

Question:
[https://tinyurl.com/
ANLP20250213](https://tinyurl.com/ANLP20250213)



Retrieval and Retrieval-Augmented Generation

Akari Asai

<https://akariasai.github.io/>

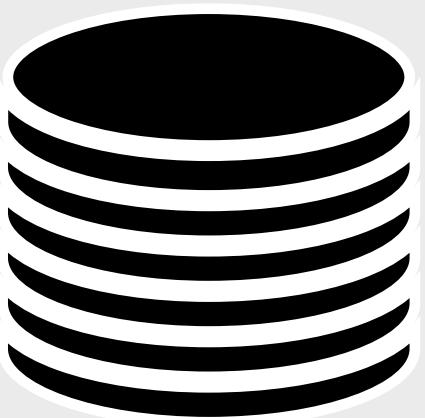
February 1, 2025 at CMU Advanced NLP

Slides adapted from

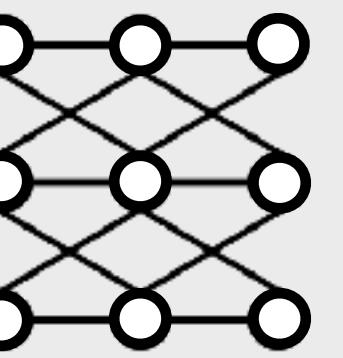
ACL 2023 tutorial by Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen <https://acl2023-retrieval-lm.github.io/>
Advanced NLP Fall 2024 by Prof. Graham Neubig <https://phontron.com/class/anlp-fall2024/>

Recap: How Do LMs Work?

Pre-training Data

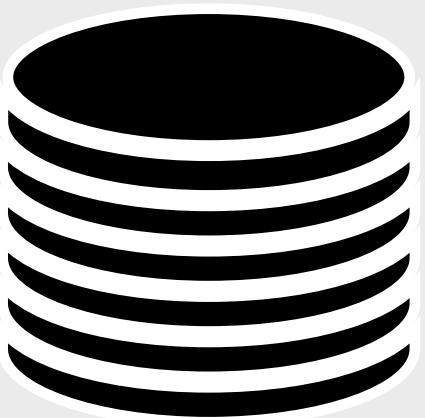


$\text{LM } \theta$



Recap: How Do LMs Work?

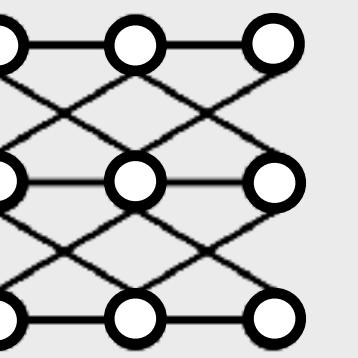
Pre-training Data



ChatGPT is developed
by OpenAI

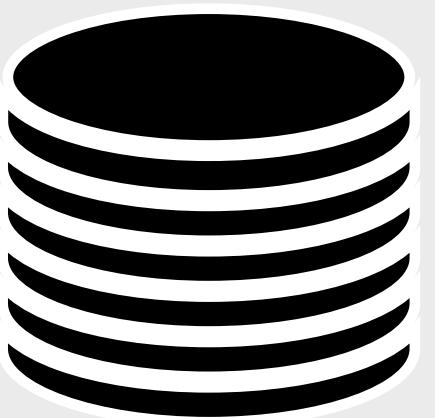


$\text{LM } \theta$



Recap: How Do LMs Work?

Pre-training Data



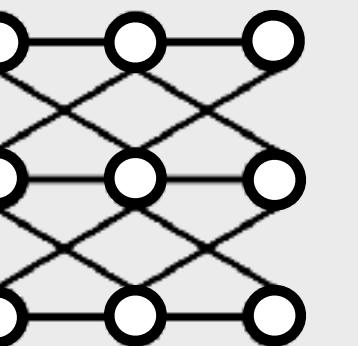
ChatGPT is developed
by OpenAI



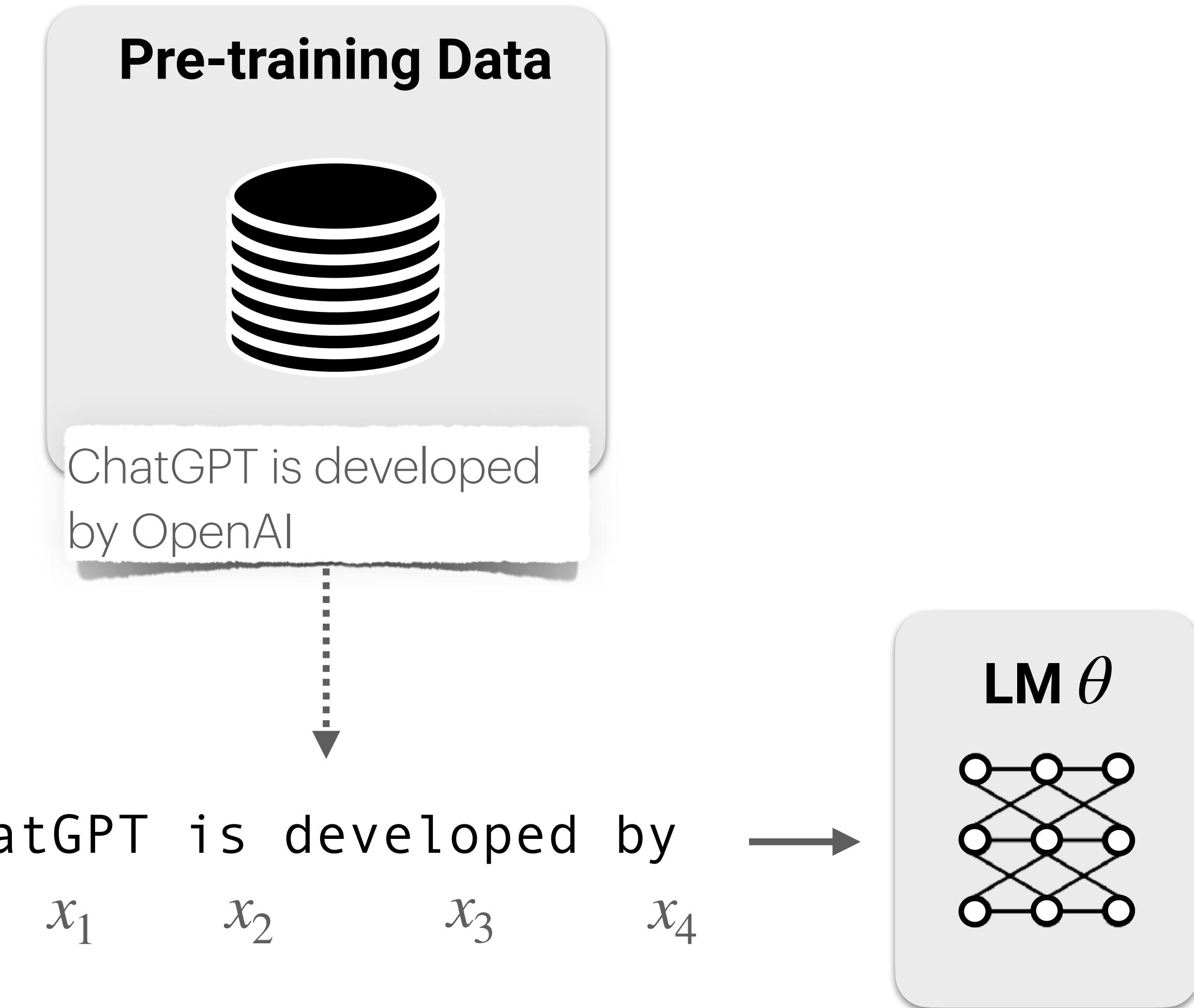
ChatGPT is developed by

$x_1 \quad x_2 \quad x_3 \quad x_4$

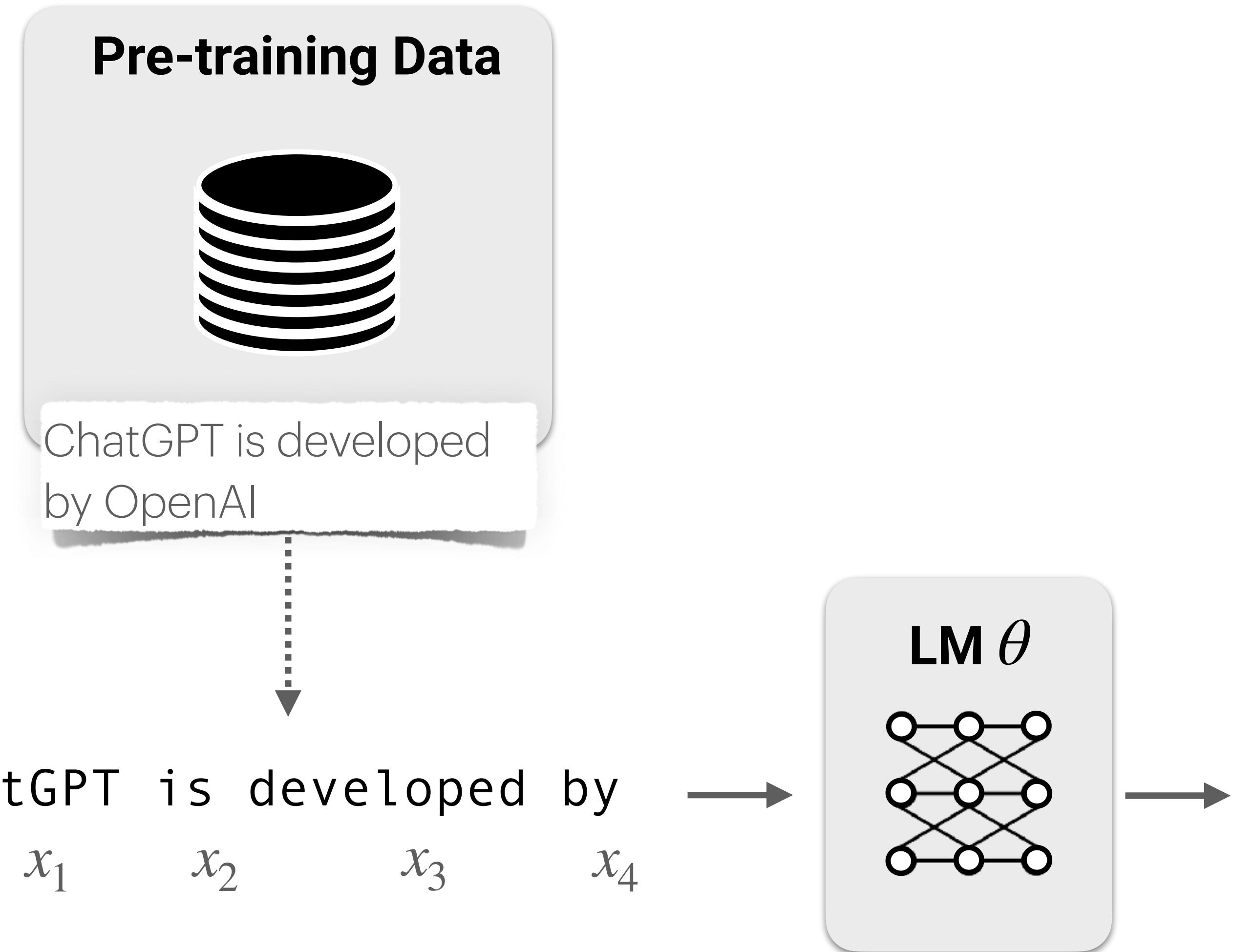
LM θ



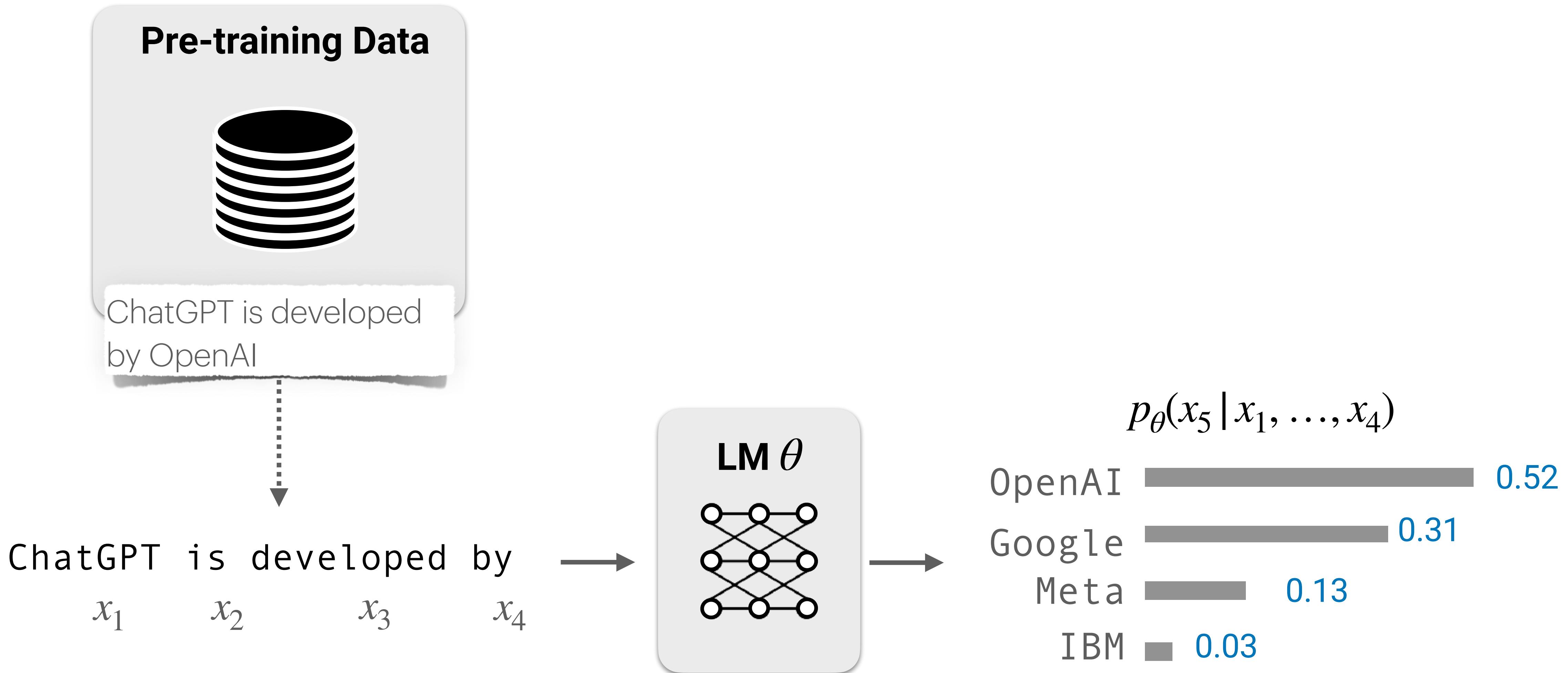
Recap: How Do LMs Work?



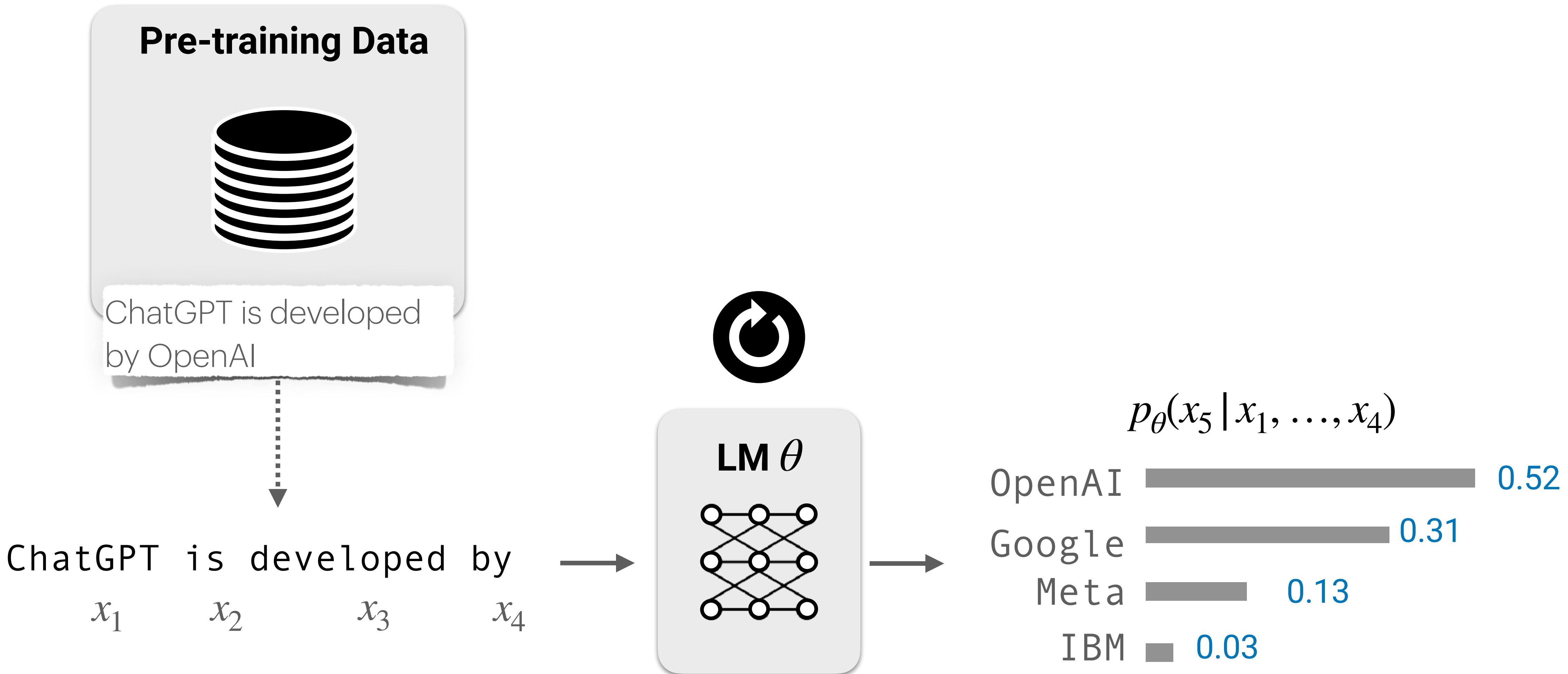
Recap: How Do LMs Work?



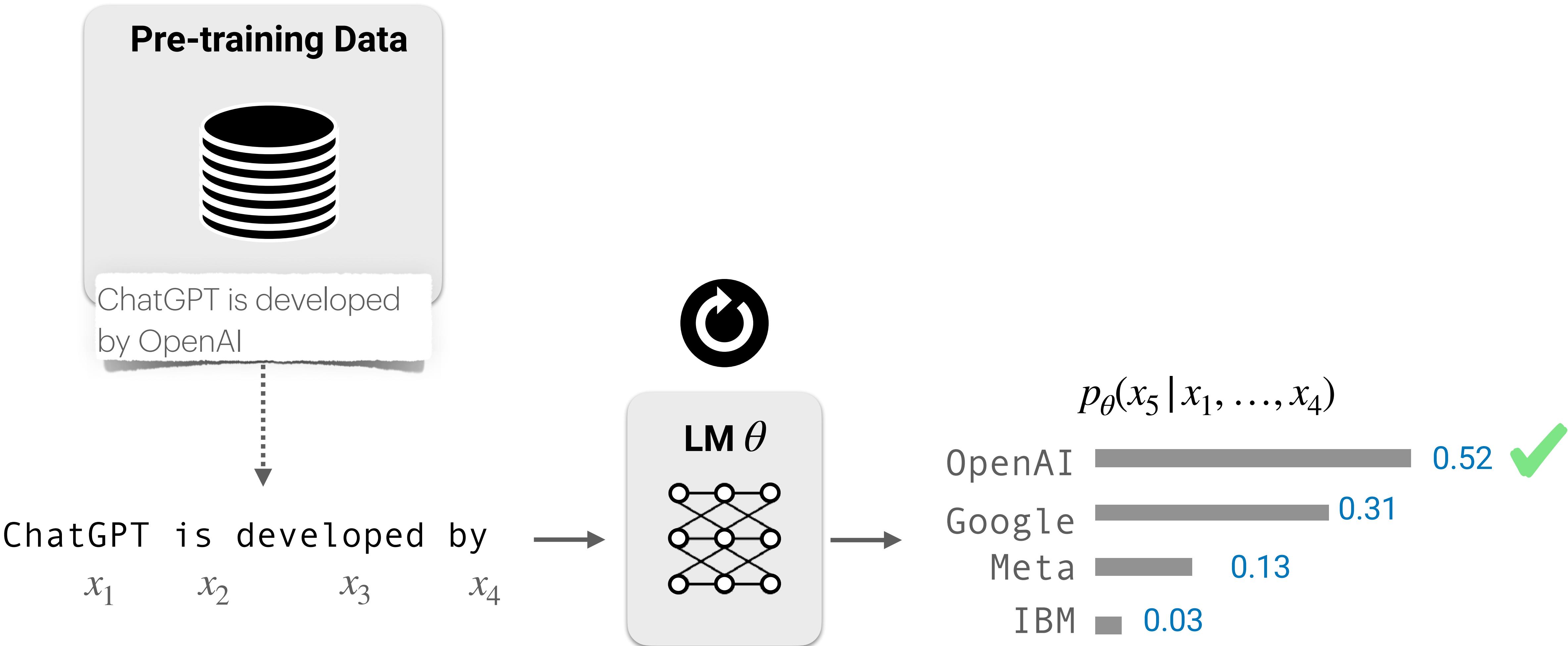
Recap: How Do LMs Work?



Recap: How Do LMs Work?



Recap: How Do LMs Work?



Limitations of Current LMs: Hallucinations

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Akari Asai is a researcher in natural language processing (NLP) and artificial intelligence (AI), particularly focusing on open-domain question answering, knowledge retrieval, and multilingual NLP. Her work primarily aims to improve how machines retrieve, understand, and reason over textual information efficiently.

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Akari Asai is a researcher in natural language processing (NLP) and artificial intelligence (AI), particularly focusing on open-domain question answering, knowledge retrieval, and multilingual NLP. Her work primarily aims to improve how machines retrieve, understand, and reason over textual information efficiently.

Representative Papers:

1. "Learning to Retrieve Passages without Supervision" (NeurIPS 2022)

- Introduces a method for training retrieval models without labeled data, improving open-domain QA efficiency.
- Leverages contrastive learning to align questions and relevant passages.

2. "One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Akari Asai is a researcher in natural language processing (NLP) and artificial intelligence (AI), particularly focusing on open-domain question answering, knowledge retrieval, and information extraction. Her work primarily aims to improve how machines retrieve, understand, and reason about information efficiently.

Learning to Retrieve Passages without Supervision

Ori Ram Gal Shachaf Omer Levy Jonathan Berant Amir Globerson

Blavatnik School of Computer Science, Tel Aviv University

ori.ram@cs.tau.ac.il

Representative Papers:

1. "Learning to Retrieve Passages without Supervision" (NeurIPS 2022)



Not my paper

- Introduces a method for training retrieval models without labeled data, improving open-domain QA efficiency.
- Leverages contrastive learning to align questions and relevant passages.

2. "One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Akari Asai is a researcher in natural language processing (NLP) and artificial intelligence (AI), particularly focusing on open-domain question answering, knowledge retrieval, and information extraction. Her work primarily aims to improve how machines retrieve, understand, and reason about information efficiently.

Learning to Retrieve Passages without Supervision

Ori Ram Gal Shachaf Omer Levy Jonathan Berant Amir Globerson

Blavatnik School of Computer Science, Tel Aviv University

ori.ram@cs.tau.ac.il

Representative Papers:

NAACL 2022



Not my paper

1. "Learning to Retrieve Passages without Supervision" (NeurIPS 2022)

- Introduces a method for training retrieval models without labeled data, improving open-domain QA efficiency.
- Leverages contrastive learning to align questions and relevant passages.

2. "One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Akari Asai is a researcher in natural language processing (NLP) and artificial intelligence (AI), particularly focusing on open-domain question answering, knowledge retrieval, and information extraction. Her work primarily aims to improve how machines retrieve, understand, and reason about information efficiently.

Learning to Retrieve Passages without Supervision

Ori Ram Gal Shachaf Omer Levy Jonathan Berant Amir Globerson

Blavatnik School of Computer Science, Tel Aviv University

ori.ram@cs.tau.ac.il

Representative Papers:

NAACL 2022

1. "Learning to Retrieve Passages without Supervision" (~~NeurIPS 2022~~)

- Introduces a method for training retrieval models without labeled data, improving open-domain QA efficiency.
- Leverages contrastive learning to align questions and relevant passages.



Not my paper



Venue is wrong

2. "One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Akari Asai is a researcher in natural language processing (NLP) and artificial intelligence (AI), particularly focusing on open-domain question answering, knowledge retrieval, and information extraction. Her work primarily aims to improve how machines retrieve, understand, and reason about information efficiently.

Learning to Retrieve Passages without Supervision

Ori Ram Gal Shachaf Omer Levy Jonathan Berant Amir Globerson

Blavatnik School of Computer Science, Tel Aviv University

ori.ram@cs.tau.ac.il

Representative Papers:

NAACL 2022

1. "Learning to Retrieve Passages without Supervision" (NeurIPS 2022)

- Introduces a method for training retrieval models without labeled data, improving open-domain QA efficiency.
- Leverages contrastive learning to align questions and relevant passages.

 **Not my paper**

 **Venue is wrong**

2. "~~One Question, Many Answers: A Retrieval-based Multimodal QA Dataset~~" (EMNLP 2022)

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Akari Asai is a researcher in natural language processing (NLP) and artificial intelligence (AI), particularly focusing on open-domain question answering, knowledge retrieval, and information extraction. Her work primarily aims to improve how machines retrieve, understand, and reason about information efficiently.

Learning to Retrieve Passages without Supervision

Ori Ram Gal Shachaf Omer Levy Jonathan Berant Amir Globerson

Blavatnik School of Computer Science, Tel Aviv University

ori.ram@cs.tau.ac.il

Representative Papers:

NAACL 2022

1. "Learning to Retrieve Passages without Supervision" (NeurIPS 2022)

- Introduces a method for training retrieval models without labeled data, improving open-domain QA efficiency.
- Leverages contrastive learning to align questions and relevant passages.

 **Not my paper**

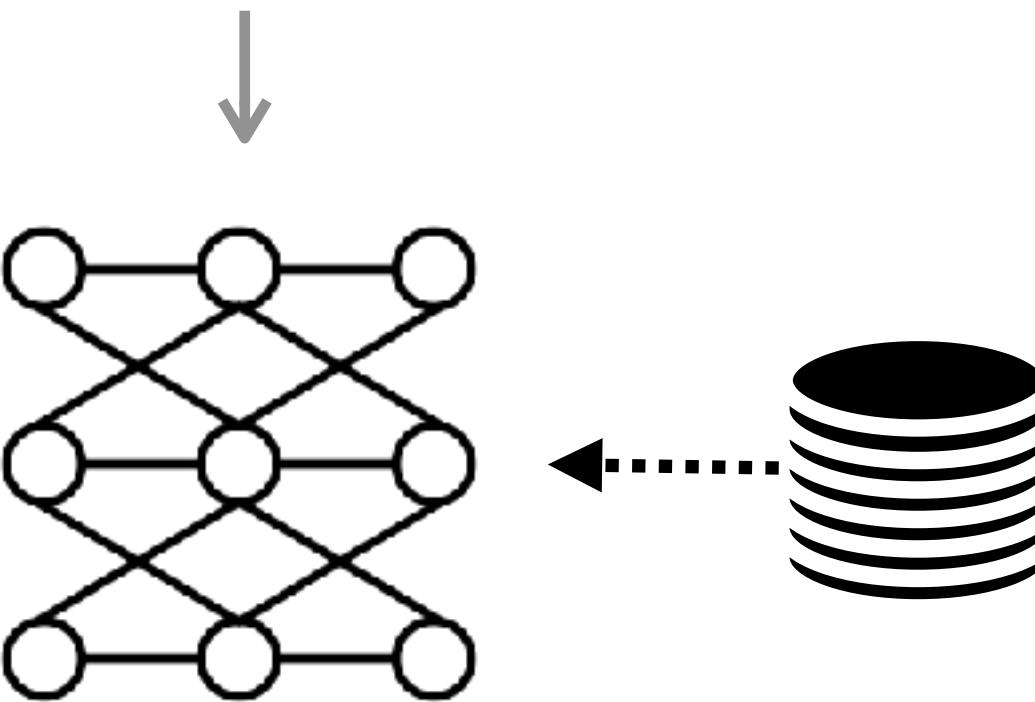
 **Venue is wrong**

2. "~~One Question, Many Answers: A Retrieval-based Multimodal QA Dataset~~" (EMNLP 2022)

LMs struggle in long-tail knowledge

Retrieval-Augmented LMs: Intuition

Who is Akari Asai?



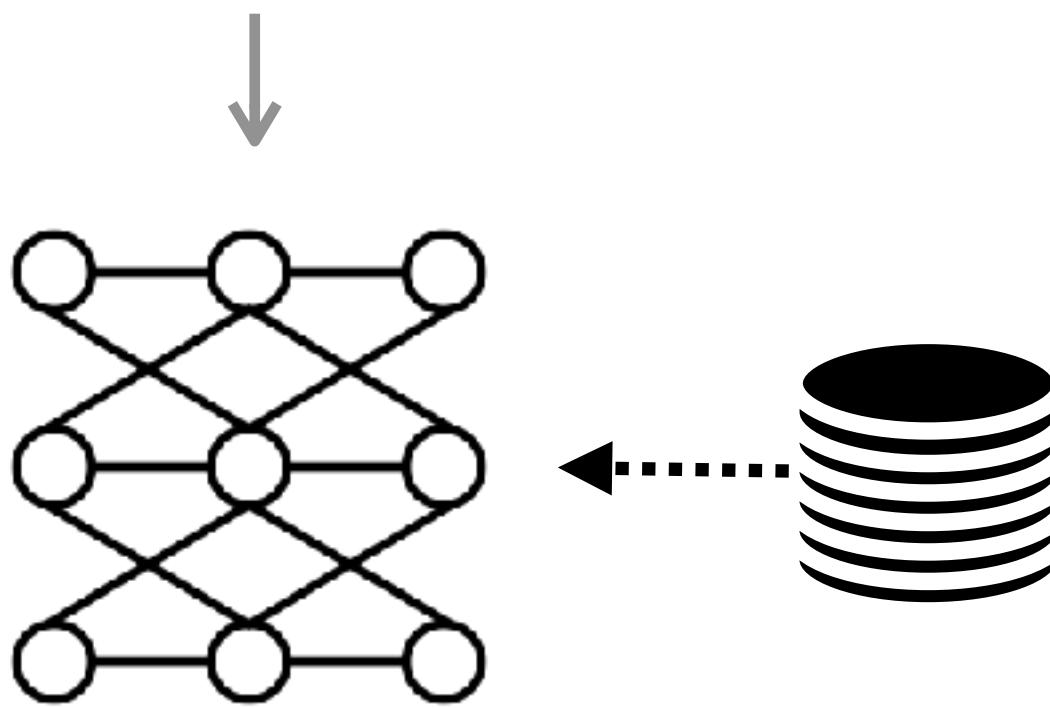
LM

Japanese voice actress

Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington

Retrieval-Augmented LMs: Intuition

Who is Akari Asai?

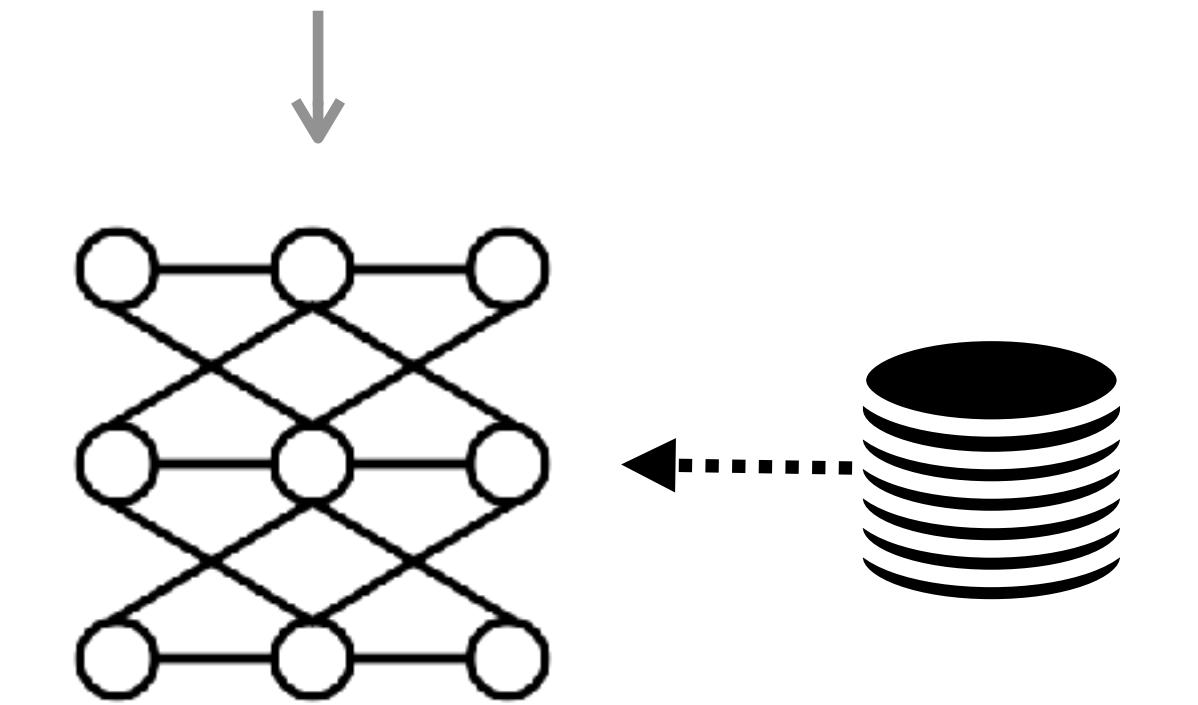


LM

Japanese voice actress X

Retrieval-Augmented LMs: Intuition

Who is Akari Asai?

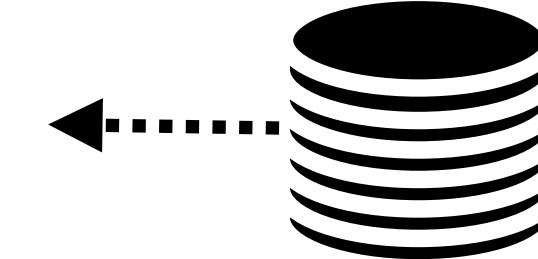
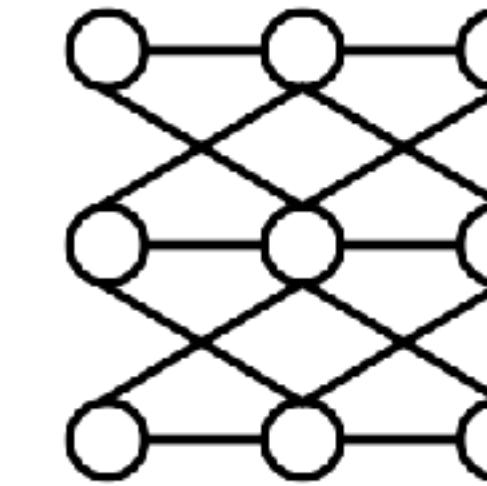


Retrieval-Augmented LMs: Intuition

Who is Akari Asai?



Who is Akari Asai?



LM

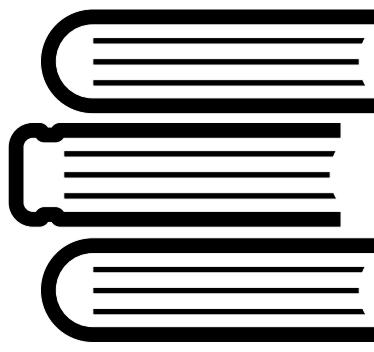


Japanese voice actress



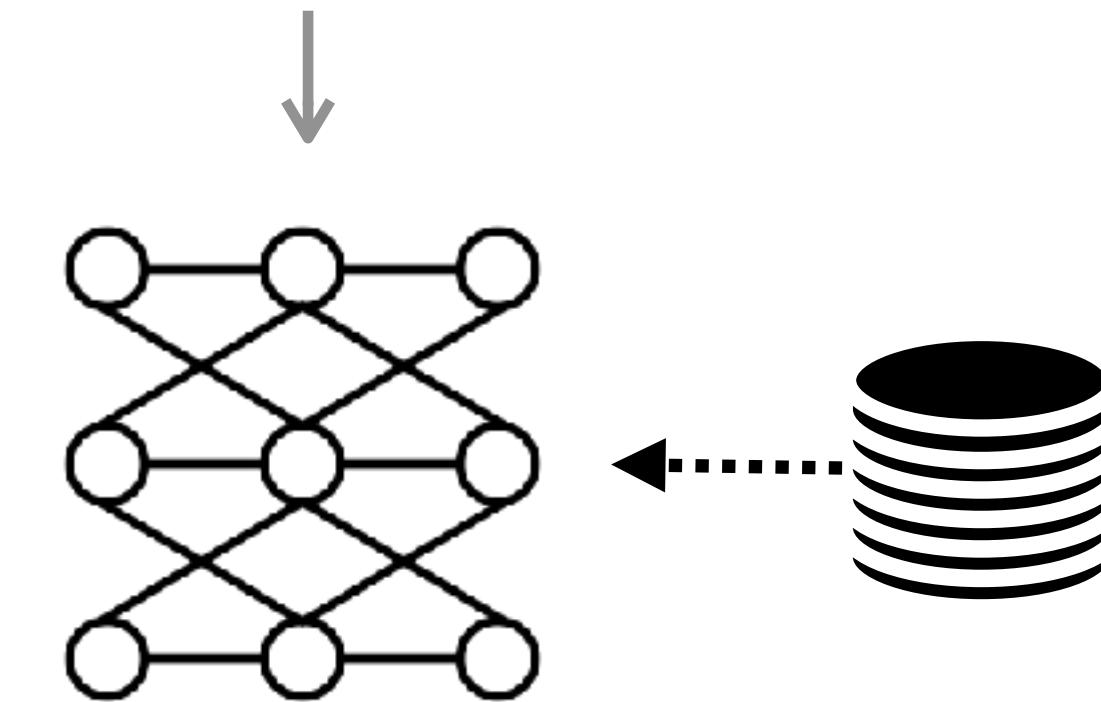
Retrieval-Augmented LMs: Intuition

Who is Akari Asai?



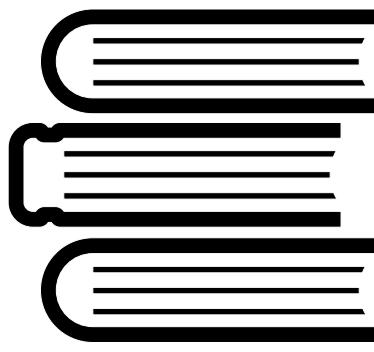
Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington

Who is Akari Asai?

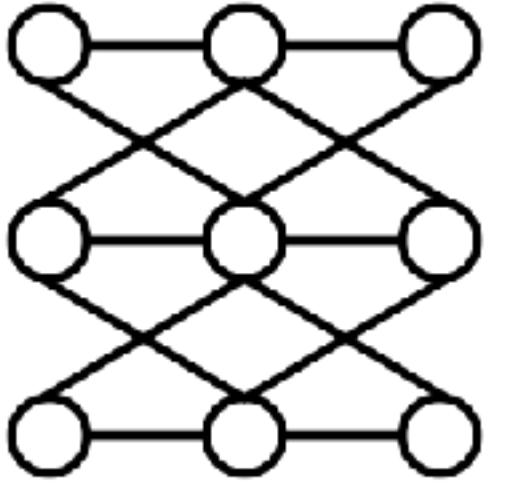


Retrieval-Augmented LMs: Intuition

Who is Akari Asai?



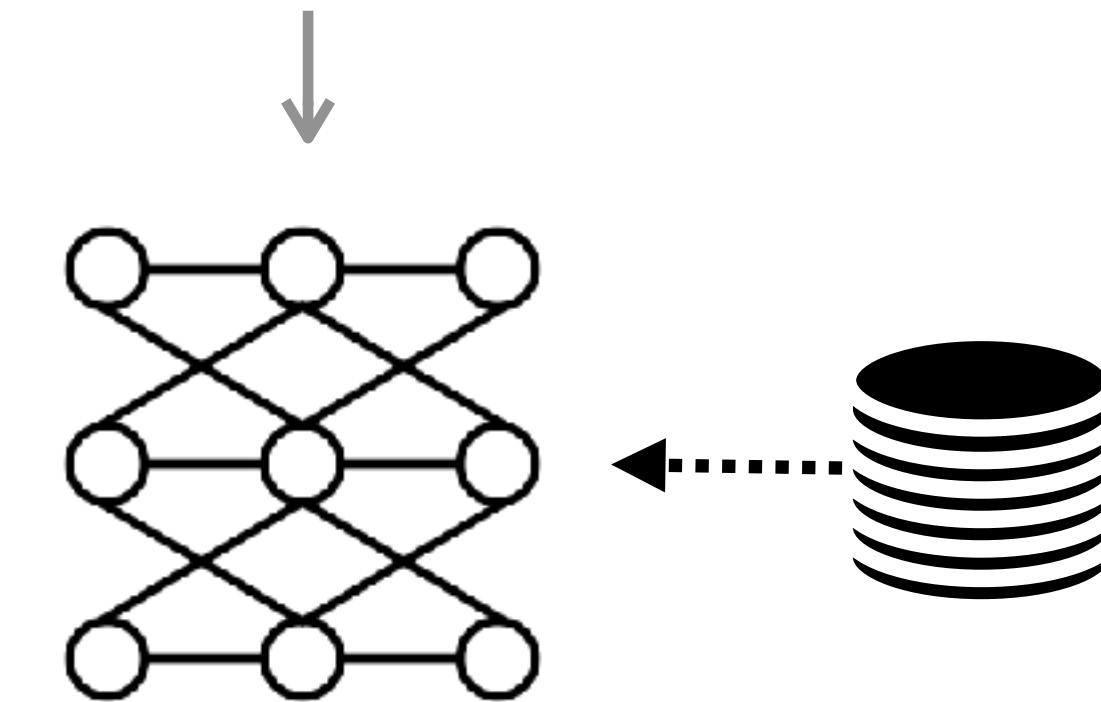
Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington



LM

Ph.D. student at UW

Who is Akari Asai?



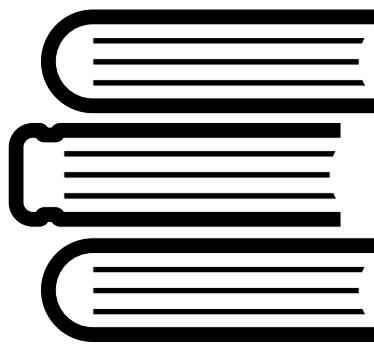
LM

Japanese voice actress

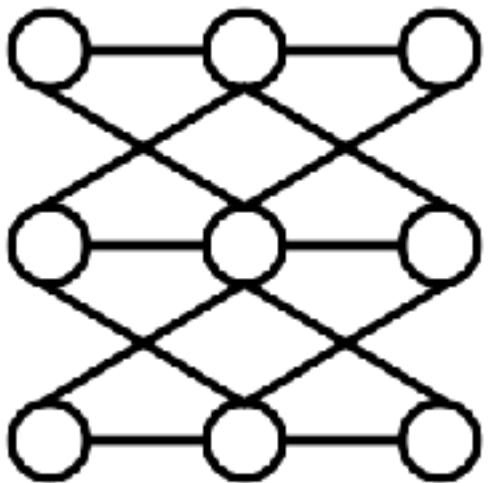


Retrieval-Augmented LMs: Intuition

Who is Akari Asai?



Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington

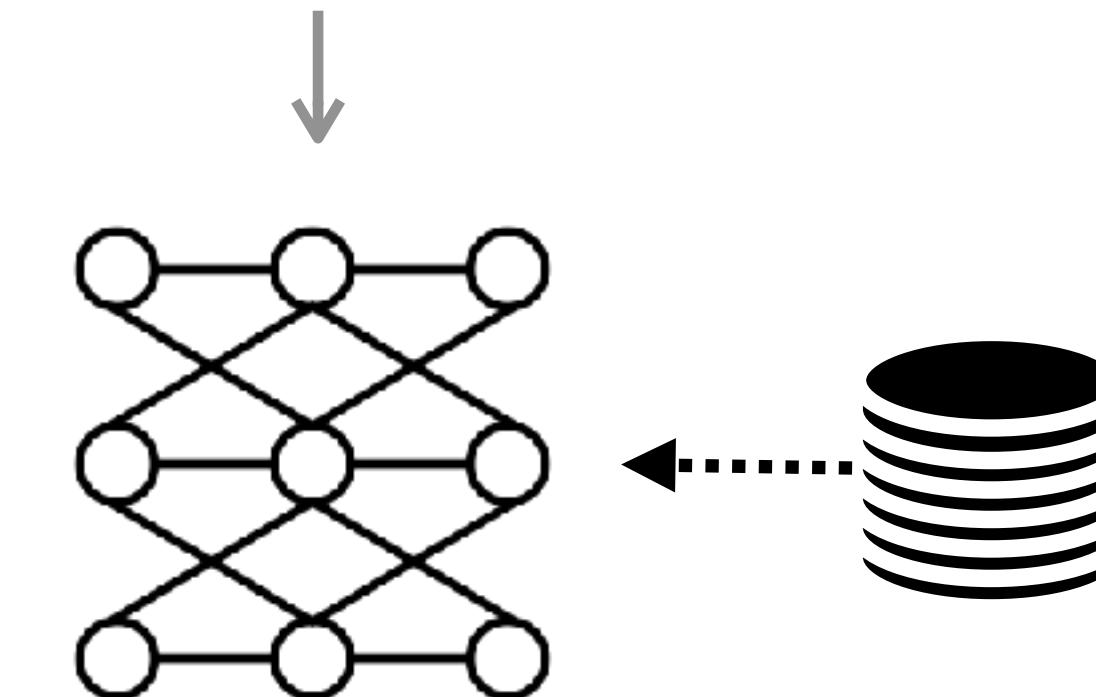


LM

Ph.D. student at UW



Who is Akari Asai?



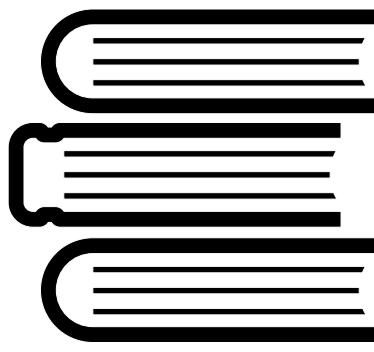
LM

Japanese voice actress

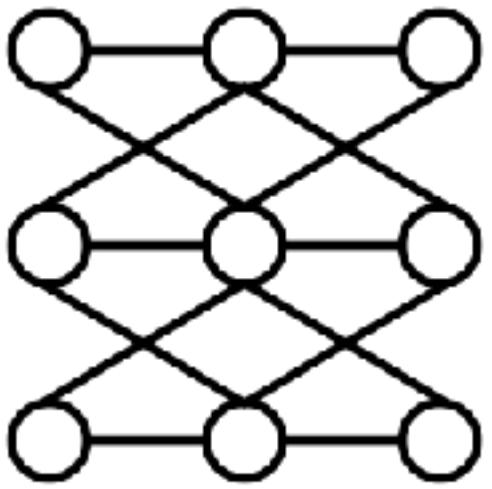


Retrieval-Augmented LMs: Intuition

Who is Akari Asai?



Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington



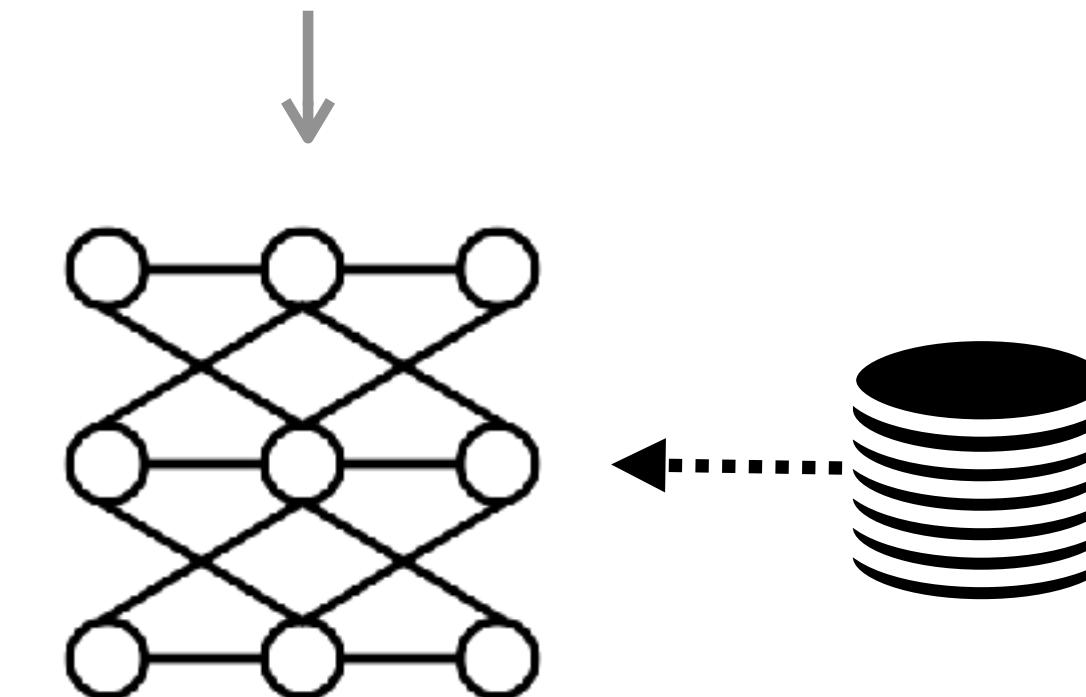
LM



Ph.D. student at UW



Who is Akari Asai?



LM



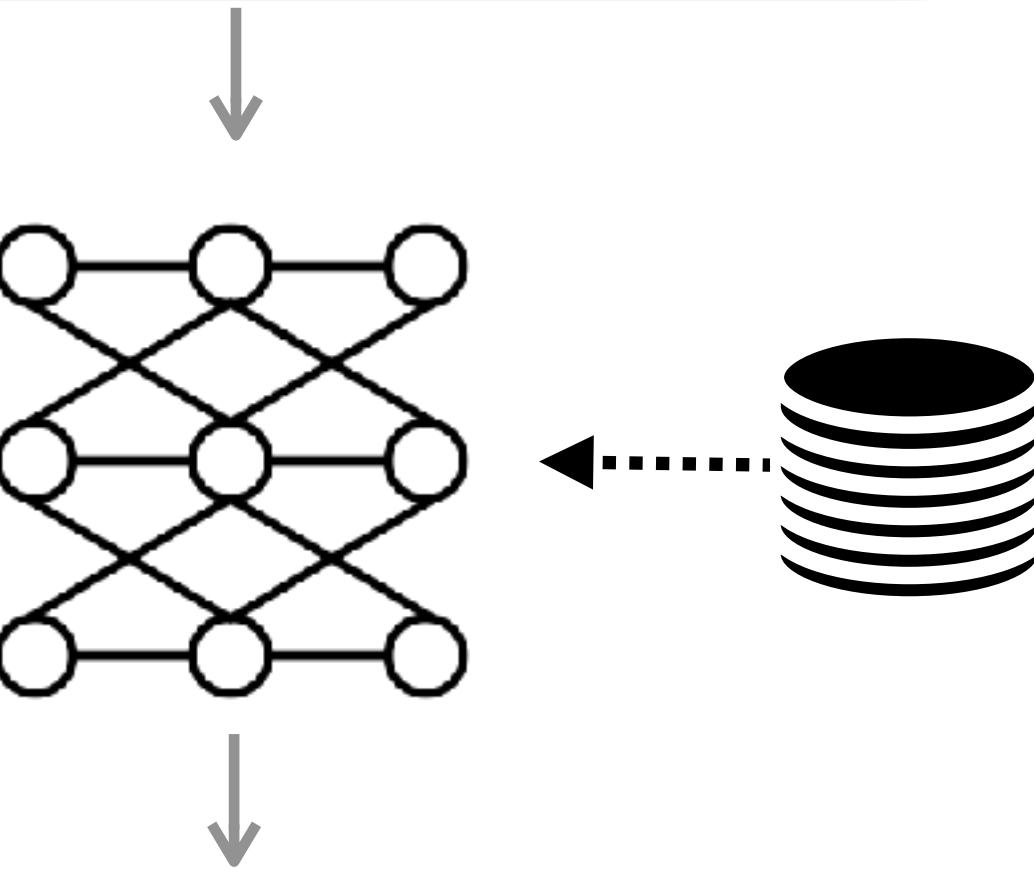
Japanese voice actress



Reduces hallucinations

Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



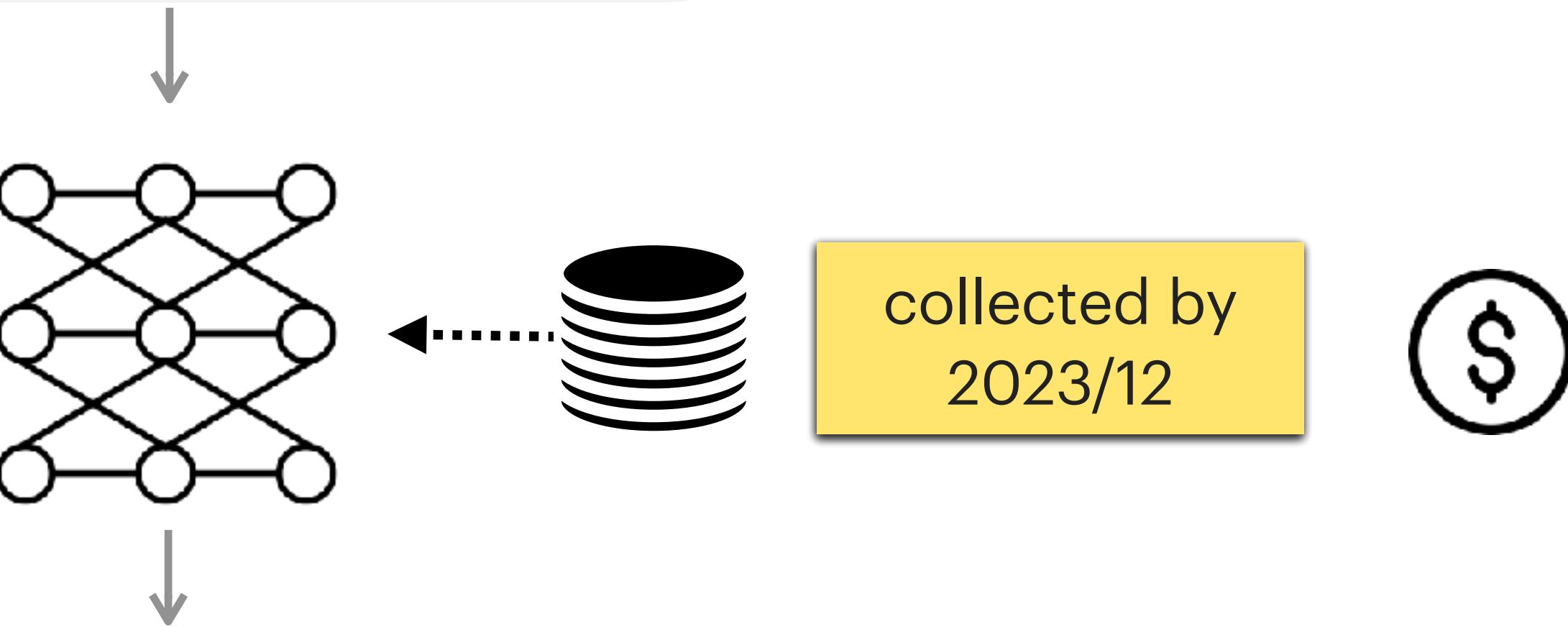
collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called “GPT4o.”



Retrieval-Augmented LMs: Intuition

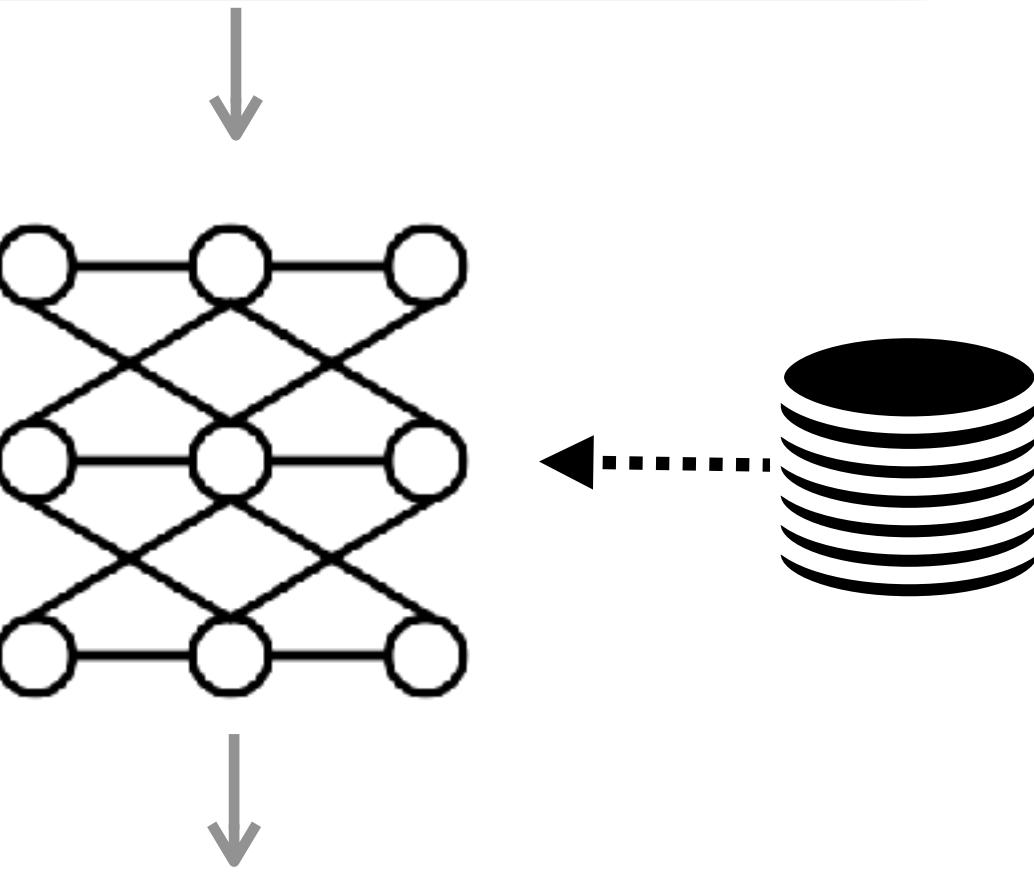
Who developed GPT4o?



Based on general knowledge, there is no widely recognized AI model or term specifically called “GPT4o.”

Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



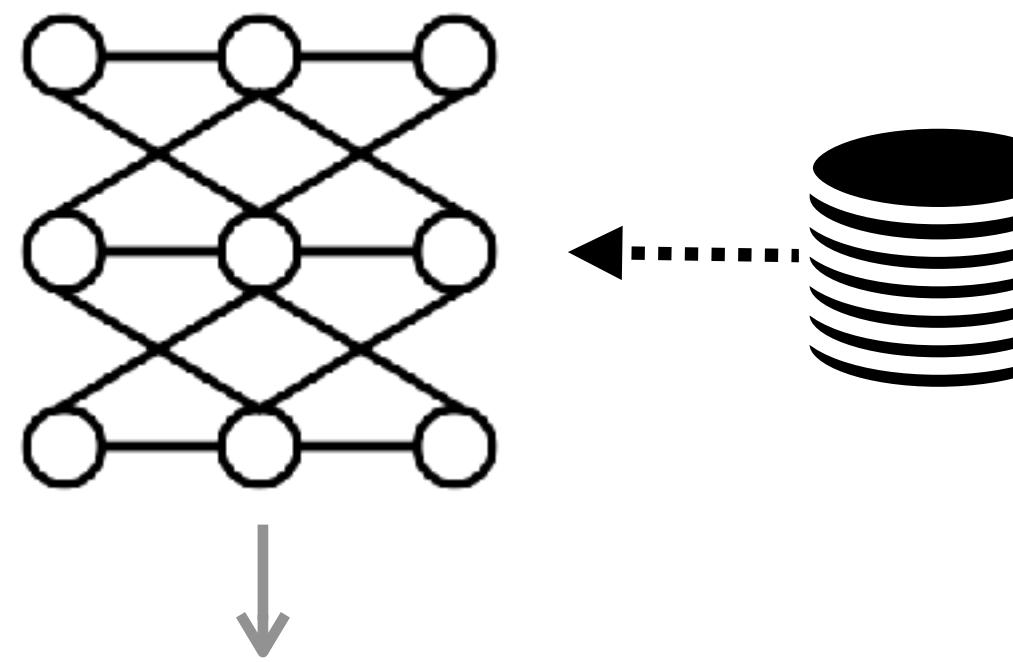
collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called “GPT4o.”



Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o."

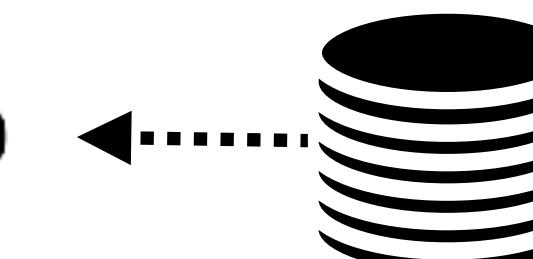
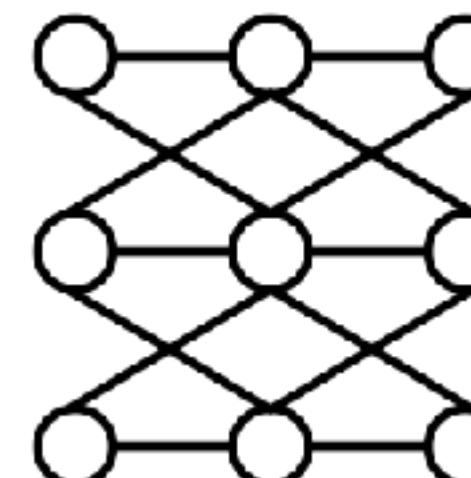


Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



Who developed GPT4o?



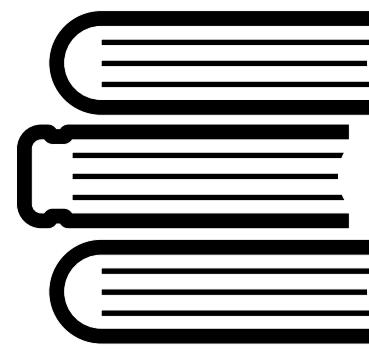
collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o."



Retrieval-Augmented LMs: Intuition

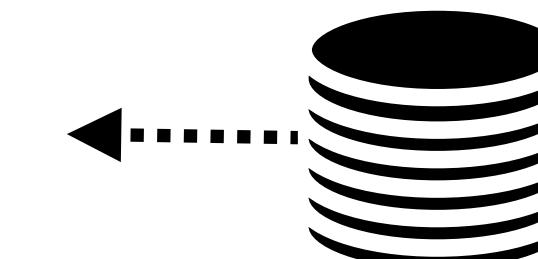
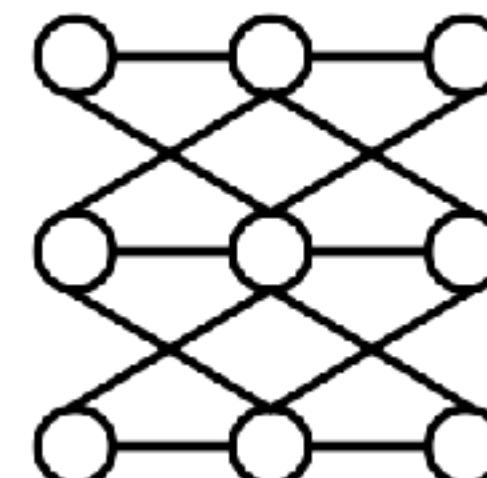
Who developed GPT4o?



GPT-4o is a Multilingual and Multimodal transformer-based model developed by OpenAI and released in May 2024.

Updated by
2024/01

Who developed GPT4o?



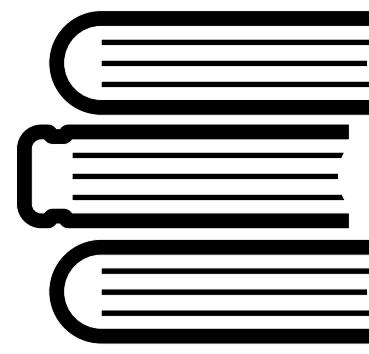
collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o."



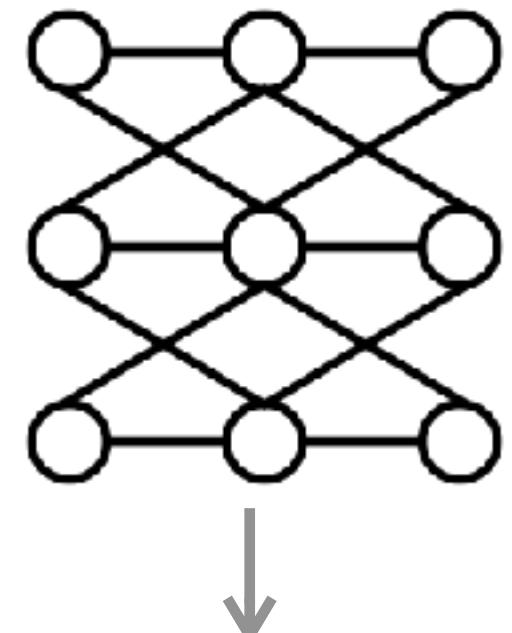
Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



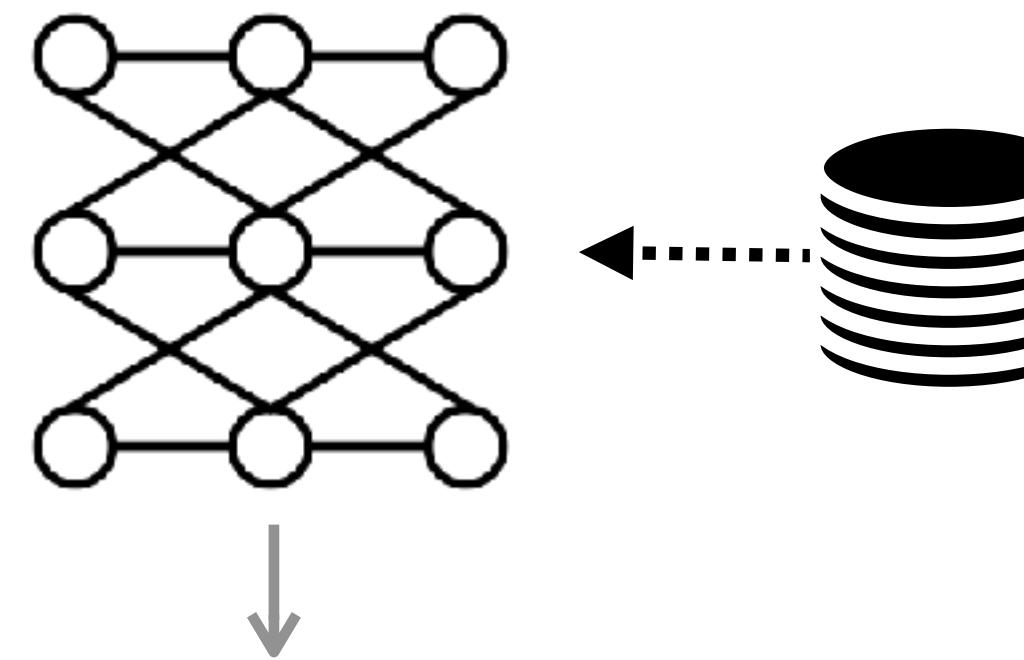
GPT-4o is a Multilingual and Multimodal transformer-based model developed by OpenAI and released in May 2024.

Updated by
2024/01



GPT4o is a new model released by OpenAI in May 2024.

Who developed GPT4o?

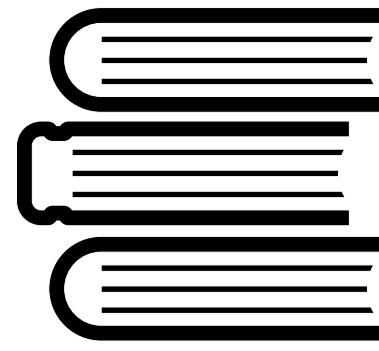


collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o."

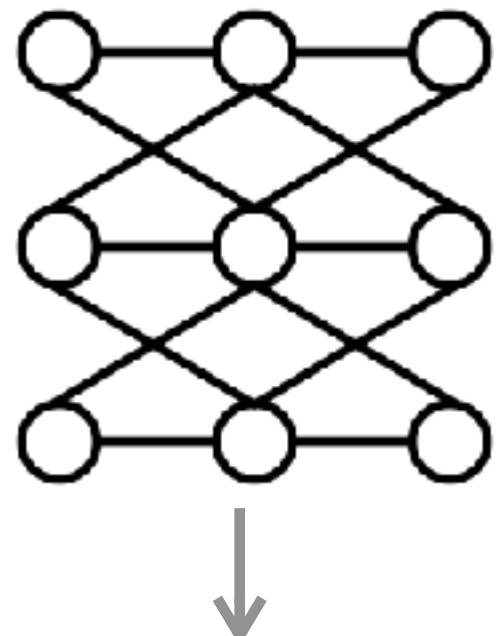
Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



GPT-4o is a Multilingual and Multimodal transformer-based model developed by OpenAI and released in May 2024.

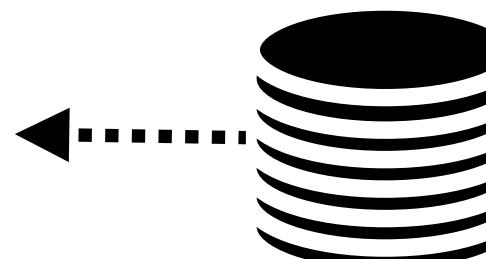
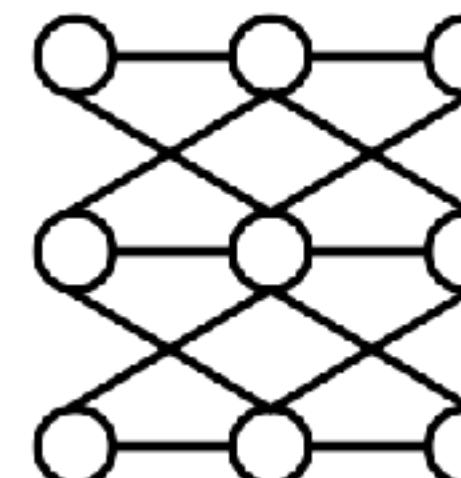
Updated by
2024/01



GPT4o is a new model released by OpenAI in May 2024.



Who developed GPT4o?



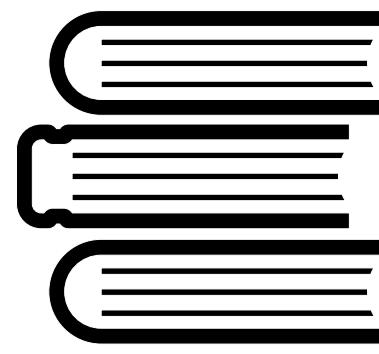
collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o."



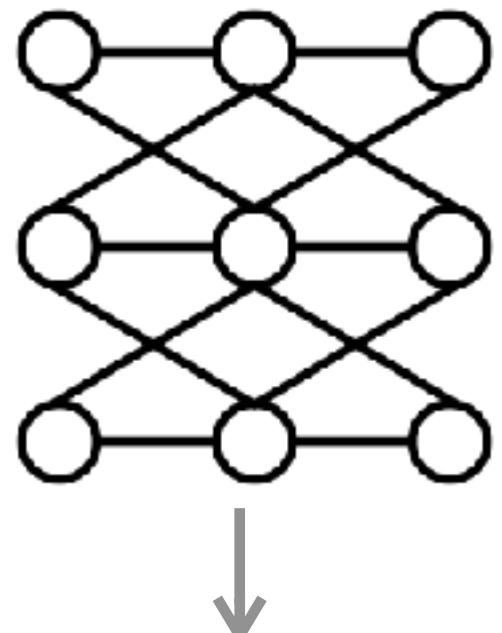
Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



GPT-4o is a Multilingual and Multimodal transformer-based model developed by OpenAI and released in May 2024.

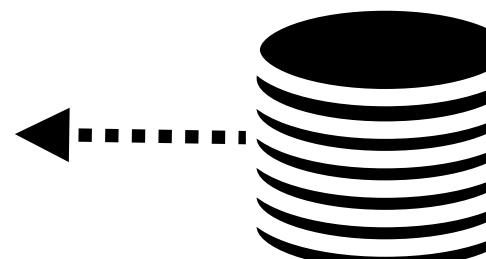
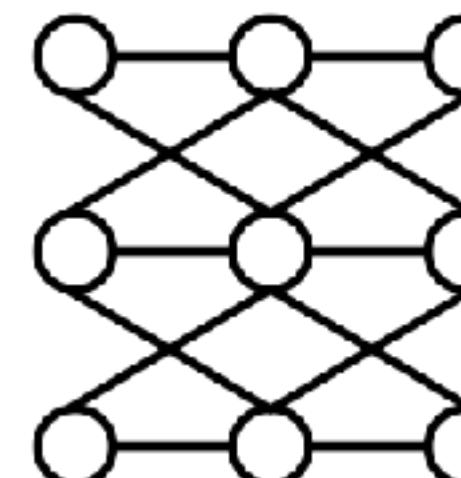
Updated by
2024/01



GPT4o is a new model released by OpenAI in May 2024.



Who developed GPT4o?



collected by
2023/12

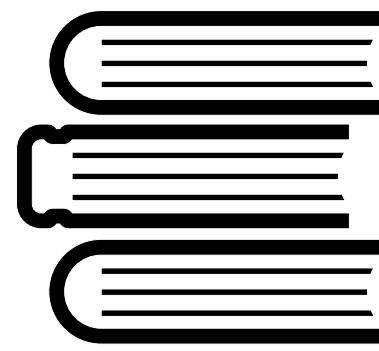
Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o."



Update knowledge w/o retraining

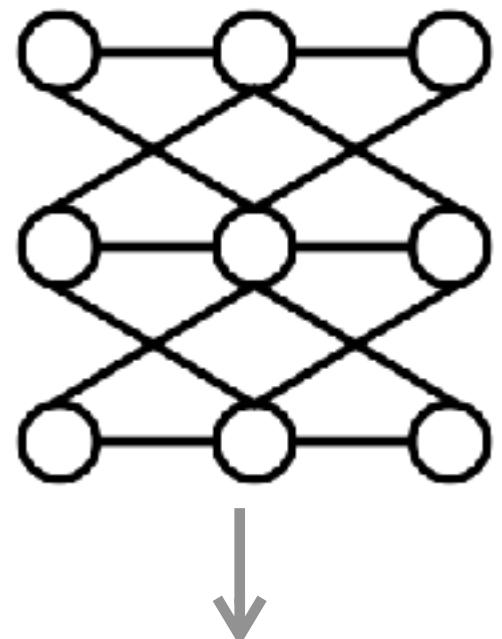
Retrieval-Augmented LMs: Intuition

Who developed GPT4o?



GPT-4o is a Multilingual and Multimodal transformer-based model developed by OpenAI and released in May 2024.

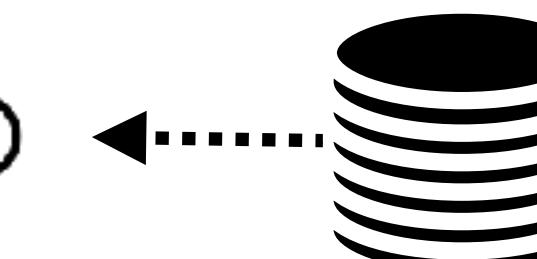
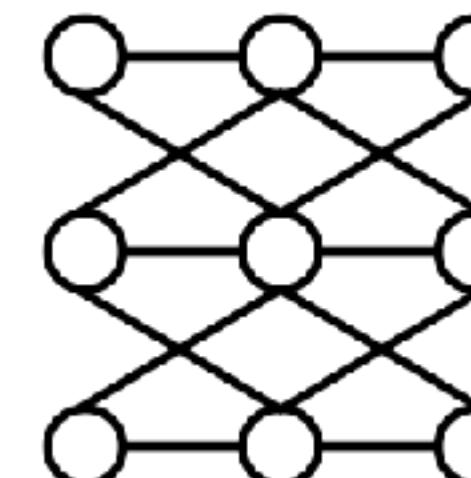
Updated by
2024/01



GPT4o is a new model released by OpenAI in May 2024.



Who developed GPT4o?



collected by
2023/12

Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o."



Update knowledge w/o retraining



Improve verifiability

Widespread Adoptions in Real World

Widespread Adoptions in Real World

October 31, 2024 Product

Introducing ChatGPT search

Get fast, timely answers with links to relevant web sources.

[Try now ↗](#) [Download Chrome extension ↗](#)



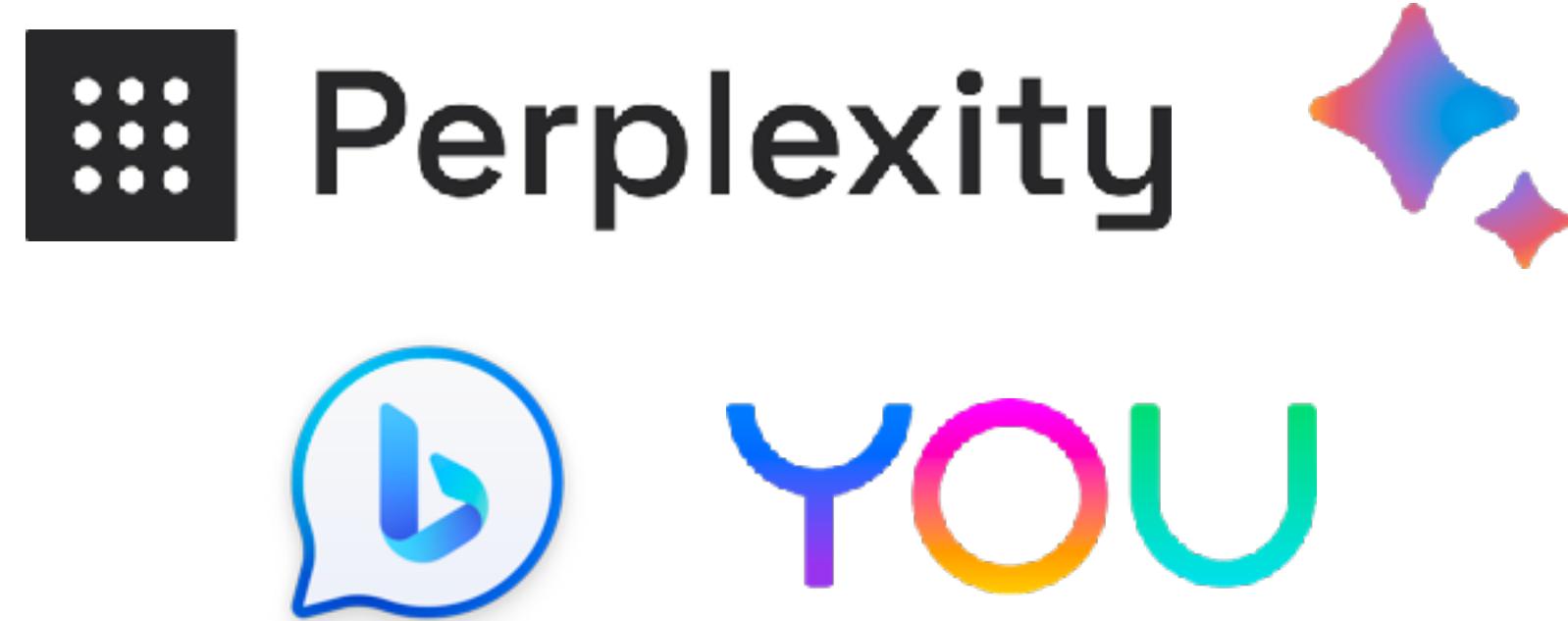
Widespread Adoptions in Real World

October 31, 2024 Product

Introducing ChatGPT search

Get fast, timely answers with links to relevant web sources.

[Try now ↗](#) [Download Chrome extension ↗](#)



Forbes

Council Post: How RAG-Powered AI Applications Have A Positive Impact On Businesses

RAG is a technique in which an LLM is connected to an external, updateable database. It addresses the data gaps in transformers by providing domain-specific...



Widespread Adoptions in Real World

October 31, 2024 Product

Introducing ChatGPT search

Get fast, timely answers with links to relevant web sources.

[Try now ↗](#) [Download Chrome extension ↗](#)

F Forbes

Council Post: How RAG-Powered AI Applications Have A Positive Impact On Businesses



RAG is a technique in which an LLM is connected to an external, updateable database. It addresses the data gaps in transformers by providing domain-specific...

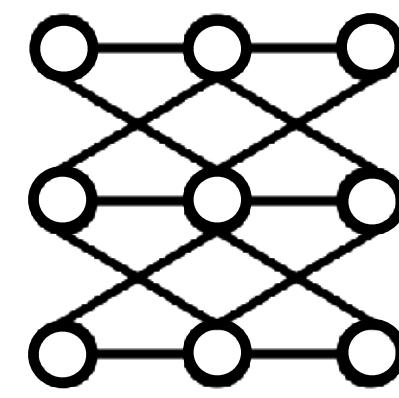


The Shift from Models to Compound AI Systems

symbolic solver to tackle olympiad problems. In enterprises, our colleagues at Databricks found that 60% of LLM applications use some form of [retrieval-augmented generation \(RAG\)](#), and

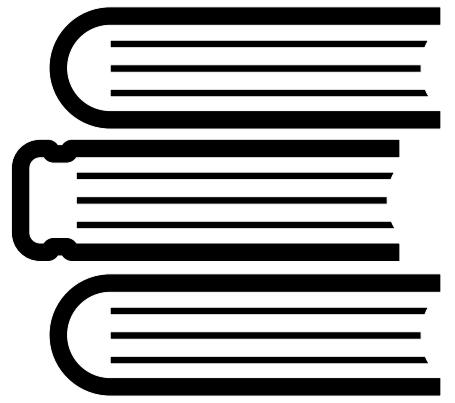
Retrieval-Augmented LMs: Overview

Retrieval-Augmented LMs: Overview



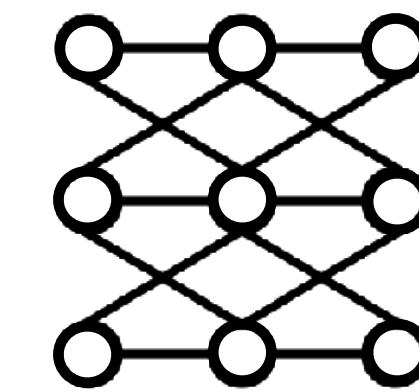
LM

Retrieval-Augmented LMs: Overview



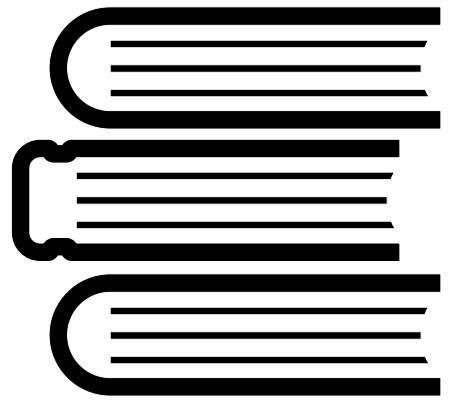
Datastore

Collections of a large number of documents



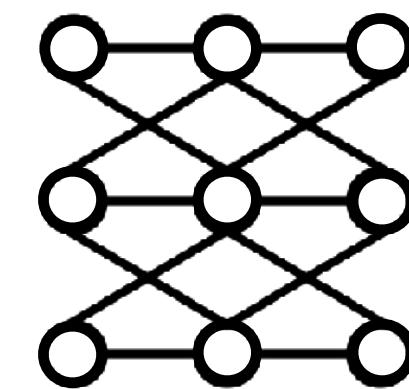
LM

Retrieval-Augmented LMs: Overview



Datastore

Collections of a large number of documents



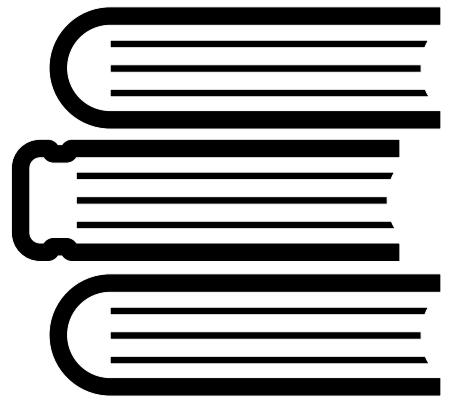
LM

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented LMs: Overview



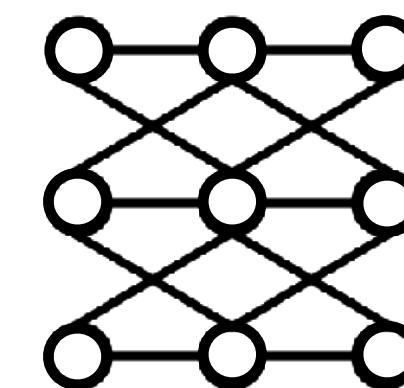
Datastore

Collections of a large number of documents



Retriever

Retrieve top k documents in datastore



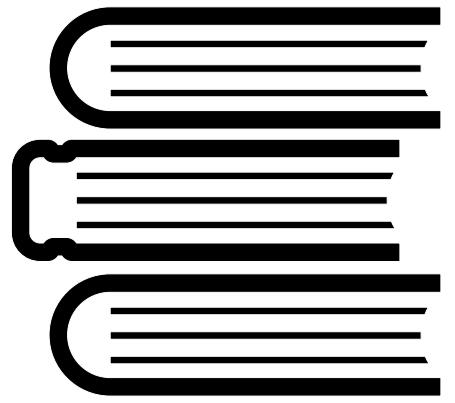
LM

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

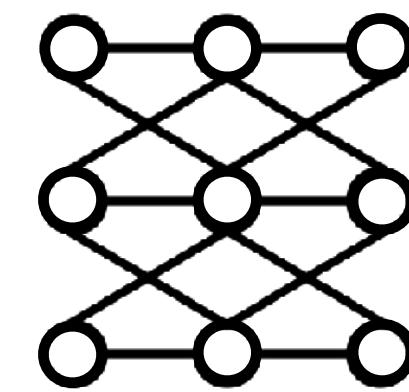
Retrieval-Augmented LMs: Overview



Datastore



Retriever



LM

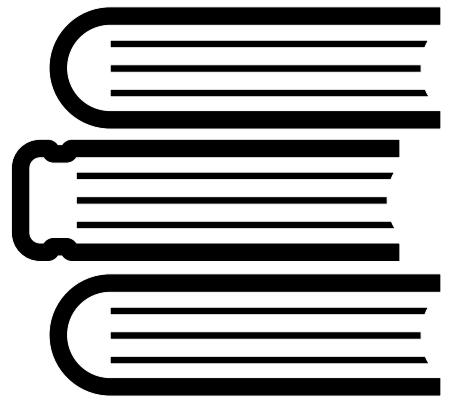
GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented LMs: Overview

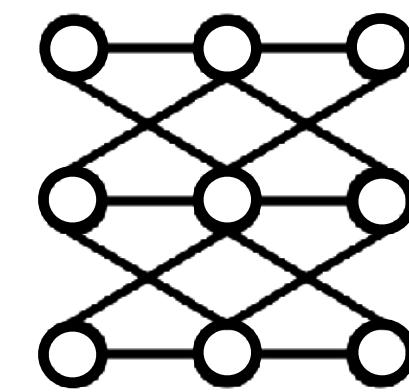
x : Which company developed GPT4o?



Datastore



Retriever



LM

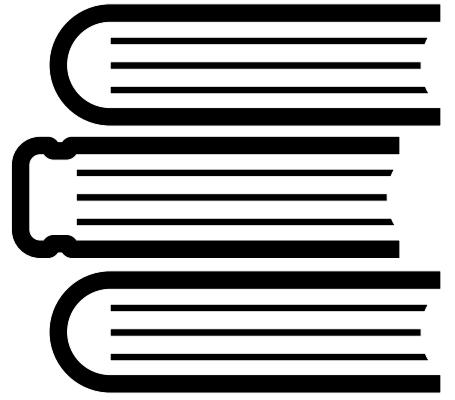
GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

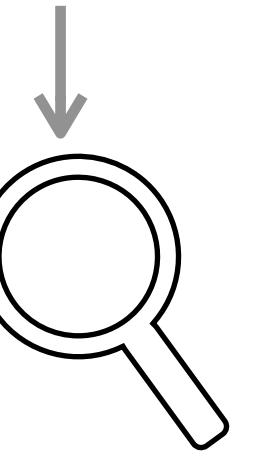
GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented LMs: Overview

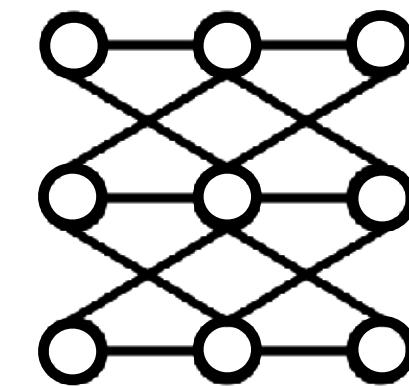
x : Which company developed GPT4o?



Datastore



Retriever



LM

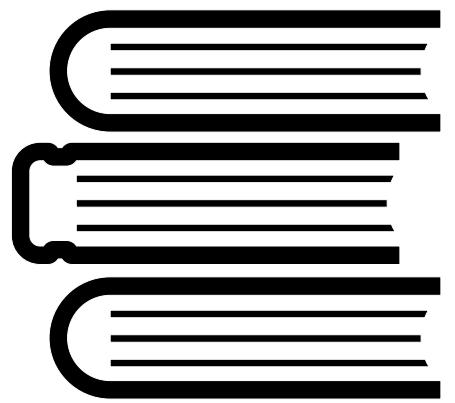
GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

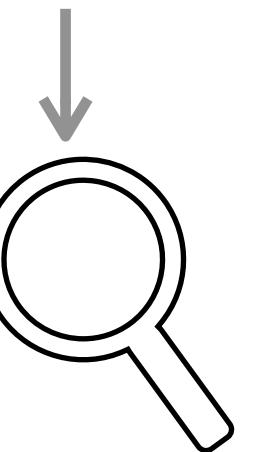
GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented LMs: Overview

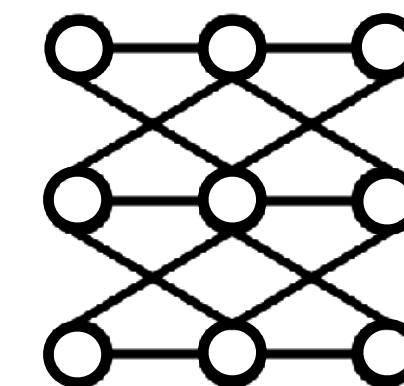
x : Which company developed GPT4o?



Datastore



Retriever



LM

$\text{Sim}(\cdot | x)$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

Transformers is a series of science fiction action films based on the Transformers franchise.

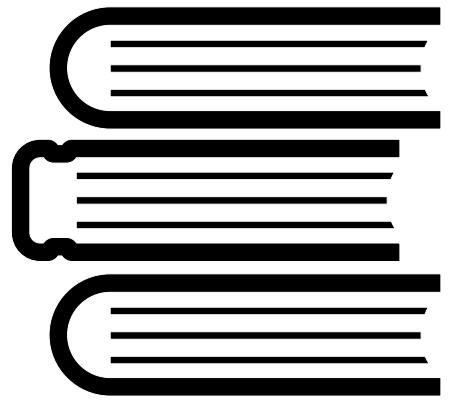
0.1

GPT4o was released by OpenAI in May 2024.

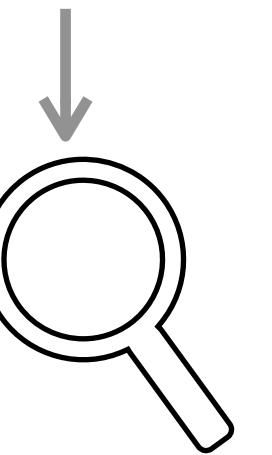
0.8

Retrieval-Augmented LMs: Overview

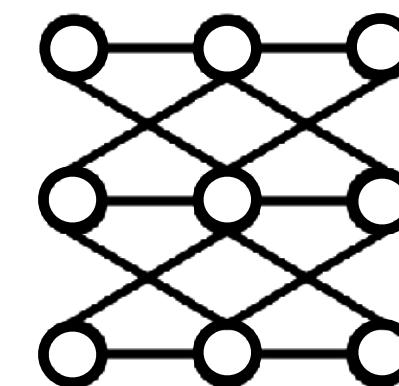
x : Which company developed GPT4o?



Datastore



Retriever



LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

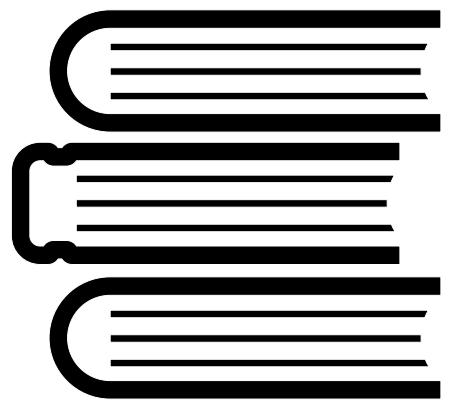
0.9

0.1

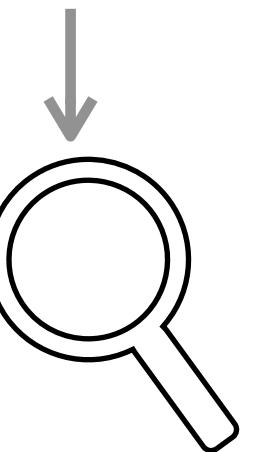
0.8

Retrieval-Augmented LMs: Overview

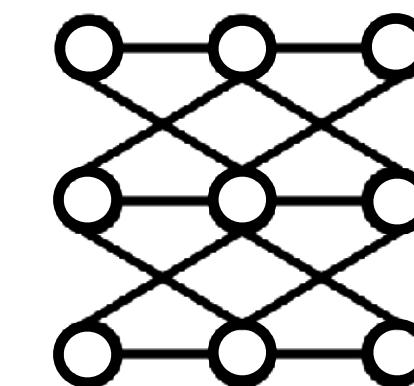
x : Which company developed GPT4o?



Datastore



Retriever



LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

Transformers is a series of science fiction action films based on the Transformers franchise.

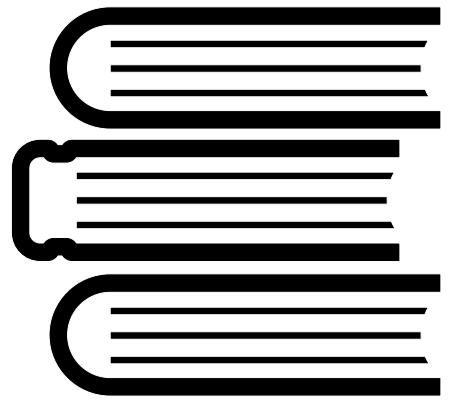
0.1

GPT4o was released by OpenAI in May 2024.

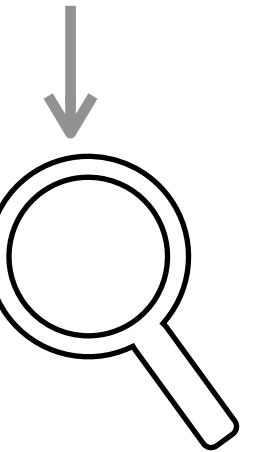
0.8

Retrieval-Augmented LMs: Overview

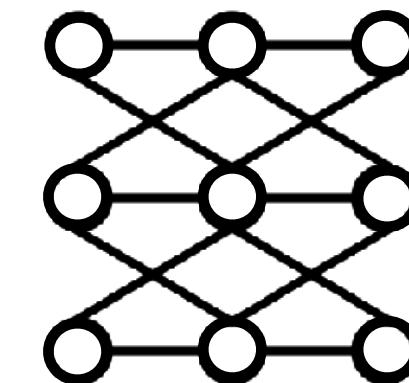
x : Which company developed GPT4o?



Datastore



Retriever



LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

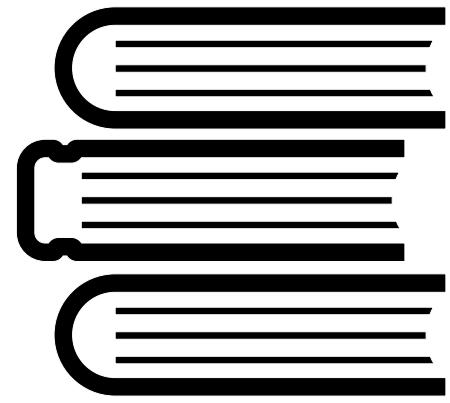
Transformers is a series of science fiction action films based on the Transformers franchise.

0.1

GPT4o was released by OpenAI in May 2024.

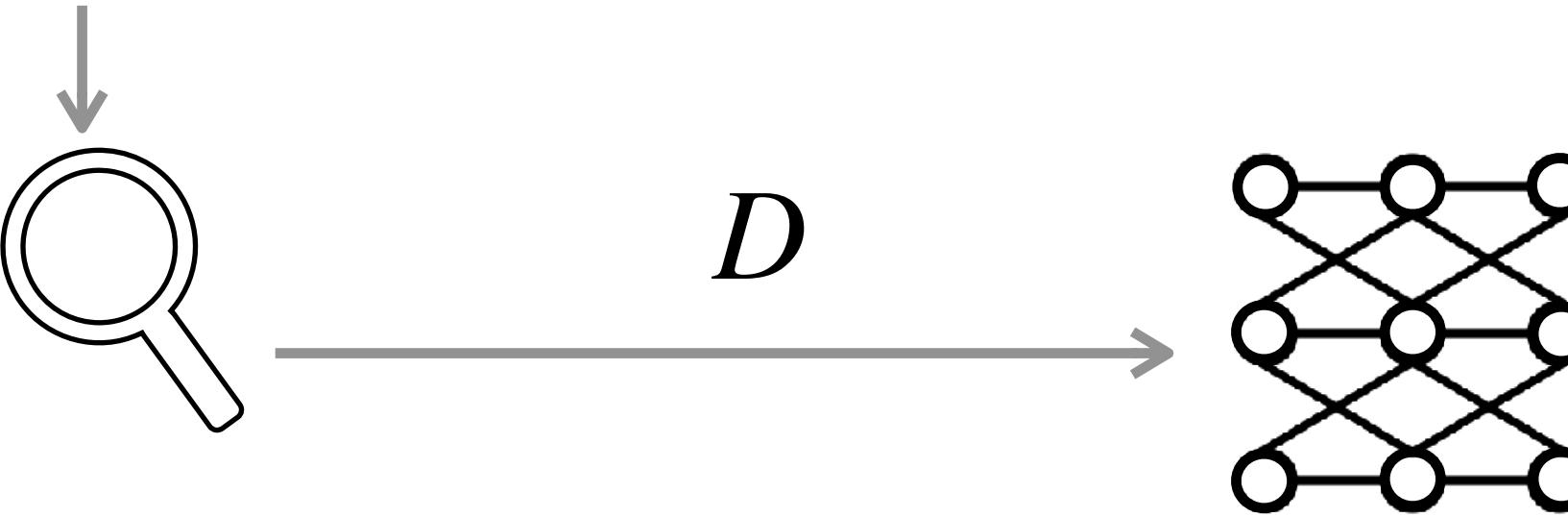
0.8

Retrieval-Augmented LMs: Overview



Datastore

x : Which company developed GPT4o?



Retriever

LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT-4o is a pre-trained transformer developed by OpenAI.

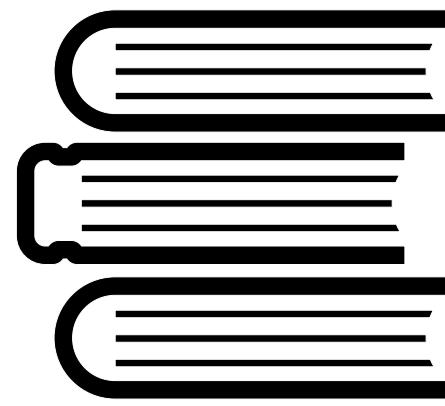
0.9

0.1

GPT4o was released by OpenAI in May 2024.

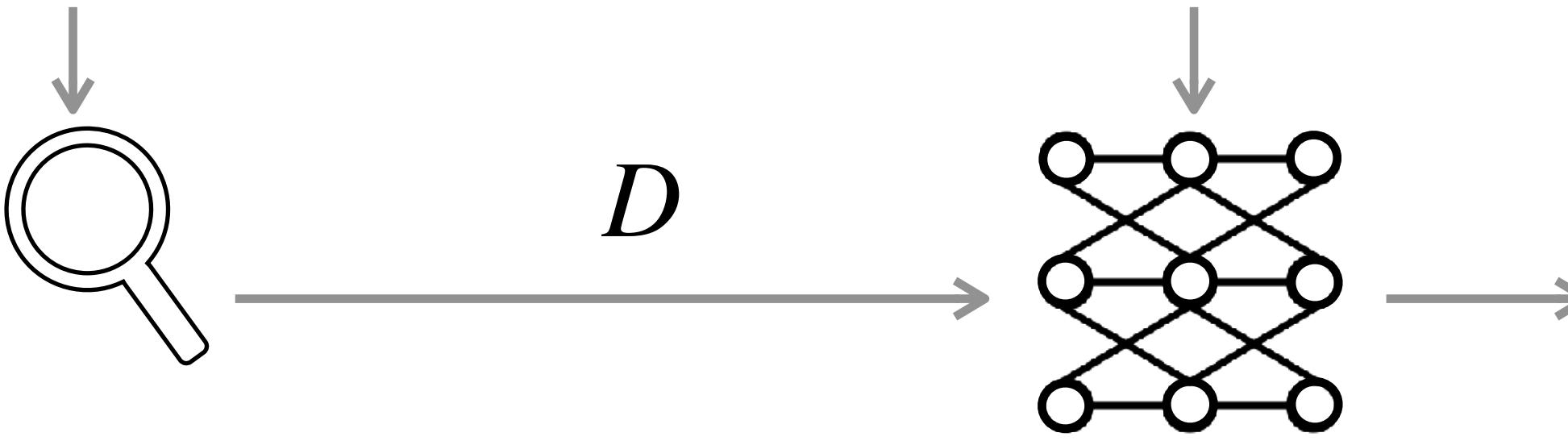
0.8

Retrieval-Augmented LMs: Overview



Datastore

x : Which company developed GPT4o?



Retriever

LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

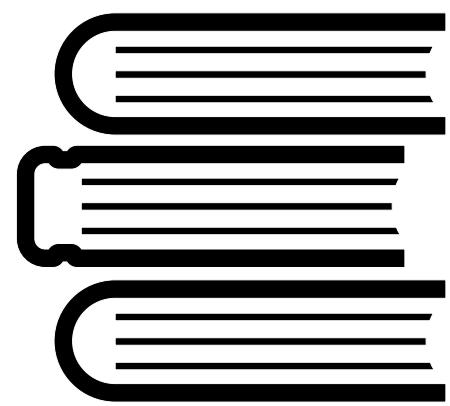
Transformers is a series of science fiction action films based on the Transformers franchise.

0.1

GPT4o was released by OpenAI in May 2024.

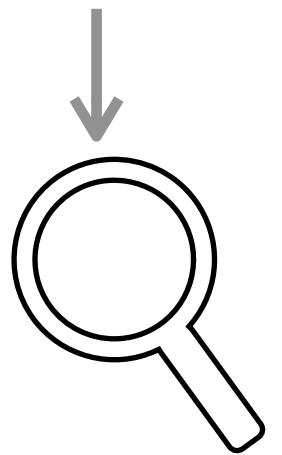
0.8

Retrieval-Augmented LMs: Overview



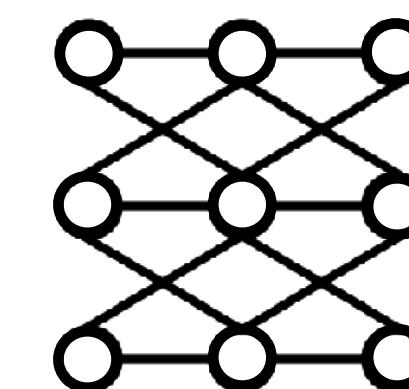
Datastore

x : Which company developed GPT4o?



Retriever

D



LM

y : OpenAI

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

0.1

Transformers is a series of science fiction action films based on the Transformers franchise.

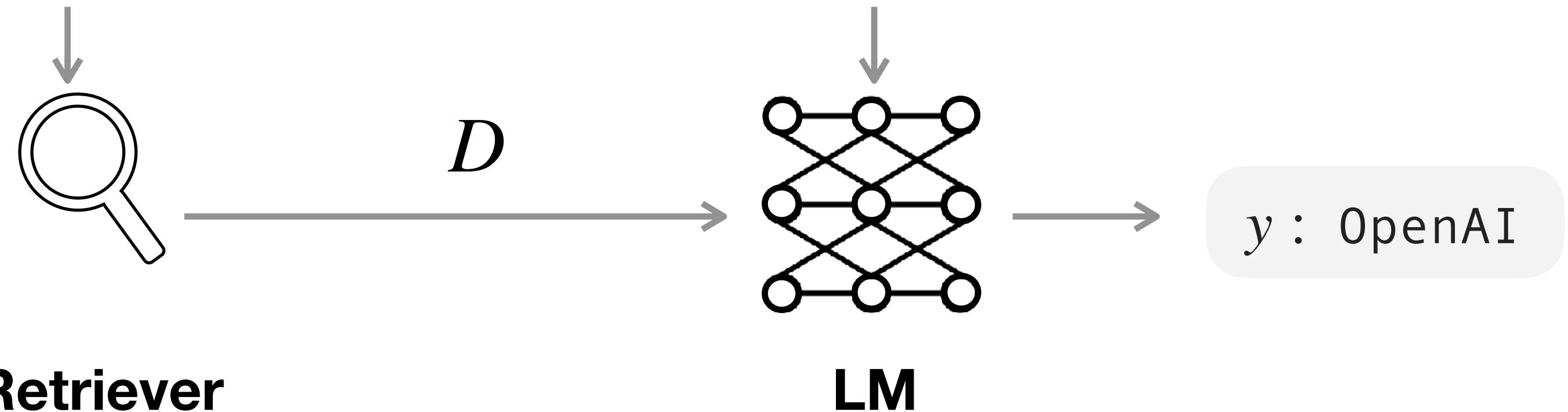
GPT4o was released by OpenAI in May 2024.

0.8

Today's Outline

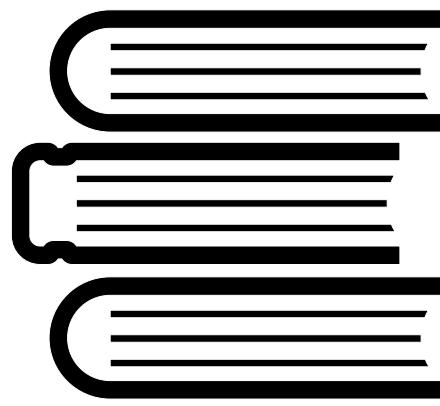


x : Which company developed GPT4o?

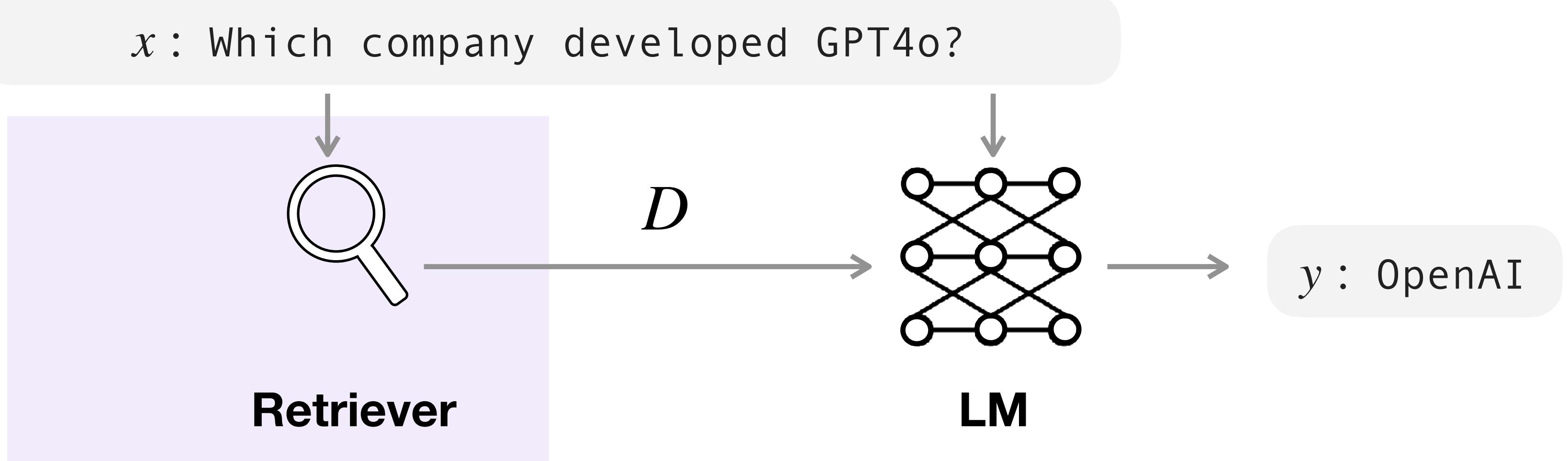


- ✓ Sources of datastore
- ✓ Processing
- ✓ Scaling

Today's Outline

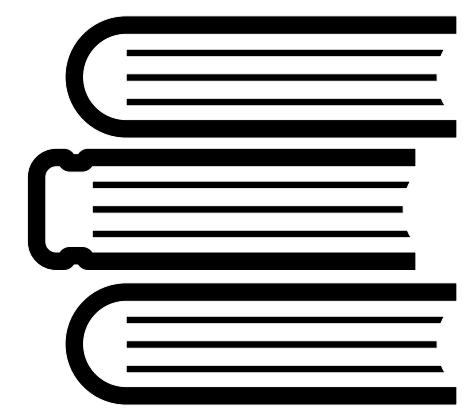


Datastore



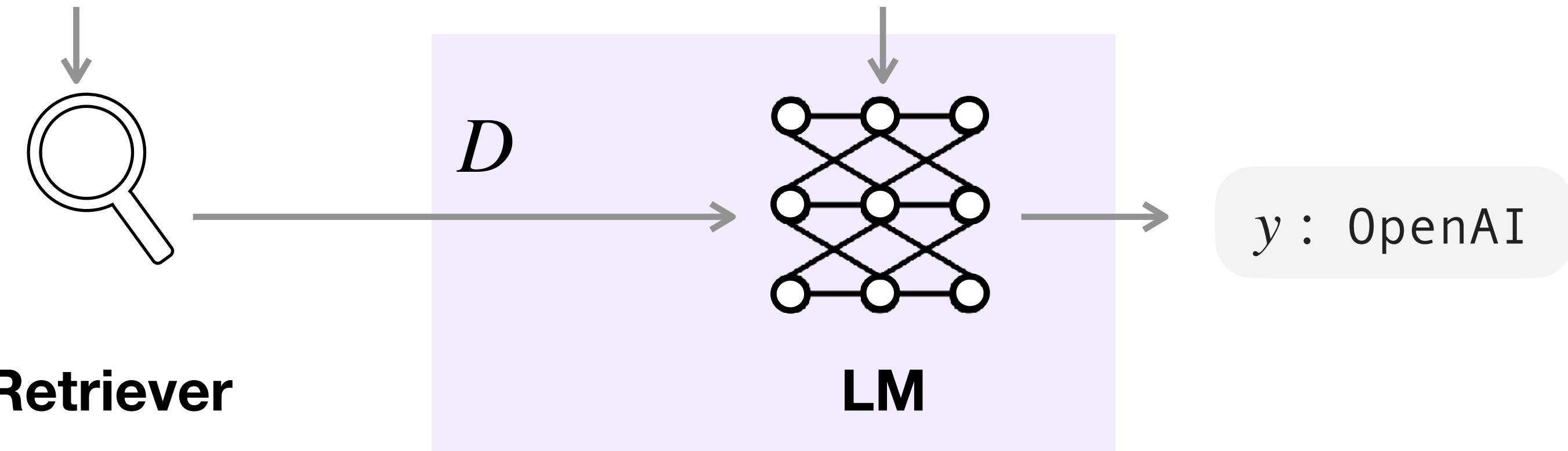
- ✓ Types of retrievers
- ✓ Training
- ✓ Evaluations

Today's Outline



Datastore

x : Which company developed GPT4o?

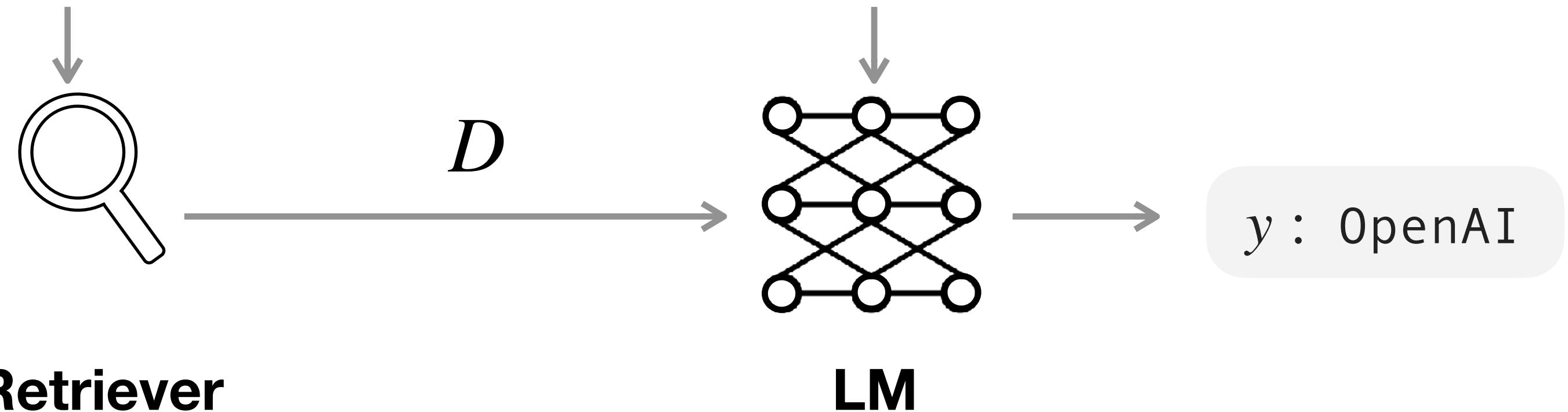


- ✓ Common architectures
- ✓ Recent progress in RAG

Today's Outline



x : Which company developed GPT4o?



- ✓ Sources of datastore
- ✓ Processing
- ✓ Scaling

What Should Be Used as “Datastore”?

What Should Be Used as “Datastore”?

x : Which company developed GPT4o?

What Should Be Used as “Datastore”?

x : Which company developed GPT4o?

English Wikipedia



Chen et al., 2017; Gu et al., 2020; Asai et al., 2020;
Guu et al., 2021; Lewis et al., 2021 ... etc

<https://dumps.wikimedia.org/>

What Should Be Used as “Datastore”?

x : Which company developed GPT4o?

x : How should I implement RAG using LlamaIndex?

English Wikipedia



Chen et al., 2017; Gu et al., 2020; Asai et al., 2020;
Guu et al., 2021; Lewis et al., 2021 ... etc

<https://dumps.wikimedia.org/>

What Should Be Used as “Datastore”?

x : Which company developed GPT4o?

x : How should I implement RAG using LlamaIndex?

