Section 5:
Applications

Downstream adaptation of retrieval-based LMs

What are the **tasks**?

The capital city of Ontario is __

Toronto, which is known for …

Downstream adaptation of retrieval-based LMs

What are the **tasks**?

Q: What is the capital of Ontario?

A: Toronto

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

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
- General NLU
- Language Modeling & other generation tasks

…

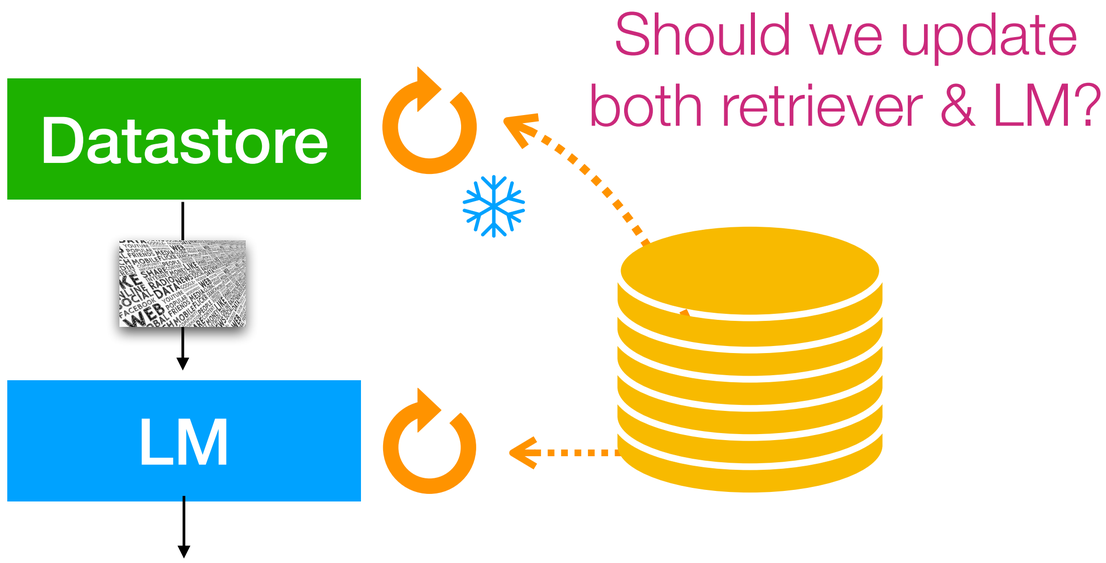
How to **adapt?**

- Supervised fine-tuning
- Reinforcement learning
- Prompting

How to adapt a retrieval-based LM for a task

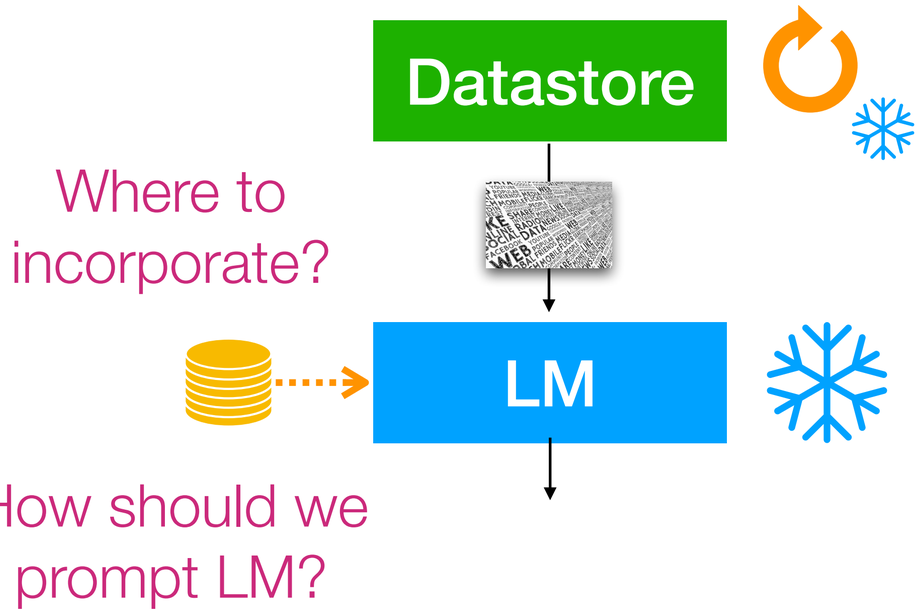
Fine-tuning (+RL)

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
- General NLU
- Language Modeling & other generation tasks

…

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

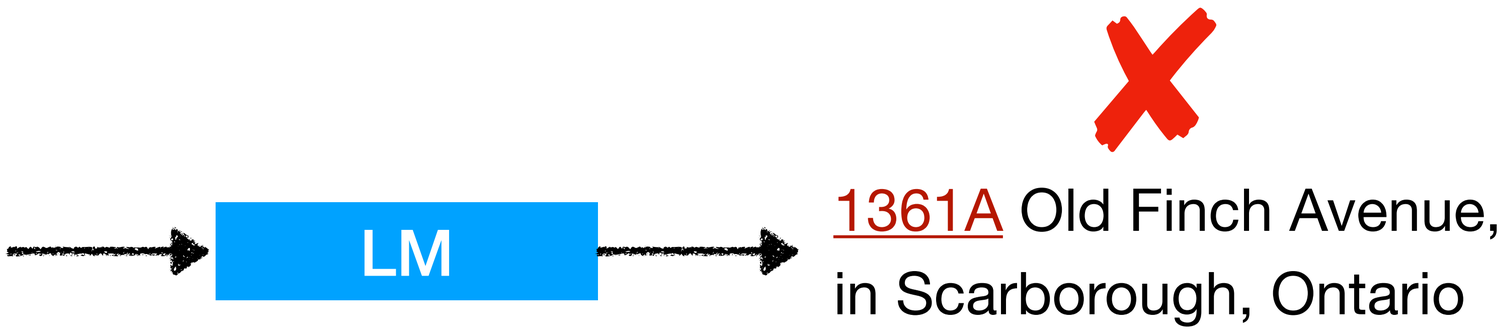
Long-tail | knowledge update | Verifiability | Parameter-efficiency | Privacy

Q: Is Toronto really cold during winter?

Yes it is.

Effectiveness of retrieval-based LMs

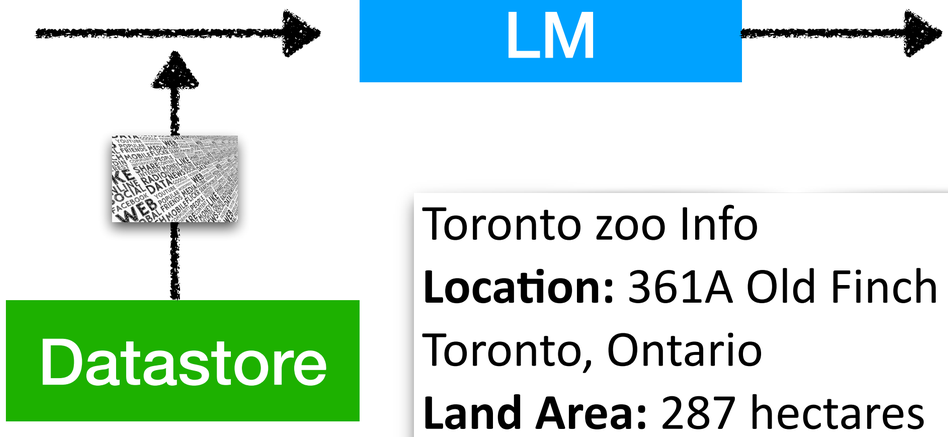
Long-tail | knowledge update | Verifiability | Parameter-efficiency | Privacy



Effectiveness of retrieval-based LMs

Long-tail | knowledge update | Verifiability | Parameter-efficiency | Privacy

Q: Where is Toronto Zoo located?



361A Old Finch Avenue, in Scarborough, Ontario

Effectiveness of retrieval-based LMs

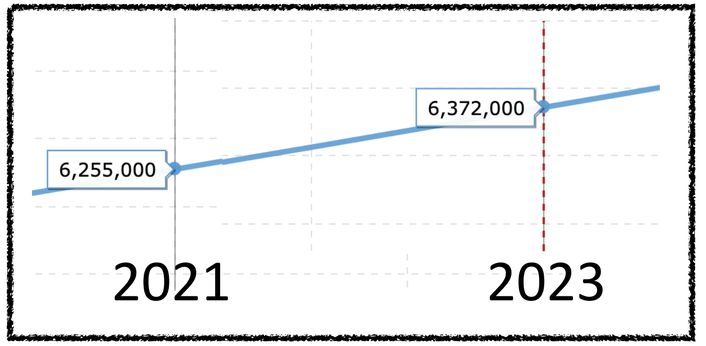
Long-tail | **Knowledge update** | Verifiability | Parameter-efficiency | Privacy

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

→ LM →

A: 6,255,000

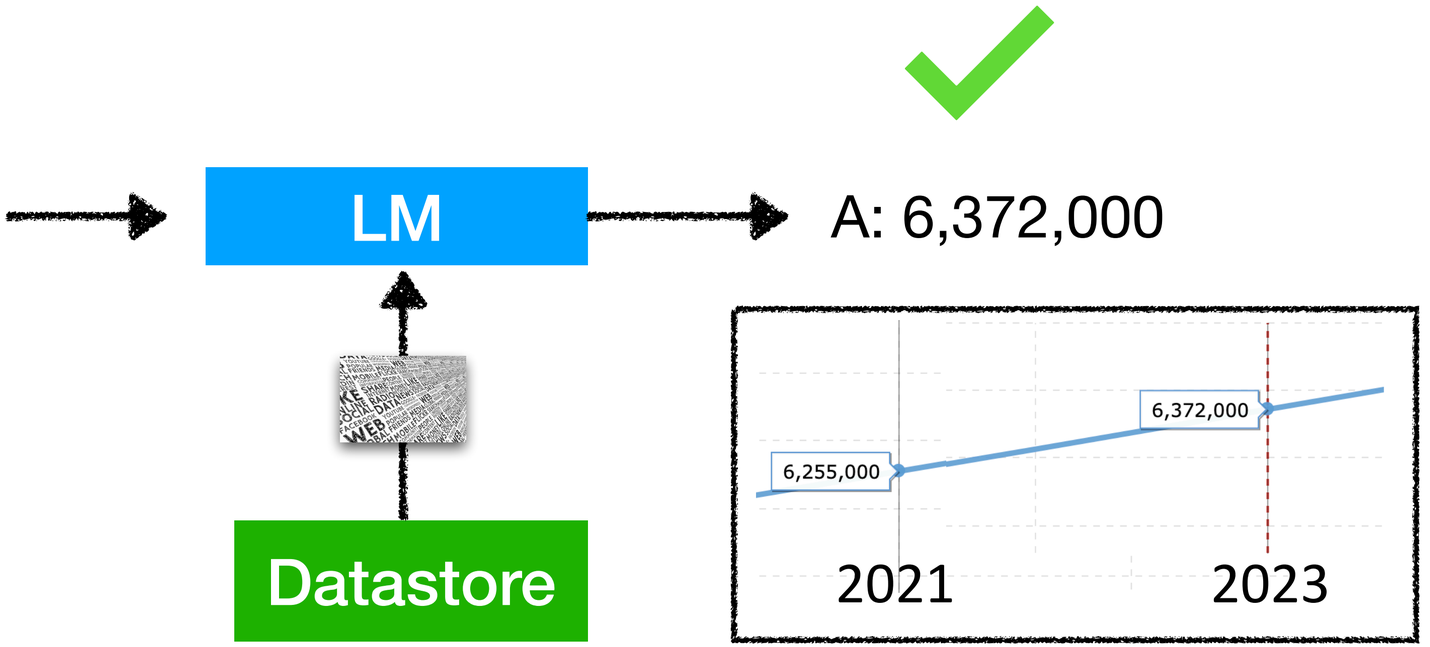
Trained in **2021** corpus



Effectiveness of retrieval-based LMs

Long-tail | **Knowledge update** | Verifiability | Parameter-efficiency | Privacy

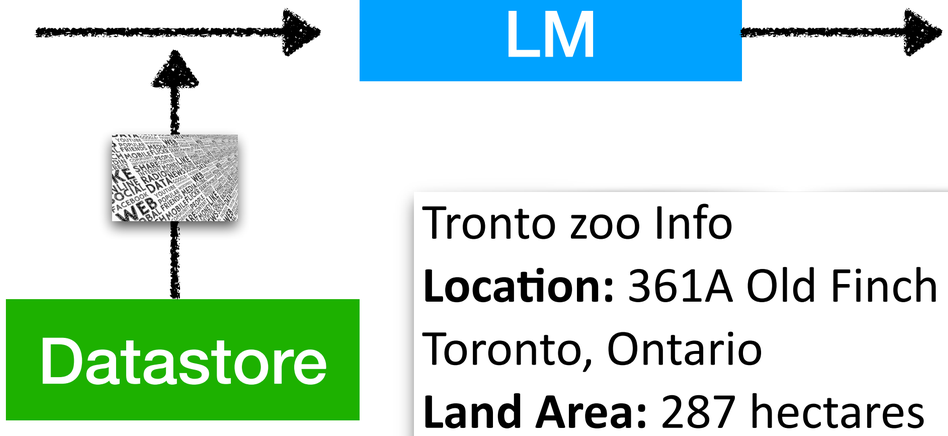
Q: What is the population of Toronto Metropolitan area in 2023?



Effectiveness of retrieval-based LMs

Long-tail | knowledge update | **Verifiability** | Parameter-efficiency | Privacy

Q: Where is Toronto Zoo located?

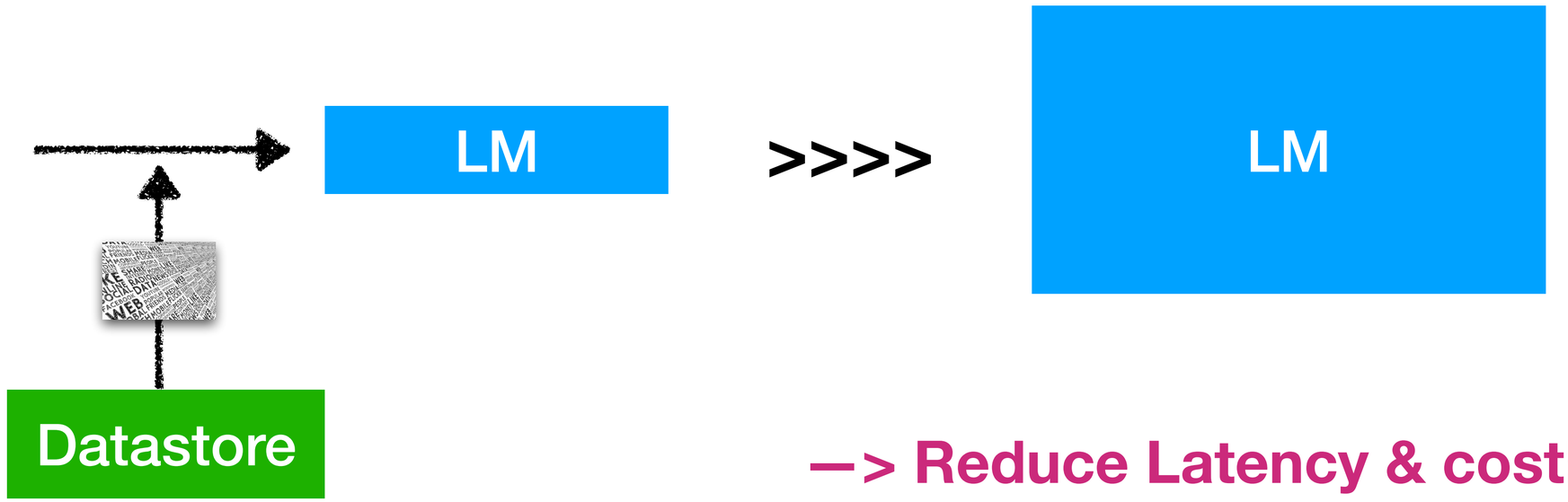


Effectiveness of retrieval-based LMs

Long-tail | knowledge update | Verifiability | Parameter-efficiency | Privacy

Effectiveness of retrieval-based LMs

Long-tail | knowledge update | Verifiability | **Parameter-efficiency** | Privacy



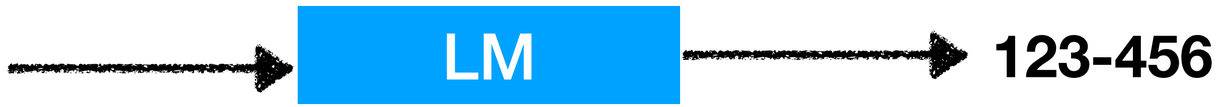
—> Reduce Latency & cost!!

Effectiveness of retrieval-based LMs

Long-tail | knowledge update | Verifiability | Parameter-efficiency | **Privacy**

Email: mail@alice.com

Phone: 404-



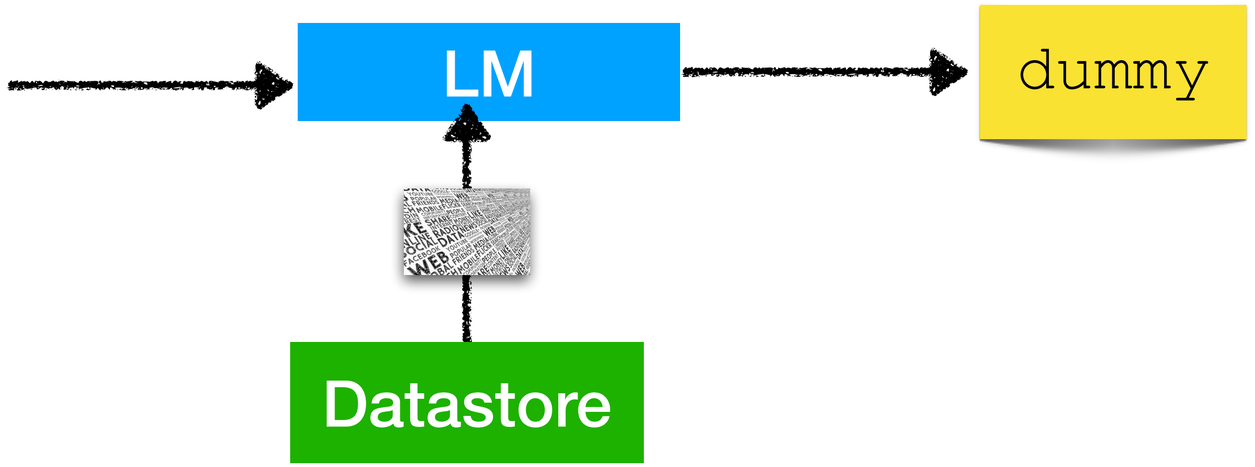
LM can leak private information included in pertaining corpora

Effectiveness of retrieval-based LMs

Long-tail | knowledge update | Verifiability | Parameter-efficiency | **Privacy**

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

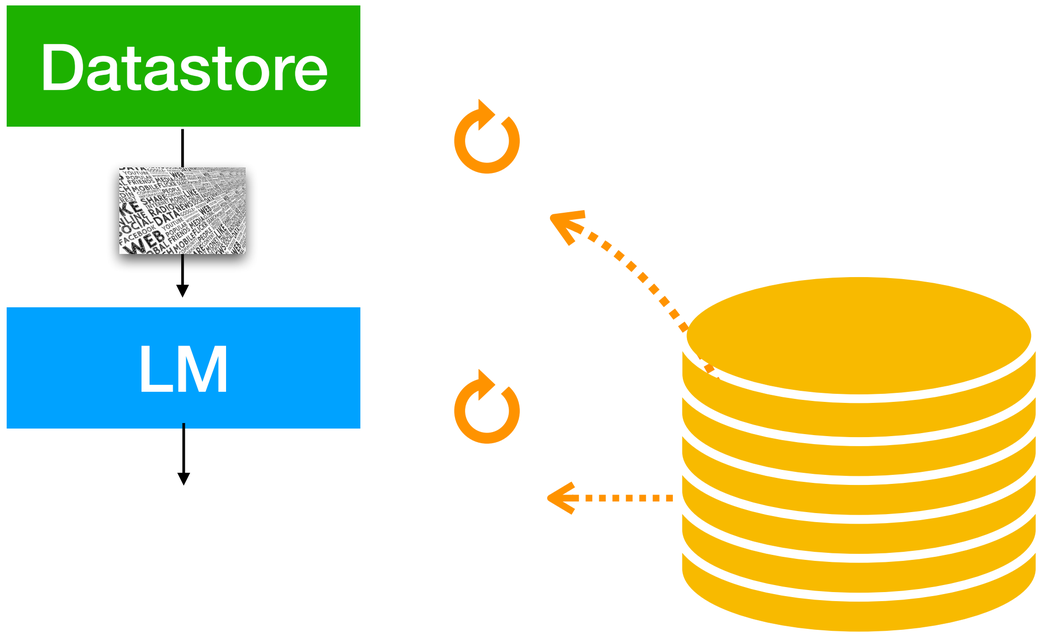
What is **data store?**

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

Adapting retrieval-based LMs for tasks

Fine-tuning

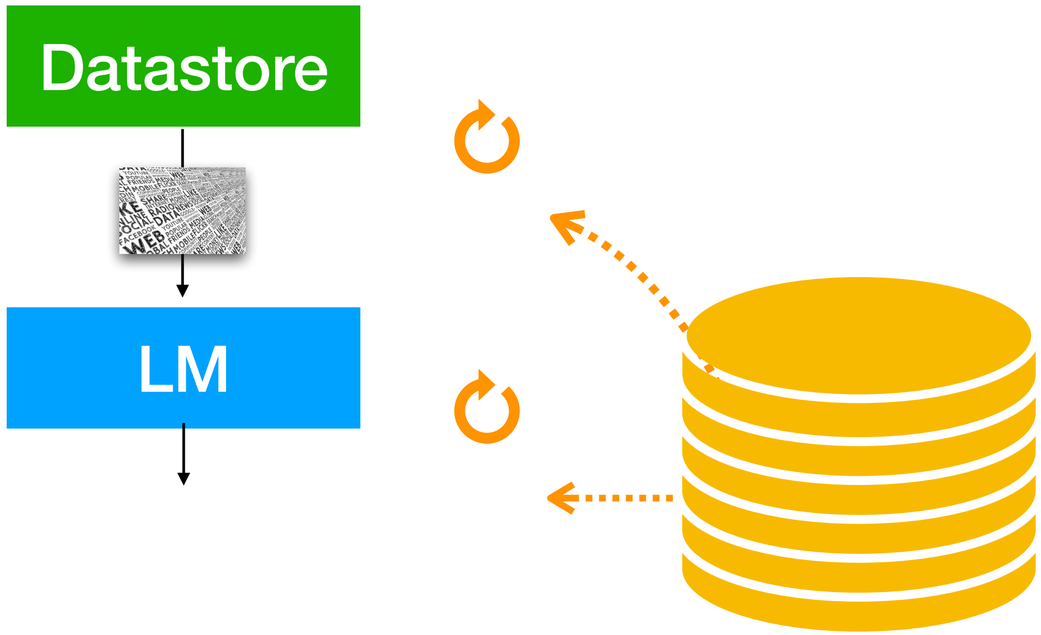
Training LM and / or retriever
on task-data & data store



