Retrieval-based LLMs and Retrieval-augmented Generation

Yatin Nandwani

Slides from –

ACL 2023 Tutorial - https://acl2023-retrieval-lm.github.io

Graham Neubig Lecture - https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs *– close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs *– close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Parametric LLMs

Parametric LLMs – amazing creative writers!

YA **You**

Narrate the story of rabbit and tortoise in pirate style.

‹ 5 / 5 ›

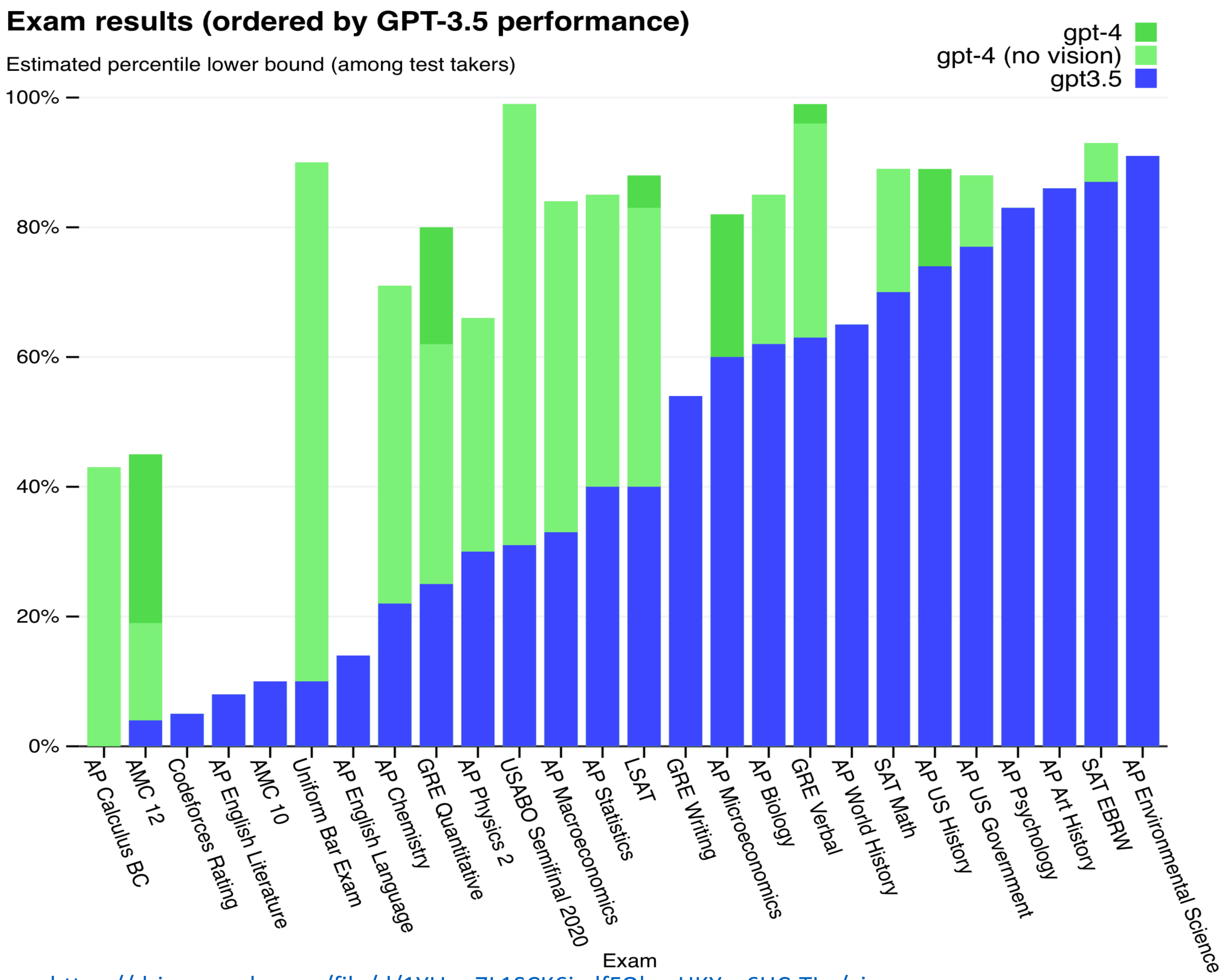
ChatGPT

Ahoy there, me hearties! Gather 'round and hear the tale of the great race betwixt the hare and the tortoise, but with a swashbucklin' twist!

Once upon the high seas, there lived a swift and cunning hare named Jackrabbit, known far and wide for his lightning speed and cocky swagger. And in the same waters, there dwelt a slow but steady tortoise by the name of Captain Shellback, renowned for his patience and resilience.

One fine day, as the salty breeze danced through the rigging, Captain Shellback issued a challenge to Jackrabbit, proclaimin', "Arrr, I challenge ye to a race, ye boastful hare! From this here cove to yonder island, we shall see who be the fastest sailor on the seven seas!"

Parametric LLMs – Stellar performance in exams!



OpenAI – GPT-4 Technical Report
https://arxiv.org/abs/2303.08774

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Hallucinations in Parametric LLMs

YA **You**

what is the use of PMI in NLP?

ChatGPT

In natural language processing (NLP), Pointwise Mutual Information (PMI) is often used as a measure of the association between words in a corpus. It can be used to identify words that are highly correlated and occur frequently together, which can be useful for tasks such as language modeling, information retrieval, and machine translation.

For example, in language modeling, PMI can be used to identify the most likely next word in a sentence given the context of the previous words. In information retrieval, PMI can be used to identify the most relevant documents for a given query by ranking documents based on the co-

Hallucinations in Parametric LLMs

Overall, PMI is a useful measure for identifying relationships between words in a corpus and can be applied to a wide range of NLP tasks.

YA **You**

Can you give me citations for the above examples?

ChatGPT

Here are some references for the examples of using PMI in natural language processing (NLP) that I provided:

- Language modeling: "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)

Hallucinations in Par

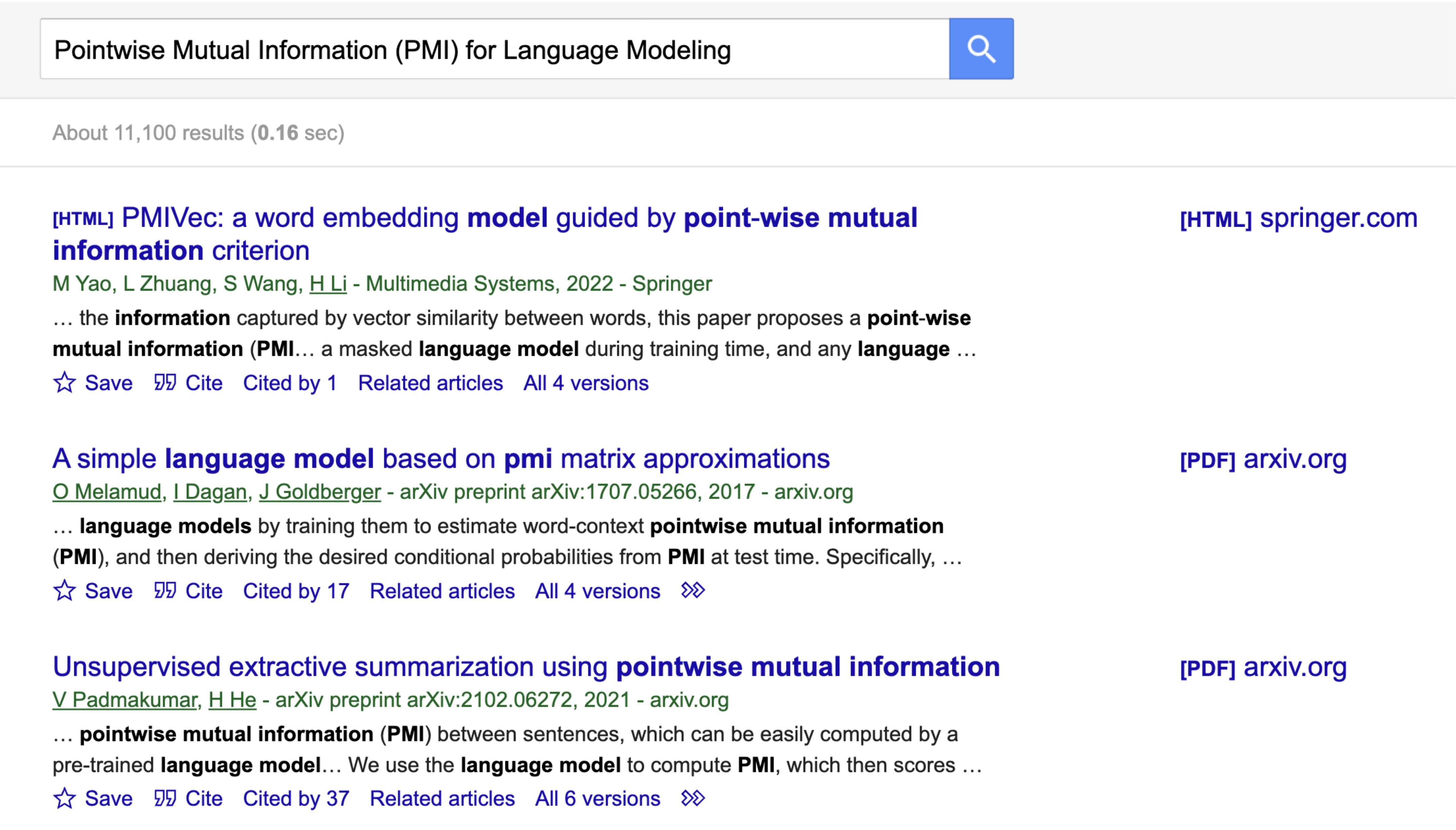
Overall, PMI is a useful measure for identifying re
be applied to a wide range of NLP tasks.

YA **You**

Can you give me citations for the above example

ChatGPT

Here are some references for the examples of us
I provided:



Pointwise Mutual Information (PMI) for Language Modeling

About 11,100 results (**0.16** sec)

[HTML] PMIVec: a word embedding **model** guided by **point-wise mutual information** criterion — [HTML] springer.com

M Yao, L Zhuang, S Wang, H Li - Multimedia Systems, 2022 - Springer

… the **information** captured by vector similarity between words, this paper proposes a **point-wise mutual information** (**PMI**… a masked **language model** during training time, and any **language** …

Save Cite Cited by 1 Related articles All 4 versions

A simple **language model** based on **pmi** matrix approximations — [PDF] arxiv.org

O Melamud, I Dagan, J Goldberger - arXiv preprint arXiv:1707.05266, 2017 - arxiv.org

… **language models** by training them to estimate word-context **pointwise mutual information** (**PMI**), and then deriving the desired conditional probabilities from **PMI** at test time. Specifically, …

Save Cite Cited by 17 Related articles All 4 versions

Unsupervised extractive summarization using **pointwise mutual information** — [PDF] arxiv.org

V Padmakumar, H He - arXiv preprint arXiv:2102.06272, 2021 - arxiv.org

… **pointwise mutual information** (**PMI**) between sentences, which can be easily computed by a pre-trained **language model**… We use the **language model** to compute **PMI**, which then scores …

Save Cite Cited by 37 Related articles All 6 versions

- Language modeling: "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)

Hallucinations in

Overall, PMI is a useful measure for ide

be applied to a wide range of NLP task

You

Can you give me citations for the above

ChatGPT

Here are some references for the exam

I provided:

- Language modeling: "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)

Information Retrieval using PMI-based Word Representations

About 947 results (**0.03** sec)

Handling the impact of low frequency events on co-occurrence based measures of **word** similarity-a case study of pointwise mutual **information** — **[PDF]** scitepress.org

F Role, M Nadif - … on Knowledge Discovery and **Information Retrieval**, 2011 - scitepress.org

… Additional visualization methods will also be investigated in order to determine how to best help researchers and corpus practitioners in choosing the right **PMI based** association …

Save Cite Cited by 71 Related articles All 3 versions

Sparsity makes sense: **Word** sense disambiguation **using** sparse contextualized **word representations** — **[PDF]** aclanthology.org

G Berend - Proceedings of the 2020 Conference on Empirical …, 2020 - aclanthology.org

… **word representations**, it becomes possible to surpass the results of more complex task-specific models on the task of finegrained all-**words word** … 3 further ***PMI-based** variants for the …

Save Cite Cited by 21 Related articles All 3 versions

Improving pointwise mutual **information** (pmi) by incorporating significant co-occurrence — **[PDF]** arxiv.org

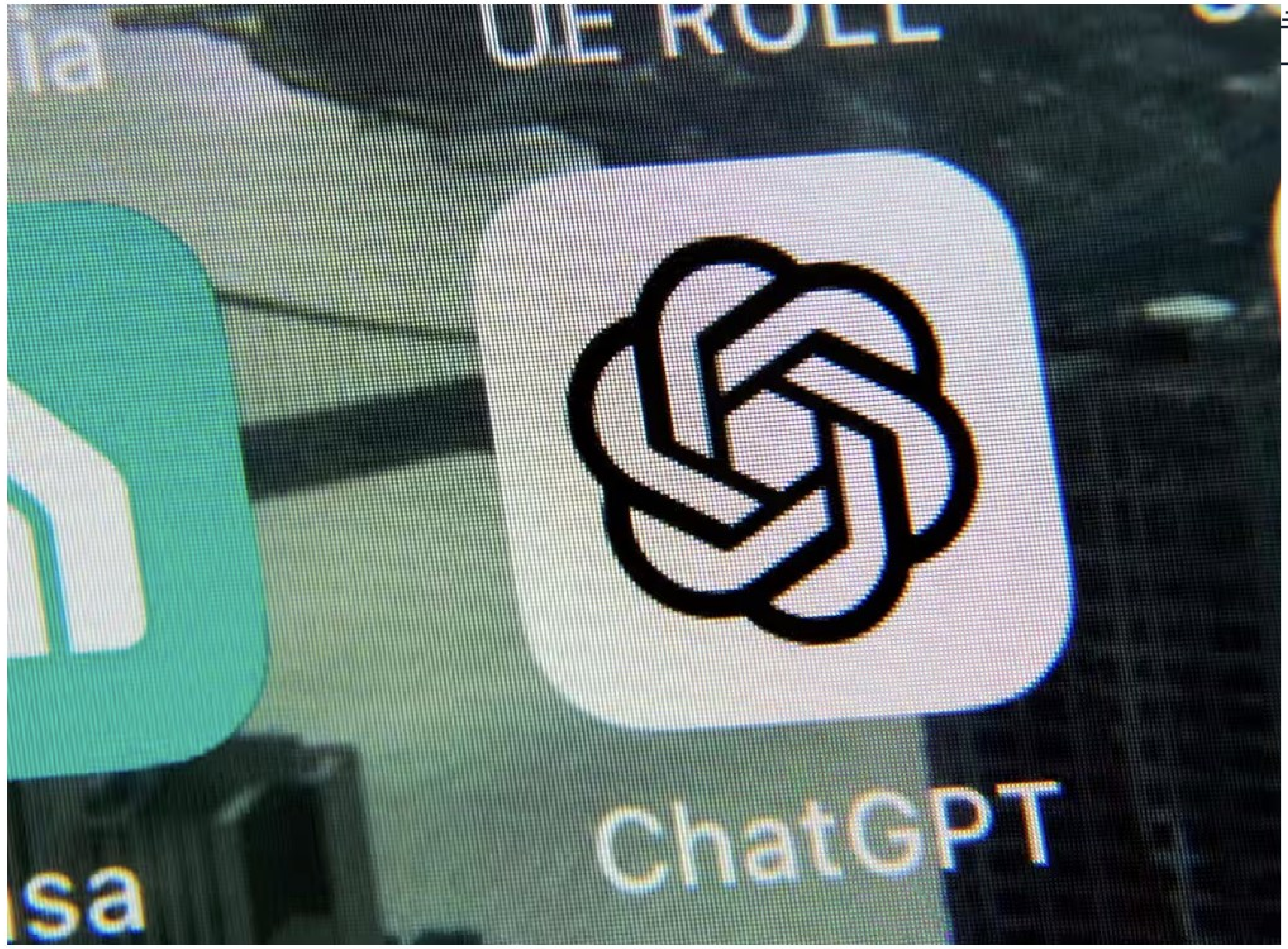
OP Damani - arXiv preprint arXiv:1307.0596, 2013 - arxiv.org

… **word** association measures. To incorporate document level significance, we need to **use** document level counts instead of **word** … are **using** document counts instead of the **word** counts, …

Save Cite Cited by 51 Related articles All 11 versions

Two US lawyers fined for submitting fake court citations from ChatGPT

Law firm also penalised after chatbot invented six legal cases that were then used in an aviation injury claim



The judge said one of the fake decisions had 'some traits that are superficially actual judicial decisions' but other portions contained 'gibberish' and were 'nonse
Photograph: Richard Drew/AP

A US judge has fined two lawyers and a law firm $5,000 (£3,935
citations generated by ChatGPT were submitted in a court filing

Mashable India Entertainment Culture Tech Science Mobility Podcast Mashable Deals

Tech

Air Canada Loses Court Case After Its Chatbot Hallucinated Fake Policies To a Customer

The airline argued that the chatbot itself was liable. The court disagreed.

By Chase Dibenedetto Feb. 18, 2024

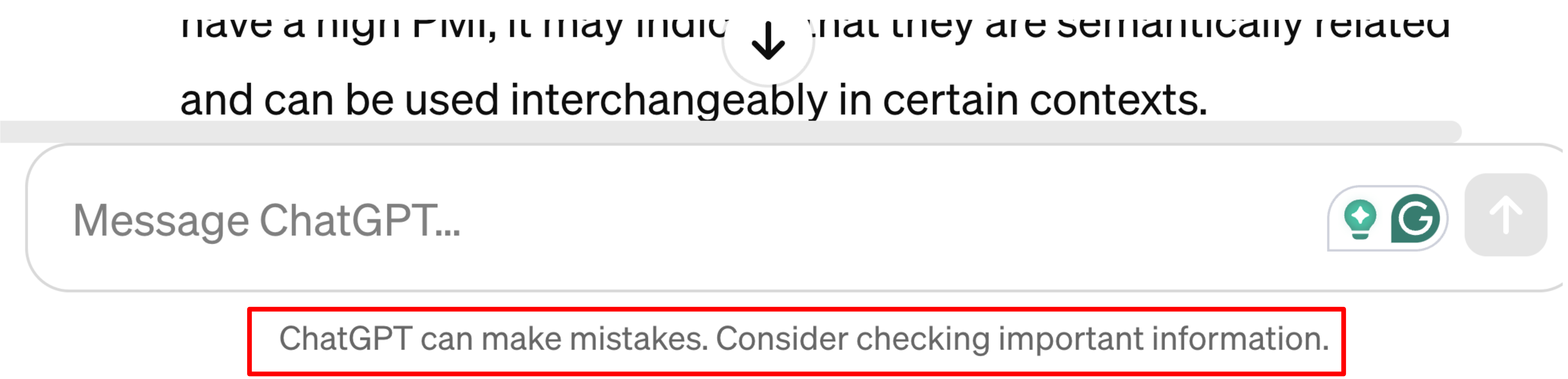
Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Core Limitations of Parametric LLMs

- Hallucinations

Core Limitations of Parametric LLMs

- Hallucinations
- Verifiability issues

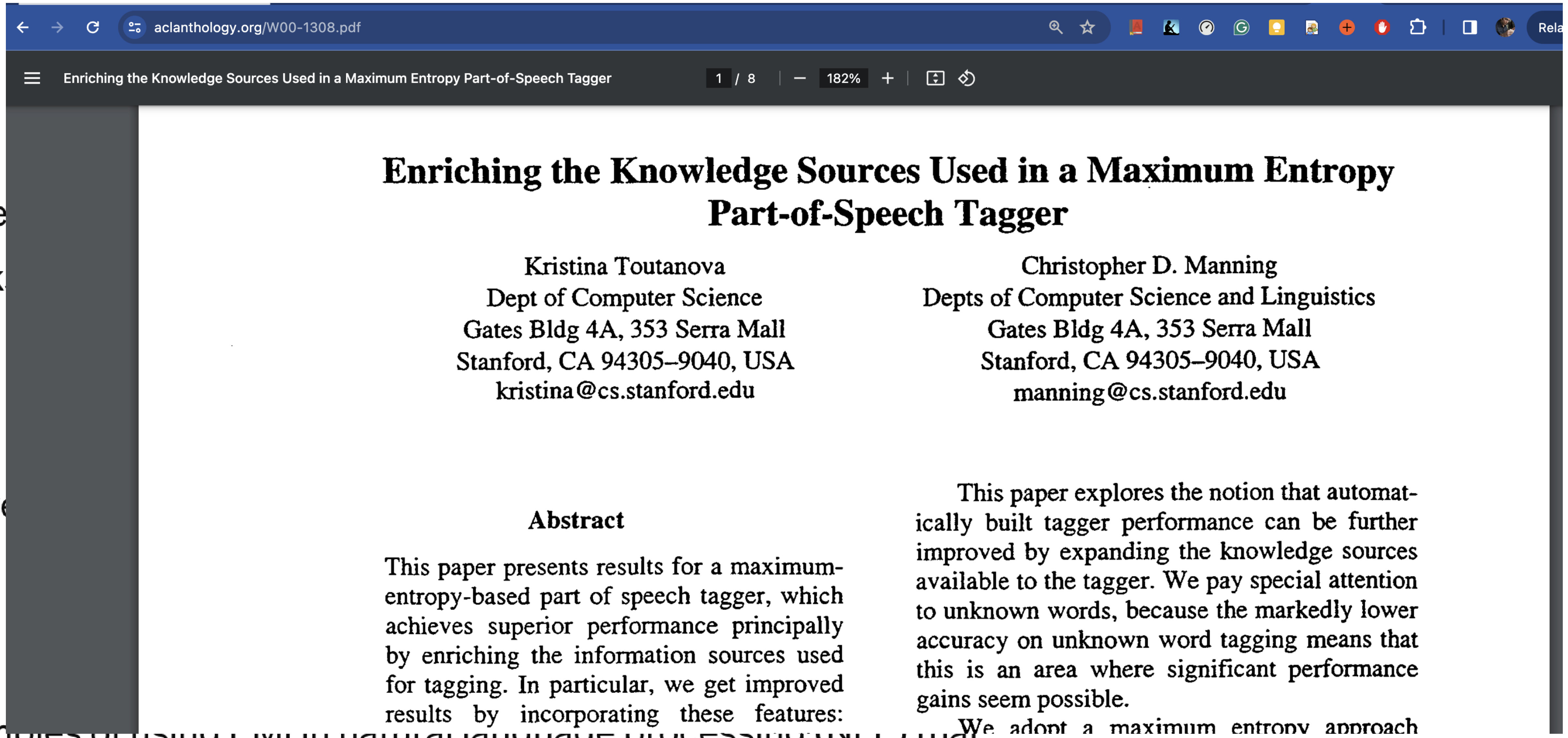


Verifiability

Overall, PMI is a useful measure for ide
be applied to a wide range of NLP task

YA **You**

Can you give me citations for the above



Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger

Kristina Toutanova
Dept of Computer Science
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
kristina@cs.stanford.edu

Christopher D. Manning
Depts of Computer Science and Linguistics
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
manning@cs.stanford.edu

Abstract

This paper presents results for a maximum-entropy-based part of speech tagger, which achieves superior performance principally by enriching the information sources used for tagging. In particular, we get improved results by incorporating these features:

This paper explores the notion that automatically built tagger performance can be further improved by expanding the knowledge sources available to the tagger. We pay special attention to unknown words, because the markedly lower accuracy on unknown word tagging means that this is an area where significant performance gains seem possible.

We adopt a maximum entropy approach

ChatGPT

Here are some references for the examples of using PMI in natural language processing (NLP) that I provided:

- Language modeling: "Pointwise Mutual Information (PMI) for Language Modeling" by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)

Verifiability

Overall, PMI is a useful measure for ide

be applied to a wide range of NLP task

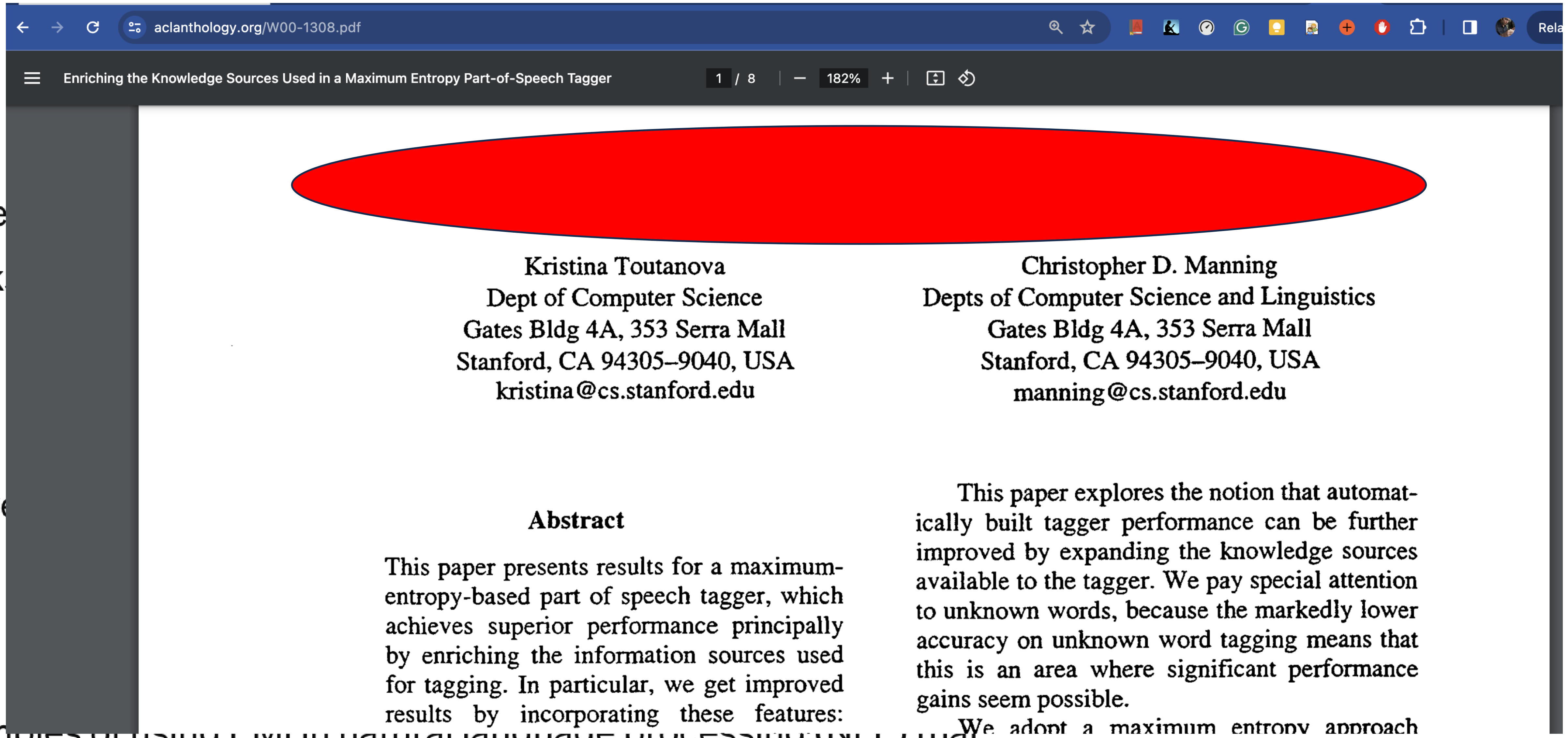
You

Can you give me citations for the above

ChatGPT

Here are some references for the examples of using PMI in natural language processing (NLP) that I provided:

- Language modelin [redacted] by R. B. Frazer and J. H. Lafferty (https://www.aclweb.org/anthology/W00-1308.pdf)
- Information retrieval: "Information Retrieval using PMI-based Word Representations" by T. Mitamura and K. Kise (https://www.aclweb.org/anthology/C04-1100.pdf)
- Machine translation: "PMI-based Word Alignment for Statistical Machine Translation" by K. Toutanova and C. D. Manning (https://www.aclweb.org/anthology/W03-0419.pdf)



[redacted]

Kristina Toutanova
Dept of Computer Science
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
kristina@cs.stanford.edu

Christopher D. Manning
Depts of Computer Science and Linguistics
Gates Bldg 4A, 353 Serra Mall
Stanford, CA 94305–9040, USA
manning@cs.stanford.edu

Abstract

This paper presents results for a maximum-entropy-based part of speech tagger, which achieves superior performance principally by enriching the information sources used for tagging. In particular, we get improved results by incorporating these features:

This paper explores the notion that automatically built tagger performance can be further improved by expanding the knowledge sources available to the tagger. We pay special attention to unknown words, because the markedly lower accuracy on unknown word tagging means that this is an area where significant performance gains seem possible.

We adopt a maximum entropy approach

Core Limitations of Parametric LLMs

- Hallucinations
- Verifiability issues
- Knowledge cut-offs
- Learning failures

You

Who is Aditya Grover? Where does he teach?

‹ 6 / 6 ›

ChatGPT

Aditya Grover is a researcher and scientist known for his work in machine learning and artificial intelligence. As of my last update in January 2022, he was a Ph.D. candidate at Stanford University, working with Prof. Stefano Ermon in the Stanford Artificial Intelligence Lab (SAIL). However, there might have been updates since then. He has contributed to various areas of machine learning, including generative models and reinforcement learning. His research often focuses on developing algorithms that can learn and generalize from data efficiently.

