THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Large Language Models for Information Retrieval

Shengyao Zhuang

Postdoc@CSIRO, PhD@UQ

s.zhuang@uq.edu.au

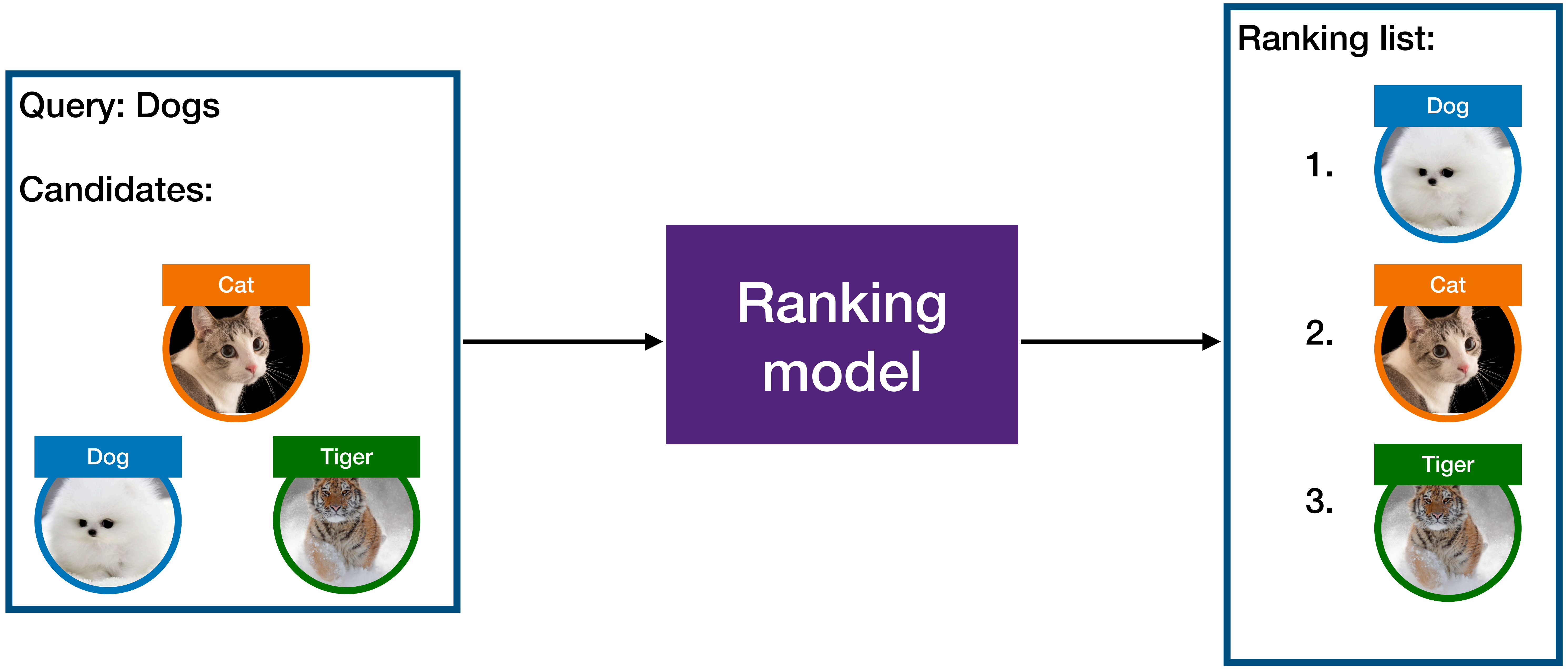
https://arvinzhuang.github.io/

ie
lab

Ranking Model (Ranker)

Ranking models are the core of search engines.

- It takes a set of candidate documents and ranks them according to their relevance to the given query.



Language model-based rankers

The University of Queensland Australia

In this talk:

- 2019 ~ 2022: BERT, T5…, less than 1B parameters
- 2022 ~ current: GPT-3/4, LLaMA…, 7B - 175B.

Traditional Ranking Models

- Bag-of-words (BoW):

Doc: *Unlike cats, dogs are usually great exercise pals….*

↓

Term frequencies: {**unlike**: 1, **cats**: 1, **dogs**: 1, **exercise**: 1, **pals**: 1, …}

Traditional Ranking Models

- Bag-of-words (BoW):

Query: dogs

Doc: *Unlike cats, dogs are usually great exercise pals….*

↓

Term frequencies: {**unlike**: 1, **cats**: 1, **dogs**: 1, **exercise**: 1, **pals**: 1, …}

Traditional Ranking Models

- Bag-of-words (BoW):

Query: Puppies

Doc: *Unlike cats, dogs are usually great exercise pals….*

↓

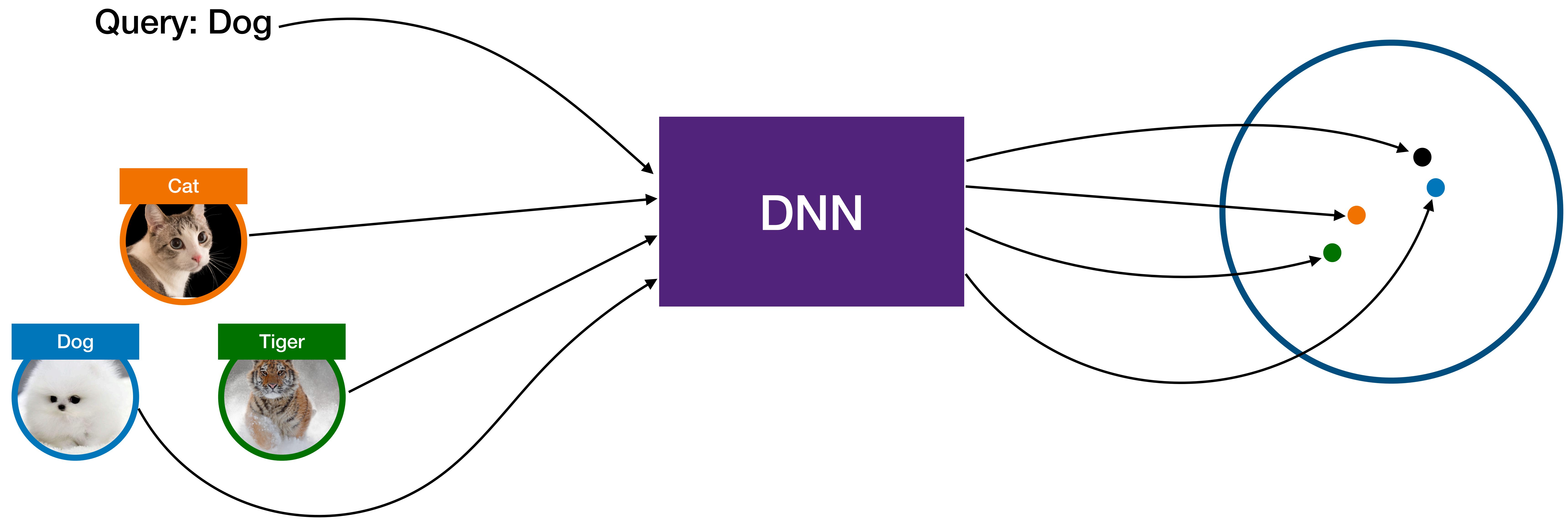
Term frequencies: {**unlike**: 1, **cats**: 1, **dogs**: 1, **exercise**: 1, **pals**: 1, …}

Vocabulary mismatch

Neural Ranking Models

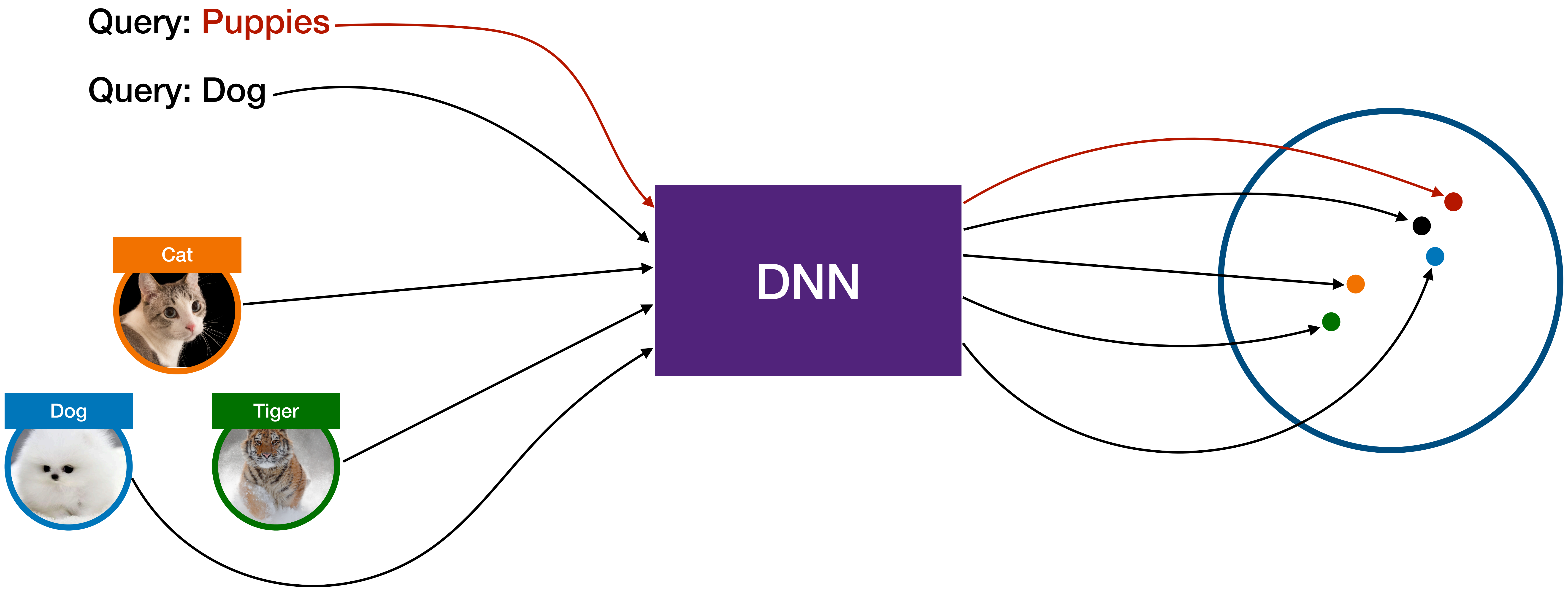
The University of Queensland Australia

Encoding query and documents with deep neural networks (DNNs).



Neural Ranking Models

Encoding query and documents with deep neural networks (DNNs).



THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Challenges of neural ranking models

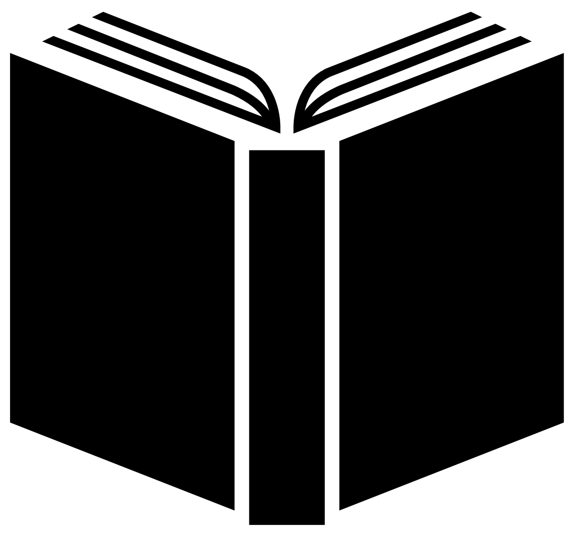
- Representation learning is hard.
- Needs lots of training data.
- Expensive to run.
- Not a great improvement over BOW.

Pre-trained Language Models (PLMs)

BERT arrived in late 2018, followed with GPTs, T5s …

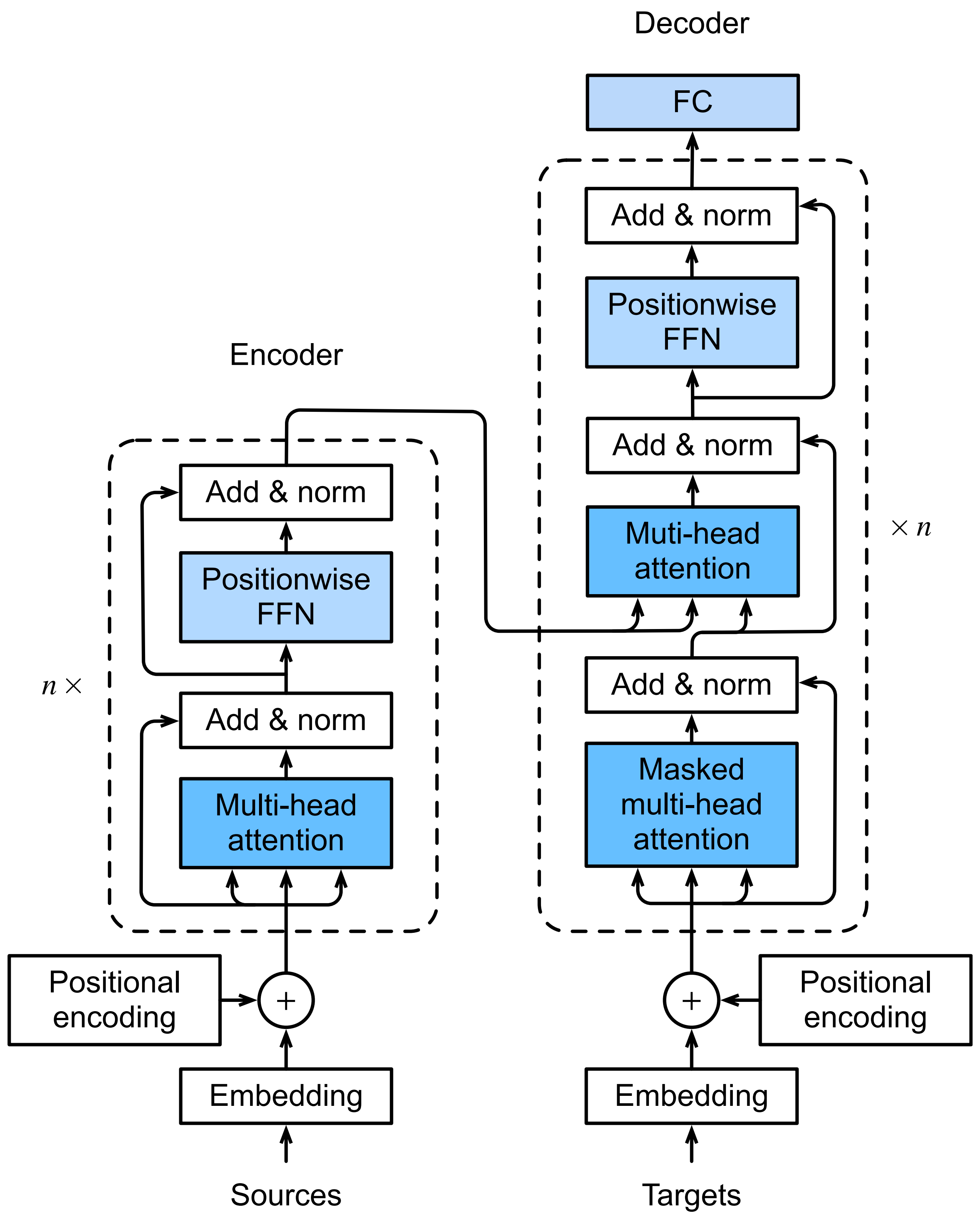
The University of Queensland Australia

Self-supervised pre-training:



Free texts

Self-supervised pre-training



Transformer

Pre-trained Language Models (PLMs)

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

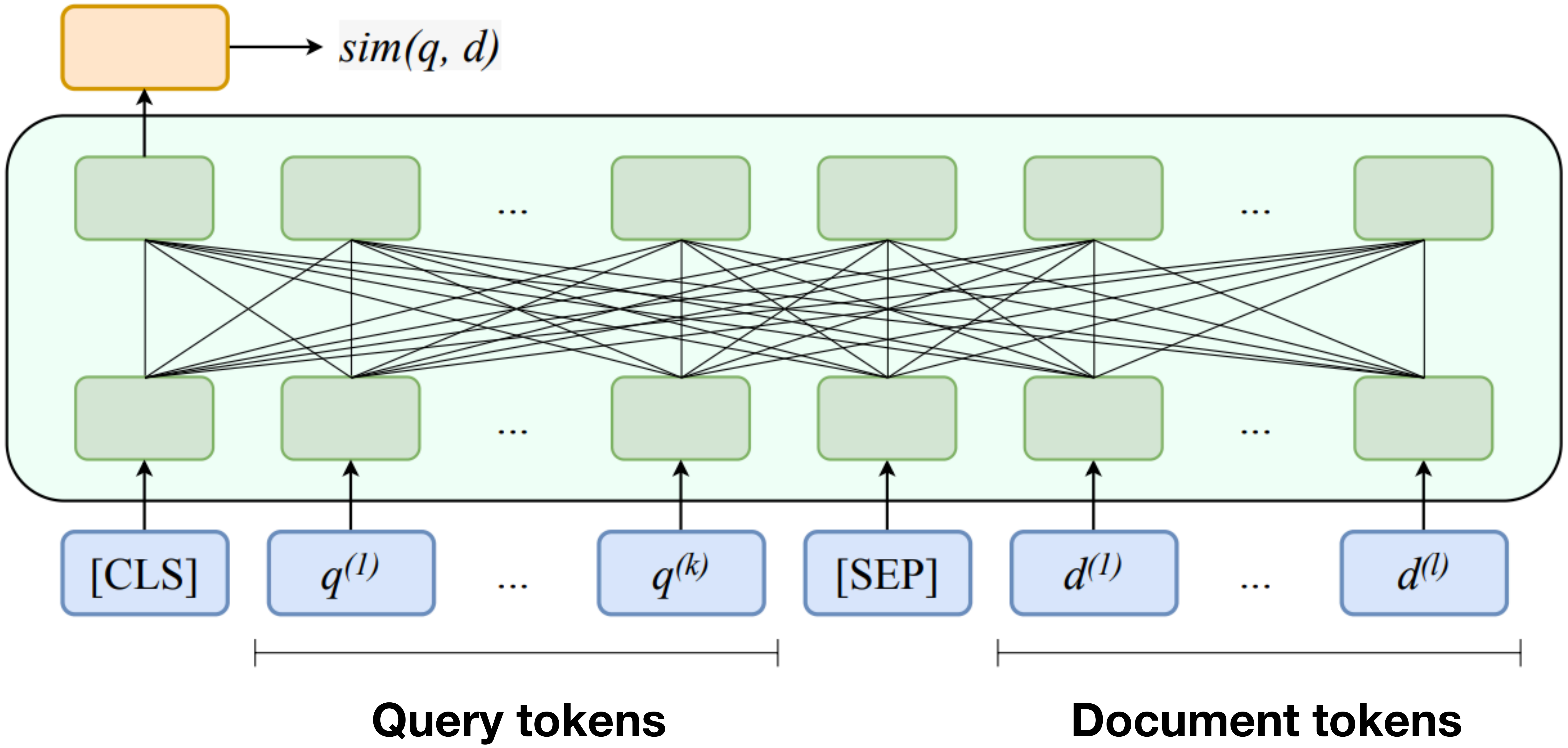
A simple adaptation of BERT for document ranking

Method		MS MARCO Passage Development MRR@10	MS MARCO Passage Test MRR@10
BM25 (Microsoft Baseline)		0.167	0.165
IRNet (Deep CNN/IR Hybrid Network)	January 2nd, 2019	0.278	0.281
BERT [Nogueira and Cho, 2019]	January 7th, 2019	0.365	0.359

J. Lin, R. Nogueira, A. Yates, Pretrained Transformers for Text Ranking: BERT and Beyond, Synthesis Lectures on Human Language Technologies 14 (4) (2021) 1–325.