Adapting retrieval-based LMs for tasks

Fine-tuning

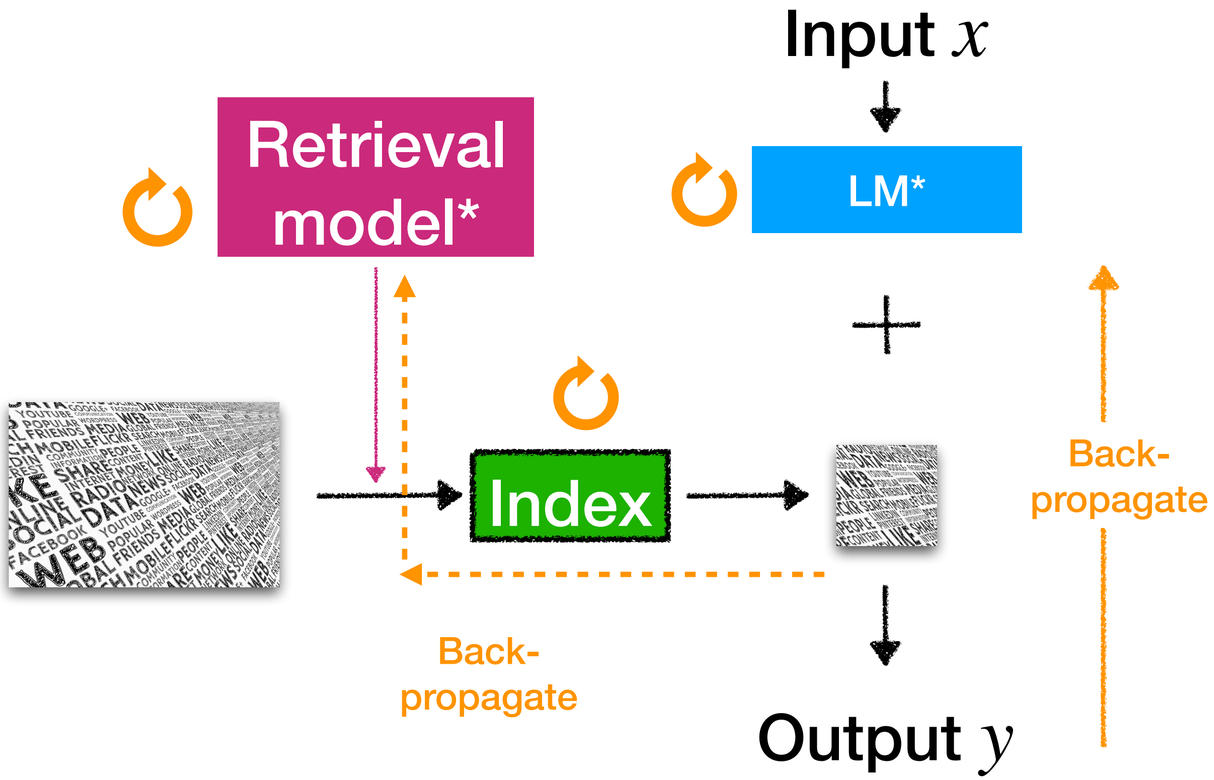
Training LM and / or retriever
on task-data & data store



Costs of retrieval-based LM training (Section 4)

Asynchronous updates (REALM)
Independent training (DPR)
…

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



Few-shot task data

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

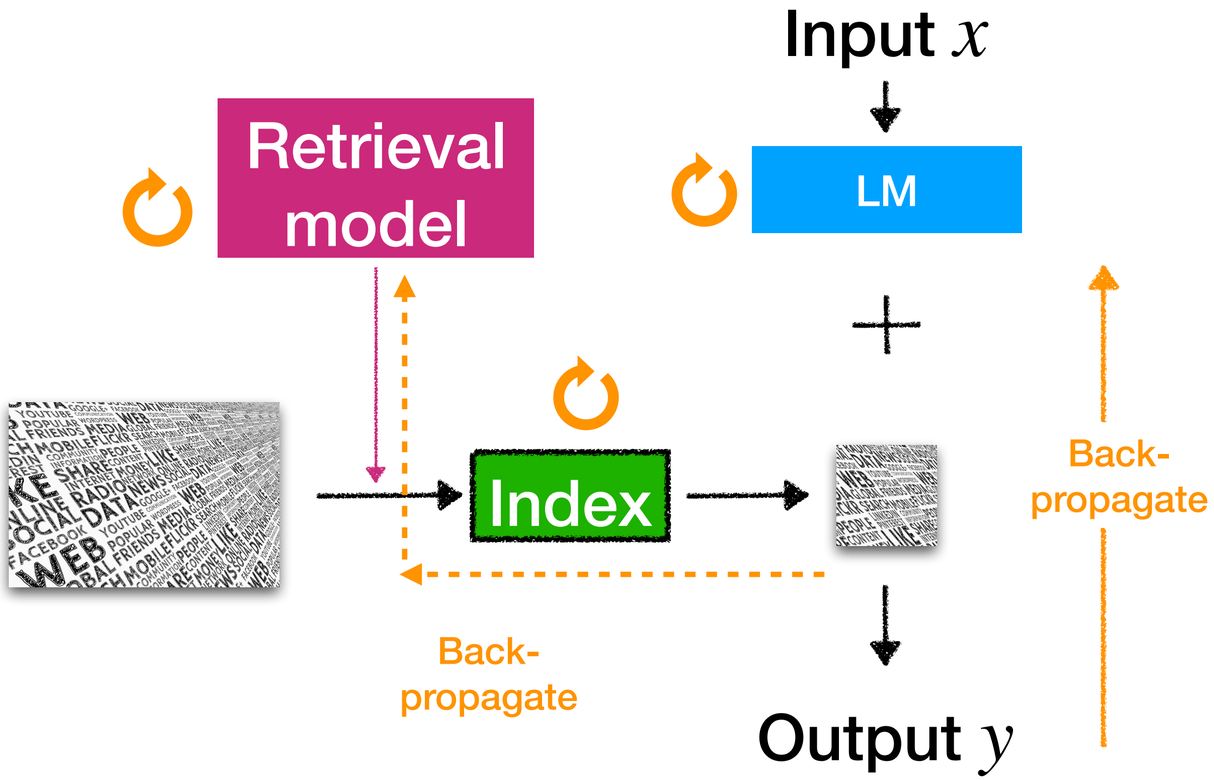
Fact Verification (x_i^{FV}, y_i^{FV})

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

*Pre-trained

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

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



Full fine-tuning w/ index update is expensive!

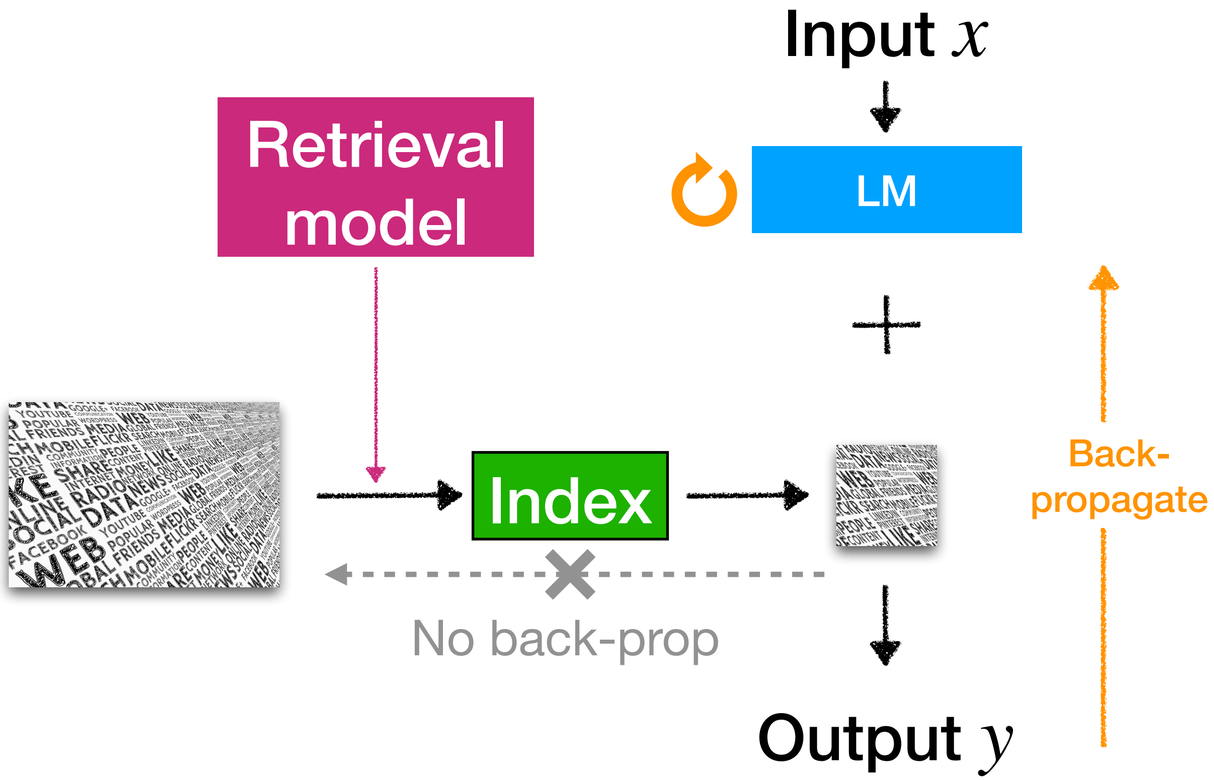
Few-shot task data

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



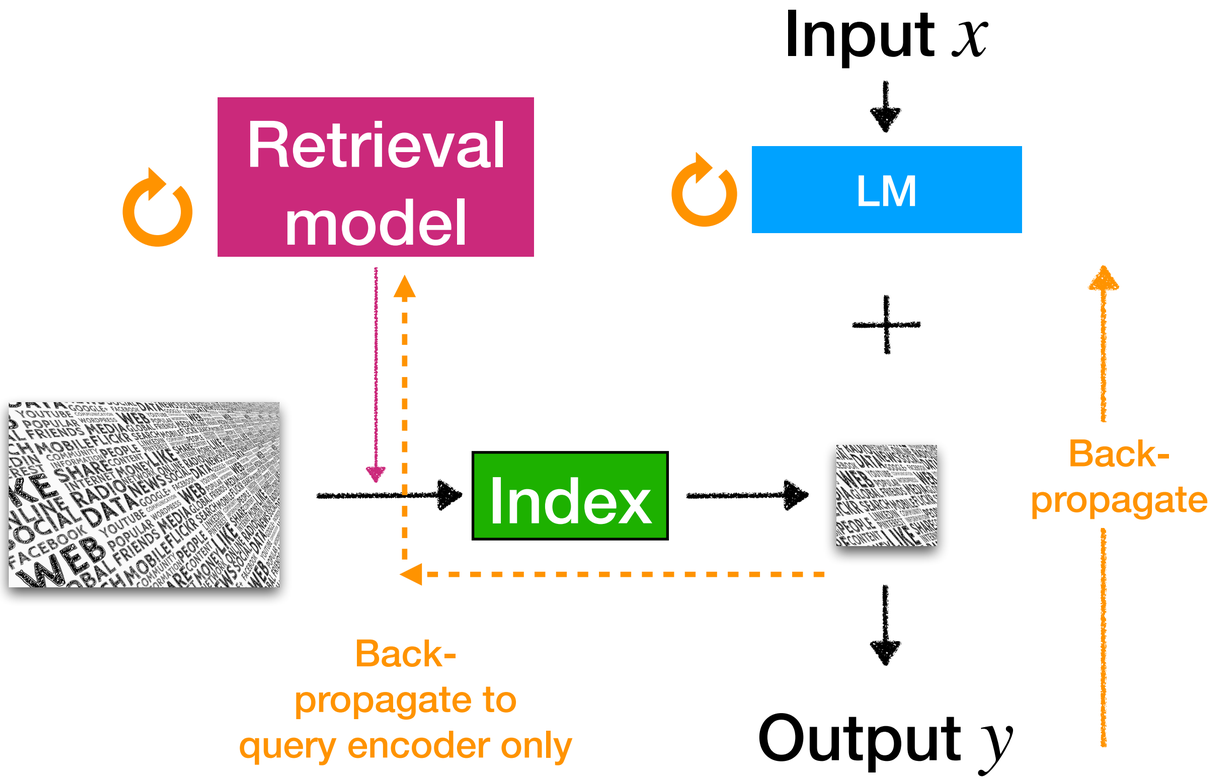
Few-shot task data

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

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

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

ATLAS: query-encoder only updates



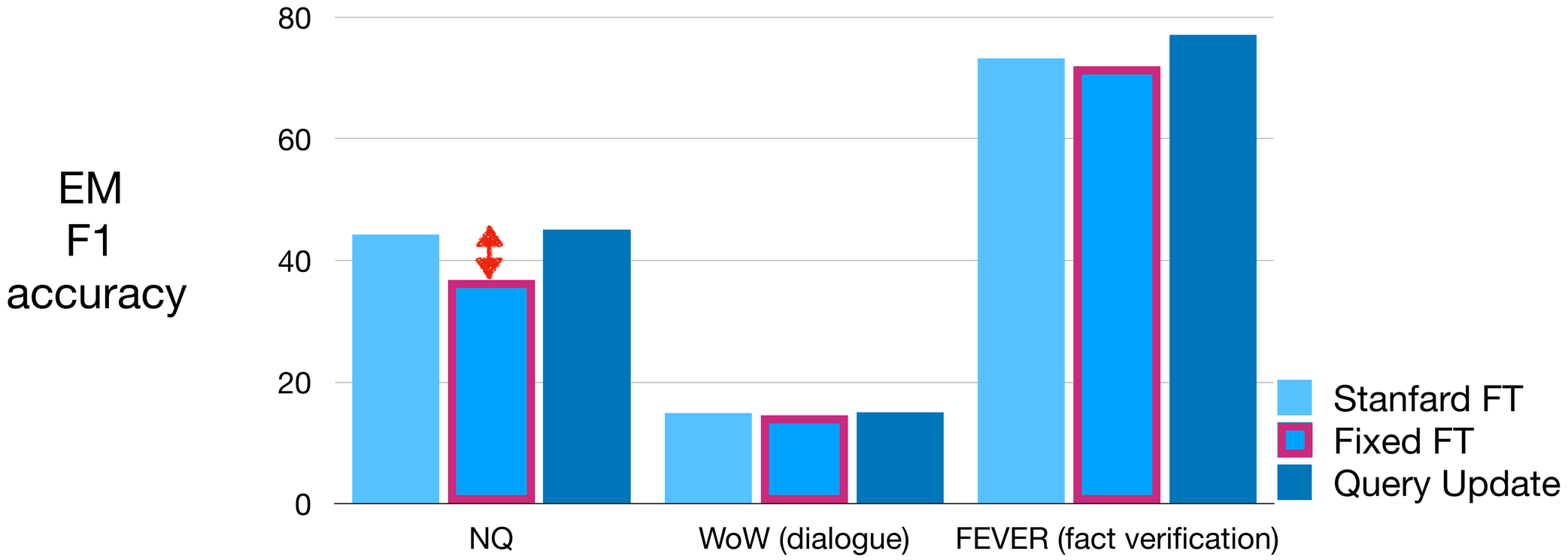
Few-shot task data

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

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

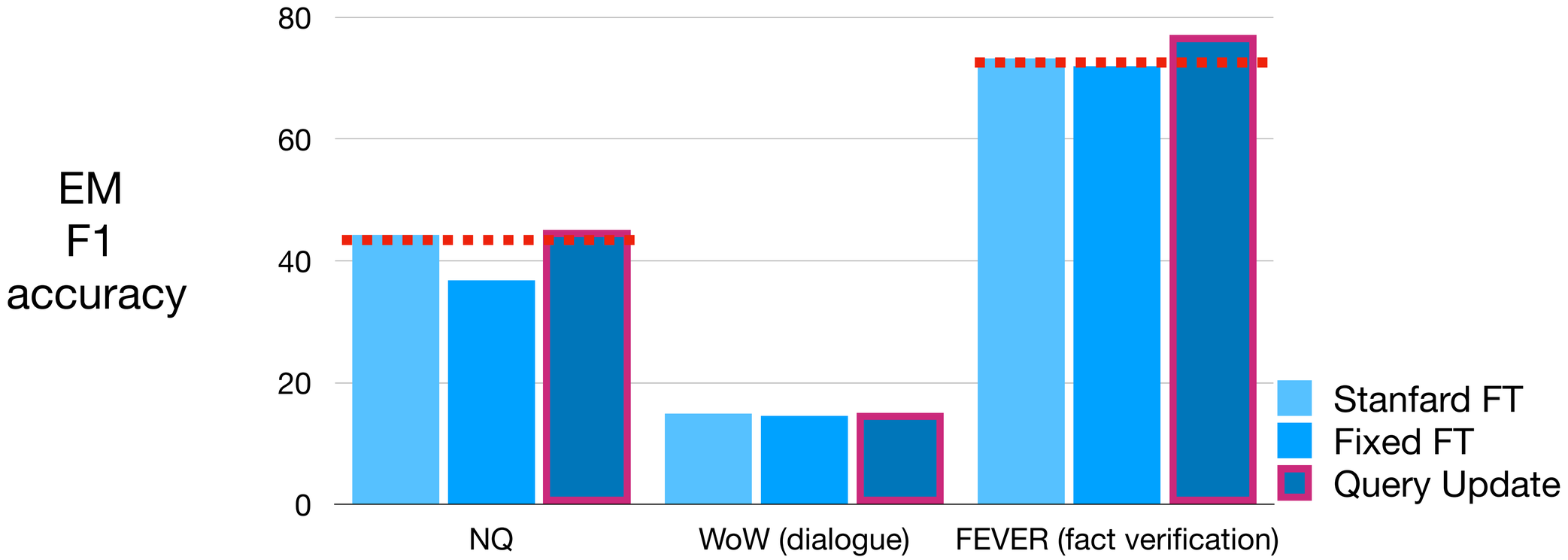
dialogue (x_i^{dial}, y_i^{dial})

Ablations of efficient retrieval training



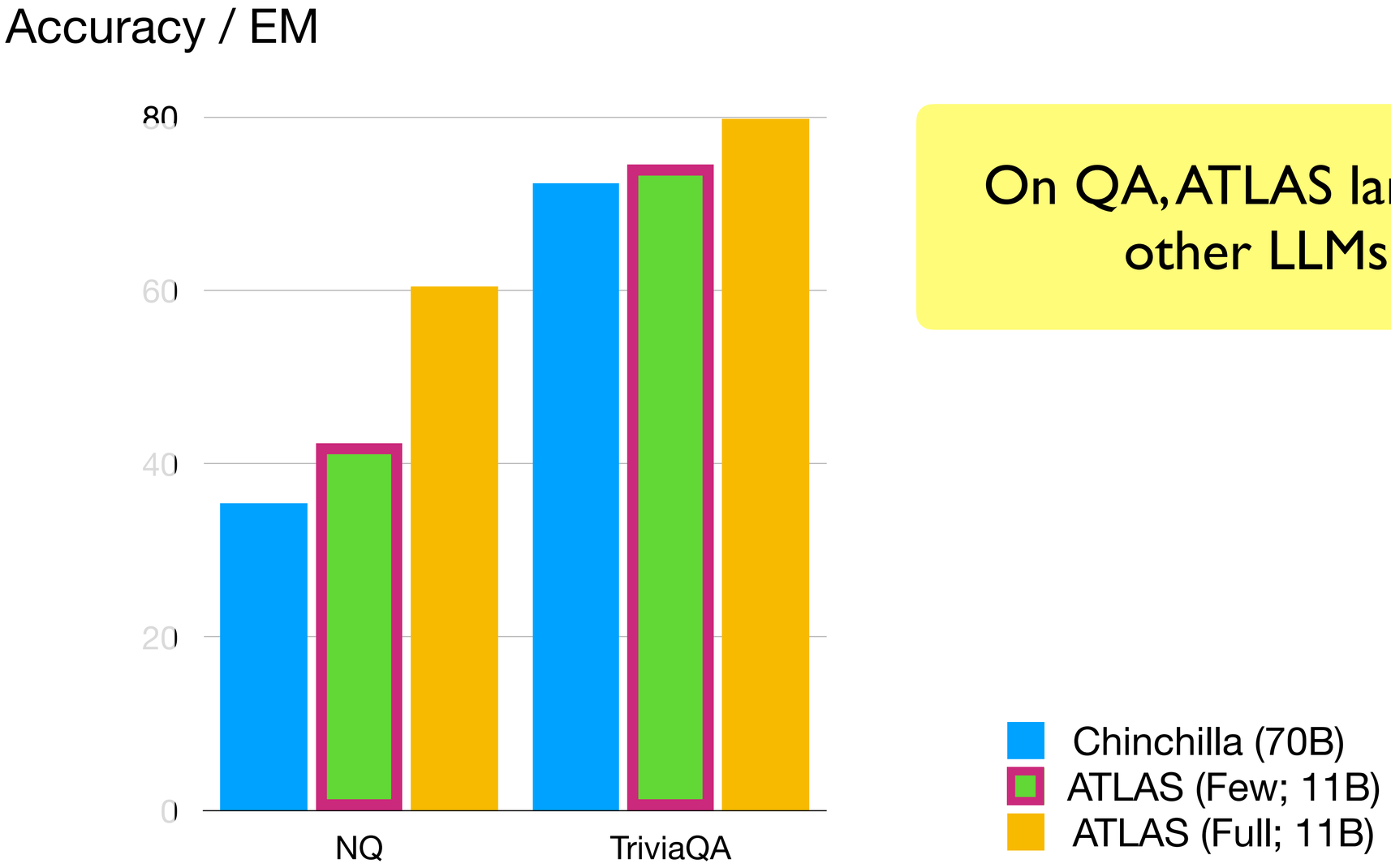
Fixed FT shows large performance drop on QA.