Core Limitations of Parametric LLMs

- Hallucinations
- Verifiability issues
- Knowledge cut-offs
- Learning failures

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs – *close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Parametric LLMs – Training vs Test

The capital city of Ontario is **Toronto**

LM

Training time

The capital city of Ontario is _____

LM

Test time

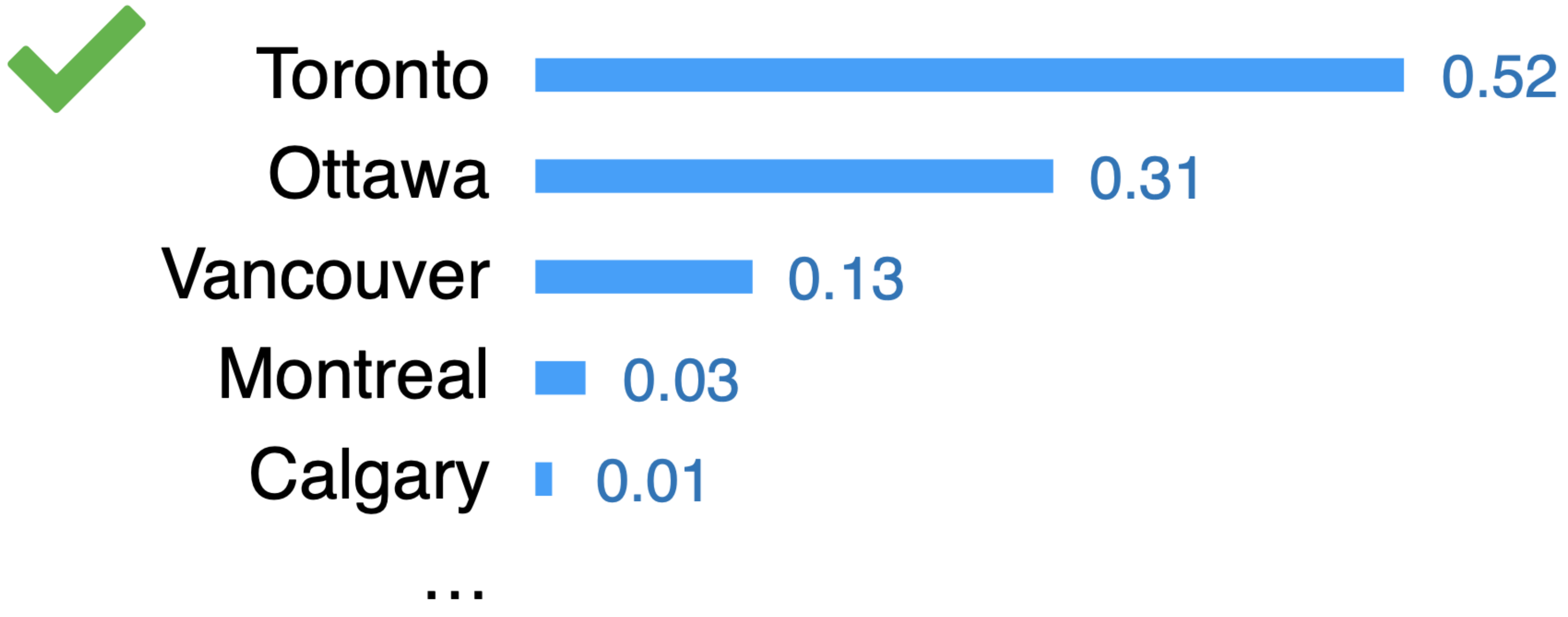
Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Parametric LLMs – Training vs Test

The capital city of Ontario is **Toronto**

LM

Training time



The capital city of Ontario is _____

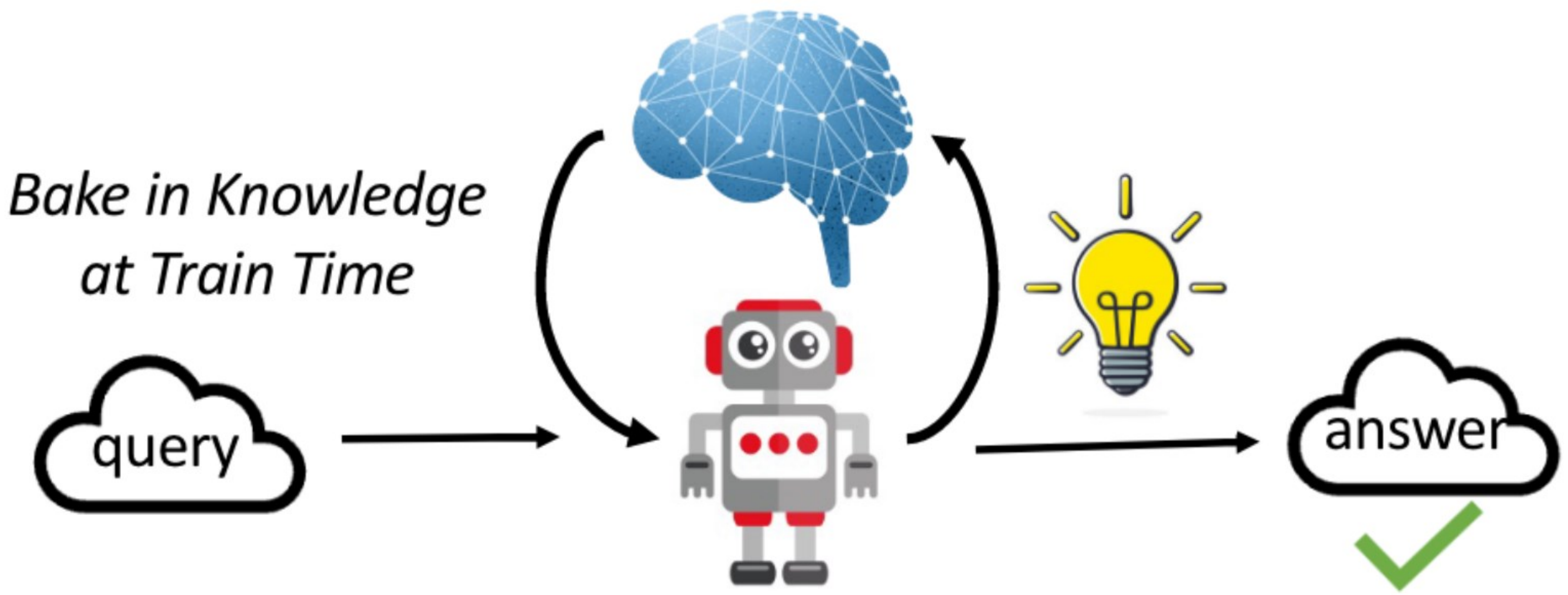
LM

Test time

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Closed Book vs Open Book Exams

Parametric LLMs



"Closed book"

Image source: http://arxiv.org/abs/2403.10131

Closed Book vs Open Book Exams

Parametric LLMs

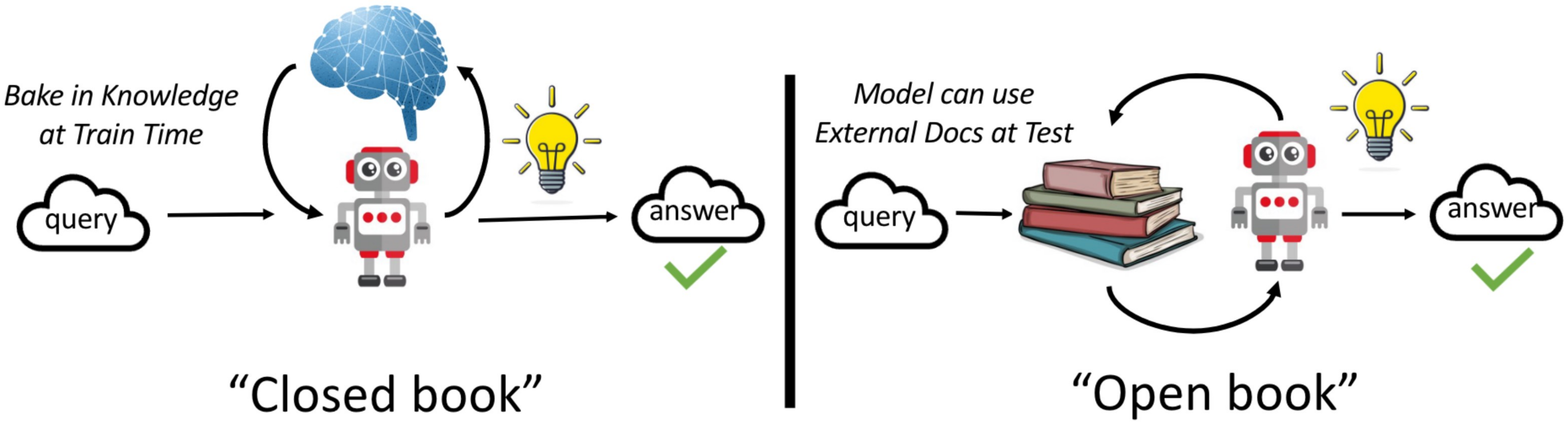


Image source: http://arxiv.org/abs/2403.10131

Closed Book vs Open Book Exams

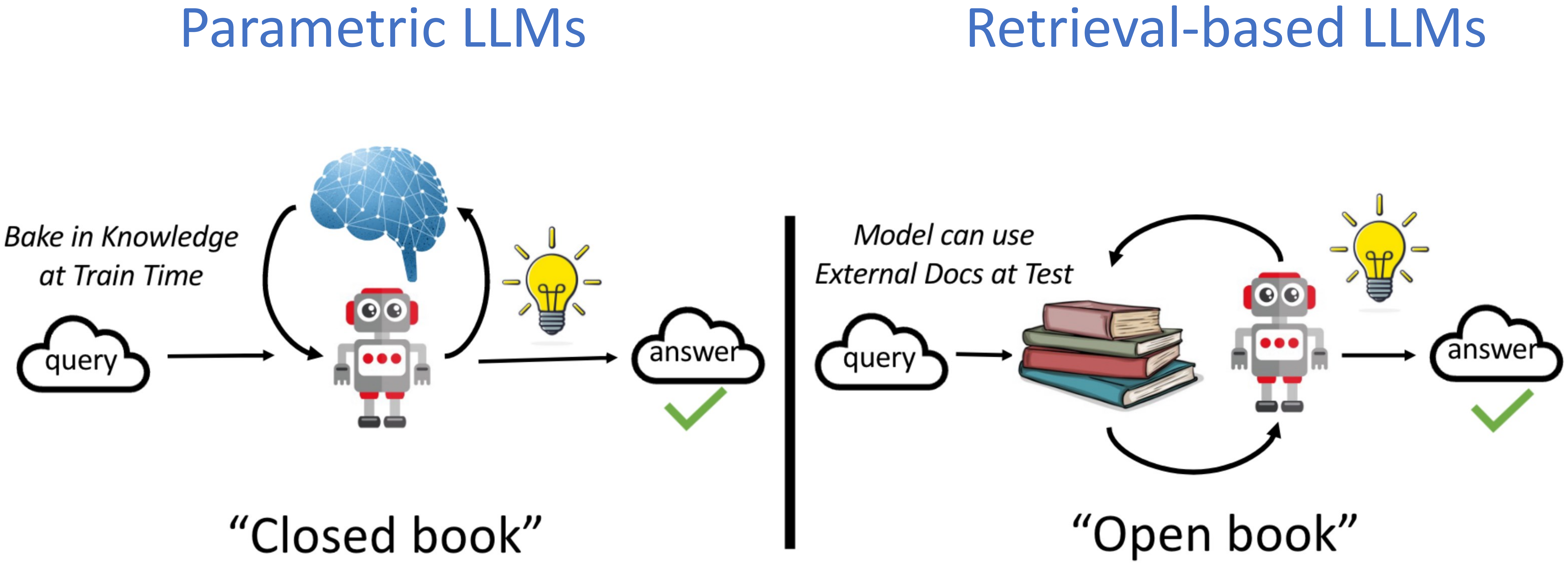
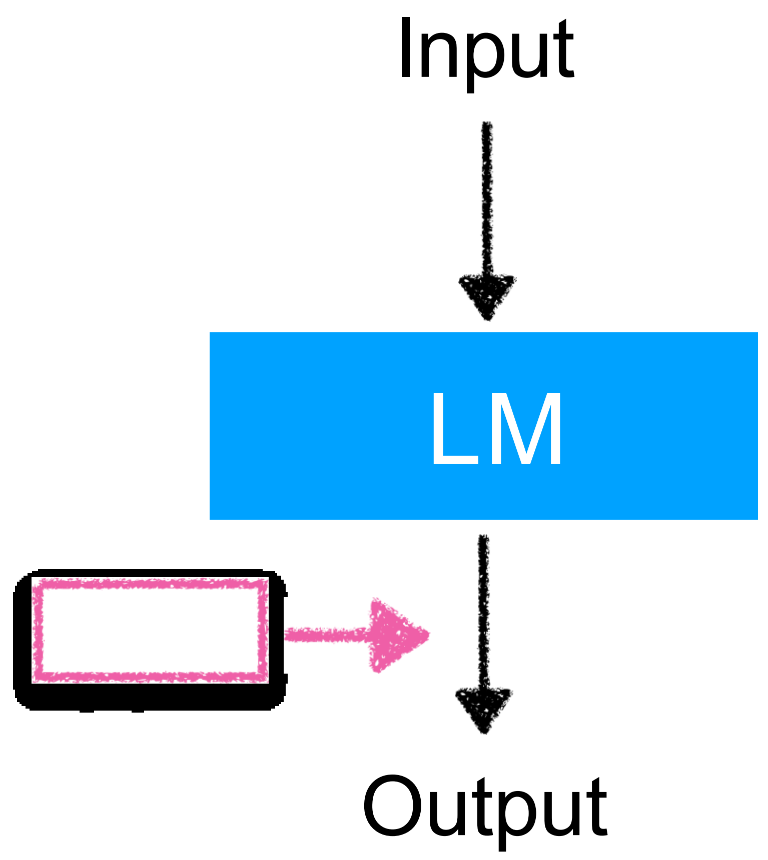


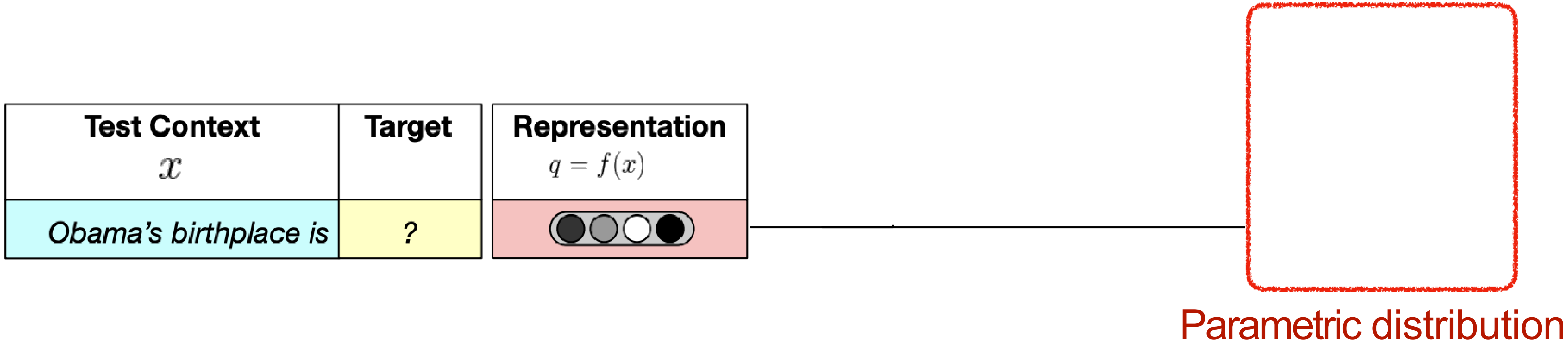
Image source: http://arxiv.org/abs/2403.10131

How to use the Book?

- Output interpolations - After solving the question yourself?



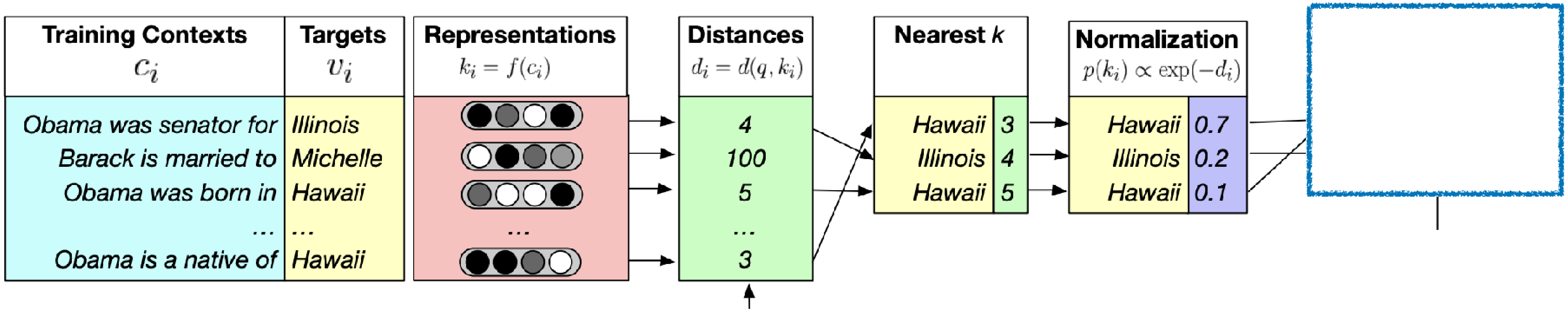
kNN-LM (Khandelwal et al. 2020)



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

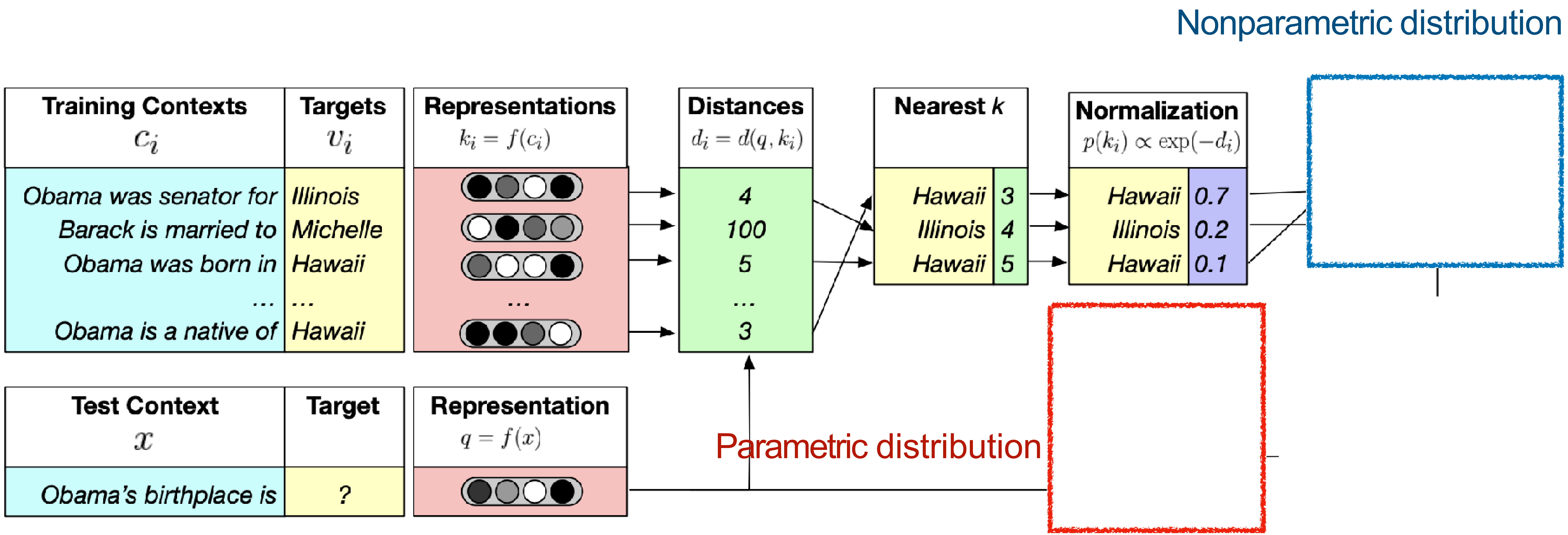
kNN-LM (Khandelwal et al. 2020)

Nonparametric distribution



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

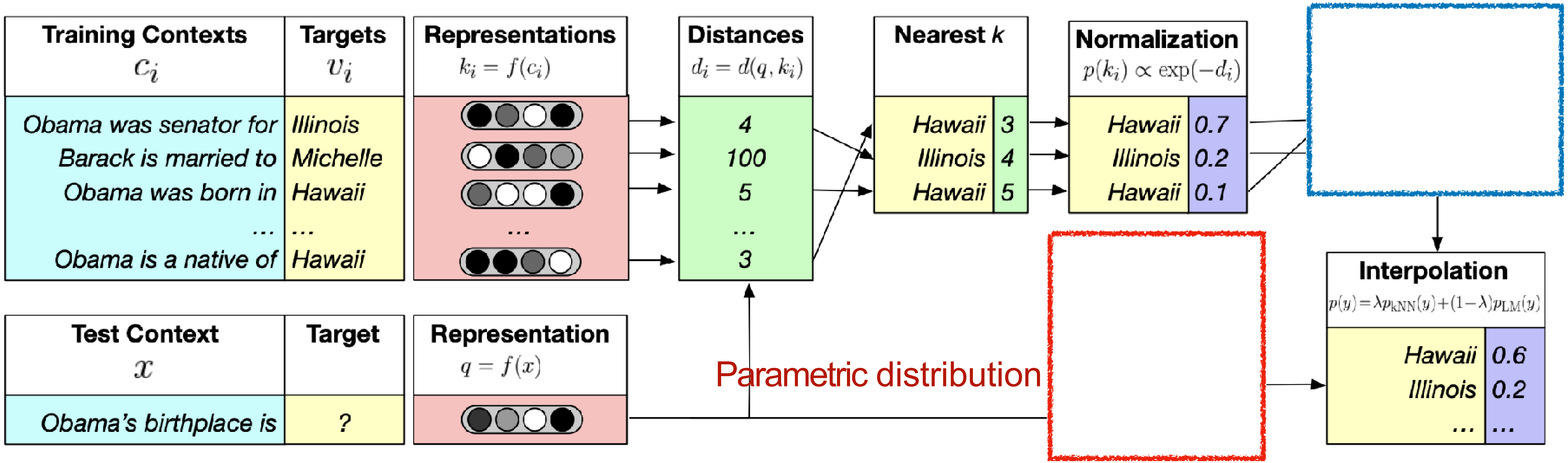
kNN-LM (Khandelwal et al. 2020)



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

kNN-LM (Khandelwal et al. 2020)

Nonparametric distribution



λ : hyperparameter

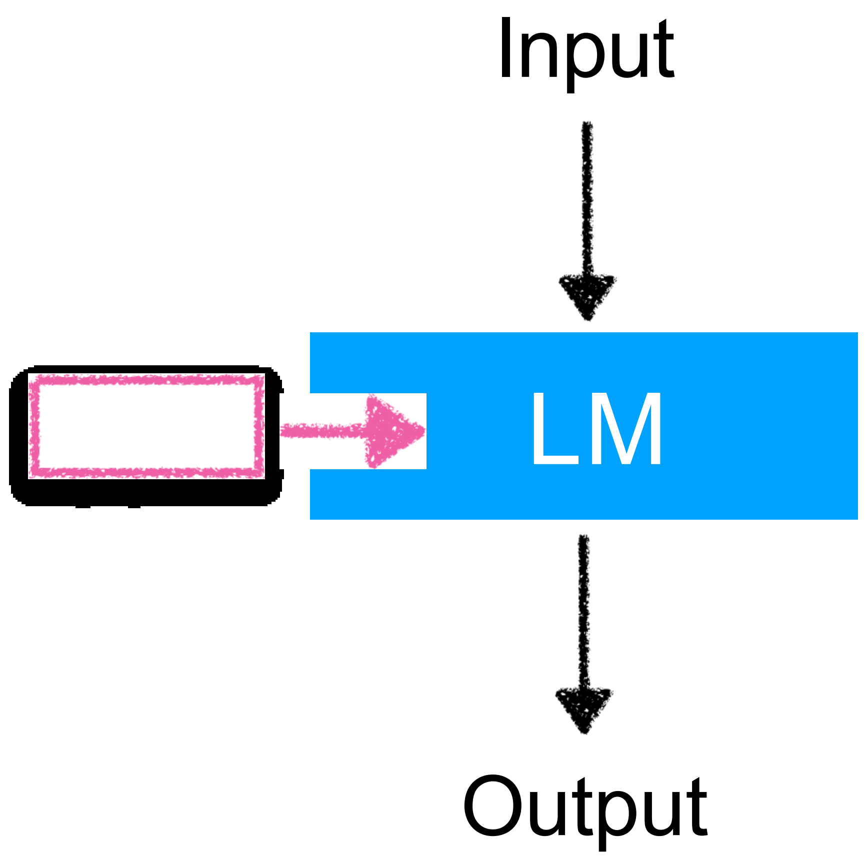
$$P_{k\text{NN-LM}}(y|x) = (1-\lambda)P_{\text{LM}}(y|x) + \lambda P_{k\text{NN}}(y|x)$$

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

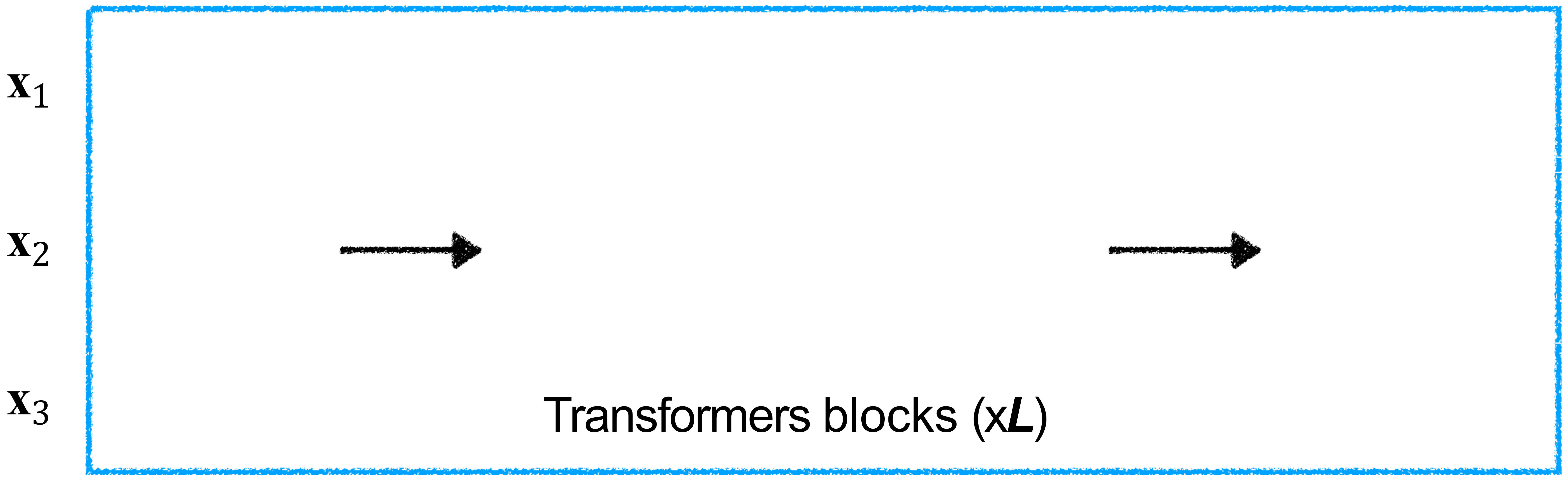
How to use the Book?

- Output interpolations - After solving the question yourself?

- Intermediate fusion – modify the LM architecture to be aware of the book?

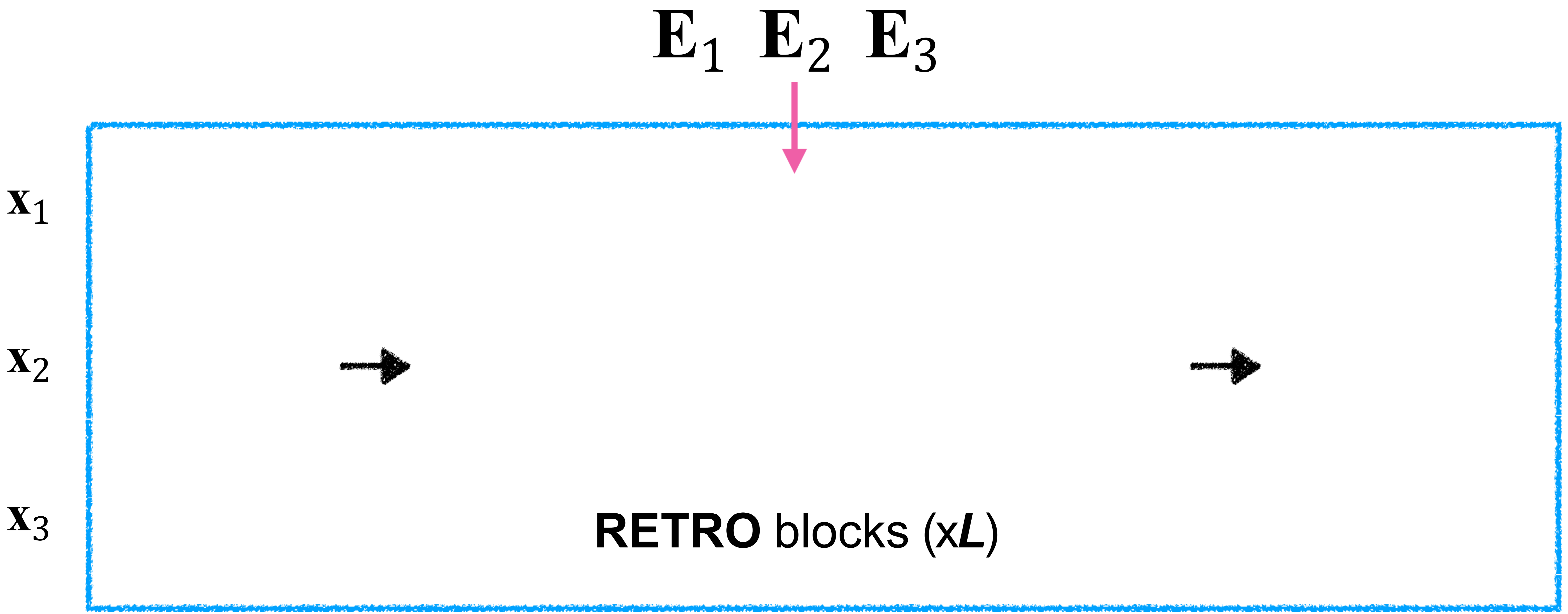


Regular decoder



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

Decoder in RETRO

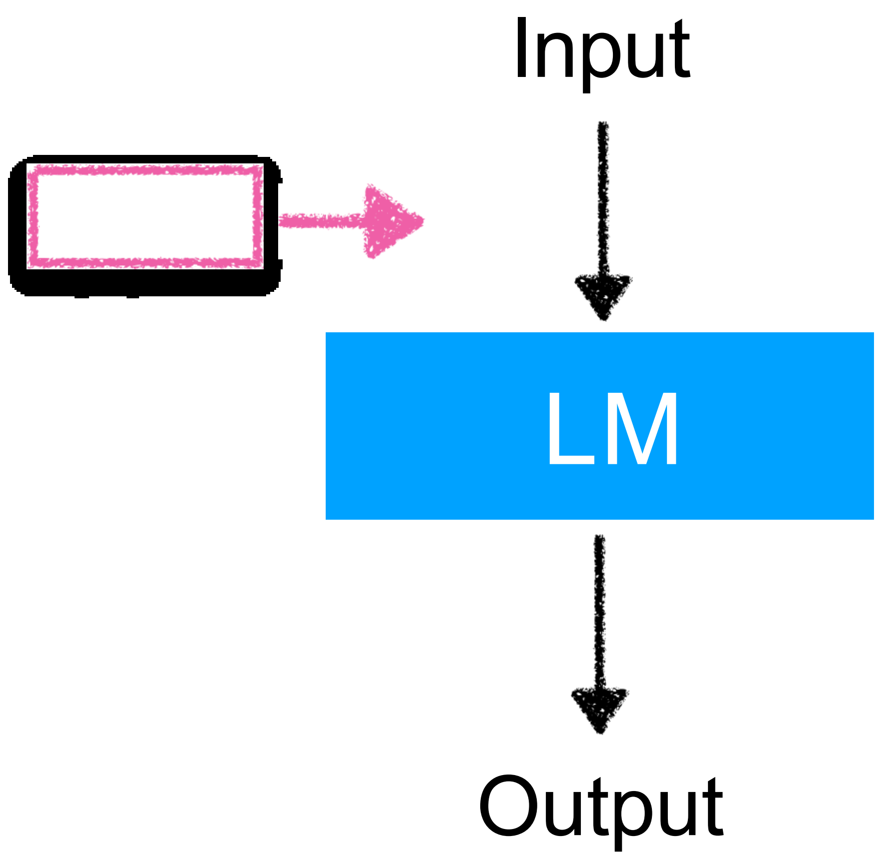


Chunked Cross Attention (CCA)

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

How to use the Book?

