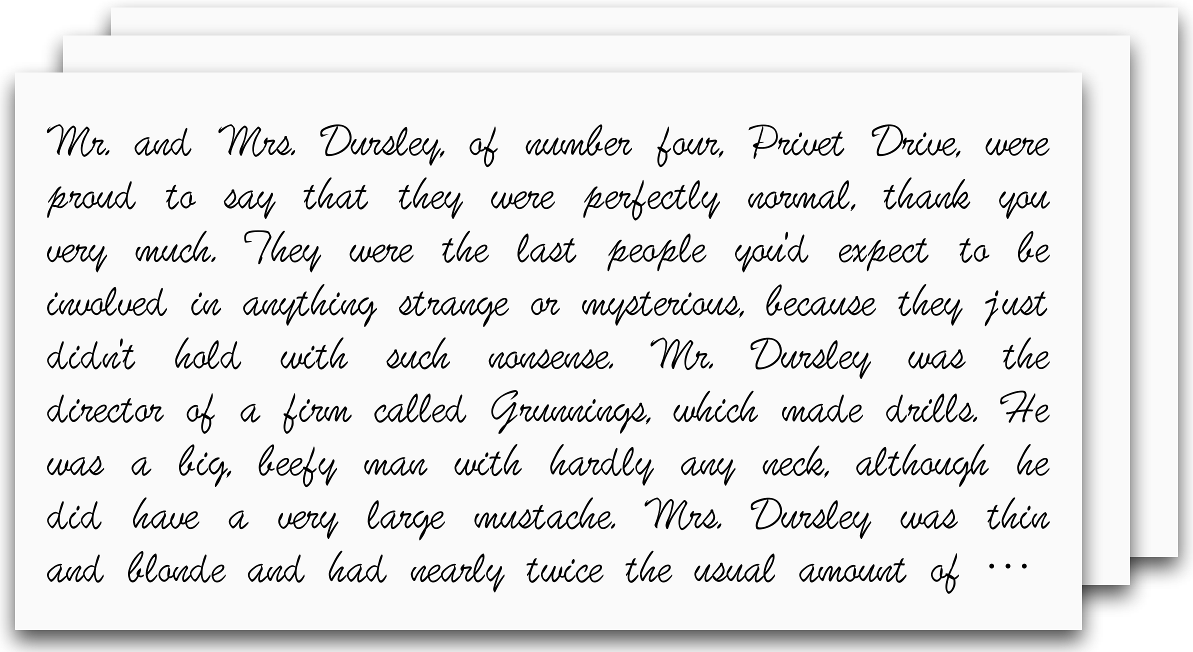
A Retrieval-based Language Model **at Scale**

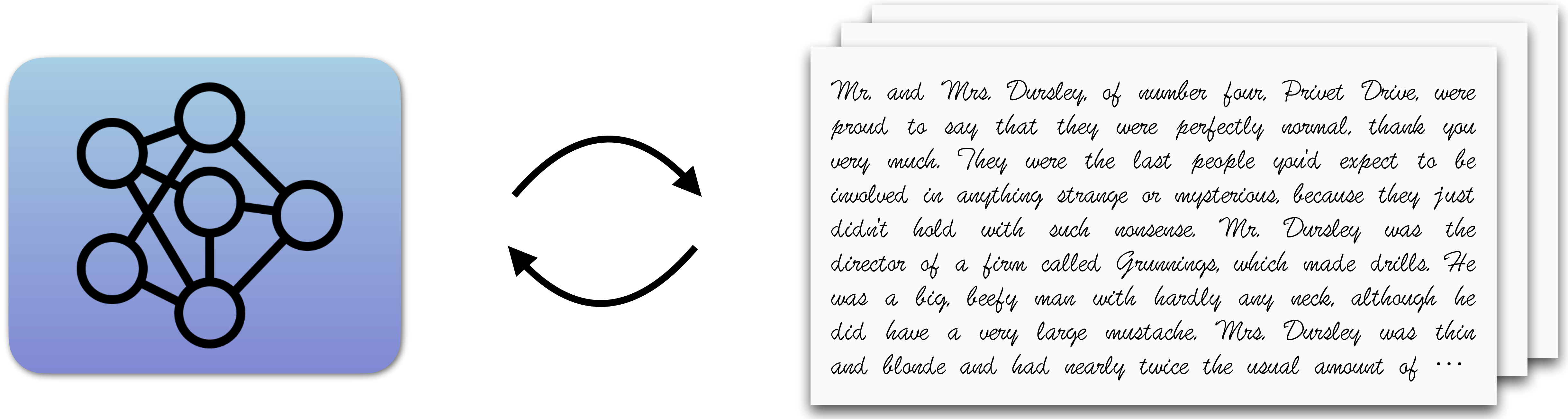
Sewon Min
sewonmin.com
Berkeley EECS BAIR Ai2

Today's LLM



Trillions of tokens

Today's LLM

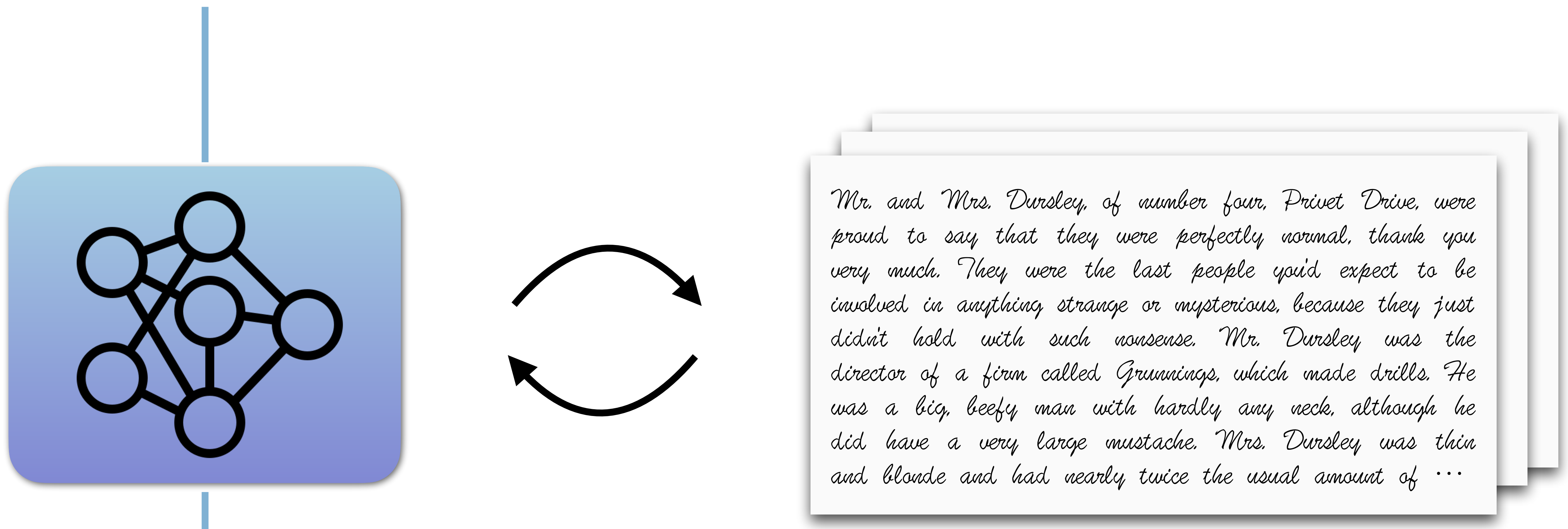


10+ billion parameters

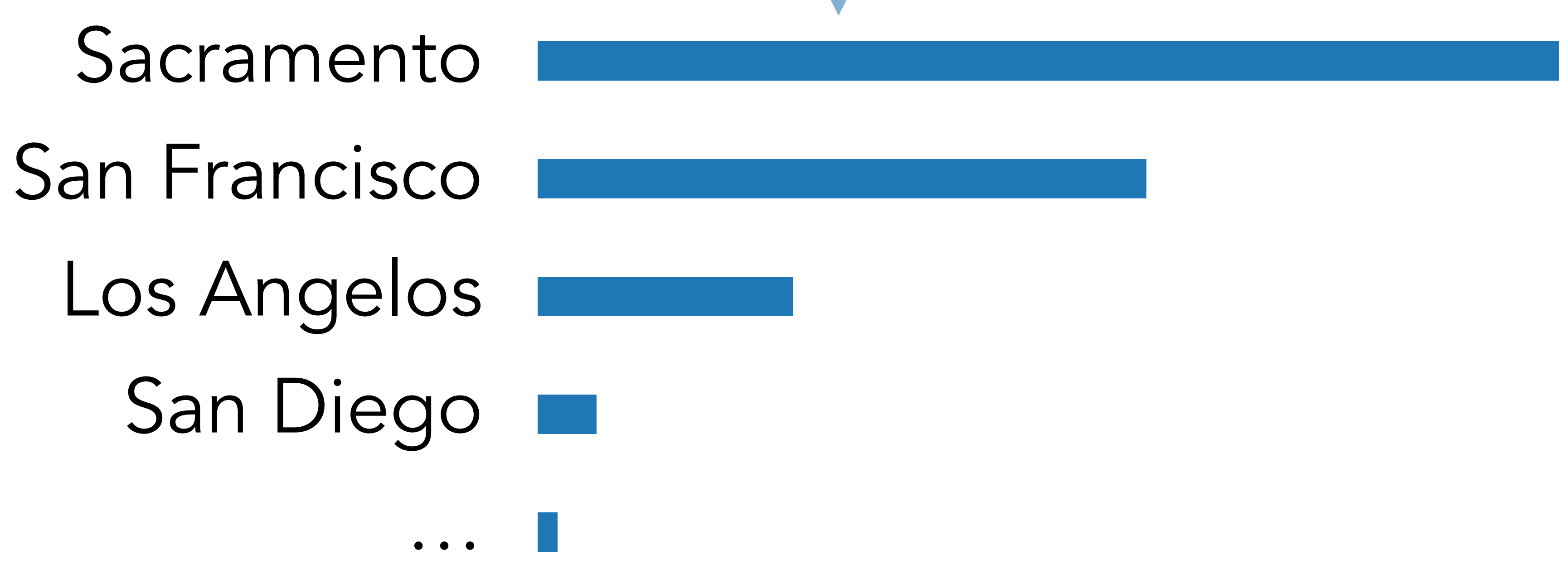
Trillions of tokens

Today's LLM

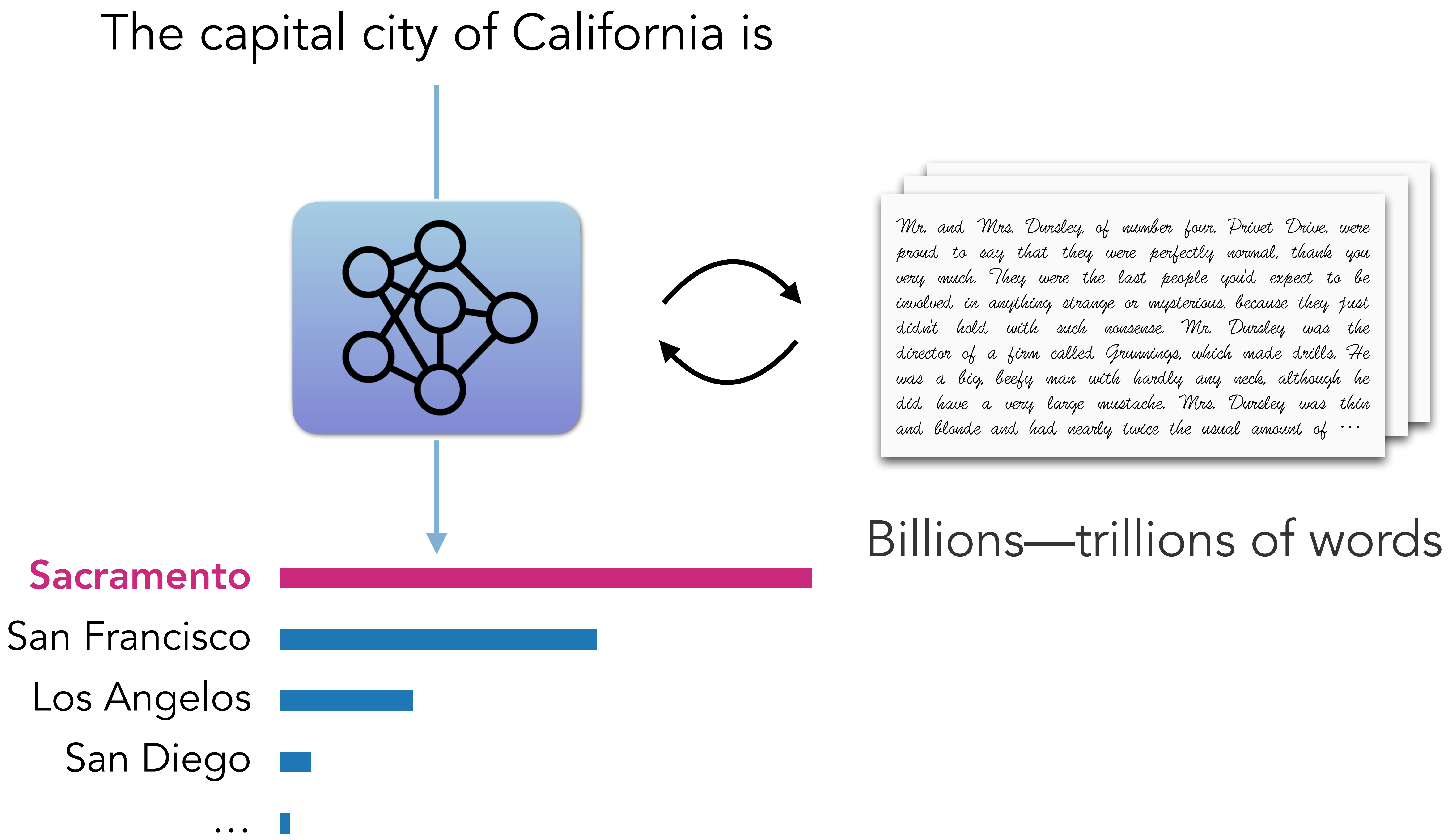
The capital city of California is



Billions—trillions of words

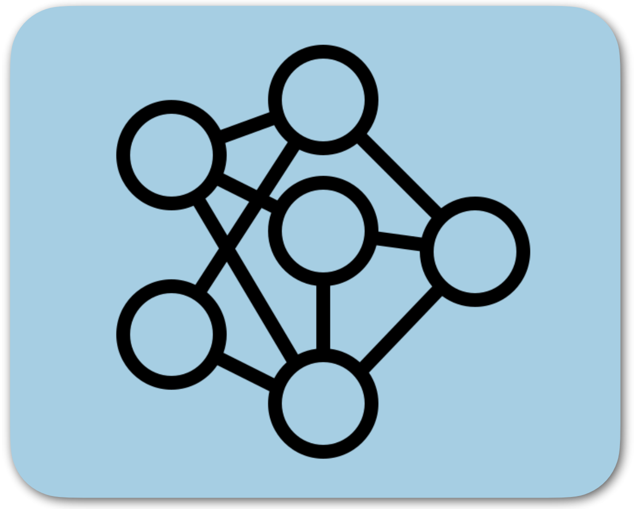


Today’s LLM



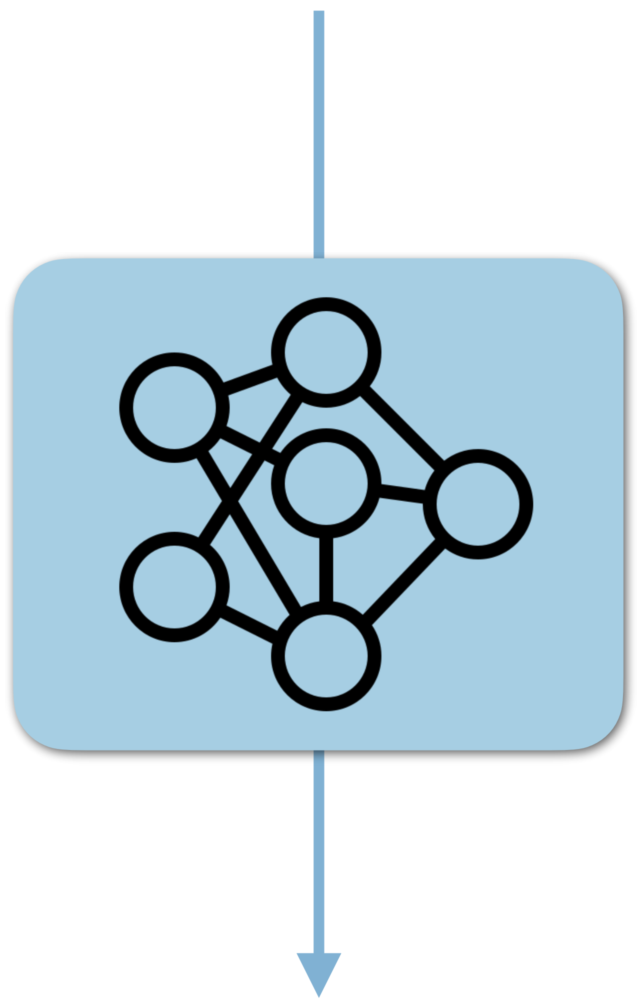
The capital city of California is
Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of …
Billions—trillions of words
Sacramento
San Francisco
Los Angelos
San Diego
…

Today's LLM



Today's LLM

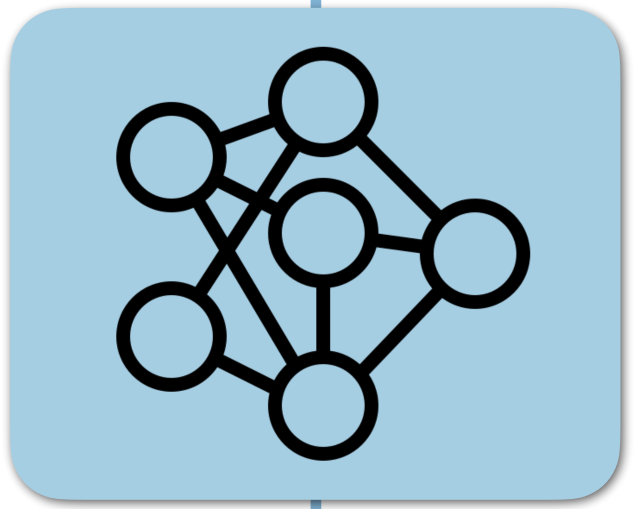
What's the capital city of California?



Sacramento

Today's LLM

What's the capital city of California?

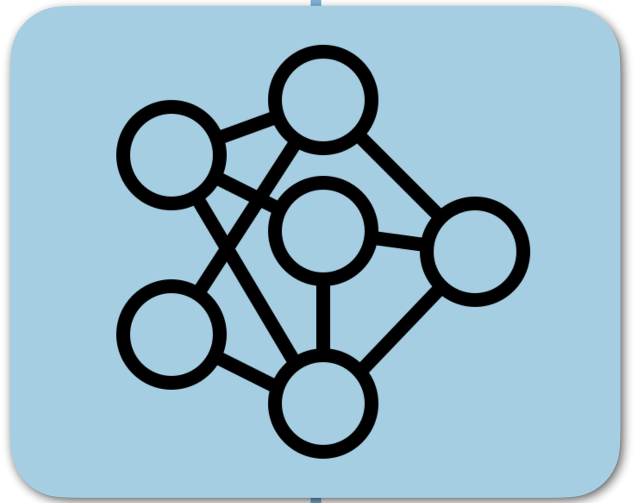


Sacramento

Works quite well in practice
(with sufficient scale)

Today's LLM

What's the capital city of California?



Sacramento

At the MIT event, Altman was asked if training GPT-4 cost $100 million; he replied, "It's more than that."

WIRED, April 17, 2023

Works quite well in practice

(with sufficient scale)

zuck 2h • Follow

Some updates on our A
vision is to build gener
source it responsibly, a
available so everyone c
bringing our two major
and GenAI) closer toge
We're currently training
Llama 3, and we're building massive compute
infrastructure to support our future roadmap,
including 350k H100s by the end of this year --
and overall almost 600k H100s equivalents of
compute if you include other GPUs. Also really
excited about our progress building new AI-
centric computing devices like Ray Ban Meta
smart glasses. Lots more to come soon.

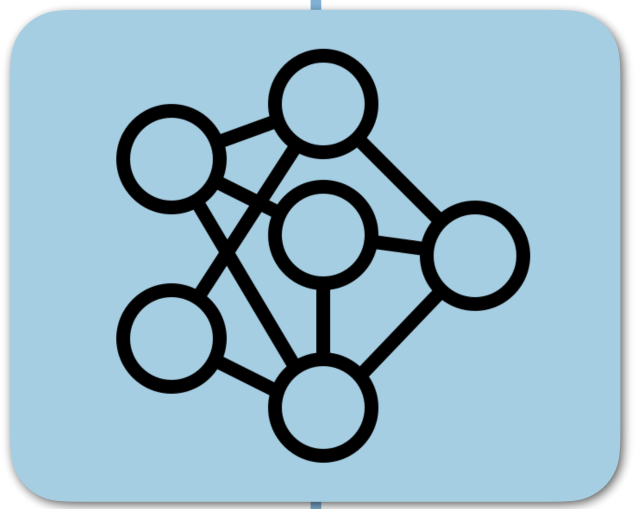
$24 billion

January 18, 2024

Today's LLM

What's the capital city of California?

Should remember every detail from the data



Sacramento

At the MIT event, Altman was asked if training GPT-4 cost $100 million; he replied, "It's more than that."

WIRED, April 17, 2023

Works quite well in practice
(with sufficient scale)

January 18, 2024

Today's LLM

ChatGPT

Tell me a bio of Sewon Min.

Sewon Min is a researcher in natural language processing (NLP) and machine learning, particularly known for her work in question answering, retrieval-augmented language models, and large-scale reasoning systems. She completed her **Ph.D. in ~~Computer Science at UC Berkeley~~**, ~~advised by **Dan Klein**~~. After her Ph.D., ~~she joined **Meta AI (FAIR)** as a researcher.~~ incorrect

[Without web browsing]

Today’s LLM

ChatGPT

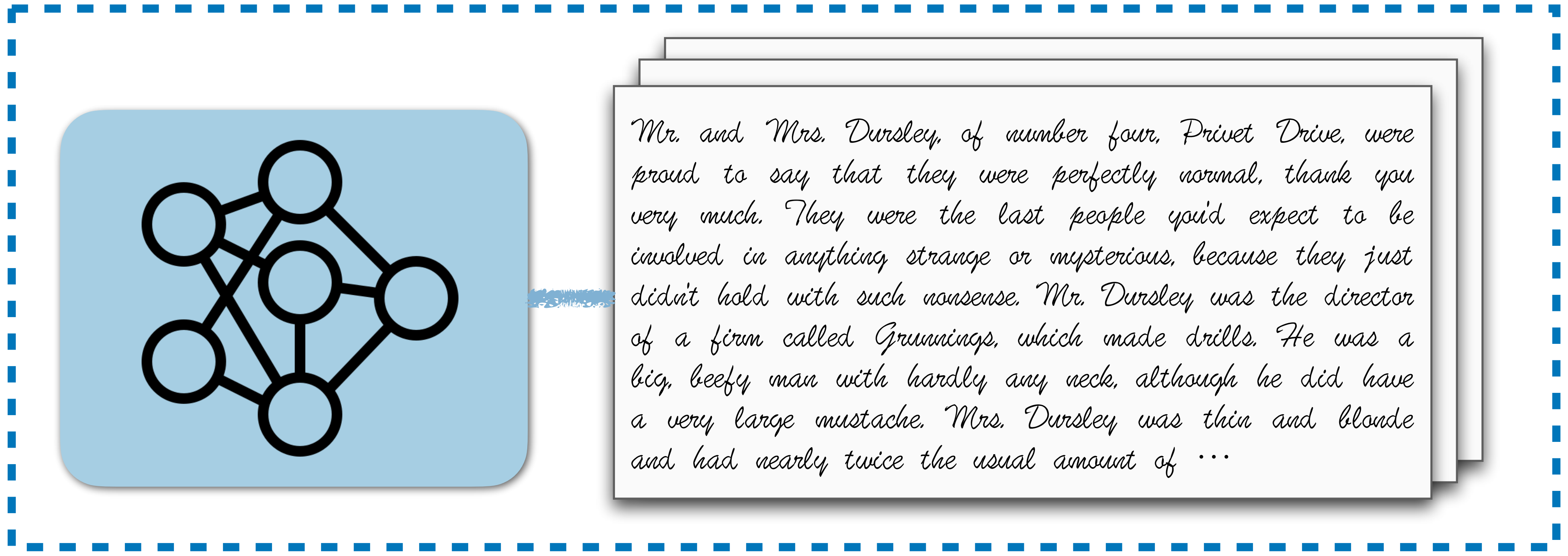
Tell me a bio of Sewon Min.

Sewon Min is a researcher in natural language processing (NLP) and machine learning, particularly known for her work in question answering, retrieval-augmented language models, and large-scale reasoning systems. She completed her **Ph.D. in ~~Computer Science at UC Berkeley~~**, ~~advised by **Dan Klein**~~. After her Ph.D., ~~she joined **Meta AI (FAIR)** as a researcher.~~ incorrect

[Without web browsing]

Expensive, still fail to remember details, fail to stay up-to-date

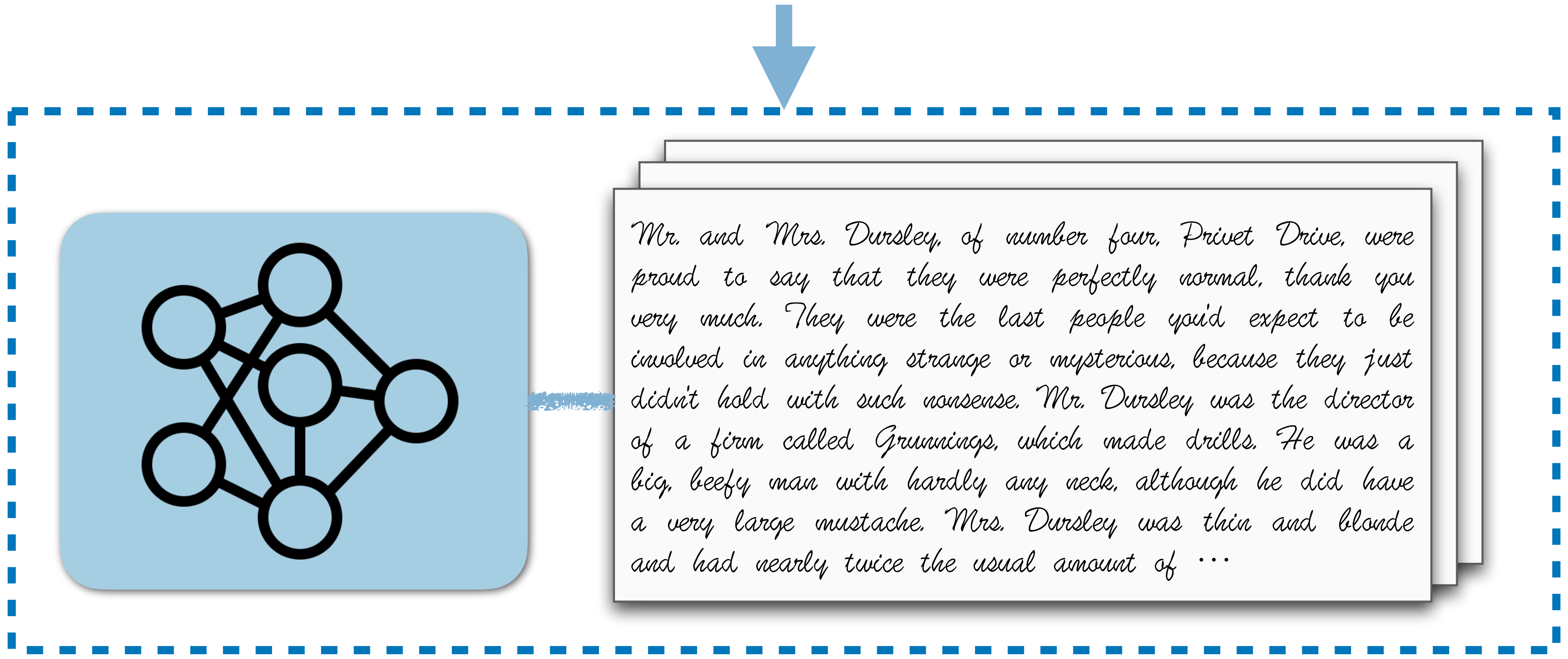
Keep the data!



data kept!

Keep the data!

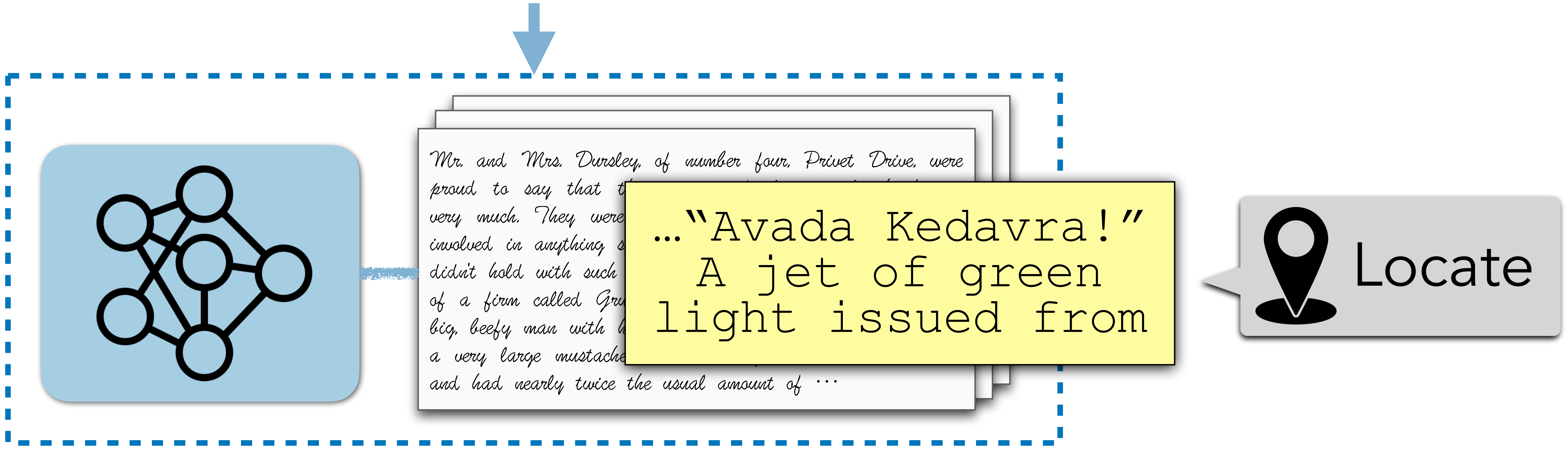
`Voldemort had raised his wand … and a flash of`



data kept!

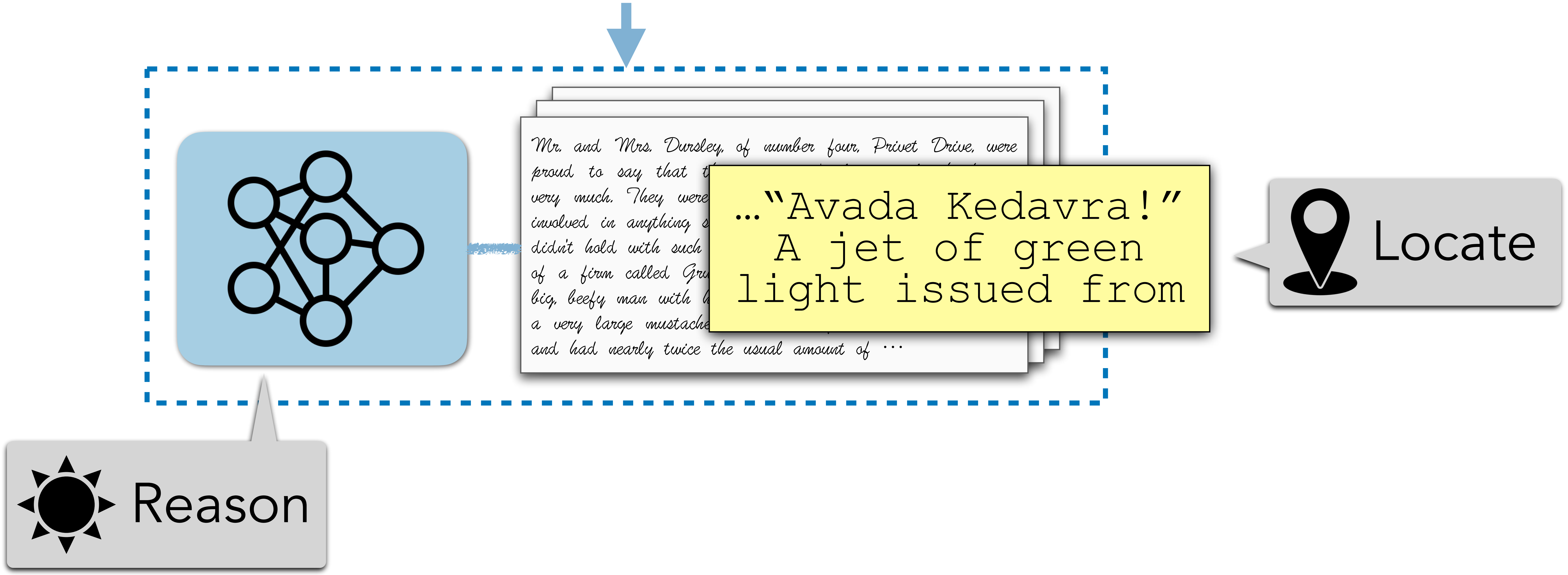
Keep the data!

Voldemort had raised his wand … and a flash of

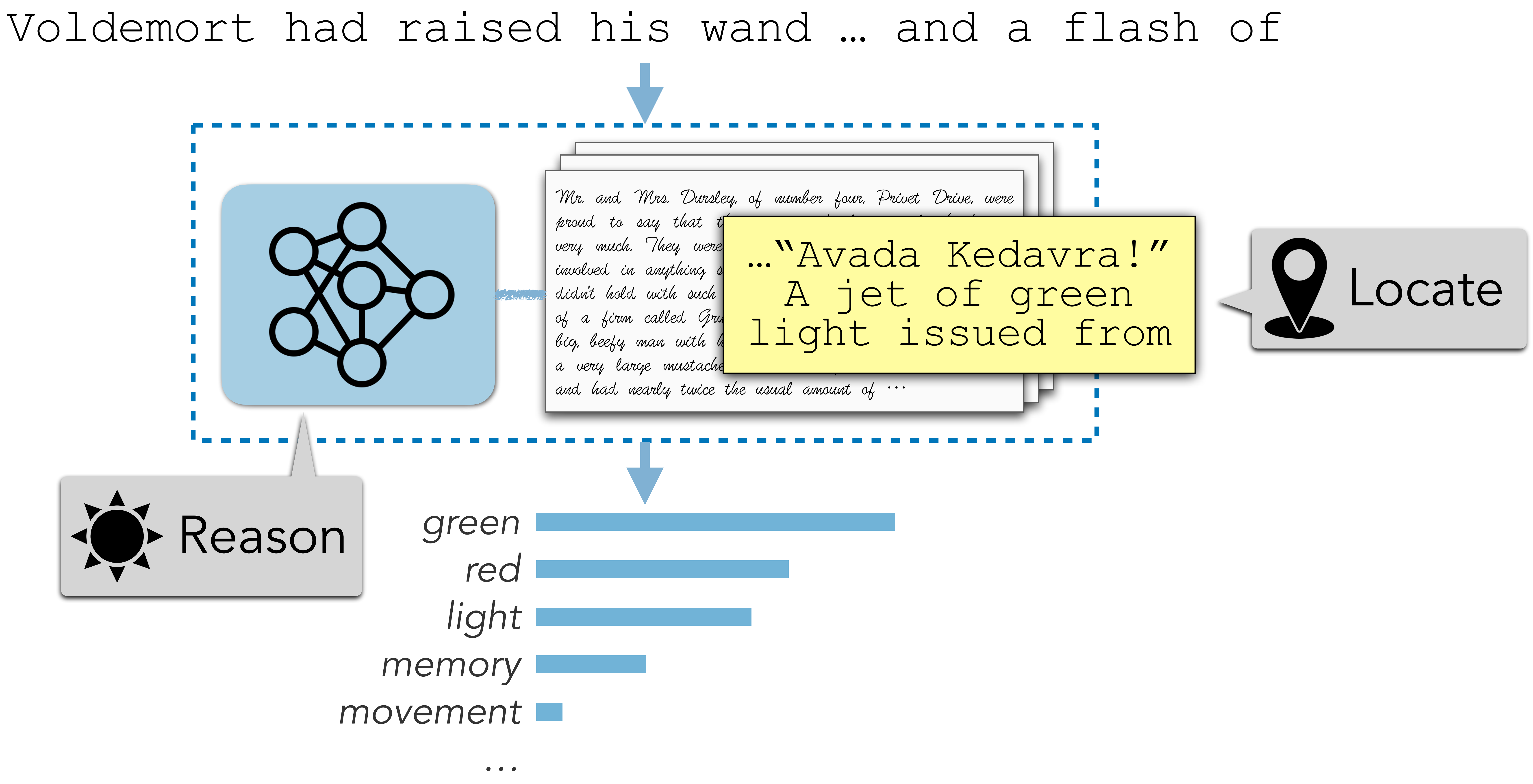


Keep the data!

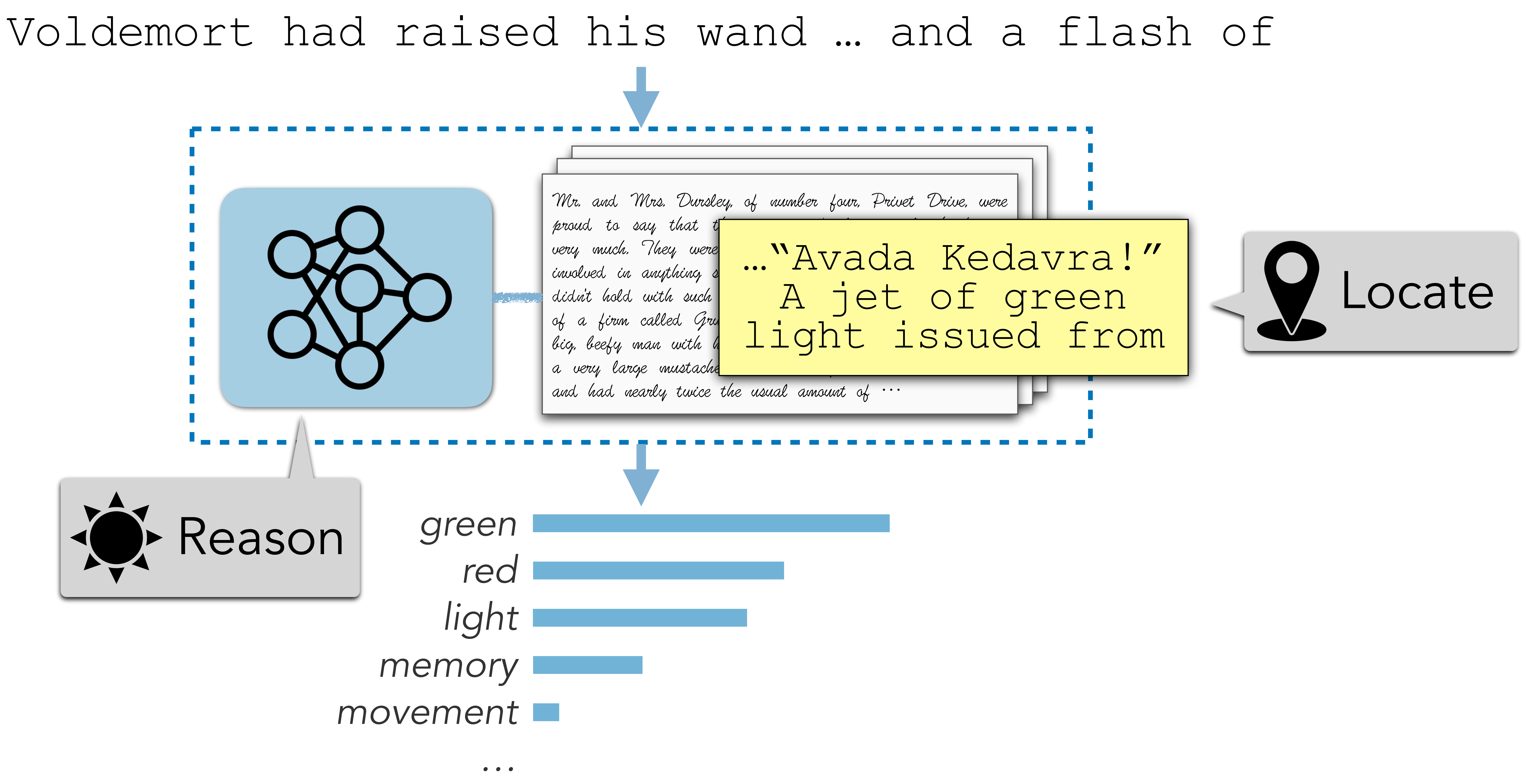
Voldemort had raised his wand … and a flash of



Keep the data!



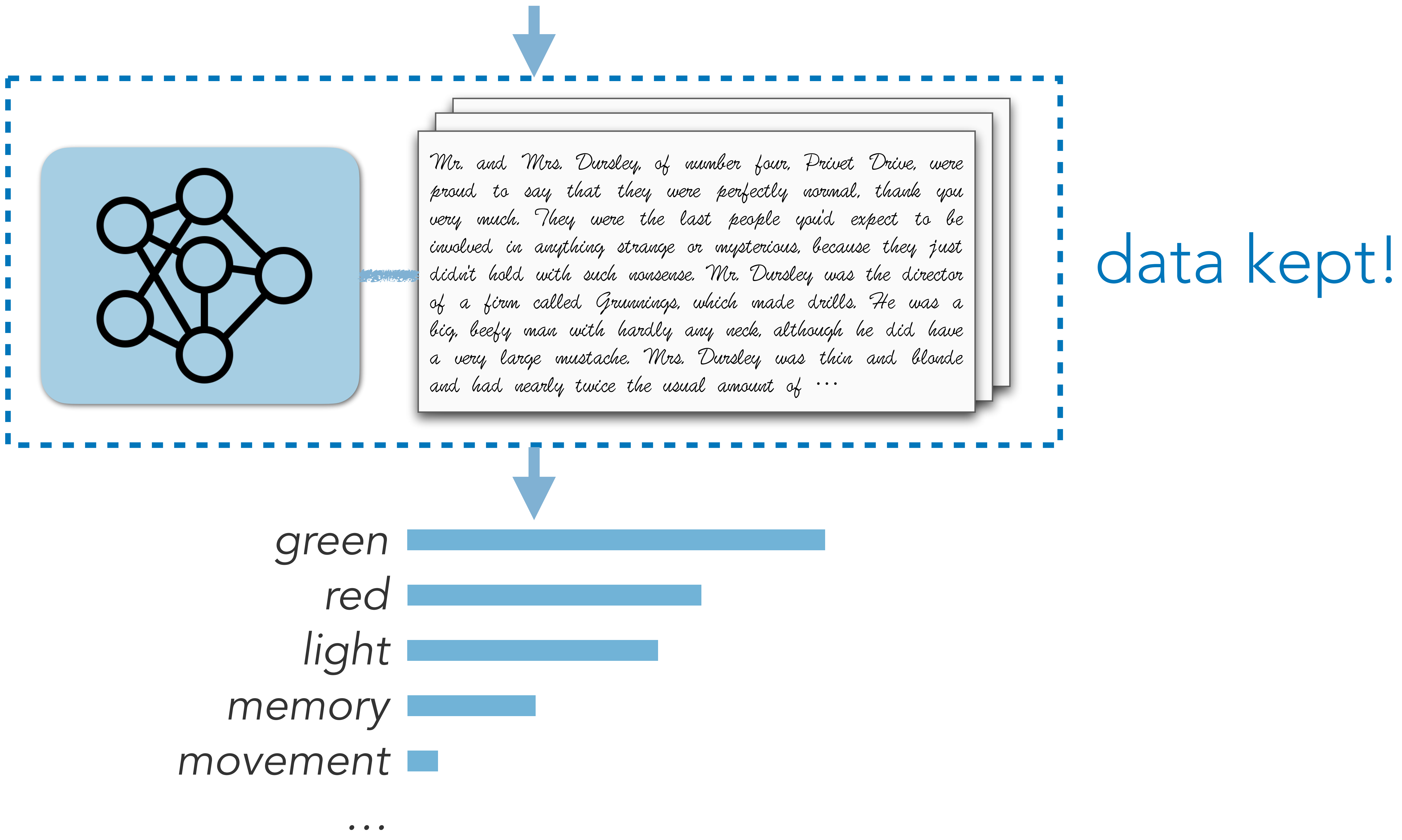
Keep the data!



No need to remember every detail

A retrieval-based LM

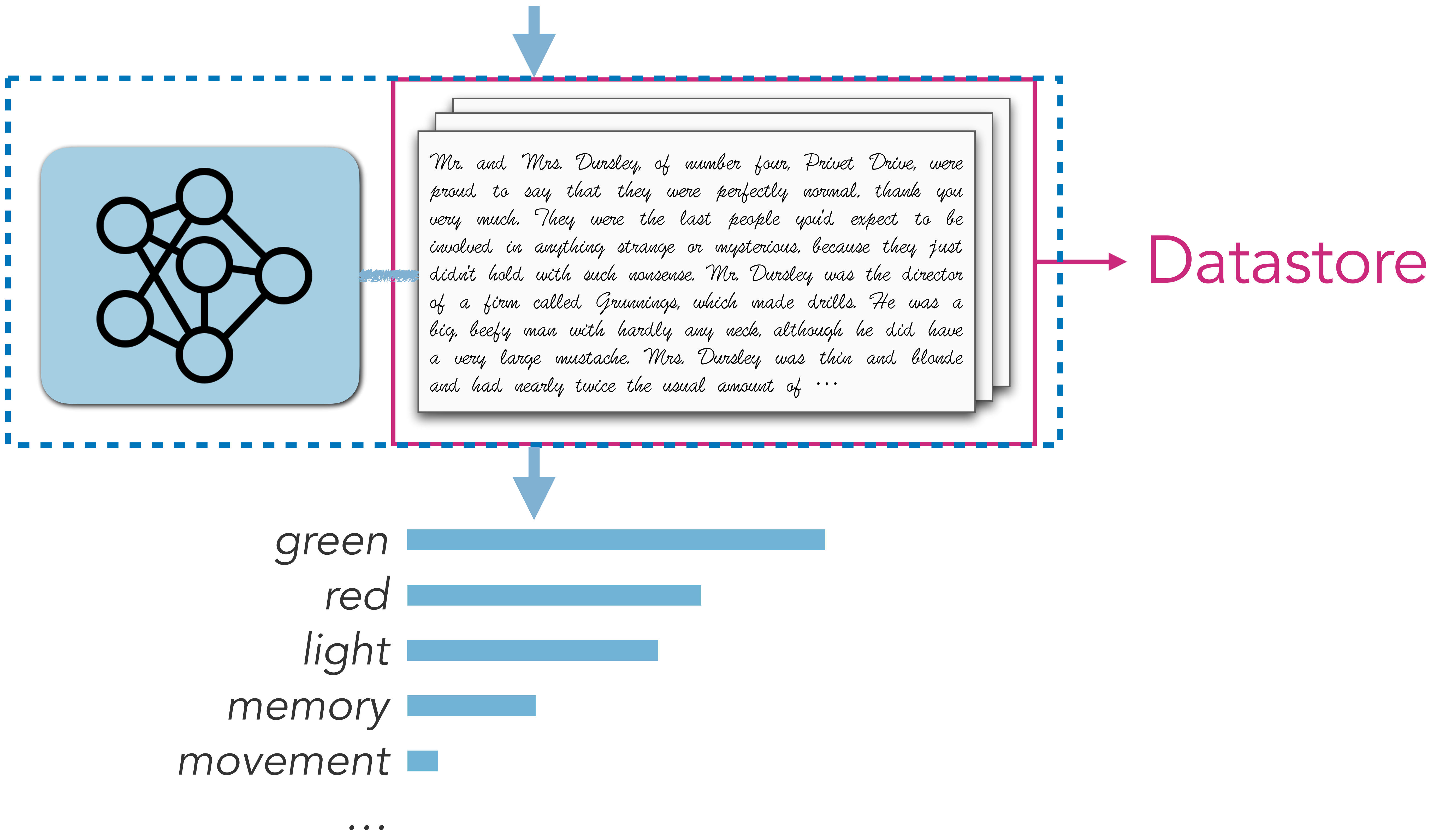
Voldemort had raised his wand … and a flash of



No need to remember every detail

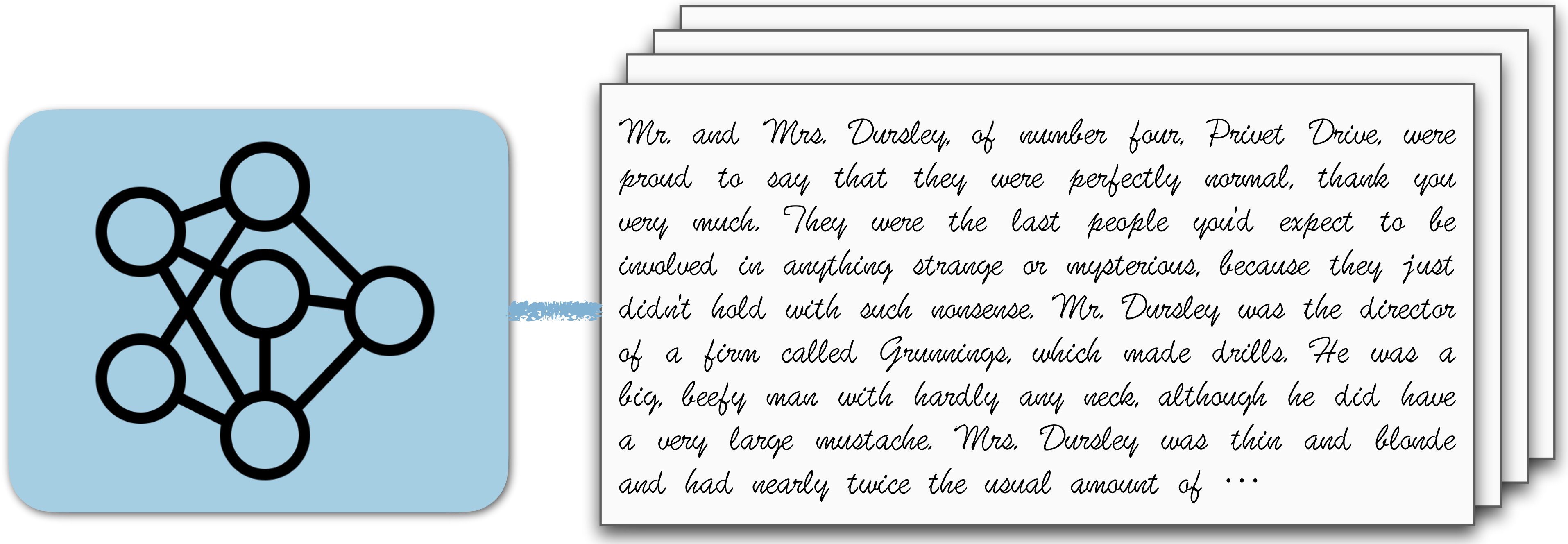
A retrieval-based LM

`Voldemort had raised his wand … and a flash of`



No need to remember every detail

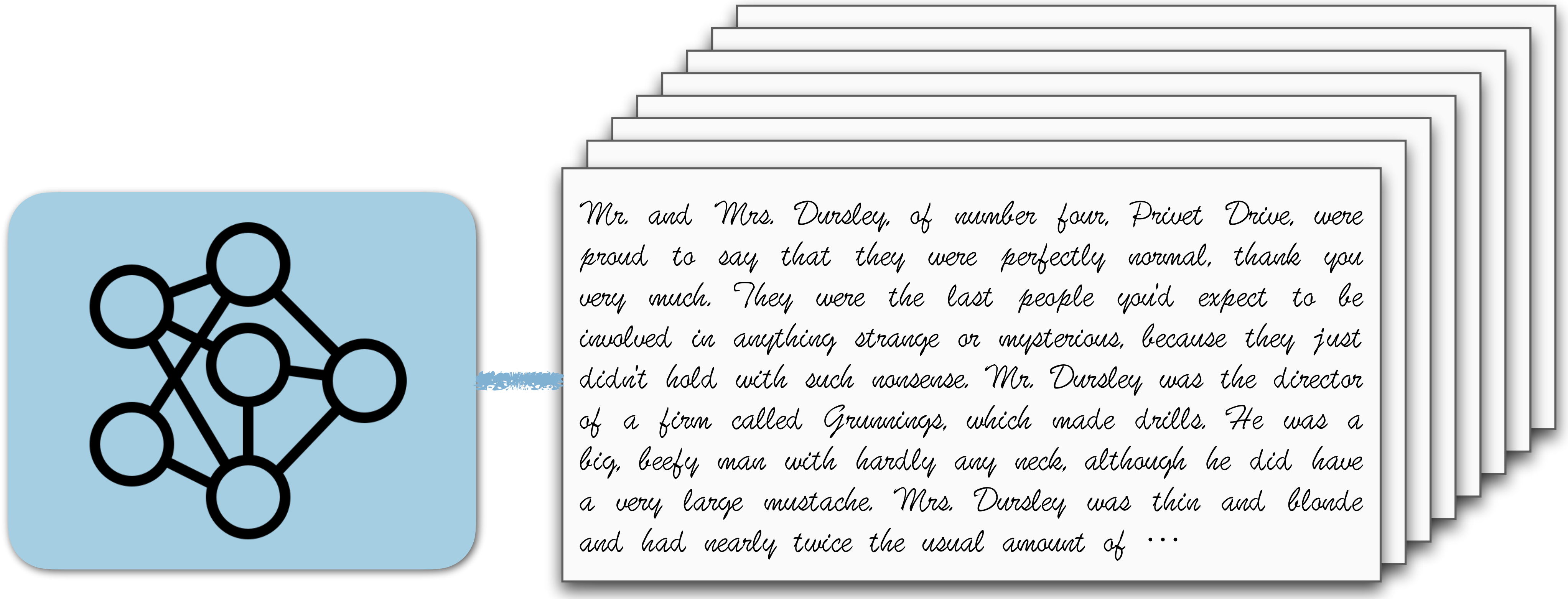
A retrieval-based LM



Datastore

No need to remember every detail, can grow & seamlessly update

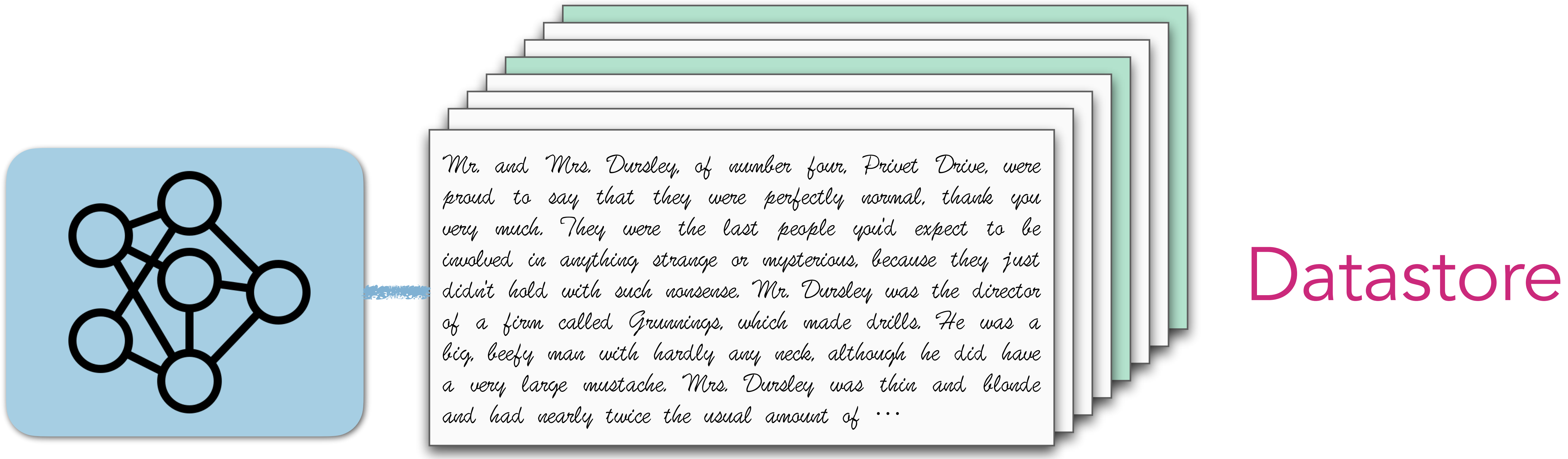
A retrieval-based LM



Datastore

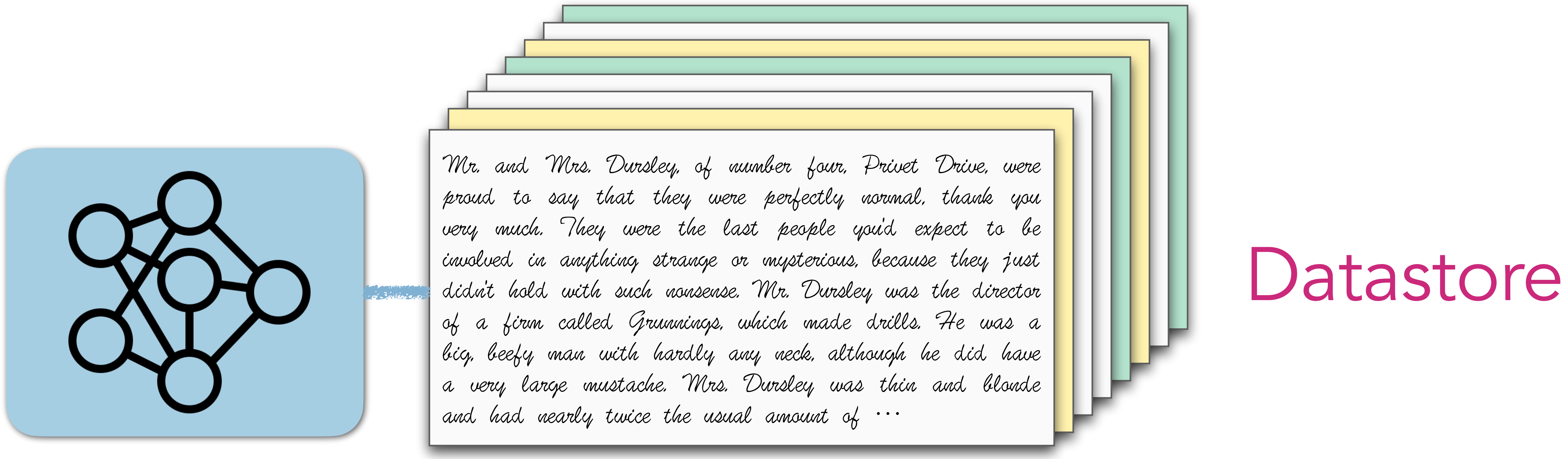
No need to remember every detail, can grow & seamlessly update

A retrieval-based LM



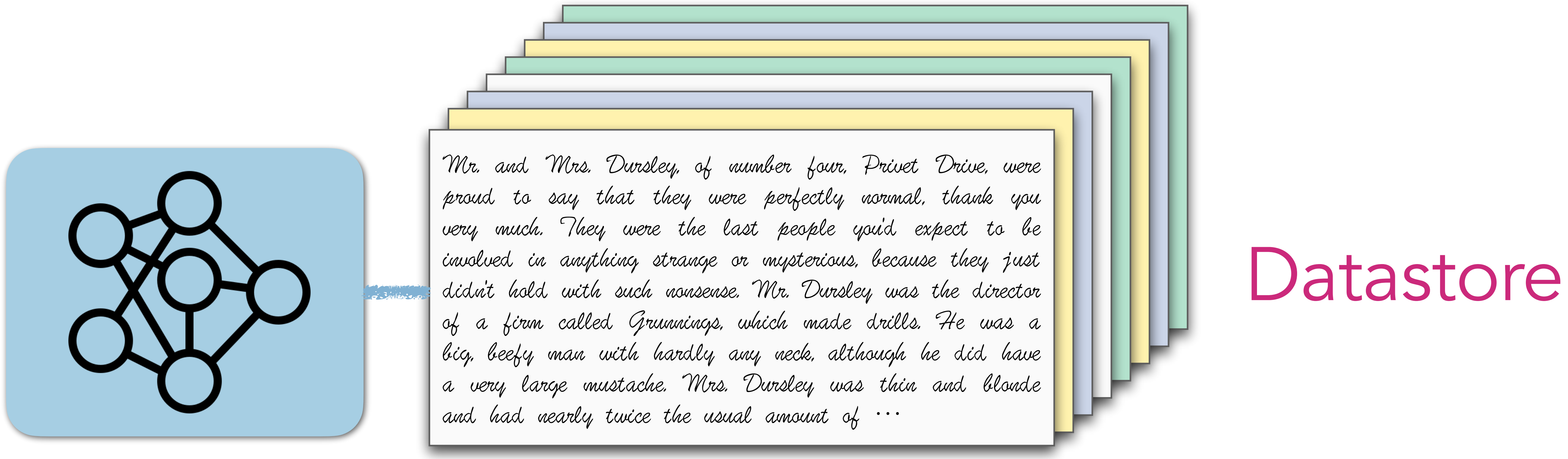
No need to remember every detail, can grow & seamlessly update

A retrieval-based LM



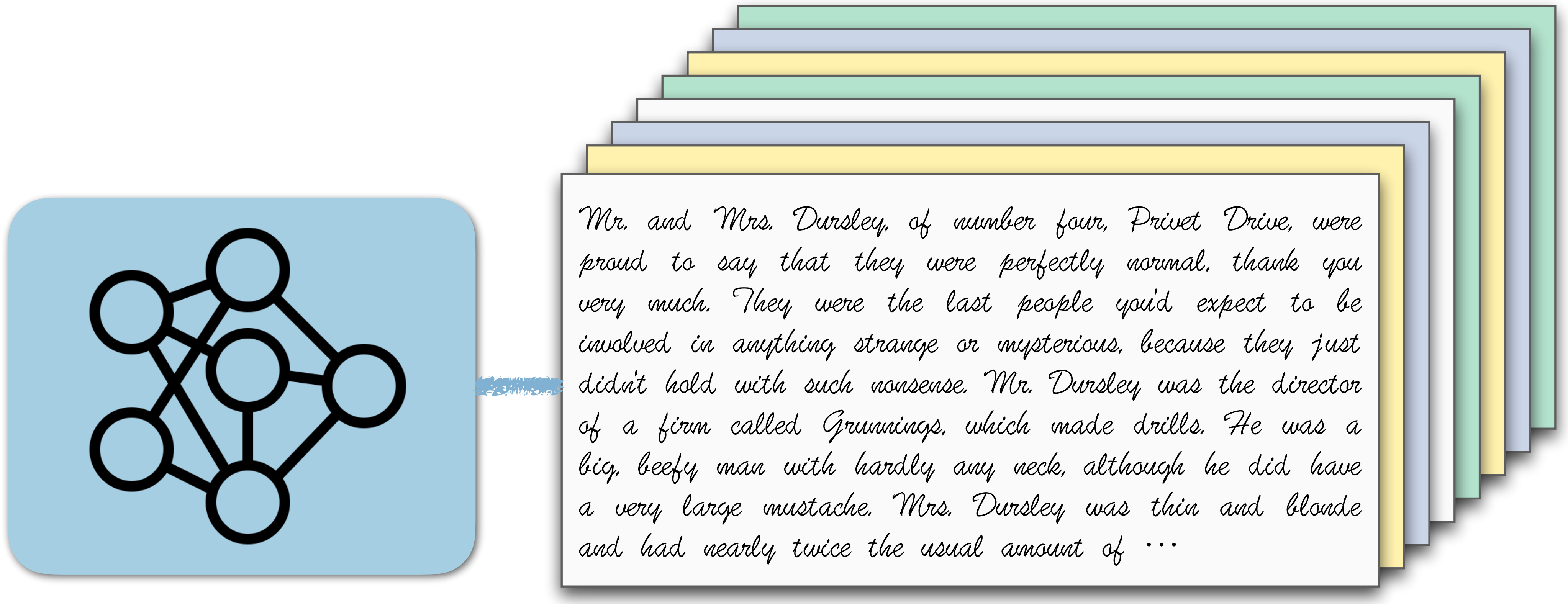
No need to remember every detail, can grow & seamlessly update

A retrieval-based LM



No need to remember every detail, can grow & seamlessly update

A retrieval-based LM

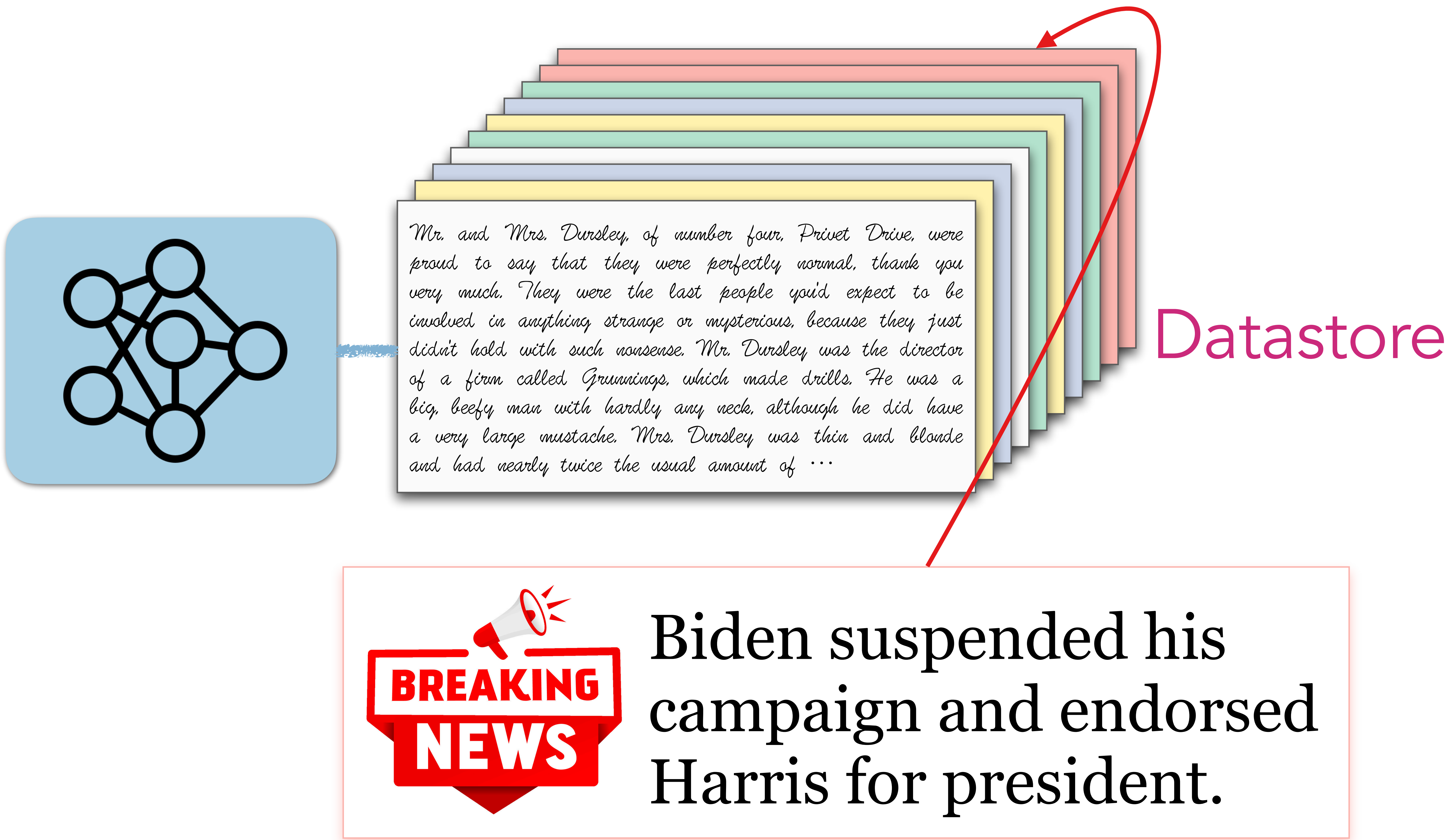


Datastore

Biden suspended his campaign and endorsed Harris for president.

No need to remember every detail, can grow & seamlessly update

A retrieval-based LM



No need to remember every detail, can grow & seamlessly update

Retrieval-based LMs: Why?

Retrieval-based LMs: Why?

☑ More performant

ChatGPT

Tell me a bio of Sewon Min.

Sewon Min is a researcher in natural language processing (NLP) and machine learning, particularly known for her work in question answering, retrieval-augmented language models, and large-scale reasoning systems. She completed her **Ph.D. in Computer Science at UC Berkeley**, advised by **Dan Klein**. After her Ph.D., she joined **Meta AI (FAIR)** as a researcher.

incorrect

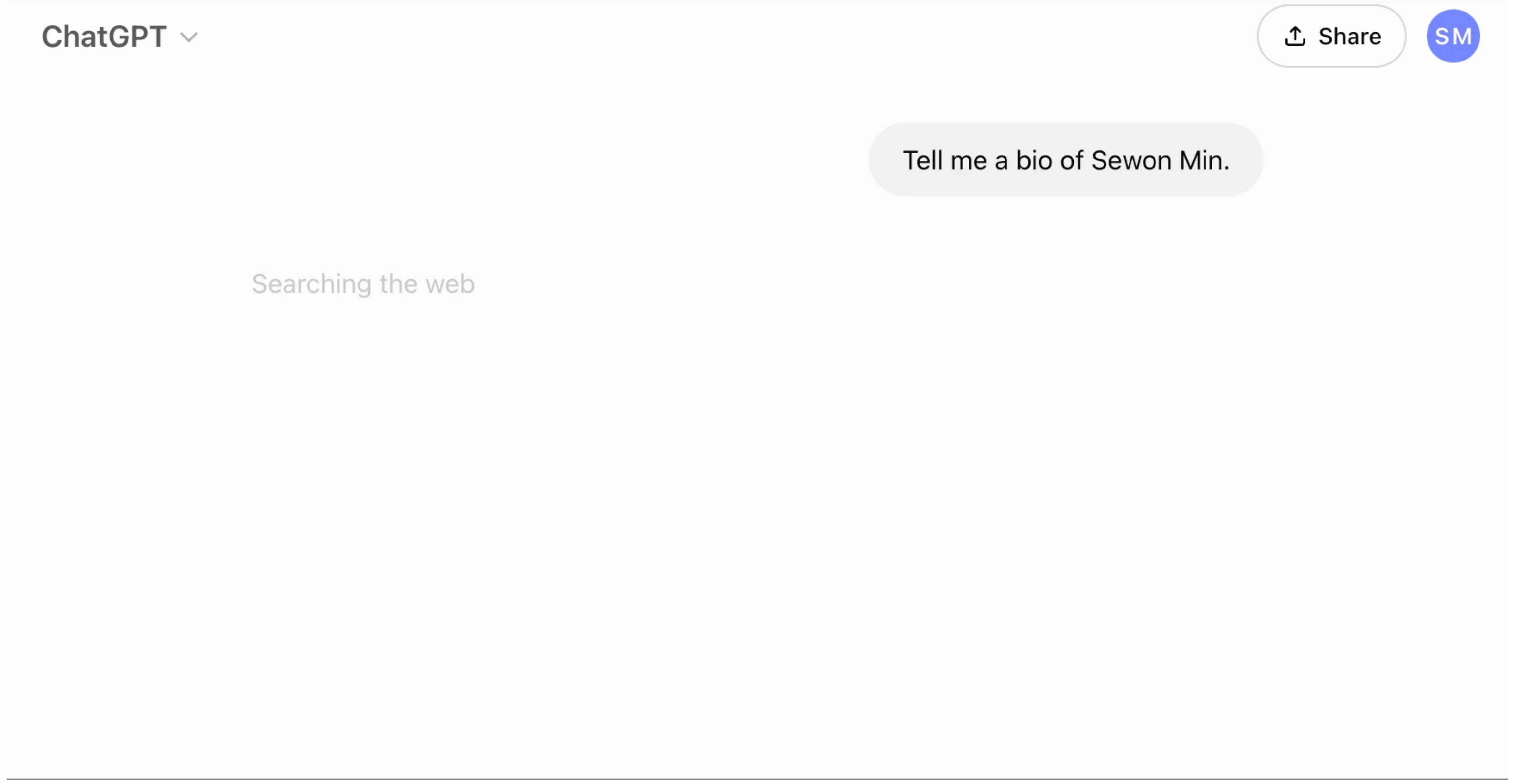
[Without web browsing]

✓ Long-tail knowledge

✓ Staying up-to-date

Retrieval-based LMs: Why?

☑ More performant



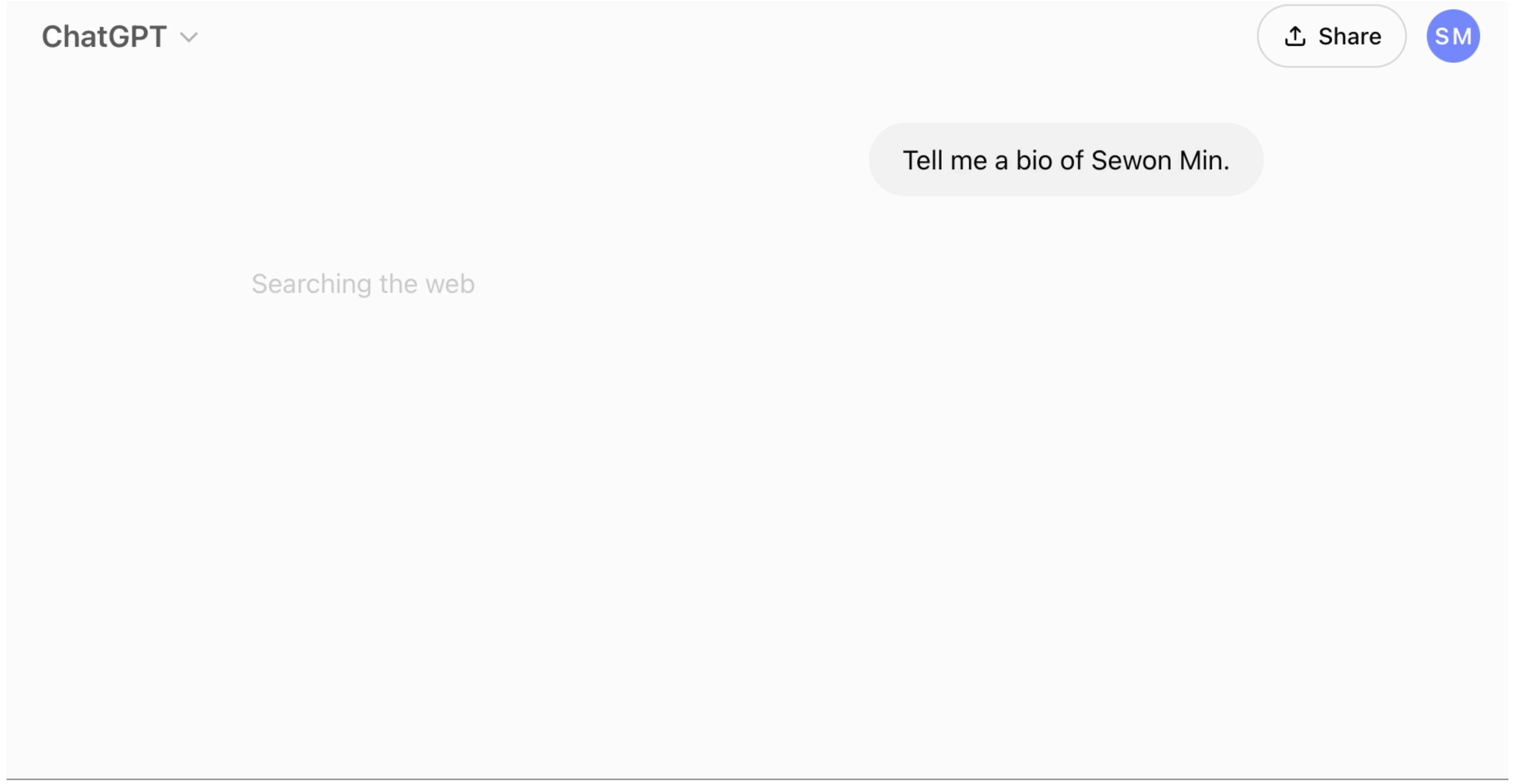
[With web browsing]

✓ Long-tail knowledge

✓ Staying up-to-date

Retrieval-based LMs: Why?

☑ More performant



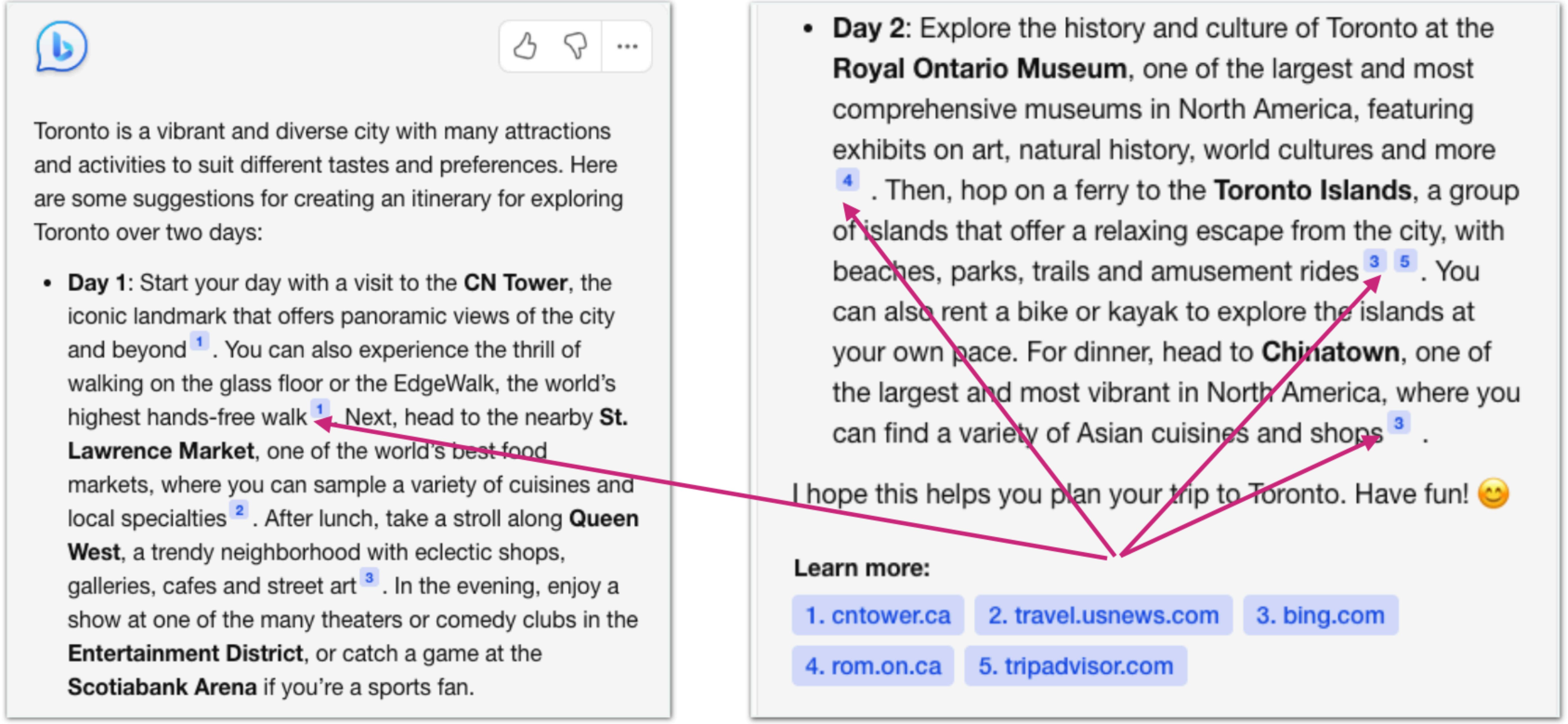
[With web browsing]

✓ Long-tail knowledge

✓ Staying up-to-date

Retrieval-based LMs: Why?

☑ More performant



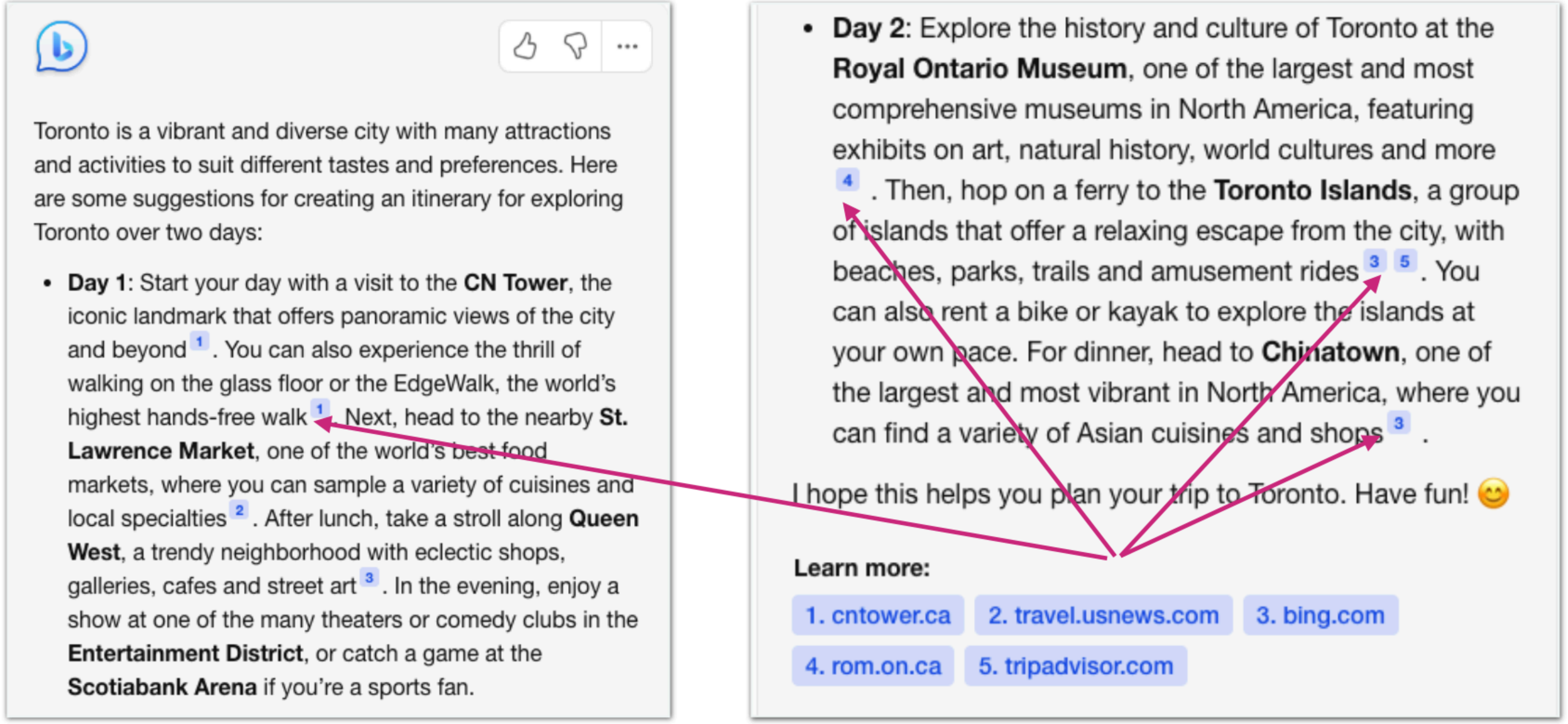
✓ Long-tail knowledge

✓ Staying up-to-date

✓ Citation (for better explainability and verification)

Retrieval-based LMs: Why?

- [x] More performant



✓ Long-tail knowledge

✓ Staying up-to-date

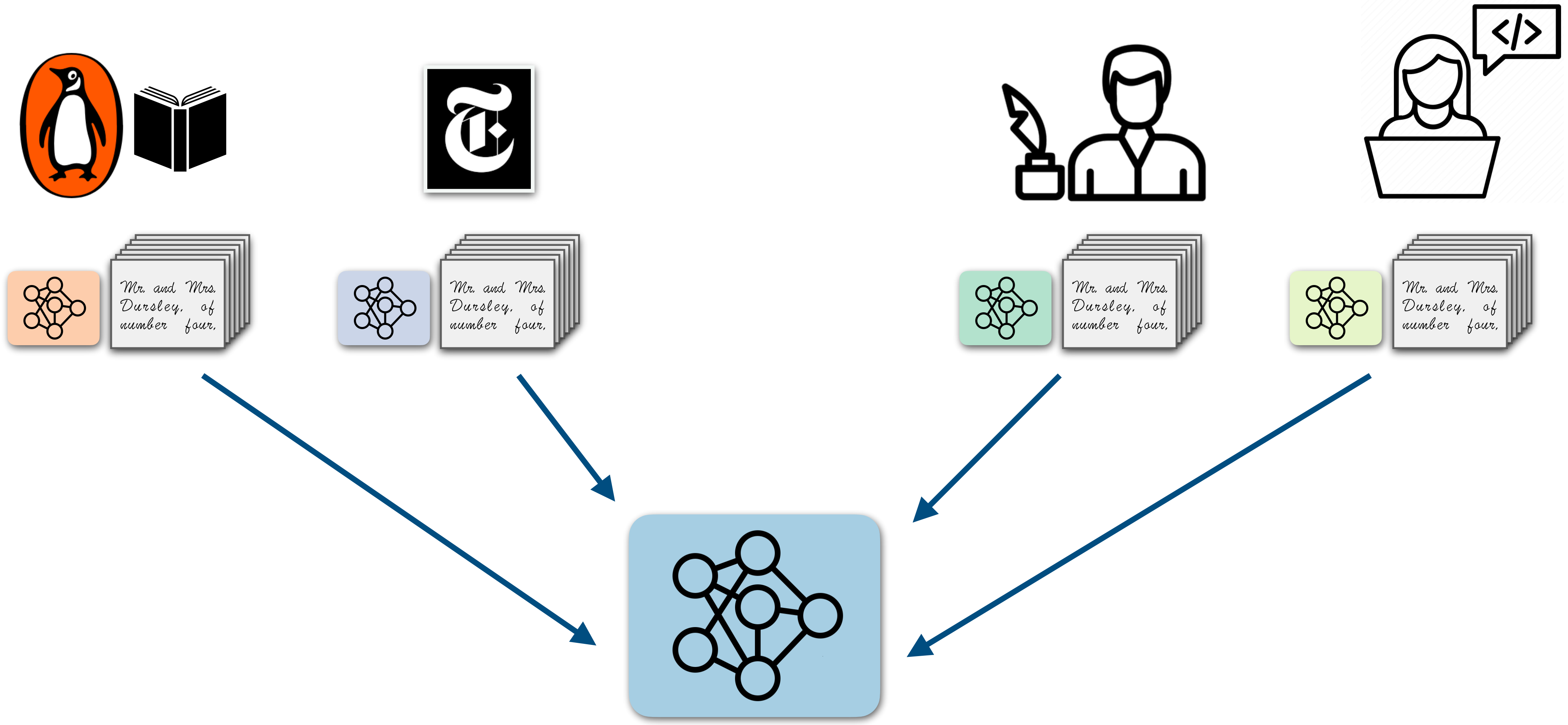
✓ Citation (for better explainability and verification)

Frontier models are already using retrieval as a key feature

Retrieval-based LMs: Why?