Ablations of efficient retrieval training



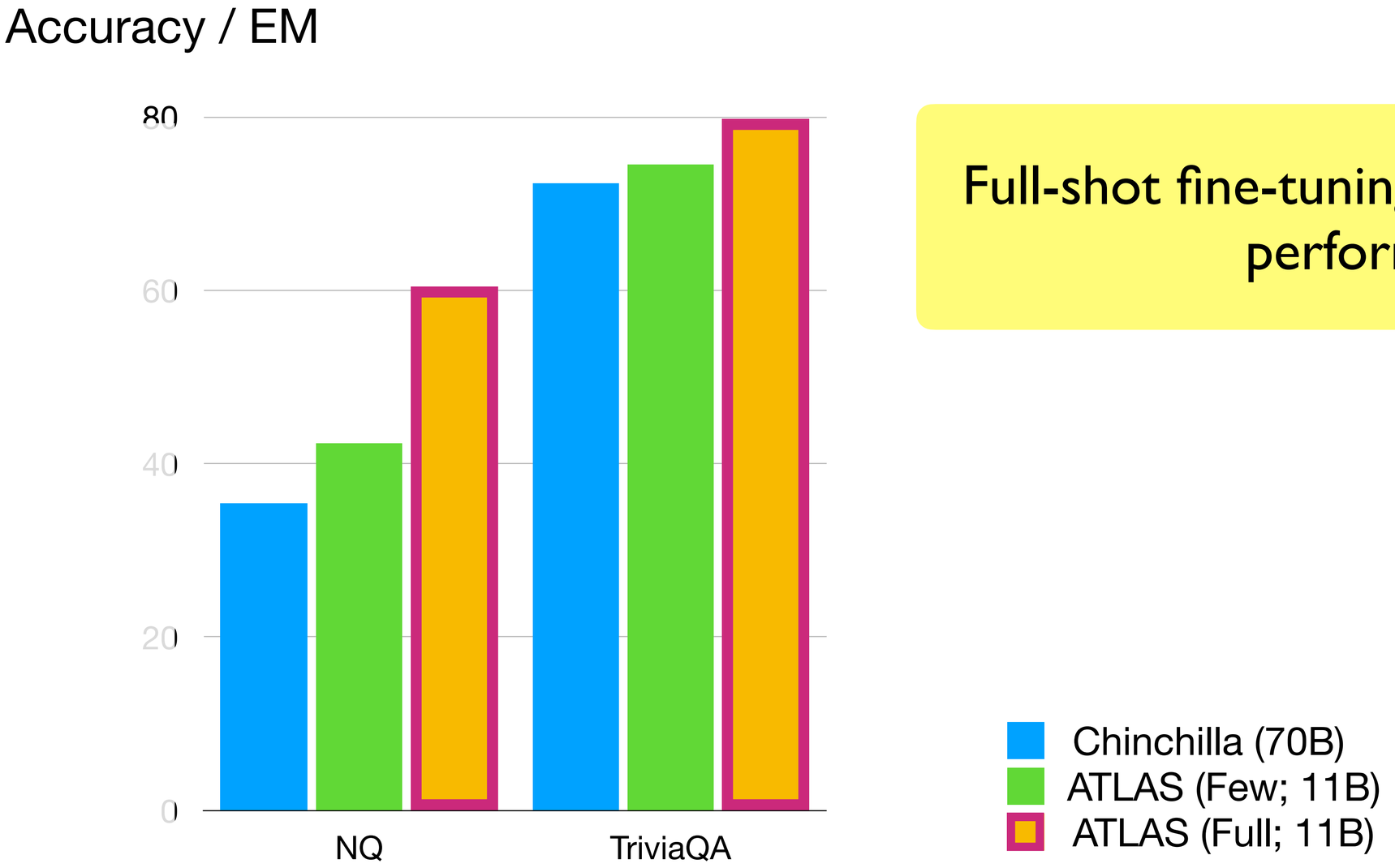
Query-side fine-tuning match or outperforms full fine-tuning

Task Results



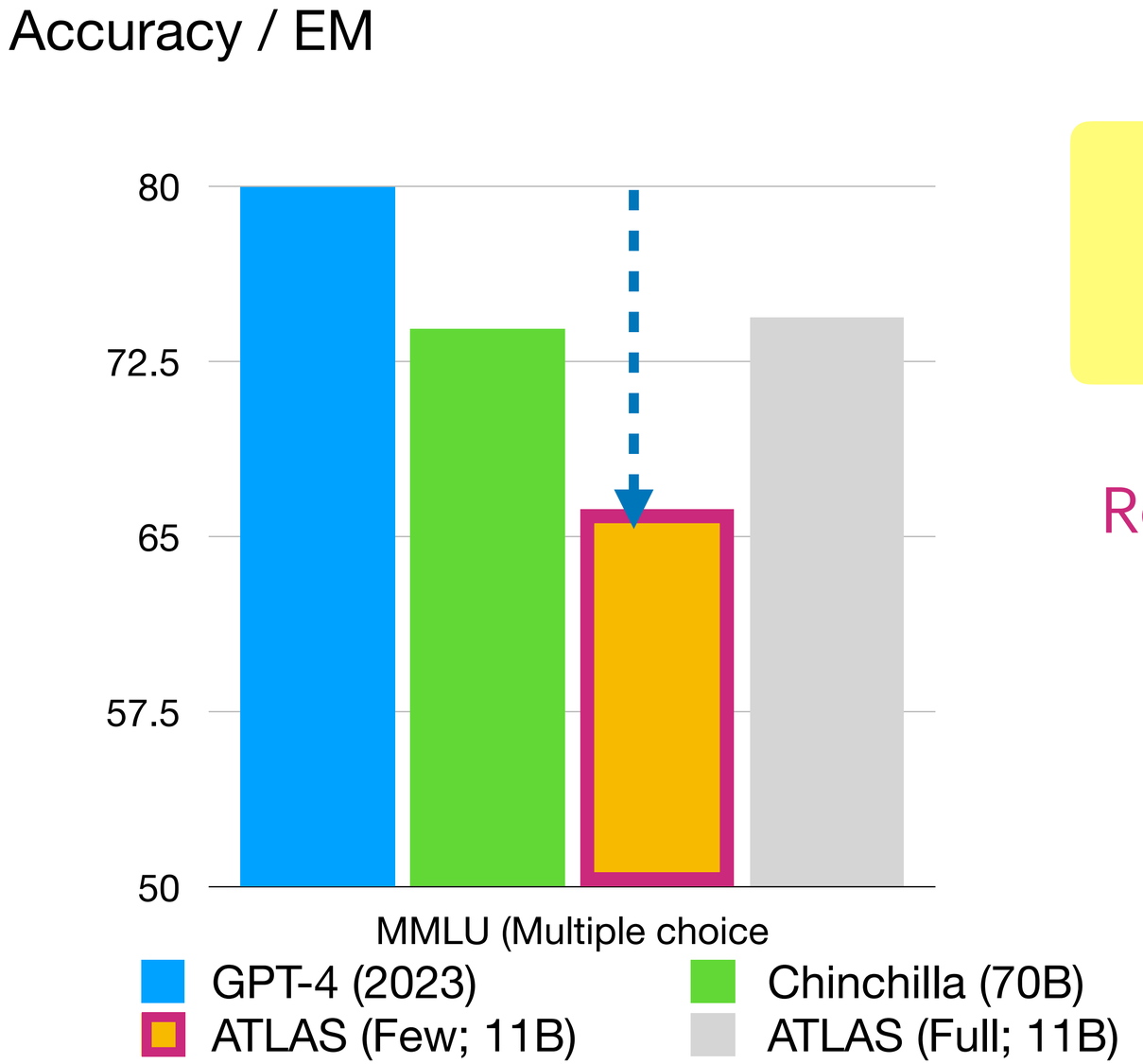
On QA, ATLAS largely outperforms other LLMs in few-shot

Task Results



Full-shot fine-tuning further improves performance

Task Results



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

Room for improvements for diverse task adaptations!

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

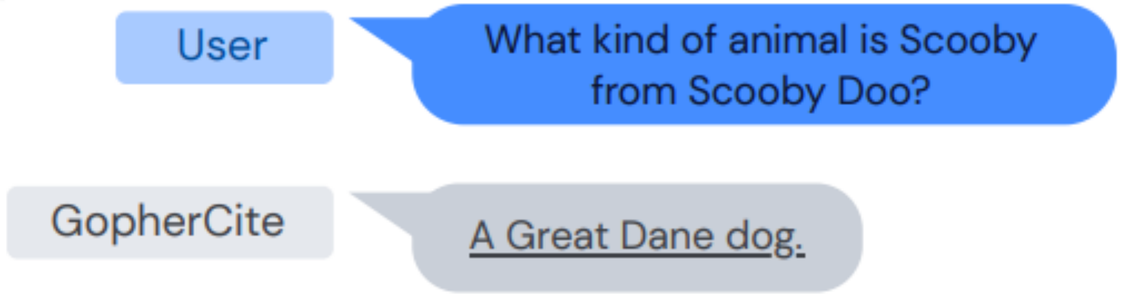
How to **adapt?**

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

What is **data store?**

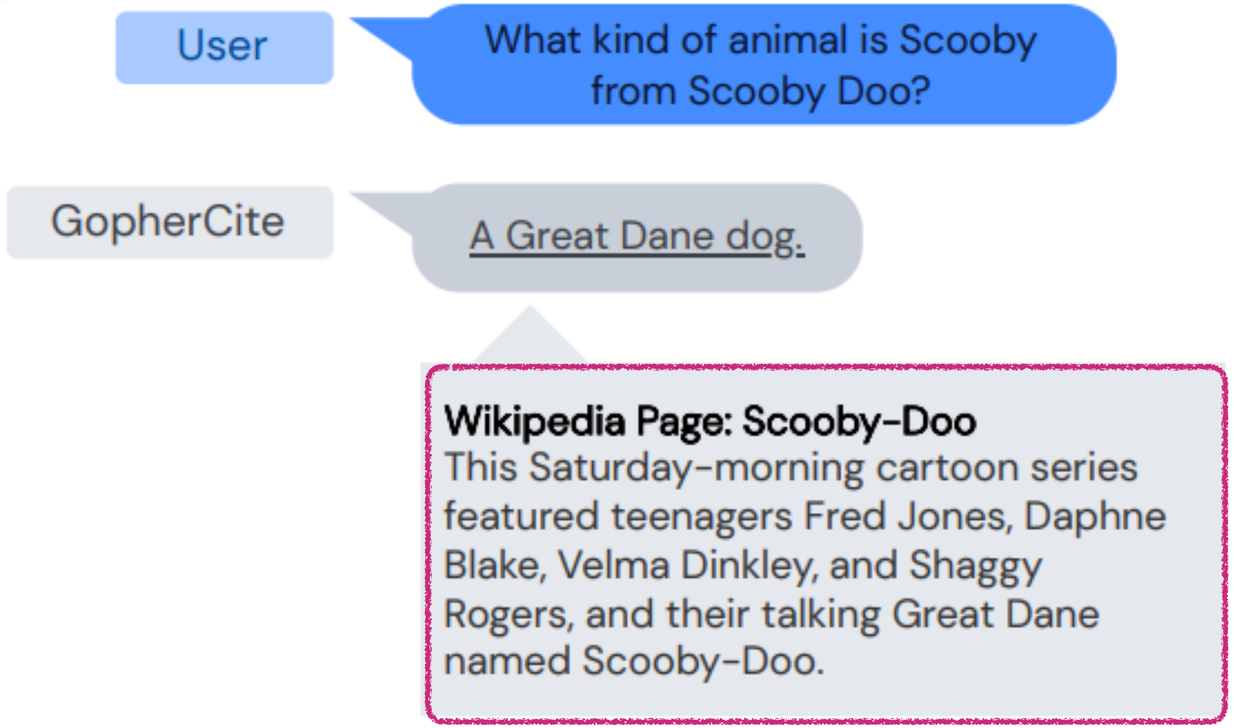
- Unlabeled Wikipedia / CC
- Web (Google / Bing Search Results)
- Training data

GopherCite (Menick et al., 2022)



Menick et al. 2022. “GopherCite: Teaching language models to support answers with verified quotes”

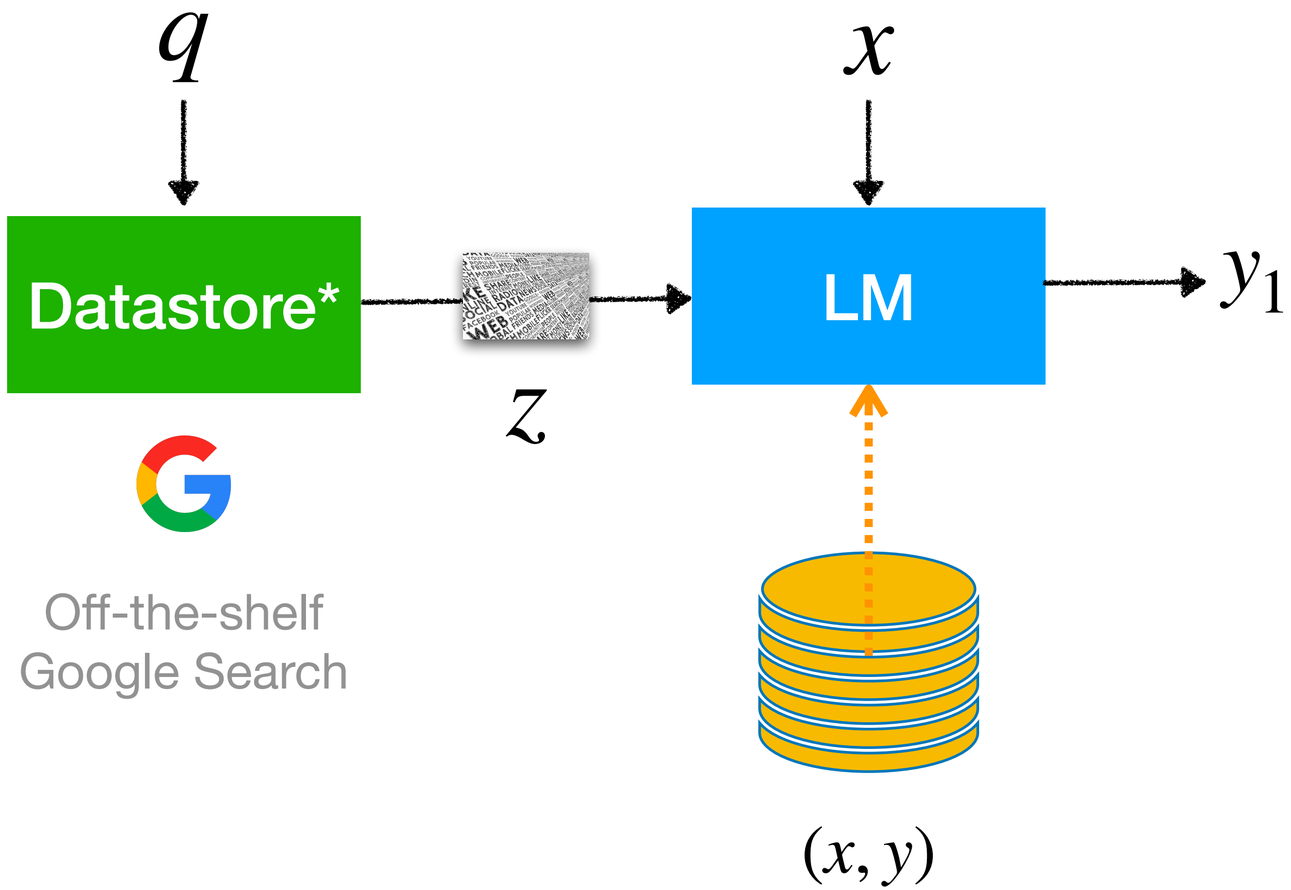
GopherCite (Menick et al., 2022)



Extract and generate a quote to support an answer

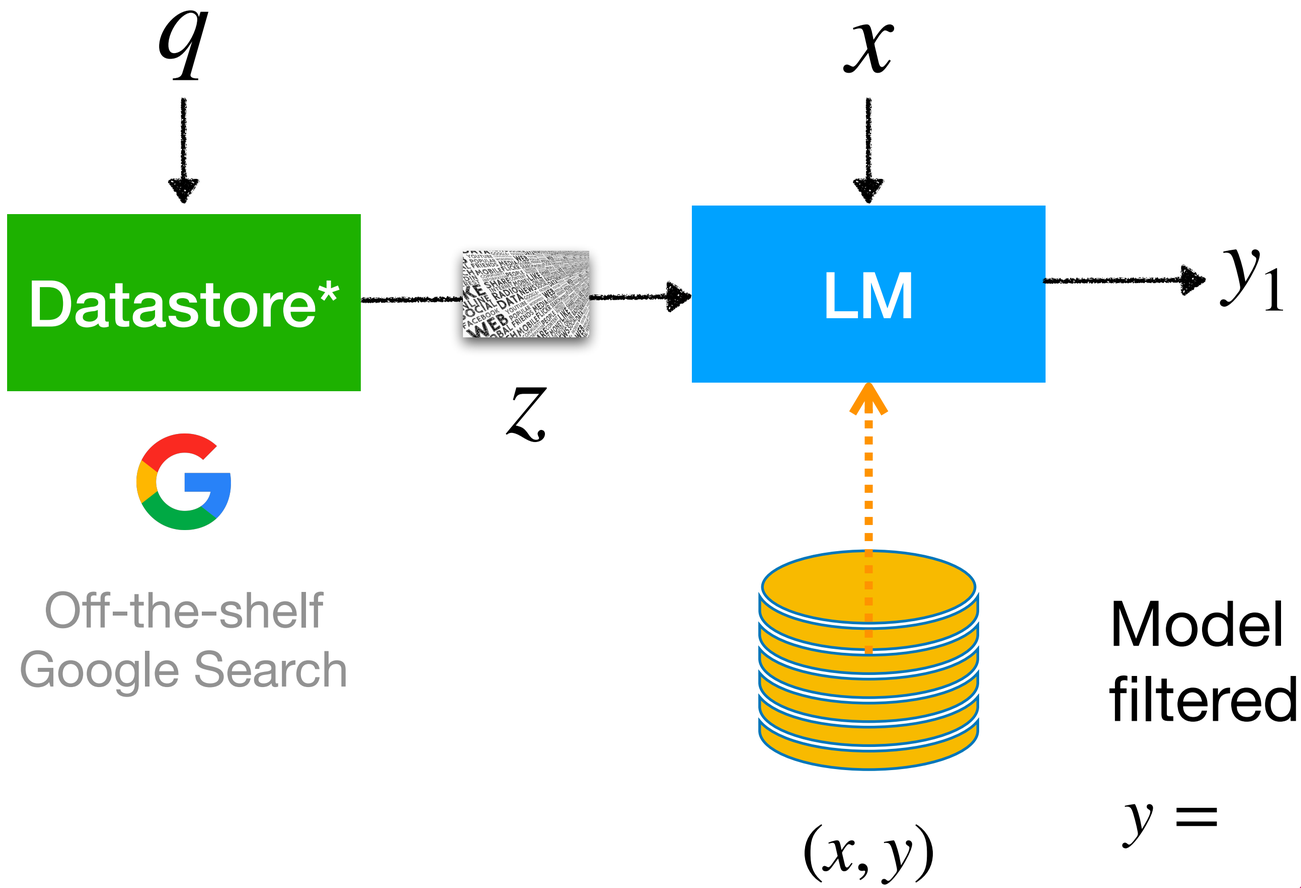
Menick et al. 2022. “GopherCite: Teaching language models to support answers with verified quotes”

GopherCite: RLHF for answering with verified quotes



Supervised fine-tuning (SFT)

GopherCite: RLHF for answering with verified quotes



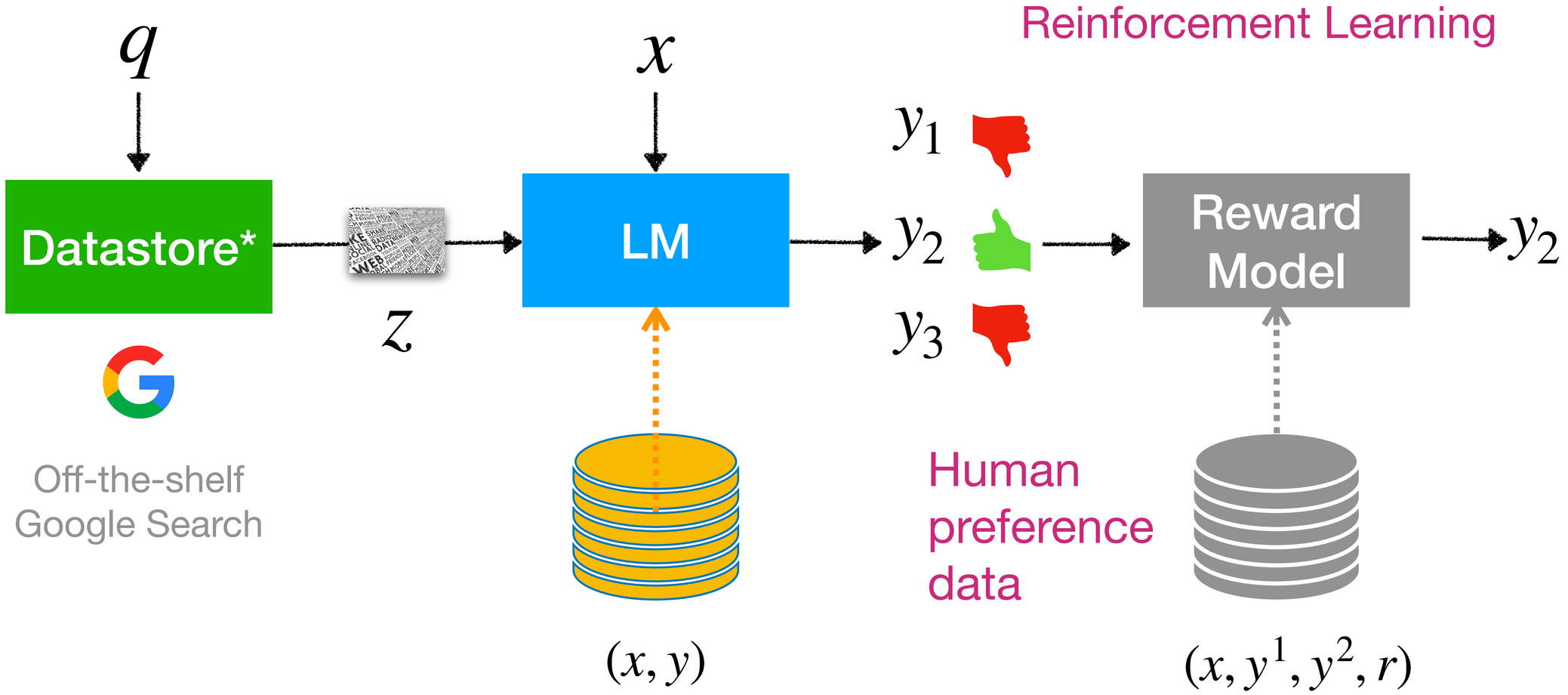
Supervised fine-tuning (SFT)

Model generated training data filtered by human

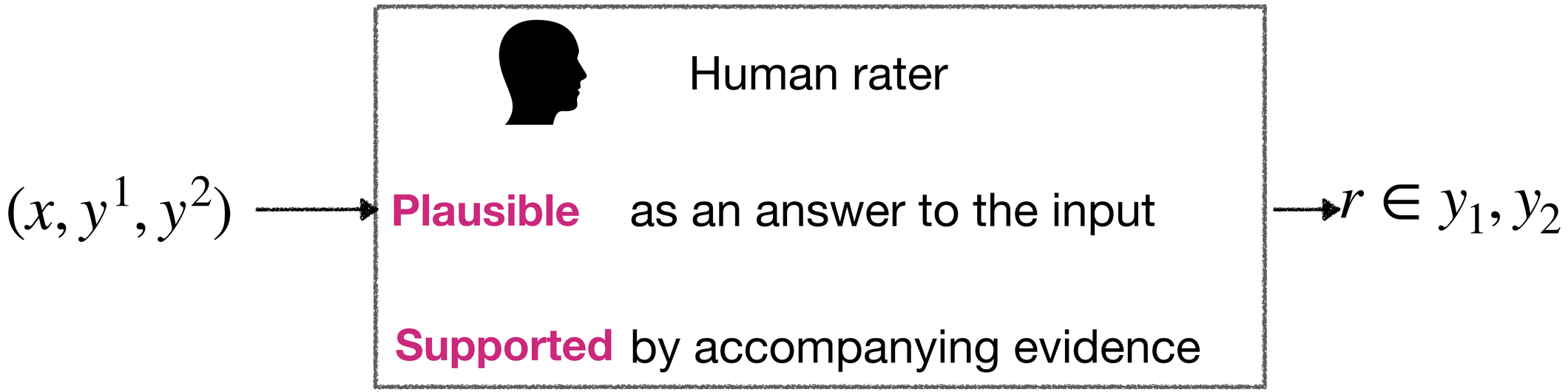
$y =$

`%<Claim>%(Document title)%[Quote from document]%`

GopherCite: RLHF for answering with verified quotes



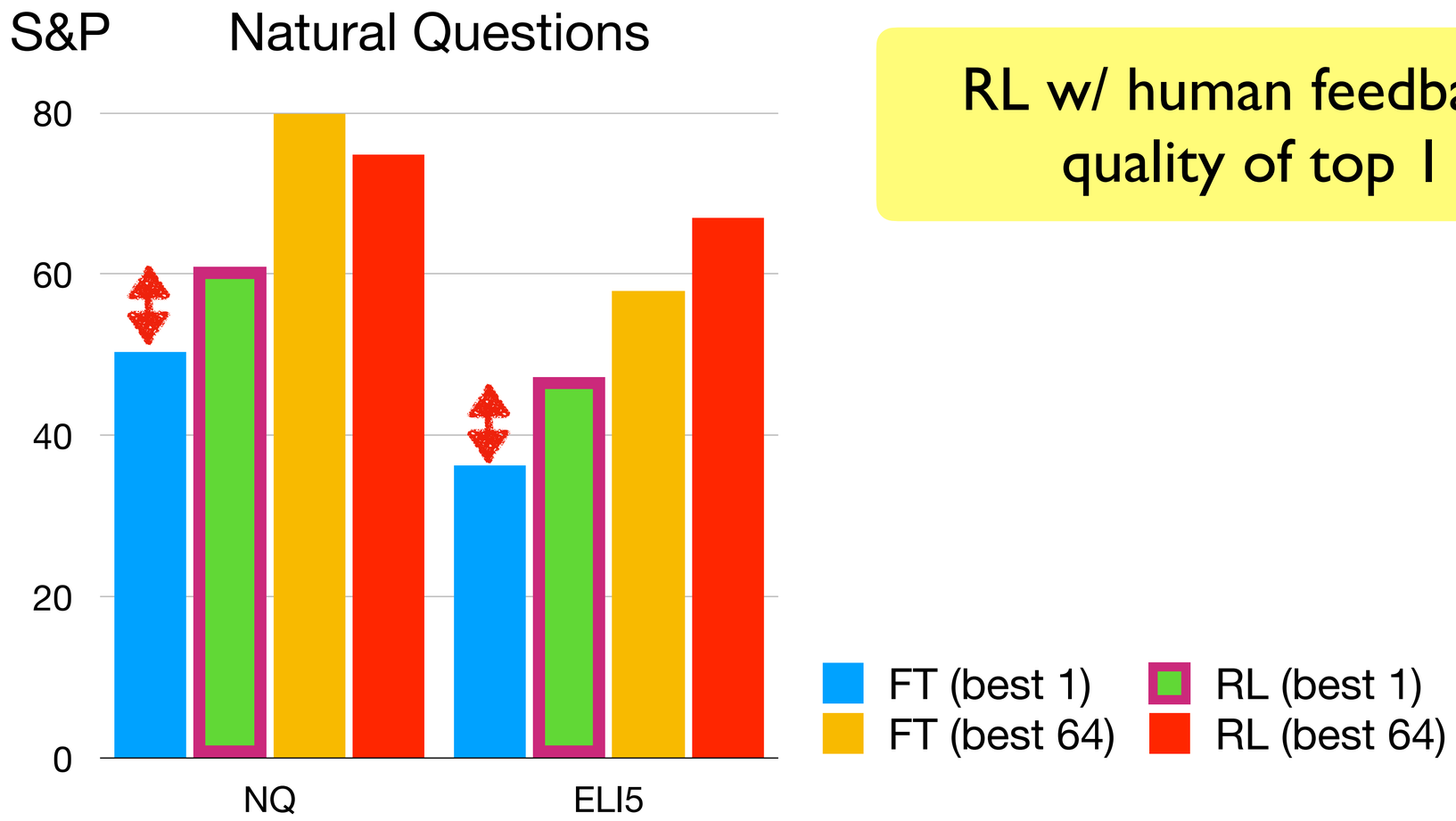
GopherCite: RLHF for answering with verified quotes