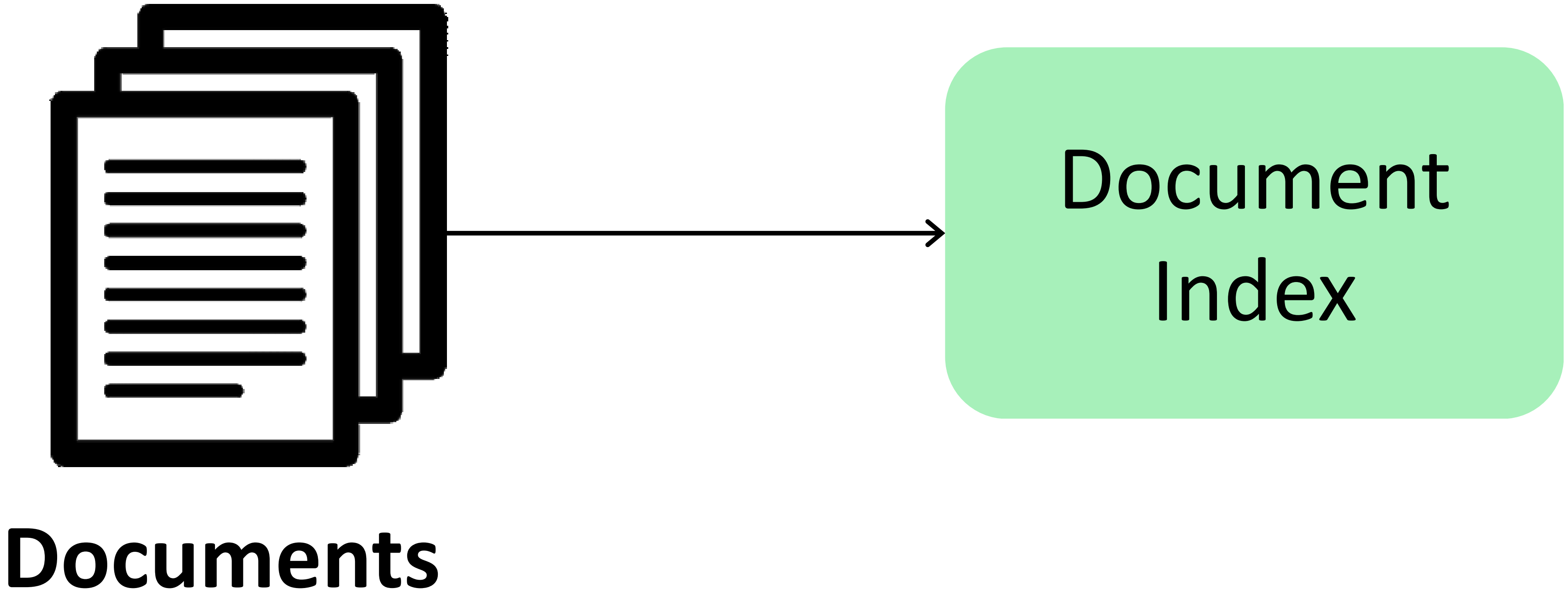
- Output interpolations - After solving the question yourself?
- Intermediate fusion – modify the LM architecture to be aware of the book?
- Input augmentation (RAG) - Before you start solving?



Outline

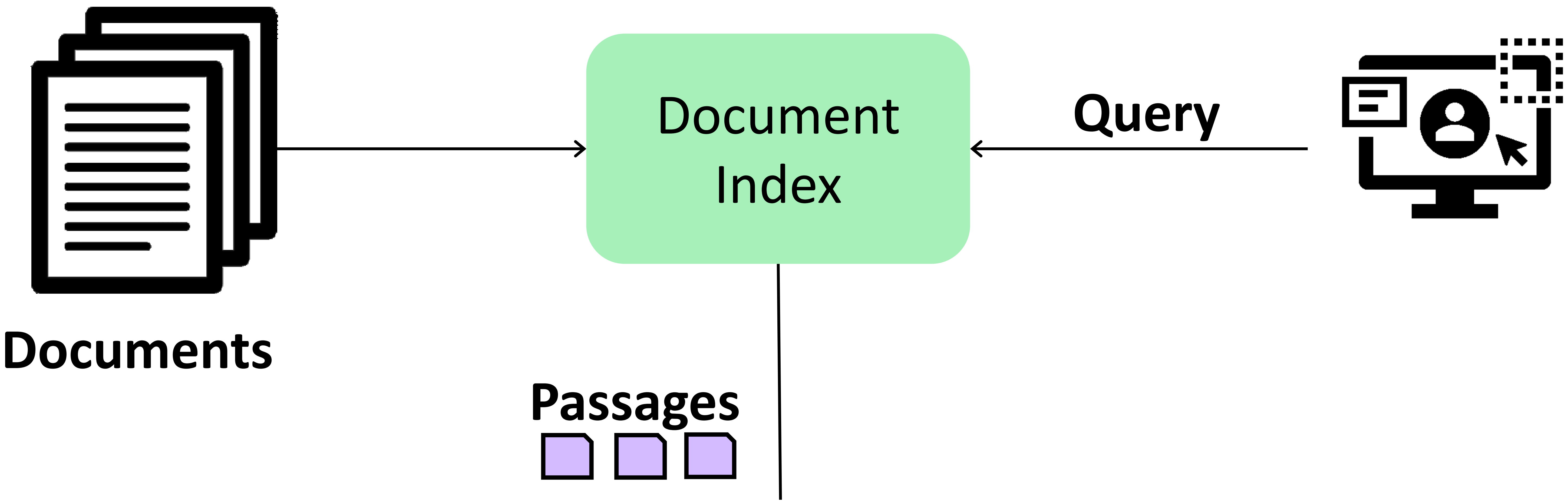
- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs *– close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Retrieval Based LLMs - Architecture

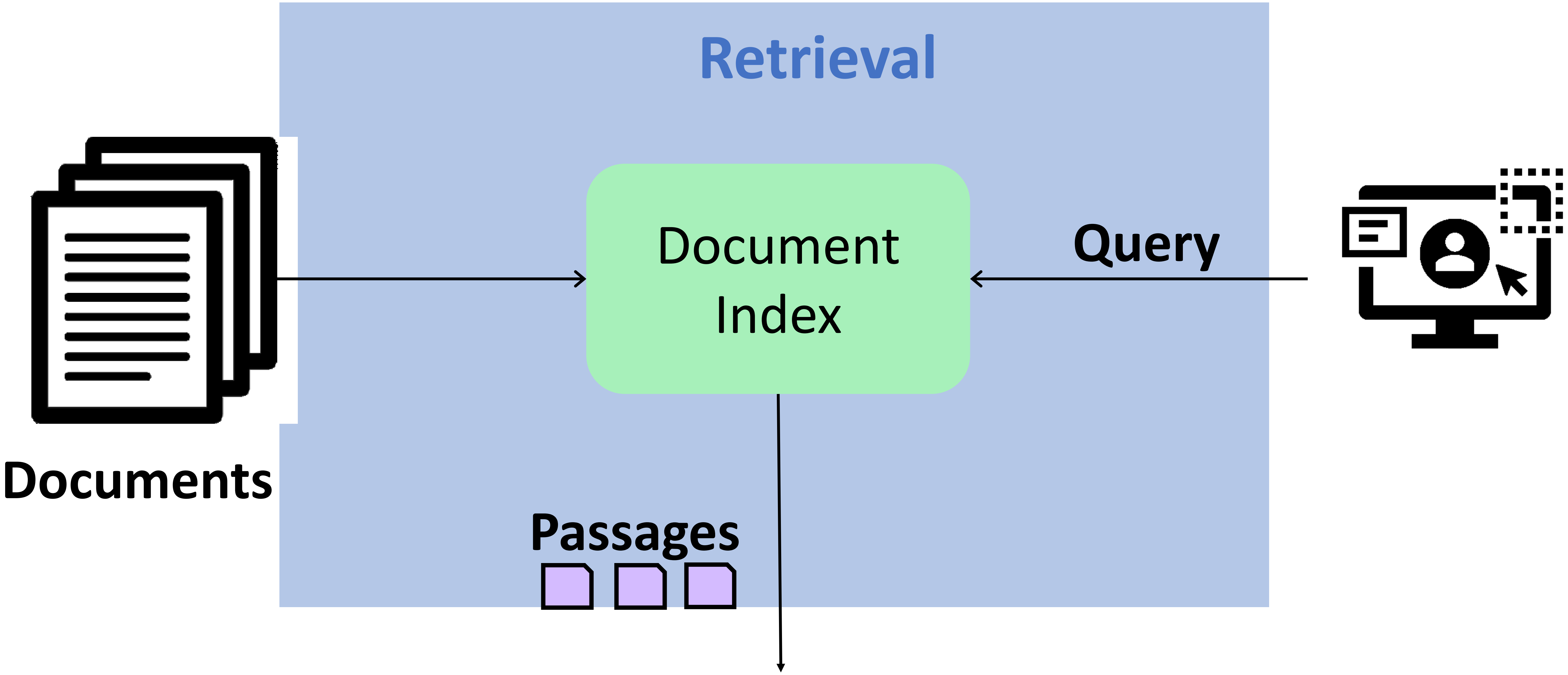


Retrieval Based LLMs - Architecture

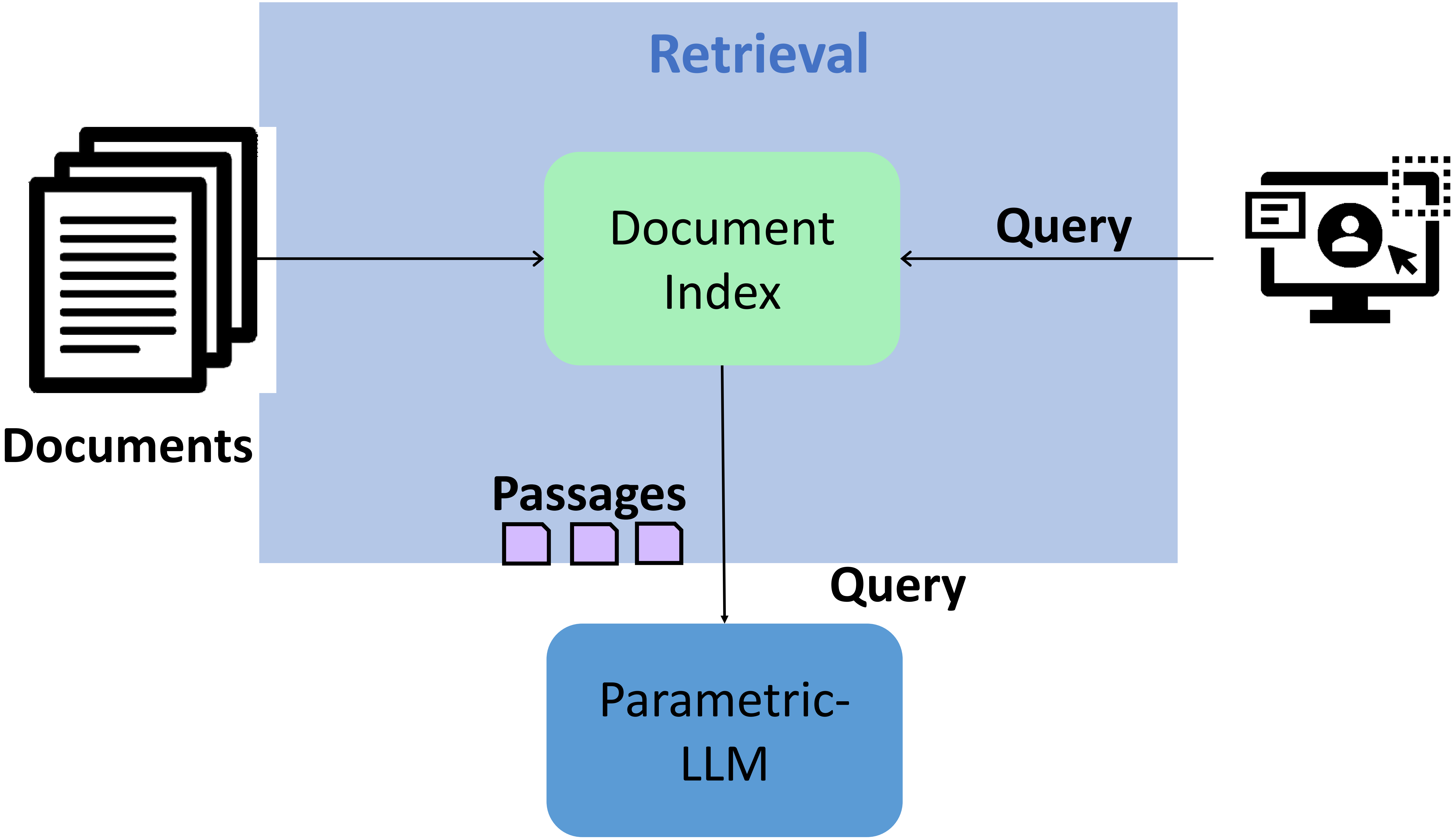
Retrieval Based LLMs - Architecture



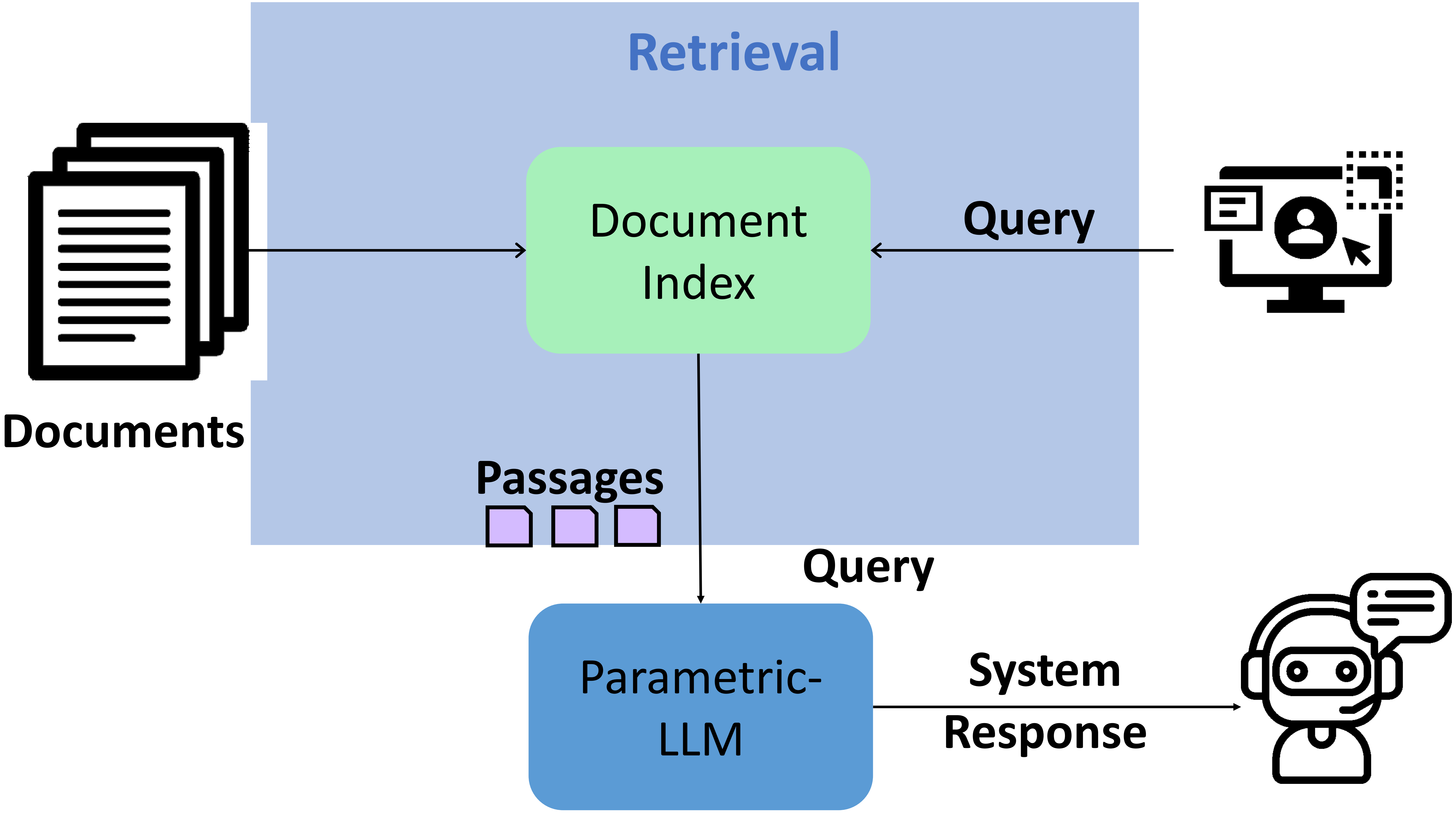
Retrieval Based LLMs - Architecture



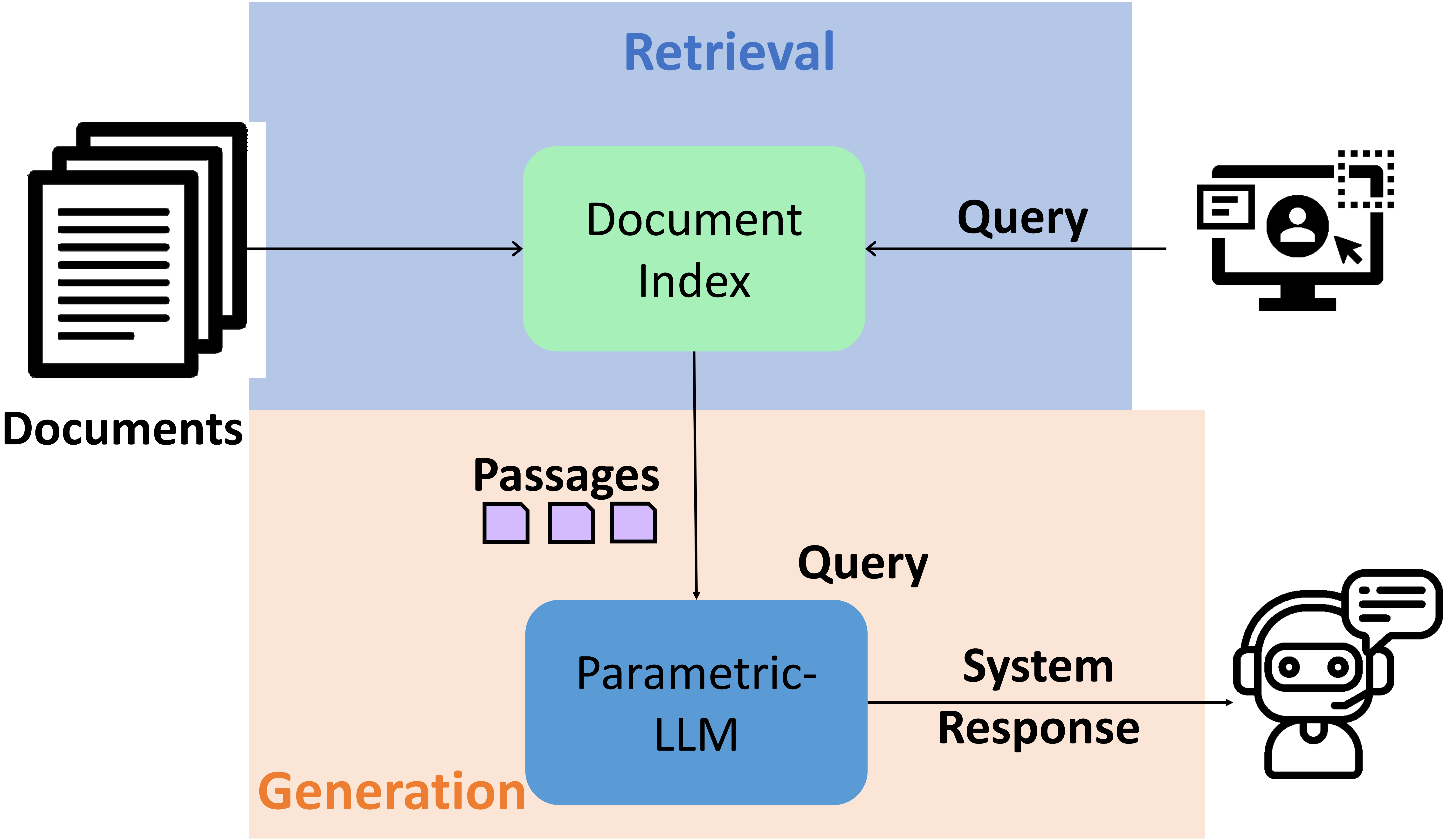
Retrieval Based LLMs - Architecture



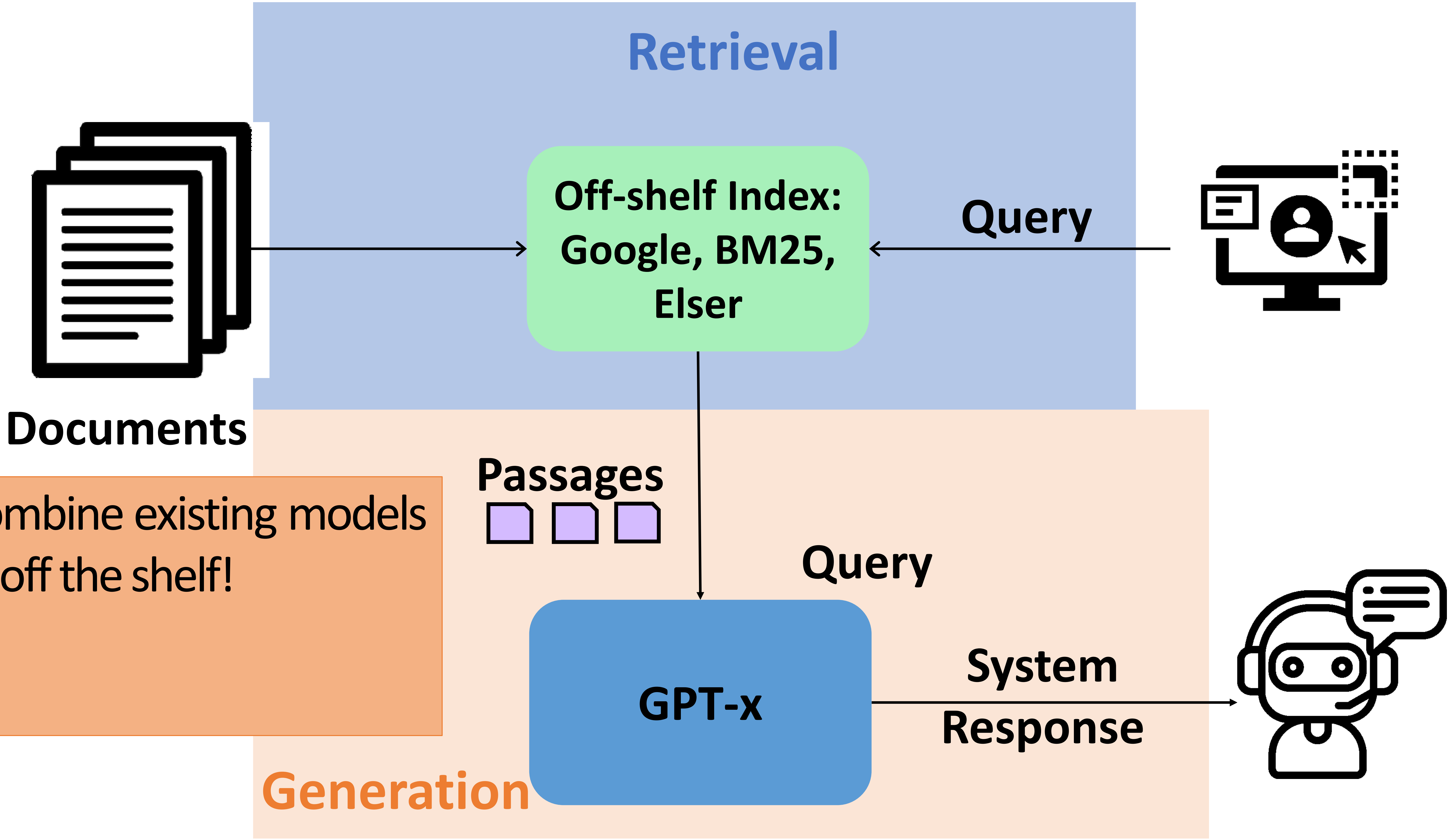
Retrieval Based LLMs - Architecture



Retrieval Based LLMs - Architecture

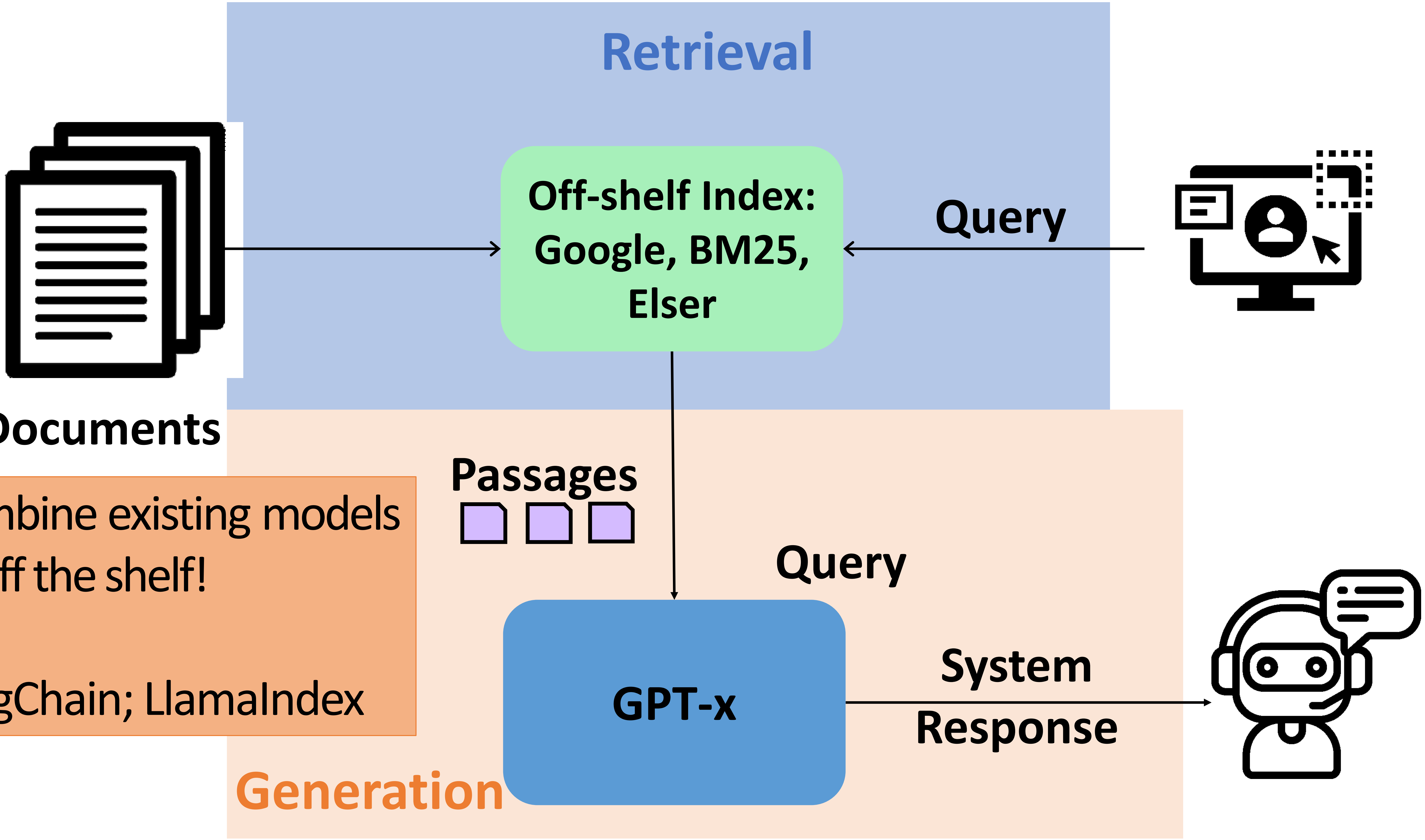


Retrieval Based LLMs - Architecture



- ✓ Simply combine existing models available off the shelf!

Retrieval Based LLMs - Architecture



- ✓ Simply combine existing models available off the shelf!
- ✓ Tools: LangChain; LlamaIndex

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs *– close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Retrieval Methods

- Sparse retrieval
- Document-level dense retrieval
- Token-level dense retrieval
- Cross-encoder reranking
- Black-box retrieval (just ask Google/Bing)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)

$$q = \text{what is nlp}$$

$$\begin{array}{c} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \ldots \end{array} \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \ldots \end{pmatrix}$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)

q=what is nlp

$d_1 =$ what is life ? candy is life !

$$\begin{array}{c} \text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \dots \end{array} \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \dots \end{pmatrix} \quad \begin{pmatrix} 0.25 \\ 0.125 \\ 0 \\ 0.25 \\ 0 \\ \dots \end{pmatrix}$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)

q=what is nlp

$d_1 =$ what is life ? candy is life !

