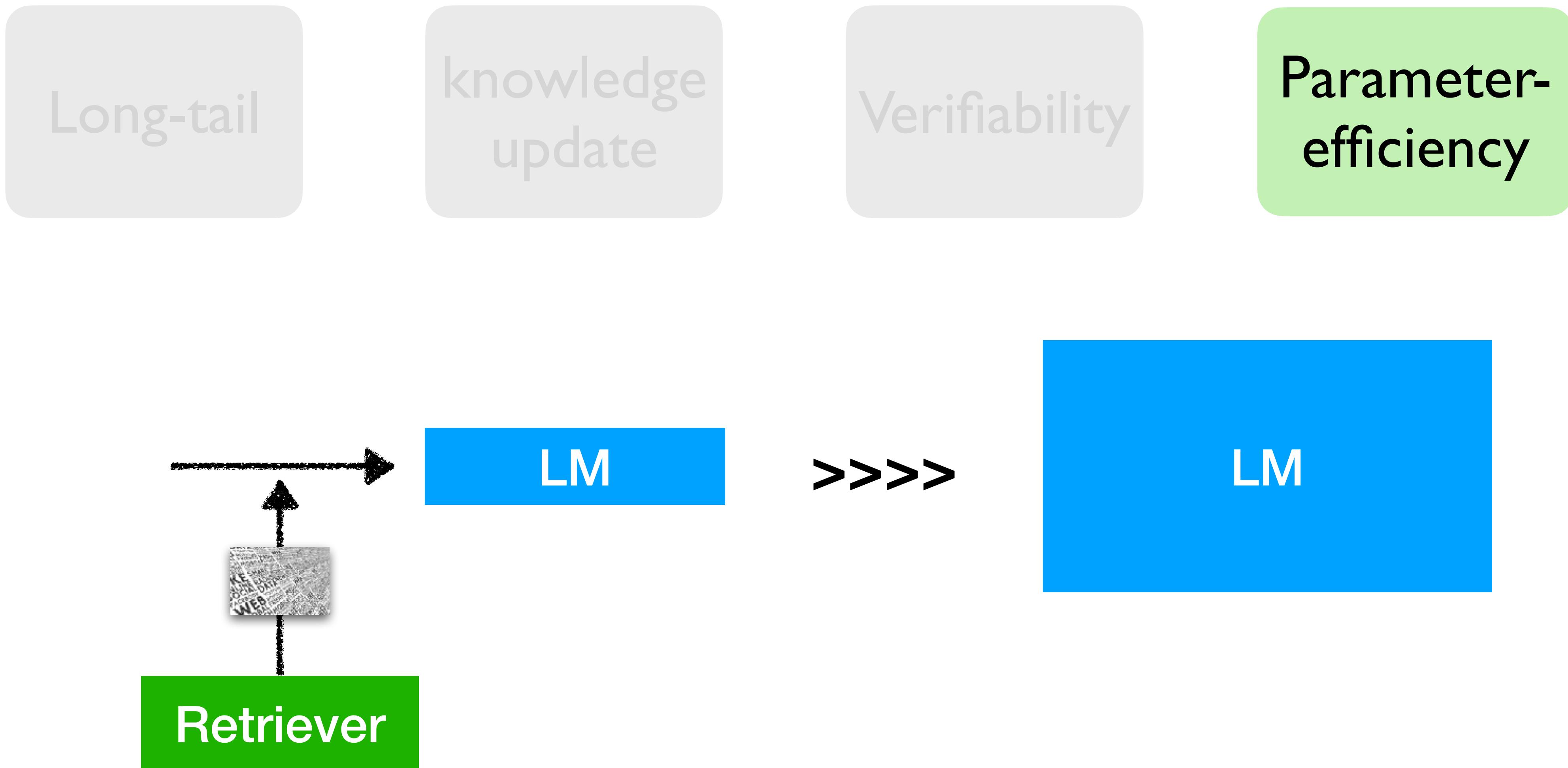
Q: Where is Toronto Zoo located?



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



Effectiveness of retrieval-based LMs



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

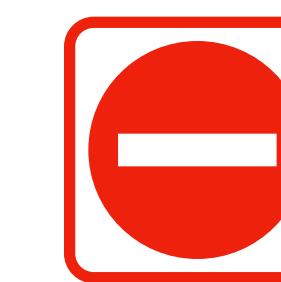
Training LM and / or retriever
on task-data & data store



Adapting retrieval-based LMs for tasks

Fine-tuning

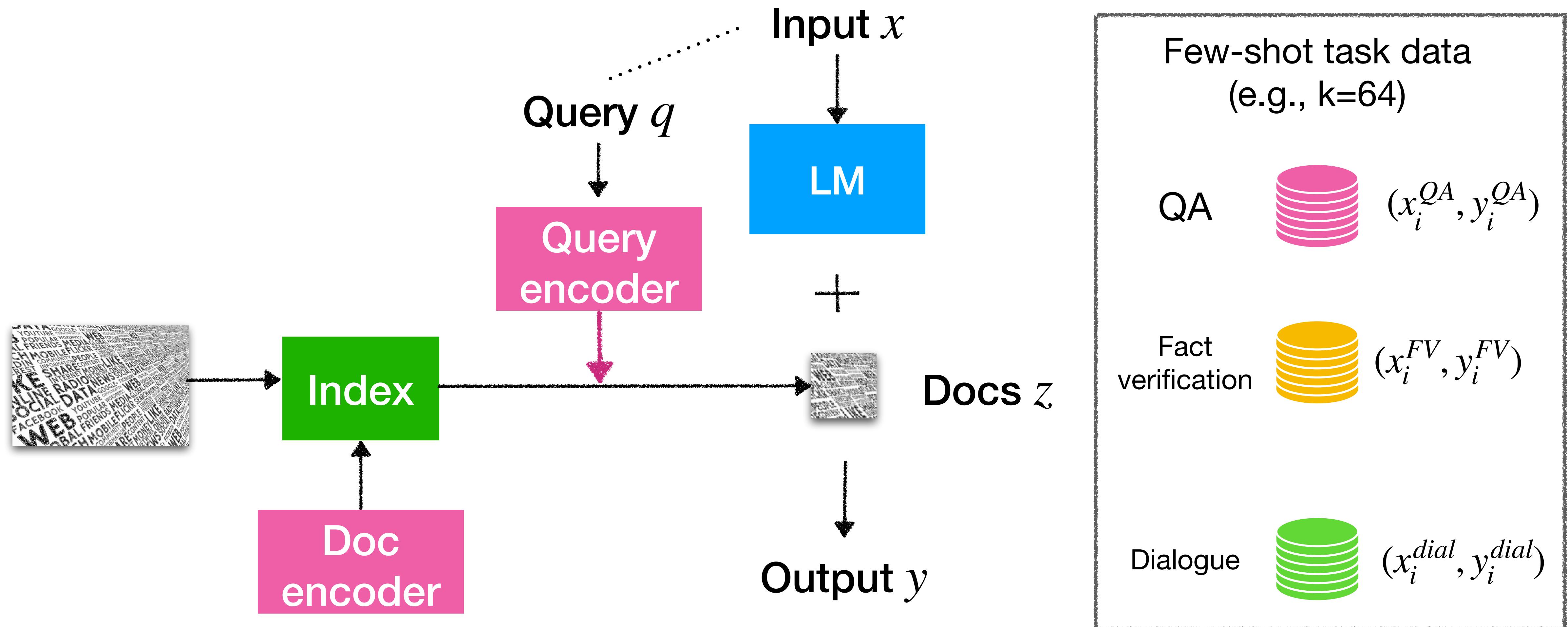
Training LM and / or retriever
on task-data & data store



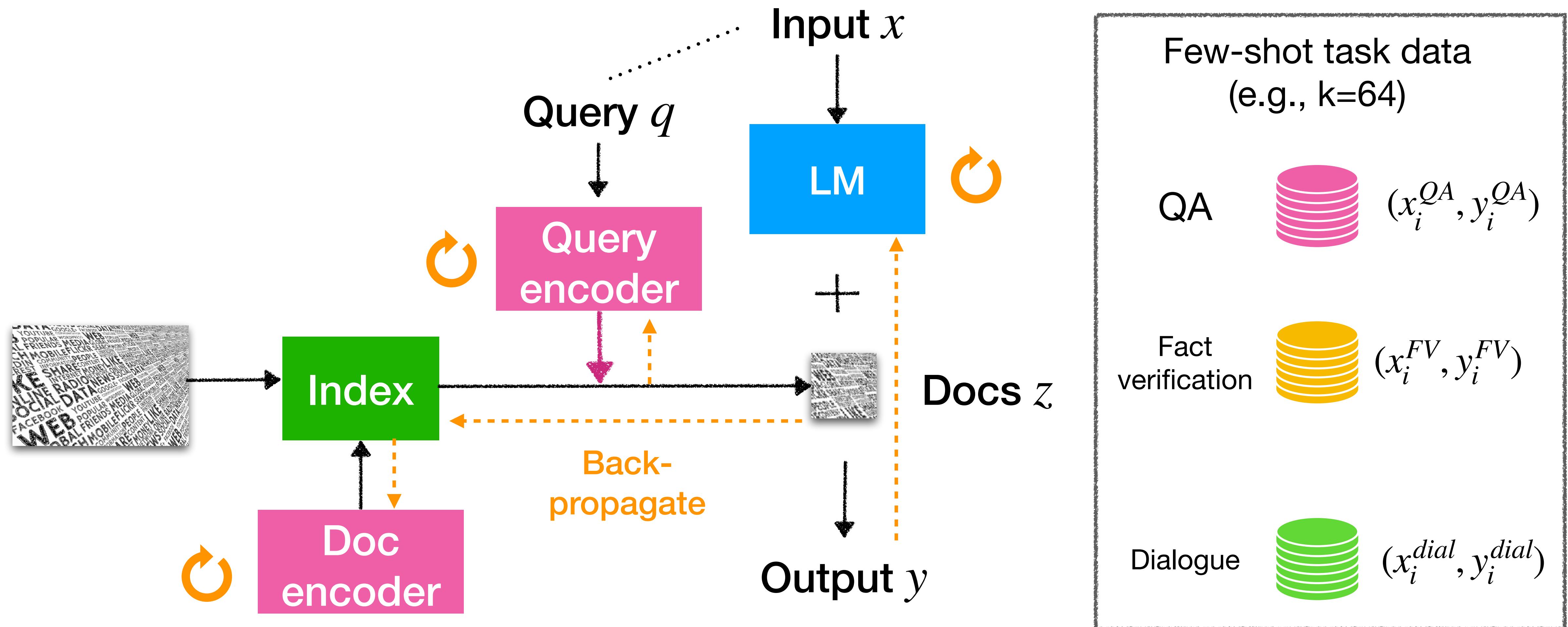
Costs of retrieval-based LM
training (Section 4)

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

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

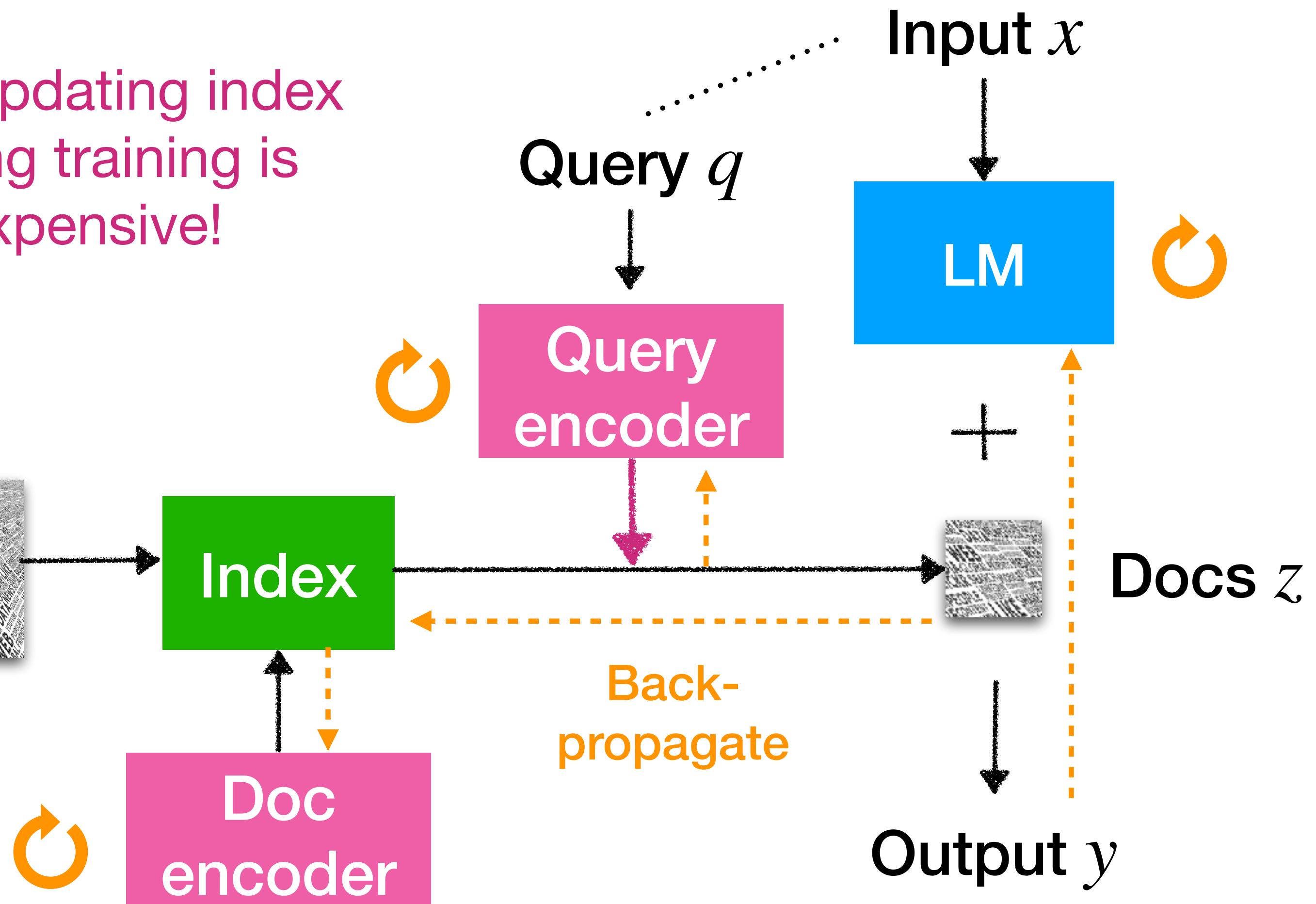


ATLAS (Izacard et al., 2022; Section 4)



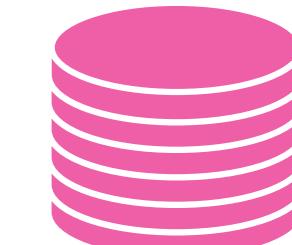
ATLAS (Izacard et al., 2022; Section 4)

Fully updating index during training is expensive!



Few-shot task data
(e.g., $k=64$)

QA



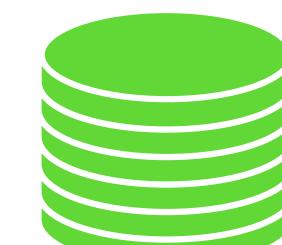
(x_i^{QA}, y_i^{QA})

Fact verification



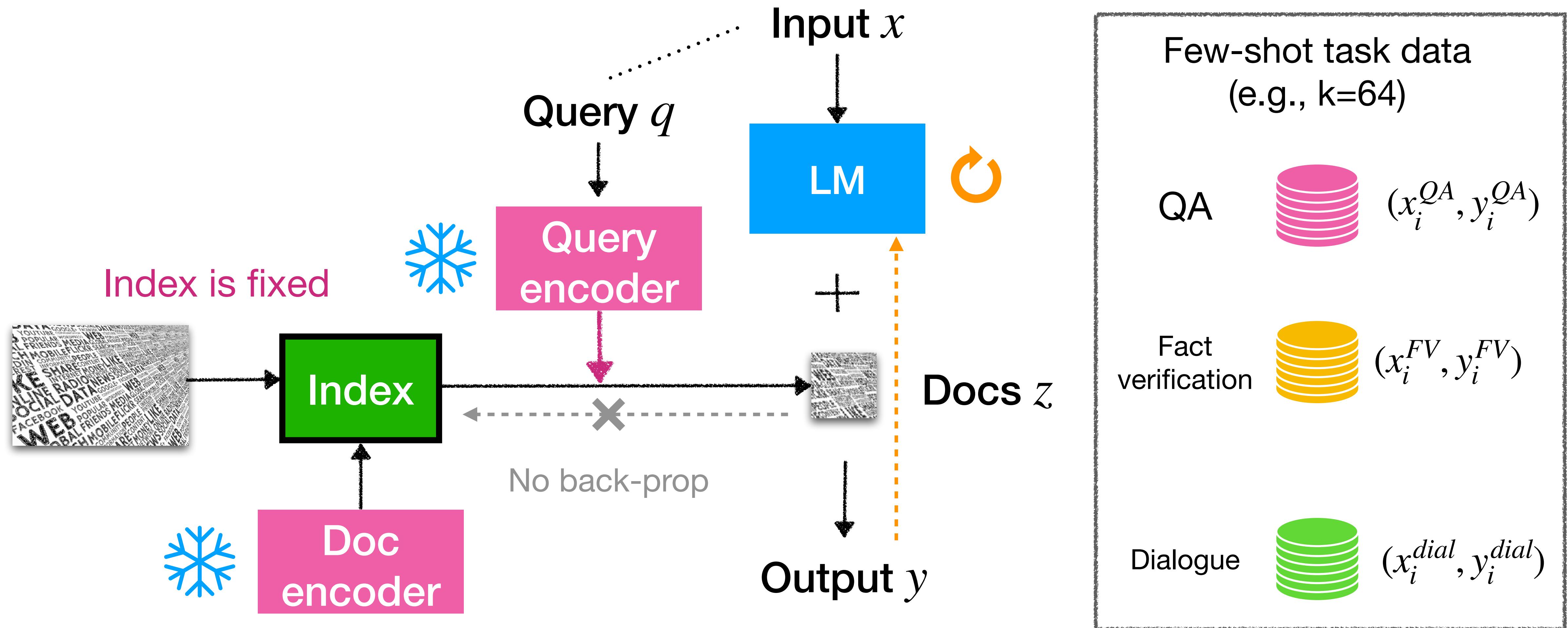
(x_i^{FV}, y_i^{FV})

Dialogue

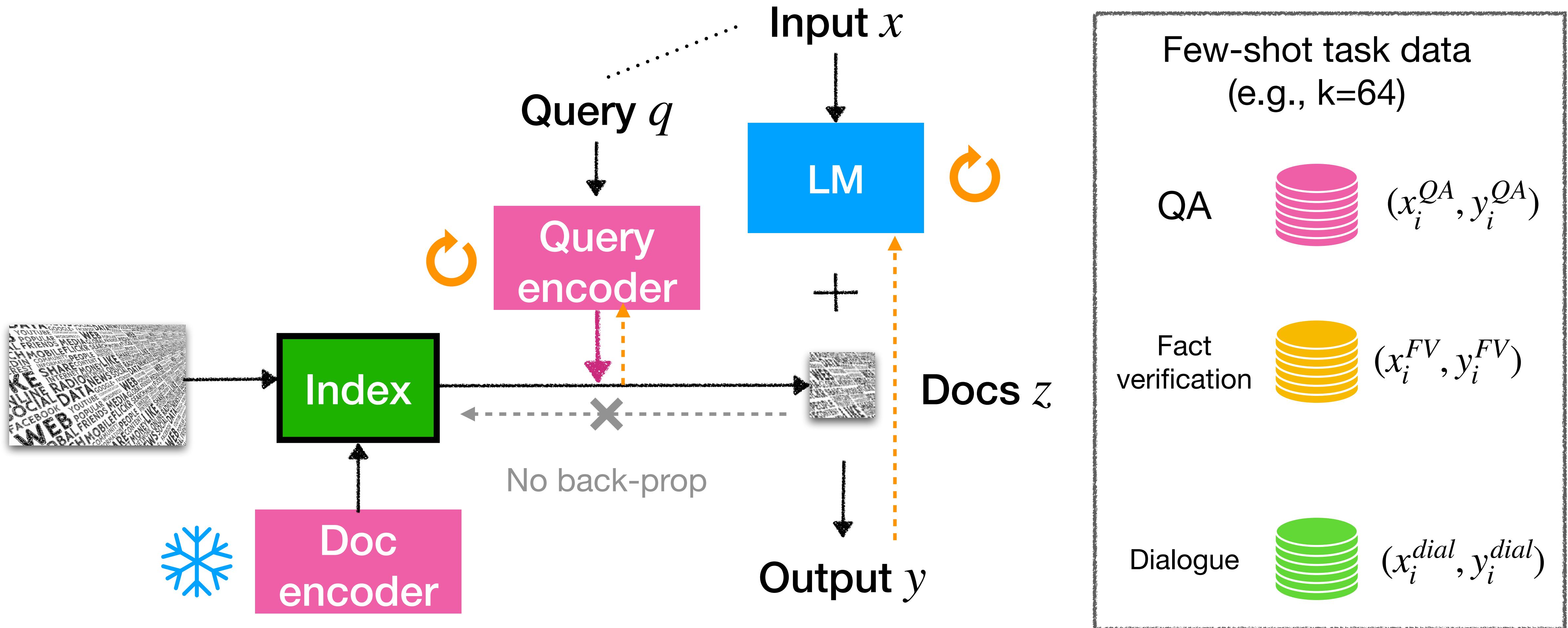


(x_i^{dial}, y_i^{dial})

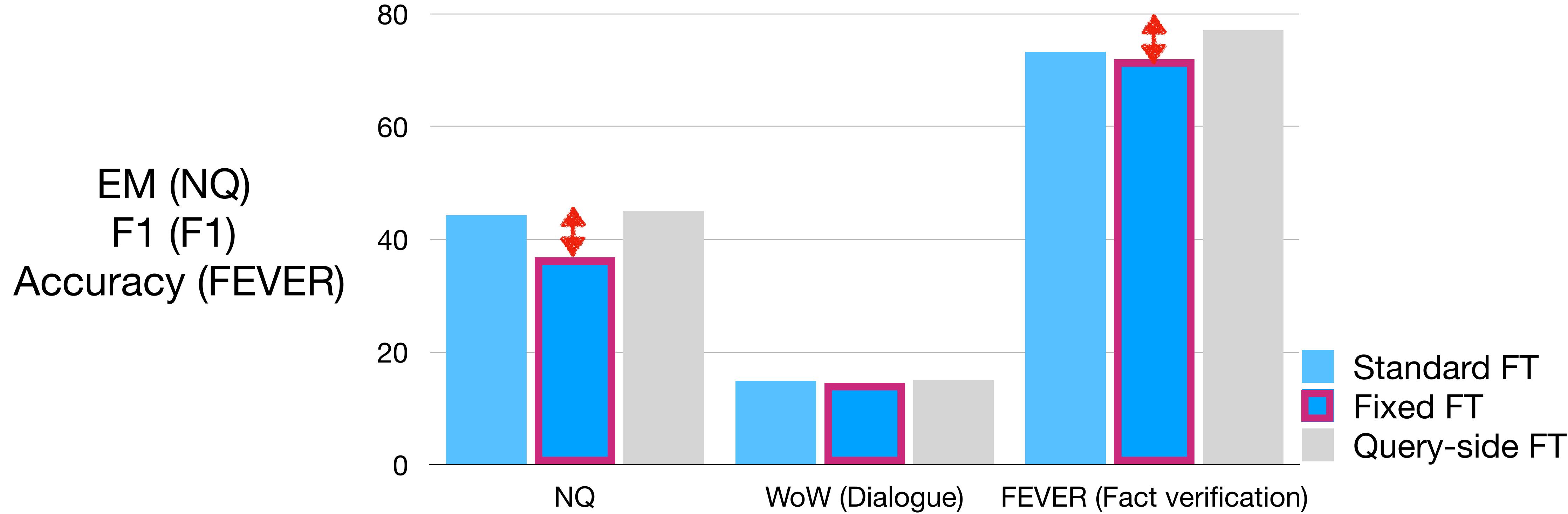
ATLAS: Fixed retrieval with fine-tuned LM