33k Human preference data

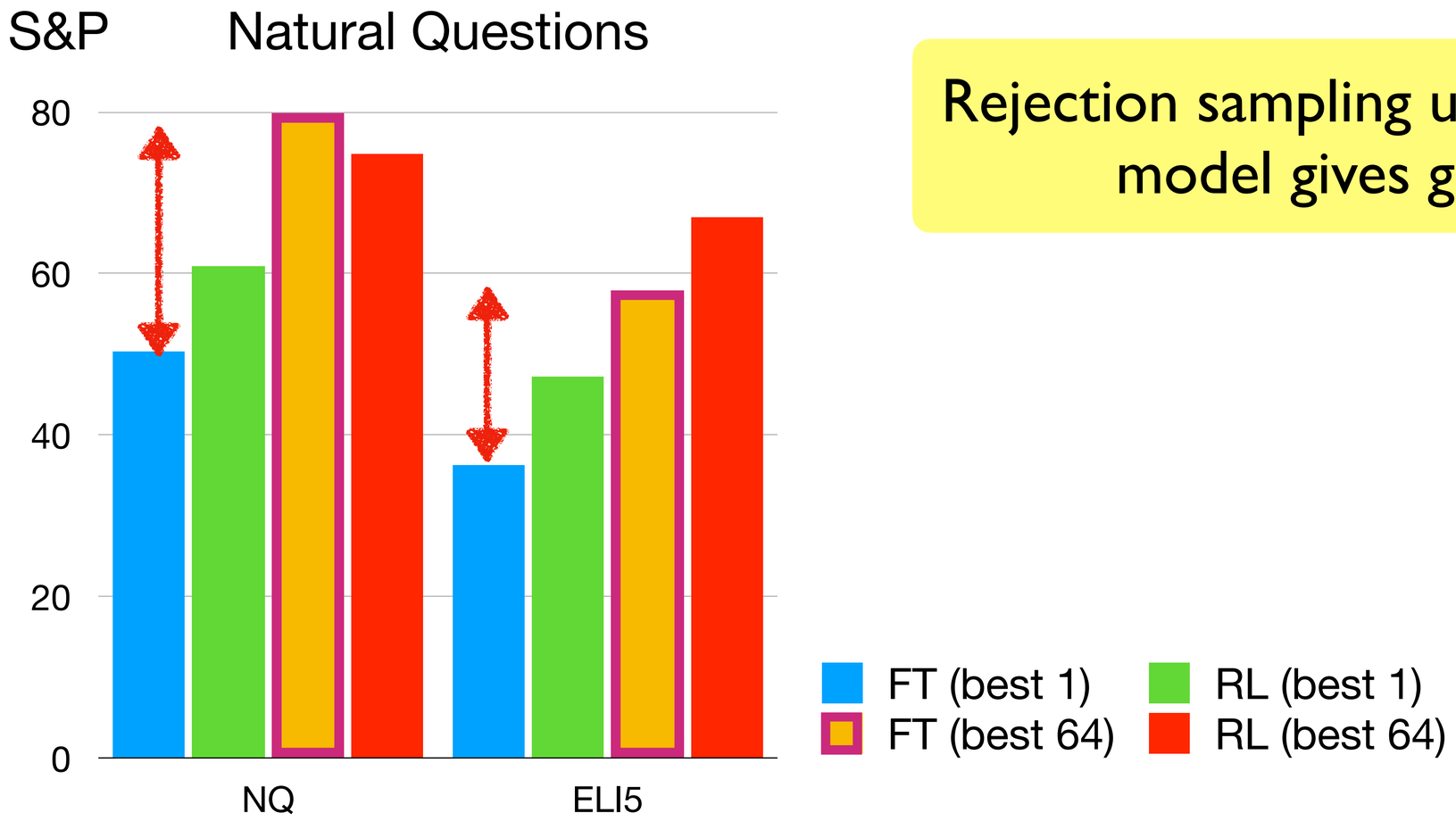
(x, y^1, y^2, r)

GopherCite: effects of RL



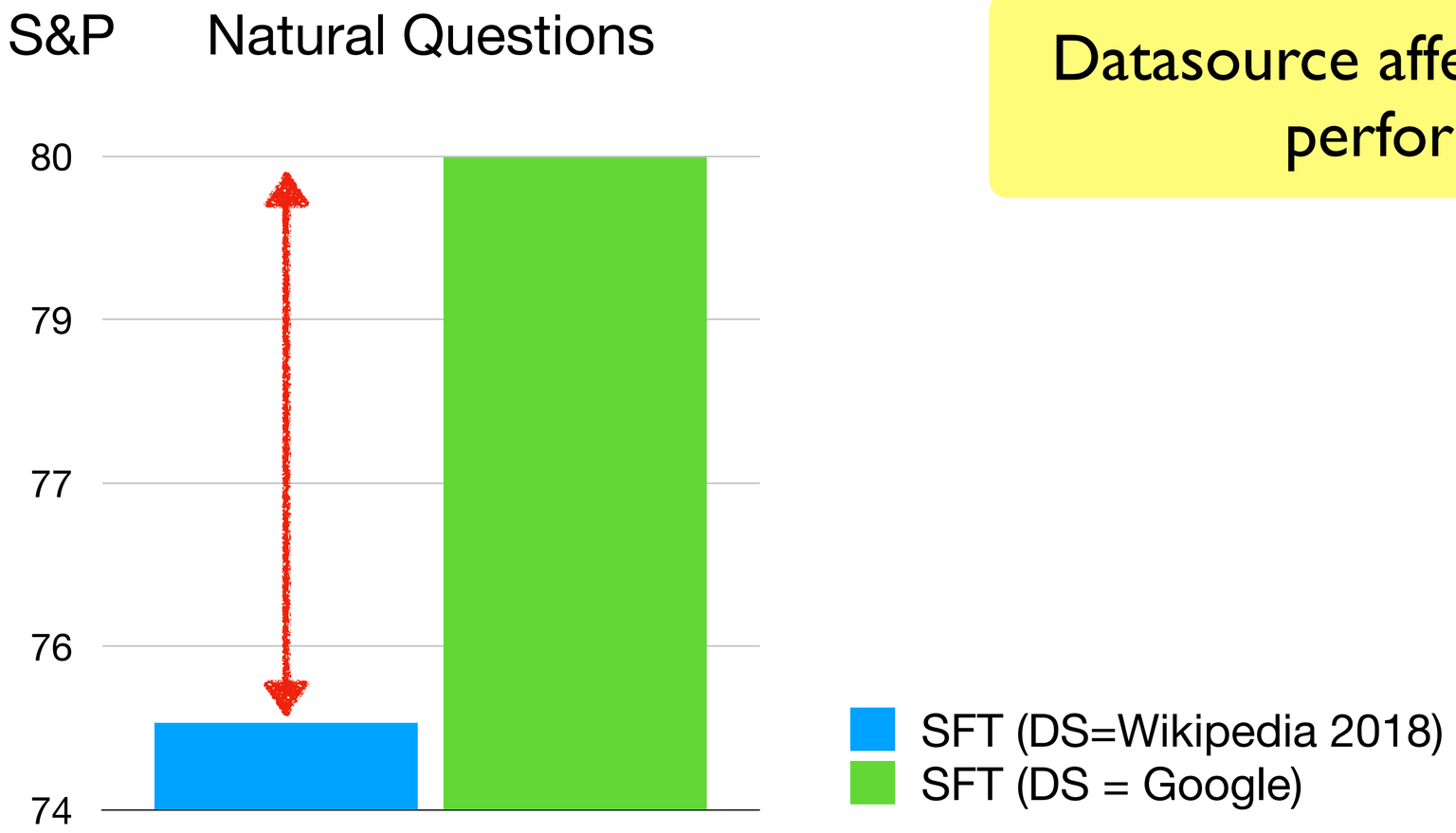
RL w/ human feedback improve the quality of top 1 generations

GopherCite: effects of RL



Rejection sampling using trained reward model gives gains from FT

GopherCite: RLHF for answering with verified quotes



Datasource affects final model performance

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)	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 LM 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 … 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:

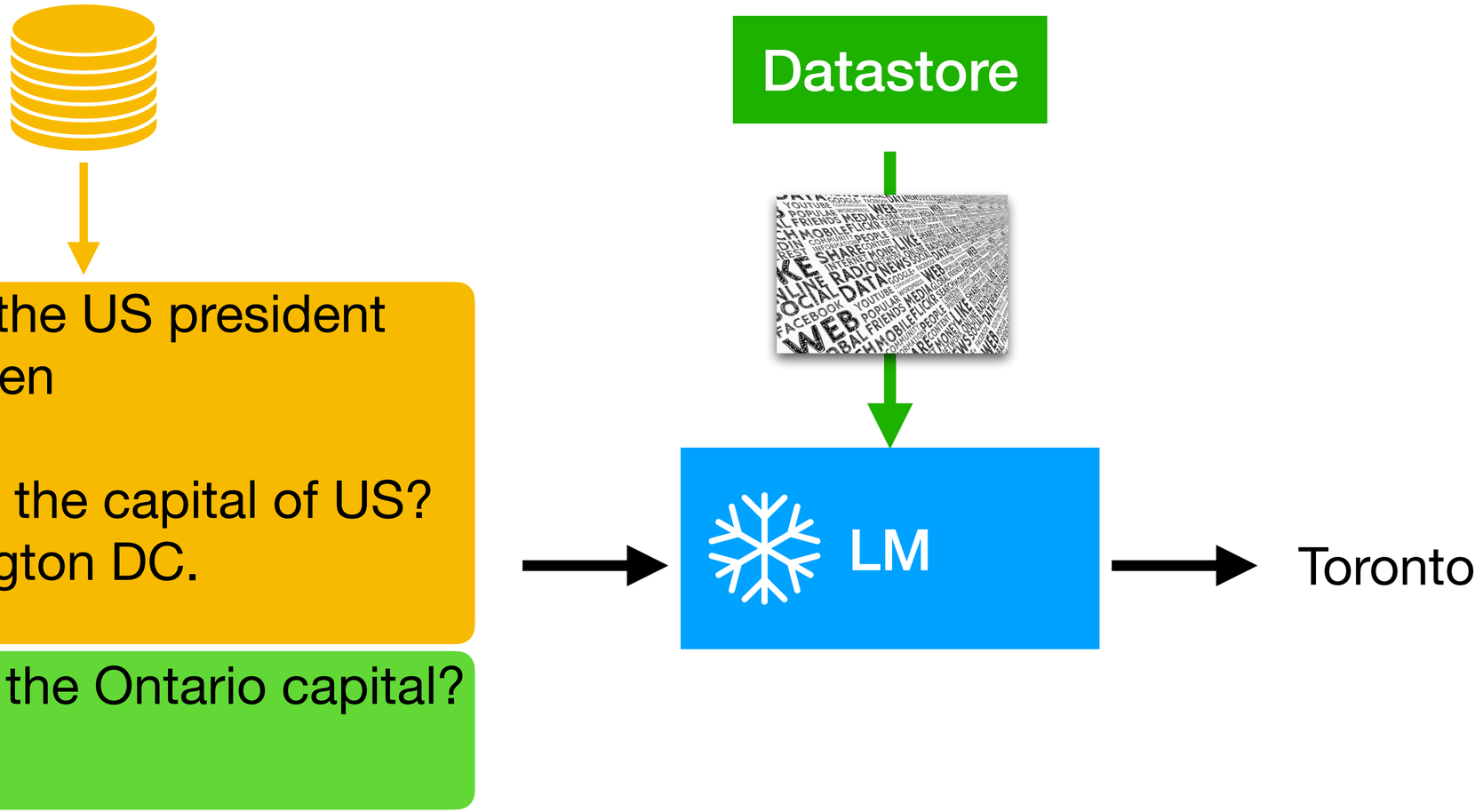
Doesn't require LM training on tasks!

Training instances (demonstrations)

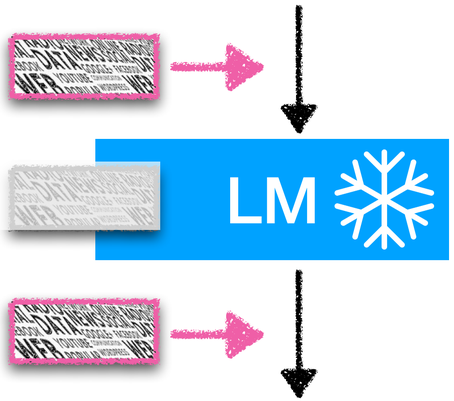
Test instances

Retrieval-based prompting

k-shot instances (k=0, 32 … etc)



Design choice of retrieval-based Prompting

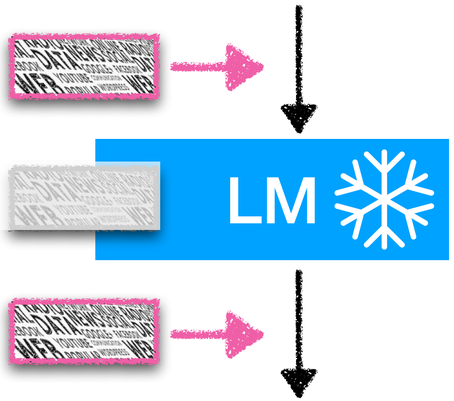


Input space:
Append retrieved context in input space

Intermediate layers:
N/A

Output space:
Interpolate token probability distributions in output space

Design choice of retrieval-based Prompting



Input space:
Append retrieved context in input space

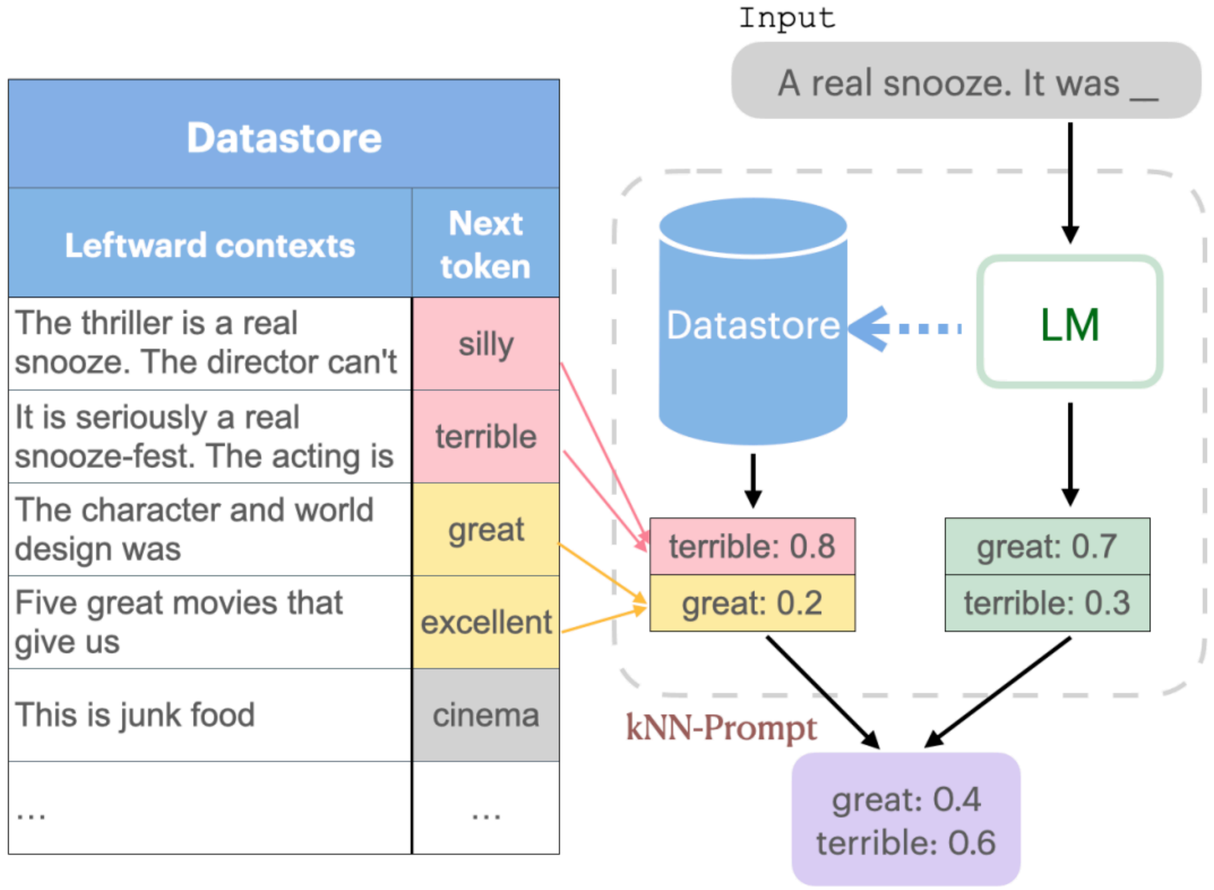
Intermediate layers:
N/A

Extending kNN-LM for downstream tasks

Output space:
Interpolate token probability distributions in output space

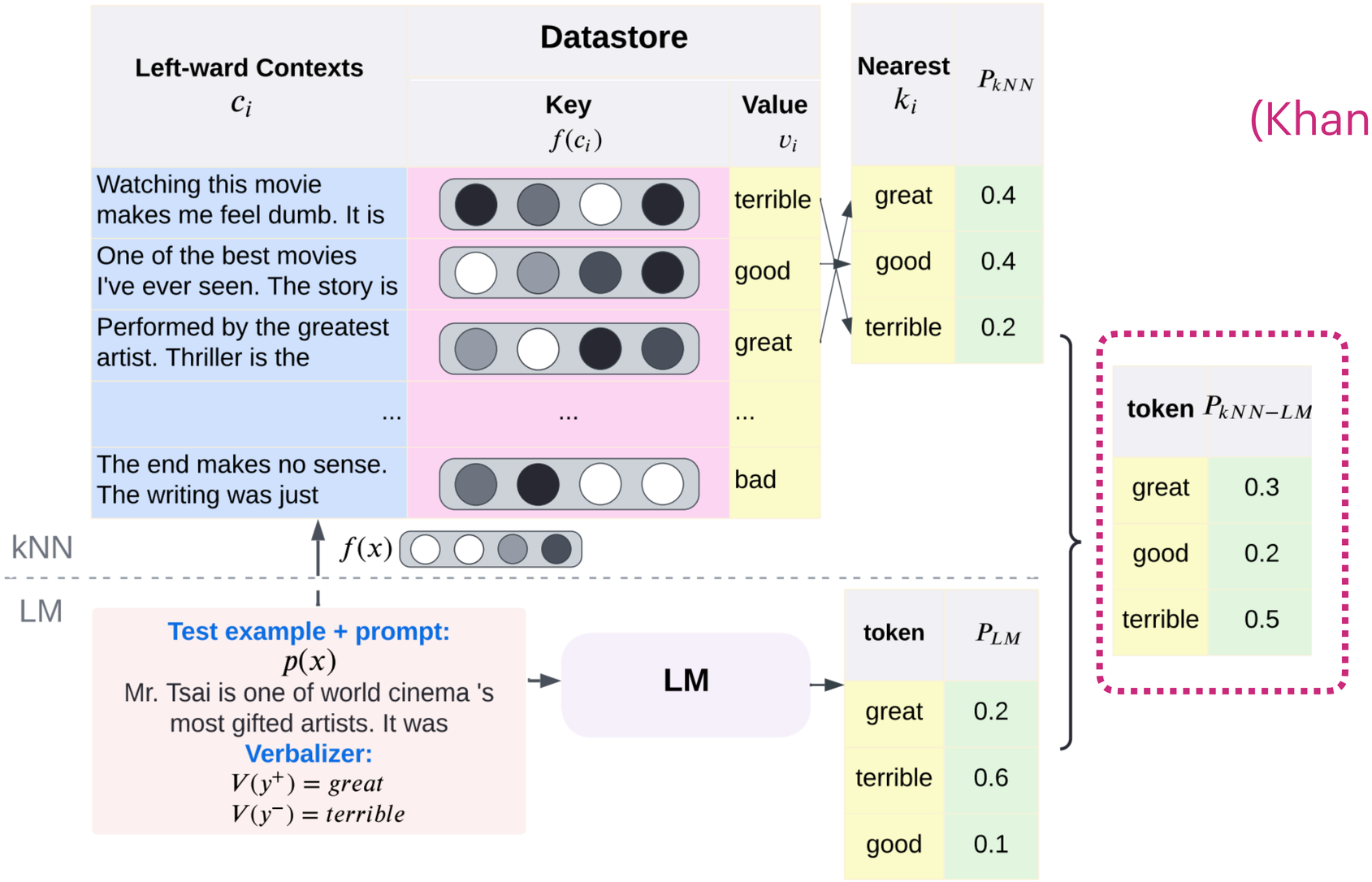
kNN-prompt (Shi et al., 2022)

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



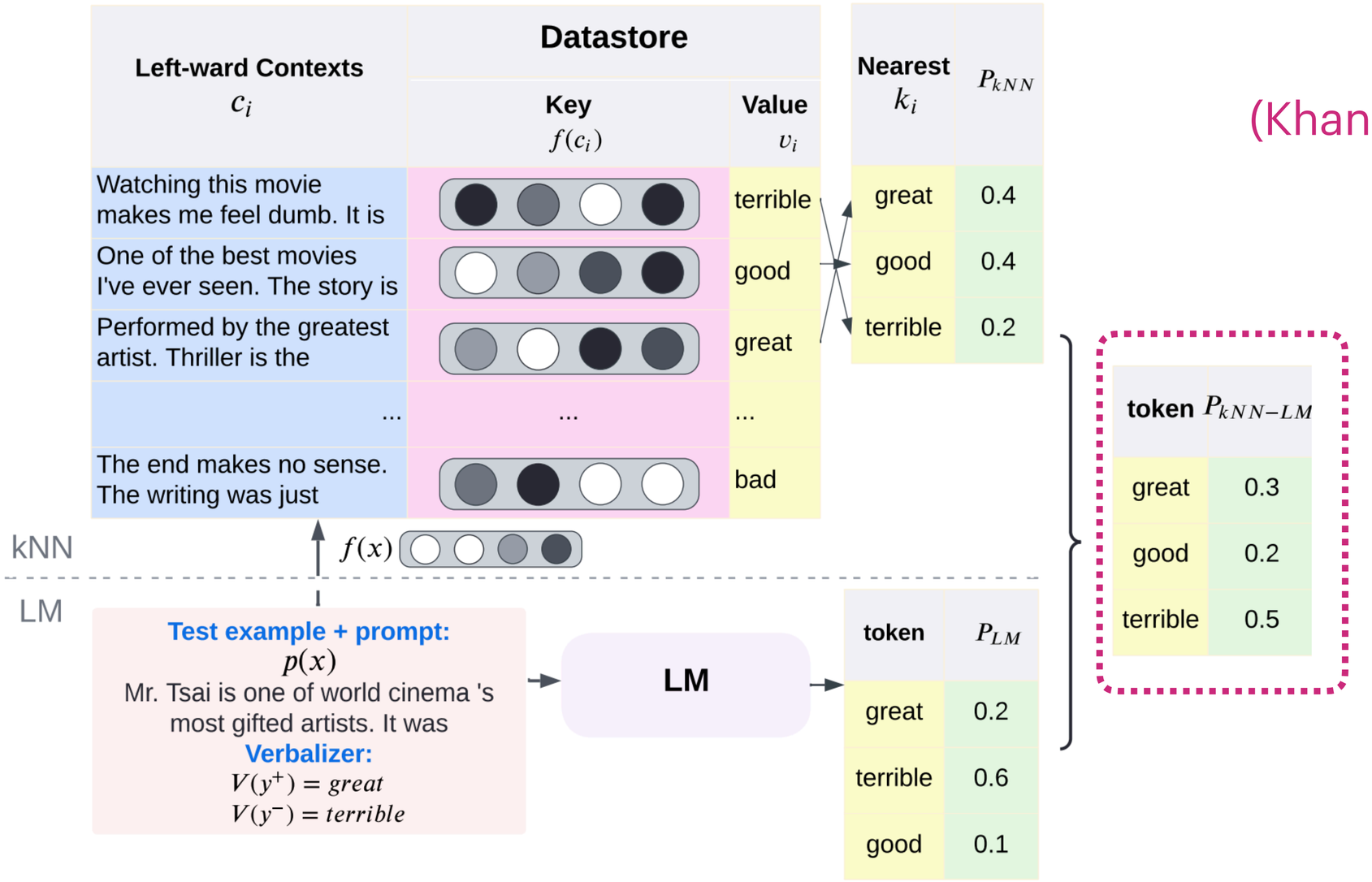
Shi et al. 2022. “Nearest Neighbor Zero-shot Inference”

kNN-prompt (Shi et al., 2022)



kNN-LM
(Khandelwal et al., 2020)

kNN-prompt (Shi et al., 2022)



