Retrieval-based Language Models

Sewon Min

University of Washington

shmsw25.github.io

CPSC 488/588 • Fall 2023 • Yale University

Adapted from ACL 2023 Tutorial w/ Akari Asai, Zexuan Zhong, & Danqi Chen

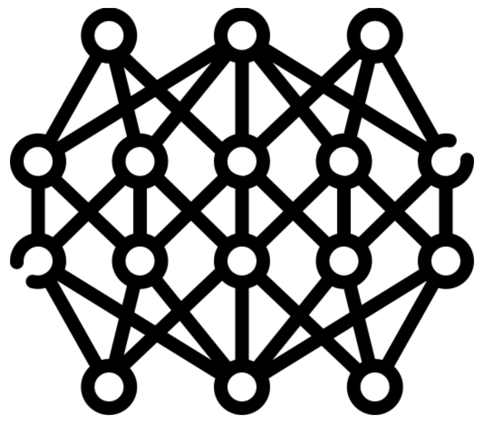
Language Models

$$P(x_n \quad x_1, x_2, \cdots, x_{n-1})$$

Language Models

$$P(x_n \;\; x_1, x_2, \cdots, x_{n-1})$$

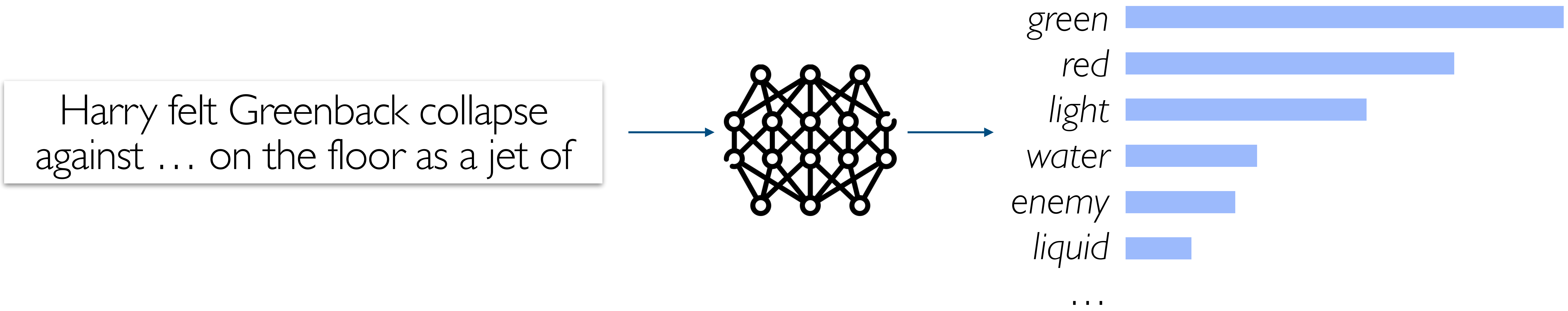
Harry felt Greenback collapse against … on the floor as a jet of



(This figure assumes autoregressive LMs, but the idea can be broadly extended to masked LMs)

Language Models

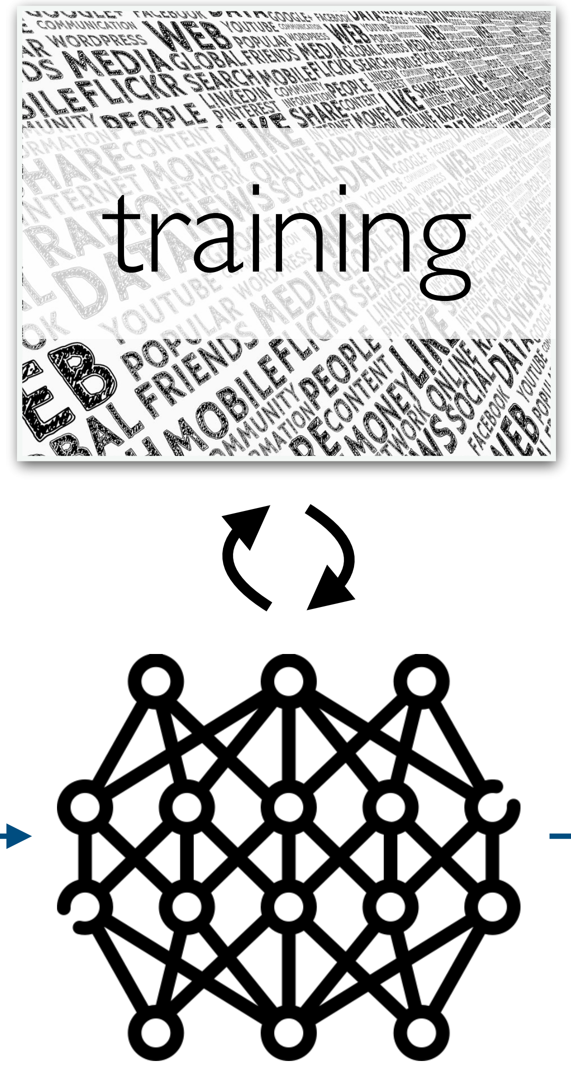
$$P(x_n \quad x_1, x_2, \cdots, x_{n-1})$$



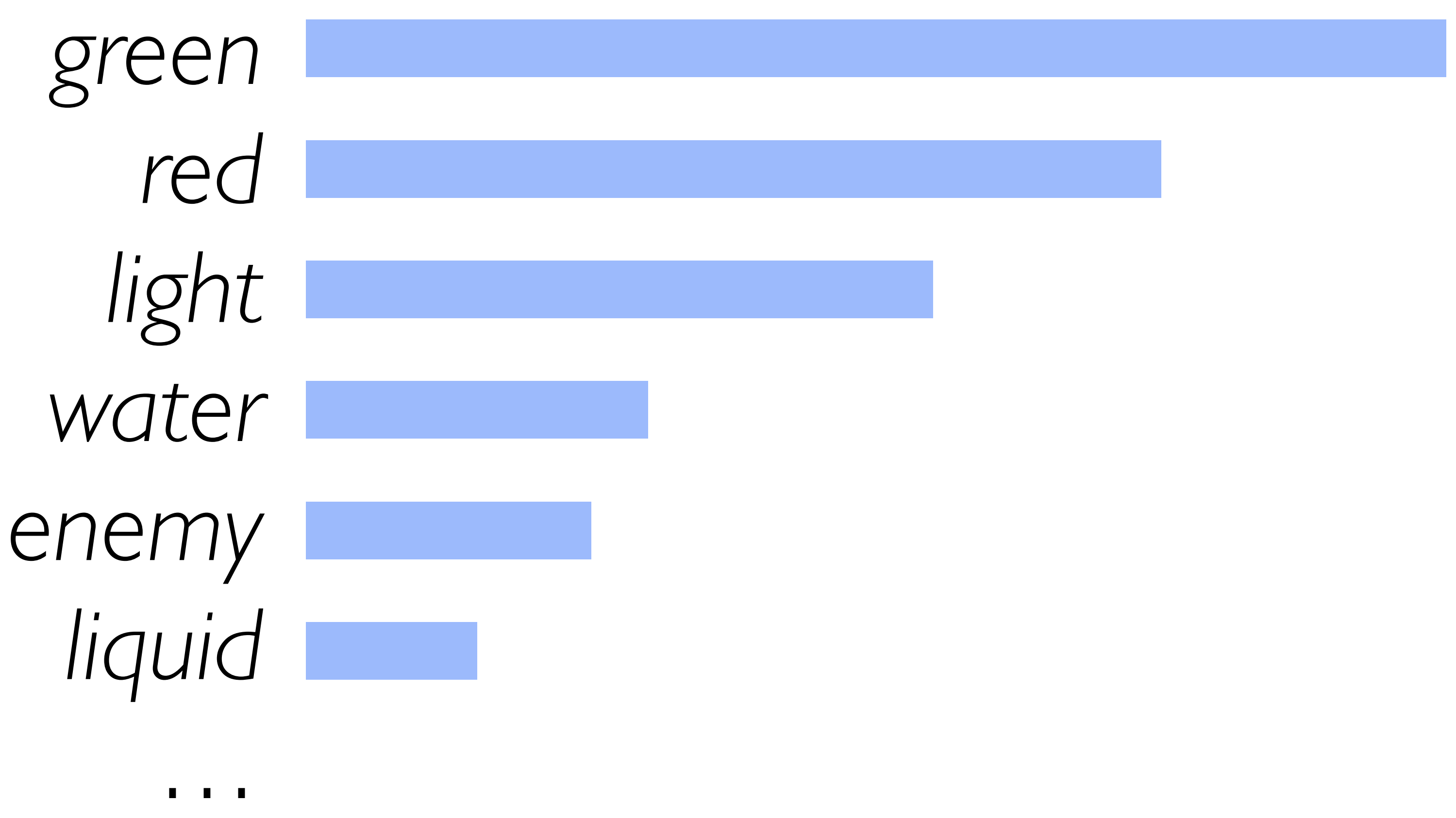
(This figure assumes autoregressive LMs, but the idea can be broadly extended to masked LMs)

Language Models

$$P(x_n \;\; x_1, x_2, \cdots, x_{n-1})$$



Harry felt Greenback collapse against … on the floor as a jet of



(This figure assumes autoregressive LMs, but the idea can be broadly extended to masked LMs)

Retrieval-based language models (LMs)

(also called semiparametric or nonparametric LMs)

(This figure assumes autoregressive LMs, but the idea can be broadly extended to masked LMs)

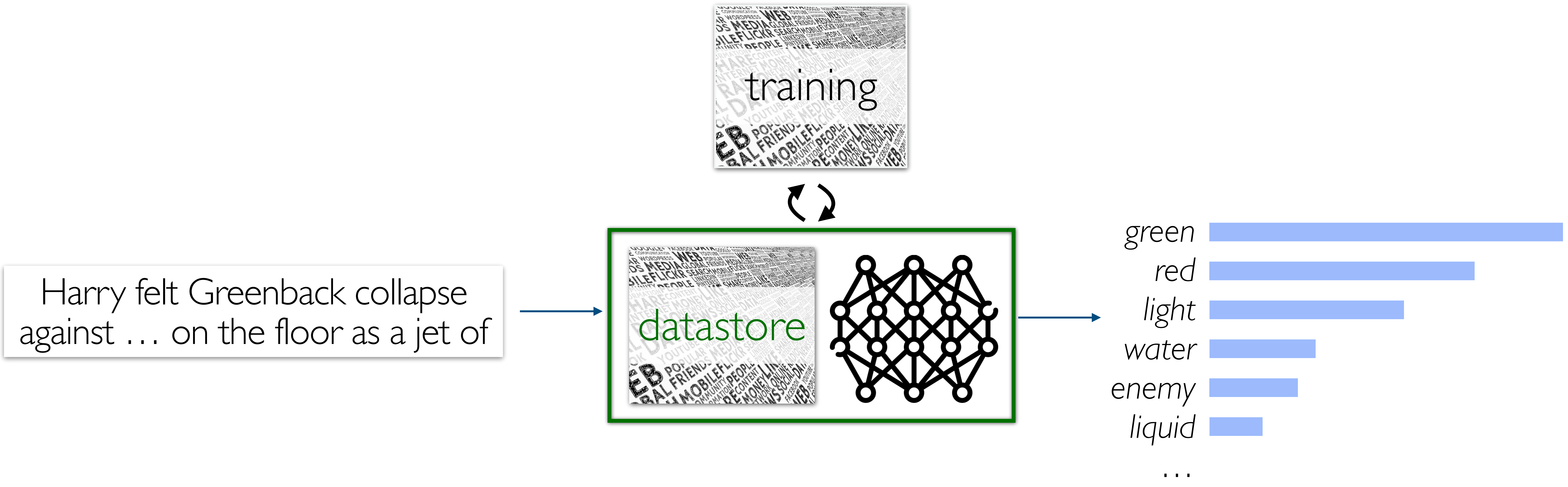
Retrieval-based language models (LMs)

(also called semiparametric or nonparametric LMs)

(This figure assumes autoregressive LMs, but the idea can be broadly extended to masked LMs)

Retrieval-based language models (LMs)

(also called semiparametric or nonparametric LMs)



(This figure assumes autoregressive LMs, but the idea can be broadly extended to masked LMs)

Overview

Why Retrieval-based LMs?

Tell me about Meta Platform.

ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Overview

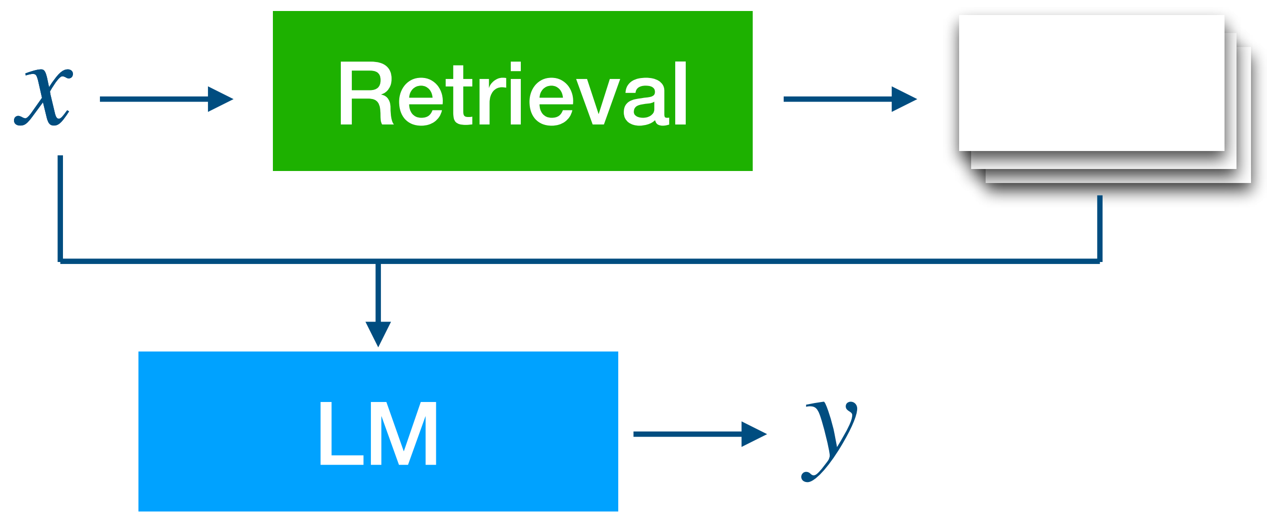
Why Retrieval-based LMs?

Tell me about Meta Platform.

ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Retrieval Augmentation



Overview

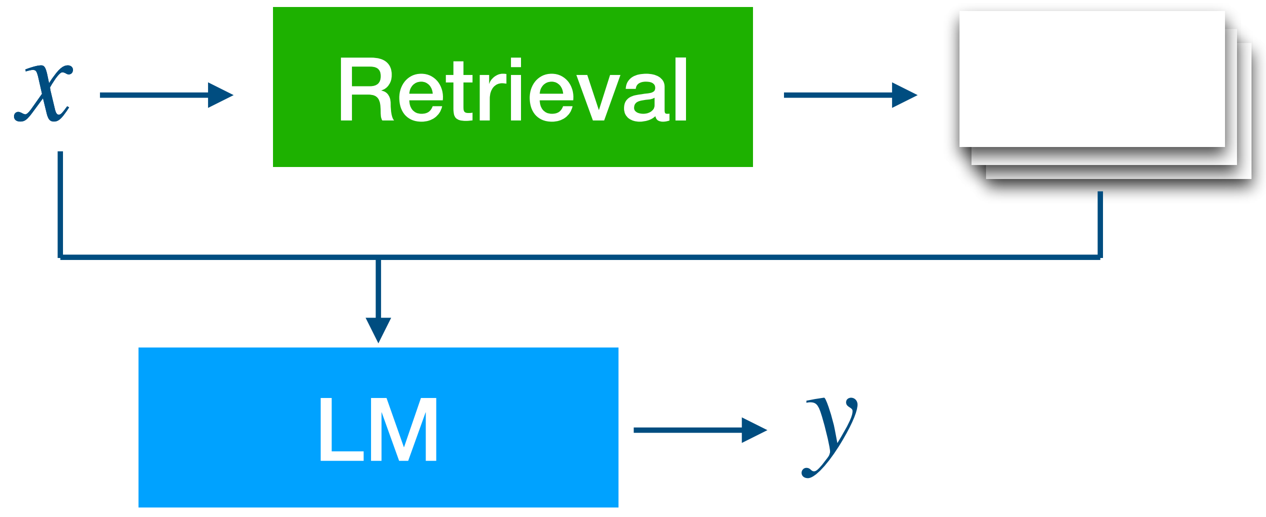
Why Retrieval-based LMs?

Tell me about Meta Platform.

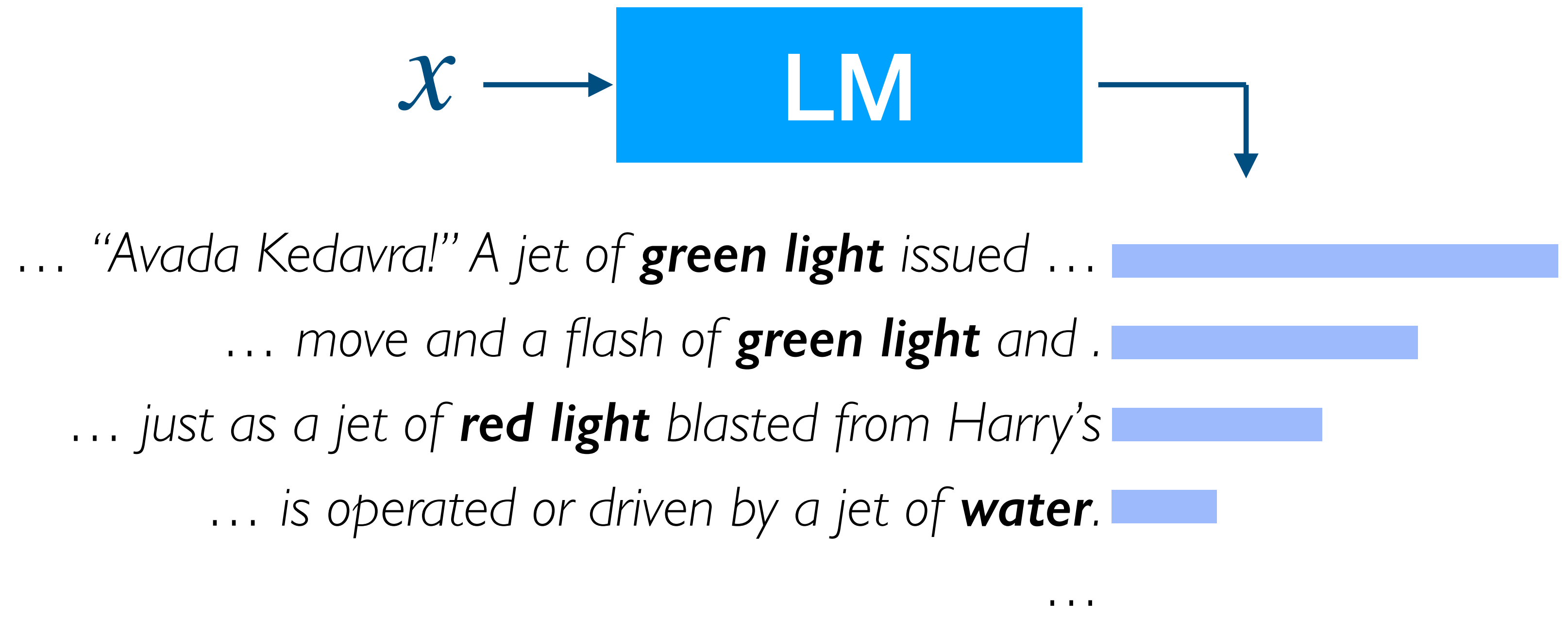
ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Retrieval Augmentation



New Retrieval-based LMs



Overview

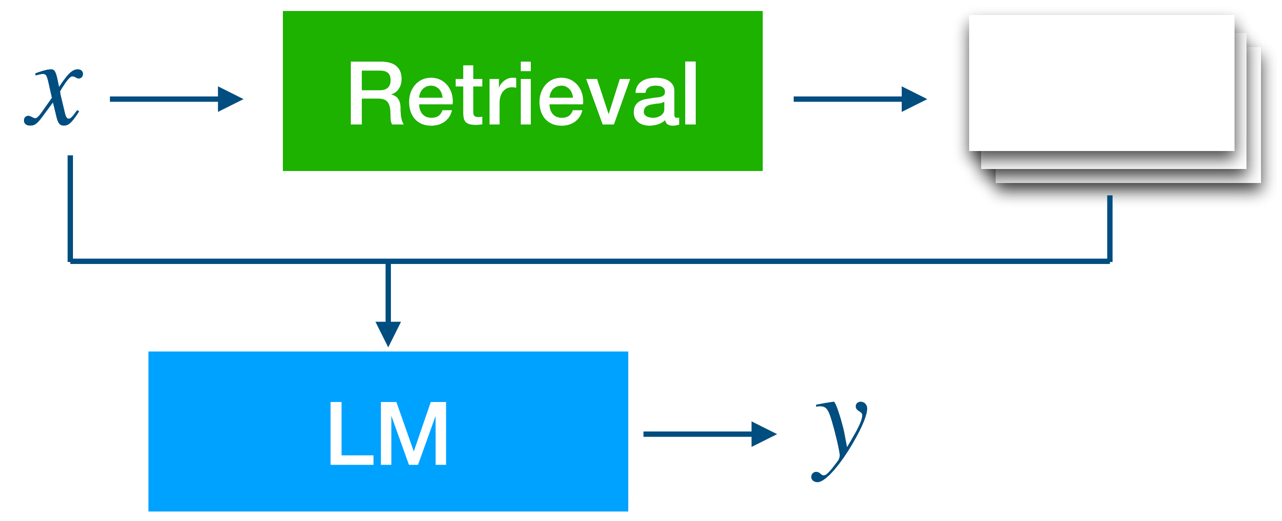
Why Retrieval-based LMs?

Tell me about Meta Platform.

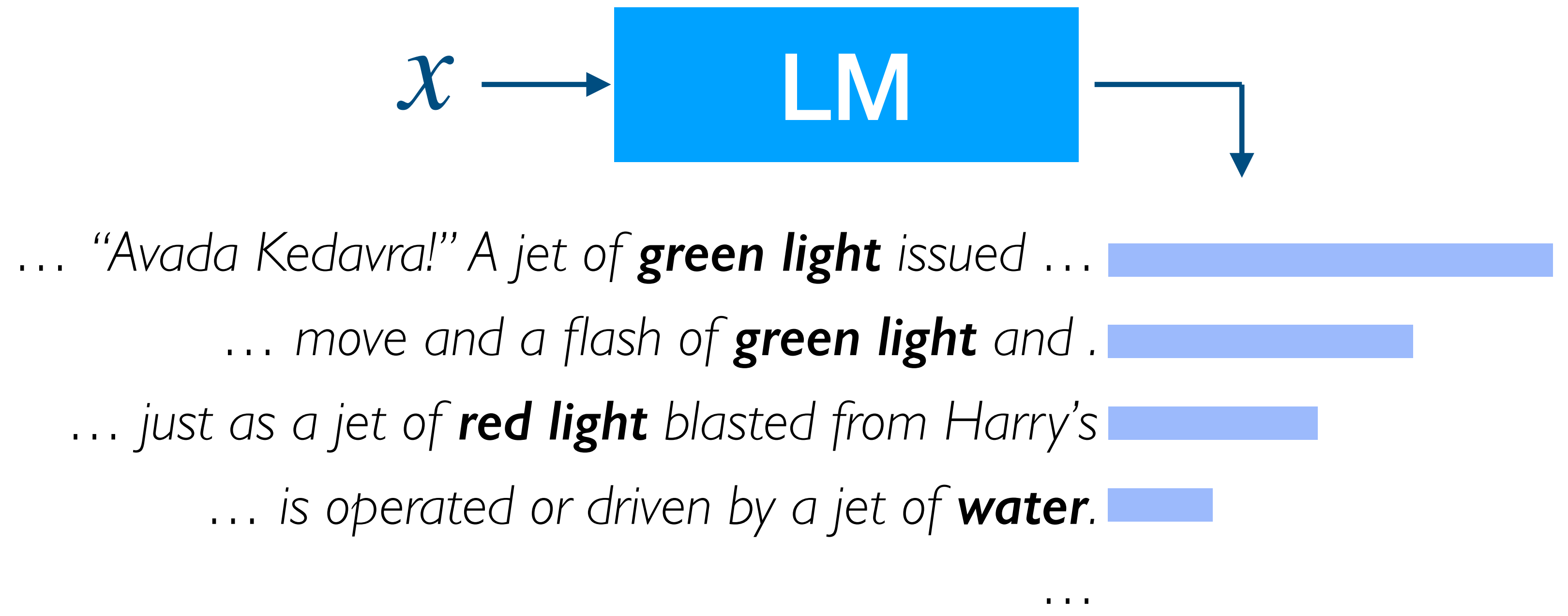
ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Retrieval Augmentation



New Retrieval-based LMs



Open Problems

Scaling **datastore** not just parameters?

Overview

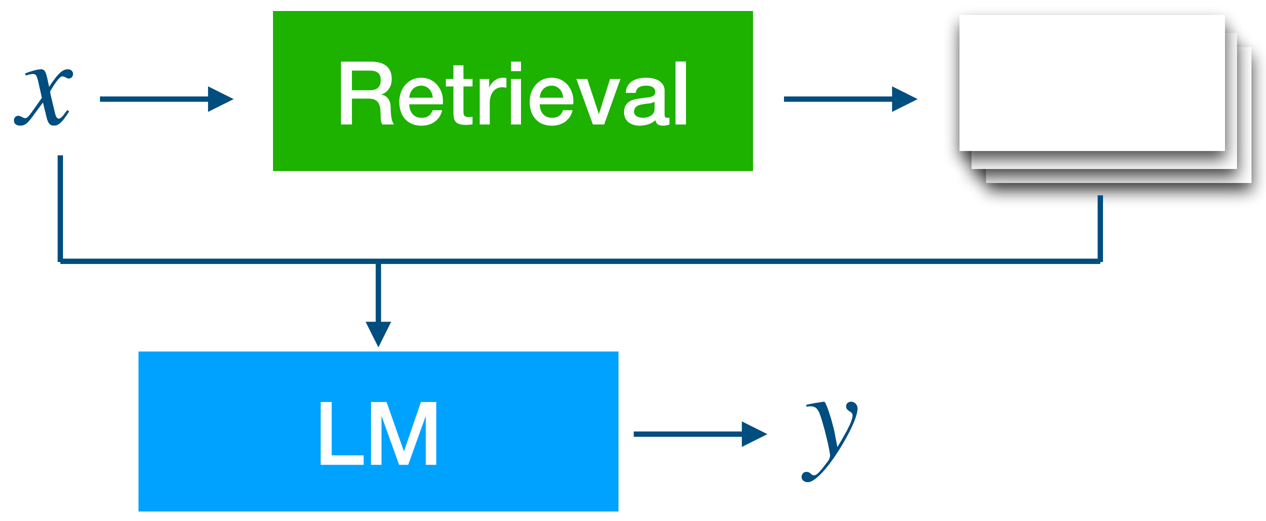
Why Retrieval-based LMs?

Tell me about Meta Platform.

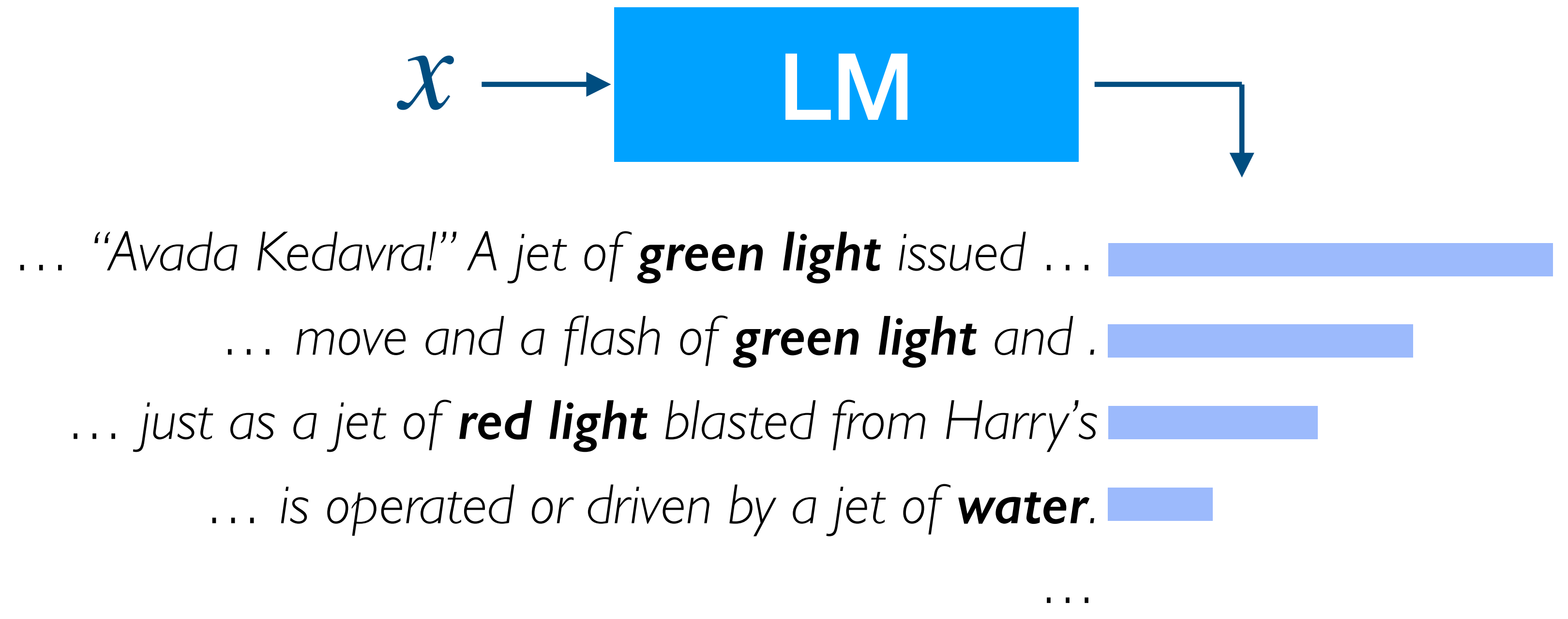
ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Retrieval Augmentation



New Retrieval-based LMs



Open Problems

Scaling **datastore** not just parameters?

Why Retrieval-based LMs?

New dimension in data use & better at long-tail

Can grow & update w/o additional training

Provide data attribution

Why Retrieval-based LMs?

New dimension in data use & better at long-tail

Can grow & update w/o additional training

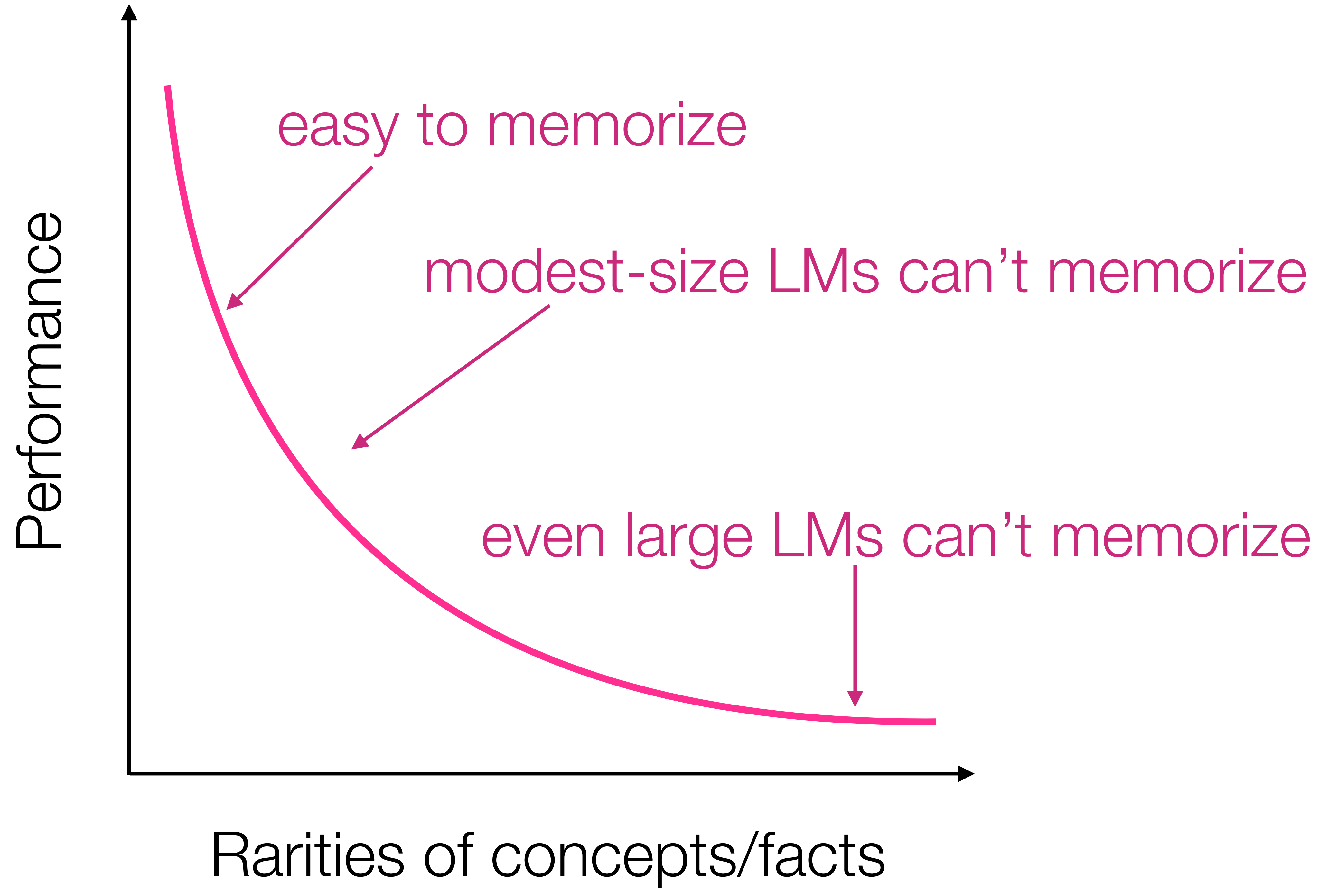
Provide data attribution

Why Retrieval-based LMs?

New dimension in data use & better at long-tail

Can grow & update w/o additional training

Provide data attribution

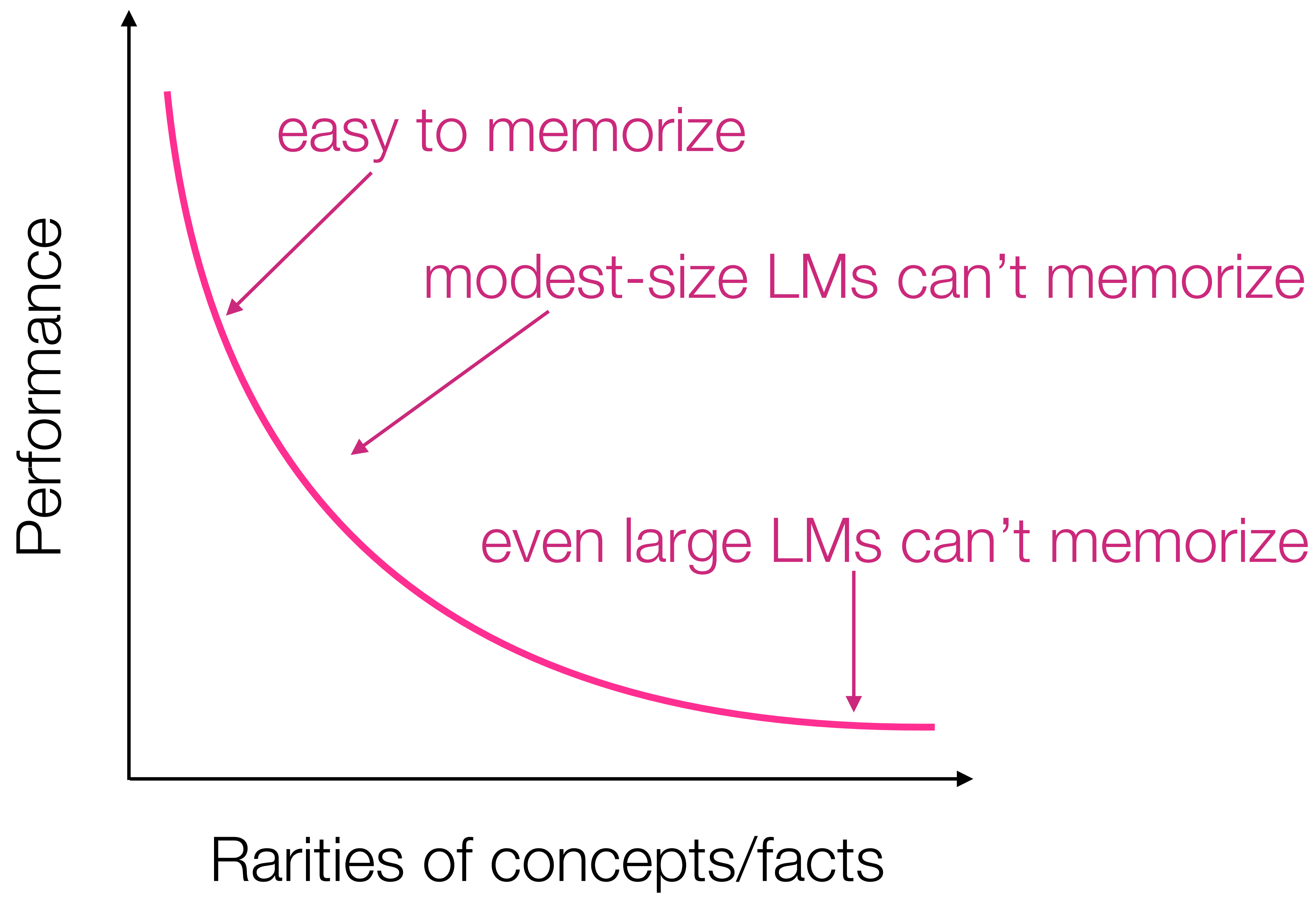


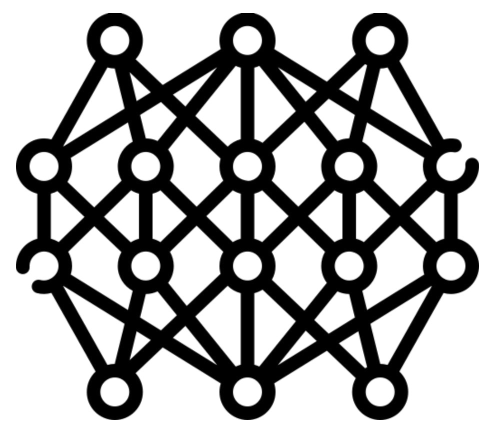
Why Retrieval-based LMs?

New dimension in data use & better at long-tail

Can grow & update w/o additional training

Provide data attribution





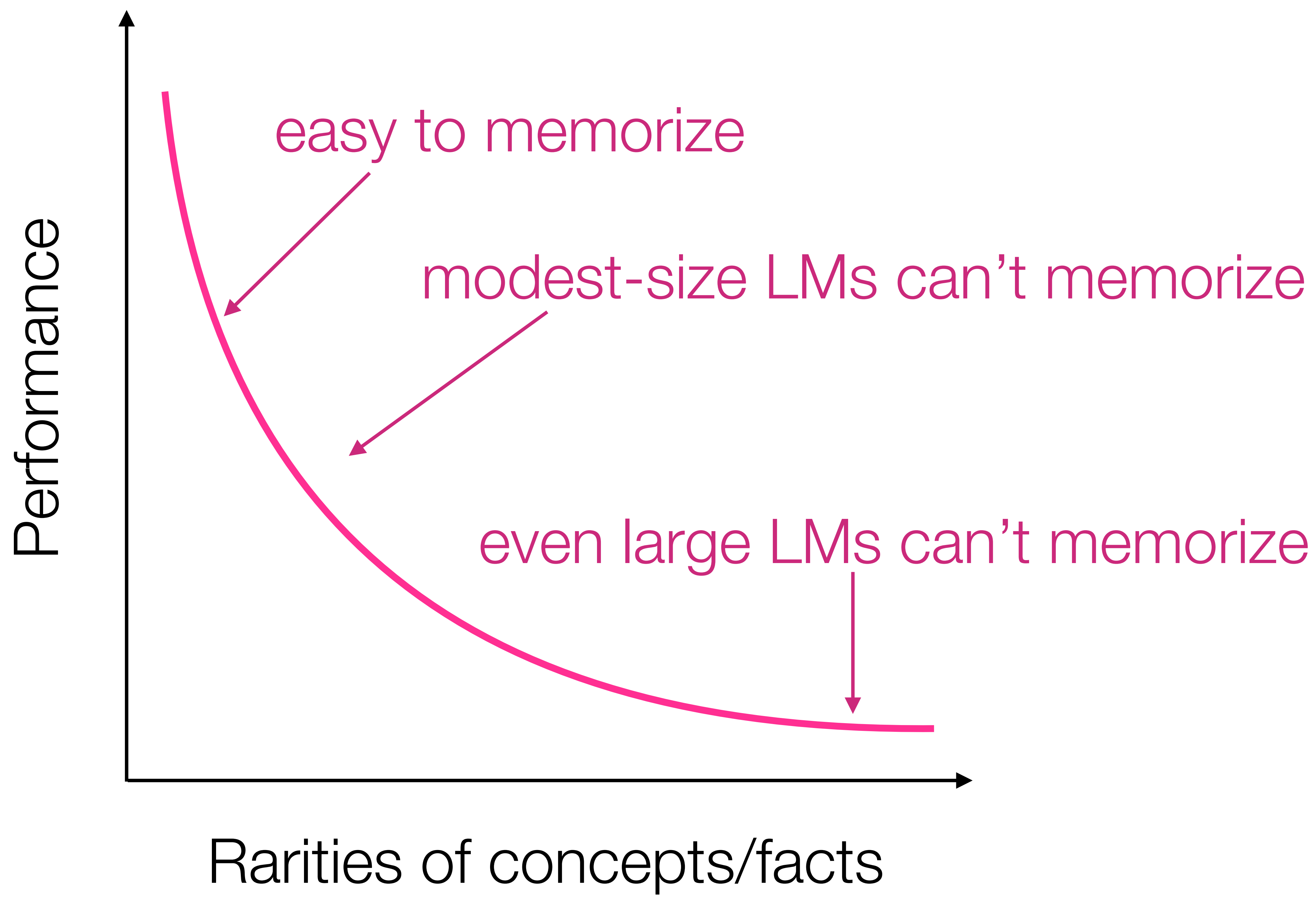
Standard LMs: Need to remember everything

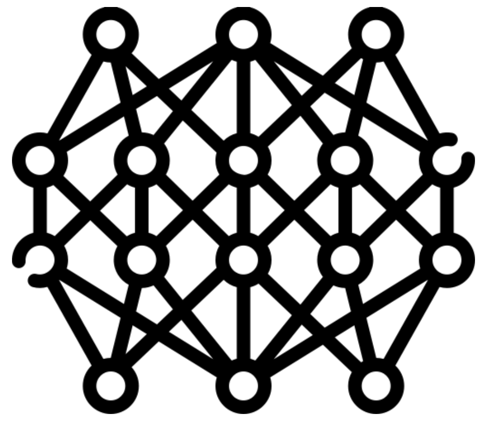
Why Retrieval-based LMs?

New dimension in data use & better at long-tail

Can grow & update w/o additional training

Provide data attribution





Standard LMs: Need to remember everything

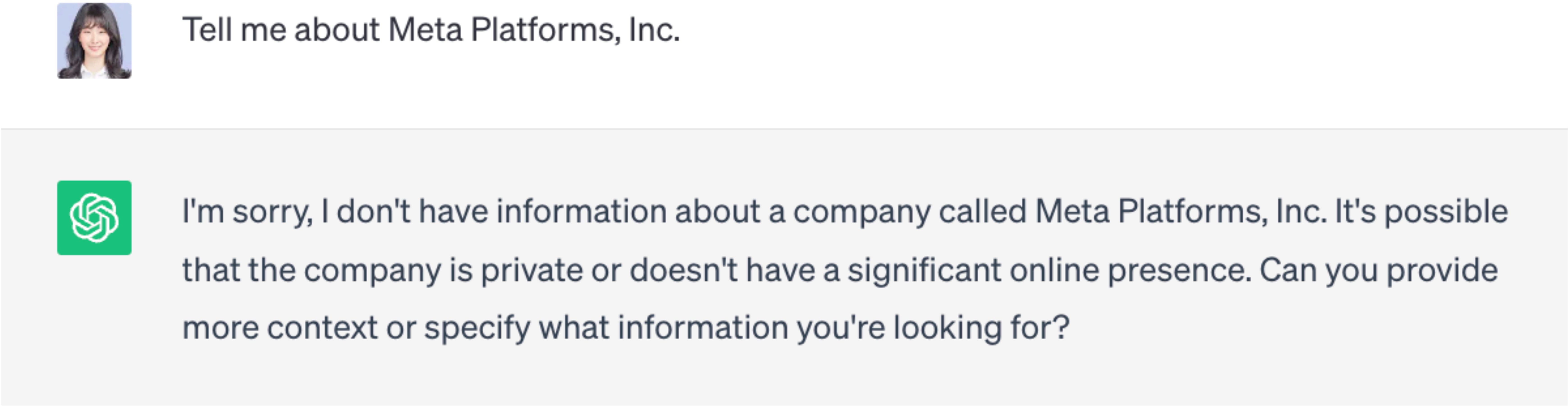
Retrieval-based LMs: Can look-up anytime

Why Retrieval-based LMs?

New dimension in data use & better at long-tail

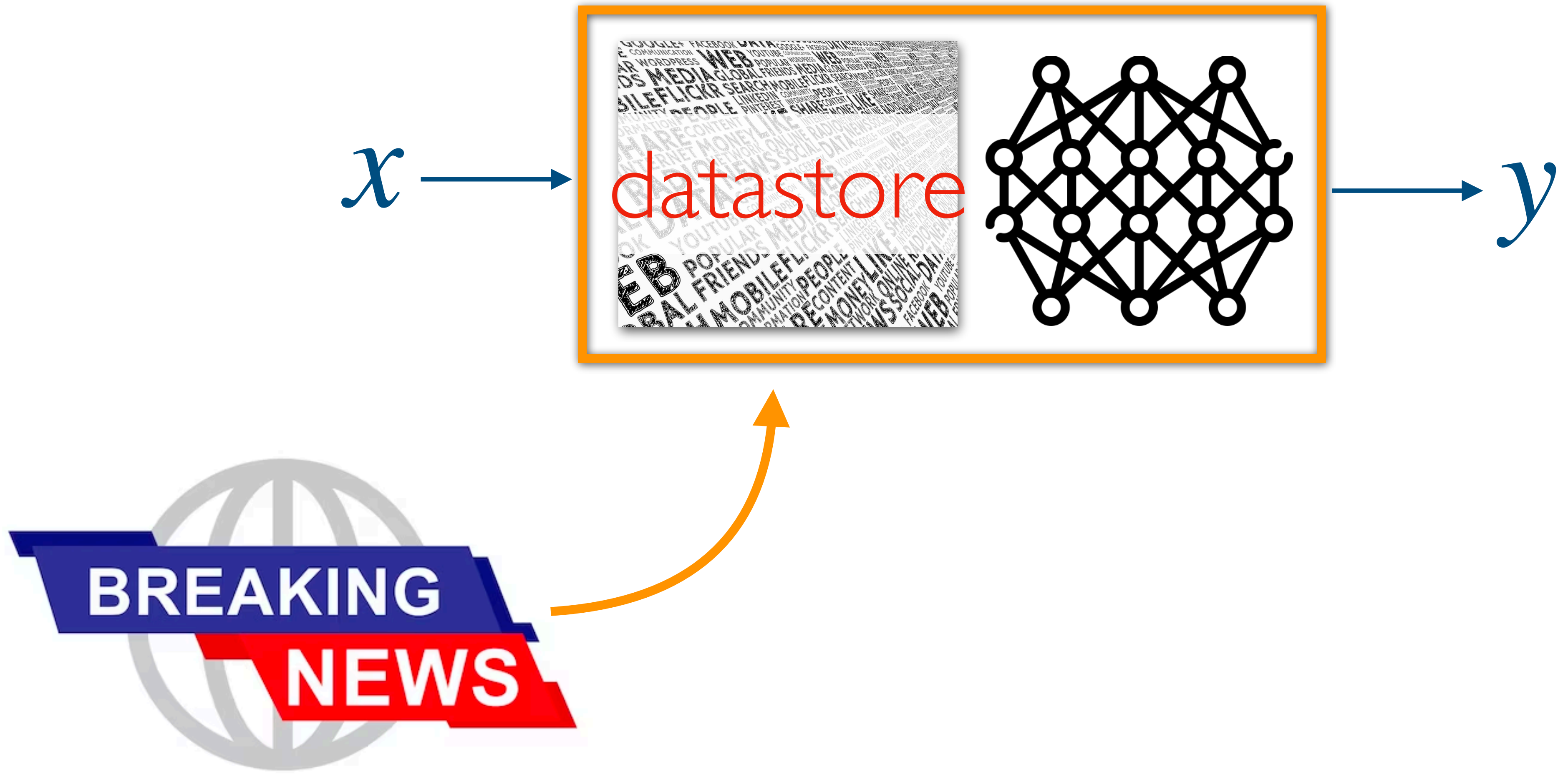
Can grow & update w/o additional training

Provide data attribution



Why Retrieval-based LMs?

New dimension in data use & better at long-tail | **Can grow & update w/o additional training** | Provide data attribution



Why Retrieval-based LMs?

New dimension in data use & better at long-tail

Can grow & update w/o additional training

Provide data attribution

Why Retrieval-based LMs?

New dimension in data use & better at long-tail | Can grow & update w/o additional training | **Provide data attribution**

List 4 important papers authored by Geoffrey Hinton

Why Retrieval-based LMs?

New dimension in data use & better at long-tail | Can grow & update w/o additional training | **Provide data attribution**

List 4 important papers authored by Geoffrey Hinton

ChatGPT

Geoffrey Hinton is a renowned computer scientist … Here are four important papers authored by him:

1. "Learning Internal Representations by Error Propagation" (with D. E. Rumelhart and R. J. Williams) - This paper, published in 1986, ..
2. "Deep Boltzmann Machines" (with R. Salakhutdinov) - Published in 2009, ..
3. "Deep Learning" (with Y. Bengio and A. Courville) - Published as a book in 2016, …
4. "Attention Is All You Need" (with V. Vaswani, N. Shazeer, et al.) - Published in 2017, this paper introduced the Transformer model,…

Overview

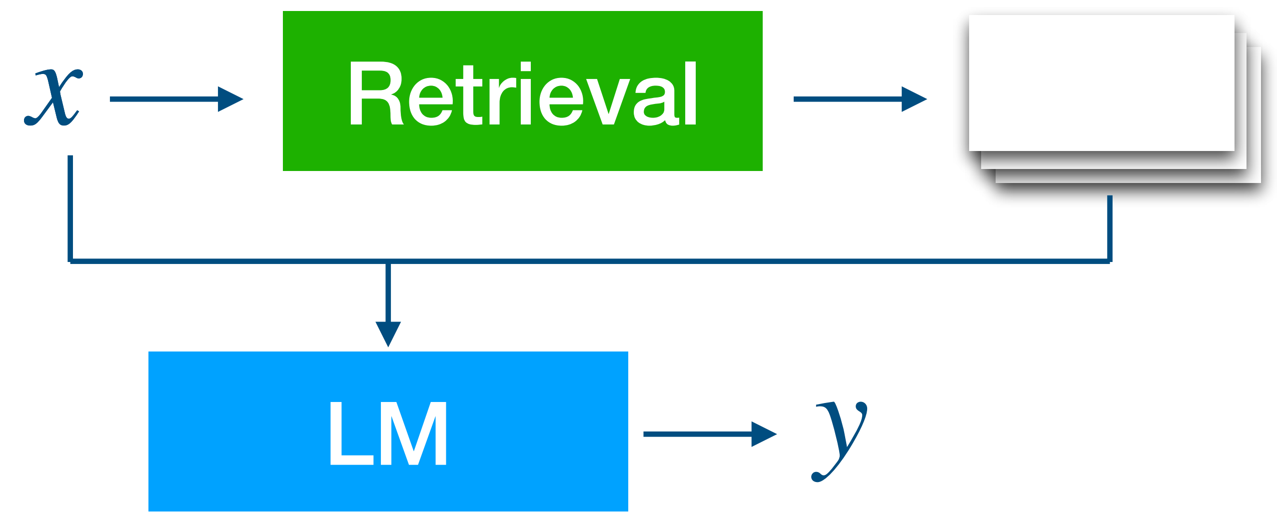
Why Retrieval-based LMs?

Tell me about Meta Platform.

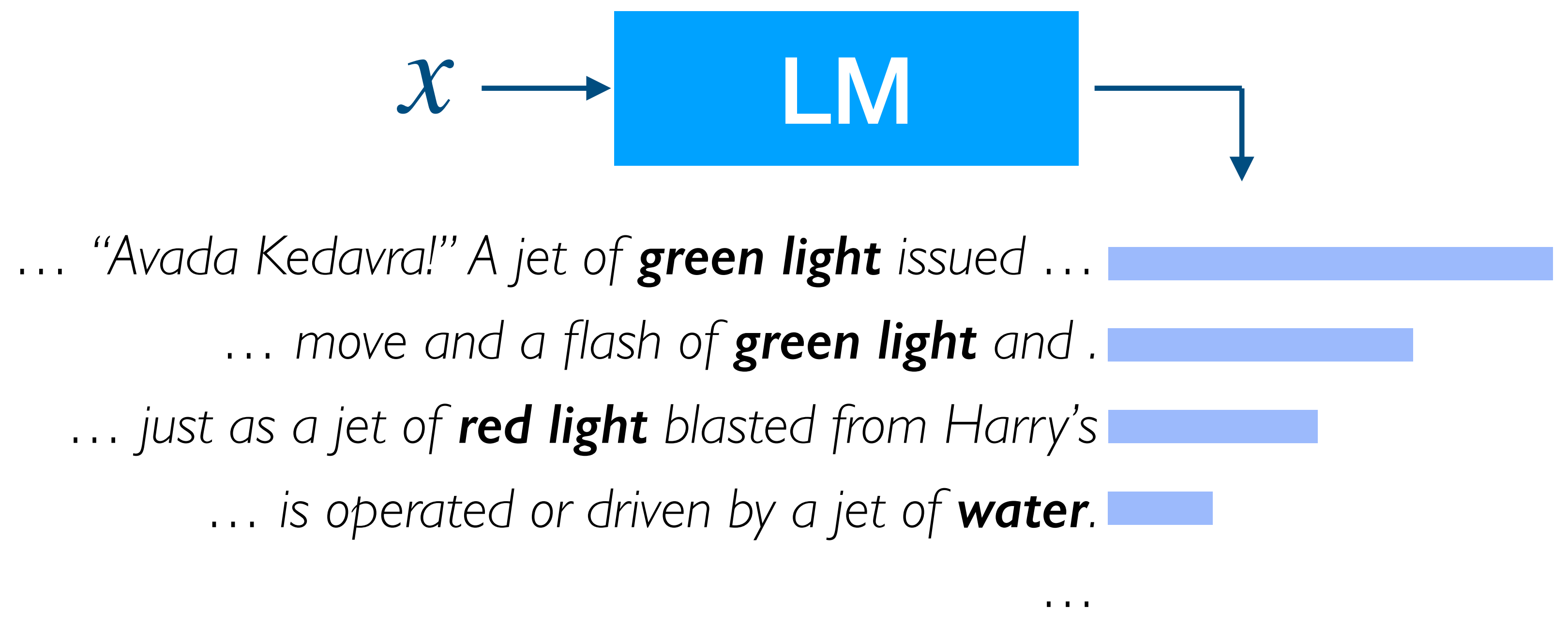
ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Retrieval Augmentation



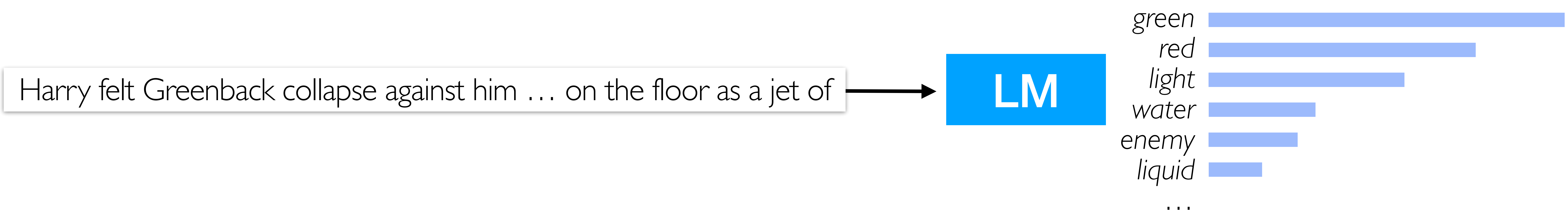
New Retrieval-based LMs



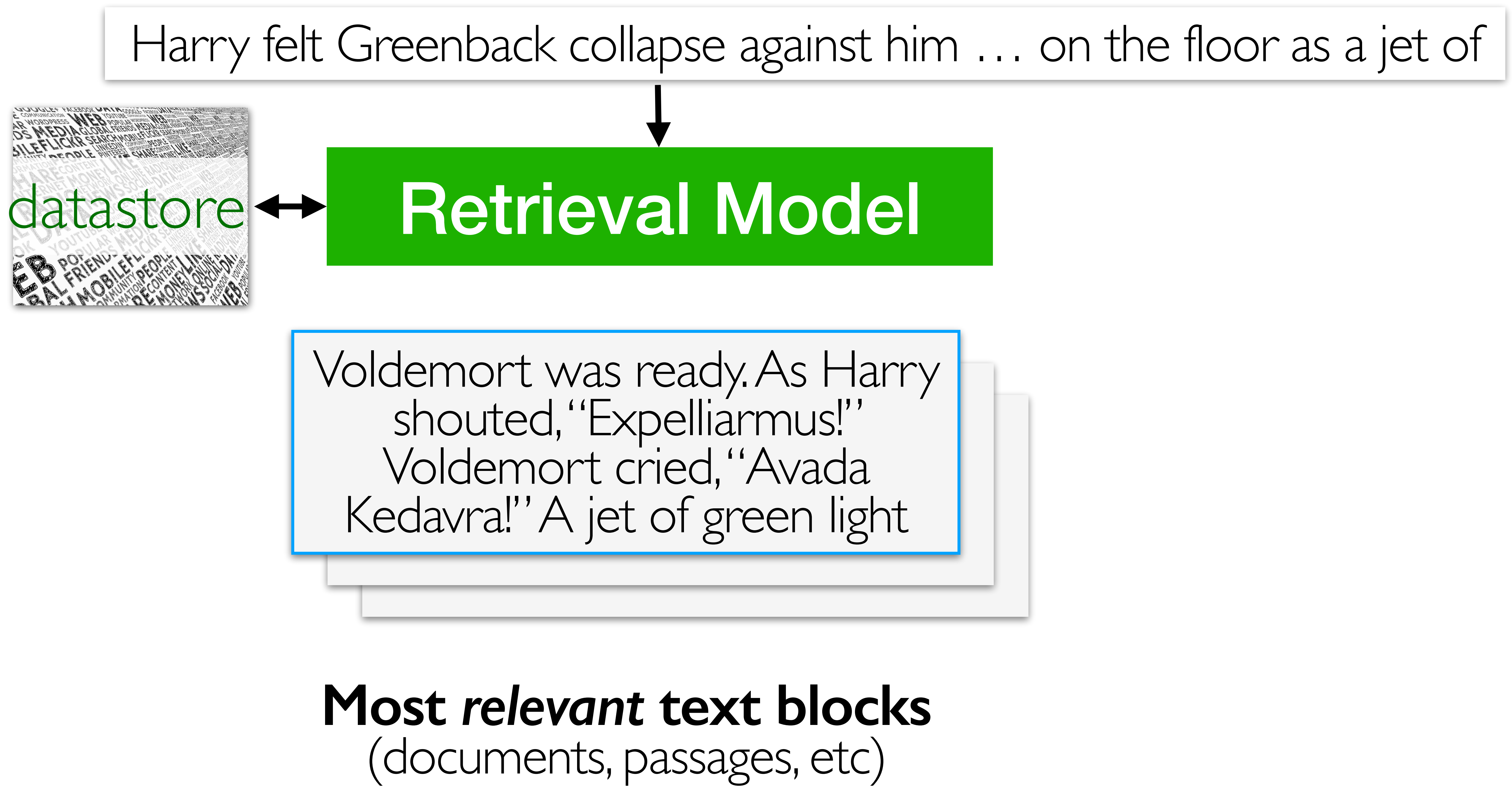
Open Problems

Scaling **datastore** not just parameters?