ATLAS: Query-side fine-tuning

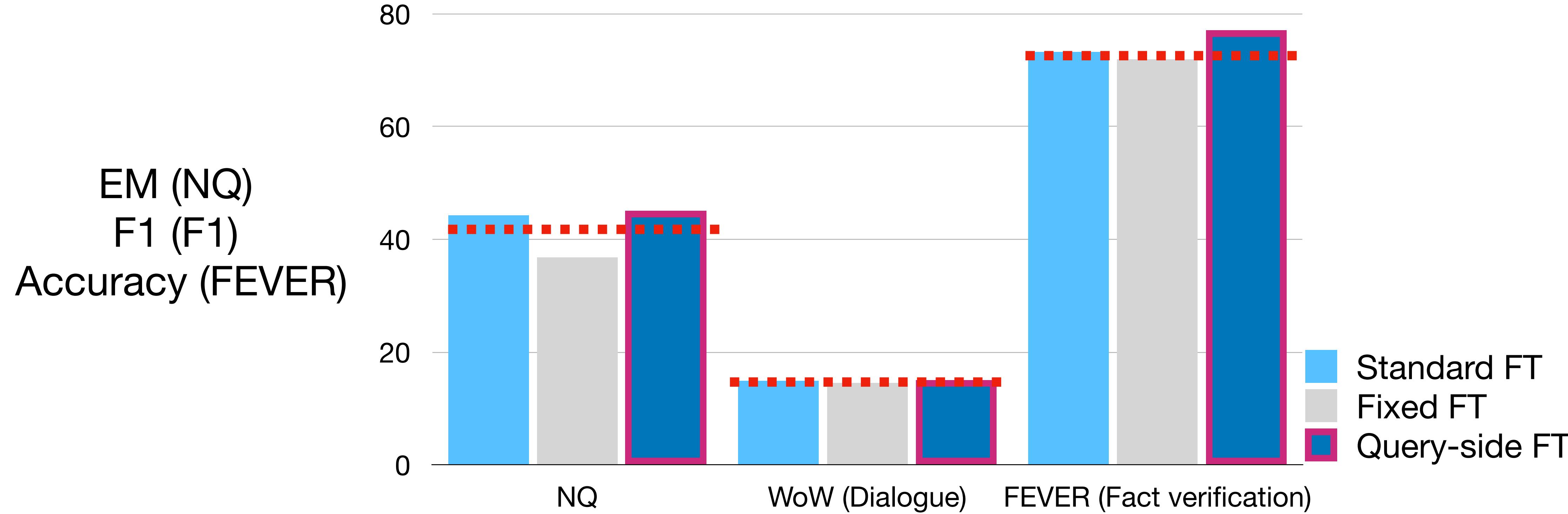


Ablations of efficient retrieval training



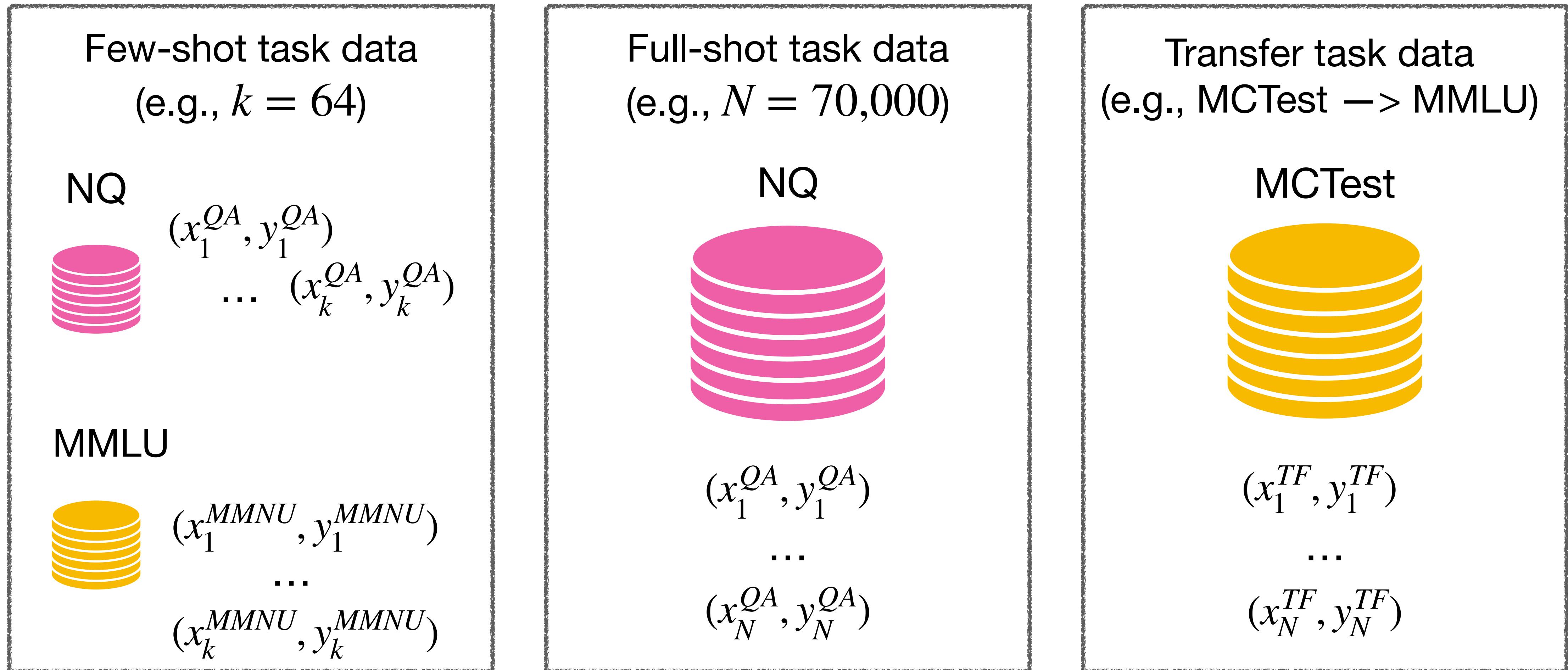
Fixed FT shows large performance drop on QA.

Ablations of efficient retrieval training



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

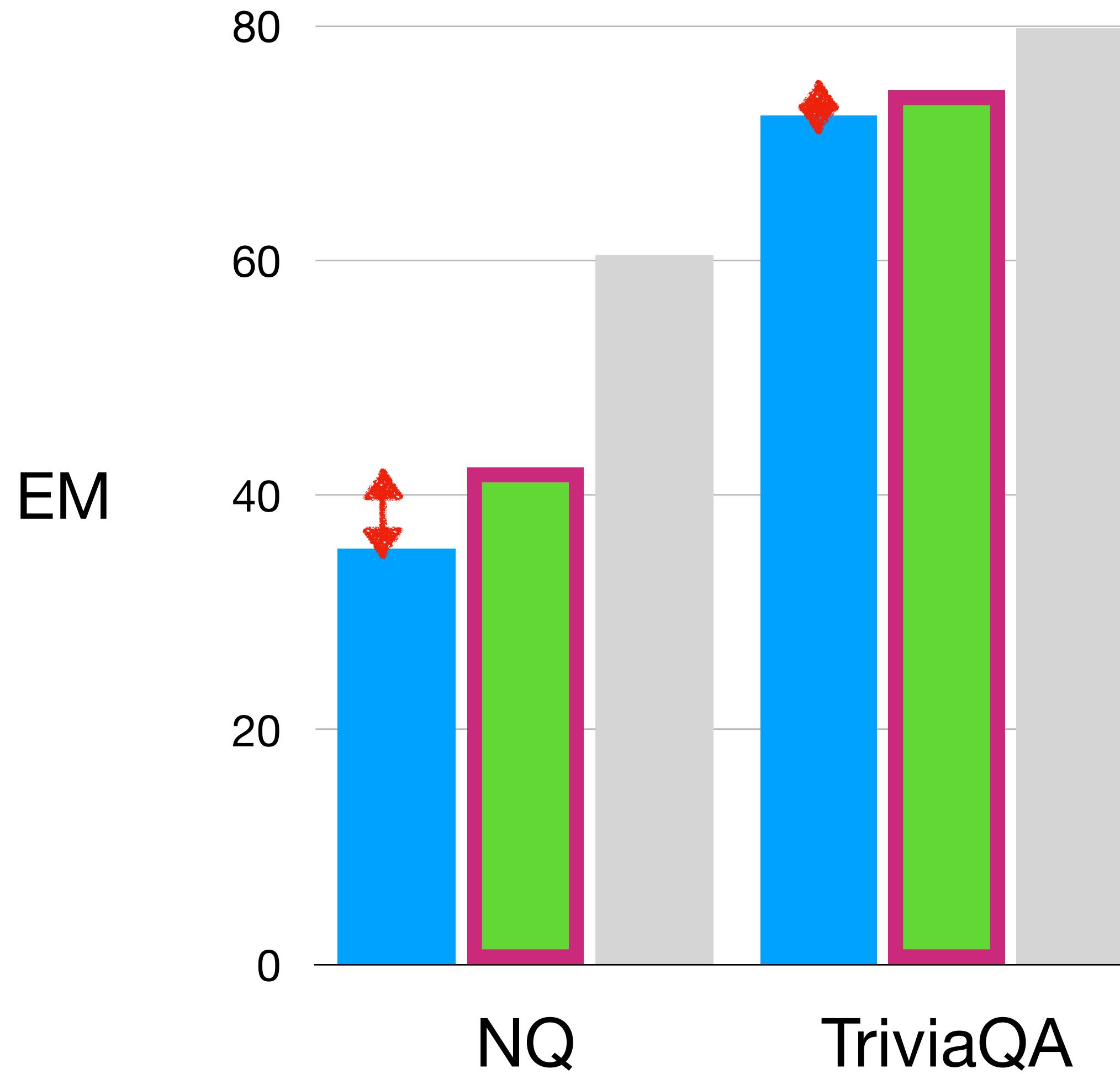
ATLAS: Few-shot v.s. full v.s. transfer setups



Kwiatkowski et al. 2019. “Natural Questions: A Benchmark for Question Answering Research”

Hendrycks et al. 2021. “Measuring Massive Multitask Language Understanding”

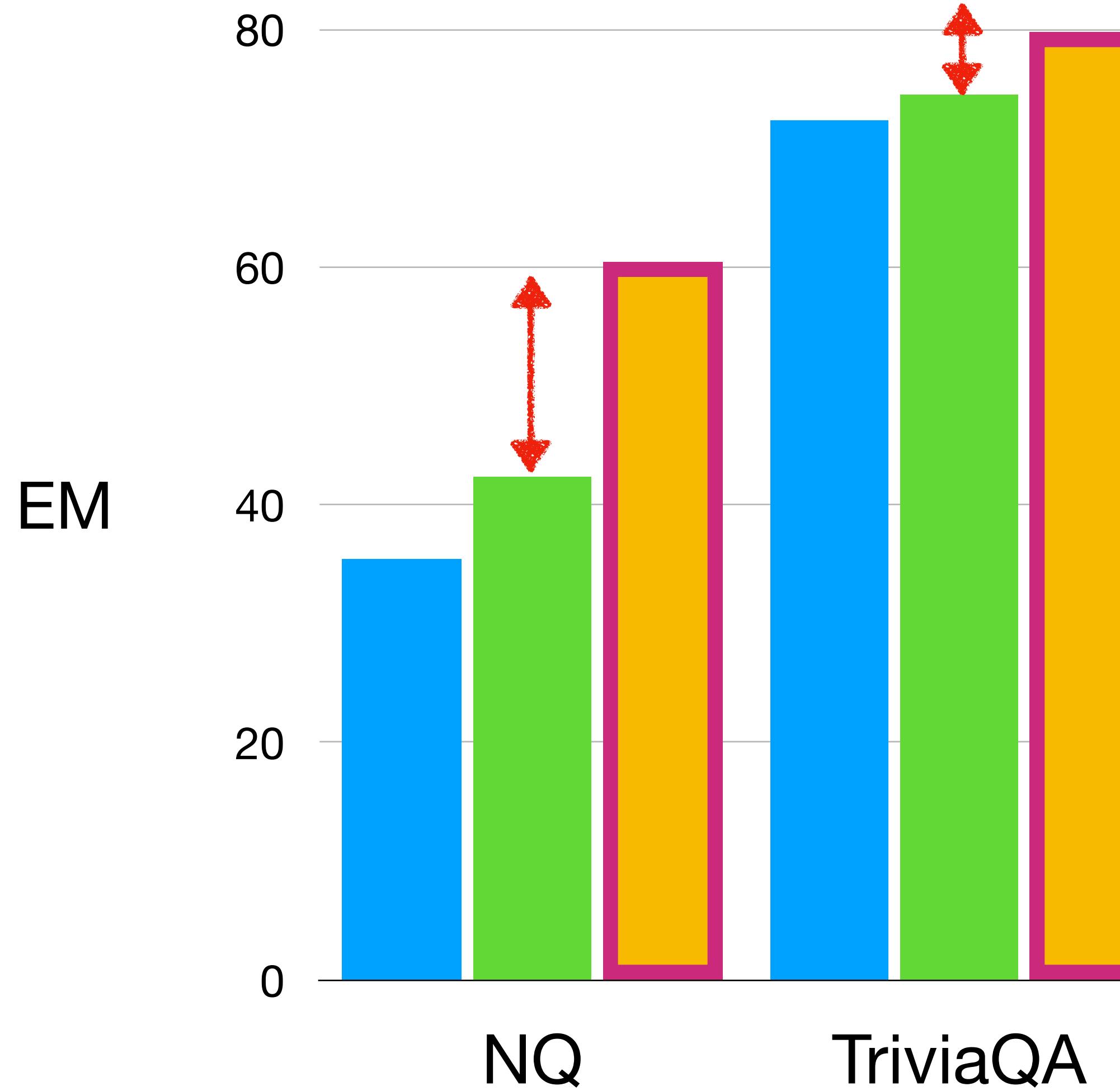
Task results



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

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

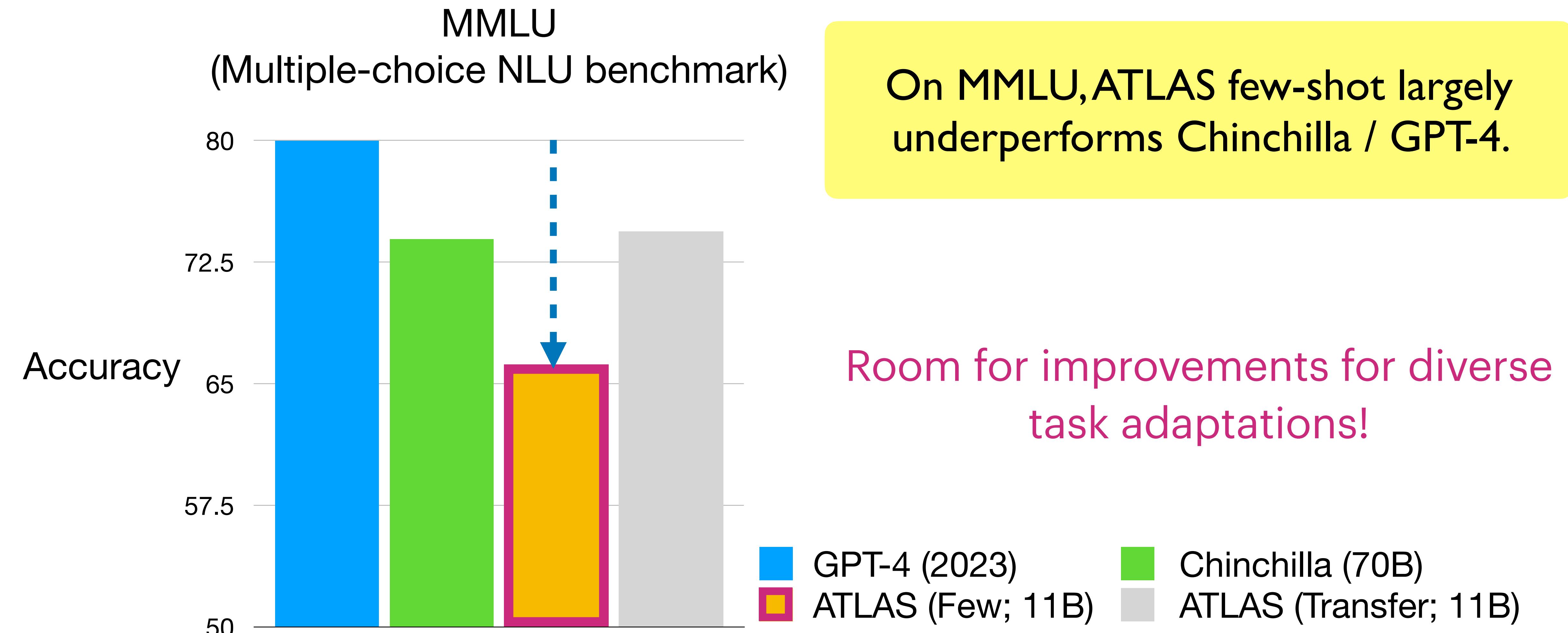
Task results



Full-shot fine-tuning further improves performance

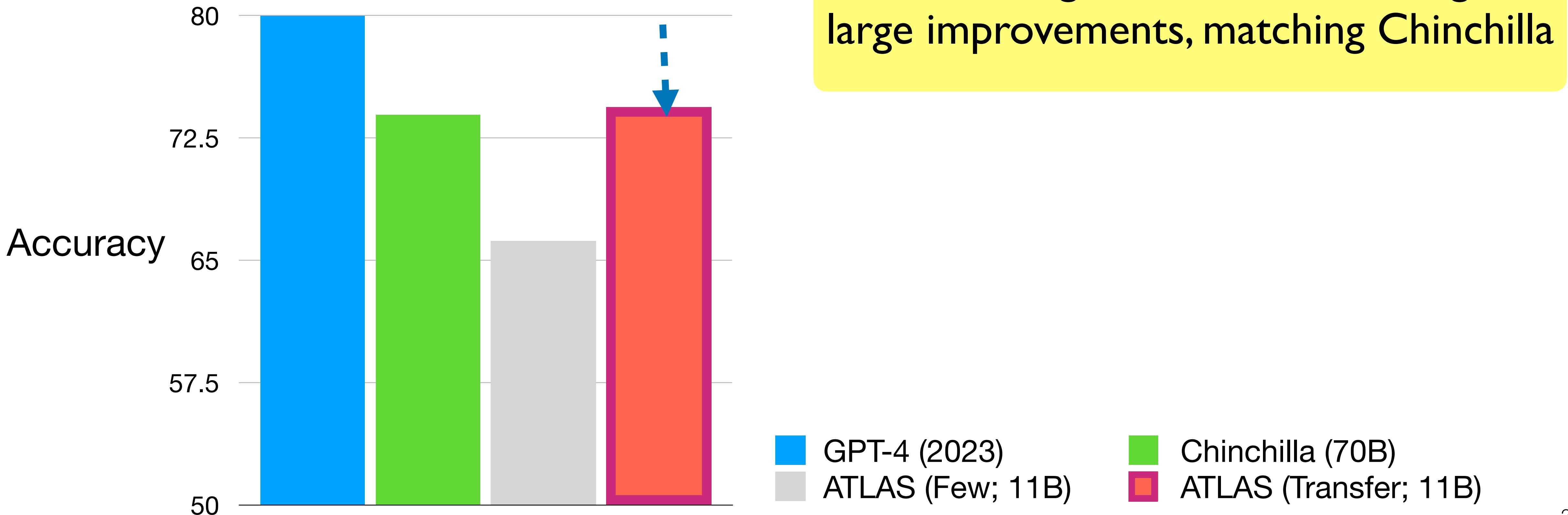
- Chinchilla (70B)
- ATLAS (Few; 11B)
- ATLAS (Full; 11B)

Task results



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 (Retriever & 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 (Retriever & 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)

User

What kind of animal is Scooby from Scooby Doo?

GopherCite

A Great Dane dog.

GopherCite (Menick et al., 2022)

User

What kind of animal is Scooby from Scooby Doo?

GopherCite

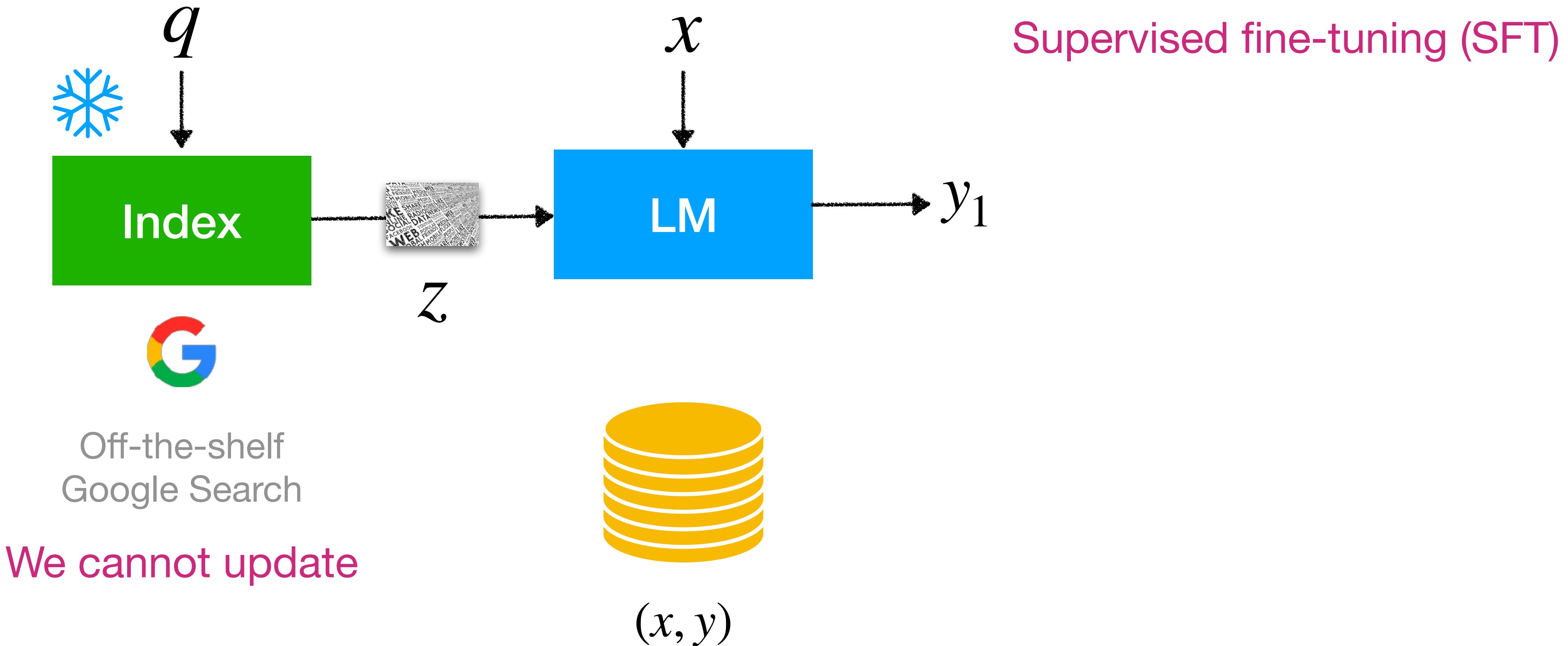
A Great Dane dog.

Wikipedia Page: Scooby-Doo

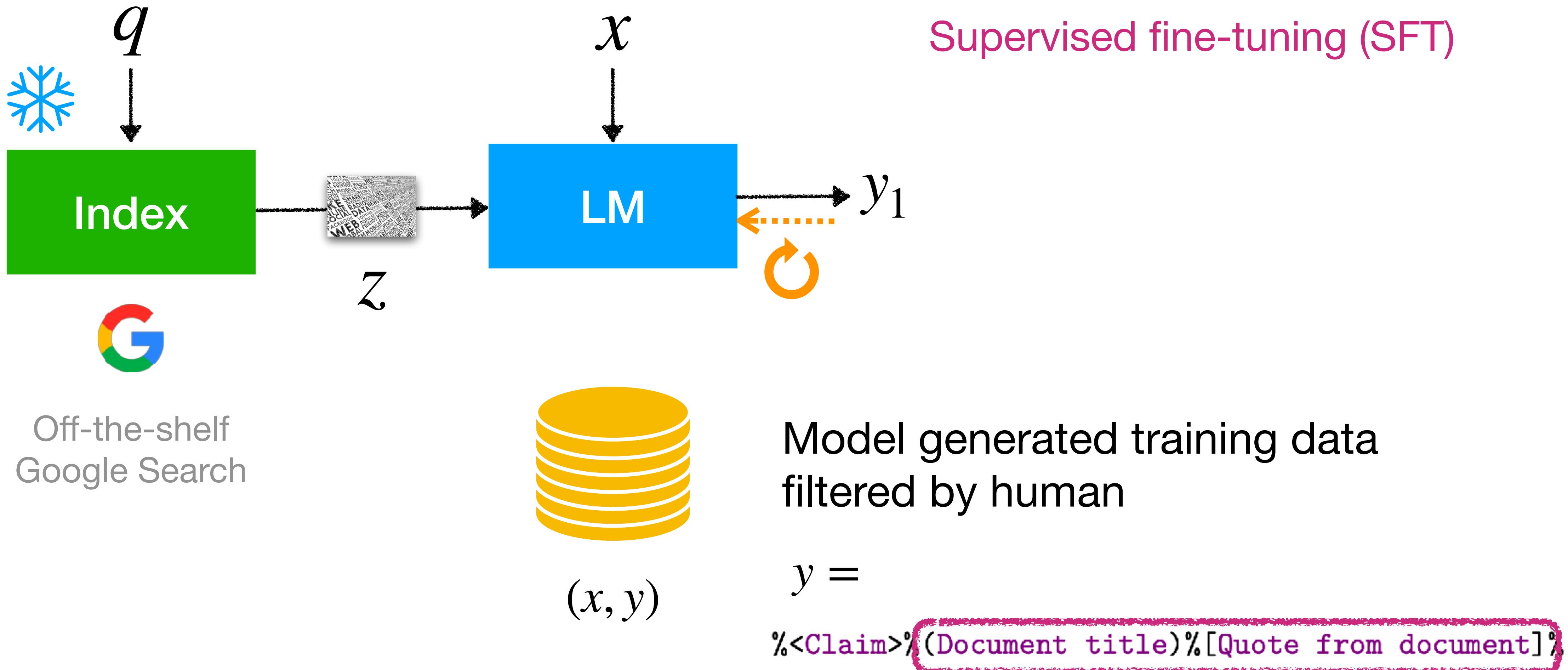
This Saturday-morning cartoon series featured teenagers Fred Jones, Daphne Blake, Velma Dinkley, and Shaggy Rogers, and their talking Great Dane named Scooby-Doo.