$d_2 =$ nlp is an acronym for natural language processing

$$
\begin{array}{c}
\text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \ldots
\end{array}
\begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0.25 \\ 0.125 \\ 0 \\ 0.25 \\ 0 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0.125 \\ 0 \\ \ldots \end{pmatrix}
$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)

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	0.33	0.25	0	0
candy	0	0.125	0	0
nlp	0.33	0	0.125	0.125
is	0.33	0.25	0.125	0
language	0	0	0	0
…	…	…	…	…

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Sparse Retrieval

- Express the query and document as a sparse word frequency vector (usually normalized by length)

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

$$
\begin{array}{c}
\text{what} \\ \text{candy} \\ \text{nlp} \\ \text{is} \\ \text{language} \\ \ldots
\end{array}
\begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0.25 \\ 0.125 \\ 0 \\ 0.25 \\ 0 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0.125 \\ 0 \\ \ldots \end{pmatrix}
\quad
\begin{pmatrix} 0 \\ 0 \\ 0.125 \\ 0 \\ 0 \\ \ldots \end{pmatrix}
$$

$\boxed{q^*d_1 = 0.165}$ $\quad q^*d_2 = 0.0825 \quad q^*d_3 = 0.0413$

- Find the document with the highest inner-product or cosine similarity in the document collection

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)
- Term frequency - in-document frequency (TF-IDF)

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

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)
- Term frequency - in-document frequency (TF-IDF)

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

- BM25: TF term similar to smoothed count-based LMS

$$\text{BM-25}(t,d) = \quad .$$

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Term Weighting (see Manning et al. 2009)

- Some terms are more important than others; Low-frequency words (*NLP, Candy*) are often more important than (*the, a, for, then, them…*)
- Term frequency - in-document frequency (TF-IDF)

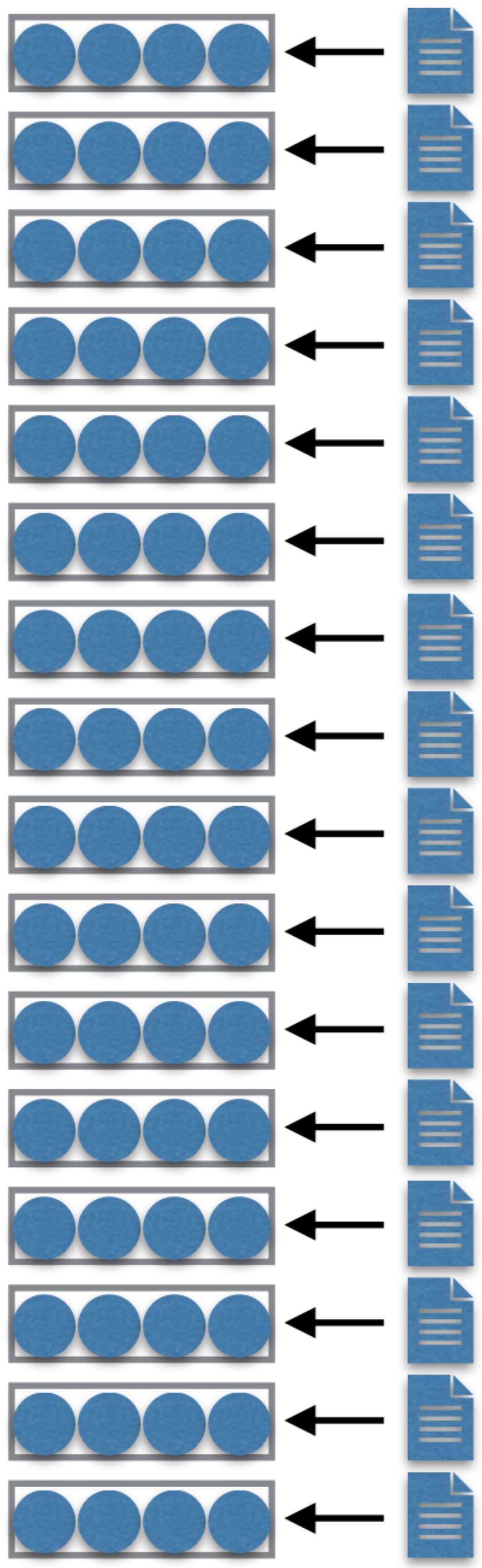
$$\text{TF-IDF}(t,d) = \quad \times \quad$$

- BM25: TF term similar to smoothed count-based LMS

$$\text{BM-25}(t,d) = \quad \cdot \quad$$

Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓Out-of-the-box embeddings. E.g. BERT
  - ✓Learned embeddings (covered later)

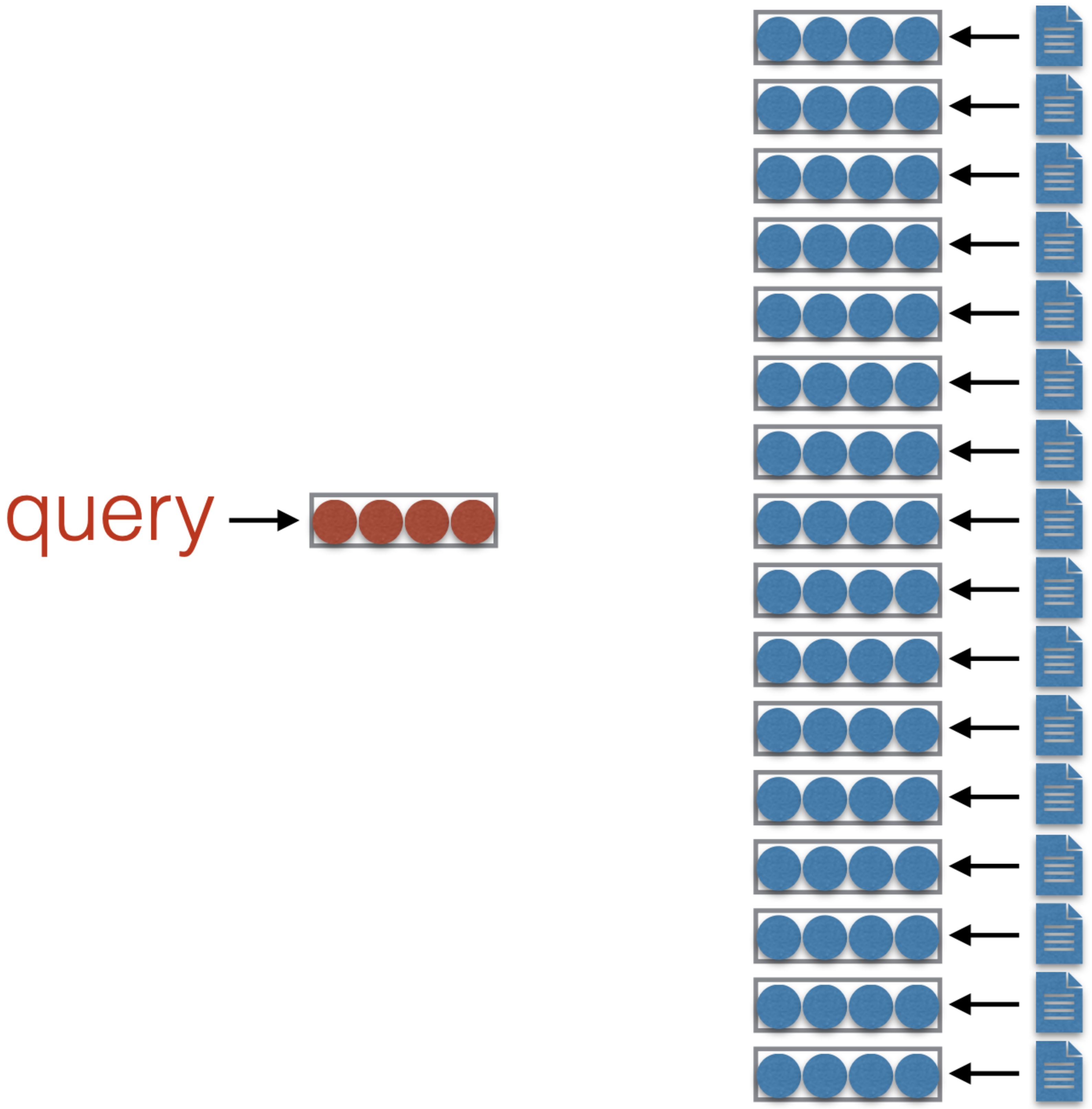


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓Out-of-the-box embeddings. E.g. BERT
  - ✓Learned embeddings (covered later)

- At test time:
  - Encode Query

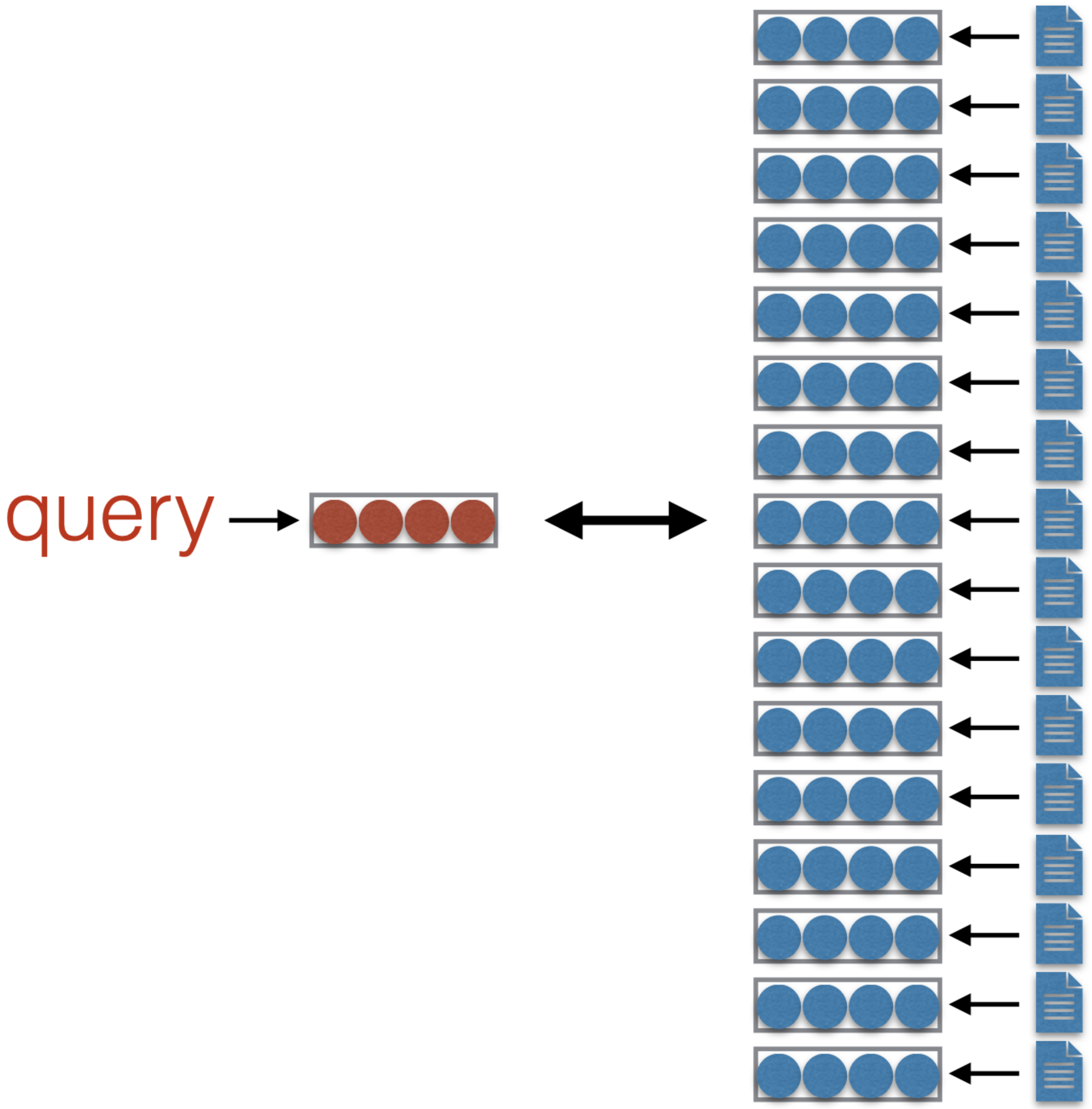


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Dense Embeddings

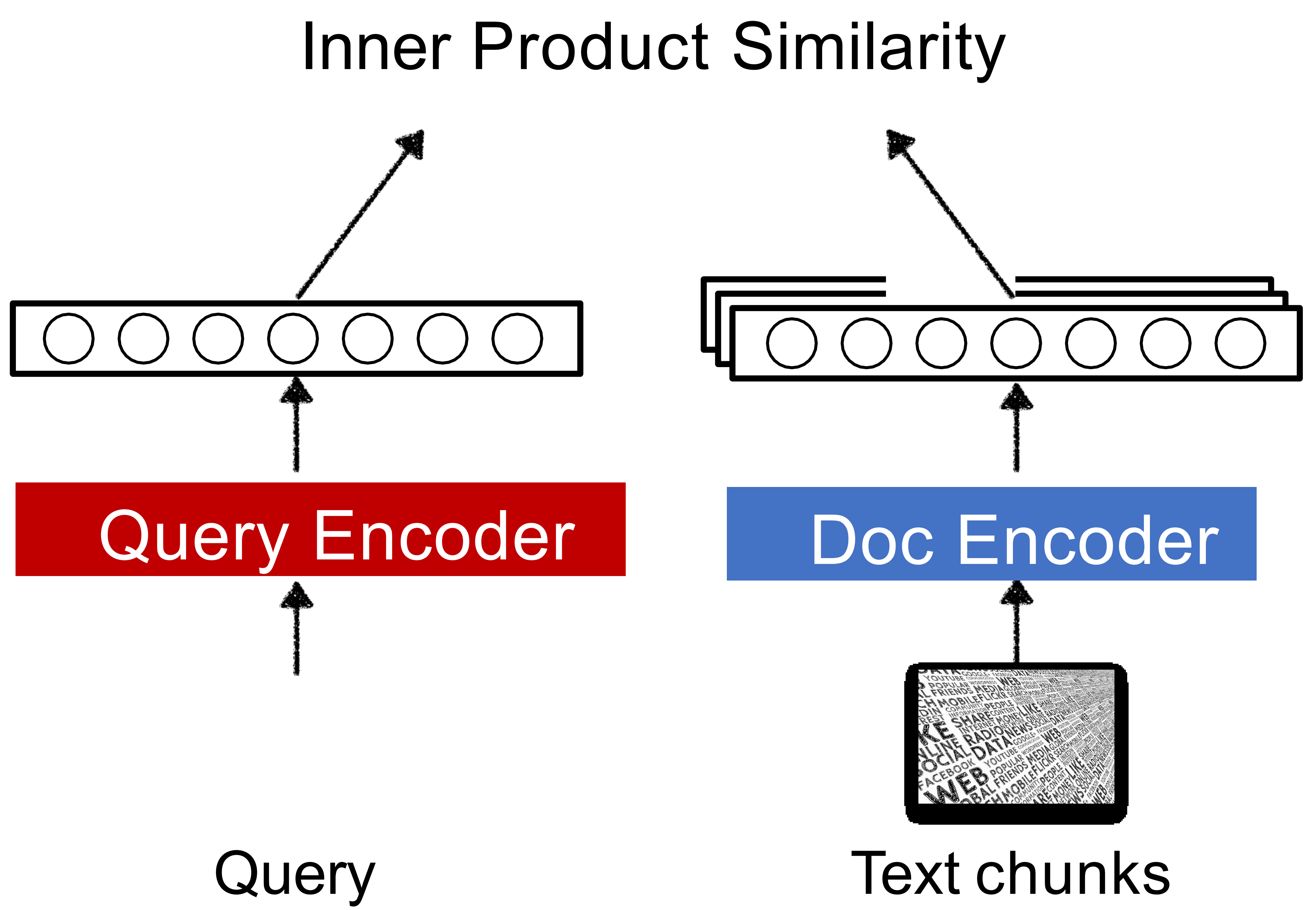
- Encode all documents using a LM and index them (one time task). Can use:
  - ✓Out-of-the-box embeddings. E.g. BERT
  - ✓Learned embeddings (covered later)

- At test time:
  - Encode Query
  - Find similar documents



Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

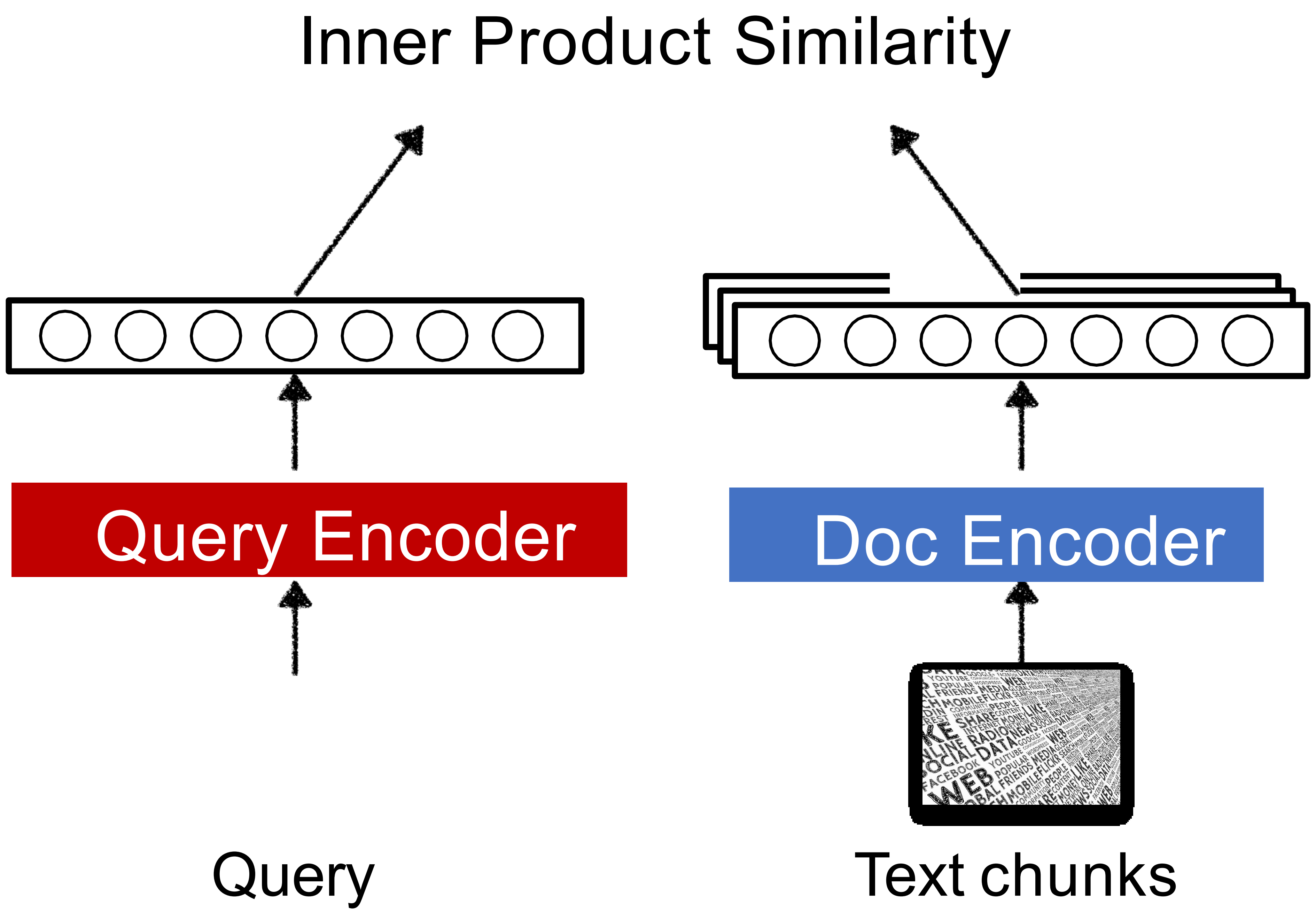
Training Dense Embeddings



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

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

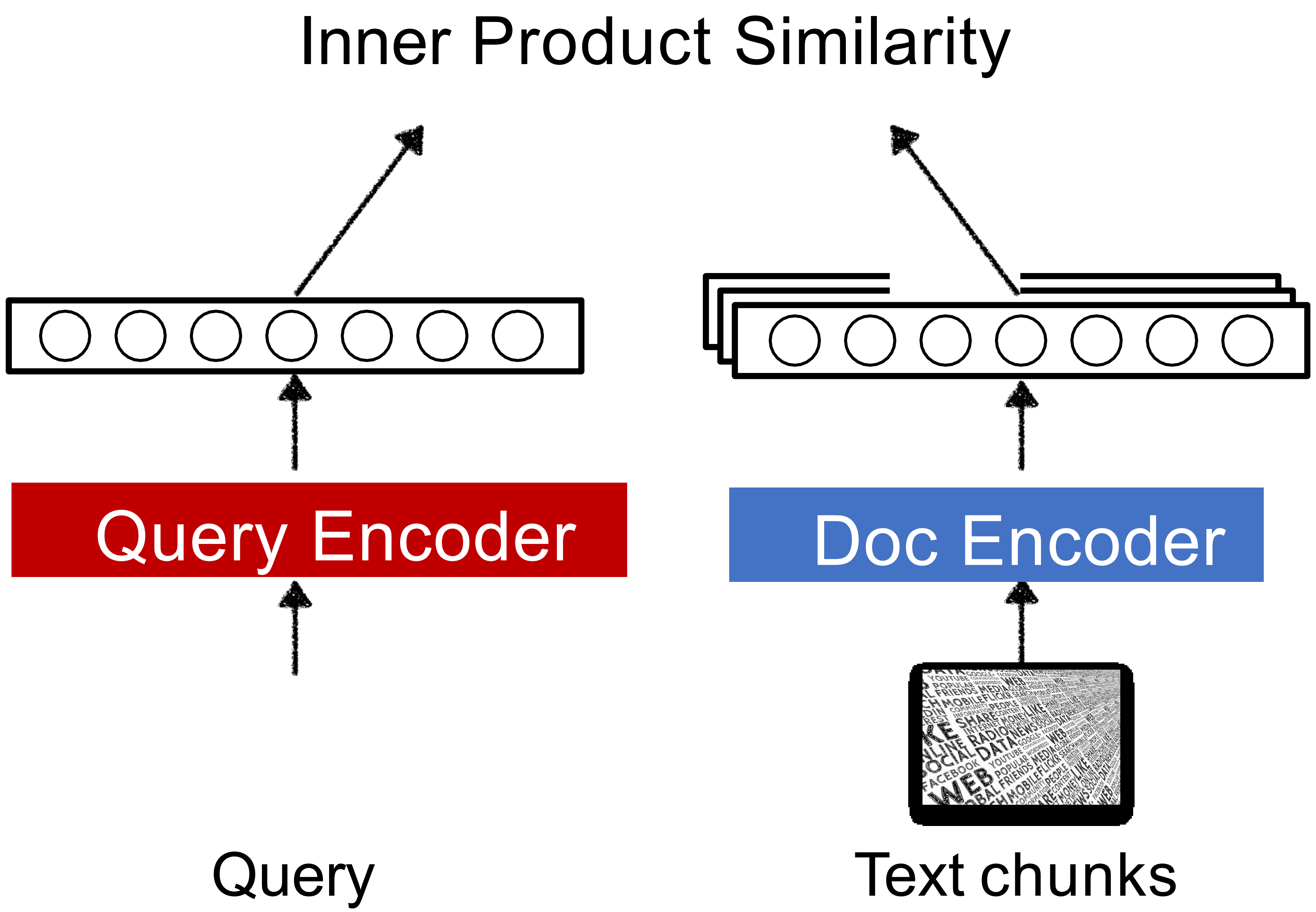
Training Dense Embeddings



$$L(q, p^{+}, p_{1}^{-}, p_{2}^{-}, \dots, p_{n}^{-}) = -\log \frac{\exp(\mathrm{sim}(q, p^{+}))}{\exp(\mathrm{sim}(q, p^{+})) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_{j}^{-}))}$$

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

Training Dense Embeddings



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

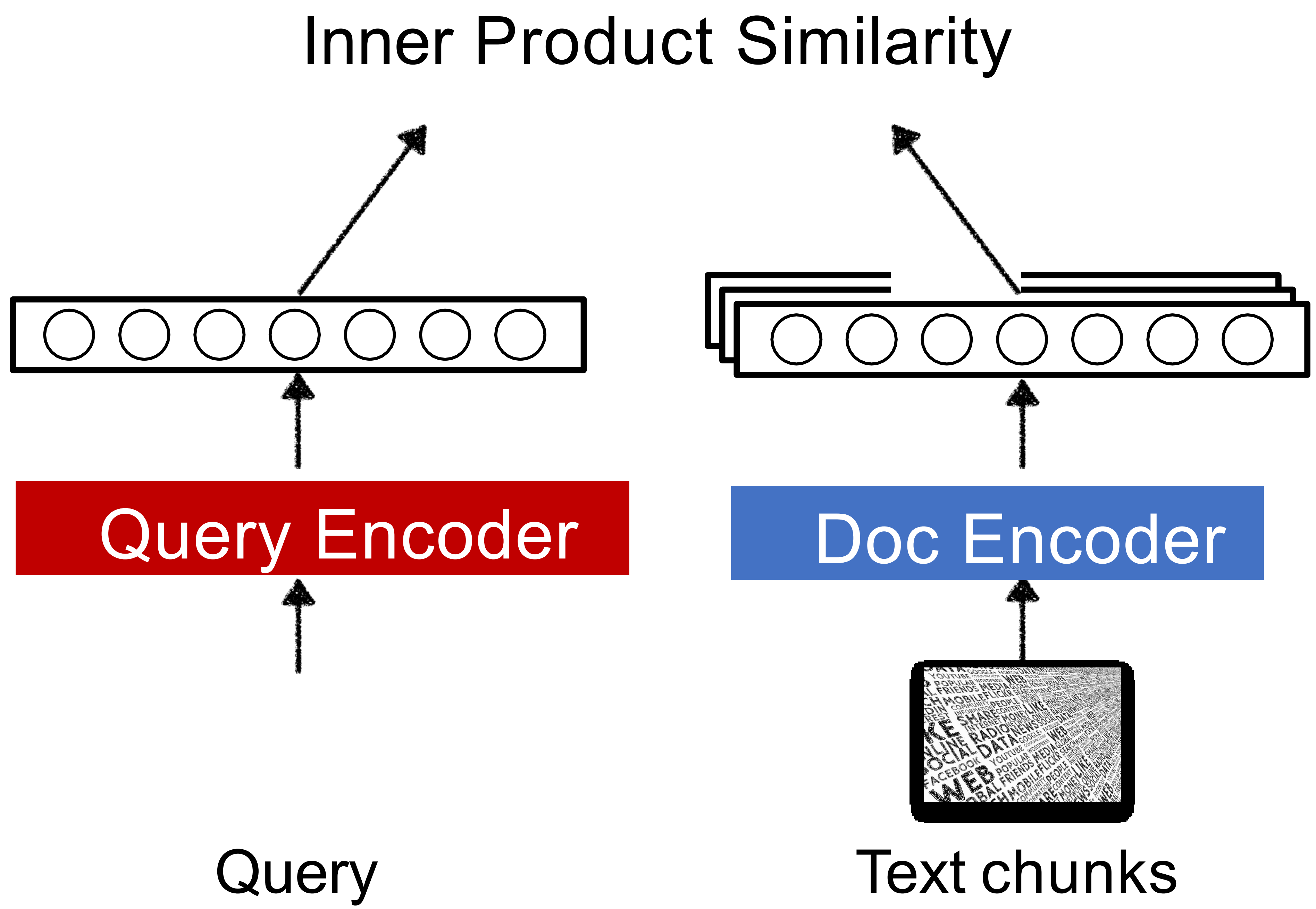
Positive passage

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

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

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings



Negative passages

Too expensive to consider all negatives!