Language Models (w/o retrieval)

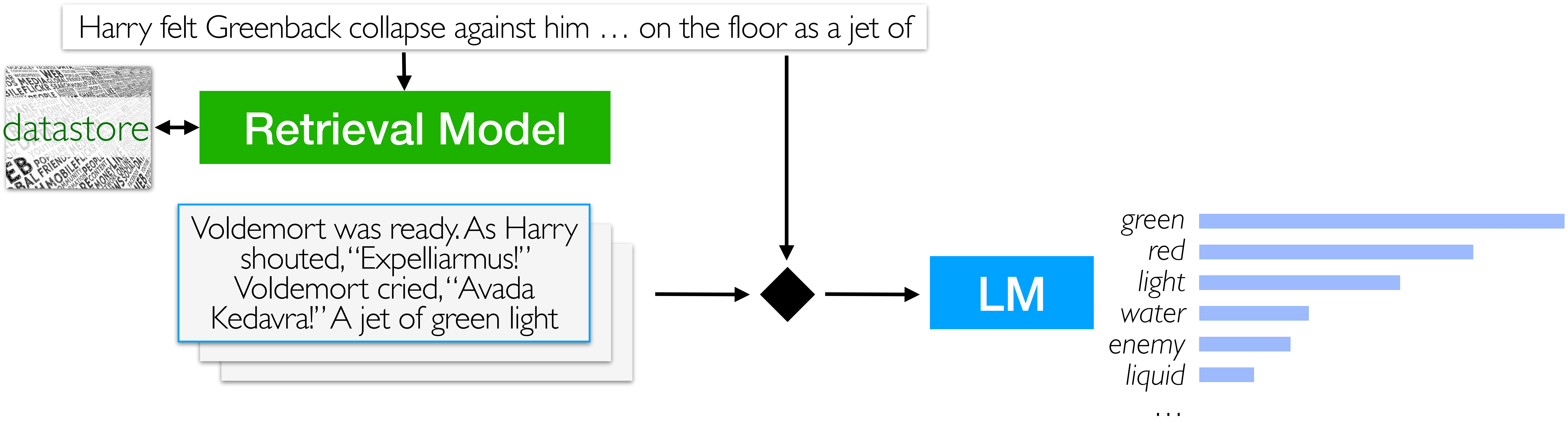


Language Models (w/ retrieval)



Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

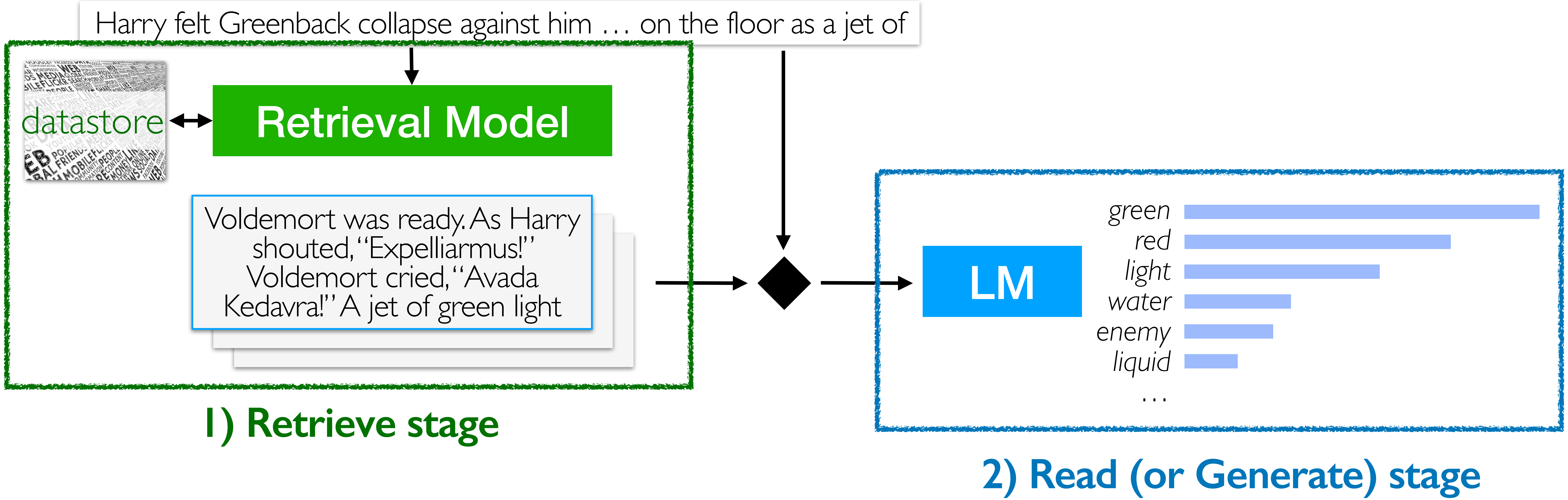
Language Models (w/ retrieval)



Most *relevant* text blocks
(documents, passages, etc)

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

Retrieval augmentation



Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

Retrieval augmentation: Overview

- Inference
- Training
- Key results

Retrieval augmentation: Overview

- **Inference**
- Training
- Key results

(1) Retrieve stage

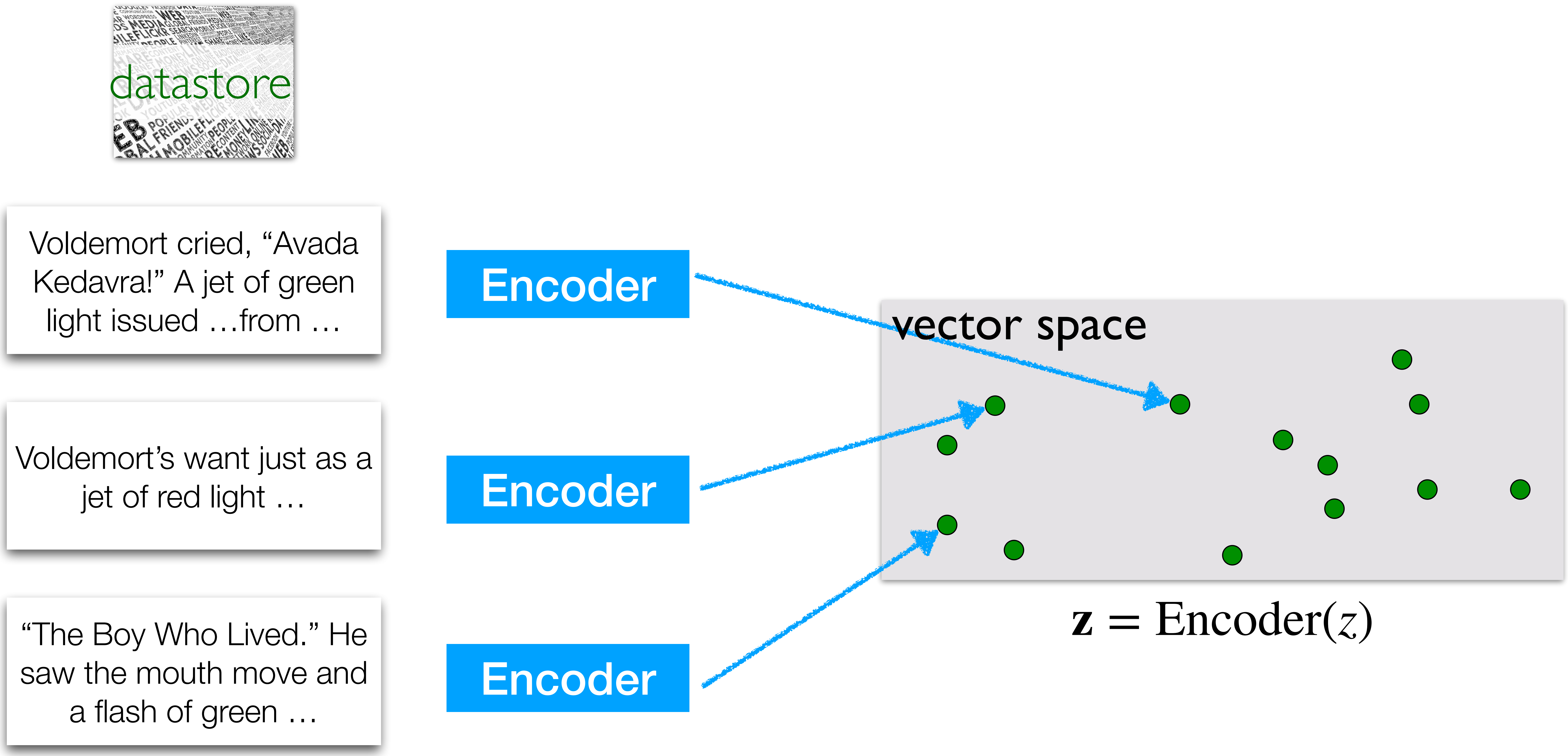
Voldemort cried, “Avada Kedavra!” A jet of green light issued …from …

Voldemort’s want just as a jet of red light …

“The Boy Who Lived.” He saw the mouth move and a flash of green …

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

(1) Retrieve stage

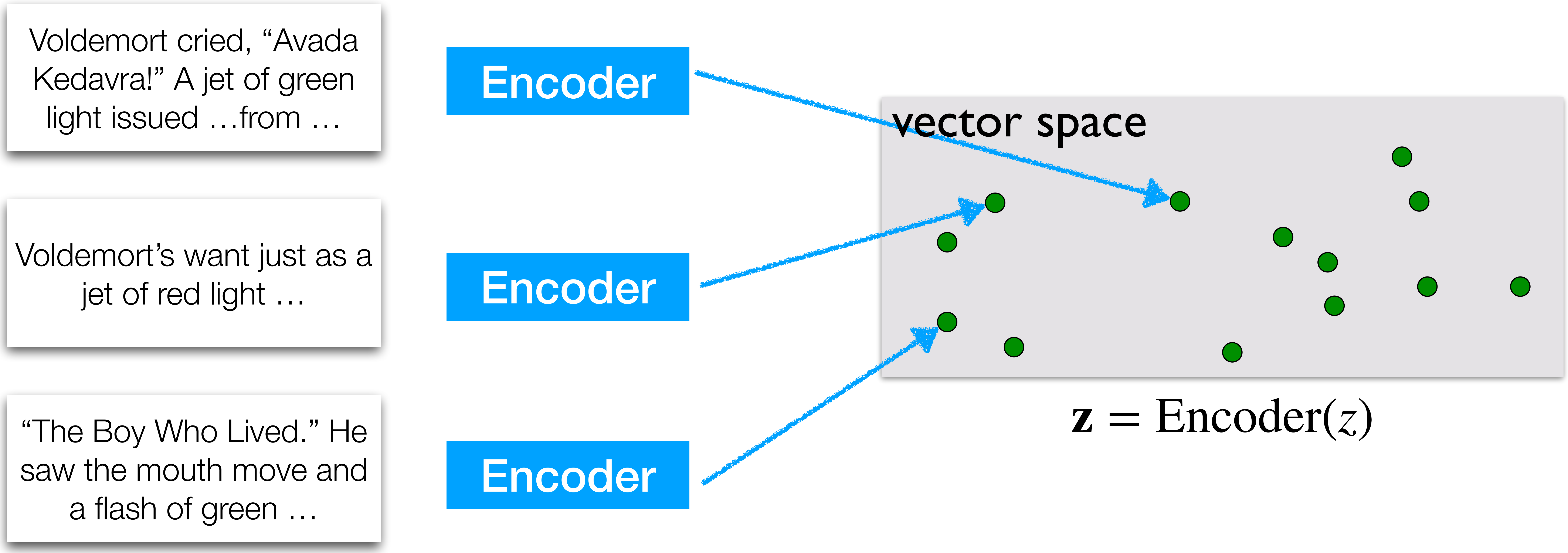


$\mathbf{z} = \text{Encoder}(z)$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

(1) Retrieve stage

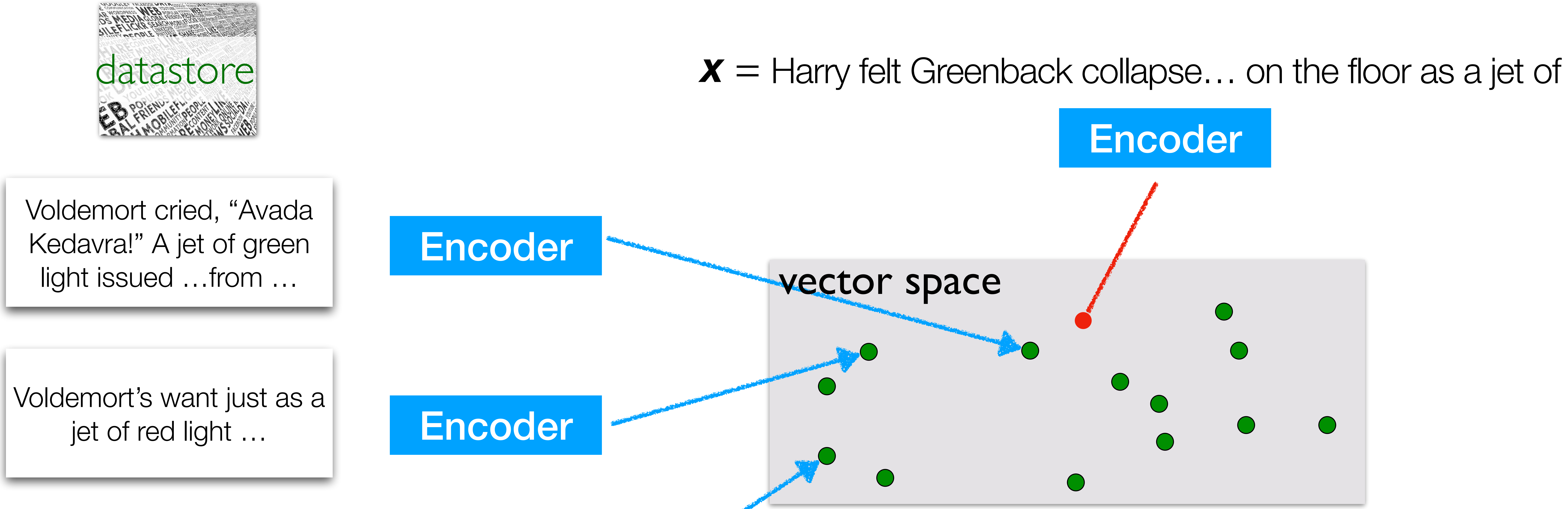
$\boldsymbol{x}$ = Harry felt Greenback collapse… on the floor as a jet of



$\mathbf{z} = \text{Encoder}(z)$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

(1) Retrieve stage

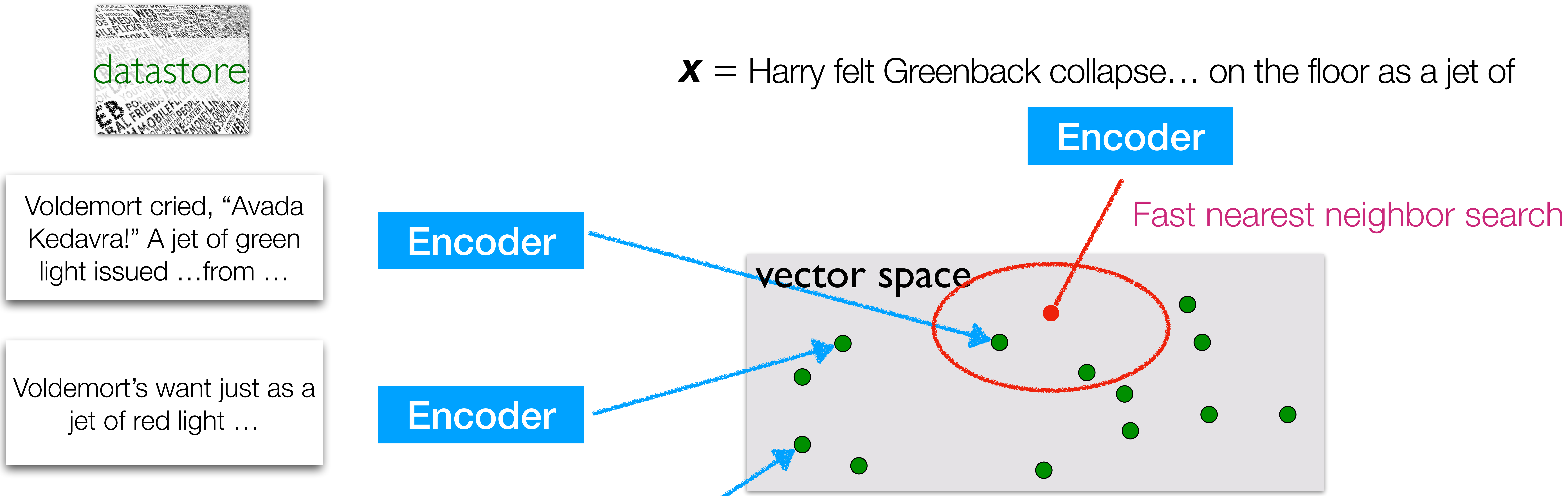


$$\mathbf{z} = \text{Encoder}(z)$$

$$\mathbf{x} = \text{Encoder}(x)$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

(1) Retrieve stage

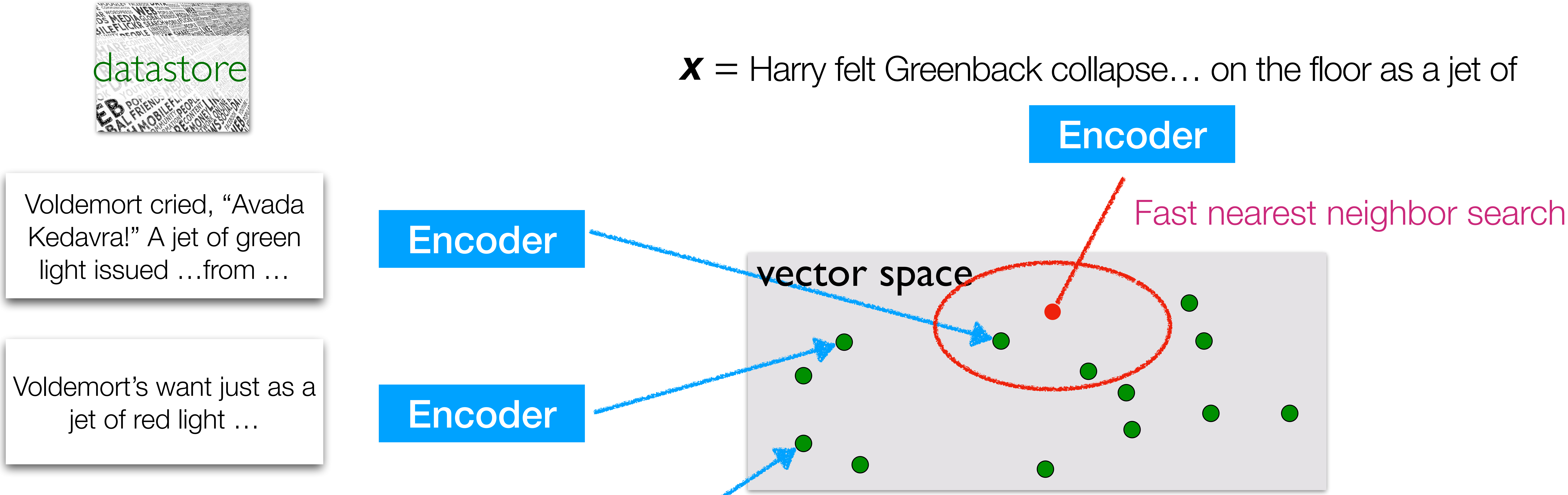


$$\mathbf{z} = \text{Encoder}(z)$$

$$\mathbf{x} = \text{Encoder}(x)$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

(1) Retrieve stage



$$\mathbf{z} = \text{Encoder}(z)$$

$$\mathbf{x} = \text{Encoder}(x)$$

$$z = \text{argmax}_{z \in \mathcal{Z}} \left(\text{sim}(x, z)\right)$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

(2) Read stage

Retrieval results *(ranked)*

Voldemort cried, “Avada Kedavra!” A jet of green light issued …from …

Voldemort’s want just as a jet of red light …

“The Boy Who Lived.” He saw the mouth move and a flash of green …

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

(2) Read stage

Retrieval results *(ranked)*

Voldemort cried, “Avada Kedavra!” A jet of green light issued …from …

Voldemort’s want just as a jet of red light …

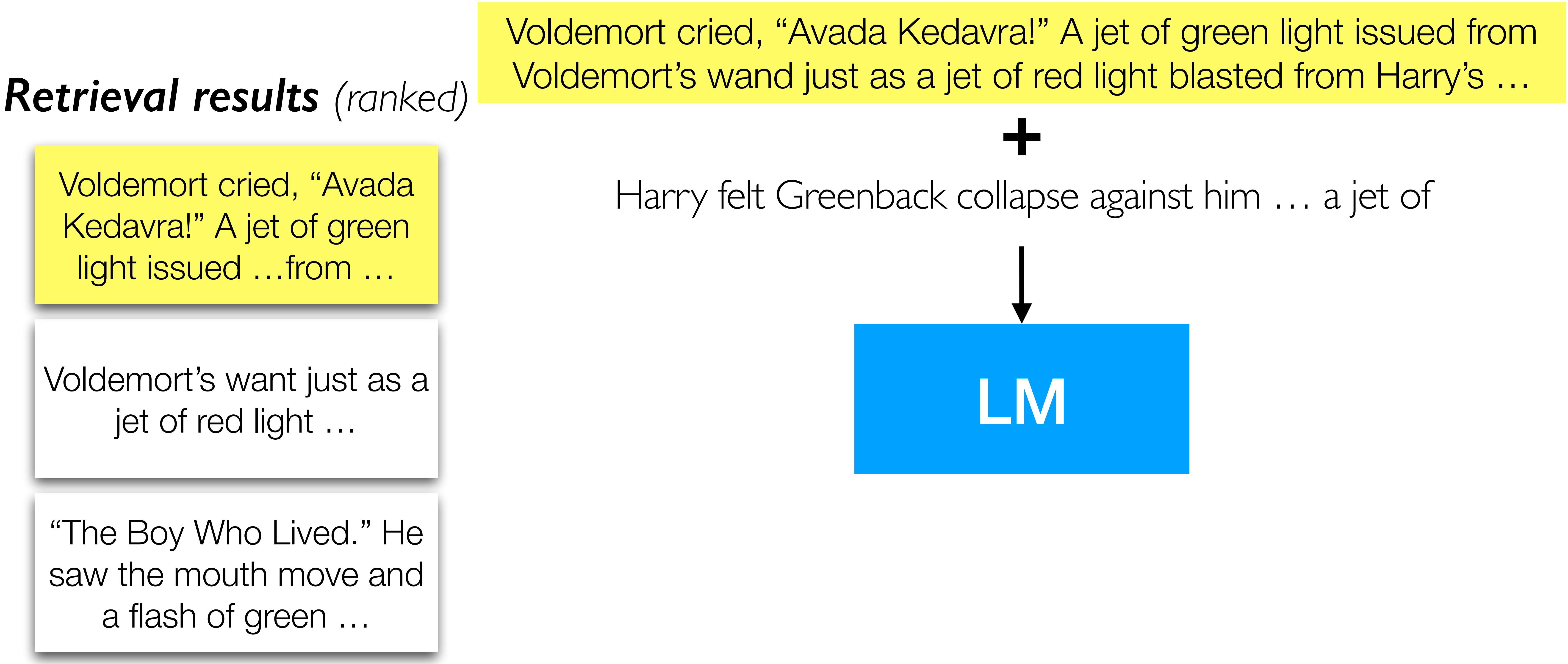
“The Boy Who Lived.” He saw the mouth move and a flash of green …

Voldemort cried, “Avada Kedavra!” A jet of green light issued from Voldemort’s wand just as a jet of red light blasted from Harry’s …

+

Harry felt Greenback collapse against him … a jet of

(2) Read stage



Retrieval results (ranked)
Voldemort cried, "Avada Kedavra!" A jet of green light issued …from …
Voldemort's want just as a jet of red light …
"The Boy Who Lived." He saw the mouth move and a flash of green …
Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's …
+
Harry felt Greenback collapse against him … a jet of
LM

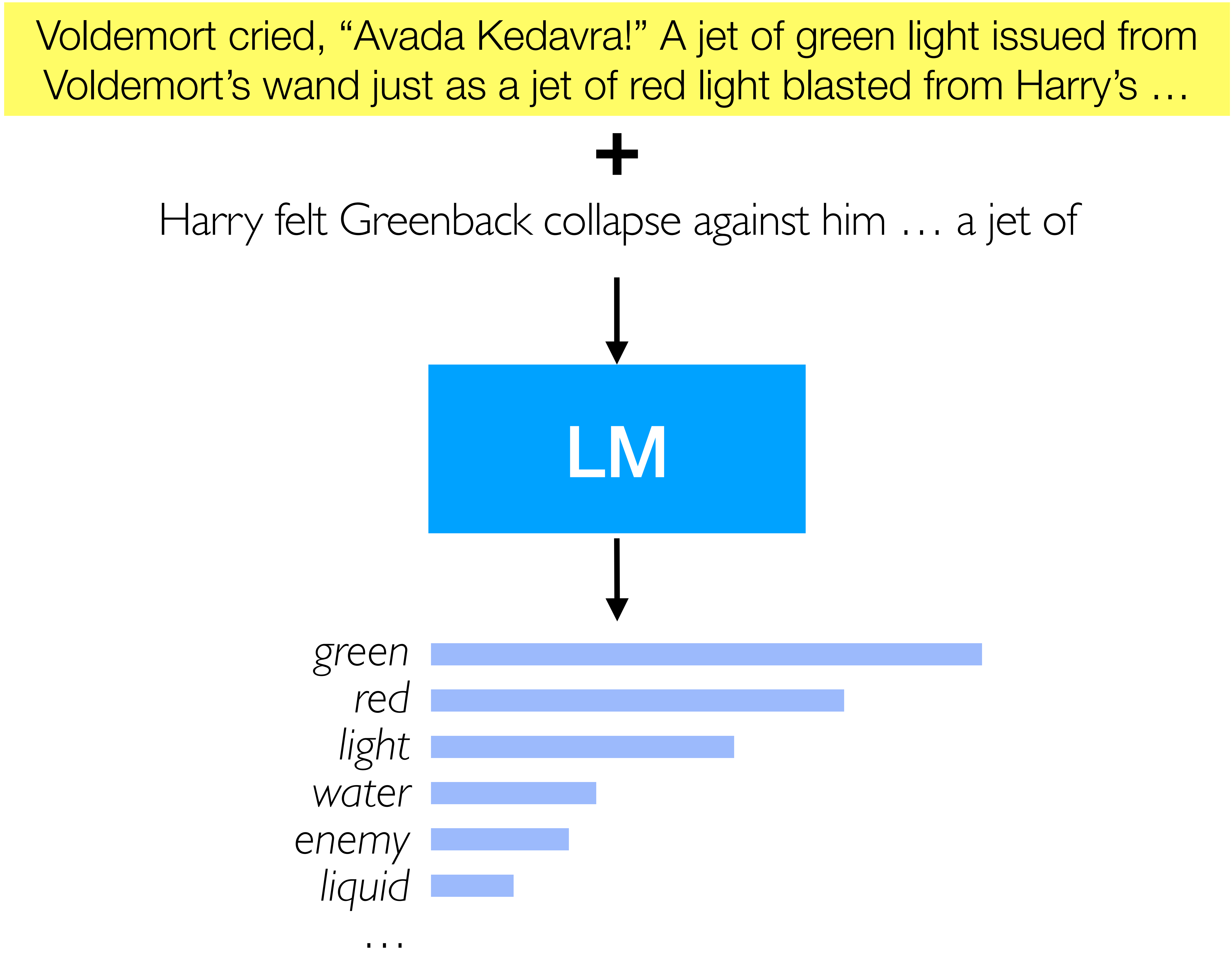
(2) Read stage

Retrieval results *(ranked)*

Voldemort cried, "Avada Kedavra!" A jet of green light issued …from …

Voldemort's want just as a jet of red light …

"The Boy Who Lived." He saw the mouth move and a flash of green …



(2) Read stage

Retrieval results *(ranked)*

Voldemort cried, "Avada Kedavra!" A jet of green light issued …from …

Voldemort's want just as a jet of red light …

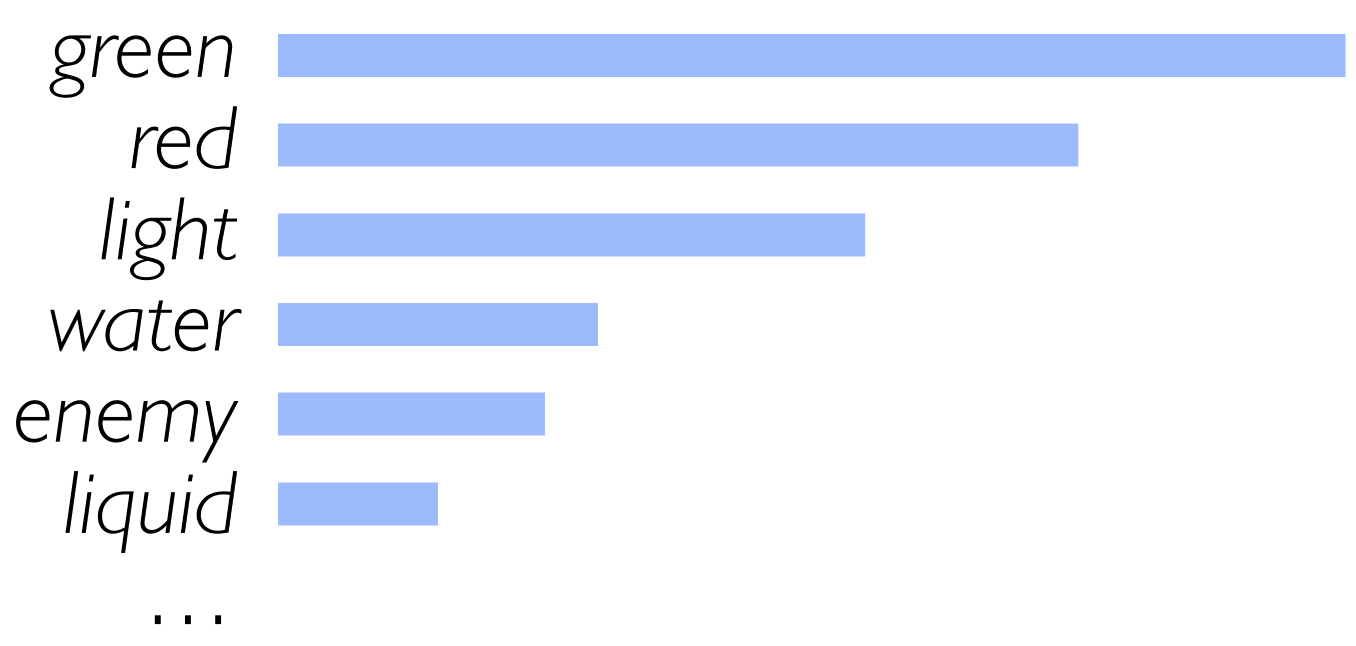
"The Boy Who Lived." He saw the mouth move and a flash of green …

Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's …

+

Harry felt Greenback collapse against him … a jet of

LM



Very simple

(You can use a black-box LM like an API!)

(2) Read stage

How to use multiple text blocks?

Retrieval results *(ranked)*

Voldemort cried, "Avada Kedavra!" A jet of green light issued …from …

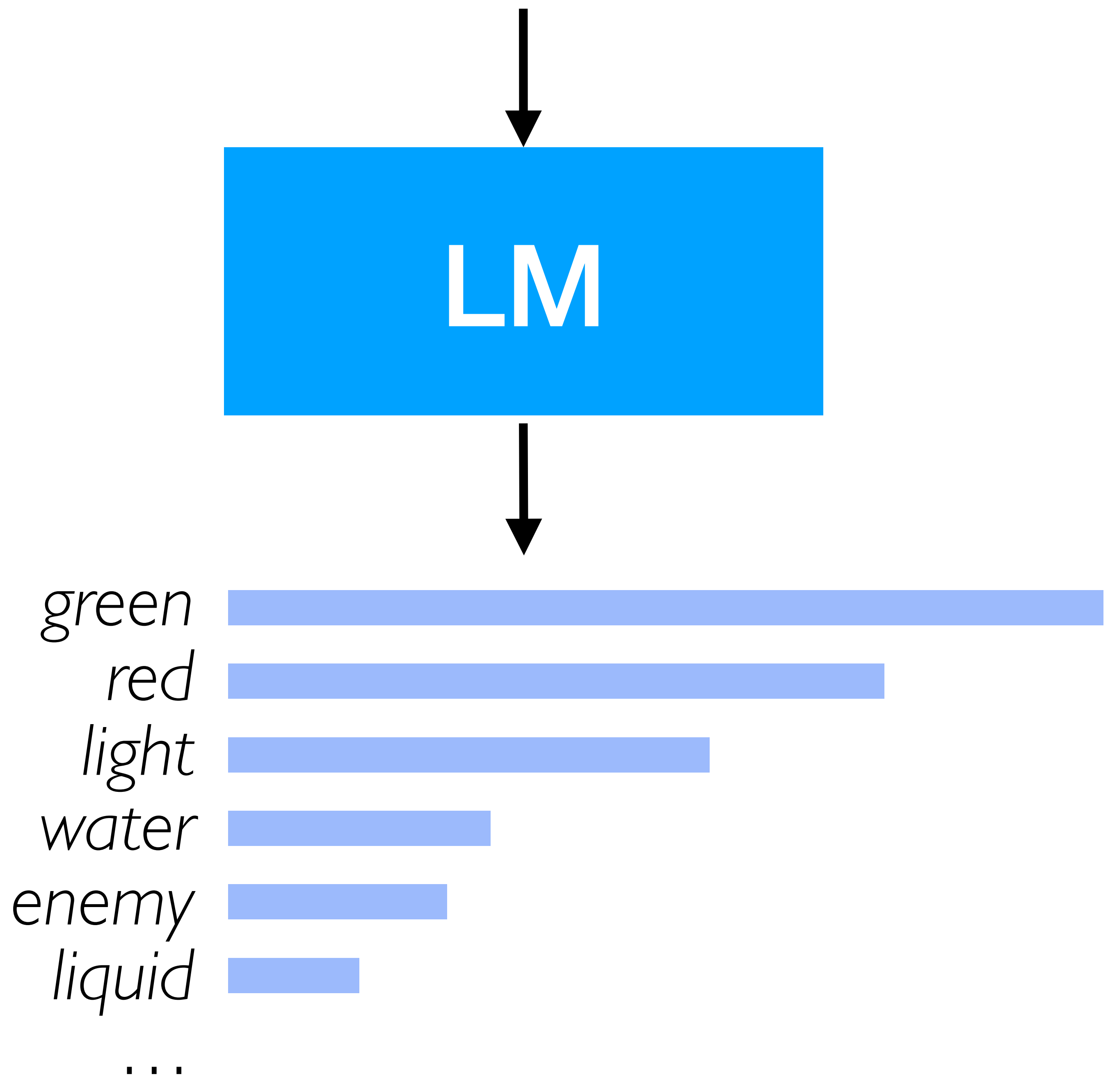
Voldemort's want just as a jet of red light …

"The Boy Who Lived." He saw the mouth move and a flash of green …

Voldemort cried, "Avada Kedavra!" A jet of green light issued from Voldemort's wand just as a jet of red light blasted from Harry's …

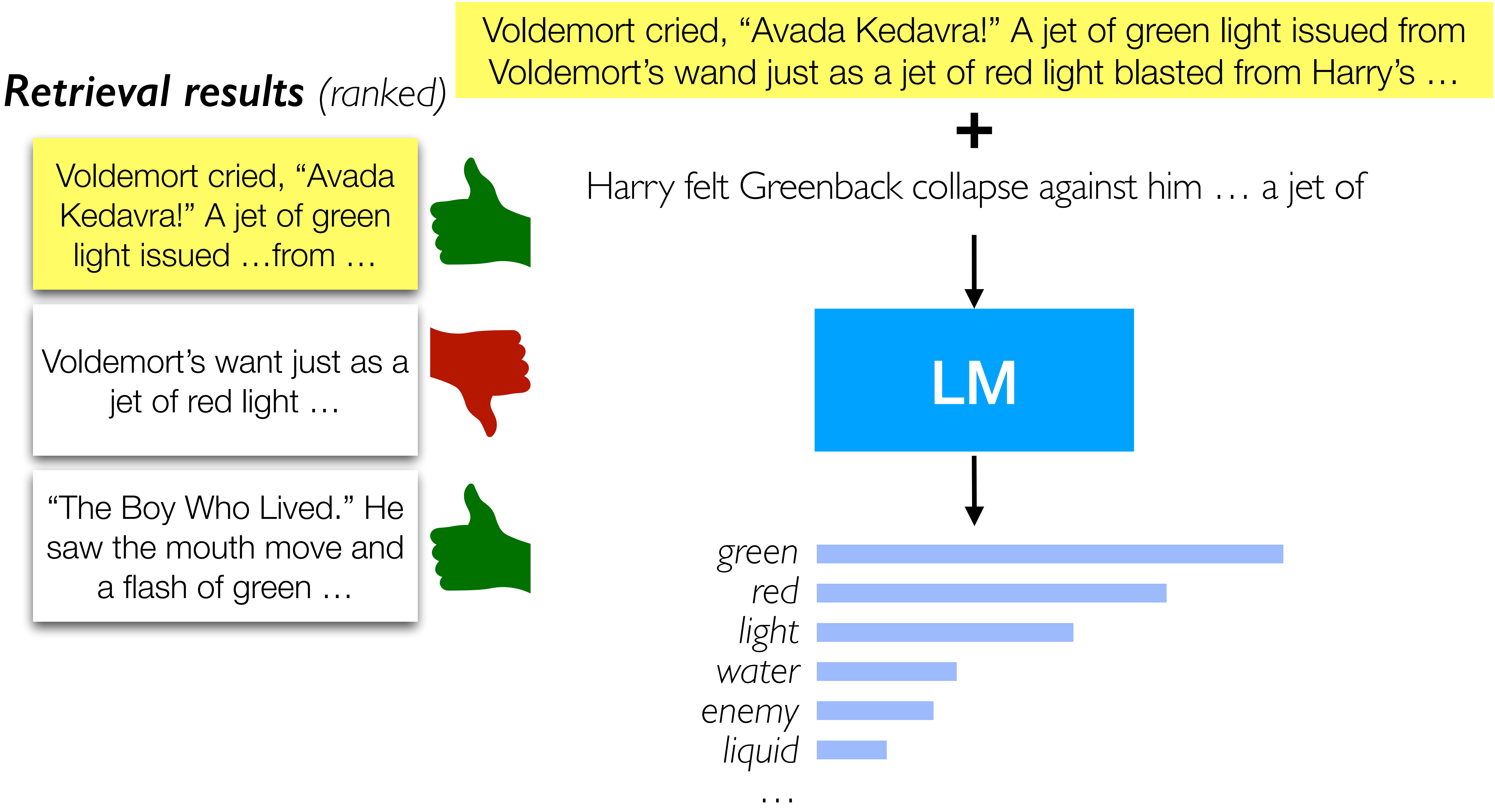
+

Harry felt Greenback collapse against him … a jet of



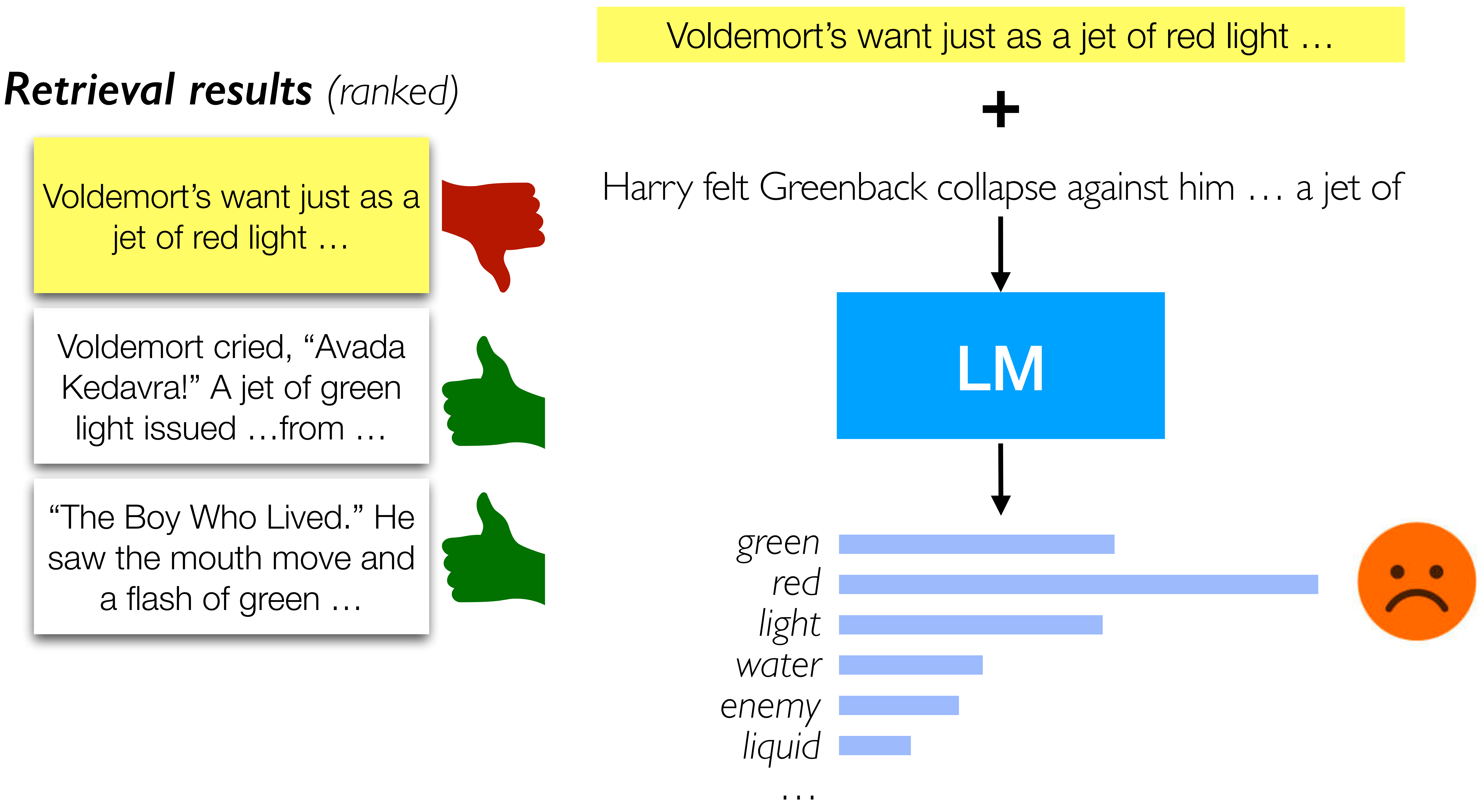
(2) Read stage

How to use multiple text blocks?



(2) Read stage

How to use multiple text blocks?



(2) Read stage

How to use multiple text blocks? **1) Concatenation**

Voldemort’s want just as a jet of red light …
Voldemort cried, “Avada Kedavra!” A jet of green light issued …from …
“The Boy Who Lived.” He saw the mouth move and a flash of green …

(2) Read stage

How to use multiple text blocks? **1) Concatenation**

Voldemort's want just as a jet of red light …
Voldemort cried, "Avada Kedavra!" A jet of green light issued …from …
"The Boy Who Lived." He saw the mouth move and a flash of green …

+

Harry felt Greenback collapse against him … a jet of

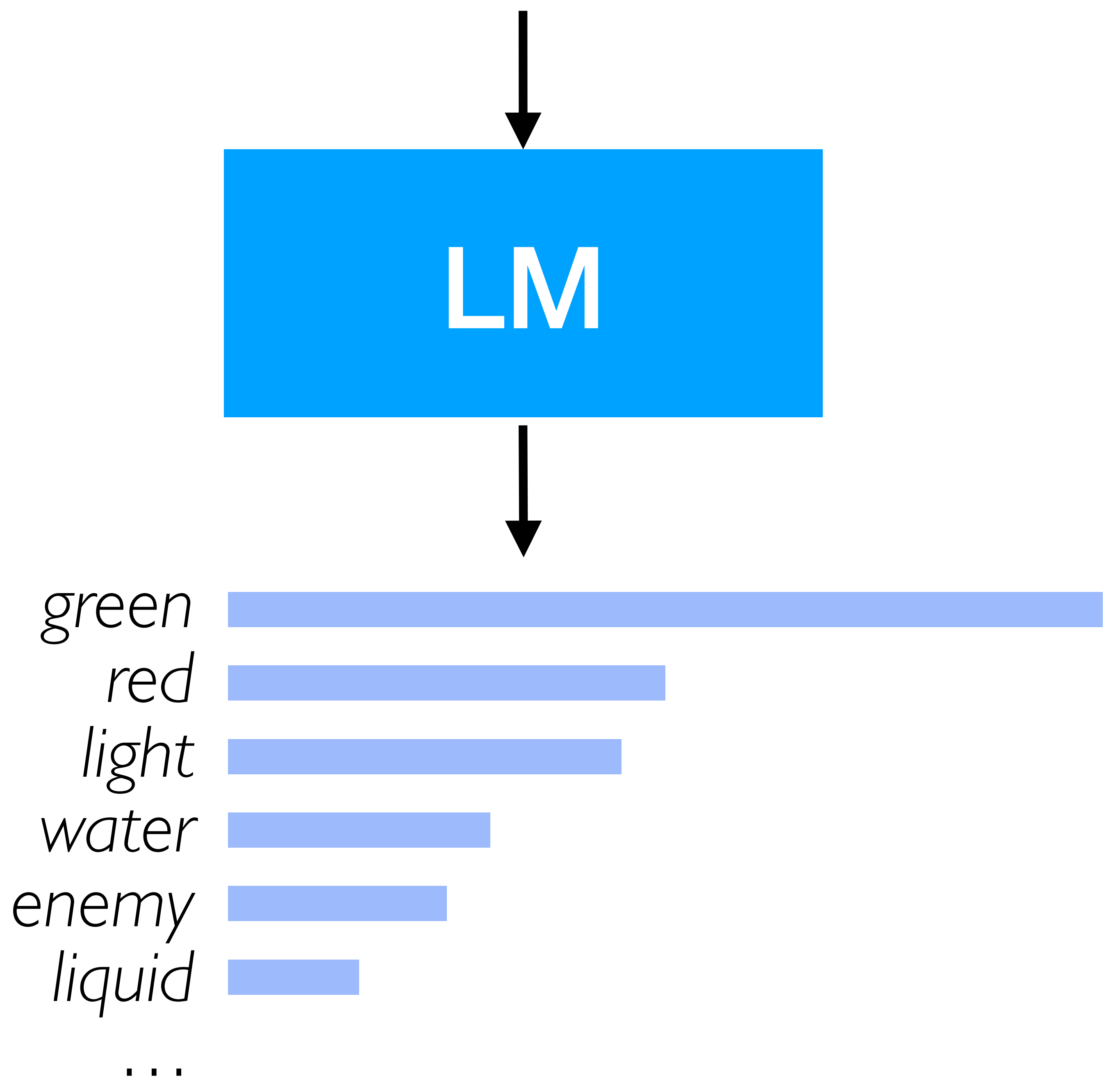
(2) Read stage

How to use multiple text blocks? **1) Concatenation**

Voldemort's want just as a jet of red light …
Voldemort cried, "Avada Kedavra!" A jet of green light issued …from …
"The Boy Who Lived." He saw the mouth move and a flash of green …

+

Harry felt Greenback collapse against him … a jet of



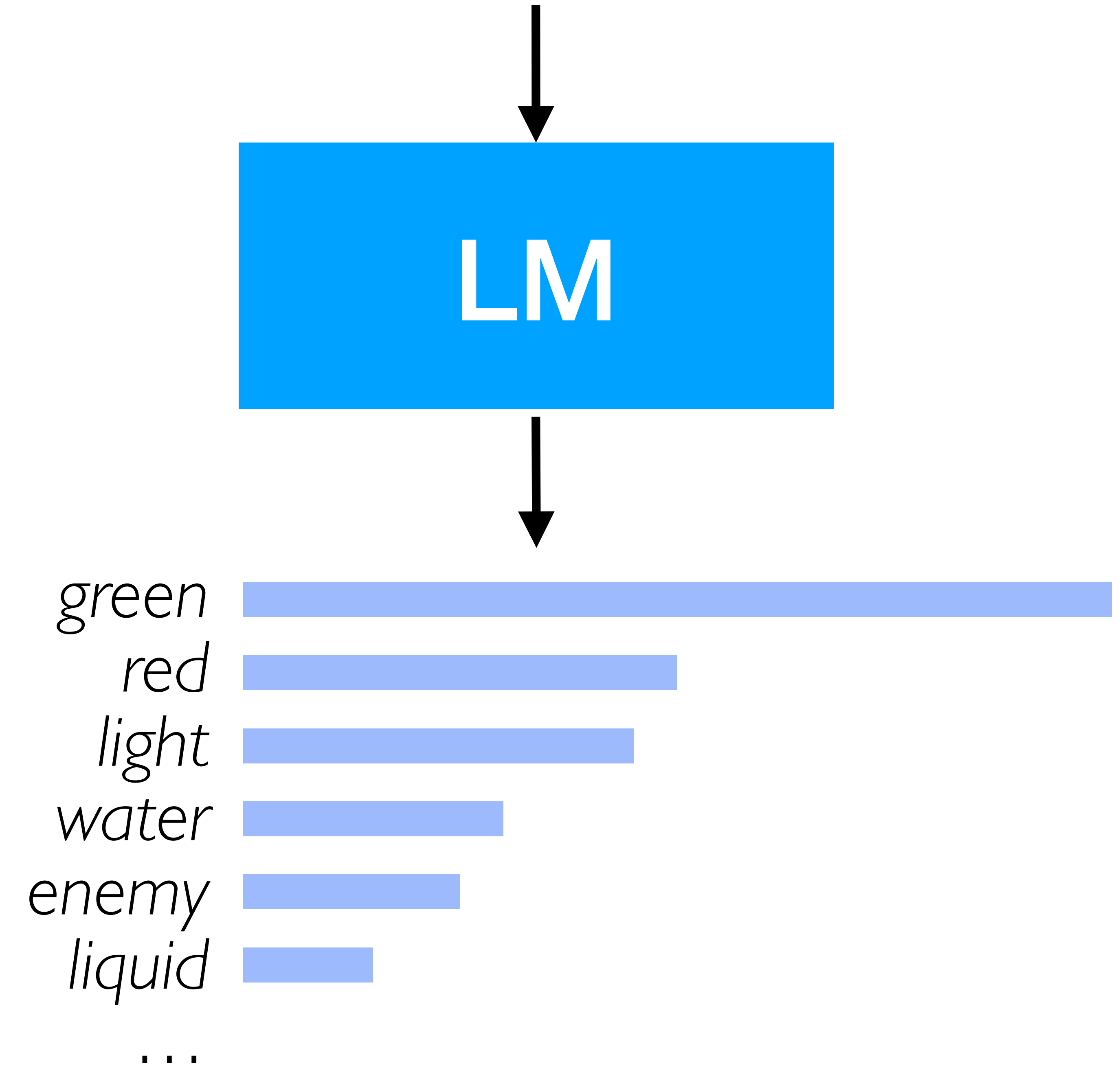
(2) Read stage

How to use multiple text blocks? **1) Concatenation**

Voldemort's want just as a jet of red light …
Voldemort cried, "Avada Kedavra!" A jet of green light issued …from …
"The Boy Who Lived." He saw the mouth move and a flash of green …

+

Harry felt Greenback collapse against him … a jet of



Simple

Increase the inference cost &
Bounded by the maximum
length limit of the LM

(2) Read stage

How to use multiple text blocks? **2) Ensembling**

Voldemort's want just as a jet of red light …

+

Harry felt Greenback collapse against him … a jet of

Voldemort cried, "Avada Kedavra!" A jet of green …

+

Harry felt Greenback collapse against him … a jet of

… saw the mouth move and a flash of green …

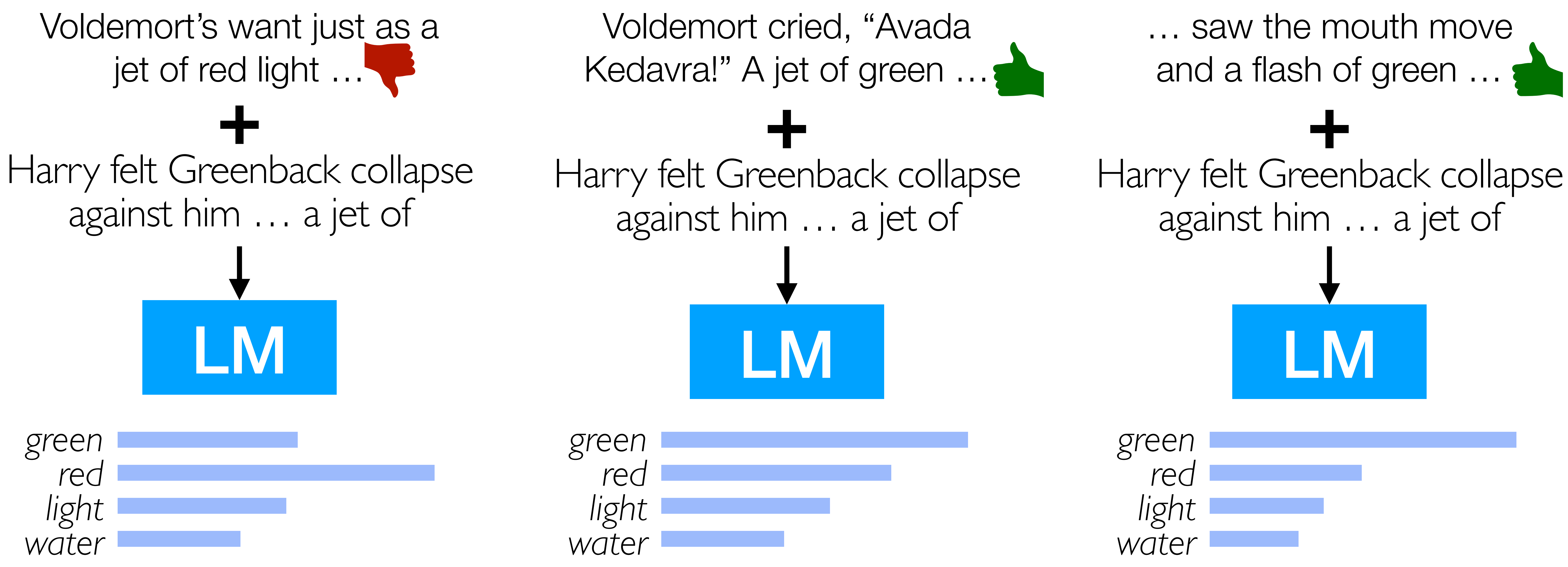
+

Harry felt Greenback collapse against him … a jet of

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training" * Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

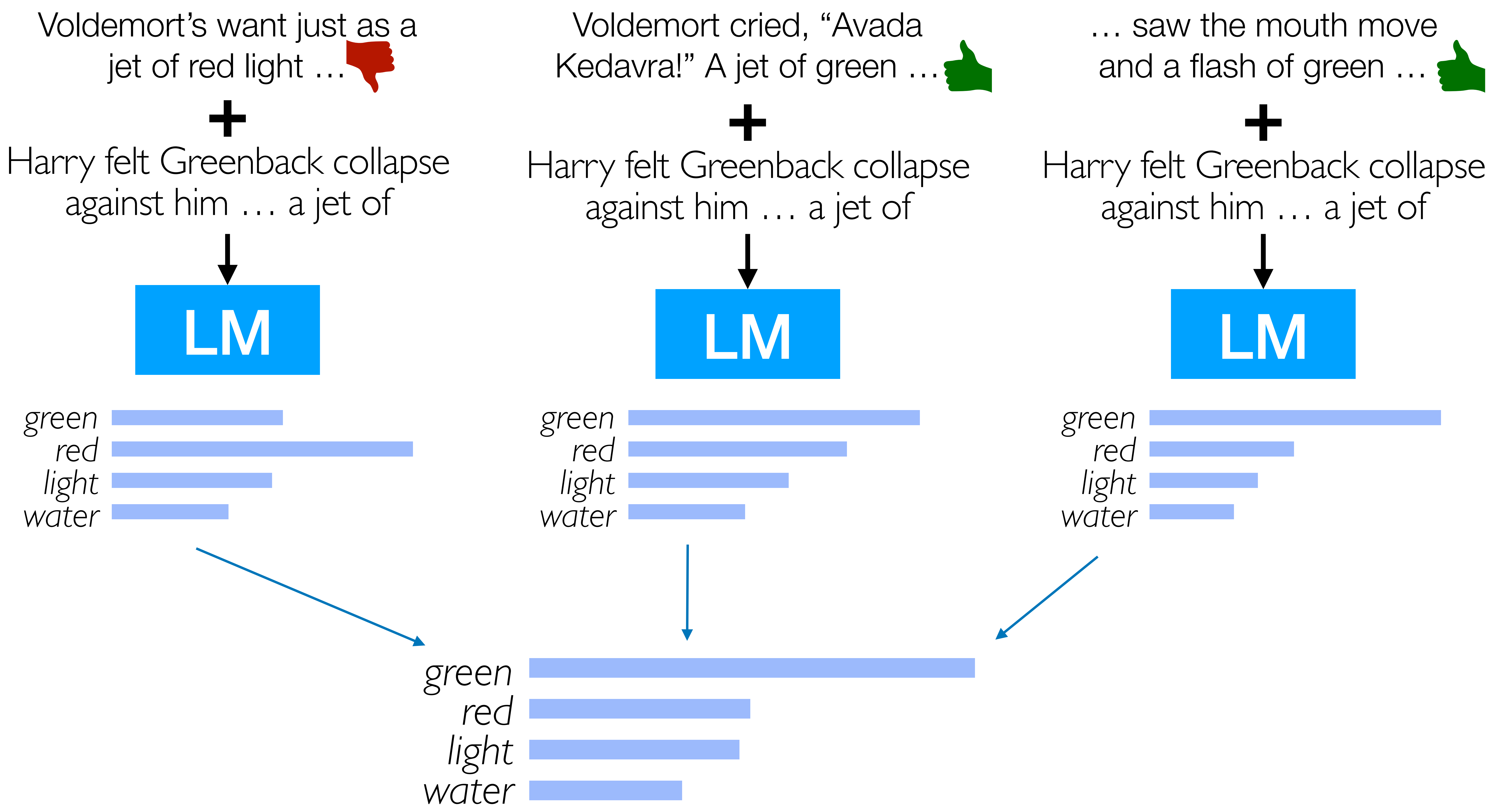
(2) Read stage

How to use multiple text blocks? **2) Ensembling**



(2) Read stage

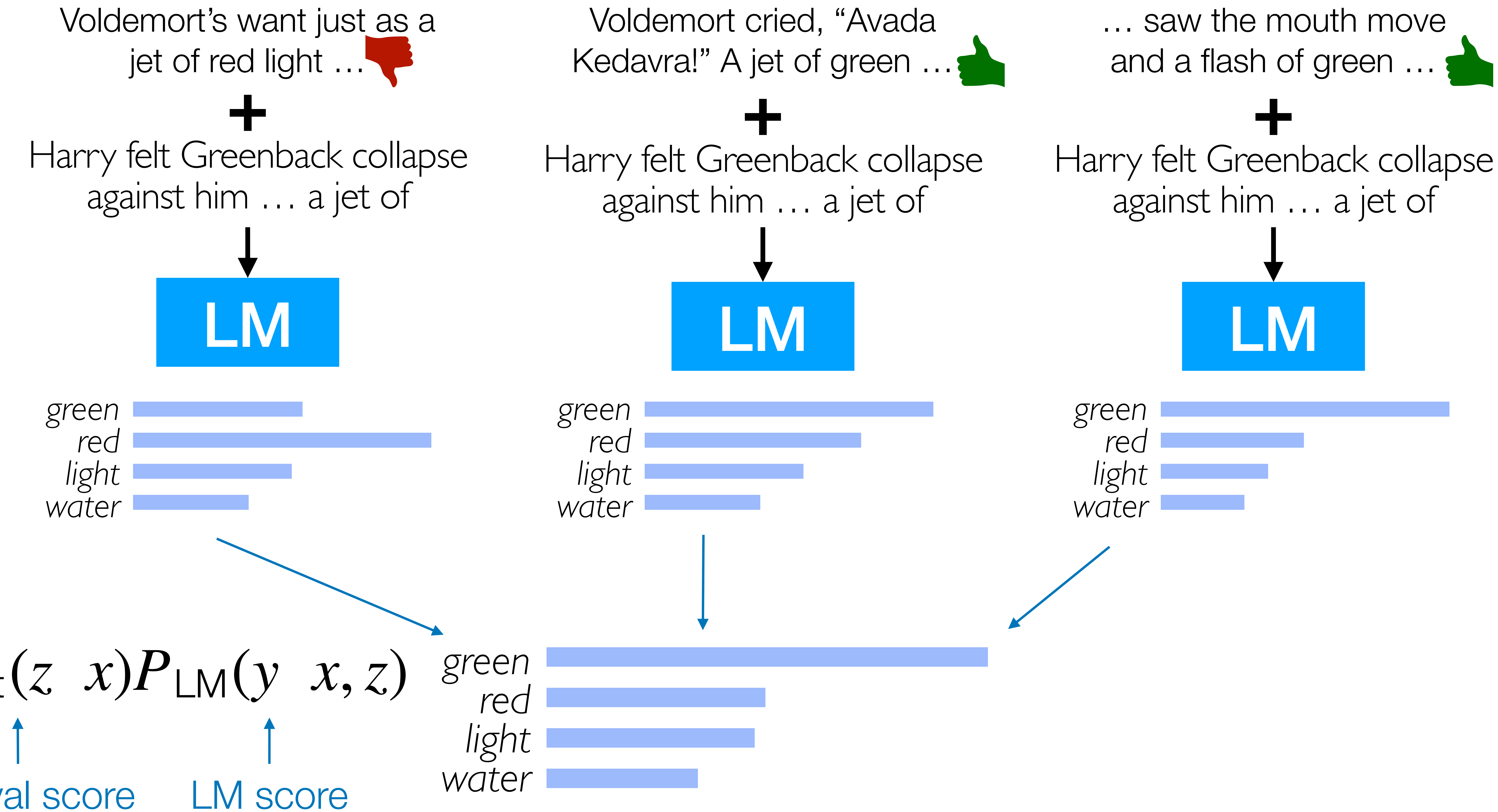
How to use multiple text blocks? **2) Ensembling**



Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training" * Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

(2) Read stage

How to use multiple text blocks? **2) Ensembling**

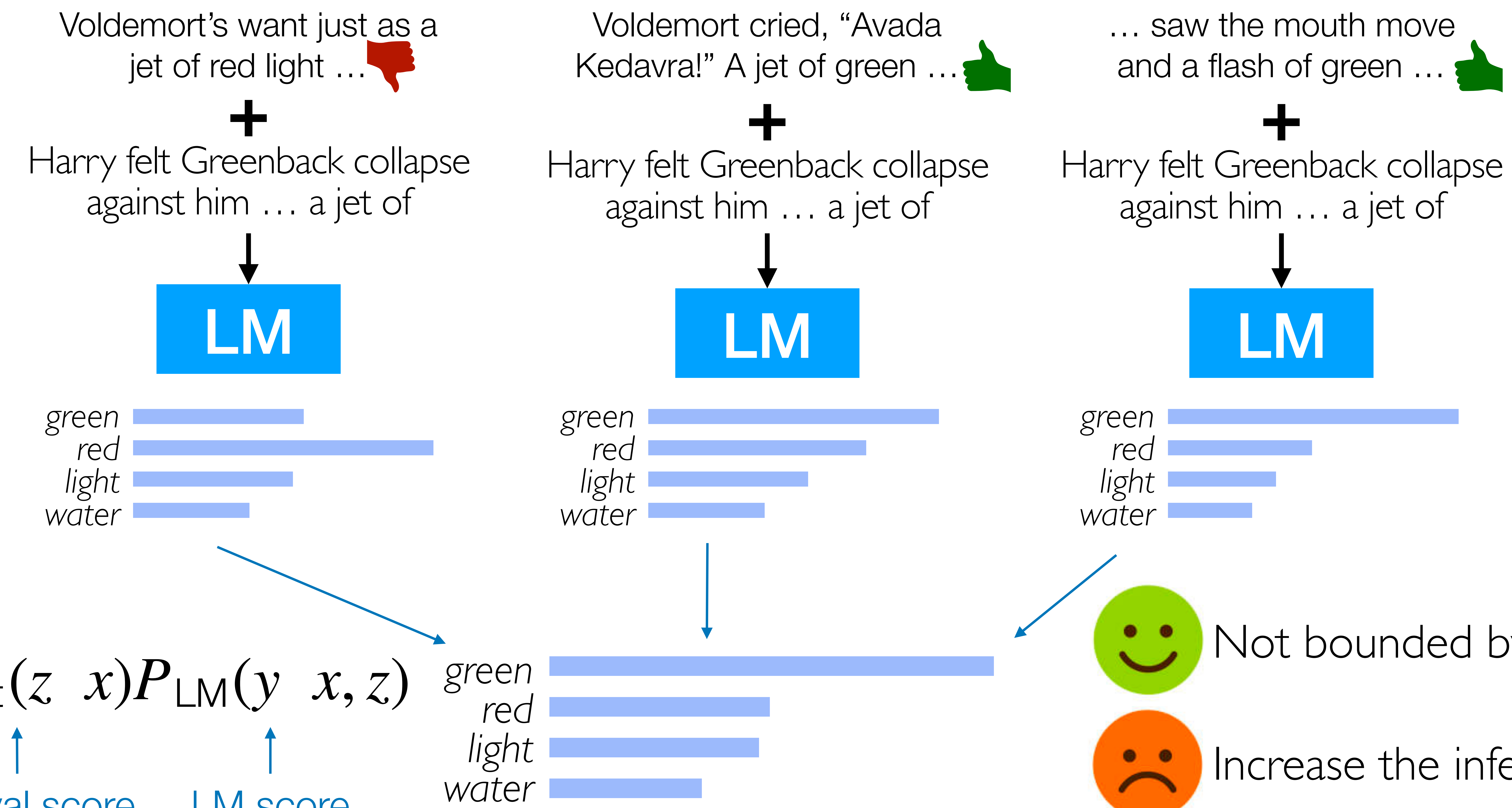


$$P(y \mid x) = \sum_{z \in \mathcal{Z}} P_{\text{ret}}(z \mid x) P_{\text{LM}}(y \mid x, z)$$

retrieval score   LM score

(2) Read stage

How to use multiple text blocks? **2) Ensembling**



$$P(y \quad x) = \sum_{z \in \mathcal{Z}} P_{\text{ret}}(z \quad x) P_{\text{LM}}(y \quad x, z)$$

retrieval score — LM score

Not bounded by the length limit

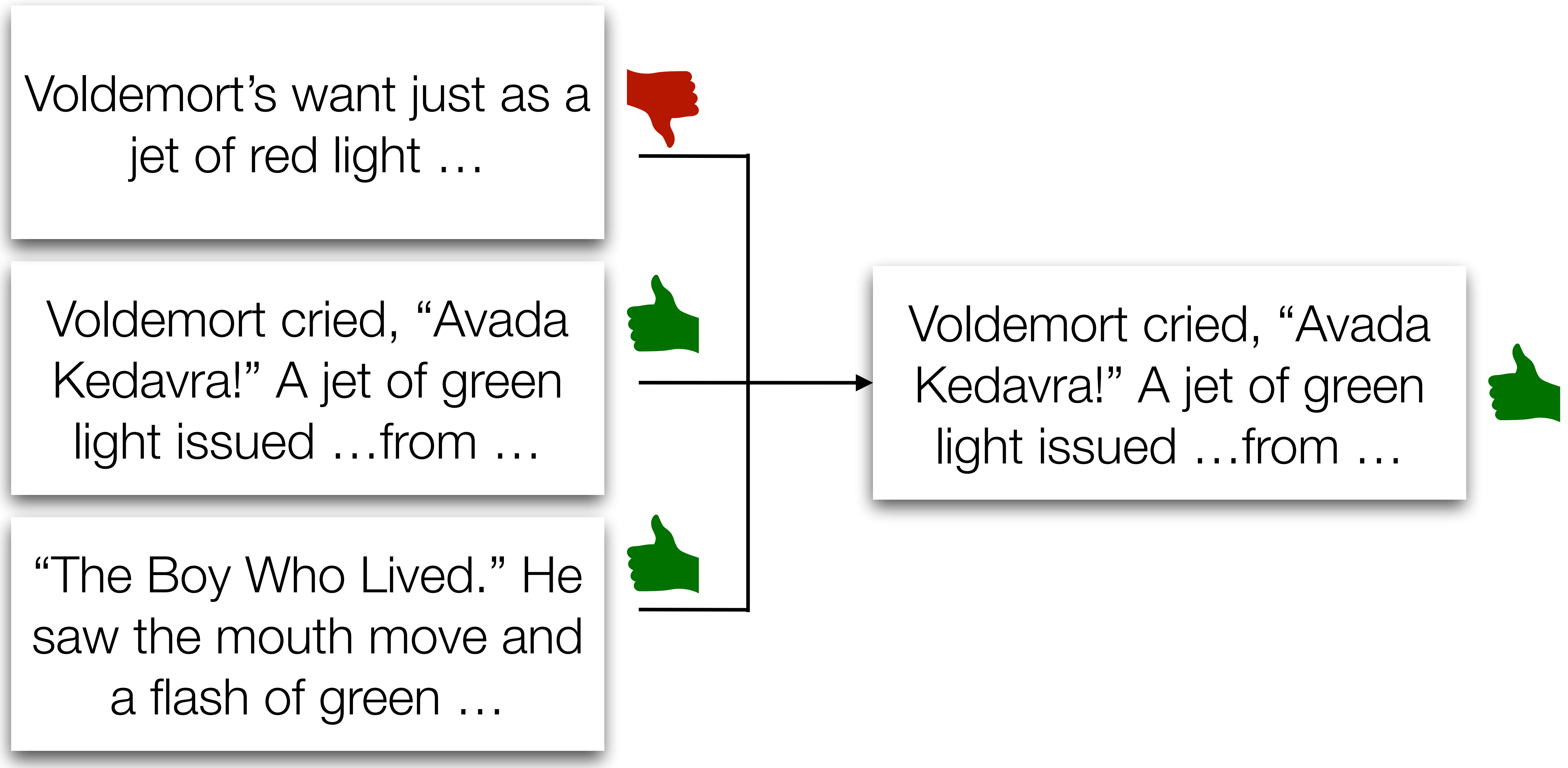
Increase the inference cost

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training" * Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