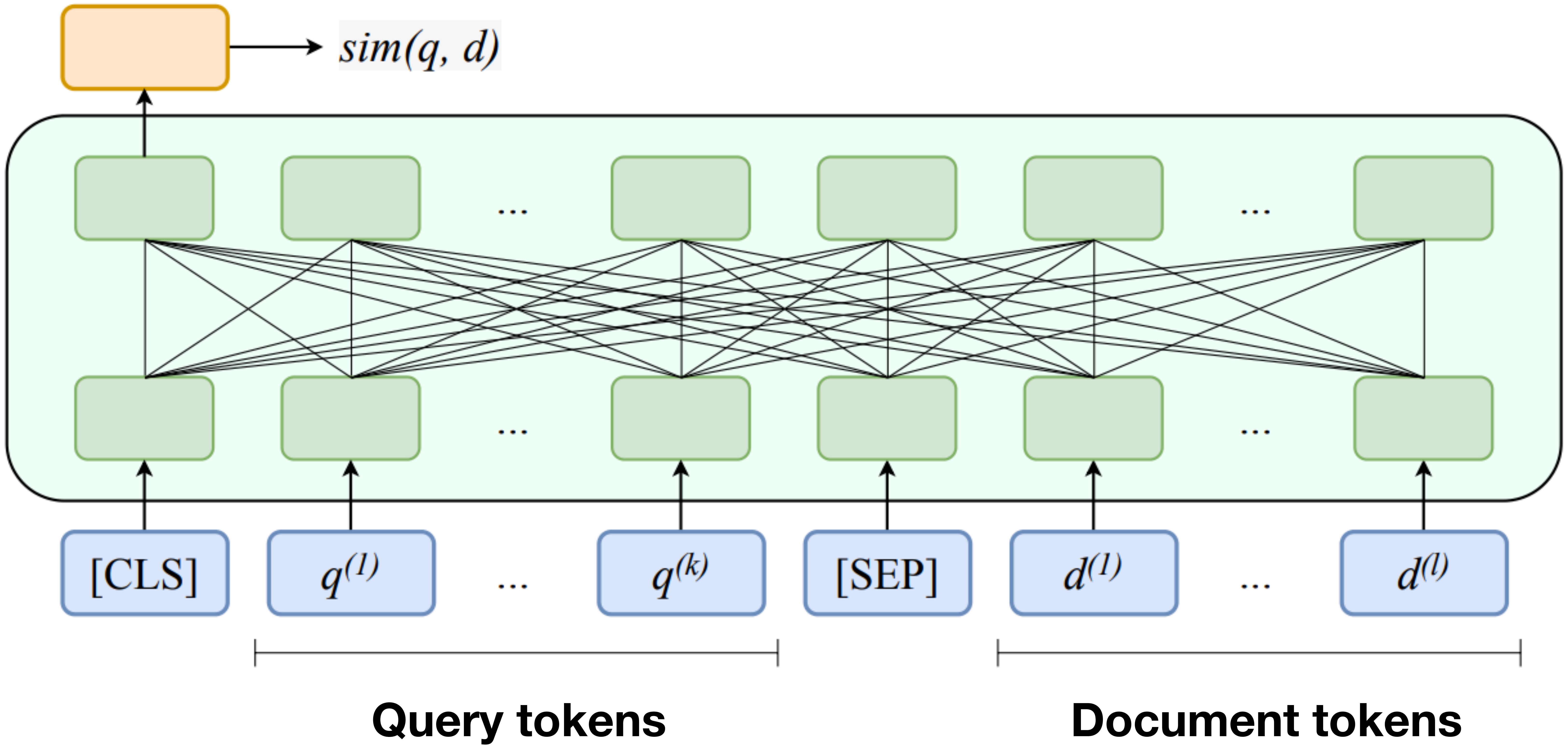
Cross-encoder ranker

A simple adaptation of BERT for document ranking



Cross-encoder ranker

A simple adaptation of BERT for document ranking

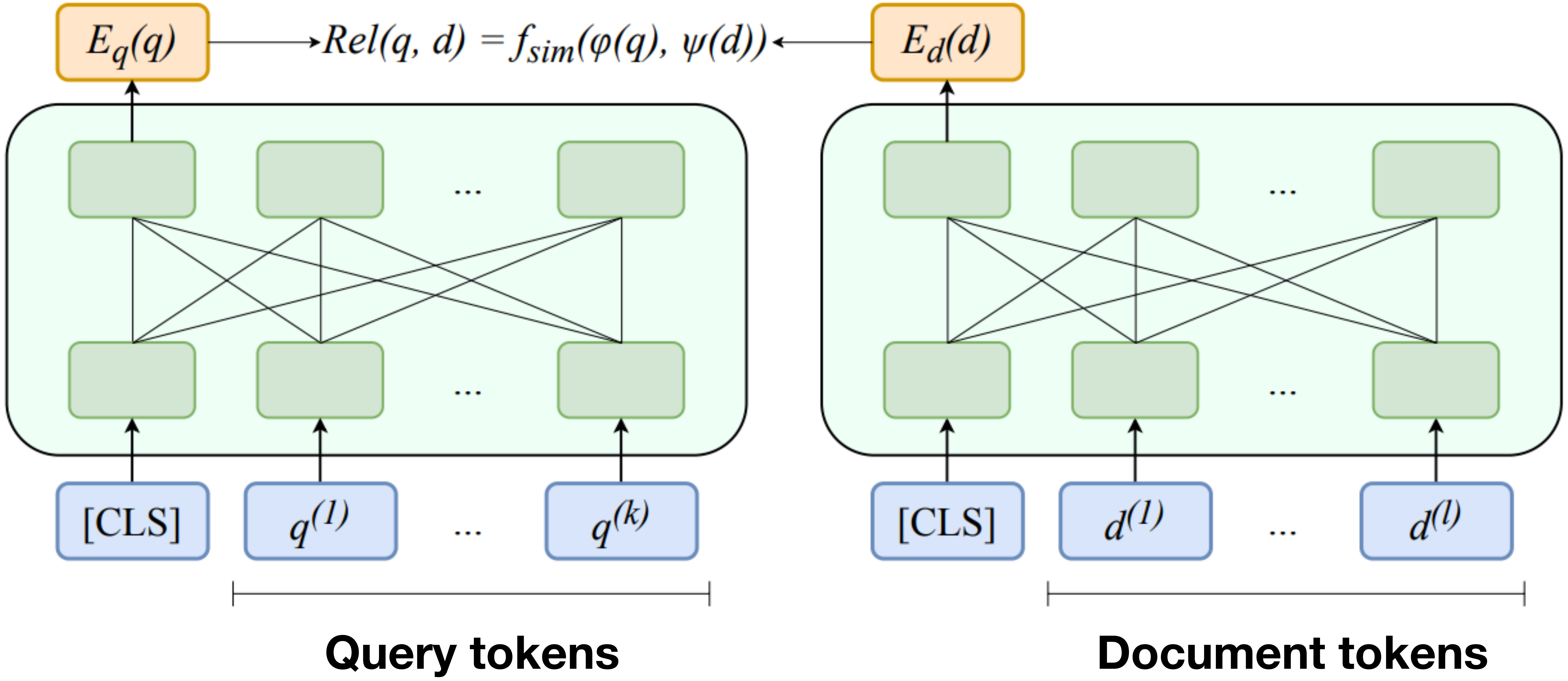


The problem: Very slow!

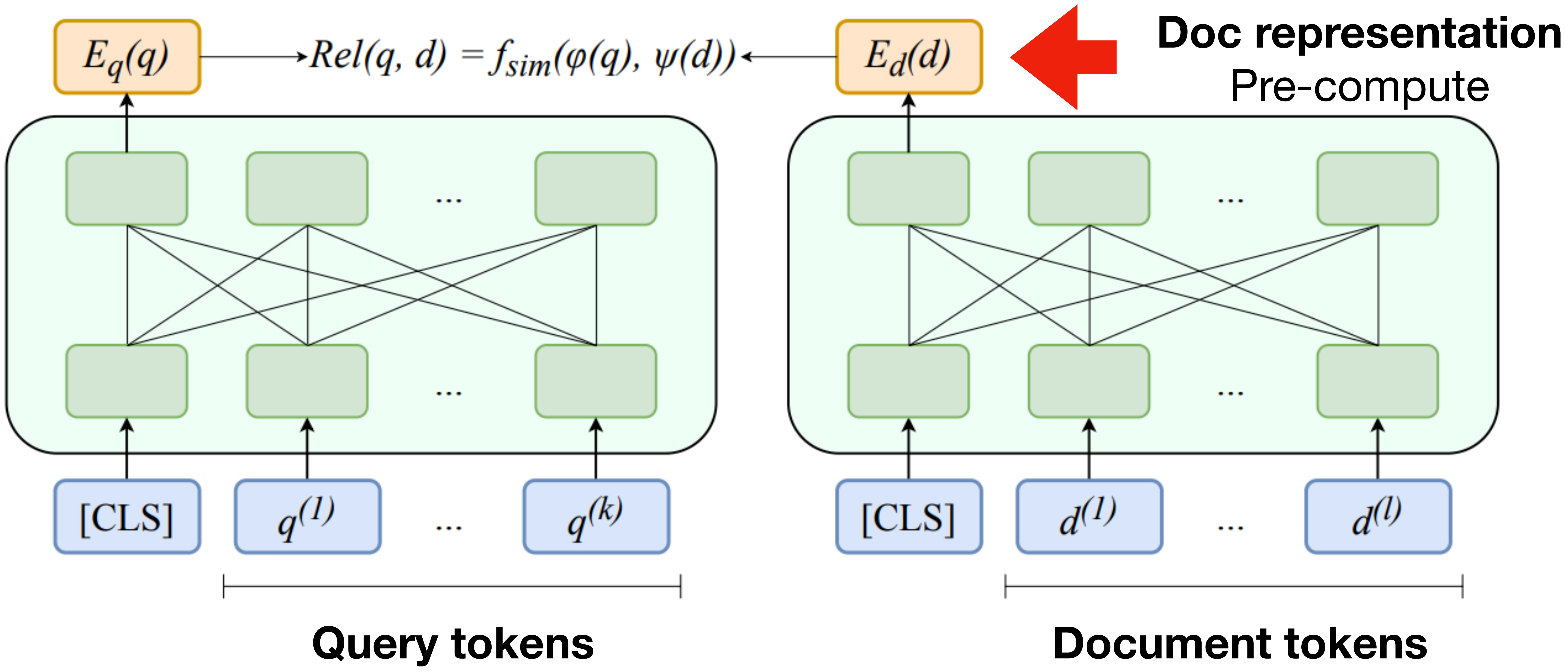
Cross-encoder ranker

Method	MRR@10	Query Latency (ms)
BM25	0.187	70
Cross-encoder (BERT large rerank BM25 top1000)	0.365	3,800 (on GPU)