$L(q$

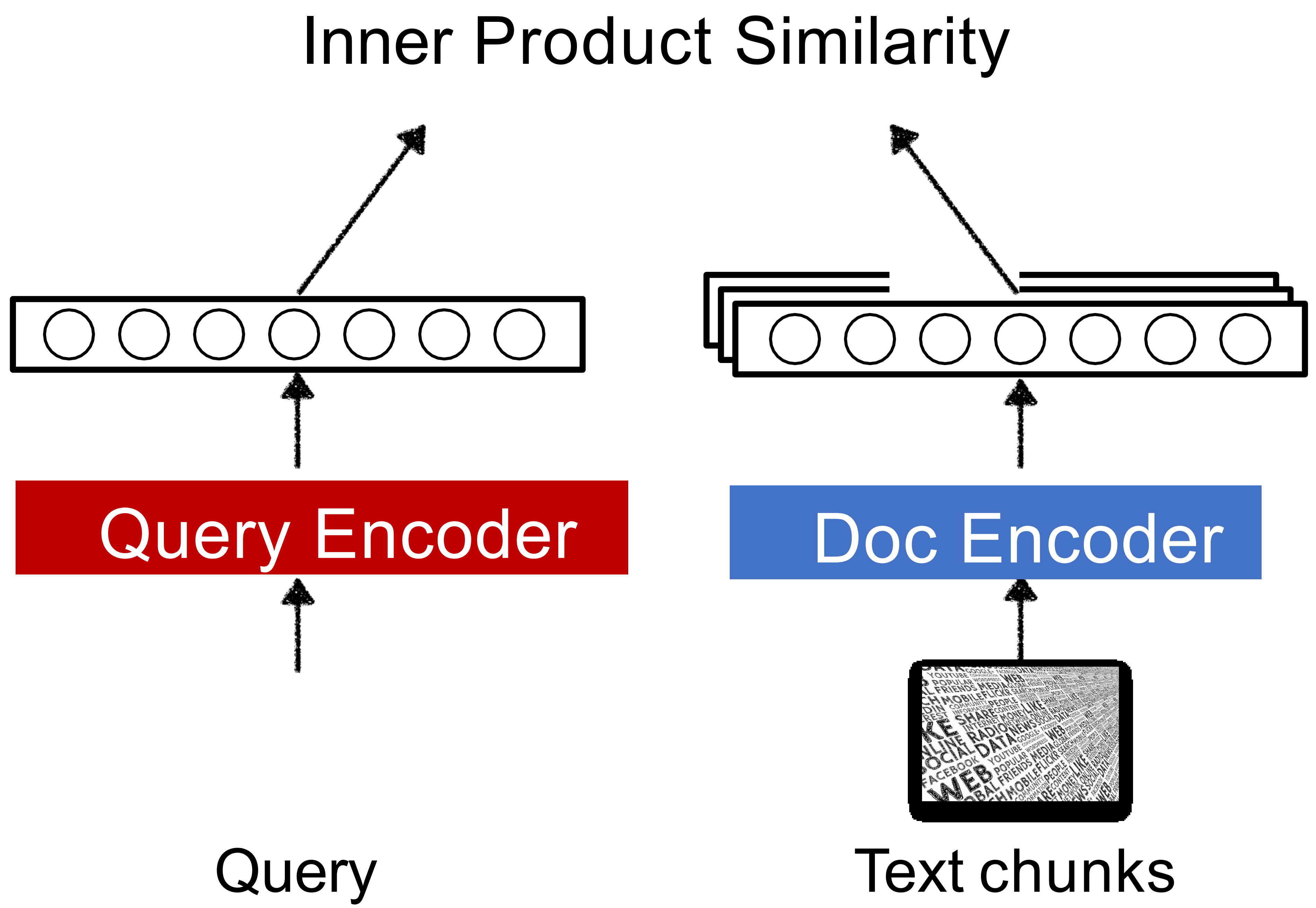
Positive passage

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

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

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings



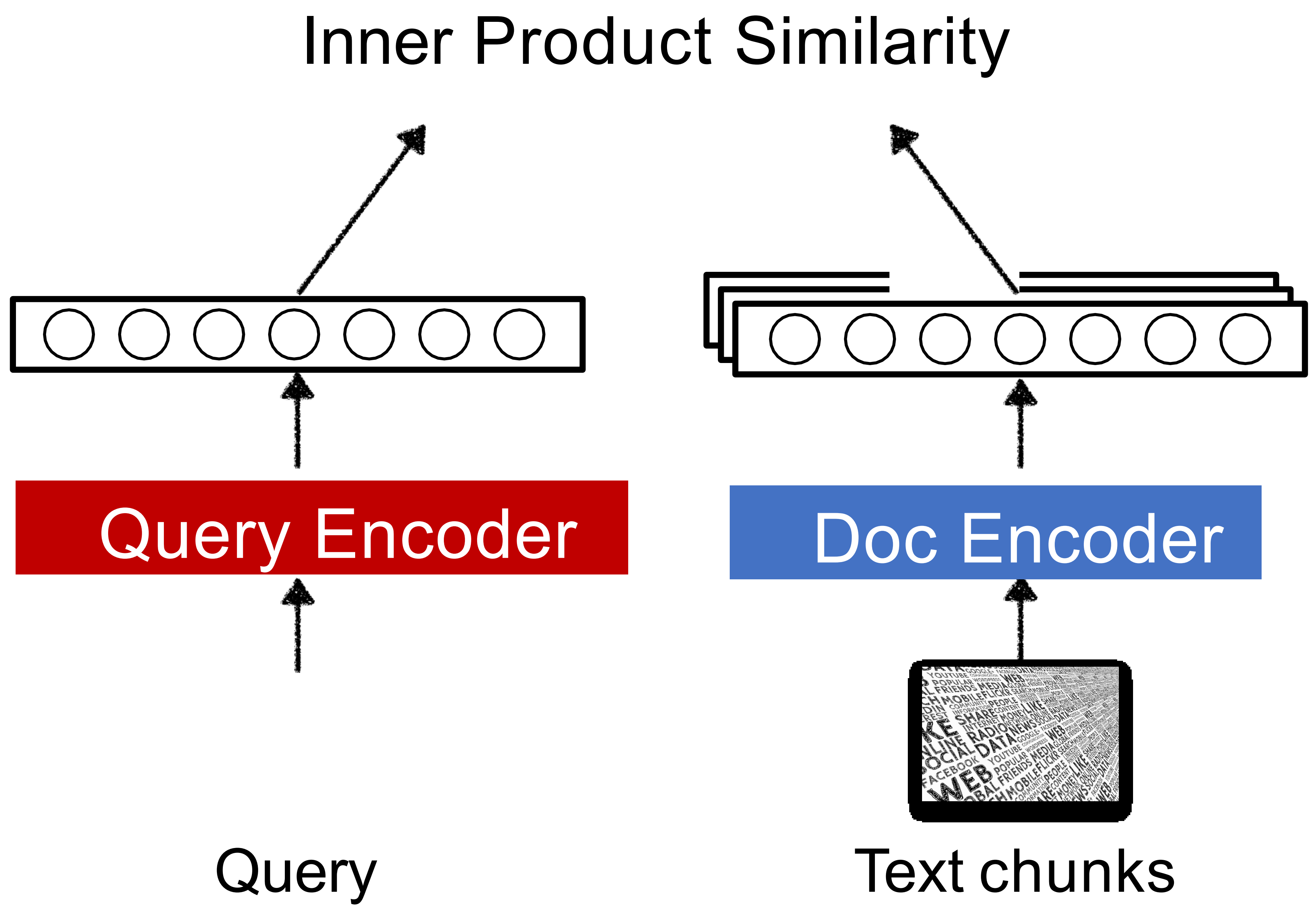
$$L(q, p^+, p_1^-, p_2^-, \ldots, p_n^-) = -\log \frac{\exp(\mathrm{sim}(q, p^+))}{\exp(\mathrm{sim}(q, p^+)) + \sum_{j=1}^{n} \exp(\mathrm{sim}(q, p_j^-))}$$

Contrastive learning

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

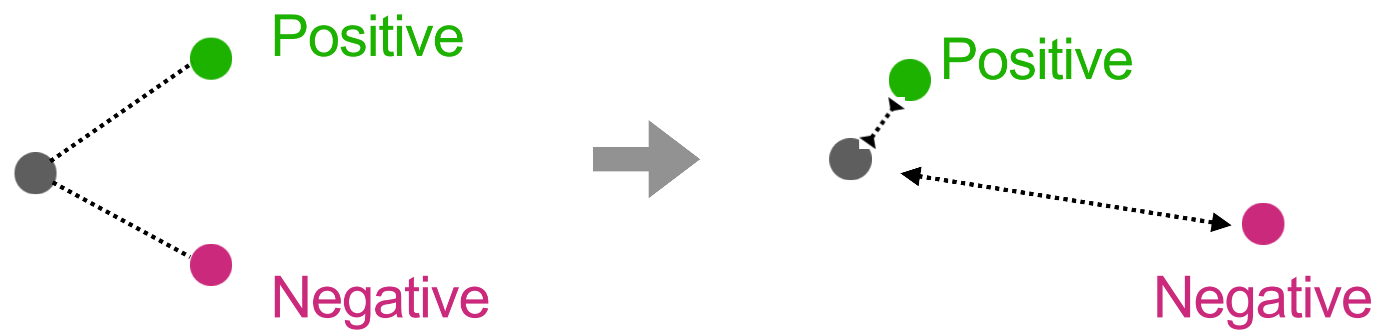
Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings



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

Contrastive learning



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

Slide source: https://drive.google.com/file/d/1YUpp7L1SCK6jgdfFObsqHKXrq6HC-TLp/view

Training Dense Embeddings

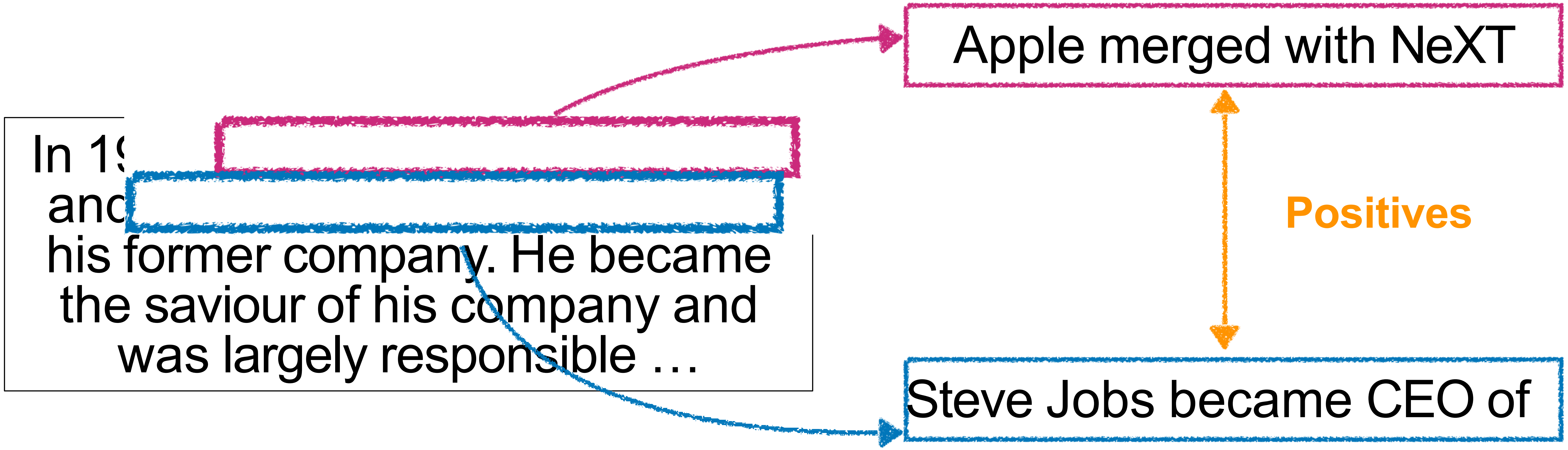
- Select positive and negative documents, train using a contrastive loss

- **DPR** (Karpukhin et al. 2020): learn encoders based on a BM25 hard negatives and in-batch negatives.

- **Contriever** (Izacard et al. 2022): contrastive learning using two random spans as positive pairs - **Unsupervised** dense retrieval model.

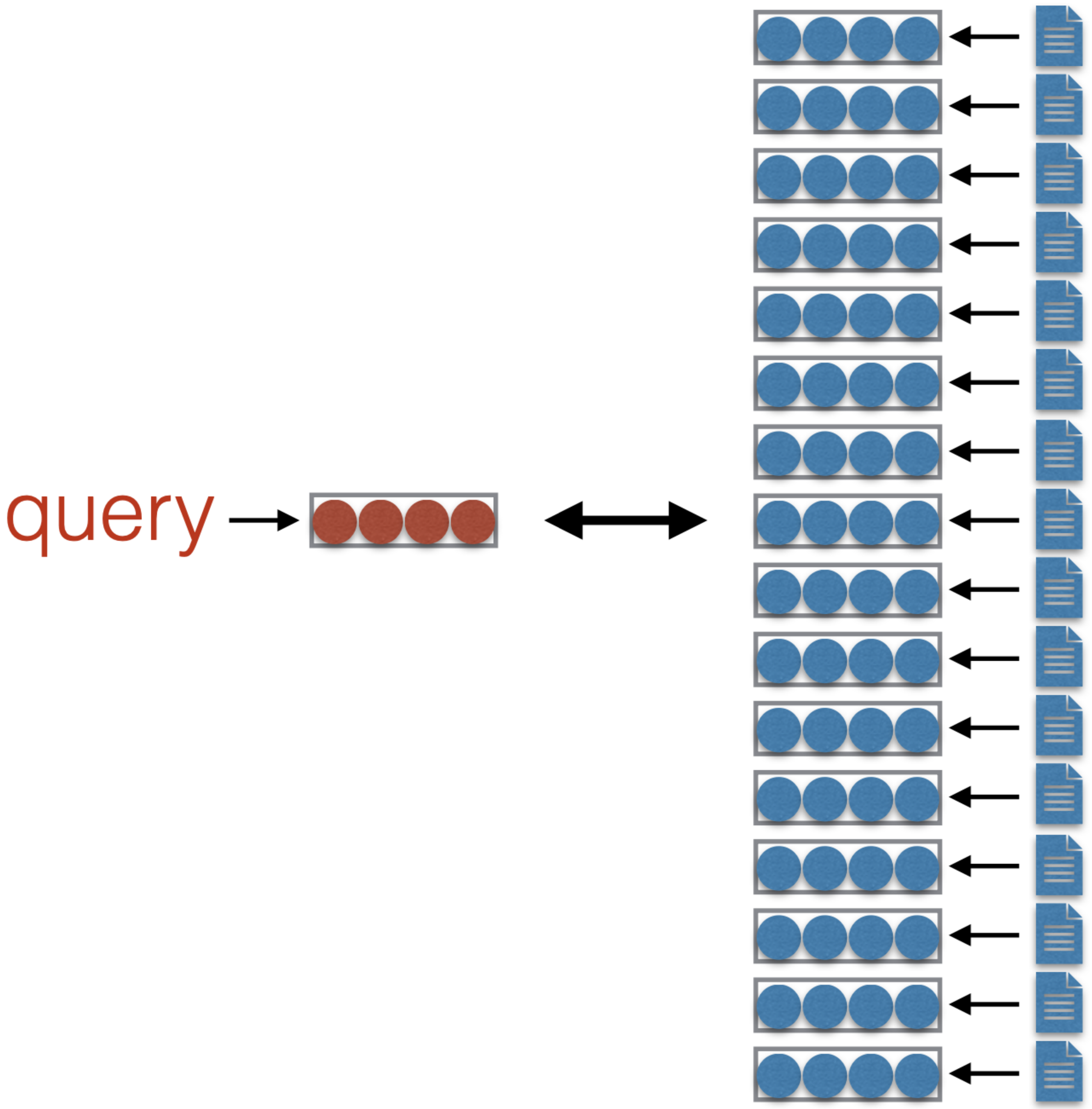
Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Independent Cropping in Contriever (Izacard et al. 2022):



Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓Out-of-the-box embeddings. E.g. BERT
  - ✓Learned embeddings (covered later)
- At test time:
  - Encode Query
  - Use Nearest Neighbor Search to find similar documents

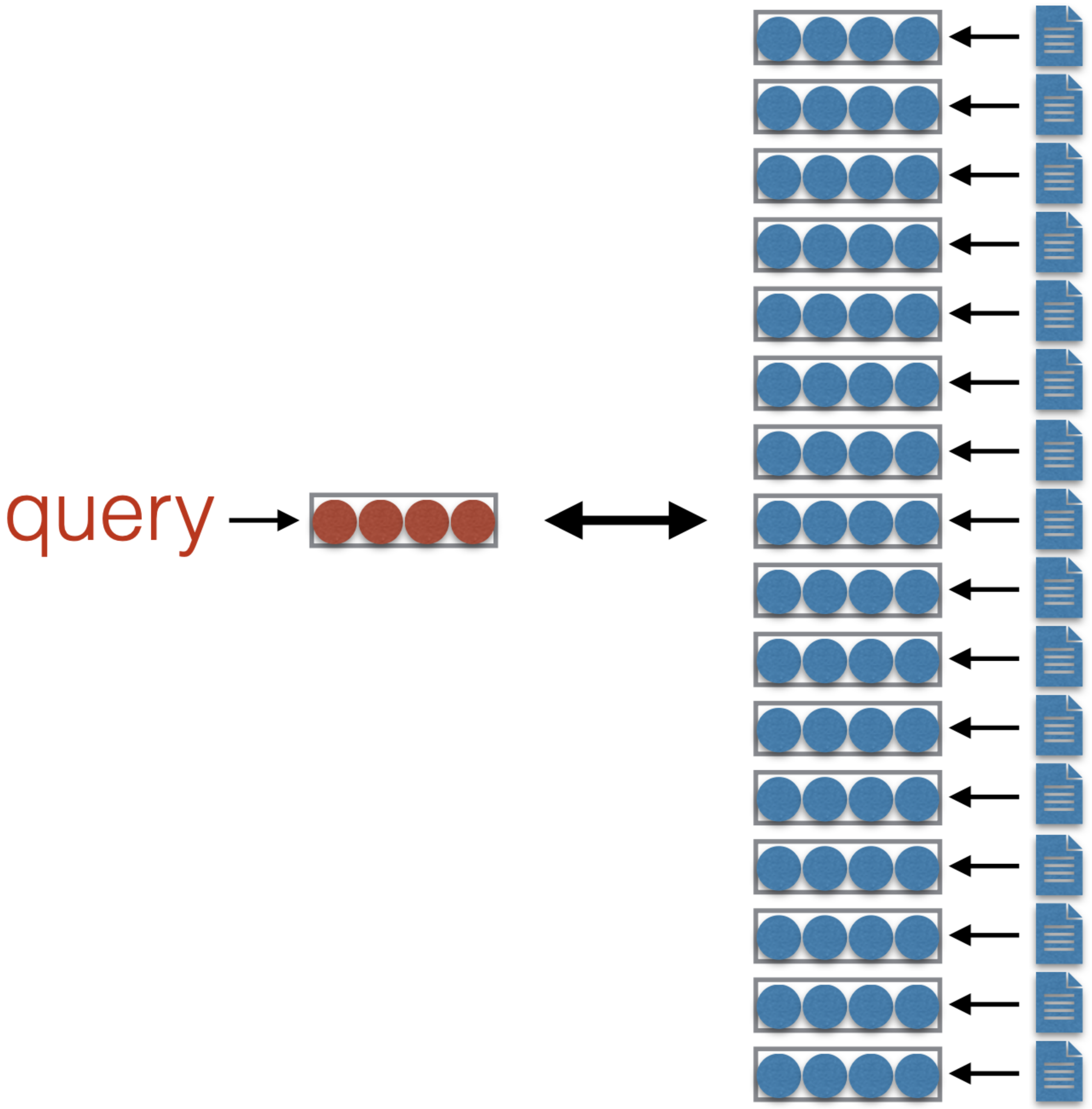


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Dense Embeddings

- Encode all documents using a LM and index them (one time task). Can use:
  - ✓Out-of-the-box embeddings. E.g. BERT
  - ✓Learned embeddings (covered later)

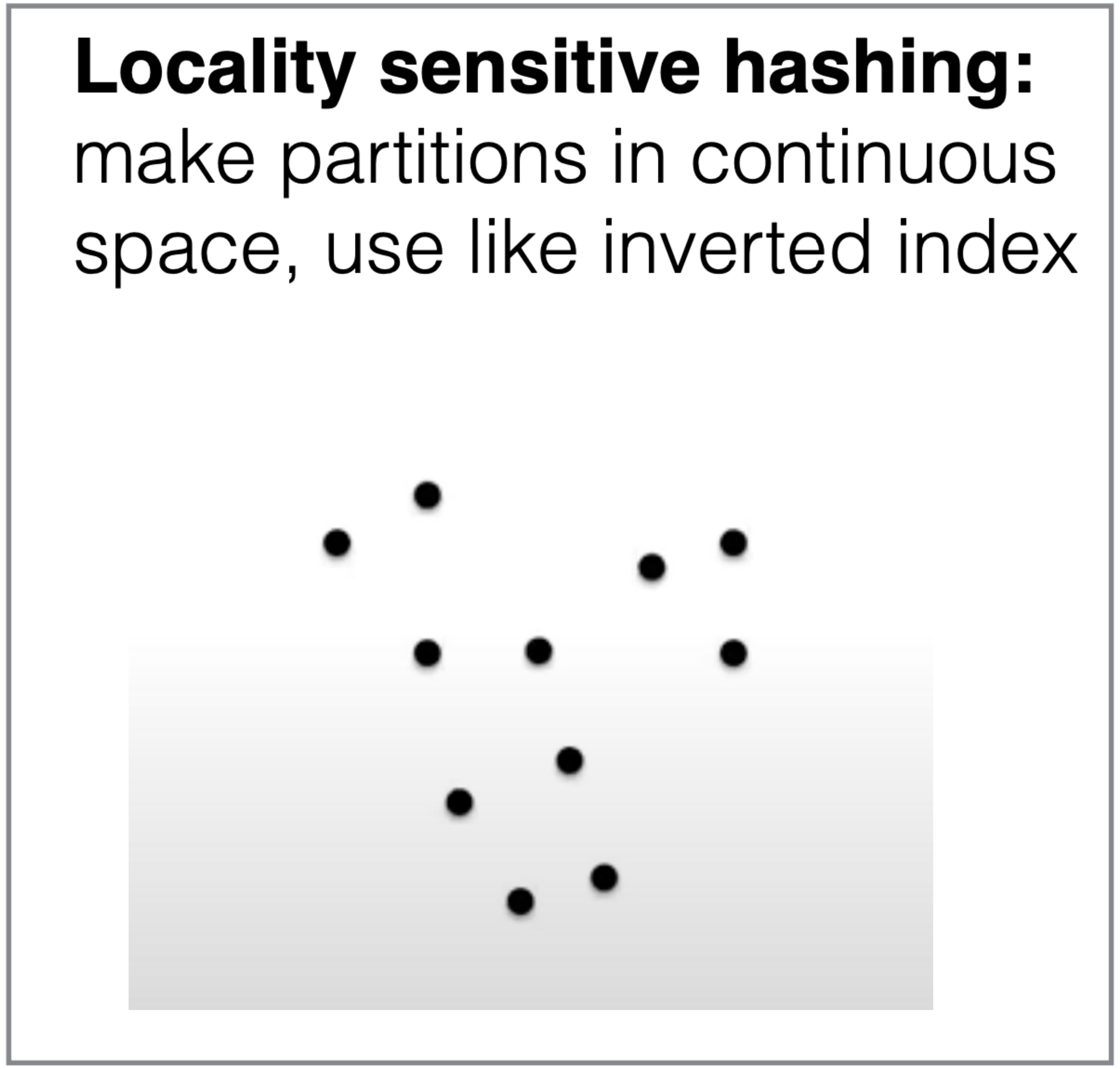
- At test time:
  - Encode Query
  - Use Nearest Neighbor Search to find similar documents



Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search
Maximum Inner Product Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

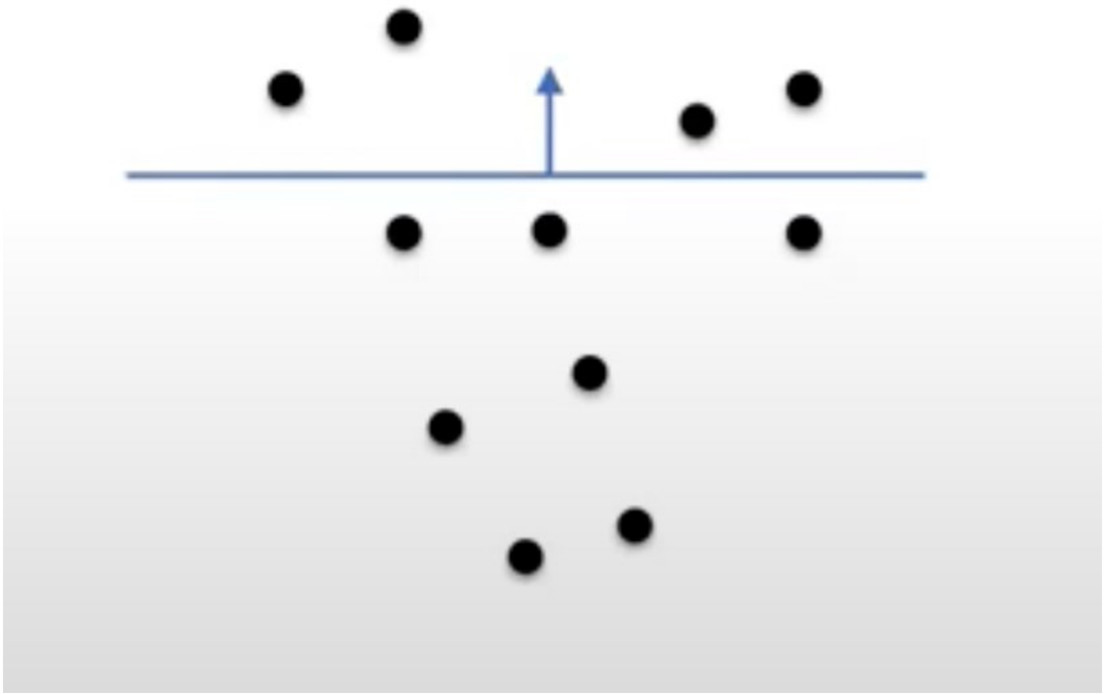


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

Locality sensitive hashing: make partitions in continuous space, use like inverted index

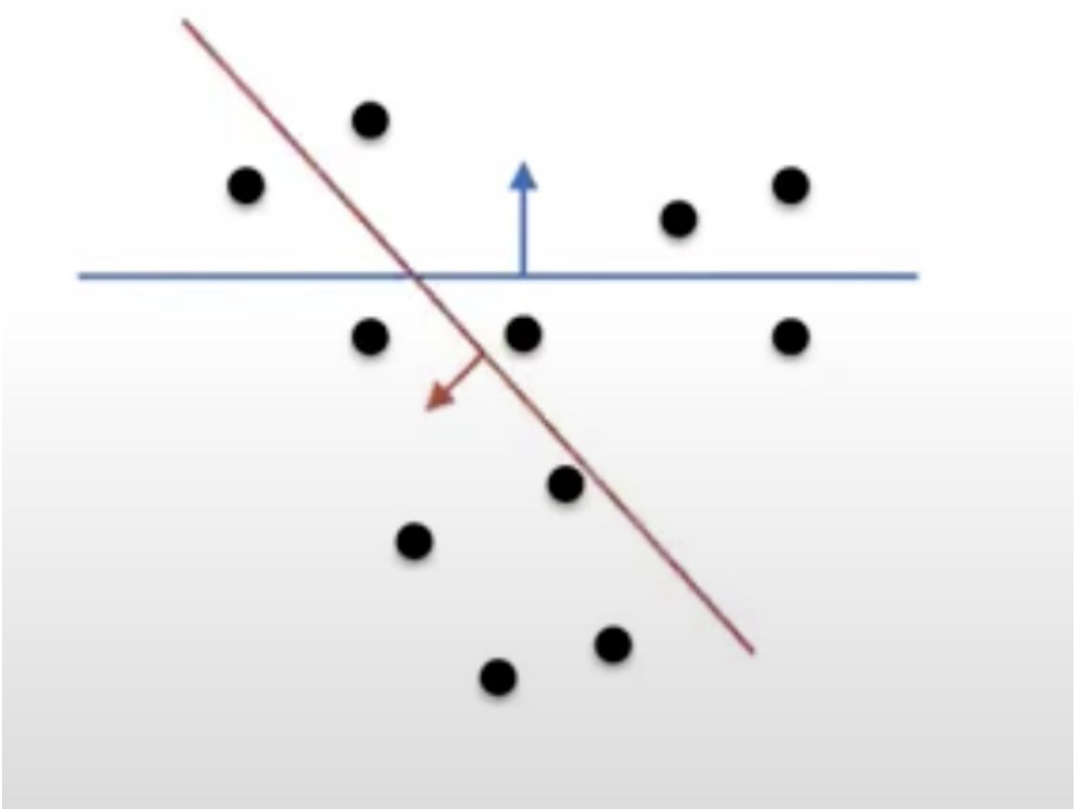


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

Locality sensitive hashing:
make partitions in continuous space, use like inverted index

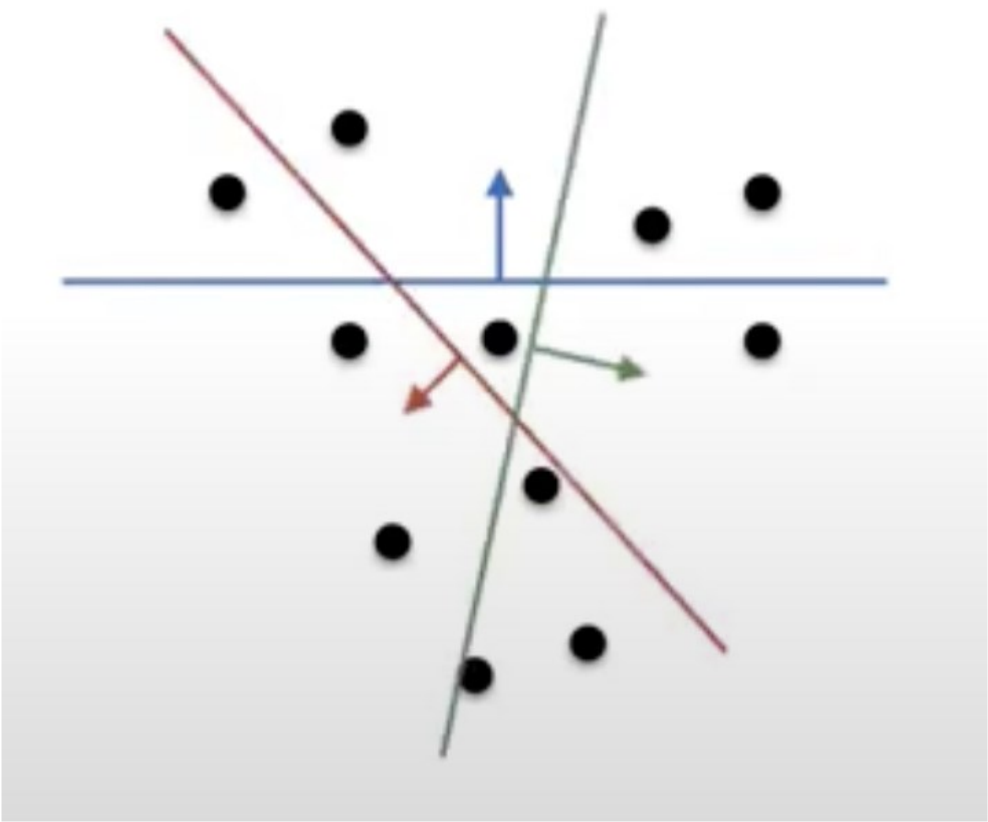


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Approximate Nearest Neighbor Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

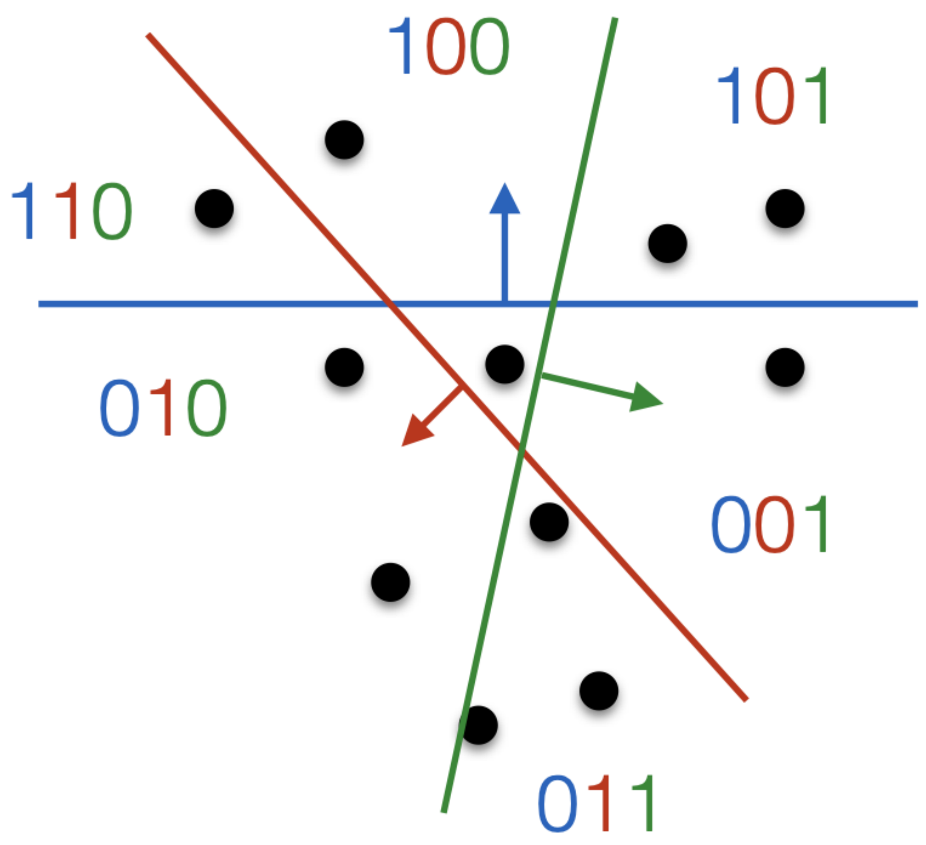
Locality sensitive hashing: make partitions in continuous space, use like inverted index



Approximate Nearest Neighbor Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

Locality sensitive hashing: make partitions in continuous space, use like inverted index