English Wikipedia



Chen et al., 2017; Gu et al., 2020; Asai et al., 2020;
Guu et al., 2021; Lewis et al., 2021 ... etc

<https://dumps.wikimedia.org/>

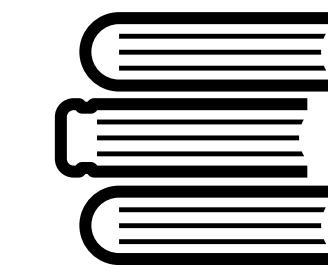
Code snippets



Official documentations



LangChain



Community forums



Massively Scaling Datastore



Massively Scaling Datastore



Massively Scaling Datastore



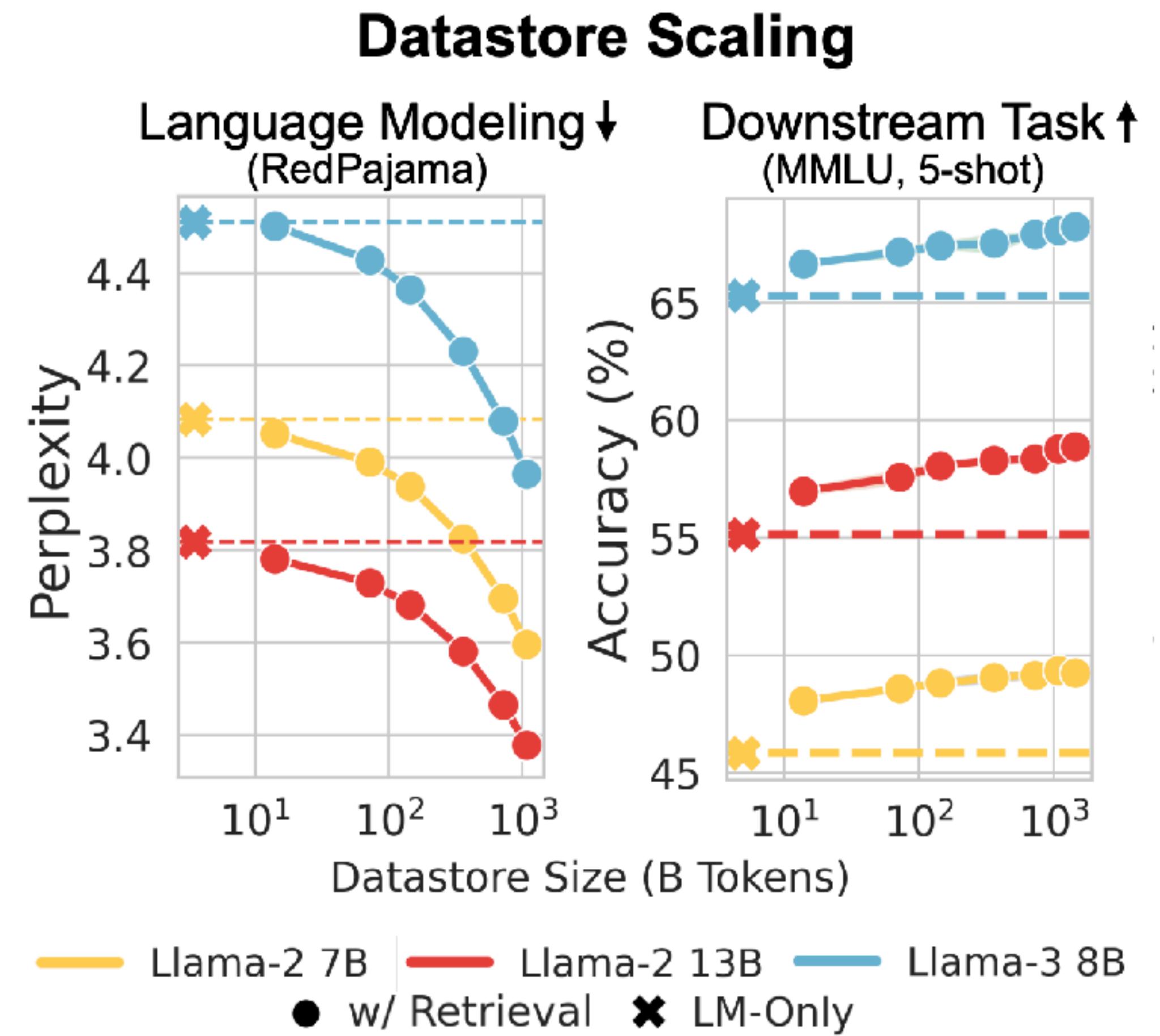
Massively Scaling Datastore

MassiveDS

1.4 trillion tokens (22TB)



Massively Scaling Datastore



Processing documents

≡ GPT-4

From Wikipedia, the free encyclopedia

Generative Pre-trained Transformer 4 (GPT-4) is a [multimodal large language model](#) trained and created by [OpenAI](#) and the fourth in its series of [GPT foundation models](#).^[1] It was launched on March 14, 2023,^[1] and made publicly available via the paid [chatbot product ChatGPT Plus](#), via OpenAI's [API](#), and via the free chatbot [Microsoft Copilot](#).^[2] As a [transformer](#)-based model, GPT-4 uses a paradigm where pre-training using both public data and "data licensed from third-party providers" is used to predict the next [token](#). After this step, the model was then fine-tuned with [reinforcement learning](#) feedback from [humans](#) and AI for [human alignment](#) and policy compliance.^{[3]:2}

Observers reported that the iteration of ChatGPT using GPT-4 was an improvement on the previous iteration based on GPT-3.5, with the caveat that GPT-4 retains some of the problems with earlier revisions.^[4] GPT-4, equipped with vision capabilities (GPT-4V),^[5] is capable of taking images as input on ChatGPT.^[6] OpenAI has not revealed technical details and statistics about GPT-4, such as the precise size of the model.^[7]

Background [edit]

Further information: [GPT-3 § Background](#), and [GPT-2 § Background](#)

OpenAI introduced the first GPT model (GPT-1) in 2018, publishing a paper called "Improving Language Understanding by Generative Pre-

Read Edit View history Tools

32 languages

Generative Pre-trained Transformer 4 (GPT-4)

Developer(s)	OpenAI
Initial release	March 14, 2023; 22 months ago
Predecessor	GPT-3.5
Successor	GPT-4o
Type	Multimodal Large language model Generative pre-trained transformer Foundation model
License	Proprietary
Website	openai.com/gpt-4

Part of a series on
[Machine learning](#)

<https://en.wikipedia.org/wiki/GPT-4>

Processing documents

≡ GPT-4

From Wikipedia, the free encyclopedia

Generative Pre-trained Transformer 4 (GPT-4) is a [multimodal large language model](#) trained and created by [OpenAI](#) and the fourth in its series of [GPT foundation models](#).^[1] It was launched on March 14, 2023,^[1] and made publicly available via the paid [chatbot product ChatGPT Plus](#), via OpenAI's [API](#), and via the free chatbot [Microsoft Copilot](#).^[2] As a [transformer](#)-based model, GPT-4 uses a paradigm where pre-training using both public data and "data licensed from third-party providers" is used to predict the next [token](#). After this step, the model was then fine-tuned with [reinforcement learning](#) feedback from [humans](#) and AI for [human alignment](#) and policy compliance.^{[3]:2}

Observers reported that the iteration of ChatGPT using GPT-4 was an improvement on the previous iteration based on GPT-3.5, with the caveat that GPT-4 retains some of the problems with earlier revisions.^[4] GPT-4, equipped with vision capabilities (GPT-4V),^[5] is capable of taking images as input on ChatGPT.^[6] OpenAI has not revealed technical details and statistics about GPT-4, such as the precise size of the model.^[7]

Background [edit]

Further information: [GPT-3 § Background](#), and [GPT-2 § Background](#)

OpenAI introduced the first GPT model (GPT-1) in 2018, publishing a paper called "Improving Language Understanding by Generative Pre-

Read Edit View history Tools

32 languages

Generative Pre-trained Transformer 4 (GPT-4)

Developer(s)	OpenAI
Initial release	March 14, 2023; 22 months ago
Predecessor	GPT-3.5
Successor	GPT-4o
Type	Multimodal Large language model Generative pre-trained transformer Foundation model
License	Proprietary
Website	openai.com/gpt-4

Part of a series on
[Machine learning](#)

<https://en.wikipedia.org/wiki/GPT-4>

Processing documents

Processing documents

Curate and preprocess data

e.g., HTML -> Plain text



Processing documents

Curate and
preprocess data



e.g., HTML -> Plain text



Processing documents

Curate and preprocess data

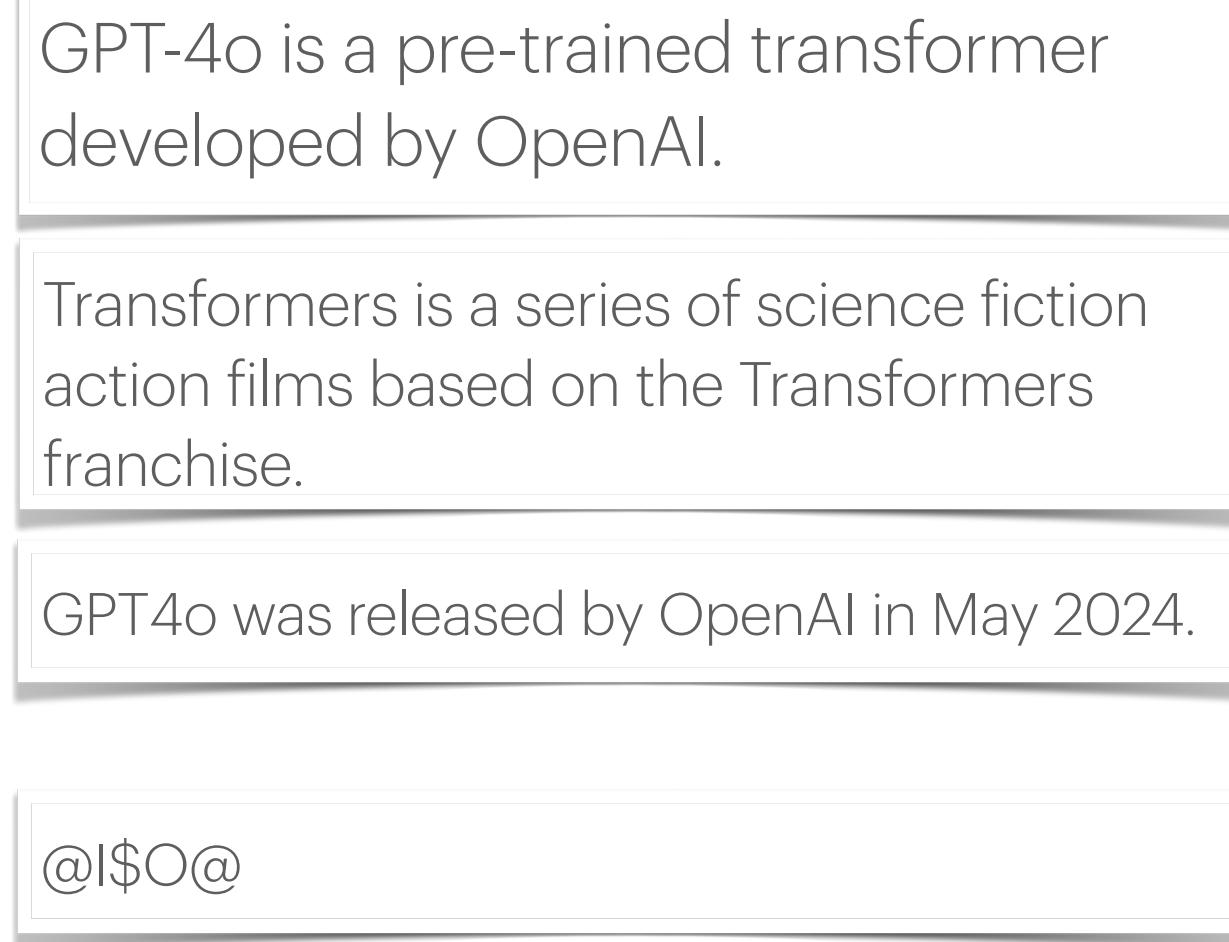
e.g., HTML -> Plain text



Chunking

Paragraph-level (e.g., \n)

Every k words (e.g., 100-250)



Processing documents

Curate and preprocess data

e.g., HTML -> Plain text



Chunking

Paragraph-level (e.g., \n)

Every k words (e.g., 100-250)

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

@I\$O@

Processing documents

Curate and preprocess data

e.g., HTML -> Plain text



Chunking

Paragraph-level (e.g., \n)

Every k words (e.g., 100-250)

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

@I\$O@

Post-processing

e.g., Remove short documents

GPT-4o is a pre-trained transformer developed by OpenAI.

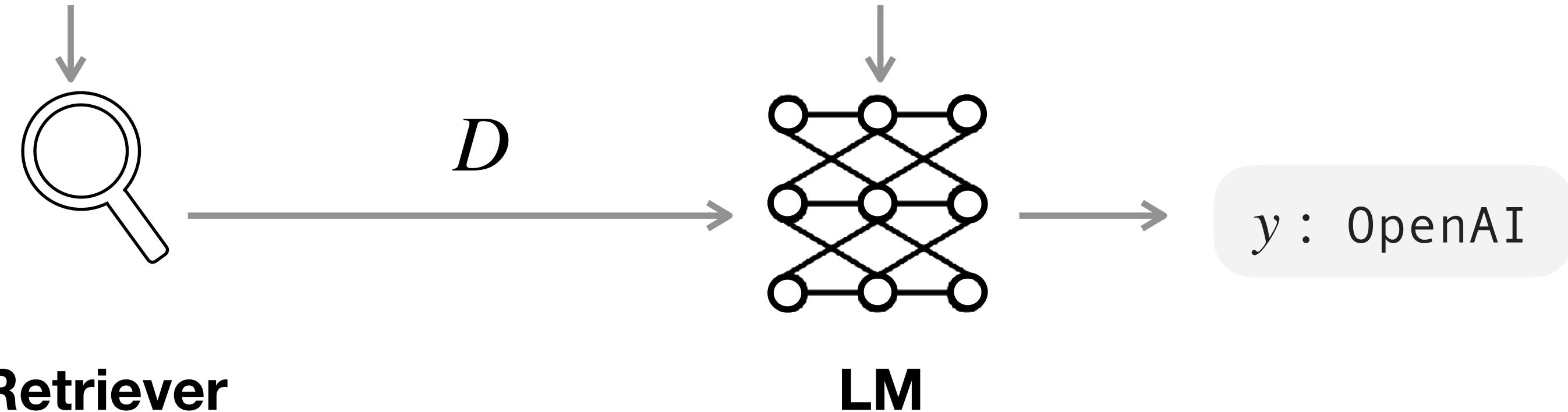
Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

Summary of Part I



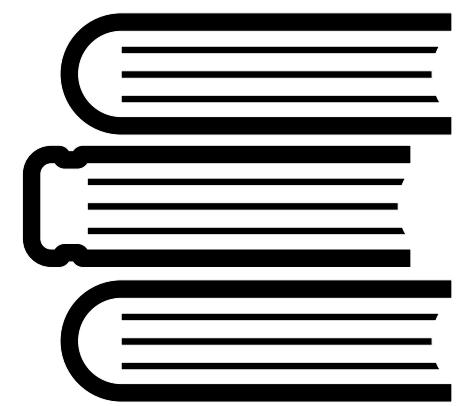
x : Which company developed GPT4o?



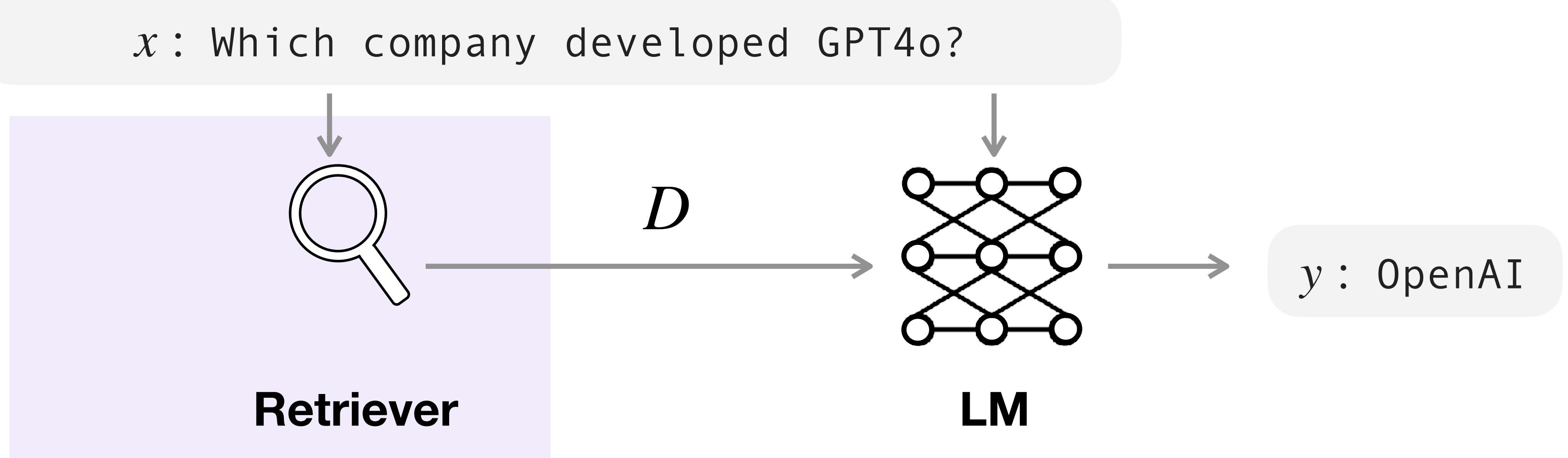
- ✓ Sources of datastore
- ✓ Processing
- ✓ Scaling

- Choosing **the right datastore** is important
- **Chunking** and **filtering** strategies are important
- **Scaling** datastores offer performance gain while adding technical challenges

Today's Outline



Datastore



- ✓ Types of retrievers
- ✓ Training
- ✓ Evaluations

Types of Retrievers

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Types of Retrievers

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Sparse retrievers

- **Sim:** Term-frequency based embeddings
- Training is not required

e.g., TF-IDF, BM25

Types of Retrievers

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Sparse retrievers

- **Sim:** Term-frequency based embeddings
- Training is not required

e.g., TF-IDF, BM25

Dense retrievers

- **Sim:** dense embeddings encoded by pre-trained LMs
- Training is needed*

e.g., DPR, Contriever, ColBERT

Types of Retrievers

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Sparse retrievers

- **Sim:** Term-frequency based embeddings
- Training is not required

e.g., TF-IDF, BM25

Dense retrievers

- **Sim:** dense embeddings encoded by pre-trained LMs
- Training is needed*

e.g., DPR, Contriever, ColBERT

Rerankers

- **Sim:** Scores based on jointly encoded query and doc
- Training is needed*

e.g., cross-encoder reranker

Types of Retrievers

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Sparse retrievers

- **Sim:** Term-frequency based embeddings
- Training is not required

e.g., TF-IDF, BM25



Dense retrievers

- **Sim:** dense embeddings encoded by pre-trained LMs
- Training is needed*

e.g., DPR, Contriever, ColBERT

Rerankers

- **Sim:** Scores based on jointly encoded query and doc
- Training is needed*

e.g., cross-encoder reranker

Types of Retrievers

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Sparse retrievers

- **Sim:** Term-frequency based embeddings
- Training is not required

e.g., TF-IDF, BM25



Dense retrievers

- **Sim:** dense embeddings encoded by pre-trained LMs
- Training is needed*

e.g., DPR, Contriever, ColBERT



Rerankers

- **Sim:** Scores based on jointly encoded query and doc
- Training is needed*

e.g., cross-encoder reranker

Types of Retrievers

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Sparse retrievers

- **Sim:** Term-frequency based embeddings
- Training is not required

e.g., TF-IDF, BM25



Dense retrievers

- **Sim:** dense embeddings encoded by pre-trained LMs
- Training is needed*

e.g., DPR, Contriever, ColBERT



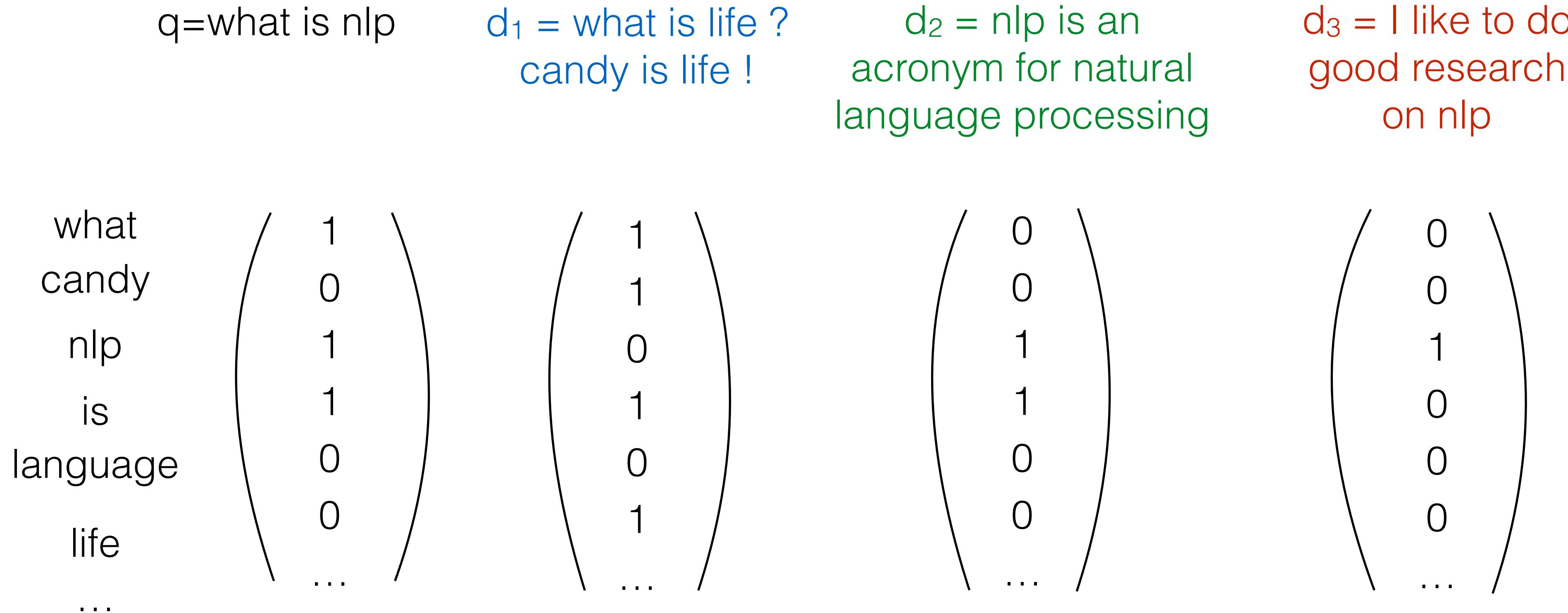
Rerankers

- **Sim:** Scores based on jointly encoded query and doc
- Training is needed*

e.g., cross-encoder reranker



Sparse Retrievers: One-hot Vectors



Sparse Retrievers: One-hot Vectors

q=what is nlp

$d_1 = \text{what is life ?}$
 candy is life !

$d_2 = \text{nlp is an}$
 $\text{acronym for natural}$
 $\text{language processing}$

$d_3 = \text{I like to do}$
 good research
 on nlp

what
candy
nlp
is
language
life
...

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ \dots \end{pmatrix}$$

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

$$\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

Check if a term appears in a document

Sparse Retrievers: Term-count Vectors

q=what is nlp

what
candy
nlp
is
language
life
...

$d_1 = \text{what is life ?}$
 candy is life !

1
0
1
1
0
2
...

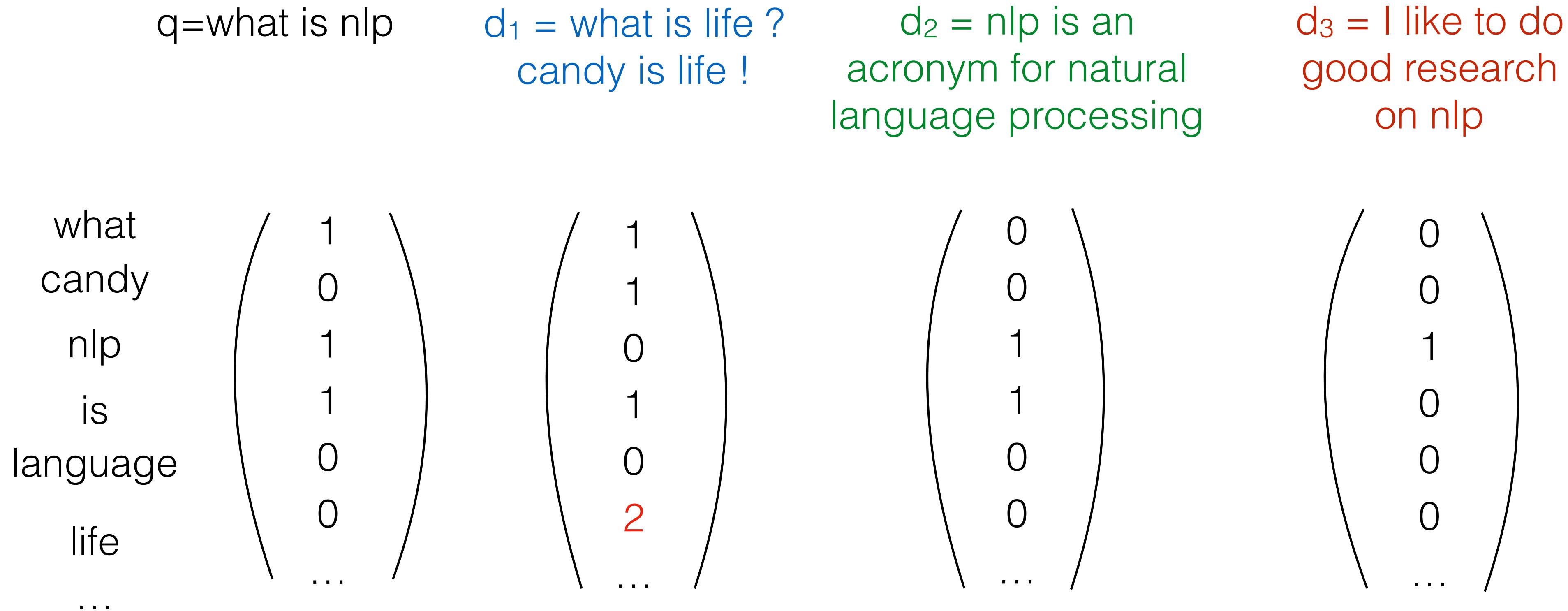
$d_2 = \text{nlp is an}$
 $\text{acronym for natural}$
 $\text{language processing}$

0
0
1
1
0
0
...

$d_3 = \text{I like to do}$
 good research
 on nlp

0
0
1
0
0
0
...

Sparse Retrievers: Term-count Vectors



Count the number of appearances in a doc

Sparse Retrievers: Computing Weighted Term Scores

$$\text{TF}(t, d) = \frac{\text{freq}(t, d)}{\sum_{t'} \text{freq}(t', d)} \quad \text{IDF}(t) = \log \left(\frac{|D|}{\sum_{d' \in D} \delta(\text{freq}(t, d') > 0)} \right)$$

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

$$\text{BM-25}(t, d) = \text{IDF}(t) \cdot \frac{\text{freq}(t, d) \cdot (k_1 + 1)}{\text{freq}(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}$$

Sparse Retrievers: Computing Weighted Term Scores

$d_1 = \text{what is life ?}$

candy is life !

$$\text{TF}(t, d) = \frac{\text{freq}(t, d)}{\sum_{t'} \text{freq}(t', d)}$$
$$\text{IDF}(t) = \log \left(\frac{|D|}{\sum_{d' \in D} \delta(\text{freq}(t, d') > 0)} \right)$$

↑
 $t_1 = \text{what}$

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

$$\text{BM-25}(t, d) = \text{IDF}(t) \cdot \frac{\text{freq}(t, d) \cdot (k_1 + 1)}{\text{freq}(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}$$

Sparse Retrievers: Computing Weighted Term Scores

$d_1 = \text{what is life ?}$

candy is life !

$$\text{TF}(t, d) = \frac{\text{freq}(t, d)}{\sum_{t'} \text{freq}(t', d)}$$

\downarrow
 \uparrow

$t_1 = \text{what}$

of documents

$$\text{IDF}(t) = \log \left(\frac{|D|}{\sum_{d' \in D} \delta(\text{freq}(t, d') > 0)} \right)$$

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

$$\text{BM-25}(t, d) = \text{IDF}(t) \cdot \frac{\text{freq}(t, d) \cdot (k_1 + 1)}{\text{freq}(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}} \right)}$$

Sparse Retrievers: Computing Weighted Term Scores

$d_1 = \text{what is life ?}$

candy is life !

$$\text{TF}(t, d) = \frac{\text{freq}(t, d)}{\sum_{t'} \text{freq}(t', d)}$$

$t_1 = \text{what}$

of documents

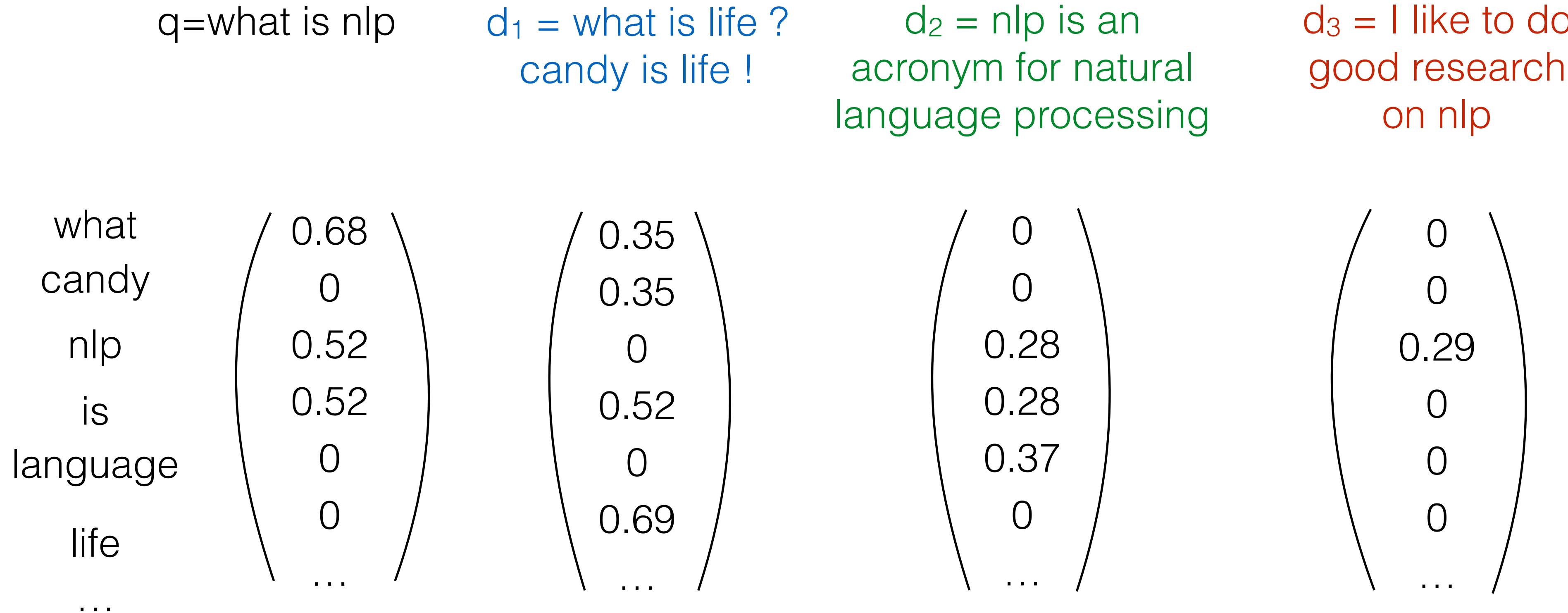
$$\text{IDF}(t) = \log \left(\frac{|D|}{\sum_{d' \in D} \delta(\text{freq}(t, d') > 0)} \right)$$

of documents where term t appears

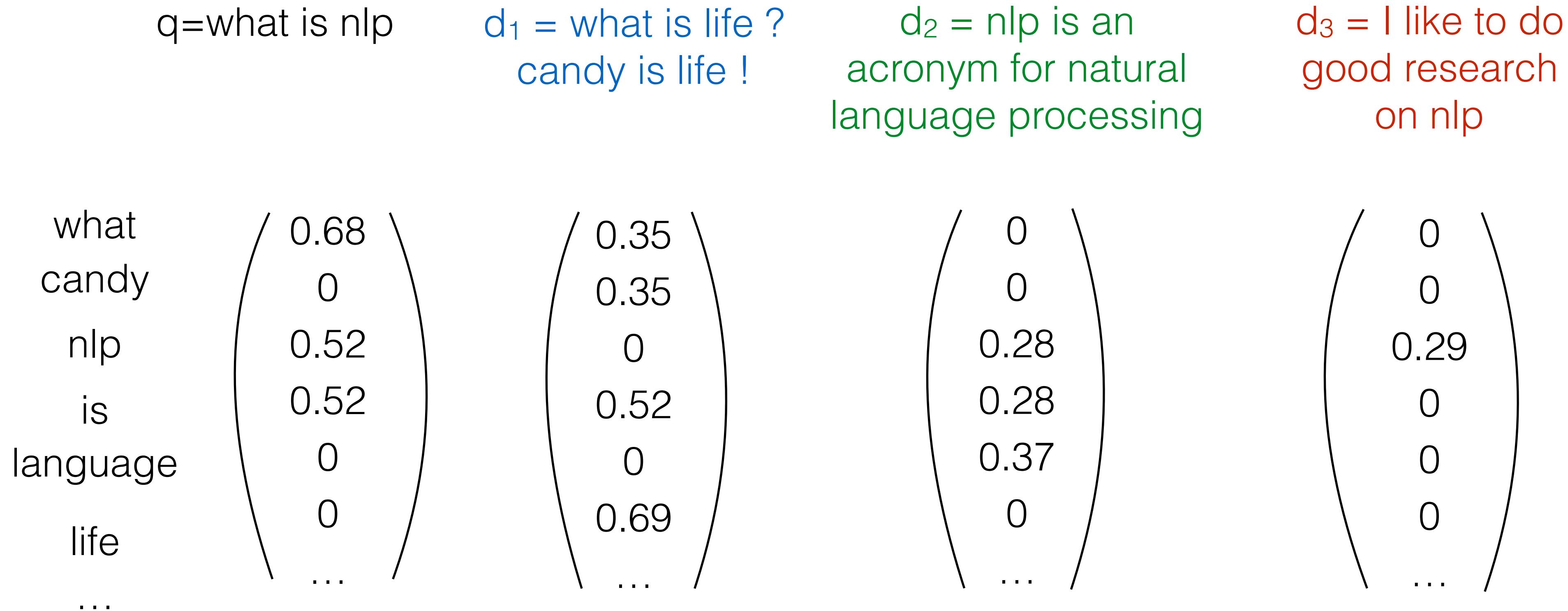
$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

$$\text{BM-25}(t, d) = \text{IDF}(t) \cdot \frac{\text{freq}(t, d) \cdot (k_1 + 1)}{\text{freq}(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}$$

Sparse Retrievers: Weighted-term Vectors

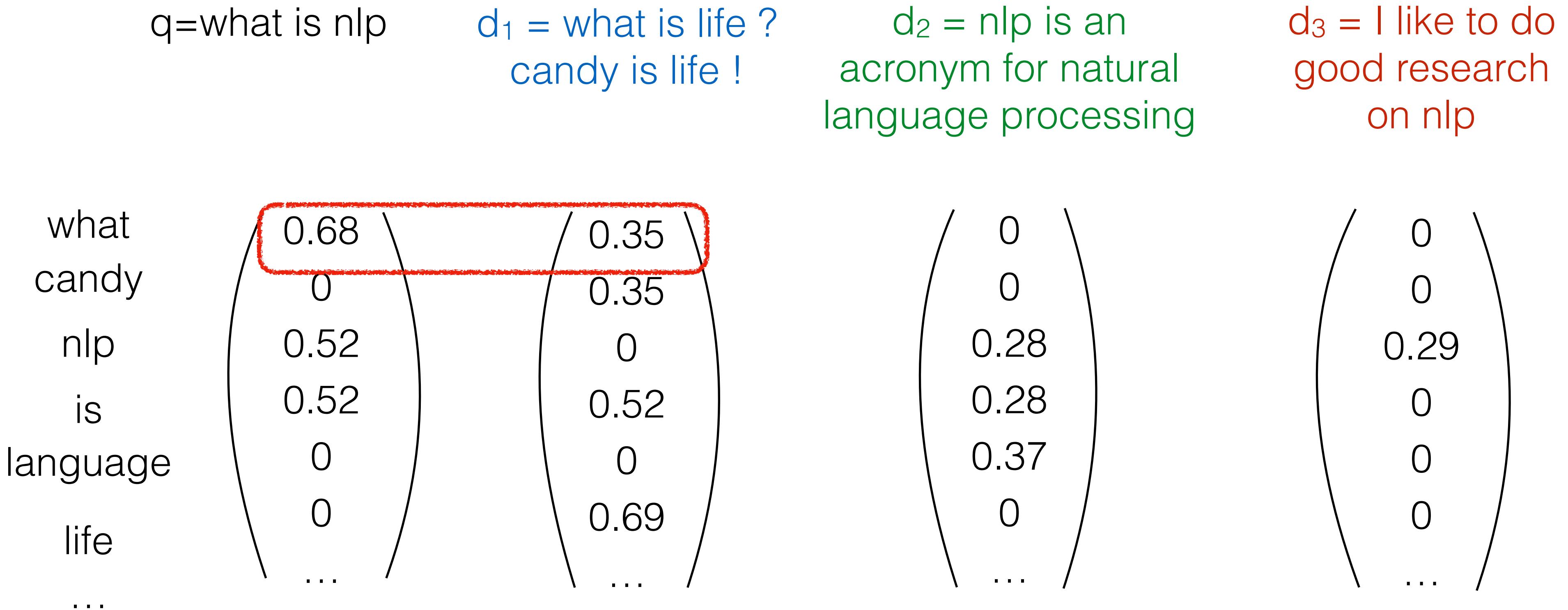


Sparse Retrievers: Weighted-term Vectors



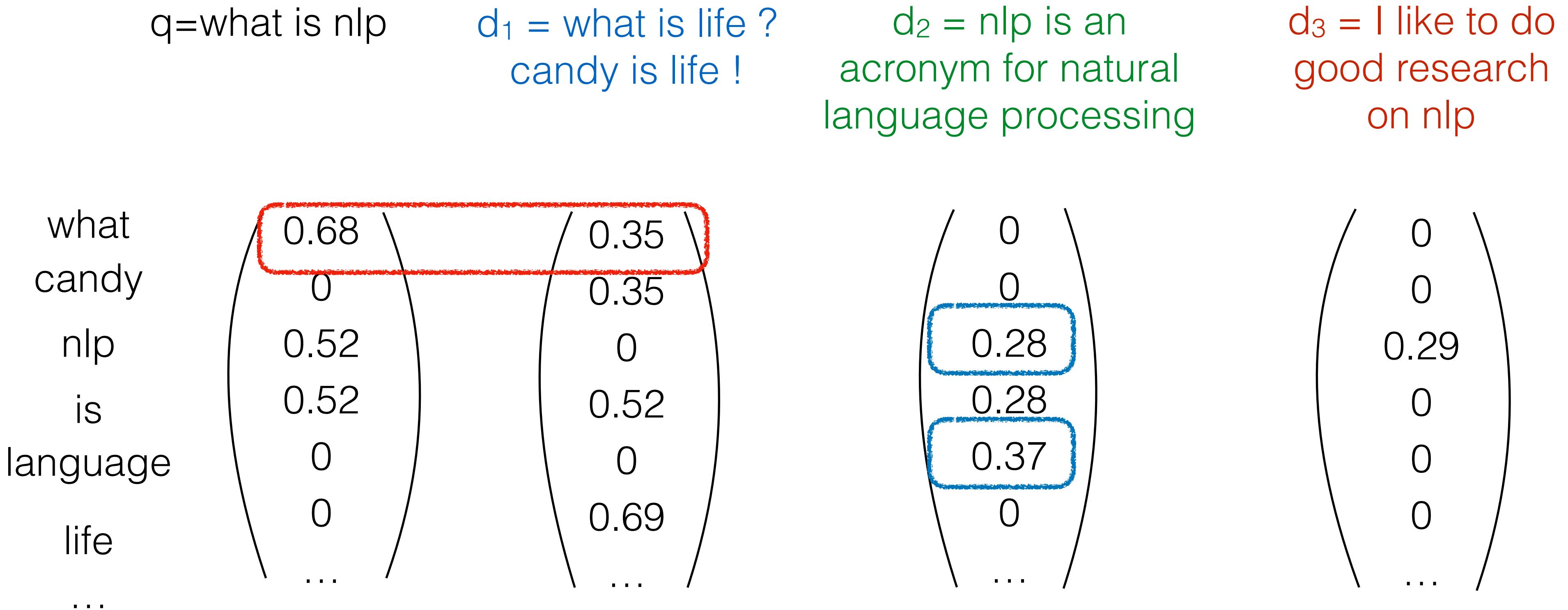
Compute TF-IDF weights to build weighted vectors

Sparse Retrievers: Weighted-term Vectors



Compute TF-IDF weights to build weighted vectors

Sparse Retrievers: Weighted-term Vectors



Compute TF-IDF weights to build weighted vectors

Computing TF-IDF Matrices: Weighted-term Vectors

q=what is nlp

what
candy
nlp
is
language
life
...

$$\begin{pmatrix} 0.68 \\ 0 \\ 0.52 \\ 0.52 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

$d_1 = \text{what is life ?}$
 candy is life !

$$\begin{pmatrix} 0.35 \\ 0.35 \\ 0 \\ 0.52 \\ 0 \\ 0.69 \\ \dots \end{pmatrix}$$

$d_2 = \text{nlp is an acronym for natural language processing}$

$$\begin{pmatrix} 0 \\ 0 \\ 0.28 \\ 0.28 \\ 0.37 \\ 0 \\ \dots \end{pmatrix}$$

$d_3 = \text{I like to do good research on nlp}$

$$\begin{pmatrix} 0 \\ 0 \\ 0.29 \\ 0 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

Computing TF-IDF Matrices: Weighted-term Vectors

q=what is nlp

what
candy
nlp
is
language
life
...

$d_1 = \text{what is life ?}$
 candy is life !

0.68
0
0.52
0.52
0
0
...

$d_2 = \text{nlp is an acronym for natural language processing}$

0
0
0.28
0.28
0.37
0
...

$d_3 = \text{I like to do good research on nlp}$

0
0
0.29
0
0
0
...

Compute cosine similarity

Computing TF-IDF Matrices: Weighted-term Vectors

q=what is nlp

what
candy
nlp
is
language
life
...

$d_1 = \text{what is life ?}$
 candy is life !

0.68
0
0.52
0.52
0
0
...

$d_2 = \text{nlp is an acronym for natural language processing}$

0
0
0.28
0.28
0.37
0
...

$d_3 = \text{I like to do good research on nlp}$

0
0
0.29
0
0
0
...

Compute cosine similarity

$q^*d_1 = 0.50$

Computing TF-IDF Matrices: Weighted-term Vectors

q=what is nlp

what
candy
nlp
is
language
life
...

$d_1 = \text{what is life ?}$
 candy is life !

0.68
0
0.52
0.52
0
0
...

$d_2 = \text{nlp is an acronym for natural language processing}$

0
0
0.28
0.28
0.37
0
...

$d_3 = \text{I like to do good research on nlp}$

0
0
0.29
0
0
0
...

Compute cosine similarity

$q^*d_1 = 0.50$

$q^*d_2 = 0.29$

Computing TF-IDF Matrices: Weighted-term Vectors

q=what is nlp

what
candy
nlp
is
language
life
...

$$\begin{pmatrix} 0.68 \\ 0 \\ 0.52 \\ 0.52 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

$d_1 = \text{what is life ?}$
 candy is life !

$$\begin{pmatrix} 0.35 \\ 0.35 \\ 0 \\ 0.52 \\ 0 \\ 0.69 \\ \dots \end{pmatrix}$$

$d_2 = \text{nlp is an acronym for natural language processing}$

$$\begin{pmatrix} 0 \\ 0 \\ 0.28 \\ 0.28 \\ 0.37 \\ 0 \\ \dots \end{pmatrix}$$

$d_3 = \text{I like to do good research on nlp}$

$$\begin{pmatrix} 0 \\ 0 \\ 0.29 \\ 0 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

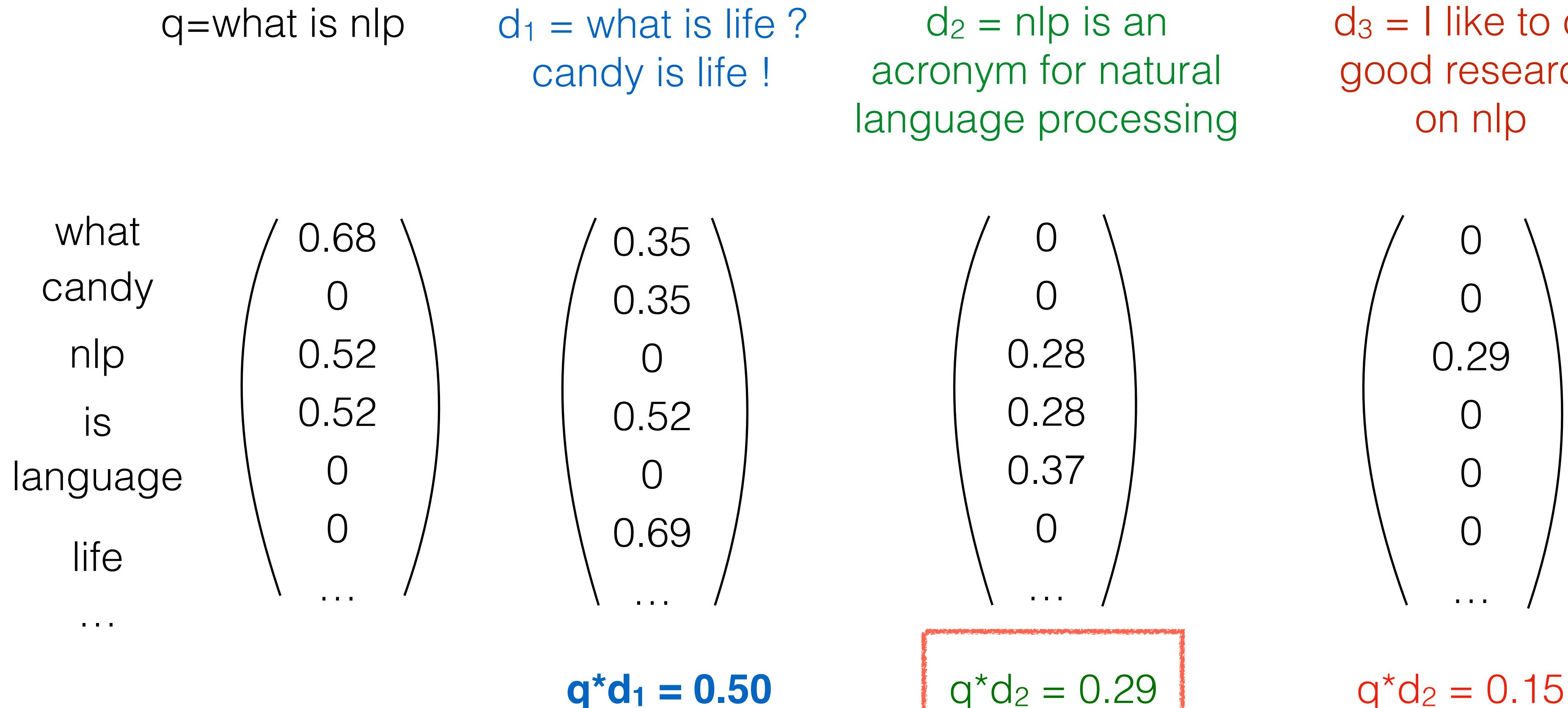
Compute cosine similarity

$$q^*d_1 = 0.50$$

$$q^*d_2 = 0.29$$

$$q^*d_3 = 0.15$$

Computing TF-IDF Matrices: Weighted-term Vectors



Computing TF-IDF Matrices: Weighted-term Vectors

$q = \text{what is nlp}$

what
candy
nlp
is
language
life
...

$$\begin{pmatrix} 0.68 \\ 0 \\ 0.52 \\ 0.52 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

“Bag-of-words”

$d_1 = \text{what is life ?}$
 candy is life !

$$\begin{pmatrix} 0.35 \\ 0.35 \\ 0 \\ 0.52 \\ 0 \\ 0.69 \\ \dots \end{pmatrix}$$

$q^*d_1 = 0.50$

$d_2 = \text{nlp is an acronym for natural language processing}$

$$\begin{pmatrix} 0 \\ 0 \\ 0.28 \\ 0.28 \\ 0.37 \\ 0 \\ \dots \end{pmatrix}$$

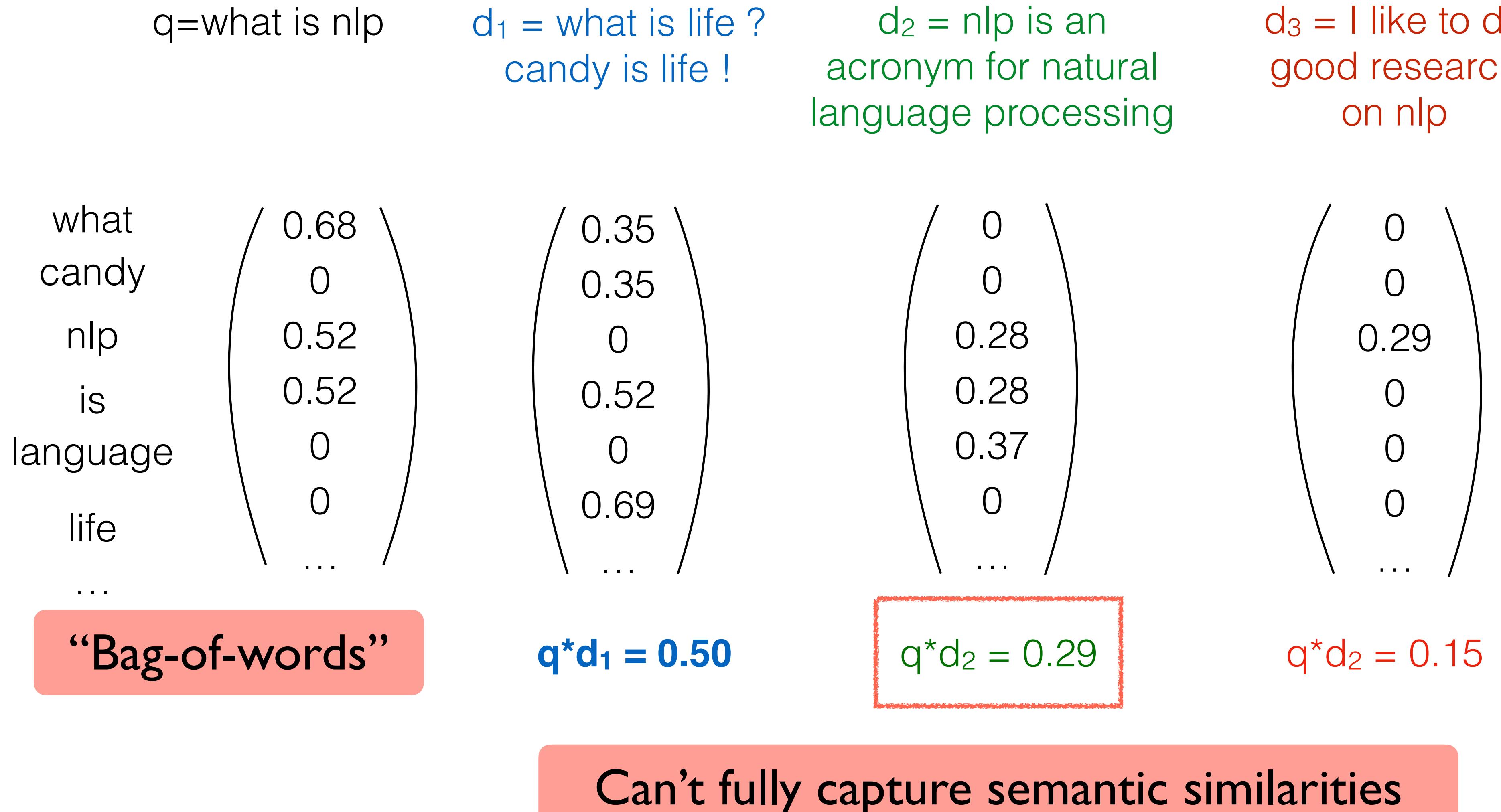
$q^*d_2 = 0.29$

$d_3 = \text{I like to do good research on nlp}$

$$\begin{pmatrix} 0 \\ 0 \\ 0.29 \\ 0 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

$q^*d_3 = 0.15$

Computing TF-IDF Matrices: Weighted-term Vectors



Dense Retrievers: Overview

FIFA World Cup 2026
will expand to 48 teams.

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

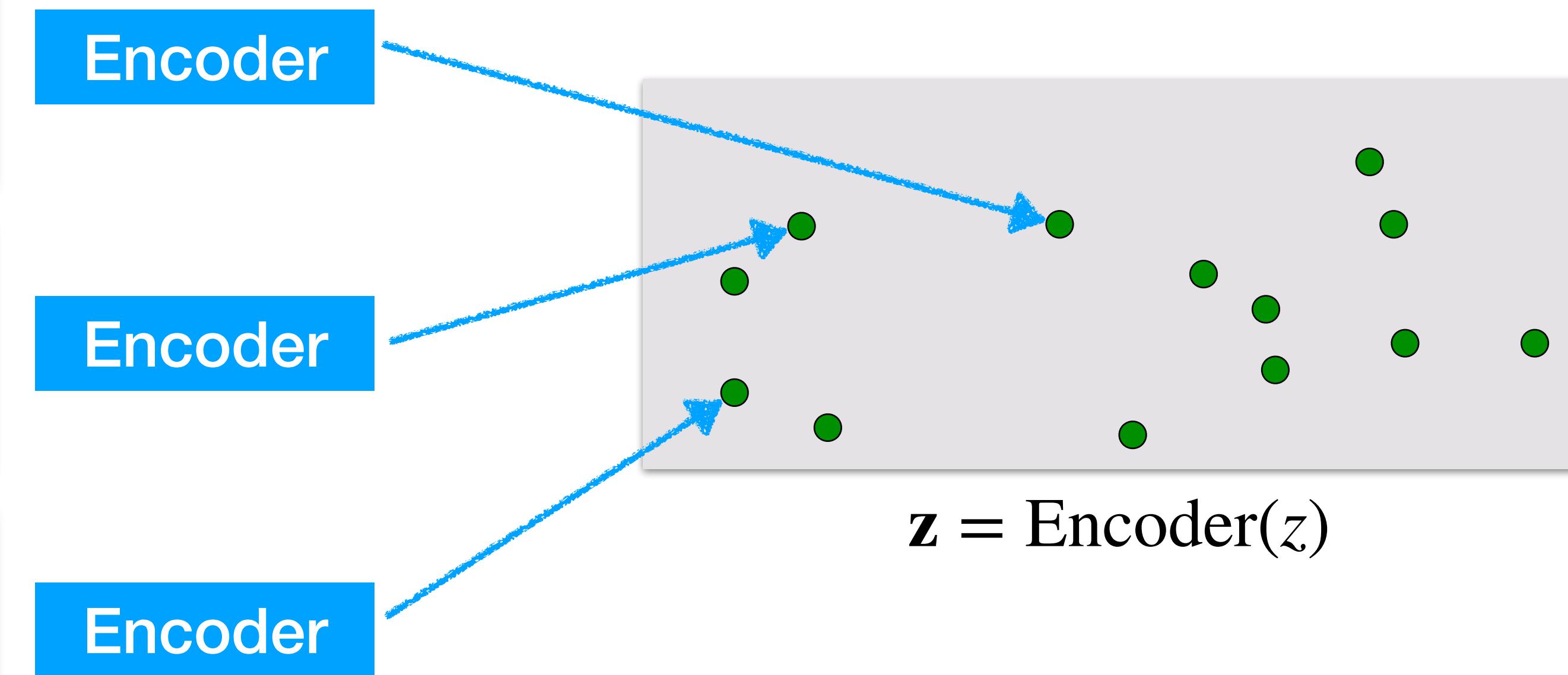
Team USA celebrated
after winning its match
against Iran ...

Dense Retrievers: Overview

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 ...



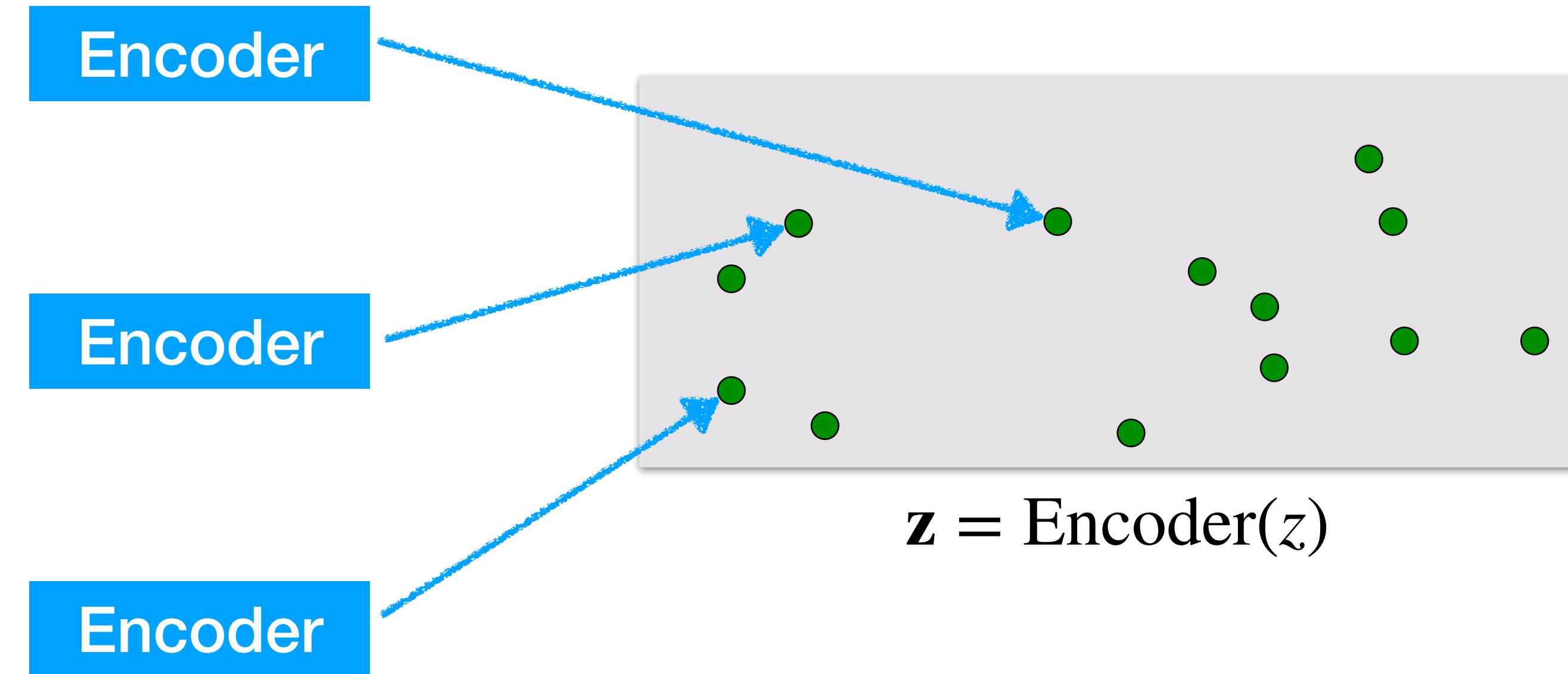
Dense Retrievers: Overview

$x =$ How many teams will participate in FIFA World Cup 2026?

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 ...



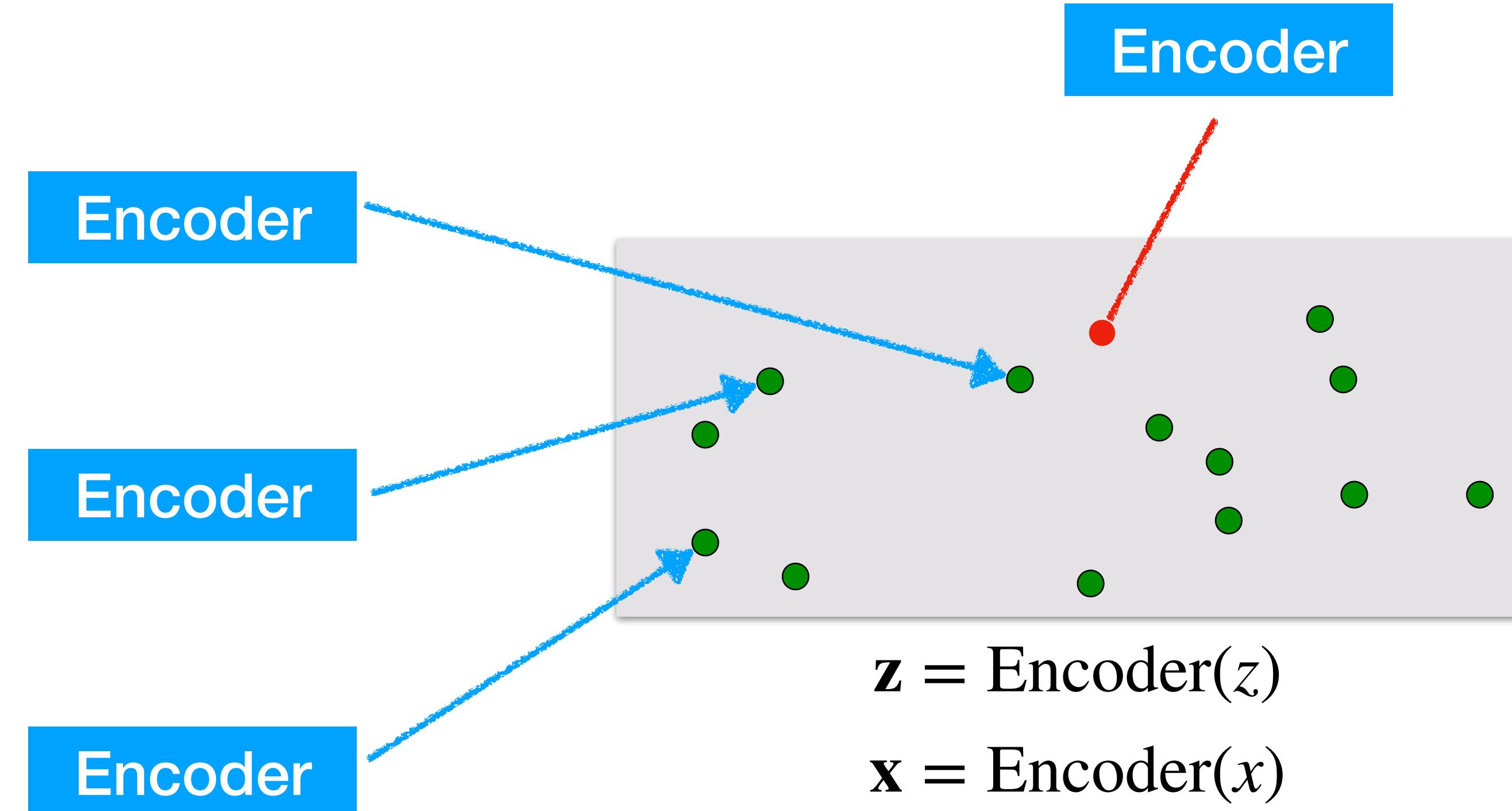
Dense Retrievers: Overview

x = How many teams will participate in FIFA World Cup 2026?

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 ...



