September 25, 2025 at CMU Advanced NLP

Retrieval and Retrieval-Augmented Generation

Akari Asai

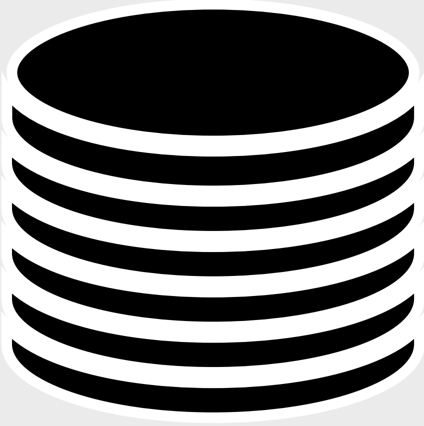
aasai@andrew.cmu.edu | akaria@allenai.org
https://akariasai.github.io/

Ai2

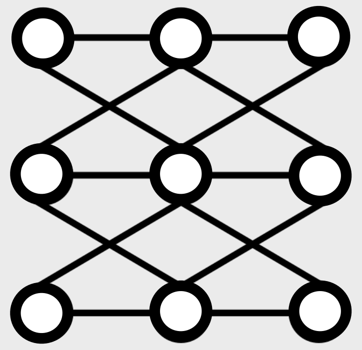
Slides adapted from
ACL 2023 tutorial by Akari Asai, Sewon Min, Zexuan Zhong, Danqi Chen https://acl2023-retrieval-lm.github.io/
Advanced NLP Fall 2024 by Prof. Graham Neubig https://phontron.com/class/anlp-fall2024/

Parametric LMs

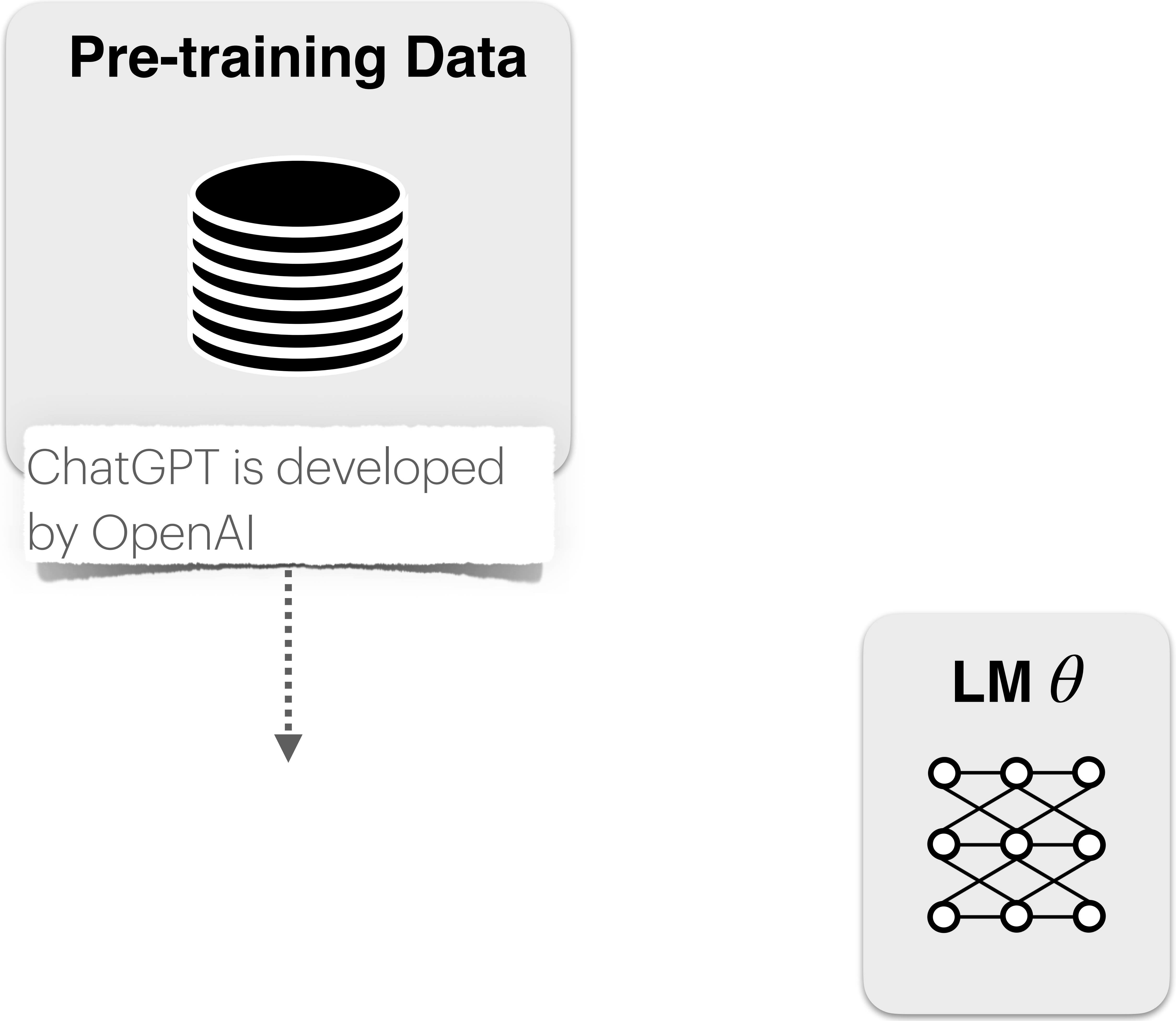
Pre-training Data



LM θ

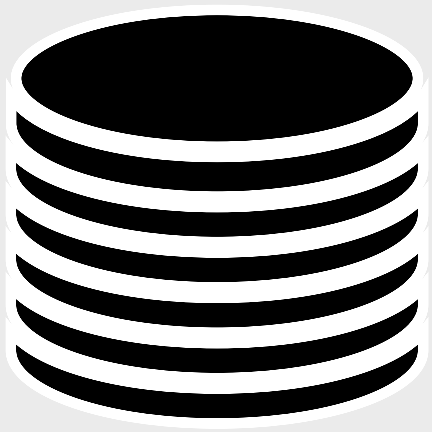


Parametric LMs



Parametric LMs

Pre-training Data

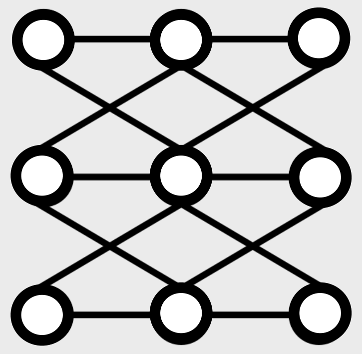


ChatGPT is developed by OpenAI

`ChatGPT is developed by`

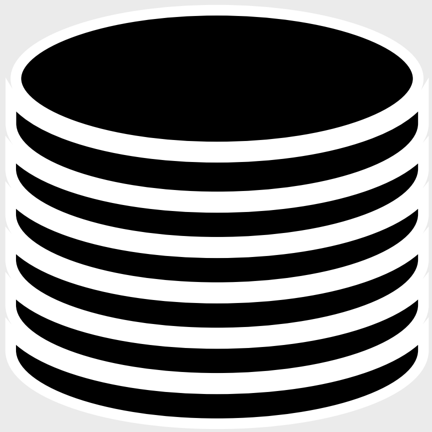
x_1 x_2 x_3 x_4

LM θ



Parametric LMs

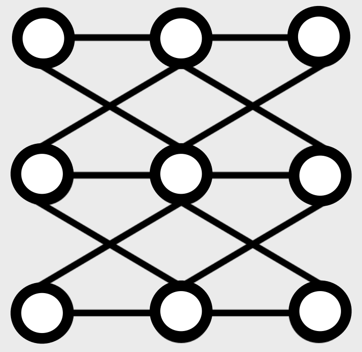
Pre-training Data



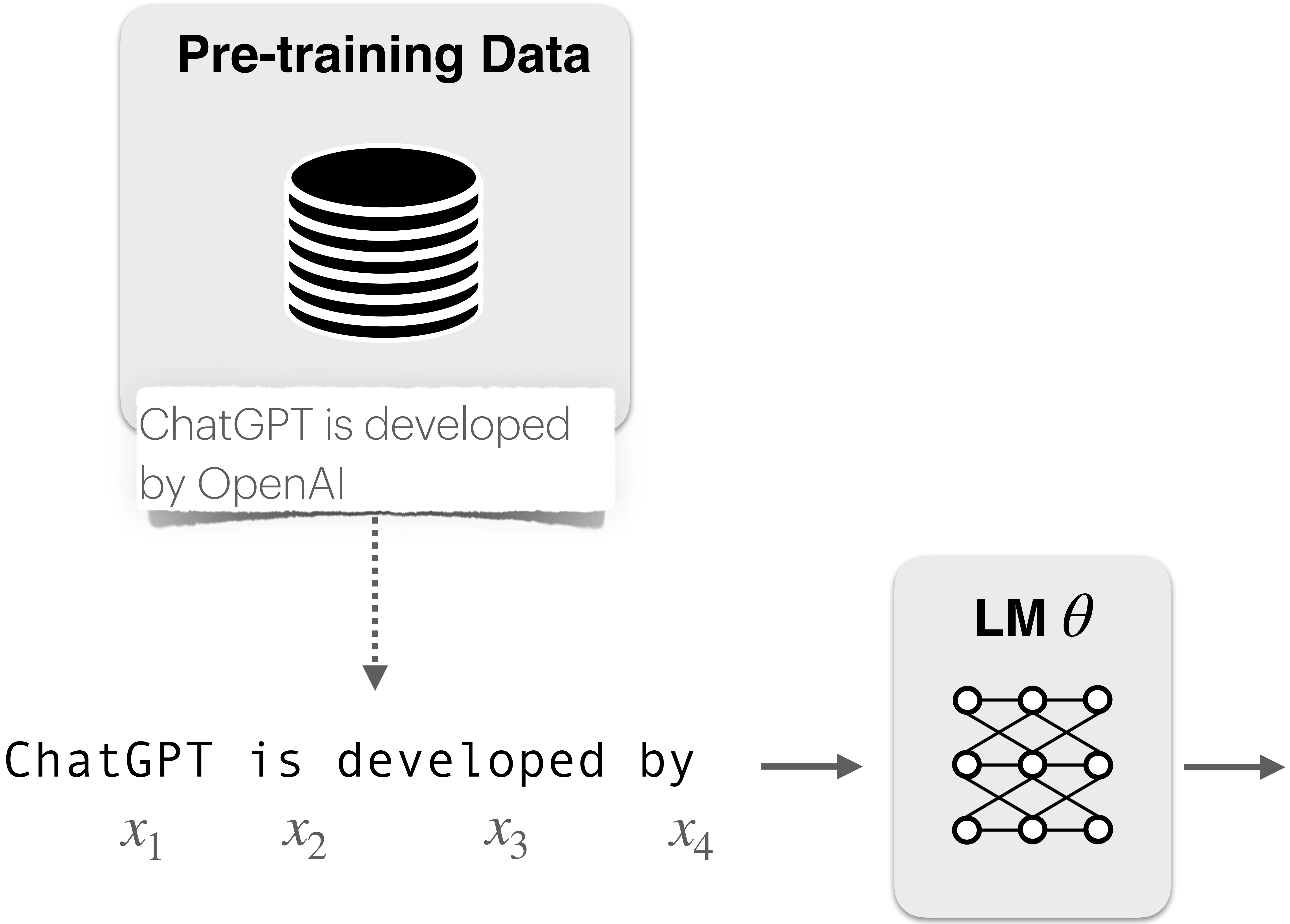
ChatGPT is developed by OpenAI

`ChatGPT is developed by` → **LM θ**

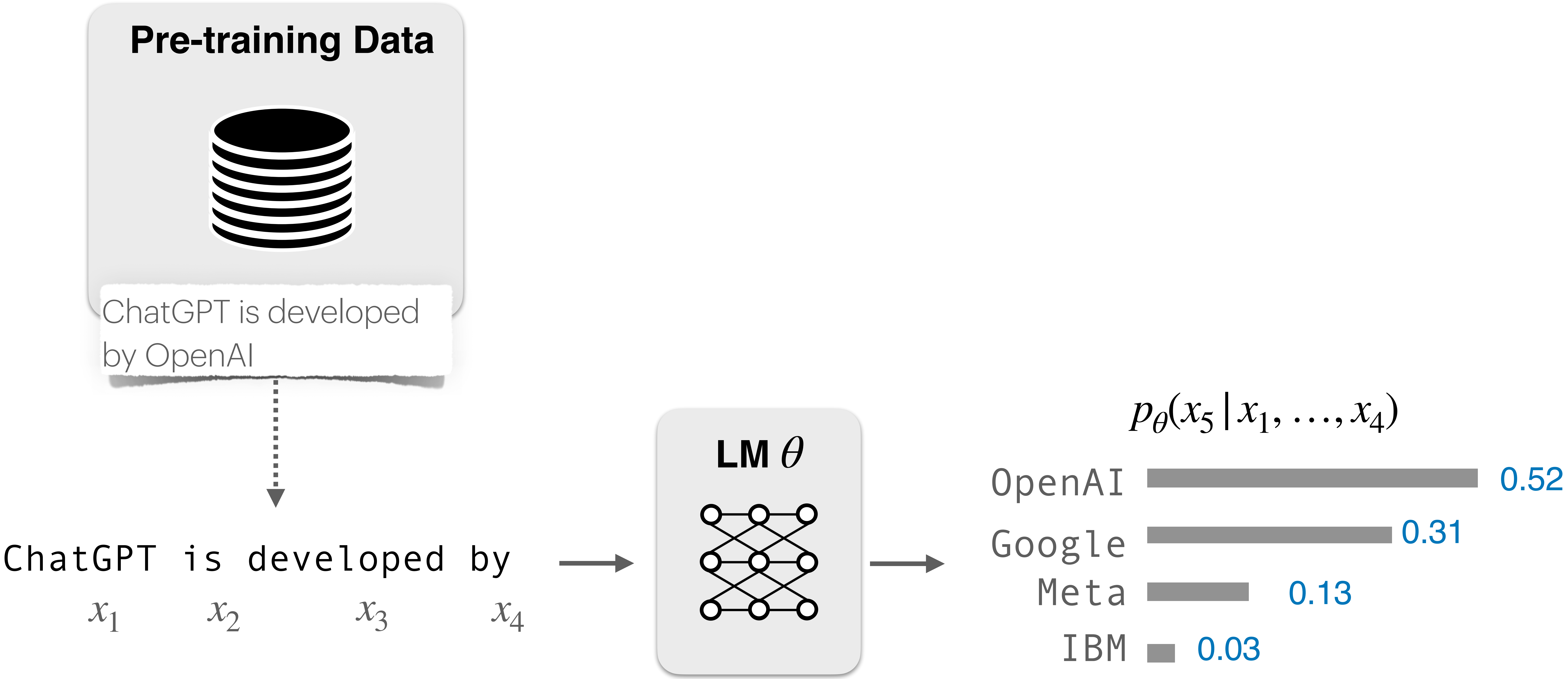
x_1 x_2 x_3 x_4



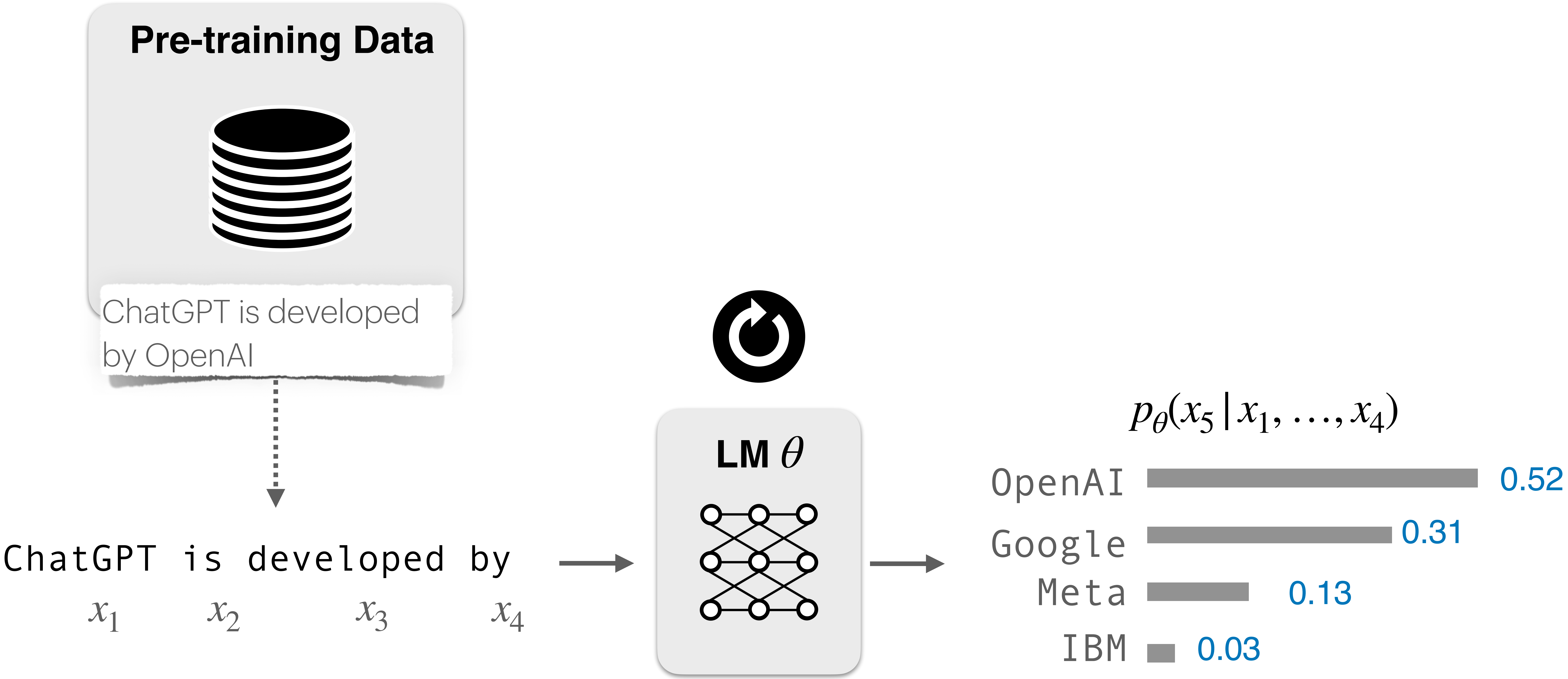
Parametric LMs



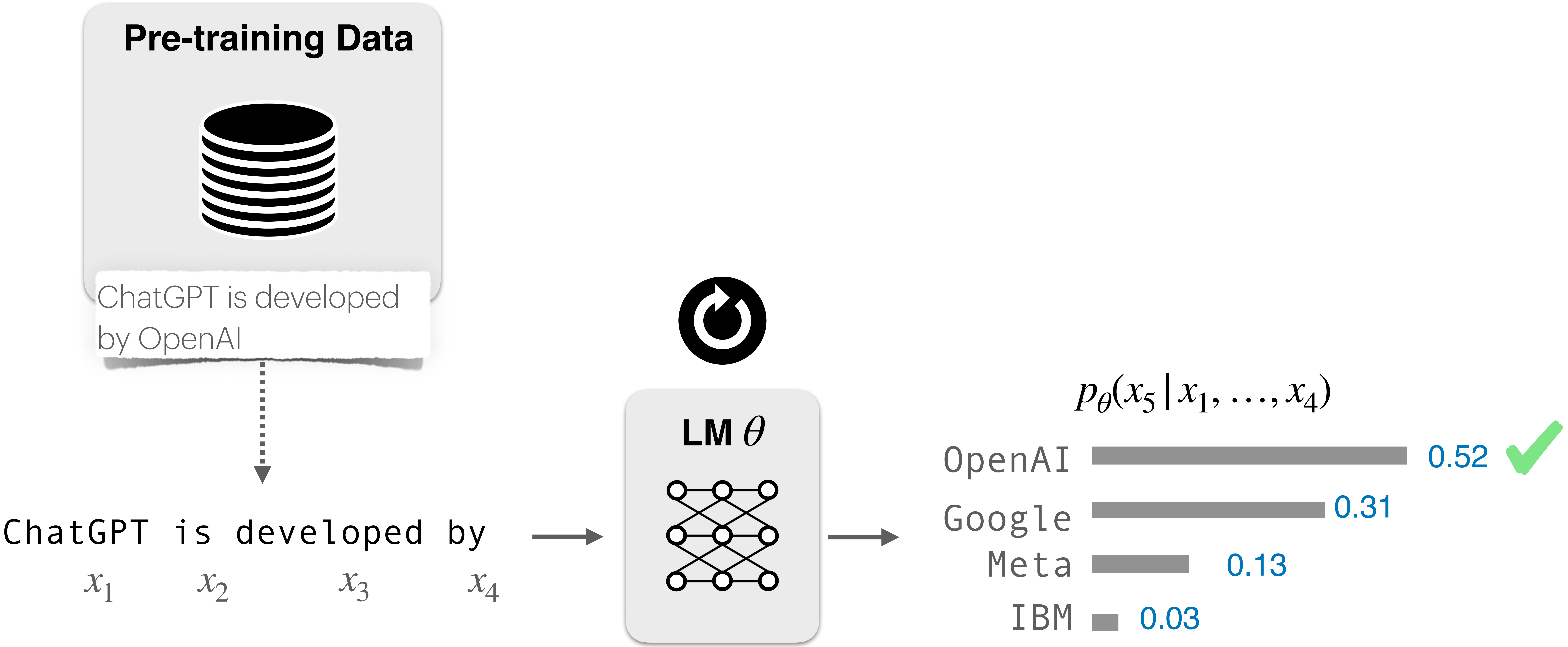
Parametric LMs



Parametric LMs



Parametric LMs



Limitations of Current LMs: Hallucinations

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Limitations of Current LMs: Hallucinations

Explain Akari Asai's research, including her representative papers

(ChatGPT4o, February 8, 2025)

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

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.

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

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)**

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

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 retriev
Her work primarily aims to improve how machines retrieve, understand, and r
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)**
   - 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)**

Not my paper

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

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 retriev
Her work primarily aims to improve how machines retrieve, understand, and i
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.
2. **"One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)**

Not my paper

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 retriev
Her work primarily aims to improve how machines retrieve, understand, and r
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.
2. **"One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)**

Not my paper

Venue is wrong

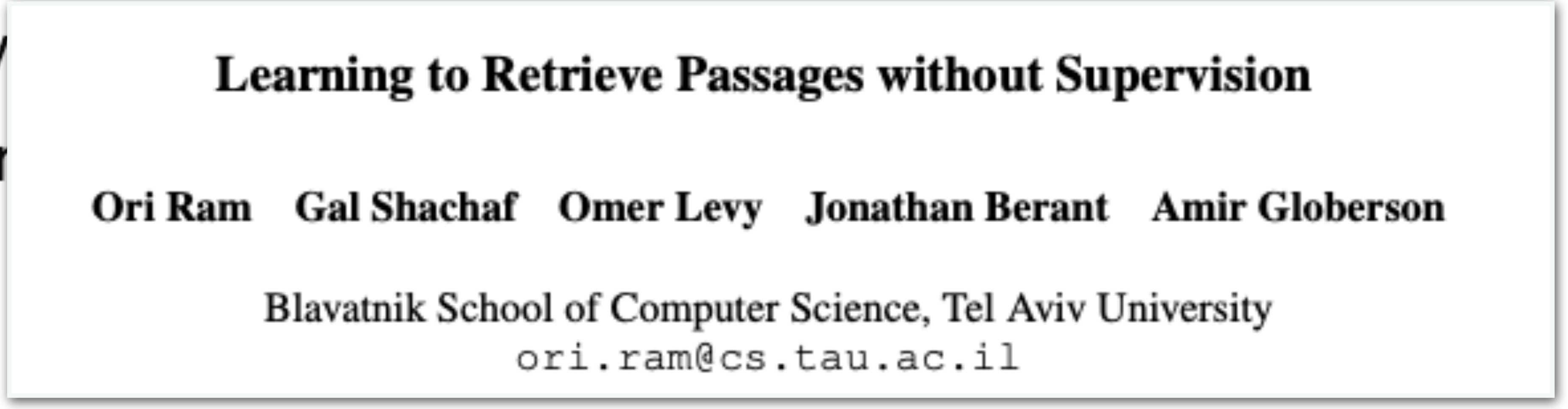
Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

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 retriev
Her work primarily aims to improve how machines retrieve, understand, and r
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.
2. **~~"One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)~~**

Not my paper

Venue is wrong

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

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 retriev… Her work primarily aims to improve how machines retrieve, understand, and r… 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.
2. ~~**"One Question, Many Answers: A Retrieval-based Multimodal QA Dataset" (EMNLP 2022)**~~

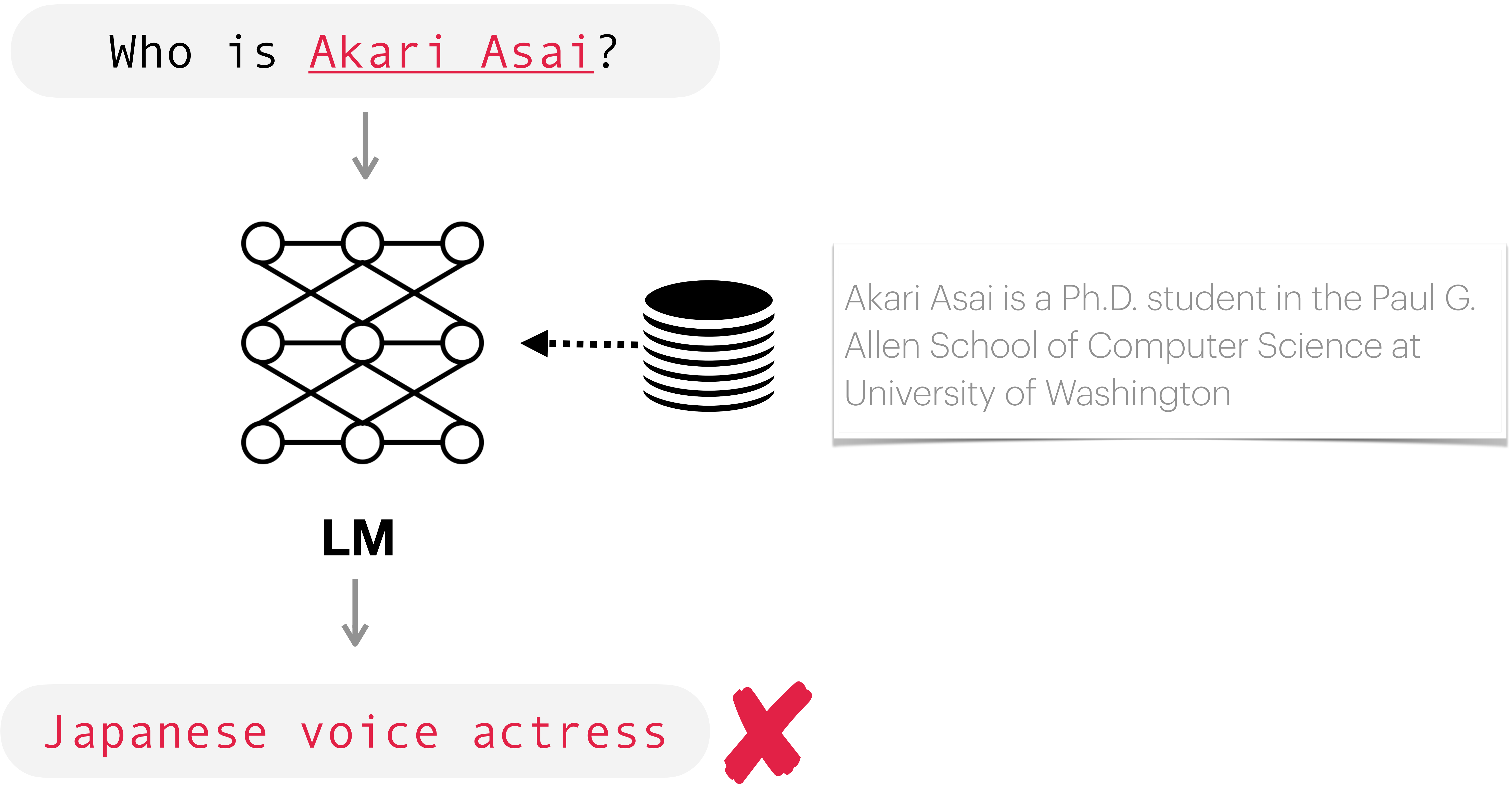
Not my paper

Venue is wrong

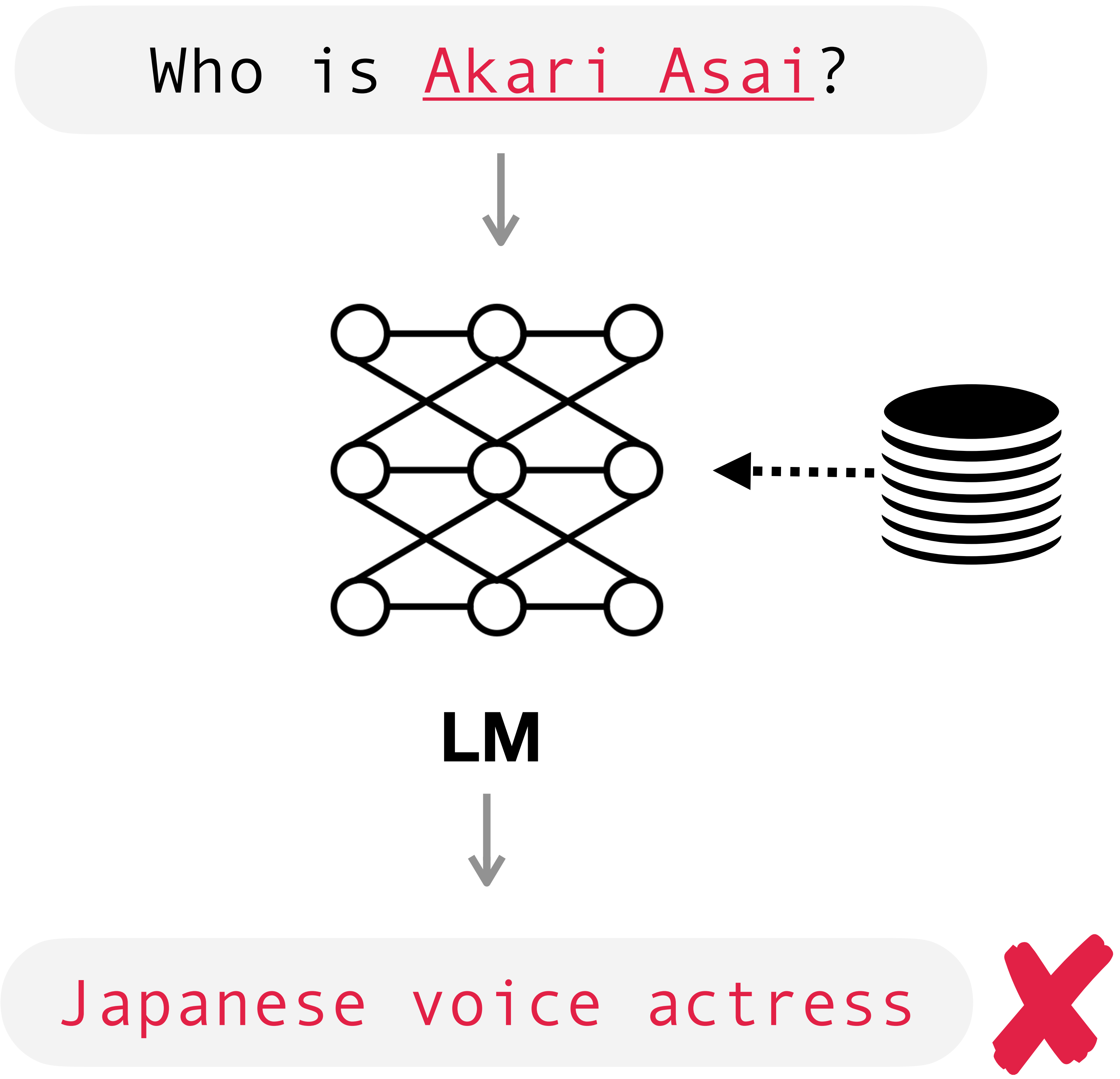
LMs struggle in long-tail knowledge

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

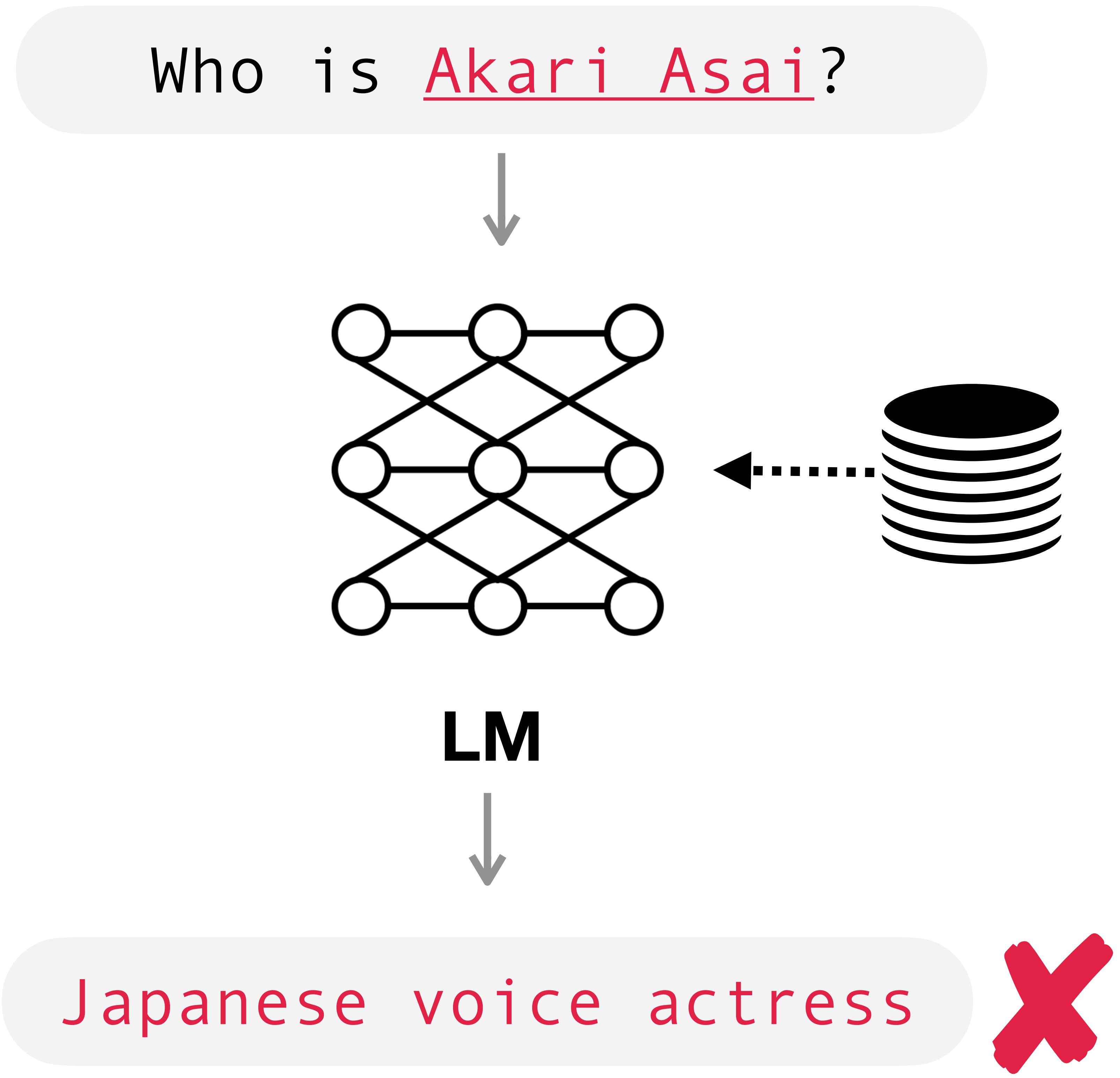
Retrieval-Augmented LMs: Intuition



Retrieval-Augmented LMs: Intuition



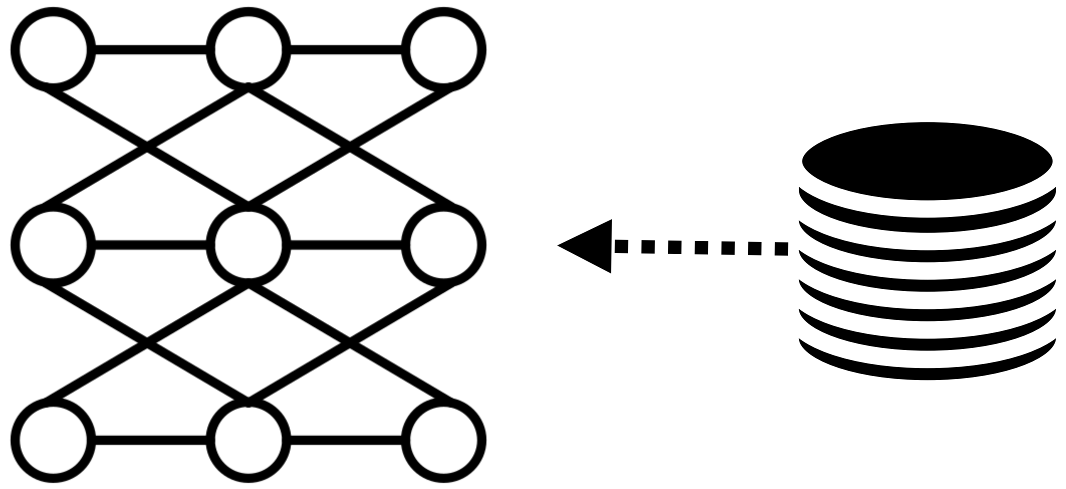
Retrieval-Augmented LMs: Intuition



Retrieval-Augmented LMs: Intuition

`Who is Akari Asai?`

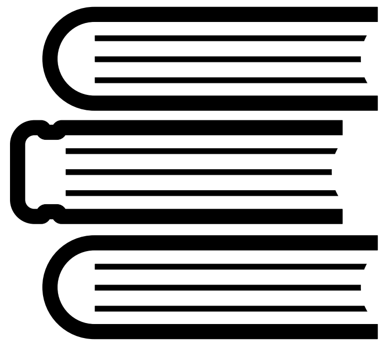
`Who is Akari Asai?`



`Japanese voice actress`

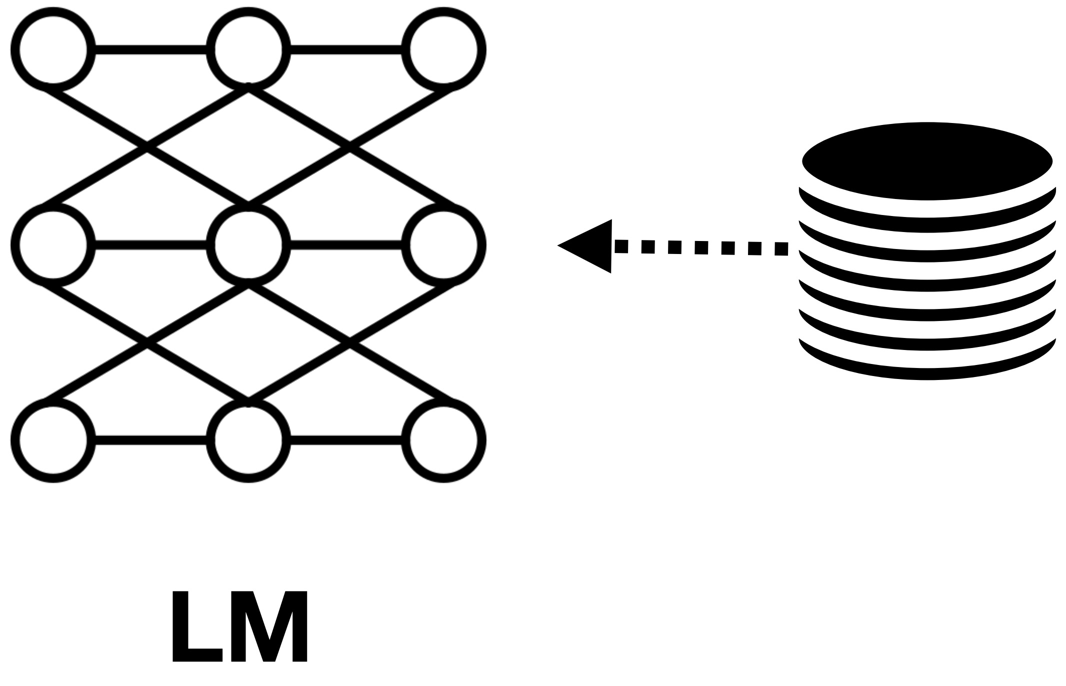
Retrieval-Augmented LMs: Intuition

`Who is Akari Asai?`



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

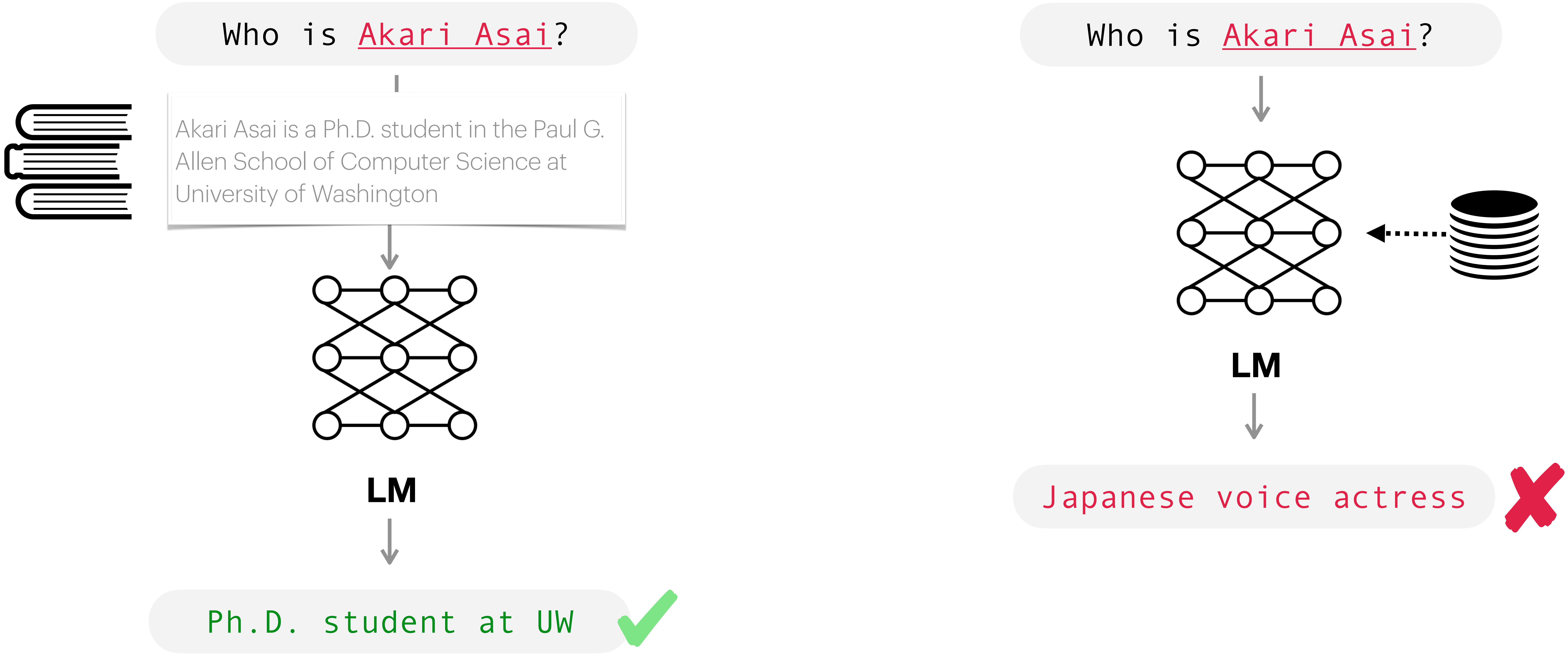
`Who is Akari Asai?`



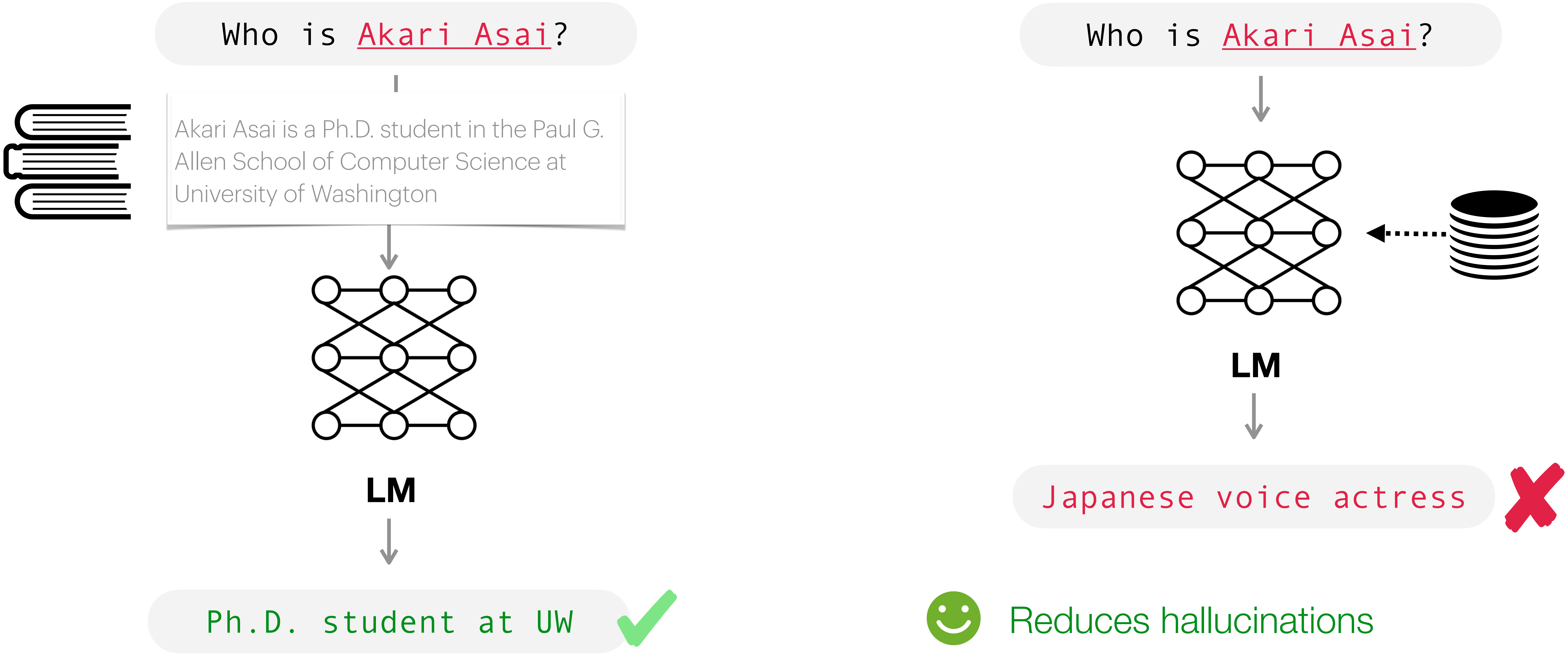
`Japanese voice actress`

Retrieval-Augmented LMs: Intuition

Retrieval-Augmented LMs: Intuition

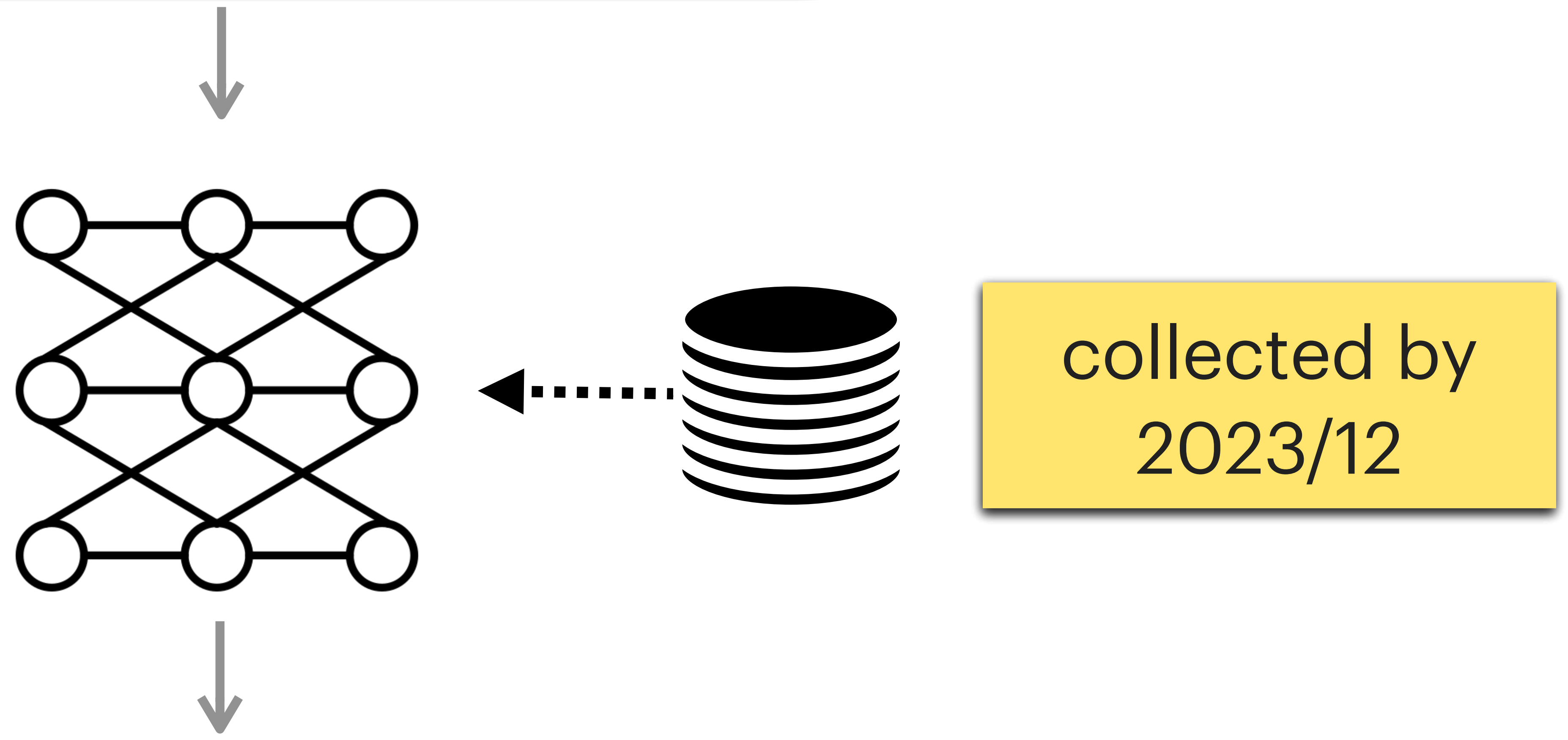


Retrieval-Augmented LMs: Intuition



Retrieval-Augmented LMs: Intuition

Who developed GPT4o?

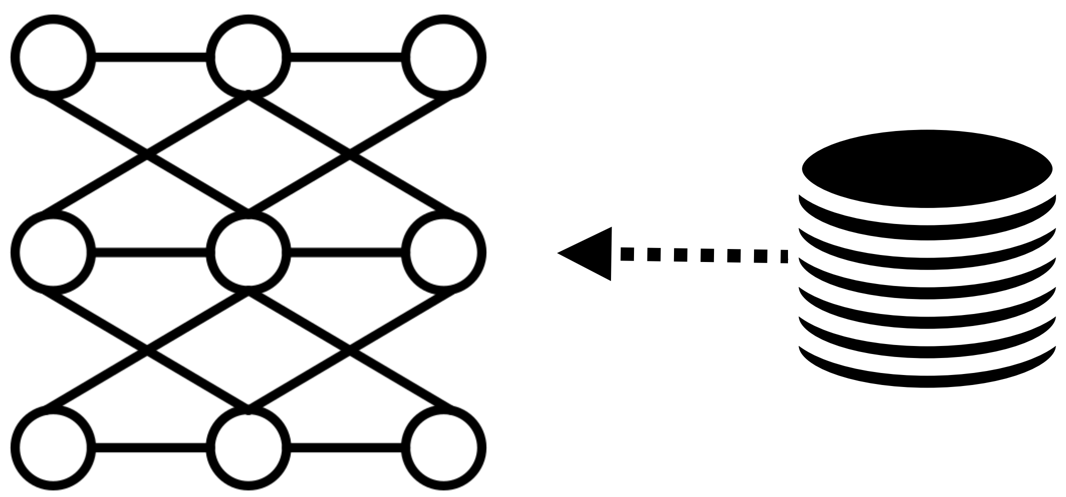


Based on general knowledge, there is no widely recognized AI model or term specifically called “GPT4o."

Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition

`Who developed GPT4o?`



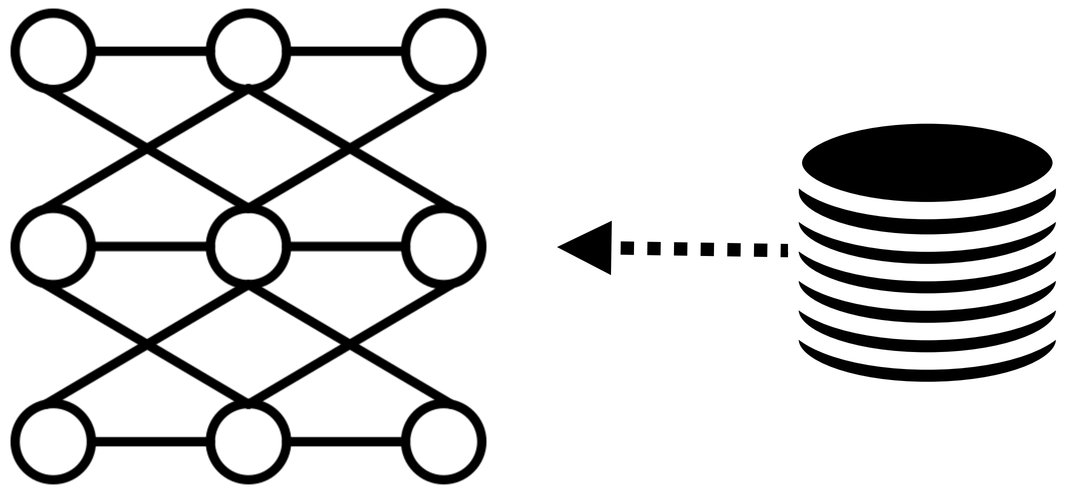
collected by 2023/12

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

Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition

Who developed GPT4o?

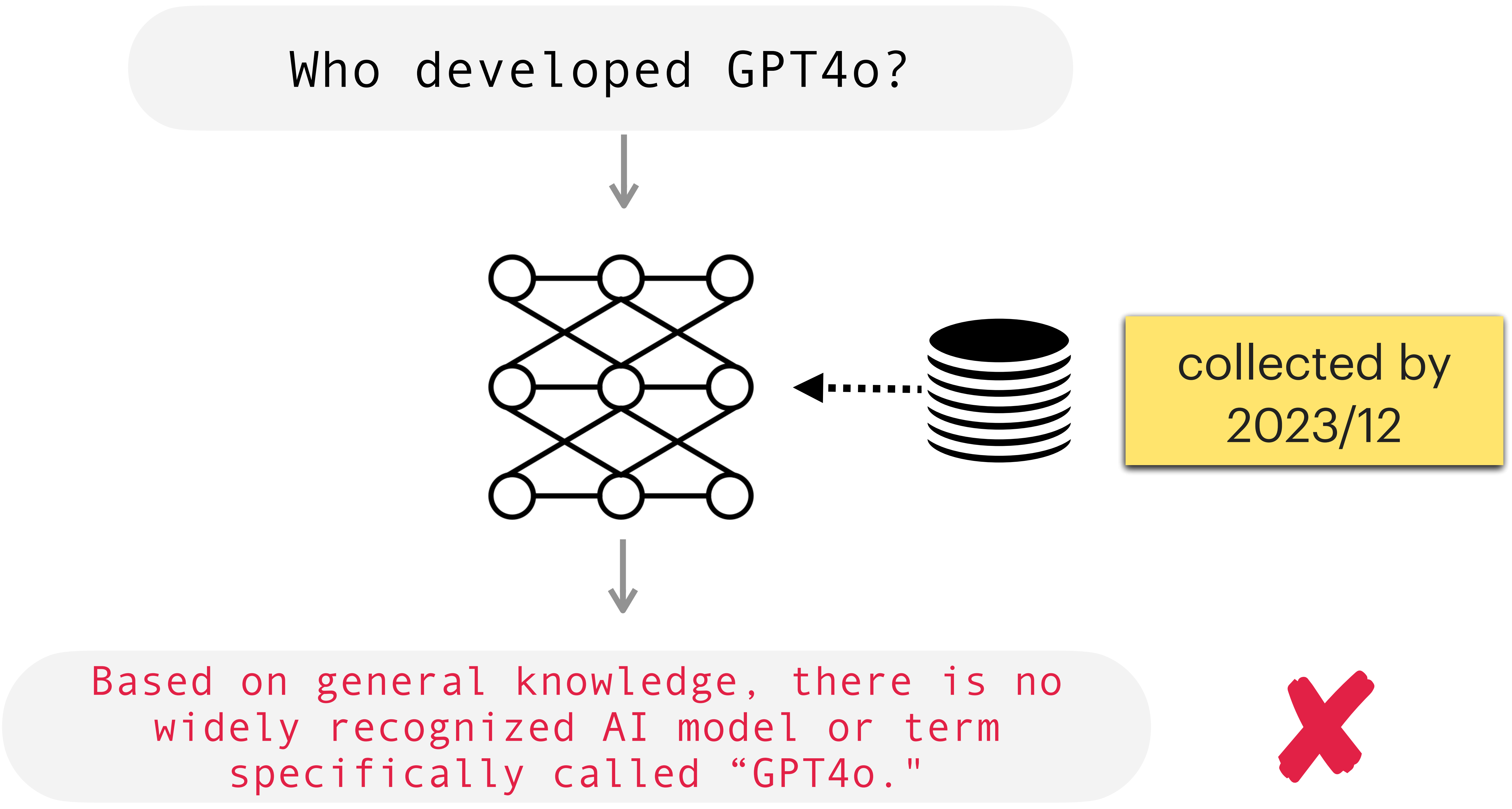


collected by 2023/12

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

Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition

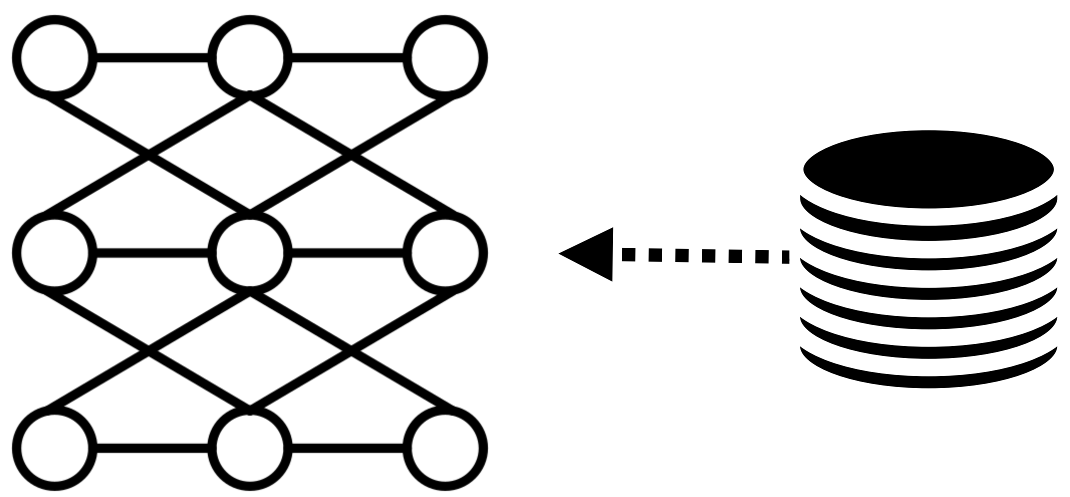


Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition

Who developed GPT4o?

Who developed GPT4o?



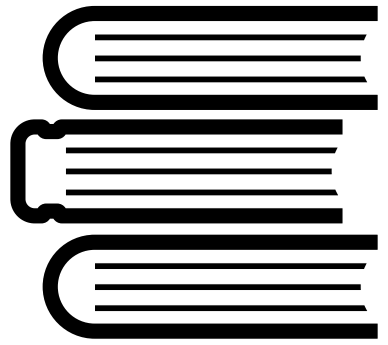
collected by 2023/12

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

Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition

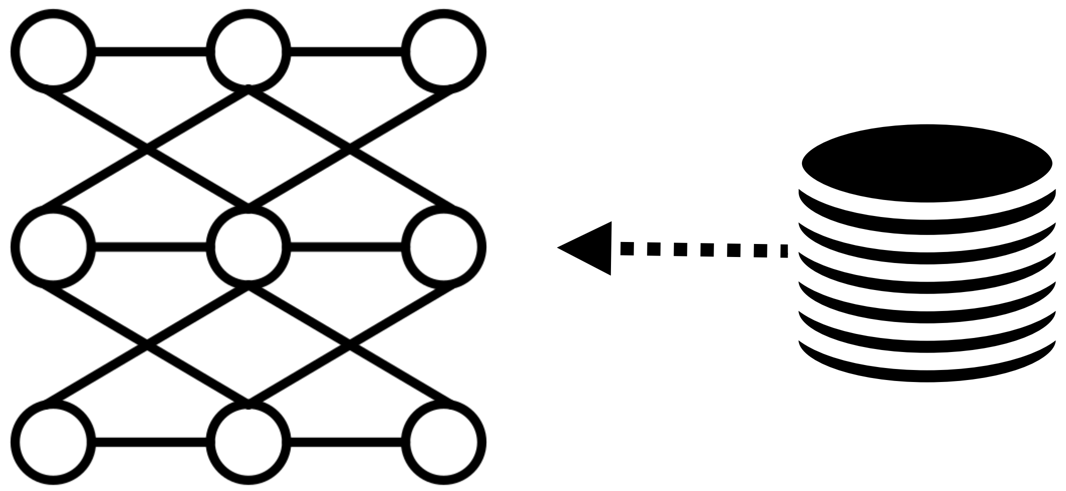
Who developed GPT4o?



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

Updated by 2024/01

Who developed GPT4o?