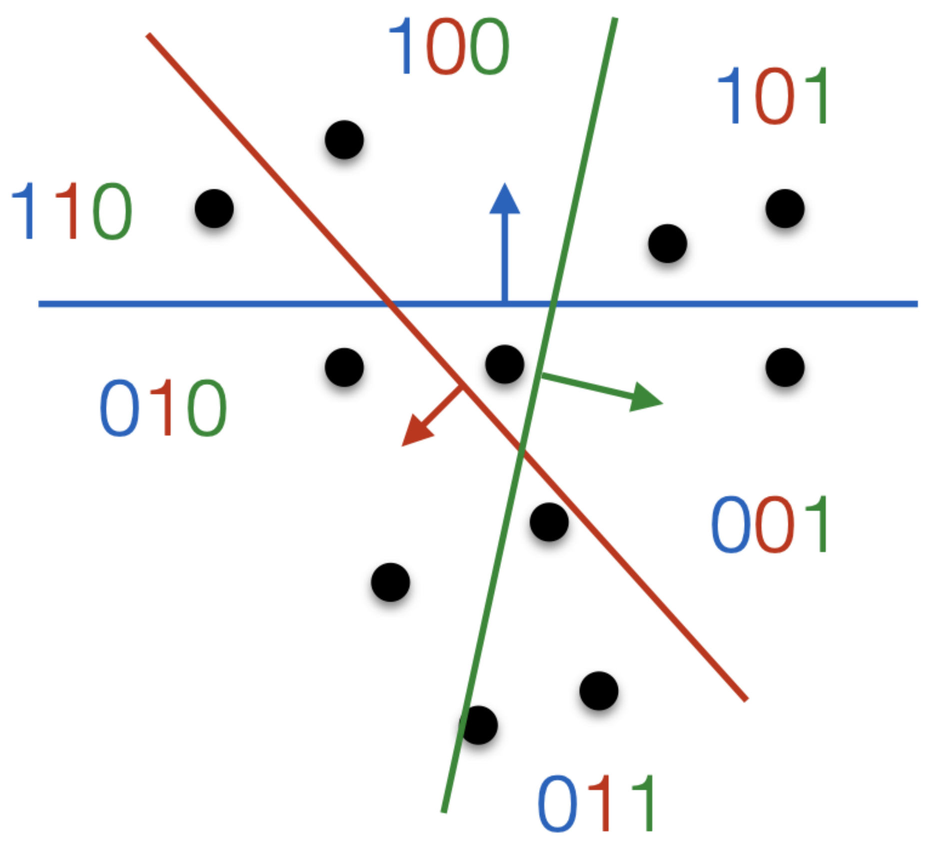


Graph-based search: create "hubs" and search from there

Approximate Nearest Neighbor Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

Locality sensitive hashing: make partitions in continuous space, use like inverted index



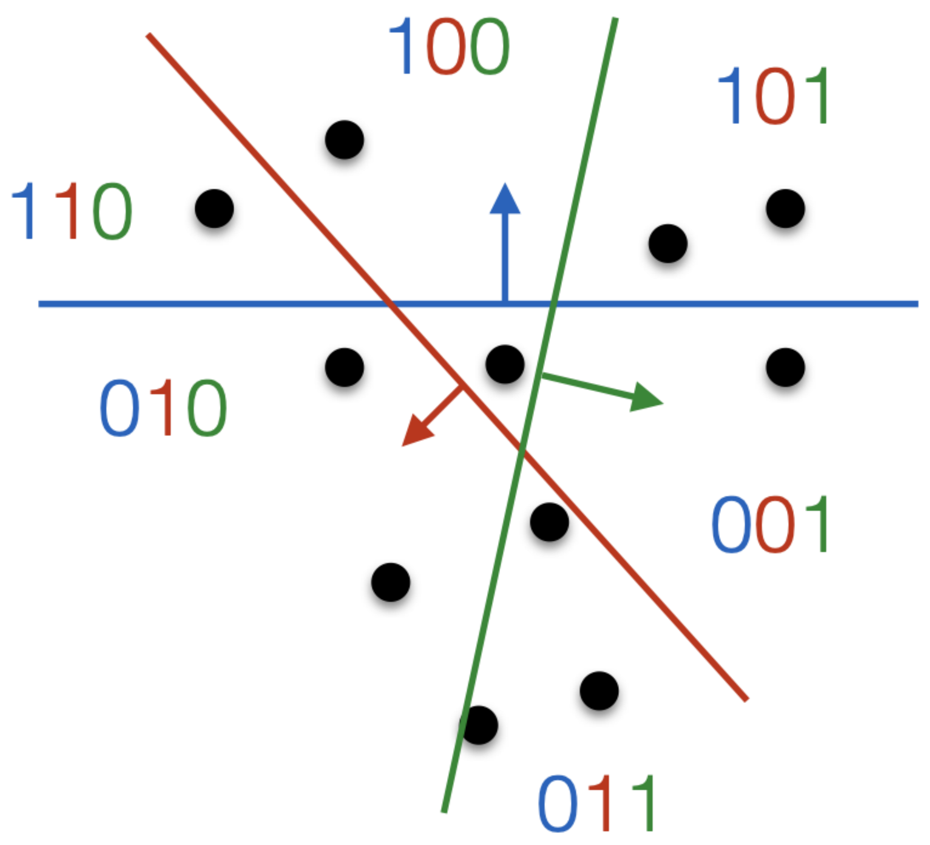
Graph-based search: create "hubs" and search from there

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

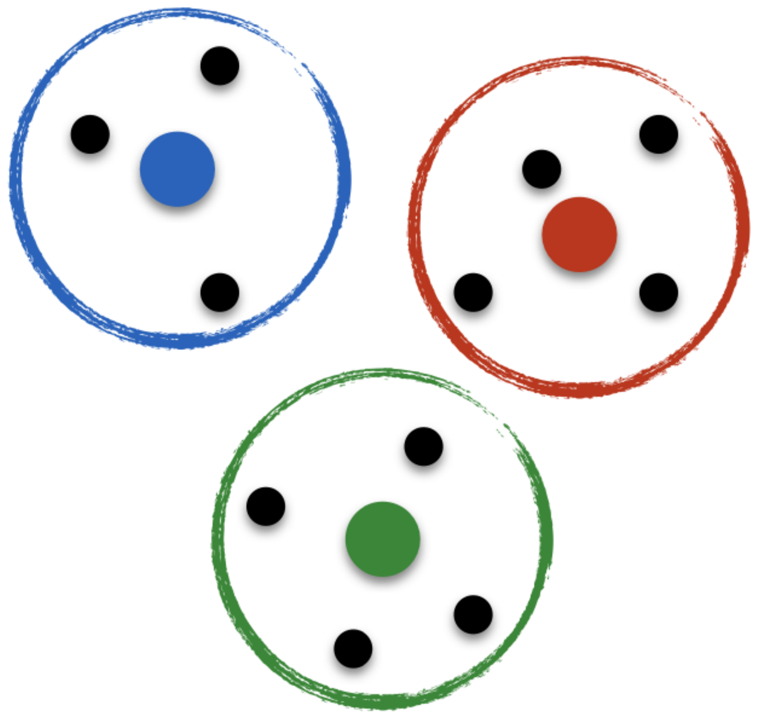
Approximate Nearest Neighbor Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

Locality sensitive hashing: make partitions in continuous space, use like inverted index



Graph-based search: create "hubs" and search from there

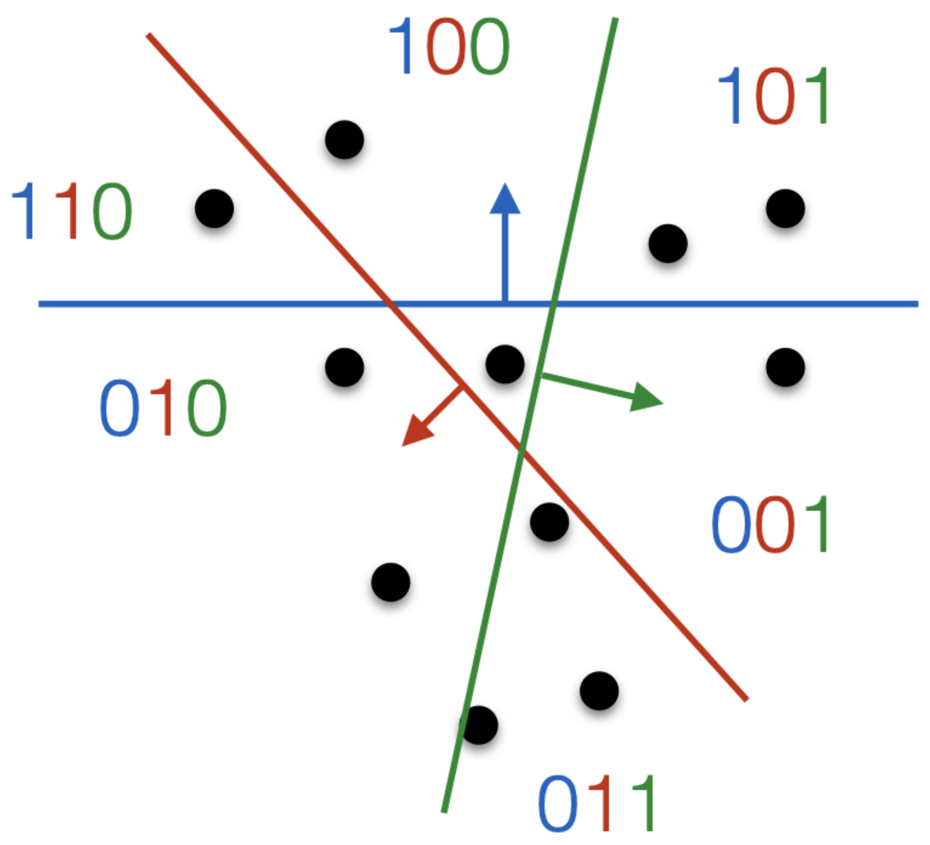


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

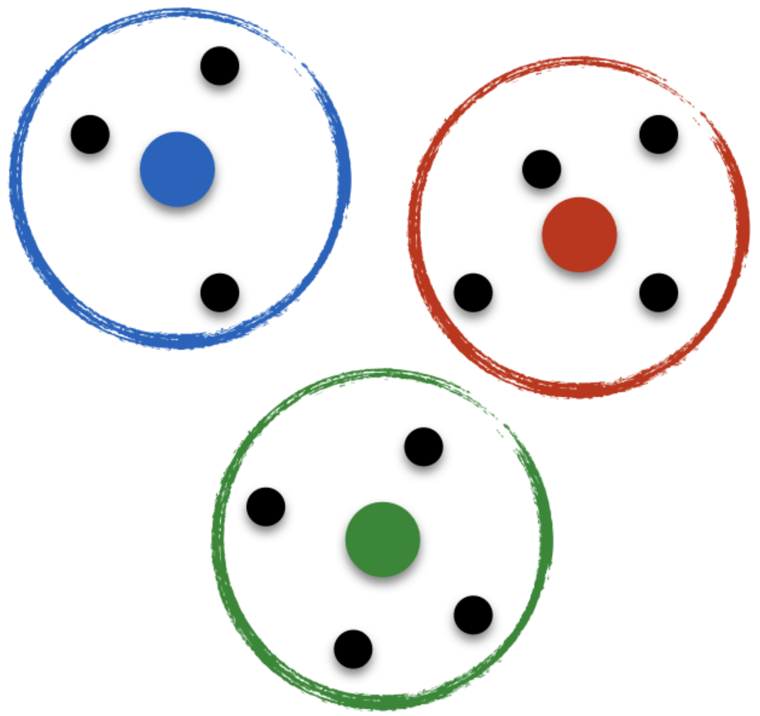
Approximate Nearest Neighbor Search (MIPS)

- Methods to retrieve embeddings in sub-linear time

Locality sensitive hashing: make partitions in continuous space, use like inverted index



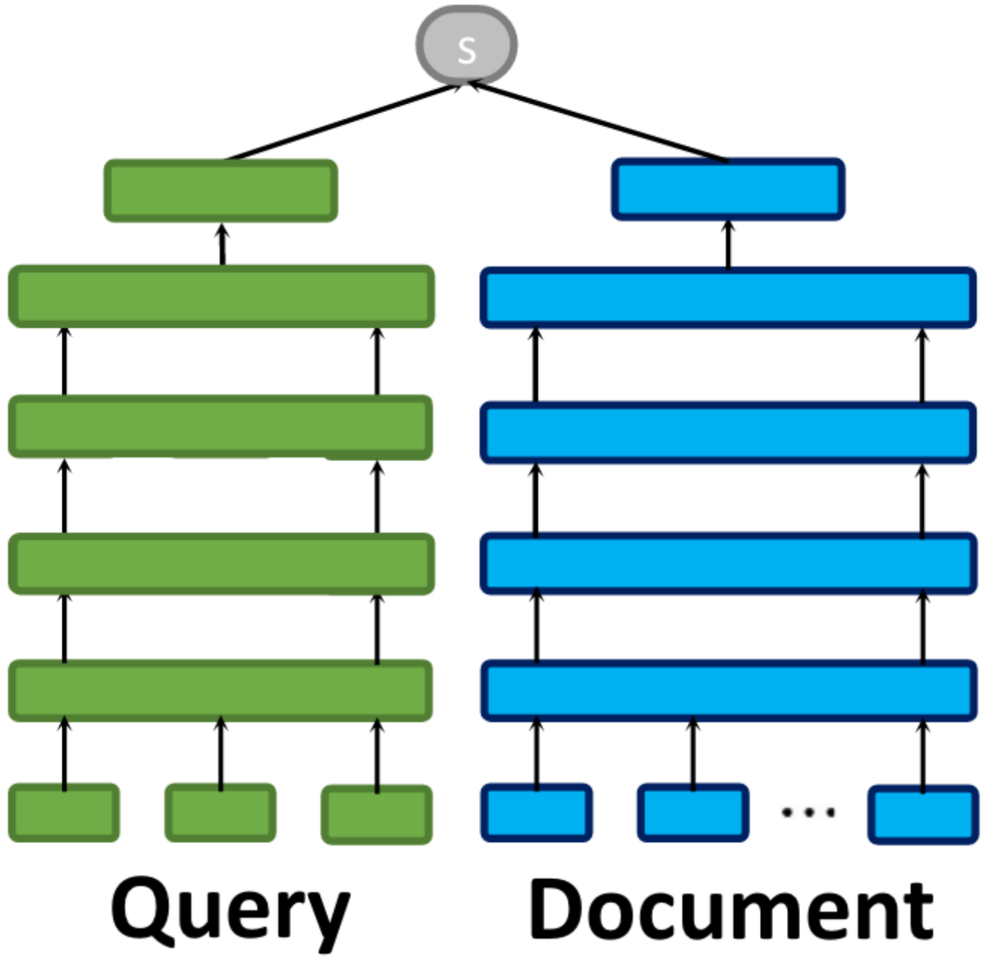
Graph-based search: create “hubs” and search from there



- Software: ANNOY (Spotify), FAISS

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Bi-Encoder Scoring

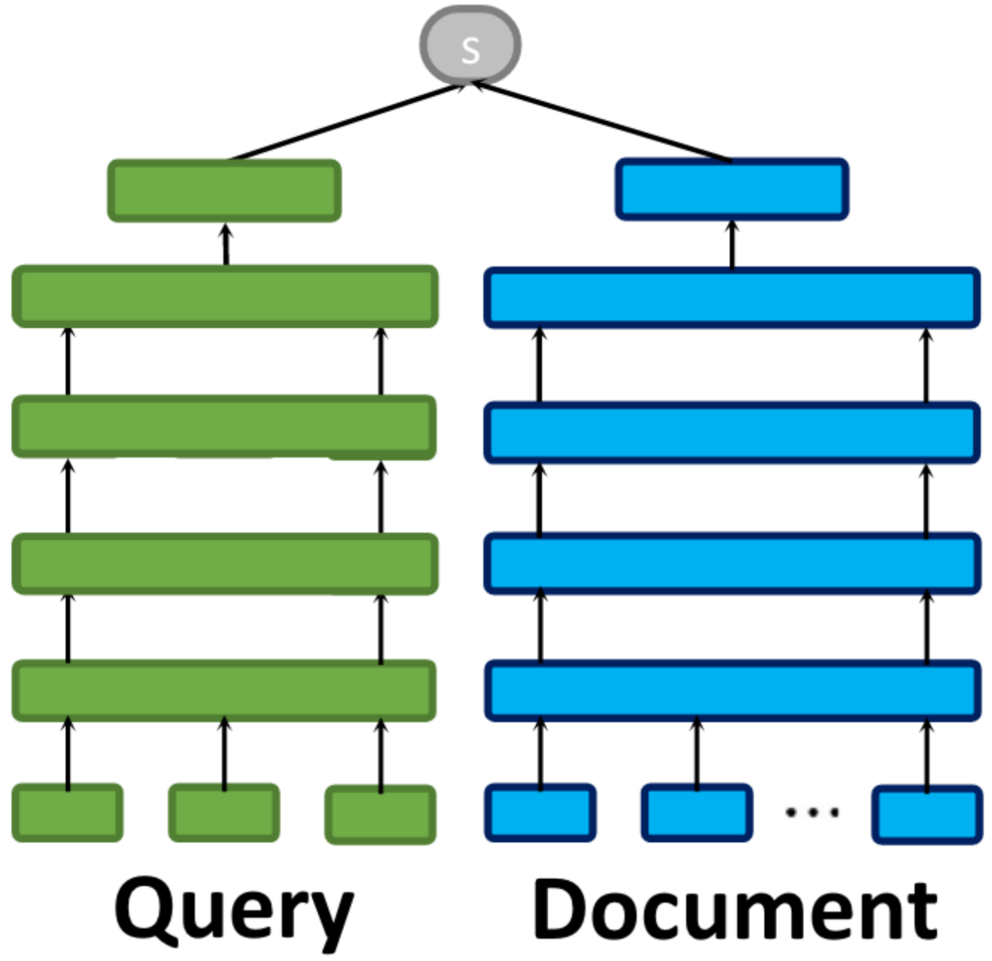


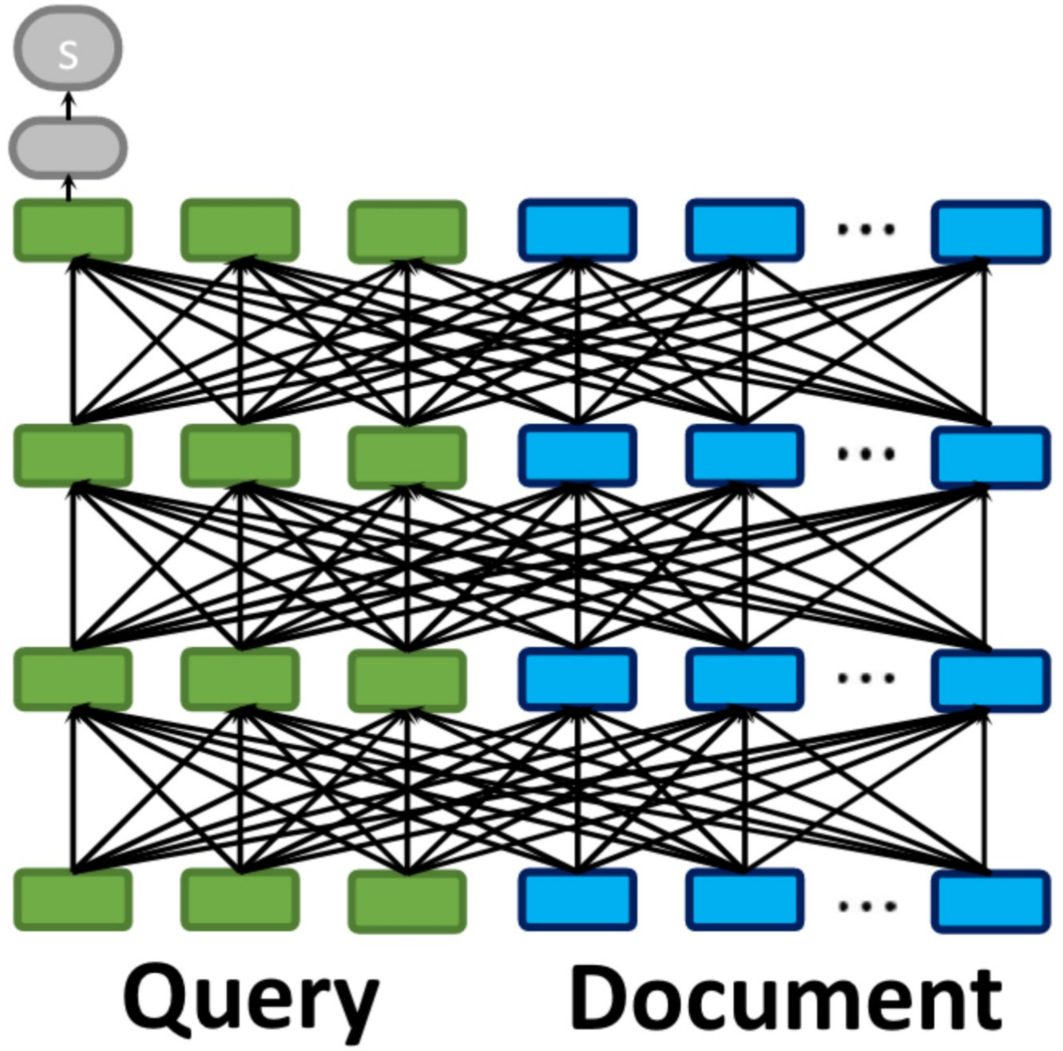
Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Cross-Encoder Reranking

- Jointly encode both queries and documents using neural model (Nogueira et al. 2019)



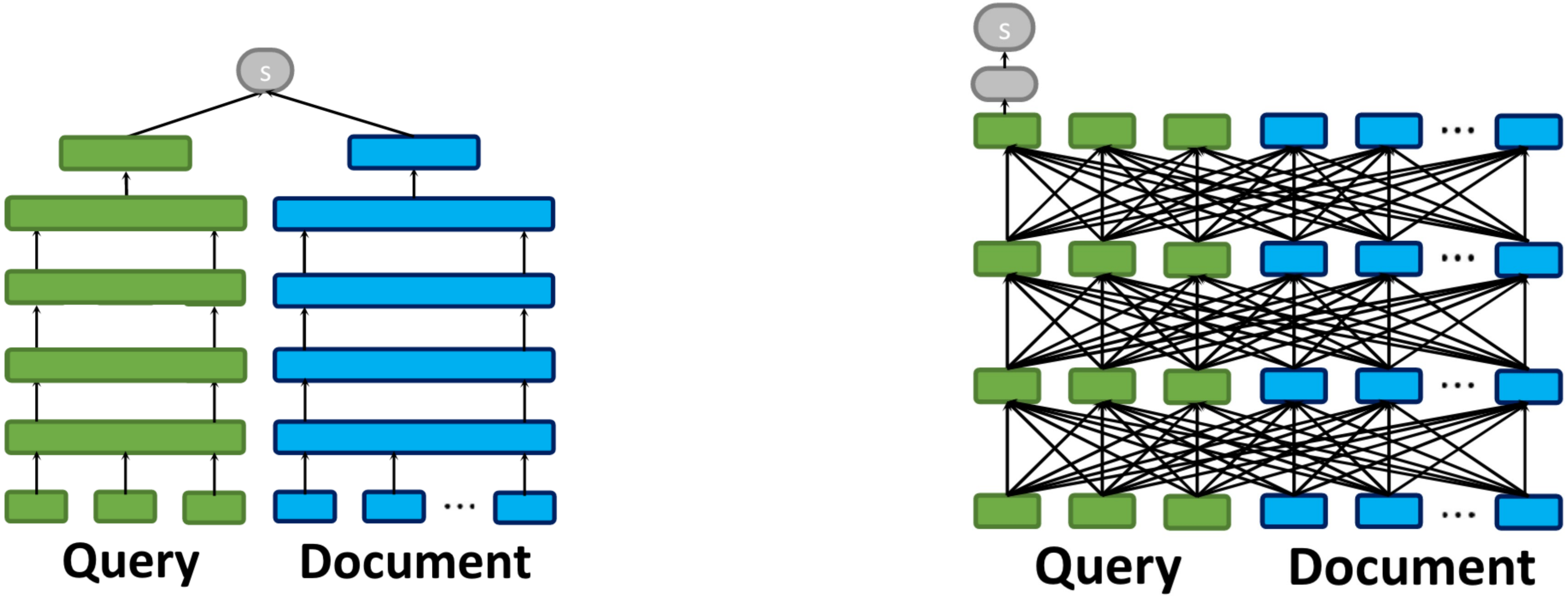


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Cross-Encoder Reranking

- Jointly encode both queries and documents using neural model (Nogueira et al. 2019)



- Precludes approximate nearest neighbour lookup, so can only be used on small number of candidates

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Token-level Dense Retrieval

ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Significantly more effective (but more costly) than single-vector retrieval

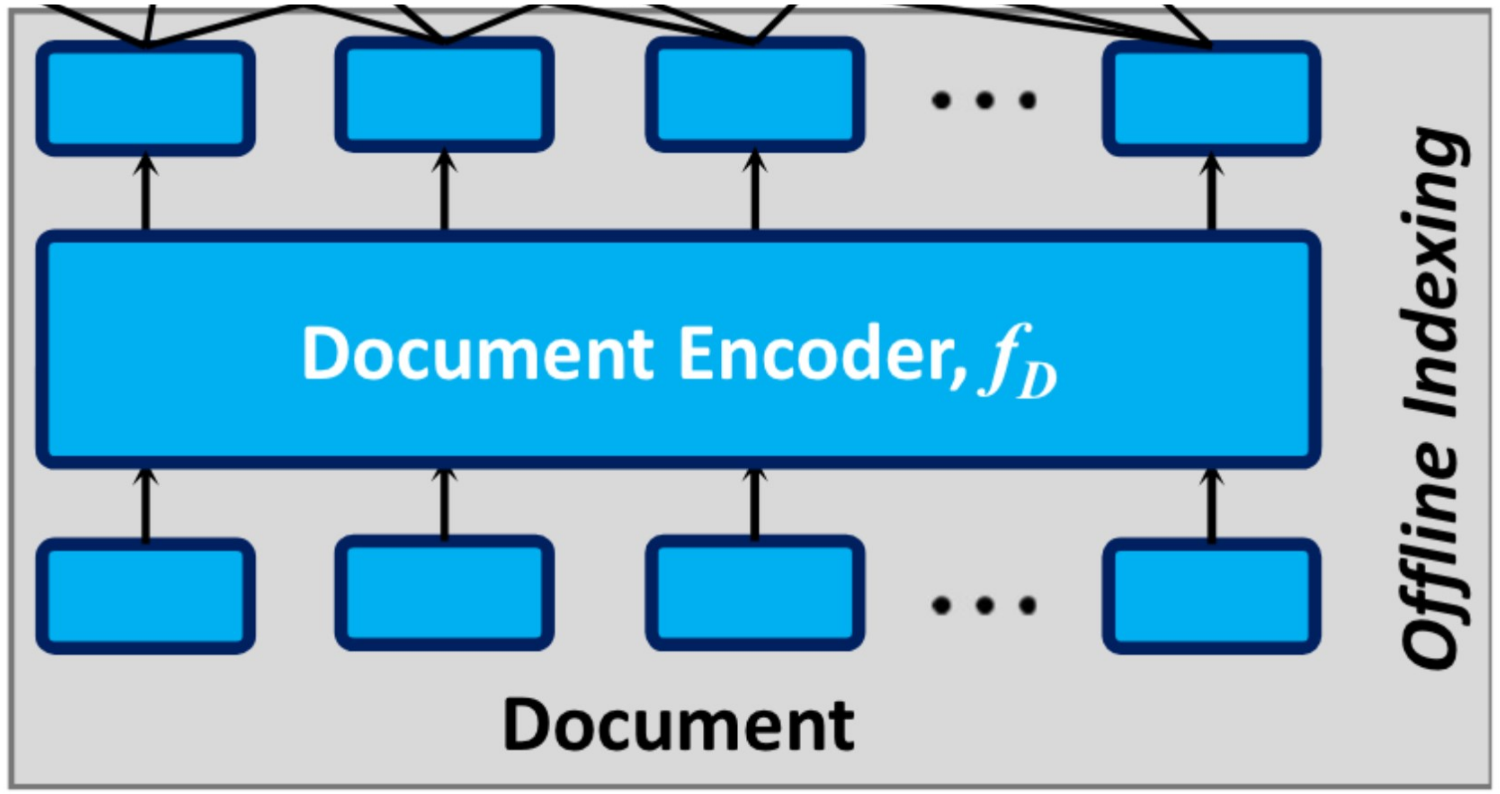


Figure from Khattab et al. (2020)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Token-level Dense Retrieval

ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Significantly more effective (but more costly) than single-vector retrieval

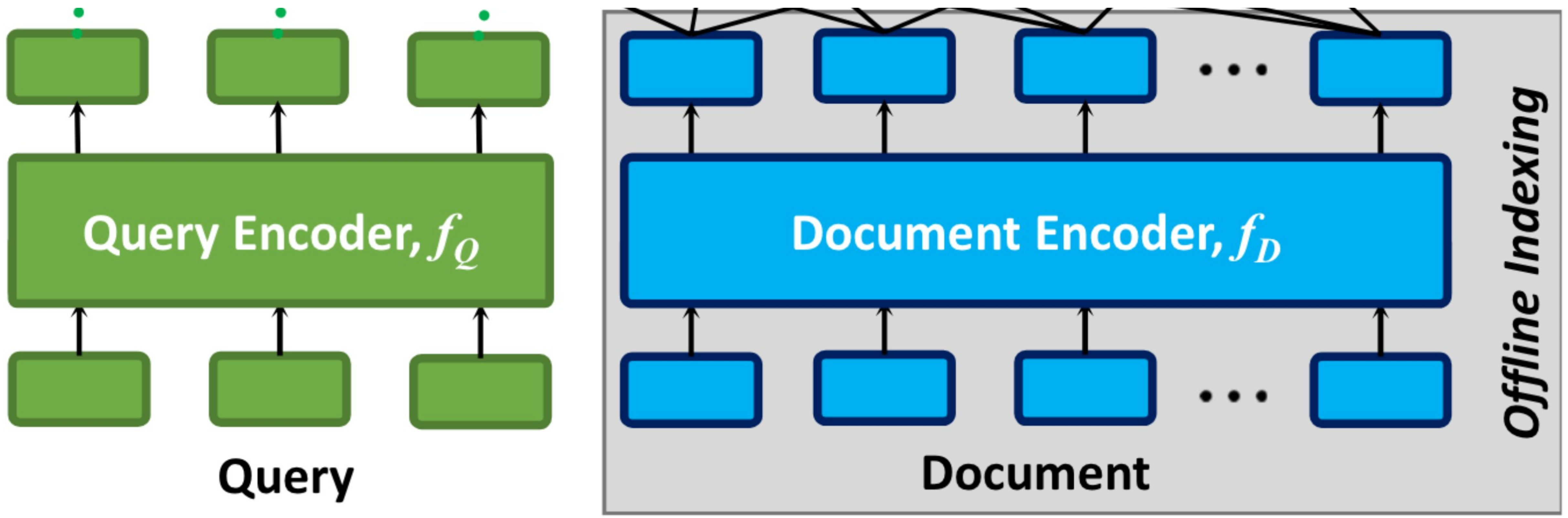


Figure from Khattab et al. (2020)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Token-level Dense Retrieval

ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Significantly more effective (but more costly) than single-vector retrieval

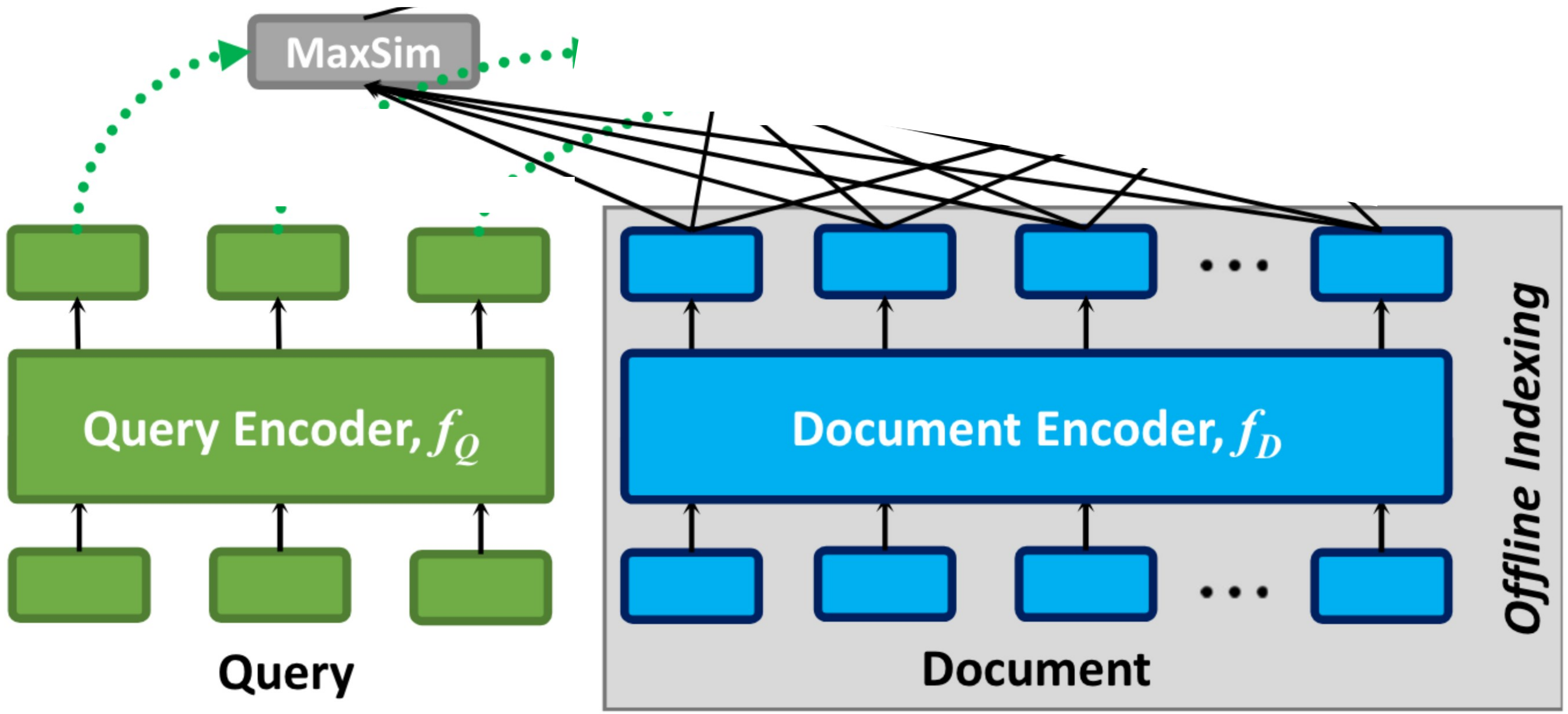


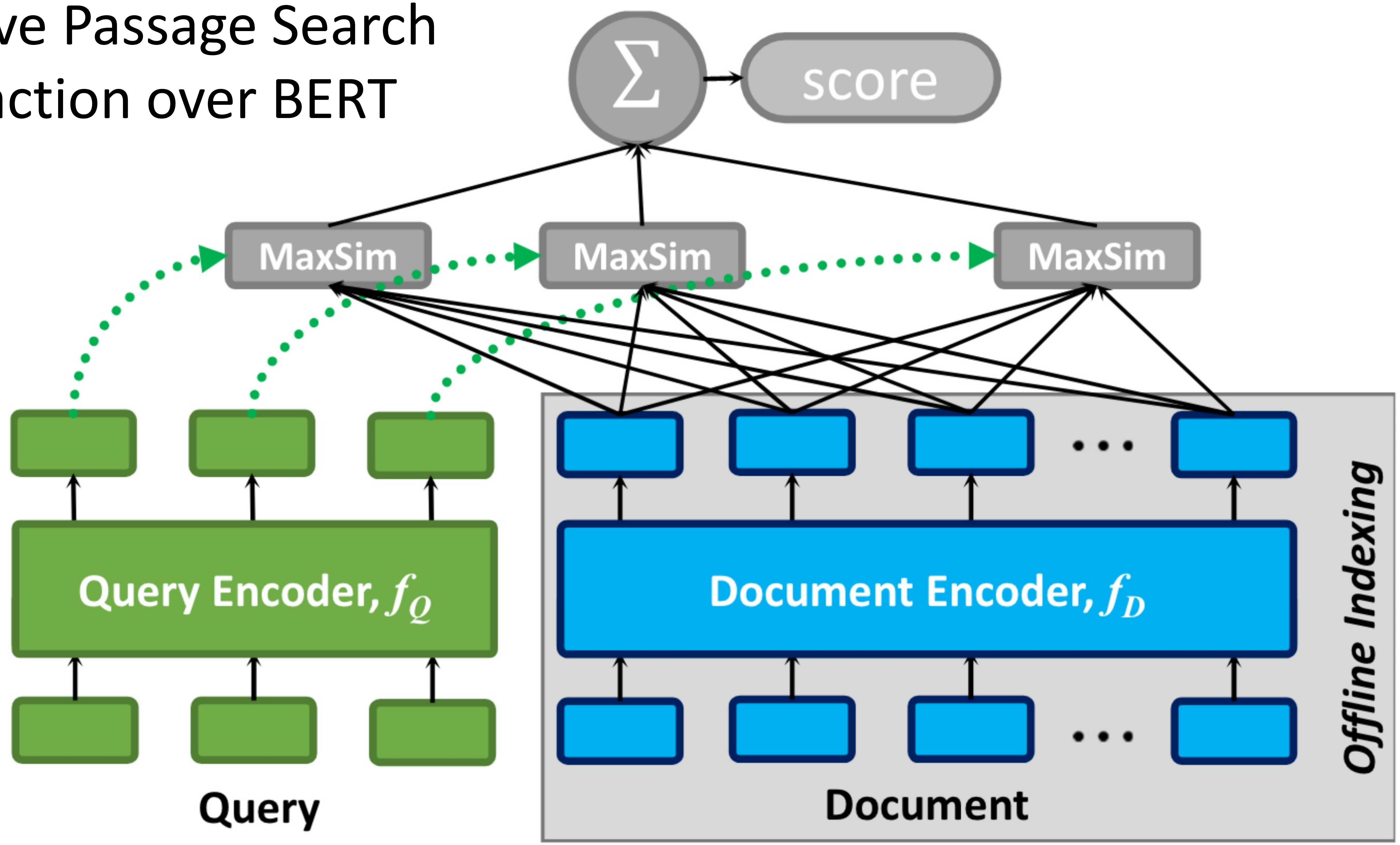
Figure from Khattab et al. (2020)

Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Token-level Dense Retrieval

ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT

Significantly more effective (but more costly) than single-vector retrieval

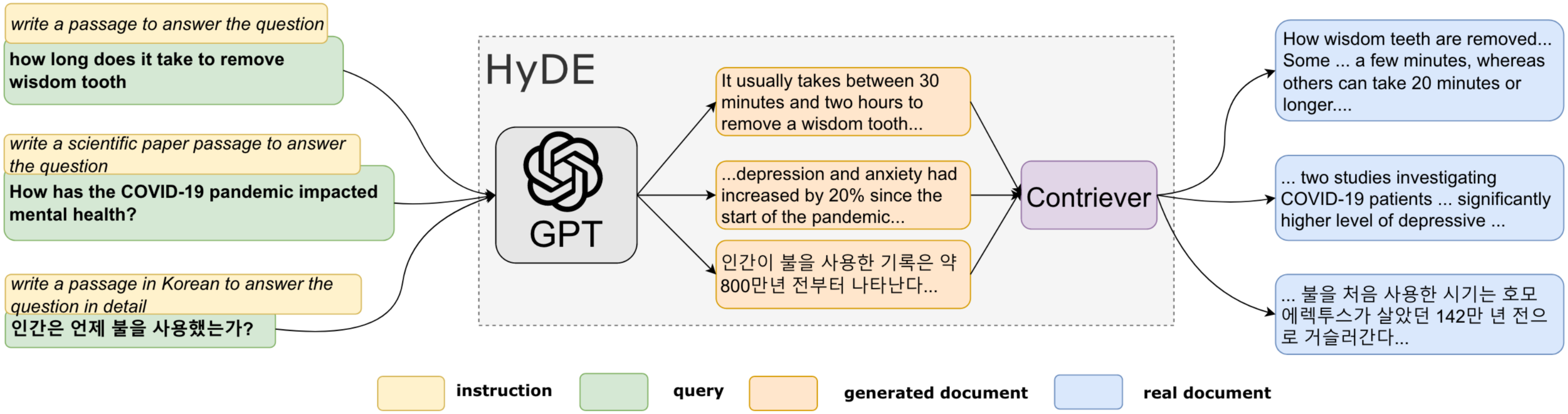


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Figure from Khattab et al. (2020)

Hypothetical Document Embeddings (Gao et al. 2023)

- Generate a "hypothetical document" for the query using an LLM, and try to look it up
- Can be easier than trying to match under-specified query

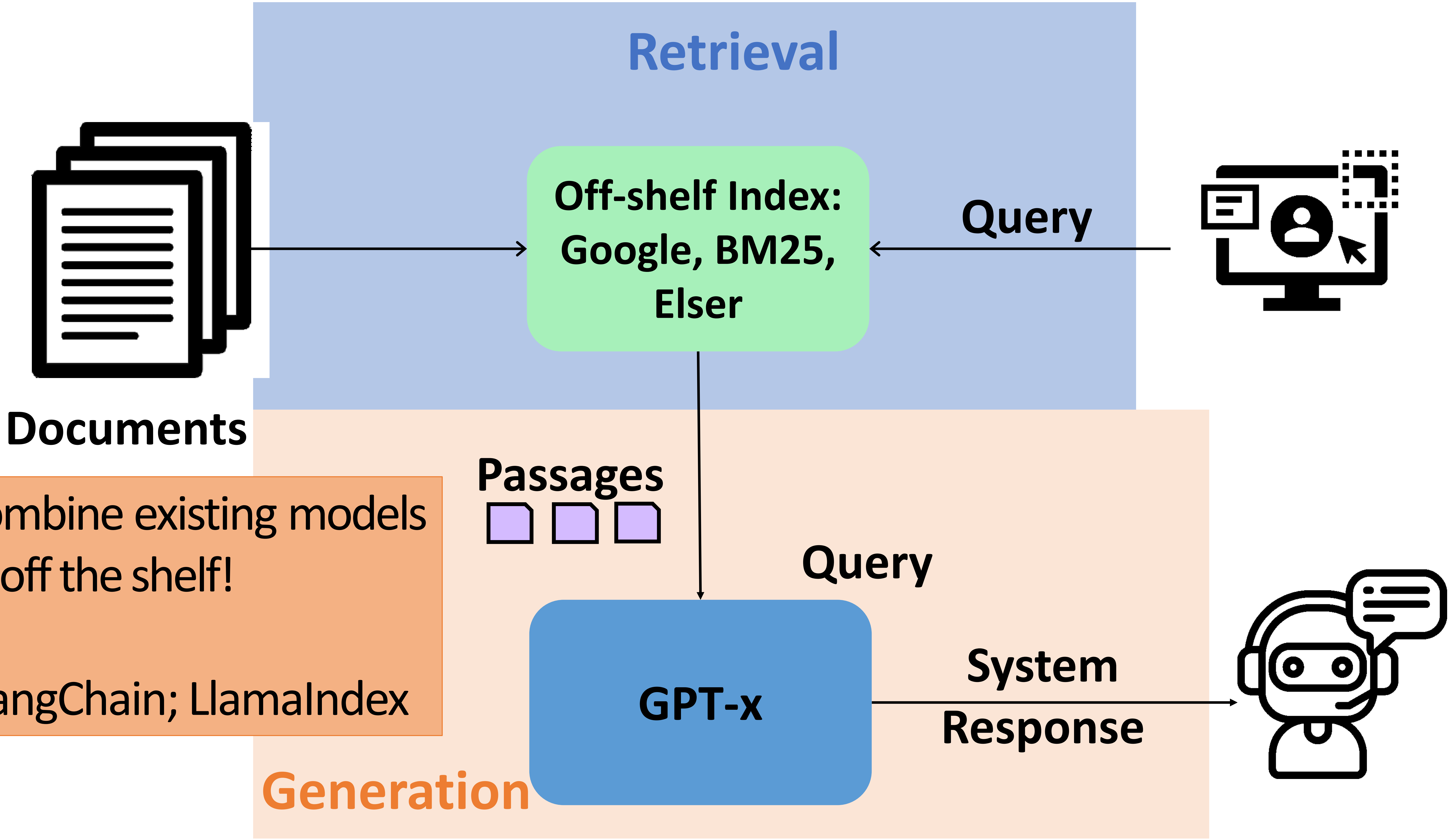


Slide source: https://phontron.com/class/anlp2024/assets/slides/anlp-10-rag.pdf

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs *– close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Retrieval Based LLMs - Architecture



- ✓ Simply combine existing models available off the shelf!
- ✓ Check: LangChain; LlamaIndex

Retrieval Based LLMs - Architecture

- REALM (Guu et al 2020): Retrieval-Augmented Language Model Pre-Training ICML 2020
- RAG (Lewis et al 2020): Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

REALM (Guu et al 2020)

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

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

↓

LM

↓

48

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

REALM (Guu et al 2020)

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

q (=x)

Retrieval

k chunks of text (passages)

FIFA World Cup 2026 will expand to 48 teams.

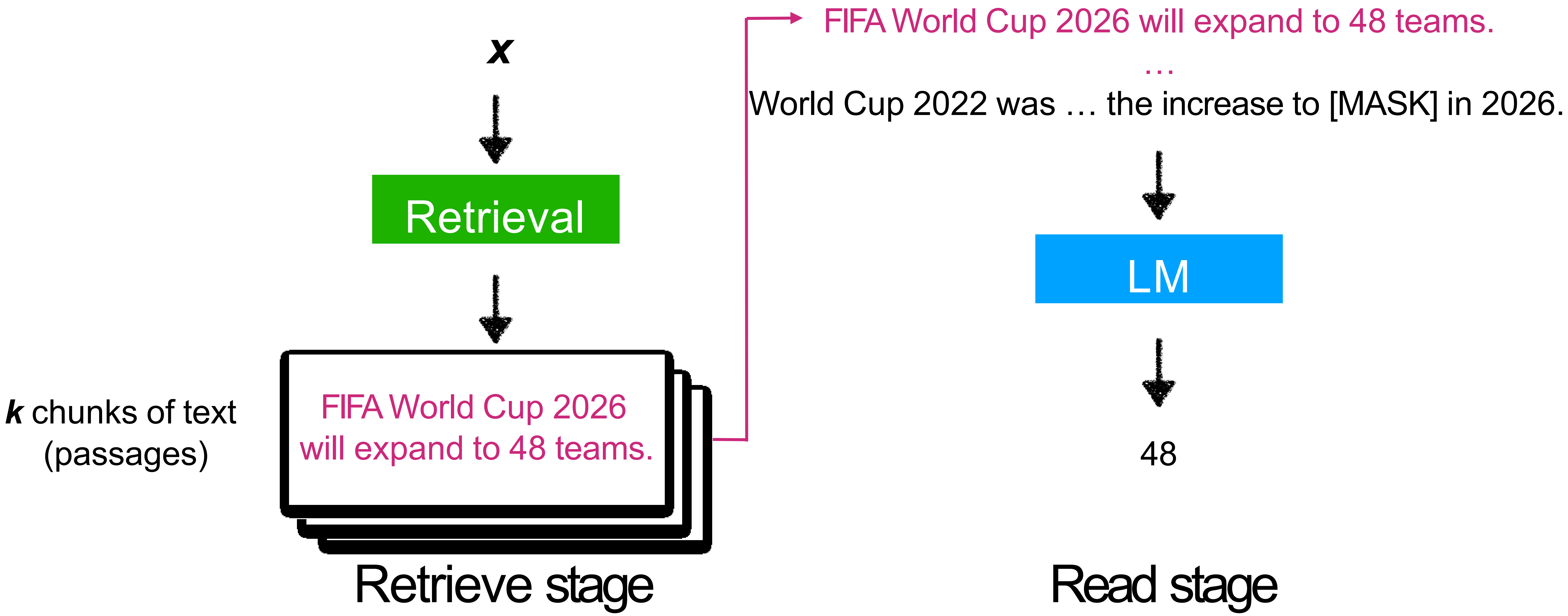
World Cup 2022 was … the increase to [MASK] in 2026.

LM

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

REALM (Guu et al 2020)

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



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

REALM (Guu et al 2020)

[CLS] z_1 [SEP] x → LM → $P(y \mid x, z_1)$

[CLS] z_2 [SEP] x → LM → $P(y \mid x, z_2)$

⋮

[CLS] z_k [SEP] x → LM → $P(y \mid x, z_k)$

Weighted average

$$P(y \mid x) = \sum \underbrace{P(z \mid x)}_{\text{from the etrieve stage}} \underbrace{P(y \mid x, z)}_{\text{from the read stage}}$$

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

REALM (Guu et al 2020)

[CLS] z_1 [SEP] x → LM → $P(y \mid x, z_1)$

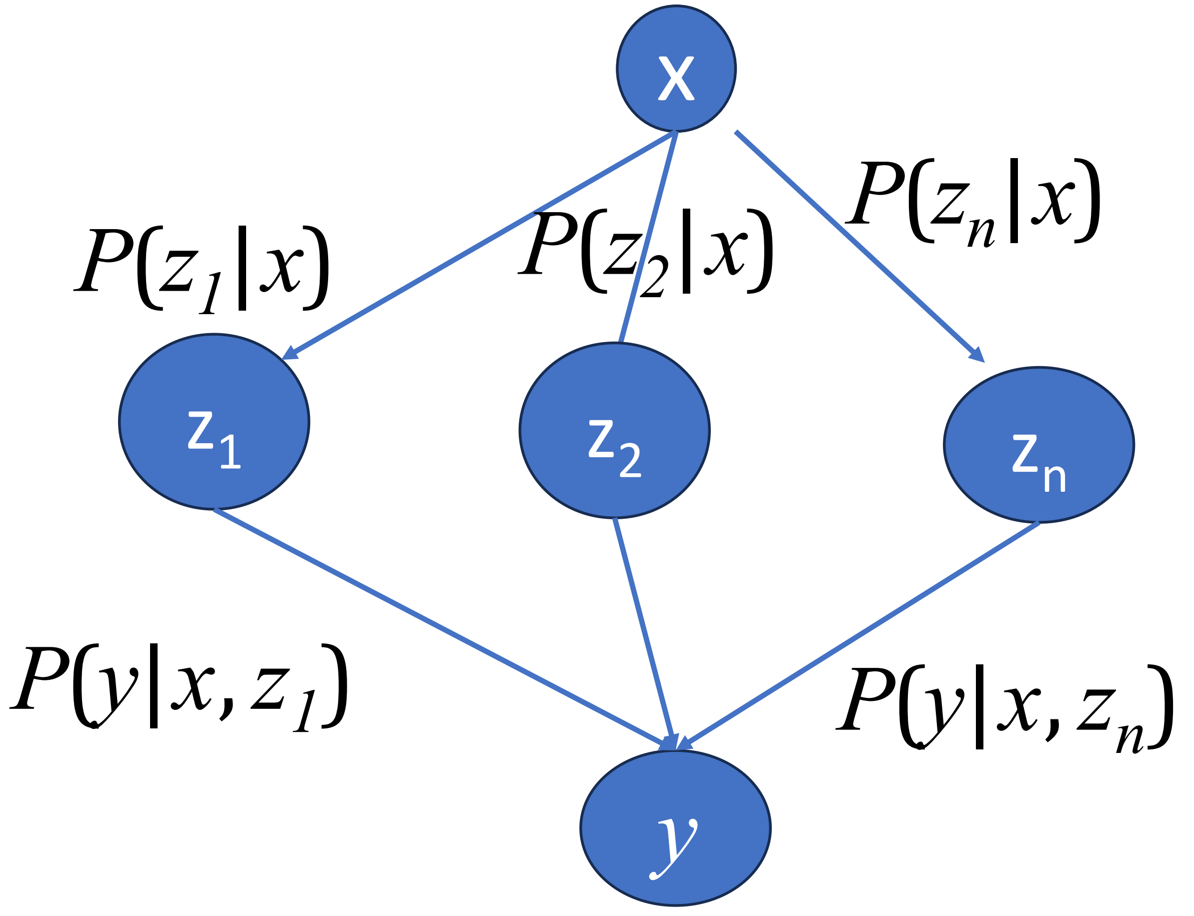
[CLS] z_2 [SEP] x → LM → $P(y \mid x, z_2)$

⋮

[CLS] z_k [SEP] x → LM → $P(y \mid x, z_k)$

Weighted average

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



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

REALM (Guu et al 2020)

[CLS] z_1 [SEP] x → LM → $P(y|x, z_1)$

[CLS] z_2 [SEP] x → LM → $P(y|x, z_2)$

⋮

[CLS] z_k [SEP] x → LM → $P(y|x, z_k)$

Weighted average

0 if not one of top *k*

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

Need to approximate
Consider top *k* chunks only

REALM: Joint Training

Trainable components

- Retriever
  - Document Encoder
  - Query Encoder

- Reader: LM

REALM: Training

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

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

REALM: Training

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

Retriever

q (=x)

↓

Index

↓

top-K retrieved chunks

The pyramidion on top allows for less material higher up the pyramid.

$P_\eta(z|x)$

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

REALM: Training

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

Retriever Reader

q (=x)

Index

top-K retrieved chunks

The pyramidion on top allows for less material higher up the pyramid.

$P_\eta(z|x)$

The pyramidion on top … the pyramid.

…

The [MASK] at the top of the pyramid.

LM

pyramid

$P_\theta(y|x,z)$

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

REALM: Training

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

REALM: Training Approximations

- Freeze top-k documents

- Freeze index (document embeddings), but search top-k documents

- Update index every T steps

REALM – Results

Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)	# params
Baselines with Frozen retriever + reranking						
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m

REALM – Results

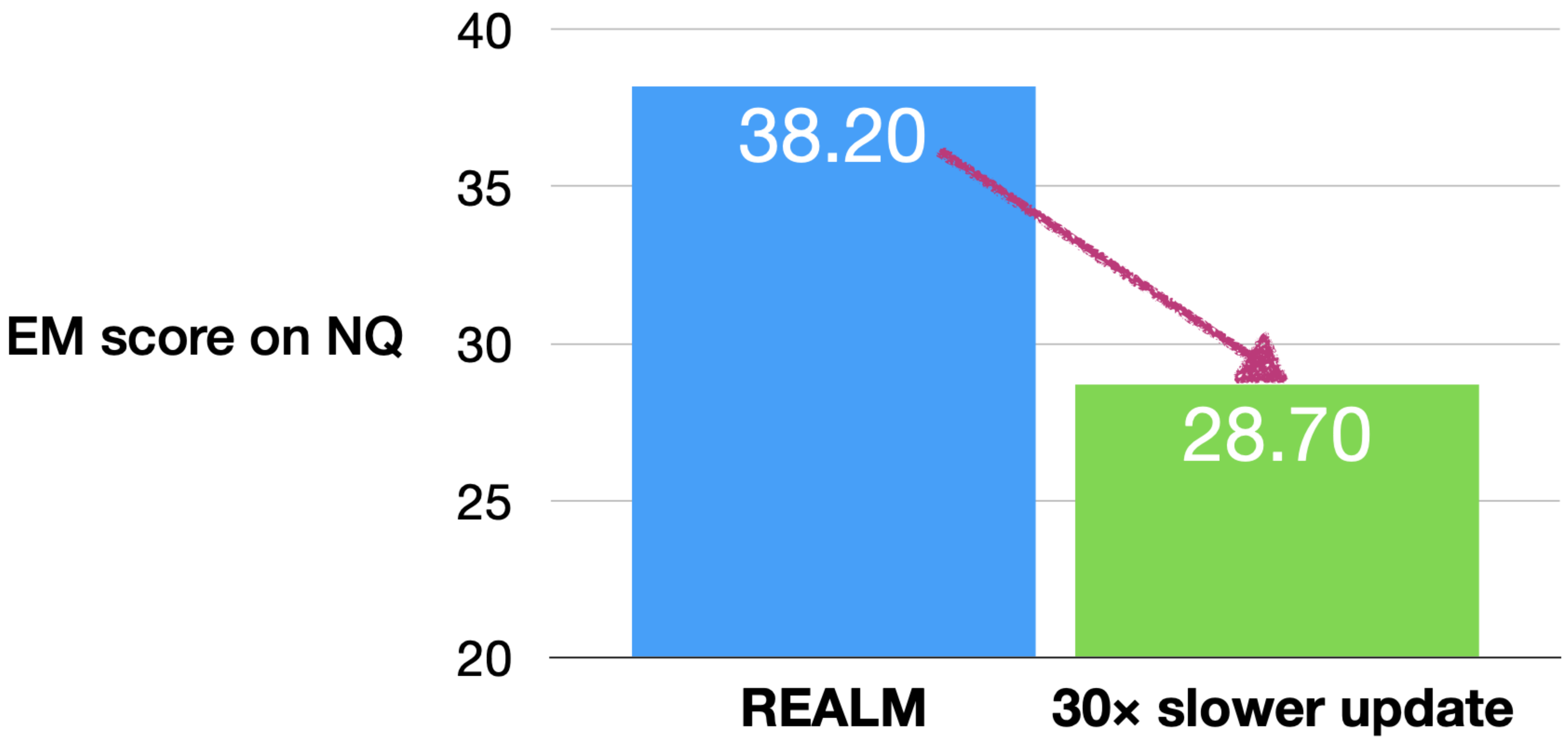
Name	Architectures	Pre-training	NQ (79k/4k)	WQ (3k/2k)	CT (1k /1k)	# params
Baselines with Frozen retriever + reranking						
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m
REALM						
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9	330m

REALM: Index update rate

How often should we update the retrieval index?