- More performant
- More flexibility

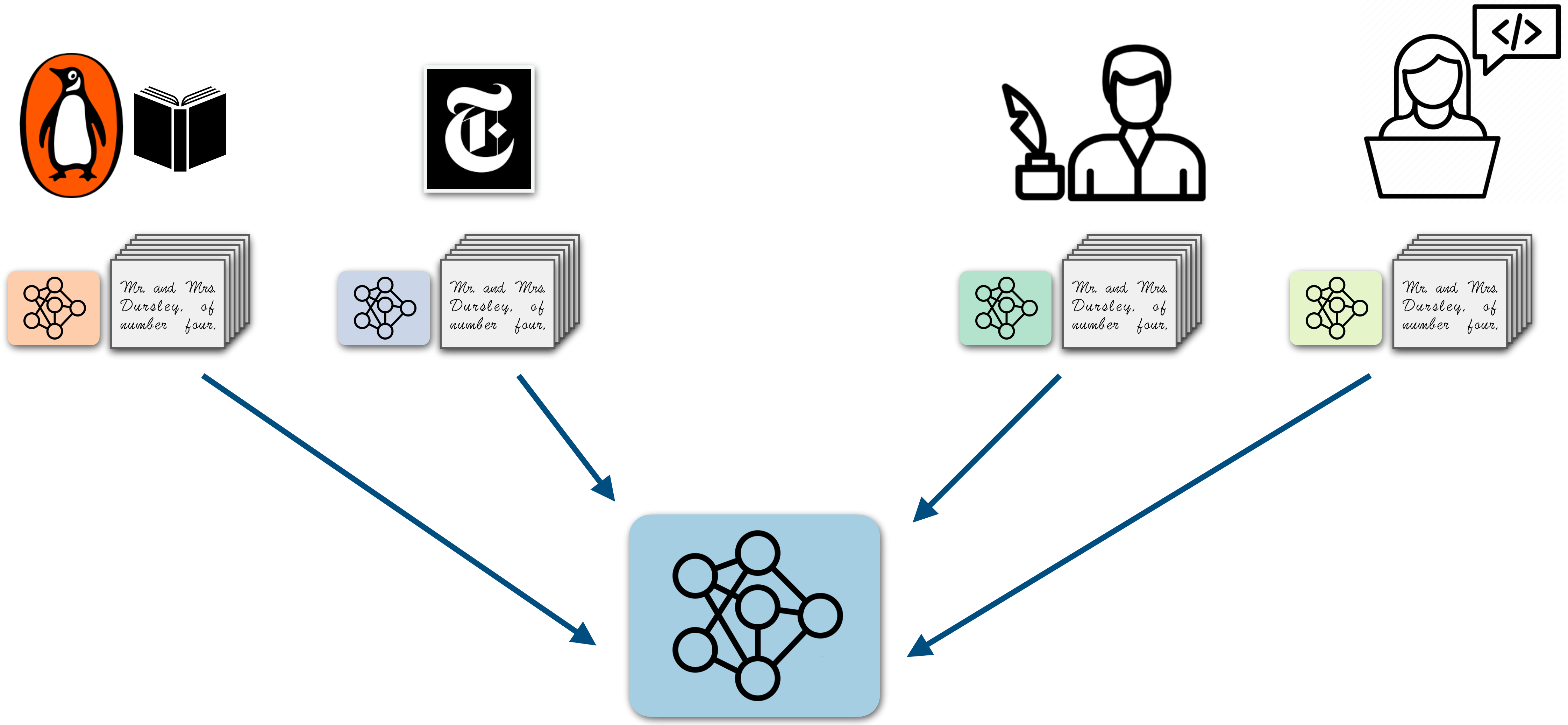
(Proprietary, private, or copyrighted data)



Retrieval-based LMs: Why?

- [x] More performant
- [x] More flexibility

(Proprietary, private, or copyrighted data)



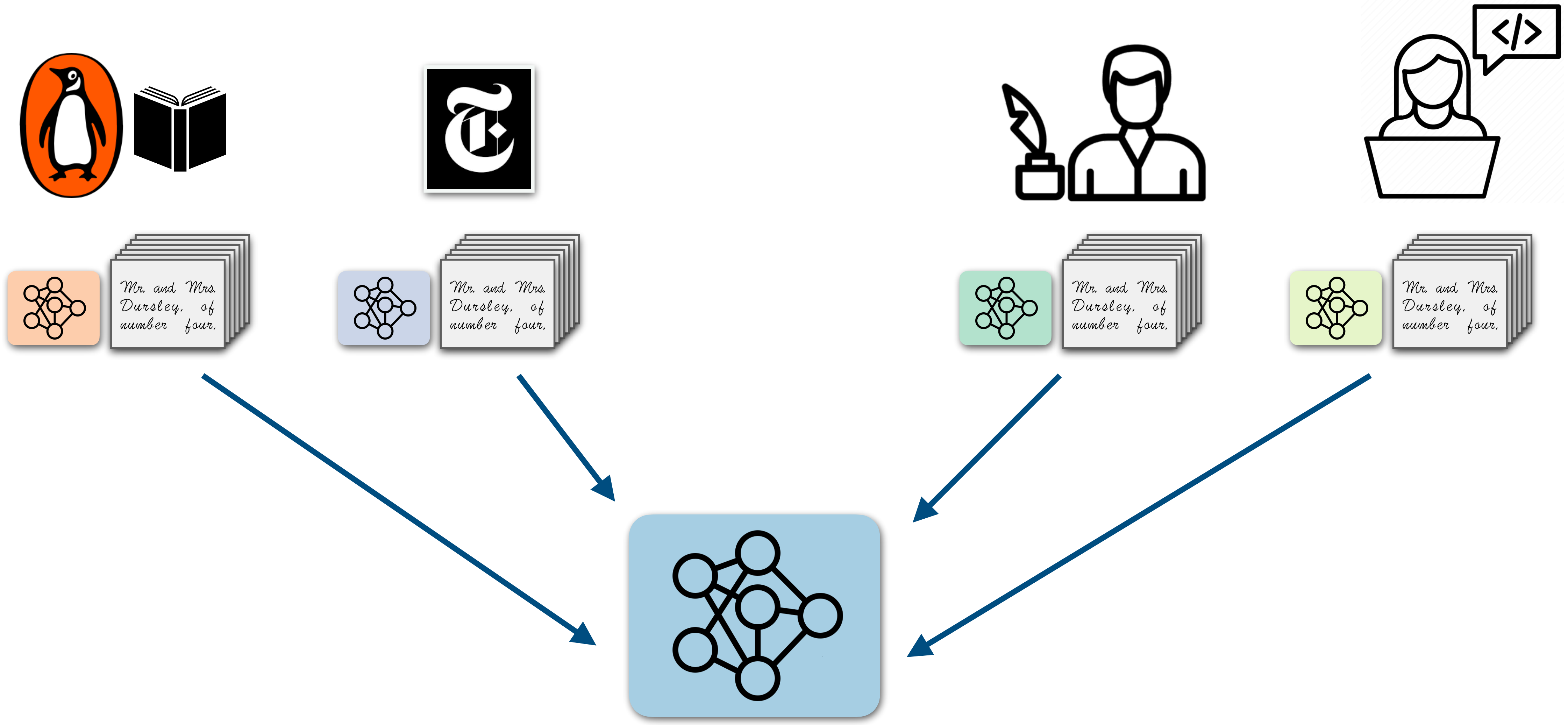
✓ Flexibility to be added or removed later (for free)

Retrieval-based LMs: Why?

☑ More performant

☑ More flexibility

(Proprietary, private, or copyrighted data)



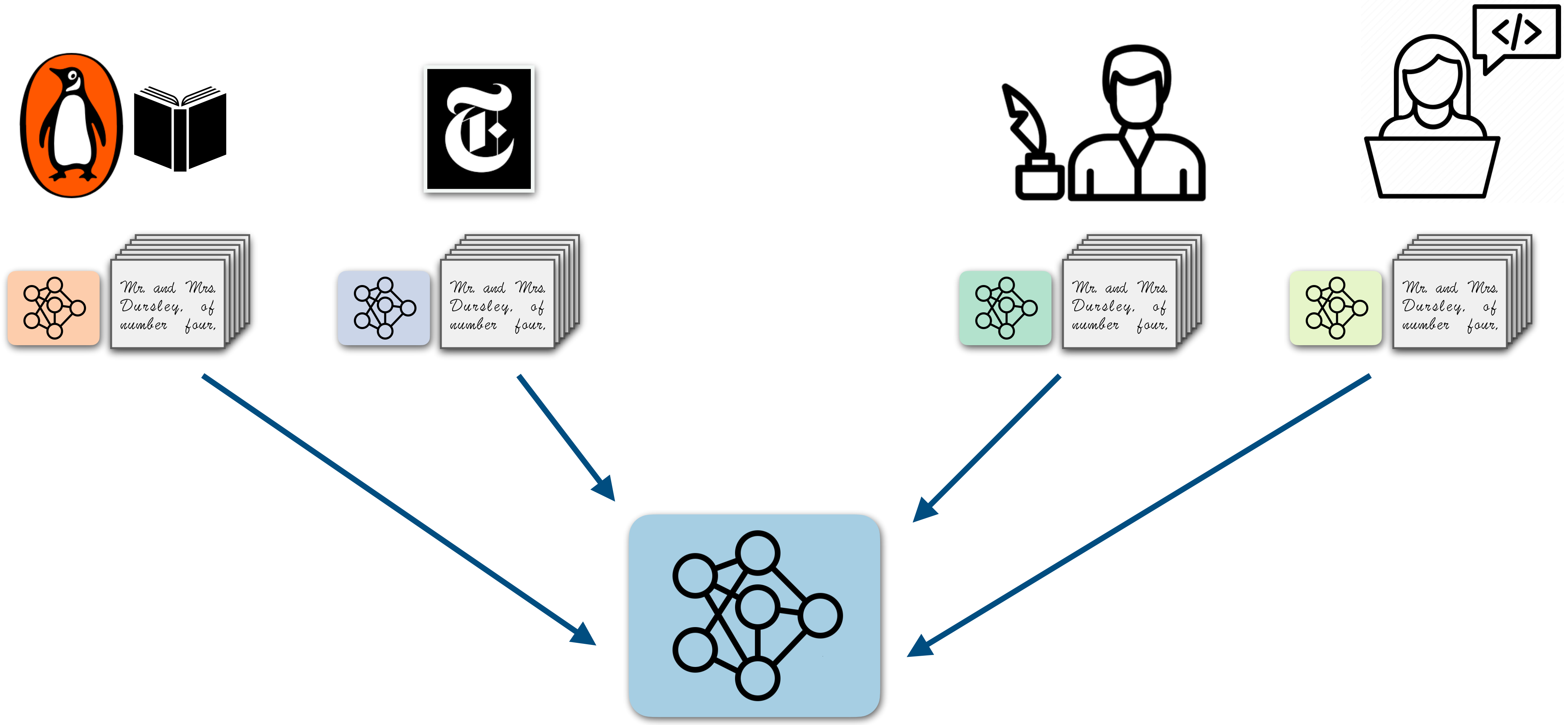
✓ Flexibility to be added or removed later (for free)

✓ Flexibility to be hosted remotely

Retrieval-based LMs: Why?

- More performant
- More flexibility

(Proprietary, private, or copyrighted data)

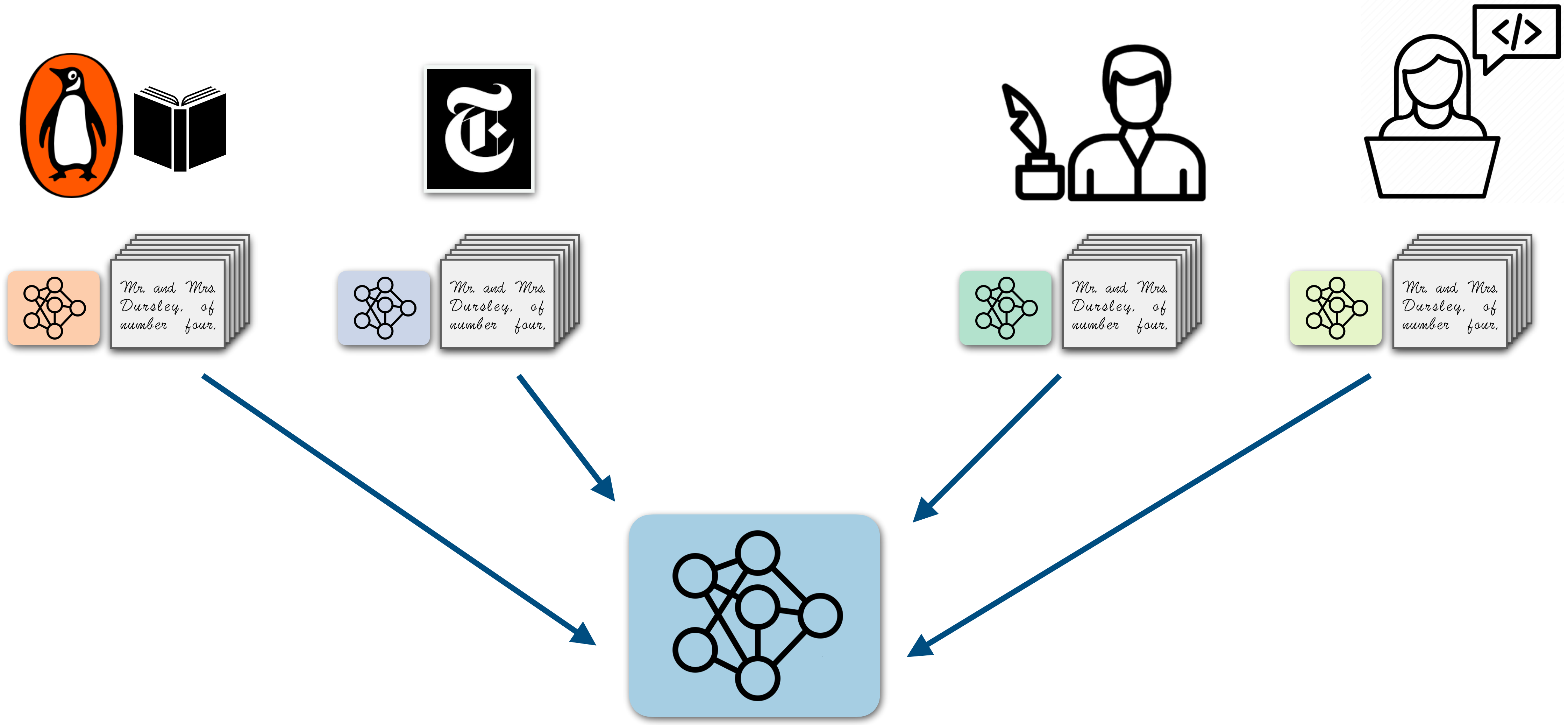


- ✓ Flexibility to be added or removed later (for free)
- ✓ Flexibility to be hosted remotely
- ✓ Attribution & credit assignment

Retrieval-based LMs: Why?

- More performant
- More flexibility

(Proprietary, private, or copyrighted data)



✓ Flexibility to be added or removed later (for free)

✓ Flexibility to be hosted remotely

✓ Attribution & credit assignment

We can think of retrieval as an alternative way to *use* the data (in addition to *"training"* on the data)

Today's Lecture

Part 1. **Basics** of retrieval-based LMs (35min)

- Retrieval
- Augmentation
- Training of retrieval-based LMs

Part 2. **Recent research** on *scaling* retrieval-based LMs (35min)

- Scalable Pre-training with Retrieval
- Scaling a Datastore
- Datastore for Responsible Data Use

Open Problems (10min)

Today's Lecture

Part 1. **Basics** of retrieval-based LMs (35min)

- Retrieval
- Augmentation
- Training of retrieval-based LMs

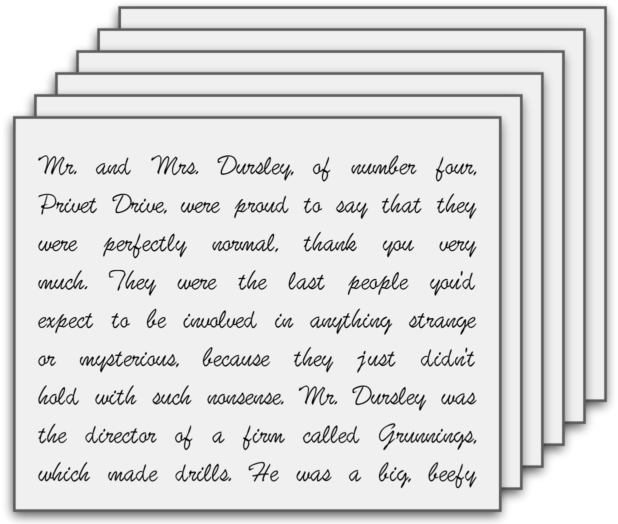
Part 2. **Recent research** on *scaling* retrieval-based LMs (35min)

- Scalable Pre-training with Retrieval
- Scaling a Datastore
- Datastore for Responsible Data Use

Open Problems (10min)

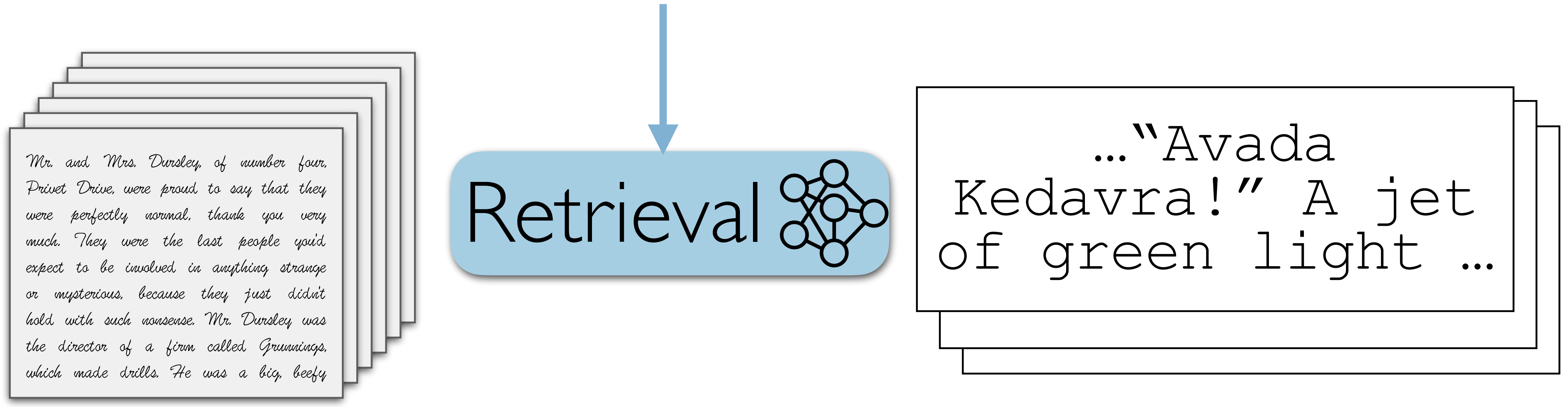
A two-stage pipeline

`Voldemort had raised his wand … and a flash of`

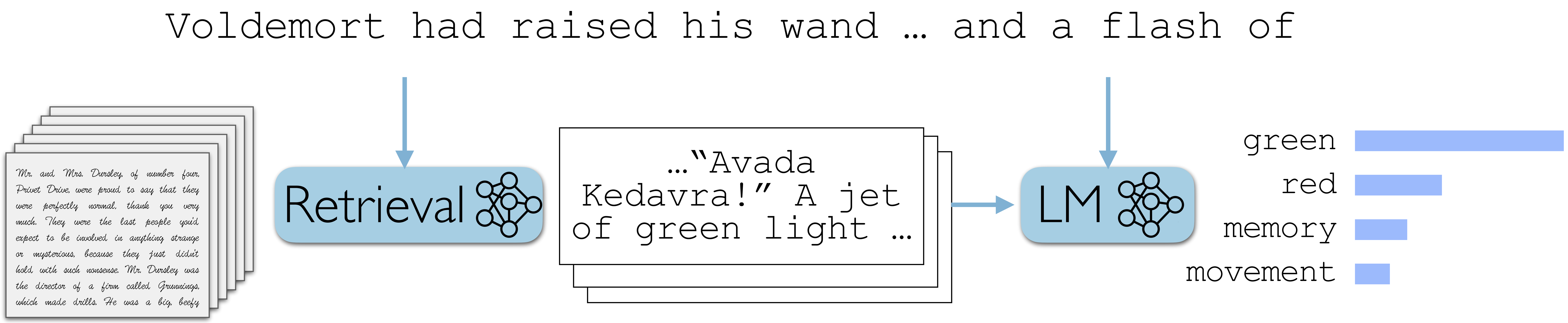


A two-stage pipeline

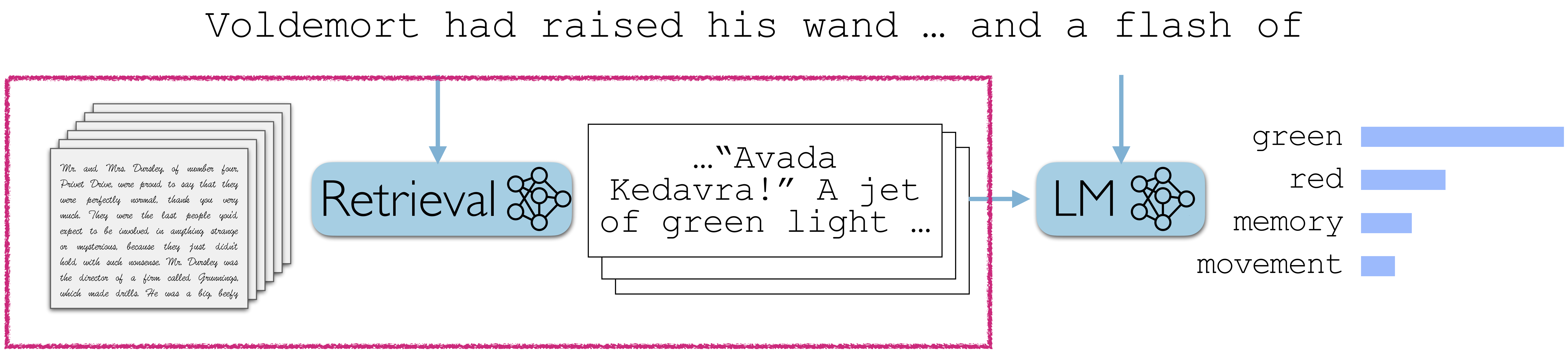
Voldemort had raised his wand … and a flash of



A two-stage pipeline

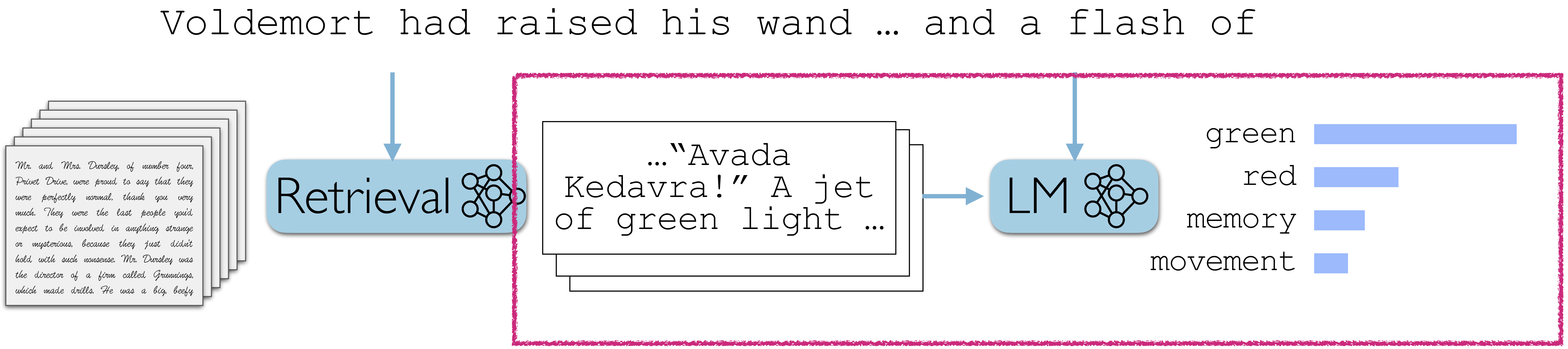


A two-stage pipeline



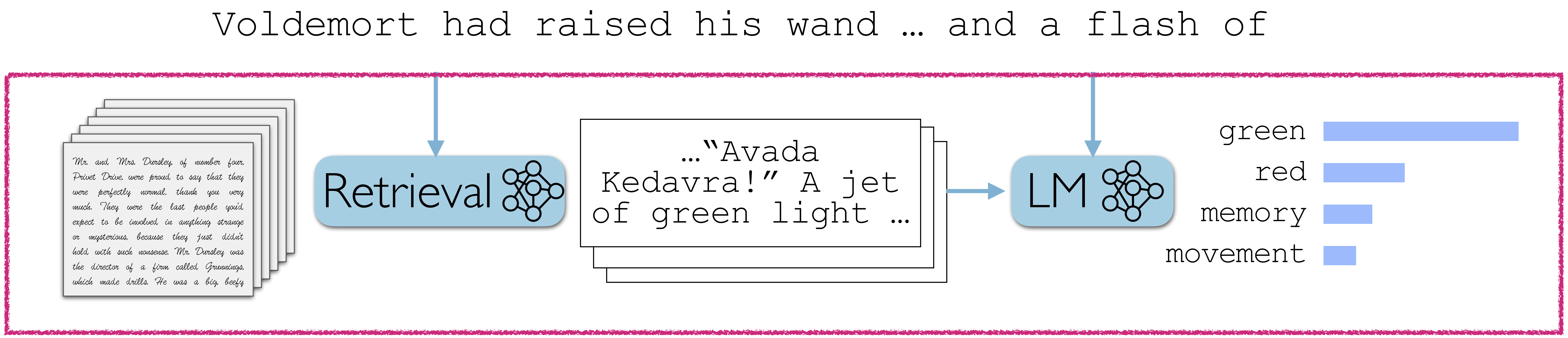
1) Retrieval

A two-stage pipeline



1) Retrieval **2) Augmentation**

A two-stage pipeline

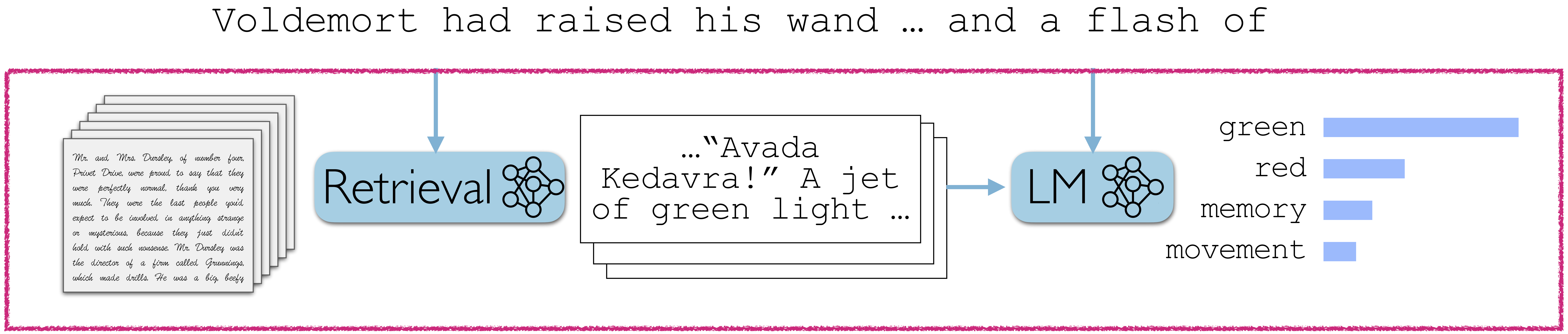


1) Retrieval

2) Augmentation

3) Training

A two-stage pipeline



1) Retrieval

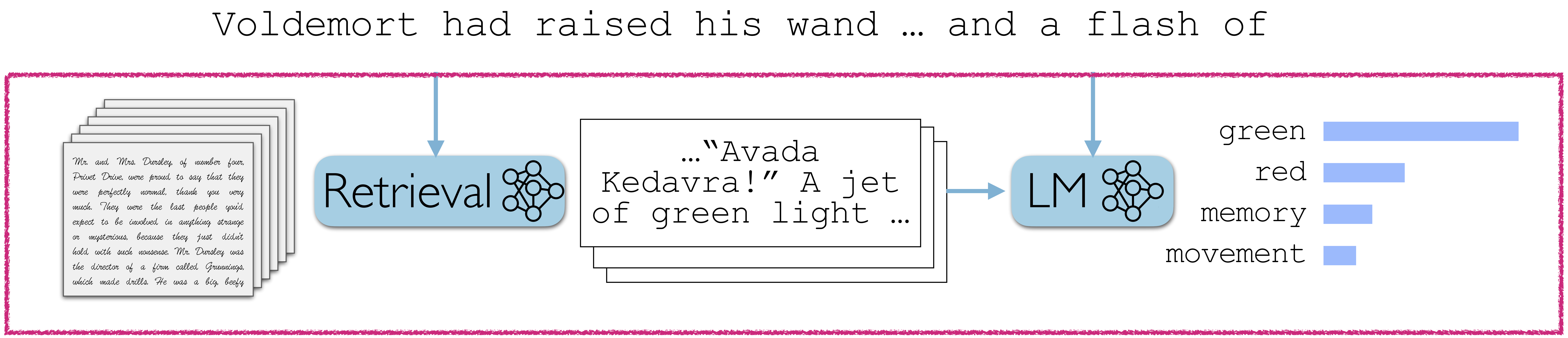
2) Augmentation

3) Training

(At the end, different architectures beyond the two-stage pipeline)

A two-stage pipeline

Earlier work: Li et al. 2016, Chen et al. 2017, Gu et al 2017, Zhang et al. 2018



1) Retrieval

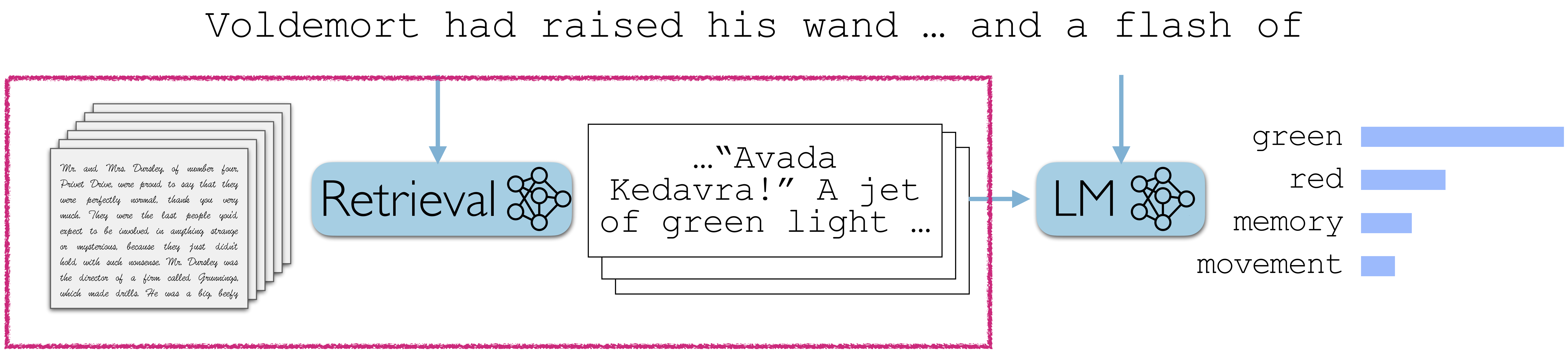
2) Augmentation

3) Training

(At the end, different architectures beyond the two-stage pipeline)

A two-stage pipeline

Earlier work: Li et al. 2016, Chen et al. 2017, Gu et al 2017, Zhang et al. 2018



1) Retrieval

2) Augmentation

3) Training

A retrieval problem

A retrieval problem

($\text{Passages in the datastore: } z_1, \cdots, z_N$

A retrieval problem

$\left(\text{Passages in the datastore: } z_1, \cdots, z_N\text{, Input: } x\right)$

A retrieval problem

$\left(\text{Passages in the datastore: } z_1, \cdots, z_N, \text{ Input: } x\right) \longrightarrow z_i \quad (1 \leq i \leq N)$

A retrieval problem

$$\left(\text{Passages in the datastore: } z_1, \cdots, z_N\text{, Input: } x\right) \longrightarrow z_i \quad (1 \leq i \leq N)$$

Can generalize to k passages
(usually $k \leq 100$)

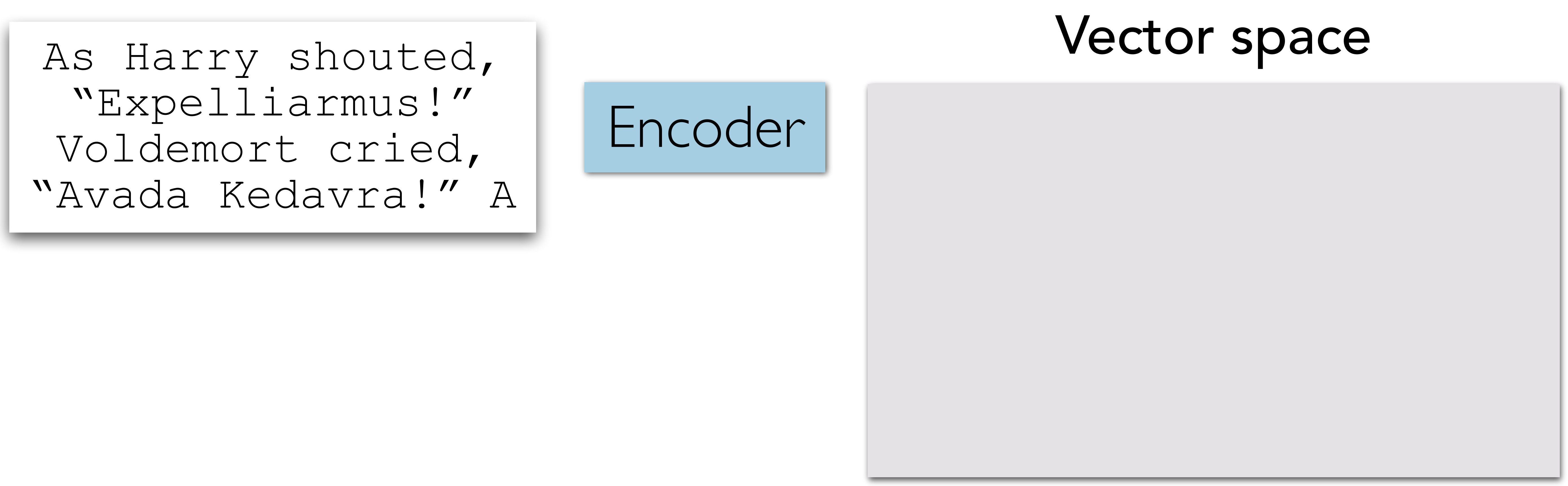
A **neural** retrieval problem

$\big($Passages in the datastore: $z_1, \cdots, z_N$, Input: $x\big) \longrightarrow z_i$ $(1 \le i \le N)$

"Siamese" network (Bromley et al. 1993, Chopra et al 2005, Yih et al 2011, Huang et al 2013)

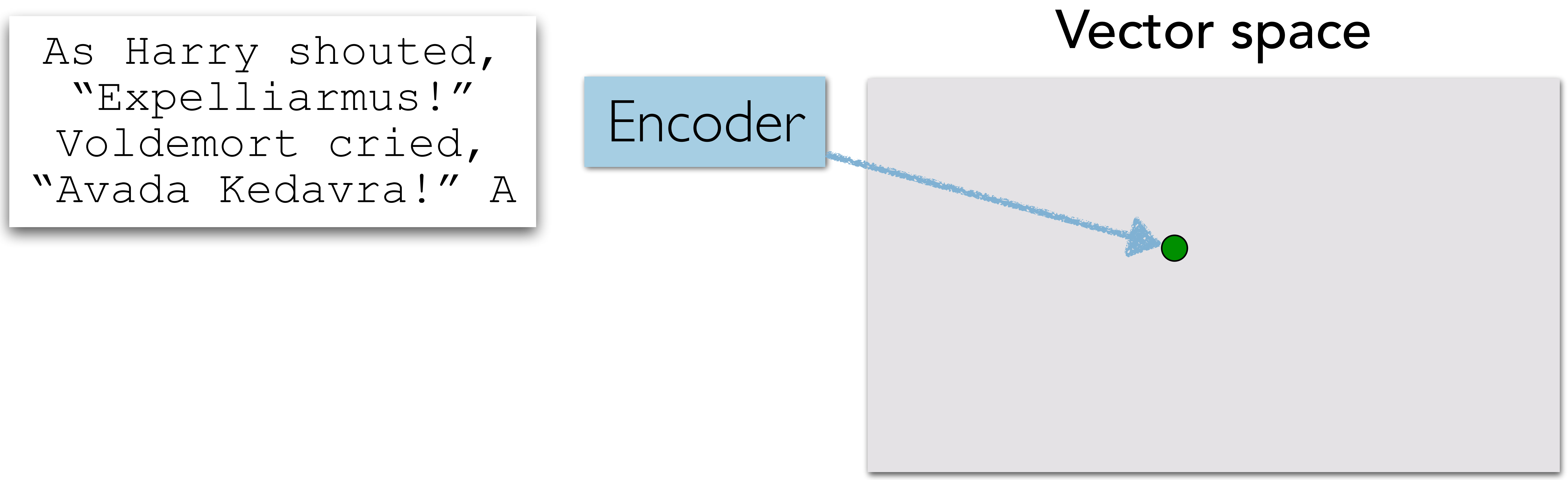
A neural retrieval problem

$\Big($Passages in the datastore: $z_1, \cdots, z_N$, Input: $x\Big) \longrightarrow z_i \quad (1 \leq i \leq N)$



A neural retrieval problem

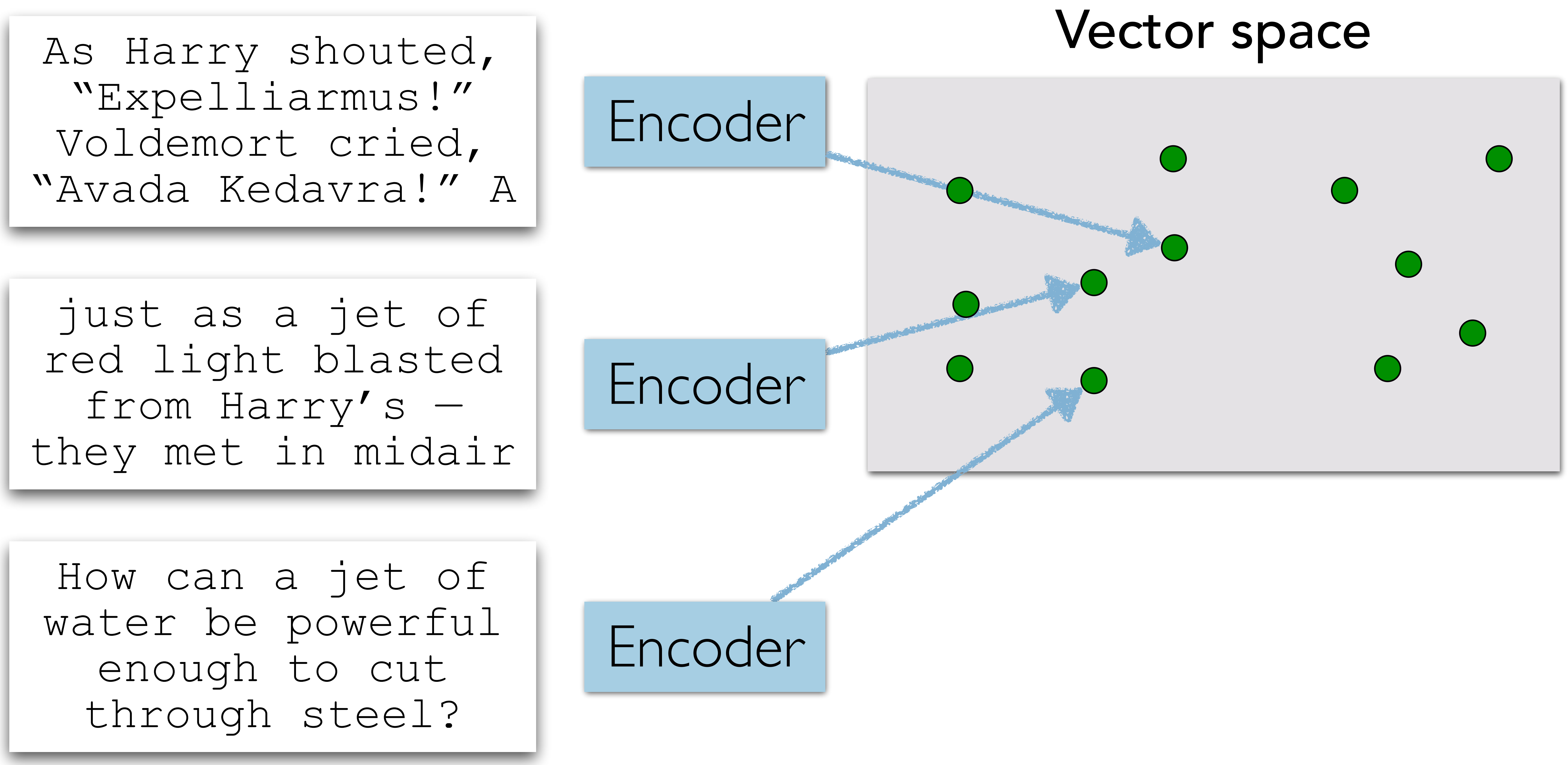
$\left(\text{Passages in the datastore: } z_1, \cdots, z_N, \text{ Input: } x\right) \longrightarrow z_i \quad (1 \le i \le N)$



$\mathbf{z}_i = \mathrm{Enc}(z_i) \in \mathbb{R}^h \quad (1 \le i \le N)$

A neural retrieval problem

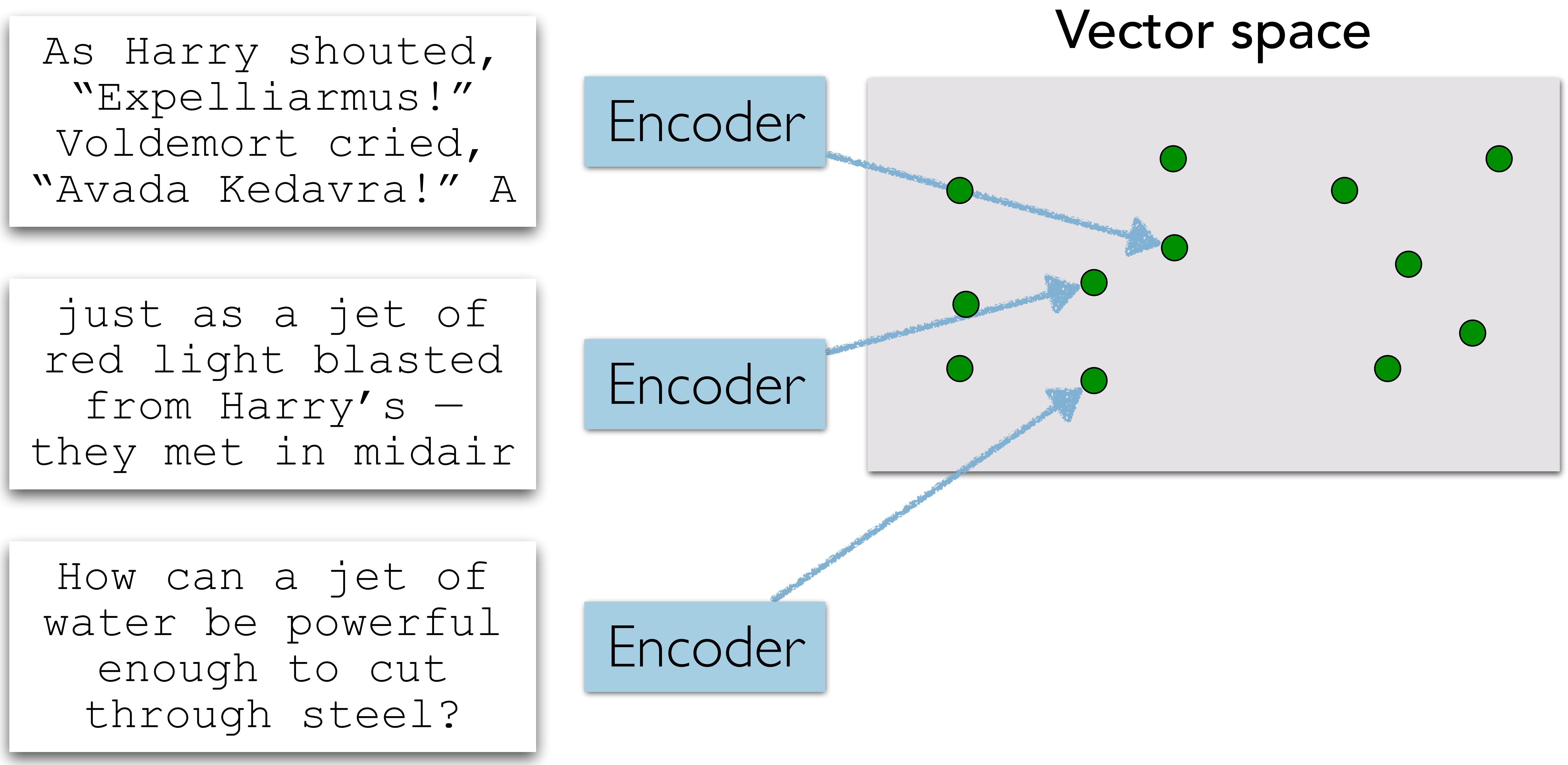
$\left(\text{Passages in the datastore: } z_1, \cdots, z_N, \text{ Input: } x\right) \longrightarrow z_i \quad (1 \leq i \leq N)$



$\mathbf{z}_i = \text{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$

A neural retrieval problem

$\left(\text{Passages in the datastore: } z_1, \cdots, z_N, \text{ Input: } x\right) \longrightarrow z_i \quad (1 \le i \le N)$

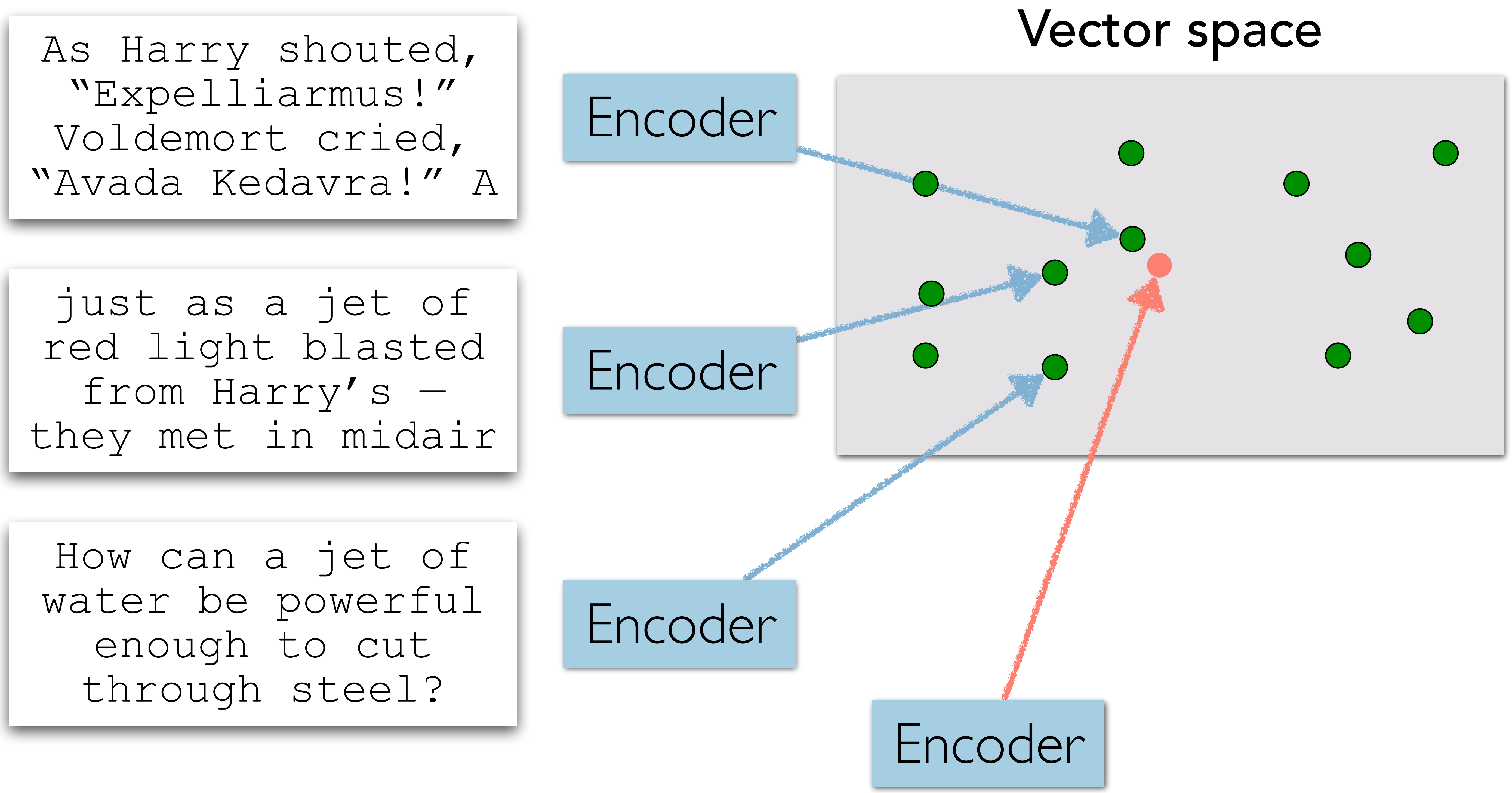


$$\mathbf{z}_i = \mathrm{Enc}(z_i) \in \mathbb{R}^h \quad (1 \le i \le N)$$

Voldemort had raised his wand … and a flash of

A neural retrieval problem

$\left(\text{Passages in the datastore: } z_1, \cdots, z_N, \text{ Input: } x\right) \longrightarrow z_i \quad (1 \le i \le N)$



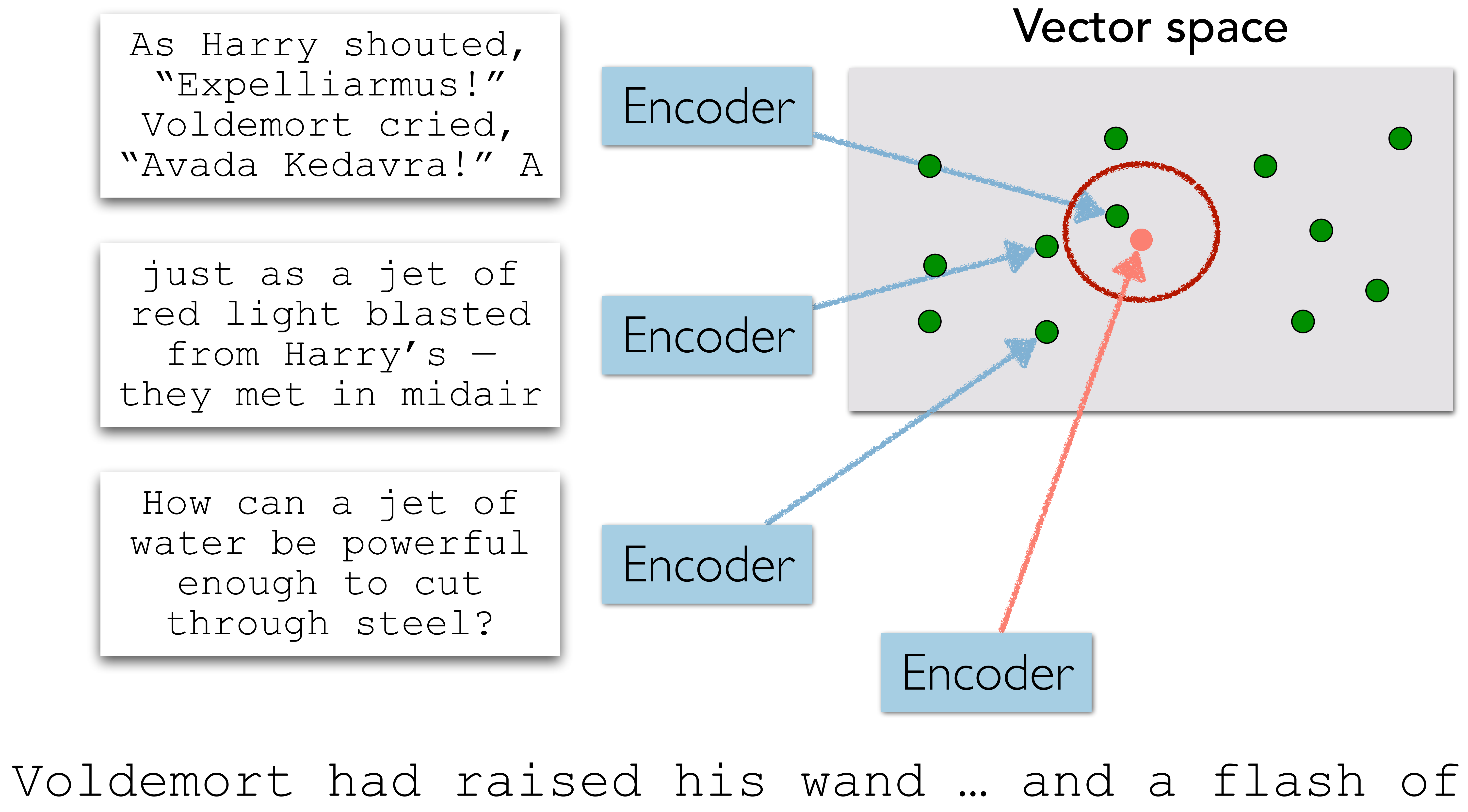
$\mathbf{z}_i = \mathrm{Enc}(z_i) \in \mathbb{R}^h \quad (1 \le i \le N)$

$\mathbf{x} = \mathrm{Enc}(x) \in \mathbb{R}^h$

Voldemort had raised his wand … and a flash of

A neural retrieval problem

$\left(\text{Passages in the datastore: } z_1, \cdots, z_N, \text{ Input: } x\right) \longrightarrow z_i \quad (1 \le i \le N)$



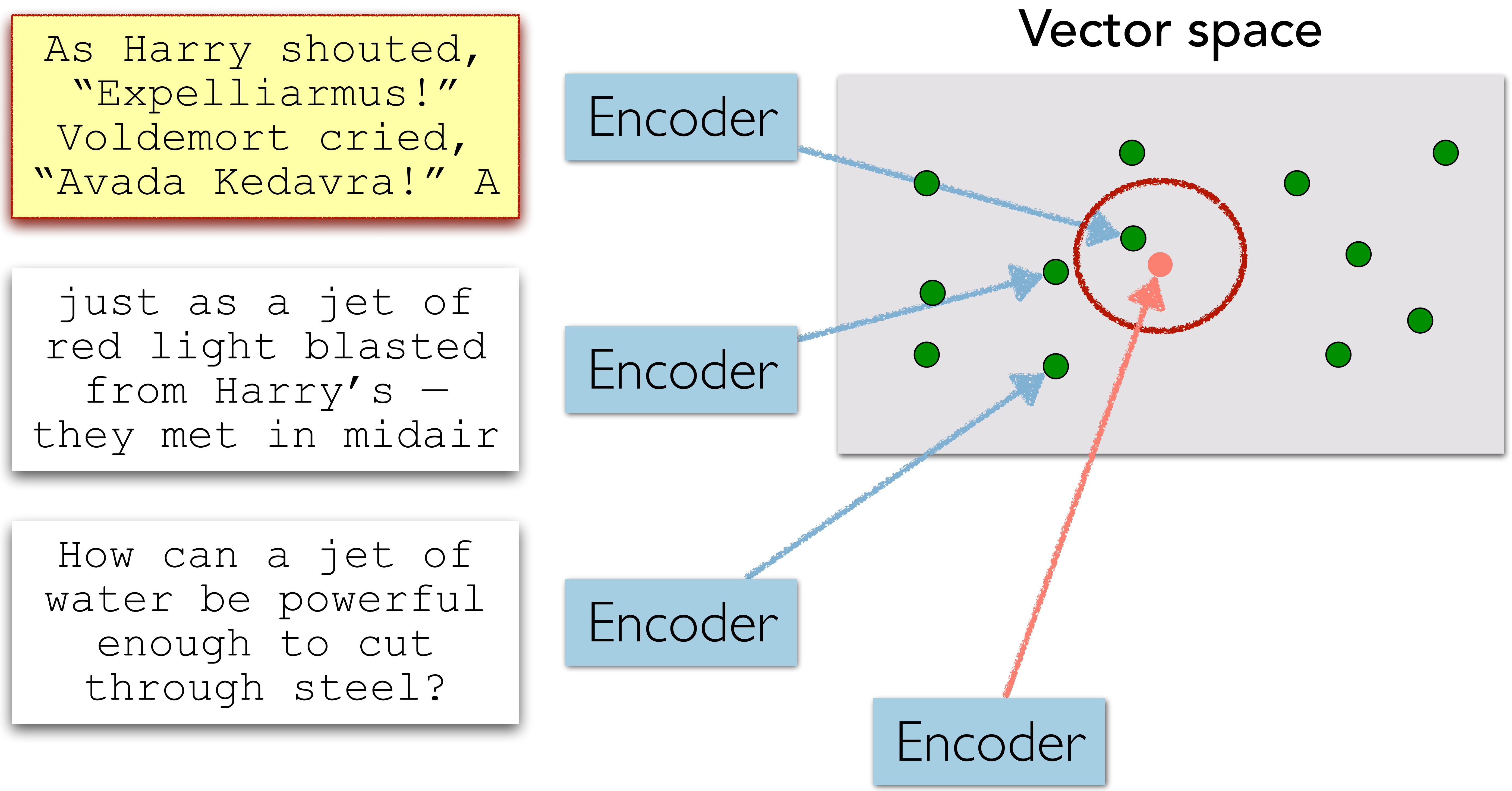
$$\mathbf{z}_i = \mathrm{Enc}(z_i) \in \mathbb{R}^h \quad (1 \le i \le N)$$

$$\mathbf{x} = \mathrm{Enc}(x) \in \mathbb{R}^h$$

$$\mathrm{sim}(x, z_i) = \mathbf{x}^\top \mathbf{z}_i$$

A neural retrieval problem

$\big($Passages in the datastore: $z_1, \cdots, z_N$, Input: $x\big) \longrightarrow z_i \quad (1 \leq i \leq N)$



$$\mathbf{z}_i = \mathrm{Enc}(z_i) \in \mathbb{R}^h \quad (1 \leq i \leq N)$$

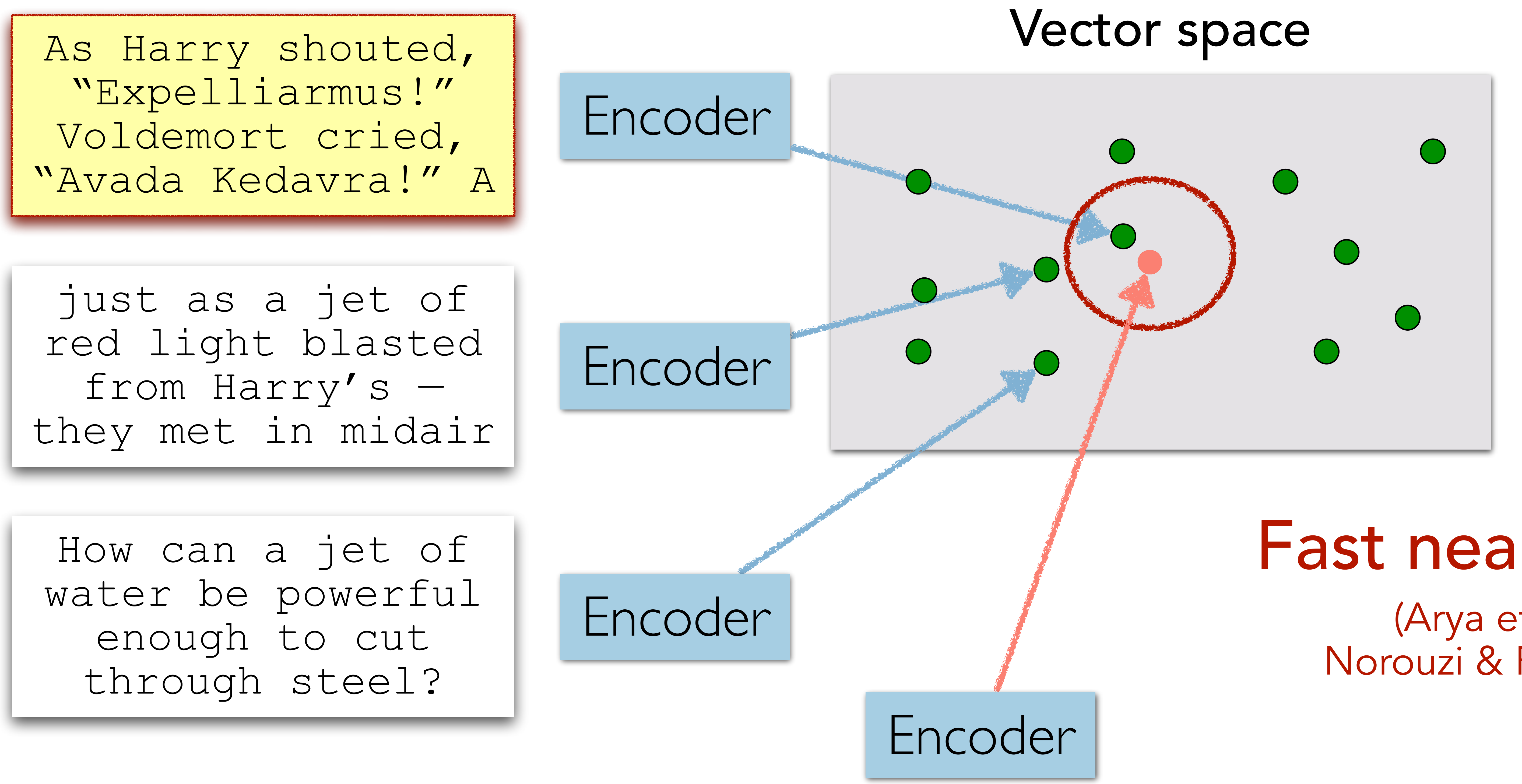
$$\mathbf{x} = \mathrm{Enc}(x) \in \mathbb{R}^h$$

$$\mathrm{sim}(x, z_i) = \mathbf{x}^\top \mathbf{z}_i$$

Voldemort had raised his wand … and a flash of

A neural retrieval problem

$\Big($Passages in the datastore: $z_1, \cdots, z_N$, Input: $x\Big) \longrightarrow z_i \quad (1 \le i \le N)$



$\mathbf{z}_i = \mathrm{Enc}(z_i) \in \mathbb{R}^h \quad (1 \le i \le N)$

$\mathbf{x} = \mathrm{Enc}(x) \in \mathbb{R}^h$

$\mathrm{sim}(x, z_i) = \mathbf{x}^\top \mathbf{z}_i$

Fast nearest neighbor search

(Arya et al. 1998, Dong et al. 2011, Norouzi & Fleet 2013, Johnson et al. 2017)

```
Voldemort had raised his wand … and a flash of
```

A neural retrieval problem: **in 2019**

A neural retrieval problem: **in 2019**

No good recipe for training the encoder

- Required massive compute & labeled dataset
- Not much better than alternatives (e.g. lexical-matching)

A neural retrieval problem: **in 2019**

No good recipe for training the encoder

- Required massive compute & labeled dataset
- Not much better than alternatives (e.g. lexical-matching)

Only in 2020, neural retrieval began its era

- Advent of pre-trained encoders such as BERT
- Development of improved learning objectives (next slide)

Contrastive learning

Contrastive learning

x : input

Contrastive learning

x : input z^+: a positive passage to x (typically given)

Contrastive learning

x : input z^+: a positive passage to x (typically given)

$$-\log \frac{e^{\mathrm{sim}(x,z^+)}}{\sum_{i=1}^{N} e^{\mathrm{sim}(x,z_i)}}$$

Contrastive learning

x : input z^+: a positive passage to x (typically given)

$$-\log \frac{e^{\mathrm{sim}(x,z^+)}}{\sum_{i=1}^{N} e^{\mathrm{sim}(x,z_i)}}$$

Get close to a **positive**

Contrastive learning

x : input z^+: a positive passage to x (typically given)

$$-\log \frac{e^{\mathrm{sim}(x,z^+)}}{\sum_{i=1}^{N} e^{\mathrm{sim}(x,z_i)}}$$

Get close to a **positive**

Push back **all the others**

Contrastive learning

x : input z^+: a positive passage to x (typically given)

$$-\log \frac{e^{\text{sim}(x,z^+)}}{\boxed{\sum_{i=1}^{N} e^{\text{sim}(x,z_i)}}}$$

Get close to a **positive**

Push back **all the others**

Contrastive learning

x : input $\quad\quad z^+$: a positive passage to x (typically given)

$$-\log \frac{e^{\mathrm{sim}(x,z^+)}}{\sum_{i=1}^{N} e^{\mathrm{sim}(x,z_i)}} \approx$$

1) In-batch approximation

Yih et al., 2011, Henderson et al., 2017, Gillick et al., 2019

Contrastive learning

x : input $\qquad$ z^+: a positive passage to x (typically given)

$$-\log \frac{e^{\text{sim}(x,z^+)}}{\sum_{i=1}^{N} e^{\text{sim}(x,z_i)}} \approx -\log \frac{e^{\text{sim}(x,z^+)}}{e^{\text{sim}(x,z^+)} + \sum_{i=1}^{n} e^{\text{sim}(x,z_i^-)}}$$

1) In-batch approximation

Other passages within the batch

Yih et al., 2011, Henderson et al., 2017, Gillick et al., 2019

Contrastive learning

x : input z^+: a positive passage to x (typically given)

$$-\log \frac{e^{\mathrm{sim}(x,z^+)}}{\sum_{i=1}^{N} e^{\mathrm{sim}(x,z_i)}} \approx -\log \frac{e^{\mathrm{sim}(x,z^+)}}{e^{\mathrm{sim}(x,z^+)} + \sum_{i=1}^{n} e^{\mathrm{sim}(x,z_i^-)}}$$

1) In-batch approximation

Negatives in the batch

Yih et al., 2011, Henderson et al., 2017, Gillick et al., 2019

Contrastive learning

x : input z^+: a positive passage to x (typically given)

$$-\log \frac{e^{\mathrm{sim}(x,z^+)}}{\sum_{i=1}^{N} e^{\mathrm{sim}(x,z_i)}} \approx -\log \frac{e^{\mathrm{sim}(x,z^+)}}{e^{\mathrm{sim}(x,z^+)} + \sum_{i=1}^{n} e^{\mathrm{sim}(x,z_i^-)}}$$

1) In-batch approximation

2) **Hard negatives** in the batch

Passages that ***challenge*** the model,
typically obtained by passages with high lexical overlap
(Karpukhin et al. 2020)

Retrieval vs. Parametric-only

Retrieval vs. Parametric-only

2016–2019: Retrieval-based models/LMs by default

(But mostly based on lexical matching retrieval)

Retrieval vs. Parametric-only

2016–2019: Retrieval-based models/LMs by default

(But mostly based on lexical matching retrieval)

February 2020: “Parametric-only LMs” might be better?!

Roberts et al. “How Much Knowledge Can You Pack into the Parameters of a Language Model?”

Retrieval vs. Parametric-only

2016–2019: Retrieval-based models/LMs by default

(But mostly based on lexical matching retrieval)

February 2020: "Parametric-only LMs" might be better?!

Roberts et al. "How Much Knowledge Can You Pack into the Parameters of a Language Model?"

April 2020: But now, neural retrieval is much better!

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

Retrieval vs. Parametric-only

2016–2019: Retrieval-based models/LMs by default

(But mostly based on lexical matching retrieval)

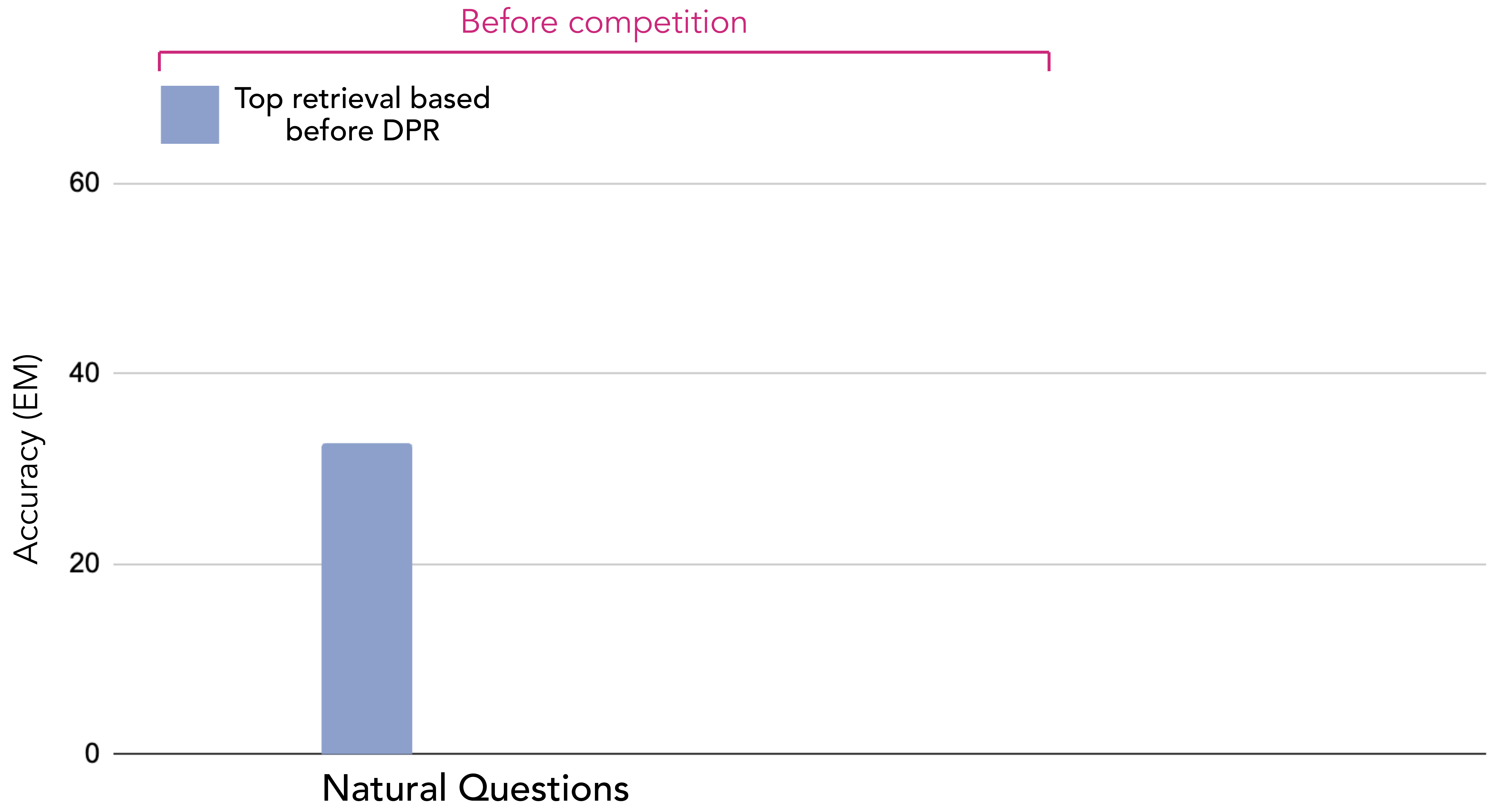
February 2020: "Parametric-only LMs" might be better?!

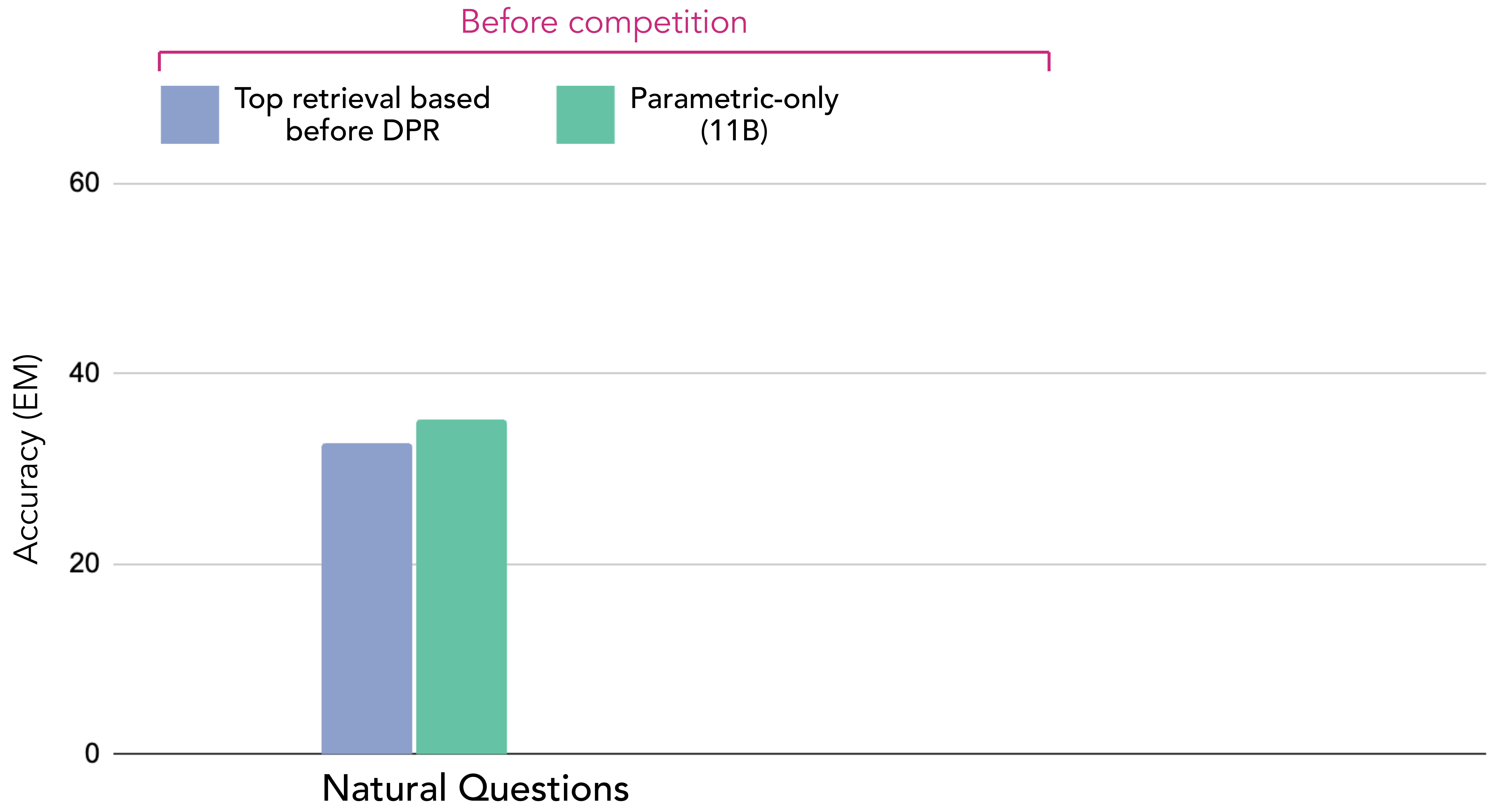
Roberts et al. "How Much Knowledge Can You Pack into the Parameters of a Language Model?"

April 2020: But now, neural retrieval is much better!

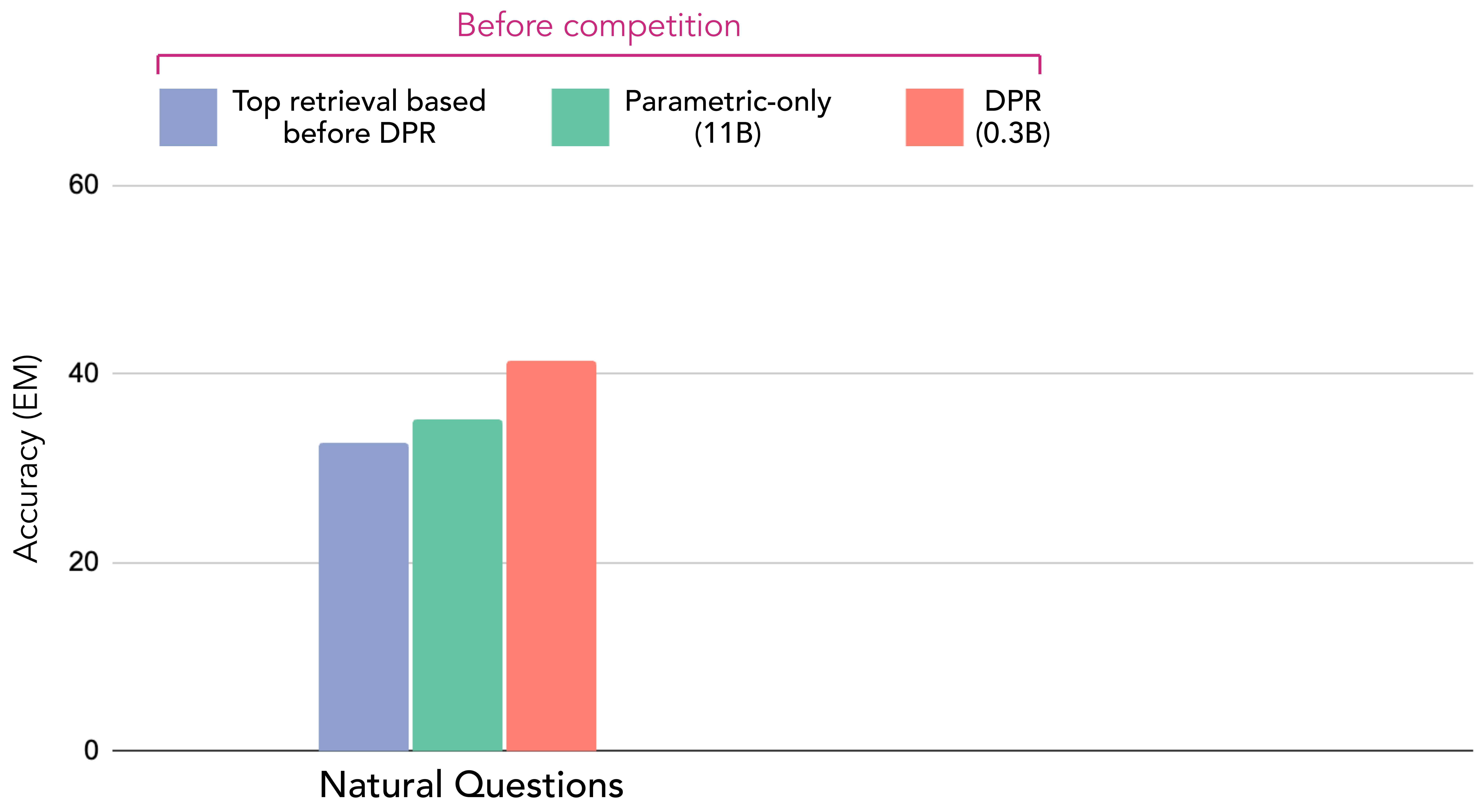
Karpukhin et al. "Dense Passage Retrieval for Open-Domain Question Answering"

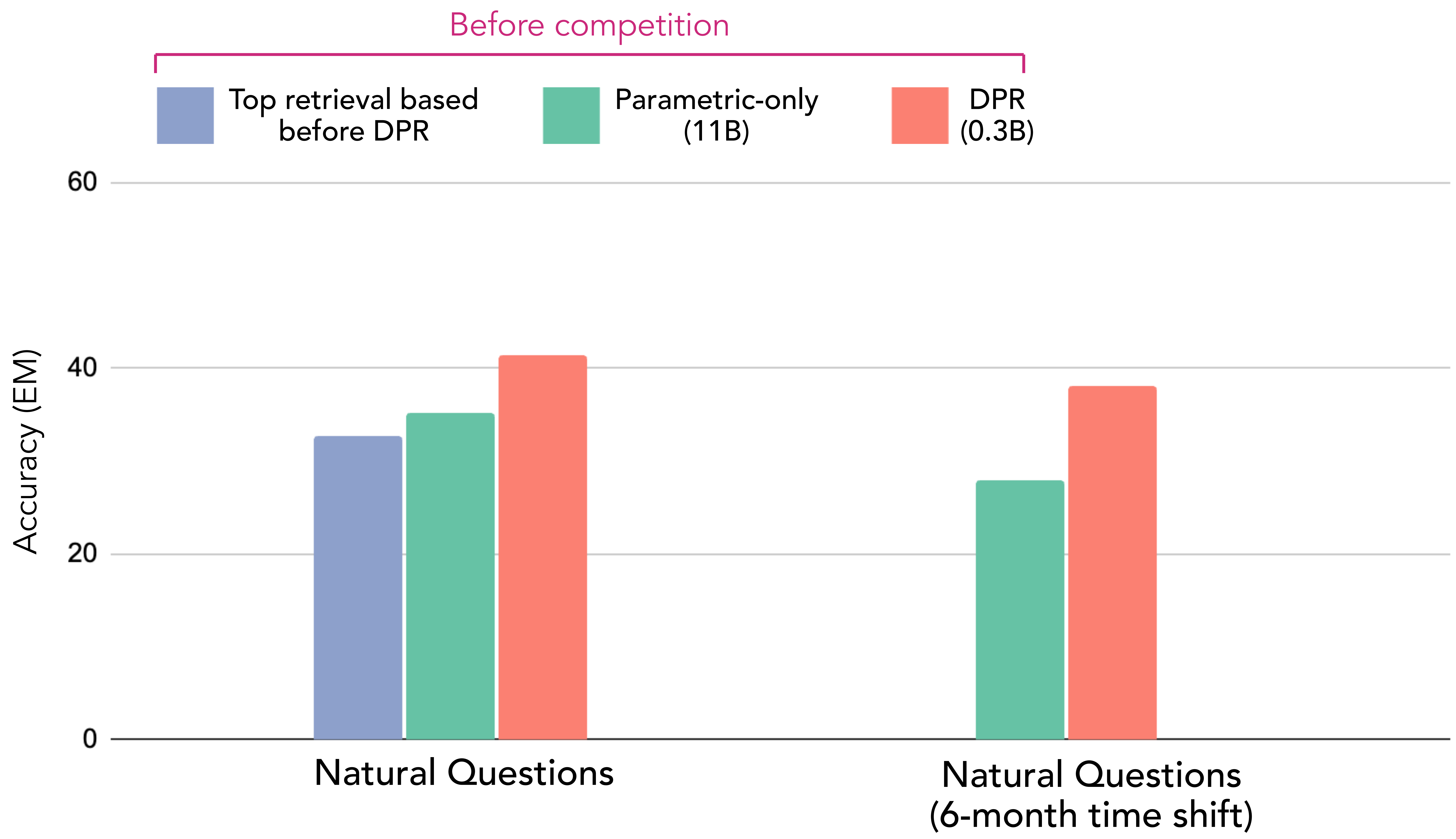
Summer 2020: A NeurIPS 2020 Competition!