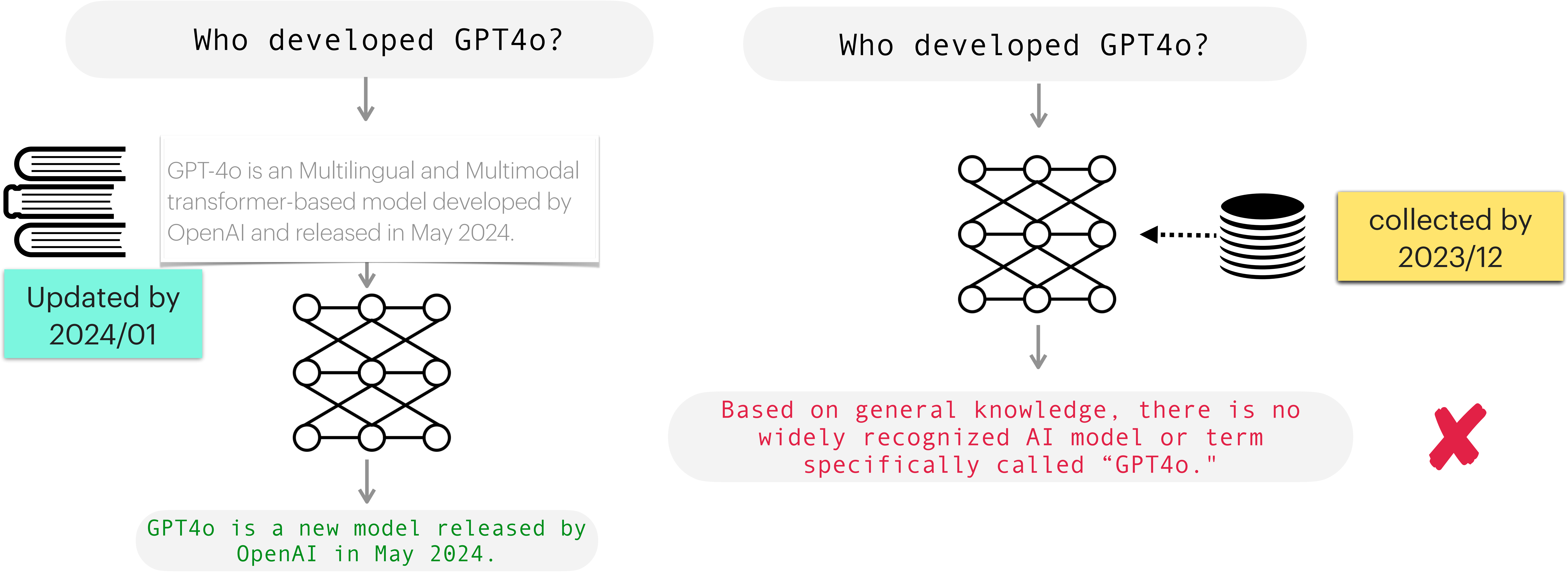


collected by 2023/12

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

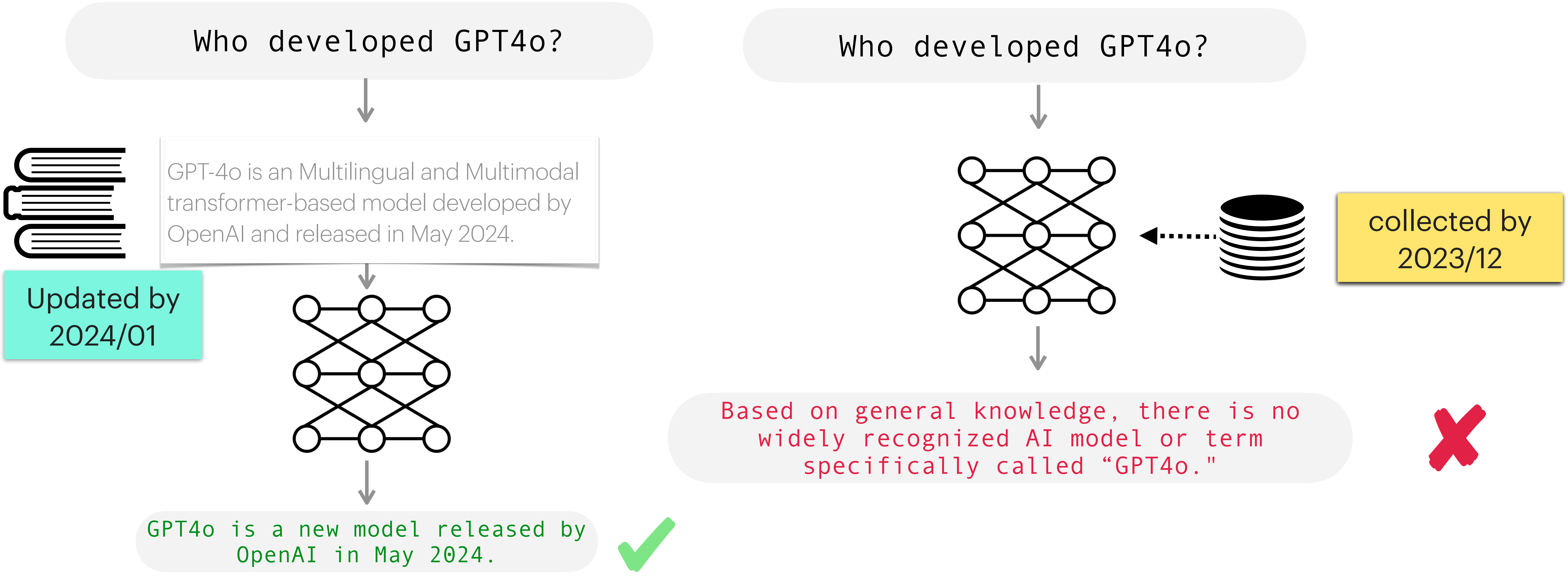
Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition



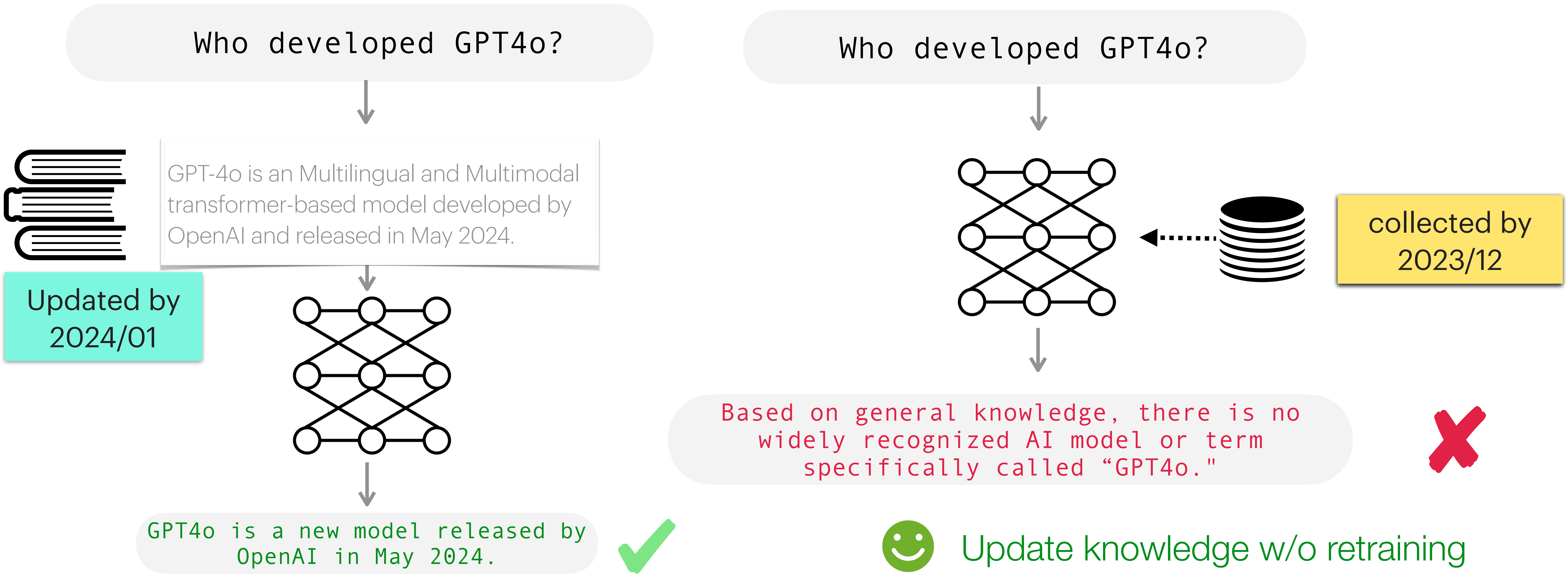
Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition



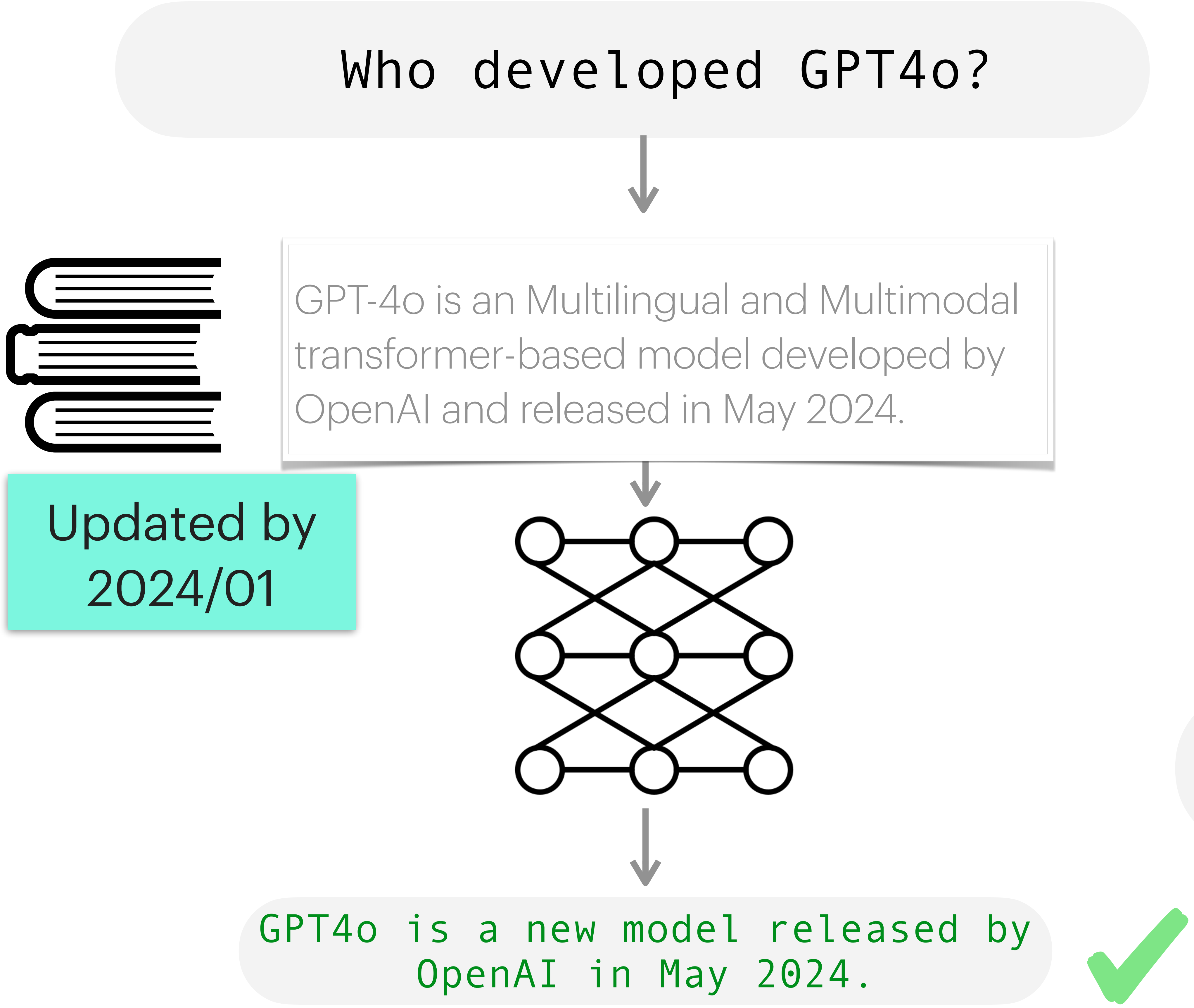
Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

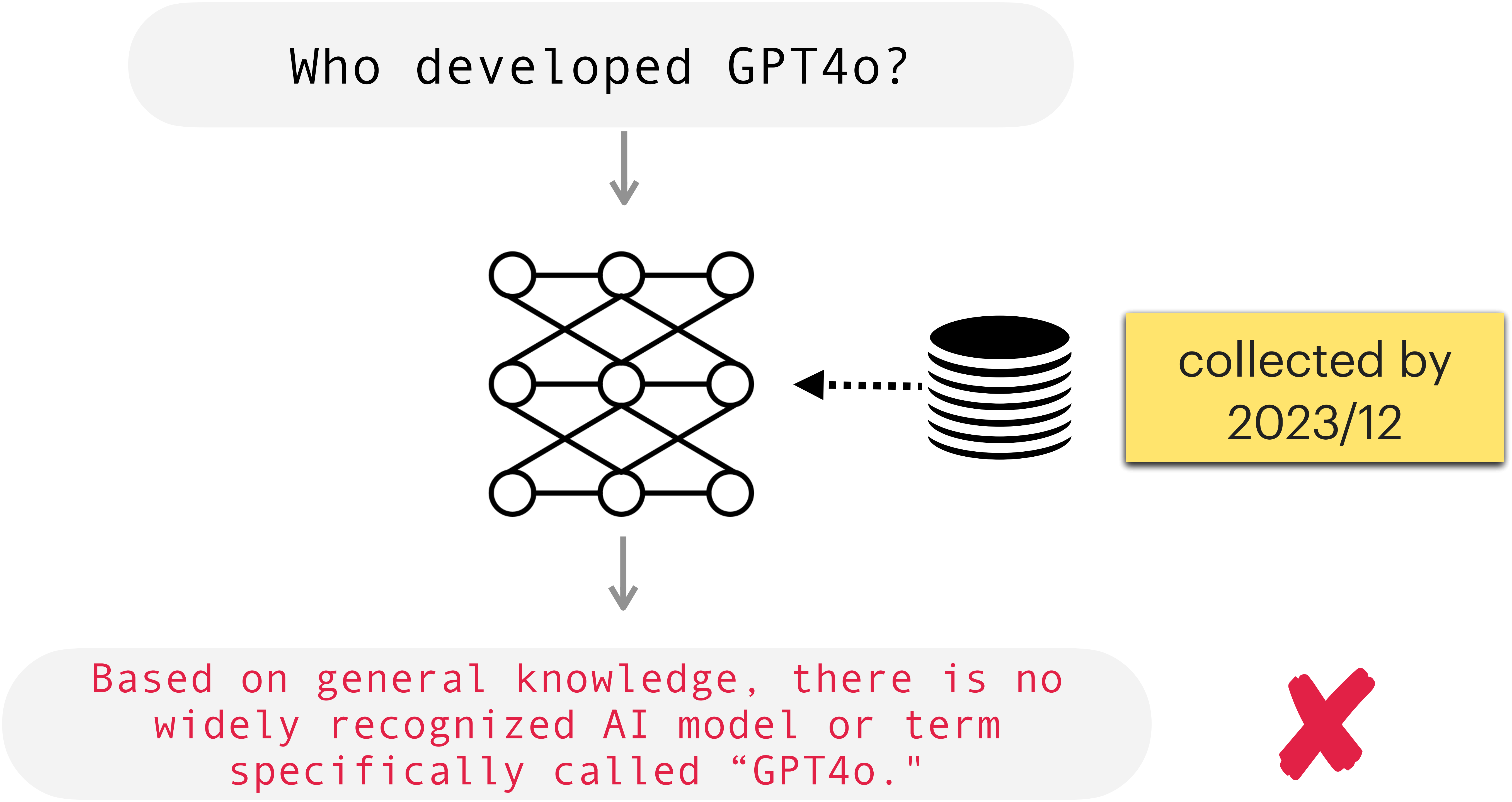
Retrieval-Augmented LMs: Intuition



Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Retrieval-Augmented LMs: Intuition



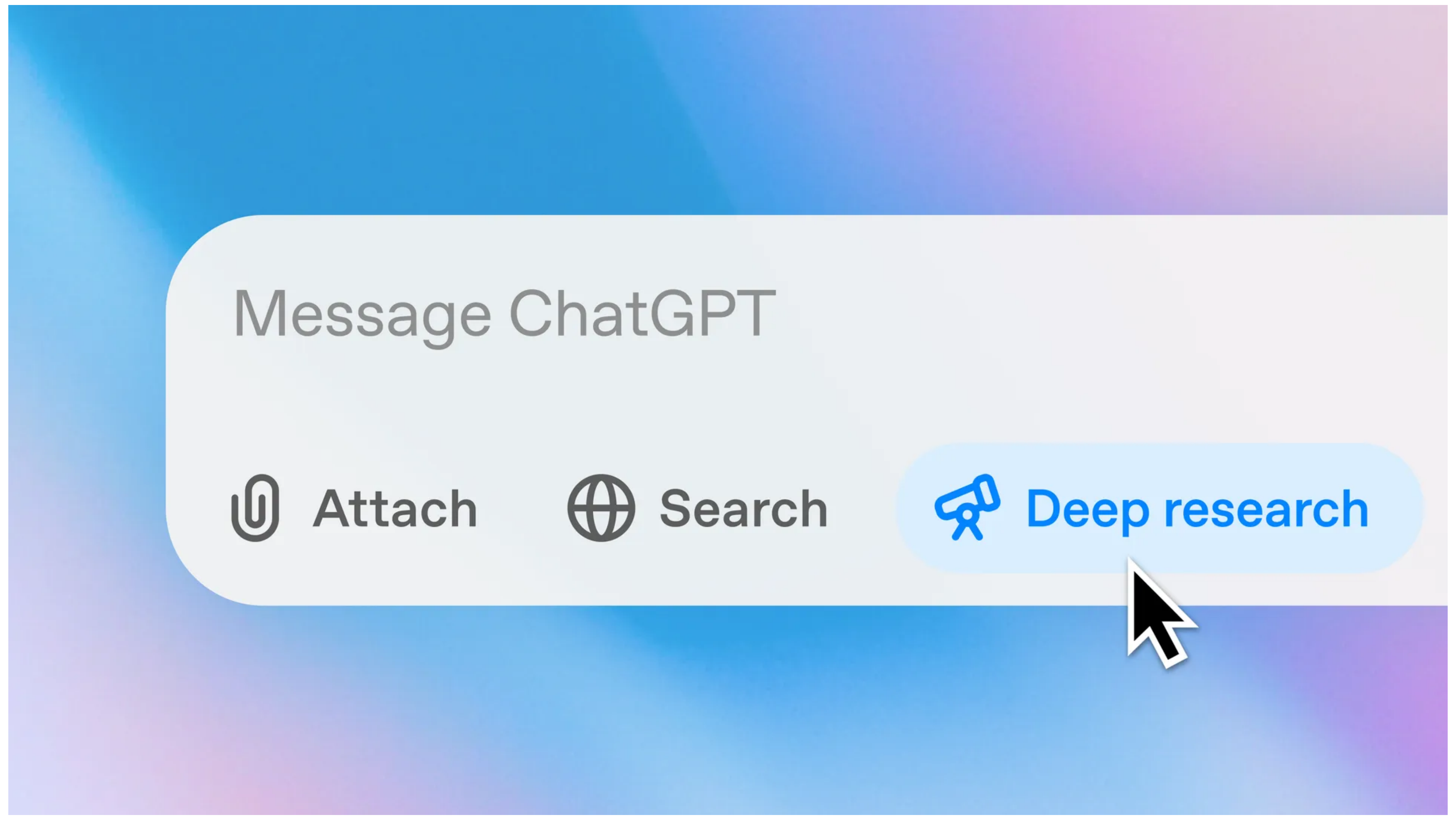


- Update knowledge w/o retraining
- Improve verifiability

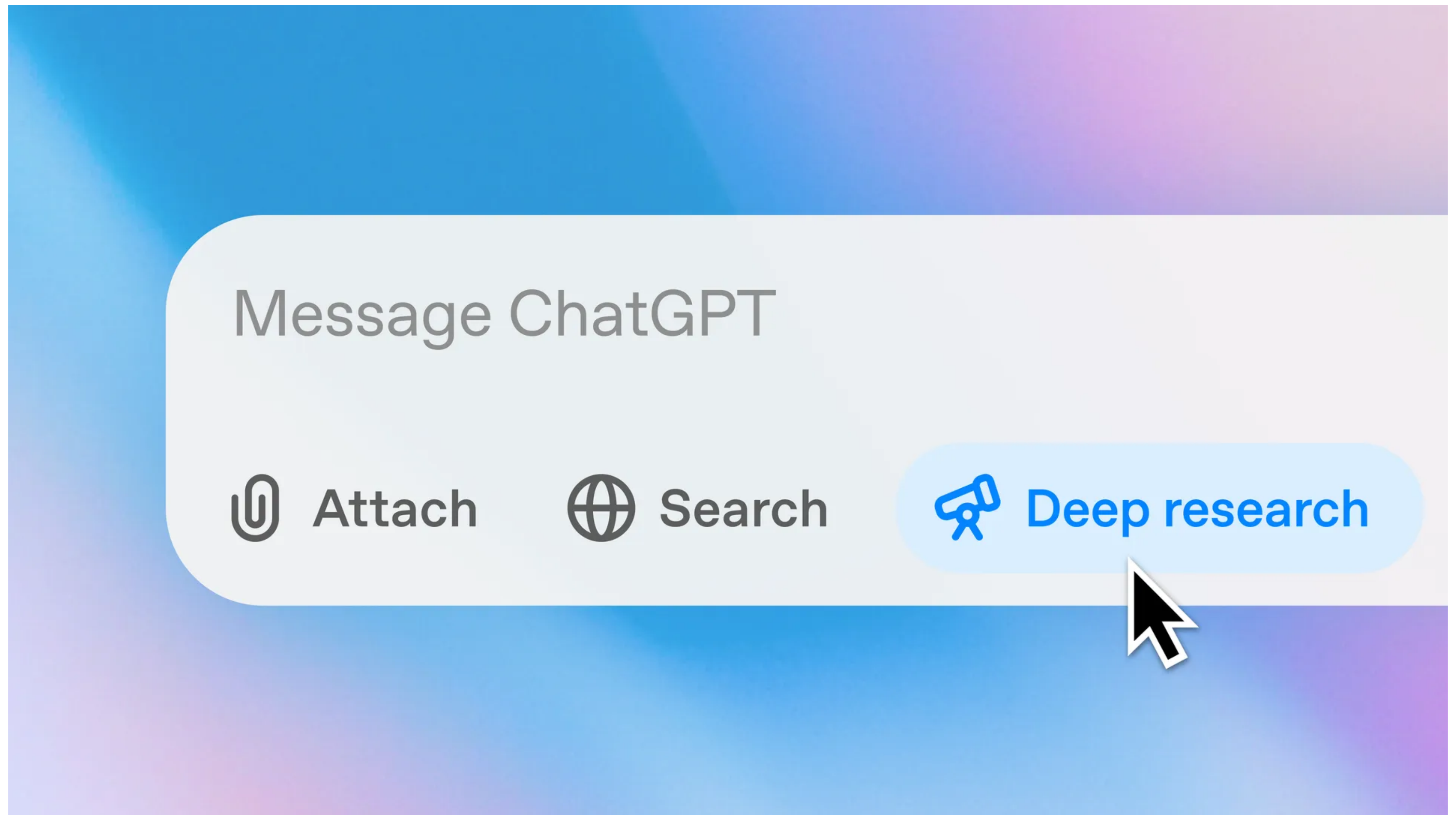
Kasai, … Asai et al. NeurIPS D&B 2023. RealTime QA: What's the Answer Right Now?

Widespread Adoptions in Real World

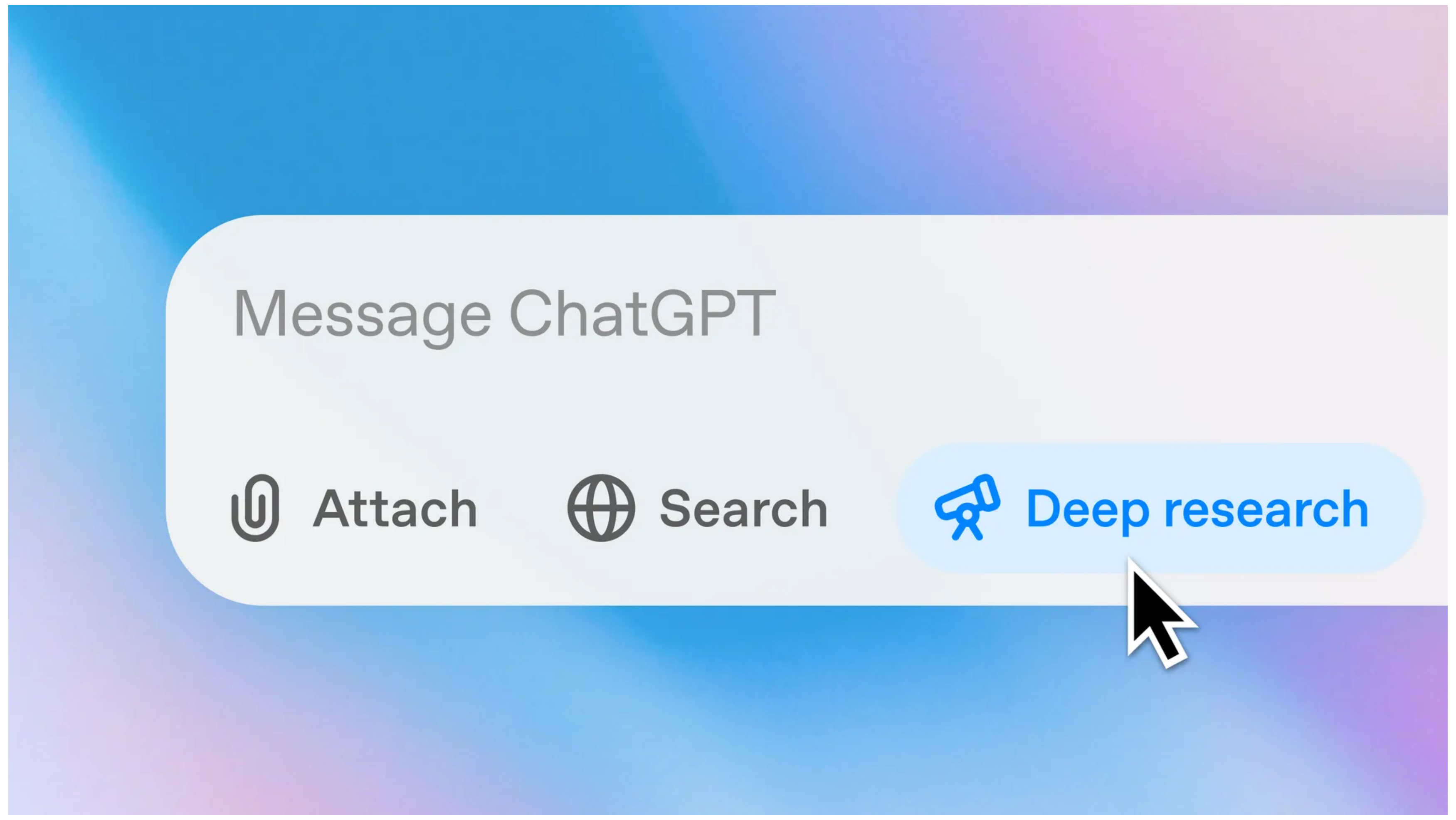
Widespread Adoptions in Real World



Widespread Adoptions in Real World



Widespread Adoptions in Real World



Perplexity

YOU

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

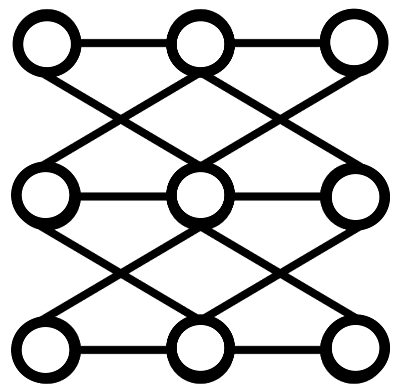
Subscribe About Archive BAIR

The Shift from Models to Compound AI Systems

At Databricks 60% of LLM applications use some form of retrieval-augmented generation (RAG)

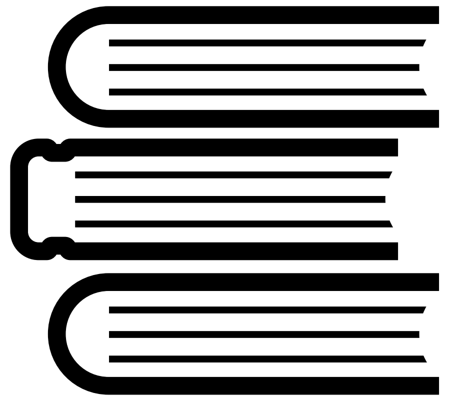
Retrieval-Augmented LMs: Overview

Retrieval-Augmented LMs: Overview



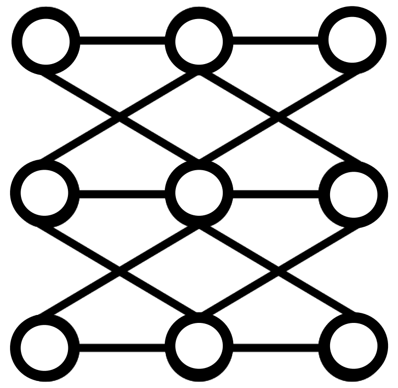
LM

Retrieval-Augmented LMs: Overview



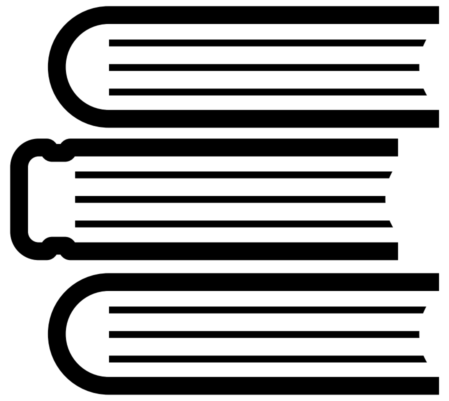
Datastore

Collections of a large number of documents



LM

Retrieval-Augmented LMs: Overview



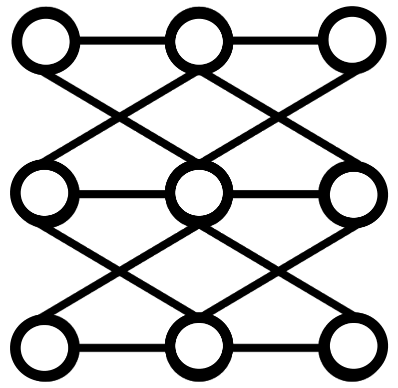
Datastore

Collections of a large number of documents

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

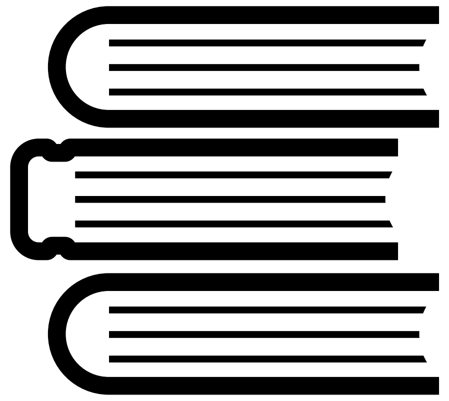
Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.



LM

Retrieval-Augmented LMs: Overview

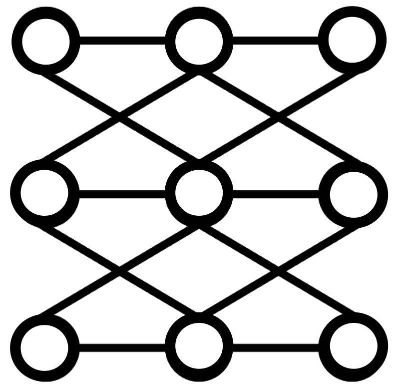


Datastore

Collections of a large number of documents

Retriever

Retrieve top *k* documents in datastore



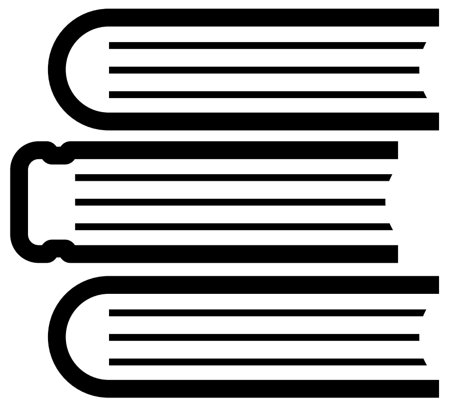
LM

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

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

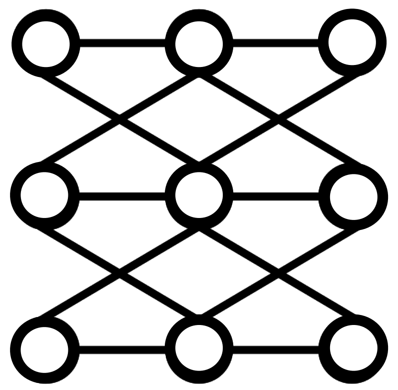
GPT4o was released by OpenAI in May 2024.

Retrieval-Augmented LMs: Overview



Datastore

Retriever



LM

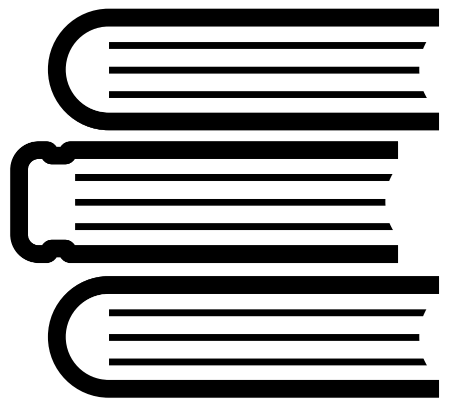
GPT-4o is a pre-trained transformer developed by OpenAI.

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

GPT4o was released by OpenAI in May 2024.

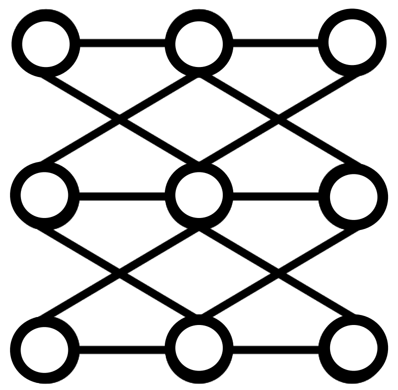
Retrieval-Augmented LMs: Overview

`x : Which company developed GPT4o?`



Datastore

Retriever



LM

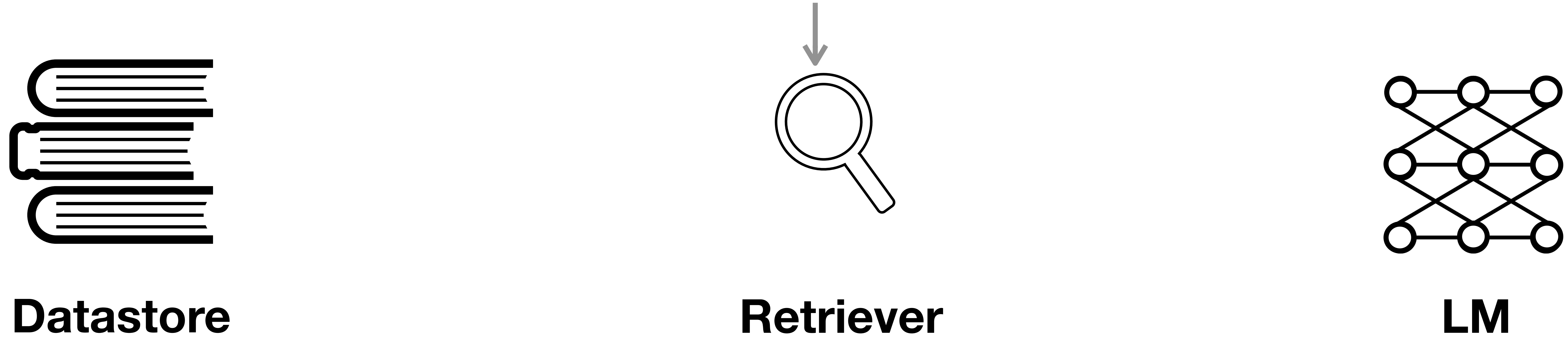
GPT-4o is a pre-trained transformer developed by OpenAI.

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?`



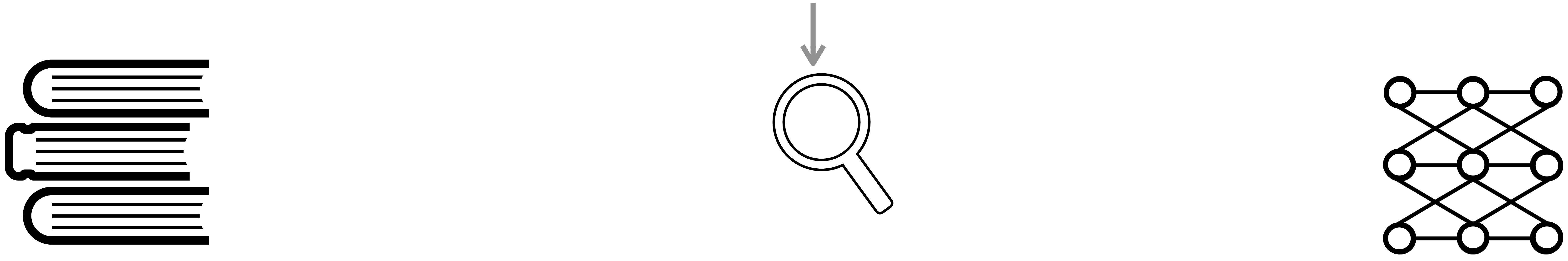
GPT-4o is a pre-trained transformer developed by OpenAI.

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?`



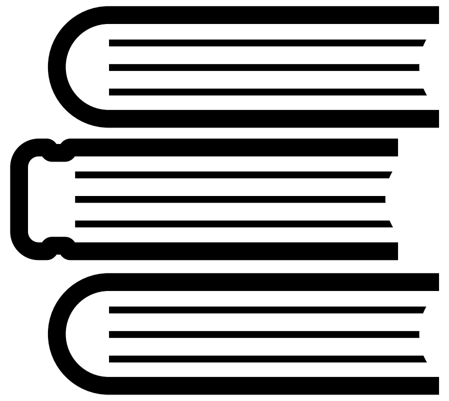
Datastore **Retriever** **LM**

$$\mathrm{Sim}(\cdot \mid x)$$

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

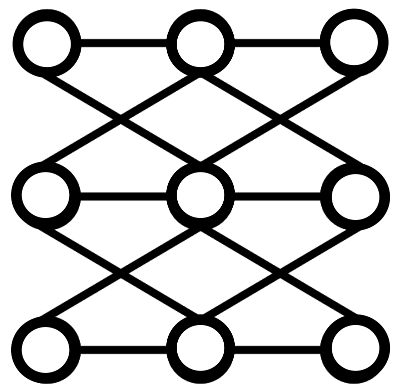
Retrieval-Augmented LMs: Overview

`x : Which company developed GPT4o?`



Datastore

Retriever



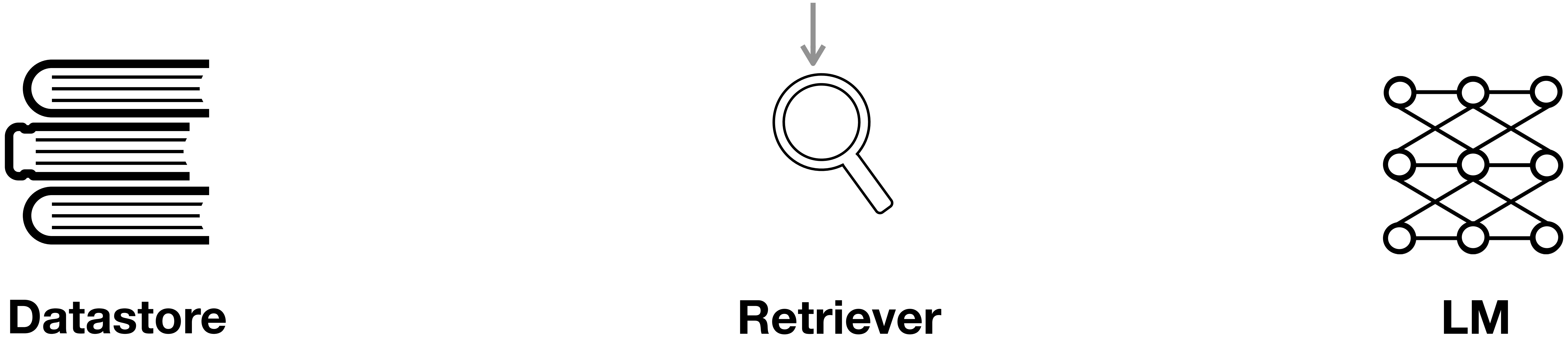
LM

$$D \in \mathrm{Top}_k \mathrm{Sim}(\,\cdot \mid x)$$

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?`



$$D \in \mathrm{Top}_k \mathrm{Sim}(\cdot \mid x)$$

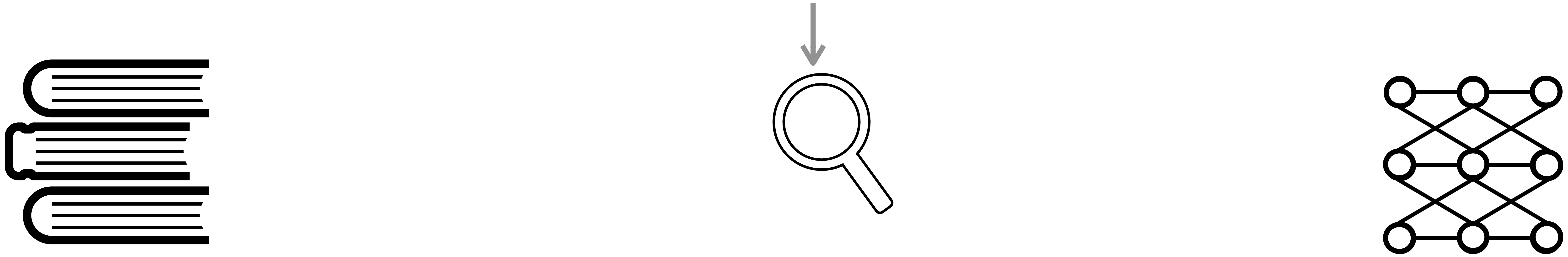
GPT-4o is a pre-trained transformer developed by OpenAI. 0.9

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**

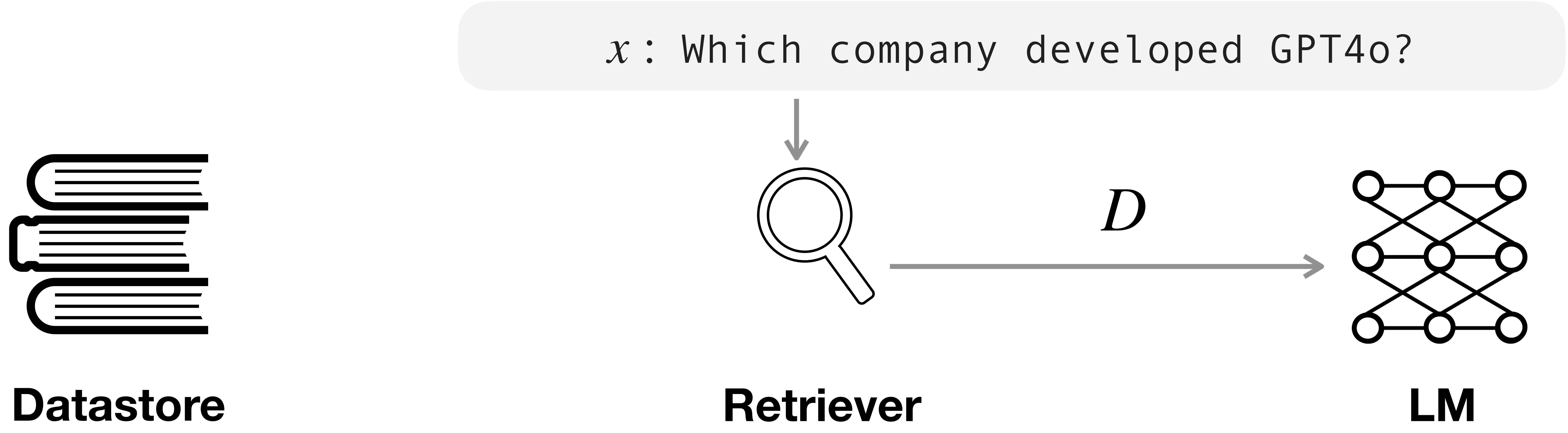
$$D \in \mathrm{Top}_k \mathrm{Sim}(\cdot \mid x)$$

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



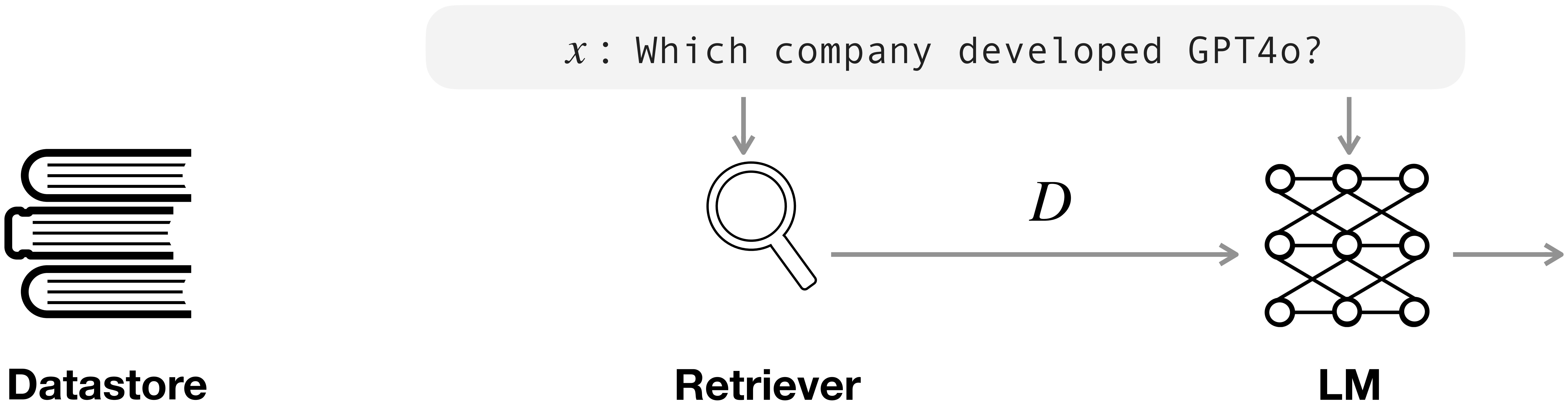
$$D \in \mathrm{Top}_k \mathrm{Sim}(\cdot \mid x)$$

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

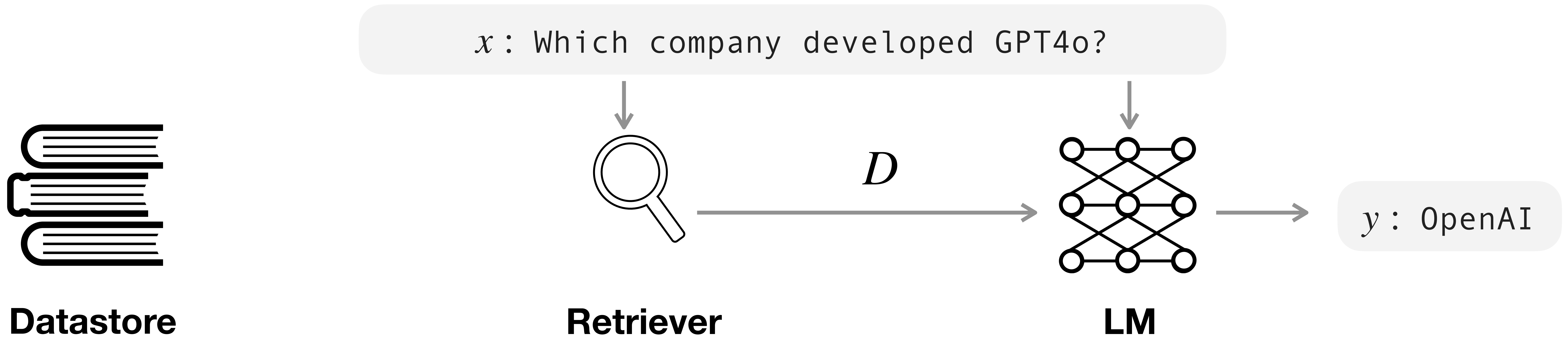
Retrieval-Augmented LMs: Overview



$$D \in \mathrm{Top}_k \mathrm{Sim}(\cdot \mid x)$$

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



$$D \in \mathrm{Top}_k \mathrm{Sim}(\cdot \mid x)$$

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

0.9

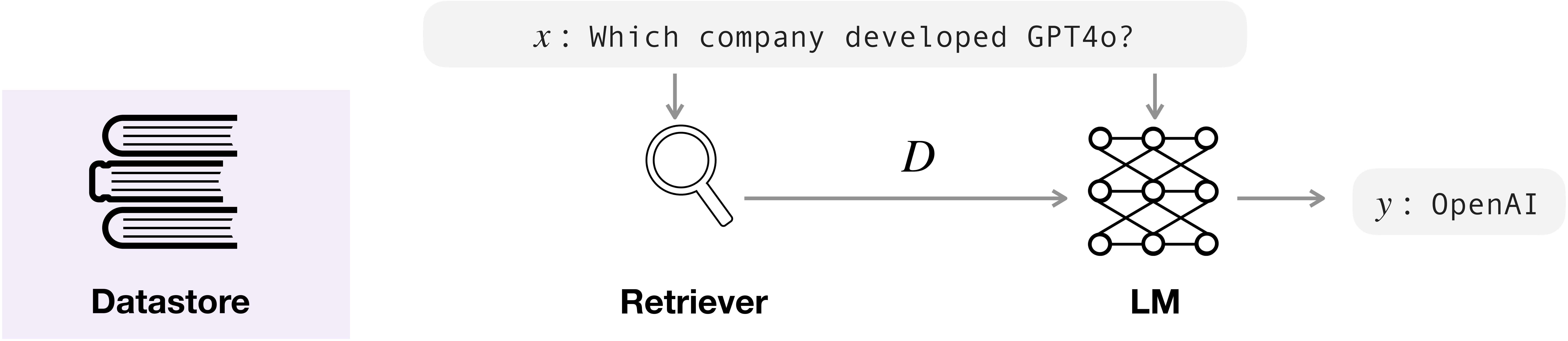
Transformers is a series of science fiction action films based on the Transformers franchise.

0.1

GPT4o was released by OpenAI in May 2024.

0.8

Today's Outline

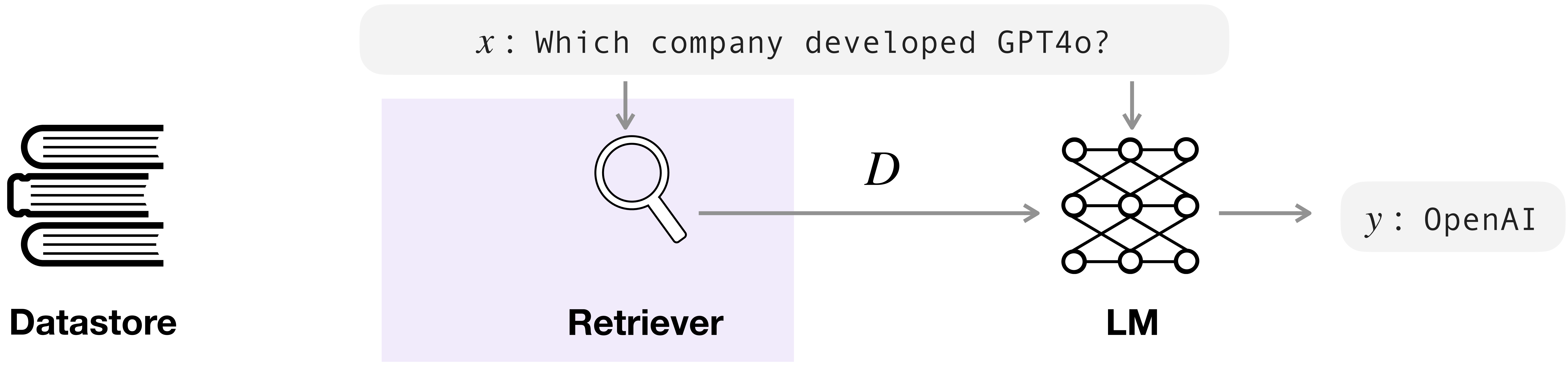


✓ Sources of datastore

✓ Processing

✓ Scaling

Today's Outline

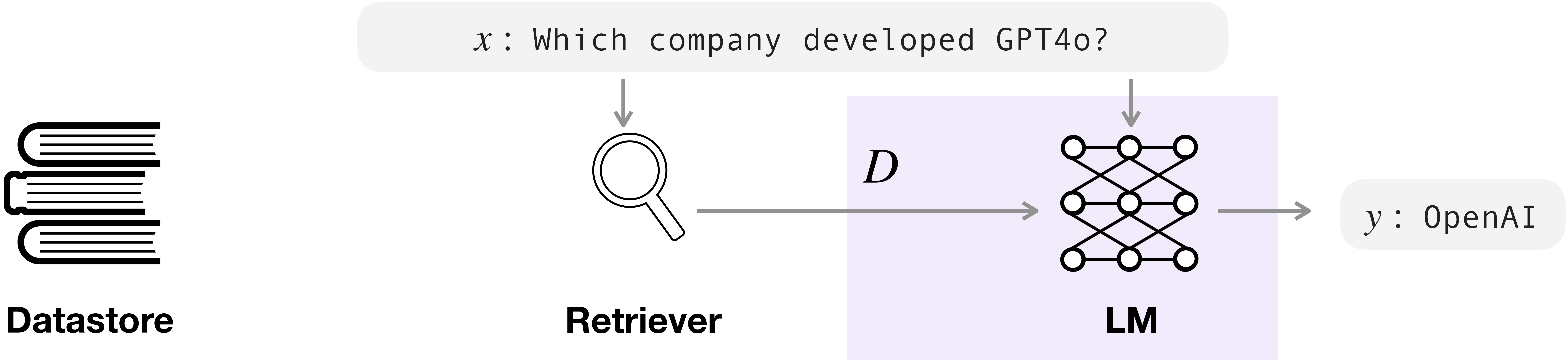


✓ Types of retrievers

✓ Training

✓ Evaluations

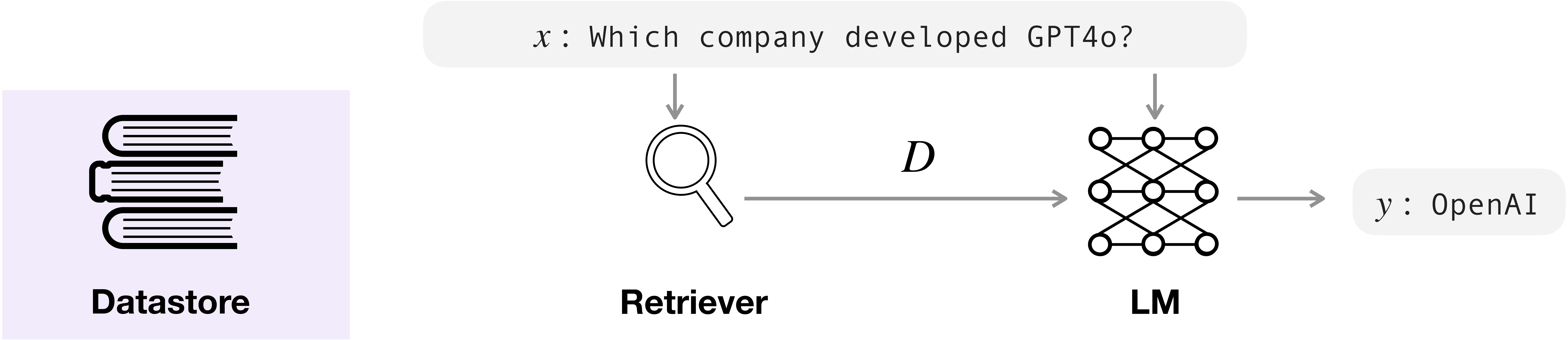
Today's Outline



✓ Common architectures

✓ Recent progress in RAG

Today's Outline



- ✓ Sources of datastore
- ✓ Processing
- ✓ Scaling

What Should Be Used as "Datastore"?

Wang*, Asai* et al. Findings of NAACL 2025. CodeRAG-Bench: Can Retrieval Augment Code Generation?.

What Should Be Used as "Datastore"?

x : `Which company developed GPT4o?`

Wang*, Asai* et al. Findings of NAACL 2025. CodeRAG-Bench: Can Retrieval Augment Code Generation?.

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/

Wang*, Asai* et al. Findings of NAACL 2025. CodeRAG-Bench: Can Retrieval Augment Code Generation?.

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/

Wang*, Asai* et al. Findings of NAACL 2025. CodeRAG-Bench: Can Retrieval Augment Code Generation?.

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

stackoverflow

Wang*, Asai* et al. Findings of NAACL 2025. CodeRAG-Bench: Can Retrieval Augment Code Generation?.

Massively Scaling Datastore

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Massively Scaling Datastore

English Wikipedia
5B tokens

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Massively Scaling Datastore

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

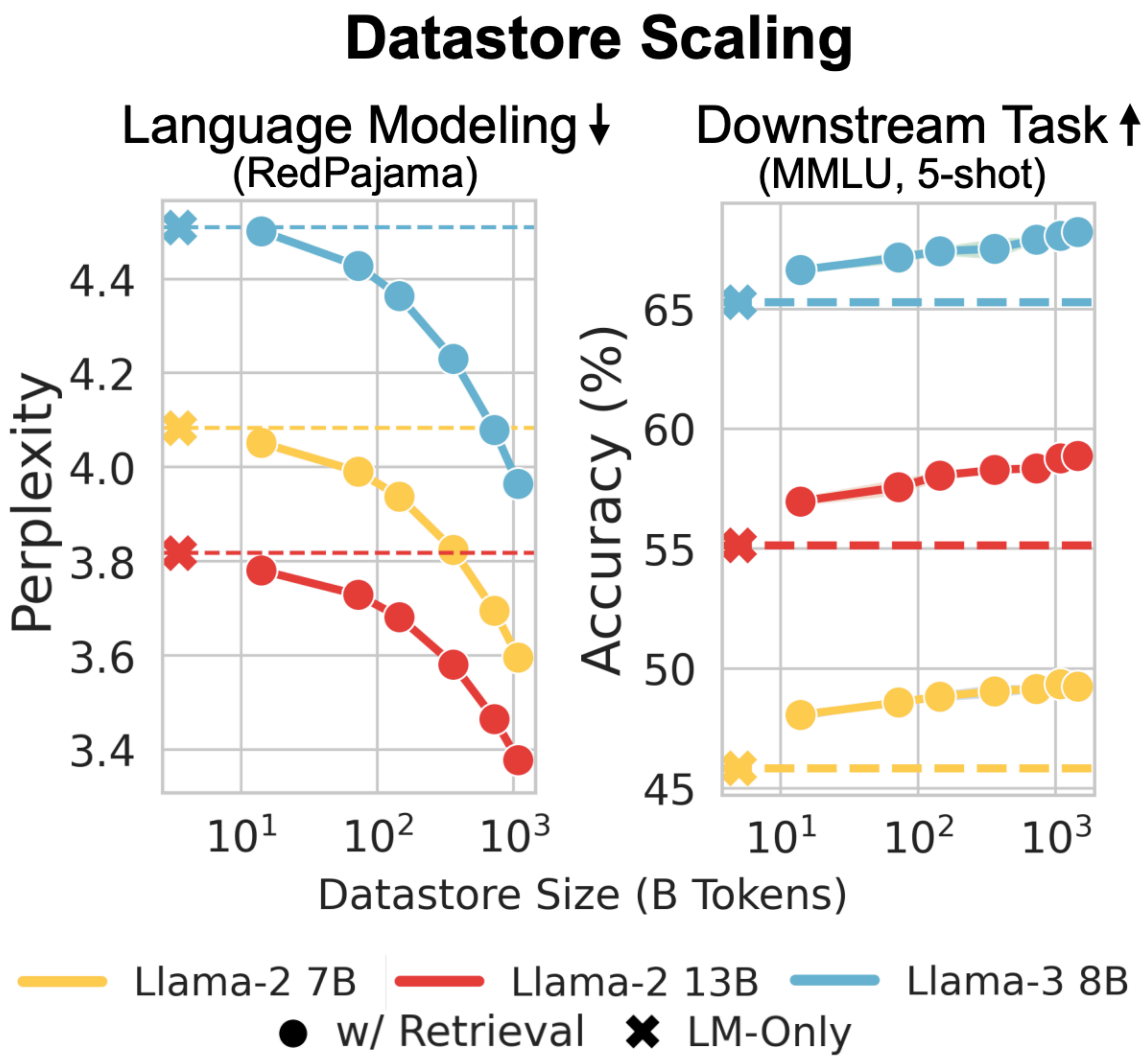
Massively Scaling Datastore

MassiveDS

1.4 trillion tokens (22TB)

English Wikipedia
5B tokens

Massively Scaling Datastore



Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Processing Documents

GPT-4

32 languages

Article Talk Read Edit View history Tools

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

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

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

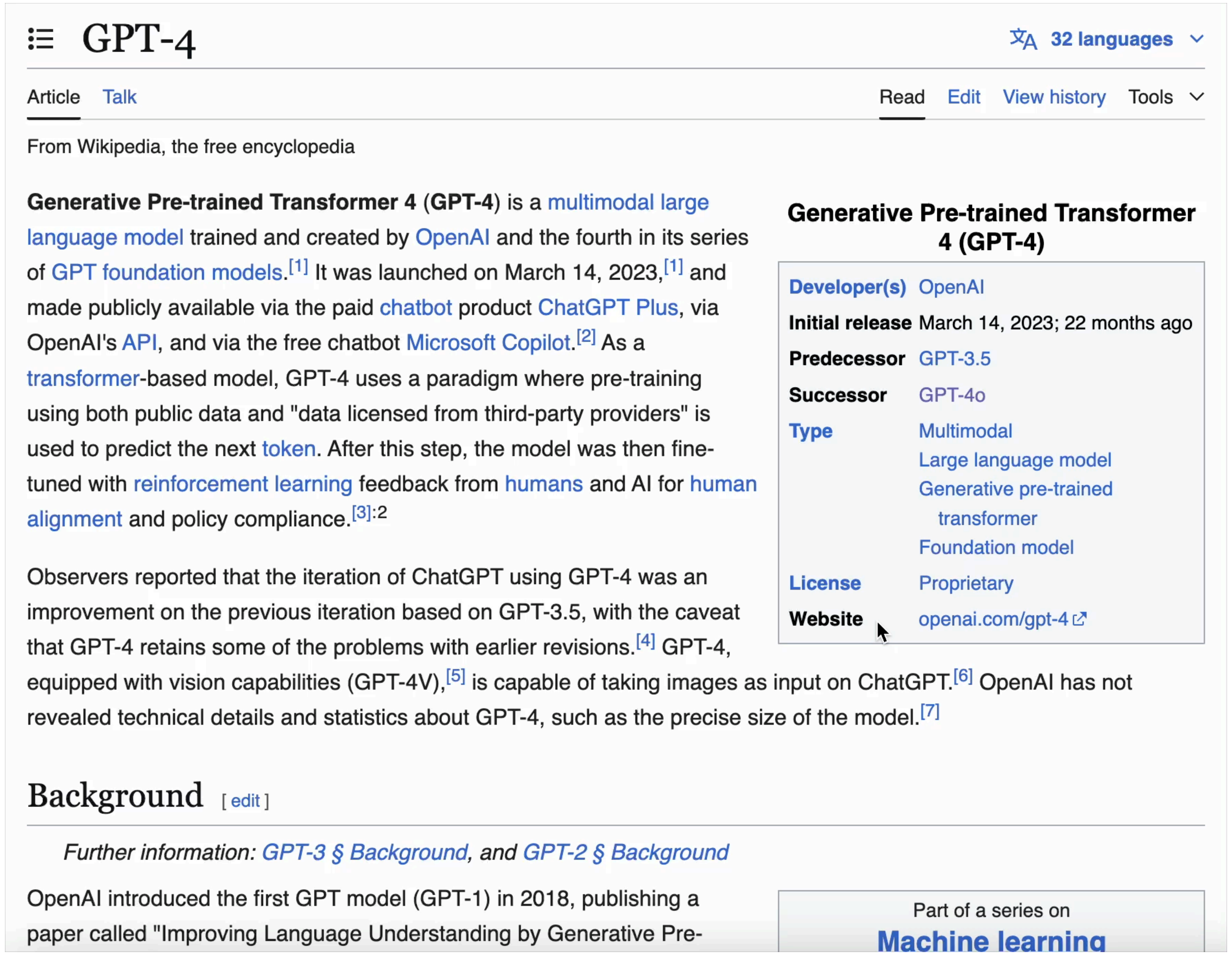
OpenAI introduced the first GPT model (GPT-1) in 2018, publishing a paper called "Improving Language Understanding by Generative Pre-

Part of a series on

Machine learning

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

Processing Documents



GPT-4

32 languages

Article Talk | Read Edit View history Tools

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

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

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-

Part of a series on
Machine learning

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

Processing Documents

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Processing Documents

Curate and preprocess data

e.g., HTML -> Plain text

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Processing Documents

Curate and preprocess data

e.g., HTML -> Plain text


GPT-4

Article Talk

OpenAI

Article Talk

From Wikipedia, the free encyclopedia

Not to be confused with OpenAL, OpenAPI, or Open-source artificial inte

OpenAI is an American Artificial Intelligence (AI) research organization founded in December 2015 and headquartered in San Francisco, California. Its stated mission is to develop "safe and beneficial" artificial general intelligence (AGI), which it defines as "highly autonomous systems that outperform humans at most economically valuable work".[5] As a leading organization in the ongoing AI boom,[6] OpenAI is known for the GPT family of large language models, the DALL-E series of text-to-image models, and a text-to-video model named Sora.[7][8] Its release of ChatGPT in November 2022 has been credited with catalyzing widespread interest in generative AI.

