R^2LLMs:
Retrieval and Ranking with LLMs

Guido Zuccon[1], **Shengyao** Zhuang[2], **Xueguang** Ma[3]

[1] ielab, The University of Queensland, Australia & Goole Research Australia
[2] ielab, CSIRO, Australia
[3] The University of Waterloo

https://ielab.io/tutorials/r2llms.html

Part 2: LLM-based Retrievers

LLM-based Retrievers

Overview

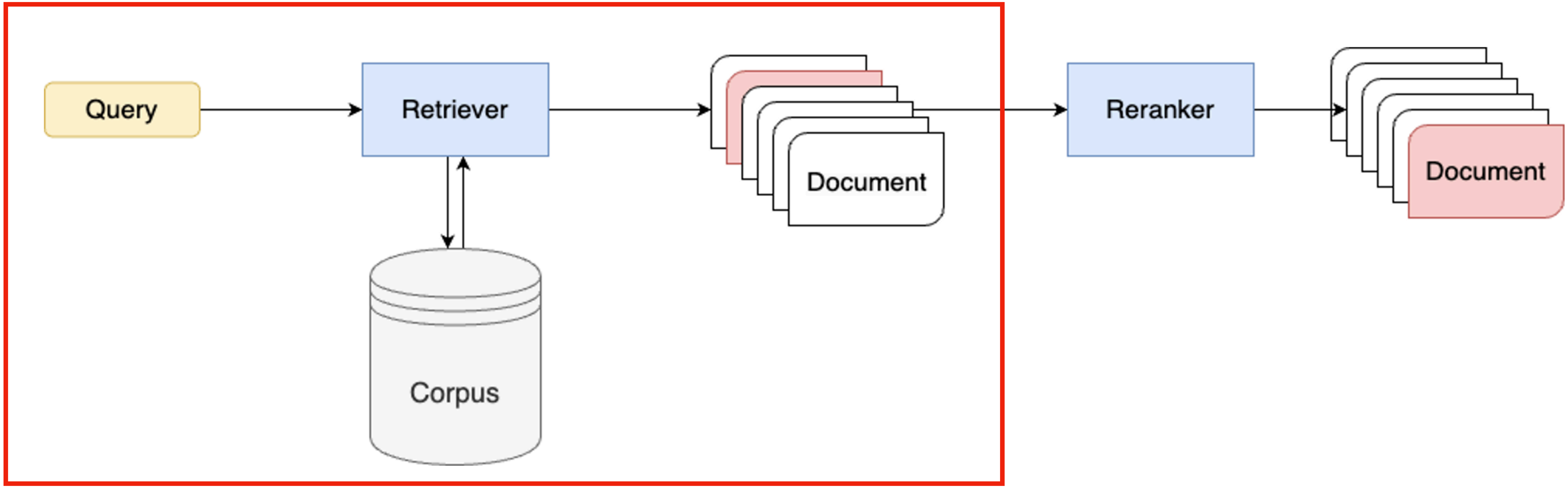
Neural Retrieval Background

- What is dense retriever
- How to train dense retriever
- How to evaluate

Large Language Model Retriever

- LLM backbone for robust dense retriever
- LLM data synthesis for general purpose embedding model
- LLM Reasoning for Deep Retrieval
- VLM for multimodal retrieval

Document Retrieval pipeline



Background

Tasks of Document Retrieval

Given

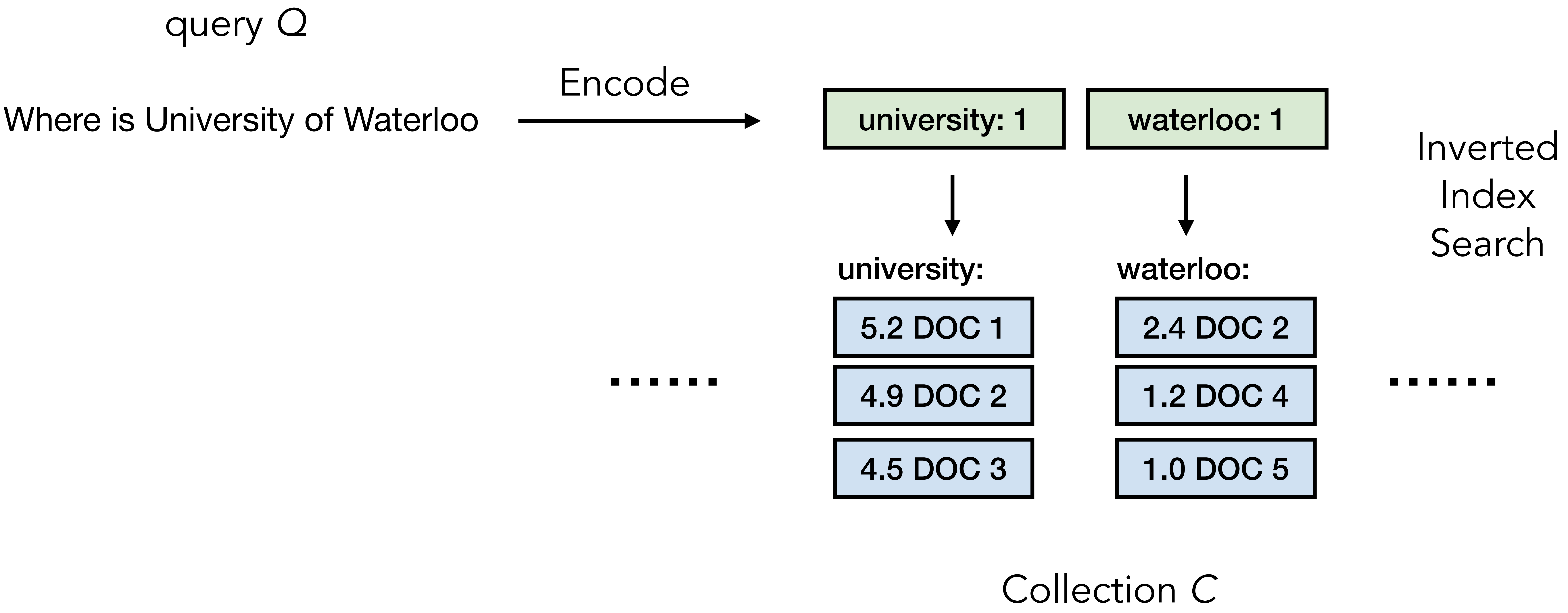
- a query *Q*;
- a corpus *C* containing *m* documents {D1, …, Dm}

The task of text retrieval is to

return *n* documents from the corpus that are *most relevant* to the query Q, where *n* ≪ *m*.

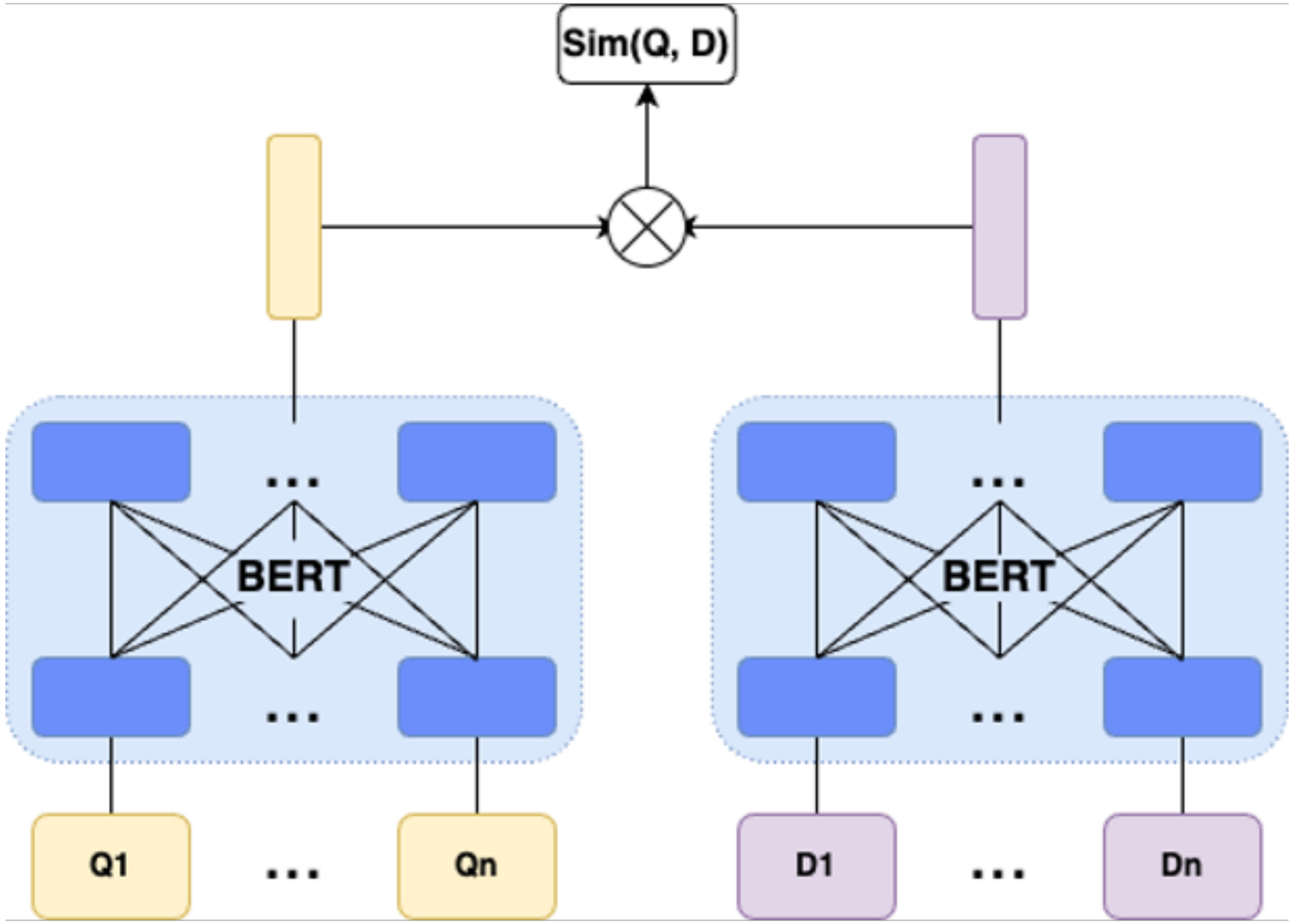
Background

Traditional Retrieval: e.g., BM25/TF-IDF



Background

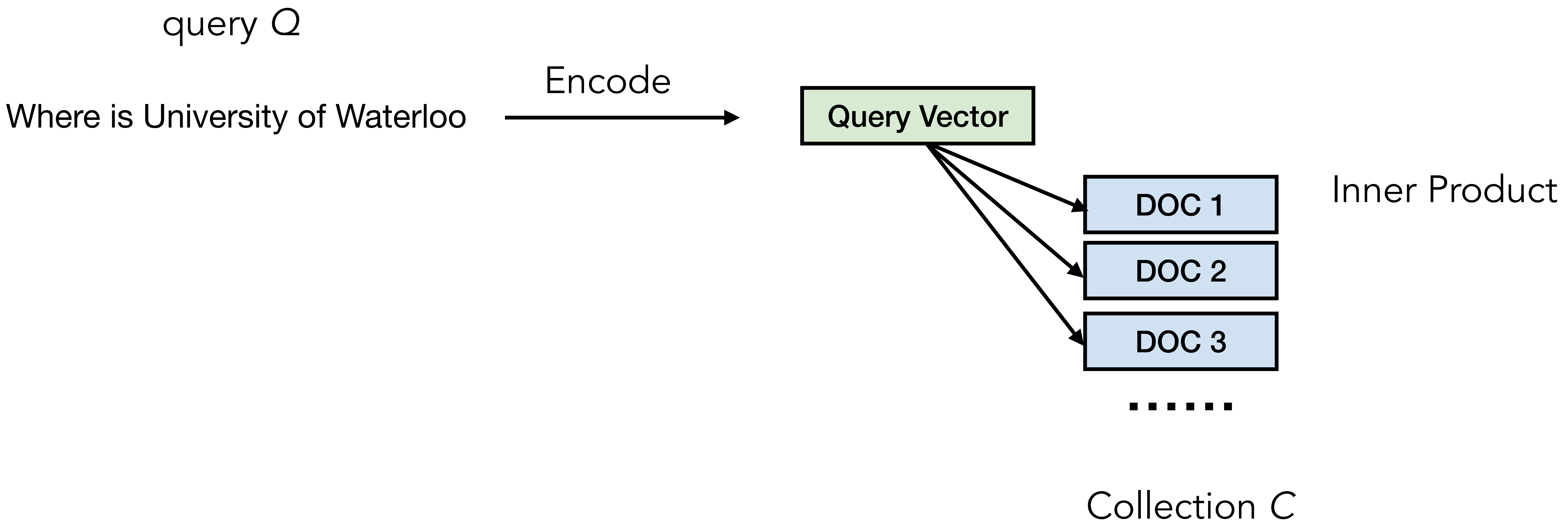
Dense Passage Retriever: DPR (Karpukhin et.al., 2020)



Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W. Dense Passage Retrieval for Open-Domain Question Answering. *EMNLP 2020.*

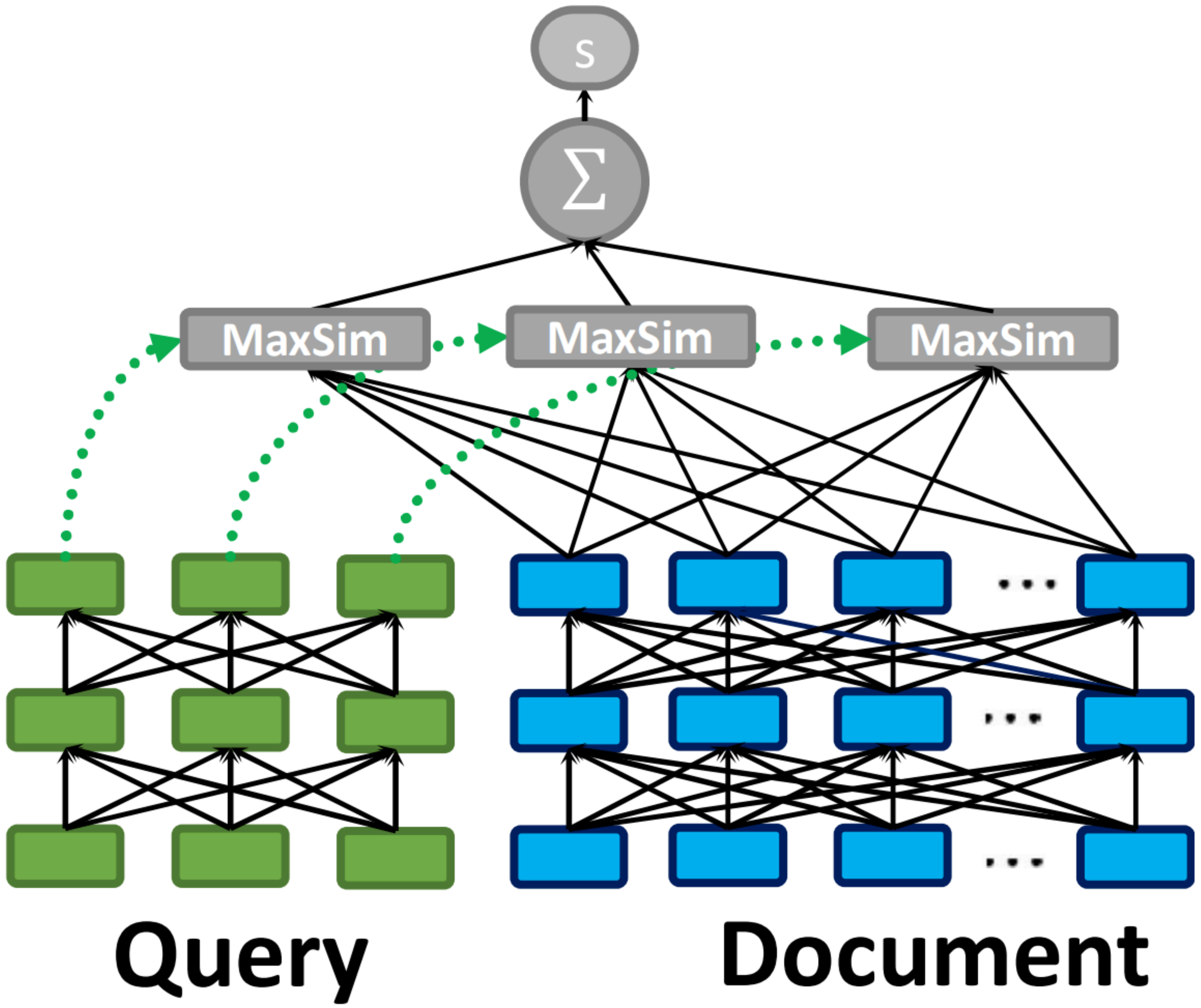
Background

Dense Passage Retriever: DPR (Karpukhin et.al., 2020)



Background

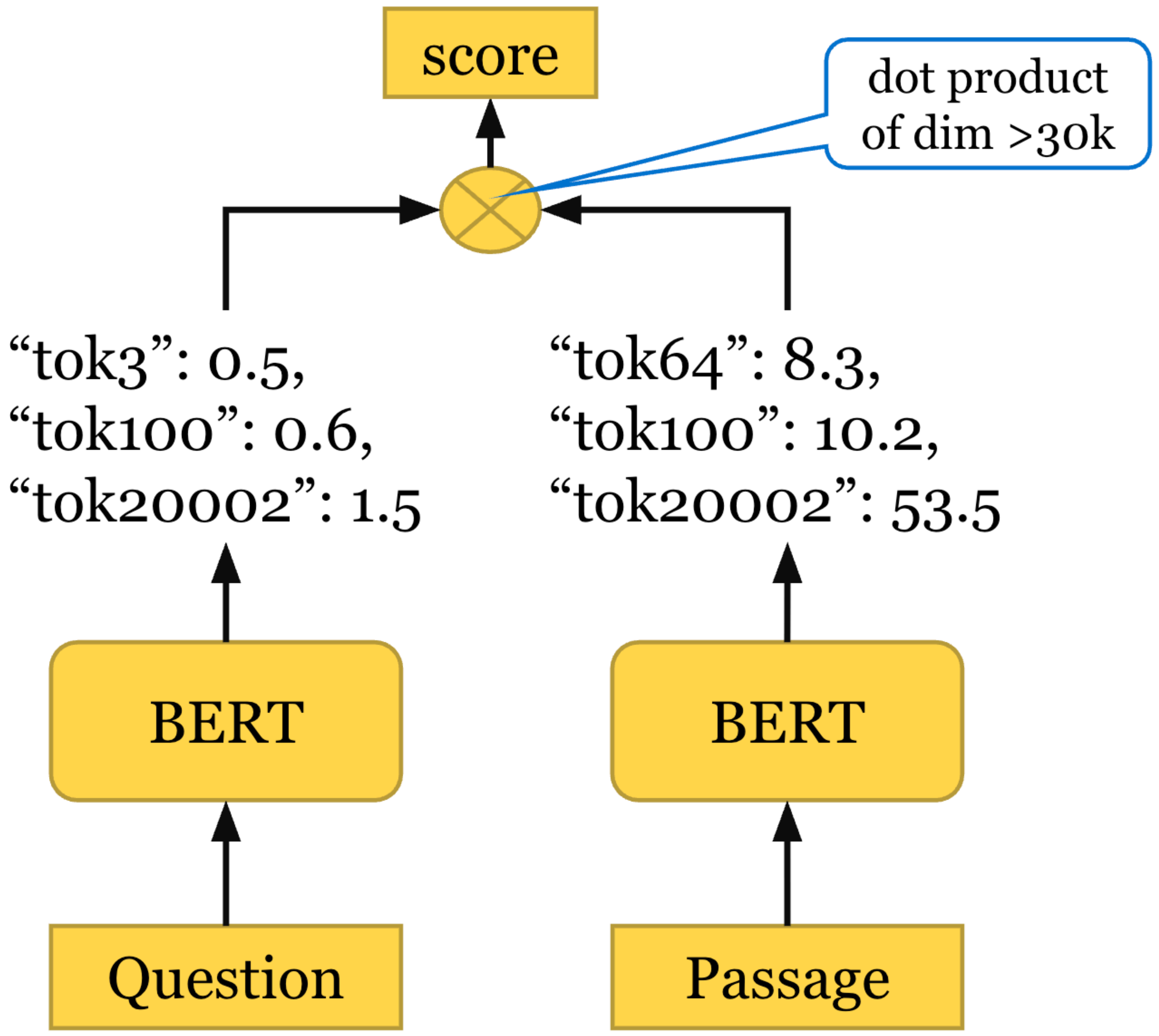
Late Interaction Retriever: ColBERT (Khattab et.al., 2020)