(2) Read stage

How to use multiple text blocks? **3) Reranking**

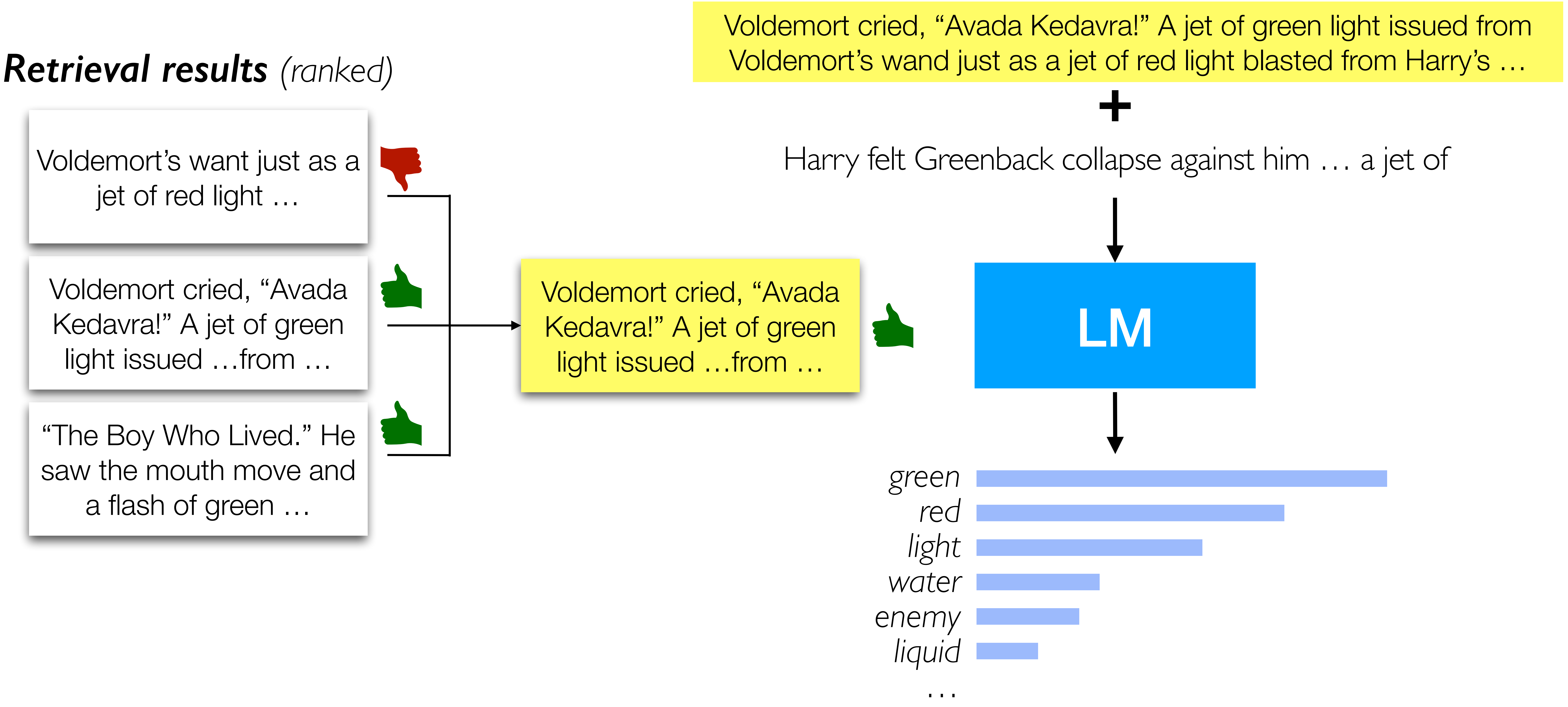
Retrieval results *(ranked)*



Ram et al. 2023. “In-Context Retrieval-Augmented Language Models”

(2) Read stage

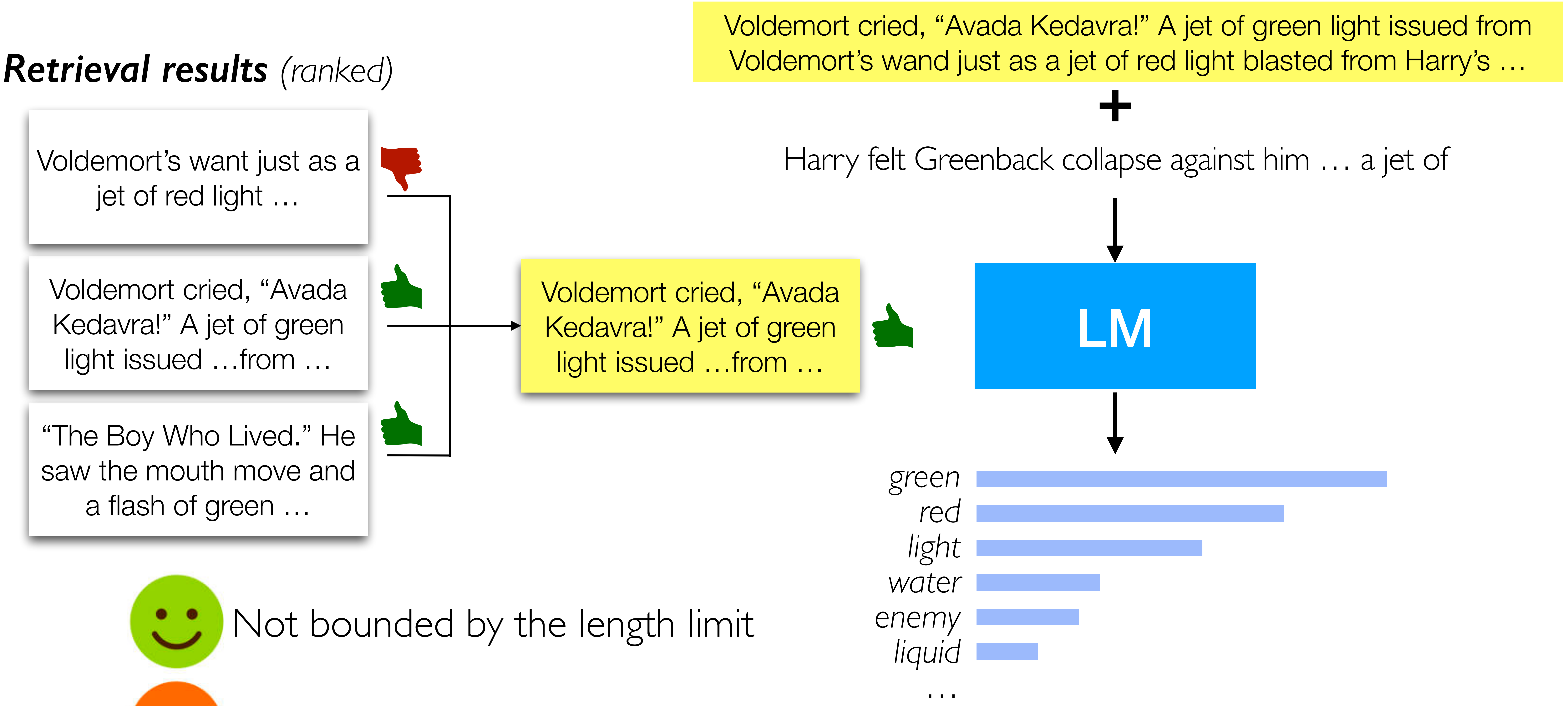
How to use multiple text blocks? **3) Reranking**



Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"

(2) Read stage

How to use multiple text blocks? **3) Reranking**



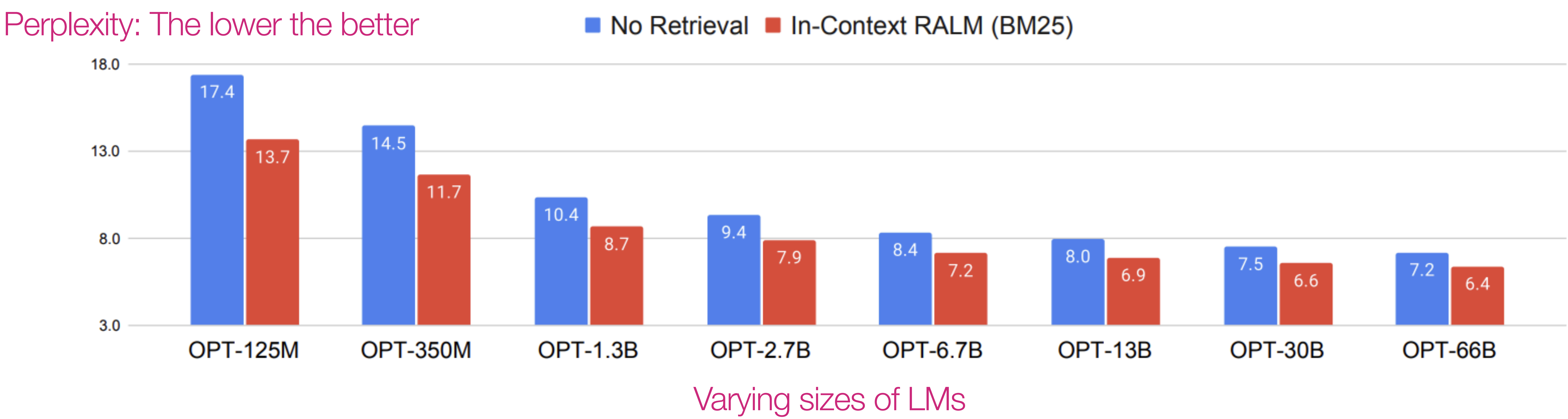
- Not bounded by the length limit
- Increase the inference cost

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"

Key results

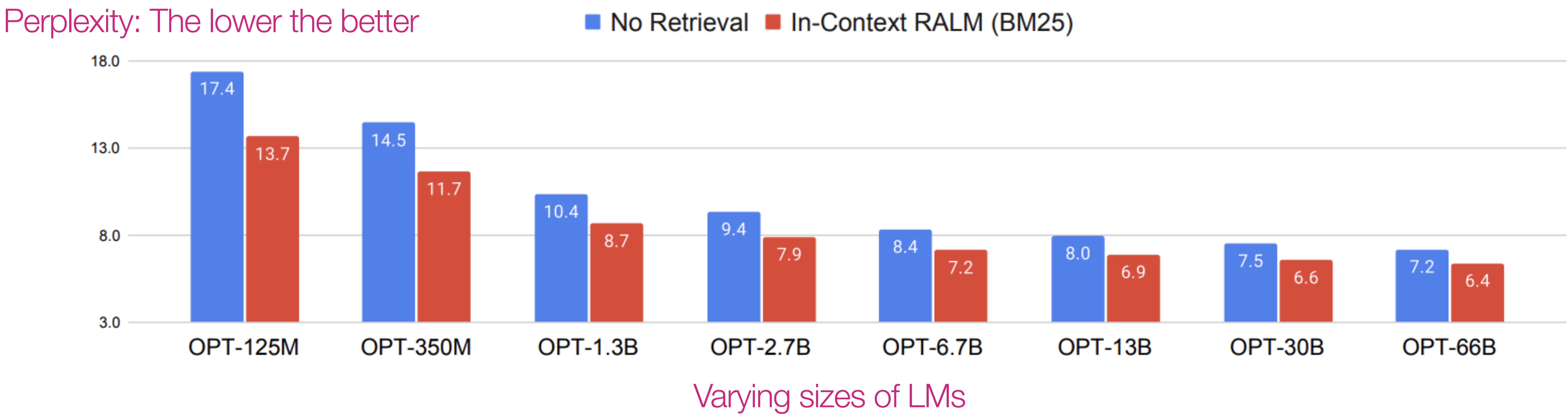
Ram et al. 2023. “In-Context Retrieval-Augmented Language Models”

Key results



Ram et al. 2023. “In-Context Retrieval-Augmented Language Models”

Key results



Retrieval helps over all sizes of LMs

Ram et al. 2023. "In-Context Retrieval-Augmented Language Models"

Retrieval augmentation: Overview

- Inference
  - **Step 1: Retrieve**
  - **Step 2: Read (Generate)**
  - **Optionally, with multiple passages: Concatenation, Ensembling, Reranking**
- Training
- Key results

Retrieval augmentation: Overview

- Inference
  - Step 1: Retrieve
  - Step 2: Read (Generate)
  - Optionally, with multiple passages: Concatenation, Ensembling, Reranking
- **Training**
- Key results

How to train it?

Retrieval Model

trained in isolation

LM

trained in isolation

How to train it?

Retrieval Model

trained in isolation

LM

trained in isolation ← GPT-2, GPT-3, LLAMA, …

How to train it?

Retrieval Model

trained in isolation

← DPR, Contriever, GTR, …

LM

trained in isolation

← GPT-2, GPT-3, LLAMA, …

How to train it?

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

How to train it?

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

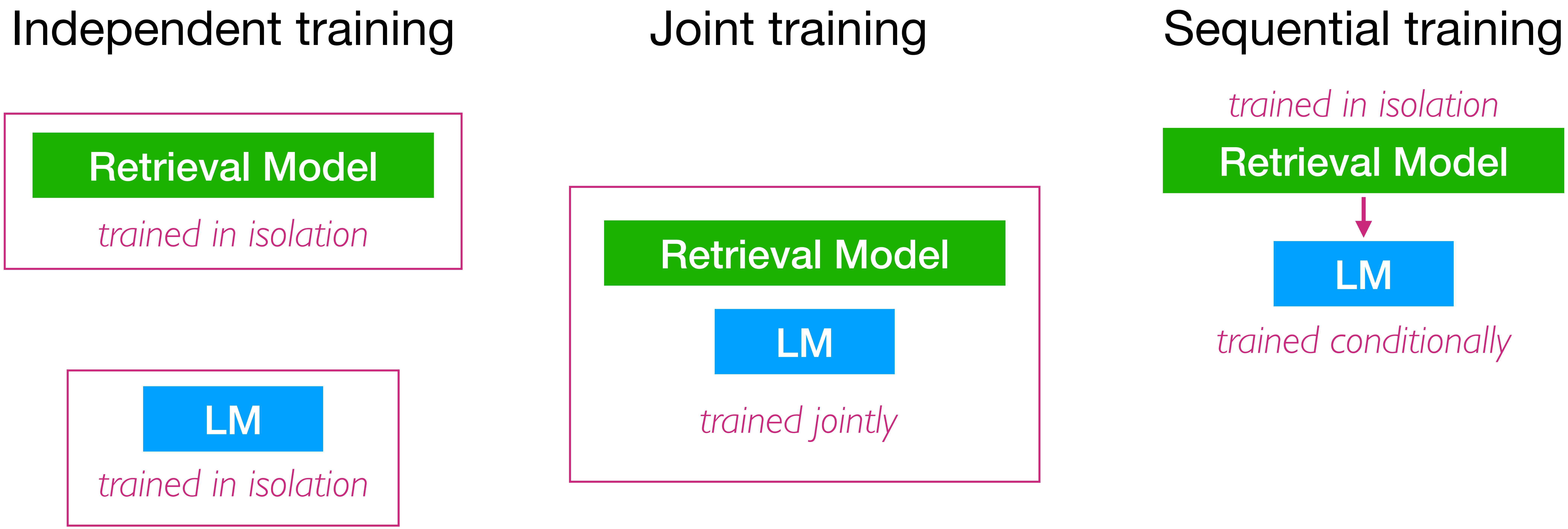
Joint training

Retrieval Model

LM

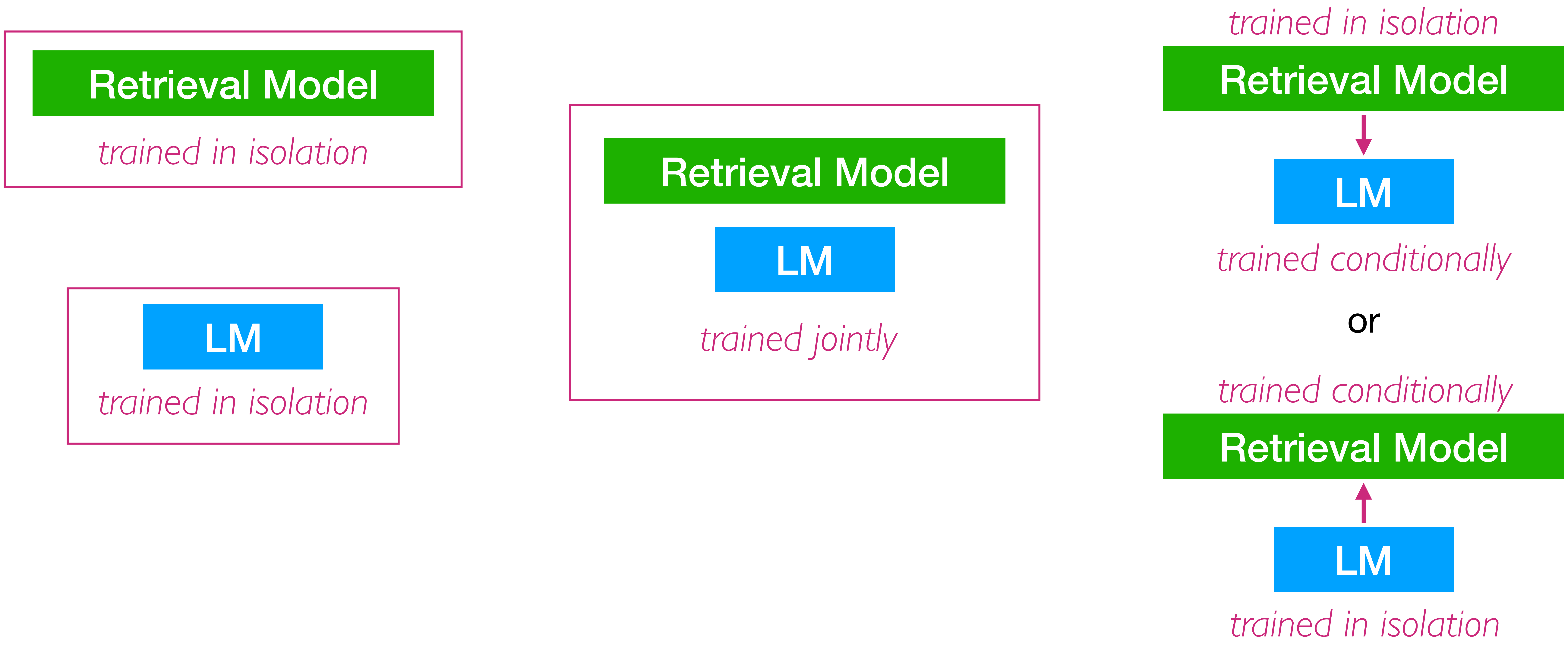
trained jointly

How to train it?



How to train it?

Independent training | Joint training | Sequential training



How to train it?

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

Joint training

(Skipping details)

Retrieval Model

LM

trained jointly

Sequential training

trained in isolation

Retrieval Model

LM

trained conditionally

or

trained conditionally

Retrieval Model

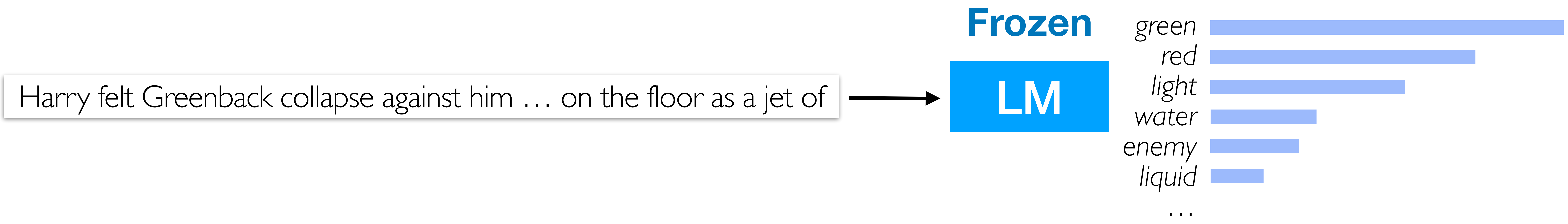
LM

trained in isolation

Sequential training: freeze LM, tune retrieval

Shi et al. 2023. “REPLUG: Retrieval-Augmented Black-Box Language Models”

Sequential training: freeze LM, tune retrieval



Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval

Harry felt Greenback collapse against him … on the floor as a jet of

Ground truth token: *green*

Frozen

LM

LM

Shi et al. 2023. “REPLUG: Retrieval-Augmented Black-Box Language Models”

Sequential training: freeze LM, tune retrieval

Harry felt Greenback collapse against him … on the floor as a jet of

Ground truth token: *green*

Retrieval Model

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light

Voldemort's want just as a jet of red light …

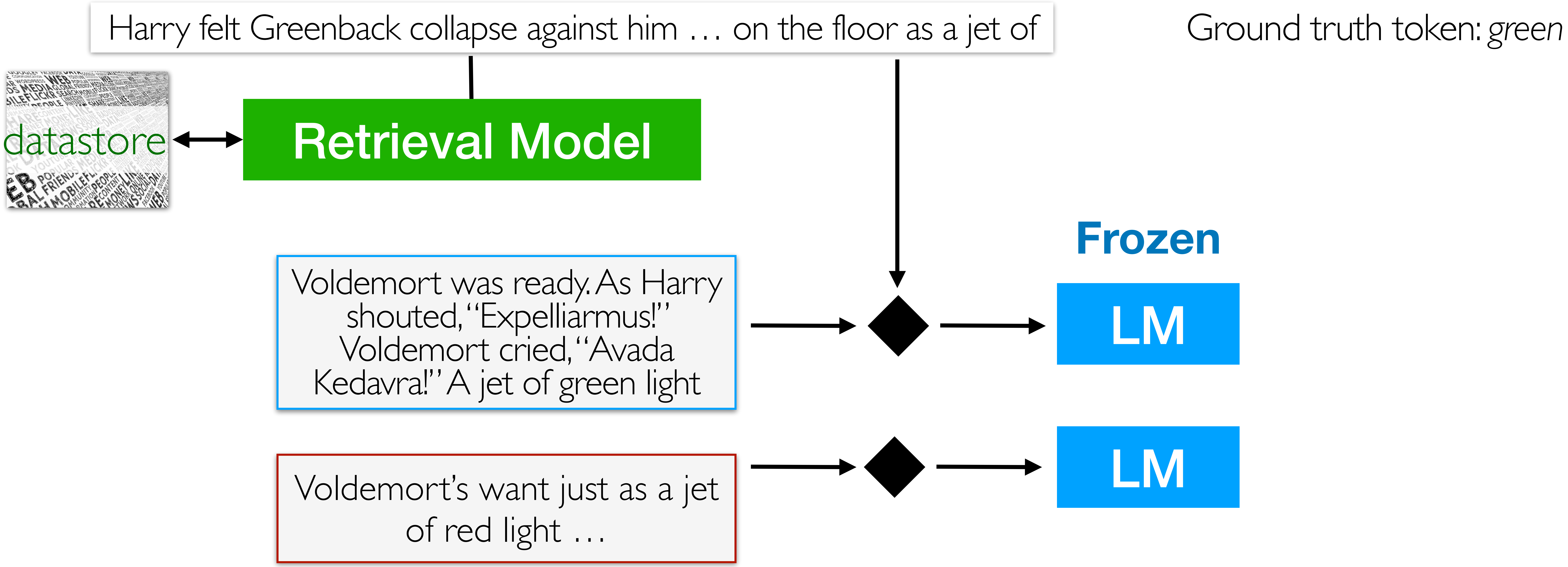
Frozen

LM

LM

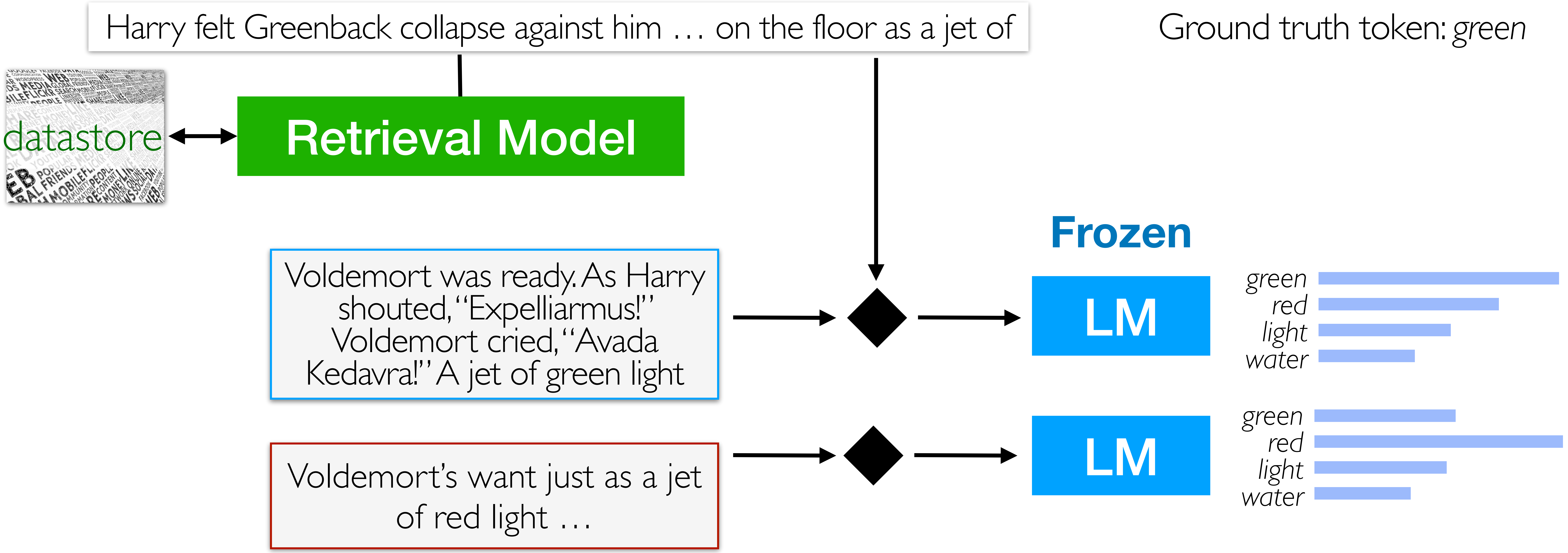
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



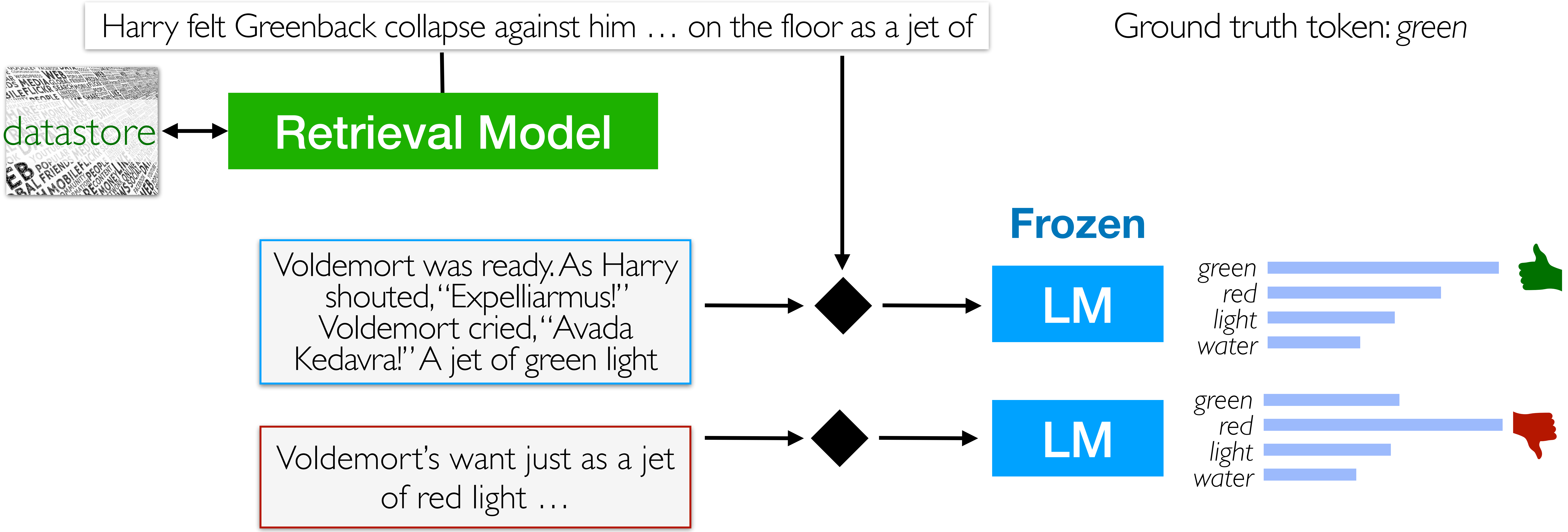
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



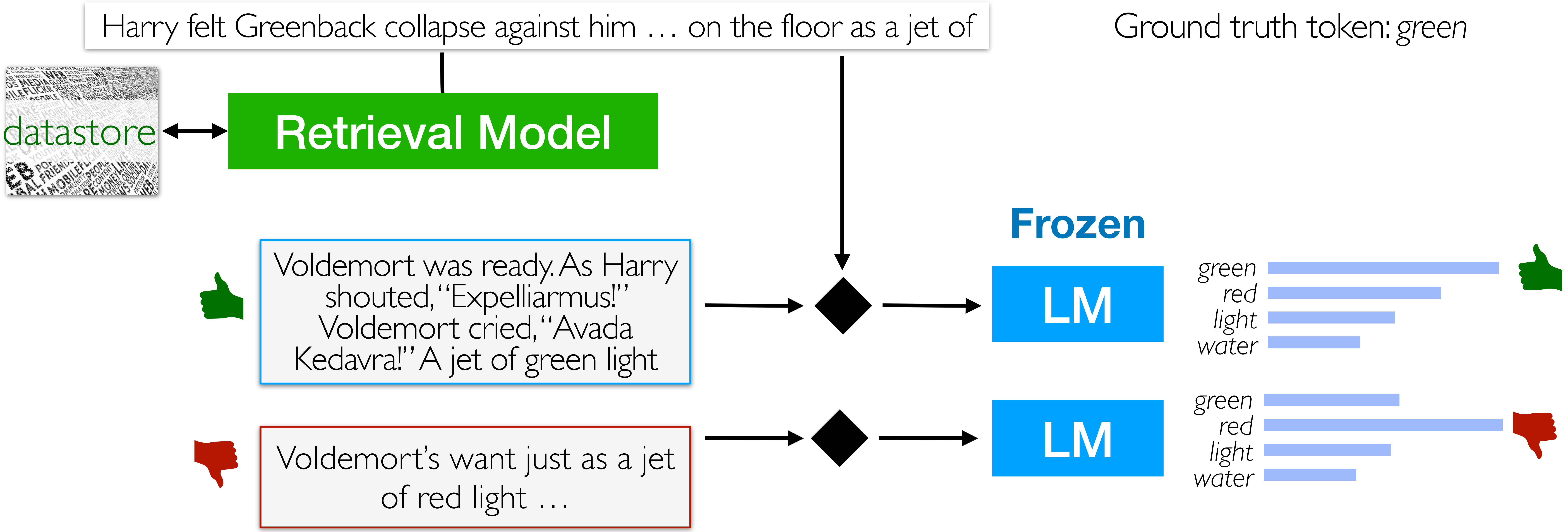
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



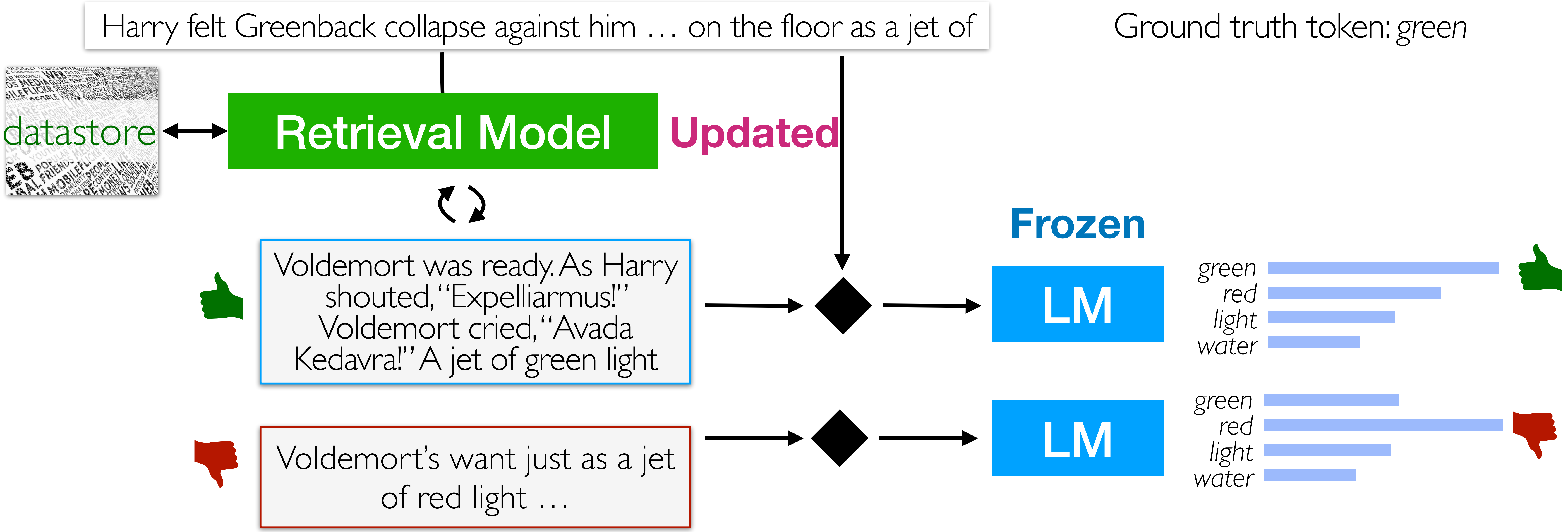
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



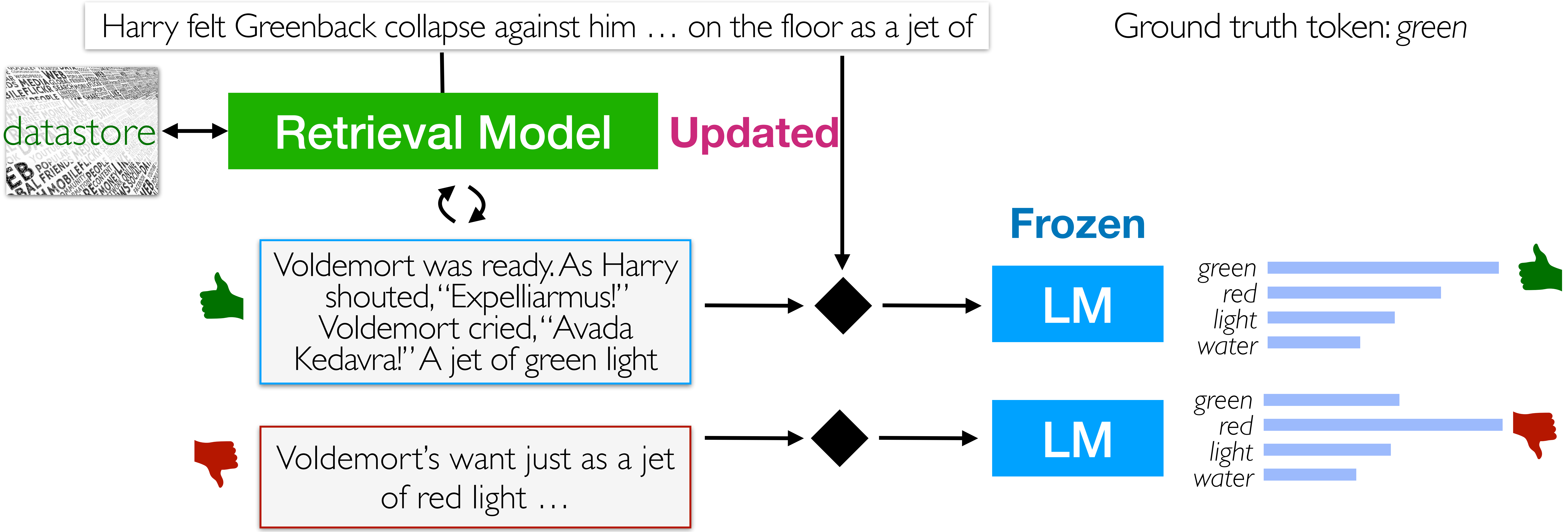
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



Updated

Maximize $P(y \mid x) = \sum_{z \in \mathcal{Z}} \boxed{P_{\text{ret}}(z \mid x)} P_{\text{LM}}(y \mid x, z)$

Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze retrieval, tune LM

Shi et al. 2023. “In-Context Pretraining: Language Modeling Beyond Document Boundaries”

Sequential training: freeze retrieval, tune LM

Harry felt Greenback collapse against him … on the floor as a jet of

Ground truth token: *green*

Shi et al. 2023. "In-Context Pretraining: Language Modeling Beyond Document Boundaries"

Sequential training: freeze retrieval, tune LM

Harry felt Greenback collapse against him … on the floor as a jet of

Ground truth token: *green*

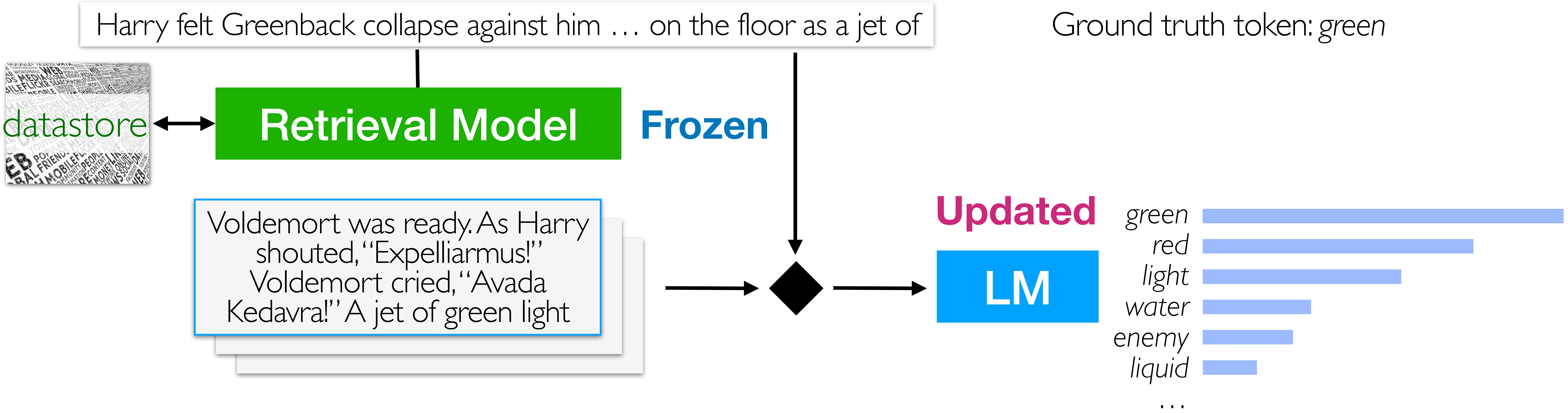
Retrieval Model

Frozen

Voldemort was ready. As Harry shouted, "Expelliarmus!" Voldemort cried, "Avada Kedavra!" A jet of green light

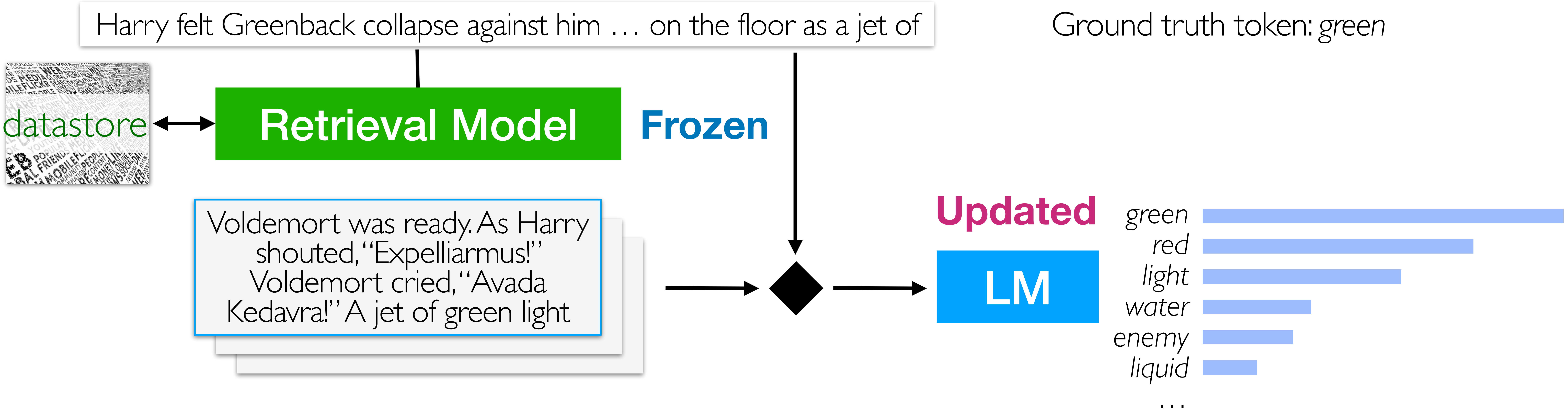
Shi et al. 2023. "In-Context Pretraining: Language Modeling Beyond Document Boundaries"

Sequential training: freeze retrieval, tune LM



Shi et al. 2023. "In-Context Pretraining: Language Modeling Beyond Document Boundaries"

Sequential training: freeze retrieval, tune LM



Updated

$$\text{Maximize } P(y \mid x) = \sum_{z \in \mathcal{Z}} P_{\text{ret}}(z \mid x) \boxed{P_{\text{LM}}(y \mid x, z)}$$

Shi et al. 2023. "In-Context Pretraining: Language Modeling Beyond Document Boundaries"

Summary: Training

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

Joint training

(Skipping details)

Retrieval Model

LM

trained jointly

Sequential training

trained in isolation

Retrieval Model

LM

trained conditionally

or

trained in isolation

Retrieval Model

LM

trained conditionally

Summary: Training

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

Joint training

(Skipping details)

Retrieval Model

LM

trained jointly

Quite difficult, essentially iterative sequential training

Sequential training

trained in isolation

Retrieval Model

LM

trained conditionally

or

trained in isolation

Retrieval Model

LM

trained conditionally

Gus et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"
Izcard et al. 2022. "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Summary: Training

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

Good enough if you want minimal effort

Joint training

(Skipping details)

Retrieval Model

LM

trained jointly

Principle way but still open question

Sequential training

trained in isolation

Retrieval Model

LM

trained conditionally

or

trained in isolation

Retrieval Model

LM

trained conditionally

Good middle ground

Retrieval augmentation: Overview

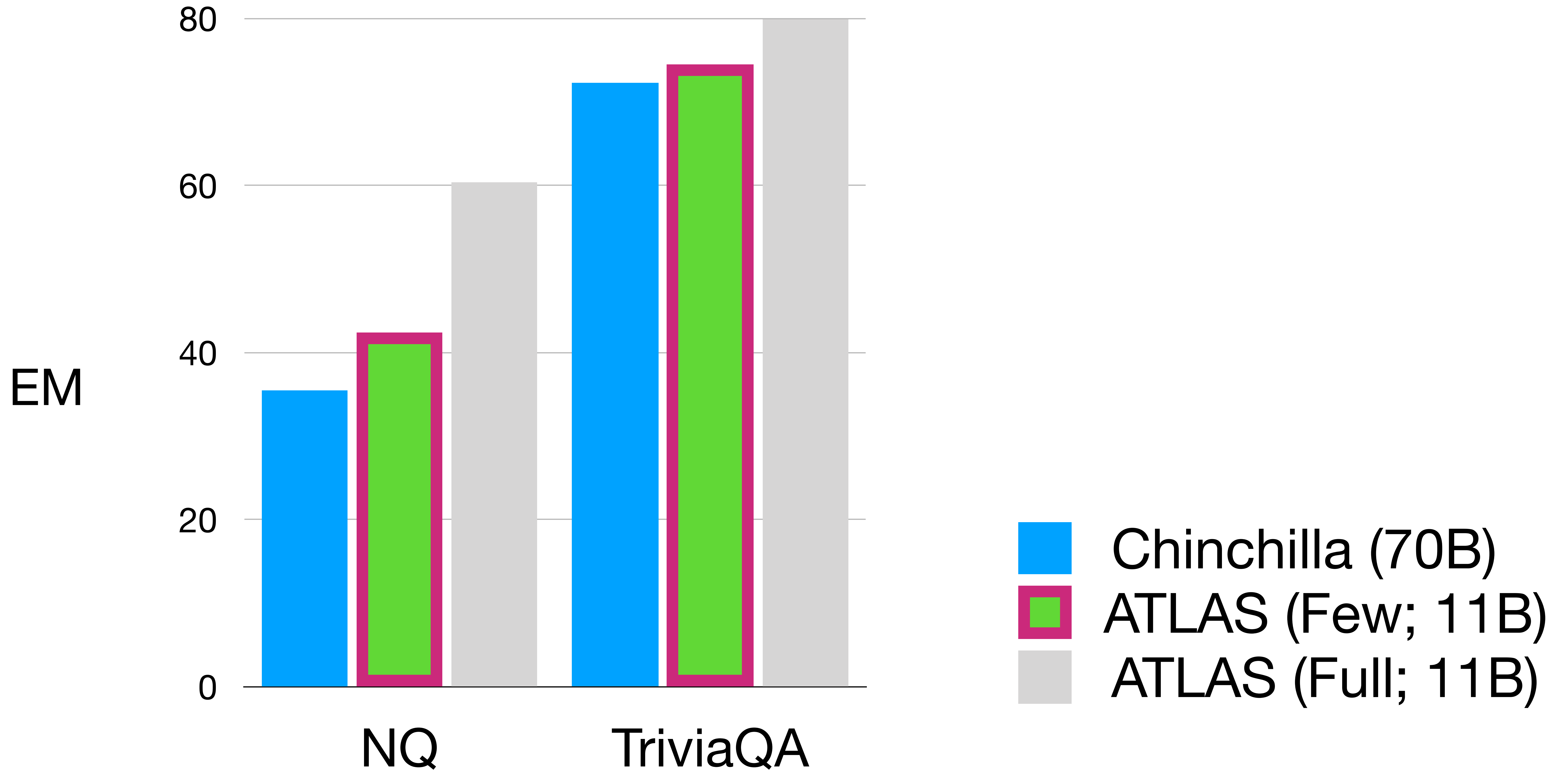
- Inference
  - Step 1: Retrieve
  - Step 2: Read (Generate)
  - Optionally, with multiple passages: Concatenation, Ensembling, Reranking
- Training
  - **Independent training, Joint training, Sequential training**
- Key results

Retrieval augmentation: Overview

- Inference
  - Step 1: Retrieve
  - Step 2: Read (Generate)
  - Optionally, with multiple passages: Concatenation, Ensembling, Reranking
- Training
  - Independent training, Joint training, Sequential training
- **Key results**

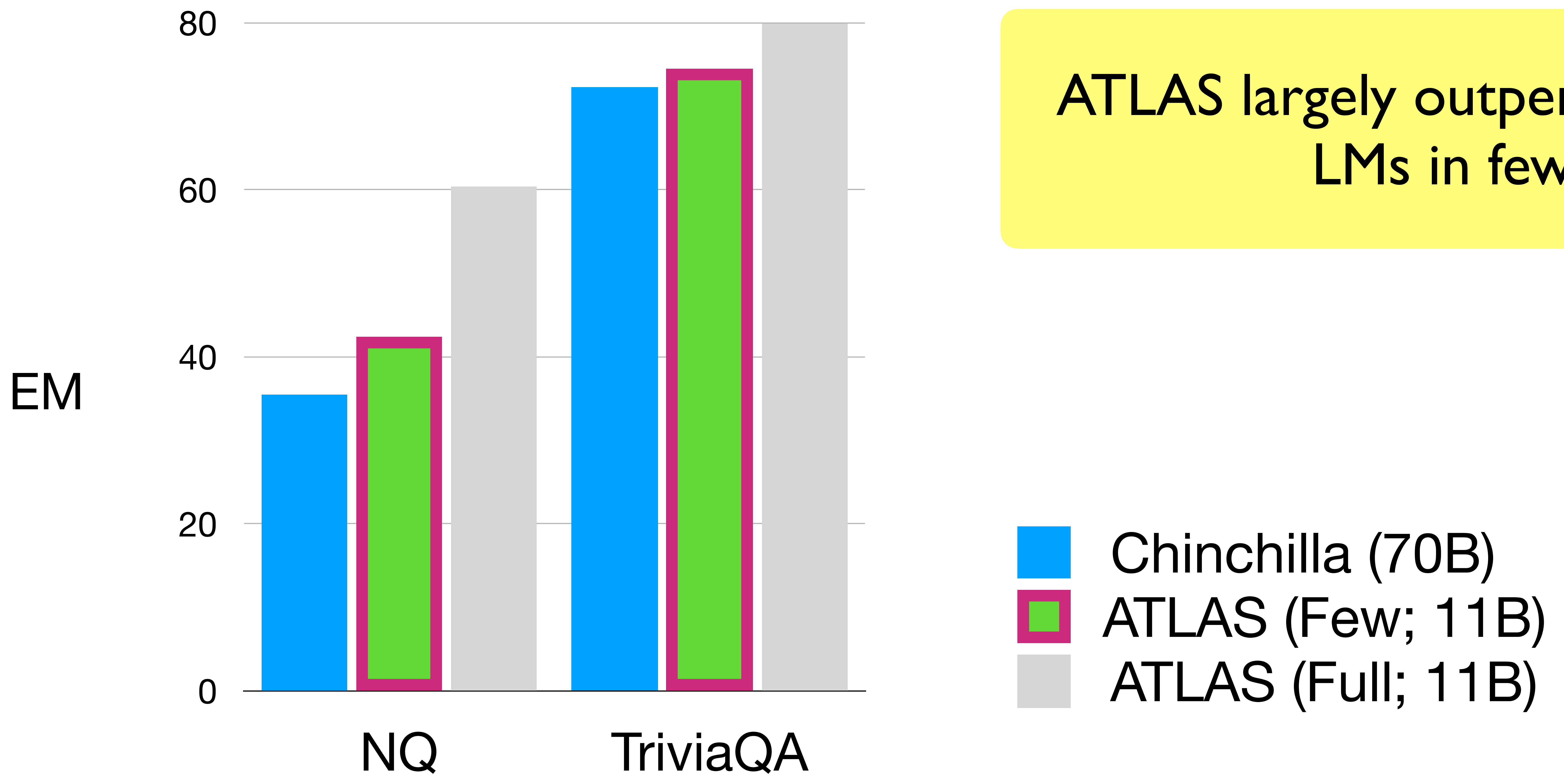
Question Answering

Question Answering



Izcard et al. “Atlas: Few-shot Learning with Retrieval Augmented Language Models”

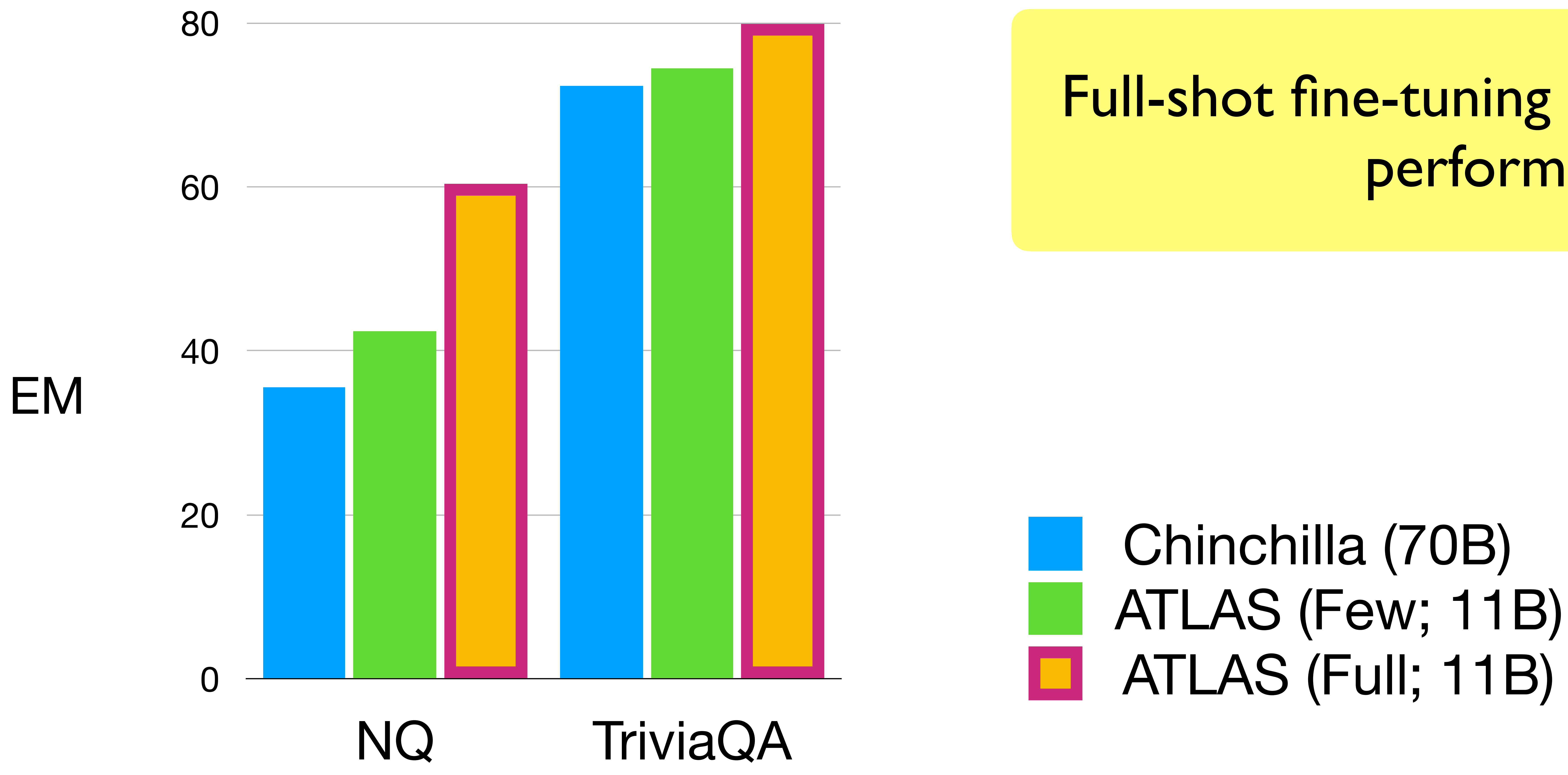
Question Answering



ATLAS largely outperforms 7x larger LMs in few-shot

Izcard et al. "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Question Answering

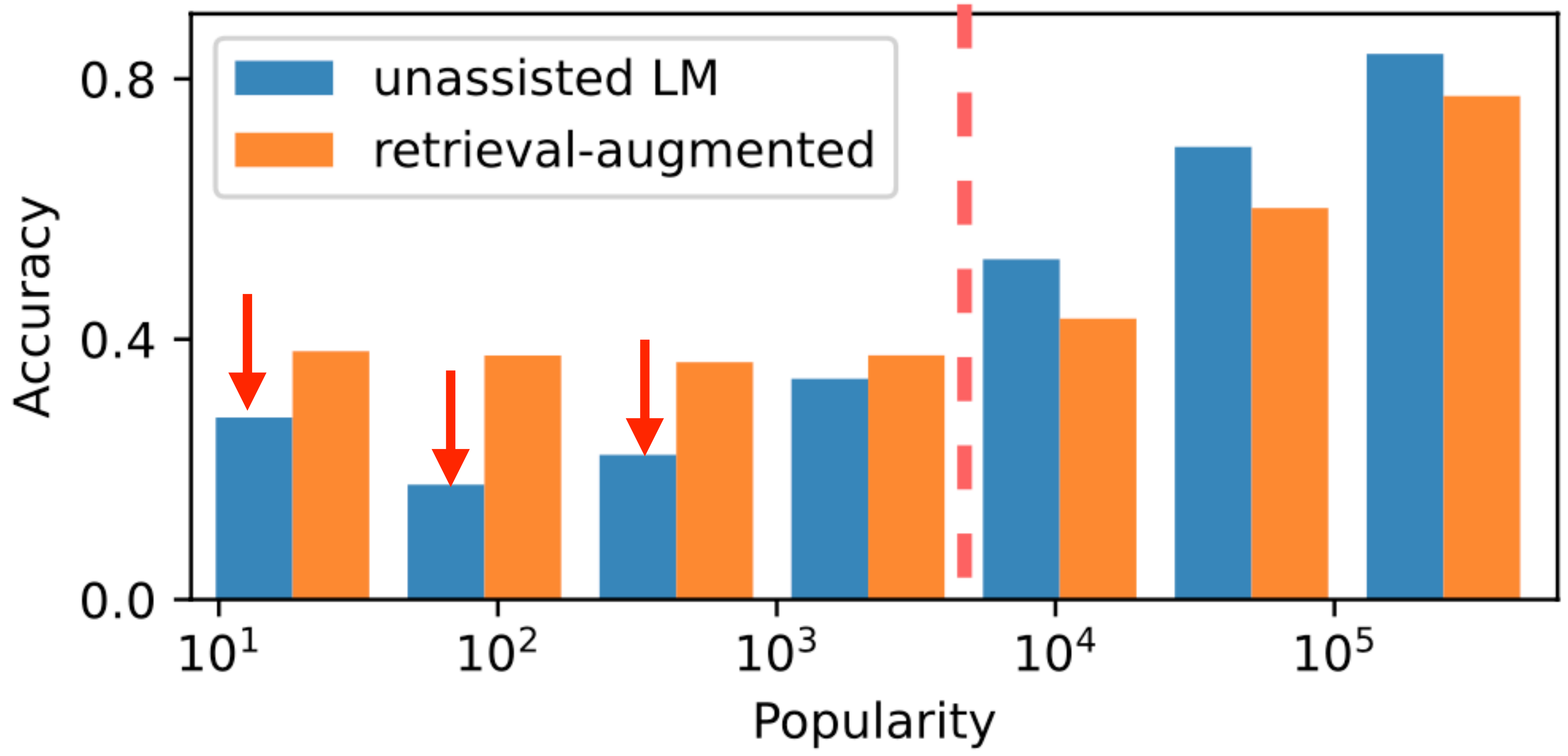


Full-shot fine-tuning further improves performance

Izcard et al. "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Question Answering

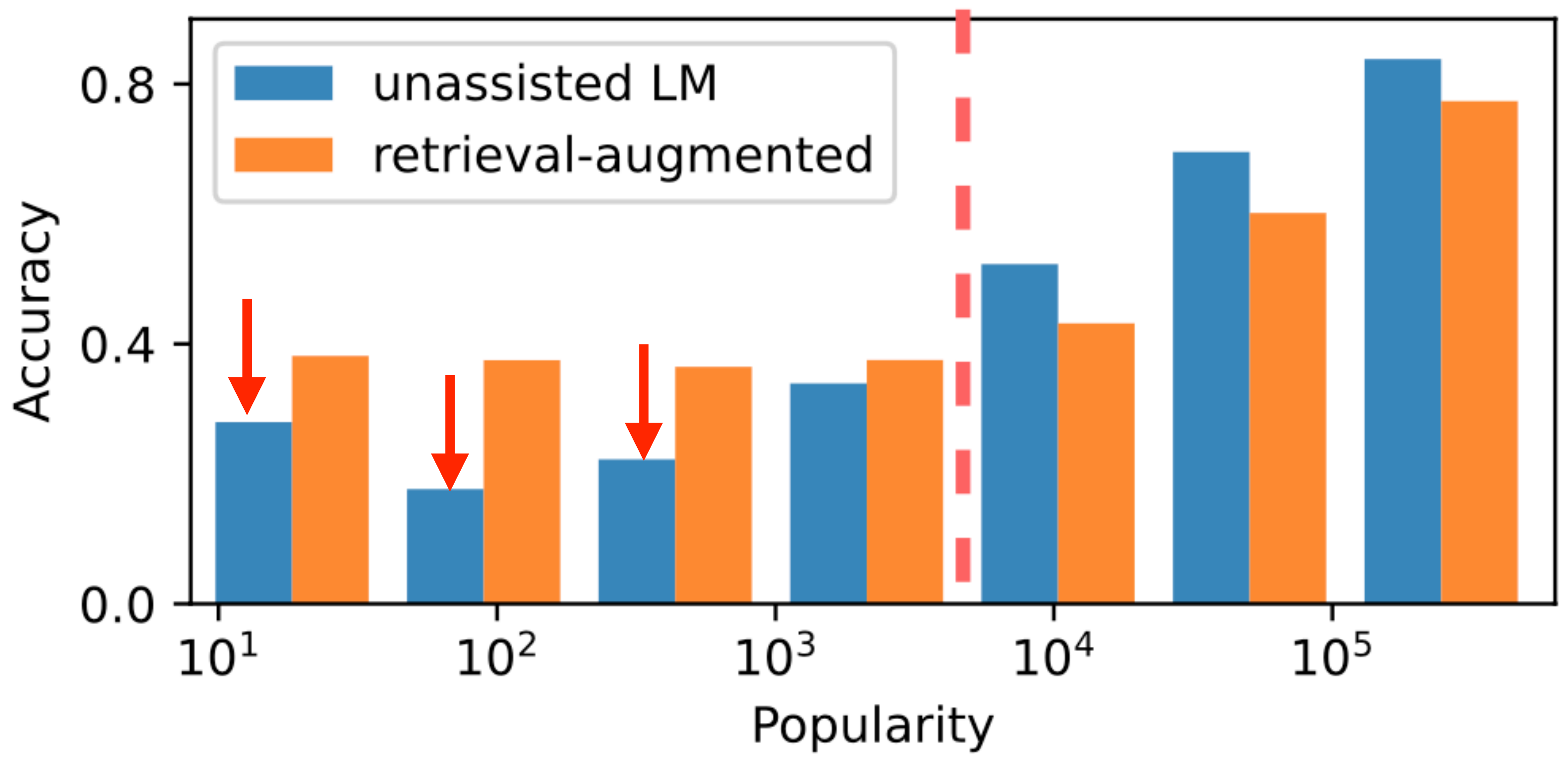
What is Kathy Saltzman's occupation?



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

Question Answering

What is Kathy Saltzman's occupation?

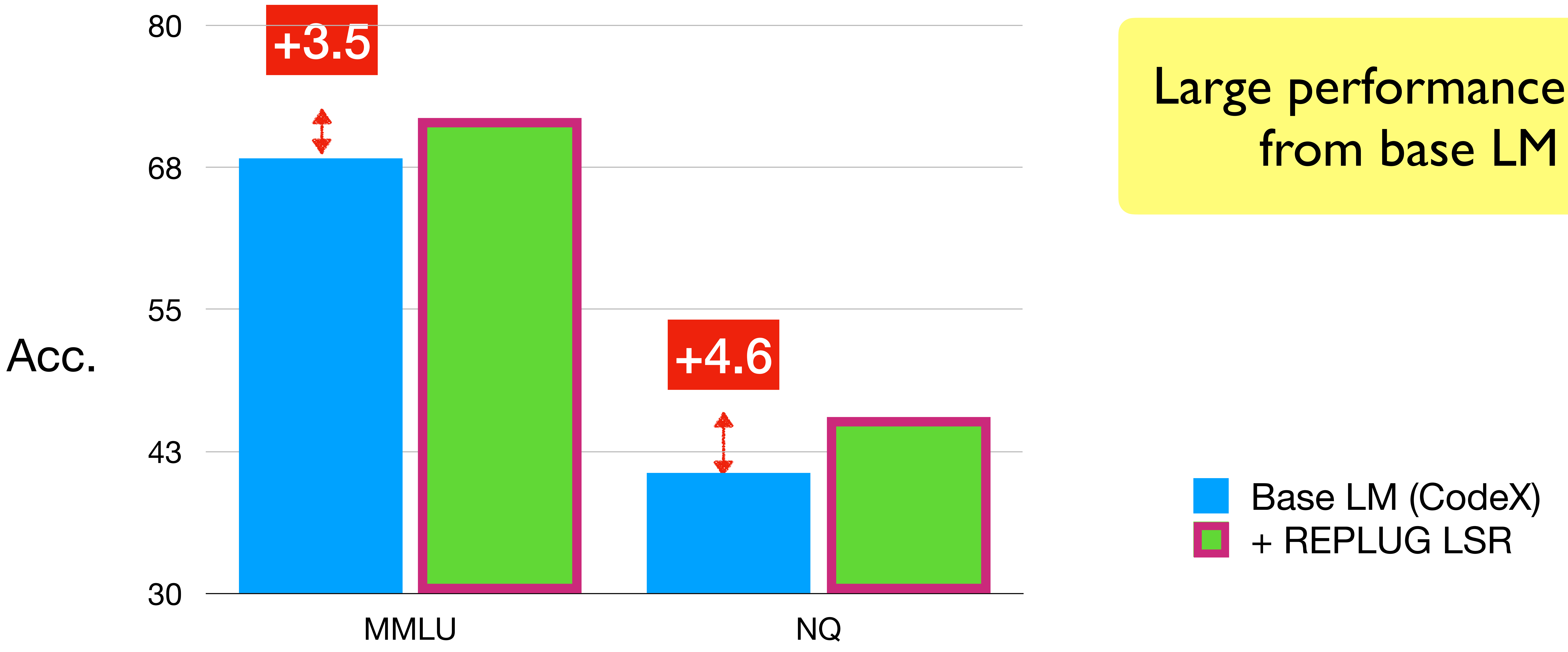


Gains increase as the rarity increases (even over GPT-3!)

