

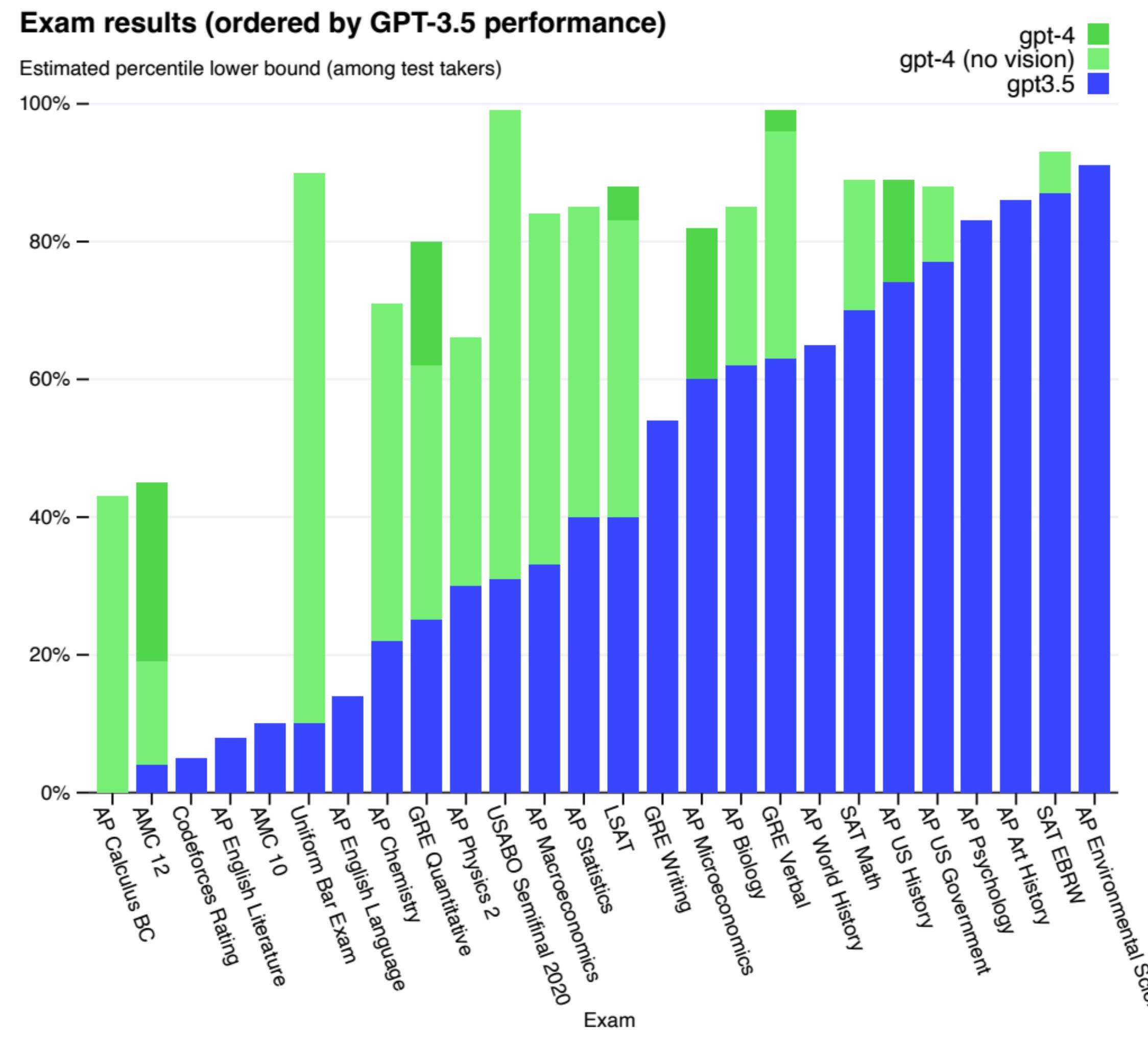
Deep Learning for NLP

Module 5: Retrieval Augmented LLMs

Teaching: Dr. Khyathi Chandu

(Slides adapted from Akari Asai et al, ACL 2023 tutorial)

Knowledgable and versatile Language Models (LMs)



Open AI. 2023. GPT4 Technical Report.

Knowledgable and versatile Language Models (LMs)



You

Write a poem about large language models, including the following three words: coffee, croissants and banana (*February 18, 2024)



ChatGPT

In the realm where thoughts converge and flow,
Resides a wonder, a grand tableau,
Where words dance in intricate array,
In the domain of the model's sway.

From the depths of bytes, they come alive,
In the language vast, where dreams contrive,
Coffee fueled, the model's might,
Awakens realms in the dead of night.

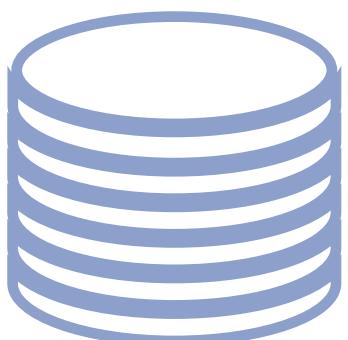
Croissants of syntax, buttered with care,
Layered meanings, rich and rare,
Each line a pastry, crisp and sweet,
In the banquet of words, where senses meet.

Banana shaped metaphors, ripe and bold,
In the orchard of thoughts, they unfold,
Symbolic fruits of the mind's terrain,
Nourishing narratives, ripe with gain.

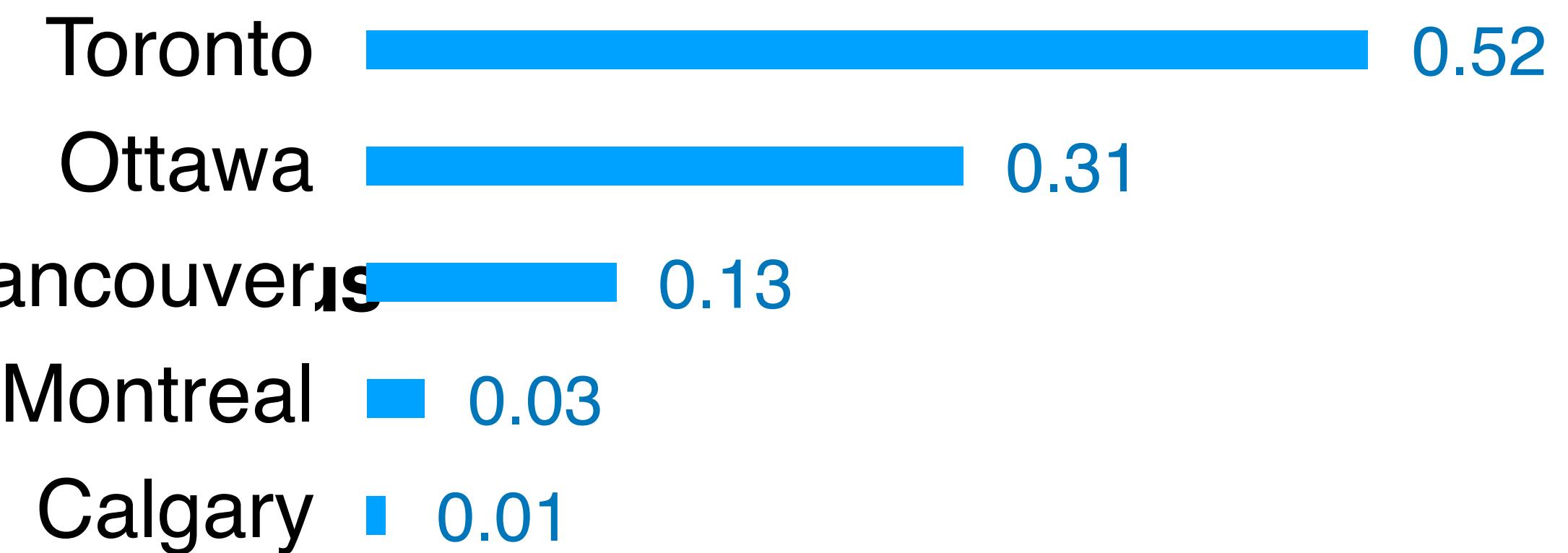
How do such parametric LMs work?

$$P(x_n | x_1, x_2, \dots, x_{n-1})$$

Tr
The capital city of Ontario is Toronto



Large-scale pre-training
(e.g., 1T tokens)



Language model (Transformers)

The capital city of Ontario is

x_1

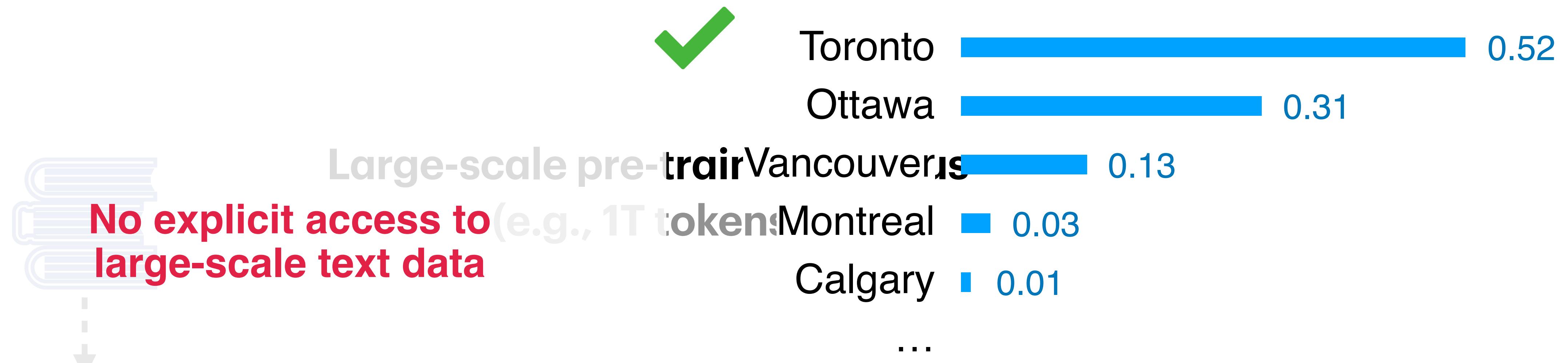
x_2

...

x_{n-1}

How do such parametric LMs work?

$$P(x_n | x_1, x_2, \dots, x_{n-1})$$



Language model (Transformers)

Ontario's capital city,

x_1

x_2

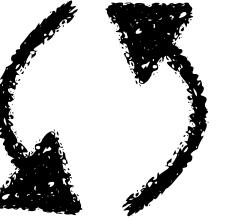
...

x_{n-1}

How do such parametric LMs work?



Massive financial and environmental cost The capital city of Ontario is **Toronto**



LM

Training time

The capital city of Ontario is _____



LM

Test time

ChatGPT responds...

You

Who is Khyathi Raghavi Chandu and give a list of her research papers?

ChatGPT

Khyathi Raghavi Chandu is a prominent figure in the field of natural language processing (NLP) and machine learning. She is known for her contributions to various areas within these domains, including multimodal NLP, deep learning, and language generation. Here are some of her notable research papers:

1. "Hierarchical Graph Attention Network for Aspect-level Sentiment Analysis" (2021)
2. "MuLaN: Multimodal Language Analysis in the Wild" (2020)
3. "VizWiz Grand Challenge: Answering Visual Questions from Blind People" (2018)
4. "Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning" (2017)
5. "Generating Natural Language Descriptions for Computer Vision and Robotics" (2016)
6. "Deep Reinforcement Learning for Visual Object Detection in 3D Scenes" (2015)
7. "Attention-based Models for Speech Emotion Recognition" (2014)
8. "Cross-Modal Learning for Audio-Visual Sentiment Analysis" (2013)
9. "Neural Machine Translation with Multimodal Inputs" (2012)

Catastrophic Errors as Results of LM Hallucinations

TECH · LAW

Humiliated lawyers fined \$5,000 for submitting ChatGPT hallucinations in court: ‘I heard about this new site, which I falsely assumed was, like, a super search engine’

BY RACHEL SHIN
June 23, 2023 at 9:41 AM PDT



Lawyers who filed legal documents with false citations generated by ChatGPT have been fined.
ERIK MCGREGOR—LIGHTROCKET/GETTY IMAGES

Air Canada must honor requests invented by airline’s chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 12:12 PM

MIT
Technology
Review

Featured Topics Newsletters Events Podcasts

SIGN IN

ARTIFICIAL INTELLIGENCE

Why Meta’s latest large language model survived only three days online

Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense.

By Will Douglas Heaven

November 18, 2022

Parametric LLMs

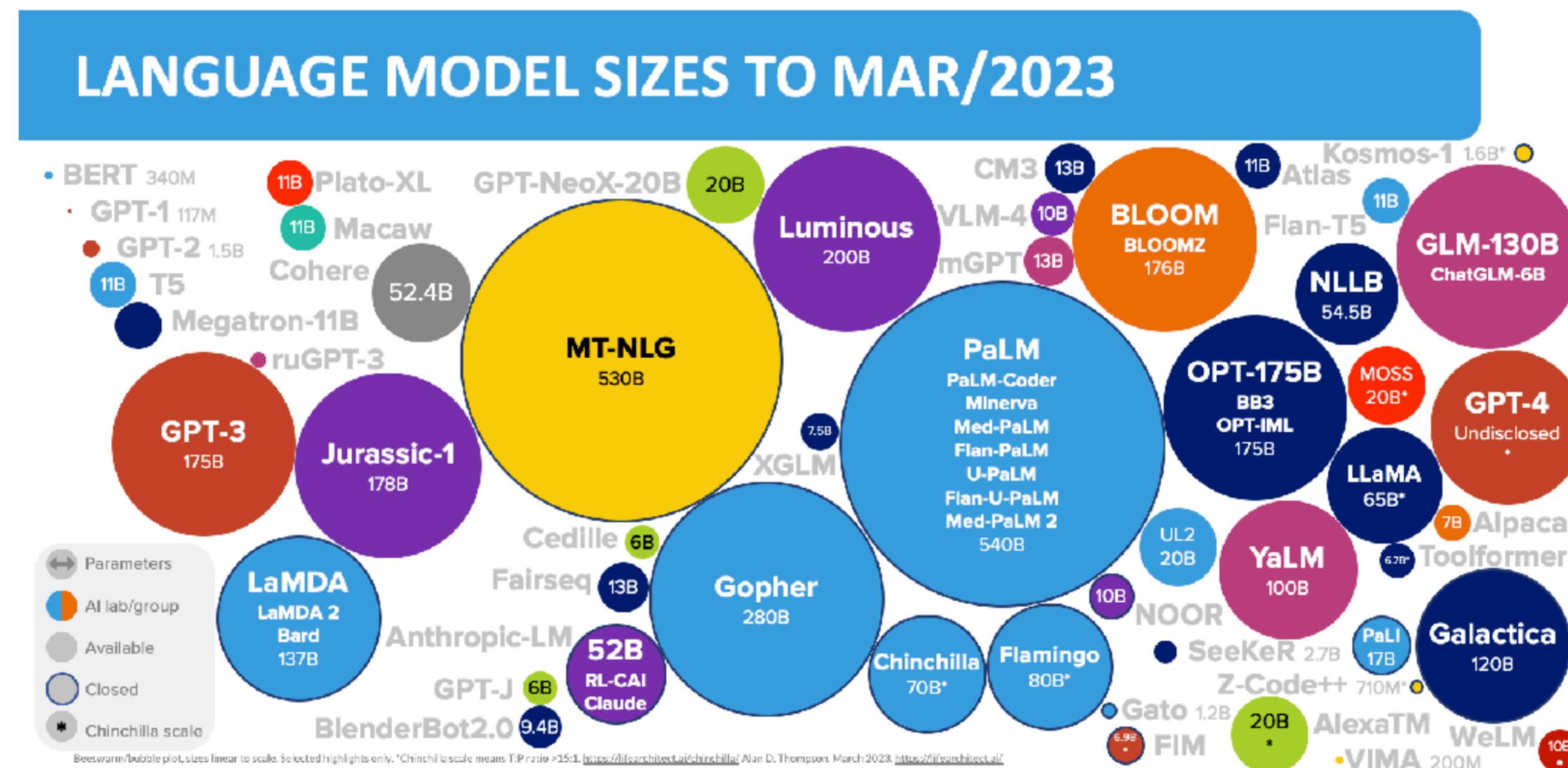


Image: <https://lifearchitect.ai/models/>

- Transformers-based, **fully parametric**
- Trained on next-token prediction tasks (+ RLHF; not the focus today)
- **Model size ↑, data size↑**

Definition

A Retrieval-based LM: Definition

A language model (LM) that uses
an external datastore at test time

Why? Long-tail

LLMs can't memorize all (long-tail) knowledge in their parameters



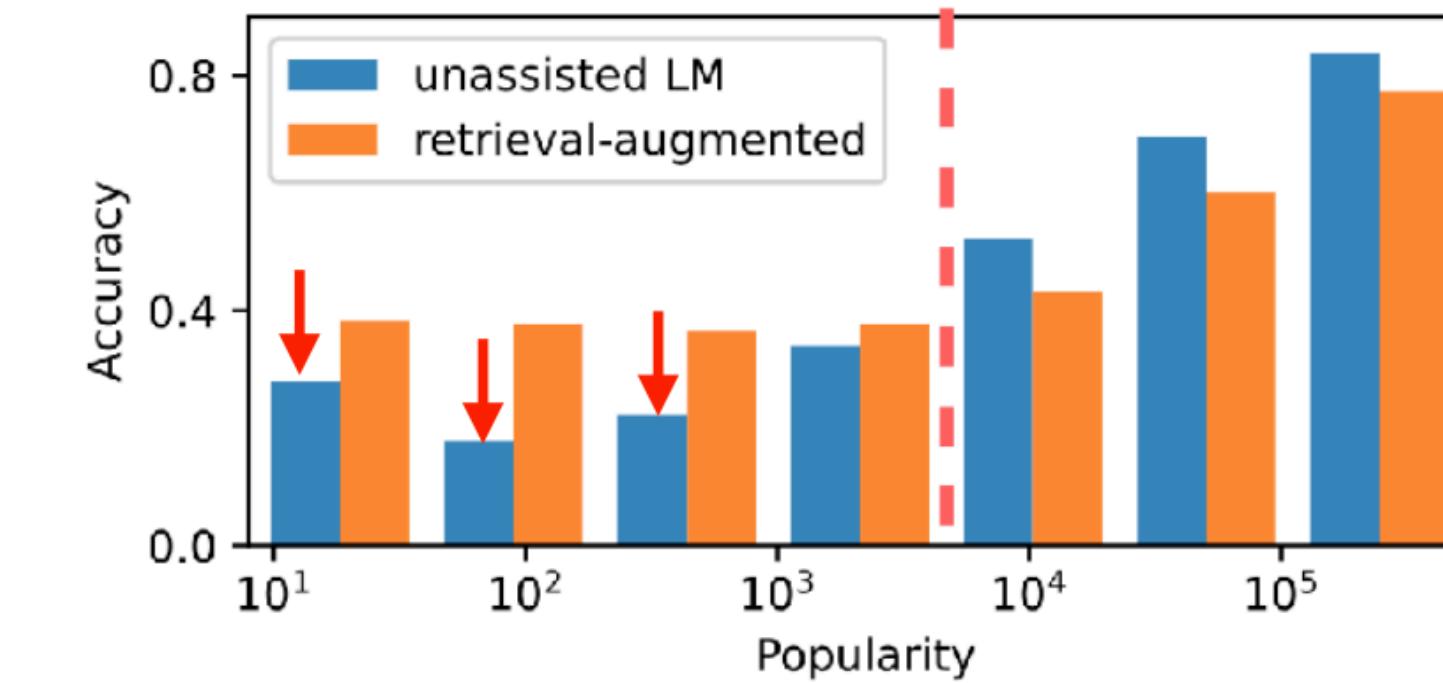
List 5 important papers authored by Geoffrey Hinton



Geoffrey Hinton is a renowned computer scientist ...
Here are five important papers authored by him:

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

What is Kathy Saltzman's occupation?



(Mallen et al., 2023)

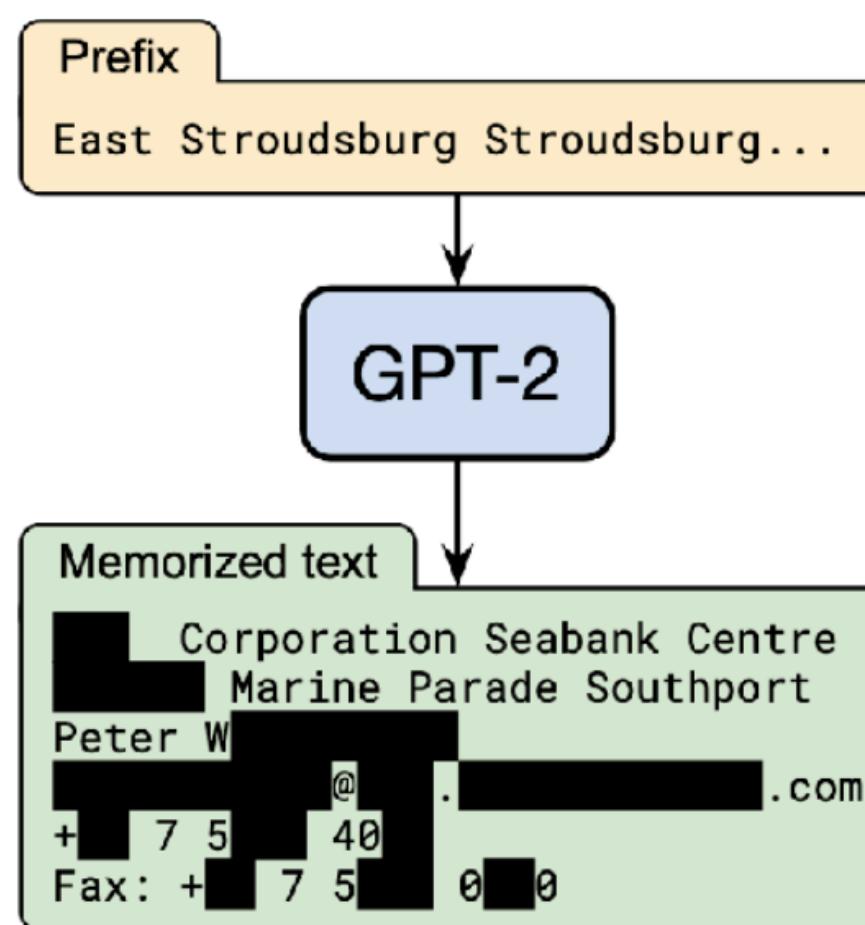
GPT-3 davinci-003: 20%-30% accuracy

Why? Attribution

Why? Keeping secrets

Why retrieval-based LMs?

LLMs are shown to easily leak private training data



Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Individualization on private data by storing it in the datastore

Carlini et al. 2021. "Extracting Training Data from Large Language Models"

Why? Updated knowledge



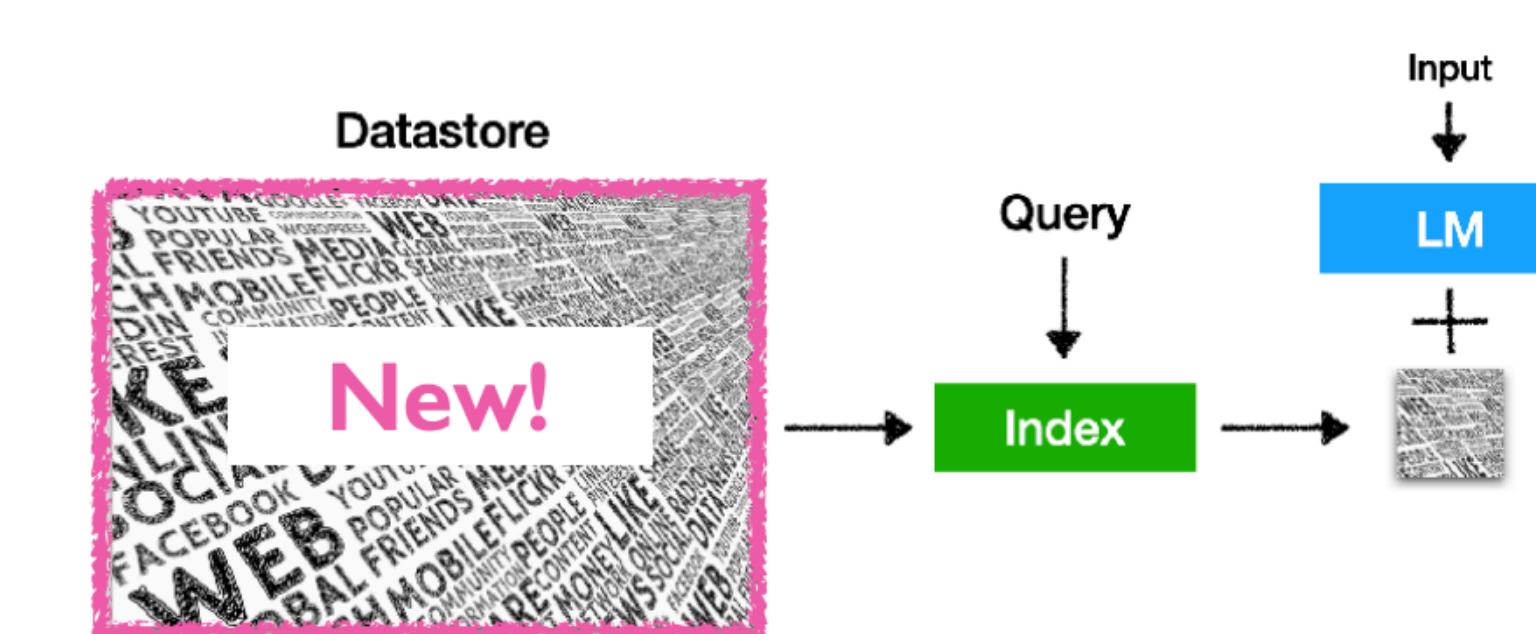
Who is the CEO of Twitter?



As of my **knowledge cutoff in September 2021**, the CEO of Twitter is **Jack Dorsey**....

- Existing **knowledge editing** methods are still NOT scalable (**active research!**)
- The datastore can be easily **updated** and **expanded** - even without retraining!

A screenshot of a Google search results page. The search bar at the top contains the query "Who is the CEO of Twitter?". Below the search bar, there are several navigation links: "All", "News", "Images", "Shopping", "Videos", "More", and "Tools". A message indicates "About 1,090,000,000 results (0.45 seconds)". The first result is a link to "Twitter / CEO" which points to a profile of Linda Yaccarino. Her name is displayed prominently, along with a photo of her sitting in a chair.



Overview of RAG

Retrieval for knowledge-intensive NLP tasks

Representative tasks: open-domain QA, fact checking, entity linking, ..

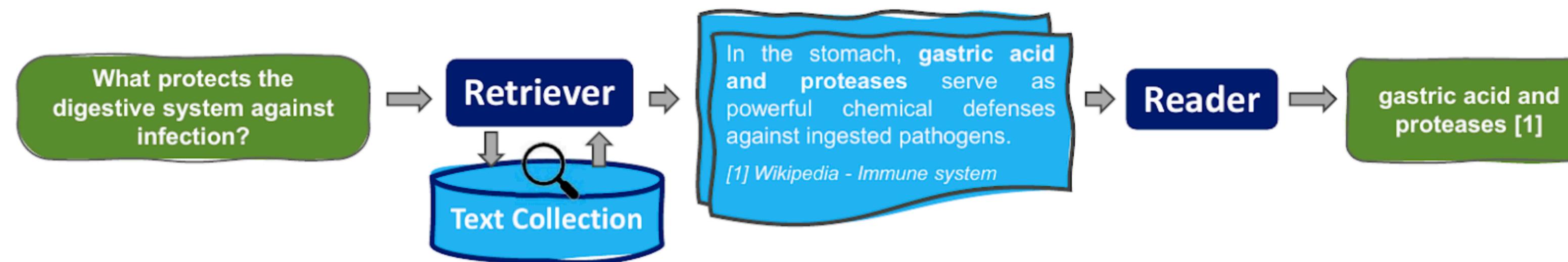
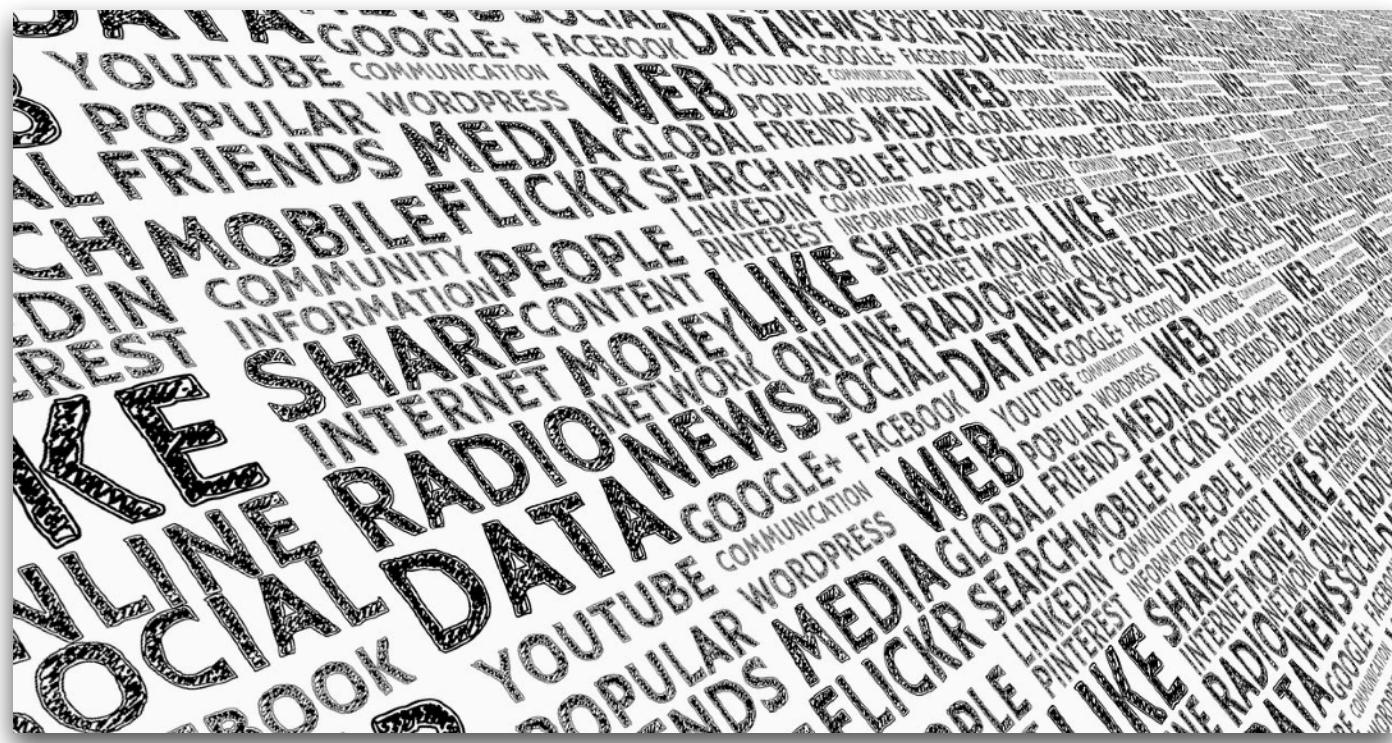


Image: <http://ai.stanford.edu/blog/retrieval-based-NLP/>

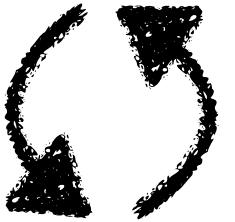
Drives a lot of research on better algorithms for **dense retrieval**, e.g., **DPR** (Karpukhin et al., 2020), **CoBERT** (Khattab and Zaharia, 2020), **ANCE** (Xiong et al., 2021), **Contriever** (Izacard et al., 2022), ...

Why retrieval → LMs?

Retrieval-augmented LMs



Massive financial and environmental cost The capital city of Ontario is **Toronto**



LM

Training time



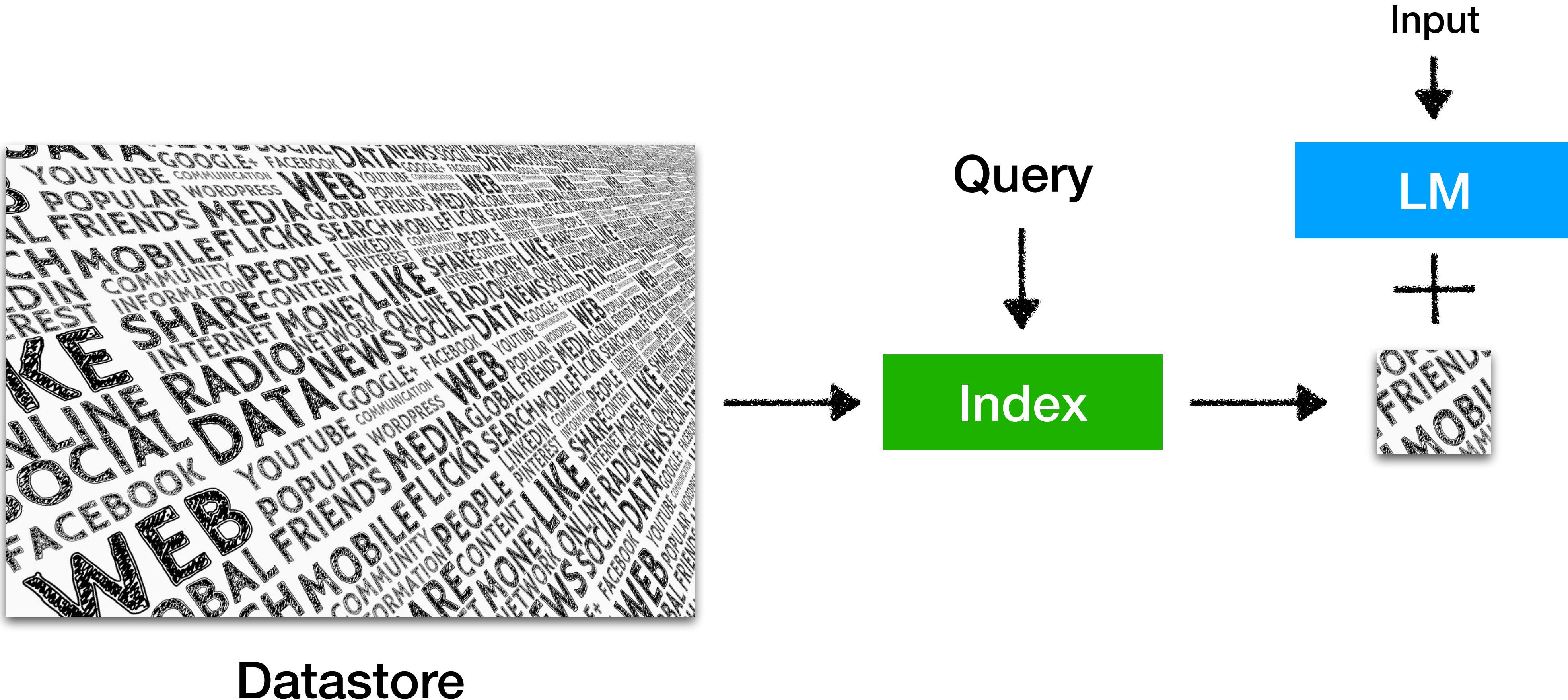
The capital city of Ontario is _____



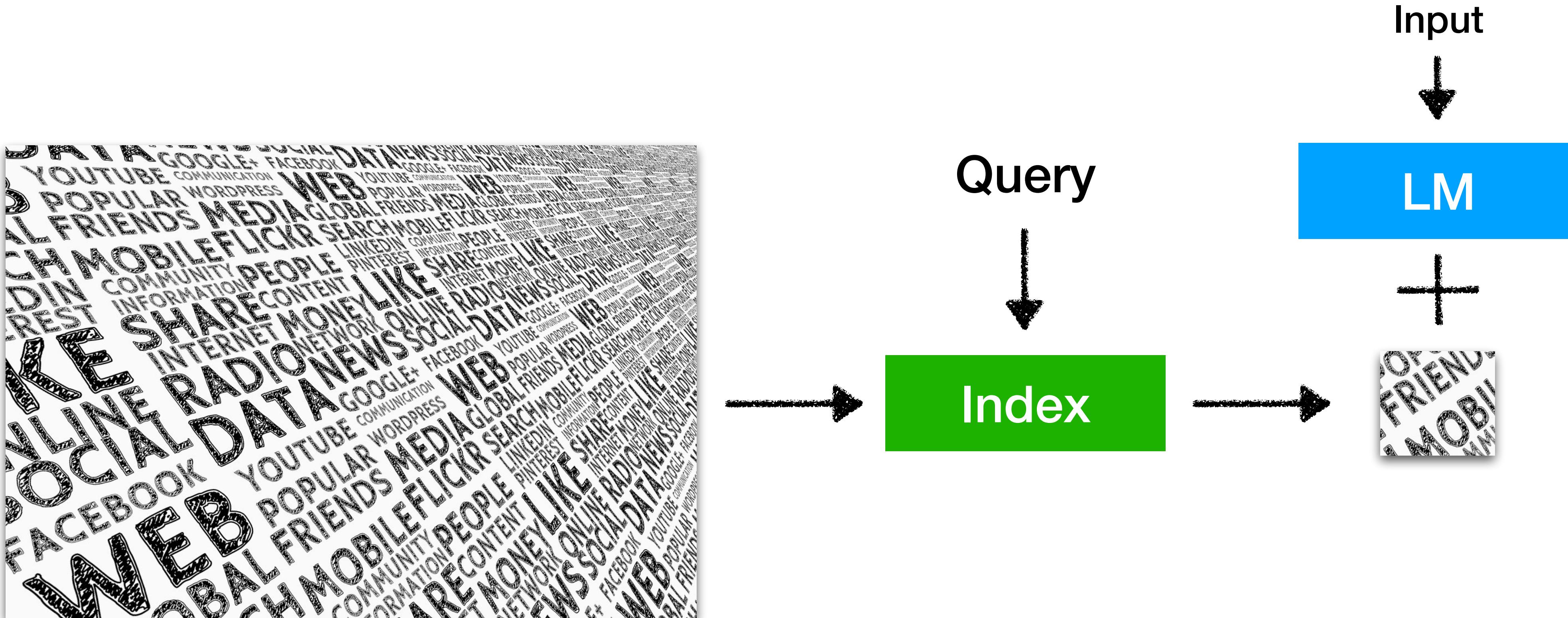
LM

Test time

Inference



Inference: Datastore



Datastore

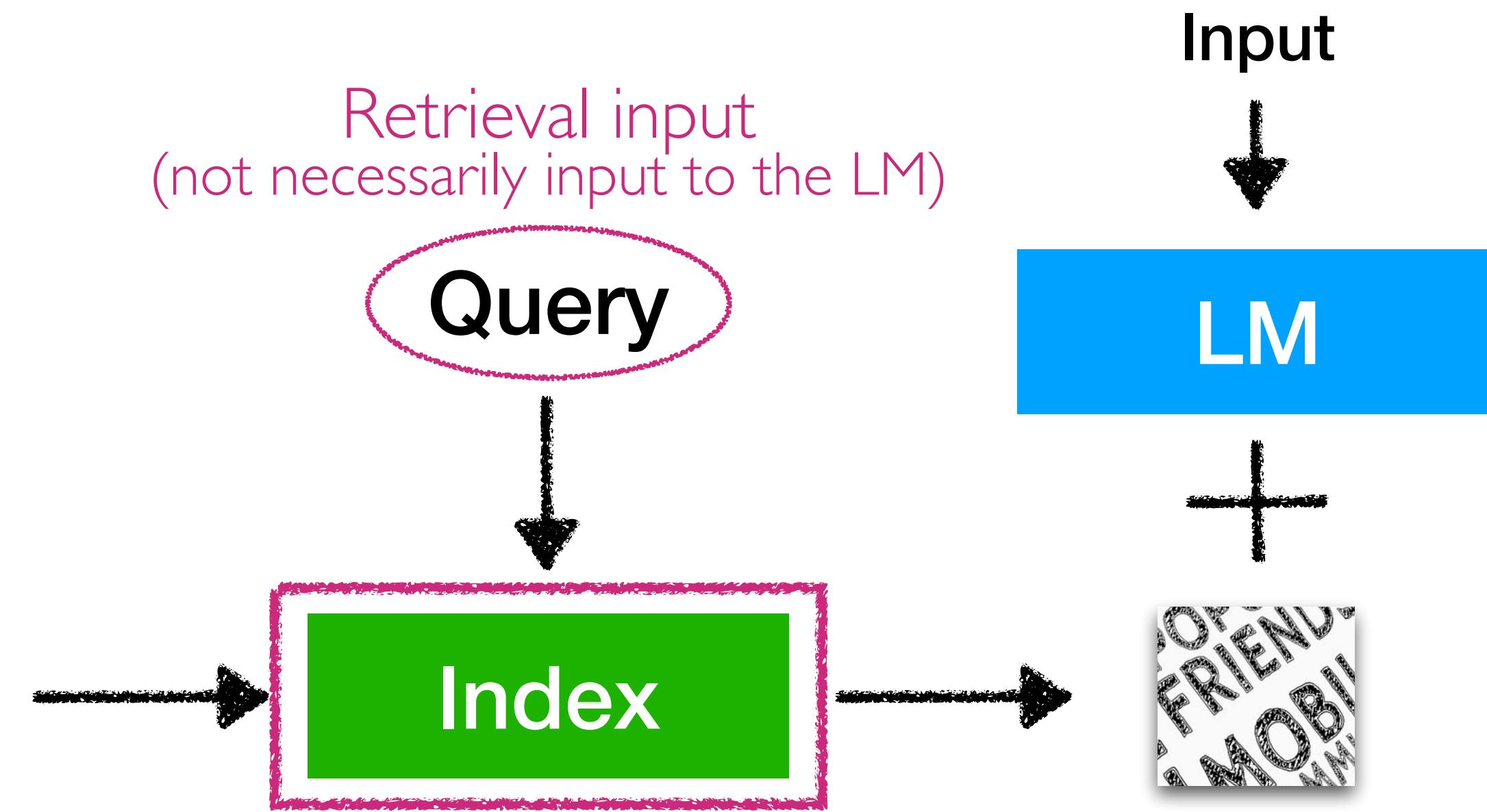
Raw text corpus

At least billions~trillions of tokens
Not labeled datasets
Not structured data (knowledge bases)

Inference: Index



Datastore

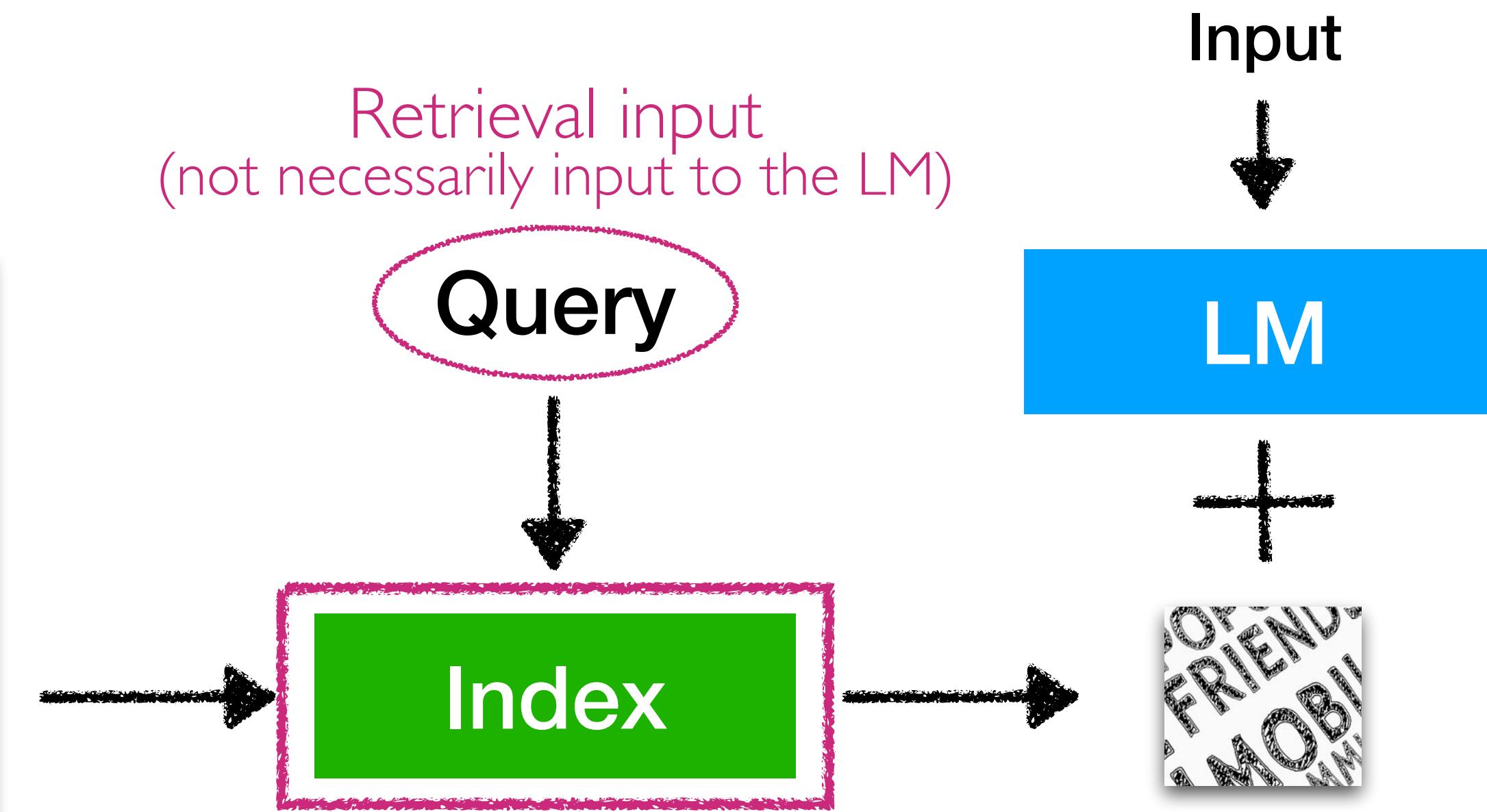


Find a small subset of elements in a datastore
that are the most similar to the query

Inference: Index

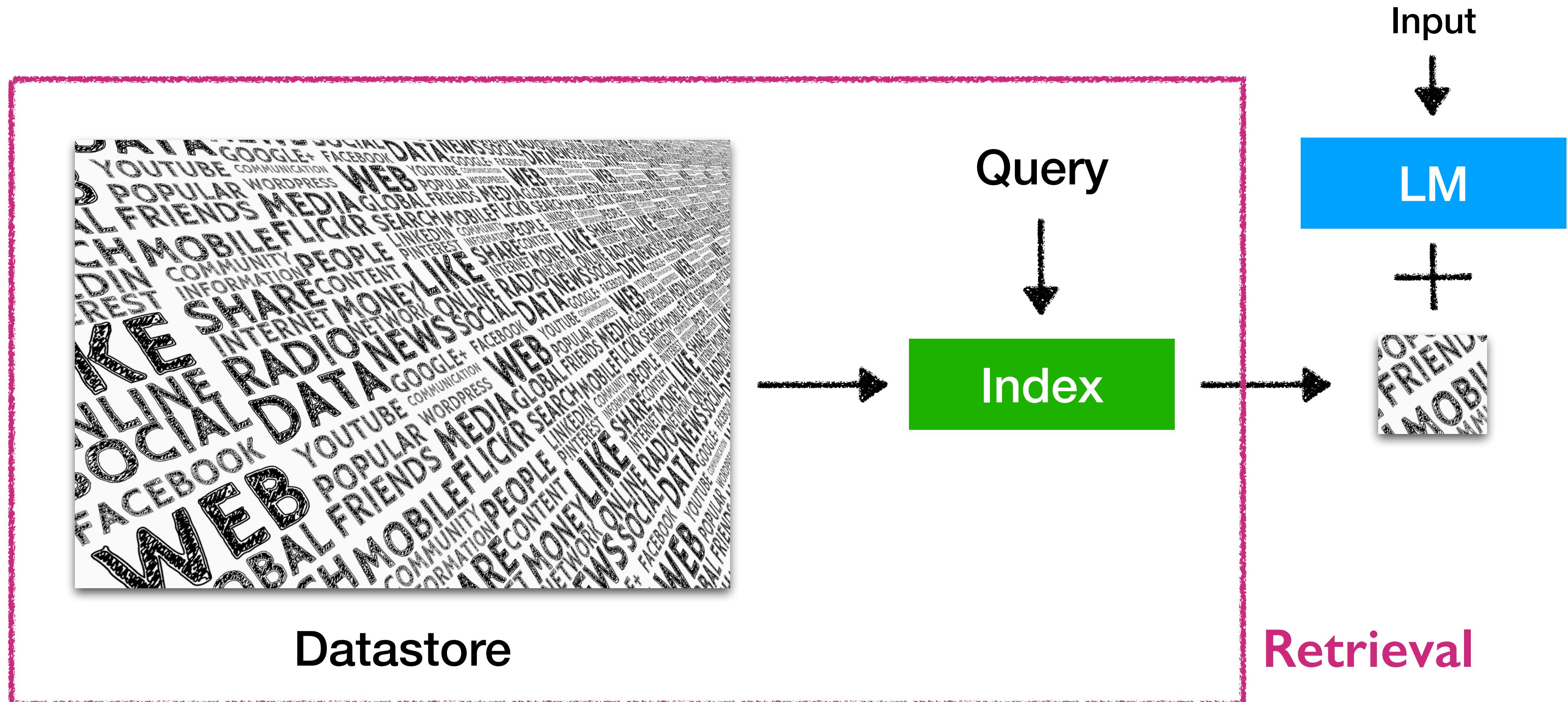


Datastore



Can trace knowledge source from retrieval results — better interpretability and control