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Processing Documents

Curate and preprocess data

e.g., HTML -> Plain text


GPT-4

Article Talk

OpenAI

Article Talk

From Wikipedia, the free encyclopedia

Not to be confused with OpenAL, OpenAPI, or Open-source artificial inte

OpenAI is an American Artificial Intelligence (AI) research organization founded in December 2015 and headquartered in San Francisco, California. Its stated mission is to develop "safe and beneficial" artificial general intelligence (AGI), which it defines as "highly autonomous systems that outperform humans at most economically valuable work".[5] As a leading organization in the ongoing AI boom,[6] OpenAI is known for the GPT family of large language models, the DALL-E series of text-to-image models, and a text-to-video model named Sora.[7][8] Its release of ChatGPT in November 2022 has been credited with catalyzing widespread interest in generative AI.

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@

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Processing Documents

Curate and preprocess data → **Chunking** →

e.g., HTML -> Plain text


GPT-4

Article Talk

From Wikipedi

Generative
language mo
of GPT foun
made public
OpenAI's AP
transformer-
using both p
used to pred
tuned with re
alignment ar

OpenAI

Article Talk

From Wikipedia, the free encyclopedia

Not to be confused with OpenAL, OpenAPI, or Open-source artificial inte

OpenAI is an American Artificial Intelligence (AI) research organization founded in December 2015 and headquartered in San Francisco, California. Its stated mission is to develop "safe and beneficial" artificial general intelligence (AGI), which it defines as "highly autonomous systems that outperform humans at most economically valuable work".[5] As a leading organization in the ongoing AI boom,[6] OpenAI is known for the GPT family of large language models, the DALL-E series of text-to-image models, and a text-to-video model named Sora.[7][8] Its release of ChatGPT in November 2022 has been credited with catalyzing widespread interest in generative AI.

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@

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

Processing Documents

Curate and preprocess data → **Chunking** → **Post-processing**

e.g., HTML -> Plain text


GPT-4

Article Talk

OpenAI

Article Talk

From Wikipedia, the free encyclopedia

Not to be confused with OpenAL, OpenAPI, or Open-source artificial inte

OpenAI is an American Artificial Intelligence (AI) research organization founded in December 2015 and headquartered in San Francisco, California. Its stated mission is to develop "safe and beneficial" artificial general intelligence (AGI), which it defines as "highly autonomous systems that outperform humans at most economically valuable work".[5] As a leading organization in the ongoing AI boom,[6] OpenAI is known for the GPT family of large language models, the DALL-E series of text-to-image models, and a text-to-video model named Sora.[7][8] Its release of ChatGPT in November 2022 has been credited with catalyzing widespread interest in generative AI.

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@

e.g., Remove short documents

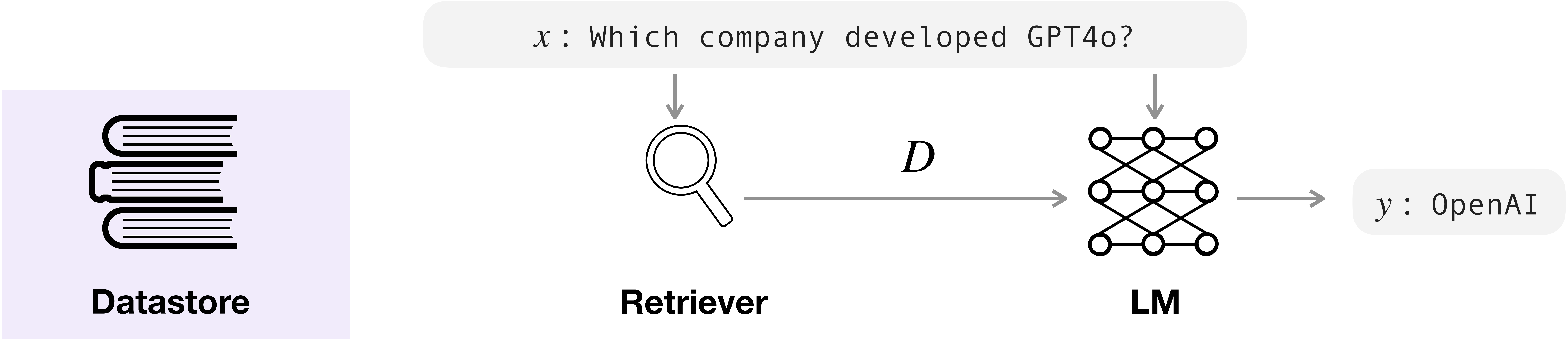
GPT-4o is a pre-trained transformer developed by OpenAI.

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

GPT4o was released by OpenAI in May 2024.

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-based Language Models with a Trillion-token Datastore

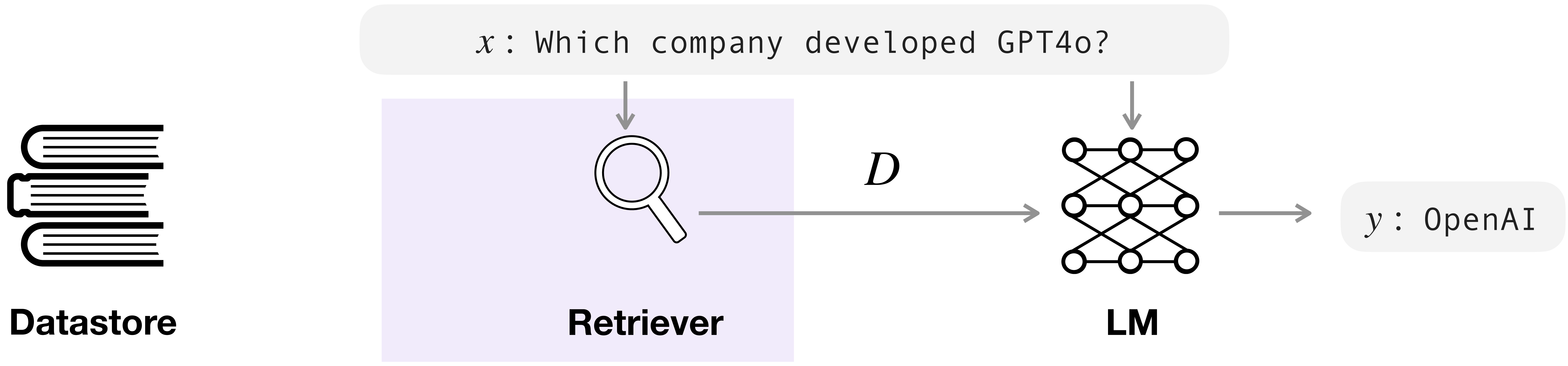
Summary of Part 1



- ✓ 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



- ✓ Types of retrievers
- ✓ Training
- ✓ Evaluations

Types of Retrievers

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

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

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

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

$$D \in \mathrm{Top}_k \mathrm{Sim}(\cdot \mid x)$$

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

$$D \in \mathrm{Top}_k \mathrm{Sim}(\,\cdot \mid x)$$

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

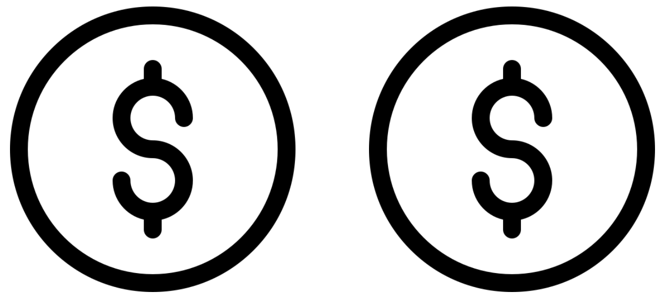
- **Sim:** Term-frequency based embeddings
- Training is not required

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

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

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

	q=what is nlp	d_1 = what is life ? candy is life !	d_2 = nlp is an acronym for natural language processing	d_3 = I like to do good research on nlp
what	1	1	0	0
candy	0	1	0	0
nlp	1	0	1	1
is	1	1	1	0
language	0	0	0	0
life	0	1	0	0
…	…	…	…	…

Sparse Retrievers: One-hot Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

$$\begin{array}{c} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \text{life} \\ \dots \end{array} \quad \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \dots \end{pmatrix} \quad \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ \dots \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \dots \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \dots \end{pmatrix}$$

Check if a term appears in a document

Sparse Retrievers: Term-count Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

$$\begin{array}{c} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \text{life} \\ \ldots \end{array} \quad \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \ldots \end{pmatrix} \quad \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 2 \\ \ldots \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \ldots \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \ldots \end{pmatrix}$$

Sparse Retrievers: Term-count Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

$$
\begin{array}{c}
\text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \text{life} \\ \cdots
\end{array}
\quad
\begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \cdots \end{pmatrix}
\quad
\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 2 \\ \cdots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ \cdots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \cdots \end{pmatrix}
$$

Count the number of appearances in a doc

Sparse Retrievers: Computing Weighted Term Scores

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

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

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

Robertson et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond.

Sparse Retrievers: Computing Weighted Term Scores

d$_1$ = what is life ?
candy is life !

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

t$_1$ = what

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

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

Robertson et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond.

Sparse Retrievers: Computing Weighted Term Scores

d_1 = what is life ?
candy is life !

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

t_1 = what

\# of documents

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

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

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

Robertson et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond.

Sparse Retrievers: Computing Weighted Term Scores

d_1 = what is life ?
candy is life !

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

t_1 = what

\# of documents

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

\# of documents **where term t appears**

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

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

Robertson et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond.

Sparse Retrievers: Weighted-term Vectors

	q=what is nlp	d_1 = what is life ? candy is life !	d_2 = nlp is an acronym for natural language processing	d_3 = I like to do good research on nlp
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

Compute TF-IDF weights to build weighted vectors

Sparse Retrievers: Weighted-term Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

	q	d_1	d_2	d_3
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

Compute TF-IDF weights to build weighted vectors

Sparse Retrievers: Weighted-term Vectors

	q=what is nlp	d_1 = what is life ? candy is life !	d_2 = nlp is an acronym for natural language processing	d_3 = I like to do good research on nlp
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

Compute TF-IDF weights to build weighted vectors

Sparse Retrievers: Weighted-term Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

	q	d_1	d_2	d_3
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

Compute TF-IDF weights to build weighted vectors

Sparse Retrievers: Weighted-term Vectors

	q=what is nlp	d_1 = what is life ? candy is life !	d_2 = nlp is an acronym for natural language processing	d_3 = I like to do good research on nlp
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

Sparse Retrievers: Weighted-term Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

	q	d_1	d_2	d_3
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

Compute cosine similarity

Sparse Retrievers: Weighted-term Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

$$
\begin{array}{l}
\text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \text{life} \\ \ldots
\end{array}
\quad
\begin{pmatrix} 0.36 \\ 0 \\ 0.13 \\ 0.13 \\ 0 \\ 0 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0.18 \\ 0.18 \\ 0 \\ 0.13 \\ 0 \\ 0.36 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.05 \\ 0.05 \\ 0.13 \\ 0 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.05 \\ 0 \\ 0 \\ 0 \\ \ldots \end{pmatrix}
$$

Compute cosine similarity

$q*d_1 = 0.44$

Sparse Retrievers: Weighted-term Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

$$
\begin{array}{c}
\text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \text{life} \\ \dots
\end{array}
\quad
\begin{pmatrix} 0.36 \\ 0 \\ 0.13 \\ 0.13 \\ 0 \\ 0 \\ \dots \end{pmatrix}
\quad
\begin{pmatrix} 0.18 \\ 0.18 \\ 0 \\ 0.13 \\ 0 \\ 0.36 \\ \dots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.05 \\ 0.05 \\ 0.13 \\ 0 \\ \dots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.05 \\ 0 \\ 0 \\ 0 \\ \dots \end{pmatrix}
$$

Compute cosine similarity

$q*d_1$ = 0.44

$q*d_2$ = 0.21

Sparse Retrievers: Weighted-term Vectors

	q=what is nlp	d_1 = what is life ? candy is life !	d_2 = nlp is an acronym for natural language processing	d_3 = I like to do good research on nlp
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

Compute cosine similarity

$q*d_1 = 0.44$

$q*d_2 = 0.21$

$q*d_3 = 0.32$

Computing TF-IDF Matrices: Weighted-term Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

	q	d_1	d_2	d_3
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

$q*d_1 = 0.44$

$\boxed{q*d_2 = 0.21}$

$q*d_3 = 0.32$

Computing TF-IDF Matrices: Weighted-term Vectors

q=what is nlp

d_1 = what is life ? candy is life !

d_2 = nlp is an acronym for natural language processing

d_3 = I like to do good research on nlp

$$
\begin{array}{c}
\text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \text{life} \\ \dots
\end{array}
\begin{pmatrix} 0.36 \\ 0 \\ 0.13 \\ 0.13 \\ 0 \\ 0 \\ \dots \end{pmatrix}
\quad
\begin{pmatrix} 0.18 \\ 0.18 \\ 0 \\ 0.13 \\ 0 \\ 0.36 \\ \dots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.05 \\ 0.05 \\ 0.13 \\ 0 \\ \dots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.05 \\ 0 \\ 0 \\ 0 \\ \dots \end{pmatrix}
$$

"Bag-of-words"

$q*d_1 = 0.44$

$\boxed{q*d_2 = 0.21}$

$q*d_3 = 0.32$

Computing TF-IDF Matrices: Weighted-term Vectors

	q=what is nlp	d_1 = what is life ? candy is life !	d_2 = nlp is an acronym for natural language processing	d_3 = I like to do good research on nlp
what	0.36	0.18	0	0
candy	0	0.18	0	0
nlp	0.13	0	0.05	0.05
is	0.13	0.13	0.05	0
language	0	0	0.13	0
life	0	0.36	0	0
…	…	…	…	…

"Bag-of-words"

$q*d_1$ = 0.44

$\boxed{q*d_2 = 0.21}$

$q*d_3$ = 0.32

Can't fully capture semantic similarities

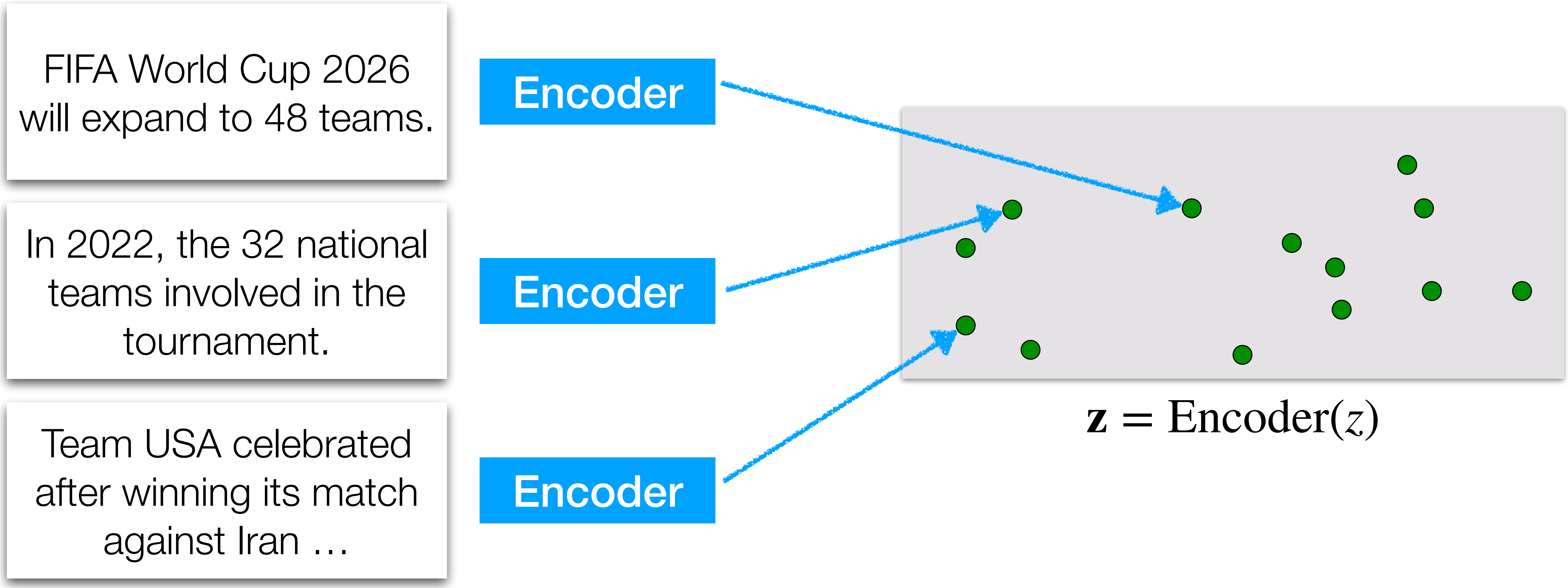
Dense Retrievers: Overview

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 …

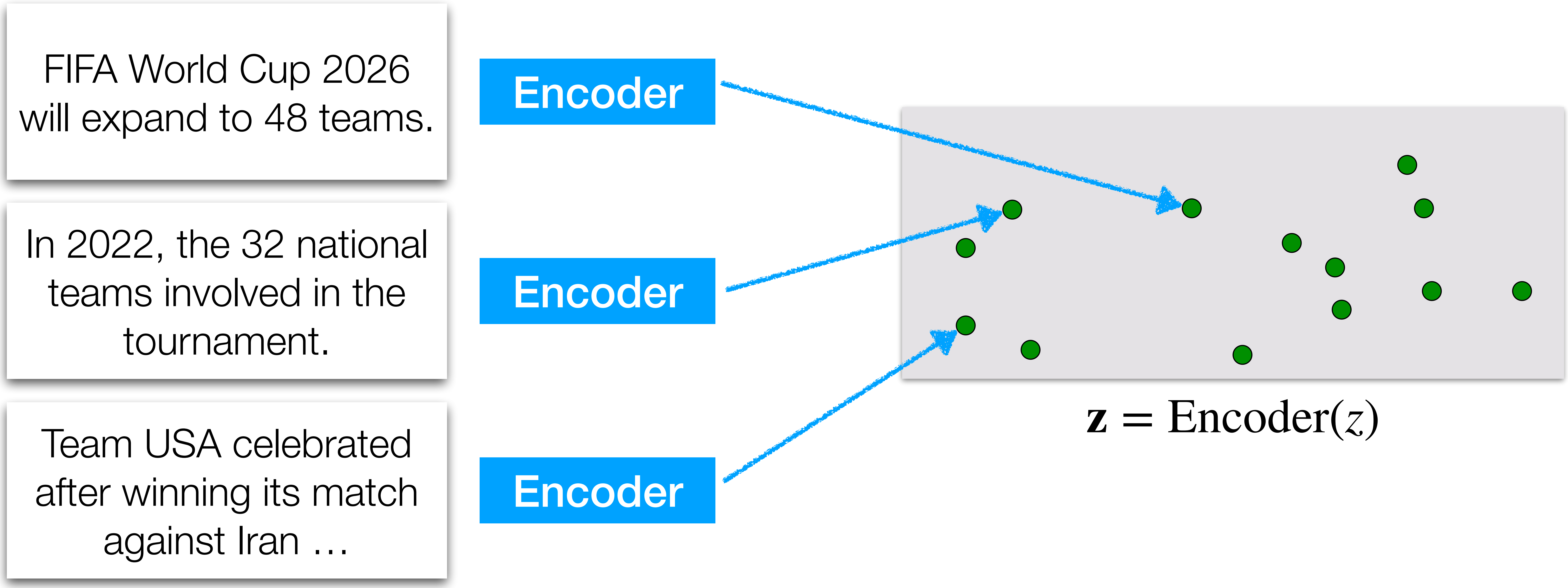
Dense Retrievers: Overview



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

Dense Retrievers: Overview

$\boldsymbol{x}$ = How many teams will participate in FIFA World Cup 2026?



Dense Retrievers: Overview

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

Encoder

FIFA World Cup 2026 will expand to 48 teams.

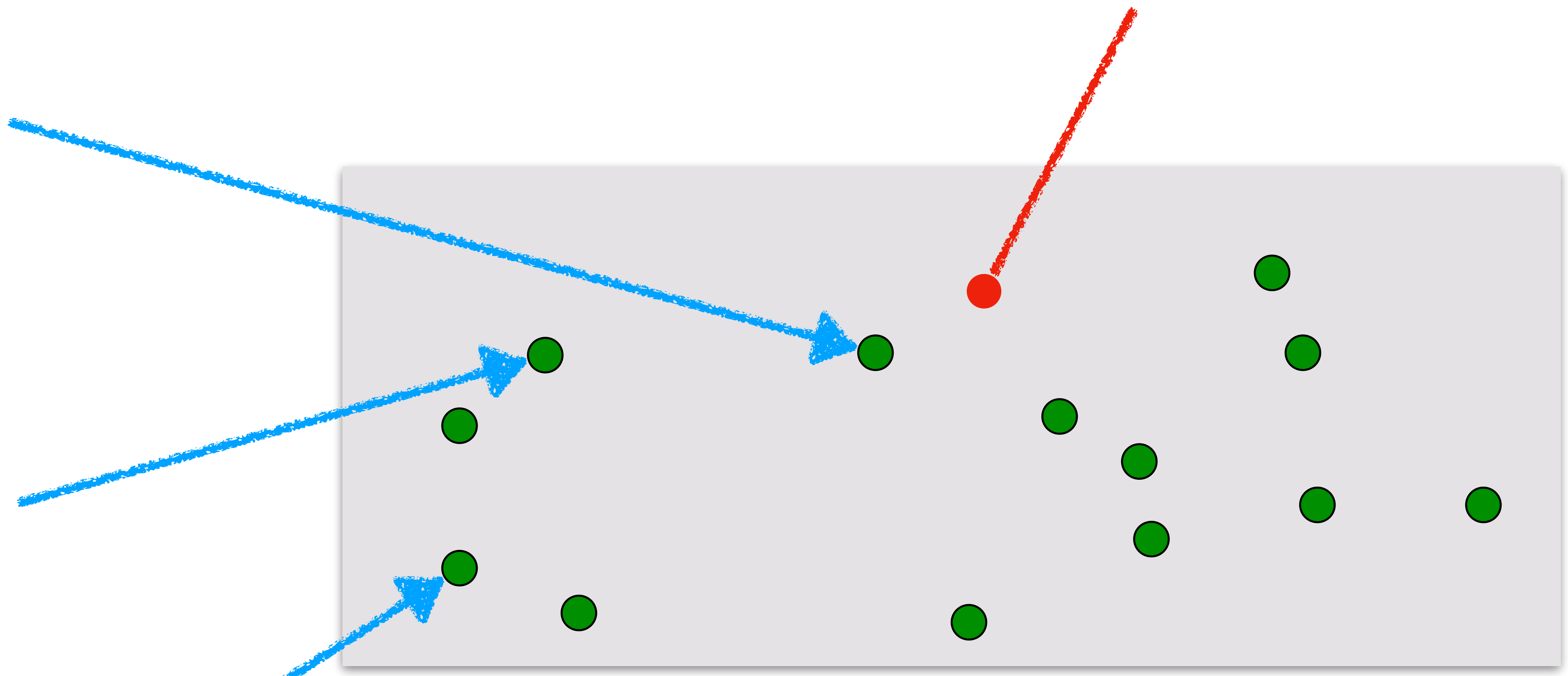
In 2022, the 32 national teams involved in the tournament.

Team USA celebrated after winning its match against Iran …

Encoder

Encoder

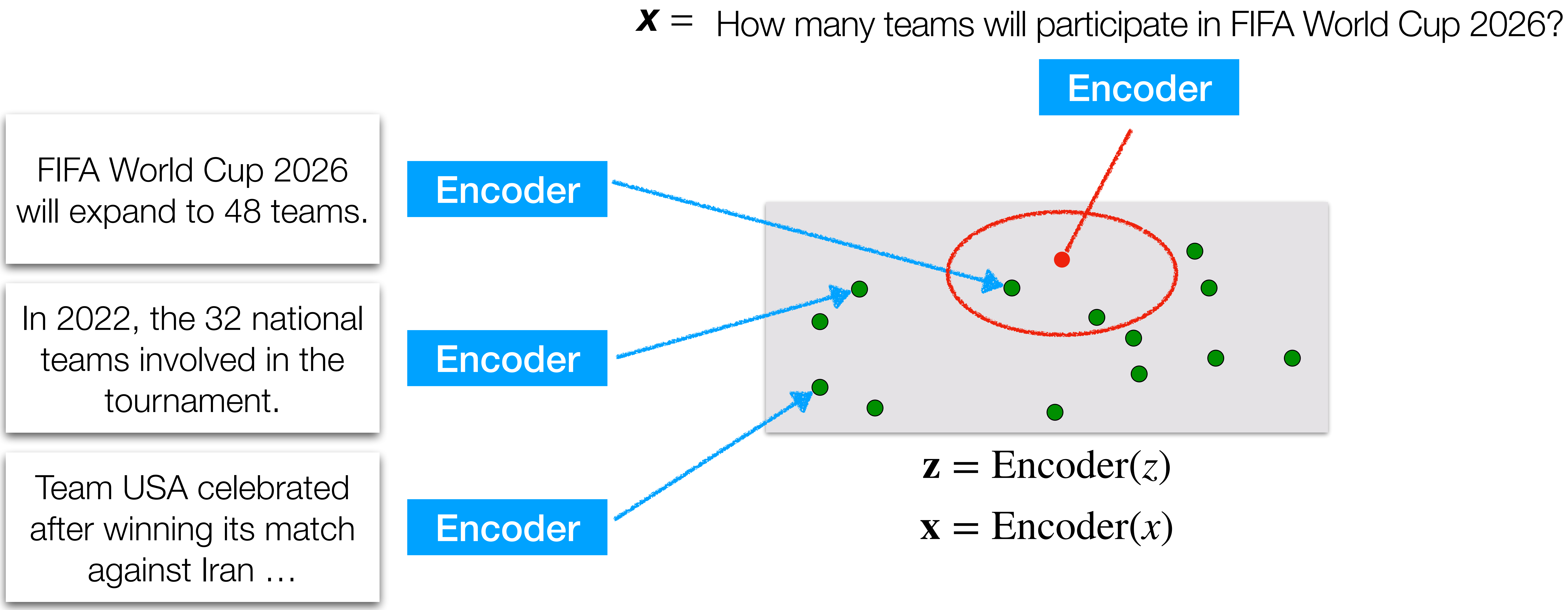
Encoder



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

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

Dense Retrievers: Overview

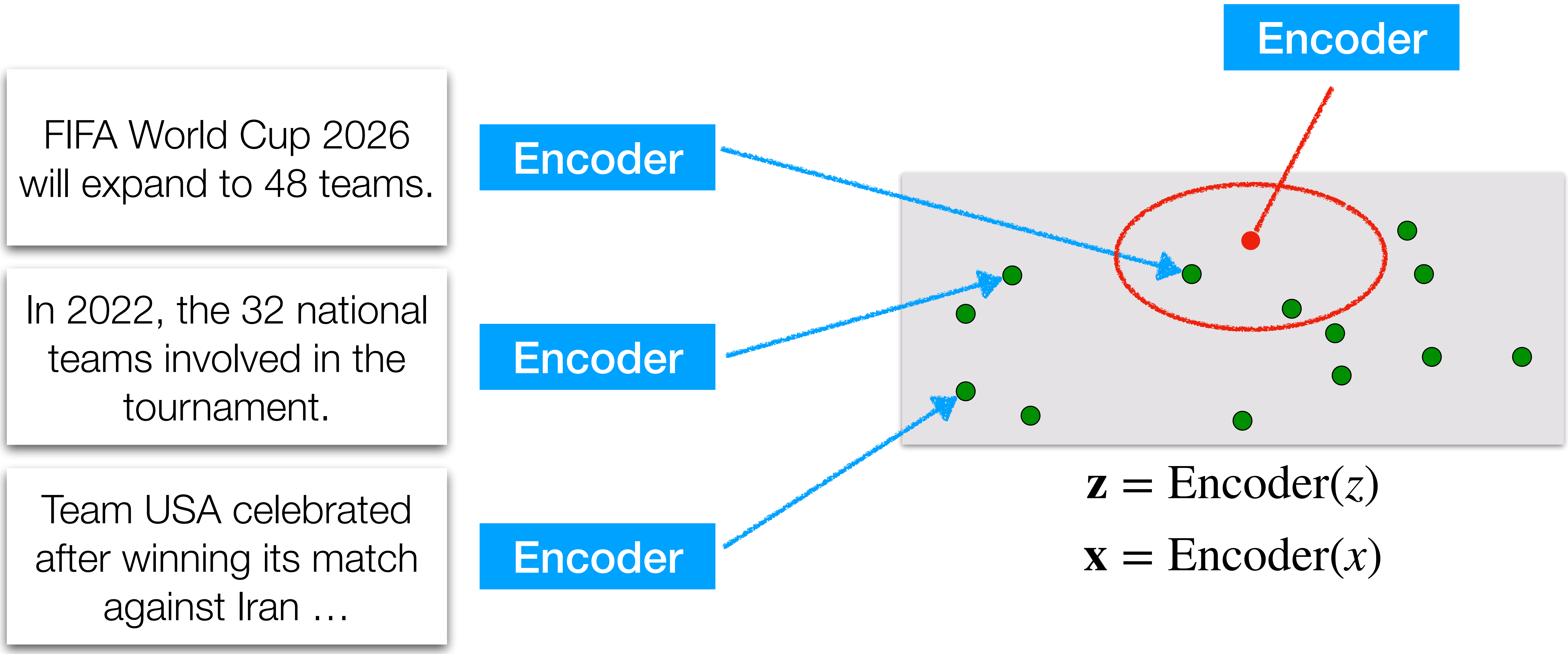


$$\mathbf{z} = \text{Encoder}(z)$$

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

Dense Retrievers: Overview

$\boldsymbol{x}$ = How many teams will participate in FIFA World Cup 2026?



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

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

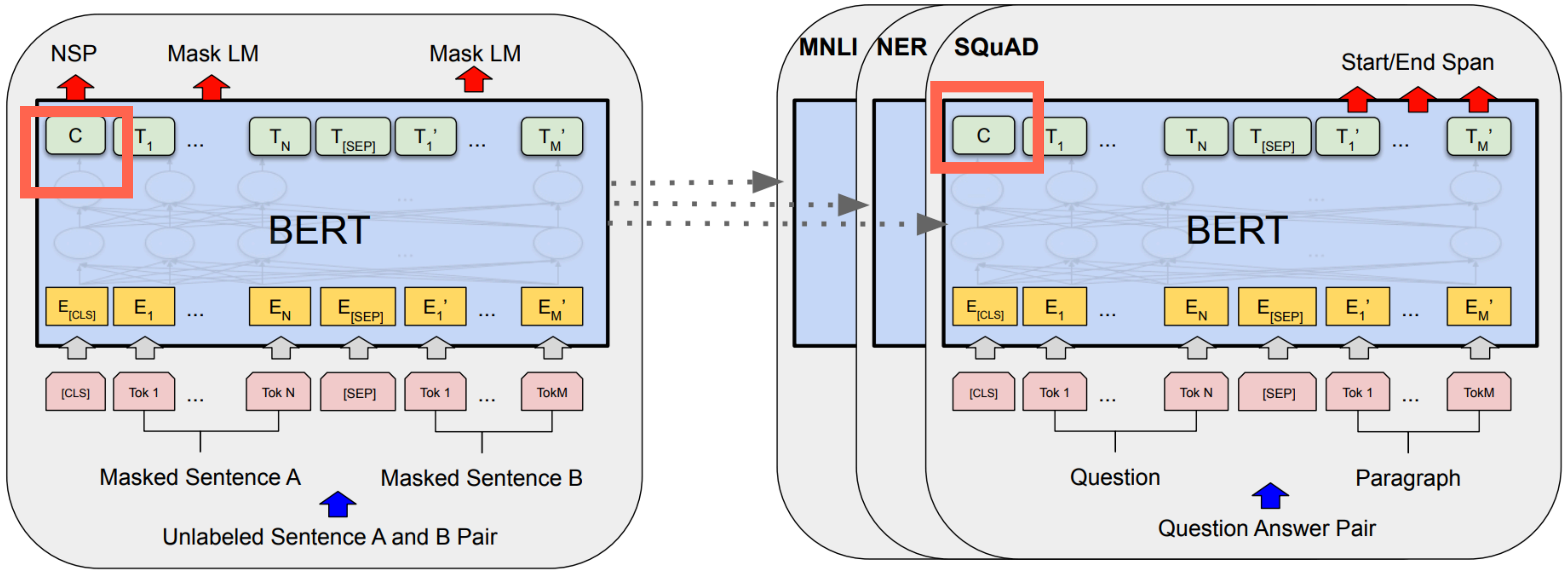
$$z_1, \ldots, z_k = \text{argTop-}k\,(\mathbf{x} \cdot \mathbf{z})$$

k retrieved chunks

Dense Retrievers: Generating *Embeddings*

- **Use output of [CLS] token in masked LMs** $\mathbb{R}^d$

e.g., DPR



Pre-training

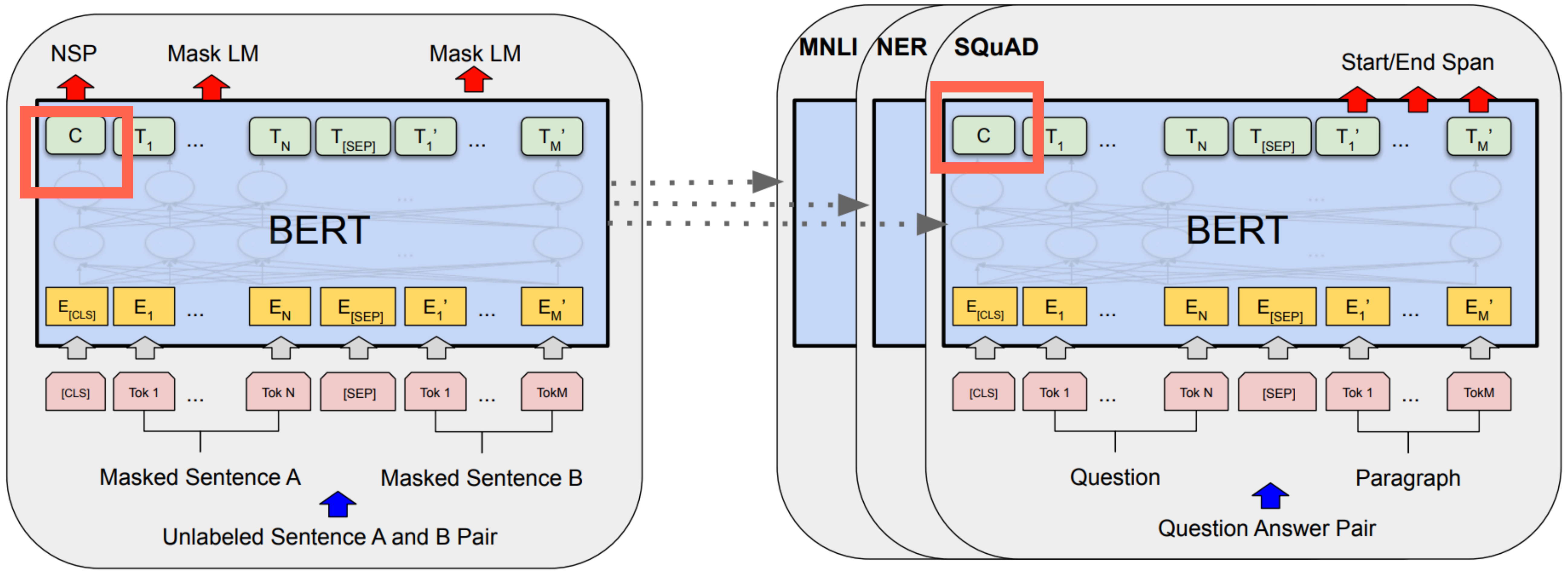
Fine-Tuning

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** $\mathbb{R}^d$

  e.g., DPR



Pre-training

Fine-Tuning

- **Mean / Max pooling of output vectors** $\mathbb{R}^{N \times d}$

  e.g., SBERT, SGPT

	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	Exhausti
Exact Search for L2	`IndexFlatL2`	`"Flat"`	`d`	`4*d`	yes
Exact Search for Inner Product	`IndexFlatIP`	`"Flat"`	`d`	`4*d`	yes
Hierarchical Navigable Small World graph exploration	`IndexHNSWFlat`	`"HNSW,Flat"`	`d`, `M`	`4*d + x * M * 2 * 4`	no
Inverted file with exact post-verification	`IndexIVFFlat`	`"IVFx,Flat"`	`quantizer`, `d`, `nlists`, `metric`	`4*d + 8`	no
Locality-Sensitive Hashing (binary flat index)	`IndexLSH`	-	`d`, `nbits`	`ceil(nbits/8)`	yes
Scalar quantizer (SQ) in flat mode	`IndexScalarQuantizer`	`"SQ8"`	`d`	`d`	yes
Product quantizer (PQ) in flat mode	`IndexPQ`	`"PQx"`, `"PQ"M"x"nbits`	`d`, `M`, `nbits`	`ceil(M * nbits / 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"y"x"nbits`	`quantizer`, `d`, `nlists`, `M`, `nbits`	`ceil(M * nbits/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	Exhausti
Exact Search for L2	`IndexFlatL2`	`"Flat"`	`d`	`4*d`	yes
Exact Search for Inner Product	`IndexFlatIP`	`"Flat"`	`d`	`4*d`	yes
Hierarchical Navigable Small World graph exploration	`IndexHNSWFlat`	`"HNSW,Flat"`	`d`, `M`	`4*d + x * M * 2 * 4`	no
Inverted file with exact post-verification	`IndexIVFFlat`	`"IVFx,Flat"`	`quantizer`, `d`, `nlists`, `metric`	`4*d + 8`	no
Locality-Sensitive Hashing (binary flat index)	`IndexLSH`	-	`d`, `nbits`	`ceil(nbits/8)`	yes
Scalar quantizer (SQ) in flat mode	`IndexScalarQuantizer`	`"SQ8"`	`d`	`d`	yes
Product quantizer (PQ) in flat mode	`IndexPQ`	`"PQx"`, `"PQ"M"x"nbits`	`d`, `M`, `nbits`	`ceil(M * nbits / 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"y"x"nbits`	`quantizer`, `d`, `nlists`, `M`, `nbits`	`ceil(M * nbits/8)+8`	no

Exact search (still fast for 10^6~10^7 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	Exhausti
Exact Search for L2	`IndexFlatL2`	`"Flat"`	`d`	`4*d`	yes
Exact Search for Inner Product	`IndexFlatIP`	`"Flat"`	`d`	`4*d`	yes
Hierarchical Navigable Small World graph exploration	`IndexHNSWFlat`	`"HNSW,Flat"`	`d`, `M`	`4*d + x * M * 2 * 4`	no
Inverted file with exact post-verification	`IndexIVFFlat`	`"IVFx,Flat"`	`quantizer`, `d`, `nlists`, `metric`	`4*d + 8`	no
Locality-Sensitive Hashing (binary flat index)	`IndexLSH`	-	`d`, `nbits`	`ceil(nbits/8)`	yes
Scalar quantizer (SQ) in flat mode	`IndexScalarQuantizer`	`"SQ8"`	`d`	`d`	yes
Product quantizer (PQ) in flat mode	`IndexPQ`	`"PQx"`, `"PQ"M"x"nbits`	`d`, `M`, `nbits`	`ceil(M * nbits / 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"y"x"nbits`	`quantizer`, `d`, `nlists`, `M`, `nbits`	`ceil(M * nbits/8)+8`	no

Exact search (still fast for 10^6~10^7 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	Exhausti
Exact Search for L2	IndexFlatL2	"Flat"	d	4*d	yes
Exact Search for Inner Product	IndexFlatIP	"Flat"	d	4*d	yes
Hierarchical Navigable Small World graph exploration	IndexHNSWFlat	"HNSW,Flat"	d, M	4*d + x * M * 2 * 4	no
Inverted file with exact post-verification	IndexIVFFlat	"IVFx,Flat"	quantizer, d, nlists, metric	4*d + 8	no
Locality-Sensitive Hashing (binary flat index)	IndexLSH	-	d, nbits	ceil(nbits/8)	yes
Scalar quantizer (SQ) in flat mode	IndexScalarQuantizer	"SQ8"	d	d	yes
Product quantizer (PQ) in flat mode	IndexPQ	"PQx", "PQ"M"x"nbits	d, M, nbits	ceil(M * nbits / 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"y"x"nbits	quantizer, d, nlists, M, nbits	ceil(M * nbits/8)+8	no

Exact search (still fast for 10^6~10^7 scale)

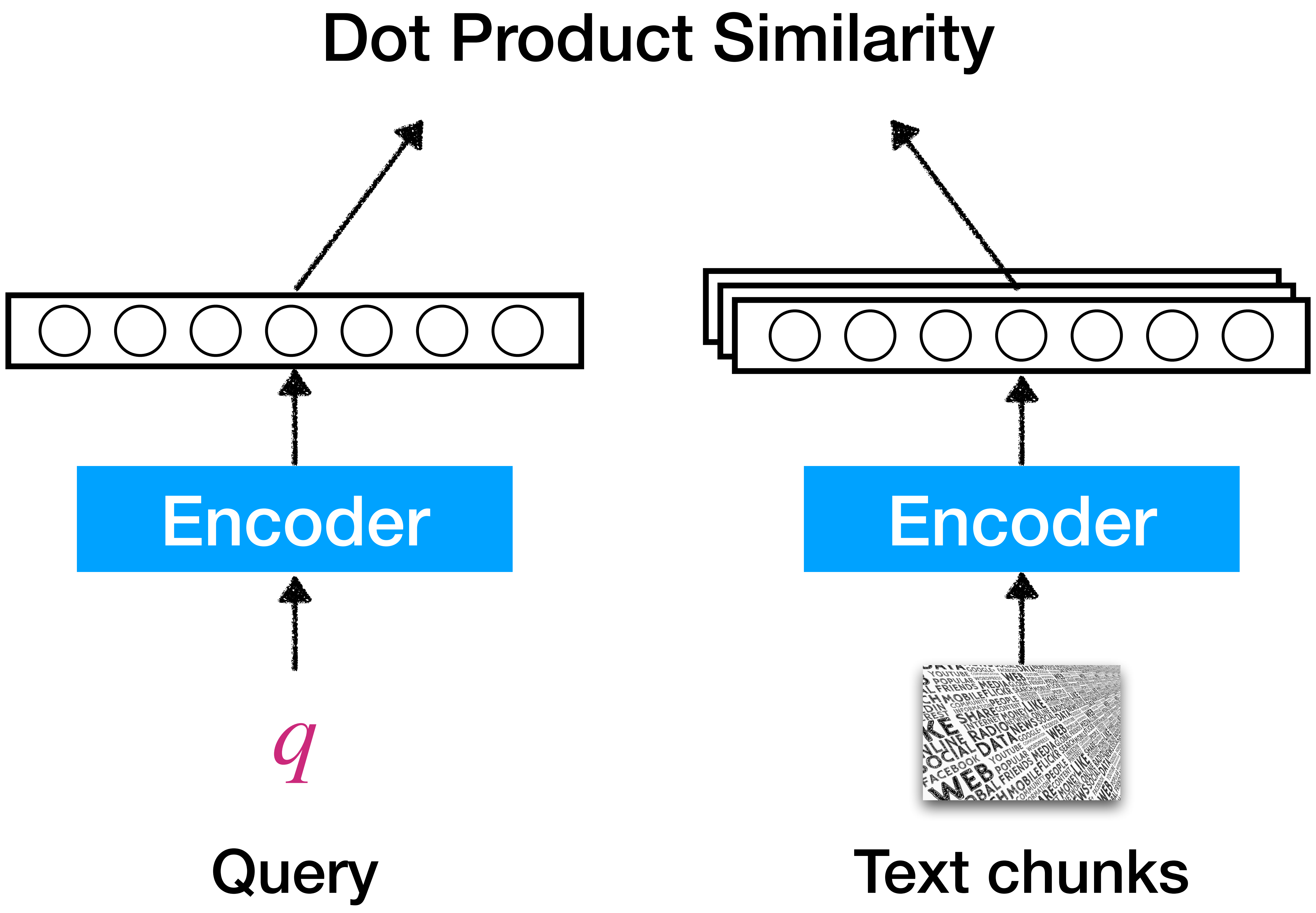
Approximate search (faster but more memory)

Reduce index size with quantization

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

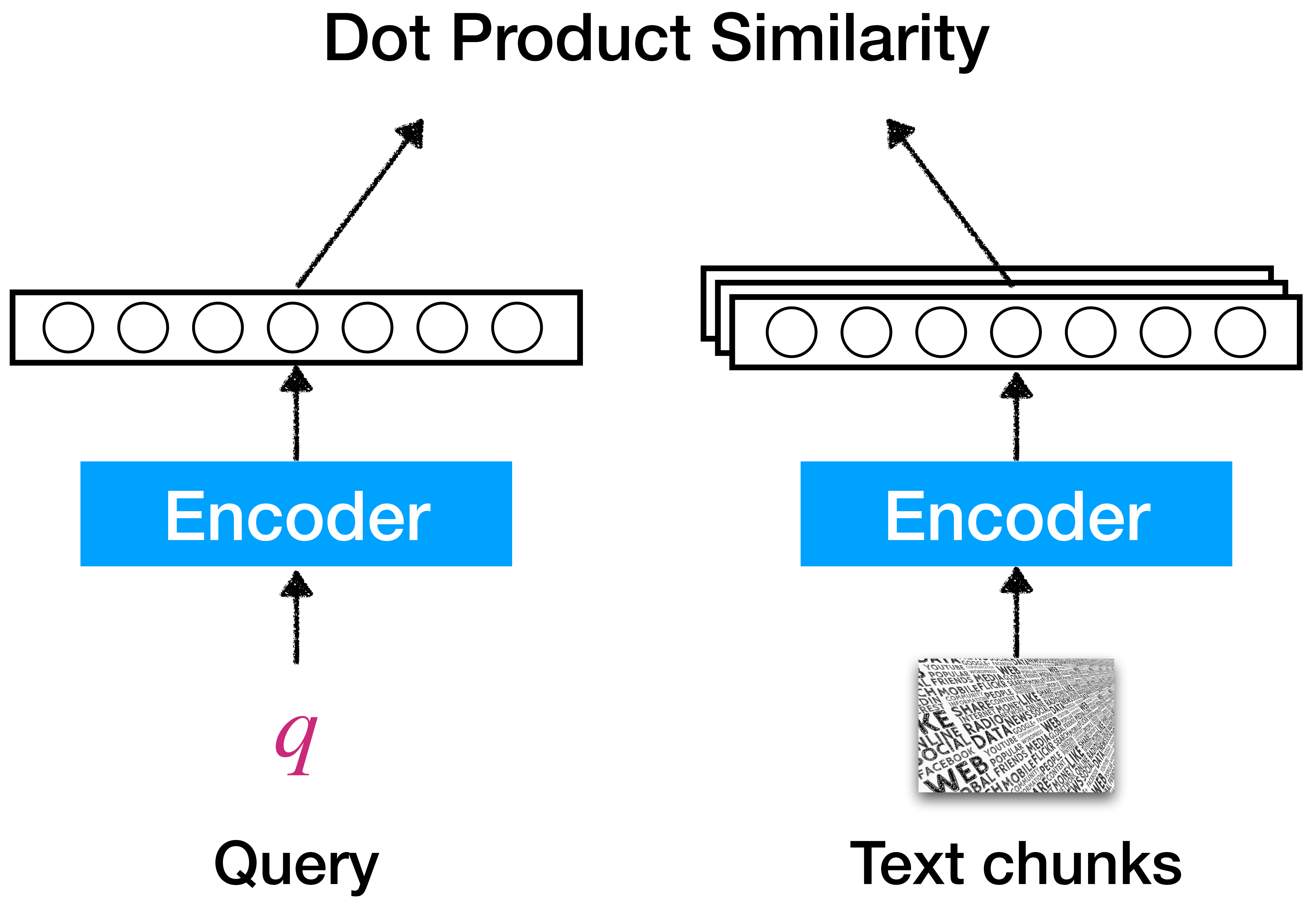
https://speakerdeck.com/matsui_528/cvpr20-tutorial-billion-scale-approximate-nearest-neighbor-search (CVPR 2020 Tutorial)

Training Dense Retriever

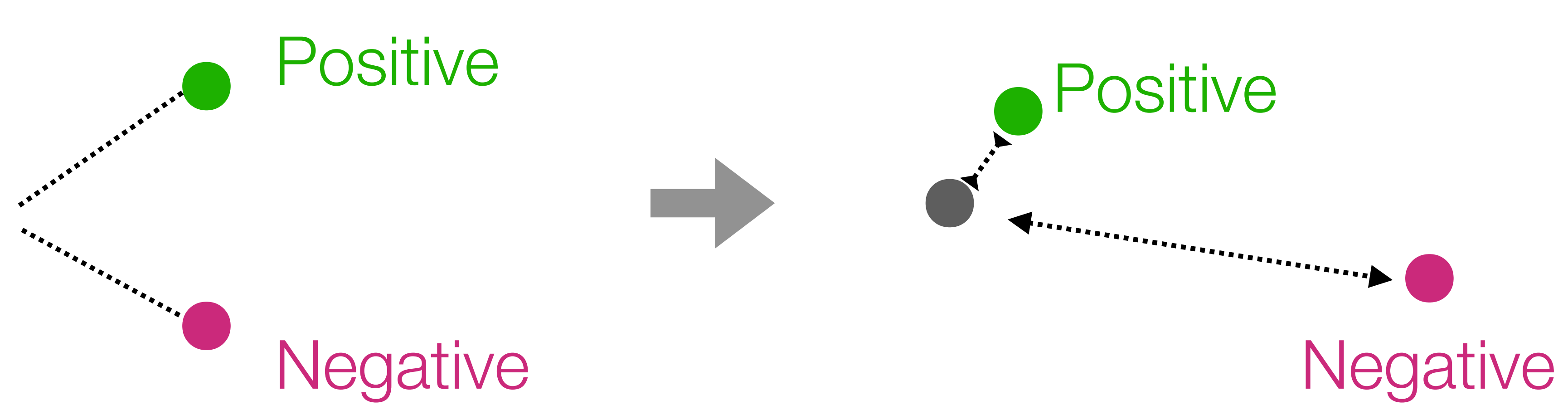


Contrastive learning

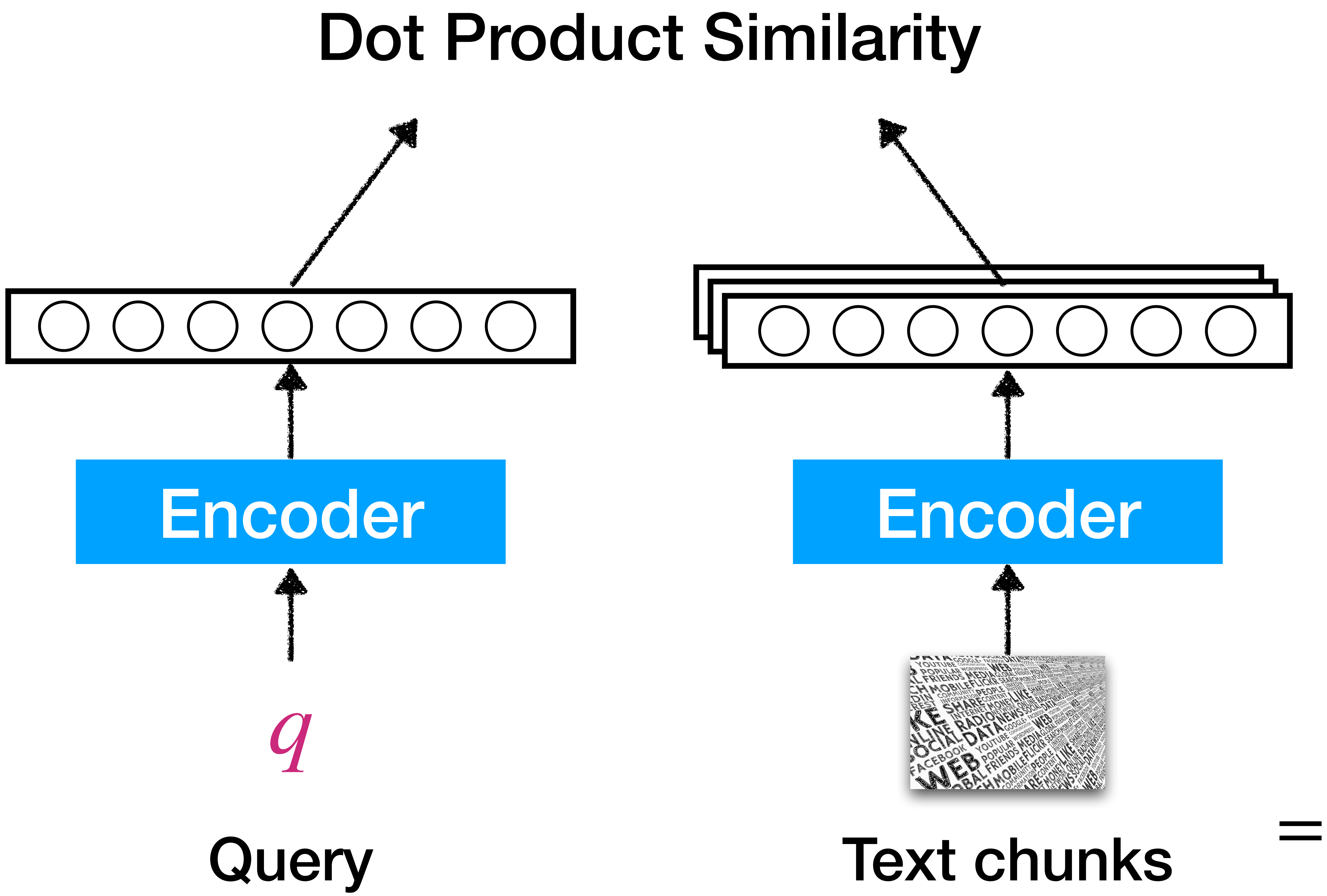
Training Dense Retriever



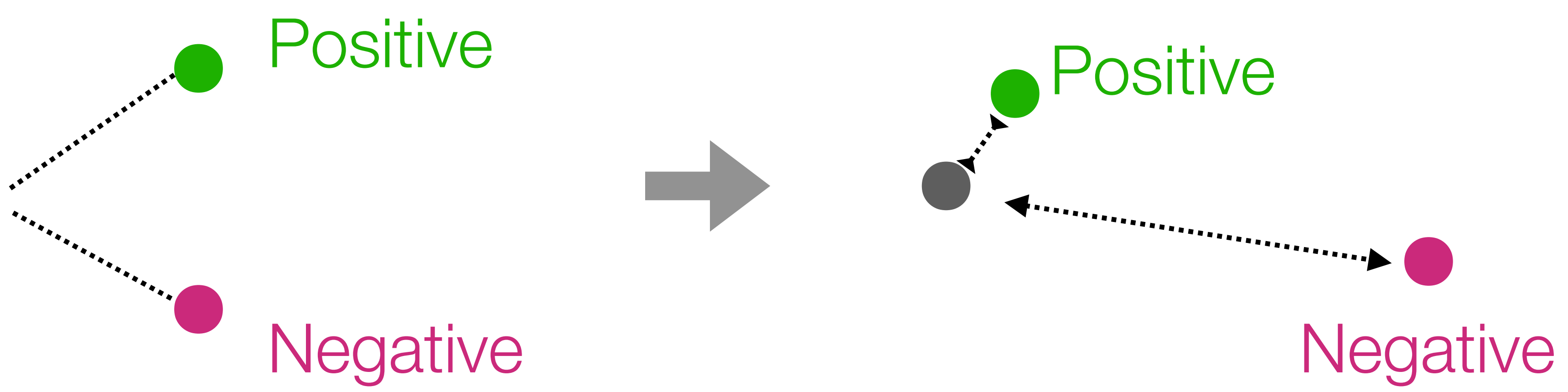
Contrastive learning



Training Dense Retriever



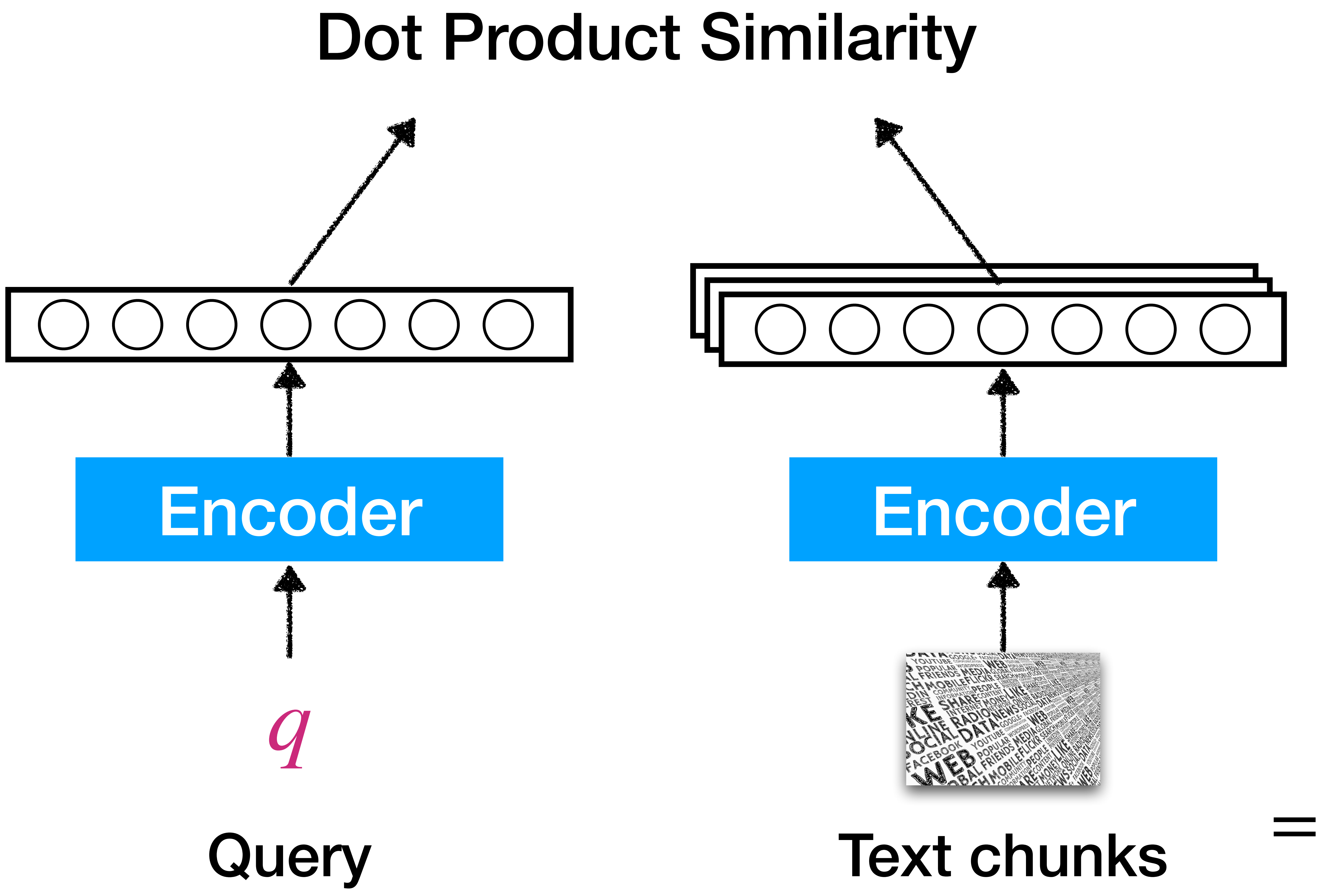
Contrastive learning



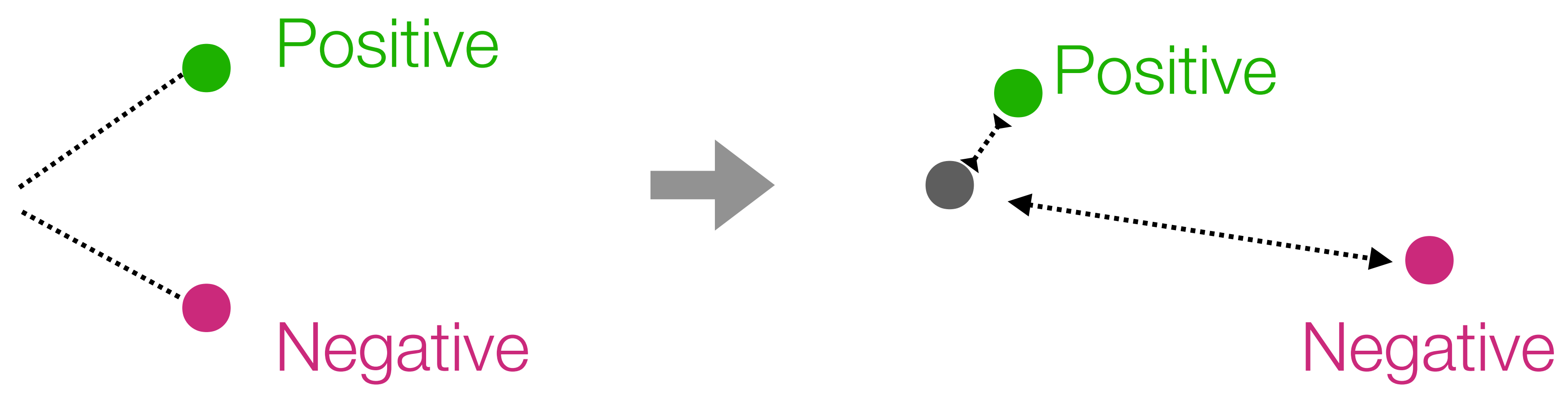
$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-)$$

$$= -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Training Dense Retriever



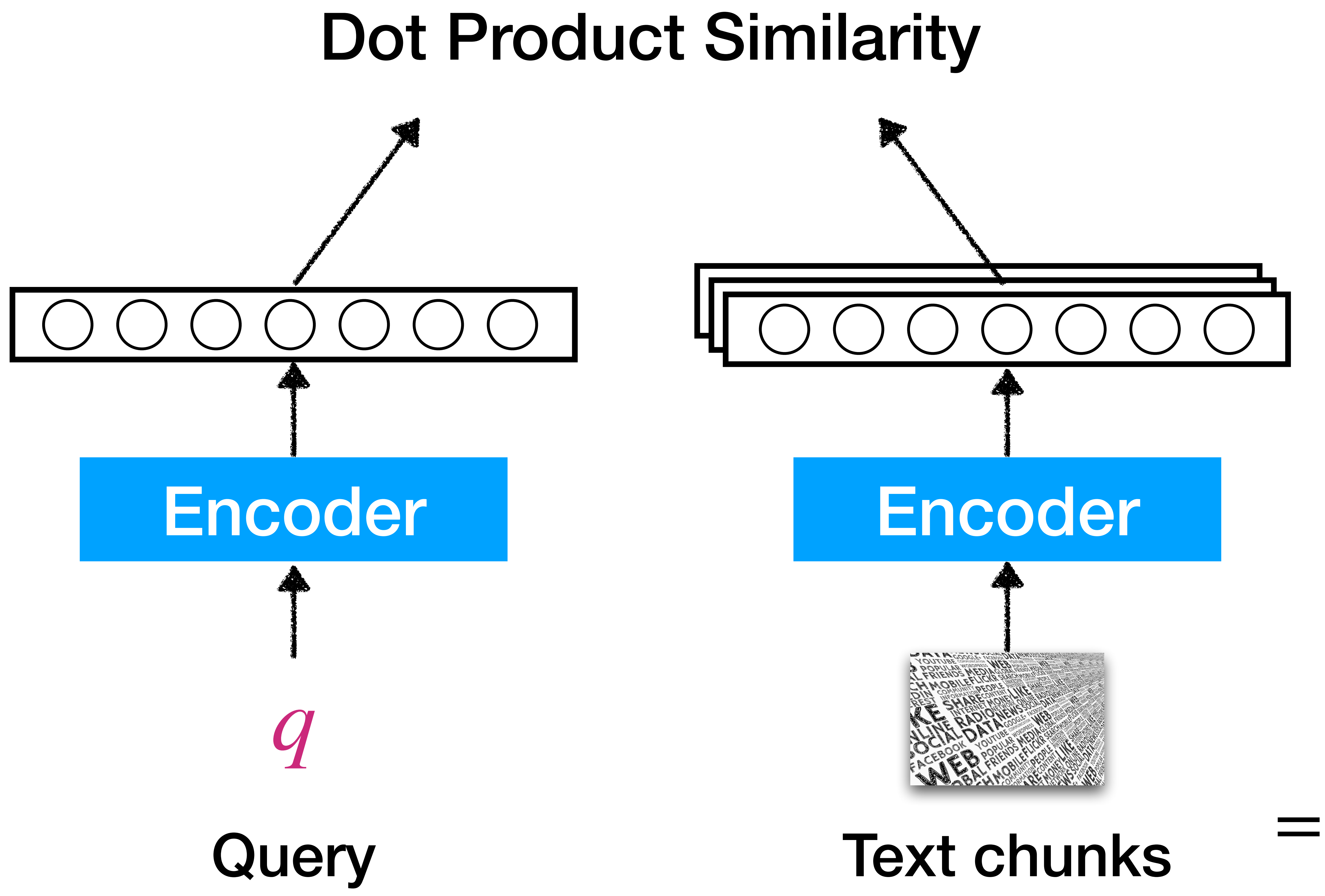
Contrastive learning



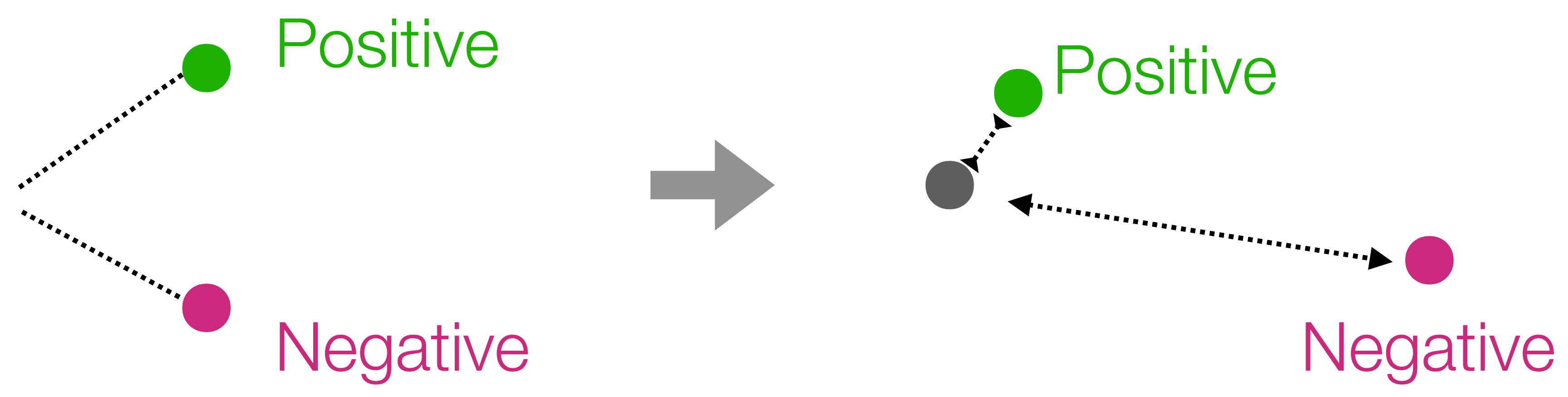
$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-)$$

$$= -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Training Dense Retriever