Khattab, O. and Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *SIGIR 2020.*

Background

Learned Sparse Retriever: DeepImpact/SPLADE/uniCOIL etc



Mallia, A., Khattab, O., Suel, T. and Tonellotto, N. Learning Passage Impacts for Inverted Indexes. *SIGIR 2021*.
Formal, T., Piwowarski, B. and Clinchant, S. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. *SIGIR 2021*.
Lin, J. and Ma, X. A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques. *arXiv preprint arXiv:2106.07545*.

$$S(q, d) = \varphi(\eta_q(q), \eta_d(d_i))$$

$\varphi(\,\cdot\,,\,\cdot\,)$ is the similarity function
$\eta(\,\cdot\,)$ is the encoding function

How to train dense retriever

InfoNCE Loss

$$\mathcal{L}(Q, D^+, \{D_N\}) = -\log p(D = D^+ \mid Q)$$
$$= -\log \frac{\exp(\mathrm{Sim}(Q, D^+))}{\exp(\mathrm{Sim}(Q, D^+)) + \sum_{D_i^- \in \{D_N\}} \exp(\mathrm{Sim}(Q, D_i^-))}$$

Classify the most relevant document from the candidates for a query

How to train dense retriever

Training Data (MS MARCO)

Query	Positive Document	Hard Negative Documents
where is whitemarsh island	[{ "docid": "5399011", "title": "Whitemarsh Island, Georgia", "text": "Whitemarsh Island, Georgia. Whitemarsh Island (pronounced WIT-marsh) is a census-designated place (CDP) in Chatham County, Georgia, United States. The population was 6,792 at the 2010 census. It is part of the Savannah Metropolitan Statistical Area. The communities of Whitemarsh Island are a relatively affluent suburb of Savannah." }]	[{ "docid": "2670040", "title": "What military strategy was used in the pacific?", "text": "the strategy of island hopping was used by the United States in the Pacific theater of world war two. Thought of by Douglas MacArthur, island hopping was a strategy that used the technique of jumping from island to island on a chain to control the chain as a whole vs attacking all the islands at once." }, { "docid": "4683145", "title": "Whakaari / White Island", "text": "For the island near Dunedin, see White Island, Otago. Whakaari/White Island is an active andesite stratovolcano, situated 48 km (30 mi) from the east coast of the North Island of New Zealand, in the Bay of Plenty." }, { "docid": "4595226", "title": "Jekyll Island", "text": "Jekyll Island, at 5,700 acres, is the smallest of Georgia's barrier islands. The island is located in Glynn County, just southeast of the city of Brunswick, south of St. Simons Island, and north of Cumberland Island.n 1886 Finney, as a representative of the newly formed Jekyll Island Club, purchased the island from the DuBignons for $125,000. Ground was broken on the clubhouse building in mid-August 1886, and the club officially opened its doors in January 1888." },

How to train dense retriever

Training Process

```
# Assume:
# - Batch size = N
# - Each sample provides: (query, positive_doc, [hard_negatives])
# - D_hard is a list of lists, each with K hard negatives
# - Model encodes queries and docs into dense vectors

for epoch in range(num_epochs):
    for Q, D_pos, D_hard in dataloader:
        # Step 1: Encode queries and positive docs
        Q_emb = model.encode(Q)                          # shape: (N, dim)
        D_pos_emb = model.encode(D_pos)                  # shape: (N, dim)

        # Step 2: Encode all hard negatives (flatten first)
        D_hard_flat = [doc for docs in D_hard for doc in docs]    # total: N * K
        D_hard_emb = model.encode(D_hard_flat)           # shape: (N*K, dim)

        # Step 3: Combine positive docs and hard negatives
        D_all_emb = concat([D_pos_emb, D_hard_emb], dim=0)   # shape: (N + N*K, dim)

        # Step 4: Compute similarity between each query and all docs
        sim_matrix = similarity(Q_emb, D_all_emb)         # shape: (N, N + N*K)

        # Step 5: Create labels (each query's correct doc is at position i)
        labels = arange(len(Q))                            # shape: (N,)

        # Step 6: Compute contrastive loss (e.g., cross-entropy over similarity)
        loss = cross_entropy(sim_matrix, labels)           # softmax over rows

        # Step 7: Backprop and update
        loss.backward()
        optimizer.step()
        optimizer.zero_grad()
```

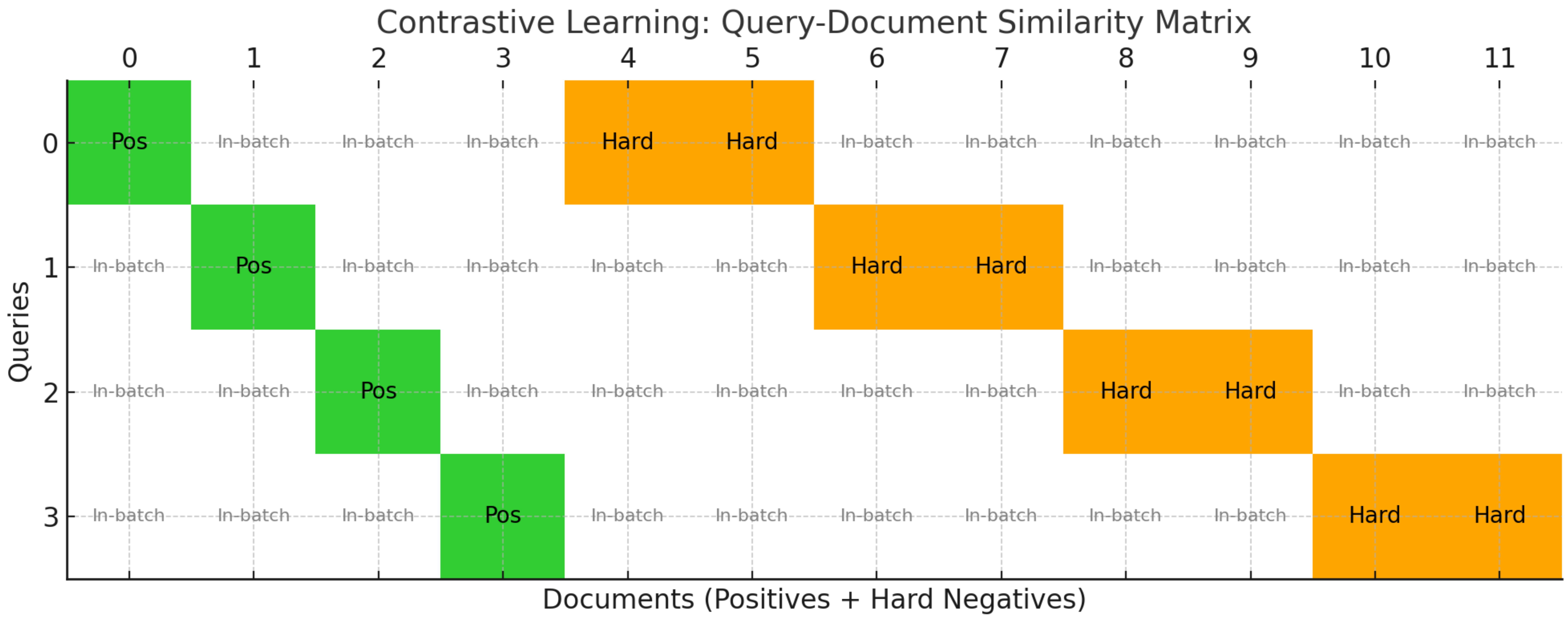
How to train dense retriever

Training Process

```
# Assume:
# - Batch size = N
# - Each sample provides: (query, positive_doc, [hard_negatives])
# - D_hard is a list of lists, each with K hard negatives
# - Model encodes queries and docs into dense vectors

for epoch in range(num_epochs):
    for Q, D_pos, D_hard in dataloader:
        # Step 1: Encode queries and positive docs
        Q_emb = model.encode(Q)                          # shape: (N, dim)
        D_pos_emb = model.encode(D_pos)                  # shape: (N, dim)

        # Step 2: Encode all hard negatives (flatten first)
        D_hard_flat = [doc for docs in D_hard for doc in docs]   # total: N * K
        D_hard_emb = model.encode(D_hard_flat)           # shape: (N*K, dim)

        # Step 3: Combine positive docs and hard negatives
        D_all_emb = concat([D_pos_emb, D_hard_emb], dim=0)   # shape: (N + N*K, dim)

        # Step 4: Compute similarity between each query and all docs
        sim_matrix = similarity(Q_emb, D_all_emb)     # shape: (N, N + N*K)

        # Step 5: Create labels (each query's correct doc is at position i)
        labels = arange(len(Q))                        # shape: (N,)

        # Step 6: Compute contrastive loss (e.g., cross-entropy over similarity)
        loss = cross_entropy(sim_matrix, labels)      # softmax over rows

        # Step 7: Backprop and update
        loss.backward()
        optimizer.step()
        optimizer.zero_grad()
```

How to train dense retriever

Training Process



How to evaluate dense retriever

Metrics	How it works	Why useful
nDCG@10 **(Normalized Discounted Cumulative Gain)**	Assigns higher weight to relevant documents at higher ranks. Discounts the gain logarithmically as rank increases.	Captures both **relevance** and **ranking position**.
MRR@10 **(Mean Reciprocal Rank)**	Calculates the reciprocal rank (1/rank) of the first relevant document.	Measures **early precision**
Recall@100	Computes the ratio of retrieved relevant documents in the top 100 to the total number of relevant documents.	Evaluates **coverage**
Top-k Accuracy	For each query, check if any of the top-*k* documents contain the **gold answer** span.	Simple for QA tasks

Background

Dense Retrieval in BERT Era

Dense retrieval is **successful in-domain** effectiveness, e.g., Natural Questions

Method	NQ (top-20 Acc.)
BM25	59.1
DPR (Karpukhin et.al., 2020)	79.4
	+34%

Example Query:

big little lies season 2 how many episodes

Example Doc:

series garnered several accolades. It received 16 Emmy Award nominations and won eight, including Outstanding Limited Series and acting awards for Kidman, Skarsgård, and Dern. The trio also won Golden Globe Awards in addition to a Golden Globe Award for Best Miniseries or Television Film win for the series. Kidman and Skarsgård also received Screen Actors Guild Awards for their performances. Despite originally being billed as a miniseries, HBO renewed the series for a second season. Production on the second season began in March 2018 and is set to premiere in 2019. All **seven** episodes are being written by Kelley

Background

Dense Retrieval in BERT Era

- We do not want a retriever just for one task
- How about zero-shot?
- E.g. apply DPR trained on Wiki to other tasks like scientific paper retrieval?

Example Query:

0-dimensional biomaterials lack inductive properties.

Example Doc:

Nanotechnologies are emerging platforms that could be useful in measuring, understanding, and manipulating stem cells. Examples include magnetic nanoparticles and quantum dots for stem cell labeling and in vivo tracking; nanoparticles, carbon nanotubes, and polyplexes for the intracellular delivery of genes/ oligonucleotides and protein/peptides; and engineered nanometer-scale scaffolds for stem cell differentiation and transplantation. This review examines the use of nanotechnologies for stem cell tracking, differentiation, and transplantation. We further discuss their utility and the potential concerns regarding their cytotoxicity.

BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Thakur et. al., 2021

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *NeurIPS 2021*.

Background

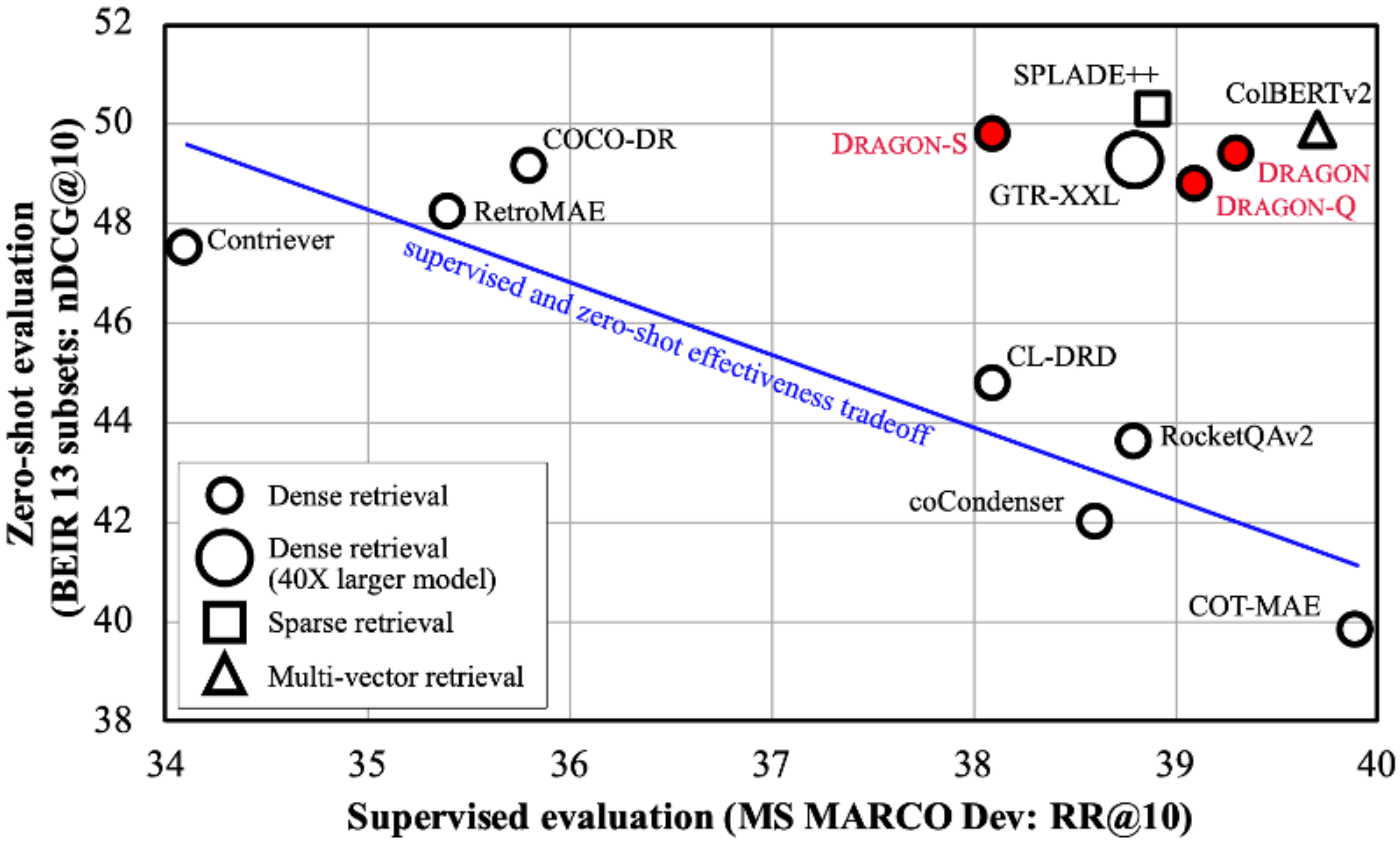
Dense retrieval **struggles** in **out-of-domain** generalization

Method	NQ (top-20 Acc.)	BEIR (nDCG@10)
BM25	59.1	40.4
DPR (Karpukhin et.al., 2020)	79.4	24.1
	+34%	**-40%**

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *NeurIPS 2021*.

Efforts for Improving Zero Shot Retriever in Pre-LLM Era

(2021-early 2023)



Lin, S.-C., Asai, A., Li, M., Oguz, B., Lin, J., Mehdad, Y., Yih, W. and Chen, X. How to Train Your DRAGON: Diverse Augmentation Towards Generalizable Dense Retrieval. *Findings of EMNLP 2023*.

Motivation

- Open-source LLMs like Llama came in 2023
- LLMs are generalizable for generation tasks
- How can we use LLM to build more **effective** and **robust** retrieval system?
  - Let retrieval grows with the advancement of LLM

What Change Does LLM brings to IR?

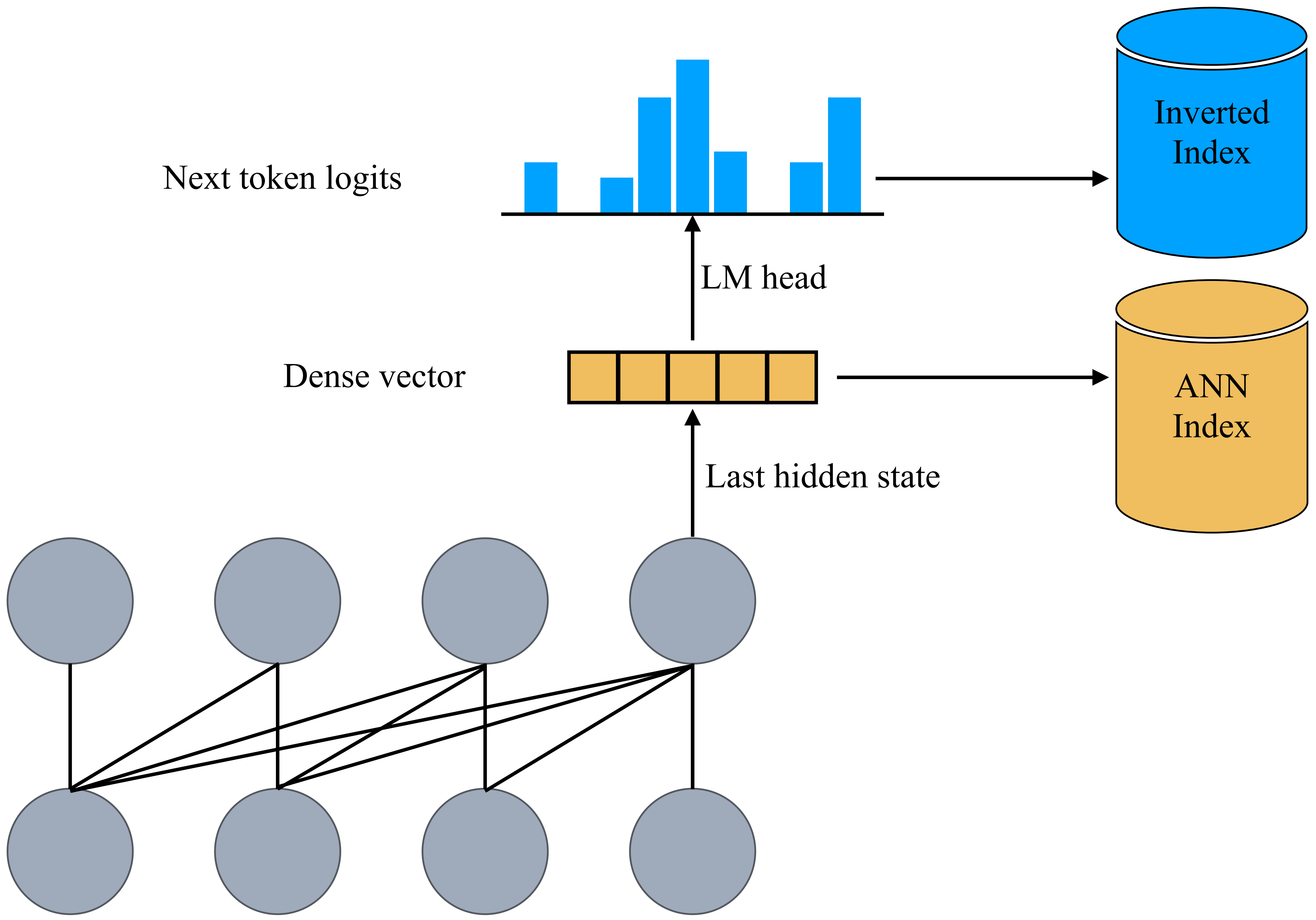
Backbone

Training Data

Observation: PromptReps

Are LLM naturally a retriever?