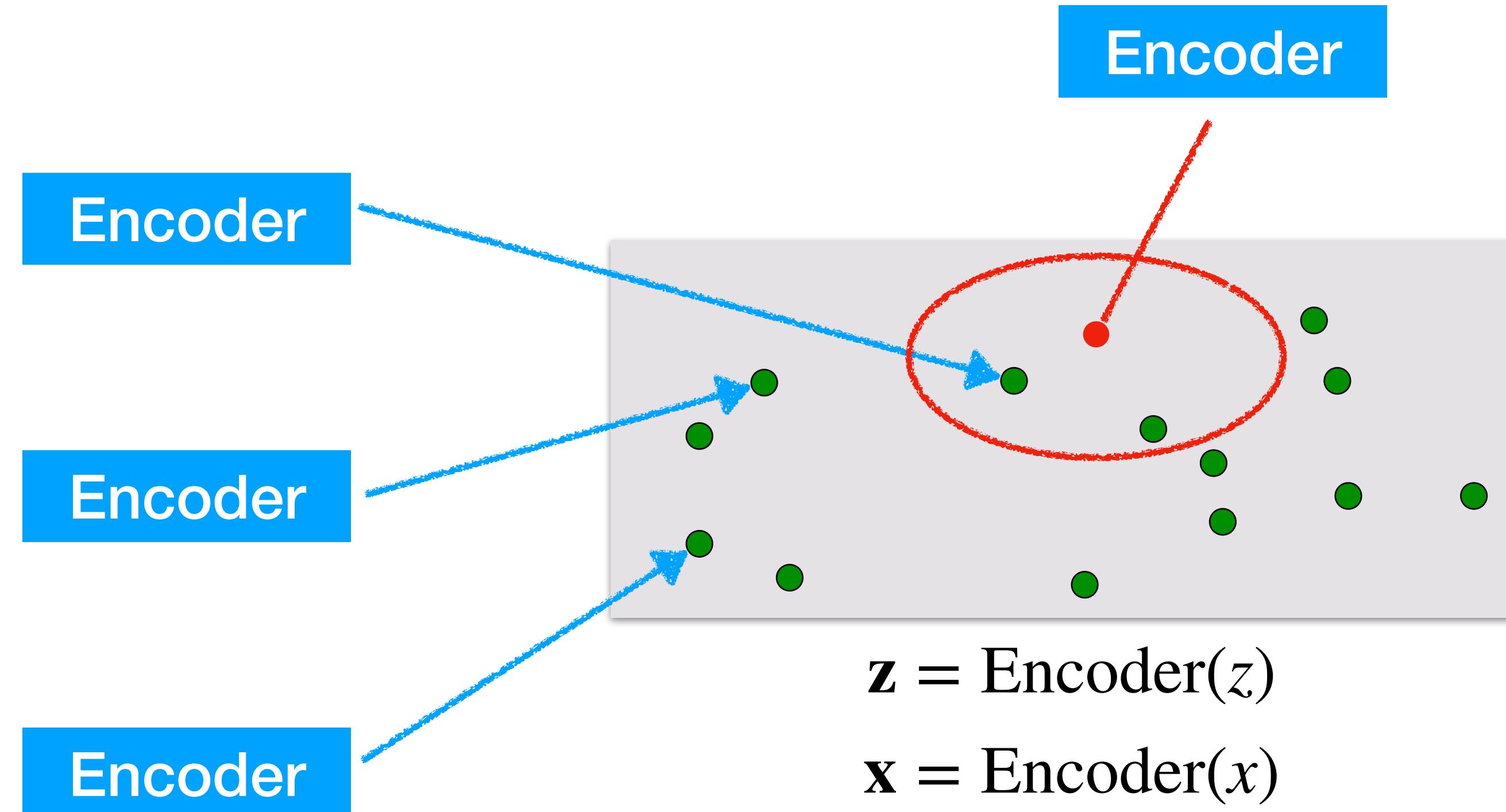
Dense Retrievers: Overview

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 ...

$x =$ How many teams will participate in FIFA World Cup 2026?



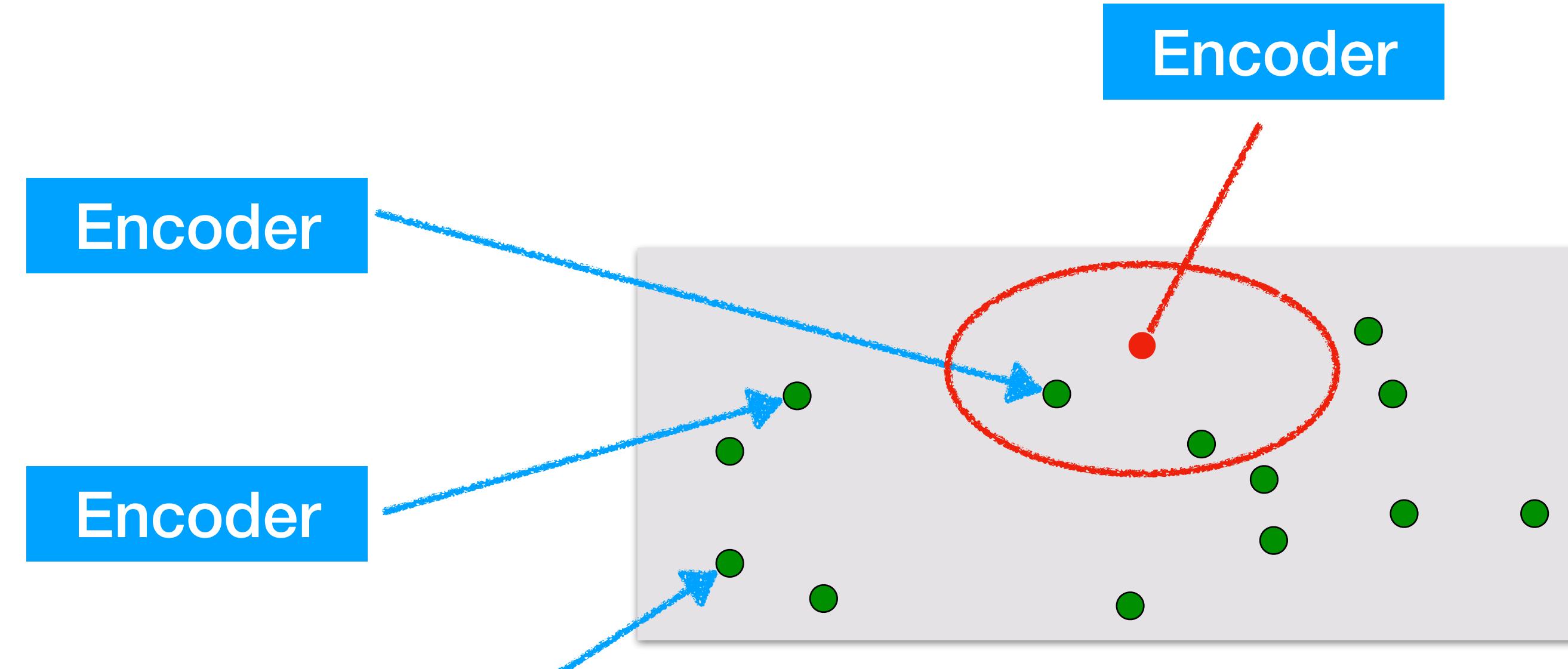
Dense Retrievers: Overview

$x =$ How many teams will participate in FIFA World Cup 2026?

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 ...



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

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

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

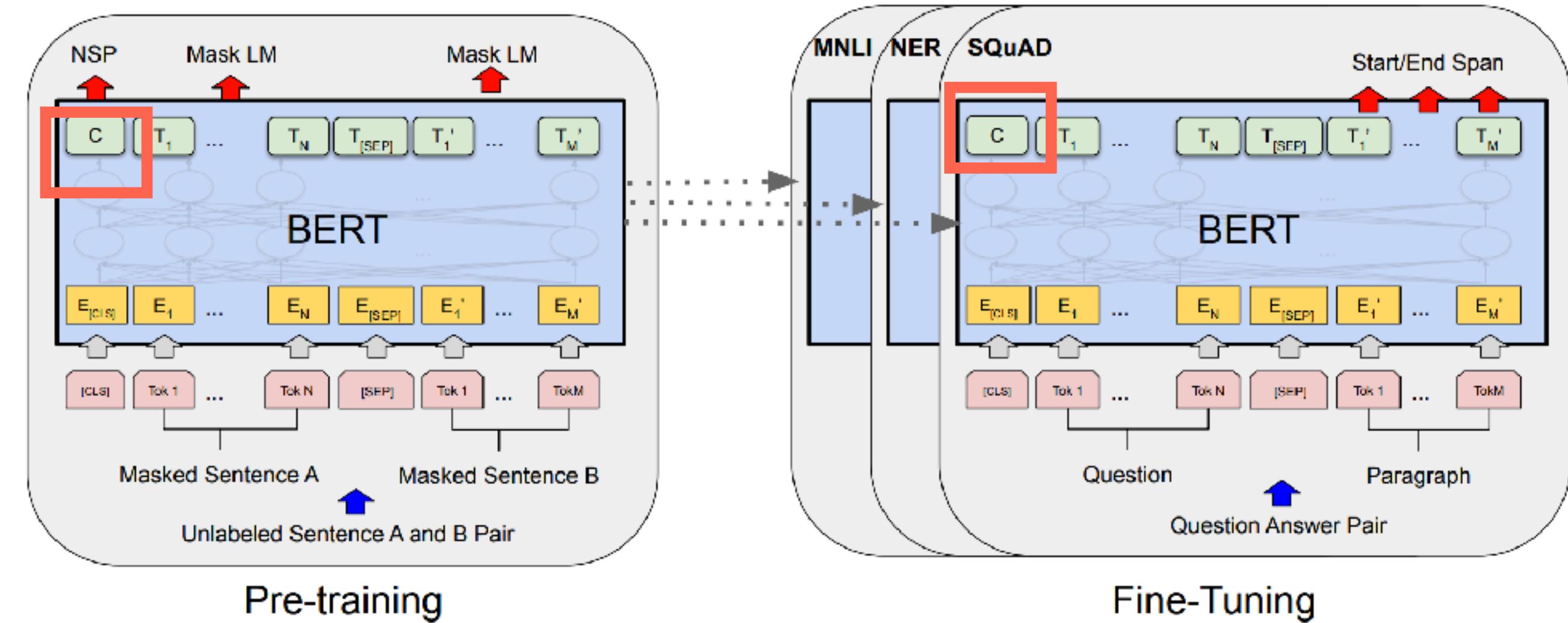
k retrieved chunks

Dense Retrievers: Generating *Embeddings*

- Use output of [CLS] token in masked LMs

e.g., DPR

$$\mathbb{R}^d$$



Karpukhin et al. EMNLP 2020. Dense Passage Retrieval for Open-Domain Question Answering.

Reimers et al. EMNLP 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

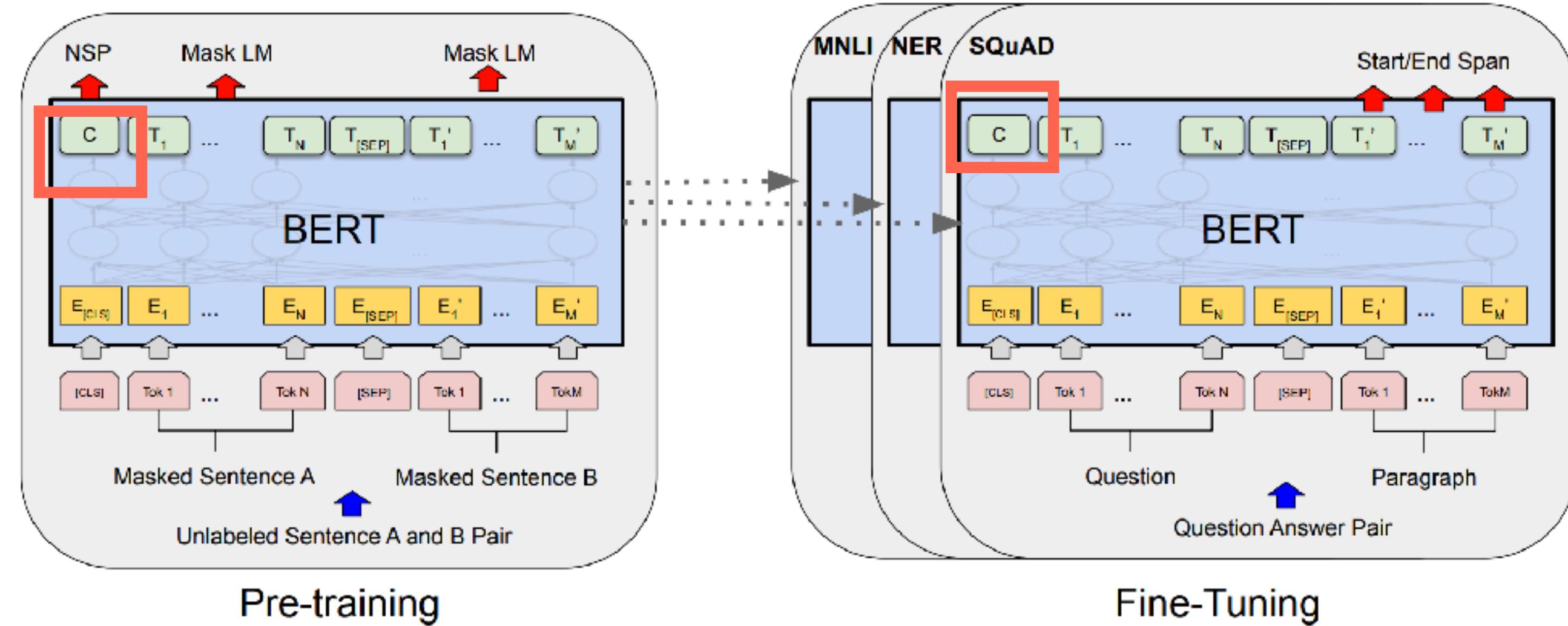
Muennighoff. 2022. SGPT: GPT Sentence Embeddings for Semantic Search.

Dense Retrievers: Generating *Embeddings*

- Use output of [CLS] token in masked LMs

e.g., DPR

$$\mathbb{R}^d$$



- Mean / Max pooling of output vectors

e.g., SBERT, SGPT

$$\mathbb{R}^{N \times d}$$

	NLI	STSB
Pooling Strategy		
MEAN	80.78	87.44
MAX	79.07	69.92
CLS	79.80	86.62

Karpukhin et al. EMNLP 2020. Dense Passage Retrieval for Open-Domain Question Answering.

Reimers et al. EMNLP 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Muennighoff. 2022. SGPT: GPT Sentence Embeddings for Semantic Search.

Fast Nearest Neighbor Search

Method	Class name	index_factory	Main parameters	Bytes/vector	Exhaustive search
Exact Search for L2	IndexFlatL2	"Flat"	d	4*c	yes
Exact Search for Inner Product	IndexFlatIP	"Flat"	d	4*c	yes
Hierarchical Navigable Small World graph exploration	IndexHNSWFlat	"HNSW,Flat"	d, M	4*c + x * M * 2 * 4	no
Inverted file with exact post-verification	IndexIVFFlat	"IVFx,Flat"	quantizer, d, nlists, metric	4*c + 8	no
Locality-Sensitive Hashing (binary flat index)	IndexLSH	-	d, nbits	ceil(nbites/8)	yes
Scalar quantizer (SQ) in flat mode	IndexScalarQuantizer	"SQ8"	d	d	yes
Product quantizer (PQ) in flat mode	IndexPQ	"PQx", "PQ" "x" nbits	d, M, nbits	ceil(M + nbites / 8)	yes
IVF and scalar quantizer	IndexIVFScalarQuantizer	"IVFx,SQ4" "IVFx,SQ8"	quantizer, d, nlists, qtype	SQfp16: 2 * d + 8, SQ8: d + 8 or SQ4: d/2 + 8	no
IVFADC (coarse quantizer+PQ on residuals)	IndexIVFPQ	"IVFx,PQ" "x" nbits	quantizer, d, nlists, M, nbits	ceil(M * nbites/8)-8	no

<https://github.com/facebookresearch/faiss/wiki>

https://speakerdeck.com/matsui_528/cvpr20-tutorial-billion-scale-approximate-nearest-neighbor-search (CVPR 2020 Tutorial)

Fast Nearest Neighbor Search

Method	Class name	index_factory	Main parameters	Bytes/vector	Exhaustive search
Exact Search for L2	IndexFlatL2	"Flat"	d	4*c	yes
Exact Search for Inner Product	IndexFlatIP	"Flat"	d	4*c	yes
Hierarchical Navigable Small World graph exploration	IndexHNSWFlat	"HNSW,Flat"	d, M	4*c + x * M * 2 * 4	no
Inverted file with exact post-verification	IndexIVFFlat	"IVFx,Flat"	quantizer, d, nlists, metric	4*c + 8	no
Locality-Sensitive Hashing (binary flat index)	IndexLSH	-	d, nbits	ceil(nbites/8)	yes
Scalar quantizer (SQ) in flat mode	IndexScalarQuantizer	"SQ8"	d	d	yes
Product quantizer (PQ) in flat mode	IndexPQ	"PQx", "PQ" "x" "nbites"	d, M, nbites	ceil(M + nbites / 8)	yes
IVF and scalar quantizer	IndexIVFScalarQuantizer	"IVFx,SQ4" "IVFx,SQ8"	quantizer, d, nlists, qtype	SQfp16: 2 * d + 8, SQ8: d + 8 or SQ4: d/2 + 8	no
IVFADC (coarse quantizer+PQ on residuals)	IndexIVFPQ	"IVFx,PQ" "x" "nbites"	quantizer, d, nlists, M, nbites	ceil(M * nbites/8)-8	no

Exact search (still fast for $10^6 \sim 10^7$ scale)

<https://github.com/facebookresearch/faiss/wiki>

https://speakerdeck.com/matsui_528/cvpr20-tutorial-billion-scale-approximate-nearest-neighbor-search (CVPR 2020 Tutorial)

Fast Nearest Neighbor Search

Method	Class name	index_factory	Main parameters	Bytes/vector	Exhaustive search
Exact Search for L2	IndexFlatL2	"Flat"	d	4*c	yes
Exact Search for Inner Product	IndexFlatIP	"Flat"	d	4*c	yes
Hierarchical Navigable Small World graph exploration	IndexHNSWFlat	"HNSW,Flat"	d, M * 2 * 4	4*c + x * M * 2 * 4	no
Inverted file with exact post-verification	IndexIVFFlat	"IVFx,Flat"	quantizer, d, nlists, metric	4*c + 8	no
Locality-Sensitive Hashing (binary flat index)	IndexLSH	-	d, nbits	ceil(nbites/8)	yes
Scalar quantizer (SQ) in flat mode	IndexScalarQuantizer	"SQ8"	d	d	yes
Product quantizer (PQ) in flat mode	IndexPQ	"PQx", "PQ" "x" "nbits	d, M, nbits	ceil(M + nbites / 8)	yes
IVF and scalar quantizer	IndexIVFScalarQuantizer	"IVFx,SQ4" "IVFx,SQ8"	quantizer, d, nlists, qlype	SQfp16: 2 * d + 8, SQ8: d + 8 or SQ4: d/2 + 8	no
IVFADC (coarse quantizer+PQ on residuals)	IndexIVFPQ	"IVFx,PQ" "x" "nbits	quantizer, d, nlists, M, nbits	ceil(M * nbites/8)-8	no

Exact search (still fast for $10^6 \sim 10^7$ scale)

Approximate search (faster but more memory)

<https://github.com/facebookresearch/faiss/wiki>

https://speakerdeck.com/matsui_528/cvpr20-tutorial-billion-scale-approximate-nearest-neighbor-search (CVPR 2020 Tutorial)

Fast Nearest Neighbor Search

Method	Class name	index_factory	Main parameters	Bytes/vector	Exhaustive search
Exact Search for L2	IndexFlatL2	"Flat"	d	4*c	yes
Exact Search for Inner Product	IndexFlatIP	"Flat"	d	4*c	yes
Hierarchical Navigable Small World graph exploration	IndexHNSWFlat	"HNSW,Flat"	d, M * 2 * 4	4*c + x * M * 2 * 4	no
Inverted file with exact post-verification	IndexIVFFlat	"IVFx,Flat"	quantizer, d, nlists, metric	4*c + 8	no
Locality-Sensitive Hashing (binary flat index)	IndexLSH	-	d, nbits	ceil(nbites/8)	yes
Scalar quantizer (SQ) in flat mode	IndexScalarQuantizer	"SQ8"	d	d	yes
Product quantizer (PQ) in flat mode	IndexPQ	"PQx", "PQ" "x" nbits	d, M, nbits	ceil(M + nbites / 8)	yes
IVF and scalar quantizer	IndexIVFScalarQuantizer	"IVFx,SQ4" "IVFx,SQ8"	quantizer, d, nlists, qlype	SQfp16: 2 * d + 8, SQ8: d + 8 or SQ4: d/2 + 8	no
IVFADC (coarse quantizer+PQ on residuals)	IndexIVFPQ	"IVFx,PQ" "x" nbits	quantizer, d, nlists, M, nbits	ceil(M * nbites/8)-8	no

Exact search (still fast for $10^6 \sim 10^7$ scale)

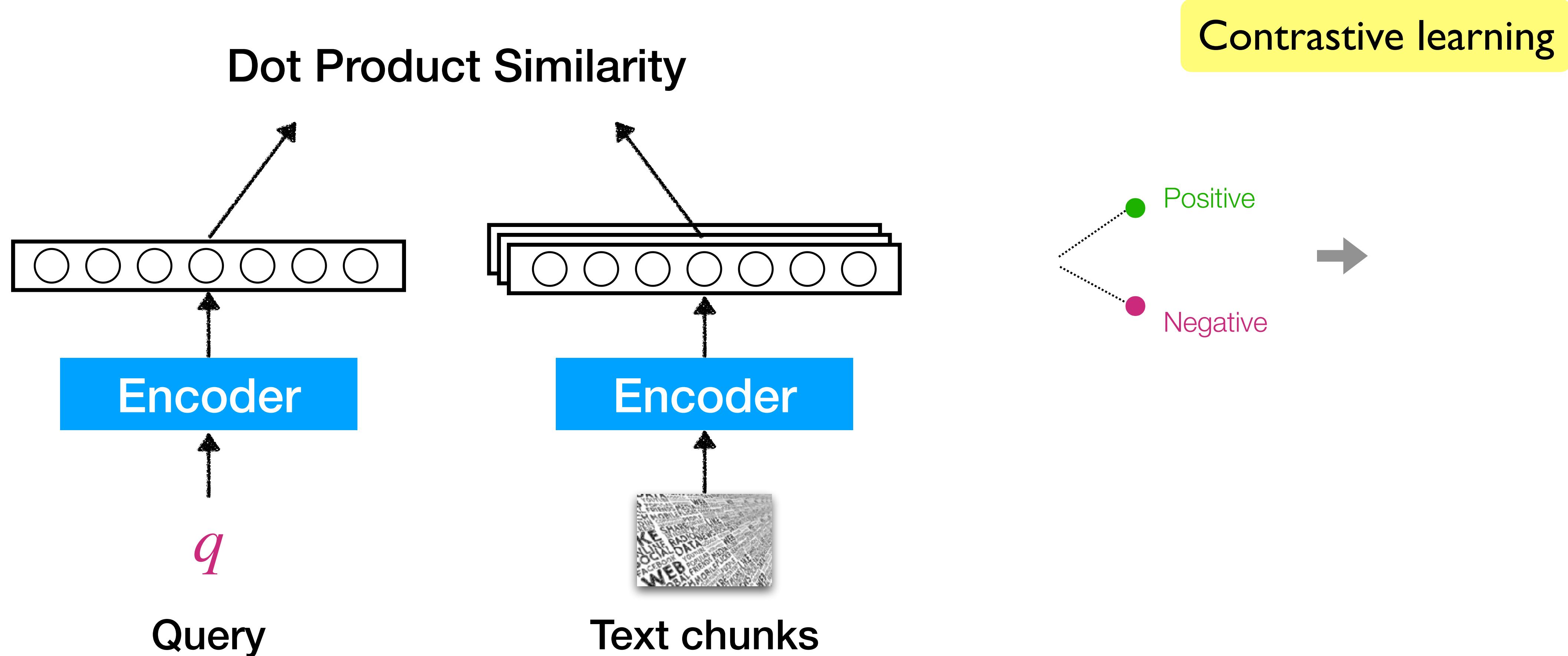
Approximate search (faster but more memory)

Reduce index size with quantization

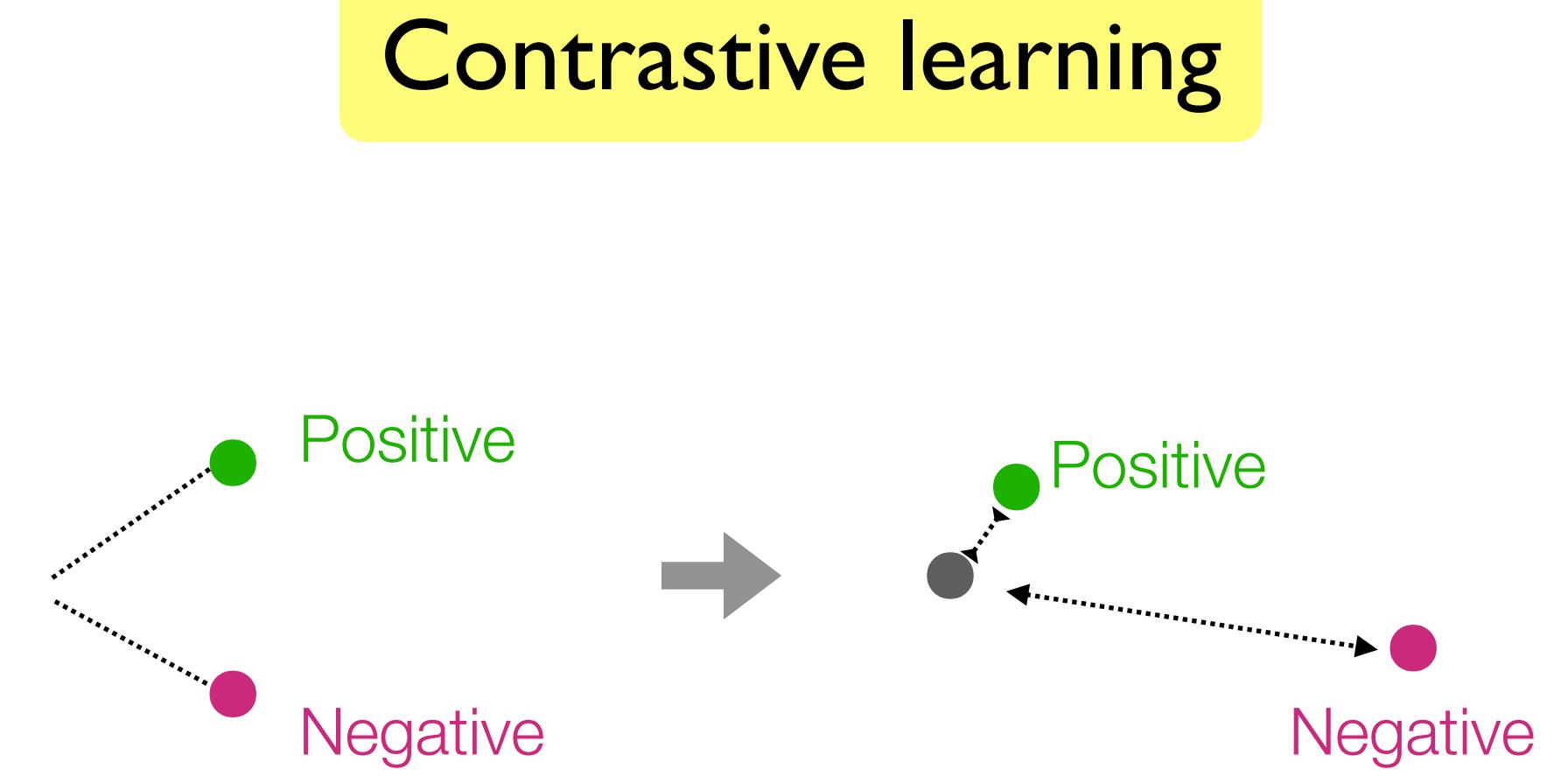
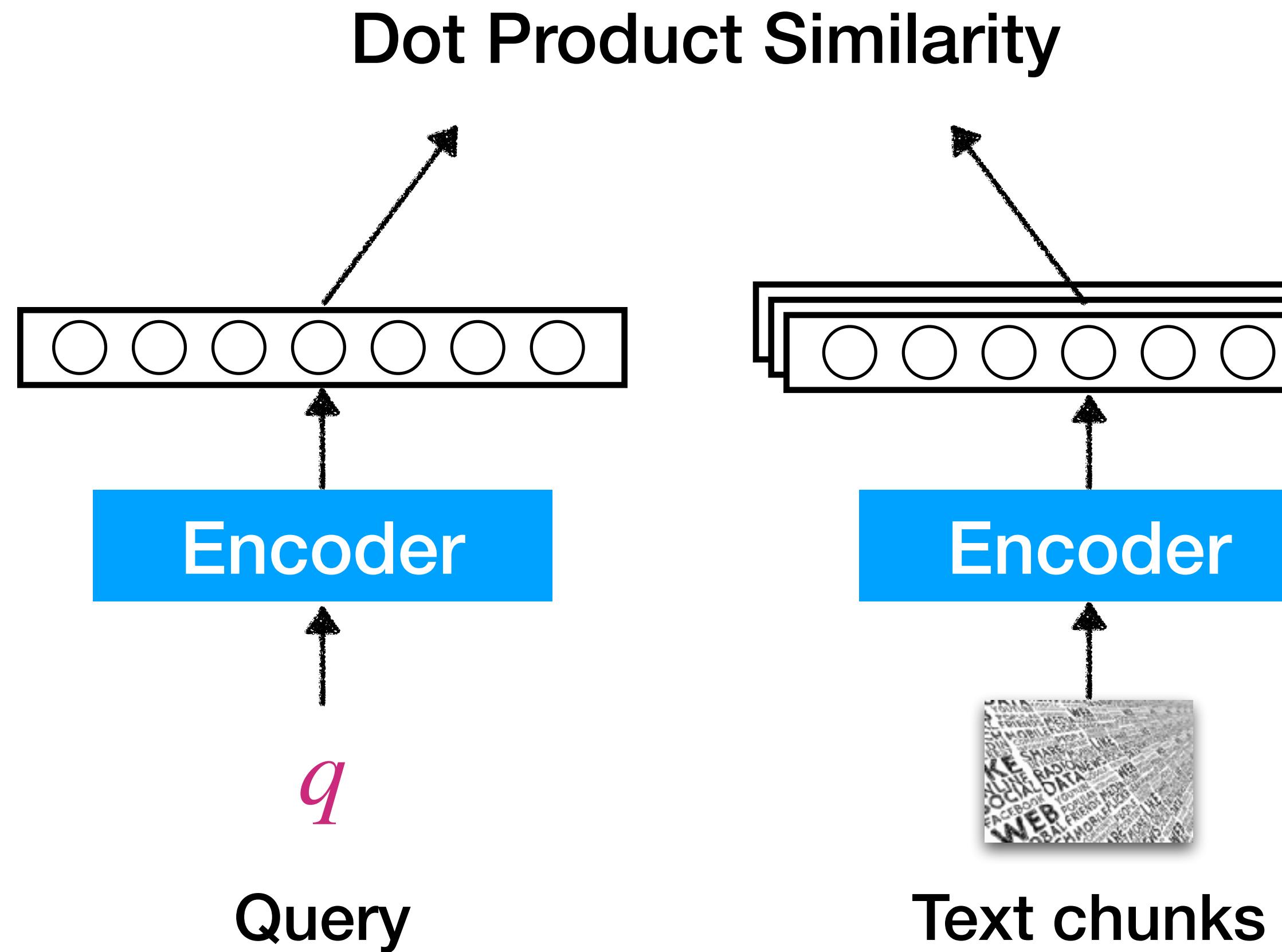
<https://github.com/facebookresearch/faiss/wiki>

https://speakerdeck.com/matsui_528/cvpr20-tutorial-billion-scale-approximate-nearest-neighbor-search (CVPR 2020 Tutorial)

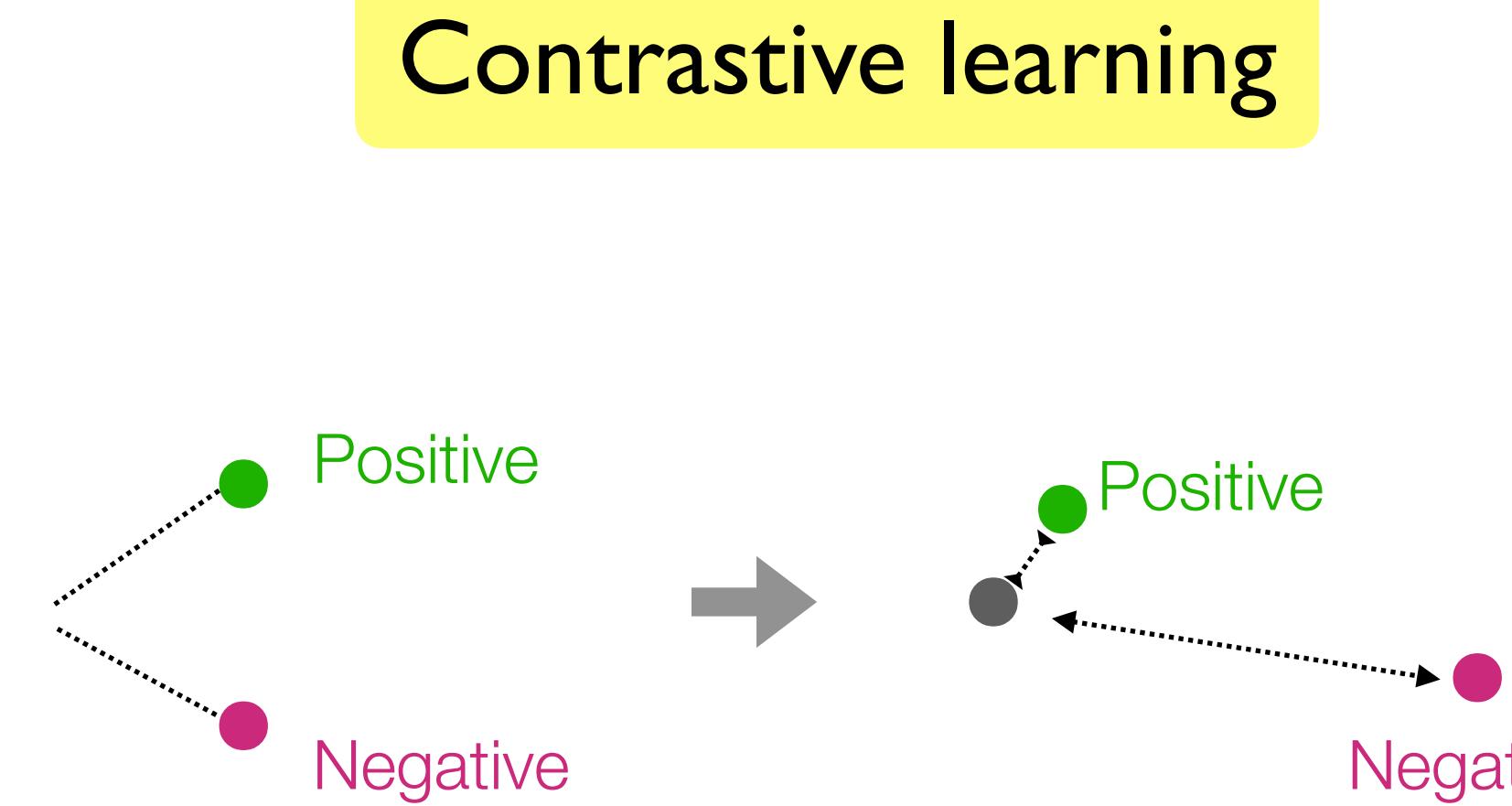
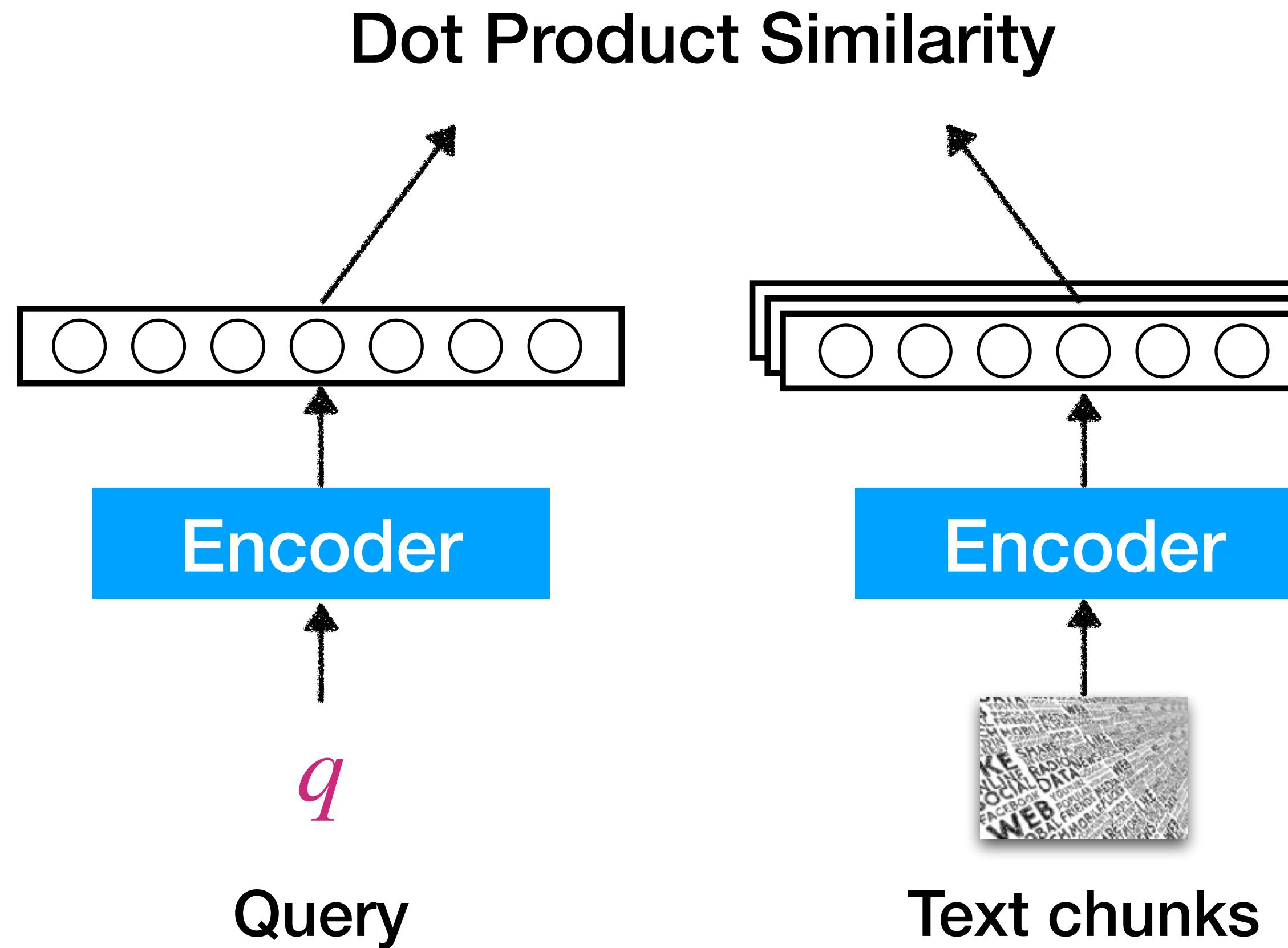
Training Dense Retriever



Training Dense Retriever

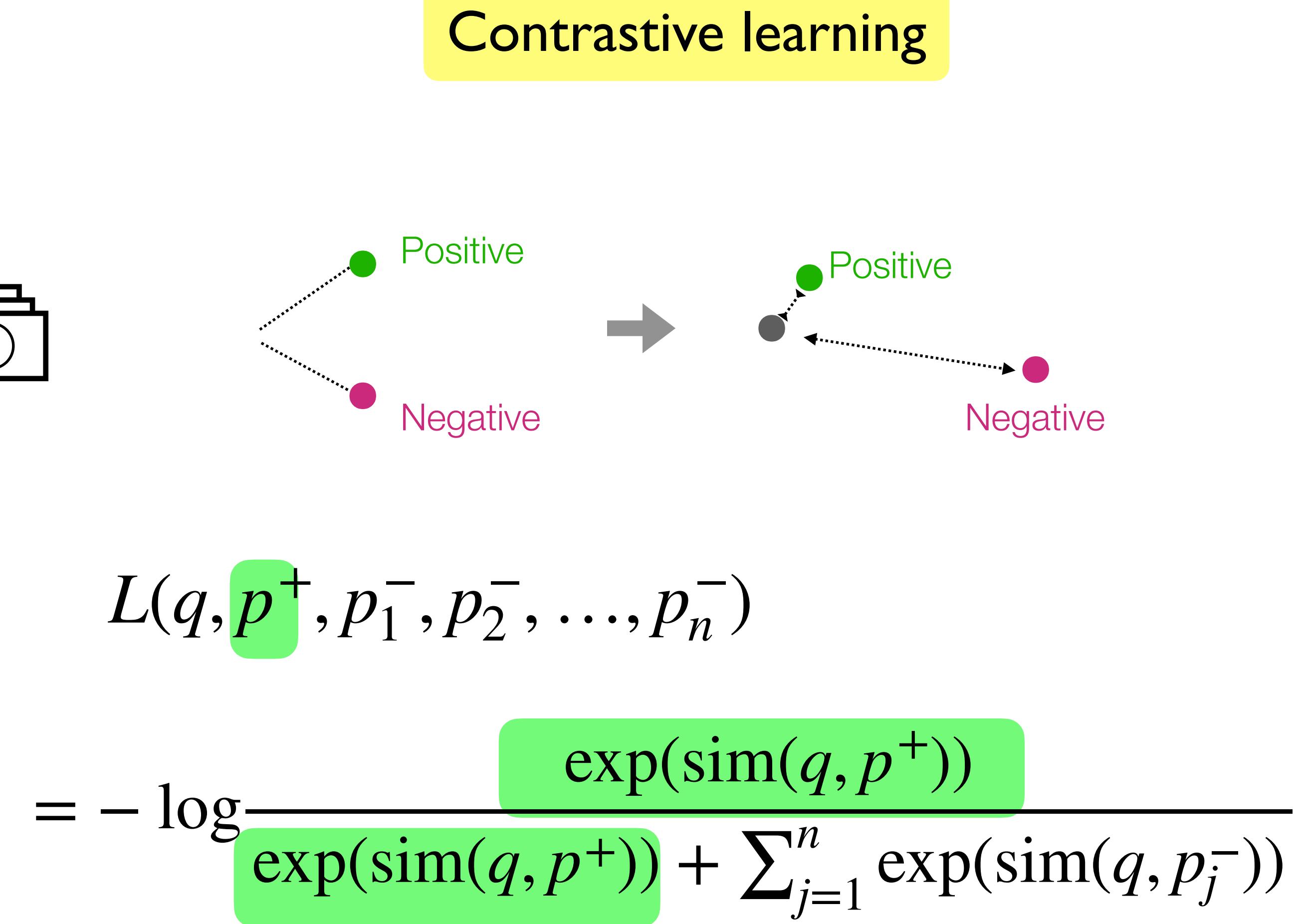
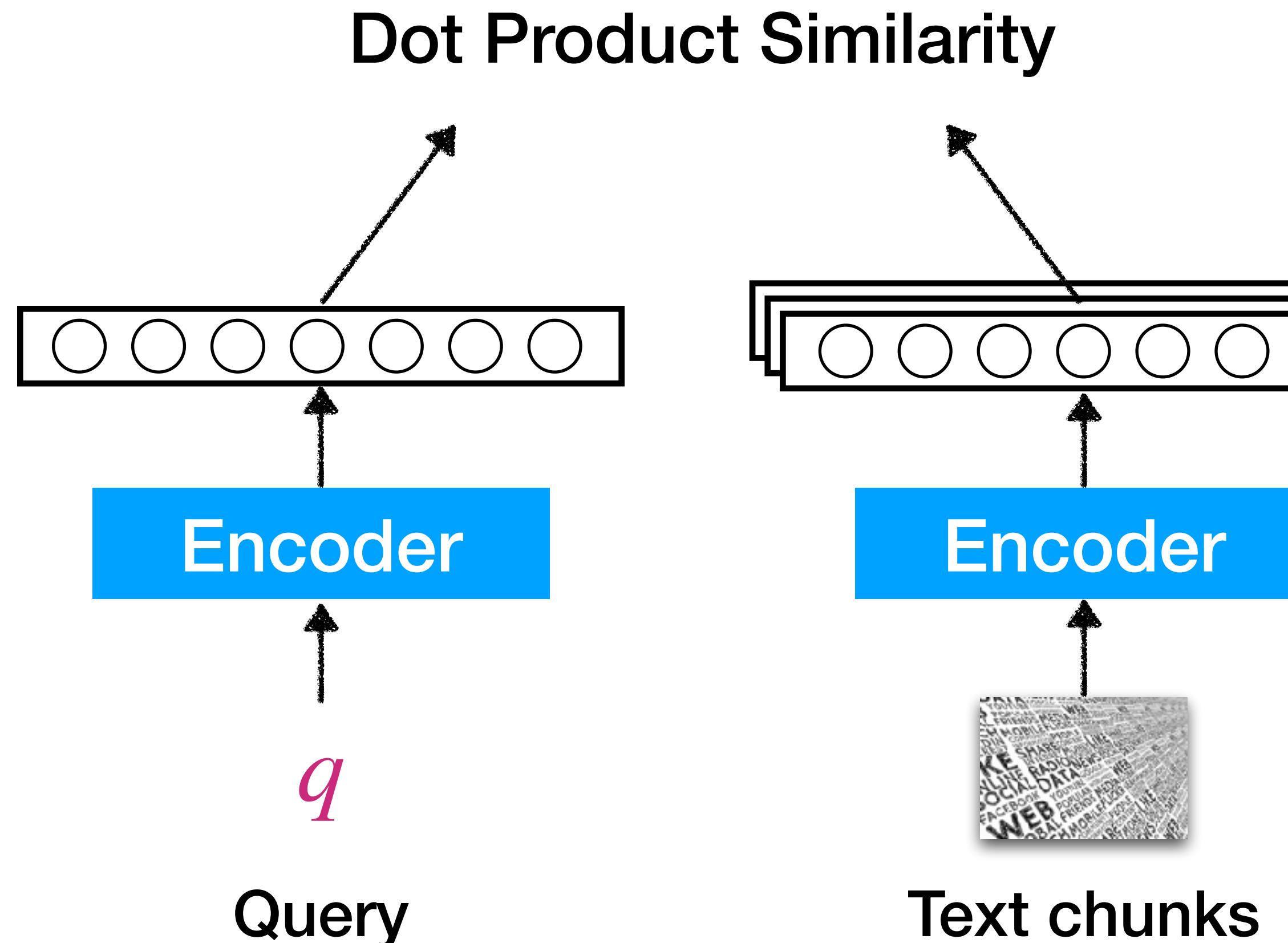


Training Dense Retriever

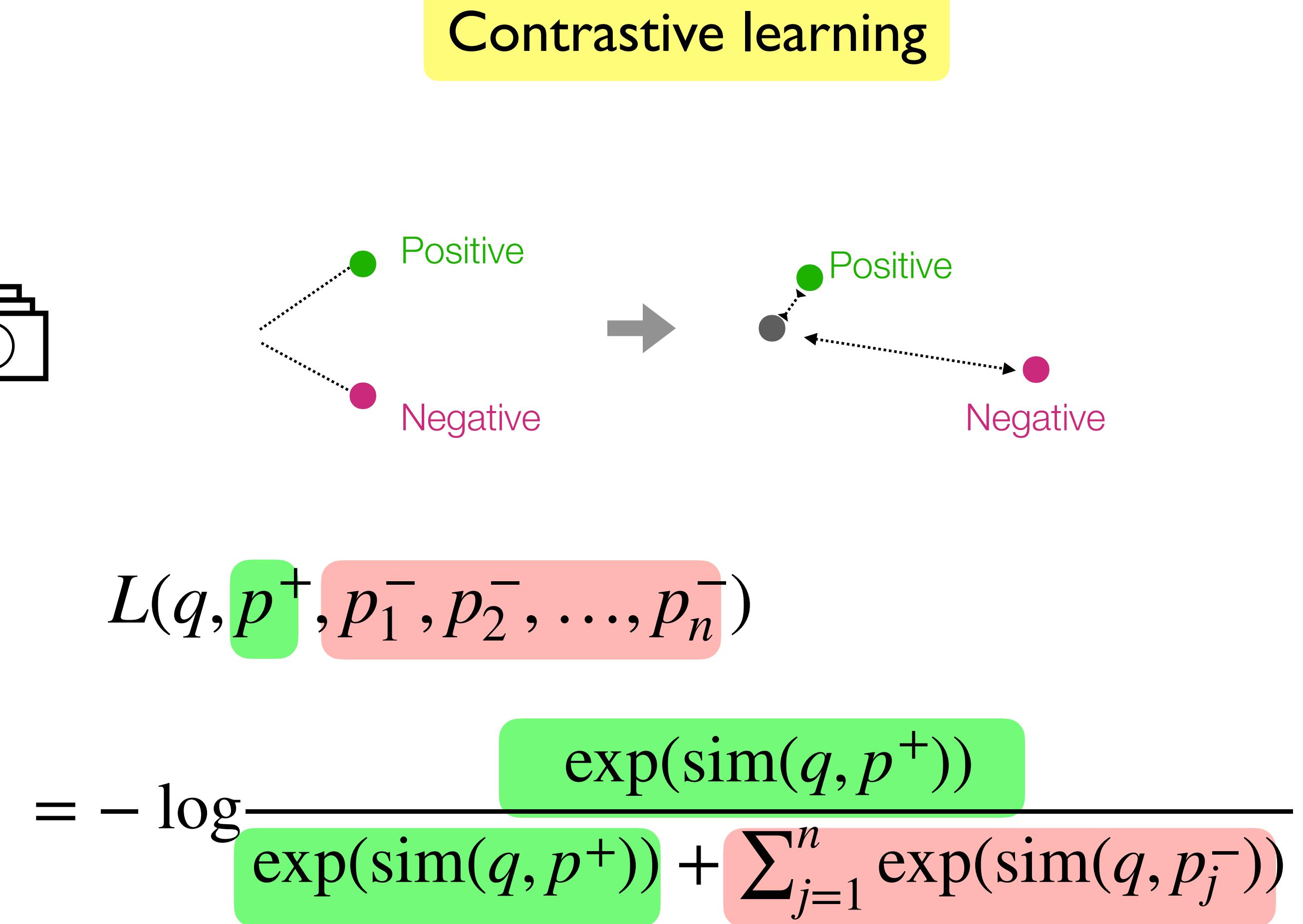
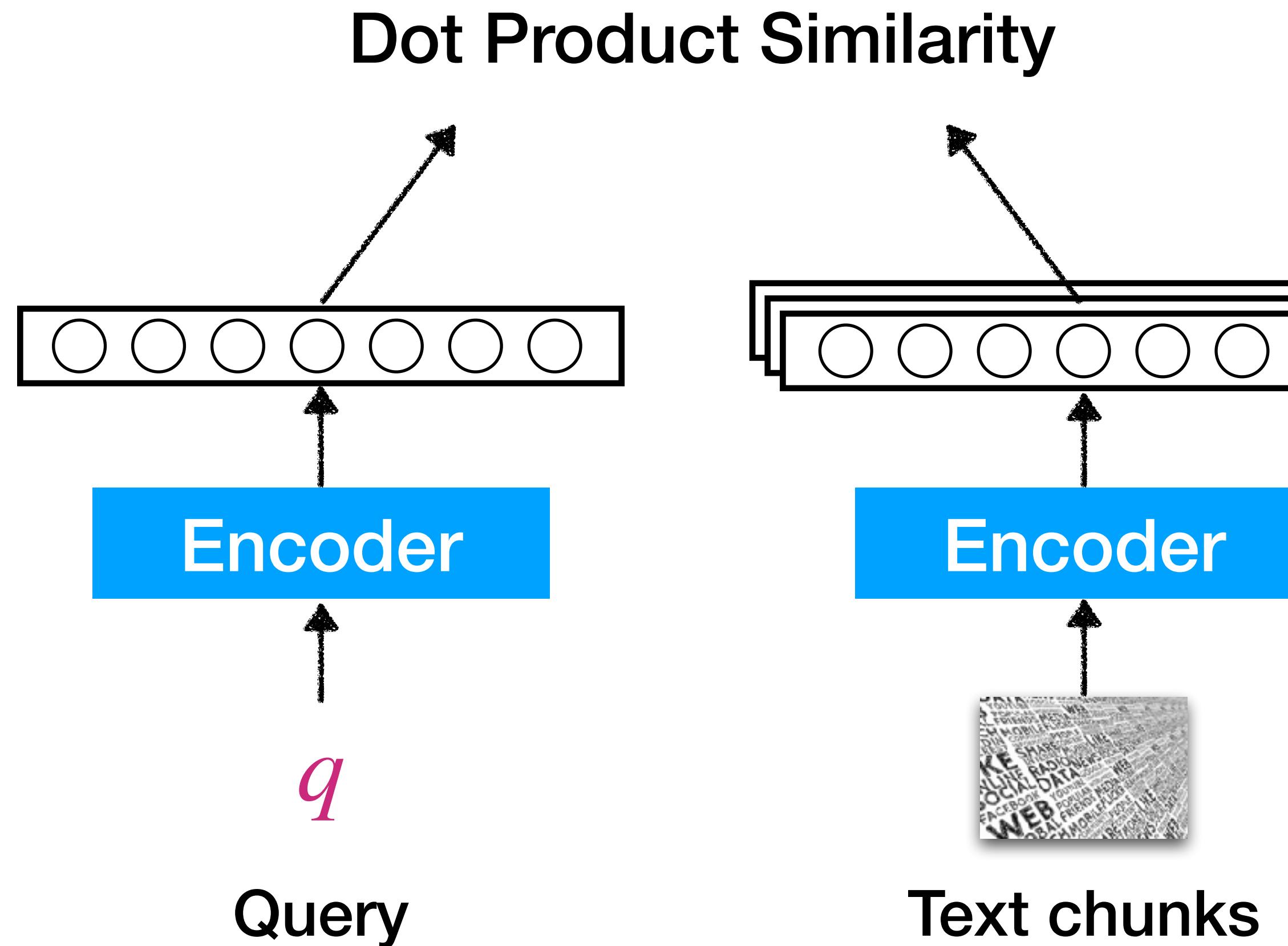


$$L(q, p^+, p_1^-, p_2^-, \dots, p_n^-) = - \log \frac{\exp(\text{sim}(q, p^+))}{\exp(\text{sim}(q, p^+)) + \sum_{j=1}^n \exp(\text{sim}(q, p_j^-))}$$

Training Dense Retriever



Training Dense Retriever

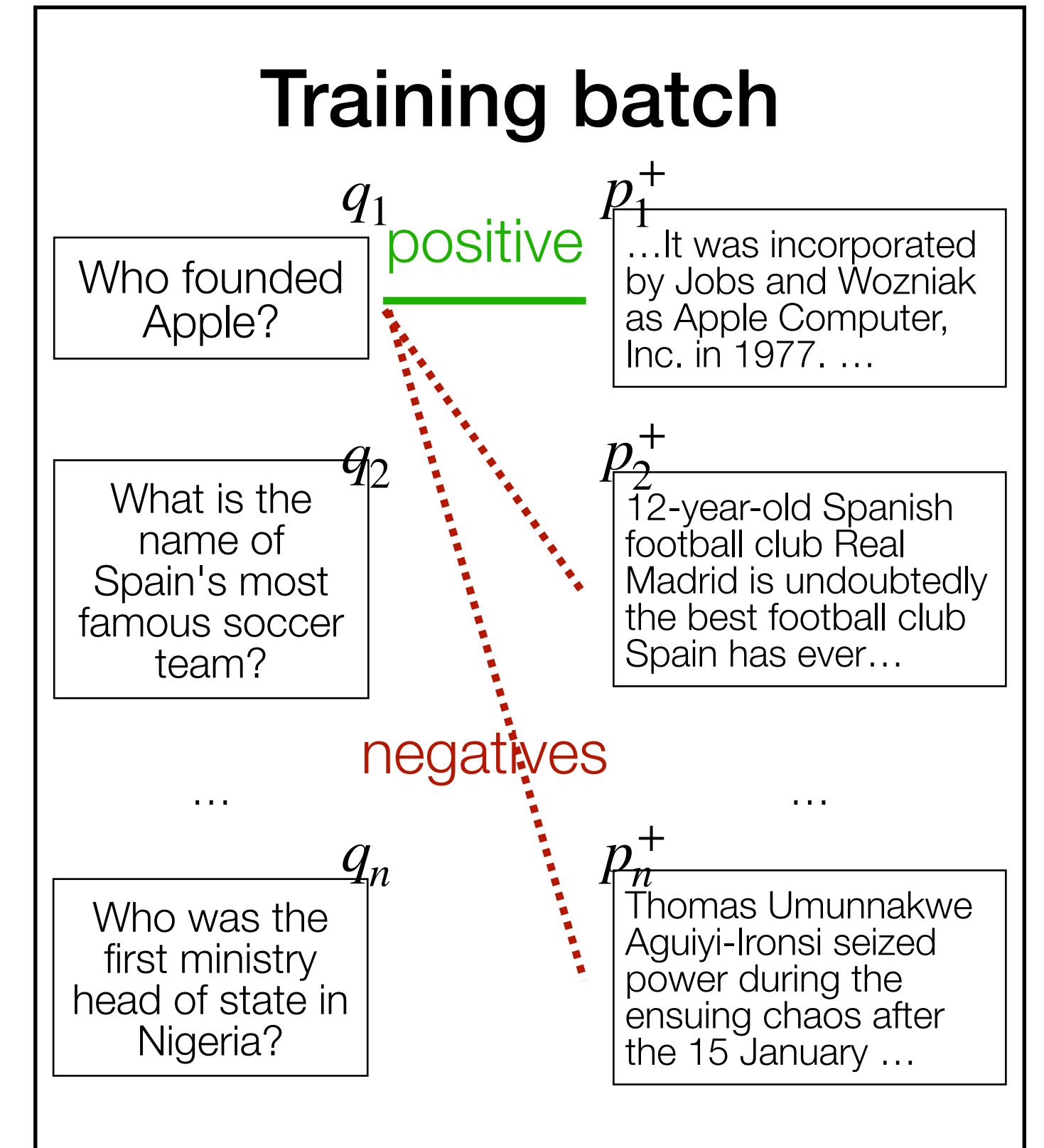


Training Dense Retriever

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

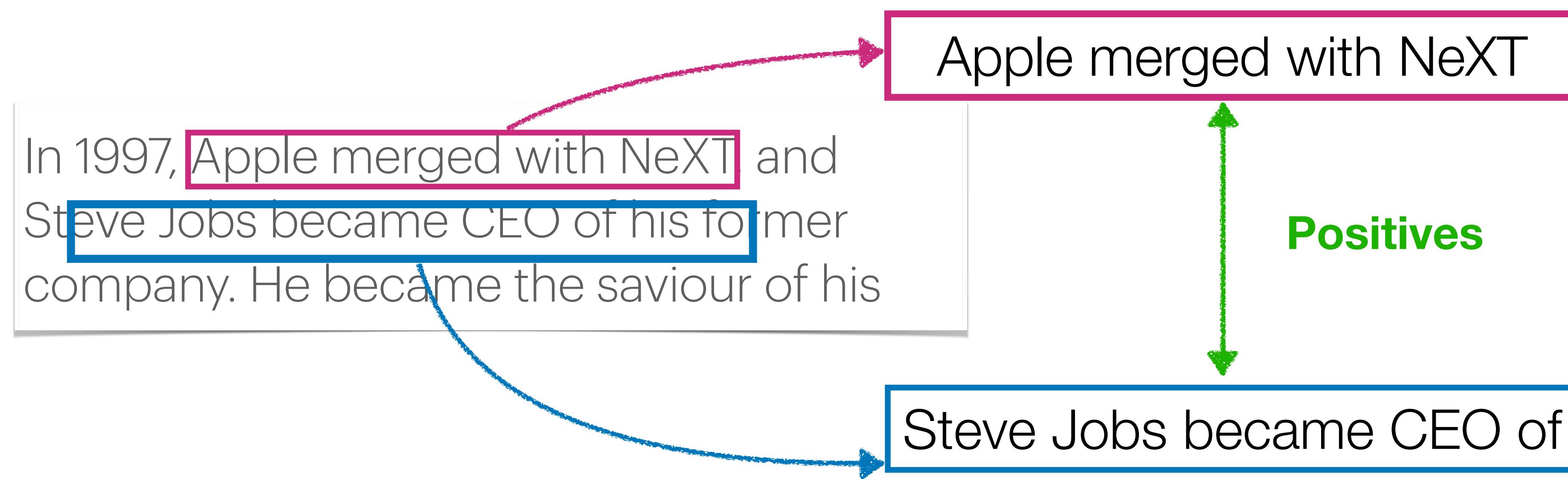
In-batch negatives

Hard negative retrieved by the same / another model



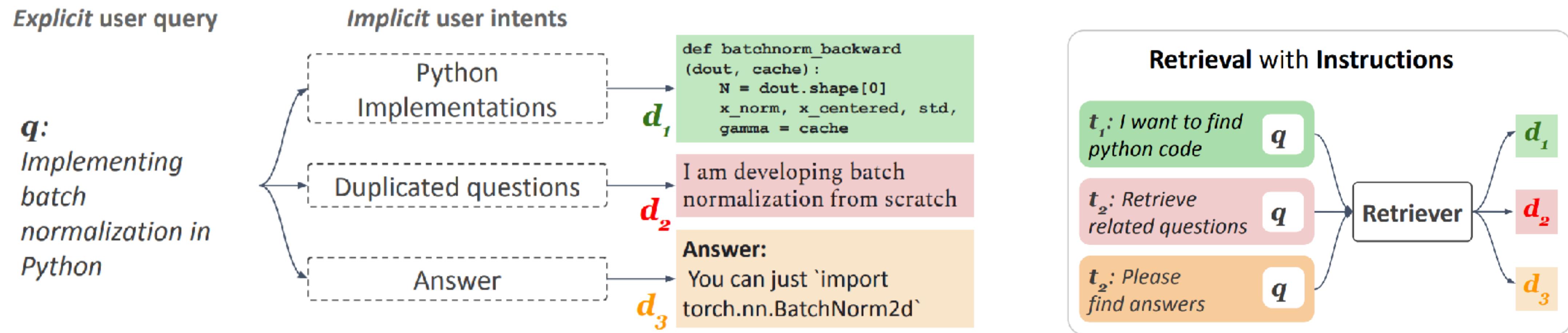
Unsupervised Training

Independent Cropping

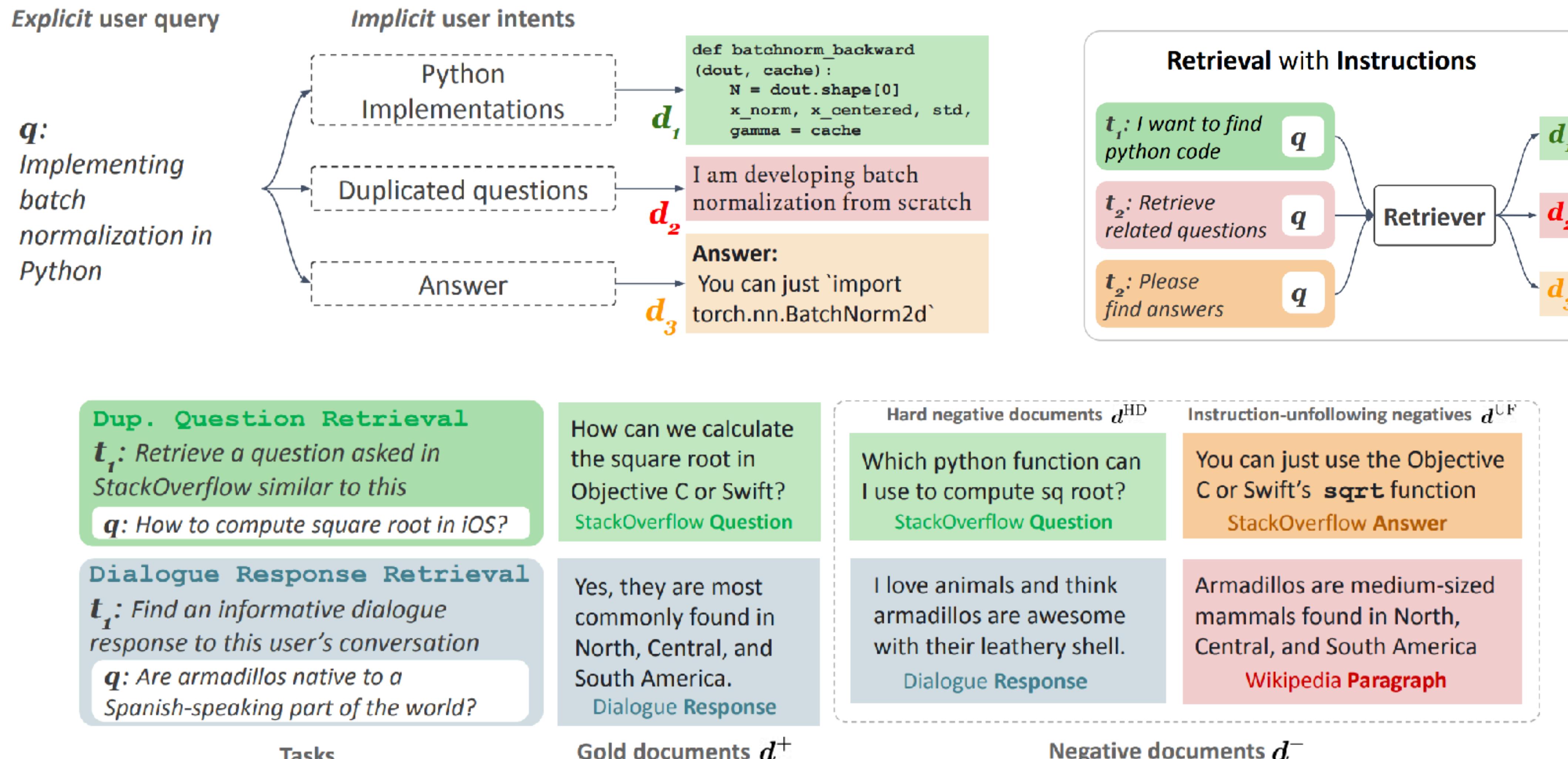


Unsupervised dense retrieval model!