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

Reasoning (MMLU)

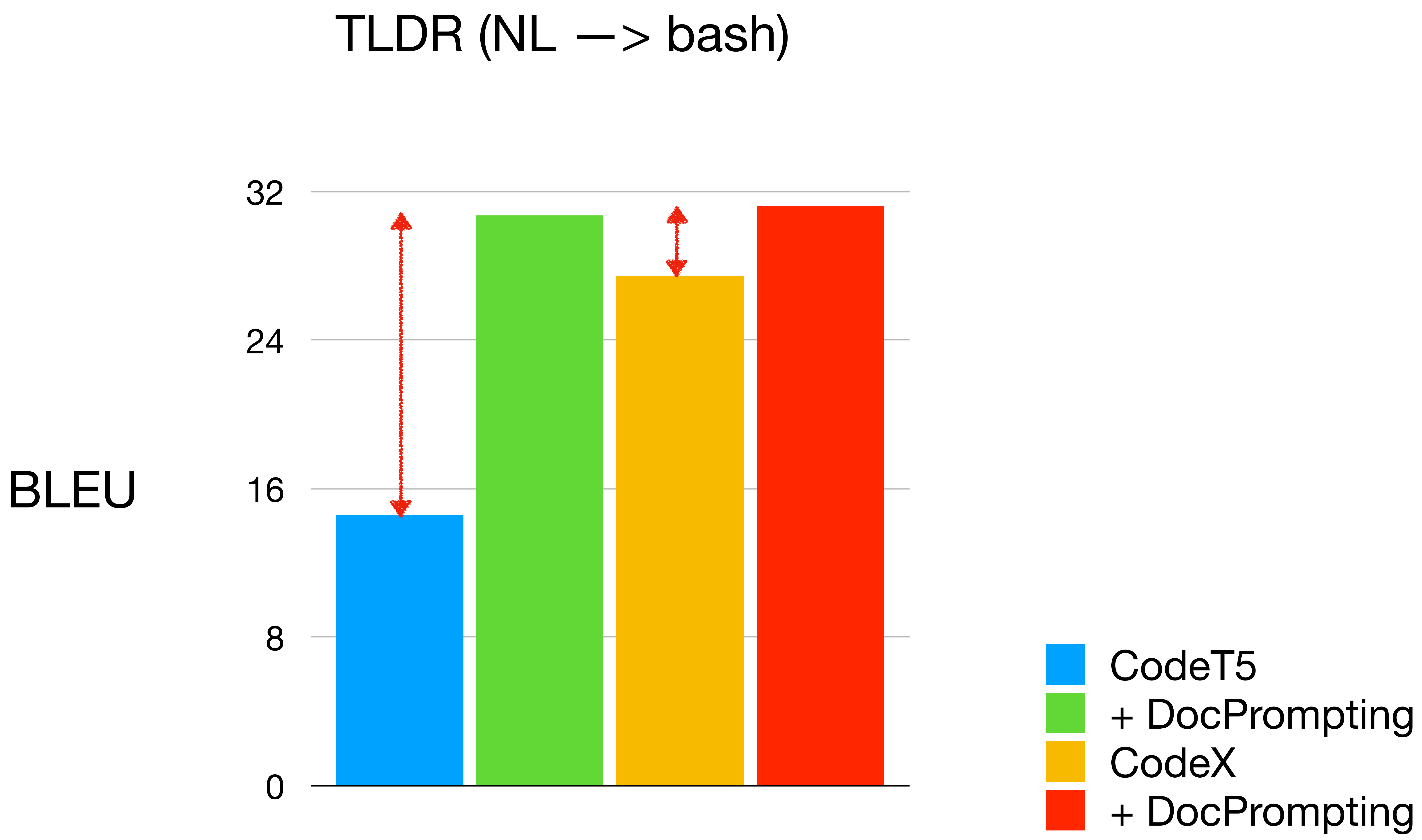
Reasoning (MMLU)



Large performance gain from base LM

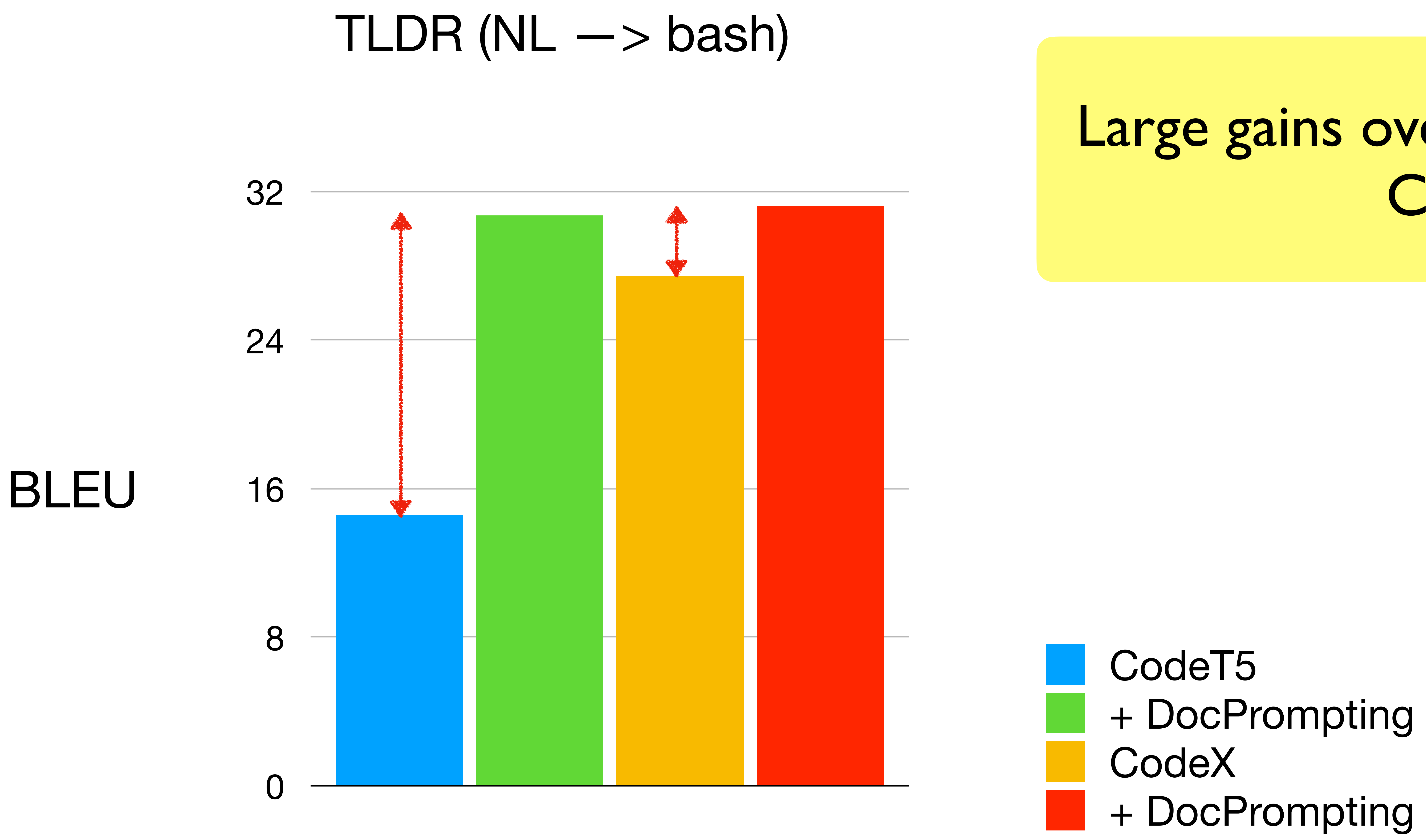
Shi et al. 2023. “REPLUG: Retrieval-Augmented Black-Box Language Models”

Code generation



Zhou et al. 2023. “DocPrompting: Generating Code by Retrieving the Docs”

Code generation

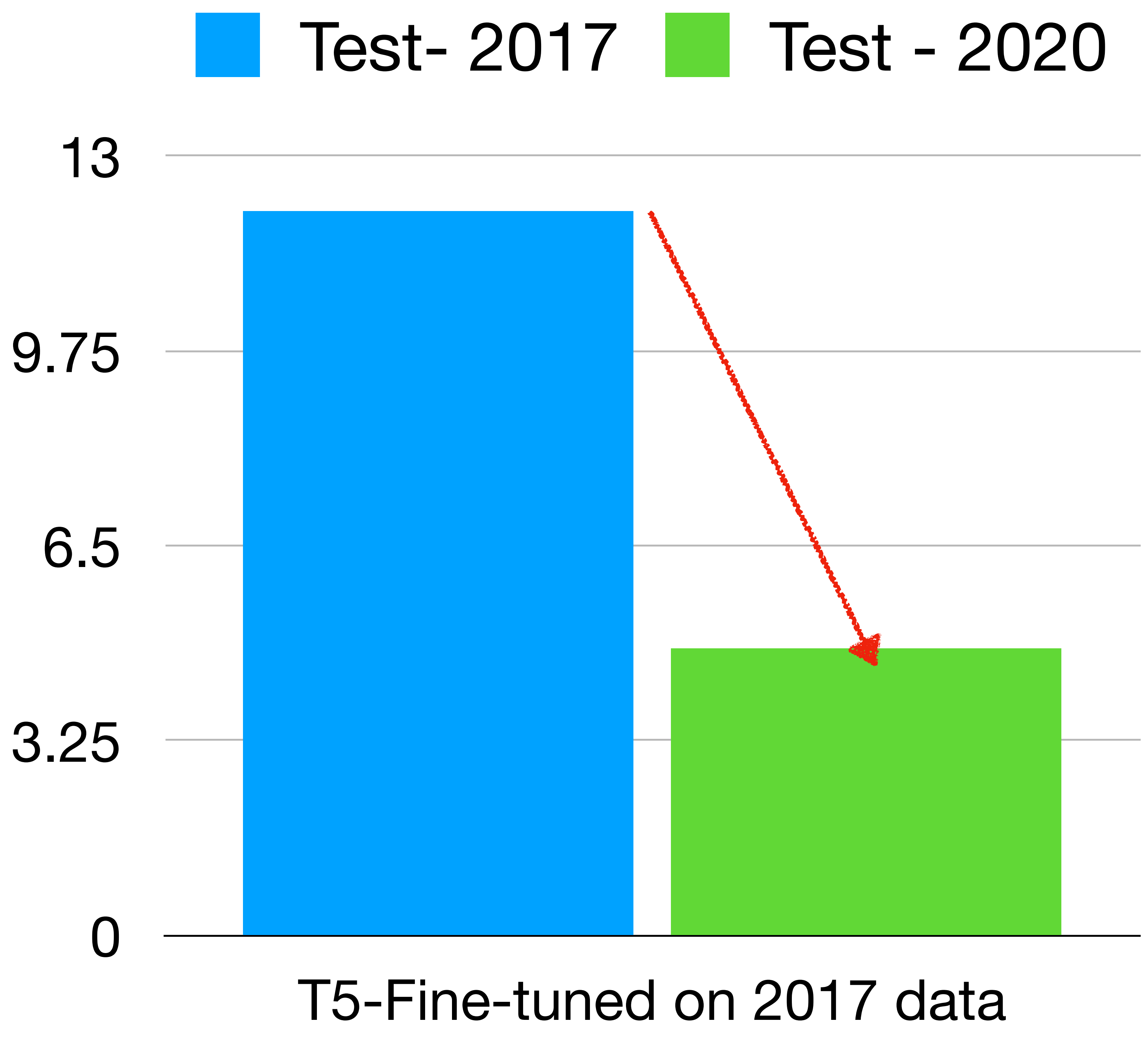


Large gains over both CodeT5 & CodeX

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

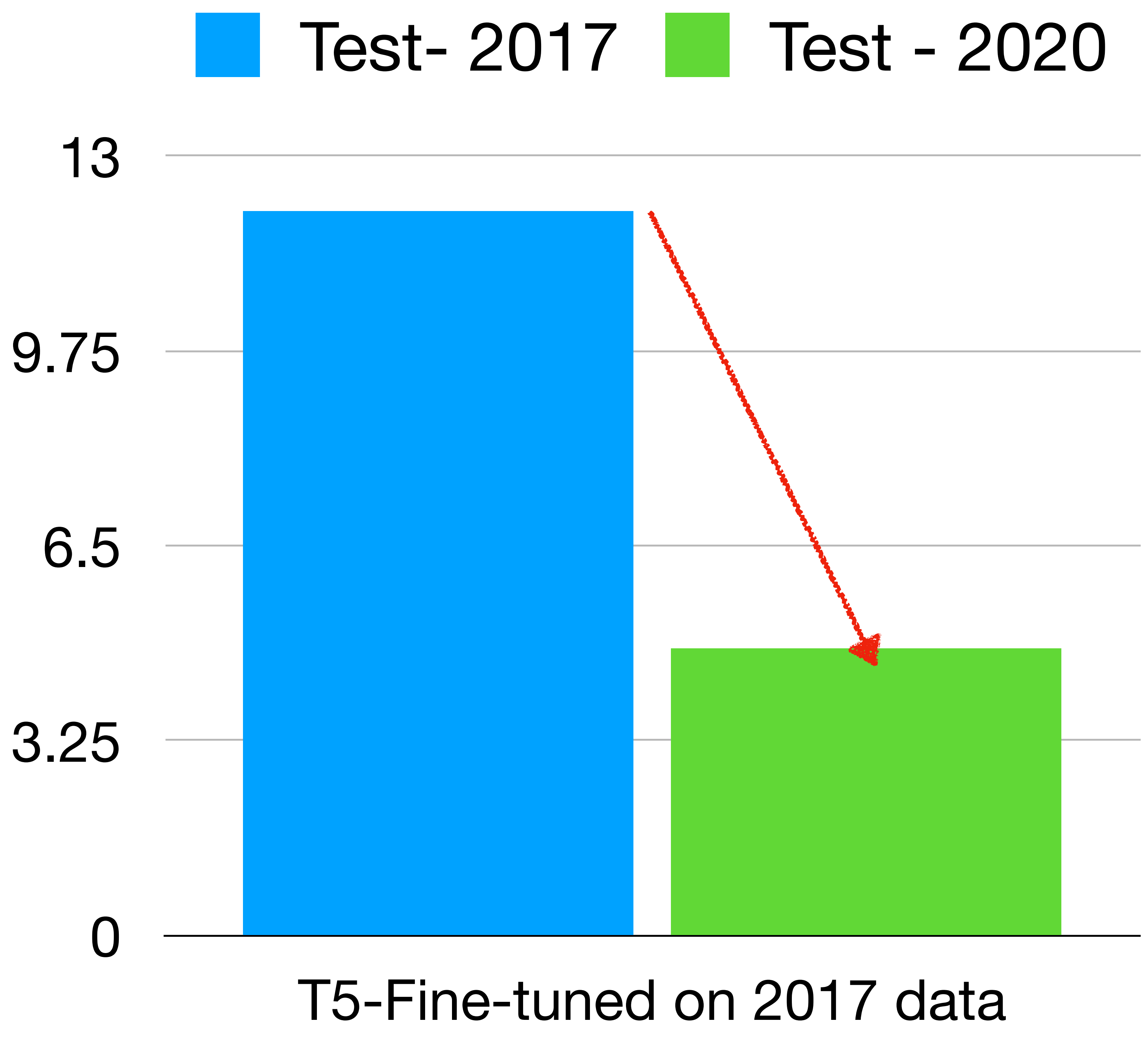
Can update effectively

Can update effectively



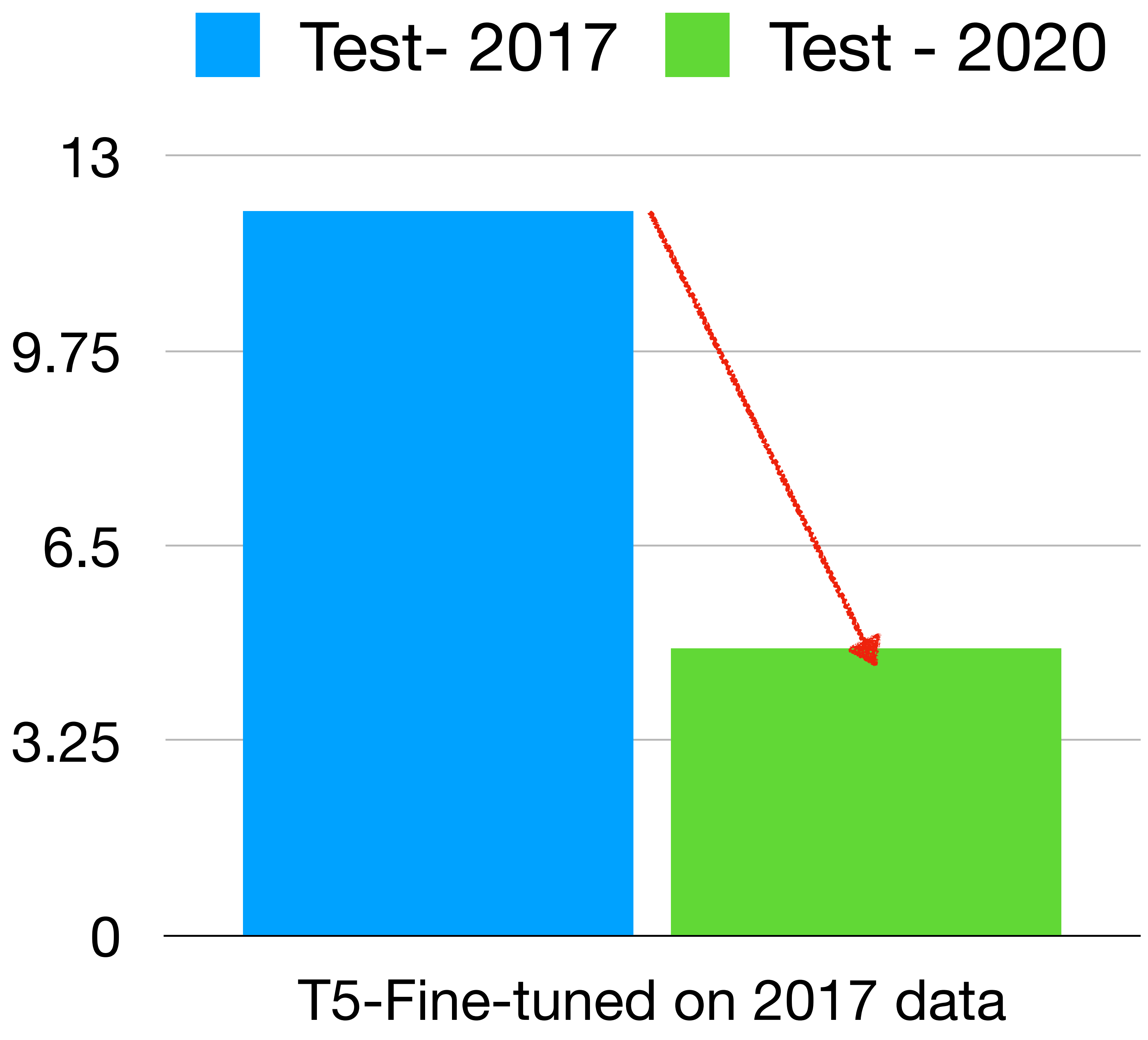
Izcard et al. “Atlas: Few-shot Learning with Retrieval Augmented Language Models”

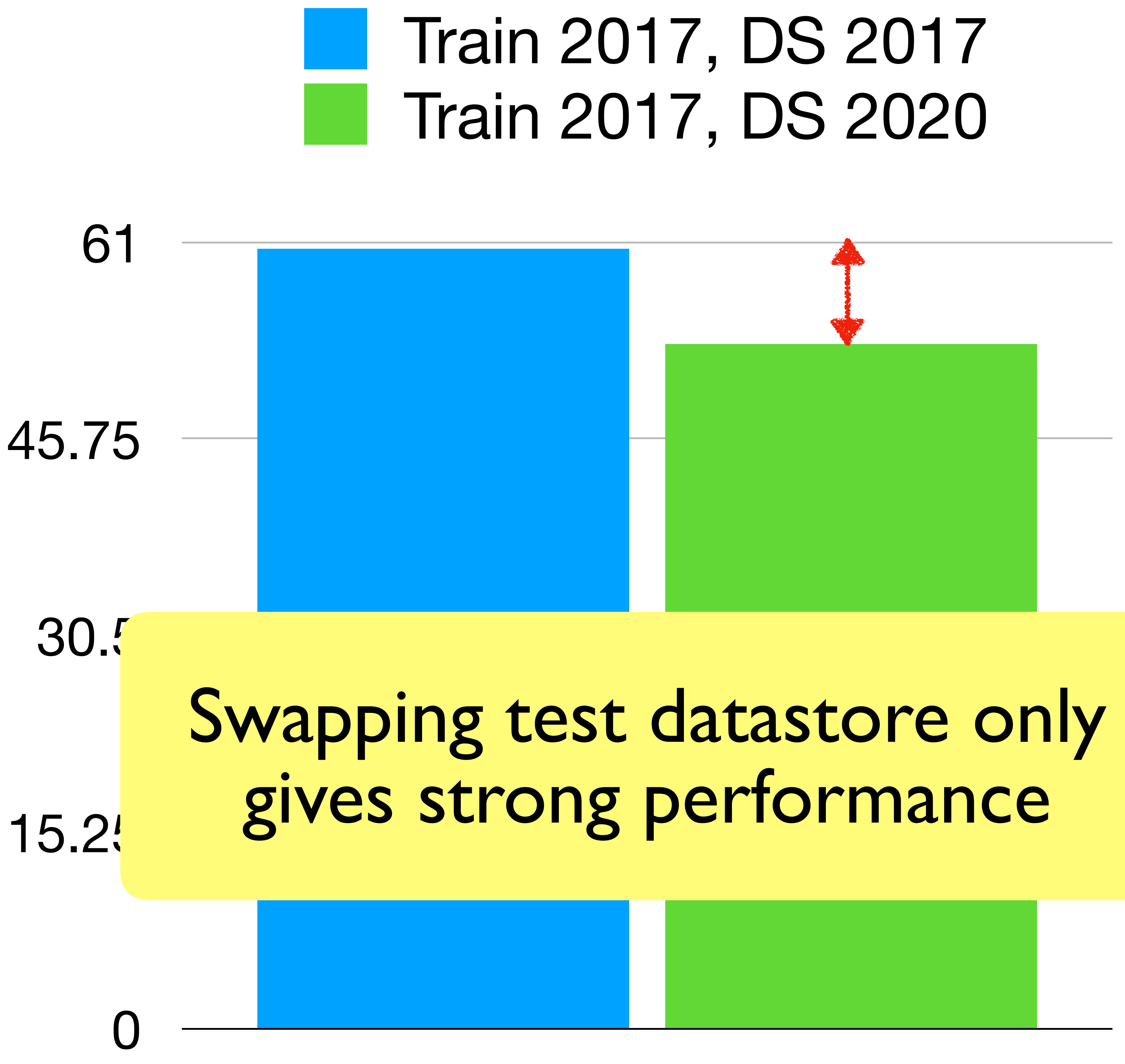
Can update effectively



Izcard et al. "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Can update effectively



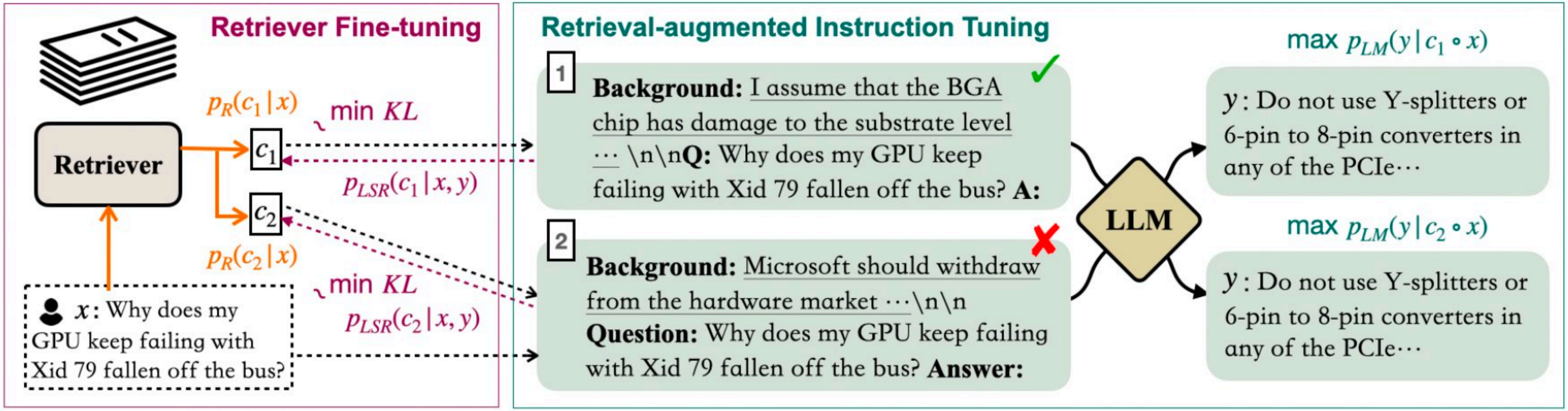


Swapping test datastore only gives strong performance

Izcard et al. "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

Instruction-tuning

Instruction-tuning



Lin et al. 2023. RA-DIT: Retrieval-Augmented Dual Instruction Tuning

Retrieval augmentation: Summary

- Inference
  - Step 1: Retrieve
  - Step 2: Read (Generate)
  - Optionally, with multiple passages: Concatenation, Ensembling, Reranking
- Training
  - Independent training, Joint training, Sequential training
- Key results
  - QA, Reasoning, Code generation, etc
  - Update effectively
  - You can also do instruction tuning

Retrieval Augmented LMs are already being used!

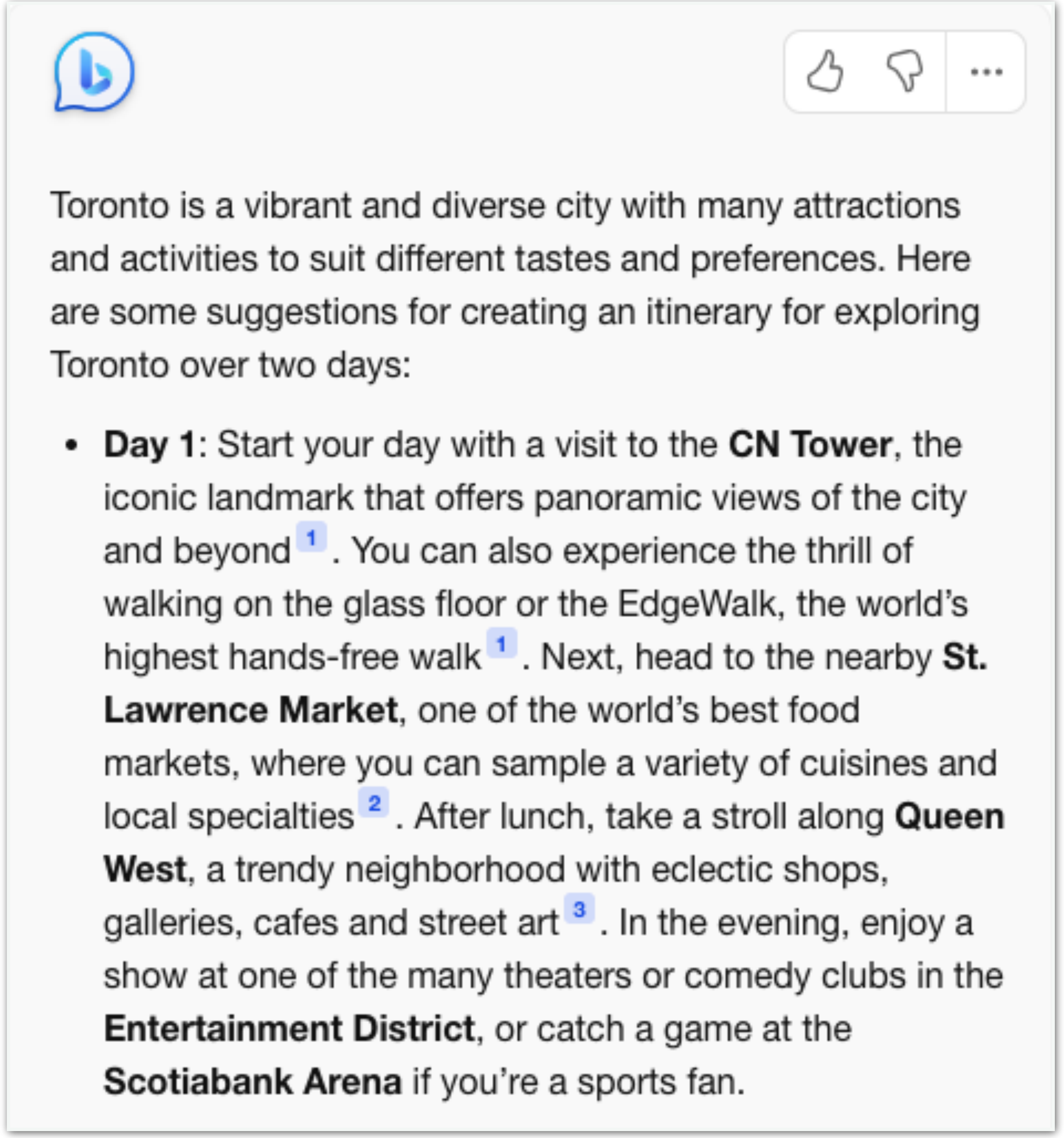
Chat GPT Extension

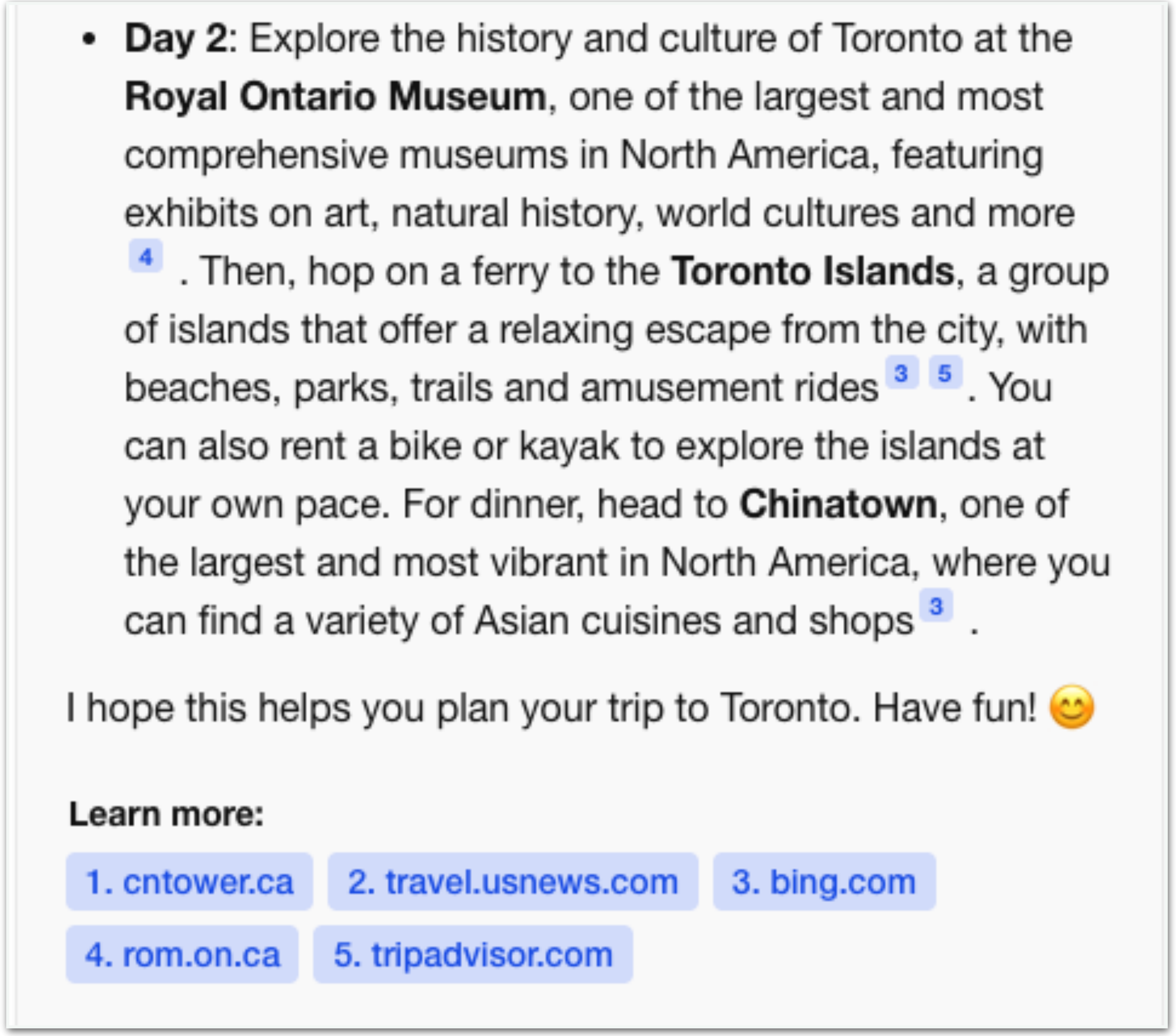
YOU

Perplexity

llamaindex

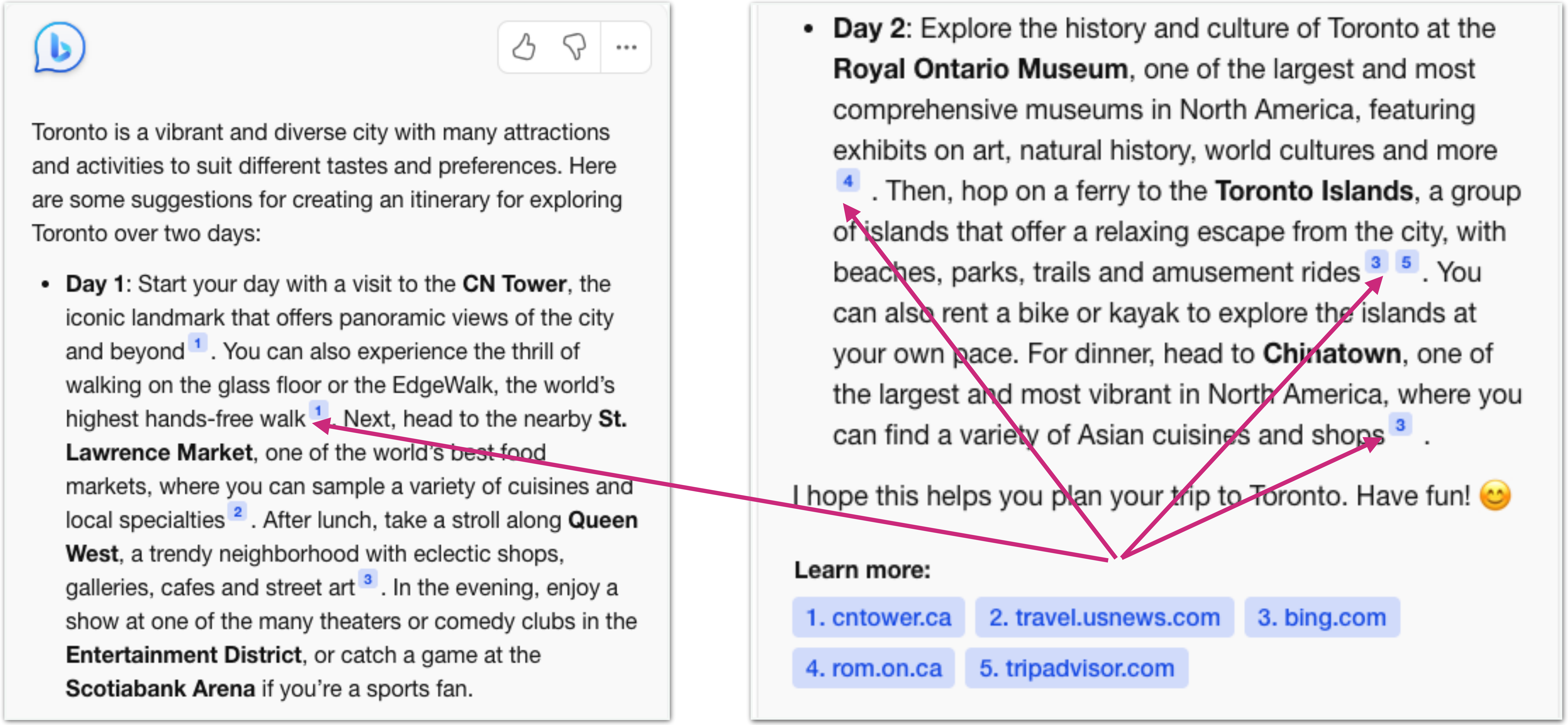
Retrieval Augmented LMs are already being used!





Liu et al. 2023. "Evaluating Verifiability in Generative Search Engines"
Min et al. 2023. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation"

Retrieval Augmented LMs are already being used!



Liu et al. 2023. "Evaluating Verifiability in Generative Search Engines"
Min et al. 2023. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation"

Overview

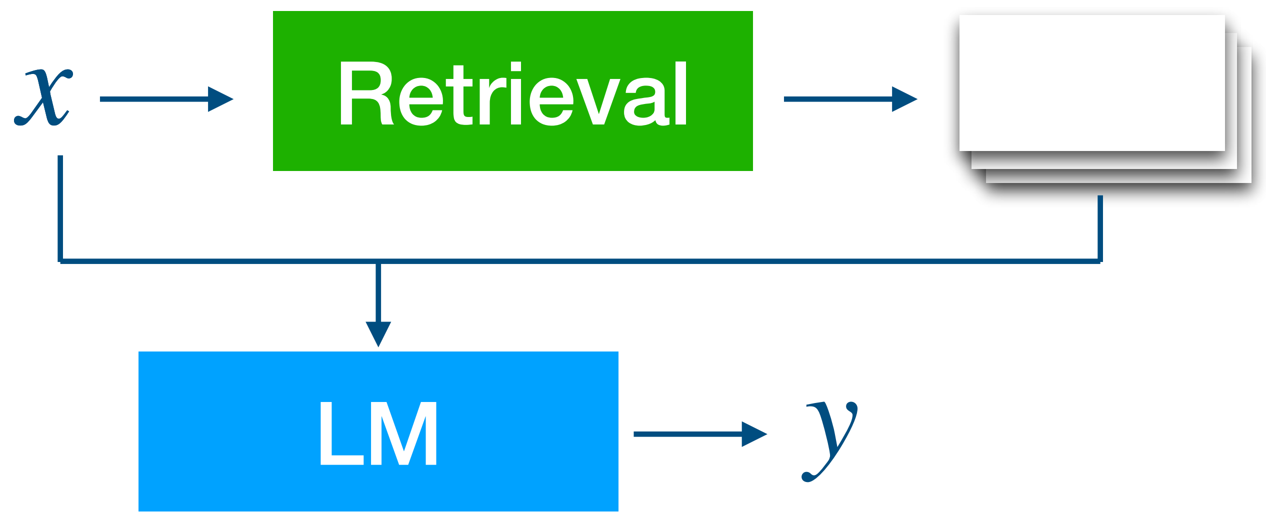
Why Retrieval-based LMs?

Tell me about Meta Platform.

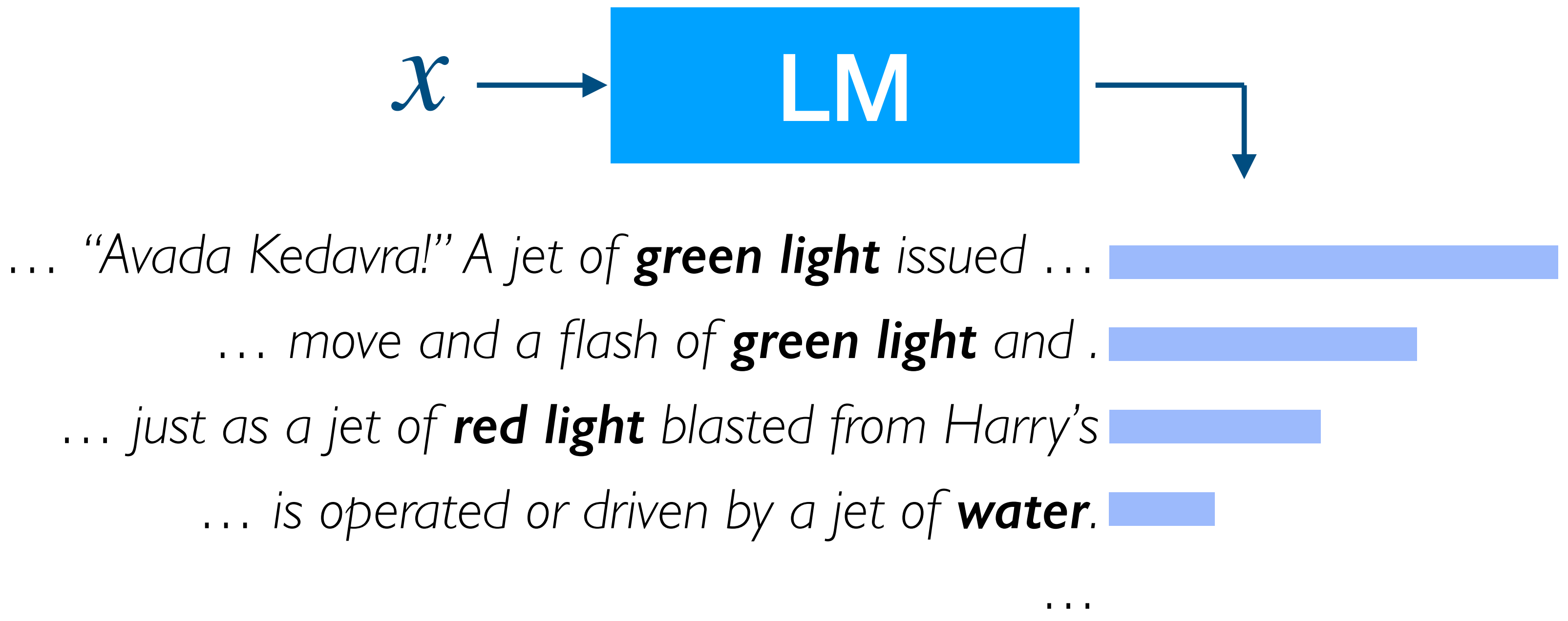
ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Retrieval Augmentation



New Retrieval-based LMs



Open Problems

Scaling **datastore** not just parameters?

New Retrieval-based LMs

- New Methodology 1 — Designing a new Transformer
- New Methodology 2 — Designing a new Softmax
- New LM Design — Mitigating fairness & legality issues

New Retrieval-based LMs

1. How to overcome sequence length limit issue?

2. How to overcome efficiency issue when retrieving ***many*** blocks, ***frequently***?

- **New Methodology 1 — Designing a new Transformer**
- New Methodology 2 — Designing a new Softmax
- New LM Design — Mitigating fairness & legality issues

RETRO (Borgeaud et al. 2021)

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

RETRO (Borgeaud et al. 2021)

New Transformers layers, designed to read *many* text blocks, *frequently*, more *efficiently*

Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

RETRO (Borgeaud et al. 2021)

x = World Cup 2022 was the last with 32 teams, before the increase to

Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

RETRO (Borgeaud et al. 2021)

x = World Cup 2022 was / the last with 32 teams, / before the increase to

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

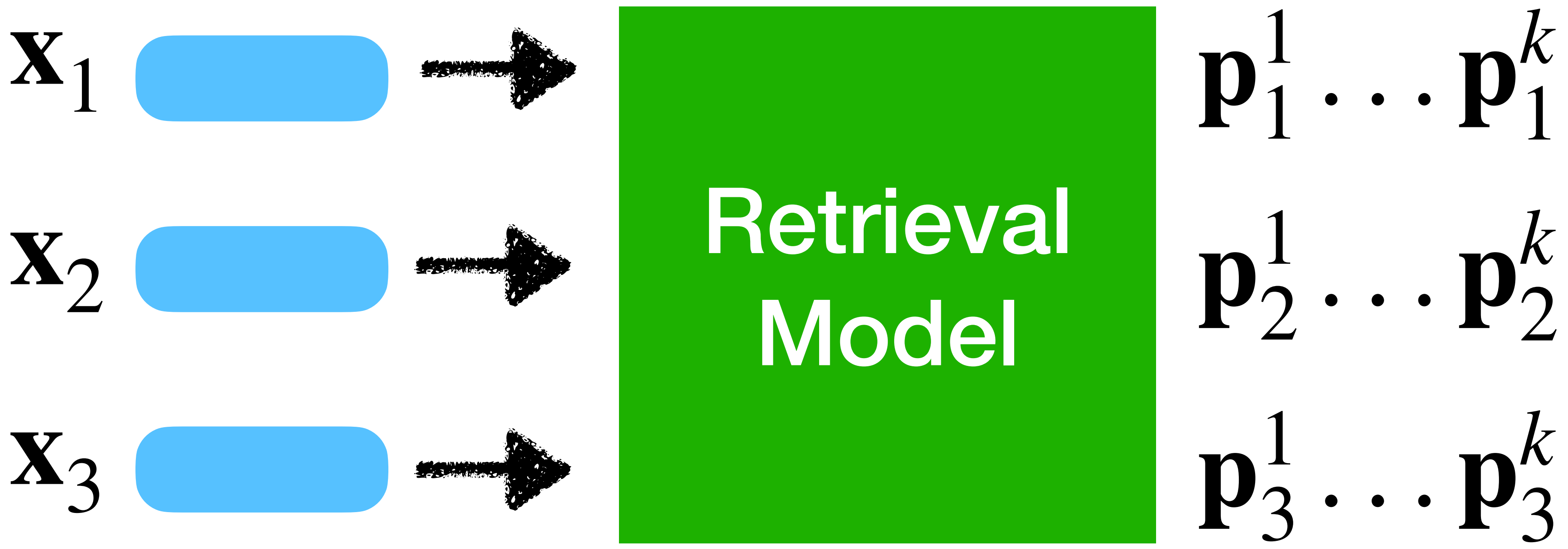
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

RETRO (Borgeaud et al. 2021)

$\boldsymbol{x}$ = World Cup 2022 was | the last with 32 teams, | before the increase to

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

(k text blocks per split)



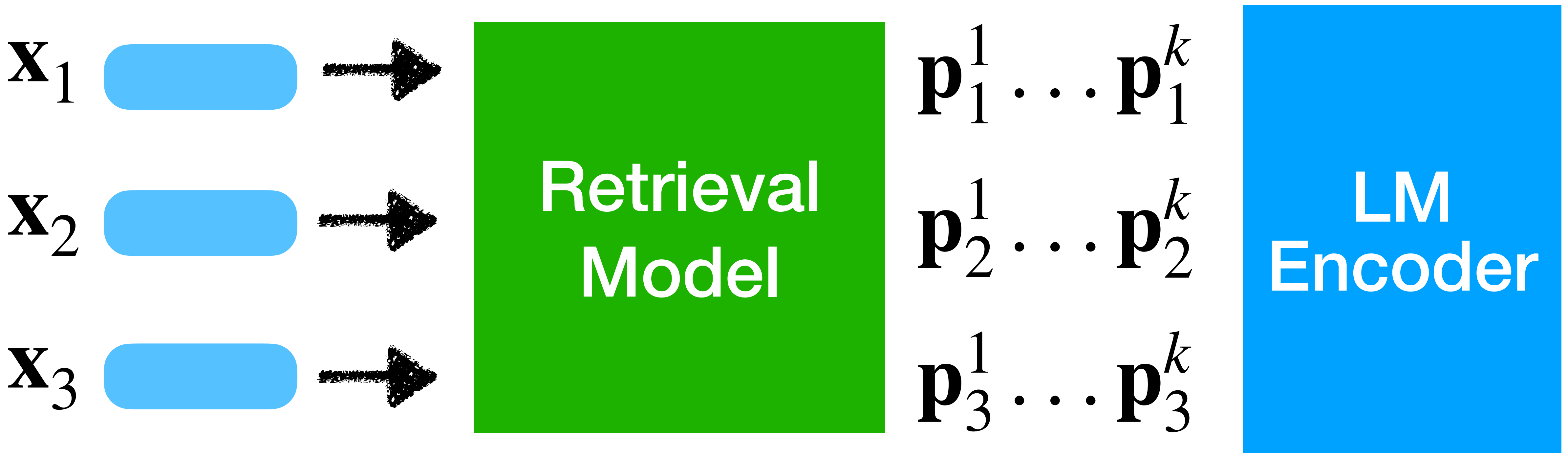
Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

RETRO (Borgeaud et al. 2021)

$\boldsymbol{x}$ = World Cup 2022 was the last with 32 teams, before the increase to

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

(k text blocks per split)



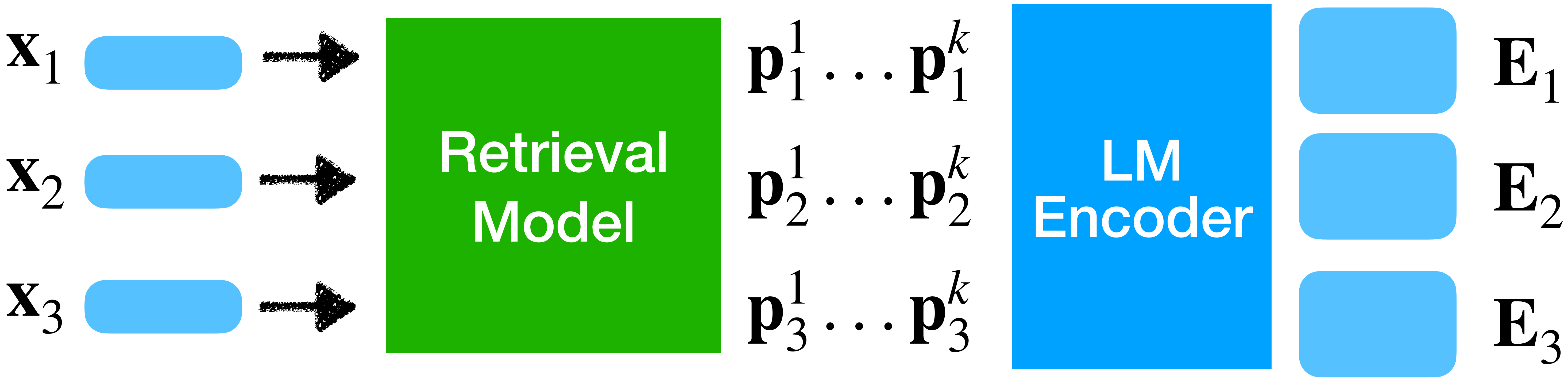
Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

RETRO (Borgeaud et al. 2021)

$\boldsymbol{x}$ = World Cup 2022 was the last with 32 teams, before the increase to

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

(k text blocks per split)



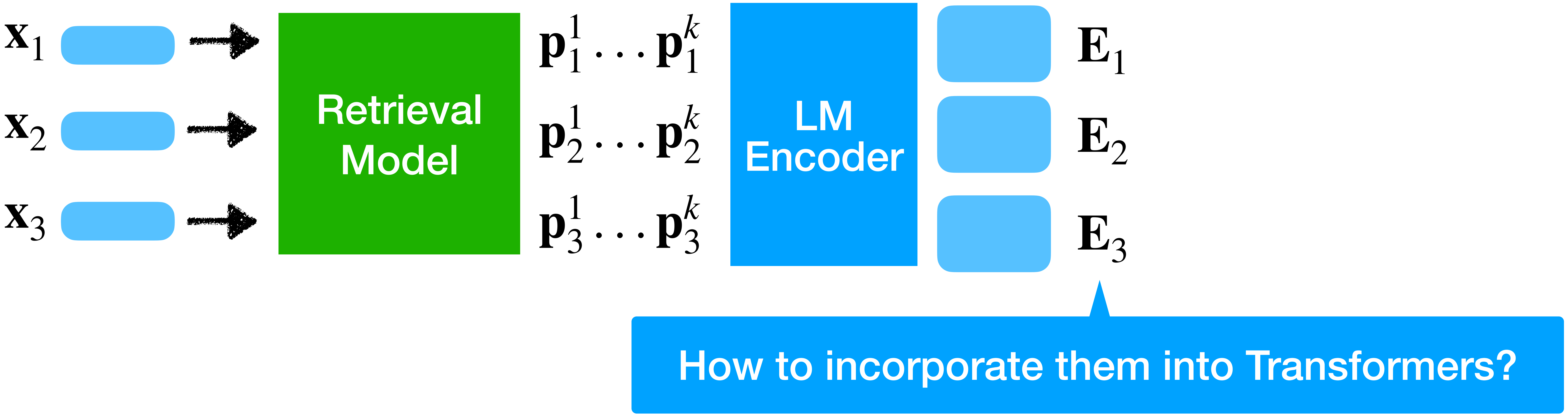
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

RETRO (Borgeaud et al. 2021)

$\boldsymbol{x}$ = World Cup 2022 was / the last with 32 teams, / before the increase to

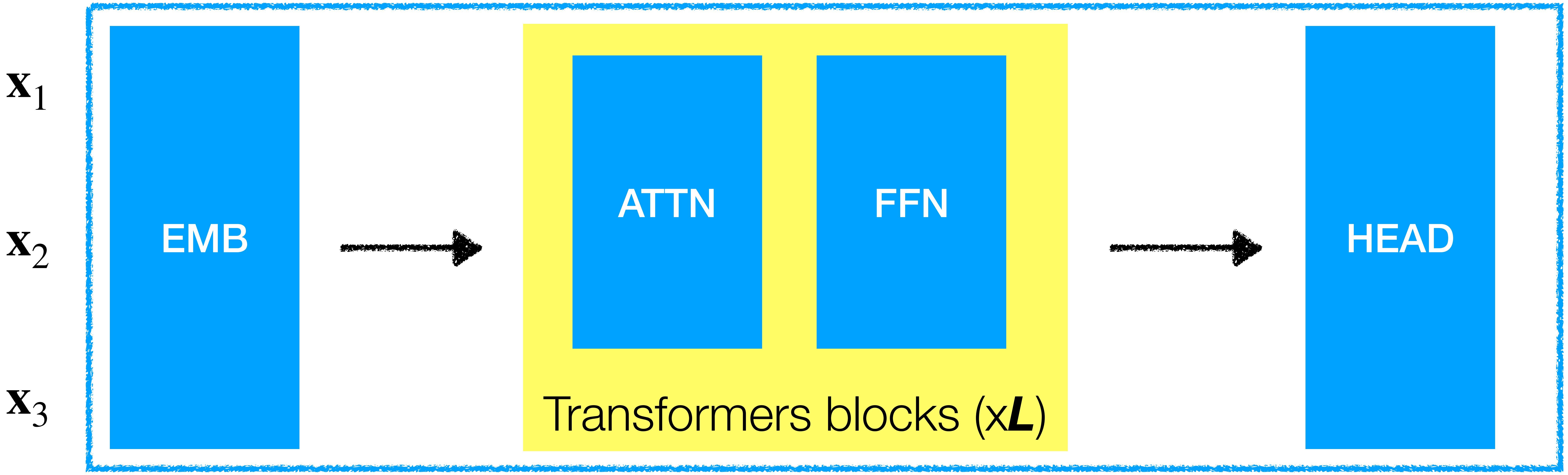
$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

(k text blocks per split)



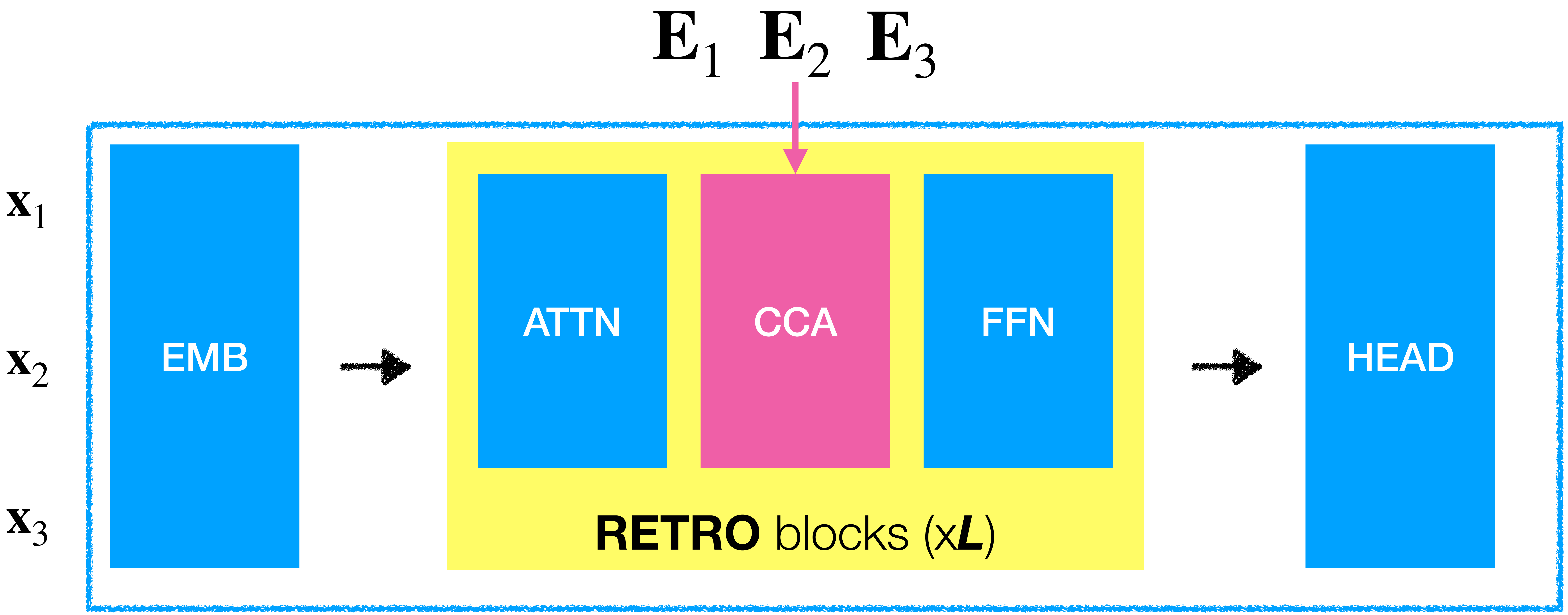
Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

Regular Transformers



Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

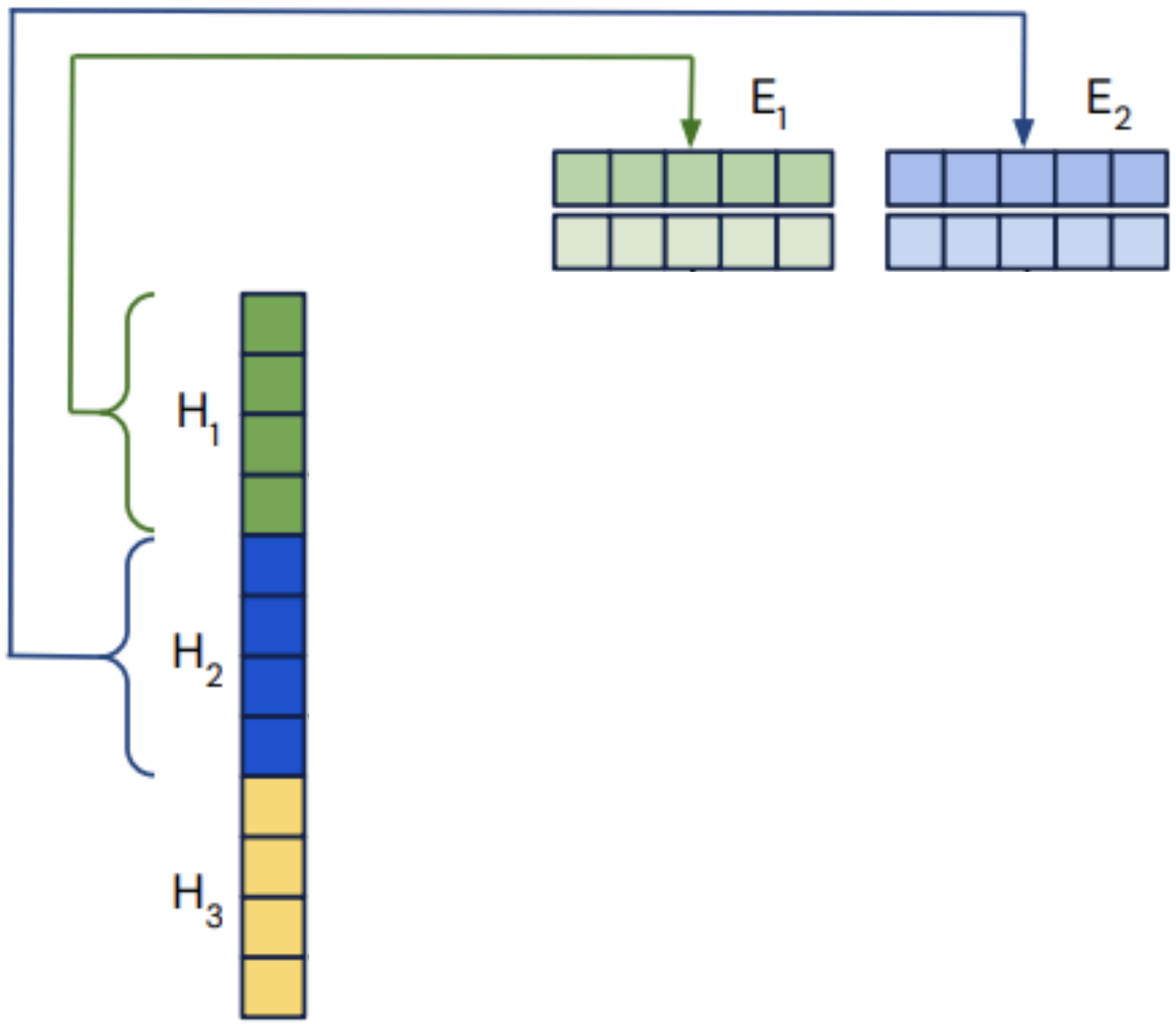
RETRO Transformers



Chunked Cross Attention (CCA)