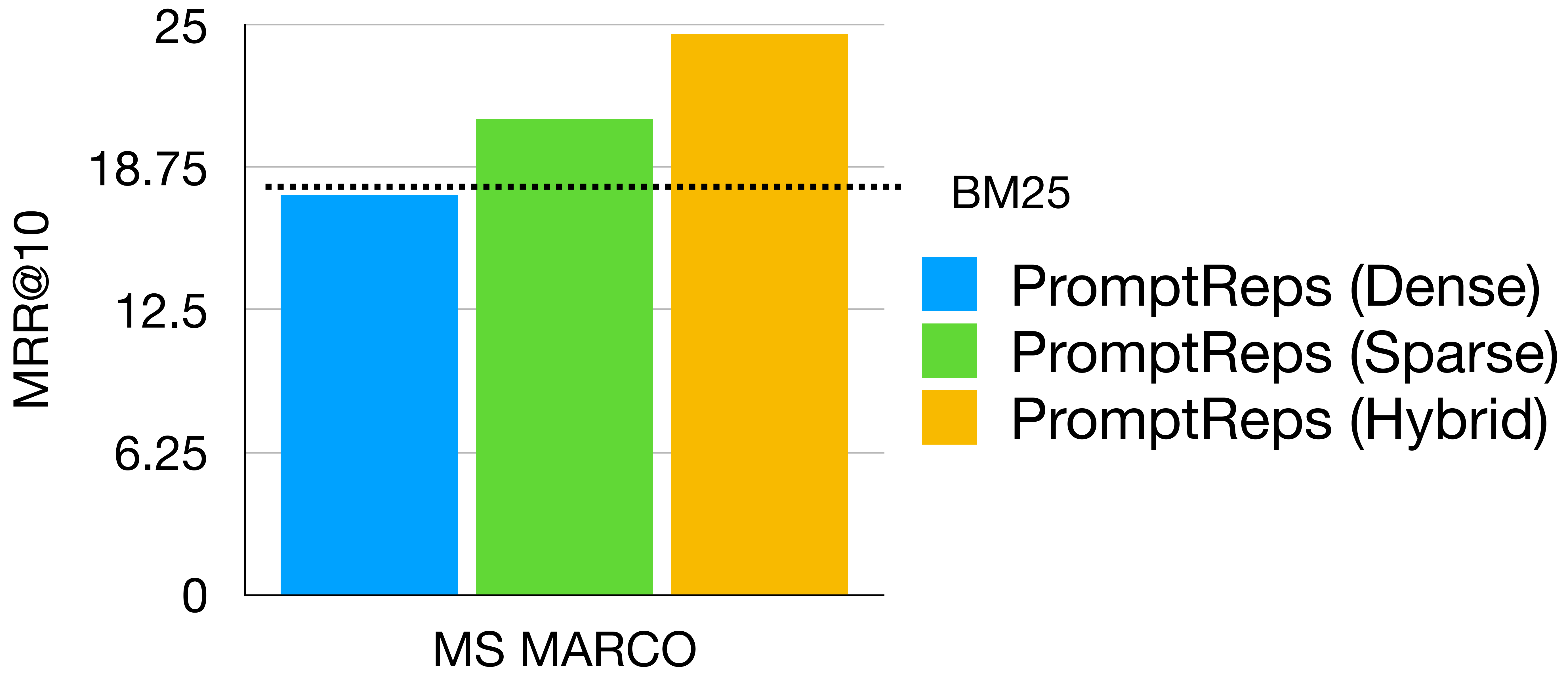
PromptReps



<User> *Passage: "***[text]***". Use one word to represent the passage in a retrieval task. Make sure your word is in lowercase.*
<Assistant> *The word is: "*

Zhuang, S., Ma, X., Koopman, B., Lin, J. and Zuccon, G., 2024. PromptReps: Prompting Large Language Models to Generate Dense and Sparse Representations for Zero-Shot Document Retrieval. *arXiv preprint arXiv:2404.18424*.

PromptReps

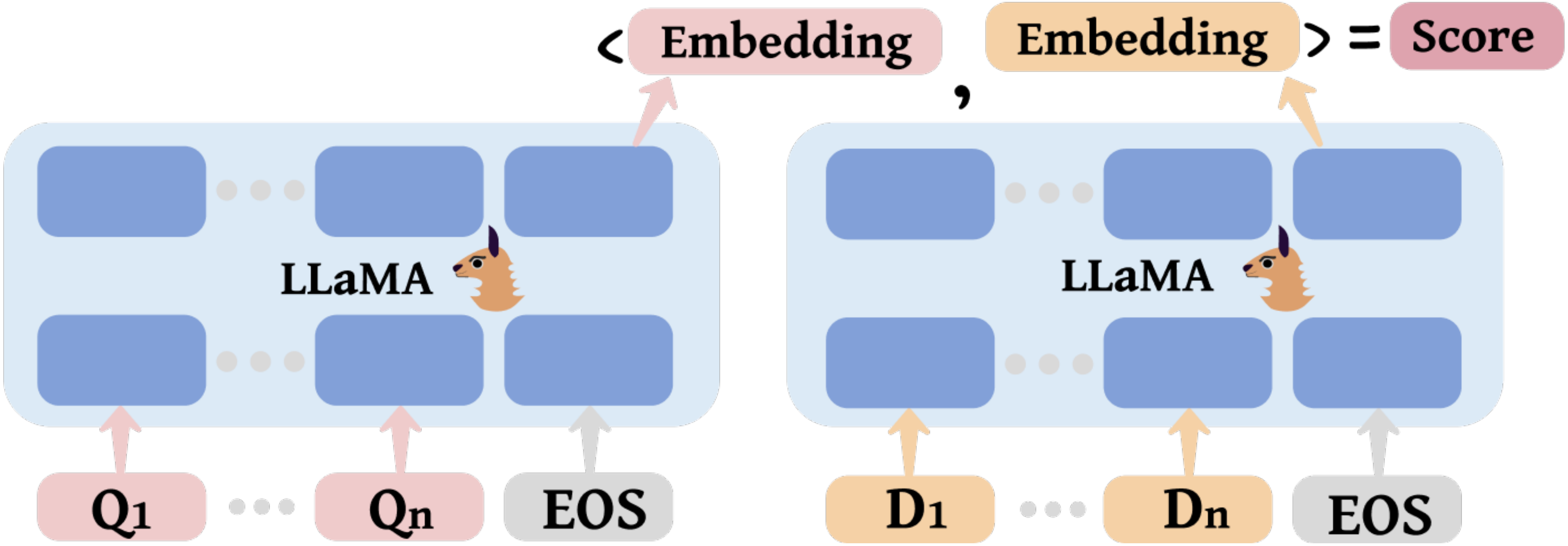


Combining Dense & Sparse representation allows effective zero-shot retrieval (no training of representations)

Zhuang, S., Ma, X., Koopman, B., Lin, J. and Zuccon, G., 2024. PromptReps: Prompting Large Language Models to Generate Dense and Sparse Representations for Zero-Shot Document Retrieval. *arXiv preprint arXiv:2404.18424*.

What if we directly fine-tune LLM as dense retriever?

RepLlama: Fine-Tuning LLaMA for Dense Retrieval



Ma, X., Wang, L., Yang, N., Wei, F. and Lin, J. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. *SIGIR 2024*.

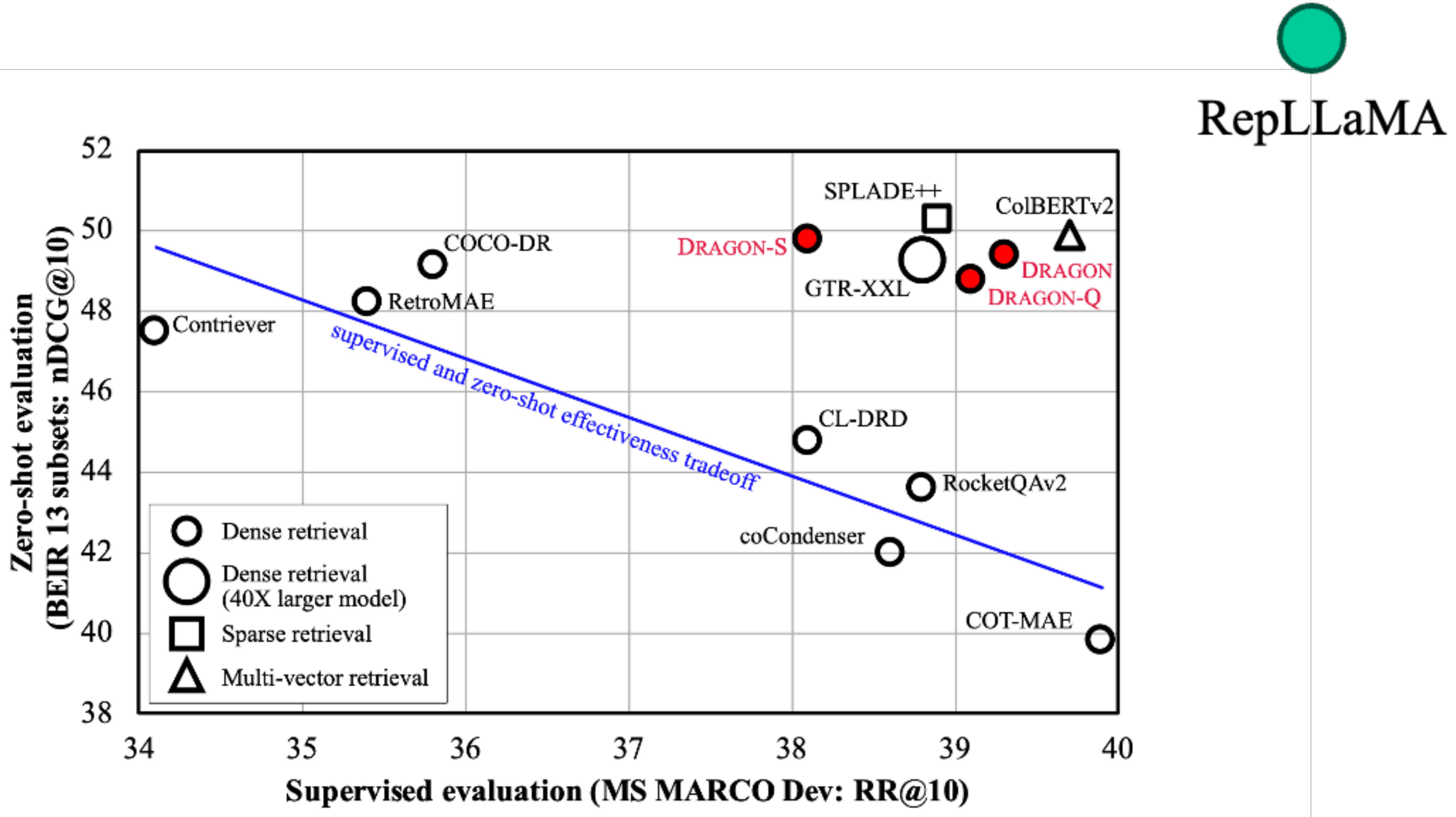
RepLlama: Fine-Tuning LLaMA for Dense Retrieval

Technical Details

- Directly replace original dense retrieval paradigm with decoder-only LLM (i.e. Llama2)
- Training on MS MARCO with LoRA fine-tuning
- 1 positive document paired with 1 positive document and 15 hard negative documents

Ma, X., Wang, L., Yang, N., Wei, F. and Lin, J. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. *SIGIR 2024*.

RepLLaMA v.s. Pre-LLM-Era Neural Retrievers on Zero-Shot Evaluation

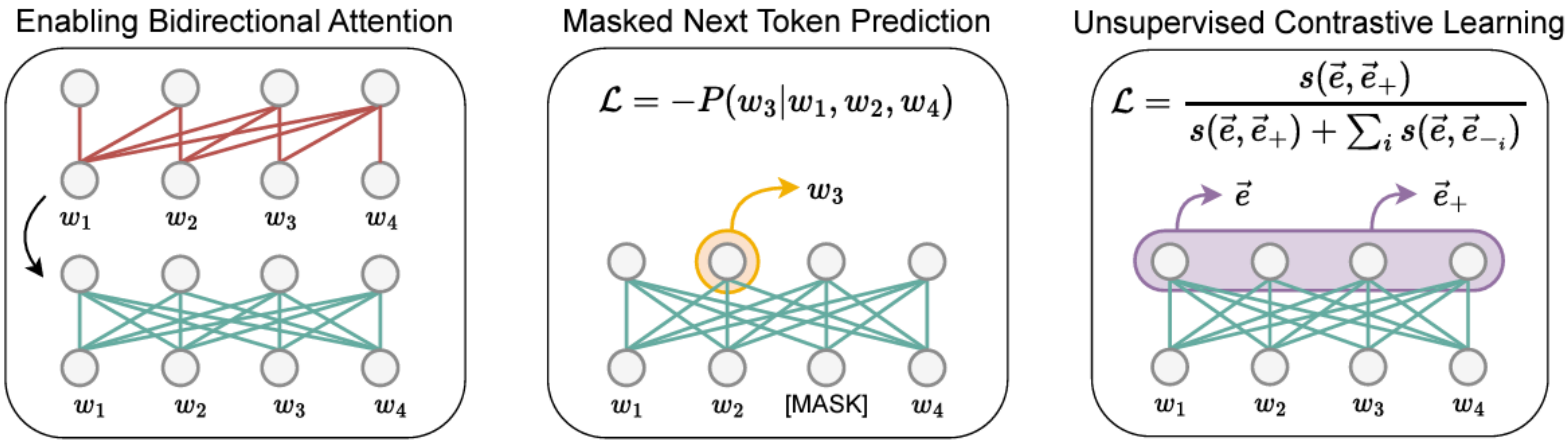


Ma, X., Wang, L., Yang, N., Wei, F. and Lin, J. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. *SIGIR 2024*.

LLM retriever is more robust than BERT-era retriever

Is decoder-only LLM (sub-)optimal for encoding task?

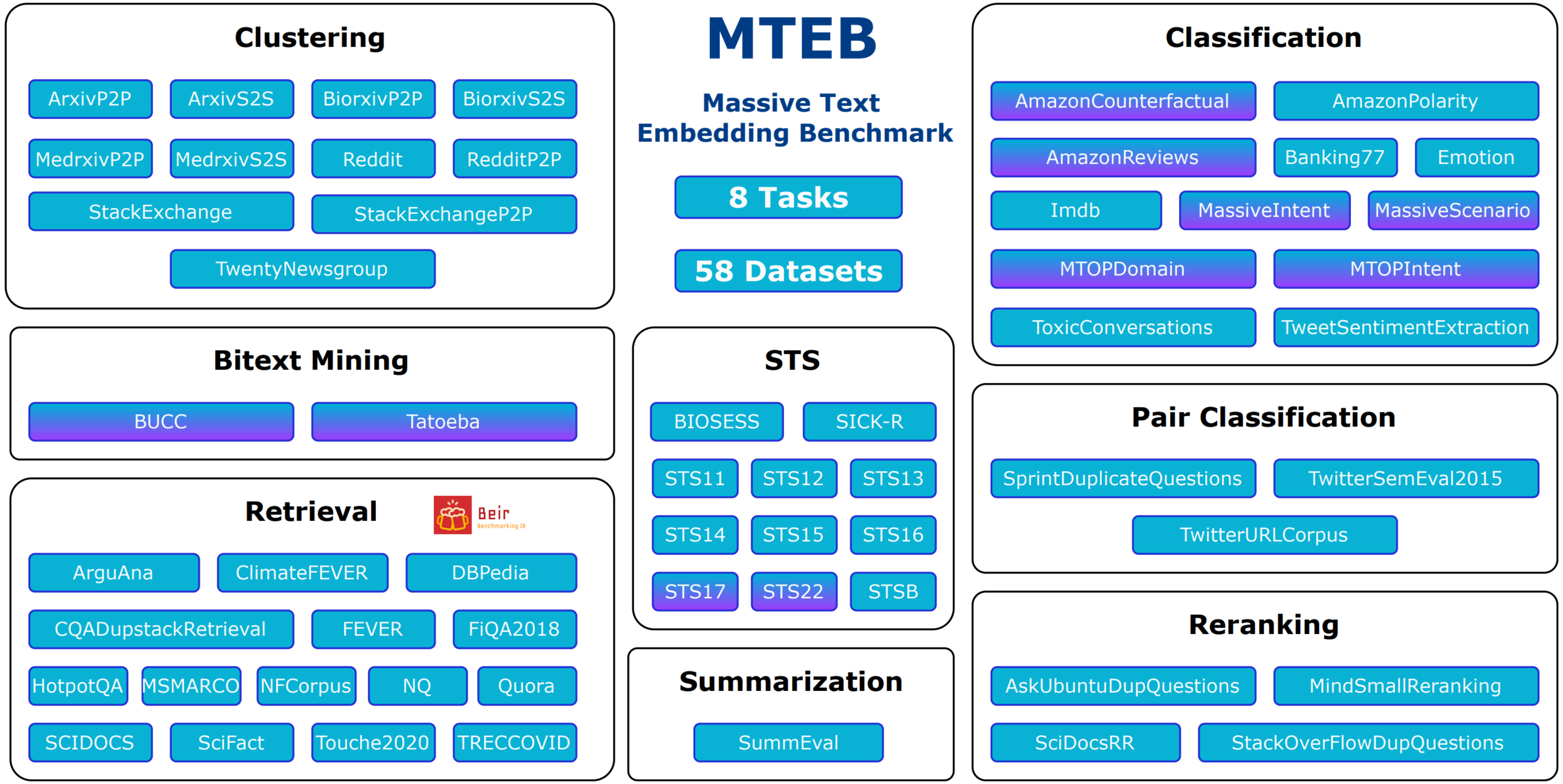
LLM2Vec: Make Decoder model Strong Text Encoder



BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N. and Reddy, S., 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

LLM retriever is robust across domains.