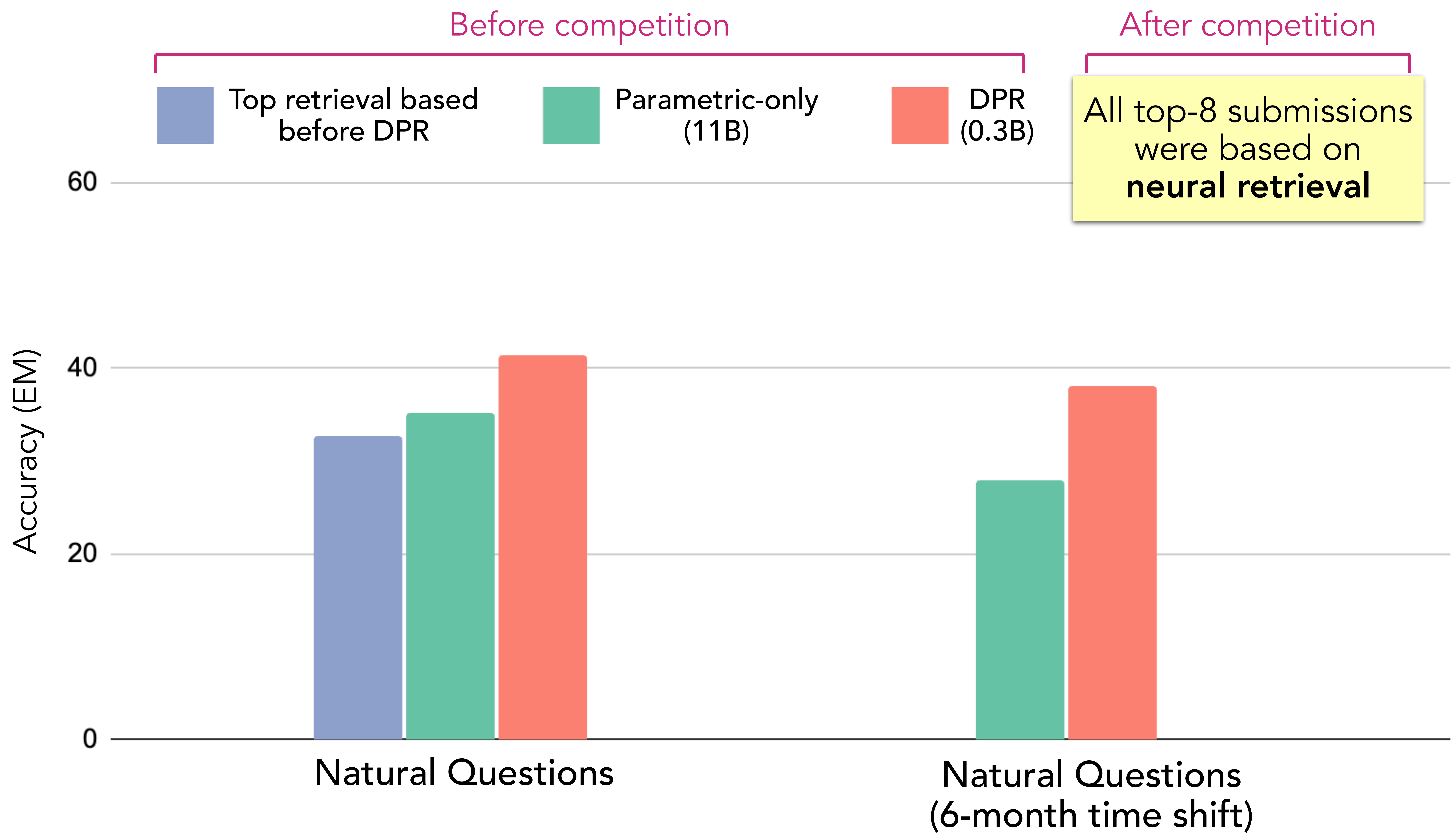

Before competition
Top retrieval based before DPR
60
40
20
0
Accuracy (EM)
Natural Questions

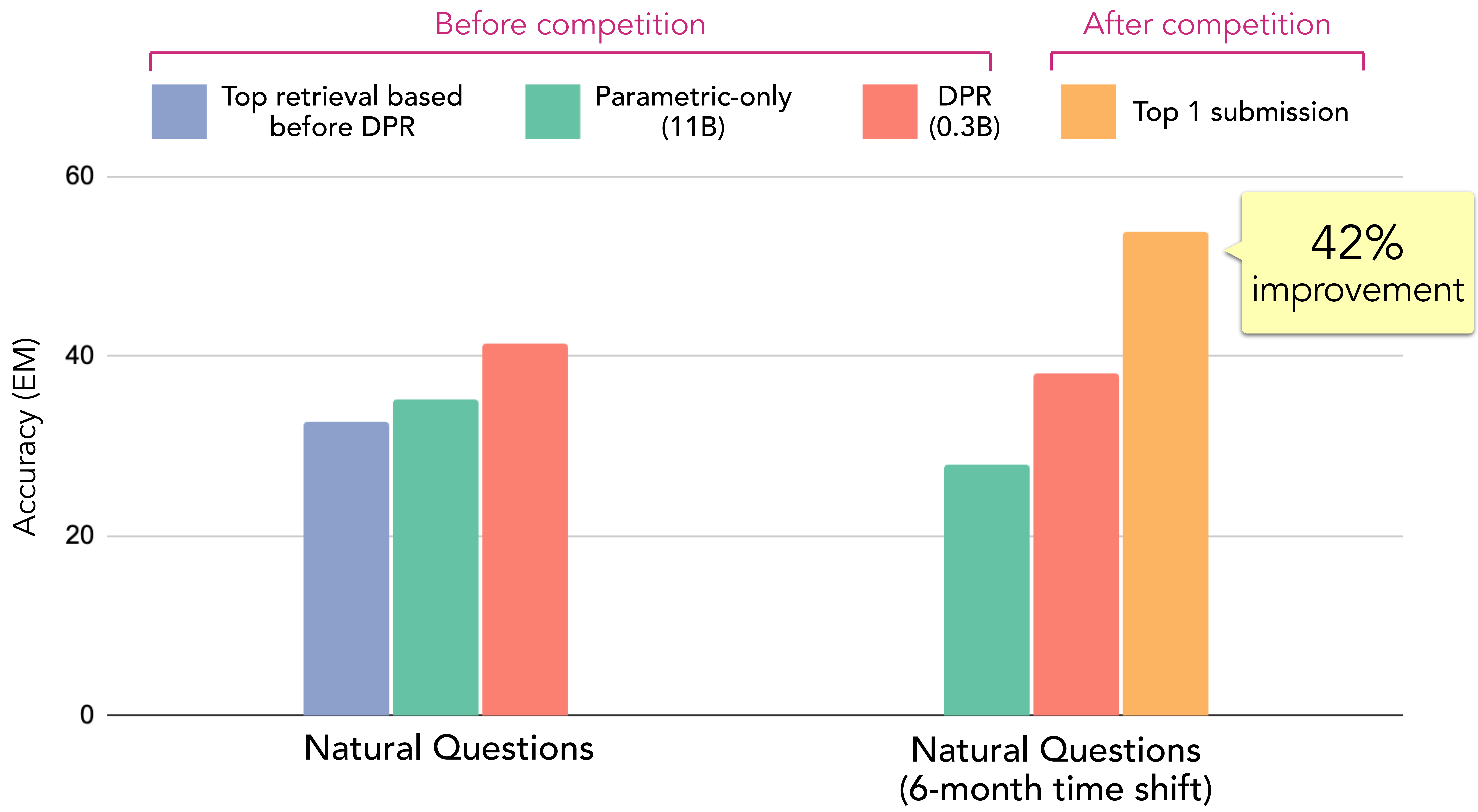


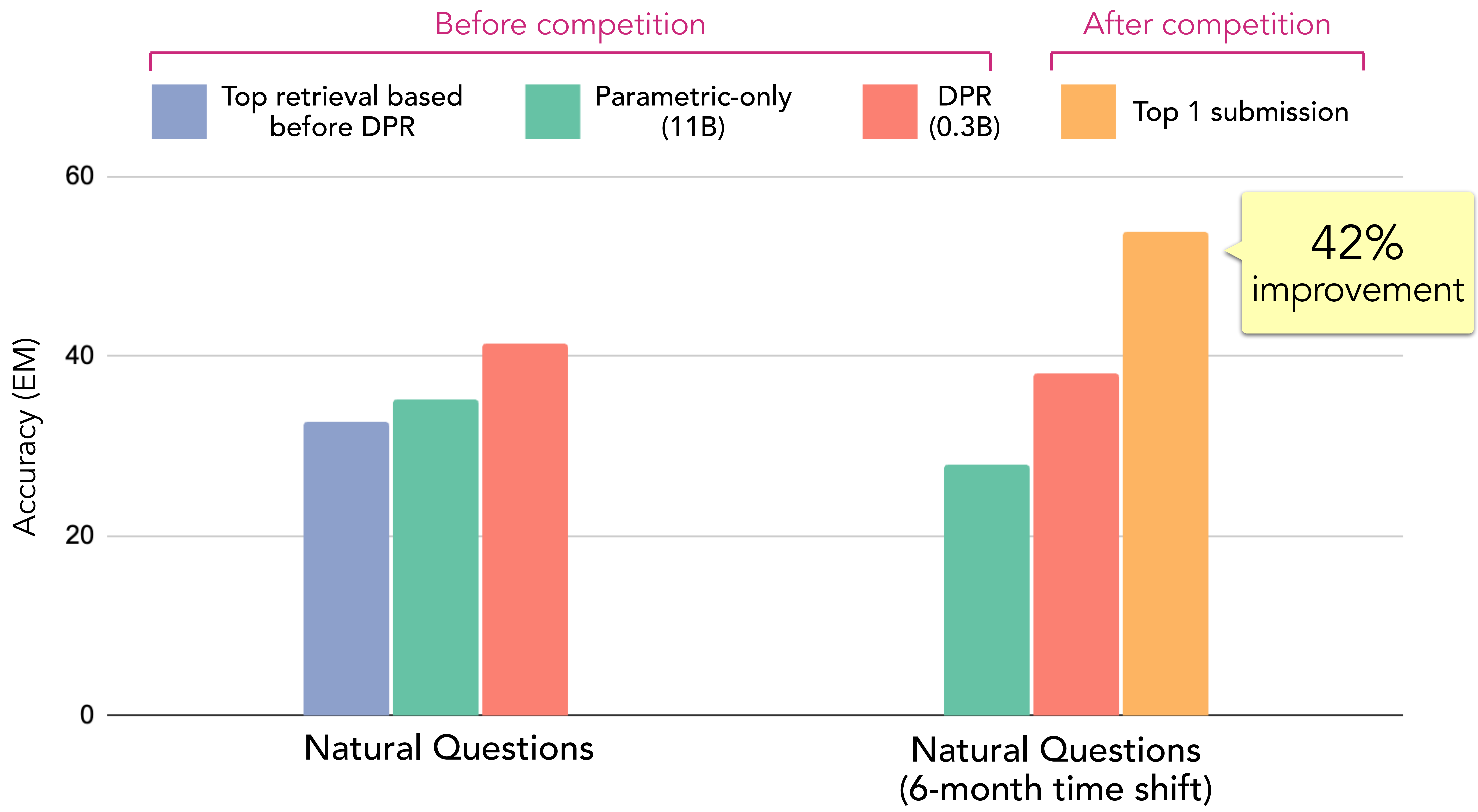
Before competition
Top retrieval based before DPR
Parametric-only (11B)
60
40
20
0
Accuracy (EM)
Natural Questions


Before competition
Top retrieval based before DPR
Parametric-only (11B)
DPR (0.3B)
60
40
20
0
Accuracy (EM)
Natural Questions


Before competition
Top retrieval based before DPR
Parametric-only (11B)
DPR (0.3B)
Accuracy (EM)
60
40
20
0
Natural Questions
Natural Questions (6-month time shift)

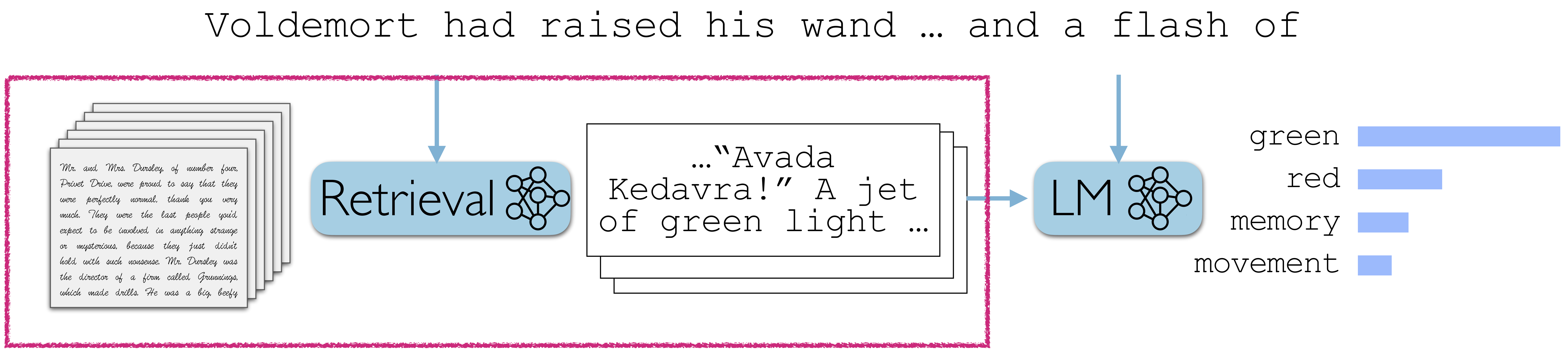

Before competition
After competition
Top retrieval based before DPR
Parametric-only (11B)
DPR (0.3B)
All top-8 submissions were based on **neural retrieval**
60
40
20
0
Accuracy (EM)
Natural Questions
Natural Questions (6-month time shift)


Before competition
After competition
Top retrieval based before DPR
Parametric-only (11B)
DPR (0.3B)
Top 1 submission
60
40
20
0
Accuracy (EM)
42% improvement
Natural Questions
Natural Questions (6-month time shift)



Takeaway: The quality of retrieval-based LMs depends on the quality of **retrieval**

A two-stage pipeline

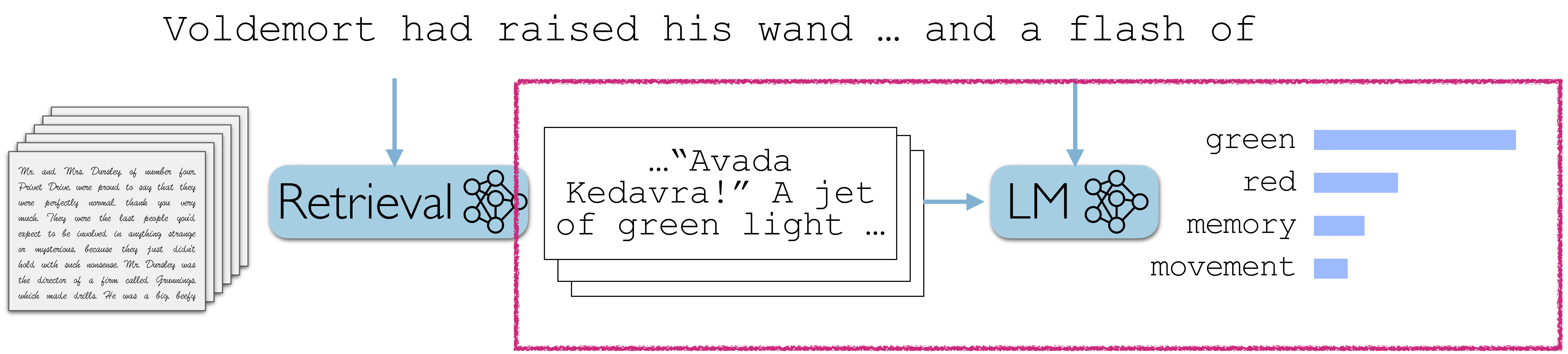


1) Retrieval

2) Augmentation

3) Training

A two-stage pipeline



1) Retrieval

2) Augmentation

3) Training

Augmentation

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 …

Augmentation

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

+

Harry felt Greenback collapse against him … a jet of

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 …

Augmentation

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 …

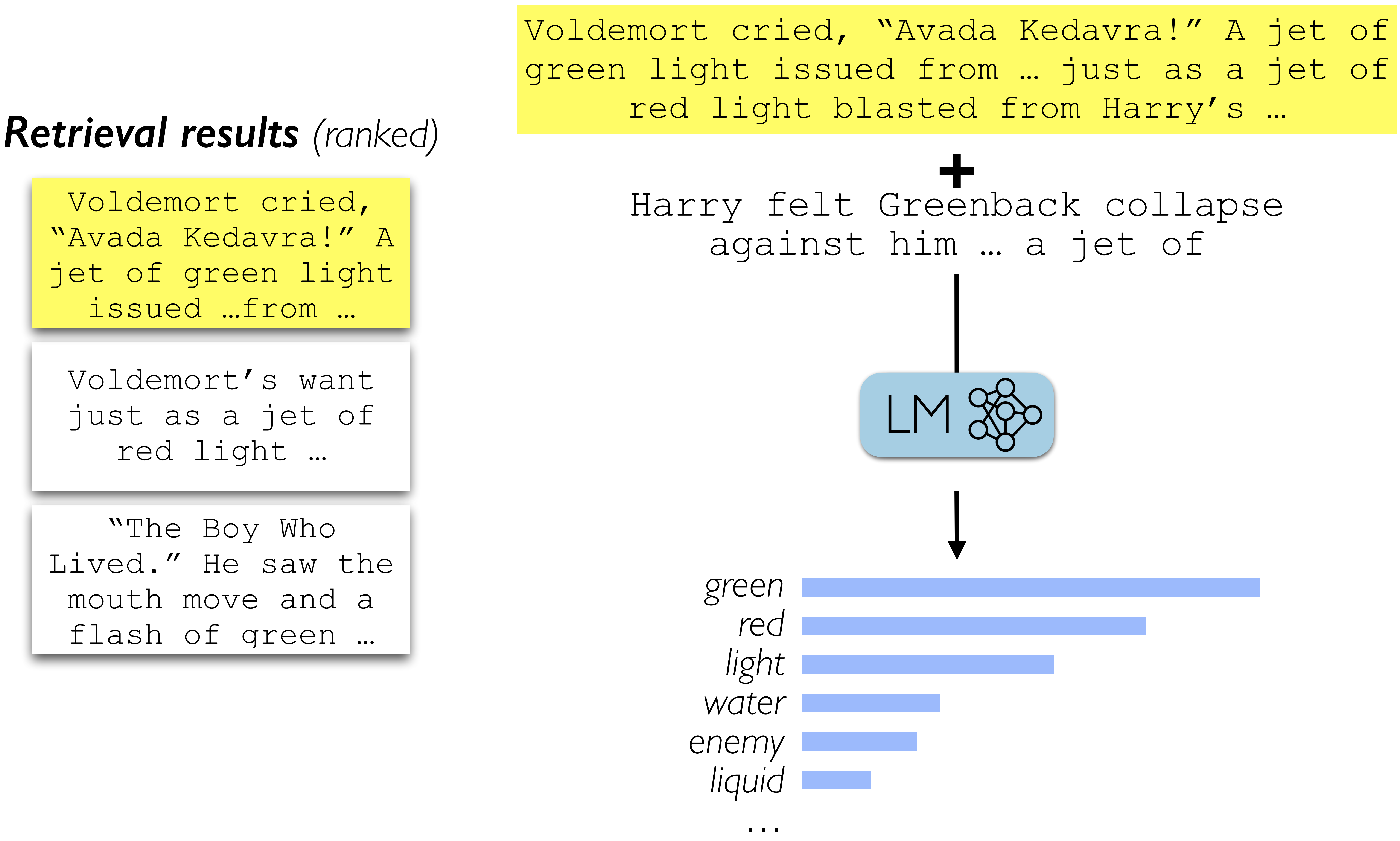
Voldemort cried, "Avada Kedavra!" A jet of green light issued from … just as a jet of red light blasted from Harry's …

\+

Harry felt Greenback collapse against him … a jet of

LM

Augmentation



Augmentation

***Retrieval results** (ranked)*

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

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

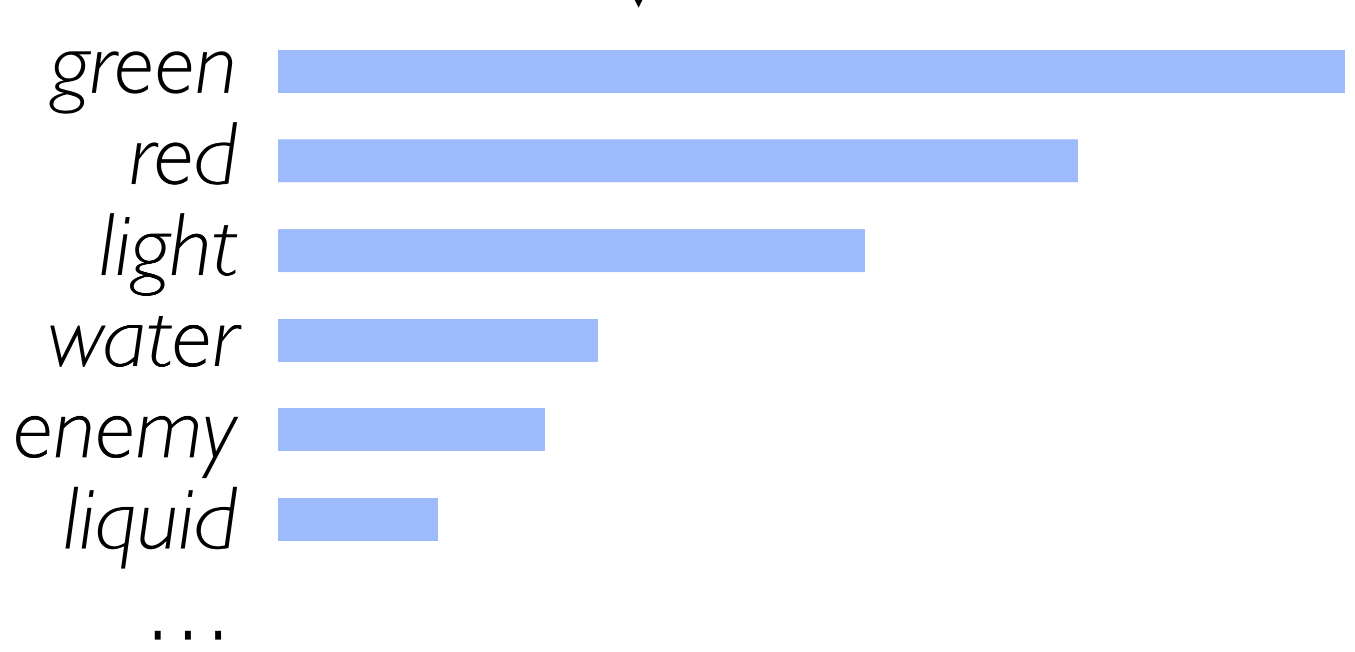
"The Boy Who Lived." He saw the mouth move and a flash of green …

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

Augmentation

Retrieval results *(ranked)*

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

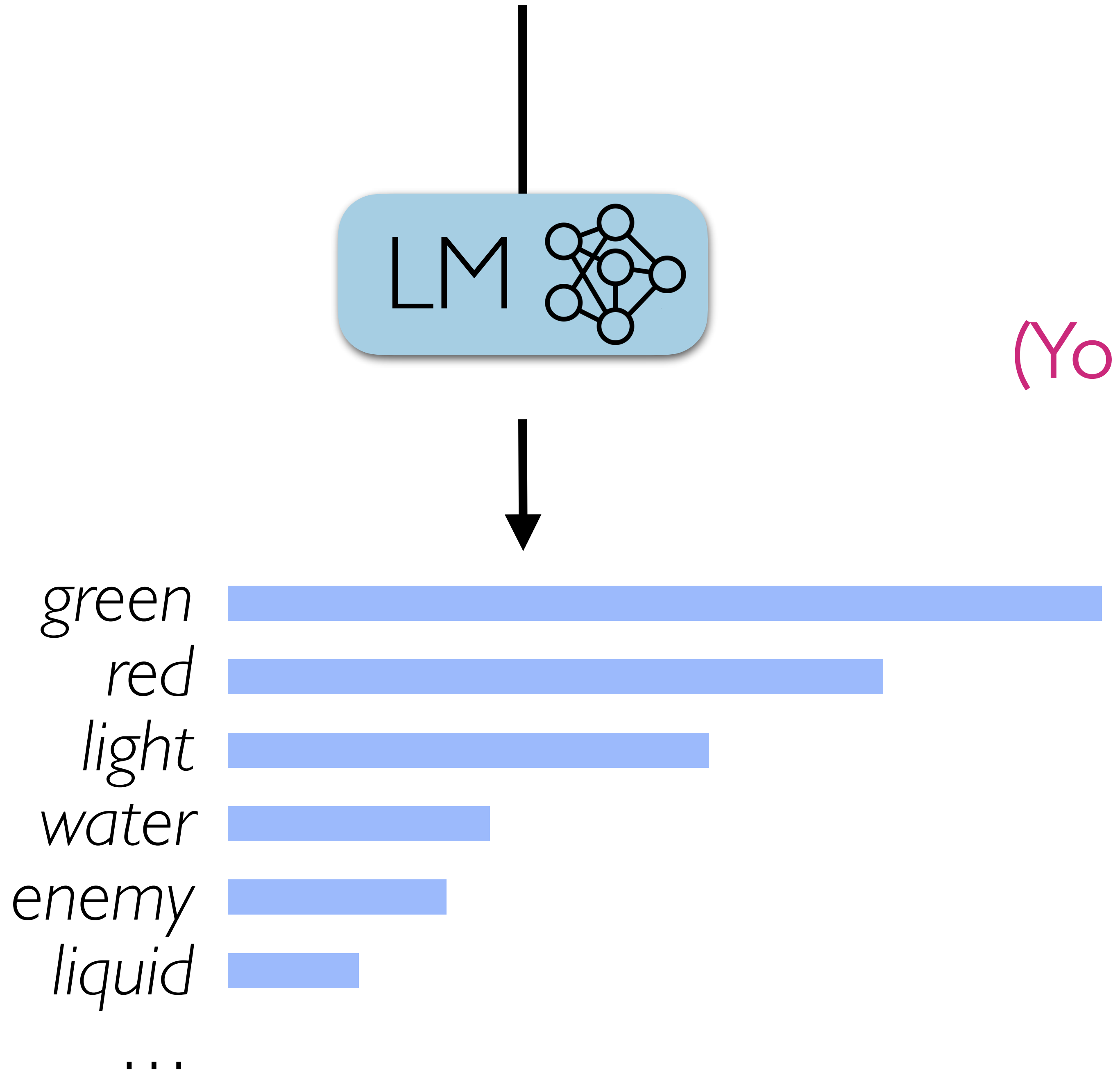
Voldemort's want just as a jet of red light …

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

+

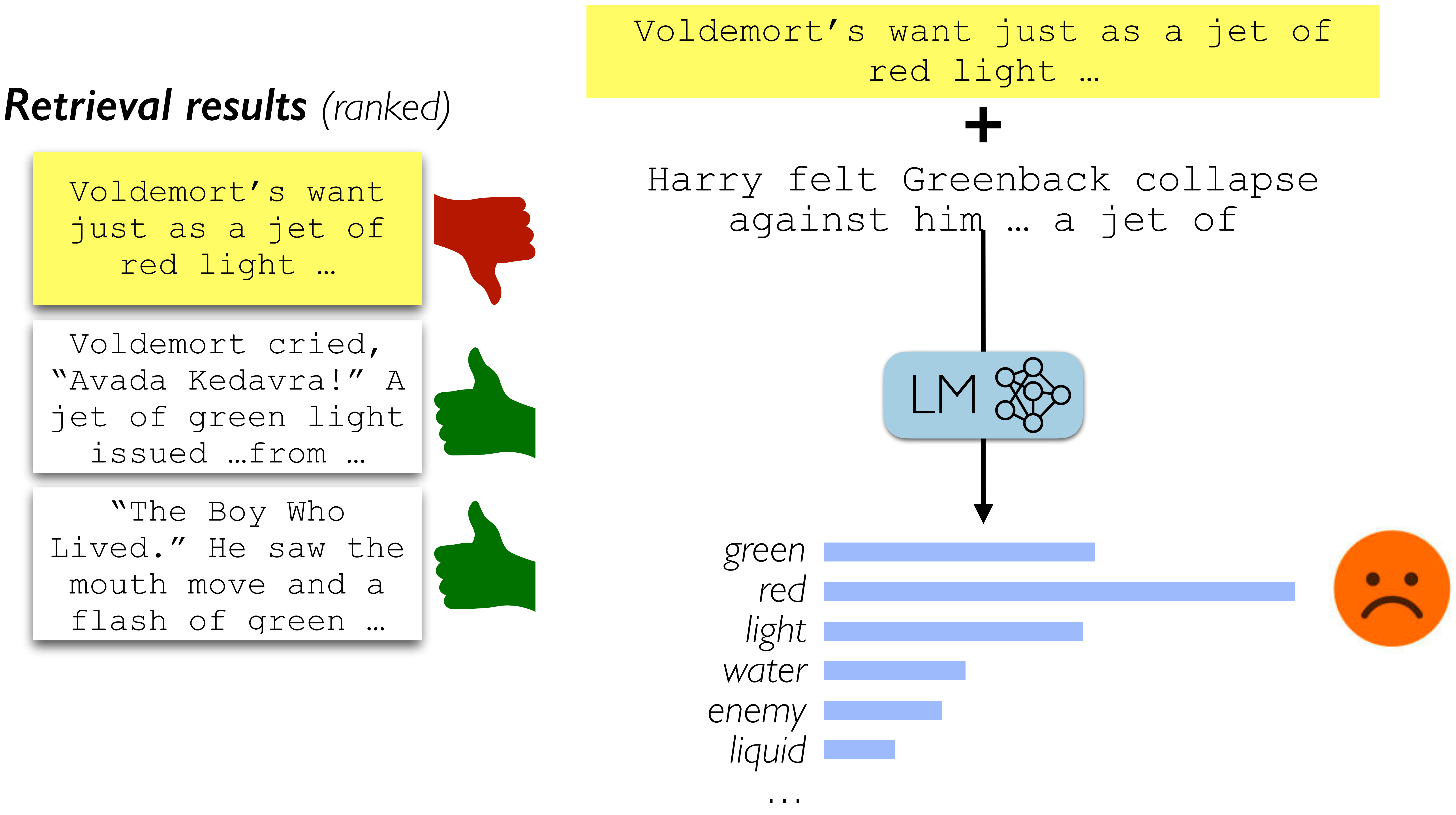
Harry felt Greenback collapse against him … a jet of



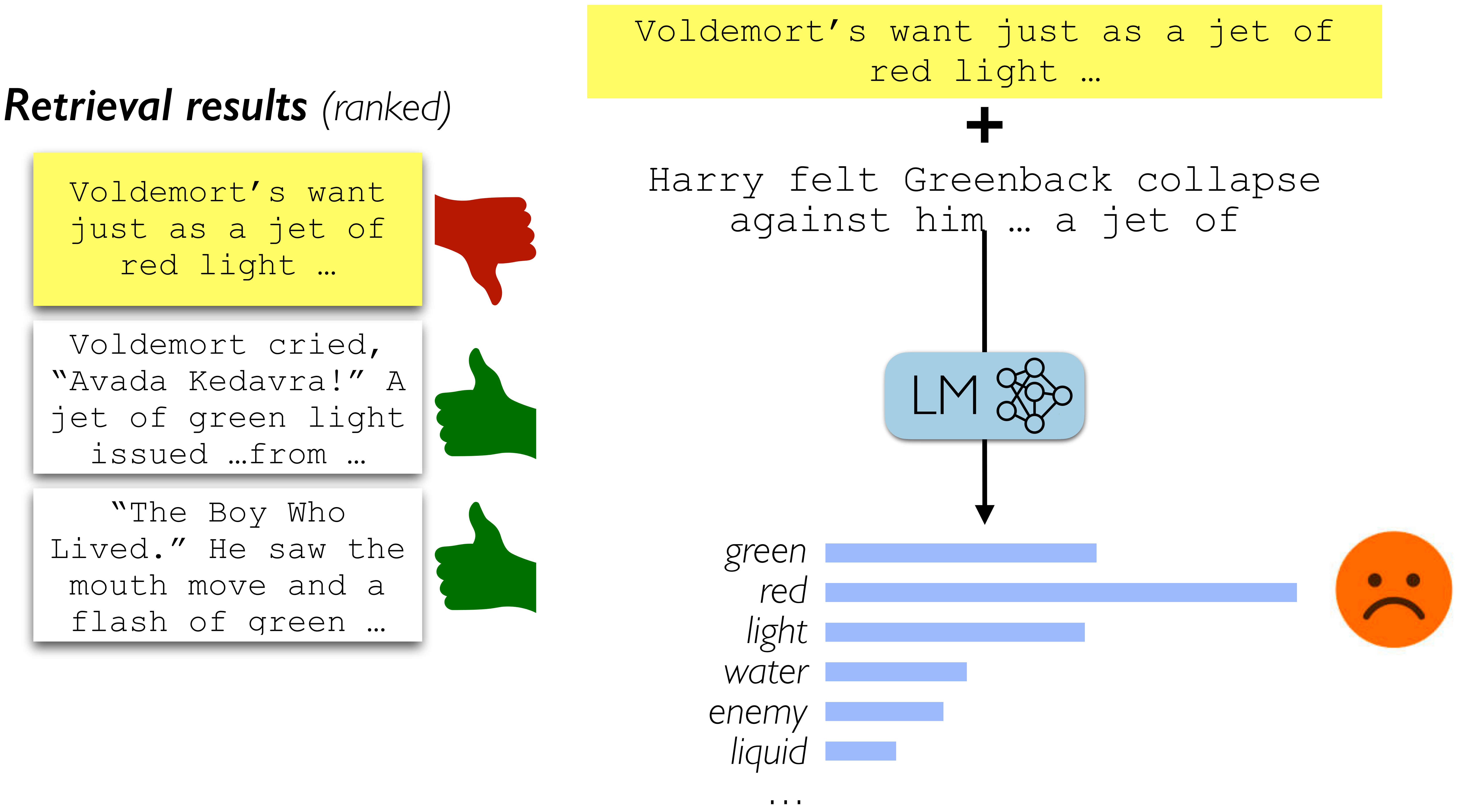
Very simple

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

Augmentation



Augmentation



Q: How do we use multiple passages?

Augmentation (1): Concatenation

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

Augmentation (1): Concatenation

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

+

Harry felt Greenback collapse against him … a jet of

Augmentation (1): Concatenation

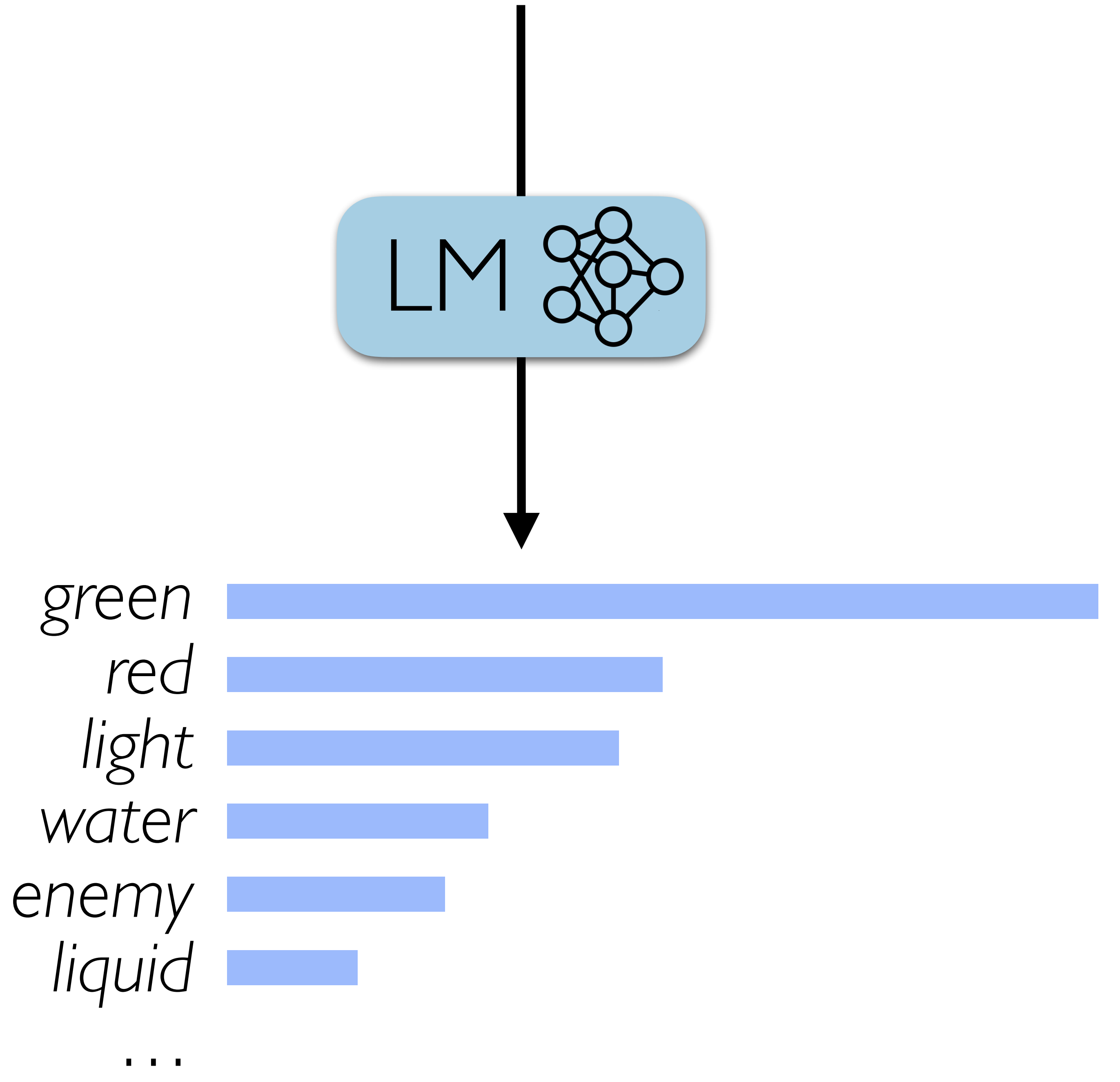
Voldemort's want just as a jet of red light …

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

… He saw the mouth move and a flash of green …

+

Harry felt Greenback collapse against him … a jet of

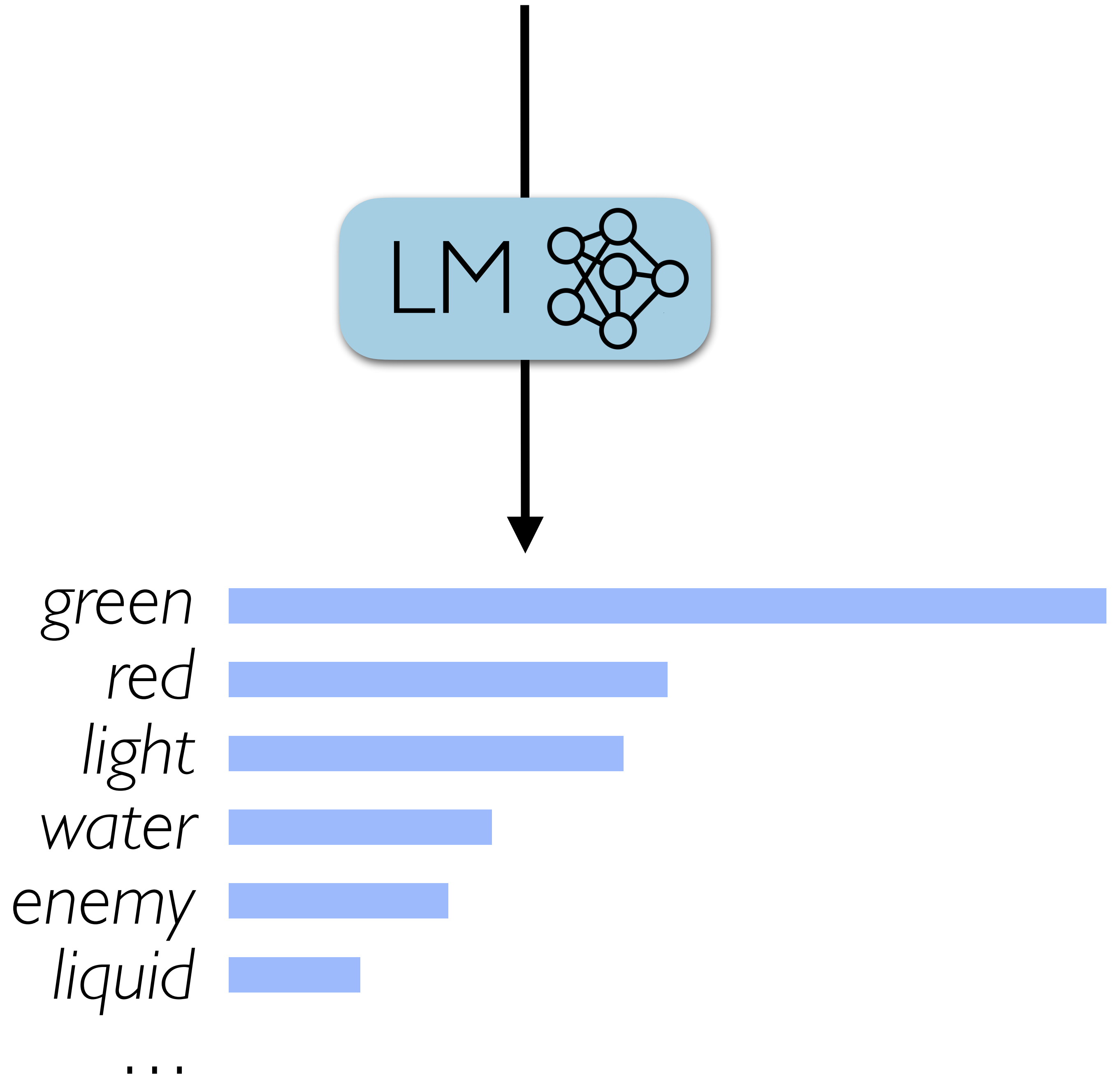


Augmentation (1): Concatenation

```
Voldemort's want just as a jet of red light …
… "Avada Kedavra!" A jet of green light issued …from …
… He saw the mouth move and a flash of green …
```

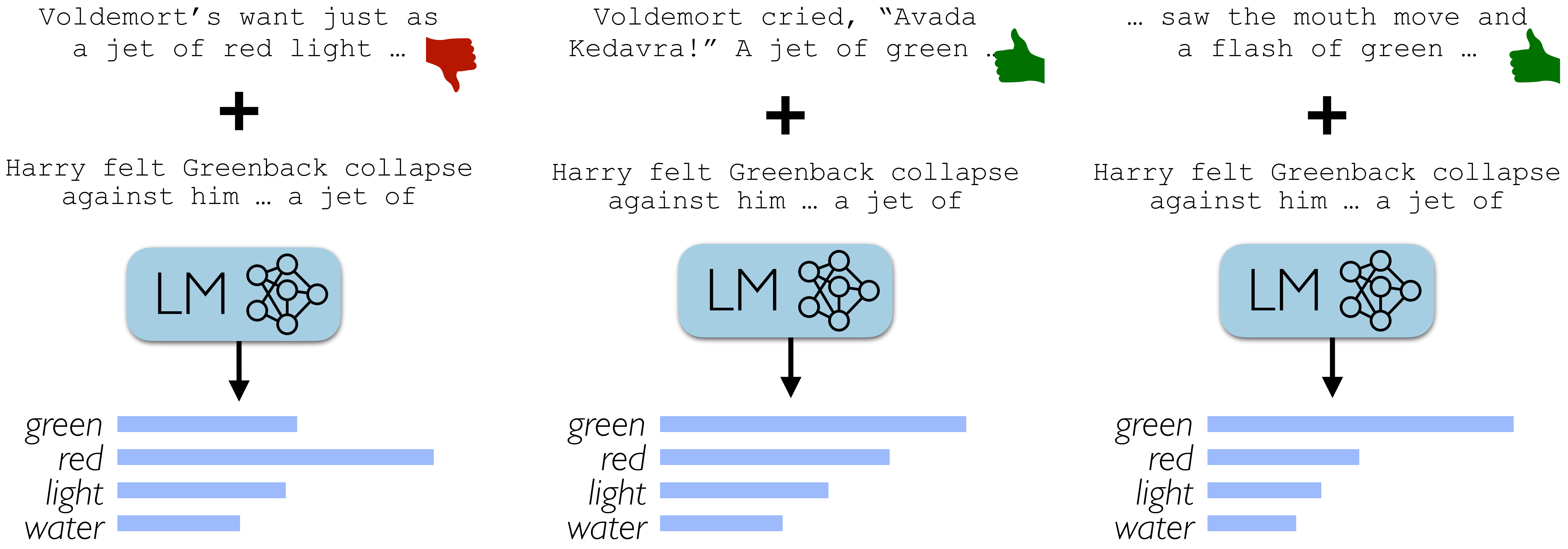
+

```
Harry felt Greenback collapse against him … a jet of
```

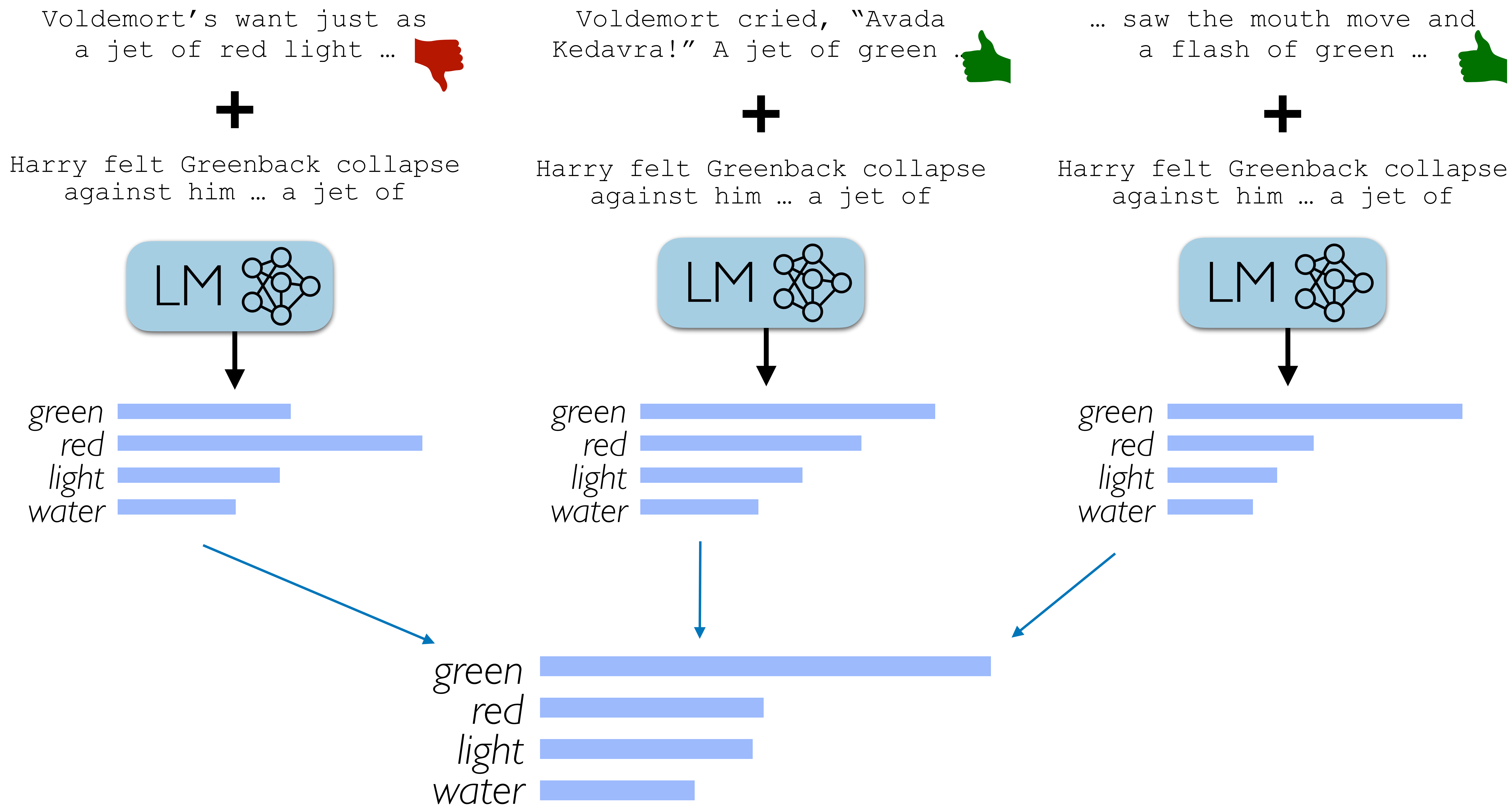


- Simple
- Increase the inference cost & Bounded by the maximum length limit of the LM

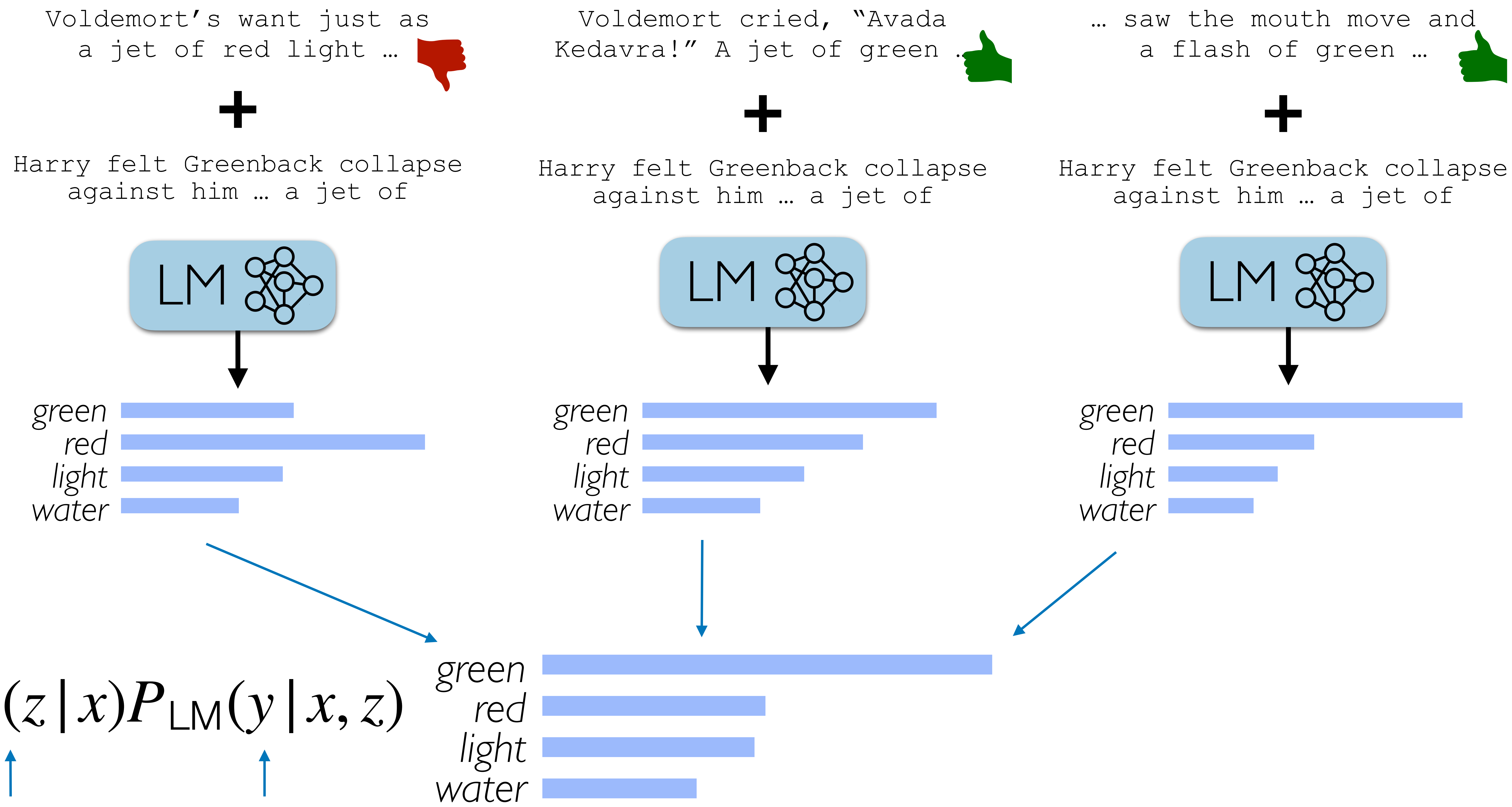
Augmentation (2): Ensembling



Augmentation (2): Ensembling



Augmentation (2): Ensembling

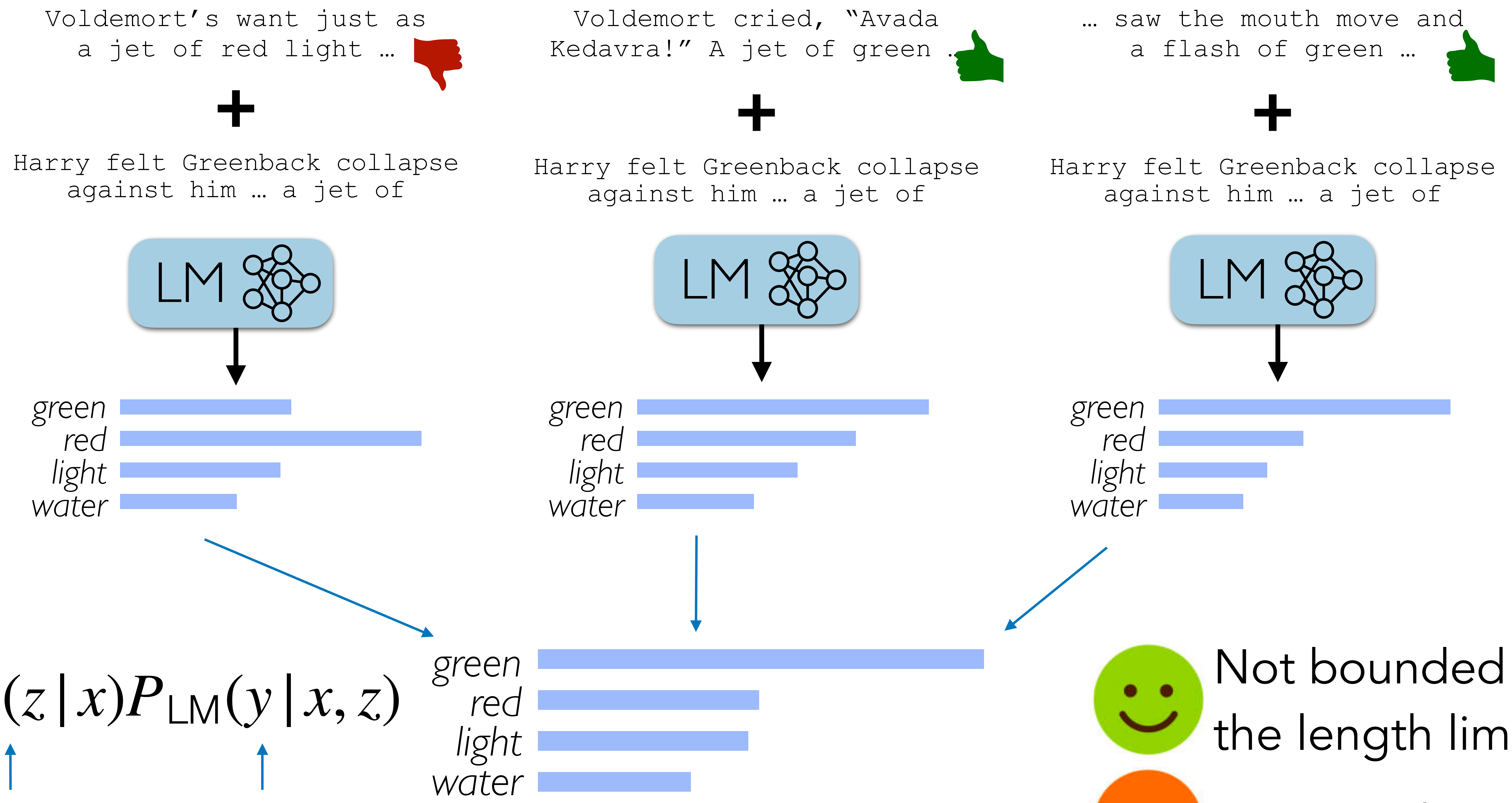


$$P(y \mid x) = \sum_{z \in \mathcal{Z}} P_{\text{ret}}(z \mid x) P_{\text{LM}}(y \mid x, z)$$

retrieval score

LM score

Augmentation (2): Ensembling



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

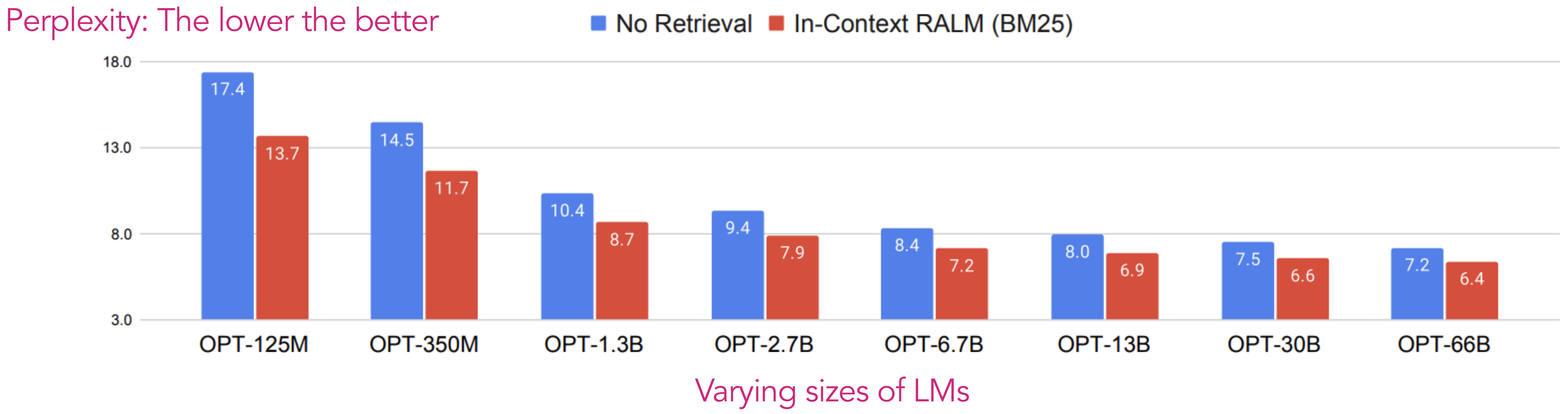
retrieval score

LM score

Not bounded by the length limit

Increase the inference cost

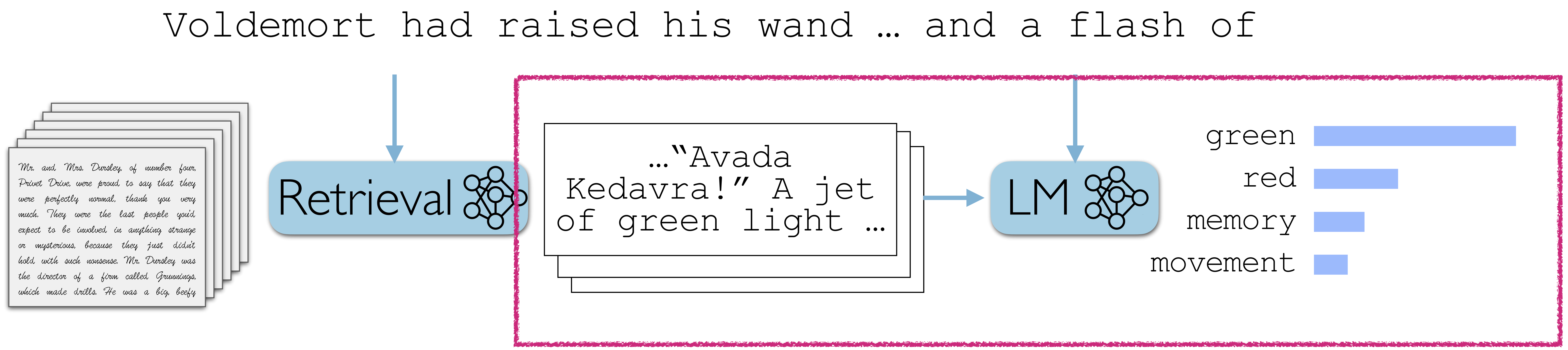
Results



Retrieval helps over all sizes of LMs

Graphs from Ram et al. 2023

A two-stage pipeline

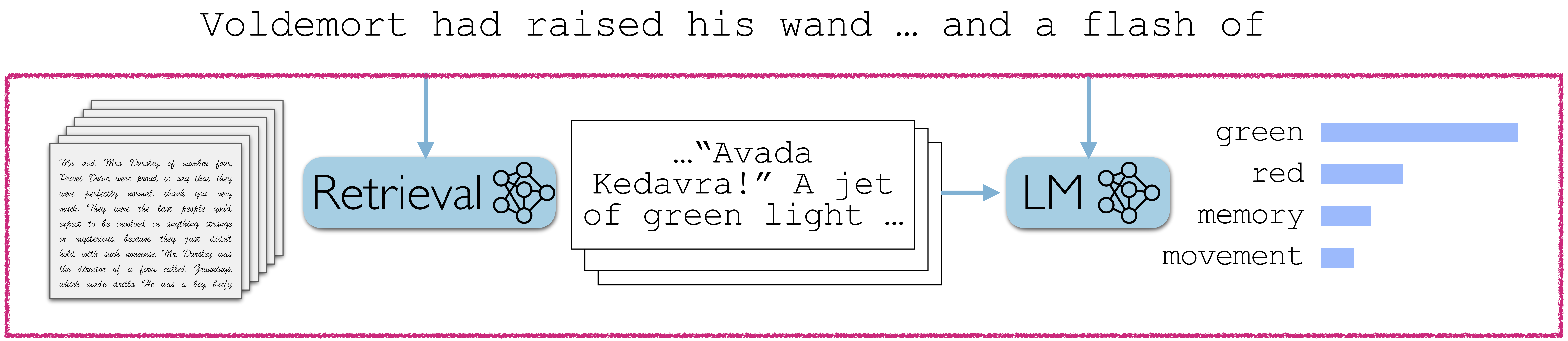


1) Retrieval

2) Augmentation

3) Training

A two-stage pipeline



1) Retrieval

2) Augmentation

3) Training

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-3/4, LLAMA, Qwen, etc…

How to train it?

Retrieval Model

trained in isolation

DPR, Contriever, GTR, GRIT…

LM

trained in isolation

GPT-3/4, LLAMA, Qwen, etc…

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

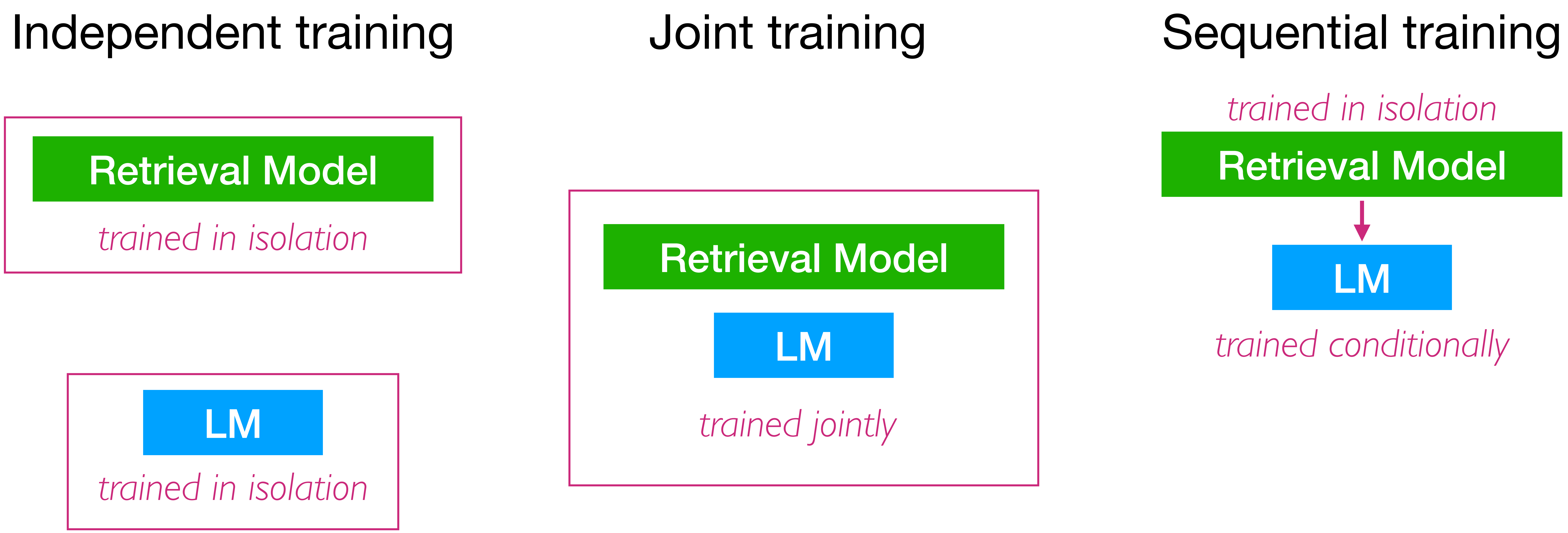
Joint training

Retrieval Model

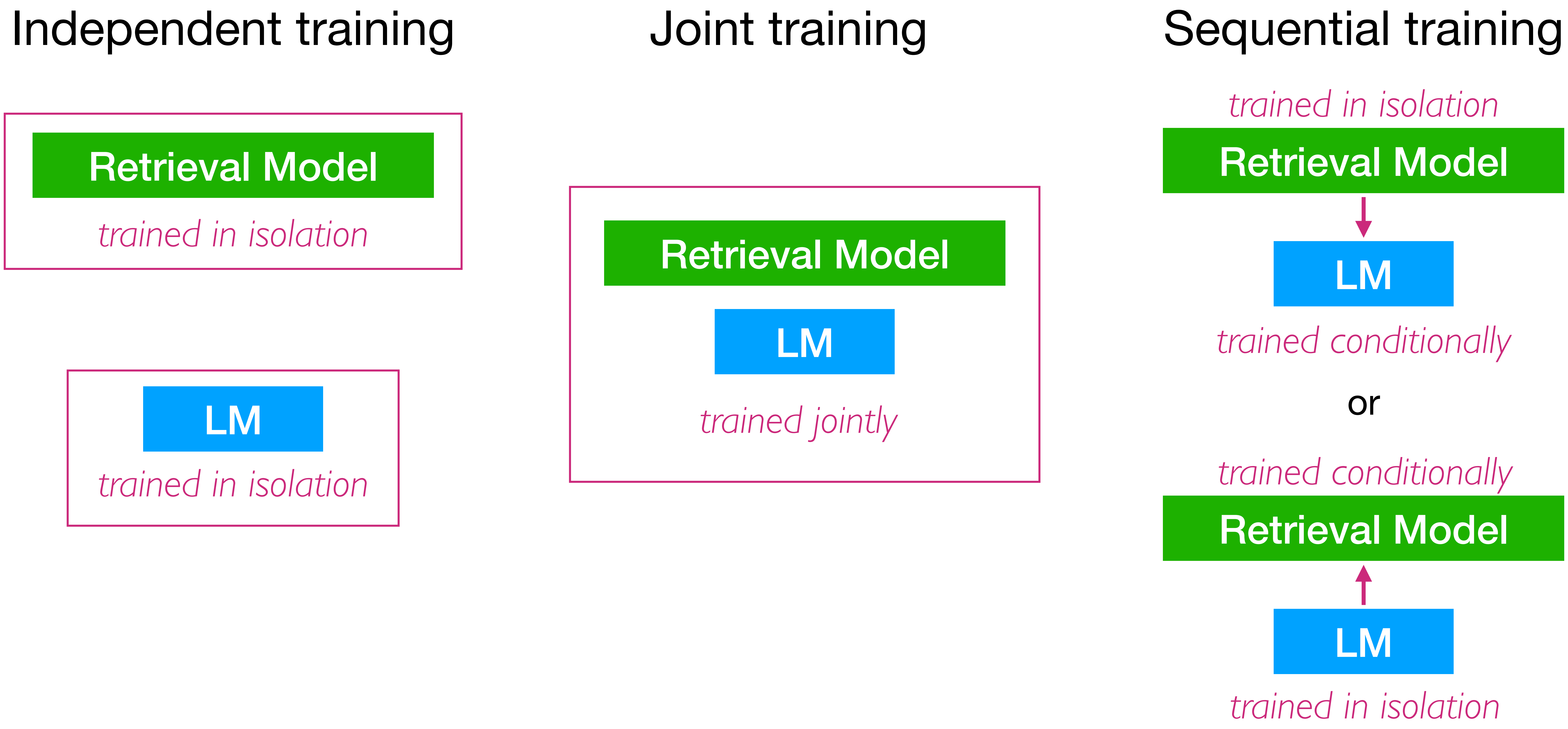
LM

trained jointly

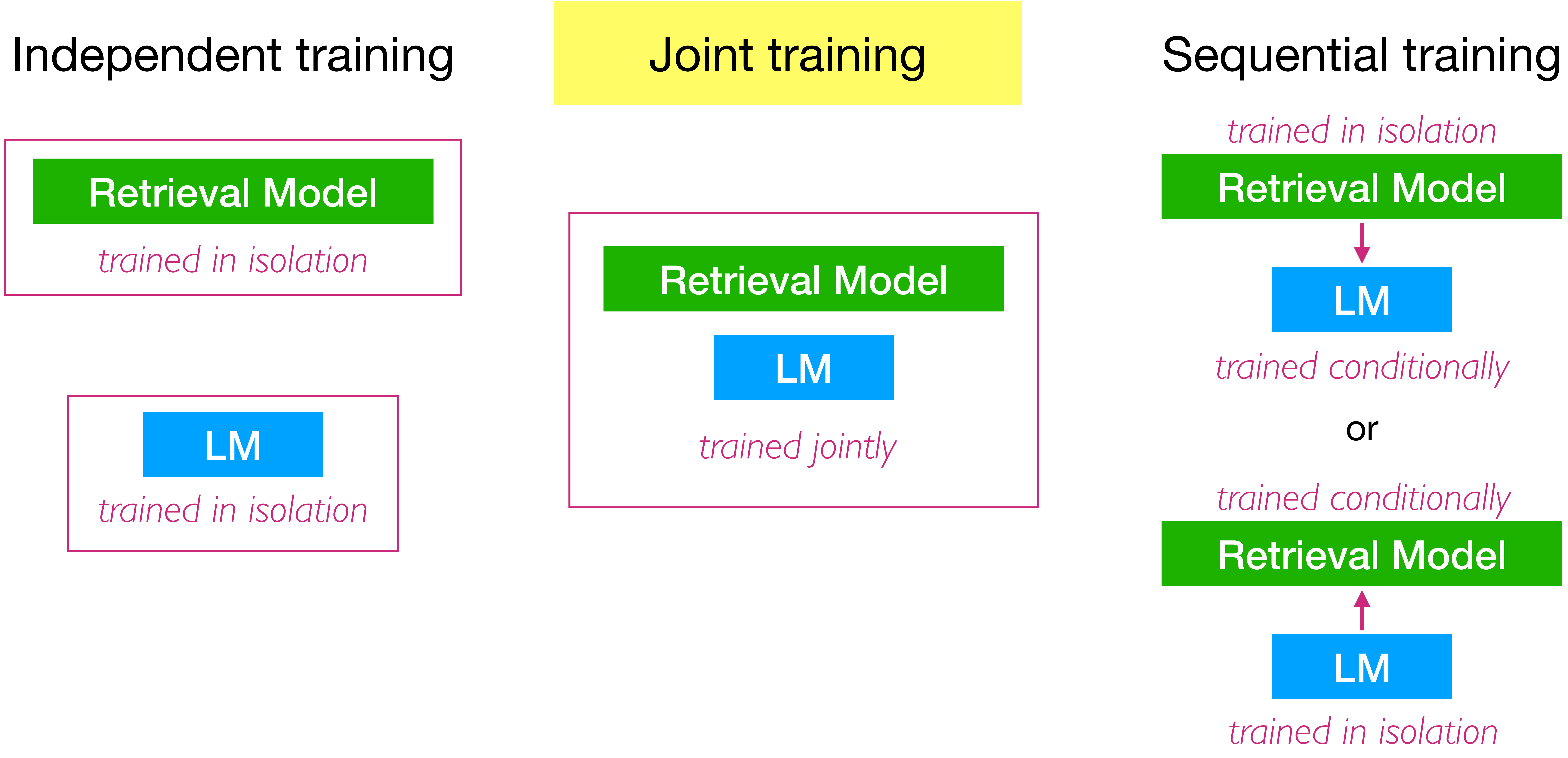
How to train it?



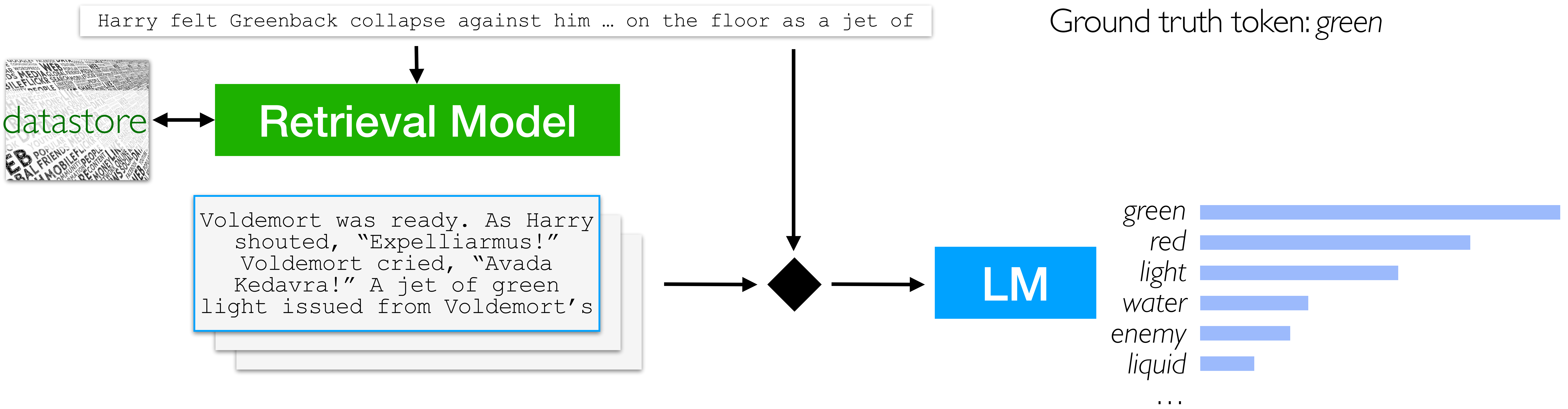
How to train it?



How to train it?

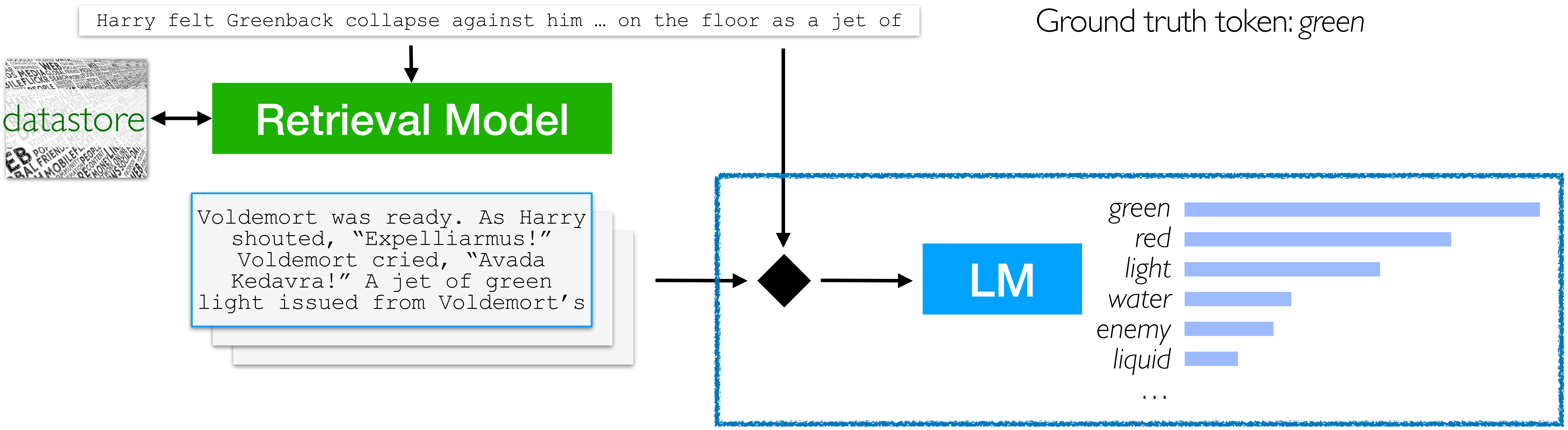


Joint training



REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

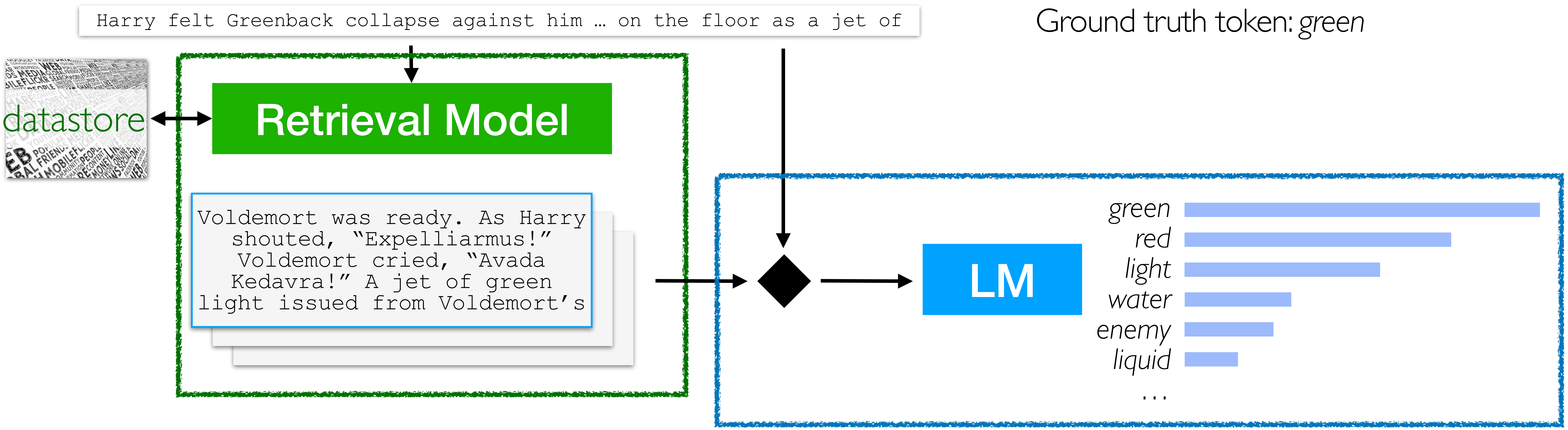
Joint training



⚠️ Training LMs can be very expensive!

REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

Joint training



⚠️ Too large! Expensive to update index during training!

⚠️ Training LMs can be very expensive!

REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

Joint training

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

Ground truth token: *green*

Retrieval Model

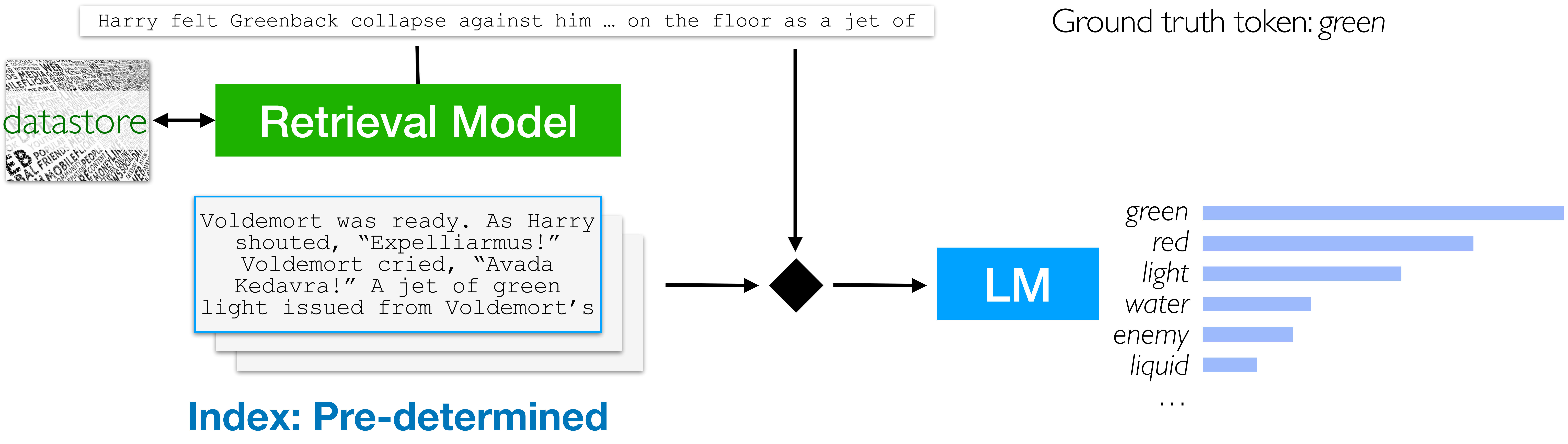
Frozen

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

Index: Pre-determined

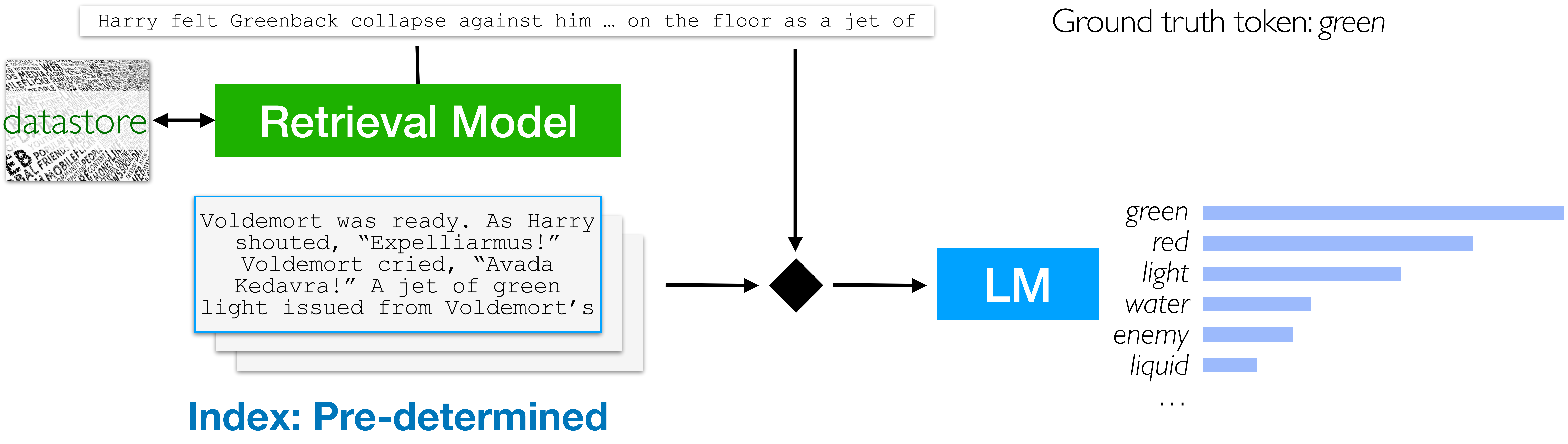
REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

Joint training



REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

Joint training

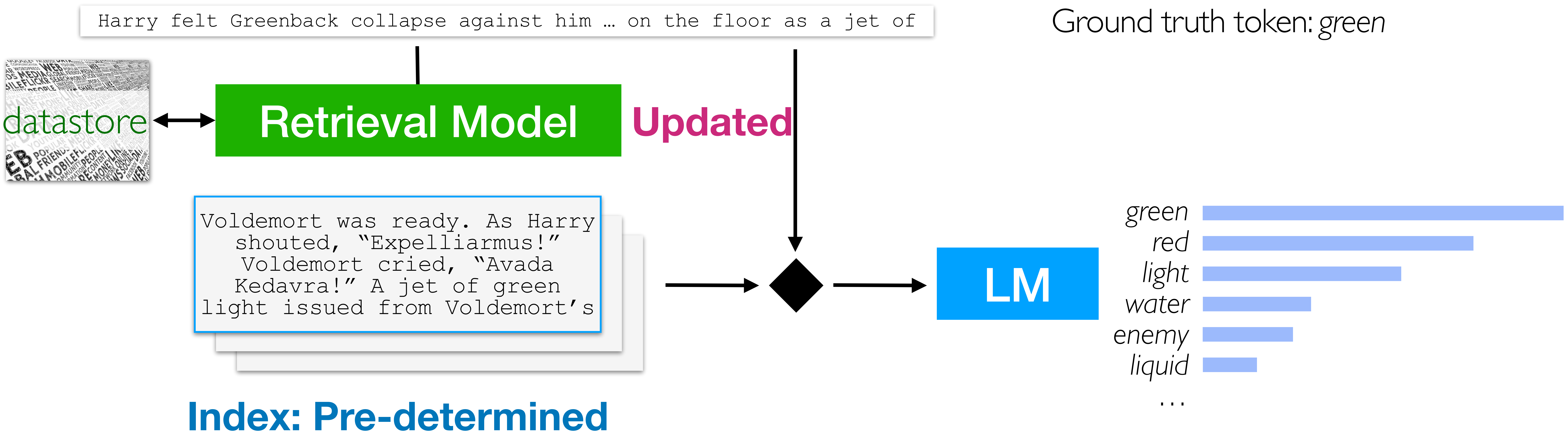


Maximize $P(y|x) = \sum_{z \in \mathcal{Z}} P_{\text{ret}}(z|x) P_{\text{LM}}(y|x,z)$

retrieval score LM score

REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

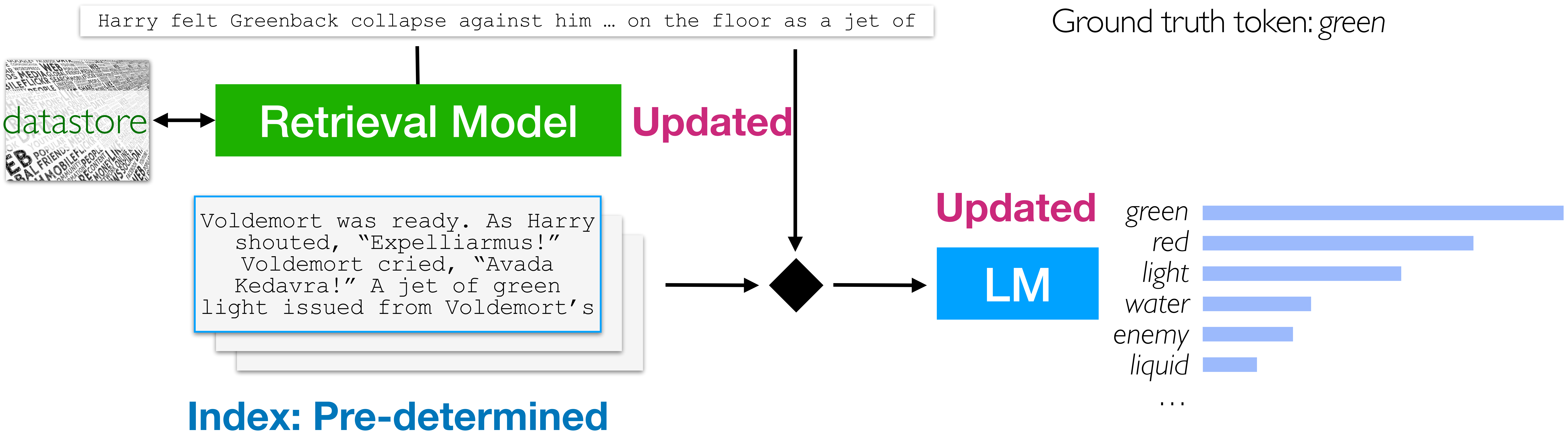
Joint training



$$\text{Maximize } P(y|x) = \sum_{z \in \mathcal{Z}} \underbrace{P_{\text{ret}}(z|x)}_{\text{retrieval score}} \underbrace{P_{\text{LM}}(y|x,z)}_{\text{LM score}}$$

REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

Joint training

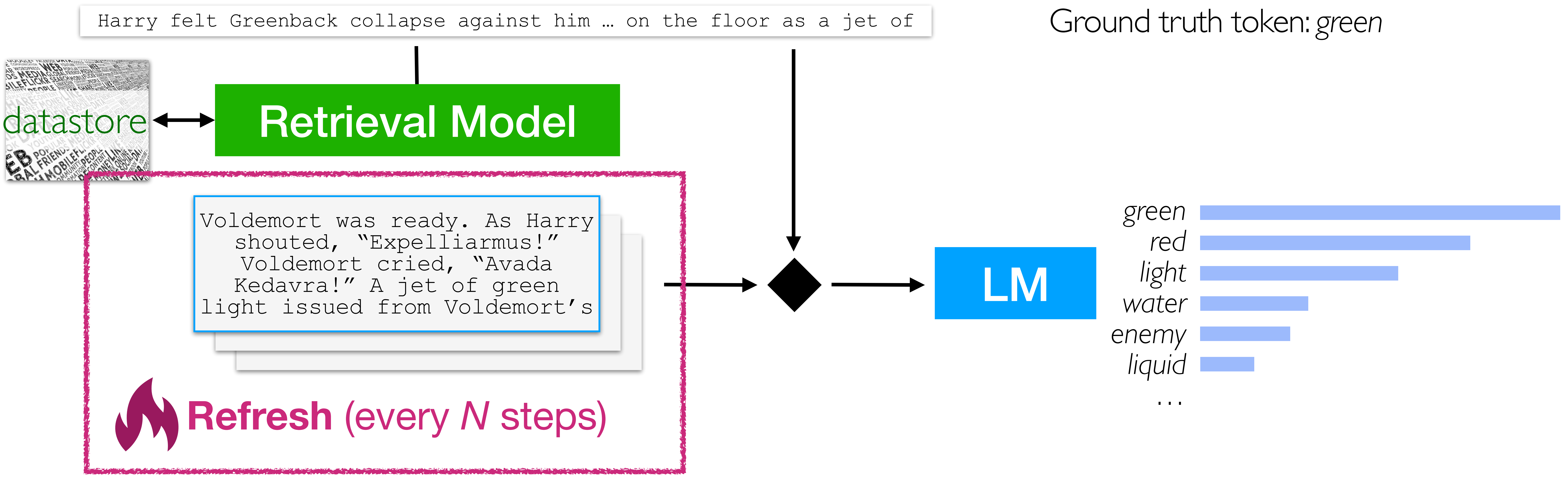


Maximize $P(y \mid x) = \sum_{z \in \mathcal{Z}} P_{\text{ret}}(z \mid x) P_{\text{LM}}(y \mid x, z)$

retrieval score — LM score

REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

Joint training



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

retrieval score LM score

REALM (Guu et al 2020), RAG (Lewis et al 2020), Atlas (Izcard et al 2022)

How to train it?