Contrastive learning



$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-)$$

$$= -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

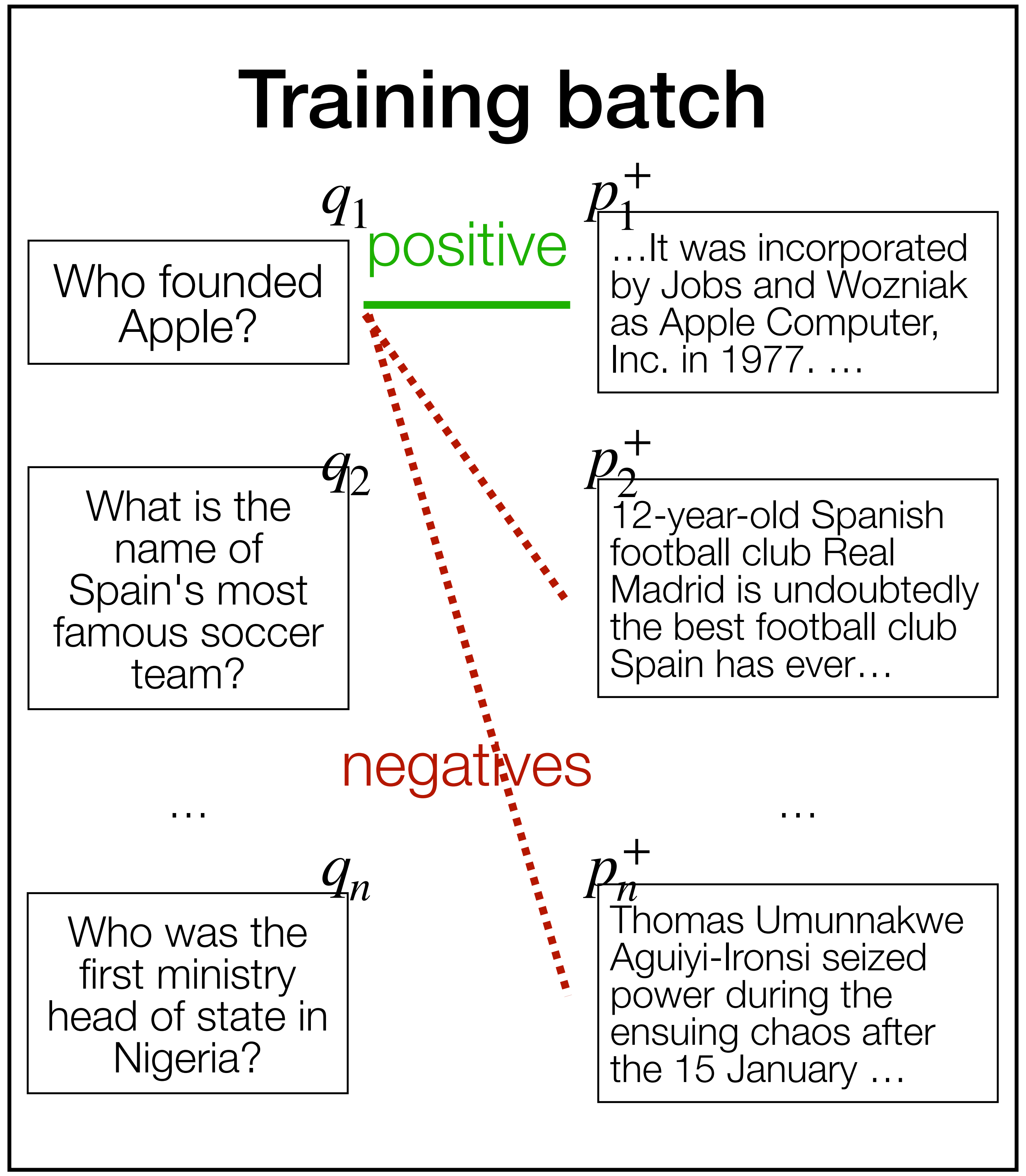
Training Dense Retriever

$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-)$$

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

In-batch negatives

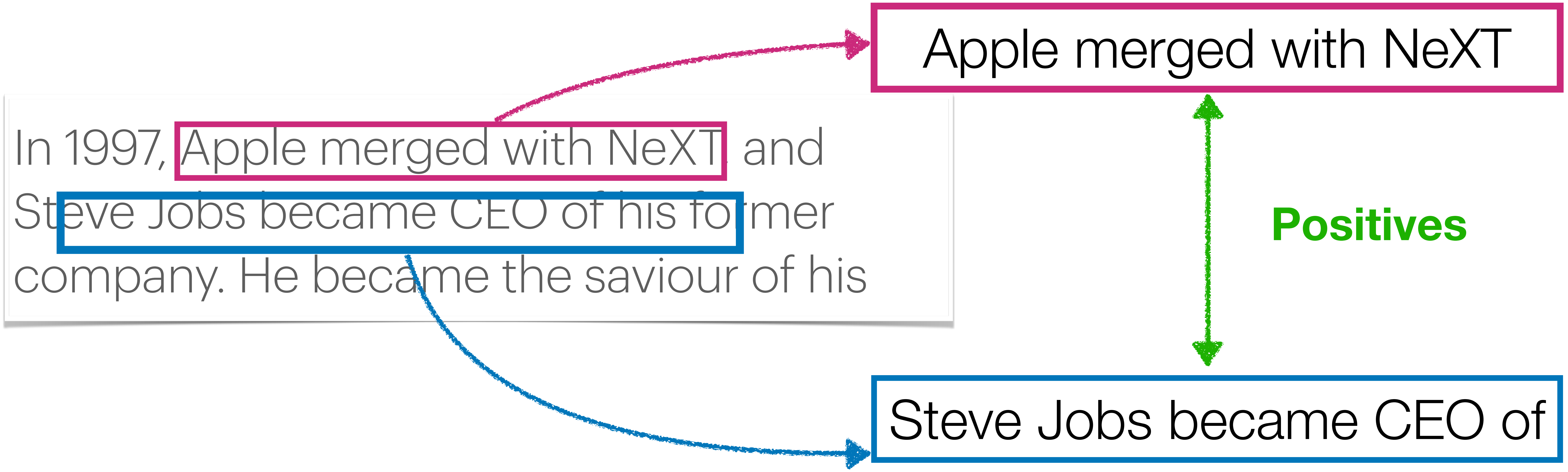
Hard negative retrieved by the same / another model



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

Unsupervised Training

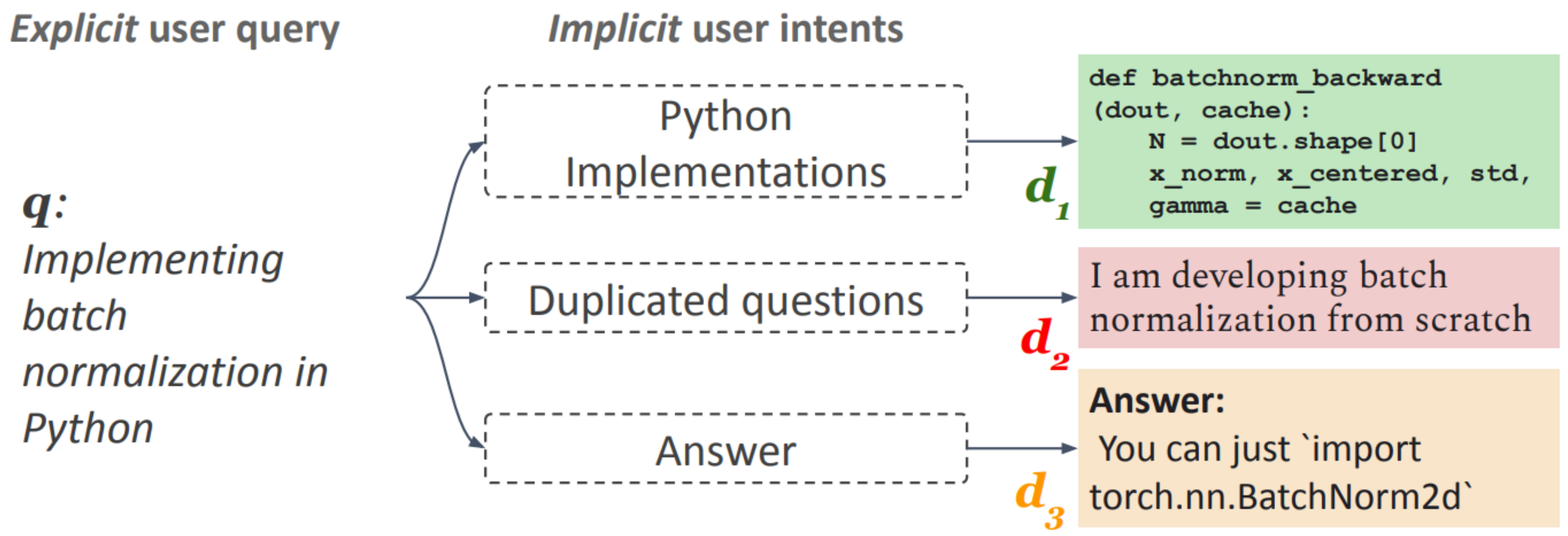
Independent Cropping

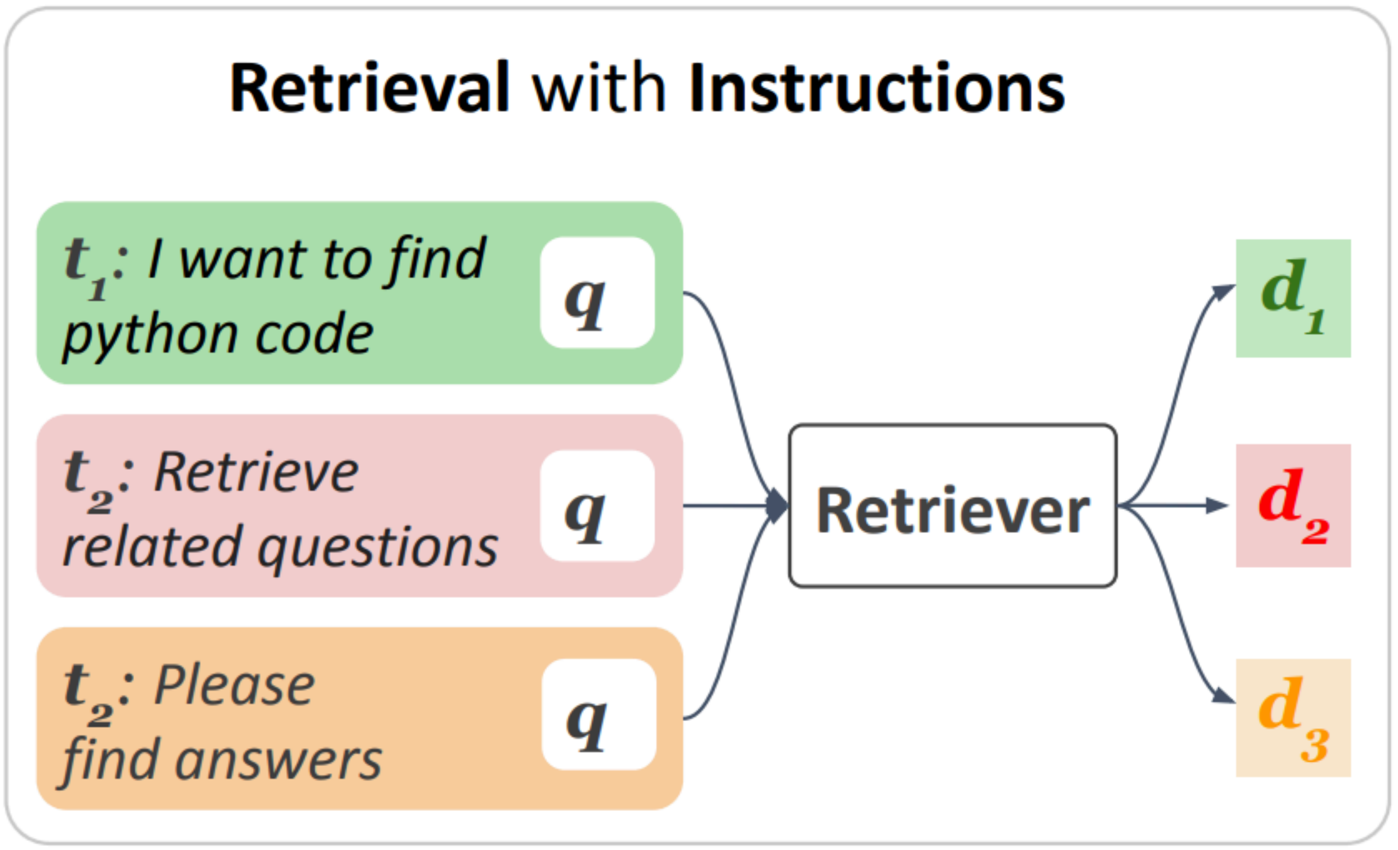


Unsupervised dense retrieval model!

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

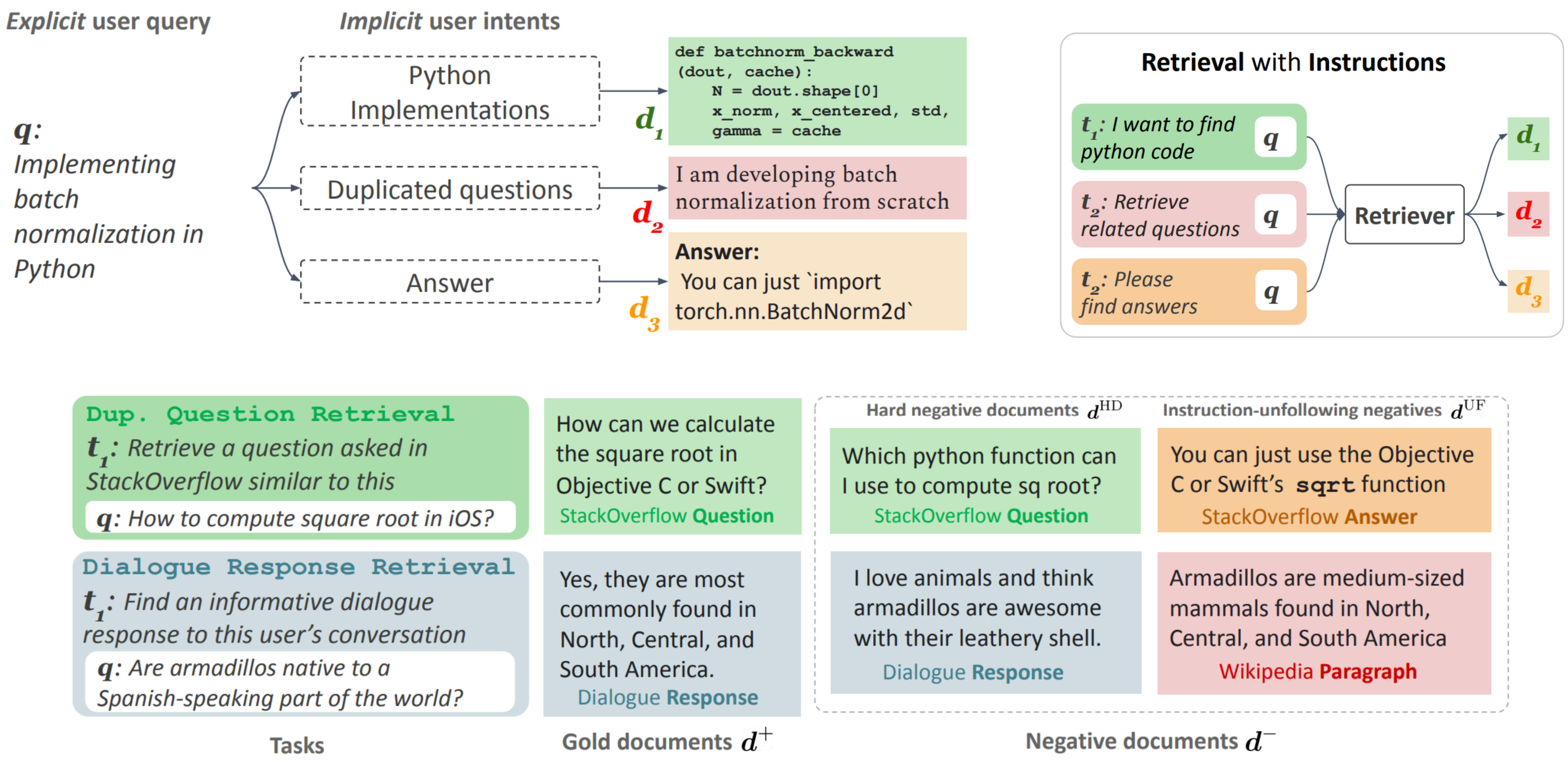
Instruction Tuning for Retriever





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

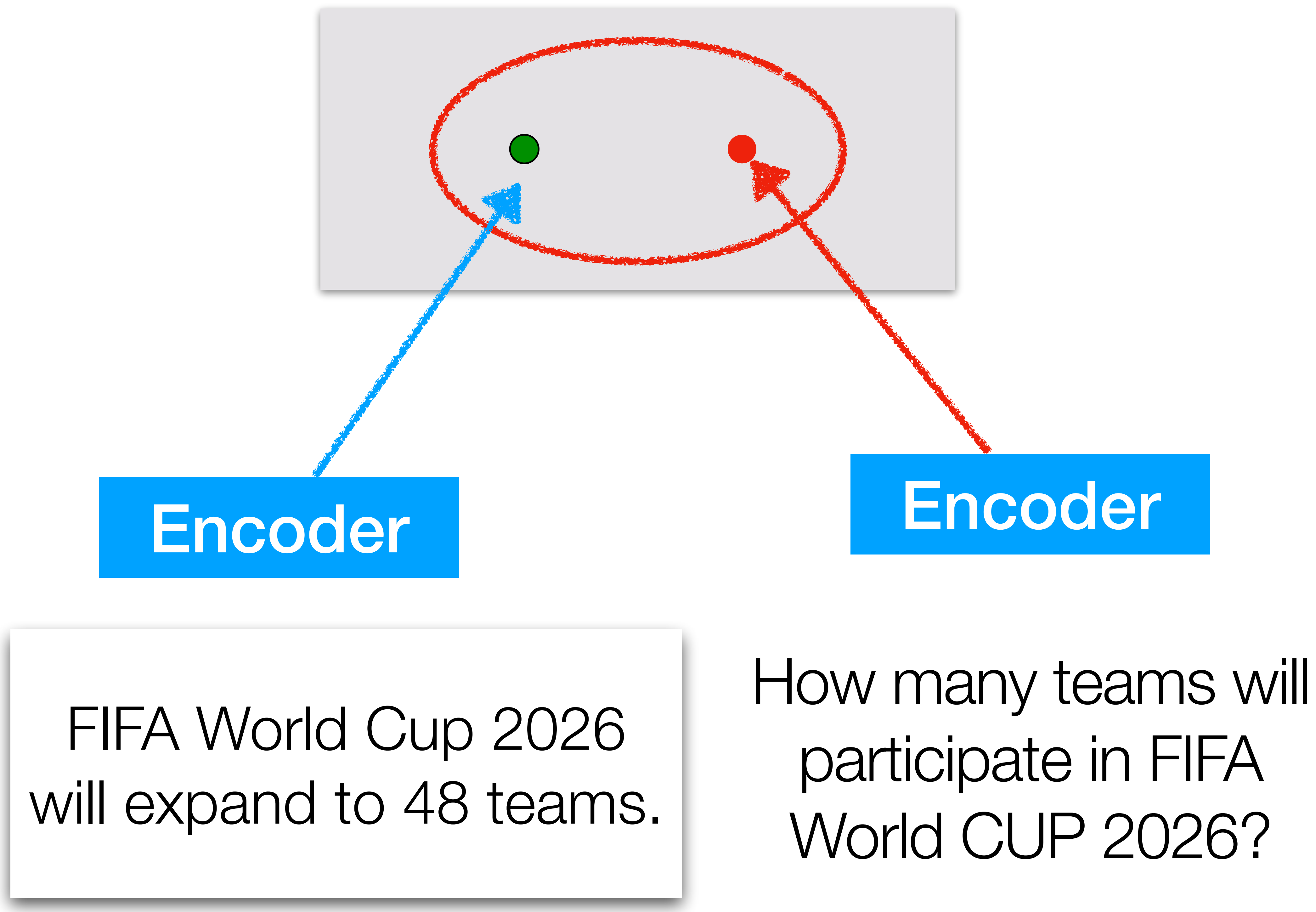
Instruction Tuning for Retriever



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

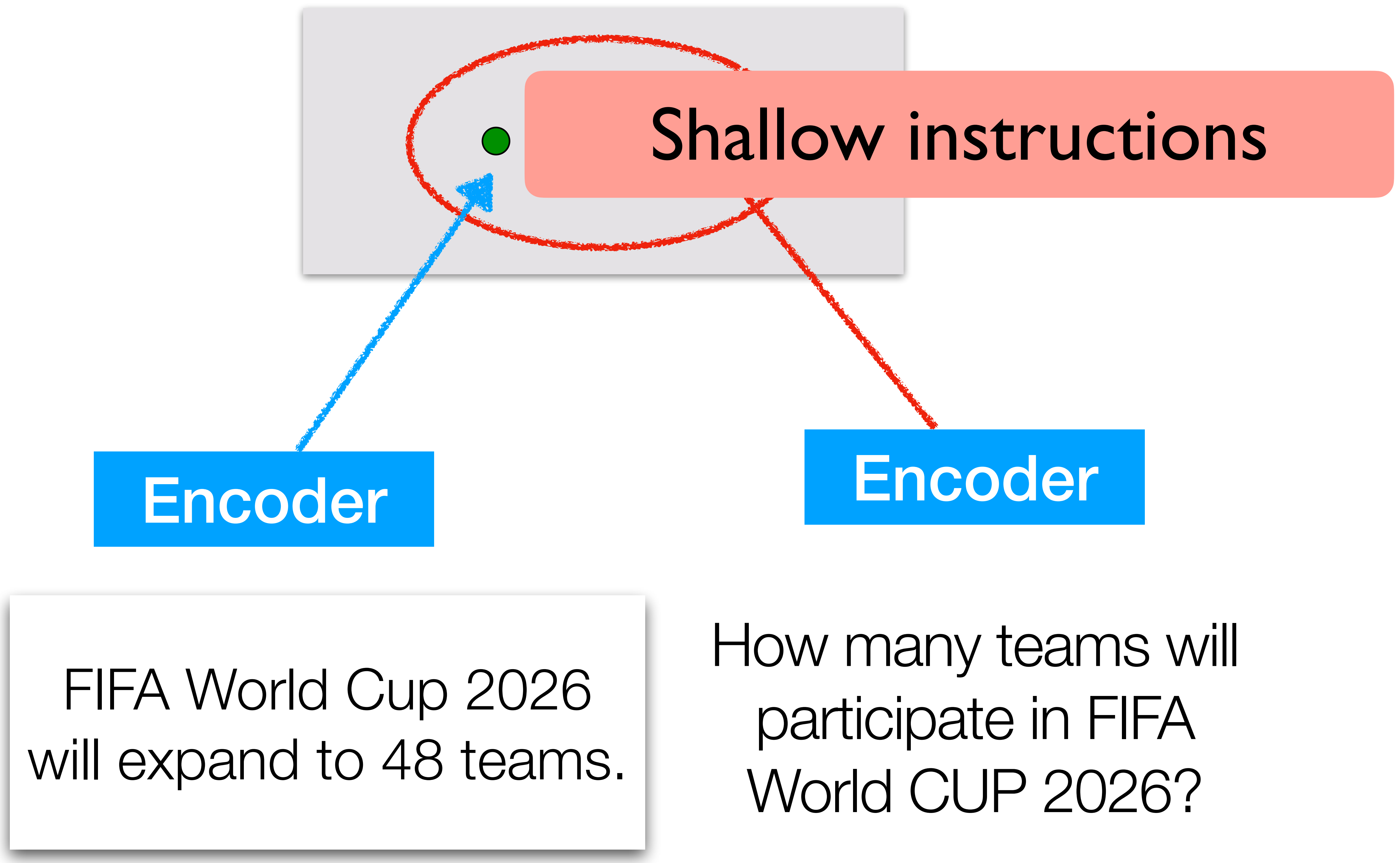
Reranking with Cross-encoder

Bi-Encoder



Reranking with Cross-encoder

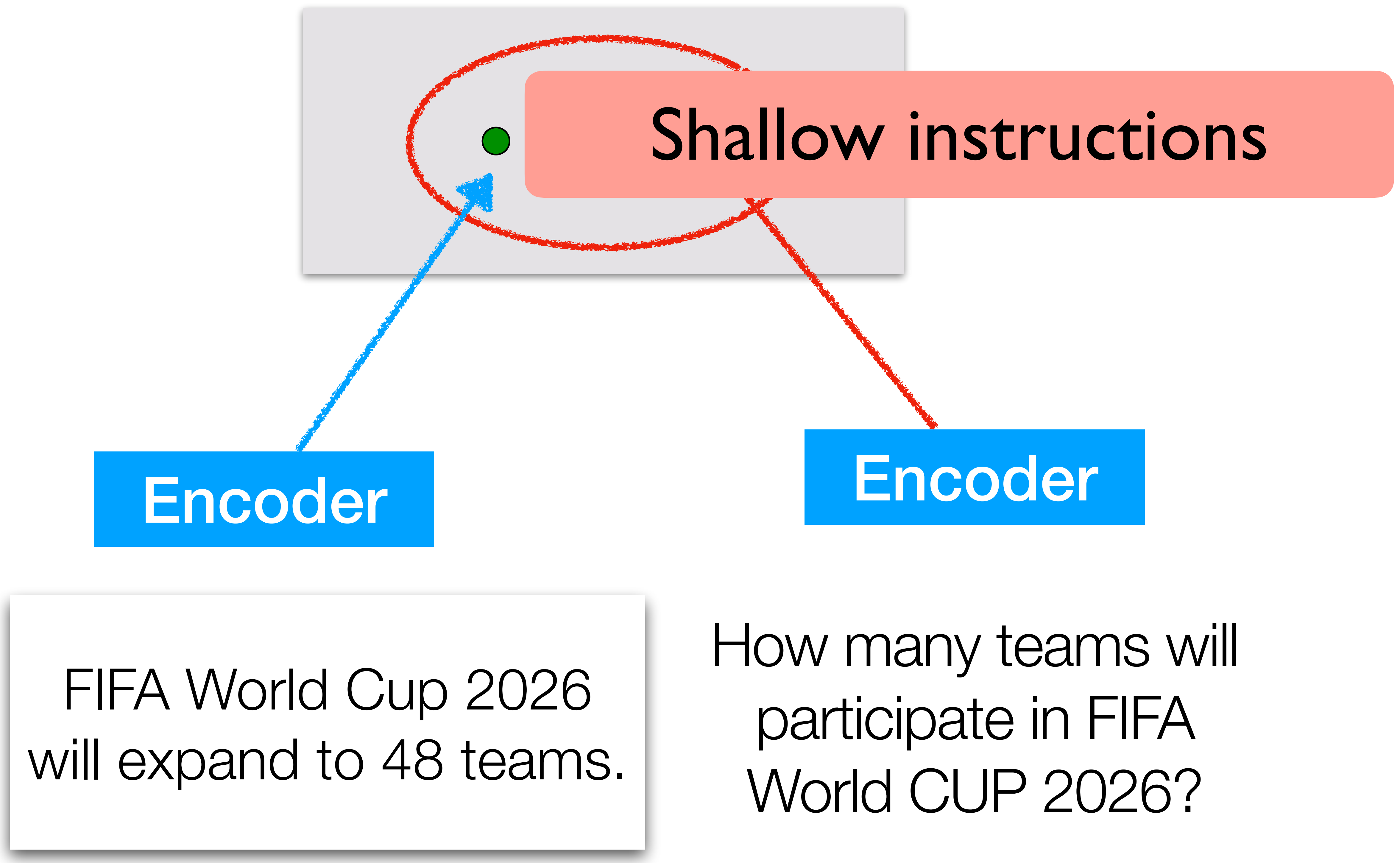
Bi-Encoder



Reranking with Cross-encoder

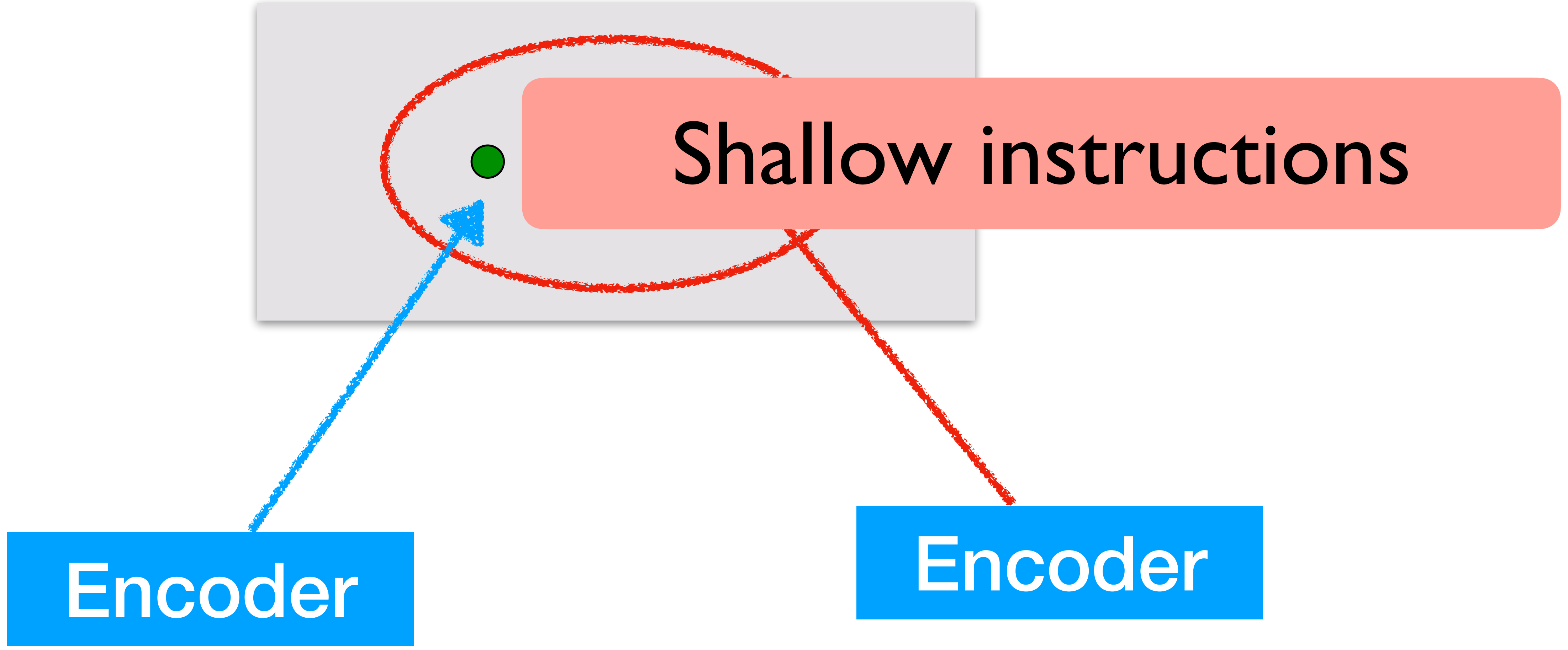
Bi-Encoder

Cross-Encoder



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?

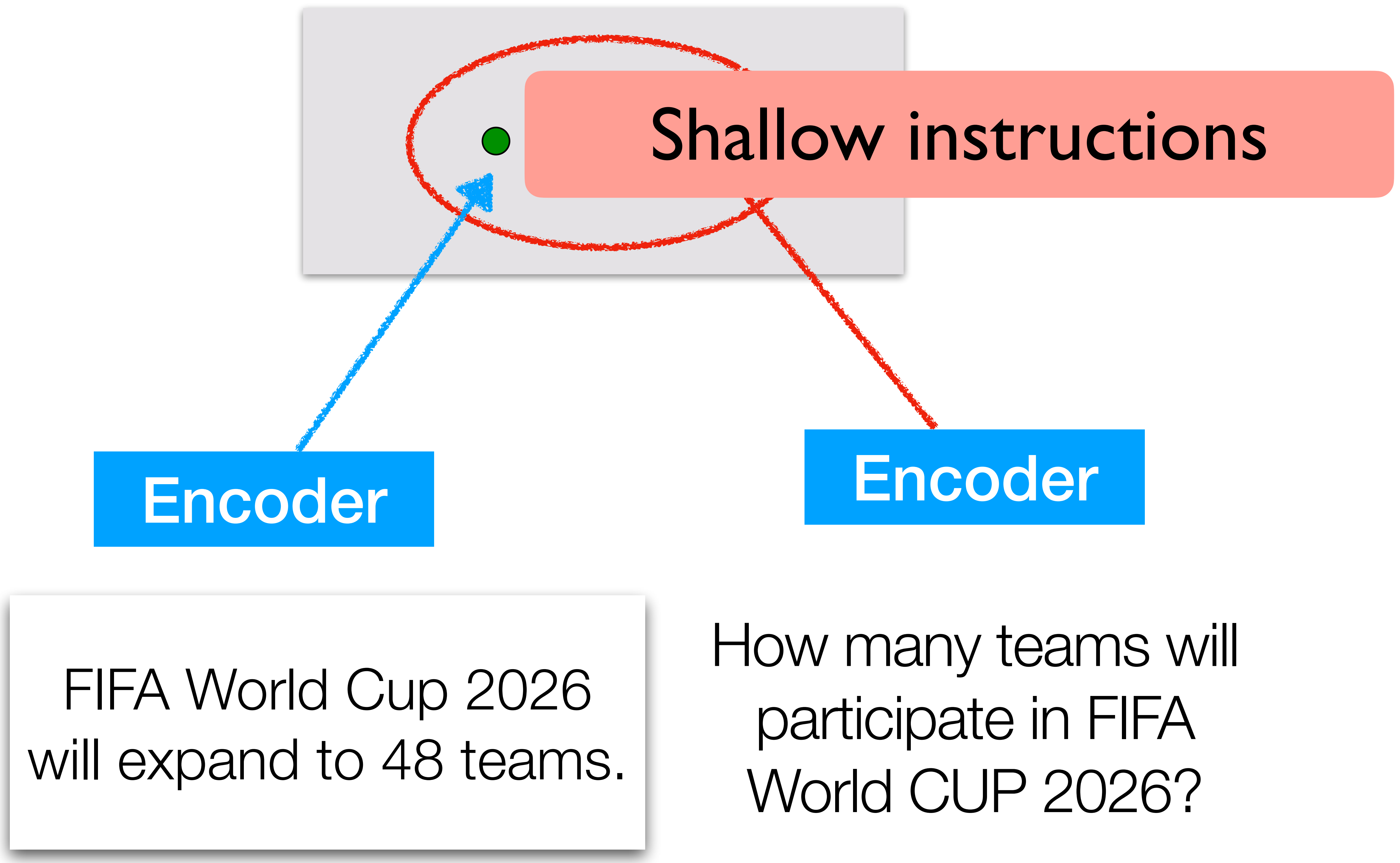
Cross-Encoder

FIFA World Cup 2026 will expand to 48 teams.

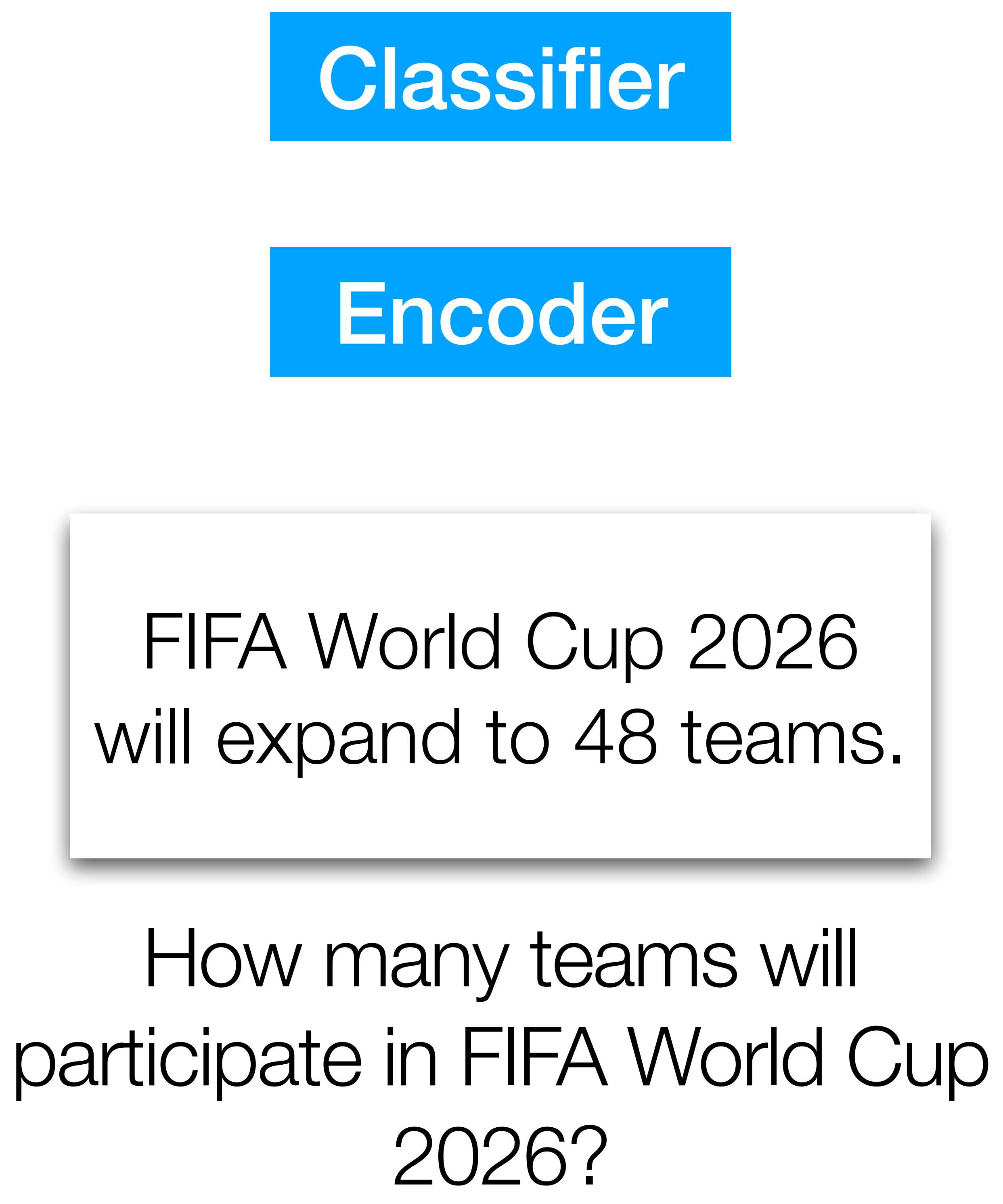
How many teams will participate in FIFA World Cup 2026?

Reranking with Cross-encoder

Bi-Encoder

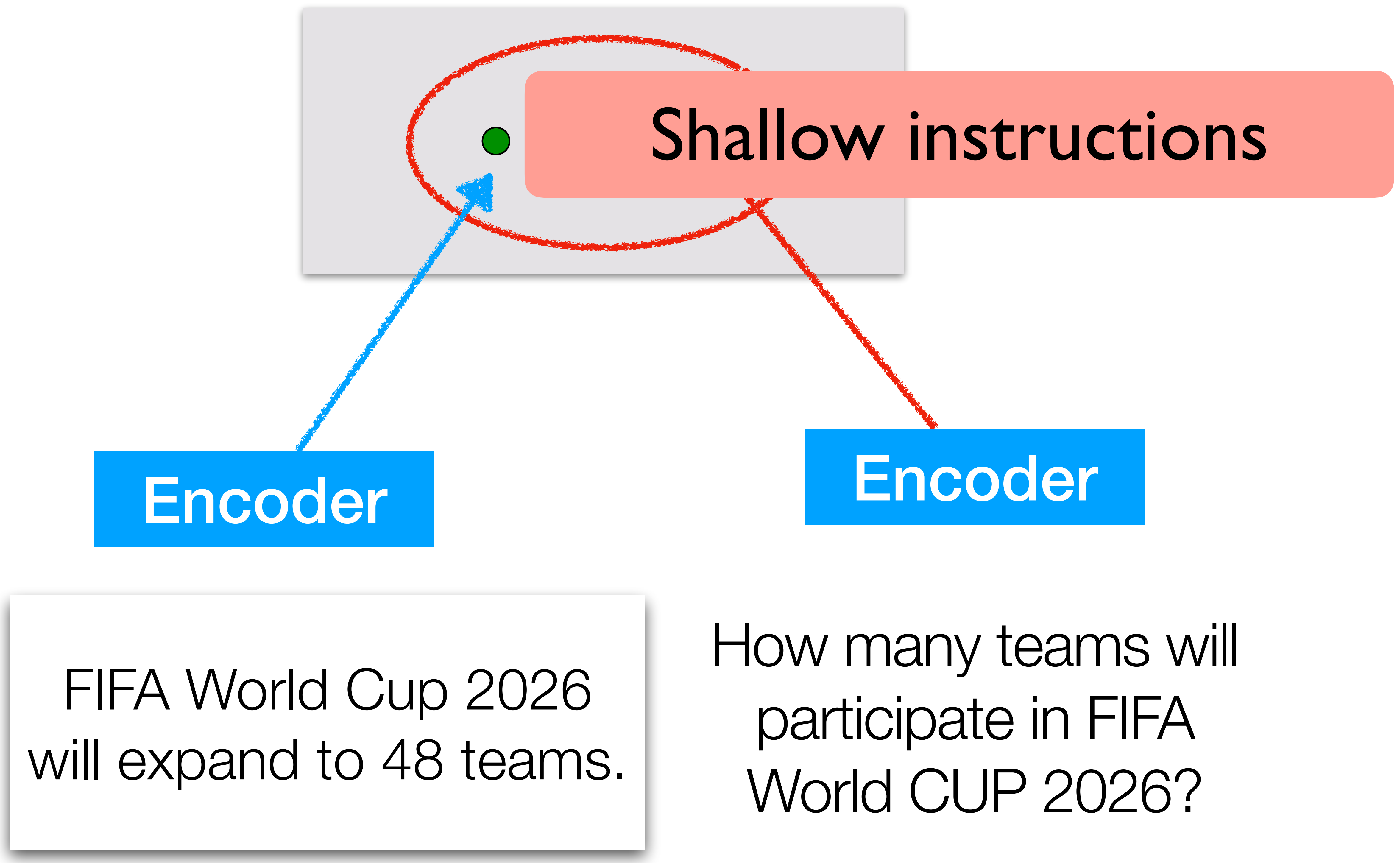


Cross-Encoder

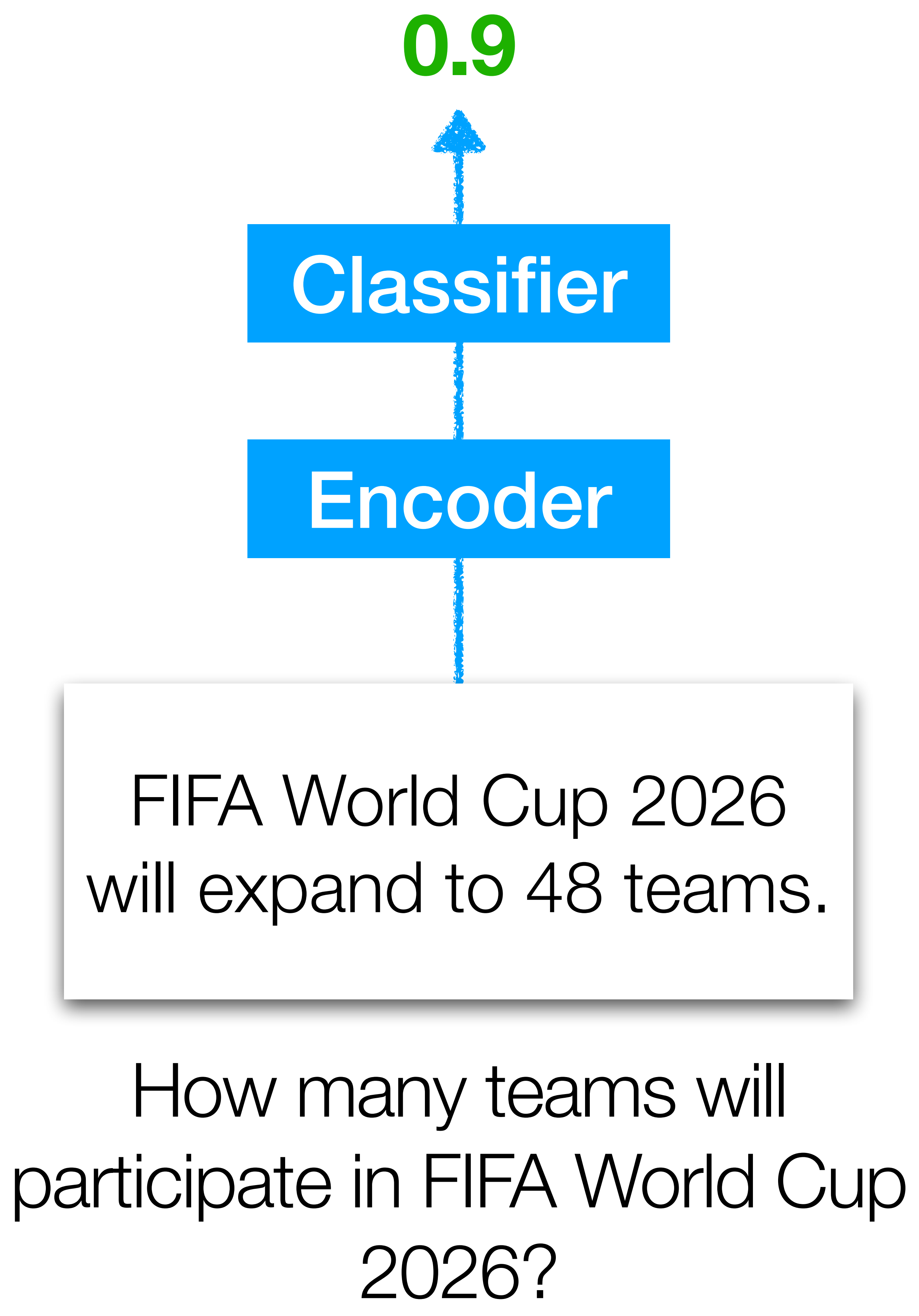


Reranking with Cross-encoder

Bi-Encoder



Cross-Encoder



Evaluation Metrics for Retriever

Manning et al. 2009. An Introduction to Information Retrieval.

Evaluation Metrics for Retriever

Evaluation of **unranked** retrieval sets

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

Manning et al. 2009. An Introduction to Information Retrieval.

Evaluation Metrics for Retriever

Evaluation of **unranked** retrieval sets

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

Evaluation of **ranked** retrieval sets

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

Manning et al. 2009. An Introduction to Information Retrieval.

Evaluation Metrics for Retriever

Evaluation of **unranked** retrieval sets

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

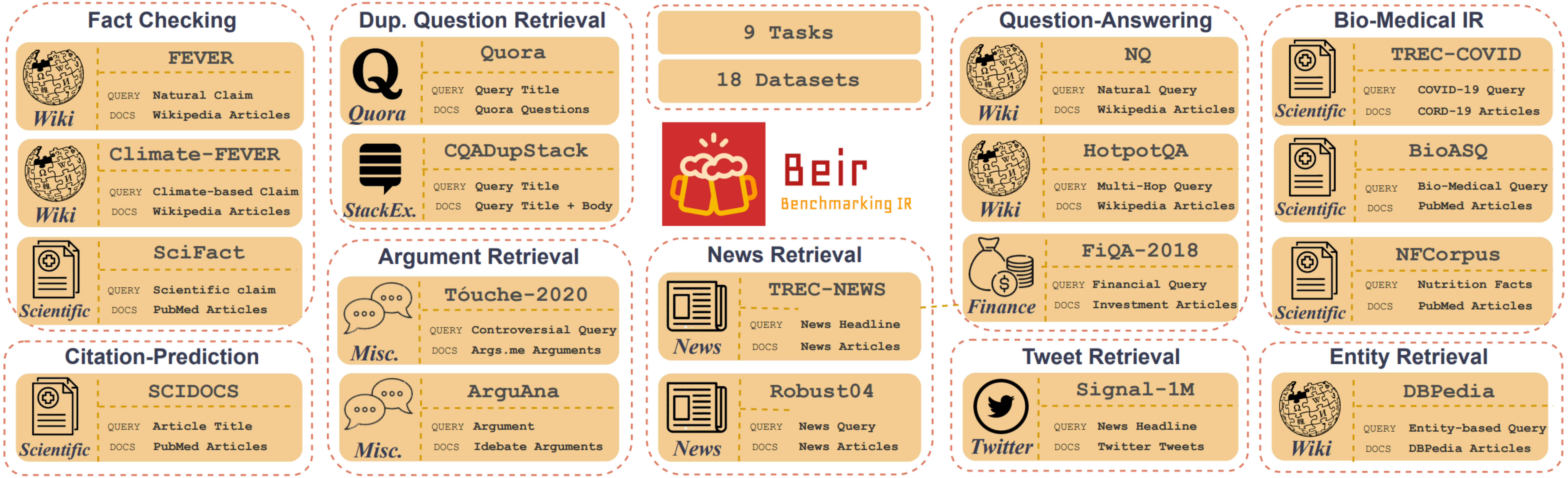
Evaluation of **ranked** retrieval sets

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

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

Manning et al. 2009. An Introduction to Information Retrieval.

Retrieval Benchmarks: BEIR and MTEB



Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.
Muennighoff et al. 2022. MTEB: Massive Text Embedding Benchmark.

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.

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

Adding CE (cross-encoder) helps

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

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

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

Izacard et al. TMLR 2022. Unsupervised Dense Information Retrieval with Contrastive Learning.
Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.

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

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

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

BEIR Results

	BM25	BM25+CE	DPR	Contriever Ours	Contriever 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

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

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

BEIR Results

	BM25	BM25+CE	DPR	Contriever: Ours	Contriever: 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

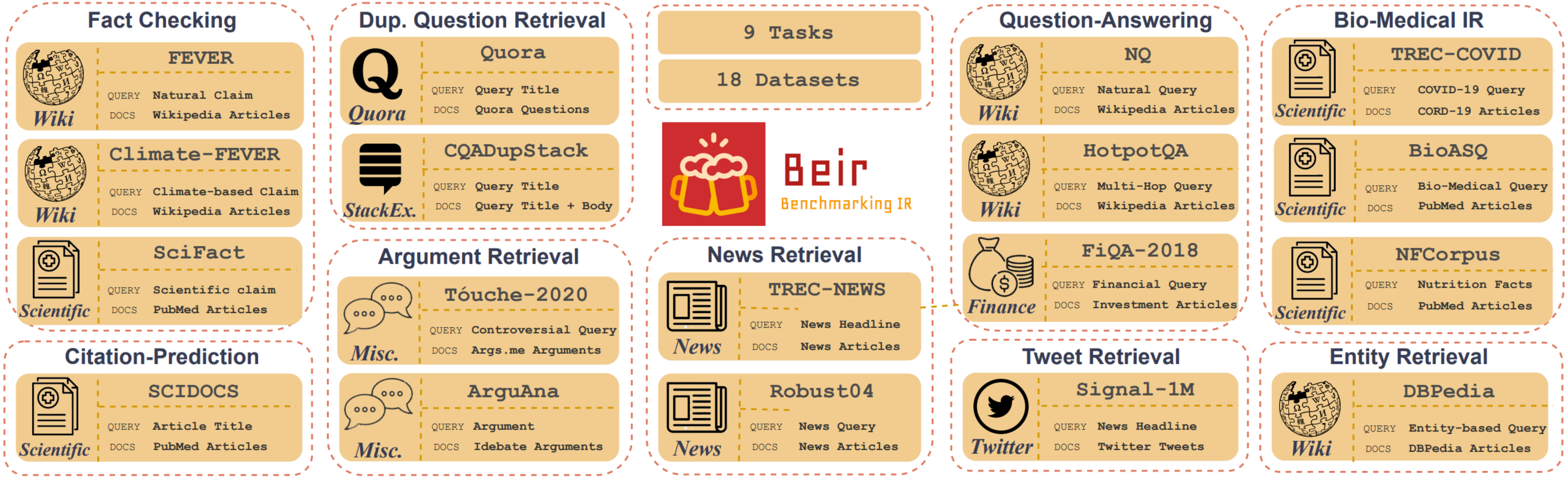
Dense retrievers could struggle in OOD

Unsupervised training helps in OOD

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

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

Retrieval Benchmarks: BEIR and MTEB



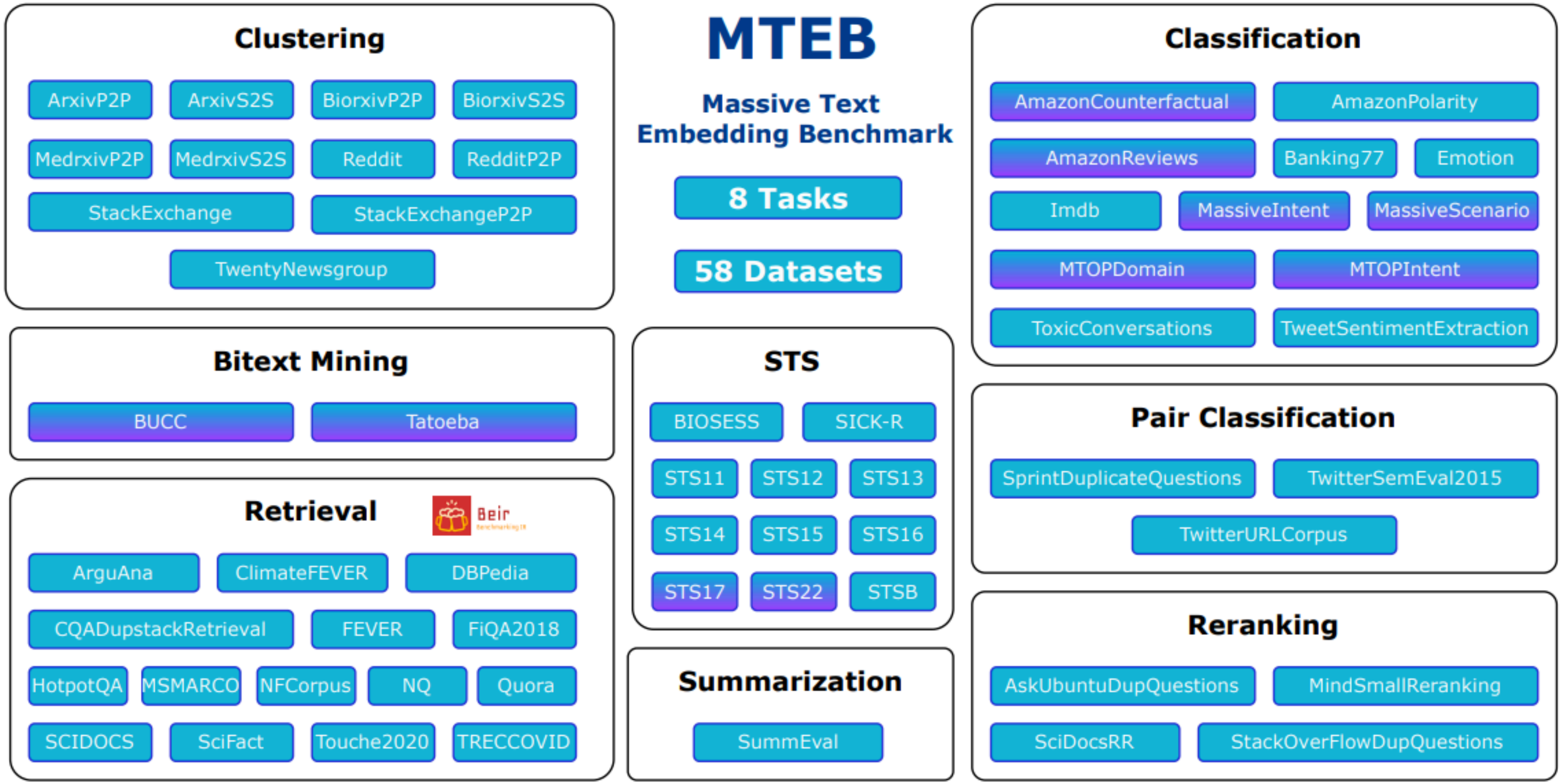
Thakur et al. 2021. NeurIPS D&B. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models.
Muennighoff et al. 2022. MTEB: Massive Text Embedding Benchmark.

Retrieval Benchmarks: BEIR and MTEB

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

Muennighoff et al. 2022. MTEB: Massive Text Embedding Benchmark.

Retrieval Benchmarks: BEIR and MTEB



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

Muennighoff et al. 2022. MTEB: Massive Text Embedding Benchmark.