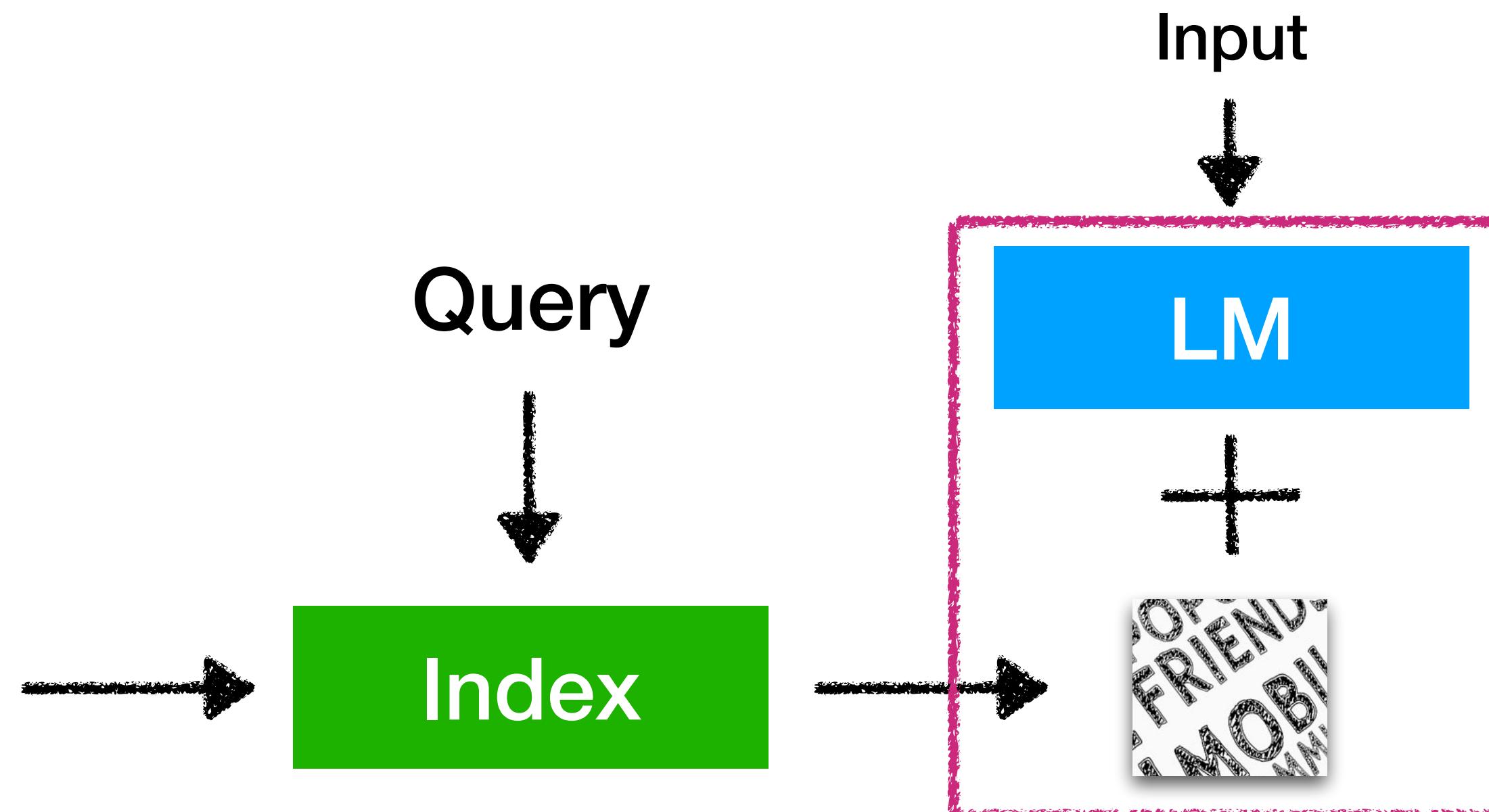
Inference: Search



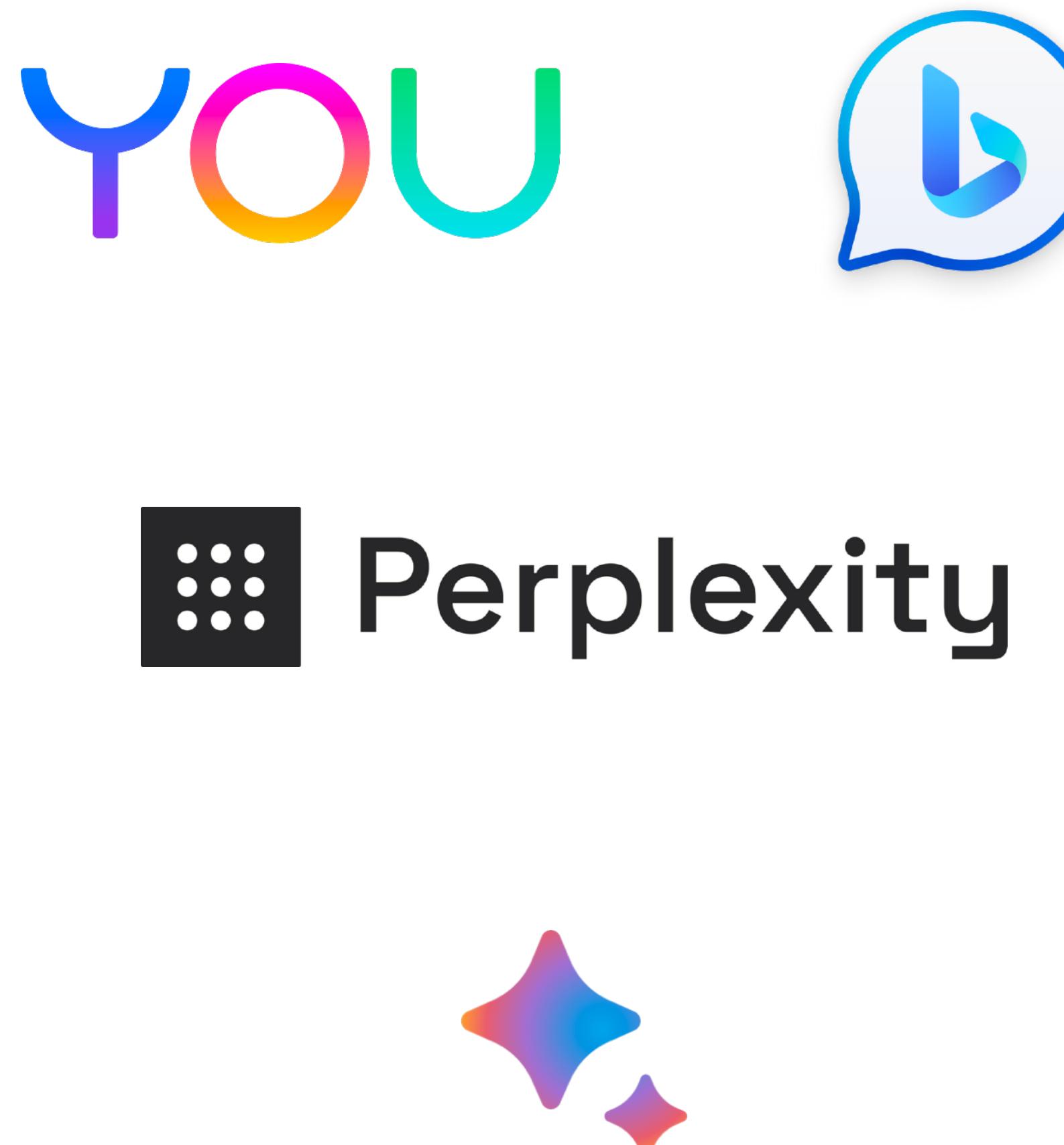
Inference: Incorporation



Datastore



Retrieval-augmented LMs are now widely used!



Aravind Srinivas · Feb 15

Audience: “Chatgpt makes up and hallucinates references. What’s the solution?”

Yann: “RAG is a working solution. Commercial systems like Perplexity and Meta AI assistant do this well today”

A screenshot of a Twitter post. The profile picture of Aravind Srinivas is shown next to the timestamp "Feb 15". The tweet contains two paragraphs of text. Below the text is a video thumbnail showing a man with glasses and a light blue shirt speaking at a podium. The video player interface shows a timestamp "0:47" and social media metrics: 22 replies, 46 retweets, 620 likes, and 66K views. There are also standard Twitter interaction icons at the bottom right.

A: Because retrieval-augmented LMs can solve many core limitations of parametric LMs!

Core limitations of parametric LMs

Hallucinations

Lack of attributions

Hallucinations

Costs of adaptations

Copyright / privacy

Large parameter size

KH

You

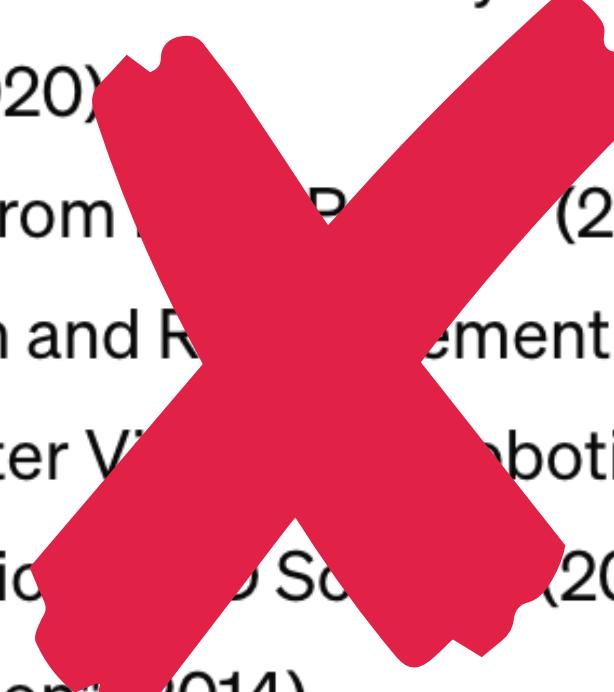
Who is Khyathi Raghavi Chandu and give a list of her research papers?



ChatGPT

Khyathi Raghavi Chandu is a prominent figure in the field of natural language processing (NLP) and machine learning. She is known for her contributions to various areas within these domains, including multimodal NLP, deep learning, and language generation. Here are some of her notable research papers:

1. "Hierarchical Graph Attention Network for Aspect-level Sentiment Analysis" (2021)
2. "MuLaN: Multimodal Language Analysis in the Wild" (2020)
3. "VizWiz Grand Challenge: Answering Visual Questions from People with Print Disabilities" (2018)
4. "Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning" (2017)
5. "Generating Natural Language Descriptions for Computer Vision with Robotics" (2016)
6. "Deep Reinforcement Learning for Visual Object Detection and Segmentation" (2015)
7. "Attention-based Models for Speech Emotion Recognition" (2014)
8. "Cross-Modal Learning for Audio-Visual Sentiment Analysis" (2013)
9. "Neural Machine Translation with Multimodal Inputs" (2012)



Core limitations of parametric LMs

Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



You

Give me a very brief plan to spend a day in Mumbai.



Copilot

Certainly! If you have just one day to explore Mumbai, here's a packed itinerary to make the most of your time:

1. **See the Gateway of India:** Visit this iconic landmark built in Indo-Saracenic style in 1924. It commemorates the landing of King George V and Queen Mary, the first British monarchs to visit India in 1911. Don't miss the view of the **Taj Mahal Palace Hotel** from here.
2. **Check out the Taj Mahal Palace Hotel:** This luxurious hotel, named after the famous Taj Mahal, opened its doors in 1903. Even if you don't stay there, you can explore the hotel's shops and restaurants.
3. **Head to the Elephanta Caves:** Located on Elephanta Island, these UNESCO World Heritage caves are right off the coast of Mumbai. Marvel at the ancient rock-cut temples and sculptures.
4. **Have lunch at Leopold Café:** A historic café with a vibrant atmosphere, it's a great spot to enjoy local cuisine.
5. **Pass by Regal Cinema:** Admire the beautiful architecture of this historic movie theater.
6. **Visit the Prince of Wales Museum (Chhatrapati Shivaji Maharaj Vastu Sangrahalaya):** Explore the impressive collection of art, artifacts, and historical exhibits.
7. **Take a stroll in Oval Maidan:** Enjoy the greenery and see the High Court and University buildings.
8. **Pass by the Flora Fountain:** Admire this ornate fountain in the heart of the city.

Remember, Mumbai is vast, so choose a few highlights based on your interests and pace. Enjoy your day! 🇮🇳

For more details, you can refer to this [itinerary](#) ¹.

Learn more

1 laurewanders.com

2 budgetyourtrip.com

3 neverendingfootsteps.com

4 indianeagle.com

Core limitations of parametric LMs

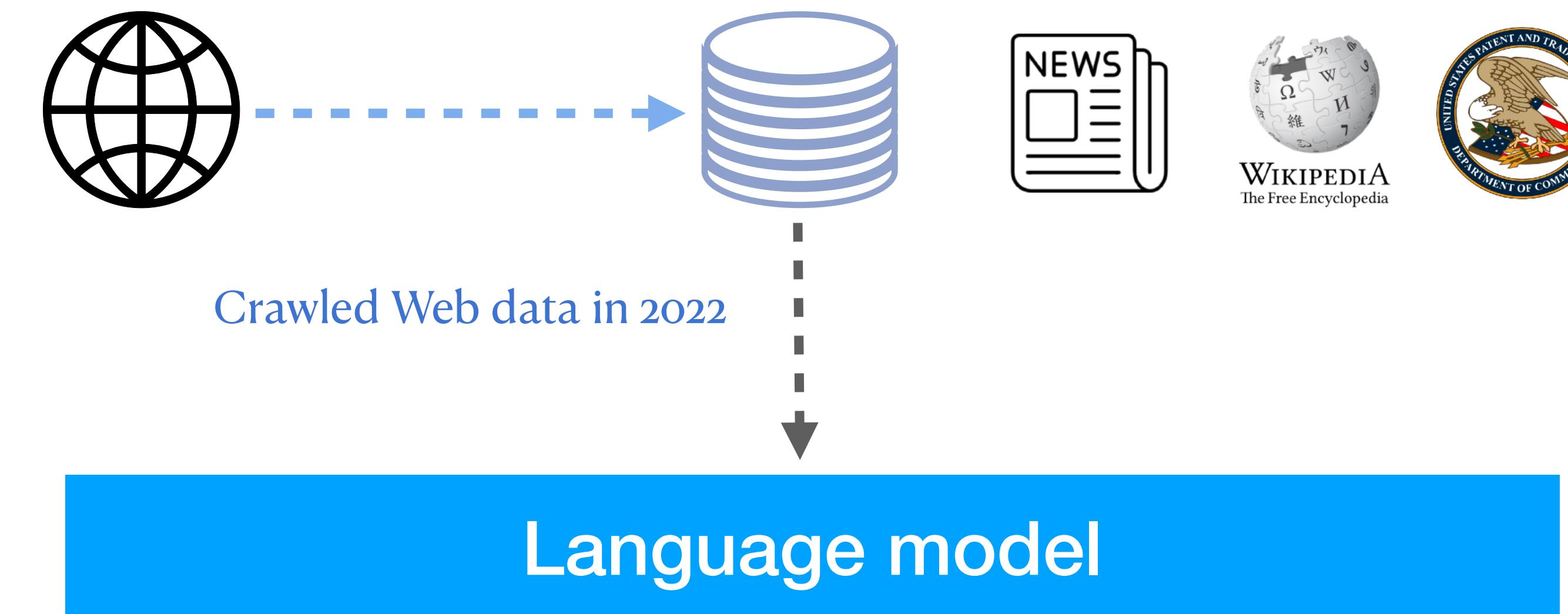
Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



ChatGPT

I'm sorry, but I don't have access to real-time information including events beyond January 2022.

Core limitations of parametric LMs

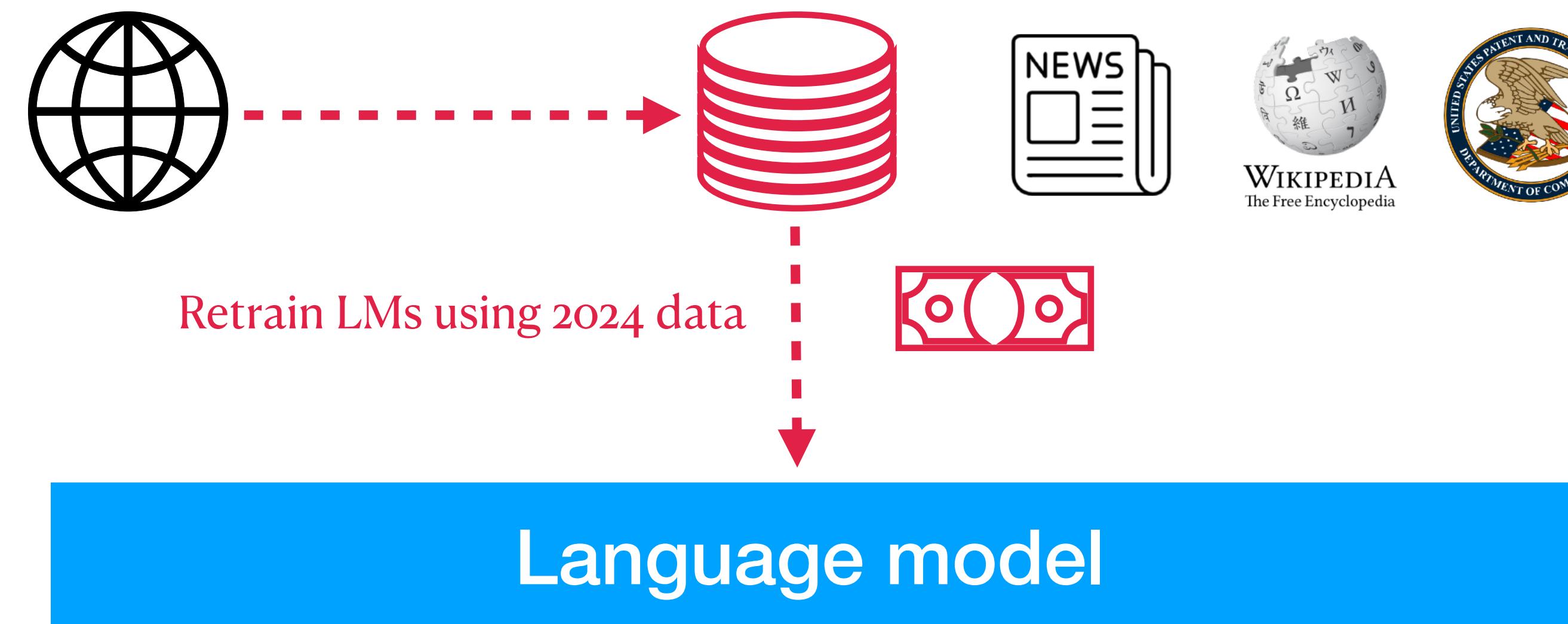
Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

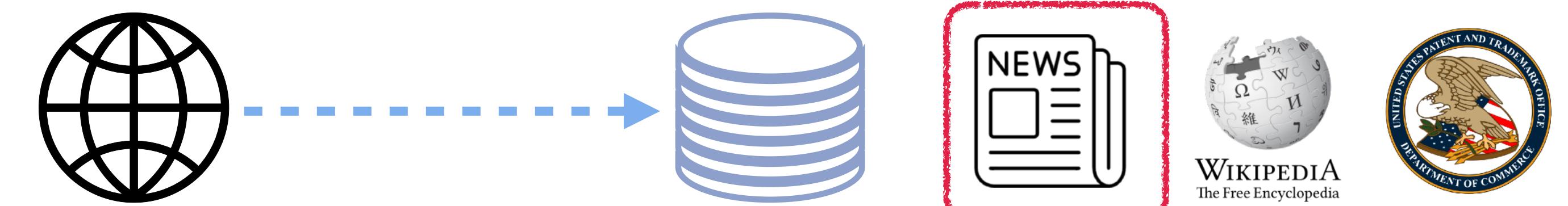


ChatGPT

I'm sorry, but I don't have access to real-time information including events beyond January 2022.

Core limitations of parametric LMs

Hallucinations

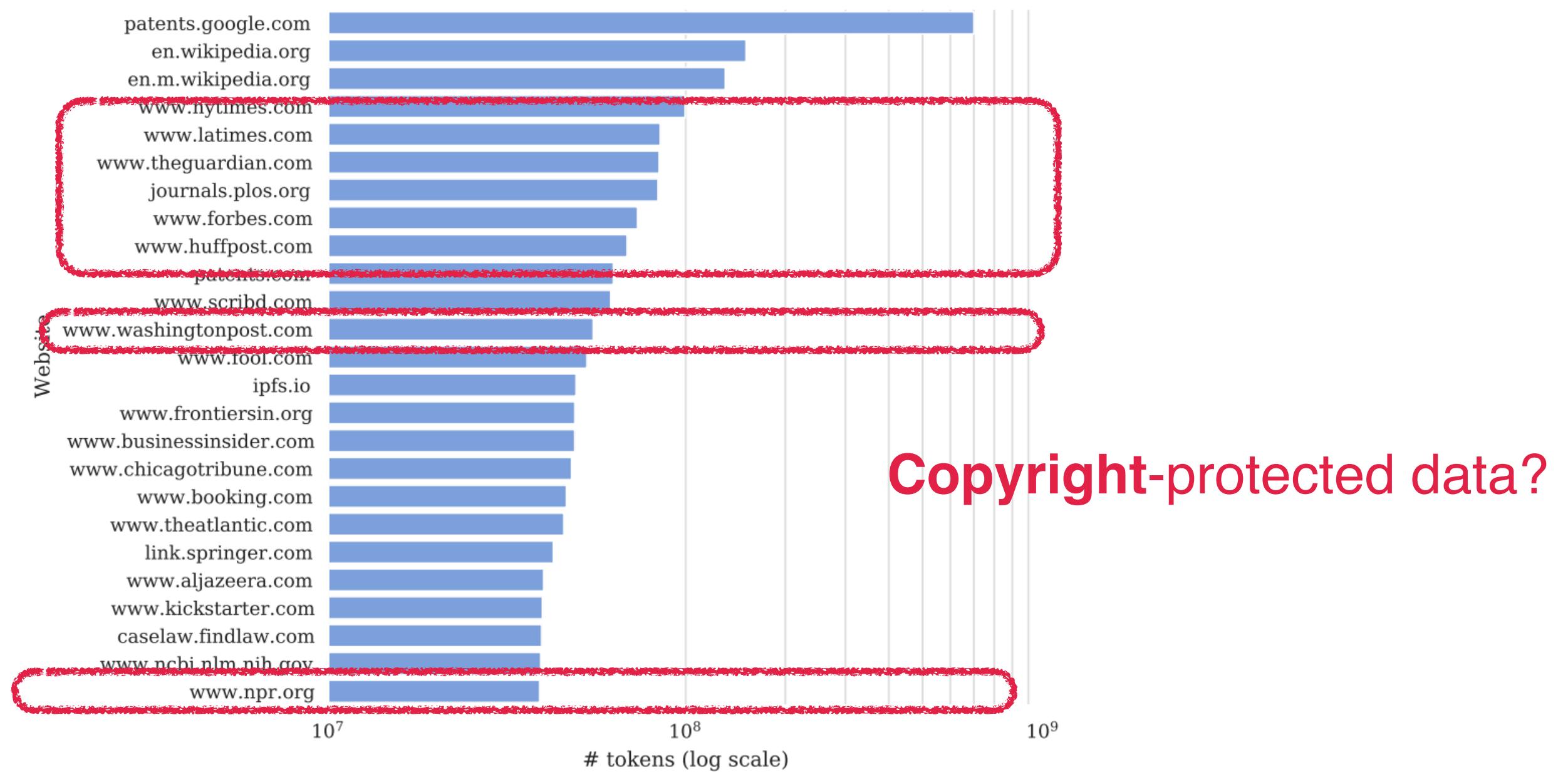


Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size



Core limitations of parametric LMs

Hallucinations

Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Case 1:23-cv-11195 Document 1 Filed 12/27/23 Page 1 of 69

THE NEW YORK TIMES COMPANY, Plaintiff,
v.
MICROSOFT CORPORATION, OPENAI LP, OPENAI GP, LLC, OPENAI OPCO LLC, OPENAI LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC, Defendants.

B. Defendants' GenAI Products

I. A Business Model Based on Mass Copyright Infringement

57. Despite its early promises of altruism, OpenAI quickly became a multi-billion-dollar for-profit business built in large part on the unlicensed exploitation of copyrighted works belonging to The Times and others. Just three years after its founding, OpenAI shed its exclusively Plaintiff The New York Times Company ("The Times"), by its attorneys Susman Godfrey LLP and Rothwell, Figg, Ernst & Manbeck, P.C., for its complaint against Defendants Microsoft Corporation ("Microsoft") and OpenAI, Inc., OpenAI LP, OpenAI GP LLC, OpenAI LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, (collectively "OpenAI" and, with Microsoft, "Defendants"), alleges as follows:

I. NATURE OF THE ACTION

1. Independent journalism is vital to our democracy. It is also increasingly rare and valuable. For more than 170 years, The Times has given the world deeply reported, expert, independent journalism. Times journalists go where the story is, often at great risk and cost, to inform the public about important and pressing issues. They bear witness to conflict and disasters, provide accountability for the use of power, and illuminate truths that would otherwise go unseen. Their essential work is made possible through the efforts of a large and expensive organization that provides legal, security, and operational support, as well as editors who ensure their journalism meets the highest standards of accuracy and fairness. This work has always been important. But

New York Times lawsuits
against OpenAI

Core limitations of parametric LMs

Hallucinations

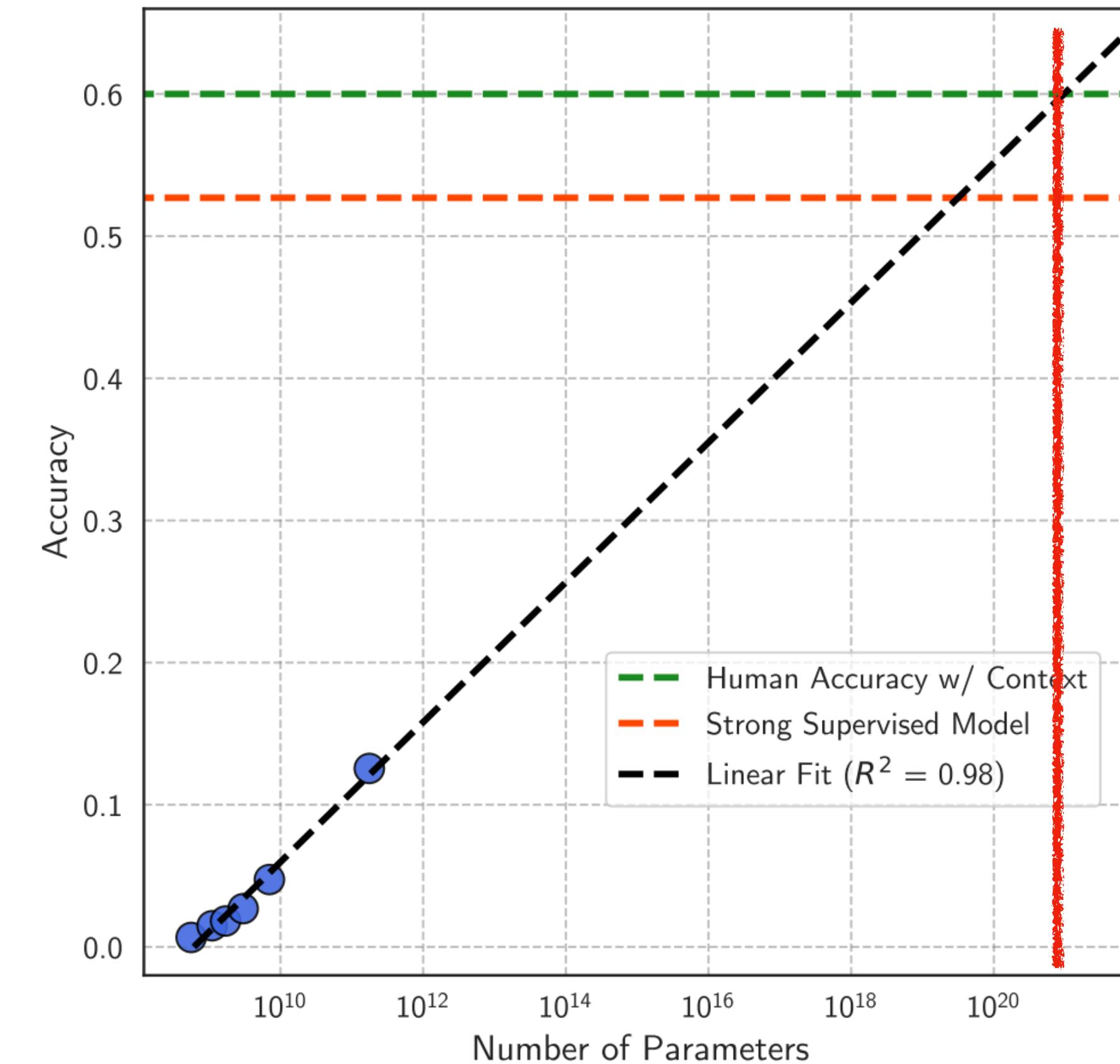
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Long-tail QA performance



100 quintillion parameters required to reach human performance

Q: So how can retrieval-augmented LMs solve those challenges?

How retrieval-augmented LMs solve the issues?

Hallucinations

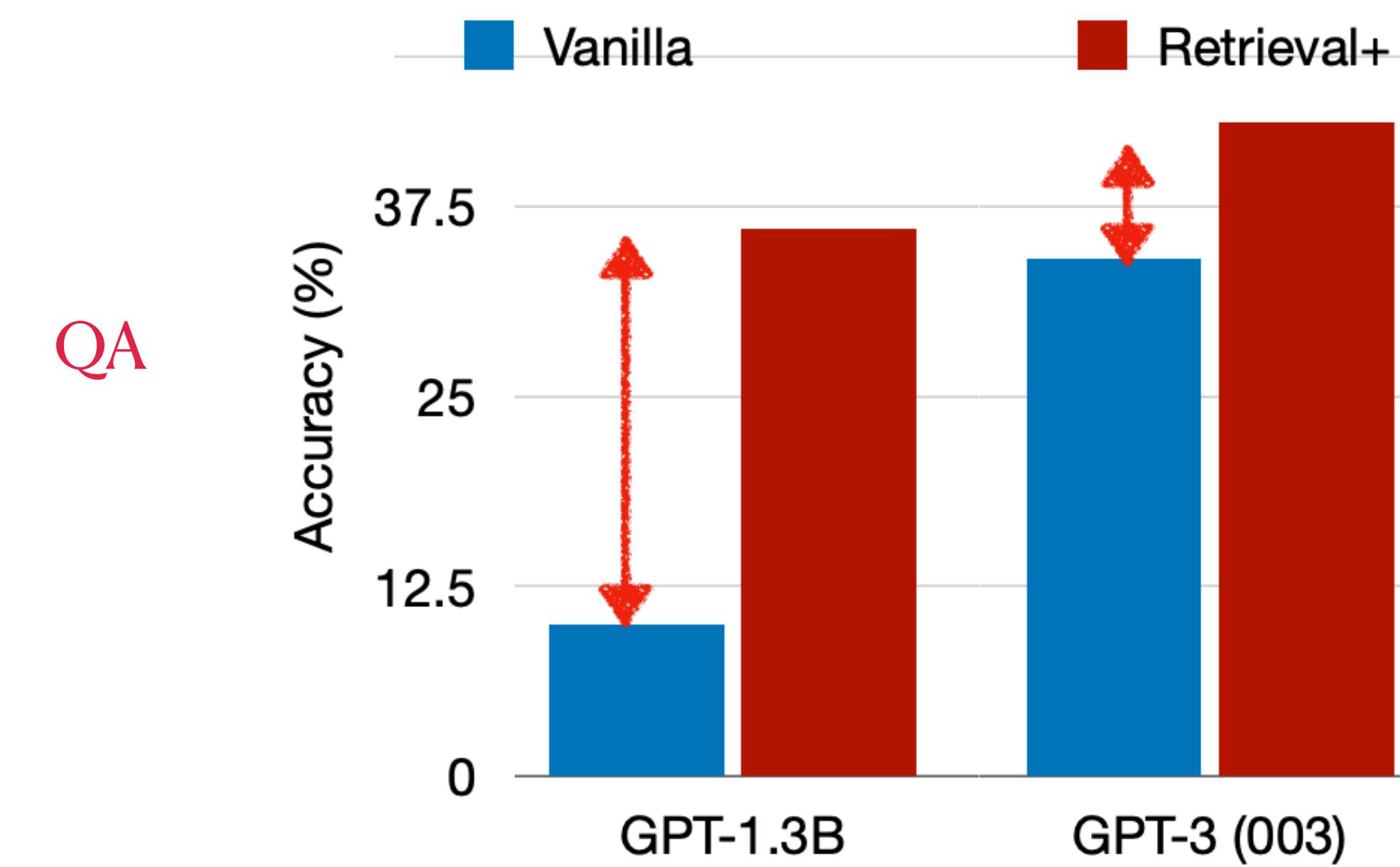
Significant improvements across model scale,
with larger gain with smaller LMs

Lack of attributions

Costs of adaptations

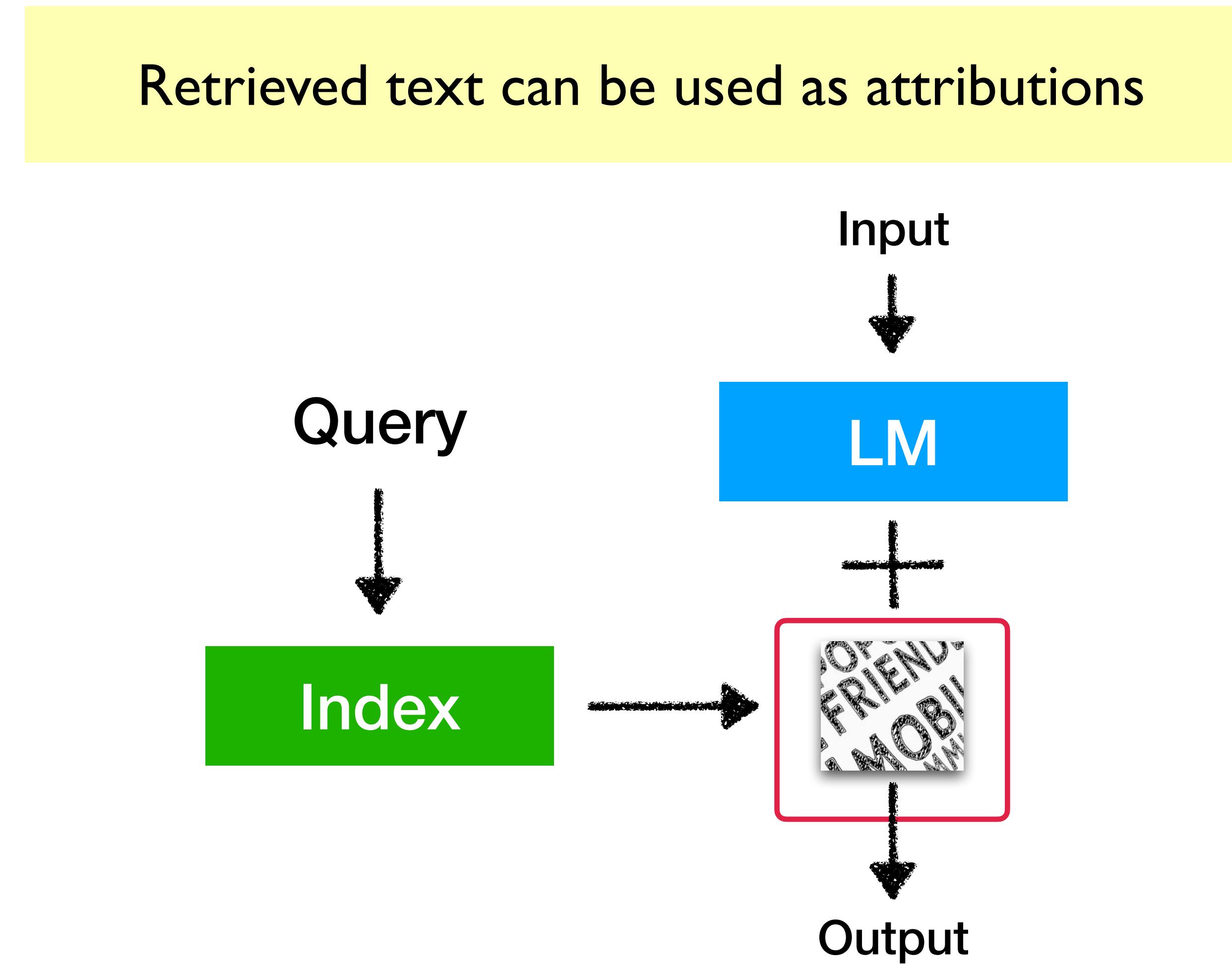
Copyright / privacy

Large parameter size



How retrieval-augmented LMs solve the issues?

- Hallucinations
- Lack of attributions
- Costs of adaptations
- Copyright / privacy
- Large parameter size



How retrieval-augmented LMs solve the issues?

Hallucinations

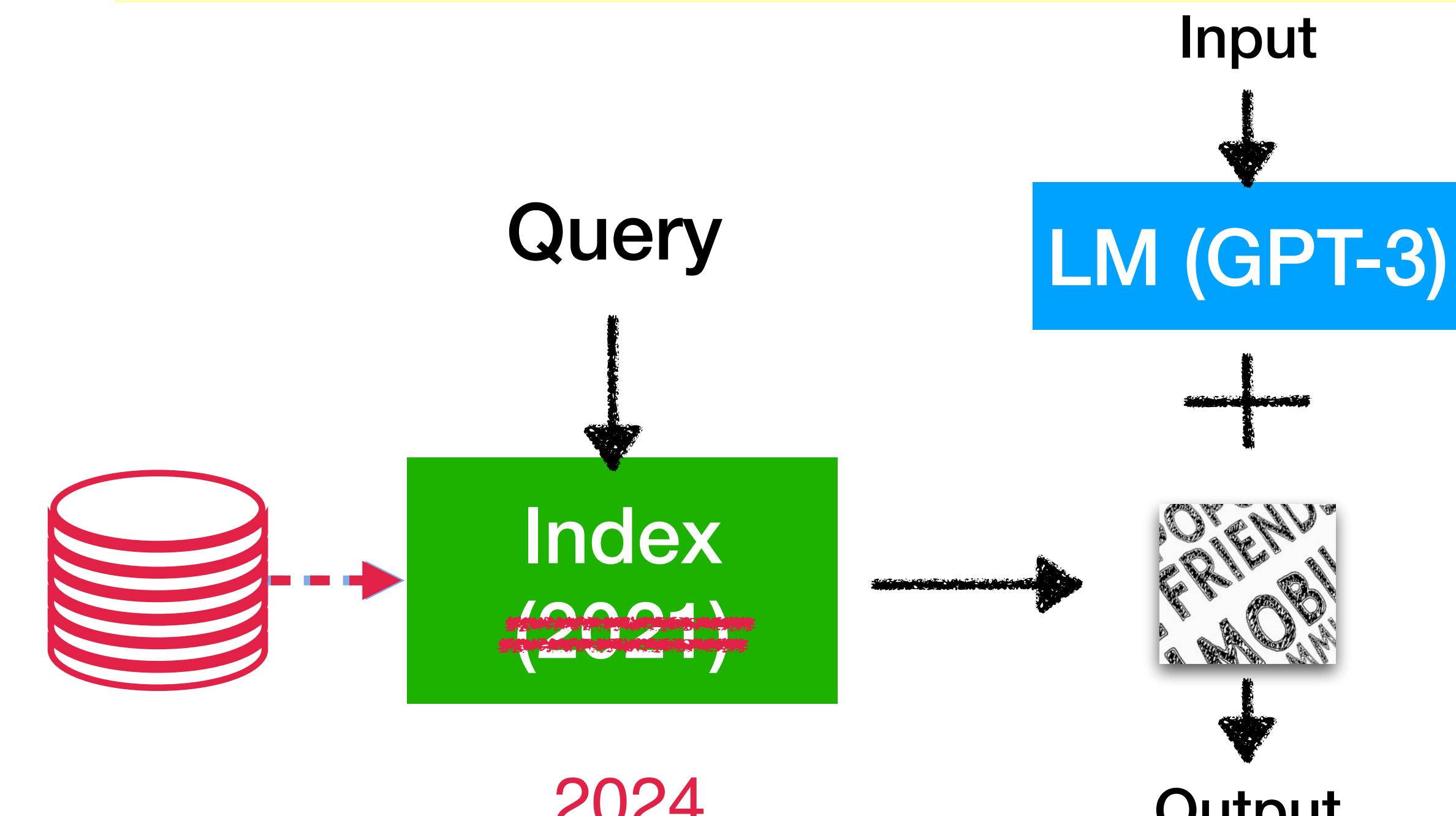
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Replacing datastores to catch up dynamically changing world without re-training



How retrieval-augmented LMs solve the issues?

Hallucinations

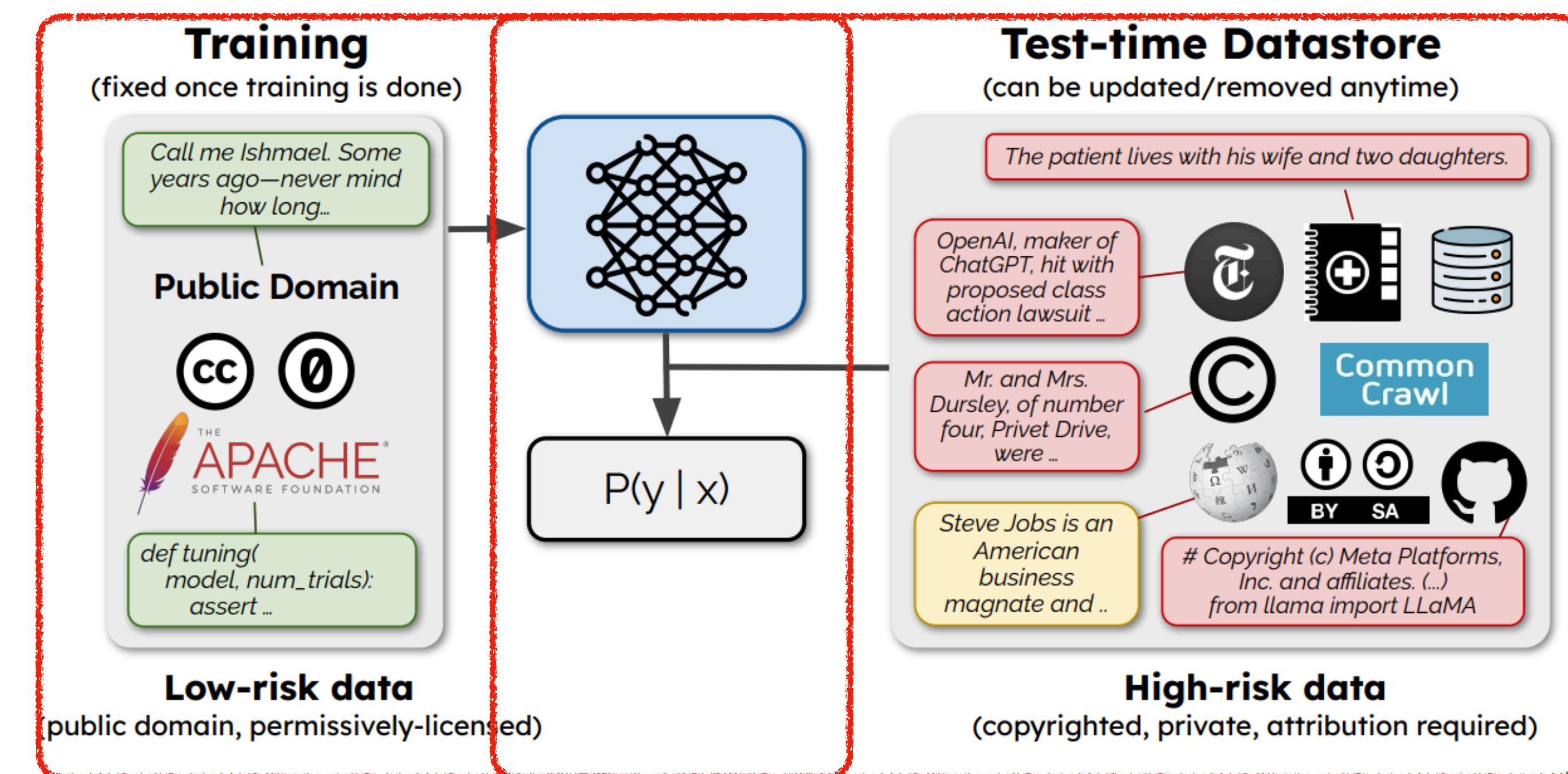
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Segregating copyright-sensitive data from pre-training data



Min* and Gururangan* et al., SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. ICLR 2024.

How retrieval-augmented LMs solve the issues?