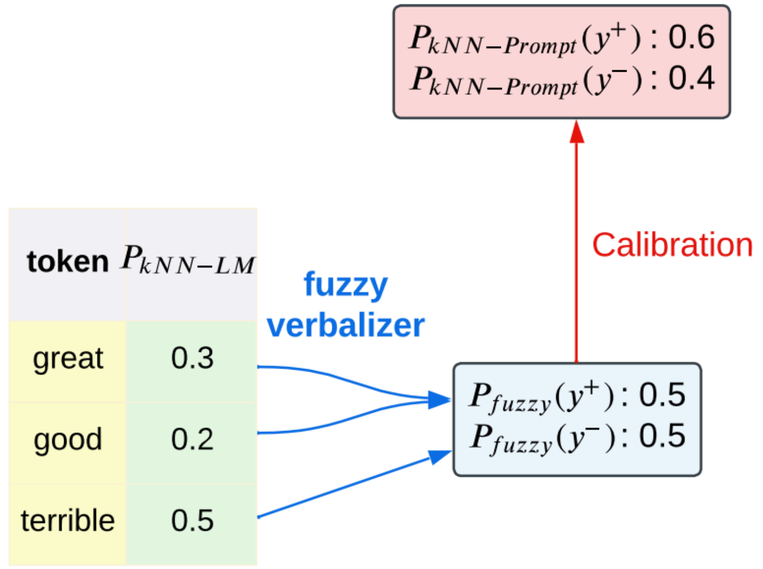
kNN-LM
(Khandelwal et al., 2020)

The kNN token distributions are quite sparse

kNN-prompt (Shi et al., 2022)

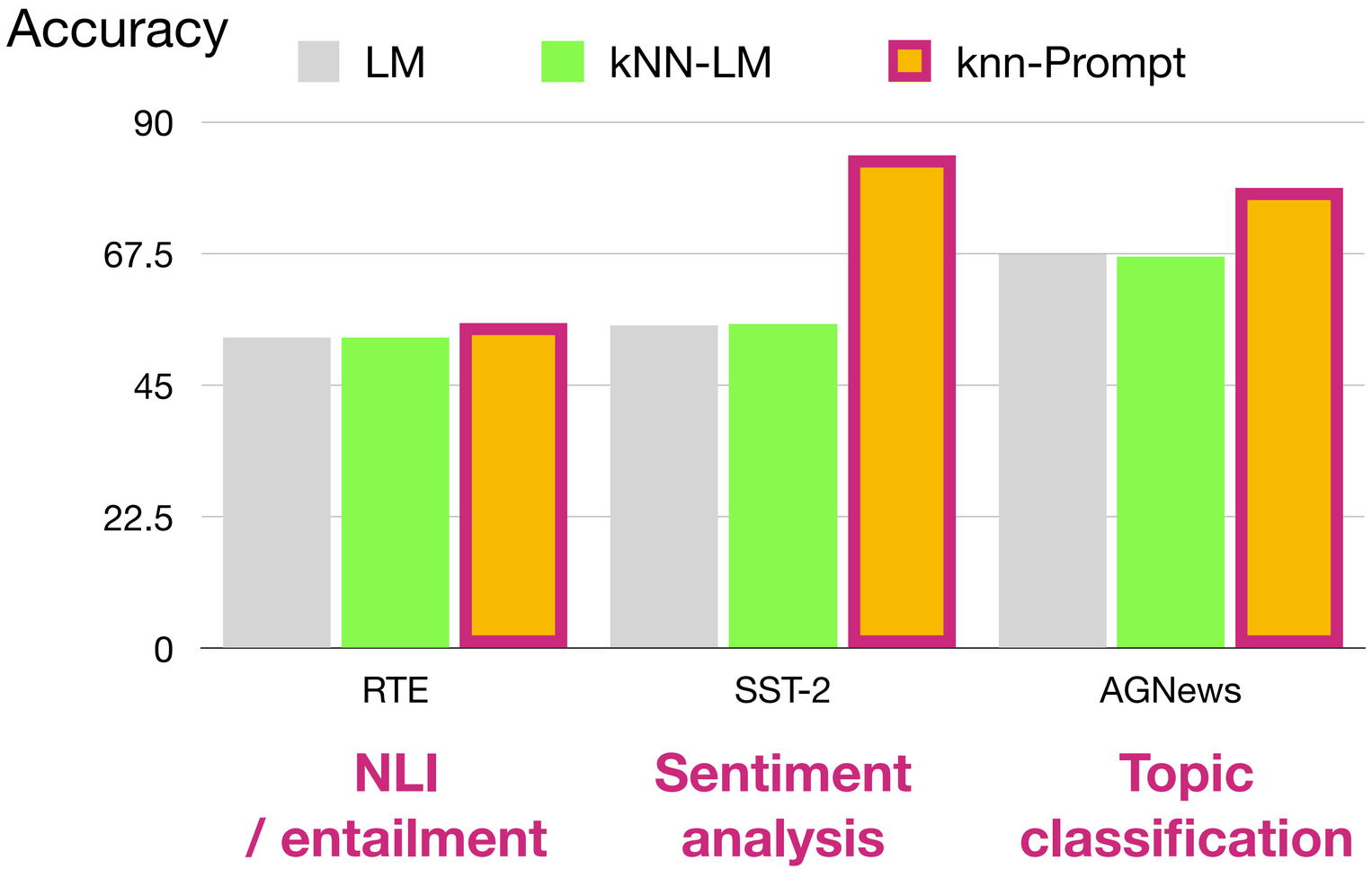
$$P_{FV}(y \mid x) \propto \sum_{v_i \in \mathcal{N}(v)} P(v_i \mid p(x))$$

Find similar tokens using GloVe & ConceptNet



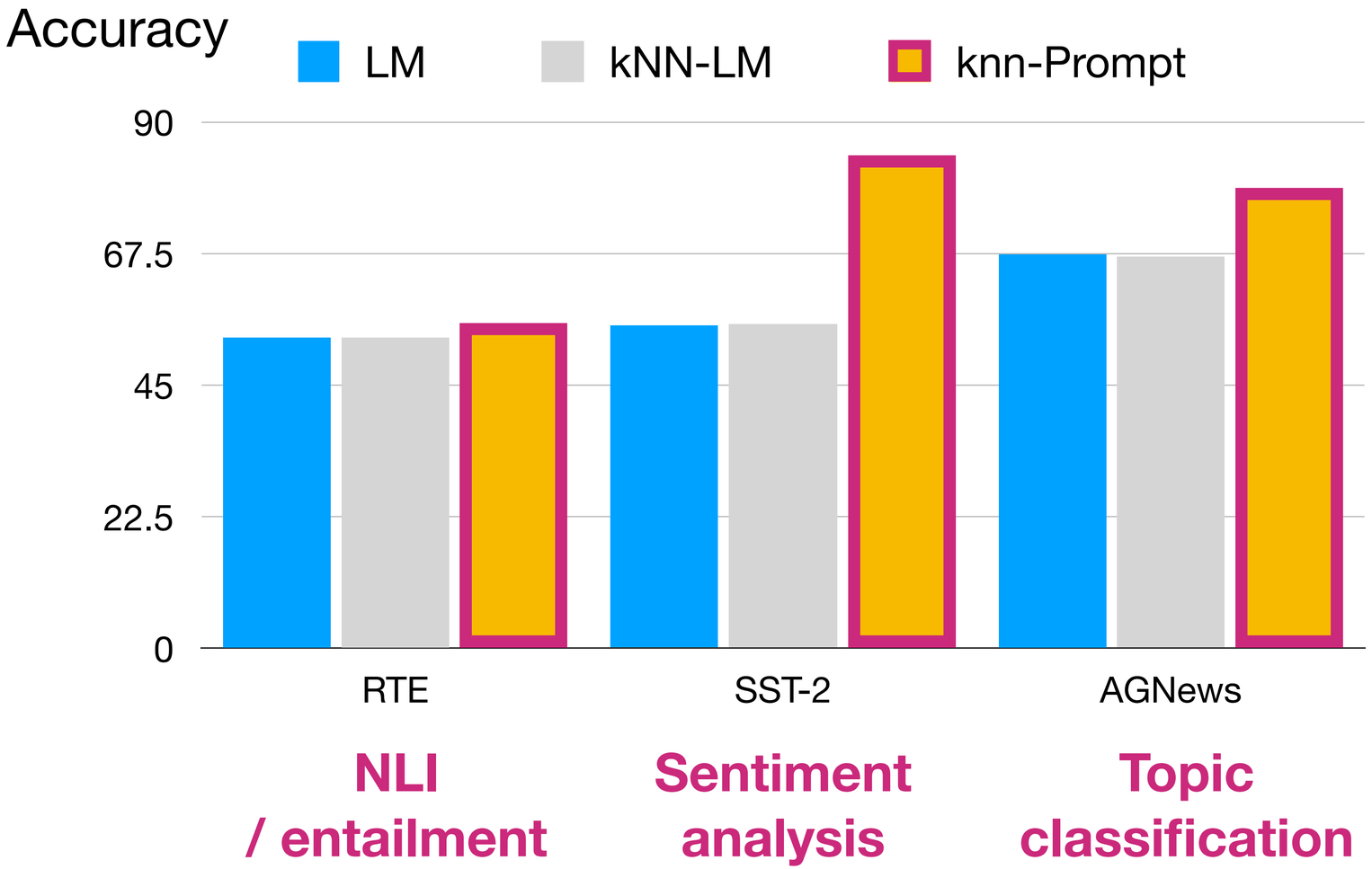
fuzzy verbalizer maps token probability to target class labels

Results on diverse classification tasks



Significant gains from kNN-LM

Results on diverse classification tasks



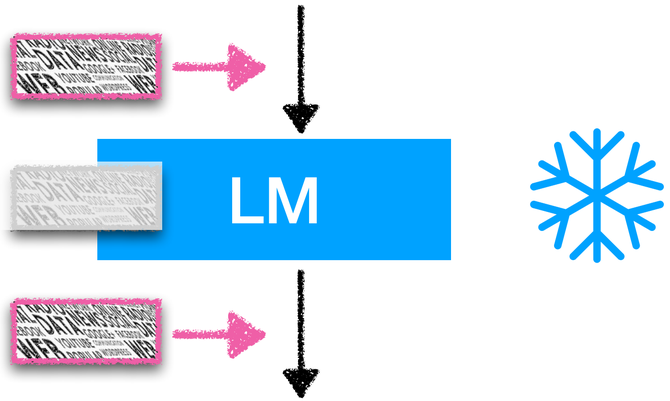
kNN prompt largely outperforms vanilla LM in zero-shot classification

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

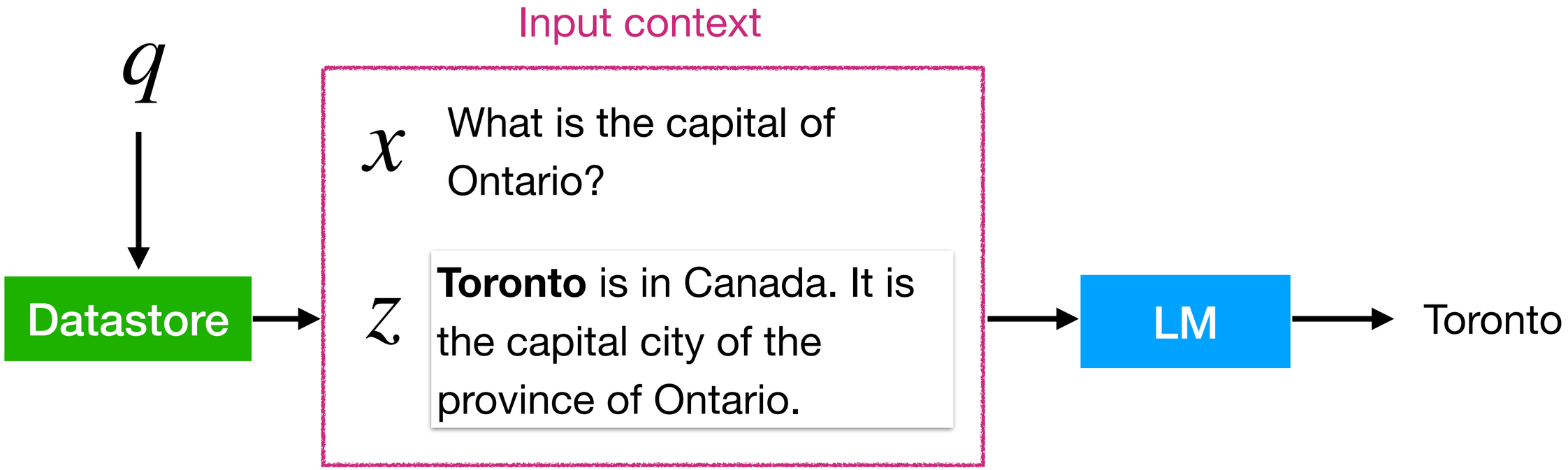


Input space:
Append retrieved context in input space

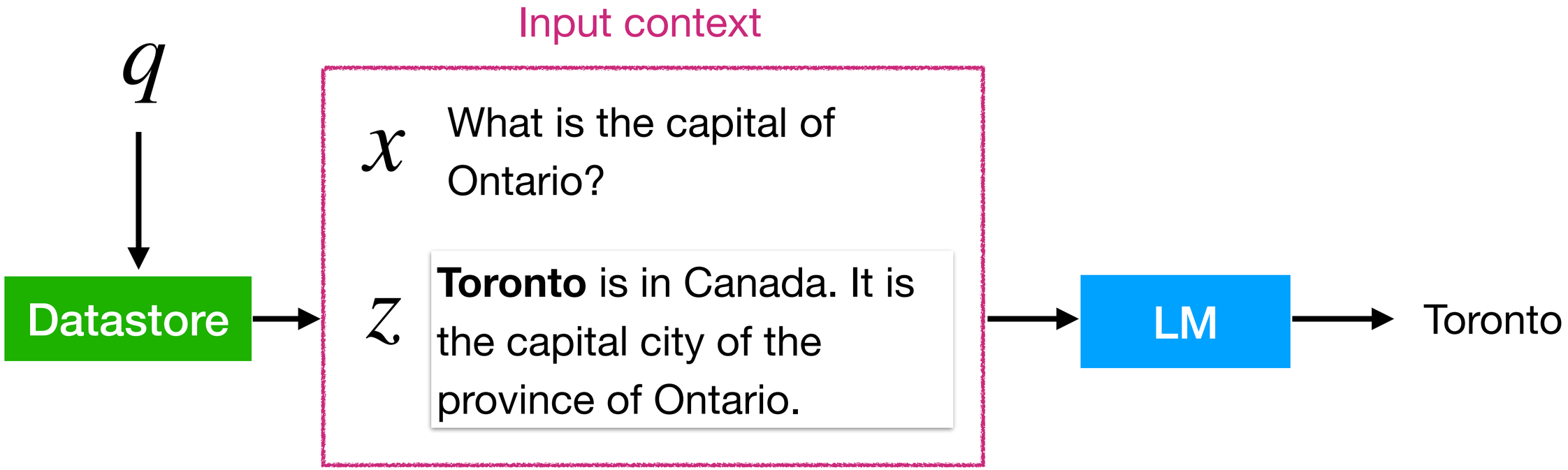
Intermediate layers:
N/A

Output space:
Interpolate token probability distributions in output space

Retrieval-in-context

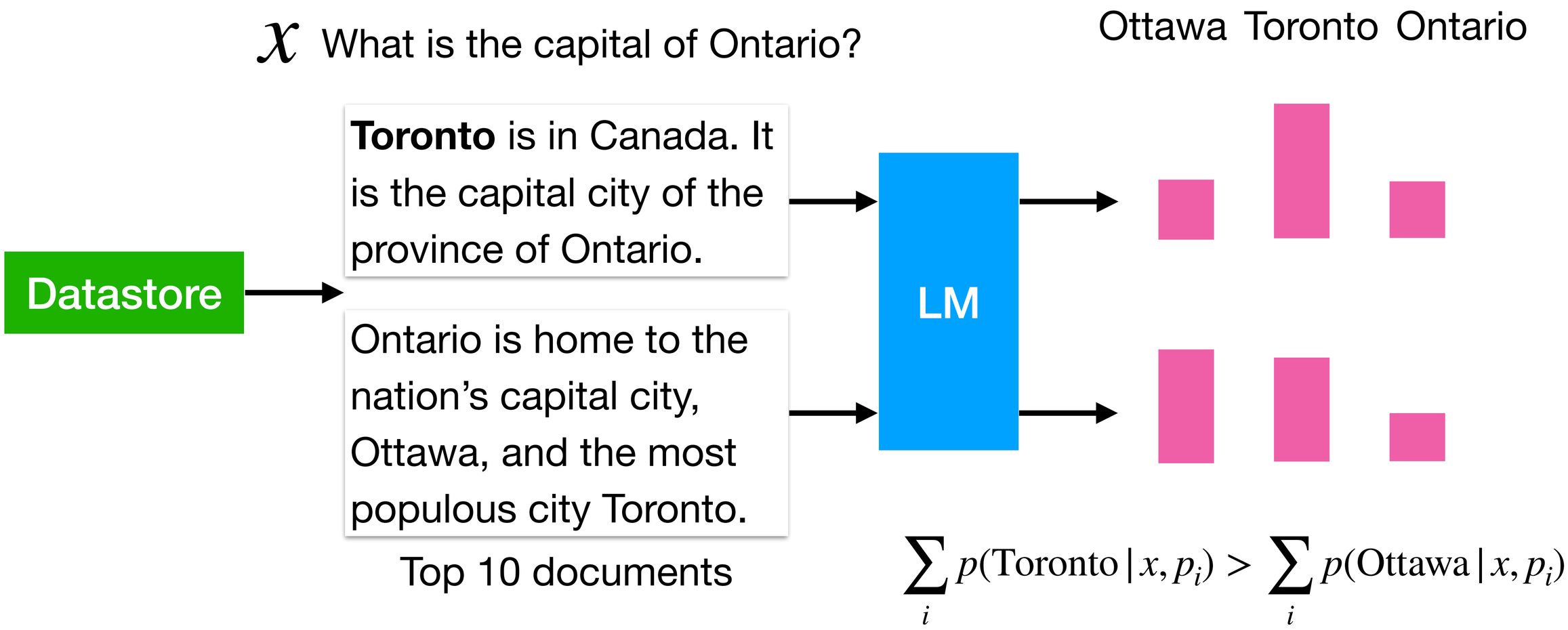


Retrieval-in-context

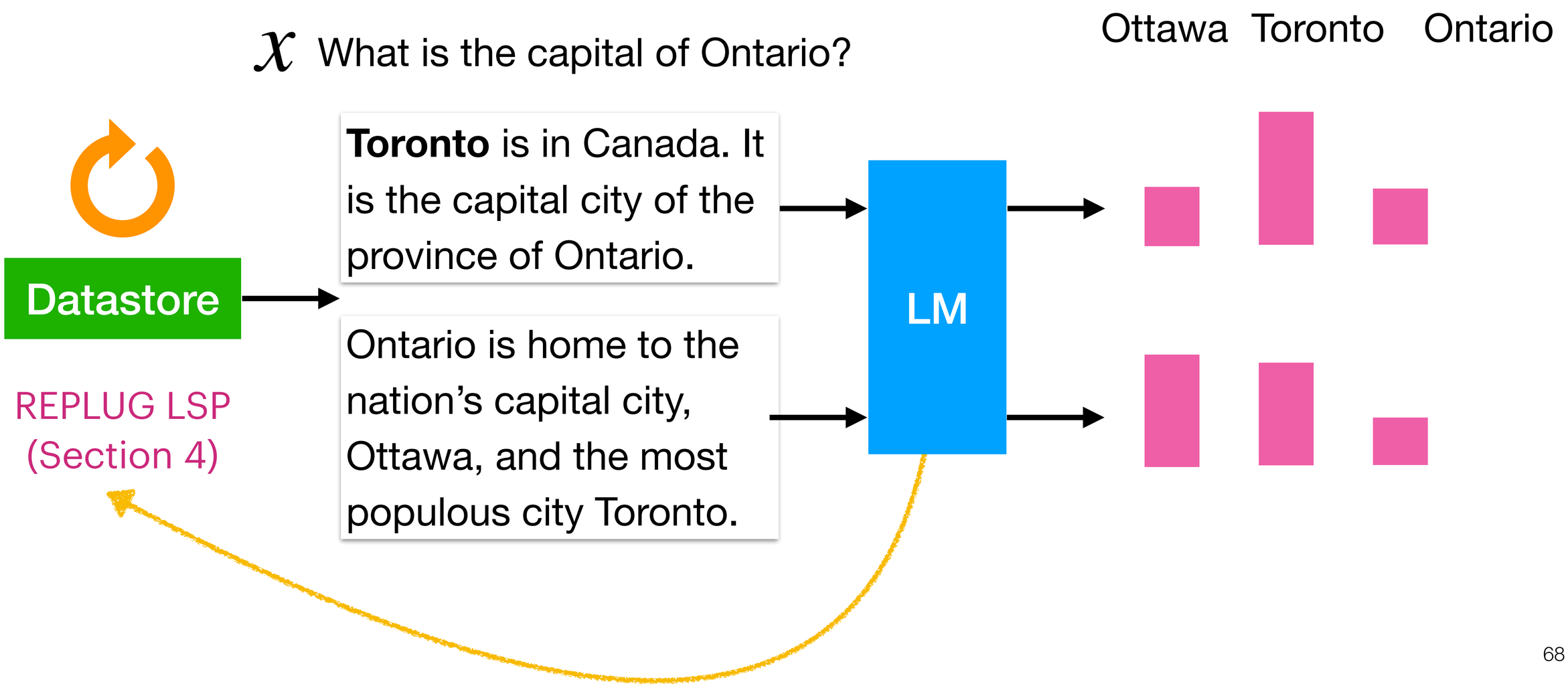


(Shi et al., 2023; Ram et al., 2022; Mallen et al., 2022; Yu et al., 2022; Press et al., 2022; *inter alia*)