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

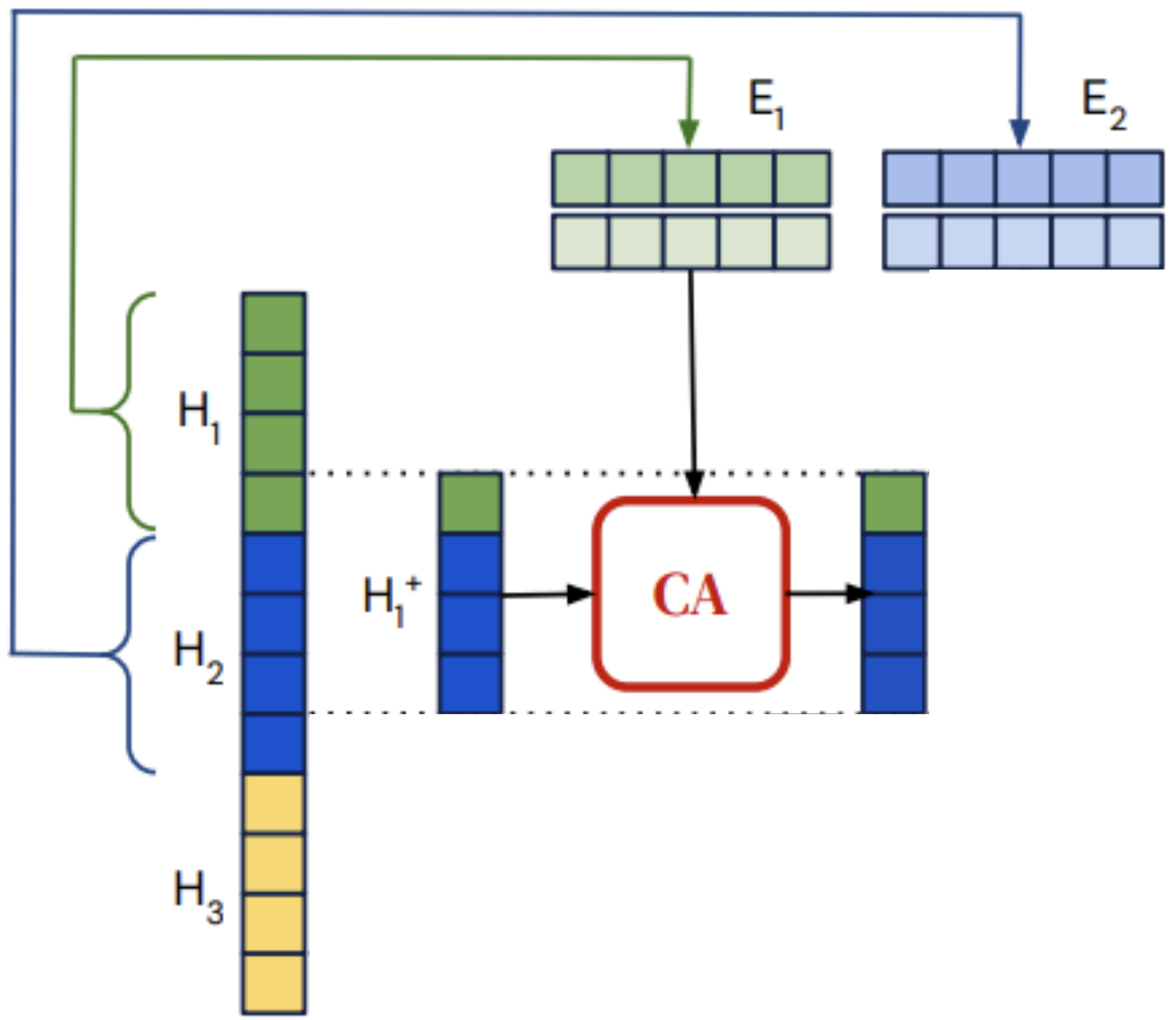
Chunked Cross Attention



Outputs from the previous layer H

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

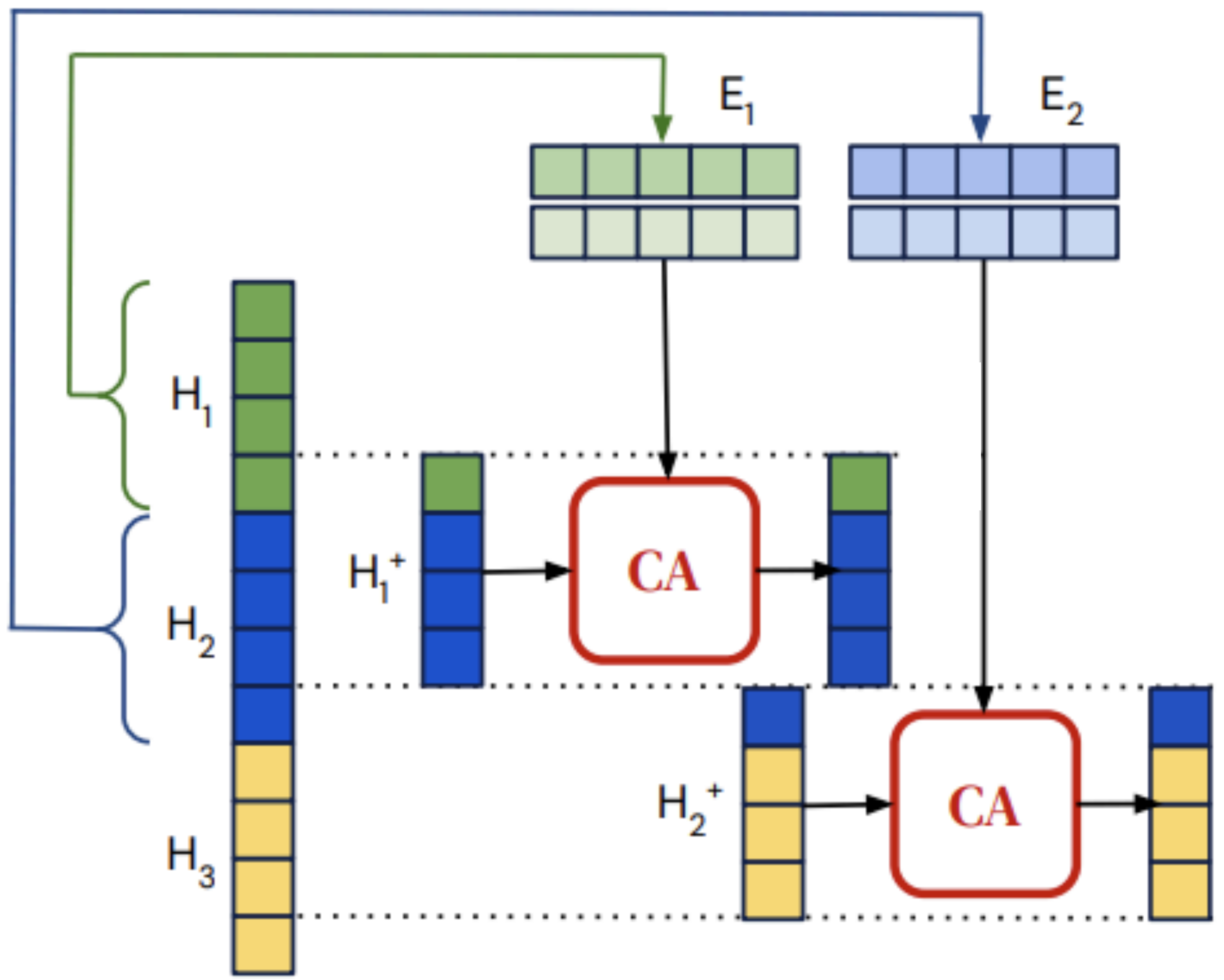
Chunked Cross Attention



Outputs from the previous layer H

Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

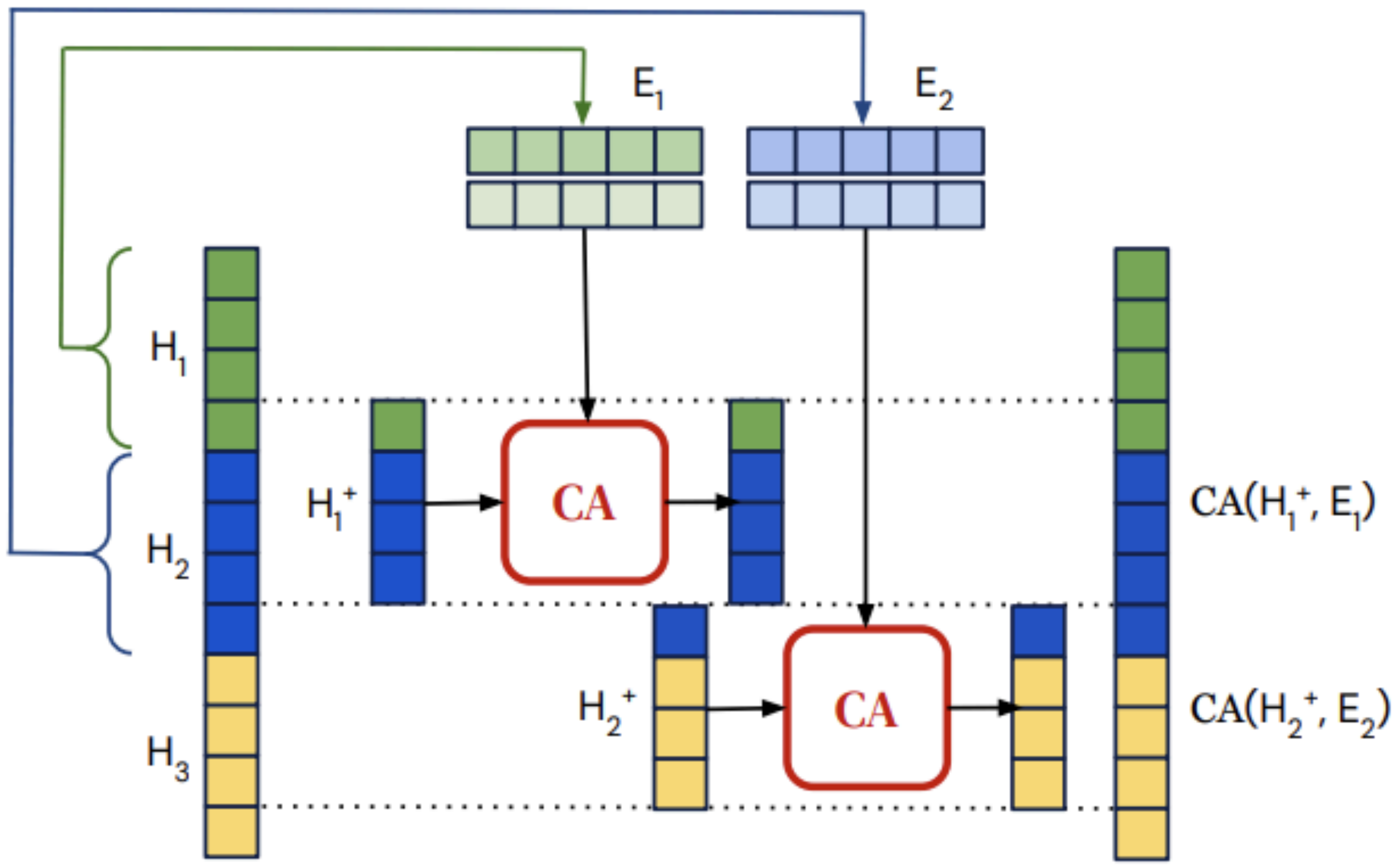
Chunked Cross Attention



Outputs from the previous layer H

Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

Chunked Cross Attention

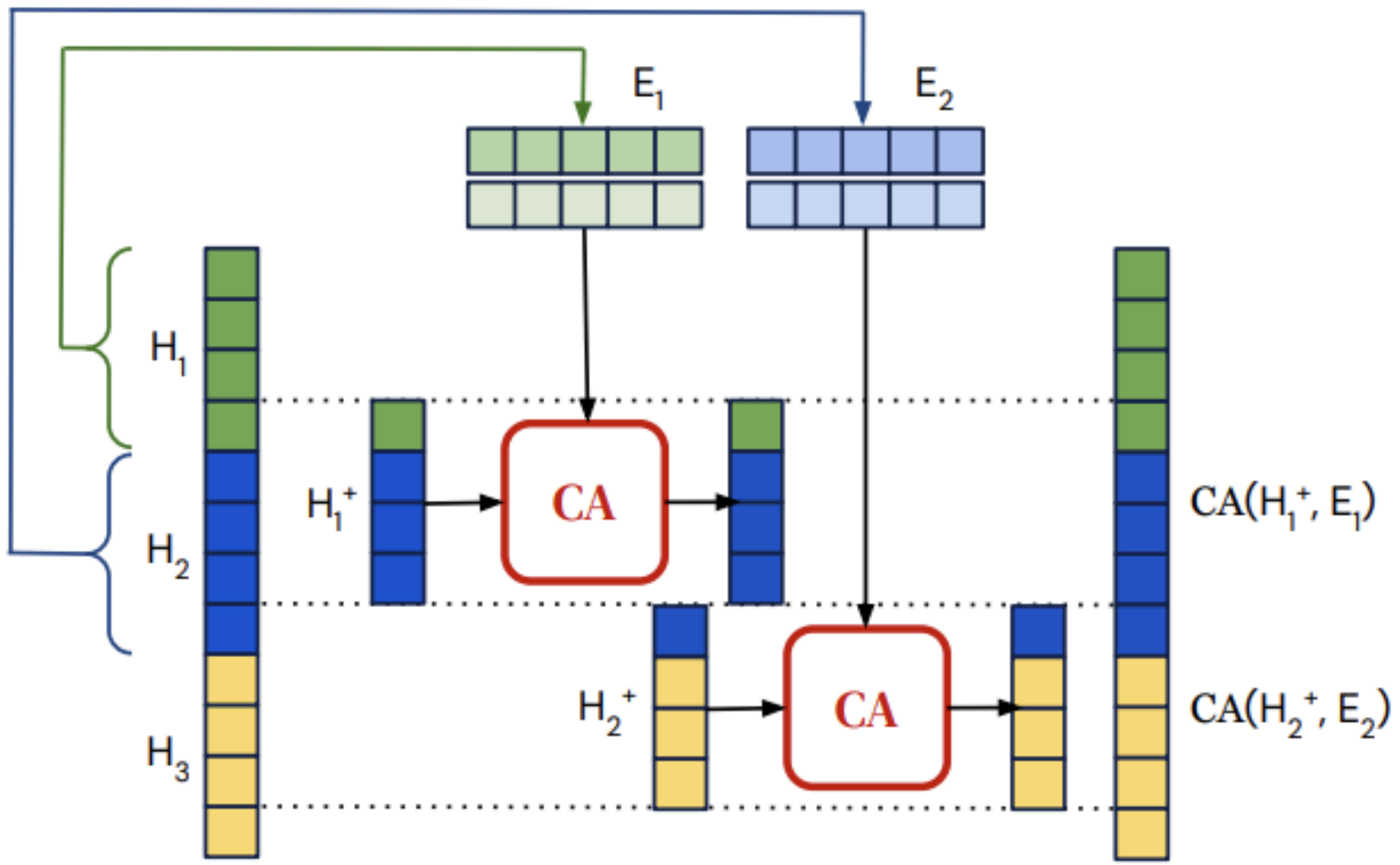


Outputs from the previous layer H

Inputs to the next layer

Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

Chunked Cross Attention



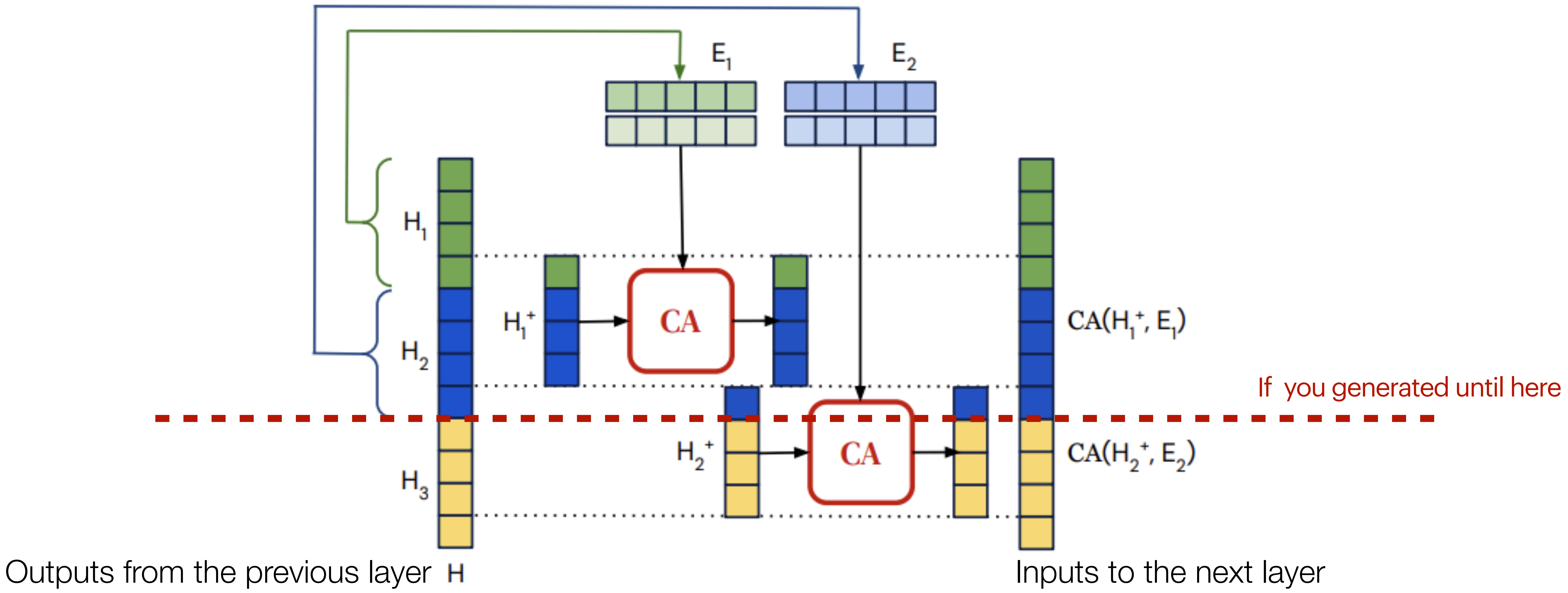
Outputs from the previous layer H

Inputs to the next layer

✓Cross-attention can be computed ***in parallel*, and be *re-used***

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

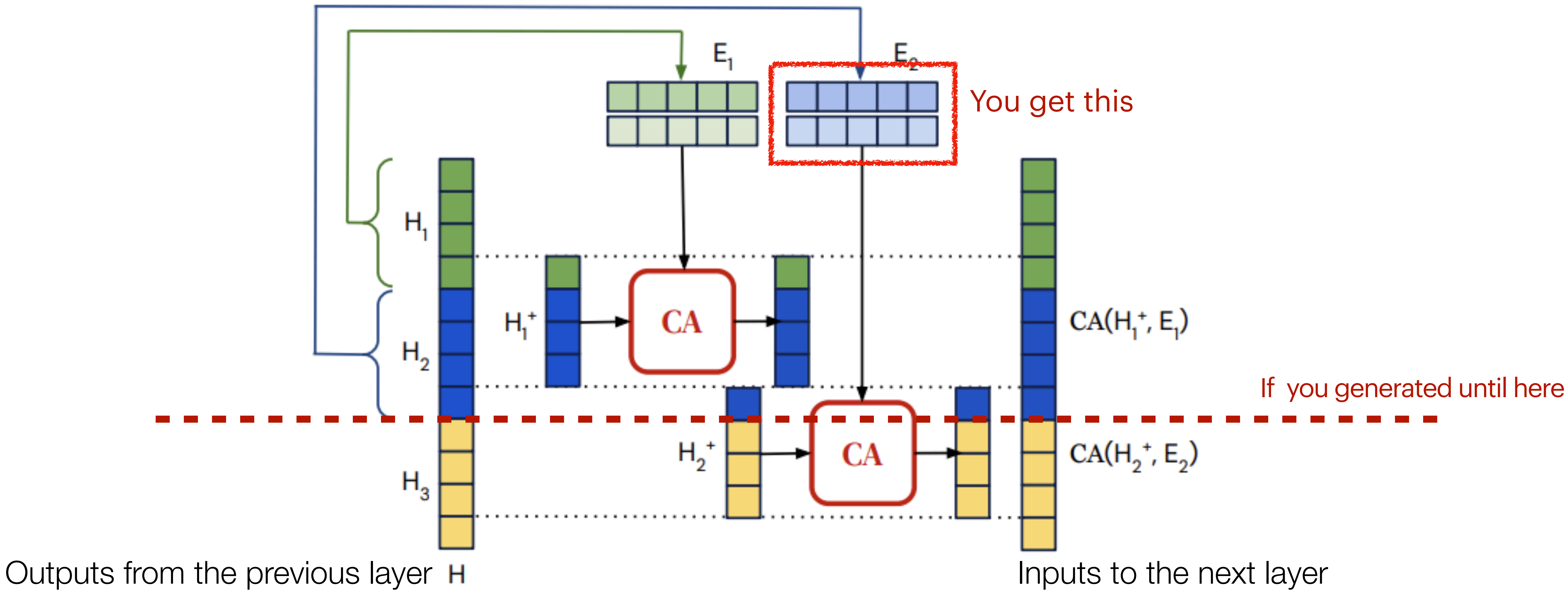
Chunked Cross Attention



✓Cross-attention can be computed ***in parallel*, and be *re-used***

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

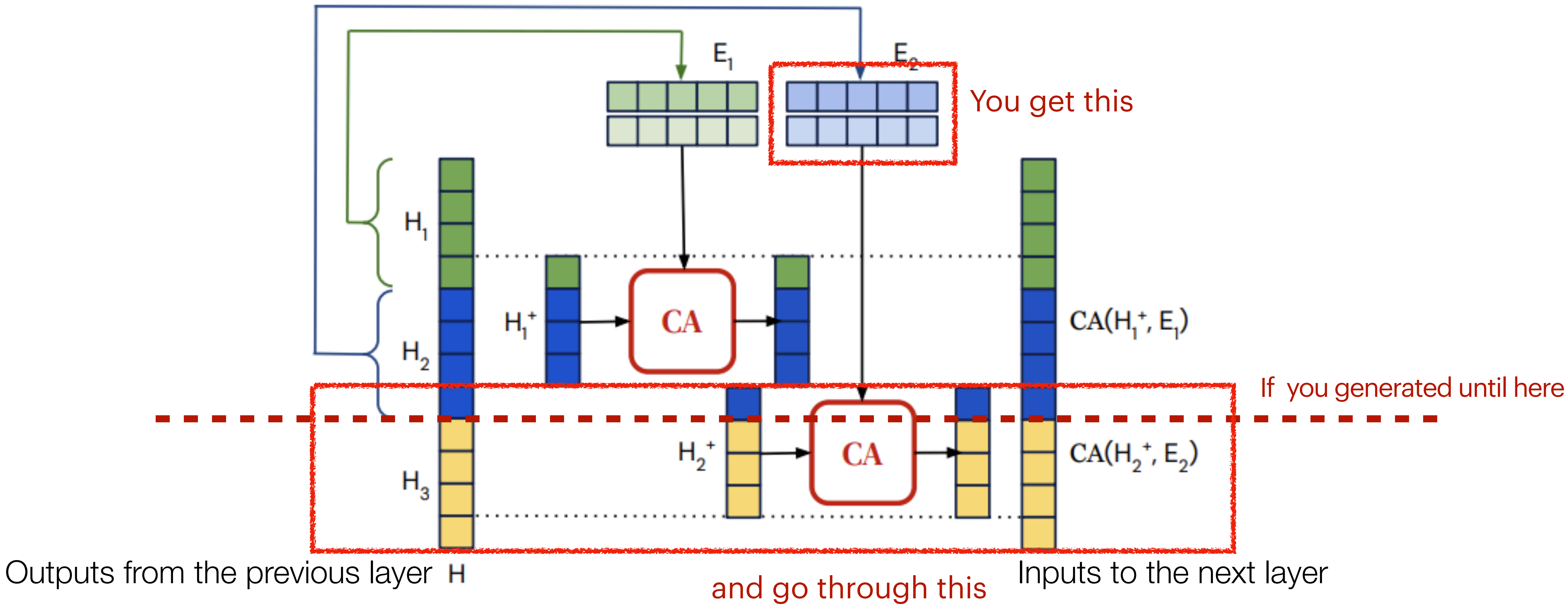
Chunked Cross Attention



✓ Cross-attention can be computed ***in parallel*, and be *re-used***

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

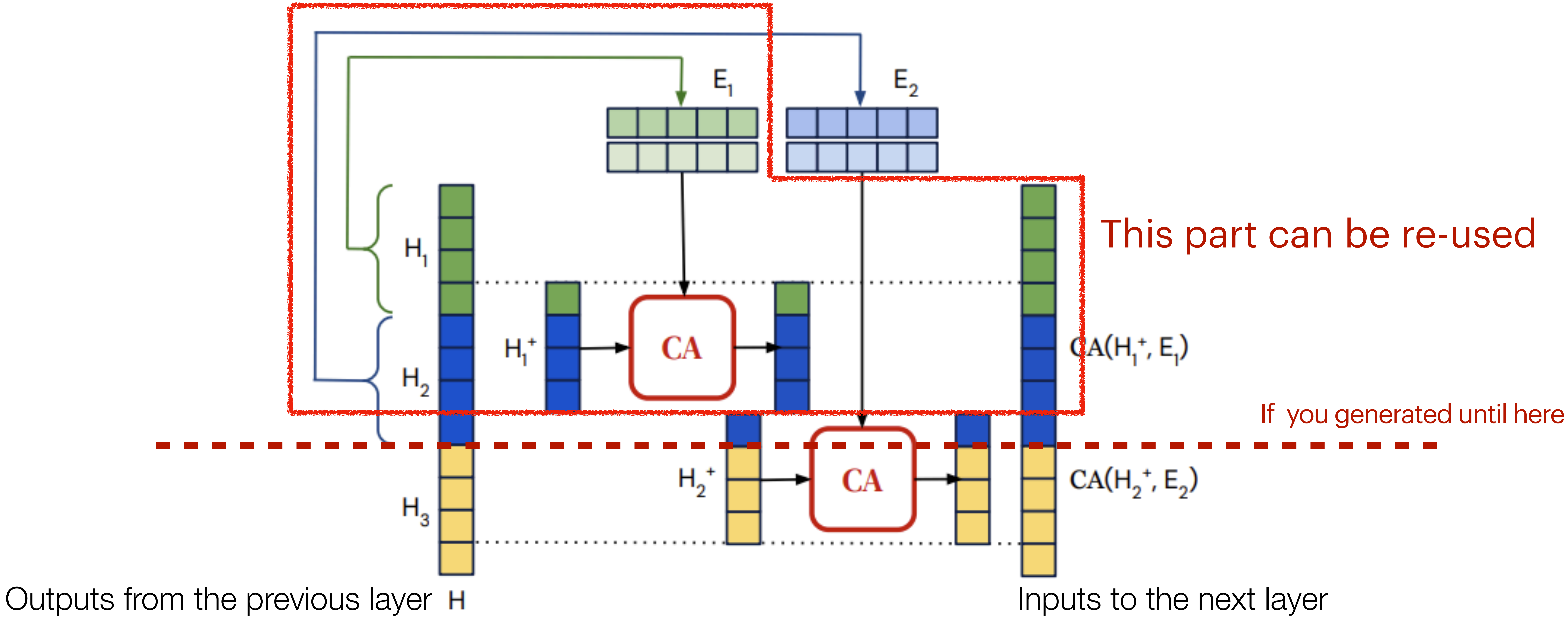
Chunked Cross Attention



✓ Cross-attention can be computed ***in parallel*, and be *re-used***

Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

Chunked Cross Attention



✓Cross-attention can be computed ***in parallel*, and be *re-used***

 Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

Results

Perplexity: The lower the better

Model	Retrieval Set	#Database tokens	#Database keys	Valid	Test
Adaptive Inputs (Baevski and Auli, 2019)	-	-	-	17.96	18.65
SPALM (Yogatama et al., 2021)	Wikipedia	3B	3B	17.20	17.60
*k*NN-LM (Khandelwal et al., 2020)	Wikipedia	3B	3B	16.06	16.12
Megatron (Shoeybi et al., 2019)	-	-	-	-	10.81
Baseline transformer (ours)	-	-	-	21.53	22.96
*k*NN-LM (ours)	Wikipedia	4B	4B	18.52	19.54
RETRO	Wikipedia	4B	0.06B	18.46	18.97
RETRO	C4	174B	2.9B	12.87	10.23
RETRO	MassiveText (1%)	18B	0.8B	18.92	20.33
RETRO	MassiveText (10%)	179B	4B	13.54	14.95
RETRO	MassiveText (100%)	1792B	28B	**3.21**	**3.92**

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

Results

Perplexity: The lower the better

Model	Retrieval Set	#Database tokens	#Database keys	Valid	Test
Adaptive Inputs (Baevski and Auli, 2019)	-	-	-	17.96	18.65
SPALM (Yogatama et al., 2021)	Wikipedia	3B	3B	17.20	17.60
*k*NN-LM (Khandelwal et al., 2020)	Wikipedia	3B	3B	16.06	16.12
Megatron (Shoeybi et al., 2019)	-	-	-	-	10.81
Baseline transformer (ours)	-	-	-	21.53	22.96
*k*NN-LM (ours)	Wikipedia	4B	4B	18.52	19.54
RETRO	Wikipedia	4B	0.06B	18.46	18.97
RETRO	C4	174B	2.9B	12.87	10.23
RETRO	MassiveText (1%)	18B	0.8B	18.92	20.33
RETRO	MassiveText (10%)	179B	4B	13.54	14.95
RETRO	MassiveText (100%)	1792B	28B	**3.21**	**3.92**

Borgeaud et al. 2021. “Improving language models by retrieving from trillions of tokens”

Results

Perplexity: The lower the better

Model	Retrieval Set	#Database tokens	#Database keys	Valid	Test
Adaptive Inputs (Baevski and Auli, 2019)	-	-	-	17.96	18.65
SPALM (Yogatama et al., 2021)	Wikipedia	3B	3B	17.20	17.60
*k*NN-LM (Khandelwal et al., 2020)	Wikipedia	3B	3B	16.06	16.12
Megatron (Shoeybi et al., 2019)	-	-	-	-	10.81
Baseline transformer (ours)	-	-	-	21.53	22.96
*k*NN-LM (ours)	Wikipedia	4B	4B	18.52	19.54
RETRO	Wikipedia	4B	0.06B	18.46	18.97
RETRO	C4	174B	2.9B	12.87	10.23
RETRO	MassiveText (1%)	18B	0.8B	18.92	20.33
RETRO	MassiveText (10%)	179B	4B	13.54	14.95
RETRO	MassiveText (100%)	1792B	28B	**3.21**	**3.92**

Significant improvements by retrieving from 1.8 trillion tokens
(We'll talk more about the importance of the **datastore size** later)

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

Results

Perplexity: The lower the better

Model	Retrieval Set	#Database tokens	#Database keys	Valid	Test
Adaptive Inputs (Baevski and Auli, 2019)	-	-	-	17.96	18.65
SPALM (Yogatama et al., 2021)	Wikipedia	3B	3B	17.20	17.60
*k*NN-LM (Khandelwal et al., 2020)	Wikipedia	3B	3B	16.06	16.12
Megatron (Shoeybi et al., 2019)	-	-	-	-	10.81
Baseline transformer (ours)	-	-	-	21.53	22.96
*k*NN-LM (ours)	Wikipedia	4B	4B	18.52	19.54
RETRO	Wikipedia	4B	0.06B	18.46	18.97
RETRO	C4	174B	2.9B	12.87	10.23
RETRO	MassiveText (1%)	18B	0.8B	18.92	20.33
RETRO	MassiveText (10%)	179B	4B	13.54	14.95
RETRO	MassiveText (100%)	1792B	28B	**3.21**	**3.92**

Significant improvements by retrieving from 1.8 trillion tokens
(We'll talk more about the importance of the **datastore size** later)

Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

New Retrieval-based LMs: Overview

- New Methodology 1 — Designing a new Transformer
    - **New attention layers to incorporate more blocks (RETRO)**
    - Possibly combine with long-range Transformers
- New Methodology 2 — Designing a new Softmax
- New LM Design — Mitigating fairness & legality issues

New Retrieval-based LMs: Overview

- New Methodology 1 — Designing a new Transformer
  - New attention layers to incorporate more blocks (RETRO)
  - **Possibly combine with long-range Transformers**
- New Methodology 2 — Designing a new Softmax
- New LM Design — Mitigating fairness & legality issues

Solve length limit issue in retrieval augmentation

(and probably simpler than RETRO?!)

New Retrieval-based LMs: Overview

- New Methodology 1 — Designing a new Transformer
  - New attention layers to incorporate more blocks (RETRO)
  - Possibly combine with long-range Transformers
- **New Methodology 2 — Designing a new Softmax**
- New LM Design — Mitigating fairness & legality issues

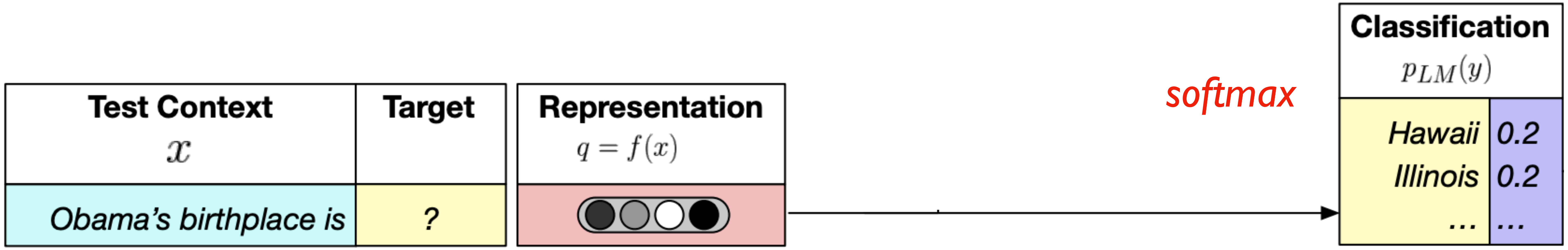
Nonparametric softmax?

kNN-LM

Test Context x	Target
Obama's birthplace is	?

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

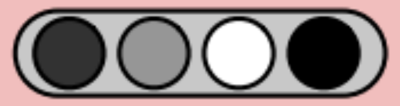
kNN-LM



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

kNN-LM

... Obama was senator for Illinois from 1997 to 2005, ... Barack is Married to Michelle and their first daughter, ... Obama was born in Hawaii, and graduated from Columbia University. ... Obama is a native of Hawaii, ...

Test Context x	**Target**	**Representation** $q = f(x)$
Obama's birthplace is	?	

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

kNN-LM

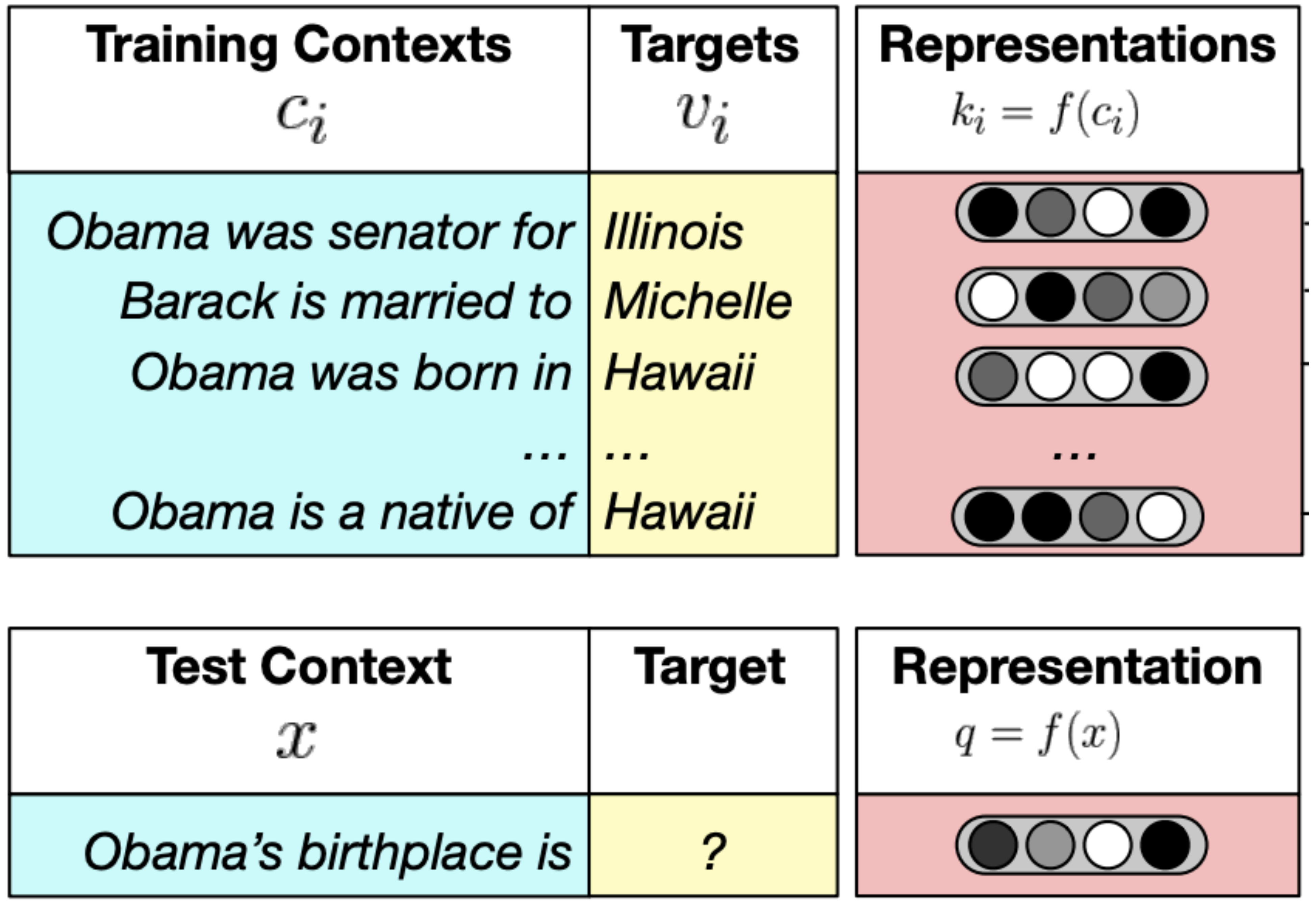
Training Contexts c_i	**Targets** v_i
Obama was senator for	*Illinois*
Barack is married to	*Michelle*
Obama was born in	*Hawaii*
…	…
Obama is a native of	*Hawaii*

… Obama was senator for Illinois from 1997 to 2005, …. Barack is Married to Michelle and their first daughter, … Obama was born in Hawaii, and graduated from Columbia University. … Obama is a native of Hawaii, ….

Test Context x	**Target**	**Representation** $q = f(x)$
Obama's birthplace is	?	

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

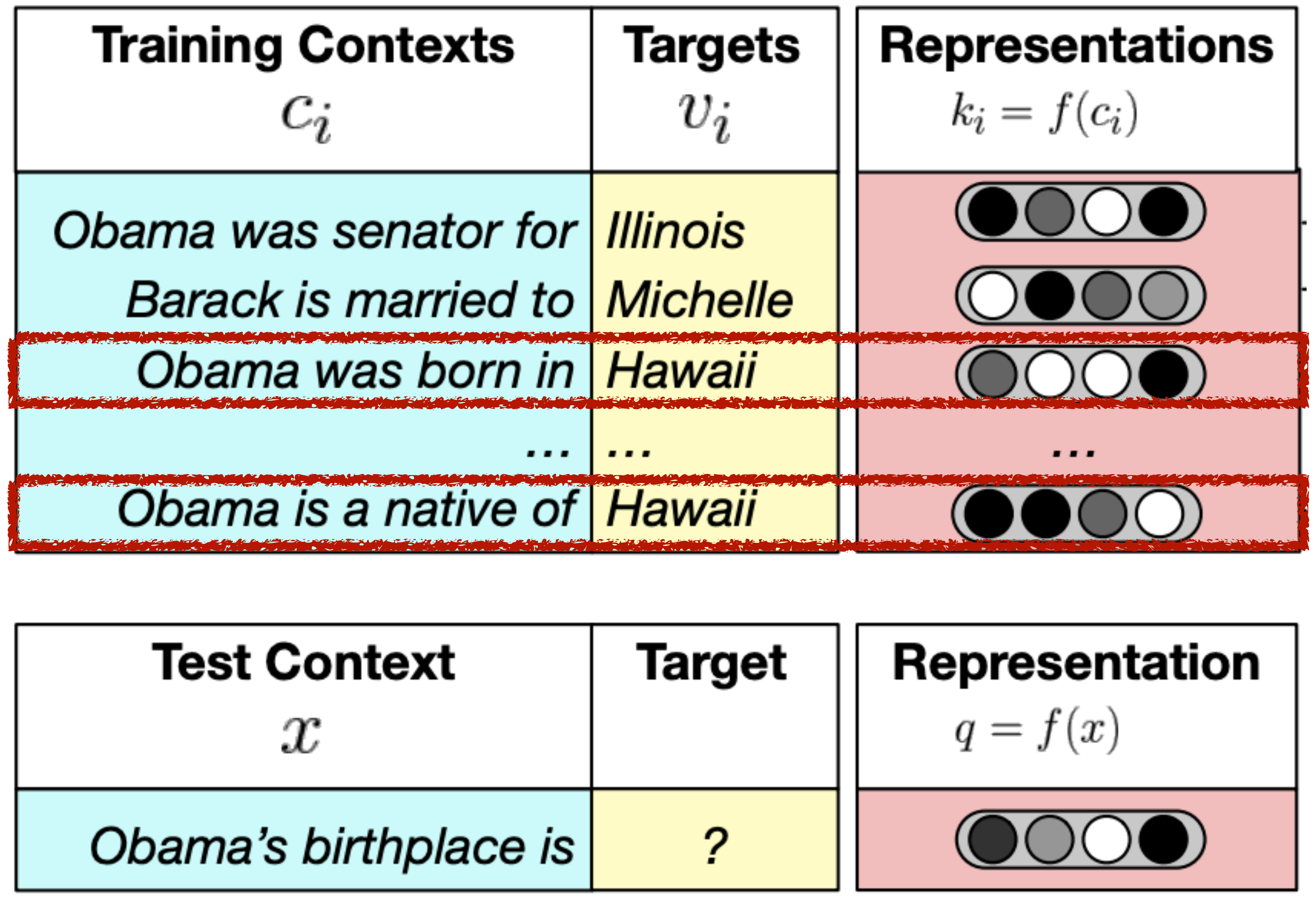
kNN-LM



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

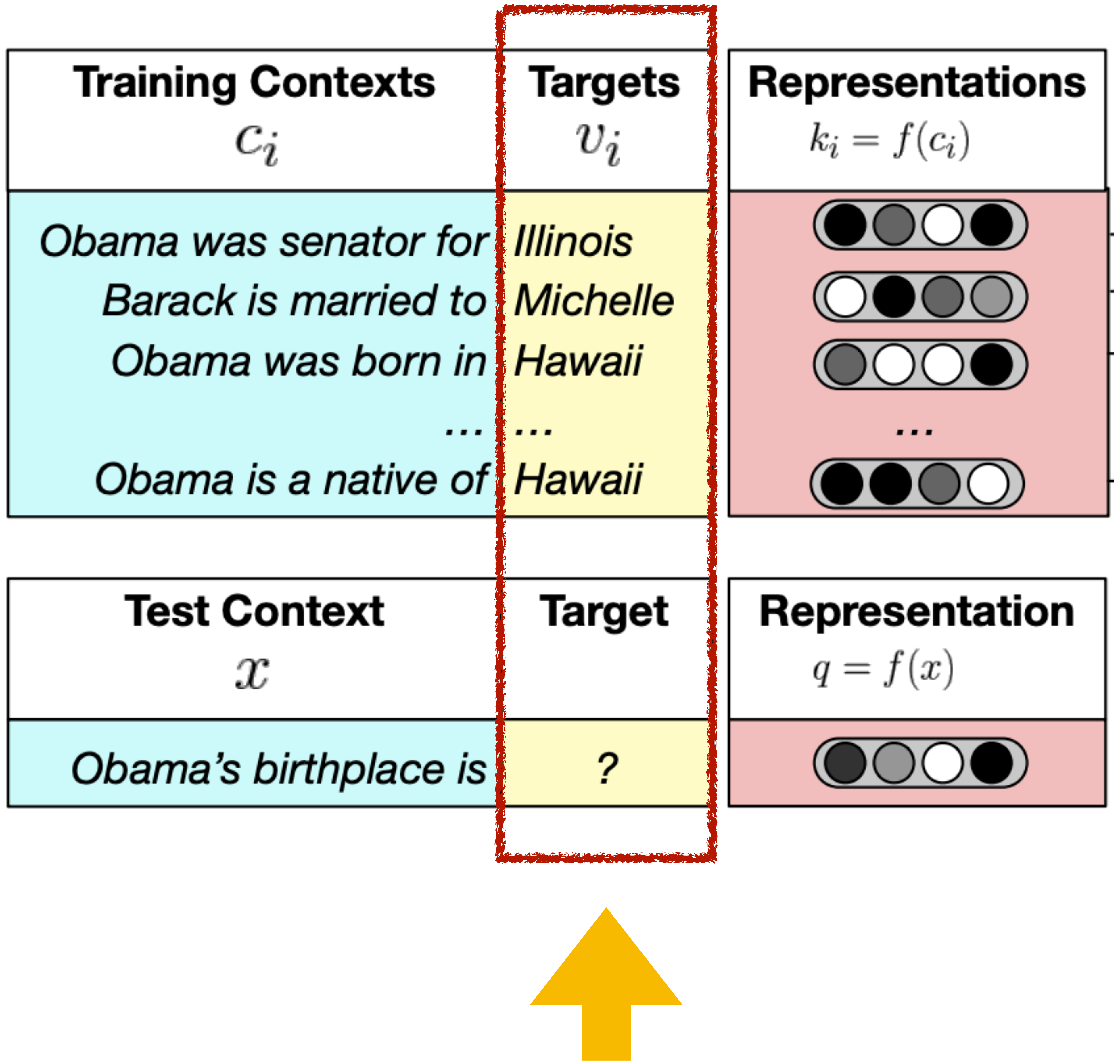
kNN-LM

# of vectors = # of tokens in the corpus (>1B)



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

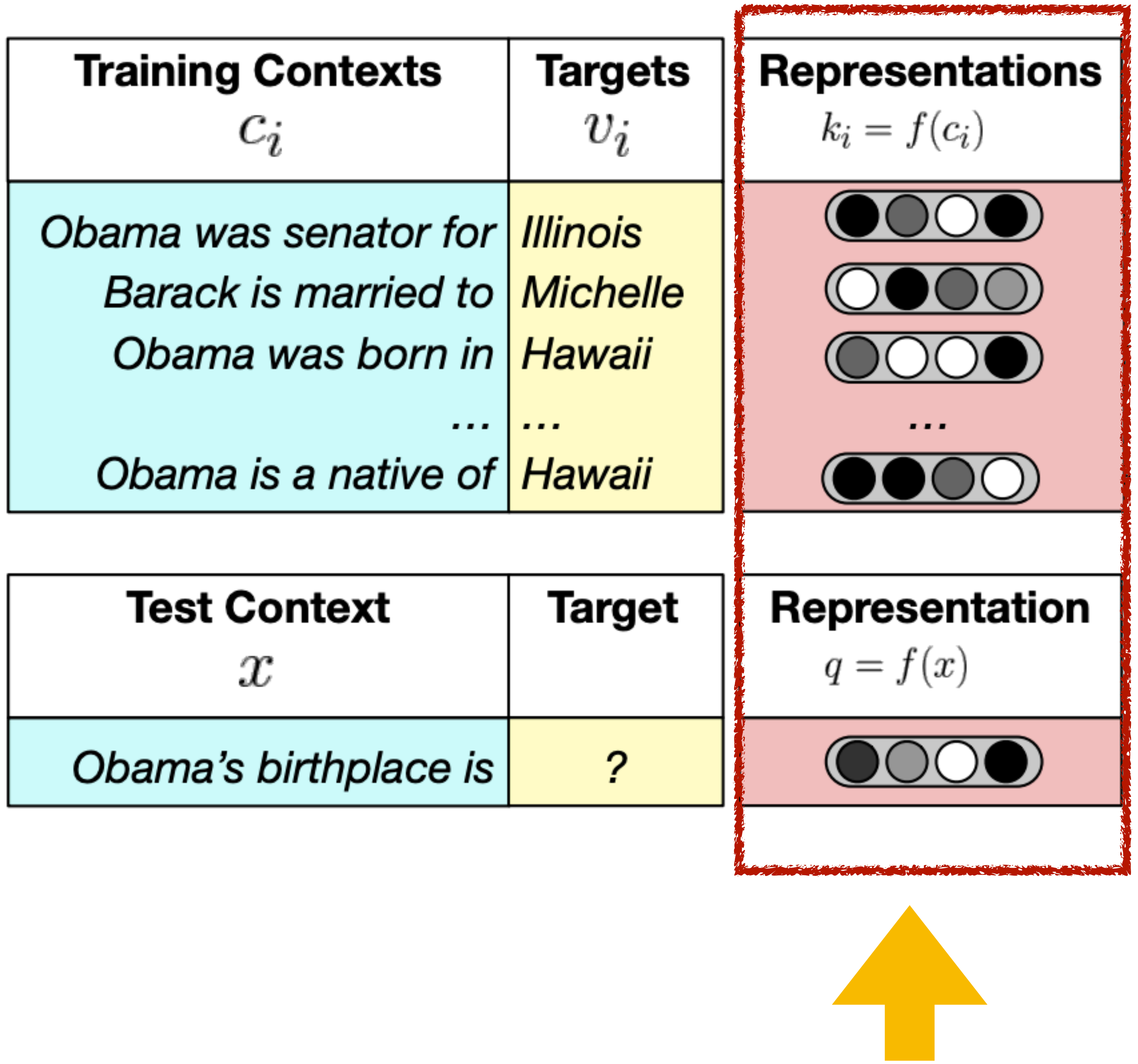
kNN-LM



Which tokens in a datastore are close to the next token?

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

kNN-LM



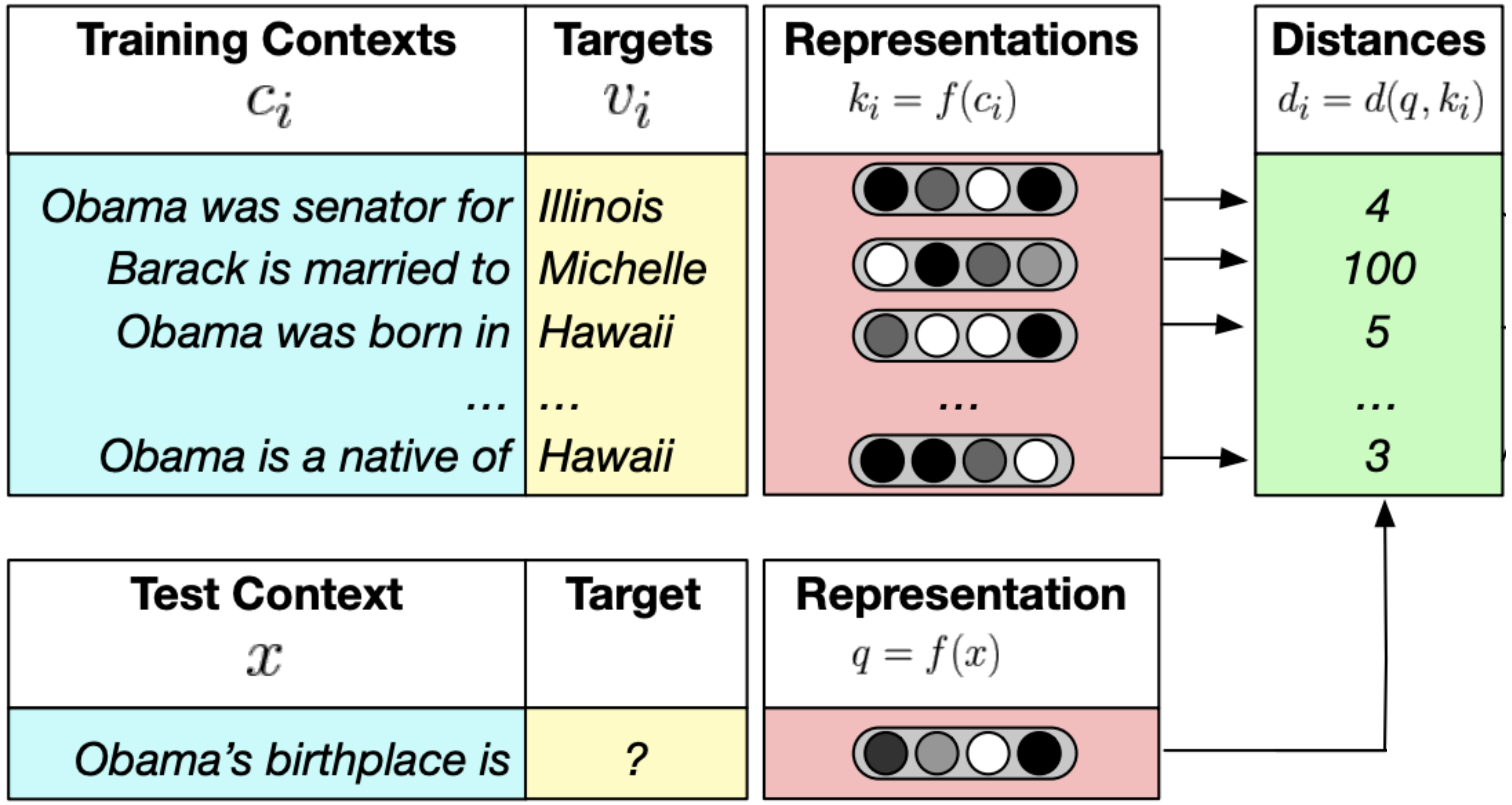
Which tokens <u>in a datastore</u> are close to the next token?

=

Which vectors in a datastore are close to the vector we have?

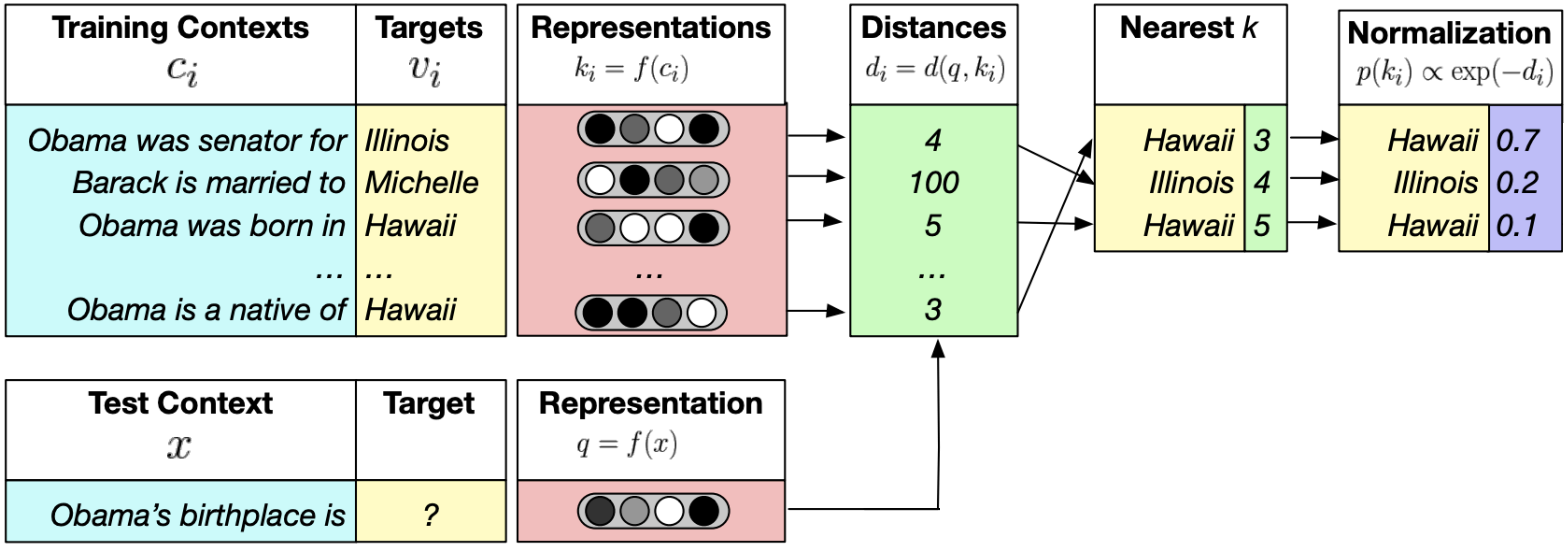
Khandelwal et al. 2020. "Generalization through Memorization: Nearest Neighbor Language Models"

kNN-LM



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

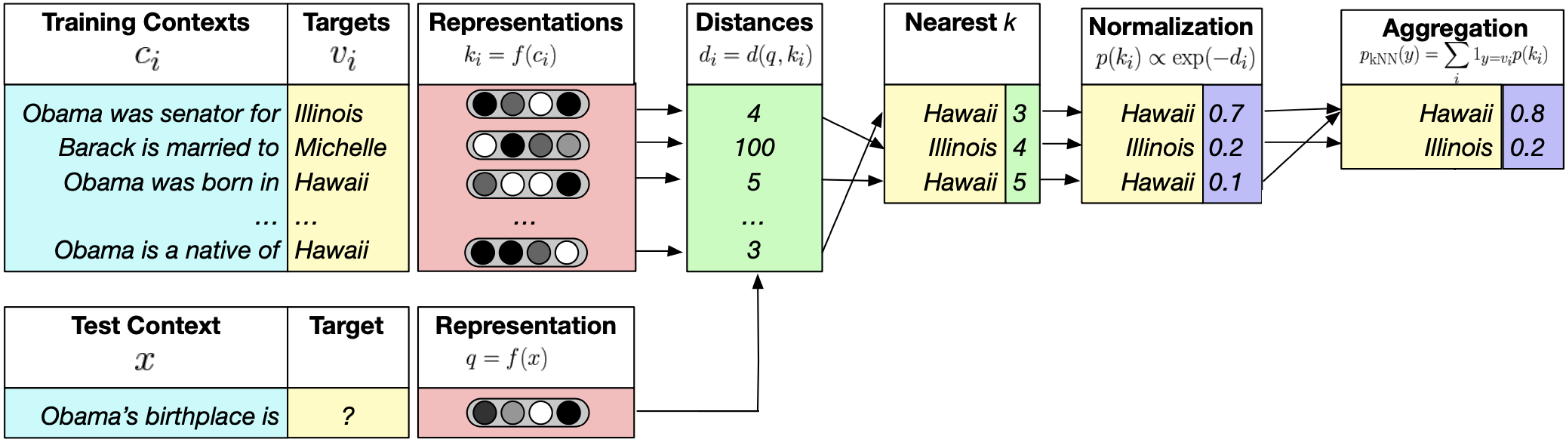
kNN-LM



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

kNN-LM

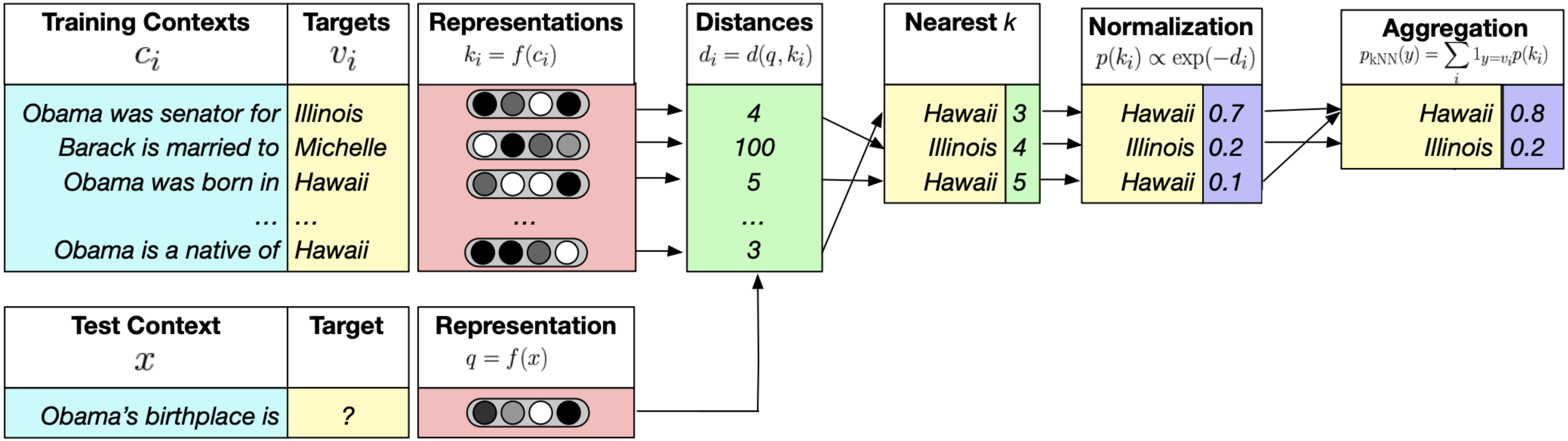
Nonparamatric softmax



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

kNN-LM

Nonparamatric softmax

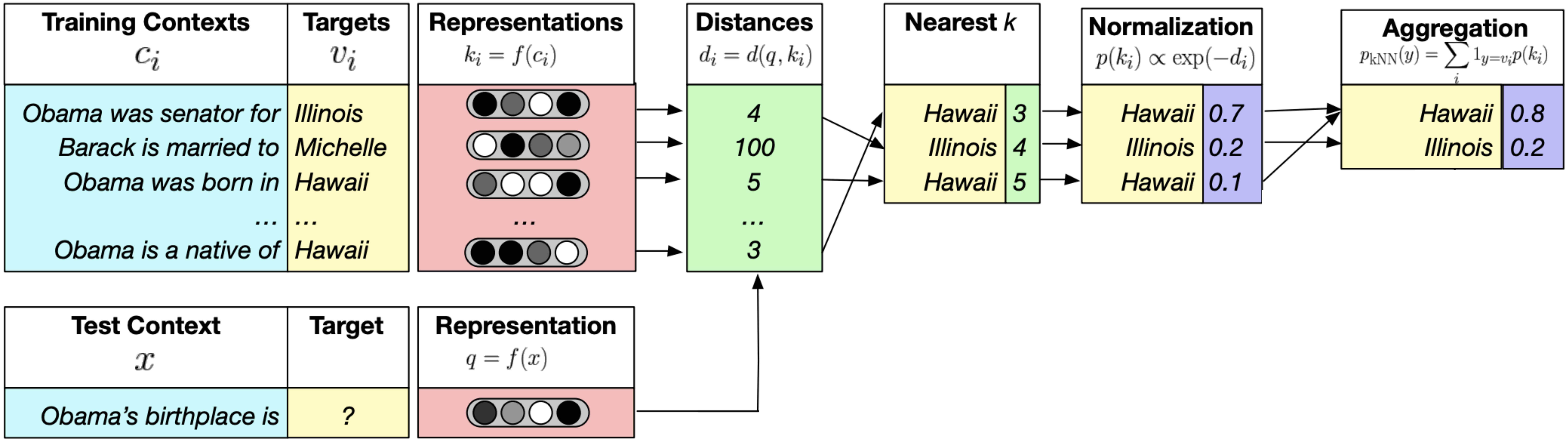


$$P_{k\text{NN}}(y \mid x) \propto \sum_{(k,v)\in\mathcal{D}} \mathbb{1}[v = y]e^{\text{sim}(k,x)}$$

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

kNN-LM

Nonparametric softmax

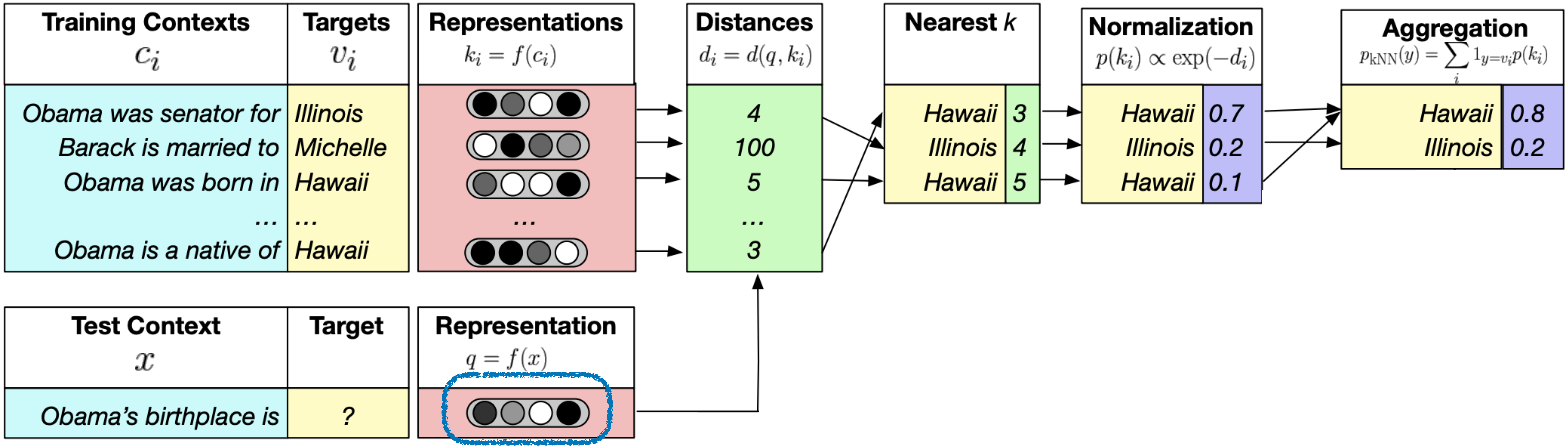


$$P_{kNN}(y\ \ x) \propto \sum_{(k,v)\in\mathcal{D}} \mathbb{1}[v = y]e^{\mathrm{sim}(k,x)} \qquad \mathrm{sim}(k, x) = -d(\mathrm{Enc}(k), \mathrm{Enc}(x))$$

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

kNN-LM

Nonparamatric softmax

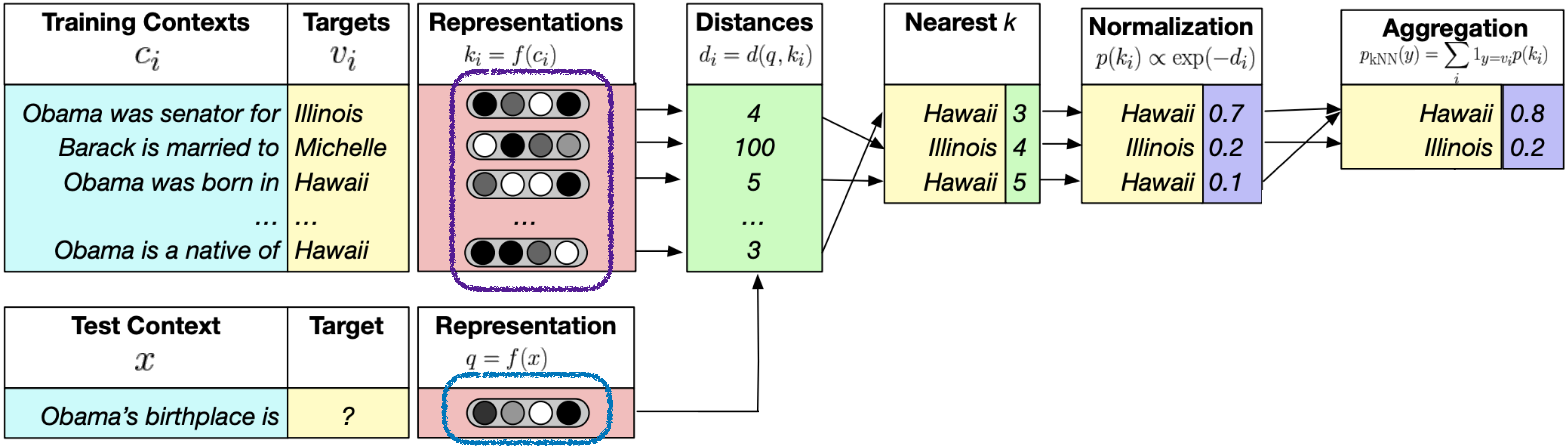


$$P_{k\mathrm{NN}}(y\ \ x) \propto \sum_{(k,v)\in\mathcal{D}} \mathbb{I}[v = y]e^{\mathrm{sim}(k,x)} \qquad \mathrm{sim}(k, x) = -\,d(\mathrm{Enc}(k), \mathrm{Enc}(x))$$

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

kNN-LM

Nonparamatric softmax

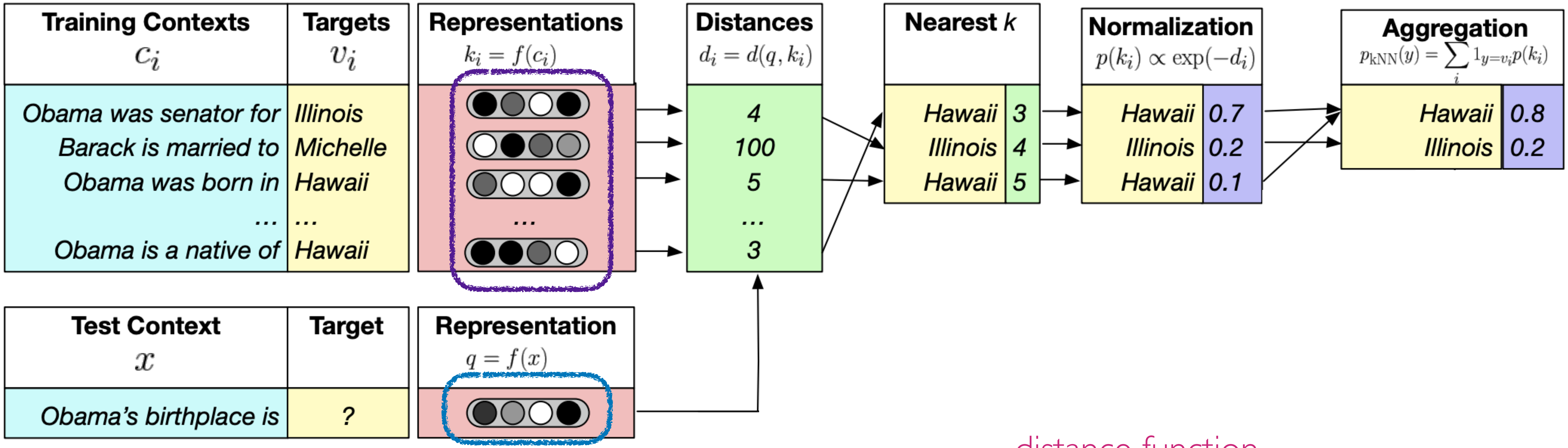


$$P_{kNN}(y \quad x) \propto \sum_{(k,v)\in\mathcal{D}} \mathbb{1}[v = y]e^{\mathrm{sim}(k,x)} \qquad \mathrm{sim}(k, x) = -\, d(\underline{\mathrm{Enc}(k)}, \underline{\mathrm{Enc}(x)})$$