Extract and generate a quote to support an answer

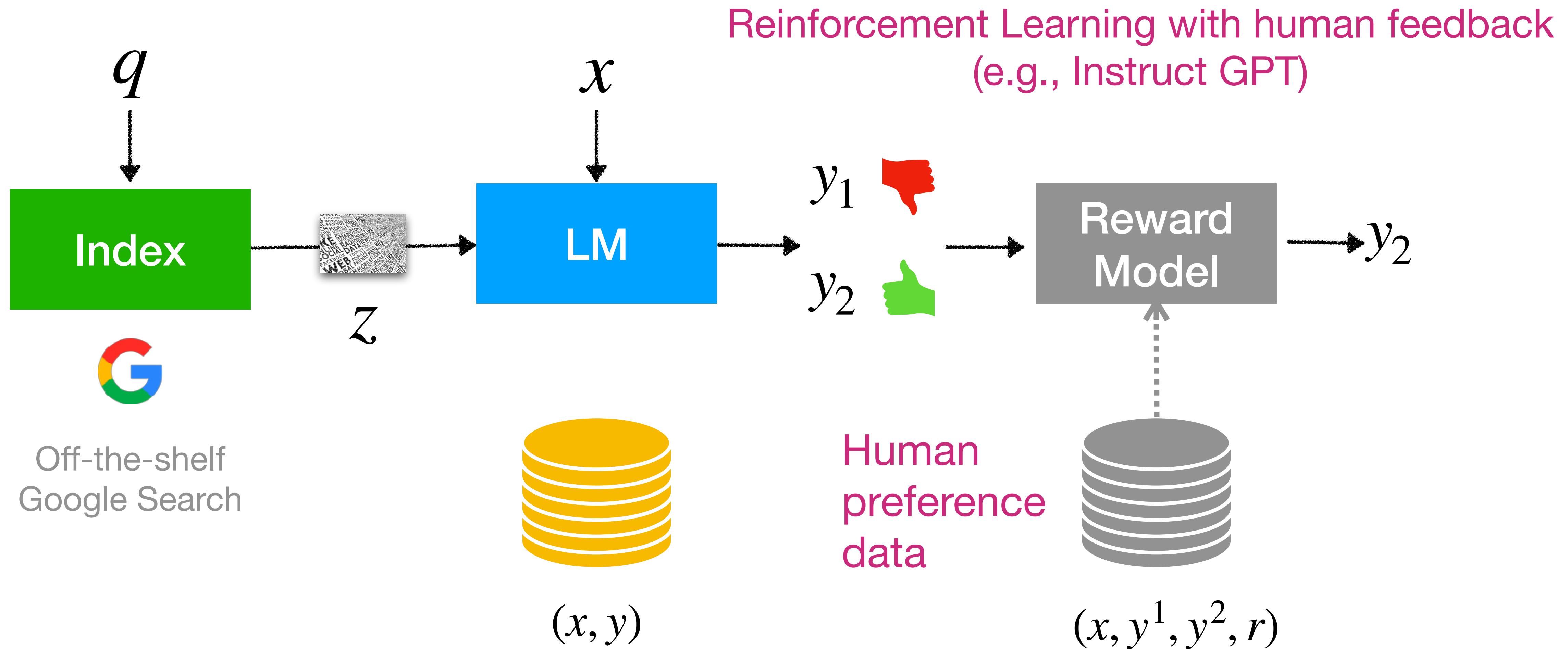
GopherCite: RLHF for answering with verified quotes



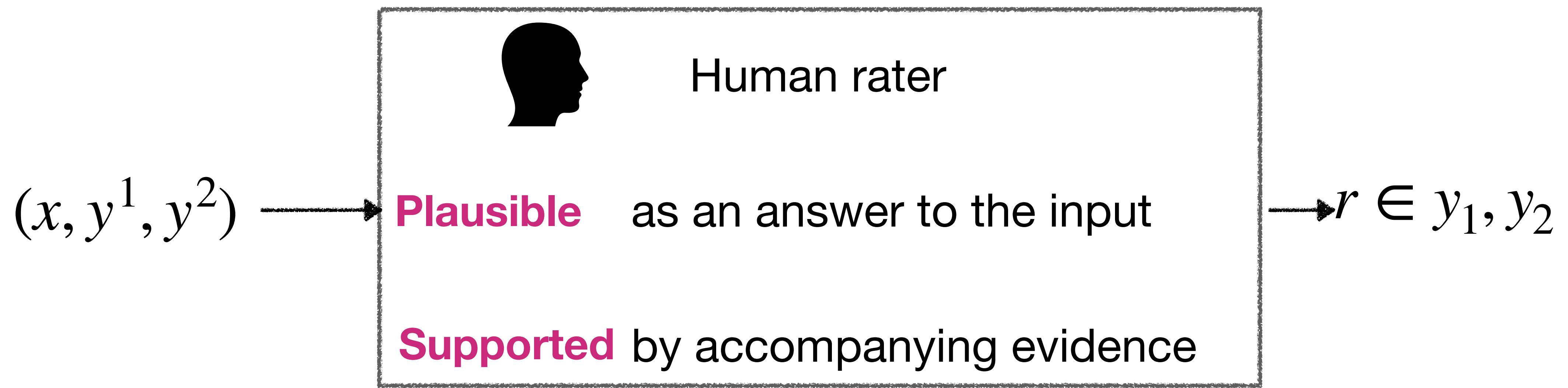
GopherCite: RLHF for answering with verified quotes



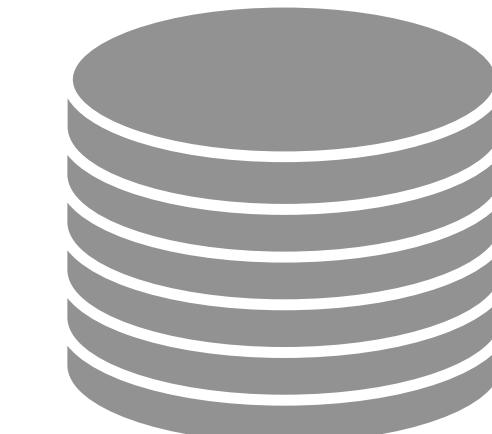
GopherCite: RLHF for answering with verified quotes



GopherCite: RLHF for answering with verified quotes



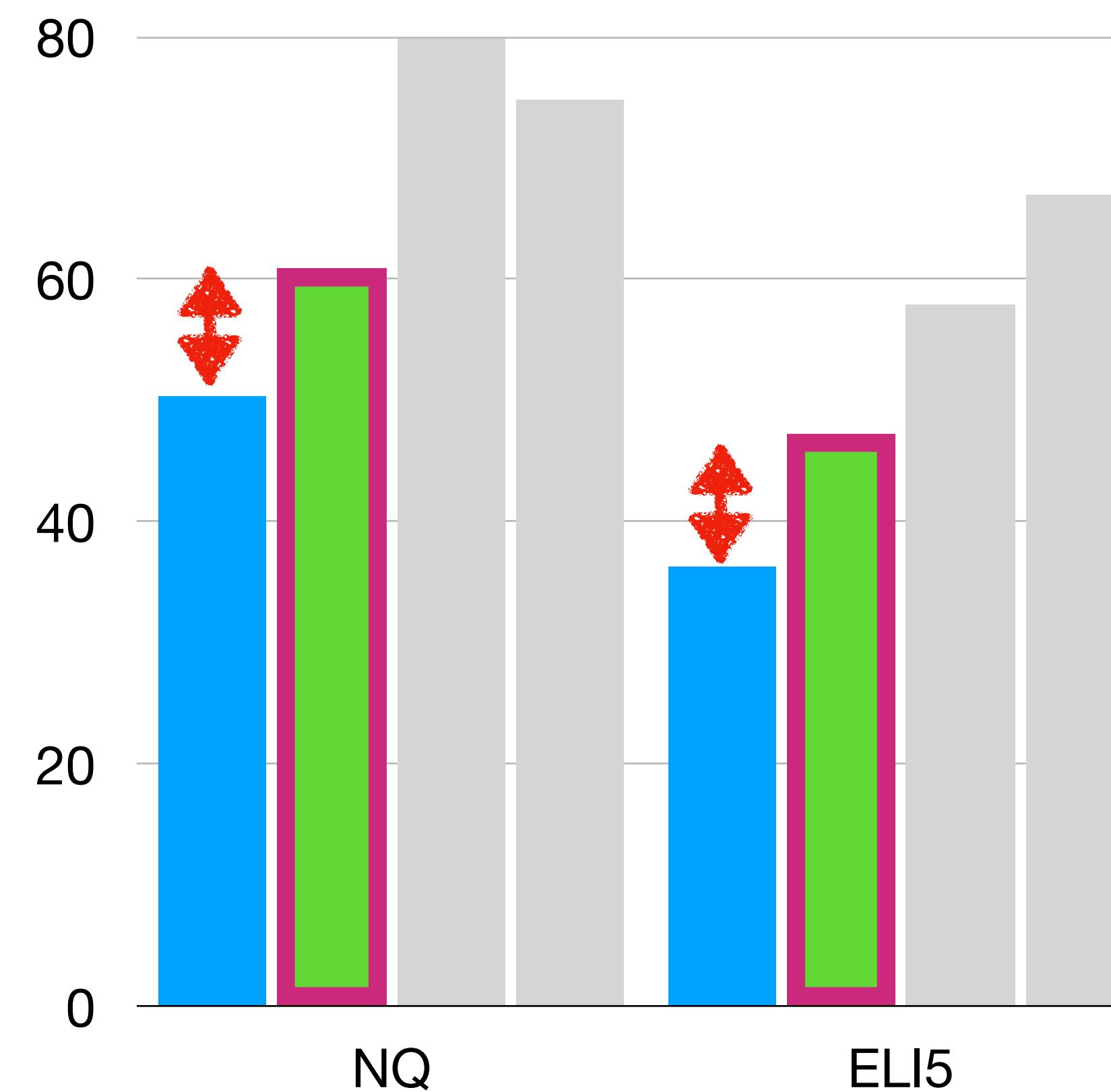
33k Human preference data



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

Effects of RL

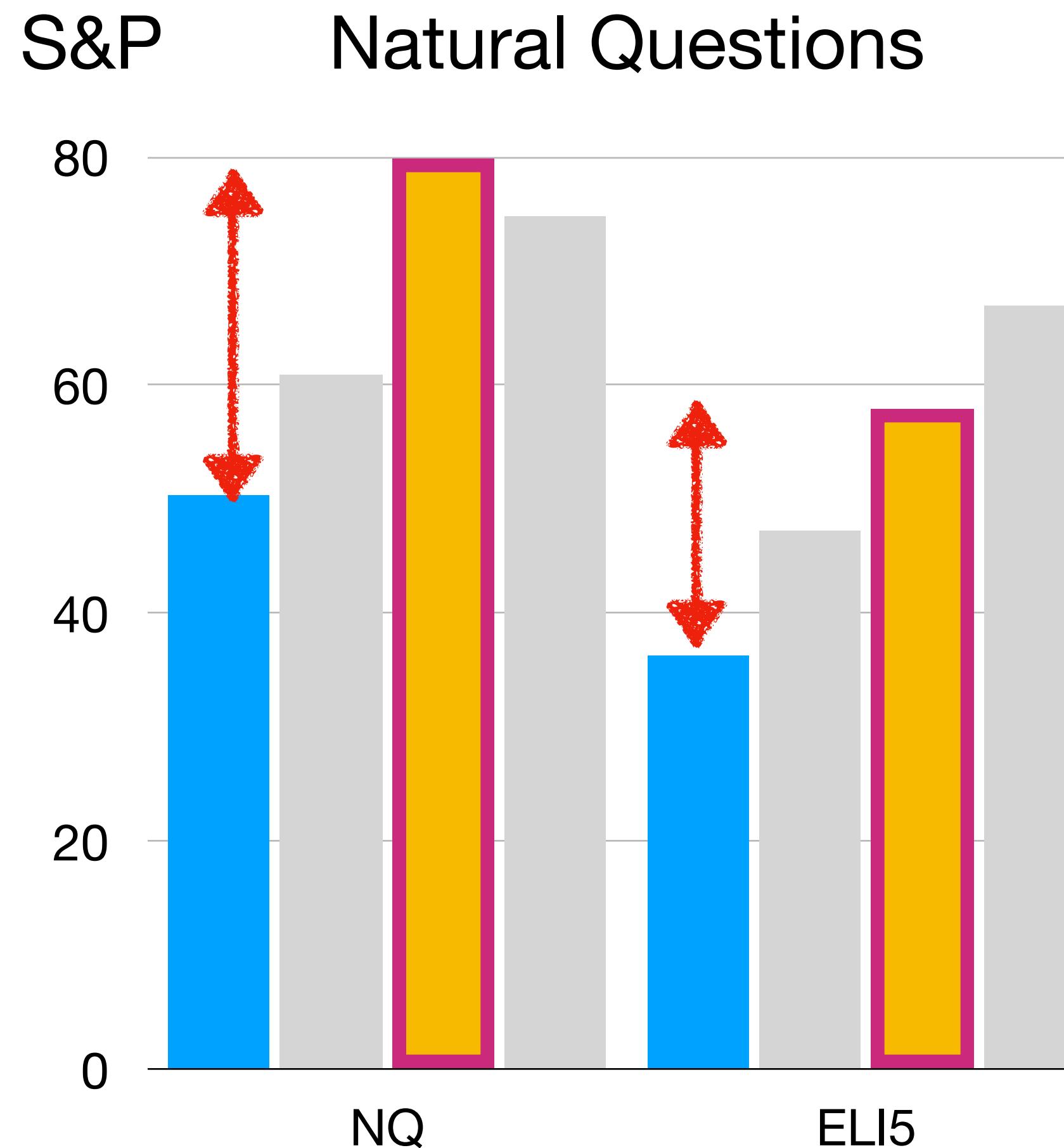
S&P Natural Questions



RL w/ human feedback improves the quality of top 1 generations

(S&P = Supported & Plausible)

Effects of RL



Sampling & reranking many generations
using a reward model gives gains from Top 1

FT (best 1) RL (best 1)
FT (best 64) RL (best 64)

Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (Retriever & 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 **fine-tuning**



Customizable



Competitive w/ more data



Requiring training

Summary of downstream adaptations

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



Requiring additional data collection (preference)

Summary of downstream adaptations

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

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.

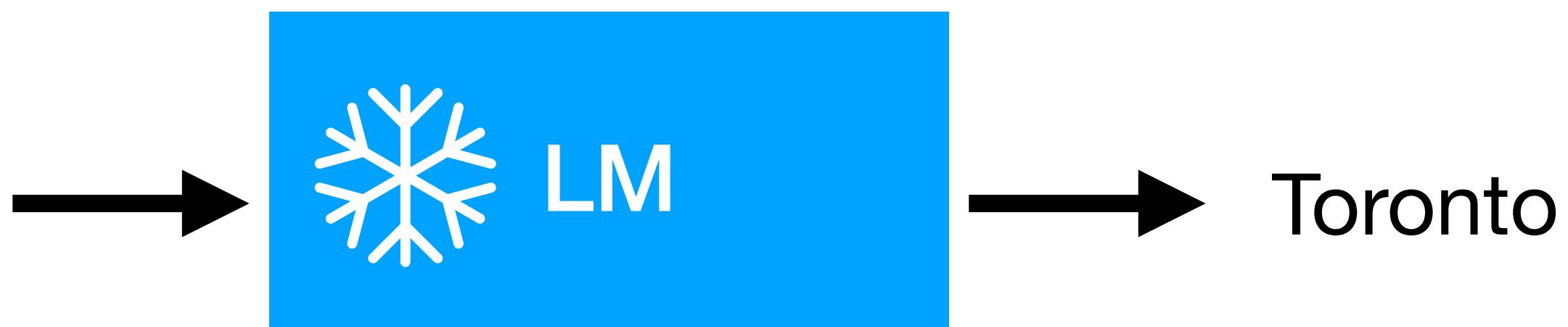
##

Q: what is the Ontario capital?

A:

Doesn't require LM training on tasks!

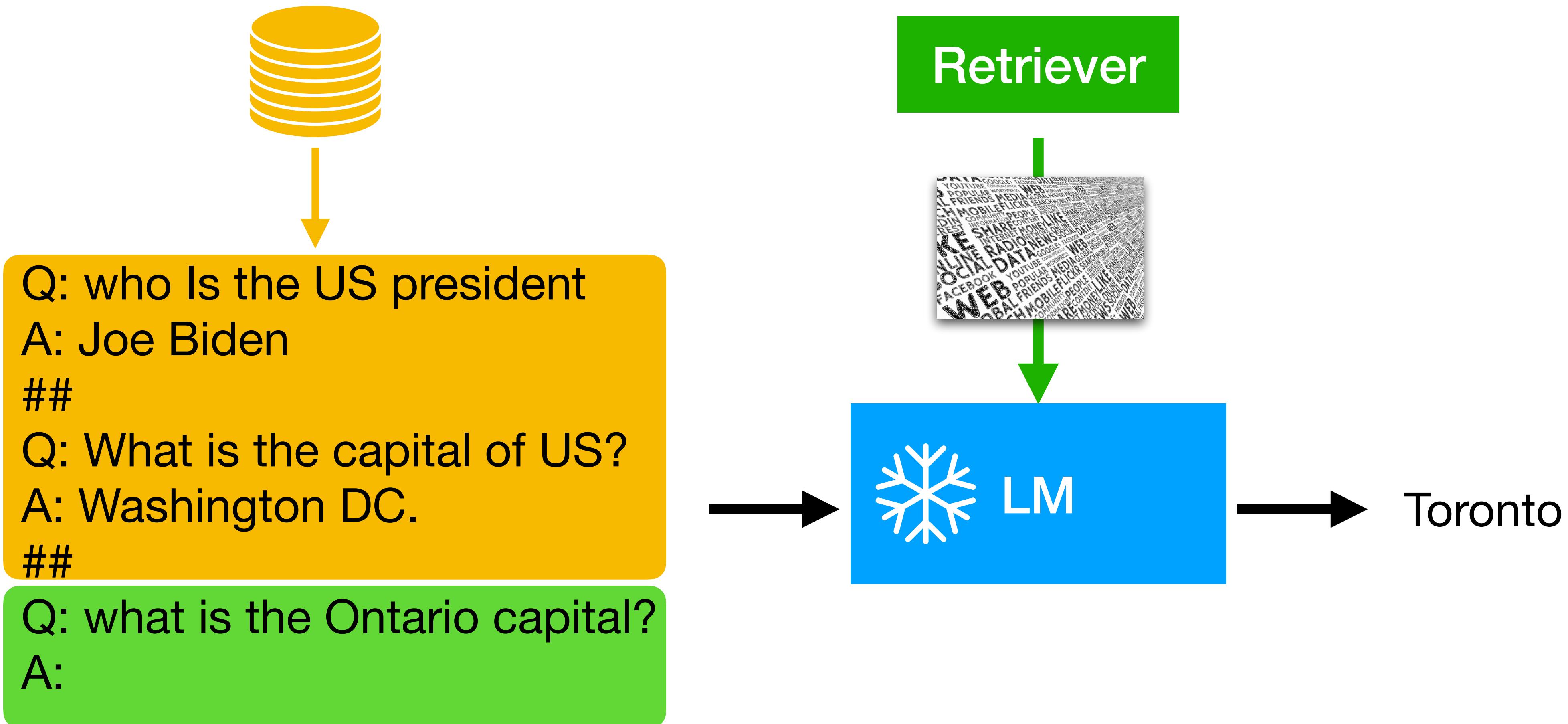
Training instances (demonstrations)



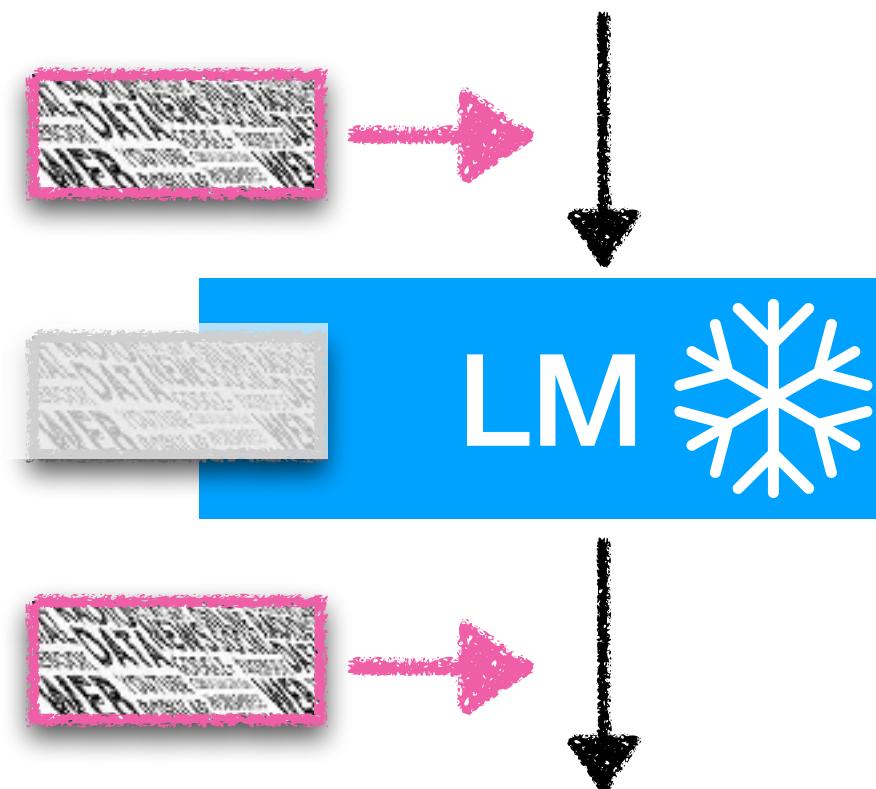
Test instances

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

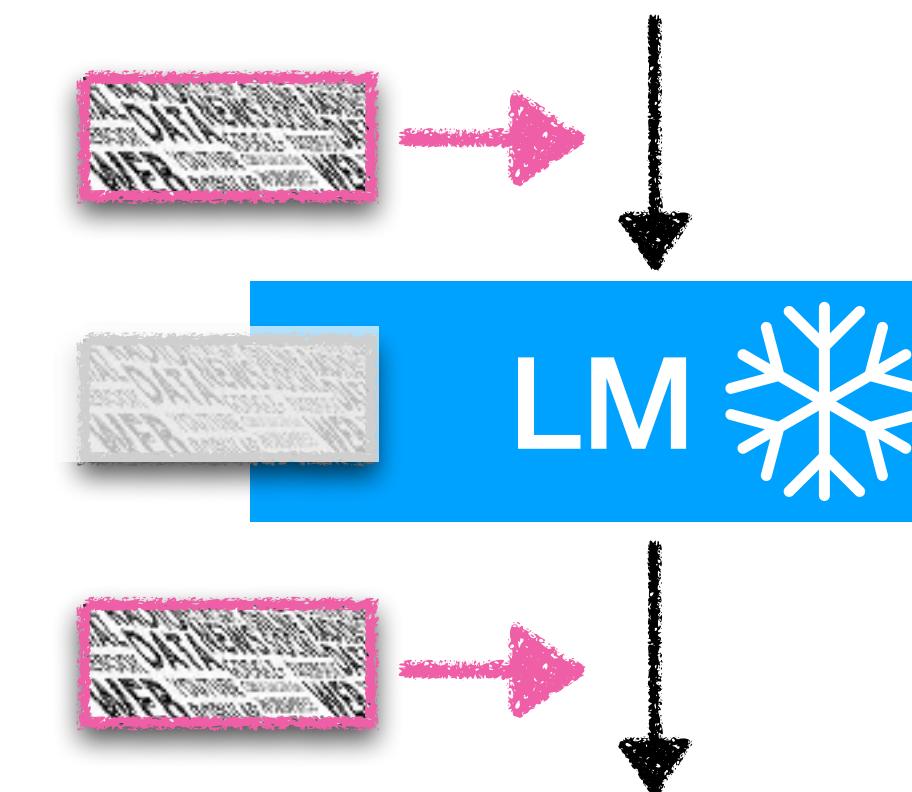
Intermediate layers:

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

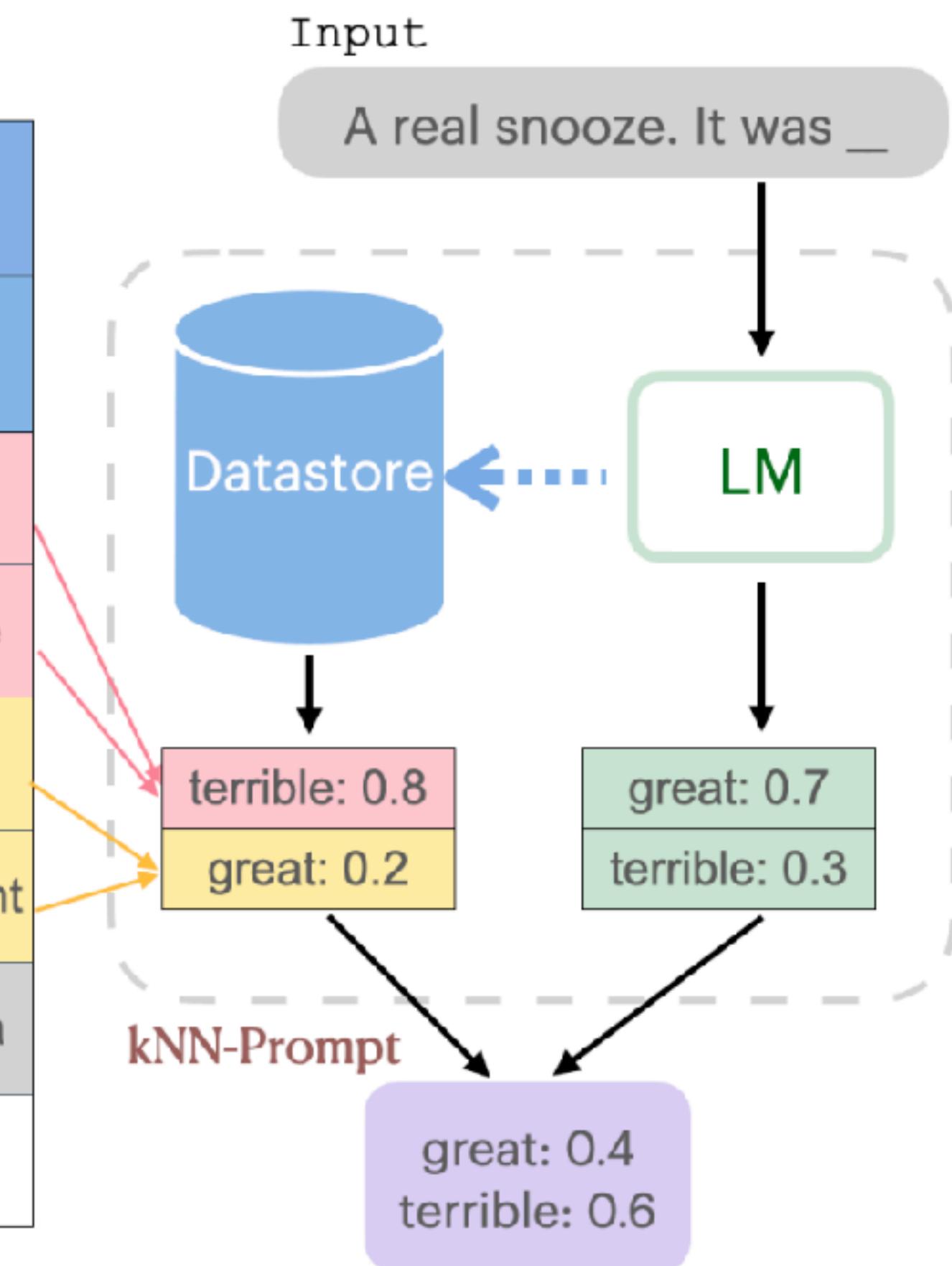
Intermediate layers:
N/A

Output space:
Interpolate token probability distributions in output space

kNN-Prompt (Shi et al., 2022)

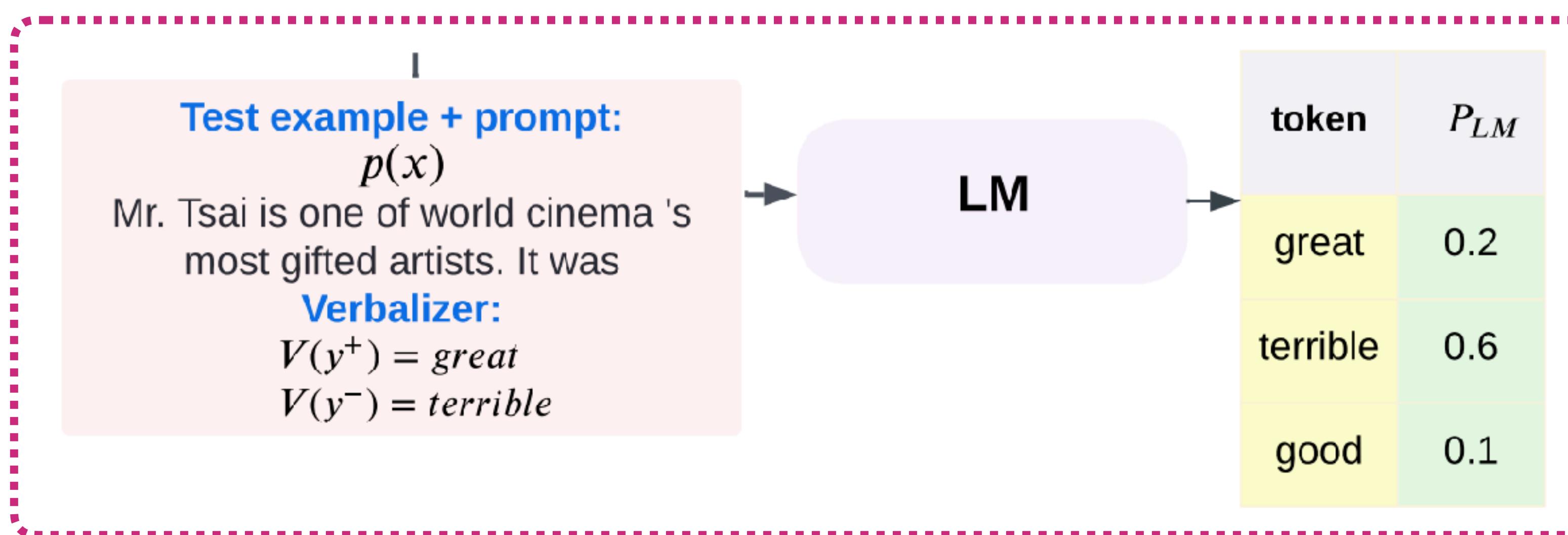
kNN LM with fuzzy verbalizers
for zero-/few-shot **classification**

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)

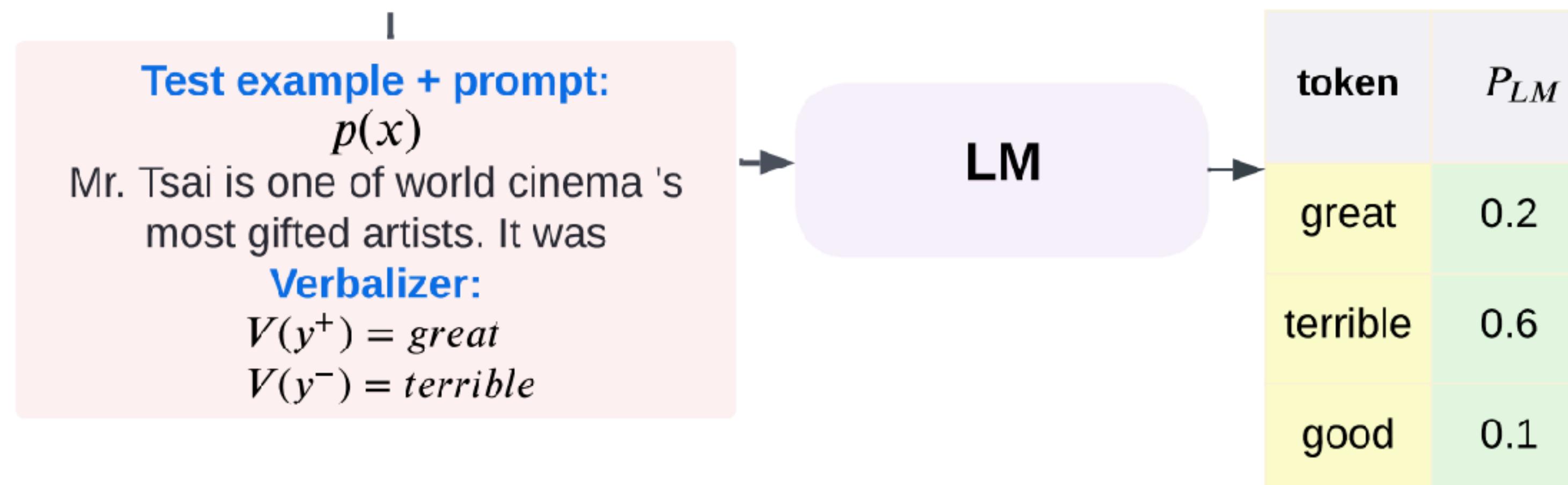
LM predicts next token



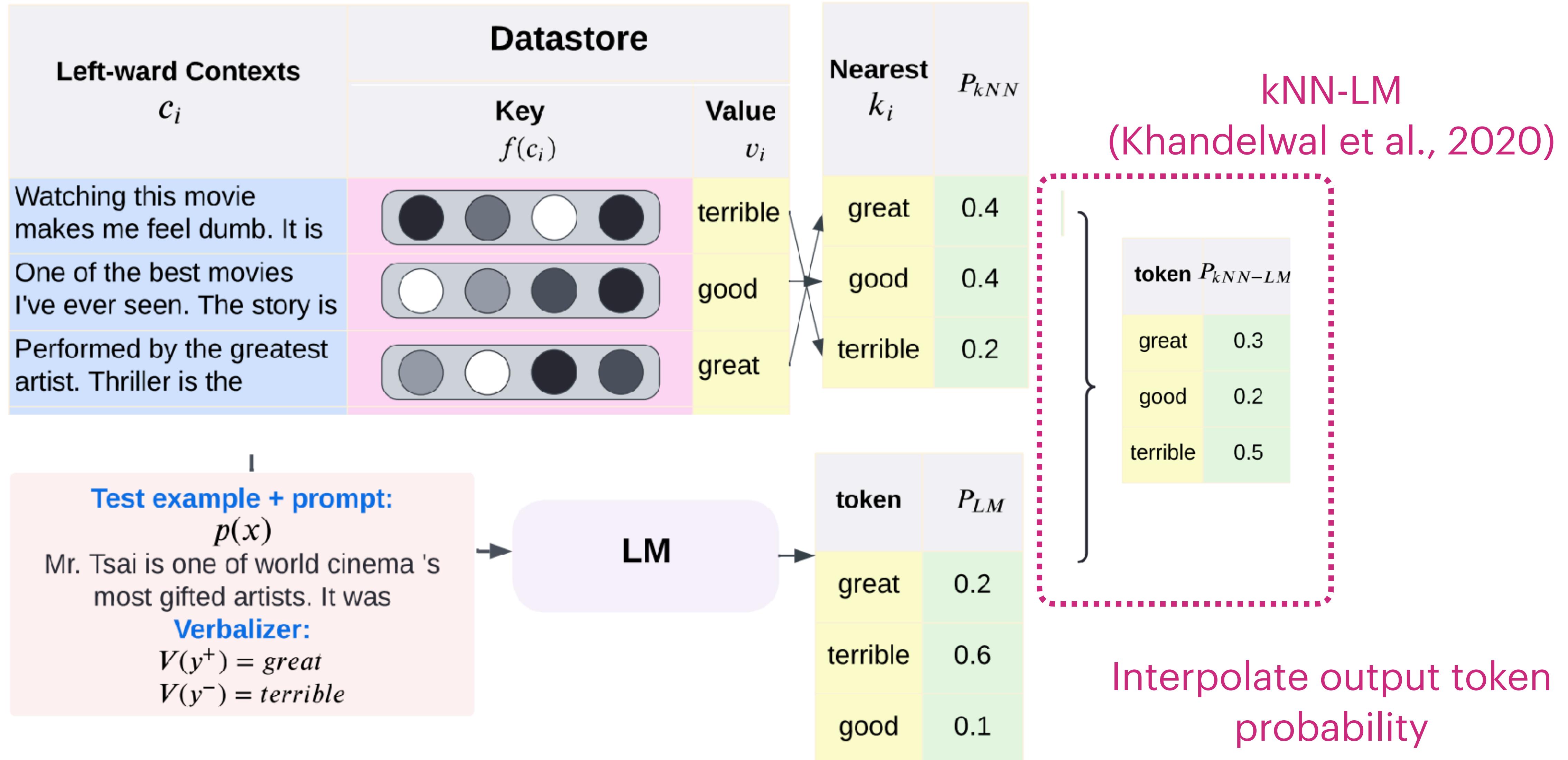
kNN-Prompt (Shi et al., 2022)

Left-ward Contexts		Datastore		Nearest k_i	P_{kNN}
c_i		Key $f(c_i)$	Value v_i		
Watching this movie makes me feel dumb. It is			terrible	great	0.4
One of the best movies I've ever seen. The story is			good	good	0.4
Performed by the greatest artist. Thriller is the			great	terrible	0.2

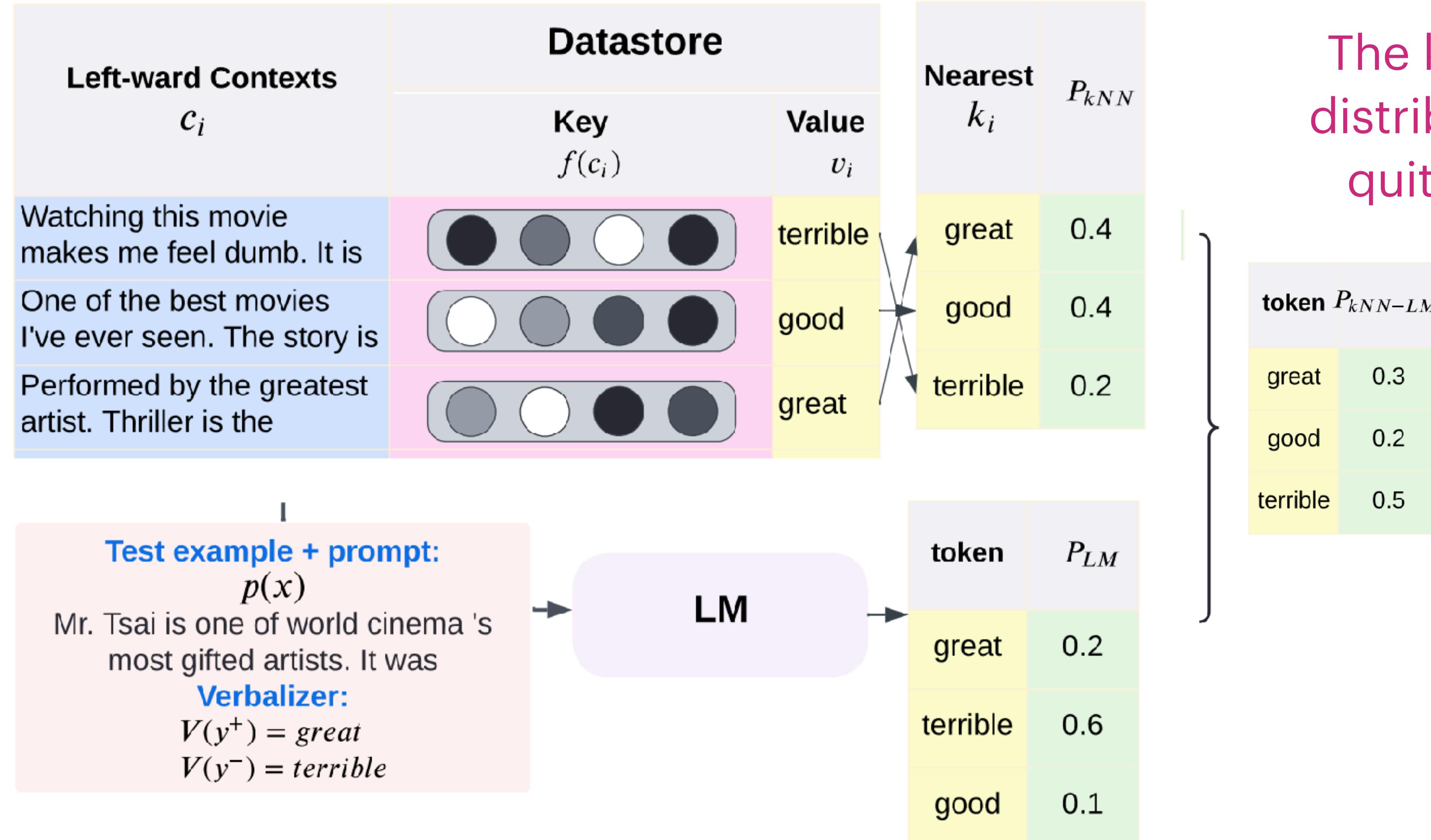
kNN predicts next tokens



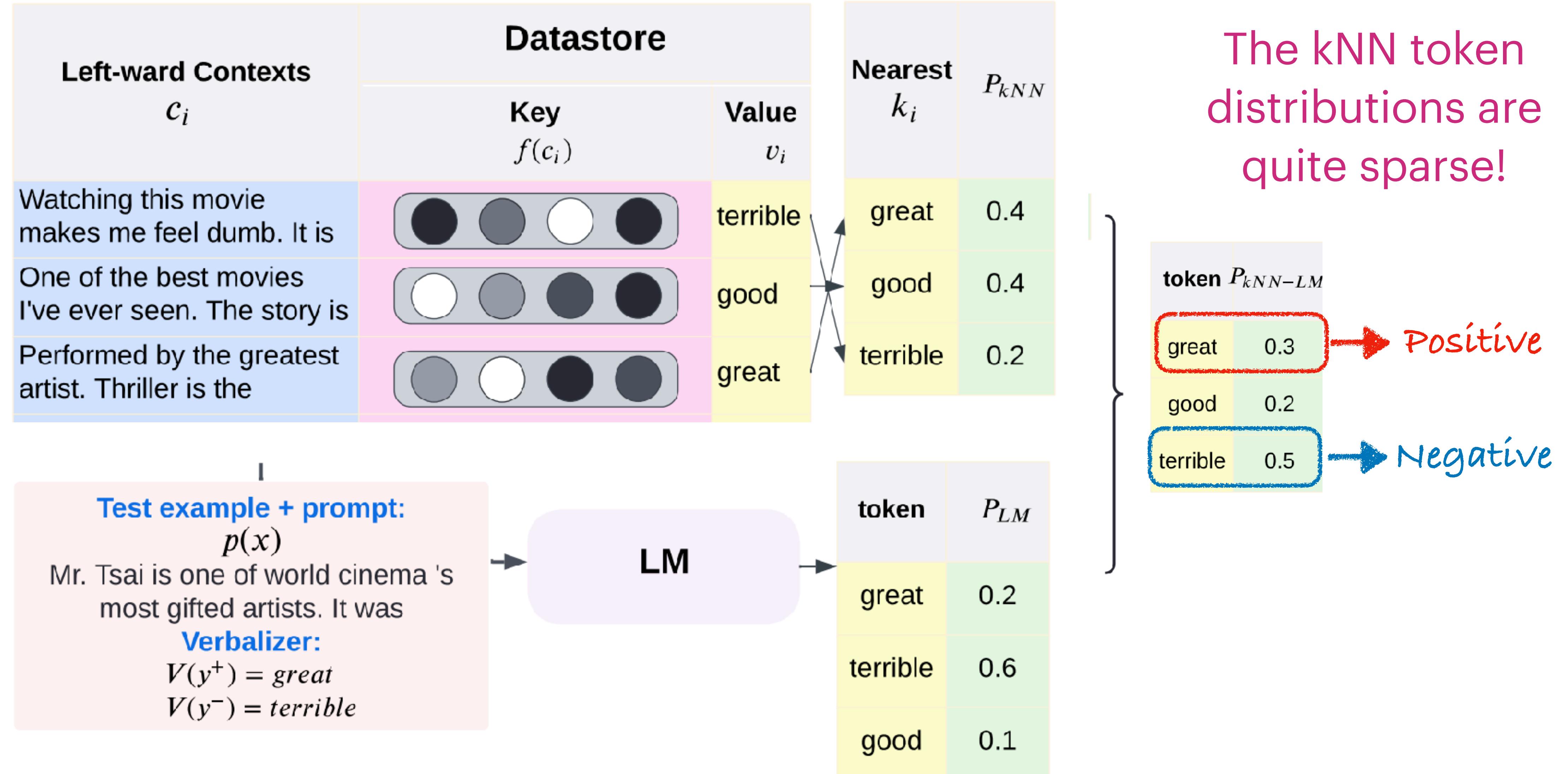
kNN-Prompt (Shi et al., 2022)



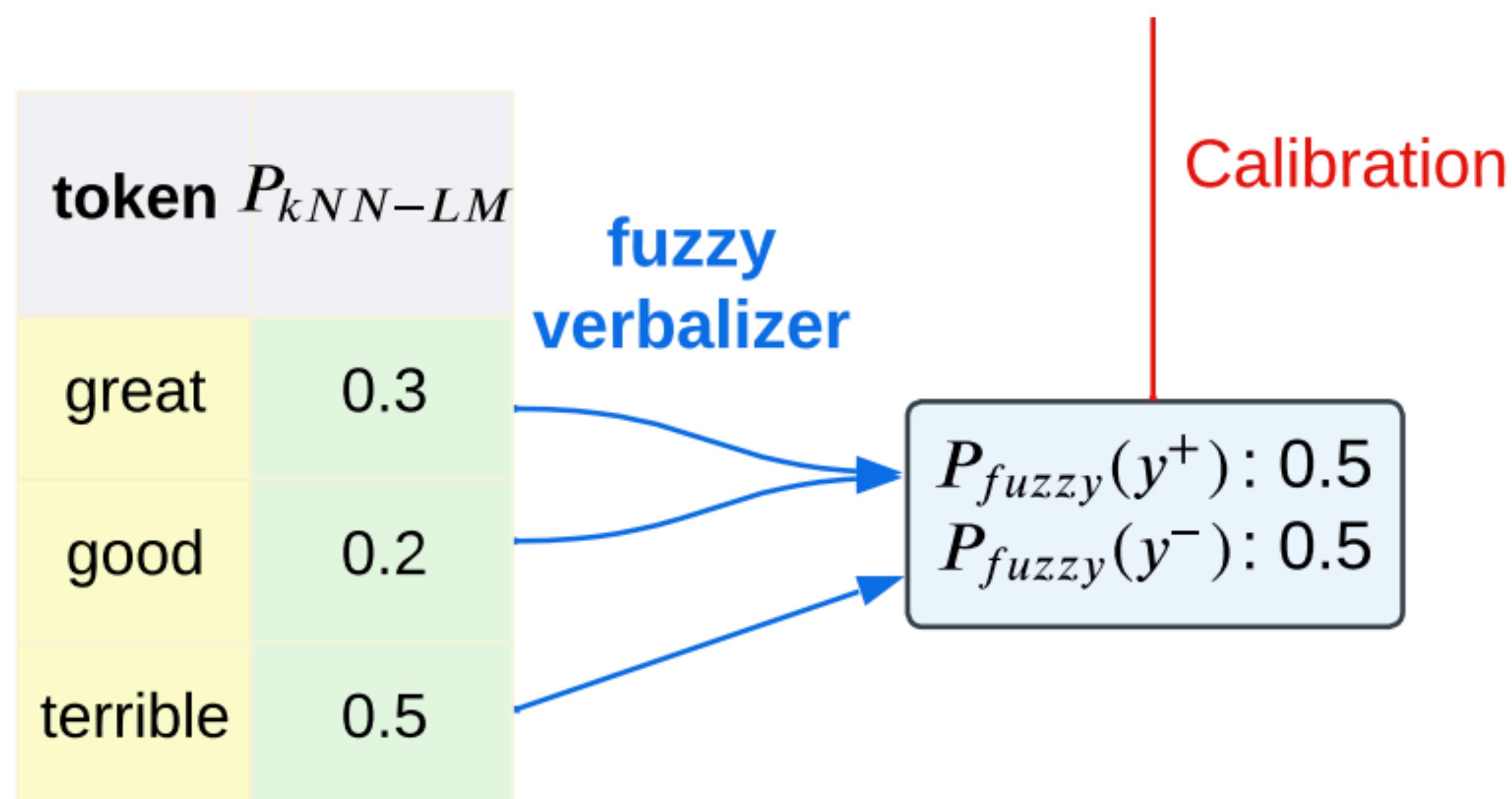
kNN-Prompt (Shi et al., 2022)



kNN-Prompt (Shi et al., 2022)



kNN-Prompt (Shi et al., 2022)