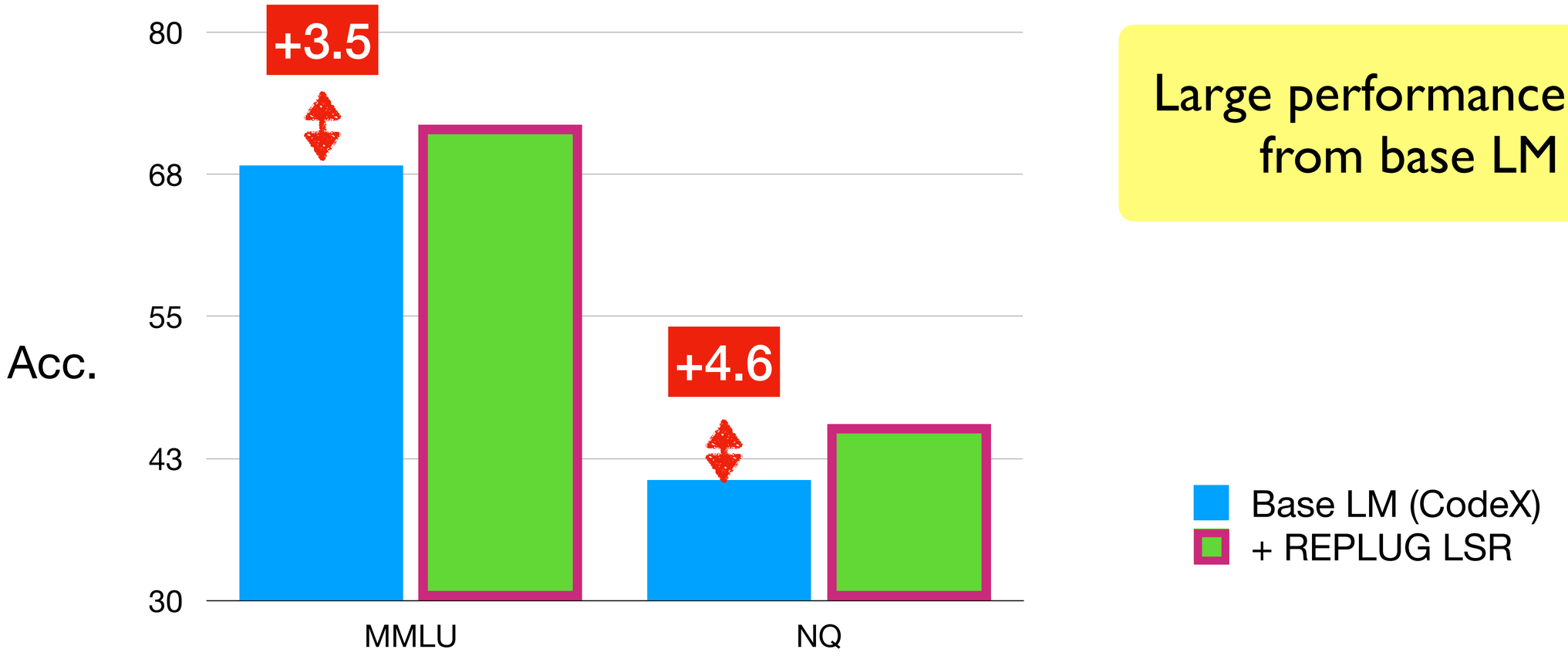
REPLUG (Shi et al., 2023; Section 3&4)



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

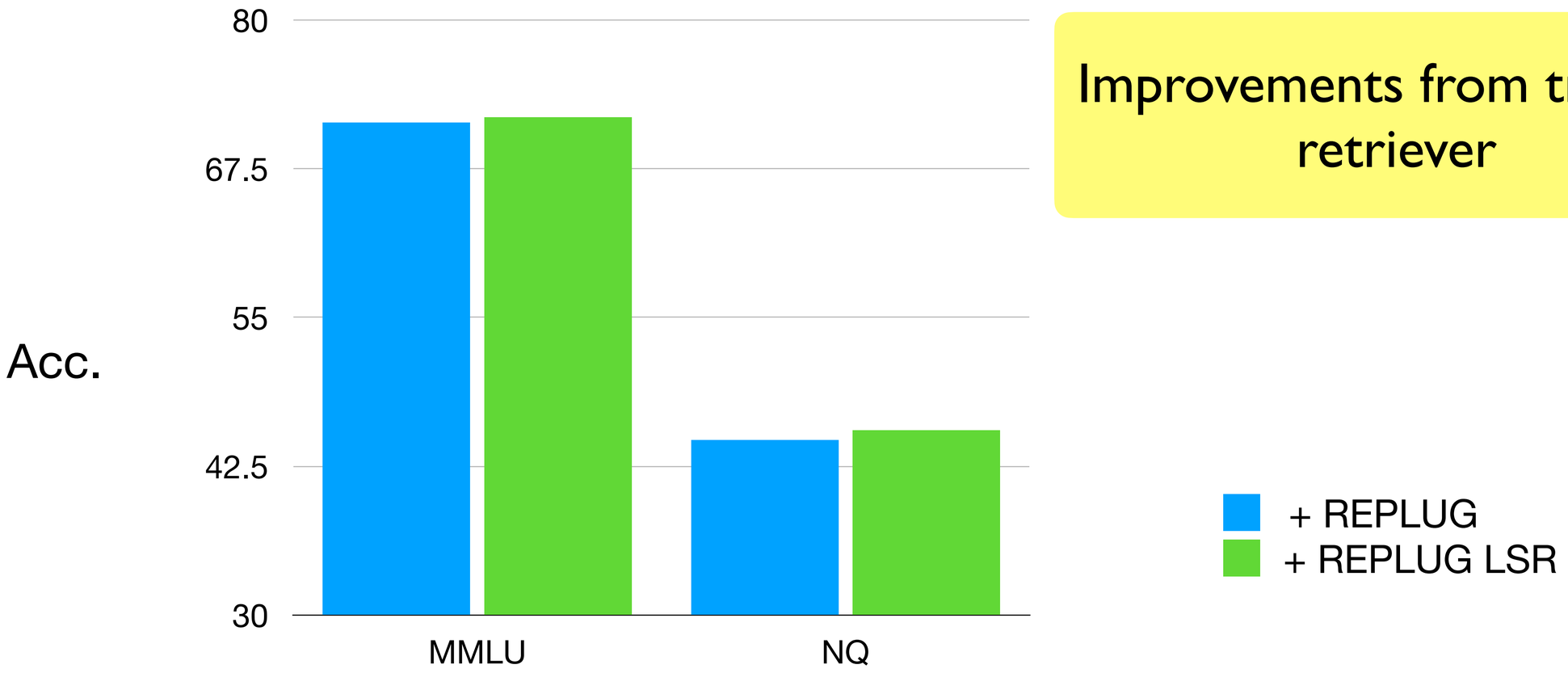


REPLUG: results on QA & MMLU



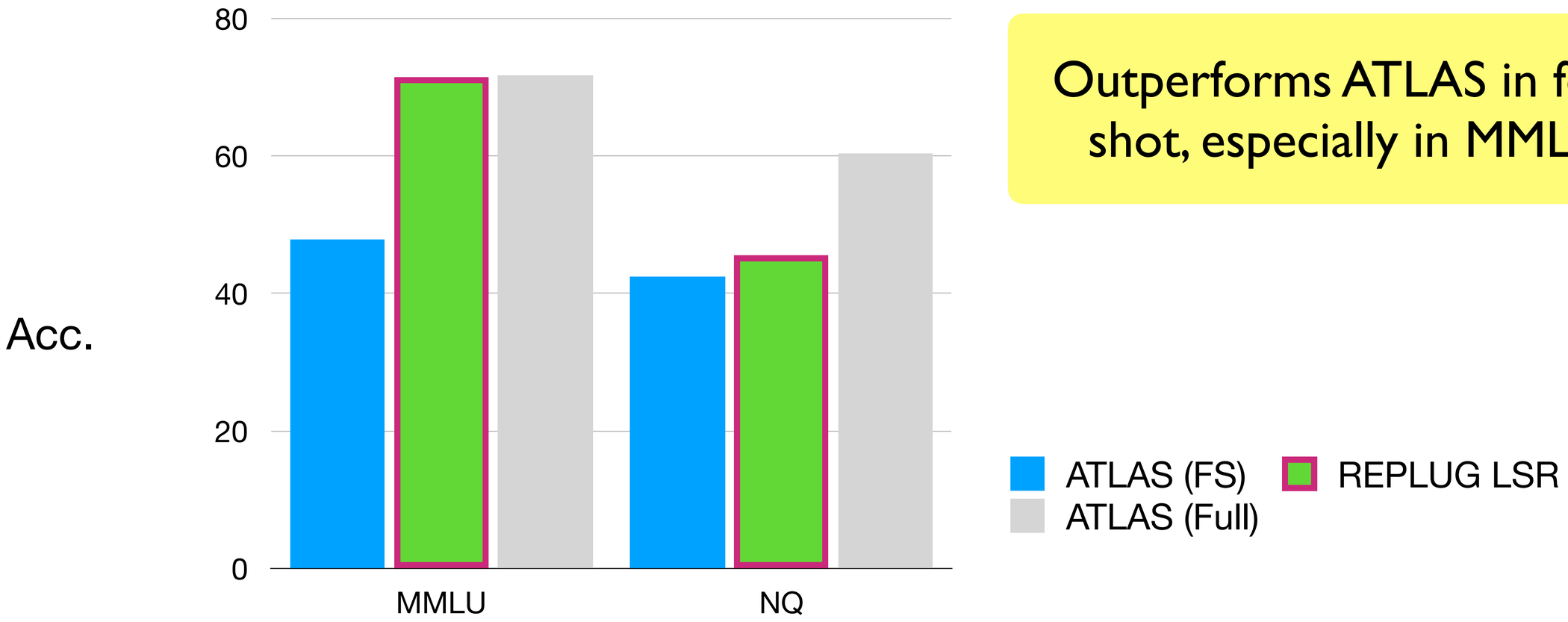
Large performance gain from base LM

REPLUG ablations of retriever training



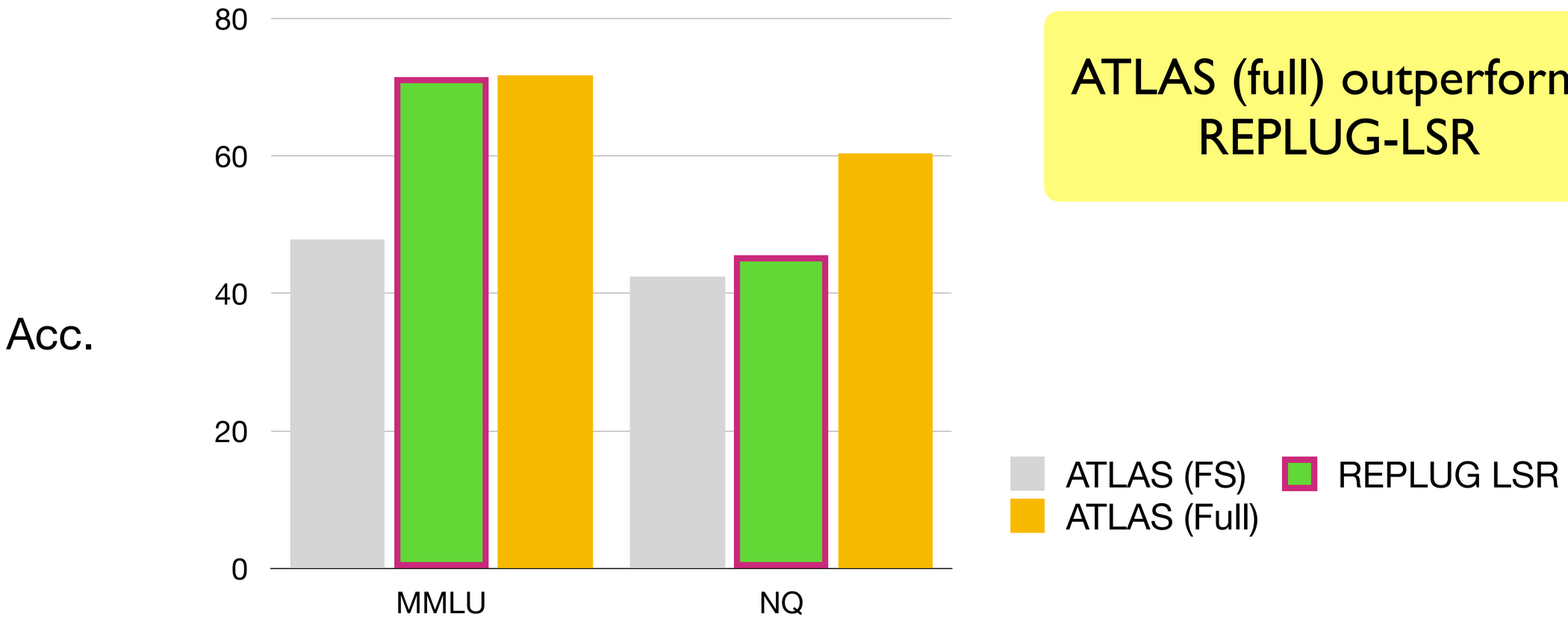
Improvements from training retriever

REPLUG: comparison with ATLAS



Outperforms ATLAS in few-shot, especially in MMLU

REPLUG: comparison with ATLAS



ATLAS (full) outperforms REPLUG-LSR

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

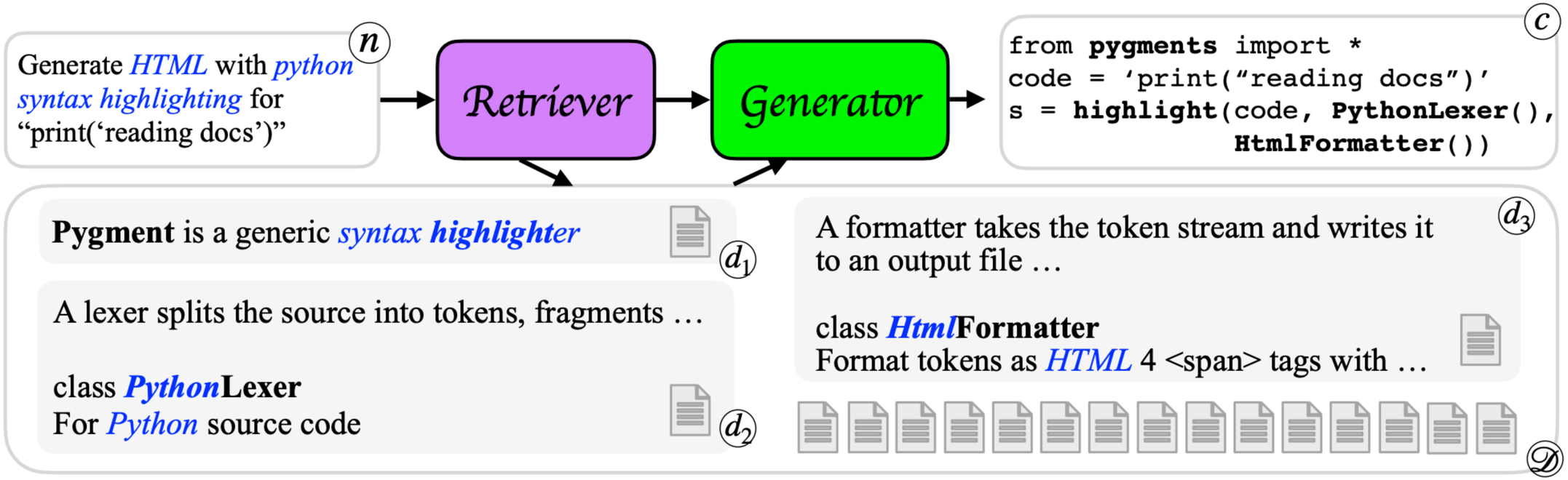
Hard to control, underperforming full FT model

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?

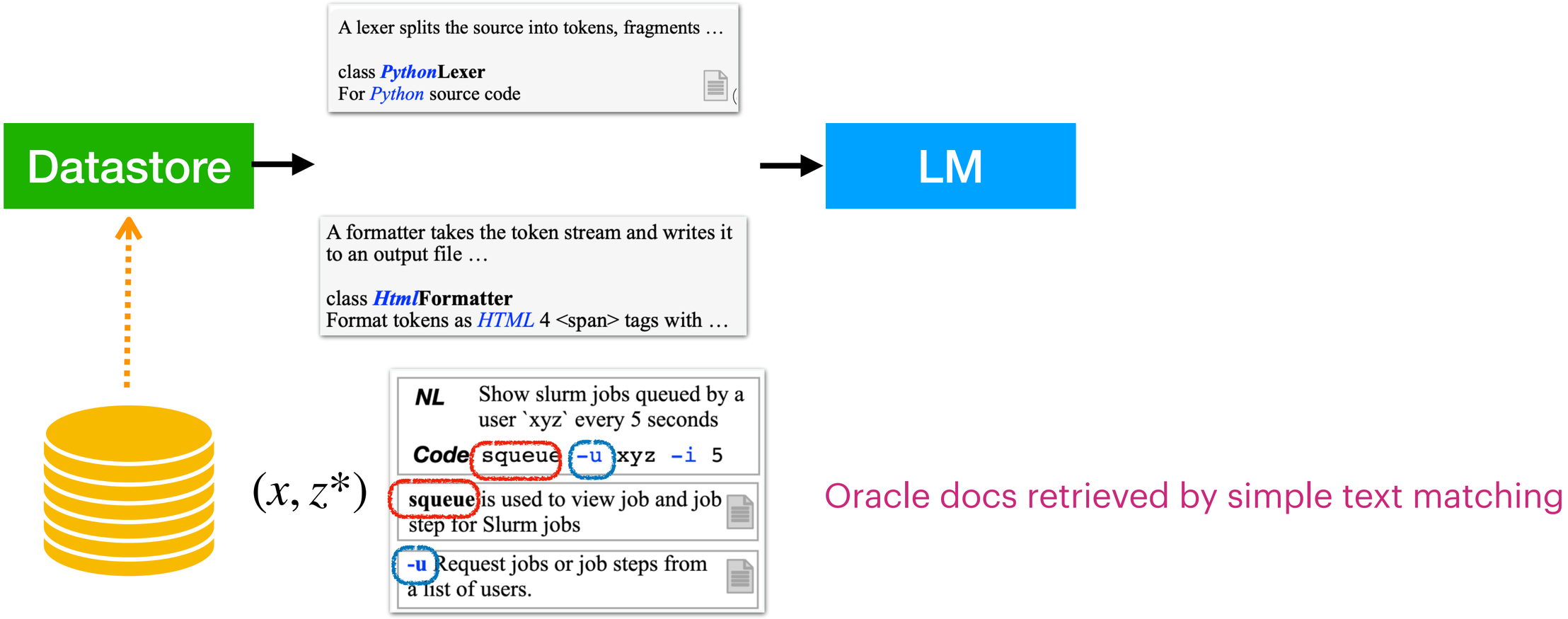
DocPrompting (Zhou et al., 2023)



Retrieve **code documentations** about related functions

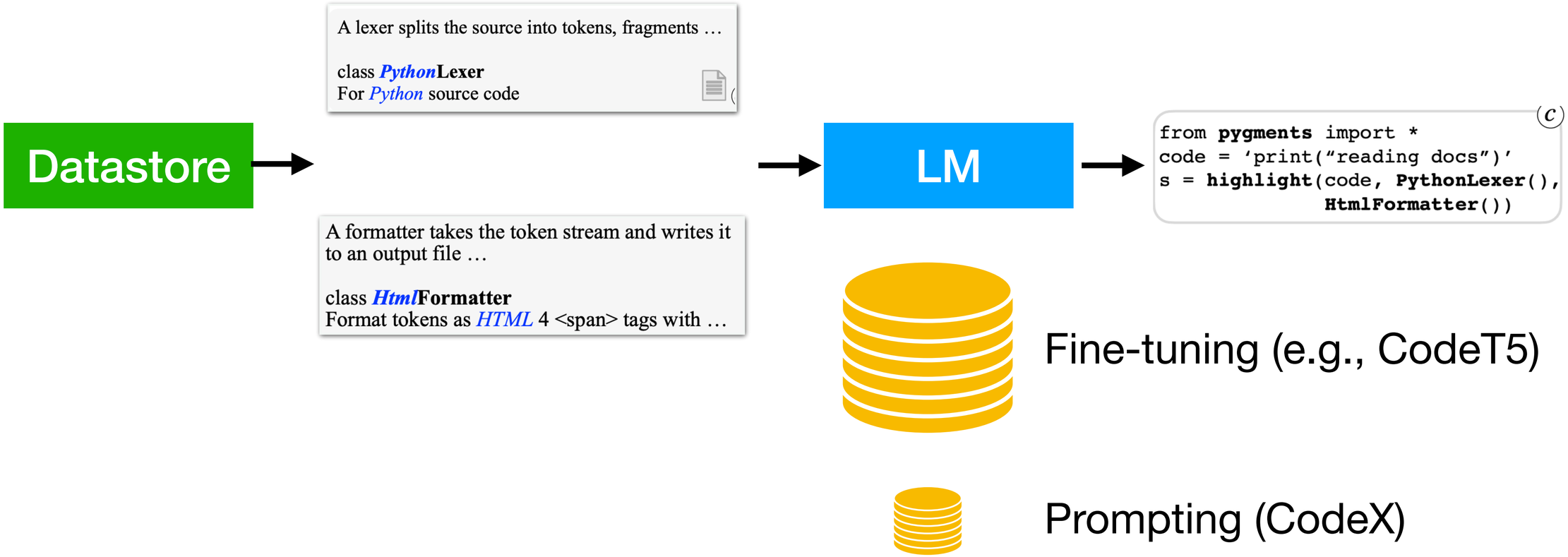
Zhou et al. 2023. "DocPrompting: Generating the Codes by Retrieving the Docs"

DocPrompting (Zhou et al., 2023)

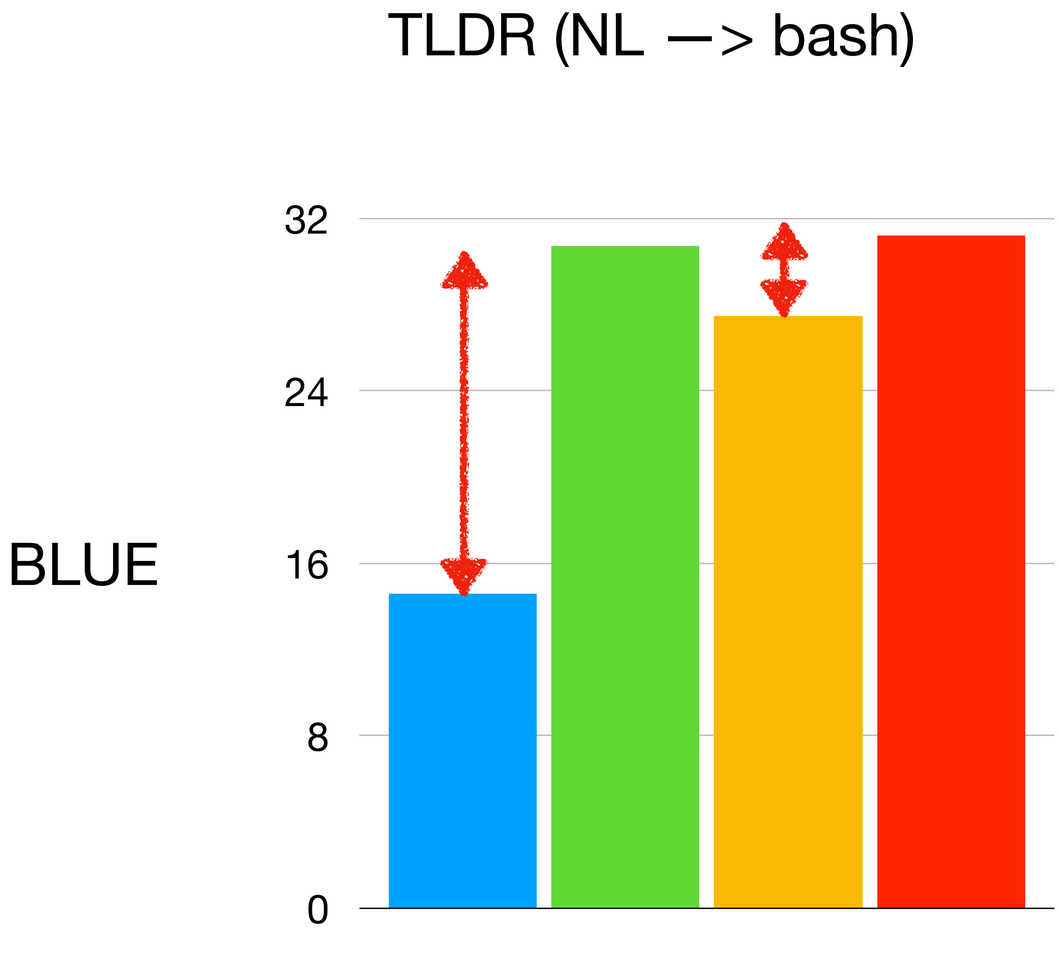


Oracle docs retrieved by simple text matching

DocPrompting (Zhou et al., 2023)