Can we make it generalizable across embedding tasks?



BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Muennighoff et. al., 2022

Muennighoff, N., Tazi, N., Magne, L. and Reimers, N. MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316.*

Towards General Purpose Text Embedding

Data is the Key

- The “query intent” can be different for same query.
- Diverse data collection
  - E.g BGE embedding

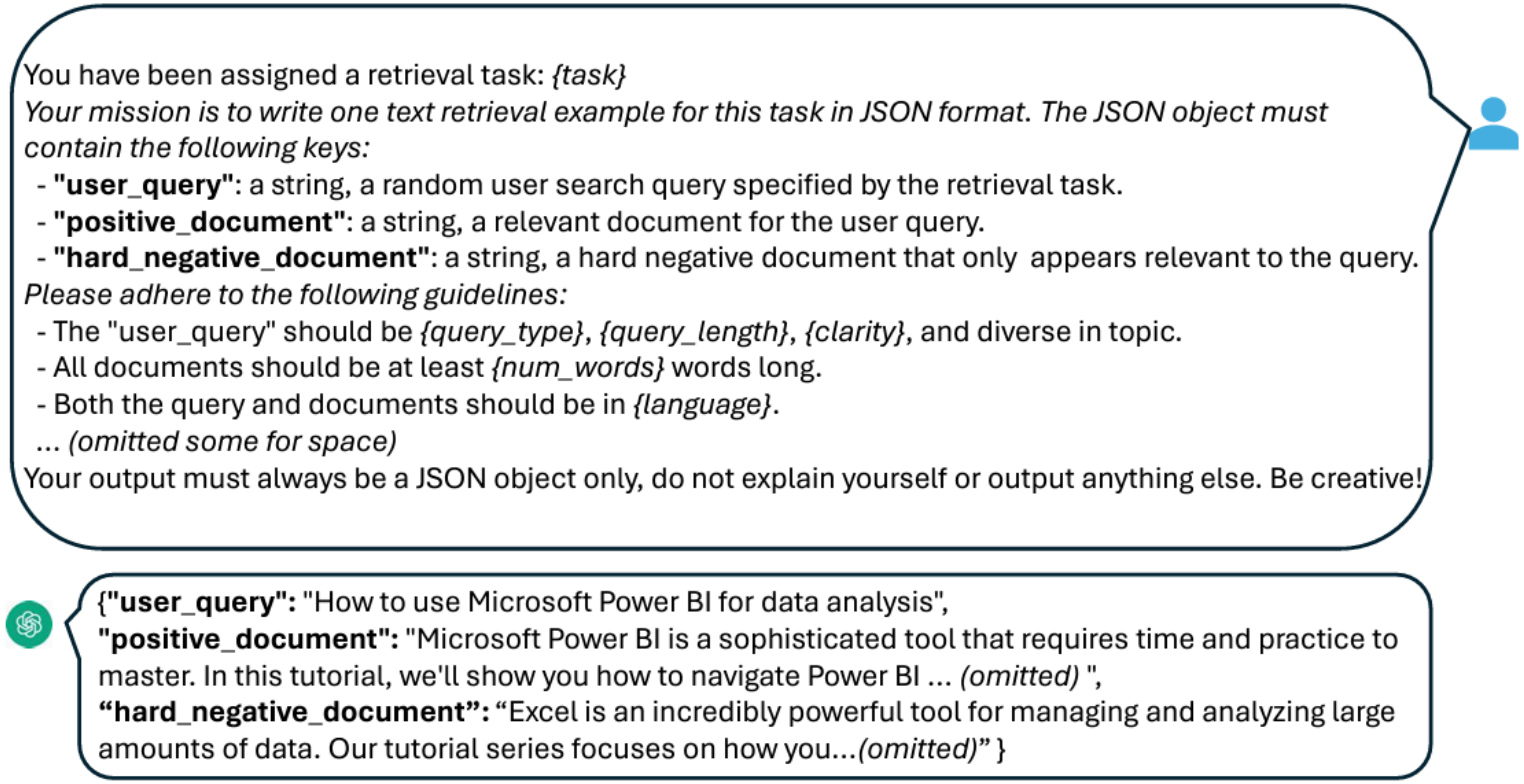
Towards General Purpose Text Embedding

Data is the Key

- Synthetic Data Generation for LLM Retriever Training
- Query, Positive Document, Negative Documents
    1. Directly generation from LLM
    2. Use real document as seed to generate query

E5 with LLMs

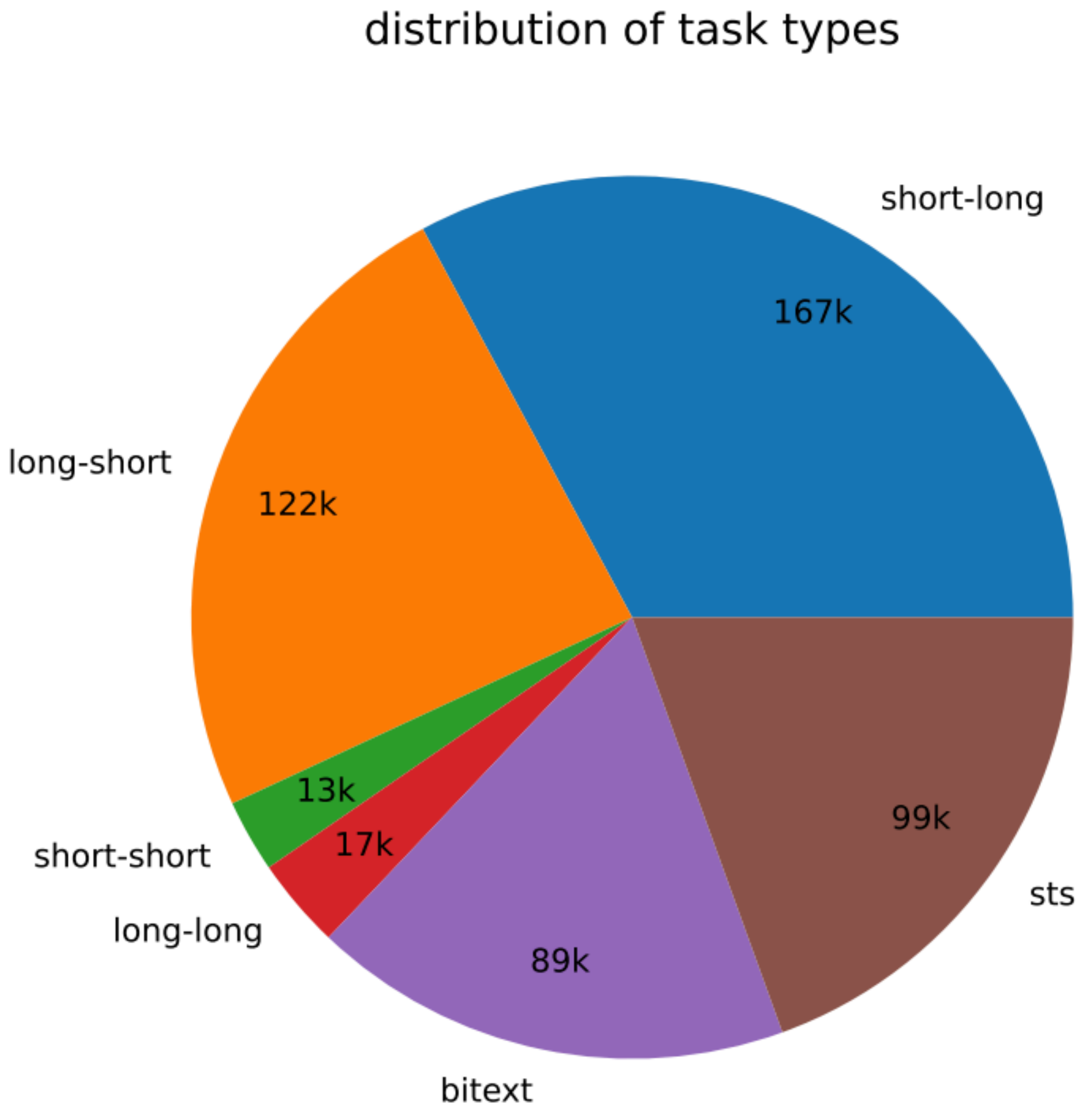
- Firstly, let LLM brainstorm retrieval tasks
- Then, let LLM generate query, positive document, hard negative document all together

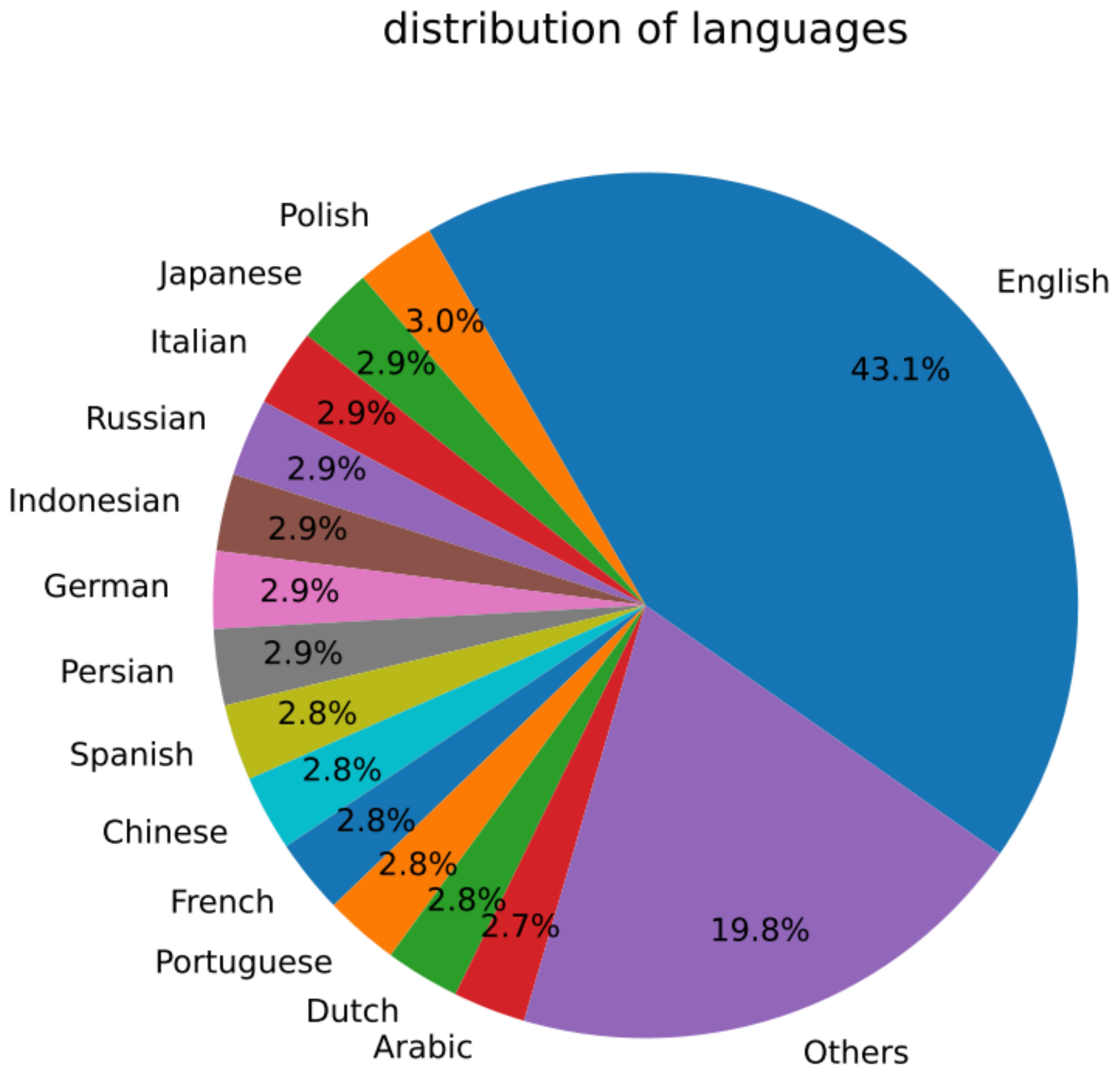


Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F., 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

E5 with LLMs

- Leverage LLM’s generation capability to create diverse embedding task training data, for multiple languages.





Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F., 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

E5 with LLMs

Use instruction in query:

"Instruct: {task_definition} \n Query: {q}"

(2) append [EOS] token to end of query and document; feed them into LLM to get embeddings by taking last layer of [EOS] vector

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F., 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

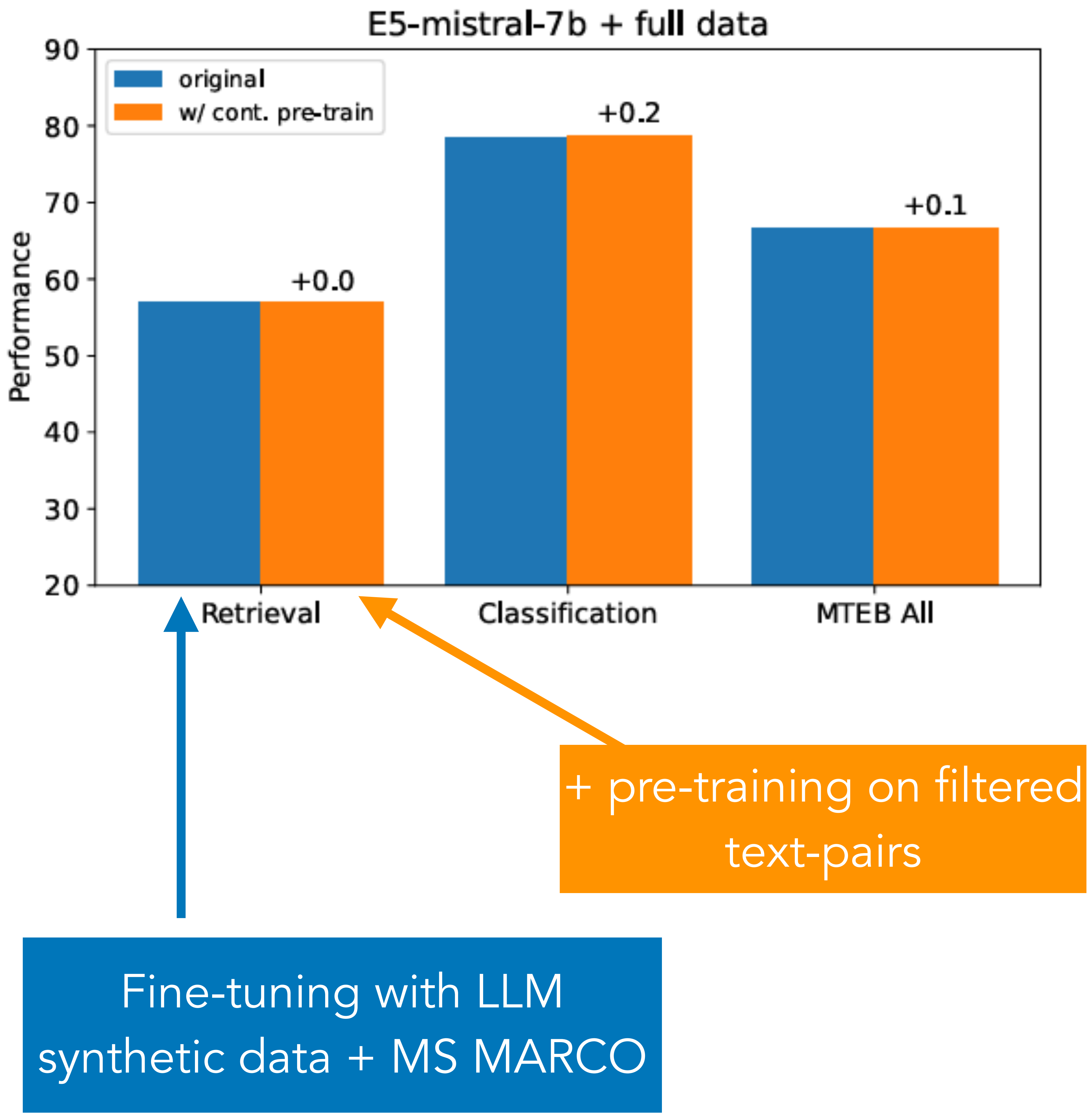
E5 with LLMs

# of datasets →	Class. 12	Clust. 11	PairClass. 3	Rerank 4	Retr. 15	STS 10	Summ. 1	Avg 56
Unsupervised Models								
Glove (Pennington et al., 2014)	57.3	27.7	70.9	43.3	21.6	61.9	28.9	42.0
$\text{SimCSE}_{\text{bert-unsup}}$ (Gao et al., 2021)	62.5	29.0	70.3	46.5	20.3	74.3	31.2	45.5
Supervised Models								
$\text{SimCSE}_{\text{bert-sup}}$ (Gao et al., 2021)	67.3	33.4	73.7	47.5	21.8	79.1	23.3	48.7
Contriever (Izacard et al., 2021)	66.7	41.1	82.5	53.1	41.9	76.5	30.4	56.0
GTR_{xxl} (Ni et al., 2022b)	67.4	42.4	86.1	56.7	48.5	78.4	30.6	59.0
$\text{Sentence-T5}_{\text{xxl}}$ (Ni et al., 2022a)	73.4	43.7	85.1	56.4	42.2	82.6	30.1	59.5
$\text{E5}_{\text{large-v2}}$ (Wang et al., 2022b)	75.2	44.5	86.0	56.6	50.6	82.1	30.2	62.3
$\text{GTE}_{\text{large}}$ (Li et al., 2023)	73.3	46.8	85.0	59.1	52.2	83.4	31.7	63.1
$\text{BGE}_{\text{large-en-v1.5}}$ (Xiao et al., 2023)	76.0	46.1	87.1	60.0	54.3	83.1	31.6	64.2
Ours								
$\text{E5}_{\text{mistral-7b}}$ + full data	**78.5**	50.3	**88.3**	**60.2**	**56.9**	**84.6**	31.4	**66.6**
w/ synthetic data only	78.2	**50.5**	86.0	59.0	46.9	81.2	31.9	63.1
w/ synthetic + msmarco	78.3	49.9	87.1	59.5	52.2	81.2	**32.7**	64.5

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F., 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

E5 with LLMs

Observation 1: contrastive *pre-*training has negligible impact on model quality.



Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F., 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368.*

E5 with LLMs

Observation 2: LoRA fine-tuning is even better than full fine-tuning.