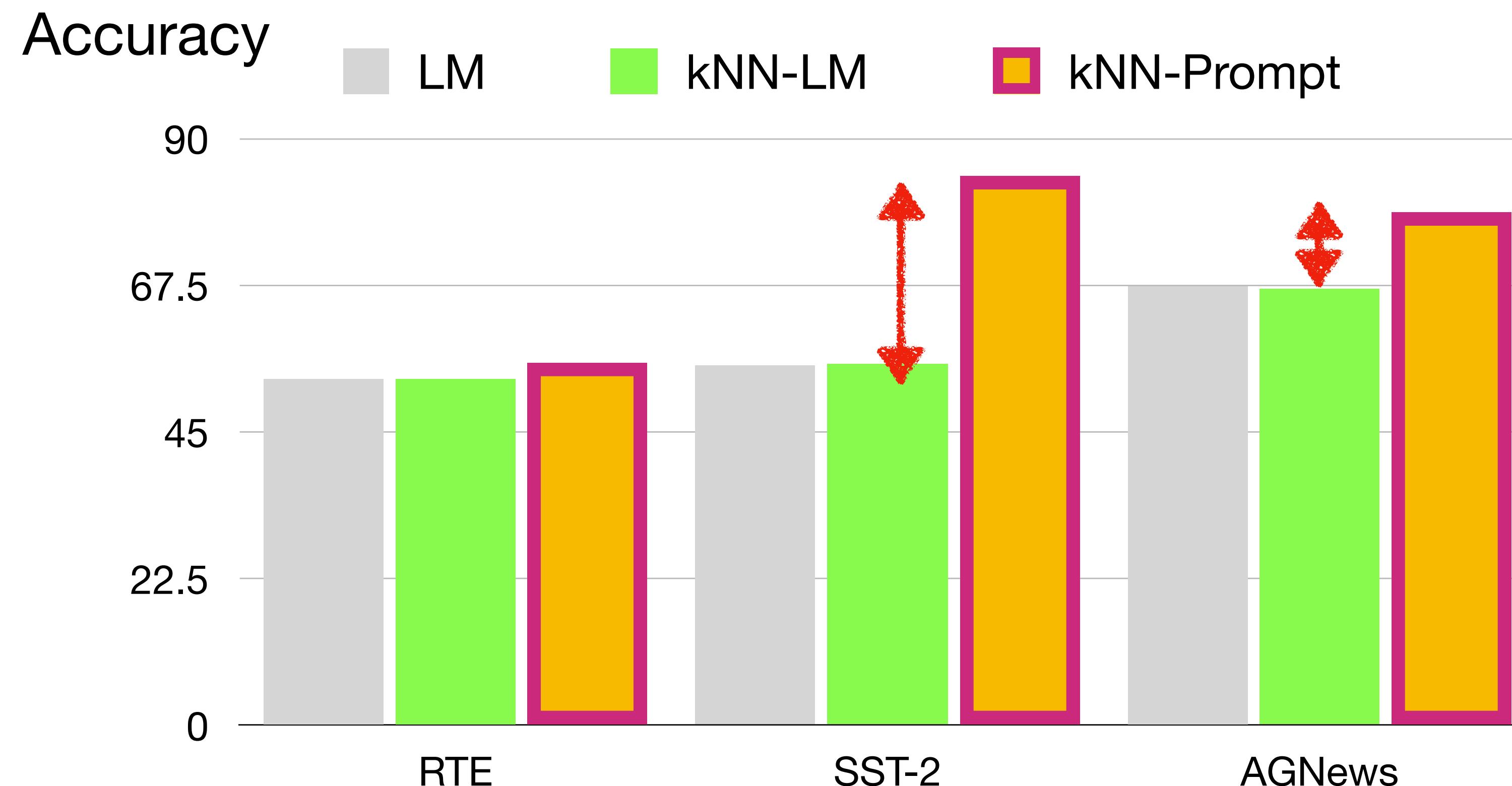


Fuzzy verbalizer maps token probability to target class labels

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

Find similar tokens using GloVe & ConceptNet

Results on diverse classification tasks



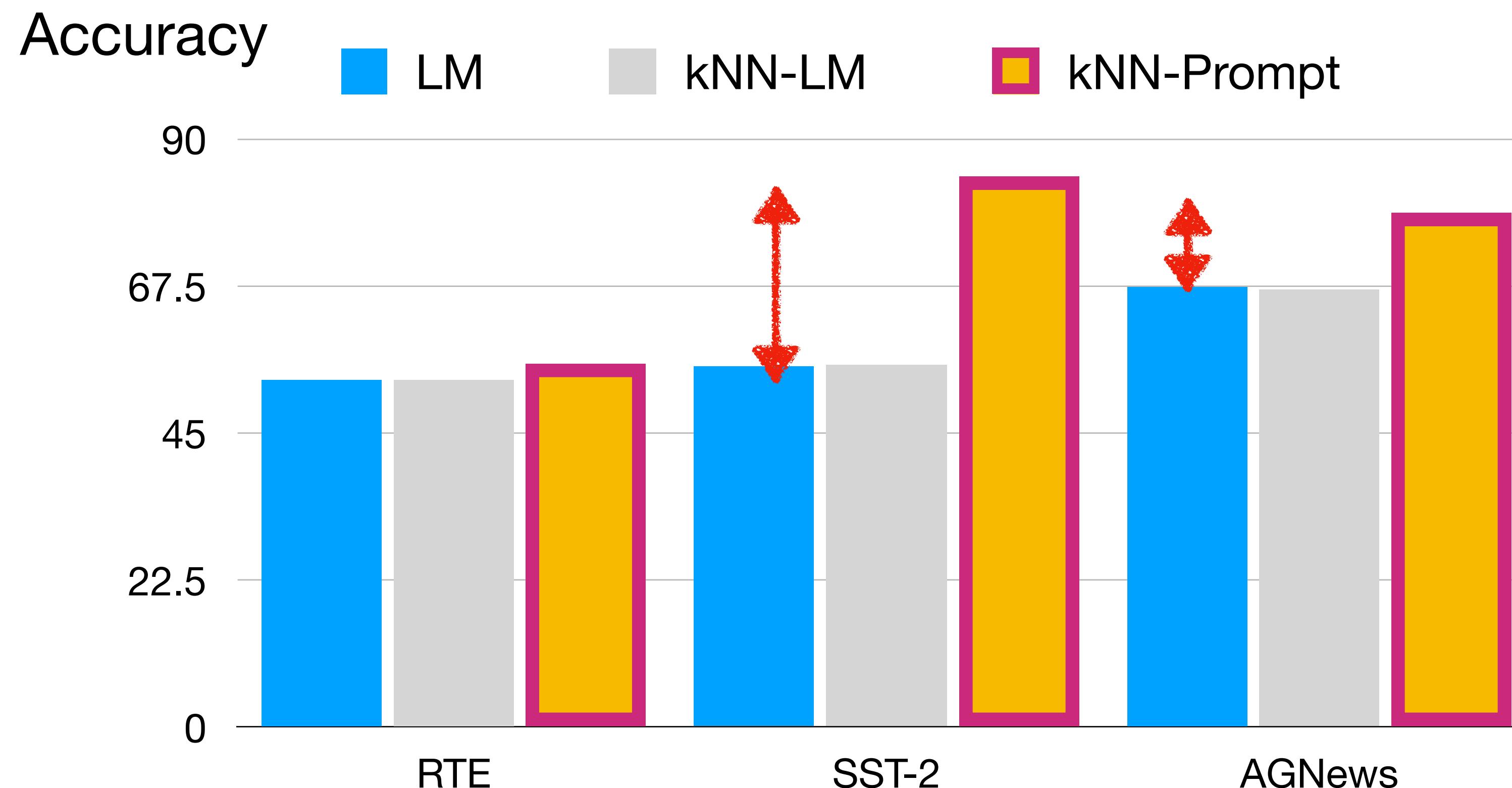
NLI
/ entailment

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

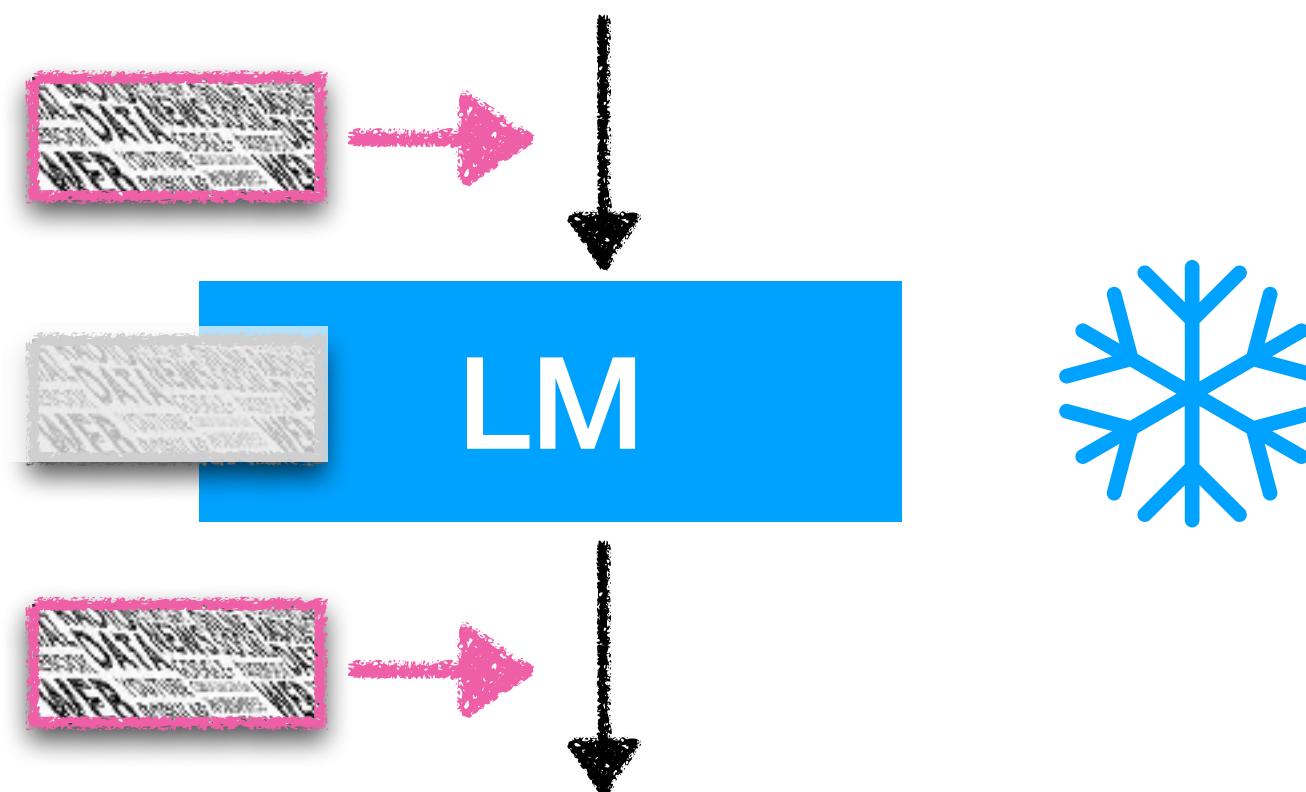
Topic
classification

Summary of downstream adaptations

	Target task	Adaptation method	Datastore
ATLAS (Izacard et al., 2022)	Knowledge-intensive	Fine-tuning (Retriever & 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 Pile

Retrieval-based LMs are effective in general NLU tasks!

Retrieval-based Prompting

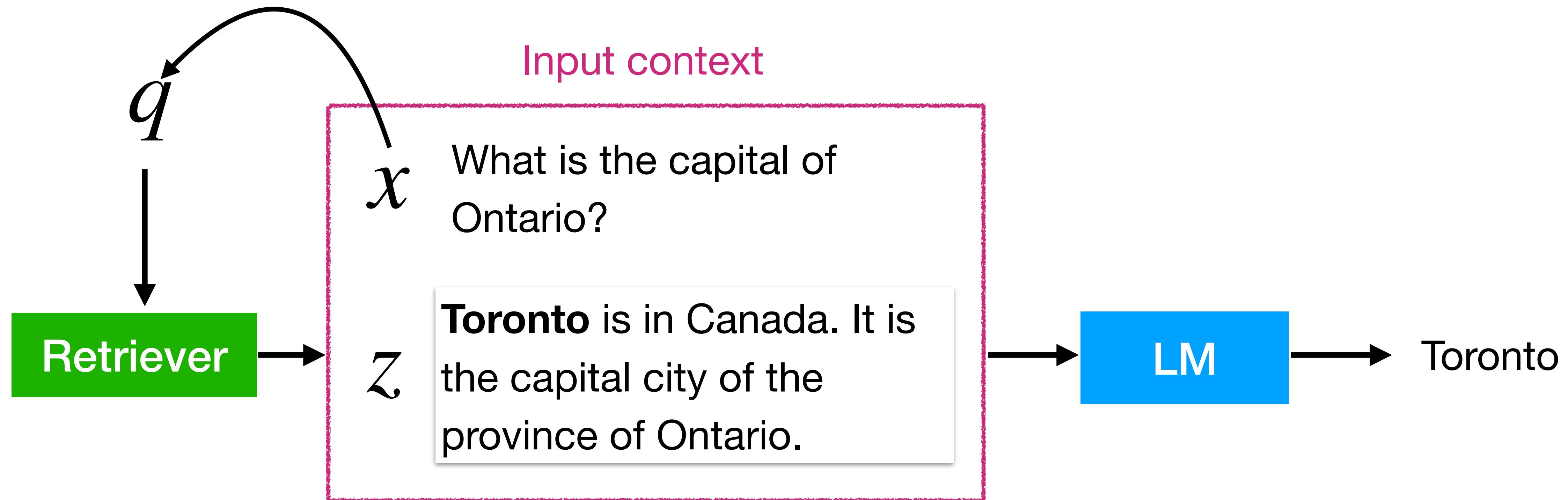


Input space:
Incorporate retrieved context in input space

Intermediate layers:
N/A

Output space:
Interpolate token probability
distributions in output space

Retrieval-in-context

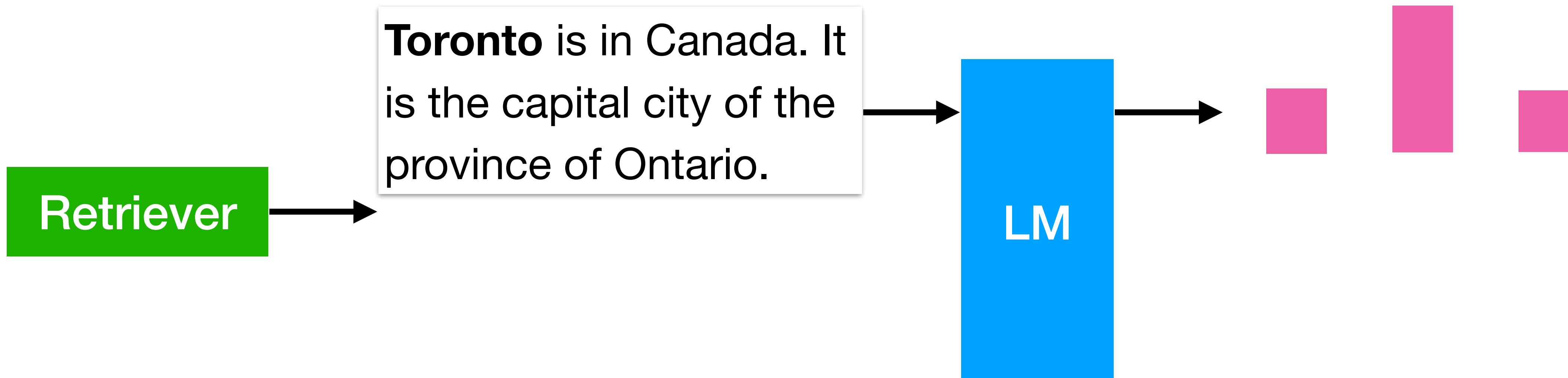


(Shi et al., 2023; Ram et al., 2022; Mallen et al., 2022; Yu et al., 2022; Press et al., 2022; *inter alia*)

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

✗ What is the capital of Ontario?

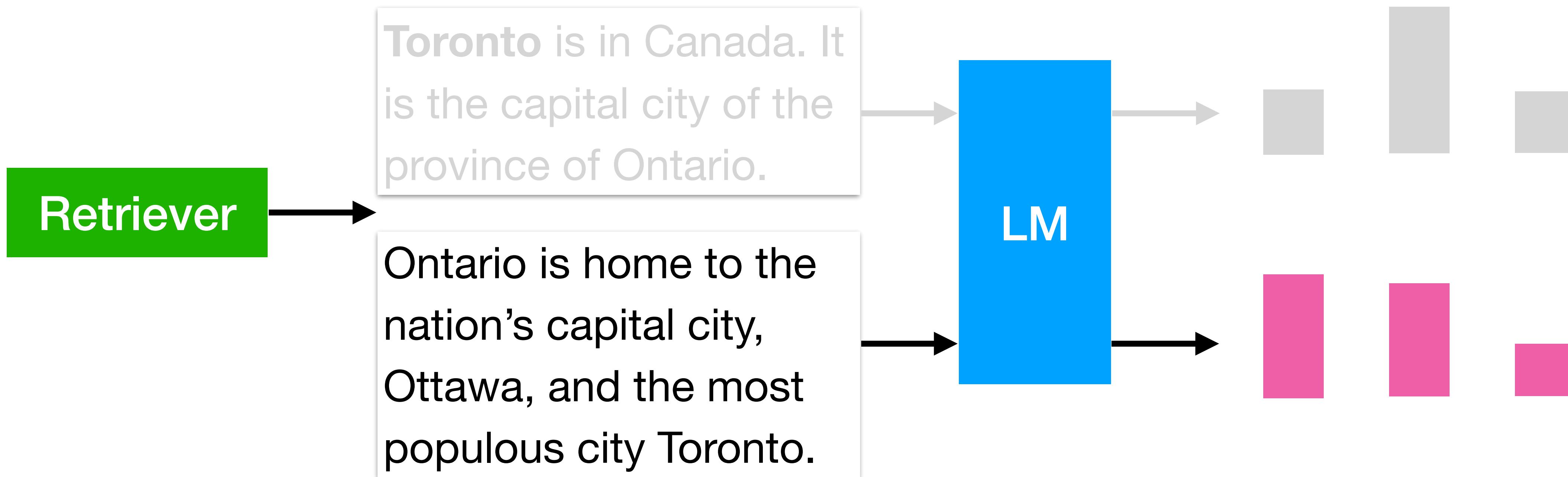
Ottawa **Toronto** Ontario



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

✗ What is the capital of Ontario?

Ottawa Toronto Ontario



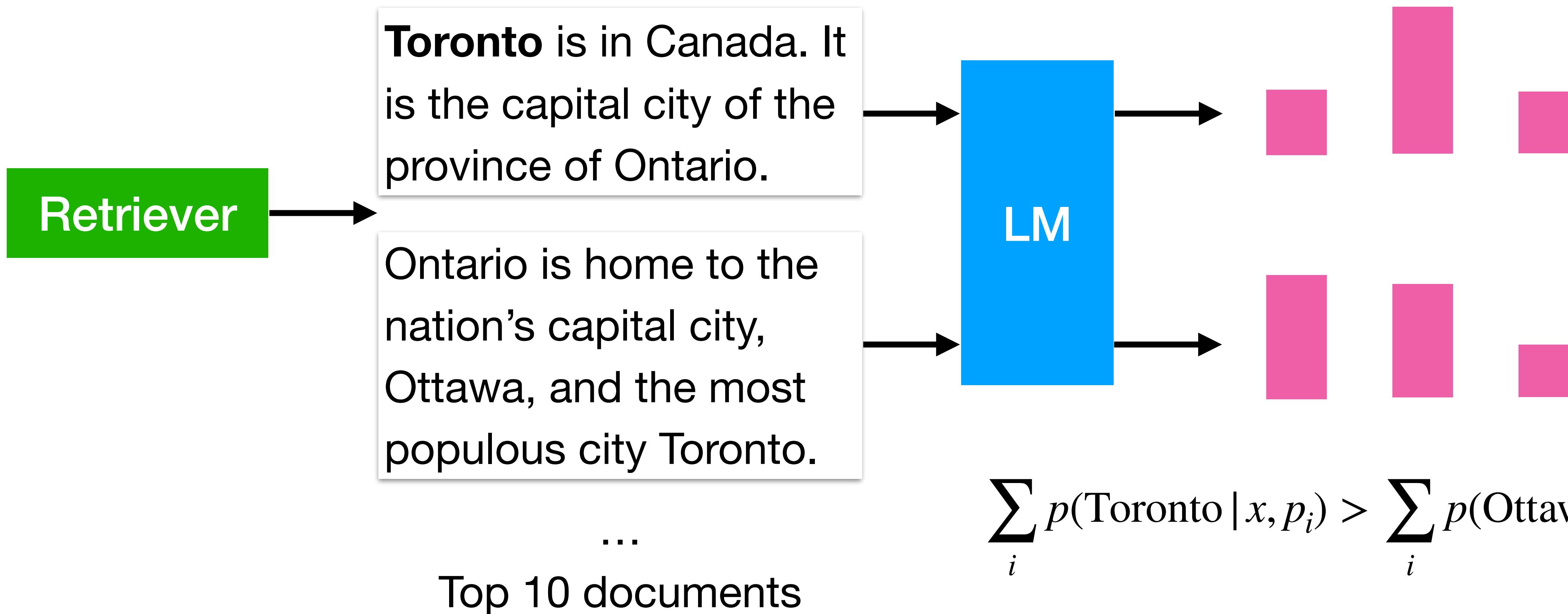
...

Top 10 documents

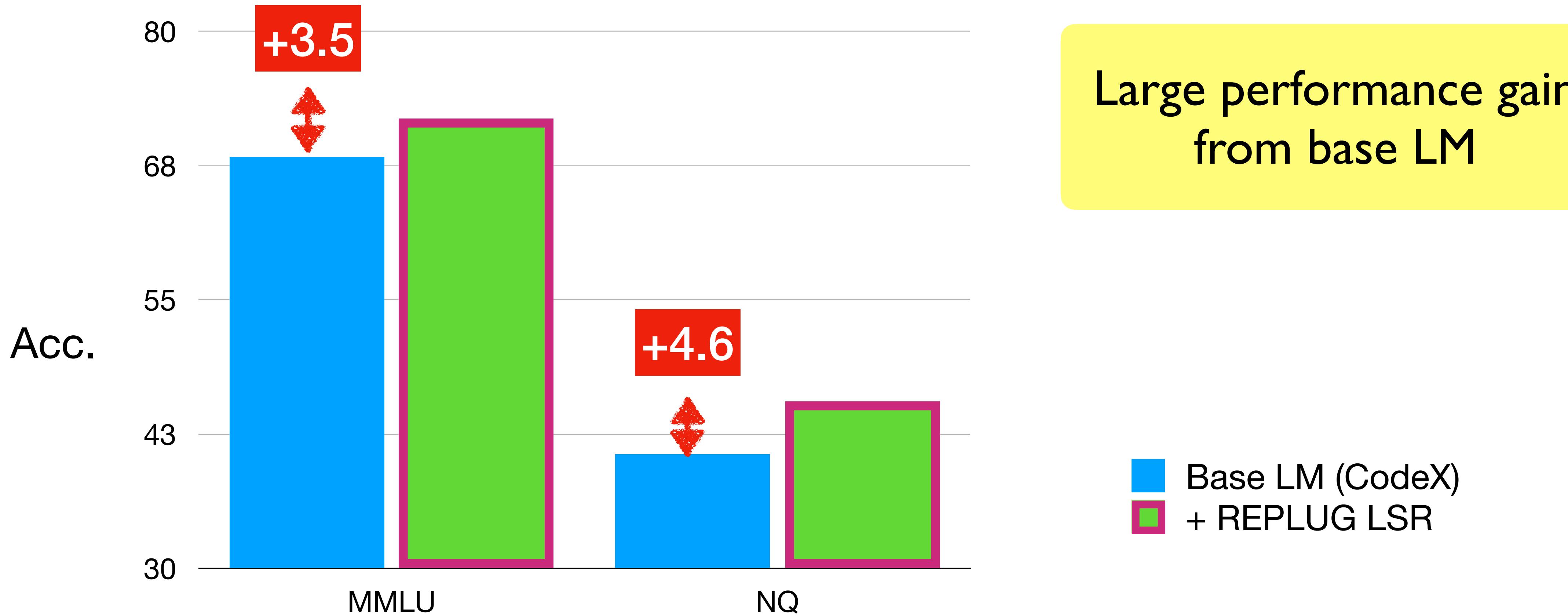
REPLUG (Shi et al., 2023; Section 3&4)

✗ What is the capital of Ontario?