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

kNN-LM

Nonparamatric softmax



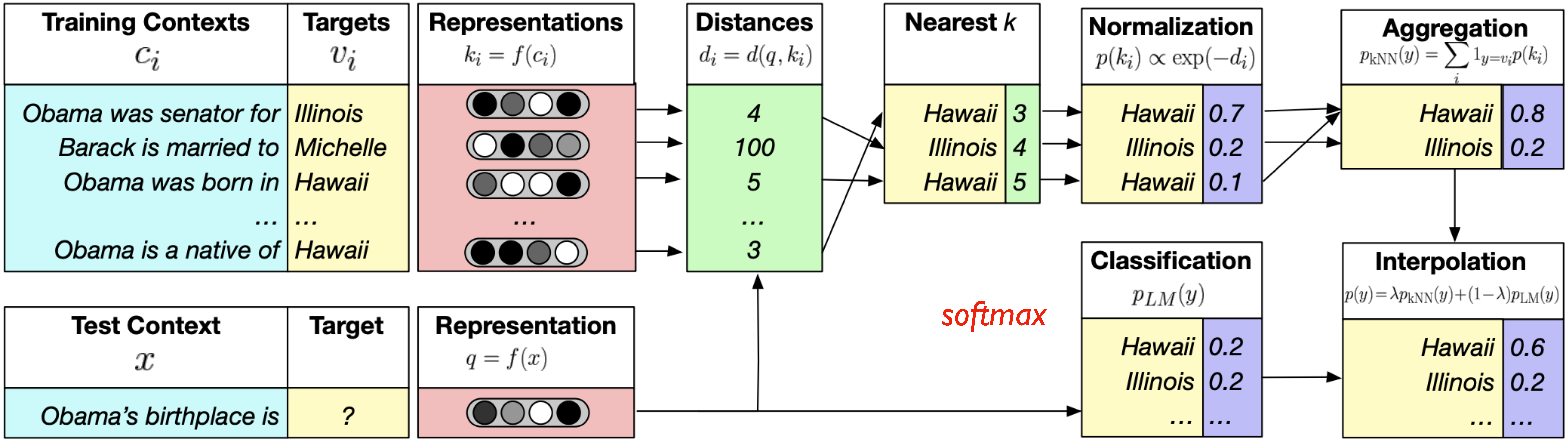
distance function

$$P_{k\text{NN}}(y \mid x) \propto \sum_{(k,v)\in\mathcal{D}} \mathbb{1}[v = y]e^{\text{sim}(k,x)} \qquad \text{sim}(k, x) = -d(\text{Enc}(k), \text{Enc}(x))$$

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

kNN-LM

Nonparamatric softmax

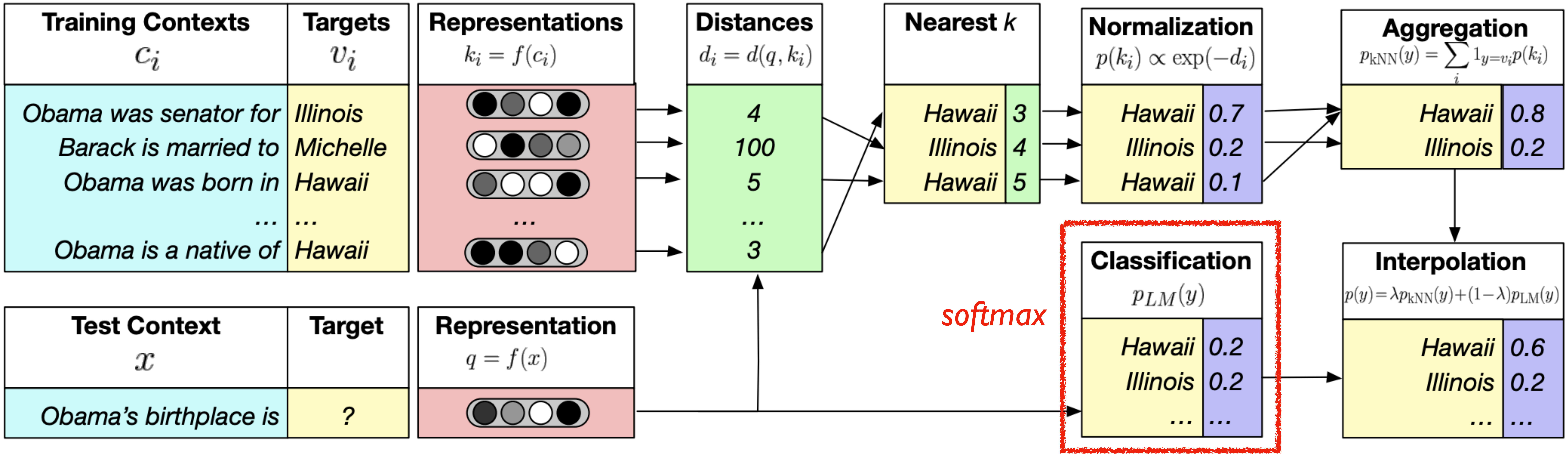


$$P_{k\mathrm{NN-LM}}(y \quad x) = (1-\lambda)P_{\mathrm{LM}}(y \quad x) + \lambda P_{k\mathrm{NN}}(y \quad x)$$

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

kNN-LM

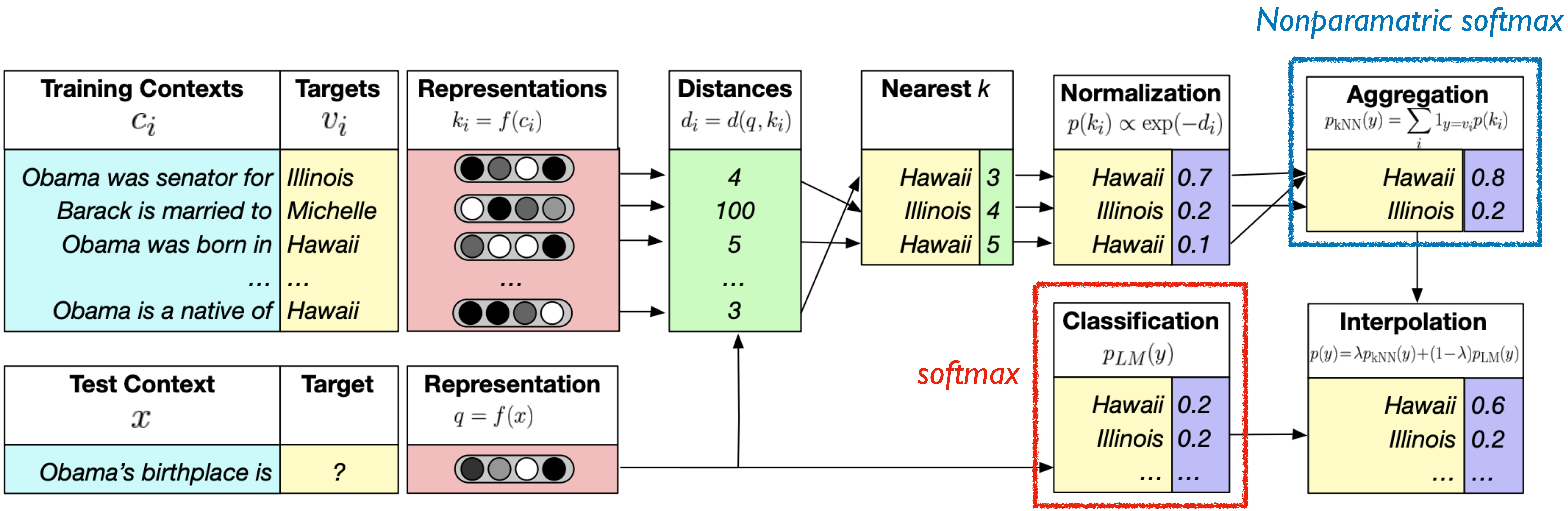
Nonparametric softmax



$$P_{k\text{NN}-\text{LM}}(y \mid x) = (1-\lambda)P_{\text{LM}}(y \mid x) + \lambda P_{k\text{NN}}(y \mid x)$$

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

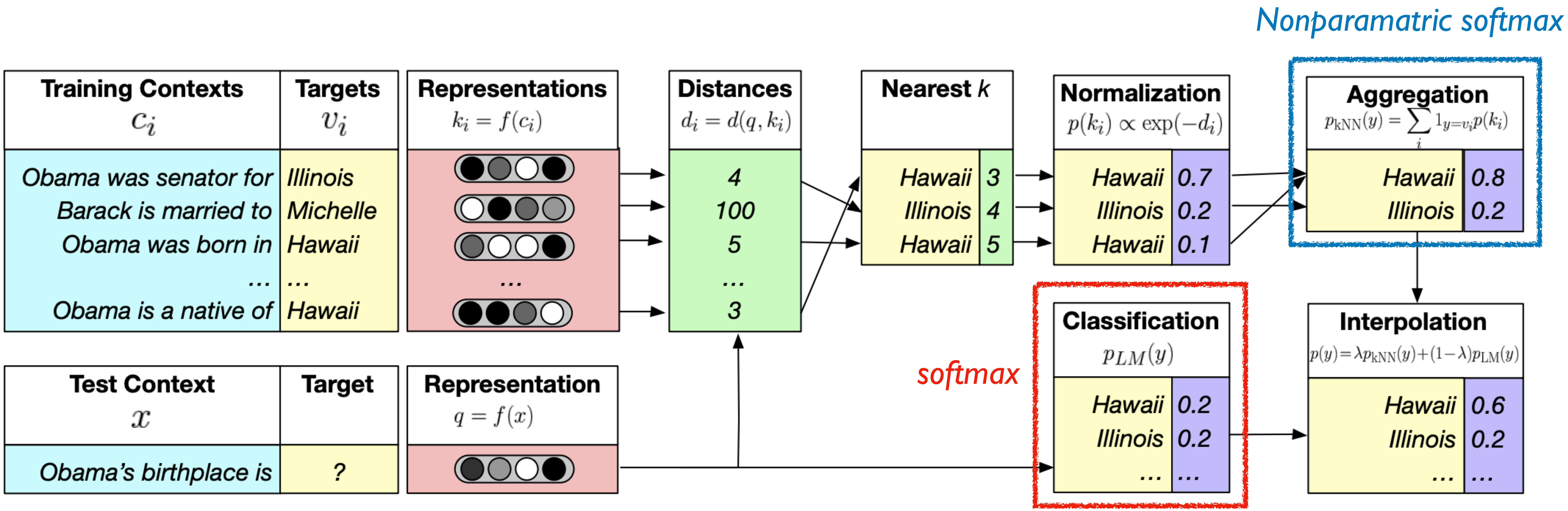
kNN-LM



$$P_{k\text{NN-LM}}(y \quad x) = (1 - \lambda) P_{\text{LM}}(y \quad x) + \lambda P_{k\text{NN}}(y \quad x)$$

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

kNN-LM



$$P_{k\mathrm{NN-LM}}(y \mid x) = (1-\lambda)\underline{P_{\mathrm{LM}}(y \mid x)} + \lambda \underline{P_{k\mathrm{NN}}(y \mid x)}$$

λ: hyperparameter

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

Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	*cheap*
Item delivered broken. Very	*cheap*
To check the version of PyTorch, you can use	*torch*
You are permitted to bring a	*torch*
A group of infections … one of the	*torch*

Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	***cheap***
Item delivered broken. Very	***cheap***
To check the version of PyTorch, you can use	*torch*
You are permitted to bring a	*torch*
A group of infections … one of the	*torch*

Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	***cheap***
Item delivered broken. Very	***cheap***
To check the version of PyTorch, you can use	*torch*
You are permitted to bring a	*torch*
A group of infections … one of the	*torch*

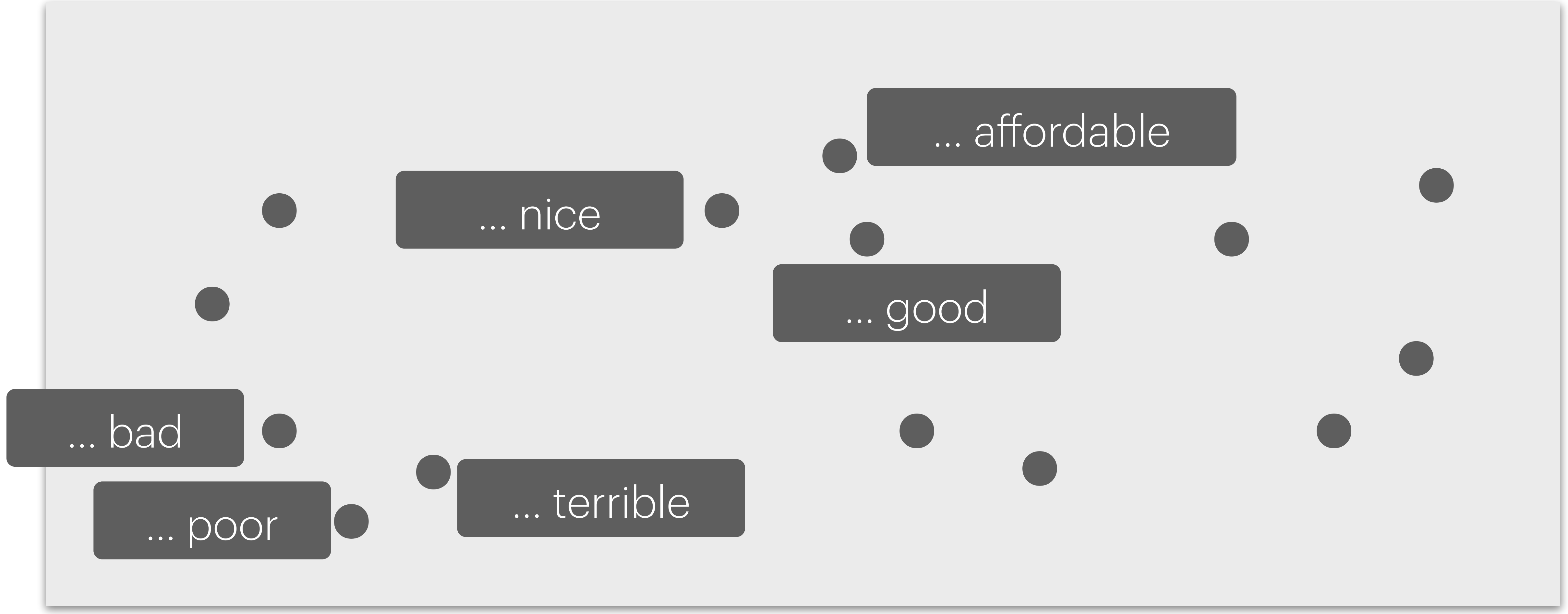
Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	***cheap***
Item delivered broken. Very	***cheap***
To check the version of PyTorch, you can use	*torch*
You are permitted to bring a	*torch*
A group of infections … one of the	*torch*

Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	***cheap***
Item delivered broken. Very	***cheap***
To check the version of PyTorch, you can use	*torch*
You are permitted to bring a	*torch*
A group of infections … one of the	*torch*

Dense vector space



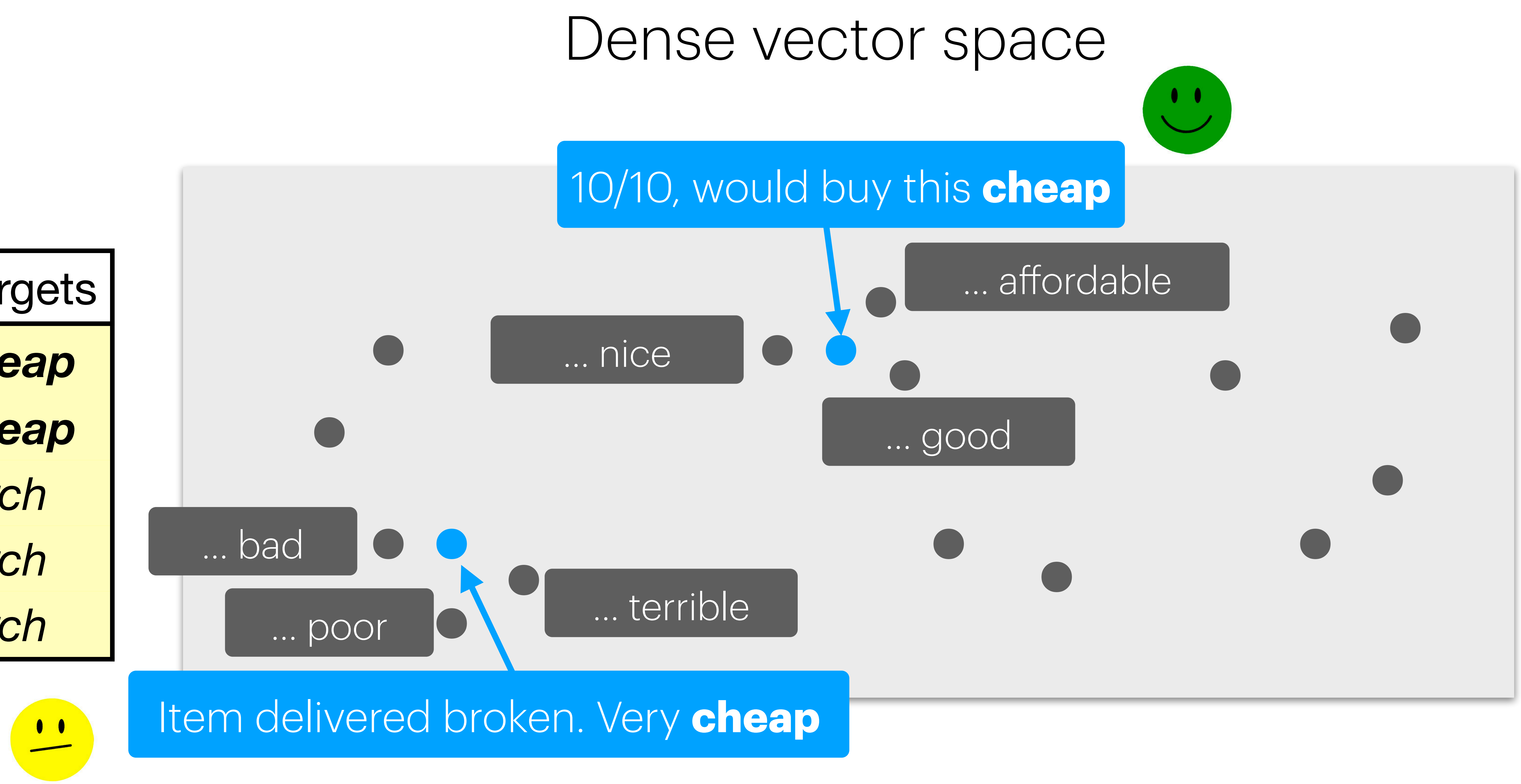
Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	***cheap***
Item delivered broken. Very	***cheap***
To check the version of PyTorch, you can use	*torch*
You are permitted to bring a	*torch*
A group of infections … one of the	*torch*

Dense vector space

Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	***cheap***
Item delivered broken. Very	***cheap***
To check the version of PyTorch, you can use	*torch*
You are permitted to bring a	*torch*
A group of infections … one of the	*torch*



Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	*cheap*
Item delivered broken. Very	*cheap*
To check the version of PyTorch, you can use	***torch***
You are permitted to bring a	***torch***
A group of infections … one of the	***torch***

Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	*cheap*
Item delivered broken. Very	*cheap*
To check the version of PyTorch, you can use	***torch***
You are permitted to bring a	***torch***
A group of infections … one of the	***torch***

Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	*cheap*
Item delivered broken. Very	*cheap*
To check the version of PyTorch, you can use	***torch***
You are permitted to bring a	***torch***
A group of infections … one of the	***torch***

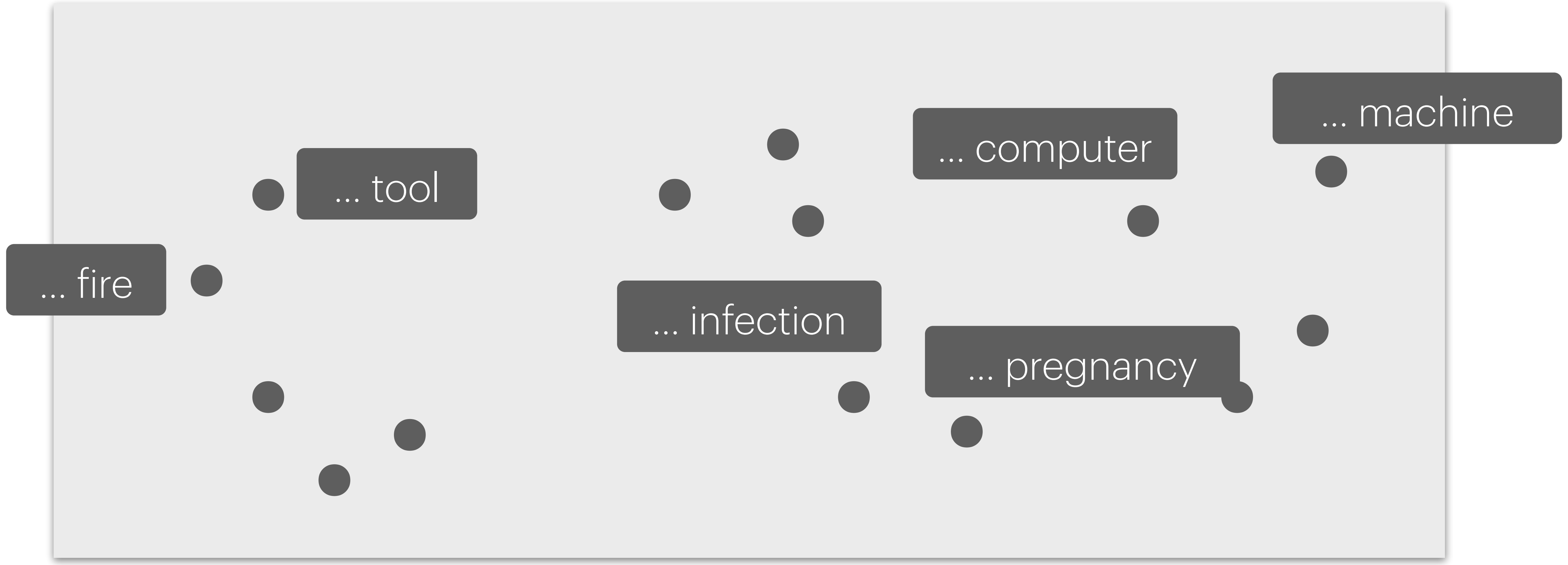
Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	*cheap*
Item delivered broken. Very	*cheap*
To check the version of PyTorch, you can use	***torch***
You are permitted to bring a	***torch***
A group of infections … one of the	***torch***

Why nonparametric softmax?

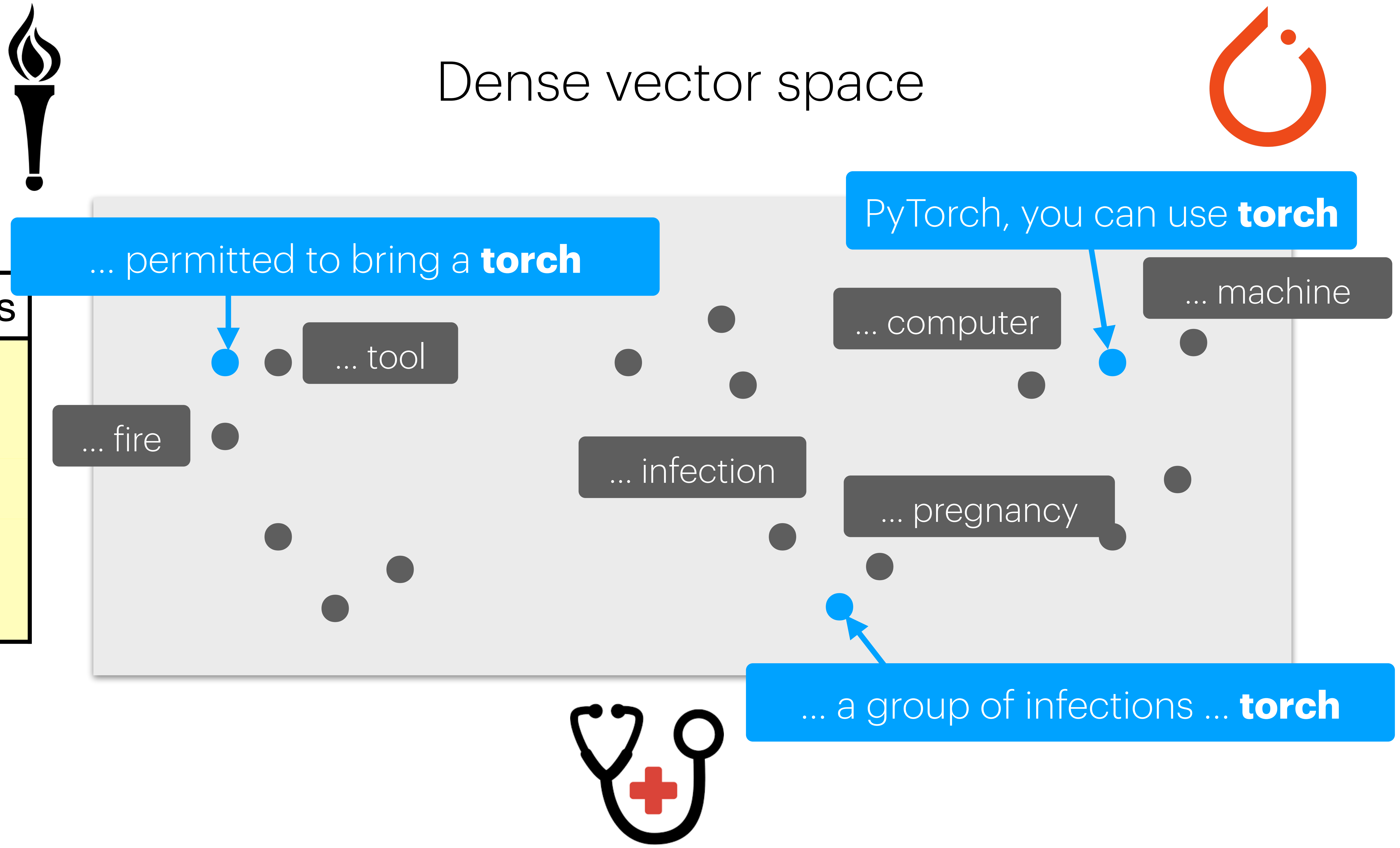
Training contexts	Targets
10/10, would buy this	*cheap*
Item delivered broken. Very	*cheap*
To check the version of PyTorch, you can use	***torch***
You are permitted to bring a	***torch***
A group of infections … one of the	***torch***

Dense vector space



Why nonparametric softmax?

Training contexts	Targets
10/10, would buy this	*cheap*
Item delivered broken. Very	*cheap*
To check the version of PyTorch, you can use	***torch***
You are permitted to bring a	***torch***
A group of infections … one of the	***torch***



Nonparametric-only, Phrase-level (NPM)

(If you can train the model…)

Min et al. 2023. Nonparametric Masked Language Modeling

Nonparametric-only, Phrase-level (NPM)

(If you can train the model…)

Nonparametric-only, Phrase-level (NPM)

(If you can train the model…)

just as a jet of **red light** blasted from Harry's …

Voldemort cried, "Avada Kedavra!" A jet of **green light** issued …

"The Boy Who Lived." He saw the mouth move and a flash of **green light**, and everything was gone.

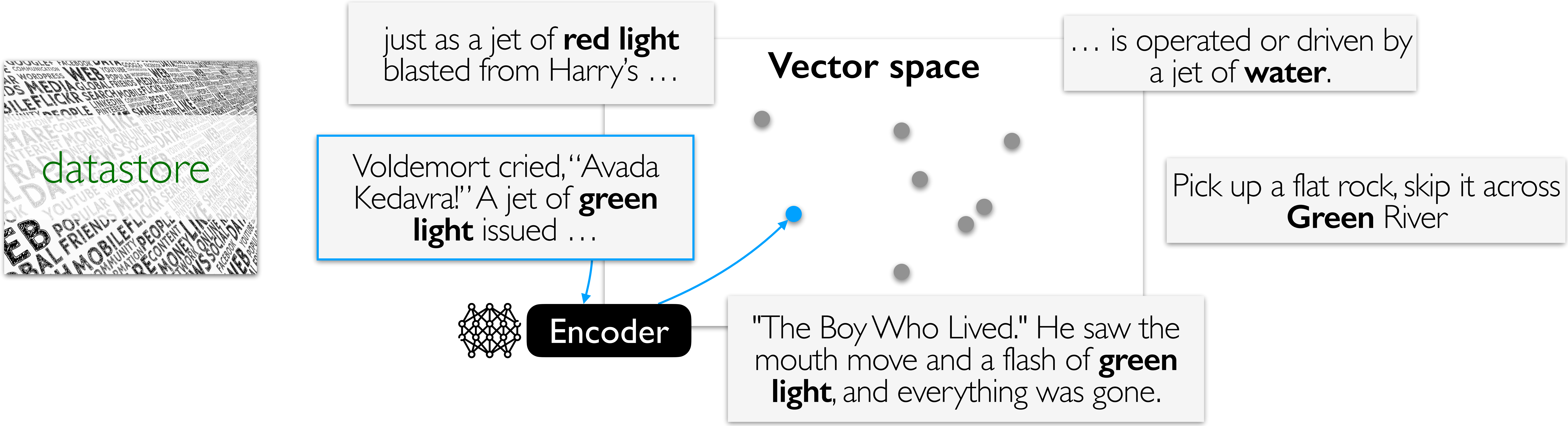
… is operated or driven by a jet of **water**.

Pick up a flat rock, skip it across **Green** River

Min et al. 2023. Nonparametric Masked Language Modeling

Nonparametric-only, Phrase-level (NPM)

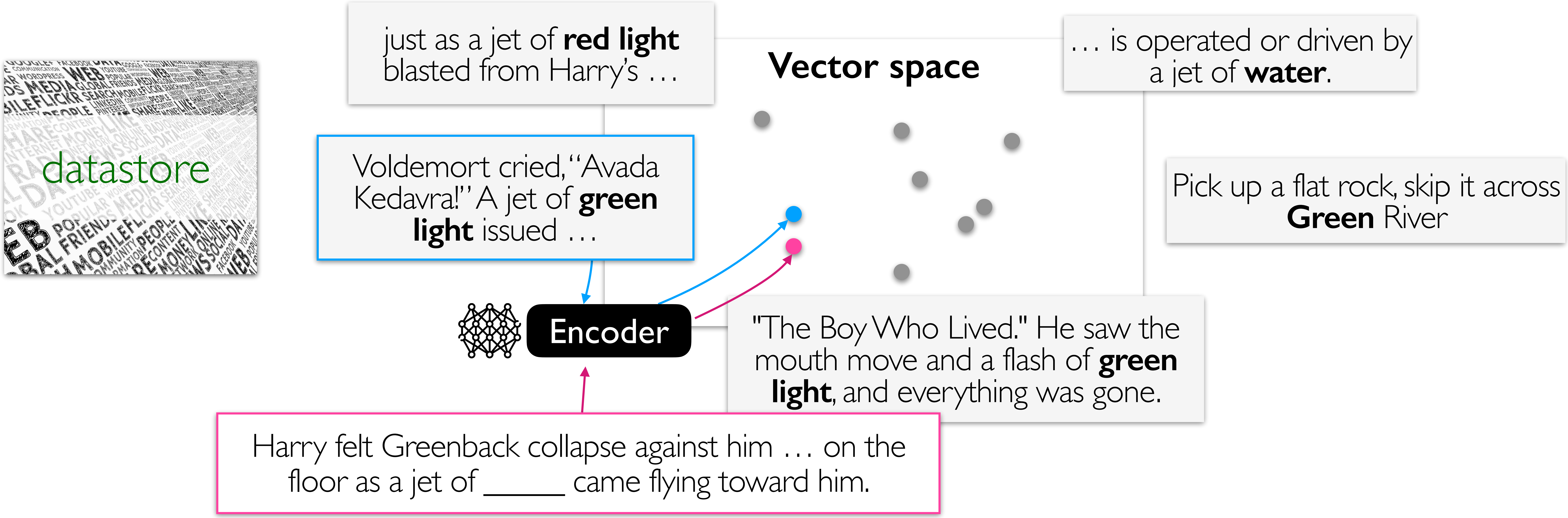
(If you can train the model…)



Min et al. 2023. Nonparametric Masked Language Modeling

Nonparametric-only, Phrase-level (NPM)

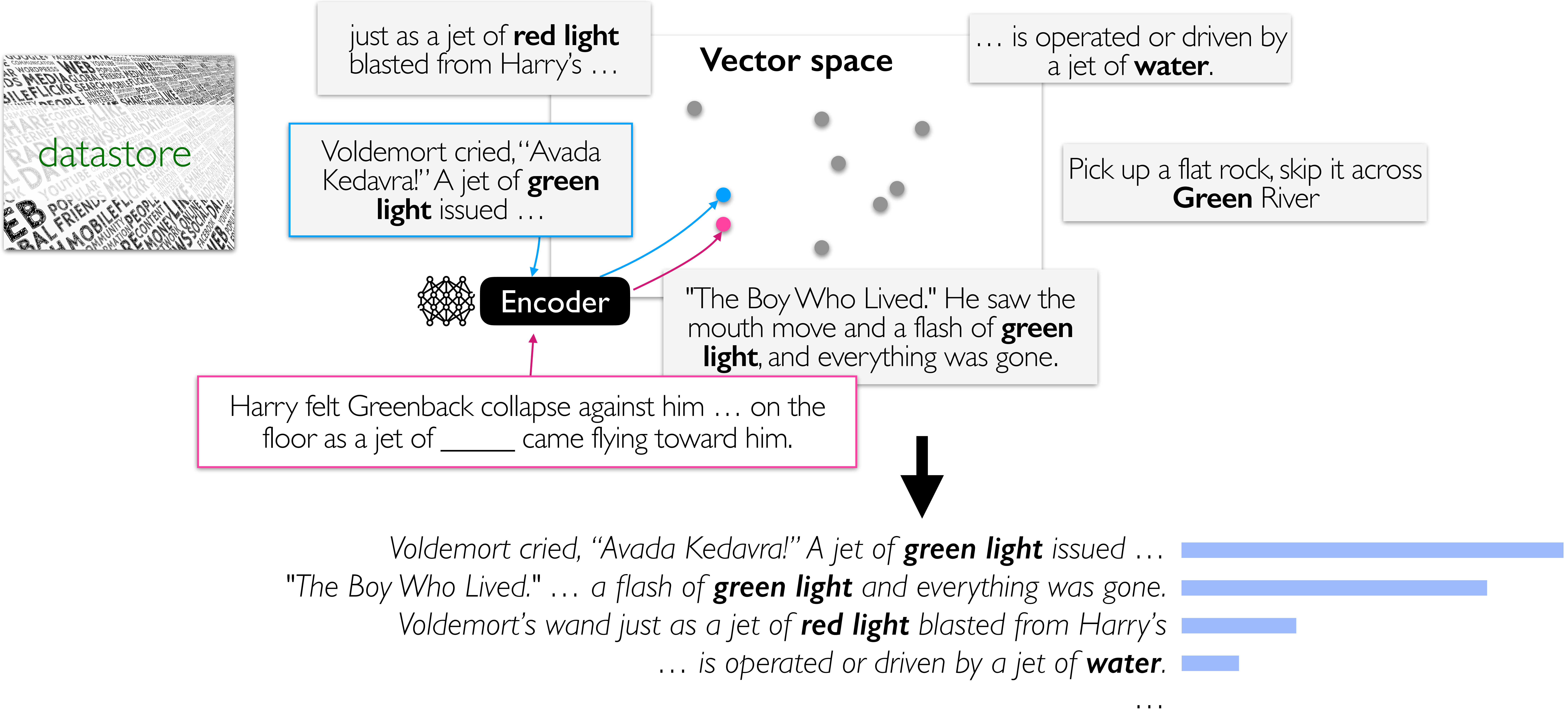
(If you can train the model…)



Min et al. 2023. Nonparametric Masked Language Modeling

Nonparametric-only, Phrase-level (NPM)

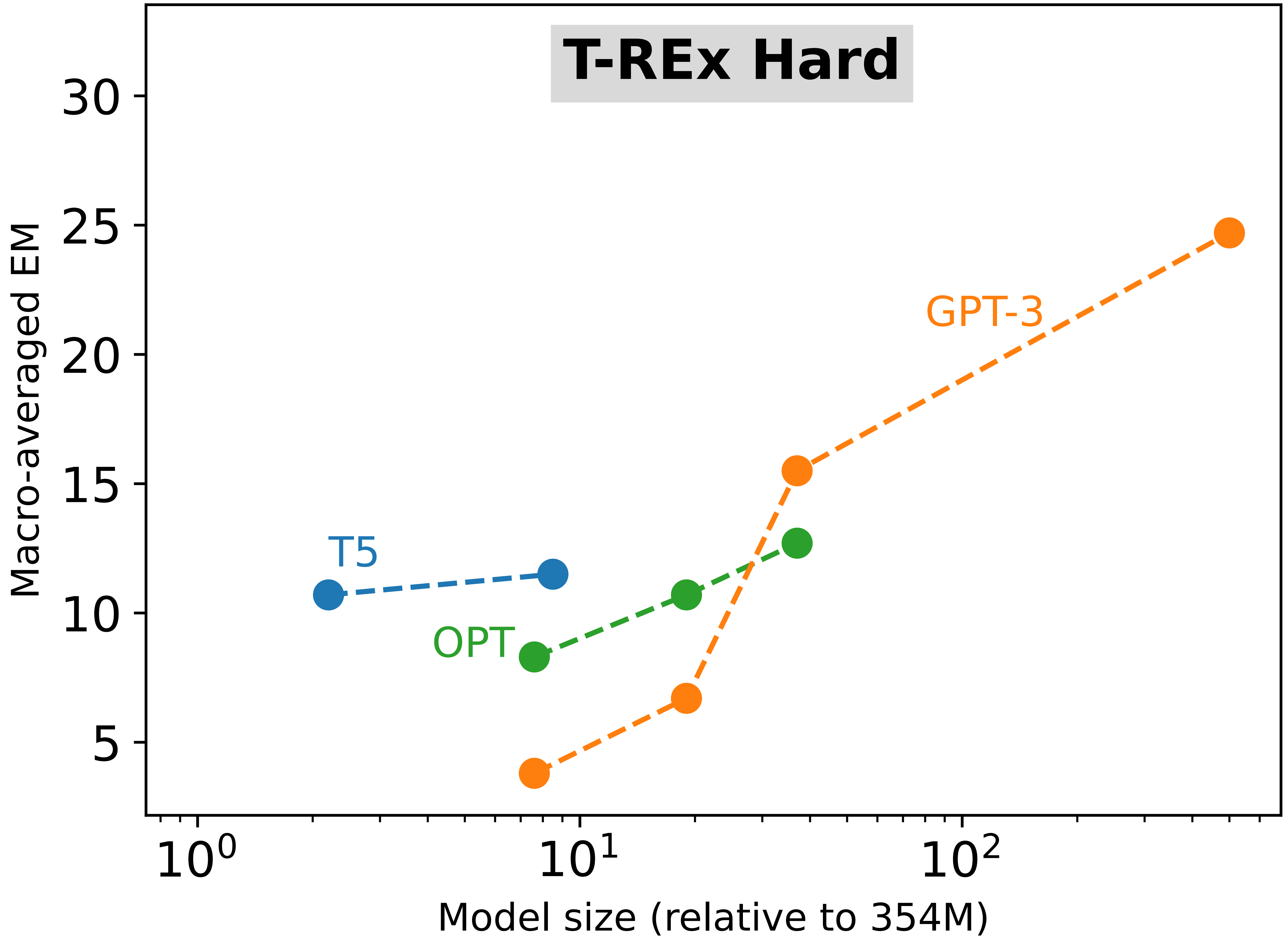
(If you can train the model…)



Min et al. 2023. Nonparametric Masked Language Modeling

NPM: Fact probing

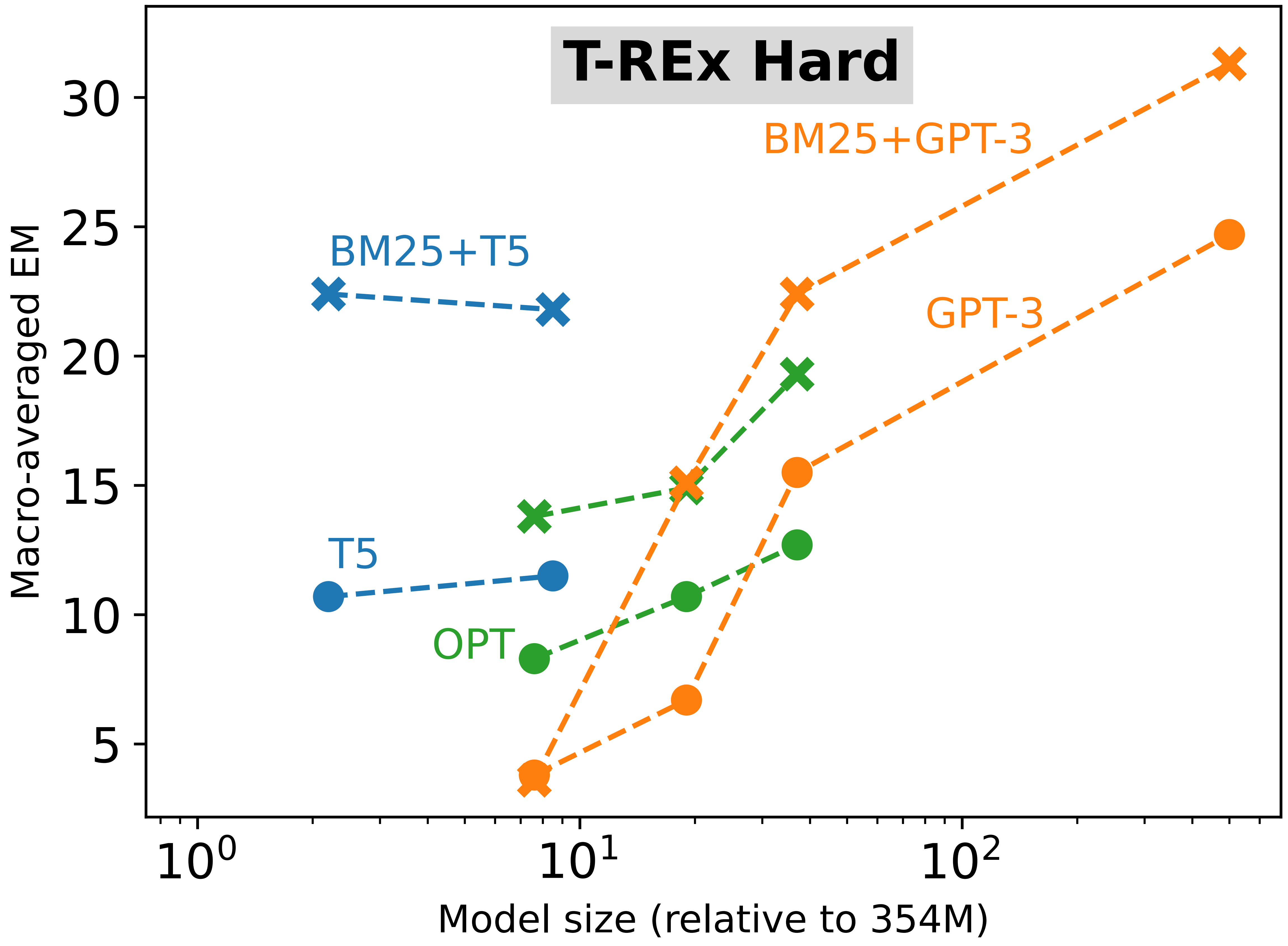
NPM: Fact probing



No-retrieval LMs are better as they get larger

Min et al. 2023. Nonparametric Masked Language Modeling

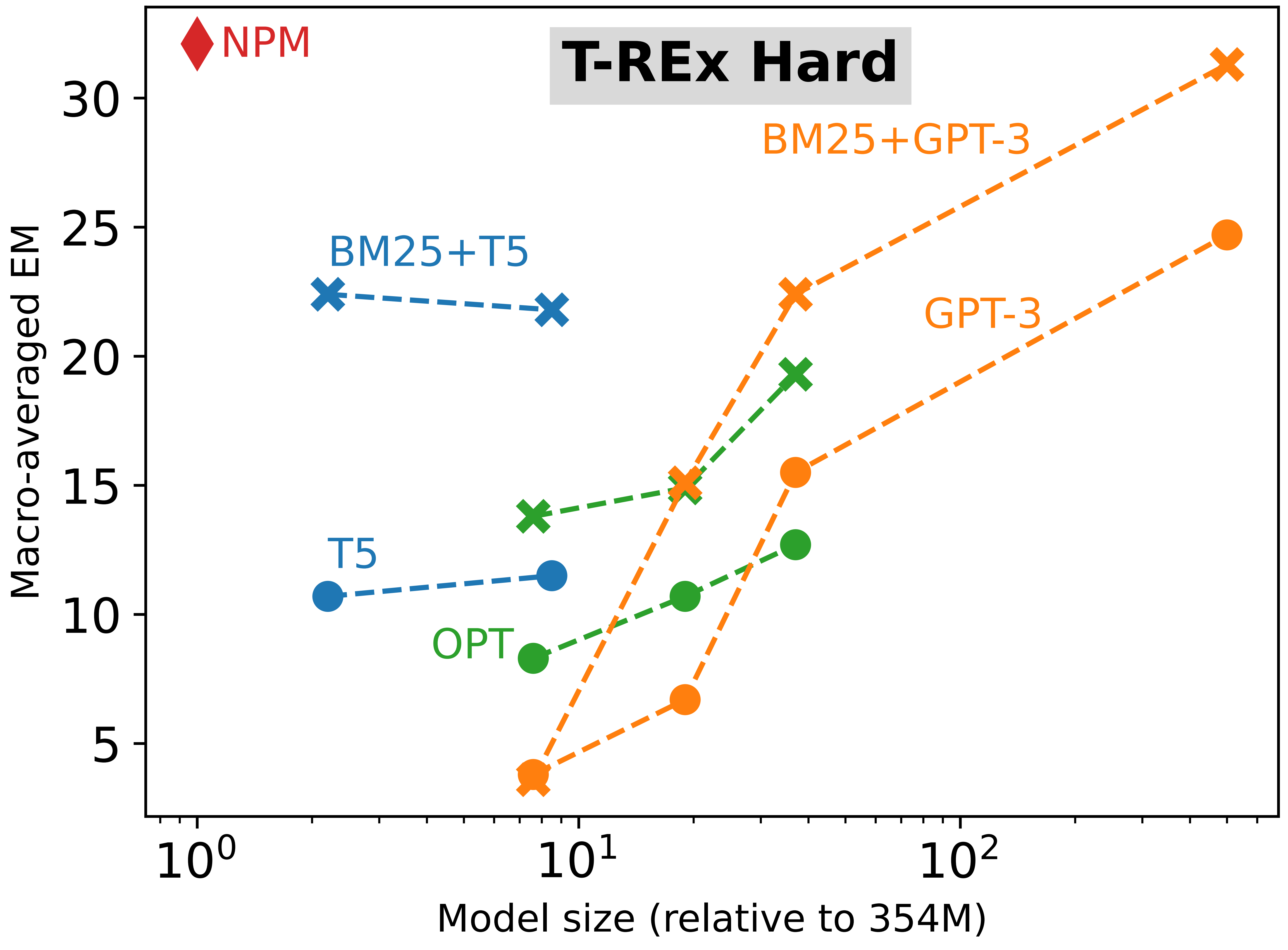
NPM: Fact probing



Retrieval augmentation helps

Min et al. 2023. Nonparametric Masked Language Modeling

NPM: Fact probing



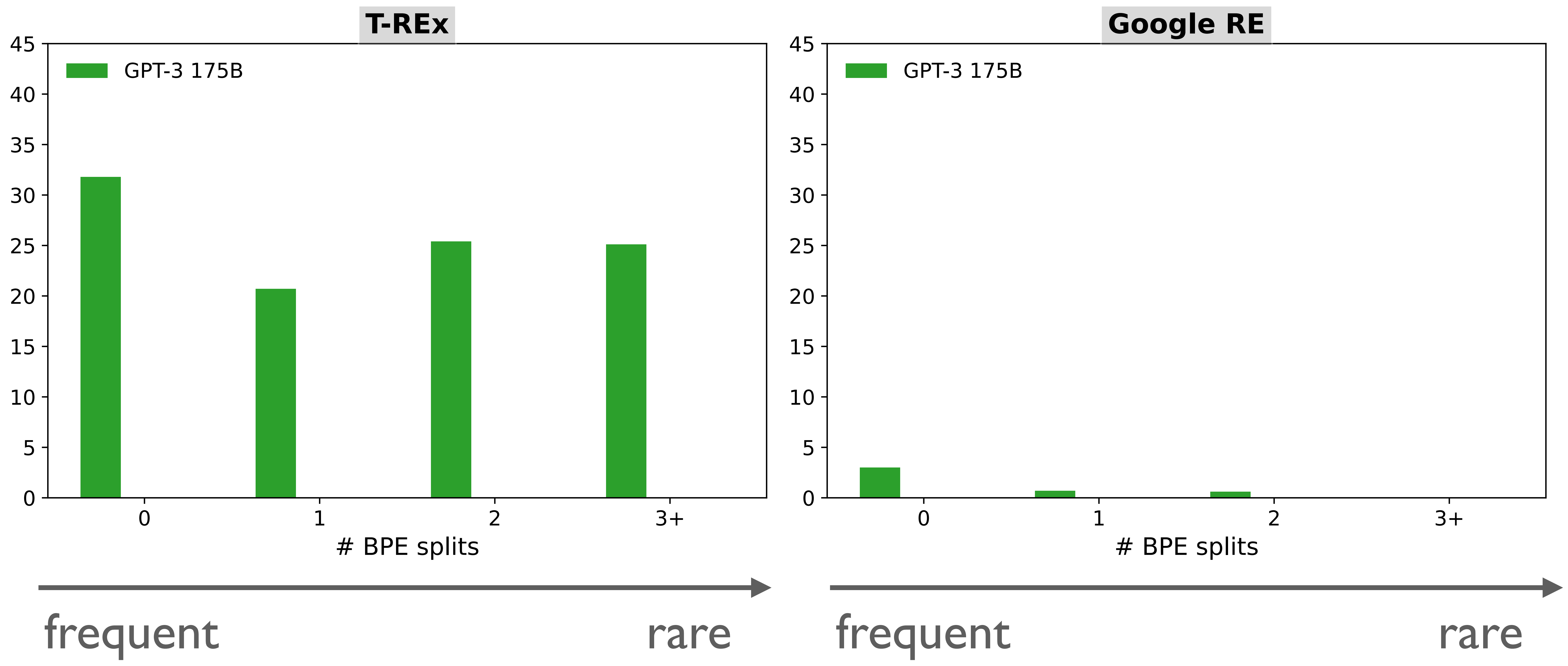
NPM is more parameter efficient

Min et al. 2023. Nonparametric Masked Language Modeling

NPM: Predicting rare entities

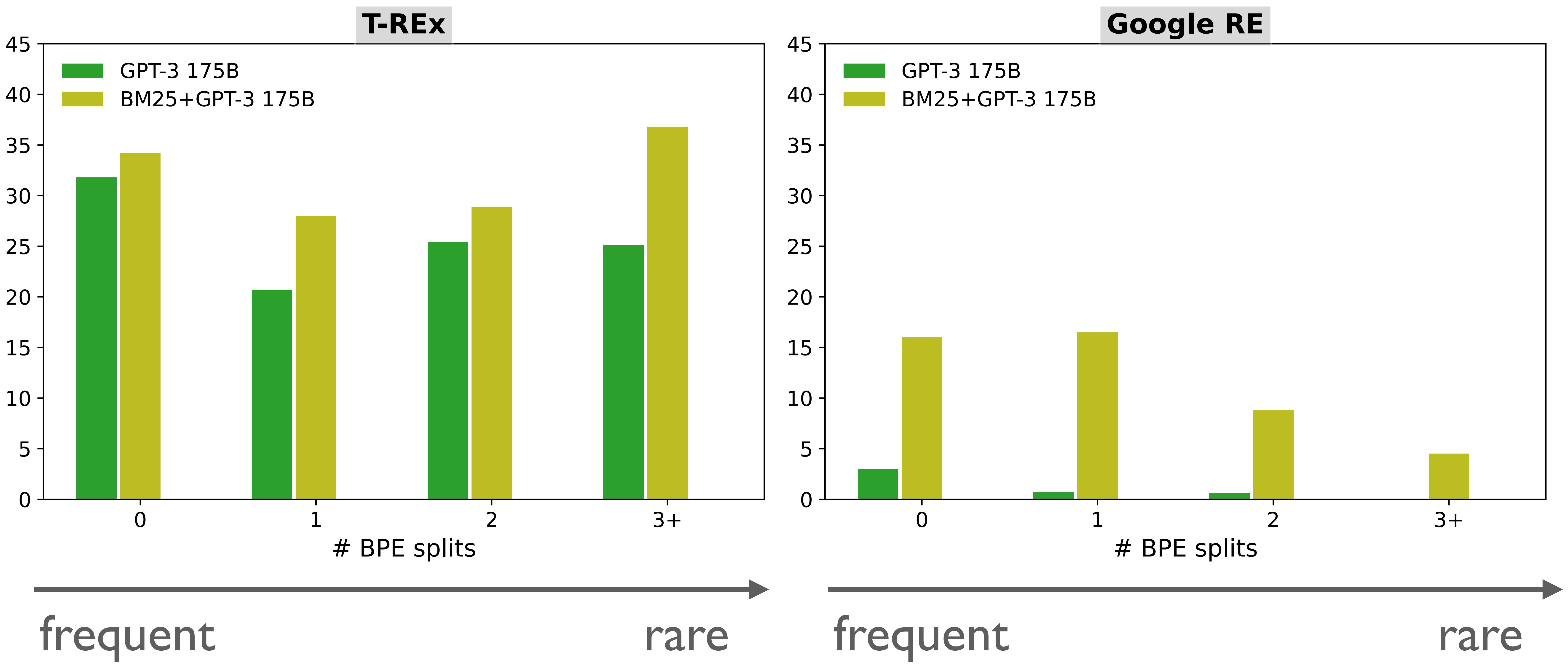
Min et al. 2023. Nonparametric Masked Language Modeling

NPM: Predicting rare entities



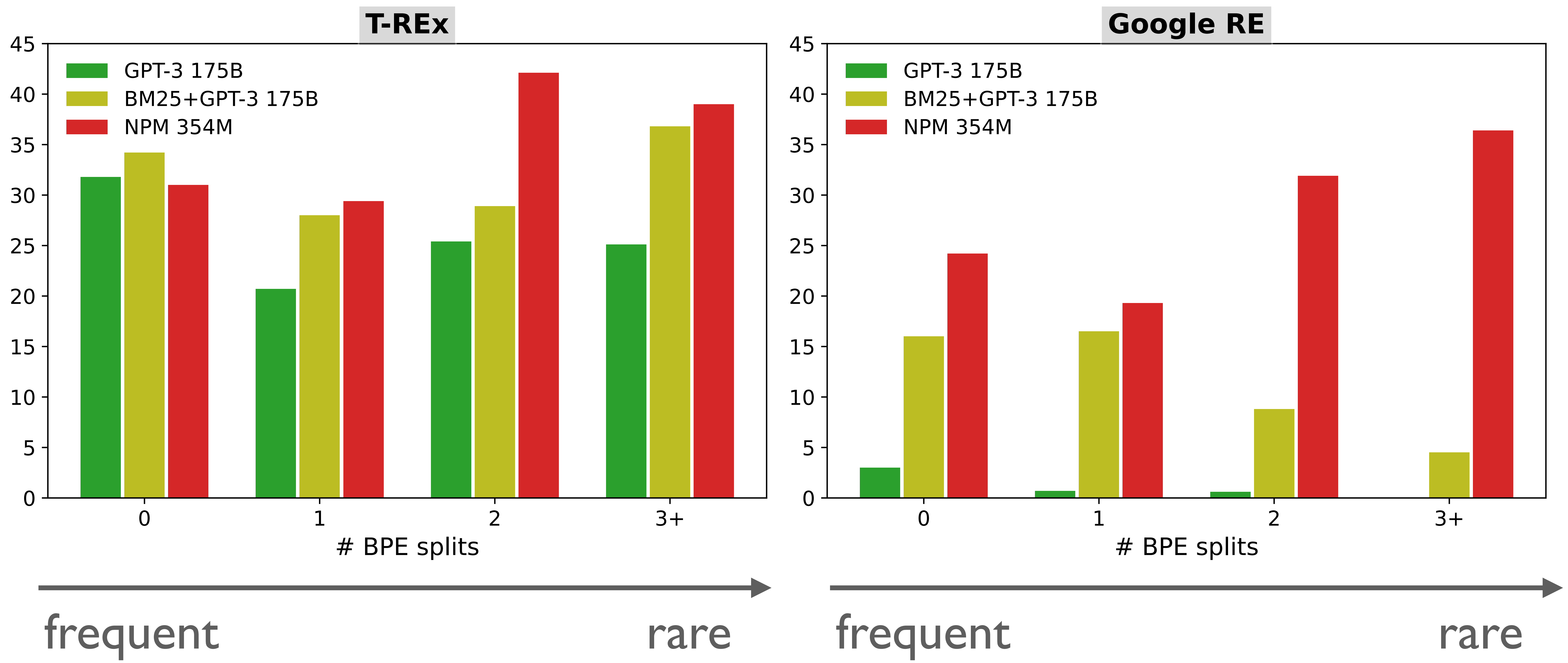
Min et al. 2023. Nonparametric Masked Language Modeling

NPM: Predicting rare entities



Min et al. 2023. Nonparametric Masked Language Modeling

NPM: Predicting rare entities



NPM outperforms by a larger margin as the rarity increases

Min et al. 2023. Nonparametric Masked Language Modeling

New Retrieval-based LMs: Overview

- New Methodology 1 — Designing a new Transformer
  - New attention layers to incorporate more blocks (RETRO)
  - Possibly combine with long-range Transformers
- New Methodology 2 — Designing a new Softmax
  - **Two softmaxes together: kNN-LM**
  - **Nonparametric softmax only, phrase-level: NPM**
- New LM Design — Mitigating fairness & legality issues

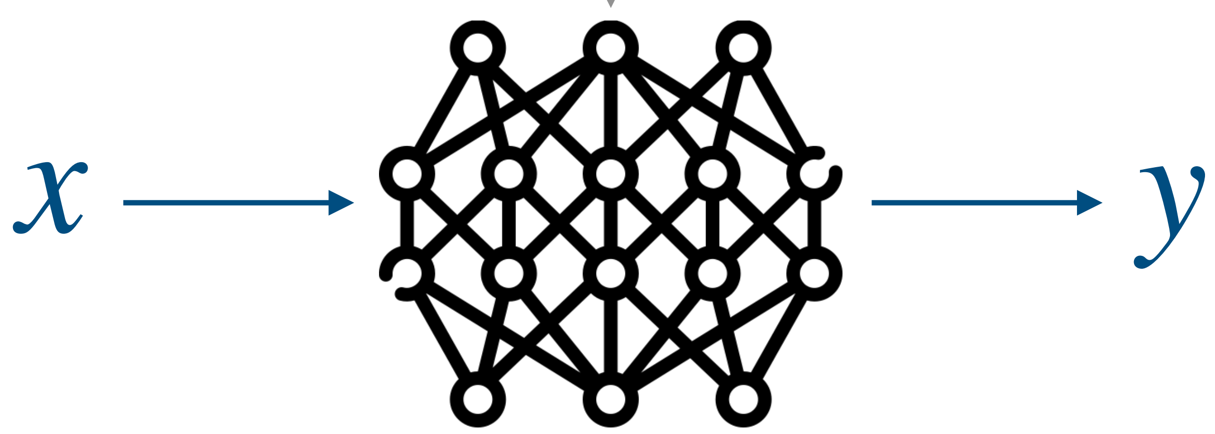
New Retrieval-based LMs: Overview

- New Methodology 1 — Designing a new Transformer
    - New attention layers to incorporate more blocks (RETRO)
    - Possibly combine with long-range Transformers
- New Methodology 2 — Designing a new Softmax
    - Two softmaxes together: kNN-LM
    - Nonparametric softmax only, phrase-level: NPM
- **New LM Design — Mitigating fairness & legality issues**

Common practice

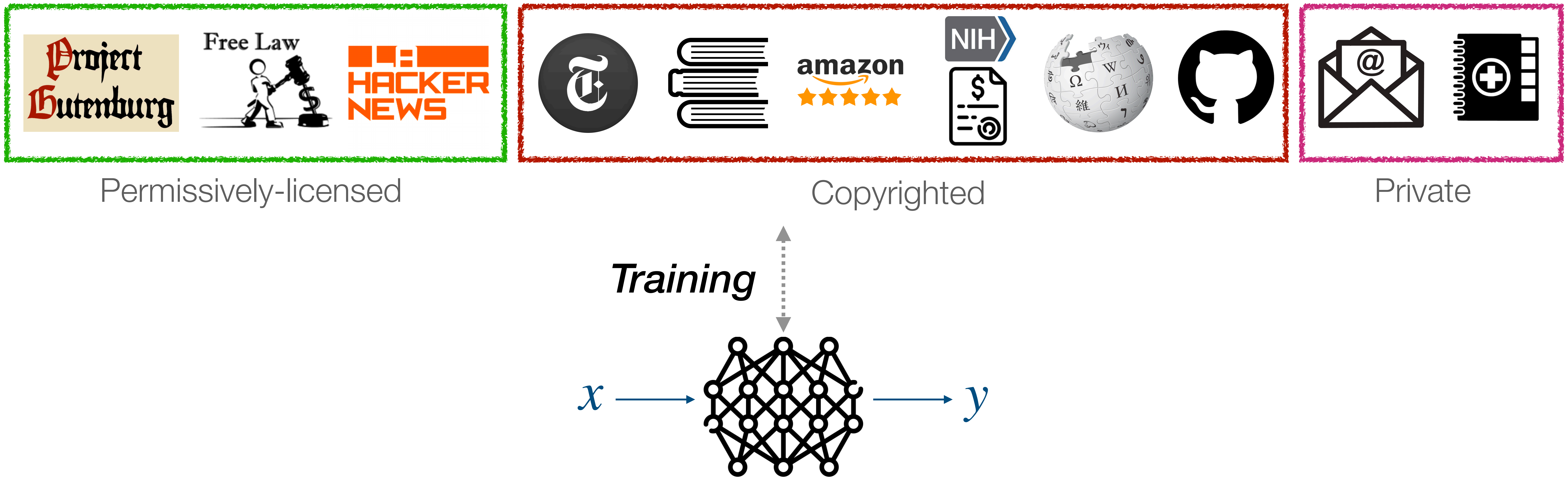
Web crawl

Training



Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

Common practice

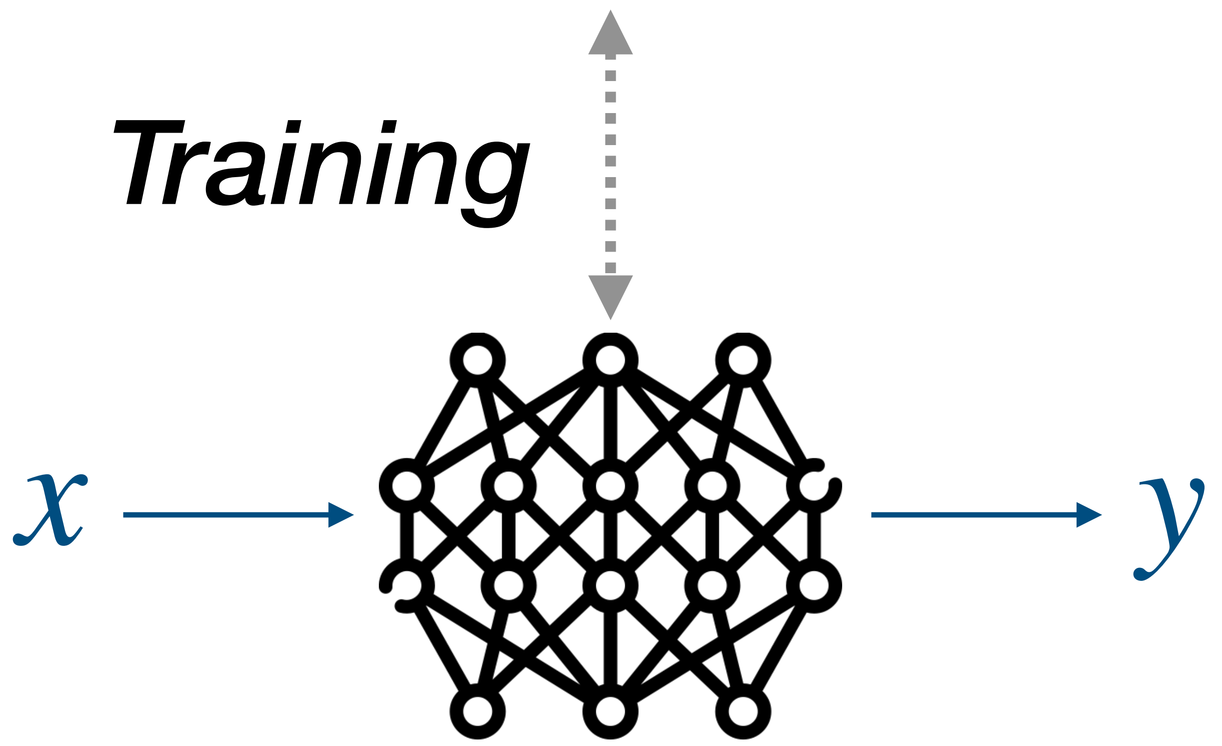


Common practice

Permissively-licensed

Copyrighted

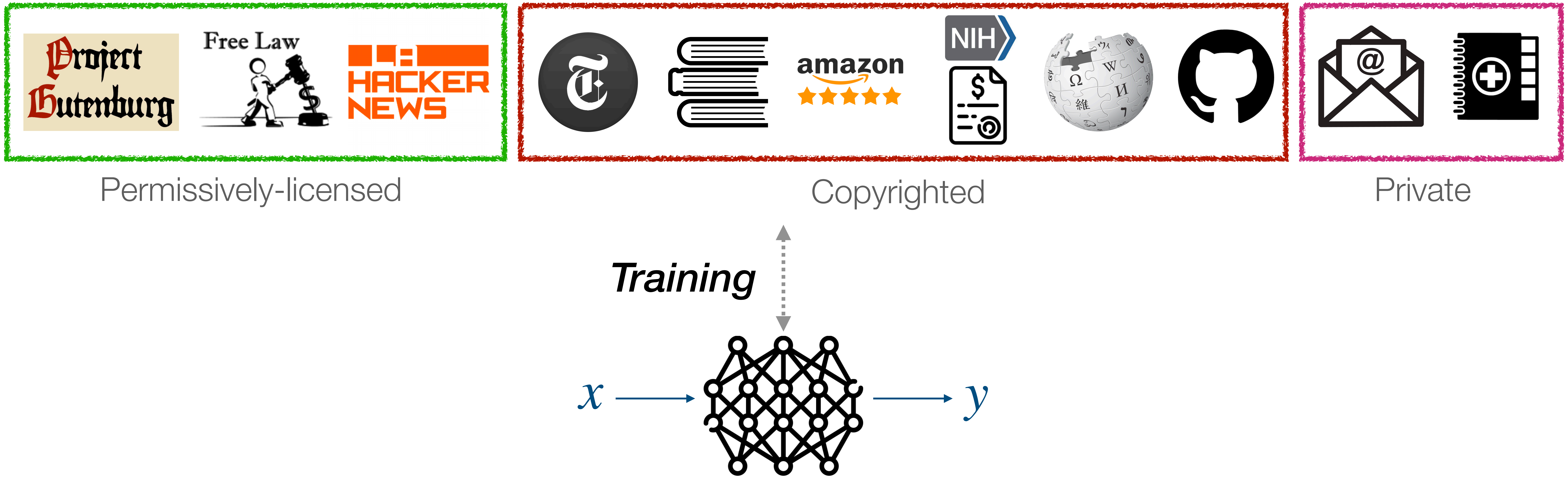
Private



Legal risk in training on copyrighted data

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

Common practice

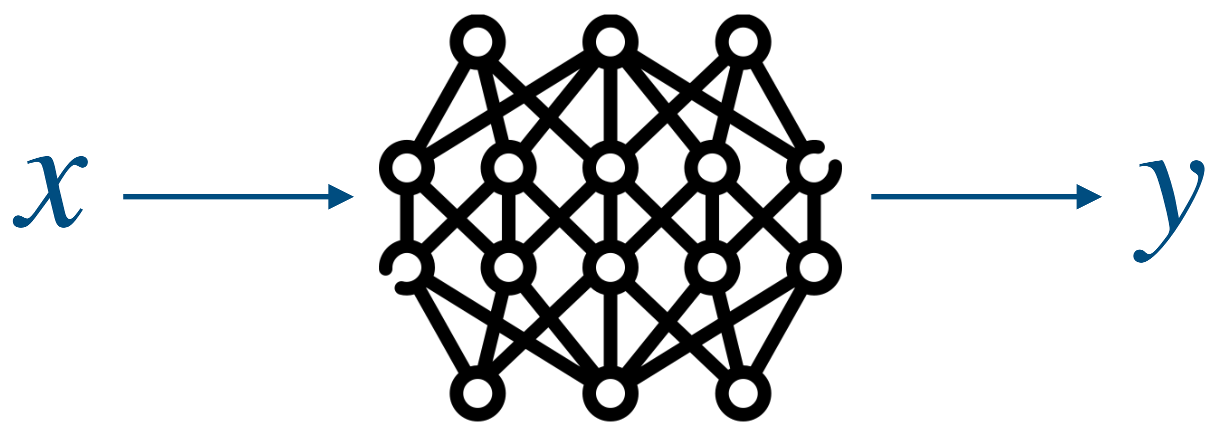


☹️ Legal risk in training on copyrighted data ☹️ Failure in crediting to data creators