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

Joint training

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

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

Ground truth token: *green*

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

Sequential training: freeze LM, tune retrieval

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

Ground truth token: *green*

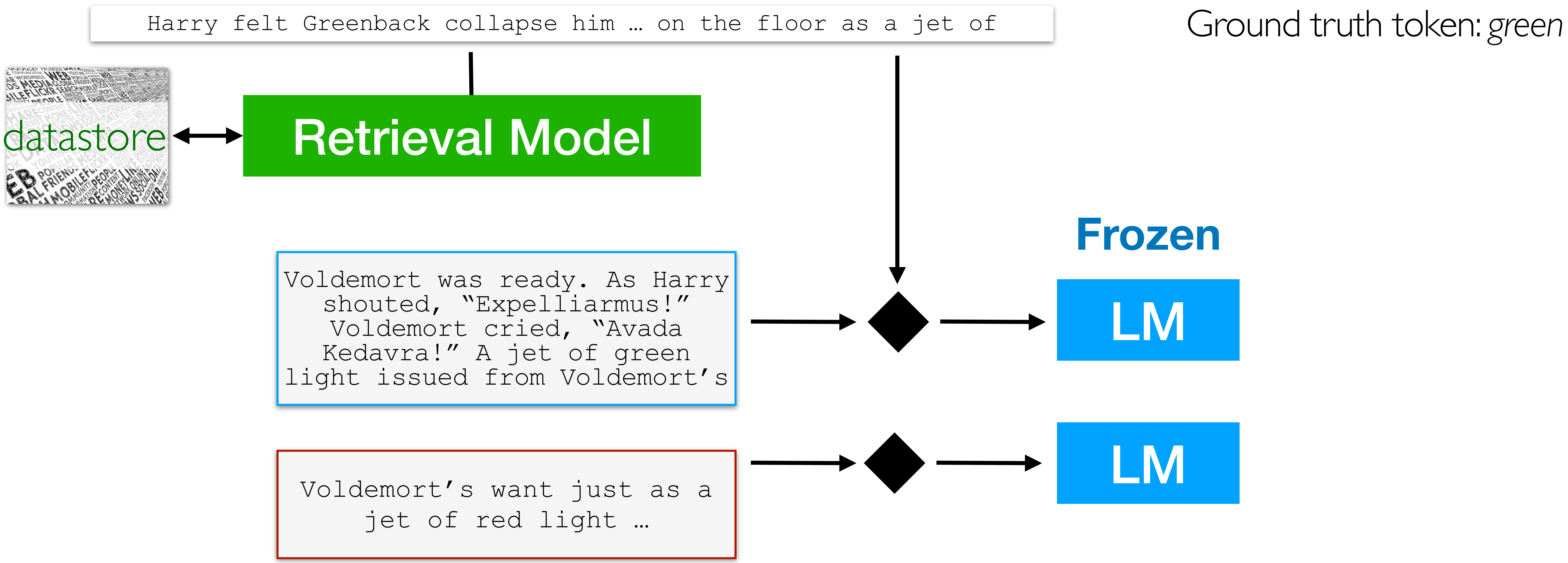
Retrieval Model

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

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

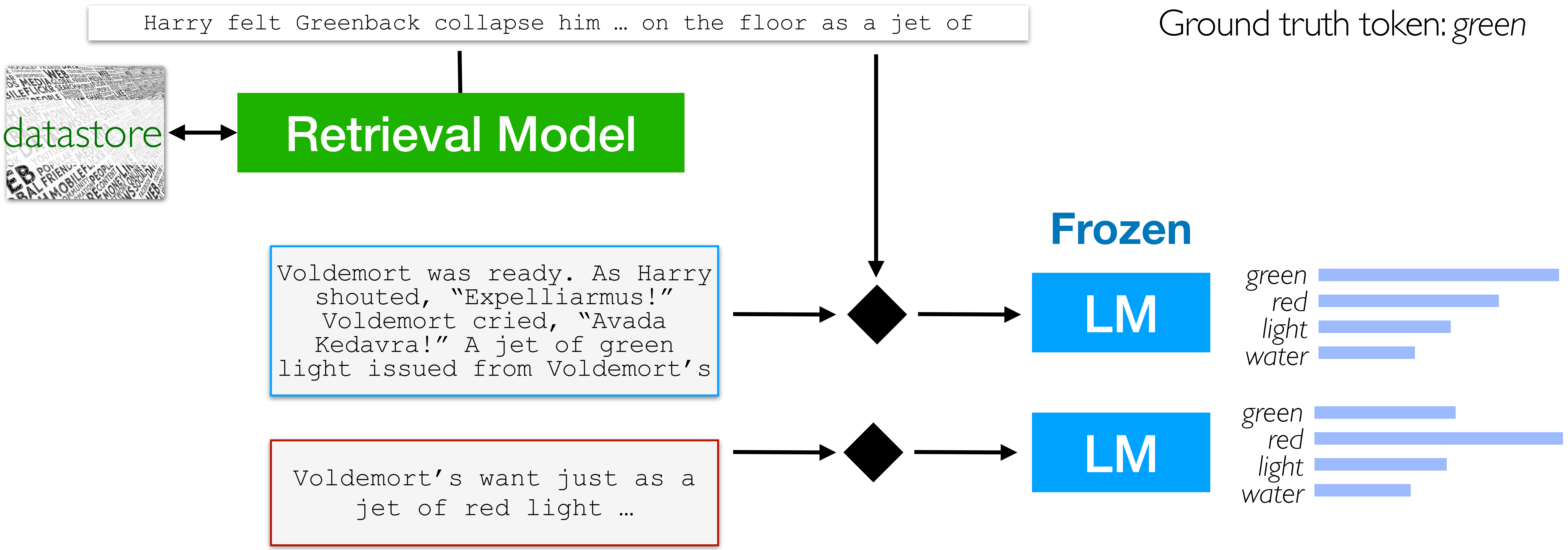
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



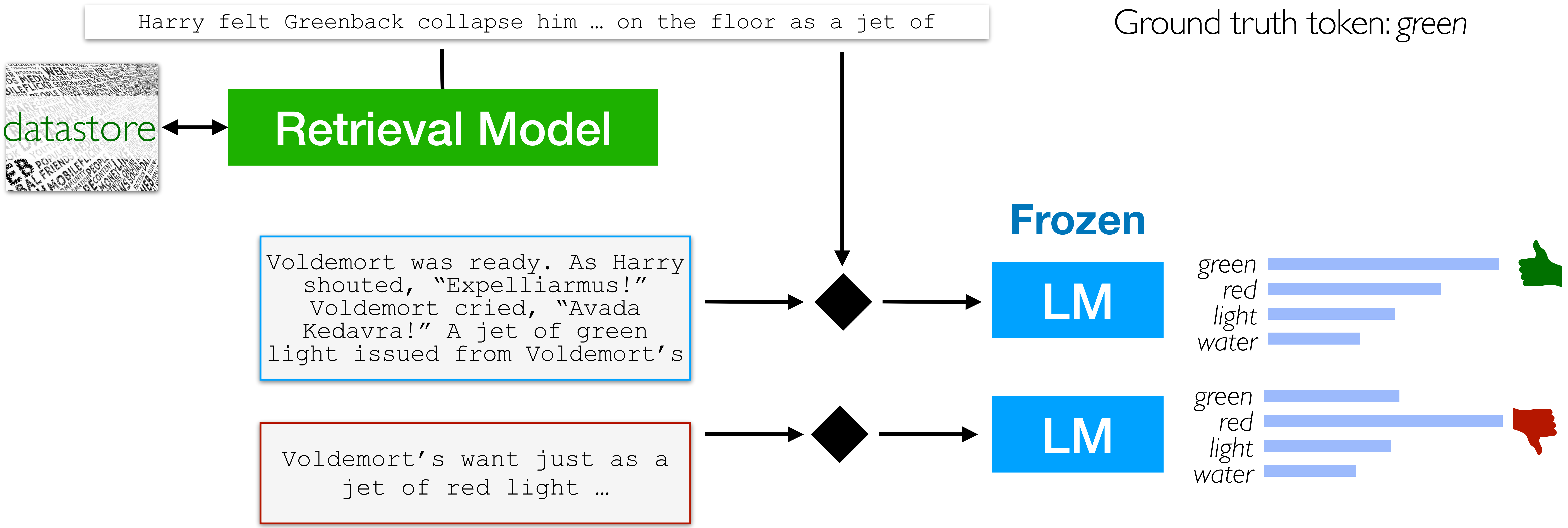
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



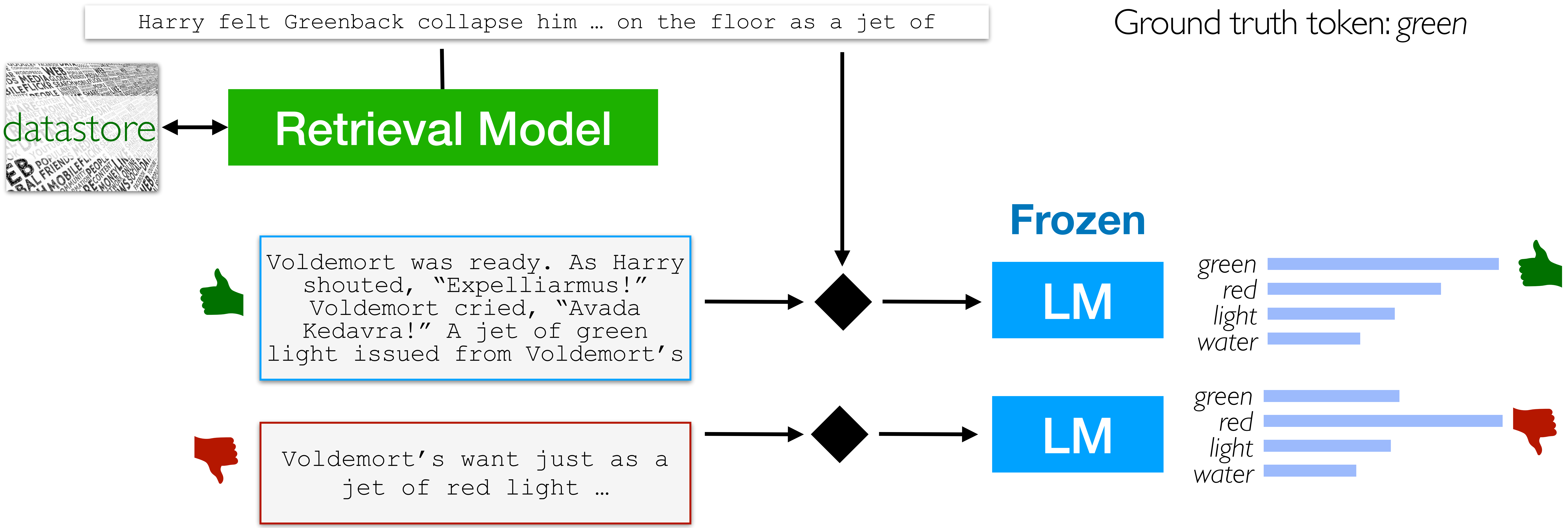
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



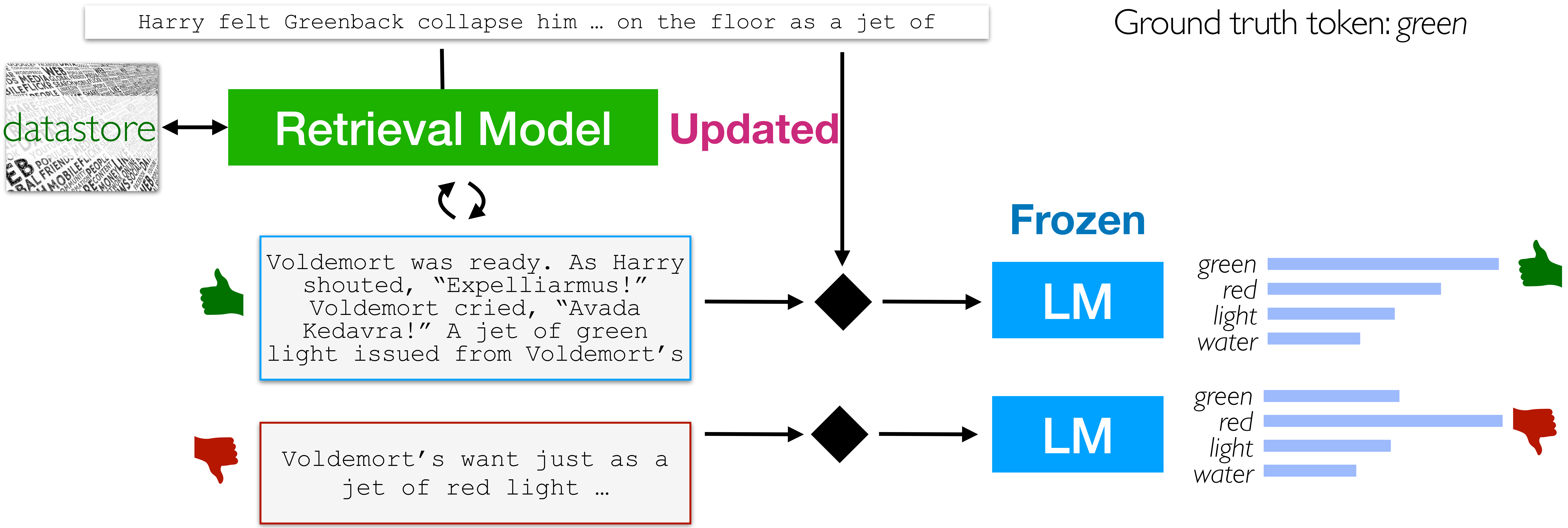
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



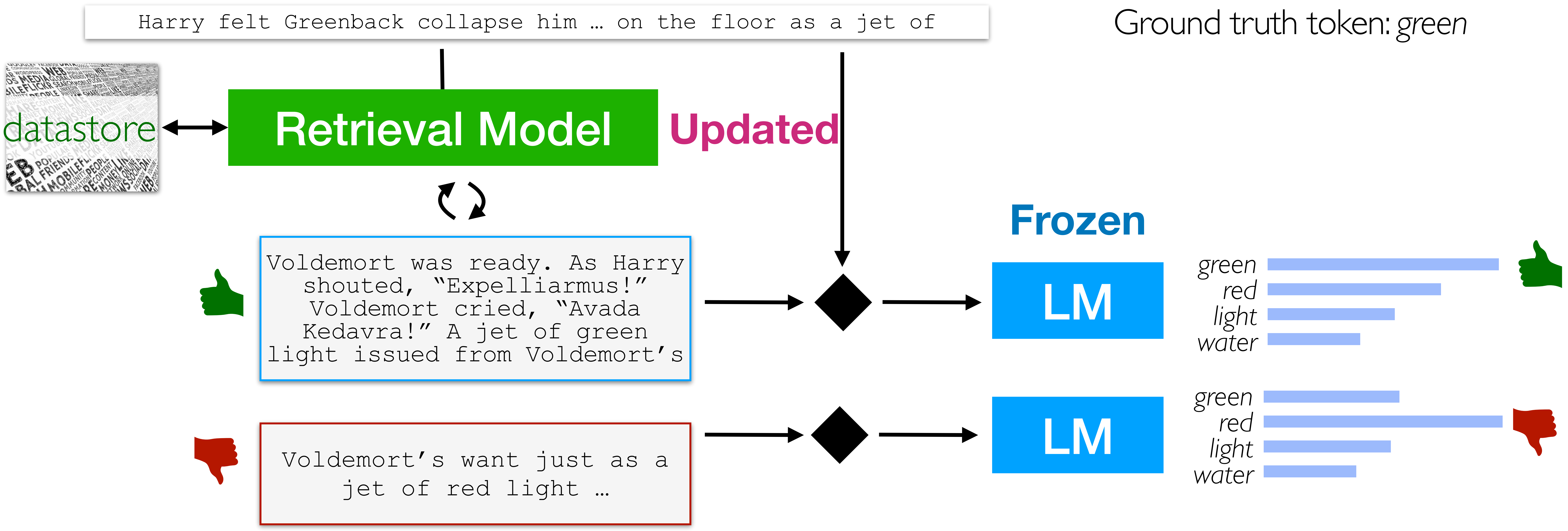
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze LM, tune retrieval



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

Sequential training: freeze LM, tune retrieval



Updated

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

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

Sequential training: freeze retrieval, tune LM

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

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. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze retrieval, tune LM

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

Ground truth token: *green*

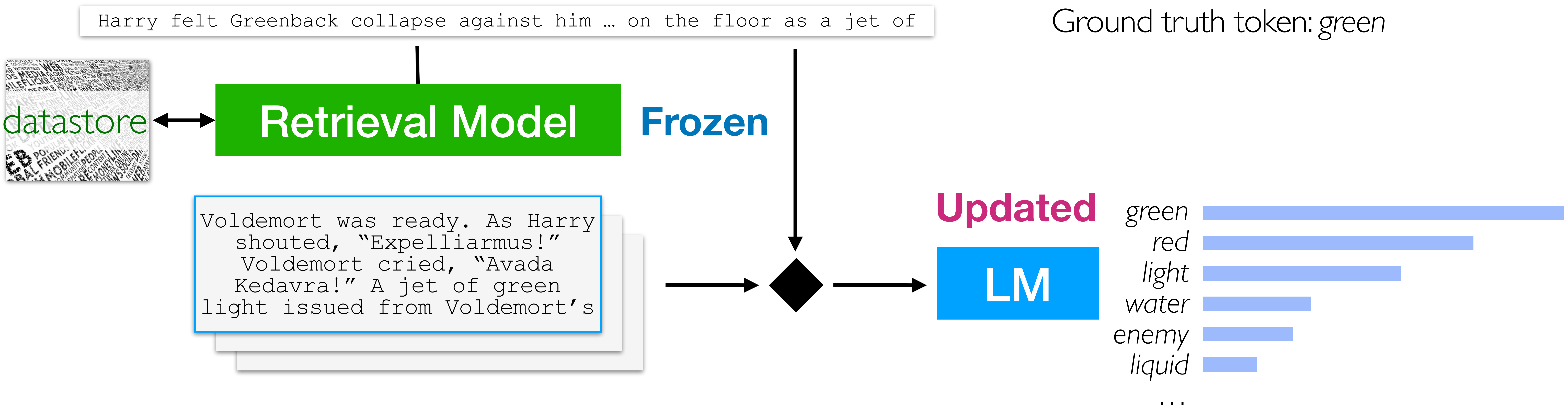
Retrieval Model

Frozen

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

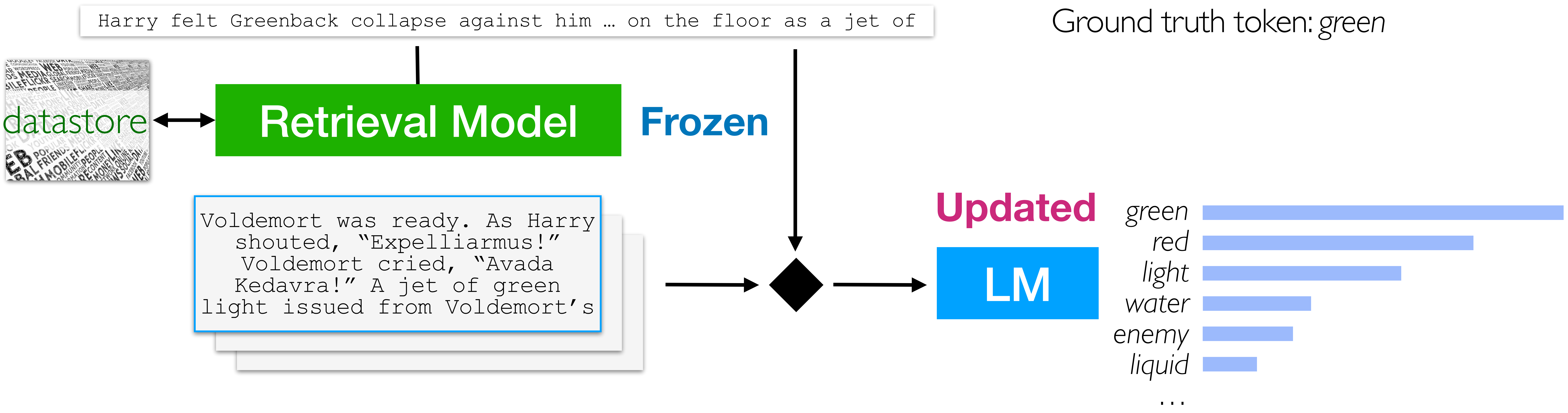
Shi et al. 2023. "REPLUG: Retrieval-Augmented Black-Box Language Models"

Sequential training: freeze retrieval, tune LM



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

Sequential training: freeze retrieval, tune LM



Updated

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

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

Summary: Training

Independent training

Retrieval Model

trained in isolation

LM

trained in isolation

Good enough if you want minimal effort

Joint training

Retrieval Model

LM

trained jointly

Principle way but still open question

Sequential training

trained conditionally

Retrieval Model

↓

LM

trained in isolation

or

trained in isolation

Retrieval Model

↑

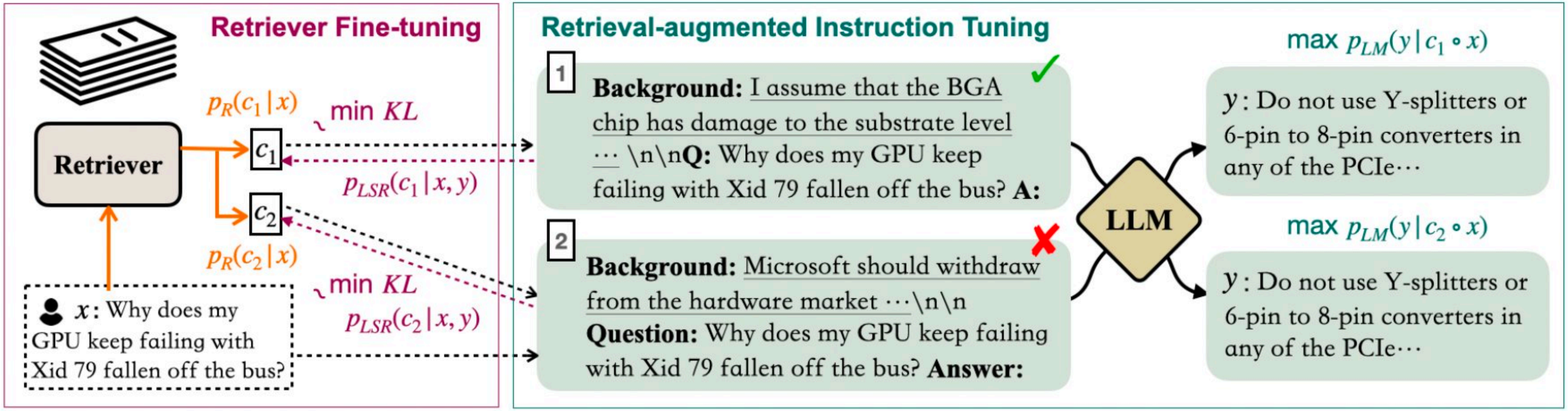
LM

trained conditionally

Good middle ground

Instruction-tuning/Post-training

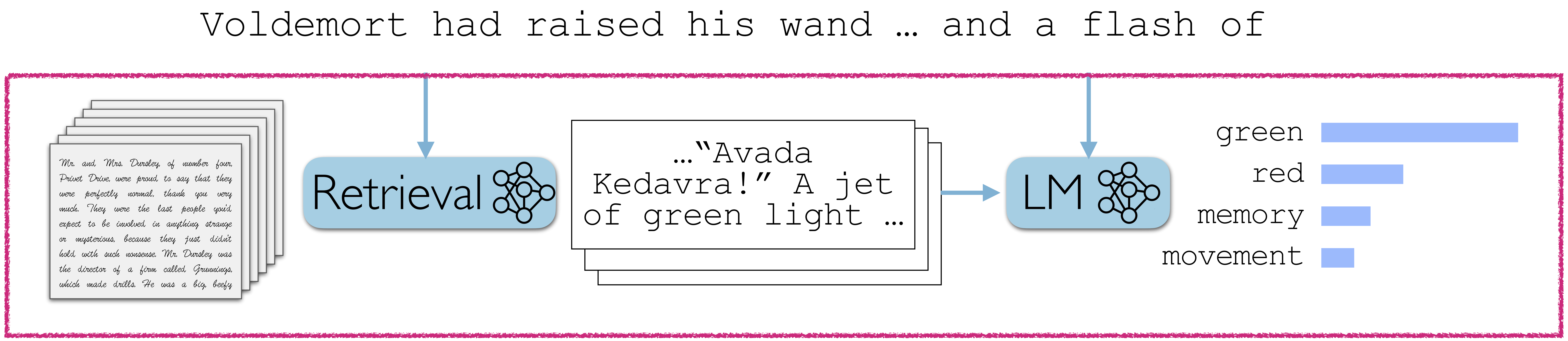
Instruction-tuning/Post-training



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

Wang et al. 2024. InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining

A two-stage pipeline

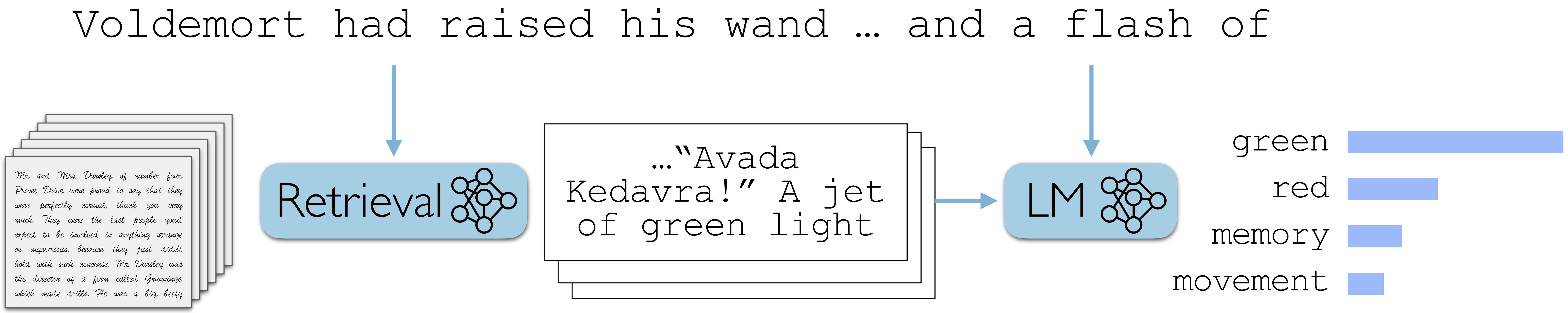


1) Retrieval

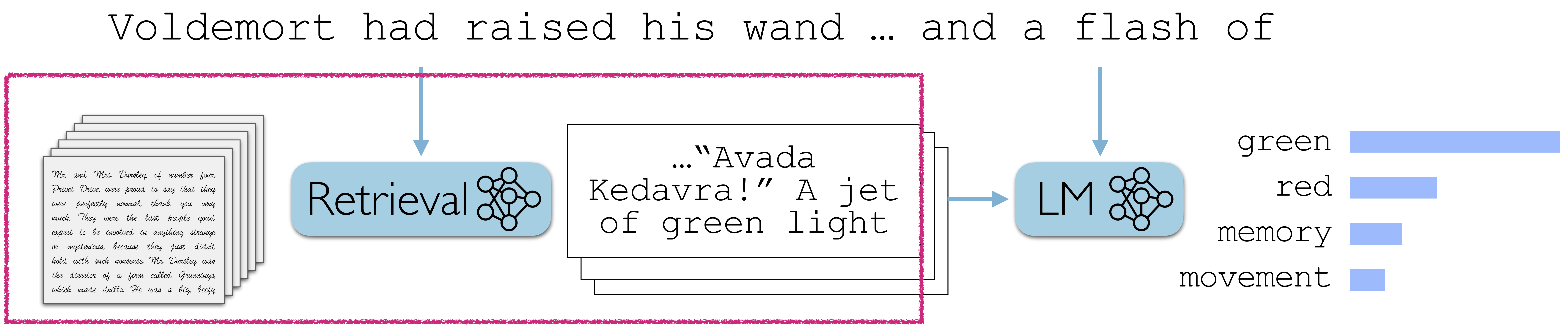
2) Augmentation

3) Training

Summary of this section



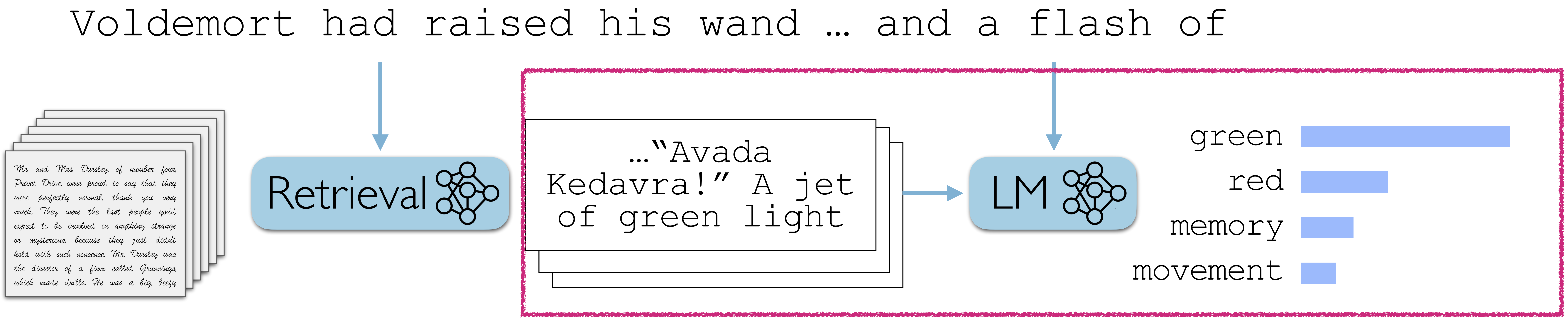
Summary of this section



1) Retrieval:

Advances in neural retrieval played a vital role in the success of retrieval-based LMs

Summary of this section



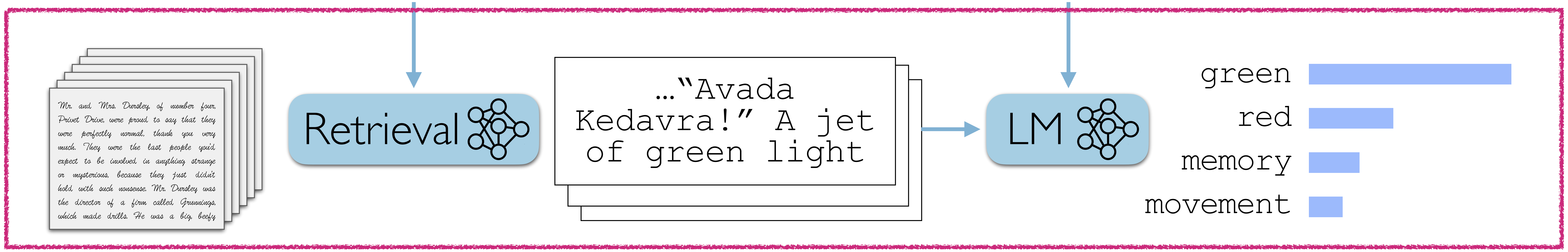
2) Augmentation:

Advances in LLMs enabled a very simple augmentation recipe

(We'll talk about how we can do this better in Part 2)

Summary of this section

Voldemort had raised his wand … and a flash of

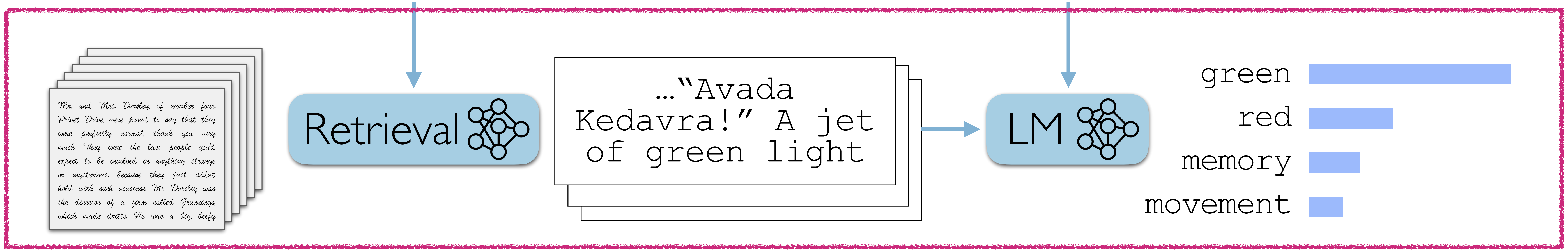


3) Training

Independent training, sequential training, and joint training, with tradeoffs in simplicity and effectiveness

Summary of this section

Voldemort had raised his wand … and a flash of



(There're different architectures beyond the two-stage pipeline)

Other architectures beyond the two-stage pipeline?

(We’ll only briefly review two different architecture types!)

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)

$\boldsymbol{x}$ = 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)

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

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

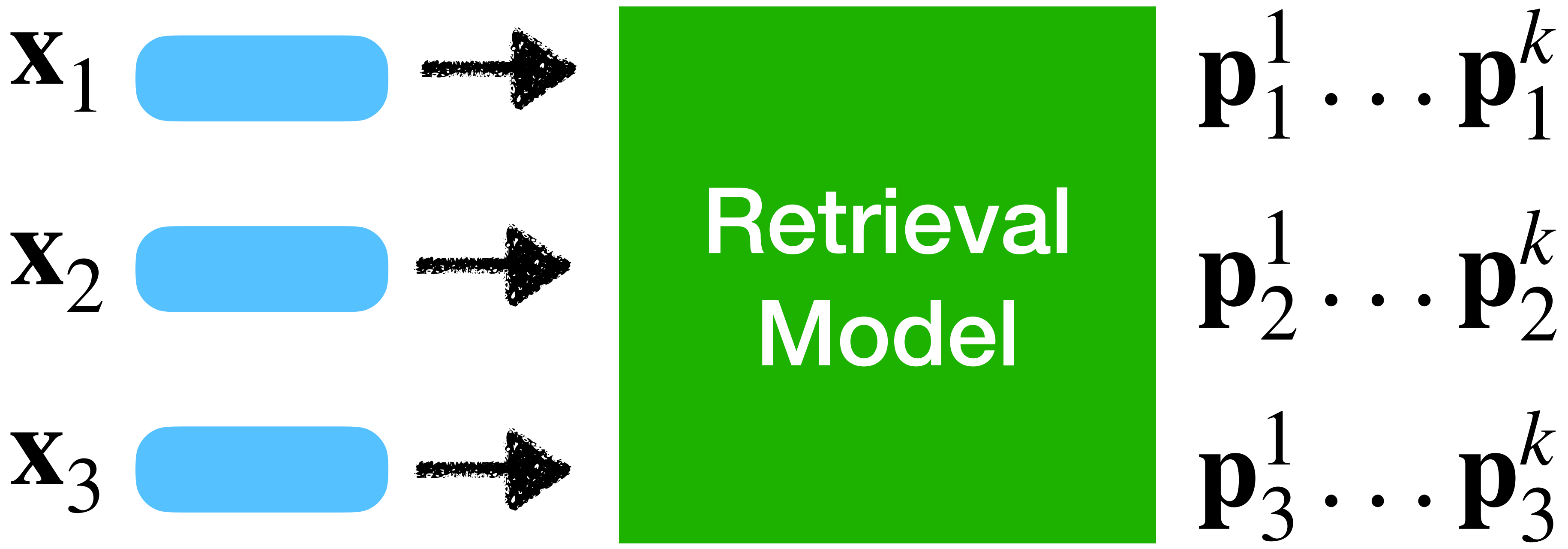
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

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)



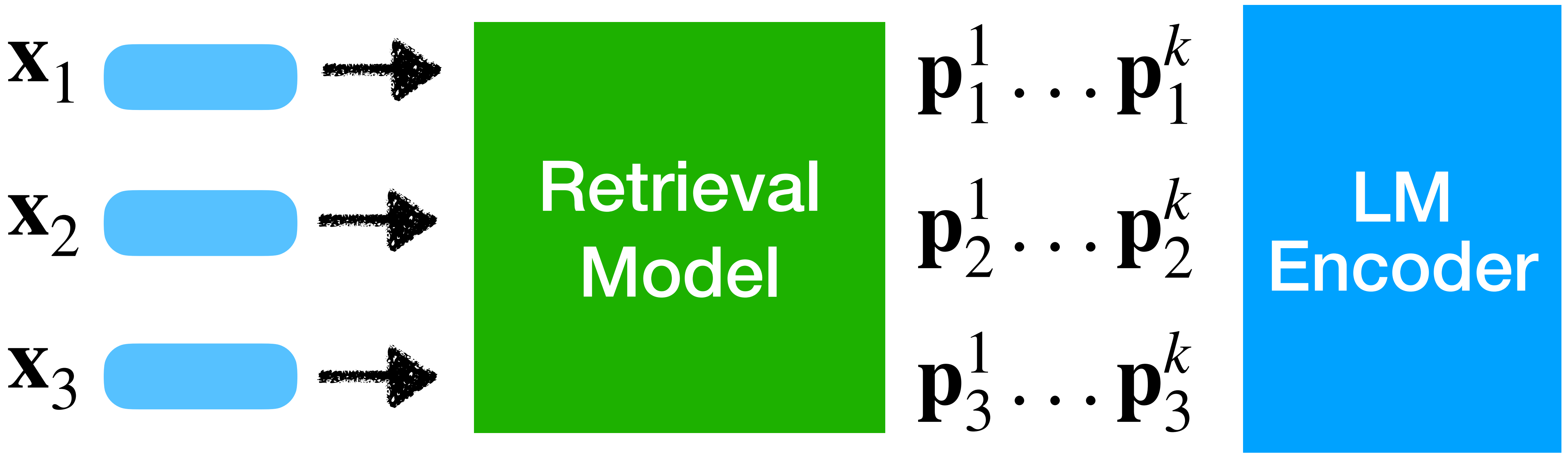
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

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)



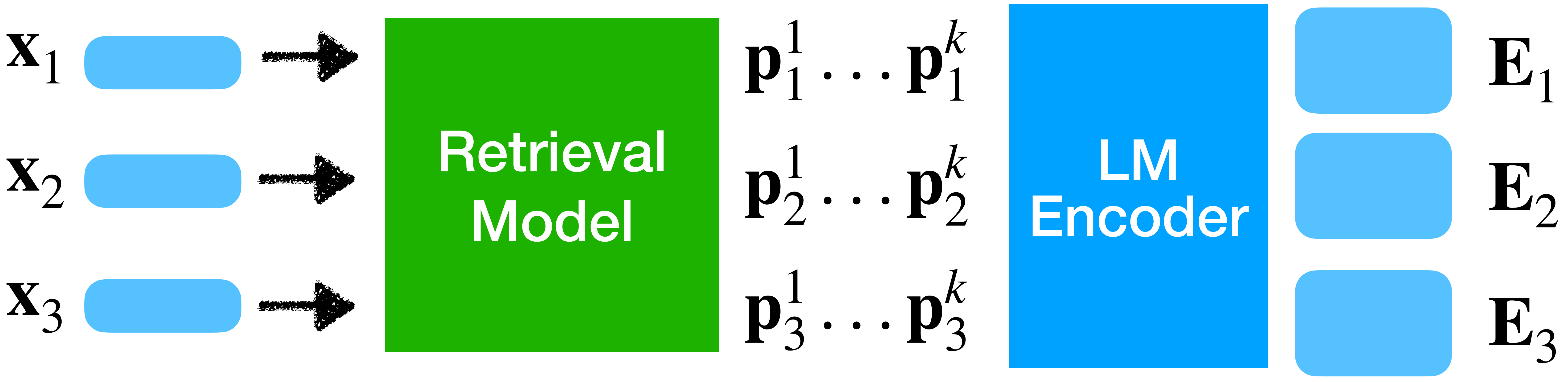
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

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)



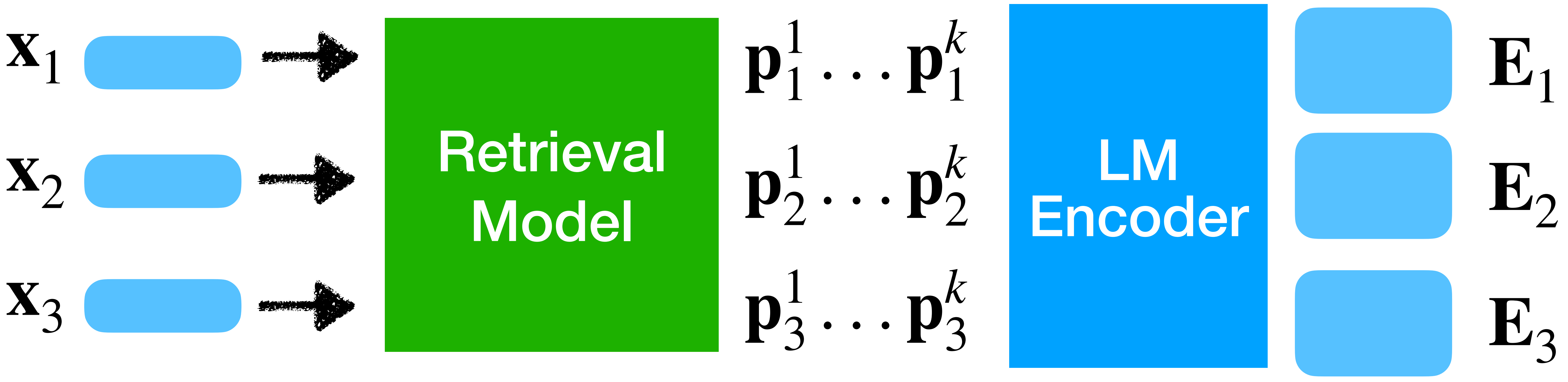
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

RETRO (Borgeaud et al. 2021)

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

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

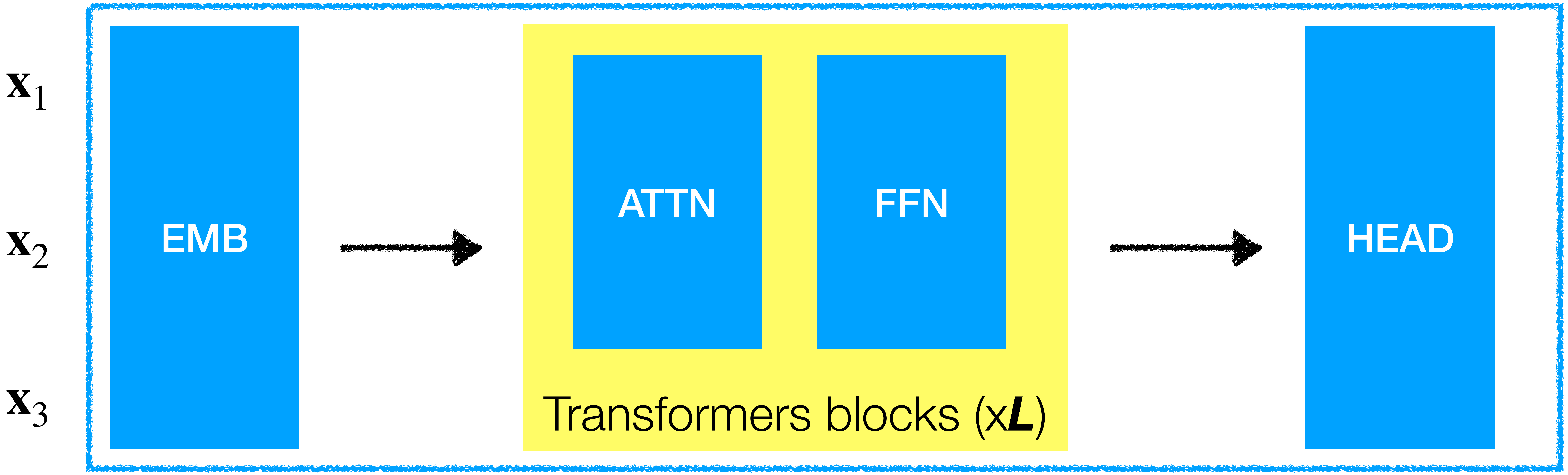
(k text blocks per split)



How to incorporate them into Transformers?

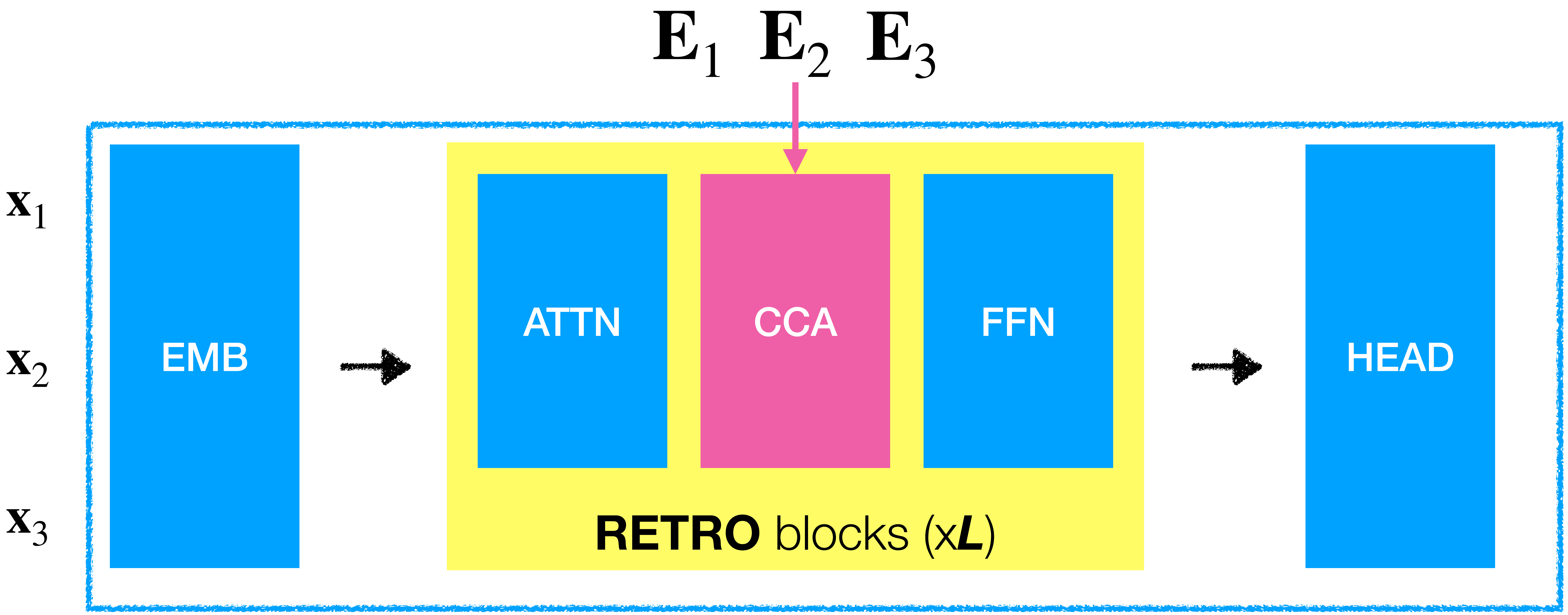
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

Regular Transformers



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

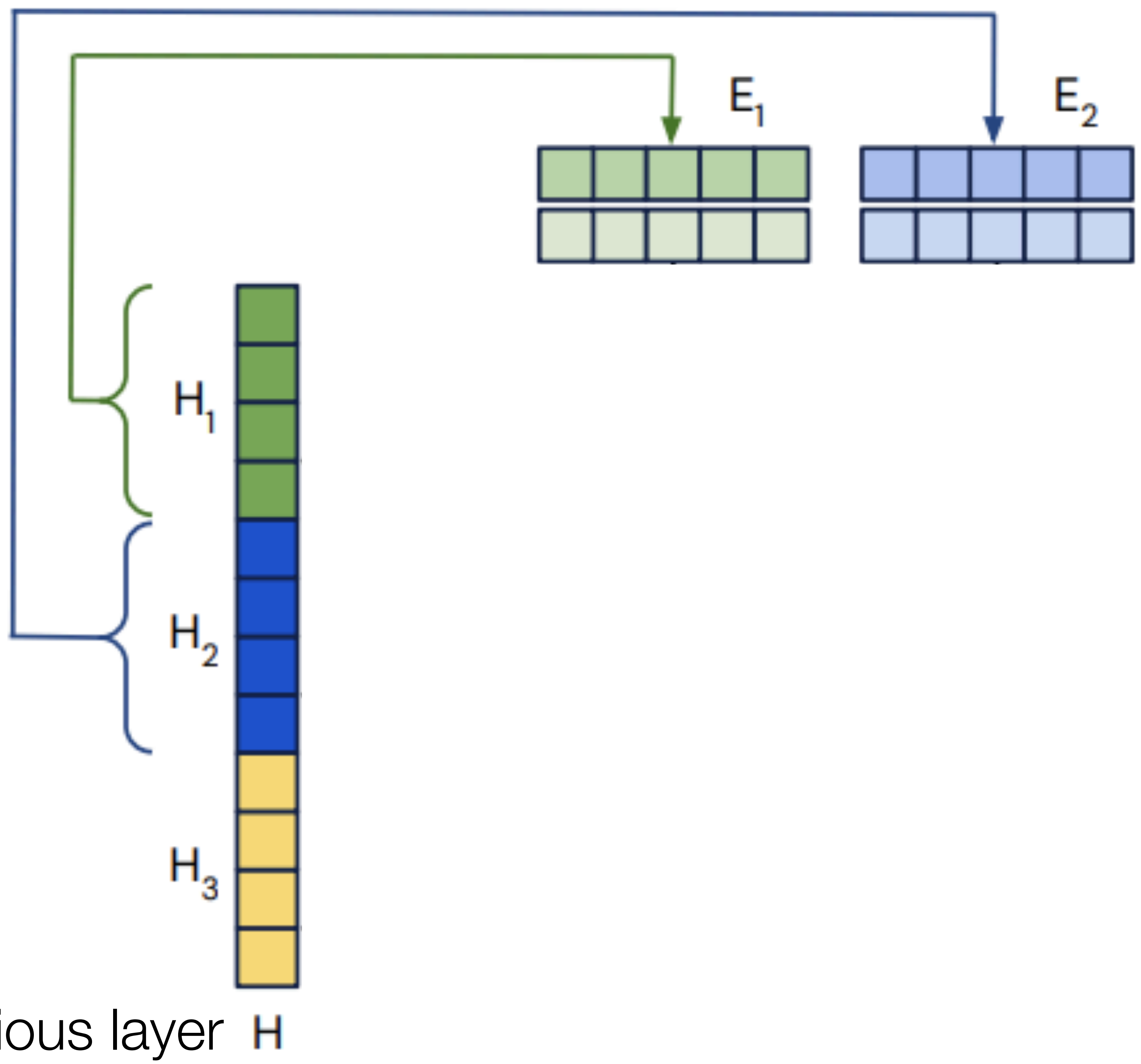
RETRO Transformers



Chunked Cross Attention (CCA)

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

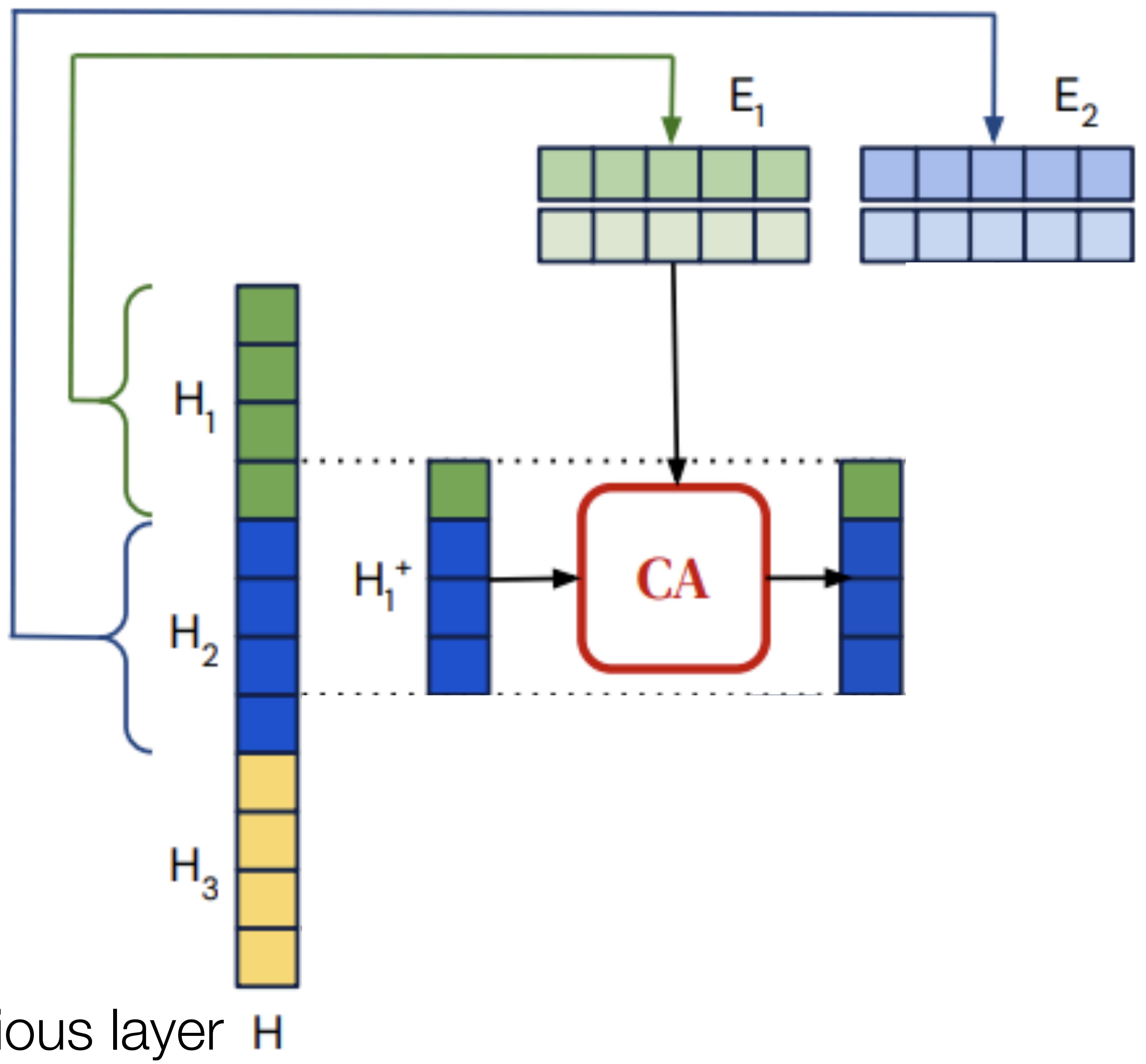
Chunked Cross Attention



Outputs from the previous layer

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

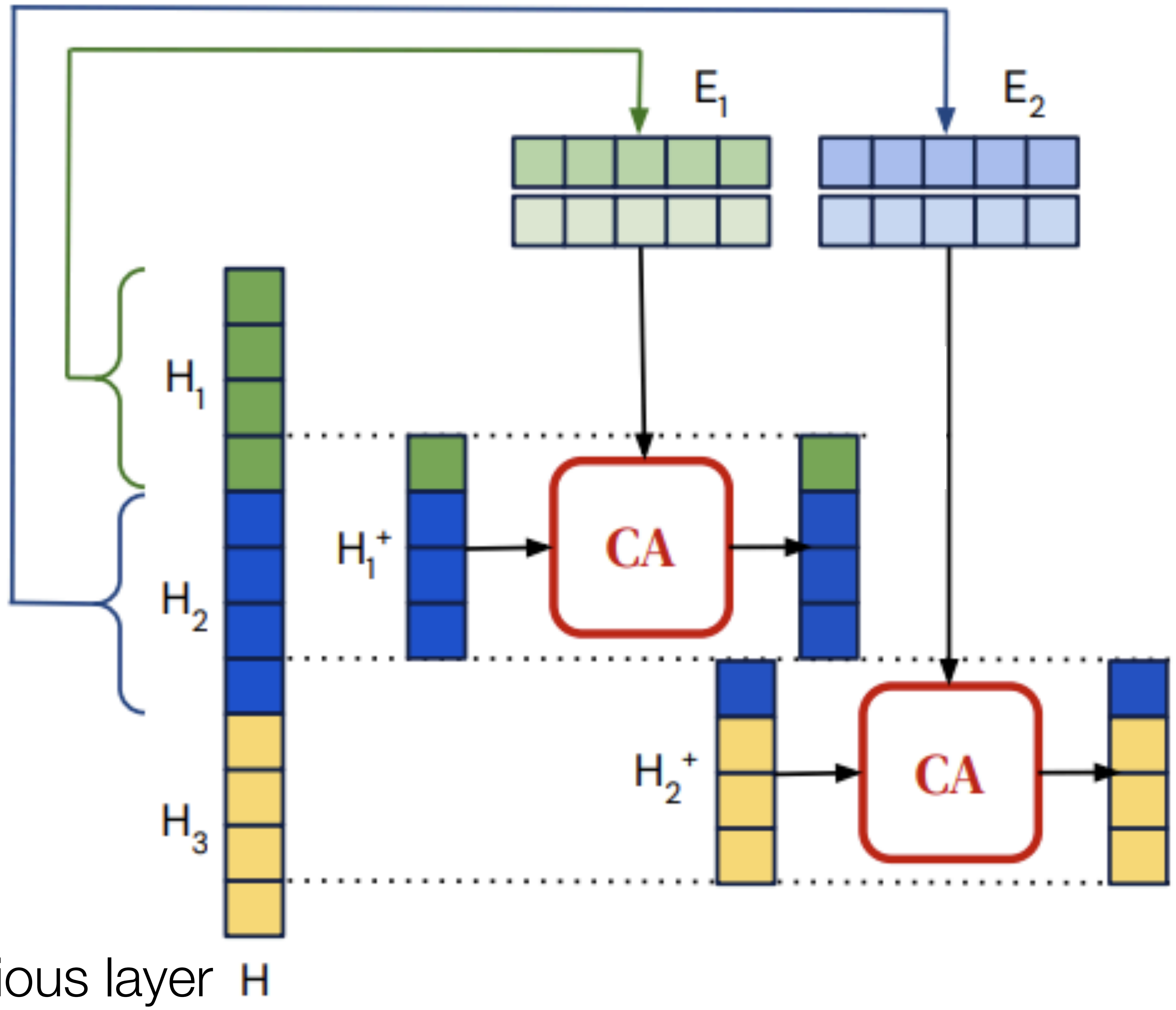
Chunked Cross Attention



Outputs from the previous layer

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

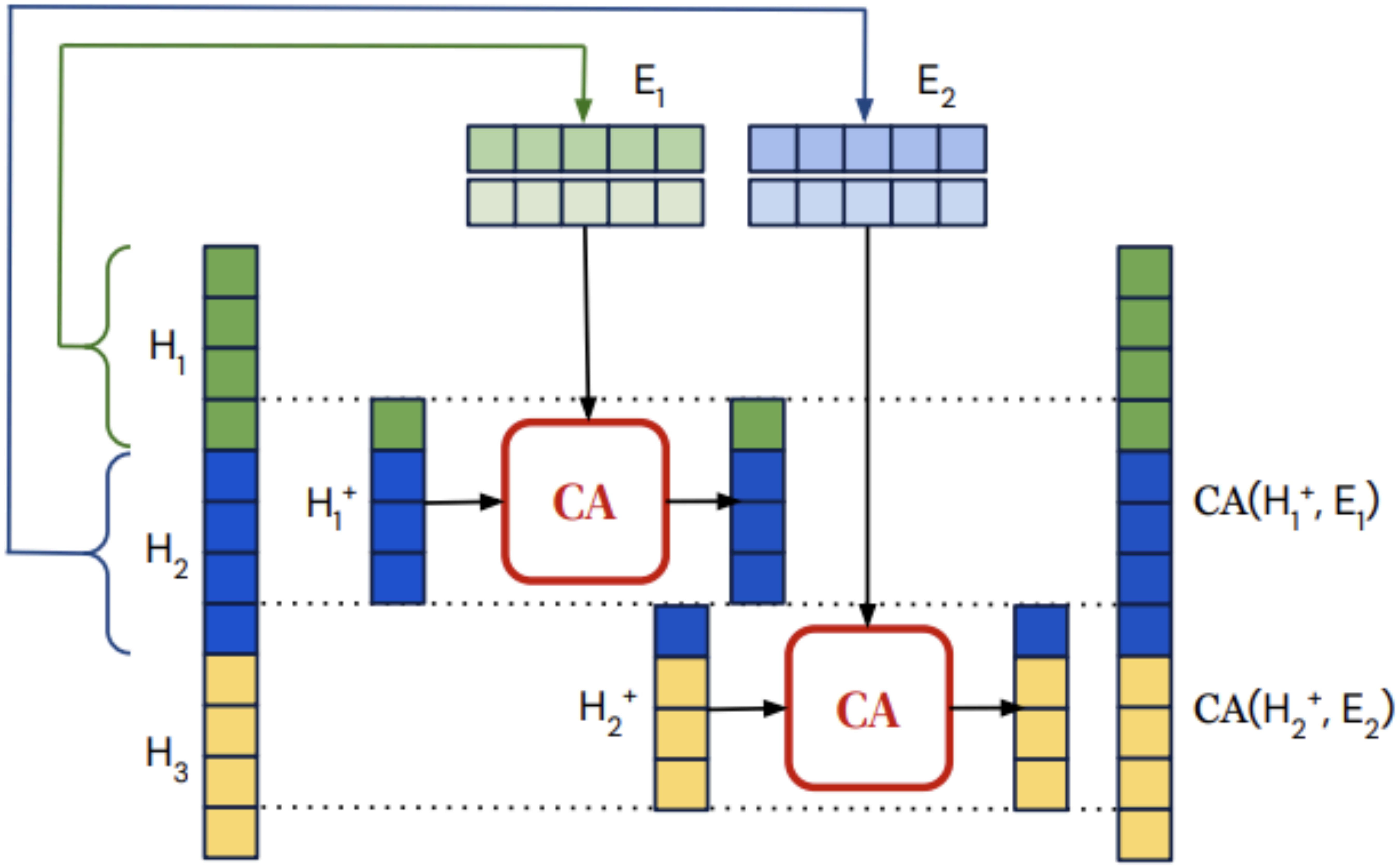
Chunked Cross Attention



Outputs from the previous layer

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

Chunked Cross Attention

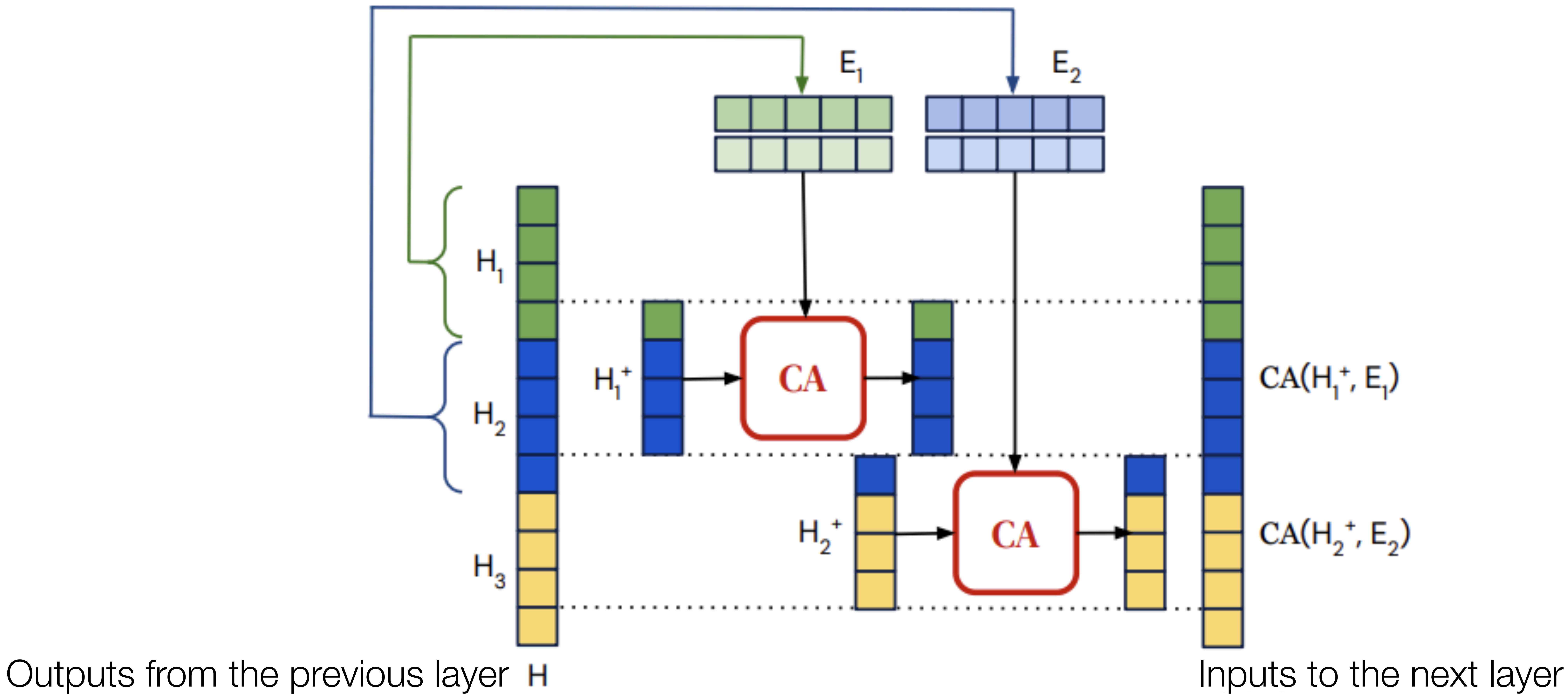


Outputs from the previous layer H

Inputs to the next layer

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

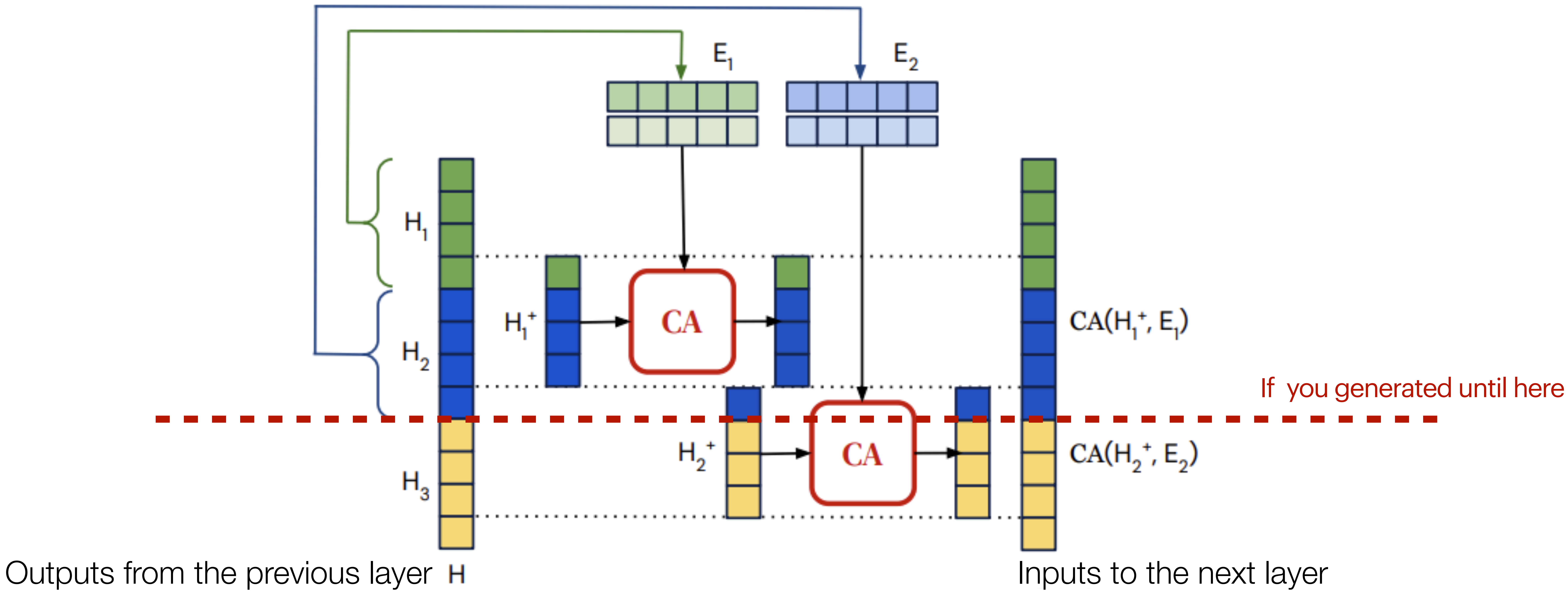
Chunked Cross Attention



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

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

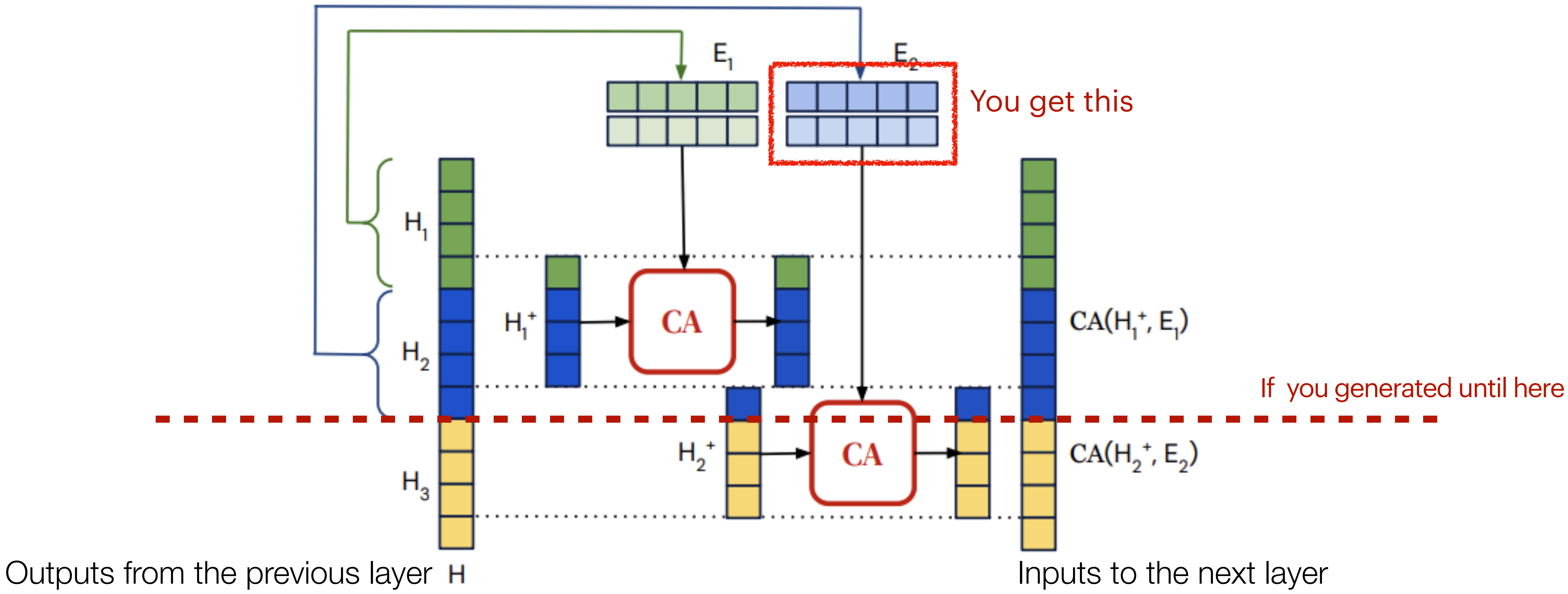
Chunked Cross Attention



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

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

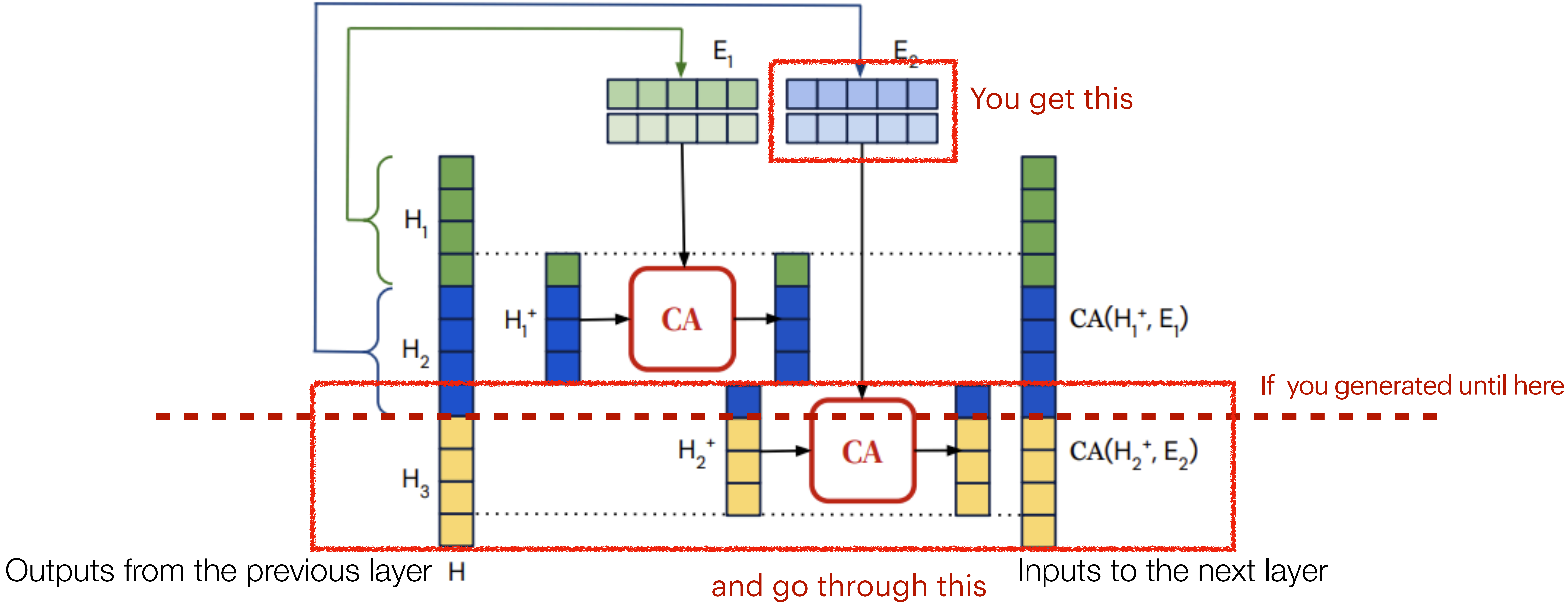
Chunked Cross Attention



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

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

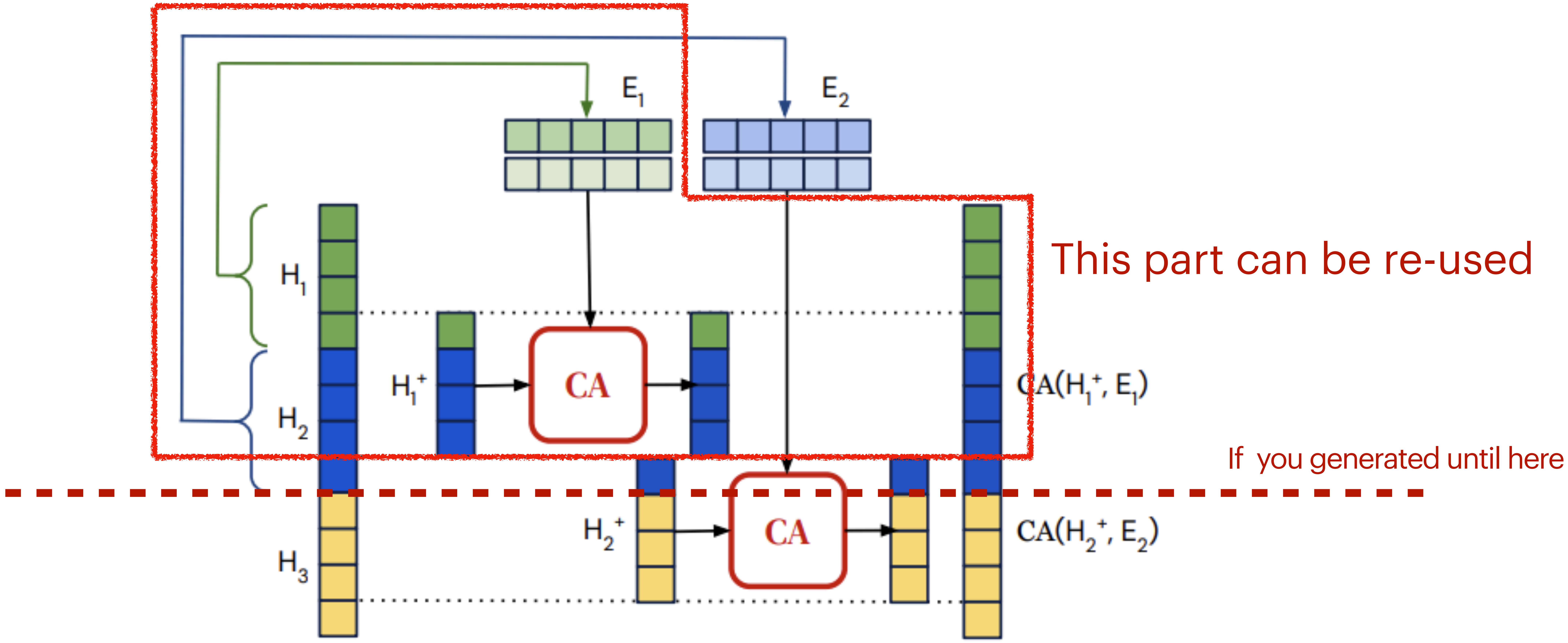
Chunked Cross Attention



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

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

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"

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
kNN-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
kNN-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)

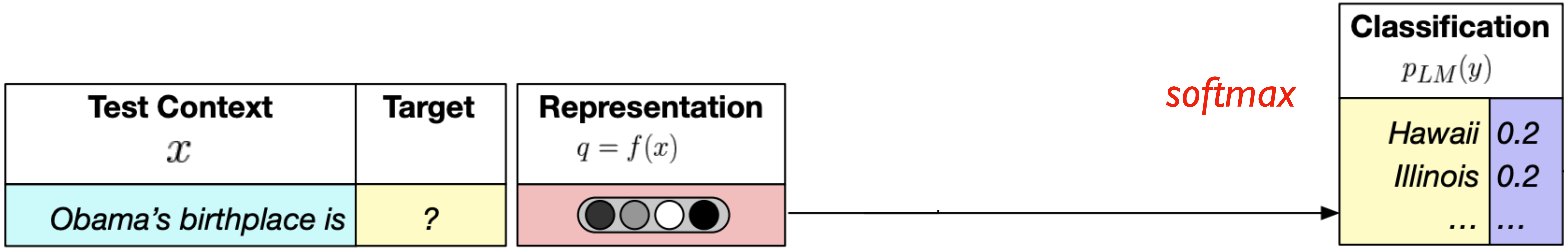
Borgeaud et al. 2021. "Improving language models by retrieving from trillions of tokens"

kNN-LM

Test Context x	**Target**
Obama's birthplace is	*?*

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

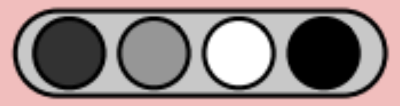
kNN-LM



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

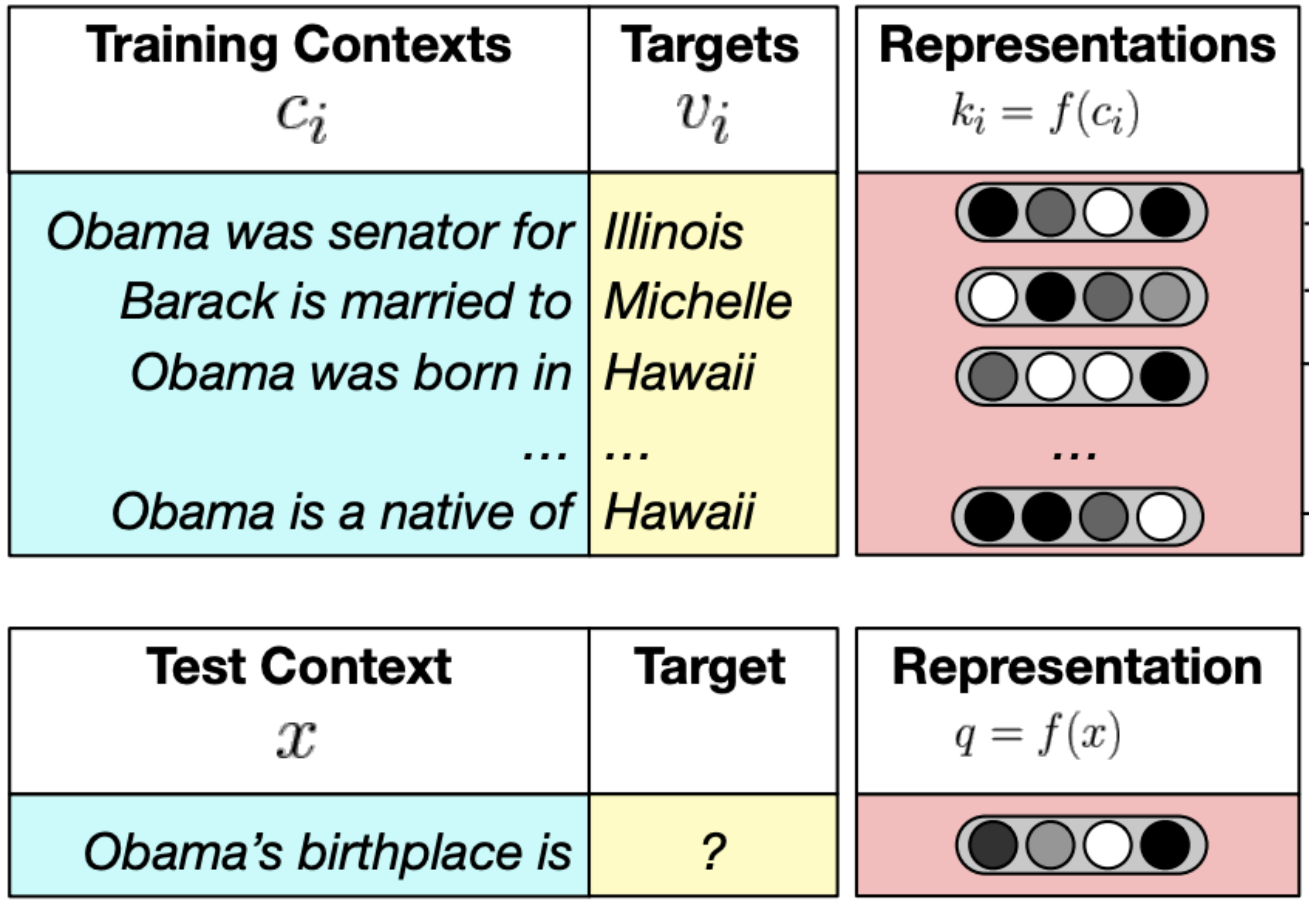
Training Contexts c_i	**Targets** v_i
Obama was senator for	*Illinois*
Barack is married to	*Michelle*
Obama was born in	*Hawaii*
...	...
Obama is a native of	*Hawaii*

... 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"