Bi-encoder ranker

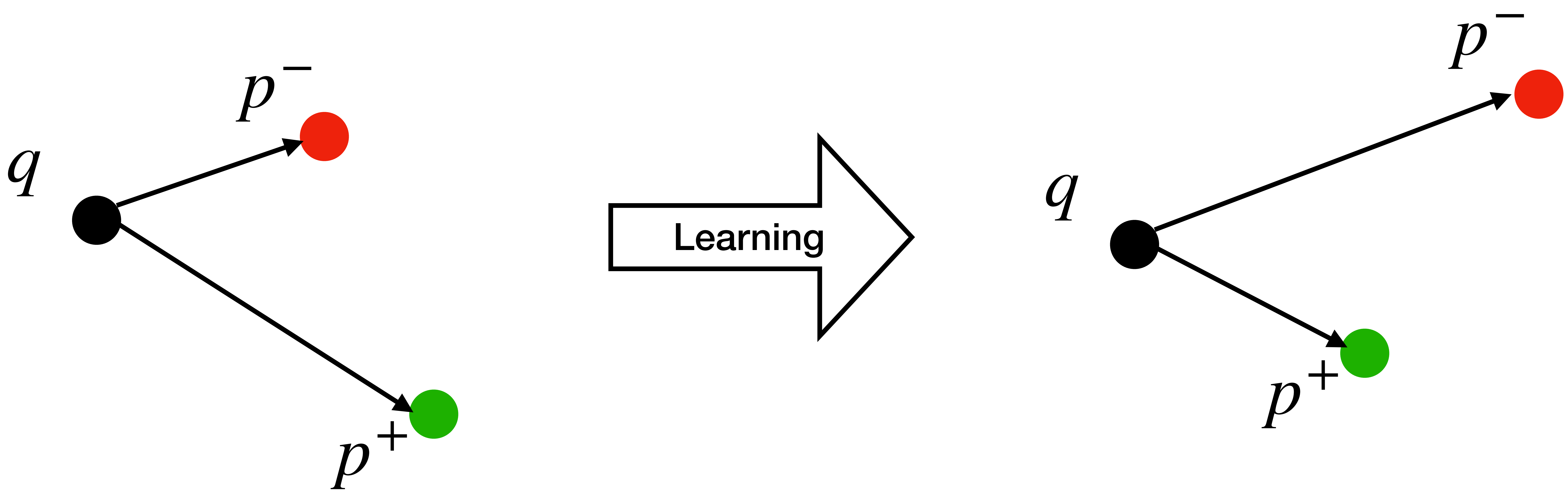


Bi-encoder ranker



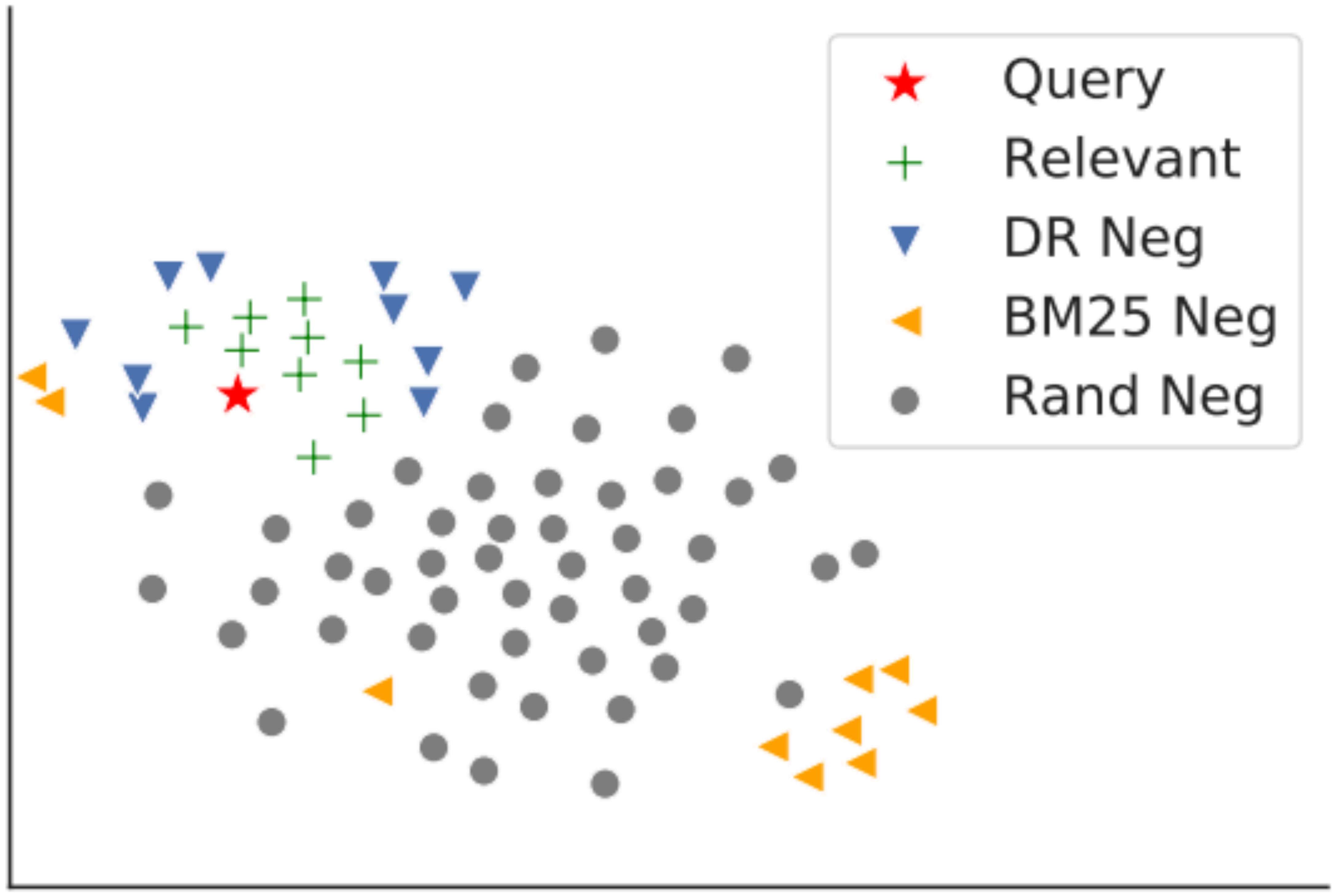
Learn the Dense Representation

Contrastive learning



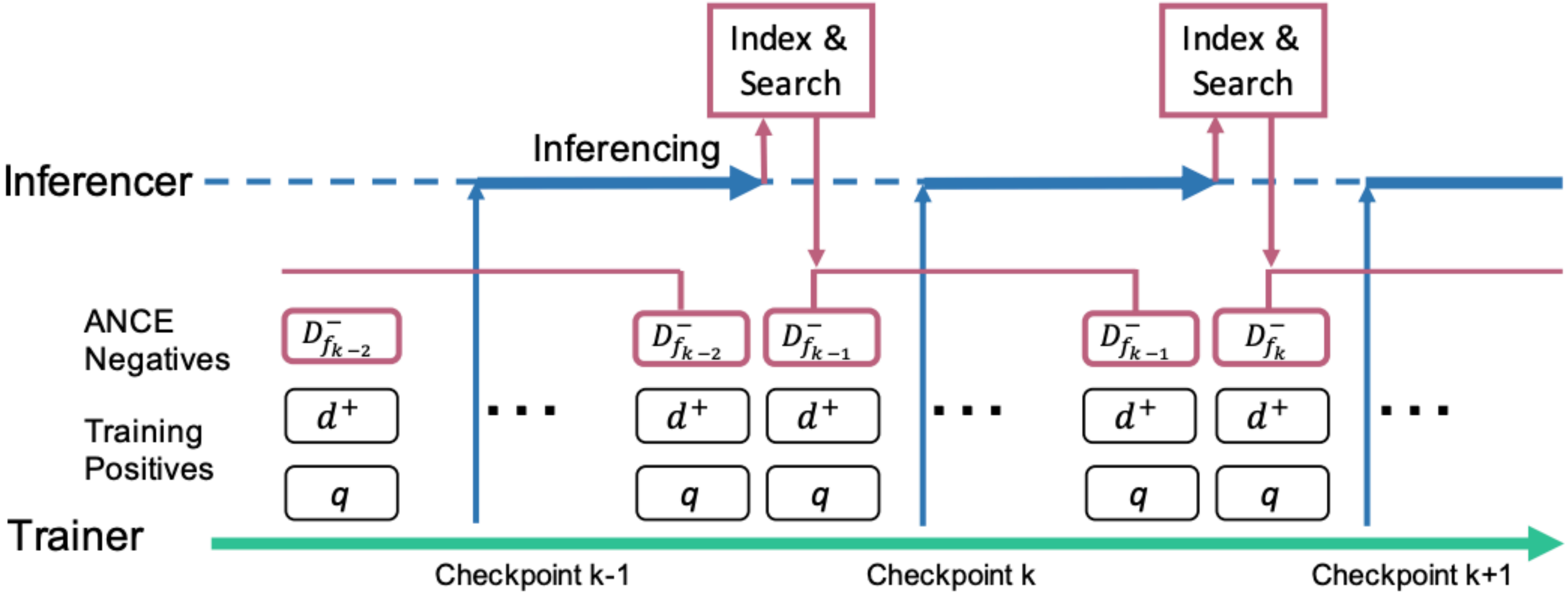
Learn the Dense Representation

Hard negatives



Learn the Dense Representation

ANCE: Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval (Xiong et al., 2019)



L Xiong, C Xiong, Y Li, K Tang, J Liu, P Bennett, J Ahmed, A Overwijk, Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval, ICLR, 2021

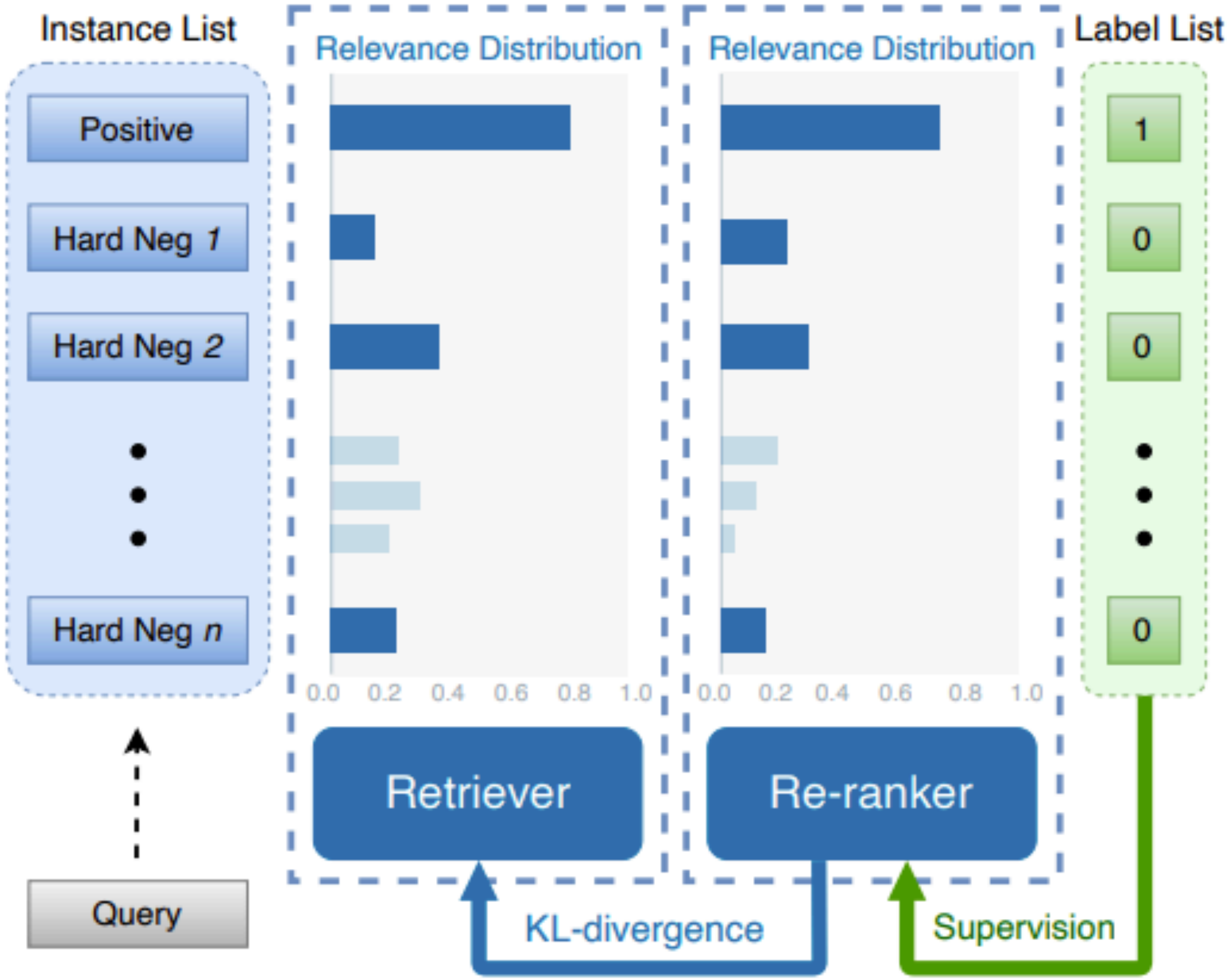
Learn the Dense Representation

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

ANCE: Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval (Xiong et al., 2019)