Hallucinations

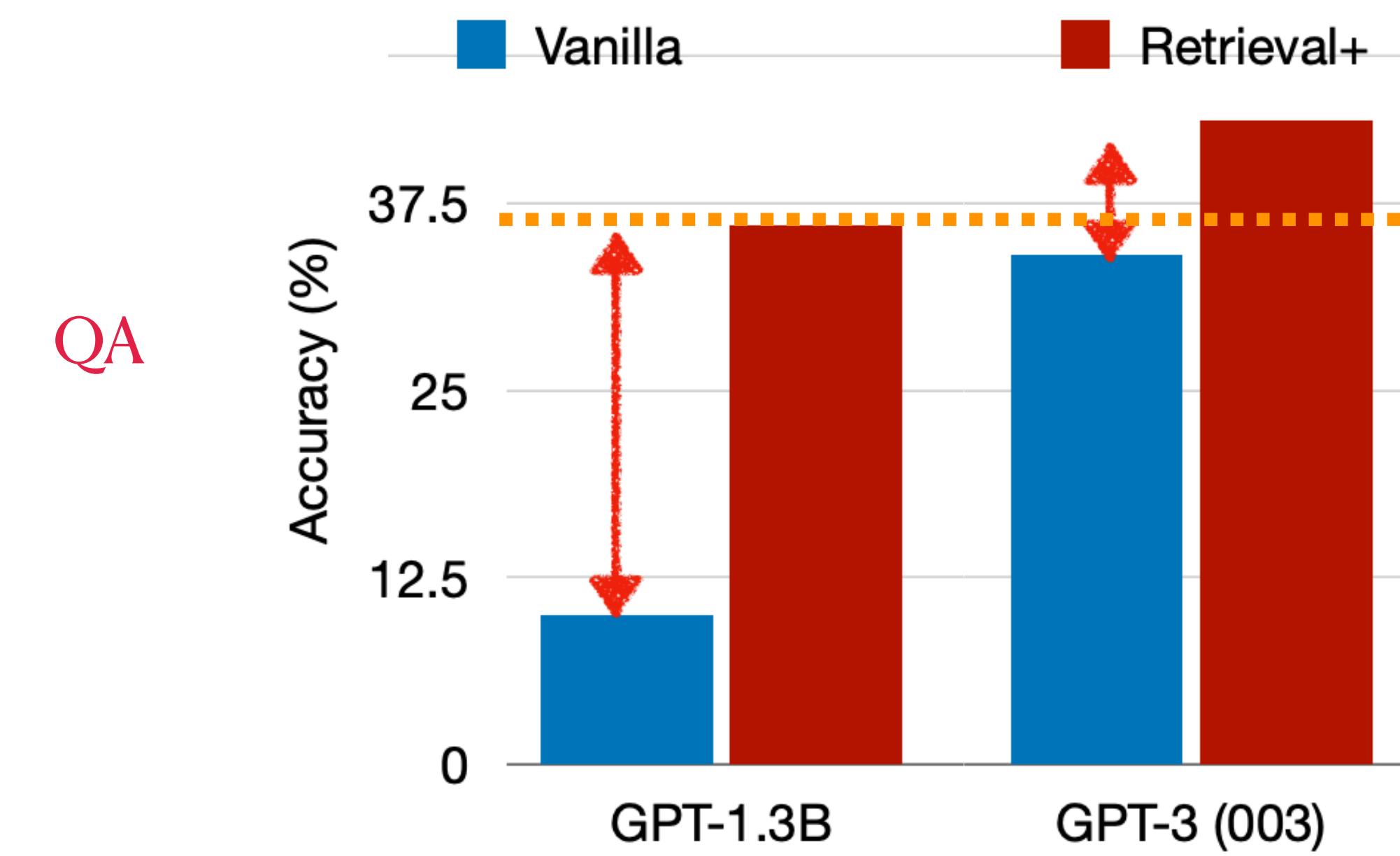
Lack of attributions

Costs of adaptations

Copyright / privacy

Large parameter size

Smaller LMs with retrieval outperform much larger LMs e.g., GPT-3



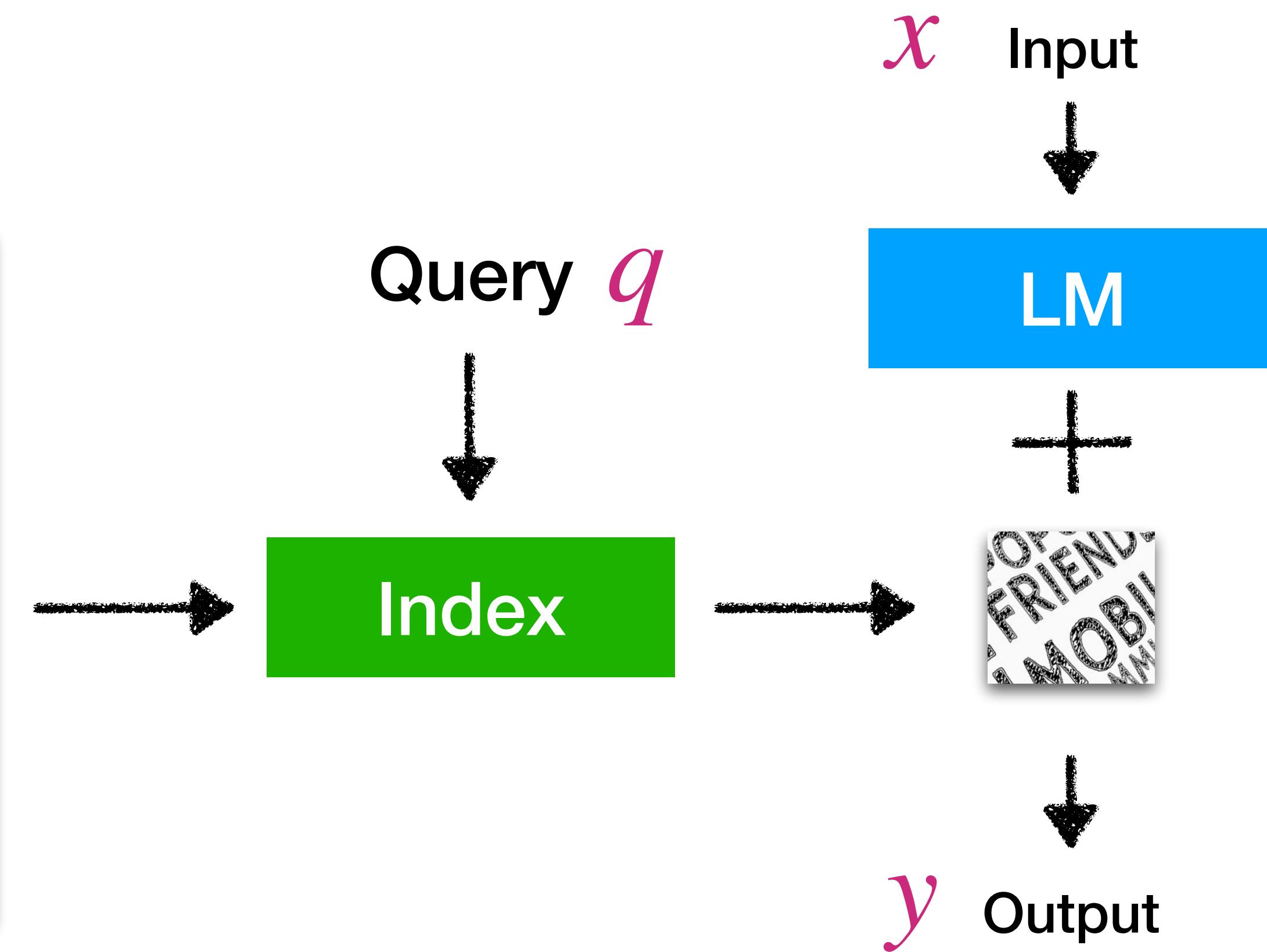
Architectures

Notations



Datastore

9



Inference: Index

Goal: find a small subset of elements in a datastore
that are the most similar to the query

sim: a similarity score between two pieces of text

Example

$$\text{sim}(i, j) = \frac{\text{tf}_{i,j} \times \log \frac{N}{\text{df}_i}}{\# \text{ of occurrences of } i \text{ in } j}$$

of total docs
of docs containing i

Example

$$\text{sim}(i, j) = \underline{\text{Encoder}(i)} \cdot \underline{\text{Encoder}(j)}$$

Maps the text into an h -dimensional vector

An entire field of
study on how to get
(or learn) the
similarity function
better
(We'll see some later!)

Inference: Index

Goal: find a small subset of elements in a datastore
that are the most similar to the query

sim: a similarity score between two pieces of text

Can be a totally separate research area on
how to do this fast & accurate

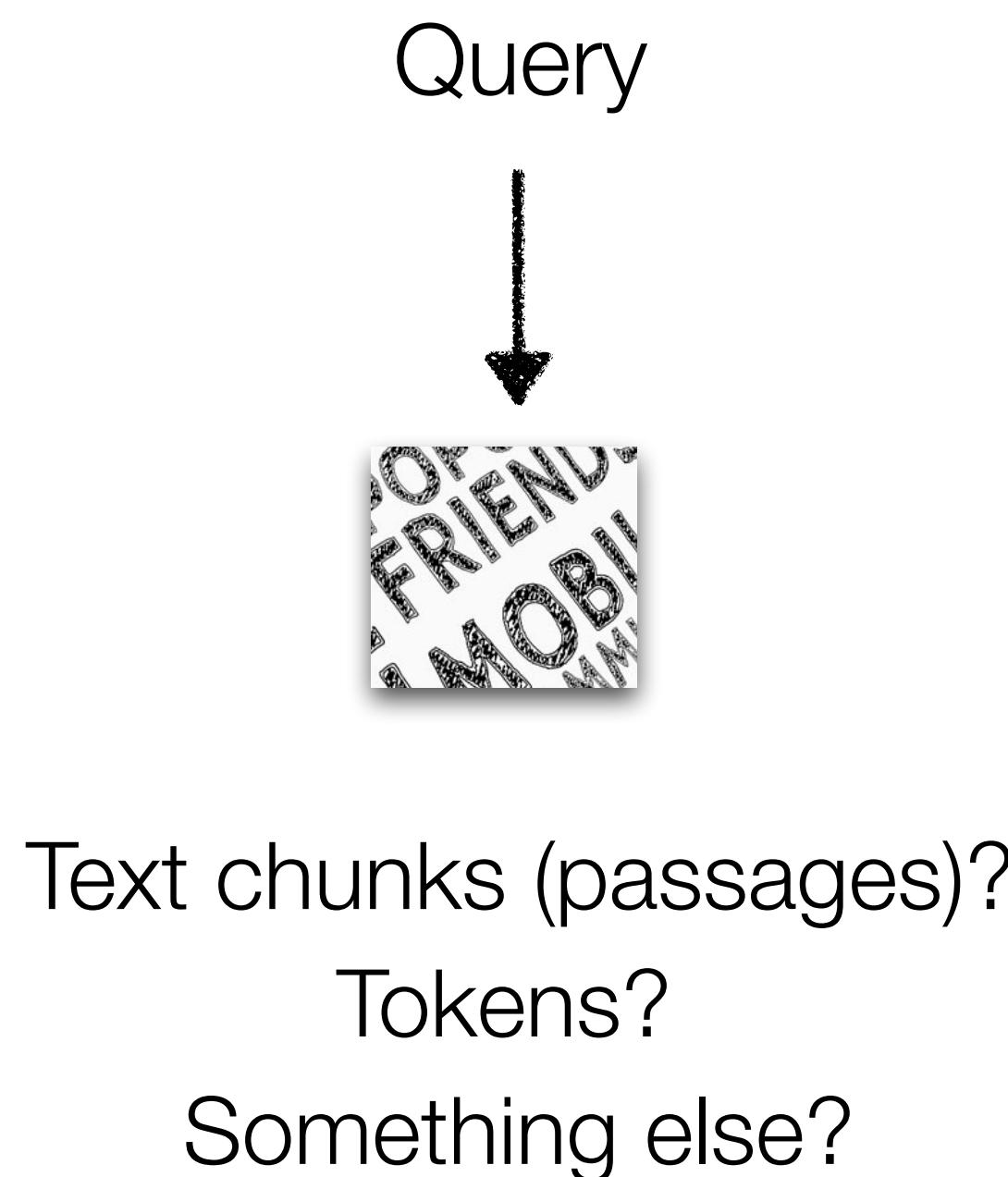
Index: given q , return $\arg \text{Top-}k_{d \in \mathcal{D}} \text{sim}(q, d)$ through fast nearest neighbor search

k elements from a datastore

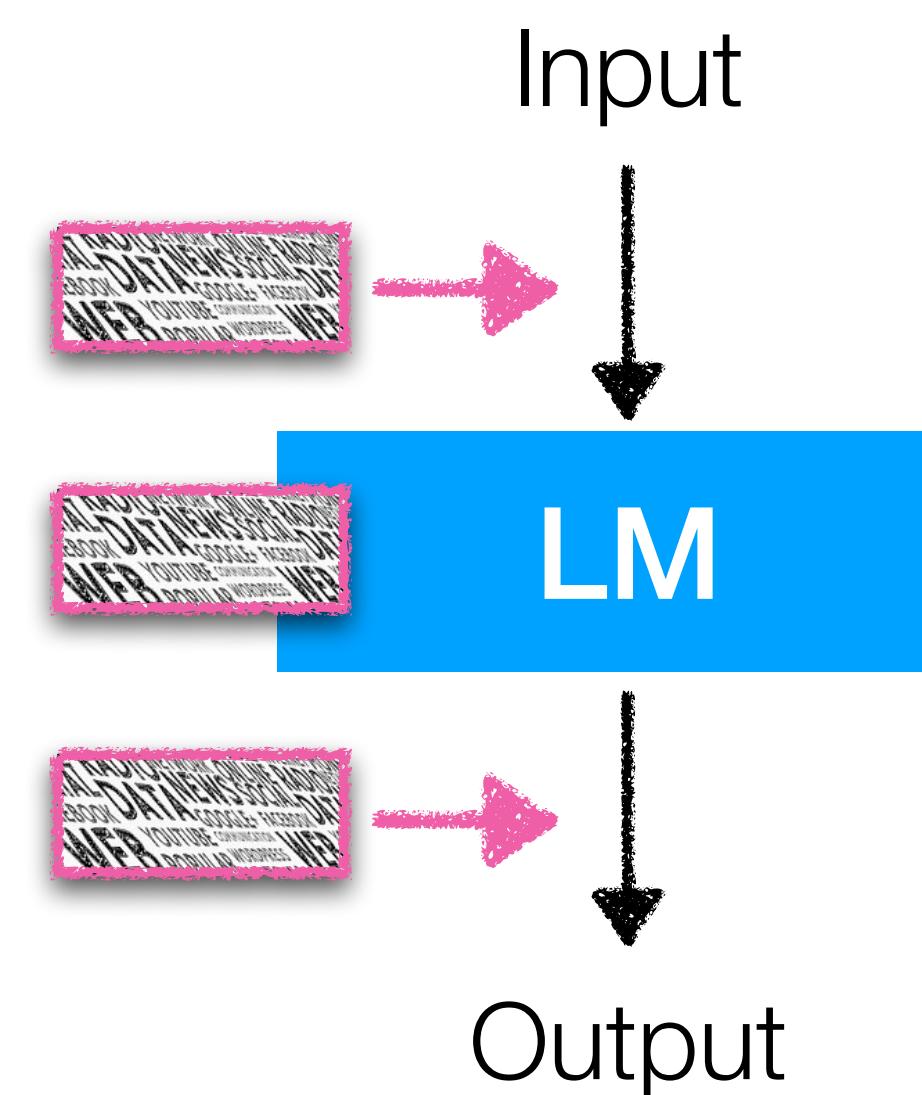
[https://github.com/
facebookresearch/faiss/wiki/](https://github.com/facebookresearch/faiss/wiki/)

Categorization of retrieval-augmented LMs

What to retrieve?



How to use retrieval?

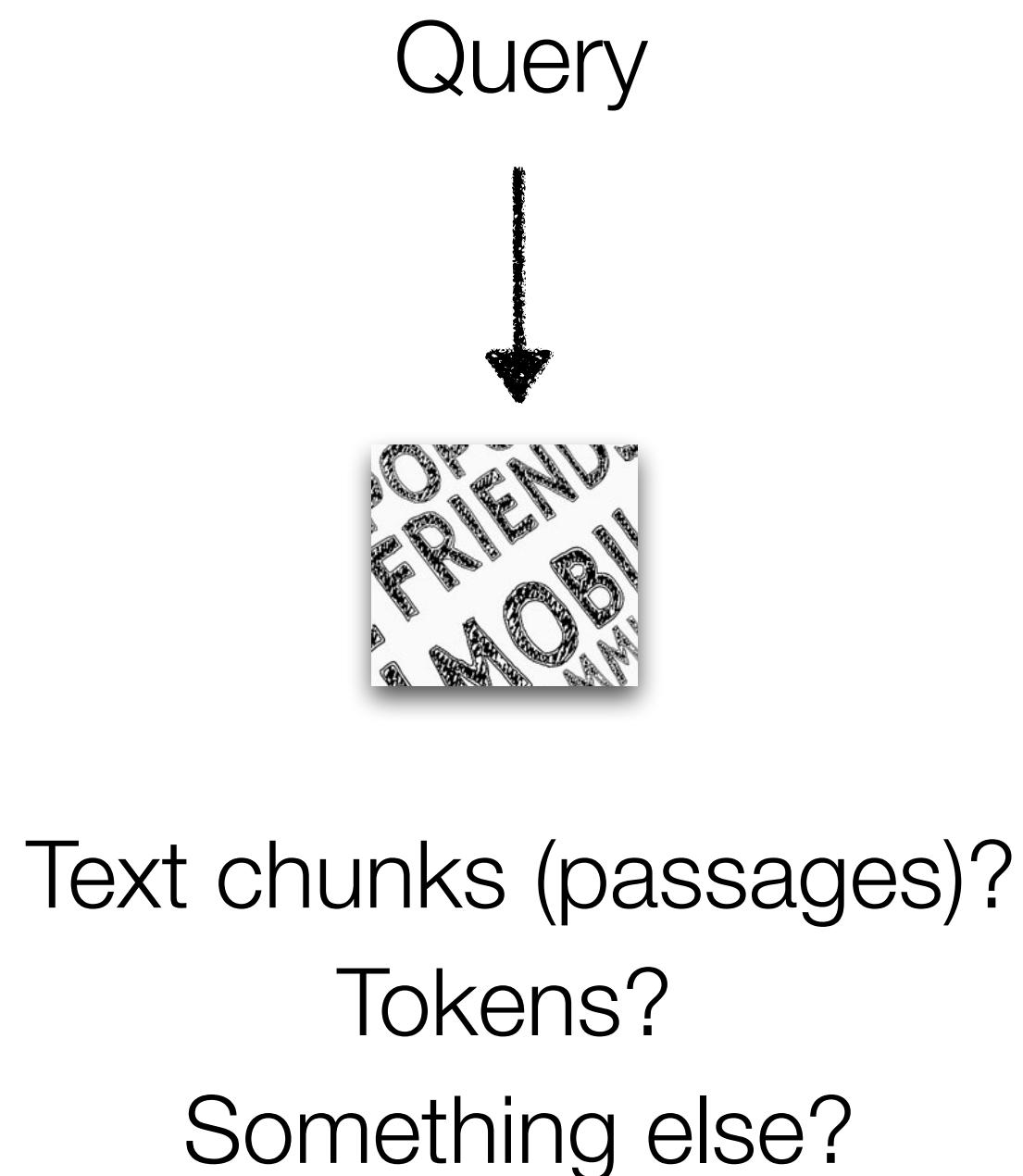


When to retrieve?

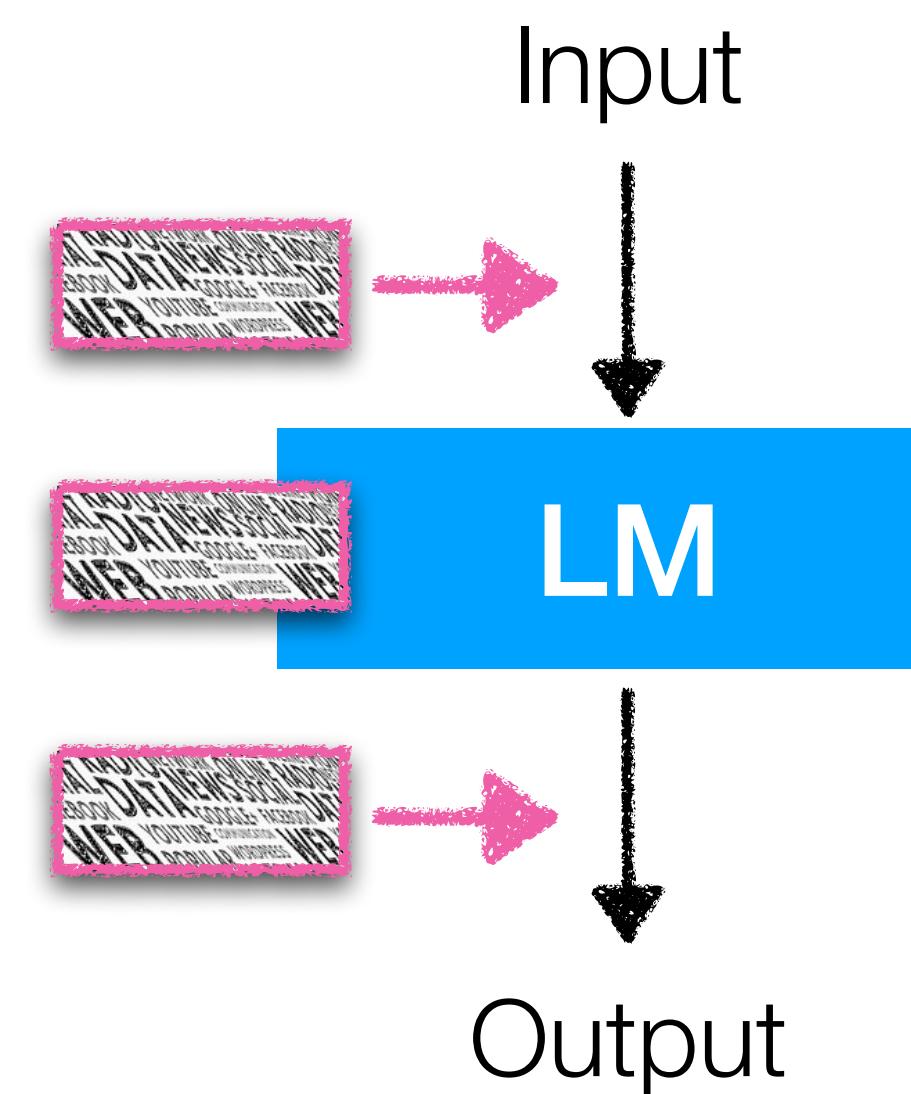


Categorization of retrieval-augmented LMs

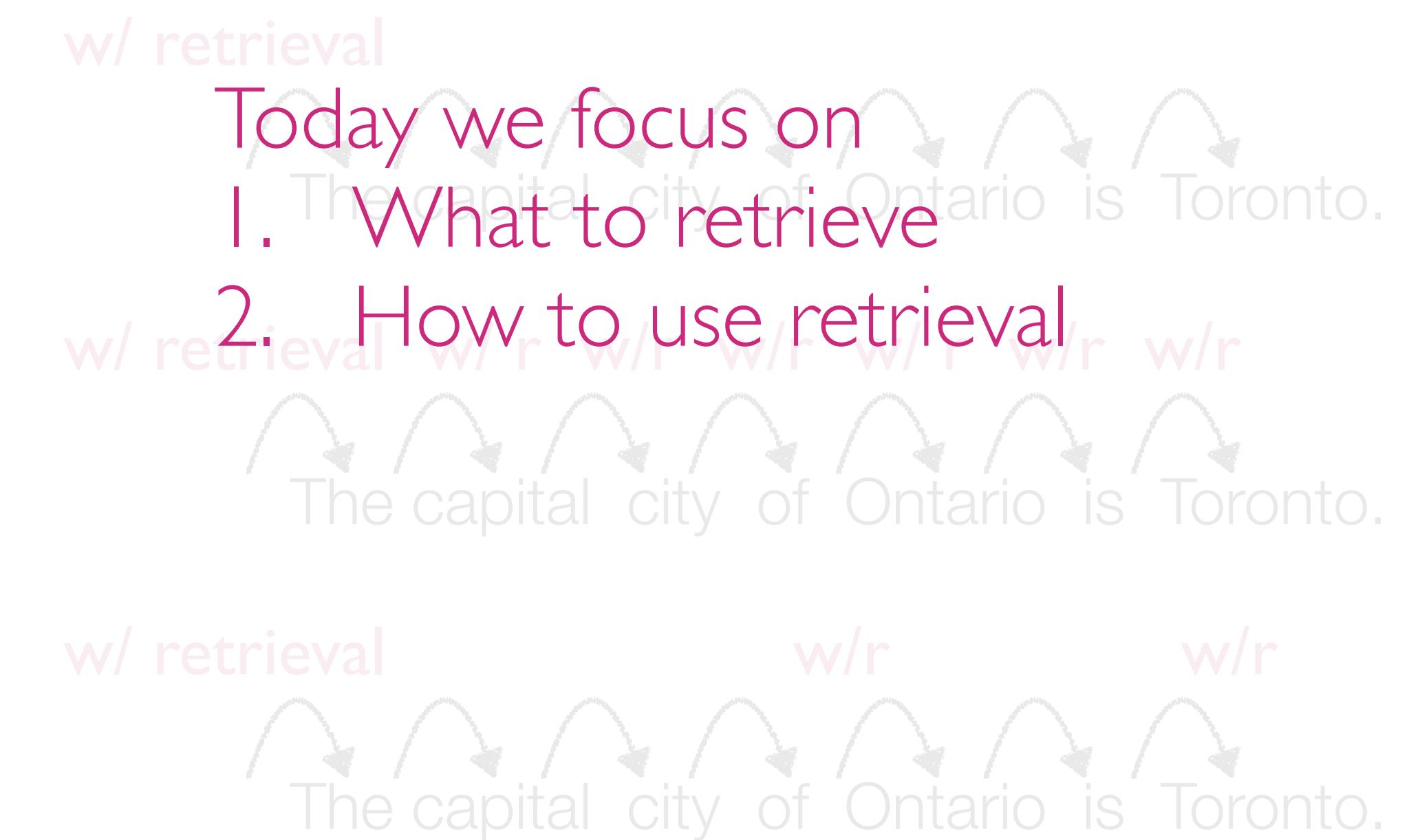
What to retrieve?



How to use retrieval?



When to retrieve?



Three representative architectures

What: Text chunks
How: Input

Input augmentation (RAG)

What: Text chunks
How: Intermediate

Intermediate fusion

What: Tokens
How: Output

Output interpolations

Three representative architectures

What: Text chunks
How: Input

REALM (Guu et al., 2020)

What: Text chunks
How: Intermediate

RETRO (Borgeaud et al., 2021)

What: Tokens
How: Output

kNN-LM (Khandelwal et al., 2020)

Three representative architectures

What: Text chunks
How: Input

REALM (Guu et al., 2020)

What: Text chunks
How: Intermediate

RETRO (Borgeaud et al., 2021)

What: Tokens
How: Output

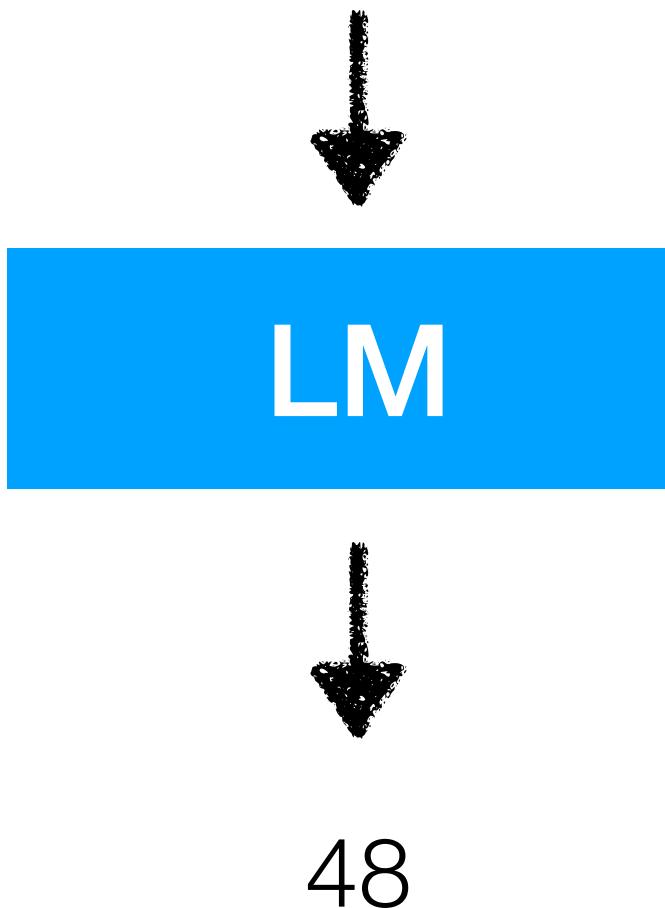
kNN-LM (Khandelwal et al., 2020)

REALM (Guu et al 2020)



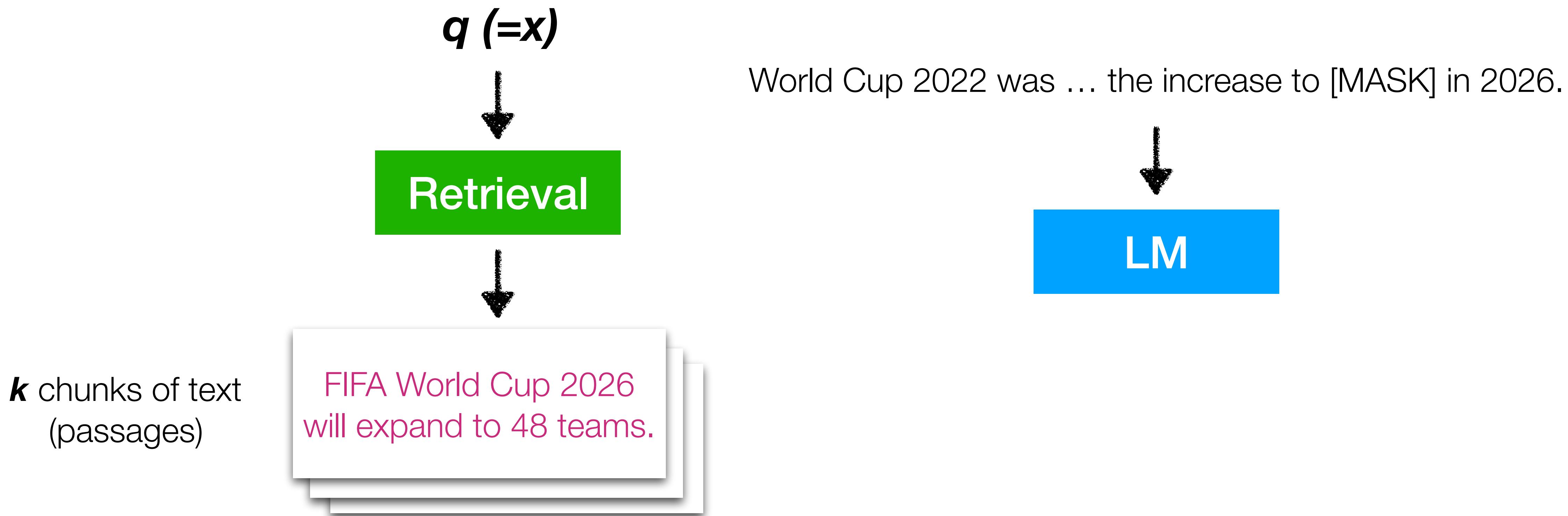
x = World Cup 2022 was the last with 32 teams before the increase to [MASK] in 2026.

World Cup 2022 was ... the increase to [MASK] in 2026.



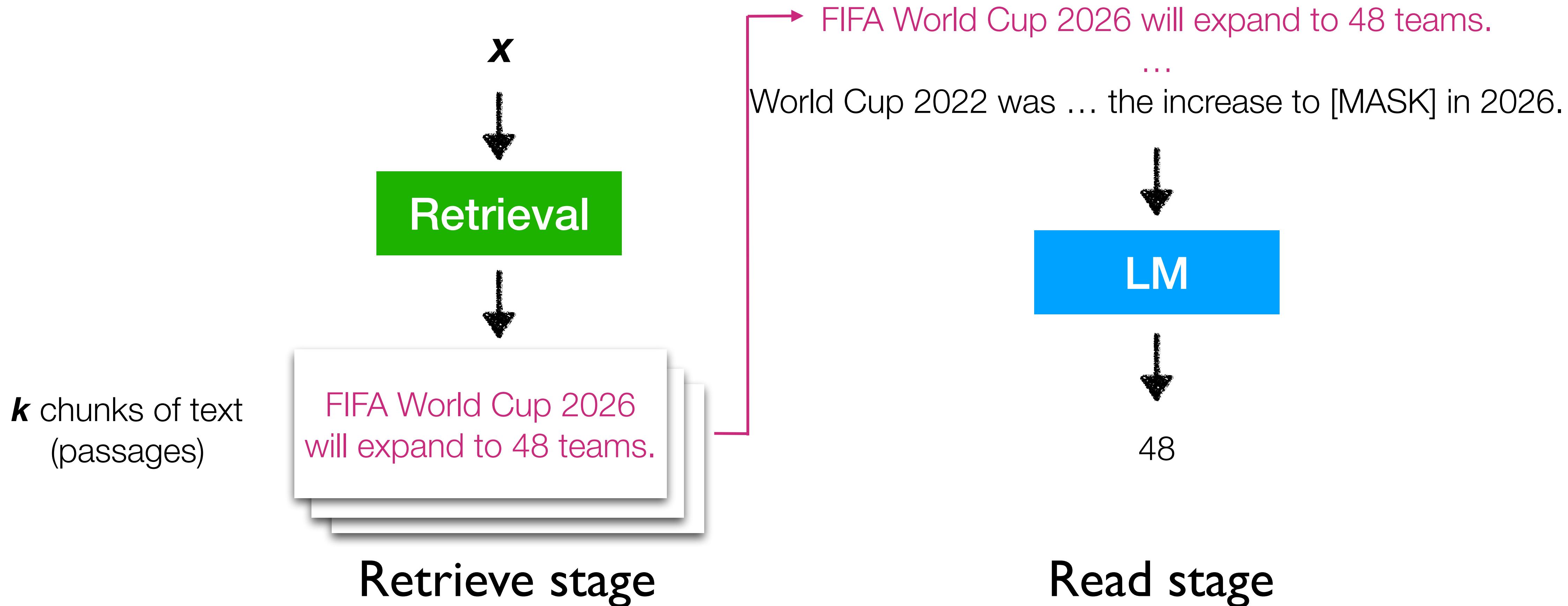
REALM (Guu et al 2020)

x = World Cup 2022 was the last with 32 teams before the increase to [MASK] in 2026.



REALM (Guu et al 2020)

x = World Cup 2022 was the last before the increase to [MASK] in the 2026 tournament.



REALM: (1) Retrieve stage

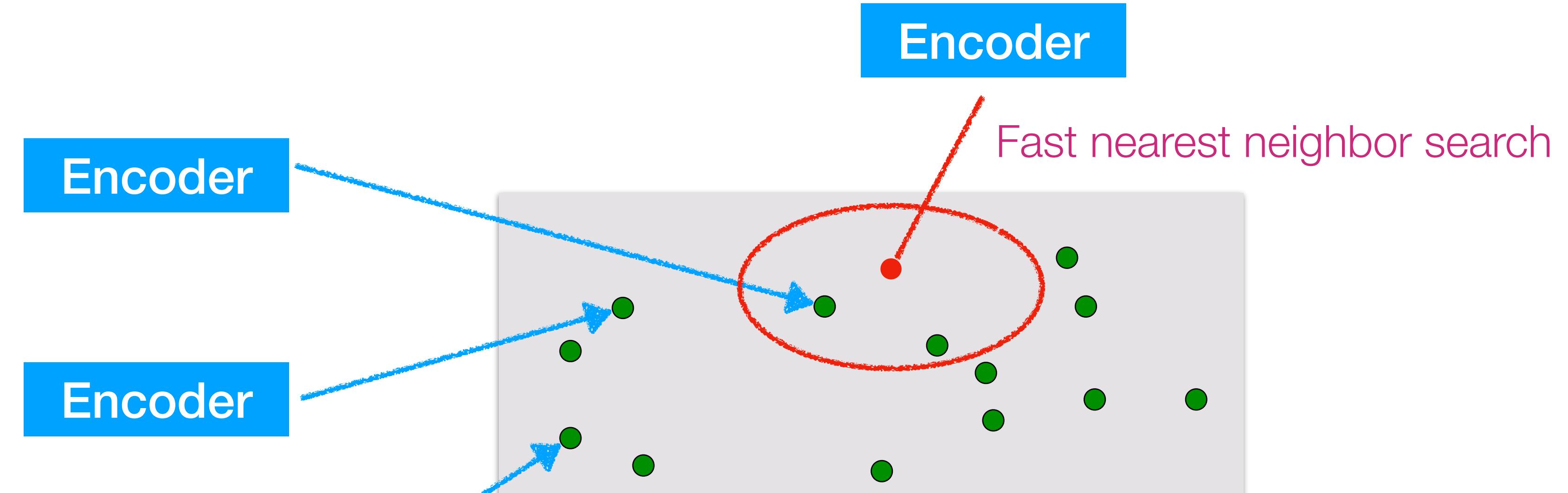
FIFA World Cup 2026 will expand to 48 teams.

In 2022, the 32 national teams involved in the tournament.

Team USA celebrated after winning its match against Iran ...

Wikipedia
13M chunks (passages)
(called *documents* in the paper)

\mathbf{x} = World Cup 2022 was ... the increase to [MASK] in 2026.



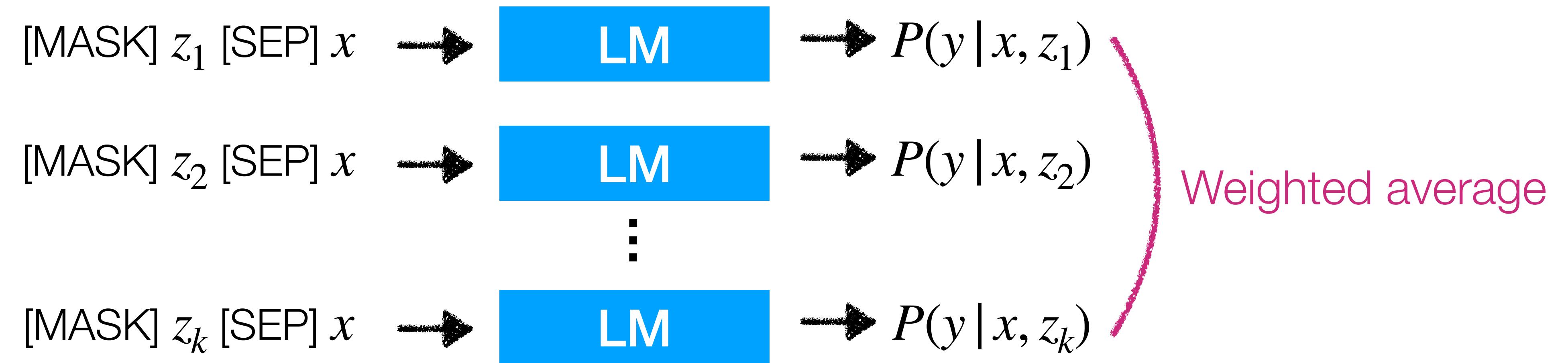
$$\mathbf{z} = \text{Encoder}(\mathbf{z})$$

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

$$z_1, \dots, z_k = \text{argTop-}k(\mathbf{x} \cdot \mathbf{z})$$

k retrieved chunks

REALM: (2) Read stage



Need to approximate
→ Consider top k chunks only

$$\sum_{z \in \mathcal{D}} \underbrace{P(z | x)}_{\text{from the retrieve stage}} \underbrace{P(y | x, z)}_{\text{from the read stage}}$$

REALM (Guu et al 2020)

What to retrieve?

- Chunks ✓
- Tokens
- Others

How to use retrieval?

- Input layer ✓
- Intermediate layers
- Output layer

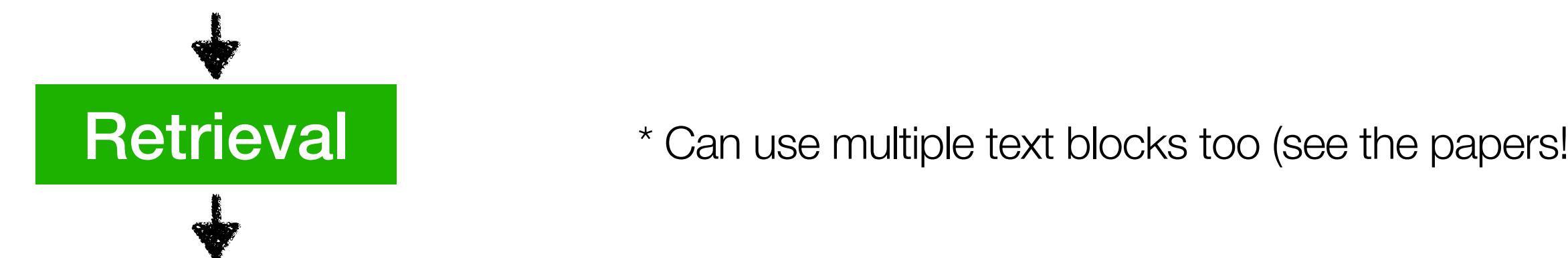
When to retrieve?

- Once ✓
- Every n tokens ($n > 1$)
- Every token

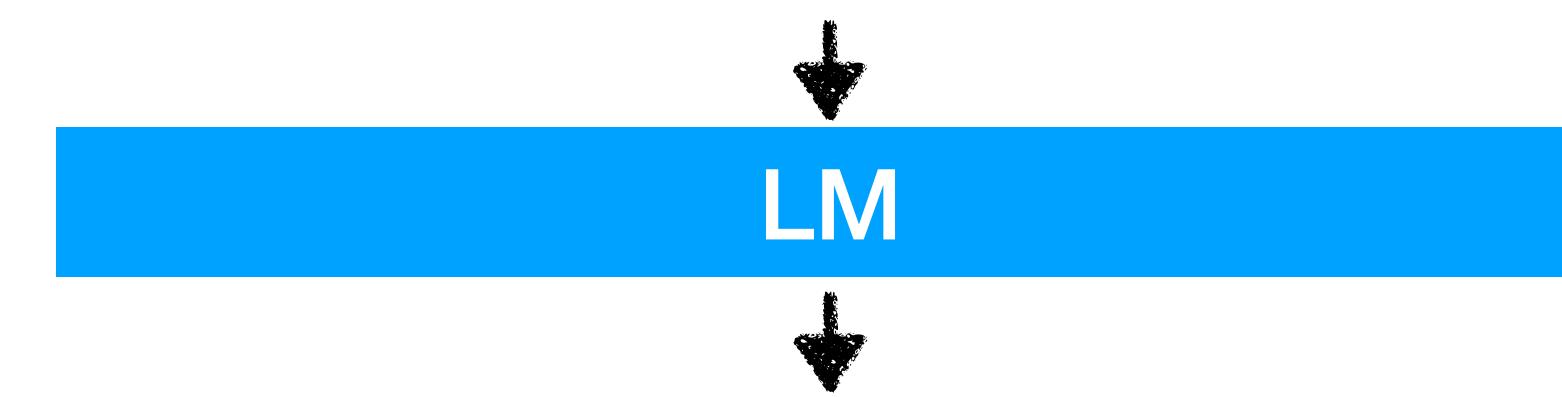
Retrieval-in-context LM

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

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



FIFA World Cup 2026 will expand to 48 teams. World Cup 2022 was the last with 32 teams, before the increase to



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

Is the whole context necessary?

Is $q=x$ necessary?

x = Team USA celebrates after winning its match against Iran at Al Thumama Stadium in Group B play of the FIFA World Cup 2022 on Nov. 29, 2022. (..) World Cup 2022 was the last with 32 teams, before the increase to

Team USA celebrates after winning its match against Iran at Al Thumama Stadium in Group B play of the FIFA World Cup 2022 on Nov. 29, 2022. (..) World Cup 2022 was the last with 32 teams, before the increase to

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

Retrieval

Retrieval

The U.S. national team defeated Iran 1-0.

FIFA World Cup 2026 will expand to 48 teams.

Does not cover “tokens that will come next”

more relevant to what will come next

Is the whole context necessary?

Is $q=x$ necessary?

x = Team USA celebrates after winning its match against Iran at Al Thumama Stadium in Group B play of the FIFA World Cup 2022 on Nov. 29, 2022. (..) World Cup 2022 was the last with 32 teams.

Team USA celebrates after winning its match against Iran at Al Thumama Stadium in Group B play of the FIFA World Cup 2022 on Nov. 29, 2022. (..) World Cup 2022 was the last with 32 teams, before the increase to

Retrieval

The U.S. national team defeated Iran 1-0.

Does not cover “tokens that will come next”

Partial context at the end is better when long contexts in input

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

Retrieval

FIFA World Cup 2026 will expand to 48 teams.

more relevant to what will come next

How frequently do we retrieve?

World Cup 2022 was the last with



Retrieval



The 2022 FIFA World Cup (...) 32 national teams involved in the tournament. World Cup 2022 was the last with

LM



32 teams before the increase to 48 in the 2026 tournament.

explained by retrieval

not really covered

How frequently do we retrieve?

World Cup 2022 was the last with

Retrieval

The 2022 FIFA World Cup (...) 32 national teams involved in the tournament. World Cup 2022 was the last with

LM

32 teams before the increase

World Cup 2022 was the last with 32 teams before the increase

Retrieval

FIFA World Cup 2026 will expand to 48 teams. World Cup 2022 was the last with 32 teams, before the increase

LM

to 48 in the 2026 tournament.

Retrieval results from a new query explain them!

How frequently do we retrieve?

World Cup 2022 was the last with

Retrieval

The 2022 FIFA World Cup (...) 32 national teams involved in the tournament. World Cup 2022 was the last with

LM

32 teams before the increase

World Cup 2022 was the last with 32 teams before the increase

Doing this repeatedly increases inference time cost

FIFA World Cup 2026 will have 48 teams, up from 32,

LM

to 48 in the 2026 tournament.

Retrieval results from a new query explain them!

Retrieval-in-context LM

What to retrieve?

- Chunks ✓
- Tokens
- Others

How to use retrieval?

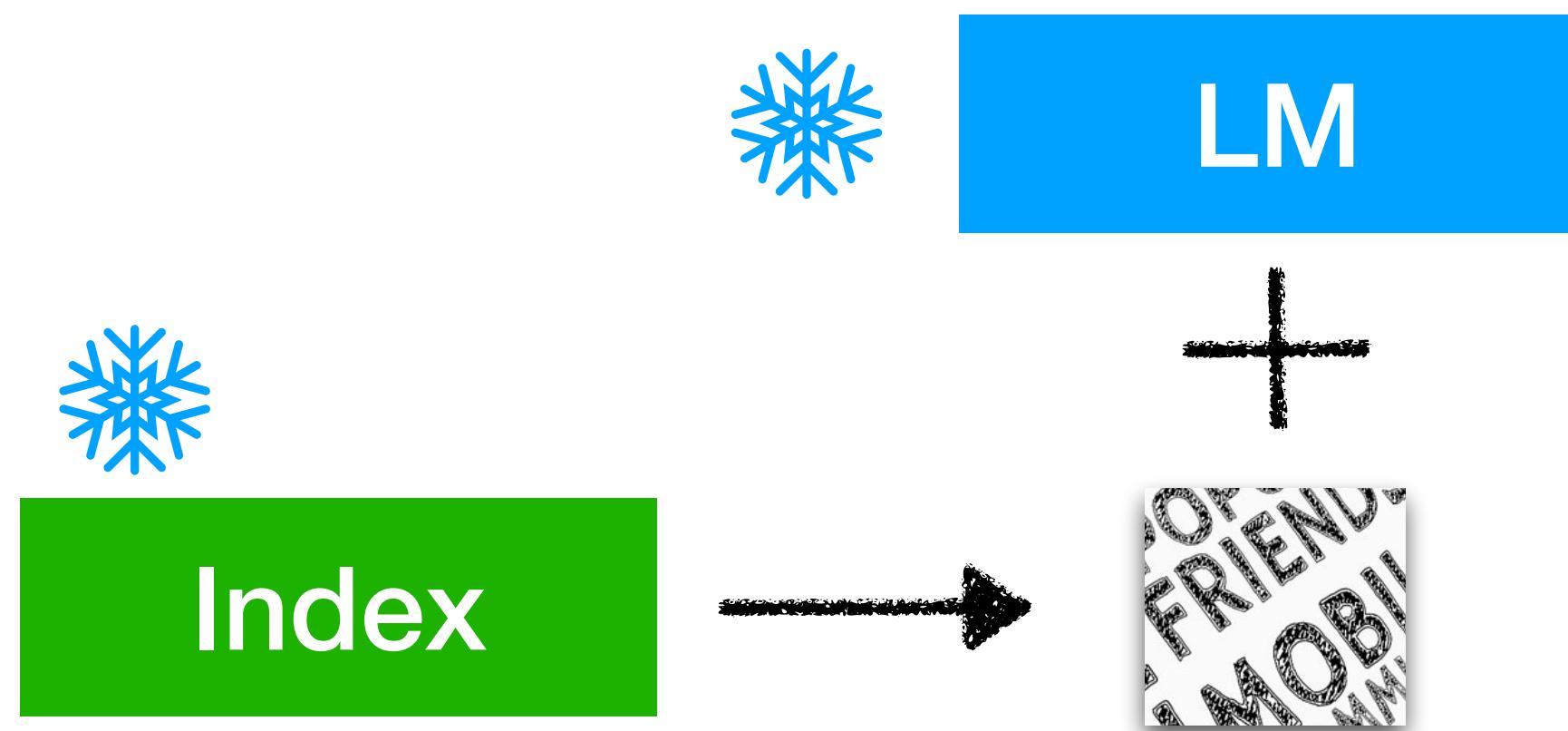
- Input layer ✓
- Intermediate layers
- Output layer

When to retrieve?

- Once
- Every n tokens ($n > 1$) ✓
- Every token

Recent trend: RAG with LLMs

Existing parametric LMs
(e.g., GPT-3)



Simply combining existing models w/o training has shown to be successful!