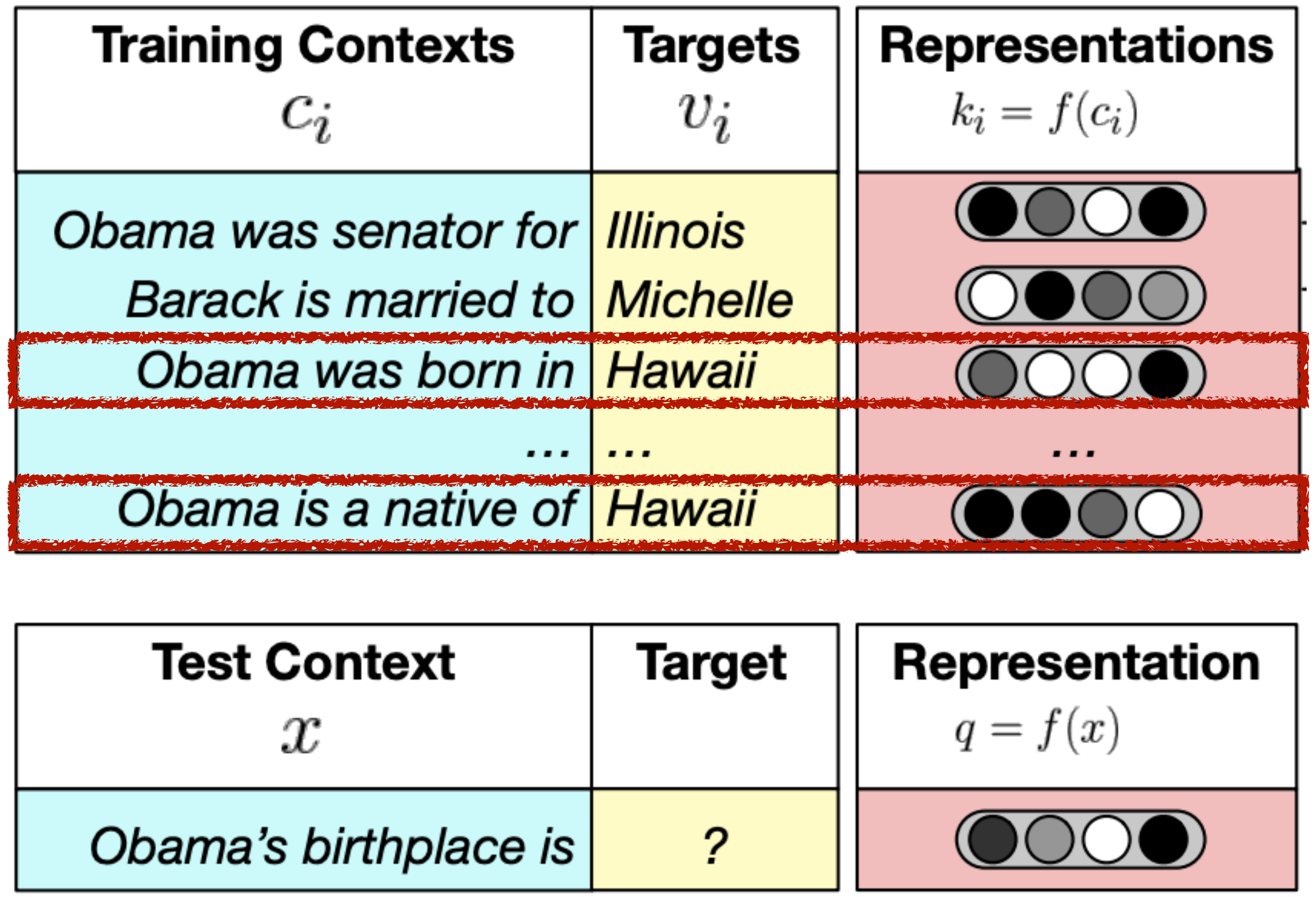
kNN-LM



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

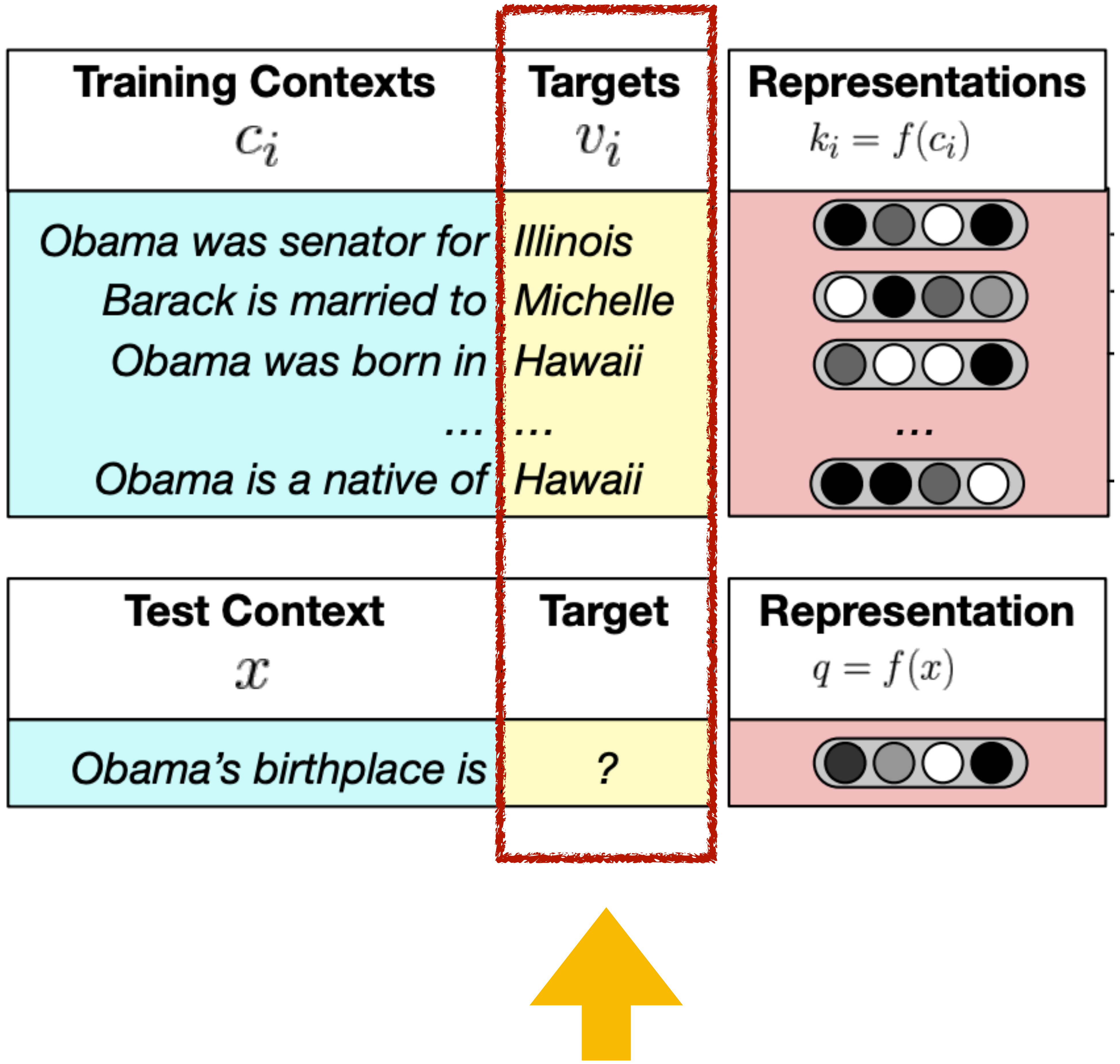
kNN-LM

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



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

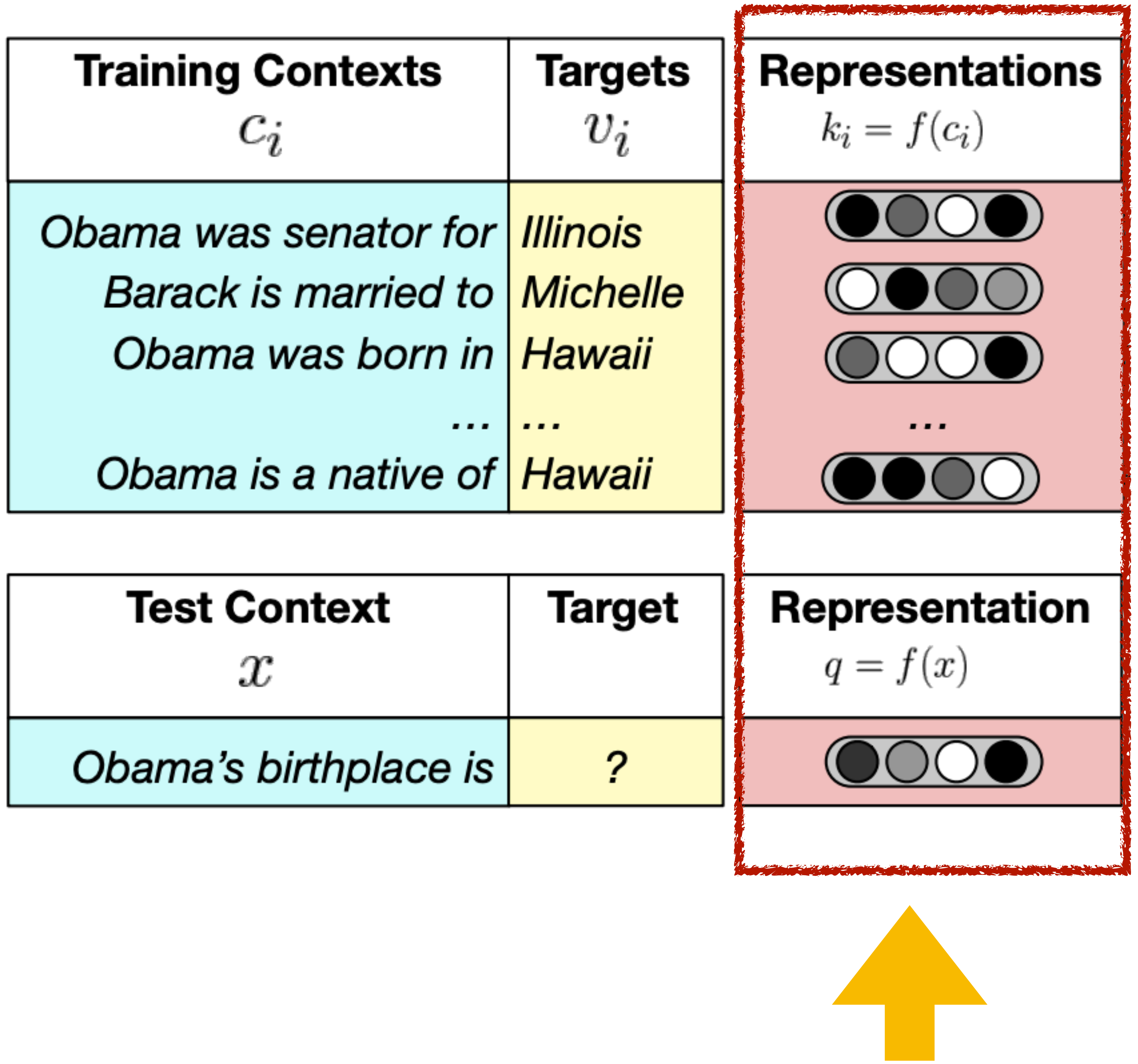
kNN-LM



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

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

kNN-LM



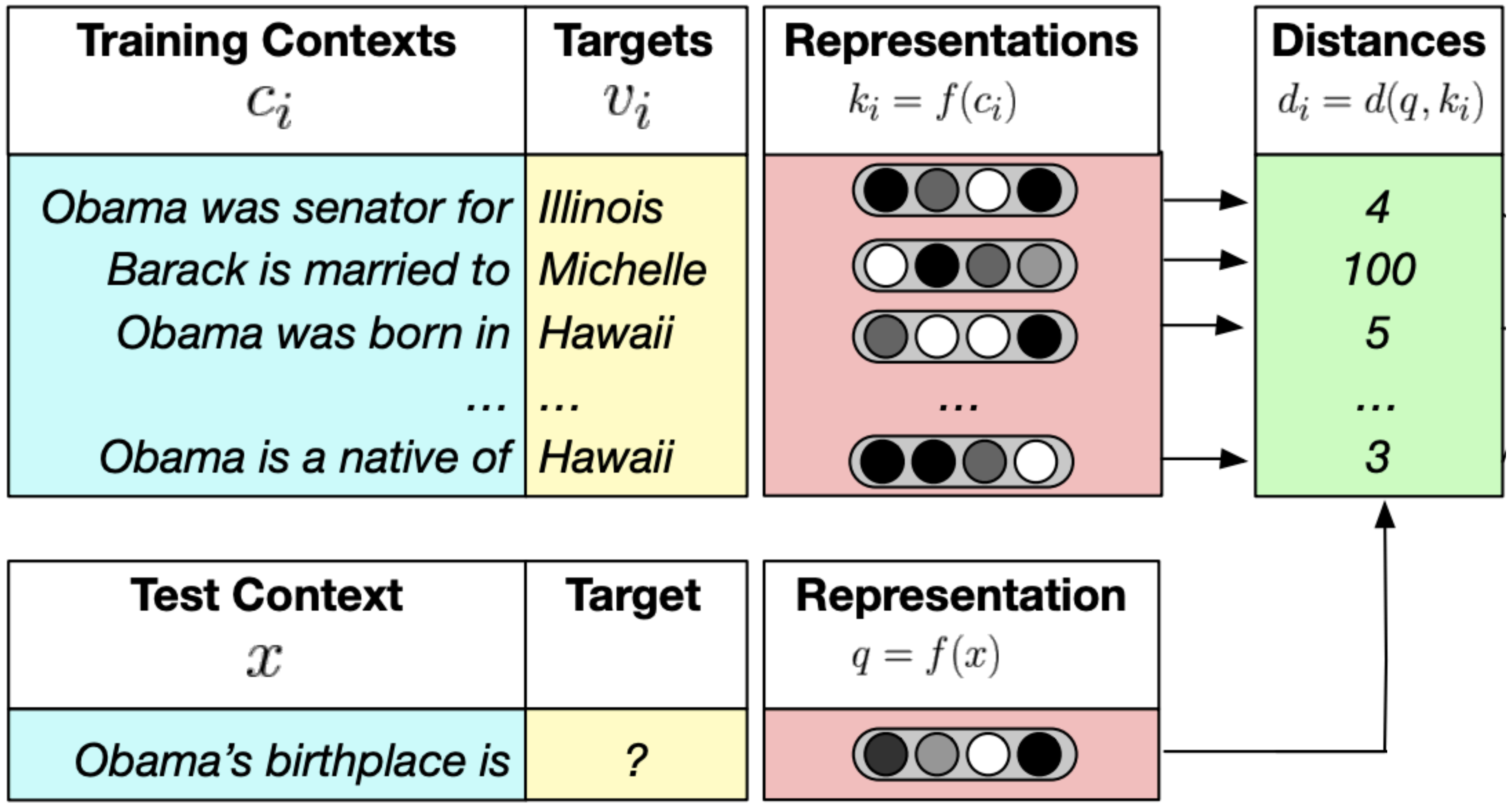
Which tokens <u>in a datastore</u> are close to the next token?

=

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

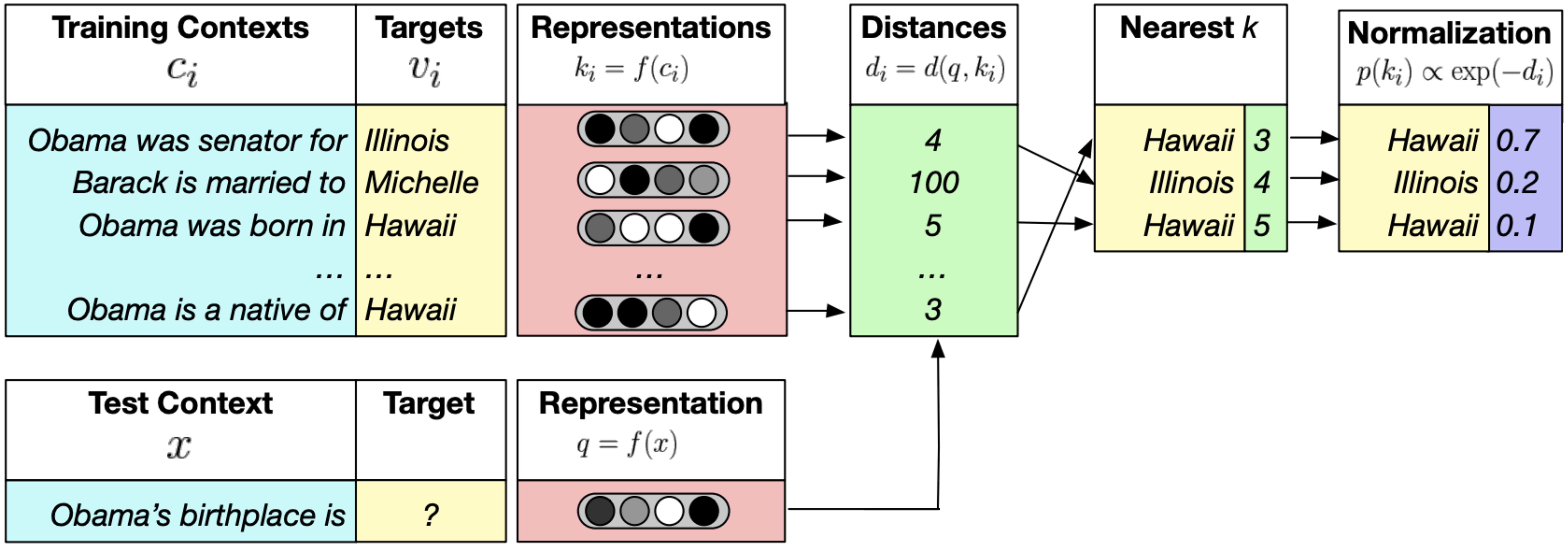
Khandelwal et al. 2020. "Generalization through Memorization: Nearest Neighbor Language Models"

kNN-LM



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

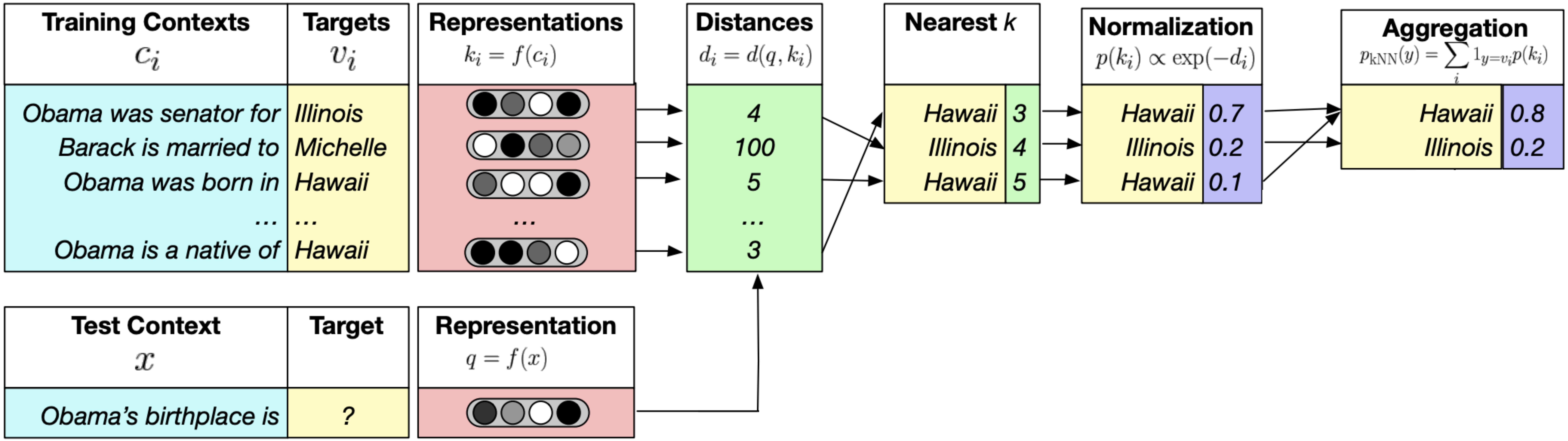
kNN-LM



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

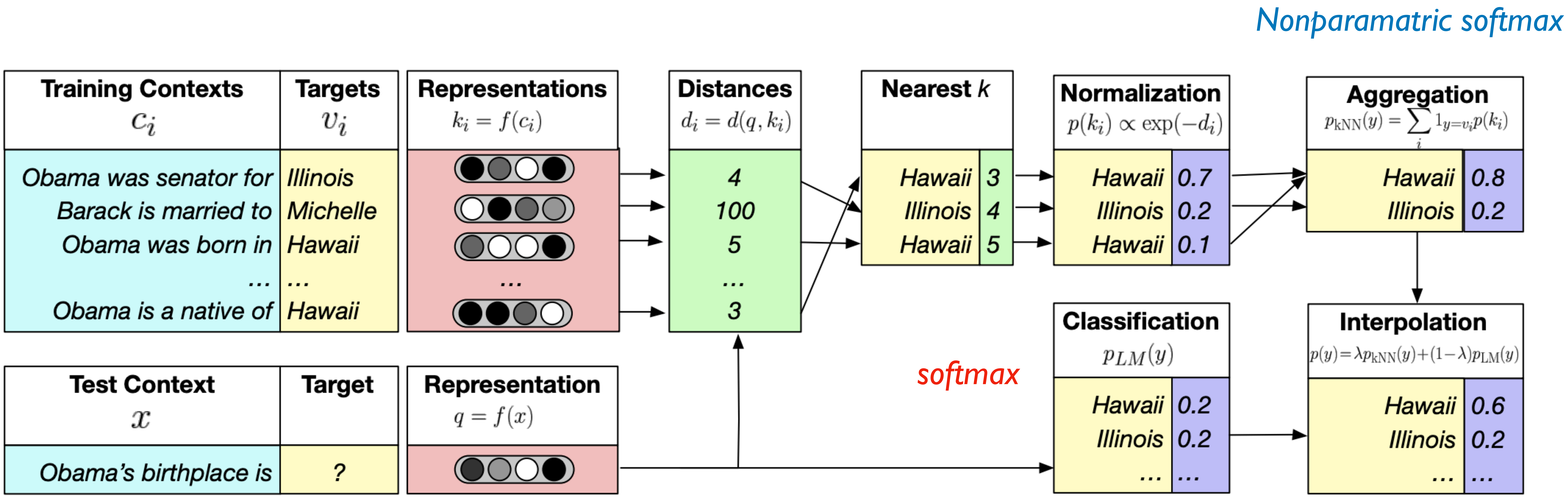
kNN-LM

Nonparamatric softmax



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

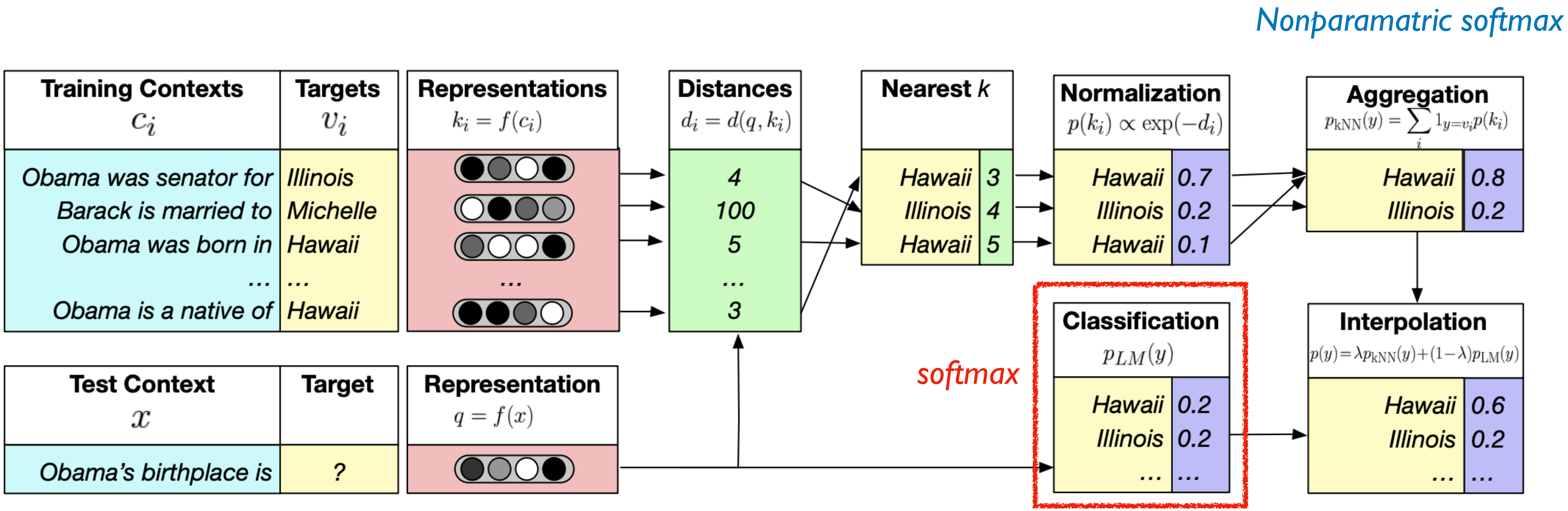
kNN-LM



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

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

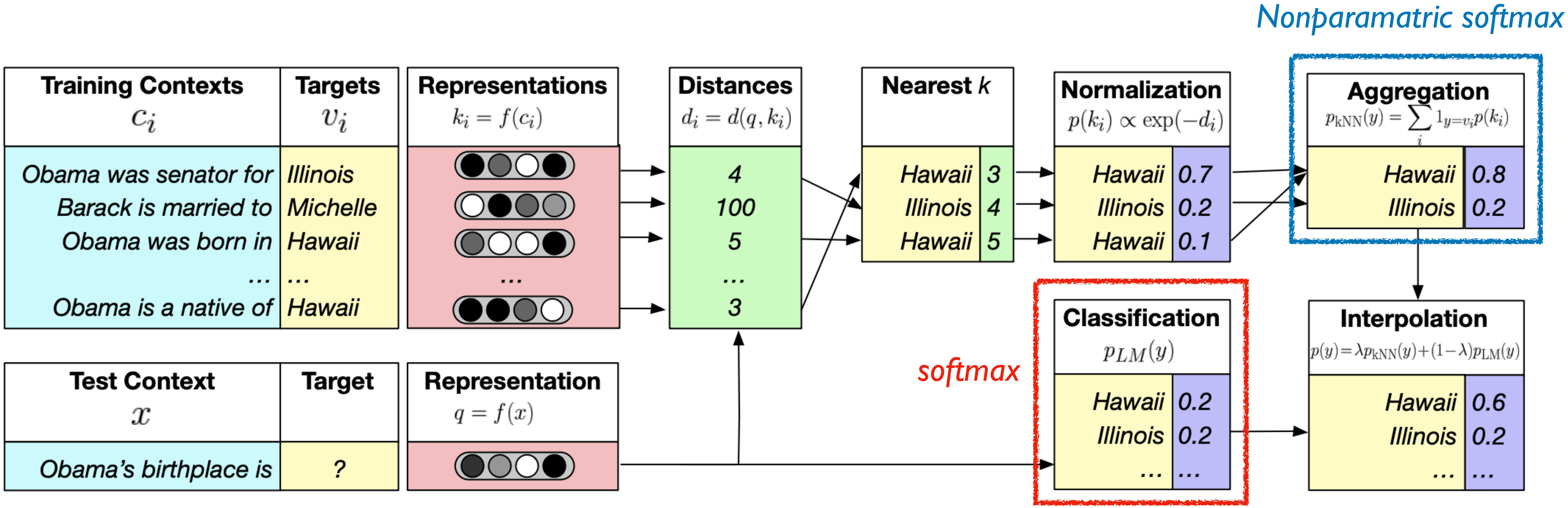
kNN-LM



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

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

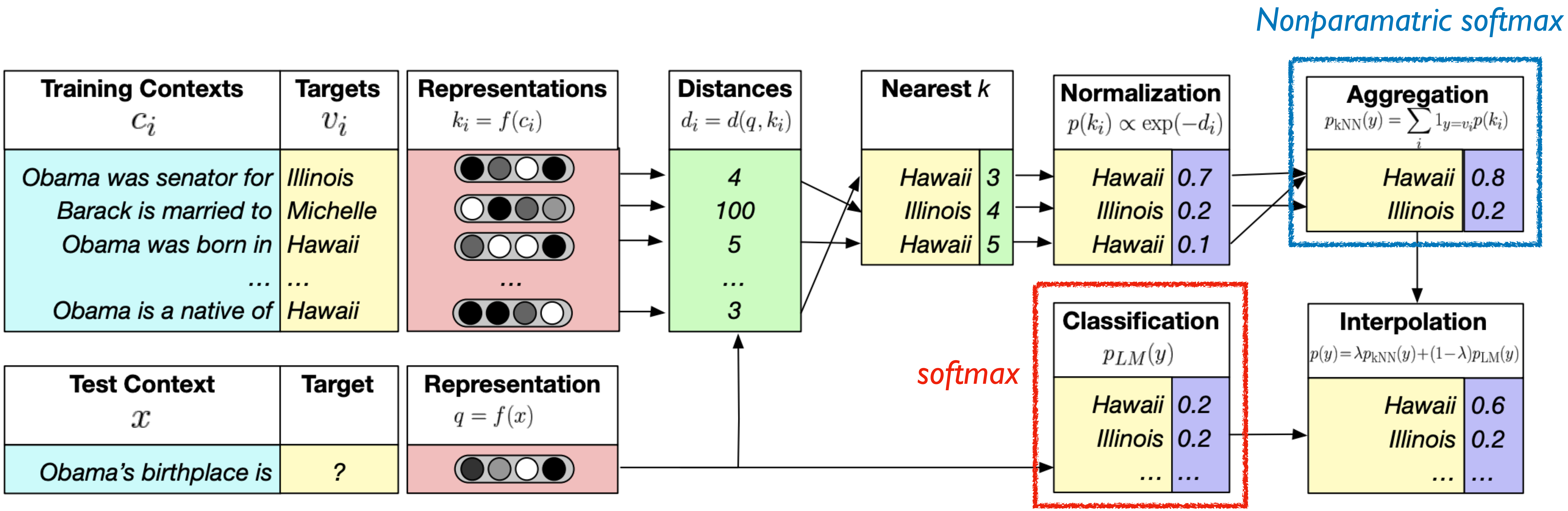
kNN-LM



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

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

kNN-LM



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

λ: hyperparameter

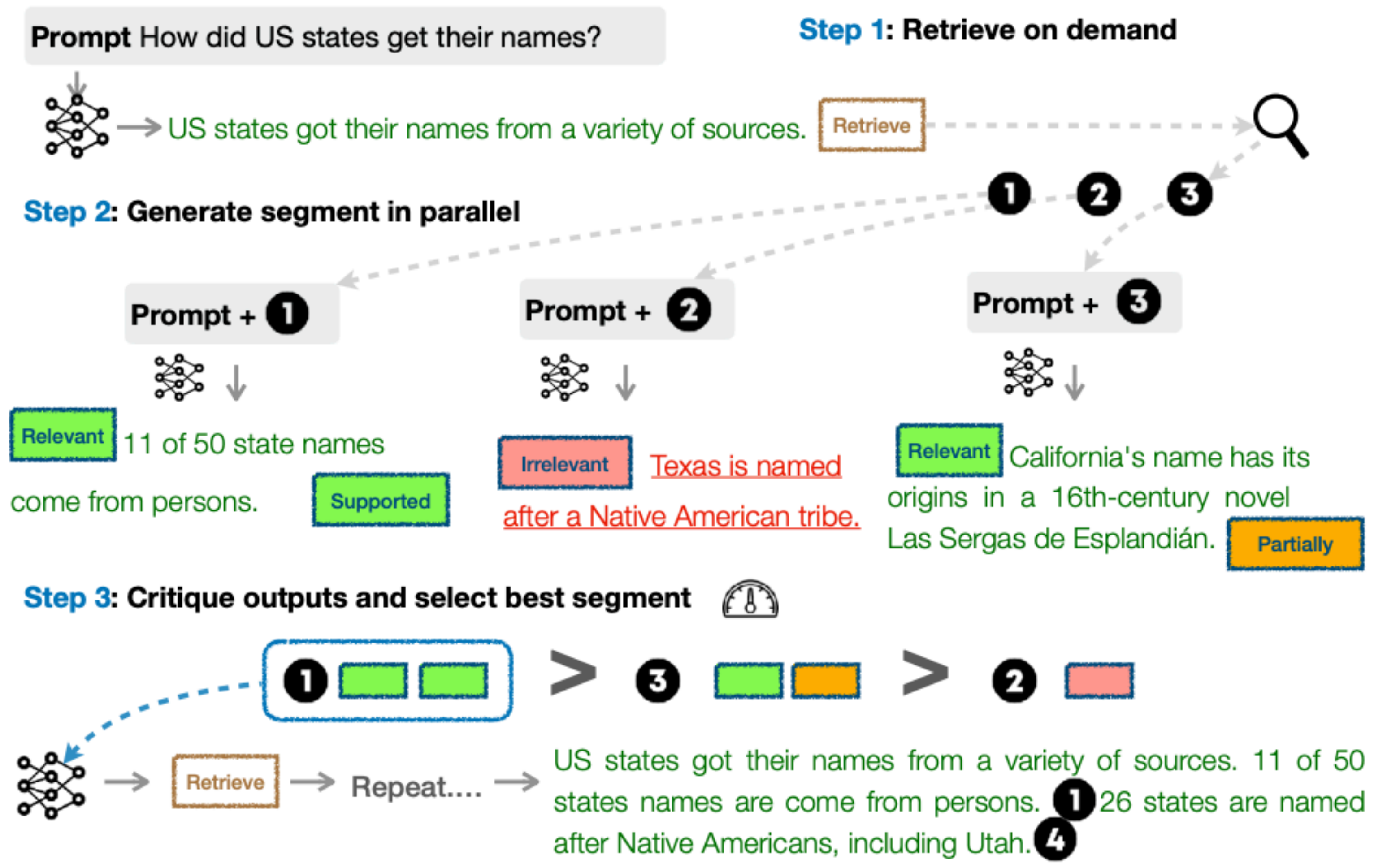
Khandelwal et al. 2020. "Generalization through Memorization: Nearest Neighbor Language Models"

Different architectures: Bigger context

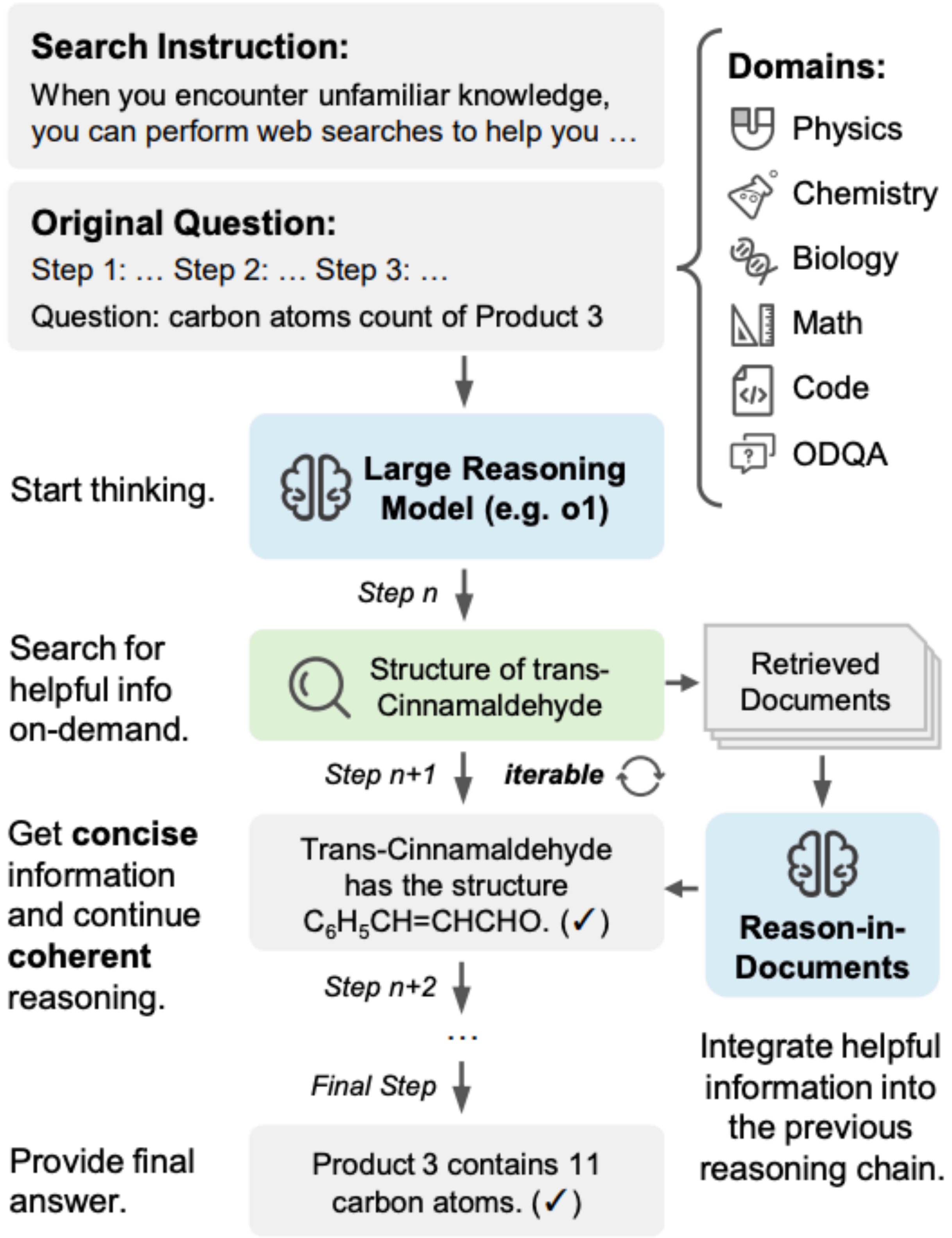
- Different architectures were proposed to address certain limitations of the two-stage pipeline, e.g., inefficiency, retrieval granularity, retrieval frequency, etc.
- As it typically modifies the architecture of Transformers, it requires training an LM extensively.
- How to train them at scale remains an open question.
- For this reason, today's most widely used retrieval-based LMs remain to be the two stage pipeline approach.

More complex pipelines

More complex pipelines

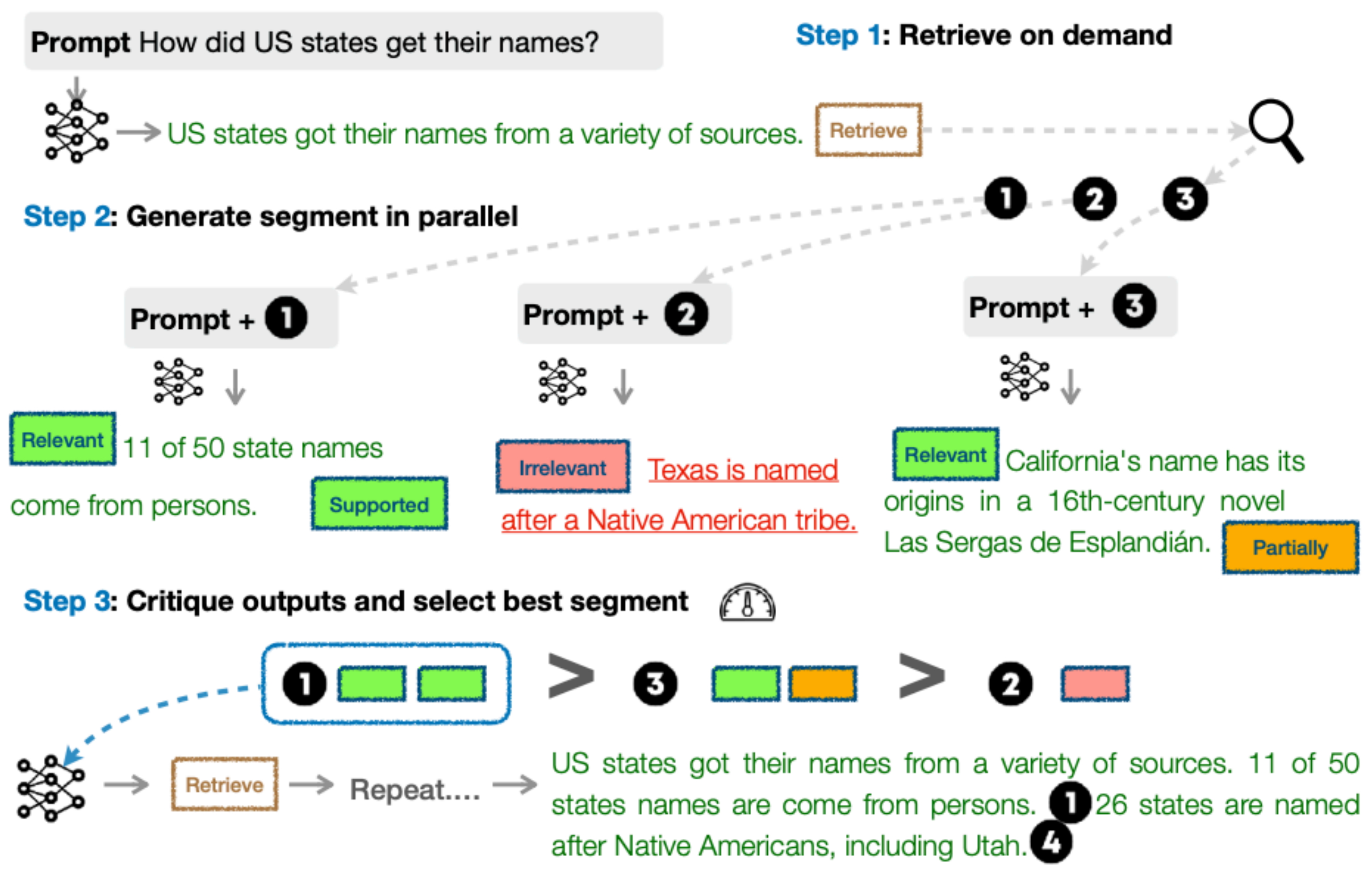


Asai at al. 2024. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection"



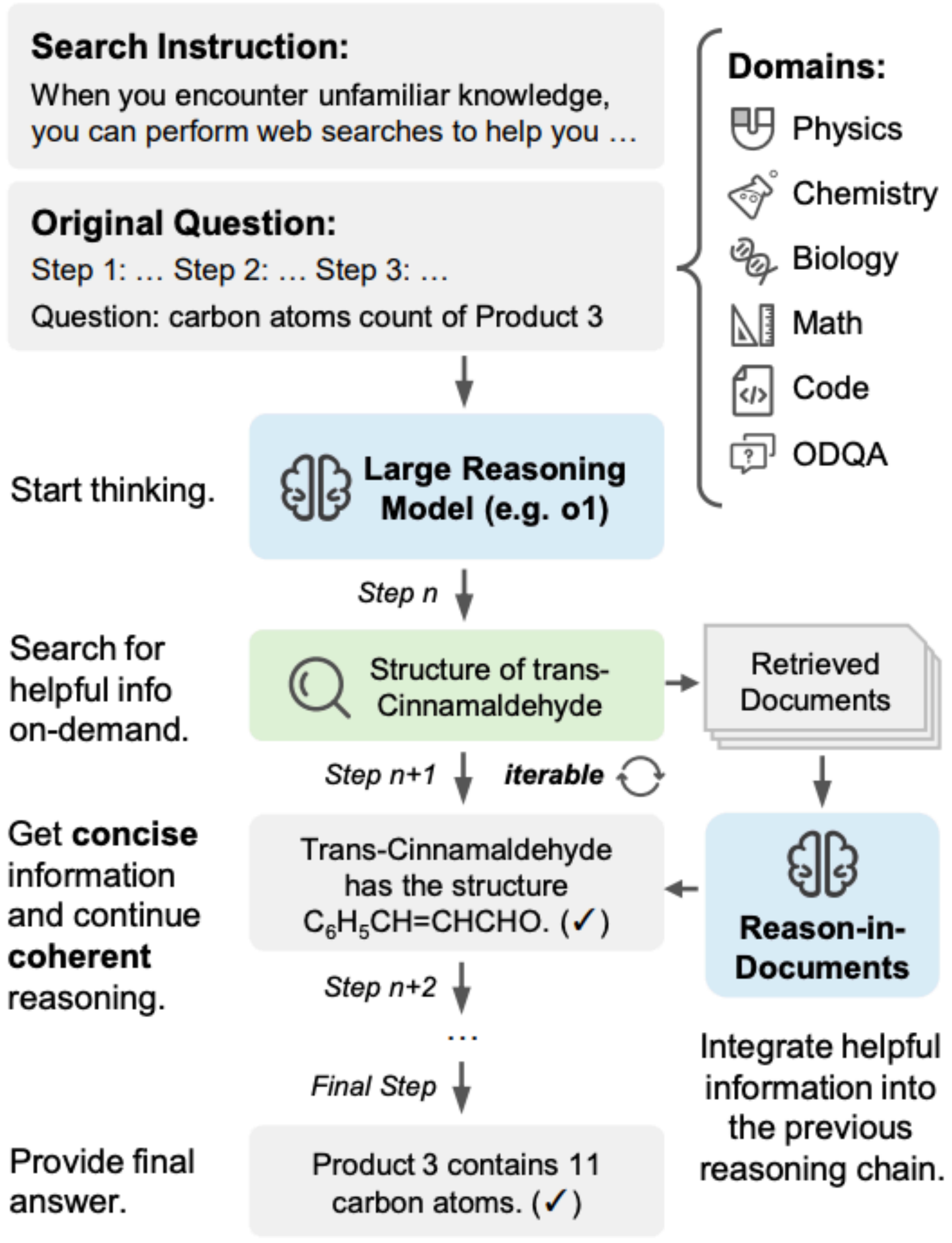
Li et al. "Search-o1: Agentic Search-Enhanced Large Reasoning Models"

More complex pipelines



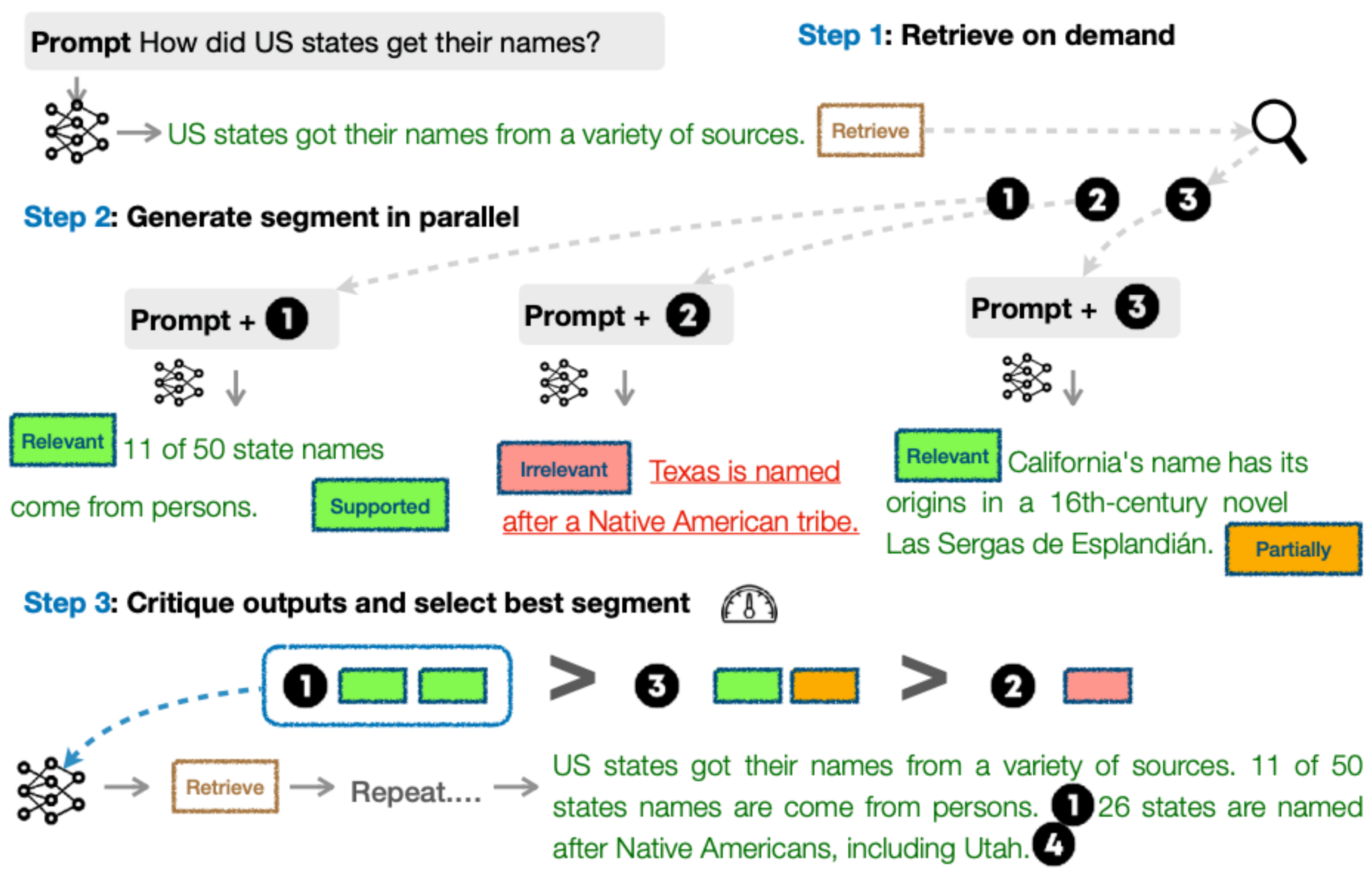
Asai at al. 2024. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection"

Key ideas: Make it an agent system with functions such as



Li et al. "Search-o1: Agentic Search-Enhanced Large Reasoning Models"

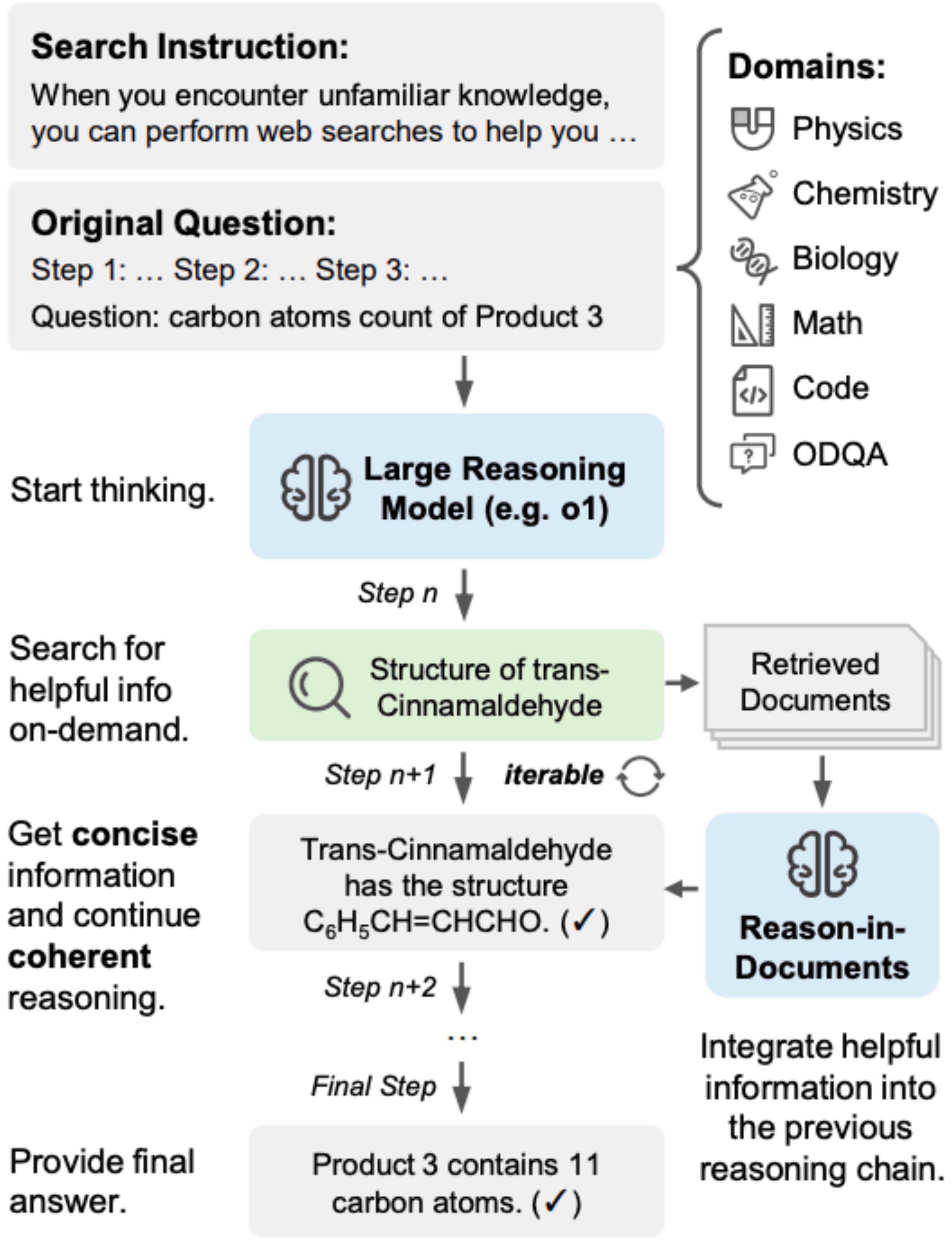
More complex pipelines



Asai at al. 2024. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection"

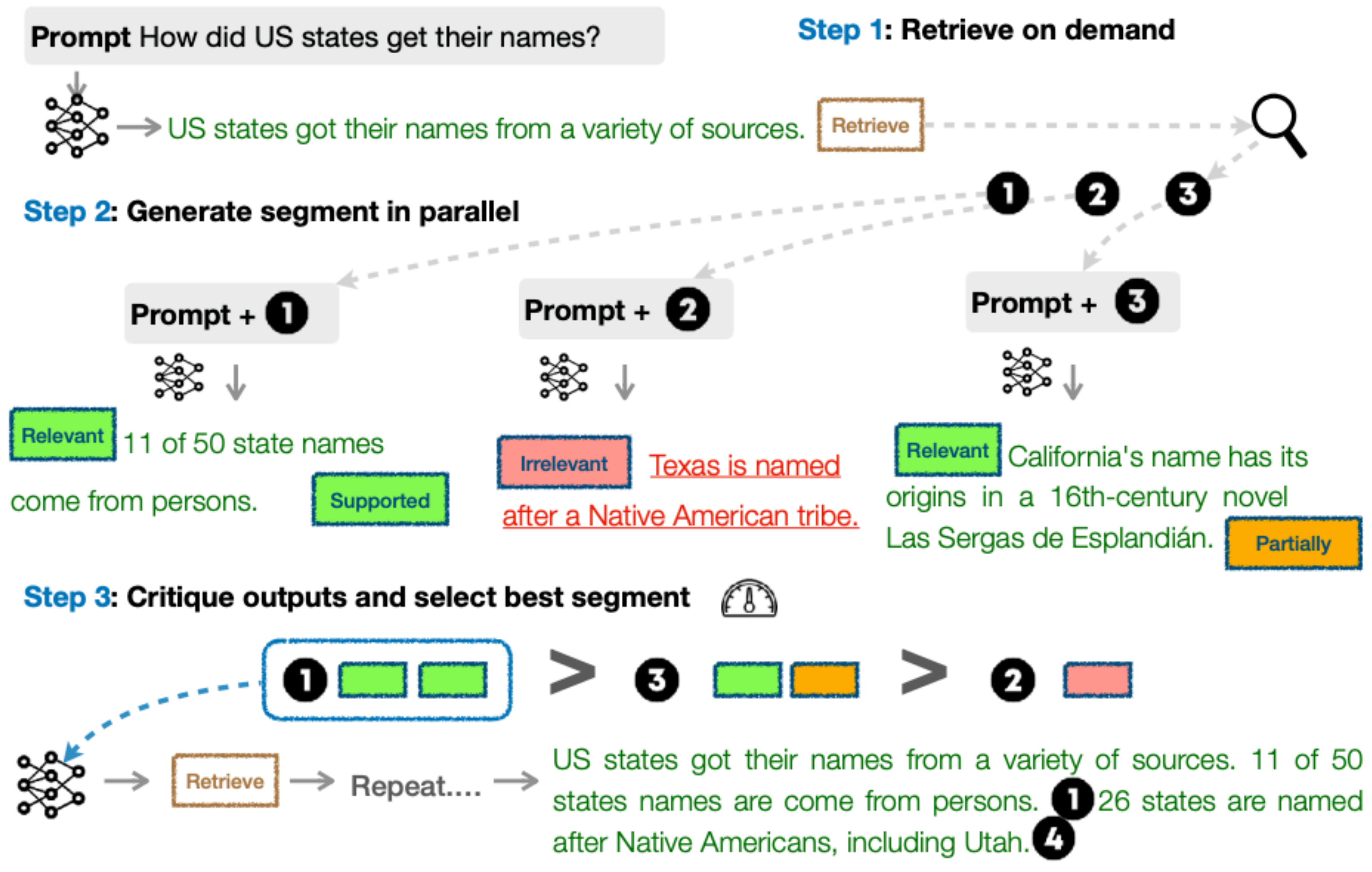
Key ideas: Make it an agent system with functions such as

1. Deciding when to use retrieval



Li et al. "Search-o1: Agentic Search-Enhanced Large Reasoning Models"

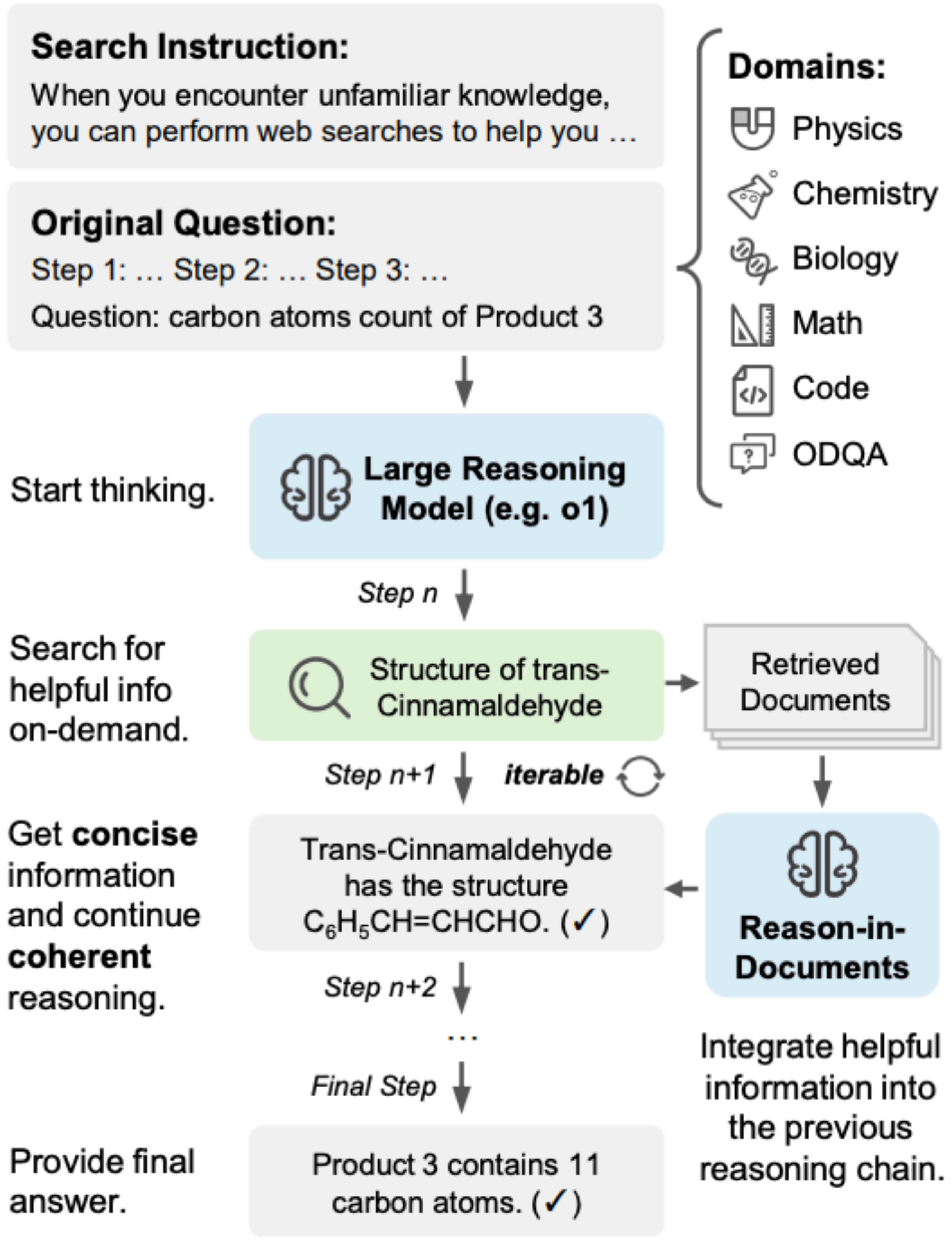
More complex pipelines



Asai at al. 2024. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection"

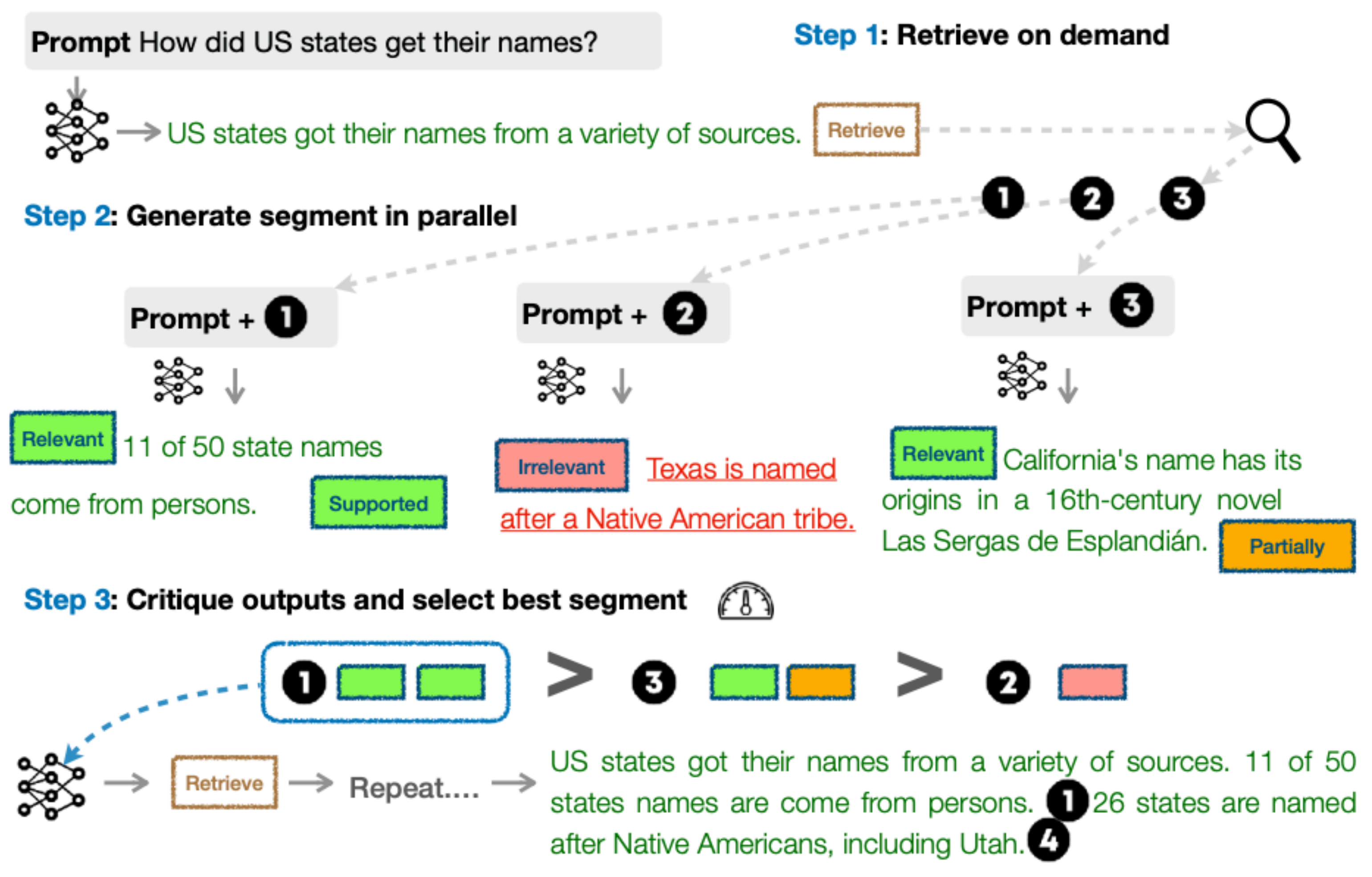
Key ideas: Make it an agent system with functions such as

1. Deciding when to use retrieval
2. Generating a retrieval query

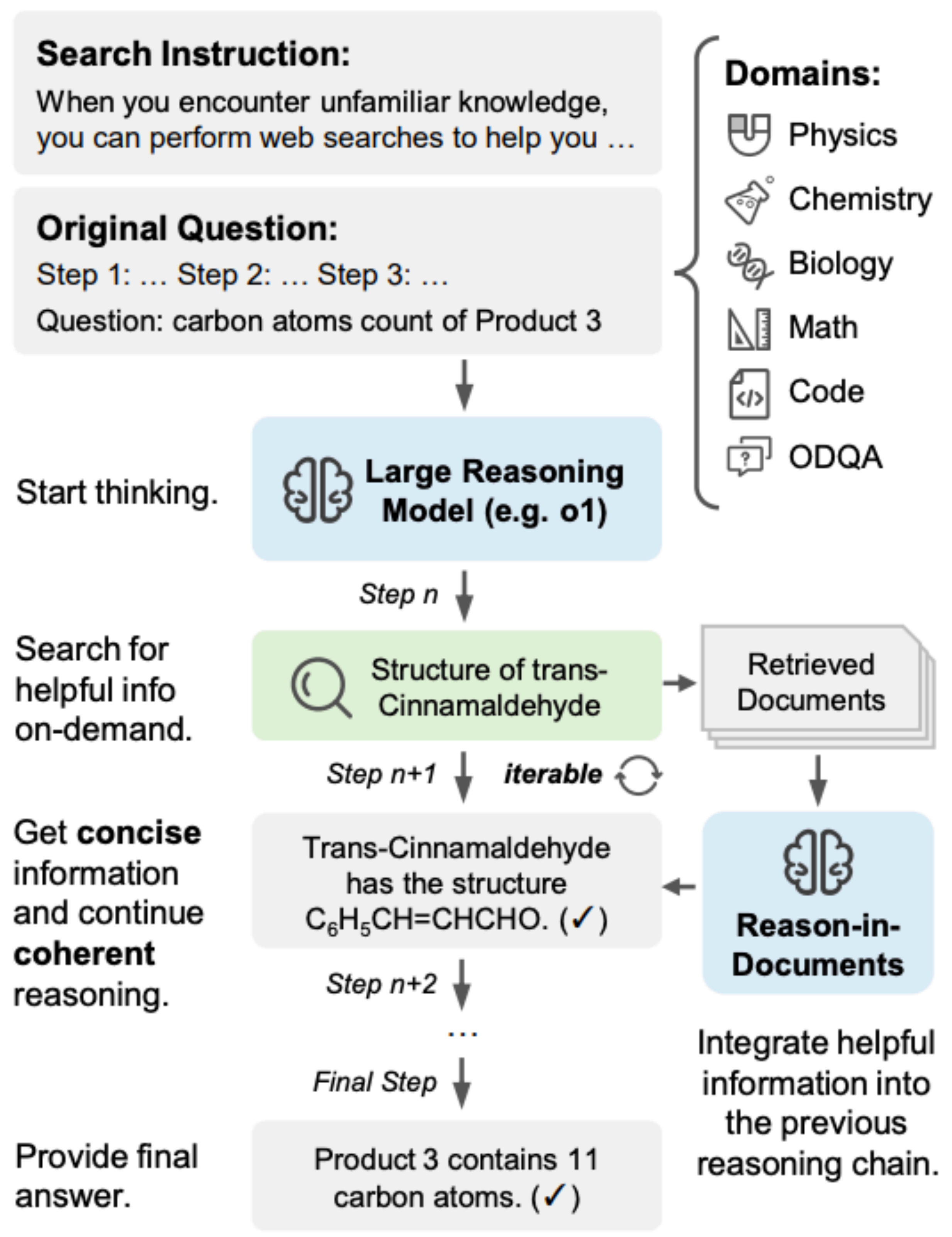


Li et al. "Search-o1: Agentic Search-Enhanced Large Reasoning Models"

More complex pipelines



Asai at al. 2024. “Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection”

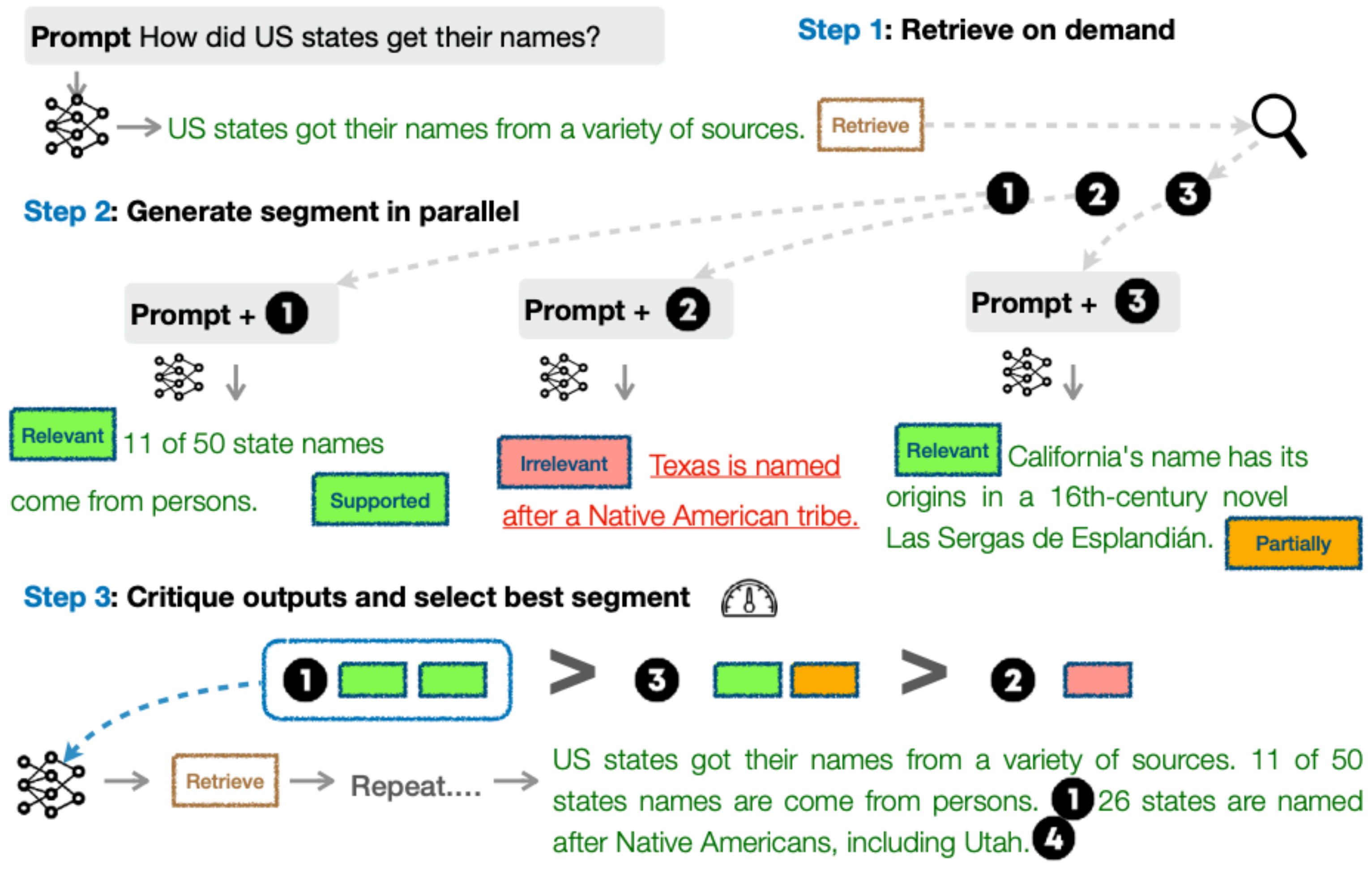


Li et al. "Search-o1: Agentic Search-Enhanced Large Reasoning Models"

Key ideas: Make it an agent system with functions such as

1. Deciding when to use retrieval
2. Generating a retrieval query
3. Reranking/adaptive adoption—use only relevant retrieved passages

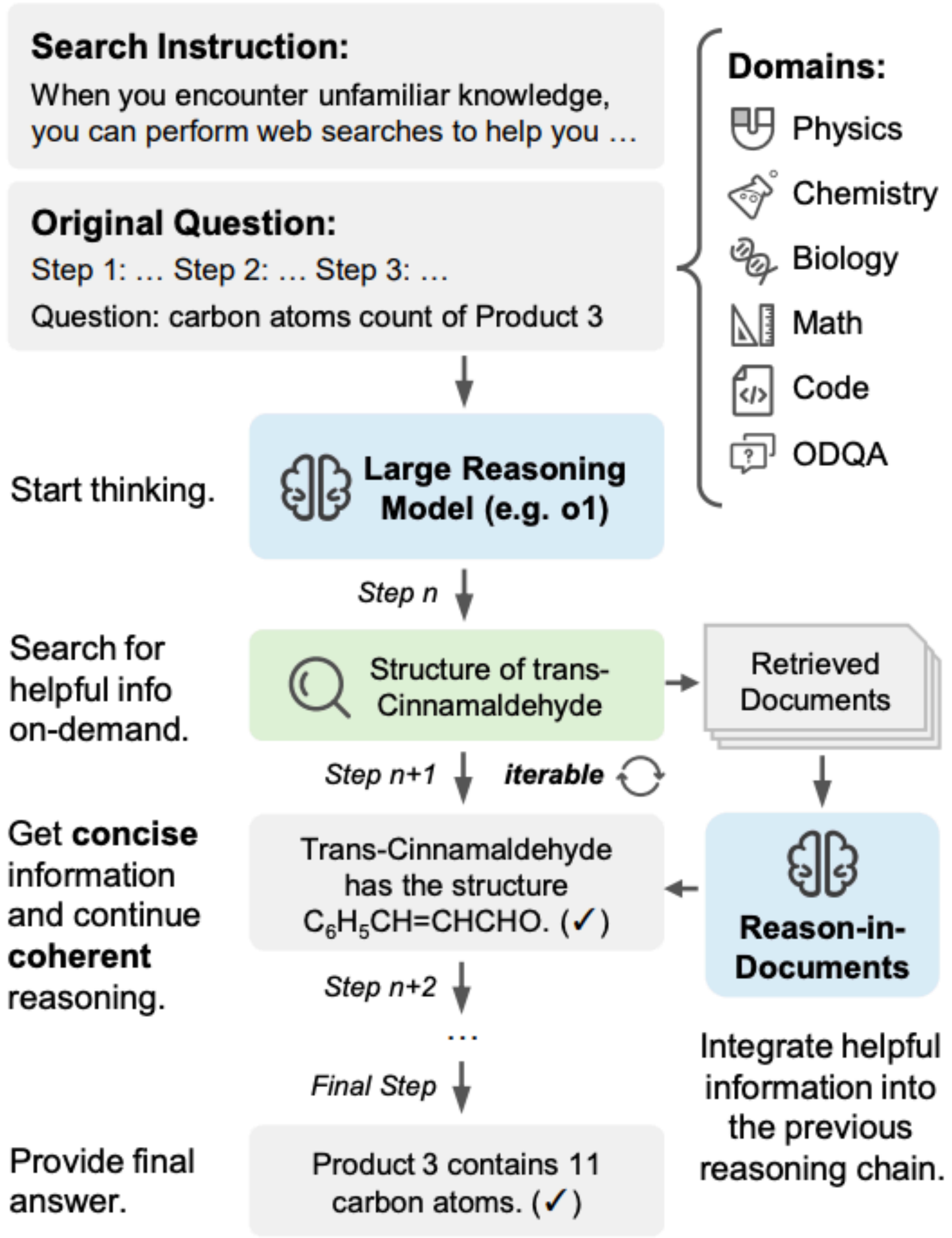
More complex pipelines



Asai at al. 2024. "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection"

Key ideas: Make it an agent system with functions such as

1. Deciding when to use retrieval
2. Generating a retrieval query
3. Reranking/adaptive adoption—use only relevant retrieved passages
4. Rewriting—make passages more comprehensible & include relevant info only



Li et al. "Search-o1: Agentic Search-Enhanced Large Reasoning Models"

QnA for Part 1

Today's Lecture

Part 1. **Basics** of retrieval-based LMs (35min)

- Retrieval
- Augmentation
- Training of retrieval-based LMs

Part 2. **Recent research** on *scaling* retrieval-based LMs (35min)

- Scalable Pre-training with Retrieval
- Scaling a Datastore
- Datastore for Responsible Data Use

Open Problems (10min)

Today's Lecture

Part 1. **Basics** of retrieval-based LMs (35min)

- Retrieval
- Augmentation
- Training of retrieval-based LMs

Part 2. **Recent research** on *scaling* retrieval-based LMs (35min)

- Scalable Pre-training with Retrieval
- Scaling a Datastore
- Datastore for Responsible Data Use

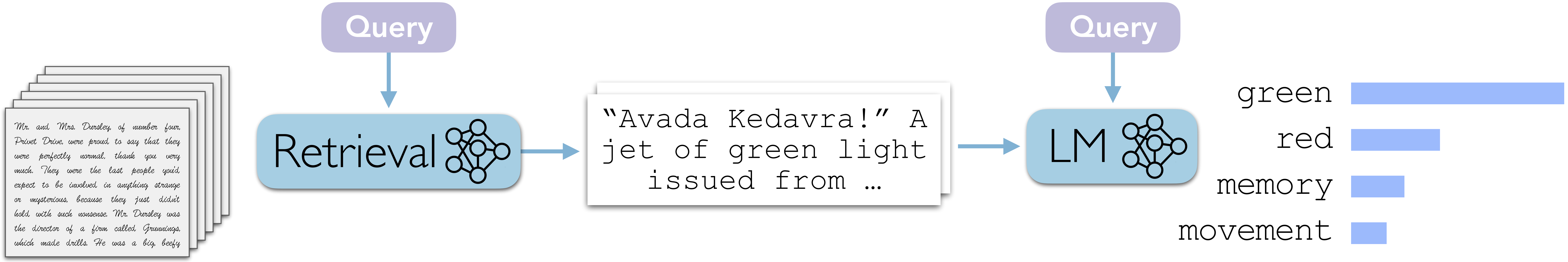
Open Problems (10min)

Is scaling important?

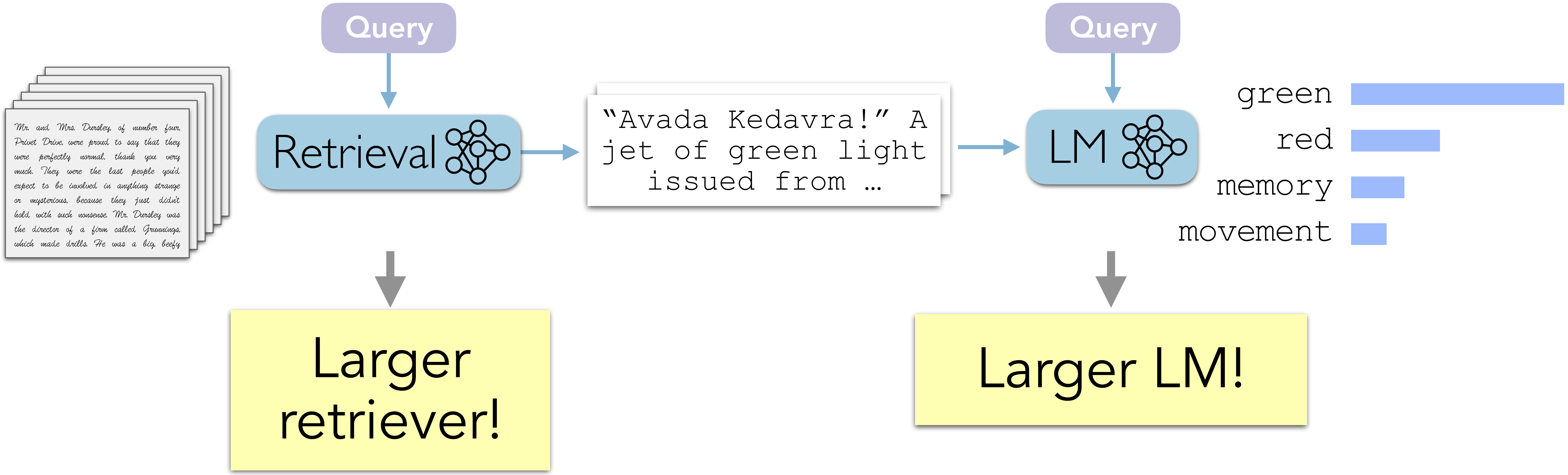
~~Is scaling important?~~

How to scale?

Scaling in retrieval-based LMs?

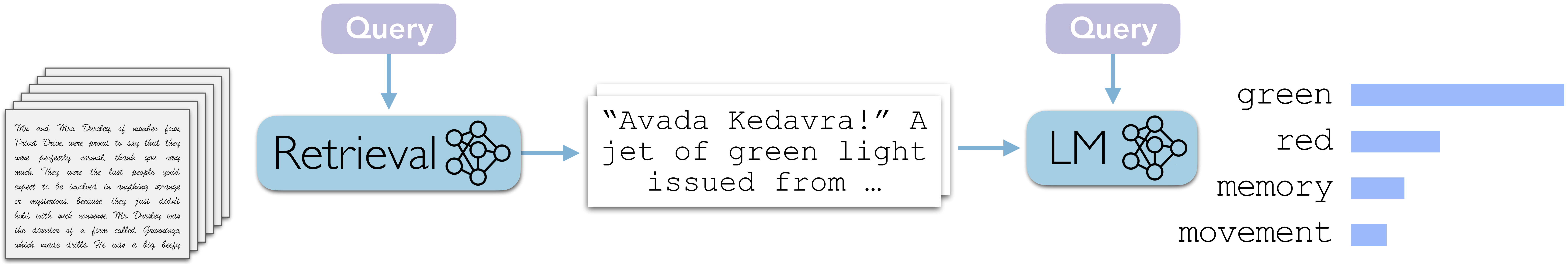


Scaling in retrieval-based LMs?

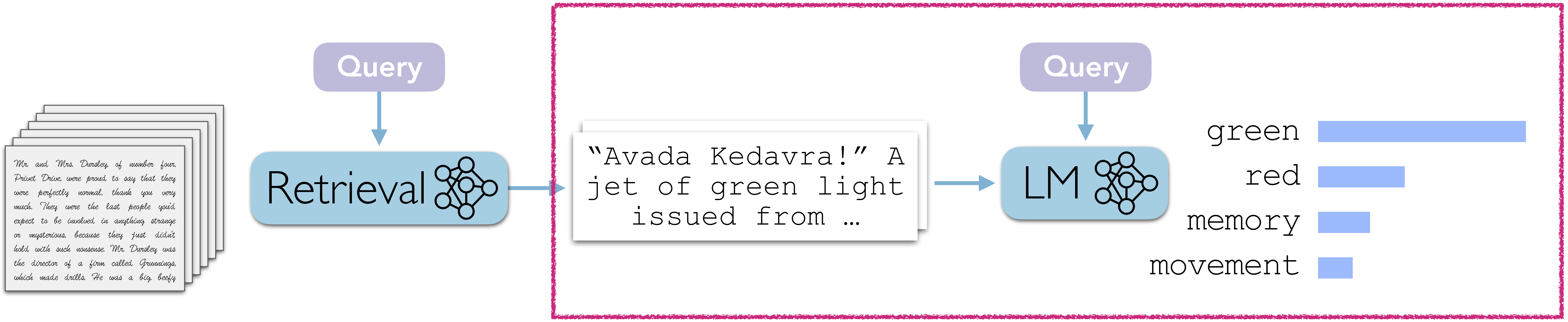


Ni et al. "Large Dual Encoders Are Generalizable Retrievers" and other work

New scaling of retrieval-based LMs

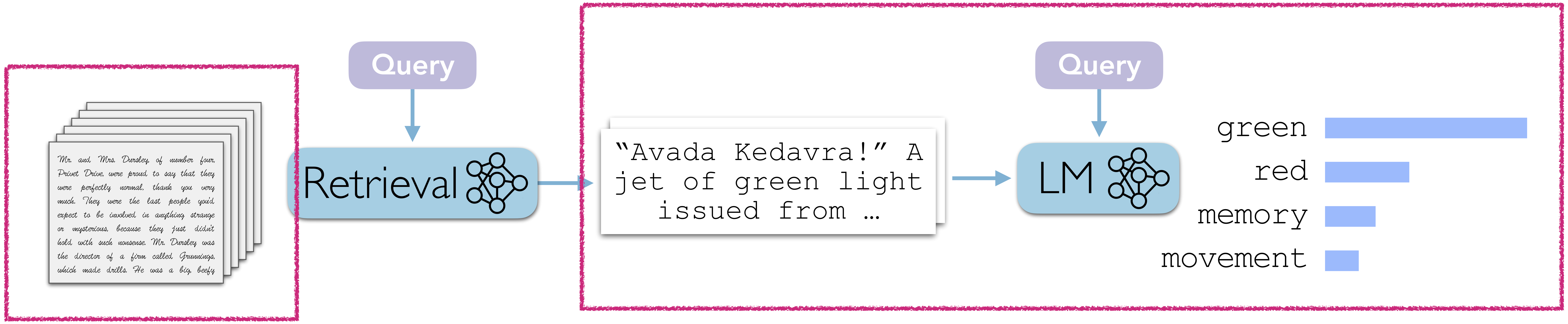


New scaling of retrieval-based LMs



1) Scaling training w/ retrieval

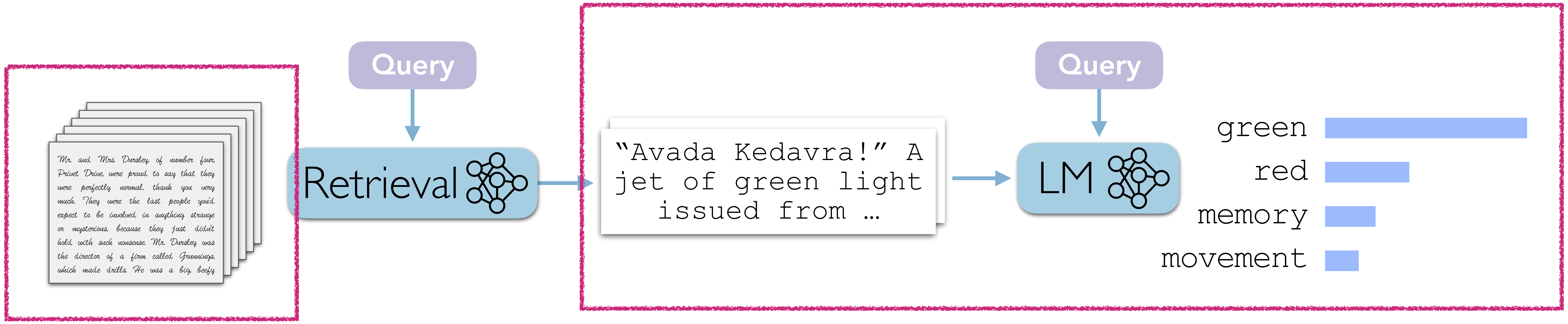
New scaling of retrieval-based LMs



2) Scaling a datastore

1) Scaling training w/ retrieval

New scaling of retrieval-based LMs

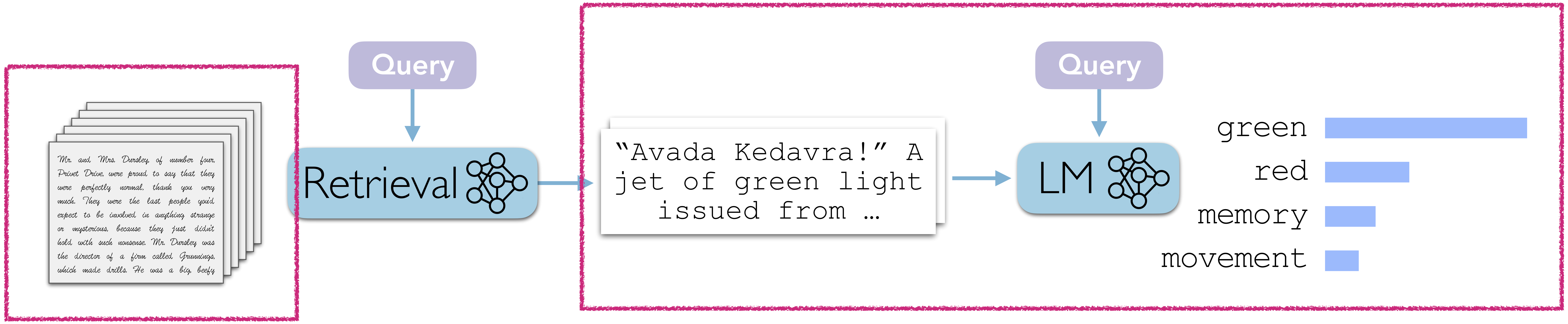


2) Scaling a datastore

1) Scaling training w/ retrieval

3) How to scale with *responsible* data use?