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

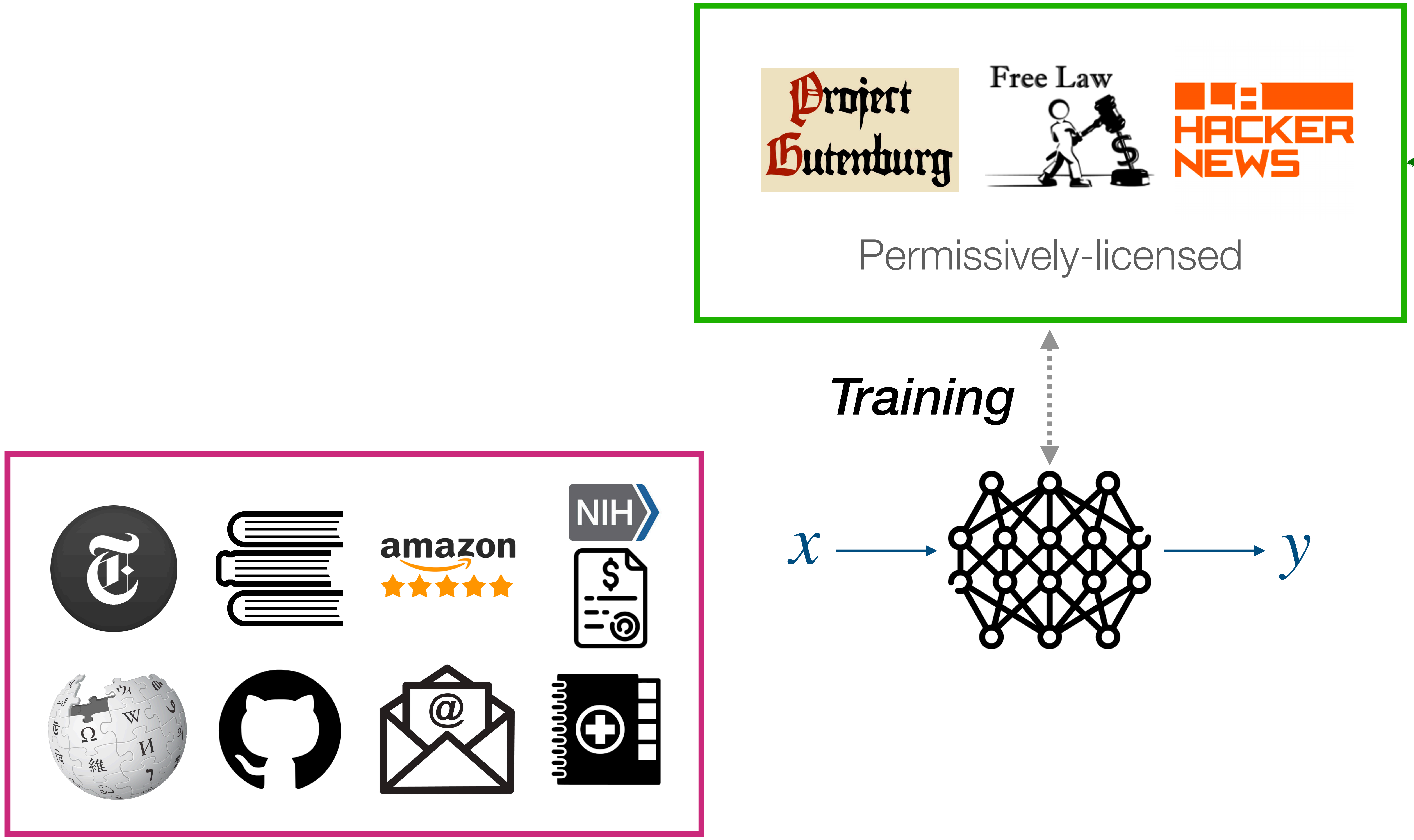
New proposal: SILO

Permissively-licensed



Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

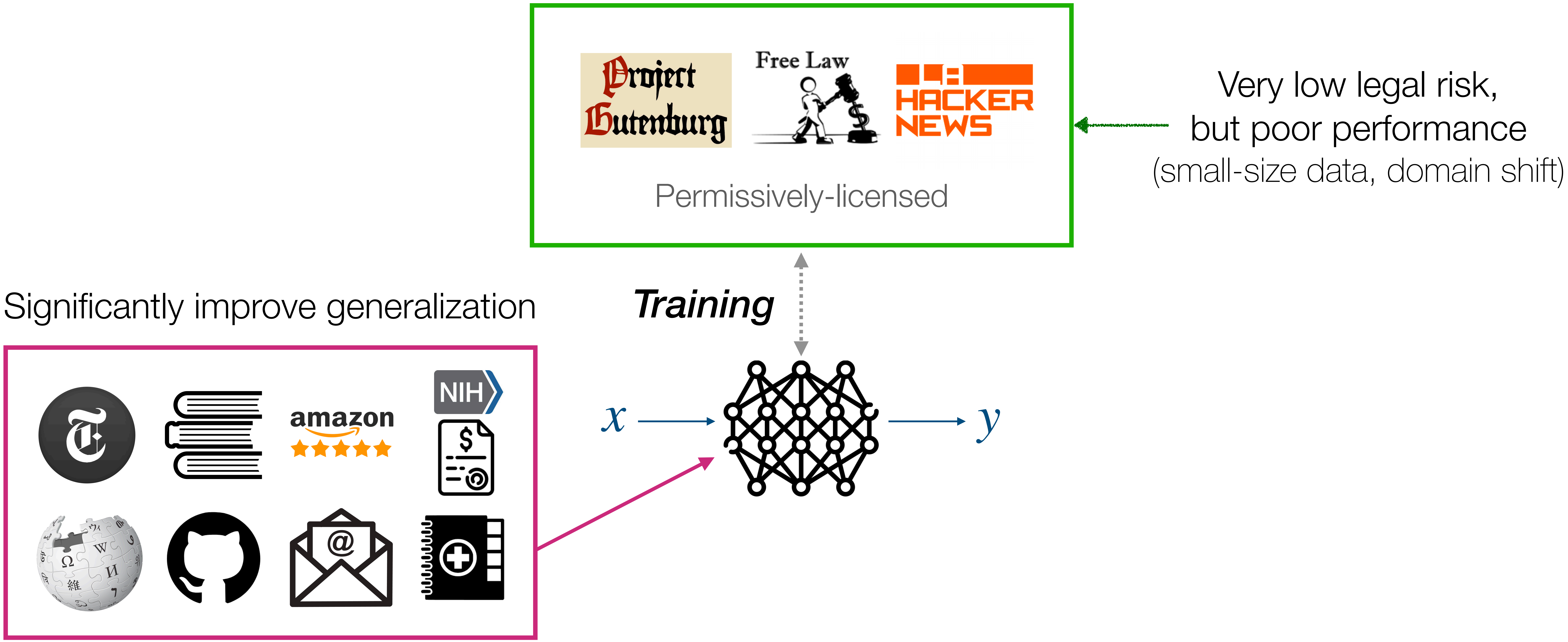
New proposal: SILO



Very low legal risk,
but poor performance
(small-size data, domain shift)

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

New proposal: SILO



Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

New proposal: SILO

Free Law

HACKER NEWS

Permissively-licensed

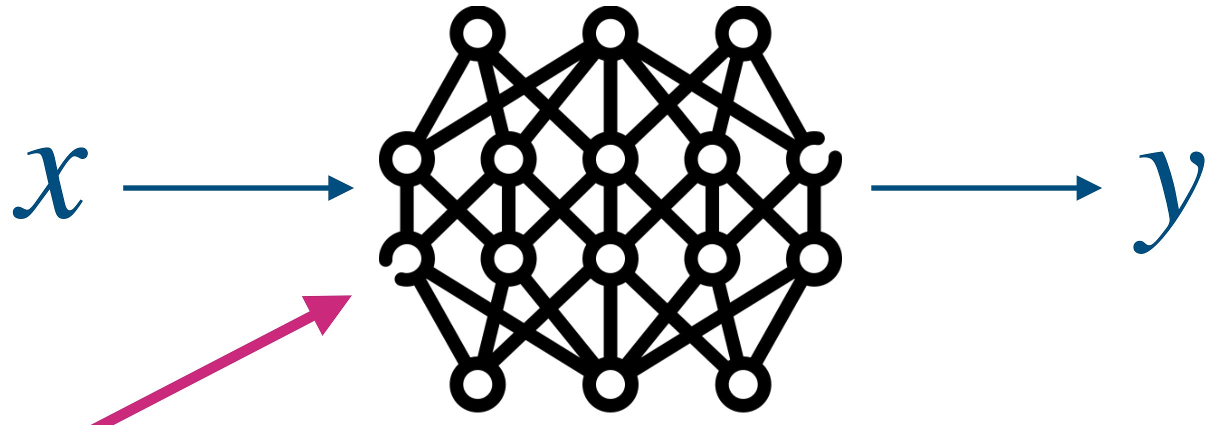
Very low legal risk,
but poor performance
(small-size data, domain shift)

Training

Significantly improve generalization

NIH

amazon

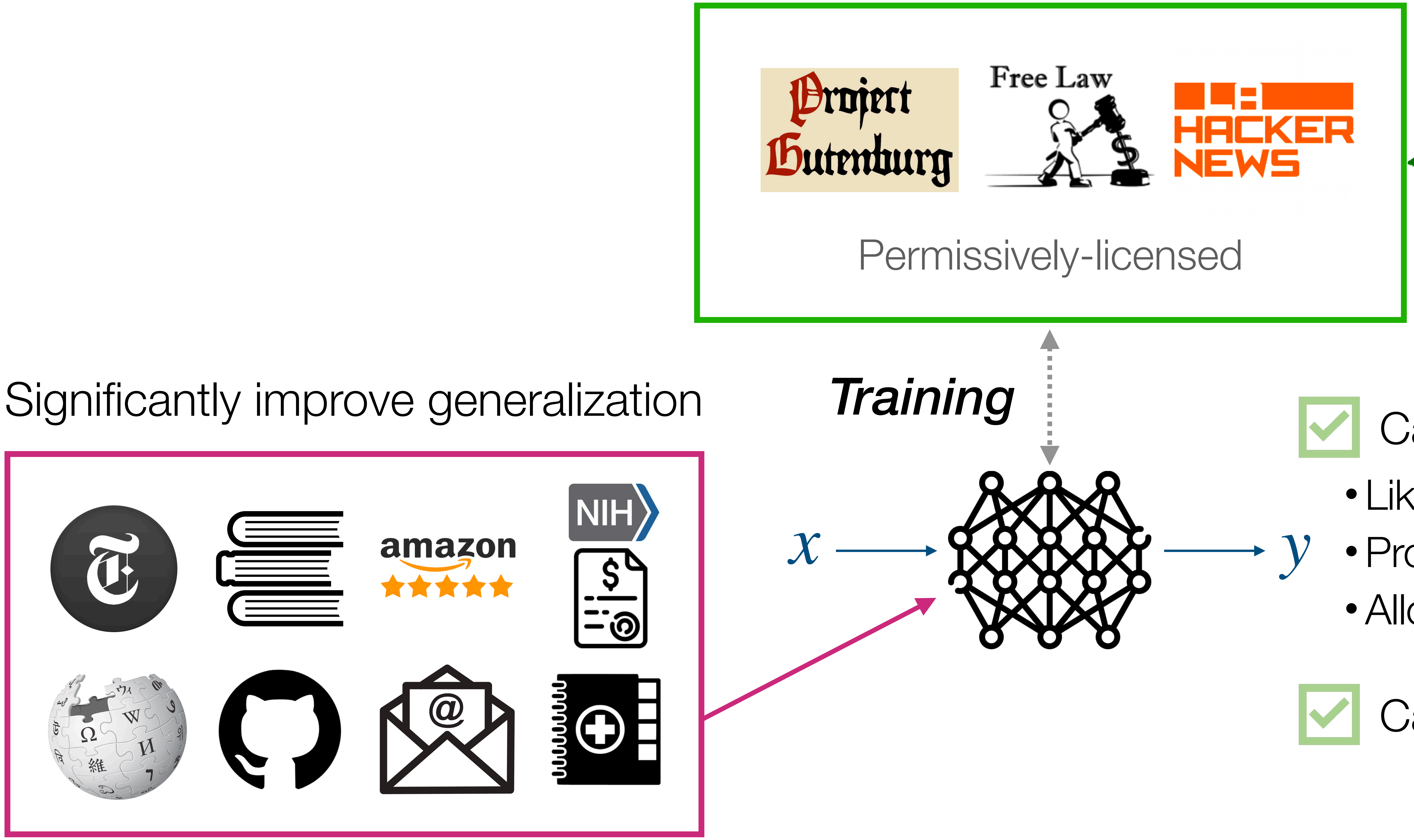


Can trace inherent attribution

Can modify the datastore at any time

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

New proposal: SILO



Very low legal risk,
but poor performance
(small-size data, domain shift)

☑ Can trace inherent attribution

- Likely defense *fair use*
- Provide copyright notice
- Allow credits (or payment) to data creators

☑ Can modify the datastore at any time

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

New proposal: SILO

Project Gutenburg
Free Law
HACKER NEWS

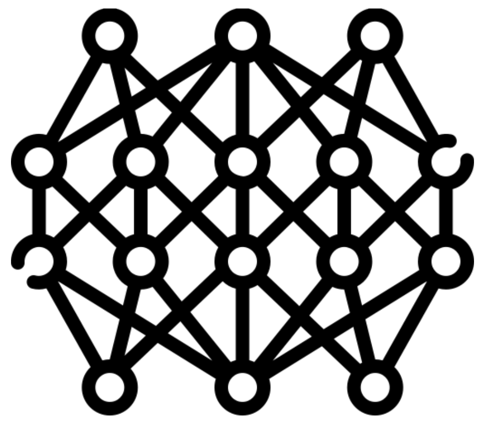
Permissively-licensed

Very low legal risk,
but poor performance
(small-size data, domain shift)

Training

Significantly improve generalization

amazon
NIH

x →  → y

Can trace inherent attribution

- Likely defense *fair use*
- Provide copyright notice
- Allow credits (or payment) to data creators

Can modify the datastore at any time

- Support removal of data at any time
- Better alignment with *GDPR*

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

SILO Attribution Example

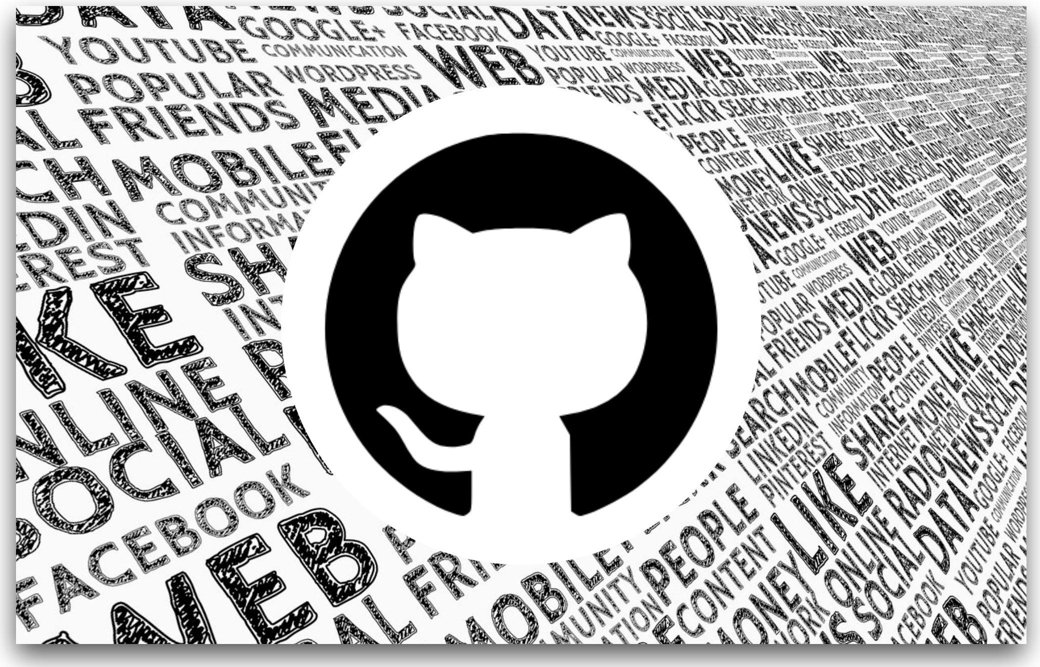
Test input:
include ‘../lib/admin.defines.php’;
include ‘../lib/admin.module.access.php’;
include ‘../lib/admin.smarty.php’;
if (! has_right (

Continuation: **[AC]**X_BILLING)) { Header …

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

SILO Attribution Example

Test input:
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (



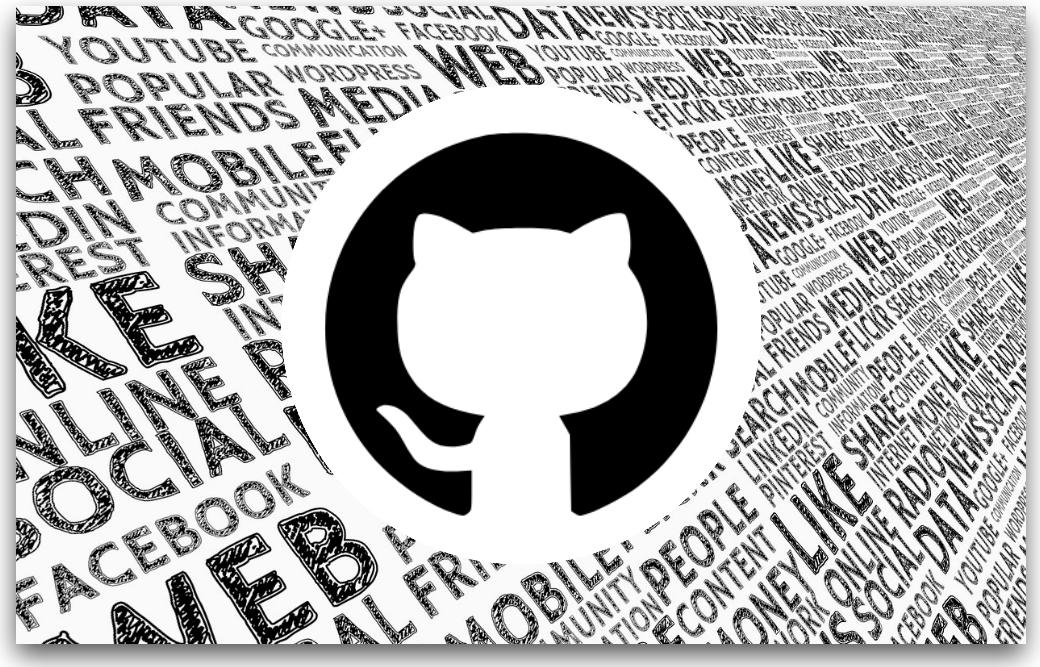
Continuation: **[AC]**X_BILLING)) { Header …

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

SILO Attribution Example

Test input:
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (

Continuation: **[AC]**X_BILLING)) { Header …



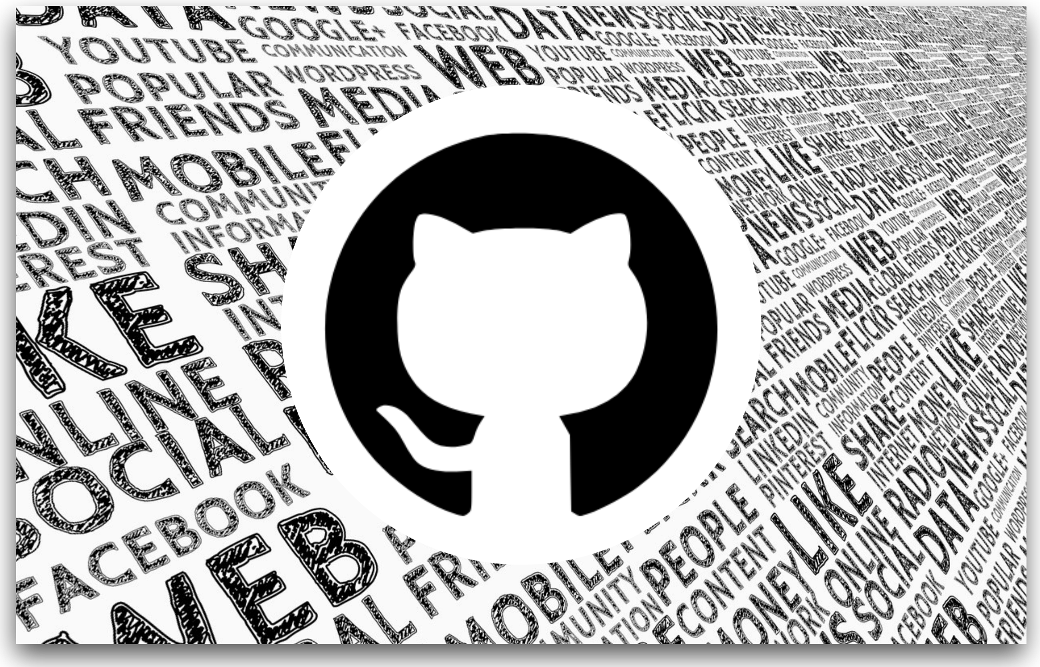
Top-1 retrieved token (in kNN-LM):
* You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
[AC]X_ACCESS)) { Header …

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

SILO Attribution Example

Test input:
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (

Continuation: **[AC]**X_BILLING)) { Header …



Top-1 retrieved token (in kNN-LM):
* You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
  [AC]X_ACCESS)) { Header …

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

New Retrieval-based LMs: Summary

- New Methodology 1 — Designing a new Transformer
  - New attention layers to incorporate more blocks (RETRO)
- New Methodology 2 — Designing a new Softmax
  - Two softmaxes together: kNN-LM
  - Nonparametric softmax only, phrase-level: NPM
- New LM Design — Mitigating fairness & legality issues
  - Train on permissive text → place copyrighted text into a datastore

Overview

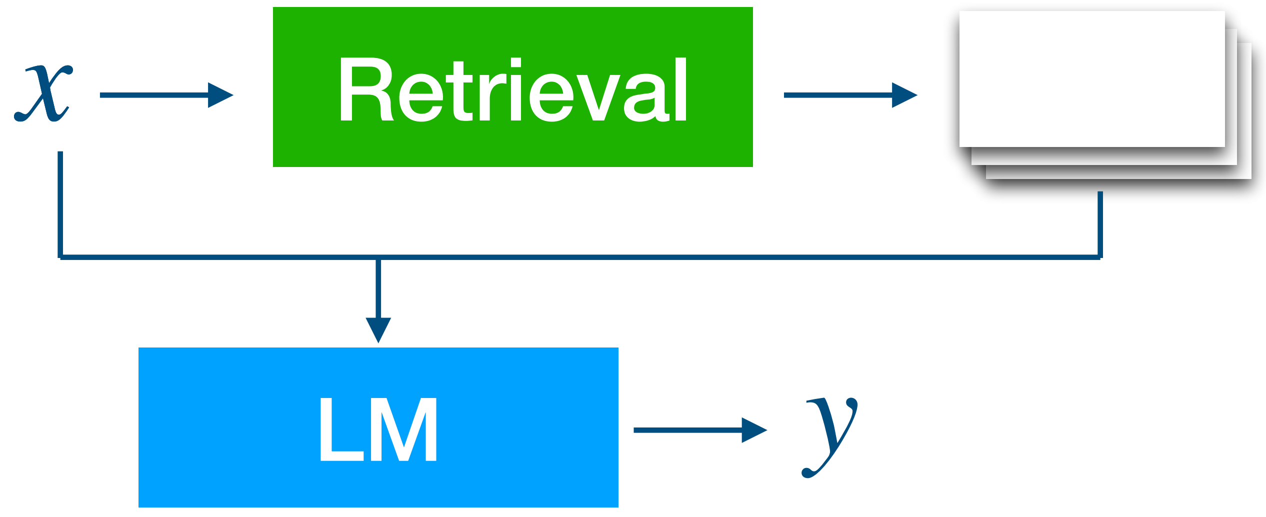
Why Retrieval-based LMs?

Tell me about Meta Platform.

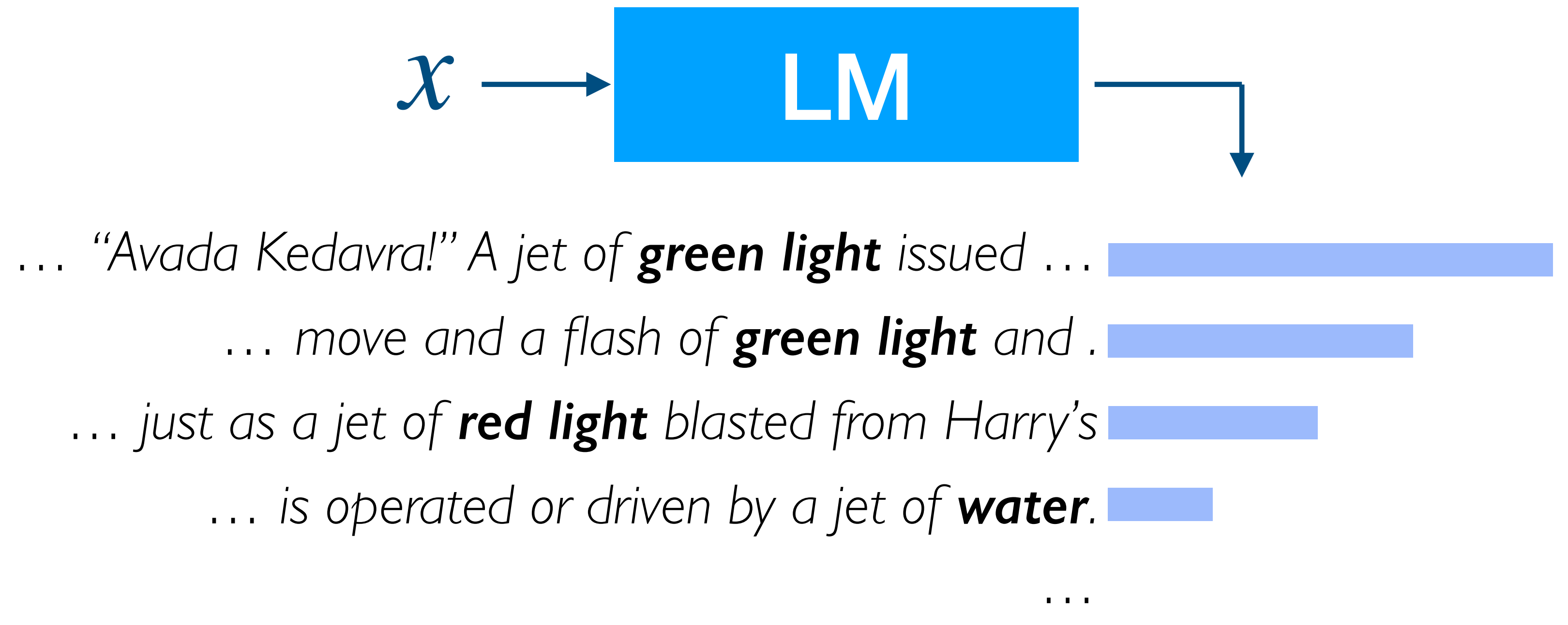
ChatGPT

I don't have any information about a company called Meta Platforms. It is possible that the company is …

Retrieval Augmentation



New Retrieval-based LMs



Open Problems

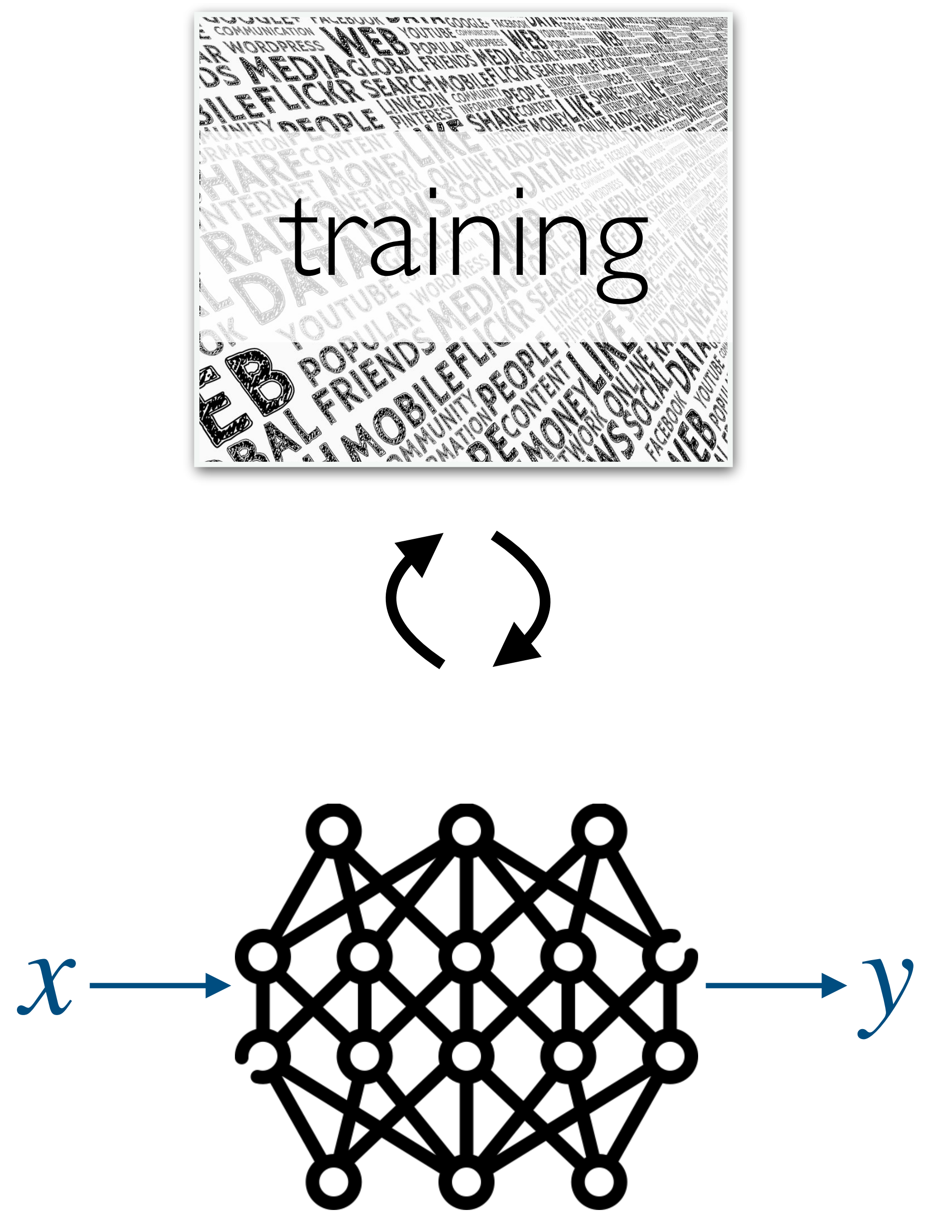
Scaling **datastore** not just parameters?

Summary

What? | **How?** | **Why?**

Summary

What? **How?** **Why?**

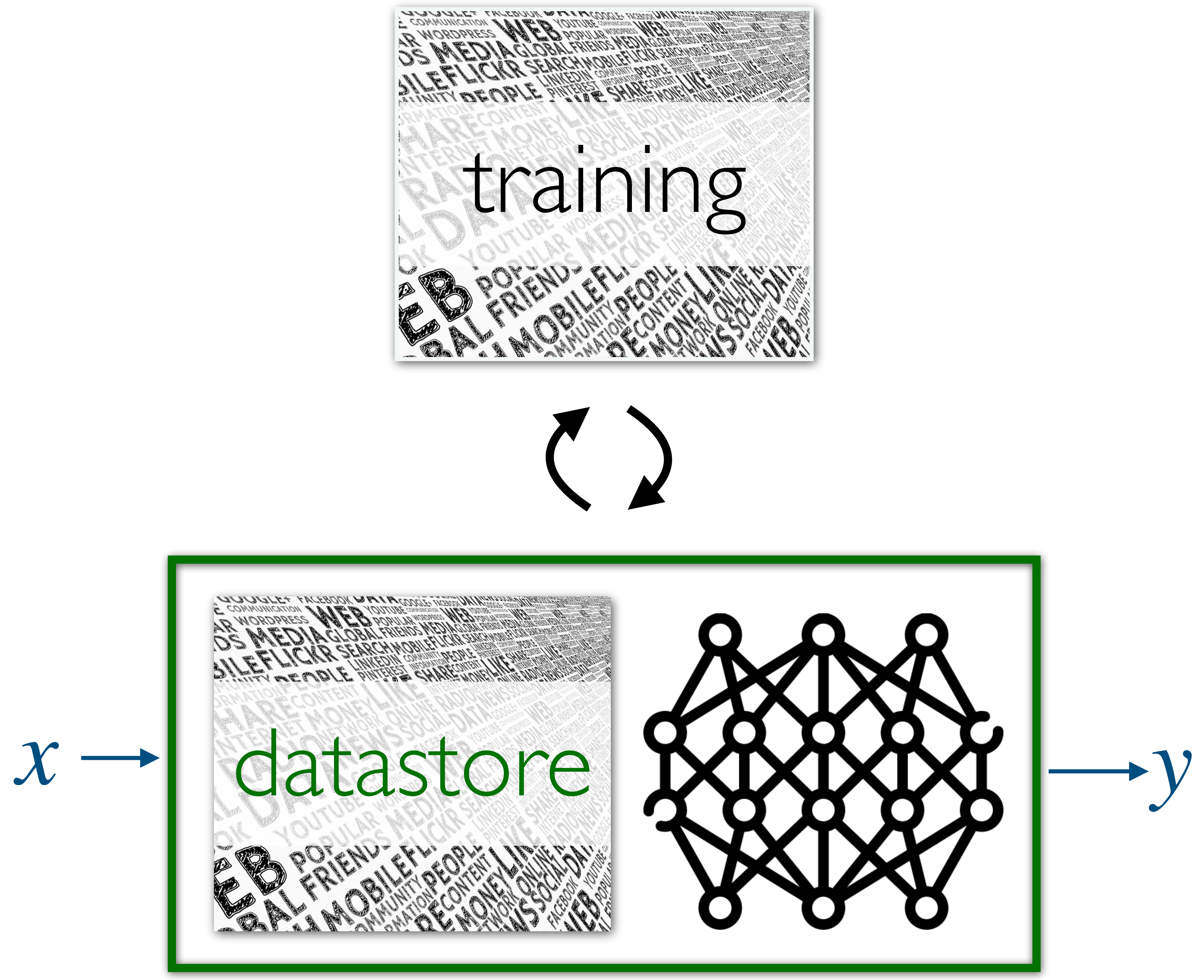


(Typical LMs)

x: test input
y: model prediction to x

Summary

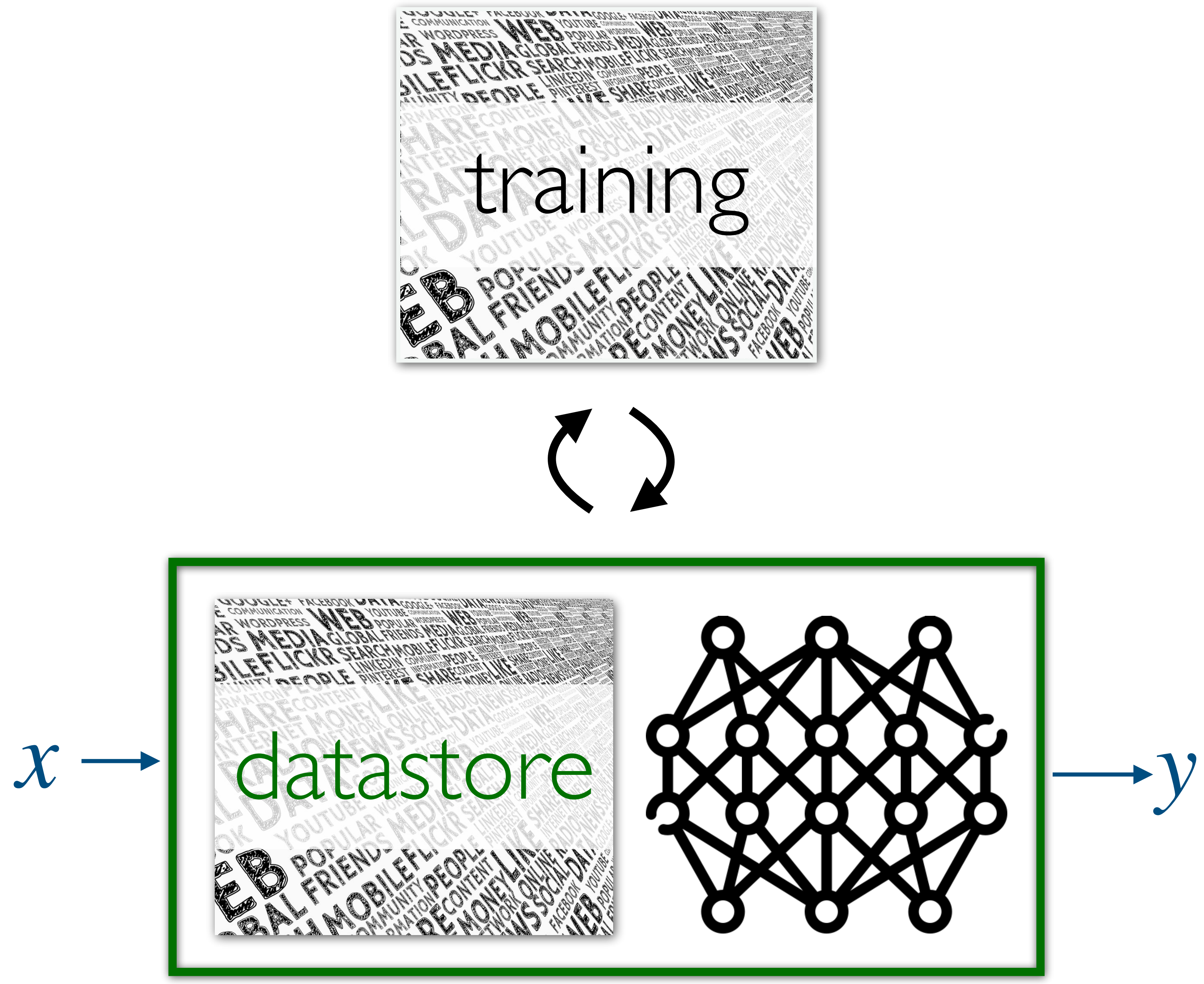
What? **How?** **Why?**



x: test input
y: model prediction to x

Summary

What?



x: test input
y: model prediction to x

How?

Retrieval augmentation

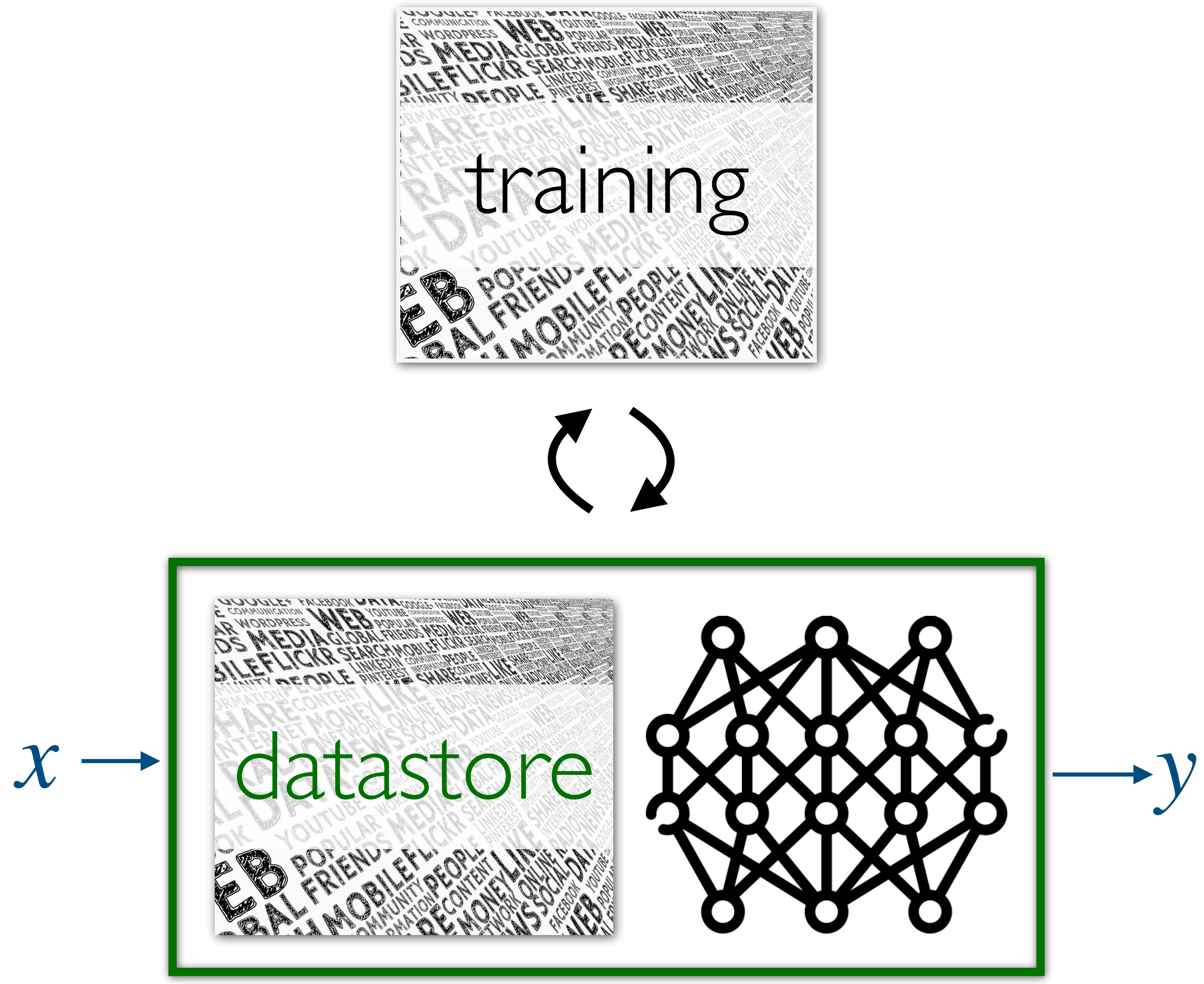
New Transformers

Nonparametric Softmax

Why?

Summary

What?



x: test input
y: model prediction to x

How?

Retrieval augmentation

New Transformers

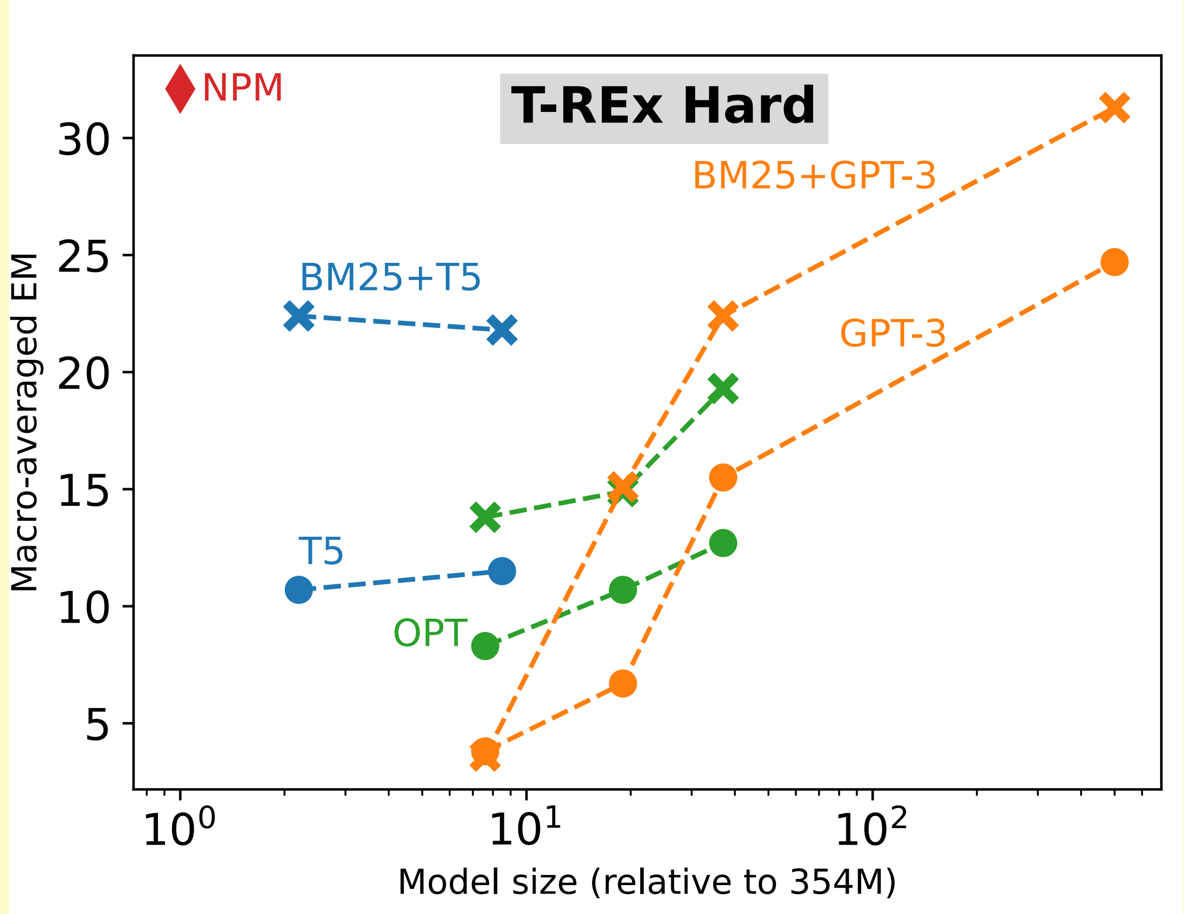
Nonparametric Softmax

Why?

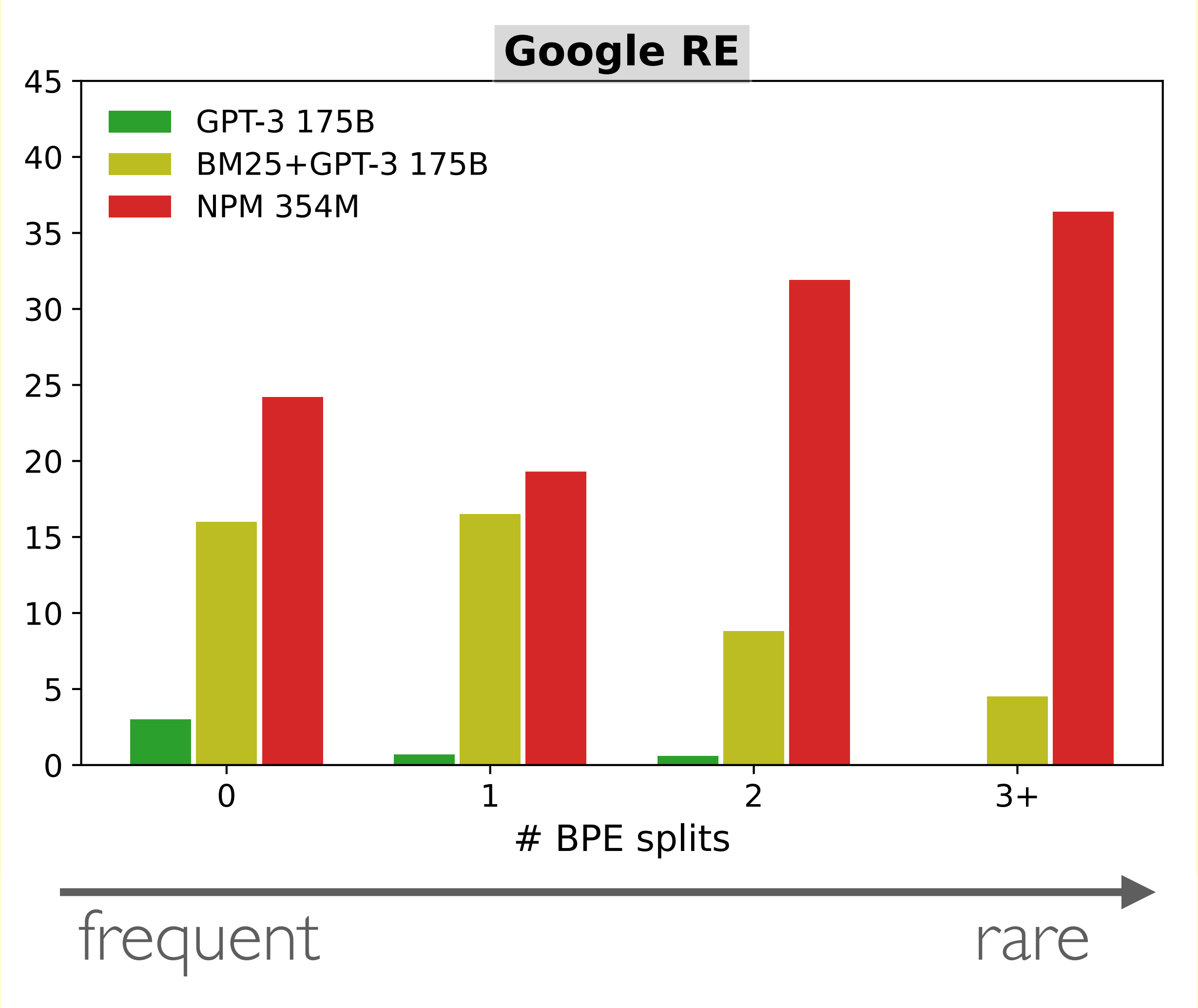
New dimension in improving LMs!

Summary

slide 86 ↓



slide 89 ↓

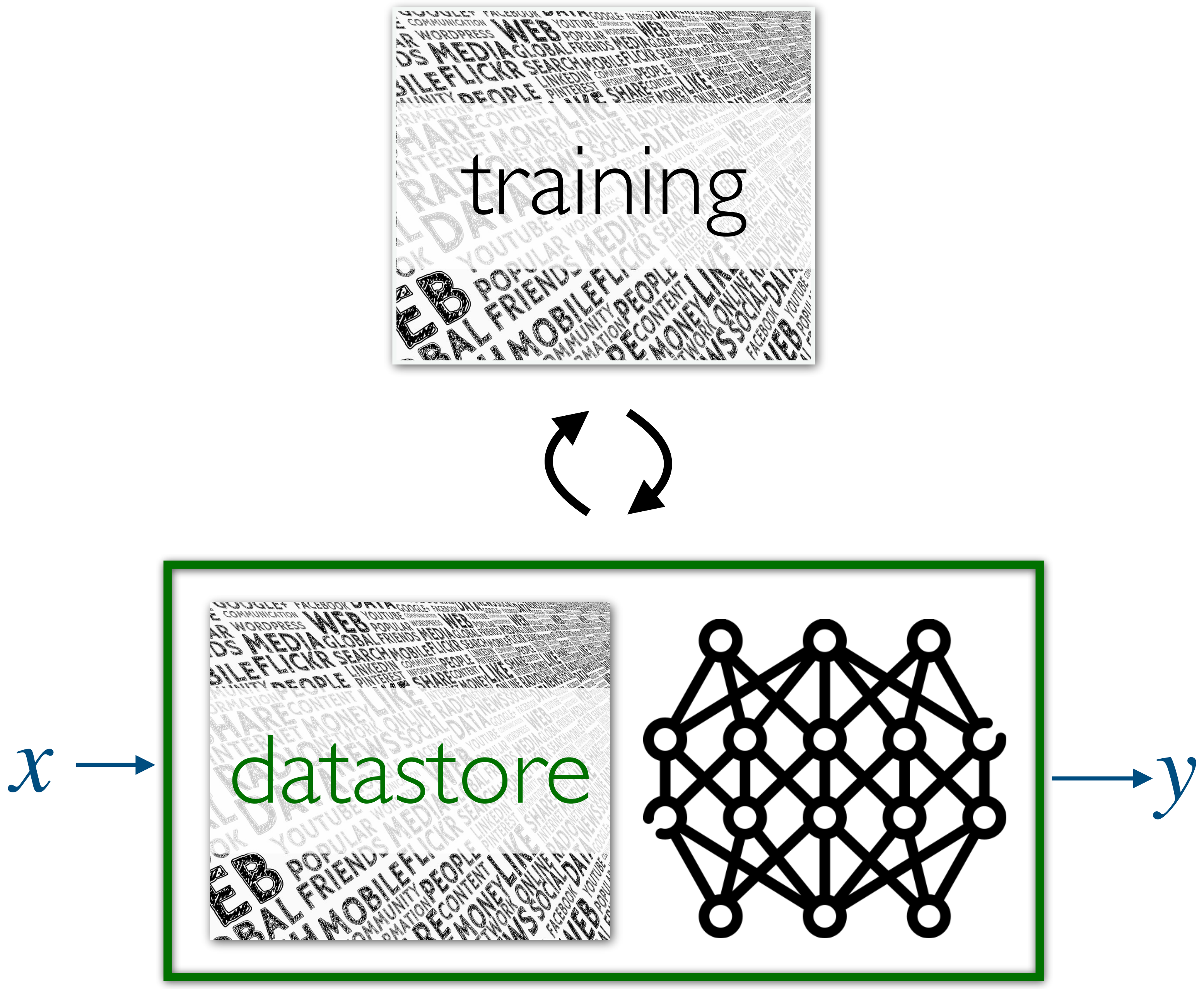


Why?

New dimension in improving LMs!

Summary

What?



x: test input
y: model prediction to x

How?

Retrieval augmentation

New Transformers

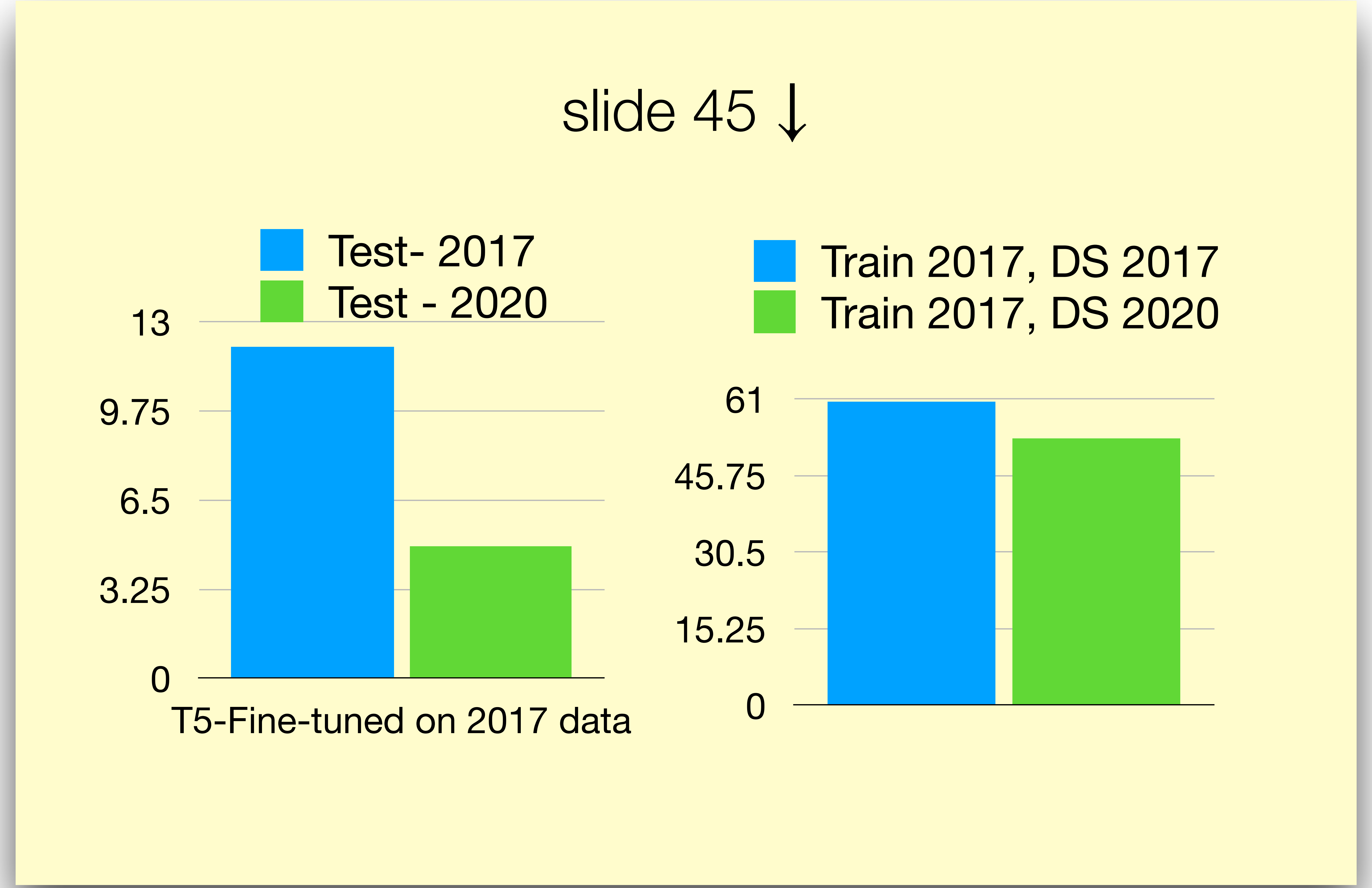
Nonparametric Softmax

Why?

New dimension in improving LMs!

Update & scale without additional training

Summary



Why?

New dimension in improving LMs!

Update & scale without additional training

Summary

slide 96 ↓

Test input:
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (

Continuation: **[AC]**X_BILLING)) { Header …

Top-1 retrieved context:
* You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
    [AC]X_ACCESS)) { Header …

Why?

New dimension in improving LMs!

Update & scale without additional training

Provide data attribution

Summary

slide 96 ↓

Test input:
include '../lib/admin.defines.php';
include '../lib/admin.module.access.php';
include '../lib/admin.smarty.php';
if (! has_right (

Continuation: **[AC]**X_BILLING)) { Header …

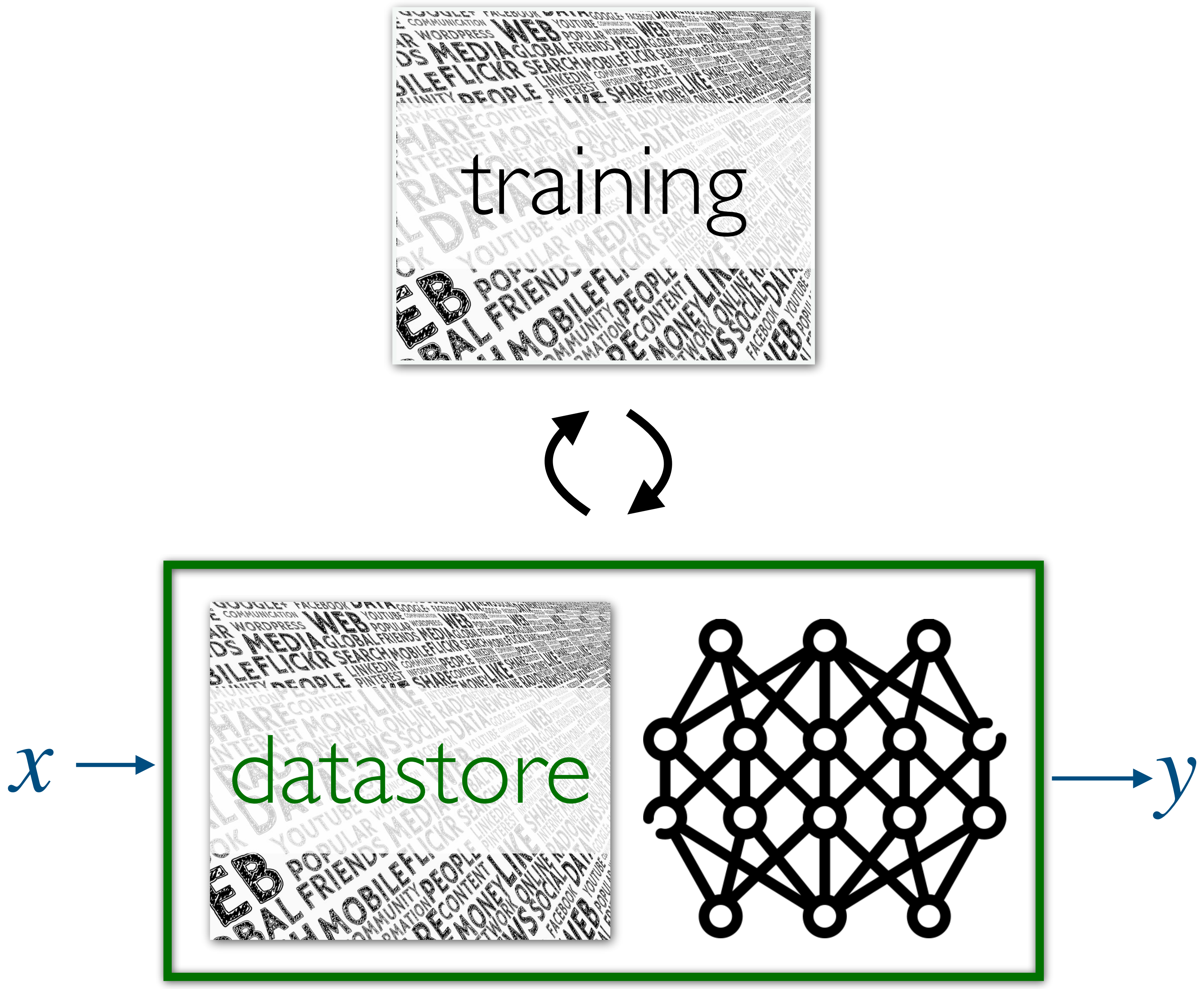
Top-1 retrieved context:
* You should have received a copy of the GNU Affero General Public License
* along with this program. If not, see <http://www.gnu.org/licenses/>.
*
*
**/
if (! has_right (
[AC]X_ACCESS)) { Header …

Why?

- New dimension in improving LMs!
- Update & scale without additional training
- Provide data attribution
- New opportunities in fairness & legality

Summary

What?



x: test input
y: model prediction to x

How?

Retrieval augmentation

New Transformers

Nonparametric Softmax

Why?

New dimension in improving LMs!