	MARCO Dev Passage Retrieval	
	MRR@10	**Recall@1k**
Dense Retrieval		
Rand Neg	0.261	0.949
NCE Neg	0.256	0.943
BM25 Neg	0.299	0.928
DPR (BM25 + Rand Neg)	0.311	0.952
BM25 $\rightarrow$ Rand	0.280	0.948
BM25 $\rightarrow$ NCE Neg	0.279	0.942
BM25 $\rightarrow$ BM25 + Rand	0.306	0.939
ANCE (FirstP)	**0.330**	**0.959**

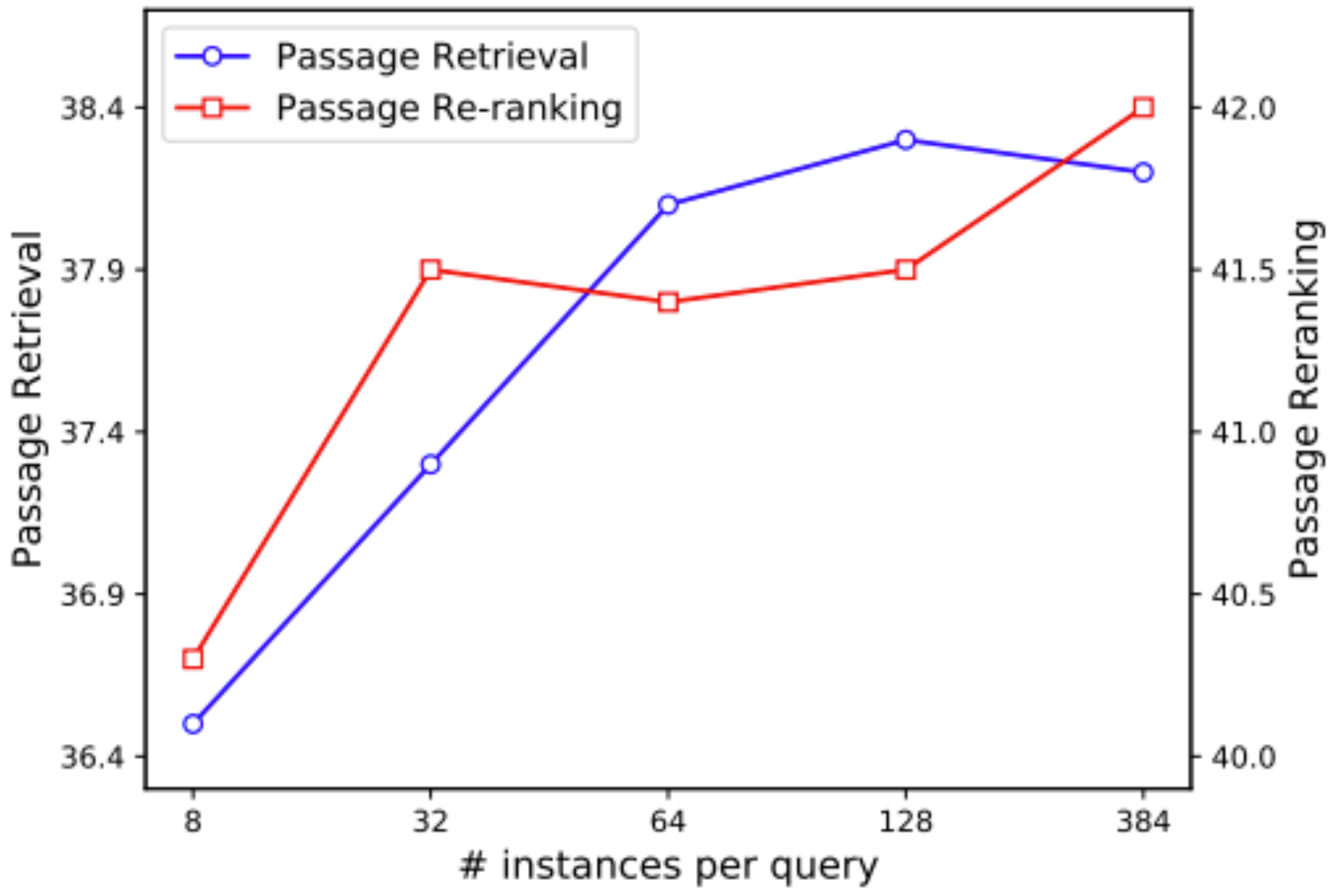
Knowledge distillation from Cross-encoder



R Ren, Y Qu, J Liu, W Zhao, Q She, H Wu, H Wang, J Wen, RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking, emnlp, 2021

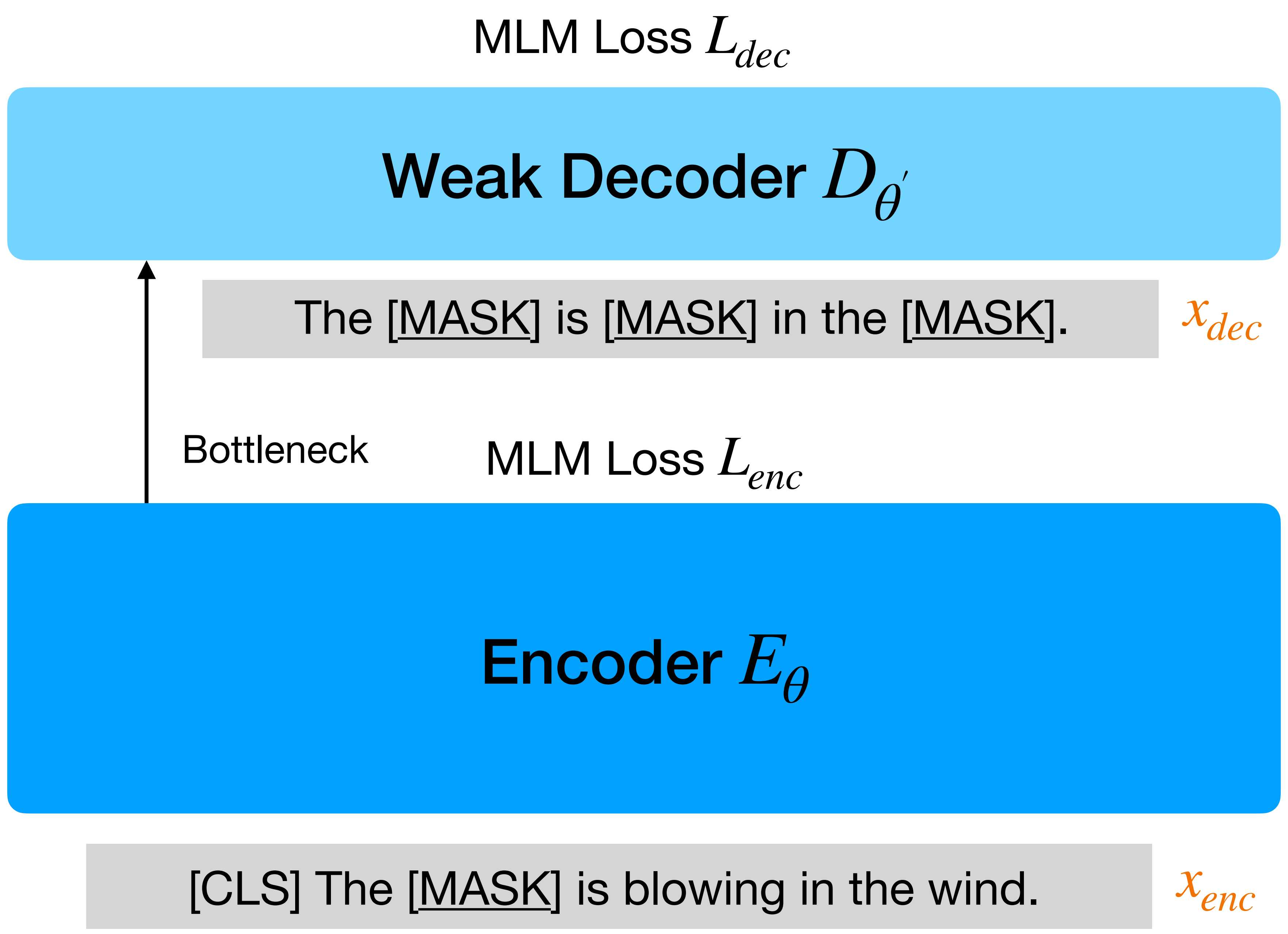
THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Knowledge distillation from Cross-encoder