New scaling of retrieval-based LMs



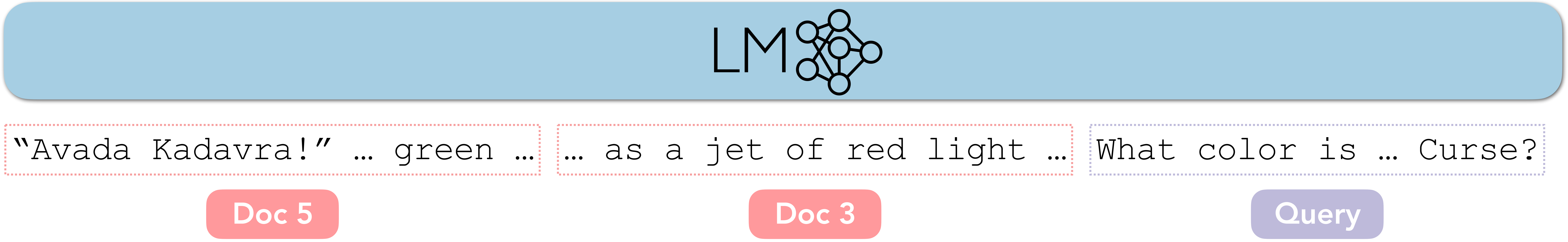
2) Scaling a datastore

1) Scaling training w/ retrieval

3) How to scale with *responsible* data use?

Motivation

Motivation


LM
"Avada Kadavra!" … green …
… as a jet of red light …
What color is … Curse?
Doc 5
Doc 3
Query

Motivation


green
red
memory
movement
LM
“Avada Kadavra!” … green …
… as a jet of red light …
What color is … Curse?
Doc 5
Doc 3
Query

Motivation

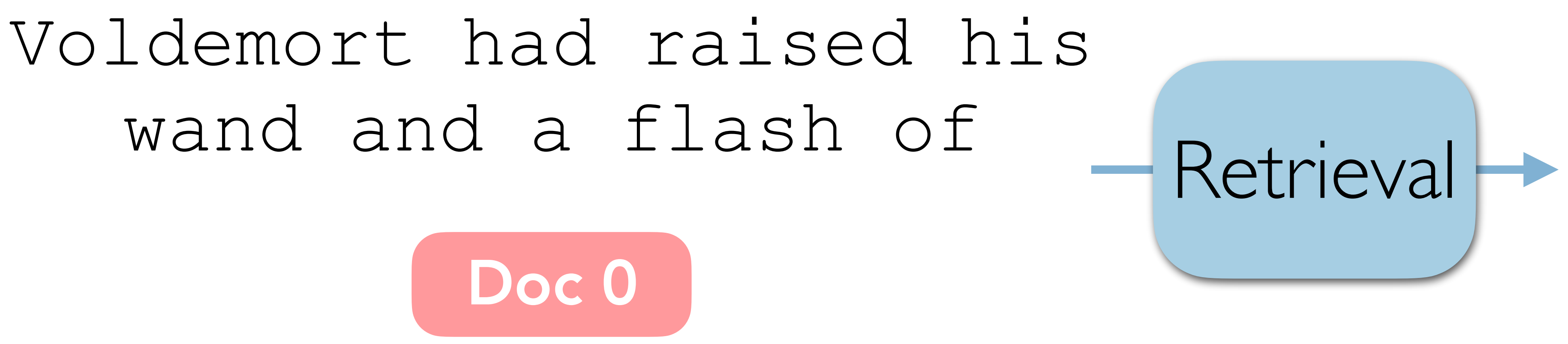
(New to LMs)

Motivation

(New to LMs)

Can we pre-train LMs to make better use of retrieval?

Pre-training w/ retrieval



Pre-training w/ retrieval

Voldemort had raised his wand and a flash of

Doc 0

Retrieval

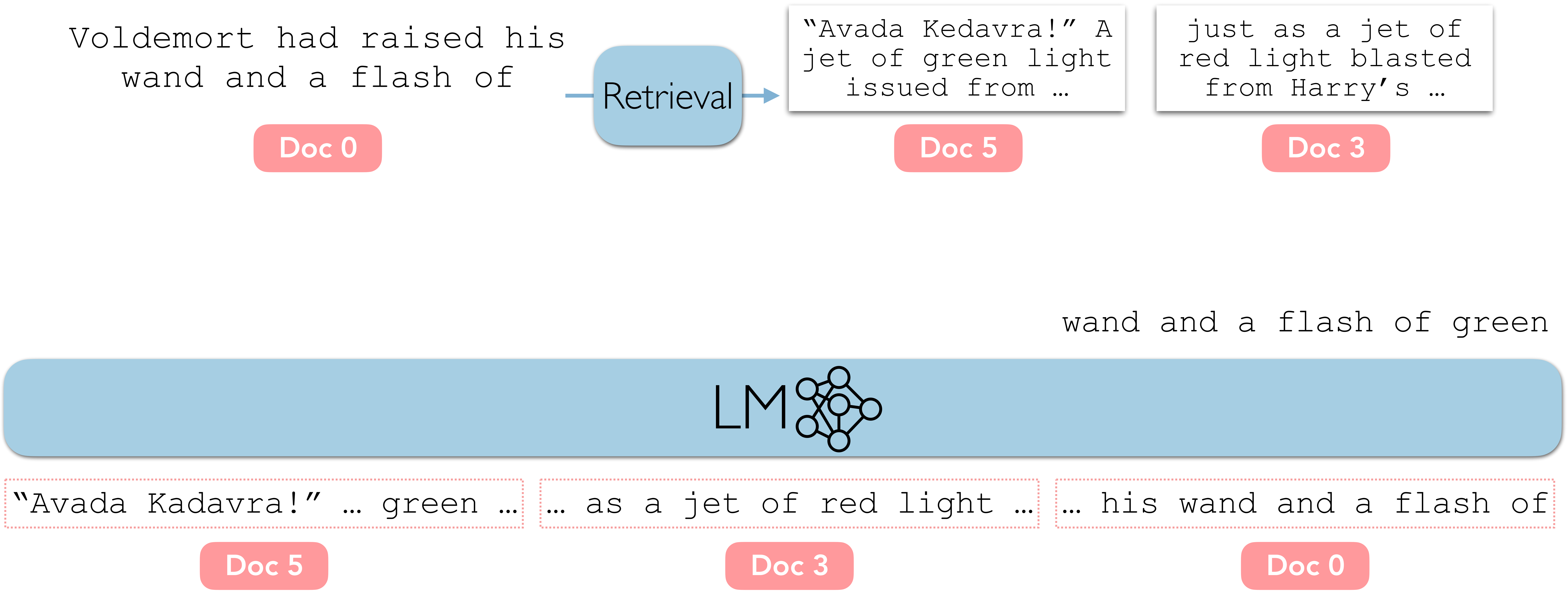
"Avada Kedavra!" A jet of green light issued from …

Doc 5

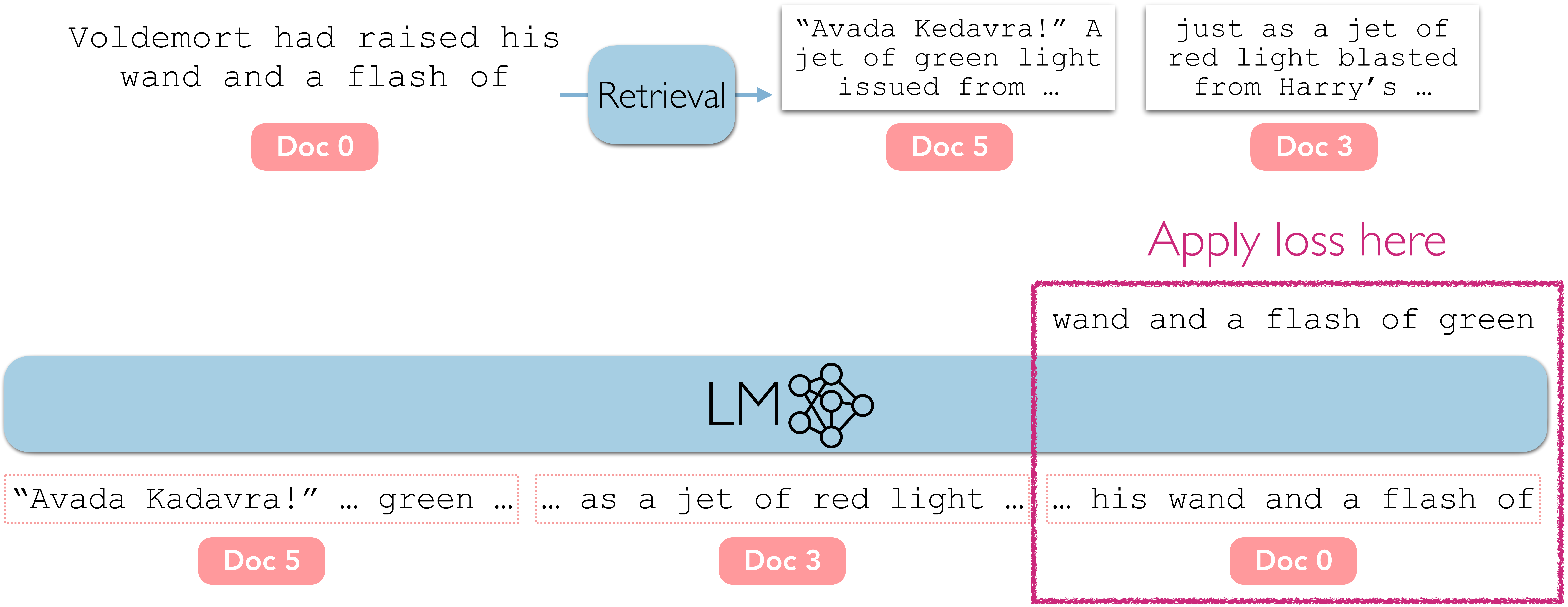
just as a jet of red light blasted from Harry's …

Doc 3

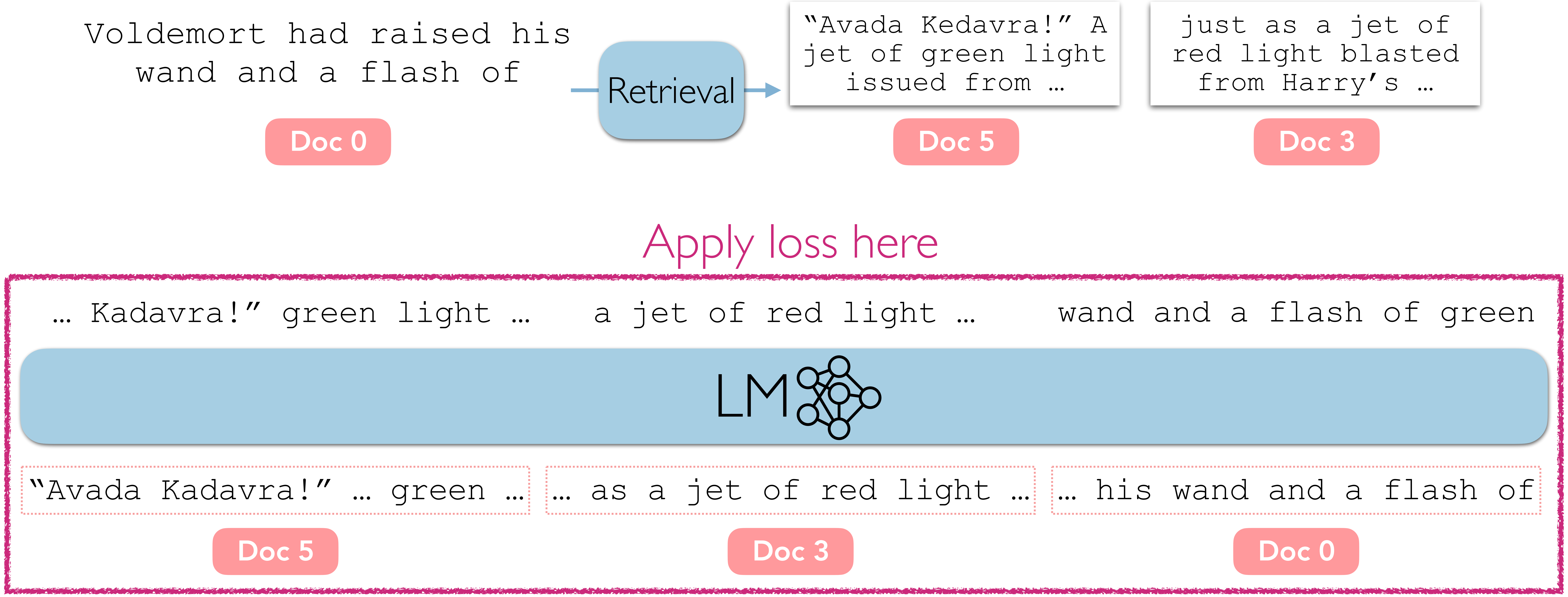
Pre-training w/ retrieval



Pre-training w/ retrieval



Pre-training w/ retrieval



Pre-training w/ retrieval: **Duplication problem**

Pre-training w/ retrieval: **Duplication problem**

Voldemort had raised his wand and a flash of

Doc 0

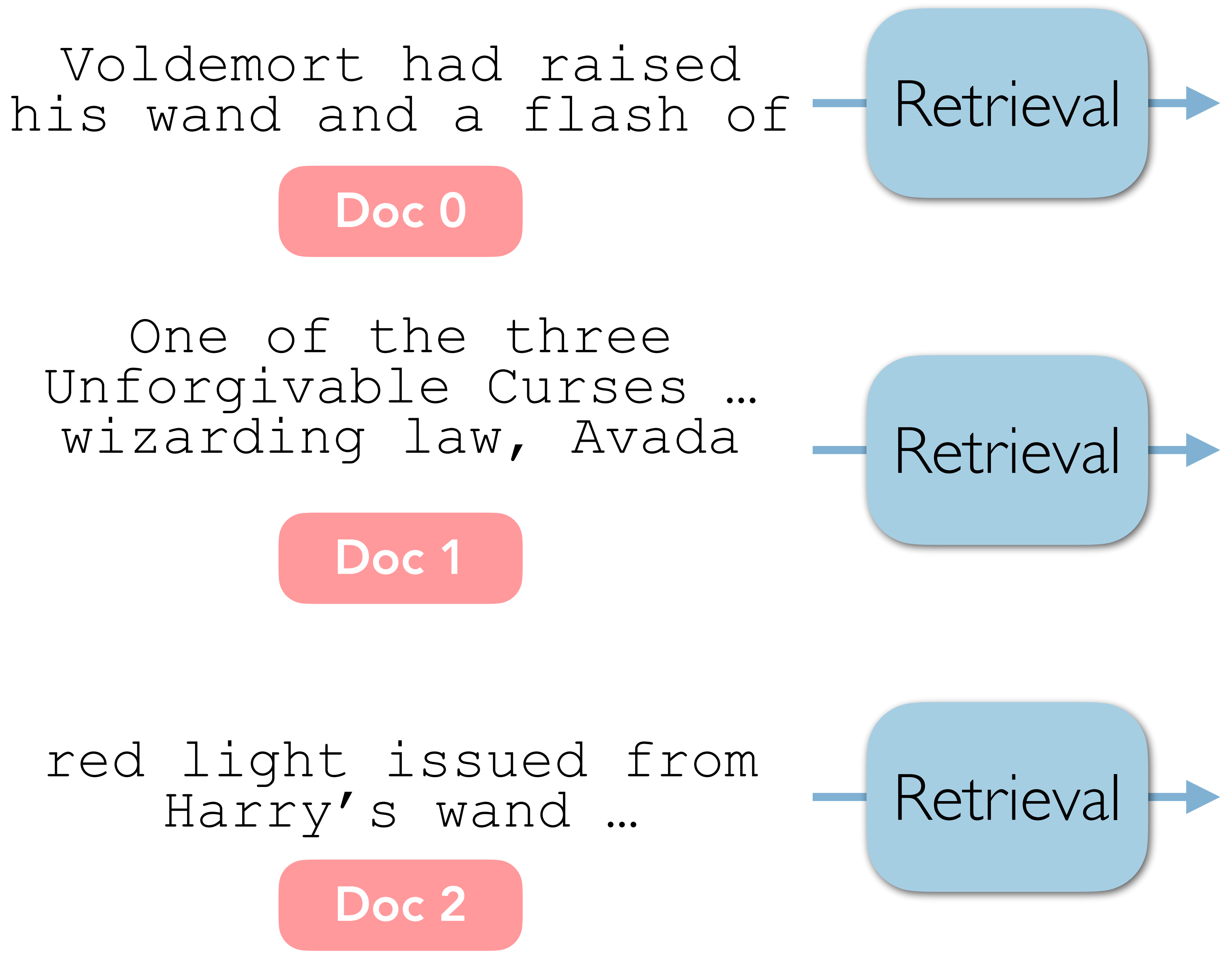
One of the three Unforgivable Curses … wizarding law, Avada

Doc 1

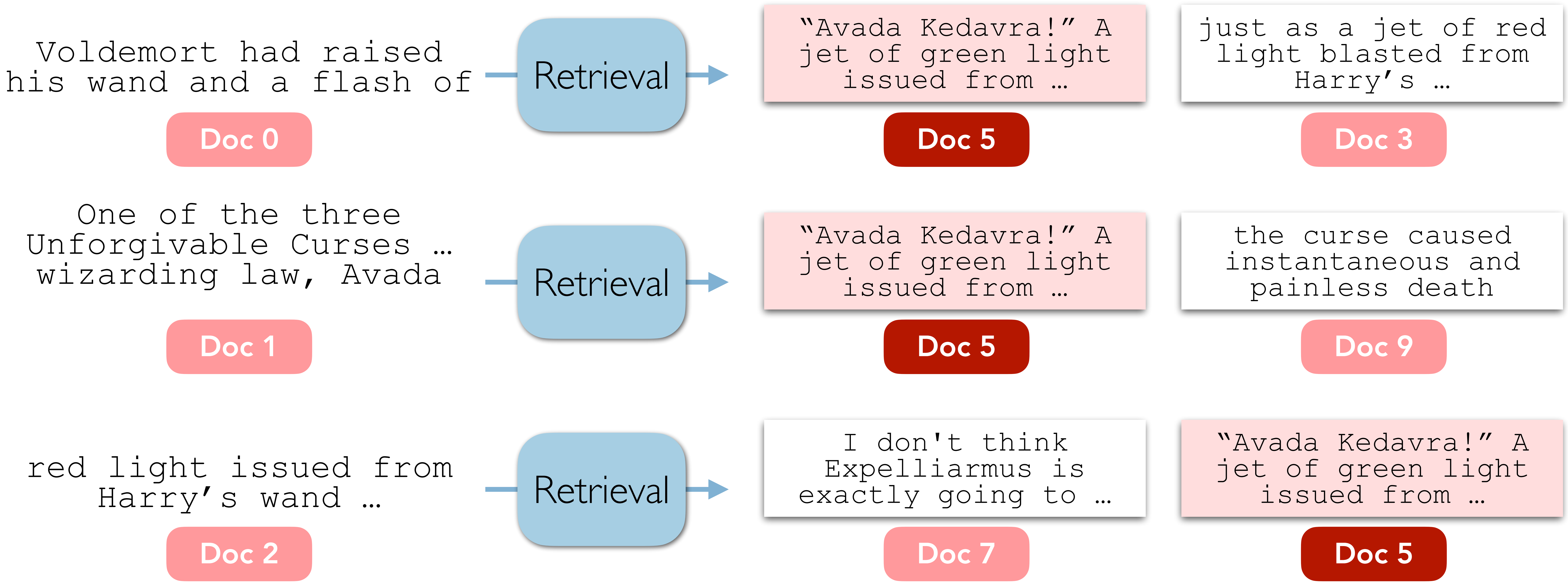
red light issued from Harry's wand …

Doc 2

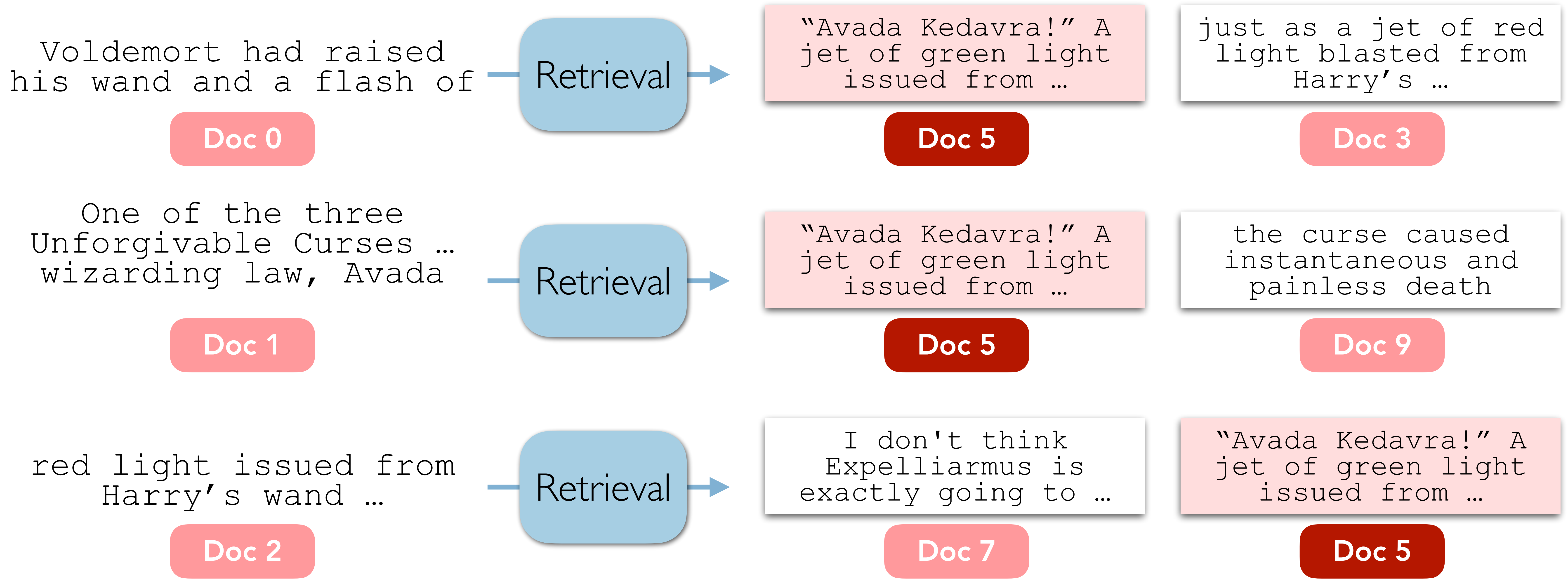
Pre-training w/ retrieval: **Duplication problem**



Pre-training w/ retrieval: **Duplication problem**



Pre-training w/ retrieval: **Duplication problem**



1) LM conditions on a set of ***relevant*** documents

2) Each document appears ***exactly once***

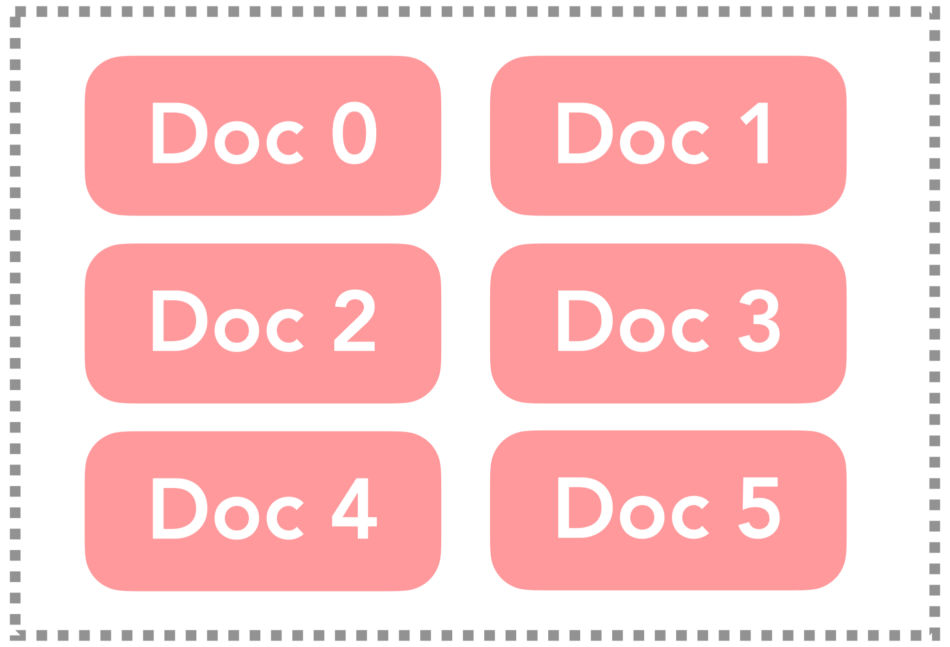
Pre-training w/ retrieval: Proposal

Document ordering problem

Pre-training w/ retrieval: Proposal

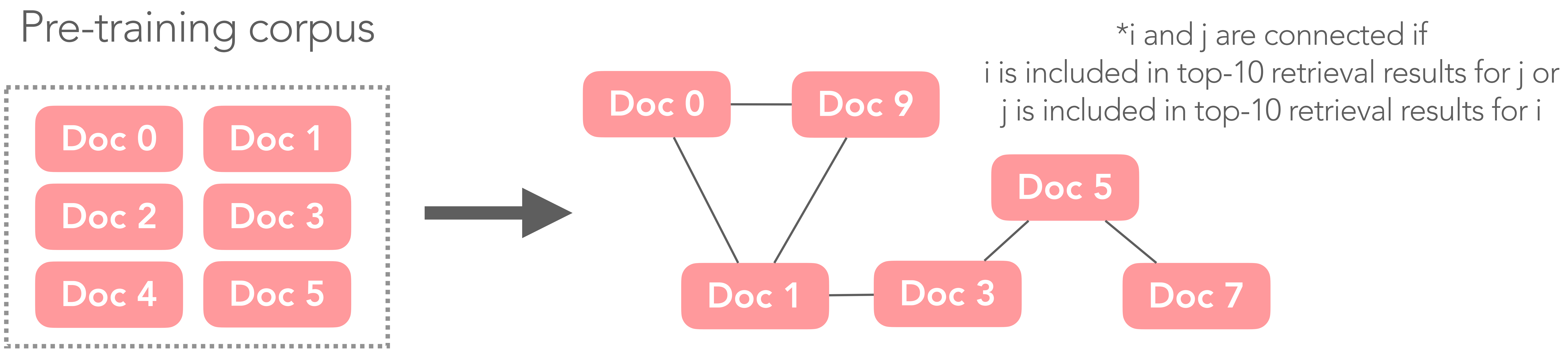
Document ordering problem

Pre-training corpus



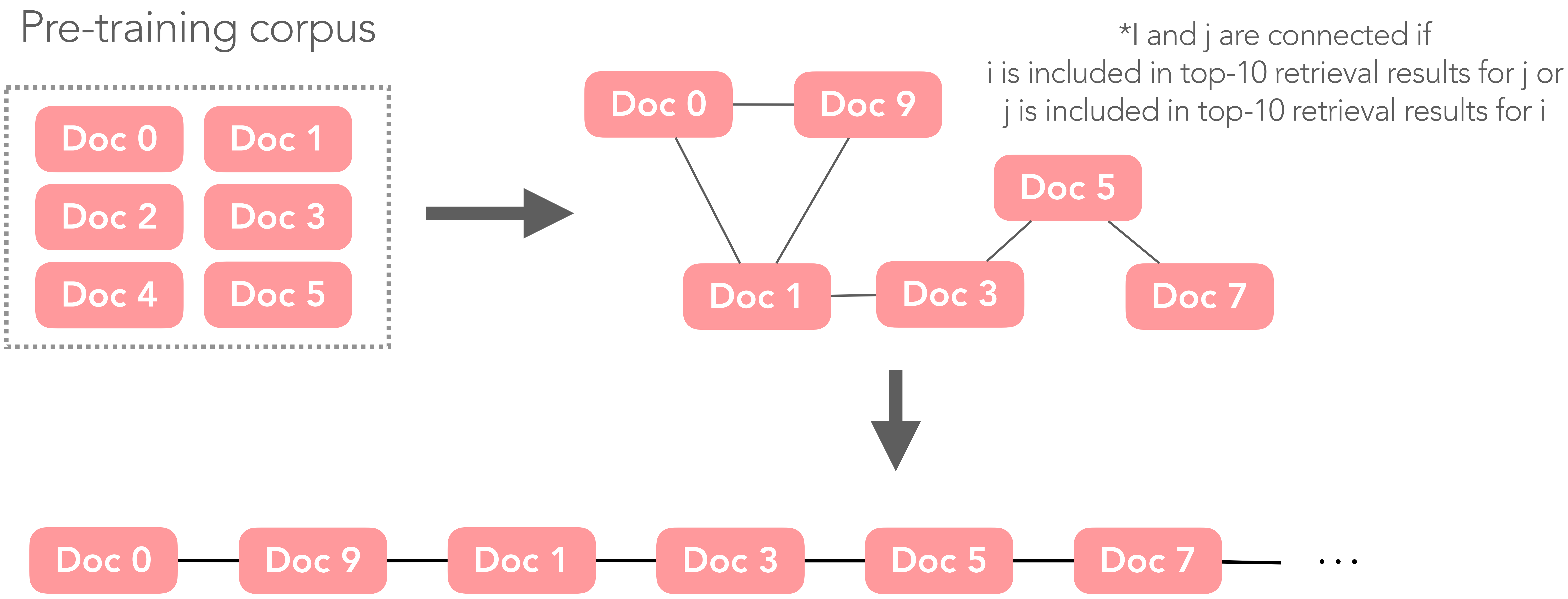
Pre-training w/ retrieval: Proposal

Document ordering problem



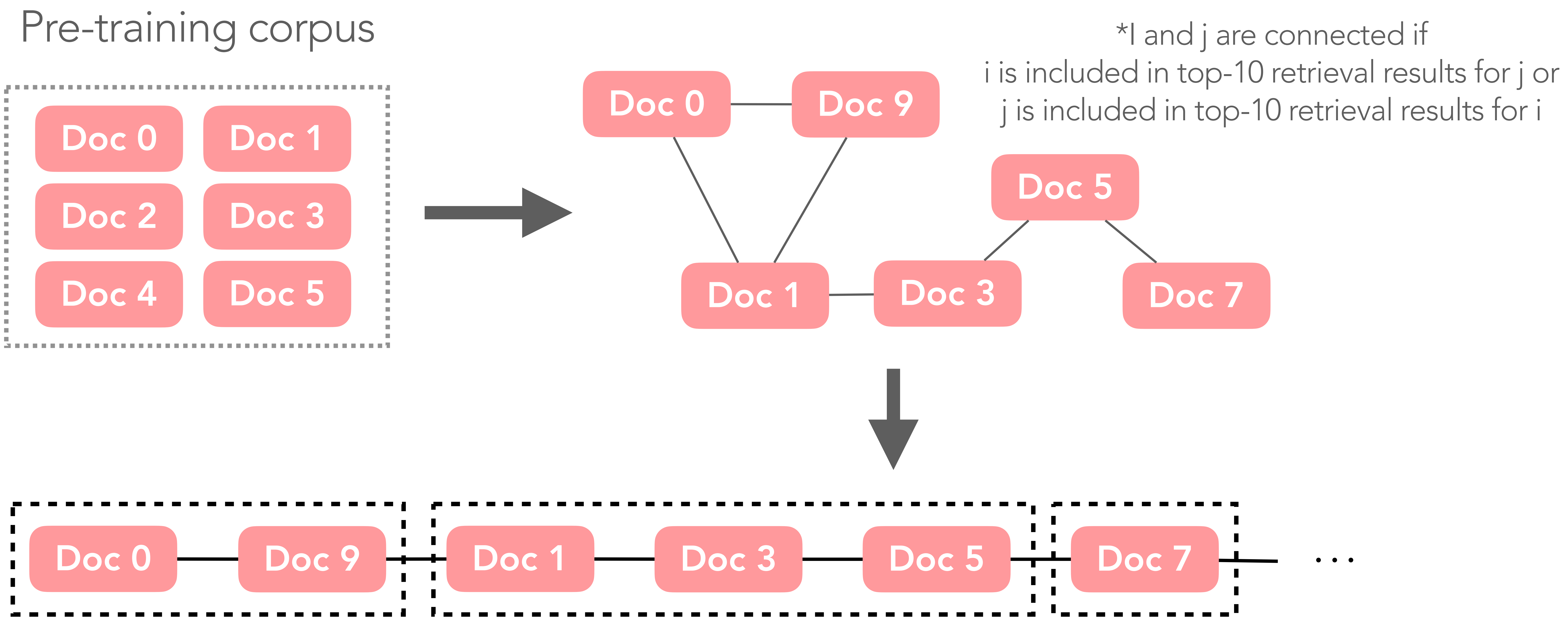
Pre-training w/ retrieval: Proposal

Document ordering problem

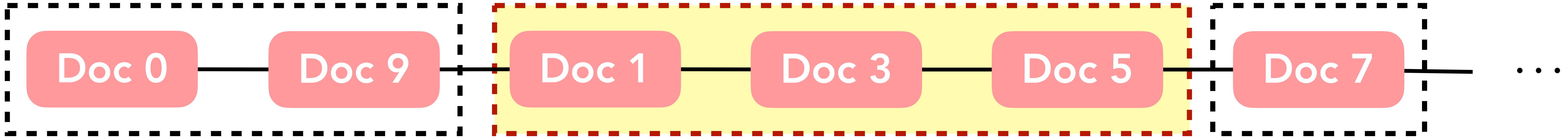


Pre-training w/ retrieval: Proposal

Document ordering problem



Pre-training w/ retrieval: Proposal



"Avada Kadavra!" … green …

Doc 1

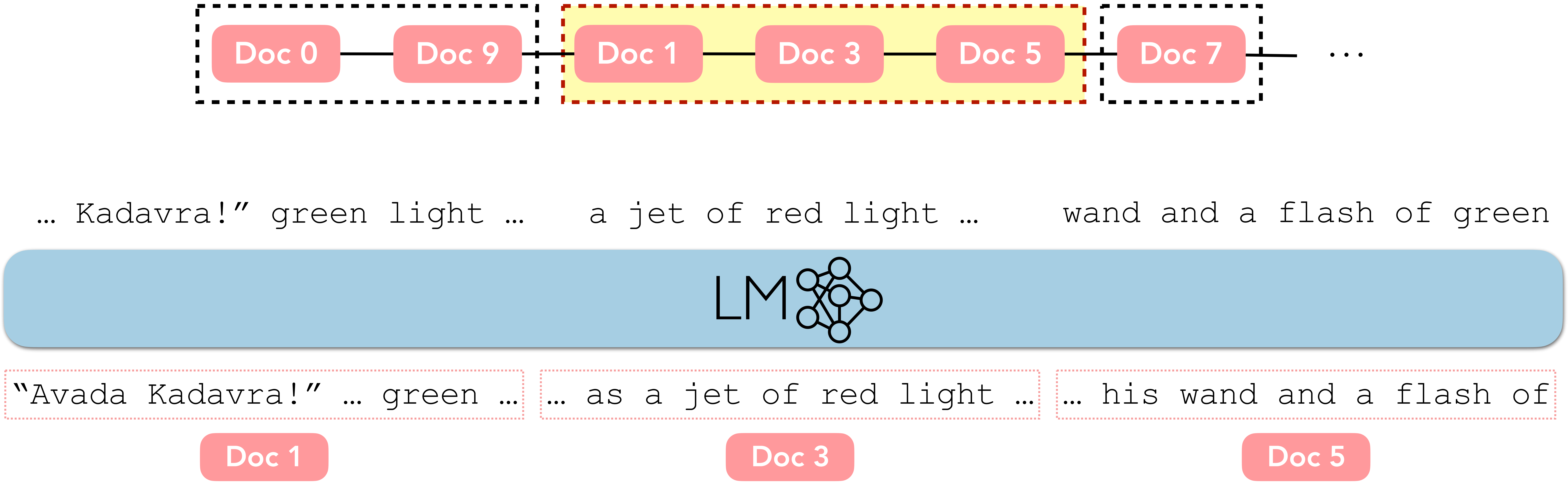
… as a jet of red light …

Doc 3

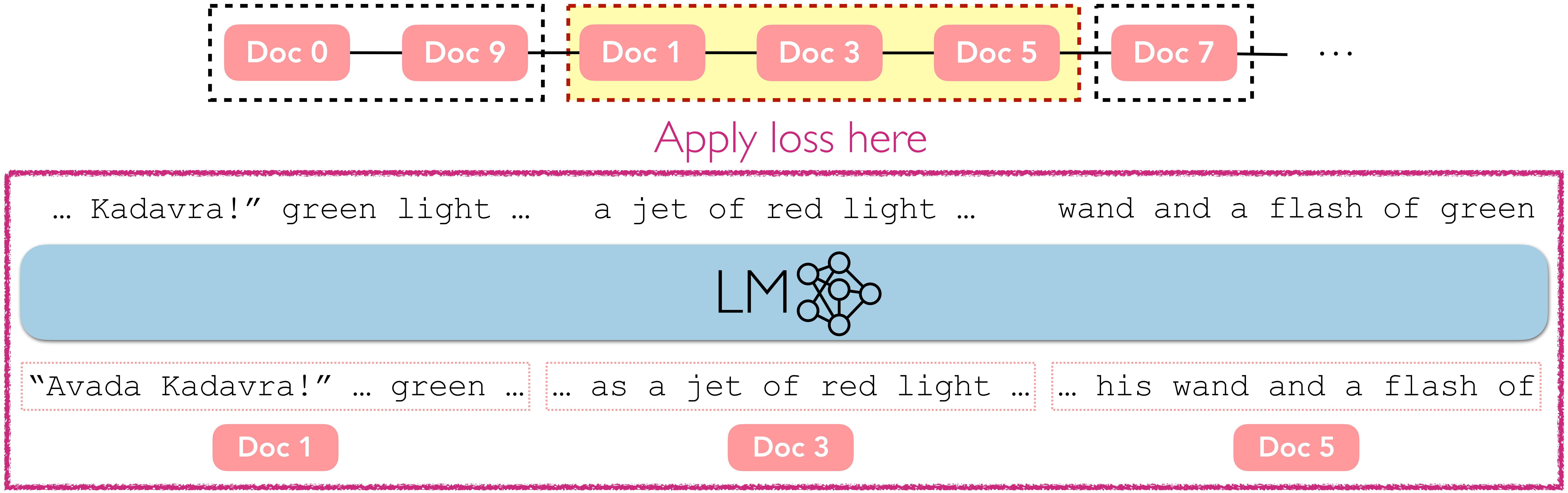
… his wand and a flash of

Doc 5

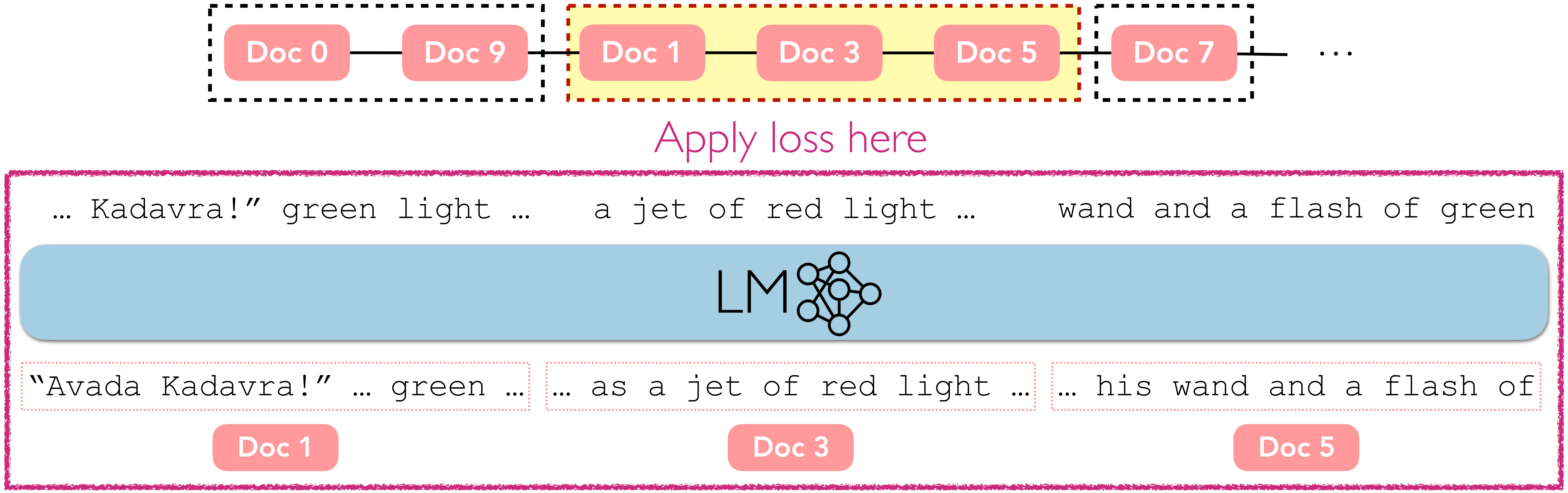
Pre-training w/ retrieval: Proposal



Pre-training w/ retrieval: Proposal



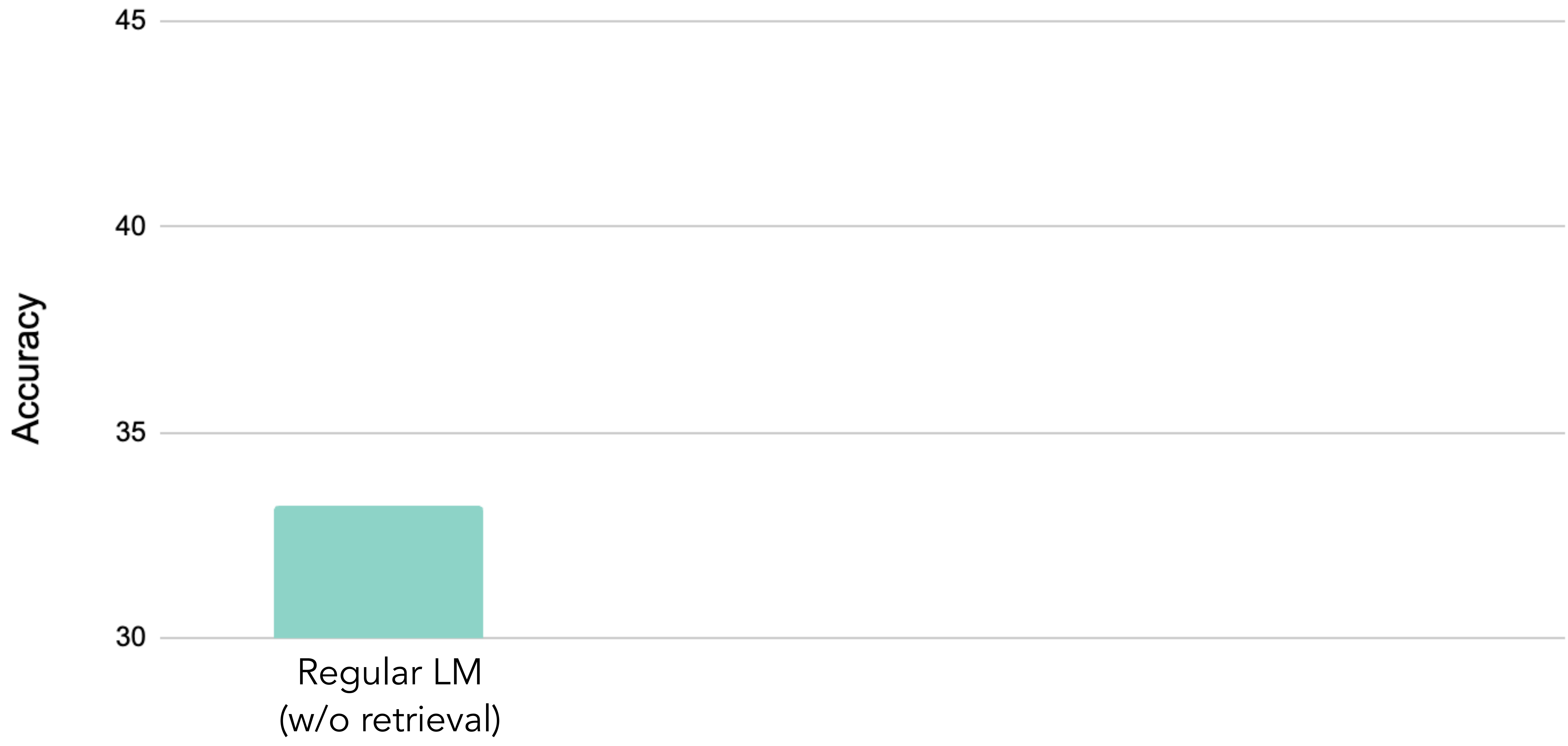
Pre-training w/ retrieval: Proposal



1) LM conditions on a set of ***relevant*** documents

2) Each document appears ***exactly once***

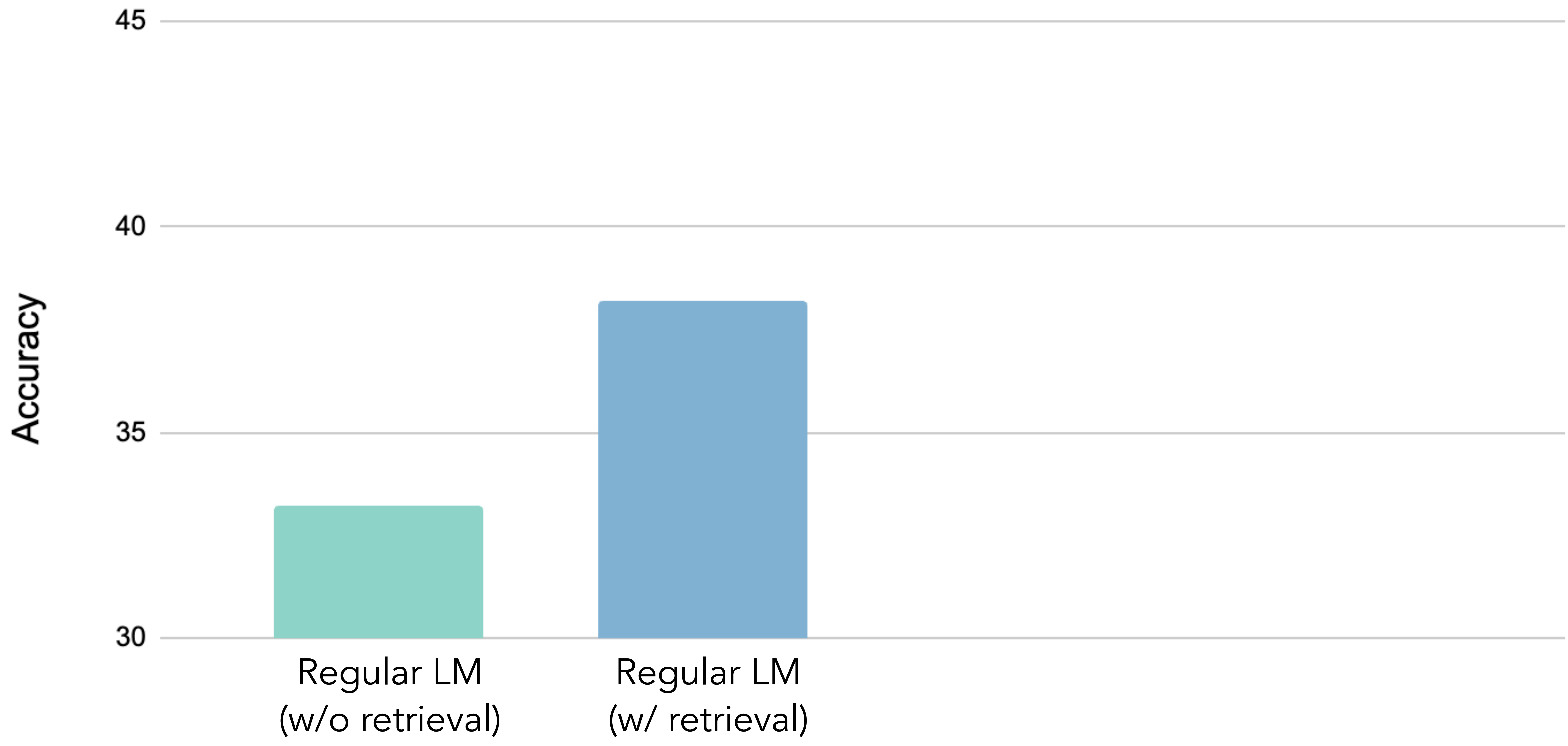
Results: Open-domain QA



Results with a 7B model trained on 300B tokens from Common Crawl

Datasets: NQ, TriviaQA

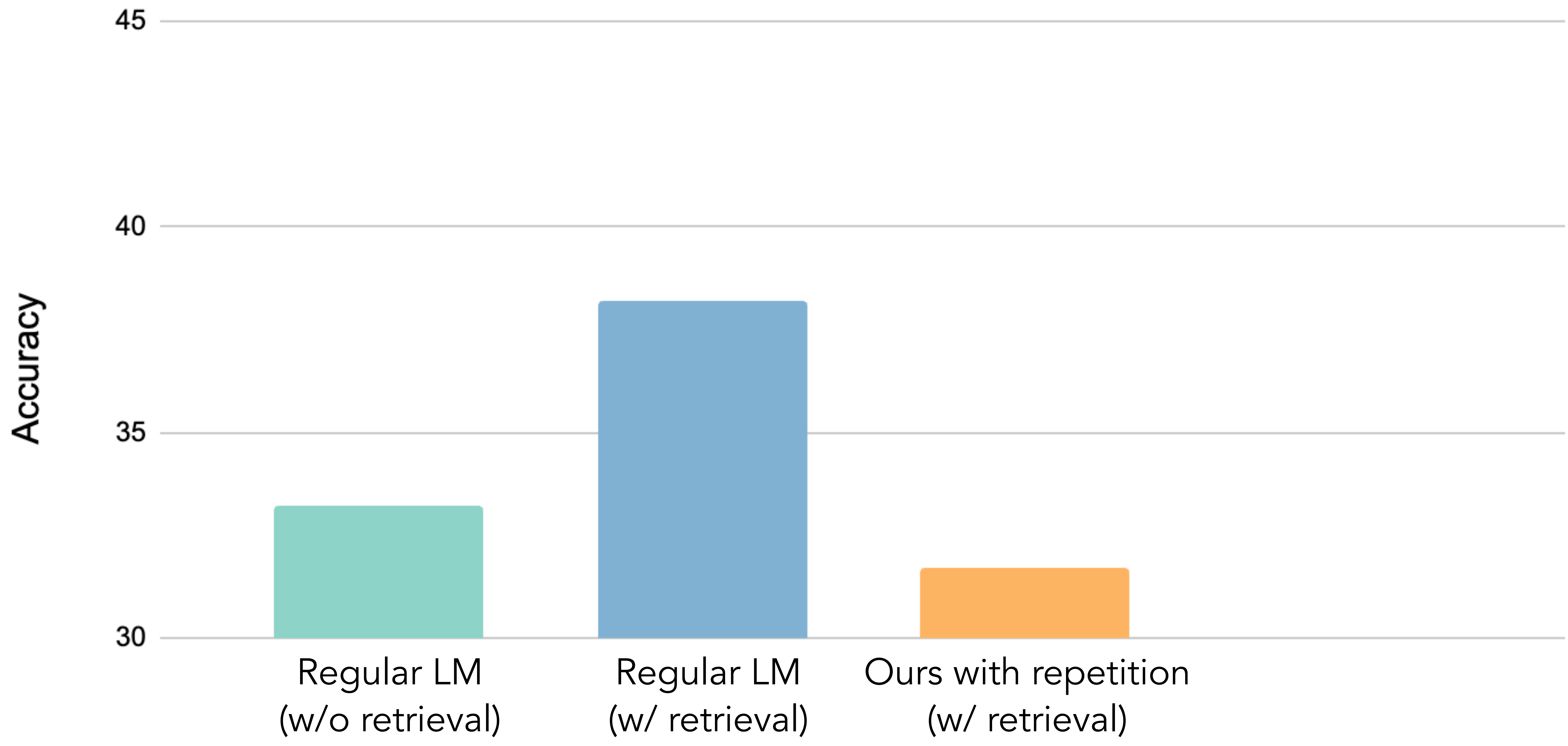
Results: Open-domain QA



Results with a 7B model trained on 300B tokens from Common Crawl

Datasets: NQ, TriviaQA

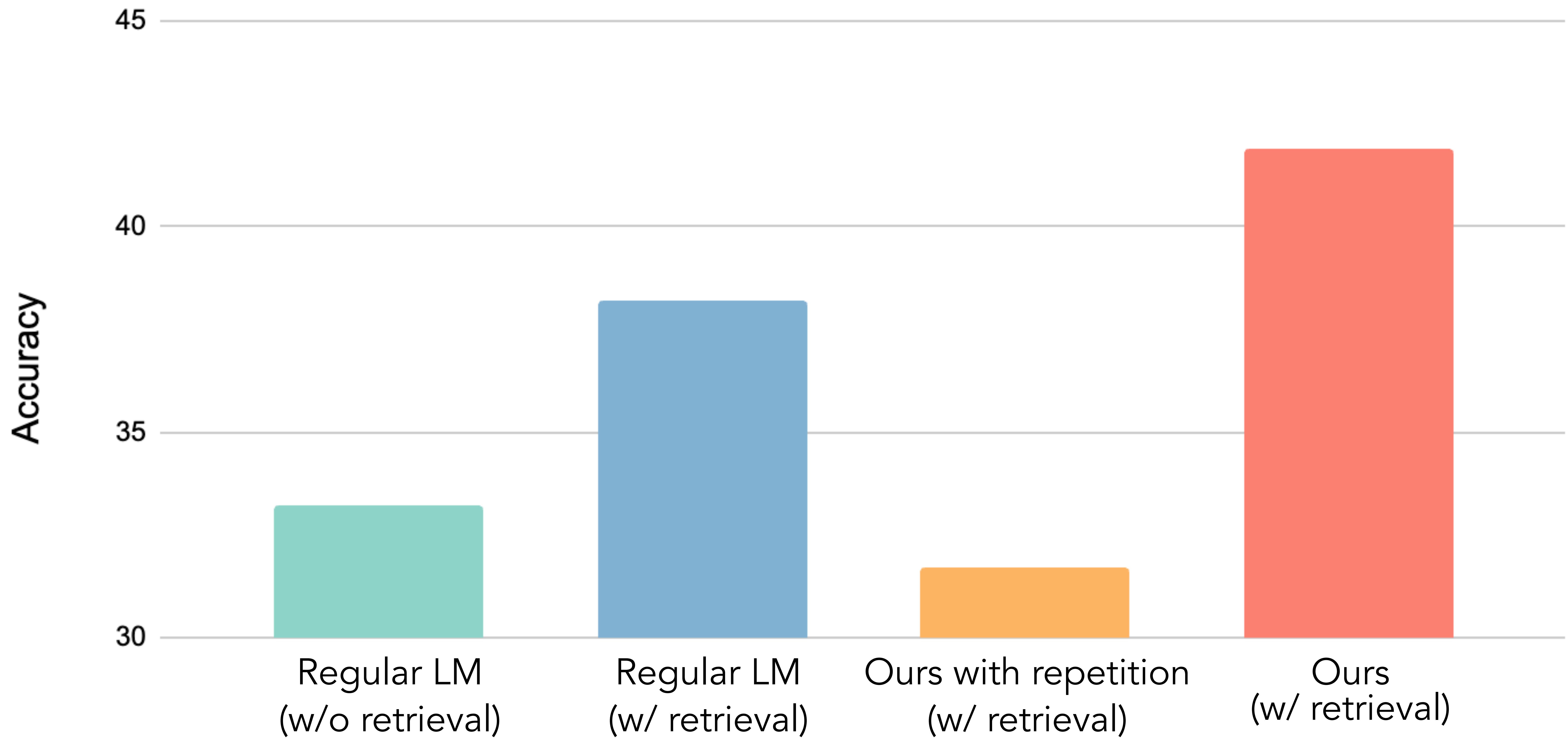
Results: Open-domain QA



Results with a 7B model trained on 300B tokens from Common Crawl

Datasets: NQ, TriviaQA

Results: Open-domain QA



Results with a 7B model trained on 300B tokens from Common Crawl

Datasets: NQ, TriviaQA

Pre-training with retrieval: Summary

Pre-training with retrieval: Summary

- Key idea: Pre-train an LM with retrieval

Pre-training with retrieval: Summary

- Key idea: Pre-train an LM with retrieval
- Naive approaches do not work due to efficiency or duplication issues
  → casted it to a document ordering problem

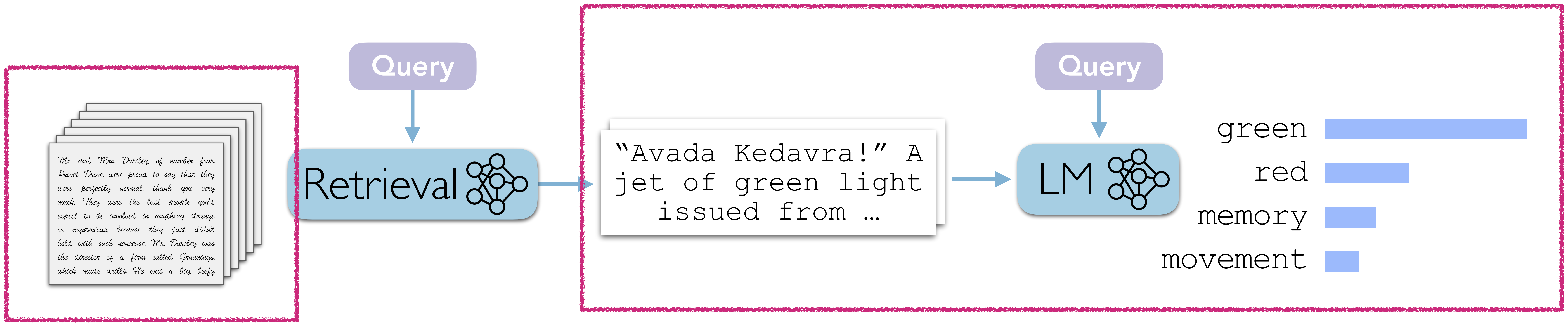
Pre-training with retrieval: Summary

- Key idea: Pre-train an LM with retrieval
- Naive approaches do not work due to efficiency or duplication issues
  → casted it to a document ordering problem
- +10.5% improvements in downstream tasks w/ retrieval on average

Pre-training with retrieval: Summary

- Key idea: Pre-train an LM with retrieval
- Naive approaches do not work due to efficiency or duplication issues
  → casted it to a document ordering problem
- +10.5% improvements in downstream tasks w/ retrieval on average
- (Not in this talk) Larger improvements (+15.9%) when retrieval results do not support answering the query

New scaling of retrieval-based LMs

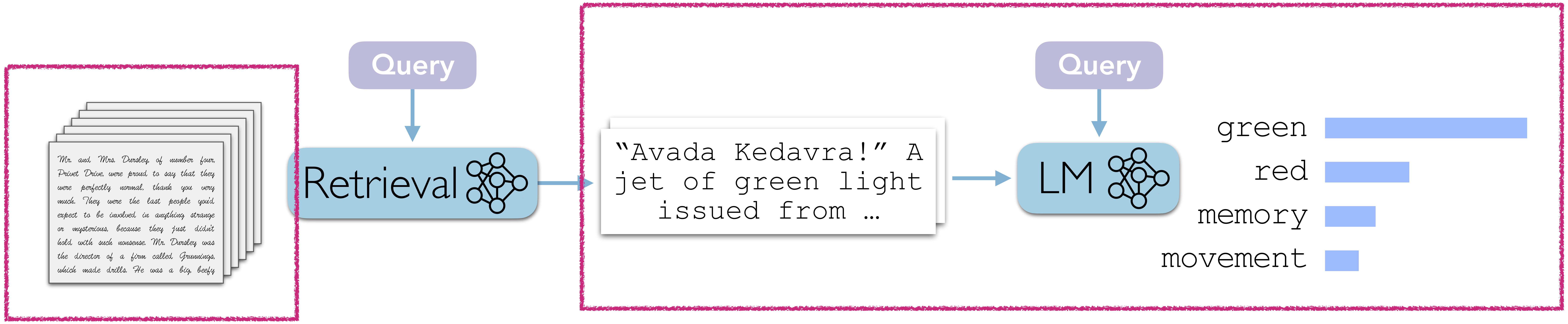


2) Scaling a datastore

1) Scaling training w/ retrieval

3) How to scale with *responsible* data use?

New scaling of retrieval-based LMs



2) Scaling a datastore

1) Scaling training w/ retrieval

3) How to scale with *responsible* data use?

The law of scaling

The law of scaling

LM capabilities = Parameter count x Training data size

The law of scaling

LM capabilities = $\frac{\text{Parameter count}}{}$ x Training data size

$\boxed{\textbf{GPU}}$ $$$$

The law of scaling

LM capabilities = Parameter count x Training data size

$\boxed{\text{GPU}}$ $$$$

The law of scaling

LM capabilities = Parameter count x Training data size x Datastore size

$\boxed{\textbf{GPU}}$ $$$$

The law of scaling

LM capabilities = Parameter count x Training data size x **Datastore size**

$\boxed{\text{GPU}}$ $$$$

The law of scaling

LM capabilities = Parameter count x Training data size x Datastore size

$\boxed{\text{GPU}}$ \$\$\$\$

$\boxed{\text{CPU \& Storage}}$ \$

The law of scaling

LM capabilities = Parameter count x Training data size x **Datastore size**

GPU $$$$ CPU & Storage $

"A small LM + a large datastore >> a large LM?"

Datastores in the literature

Reference	# tokens	Data source	Open sourced?
DPR (Karpukhin et al. 2020)	<5B	Wikipedia	O
ATLAS (Izcard et al. 2023)	<5B	Wikipedia	X
REALM (Gun et al. 2020)	<5B	Wikipedia	X
RALM (RAm et al. 2023)	<5B	Wikipedia	O
Self-RAG (Asai et al. 2024)	<5B	Wikipedia	O

Datastores in the literature

Reference	# tokens	Data source	Open sourced?
DPR (Karpukhin et al. 2020)	<5B	Wikipedia	O
ATLAS (Izcard et al. 2023)	<5B	Wikipedia	X
REALM (Gun et al. 2020)	<5B	Wikipedia	X
RALM (RAm et al. 2023)	<5B	Wikipedia	O
Self-RAG (Asai et al. 2024)	<5B	Wikipedia	O
REPLUG (Shi et al. 2023)	47B	The Pile	O
RA-DIT (Lin et al. 2024)	79B	Wikipedia + Common Crawl	x
SPHERE (Piktus et al. 2022)	90B	CC Net	O

Datastores in the literature

Reference	# tokens	Data source	Open sourced?
DPR (Karpukhin et al. 2020)	<5B	Wikipedia	O
ATLAS (Izcard et al. 2023)	<5B	Wikipedia	X
REALM (Gun et al. 2020)	<5B	Wikipedia	X
RALM (RAm et al. 2023)	<5B	Wikipedia	O
Self-RAG (Asai et al. 2024)	<5B	Wikipedia	O
REPLUG (Shi et al. 2023)	47B	The Pile	O
RA-DIT (Lin et al. 2024)	79B	Wikipedia + Common Crawl	x
SPHERE (Piktus et al. 2022)	90B	CC Net	O
RETRO (Borgeaud et al. 2022)	1.7T	MassiveText	X

Datastores in the literature

Reference	# tokens	Data source	Open sourced?
DPR (Karpukhin et al. 2020)	<5B	Wikipedia	O
ATLAS (Izcard et al. 2023)	<5B	Wikipedia	X
REALM (Gun et al. 2020)	<5B	Wikipedia	X
RALM (RAm et al. 2023)	<5B	Wikipedia	O
Self-RAG (Asai et al. 2024)	<5B	Wikipedia	O
REPLUG (Shi et al. 2023)	47B	The Pile	O
RA-DIT (Lin et al. 2024)	79B	Wikipedia + Common Crawl	x
SPHERE (Piktus et al. 2022)	90B	CC Net	O
RETRO (Borgeaud et al. 2022)	1.7T	MassiveText	X
MassiveDS (Shao et al. 2024)	**1.4T**	Common Crawl (83%), Academic papers, Code, Encyclopedia, Books, Math, Biomedical, etc	**O**

Datastores in the literature

Reference	# tokens	Data source	Open sourced?
DPR (Karpukhin et al. 2020)	<5B	Wikipedia	O
ATLAS (Izcard et al. 2023)	<5B	Wikipedia	X
REALM (Gun et al. 2020)	<5B	Wikipedia	X
RALM (RAm et al. 2023)	<5B	Wikipedia	O
Self-RAG (Asai et al. 2024)	<5B	Wikipedia	O
REPLUG (Shi et al. 2023)	47B	The Pile	O
RA-DIT (Lin et al. 2024)	79B	Wikipedia + Common Crawl	x
SPHERE (Piktus et al. 2022)	90B	CC Net	O
RETRO (Borgeaud et al. 2022)	1.7T	MassiveText	X
MassiveDS (Shao et al. 2024)	**1.4T**	Common Crawl (83%), Academic papers, Code, Encyclopedia, Books, Math, Biomedical, etc	**O**

(Lots of challenges in systems and algorithms — skipping here)

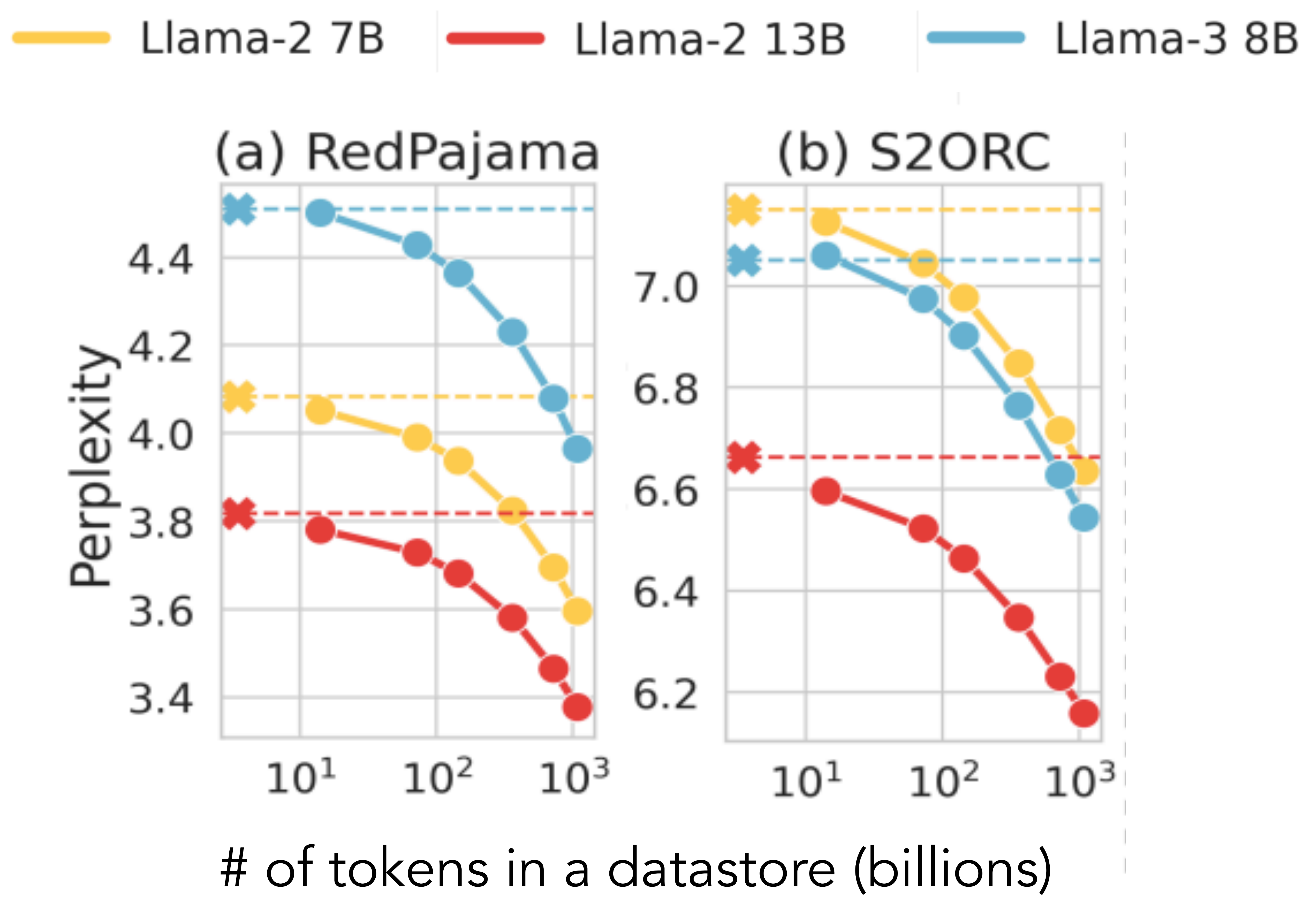
Results: Perplexity

Results: Perplexity

Llama-2 7B   Llama-2 13B   Llama-3 8B

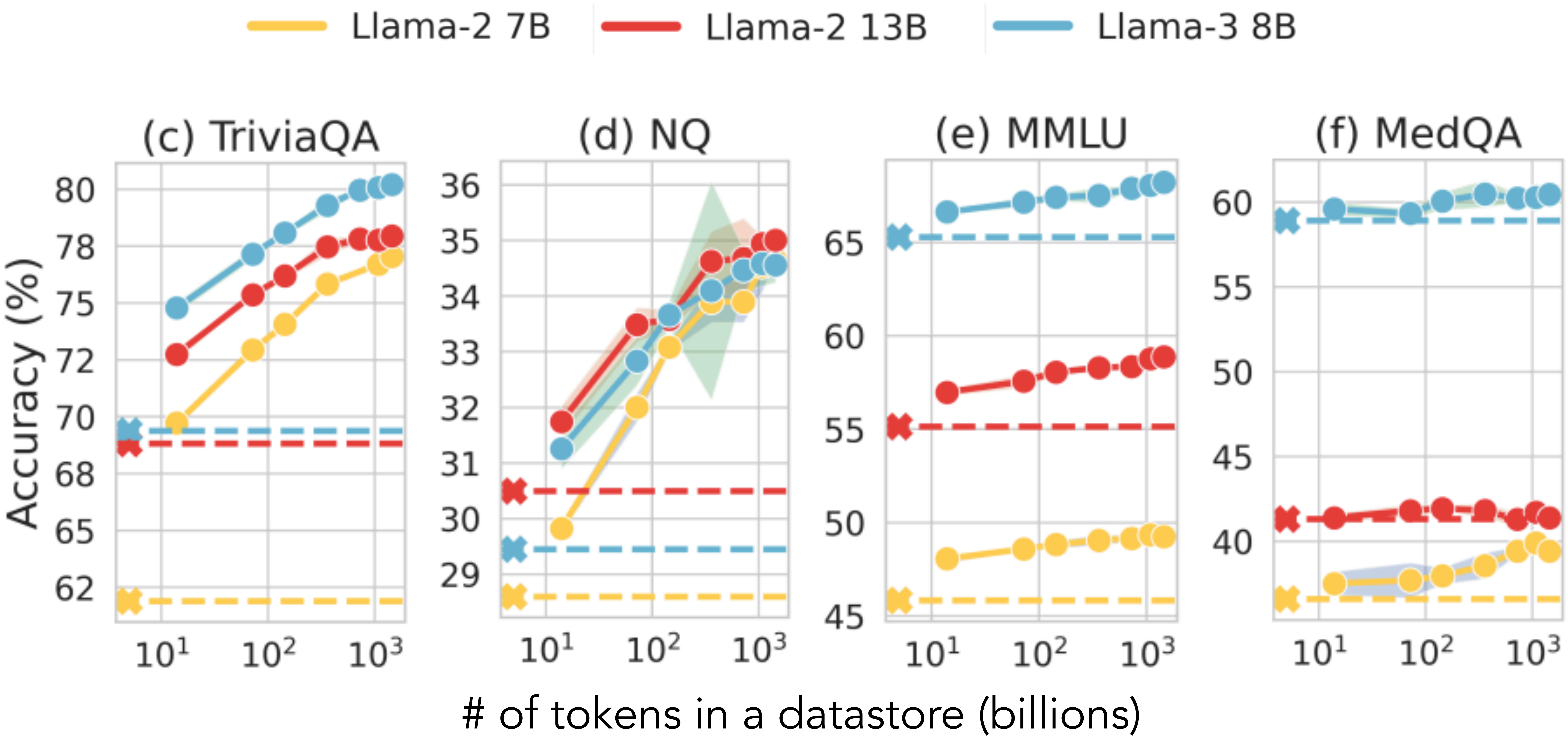
(More training compute)

Results: Perplexity

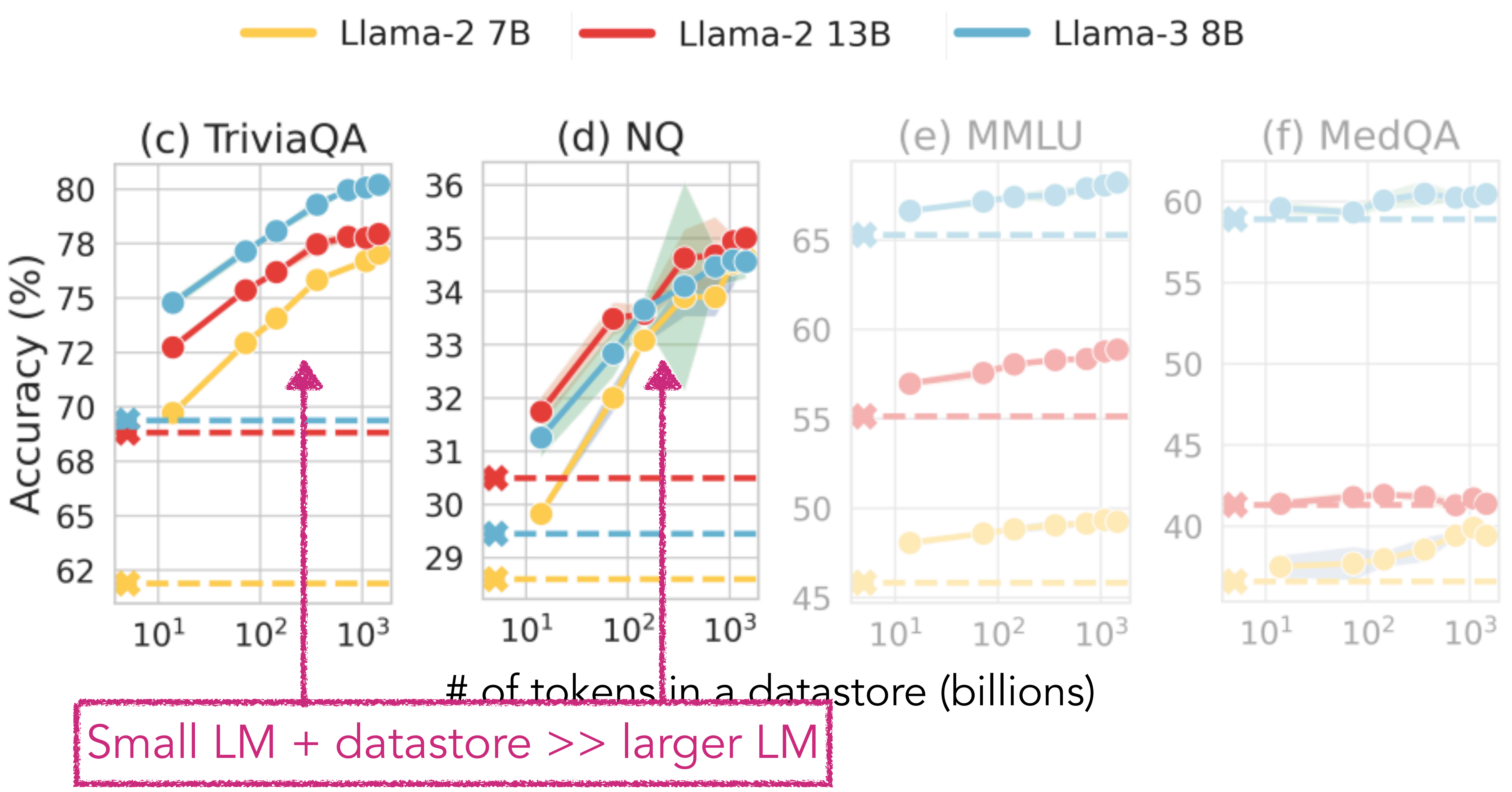


(Note: Llama-2 and Lllama-3 are not comparable in PPL!)

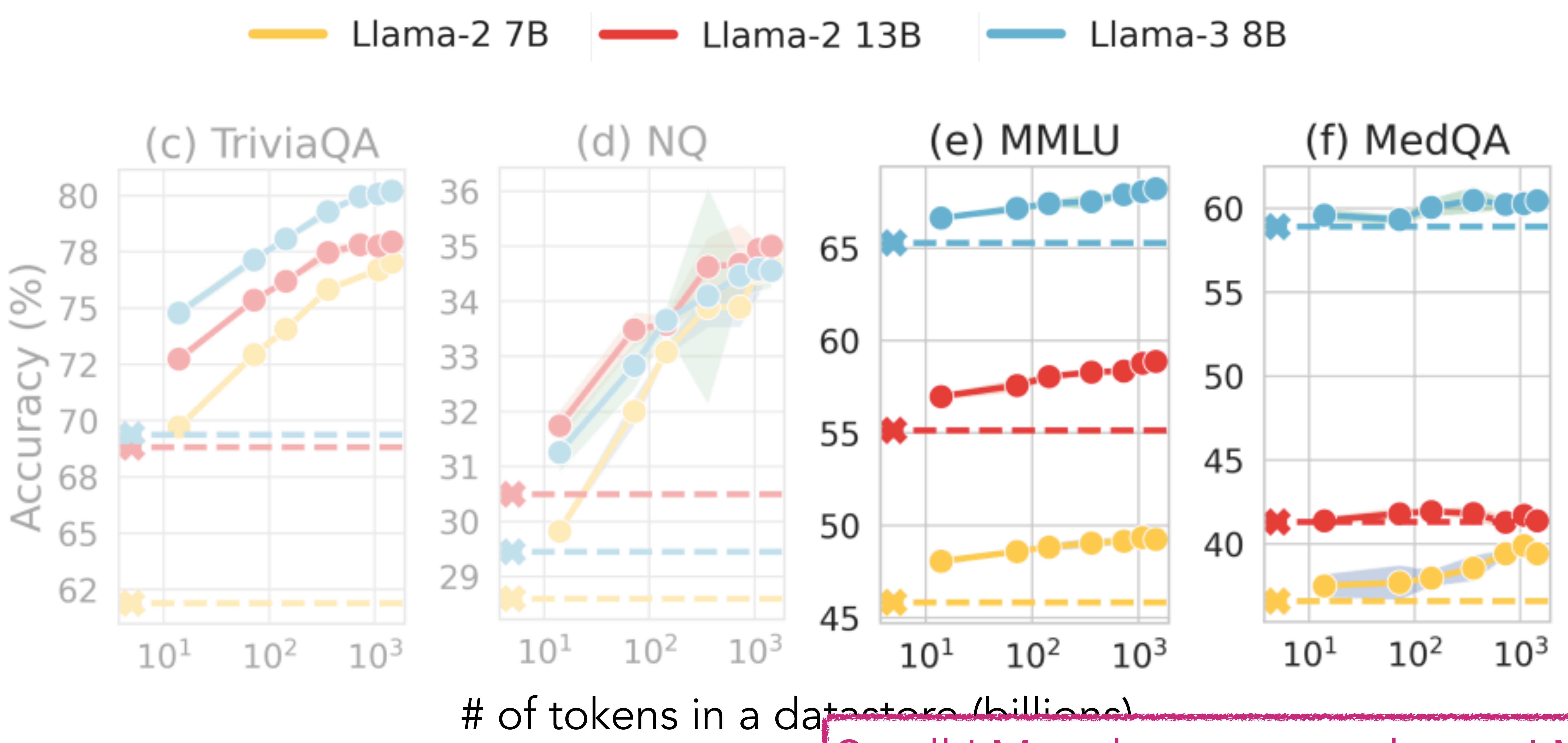
Results: Downstream tasks



Results: Downstream tasks



Results: Downstream tasks



Small LM + datastore << larger LM

Compute-Optimal Scaling

Compute-Optimal Scaling

Where to use compute — for # parameters, training data, or datastore?

Compute-Optimal Scaling

Where to use compute — for # parameters, training data, or datastore?