Update & scale without additional training

Provide data attribution

New opportunities in fairness & legality

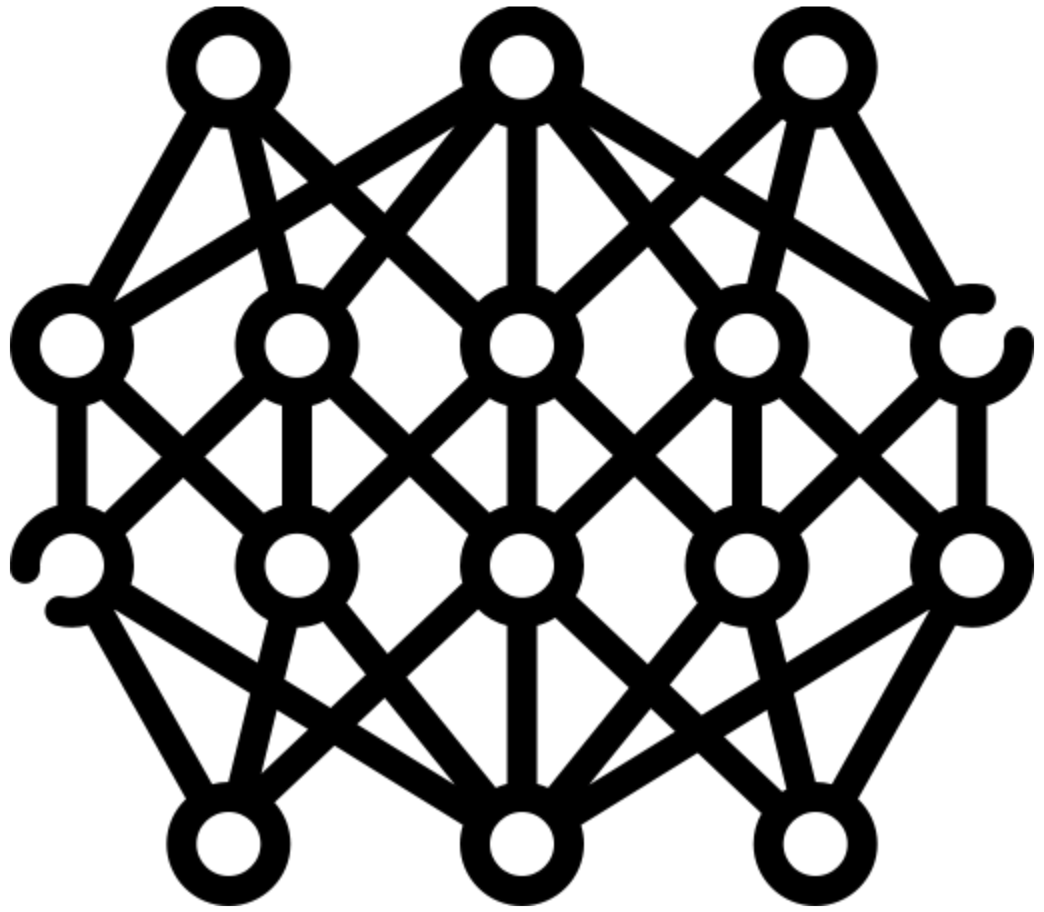
Open questions

Open question: Scaling retrieval-based LMs

Open question: Scaling retrieval-based LMs

A small LM + a large datastore ≈ a large (no-retrieval) LM?

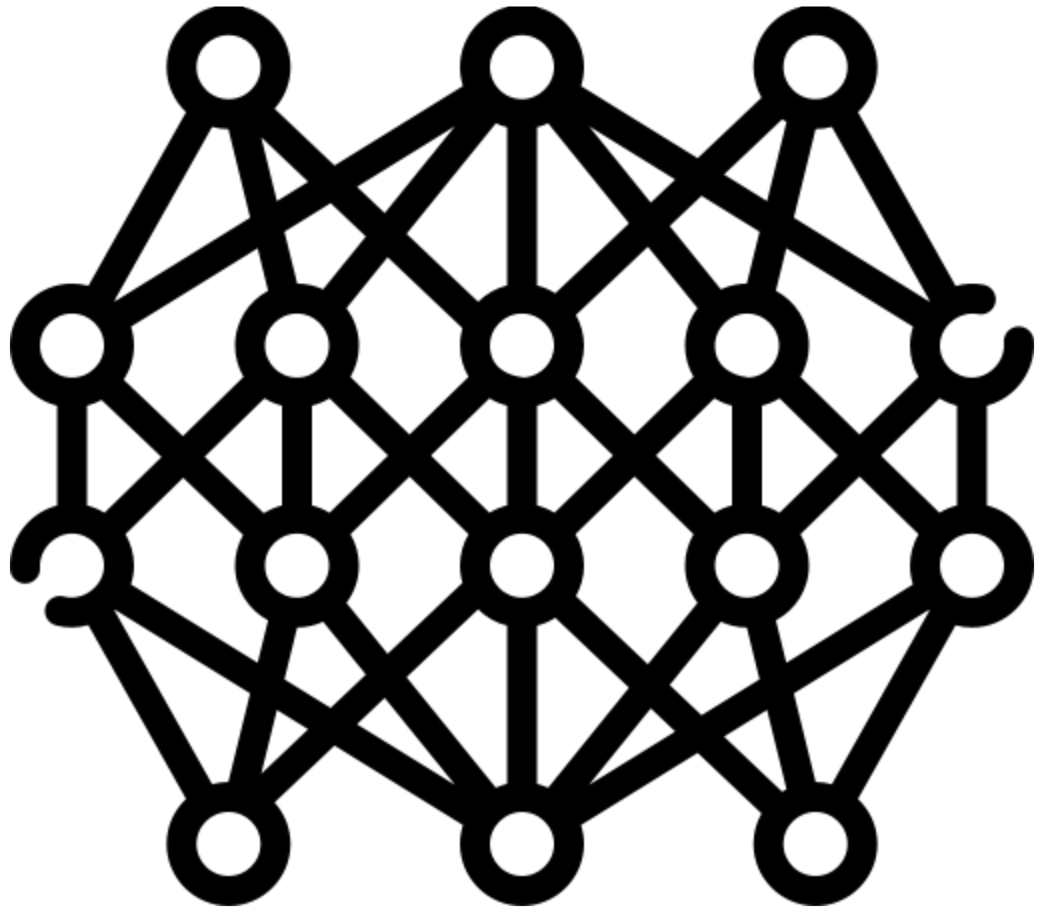
vs.



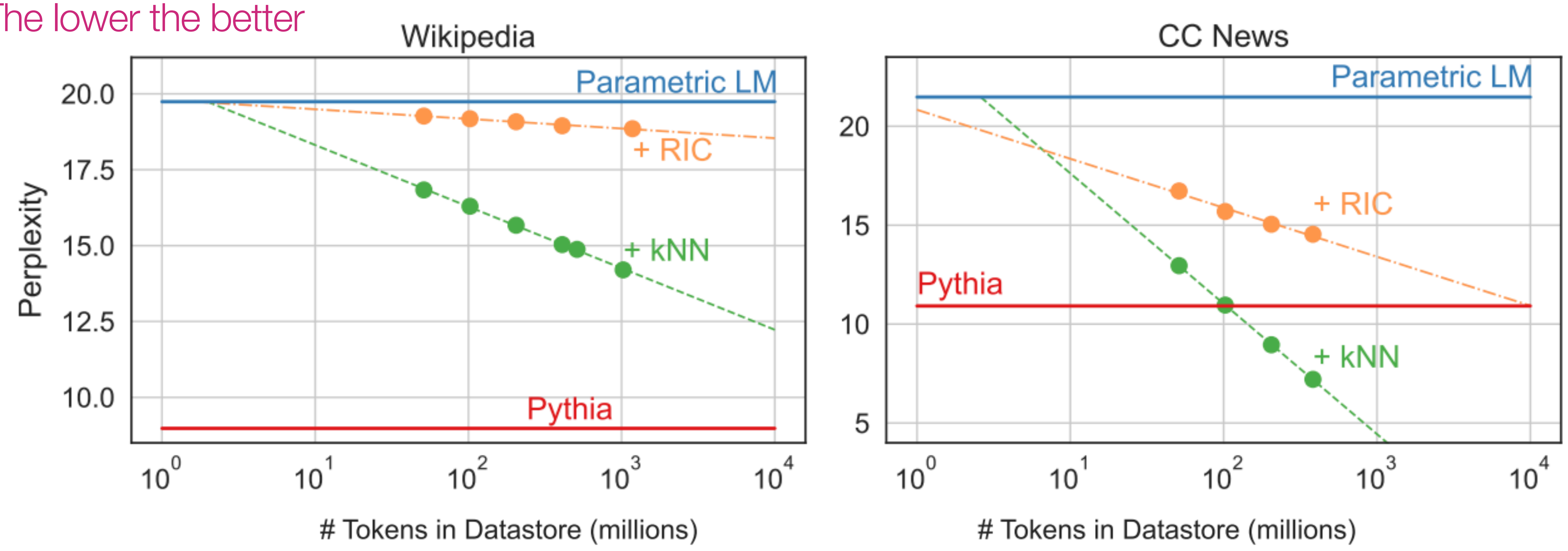
Open question: Scaling retrieval-based LMs

A small LM + a large datastore ≈ a large (no-retrieval) LM?

vs.



The lower the better



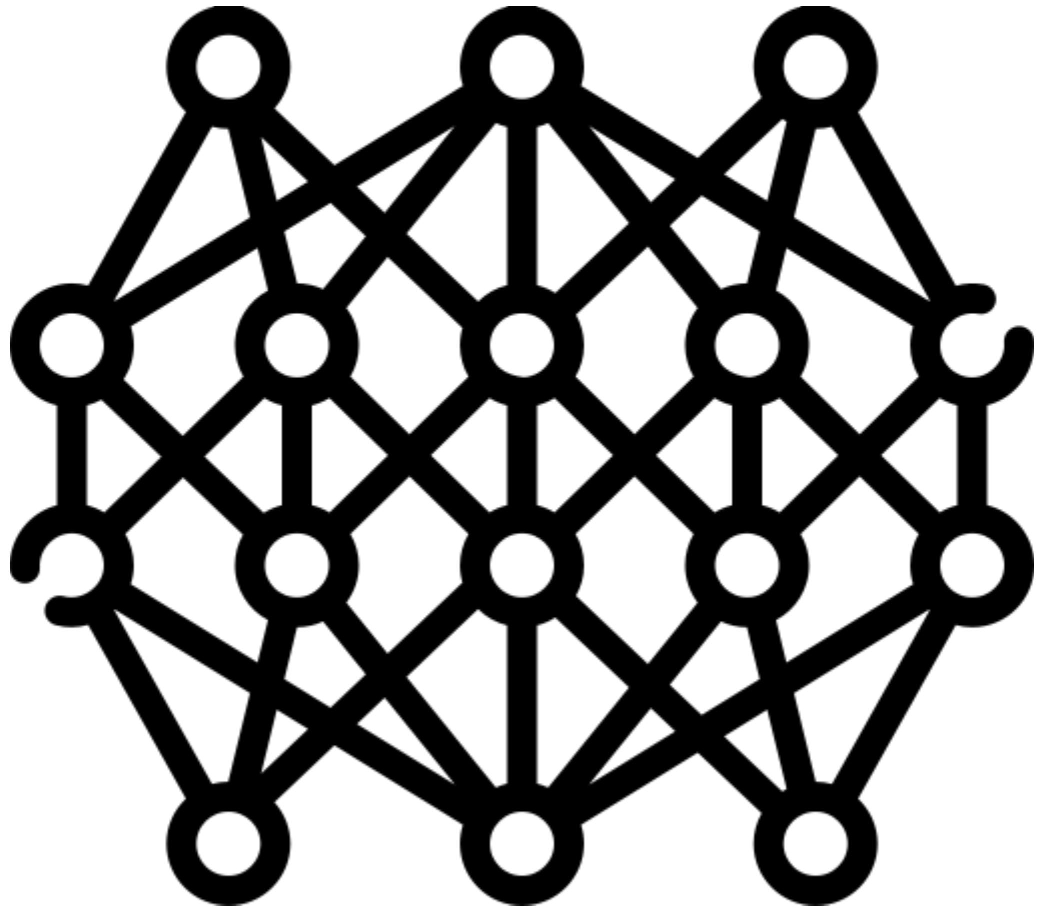
A new dimension in scaling!

Min et al. 2023. "SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore"

Open question: Scaling retrieval-based LMs

A small LM + a large datastore ≈ a large (no-retrieval) LM?

vs.

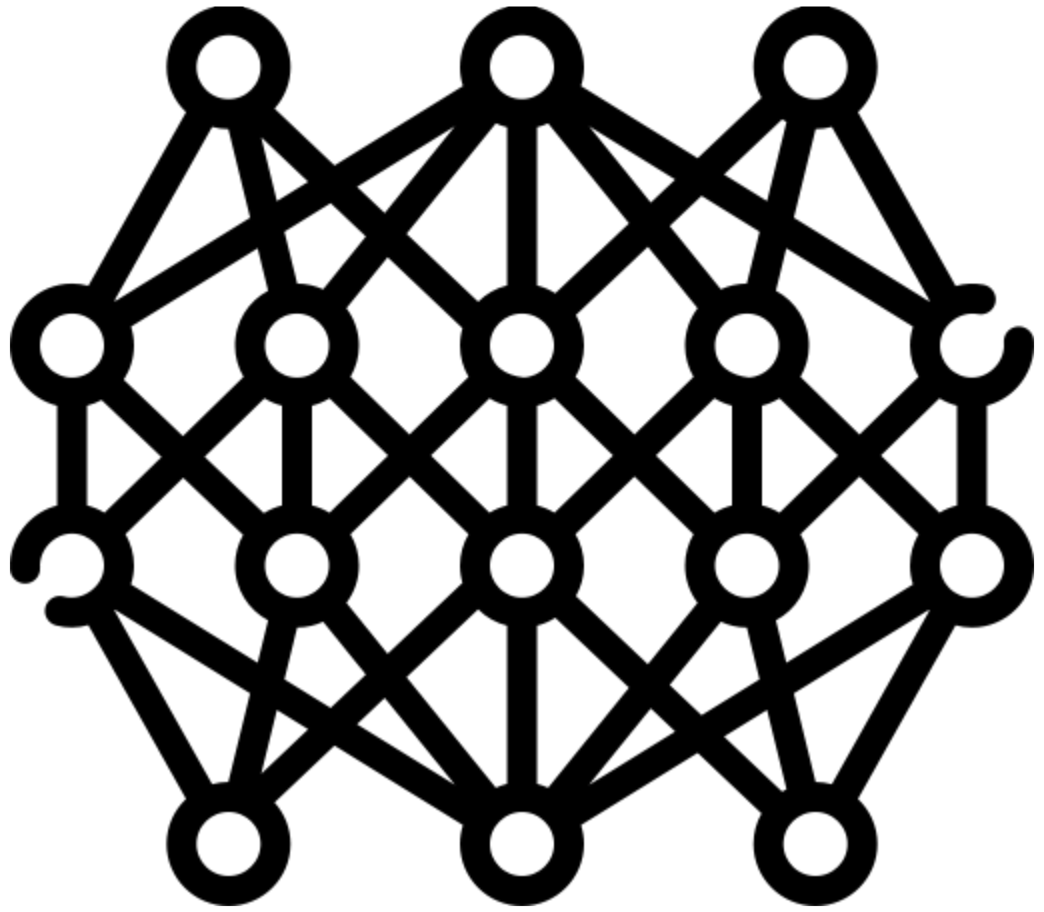


	LM	Datastore
	# of parameters	# of tokens
kNN-LM (Khandelwal et al., 2020)	250M	≤ 3B
NPM (Min et al., 2023)	350M	1B
Atlas (Izacard et al., 2022)	11B	~30B
RETRO (Borgeaud et al., 2021)	7B	2T
REPLUG (Shi et al., 2023)	≤ 175B	~5B

Open question: Scaling retrieval-based LMs

A small LM + a large datastore ≈ a large (no-retrieval) LM?

vs.



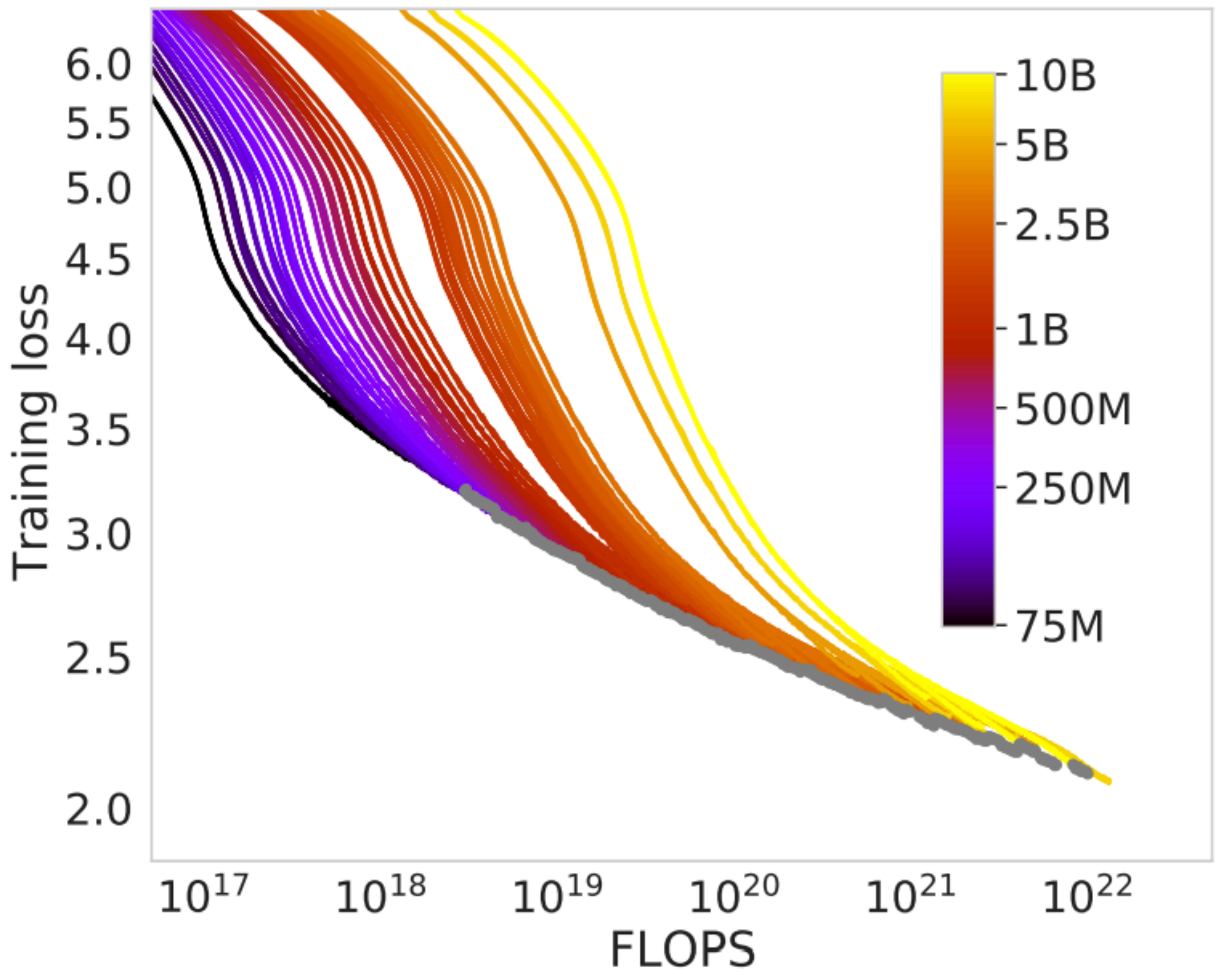
	LM	Datastore
	# of parameters	# of tokens
kNN-LM (Khandelwal et al., 2020)	250M	$\leq$ 3B
NPM (Min et al., 2023)	350M	1B
Atlas (Izacard et al., 2022)	11B	~30B
RETRO (Borgeaud et al., 2021)	7B	2T
REPLUG (Shi et al., 2023)	$\leq$ 175B	~5B

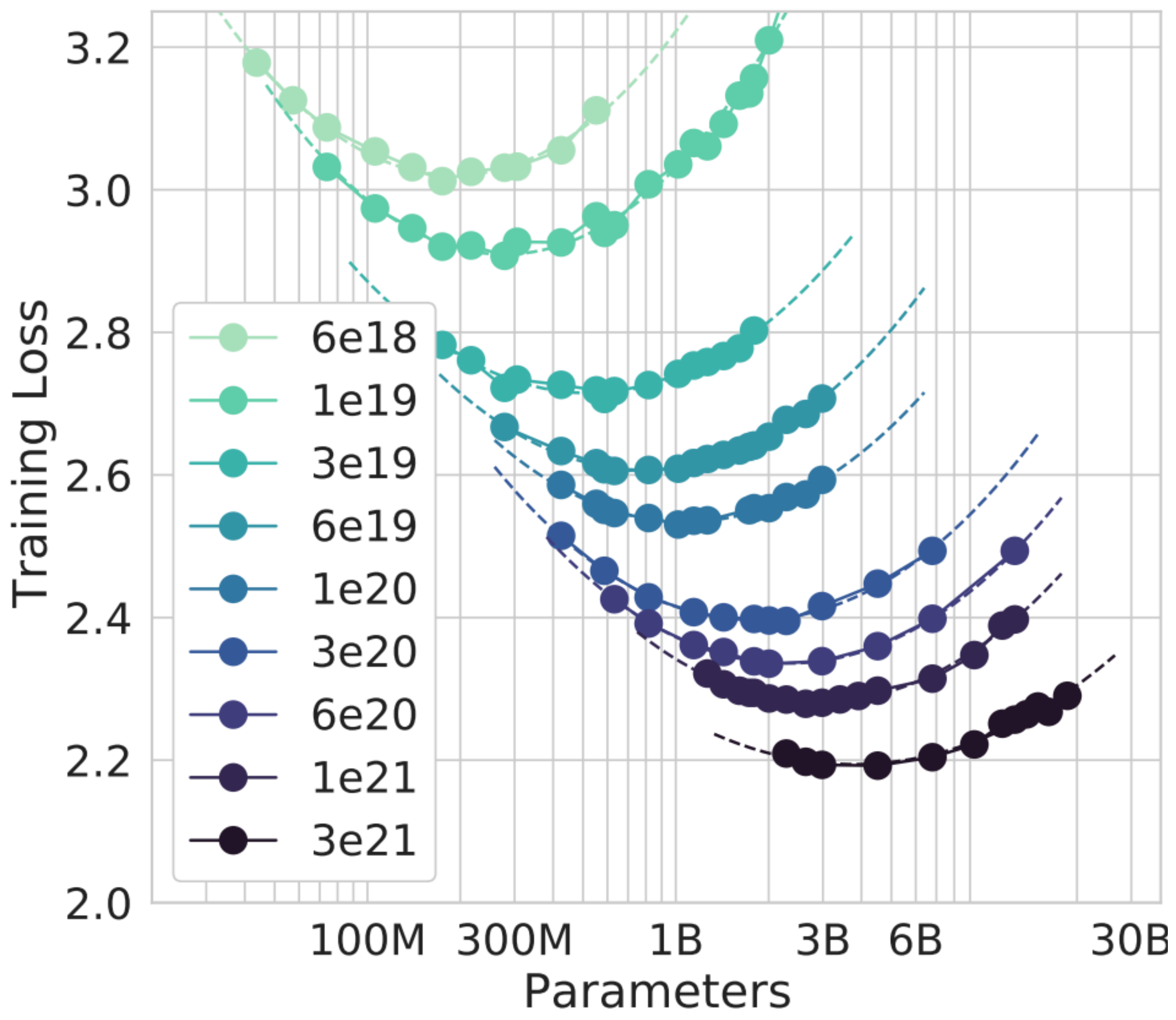
Open question: Scaling retrieval-based LMs

Scaling law?

Open question: Scaling retrieval-based LMs

Scaling law?





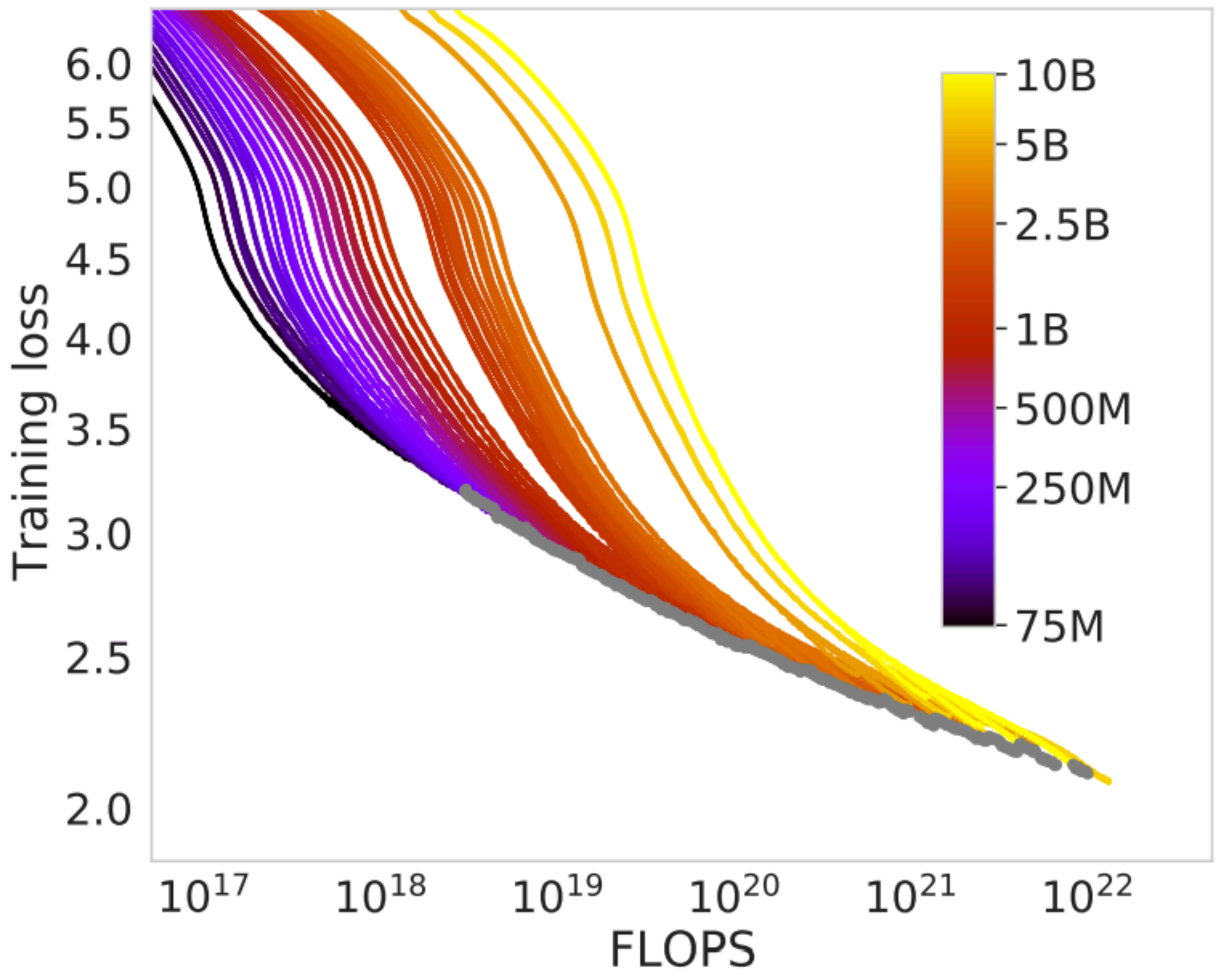
Loss as a function of:

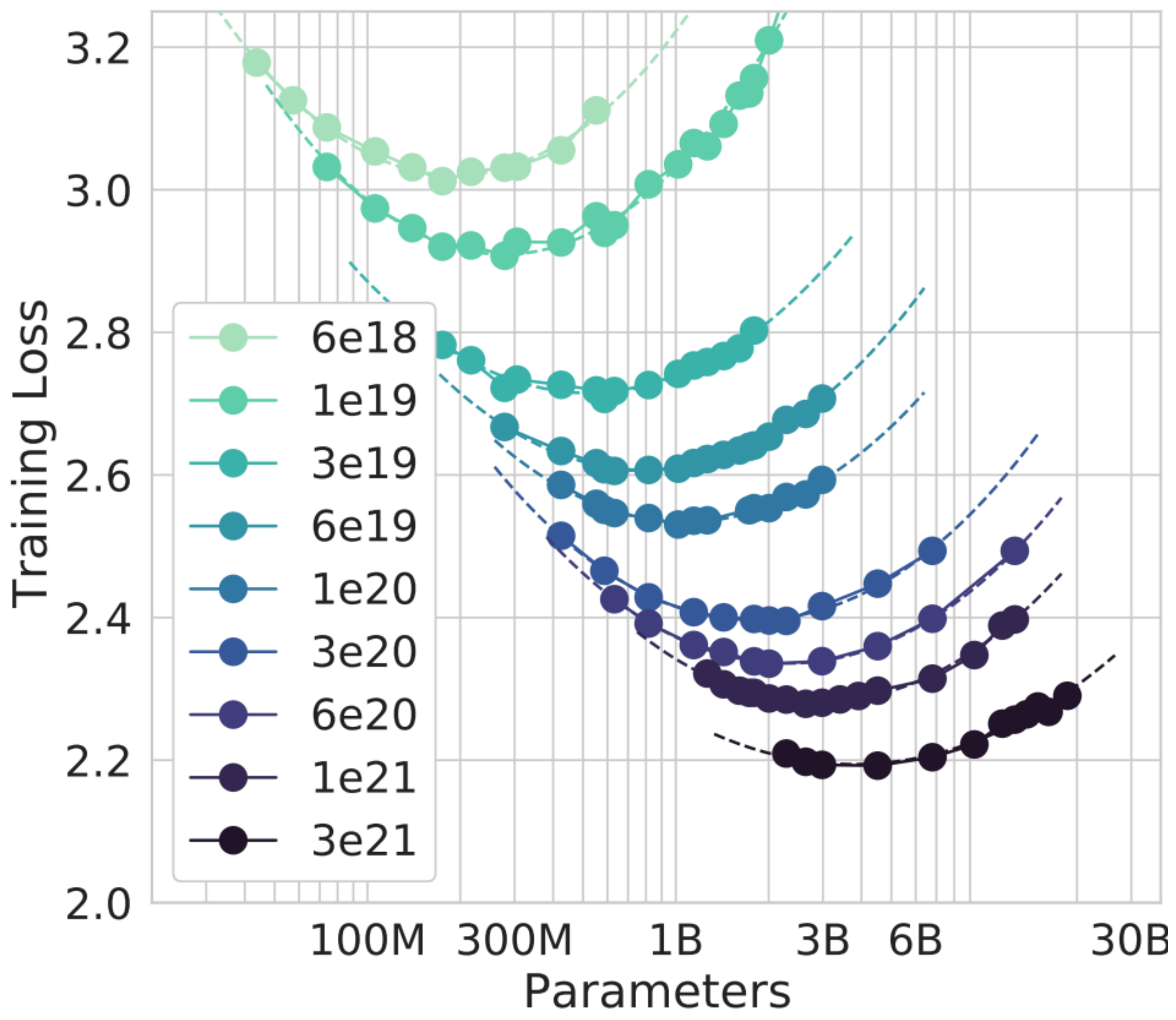
- Training data size
- # model parameters

Scaling law for parametric LMs (Kalpan et al., 2020; Hoffman et al., 2022)

Open question: Scaling retrieval-based LMs

Scaling law?





Loss as a function of:

- Training data size
- # model parameters

+ Datastore sizes?

Scaling law for parametric LMs (Kalpan et al., 2020; Hoffman et al., 2022)

Open question: Runtime efficiency

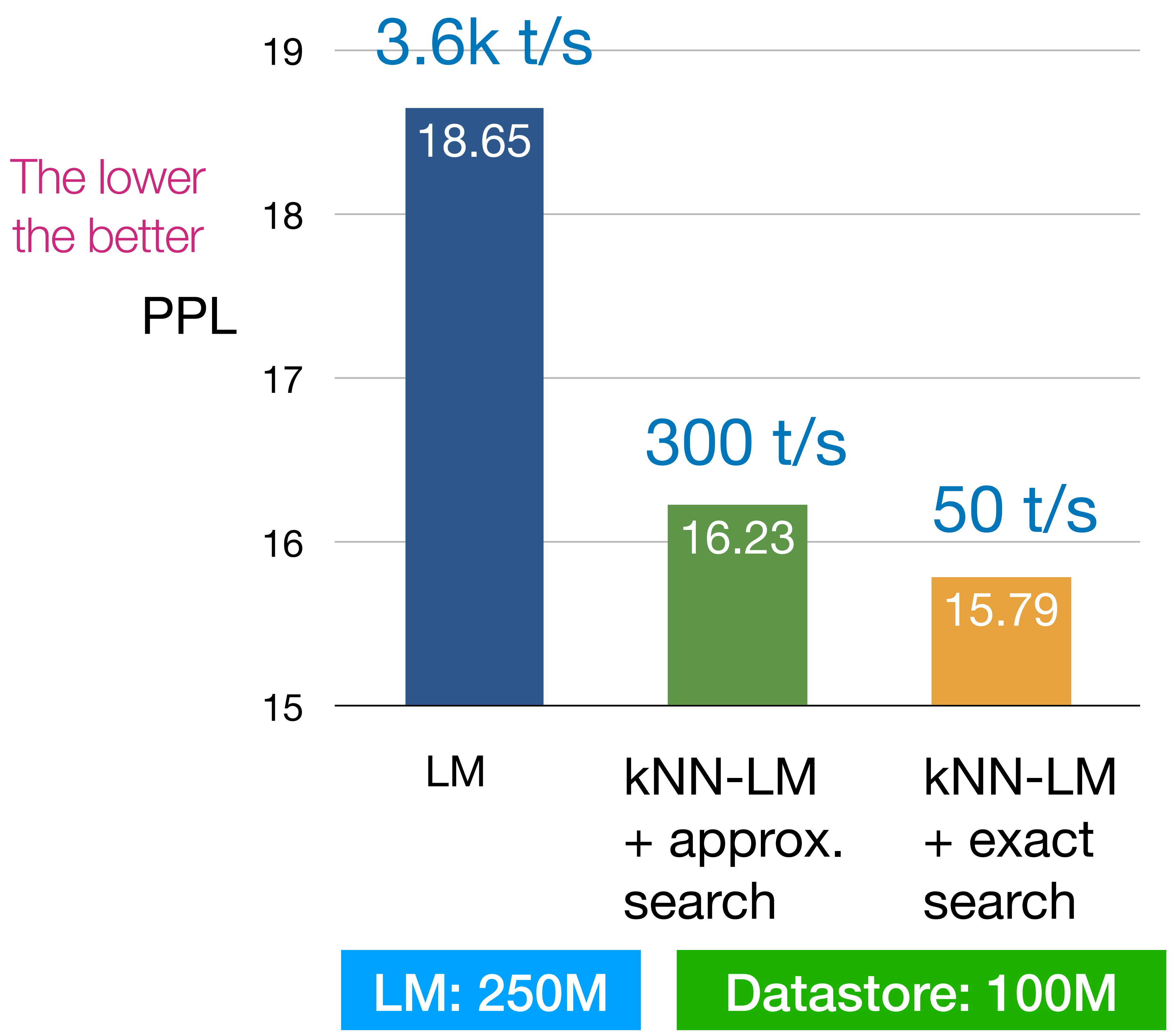
Efficiency of similarity search

Guo et al. 2020. "Accelerating Large-Scale Inference with Anisotropic Vector Quantization"

Open question: Runtime efficiency

Efficiency of similarity search

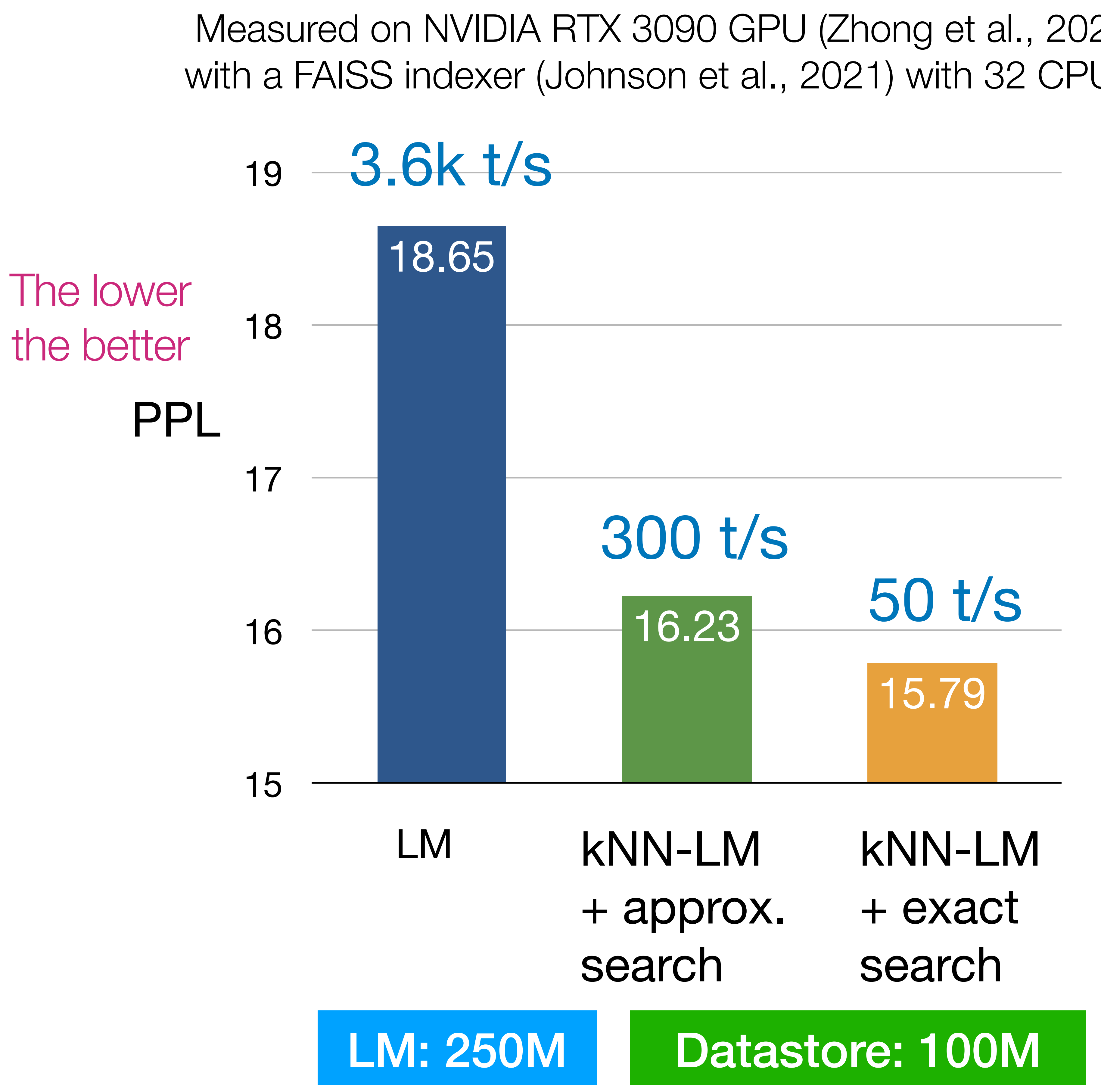
Measured on NVIDIA RTX 3090 GPU (Zhong et al., 2022)
with a FAISS indexer (Johnson et al., 2021) with 32 CPUs



Guo et al. 2020. "Accelerating Large-Scale Inference with Anisotropic Vector Quantization"

Open question: Runtime efficiency

Efficiency of similarity search



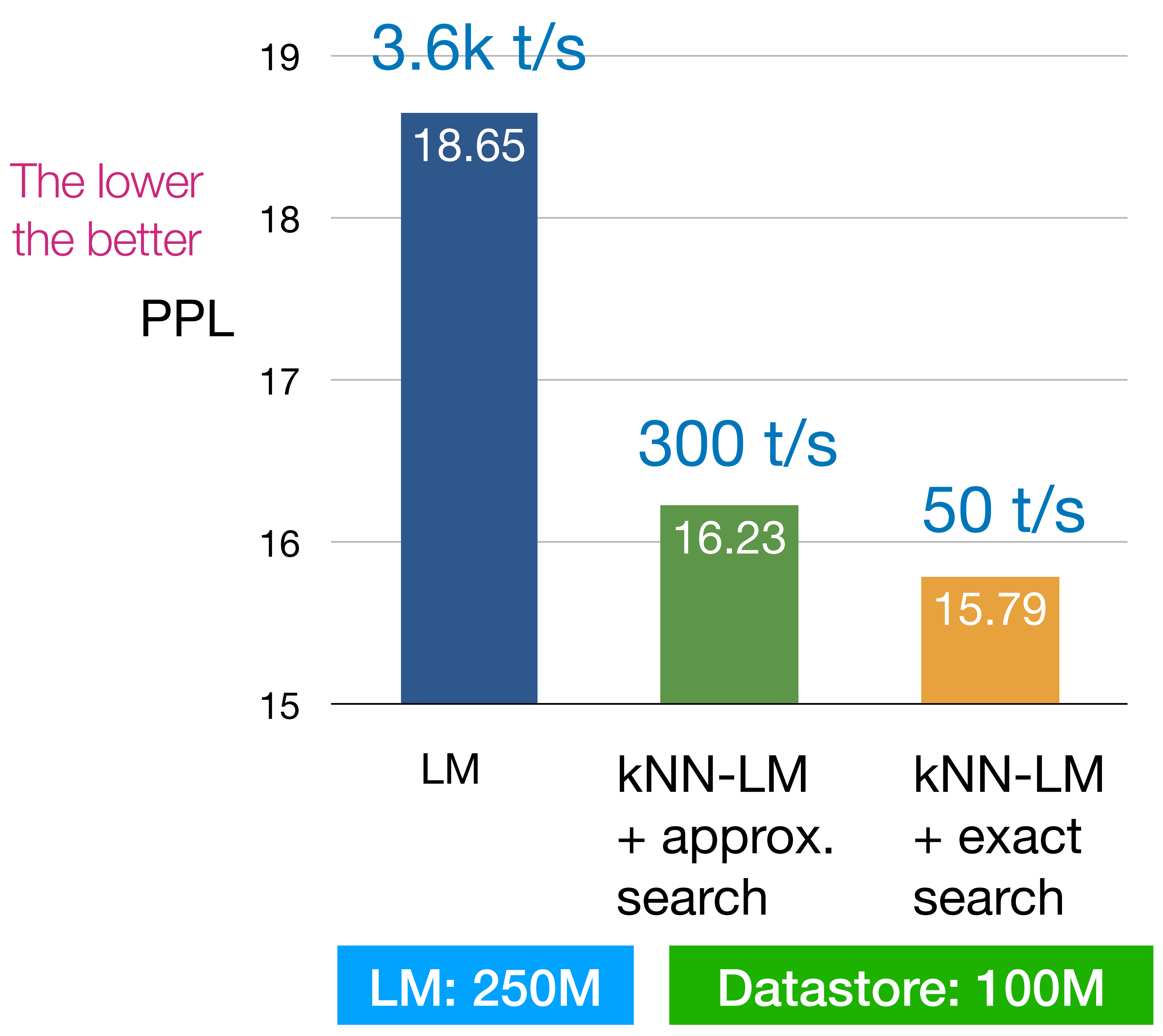
- >12 times slower with **approximate** nearest neighbor search

Guo et al. 2020. "Accelerating Large-Scale Inference with Anisotropic Vector Quantization"

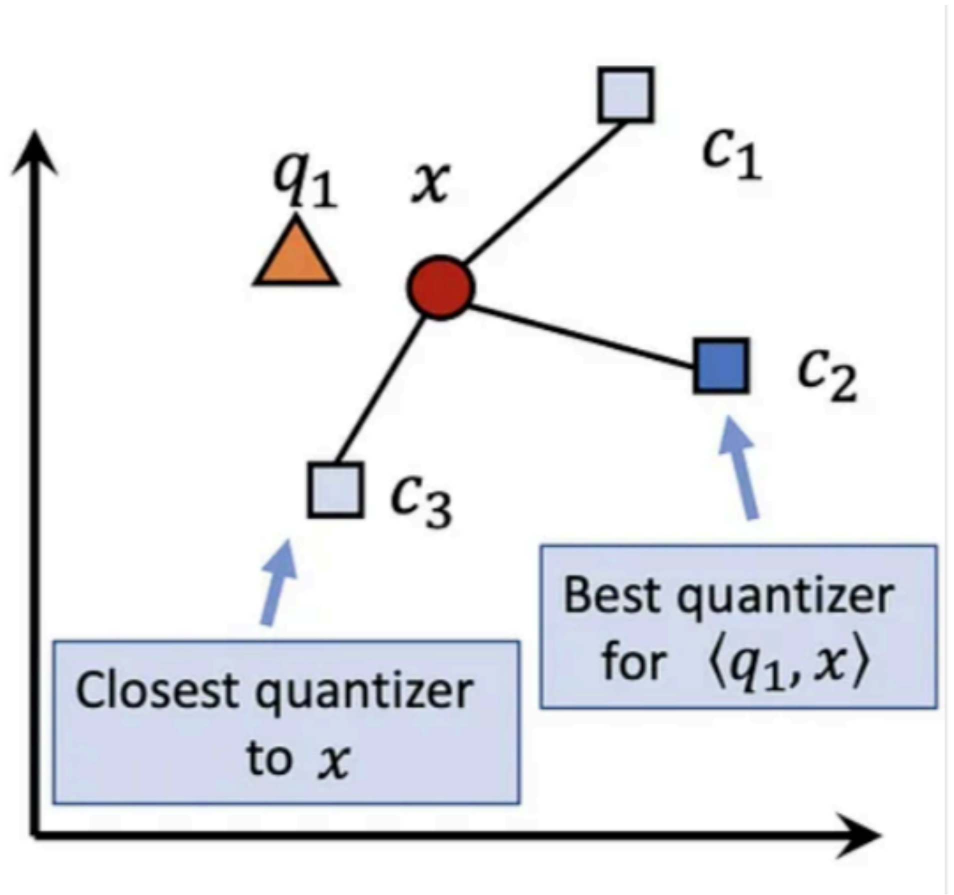
Open question: Runtime efficiency

Efficiency of similarity search

Measured on NVIDIA RTX 3090 GPU (Zhong et al., 2022)
with a FAISS indexer (Johnson et al., 2021) with 32 CPUs



- >12 times slower with **approximate** nearest neighbor search
- Efficient similarity search is an active research area (in conjunction with **systems**, **databases**, & **algorithms**)

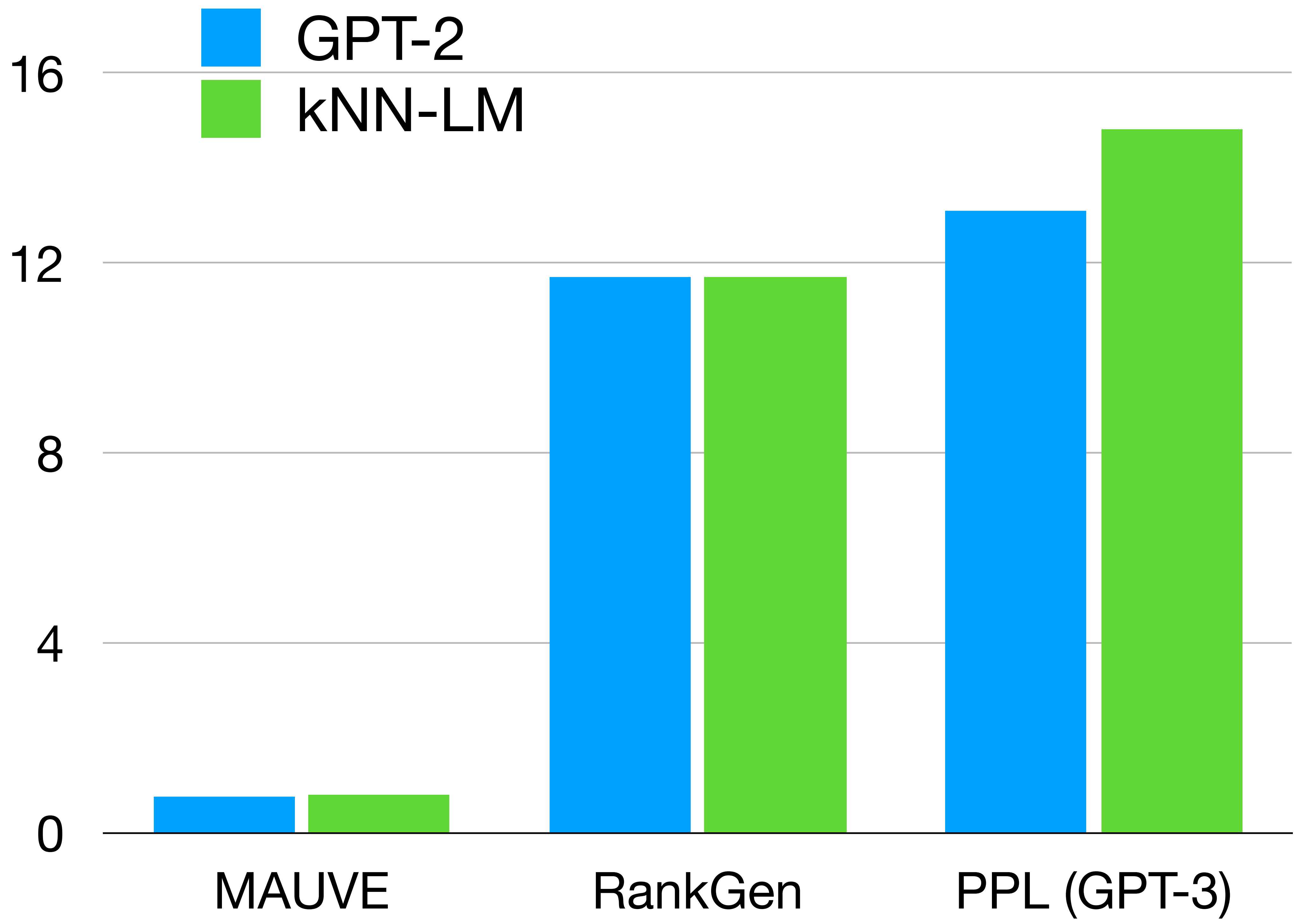


Guo et al. 2020. "Accelerating Large-Scale Inference with Anisotropic Vector Quantization"

Open question: Retrieval-based LMs for applications

Open question: Retrieval-based LMs for applications

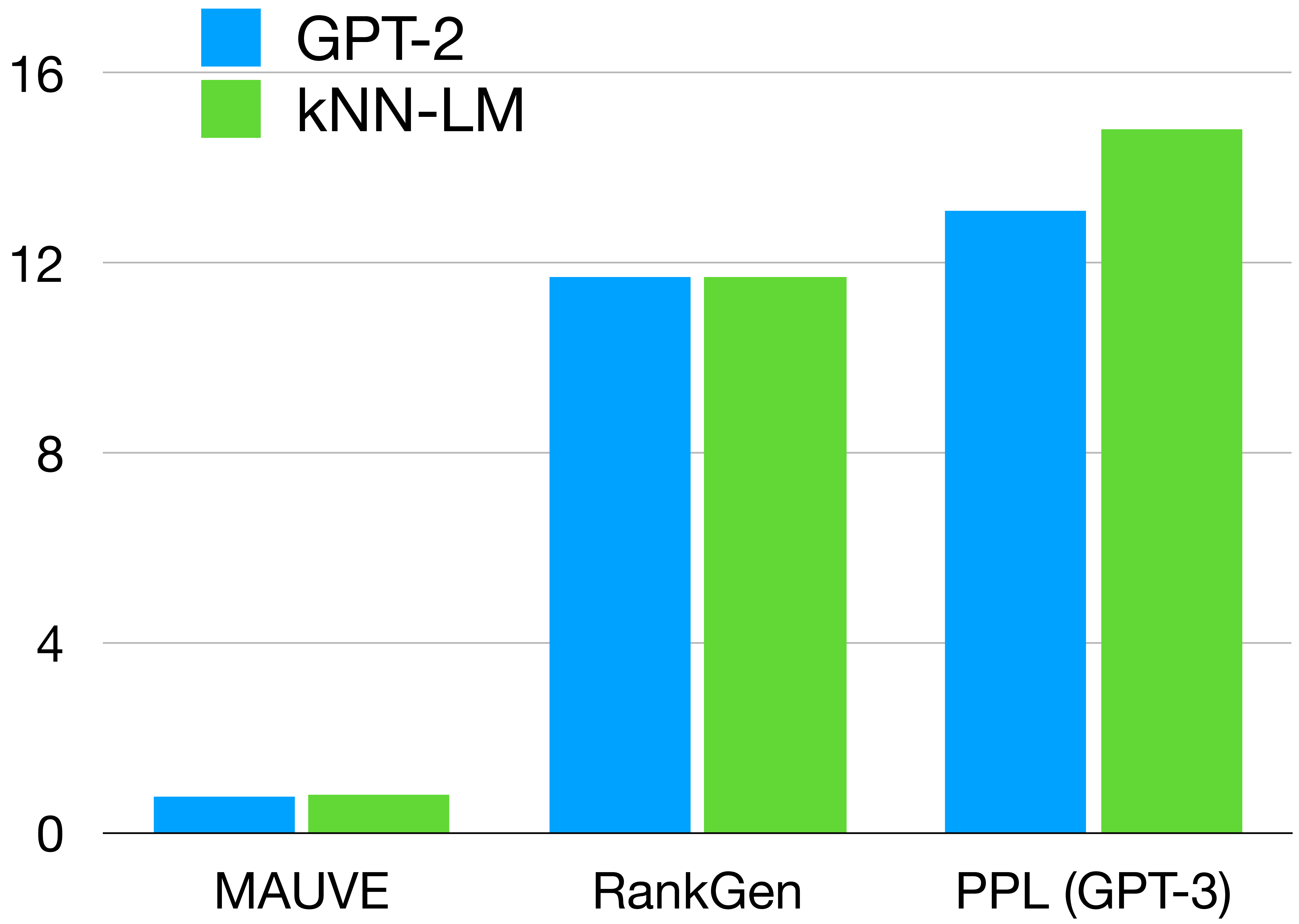
Open-ended text generation?



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

Open question: Retrieval-based LMs for applications

Open-ended text generation?



Better decoding algorithms? Better adaptation methods?

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

Open questions: Summary

Open questions: Summary

- What is the best **architecture & training method** for retrieval-based LMs in practice?

Open questions: Summary

- What is the best **architecture & training method** for retrieval-based LMs in practice?

- How to **scale the datastore** to trillions of tokens?
  [Scaling law]

Open questions: Summary

- What is the best **architecture & training method** for retrieval-based LMs in practice?
- How to **scale the datastore** to trillions of tokens? [Scaling law]
- How to improve **runtime efficiency**?

Open questions: Summary

- What is the best **architecture & training method** for retrieval-based LMs in practice?
- How to **scale the datastore** to trillions of tokens?
  [Scaling law]
- How to improve **runtime efficiency**?
- How to design **new decoding** or **adaptation methods** for downstream tasks (e.g., open-ended text generation)!

Open questions: Summary

- What is the best **architecture & training method** for retrieval-based LMs in practice?
- How to **scale the datastore** to trillions of tokens? [Scaling law]
- How to improve **runtime efficiency**?
- How to design **new decoding** or **adaptation methods** for downstream tasks (e.g., open-ended text generation)!

Q & A

Thank you for listening!

Check out ACL 2023 Tutorial on this topic (3-hour): https://acl2023-retrieval-lm.github.io/

Please leave feedback at tinyurl.com/sewon-min-talk

Extra slides (from QnA)

Validating Model Output to be Factual

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

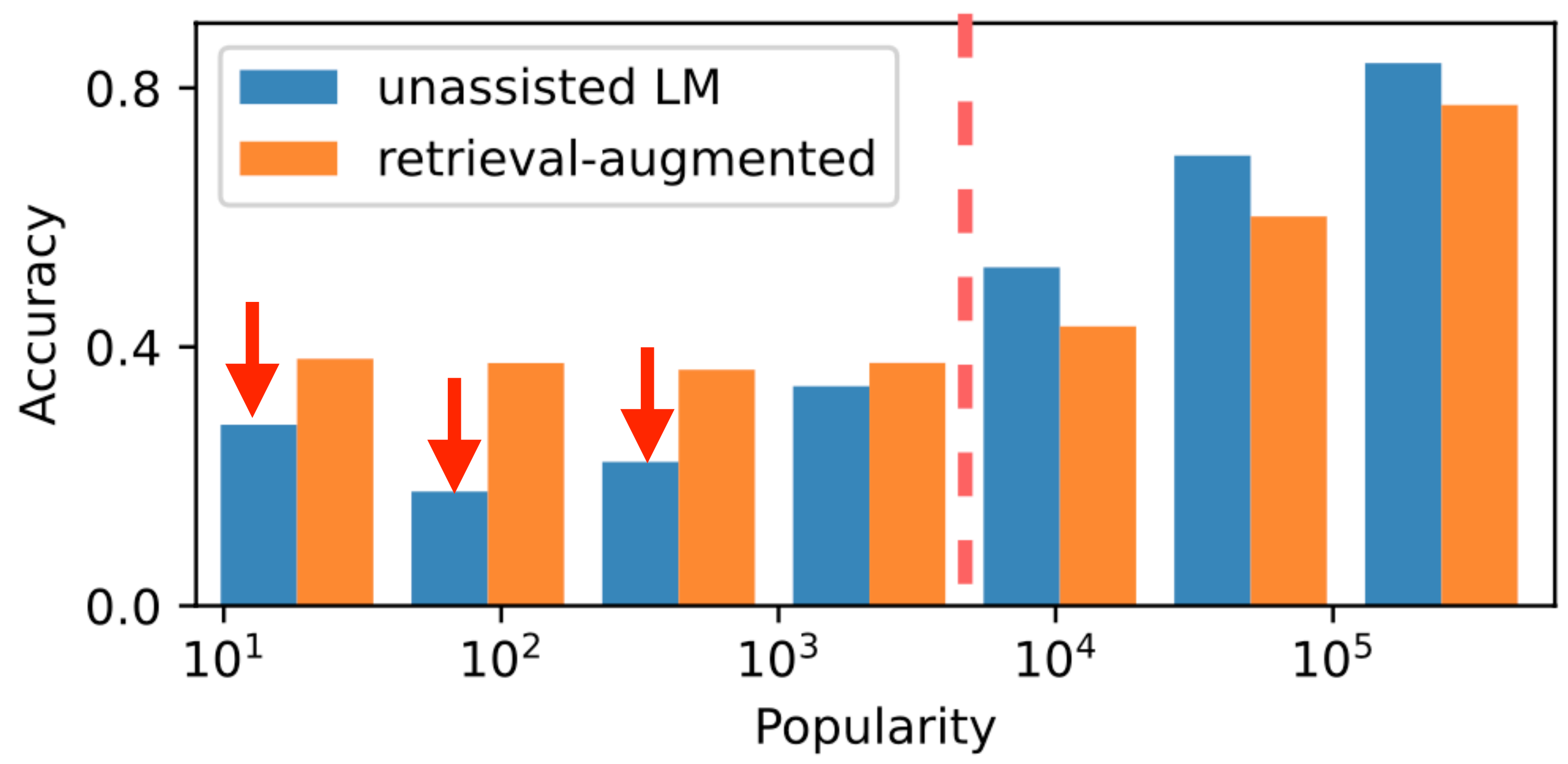
Atomic facts

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...

66.7%

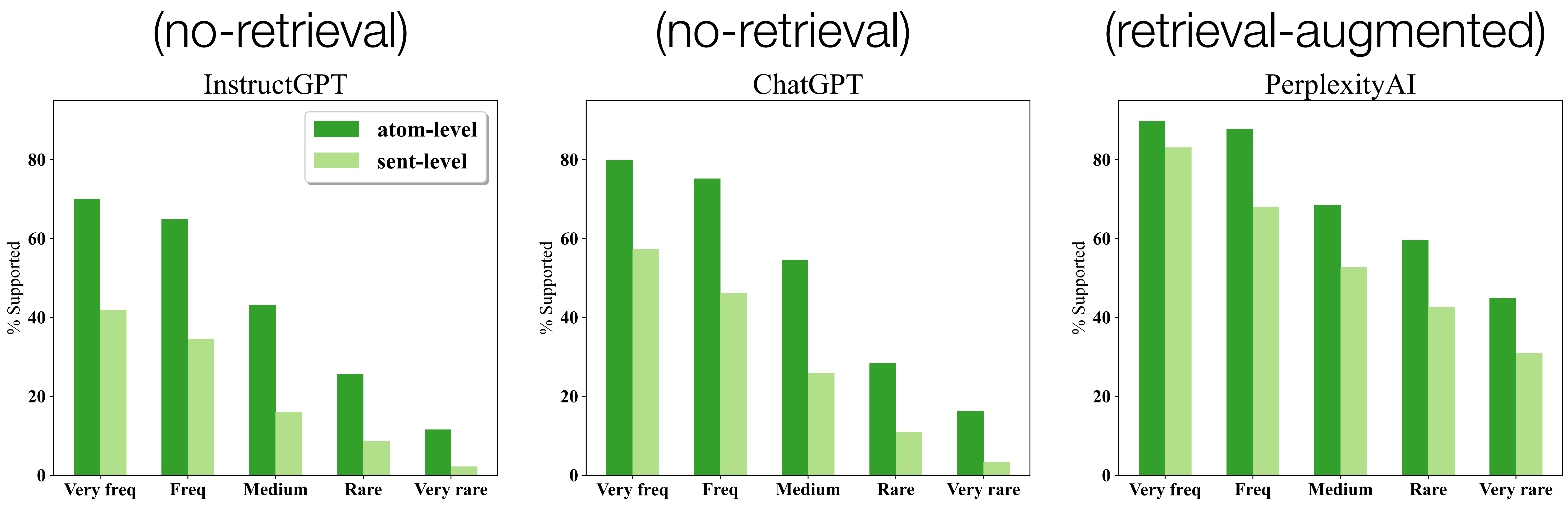
Min et al. 2023. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation"

Gains from retrieval w.r.t. frequency



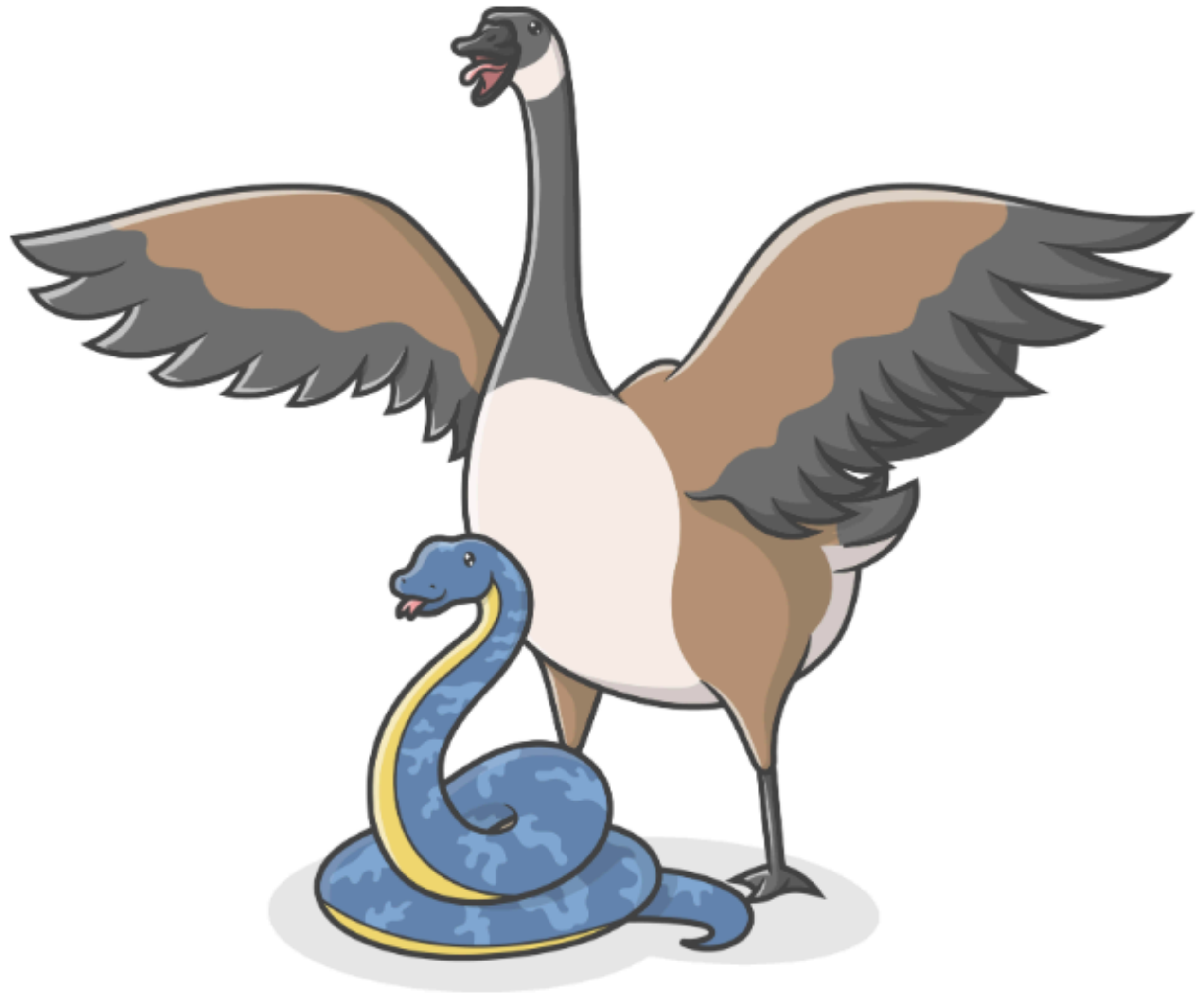
There has been mixed results about whether retrieval hurts when it comes to popular entities/facts, e.g., the top graph shows it does hurt in (short-form) question answering, and the bottom graph shows retrieval always help even with frequent entities in long-form text generation. These results are likely to depend on exact setup, e.g., the task, base LMs, and datastore, etc.

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



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

Research on information retrieval



Pyserini

pypi v0.22.1 | downloads 210k | downloads 1.8k/week | maven-central v0.22.1 | Lucene v9.5.0 | license Apache

Retrieval—including training the encoder, getting embeddings and indexing—is an active area of research. Recommend Pyserini (https://github.com/castorini/pyserini) for a set of references and also try some of them out easily.

State-of-the-art retrieval-based LMs?

- If you want the model that you can use right now — retrieval-augmentation
  - Partially because you can leverage the state-of-the-art models that industry built with no modification
  - You should use state-of-the-art retrieval (BM25, Contriever or GTR) and state-of-the-art LM (LLAMA, ChatGPT)
  - Easiest: with “independent training”, optionally with reranking
- Doesn’t mean retrieval-augmentation is the “best” under the scenario of fair comparison, e.g., when the model has exact same parameters & is trained on the exactly same data
  - The SILO paper shows kNN-LM (kNN in the graph) outperforms retrieval-augmentation (RiC in the graph), both when training data==datastore (right) and when training data!=datastore (left)
    - However, this is based on language modeling perplexity. Downstream task eval is still an open Q.