Learn the Dense Representation

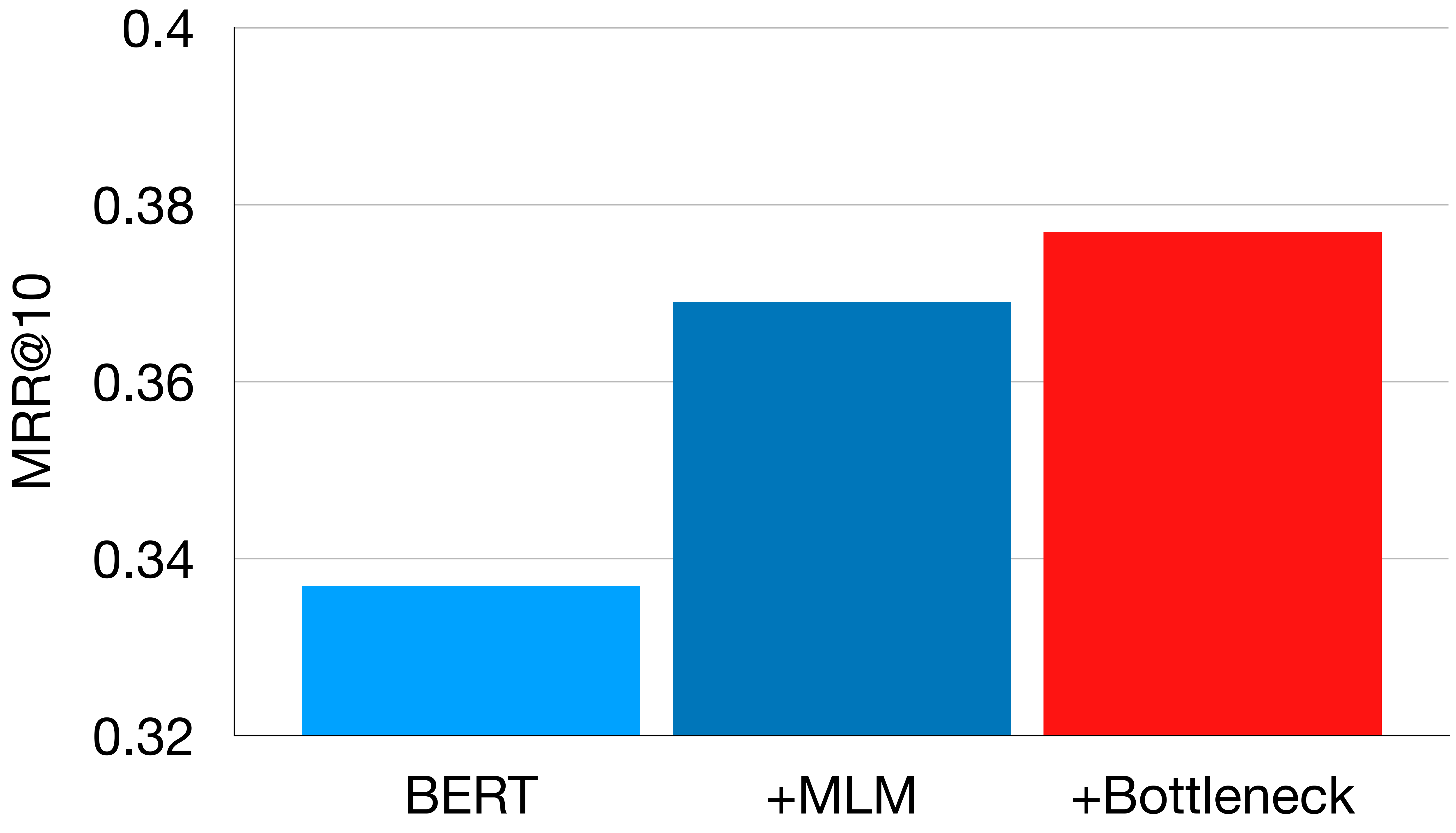
Bottlenecked pre-training



Learn the Dense Representation

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

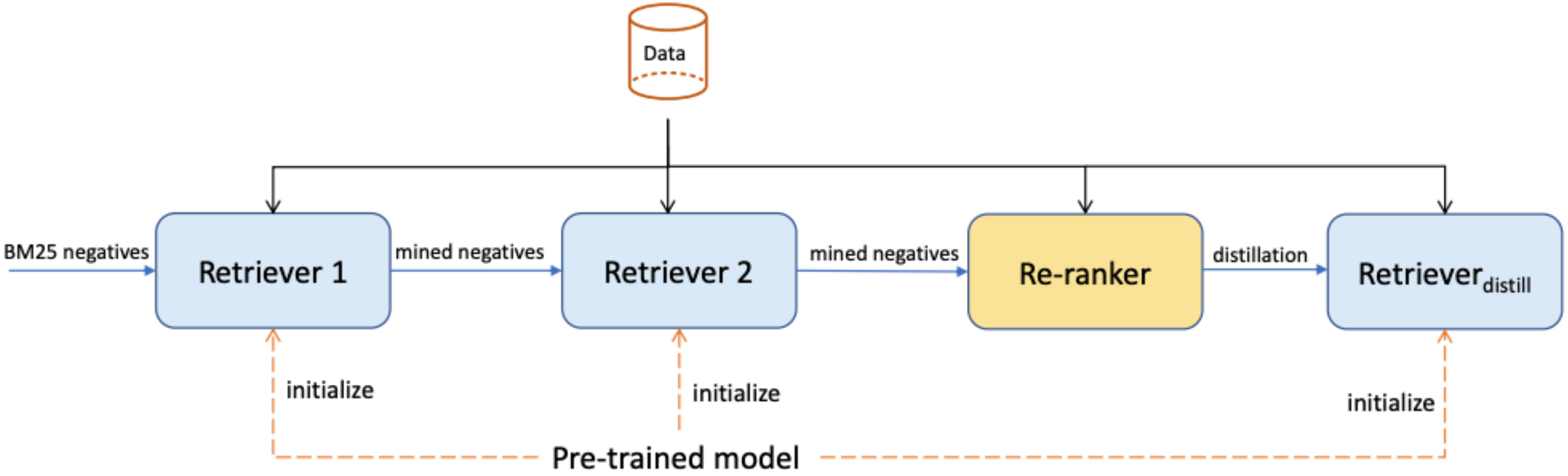
Bottlenecked pre-training



Learn the Dense Representation

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

SOTA training pipeline: SimLM (Wang et al., 2023)



L Wang, N Yang, X Huang, B Jiao, L Yang, D Jiang, R Majumder, F Wei, SimLM: Pre-training with Representation Bottleneck for Dense Passage Retrieval, ACL, 2023

Learned Sparse representation

The University of Queensland
Australia

- Bag-of-words (BoW):

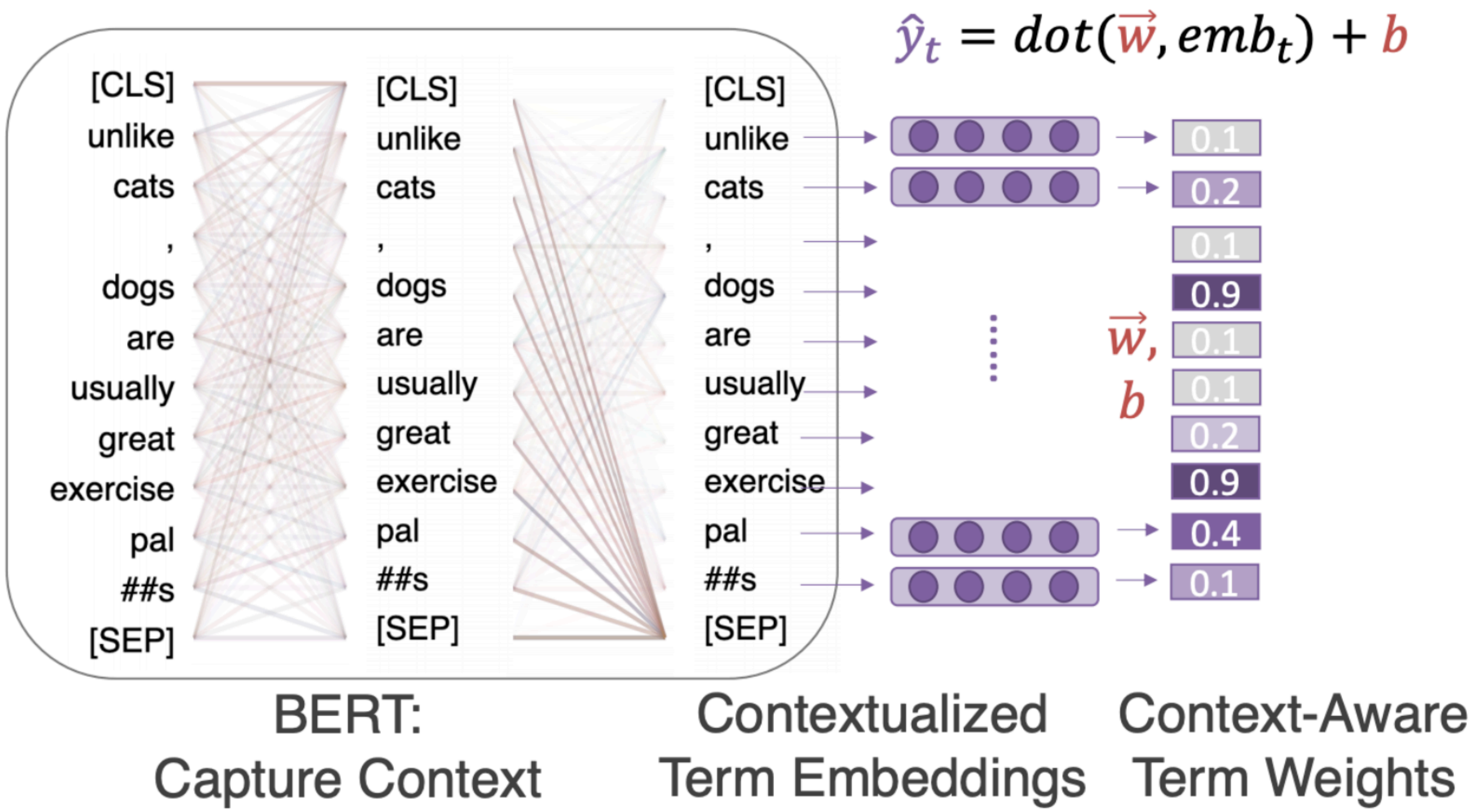
Query: dogs

Doc: *Unlike cats, dogs are usually great exercise pals….*

Term frequencies: {**unlike**: 1, **cats**: 1, **dogs**: 1, **exercise**: 1, **pals**: 1, …}

Learned Sparse representation

DeepCT (Dai and Jamie, 2020)