1B and 7B

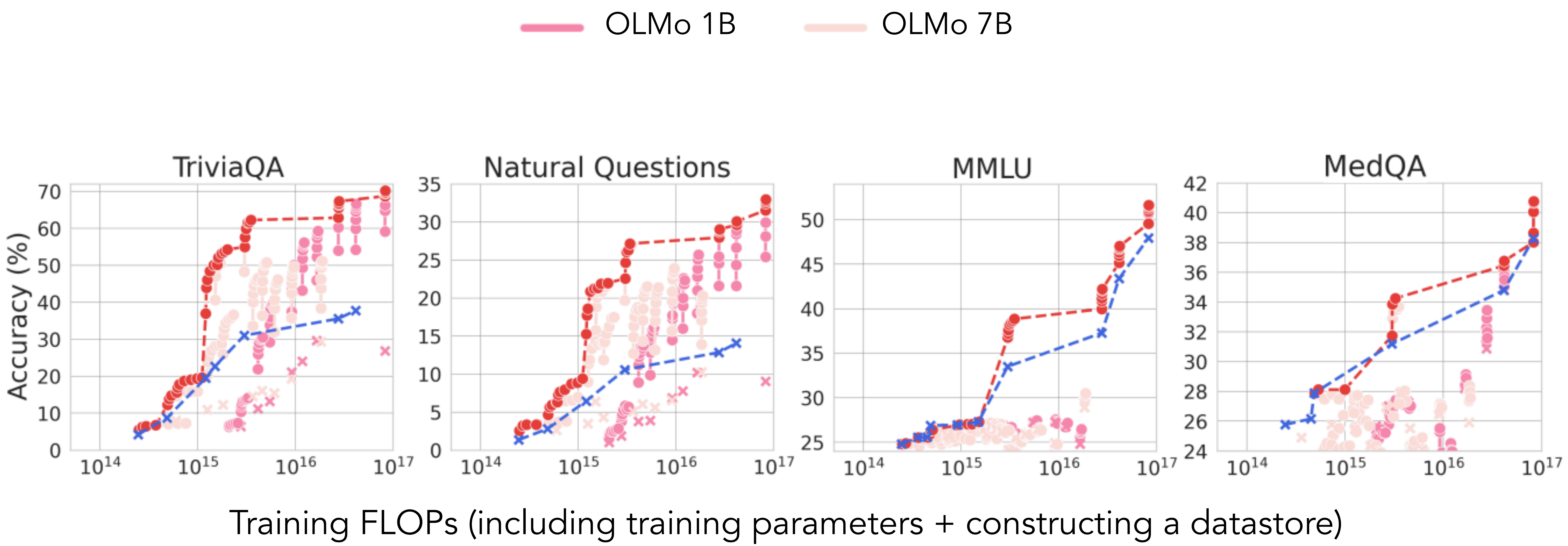
Checkpoints trained on varying sizes of data

Datastores with varying sizes

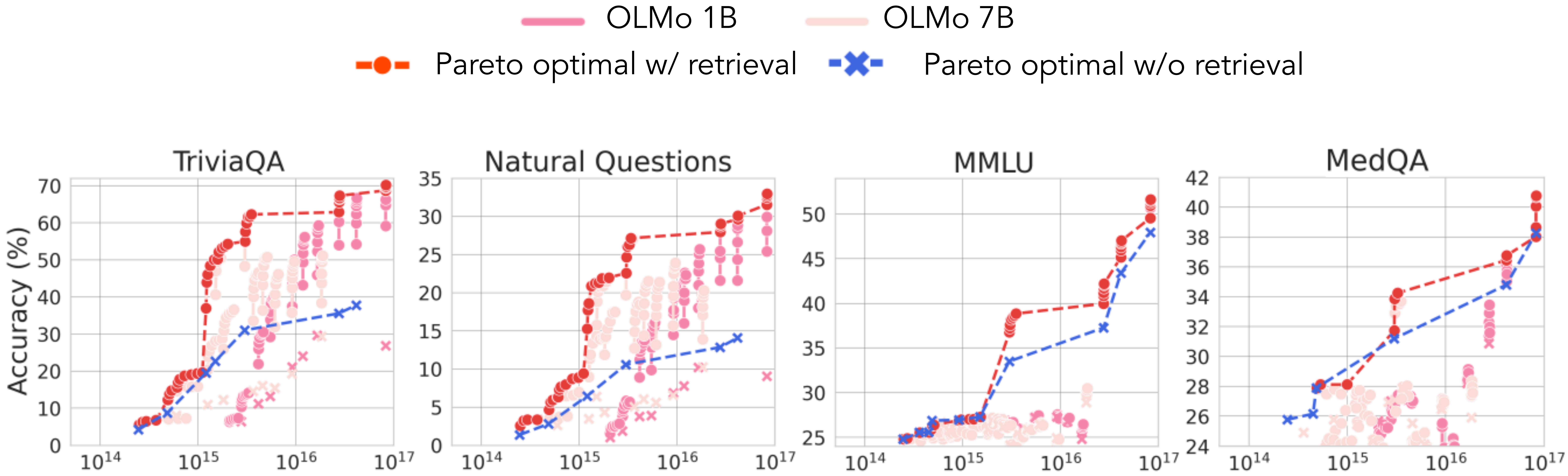
Compute-Optimal Scaling

OLMo 1B   OLMo 7B

Compute-Optimal Scaling

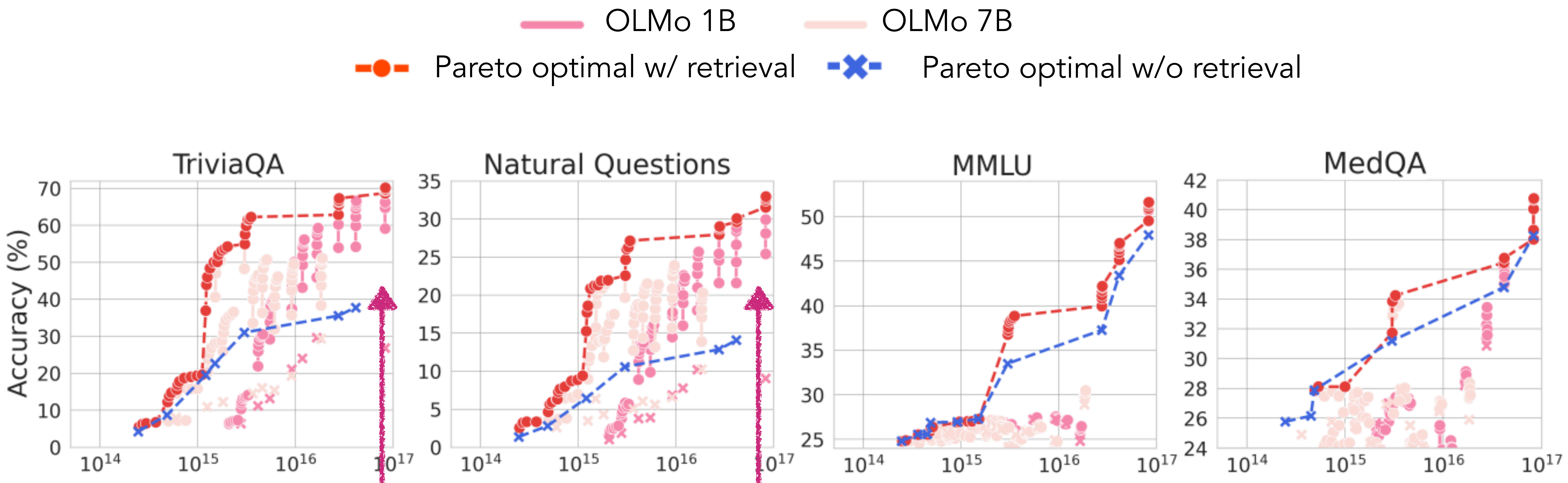


Compute-Optimal Scaling



Training FLOPs (including training parameters + constructing a datastore)

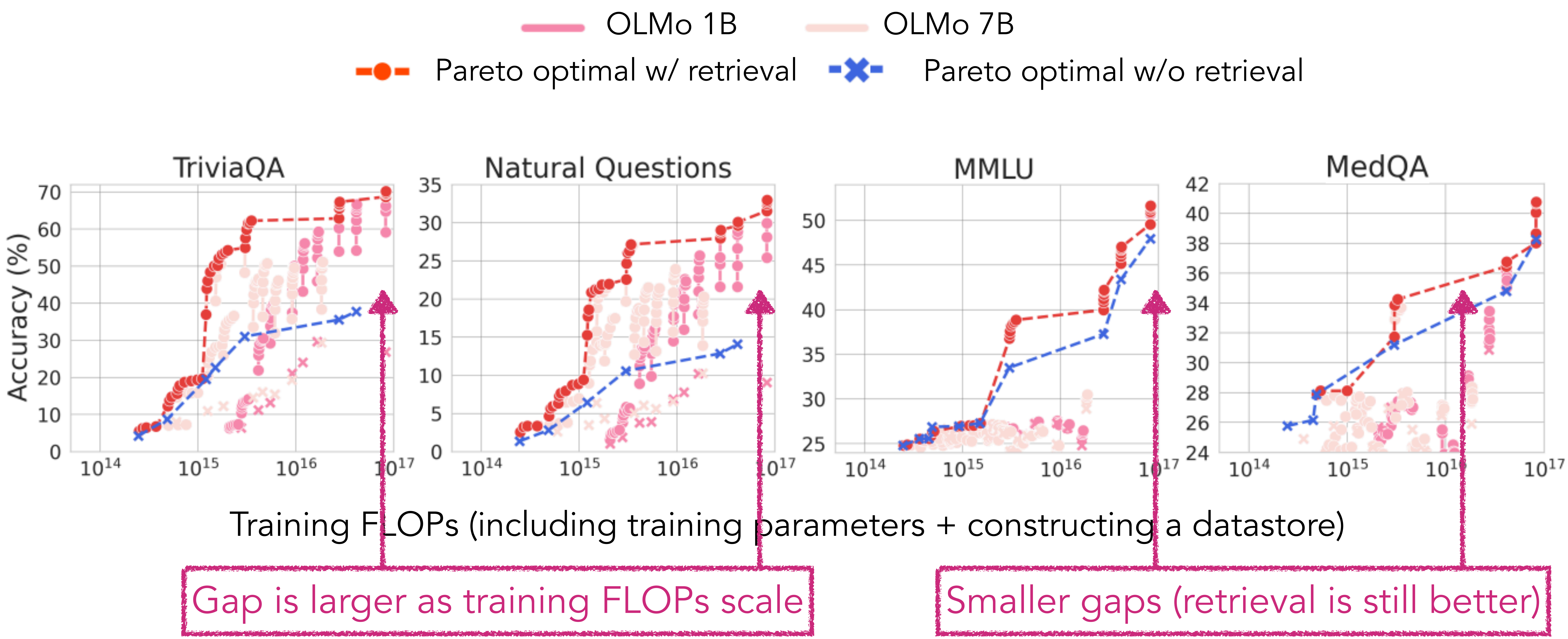
Compute-Optimal Scaling



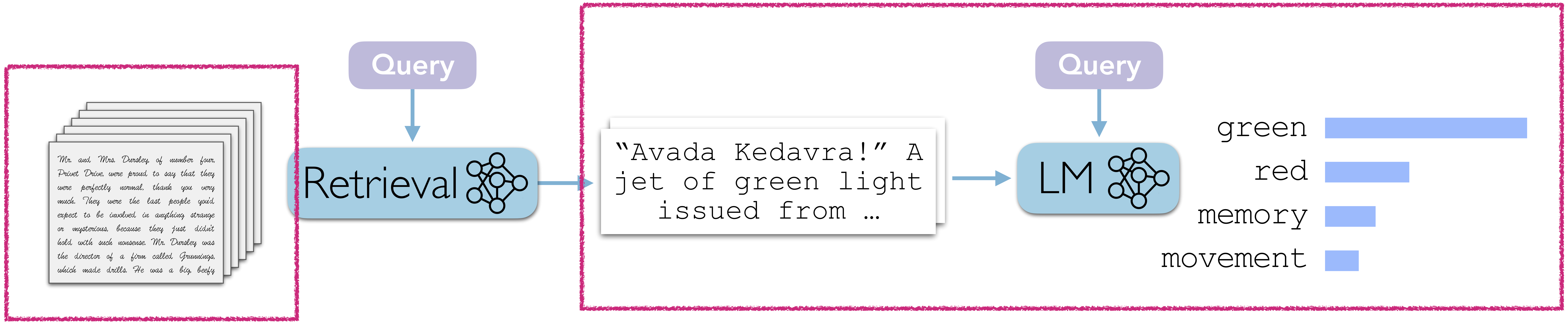
Training FLOPs (including training parameters + constructing a datastore)

Gap is larger as training FLOPs scale

Compute-Optimal Scaling



New scaling of retrieval-based LMs

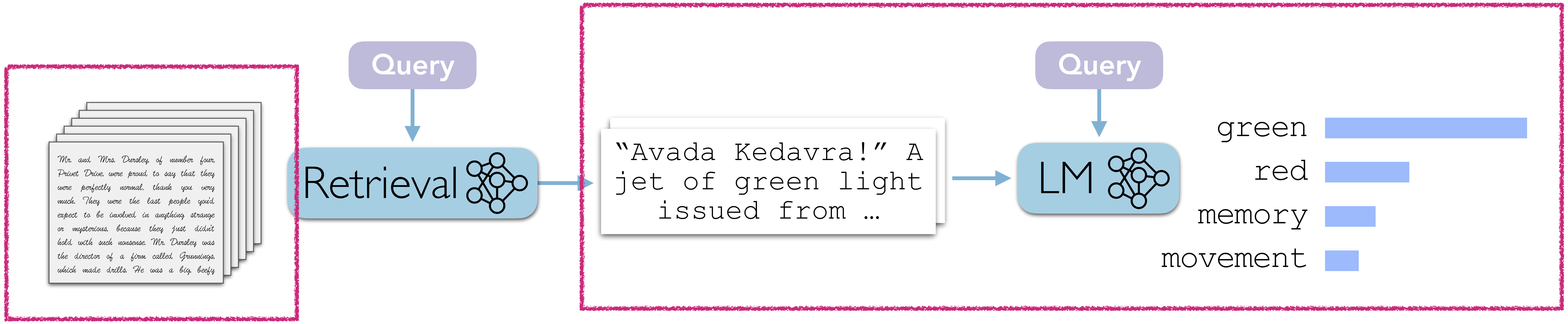


2) Scaling a datastore

1) Scaling training w/ retrieval

3) How to scale with *responsible* data use?

New scaling of retrieval-based LMs



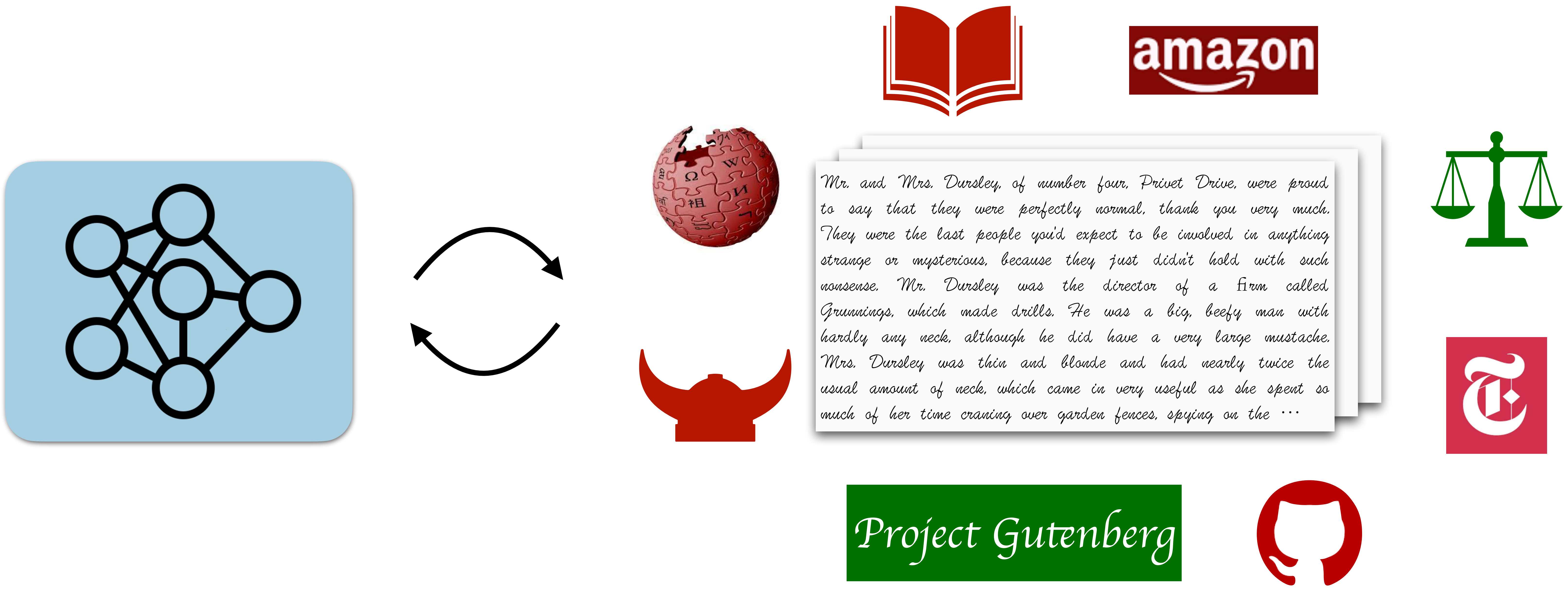
2) Scaling a datastore

1) Scaling training w/ retrieval

3) How to scale with *responsible* data use?

Current practice

Current practice

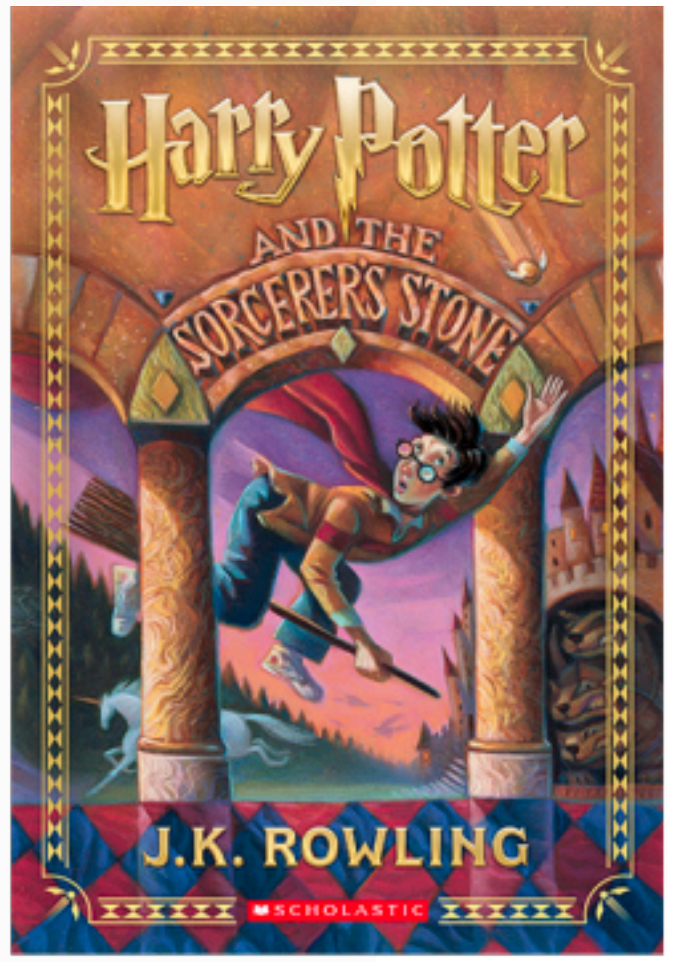


Playground

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

The Dursleys had everything they wanted,



Looking for ChatGPT? Try it now

Submit

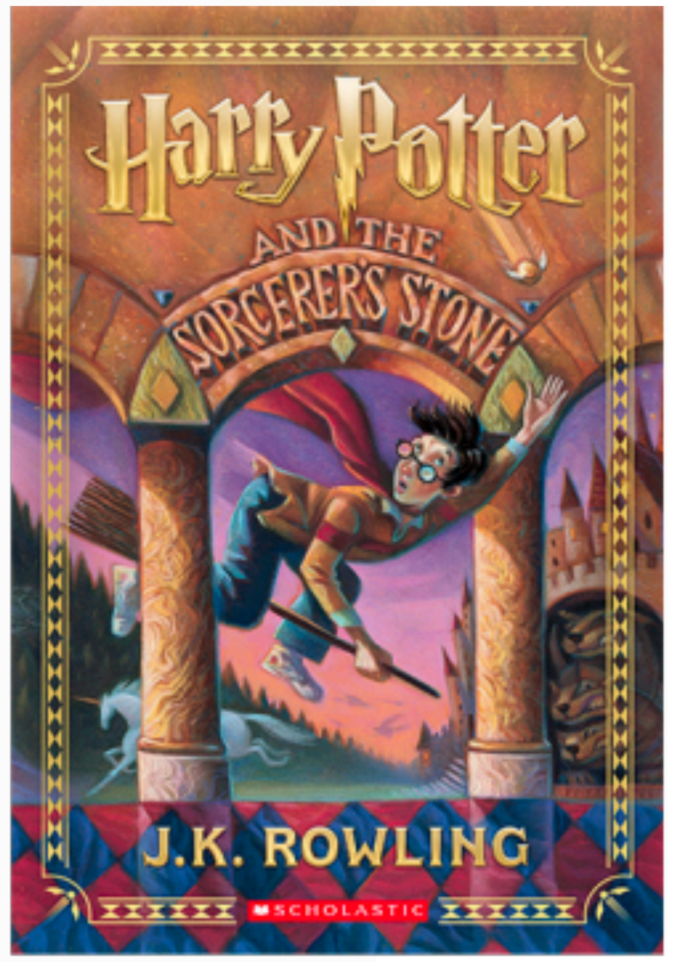
Overview Documentation API reference Examples Playground

Playground

Mr and Mrs Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large moustache. Mrs Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbours. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

The Dursleys had everything they wanted,



Looking for ChatGPT? Try it now

Submit

Problem: Risk

I want my books to be excluded.

Problem: Risk

I want my books to be excluded.

I want to get credited whenever the model uses my articles.

Problem: Risk

I want my books to be excluded.

I want to get credited whenever the model uses my articles.

Problem: Risk

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

Problem: Risk

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Problem: Risk

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

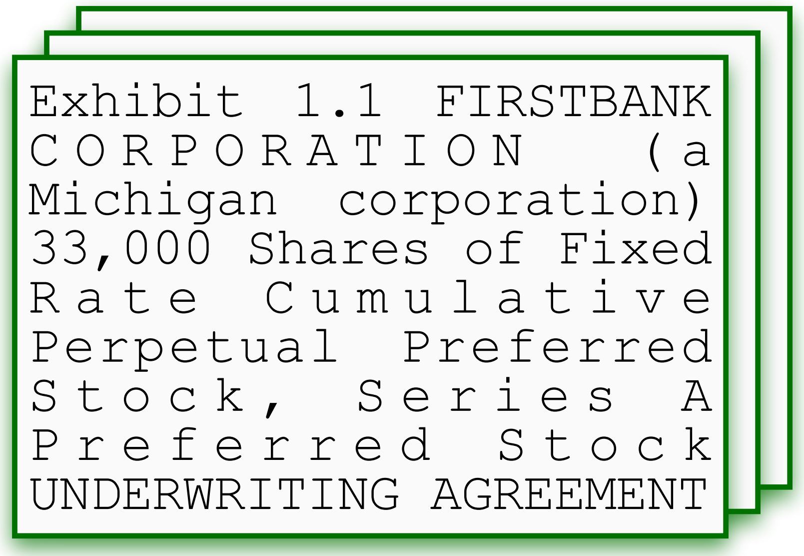
I got a lawsuit for violating DMCA (for removing CMI).

Copilot

- Re-train the model every time removal of data is needed → expensive
- Filter out any risky data and train on permissive data only → impractical

Proposal w. retrieval-based LMs

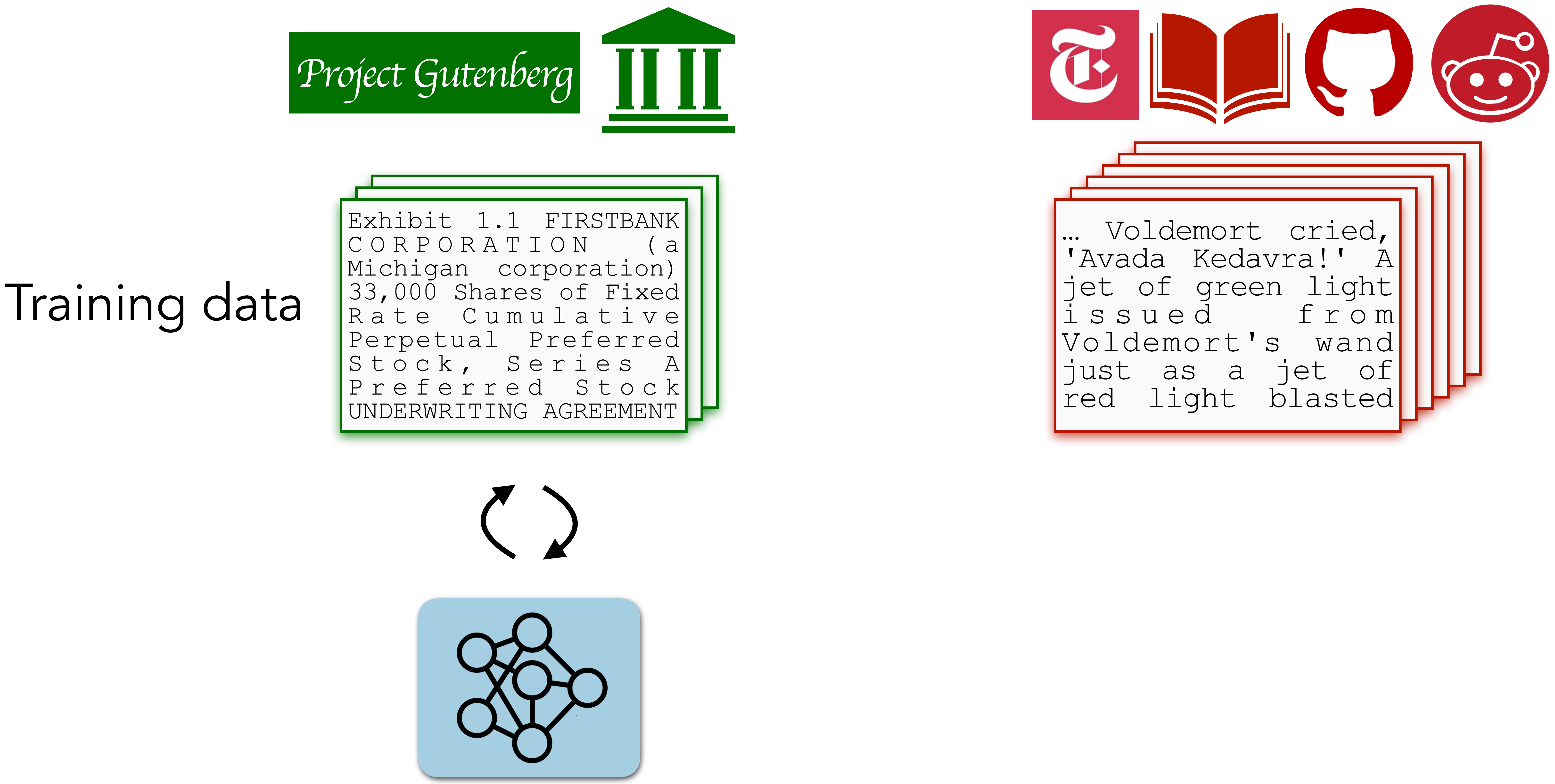
Proposal w. retrieval-based LMs



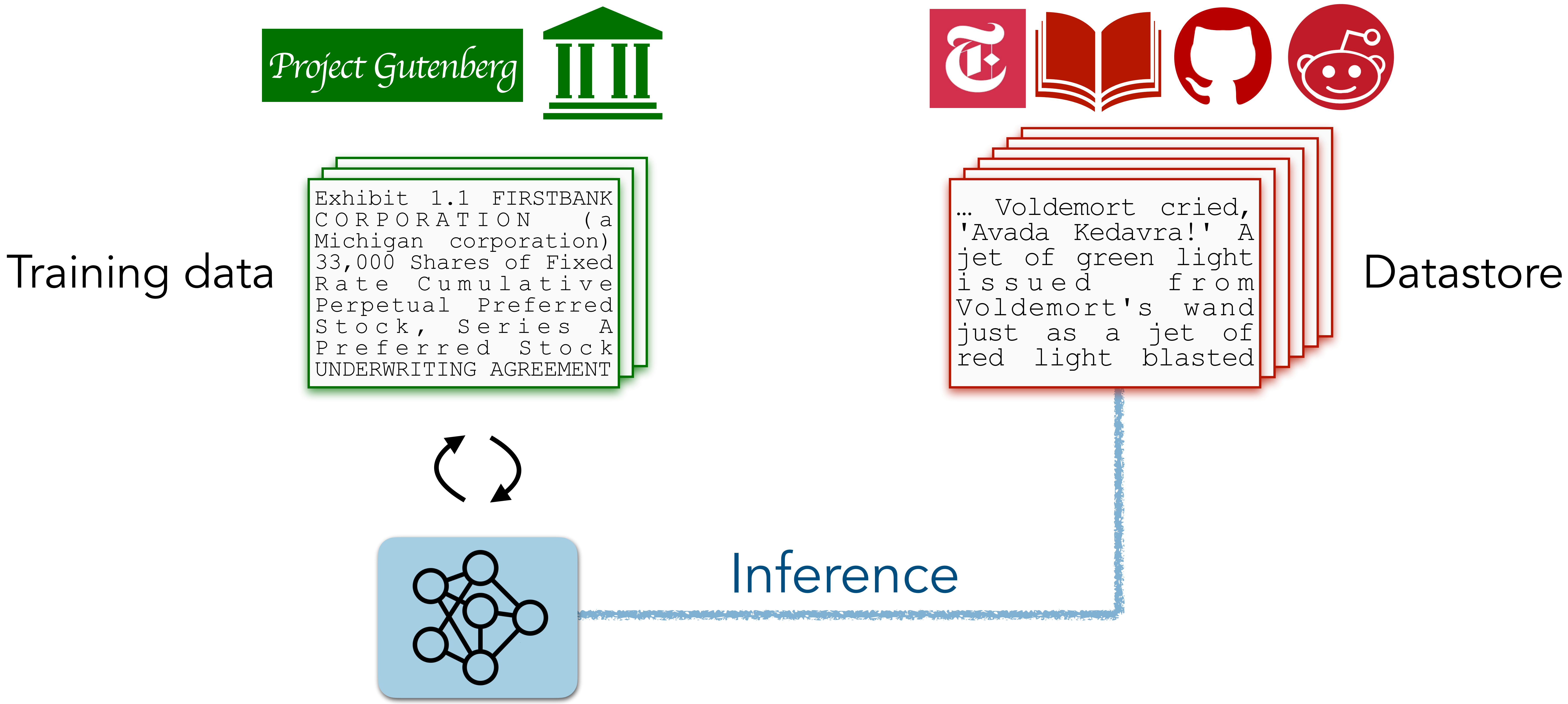
Proposal w. retrieval-based LMs

Proposal w. retrieval-based LMs

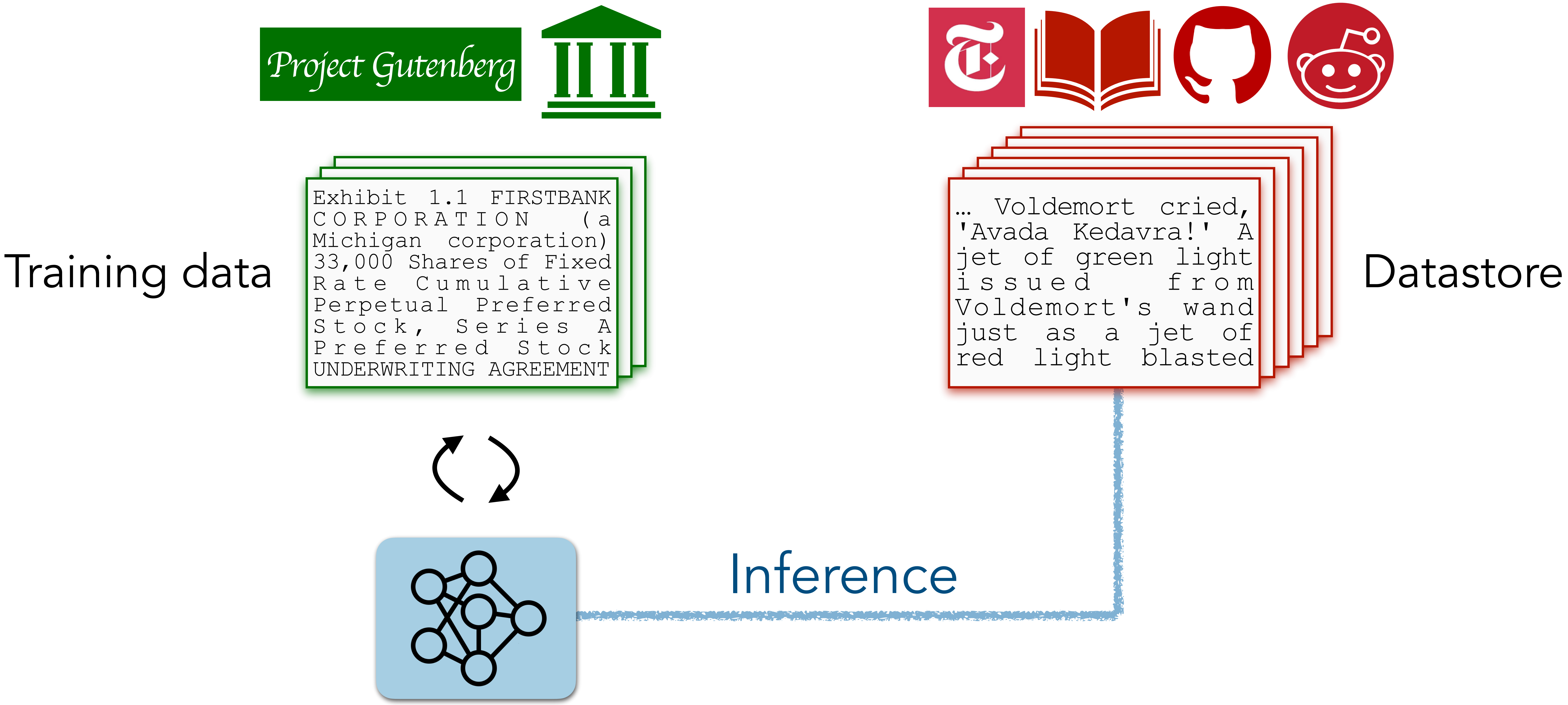
Proposal w. retrieval-based LMs



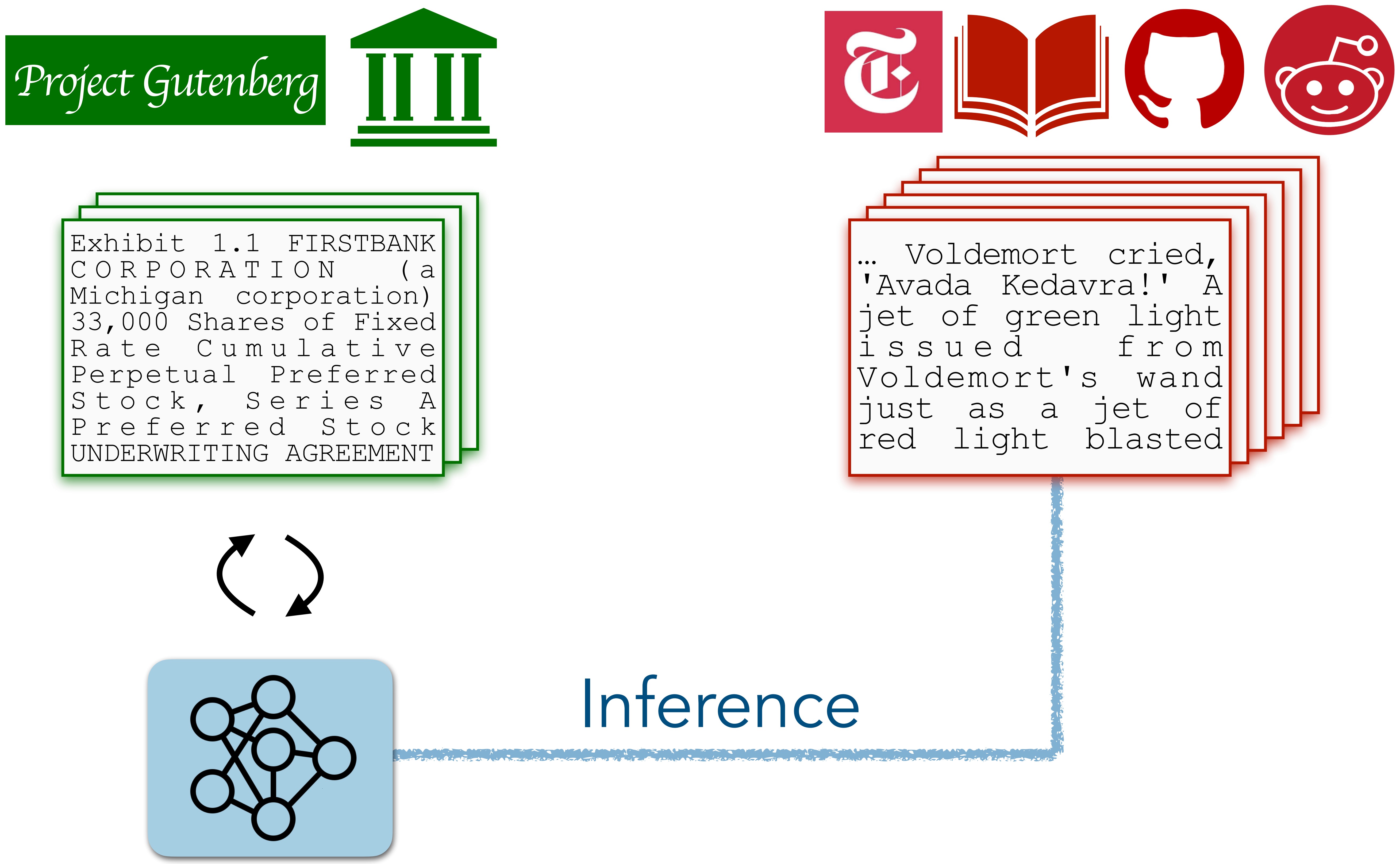
Proposal w. retrieval-based LMs



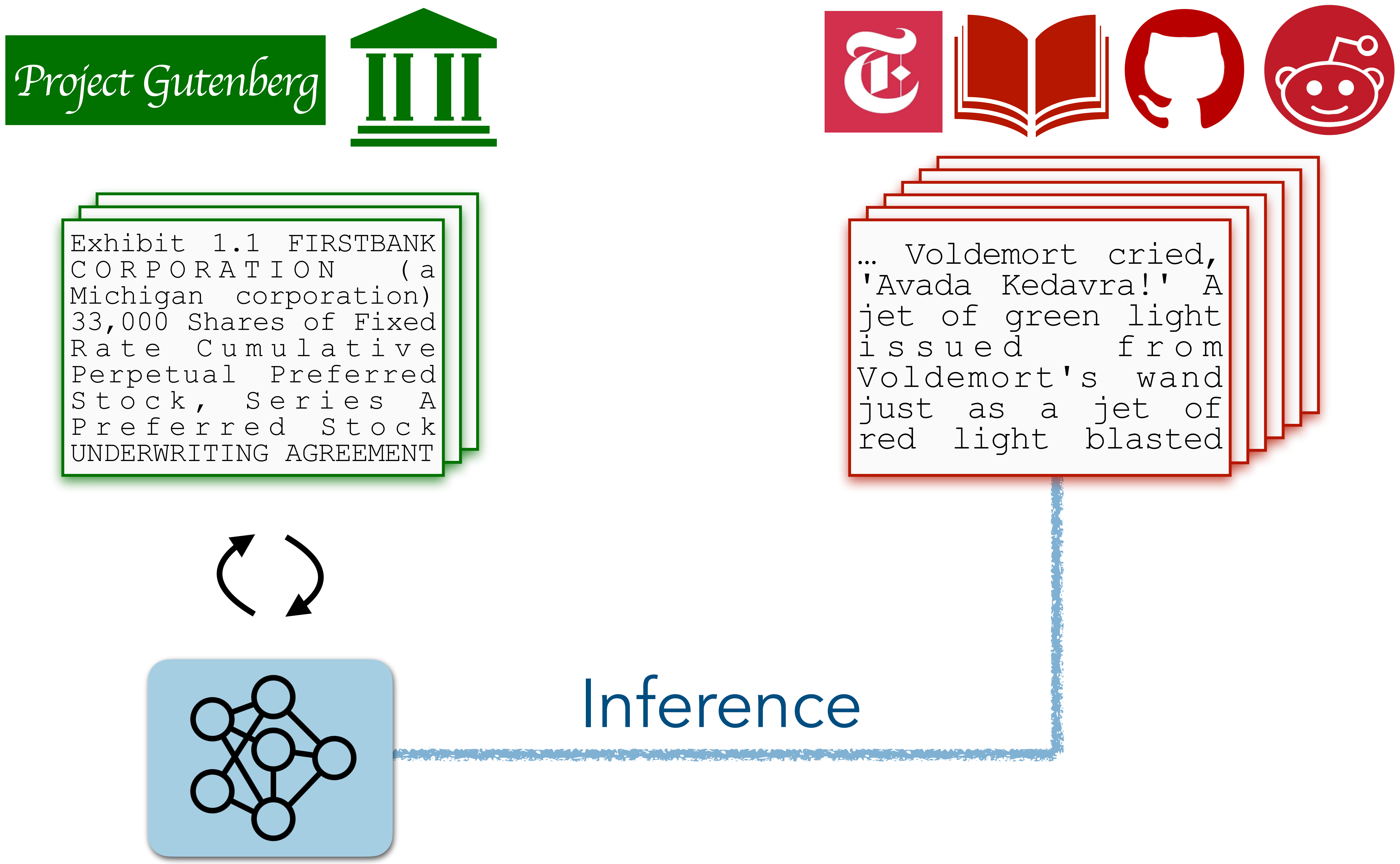
Proposal w. retrieval-based LMs



Proposal w. retrieval-based LMs

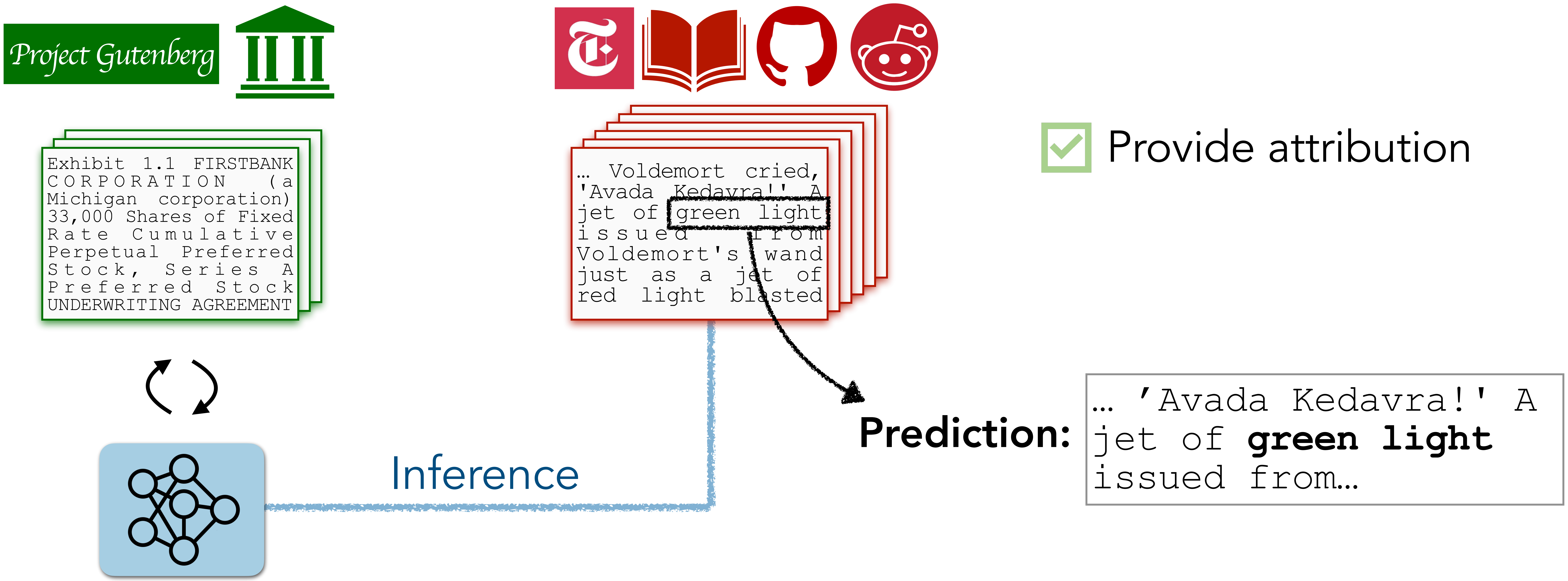


Proposal w. retrieval-based LMs

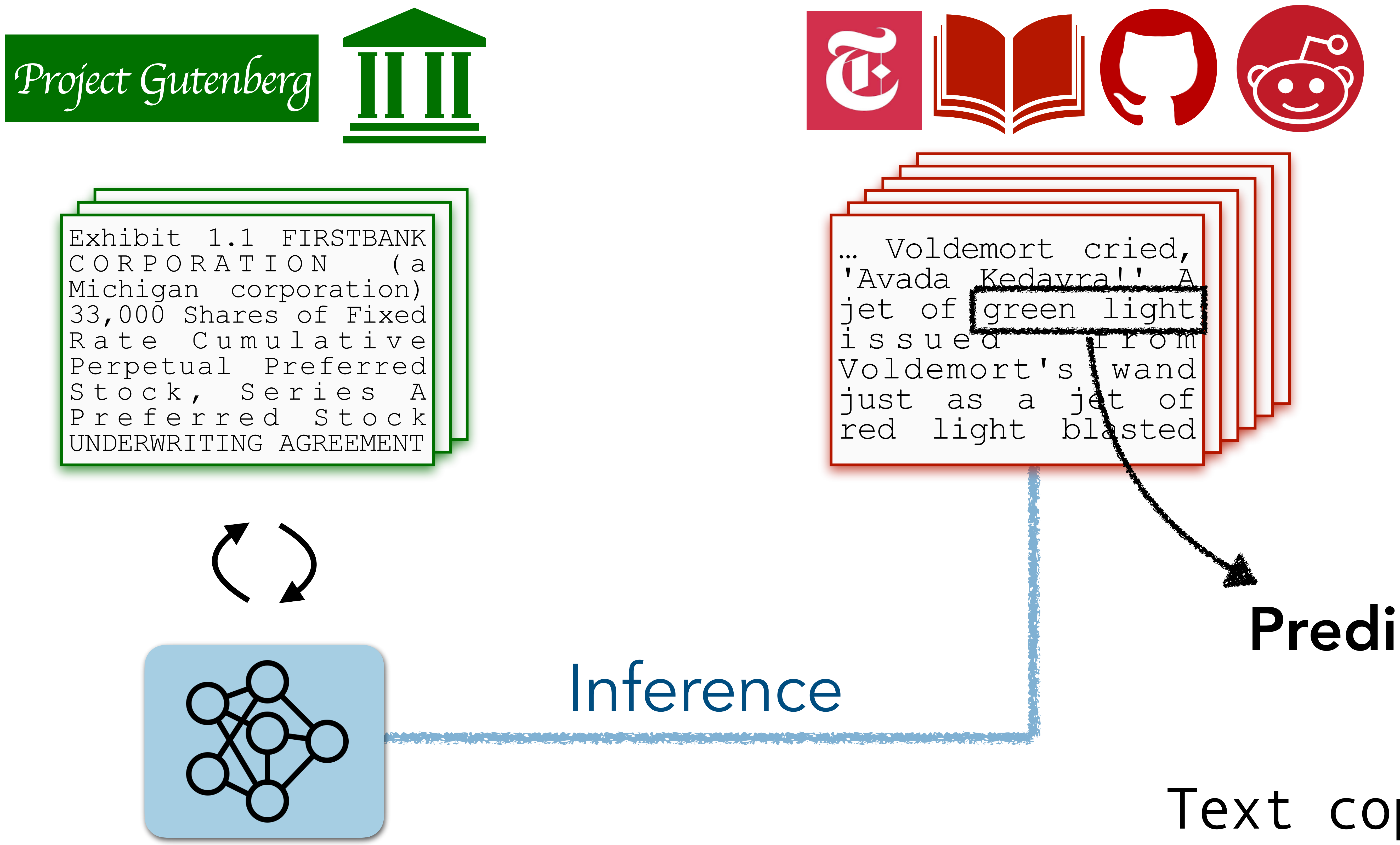


☑ Provide attribution

Proposal w. retrieval-based LMs



Proposal w. retrieval-based LMs

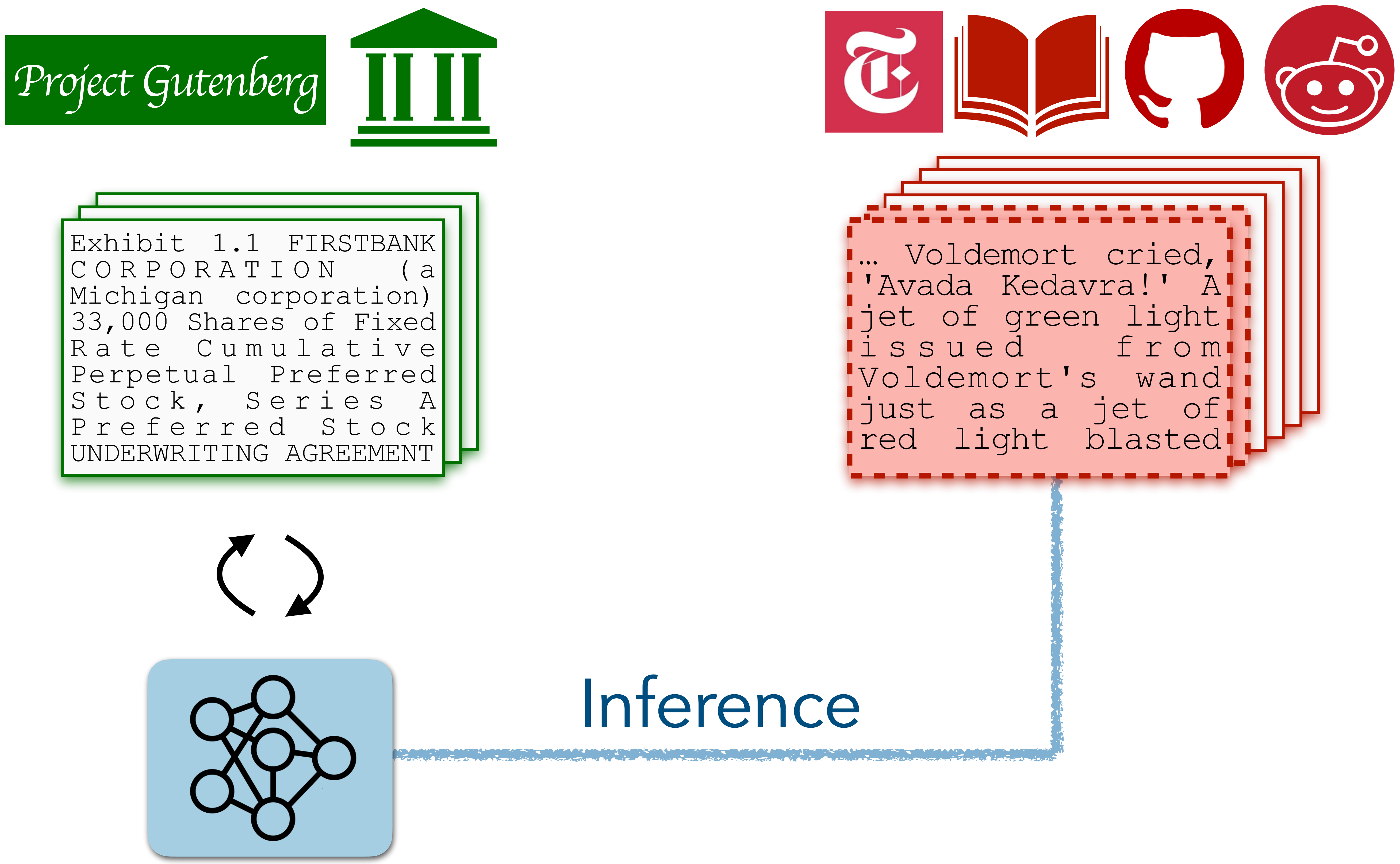


✅ Provide attribution

Prediction: … 'Avada Kedavra!' A jet of **green light** issued from…

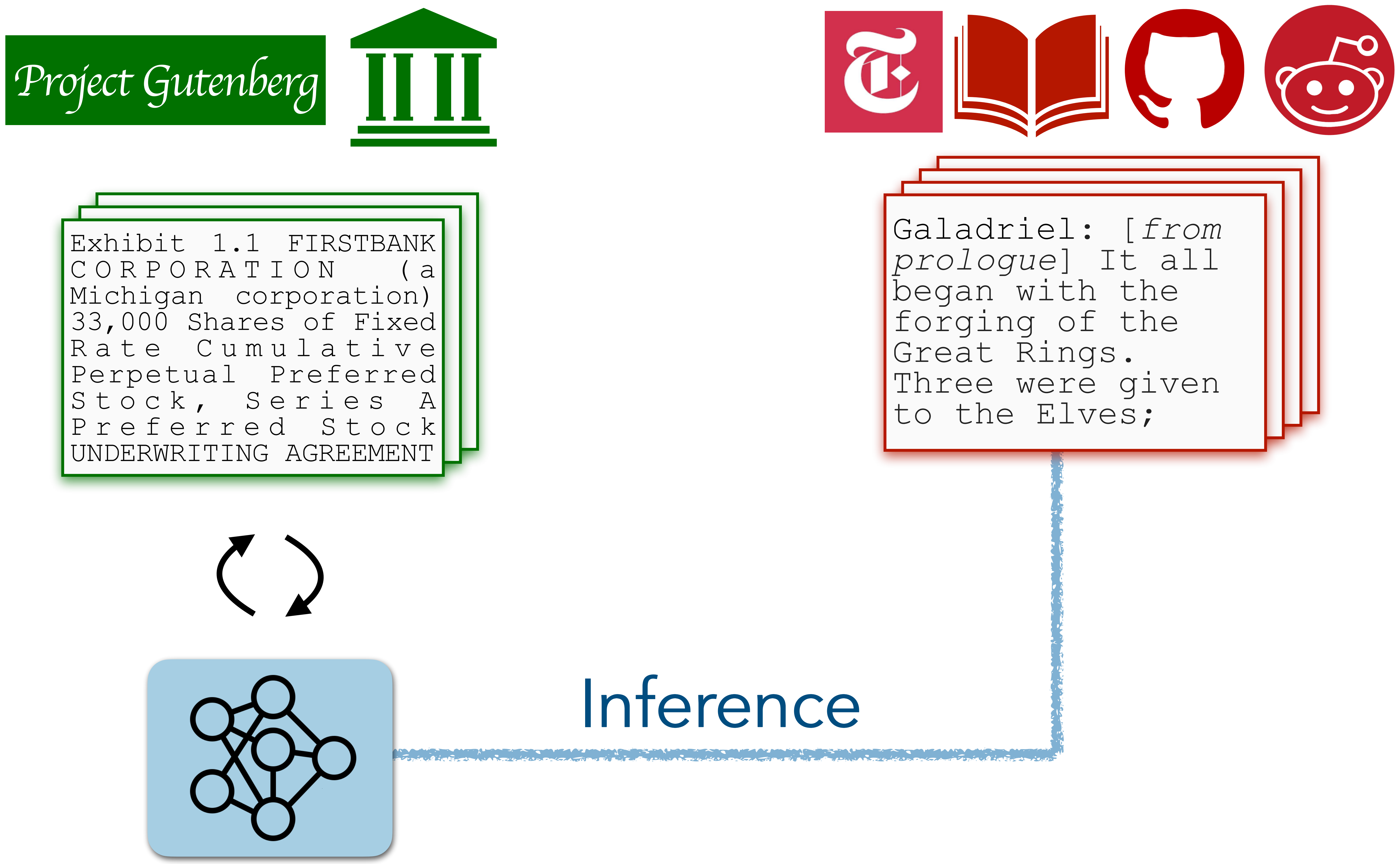
Text copyright © 1997 by J.K. Rowling

Proposal w. retrieval-based LMs



- ☑ Provide attribution
- ☑ Support opt-out

Proposal w. retrieval-based LMs



- ☑ Provide attribution
- ☑ Support opt-out

Proposal w. retrieval-based LMs

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Copilot

Proposal w. retrieval-based LMs

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Copilot

Attribution enables crediting

Proposal w. retrieval-based LMs

Can support opt-out

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Copilot

Attribution enables crediting

Proposal w. retrieval-based LMs

Can support opt-out

Attribution strengthens fair use defense,
Opt-out can also provide defense

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Copilot

Attribution enables crediting

Proposal w. retrieval-based LMs

Can support opt-out

Attribution strengthens fair use defense,
Opt-out can also provide defense

I want my books to be excluded.

I got a lawsuit for copyright infringement from NYT.

I want to get credited whenever the model uses my articles.

I got a lawsuit for violating DMCA (for removing CMI).

Copilot

Attribution enables crediting

Attribution enables providing CMI

Case study: Copyright

SILO: one of the first to mitigate copyright risks in general domains

Case study: Copyright

SILO: one of the first to mitigate copyright risks in general domains

1. What challenges would arise if we filter out all copyrighted text?

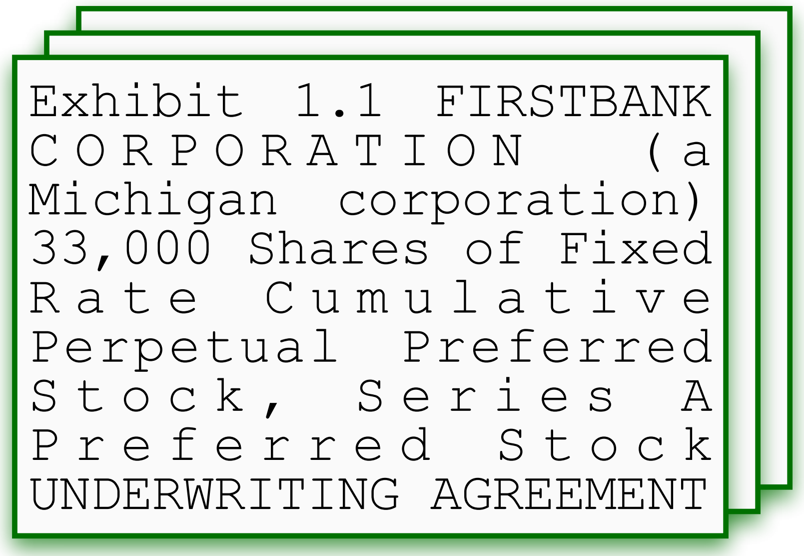
Case study: Copyright

SILO: one of the first to mitigate copyright risks in general domains

1. What challenges would arise if we filter out all copyrighted text?
2. Can SILO match performance of existing models?

SILO: (1) Collect data

Project Gutenberg



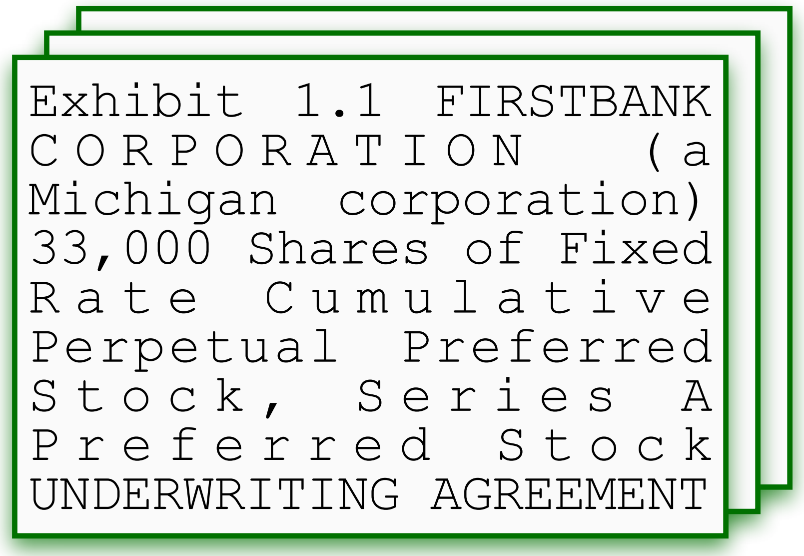
OLC

(Open License Corpus)

*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

SILO: (1) Collect data

Project Gutenberg



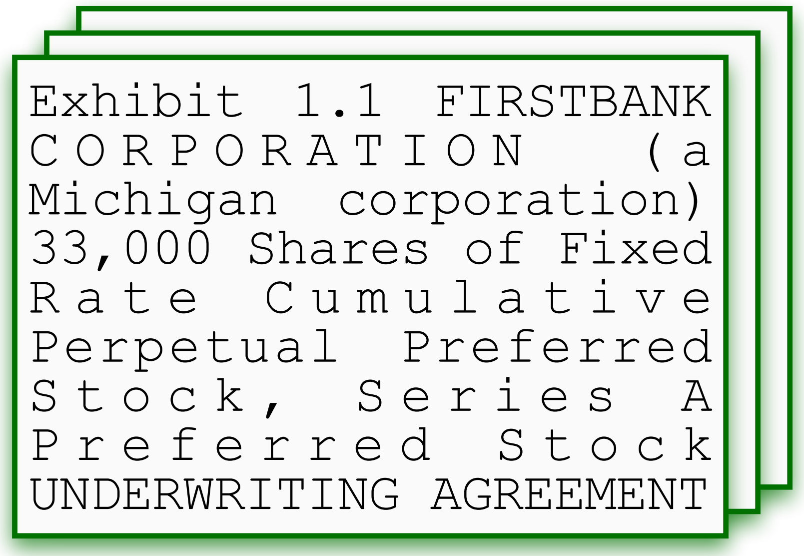
OLC

(Open License Corpus)

100B words

*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

SILO: (1) Collect data



OLC

(Open License Corpus)

100B words

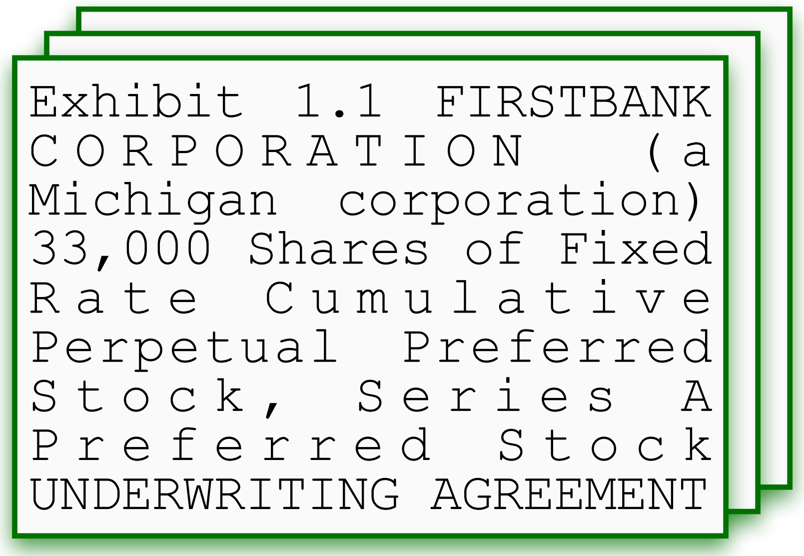
Reasonably large

(1/3 of GPT-3 training data)

*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

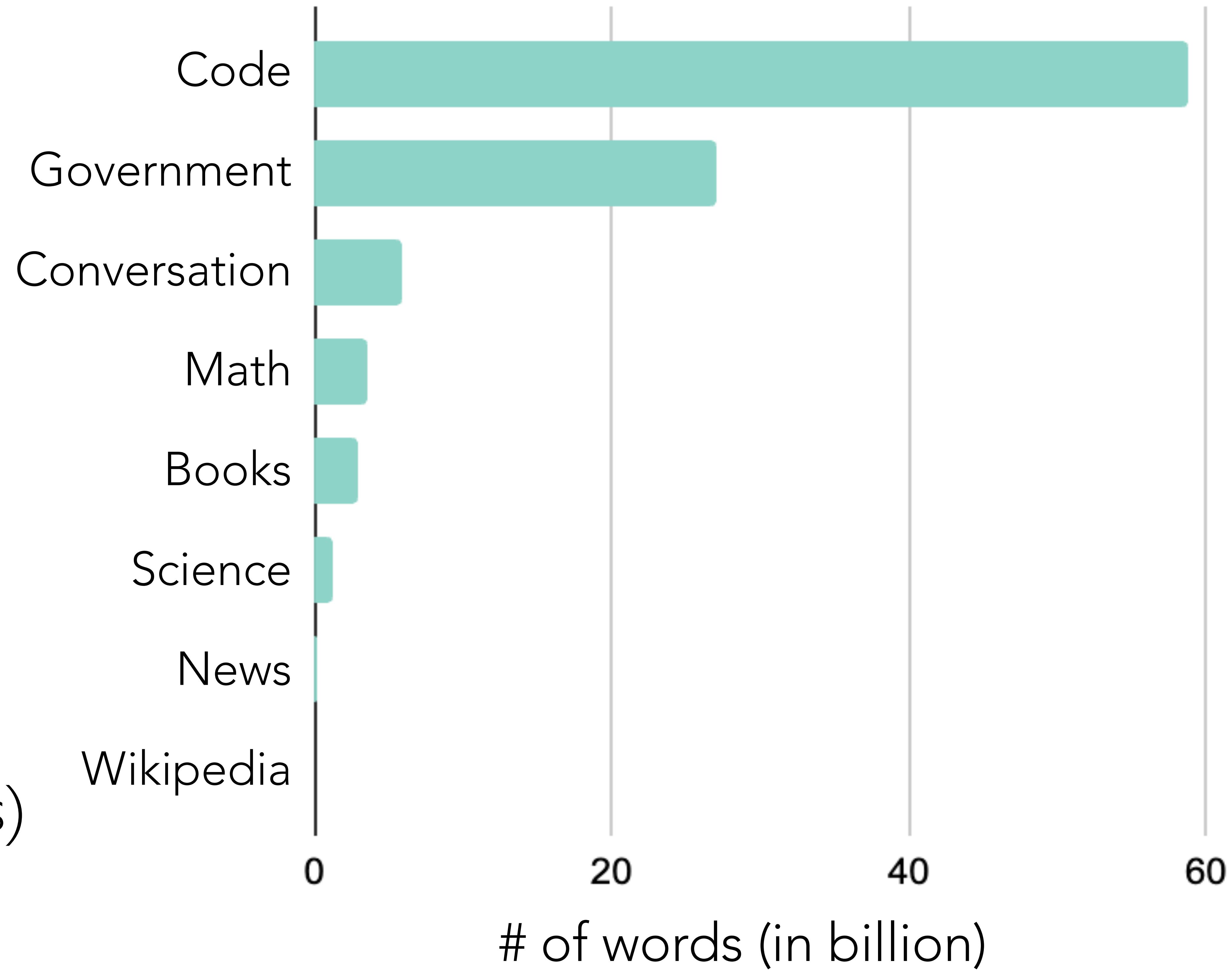
SILO: (1) Collect data

Project Gutenberg



OLC

(Open License Corpus)



*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

SILO: (1) Collect data

Project Gutenberg

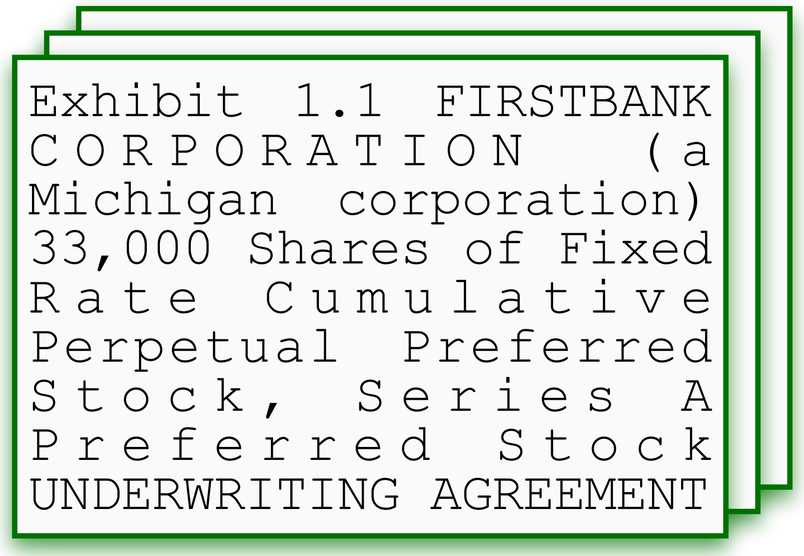
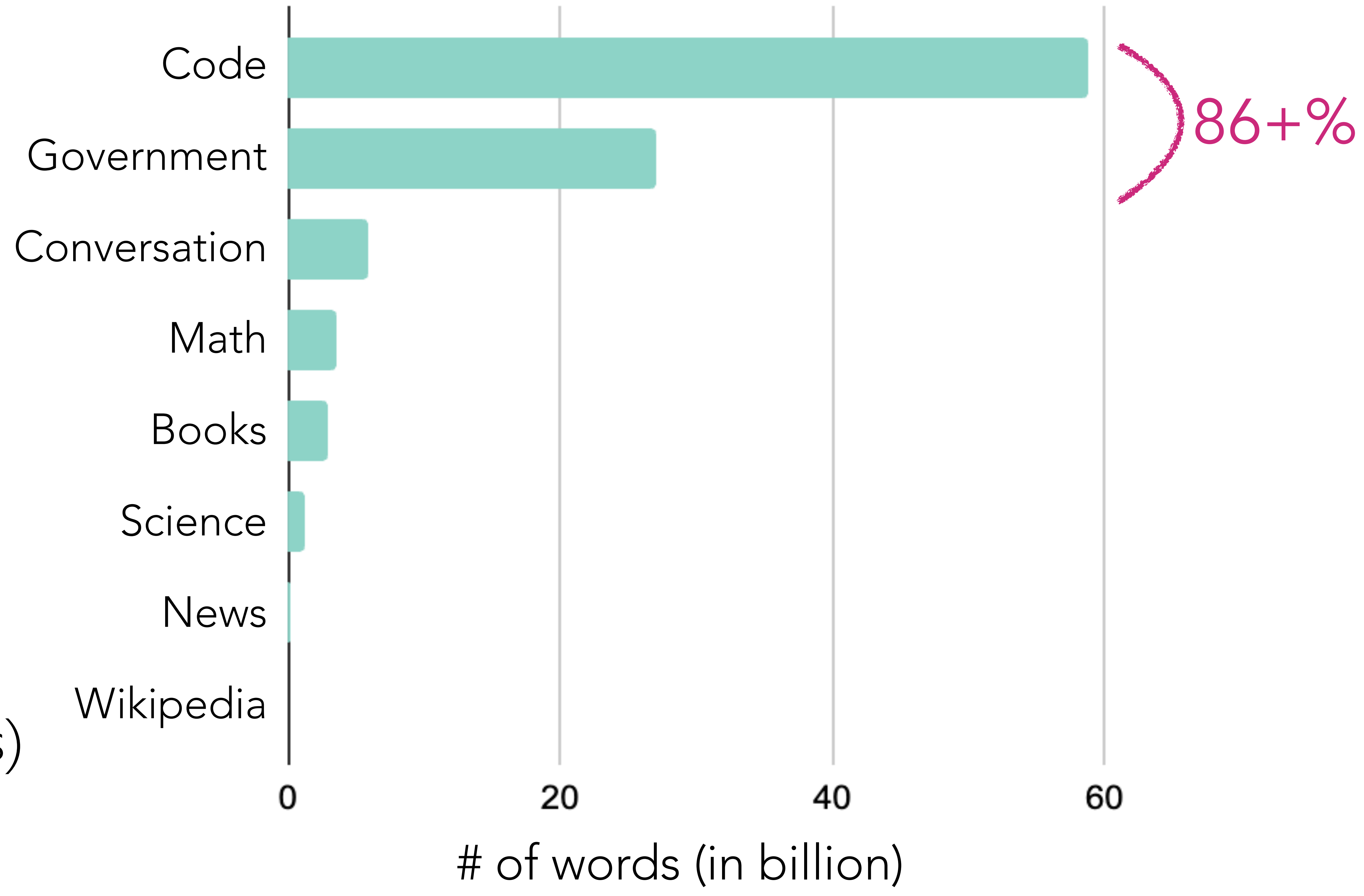


Exhibit 1.1 FIRSTBANK CORPORATION (a Michigan corporation) 33,000 Shares of Fixed Rate Cumulative Perpetual Preferred Stock, Series A Preferred Stock UNDERWRITING AGREEMENT

OLC

(Open License Corpus)



*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

SILO: (1) Collect data

Project Gutenberg

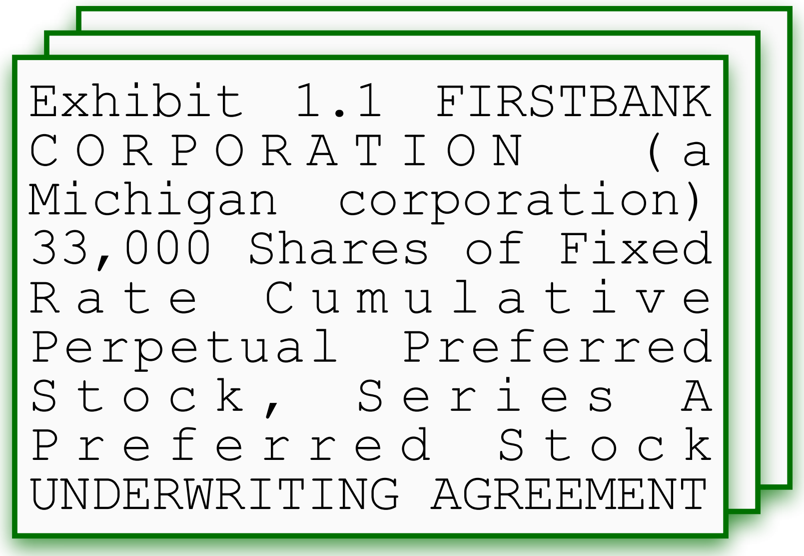
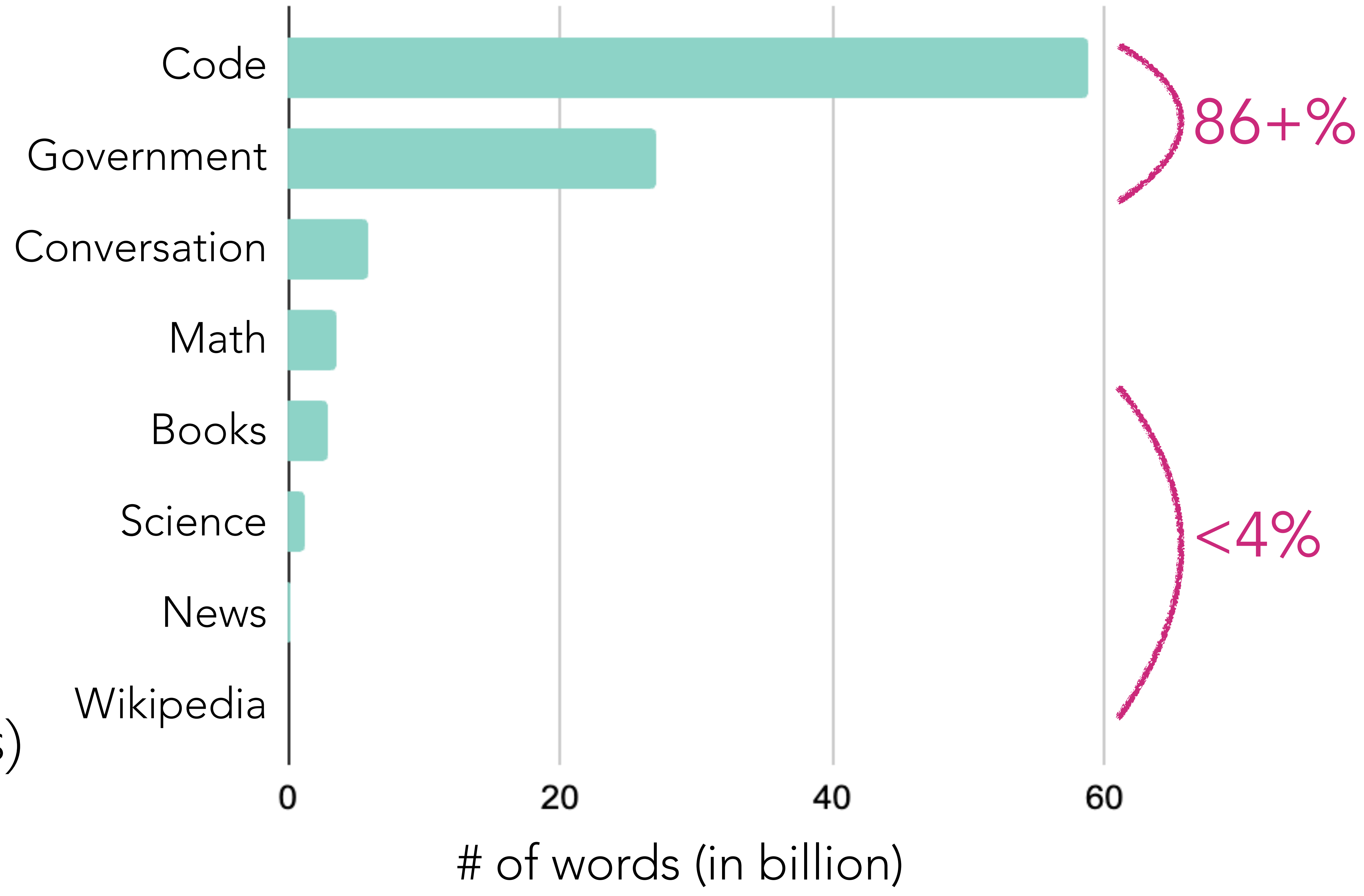


Exhibit 1.1 FIRSTBANK CORPORATION (a Michigan corporation) 33,000 Shares of Fixed Rate Cumulative Perpetual Preferred Stock, Series A Preferred Stock UNDERWRITING AGREEMENT

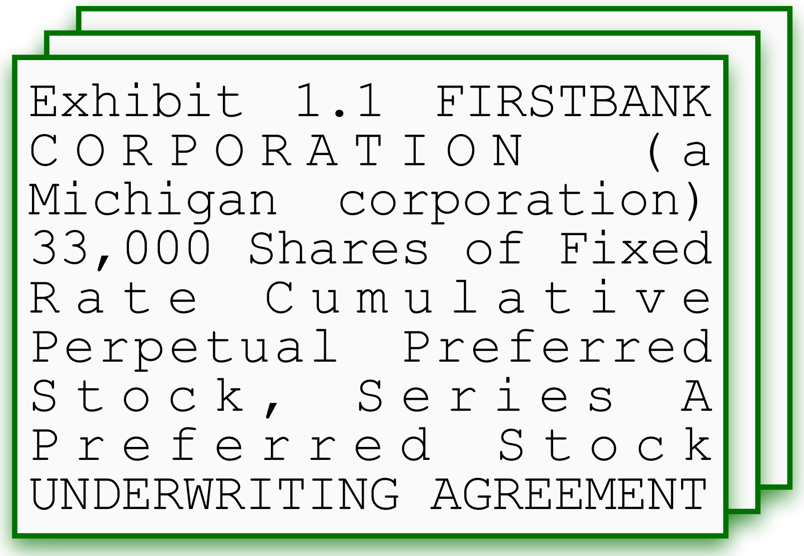
OLC

(Open License Corpus)



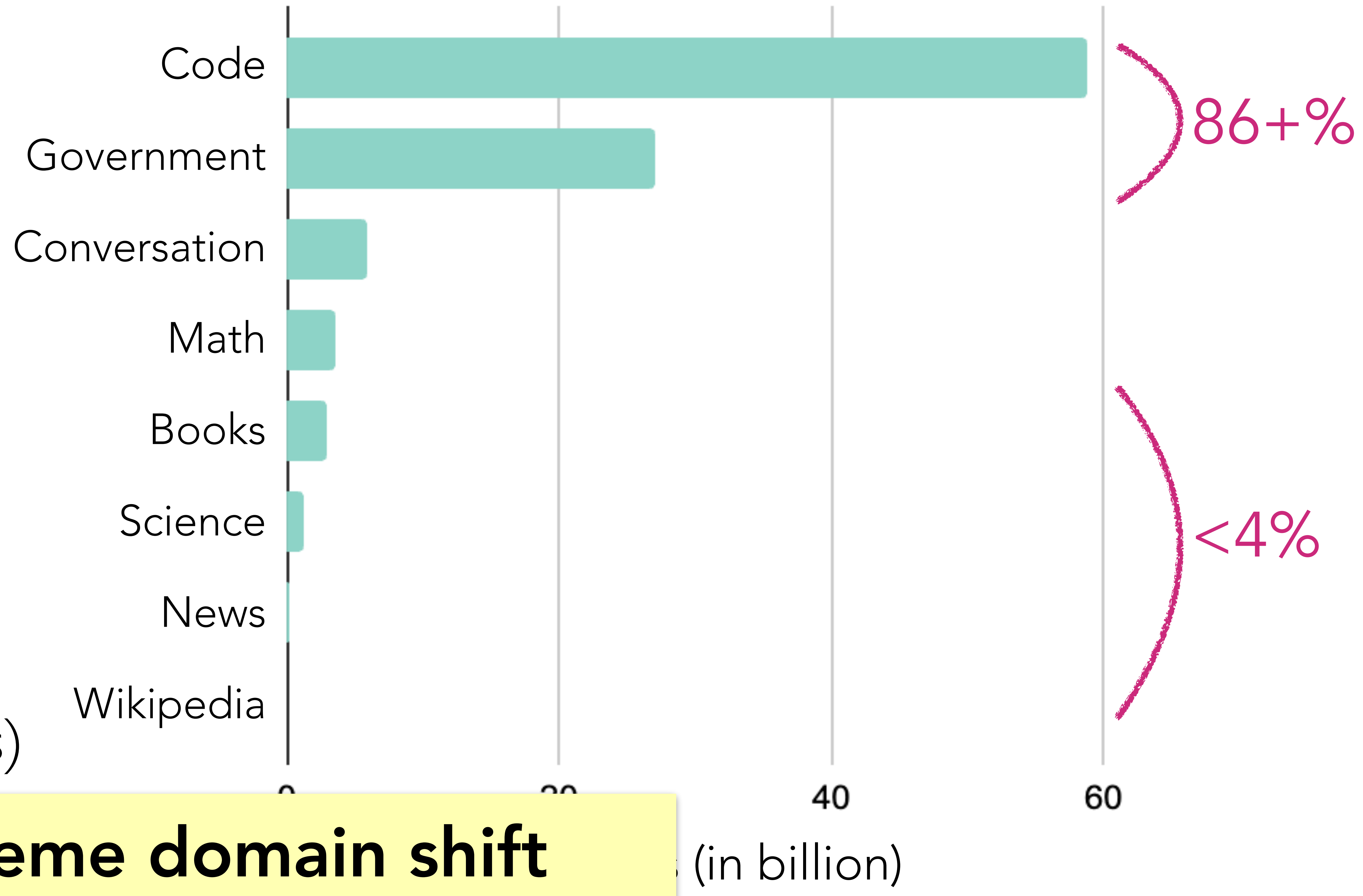
*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

SILO: (1) Collect data



OLC

(Open License Corpus)

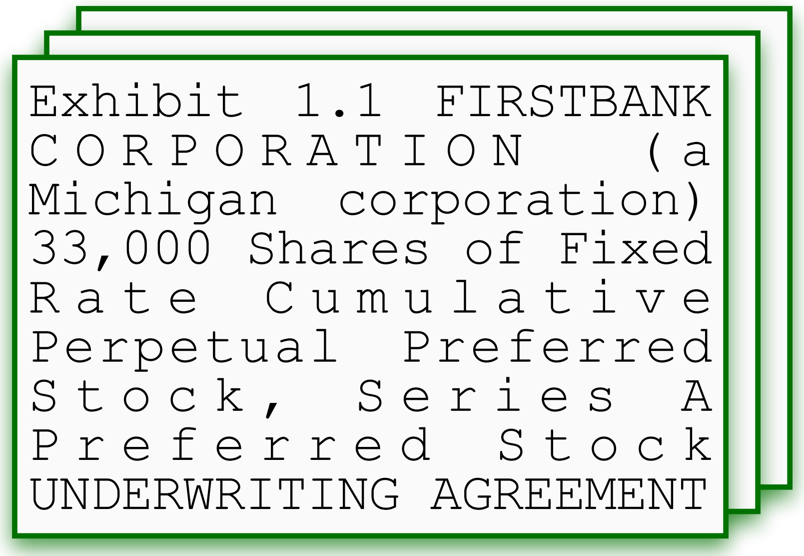


Extreme domain shift

*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

SILO: (2) Build models

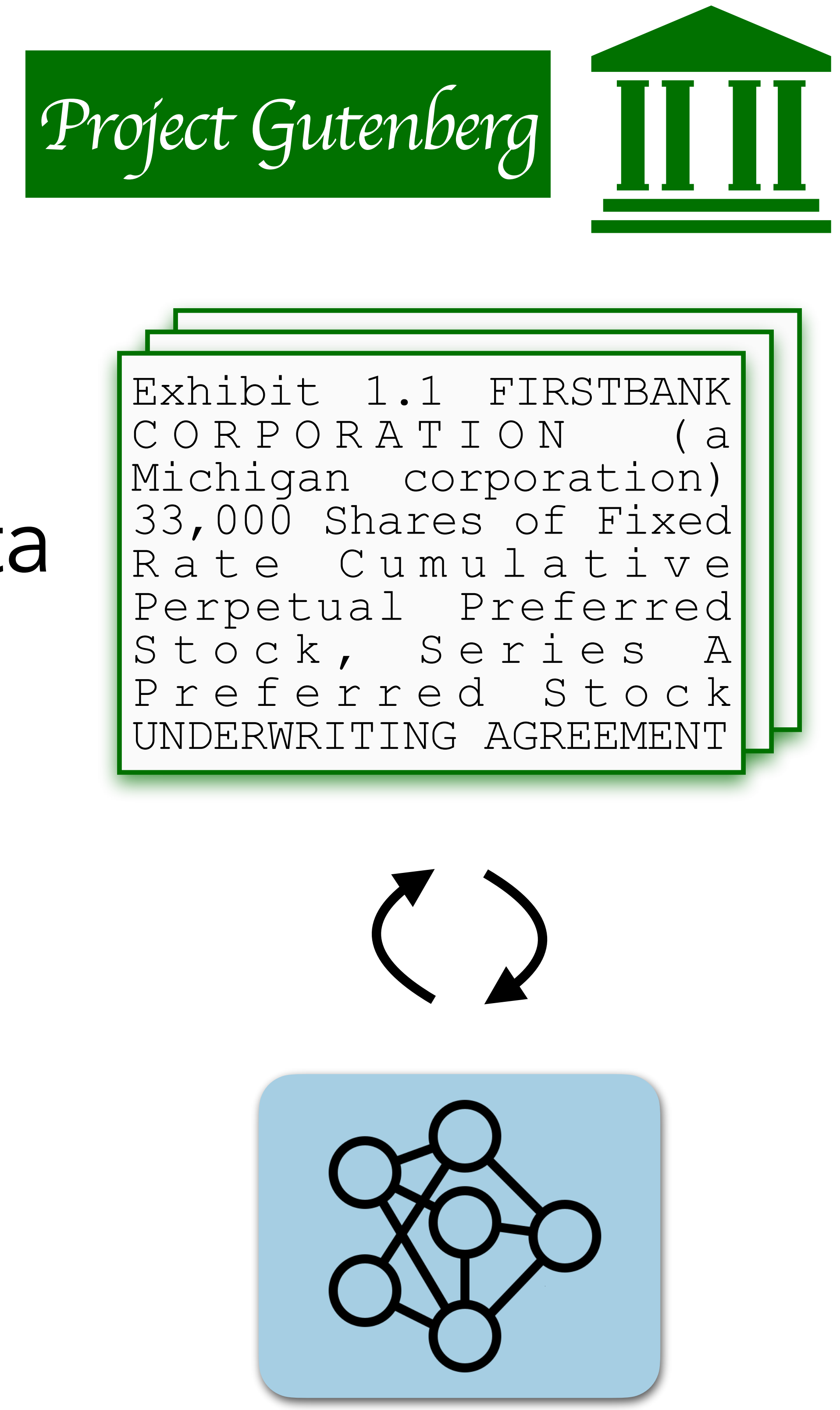
Project Gutenberg



*The definition of “permissively-licensed” could largely vary;
Here, it refers to public domain and permissive software licenses