MTEB Leaderboard

Rank (Bor…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Classificat:
1	QZhou-Embedding	53%	29070	7B	3584	8192	**75.97**	**69.52**	88.97
2	LGAI-Embedding-Preview	56%	27125	7B	4096	32768	74.12	68.40	89.97
3	Seed1.5-Embedding	56%	Unknown	Unknown	2048	32768	74.76	68.56	89.88
4	Qwen3-Embedding-8B	95%	28866	7B	4096	32768	75.22	68.71	90.43
5	Seed1.6-embedding	53%	Unknown	Unknown	2048	32768	74.07	67.98	92.42
6	Qwen3-Embedding-4B	95%	15341	4B	2560	32768	74.60	68.10	89.84
7	gemini-embedding-001	95%	Unknown	Unknown	3072	2048	73.30	67.67	90.05
8	jasper_en_vision_language_v1	56%	3802	1B	8960	131072	71.41	66.65	90.27
9	Linq-Embed-Mistral	95%	13563	7B	4096	32768	69.80	65.29	83.00
10	SFR-Embedding-Mistral	85%	13563	7B	4096	32768	69.31	64.94	80.47
11	NV-Embed-v2	56%	14975	7B	4096	32768	69.81	65.00	87.19

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

MTEB Leaderboard

Rank (Bor…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Classificat:
1	QZhou-Embedding	53%	29070	7B	3584	8192	**75.97**	**69.52**	88.97
2	LGAI-Embedding-Preview	56%	27125	7B	4096	32768	74.12	68.40	89.97
3	Seed1.5-Embedding	56%	Unknown	Unknown	2048	32768	74.76	68.56	89.88
4	Qwen3-Embedding-8B	95%	28866	7B	4096	32768	75.22	68.71	90.43
5	Seed1.6-embedding	53%	Unknown	Unknown	2048	32768	74.07	67.98	92.42
6	Qwen3-Embedding-4B	95%	15341	4B	2560	32768	74.60	68.10	89.84
7	gemini-embedding-001	95%	Unknown	Unknown	3072	2048	73.30	67.67	90.05
8	jasper_en_vision_language_v1	56%	3802	1B	8960	131072	71.41	66.65	90.27
9	Linq-Embed-Mistral								
10	SFR-Embedding-Mistral								
11	NV-Embed-v2	56%	14975	7B	4096	32768	69.81	65.00	87.19

Instruction-tuned retrievers

$$q^+_{\text{inst}} = \texttt{Instruct}:\{\texttt{task\_definition}\}\ \texttt{Query}:q^+$$

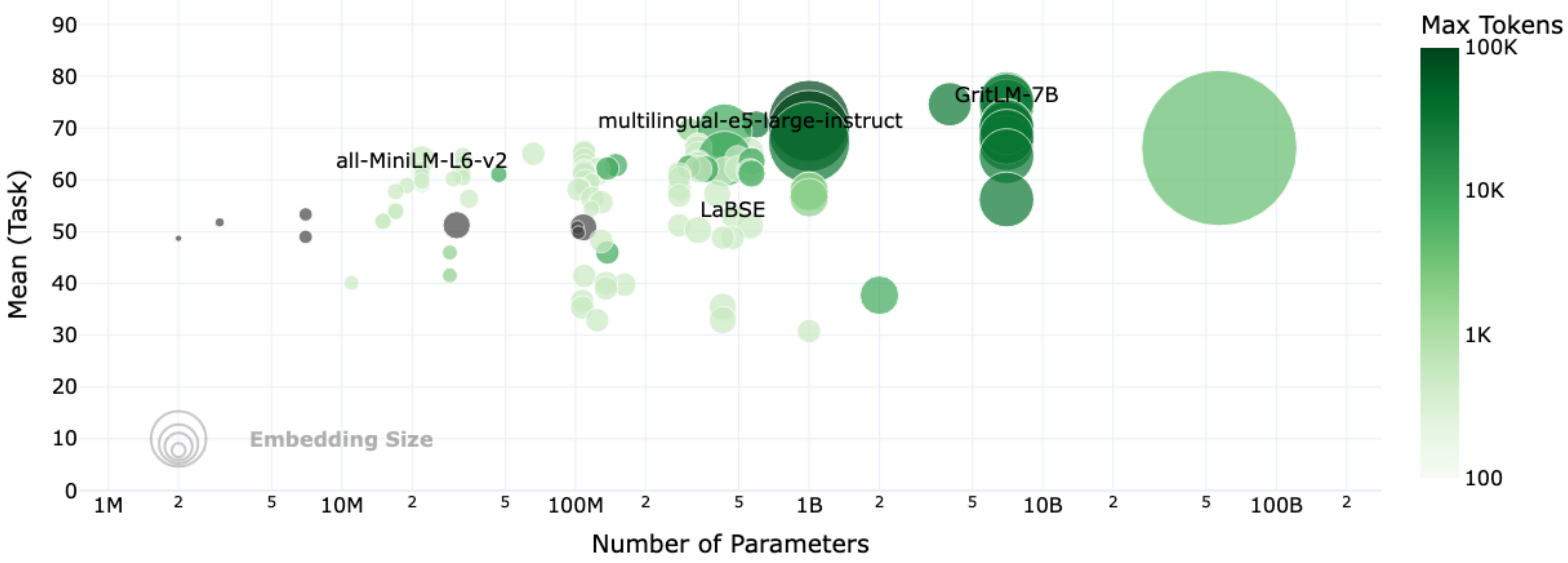
https://huggingface.co/spaces/mteb/leaderboard

MTEB Leaderboard

Rank (Bor…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Classificat:
1	QZhou-Embedding	53%	29070	7B	3584	8192	**75.97**	**69.52**	88.97
2	LGAI-Embedding-Preview	56%	27125	7B	4096	32768	74.12	68.40	89.97
3	Seed1.5-Embedding	56%	Unknown	Unknown	2048	32768	74.76	68.56	89.88
4	Qwen3-Embedding-8B	95%	28866	7B	4096	32768	75.22	68.71	90.43
5	Seed1.6-embedding	53%	Unknown	Unknown	2048	32768	74.07	67.98	92.42
6	Qwen3-Embedding-4B	95%	15341	4B	2560	32768	74.60	68.10	89.84
7	gemini-embedding-001	95%	Unknown	Unknown	3072	2048	73.30	67.67	90.05
8	jasper_en_vision_language_v1	56%	3802	1B	8960	131072	71.41	66.65	90.27
9	Linq-Embed-Mistral	95%	13563	7B	4096	32768	69.80	65.29	83.00
10	SFR-Embedding-Mistral	85%	13563	7B	4096	32768	69.31	64.94	80.47
11	NV-Embed-v2	56%	14975	7B	4096	32768	69.81	65.00	87.19

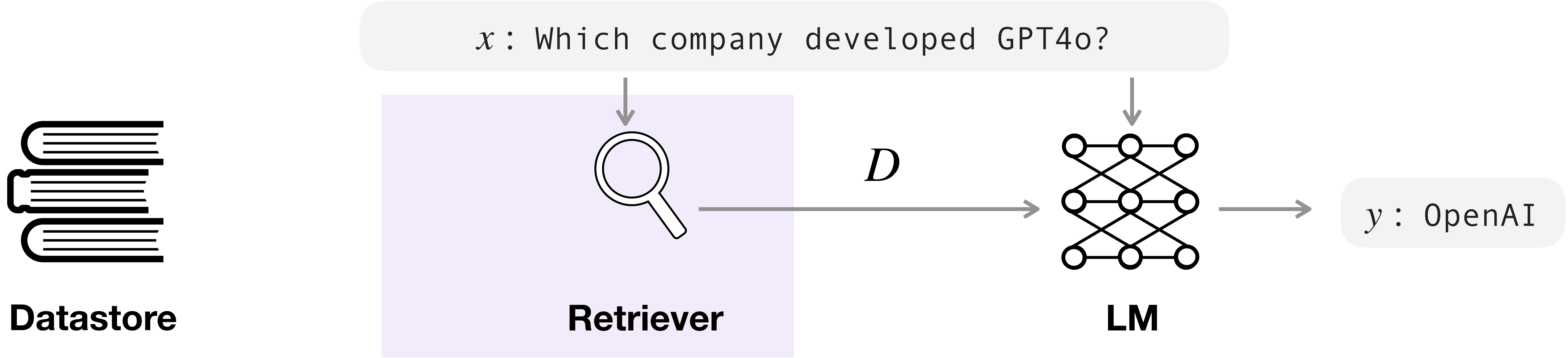
https://huggingface.co/spaces/mteb/leaderboard

MTEB Leaderboard



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

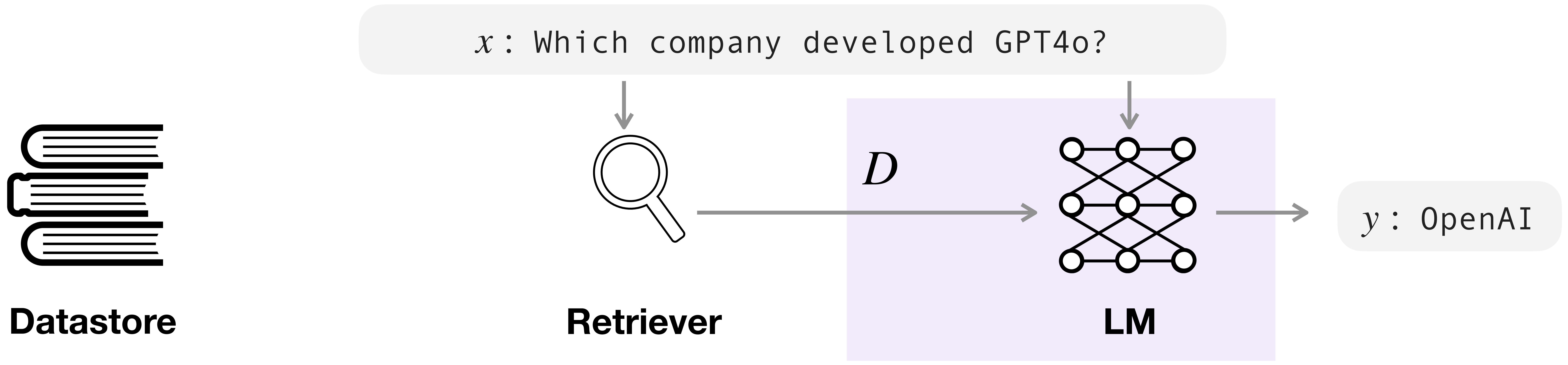
Summary of Part 2



- ✓ 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



✓ Common architectures

✓ Recent progress

Categorizing Retrieval-Augmented LMs

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query

Text chunks (passages)?

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query

Text chunks (passages)?
Tokens?

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query

Text chunks (passages)?
Tokens?
Something else?

Categorizing Retrieval-Augmented LMs

What to retrieve?

Query

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.

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 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 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 w/r w/r

The capital city of Ontario is Toronto.

How to use retrieval?

Input

LM

Output

Categorizing Retrieval-Augmented LMs

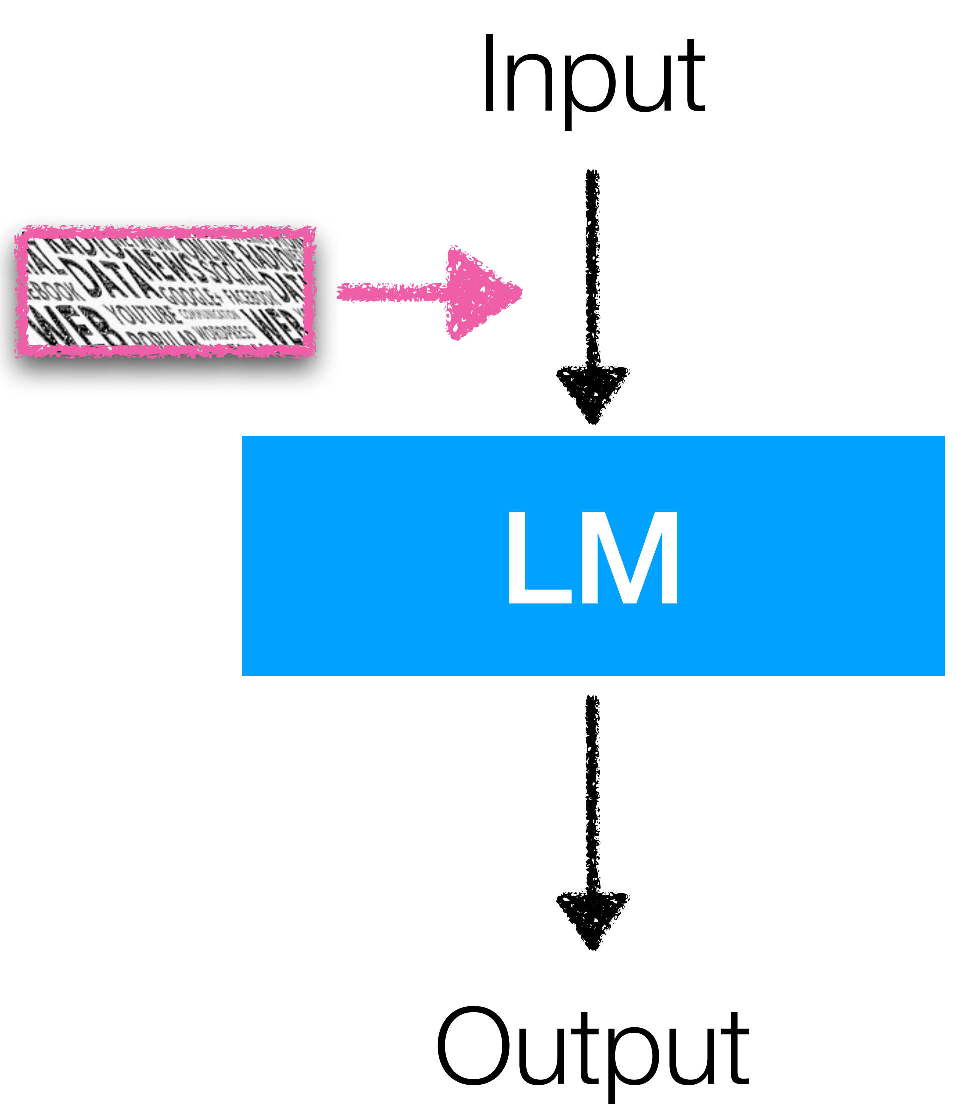
What to retrieve?

Query

Text chunks (passages)?
Tokens?
Something else?

When to retrieve?

How to use retrieval?



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.

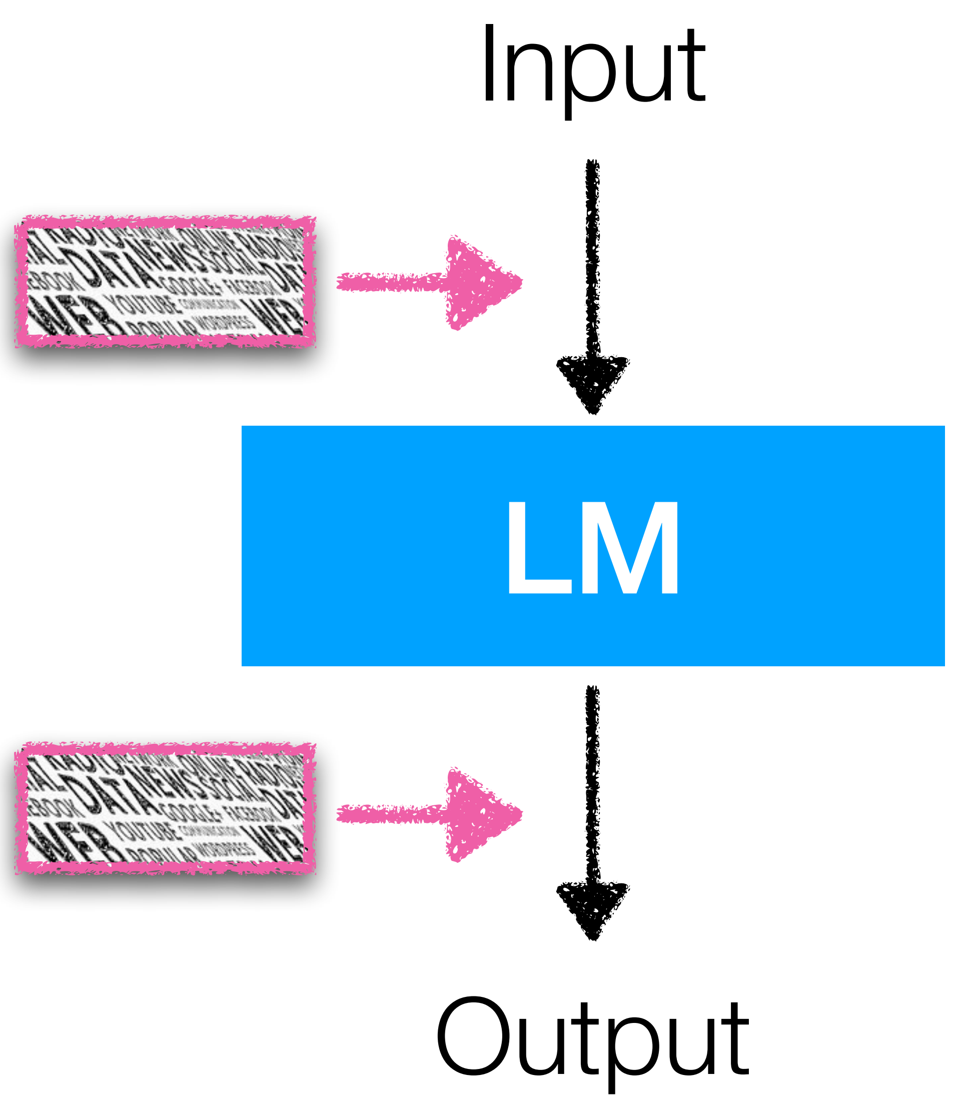
w/ retrieval w/ r w/r w/r w/ r w/r w/r

The capital city of Ontario is Toronto.

w/ retrieval w/r w/r

The capital city of Ontario is Toronto.

How to use retrieval?



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.

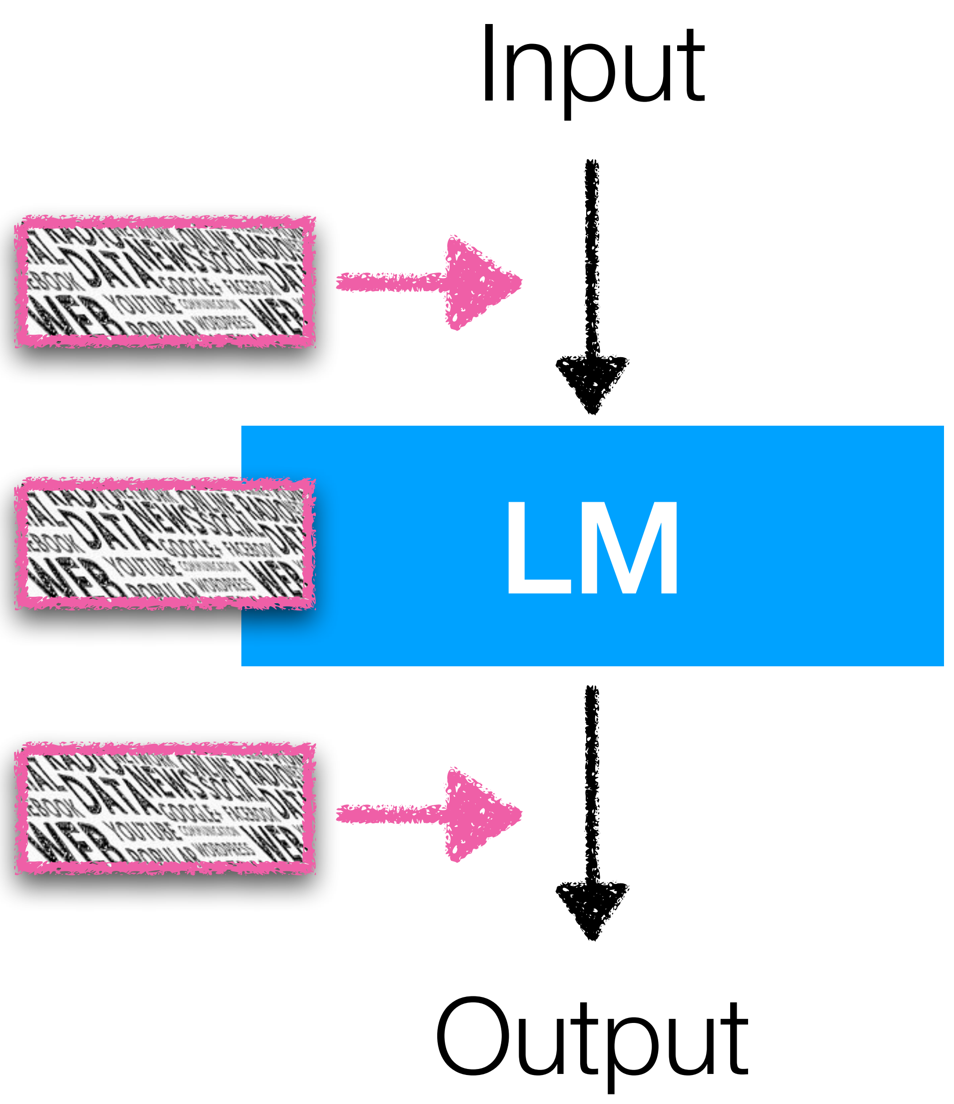
w/ retrieval w/ r w/r w/r w/ r w/r w/r

The capital city of Ontario is Toronto.

w/ retrieval w/r w/r

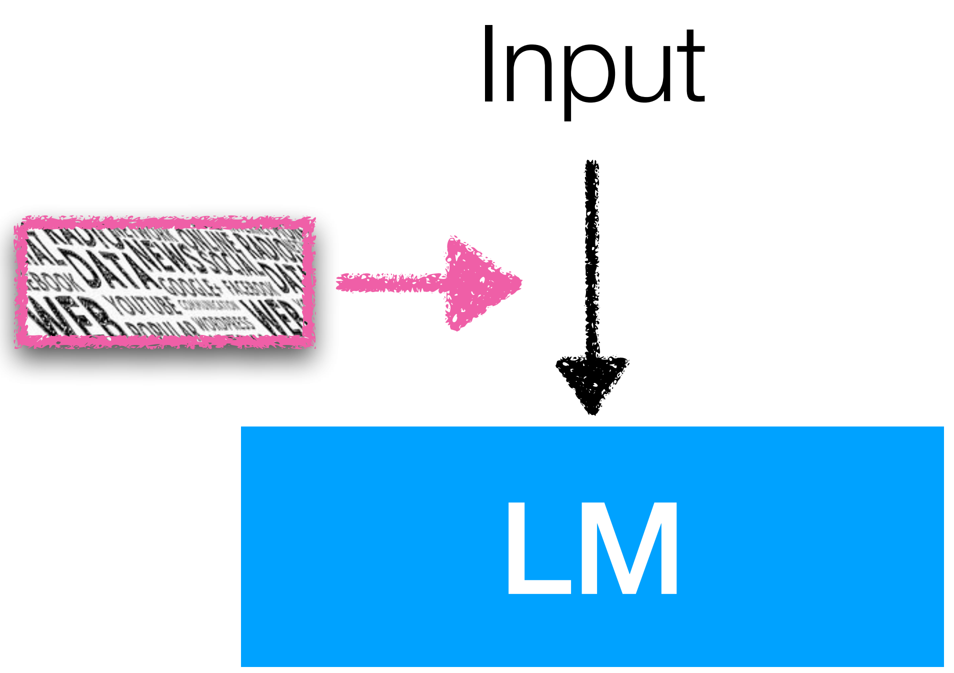
The capital city of Ontario is Toronto.

How to use retrieval?



How to Use Retrieval

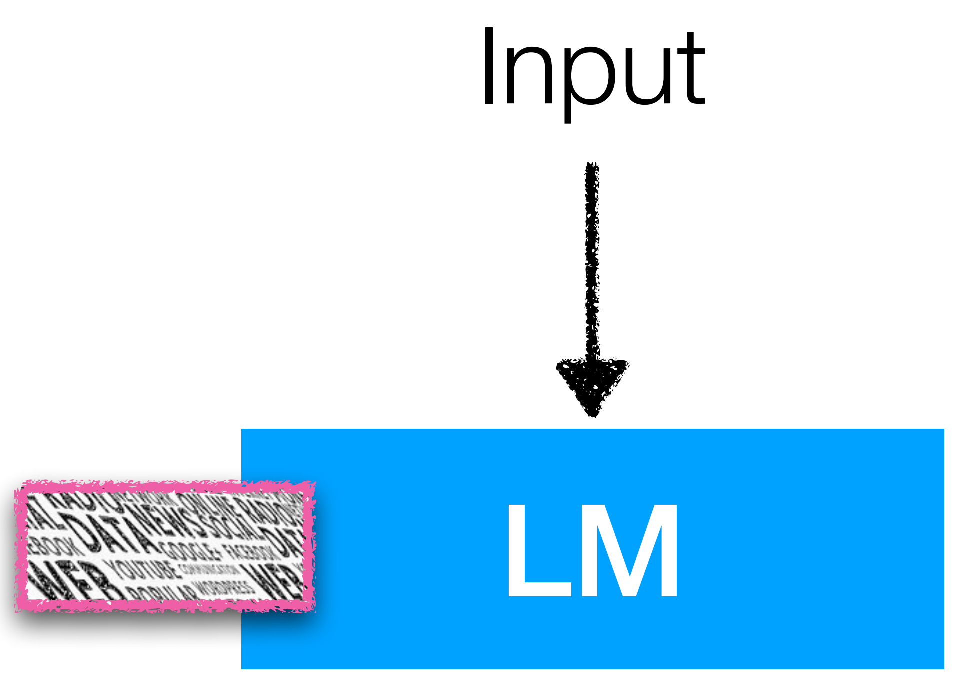
Input Augmentation



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

e.g., RAG

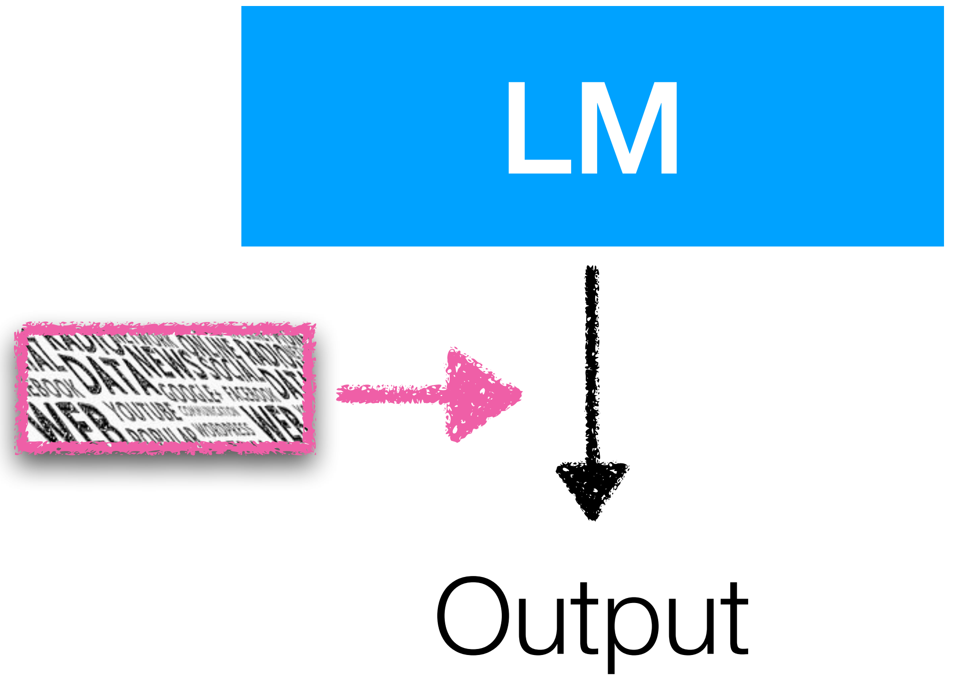
Intermediate Fusion



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

e.g., RETRO, InstructRETRO

Output Interpolation

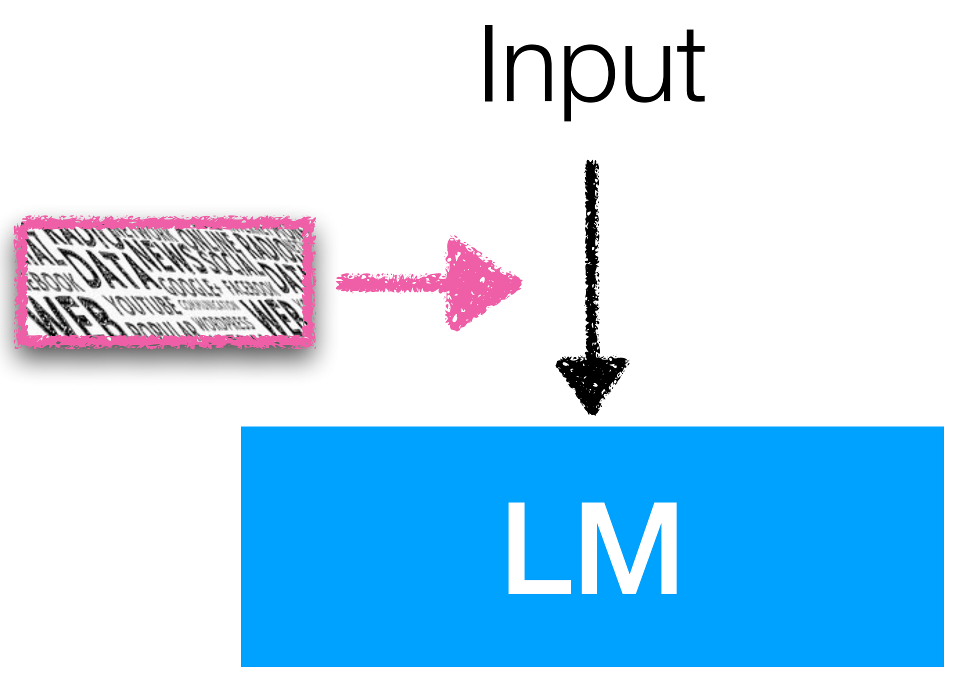


- Directly manipulate output token distributions
- No training required*
- Limited effectiveness on tasks

e.g., kNNLM

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

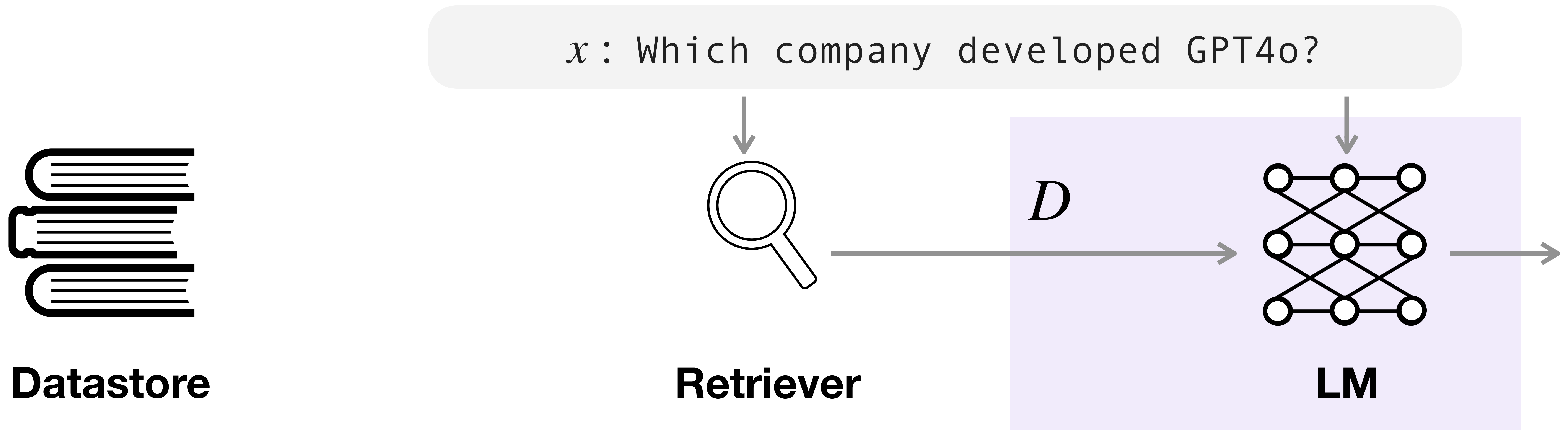
e.g., RETRO, InstructRETRO

Output Interpolation

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

e.g., kNNLM

Retrieval-Augmented Generation (Lewis et al., 2020)

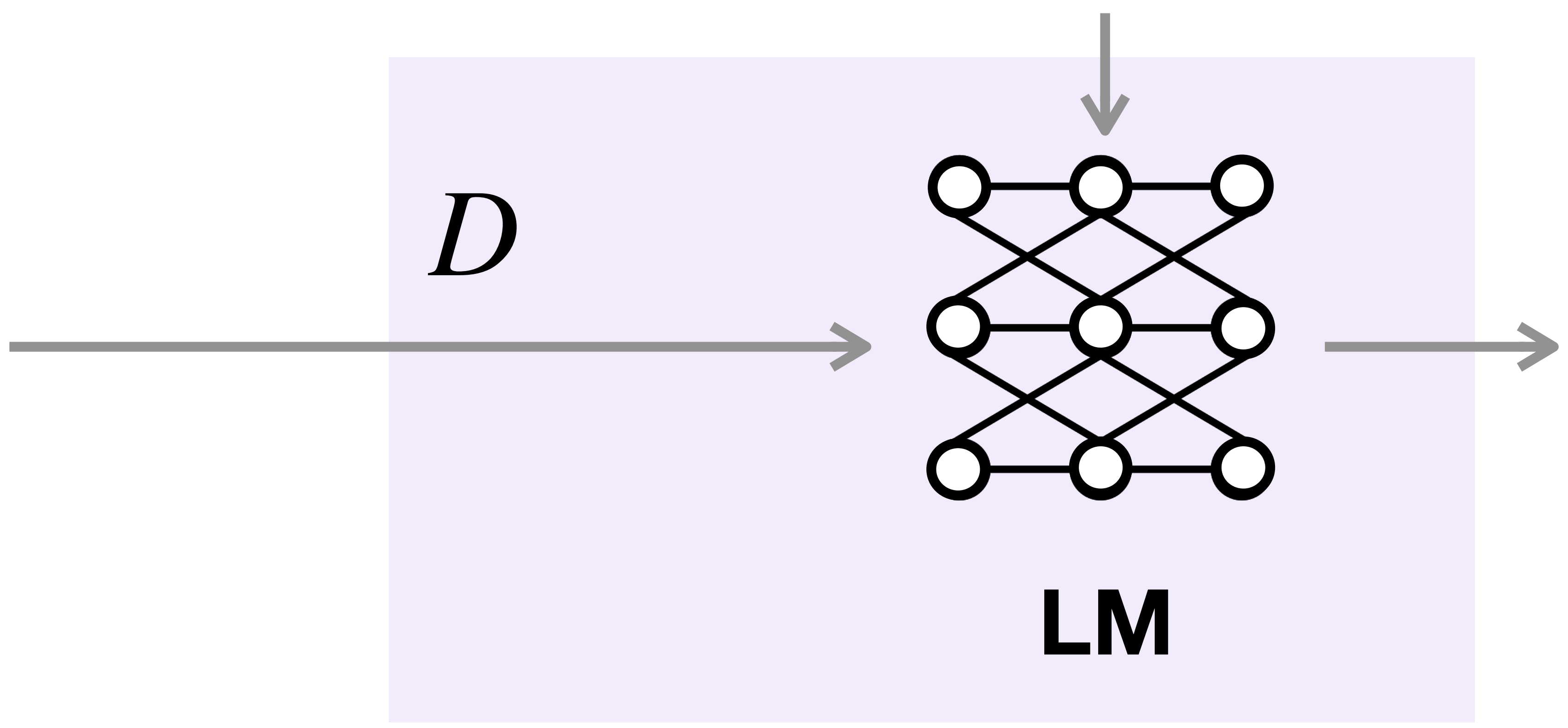


$$D \in \mathrm{Top}_k \mathrm{Sim}(\cdot \mid x)$$

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

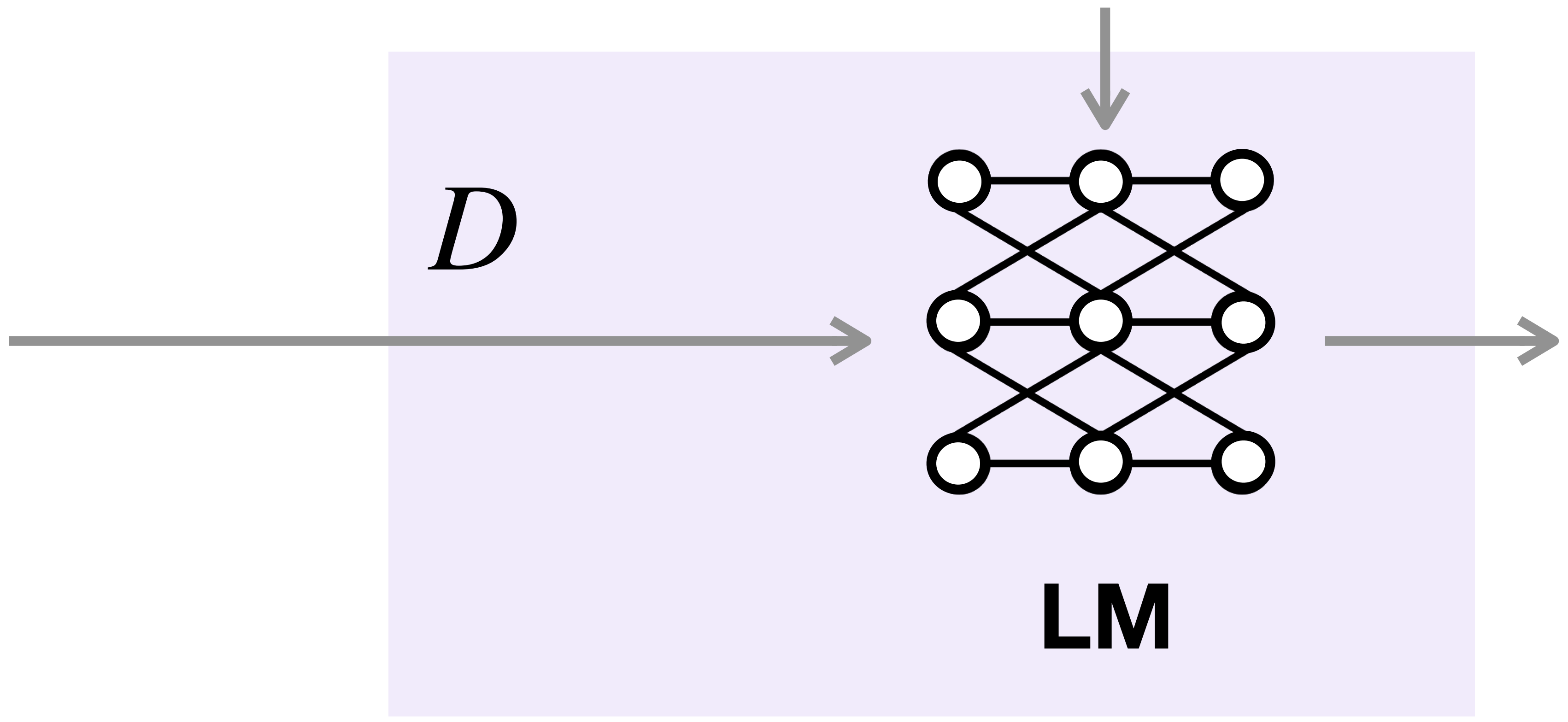
Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Retrieval-Augmented Generation (Lewis et al., 2020)



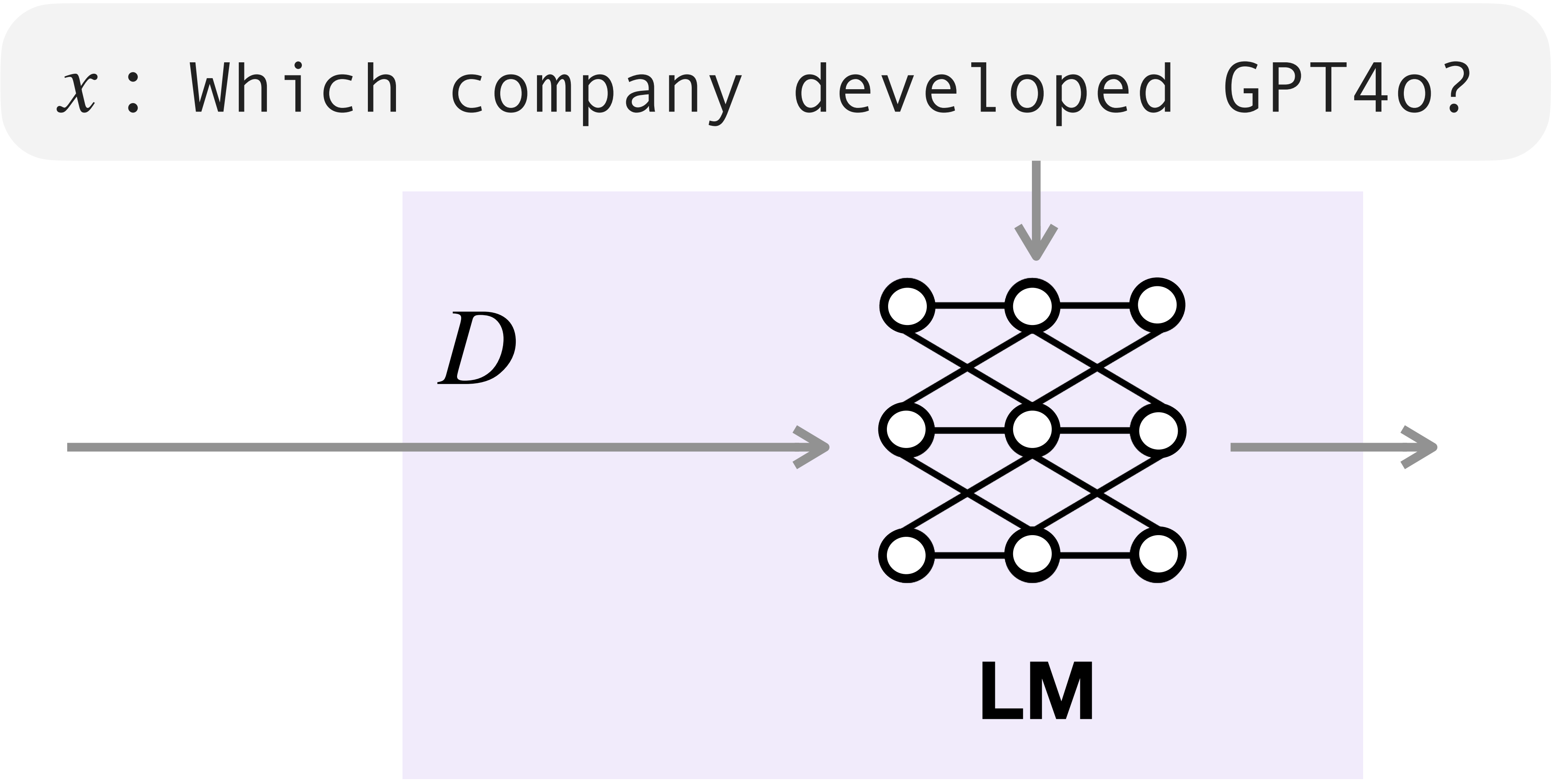
Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Retrieval-Augmented Generation (Lewis et al., 2020)



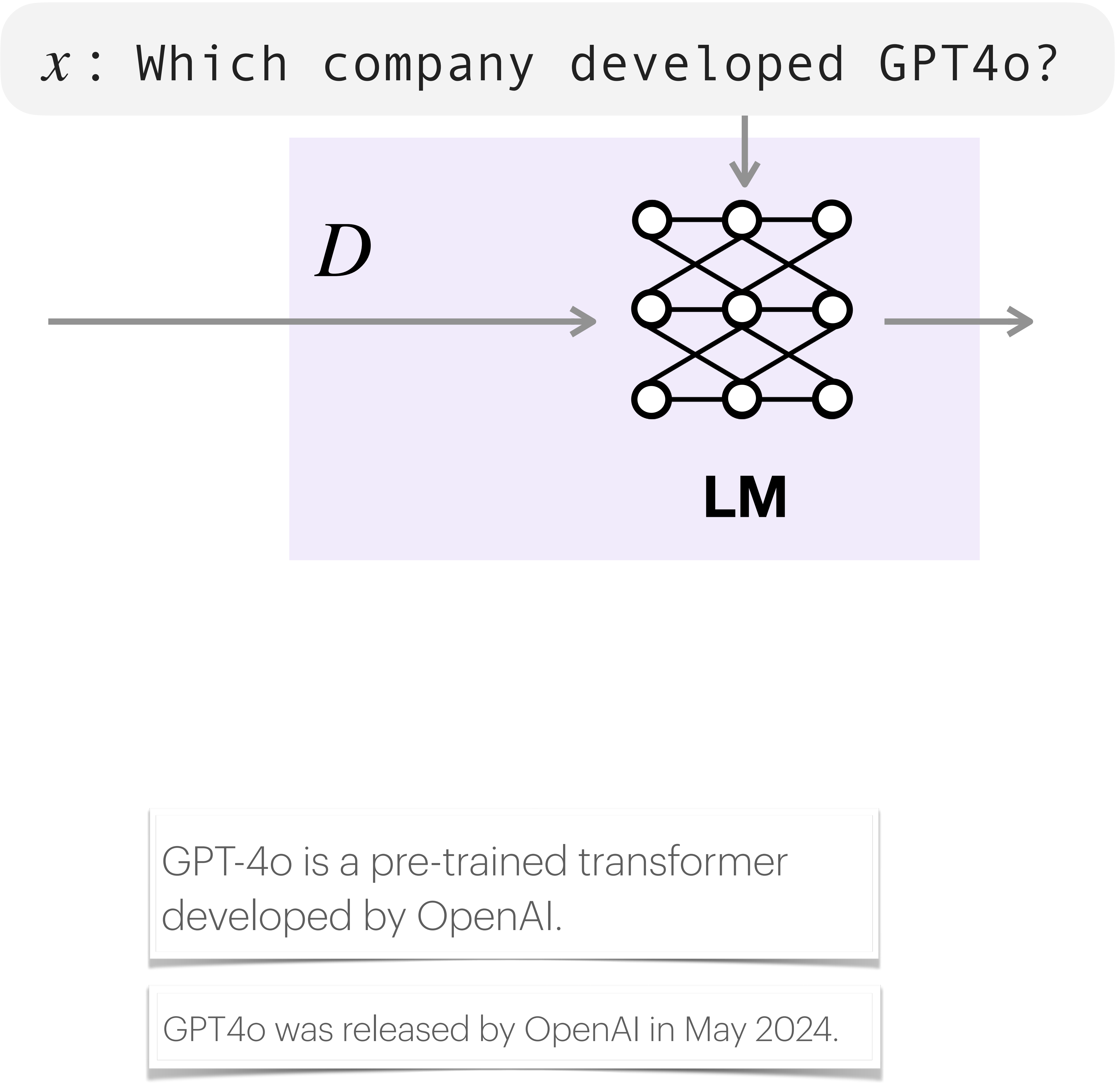
Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Retrieval-Augmented Generation (Lewis et al., 2020)



Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

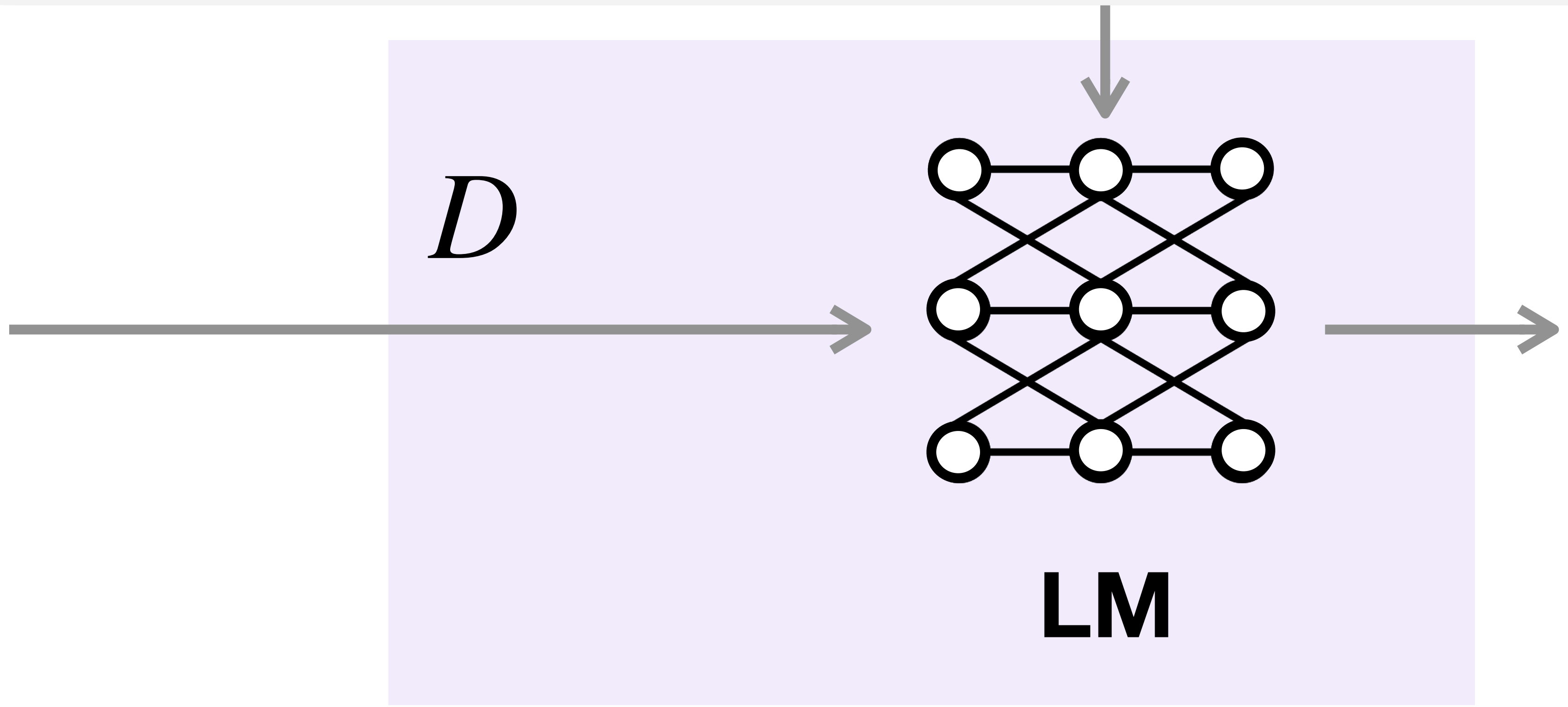
Retrieval-Augmented Generation (Lewis et al., 2020)



Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Retrieval-Augmented Generation (Lewis et al., 2020)

x : Which company developed GPT4o?



Question: Which company developed GPT4o?

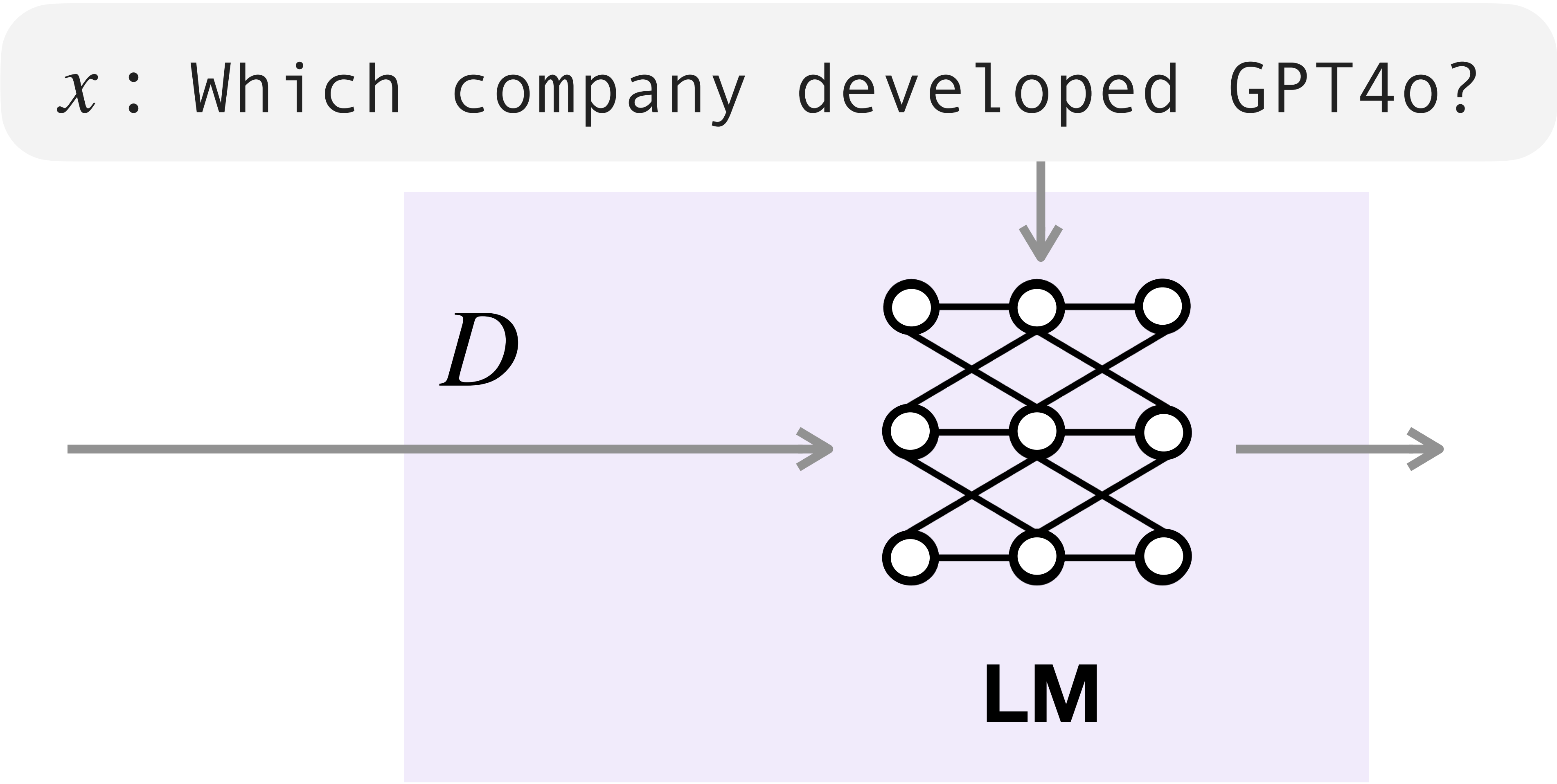
References:

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

GPT4o was released by OpenAI in May 2024.

Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Retrieval-Augmented Generation (Lewis et al., 2020)



Question: Which company developed GPT4o?

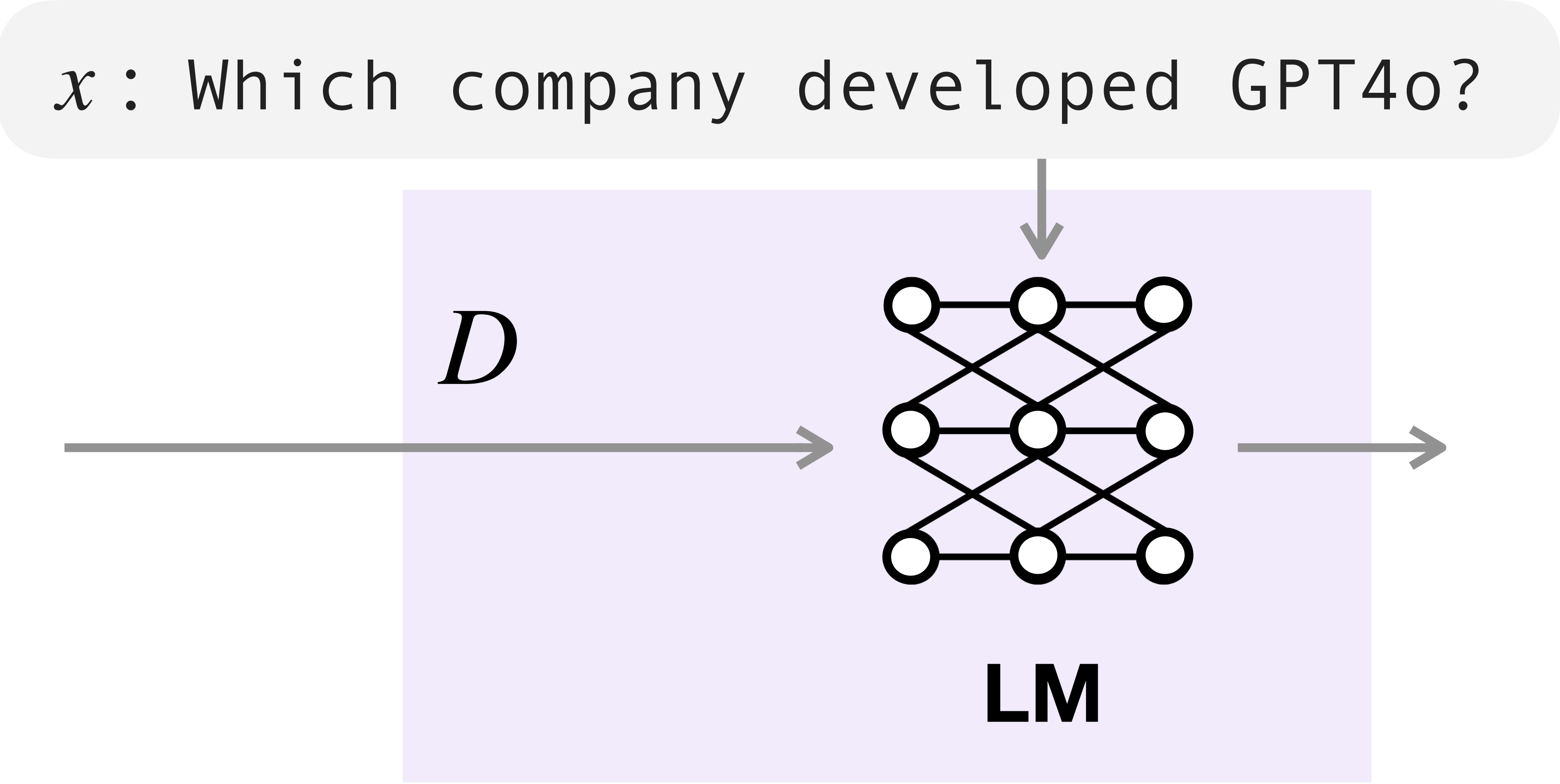
References:

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

GPT4o was released by OpenAI in May 2024.

Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Retrieval-Augmented Generation (Lewis et al., 2020)



Question: `Which company developed GPT4o?`

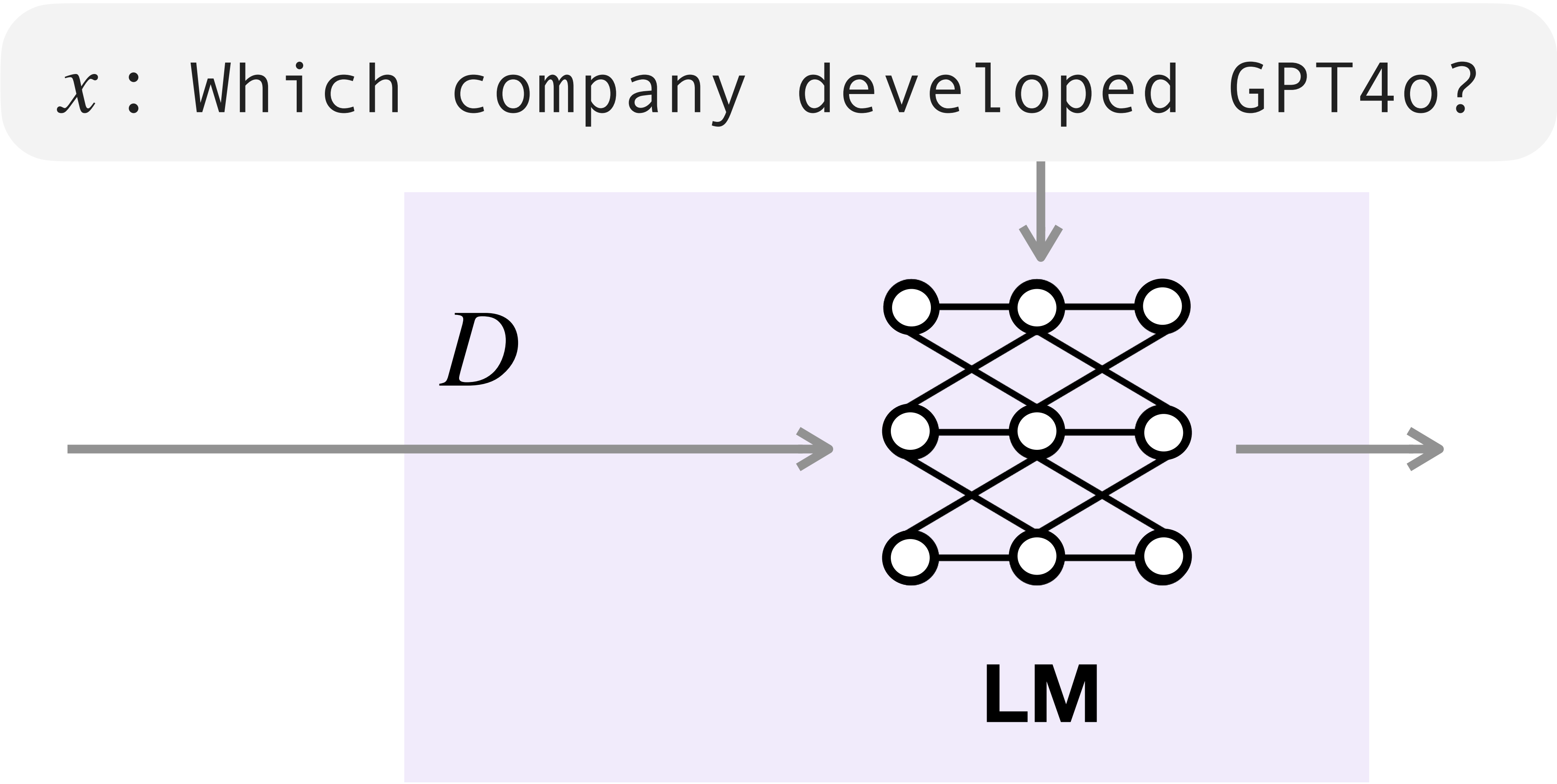
References:

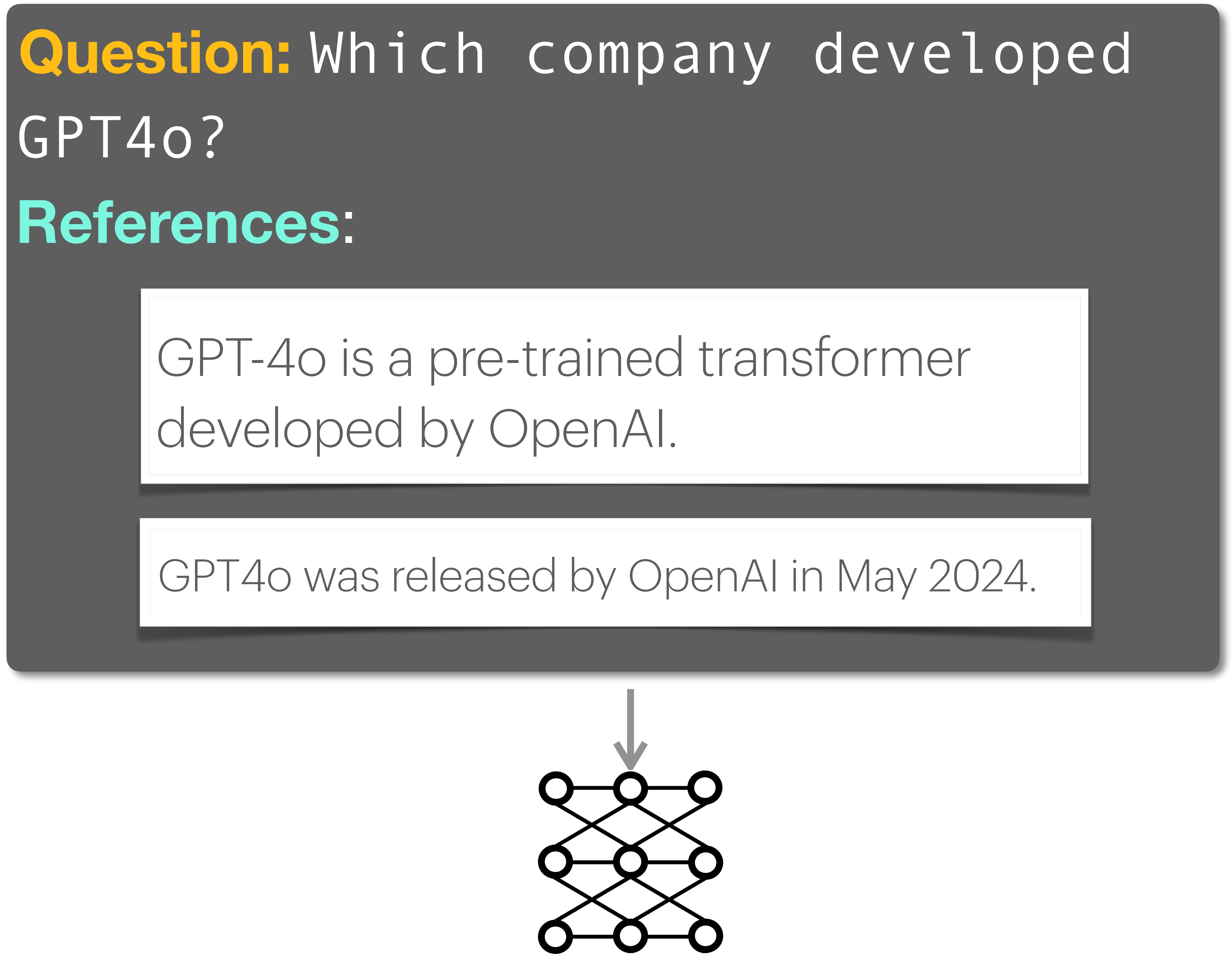
> GPT-4o is a pre-trained transformer developed by OpenAI.

> GPT4o was released by OpenAI in May 2024.

Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

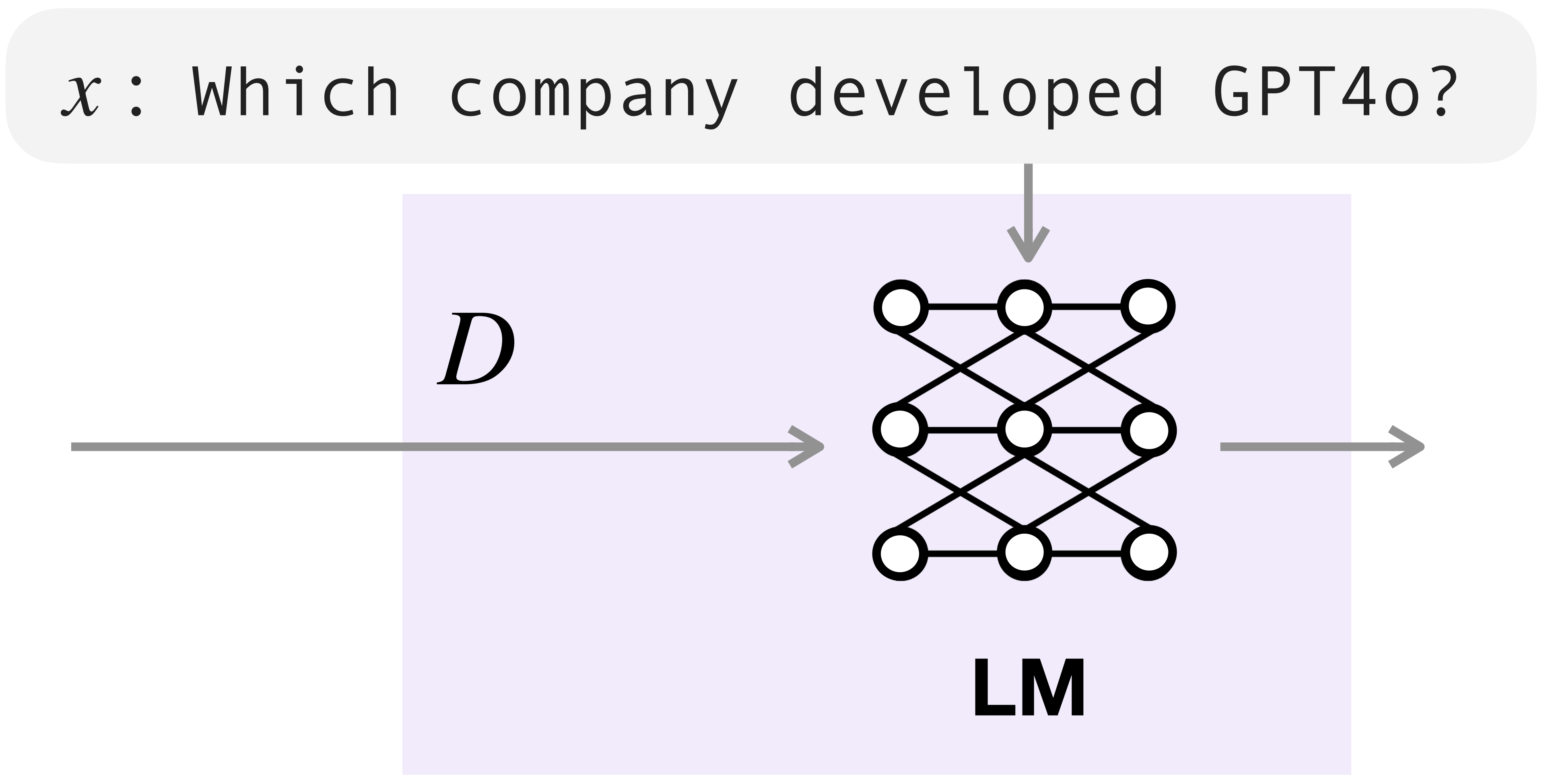
Retrieval-Augmented Generation (Lewis et al., 2020)

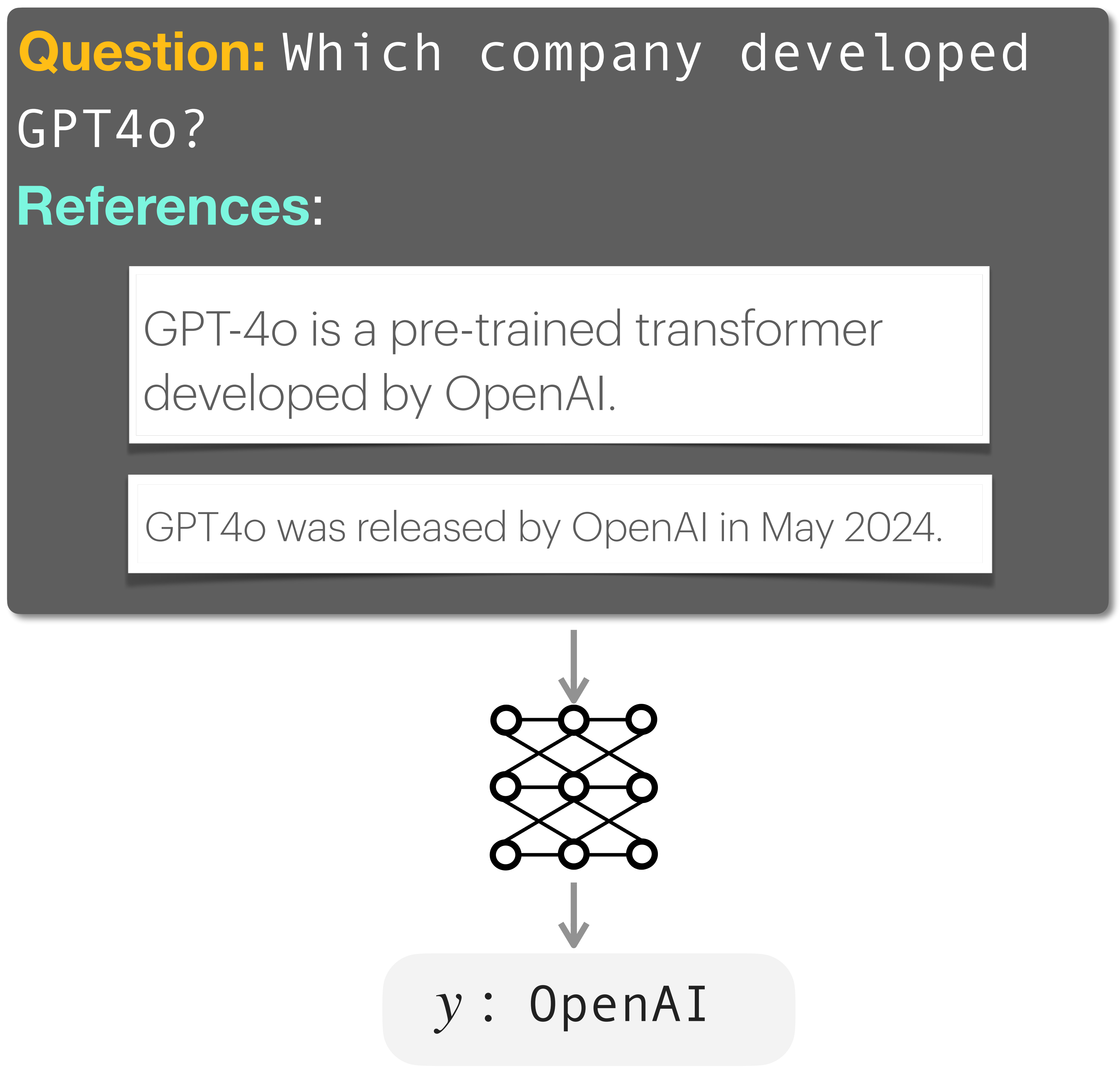




Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Retrieval-Augmented Generation (Lewis et al., 2020)

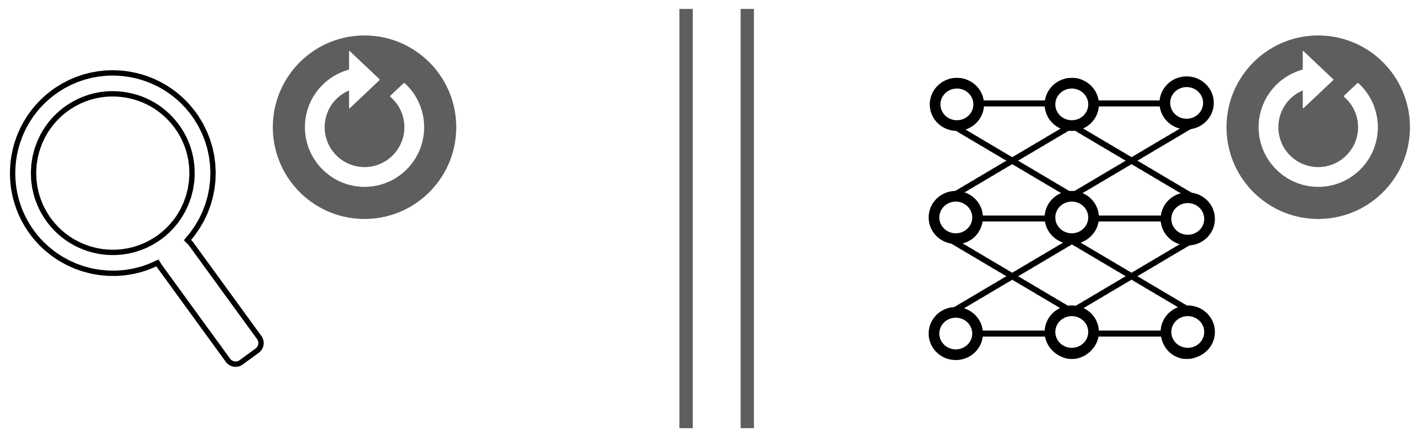




Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Training RAG

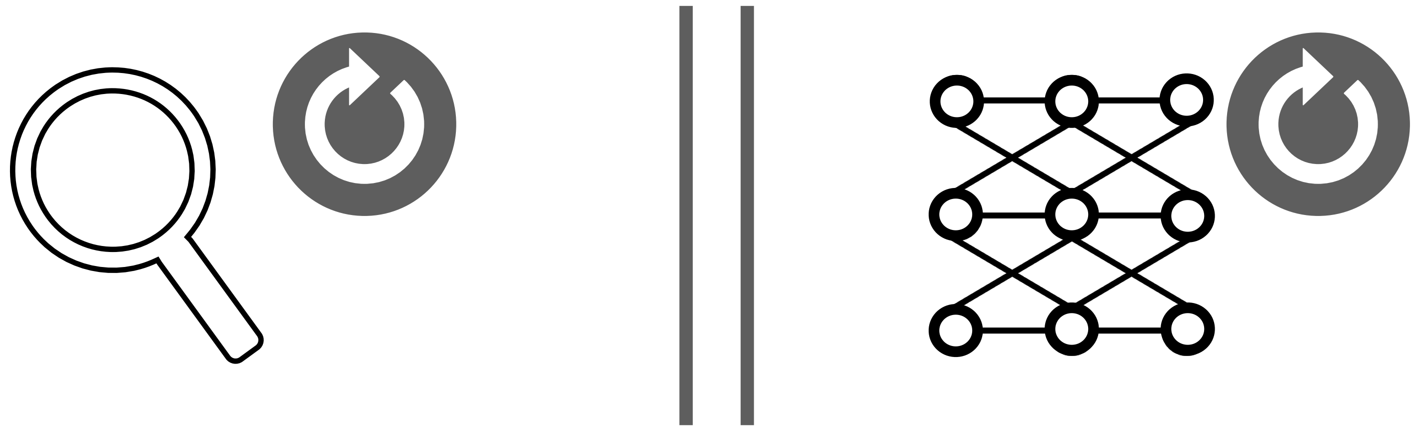
Training RAG

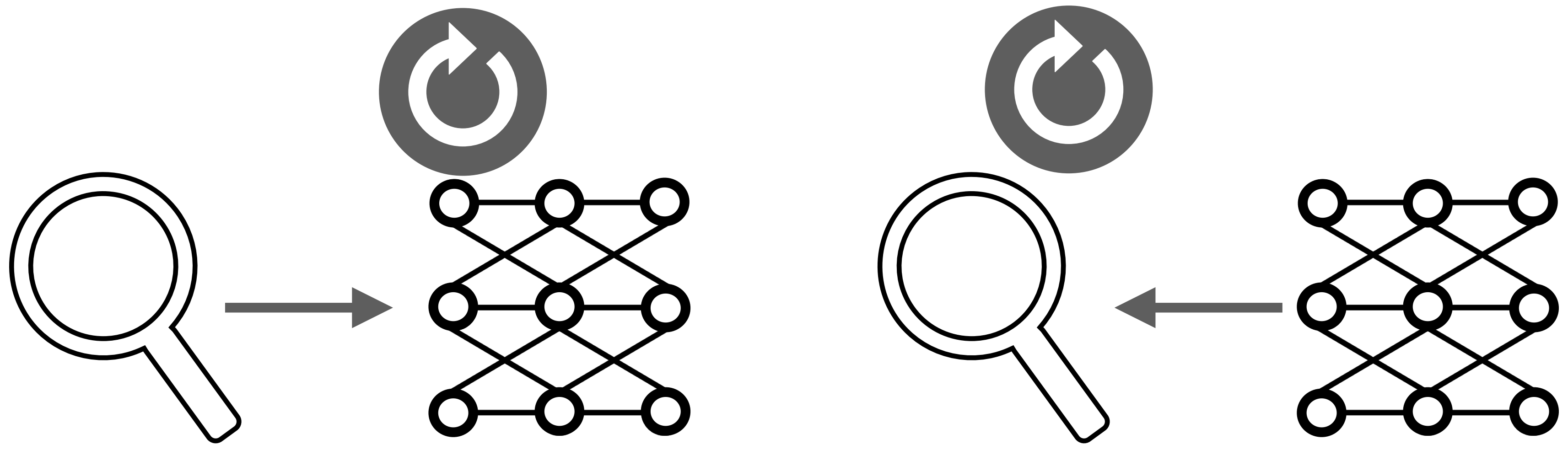


Independent training

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

Training RAG





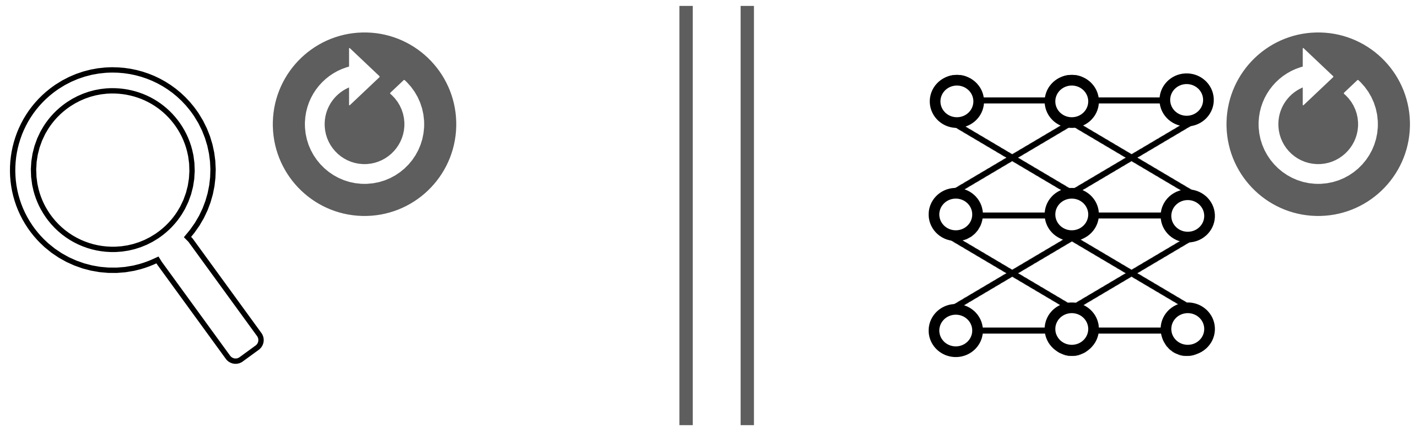
Independent training

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

Sequential training

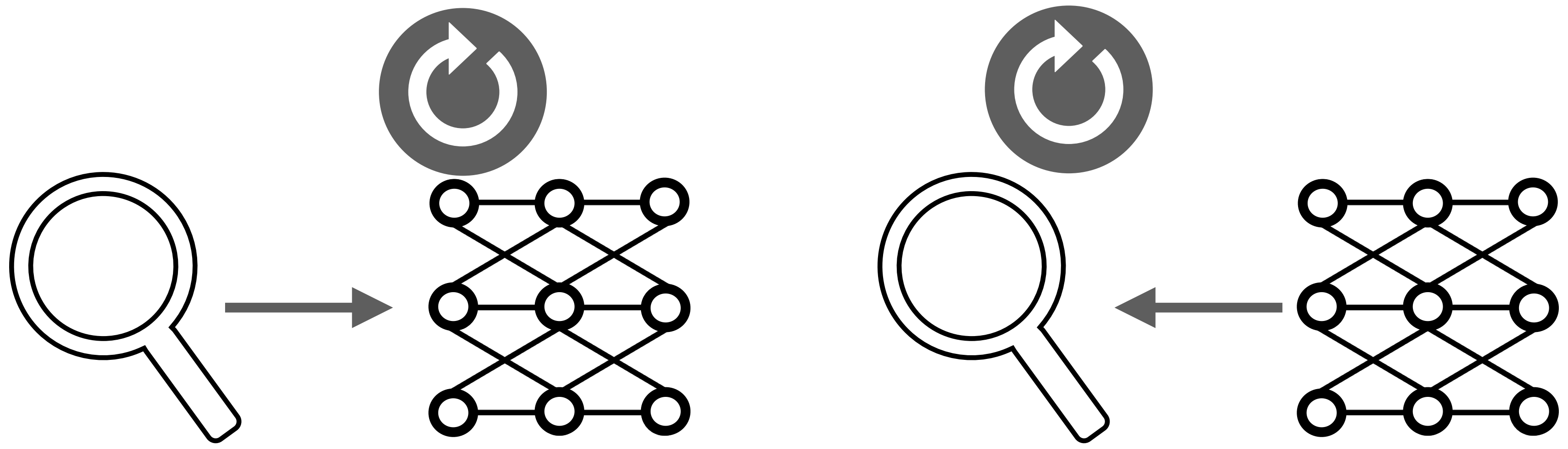
- Evidentiality Generator (Asai et al., 2023)
- REPLUG (Shi et al., 2023)

Training RAG



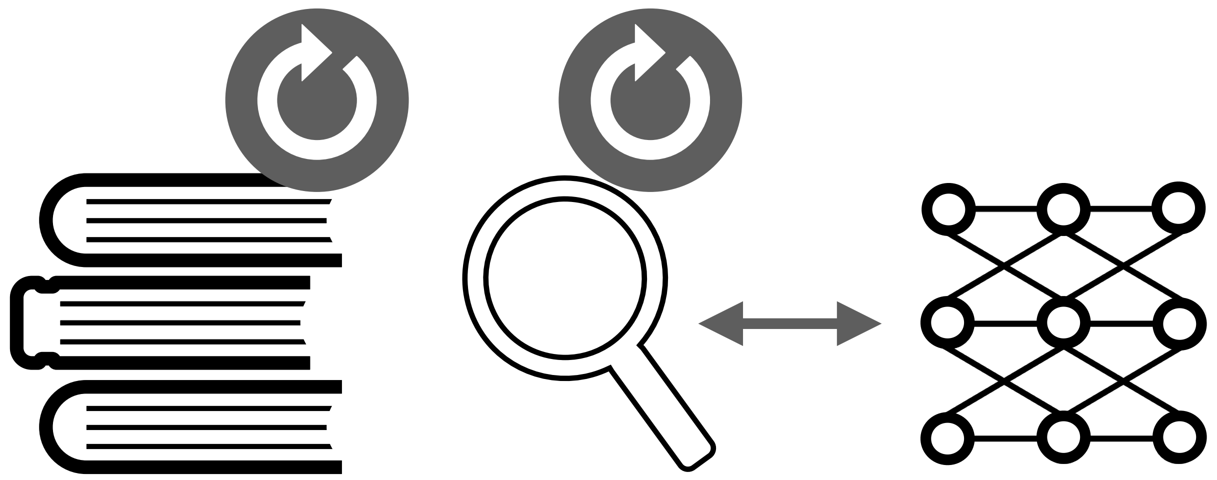
Independent training

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



Sequential training

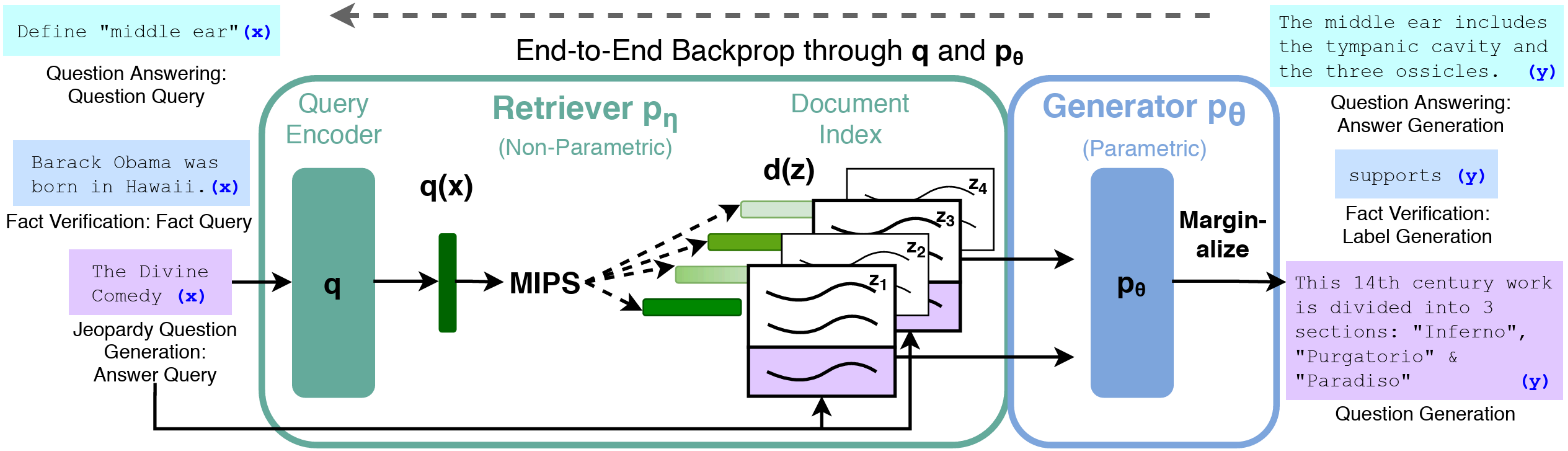
- Evidentiality Generator (Asai et al., 2023)
- REPLUG (Shi et al., 2023)



Joint training

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

End-to-end training for RAG



$$\sum_j -\log p_{RAG}(y_j \mid x_j)$$

Minimize NLL as in normal generation training

Retriever score Generator score

$$p_{\text{RAG}} \approx \prod_i \sum_{z \in \text{top-k}(p(\cdot \mid x))} p_\eta(z \mid x) p_\theta(y_i \mid x, z, y_{1:i-1})$$

Update retriever encoder and generator

Lewis et al. 2020. Retrieval-Augmented Generation for Knowledge-intensive NLP Tasks.

Training RAG

Independent training

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

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

- REPLUG (Shi et al., 2023)

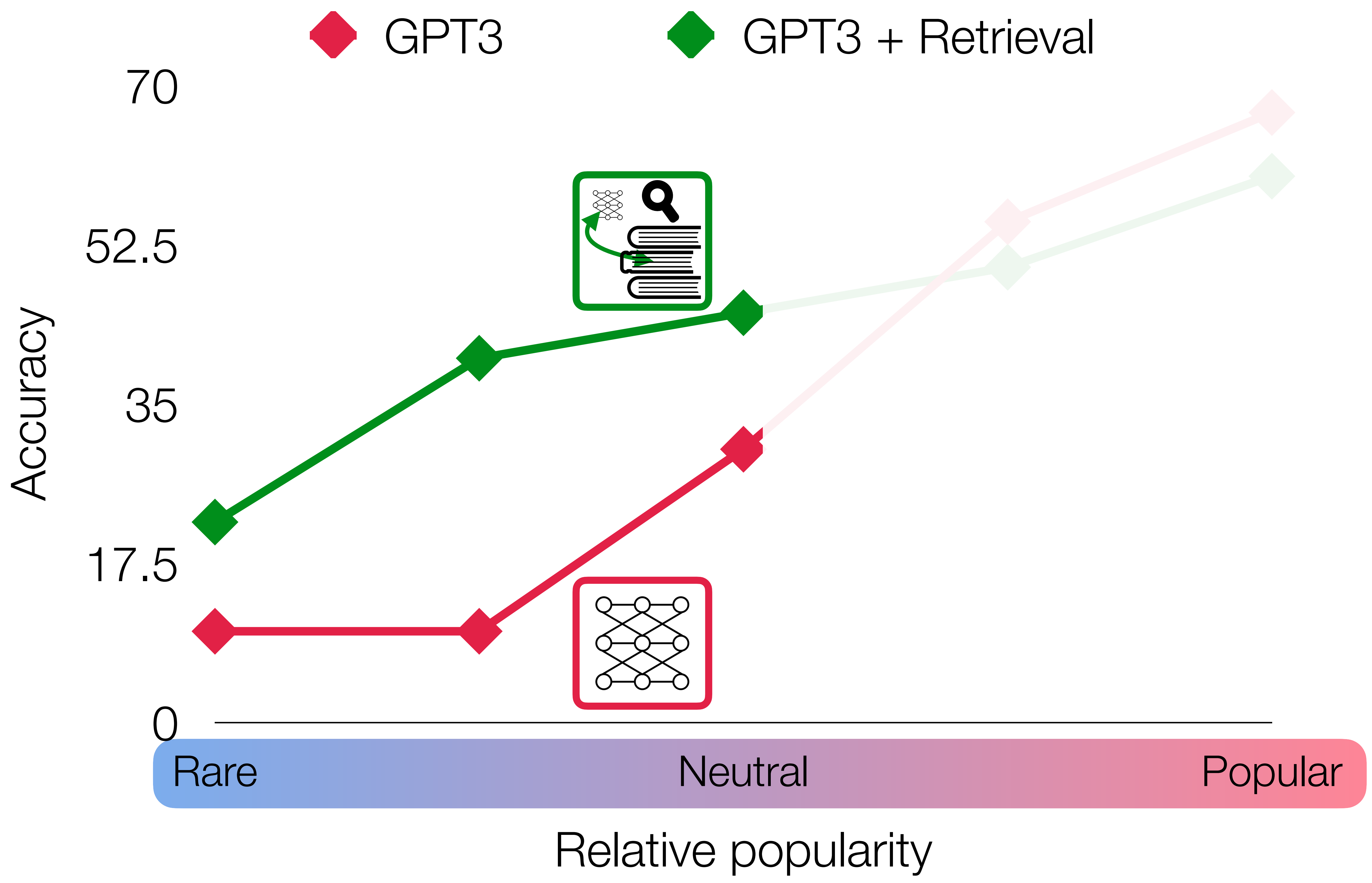
Widely referred to as ***RAG***

Joint training (async update)

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

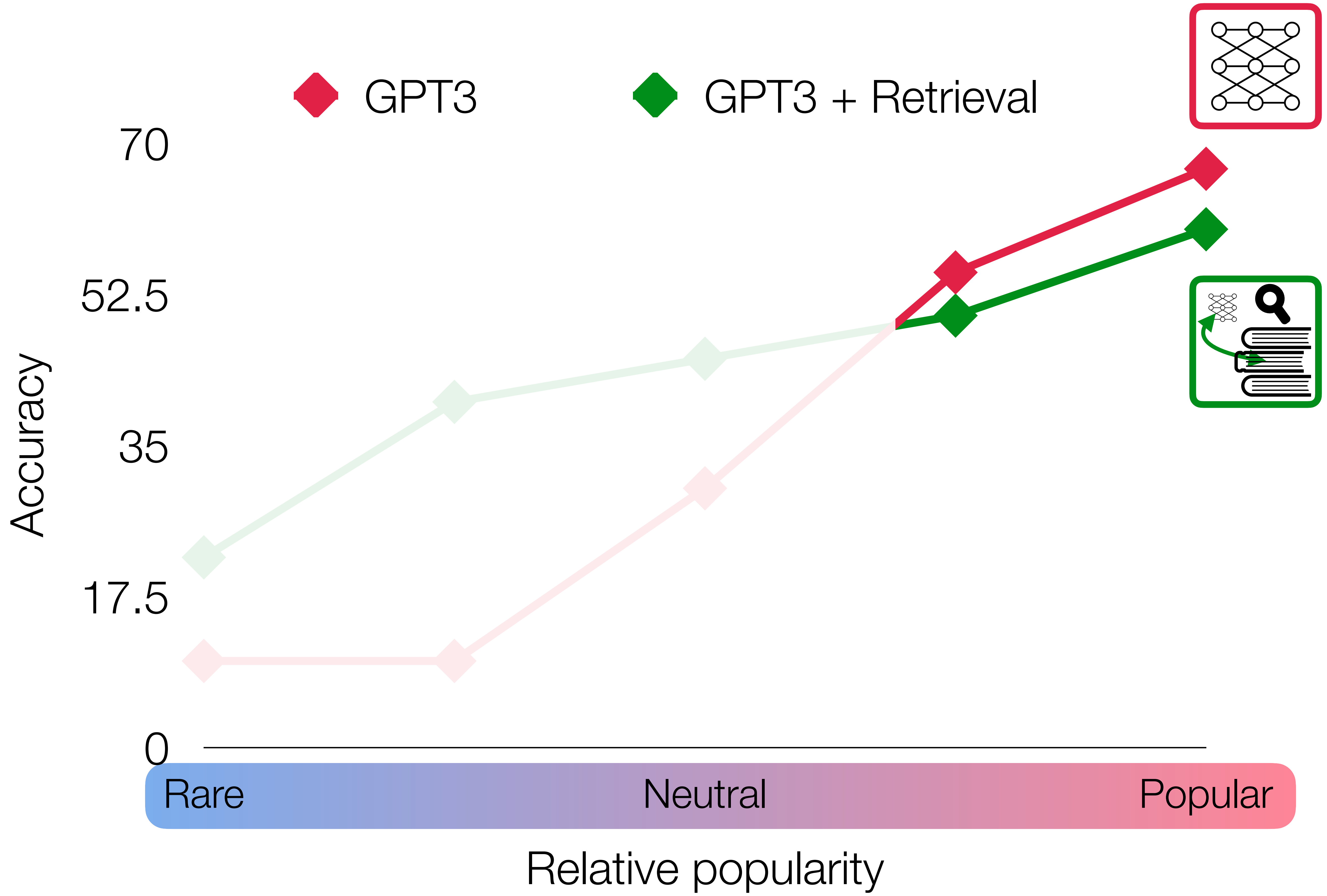
Effectiveness of Simple RAG

RAG constantly gives performance improvements esp. in long-tail



Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

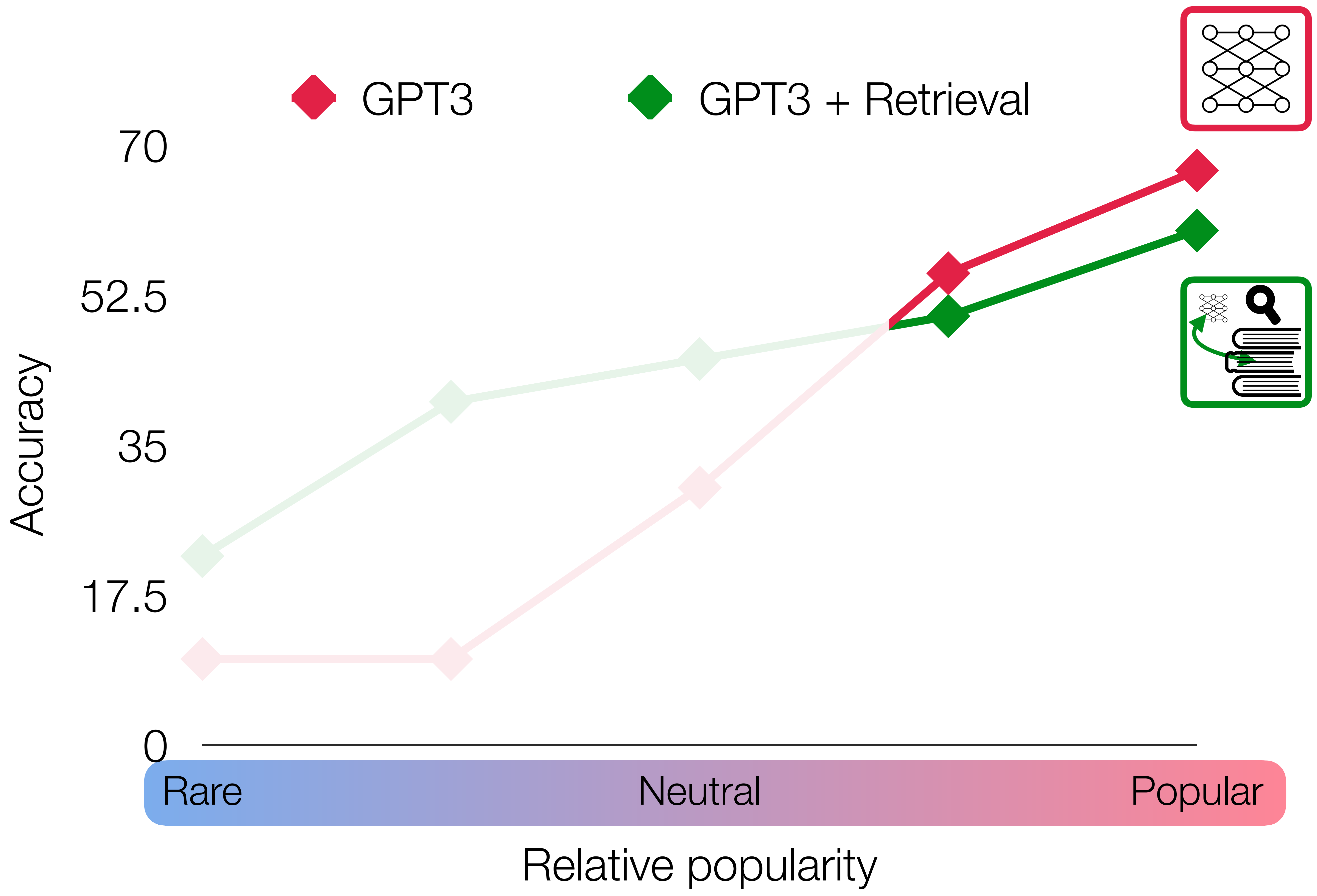
Limitations of Simple RAG

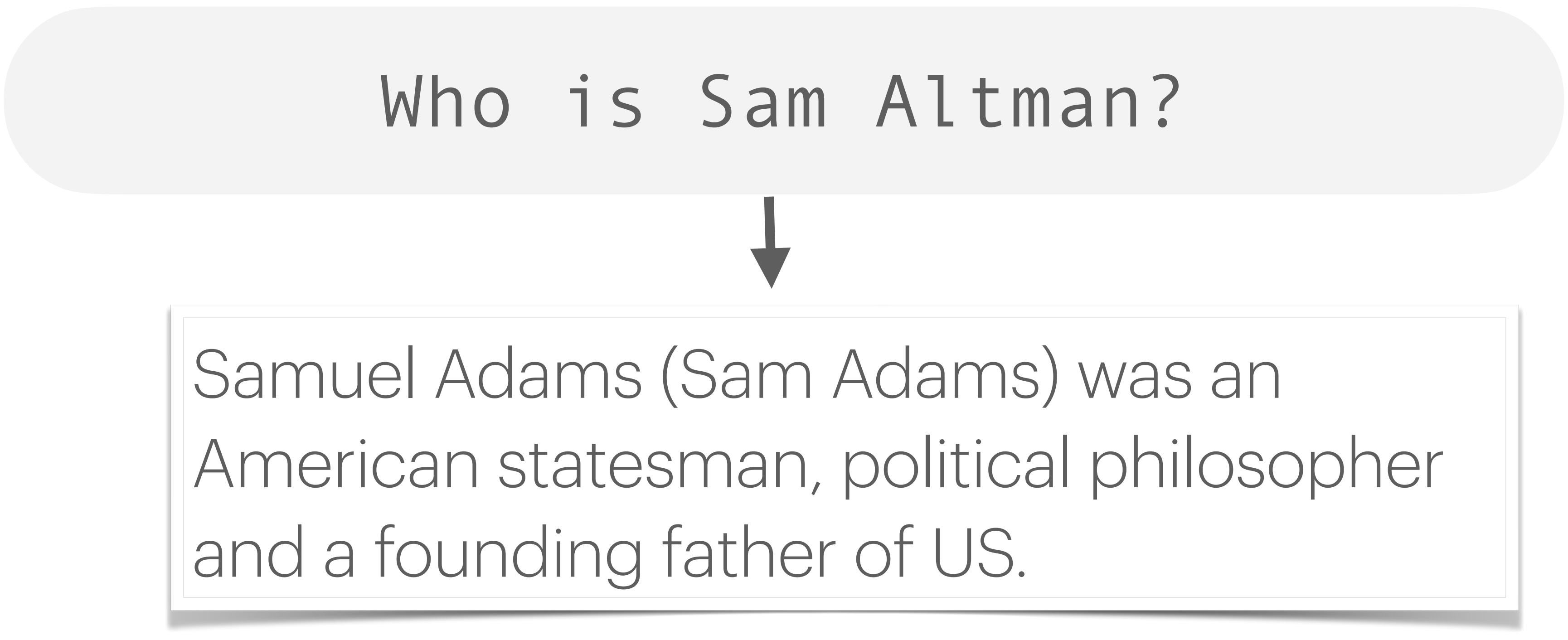


Who is Sam Altman?

Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

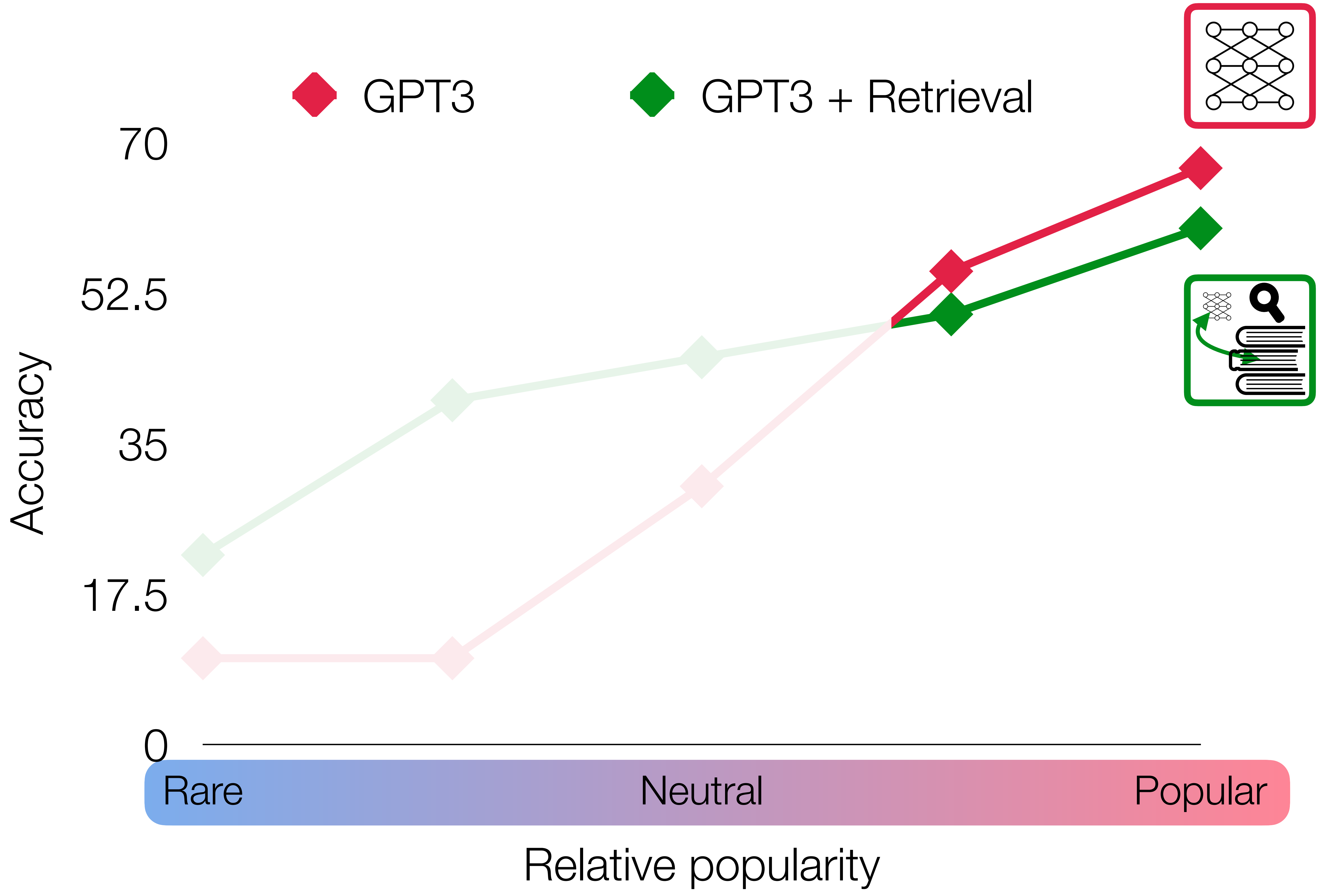
Limitations of Simple RAG

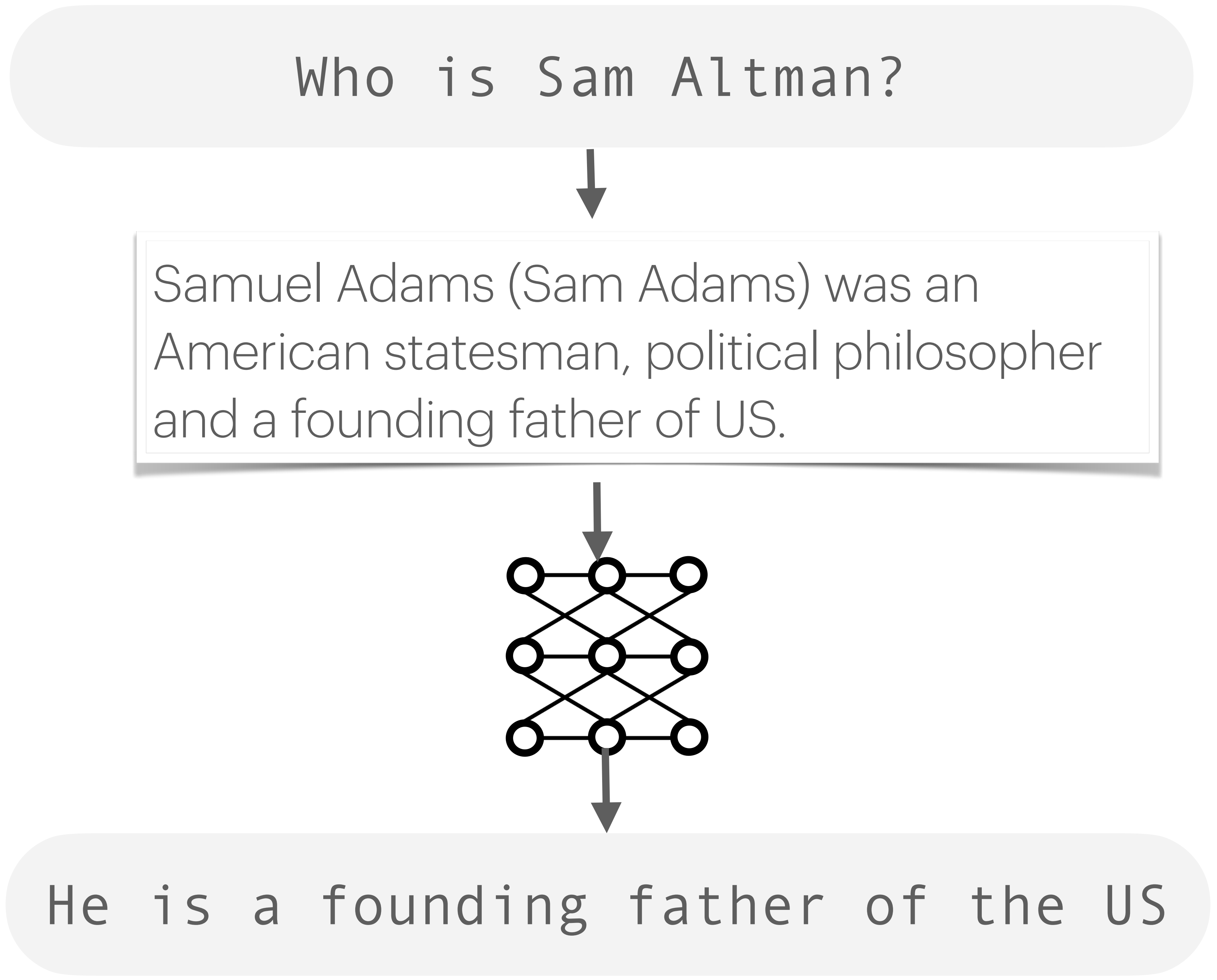




Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

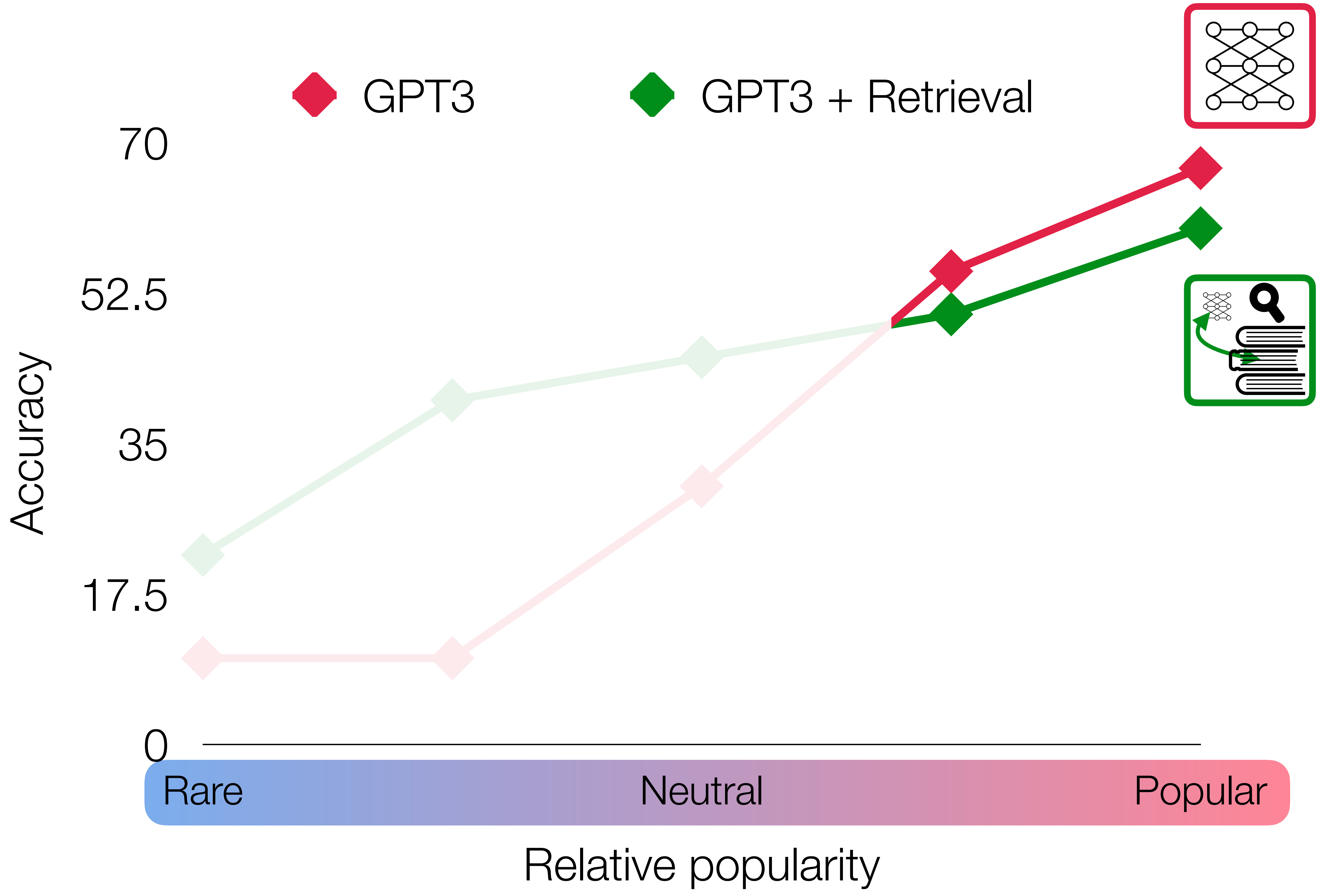
Limitations of Simple RAG

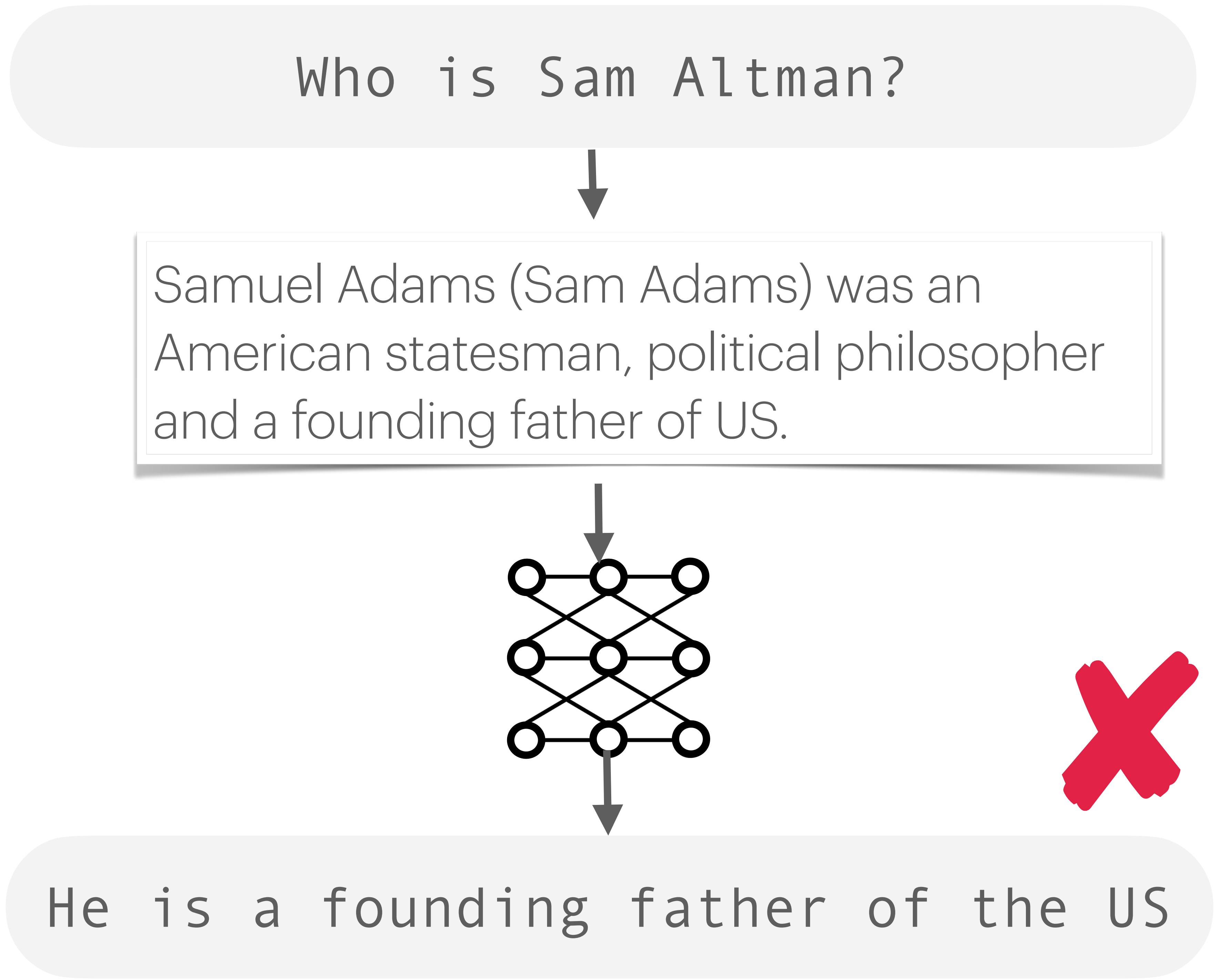




Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Limitations of Simple RAG

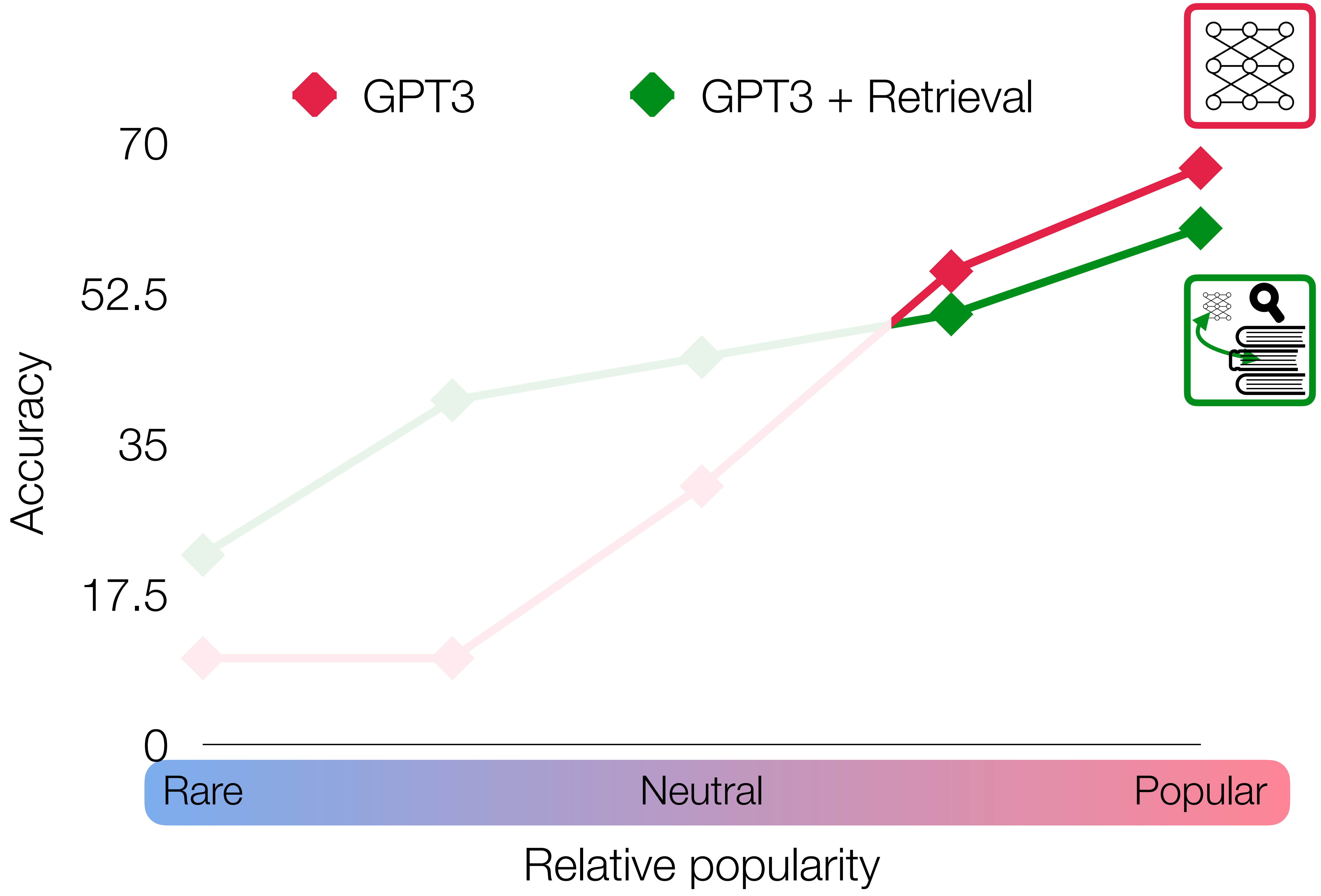


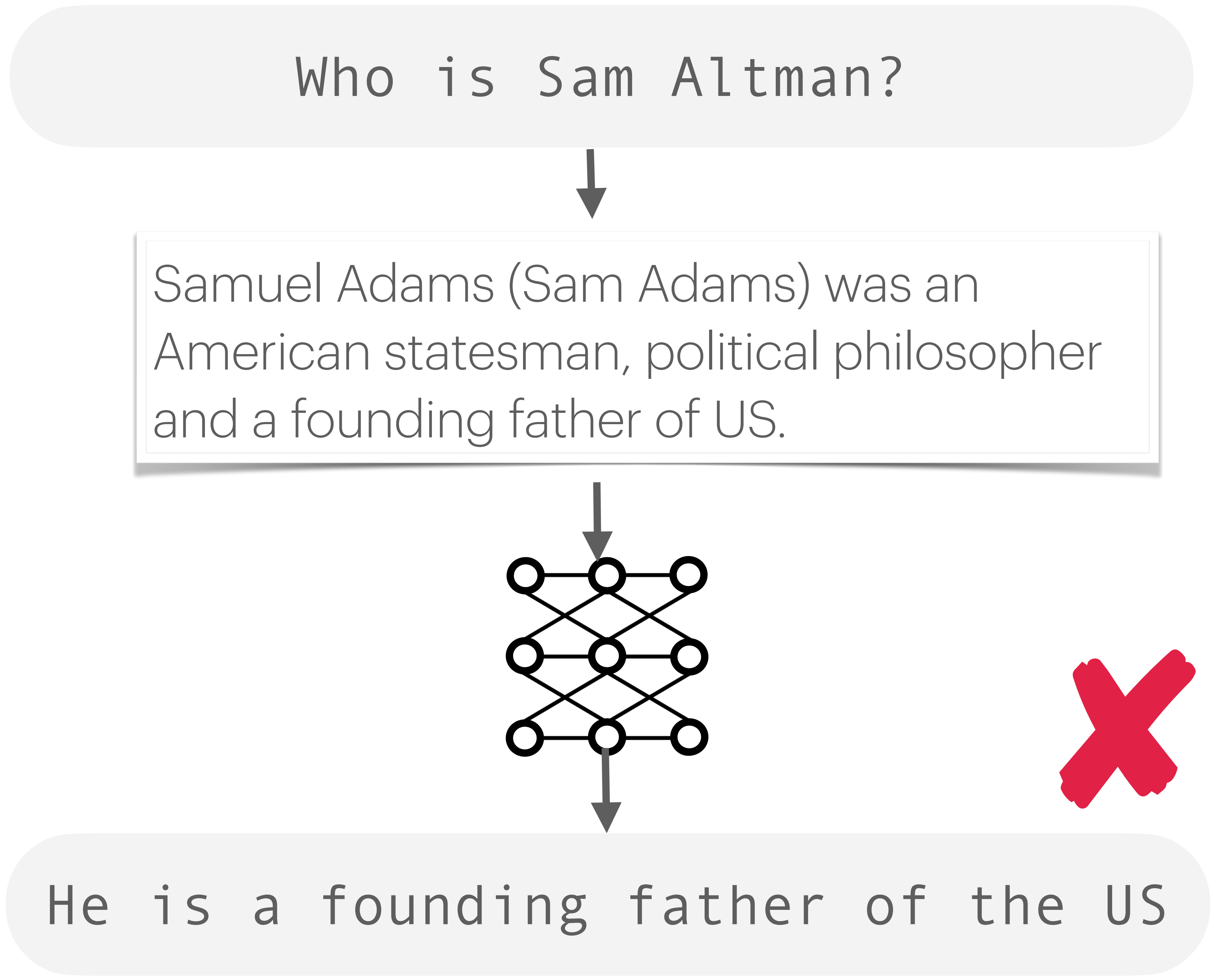


Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Limitations of Simple RAG

Unnecessary retrieval



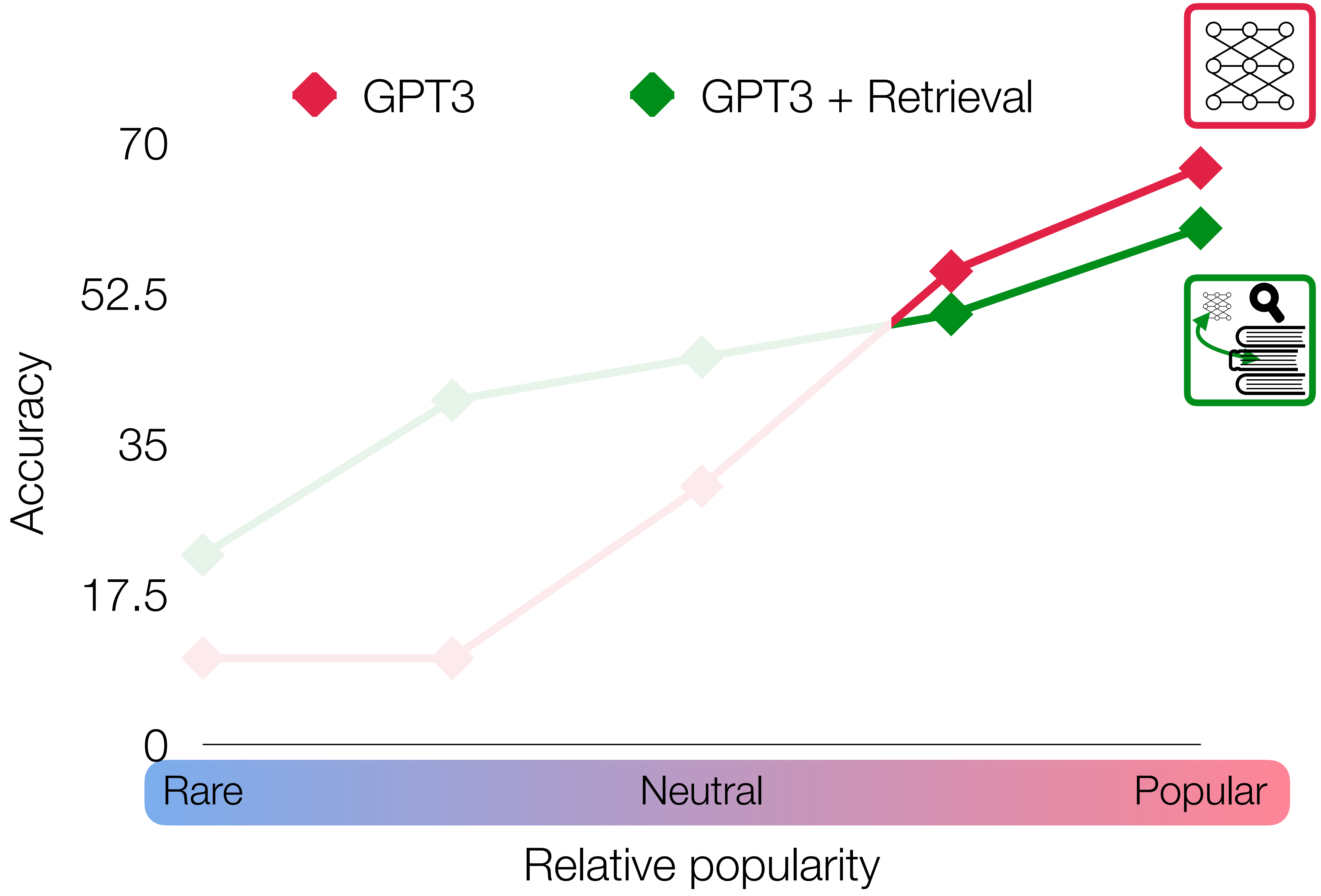


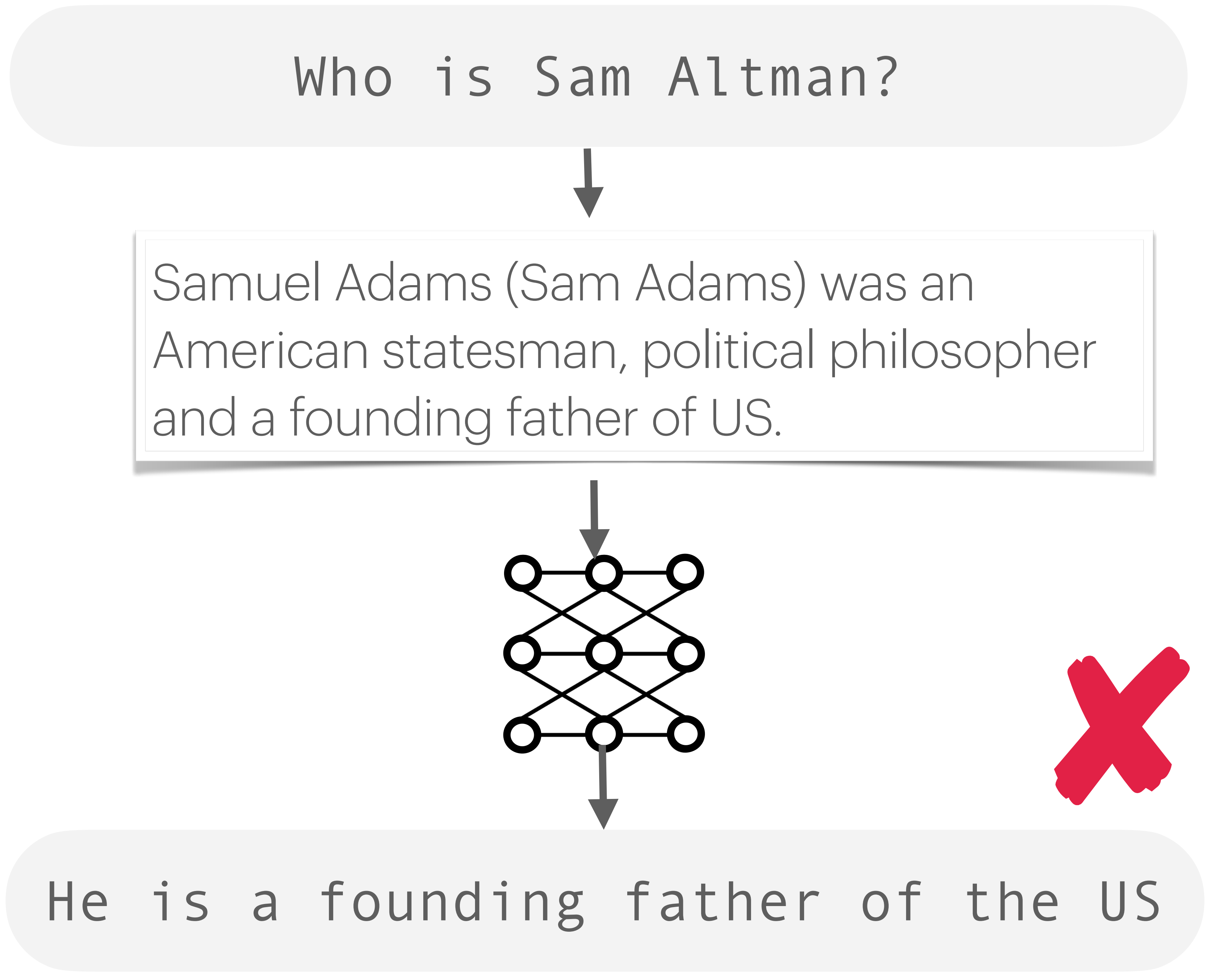
Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

Limitations of Simple RAG

Unnecessary retrieval

LMs are easily distracted by unhelpful context





Mallen*, Asai* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories

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 (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. ...