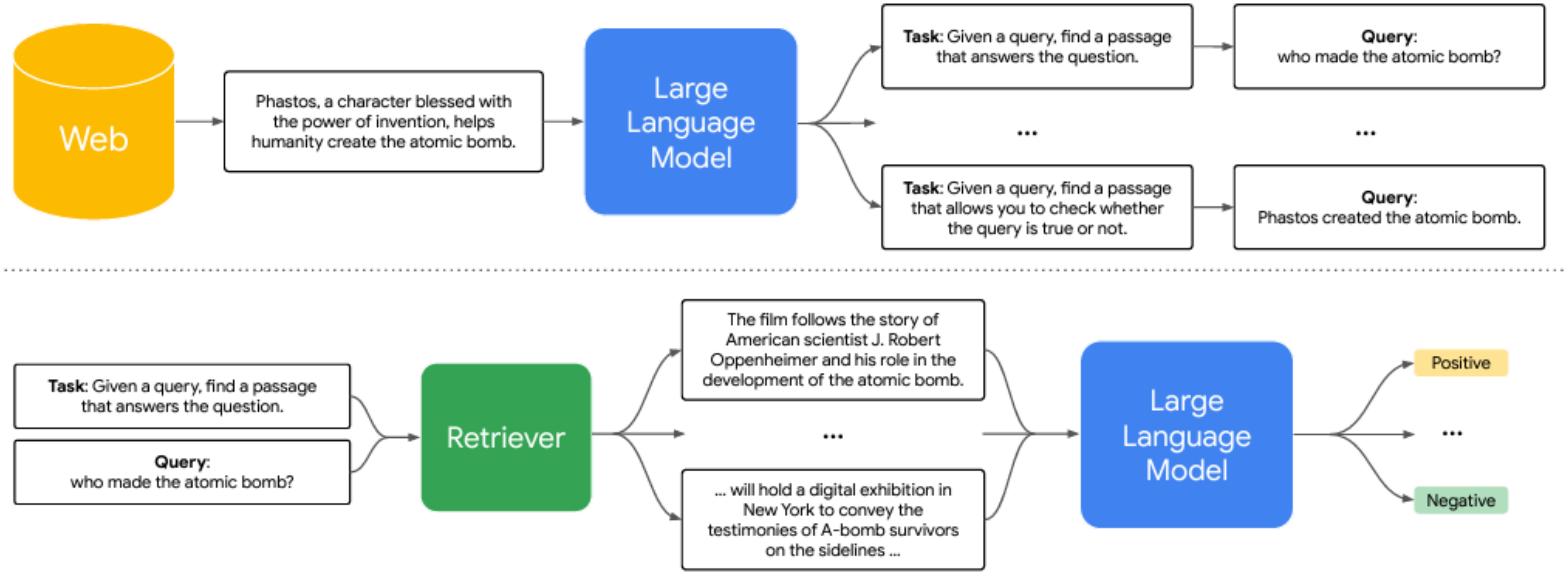
Finetuning	Retrieval nDCG@10
Full	52.2
LoRA	53.0

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. and Wei, F., 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Gecko Embeddings

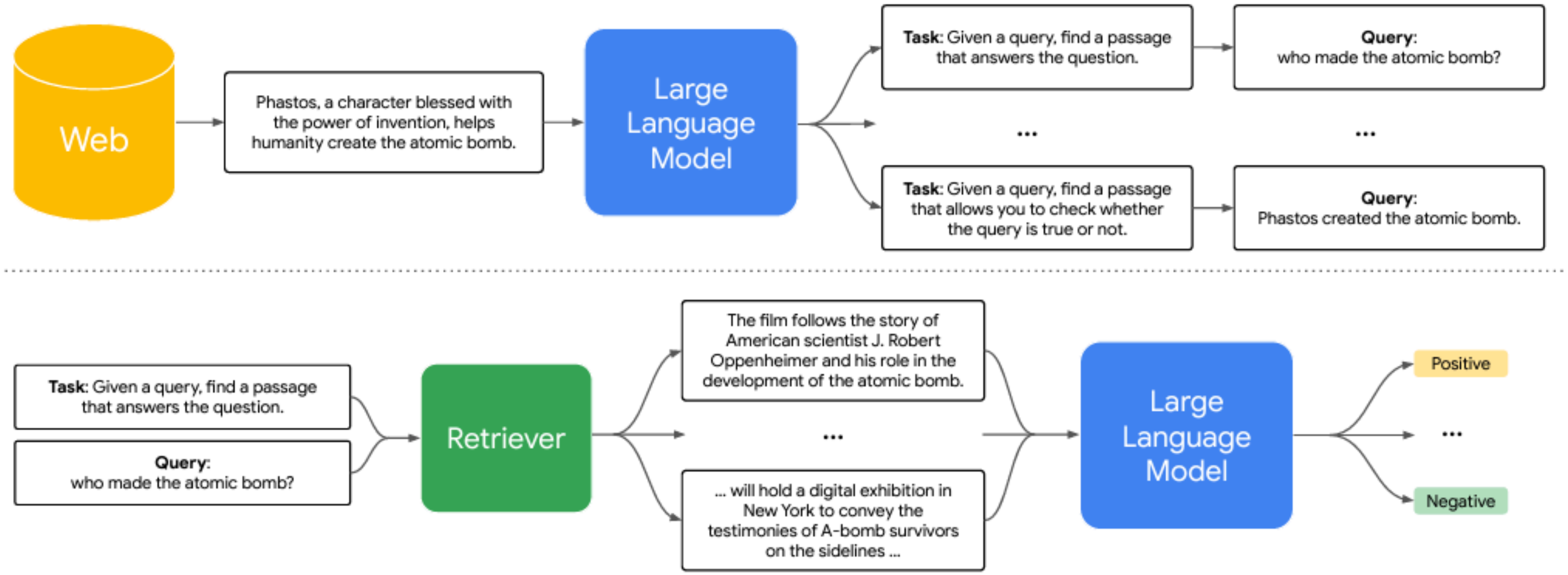
LLMs used to generate a Few-shot Prompted Retrieval dataset (FRet) for knowledge distillation from LLM into embedding through 2 tasks:

(1) diverse query generation, (2) positive & negative mining



Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J.R., Hui, K., Boratko, M., Kapadia, R., Ding, W. and Luan, Y., 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Gecko Embeddings



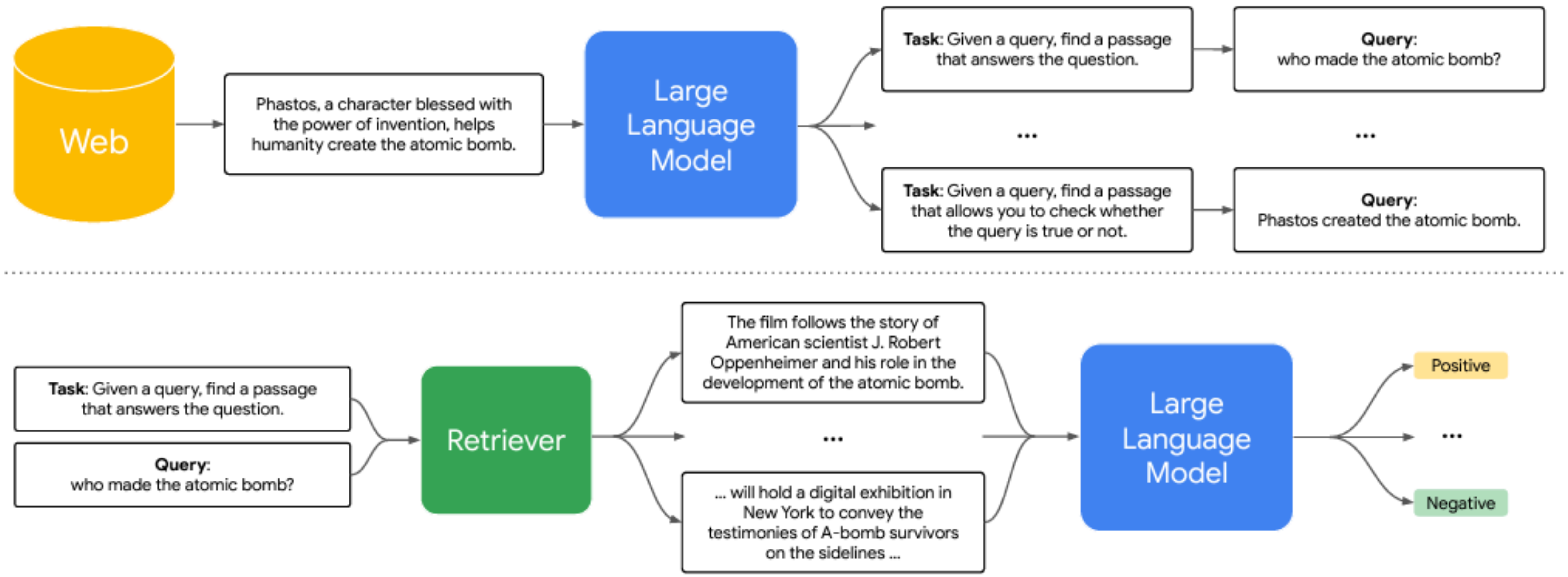
(1) Query/task generation

Prompt for task generation:

- *"Given a query, find a passage that has the answer to the query"* [question answering]
- *"Given a query, find a passage that allows you to check whether the query is true or not"* [fact checking)]

Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J.R., Hui, K., Boratko, M., Kapadia, R., Ding, W. and Luan, Y., 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Gecko Embeddings



(2) Pos/Neg Mining

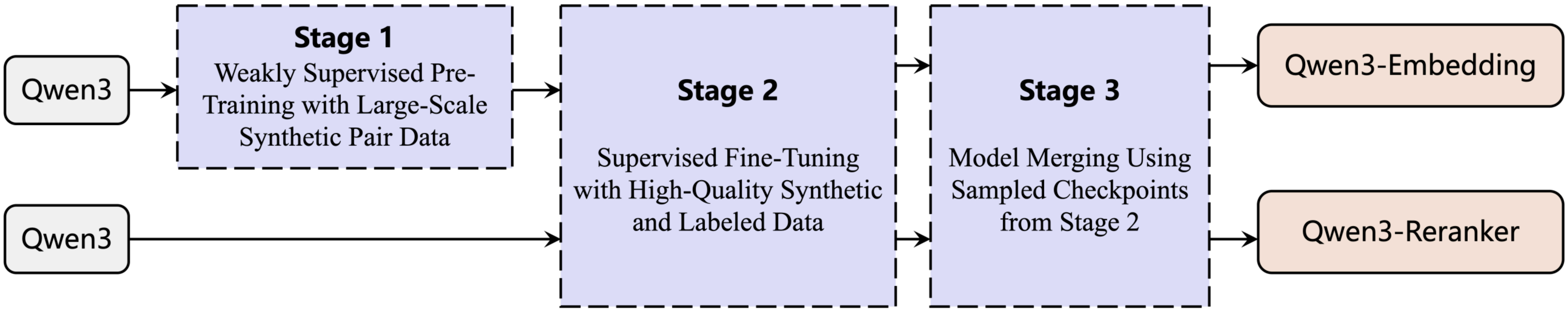
1. Use initial model trained with (q, $p$seed) pairs to retrieve top passages
2. Use an LLM to rank
3. Top ranked: p+; others p-

Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J.R., Hui, K., Boratko, M., Kapadia, R., Ding, W. and Luan, Y., 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

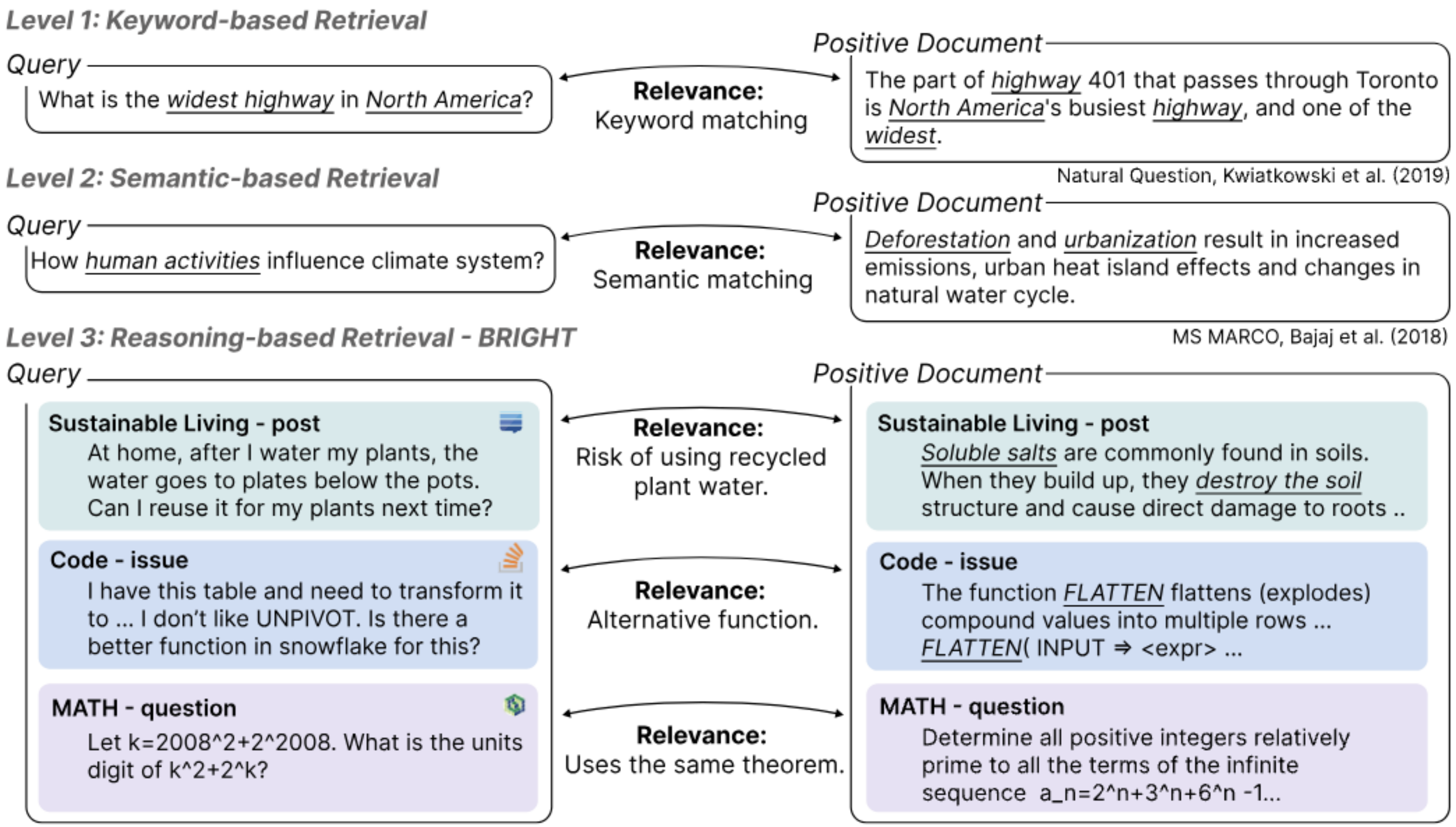
Qwen3-Embedding

Scaling synthetic training data & Multi-Stage training with synthetic data

“150 million pairs”



Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F. and Zhou, J. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.

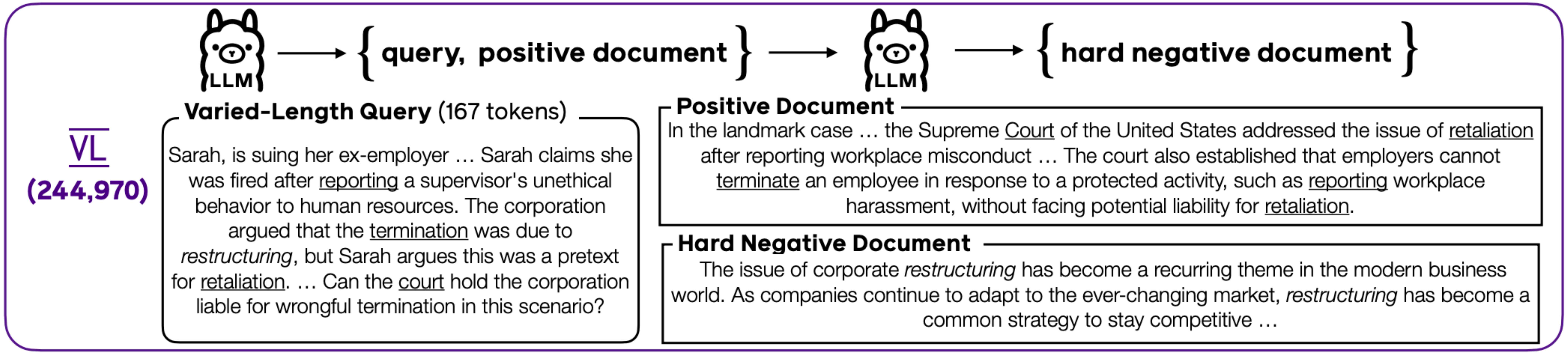


BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval

Su et. al., 2022

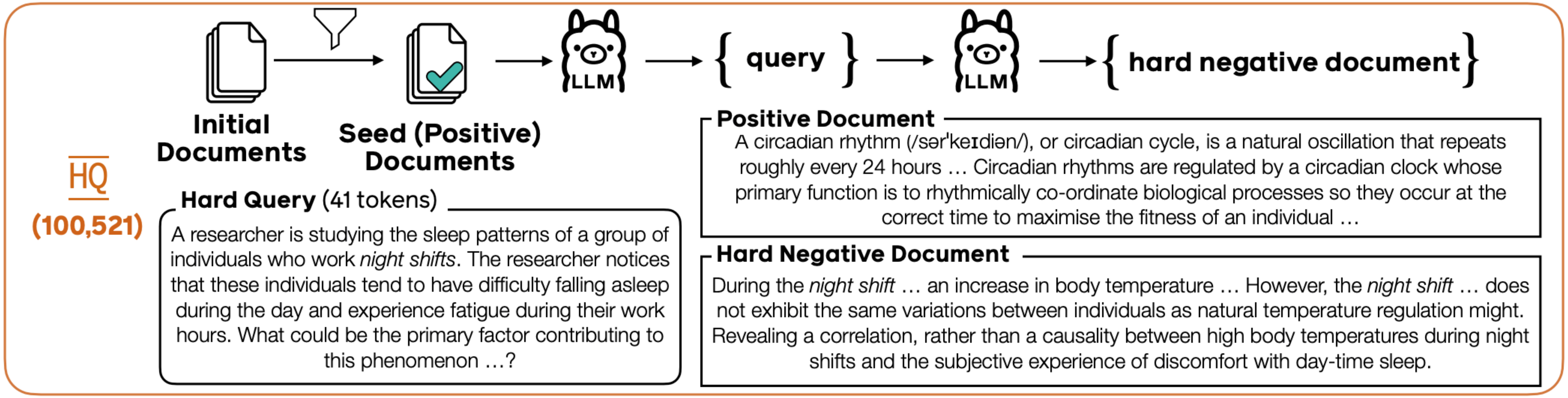
Su, H., Yen, H., Xia, M., Shi, W., Muennighoff, N., Wang, H., Haisu, L., Shi, Q., Siegel, Z. S., Tang, M., Sun, R., Yoon, J., Arik, S. O., Chen, D. and Yu, T. BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval. *ICLR 2025*.