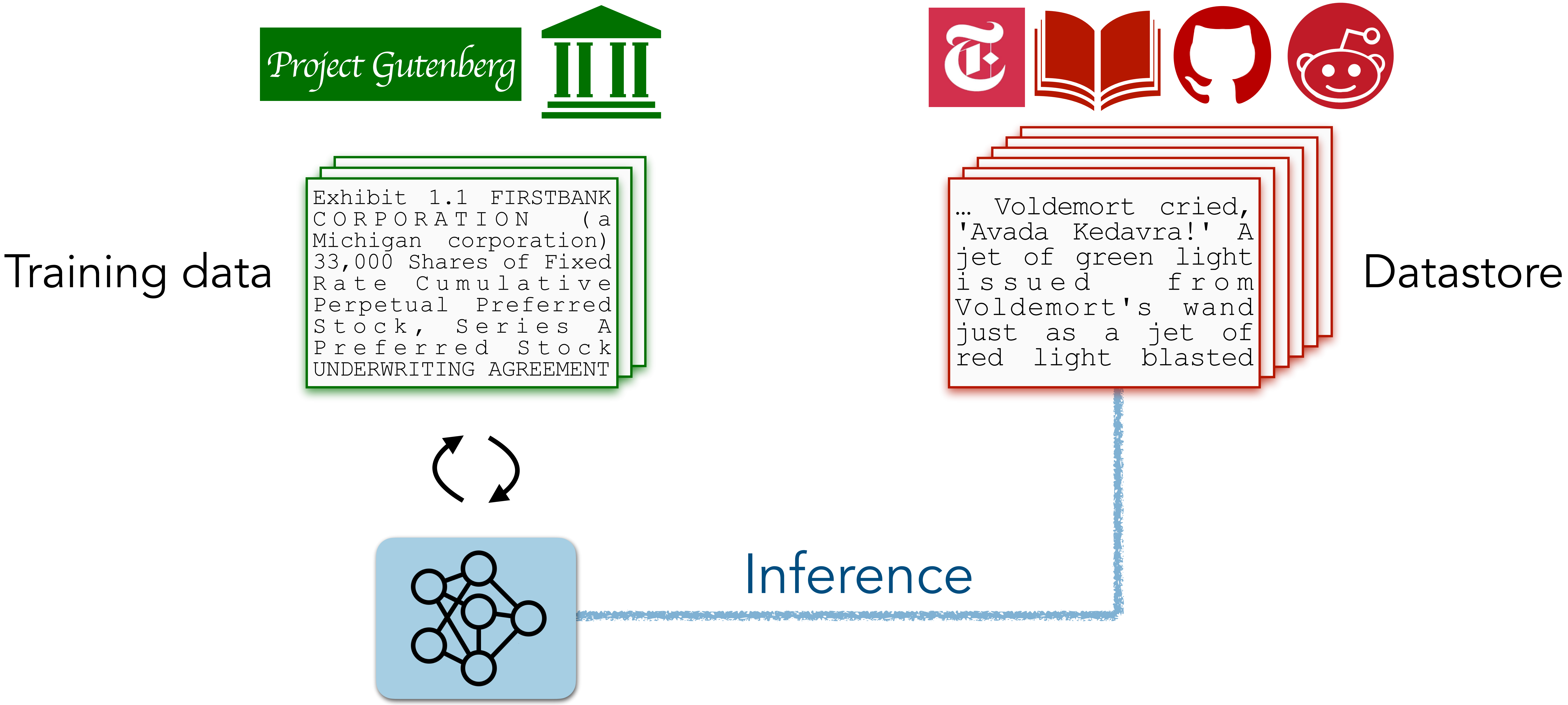
SILO: (2) Build models

Training data



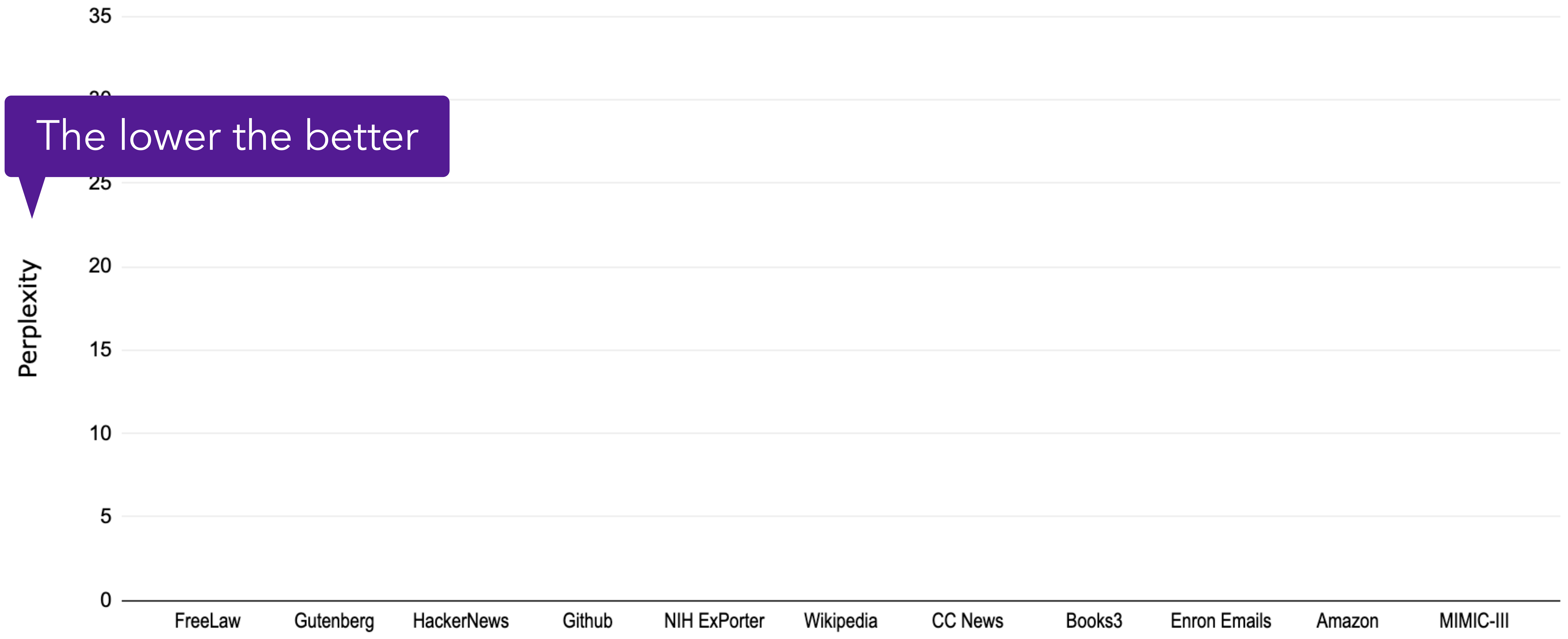
*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

SILO: (2) Build models

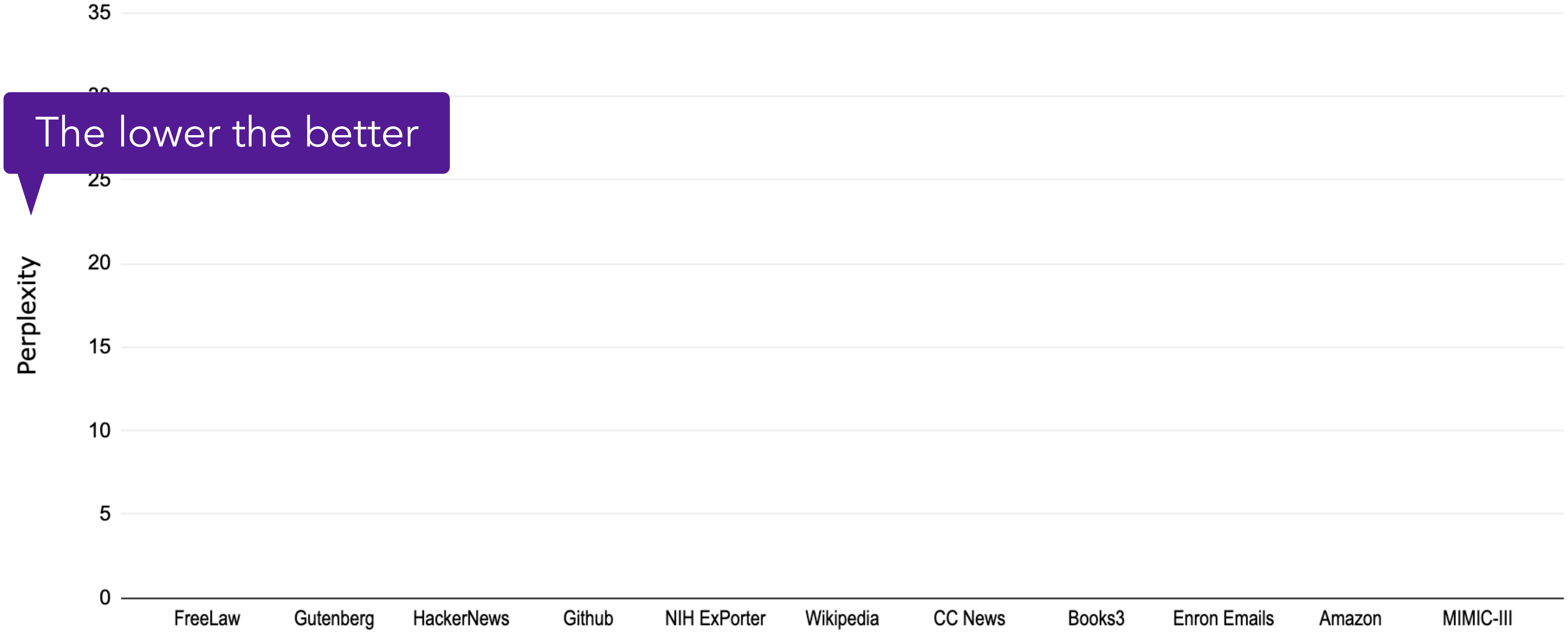


*The definition of "permissively-licensed" could largely vary;
Here, it refers to public domain and permissive software licenses

Experiments



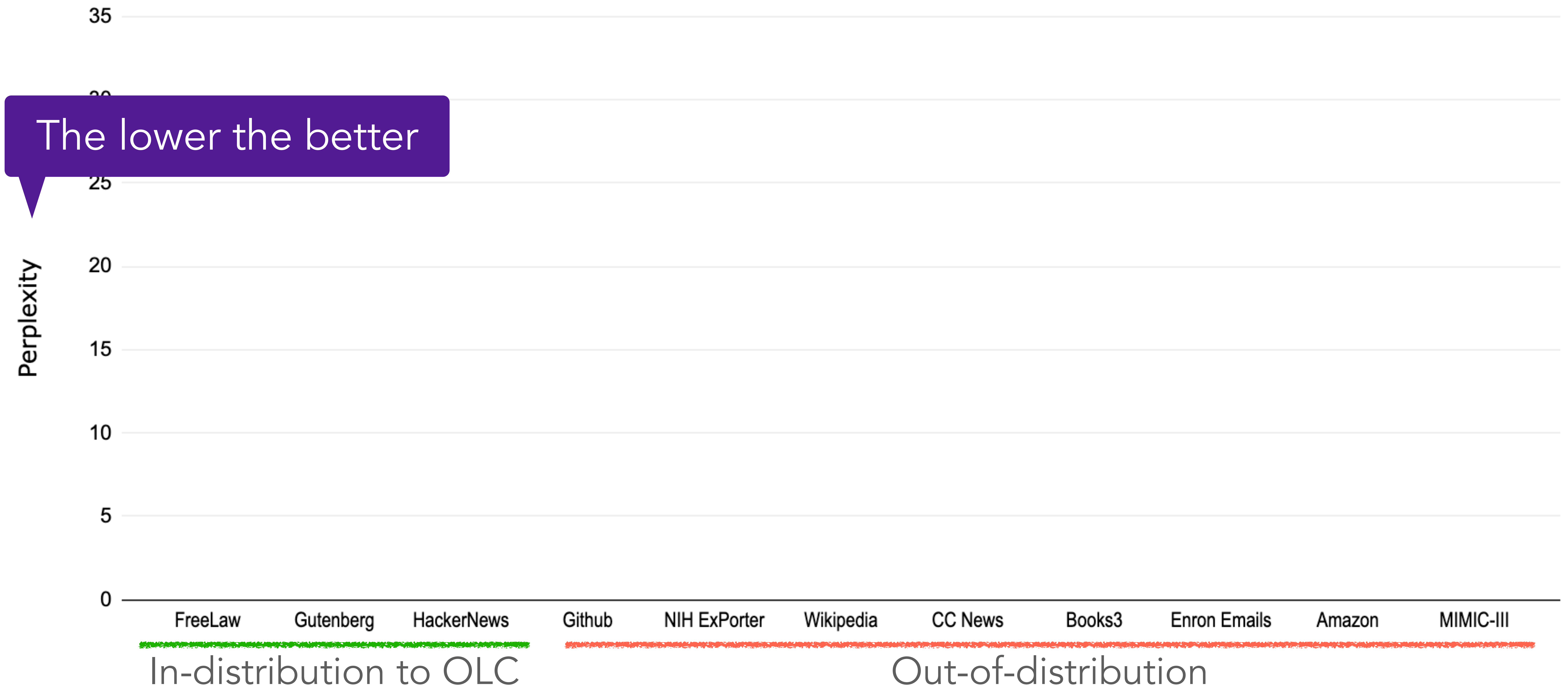
Experiments

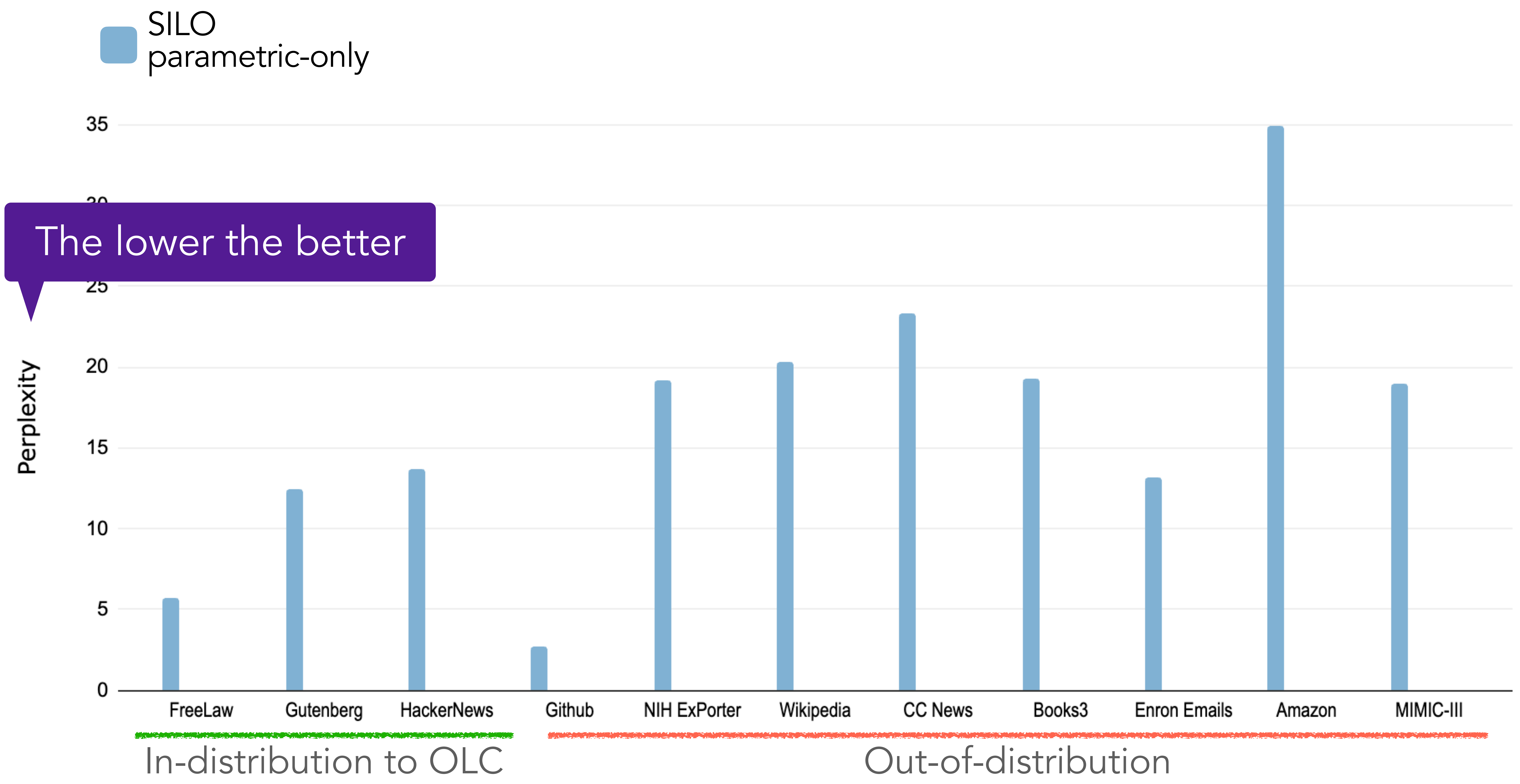


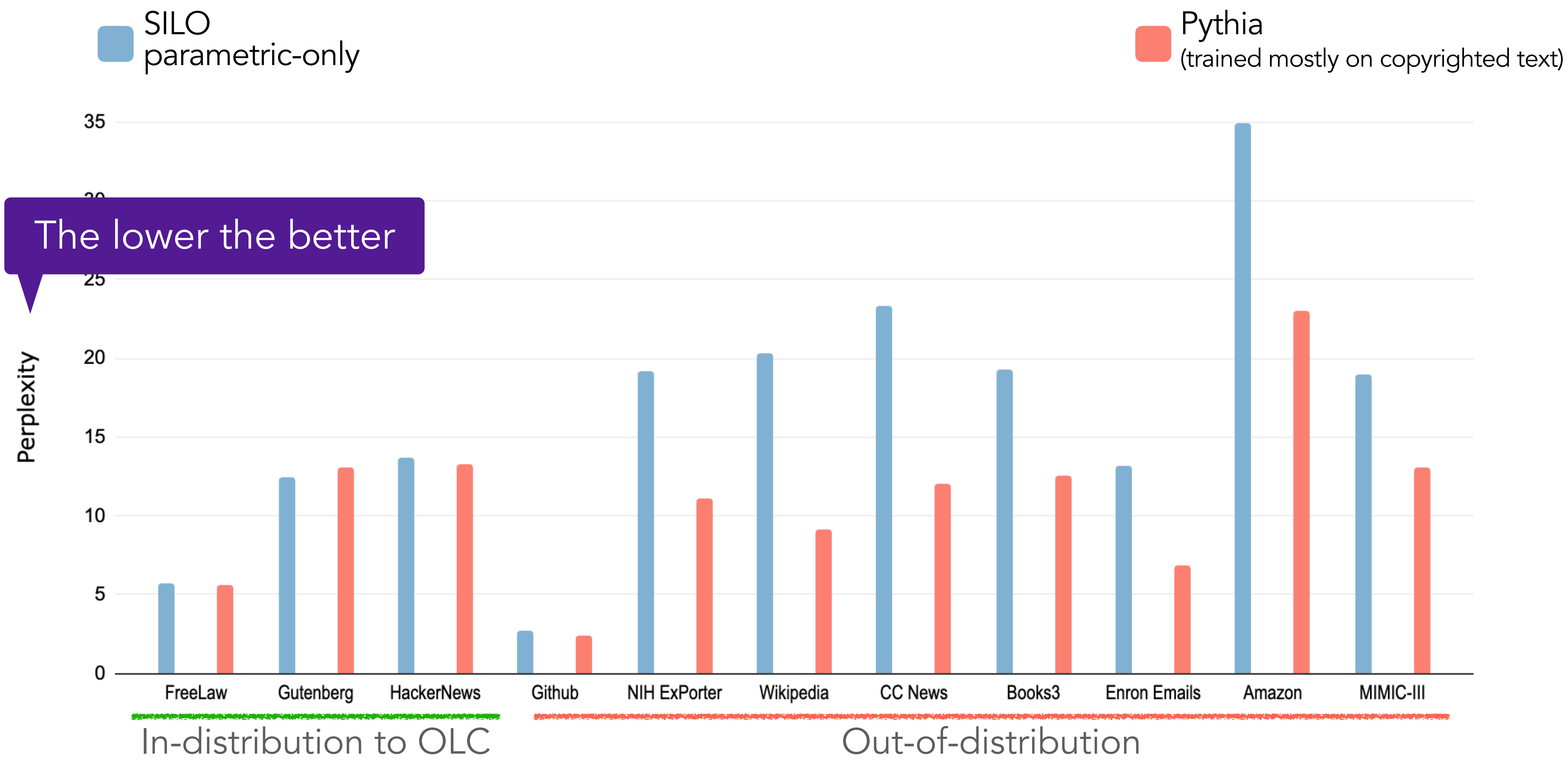
The lower the better

In-distribution to OLC

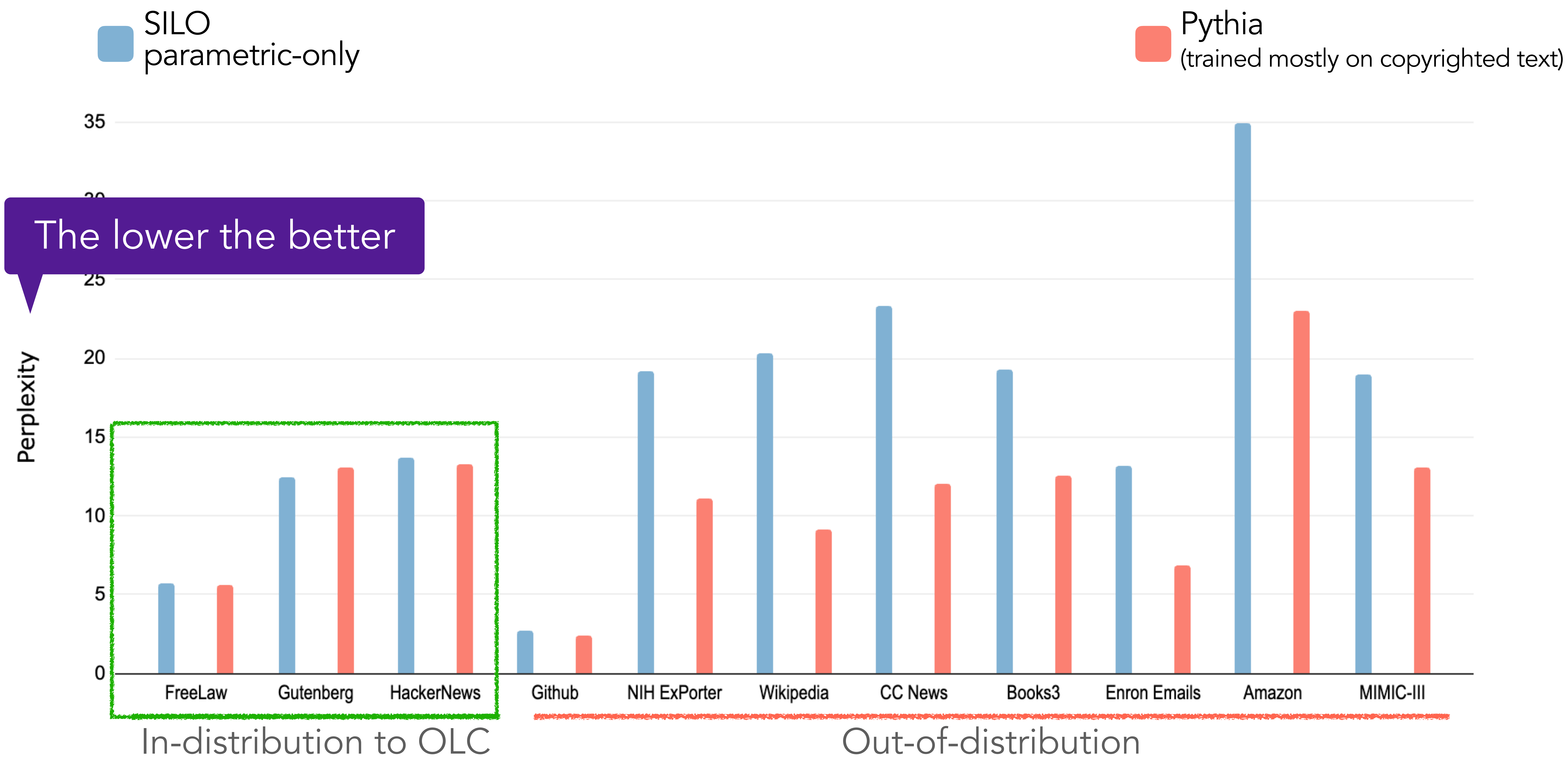
Experiments

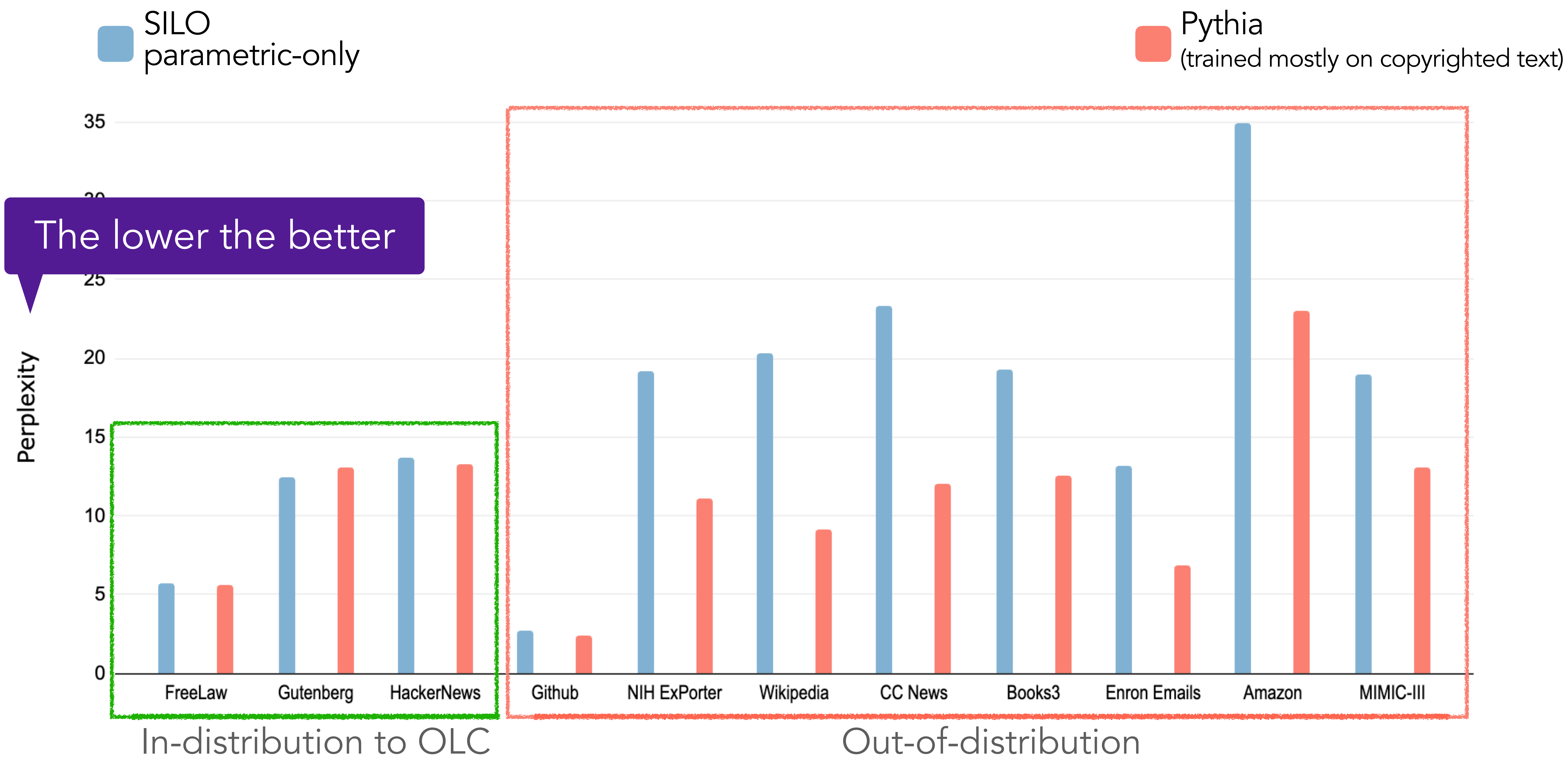



SILO
parametric-only
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution

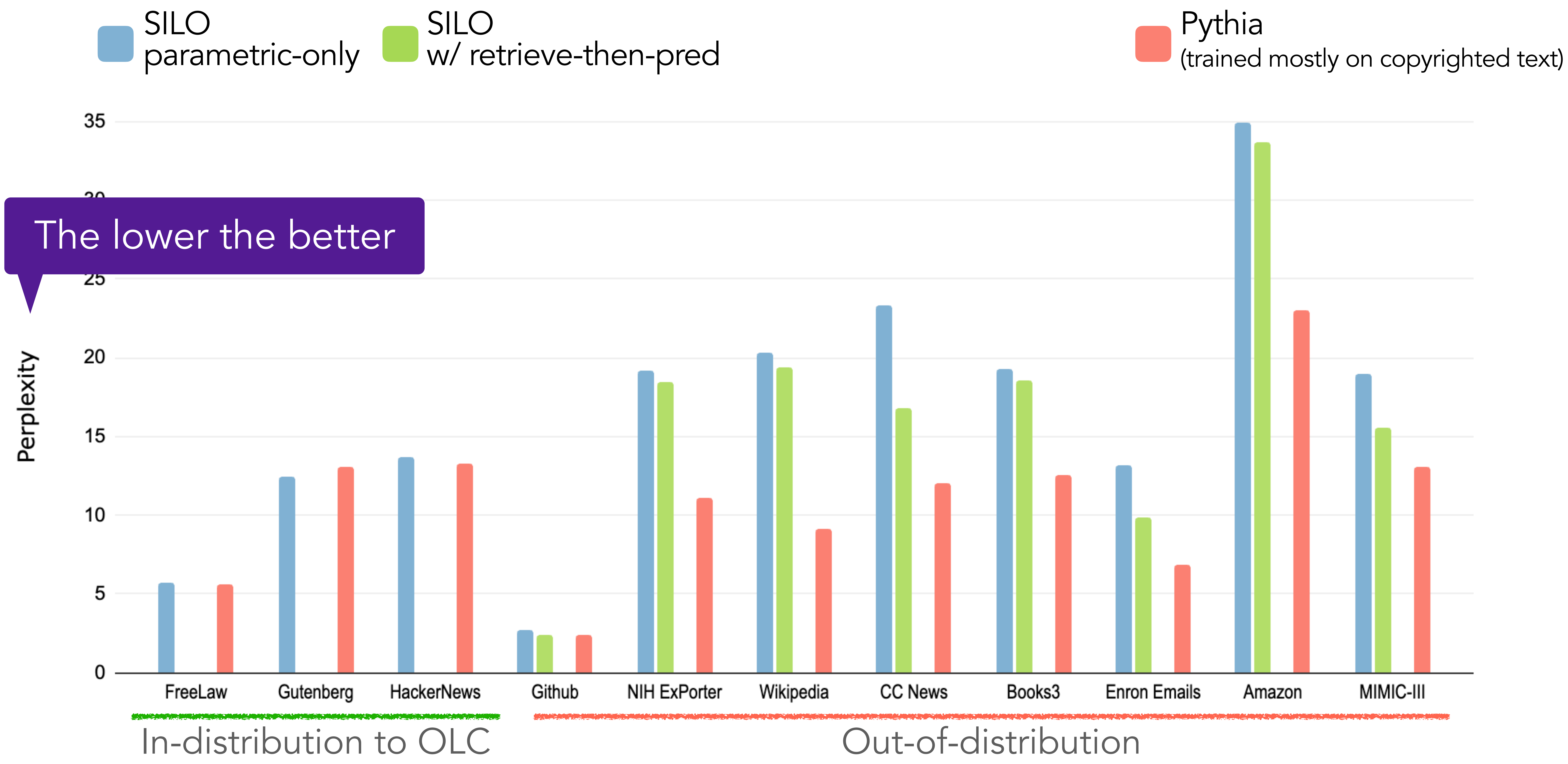


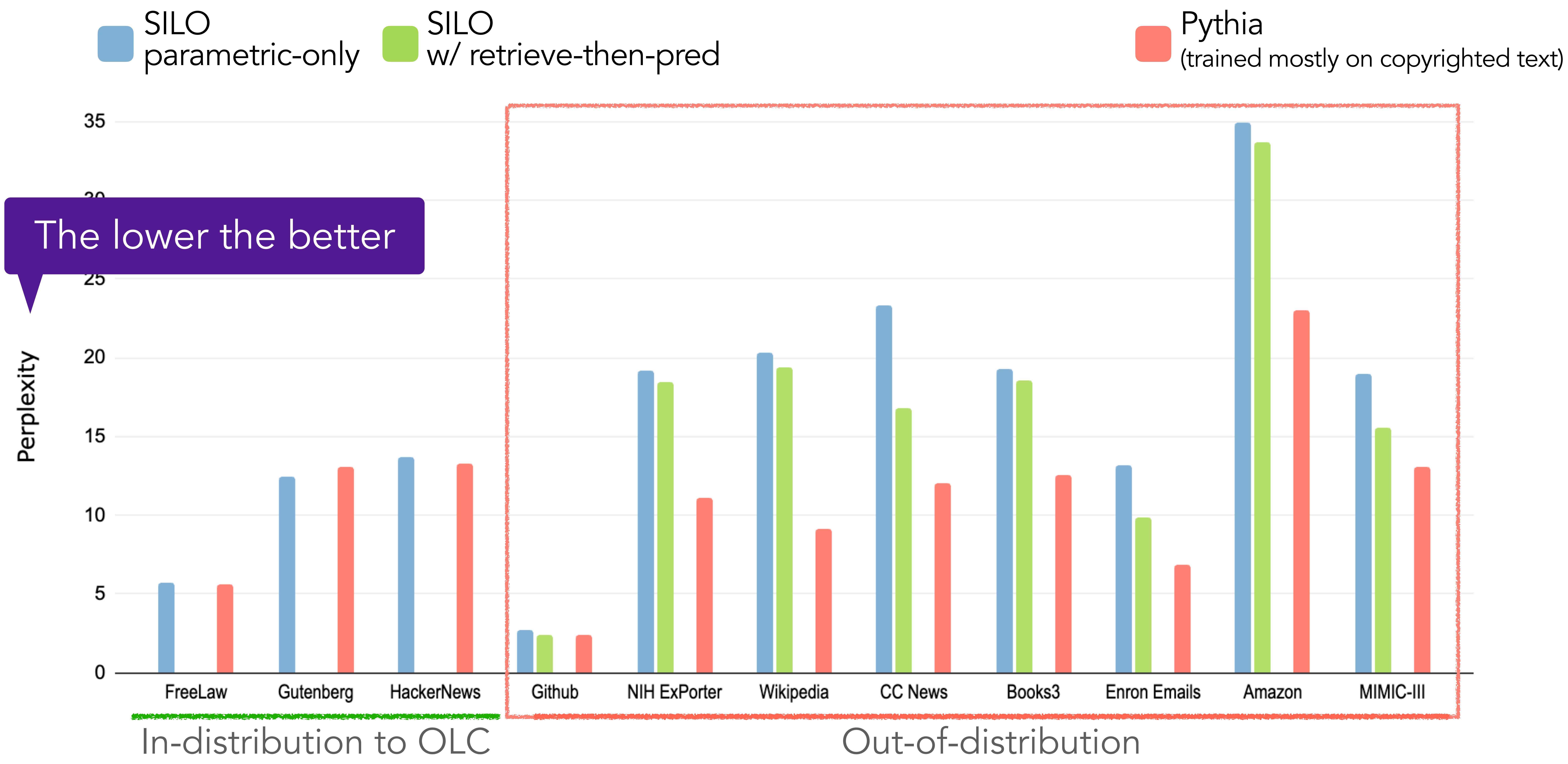
SILO
parametric-only
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
30
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution

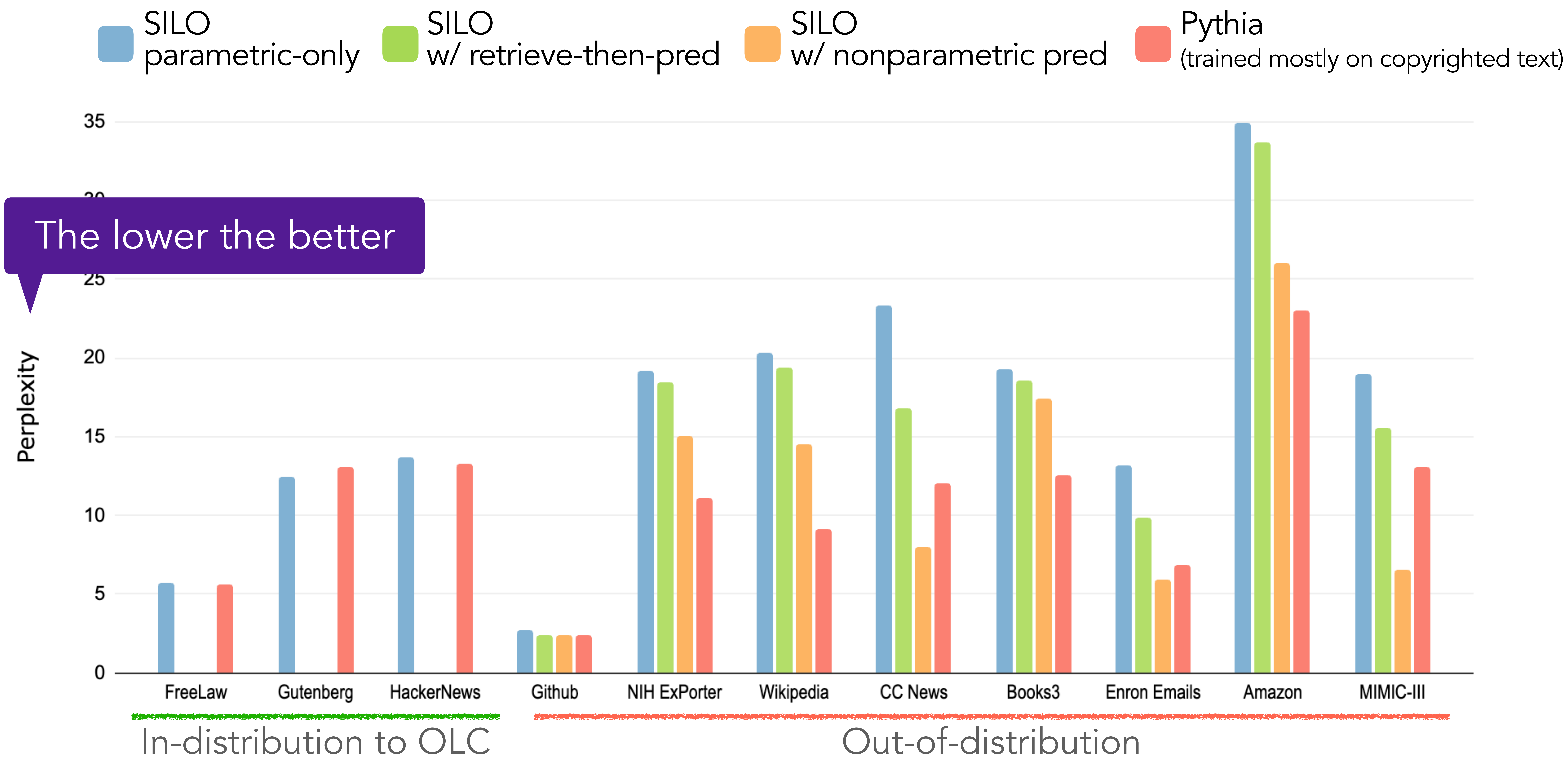

SILO
parametric-only
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution



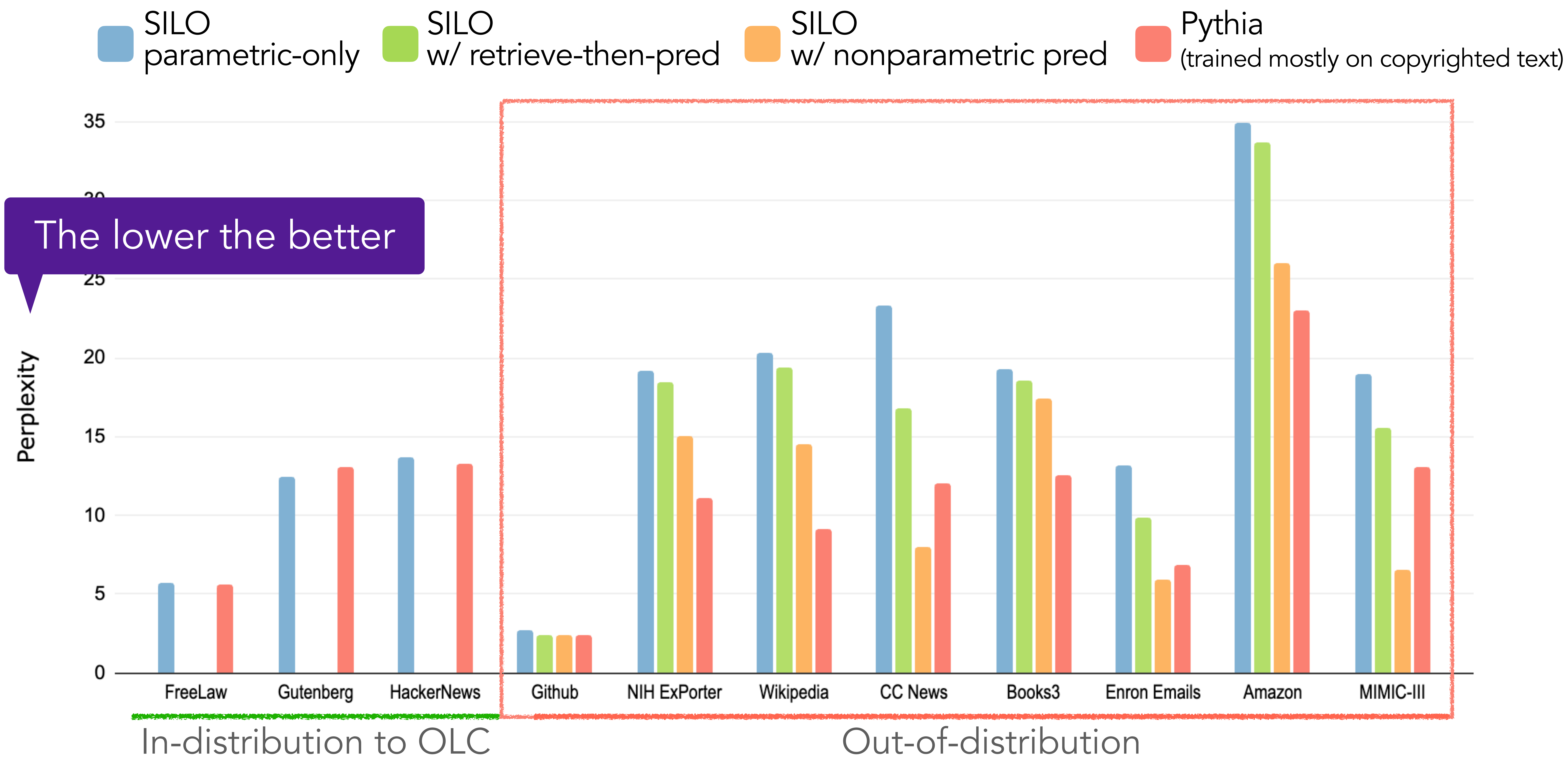
SILO
parametric-only
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
30
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution

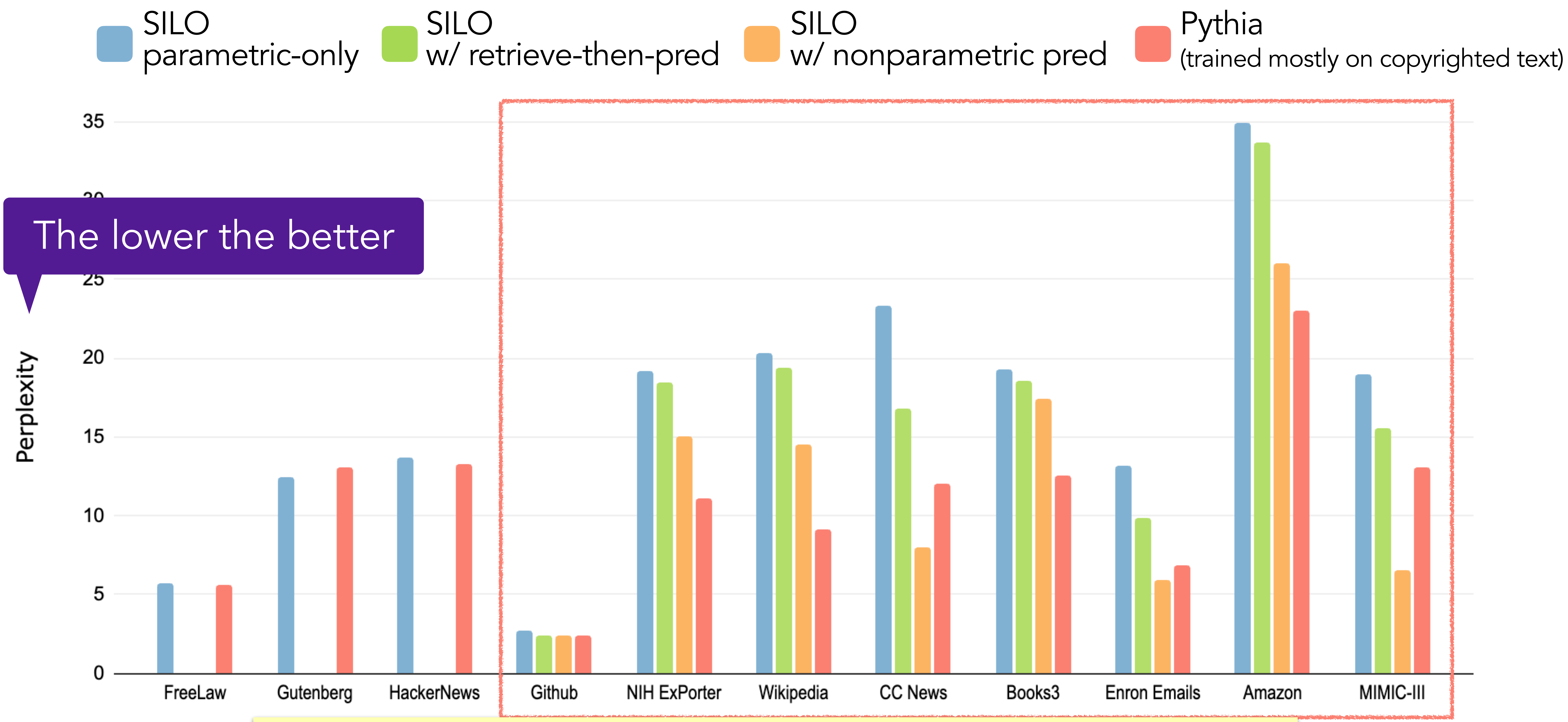

SILO parametric-only
SILO w/ retrieve-then-pred
Pythia (trained mostly on copyrighted text)
The lower the better
Perplexity
35
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution


SILO
parametric-only
SILO
w/ retrieve-then-pred
Pythia
(trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution



SILO parametric-only
SILO w/ retrieve-then-pred
SILO w/ nonparametric pred
Pythia (trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution

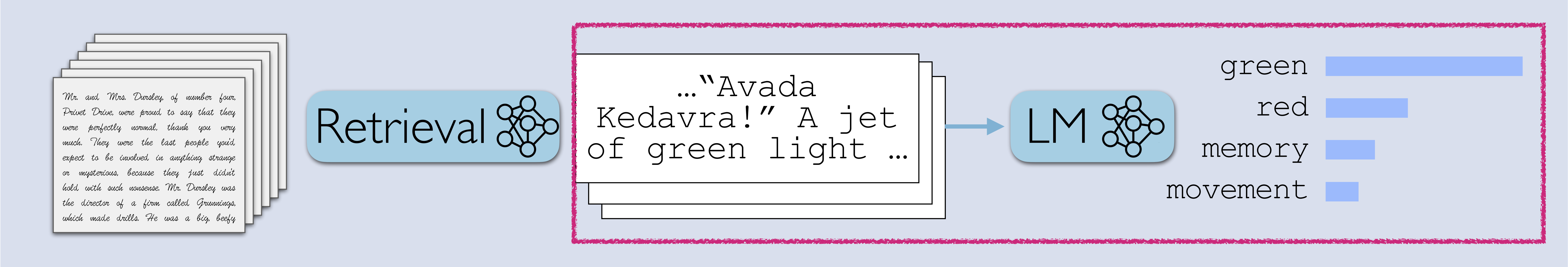

SILO parametric-only
SILO w/ retrieve-then-pred
SILO w/ nonparametric pred
Pythia (trained mostly on copyrighted text)
The lower the better
Perplexity
35
25
20
15
10
5
0
FreeLaw
Gutenberg
HackerNews
Github
NIH ExPorter
Wikipedia
CC News
Books3
Enron Emails
Amazon
MIMIC-III
In-distribution to OLC
Out-of-distribution



The lower the better

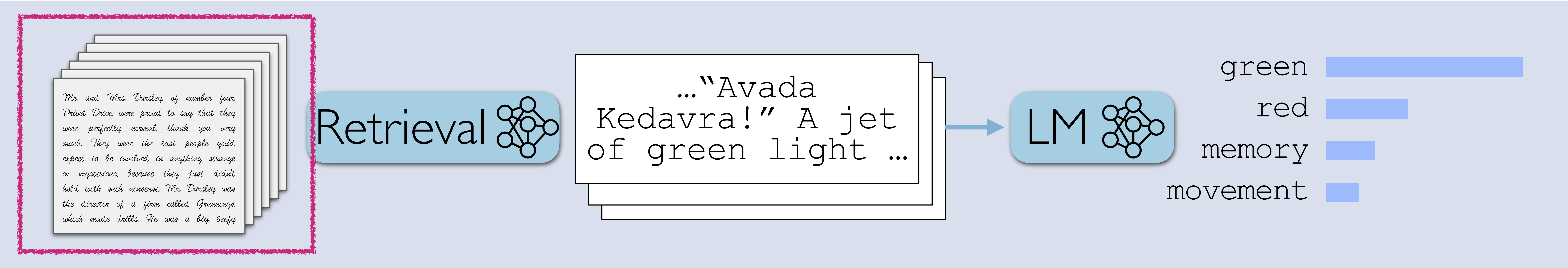
Lots of open questions on further reducing the gap + generalizing to downstream tasks

Summary of this section



Pre-training w/ retrieval: We can pre-train LLM w/ retrieval efficiently

Summary of this section

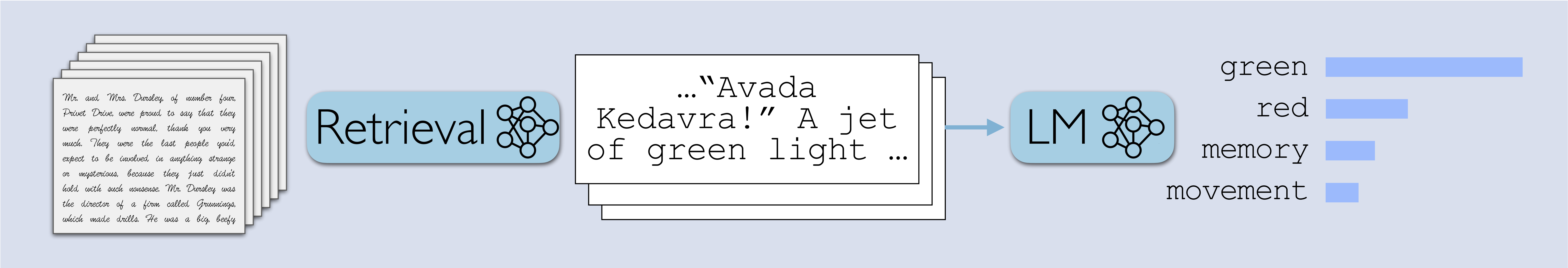


Pre-training w/ retrieval: We can pre-train LLM w/ retrieval efficiently

Scaling a datastore: A datastore size can provide a new avenue for scaling

How to ensure a wider range of tasks benefit more from retrieval is an open problem

Summary of this section



Pre-training w/ retrieval: We can pre-train LLM w/ retrieval efficiently

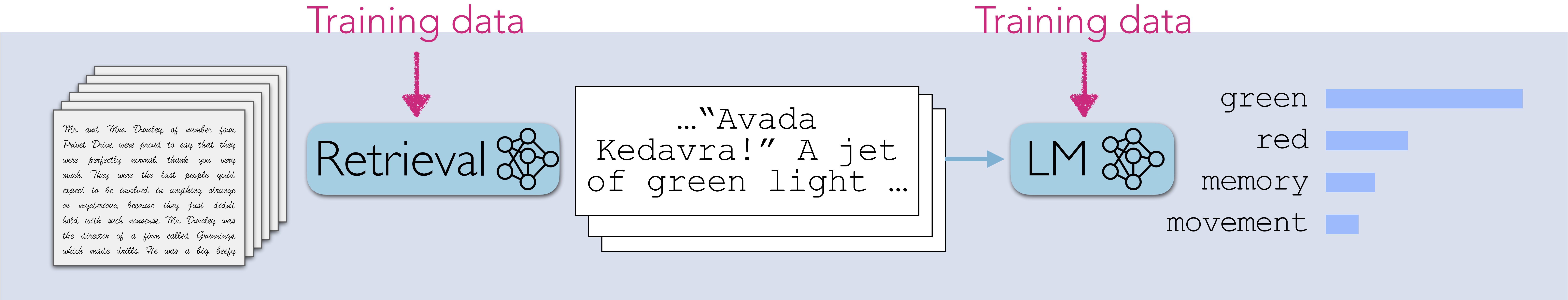
Scaling a datastore: A datastore size can provide a new avenue for scaling

How to ensure a wider range of tasks benefit more from retrieval is an open problem

SILO: the first prototype to separate permissive and restrictive data for supporting attribution and opt-out

How to ensure this model works on par with the model trained on all data remains open

Summary of this section



Pre-training w/ retrieval: We can pre-train LLM w/ retrieval efficiently

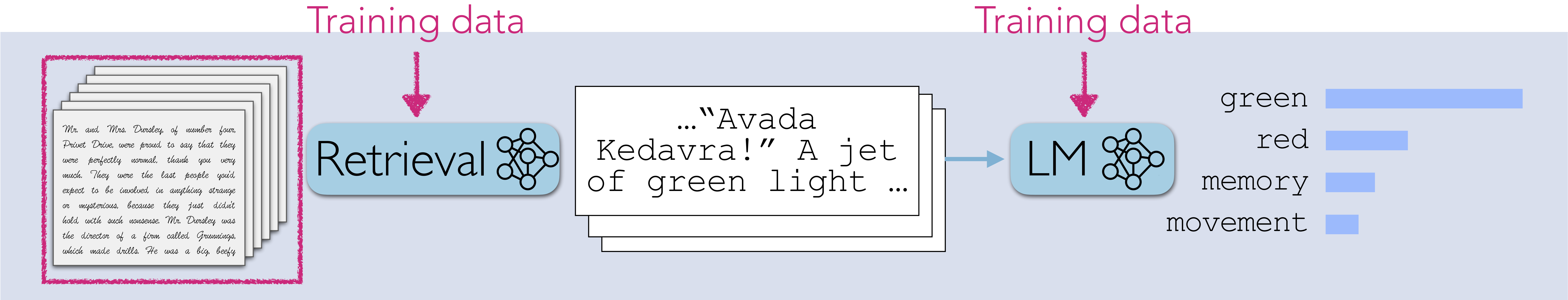
Scaling a datastore: A datastore size can provide a new avenue for scaling

How to ensure a wider range of tasks benefit more from retrieval is an open problem

SILO: the first prototype to separate permissive and restrictive data for supporting attribution and opt-out

How to ensure this model works on par with the model trained on all data remains open

Summary of this section



Pre-training w/ retrieval: We can pre-train LLM w/ retrieval efficiently

Scaling a datastore: A datastore size can provide a new avenue for scaling

How to ensure a wider range of tasks benefit more from retrieval is an open problem

SILO: the first prototype to separate permissive and restrictive data for supporting attribution and opt-out

How to ensure this model works on par with the model trained on all data remains open

QnA for Part 2

Today's Lecture

Part 1. **Basics** of retrieval-based LMs (35min)

- Retrieval
- Augmentation
- Training of retrieval-based LMs

Part 2. **Recent research** on *scaling* retrieval-based LMs (35min)

- Scalable Pre-training with Retrieval
- Scaling a Datastore
- Datastore for Responsible Data Use

Open Problems (10min)

Today's Lecture

Part 1. **Basics** of retrieval-based LMs (35min)

- Retrieval
- Augmentation
- Training of retrieval-based LMs

Part 2. **Recent research** on *scaling* retrieval-based LMs (35min)

- Scalable Pre-training with Retrieval
- Scaling a Datastore
- Datastore for Responsible Data Use

Open Problems (10min)

Open questions (1/3)

Open questions (1/3)

Q1: Can retrieval help reasoning?

Open questions (1/3)

Q1: Can retrieval help reasoning?

Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model

Parishad BehnamGhader[1] **Santiago Miret**[3] **Siva Reddy**[1,2]
[1]McGill University / Mila [2]Facebook CIFAR AI Chair [3]Intel Labs
{parishad.behnamghader, siva.reddy}@mila.quebec
santiago.miret@intel.com

Great Memory, Shallow Reasoning: Limits of kNN-LMs

Shangyi Geng **Wenting Zhao** **Alexander M Rush**
Cornell University
{sg2323, wz346, arush}@cornell.edu

Open questions (1/3)

Q1: Can retrieval help reasoning?

Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model

Parishad BehnamGhader[1] **Santiago Miret**[3] **Siva Reddy**[1,2]
[1]McGill University / Mila [2]Facebook CIFAR AI Chair [3]Intel Labs
{parishad.behnamghader, siva.reddy}@mila.quebec
santiago.miret@intel.com

Great Memory, Shallow Reasoning: Limits of kNN-LMs

Shangyi Geng **Wenting Zhao** **Alexander M Rush**
Cornell University
{sg2323, wz346, arush}@cornell.edu

Search-o1: Agentic Search-Enhanced Large Reasoning Models

Xiaoxi Li[1], **Guanting Dong**[1], **Jiajie Jin**[1], **Yuyao Zhang**[1], **Yujia Zhou**[2], **Yutao Zhu**[1], **Peitian Zhang**[1], **Zhicheng Dou**[1*]
[1]Renmin University of China [2]Tsinghua University
{xiaoxi_li, dou}@ruc.edu.cn

Project Page: https://search-o1.github.io/

Agentic Reasoning: Reasoning LLMs with Tools for the Deep Research

Junde Wu, Jiayuan Zhu, Yuyuan Liu
University of Oxford

Open questions (1/3)

Q1: Can retrieval help reasoning?

Method	Phy.	Chem.	Bio.
Direct Reasoning			
Qwen2.5-32B	57.0	33.3	52.6
Qwen2.5-Coder-32B	37.2	25.8	57.9
QwQ-32B	75.6	39.8	68.4
Qwen2.5-72B	57.0	37.6	68.4
Llama3.3-70B	54.7	31.2	52.6
GPT-4o†	59.5	40.2	61.6
o1-preview†	89.4	59.9	65.9
Retrieve/Search in Reasoning			
RAG-Qwen2.5-32B	57.0	37.6	52.6
RAG-QwQ-32B	76.7	38.7	73.7
RAgent-Qwen2.5-32B	58.1	33.3	63.2
RAgent-QwQ-32B	76.7	46.2	68.4
Search-o1	77.9	47.3	78.9
Agentic Reasoning			
Ours	**88.1**	**58.3**	**79.6**

From Wu et al. 2025

Open questions (1/3)

Q1: Can retrieval help reasoning?

Method	Phy.	Chem.	Bio.
Direct Reasoning			
Qwen2.5-32B	57.0	33.3	52.6
Qwen2.5-Coder-32B	37.2	25.8	57.9
$\boxed{\text{QwQ-32B}}$	75.6	39.8	68.4
Qwen2.5-72B	57.0	37.6	68.4
Llama3.3-70B	54.7	31.2	52.6
GPT-4o†	59.5	40.2	61.6
o1-preview†	89.4	59.9	65.9
Retrieve/Search in Reasoning			
RAG-Qwen2.5-32B	57.0	37.6	52.6
RAG-QwQ-32B	76.7	38.7	73.7
RAgent-Qwen2.5-32B	58.1	33.3	63.2
RAgent-QwQ-32B	76.7	46.2	68.4
Search-o1	77.9	47.3	78.9
Agentic Reasoning			
Ours	**88.1**	**58.3**	**79.6**

From Wu et al. 2025

Open questions (1/3)

Q1: Can retrieval help reasoning?

Method	Phy.	Chem.	Bio.
Direct Reasoning			
Qwen2.5-32B	57.0	33.3	52.6
Qwen2.5-Coder-32B	37.2	25.8	57.9
QwQ-32B	75.6	39.8	68.4
Qwen2.5-72B	57.0	37.6	68.4
Llama3.3-70B	54.7	31.2	52.6
GPT-4o†	59.5	40.2	61.6
o1-preview†	89.4	59.9	65.9
Retrieve/Search in Reasoning			
RAG-Qwen2.5-32B	57.0	37.6	52.6
RAG-QwQ-32B	76.7	38.7	73.7
RAgent-Qwen2.5-32B	58.1	33.3	63.2
RAgent-QwQ-32B	76.7	46.2	68.4
Search-o1	77.9	47.3	78.9
Agentic Reasoning			
Ours	**88.1**	**58.3**	**79.6**

+15% improvements on GPQA
(on average)

From Wu et al. 2025

Open questions (1/3)

Q1: Can retrieval help reasoning?

Method	Phy.	Chem.	Bio.
Direct Reasoning			
Qwen2.5-32B	57.0	33.3	52.6
Qwen2.5-Coder-32B	37.2	25.8	57.9
QwQ-32B	75.6	39.8	68.4
Qwen2.5-72B	57.0	37.6	68.4
Llama3.3-70B	54.7	31.2	52.6
GPT-4o†	59.5	40.2	61.6
o1-preview†	89.4	59.9	65.9
Retrieve/Search in Reasoning			
RAG-Qwen2.5-32B	57.0	37.6	52.6
RAG-QwQ-32B	76.7	38.7	73.7
RAgent-Qwen2.5-32B	58.1	33.3	63.2
RAgent-QwQ-32B	76.7	46.2	68.4
Search-o1	77.9	47.3	78.9
Agentic Reasoning			
Ours	**88.1**	**58.3**	**79.6**

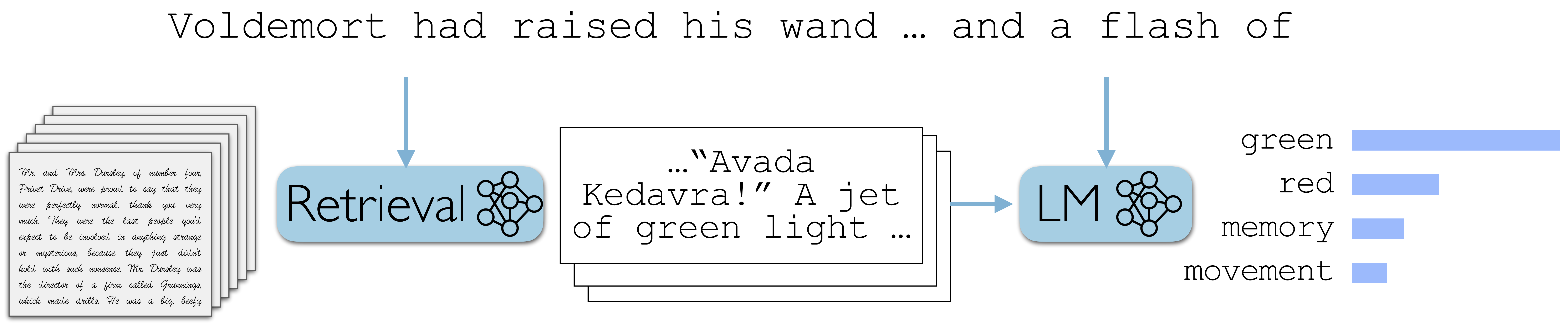
No work uses in-house retrieval
(Papers here used a search engine)

+15% improvements on GPQA
(on average)

From Wu et al. 2025

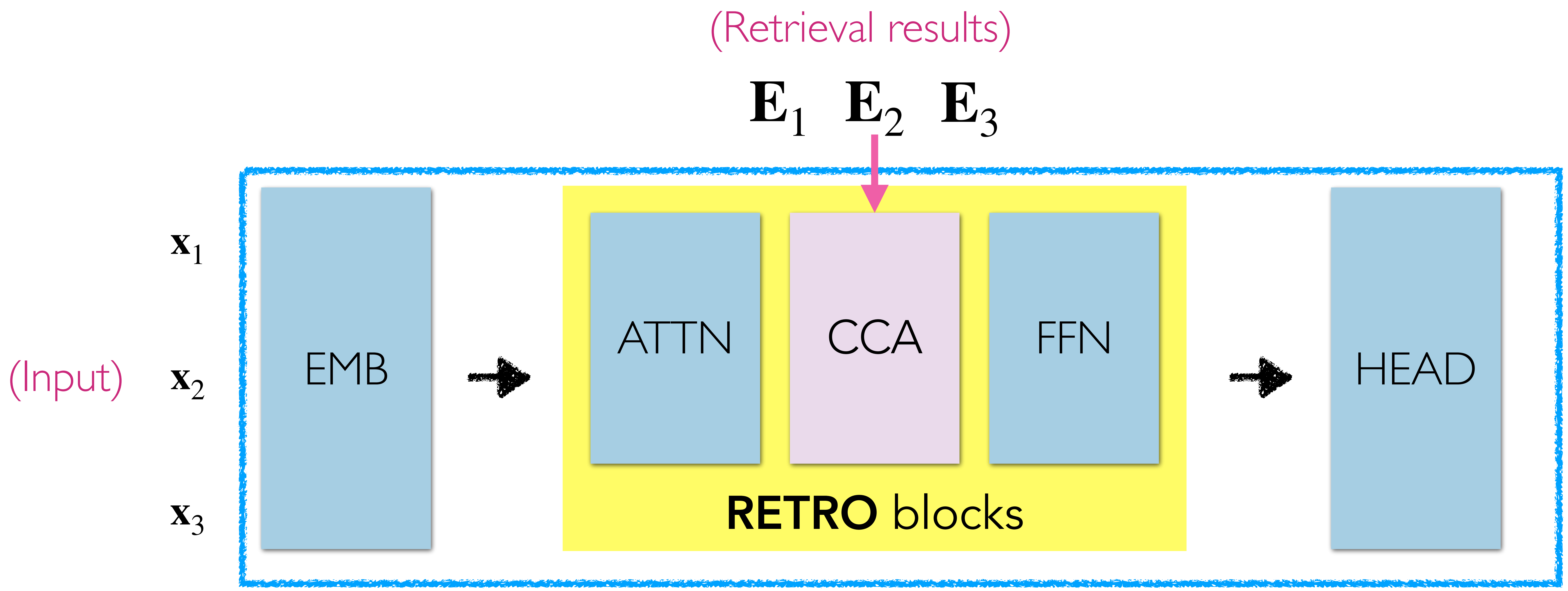
Open questions (2/3)

Q2: Alternative ways to incorporate datastore?



Open questions (2/3)

Q2: Alternative ways to incorporate datastore?

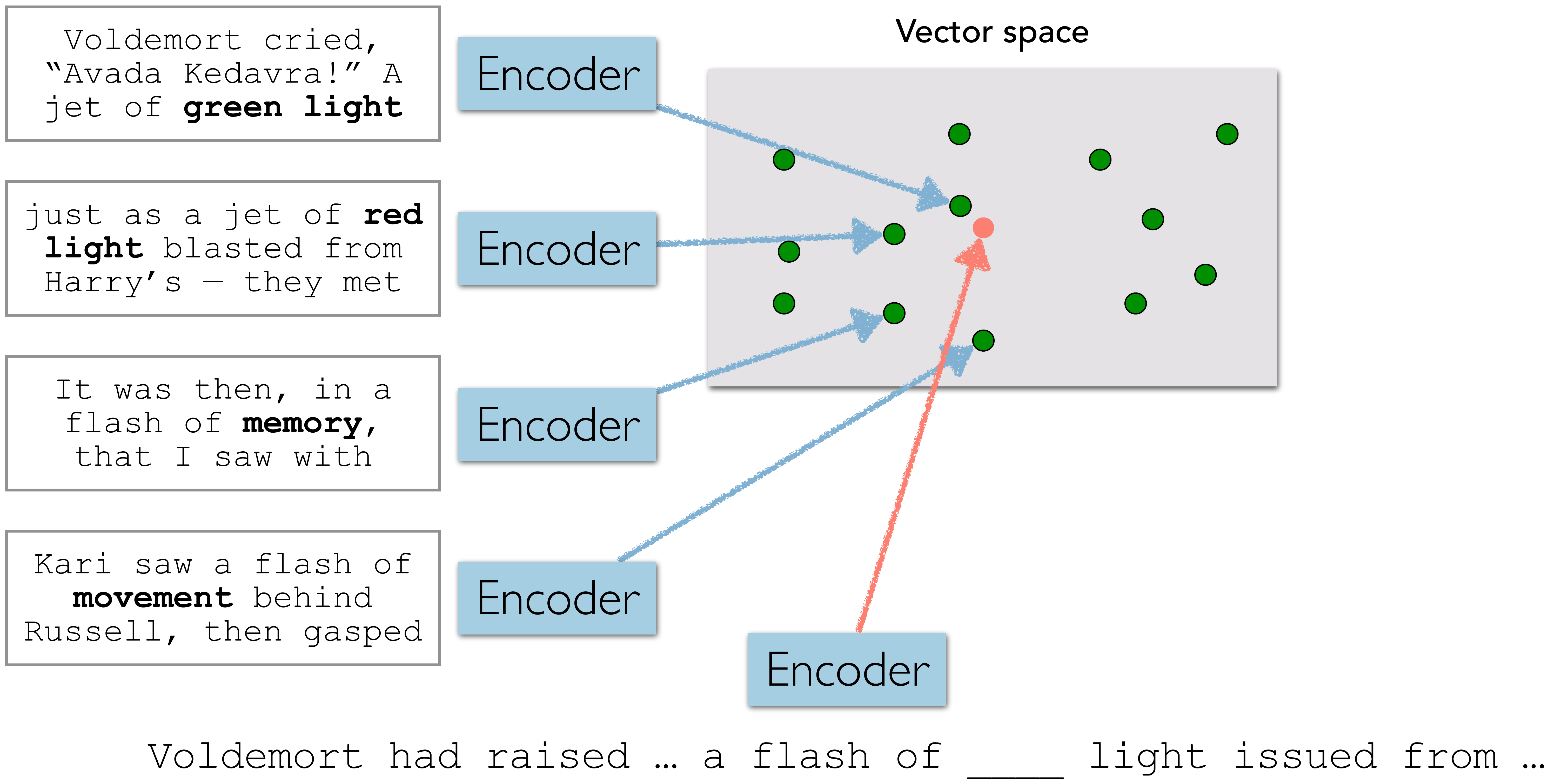


Chunked Cross Attention (CCA)

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

Open questions (2/3)

Q2: Alternative ways to incorporate datastore?



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

Open questions (2/3)

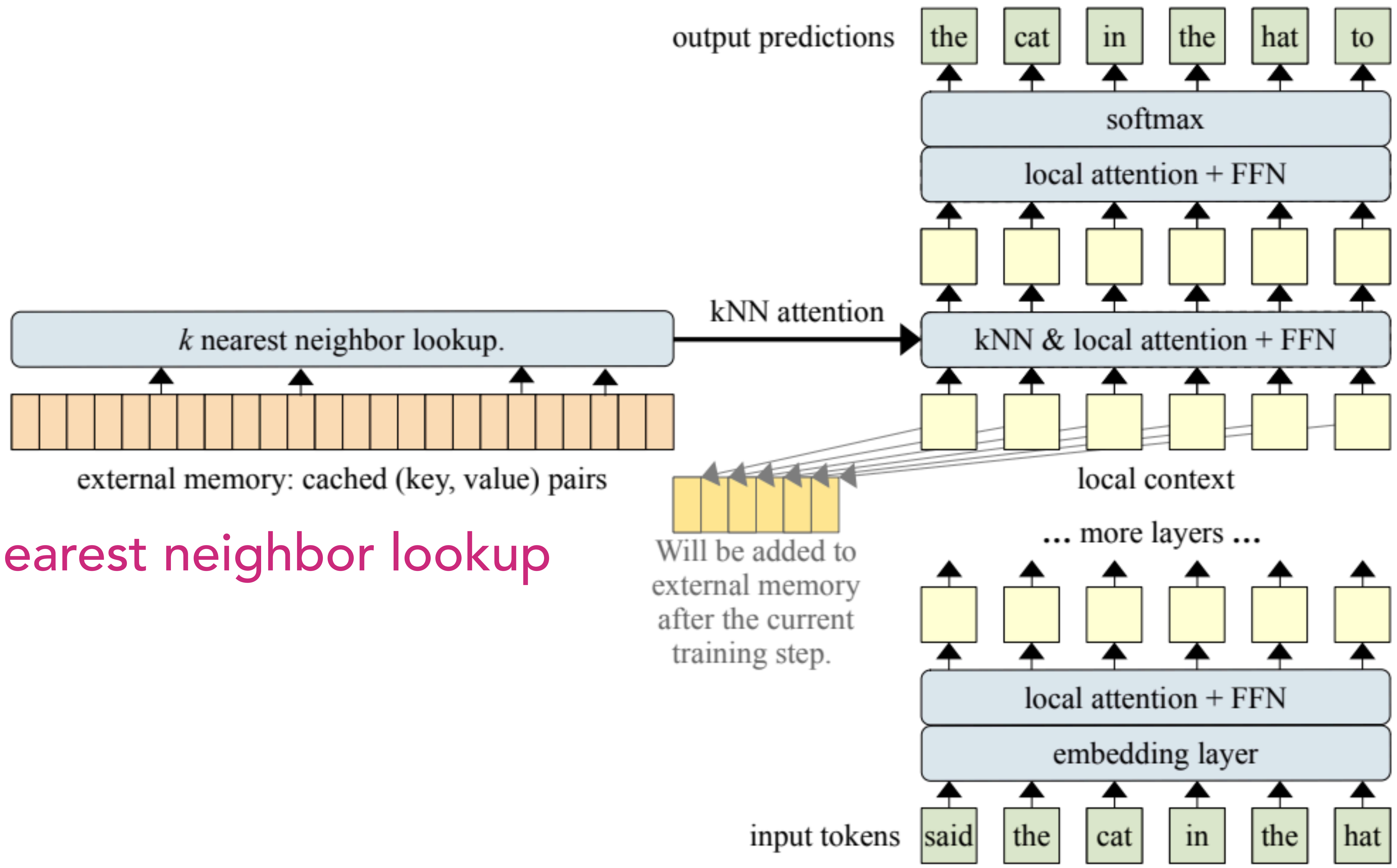
Q2: Alternative ways to incorporate datastore?

Build a million/billion-context LM that allows
"fitting the entire datastore as context"

Wu et al. "Memorizing Transformers"
Lu et al. "TurboRAG: Accelerating Retrieval-Augmented Generation with Precomputed KV Caches for Chunked Text"

Open questions (2/3)

Q2: Alternative ways to incorporate datastore?

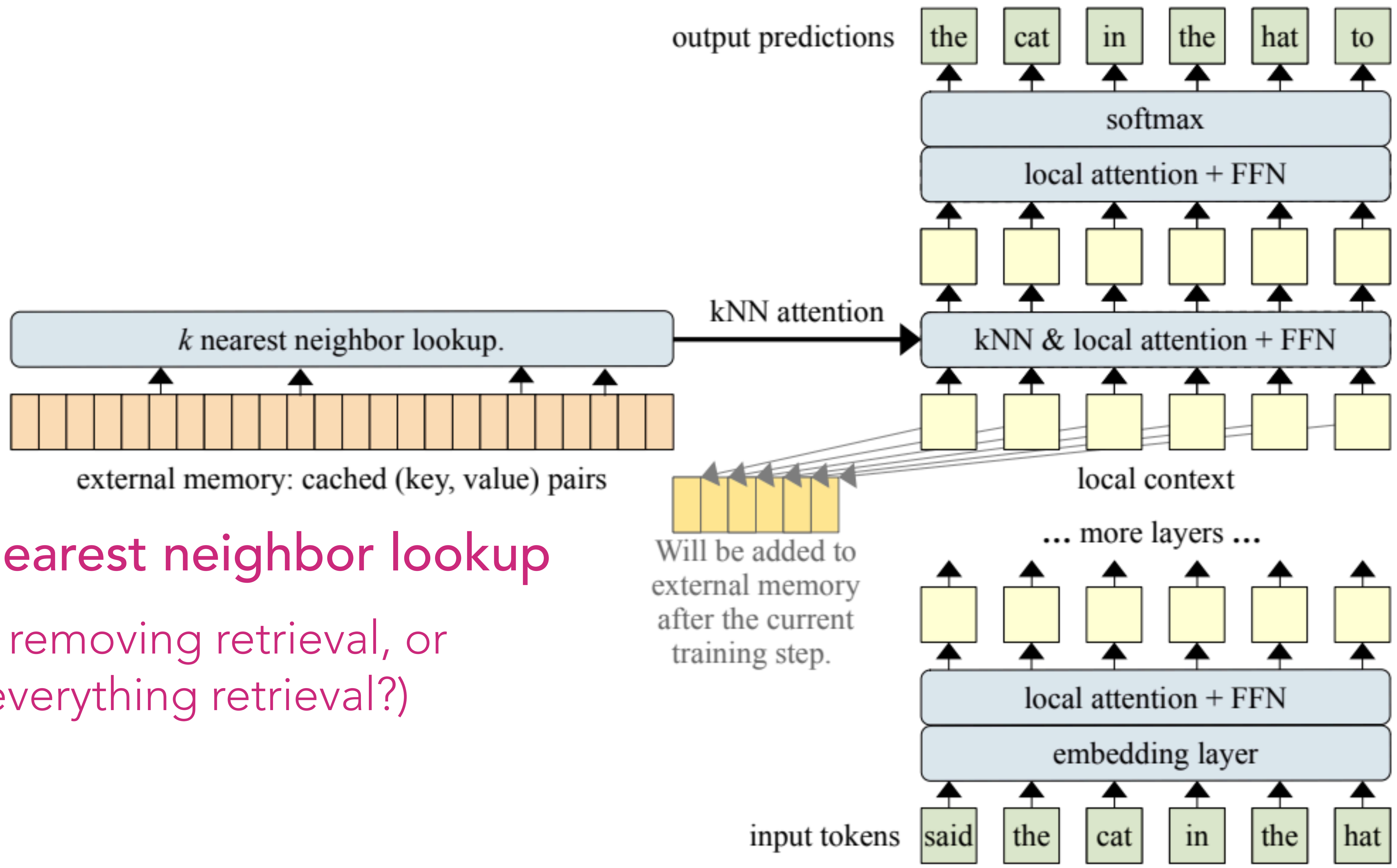


Done with *k* nearest neighbor lookup

Wu et al. "Memorizing Transformers"
Lu et al. "TurboRAG: Accelerating Retrieval-Augmented Generation with Precomputed KV Caches for Chunked Text"

Open questions (2/3)

Q2: Alternative ways to incorporate datastore?



Done with *k* nearest neighbor lookup

(Is it about removing retrieval, or making everything retrieval?)

Wu et al. "Memorizing Transformers"

Lu et al. "TurboRAG: Accelerating Retrieval-Augmented Generation with Precomputed KV Caches for Chunked Text"

Open questions (3/3)

Q3: How to optimize retrieval-based LMs (with systems point of view)?

Open questions (3/3)

Q3: How to optimize retrieval-based LMs (with systems point of view)?

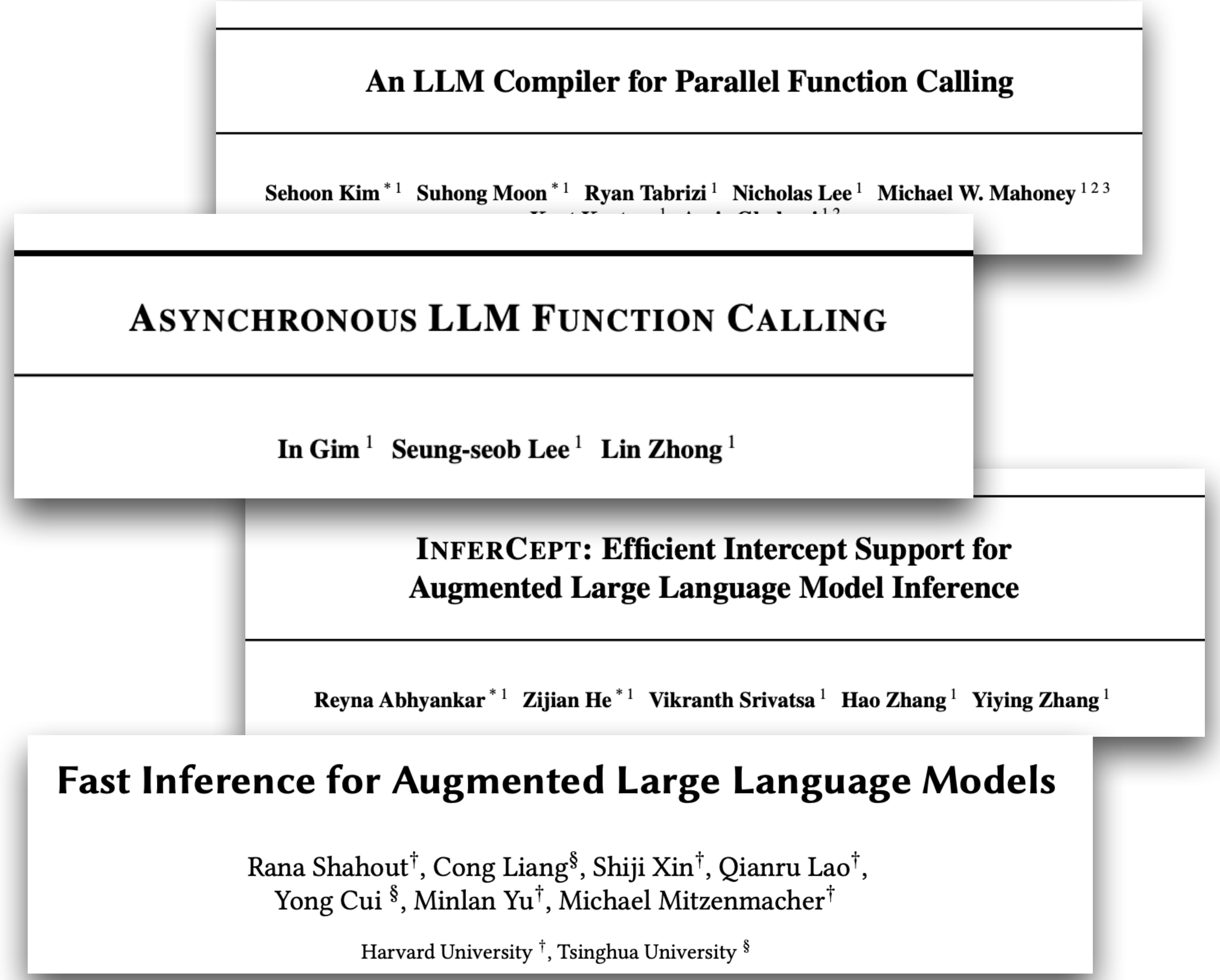
✓ Lots of work on optimizing retrieval (*k*NN)

✓ Lots of work on optimizing LLM inference

Open questions (3/3)

Q3: How to optimize retrieval-based LMs (with systems point of view)?

- ✓ Lots of work on optimizing retrieval (*k*NN)
- ✓ Lots of work on optimizing LLM inference
- ✓ (Relatively recent) Work on optimizing LLM inference with function calling, where retrieval is one of the functions



An LLM Compiler for Parallel Function Calling

Sehoon Kim [*1] Suhong Moon [*1] Ryan Tabrizi [1] Nicholas Lee [1] Michael W. Mahoney [1 2 3]

ASYNCHRONOUS LLM FUNCTION CALLING

In Gim [1] Seung-seob Lee [1] Lin Zhong [1]

INFERCEPT: Efficient Intercept Support for Augmented Large Language Model Inference

Reyna Abhyankar [*1] Zijian He [*1] Vikranth Srivatsa [1] Hao Zhang [1] Yiying Zhang [1]

Fast Inference for Augmented Large Language Models

Rana Shahout[†], Cong Liang[§], Shiji Xin[†], Qianru Lao[†], Yong Cui [§], Minlan Yu[†], Michael Mitzenmacher[†]

Harvard University [†], Tsinghua University [§]

Open questions (3/3)

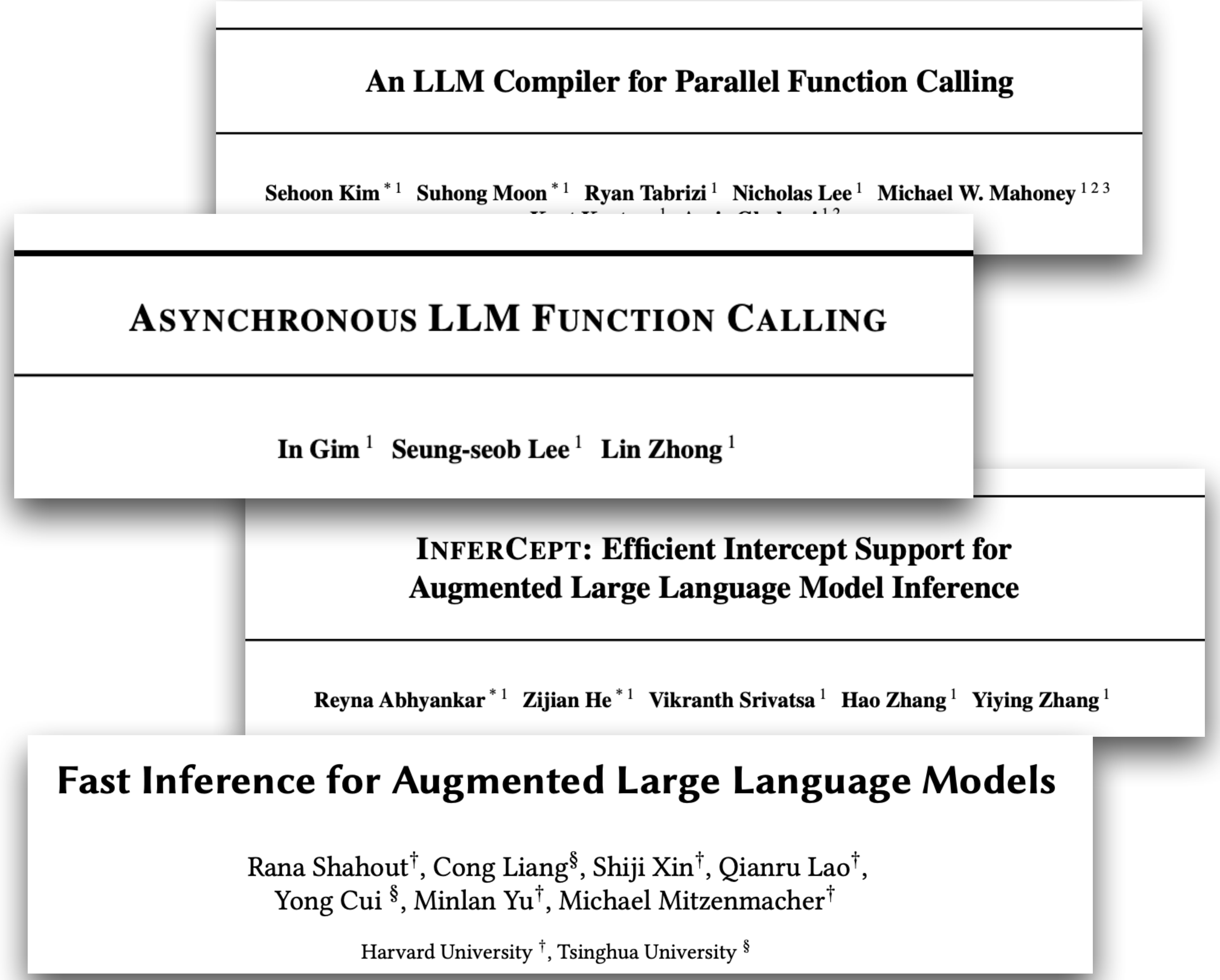
Q3: How to optimize retrieval-based LMs (with systems point of view)?

✓ Lots of work on optimizing retrieval (*k*NN)

✓ Lots of work on optimizing LLM inference

✓ (Relatively recent) Work on optimizing LLM inference with function calling, where retrieval is one of the functions

? Very little work that optimizes LMs with retrieval as a first class citizen (rather than a black box)



An LLM Compiler for Parallel Function Calling

Sehoon Kim [*1] Suhong Moon [*1] Ryan Tabrizi [1] Nicholas Lee [1] Michael W. Mahoney [1 2 3]

Asynchronous LLM Function Calling

In Gim [1] Seung-seob Lee [1] Lin Zhong [1]

InferCept: Efficient Intercept Support for Augmented Large Language Model Inference

Reyna Abhyankar [*1] Zijian He [*1] Vikranth Srivatsa [1] Hao Zhang [1] Yiying Zhang [1]

Fast Inference for Augmented Large Language Models

Rana Shahout[†], Cong Liang[§], Shiji Xin[†], Qianru Lao[†], Yong Cui [§], Minlan Yu[†], Michael Mitzenmacher[†]

Harvard University [†], Tsinghua University [§]

Thank you for listening!

sewonmin.com

sewonm@berkeley.edu

Please leave feedback at: tinyurl.com/sewonm-talk