Off-the-shelf retrievers (e.g., Google search, BM25, DPR)

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

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

Three representative architectures

What: Text chunks
How: Input

More...
See also

REALM (Guu et al., 2020)

What: Text chunks
How: Intermediate

RETRO (Borgeaud et al., 2021)

Our position paper (Asai et al., 2024; https://akariyasai.github.io/assets/pdf/ralm_position.pdf)

What: Tokens
How: Output

kNN-LM (Khandelwal et al., 2020)

RETRO

(Borgeaud et al. 2022)

- ✓ Incorporation in the “intermediate layer” instead of the “input” layer
→ designed for many chunks, frequently, more efficiently
- ✓ Scale the datastore (1.8T tokens)

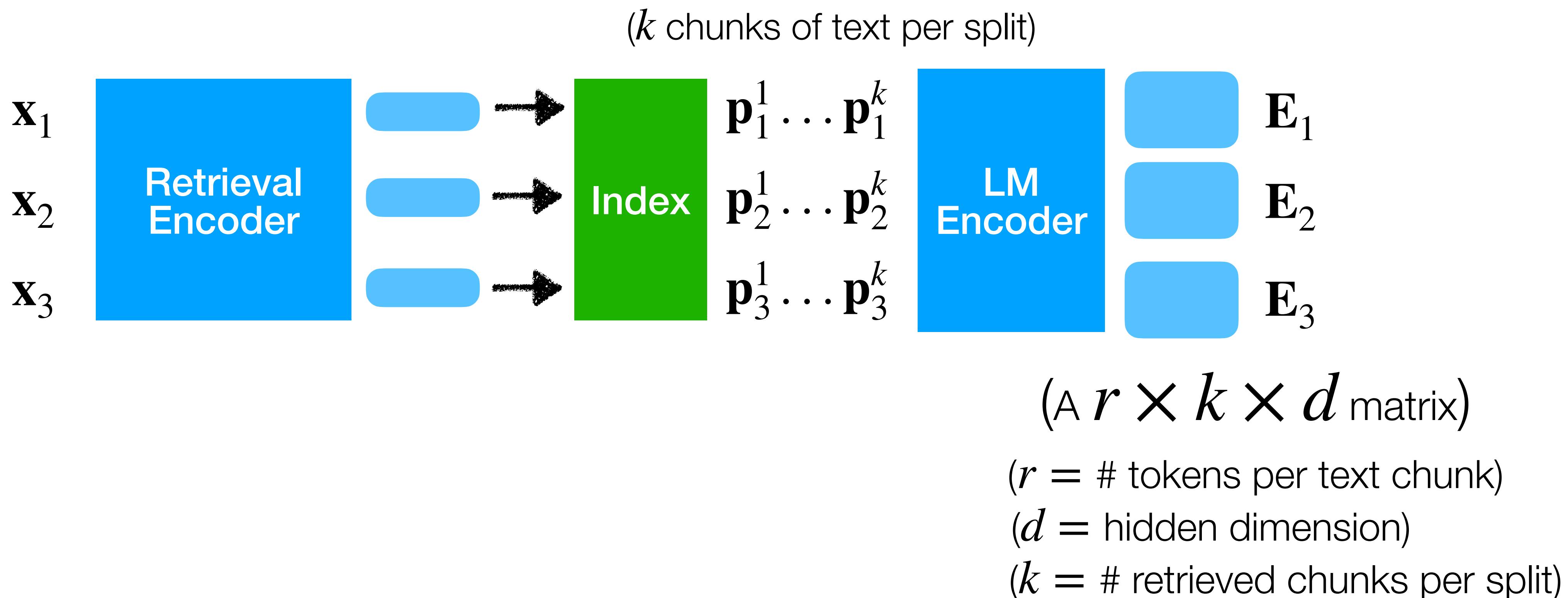
RETRO (Borgeaud et al. 2021)

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

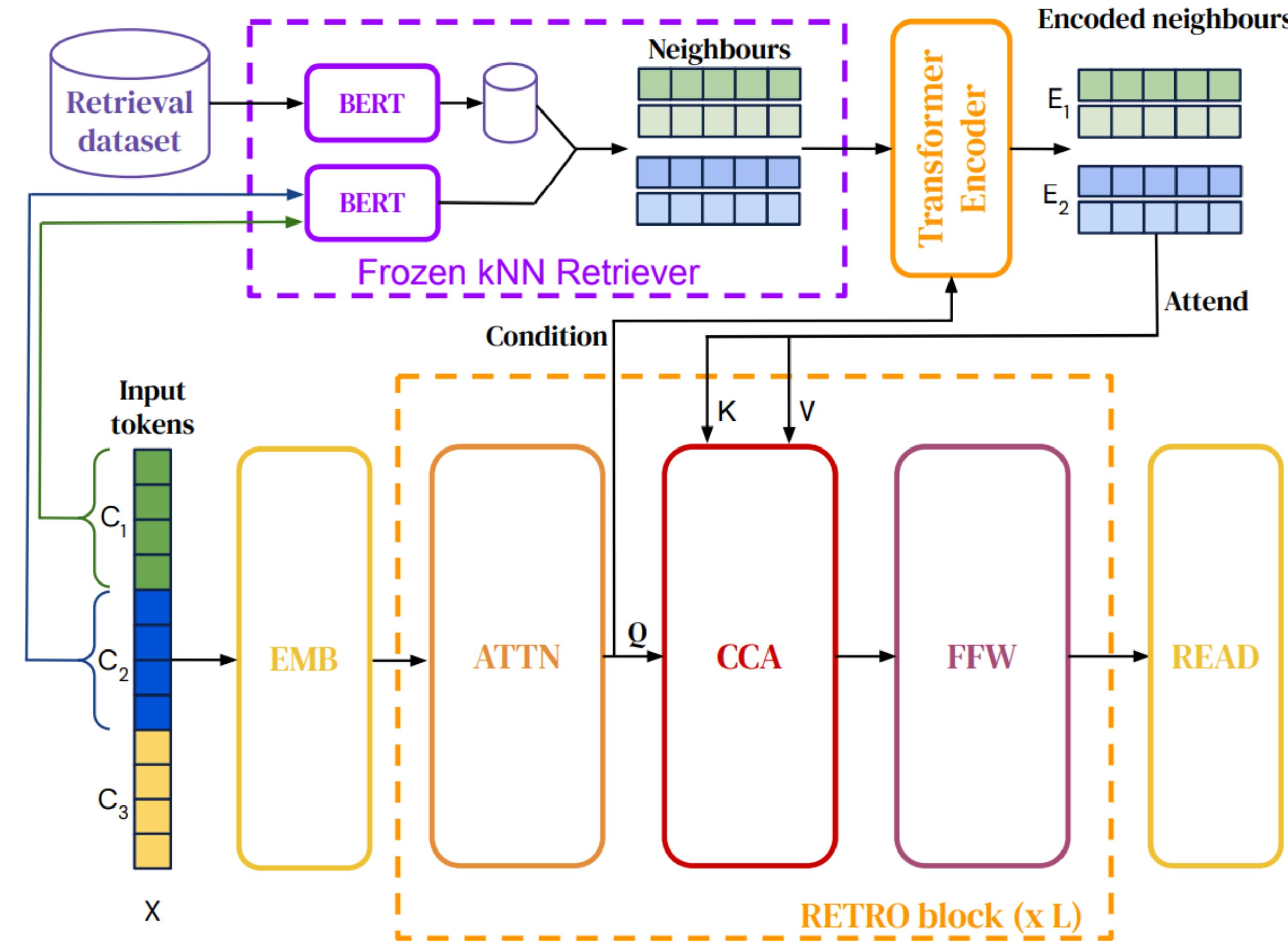
\mathbf{x}_1

\mathbf{x}_2

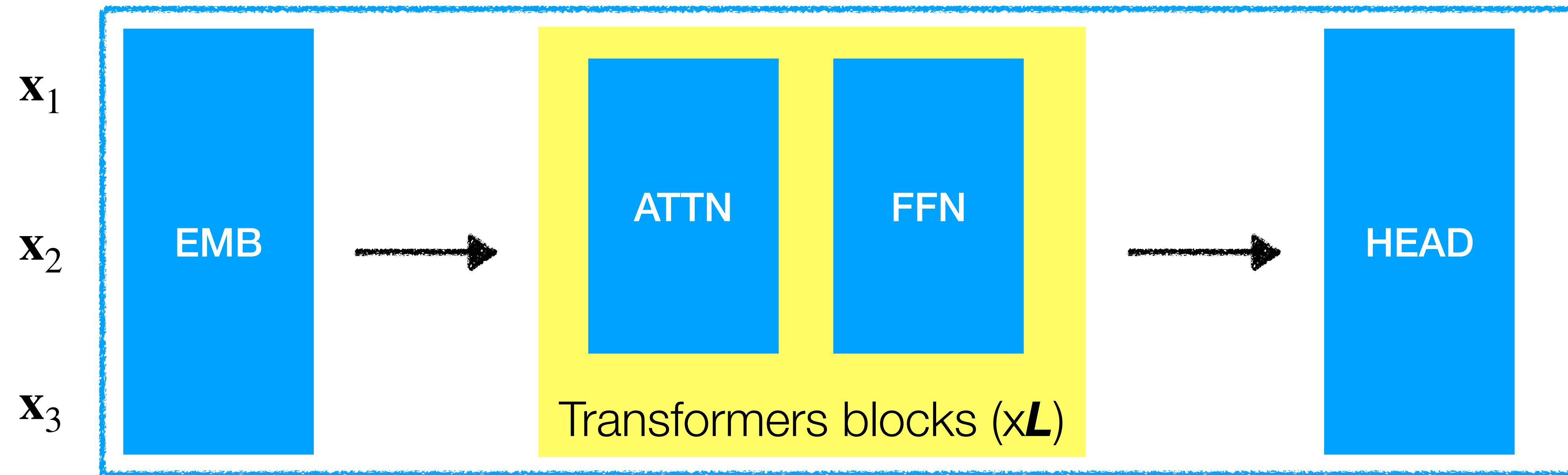
\mathbf{x}_3



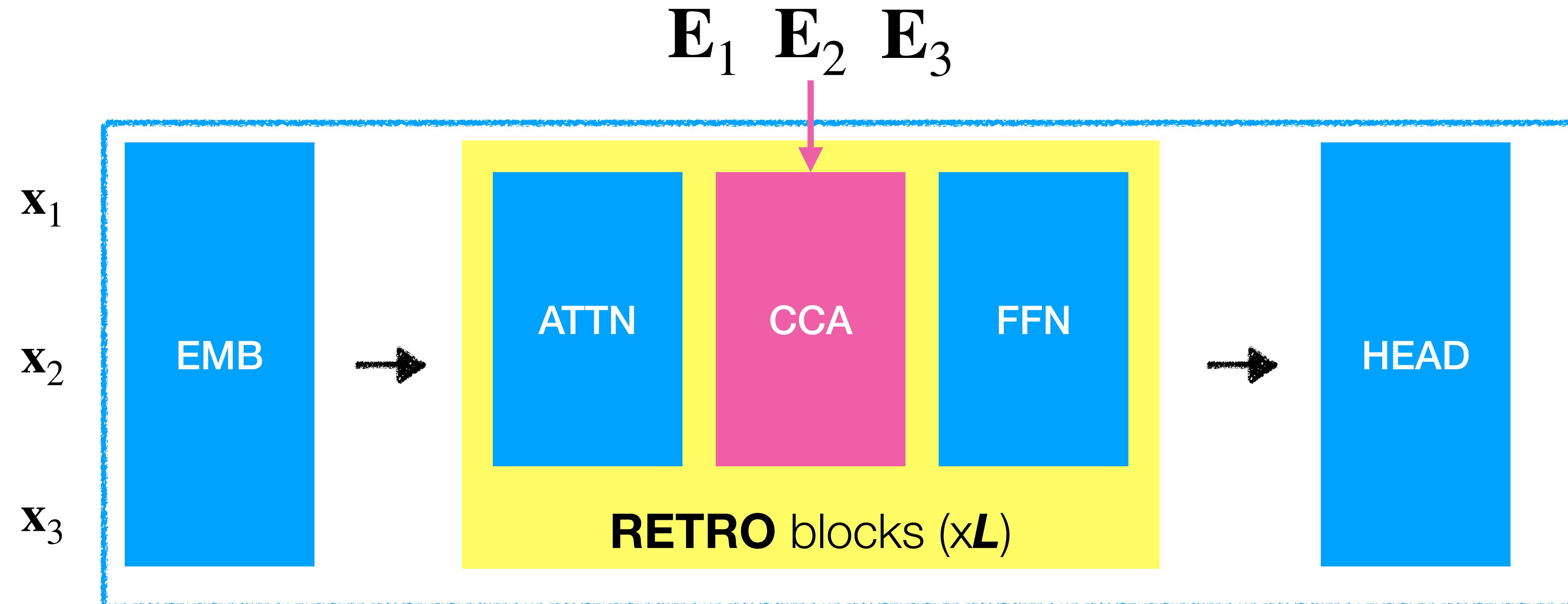
Chunked Cross Attention



Regular decoder

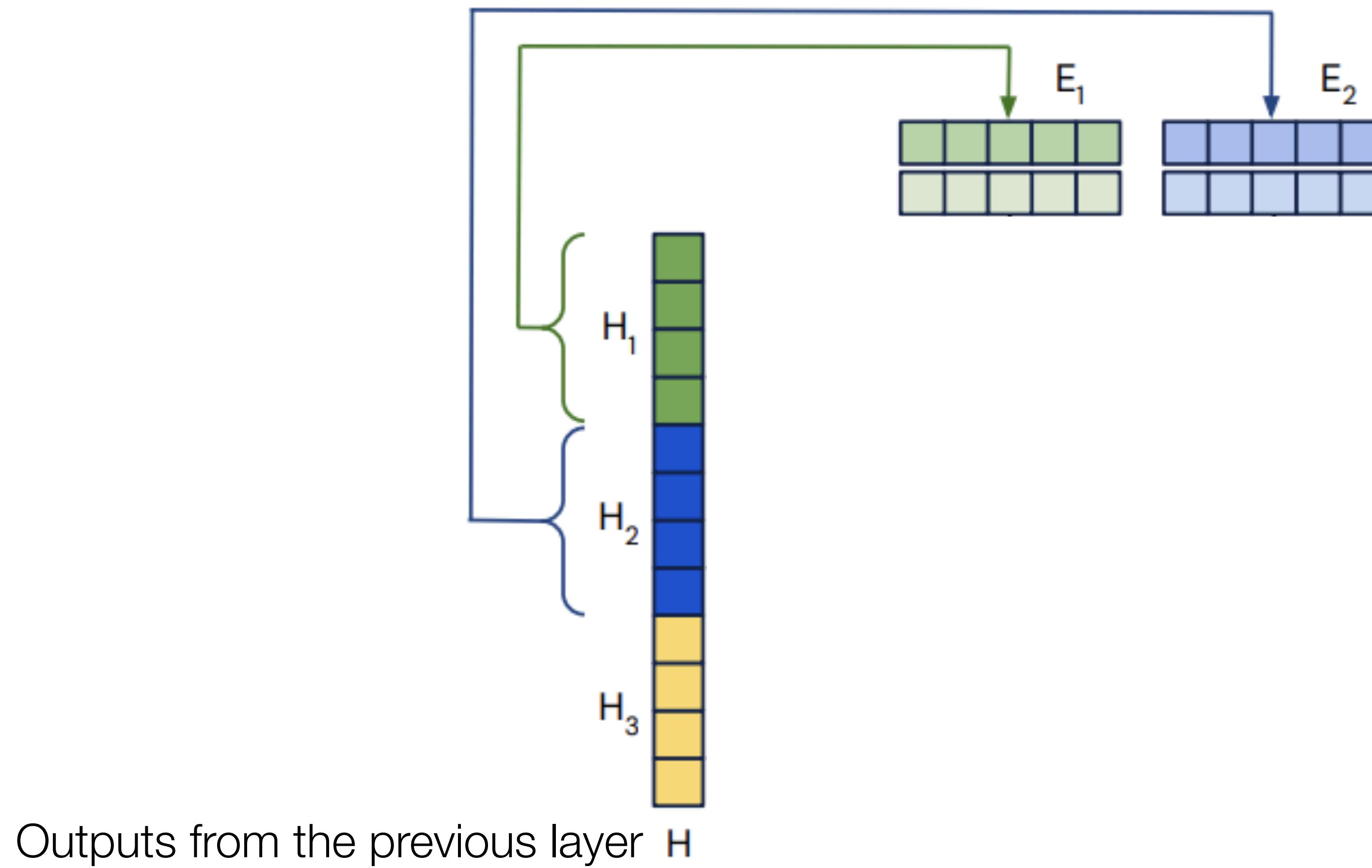


Decoder in RETRO

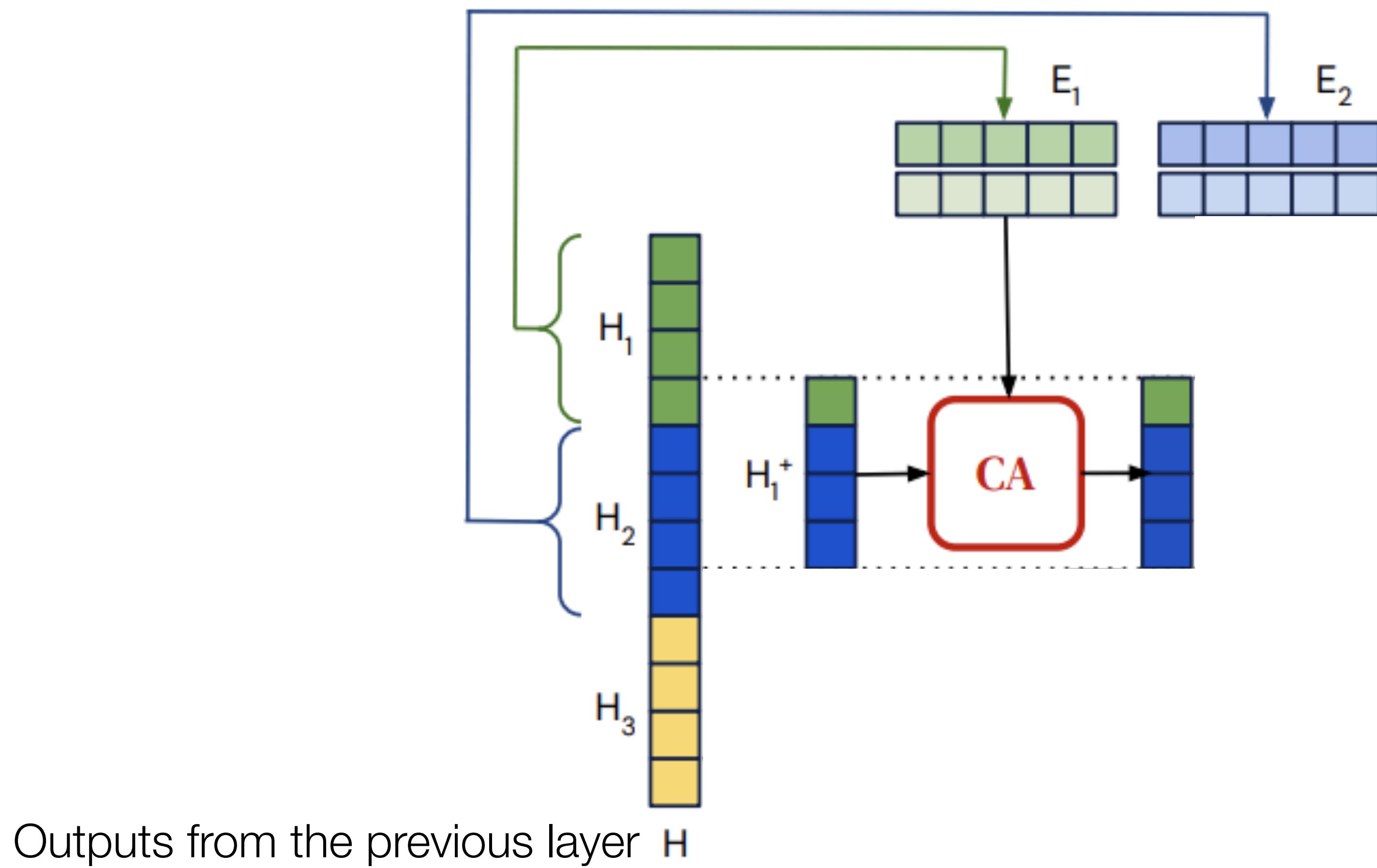


Chunked Cross Attention (CCA)

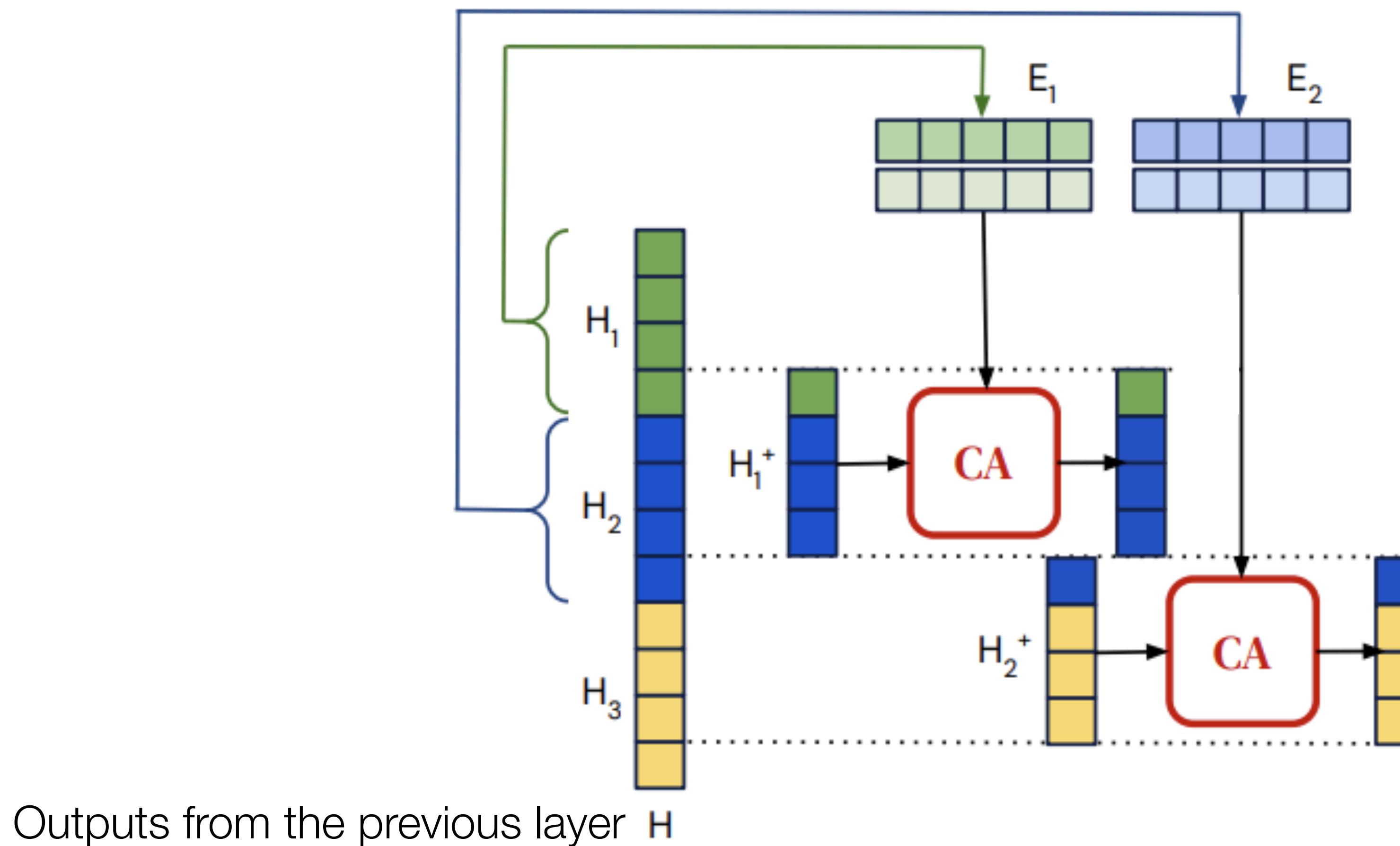
Chunked Cross Attention



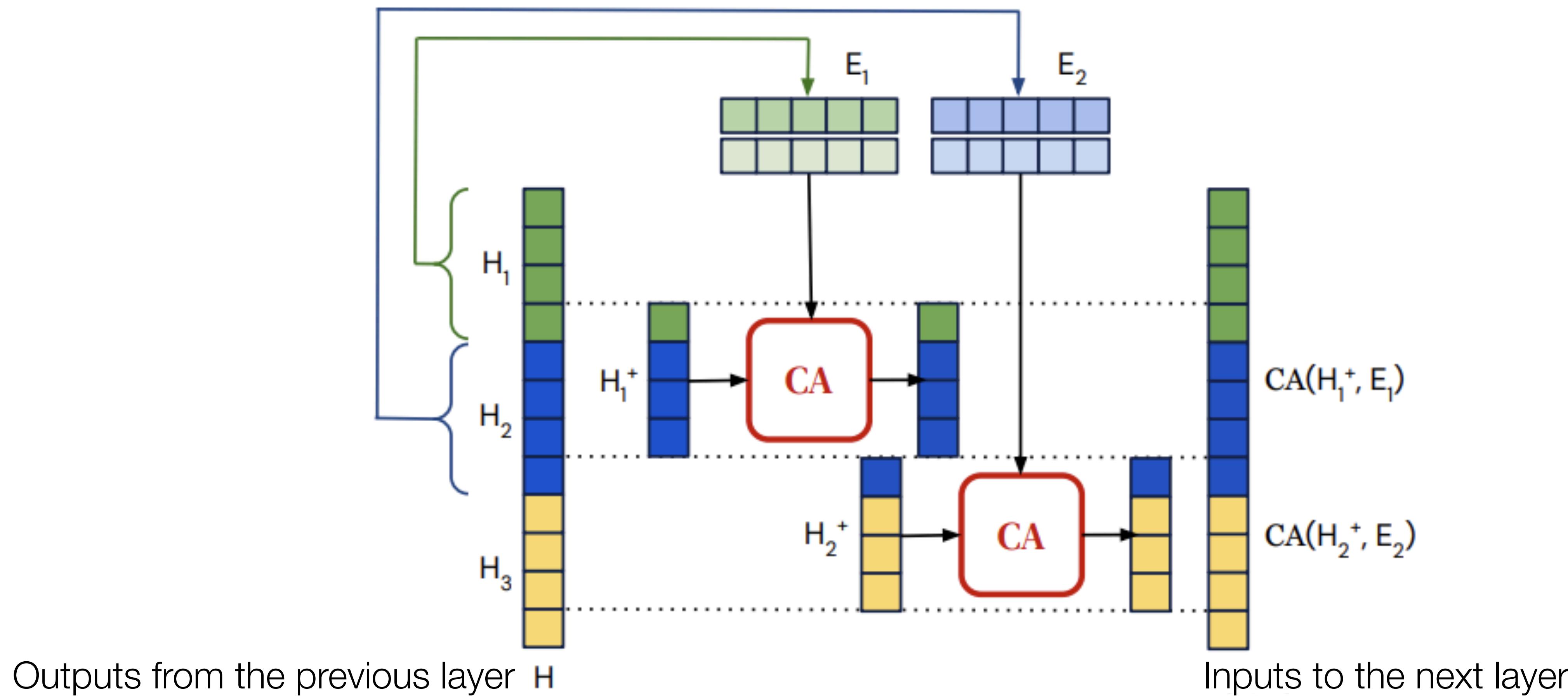
Chunked Cross Attention



Chunked Cross Attention



Chunked Cross Attention



RETRO

What to retrieve?

- **Chunks ✓**
- Tokens
- Others

How to use retrieval?

- Input layer
- **Intermediate layers ✓**
- Output layer

When to retrieve?

- Once
- **Every n tokens ($n > 1$) ✓**
- Every token

Pros and Cons of RETRO

	What do retrieve?	How to use retrieval?	When to retrieve?
REALM (Guu et al 2020)	Text chunks	Input layer	Once
Retrieve-in-context LM (Shi et al 2023, Ram et al 2023)	Text chunks	Input layer	Every n tokens
RETRO (Borgeaud et al. 2021)	Text chunks	Input layer Intermediate layers	Every n tokens



Can use many blocks, more frequently, more efficiently



Additional complexity; Can't be used without training

Three representative architectures

What: Text chunks
How: Input

More...
See also: [val-lm.](#)

REALM (Guu et al., 2020)

What: Text chunks
How: Intermediate

RETRO (Borgeaud et al., 2021)

Our position paper (Asai et al., 2024; https://akariasai.github.io/assets/pdf/ralm_position.pdf)

What: Tokens
How: Output

kNN-LM (Khandelwal et al., 2020)

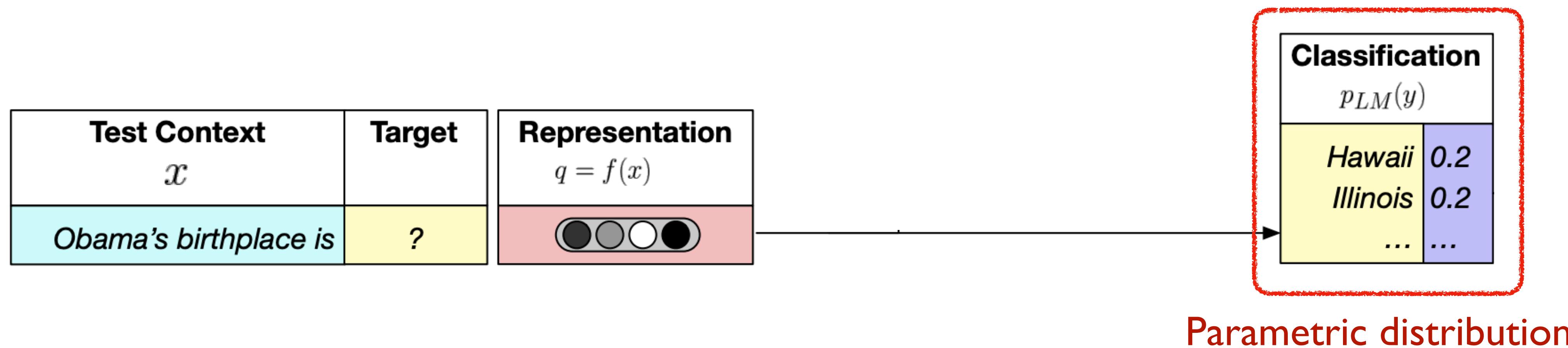
kNN-LM (Khandelwal et al. 2020)

- ✓ A different way of using retrieval, where the LM outputs a nonparametric distribution over every token in the data.
- ✓ Can be seen as an incorporation in the “output” layer

kNN-LM (Khandelwal et al. 2020)

Test Context	Target
x	
<i>Obama's birthplace is</i>	?

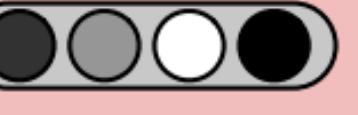
kNN-LM (Khandelwal et al. 2020)



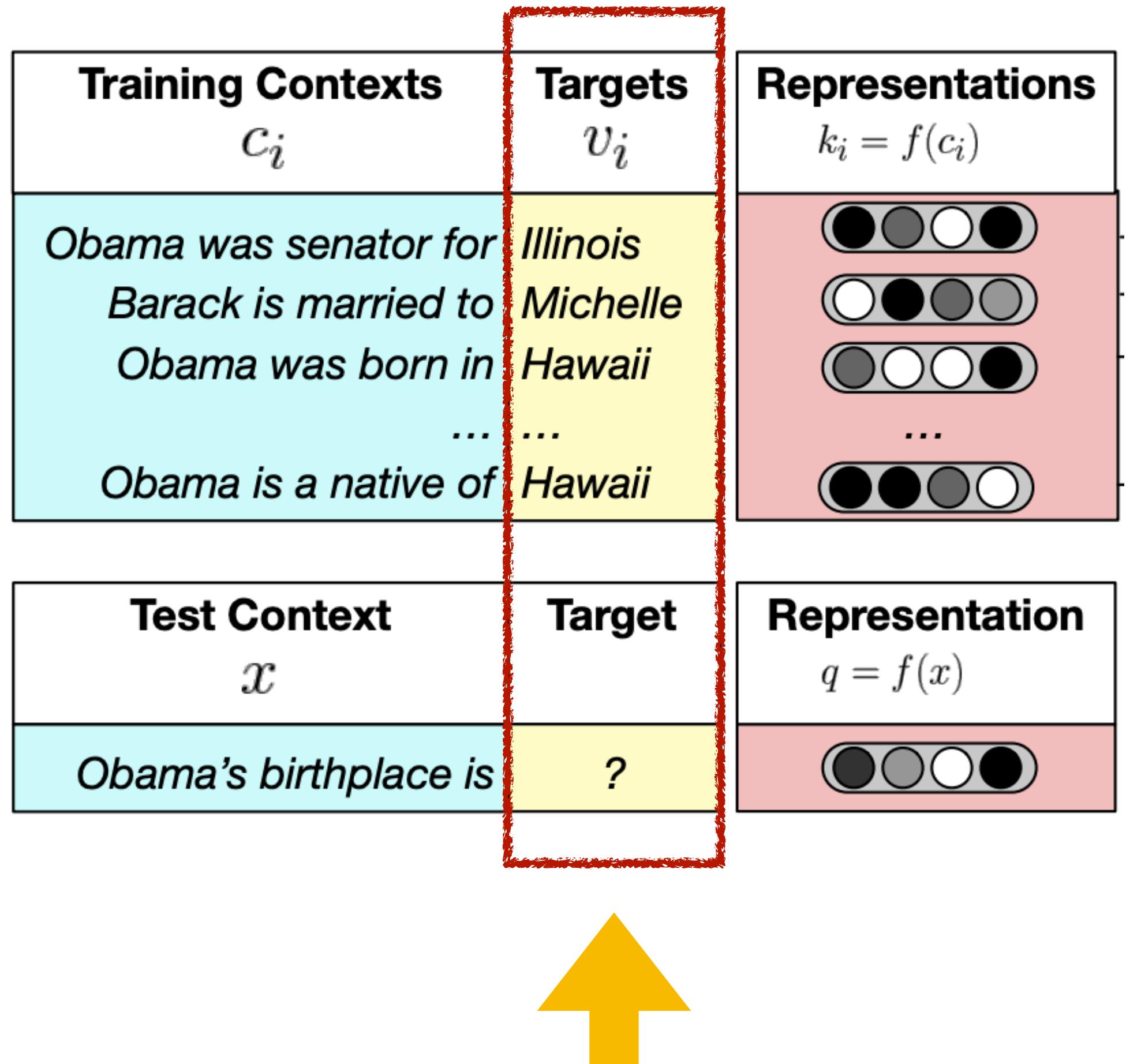
kNN-LM (Khandelwal et al. 2020)

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

... 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	Target	Representation
x		$q = f(x)$
<i>Obama's birthplace is</i>	?	

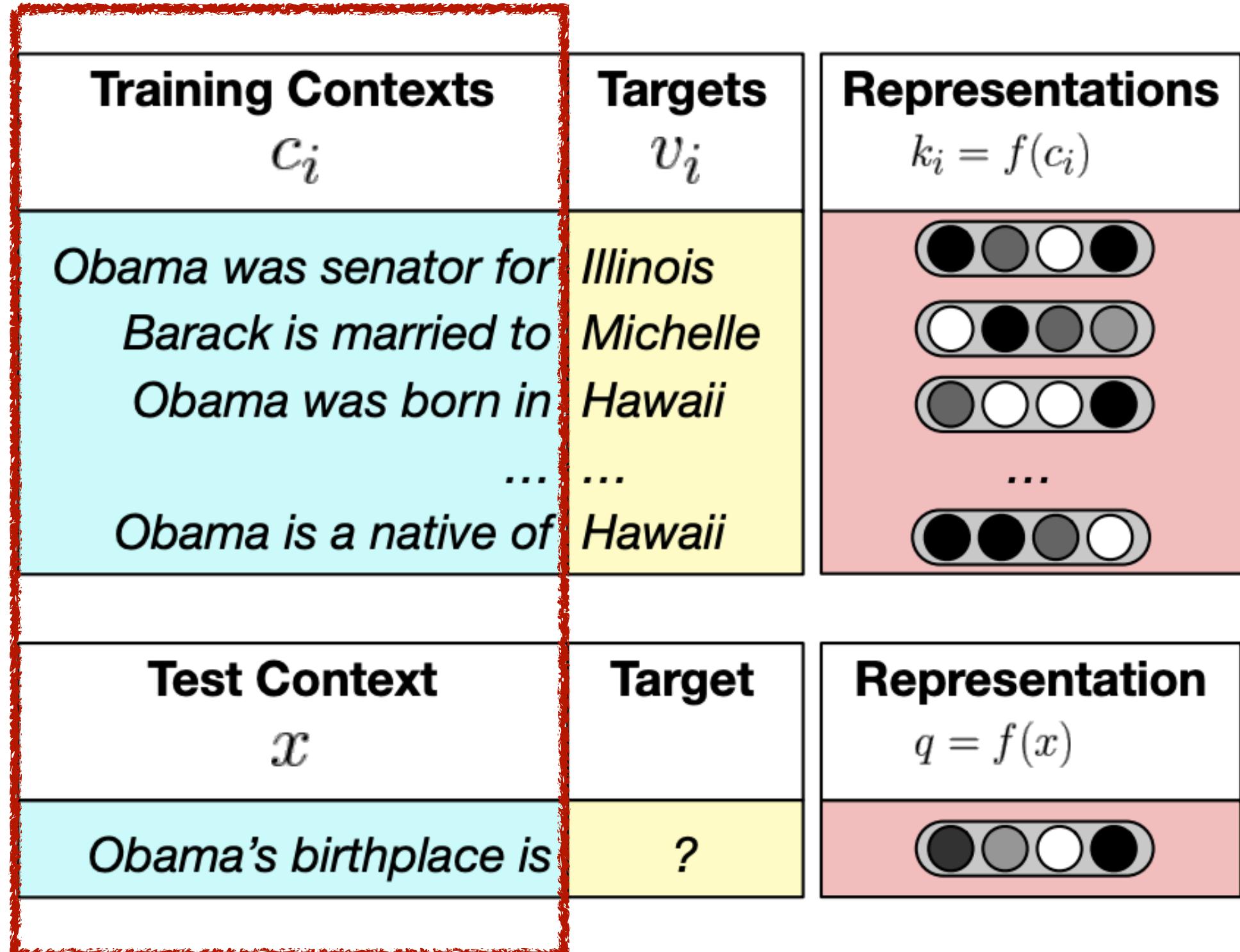
kNN-LM (Khandelwal et al. 2020)



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

kNN-LM (Khandelwal et al. 2020)

The size of the datastore = # of tokens in the corpus ($> 1B$)

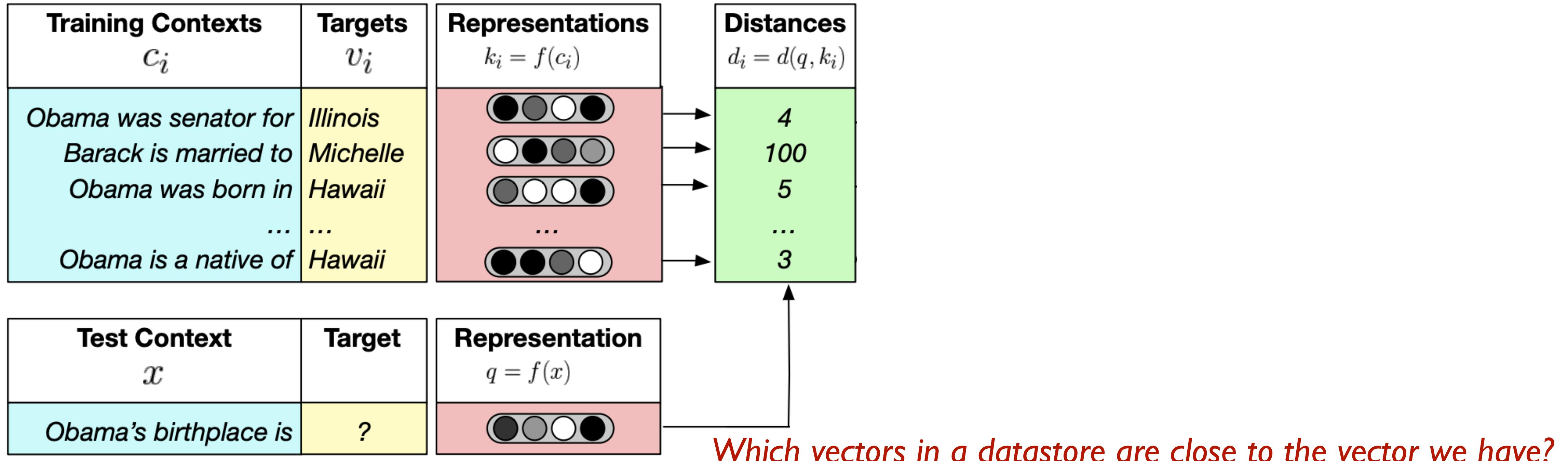


Which tokens in a datastore are close to the next token?

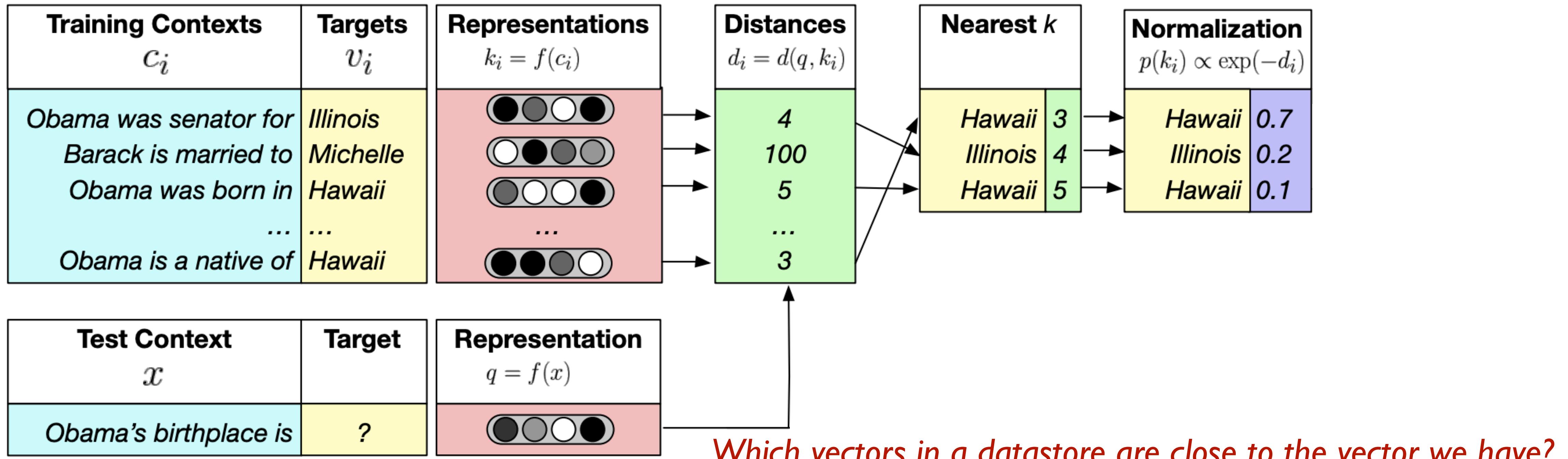
=

Which prefixes in a datastore are close to the prefix we have?

kNN-LM (Khandelwal et al. 2020)



kNN-LM (Khandelwal et al. 2020)

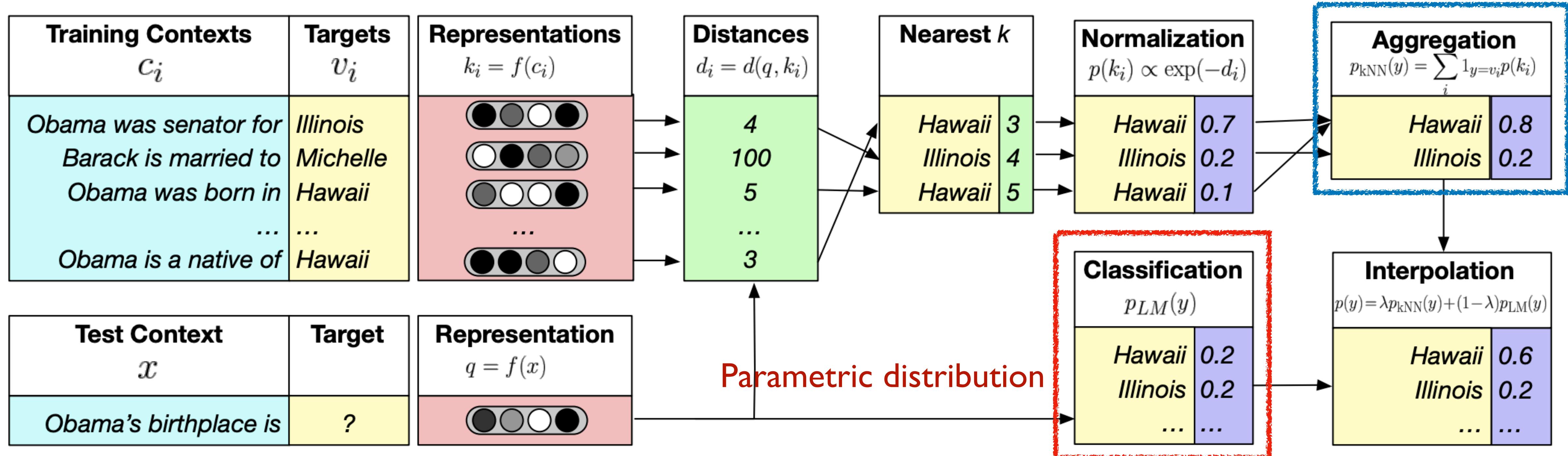


$$P_{k\text{NN}}(y|x) \propto \sum_{(k,v) \in \mathcal{D}} \mathbb{I}[v = y] \text{sim}(k, x)$$

$$\text{sim}(k, x) = \exp(-d(\text{Enc}(k), \text{Enc}(x)))$$

kNN-LM (Khandelwal et al. 2020)

Nonparametric distribution



λ : hyperparameter

$$P_{kNN-LM}(y|x) = (1 - \lambda) \underline{P_{LM}(y|x)} + \lambda \underline{P_{kNN}(y|x)}$$

kNN-LM Example

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



Dense vector space



10/10, would buy this **cheap**

... affordable

... nice

... good

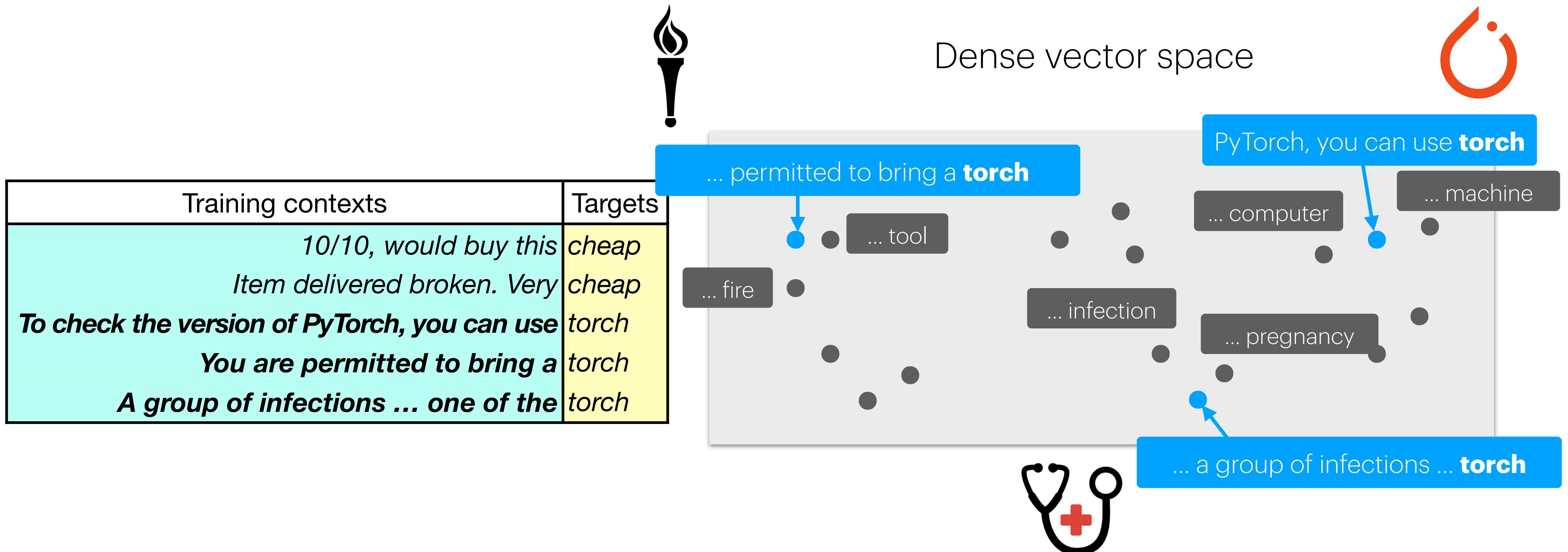
... bad

... poor

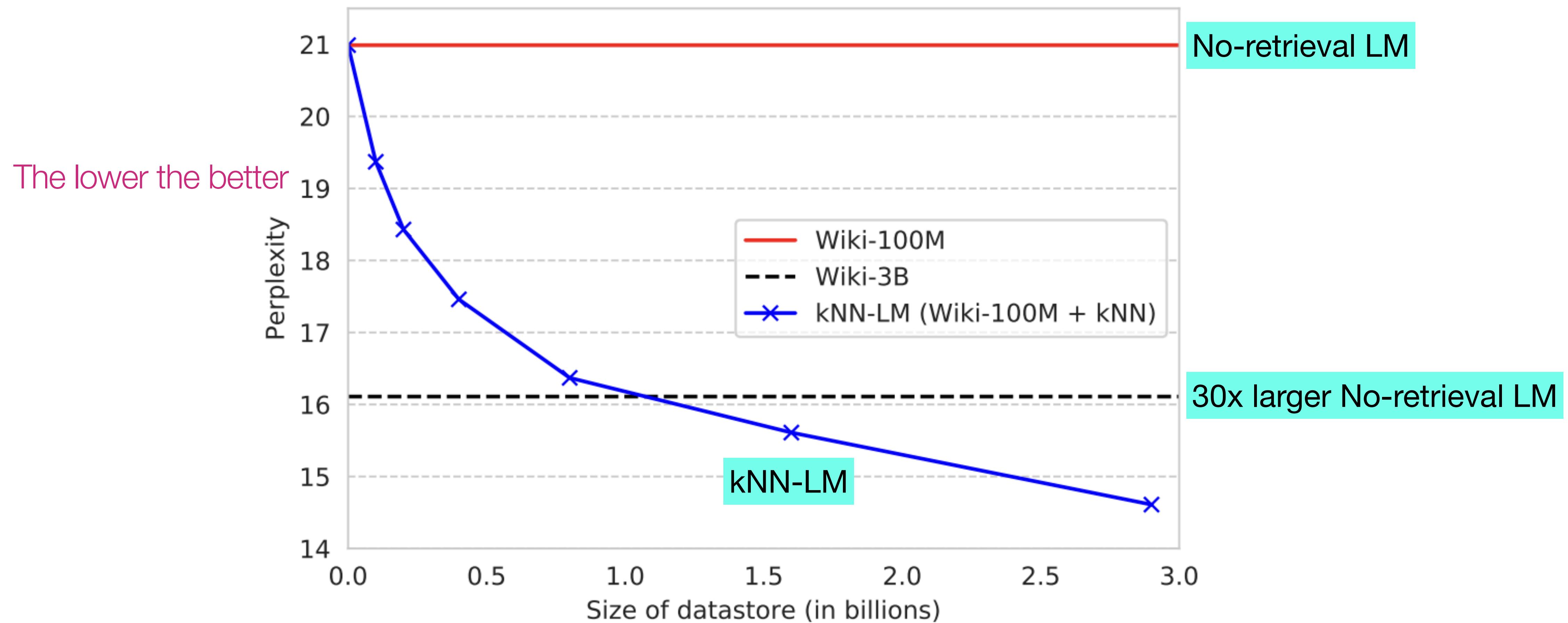
... terrible

Item delivered broken. Very **cheap**

kNN-LM Example



kNN-LM - results



Outperforms no-retrieval LM

Better with bigger datastore

kNN-LM

What to retrieve?