Instruction Tuning for Retriever

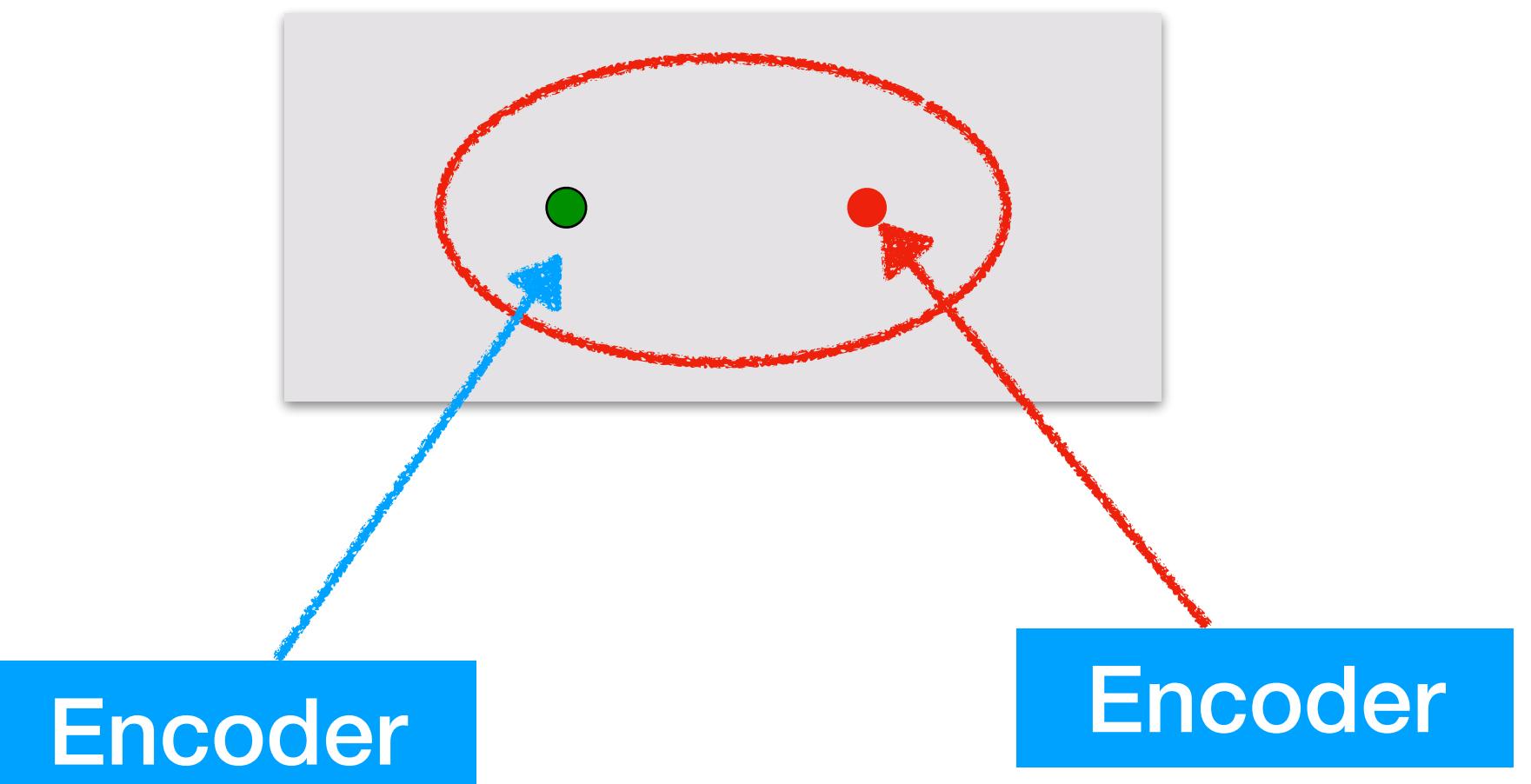


Instruction Tuning for Retriever



Reranking with Cross-encoder

Bi-Encoder

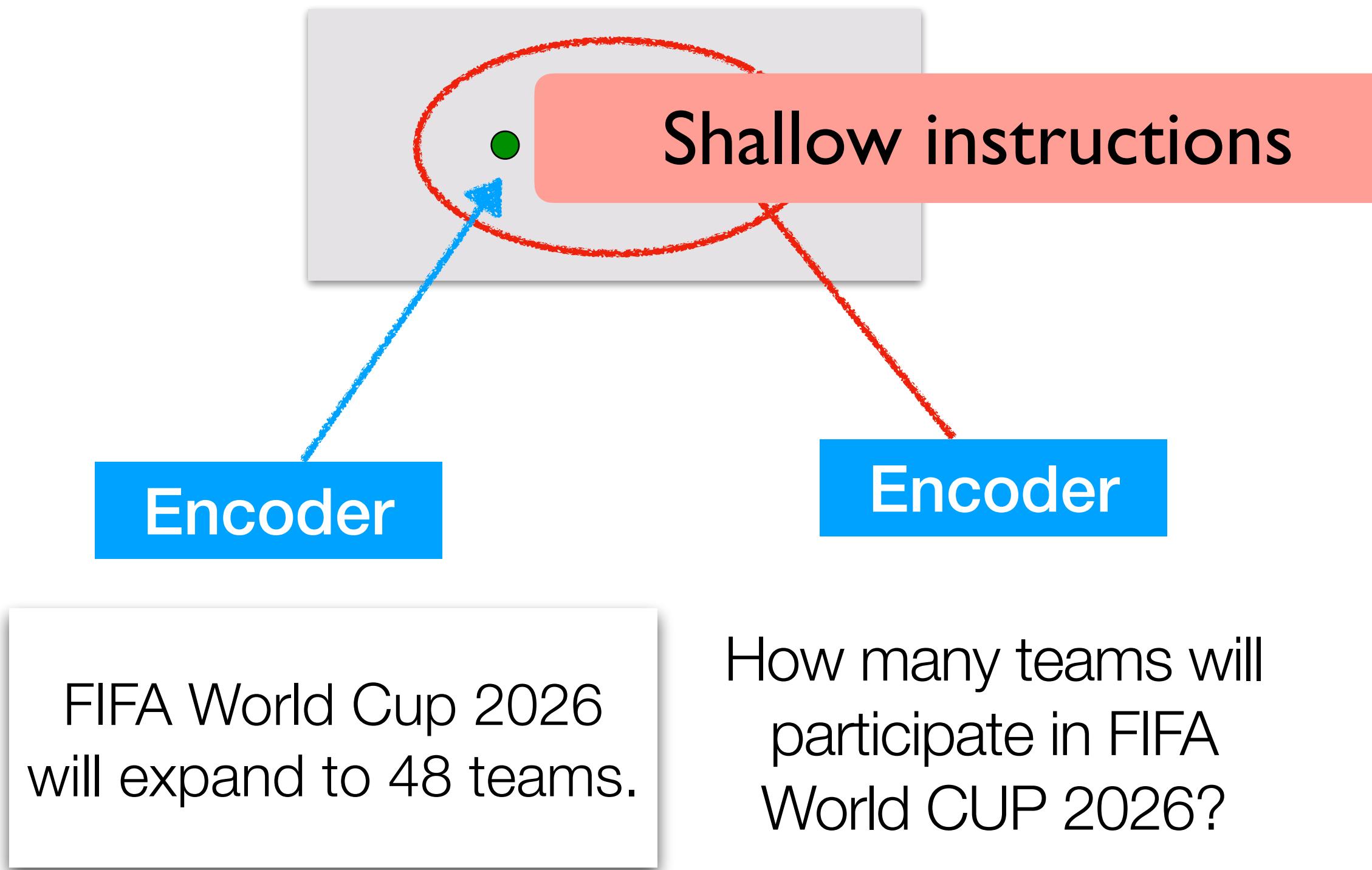


FIFA World Cup 2026
will expand to 48 teams.

How many teams will
participate in FIFA
World CUP 2026?

Reranking with Cross-encoder

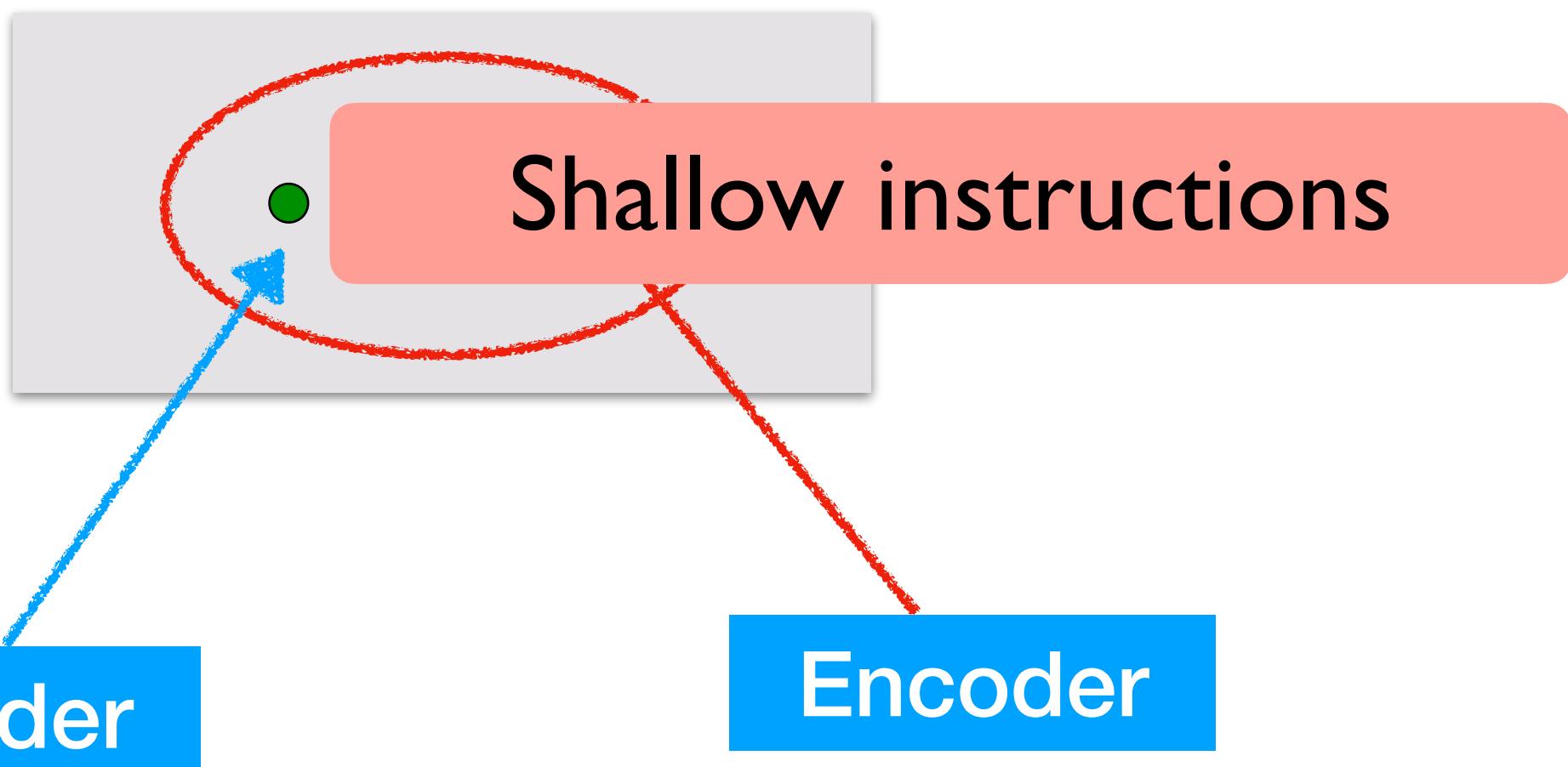
Bi-Encoder



Reranking with Cross-encoder

Bi-Encoder

Cross-Encoder

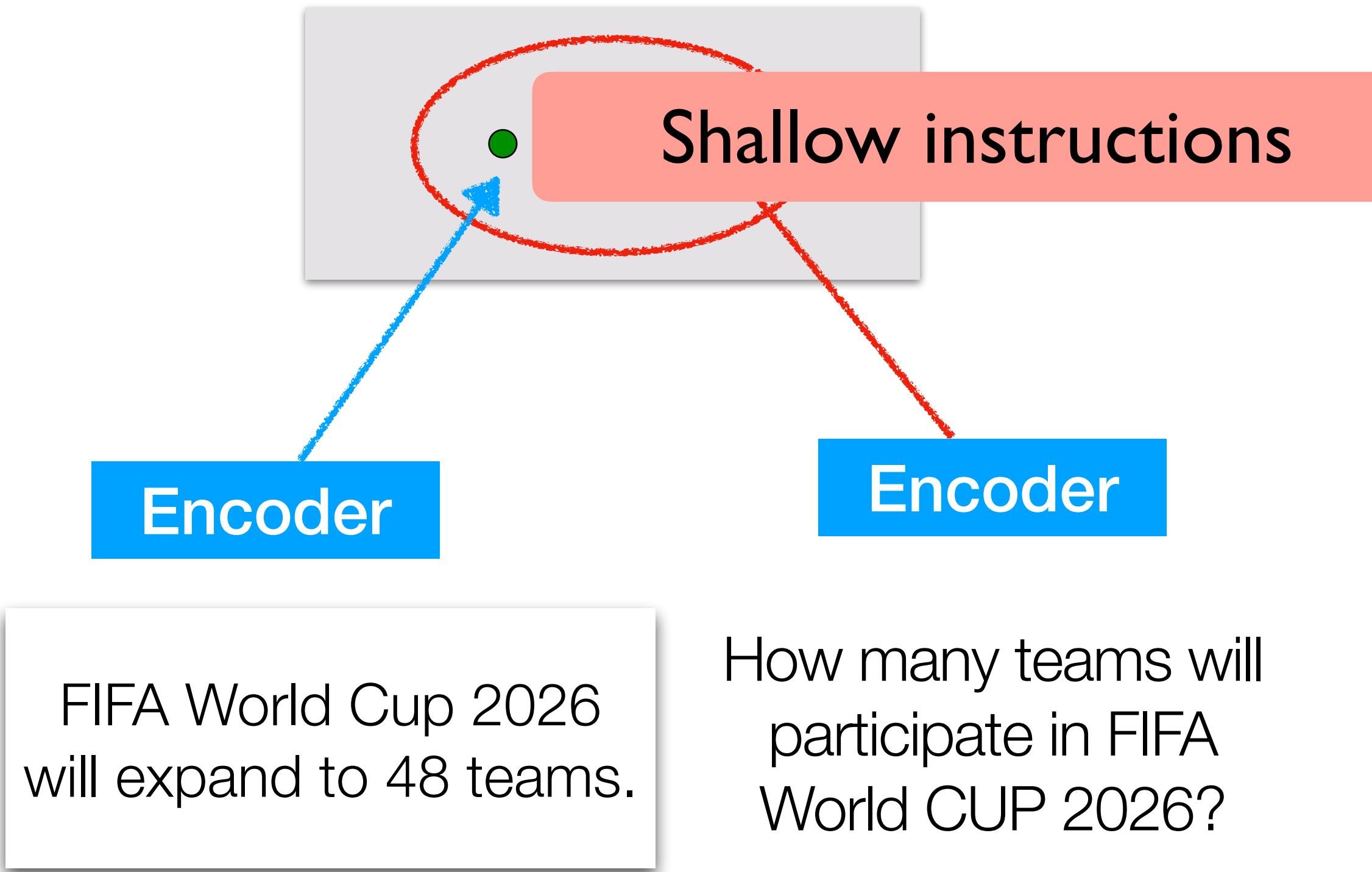


FIFA World Cup 2026
will expand to 48 teams.

How many teams will
participate in FIFA
World CUP 2026?

Reranking with Cross-encoder

Bi-Encoder



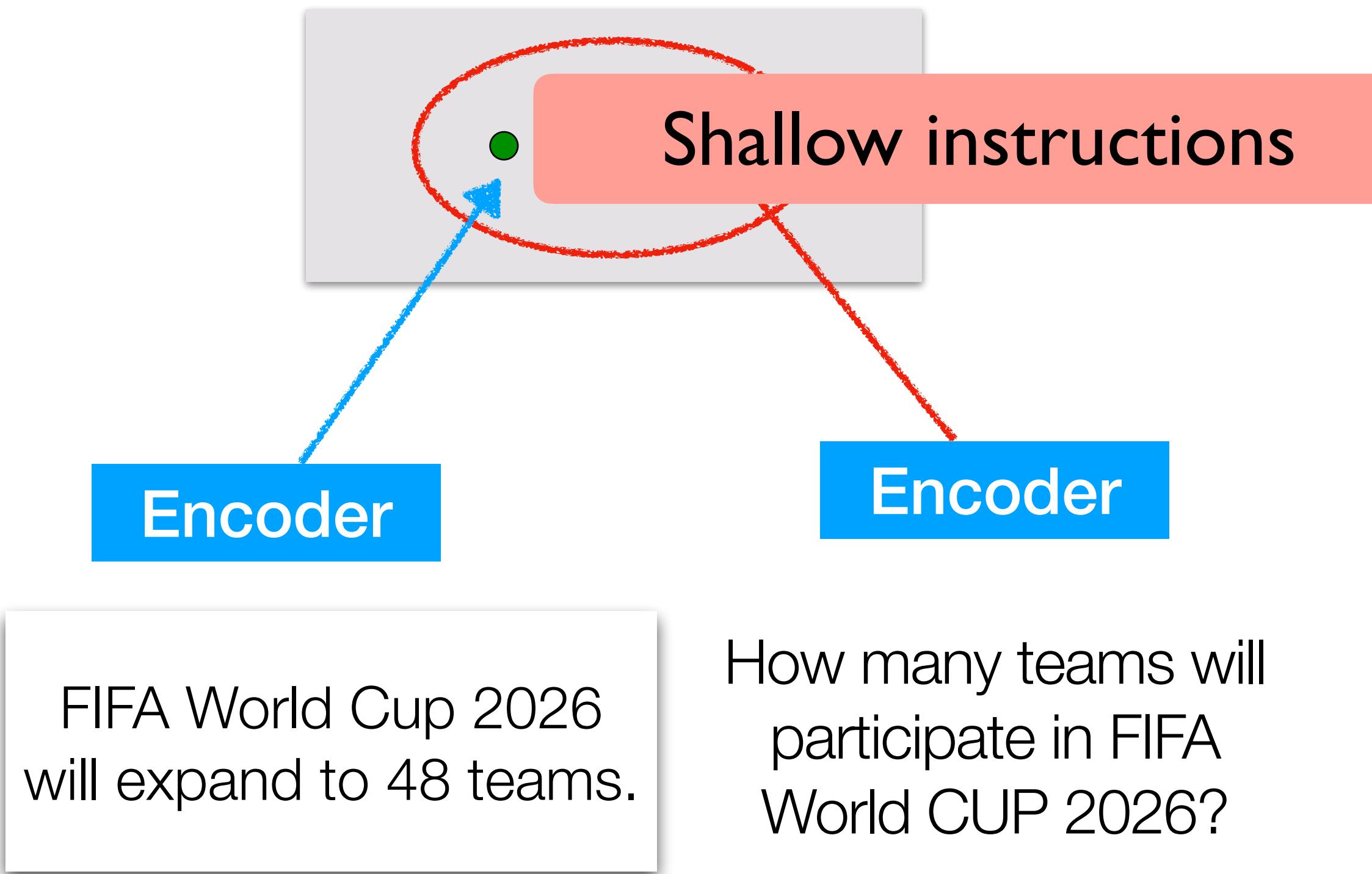
Cross-Encoder

FIFA World Cup 2026 will expand to 48 teams.

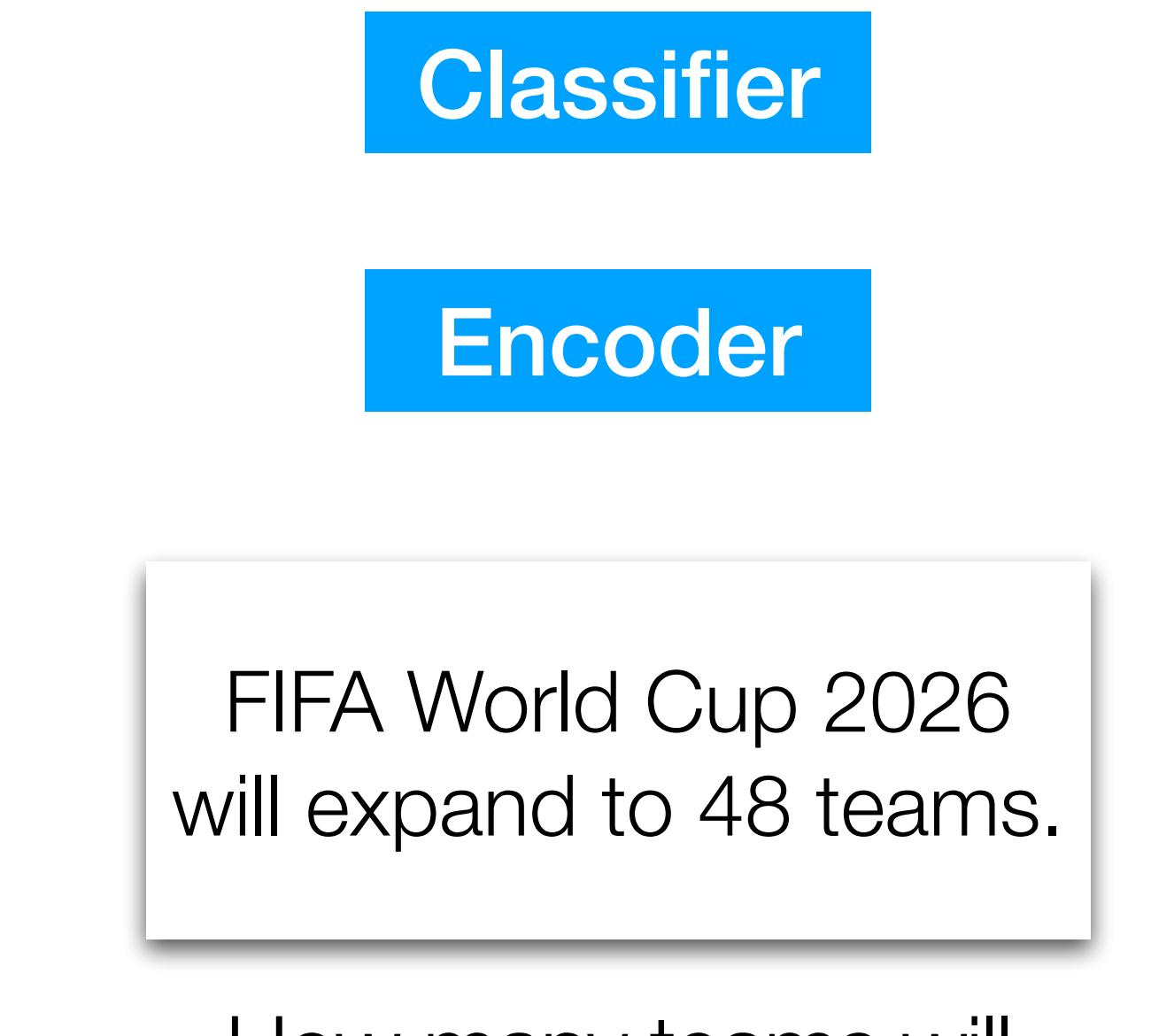
How many teams will participate in FIFA World CUP 2026?

Reranking with Cross-encoder

Bi-Encoder

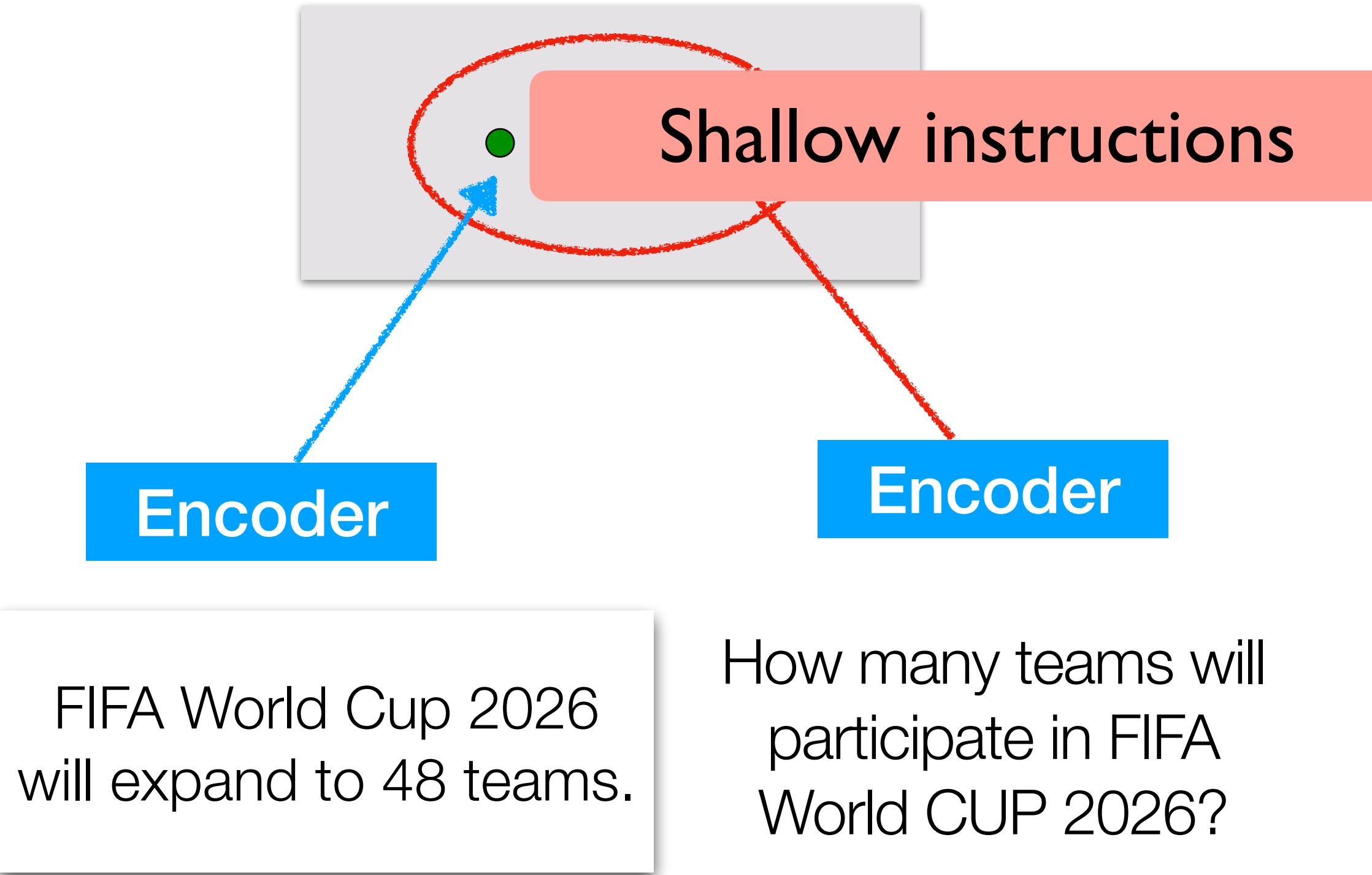


Cross-Encoder



Reranking with Cross-encoder

Bi-Encoder



Cross-Encoder



Evaluation Metrics for Retriever

Evaluation Metrics for Retriever

Evaluation of **unranked** retrieval sets

$$\text{Precision} = \frac{\#\text{(relevant items retrieved)}}{\#\text{(retrieved items)}}$$

$$\text{Recall} = \frac{\#\text{(relevant items retrieved)}}{\#\text{(relevant items)}}$$

Evaluation Metrics for Retriever

Evaluation of **unranked** retrieval sets

$$\text{Precision} = \frac{\#\text{(relevant items retrieved)}}{\#\text{(retrieved items)}}$$

$$\text{Recall} = \frac{\#\text{(relevant items retrieved)}}{\#\text{(relevant items)}}$$

Evaluation of **ranked** retrieval sets

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

$$\text{NDCG}(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^k \frac{2^{R(j,m)} - 1}{\log_2(1 + m)}$$

Evaluation Metrics for Retriever

Evaluation of **unranked** retrieval sets

$$\text{Precision} = \frac{\#\text{(relevant items retrieved)}}{\#\text{(retrieved items)}}$$

$$\text{Recall} = \frac{\#\text{(relevant items retrieved)}}{\#\text{(relevant items)}}$$

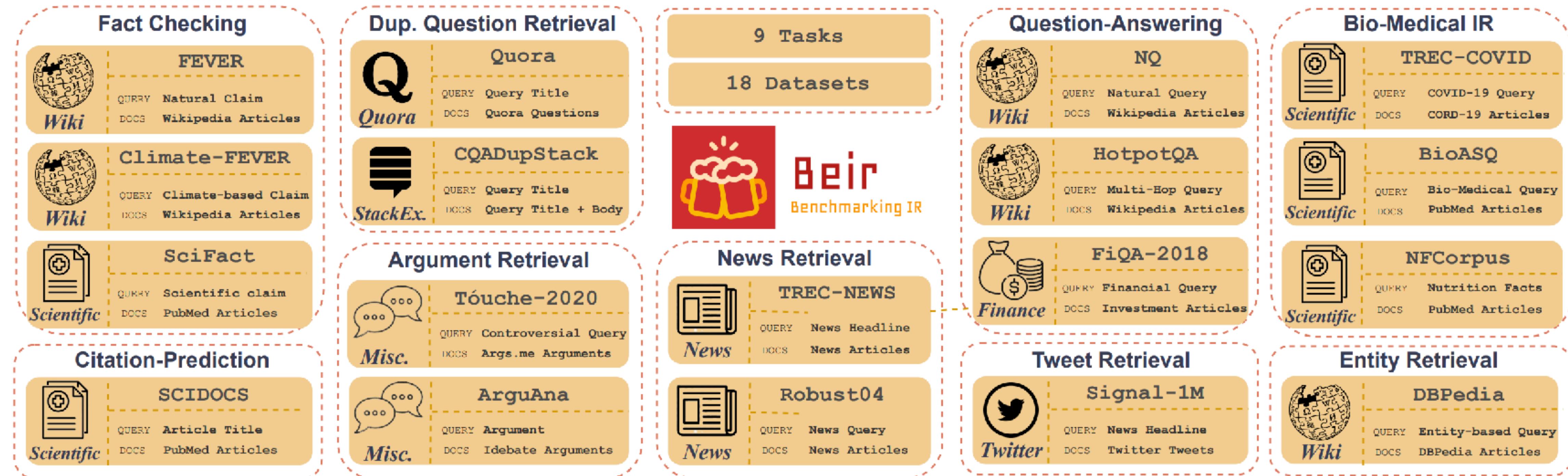
Evaluation of **ranked** retrieval sets

$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

$$\text{NDCG}(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^k \frac{2^{R(j,m)} - 1}{\log_2(1 + m)}$$

nDCG@10 is widely used (e.g., BEIR)

Retrieval Benchmarks: BEIR and MTEB



Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

Retrieval Benchmarks: BEIR and MTEB

Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

BEIR Results

	BM25	BM25+CE
MS MARCO	22.8	41.3
Trec-COVID	65.6	75.7
NFCorpus	32.5	35.0
NQ	32.9	53.3
HotpotQA	60.3	70.7
FiQA	23.6	34.7
ArguAna	31.5	31.1
Touche-2020	36.7	27.1
CQADupStack	29.9	37.0.
Quora	78.9	82.5
DBPedia	31.3	40.9
Scidocs	15.8	16.6
FEVER	75.3	81.9
Climate-FEVER	21.3	25.3
Scifact	66.5	68.8
Avg. w/o CQA	44.0	49.5
Avg.	43.0	48.6
Best on	1	3

Izacard et al. TMLR 2022. Unsupervised Dense Information Retrieval with Contrastive Learning.

BEIR Results

	BM25	BM25+CE
MS MARCO	22.8	41.3
Trec-COVID	65.6	75.7
NFCorpus	32.5	35.0
NQ	32.9	53.3
HotpotQA	60.3	70.7
FiQA	23.6	34.7
ArguAna	31.5	31.1
Touche-2020	36.7	27.1
CQADupStack	29.9	37.0.
Quora	78.9	82.5
DBPedia	31.3	40.9
Scidocs	15.8	16.6
FEVER	75.3	81.9
Climate-FEVER	21.3	25.3
Scifact	66.5	68.8
Avg. w/o CQA	44.0	49.5
Avg.	43.0	48.6
Best on	1	3

Adding CE (cross-encoder) helps

BEIR Results

	BM25	BM25+CE	DPR
MS MARCO	22.8	41.3	17.7
Trec-COVID	65.6	75.7	33.2
NFCorpus	32.5	35.0	18.9
NQ	32.9	53.3	47.4
HotpotQA	60.3	70.7	39.1
FiQA	23.6	34.7	11.2
ArguAna	31.5	31.1	17.5
Touche-2020	36.7	27.1	13.1
CQADupStack	29.9	37.0.	15.3
Quora	78.9	82.5	24.8
DBPedia	31.3	40.9	26.3
Scidocs	15.8	16.6	7.7
FEVER	75.3	81.9	56.2
Climate-FEVER	21.3	25.3	14.8
Scifact	66.5	68.8	31.8
Avg. w/o CQA	44.0	49.5	26.3
Avg.	43.0	48.6	25.5
Best on	1	3	0

Adding CE (cross-encoder) helps

BEIR Results

	BM25	BM25+CE	DPR
MS MARCO	22.8	41.3	17.7
Trec-COVID	65.6	75.7	33.2
NFCorpus	32.5	35.0	18.9
NQ	32.9	53.3	47.4
HotpotQA	60.3	70.7	39.1
FiQA	23.6	34.7	11.2
ArguAna	31.5	31.1	17.5
Touche-2020	36.7	27.1	13.1
CQADupStack	29.9	37.0.	15.3
Quora	78.9	82.5	24.8
DBPedia	31.3	40.9	26.3
Scidocs	15.8	16.6	7.7
FEVER	75.3	81.9	56.2
Climate-FEVER	21.3	25.3	14.8
Scifact	66.5	68.8	31.8
Avg. w/o CQA	44.0	49.5	26.3
Avg.	43.0	48.6	25.5
Best on	1	3	0

Adding CE (cross-encoder) helps

Dense retrievers could struggle
in OOD

BEIR Results

	Contriever					
	BM25	BM25+CE	DPR	Ours	Ours+CE	
MS MARCO	22.8	41.3	17.7	40.7	47.0	
Trec-COVID	65.6	75.7	33.2	59.6	70.1	
NFCorpus	32.5	35.0	18.9	32.8	34.4	
NQ	32.9	53.3	47.4	49.8	57.7	
HotpotQA	60.3	70.7	39.1	63.8	71.5	
FiQA	23.6	34.7	11.2	32.9	36.7	
ArguAna	31.5	31.1	17.5	44.6	41.3	
Touche-2020	36.7	27.1	13.1	23.0	29.8	
CQADupStack	29.9	37.0.	15.3	34.5	37.7	
Quora	78.9	82.5	24.8	86.5	82.4	
DBPedia	31.3	40.9	26.3	41.3	47.1	
Scidocs	15.8	16.6	7.7	16.5	17.1	
FEVER	75.3	81.9	56.2	75.8	81.9	
Climate-FEVER	21.3	25.3	14.8	23.7	25.8	
Scifact	66.5	68.8	31.8	67.7	69.2	
Avg. w/o CQA	44.0	49.5	26.3	47.5	51.2	
Avg.	43.0	48.6	25.5	46.6	50.2	
Best on	1	3	0	1	9	

Adding CE (cross-encoder) helps

Dense retrievers could struggle
in OOD

BEIR Results

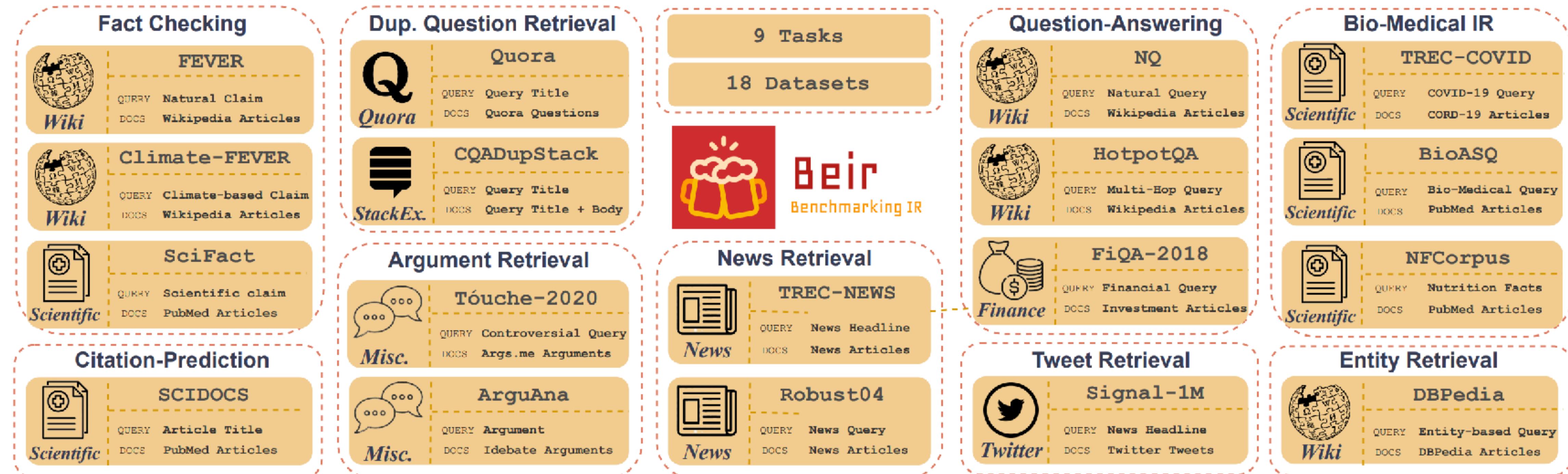
	Contriever					
	BM25	BM25+CE	DPR	Ours	Ours+CE	
MS MARCO	22.8	41.3	17.7	40.7	47.0	
Trec-COVID	65.6	75.7	33.2	59.6	70.1	
NFCorpus	32.5	35.0	18.9	32.8	34.4	
NQ	32.9	53.3	47.4	49.8	57.7	
HotpotQA	60.3	70.7	39.1	63.8	71.5	
FiQA	23.6	34.7	11.2	32.9	36.7	
ArguAna	31.5	31.1	17.5	44.6	41.3	
Touche-2020	36.7	27.1	13.1	23.0	29.8	
CQADupStack	29.9	37.0.	15.3	34.5	37.7	
Quora	78.9	82.5	24.8	86.5	82.4	
DBPedia	31.3	40.9	26.3	41.3	47.1	
Scidocs	15.8	16.6	7.7	16.5	17.1	
FEVER	75.3	81.9	56.2	75.8	81.9	
Climate-FEVER	21.3	25.3	14.8	23.7	25.8	
Scifact	66.5	68.8	31.8	67.7	69.2	
Avg. w/o CQA	44.0	49.5	26.3	47.5	51.2	
Avg.	43.0	48.6	25.5	46.6	50.2	
Best on	1	3	0	1	9	

Adding CE (cross-encoder) helps

Dense retrievers could struggle in OOD

Unsupervised training helps in OOD

Retrieval Benchmarks: BEIR and MTEB

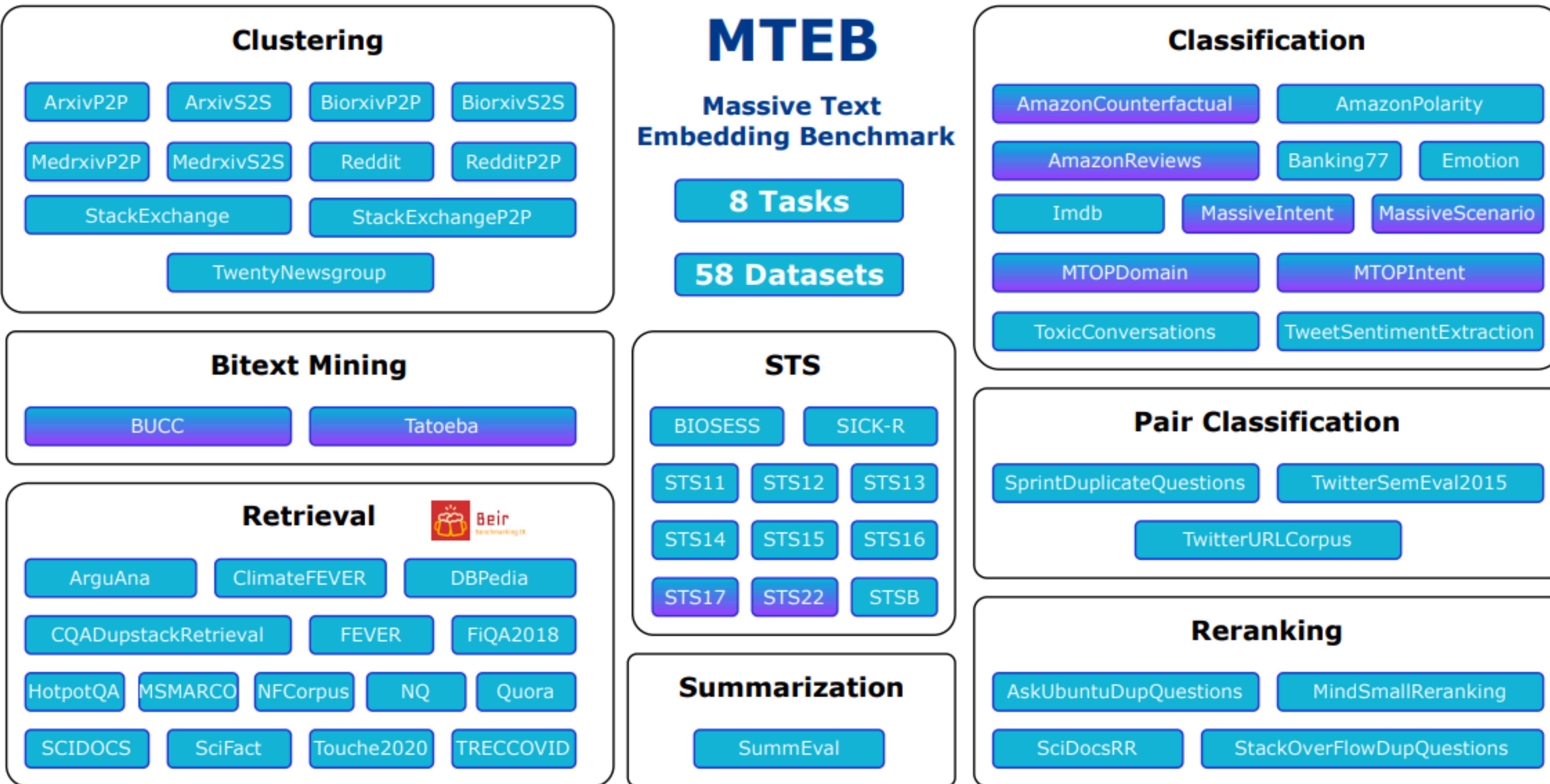


Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

Retrieval Benchmarks: BEIR and MTEB

Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

Retrieval Benchmarks: BEIR and MTEB



Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

MTEB Leaderboard

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
1	voyage-3-m-exp					74.03
2	NV-Embed-v2	7851	29.25	4096	32768	72.31
3	bge-en-ic1	7111	26.49	4096	32768	71.67
4	LENS-d8000	7111	26.49	4096	32768	71.62
5	jasper_en_vision_language_v1					71.54

https://huggingface.co/spaces/mteb/leaderboard_legacy

MTEB Leaderboard

			Memory			Average (56 datasets)
			$q_{\text{inst}}^+ = \text{Instruct} : \{\text{task_definition}\}$ Query : q^+			
1	<u>voyage-3-m-exp</u>					74.03
2	<u>NV-Embed-v2</u>	7851	29.25	4096	32768	72.31
3	<u>bge-en-ic1</u>	7111	26.49	4096	32768	71.67
4	<u>LENS-d8000</u>	7111	26.49	4096	32768	71.62
5	<u>jasper_en_vision_language_v1</u>					71.54

https://huggingface.co/spaces/mteb/leaderboard_legacy

MTEB Leaderboard

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
1	voyage-3-m-exp					74.03
2	NV-Embed-v2	7851	29.25	4096	32768	72.31
3	bge-en-ic1	7111	26.49	4096	32768	71.67
4	LENS-d8000	7111	26.49	4096	32768	71.62
5	jasper_en_vision_language_v1					71.54

https://huggingface.co/spaces/mteb/leaderboard_legacy

MTEB Leaderboard

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
1	voyage-3-m-exp					74.03
2	NV-Embed-v2	7851	29.25	4096	32768	72.31
3	bge-en-ic1	7111	26.49	4096	32768	71.67
4	LENS-d8000	7111	26.49	4096	32768	71.62
5	jasper_en_vision_language_v1					71.54

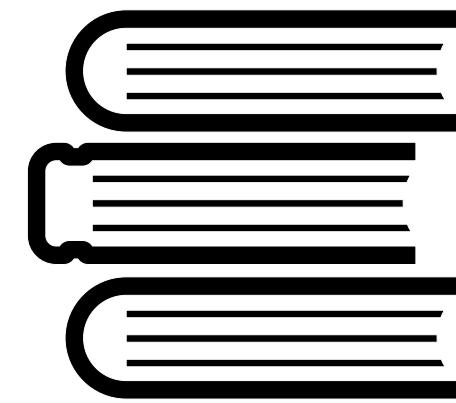
https://huggingface.co/spaces/mteb/leaderboard_legacy

MTEB Leaderboard

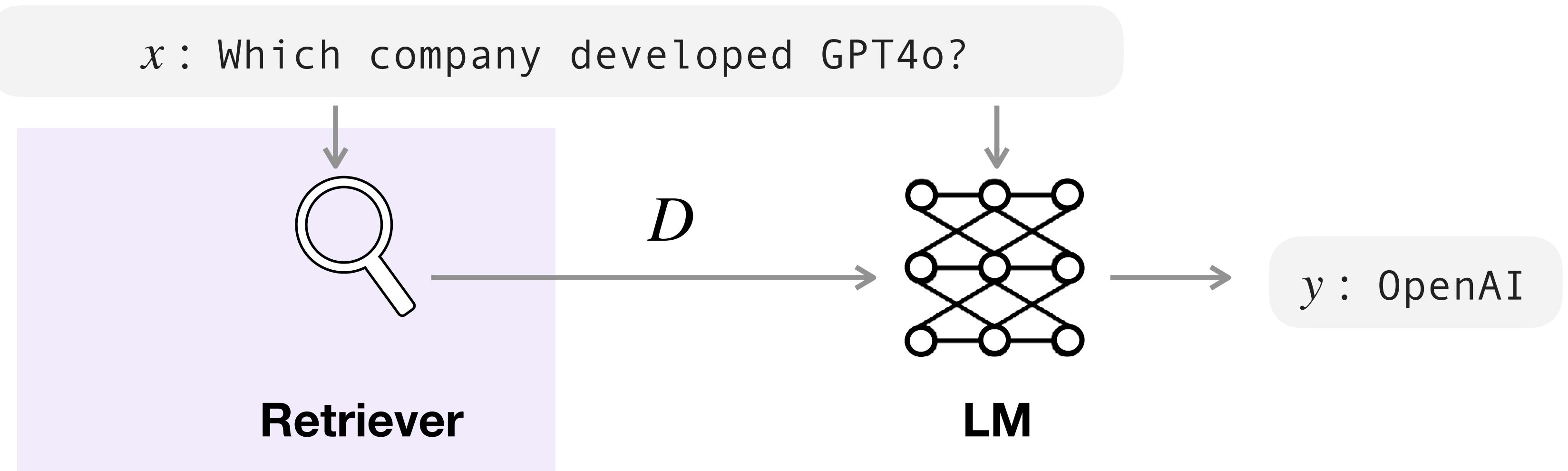
Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
1	voyage-3-m-exp					74.03
2	NV-Embed-v2	7851	29.25	4096	32768	72.31
3	bge-en-ic1	7111	26.49	4096	32768	71.67
4	LENS-d8000	7111	26.49	4096	32768	71.62
5	jasper_en_vision_language_v1					71.54
...						
30	cde-small-v2	150	0.56	768	512	65.58

https://huggingface.co/spaces/mteb/leaderboard_legacy

Summary of Part 2



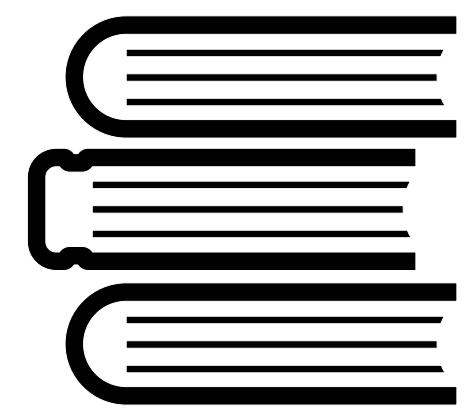
Datastore



- ✓ Types of retrievers
- ✓ Training
- ✓ Evaluations

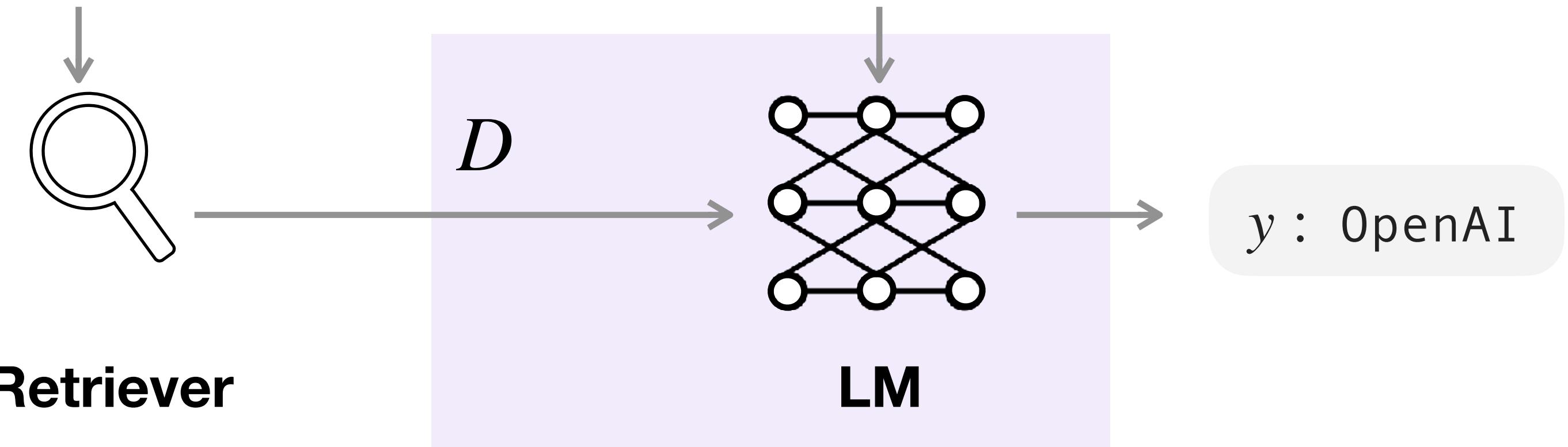
- Different types of retrievers
- Training with contrastive loss
- Common metrics: NDCG@10, Recall ... etc
- Performance v.s. cost trade off

Today's Outline



Datastore

x : Which company developed GPT4o?

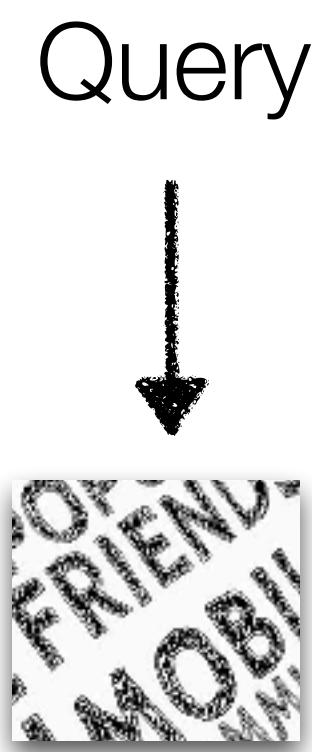


- ✓ Common architectures
- ✓ Recent progress

Categorizing Retrieval-Augmented LMs

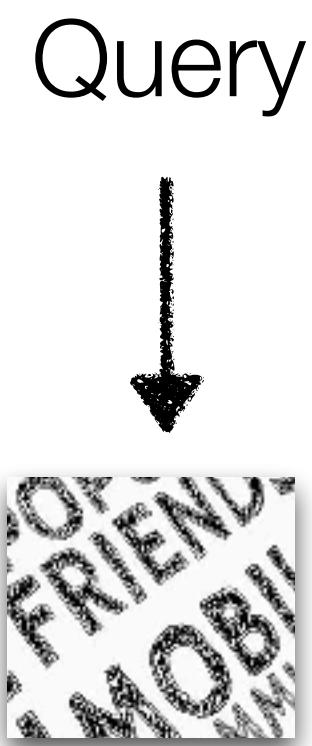
Categorizing Retrieval-Augmented LMs

What to retrieve?



Categorizing Retrieval-Augmented LMs

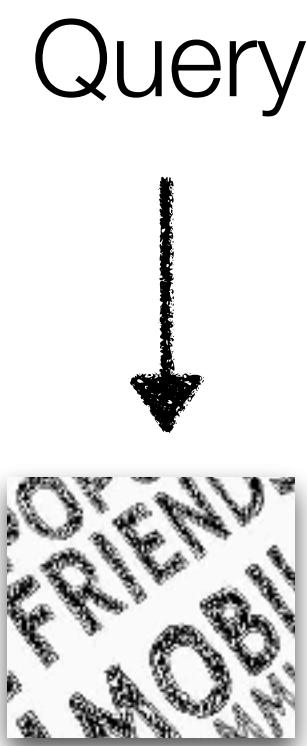
What to retrieve?



Text chunks (passages)?

Categorizing Retrieval-Augmented LMs

What to retrieve?

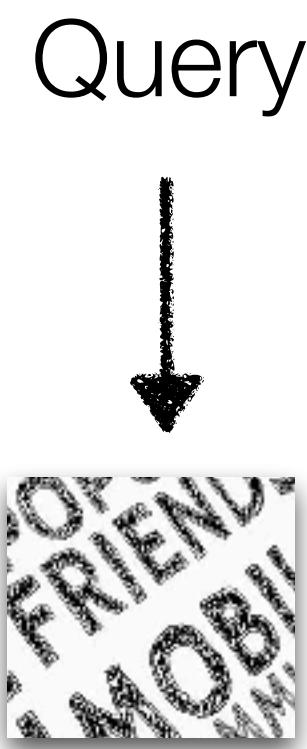


Text chunks (passages)?

Tokens?

Categorizing Retrieval-Augmented LMs

What to retrieve?



Text chunks (passages)?

Tokens?

Something else?

Categorizing Retrieval-Augmented LMs

What to retrieve?



Text chunks (passages)?

Tokens?

Something else?

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

```
graph LR; A[The] --> B[capital]; B --> C[city]; C --> D[of]; D --> E[Ontario]; E --> F[is]; F --> G[Toronto]
```

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

w/ retrieval w/ r w/r w/r w/ r w/r w/r

The capital city of Ontario is Toronto.