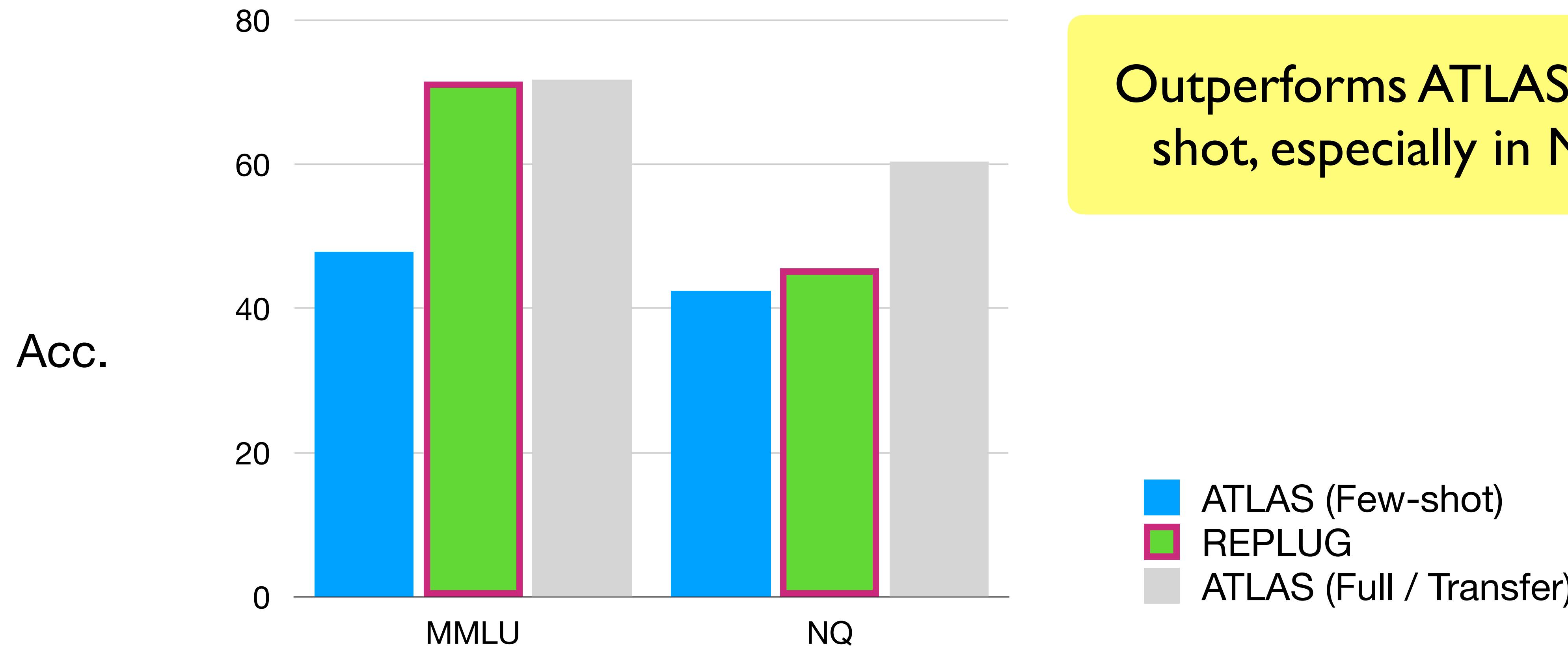
Ottawa **Toronto** Ontario



REPLUG: Results on QA & MMLU

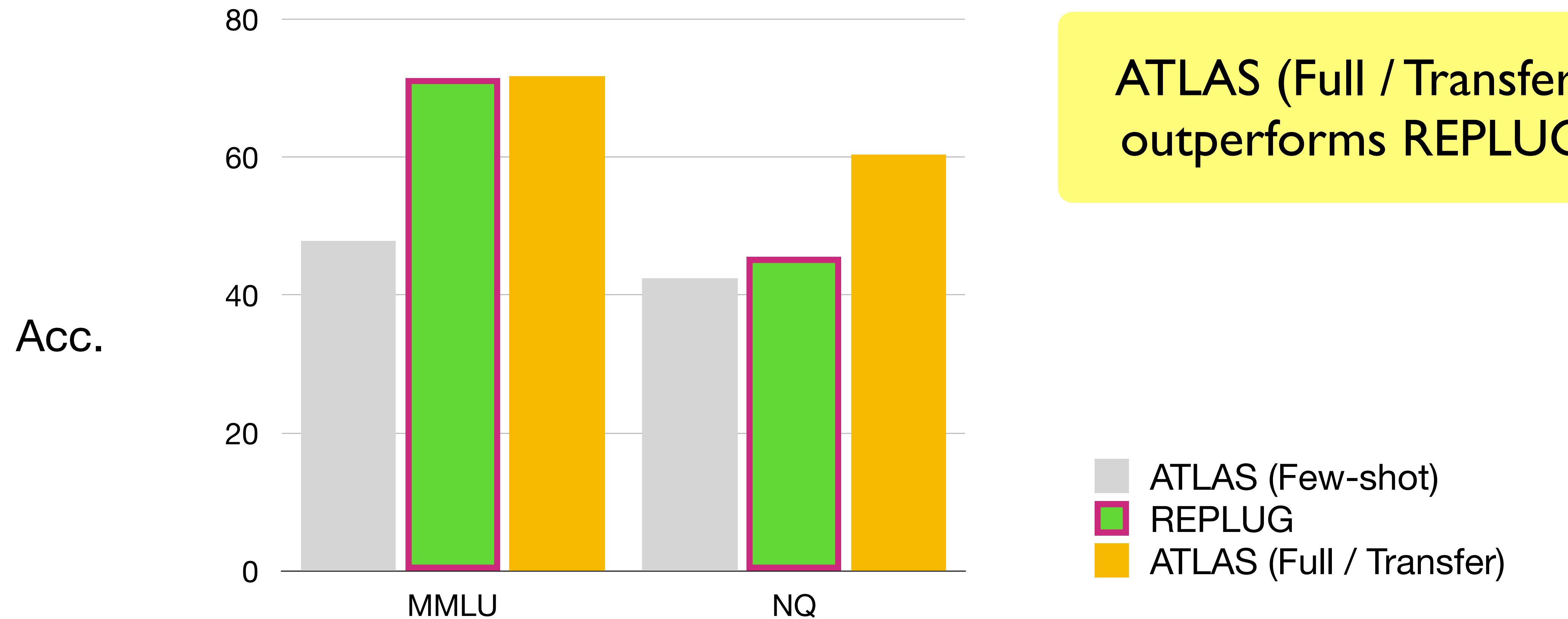


REPLUG: Comparison with ATLAS



Outperforms ATLAS in few-shot, especially in MMLU

REPLUG: Comparison with ATLAS



Summary of downstream adaptations

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



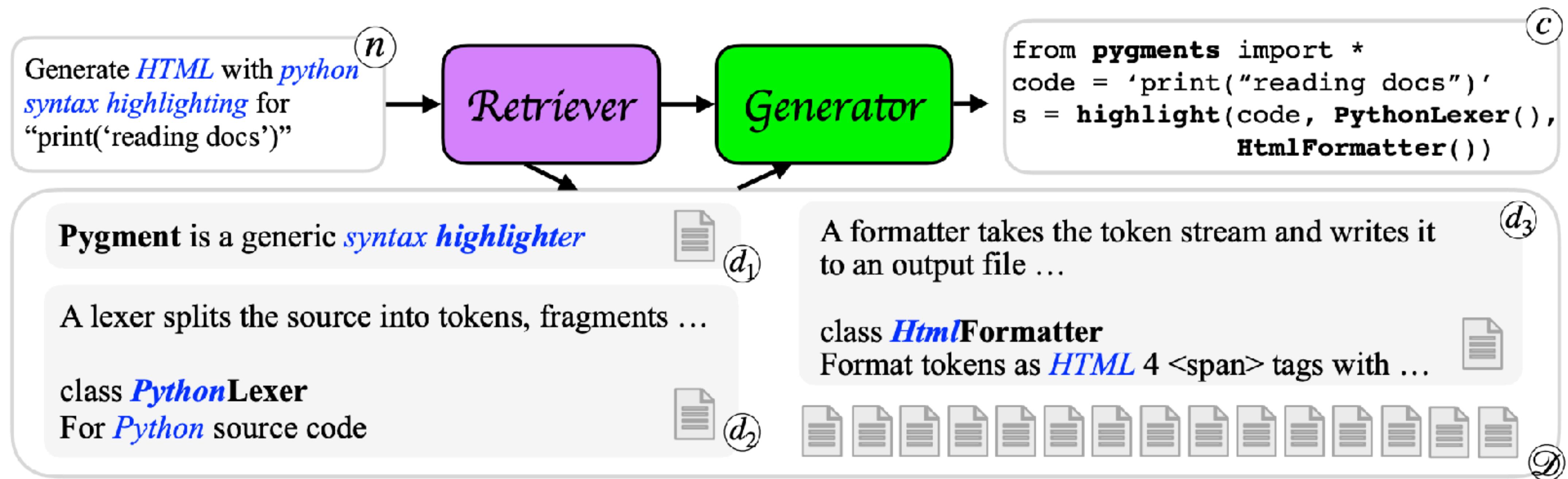
Hard to control, underperforming full FT model

Summary of downstream adaptations

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

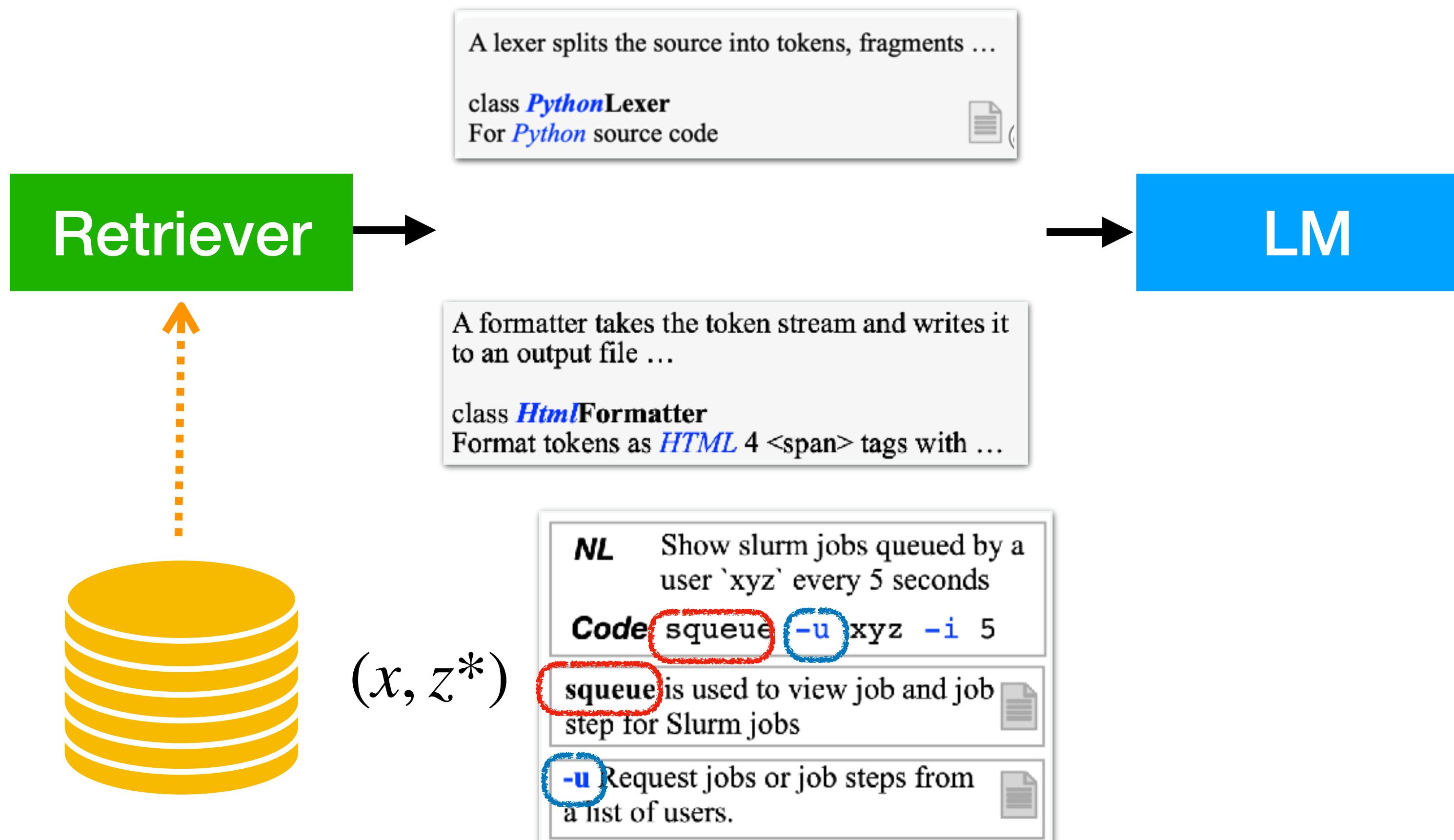
What can be other types of datastores?

DocPrompting (Zhou et al., 2023)

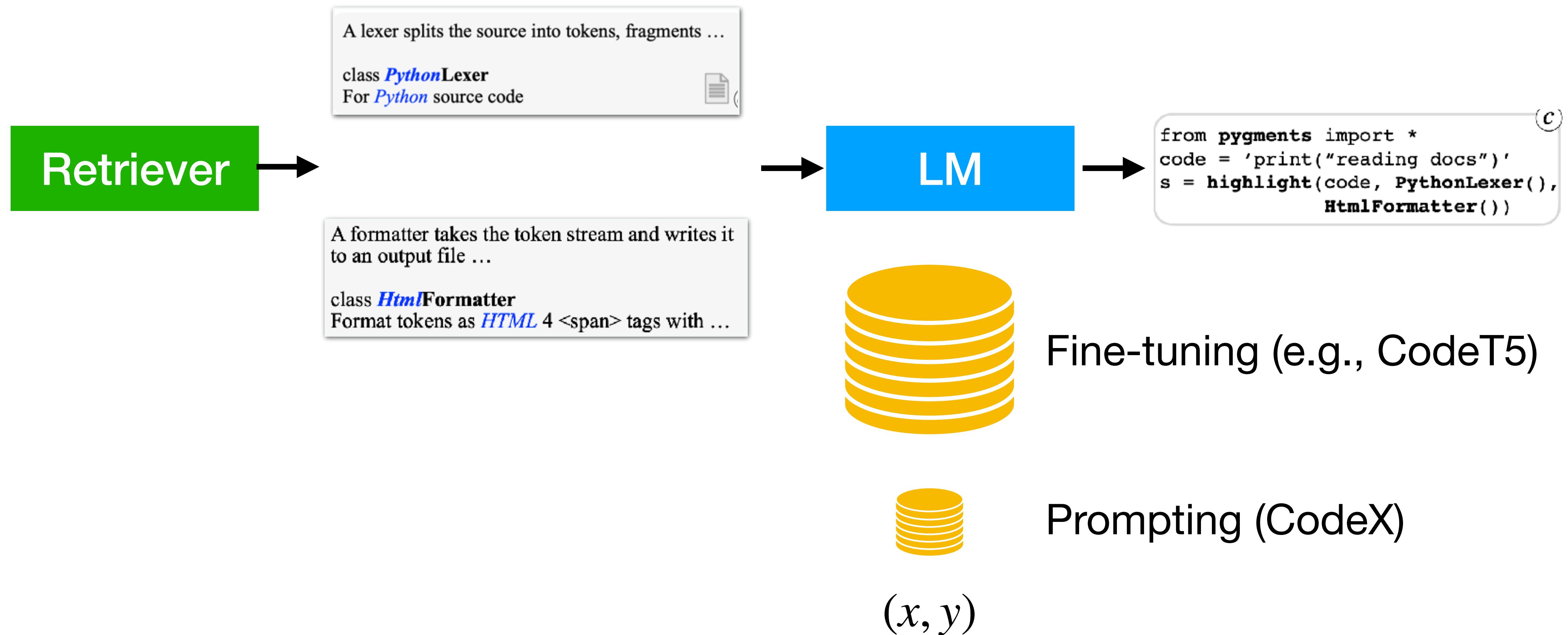


Retrieve **code documentations** about related functions

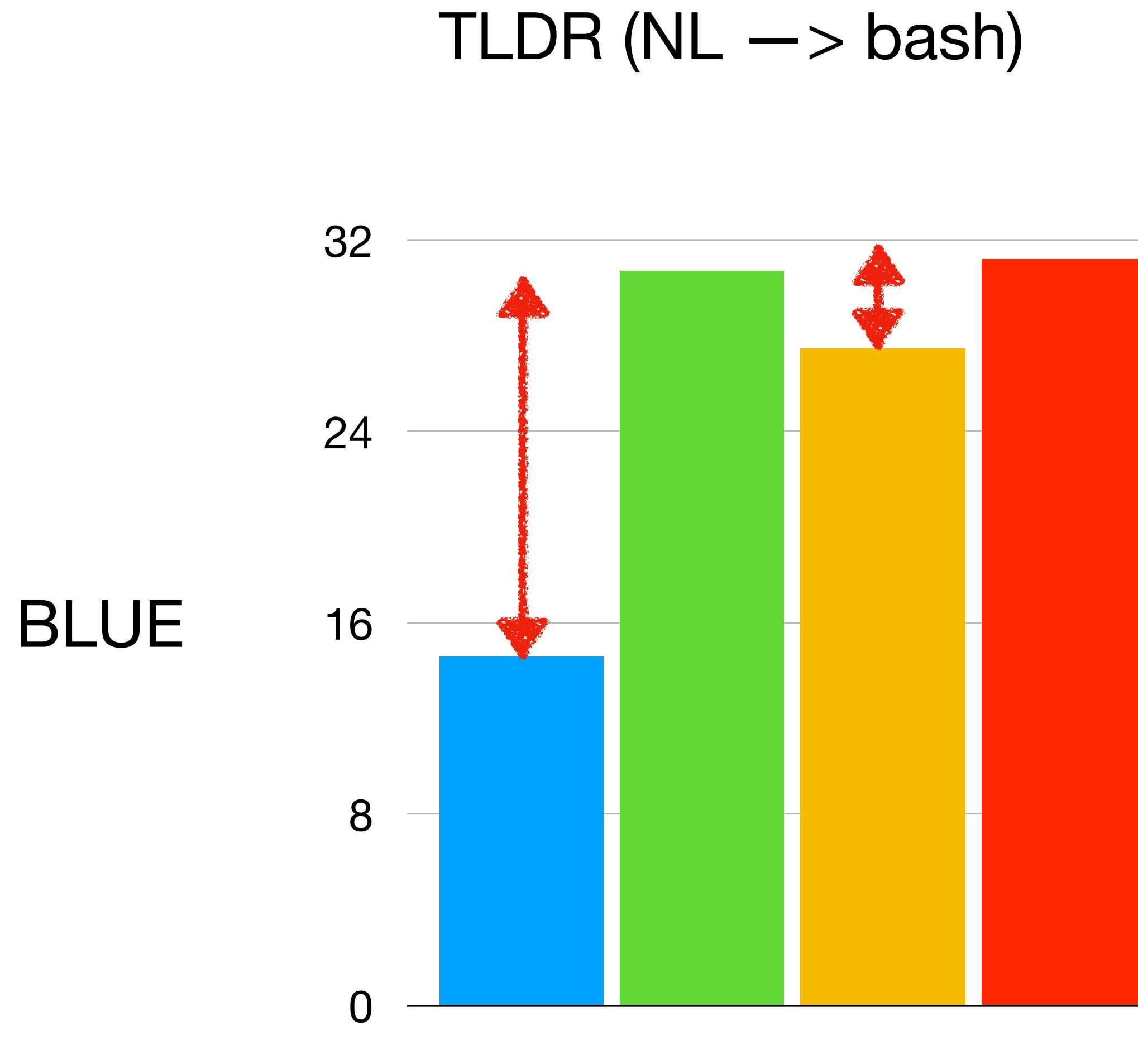
DocPrompting (Zhou et al., 2023)



DocPrompting (Zhou et al., 2023)

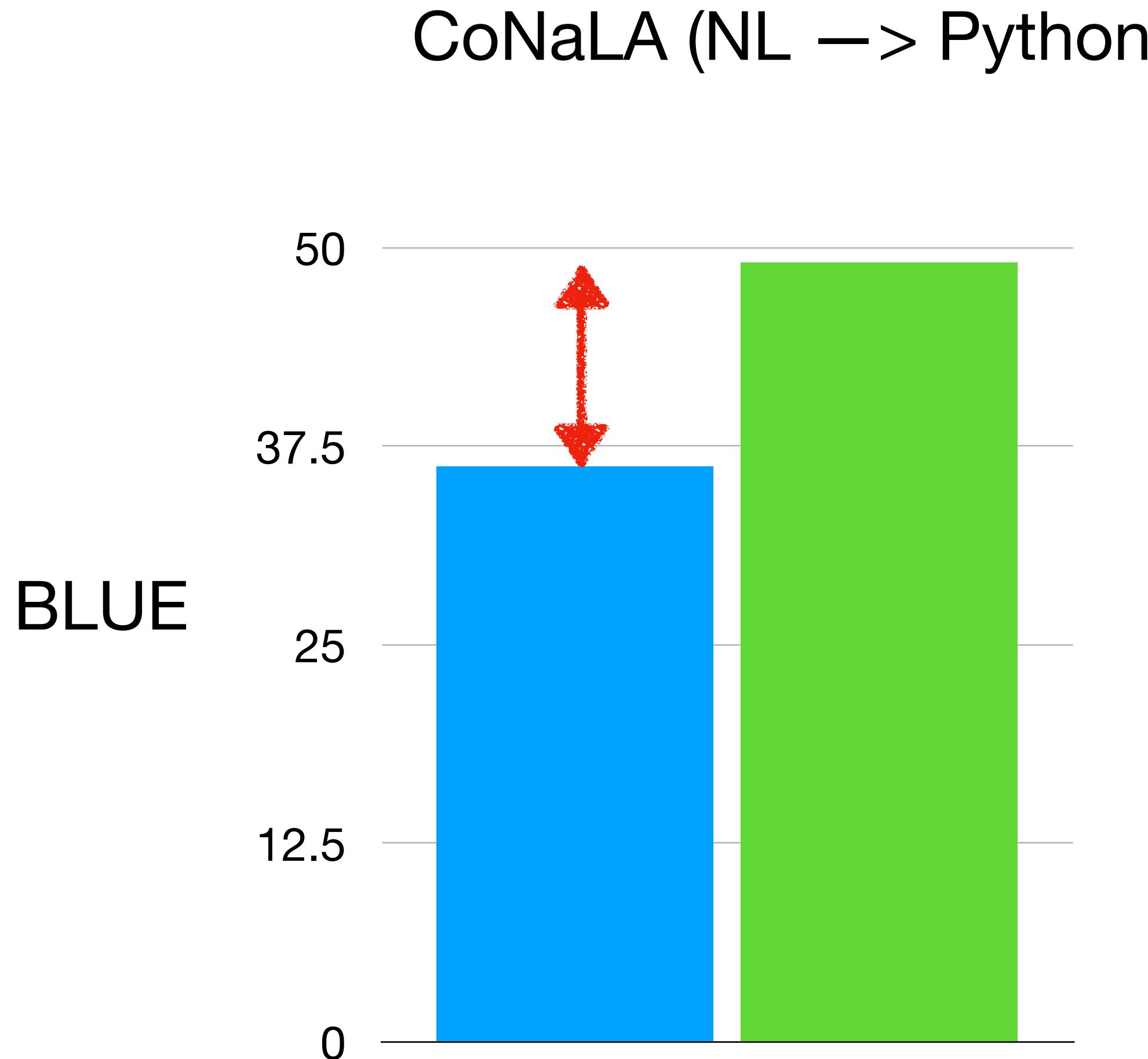


DocPrompting (Zhou et al., 2023)



Large gain given by DocPrompting
for both CodeT5 & CodeX

DocPrompting (Zhou et al., 2023)



Room for improvement in the retrieval component

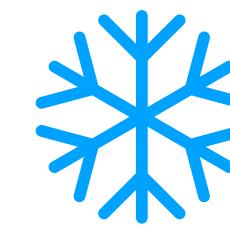
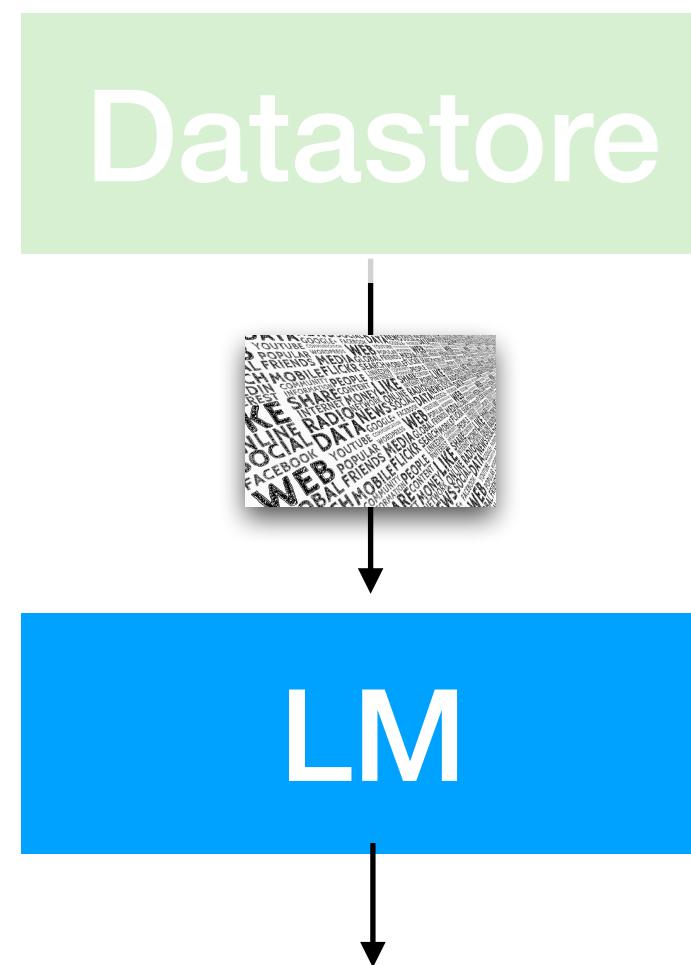
Active research in OOD / Zero-shot retrieval!
(BEIR; Thakur et al., 2021)

- + DocPrompting
- + DocPrompting (Oracle)