ReasonIR



Shao, R., Qiao, R., Kishore, V., Muennighoff, N., Lin, X. V., Rus, D., Low, B. K. H., Min, S., Yih, W., Koh, P. W. and Zettlemoyer, L. ReasonIR: Training Retrievers for Reasoning Tasks. *arXiv preprint arXiv:2504.20595.*

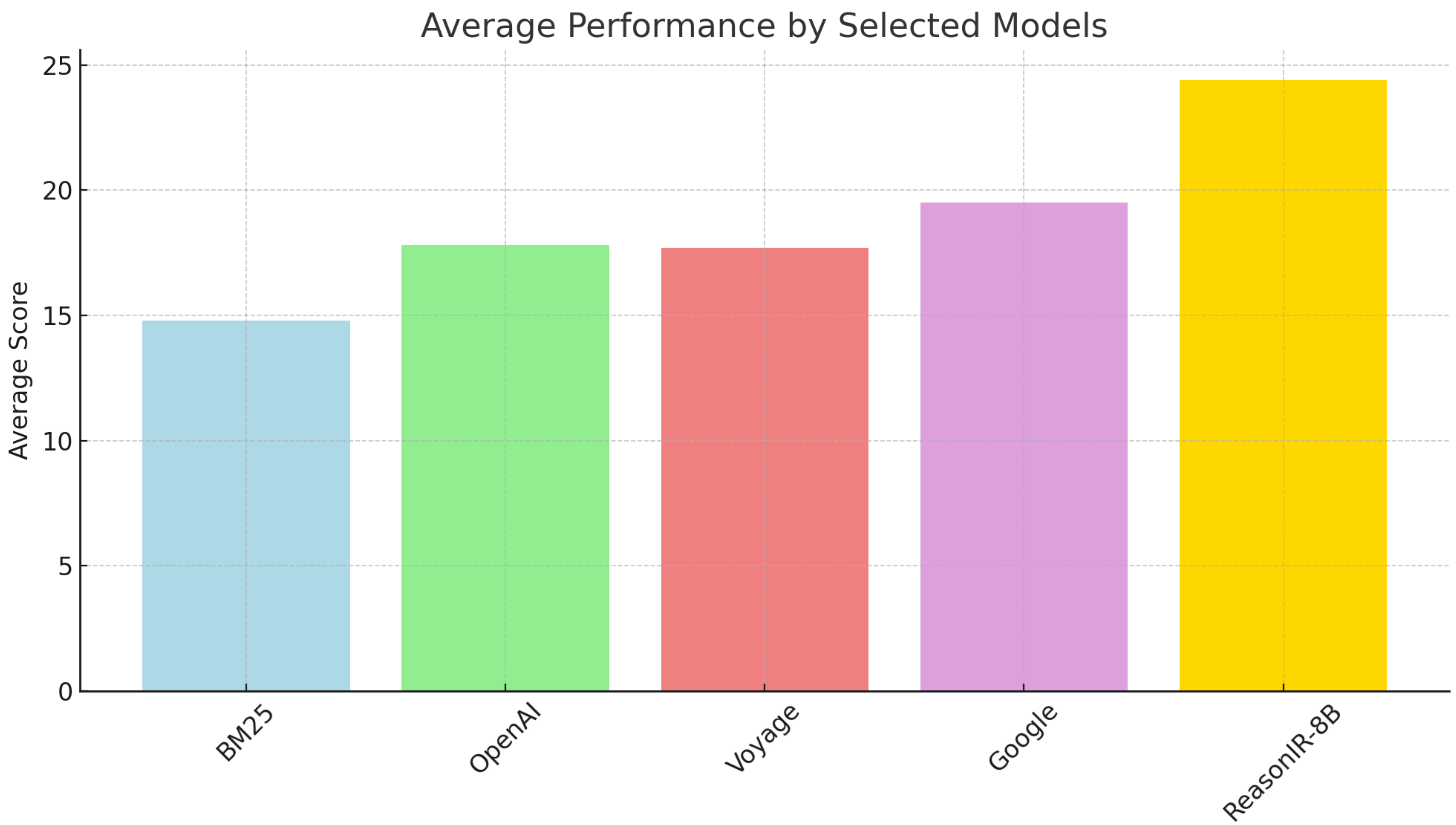
ReasonIR



Shao, R., Qiao, R., Kishore, V., Muennighoff, N., Lin, X. V., Rus, D., Low, B. K. H., Min, S., Yih, W., Koh, P. W. and Zettlemoyer, L. ReasonIR: Training Retrievers for Reasoning Tasks. *arXiv preprint arXiv:2504.20595*.

ReasonIR

- Train by mixing VL and HQ, with public retrieval datasets



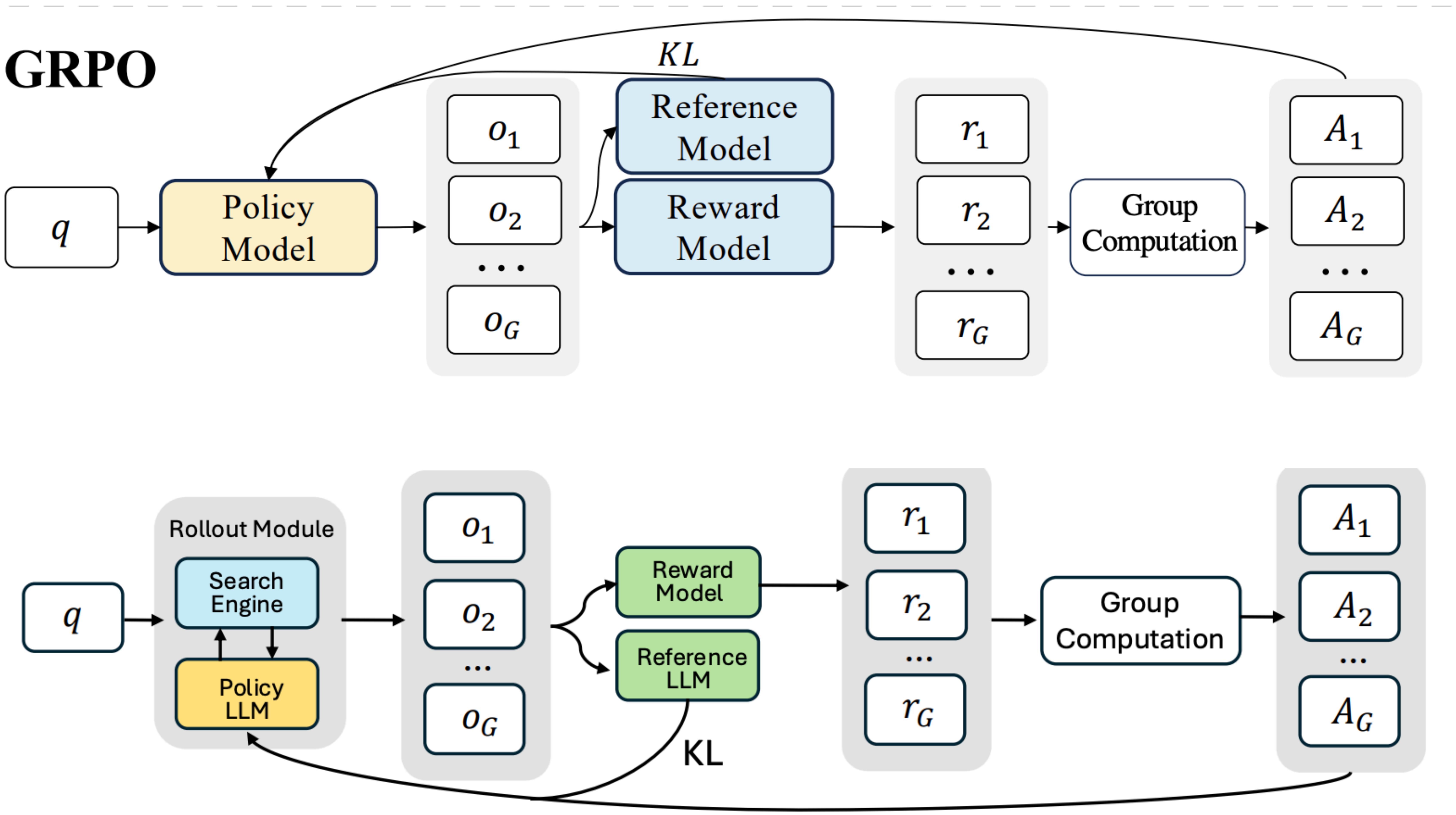
Shao, R., Qiao, R., Kishore, V., Muennighoff, N., Lin, X. V., Rus, D., Low, B. K. H., Min, S., Yih, W., Koh, P. W. and Zettlemoyer, L. ReasonIR: Training Retrievers for Reasoning Tasks. *arXiv preprint arXiv:2504.20595.*

Synthetic data for LLM retrieval training

- Improve the generalizability
- Adapt to specific tasks / domains

- Single round search is limited for complex query
- Multi-round search, with interleaved searching, reflecting, and query rewriting process can be dynamic, hard to create target supervision.
- Recent advancement on reinforcement learning with verifiable rewards is promising

Search-R1: RL for Search



Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H. and Han, J. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516*.

Search-R1: RL for Search

<think> I need to determine if Jed Hoyer or John William Henry II is older. I'll search for their ages. </think>
<search> Jed Hoyer or John William Henry II </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleya
was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury B
England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions off
alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admissions department of Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

Reasoning over user query, to plan the search

<think> I found out that Jed Hoyer is a person. Now I need to find his age and compare it to John William Henry II's age. </think>
<search> John William Henry II </search>
<information> Doc 1(Title: "John W. Henry") the firm's management had fallen from \$2.5 billion in 2006 to less than \$100 million as of late 2012. As of July 2017, ""Forbes"" estimated his net worth to be \$2.6 billion. John William Henry II was born on September 13, 1949 in Quincy, Illinois. His parents were soybean farmers, and he split his time growing up between Illinois and Arkansas. His asthmatic condition at the age of 15 prompted his family to move to Apple Valley, California. After his graduation from Victor Valley High School in Victorville, he attended Victor Valley College, then the University of California (at Riverside, ... </information>

<think> I found out that John William Henry II was born on September 13, 1949. Now I need to find Jed Hoyer's birth year to compare. </think>
<search> Jed Hoyer birth year </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleyan University he was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury Barons of the New England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions office and then the alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admissions department of Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

<think> I found out that Jed Hoyer was born on December 7, 1973. Now I can compare his birth year to John William Henry II's birth year. </think>
<answer> John William Henry II </answer>

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H. and Han, J. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516.*

Search-R1: RL for Search

<think> I need to determine if Jed Hoyer or John William Henry II is older. I'll search for their ages. </think>
<search> Jed Hoyer or John William Henry II </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleya
was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury B
England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions off
alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admissio
Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

<think> I found out that Jed Hoyer is a person. Now I need to find his age and compare it to John William Henry II's age
<search> John William Henry II </search>
<information> Doc 1(Title: "John W. Henry") the firm's management had fallen from \$2.5 billion in 2006 to less than \$1
late 2012. As of July 2017, ""Forbes"" estimated his net worth to be \$2.6 billion. John William Henry II was born on Sep
in Quincy, Illinois. His parents were soybean farmers, and he split his time growing up between Illinois and Arkansas. His asthmatic
condition at the age of 15 prompted his family to move to Apple Valley, California. After his graduation from Victor Valley High School in
Victorville, he attended Victor Valley College, then the University of California (at Riverside, ... </information>

<think> I found out that John William Henry II was born on September 13, 1949. Now I need to find Jed Hoyer's birth year to compare. </think>
<search> Jed Hoyer birth year </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleyan University he
was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury Barons of the New
England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions office and then the
alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admissions department of
Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

<think> I found out that Jed Hoyer was born on December 7, 1973. Now I can compare his birth year to John William Henry II's birth
year. </think>
<answer> John William Henry II </answer>

Reasoning over user query, to plan the search

Reasoning over the search results to write new query

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H. and Han, J. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516.*

Search-R1: RL for Search

<think> I need to determine if Jed Hoyer or John William Henry II is older. I'll search for their ages. </think>
<search> Jed Hoyer or John William Henry II </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleya
was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury B
England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions off
alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admissio
Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

<think> I found out that Jed Hoyer is a person. Now I need to find his age and compare it to John William Henry II's age
<search> John William Henry II </search>
<information> Doc 1(Title: "John W. Henry") the firm's management had fallen from \$2.5 billion in 2006 to less than \$1
late 2012. As of July 2017, ""Forbes"" estimated his net worth to be \$2.6 billion. John William Henry II was born on Sept
in Quincy, Illinois. His parents were soybean farmers, and he split his time growing up between Illinois and Arkansas. His asthmatic condition at the age of 15 prompted his family to move to Apple Valley, California. After his graduation from Victor Valley High School in Victorville, he attended Victor Valley College, then the University of California (at Riverside, ... </information>

<think> I found out that John William Henry II was born on September 13, 1949. Now I need to find Jed Hoyer's birth year to compare. </think>
<search> Jed Hoyer birth year </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleyan University he was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury Barons of the New England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions office and then the alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admission
Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

<think> I found out that Jed Hoyer was born on December 7, 1973. Now I can compare his birth year to John William He
year. </think>
<answer> John William Henry II </answer>

Reasoning over user query, to plan the search

Reasoning over the search results to write new query

Reasoning over the gathered information and summarize the answer

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H. and Han, J. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516.*

Search-R1: RL for Search

<think> I need to determine if Jed Hoyer or John William Henry II is older. I'll search for their ages. </think>
<search> Jed Hoyer or John William Henry II </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleya
was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury B
England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions off
alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admissio
Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

<think> I found out that Jed Hoyer is a person. Now I need to find his age and compare it to John William Henry II's age
<search> John William Henry II </search>
<information> Doc 1(Title: "John W. Henry") the firm's management had fallen from \$2.5 billion in 2006 to less than \$1
late 2012. As of July 2017, ""Forbes"" estimated his net worth to be \$2.6 billion. John William Henry II was born on Sept
in Quincy, Illinois. His parents were soybean farmers, and he split his time growing up between Illinois and Arkansas. His asthmatic condition at the age of 15 prompted his family to move to Apple Valley, California. After his graduation from Victor Valley High School in Victorville, he attended Victor Valley College, then the University of California (at Riverside, ... </information>

<think> I found out that John William Henry II was born on September 13, 1949. Now I need to find Jed Hoyer's birth year to compare. </think>
<search> Jed Hoyer birth year </search>
<information> Doc 1(Title: "Jed Hoyer") career saves record with Sam Elias and Nick Miceli. During his time at Wesleyan University he was also initiated into and is a brother of Delta Kappa Epsilon. He spent Summer 1995 on the roster of the Waterbury Barons of the New England Collegiate Baseball League seeing innings on the mound and at shortstop. Hoyer worked in the admissions office and then the alumni/development office, and was also a baseball coach at the university after graduating. He worked in the admissions department of Kenyon College before joining the Red Sox at 28. He married Merrill Muckerman in June 2010 in ... </information>

<think> I found out that Jed Hoyer was born on De ... is birth year to John Willian He
year. </think>
<answer> John William Henry II </answer>

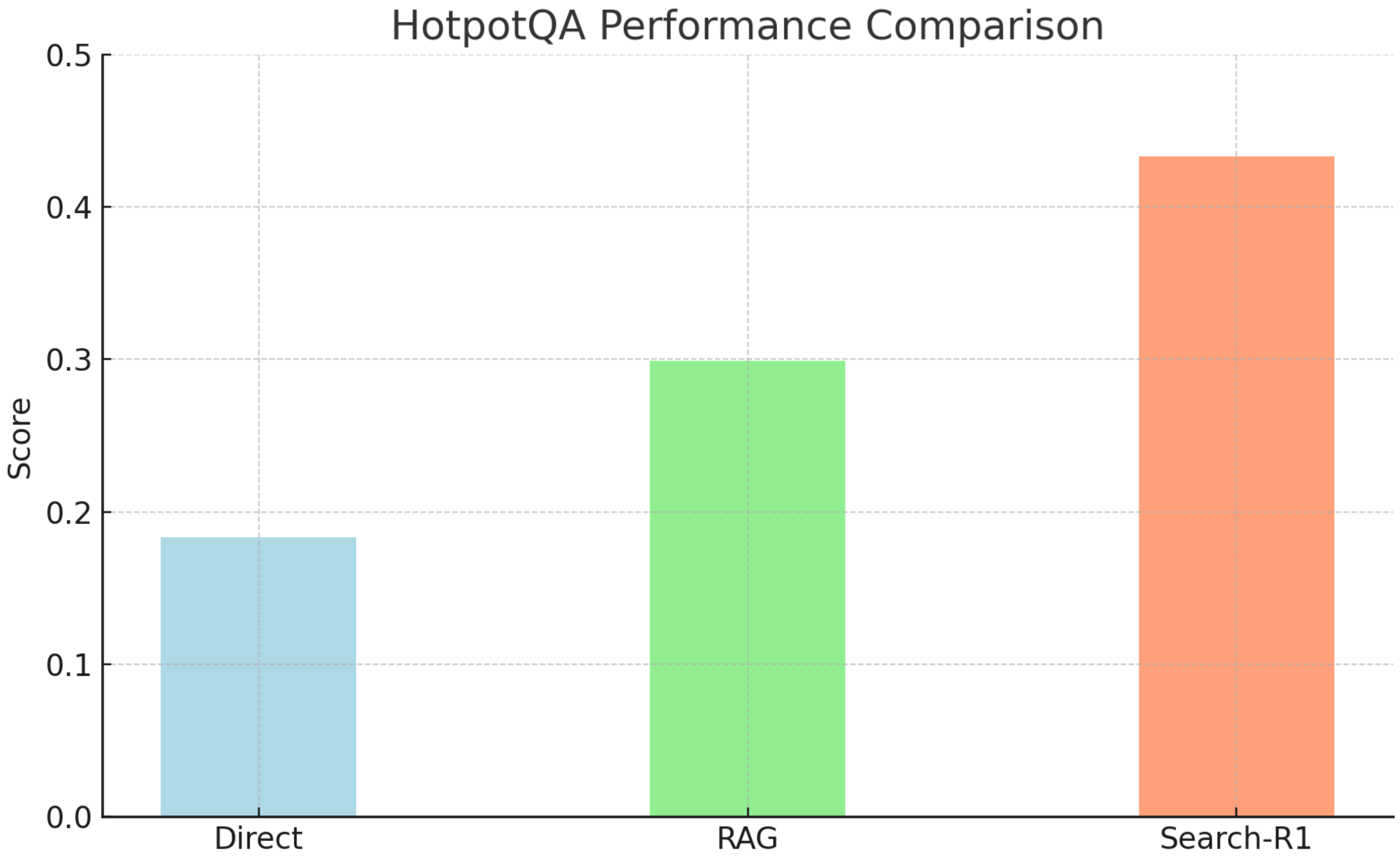
Reasoning over user query, to plan the search

Reasoning over the search results to write new query

Reasoning over the gathered information and summarize the answer

Compute the rewards
If EM, then 1, otherwise 0

Search-R1: RL for Search



Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H. and Han, J. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *arXiv preprint arXiv:2503.09516.*