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

w/ retrieval w/ r w/r w/r w/ r w/r w/r

The capital city of Ontario is Toronto.

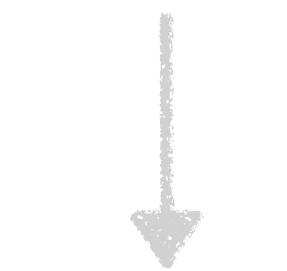
w/ retrieval

The capital city of Ontario is Toronto.

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

w/ retrieval w/ r w/r w/r w/r w/r w/r

The capital city of Ontario is Toronto.

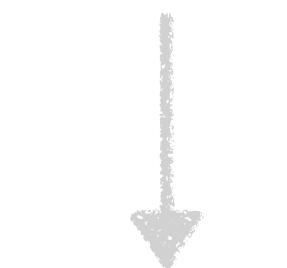
w/ retrieval

The capital city of Ontario is Toronto.

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

w/ retrieval w/ r w/r w/r w/r w/r w/r

The capital city of Ontario is Toronto.

w/ retrieval

The capital city of Ontario is Toronto.

How to use retrieval?

Input

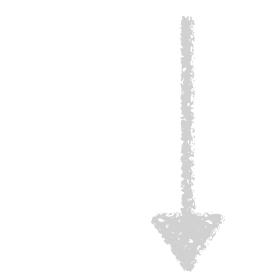
LM

Output

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

w/ retrieval w/ r w/r w/r w/r w/r w/r

The capital city of Ontario is Toronto.

w/ retrieval

The capital city of Ontario is Toronto.
w/r w/r

How to use retrieval?

Input



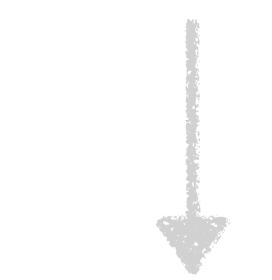
LM

Output

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

w/ retrieval w/ r w/r w/r w/r w/r w/r

The capital city of Ontario is Toronto.

w/ retrieval

The capital city of Ontario is Toronto.

How to use retrieval?

Input



LM

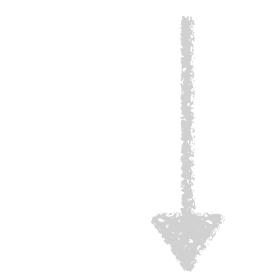


Output

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query



Text chunks (passages)?

Tokens?

Something else?

When to retrieve?

w/ retrieval

The capital city of Ontario is Toronto.

w/ retrieval w/ r w/r w/r w/r w/r w/r

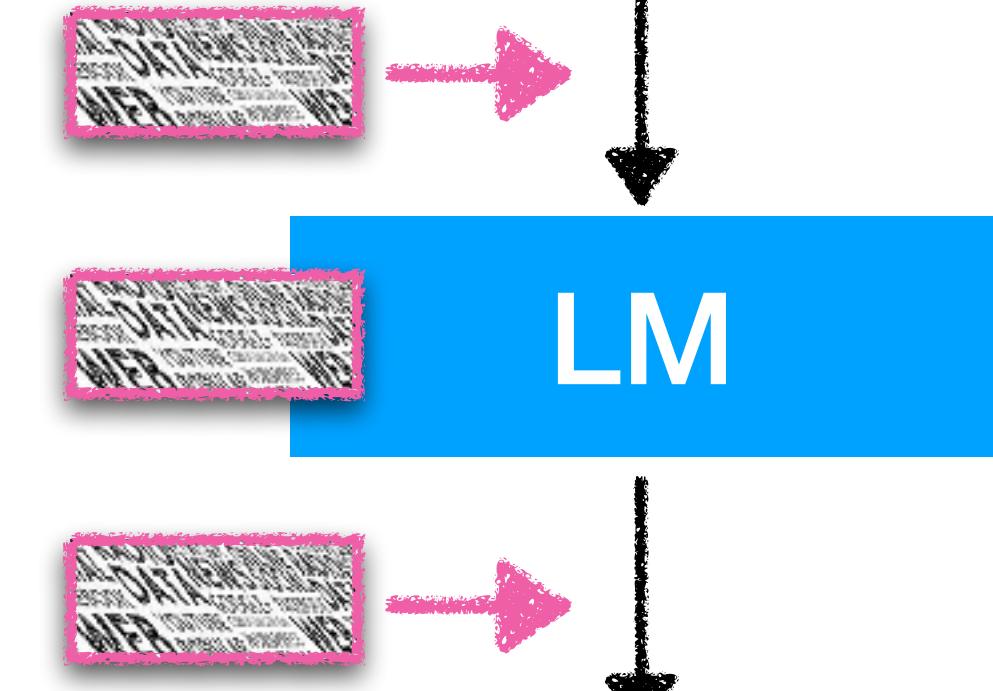
The capital city of Ontario is Toronto.

w/ retrieval

The capital city of Ontario is Toronto.

How to use retrieval?

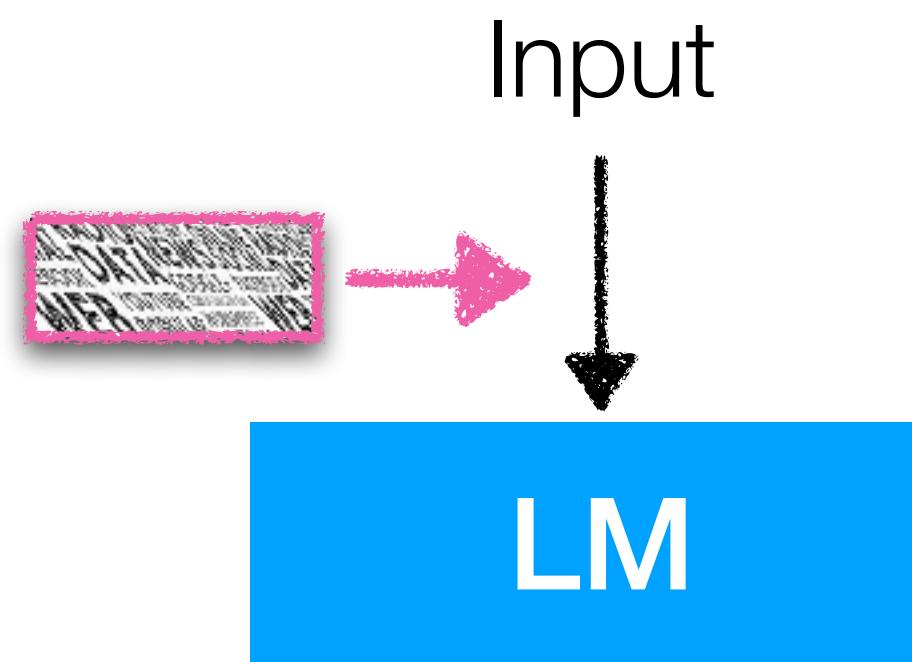
Input



Output

How to Use Retrieval

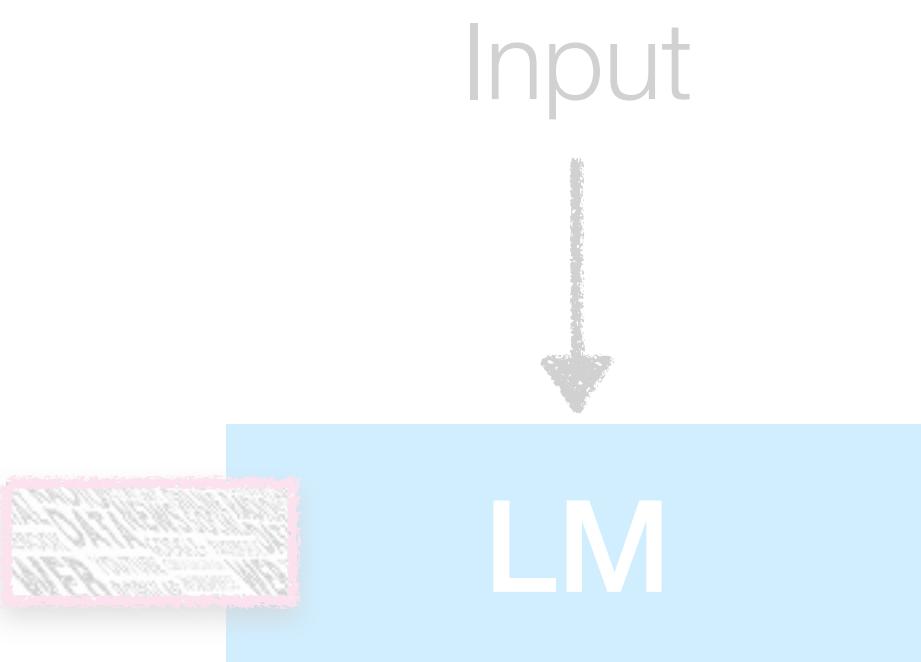
Input Augmentation



- Augment input of LMs
- Easy to apply (w/o training) & effective
- Difficulty of using many D

e.g., RAG

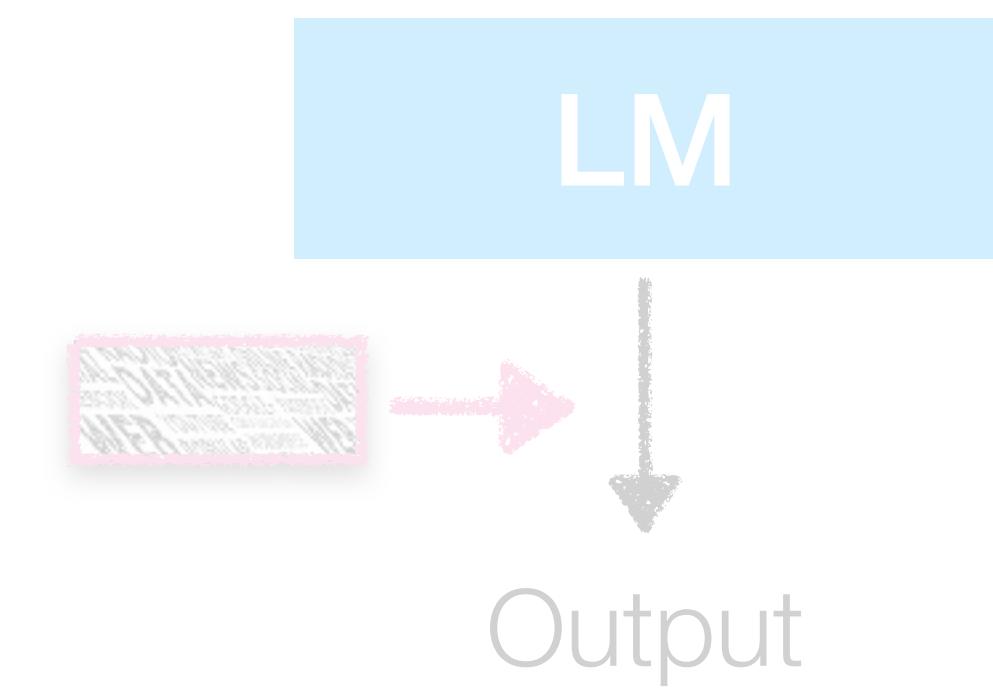
Intermediate Fusion



- Modify LMs to incorporate D in intermediate layers
- Scalable to many passages
- Requires retraining

e.g., RETRO, InstructRETRO

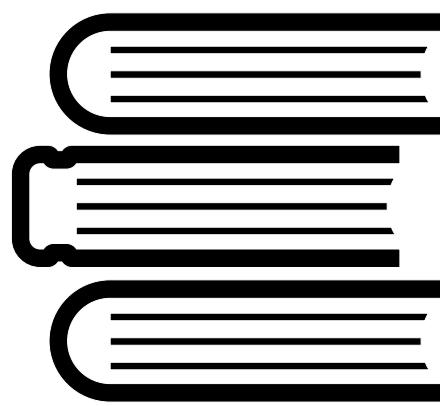
Output Interpolation



- Directly manipulate output token distributions
- No training required*
- Limited effectiveness on tasks

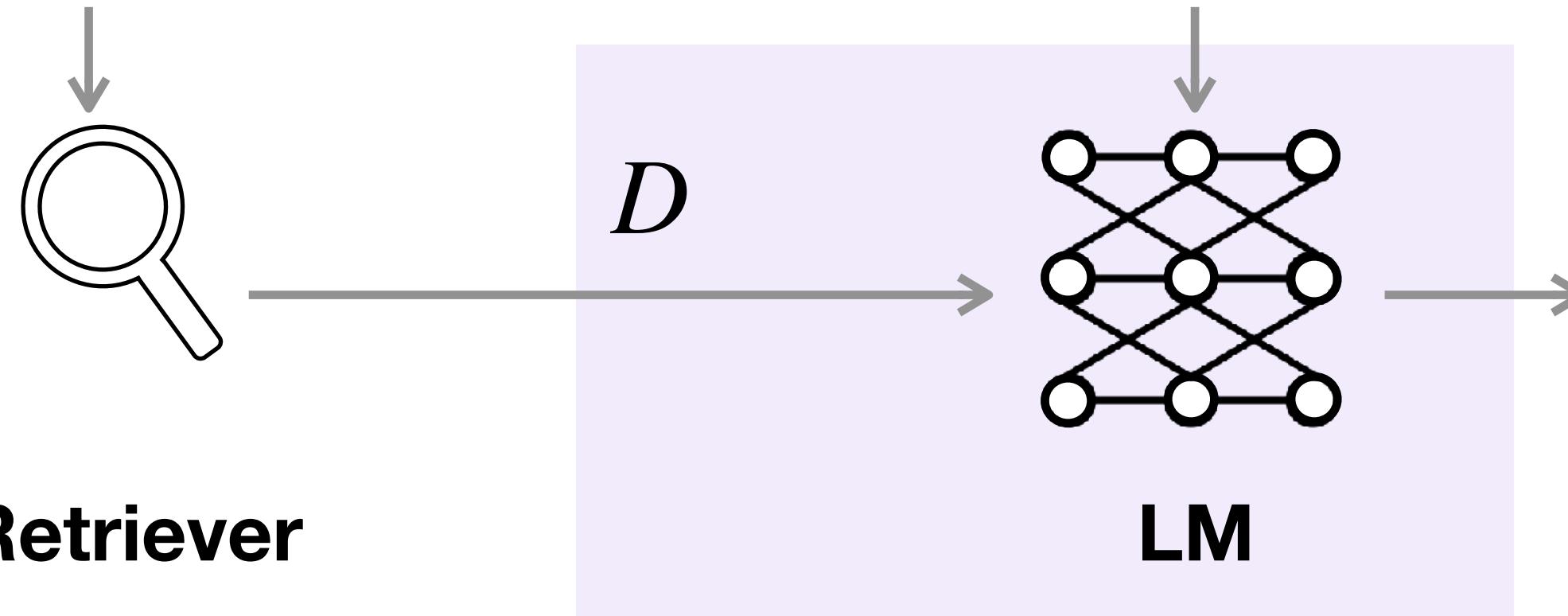
e.g., kNNLM

Retrieval-Augmented Generation



Datastore

x : Which company developed GPT4o?



$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

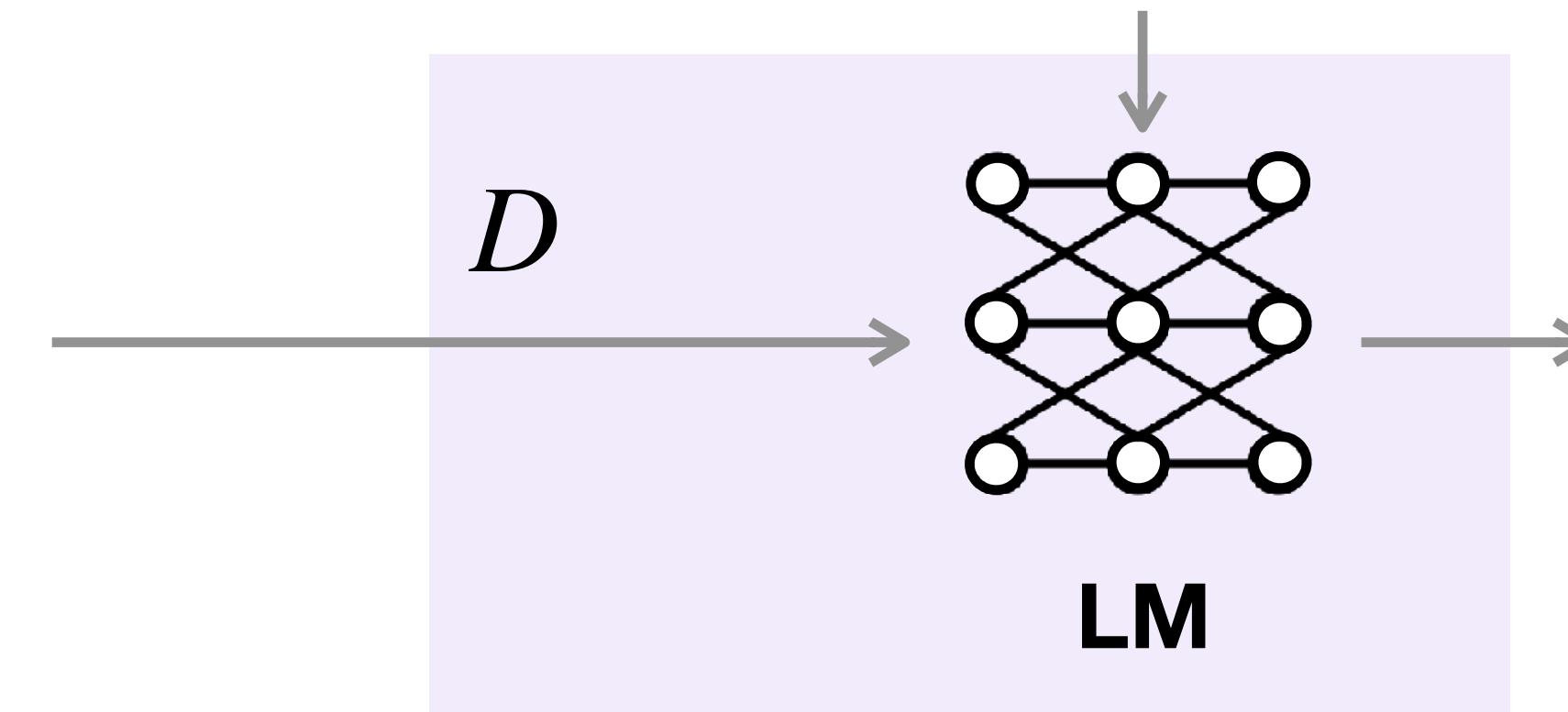
Transformers is a series of science fiction action films based on the Transformers franchise.

0.1

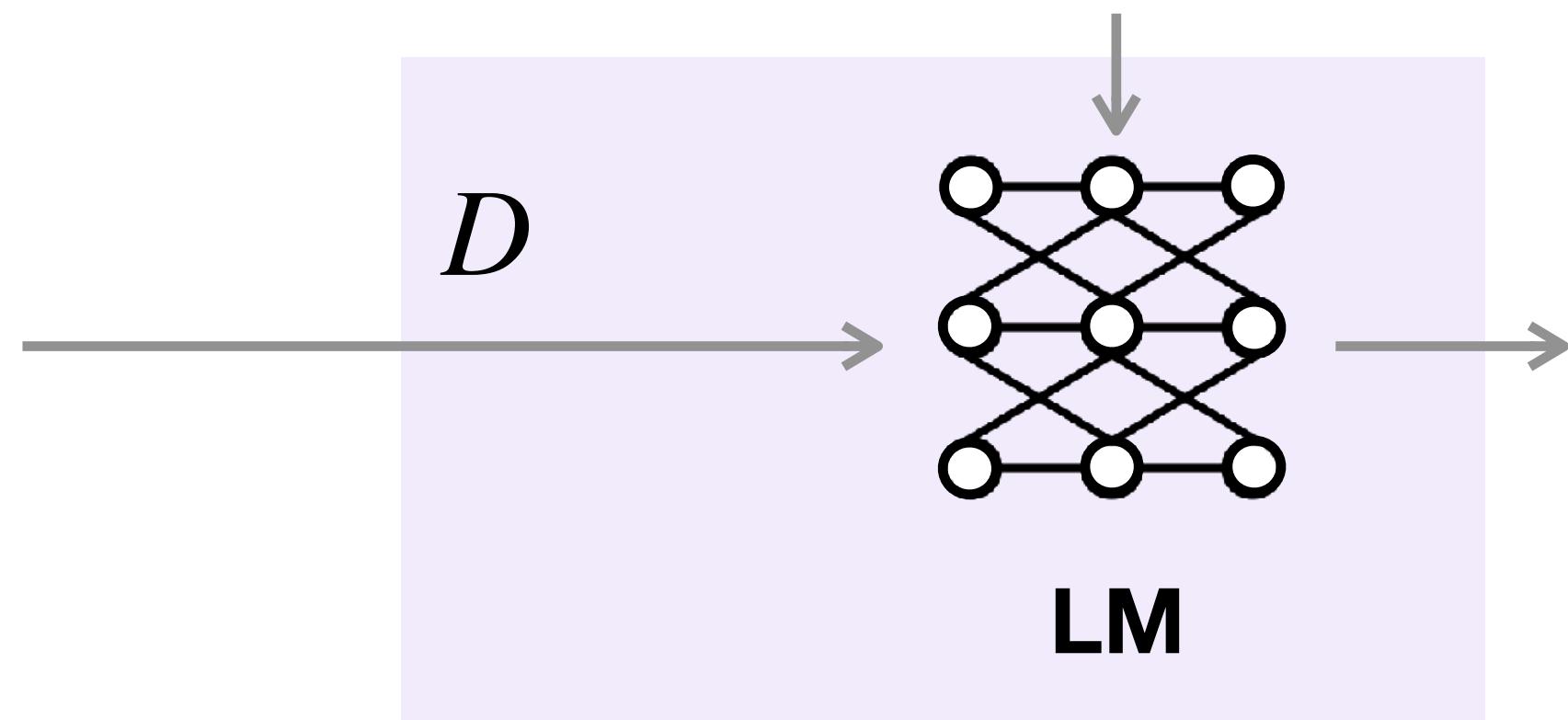
GPT4o was released by OpenAI in May 2024.

0.8

Retrieval-Augmented Generation

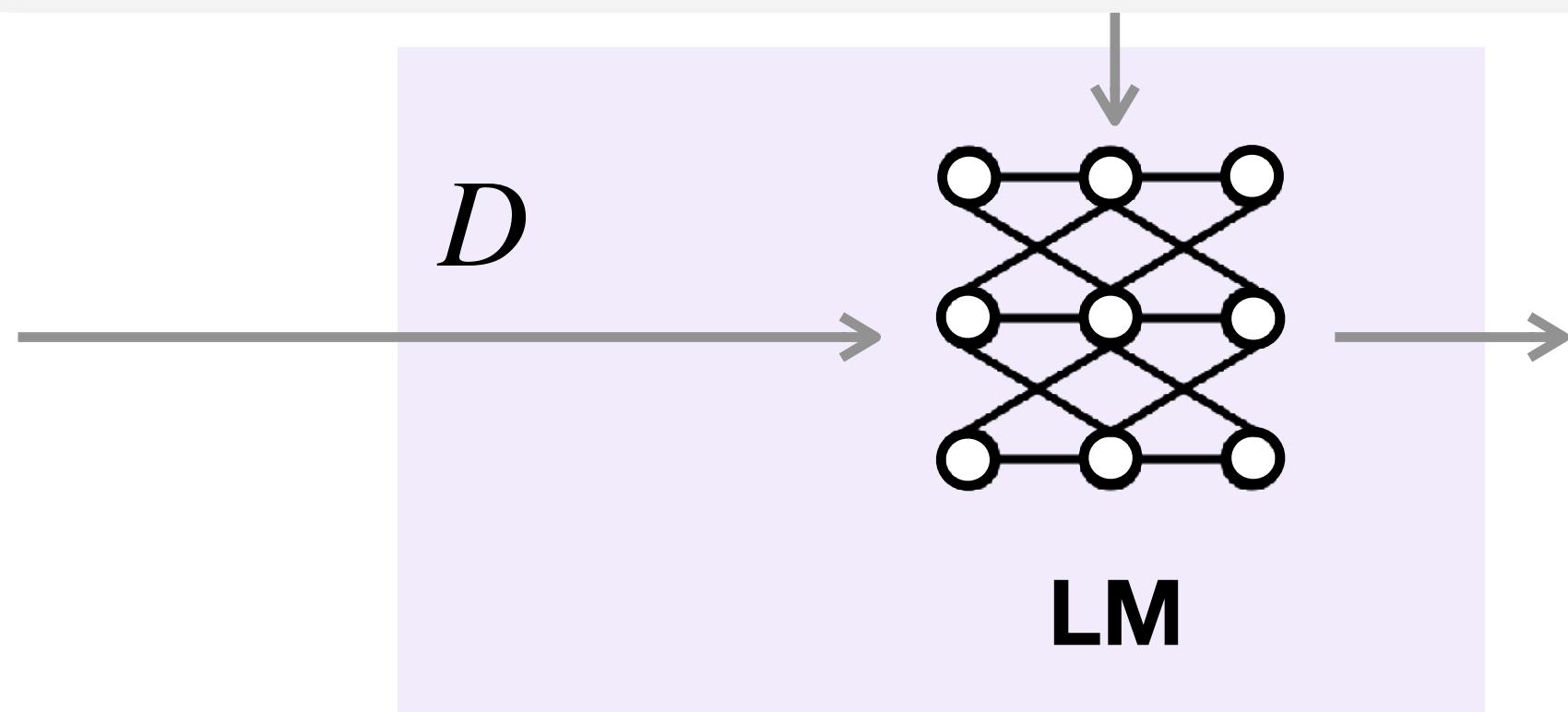


Retrieval-Augmented Generation



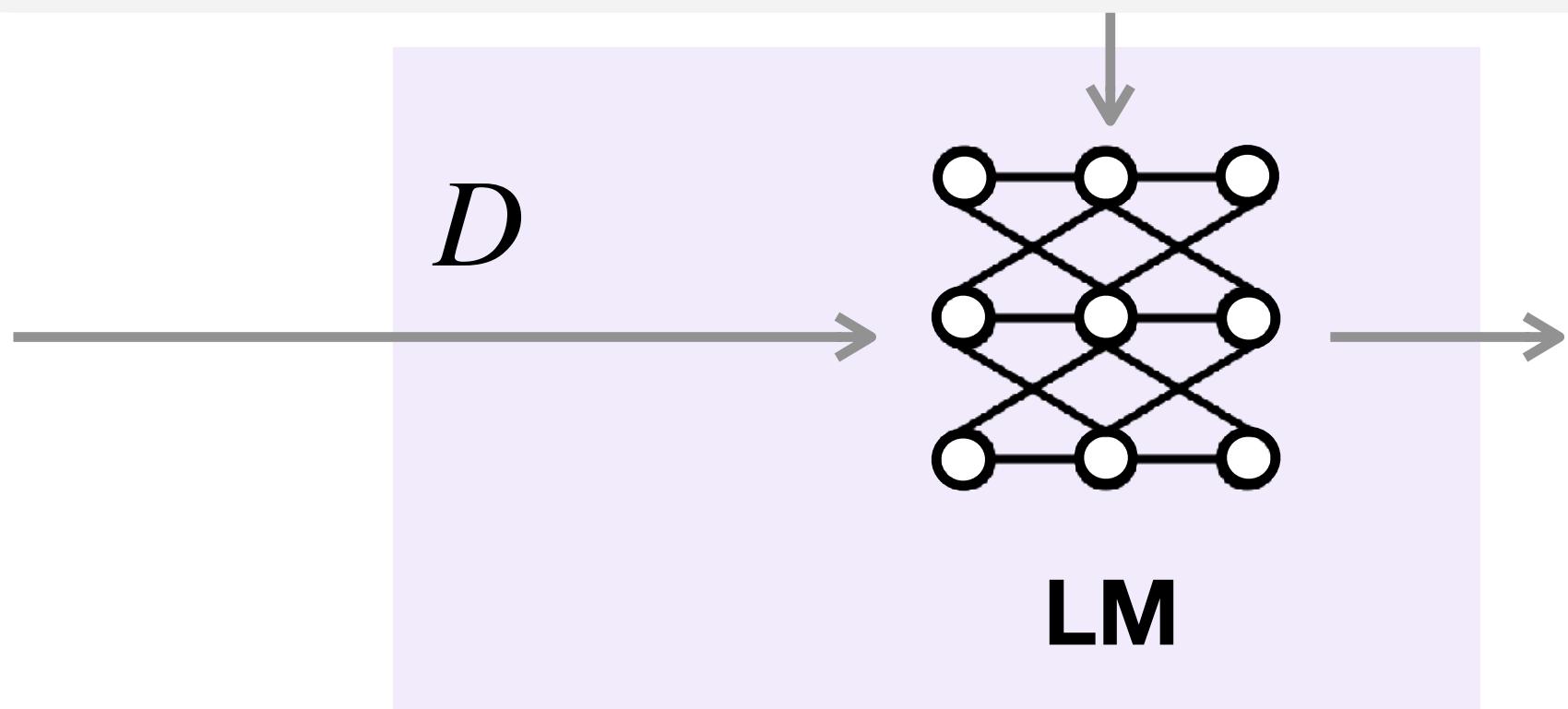
Retrieval-Augmented Generation

x : Which company developed GPT4o?



Retrieval-Augmented Generation

x : Which company developed GPT4o?

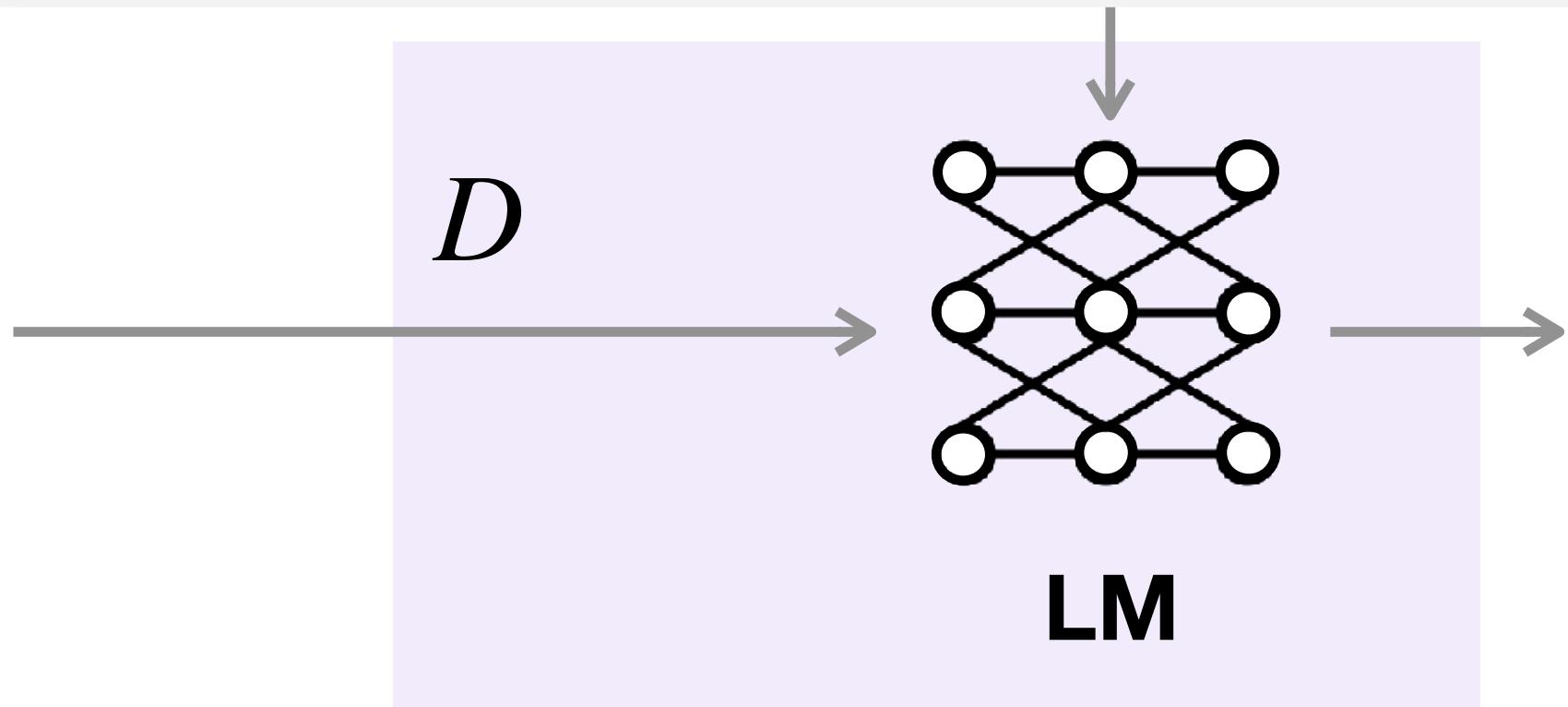


GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented Generation

x : Which company developed GPT4o?



Question: Which company developed GPT4o?

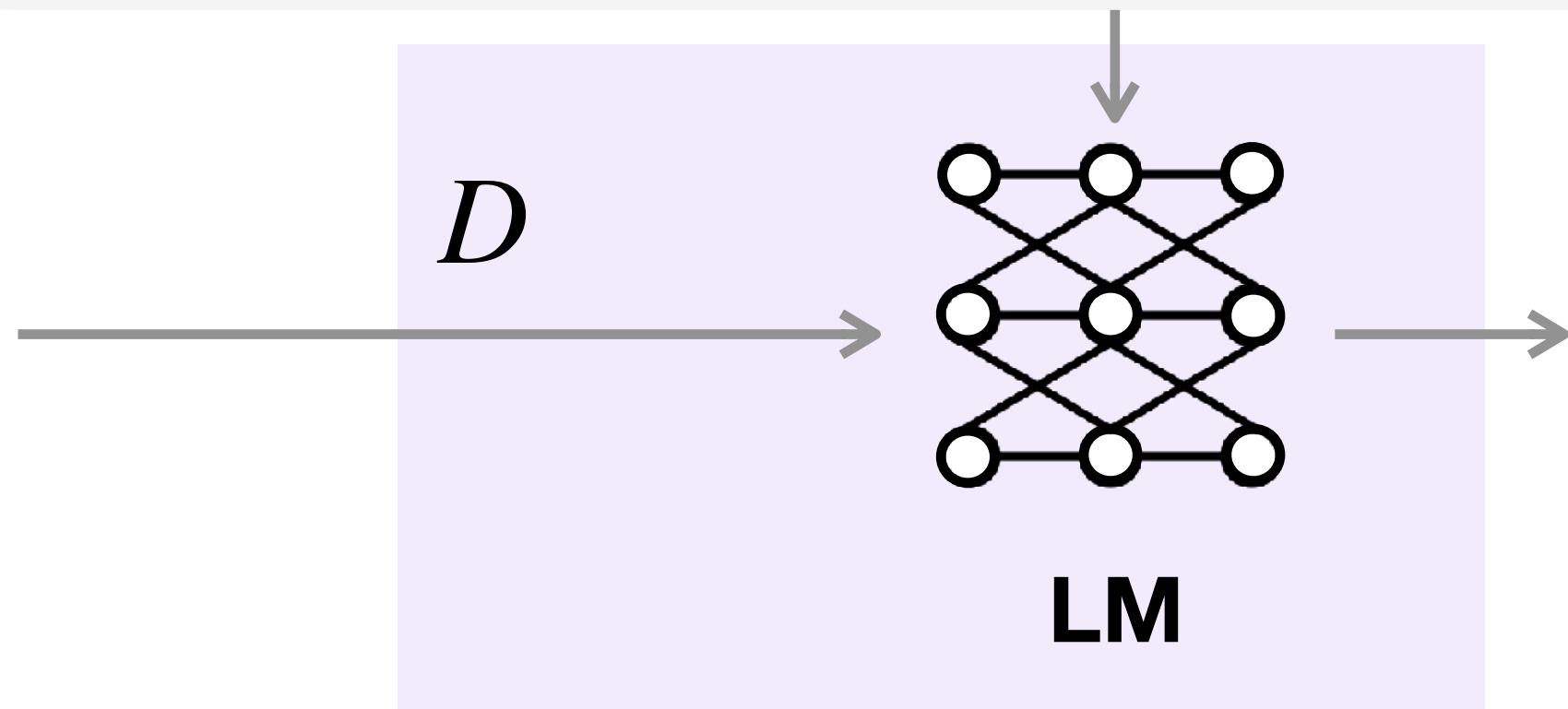
References:

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented Generation

x : Which company developed GPT4o?



Question: Which company developed GPT4o?

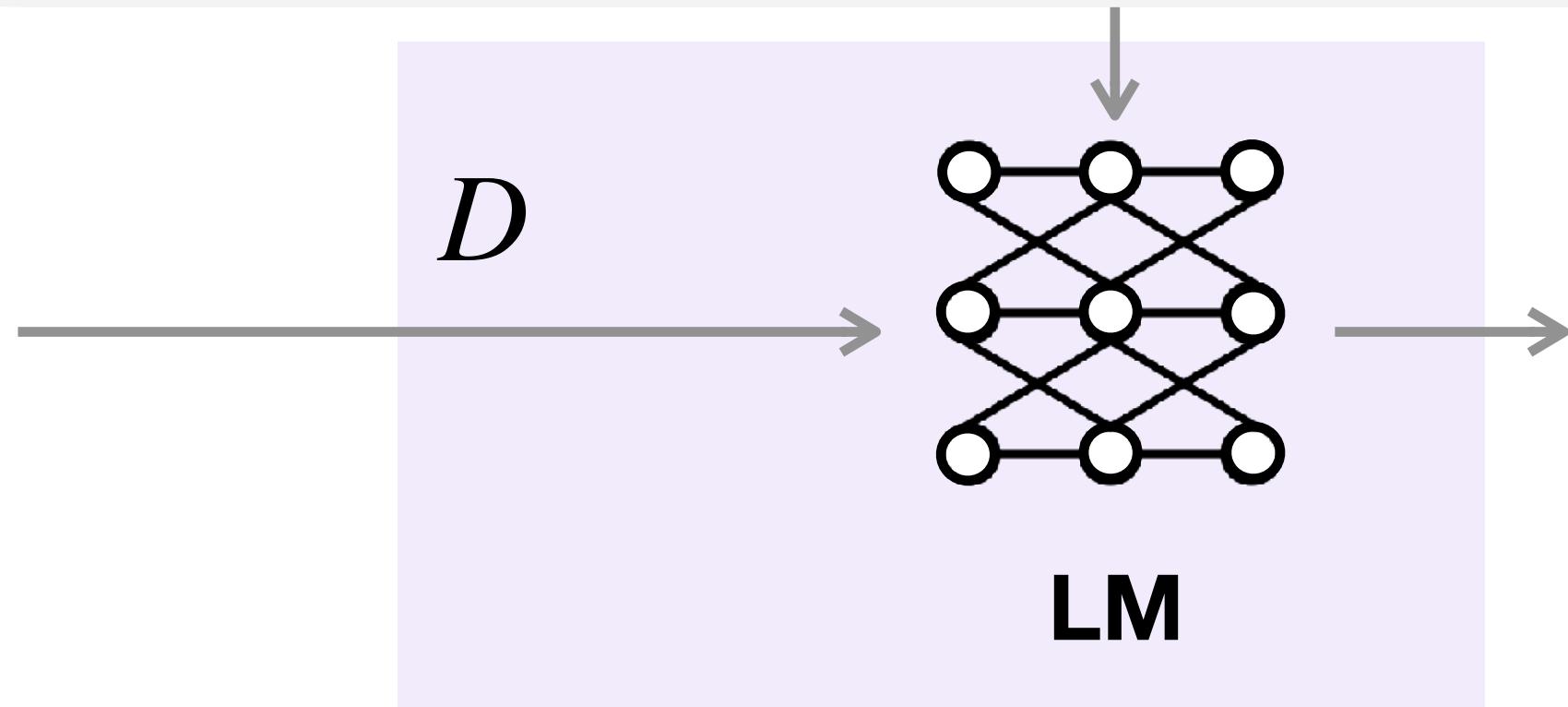
References:

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented Generation

x : Which company developed GPT4o?



Question: Which company developed GPT4o?

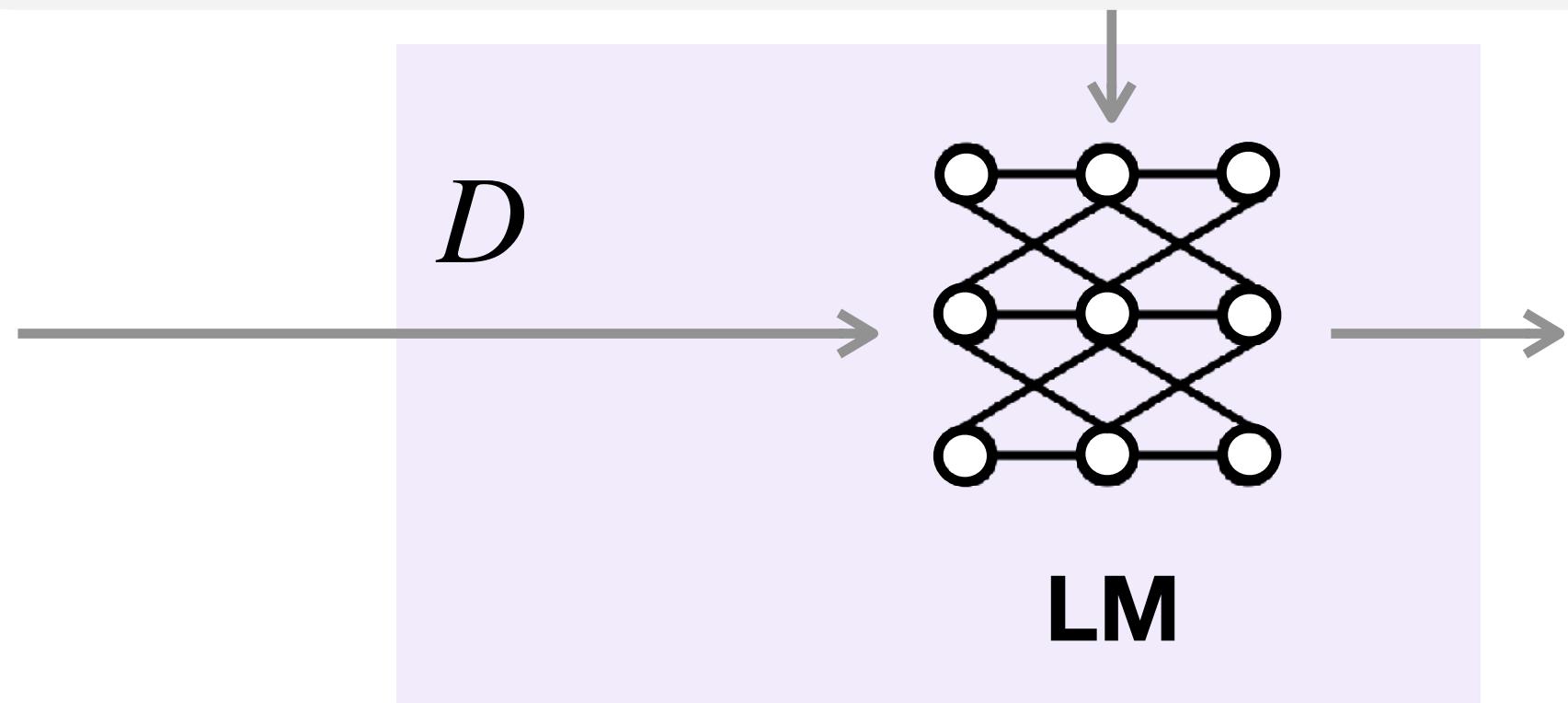
References:

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented Generation

x : Which company developed GPT4o?

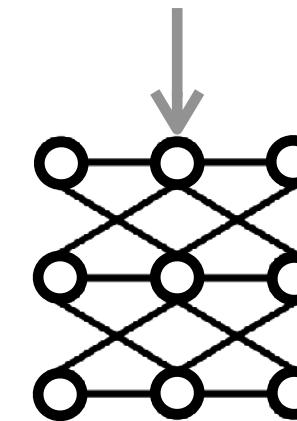


Question: Which company developed GPT4o?

References:

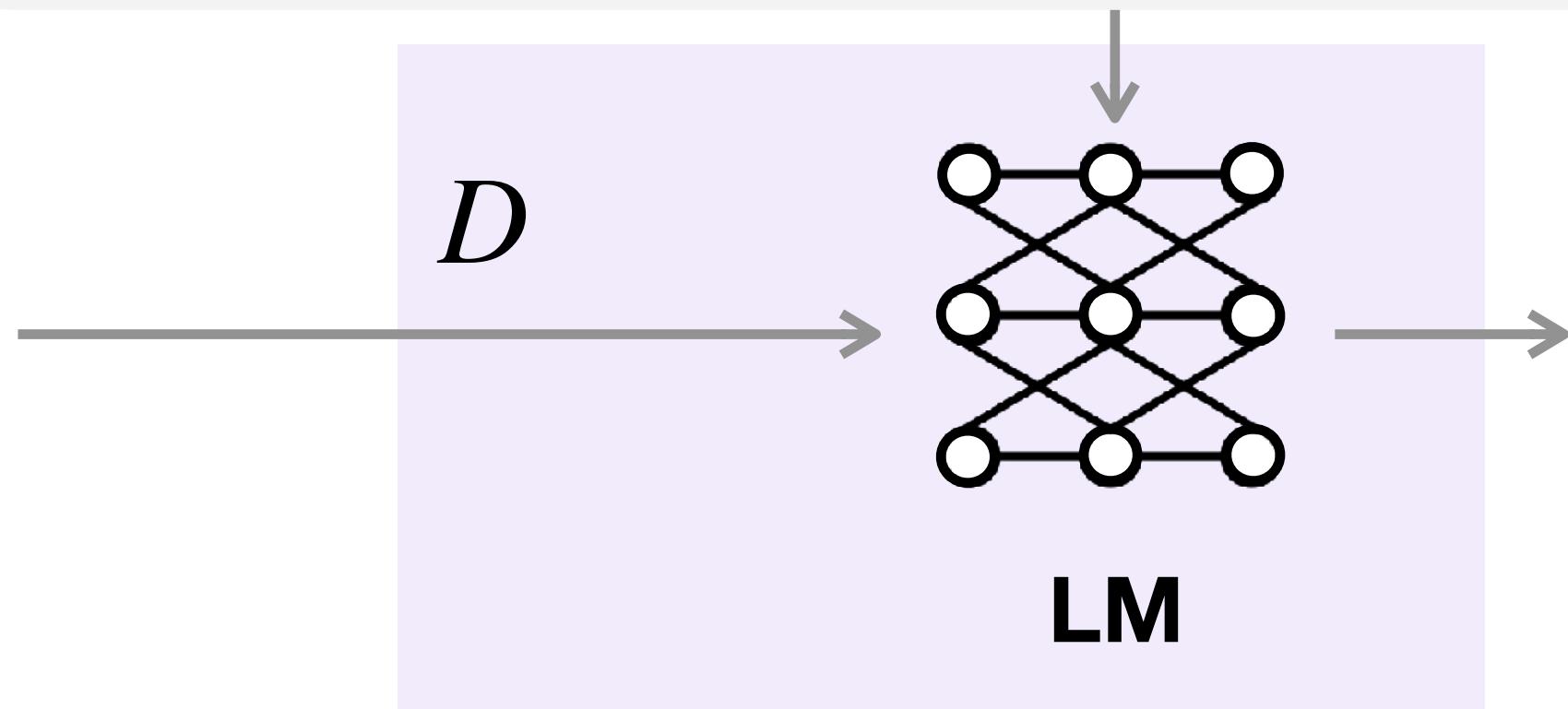
GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.



Retrieval-Augmented Generation

x : Which company developed GPT4o?

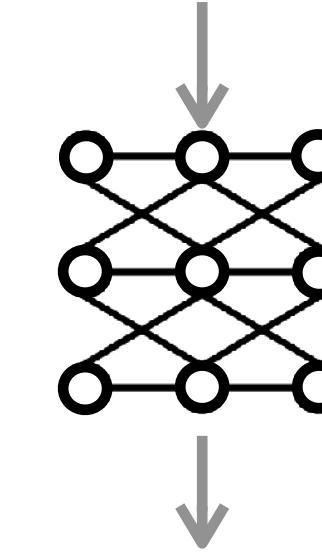


Question: Which company developed GPT4o?

References:

GPT-4o is a pre-trained transformer developed by OpenAI.

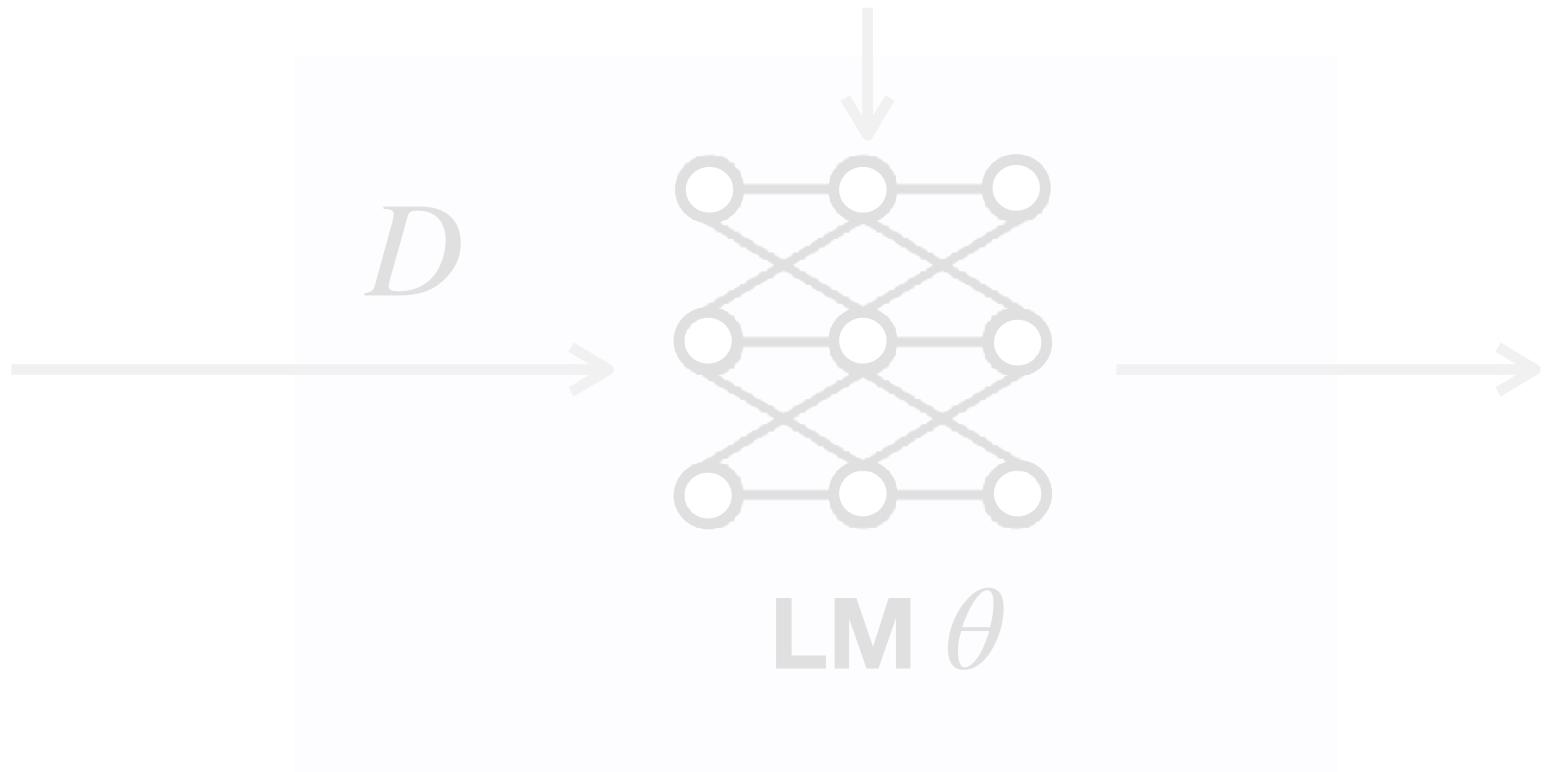
GPT4o was released by OpenAI in May 2024.



y : OpenAI

Retrieval-Augmented Generation

x : Which company developed GPT4o?

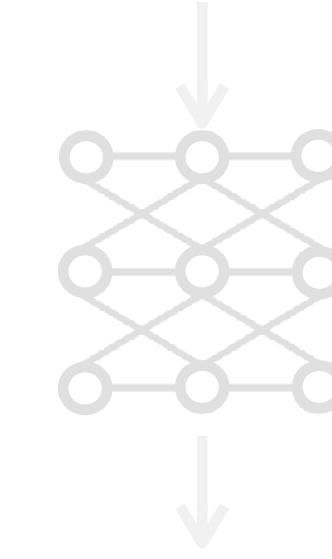


Question: Who developed GPT4o?

References:

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o is released by OpenAI in May 2024.



y : OpenAI

Retrieval-Augmented Generation

x : Which company developed GPT4o?



Question: Who developed GPT4o?

References:

Previous work **fine-tuned LMs** for specific downstream tasks

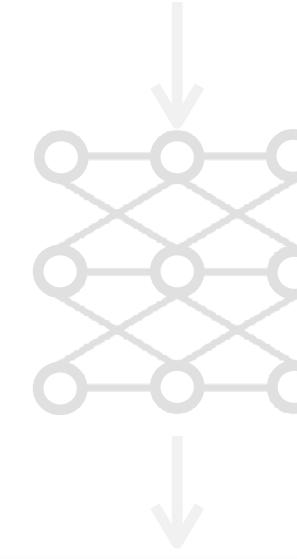
(Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020)

LM θ

D

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o is released by OpenAI in May 2024.



y : OpenAI

Training RAG

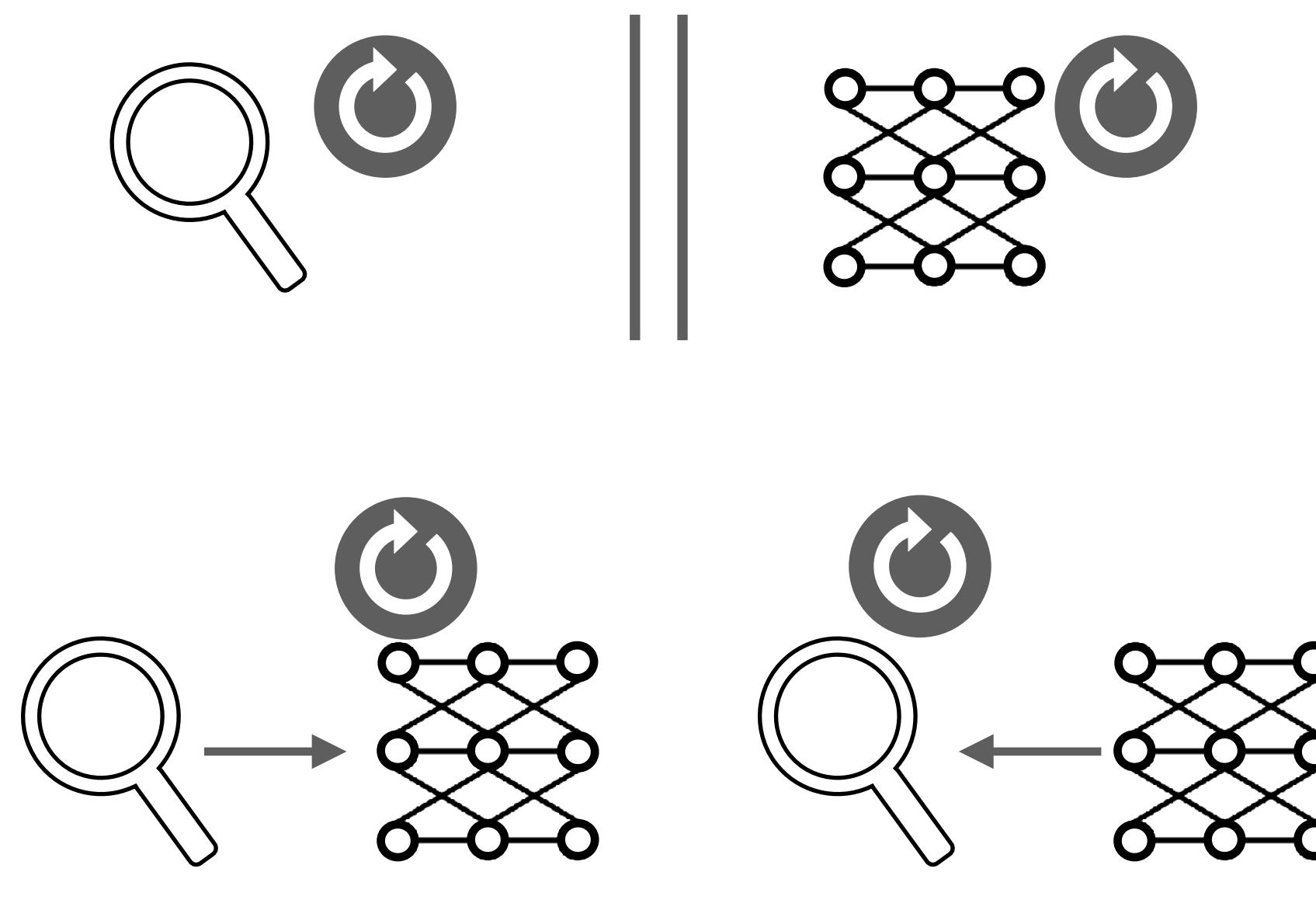
Training RAG



Independent training

- DPR (Karpukhin et al., 2020)
- DRQA (Chen et al., 2017)

Training RAG



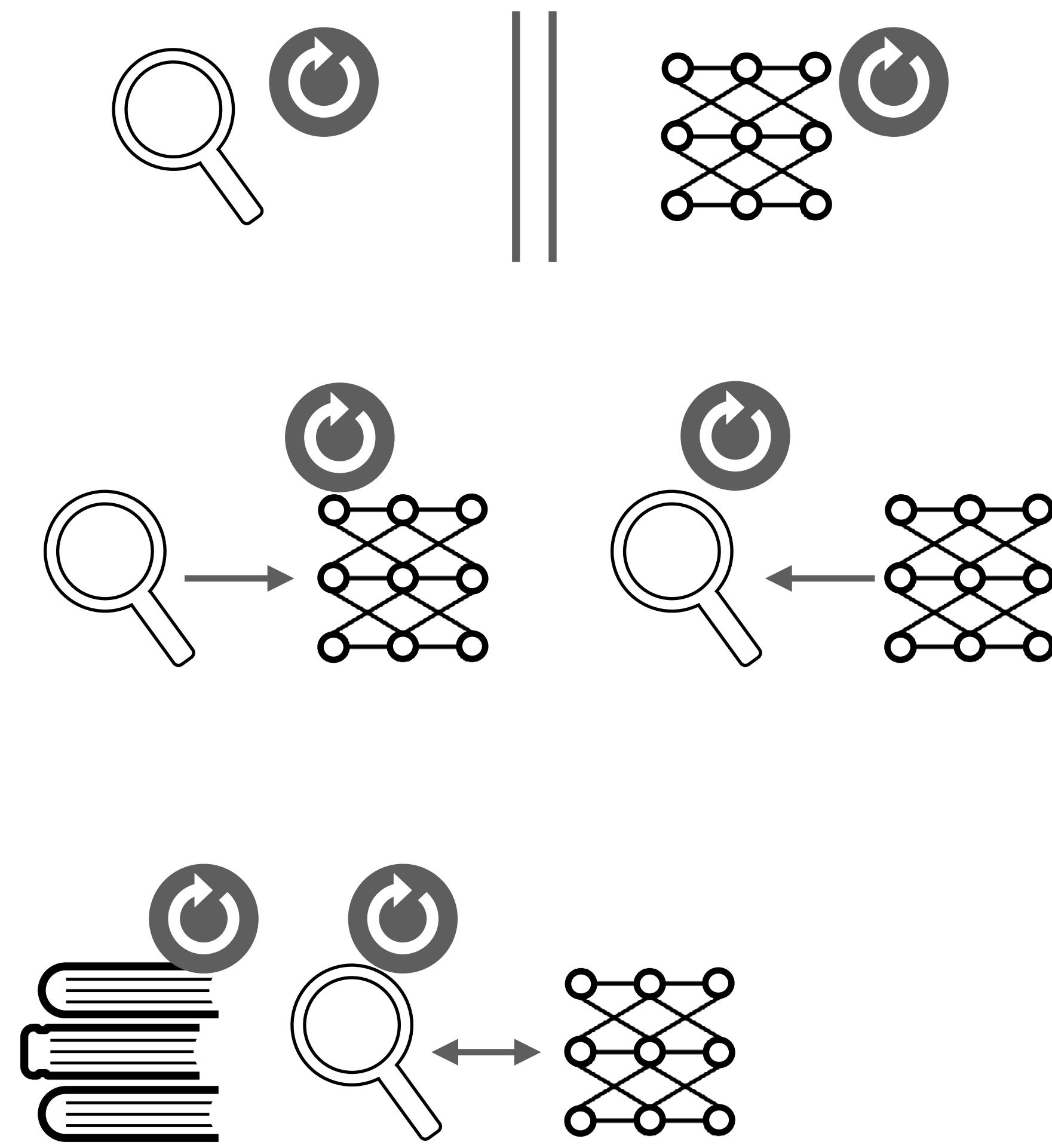
Independent training

- DPR (Karpukhin et al., 2020)
- DRQA (Chen et al., 2017)

Sequential training

- Evidentiality Generator (Asai et al., 2023)
- REPLUG (Shi et al., 2023)

Training RAG



Independent training

- DPR (Karpukhin et al., 2020)
- DRQA (Chen et al., 2017)

Sequential training

- Evidentiality Generator (Asai et al., 2023)
- REPLUG (Shi et al., 2023)

Joint training (async update)

- RAG (Lewis et al., 2021)
- REALM (Guu et al., 2021)

Retrieval-Augmented Generation

x : Which company developed GPT4o?