Summary of downstream adaptations

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

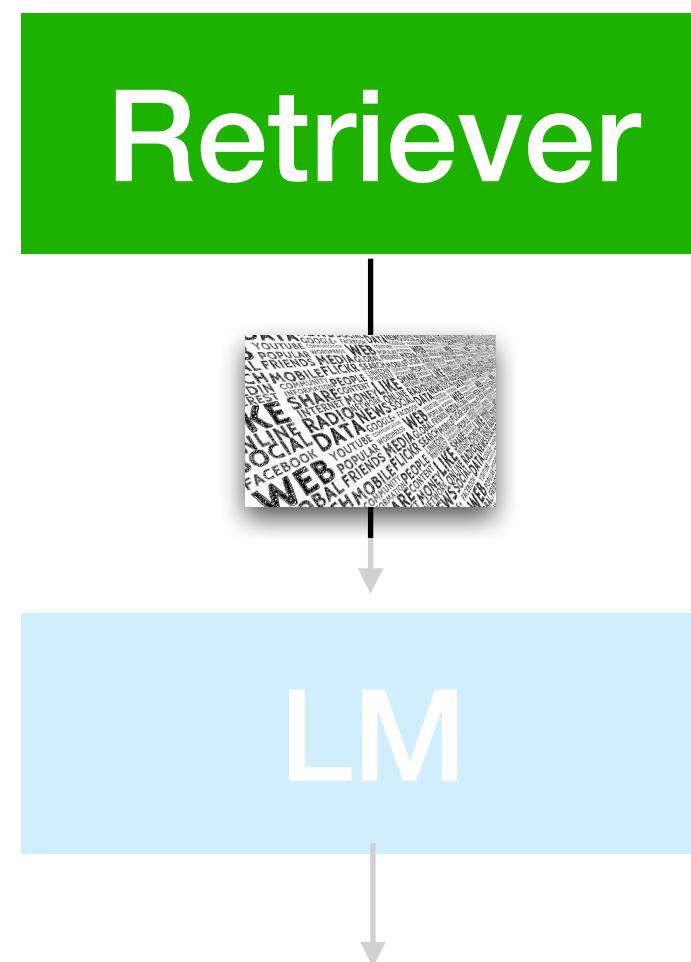


Retrieval-based prompting is easy and simple; no need to train but has higher variance



Fine-tuning (+ RL) requires training but less variance & is competitive with more data

How to adapt a retrieval-based LM for a task



Training a **retriever** on downstream tasks helps

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

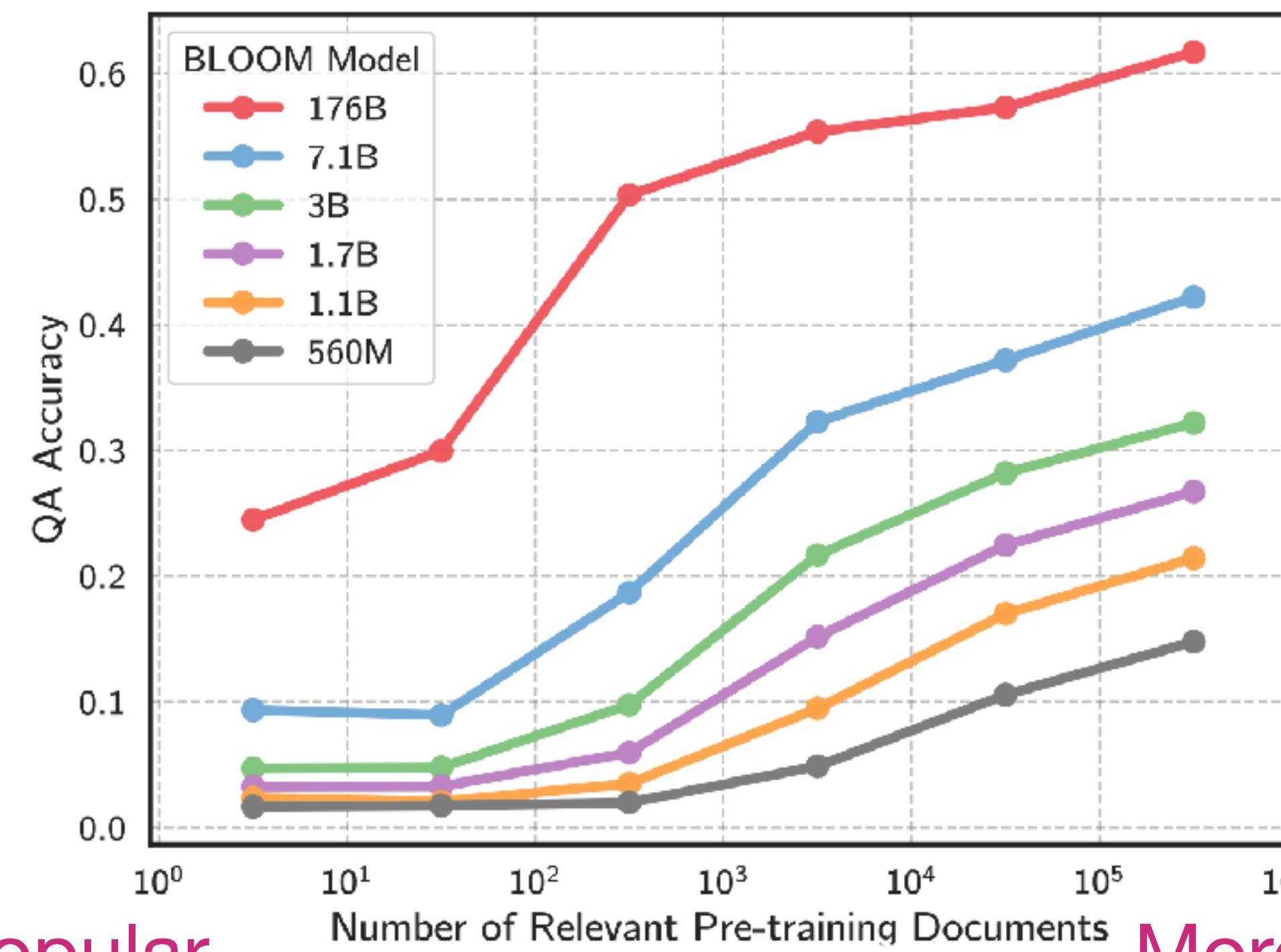
Verifiability

Parameter-
efficiency

Key effectiveness in downstream tasks

Long-tail

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



<— less popular

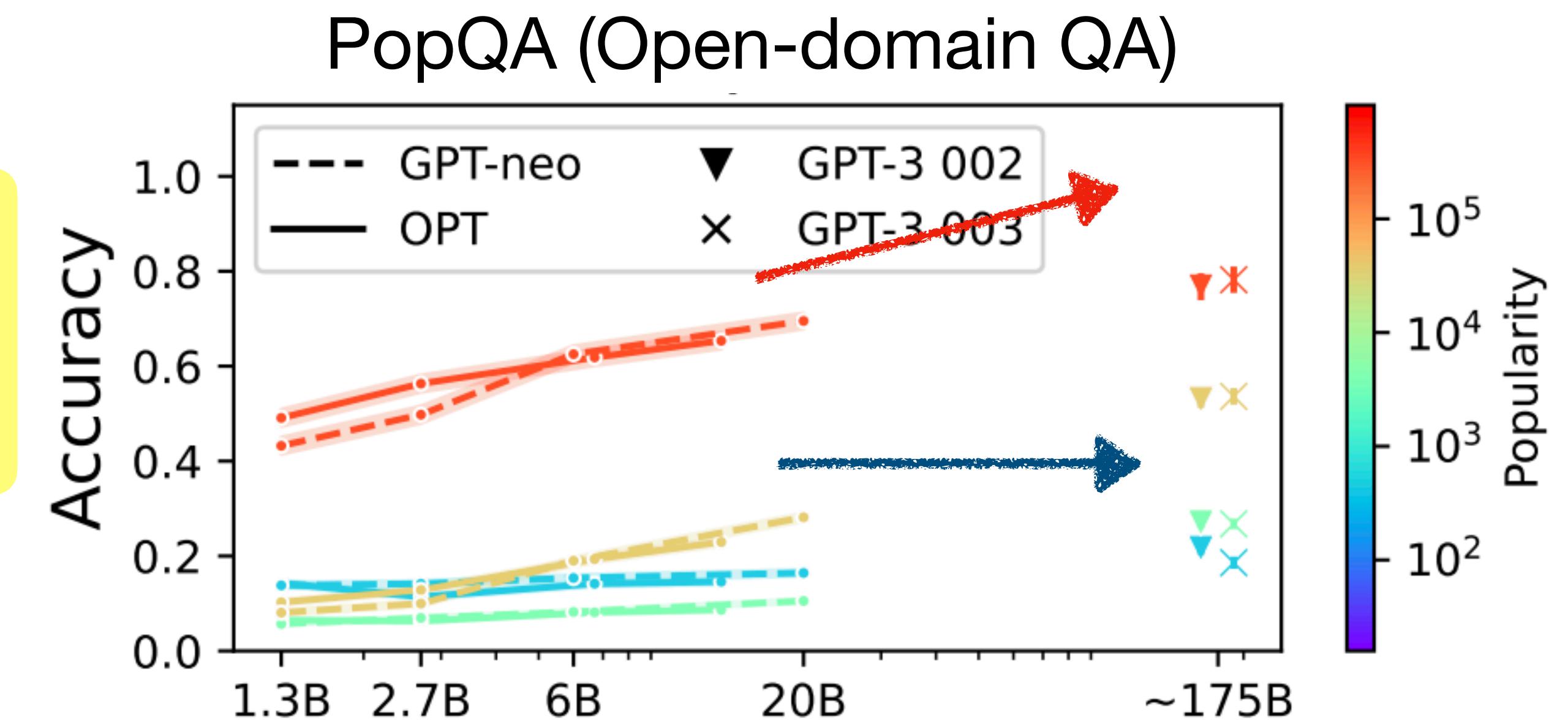
More popular —>

Key effectiveness in downstream tasks

Long-tail

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

Performance on less popular questions (blue) doesn't improve over scale

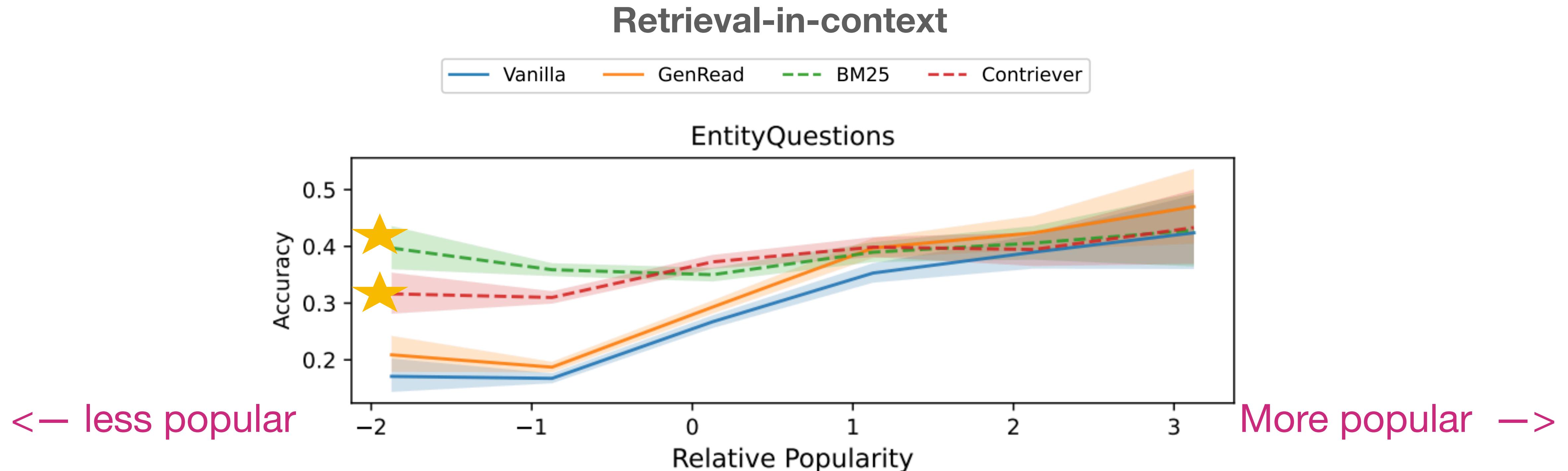


Mallen* and Asai* et al. 2023. “When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories”

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

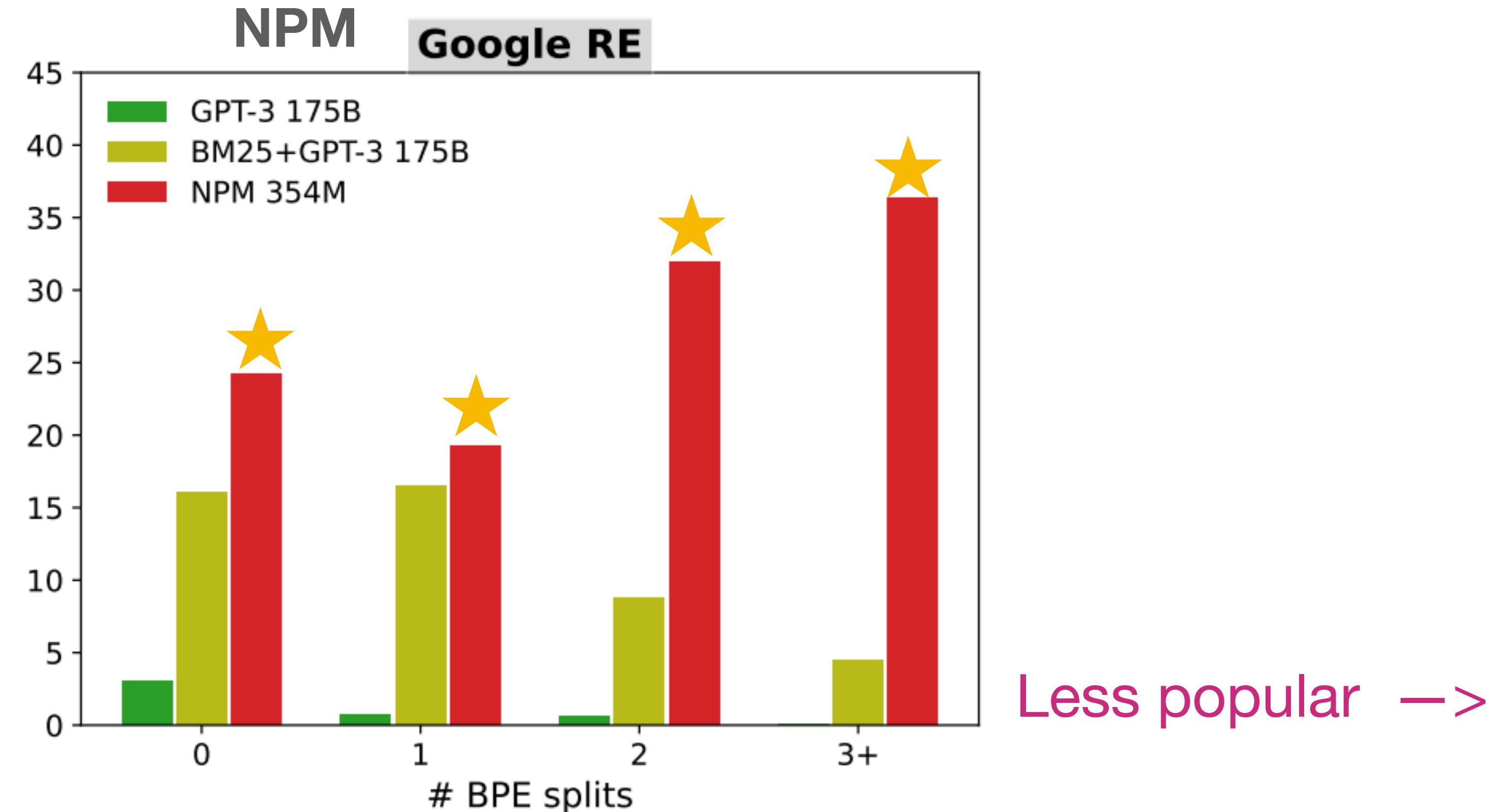
Long-tail

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

Google RE

Joshua Mathiot died in [MASK].

<— more popular

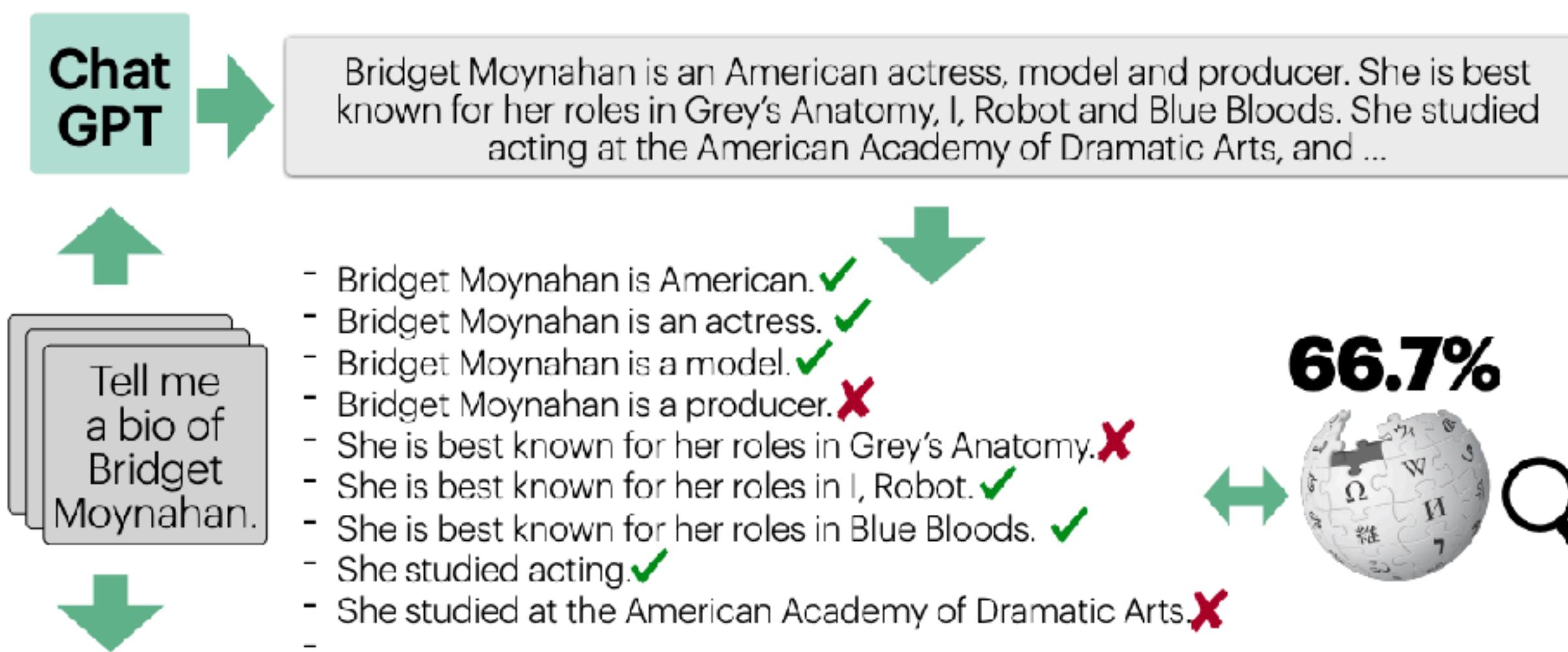


Less popular —>

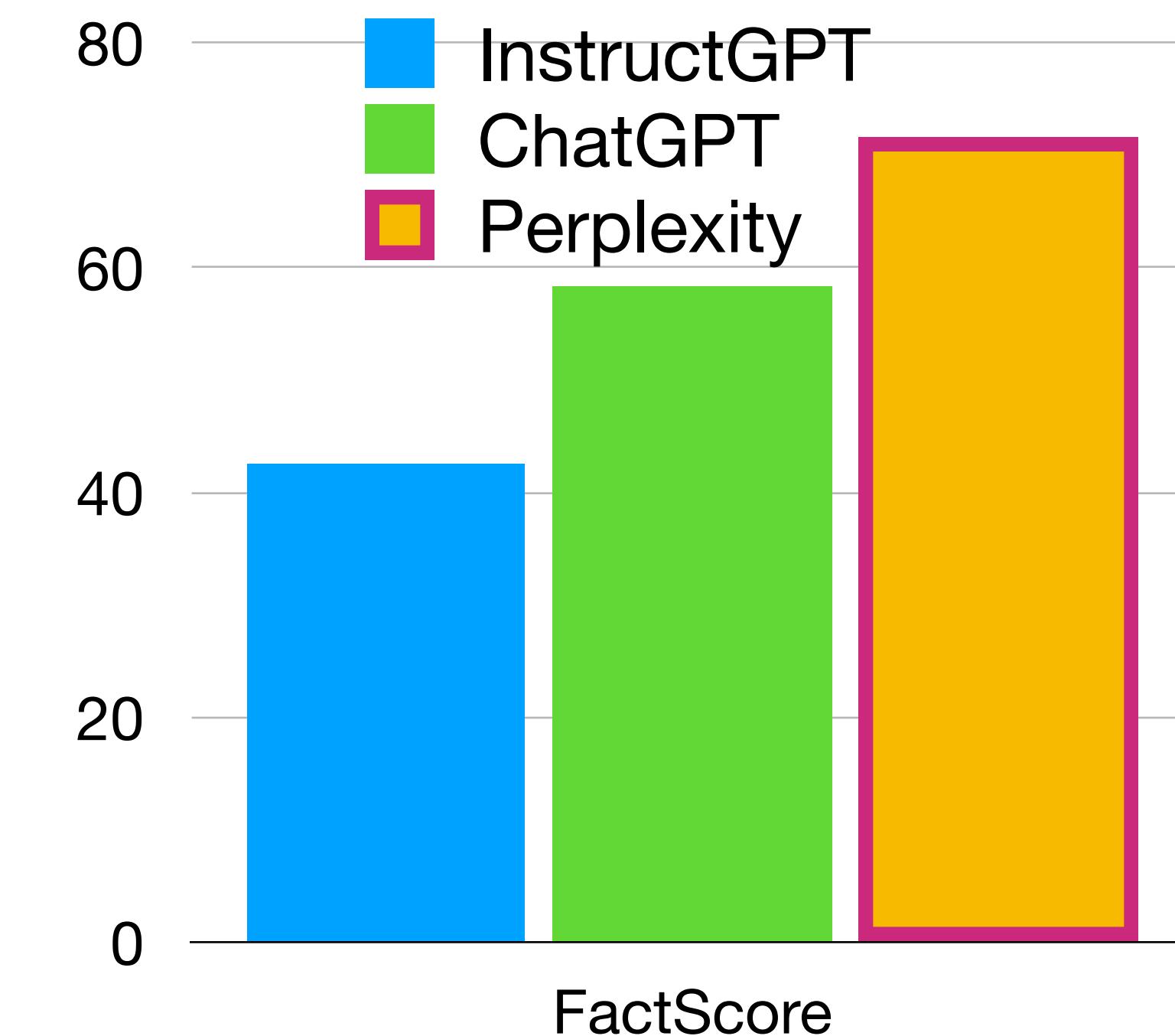
Key effectiveness in downstream tasks

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

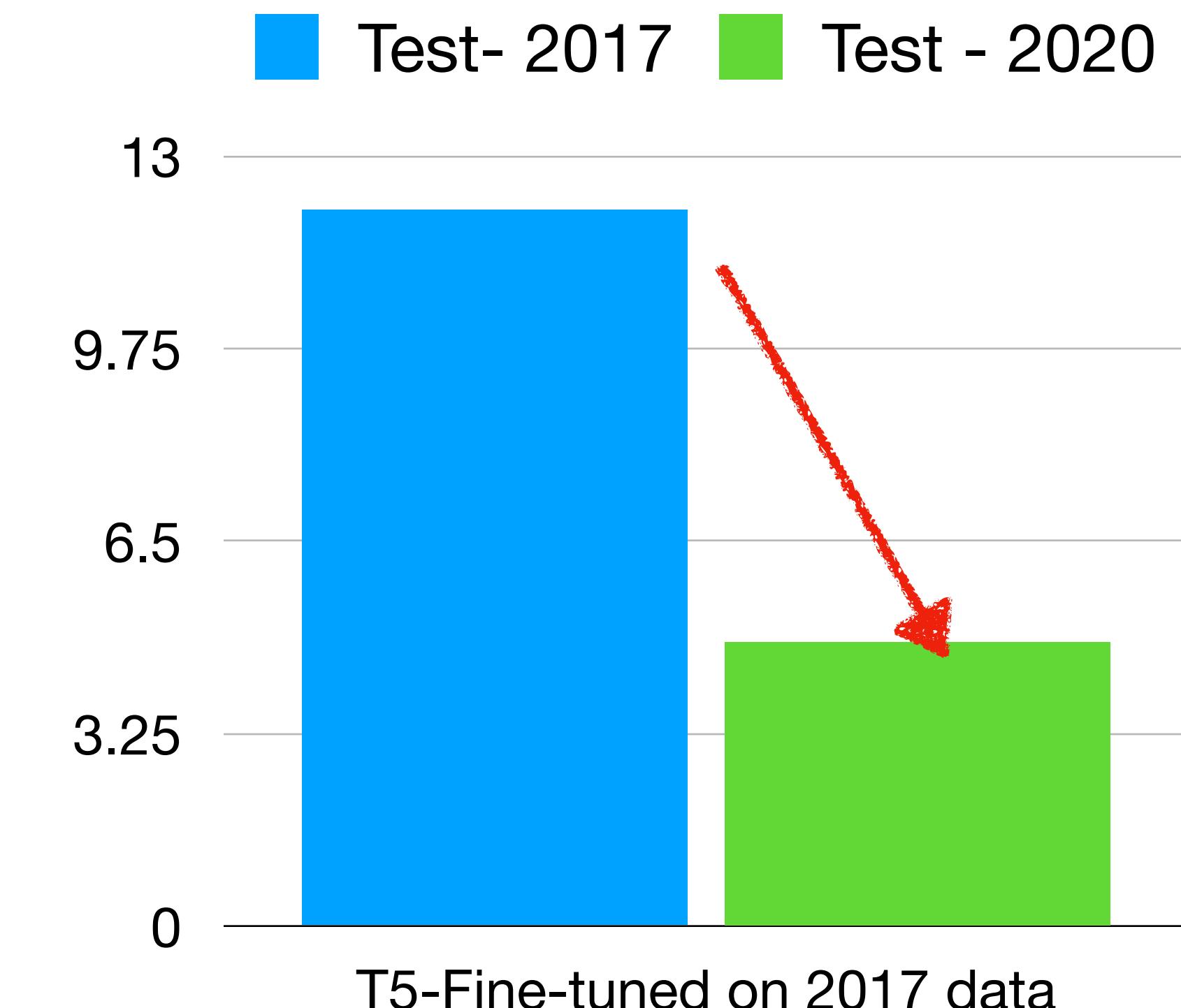
Temp LAMA

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”