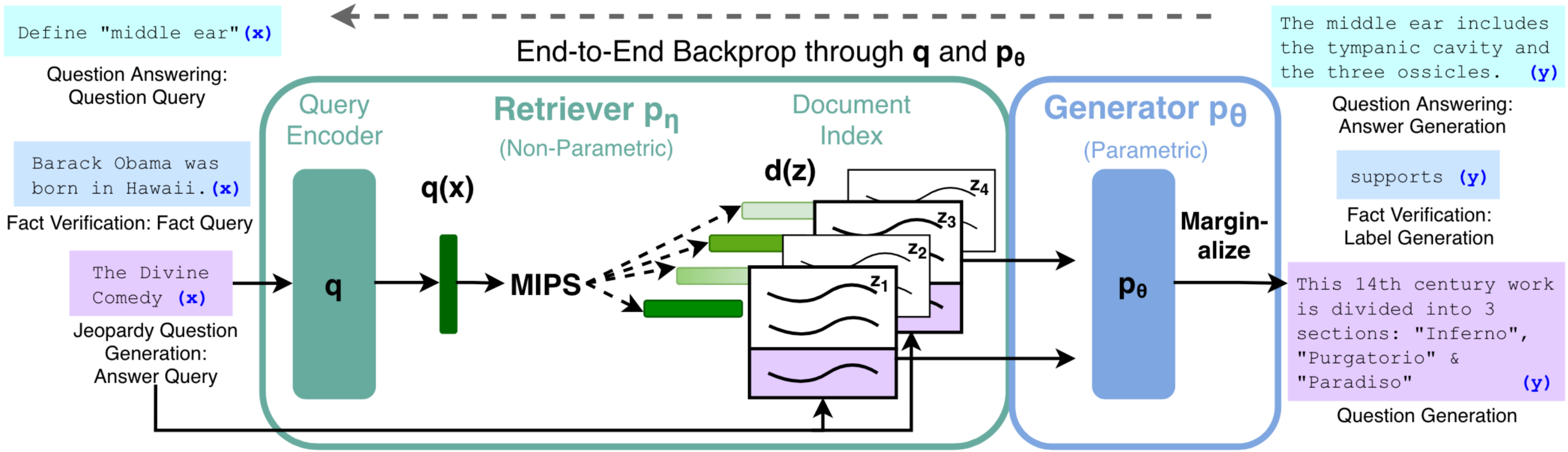
- Frequency too high: expensive
- Frequency too slow: out-dated

REALM: updating the index every 500 training steps



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

RAG: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al. 2020)



RAG: Joint Training Equation (Lewis et al. 2020)

RAG-Token Model

Same as REALM

$$p_{\text{RAG-Token}}(y|x) \approx \boxed{\sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})}$$

RAG: Joint Training Equation (Lewis et al. 2020)

RAG-Token Model

Same as REALM

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \boxed{\sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})}$$

RAG: Joint Training Equation (Lewis et al. 2020)

RAG-Token Model

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

RAG: Joint Training Equation (Lewis et al. 2020)

RAG-Token Model

$$p_{\text{RAG-Token}}(y|x) \approx \prod_{i}^{N} \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x, z_i, y_{1:i-1})$$

RAG-Sequence Model

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x, z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_{i}^{N} p_\theta(y_i|x, z, y_{1:i-1})$$

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs *– close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Training methods for retrieval-augmented LMs

- Independent training
- Sequential training
- Joint training

Training methods for retrieval-augmented LMs

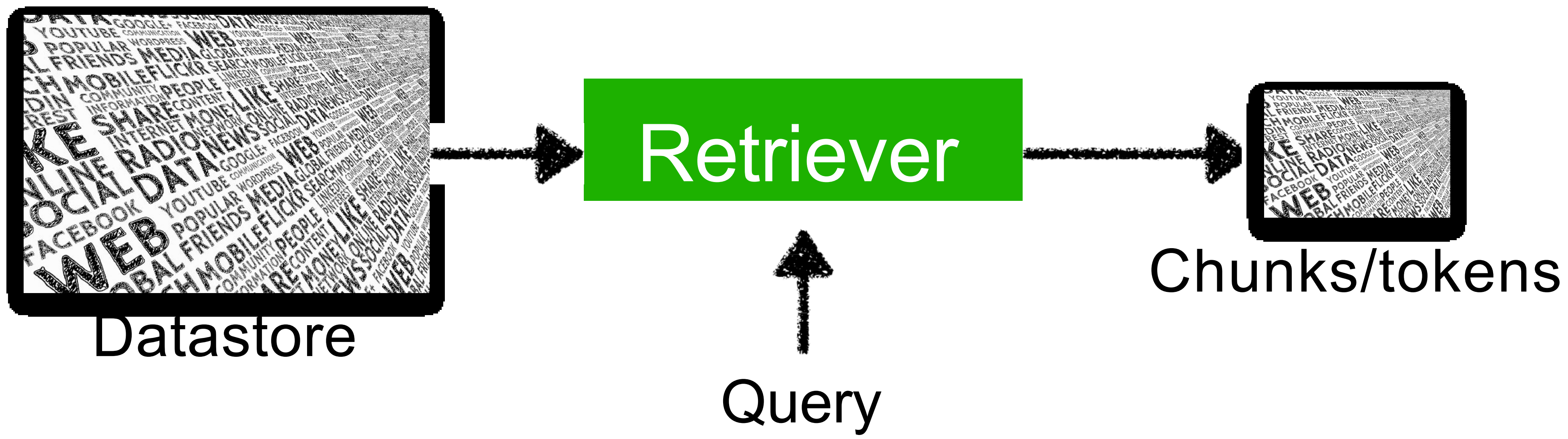
- **Independent training**
- Sequential training
- Joint training

Independent Training

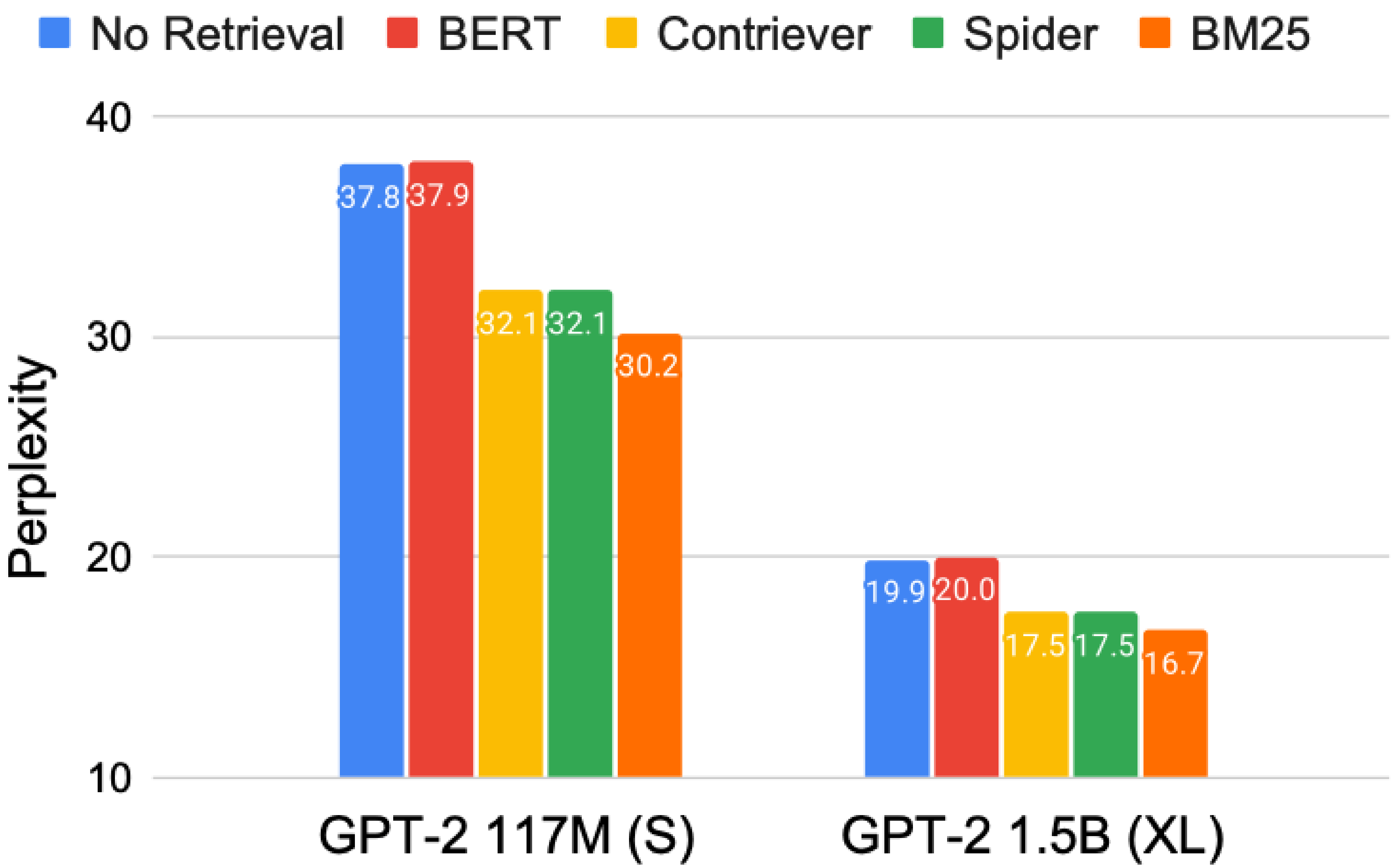
Retrieval models and language models are trained **independently**

- Training language models

- Training retrieval models



RAG with LMs using different retrievers



Better retrieval model

Better base LMs

→ Better **retrieval-based LMs**

Each component can be improved separately

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

Independent Training

- Work with off-the-shelf models (no extra training required)
- Each part can be improved independently

Independent Training

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

Training methods for retrieval-augmented LMs

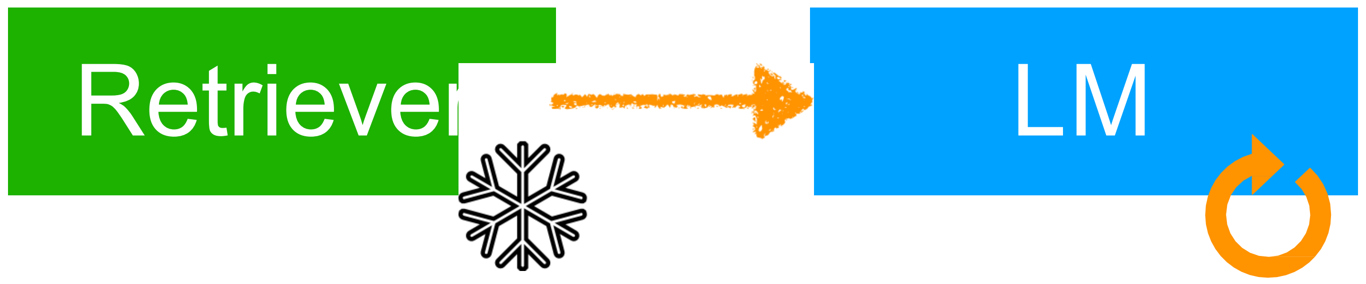
- Independent training
- **Sequential training**
- Joint training

Sequential Training

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

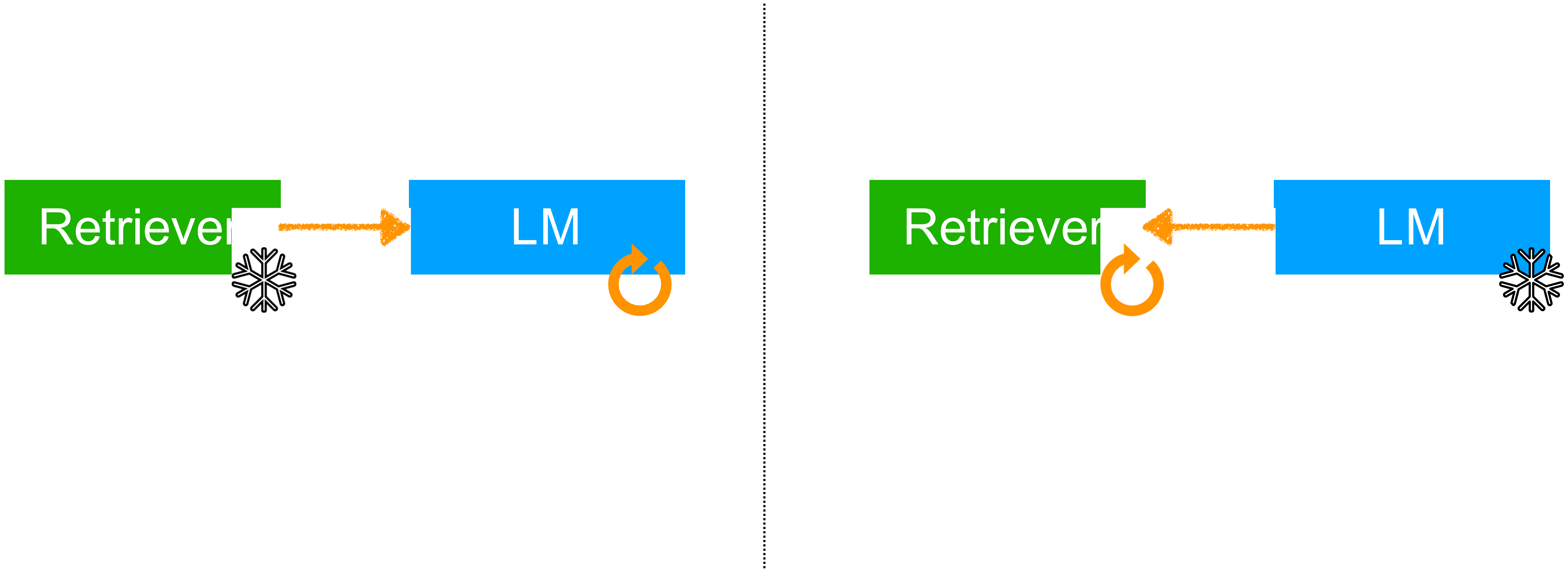
Sequential Training

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



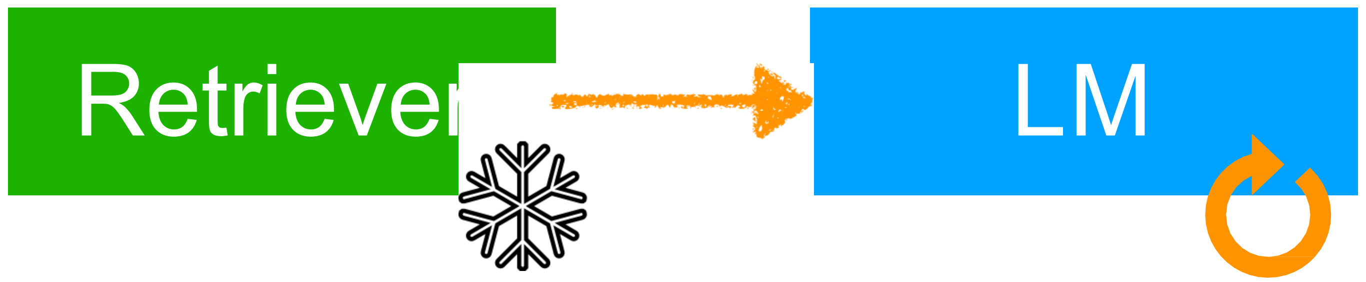
Sequential Training

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



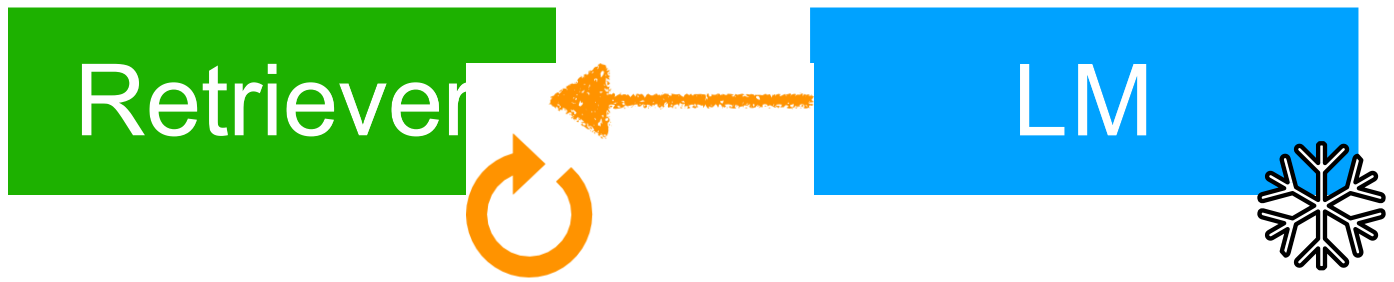
Sequential Training

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



RETRO (Borgeaud et al., 2021)

"Improving language models by retrieving from trillions of tokens"



REPLUG (Shi et al., 2023)

REPLUG: Retrieval-Augmented Black-Box Language Models

Sequential Training

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

Sequential Training

- Work with off-the-shelf components (either a large index or a powerful LM)
- LMs are trained to effectively leverage retrieval results.
- Retrievers are trained to provide text that helps LMs the most.

Let's jointly train retrieval models and LMs!

Training methods for retrieval-augmented LMs

- Independent training
- Sequential training
- **Joint training**

Joint Training

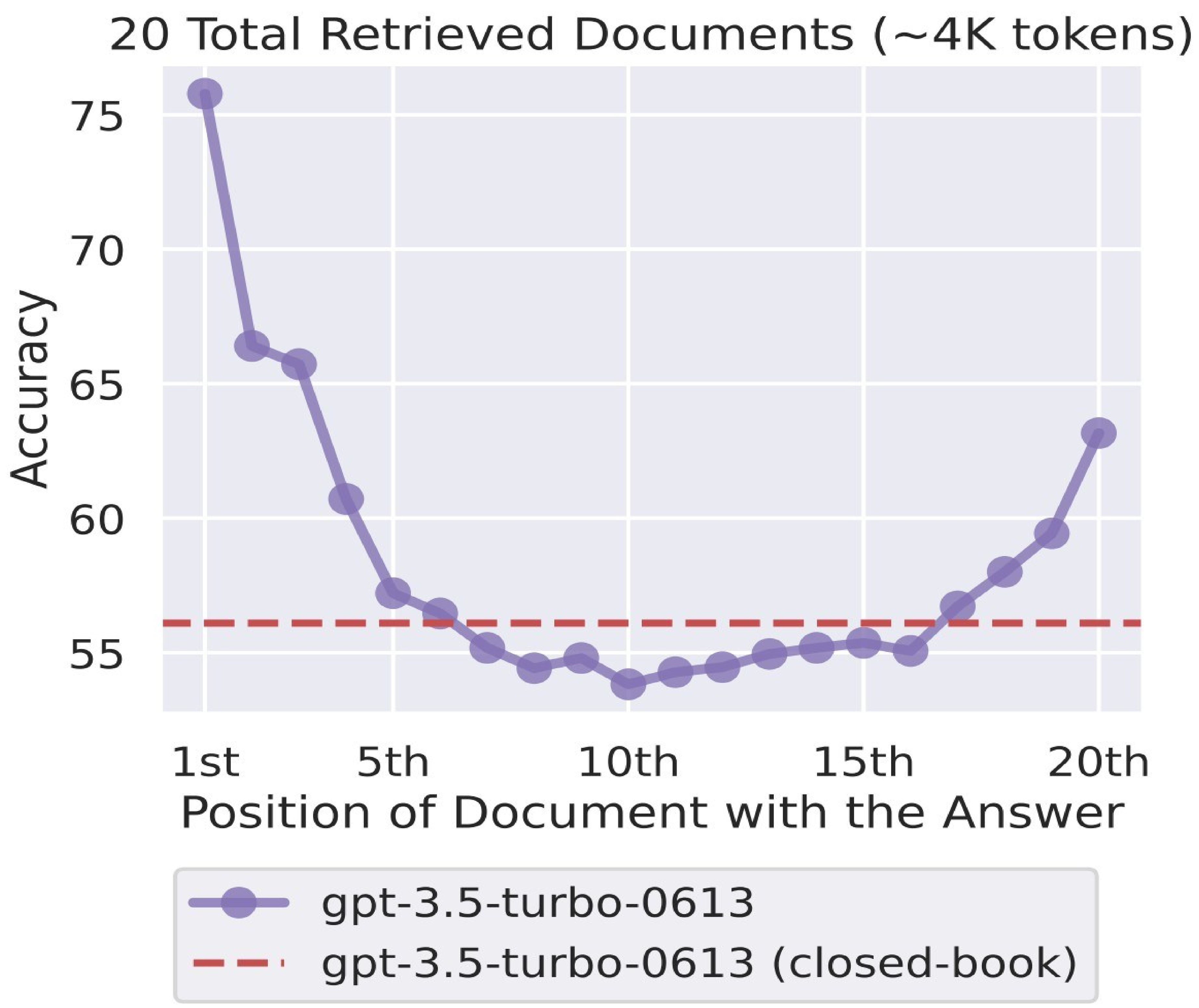
- End-to-end trained — each component is optimized
- Good performance
- Training is more complicated
(async update, overhead, data batching, etc)
- Train-test discrepancy still remains

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs – *close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*

Lost in the Middle!

- As Context Increases, Models Miss Relevant Info
- “lost-in-the- middle” (Liu et al. 2023) demonstrates that models pay less attention to things in the middle of context windows



Retrieval-augmented LMs can still hallucinate

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

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

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

Cited Webpages

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

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

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

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

124

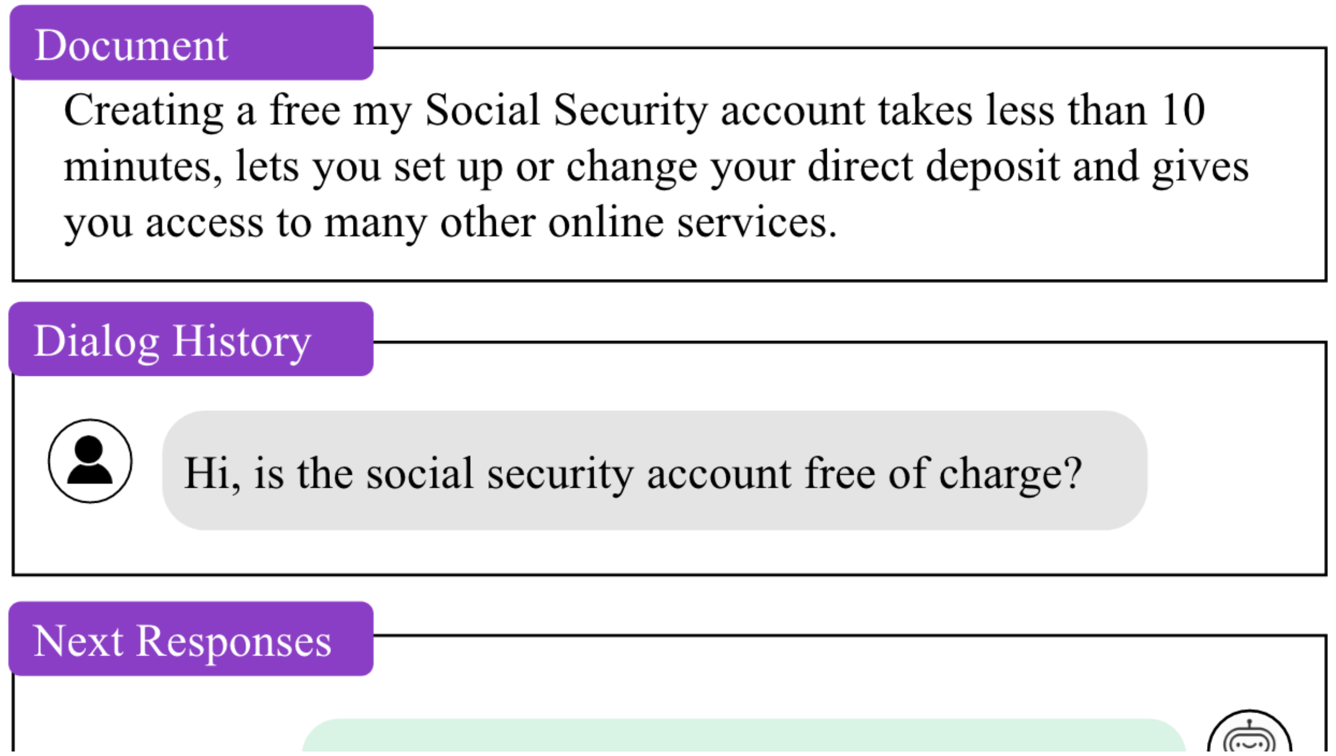
Quantifying Hallucination

Pointwise Mutual Information Based Metric and Decoding Strategy for Faithful Generation in Document Grounded Dialogs

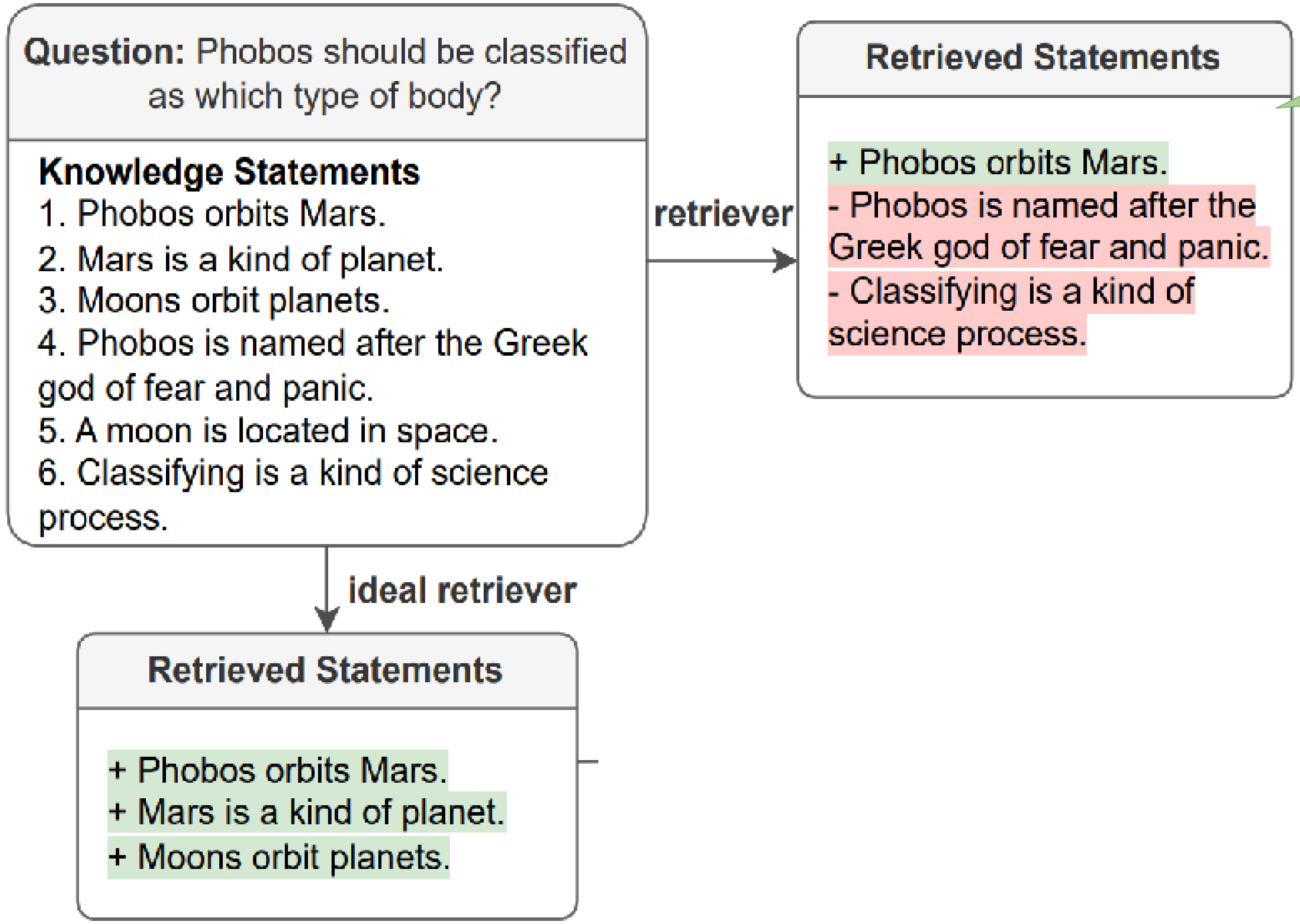
Yatin Nandwani, Vineet Kumar, Dinesh Raghu, Sachindra Joshi and **Luis A. Lastras**
IBM Research, AI
{yatin.nandwani@, vineeku6@in, diraghu1@in, jsachind@in, lastrasl@us}.ibm.com

Abstract

A major concern in using deep learning based generative models for document-grounded dialogs is the potential generation of responses that are not *faithful* to the underlying document. Existing automated metrics used for evaluating the faithfulness of response with respect to the grounding document measure the degree of similarity between the generated response and the document's content. However, these automated



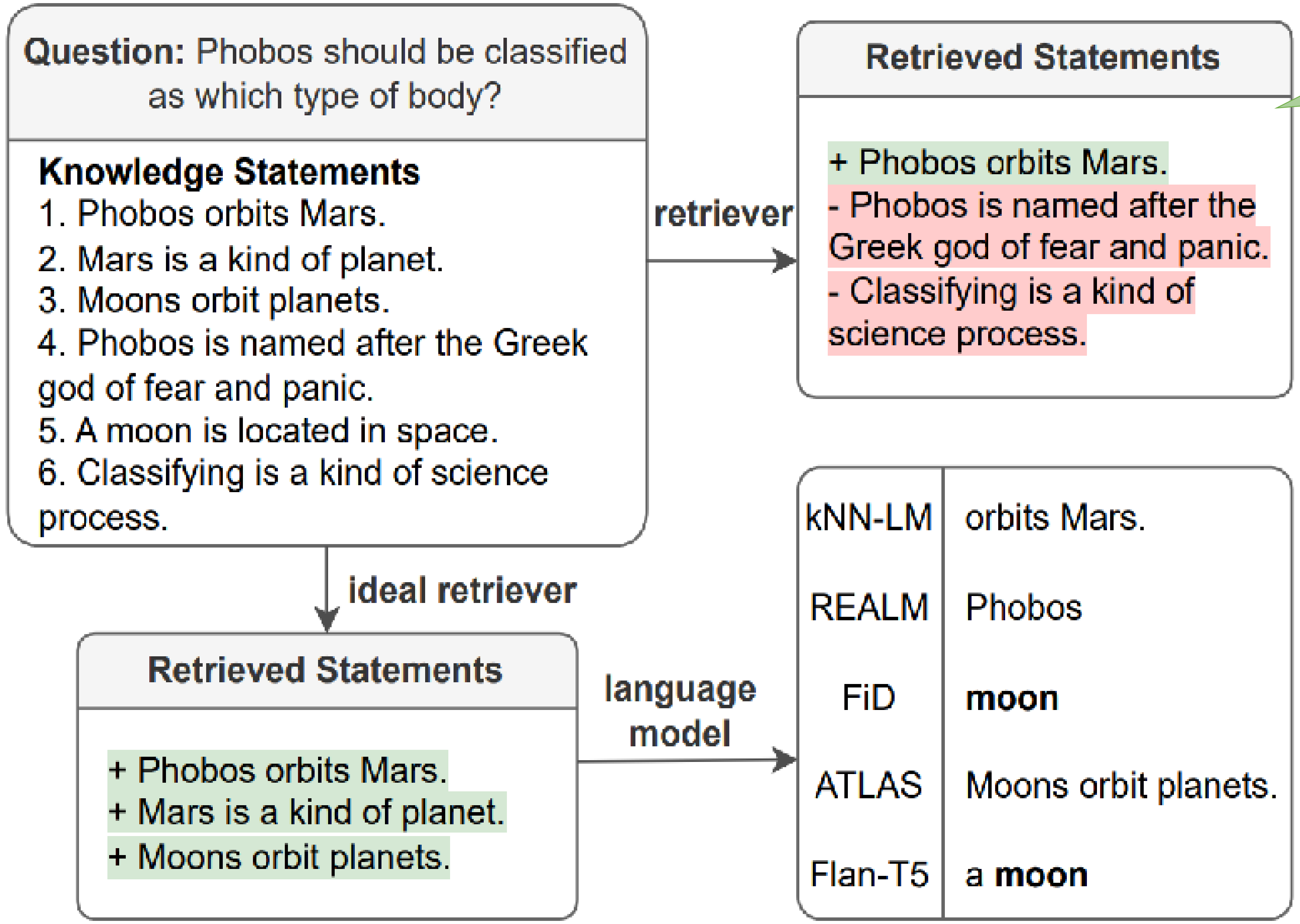
Retrieval Failures



- Retrieval fails to fetch correct information.

BehnamGhader et al. Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model. EMNLP Findings 2023.

Reasoning Failures



- Retrieval fails to fetch correct information.

- Even with ideal retriever, LM fails to give right answer.

BehnamGhader et al. Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model. EMNLP Findings 2023.

Adapt LM to Domain Corpus?

RAFT: Adapting Language Model to Domain Specific RAG

Tianjun Zhang Shishir G. Patil Naman Jain Sheng Shen Matei Zaharia Ion Stoica Joseph E. Gonzalez

tianjunz@berkeley.edu, shishirpatil@berkeley.edu

UC Berkeley

Abstract

Pretraining Large Language Models (LLMs) on large corpora of textual data is now a standard paradigm. When using these LLMs for many downstream applications, it is common to additionally bake in new knowledge (e.g., time-critical news, or private domain knowledge) into the pretrained model either through RAG-based-prompting, or finetuning. However, the optimal methodology for the model to gain such new knowledge remains an open question. In this pa-

ments). In these settings, general knowledge reasoning is less critical but instead, the primary goal is to maximize accuracy based on a given set of documents. Indeed, adapting LLMs to the specialized domains (e.g., recent news, enterprise private documents, or program resources constructed after the training cutoff) is essential to many emerging applications (Vu et al., 2023; Lazaridou et al., 2022) and is the focus of this work.

This paper studies the following question – *How to adapt pre-trained LLMs for Retrieval Augmented Generation (RAG) in specialized domains?*

Important Resources

- LangChain ; LlamaIndex – *overall frameworks*
- Lucene – *BM25 sparse retriever*
- ANNOY, FAISS, CromaDB - *dense embeddings and retrievers*
- Comprehensive RAG (CRAG) Benchmark *– KDD Cup 2024*

Outline

- Motivation
  - Drawbacks of Parametric LLMs – *hallucination, verification …*
  - Motivating Retrieval-based LLMs *– close book vs open book*
- Major components of Retrieval-based LLMs – *index, retrieve, read …*
- Retrieval Methods – *sparse, dense, reranking, black-box*
- REALM, RAG – *seminal works*
- Overview of Training Techniques – *independent, sequential, joint training ...*
- Limitations – *lost in the middle, still hallucinating, retriever failures ...*