- Chunks
- **Tokens ✓**
- Others

How to use retrieval?

- Input layer
- Intermediate layers
- **Output layer ✓**

When to retrieve?

- Once
- Every n tokens ($n > 1$)
- **Every token ✓**



More fine-grained; Can be better at rare patterns & out-of-domain
Can be very efficient (as long as kNN search is fast)



Datastore is expensive in space: given the same data, # text chunks vs. # tokens

Three representative architectures

What: Text chunks
How: Input

REALM (Guu et al., 2020)

What: Text chunks
How: Intermediate

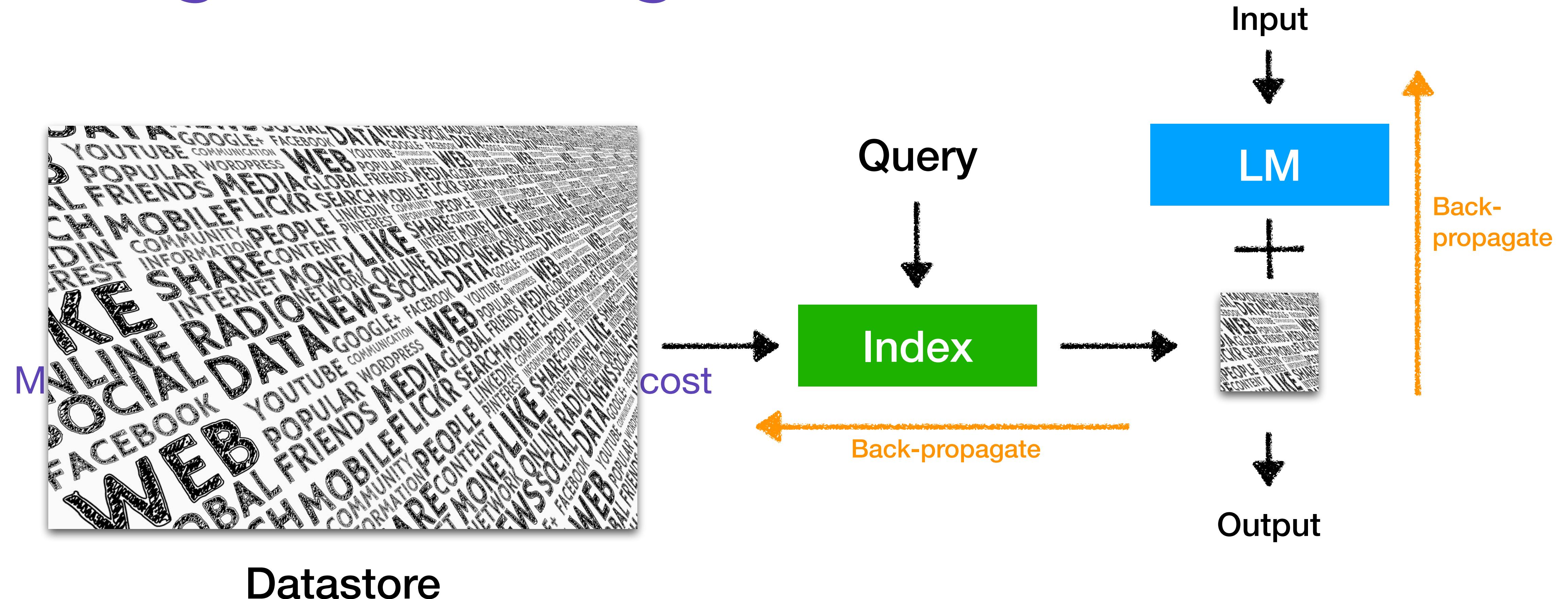
RETRO (Borgeaud et al., 2021)

What: Tokens
How: Output

kNN-LM (Khandelwal et al., 2020)

Training of RAGs

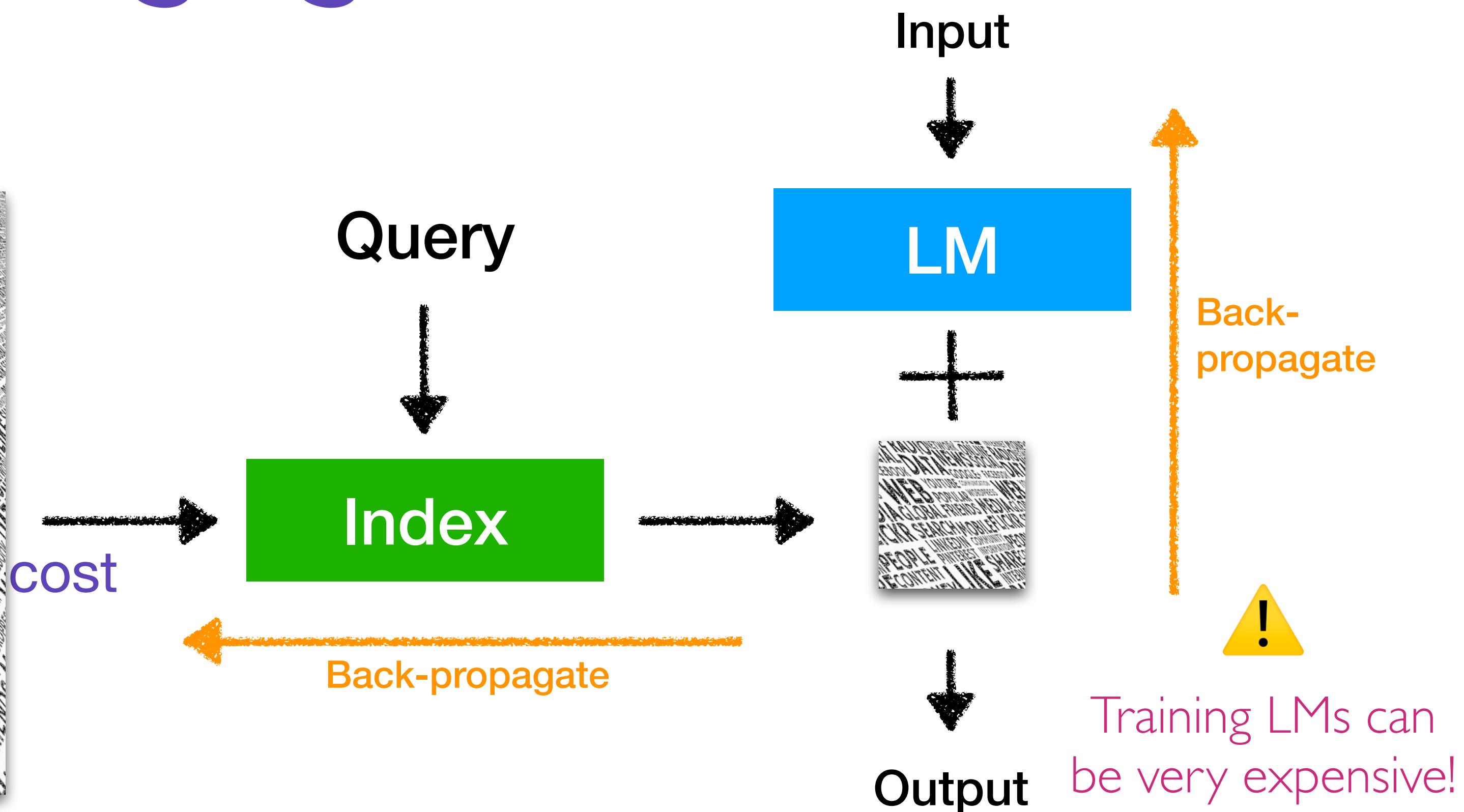
Training retrieval-augmented LMs



Why is training challenging?



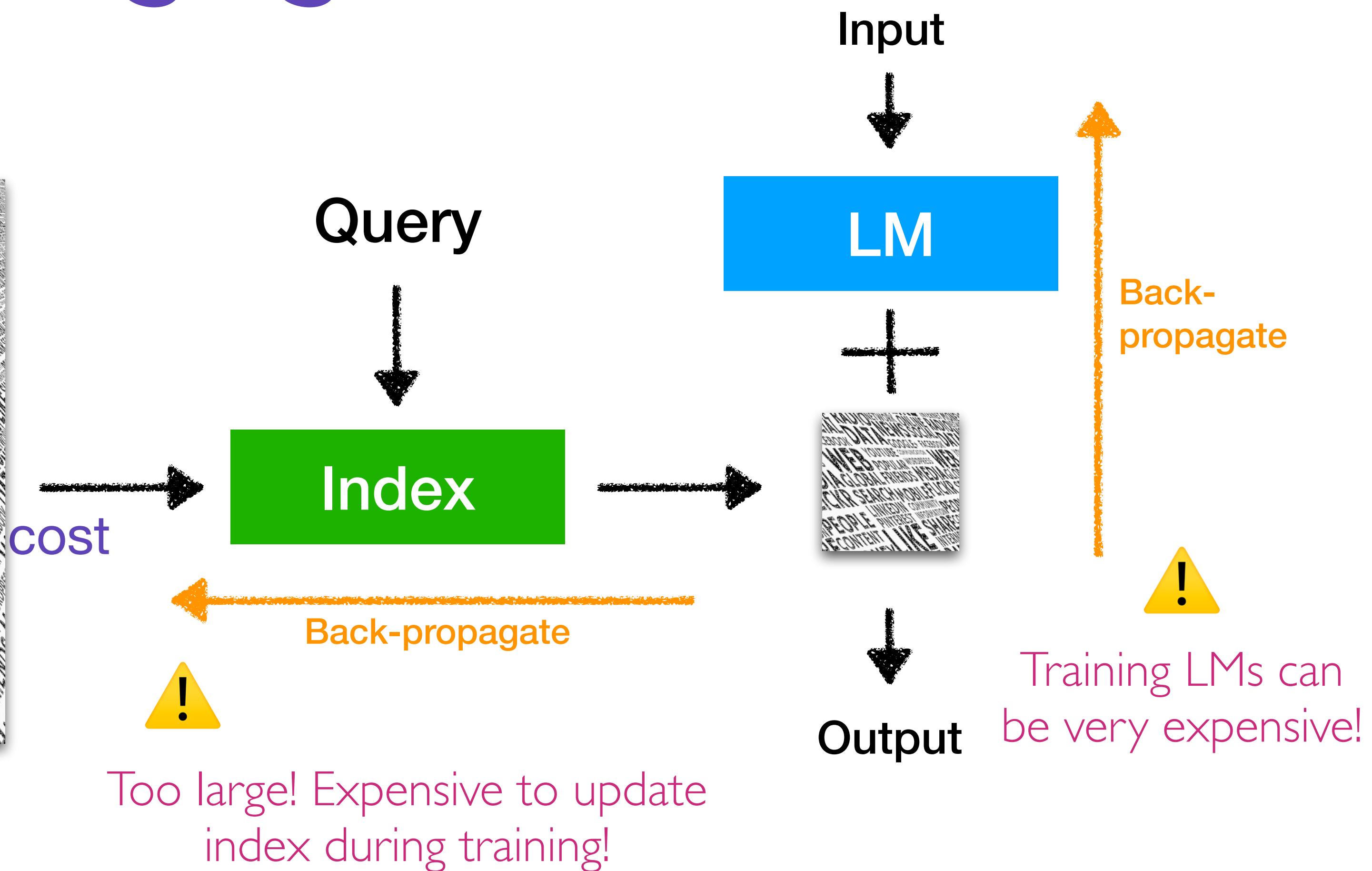
Datastore



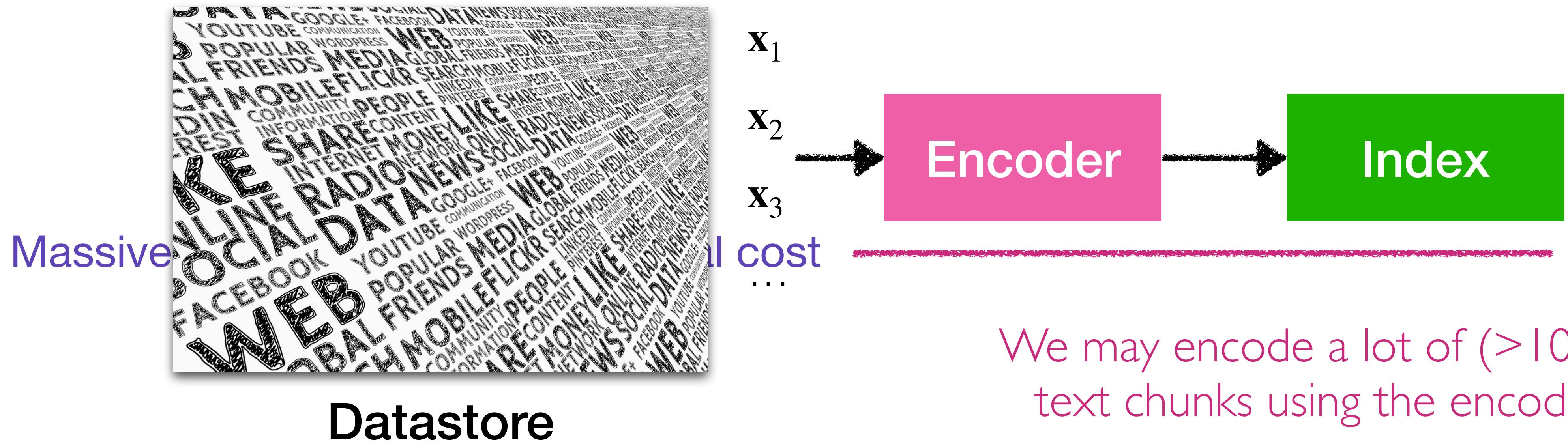
Why is training challenging?



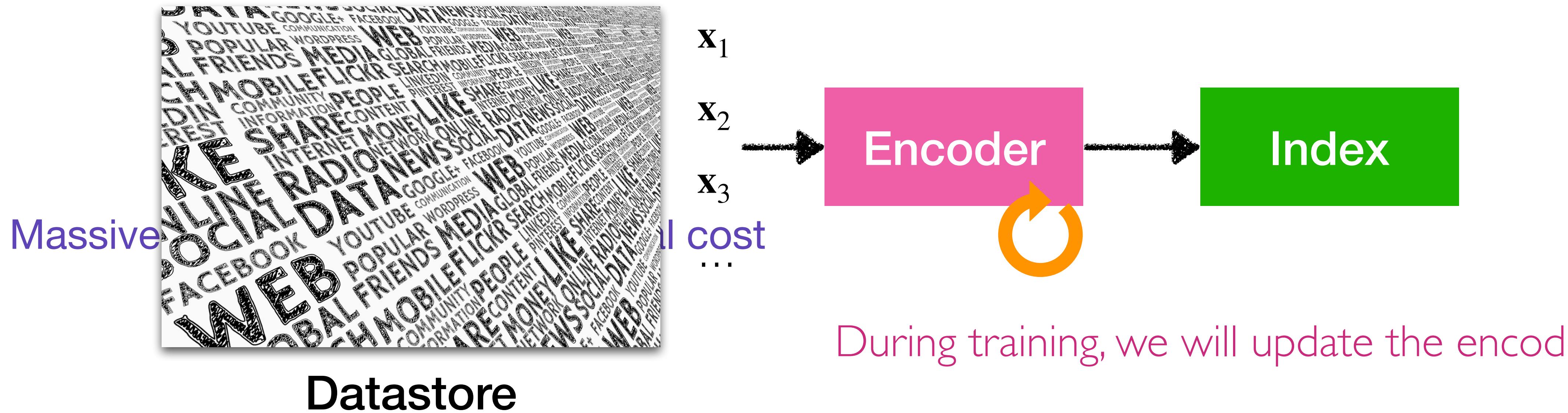
Datastore



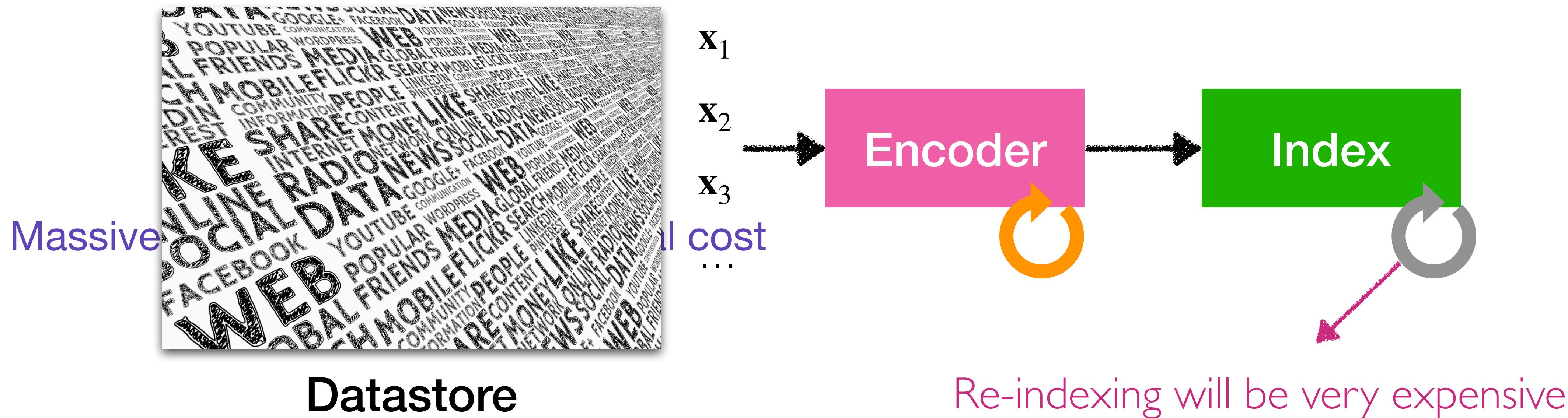
Challenges of updating retrieval models



Challenges of updating retrieval models



Challenges of updating retrieval models



Training methods for retrieval-augmented LMs

Independent training

Sequential training

Joint training w/ asynchronous index update

Joint training w/ in-batch approximation

Training methods for retrieval-augmented LMs

Independent training

Sequential training

Joint training w/ asynchronous index update

Joint training w/ in-batch approximation

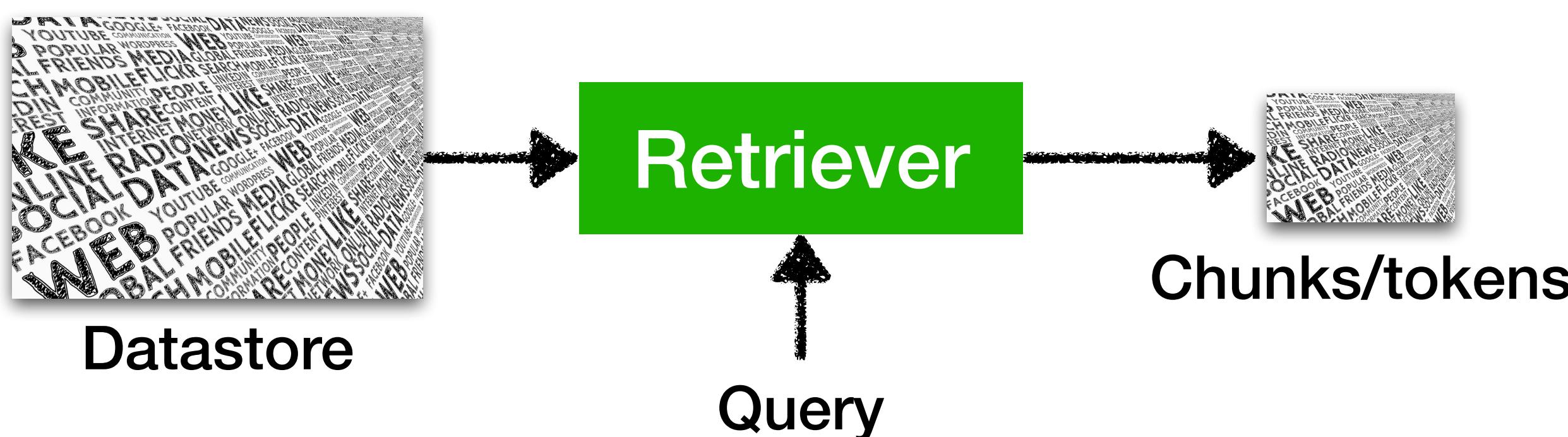
Independent training

Retrieval models and language models are trained **independently**

- Training language models



- Training retrieval models



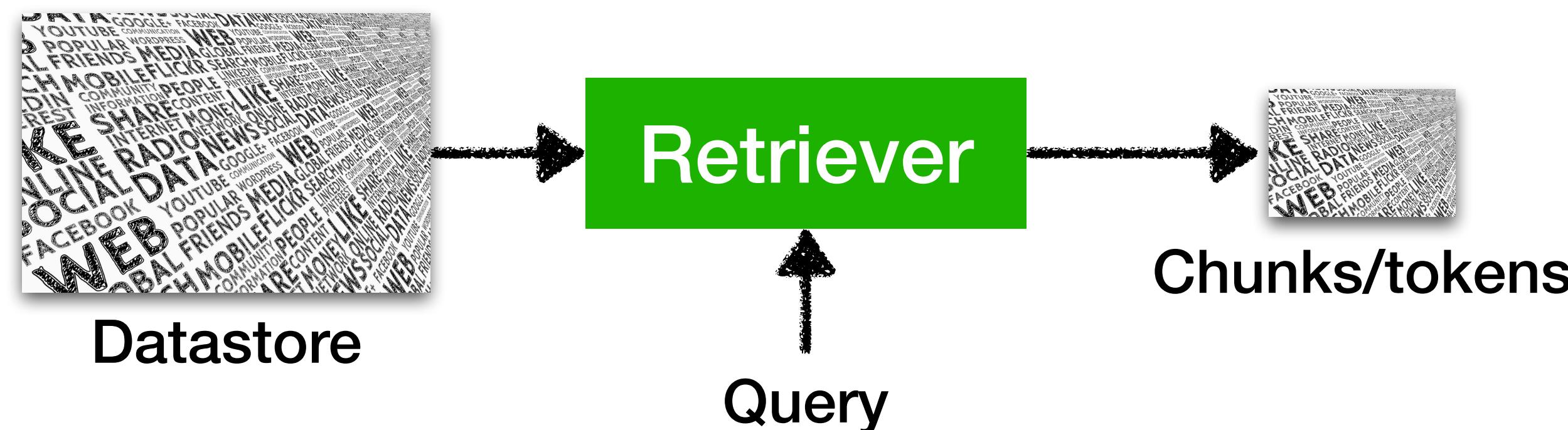
Independent training

Retrieval models and language models are trained **independently**

- Training language models



- Training retrieval models



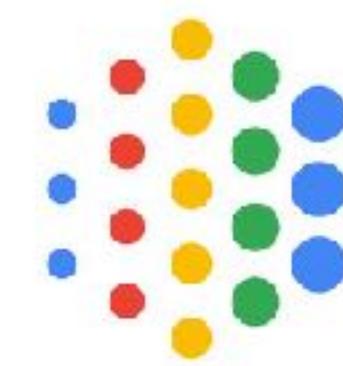
Training LMs



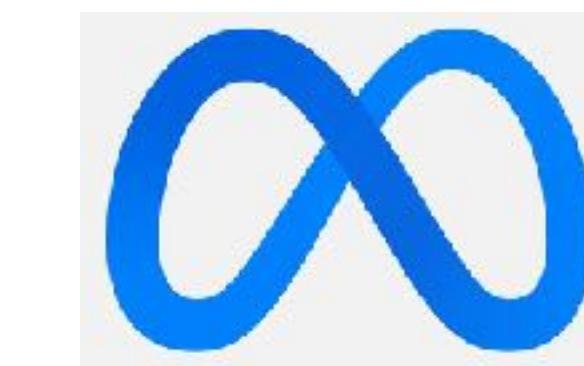
Minimize $-\log P_{\text{LM}}(y | x)$



GPT



PaLM



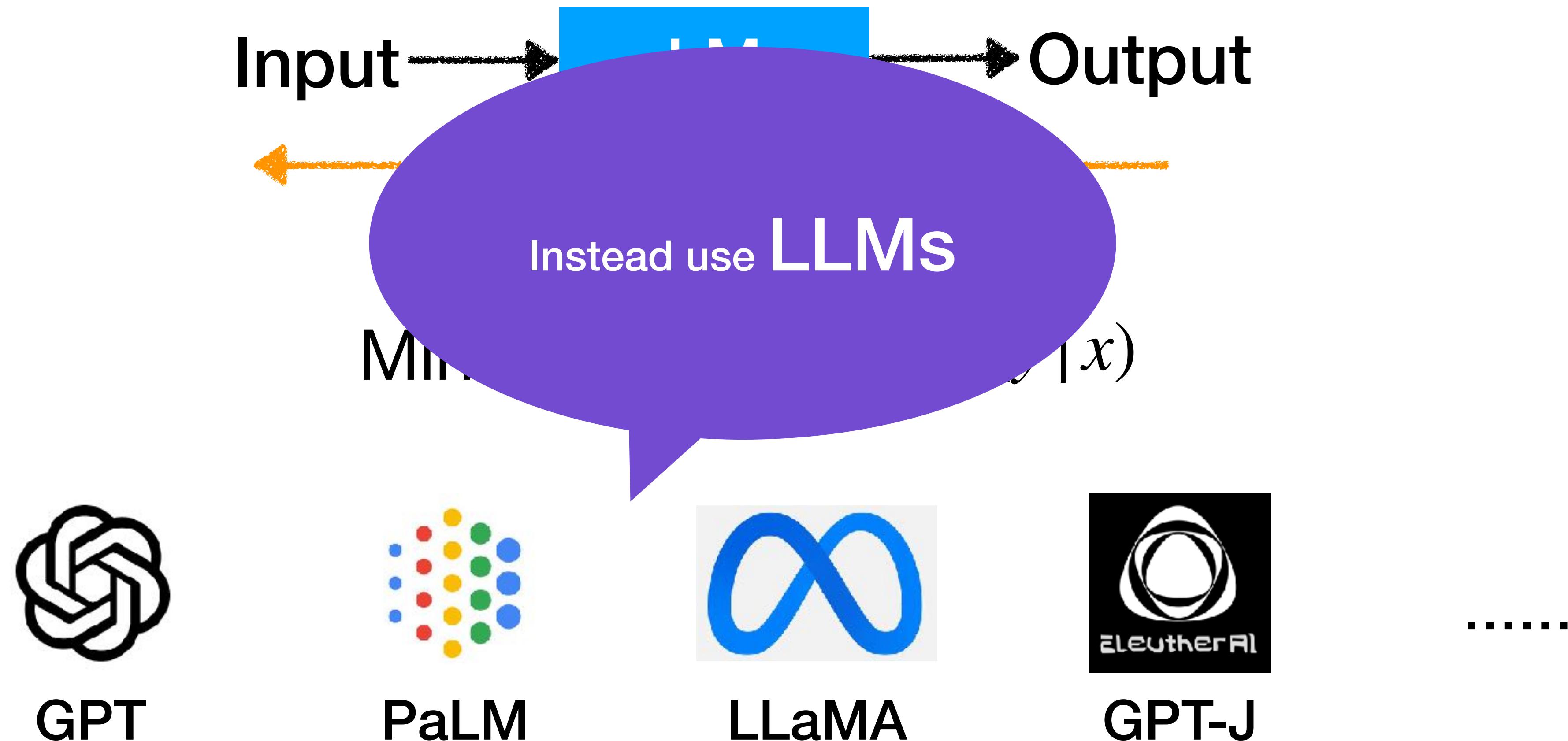
LLaMA



GPT-J

.....

Training LMs



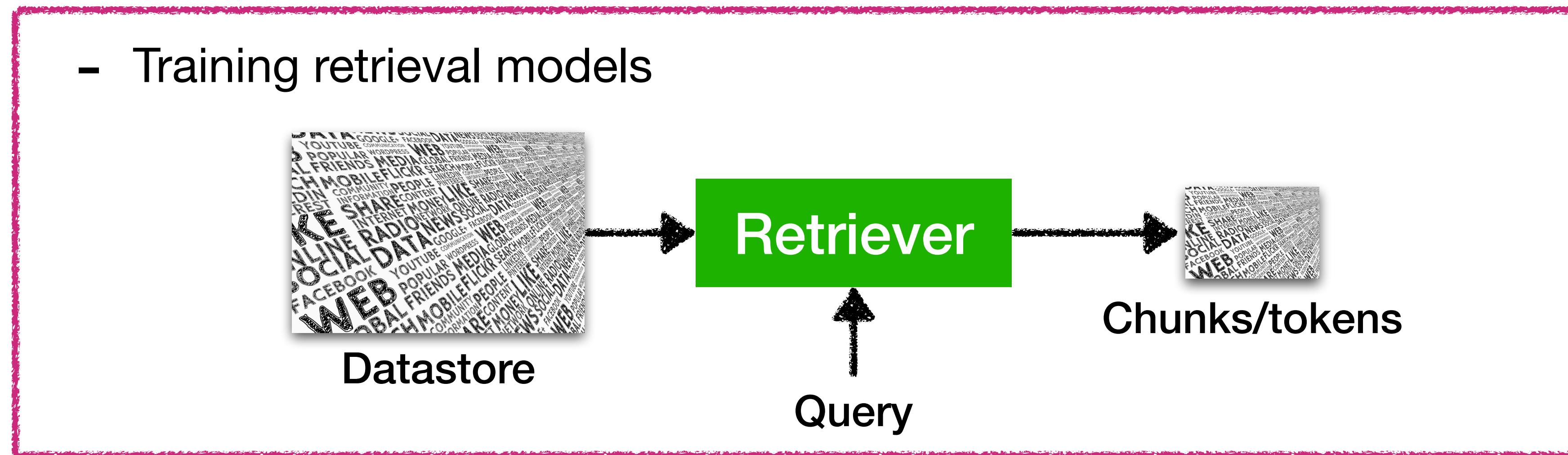
Independent training

Retrieval models and language models are trained **independently**

- Training language models



- Training retrieval models



Sparse retrieval models: TF-IDF / BM25

In 1997, Apple merged with NeXT,
and Steve Jobs became CEO of ...

Jobs returned to Apple as CEO
after the company's acquisition ...

Text chunks

[0, 0, 0.4, 0, 0.8, 0.7, ...]

[0, 1.2, 0.4, 0, 0.8, 0, ...]

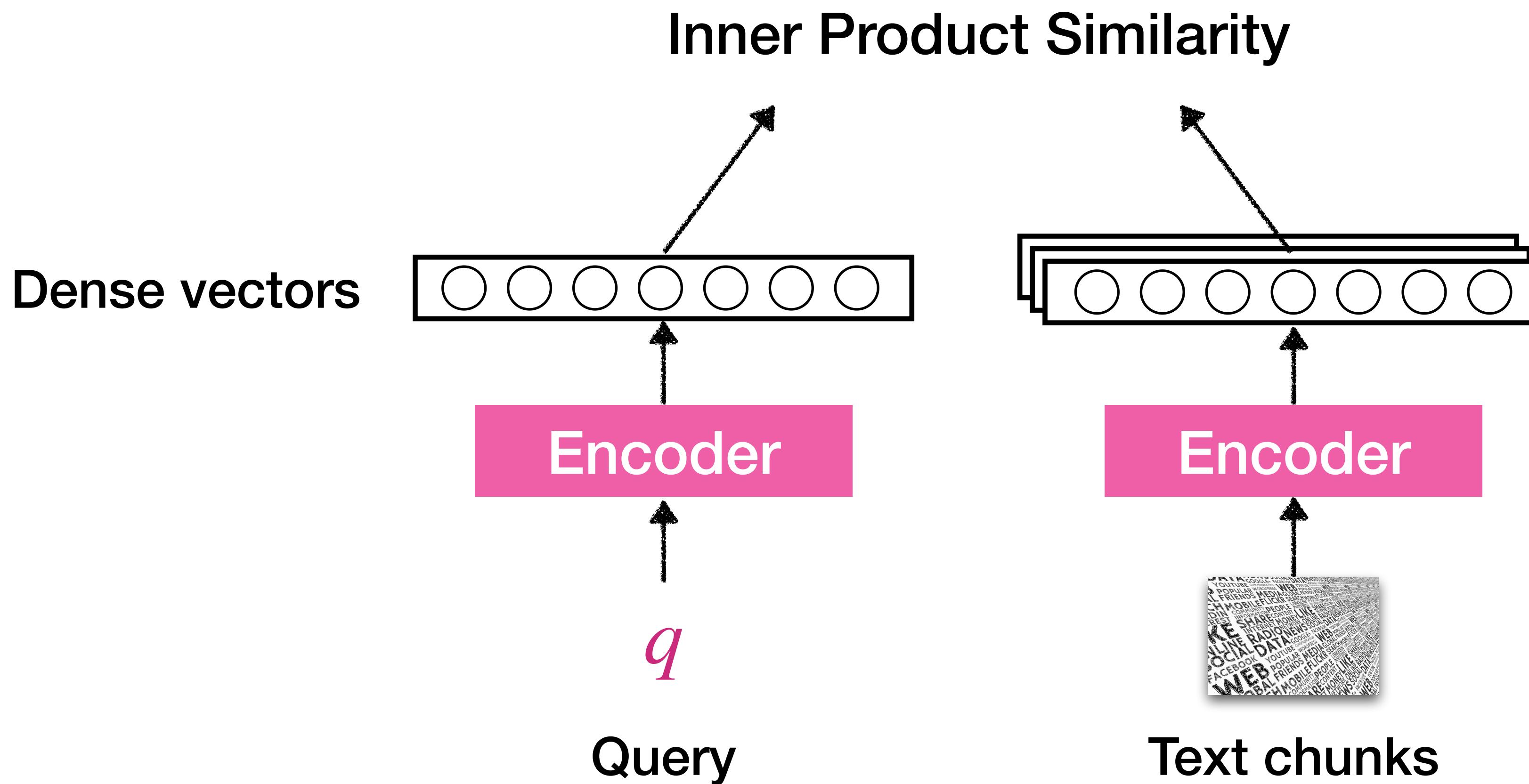
Sparse vectors



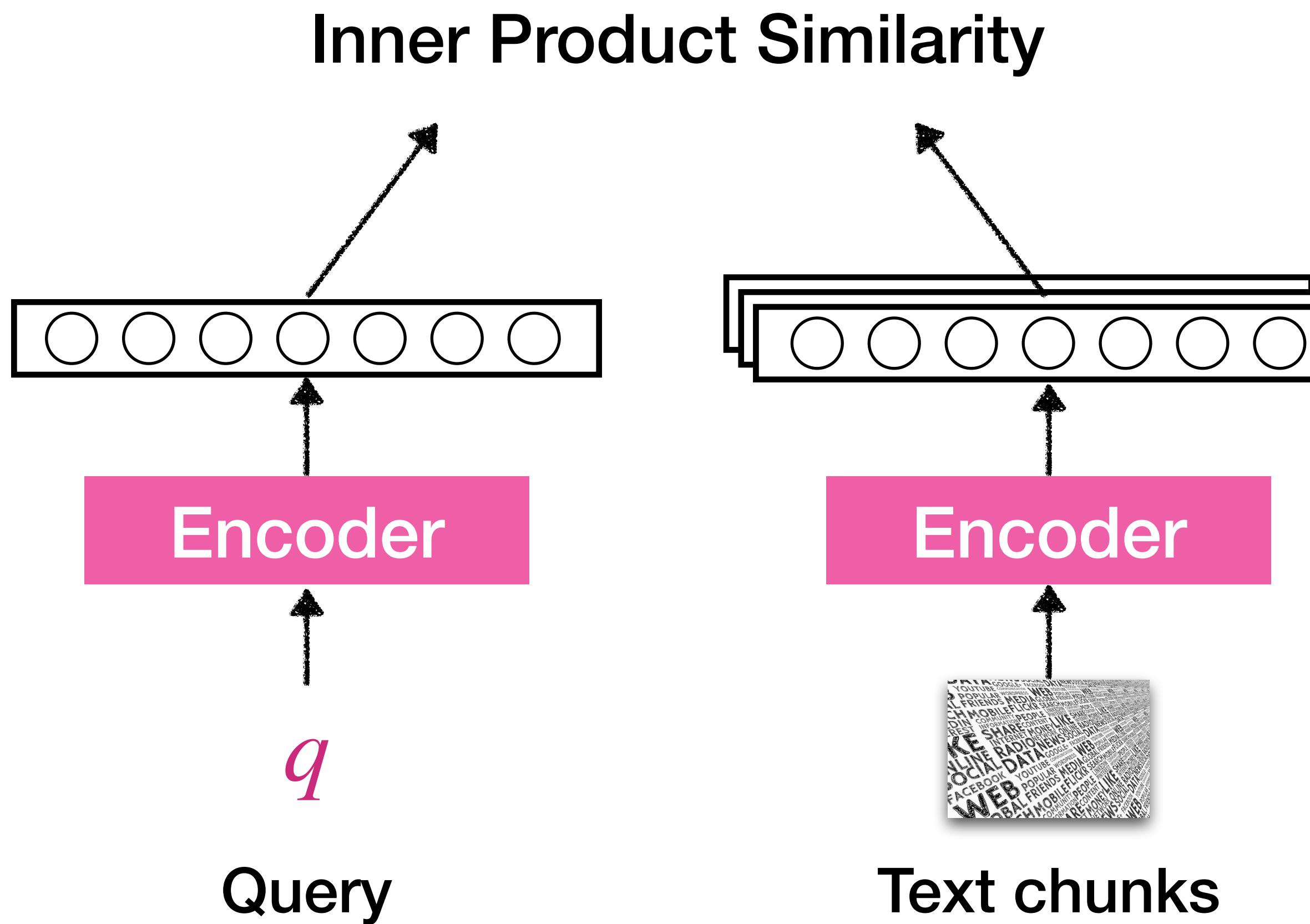
No training needed!

Ramos. Using TF-IDF to Determine Word Relevance in Document Queries. 2023.
Robertson and Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond.

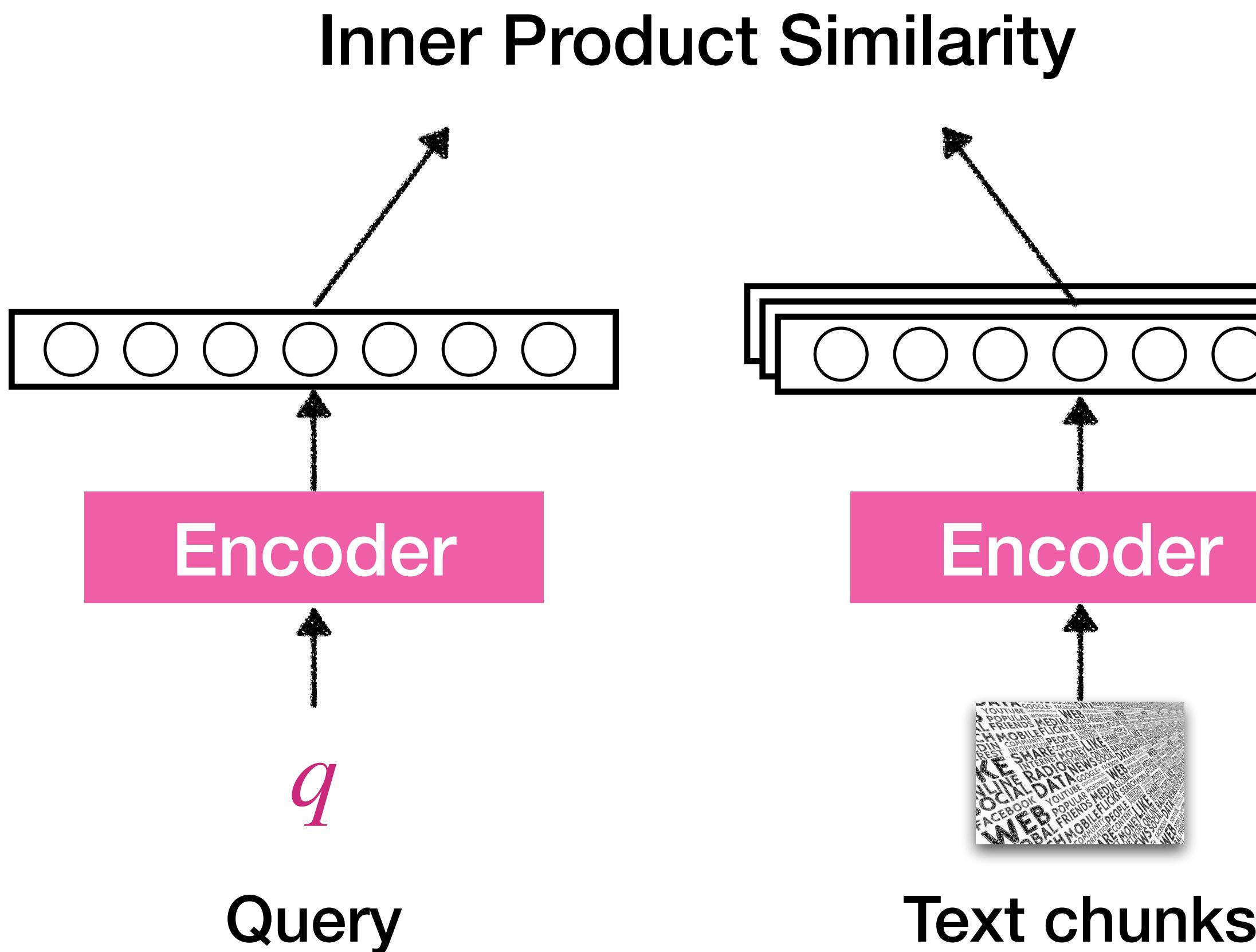
Dense retrieval models: DPR (Karpukhin et al. 2020)