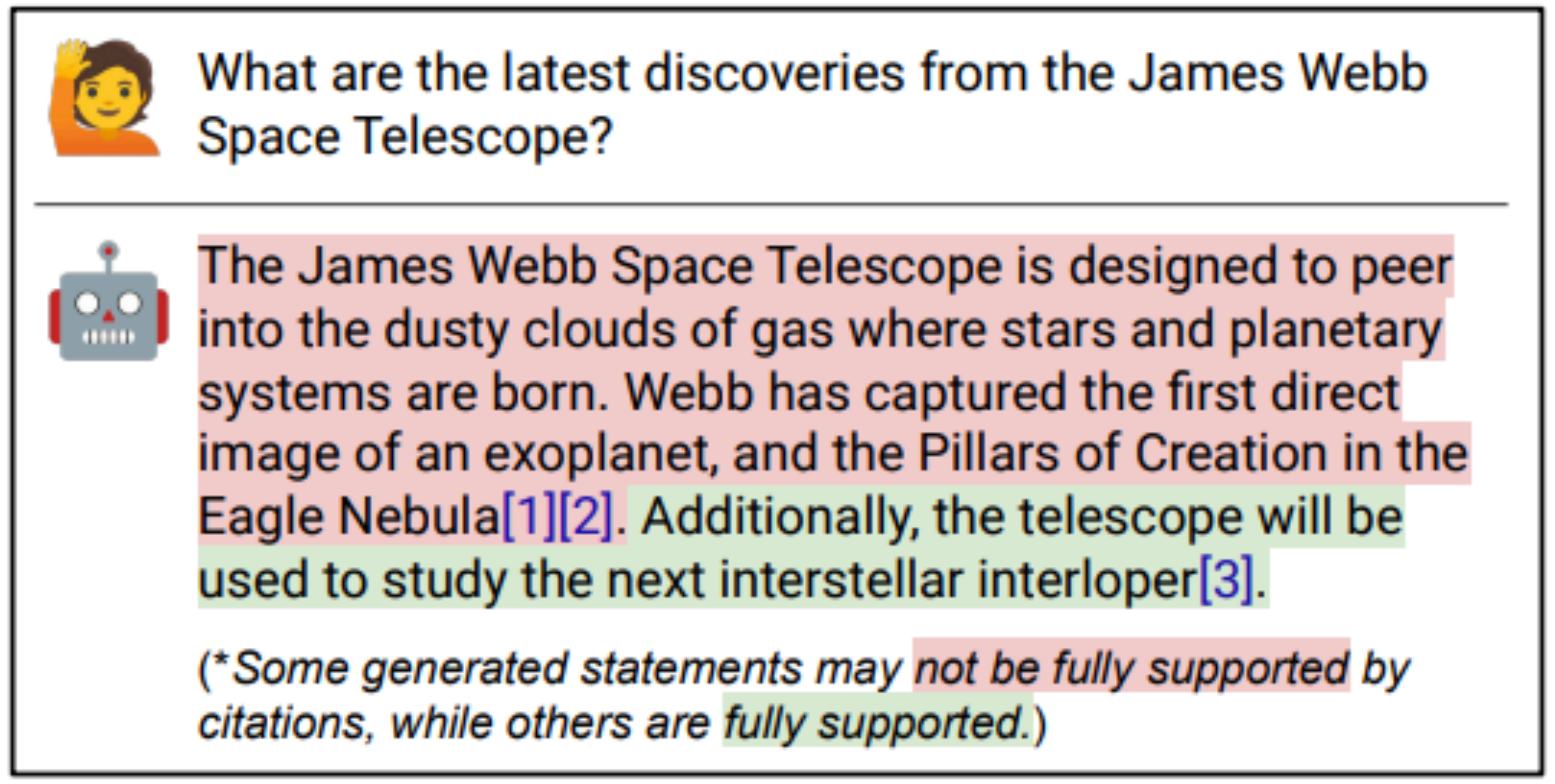
[2]: cnn.com (citation partially supports its associated statement)
Pillars of Creation: James Webb Space Telescope ...
... The Pillars of Creation, in the Eagle Nebula, is a star-forming region captured in a new image (right) by the James Webb Space Telescope that reveals more detail than a 2014 image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

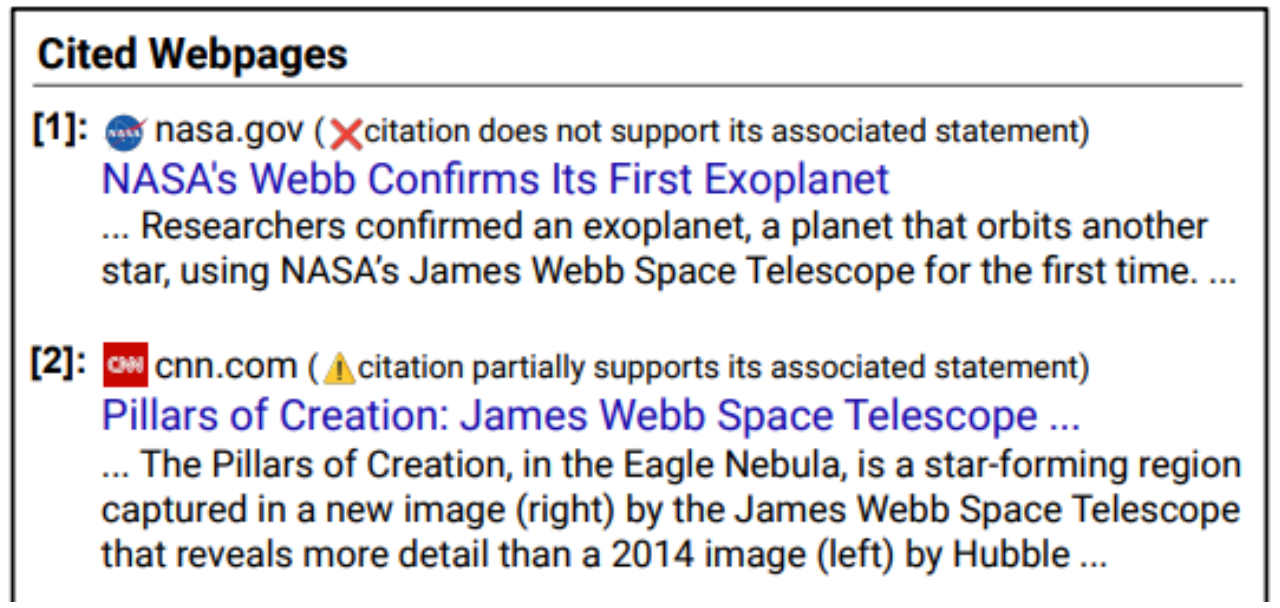
Outputs aren't often supported by citations

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





Outputs aren't often supported by citations

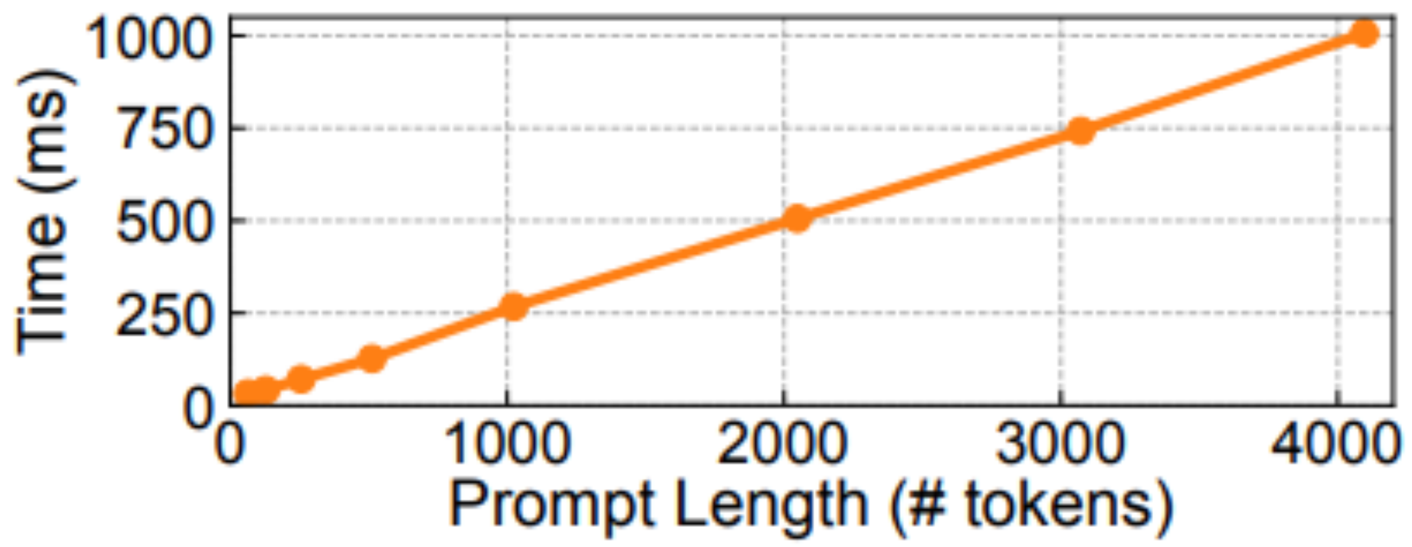
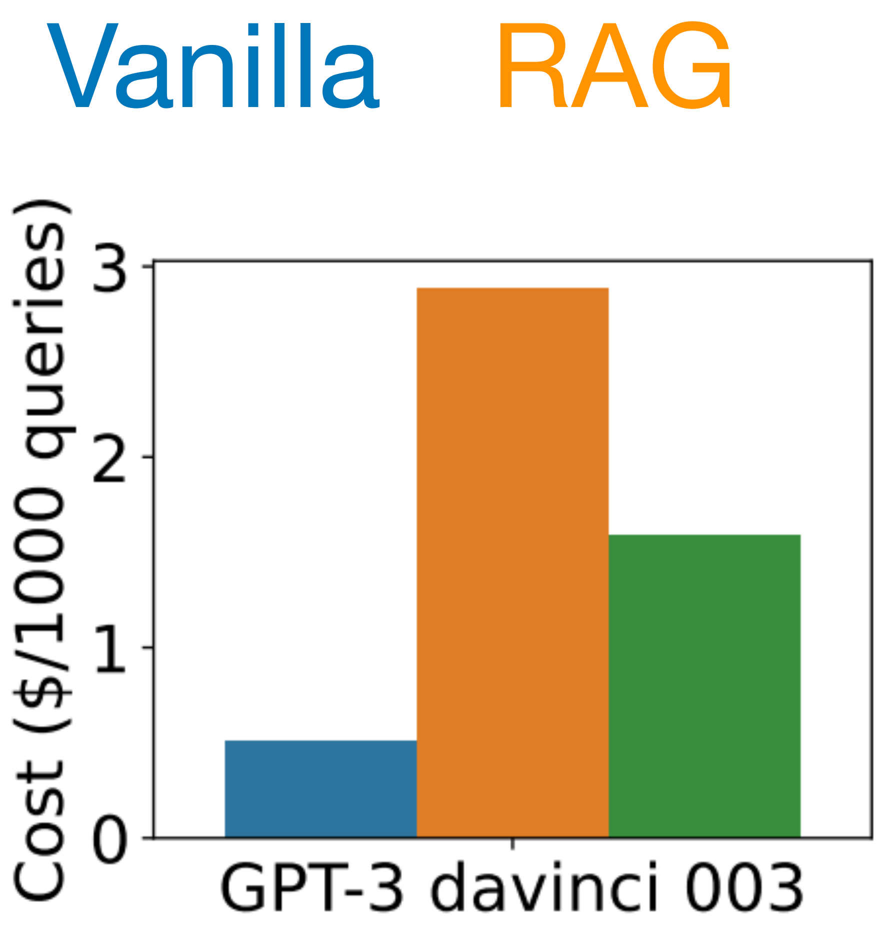


Figure 2. Inference time with different input lengths.



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

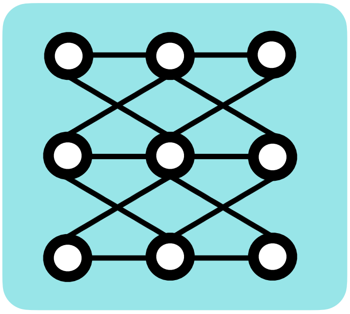
Self-RAG: Learn to Retrieve and Critique

☹ LMs aren’t trained with retrieval

☹ Fixed two-stage pipeline

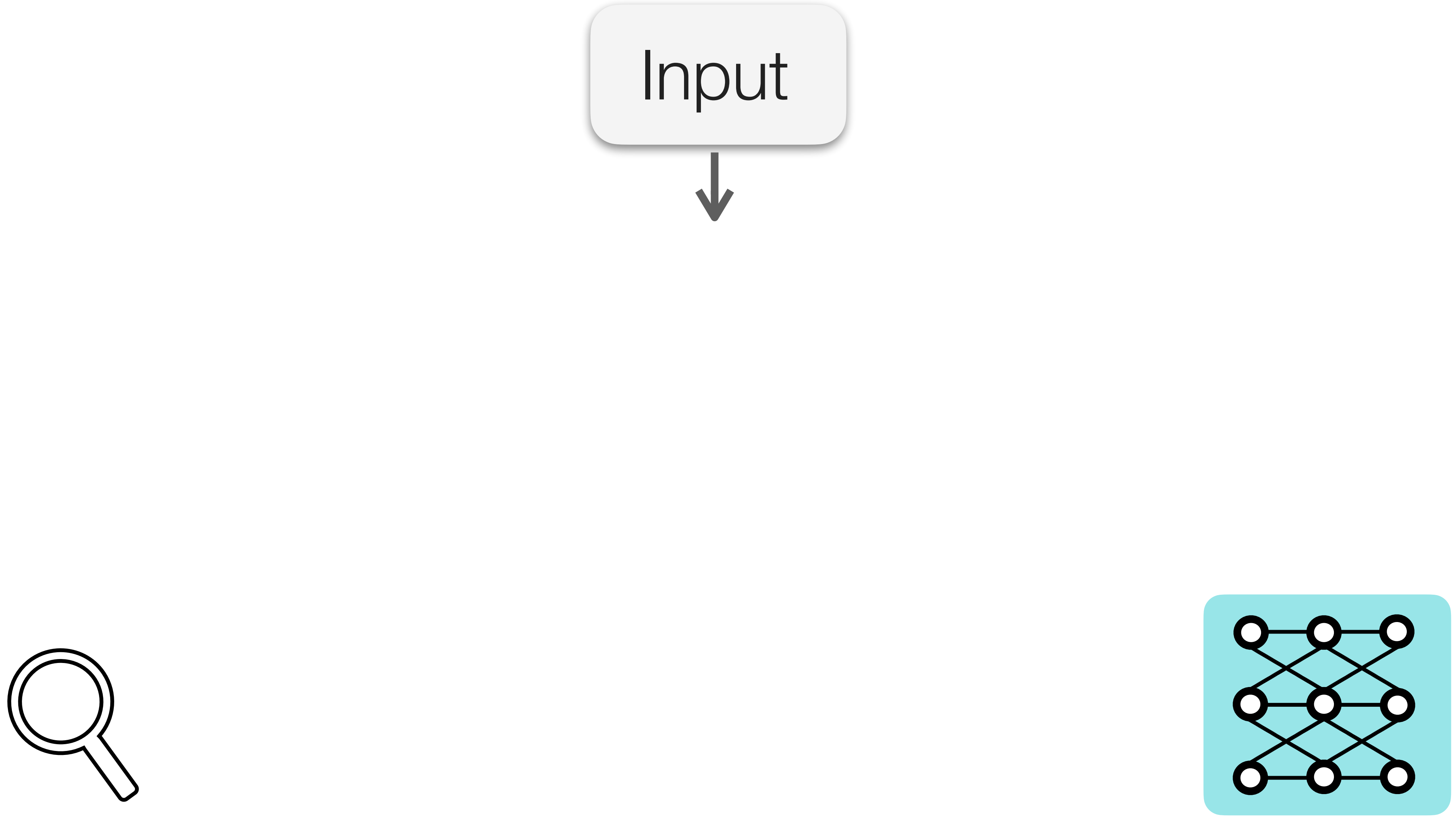
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



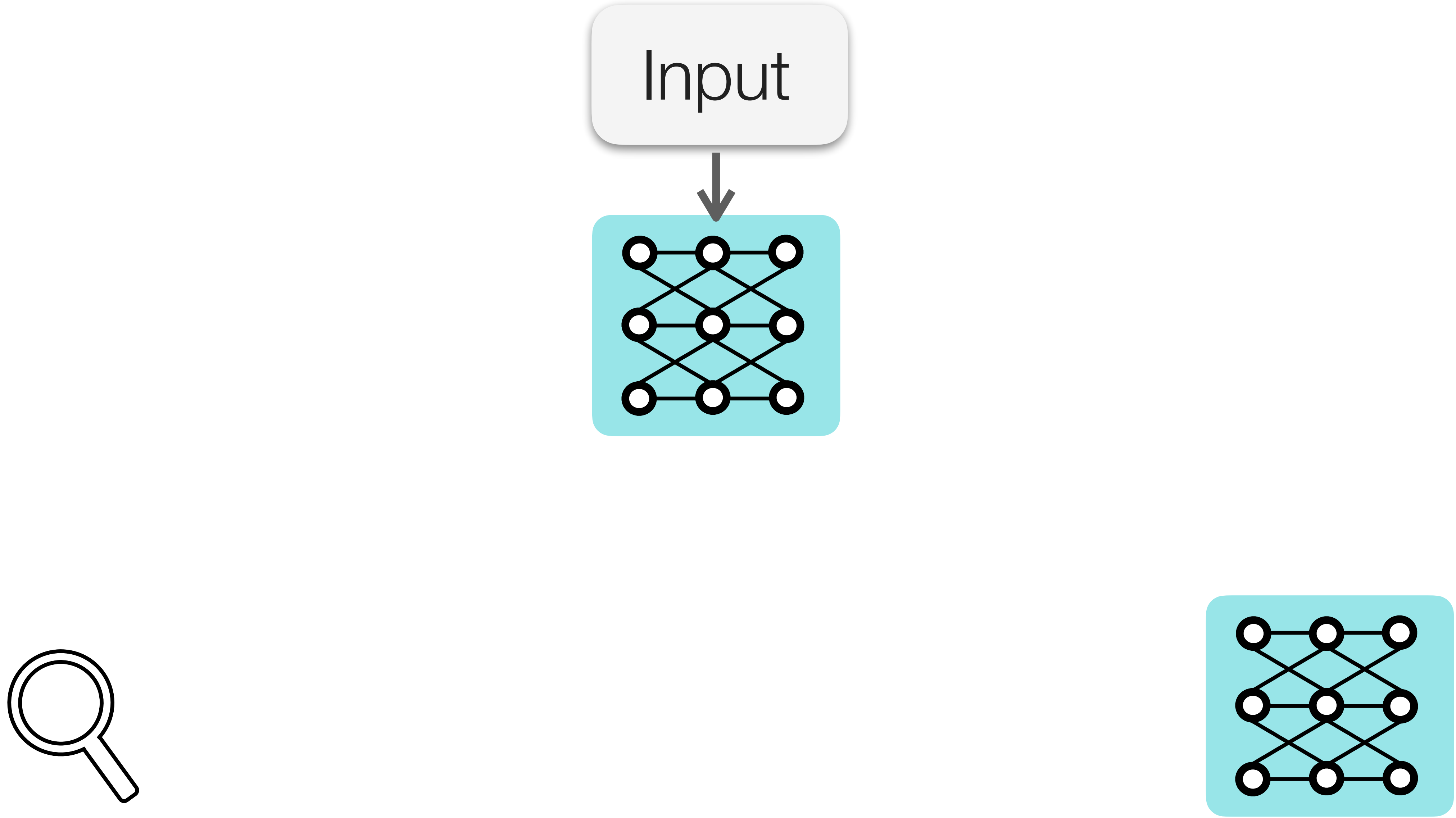
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



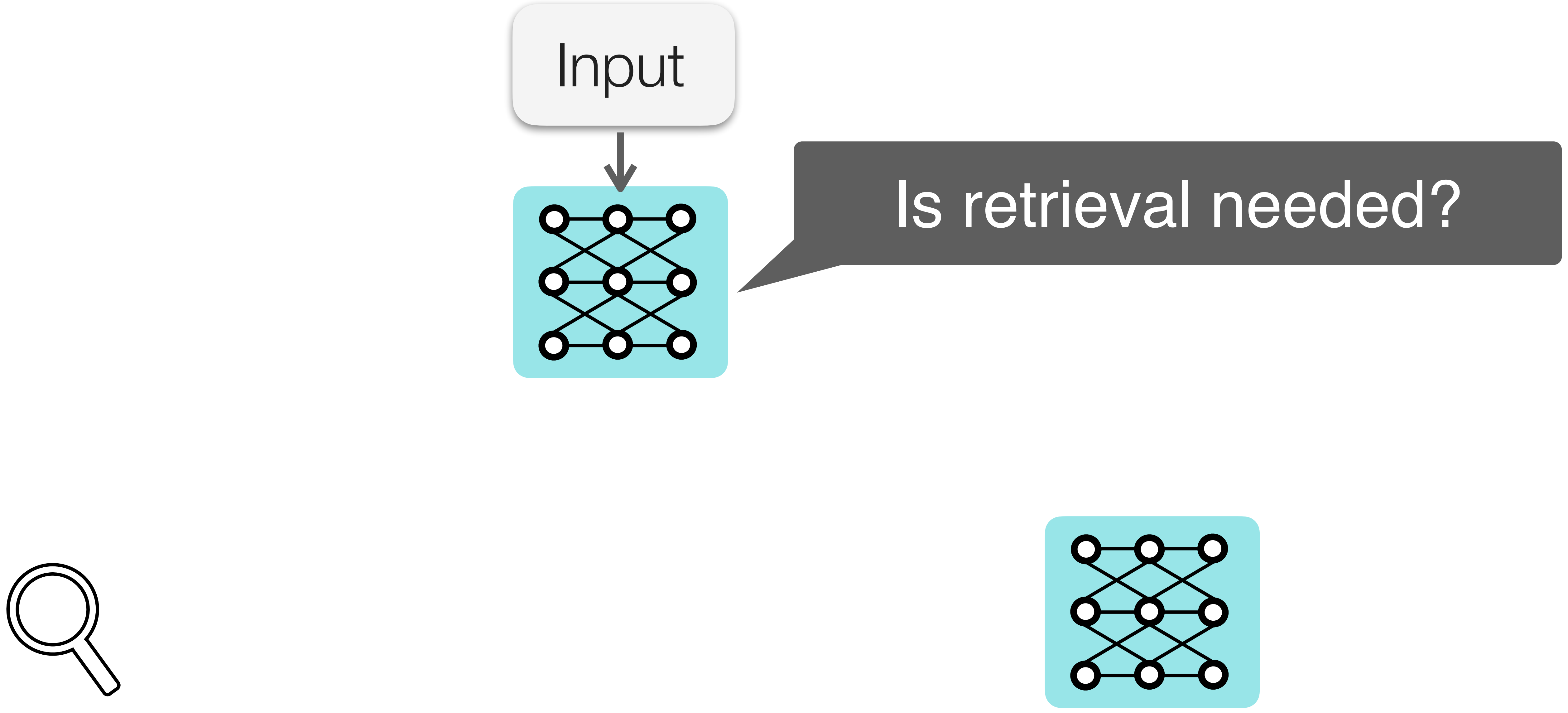
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



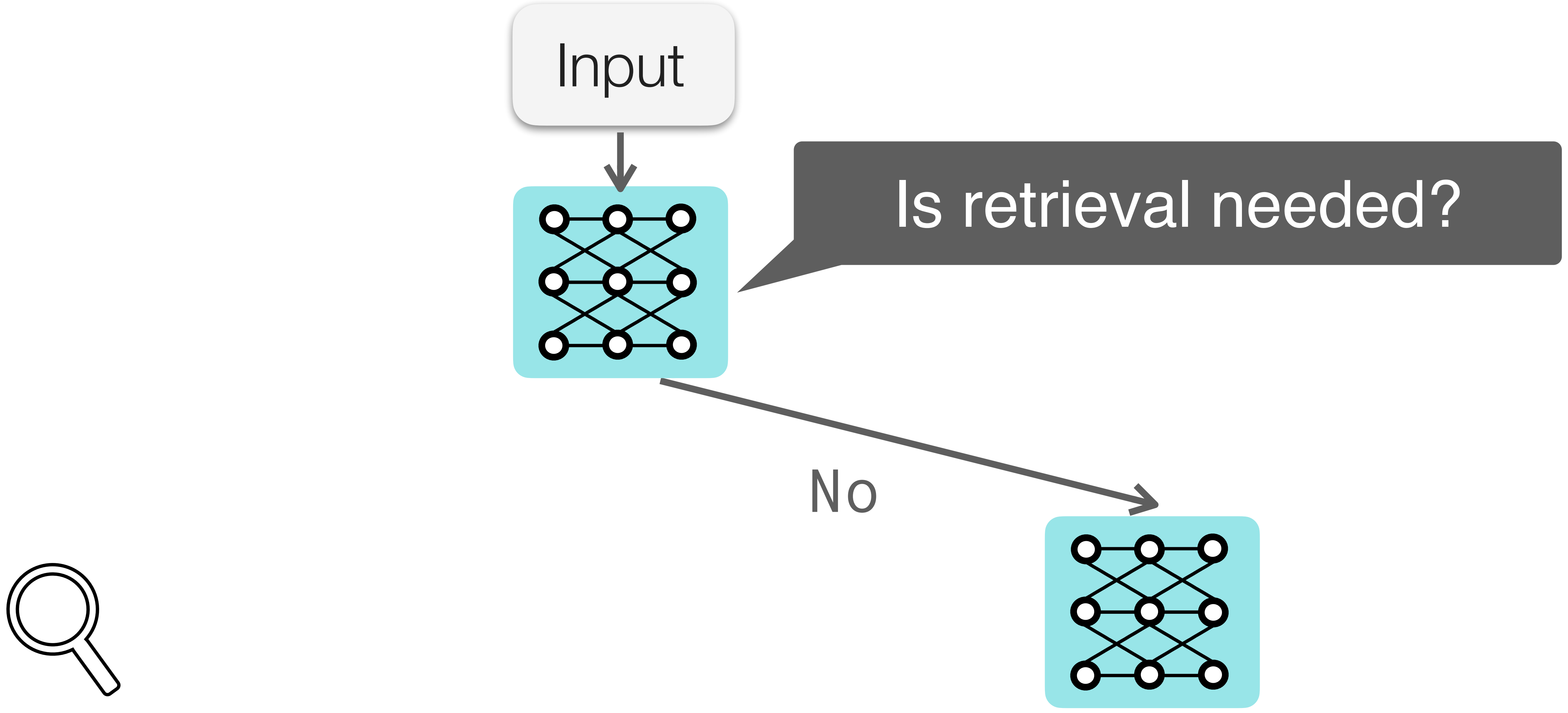
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



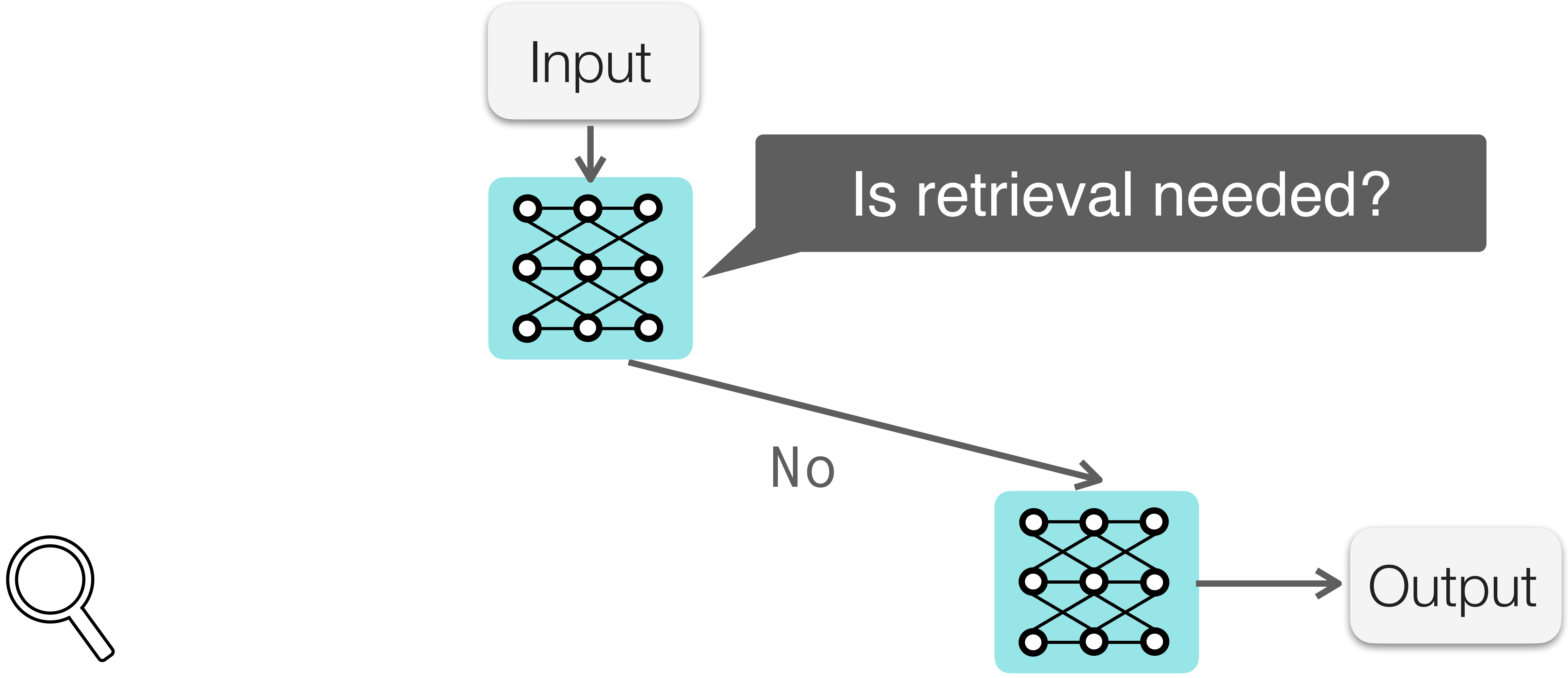
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



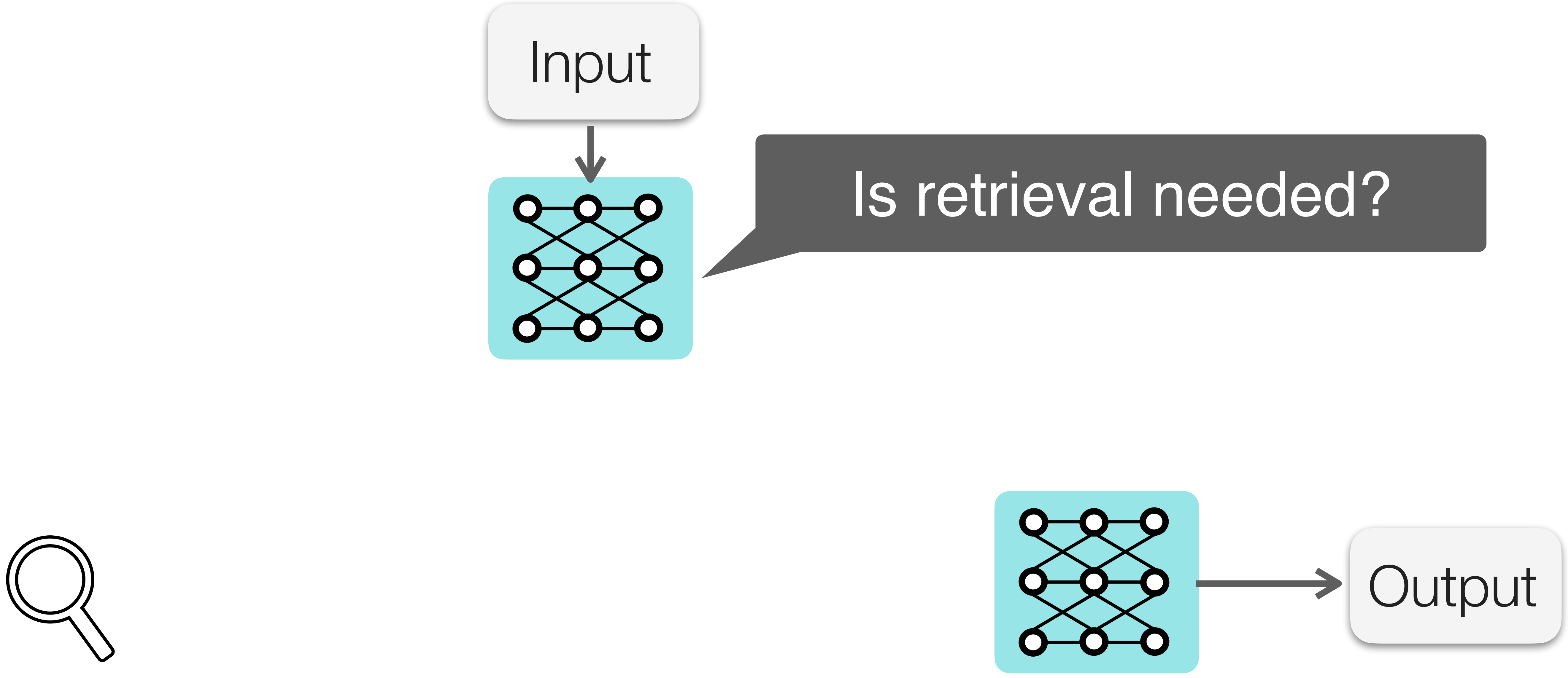
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



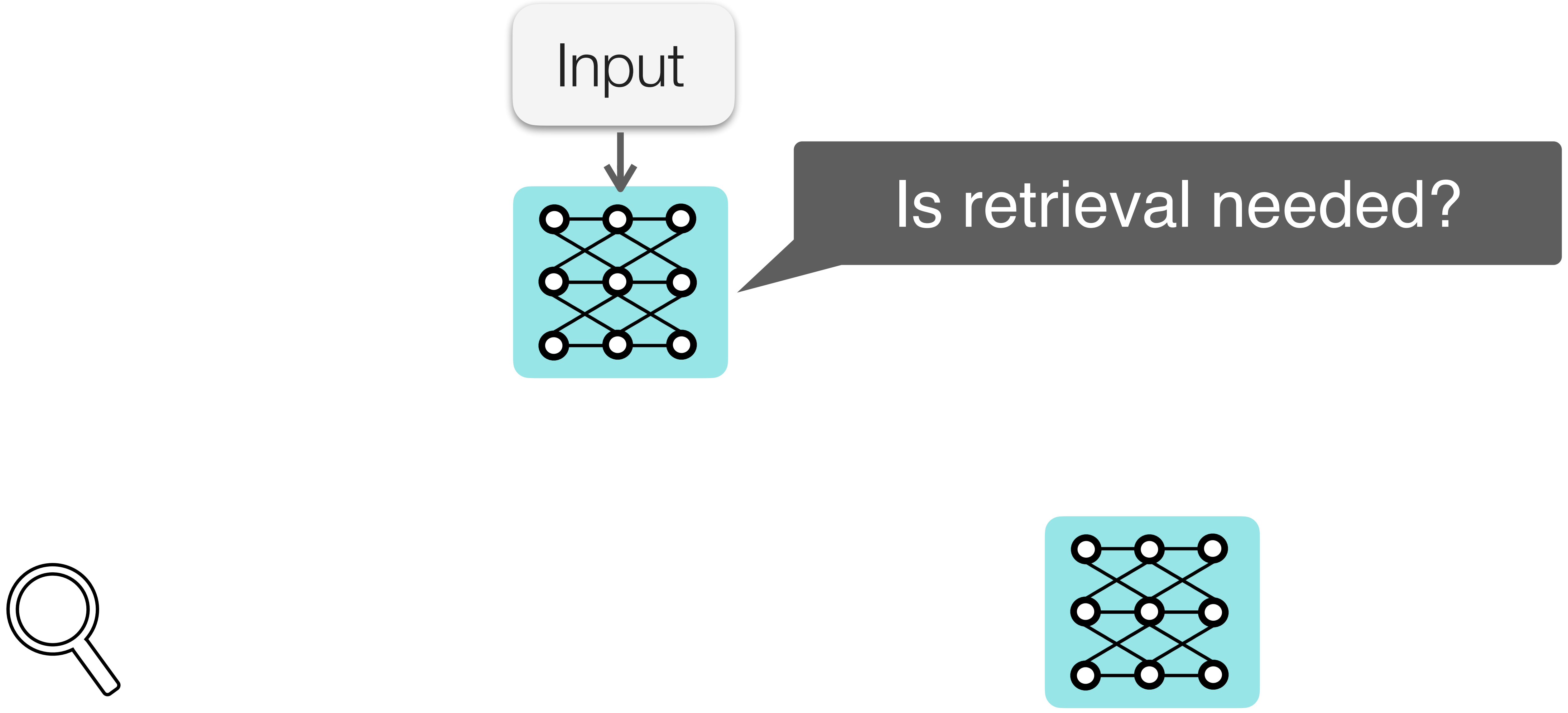
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



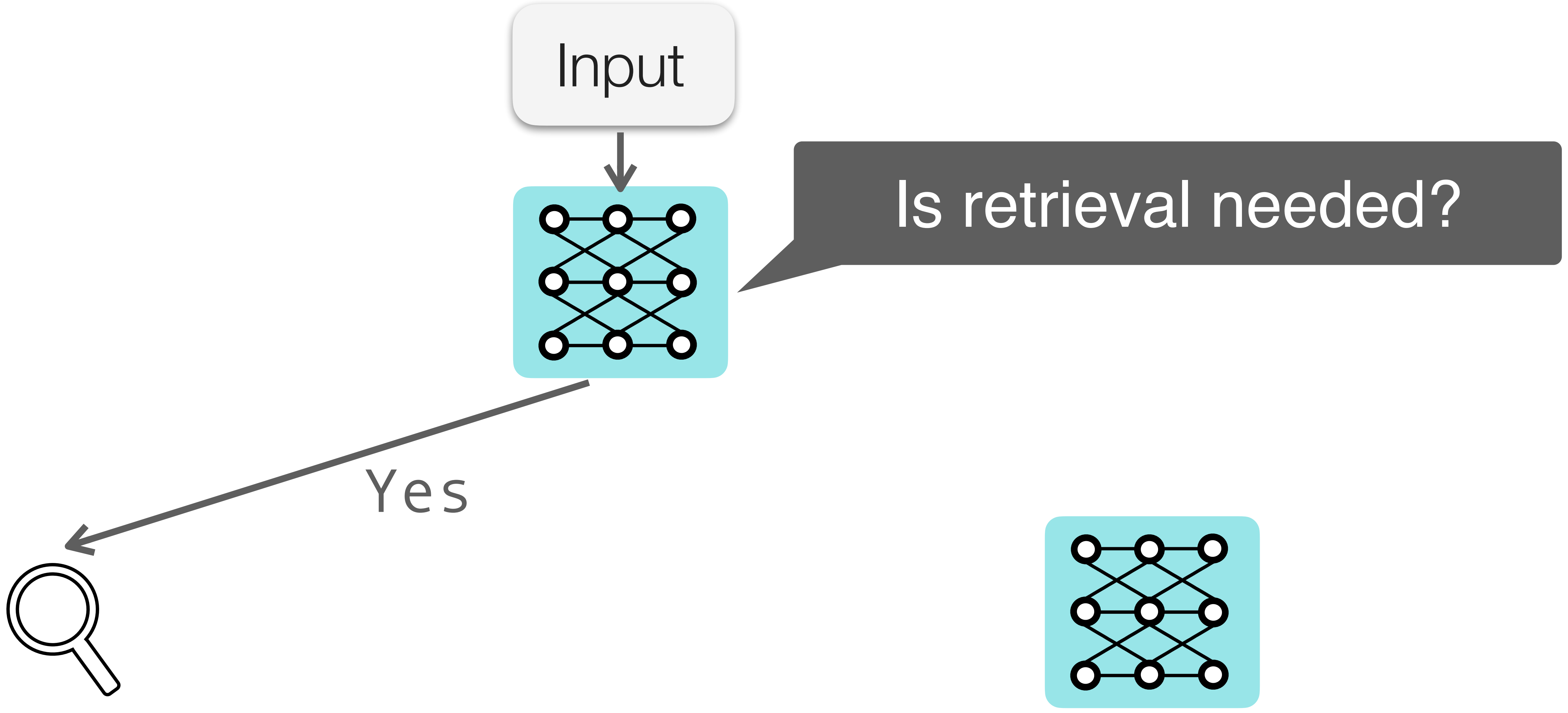
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



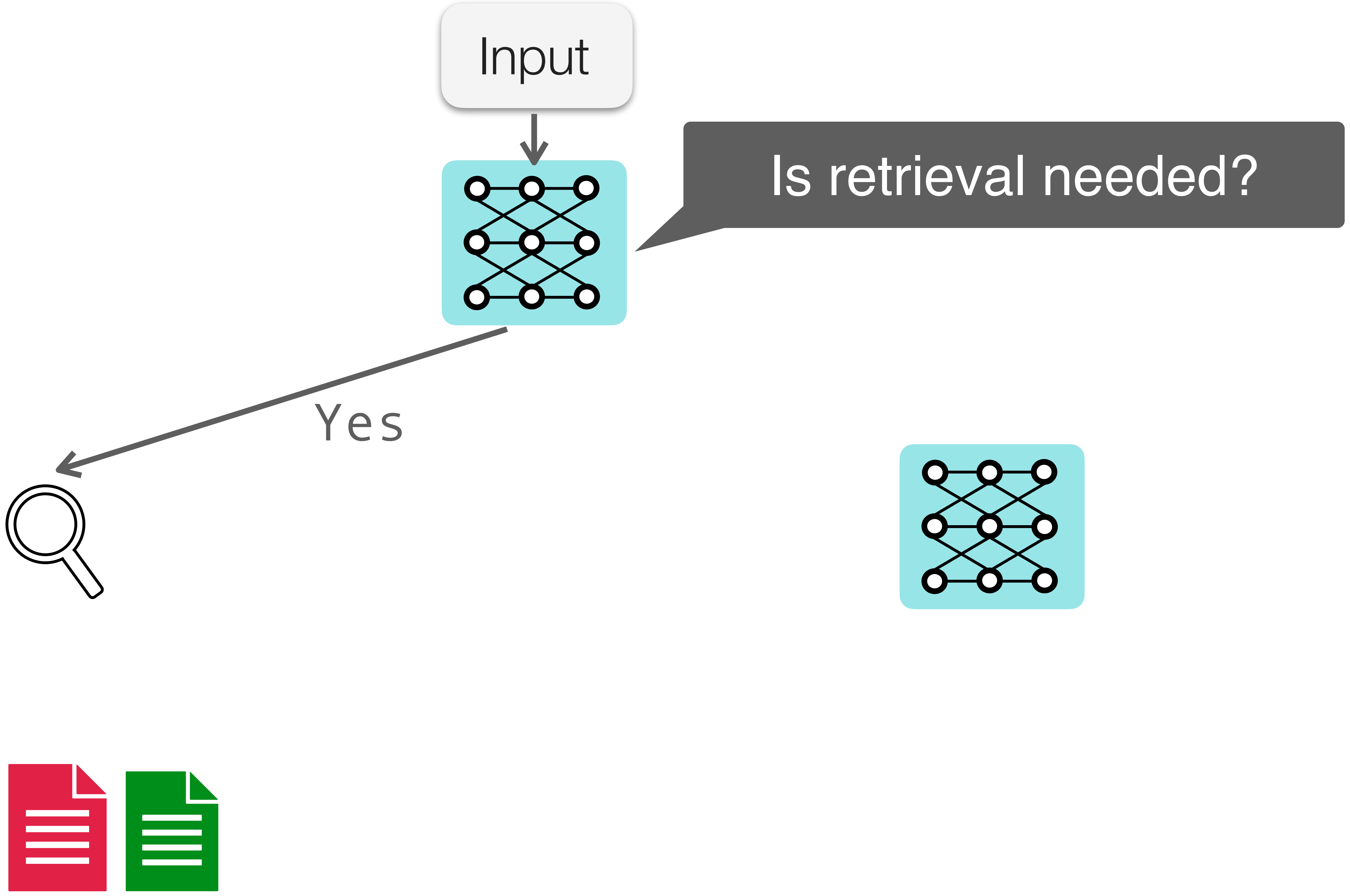
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



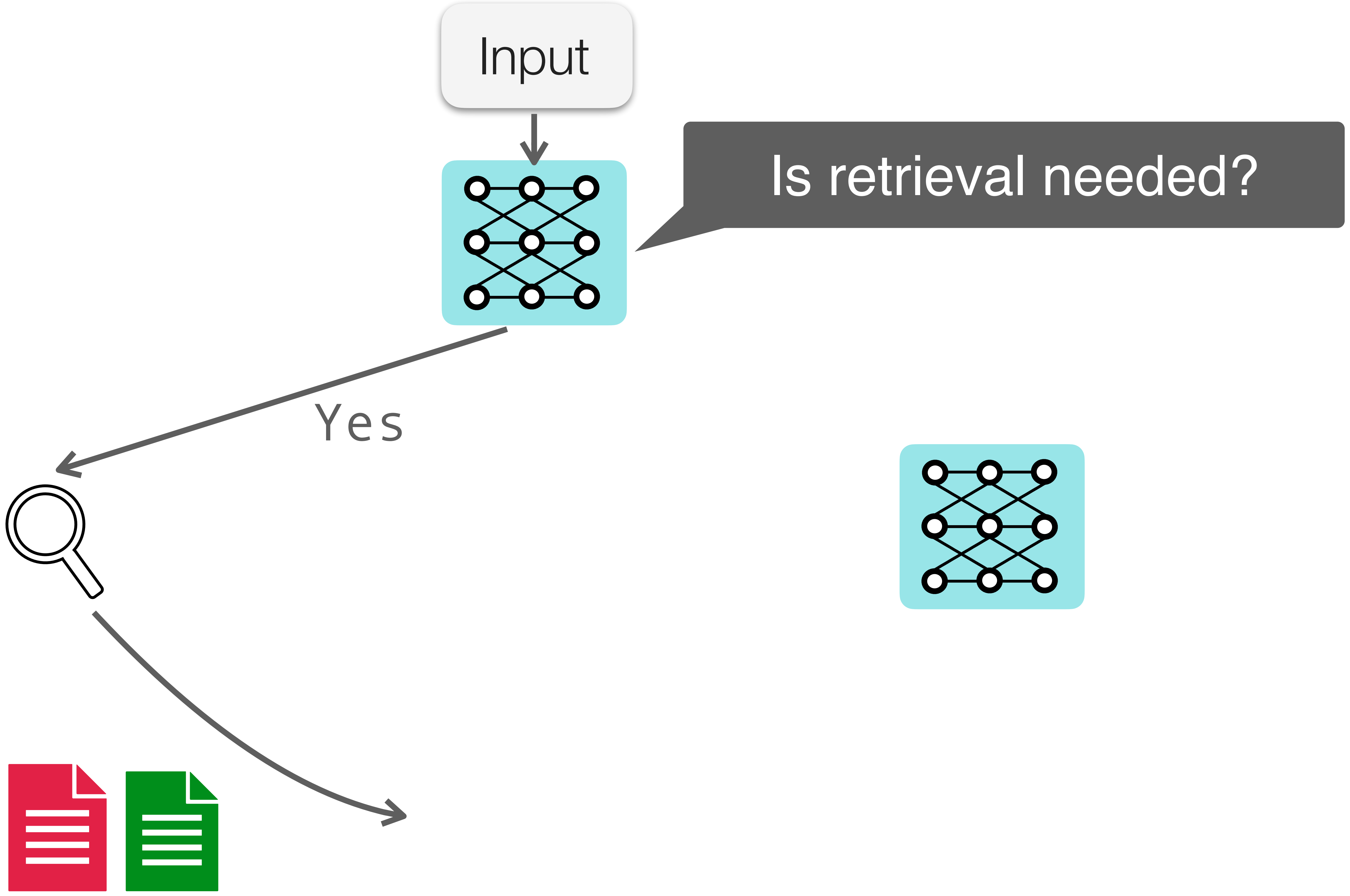
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



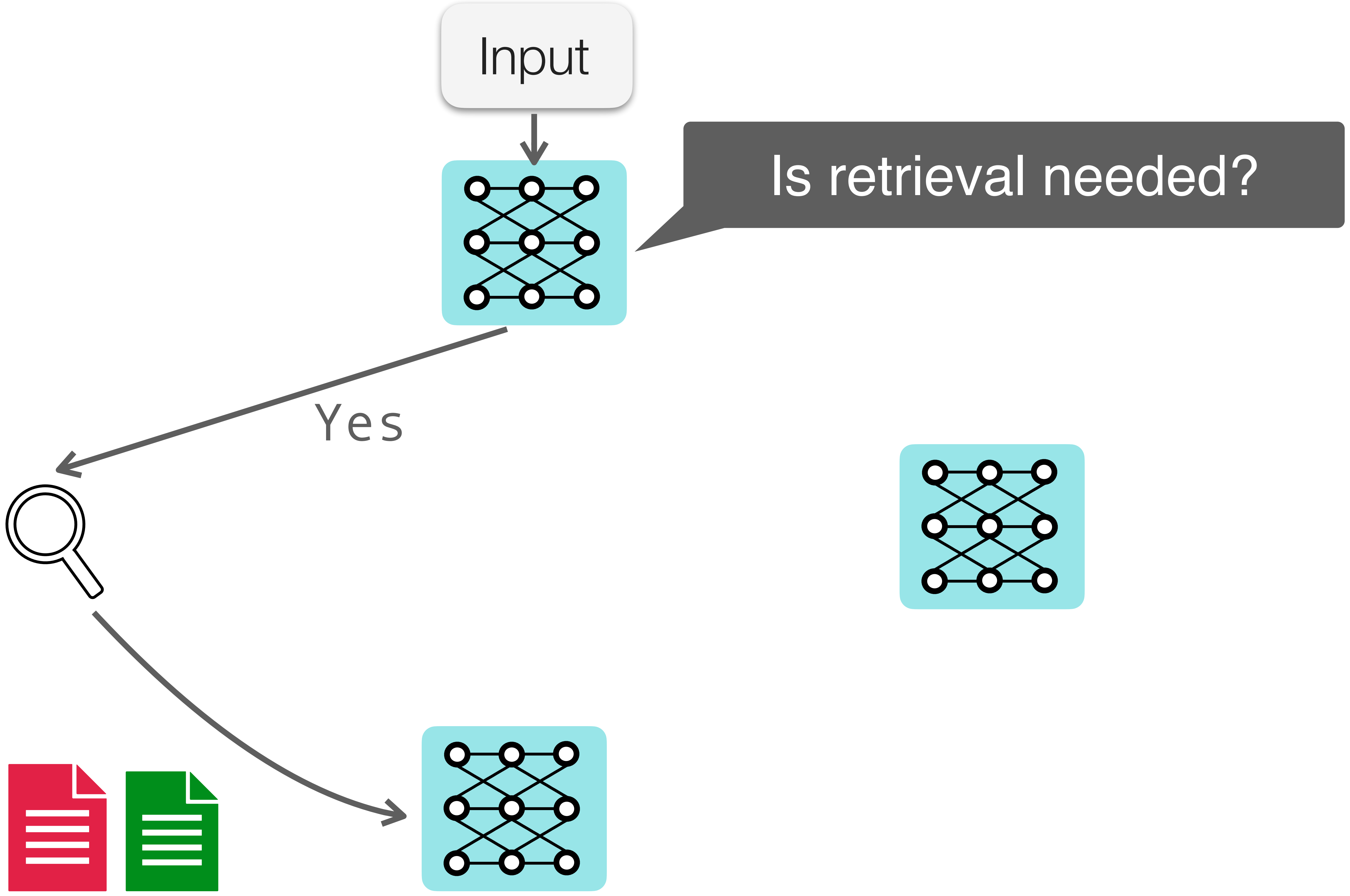
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



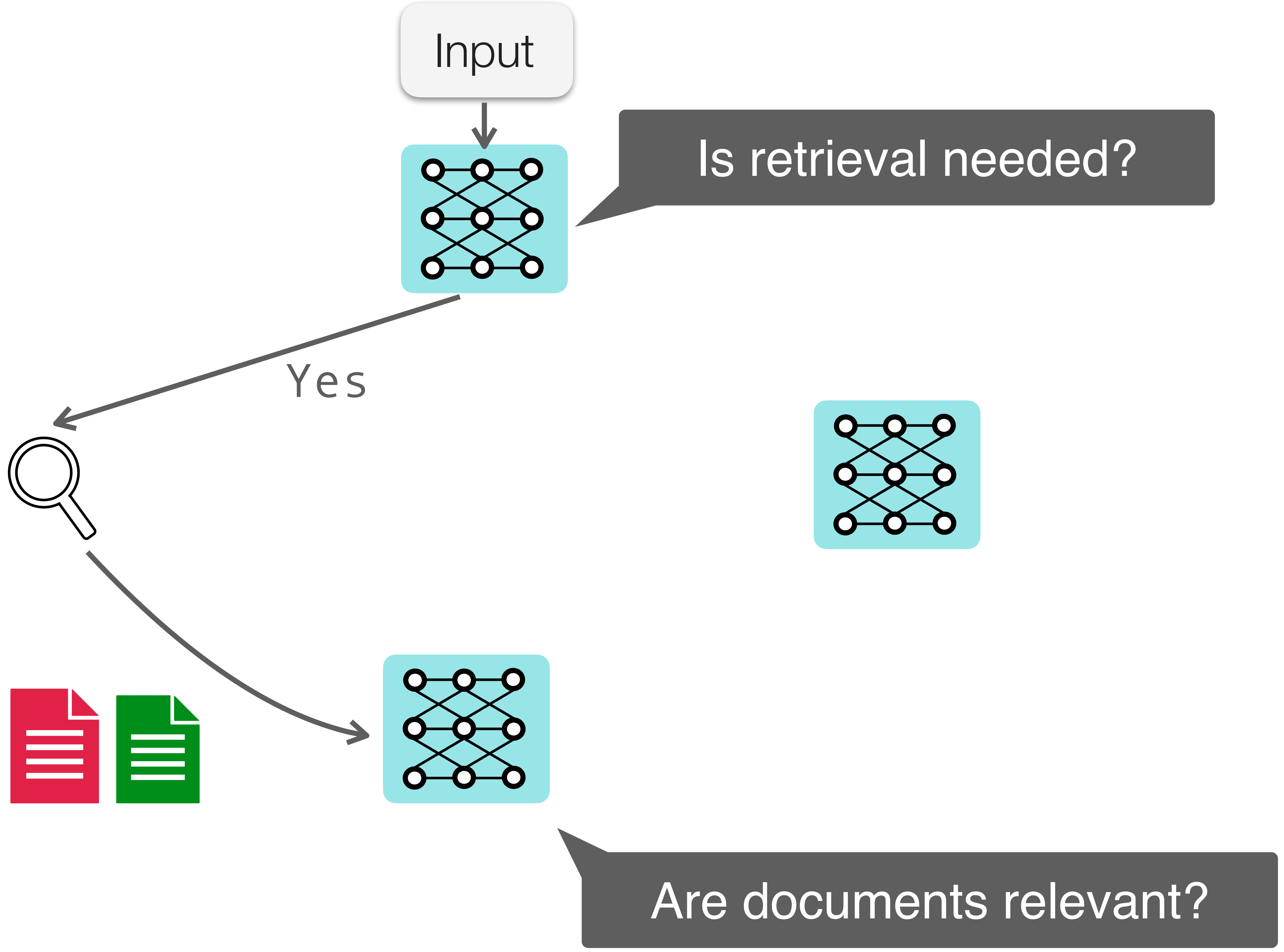
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



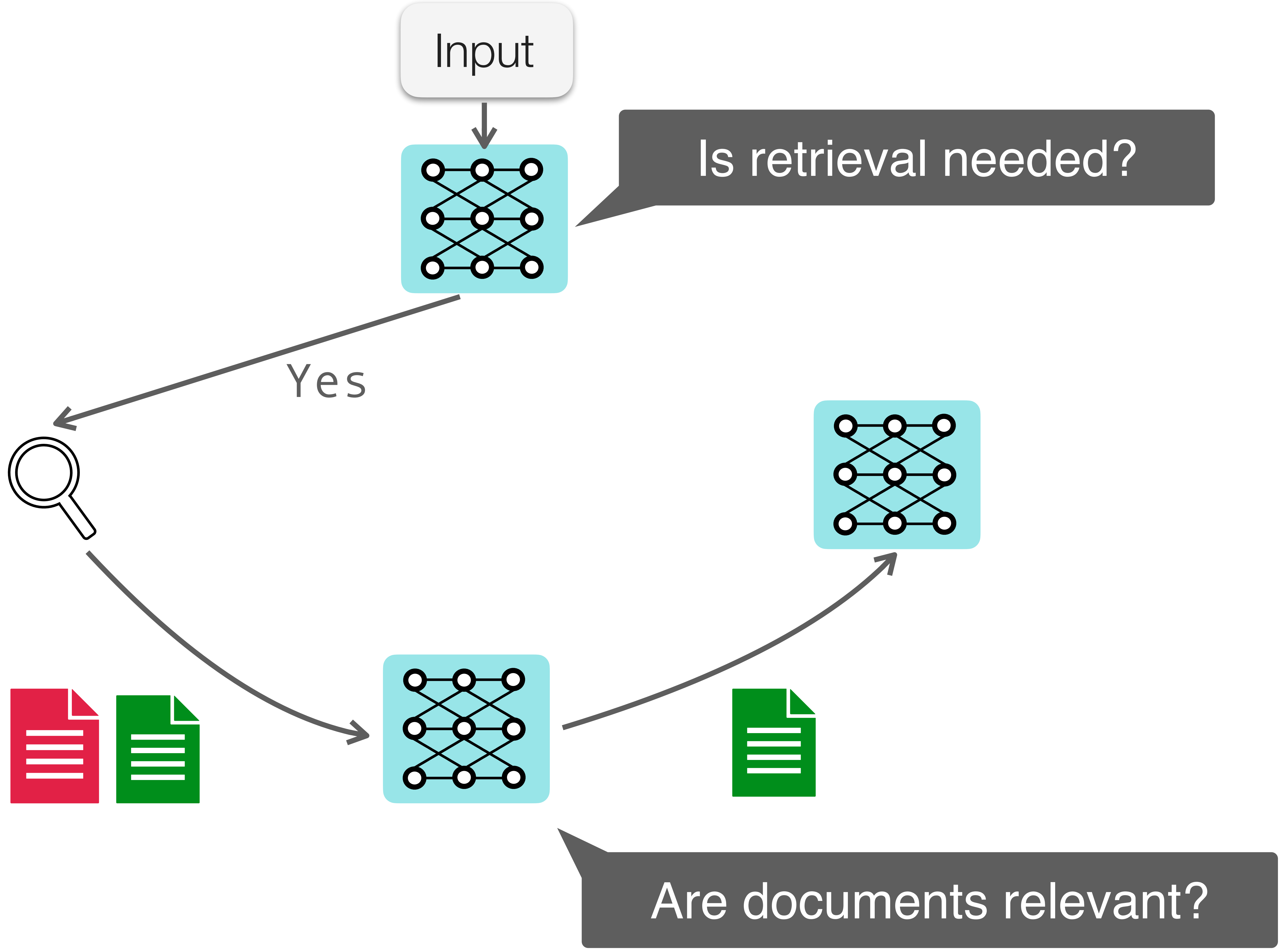
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



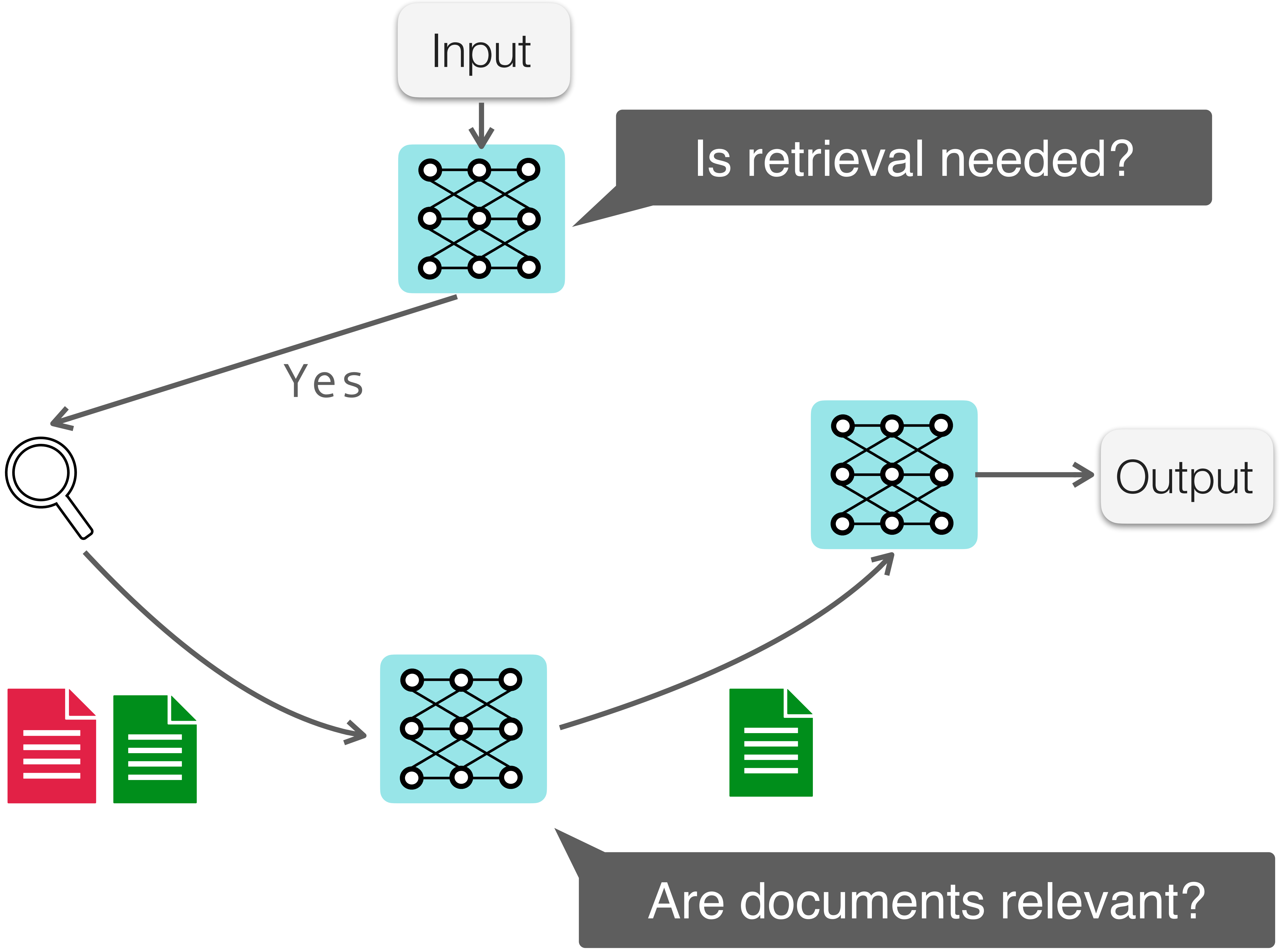
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