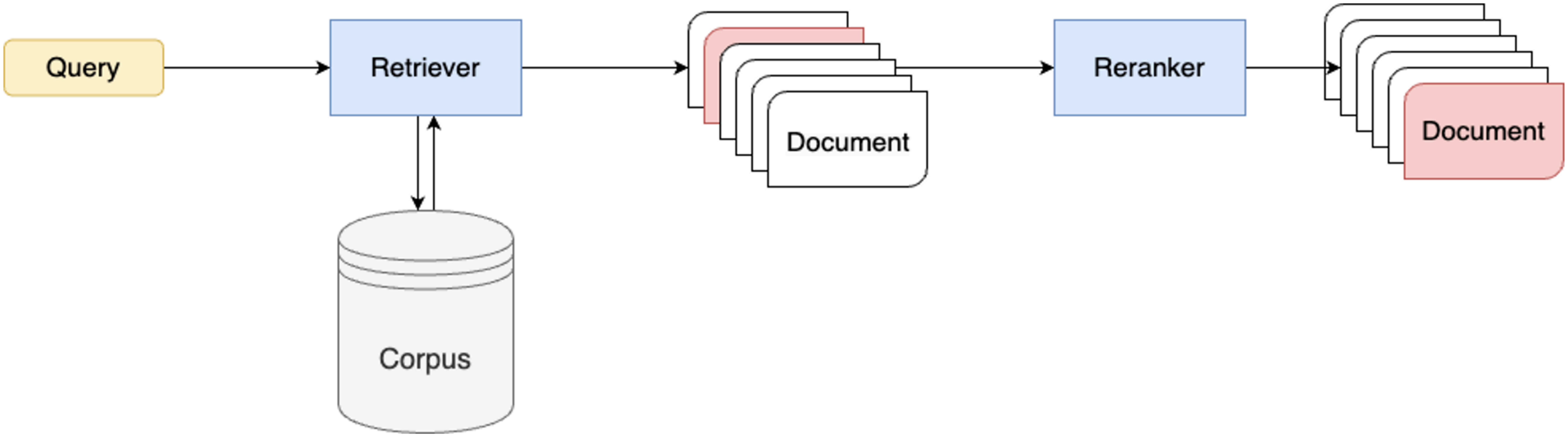
LLM backbone makes retriever robust

LLM data synthesis makes LLM general purpose embedding

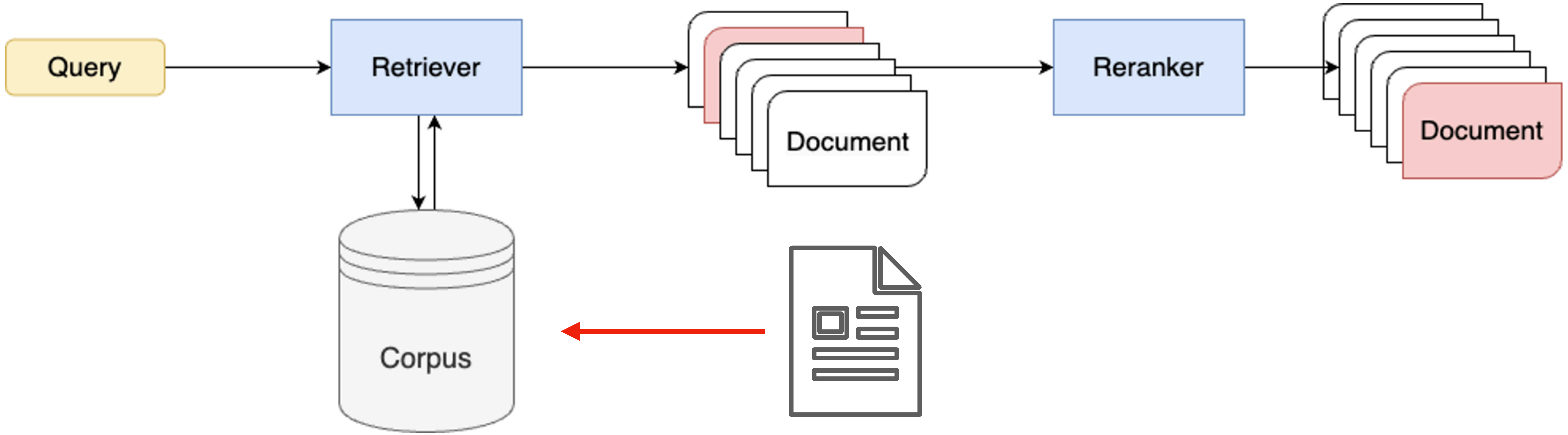
LLM-Reasoning makes search agent smart

However, Information is not just text…

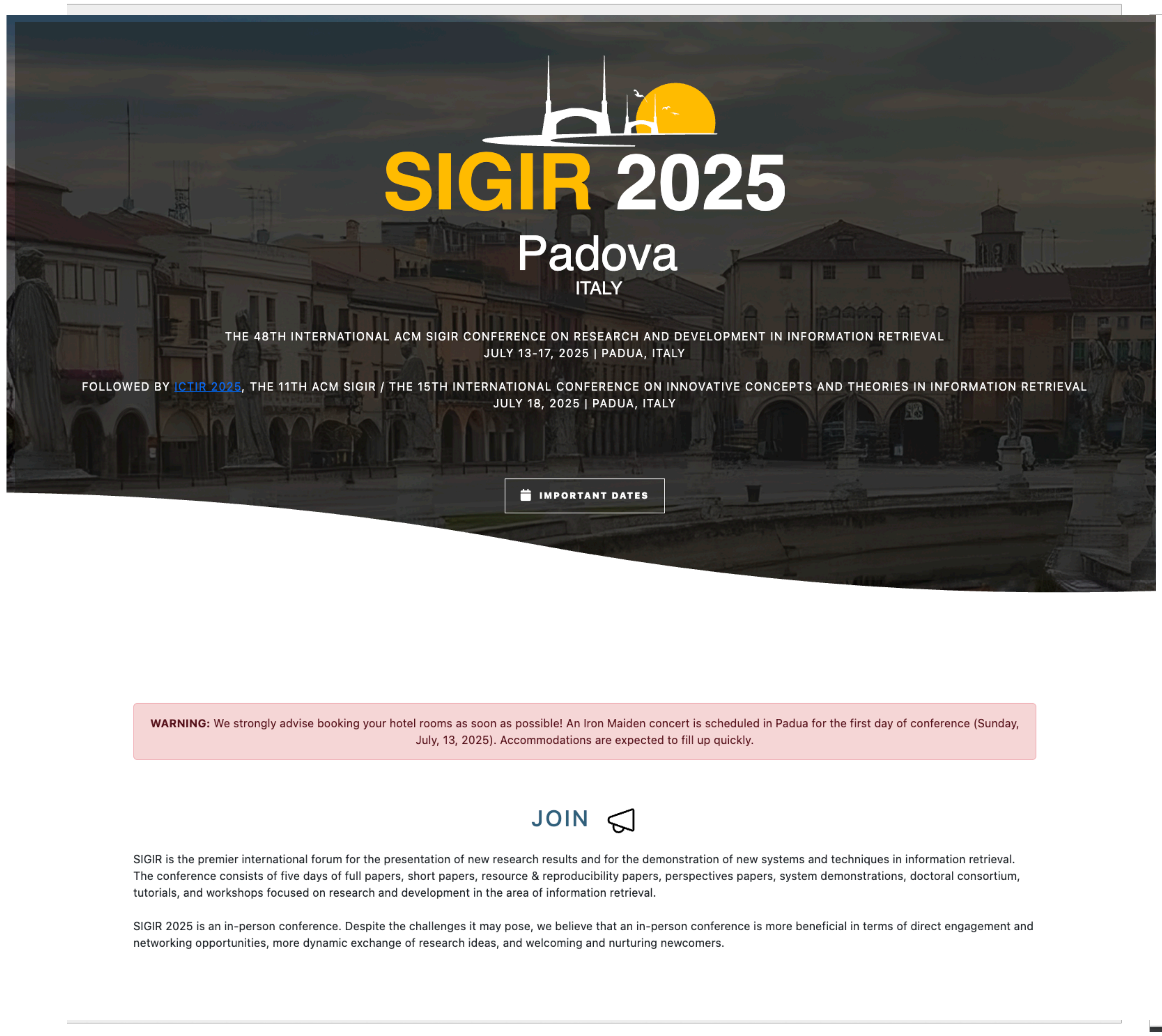
Rethink Document Retrieval Pipeline

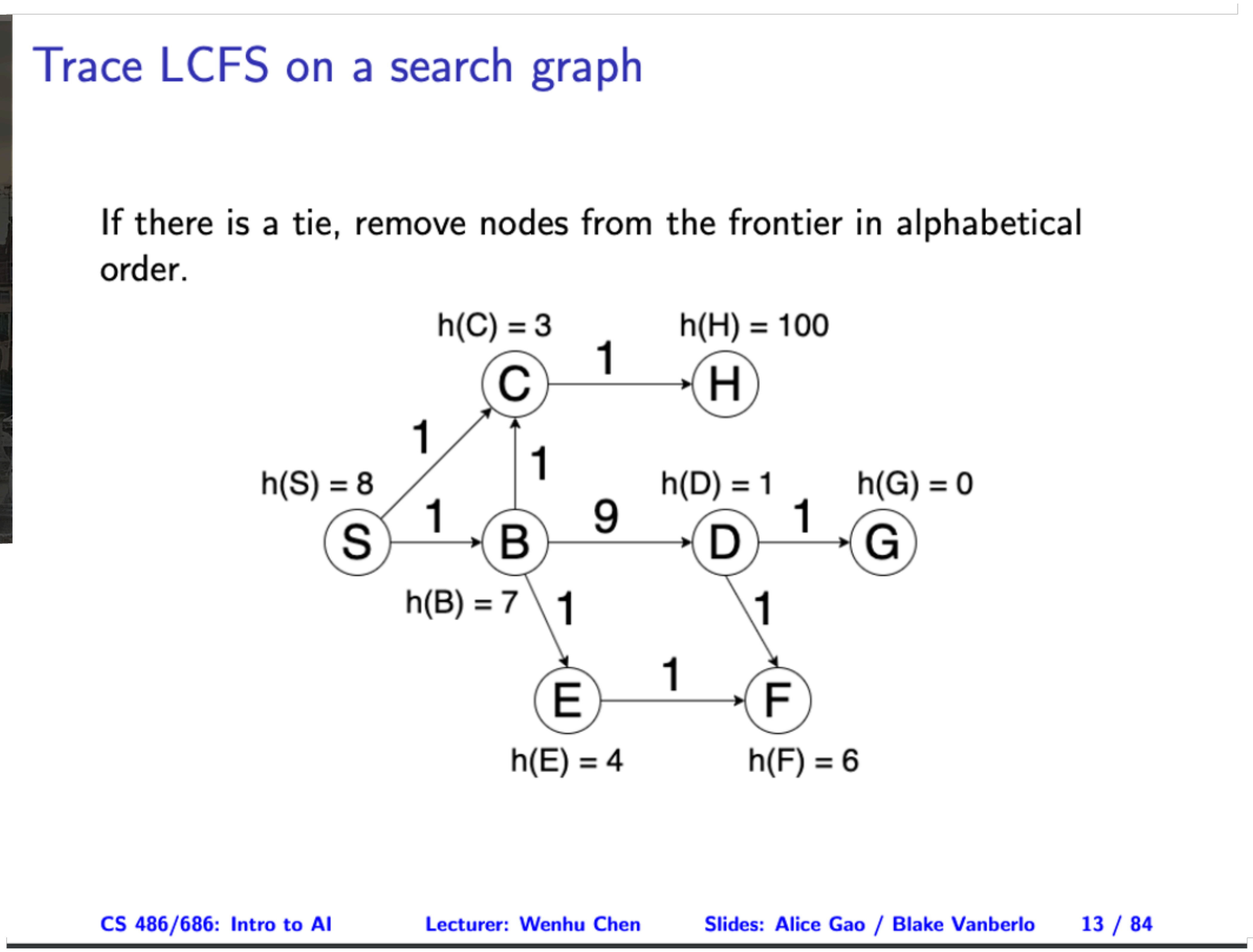


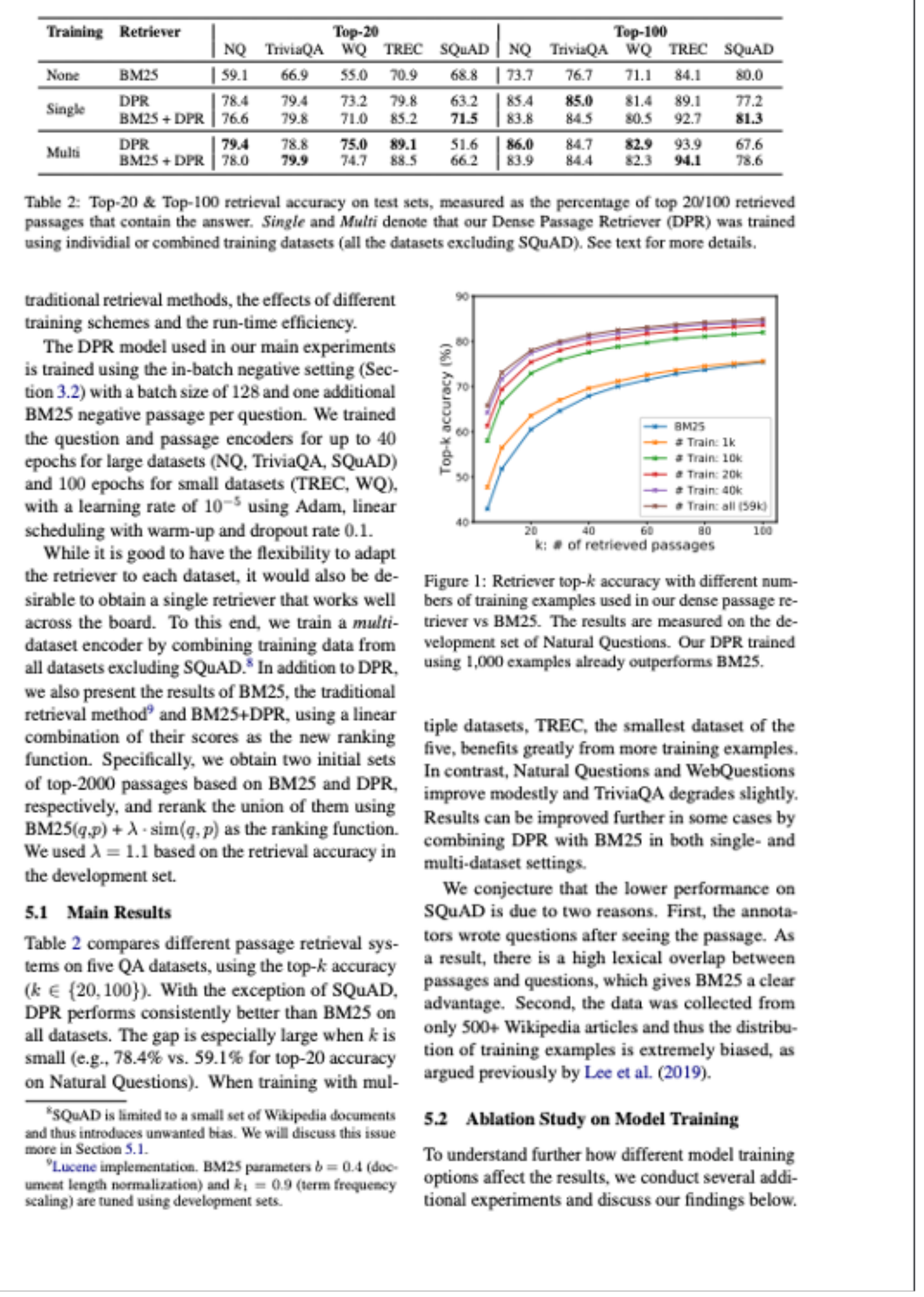
Gap in Document Processing



Gap in Document Processing







Training	Retriever	Top-20					Top-100				
		NQ	TriviaQA	WQ	TREC	SQuAD	NQ	TriviaQA	WQ	TREC	SQuAD
None	BM25	59.1	66.9	55.0	70.9	68.8	73.7	76.7	71.1	84.1	80.0
Single	DPR	78.4	79.4	73.2	79.8	63.2	85.4	**85.0**	81.4	89.1	77.2
	BM25 + DPR	76.6	79.8	71.0	85.2	**71.5**	83.8	84.5	80.5	92.7	**81.3**
Multi	DPR	**79.4**	78.8	**75.0**	**89.1**	51.6	**86.0**	84.7	**82.9**	93.9	67.6
	BM25 + DPR	78.0	**79.9**	74.7	88.5	66.2	83.9	84.4	82.3	**94.1**	78.6

Table 2: Top-20 & Top-100 retrieval accuracy on test sets, measured as the percentage of top 20/100 retrieved passages that contain the answer. *Single* and *Multi* denote that our Dense Passage Retriever (DPR) was trained using individial or combined training datasets (all the datasets excluding SQuAD). See text for more details.

traditional retrieval methods, the effects of different training schemes and the run-time efficiency.

The DPR model used in our main experiments is trained using the in-batch negative setting (Section 3.2) with a batch size of 128 and one additional BM25 negative passage per question. We trained the question and passage encoders for up to 40 epochs for large datasets (NQ, TriviaQA, SQuAD) and 100 epochs for small datasets (TREC, WQ), with a learning rate of 10^{-5} using Adam, linear scheduling with warm-up and dropout rate 0.1.

While it is good to have the flexibility to adapt the retriever to each dataset, it would also be desirable to obtain a single retriever that works well across the board. To this end, we train a *multi*-dataset encoder by combining training data from all datasets excluding SQuAD.[8] In addition to DPR, we also present the results of BM25, the traditional retrieval method[9] and BM25+DPR, using a linear combination of their scores as the new ranking function. Specifically, we obtain two initial sets of top-2000 passages based on BM25 and DPR, respectively, and rerank the union of them using $\mathrm{BM25}(q,p) + \lambda \cdot \mathrm{sim}(q,p)$ as the ranking function. We used $\lambda = 1.1$ based on the retrieval accuracy in the development set.

5.1 Main Results

Table 2 compares different passage retrieval systems on five QA datasets, using the top-k accuracy ($k \in \{20, 100\}$). With the exception of SQuAD, DPR performs consistently better than BM25 on all datasets. The gap is especially large when k is small (e.g., 78.4% vs. 59.1% for top-20 accuracy on Natural Questions). When training with multiple datasets, TREC, the smallest dataset of the five, benefits greatly from more training examples. In contrast, Natural Questions and WebQuestions improve modestly and TriviaQA degrades slightly. Results can be improved further in some cases by combining DPR with BM25 in both single- and multi-dataset settings.

[8] SQuAD is limited to a small set of Wikipedia documents and thus introduces unwanted bias. We will discuss this issue more in Section 5.1.

[9] Lucene implementation. BM25 parameters $b = 0.4$ (document length normalization) and $k_1 = 0.9$ (term frequency scaling) are tuned using development sets.

Figure 1: Retriever top-k accuracy with different numbers of training examples used in our dense passage retriever vs BM25. The results are measured on the development set of Natural Questions. Our DPR trained using 1,000 examples already outperforms BM25.

We conjecture that the lower performance on SQuAD is due to two reasons. First, the annotators wrote questions after seeing the passage. As a result, there is a high lexical overlap between passages and questions, which gives BM25 a clear advantage. Second, the data was collected from only 500+ Wikipedia articles and thus the distribution of training examples is extremely biased, as argued previously by Lee et al. (2019).

5.2 Ablation Study on Model Training

To understand further how different model training options affect the results, we conduct several additional experiments and discuss our findings below.

Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.

Gap in Document Processing

Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.

Gap in Document Processing

Document Processing are prone to errors and information loss

- It breaks original look which can affects users relevance preference
- OCR etc. can have errors

Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.