Training dense retrieval models: DPR

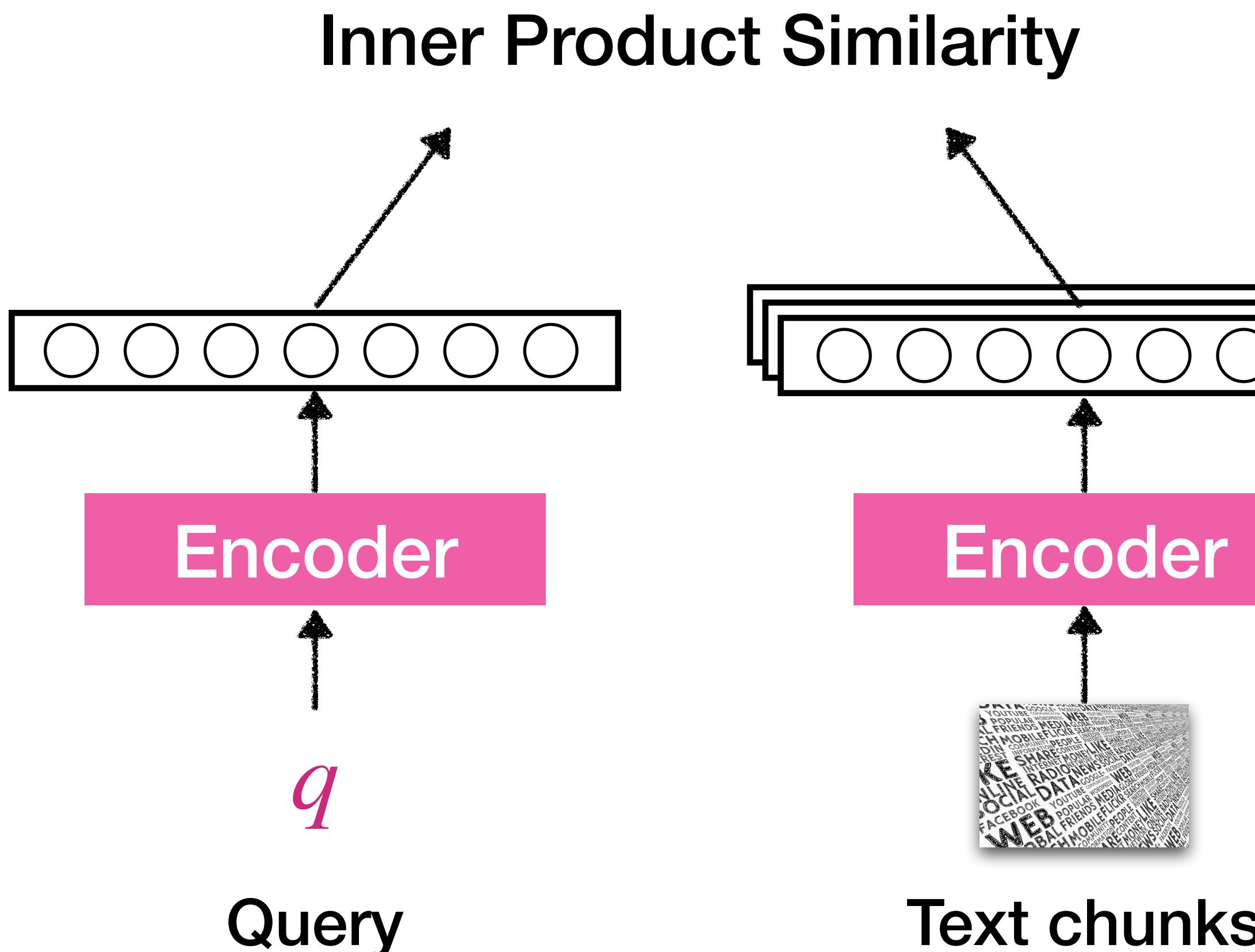


Training dense retrieval models: DPR



$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

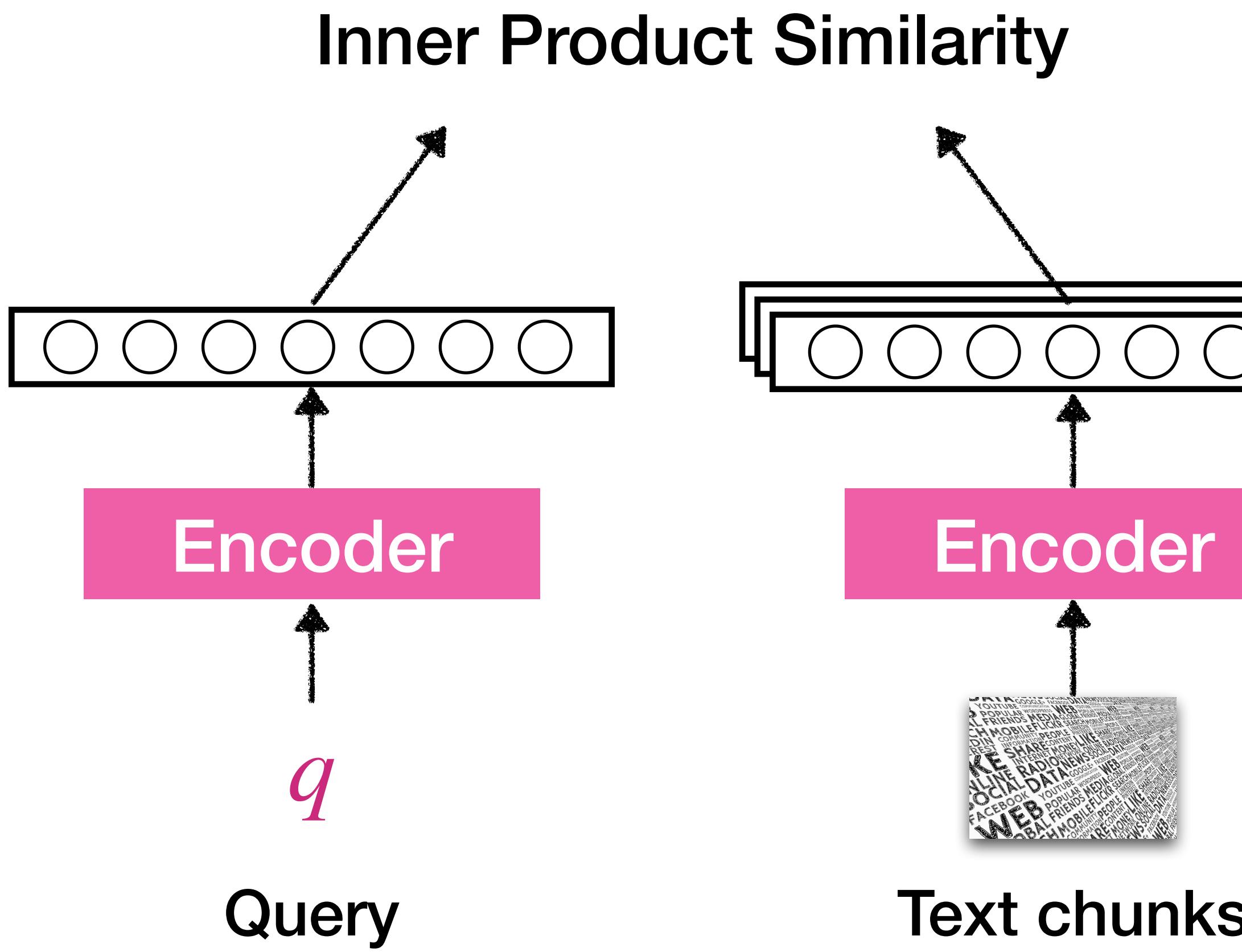
Training dense retrieval models: DPR



$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

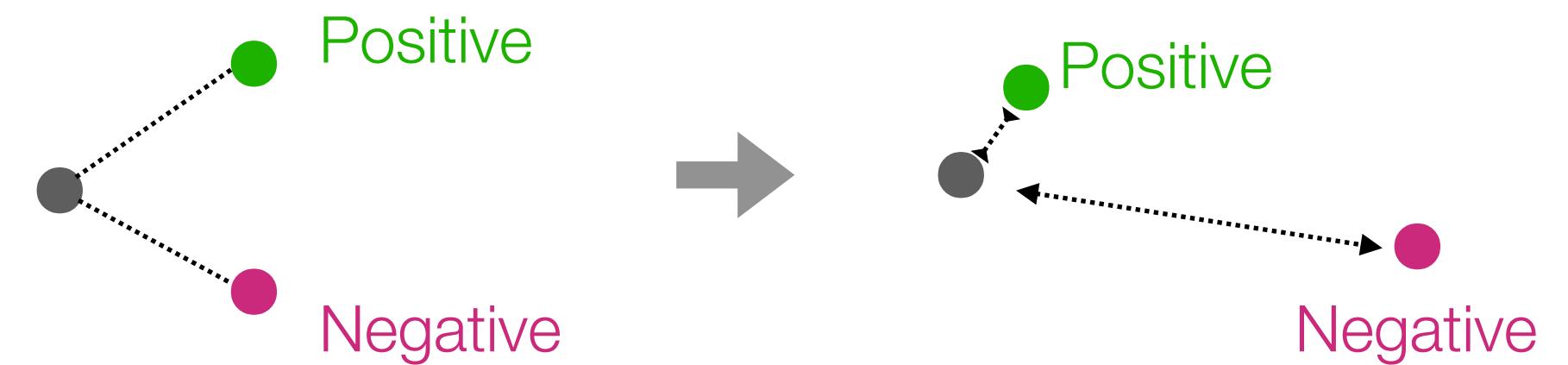
Contrastive learning

Training dense retrieval models: DPR

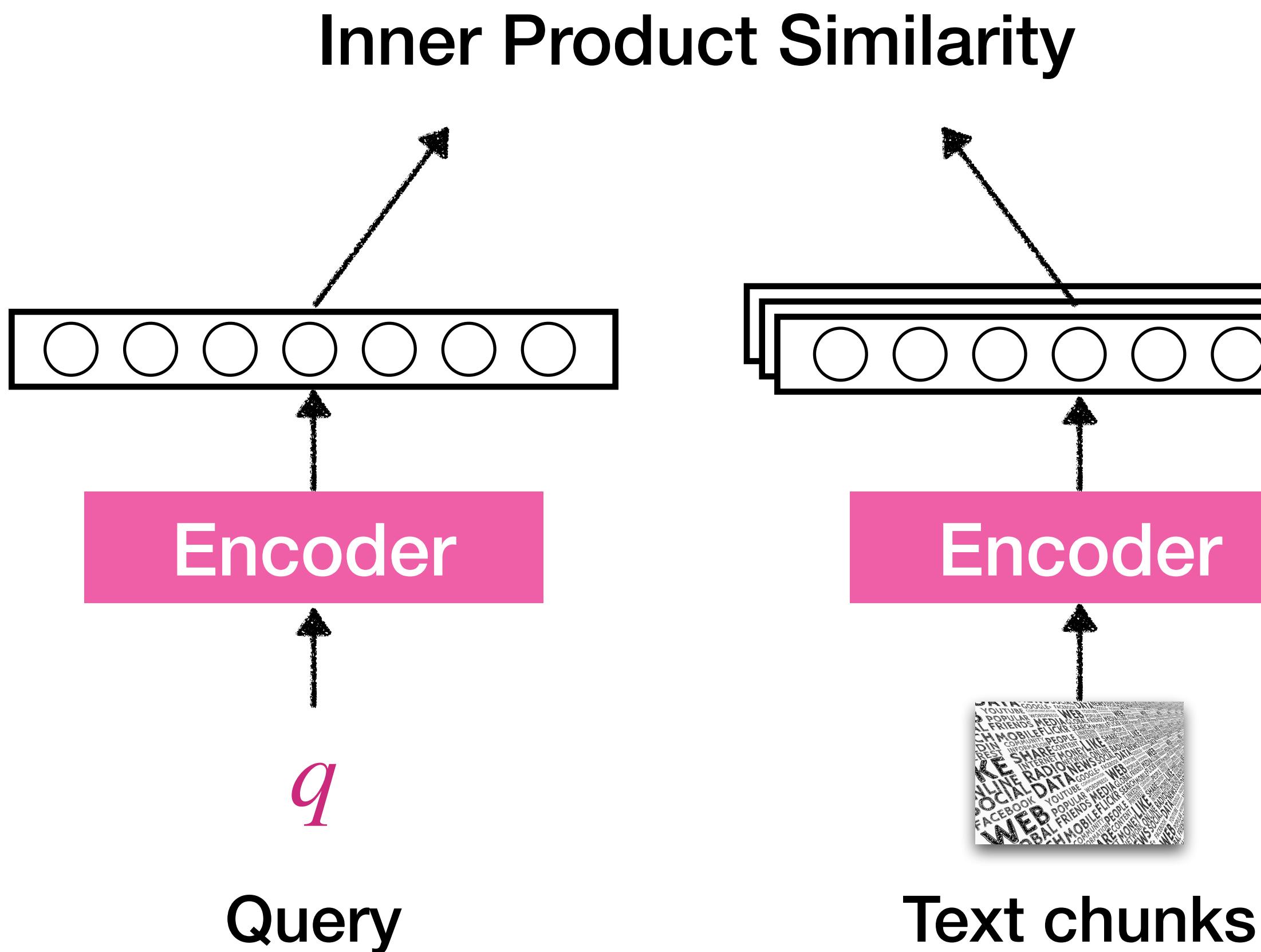


$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Contrastive learning



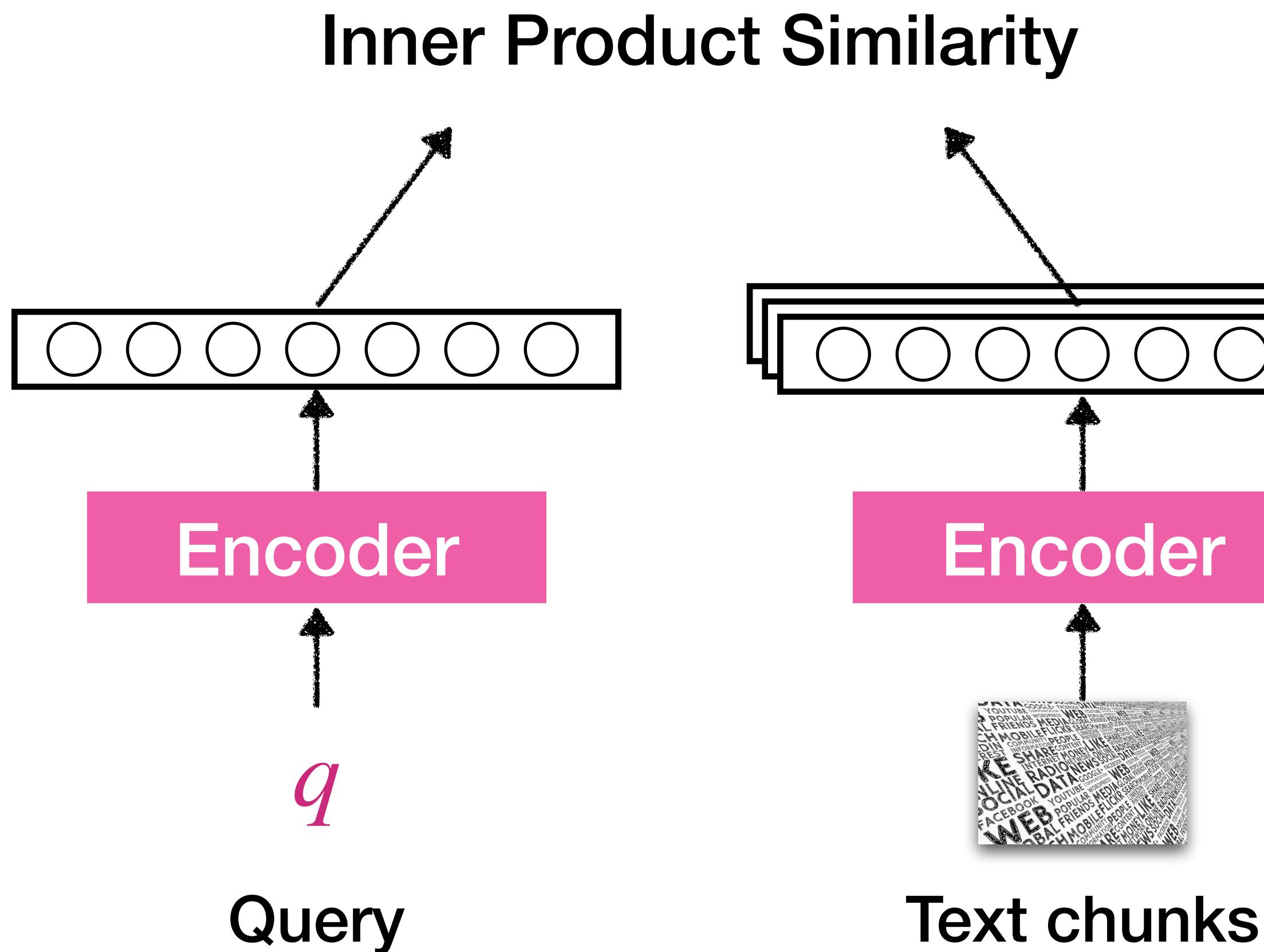
Training dense retrieval models: DPR



$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Positive passage

Training dense retrieval models: DPR



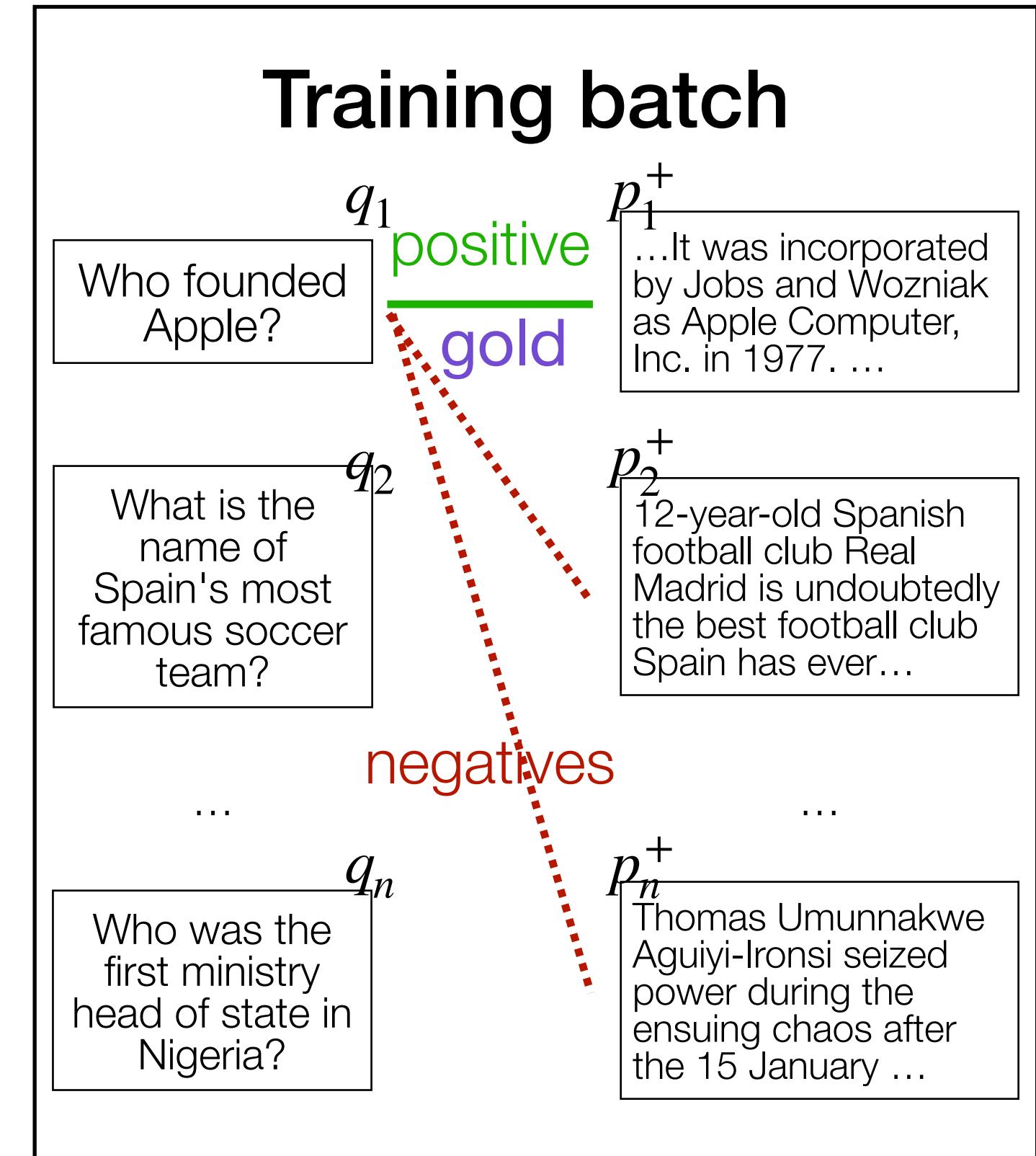
Negative passages
Too expensive to consider all negatives!

$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Training with “in-batch” negatives

$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = -\log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Back-propagation to all in-batch negatives!



Independent training



Work with off-the-shelf models (no extra training required)



Each part can be improved independently

Independent training

-  Work with off-the-shelf models (no extra training required)
-  Each part can be improved independently
-  LMs are not trained to leverage retrieval
-  Retrieval models are not optimized for LM tasks/domains

Training methods for retrieval-augmented LMs

Independent training (Retrieval in-context, kNN-LM)

Sequential training

Joint training w/ asynchronous index update

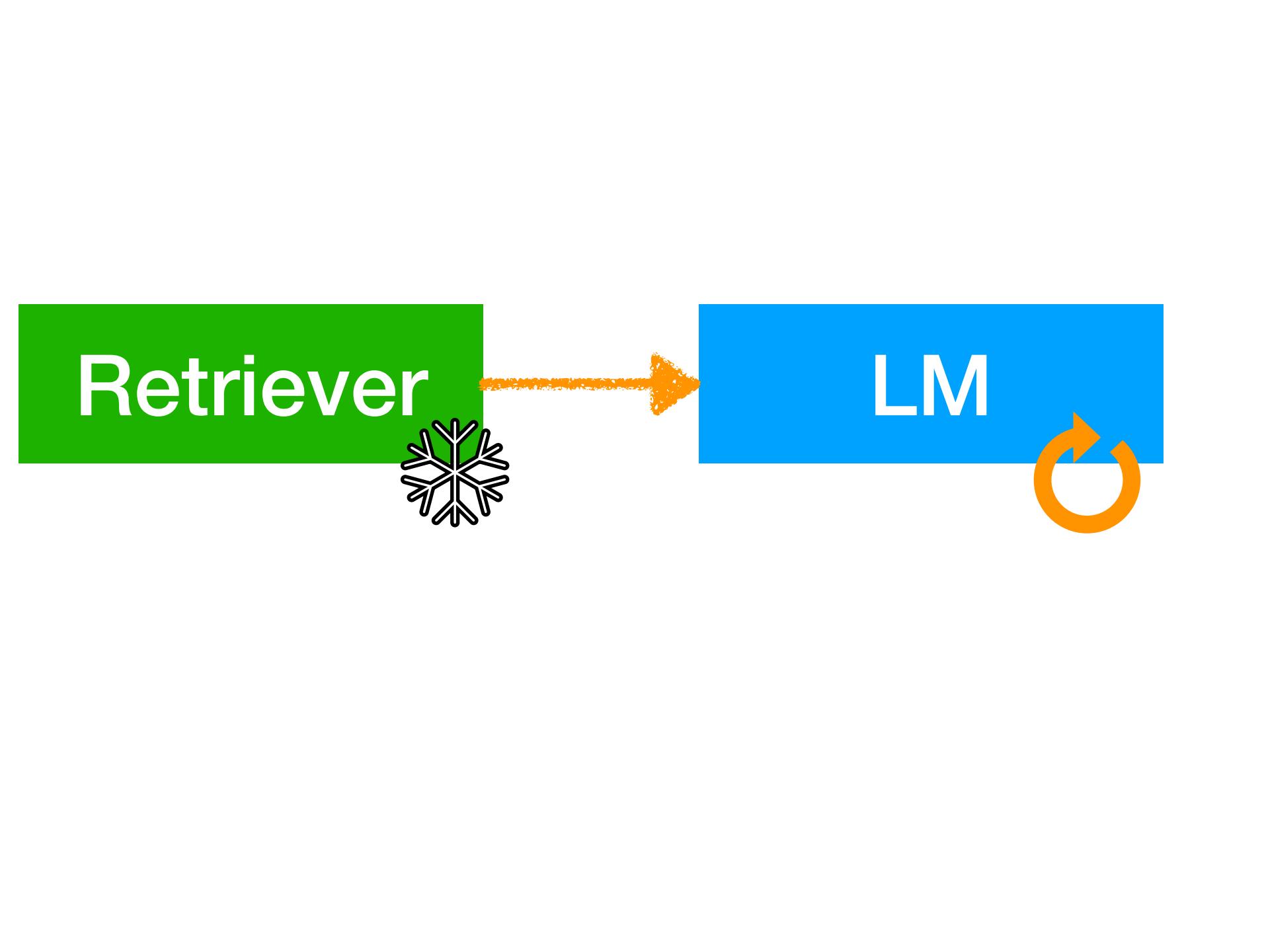
Joint training w/ in-batch approximation

Sequential training

- One component is first trained independently and then fixed
- The other component is trained with an objective that depends on the first one

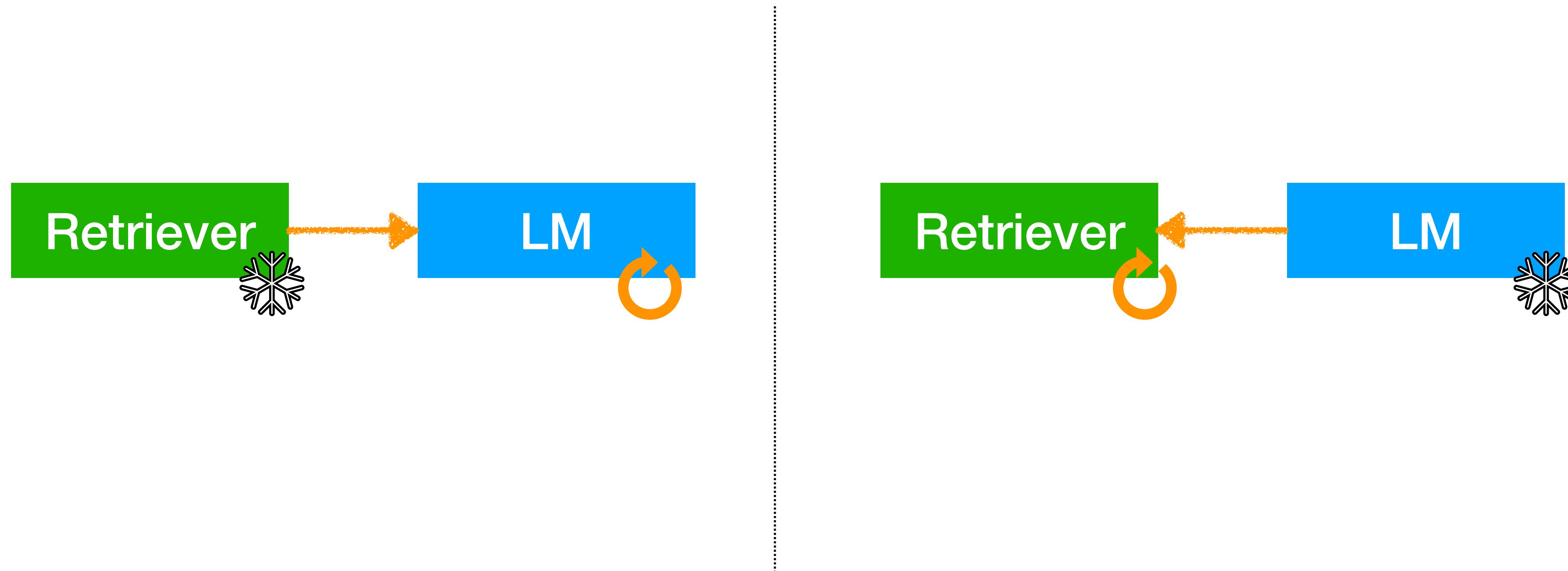
Sequential training

- One component is first trained independently and then fixed
- The other component is trained with an objective that depends on the first one



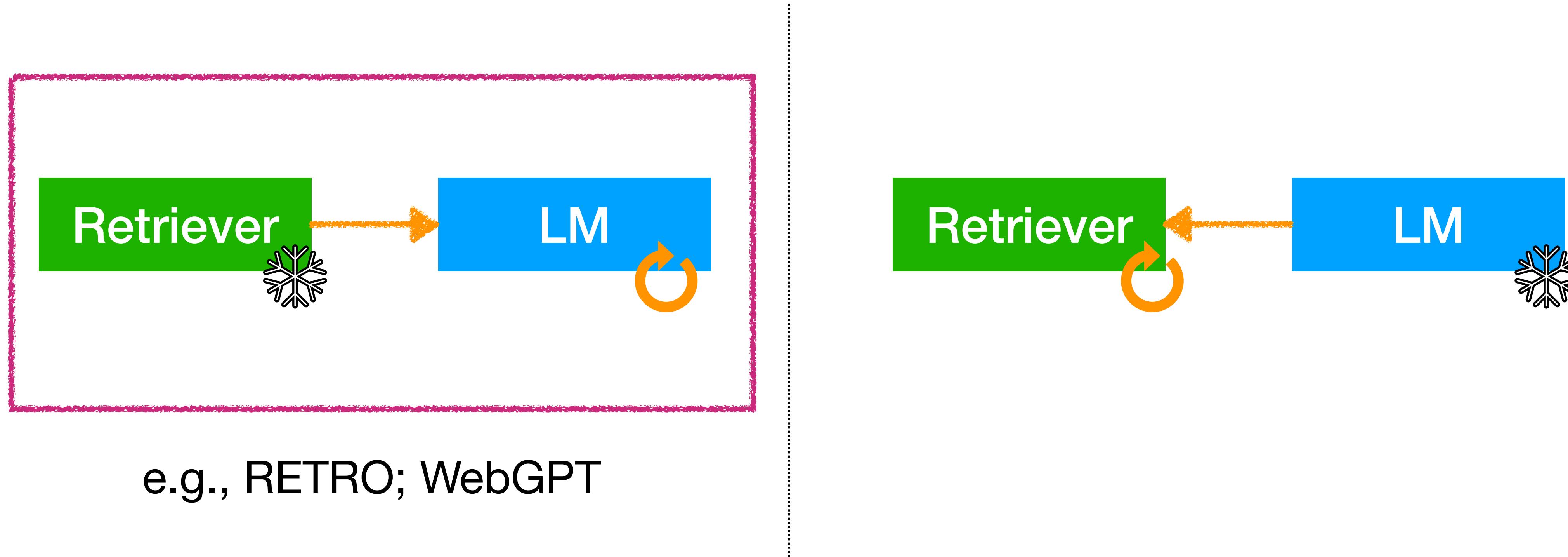
Sequential training

- One component is first trained independently and then fixed
- The other component is trained with an objective that depends on the first one



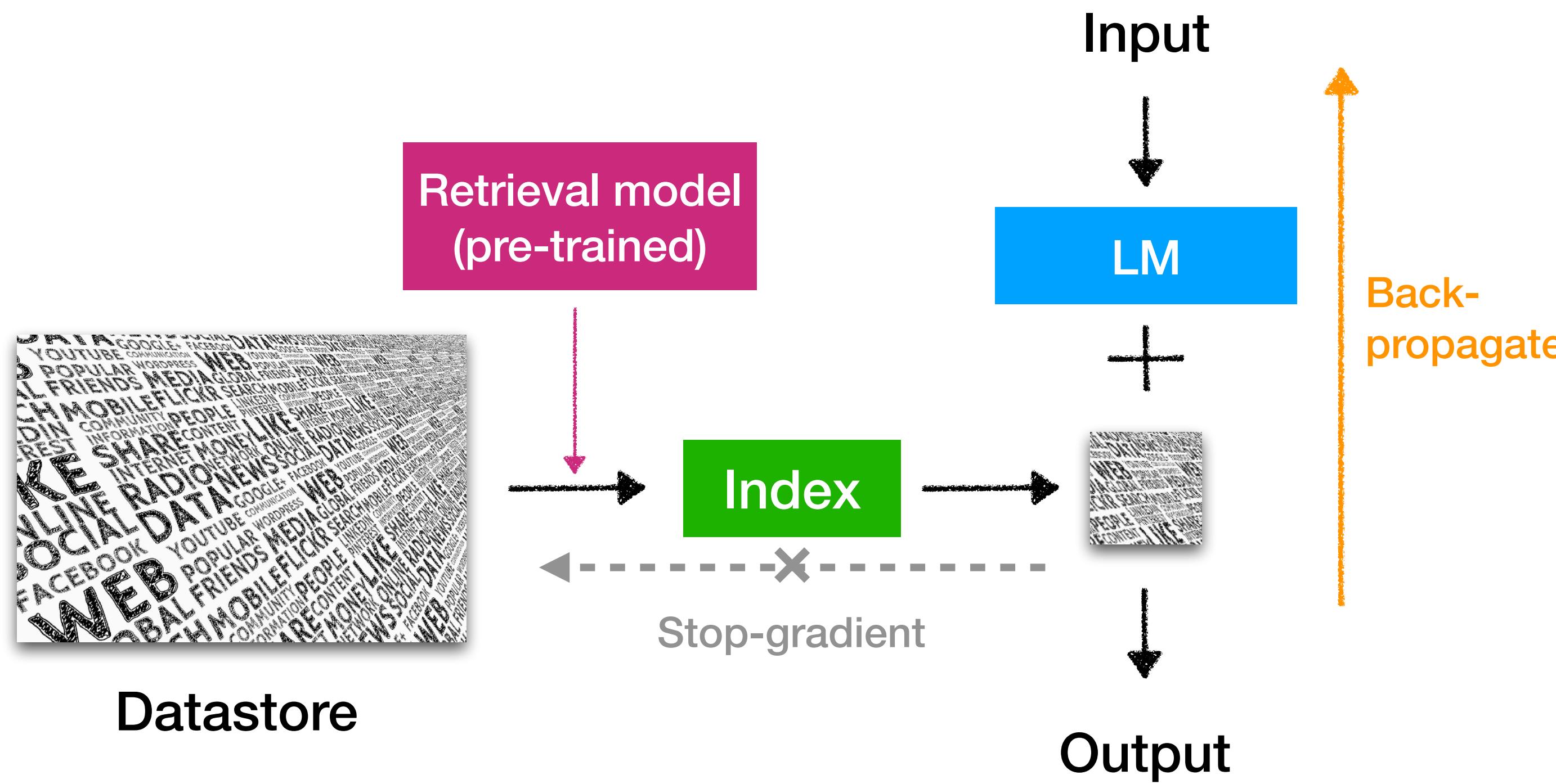
Sequential training

- One component is first trained independently and then fixed
- The other component is trained with an objective that depends on the first one

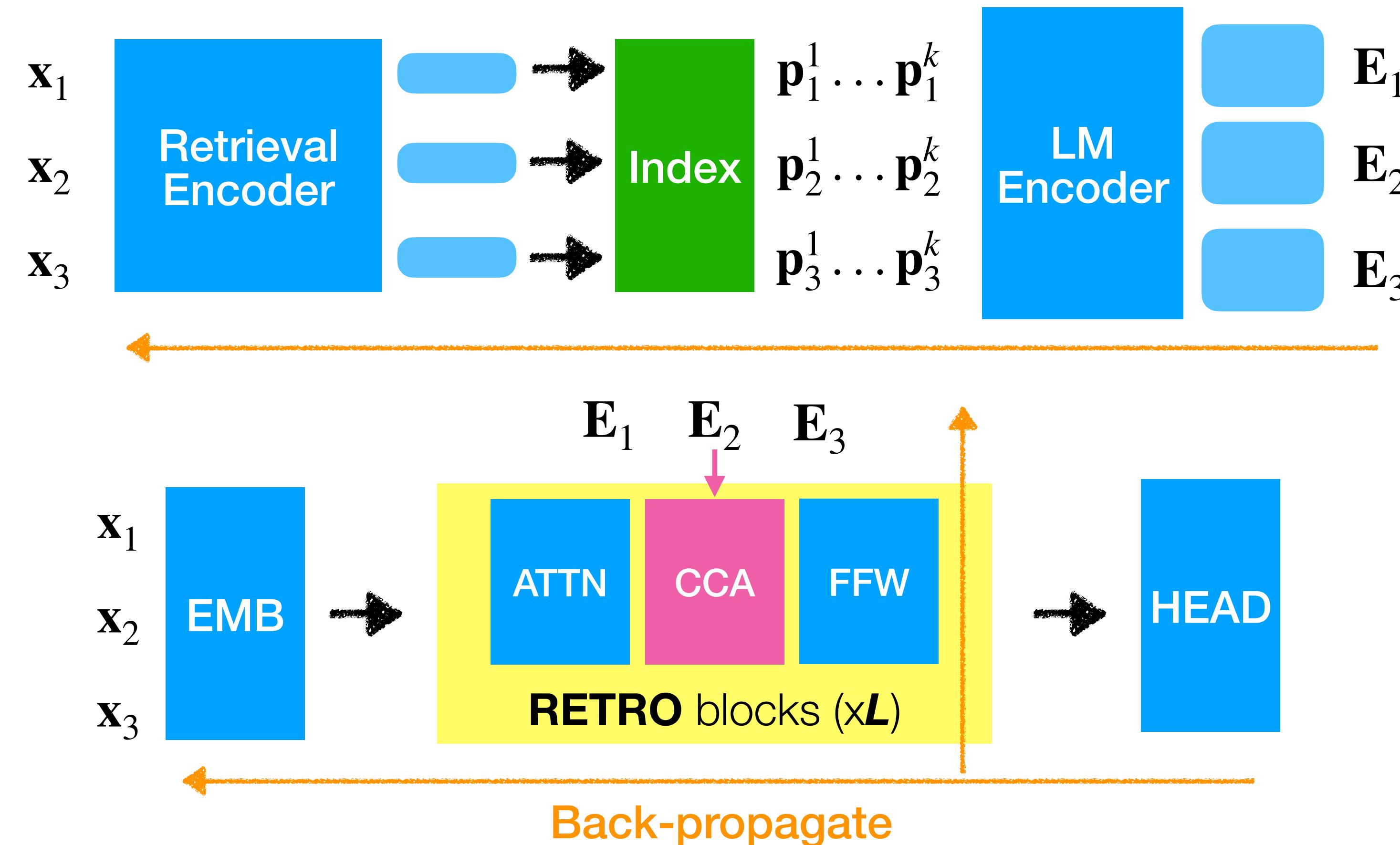


Sequential: train LM

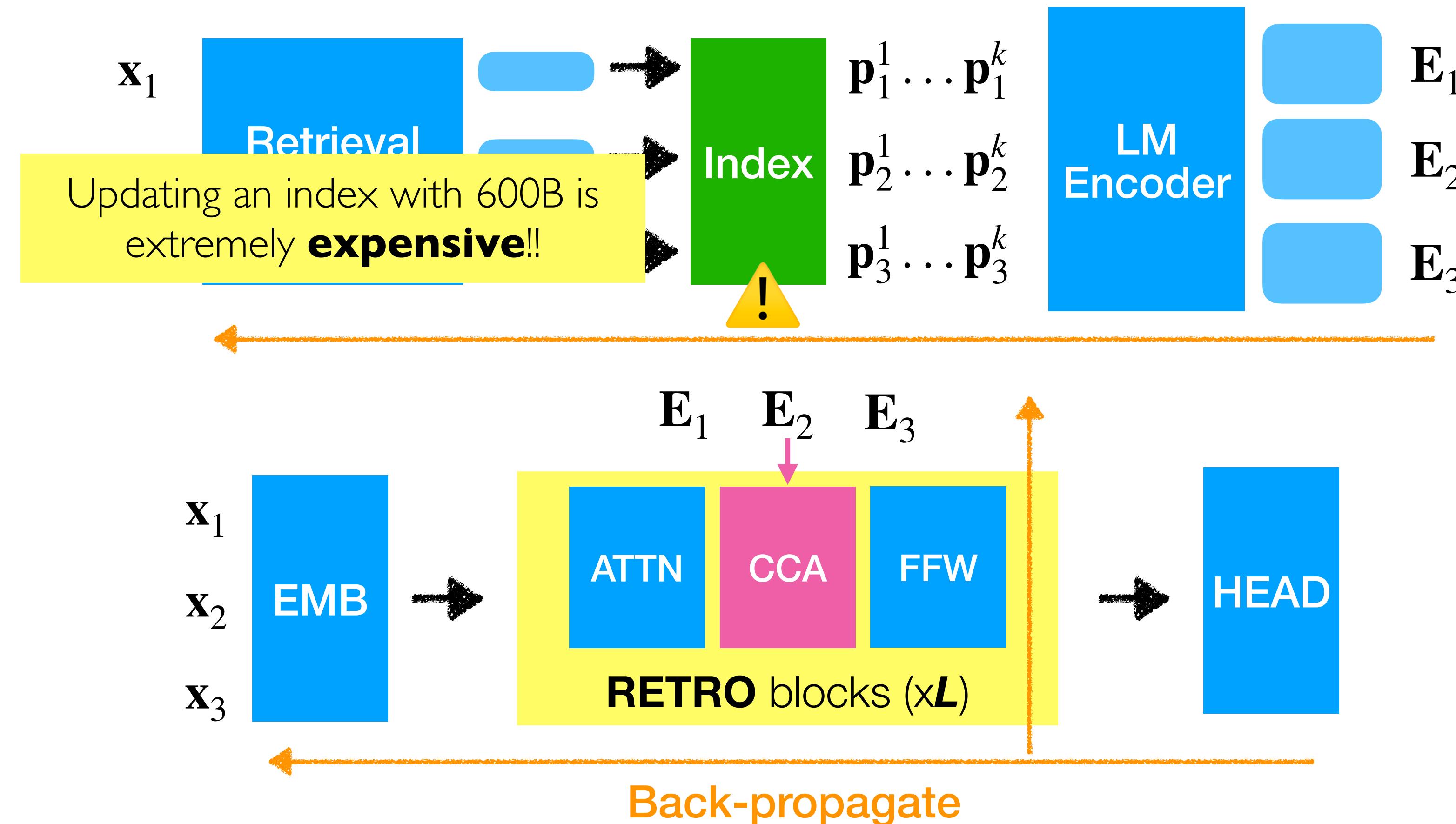
- Retrieval models are first trained independently and then fixed
 - Language models are trained with an objective that depends on the retrieval



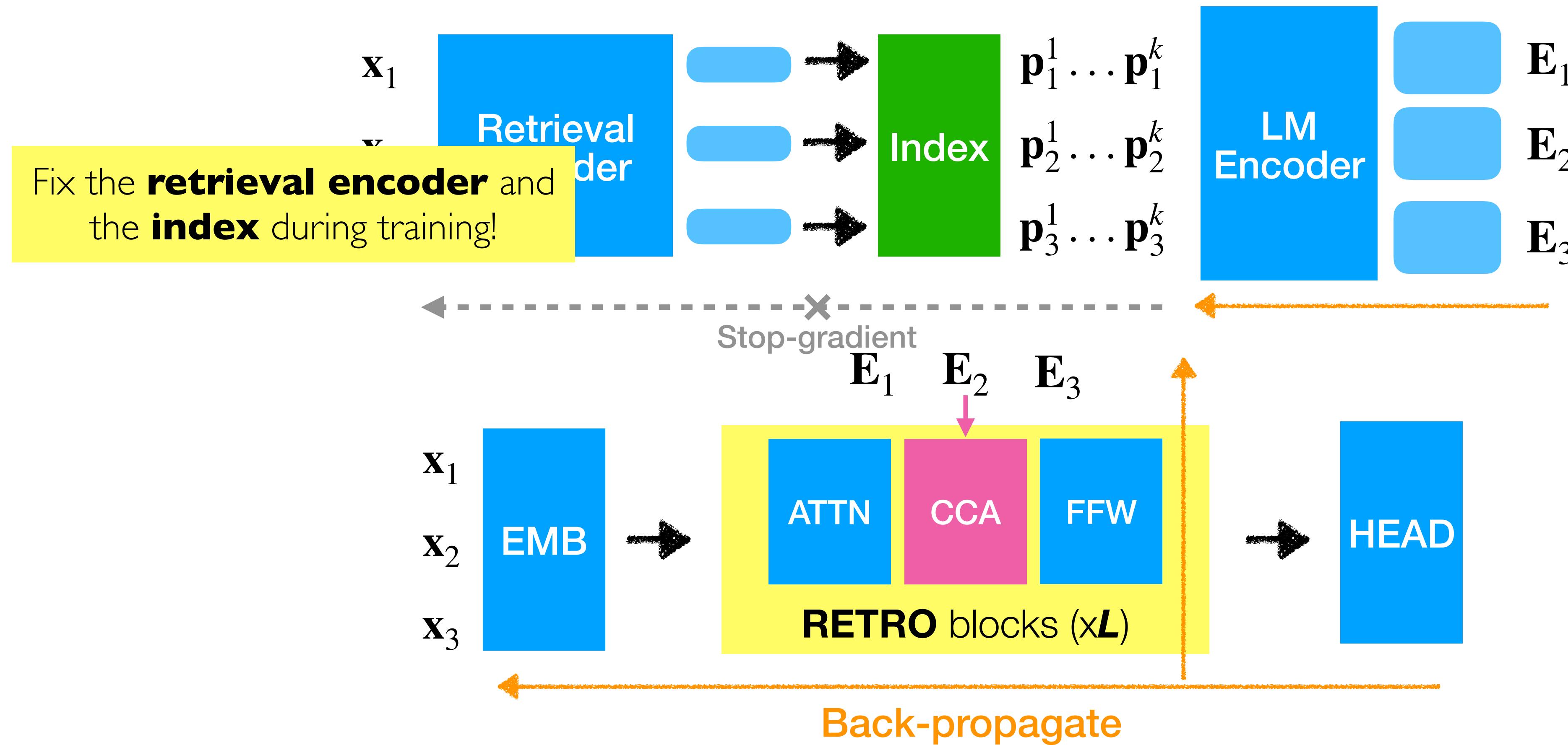
RETRO: Training



RETRO: Training (train LM)

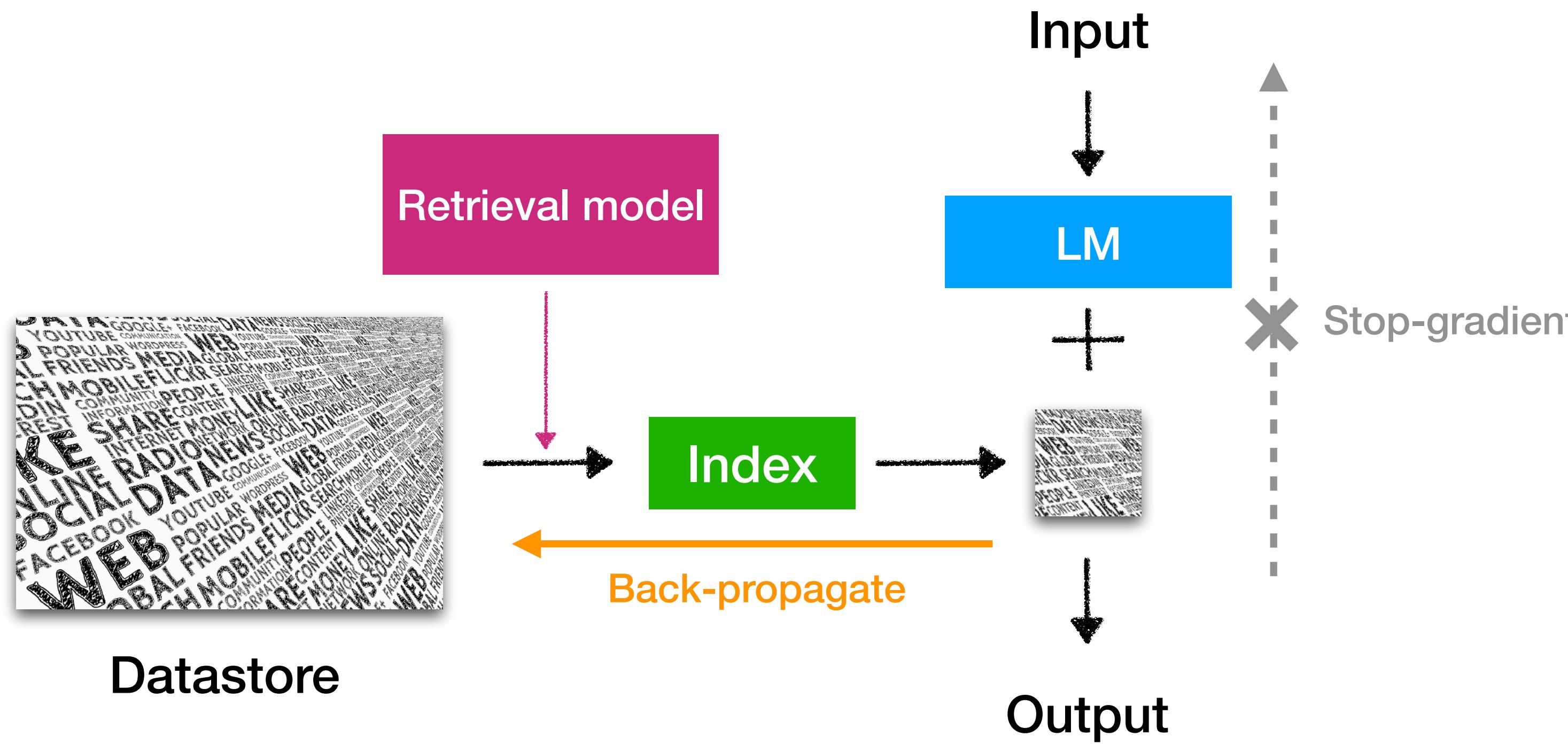


RETRO: Training (train LM)



Sequential: train Retrieval

- Language models are first trained independently and then fixed
- Retrieval models are trained/fine-tuned with supervisions from LMs



Sequential training

-  Work with off-the-shelf components (either a large index or a powerful LM)
-  LMs are trained to effectively leverage retrieval results
-  Retrievers are trained to provide text that helps LMs the most
-  One component is still fixed and not trained

Sequential training

-  Work with off-the-shelf components (either a large index or a powerful LM)
-  LMs are trained to effectively leverage retrieval results
-  Retrievers are trained to provide text that helps LMs the most
-  One component is still fixed and not trained

Let's jointly train retrieval models and LMs!