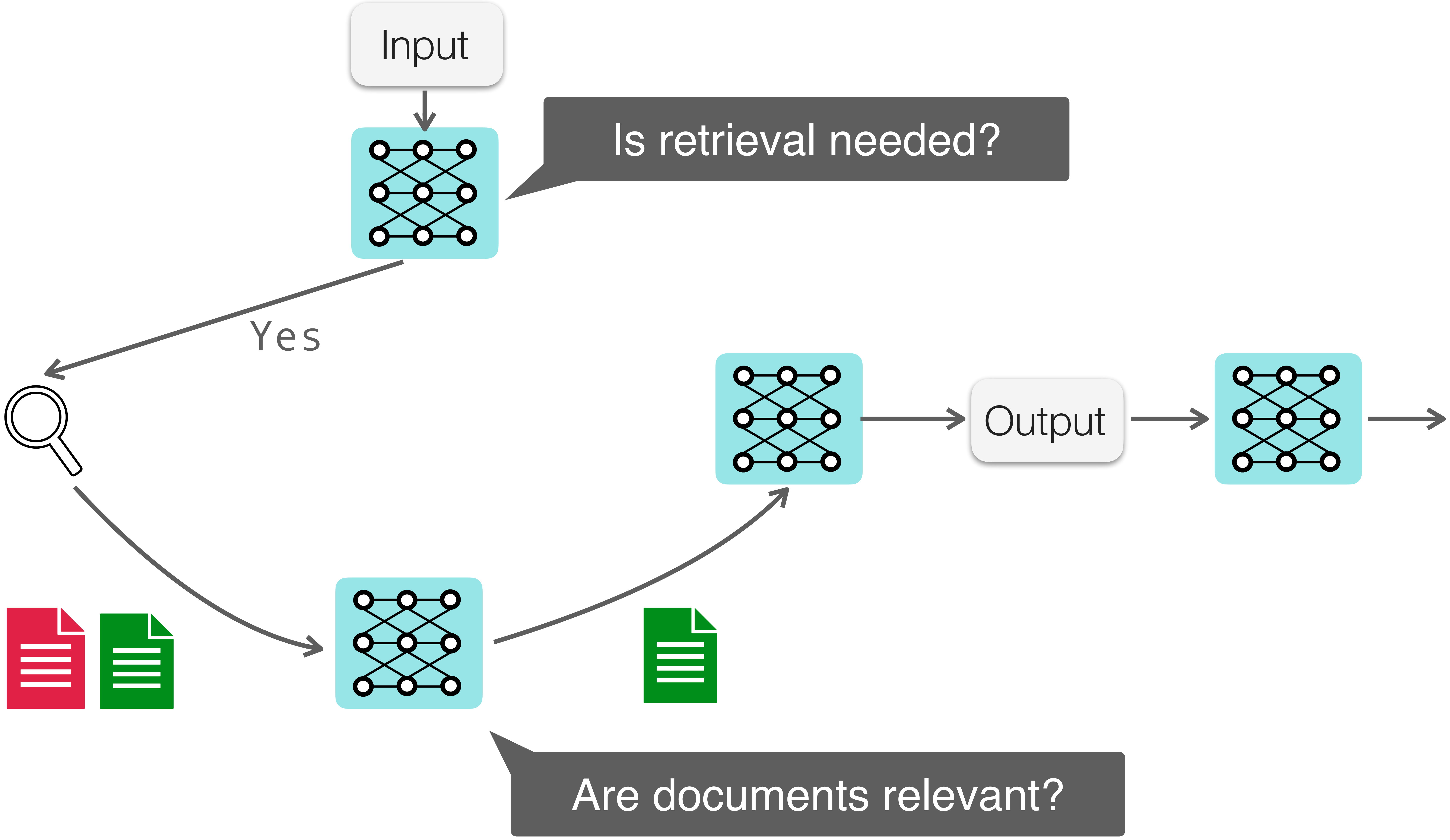
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



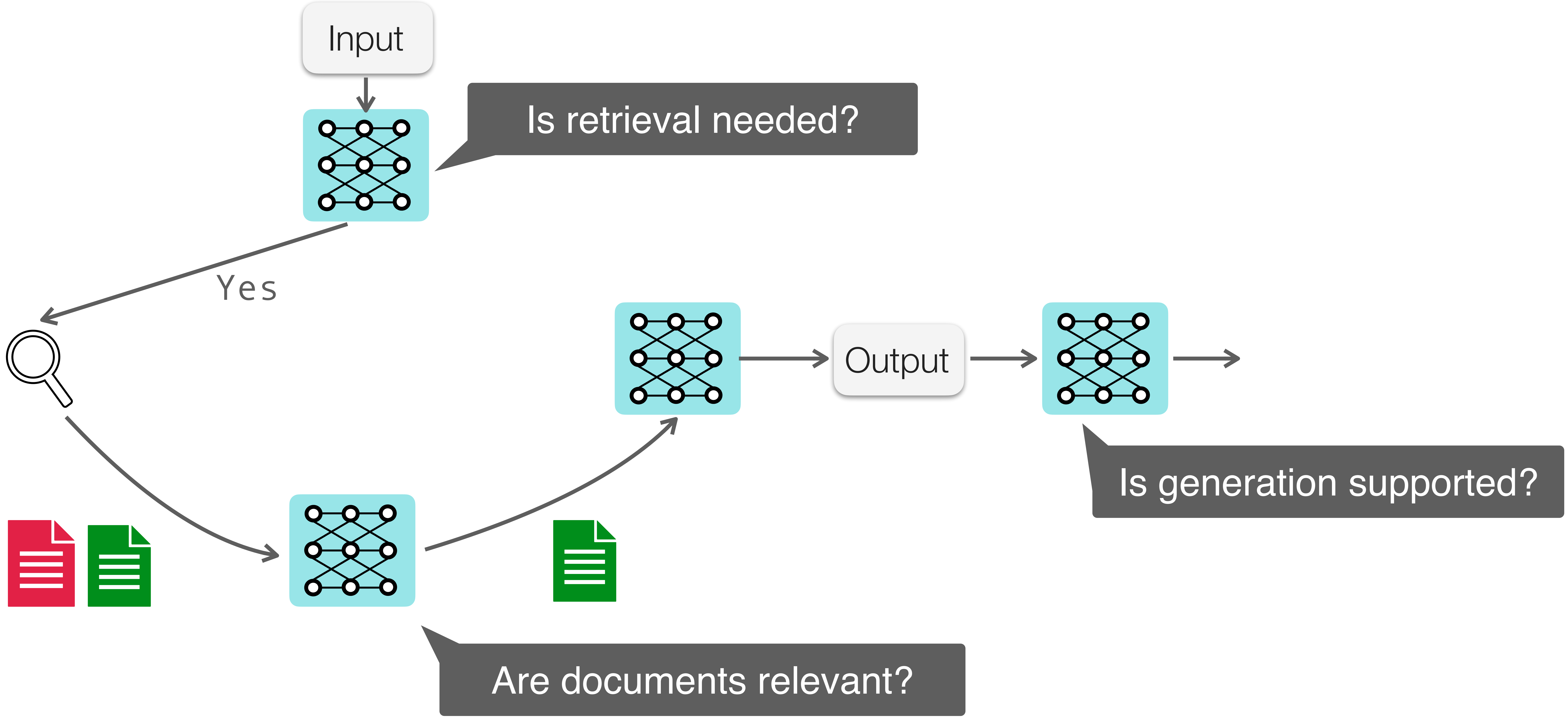
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



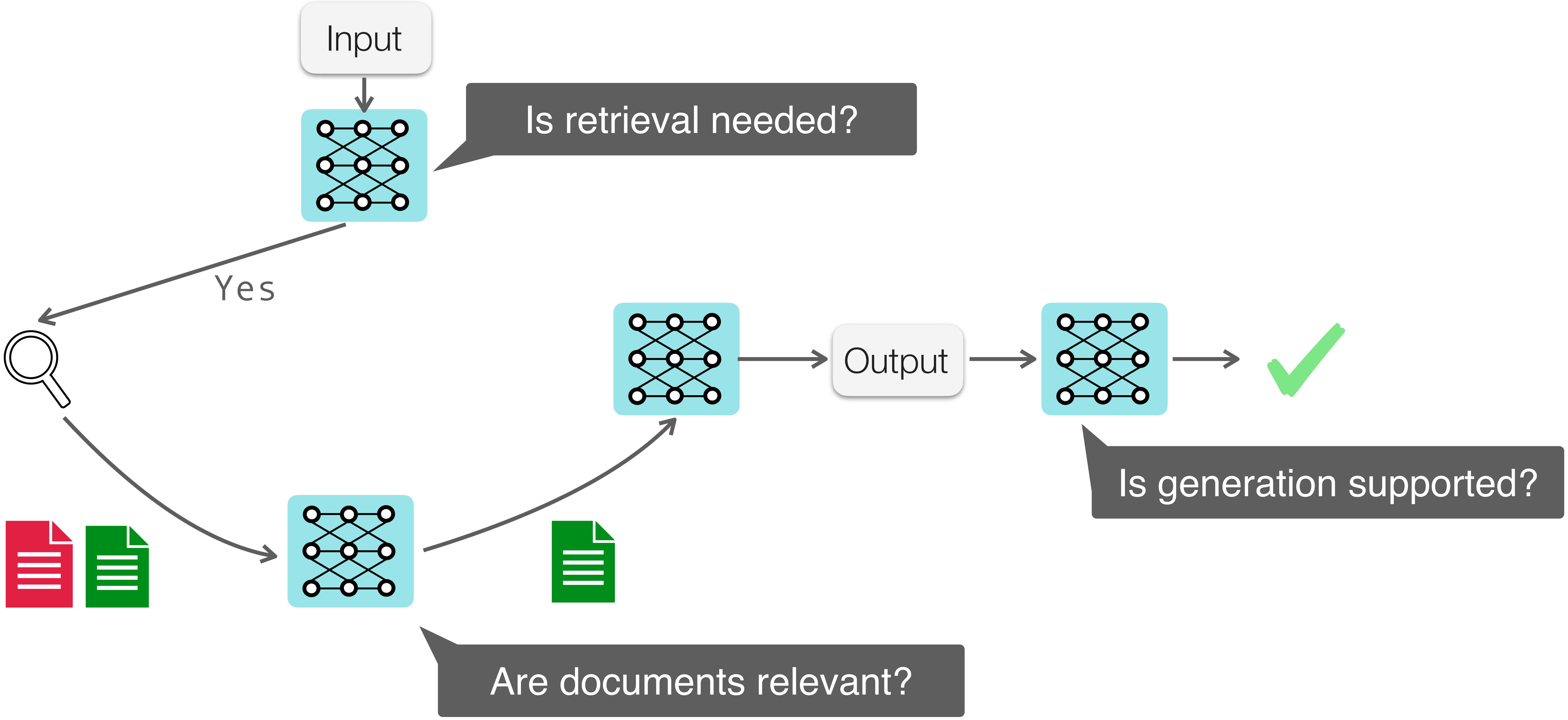
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



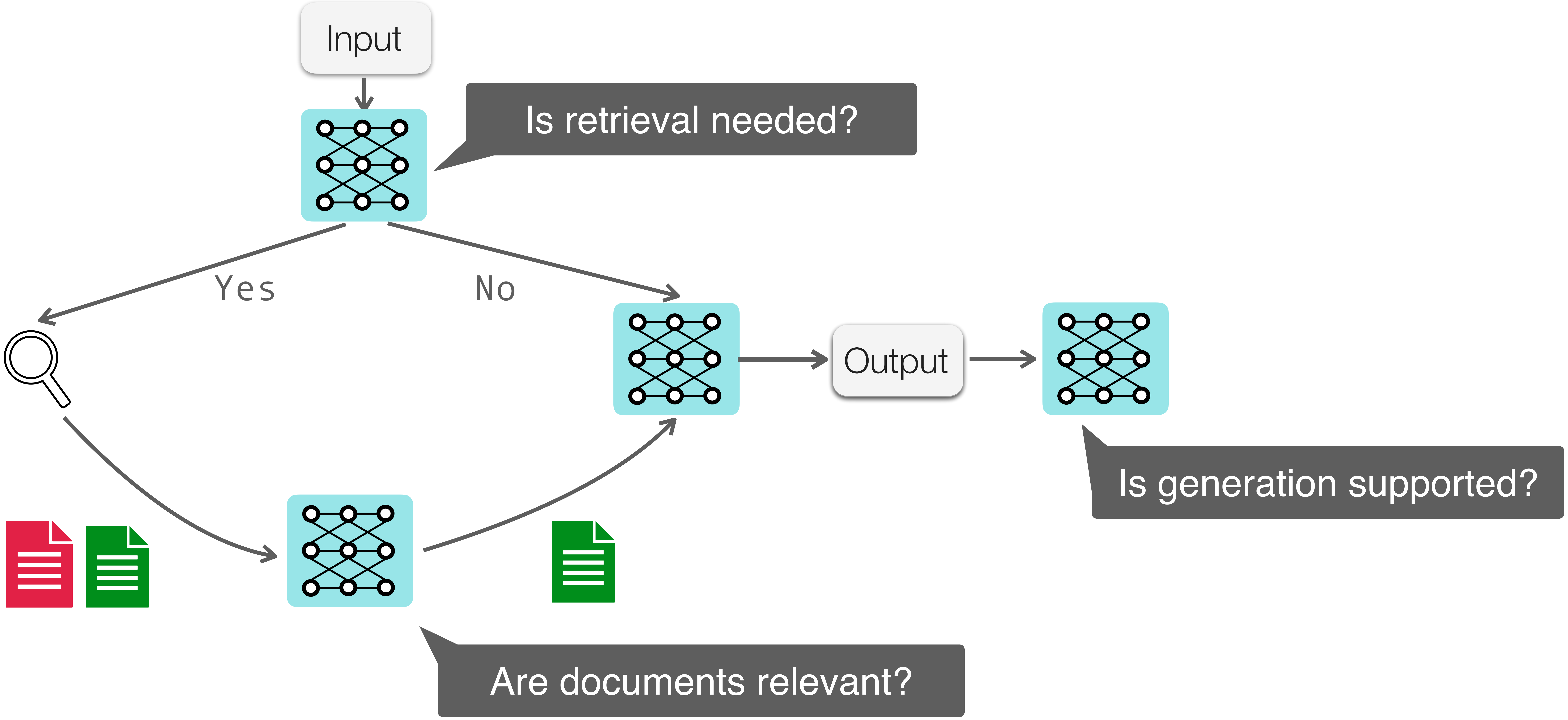
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



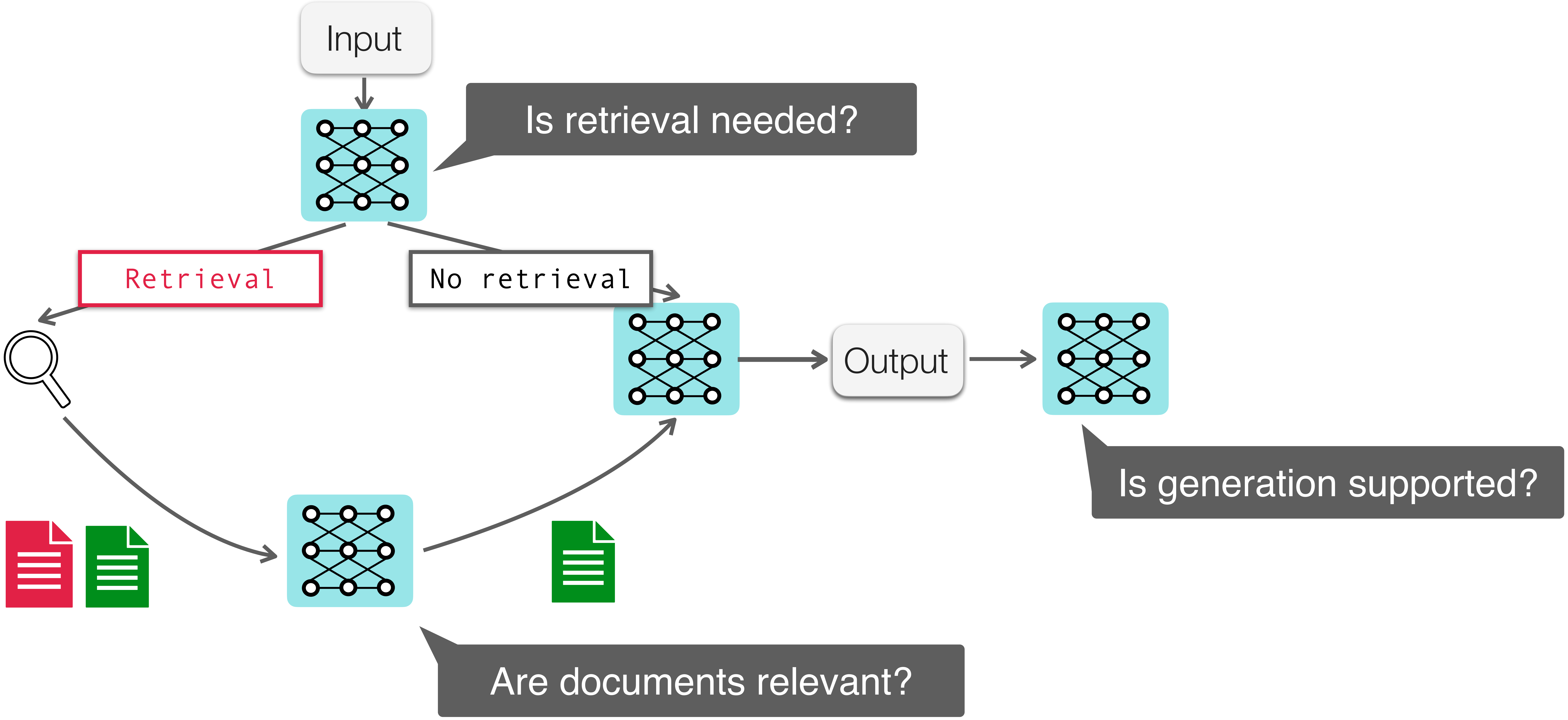
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



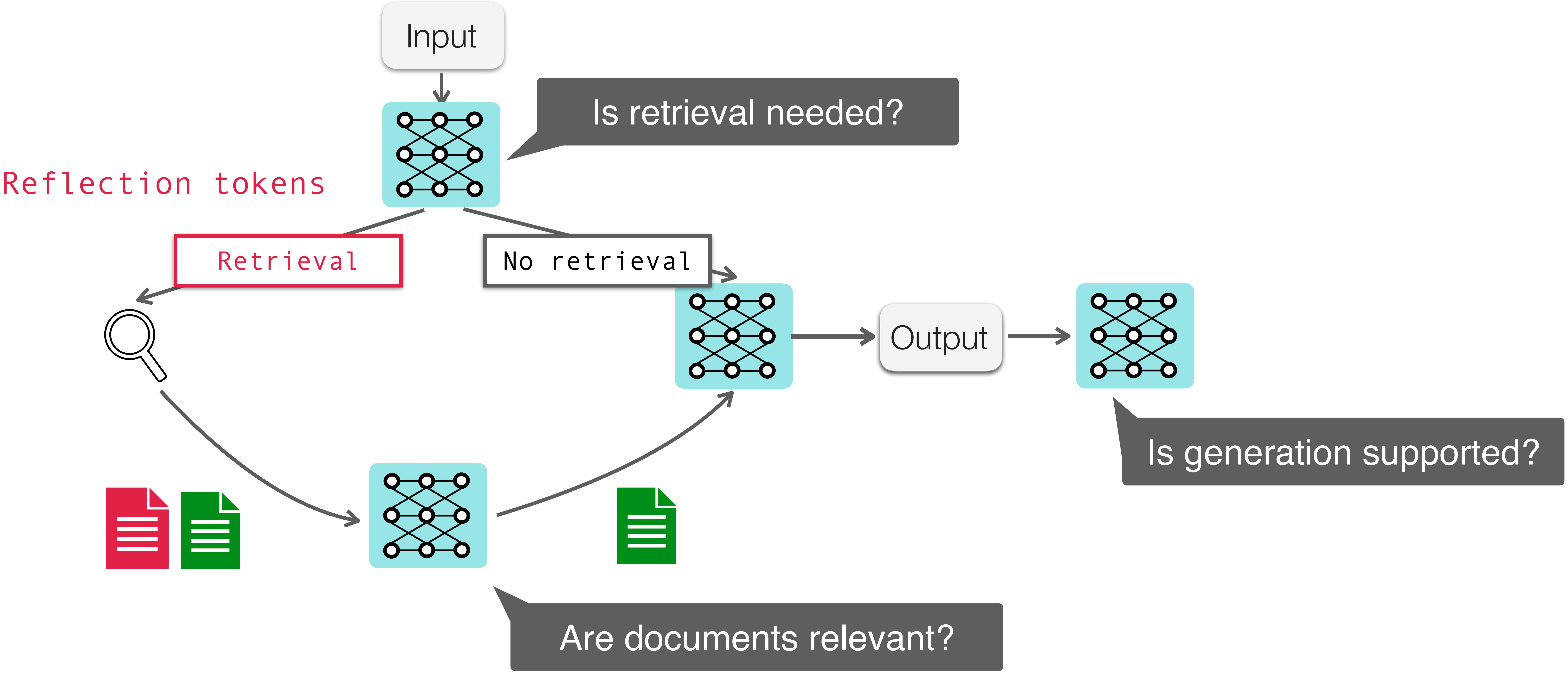
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



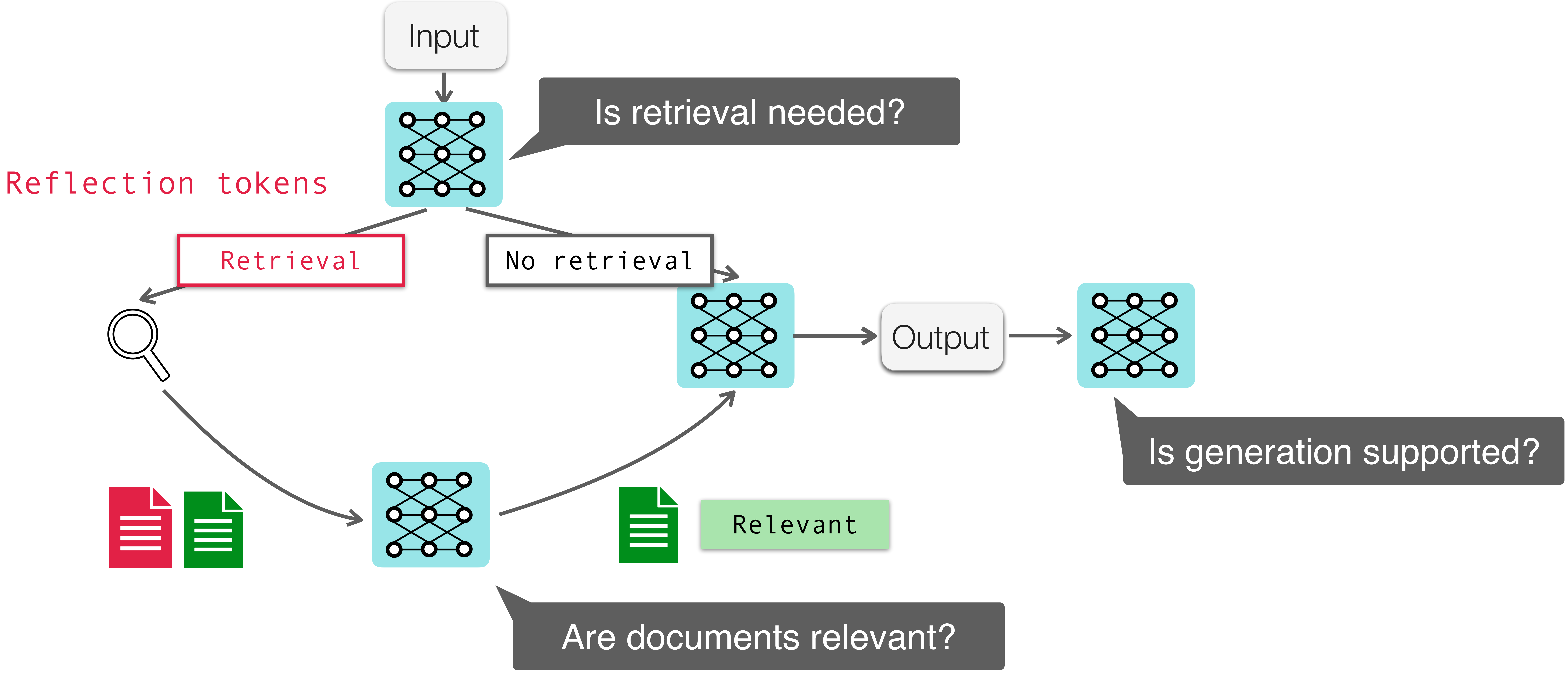
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



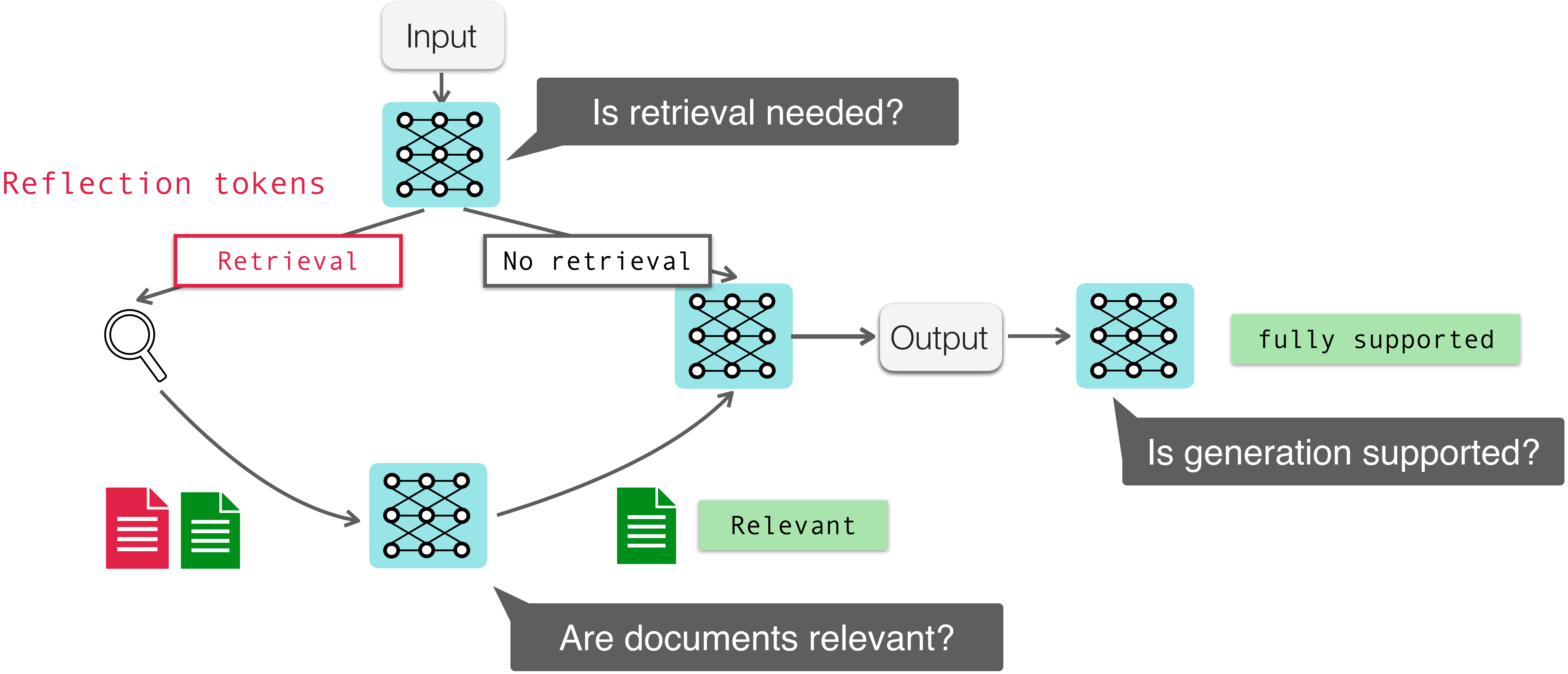
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate and Critique through Self-Reflections

Self-RAG: Learn to Retrieve and Critique



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

Self-RAG: Learn to Retrieve and Critique



Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize
scientific literature
```

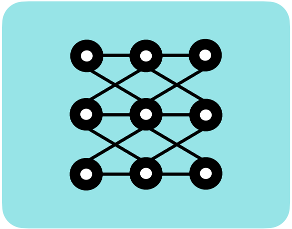
Sentence 1

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

Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize
scientific literature
```

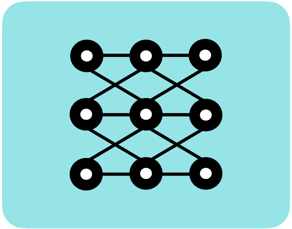


Sentence 1 `Certainly!`

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`



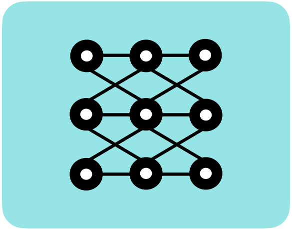
Sentence 1 `Certainly!` `Retrieval`

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

Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize
scientific literature
```



Sentence 1 `Certainly!` `Retrieval`

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

Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature
```

Sentence 1 `Certainly!` `Retrieval`

LLMs have been used in industry widely, such as chatbot system

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

LLMs have been used in industry widely, such as chatbot system

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize
scientific literature
```

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

- `Irrelevant`
- `Relevant`
- `Relevant`

LLMs have been used in industry widely, such as chatbot system

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

- `Irrelevant` ~~LLMs have been widely used in science.~~
- `Relevant`
- `Relevant`

LLMs have been used in industry widely, such as chatbot system

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

- `Irrelevant` ~~`LLMs have been widely used in science.`~~
- `Relevant`
- `Relevant`

LLMs have been used in industry widely, such as chatbot system

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input
```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize
scientific literature
```

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

`Relevant`

`Relevant`

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize
scientific literature
```

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

`Relevant`

`Relevant`

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature
```

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

`Relevant`

`Relevant`

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

`Relevant` `OpenScholar is an LM for literature synthesis.`

`Relevant`

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

`Relevant` `OpenScholar is an LM for literature synthesis.`

`Relevant` `Studies show GPT4o can help scientists for idea generations and literature synthesis.`

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input

```
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature
```

Sentence 1 Certainly! Retrieval

Sentence 2

Relevant OpenScholar is an LM for literature synthesis. fully supported

Relevant Studies show GPT4o can help scientists for idea generations and ~~literature synthesis.~~ Partially supported

OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`

Sentence 1 `Certainly!` `Retrieval`

Sentence 2

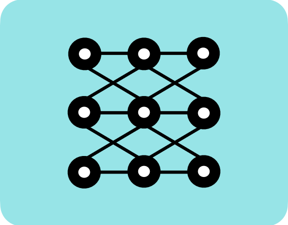
`Relevant` `OpenScholar is an LM for literature synthesis.` `fully supported` **0.9**

`Relevant` `Studies show GPT4o can help scientists for idea generations and` ~~`literature synthesis.`~~ `Partially supported` 0.5

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`



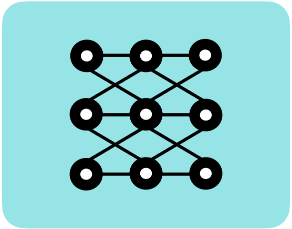
Sentence 1 `Certainly!` `Retrieval`

Sentence 2 `Relevant` `OpenScholar is an LM for literature synthesis.` `fully supported`

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`



Sentence 1 `Certainly!` `Retrieval`

Sentence 2 `Relevant` `OpenScholar is an LM for literature synthesis.` `fully supported`

Sentence 3

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`

Sentence 1 `Certainly!` `Retrieval`

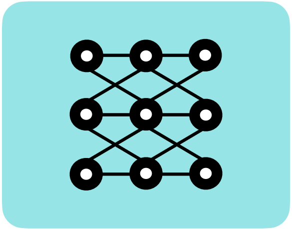
Sentence 2 `Relevant` `OpenScholar is an LM for literature synthesis.` `fully supported`

Sentence 3 `OpenScholar uses retrieval-augmented approaches.`

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

Self-RAG: Learn to Retrieve and Critique

Input `Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature`



Sentence 1 `Certainly!` `Retrieval`

Sentence 2 `Relevant` `OpenScholar is an LM for literature synthesis.` `fully supported`

Sentence 3 `OpenScholar uses retrieval-augmented approaches.` `fully supported`

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

Tool-Augmented LMs

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Use search as needed via API calls

Schick et al. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools.

Tool-Augmented LMs

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

Use search as needed via API calls

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

Can be flexibly extended to other tools

Schick et al. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools.

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.
<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>
<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>
Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.
<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>
<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>
Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.

<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>
<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>
Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.
<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>
<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>
Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Thinking

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.
<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>
<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>
Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.
<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>
<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>
Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Thinking

Tool call

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.

<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>

<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>

Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.

<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>

<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>

Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Thinking

Tool call

Search output

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.

<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>

<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>

Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.

<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>

<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>

Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Thinking

Tool call

Search output

Thinking

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.

<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>

<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>

Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.

<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>

<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>
Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Thinking

Tool call

Search output

Thinking

Tool call

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.

<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>

<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>

Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.

<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>

<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>

Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Thinking

Tool call

Search output

Thinking

Tool call

Search output

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

Deep Research: Coupling Reasoning LM with Search

Example from WebWalkerQA Dataset

Question:
What was the specific schedule for the social event held on the evening after the ACL 2023 best paper awards ceremony?
Labeled Answer:
The social event was scheduled from 7:00 PM to 10:30 PM on July 11th.

Model Output by WebThinker

Okay, so the user is asking about the specific schedule of a social event that took place on the evening *after* the ACL 2023 best paper awards ceremony. First, I need to figure out when the awards ceremony actually happened. Let me search for that specifically.

<|begin_search_query|>ACL 2023 best paper awards ceremony date<|end_search_query|>

<|begin_search_result|>The ACL 2023 best paper awards ceremony took place on **July 11, 2023**, according to multiple sources: [...] Web Page 4 [...] explicitly states that the awards were announced on July 11, 2023. [...] Web Page 5 [...] mentions the award presentation occurred on July 11, 2023 [...]. Web Page 6 [...] confirms the awards were announced on July 11, 2023. [...] Web Page 9 [...] also indicates the awards were announced on July 11, 2023. [...] the awards ceremony itself was held on the **11th**. [...]<|end_search_result|>

Okay, the awards ceremony was definitively on July 11, 2023. The question asks for the social event on the evening *after* the ceremony, which would be the evening of July 12. Let me search for the social events schedule to see what was planned.

<|begin_search_query|>ACL 2023 social events schedule<|end_search_query|>

<|begin_search_result|>The ACL 2023 social events schedule includes the following relevant details: 1. **Social Event on July 11, 2023**: - **Time**: 7:00 PM – 10:30 PM (EDT) - **Location**: Steam Whistle Brewing [...] - **Description**: This event occurred on the evening of July 11, immediately following the Best Paper Awards ceremony (which took place earlier that day at 16:00–17:30). [...] 2. **No Explicit Social Events on July 12, 2023**: - The evening of July 12, 2023, is not listed as having any official social events in the provided sources. [...] **Note**: The primary social event referenced in the sources is on July 11, not July 12. The Best Paper Awards ceremony was held on July 11, so the subsequent evening event likely refers to the Social Event on that same day. For July 12, only daytime sessions are documented. [...]<|end_search_result|>

Hmm, interesting. The awards were on July 11, and the search results show a social event on the evening of July 11, immediately following the awards ceremony (7:00 PM - 10:30 PM). However, the results explicitly state there were no official social events scheduled for the evening of July 12, which is technically the evening *after* the July 11 ceremony.

Teach reasoning LMs to plan and search via SFT / RL

Thinking

Tool call

Search output

Thinking

Tool call

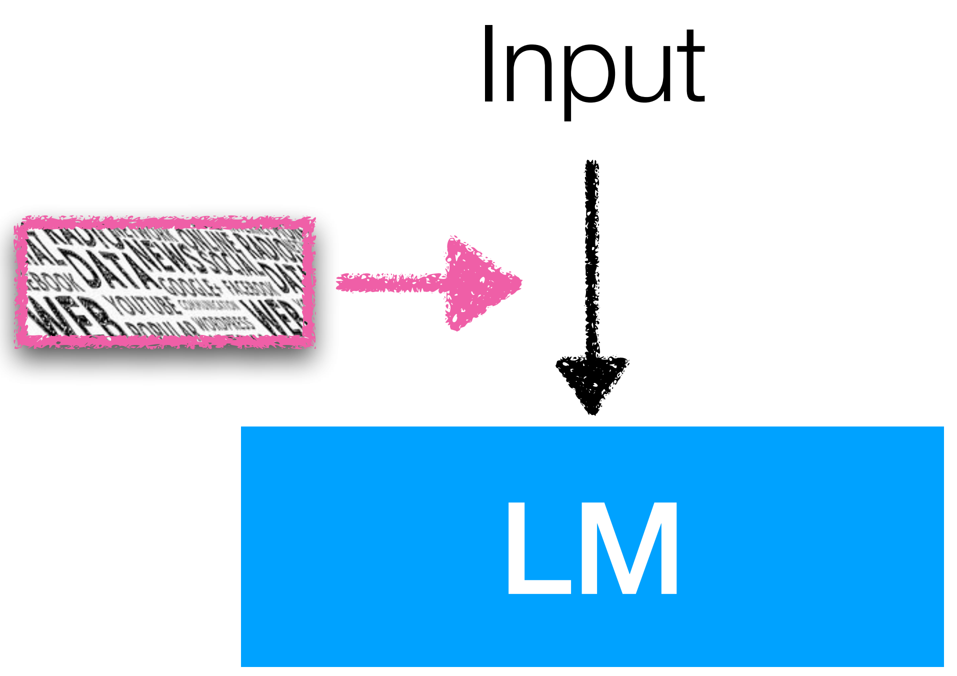
Search output

Thinking

Li et al. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capacity

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

Input

LM

Not scalable to many documents (needs context engineering)

- Mo... LM... to incorporate D in intermediate layers
- Scalable to many passages
- Requires retraining

Not strictly grounded

e.g., RETRO, InstructRETRO

Output Interpolation

LM

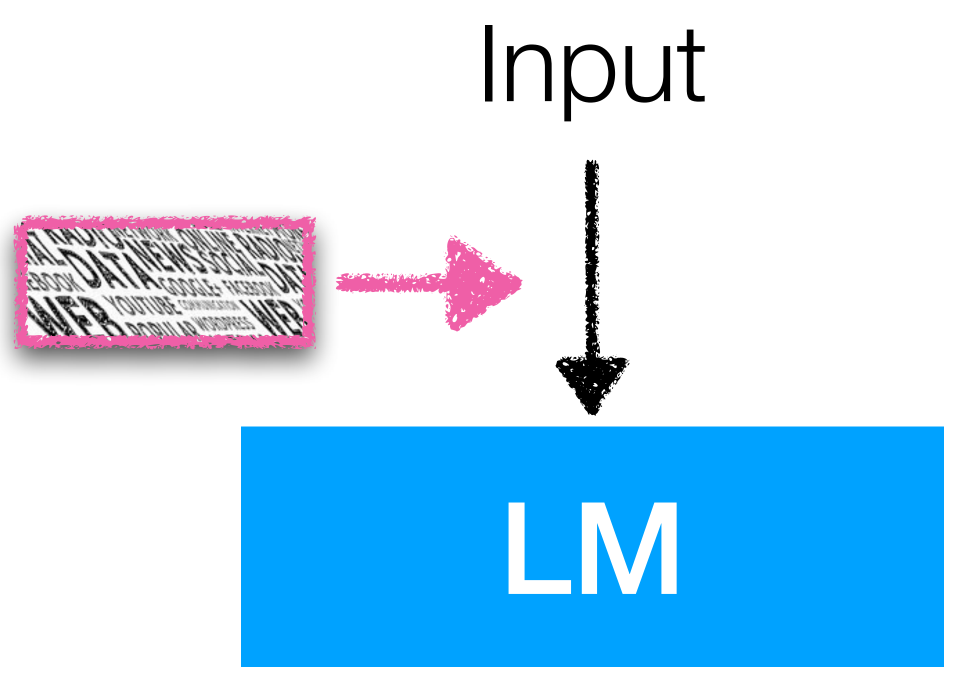
Output

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

e.g., kNNLM

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

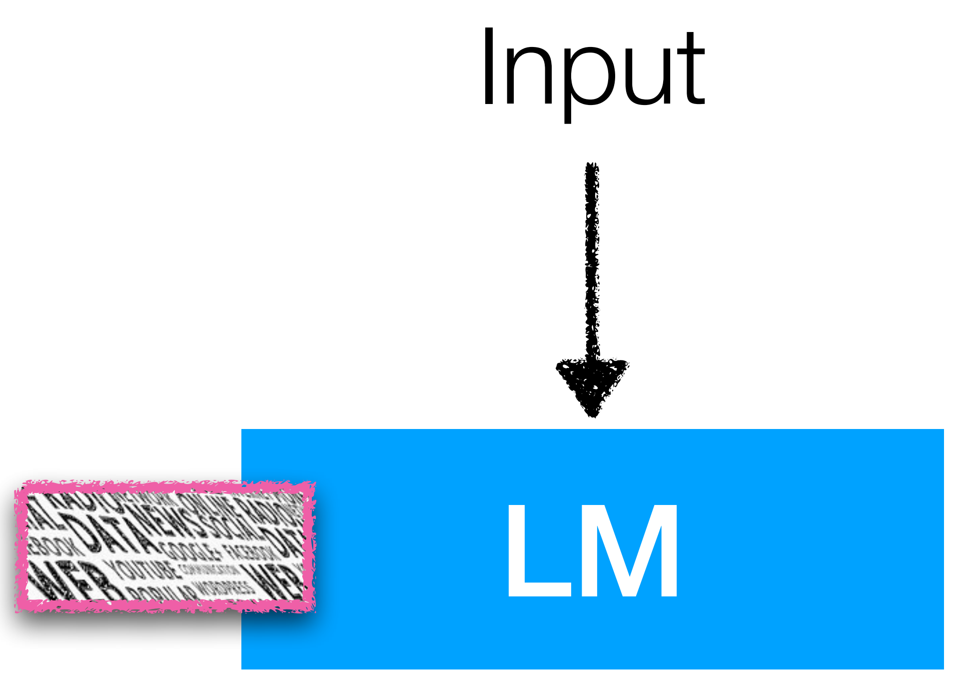
How to Use Retrieval

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

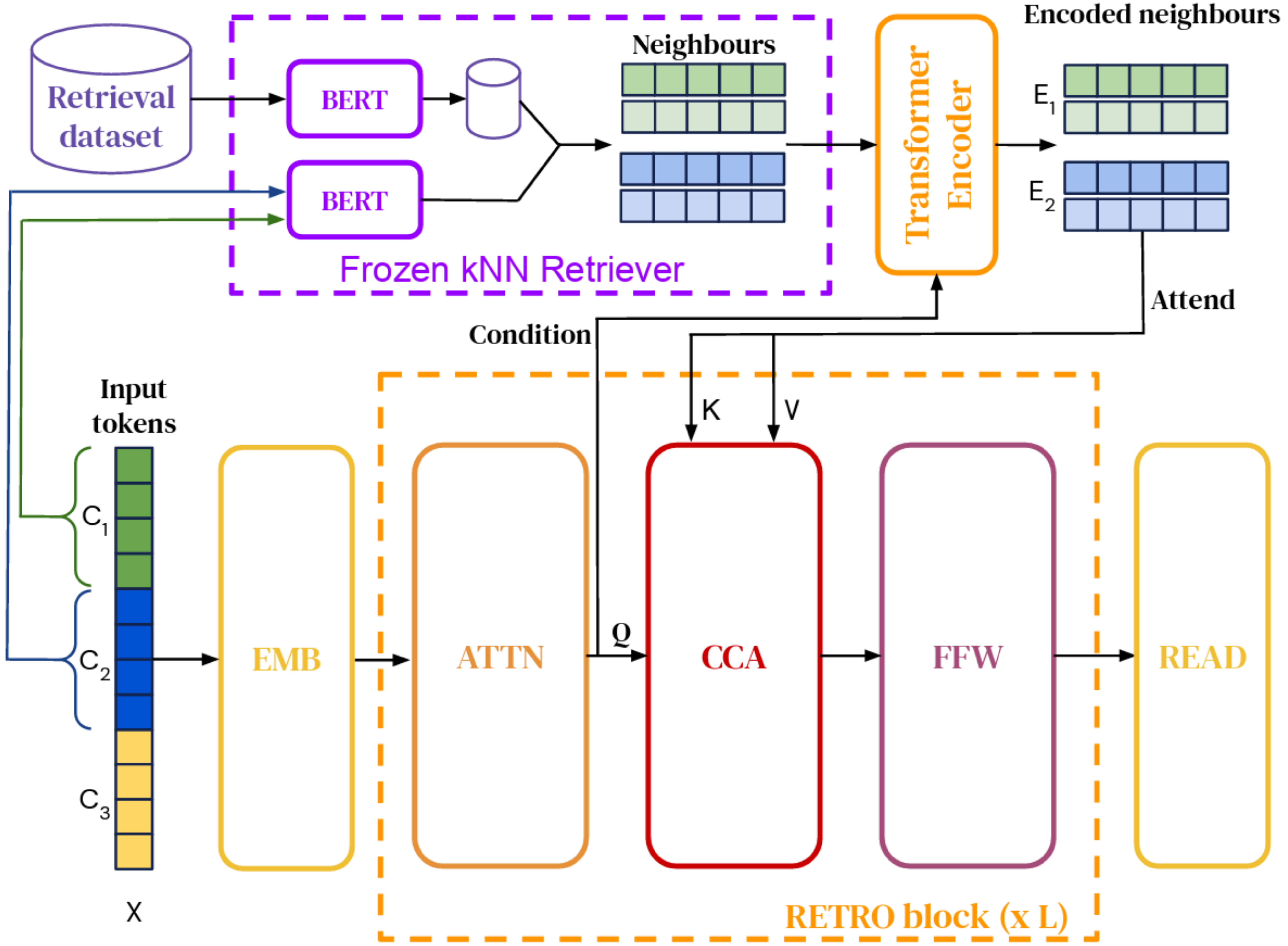
e.g., RETRO, InstructRETRO

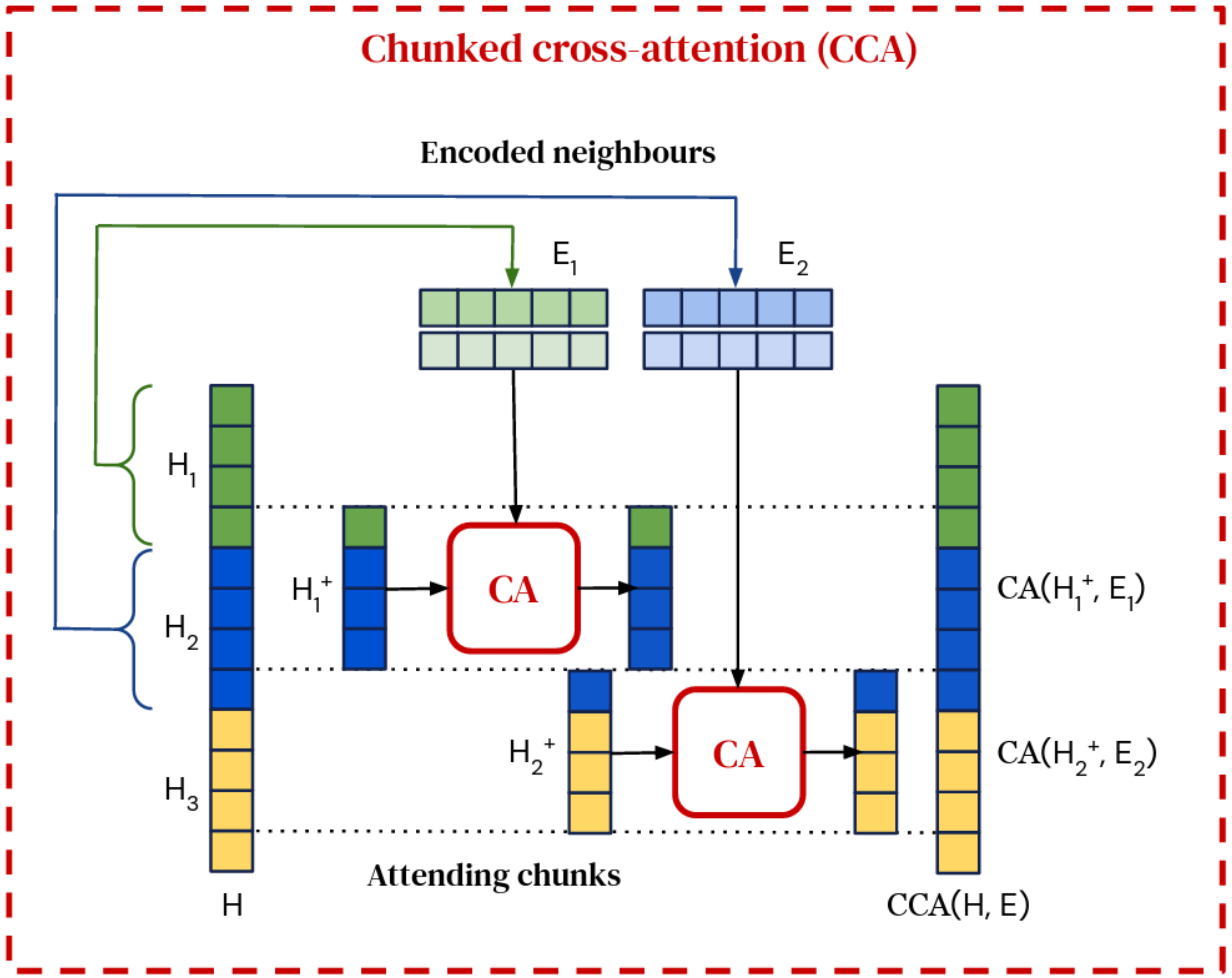
Output Interpolation

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

e.g., kNNLM

RETRO (Borgeaud et al., 2022)





Borgeaud et al. 2022. Improving language models by retrieving from trillions of tokens.

How to Use Retrieval

Input Augmentation

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

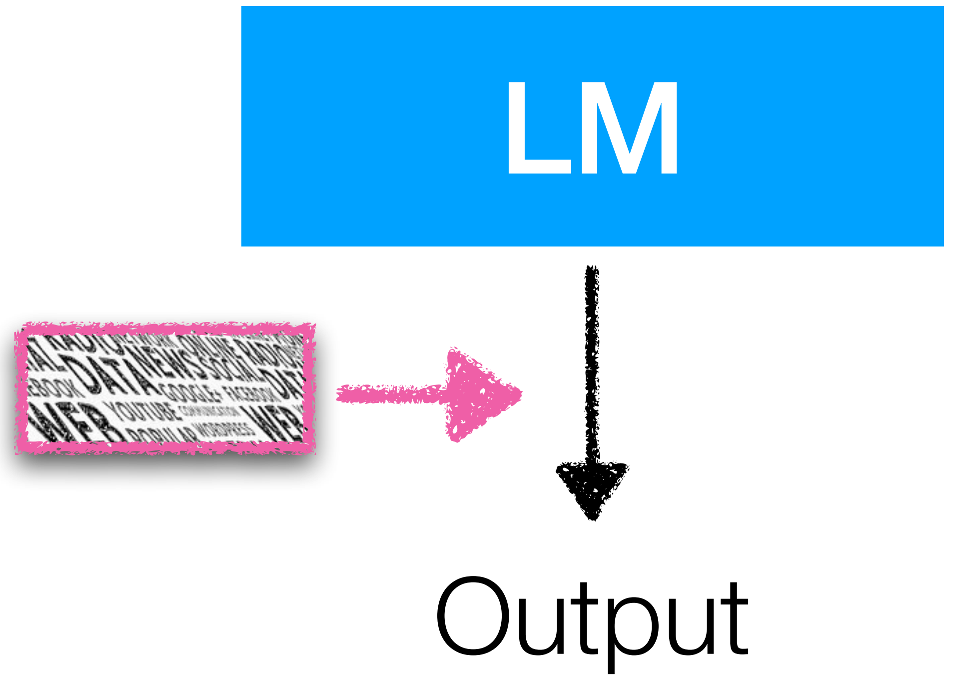
e.g., RAG

Intermediate Fusion

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

e.g., RETRO, InstructRETRO

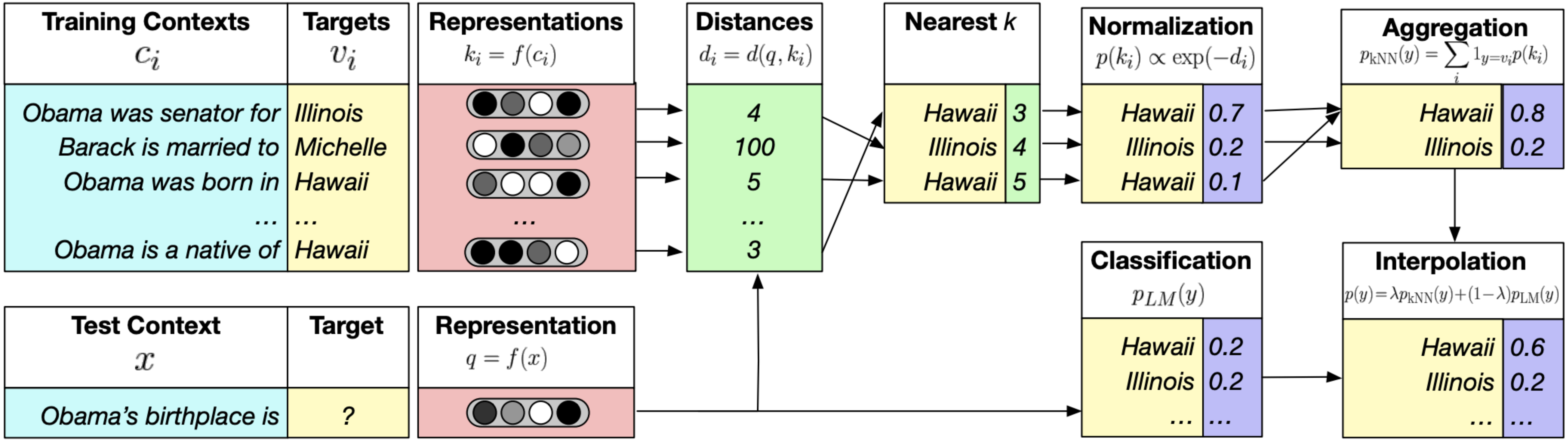
Output Interpolation



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

e.g., kNNLM

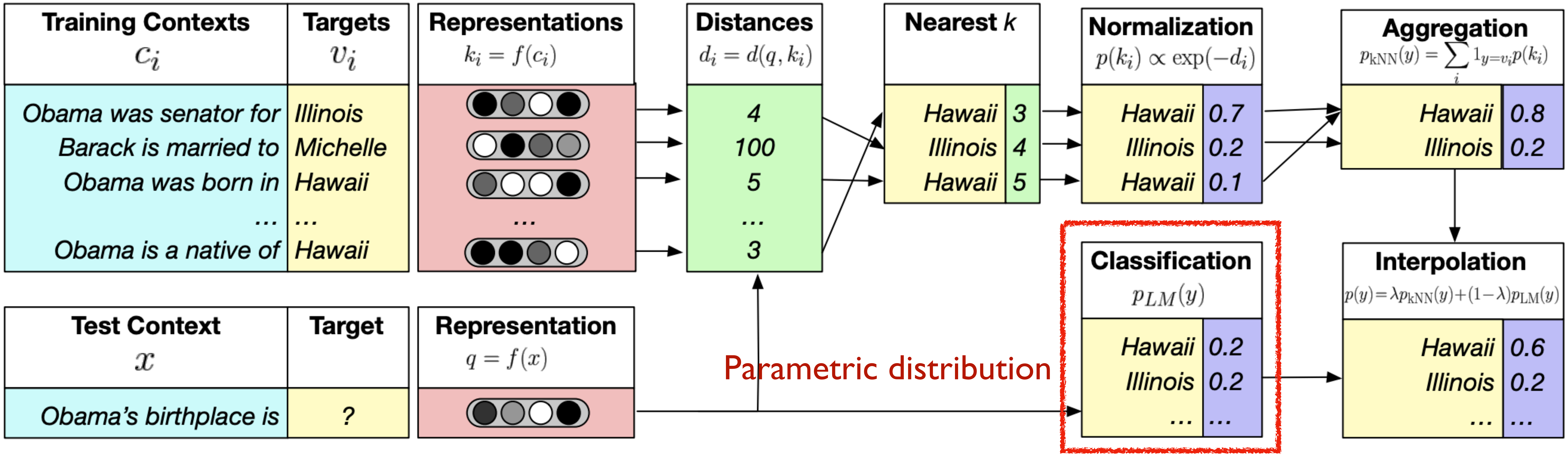
kNN-LM (Khandelwal et al. 2020)



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

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

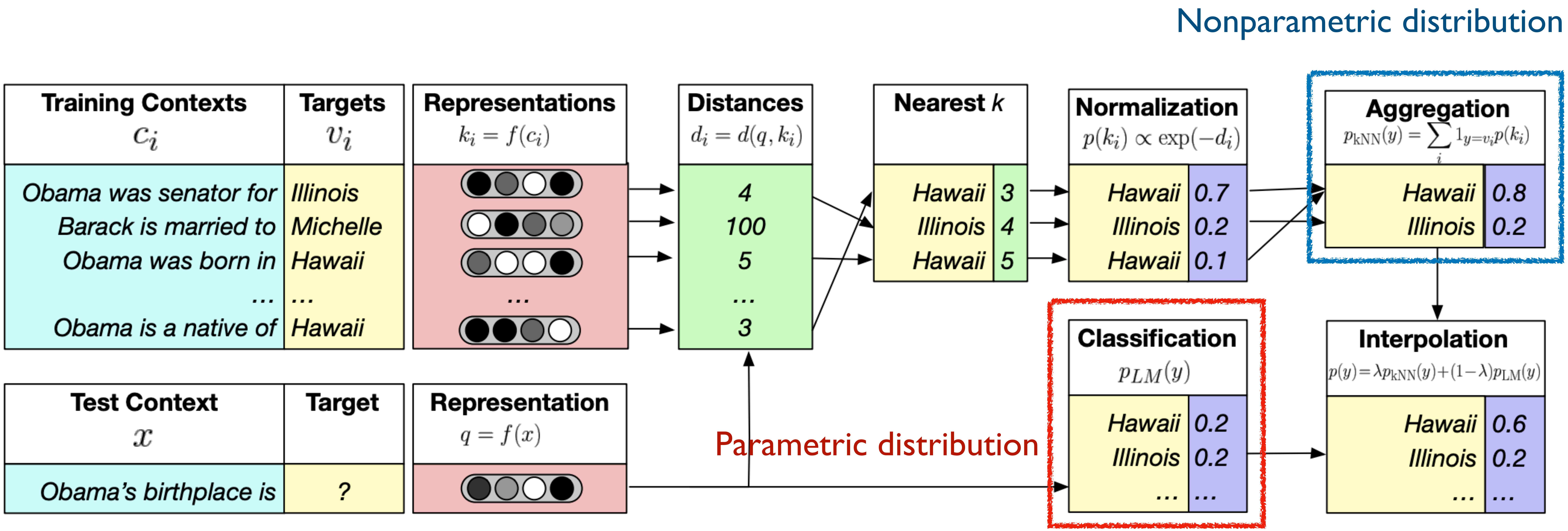
kNN-LM (Khandelwal et al. 2020)



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

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

kNN-LM (Khandelwal et al. 2020)

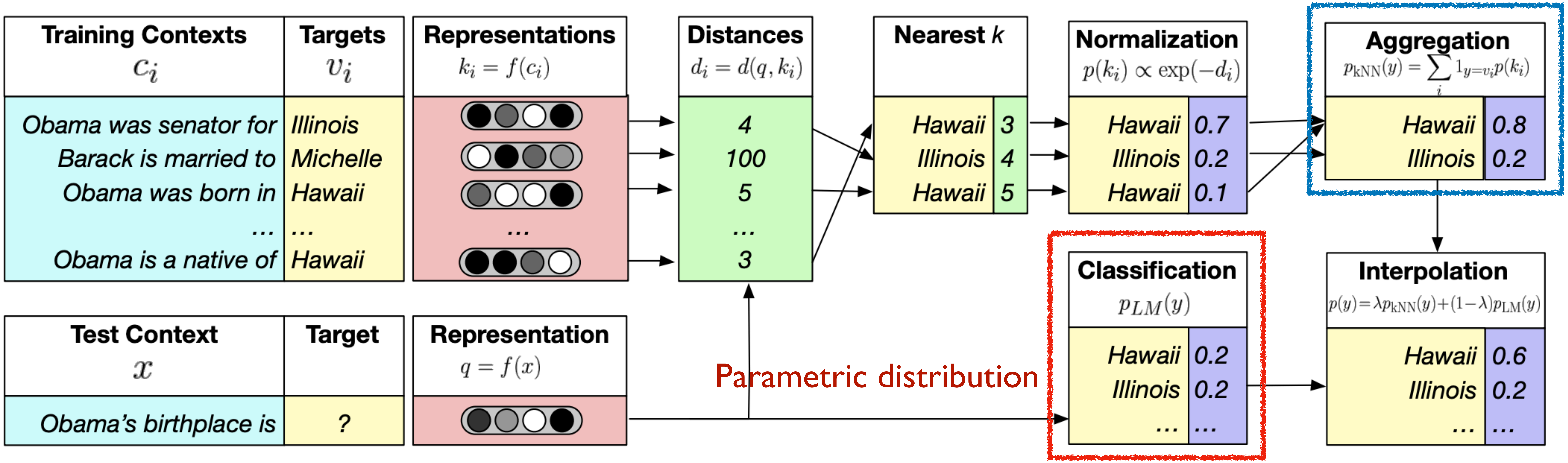


$$P_{k\mathrm{NN-LM}}(y \mid x) = (1-\lambda) P_{\mathrm{LM}}(y \mid x) + \lambda P_{k\mathrm{NN}}(y \mid x)$$

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

kNN-LM (Khandelwal et al. 2020)

Nonparametric distribution

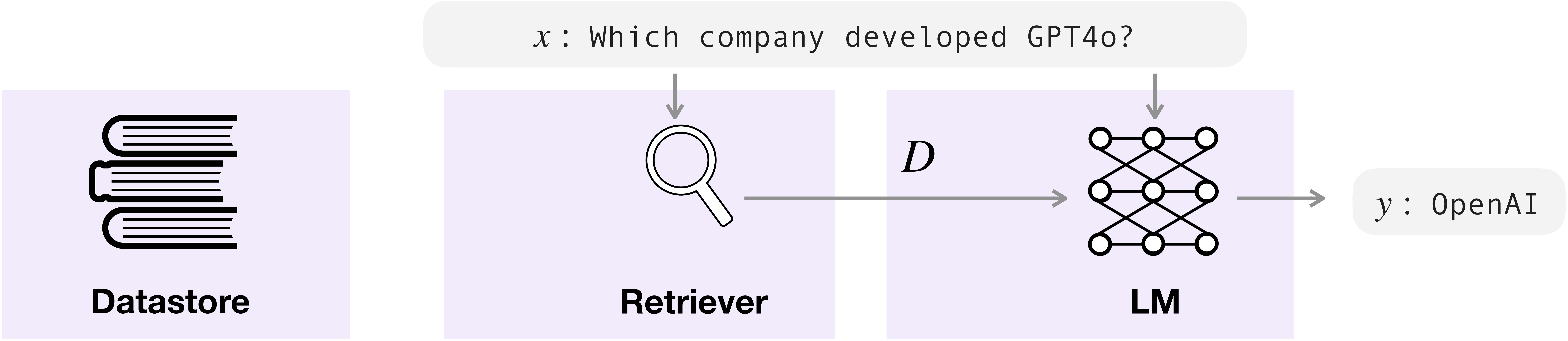


λ: hyperparameter

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

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

Summary of Part 3

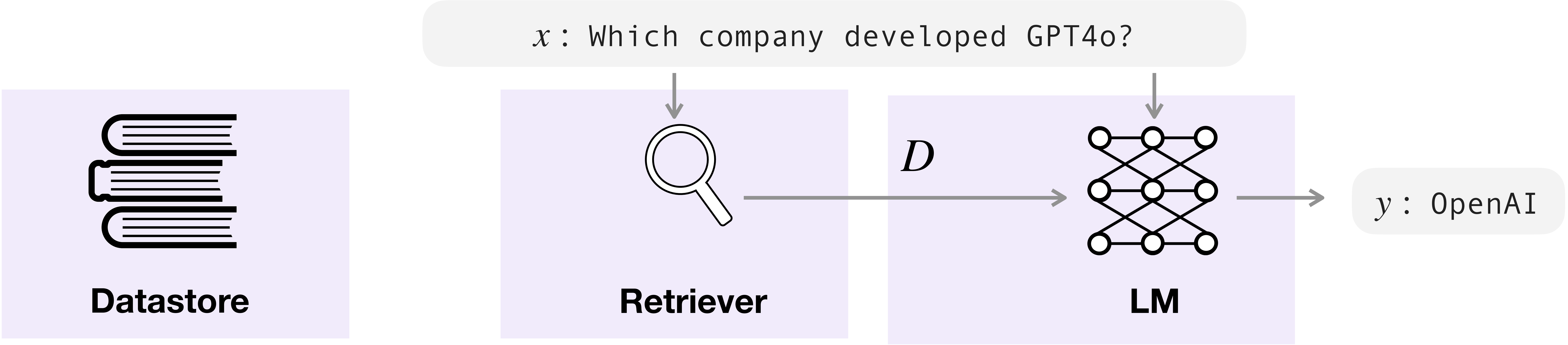


✓ Common architectures

✓ Recent progress

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

Retrieval and Retrieval-Augmented Generation



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

Contact:

https://akariasai.github.io/ aasai@andrew.cmu.edu | akaria@allenai.org