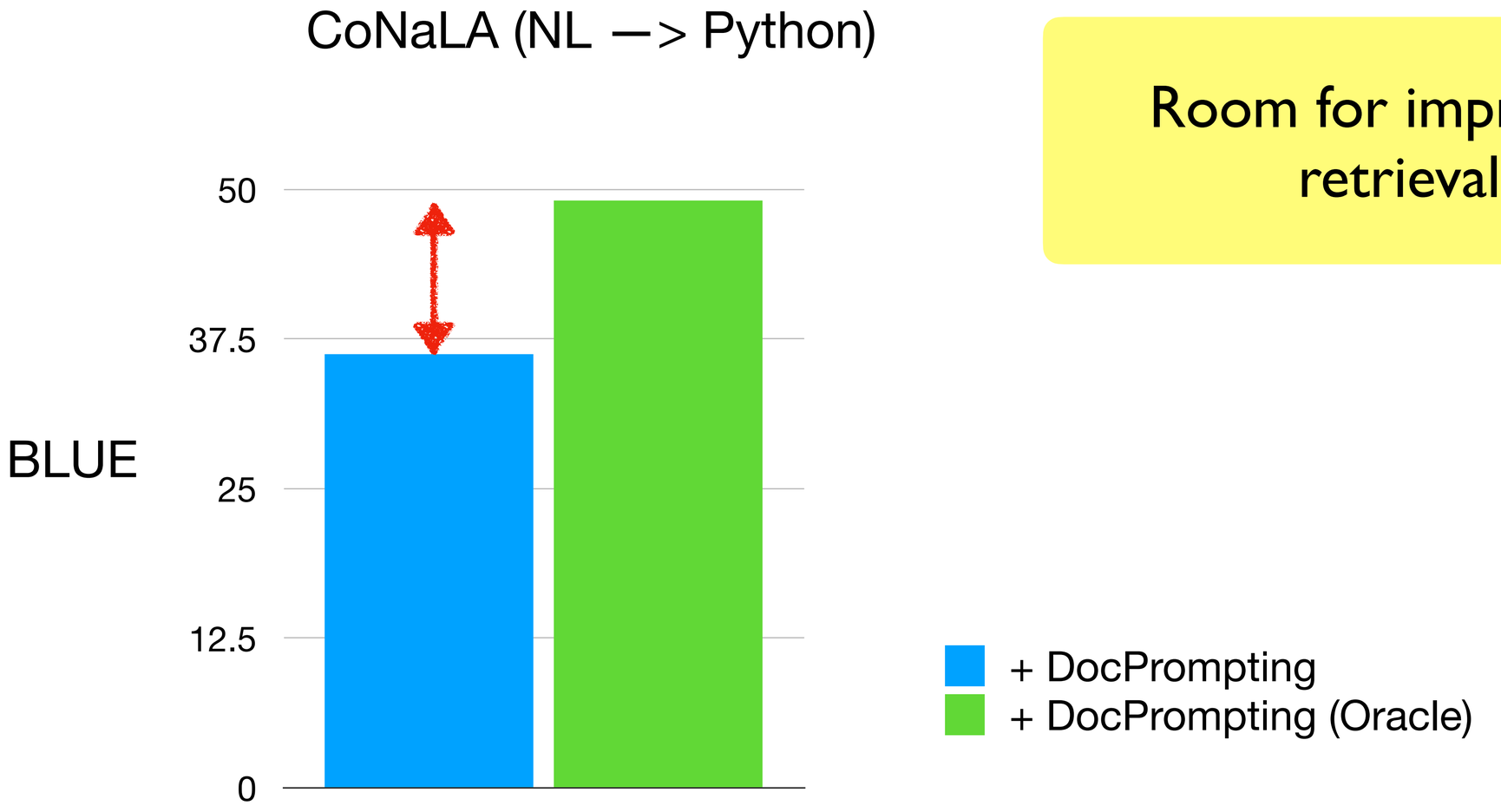


DocPrompting (Zhou et al., 2023)



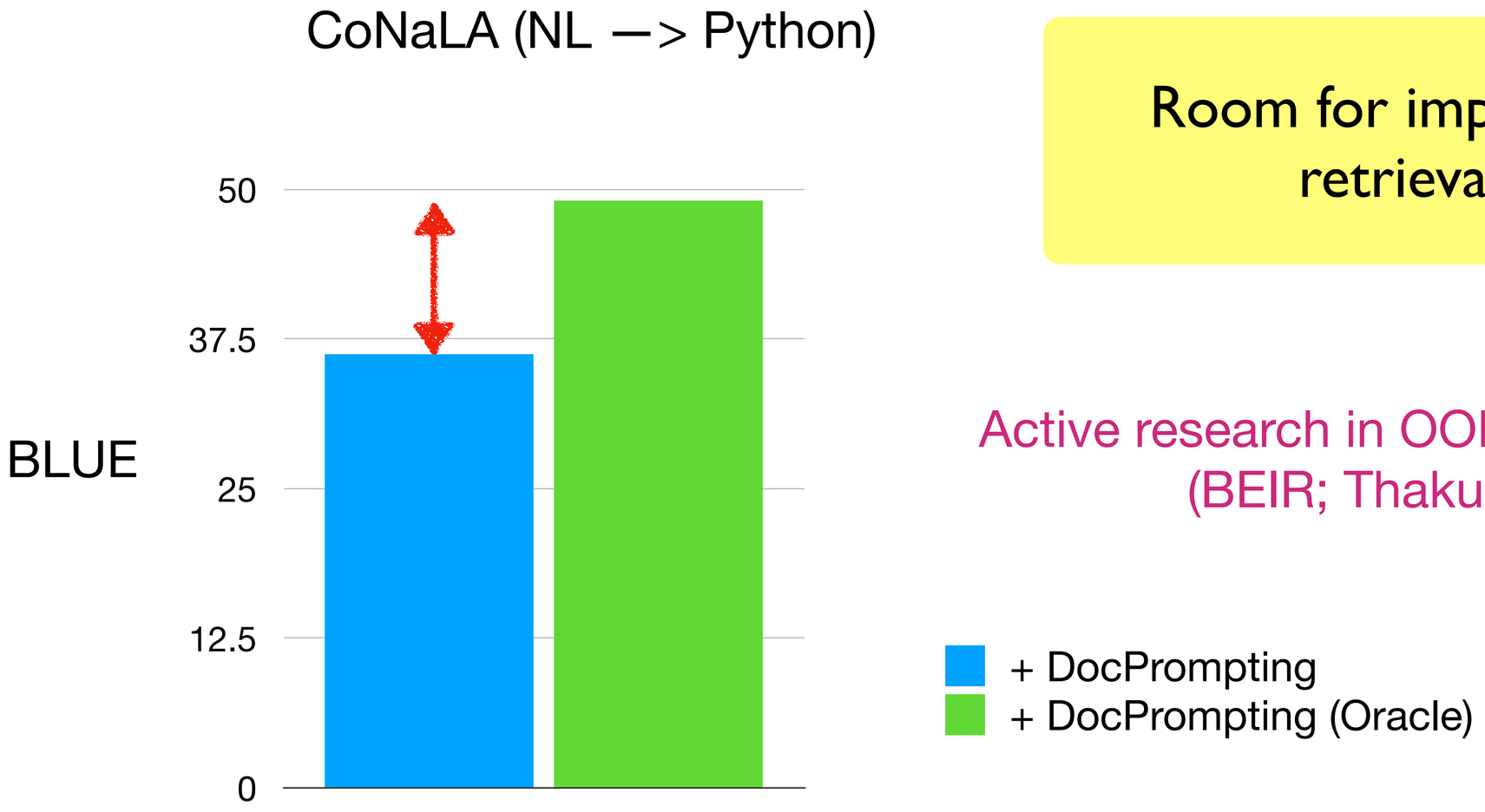
Large gain given by DocPrompting for both CodeT5 & CodeX

DocPrompting (Zhou et al., 2023)



Room for improvement for retrieval model

DocPrompting (Zhou et al., 2023)



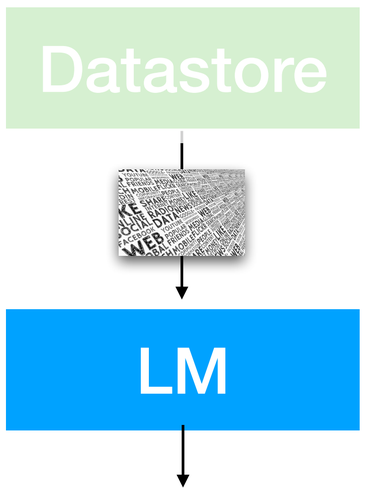
Room for improvement for retrieval model

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

Summary of downstream adaptations

	Target task	Adaptation ethod	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

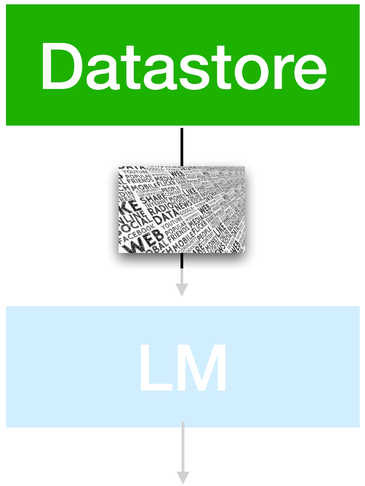
How to adapt a retrieval-based LM for a task



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

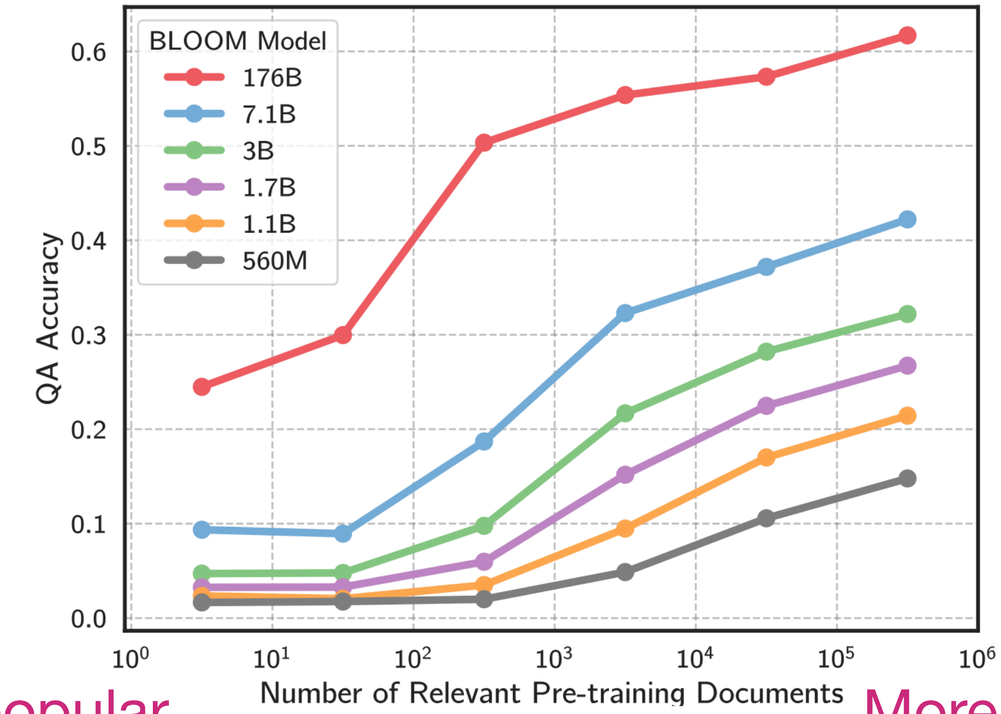
Parameter-efficiency

Privacy

Key effectiveness in downstream tasks

Long-tail

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



<— less popular

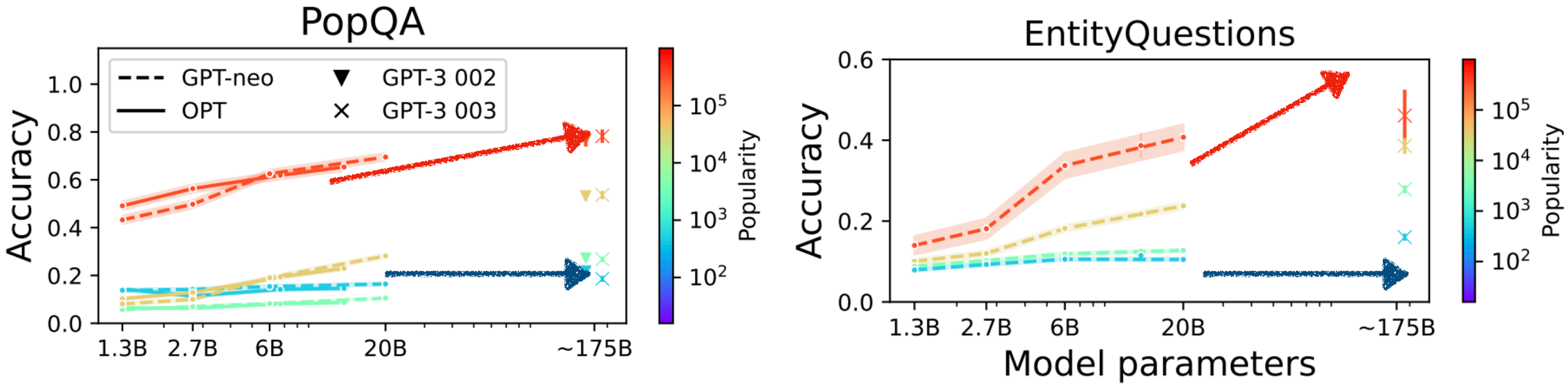
More popular —>

Kandpal et al. 2023. “Large language models struggle to learn long-tail knowledge”

Key effectiveness in downstream tasks

Long-tail

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

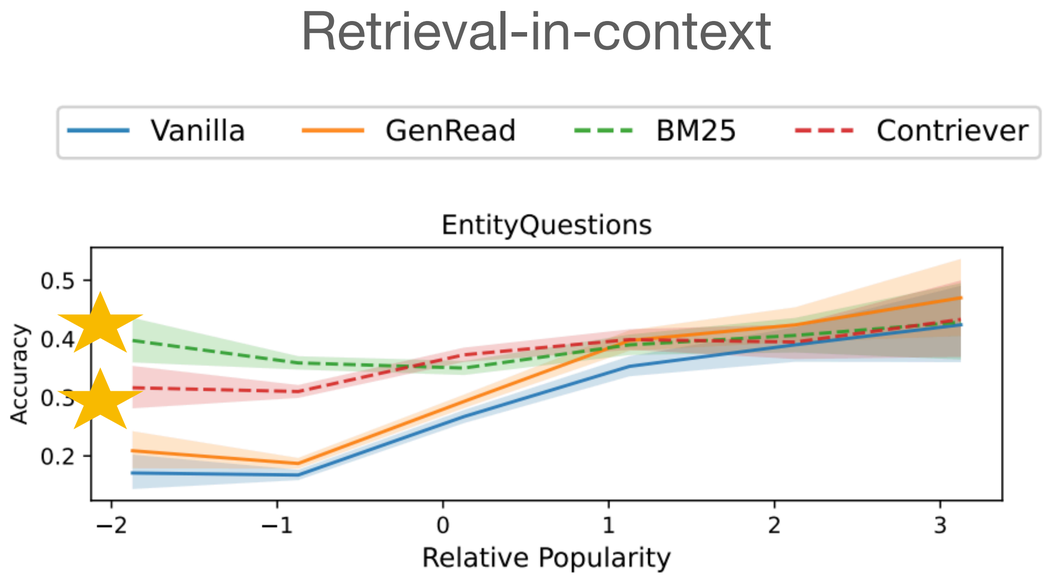


Mallen* and Asai* et al. 2023. “When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories”

Key effectiveness in downstream tasks

Long-tail

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



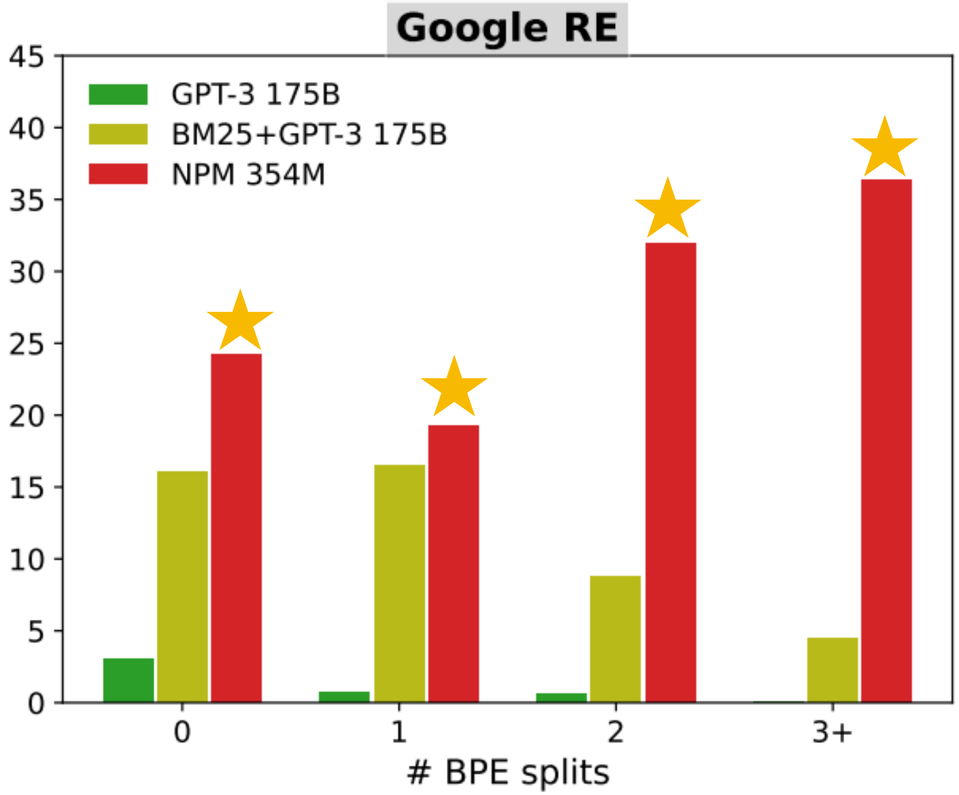
<— less popular More popular —>

Key effectiveness in downstream tasks

Long-tail

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

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



<— more popular

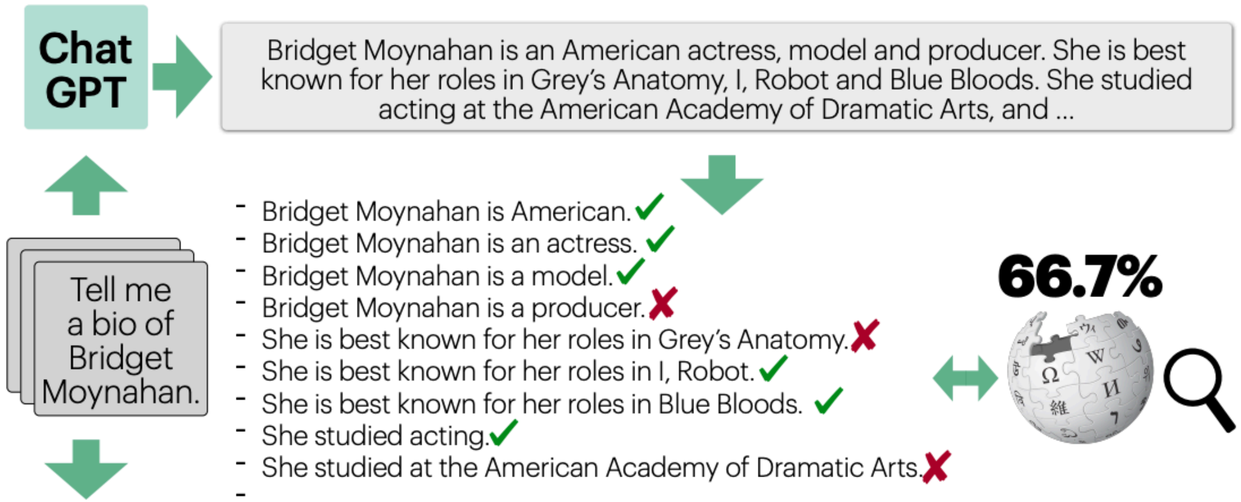
Less popular —>

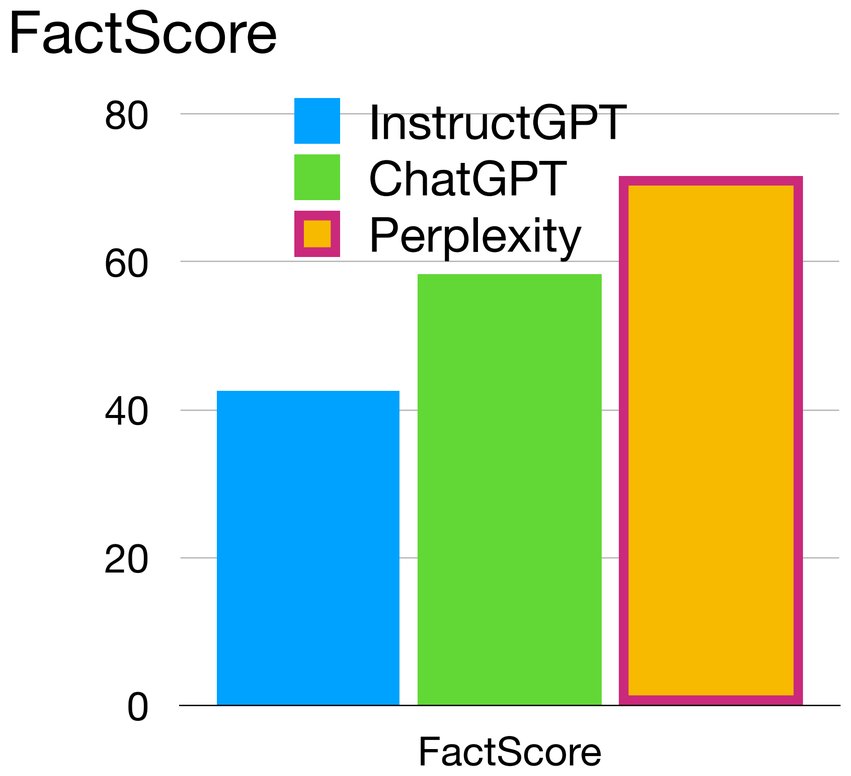
Min et al. 2023. “Nonparametric Masked Language Modeling”

Key effectiveness in downstream tasks

Long-tail

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





Min et al. 2023. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation"