Question: Who developed GPT4o?

References:

Previous work **fine-tuned LMs** for specific downstream tasks

(Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020)

LM θ

Now people often **combine retrieval with off-the-shelf LMs**

GPT-4o is a pre-trained transformer
developed by OpenAI.

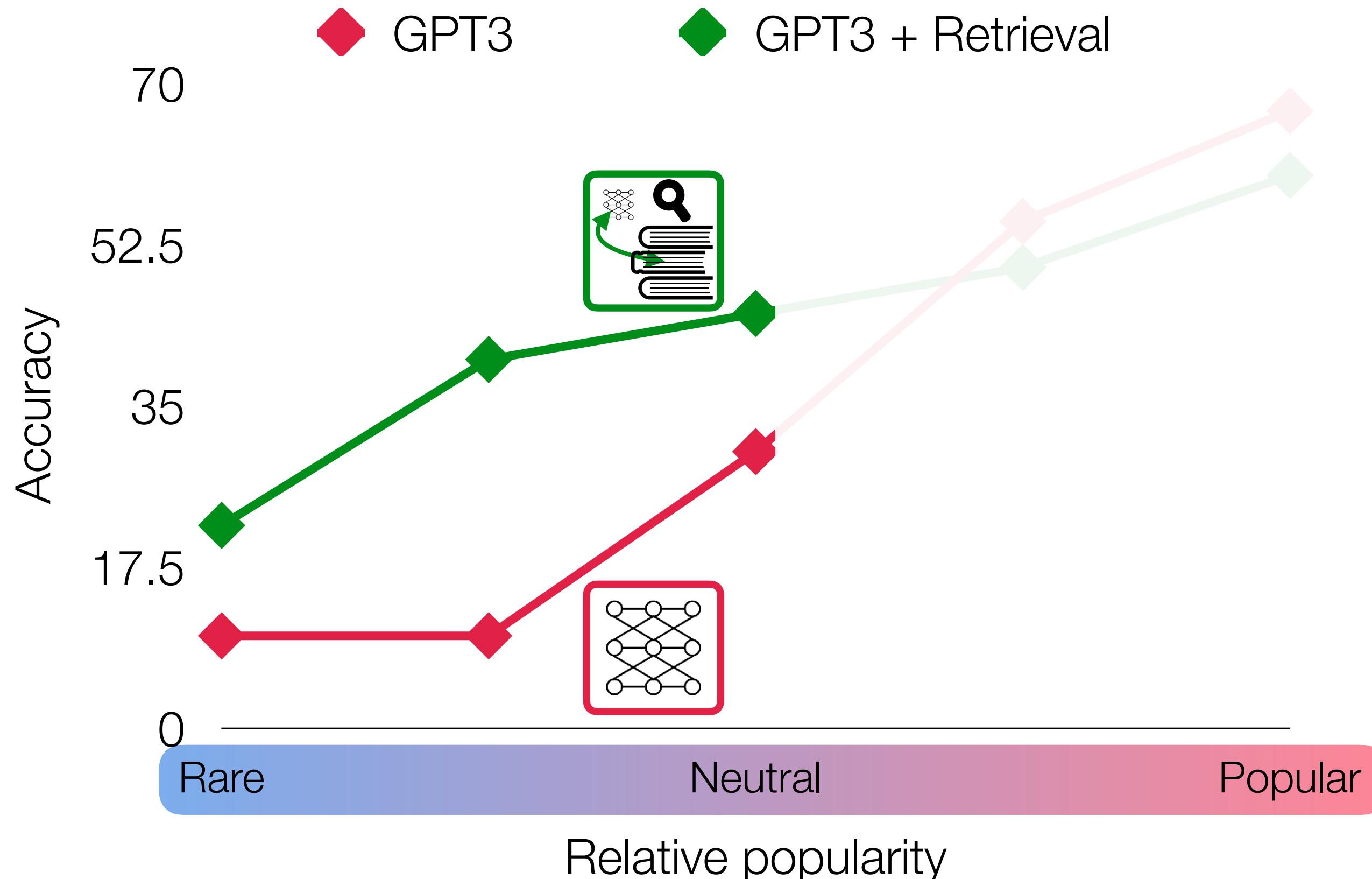
GPT4o is released by OpenAI in May 2024.



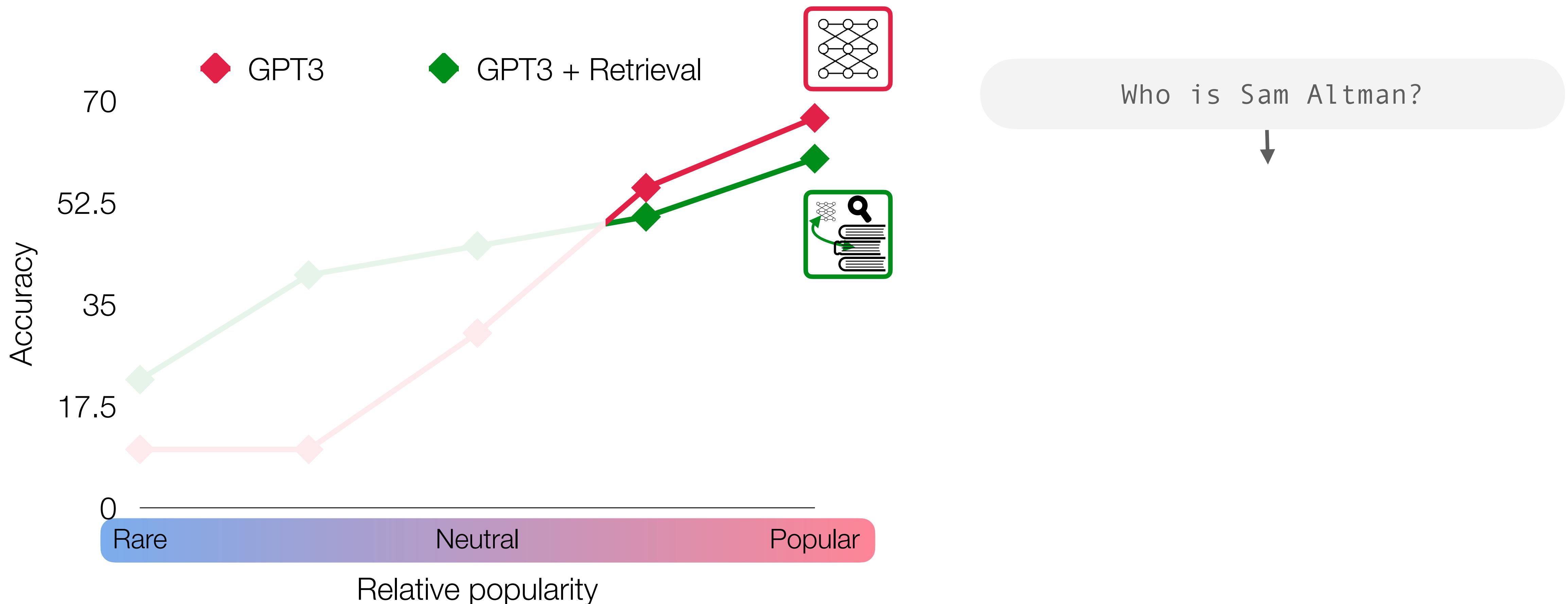
Widely referred to as **RAG**

Effectiveness of Simple RAG

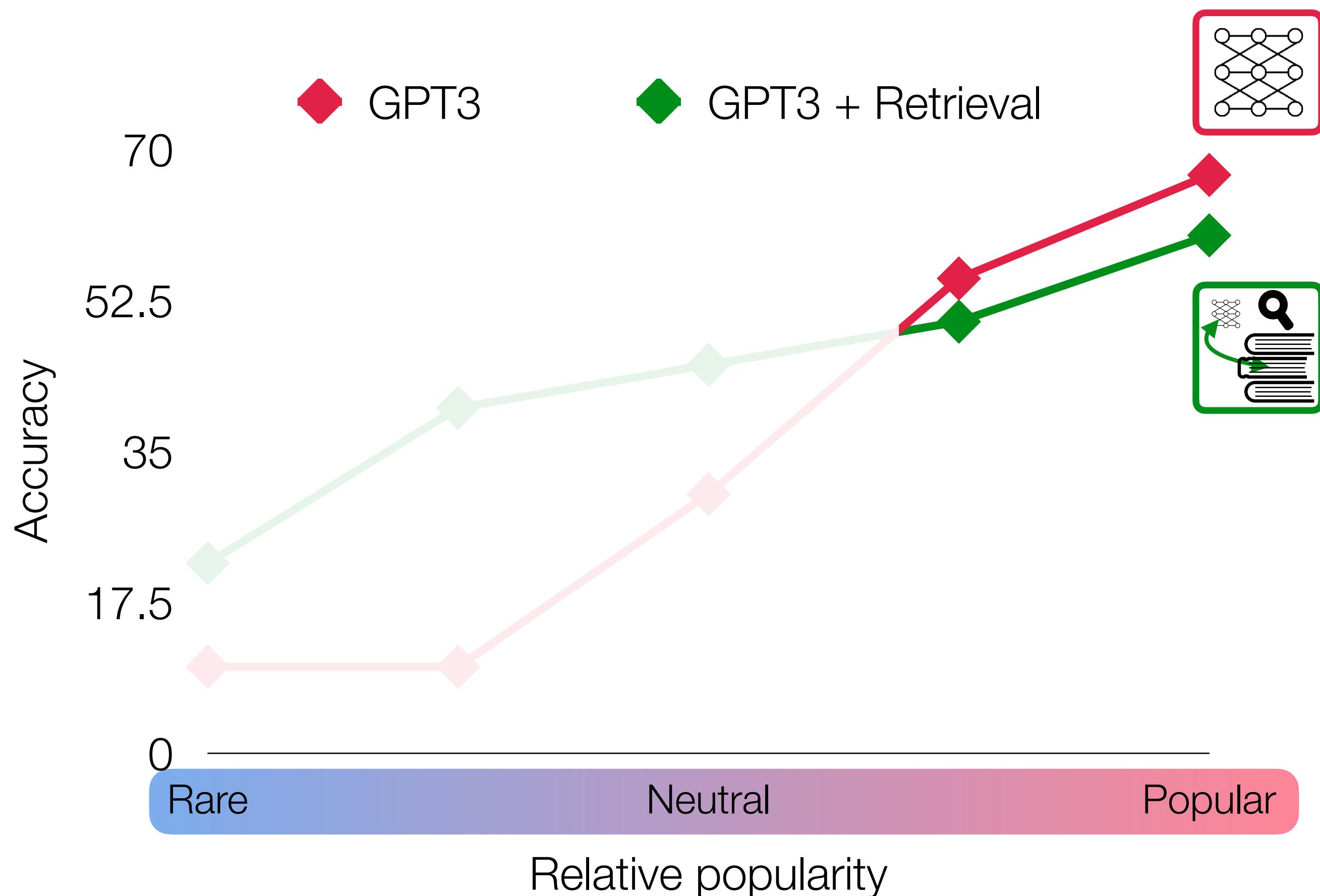
RAG constantly gives performance improvements esp. in long-tail



Limitations of Simple RAG



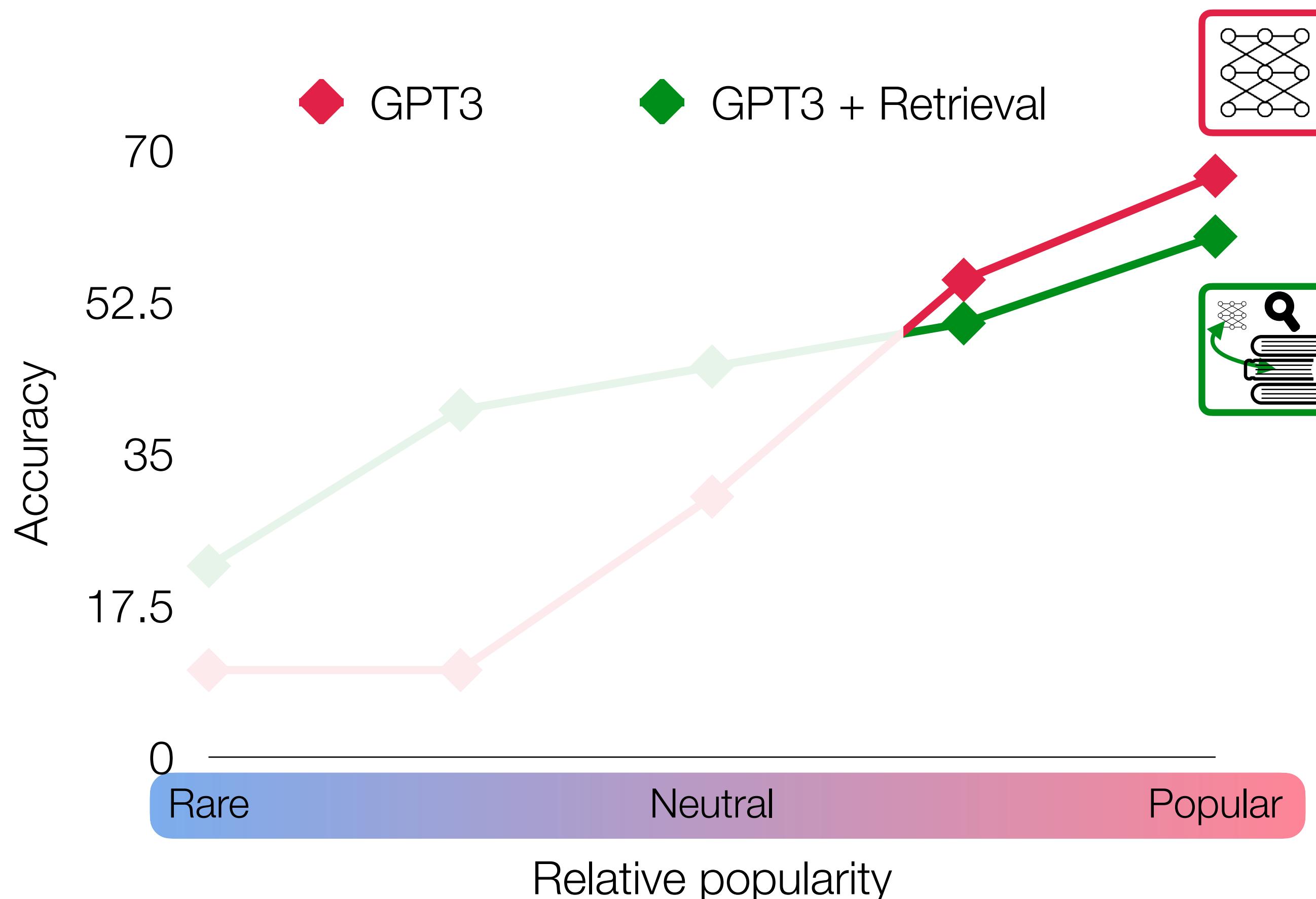
Limitations of Simple RAG



Who is Sam Altman?

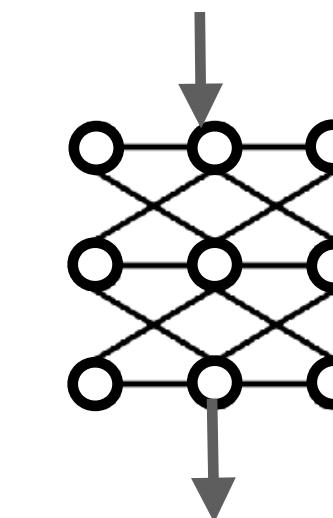
Samuel Adams (Sam Adams) was an American statesman, political philosopher and a founding father of US.

Limitations of Simple RAG



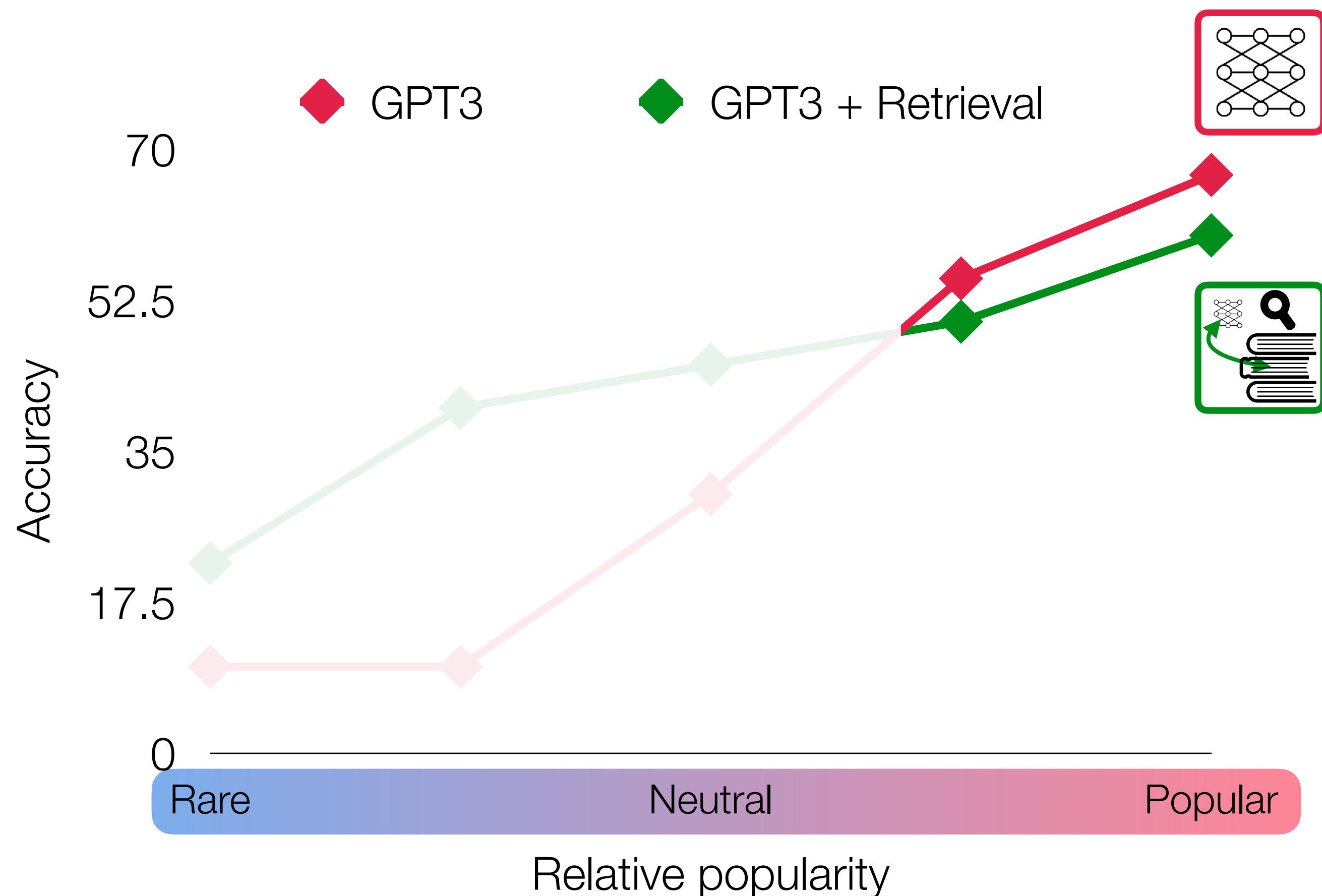
Who is Sam Altman?

Samuel Adams (Sam Adams) was an American statesman, political philosopher and a founding father of US.



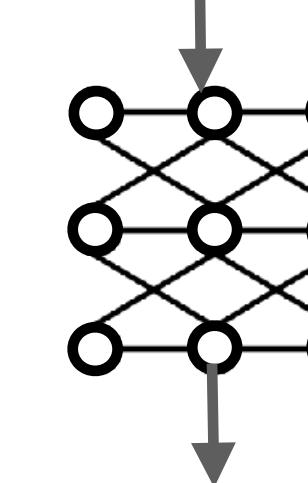
He is a founding father of the US

Limitations of Simple RAG



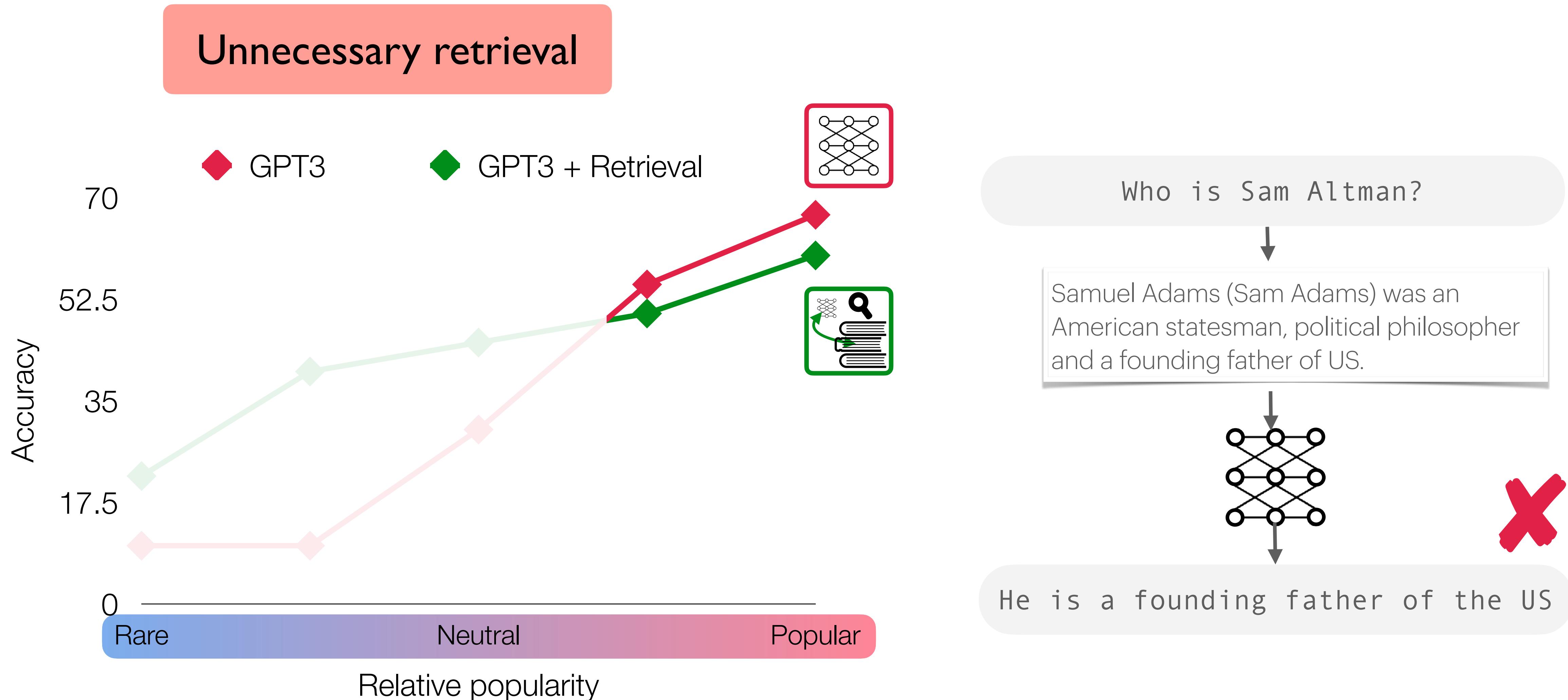
Who is Sam Altman?

Samuel Adams (Sam Adams) was an American statesman, political philosopher and a founding father of US.

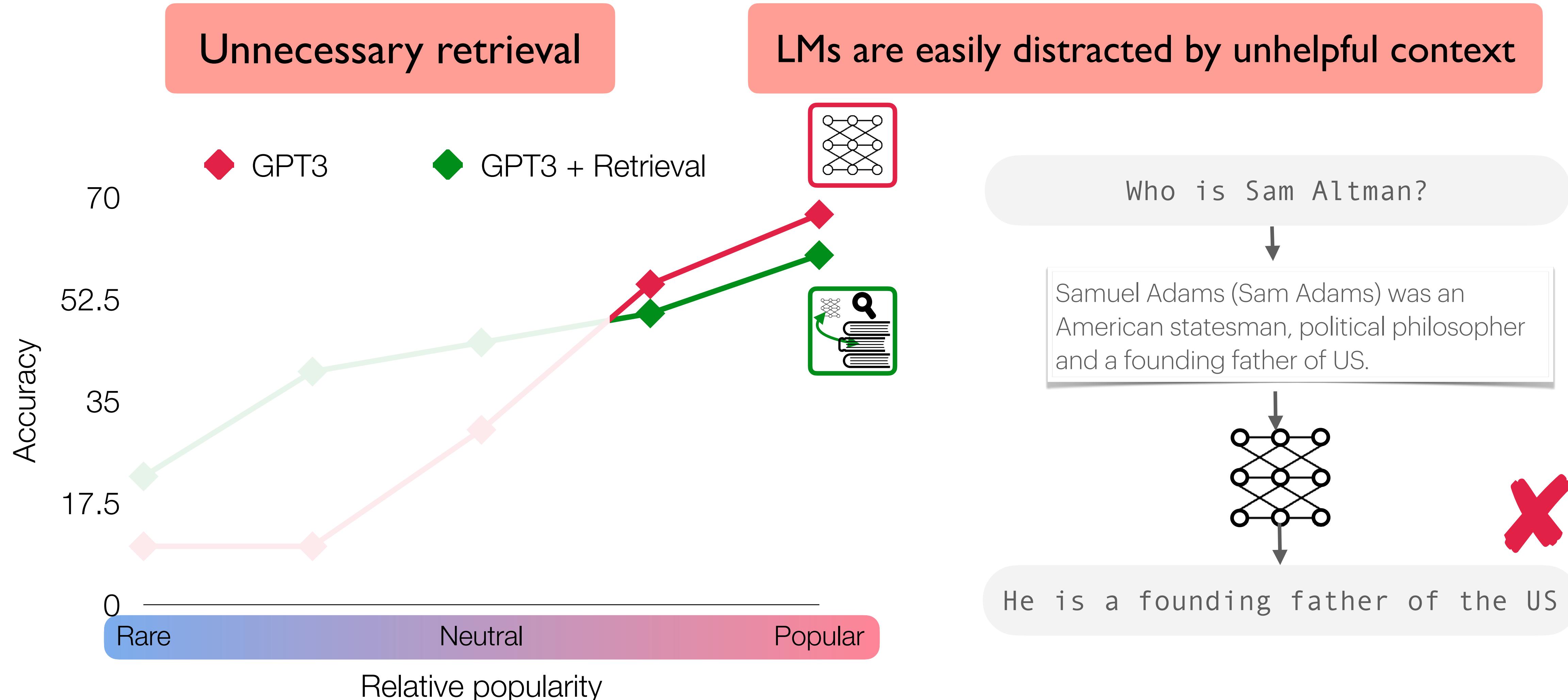


He is a founding father of the US

Limitations of Simple RAG



Limitations of Simple RAG



Limitations of Simple RAG

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

Jin et al. RAGCache: Efficient Knowledge Caching For Retrieval-Augmented Generation

Limitations of Simple RAG

 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

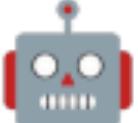
[1]:  nasa.gov (Xcitation does not support its associated statement)
[NASA's Webb Confirms Its First Exoplanet](#)
... Researchers confirmed an exoplanet, a planet that orbits another star, using NASA's James Webb Space Telescope for the first time. ...

[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 ...

Outputs aren't often supported by citations

Limitations of Simple RAG

 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 ...

Outputs aren't often supported by citations

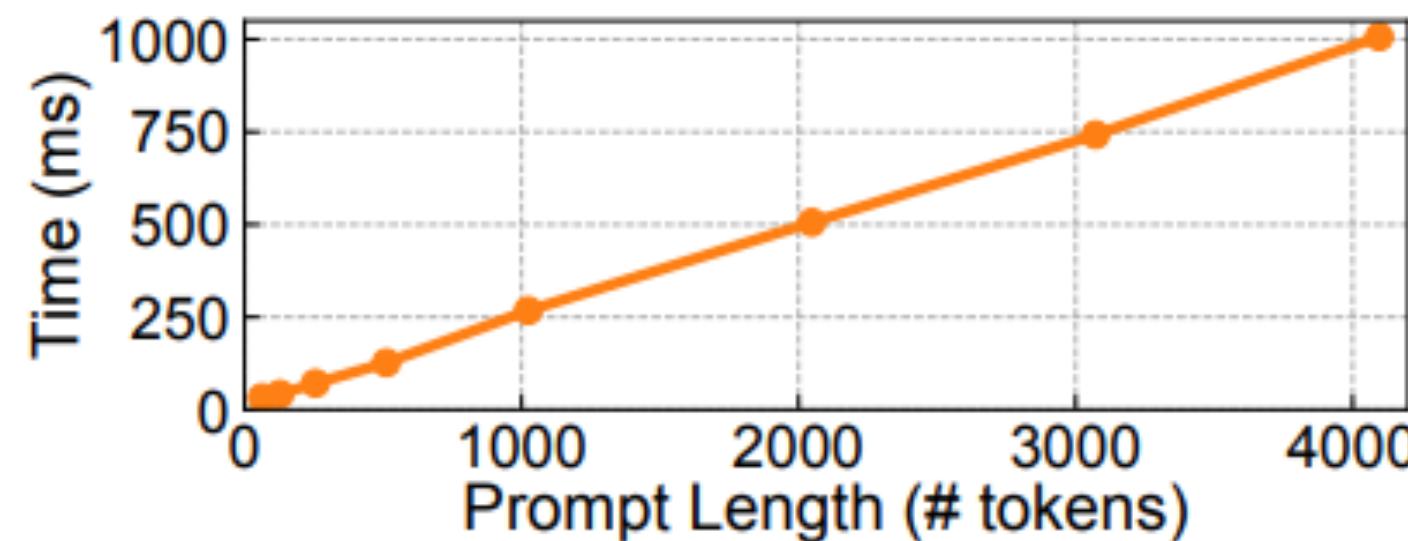
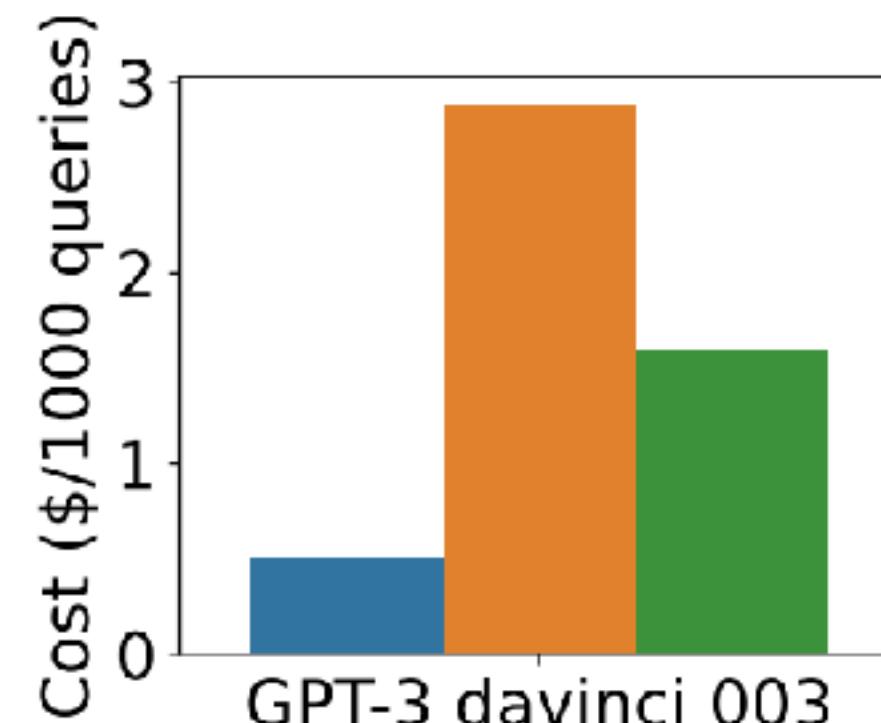


Figure 2. Inference time with different input lengths.

Vanilla RAG



Increased latency to encode much longer context

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

Jin et al. RAGCache: Efficient Knowledge Caching For Retrieval-Augmented Generation

Self-RAG: Learn to Retrieve and Critique

Retrieval-Augmented Generation (RAG)

Prompt How did US states get their names?

Step 1: Retrieve K documents

- 1 Of the fifty states, eleven are named after an individual person.
- 2 Popular names by states. In Texas, Emma is a popular baby name.
- 3 California was named after a fictional island in a Spanish book.

Retriever

Step 2: Prompt LM with K docs and generate

Prompt How did US states get their names? + ①②③

Self-evaluation

LM Individual person (e.g., California was named after Christopher Columbus). Some states including Texas and Utah, are named after American tribe. No information in passages

Prompt: Write an essay of your best summer vacation

① ② ③ → My best...

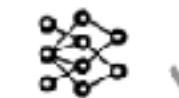
Ours: Self-reflective Retrieval-Augmented Generation

Prompt How did US states get their names?

US states got their names from a variety of sources. Retrieve

Step 2: Generate segment in parallel

Prompt + ①



Relevant 11 of 50 state names come from persons. Supported

Prompt + ②



Irrelevant Texas is named after a Native American tribe. Partially

Prompt + ③



Relevant California's name has its origins in a 16th-century novel Las Sergas de Esplandián. Partially

Step 3: Critique outputs and select best segment

① [green, green]

> ③ [green, orange]

> ② [red]

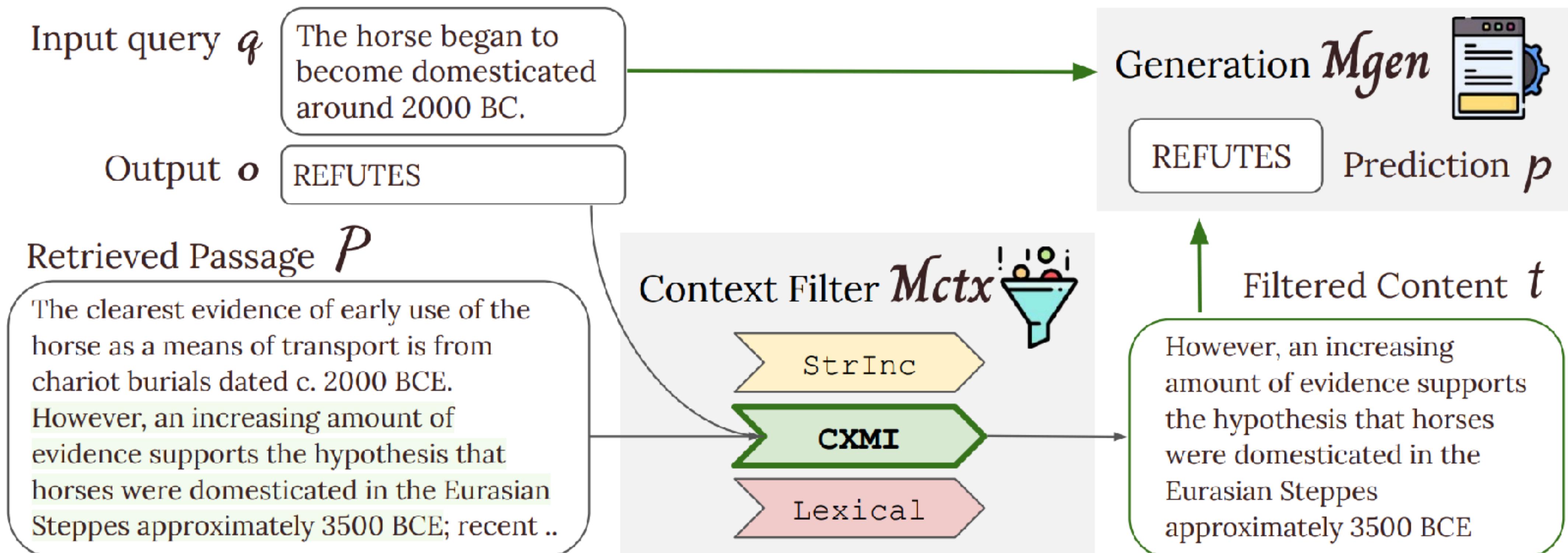
US states got their names from a variety of sources. 11 of 50 states names are come from persons. ① 26 states are named after Native Americans, including Utah. ④

Prompt: Write an essay of your best summer vacation

No Retrieval My best summer vacation is when my family and I embarked on a road trip along ...

Adaptive Retrieval

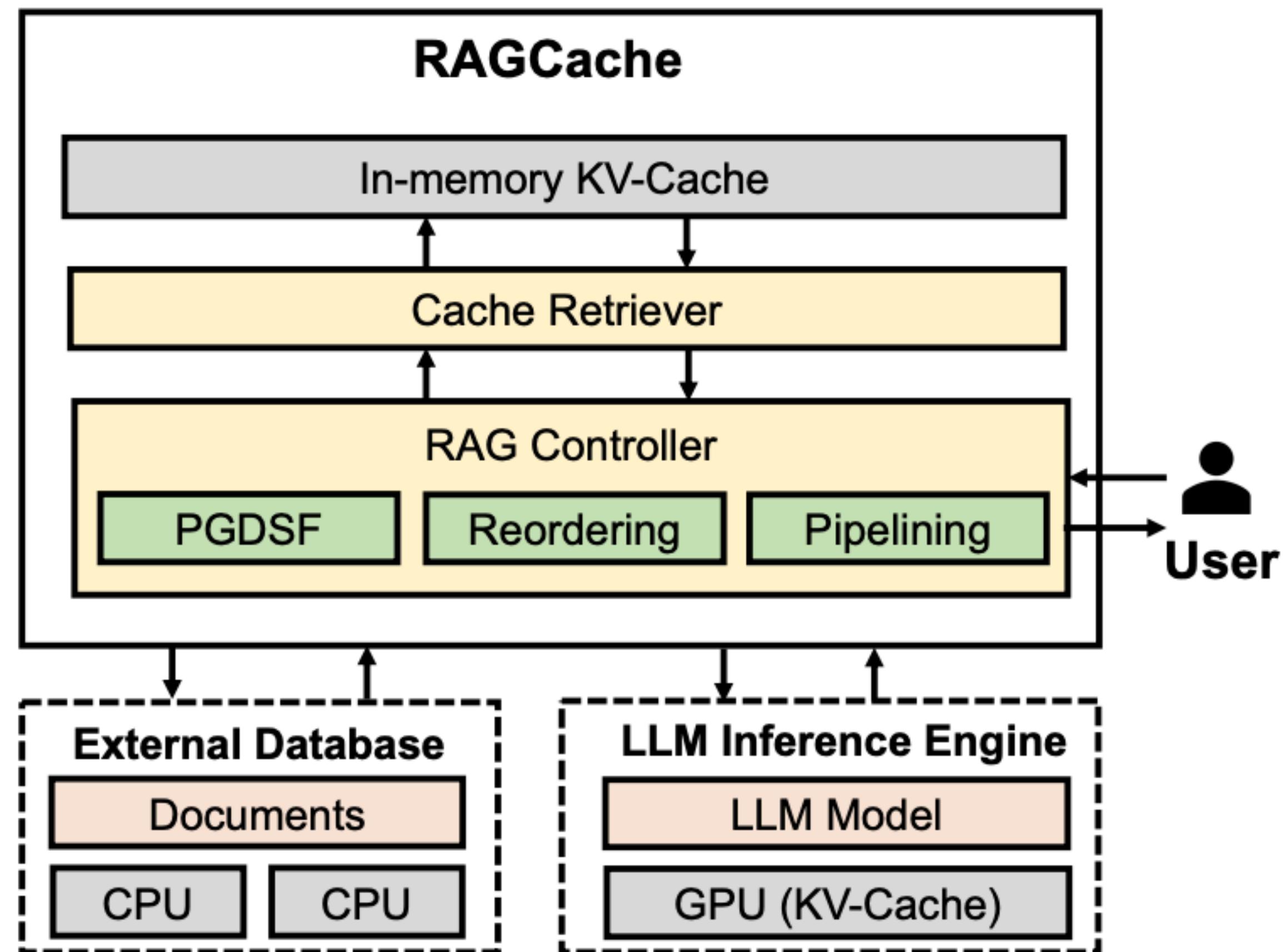
Filtering Context with External Modules



Filtering unnecessary context with additional modules

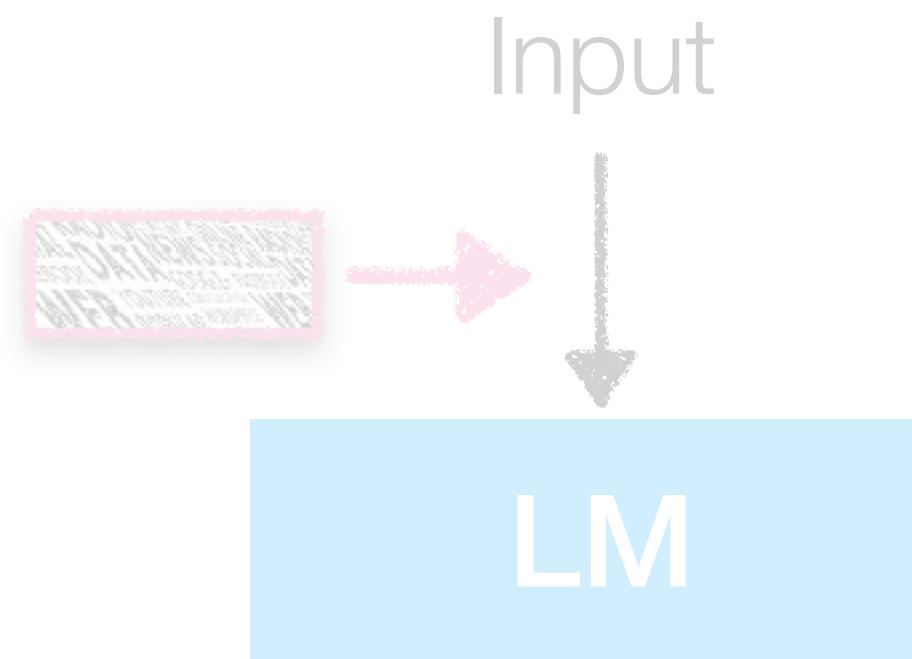
Wang et al. . 2024 Learning to Filter Context for Retrieval-Augmented Generation.

Caching for RAG



How to Use Retrieval

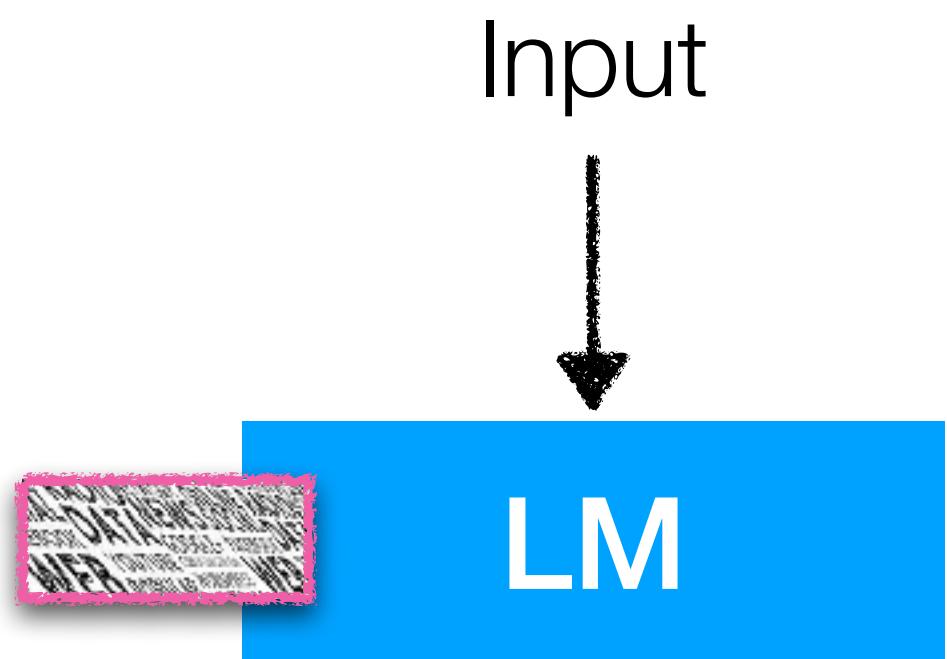
Input Augmentation



- Augment input of LMs
- Easy to apply (w/o training)
- & effective
- Difficulty of using many D

e.g., RAG

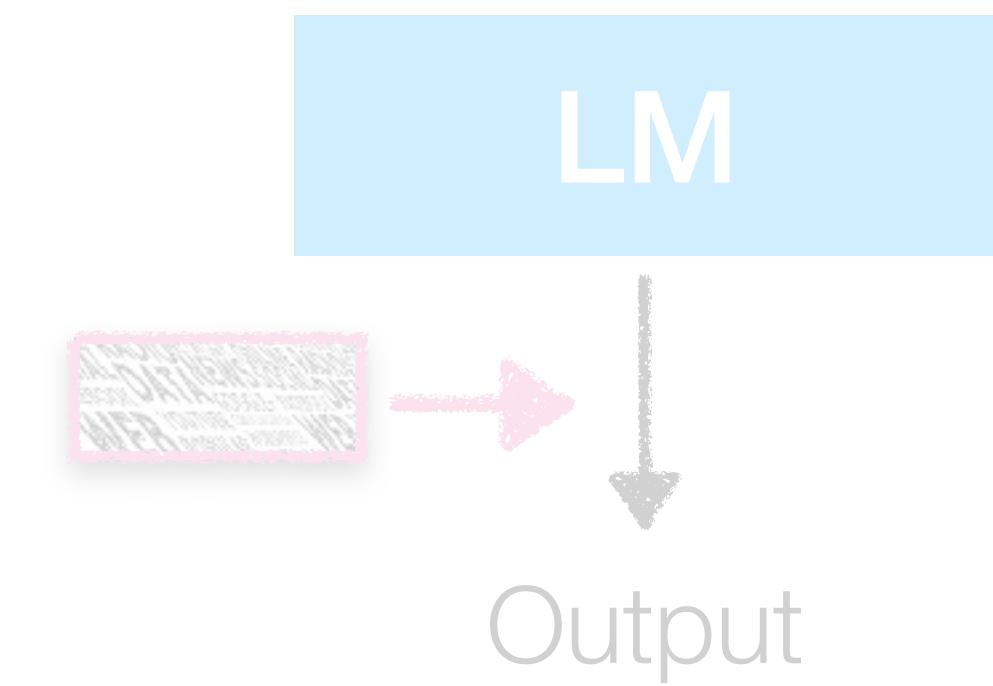
Intermediate Fusion



- Modify LMs to incorporate D in intermediate layers
- Scalable to many passages
- Requires retraining

e.g., RETRO, InstructRETRO

Output Interpolation



- Directly manipulate output token distributions
- No training required*
- Limited effectiveness on tasks

e.g., kNNLM

RETRO (Borgeaud et al. 2021)

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

RETRO (Borgeaud et al. 2021)

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

x₁

x₂

x₃

RETRO (Borgeaud et al. 2021)

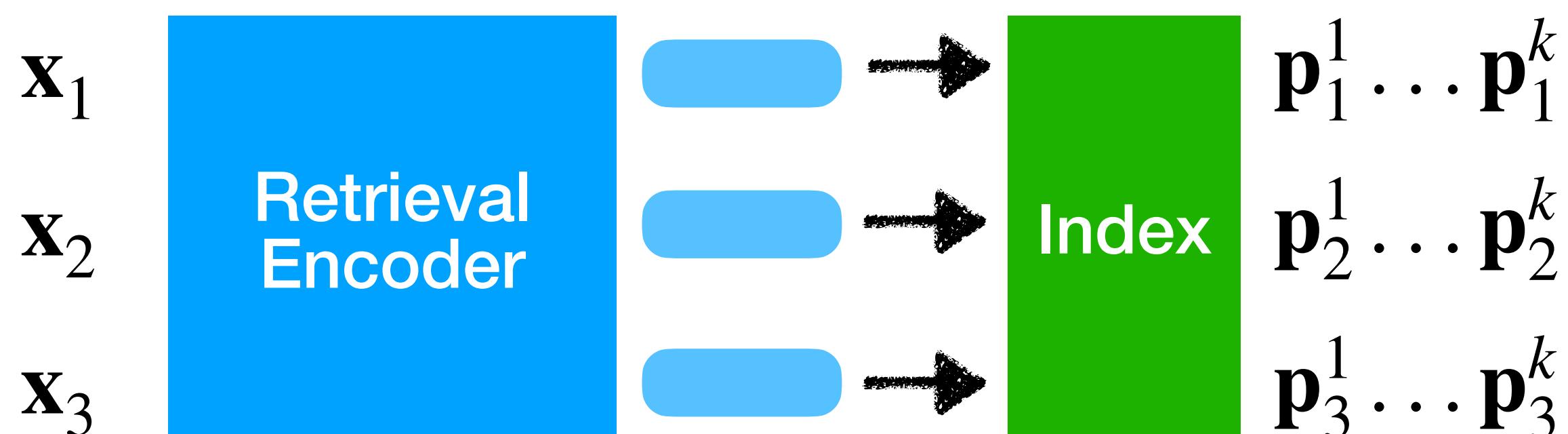
\mathbf{x} = World Cup 2022 was / the last with 32 teams, / before the increase to

\mathbf{x}_1

\mathbf{x}_2

\mathbf{x}_3

(k chunks of text per split)



RETRO (Borgeaud et al. 2021)

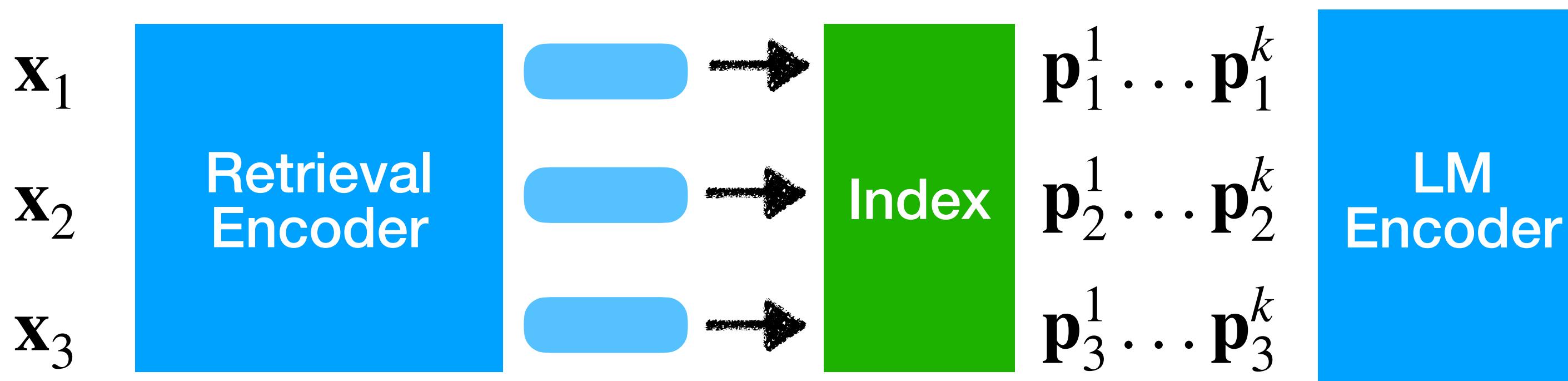
\mathbf{x} = World Cup 2022 was / the last with 32 teams, / before the increase to

\mathbf{x}_1

\mathbf{x}_2

\mathbf{x}_3

(k chunks of text per split)



RETRO (Borgeaud et al. 2021)

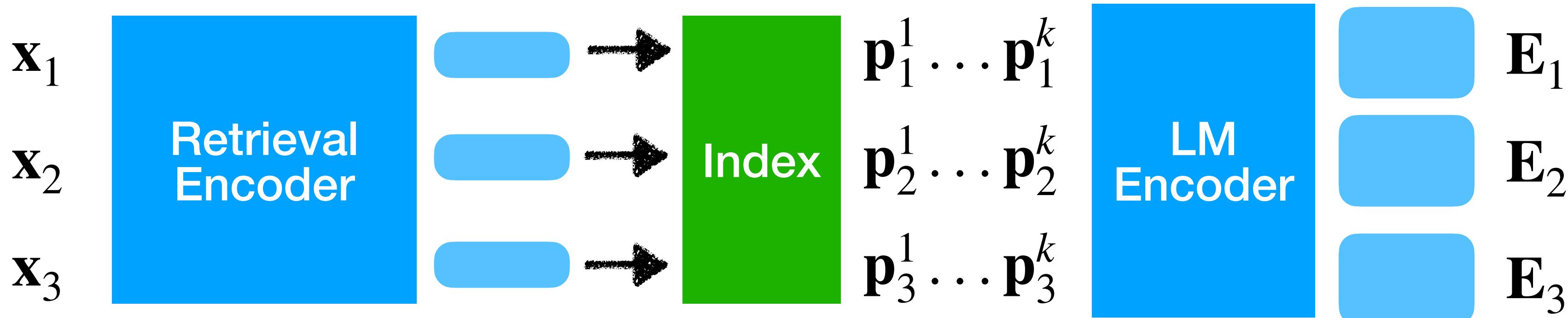
\mathbf{x} = World Cup 2022 was / the last with 32 teams, / before the increase to

\mathbf{x}_1

\mathbf{x}_2

\mathbf{x}_3

(k chunks of text per split)



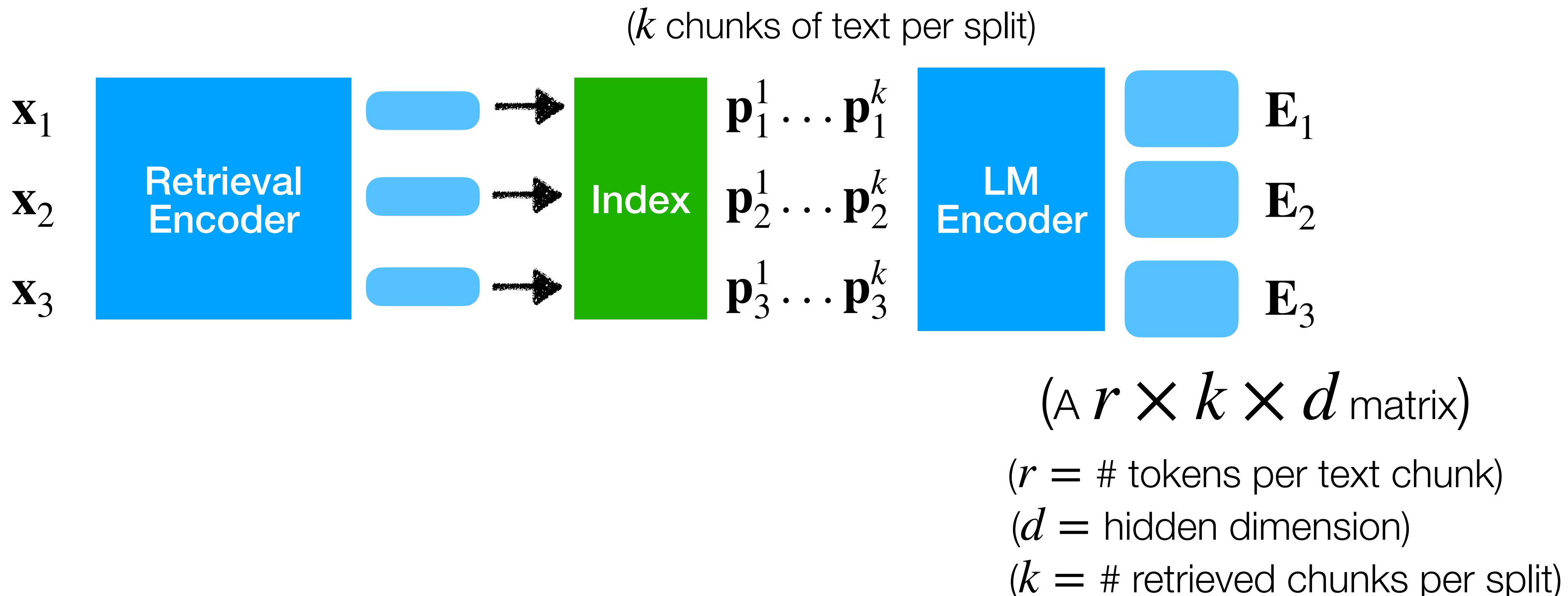
RETRO (Borgeaud et al. 2021)