Training methods for retrieval-augmented LMs

Independent training

Sequential training

Joint training w/ asynchronous index update

Joint training w/ in-batch approximation

Training methods for retrieval-augmented LMs

Independent training

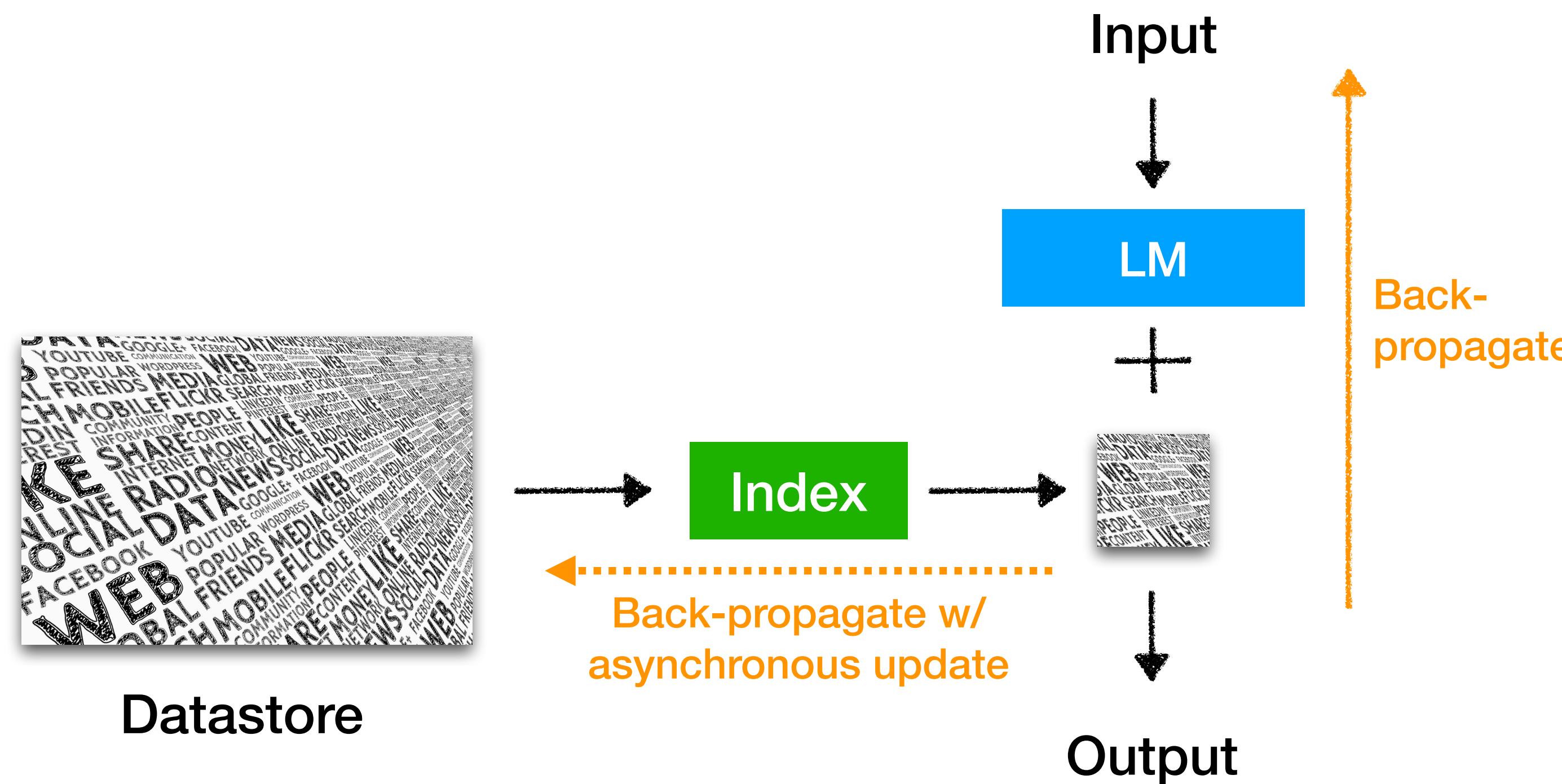
Sequential training

Joint training w/ asynchronous index update

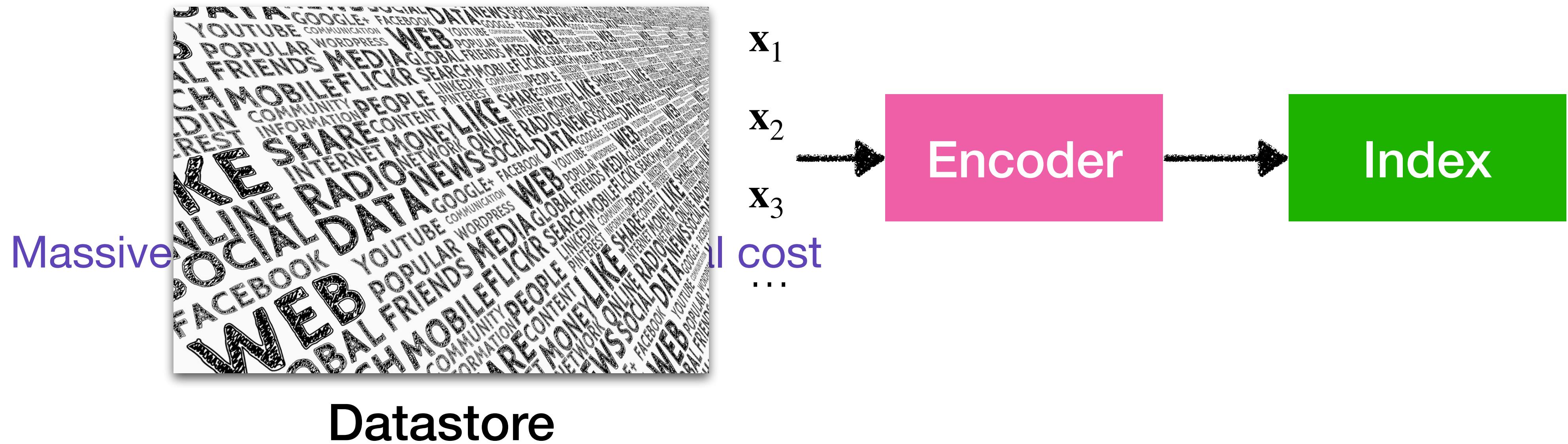
Joint training w/ in-batch approximation

Joint training w/ asynchronous index update

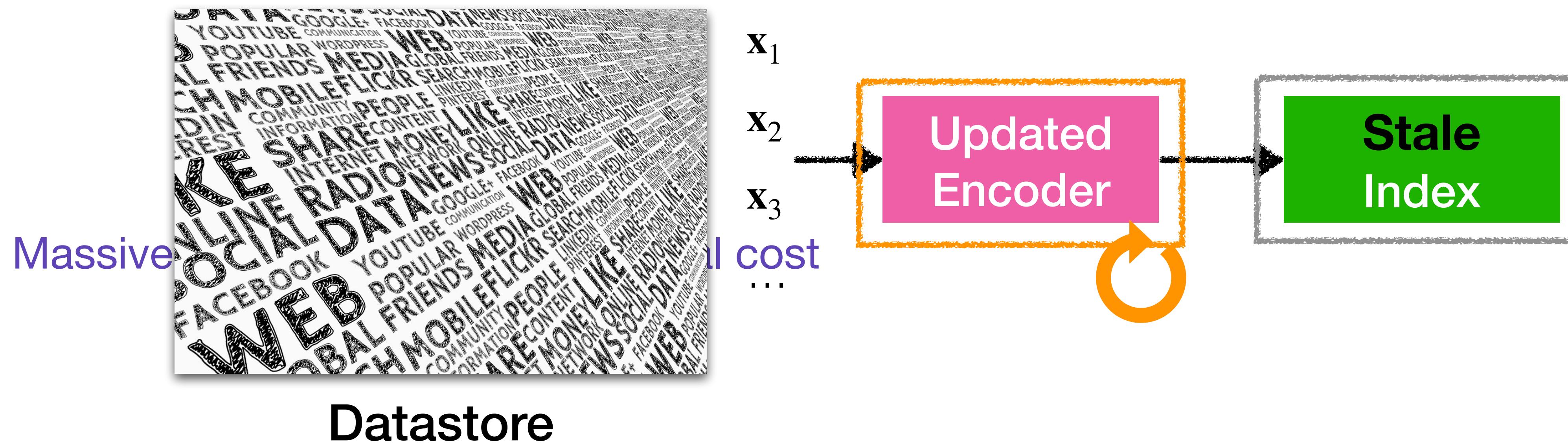
- Retrieval models and language models are trained jointly
 - Allow the index to be “stale”; rebuild the retrieval index every T steps



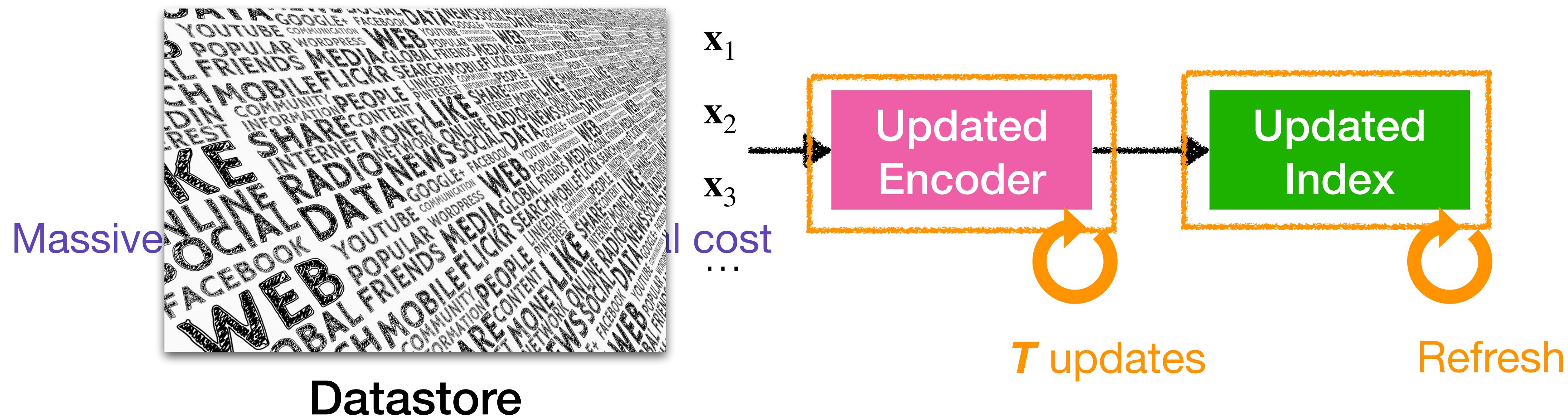
Asynchronous index update



Asynchronous index update

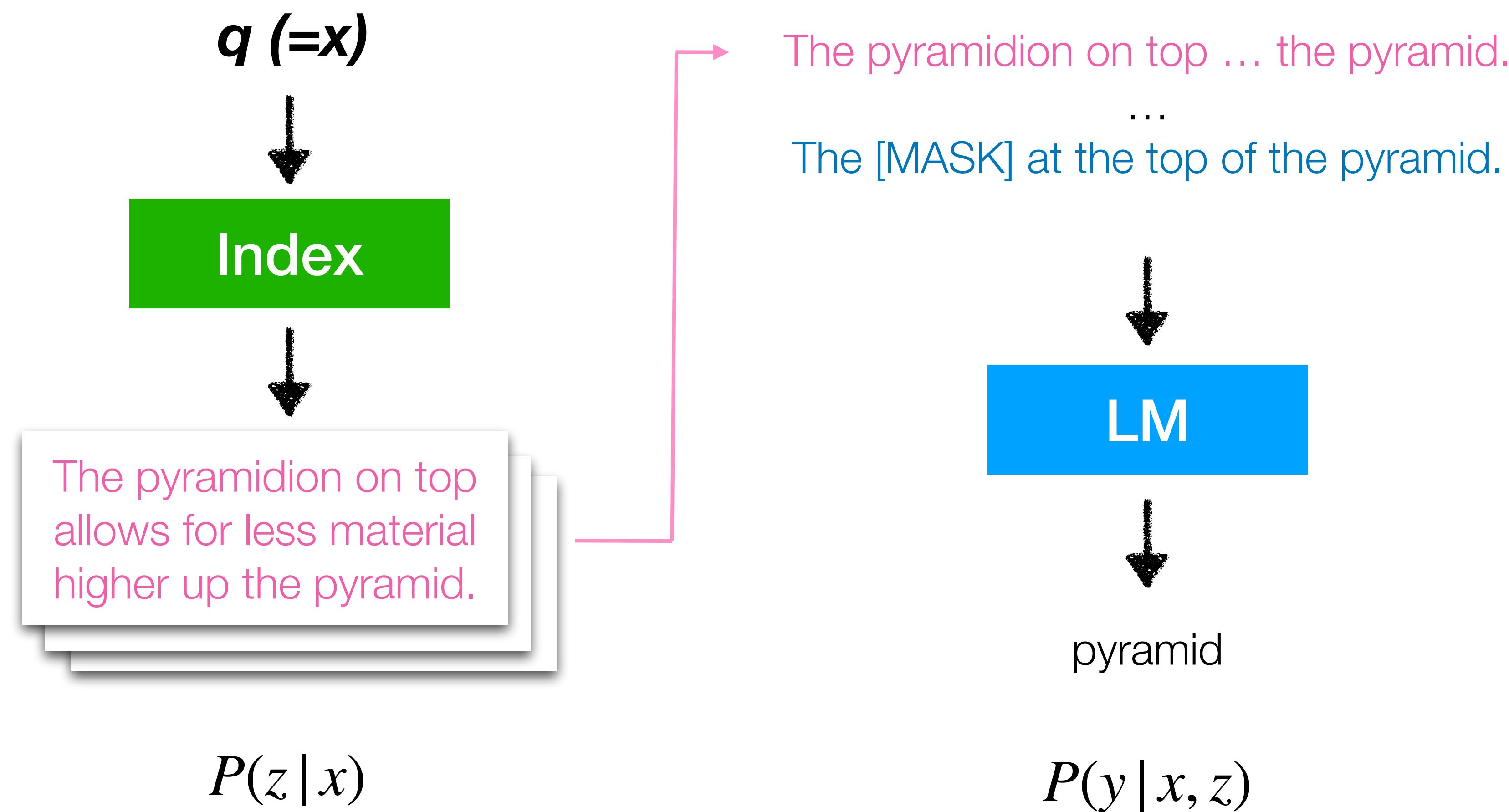


Asynchronous index update



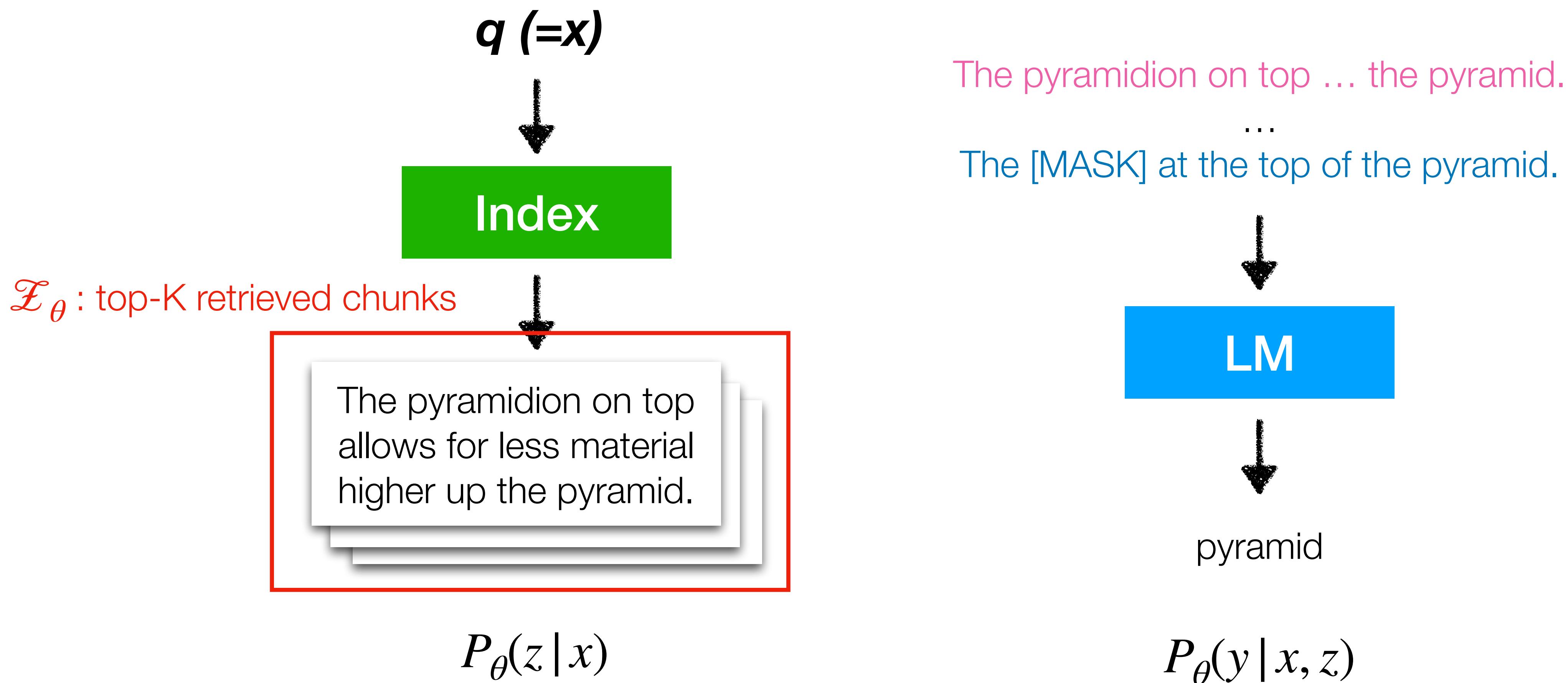
REALM (Guu et al. 2020)

x = The [MASK] at the top of the pyramid.



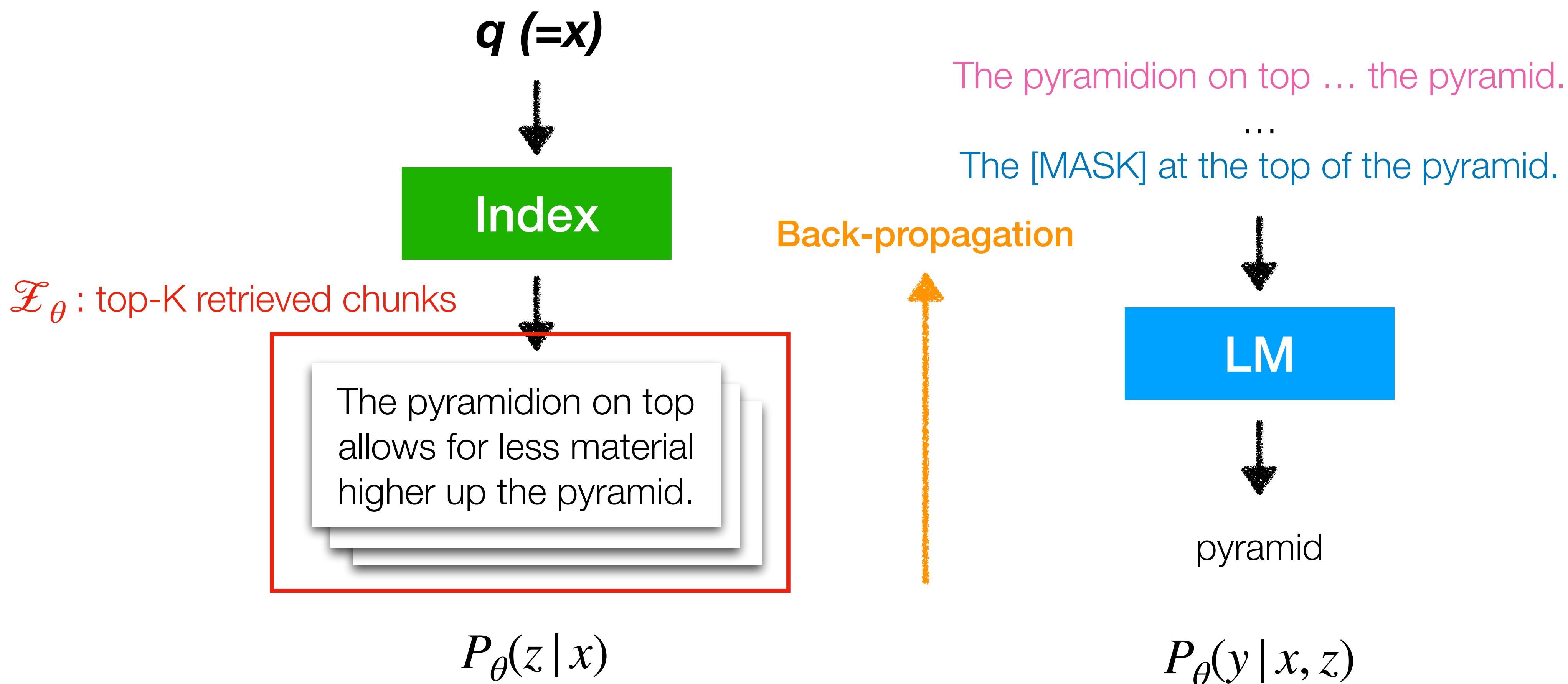
REALM: Training

Objective: maximize $\sum_{z \in \mathcal{Z}_\theta} P_\theta(z | q)P_\theta(y | q, z)$



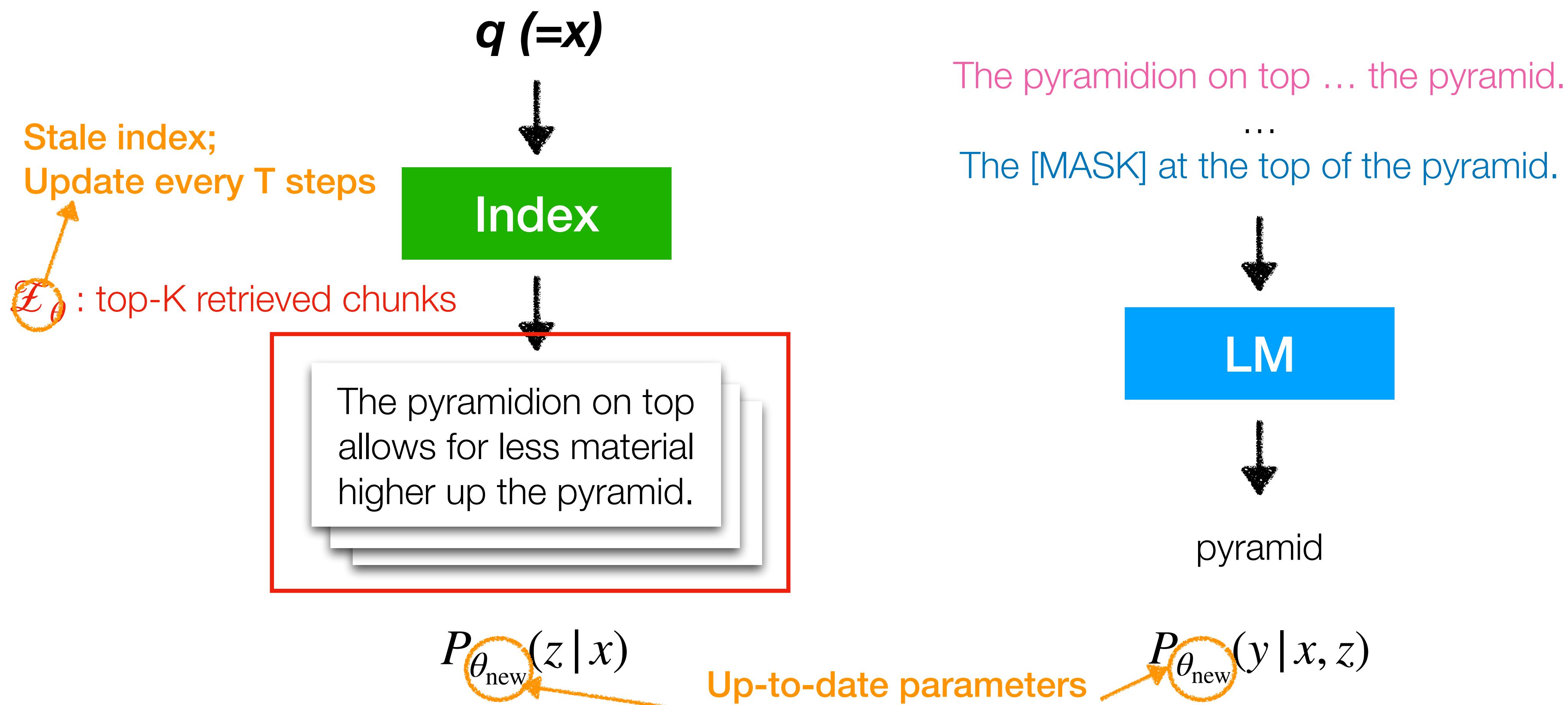
REALM: Training

Objective: maximize $\sum_{z \in \mathcal{Z}_\theta} P_\theta(z | q)P_\theta(y | q, z)$



REALM: Training

Objective: maximize $\sum_{z \in \mathcal{Z}_\theta} P_\theta(z | q)P_\theta(y | q, z)$

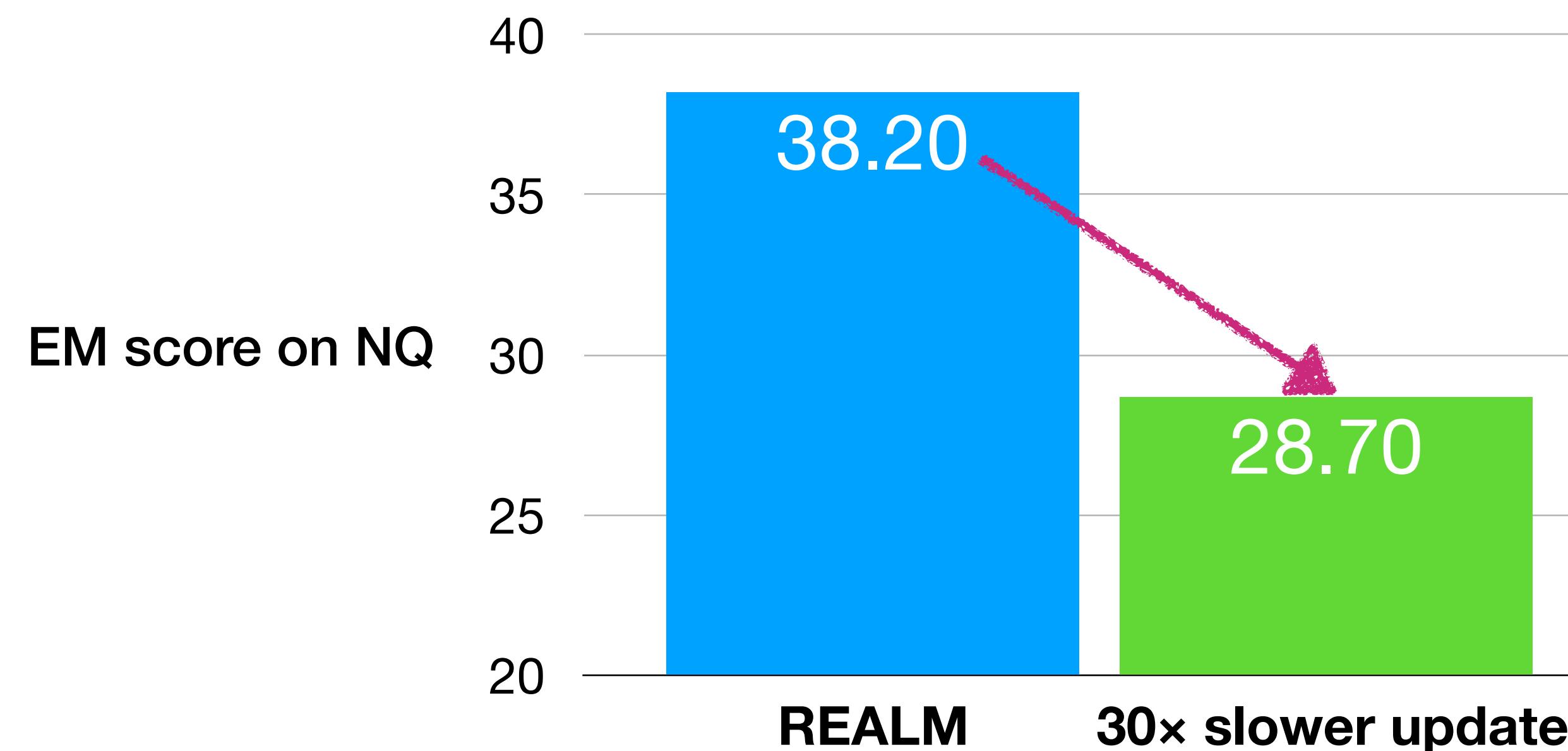


REALM: Index update rate

How often should we update the retrieval index?

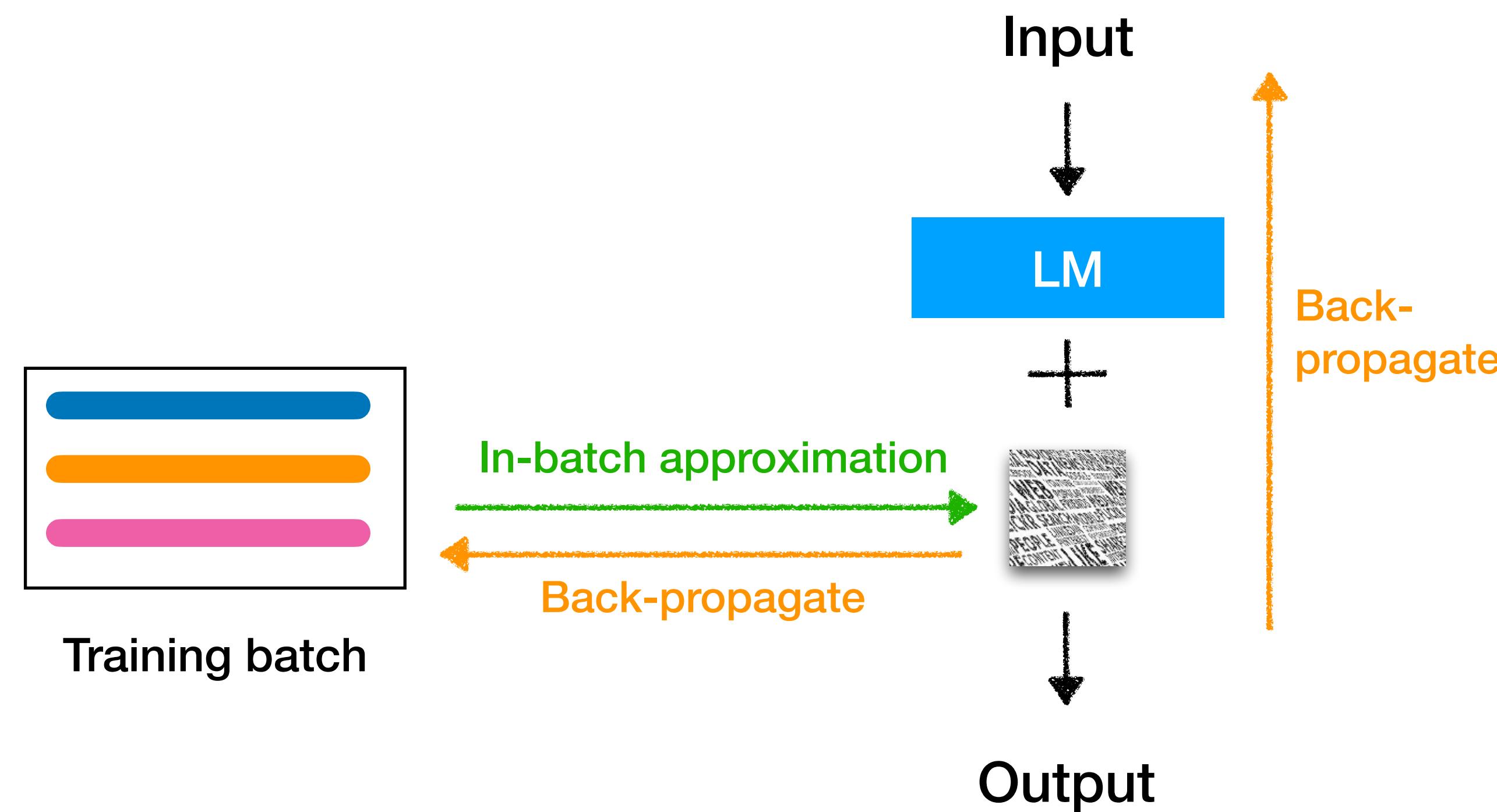
- Frequency too high: expensive
- Frequency too slow: out-dated

REALM: updating the index every 500 training steps

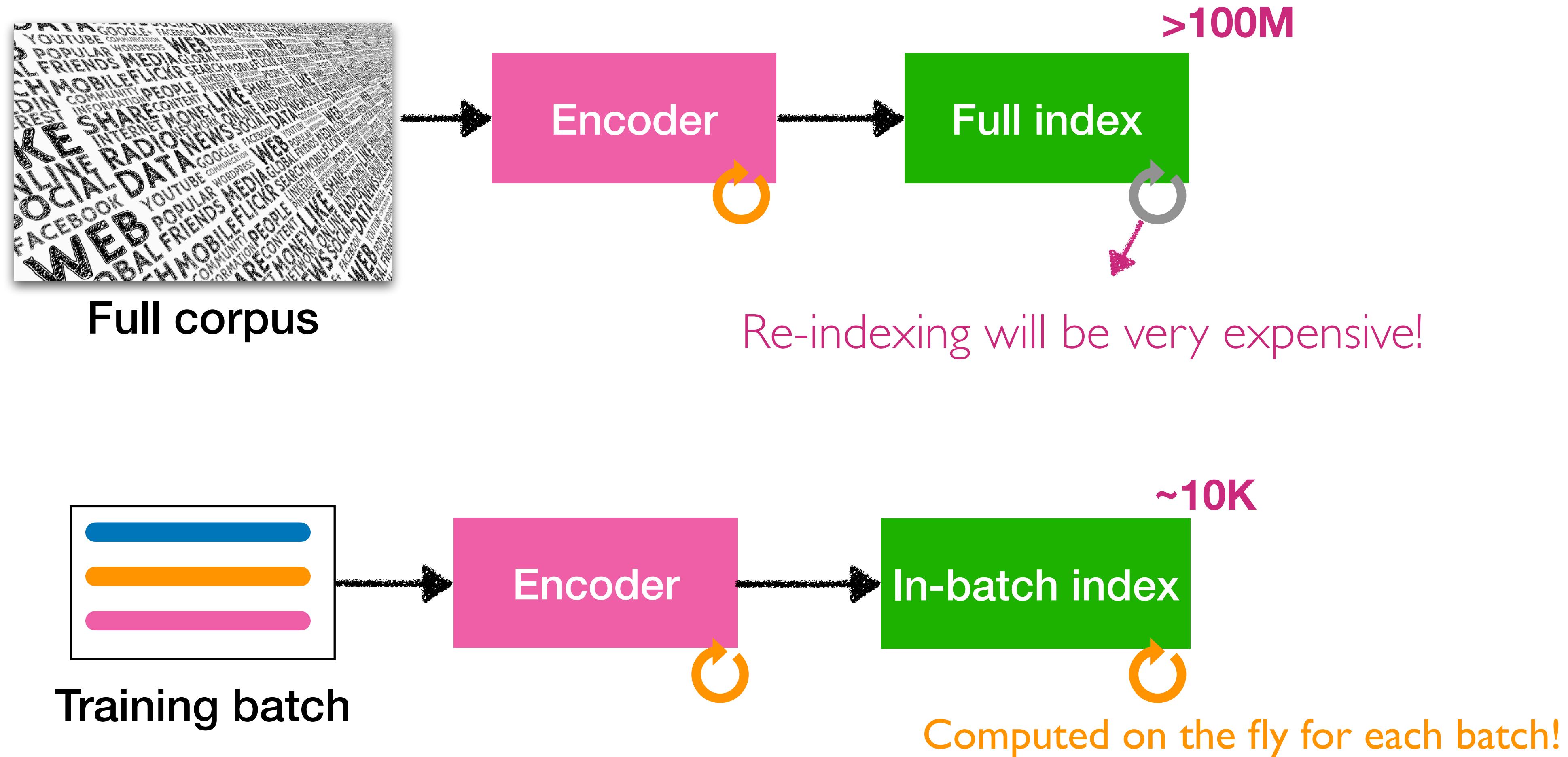


Joint training with in-batch approximation

- Retrieval models and language models are trained jointly
- Use “in-batch index” instead of full index



Joint training with in-batch approximation



Joint training

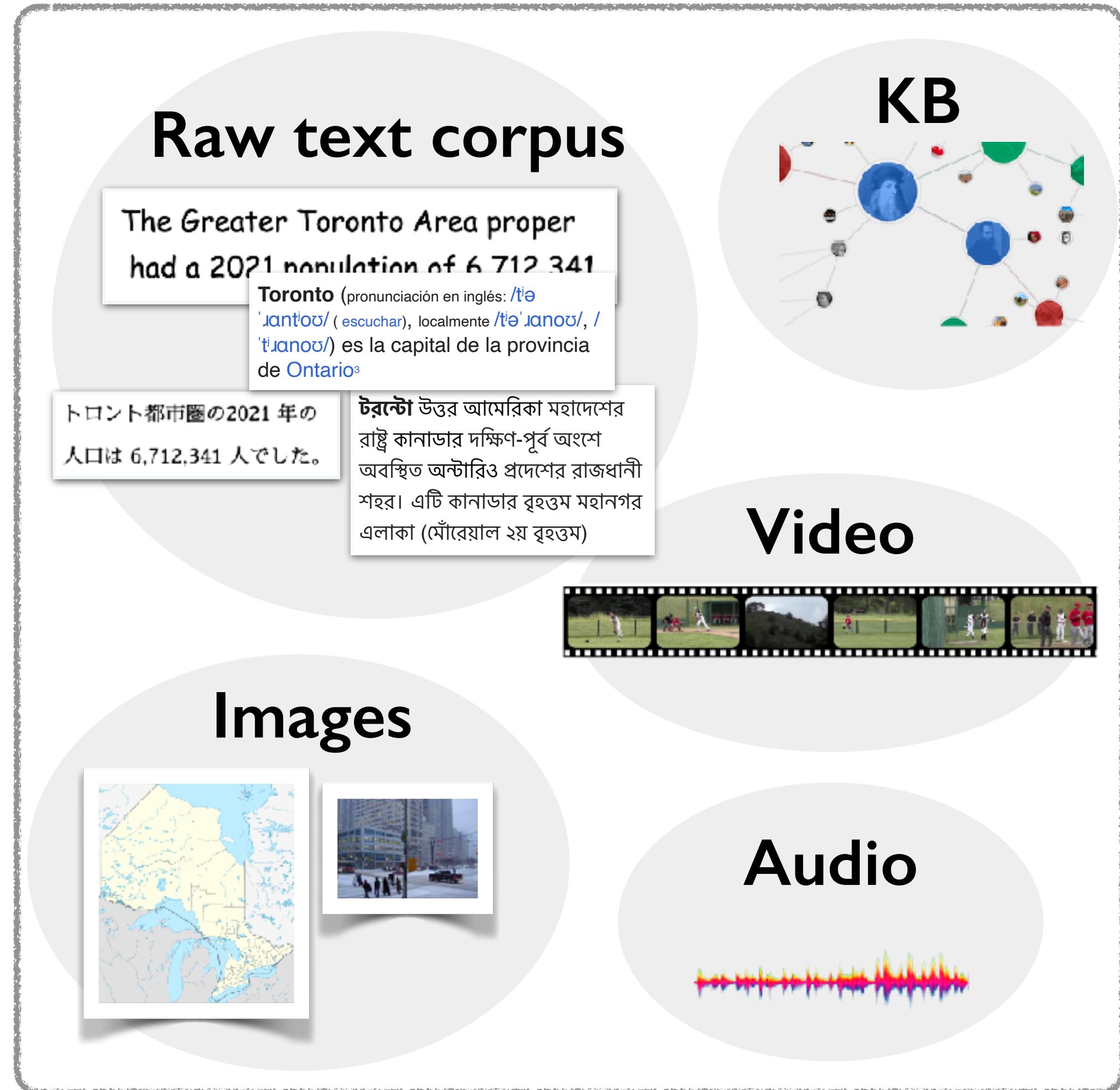
-  End-to-end trained – each component is optimized
-  Good performance
-  Training is more complicated
(async update, overhead, data batching, etc)
-  Train-test discrepancy still remains

Summary of training

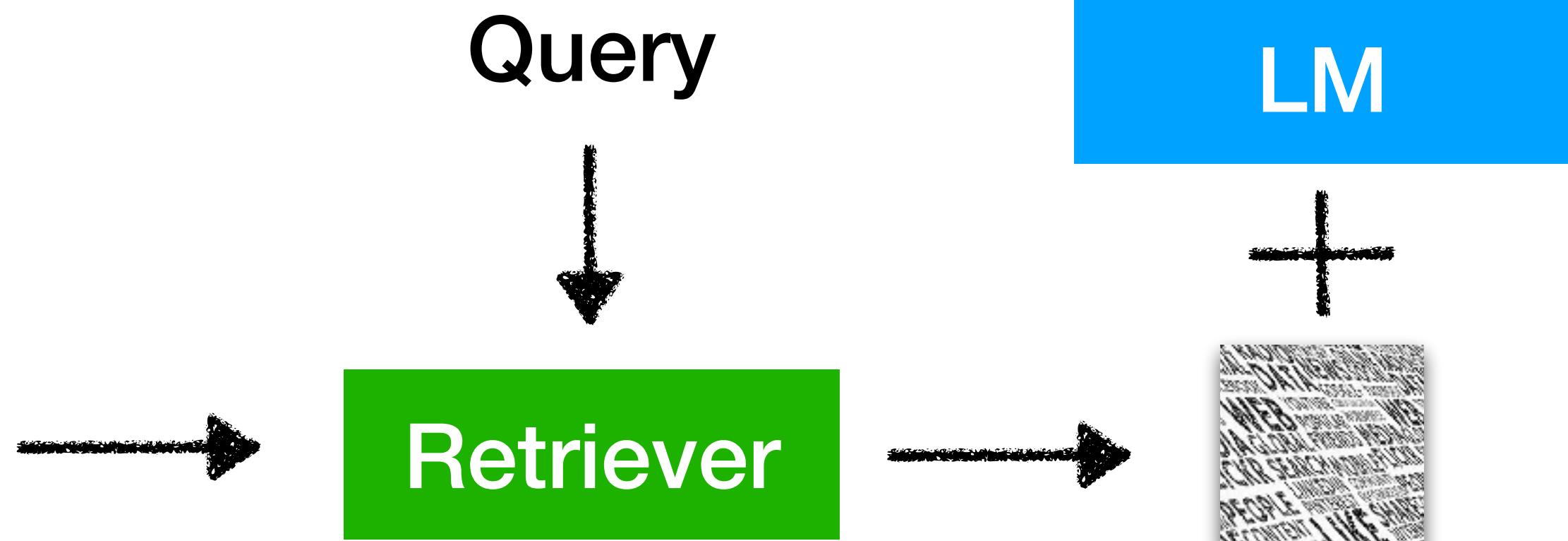
Training method		
Independent training (Ram et al 2023; Khandelwal et al 2020)	<ul style="list-style-type: none">* Easy to implement: off-the-shelf models* Easy to improve: sub-module can be separately improved	<ul style="list-style-type: none">* Models are not end-to-end trained — suboptimal performance
Sequential training (Borgeaud et al 2021; Shi et al 2023)		
Joint training: async update (Guu et al 2020; Izacard et al 2022)		
Joint training: in-batch approx (Zhong et al 2022; Min et al 2023; Rubin and Berant 2023)	<ul style="list-style-type: none">* End-to-end trained — very good performance!	<ul style="list-style-type: none">* Training may be complicated (overhead, batching methods, etc)* Train-test discrepancy still remains

Extensions: multilingual and multimodal

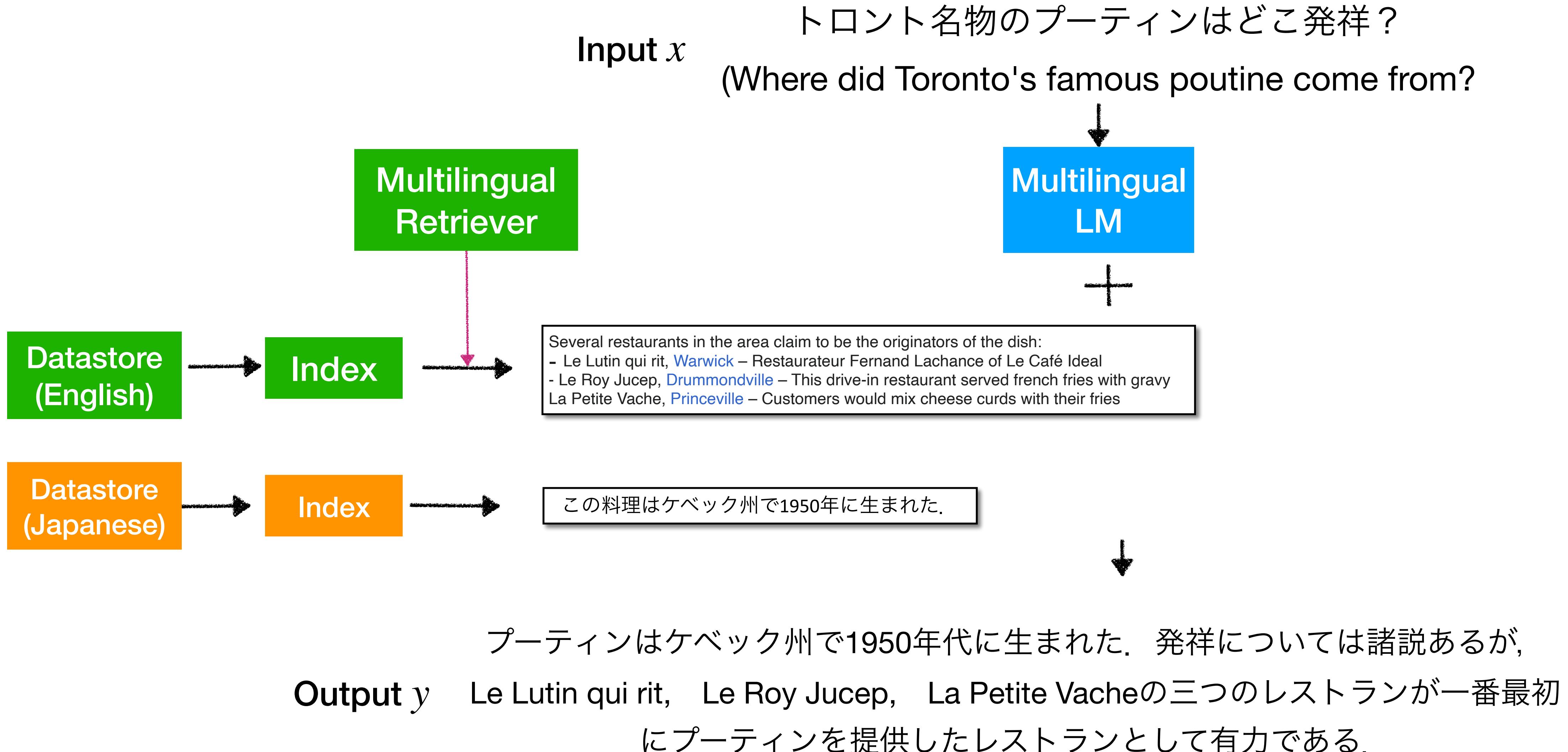
Diverse Knowledge Sources



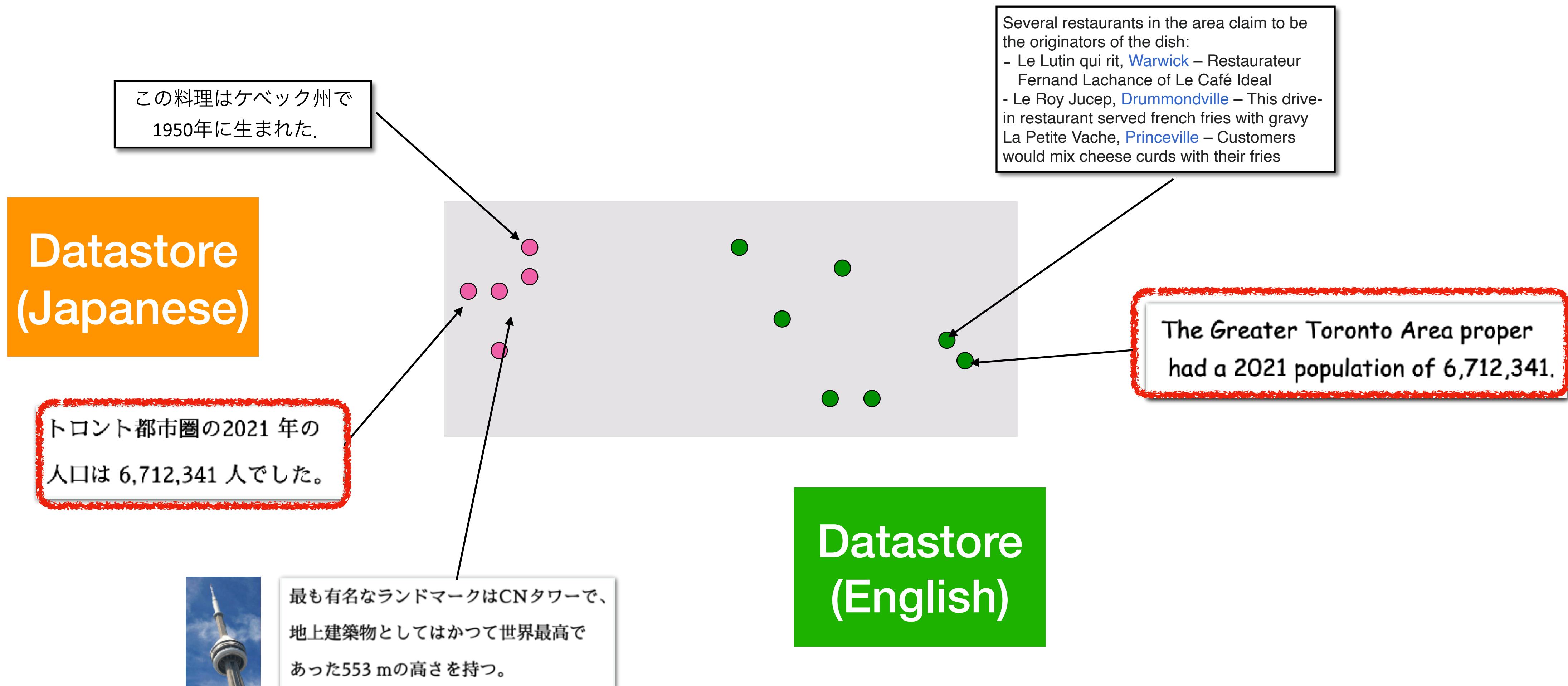
Datastore



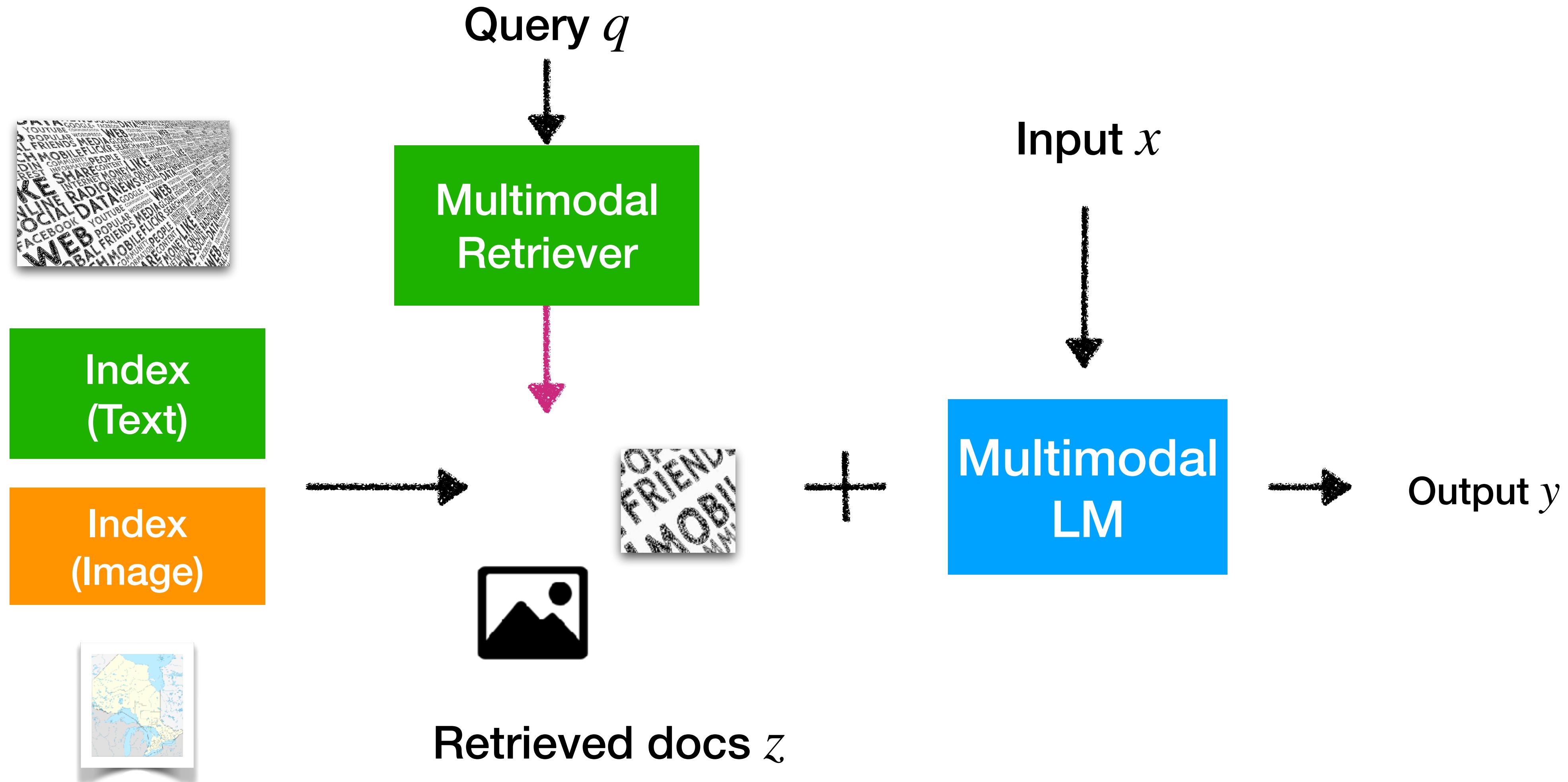
Multilingual Retriever



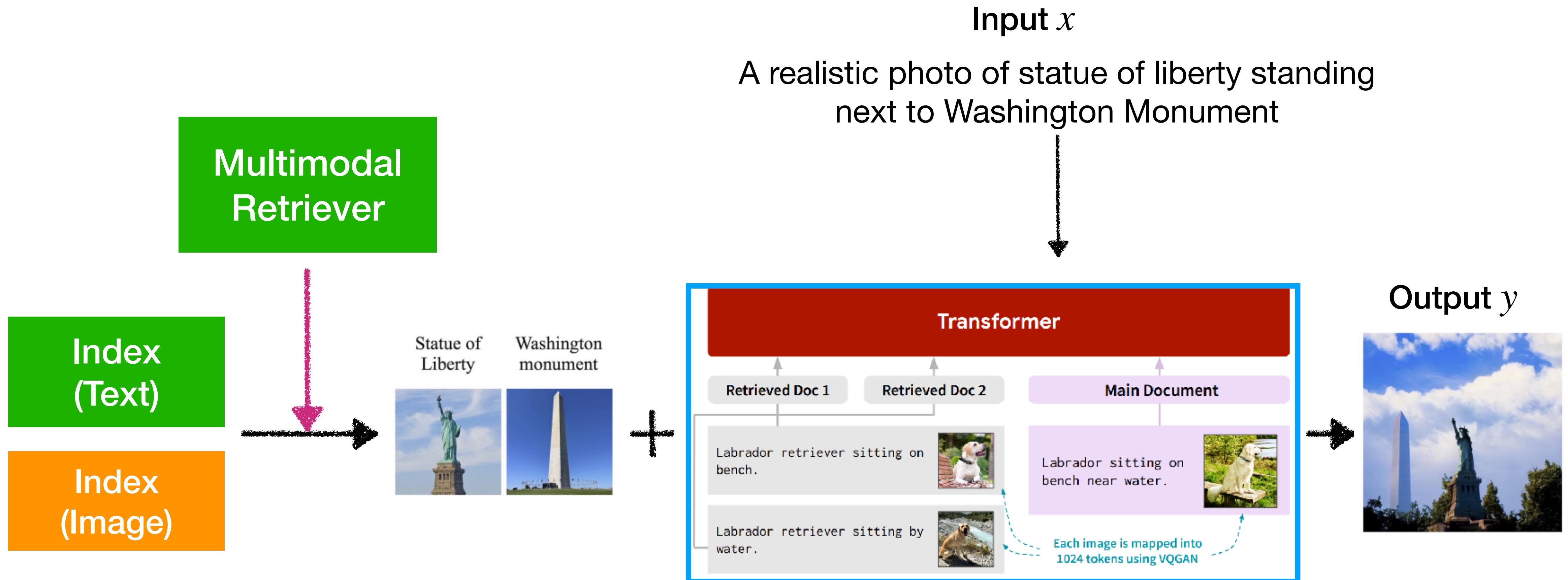
Language Biases in Representation Spaces