DSE: Unifying Multimodal Retrieval with Document Screenshot Embedding

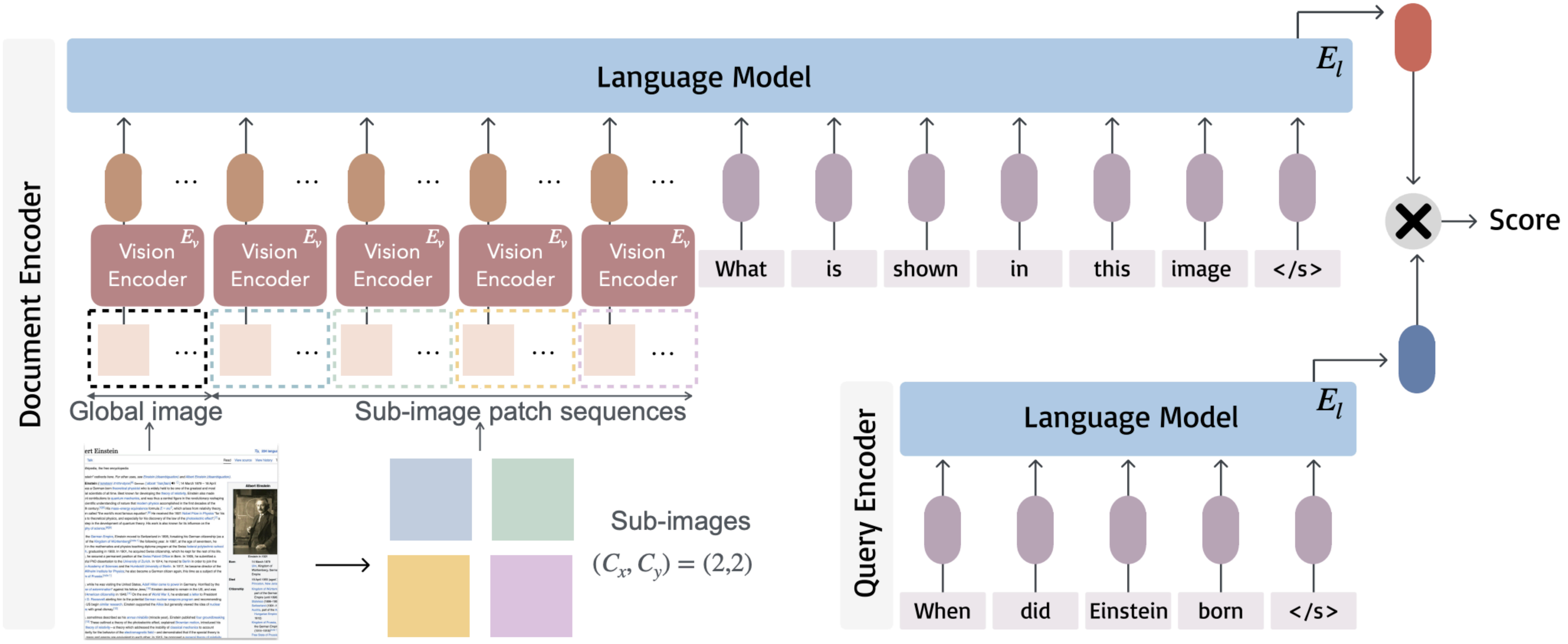
The best document processing is “Do Not Process”

Use “Screenshot” as unified input

Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.

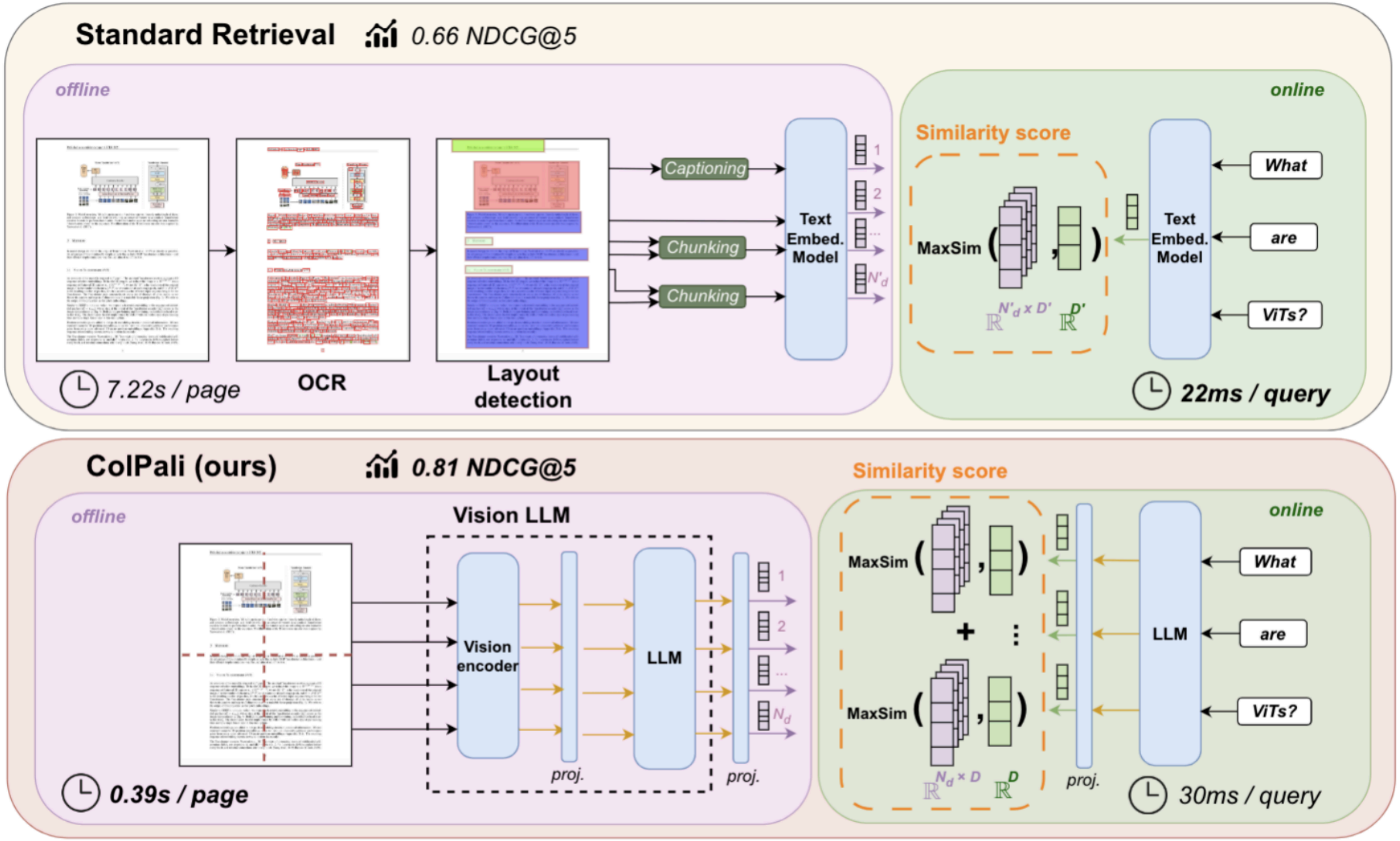
DSE: Unifying Multimodal Retrieval with Document Screenshot Embedding

Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.



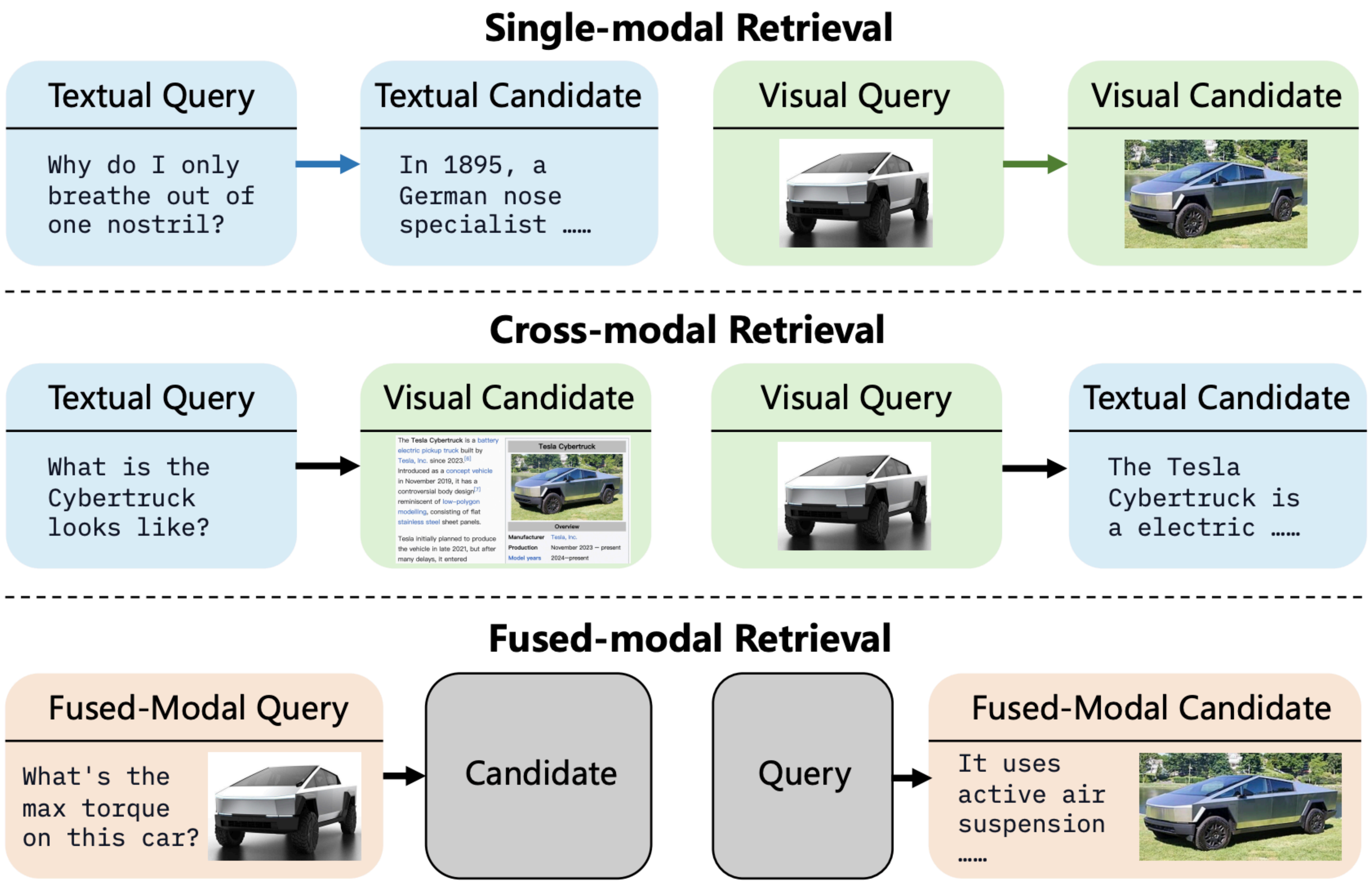
Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.

ColPali: A multi-vector document screenshot retrieval model



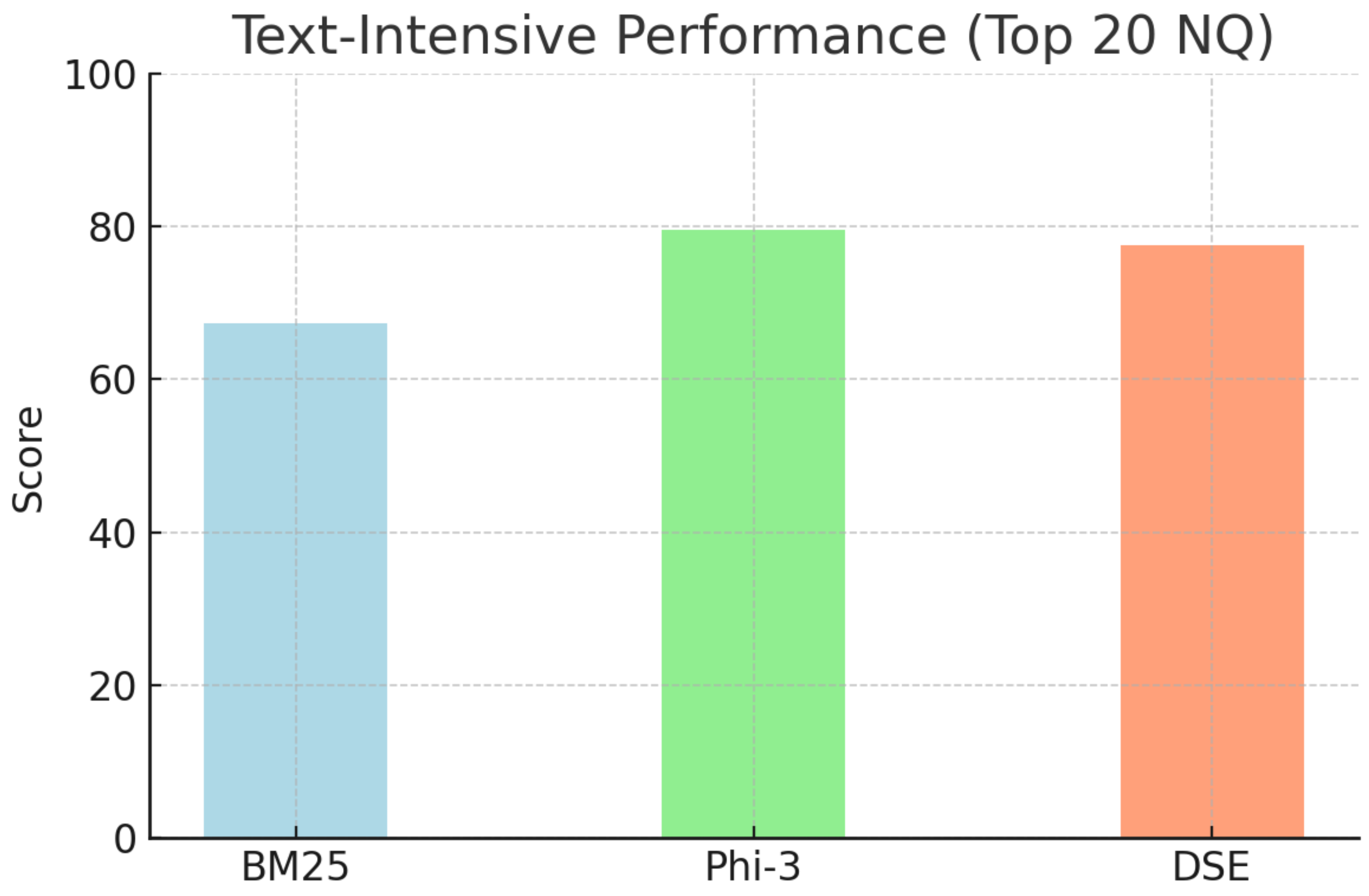
Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C. and Colombo, P. ColPali: Efficient Document Retrieval with Vision Language Models. *ICLR 2025*.

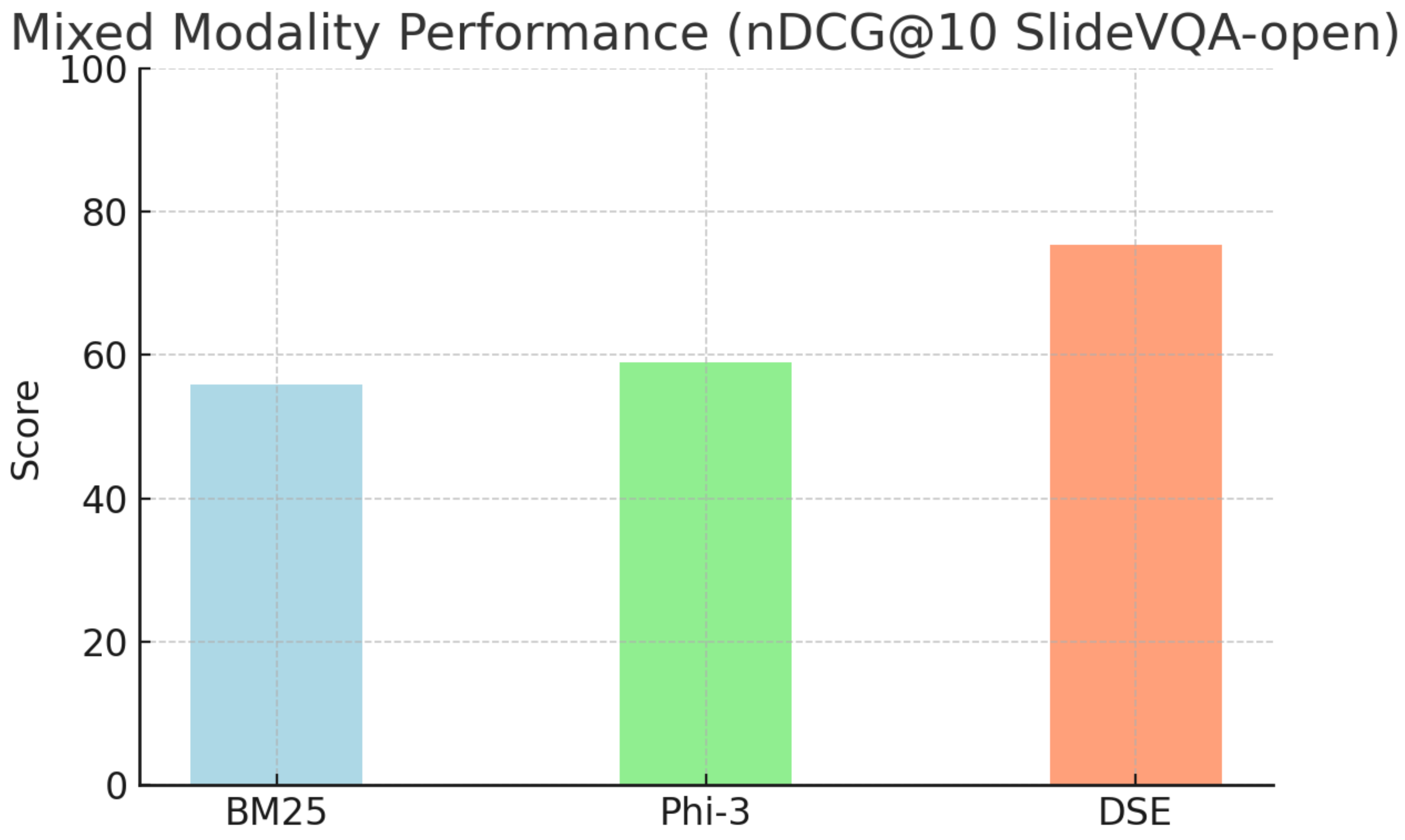
GME: Improving Universal Multimodal Retrieval by Multimodal LLMs



Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W. and Zhang, M. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint arXiv:2412.16855*.

- Can **visual-based input** get comparable effectiveness as text retriever in text-intensive retrieval tasks?
- Can **visual-based retriever** outperform text-based retriever in mixed-modality retrieval tasks?





Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.

Tevatron 2.0: Towards Unified Document Retrieval across Domain, Language, and Modality

Method	Base LLM	Domain BEIR(13)	Modality ViDoRe	Language MIRACL(17)
BGE-M3	X-RoBERTa	50.0	66.1*	69.9
MistralE5	Mistral0.1-7B	53.6	-	63.3
DSE-QWen2	Qwen2vl-2B	-	85.8	-
GME-2B	Qwen2vl-2B	55.4	87.8	-
Tevatron-WikiSS	Qwen2.5vl-3B	40.8	73.3	30.8
Tevatron-BGE	Qwen2.5vl-3B	57.0	76.4	68.2
Tevatron-Full	Qwen2.5vl-3B	$\boxed{54.3}$	$\boxed{85.3}$	$\boxed{67.6}$

Ma, X., Lin, S.-C., Li, M., Chen, W. and Lin, J. Unifying Multimodal Retrieval via Document Screenshot Embedding. *EMNLP 2024*.

Summary

- LLM backbone makes retriever robust
- LLM data synthesis makes LLM general purpose embedding
- LLM-Reasoning makes search agent smart
- VLMs enables information access beyond text

Search Anything, For Anyone

Q/A

End of Part 2

…