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

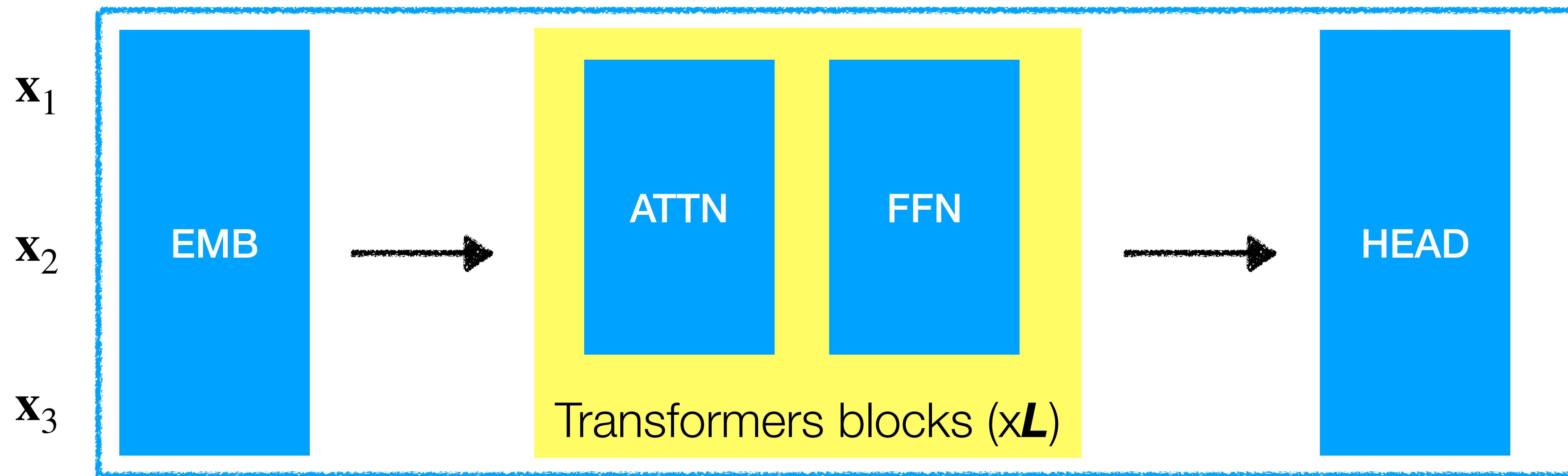
\mathbf{x}_1

\mathbf{x}_2

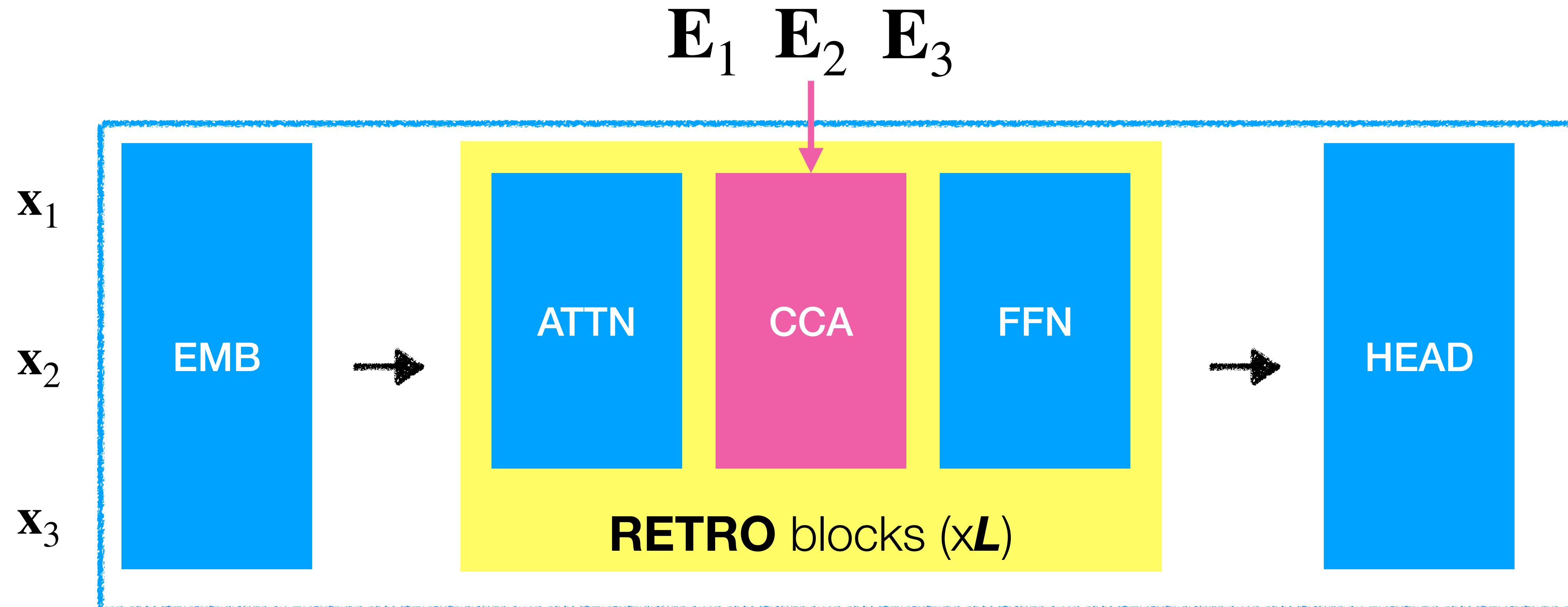
\mathbf{x}_3



Regular Decoder

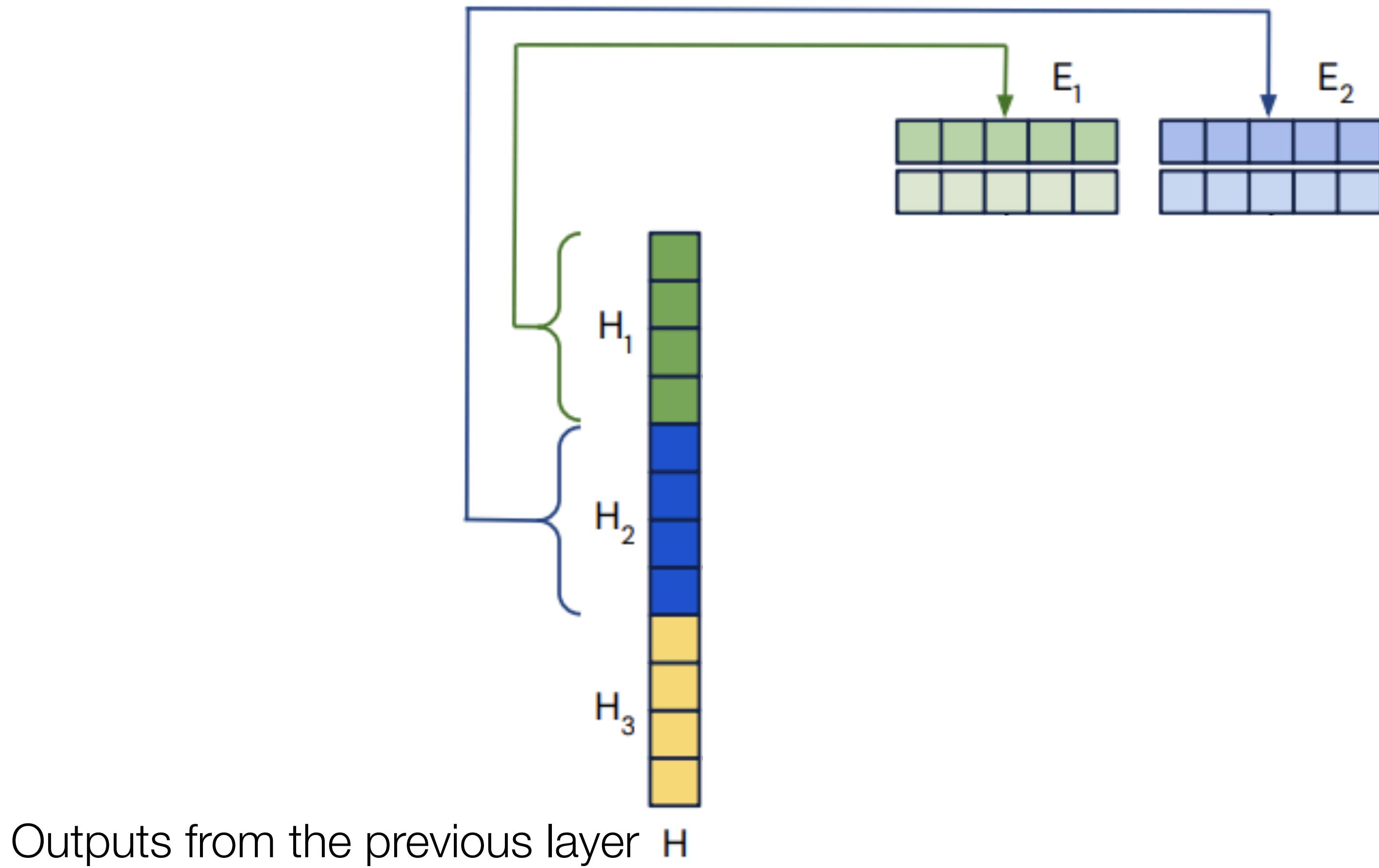


Decoder in RETRO

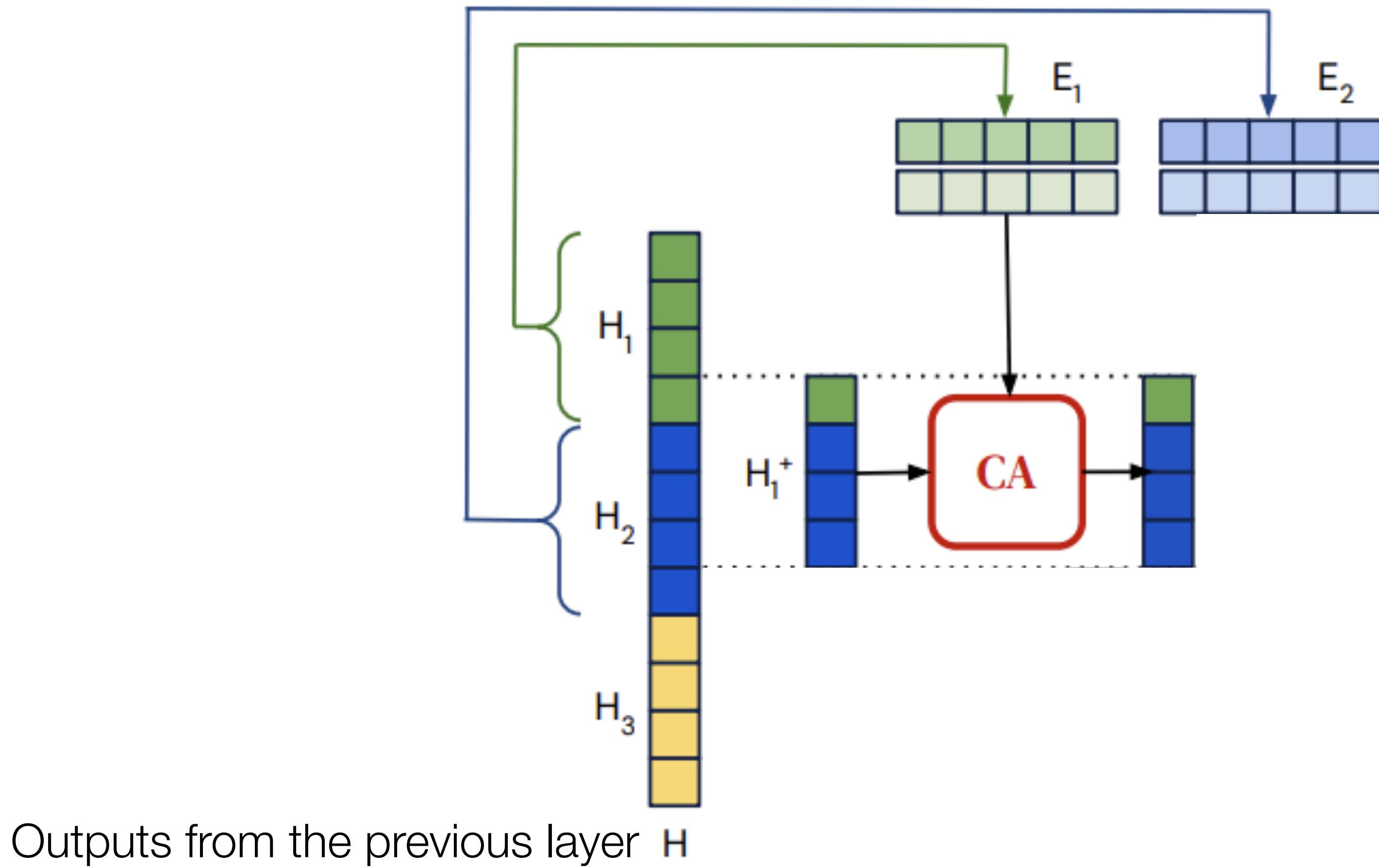


Chunked Cross Attention (CCA)

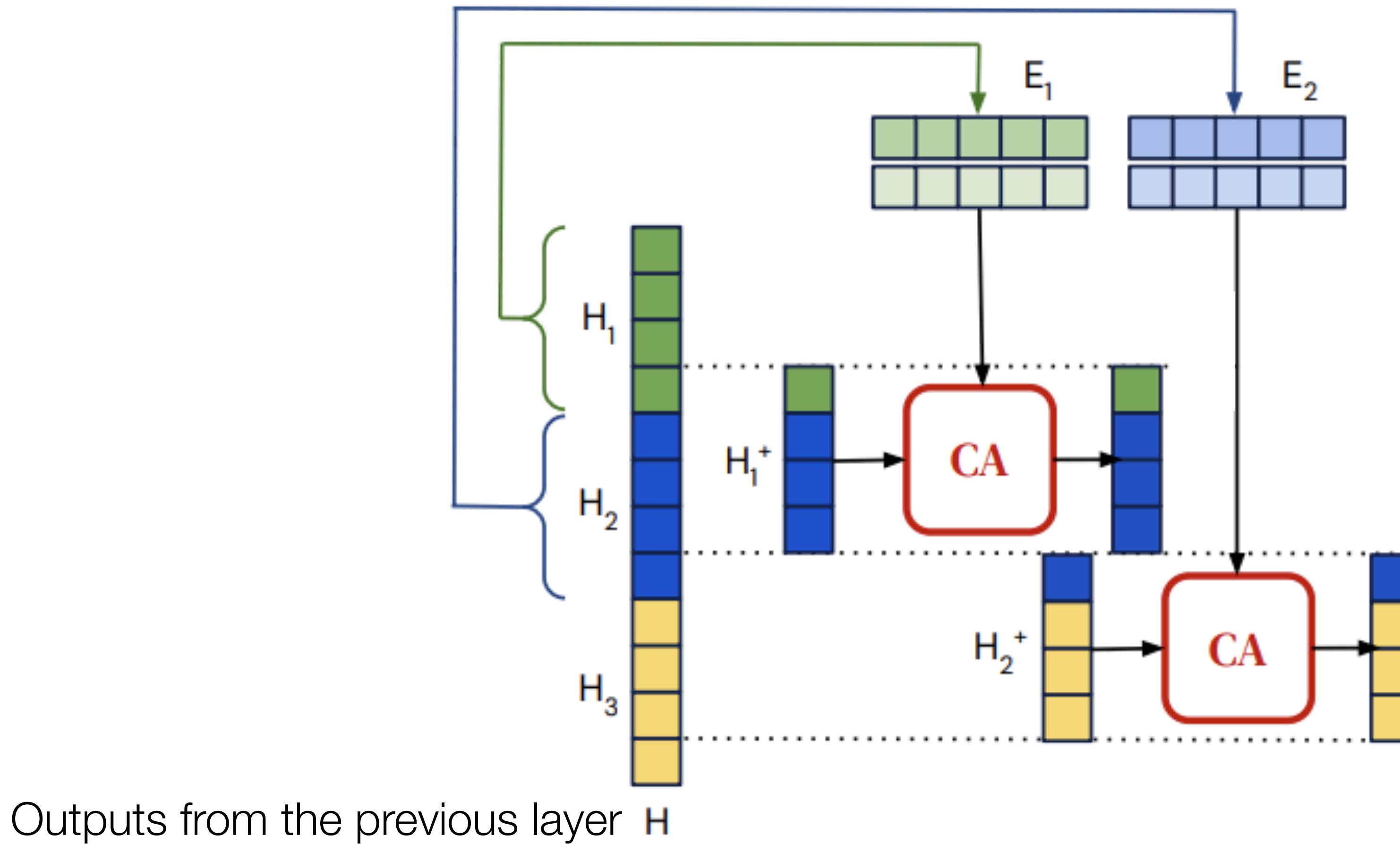
Chunked Cross Attention



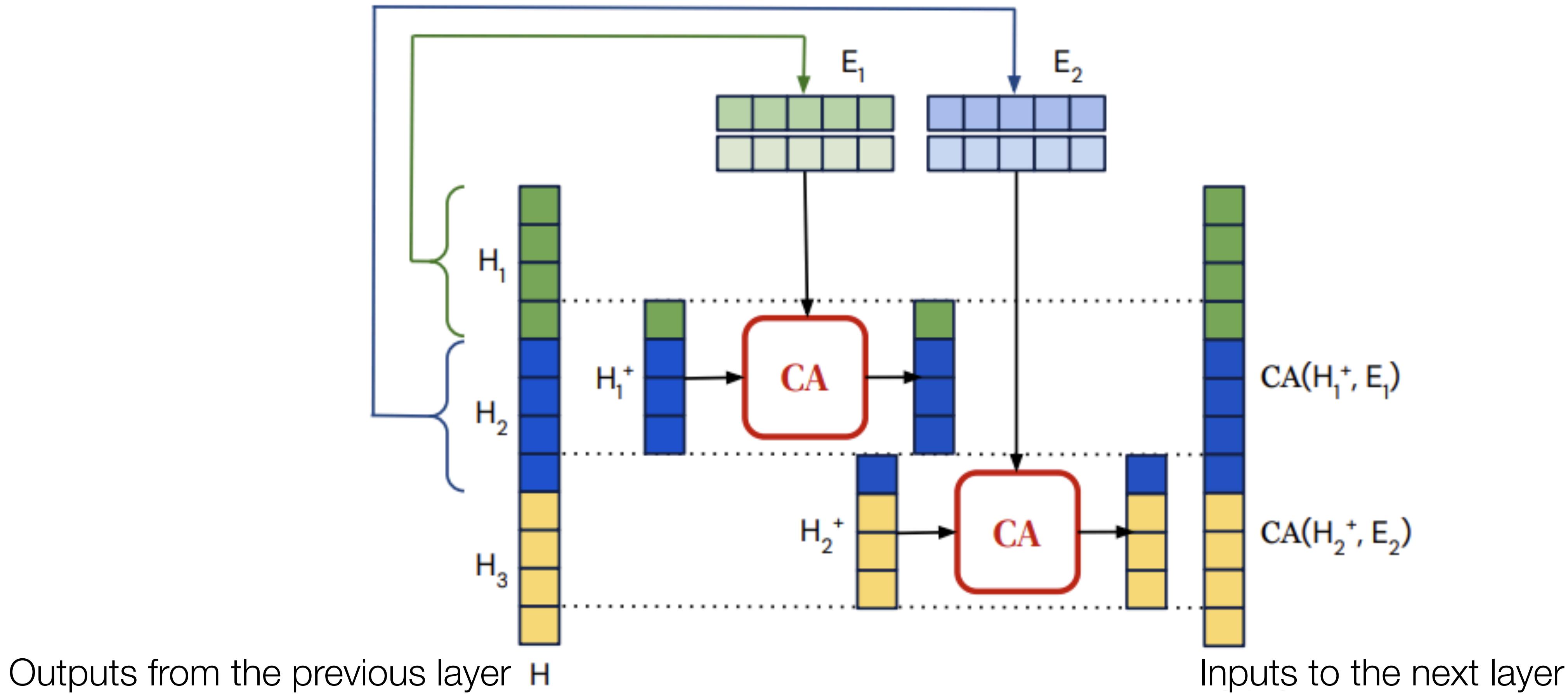
Chunked Cross Attention



Chunked Cross Attention

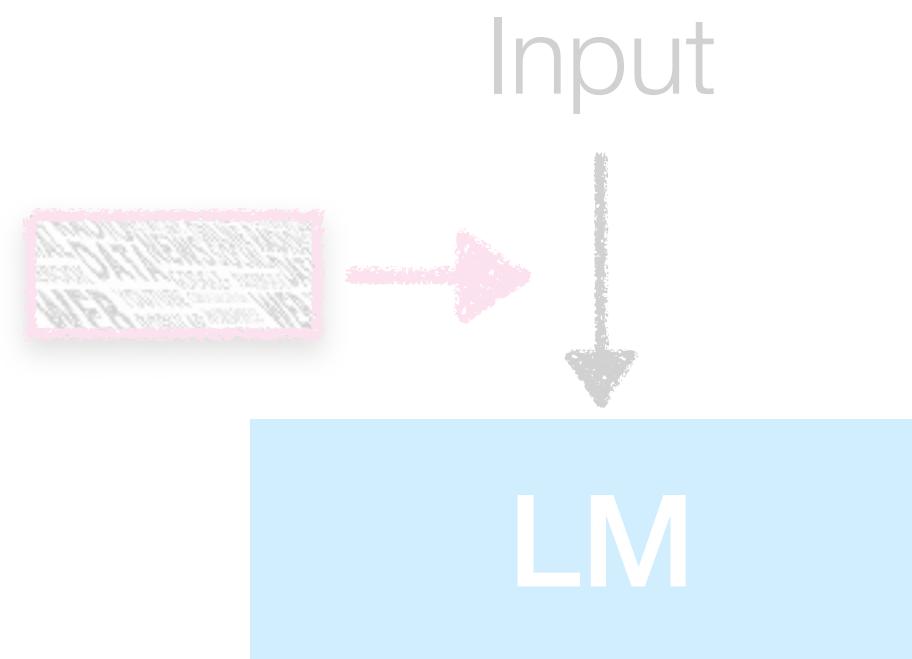


Chunked Cross Attention



How to Use retrieval

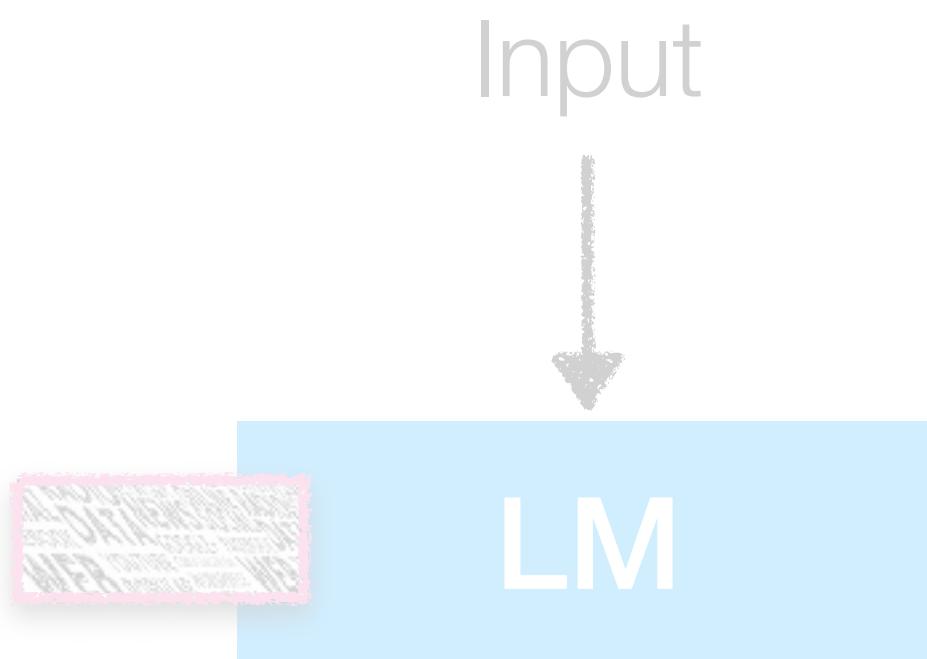
Input Augmentation



- Augment input of LMs
- Easy to apply (w/o training)
- & effective
- Difficulty of using many D

e.g., RAG

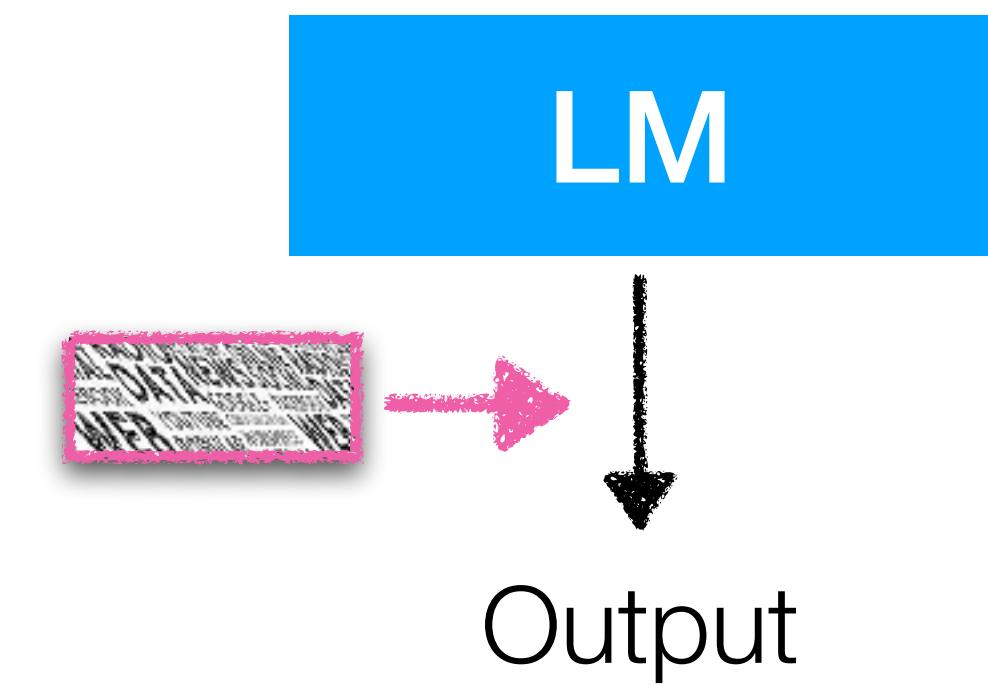
Intermediate Fusion



- Modify LMs to incorporate D in intermediate layers
- Scalable to many passages
- Requires retraining

e.g., RETRO, InstructRETRO

Output Interpolation



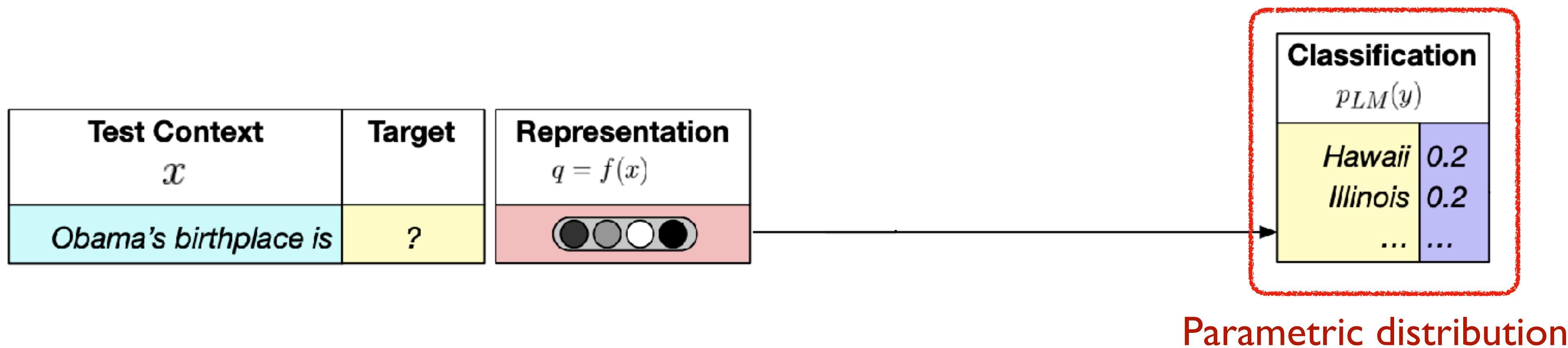
- Directly manipulate output token distributions
- No training required*
- Limited effectiveness on tasks

e.g., kNNLM

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)

... 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)

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,

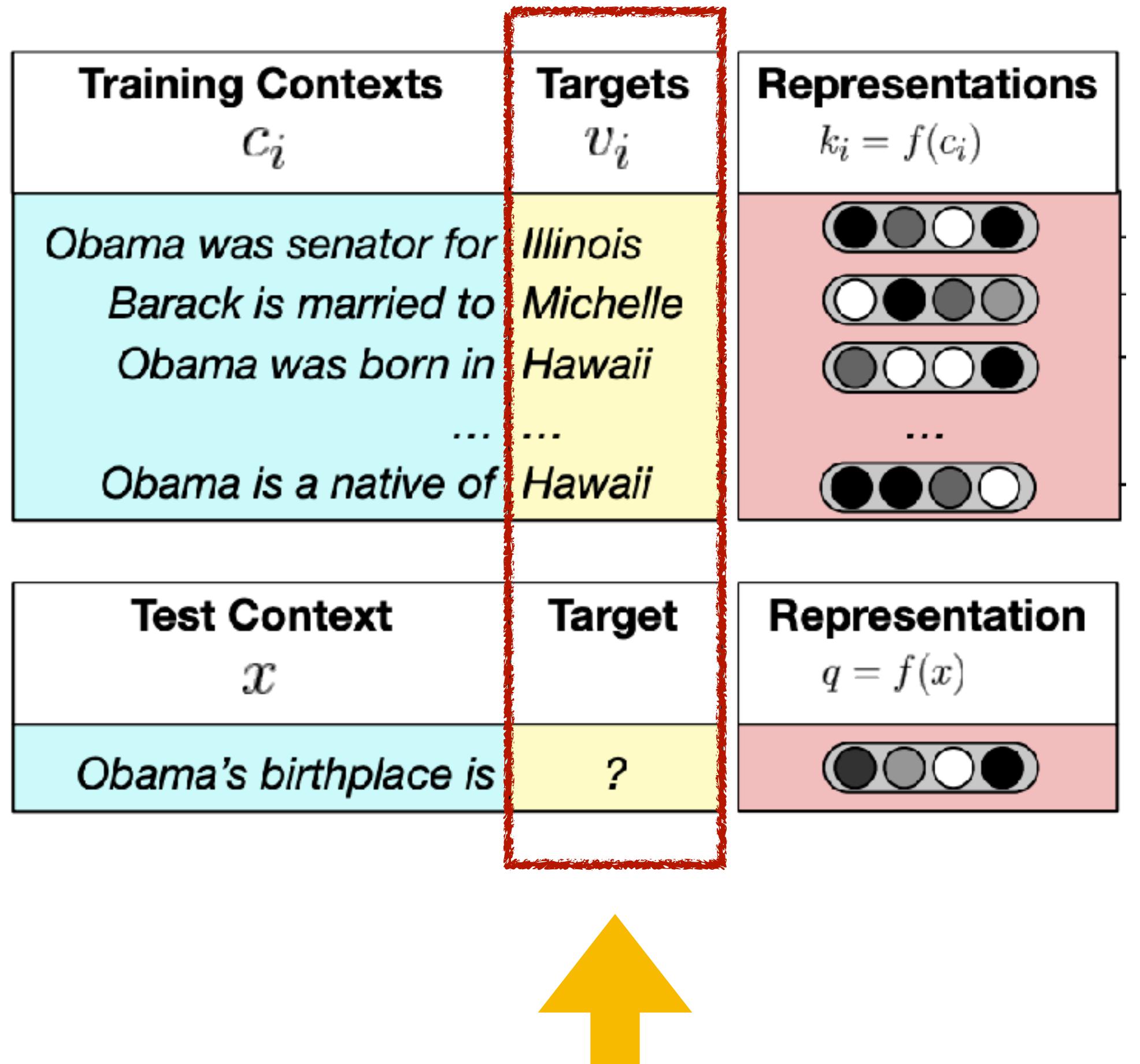
Test Context	Target	Representation
x		$q = f(x)$
<i>Obama's birthplace is</i>	?	

kNN-LM (Khandelwal et al. 2020)

Training Contexts	Targets	Representations
c_i	v_i	$k_i = f(c_i)$
<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>	

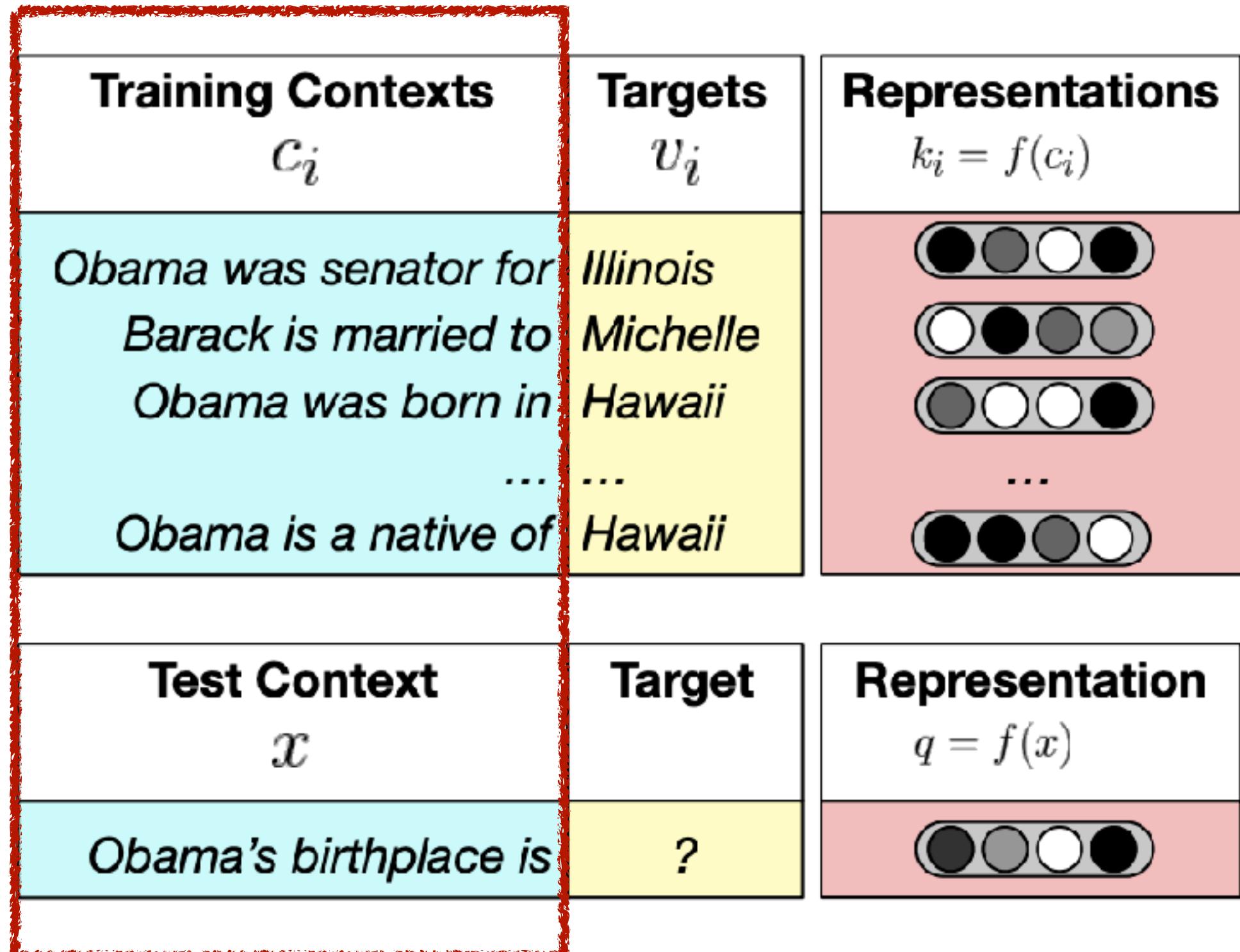
Test Context	Target	Representation
x		$q = f(x)$
<i>Obama's birthplace is</i>	?	

kNN-LM (Khandelwal et al. 2020)



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

kNN-LM (Khandelwal et al. 2020)



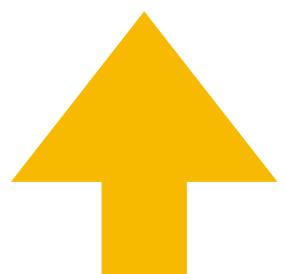
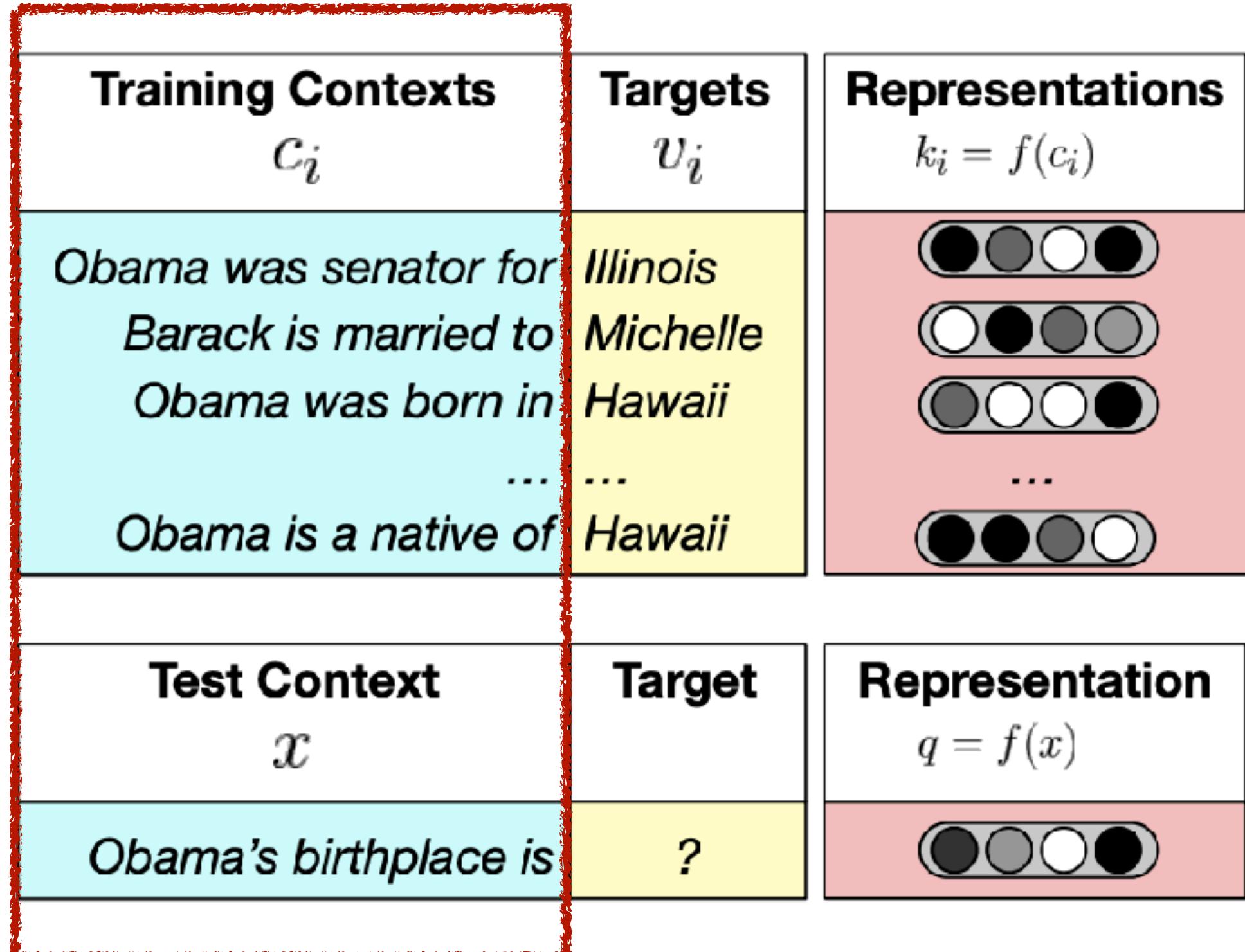
Which tokens in a datastore are close to the next token?

=

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

kNN-LM (Khandelwal et al. 2020)

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

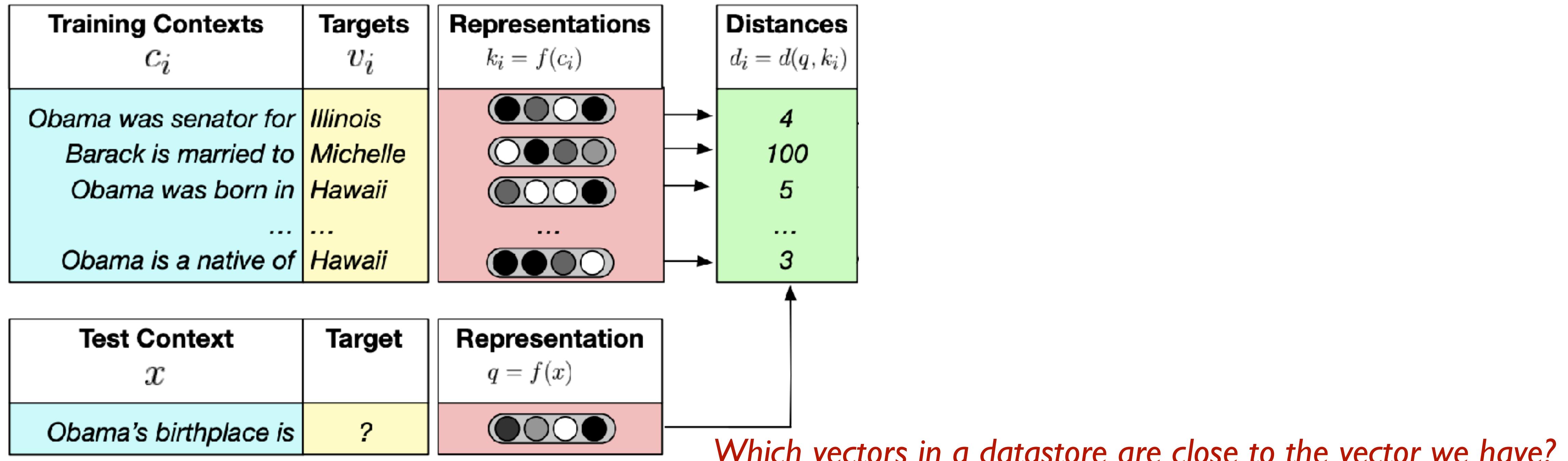


Which tokens in a datastore are close to the next token?

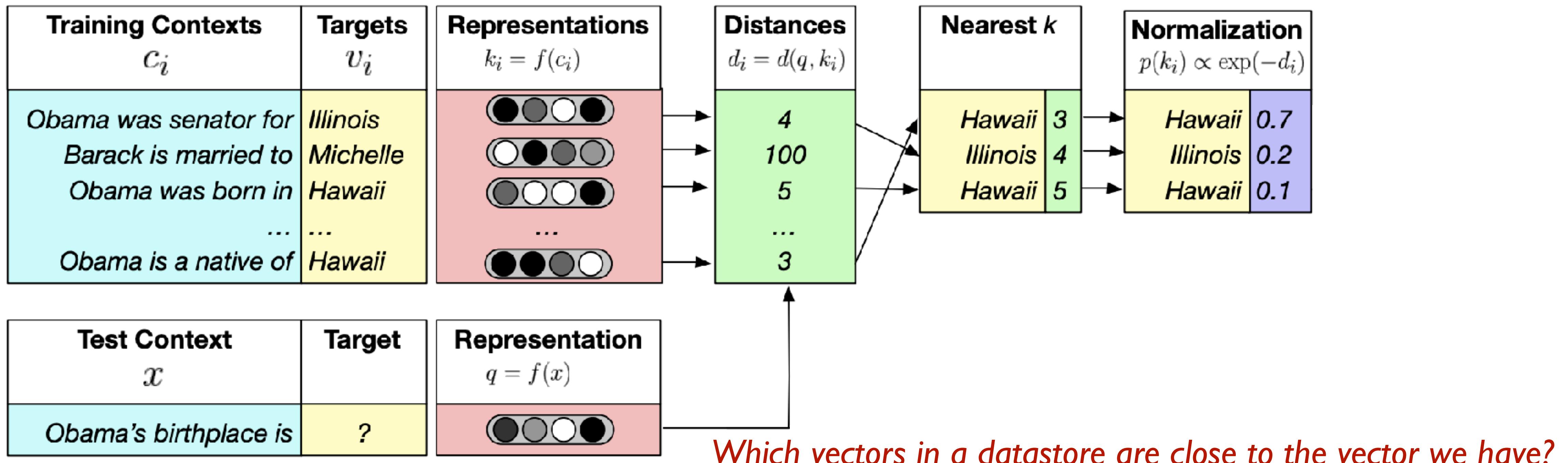
=

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

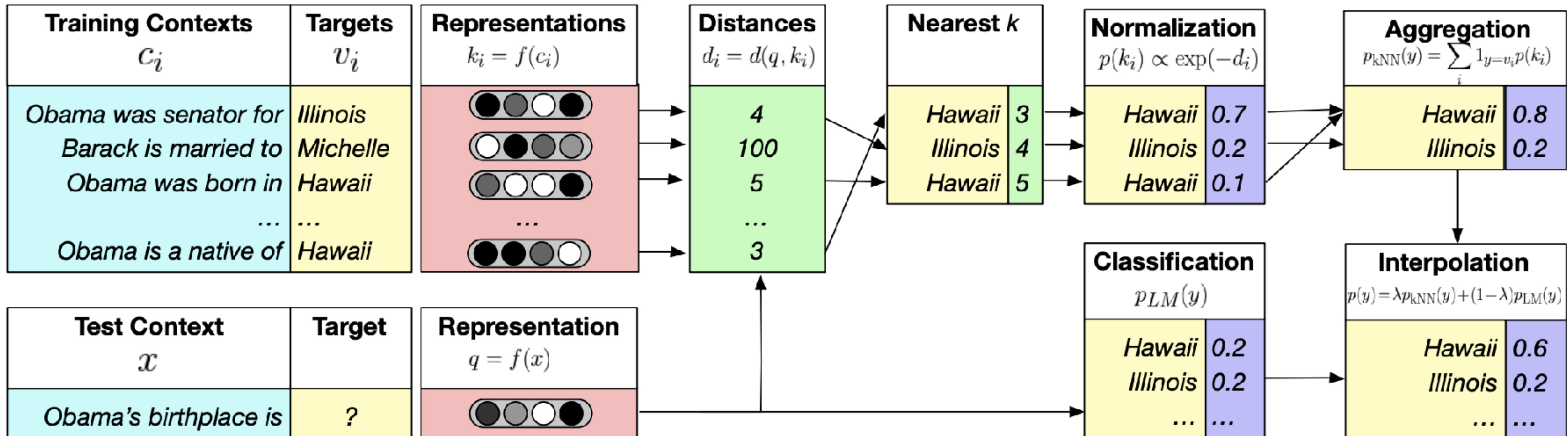
kNN-LM (Khandelwal et al. 2020)



kNN-LM (Khandelwal et al. 2020)



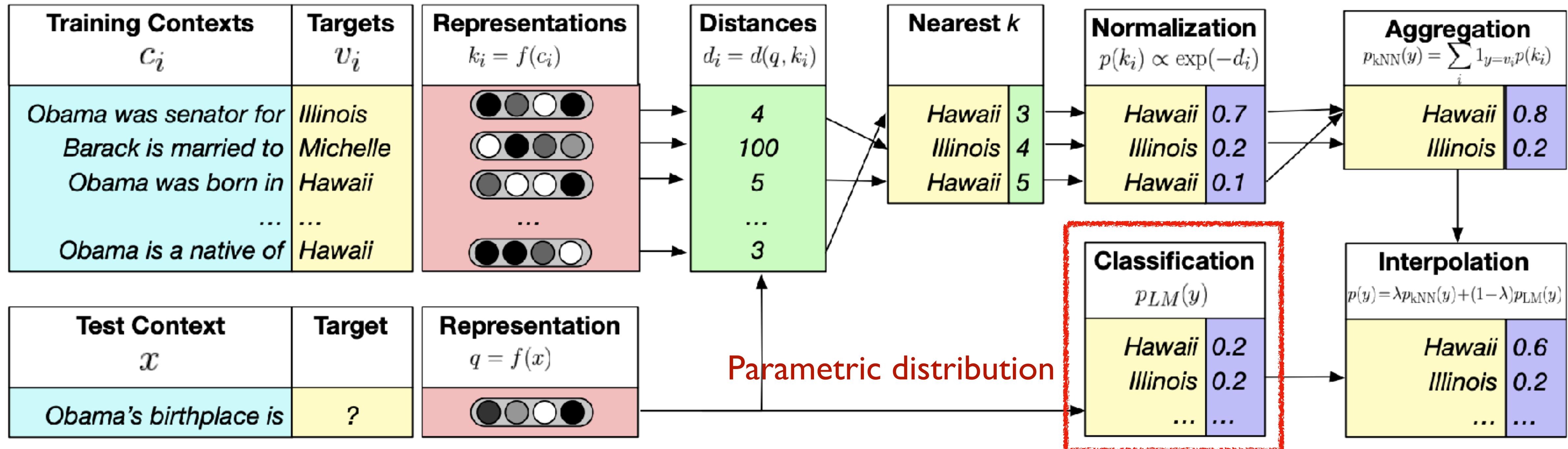
kNN-LM (Khandelwal et al. 2020)



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

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

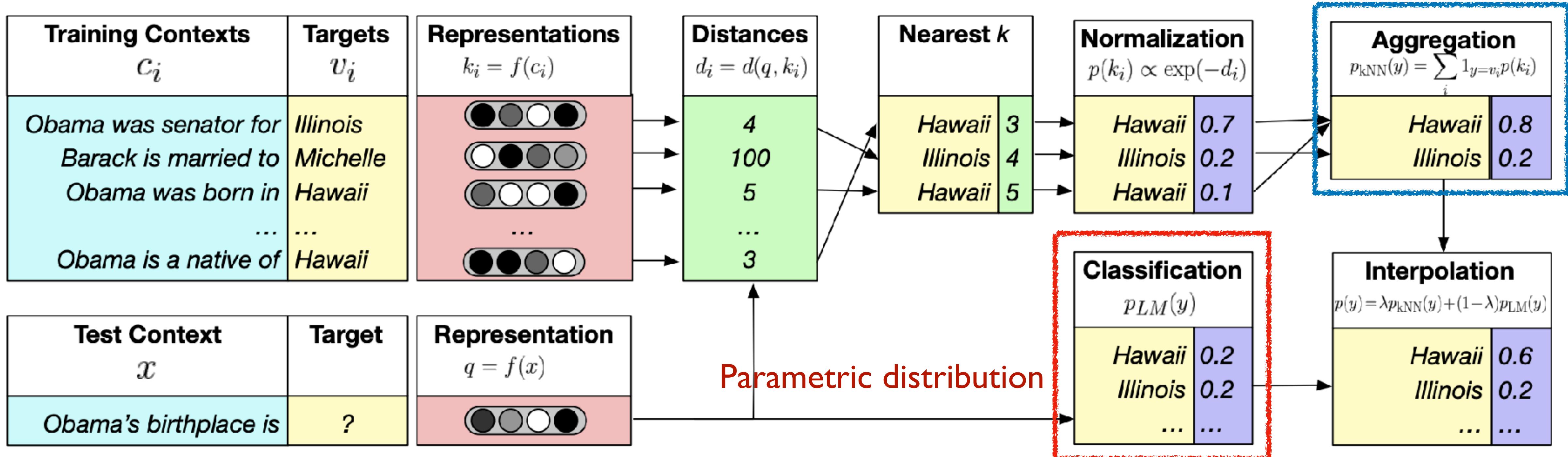
kNN-LM (Khandelwal et al. 2020)



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

kNN-LM (Khandelwal et al. 2020)

Nonparametric distribution

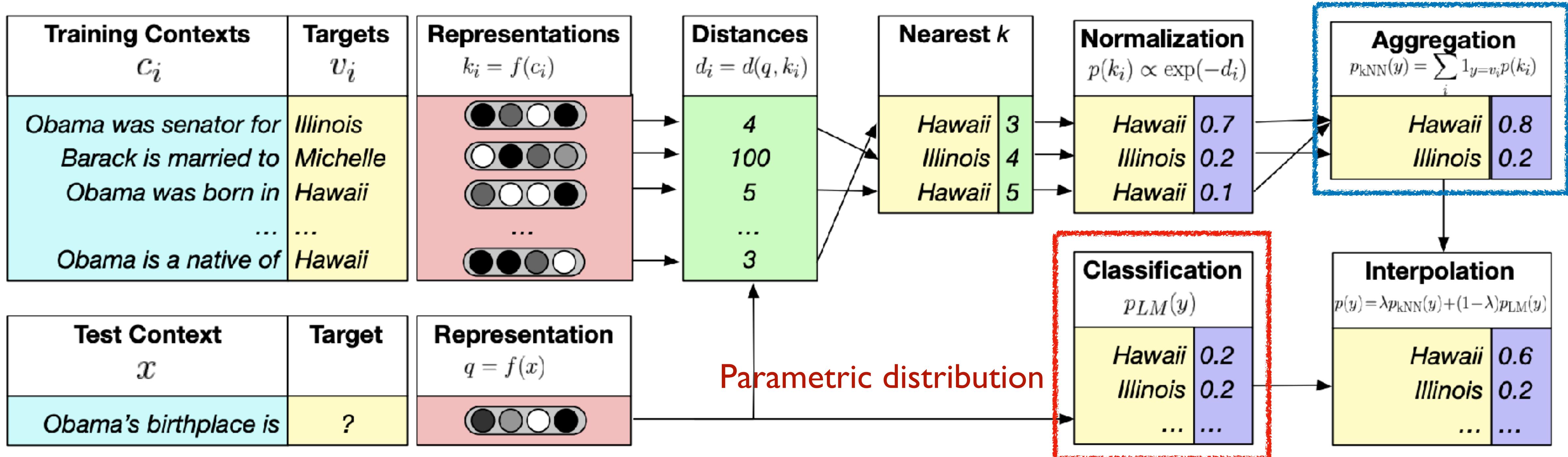


$$P_{kNN-LM}(y|x) = (1 - \lambda)P_{LM}(y|x) + \lambda P_{kNN}(y|x)$$

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

kNN-LM (Khandelwal et al. 2020)

Nonparametric distribution



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

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

RETRO w/ 1.8T datastores achieves SOTA

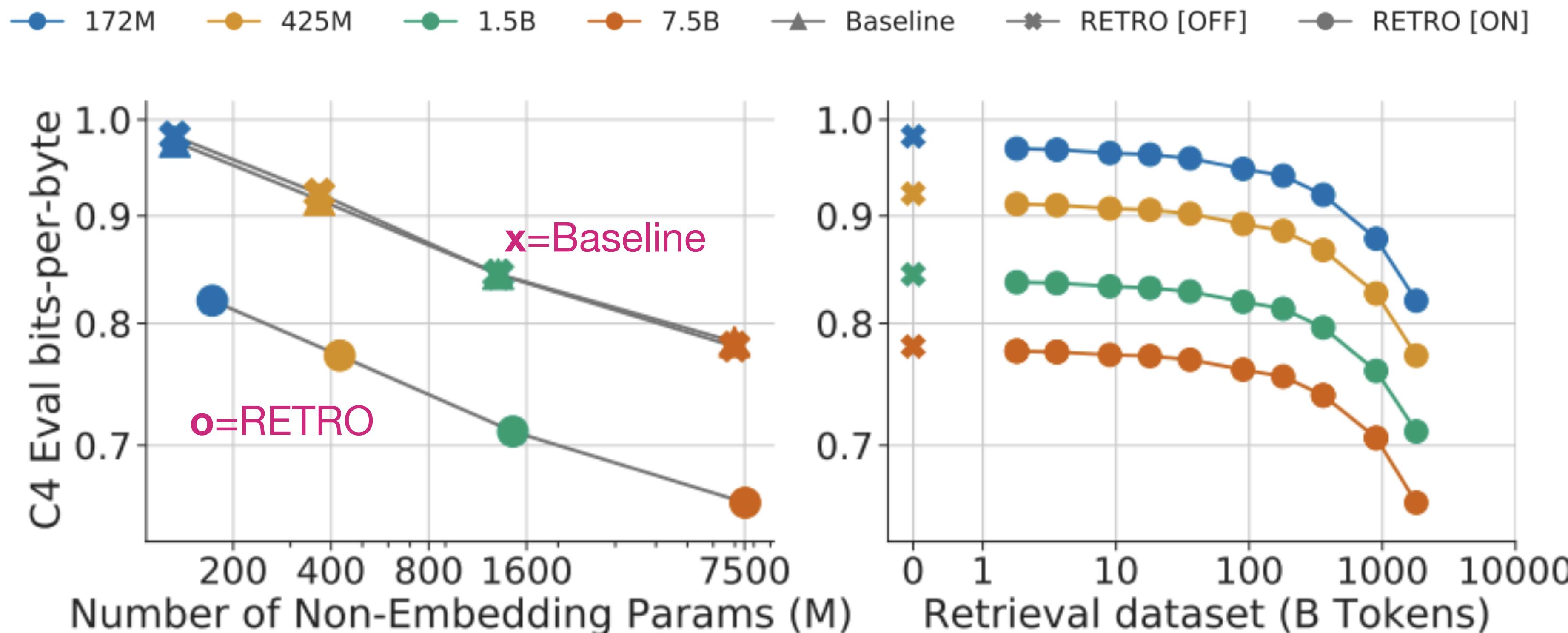
Results

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

RETRO w/ 1.8T datastores achieves SOTA

Results



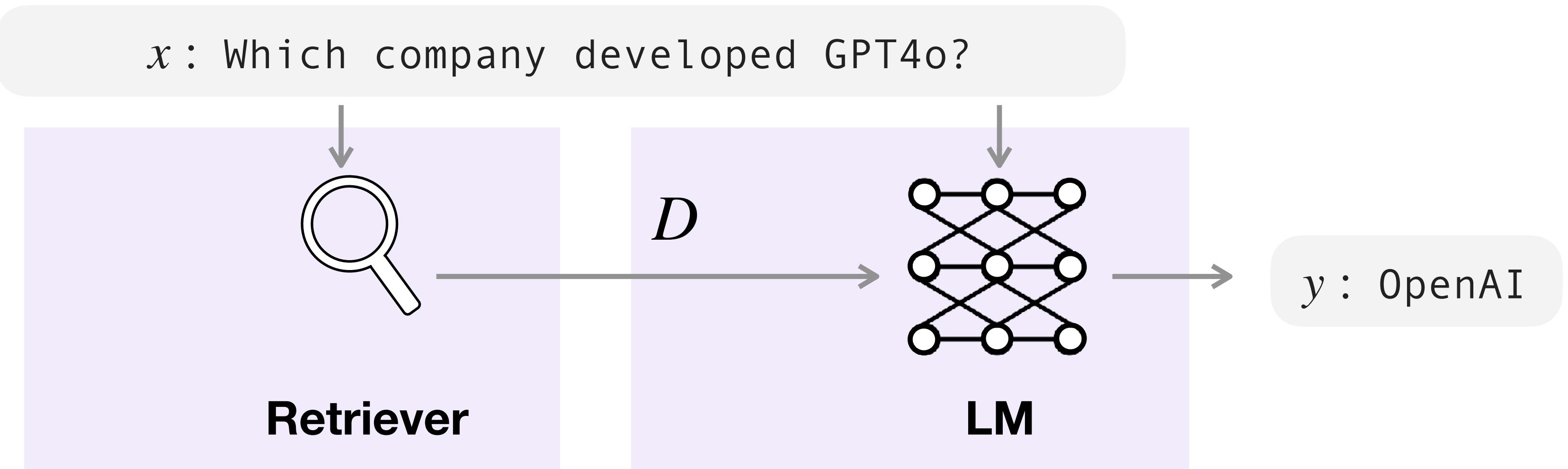
Gains are constant with model scale

The larger datastore is, the better

Are They Better than RAG?

Task	NQ	TriviaQA	NewsQA	SQuAD 2.0	SQuAD 1.1	Quoref	NarrativeQA	DROP
Metric	EM	EM	F1	F1 / EM	F1 / EM	F1	F1	F1
InstructGPT _{RAG} 43B	37.0	78.1	52.4	70.7 / 64.3	72.4 / 65.8	71.5	53.9	51.8
InstructRetro 43B (w/o encoder, Avg: +7%)	38.9 (+5.14%)	78.3 (+0.26%)	57.4 (+9.54%)	75.6 / 69.3 (+6.93%)	77.1 / 70.4 (+6.49%)	76.2 (+6.57%)	60.0 (+11.32%)	54.8 (+5.79%)
InstructRetro 48B (w/ encoder, Avg: +6%)	38.6 (+4.32%)	77.8 (-0.38%)	57.0 (+8.78%)	74.8 / 67.7 (+5.80%)	76.4 / 69.0 (+5.52%)	76.1 (+6.43%)	59.8 (+10.95%)	54.6 (+5.41%)

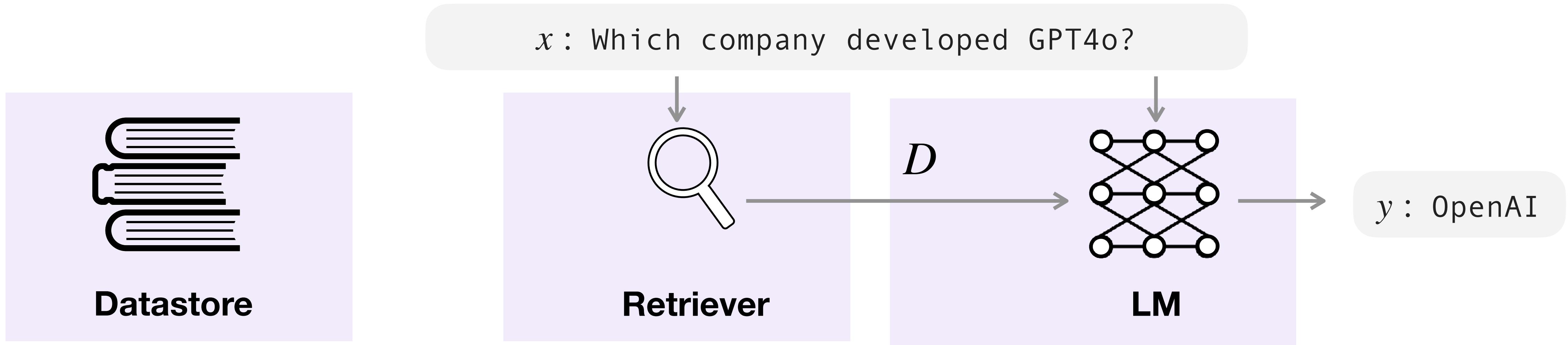
Summary of Part 2



- ✓ Common architectures
- ✓ Recent progress

- **RAG** is widely used but several limitations
- Recent progress to overcome such shortcomings
- Other architectures: **intermediate incorporations** or **output interpolation**

Today's Outline



- ✓ Sources of datastore
- ✓ Processing
- ✓ Scaling
- ✓ Types of retrievers
- ✓ Training
- ✓ Evaluations
- ✓ Common architectures
- ✓ Recent progress in RAG