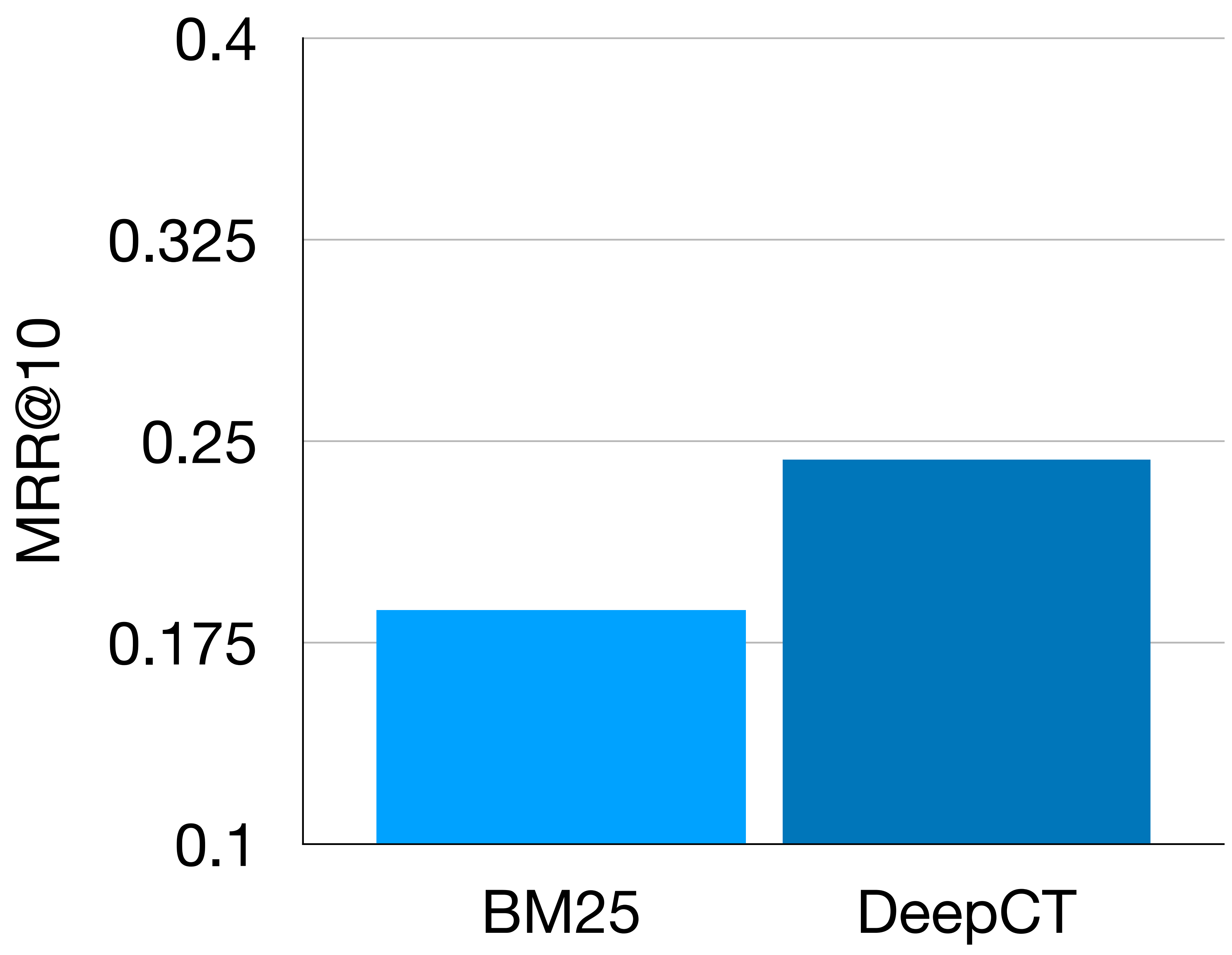


"Unlike cats, dogs are usually great exercise pals. Many breeds enjoy running and hiking, and will happily trek along on any trip. Exercise time varies…"

Z Dai, and J Callan, Context-Aware Term Weighting For First Stage Passage Retrieval, SIGIR, 2022

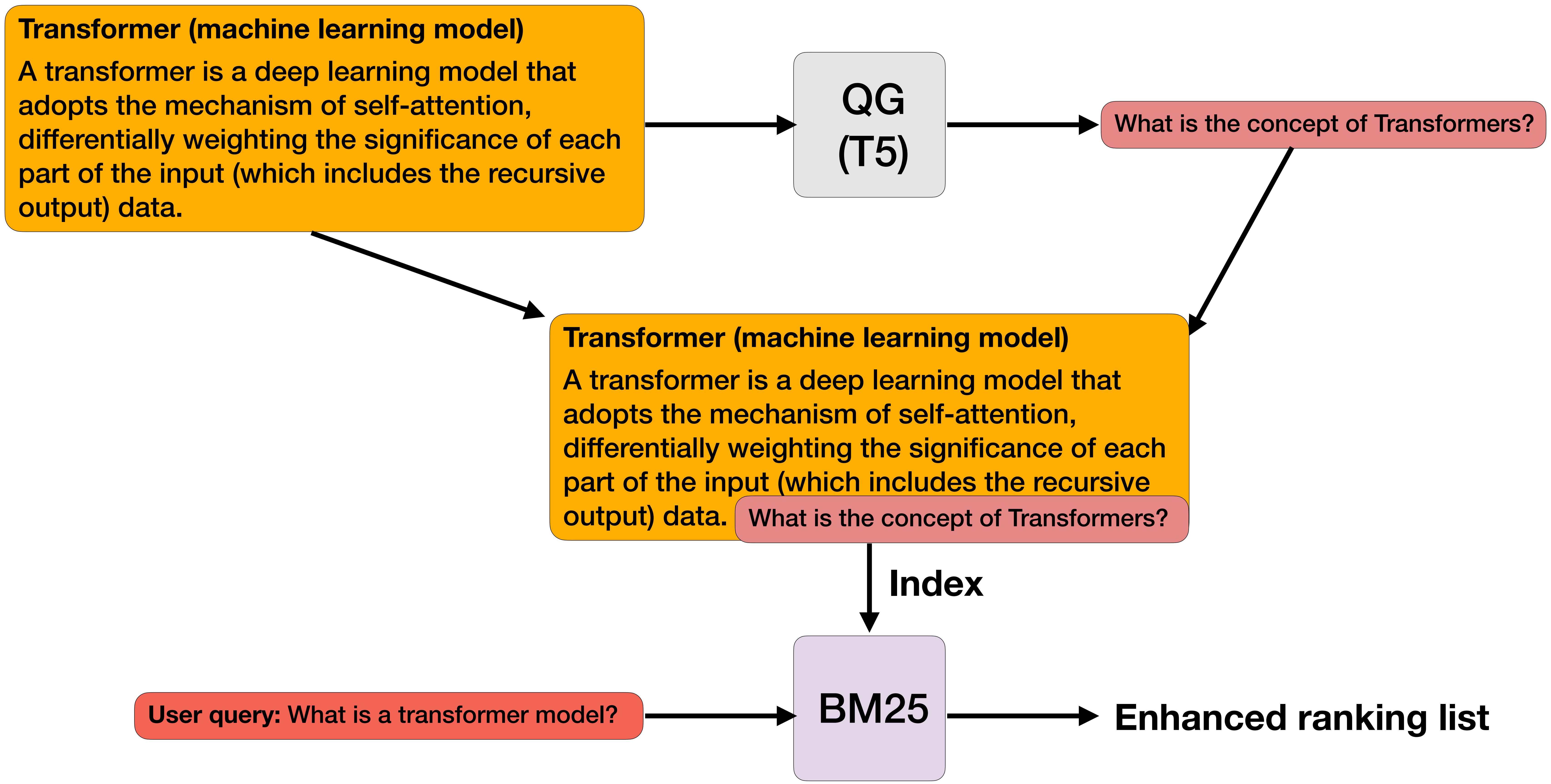
Learned Sparse representation

DeepCT (Dai and Jamie, 2020)



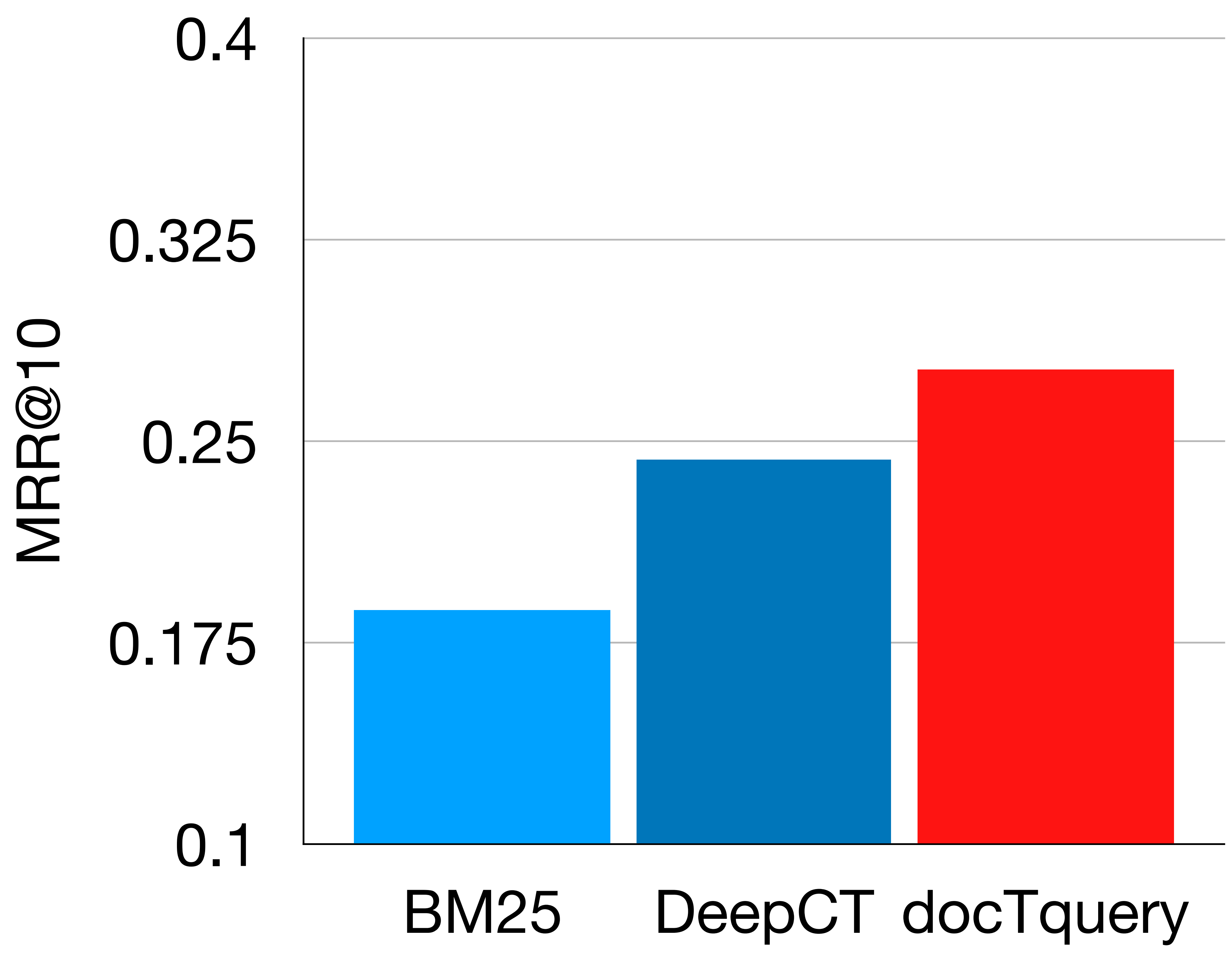
Learned Sparse representation

docTquery (Nogueira et al., 2019)