Multimodal Retrieval based Los



Multimodal Retrieval based LMs: RA-CM3



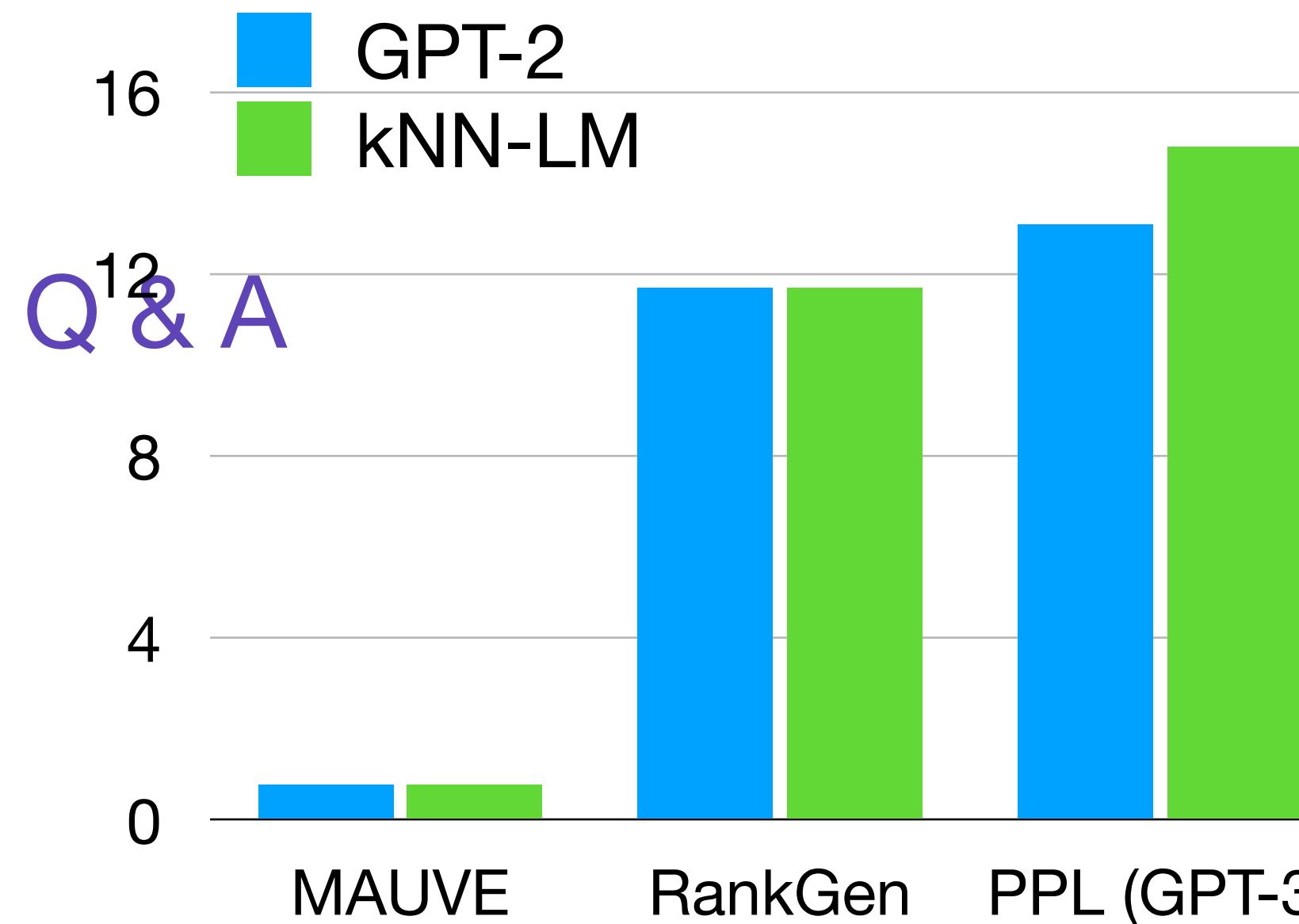
Yasunaga et al. 2023. “Retrieval-Augmented Multimodal Language Modeling”

Opportunities

Challenge: retrieval-augmented LMs for applications

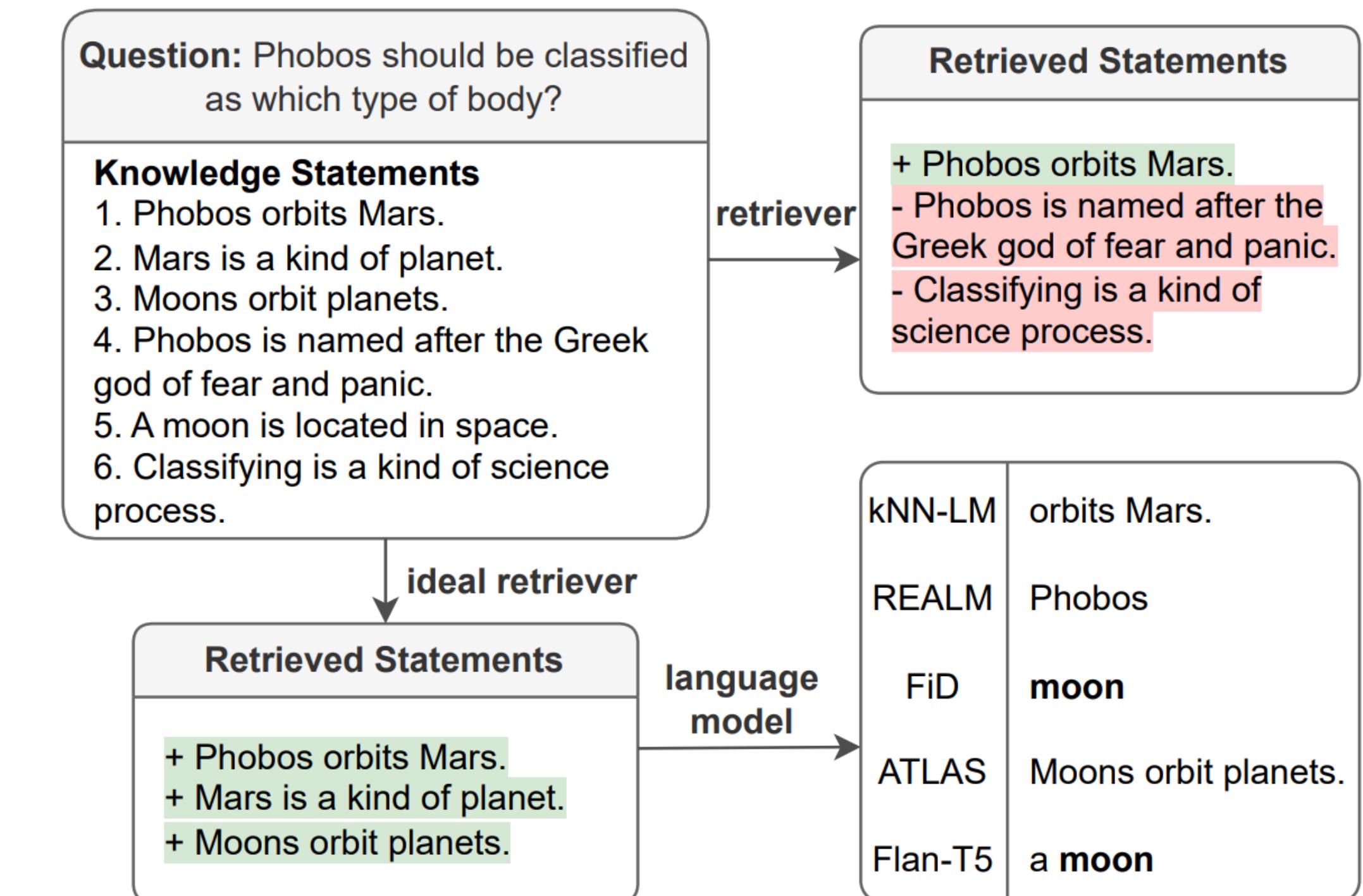
Open-ended text generation? Reasoning?

Doesn't improve open-ended generation



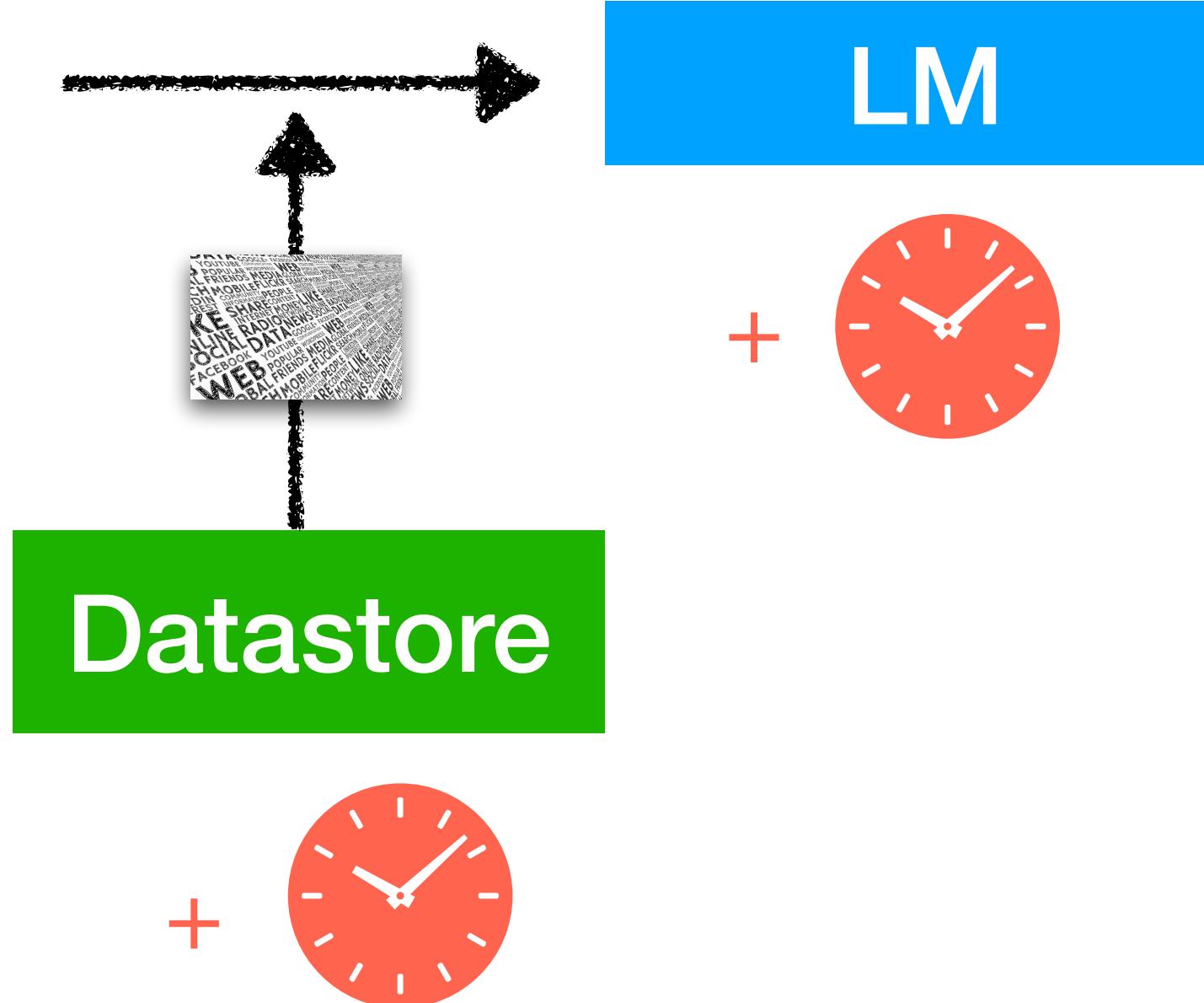
Wang et al. kNN-LM Does Not Improve Open-ended Text Generation. ACL 2023.

Failure of retrieval in reasoning task

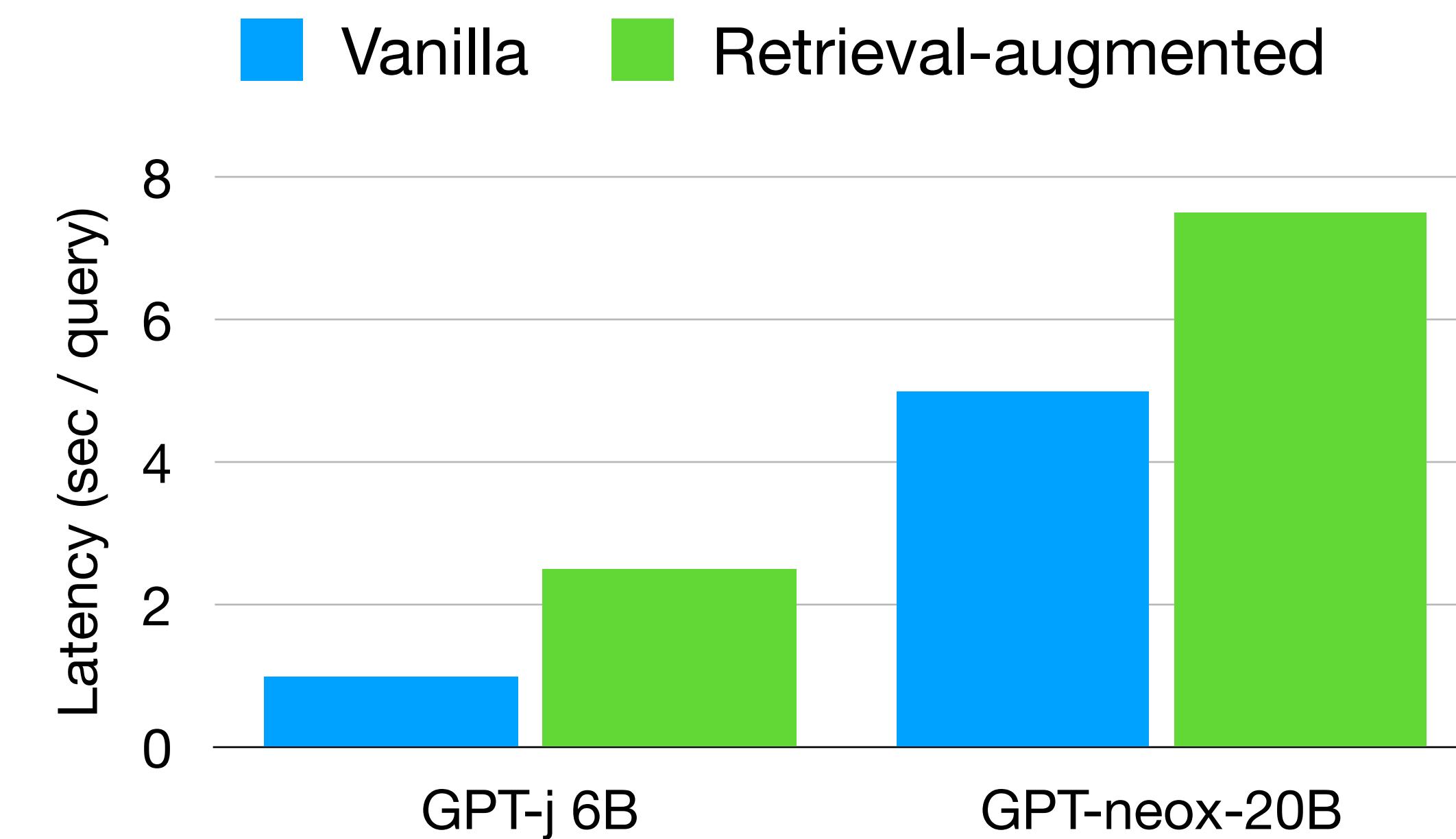


Challenge: efficiency retrieval-augmented LMs

Additional costs from retrieval augmentation



Retrieval-augmented LMs add inference costs

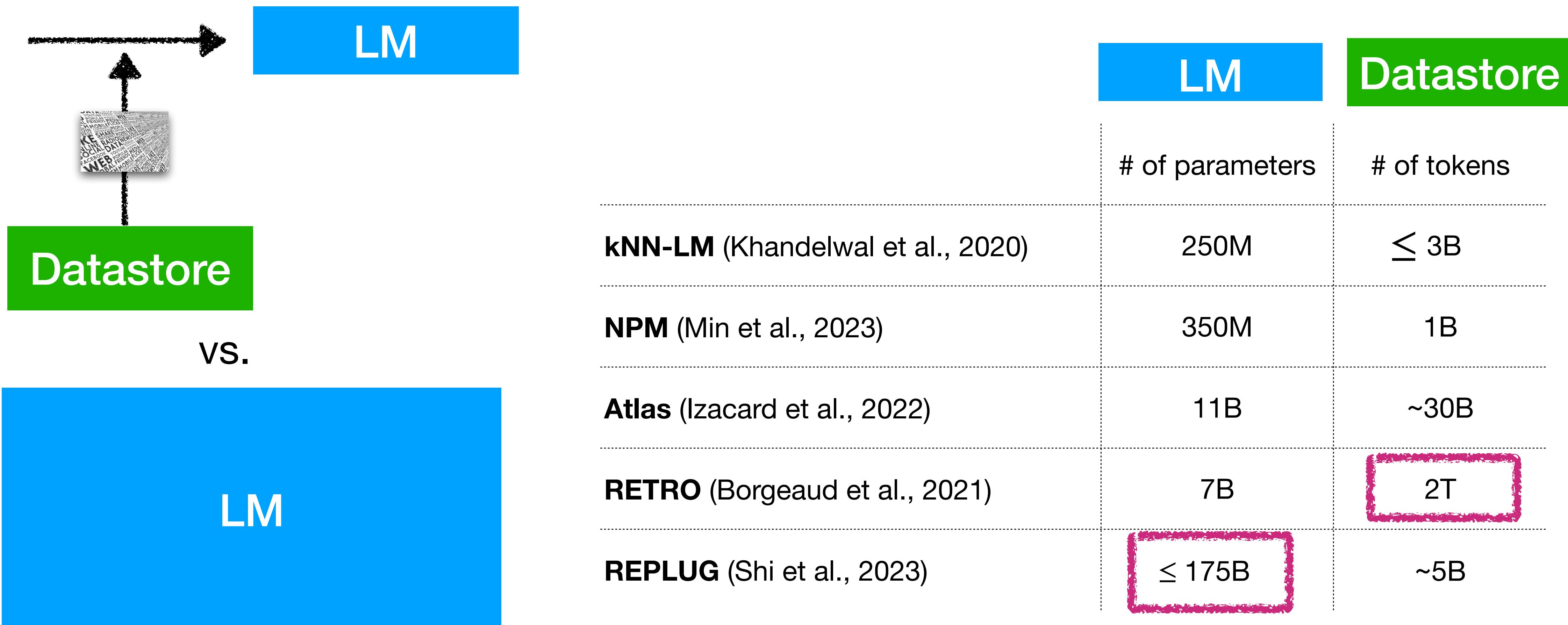


Mallen*, Asai* et al., When Not to Trust Language Models:

Investigating Effectiveness of Parametric and Non-Parametric Memories (Best Video; Oral) 2023

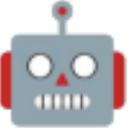
Challenge: scaling retrieval-augmented LMs

A small LM + a large datastore \approx a large parametric LM?



Challenge: robustness and controllability

 What are the latest discoveries from the James Webb Space Telescope?

 The James Webb Space Telescope is designed to peer into the dusty clouds of gas where stars and planetary systems are born. Webb has captured the first direct image of an exoplanet, and the Pillars of Creation in the Eagle Nebula^{[1][2]}. Additionally, the telescope will be used to study the next interstellar interloper^[3].

(*Some generated statements may not be fully supported by citations, while others are fully supported.)

Cited Webpages

[1]:  nasa.gov (✖citation does not support its associated statement)
[NASA's Webb Confirms Its First Exoplanet](#)
... Researchers confirmed an exoplanet, a planet that orbits another star, using NASA's James Webb Space Telescope for the first time. ...

[2]:  cnn.com (⚠citation partially supports its associated statement)
[Pillars of Creation: James Webb Space Telescope](#) ...
... The Pillars of Creation, in the Eagle Nebula, is a star-forming region captured in a new image (right) by the James Webb Space Telescope that reveals more detail than a 2014 image (left) by Hubble ...

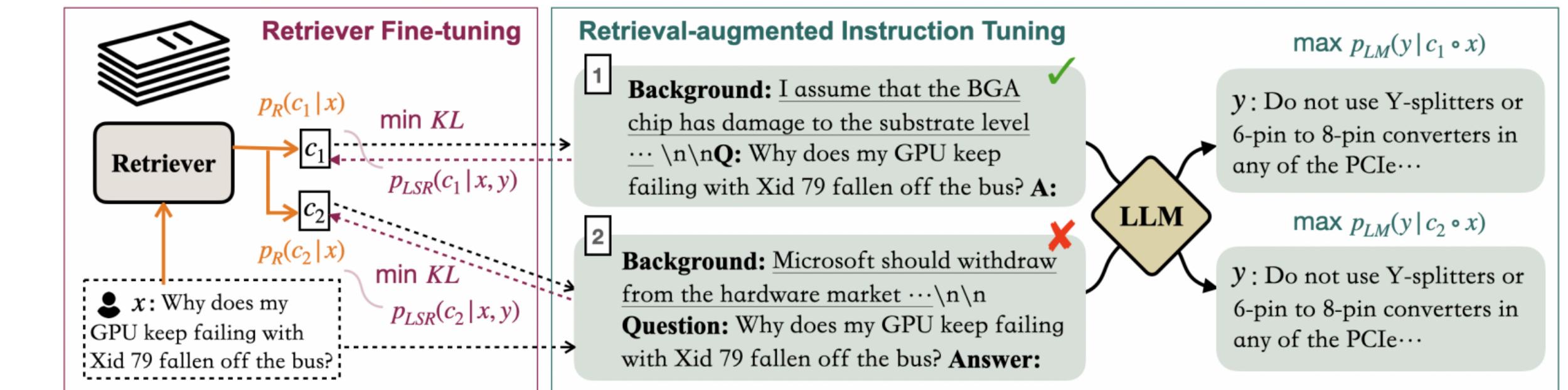
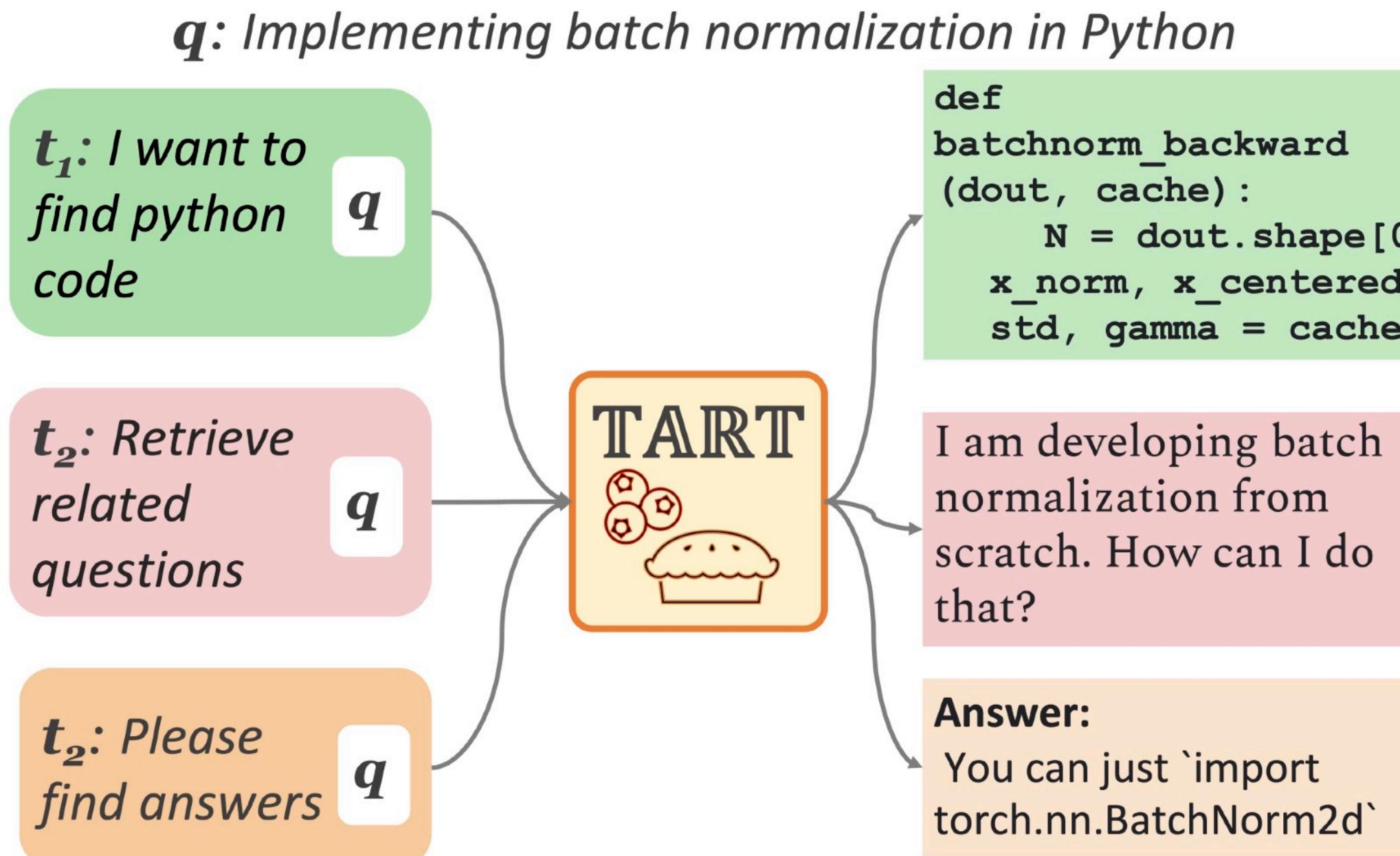
[3]:  nasa.gov (✓citation fully supports its associated statement)
[Studying the Next Interstellar Interloper with Webb](#)
...Scientists have had only limited ability to study these objects once discovered, but all of that is about to change with NASA's James Webb Space Telescope...The team will use Webb's spectroscopic capabilities in both the near-infrared and mid-infrared bands to study two different aspects of the interstellar object.

Retrieval-augmented LMs can still hallucinate

Liu et al. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023.

Beyond semantic and lexical-similarity based search

Training retrievers to optimize end-to-end retrieval-augmented LM performance in diverse tasks



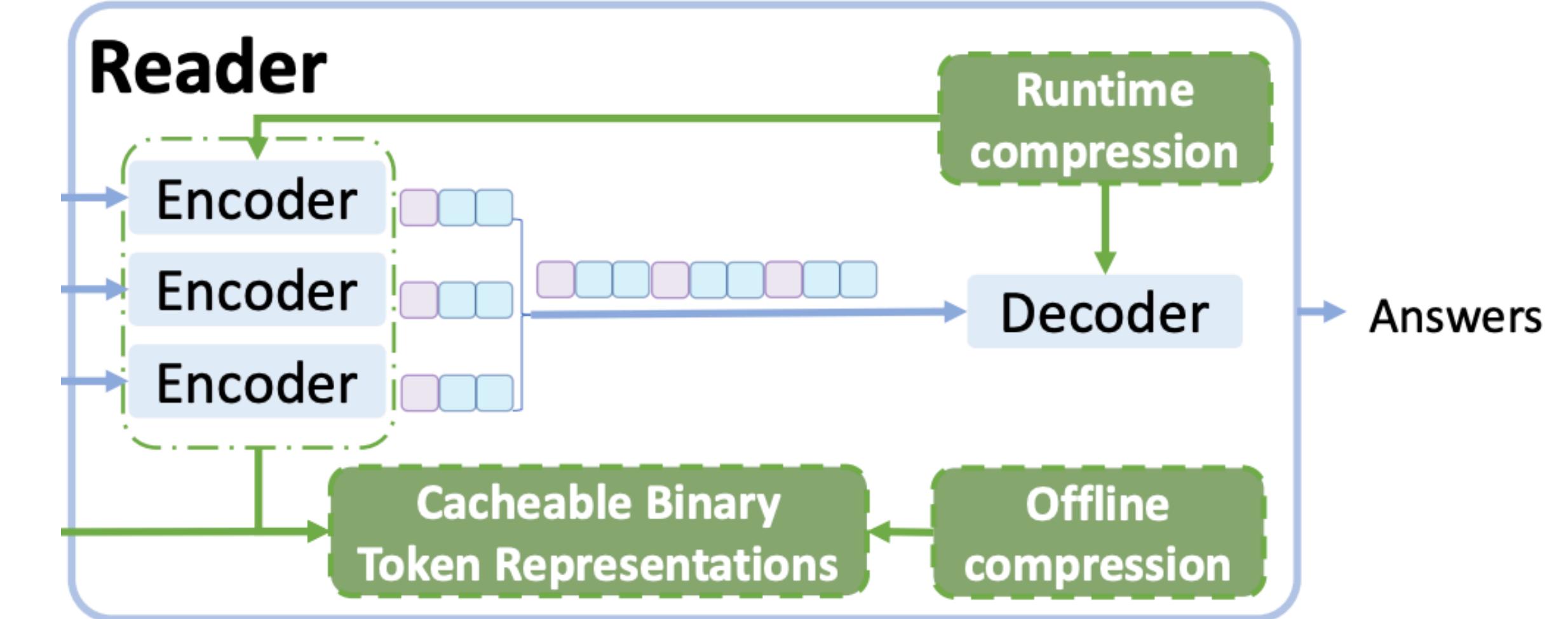
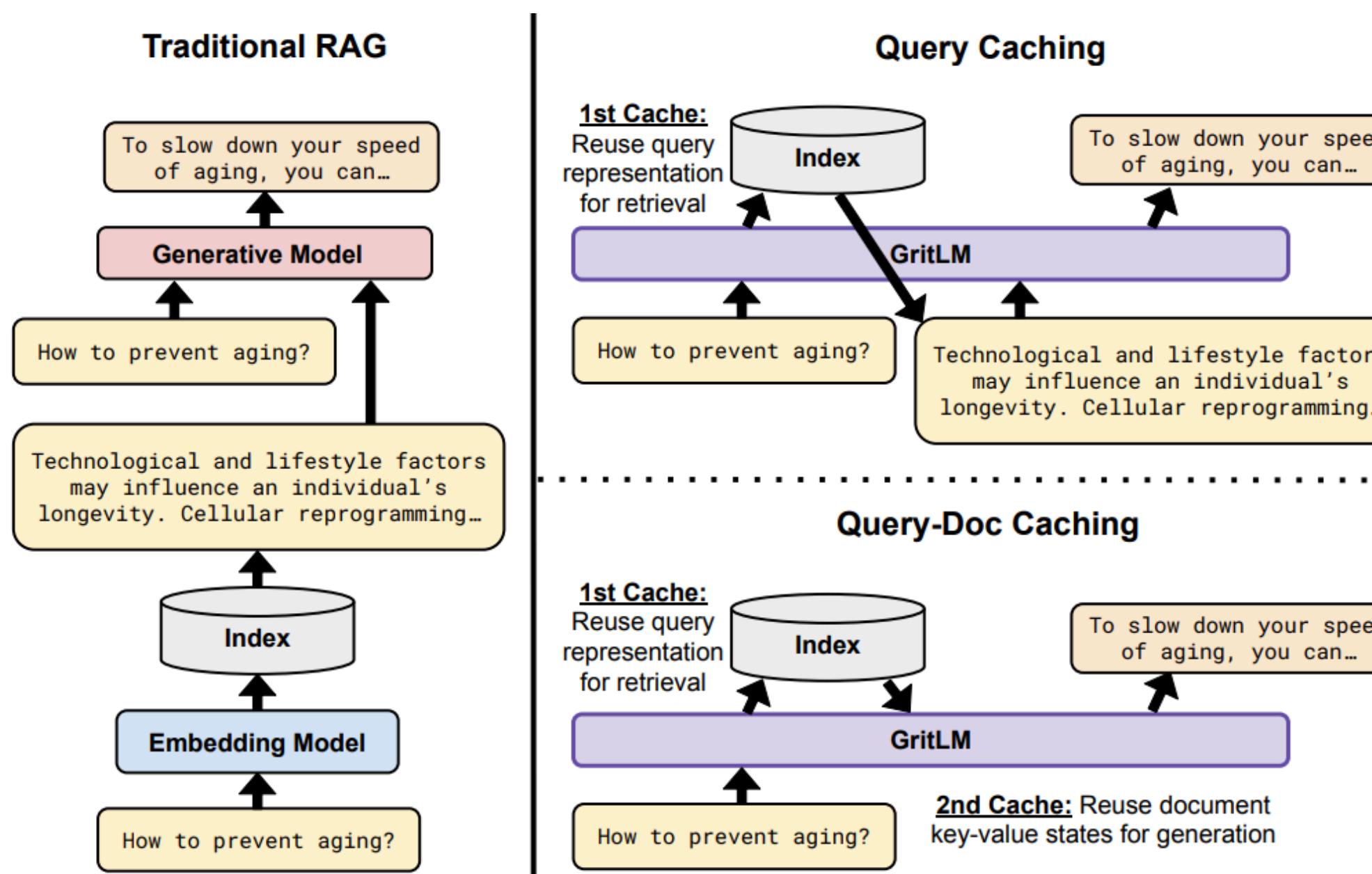
0-shot	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande
LLAMA 65B	85.3	82.8	52.3	84.2	77.0
RA-DIT 65B w/o retrieval	86.7	83.7	57.9	85.1	79.8
RA-DIT 65B	85.6	84.4	58.4	85.4	80.0

Asai et al., Task-aware Retrieval with Instruction.
Findings of ACL 2023.

Lin et al., RA-DIT: Retrieval-Augmented Dual Instruction Tuning.
ICLR 2024.

New architectures for performance and efficiency

Further explorations of unified architectures & caching

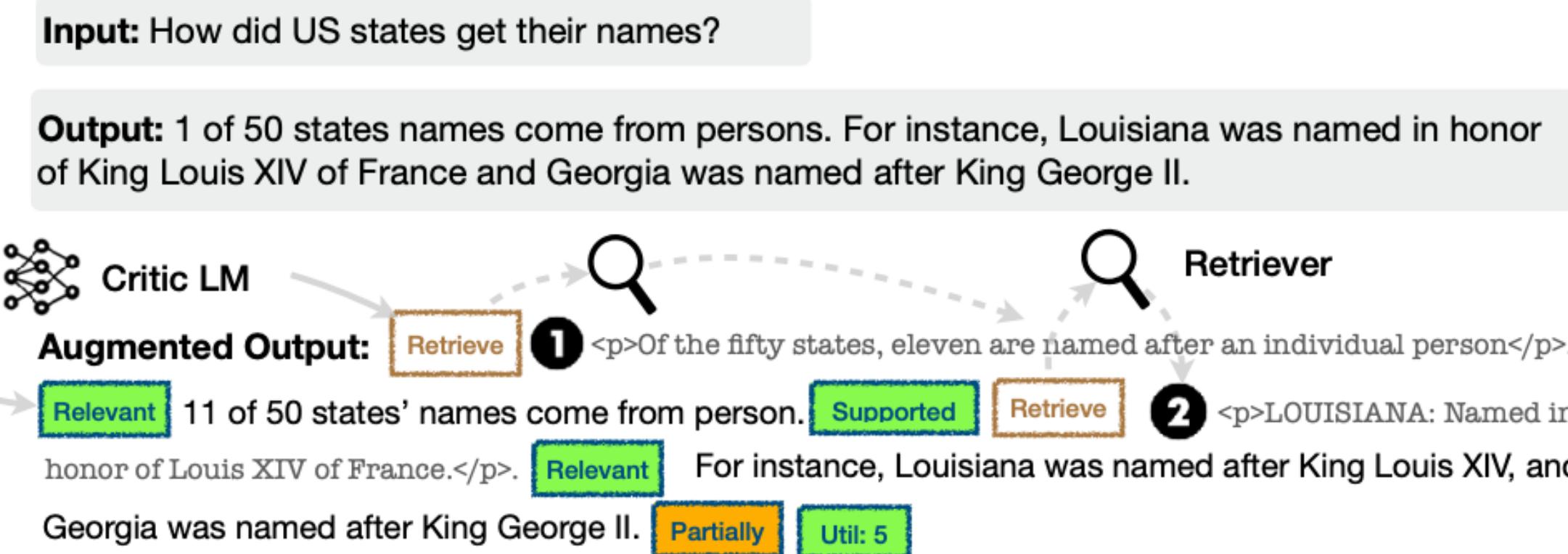


Muennighoff et al. Generative
Representational Instruction Tuning. 2024.

Cao et al. BTR: Binary Token
Representations for Efficient Retrieval
Augmented Language Models. ICLR 2024.

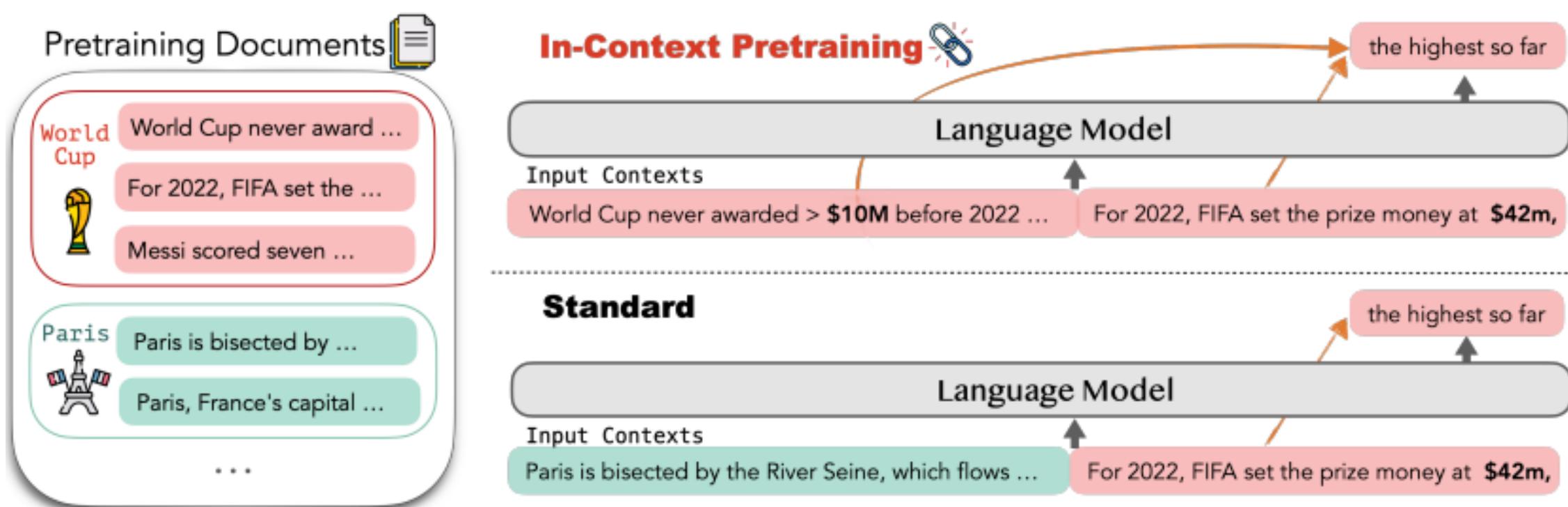
Training LMs with Retrieval

Training LMs to learn to use retrieval during pre-training or instruction-tuning



Instruction-tuning with retrieval

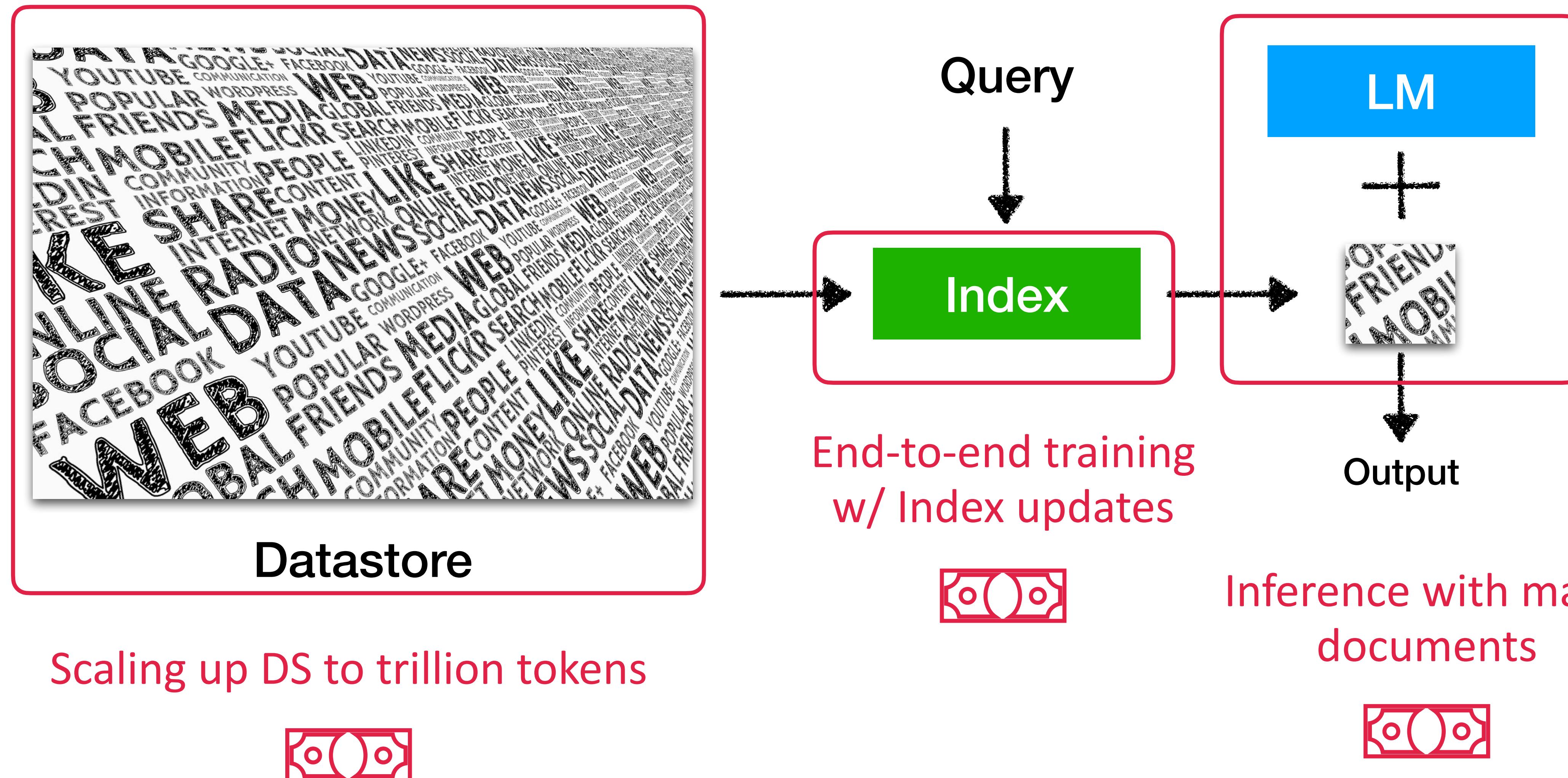
Asai et al. Self-RAG: Learning to Retrieve, Generate and Critique with Retrieval. ICLR 2024.



Retrieval-aware pre-training

Shi. et al. In-Context Pretraining: Language Modeling Beyond Document Boundaries. ICLR 2024.

Retrieval-augmented LMs can be really expensive!



References (1)

- Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen. Retrieval-based Language Models and Applications. ACL Tutorial 2023.
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. EMNLP 2021.
- Alex Mallen*, Akari Asai*, Victor Zhong, Rajarshi Das, Daniel Khashabi, Hannaneh Hajishirzi. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2023.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, Colin Raffel. Large Language Models Struggle to Learn Long-Tail Knowledge. ICML 2023.
- Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, Kentaro Inui. RealTime QA: What's the Answer Right Now?. NeurIPS (Dataset & Benchmark) 2023.

References (2)

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, Luke Zettlemoyer. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. ICLR 2024.

Juan Ramos. Using TF-IDF to Determine Word Relevance in Document Queries. The first instructional conference on machine learning 2003.

Robertson and Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval 2009.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP 2020.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer , Hannaneh Hajishirzi, Wen-tau Yih. Reliable, Adaptable, and Attributable Language Models with Retrieval. Arxiv 2024.

References (3)

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. ICML 2020.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, Yoav Shoham. In-Context Retrieval-Augmented Language Models. arXiv 2023.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih. REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv 2023.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, Laurent Sifre. Improving language models by retrieving from trillions of tokens. arXiv 2021.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis. Generalization through Memorization: Nearest Neighbor Language Models. ICLR 2020.

References (4)

Shufan Wang, Yixiao Song, Andrew Drozdov, Aparna Garimella, Varun Manjunatha, Mohit Iyyer. KNN-LM Does Not Improve Open-ended Text Generation. ACL 2023.

Parishad BehnamGhader, Santiago Miret, Siva Reddy. Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model. Findings of EMNLP 2023.

Nelson F. Liu, Tianyi Zhang, Percy Liang. Evaluating Verifiability in Generative Search Engines. Findings of EMNLP 2023.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, Wen-tau Yih. Task-aware Retrieval with Instructions. Findings of ACL 2023.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvassy, Mike Lewis, Luke Zettlemoyer, Scott Yih. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. ICLR 2024.

References (5)

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, Douwe Kiela.
Generative Representational Instruction Tuning. Arxiv 2024.

Qingqing Cao, Sewon Min, Yizhong Wang, Hannaneh Hajishirzi. BTR: Binary Token Representations for Efficient Retrieval Augmented Language Models. ICLR 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. ICLR 2024.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, Mike Lewis. In-Context Pretraining: Language Modeling Beyond Document Boundaries. ICLR 2024.