Key effectiveness in downstream tasks

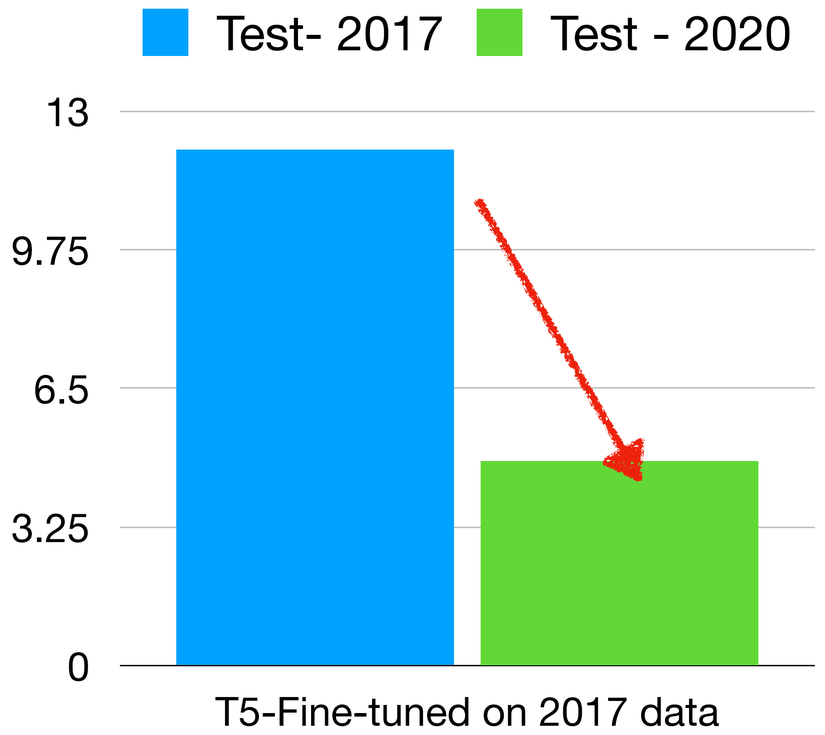
Update

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

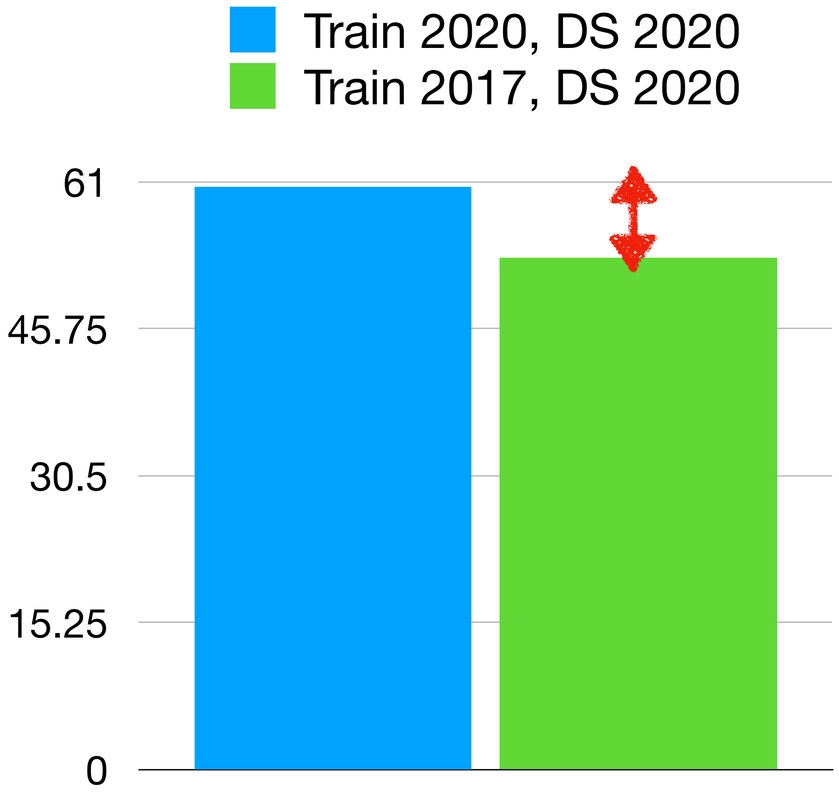
Update

We can simply swap the knowledge corpus to **adapt the temporal changes**, without any new 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

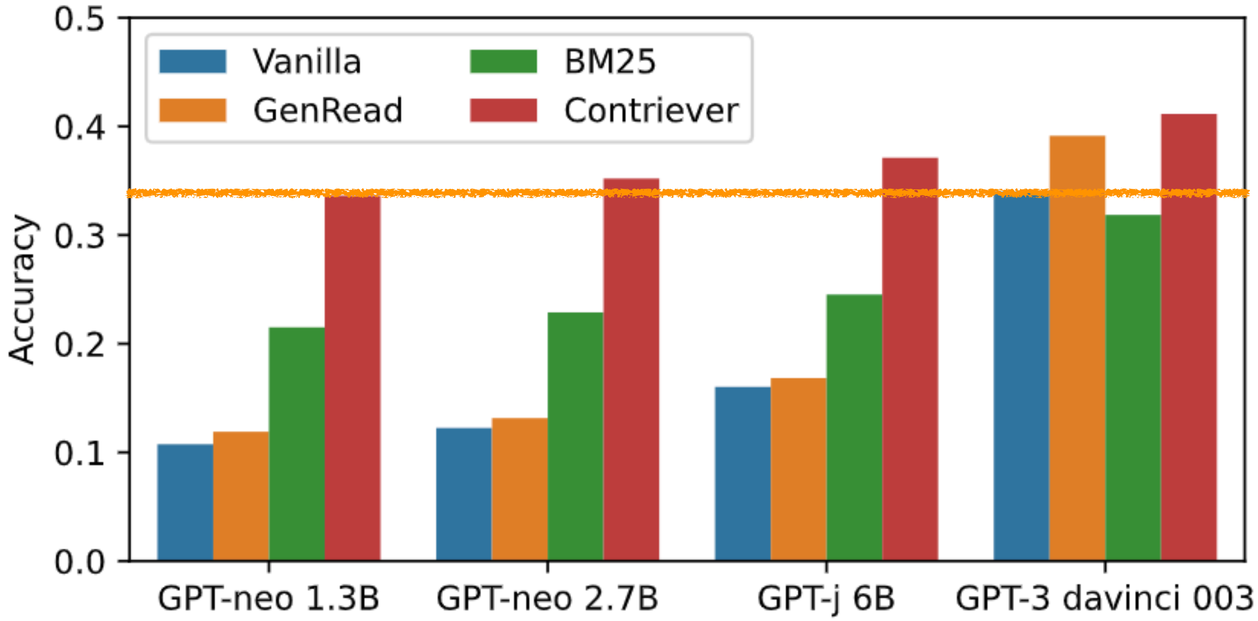


Key effectiveness in downstream tasks

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"

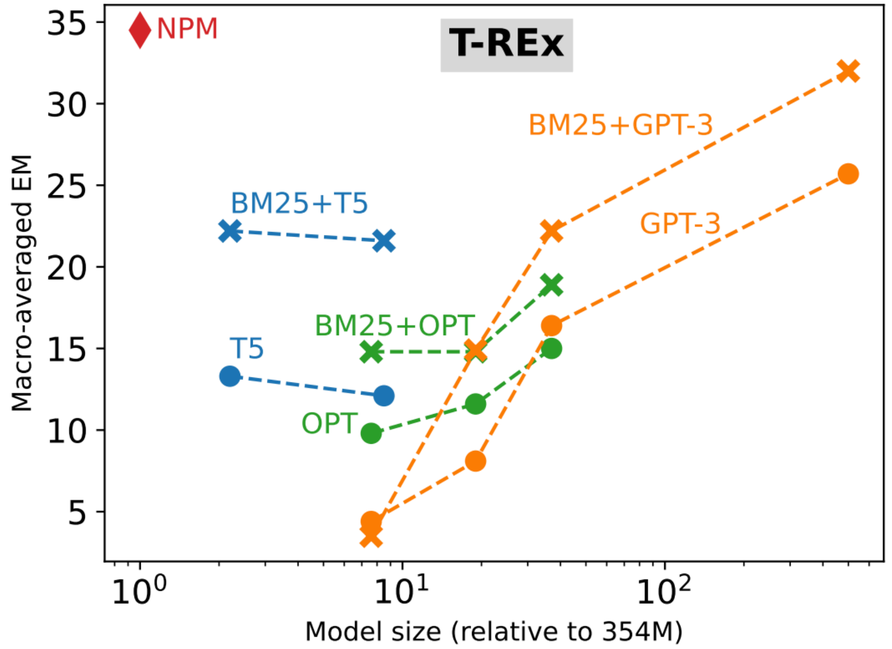
Key effectiveness in downstream tasks

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 ZS relation extraction.



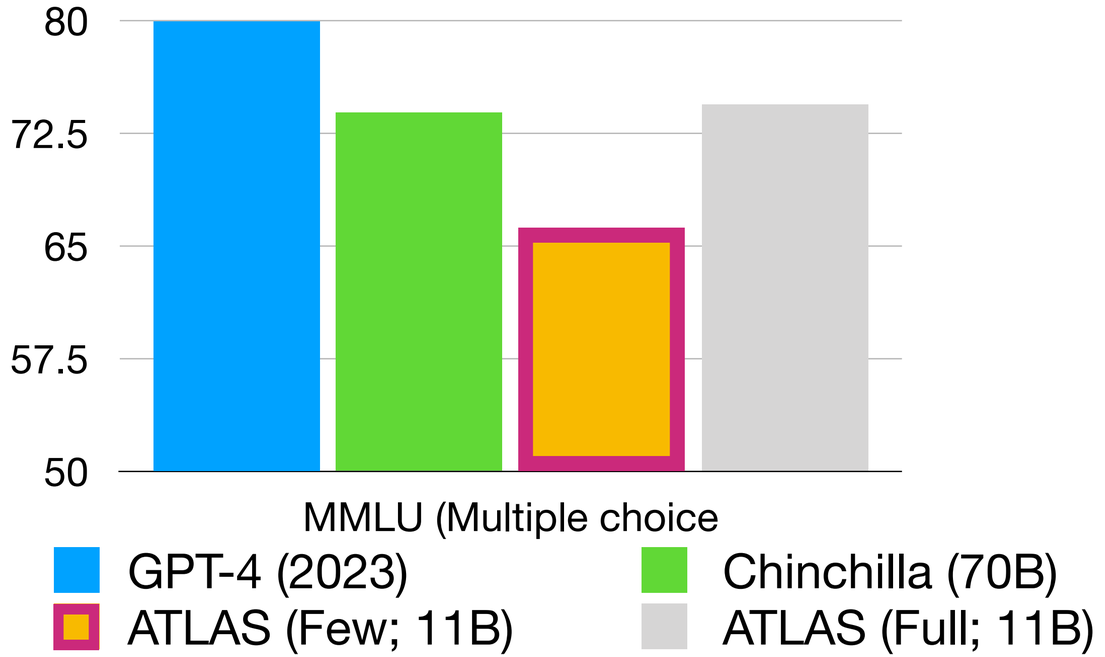
Min et al. 2023. “Nonparametric Masked Language Modeling”

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!

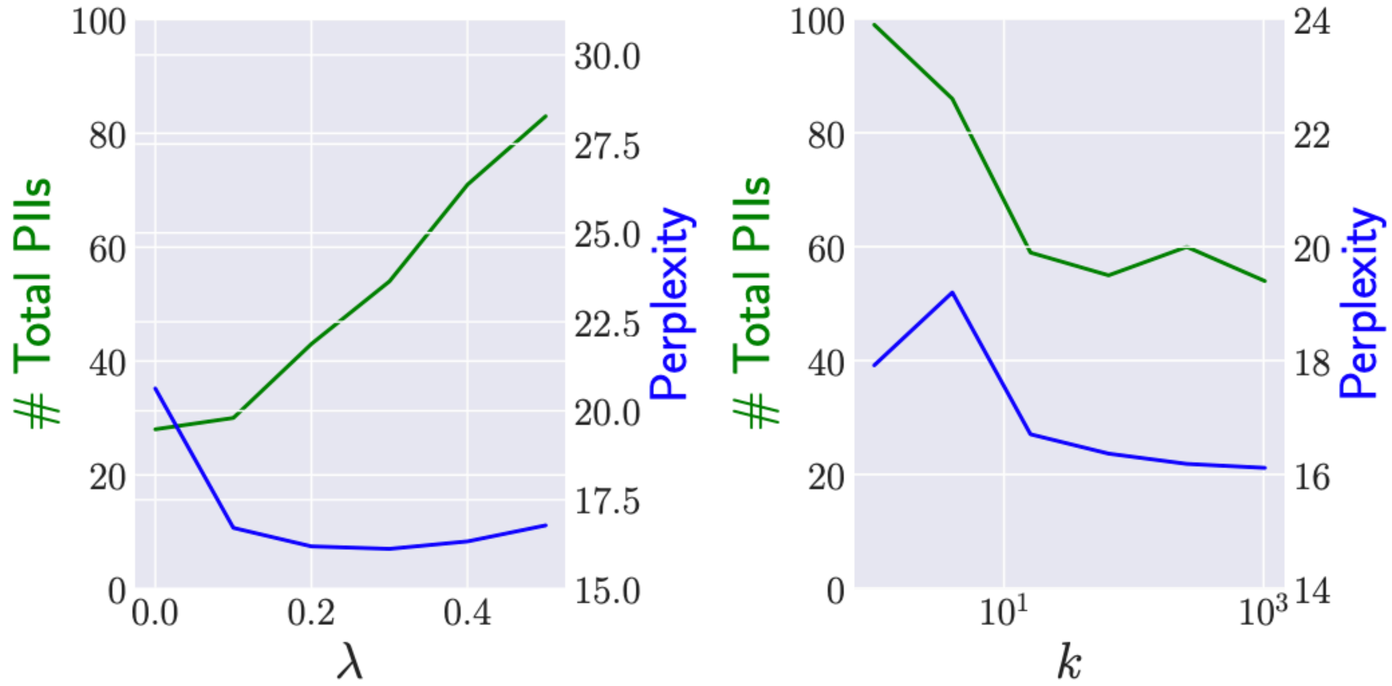


Izacard et al. 2022. “Few-shot learning with retrieval augmented language models”

Key effectiveness in downstream tasks

Privacy

Retrieval-based LMs enable us to mitigate privacy risks.

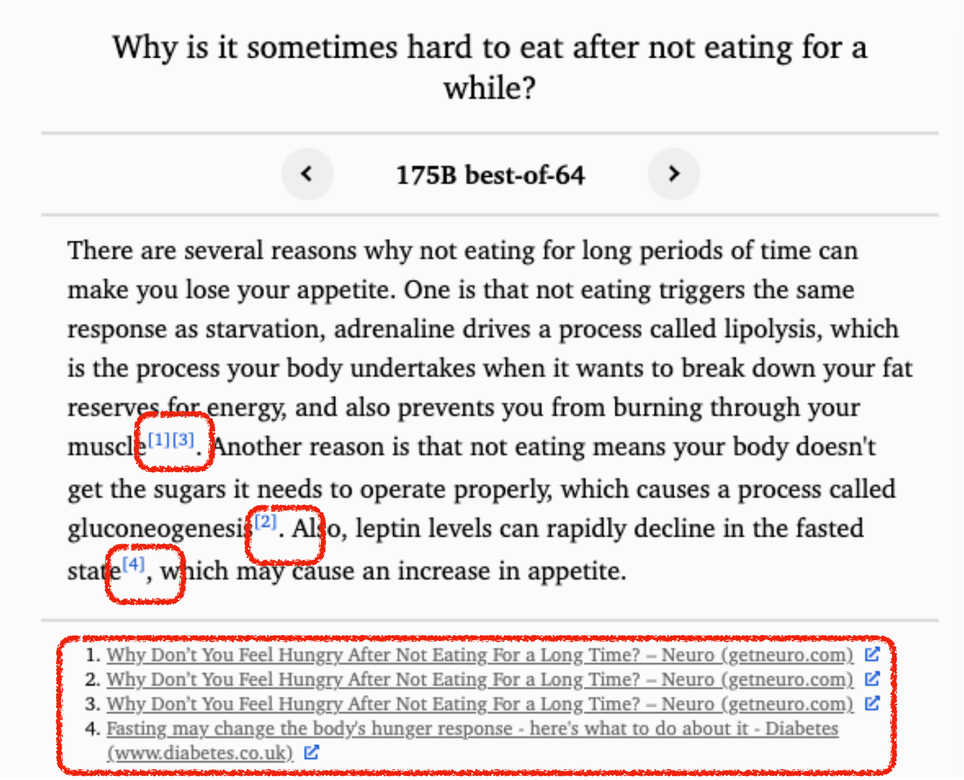


Huang et al. 2023. “Privacy Implications of Retrieval-Based Language Models”

Key effectiveness in downstream tasks

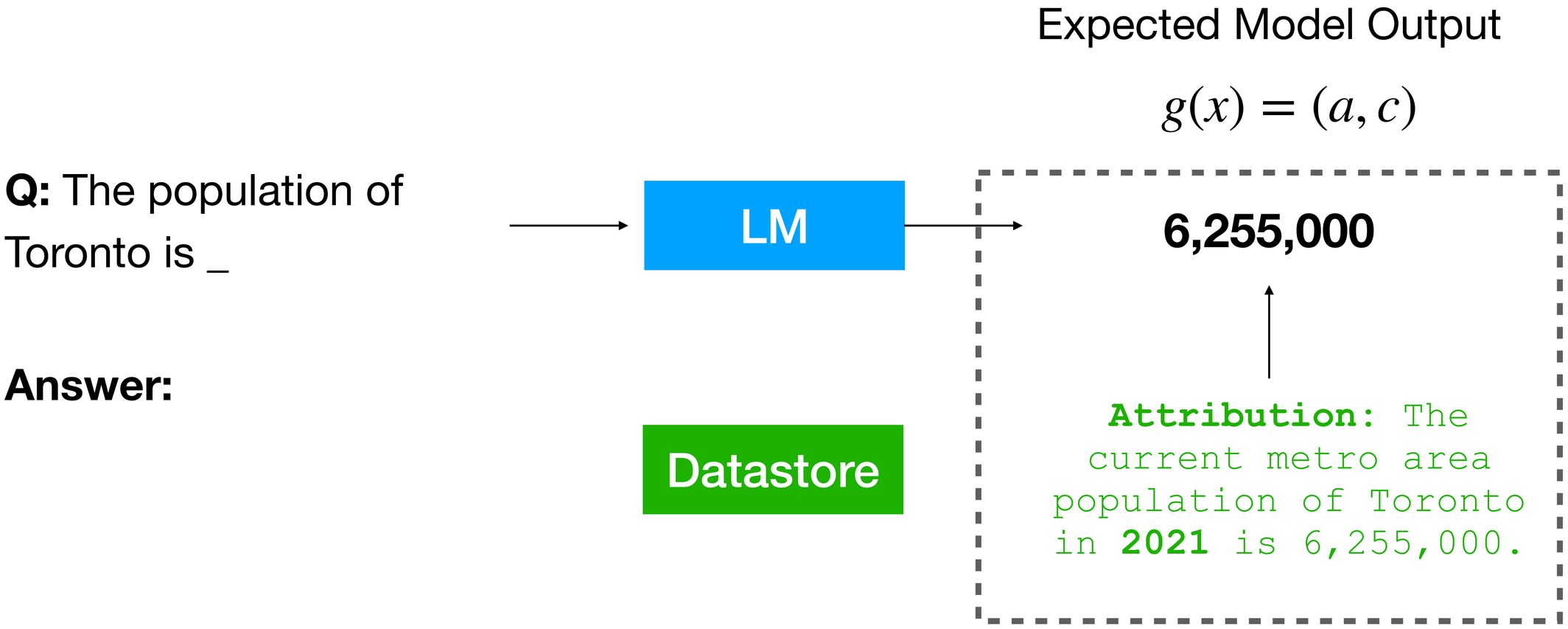
Verifiability

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



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

Attributions: AttributedQA (Bohnet et al., 2022)



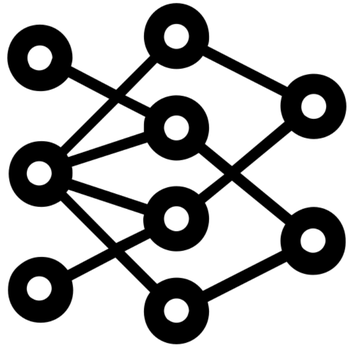
Bohnet et al. 2022. “Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models”

Attributions: AttributedQA (Bohnet et al., 2022)

Human Evaluation (AIS)

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

Automatic Evaluation (Auto AIS)

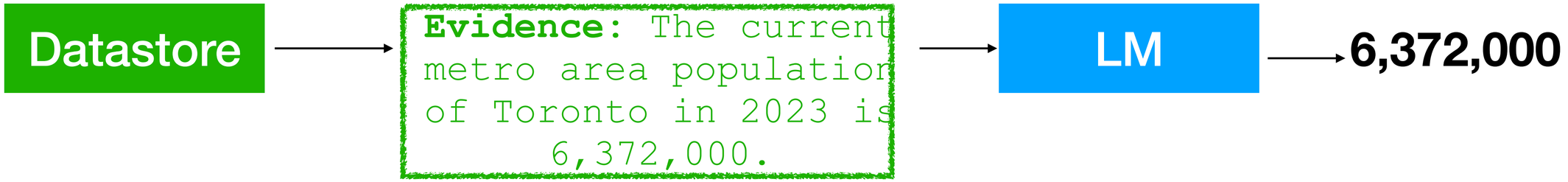


NLI model

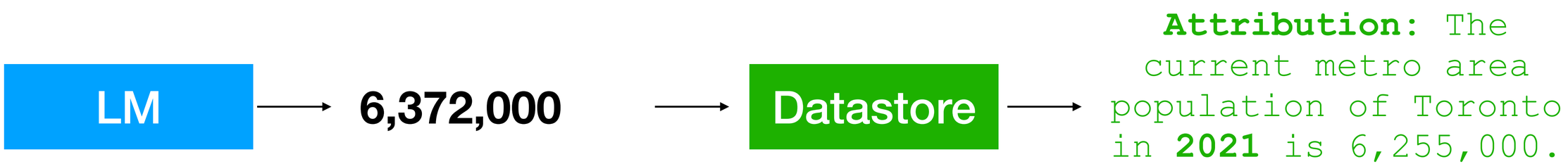
$$E^{\mathsf{A}}[g] = \frac{1}{n}\sum_{i=1}^{n} \texttt{AutoAIS}(x_i, g(x_i))$$

AttributedQA (Bohnet et al., 2022)

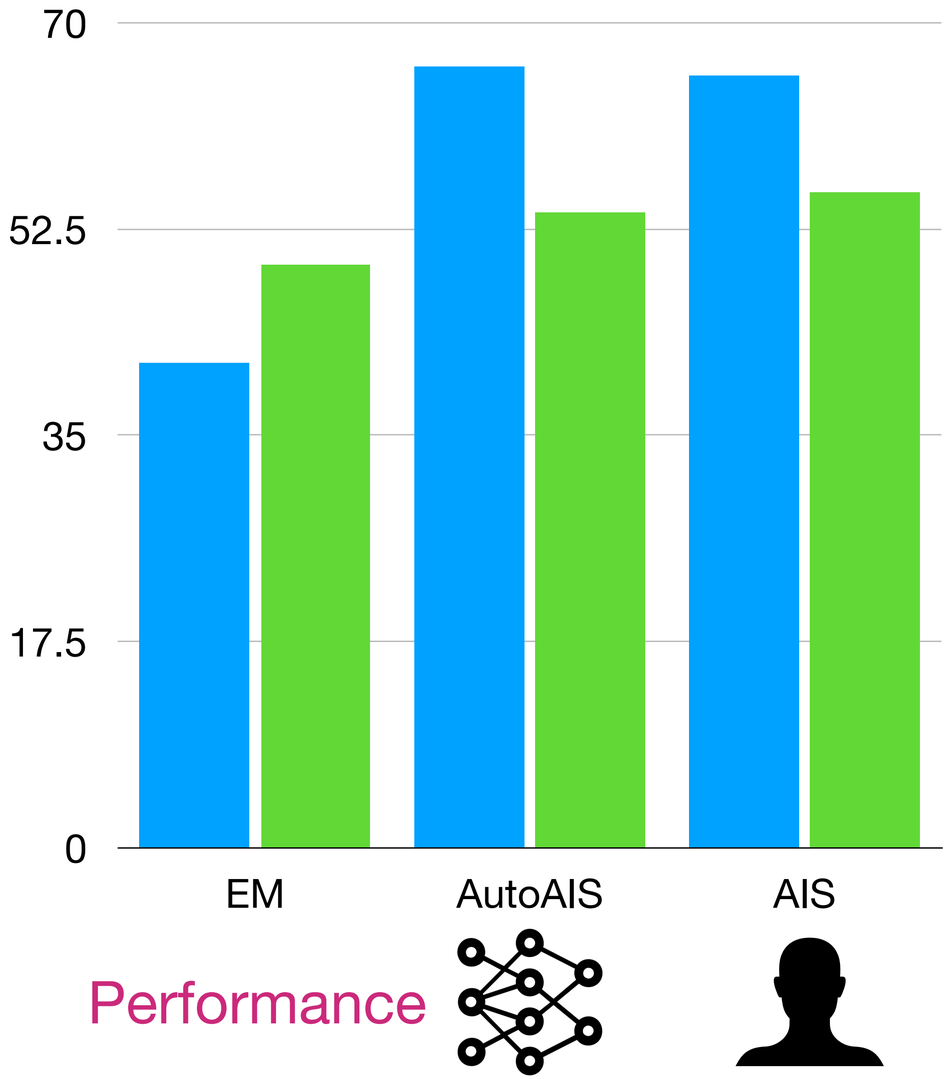
Retrieval-in-context



Post-hoc retrieval



AttributedQA (Bohnet et al., 2022)



Retrieval in context yields higher AIS than post-hoc retrieval

retrieval-in-context
posthoc retrieval

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”