Dhingra et al. 2022. “Time-Aware language models as temporal knowledge bases”



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



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

Dhingra et al. 2022. “Time-Aware language models as temporal knowledge bases”

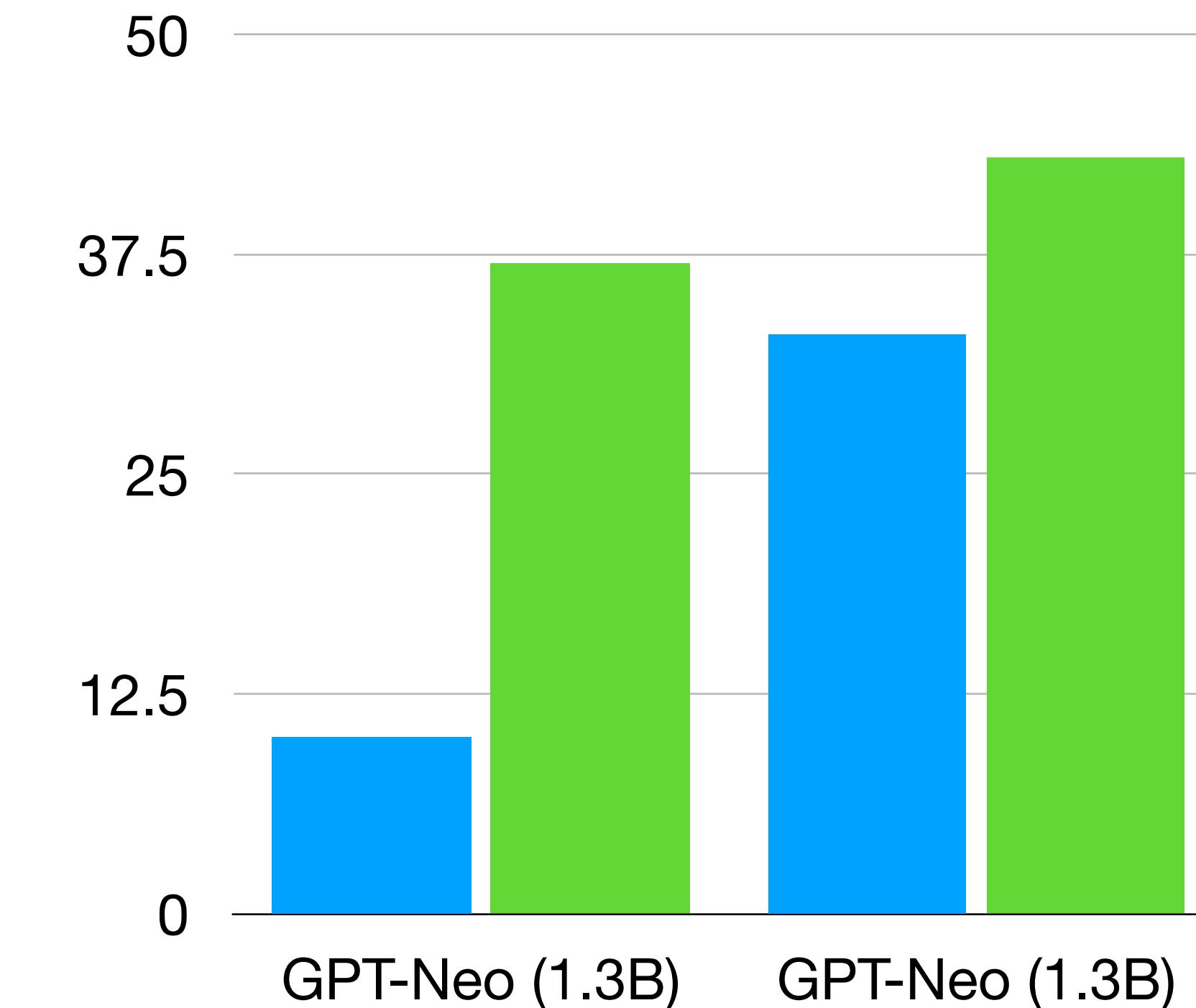
Key effectiveness in downstream tasks

Parameter-
efficiency

Much smaller LMs with retrieval can outperforms
much larger LMs in fact completions.

Retrieval + GPT-Neo 1.3B outperforms
vanilla GPT3 on PopQA

w/o retrieval
w/ Contriever



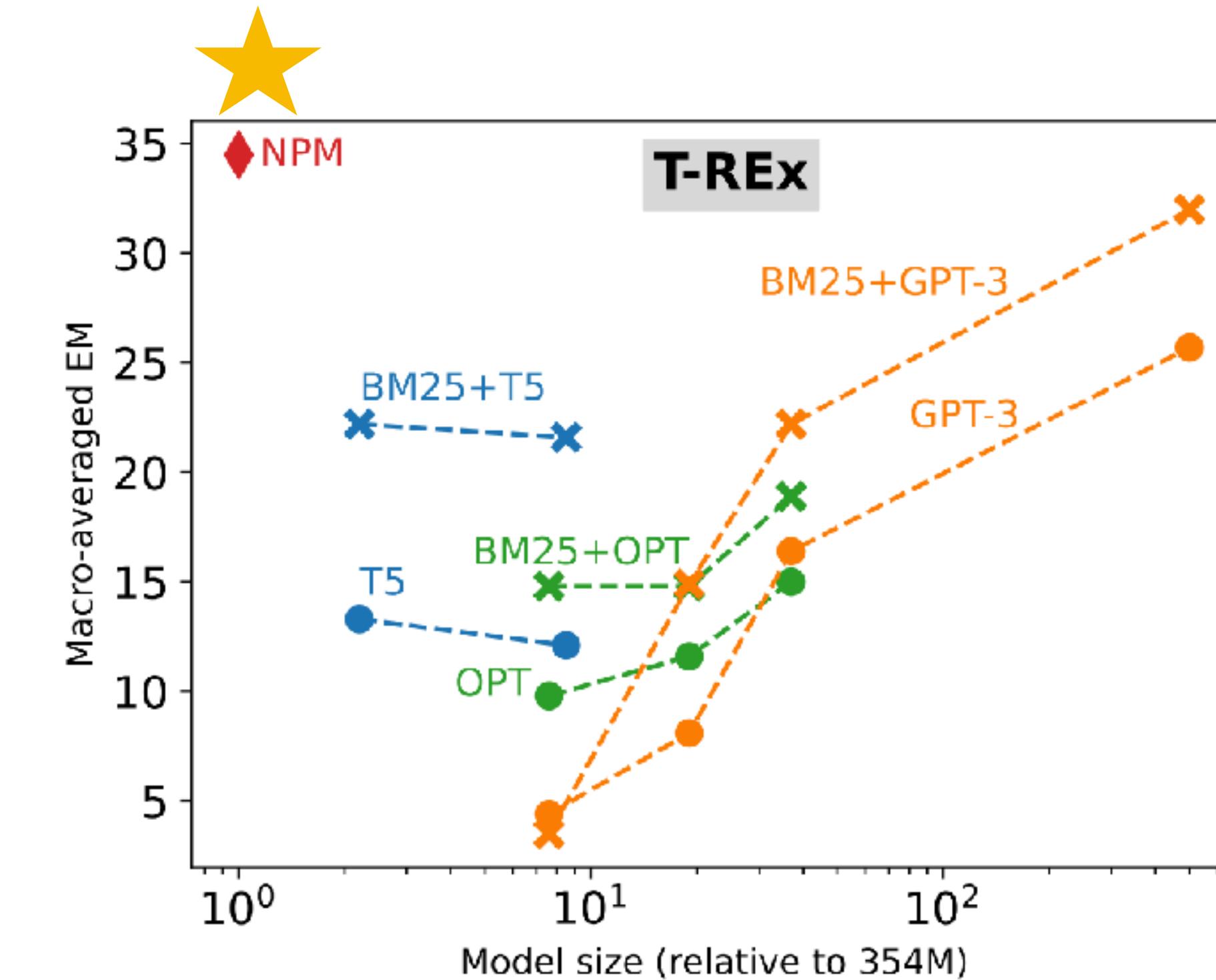
Mallen* and Asai* et al. 2023. “When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories”

Key effectiveness in downstream tasks

Parameter-
efficiency

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

Much smaller LMs with retrieval can outperforms much larger LMs in fact completions.



Min et al. 2023. “Nonparametric Masked Language Modeling”
Petroni et al. 2019. “Language Models as Knowledge Bases?”

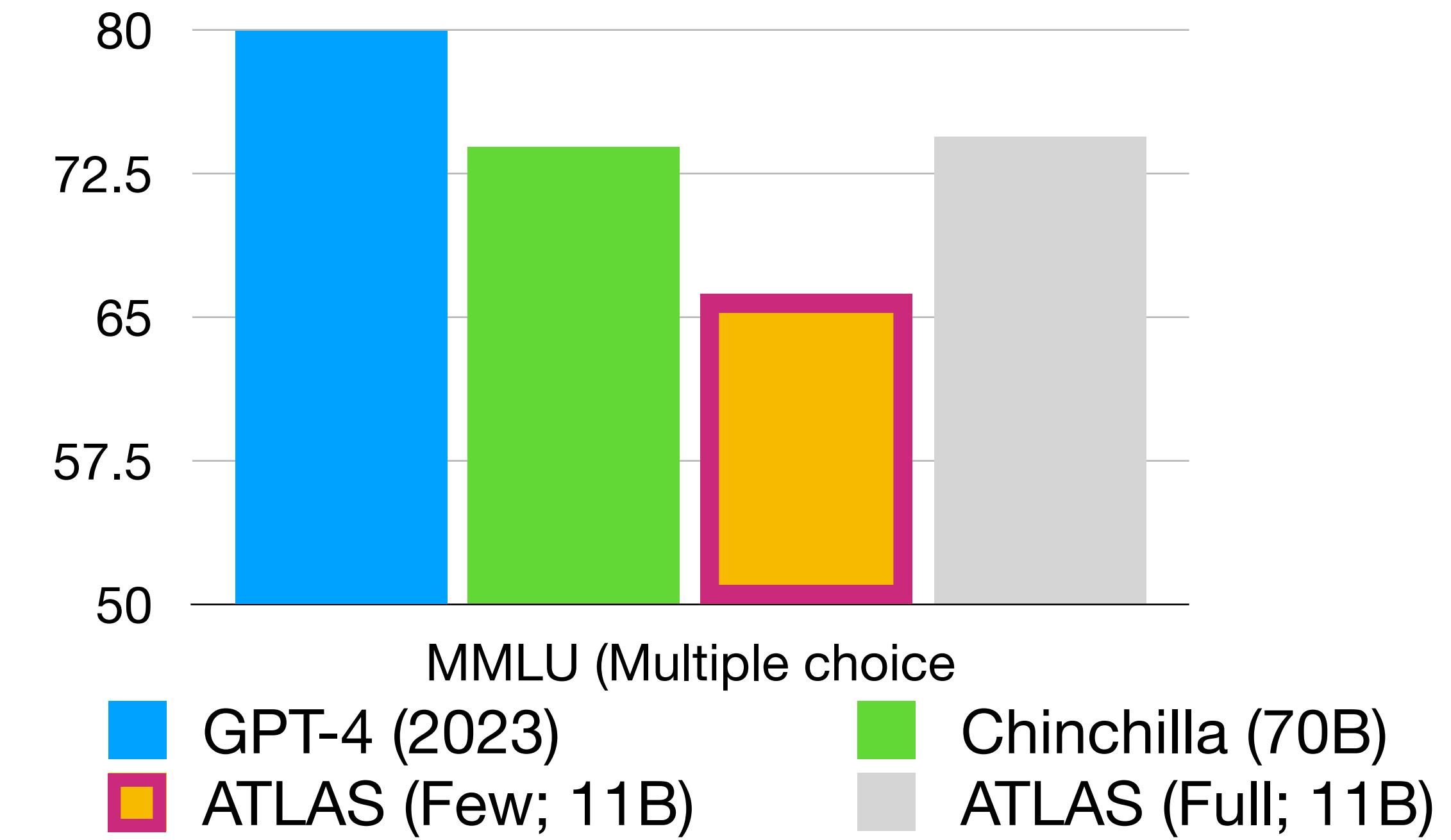
Key effectiveness in downstream tasks

Parameter-
efficiency

Much smaller LMs with retrieval can outperforms
much larger LMs in fact completions.

Room for improvements for
diverse task adaptations!

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



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?

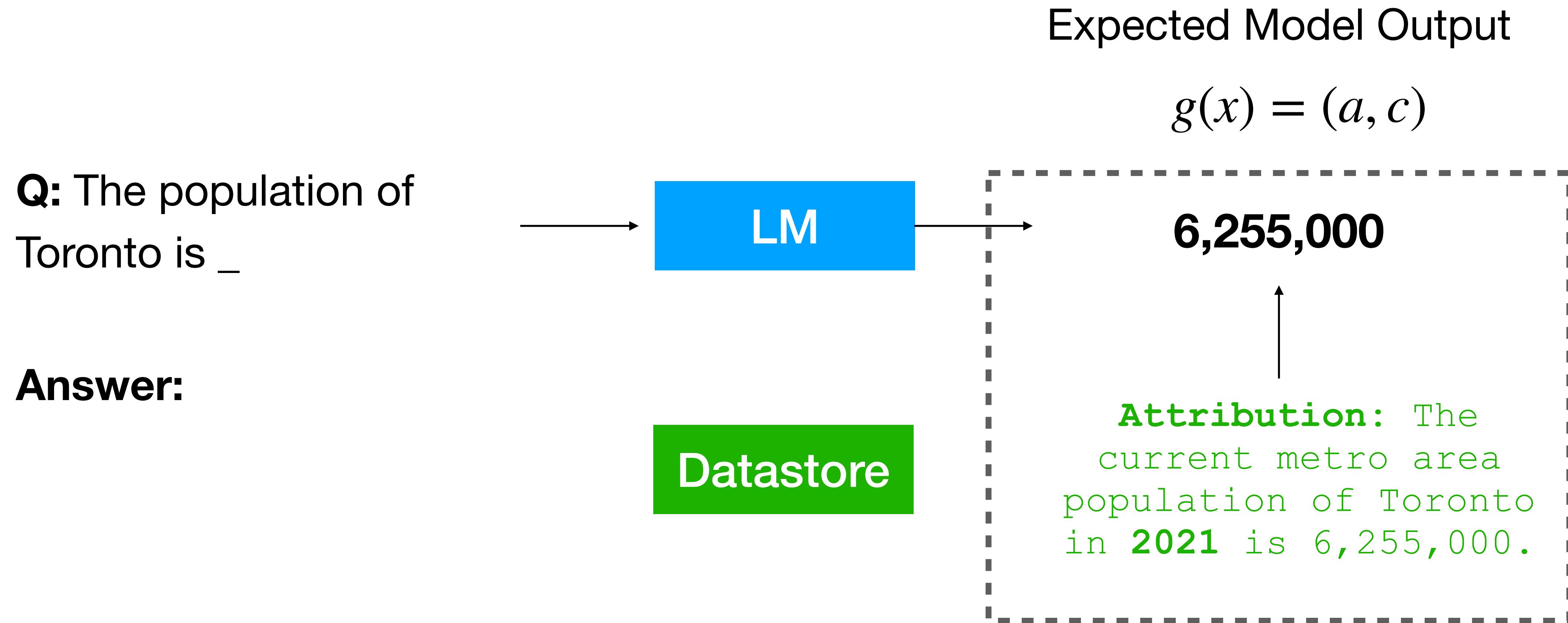
175B best-of-64

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 [1][3]. 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 [2]. Also, leptin levels can rapidly decline in the fasted state [4], 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 \(www.diabetes.co.uk\)](#) ↗

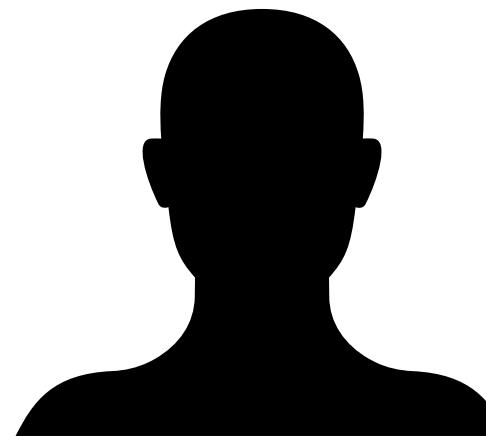
Nakano et al. 2021. “WebGPT: Browser-assisted question-answering with human feedback”

Attributions: AttributedQA (Bohnet et al., 2022)

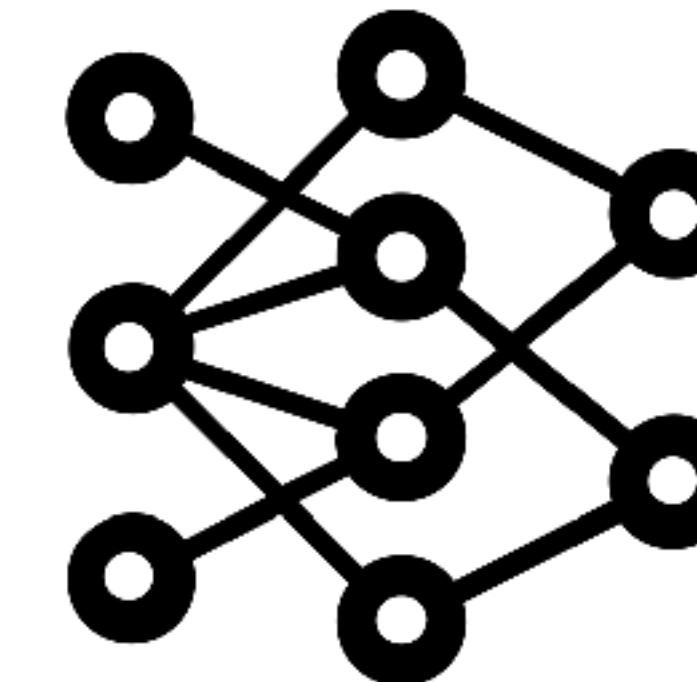


Attributions: AttributedQA (Bohnet et al., 2022)

Human Evaluation (AIS)



Automatic Evaluation (AutoAIS)



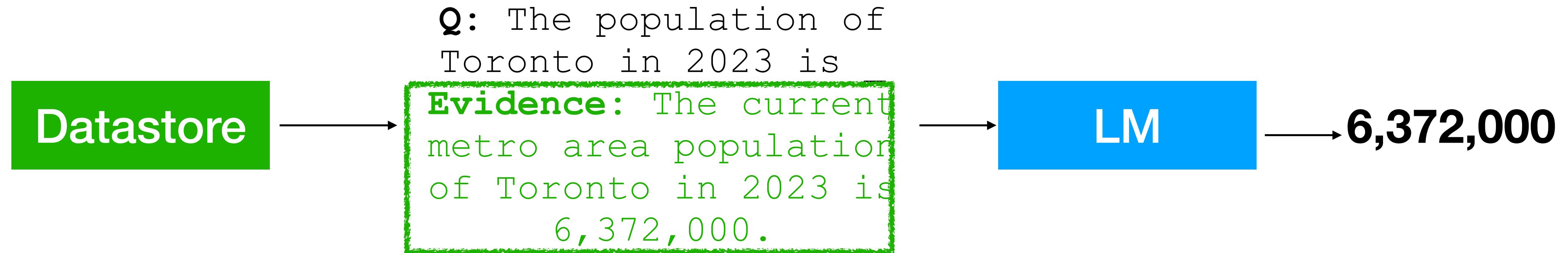
NLI model

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)

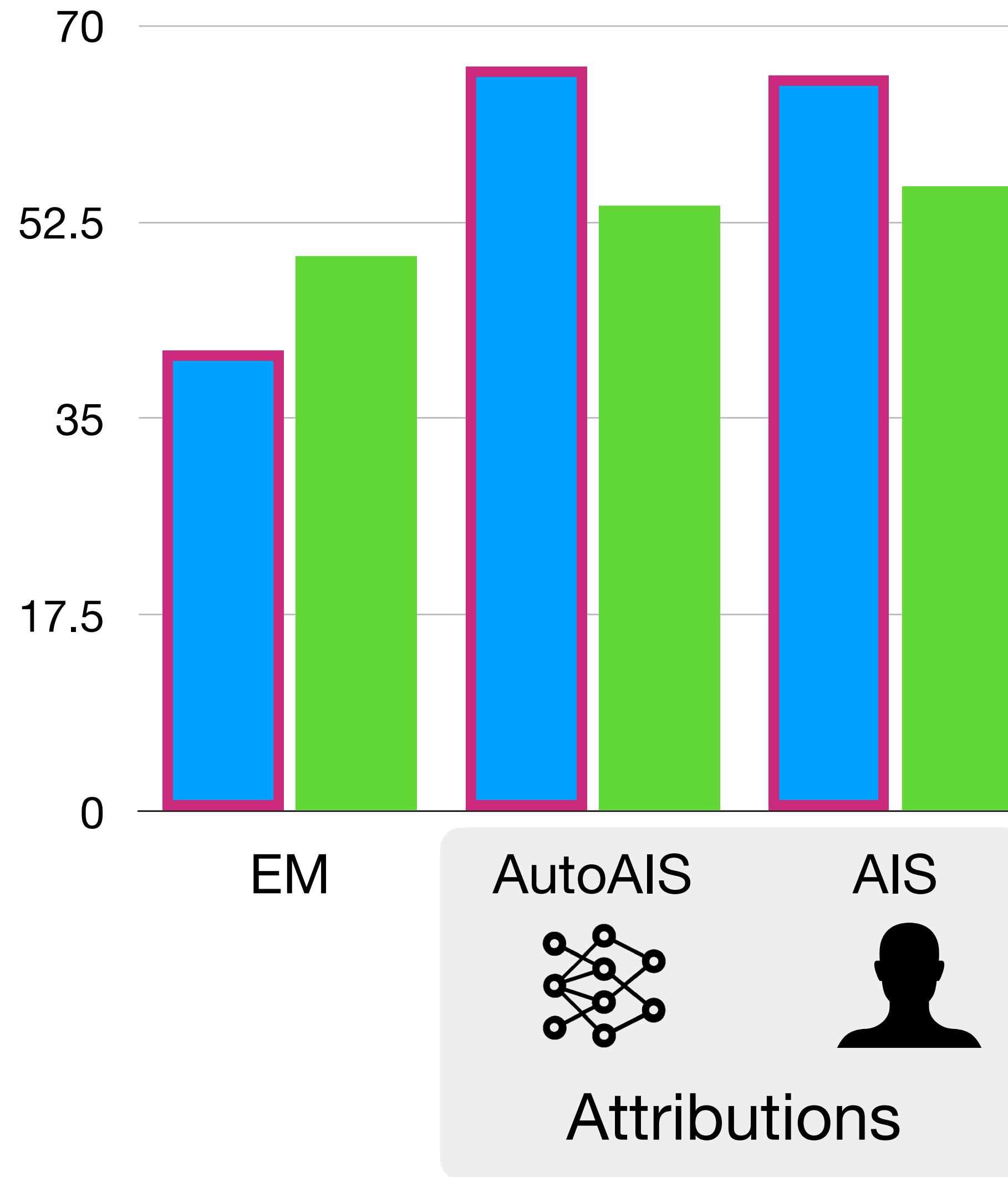
Retrieval-based LM



Post-hoc retrieval



AttributedQA (Bohnet et al., 2022)



Retrieval in context yields higher AIS than post-hoc retrieval

■ Retrieval-based LM ■ Post-hoc retrieval

When to use a retrieval-based LM

Long-tail

knowledge
update

Verifiability

Parameter-
efficiency

When to use a retrieval-based LM

Long-tail

knowledge
update

Verifiability

Parameter-
efficiency

Out of domain adaptations

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

Khandelwal, et al. 2020. “Nearest Neighbor Zero-shot Inference”

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

When to use a retrieval-based LM

Long-tail

knowledge
update

Verifiability

Parameter-
efficiency

Out of domain adaptations

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

Privacy

(Huang et al., 2023)

Khandelwal, et al. 2020. “Nearest Neighbor Zero-shot Inference”

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

Huang et al. 2023. “Nearest Neighbor Zero-shot Inference”