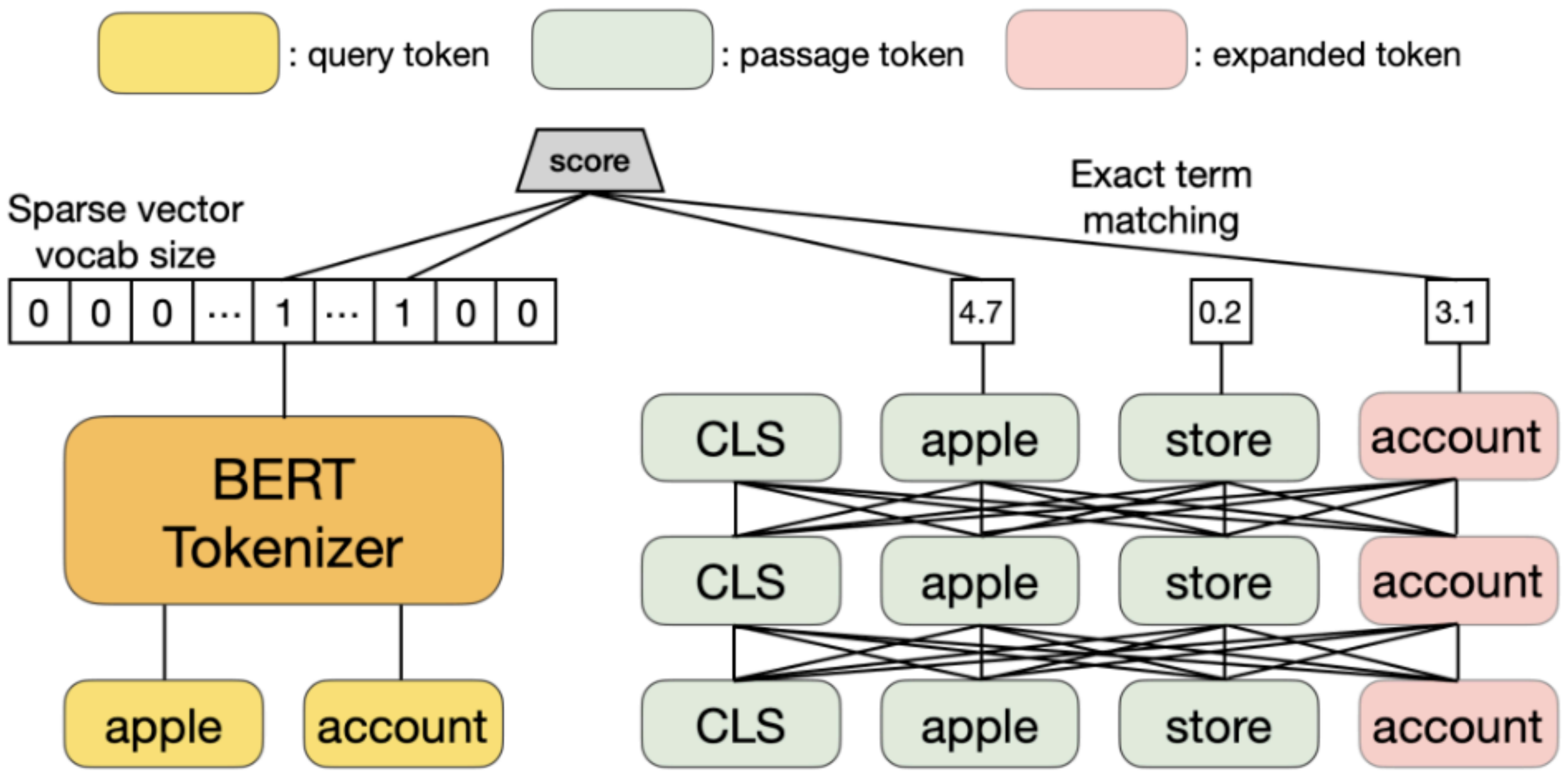
Nogueira, Rodrigo, Wei, Yang, Jimmy, Lin, and Kyunghyun, Cho. "Document expansion by query prediction".*arXiv preprint,* 2019.

docTquery (Nogueira et al., 2019)



Learned Sparse representation

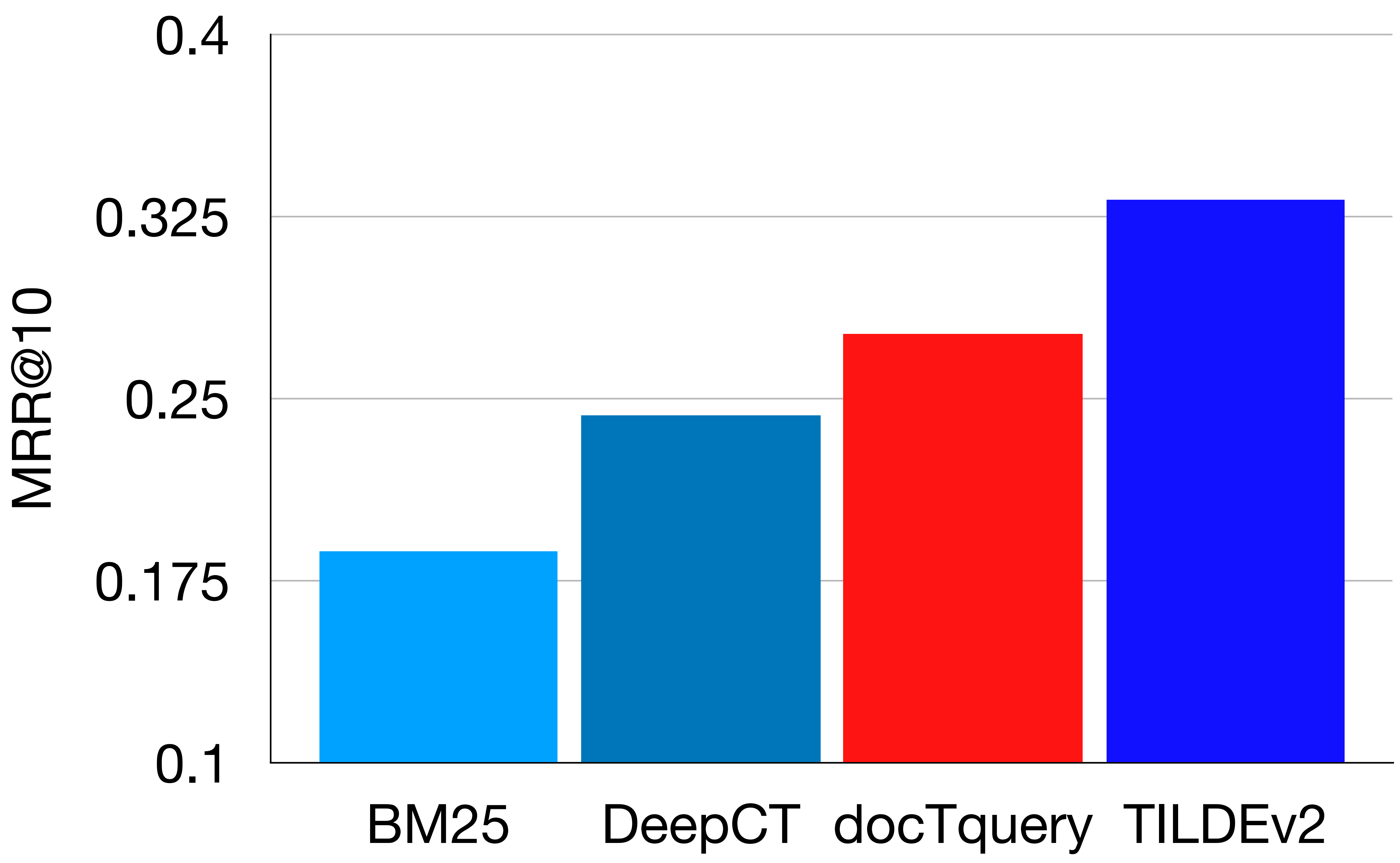
TILDEv2 (Zhuang and Guido, 2022)



S Zhuang and G Zuccon, Fast Passage Re-ranking with Contextualized Exact Term Matching and Efficient Passage Expansion, arXiv, 2022

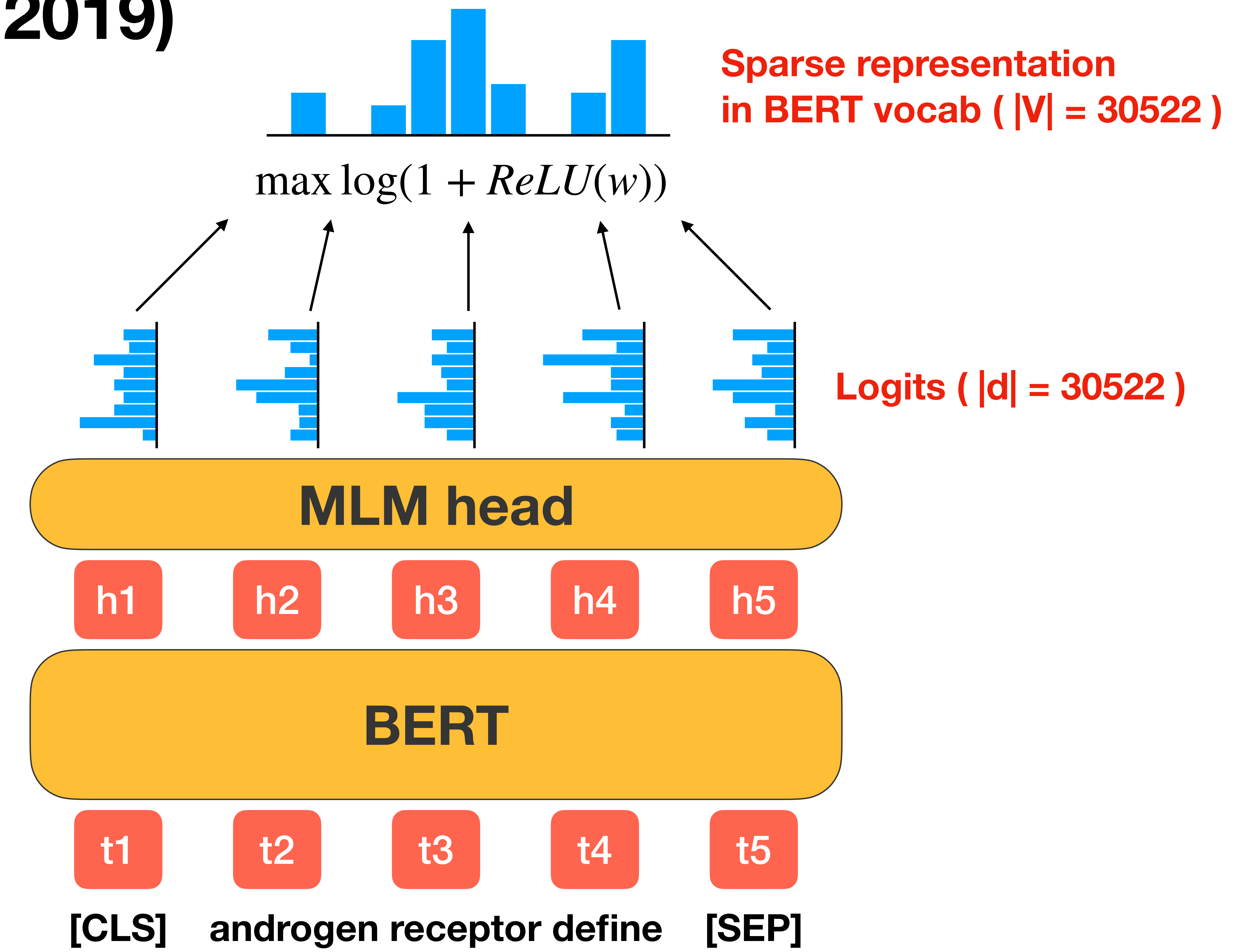
Learned Sparse representation

TILDEv2 (Zhuang and Guido, 2022)



Learned Sparse representation

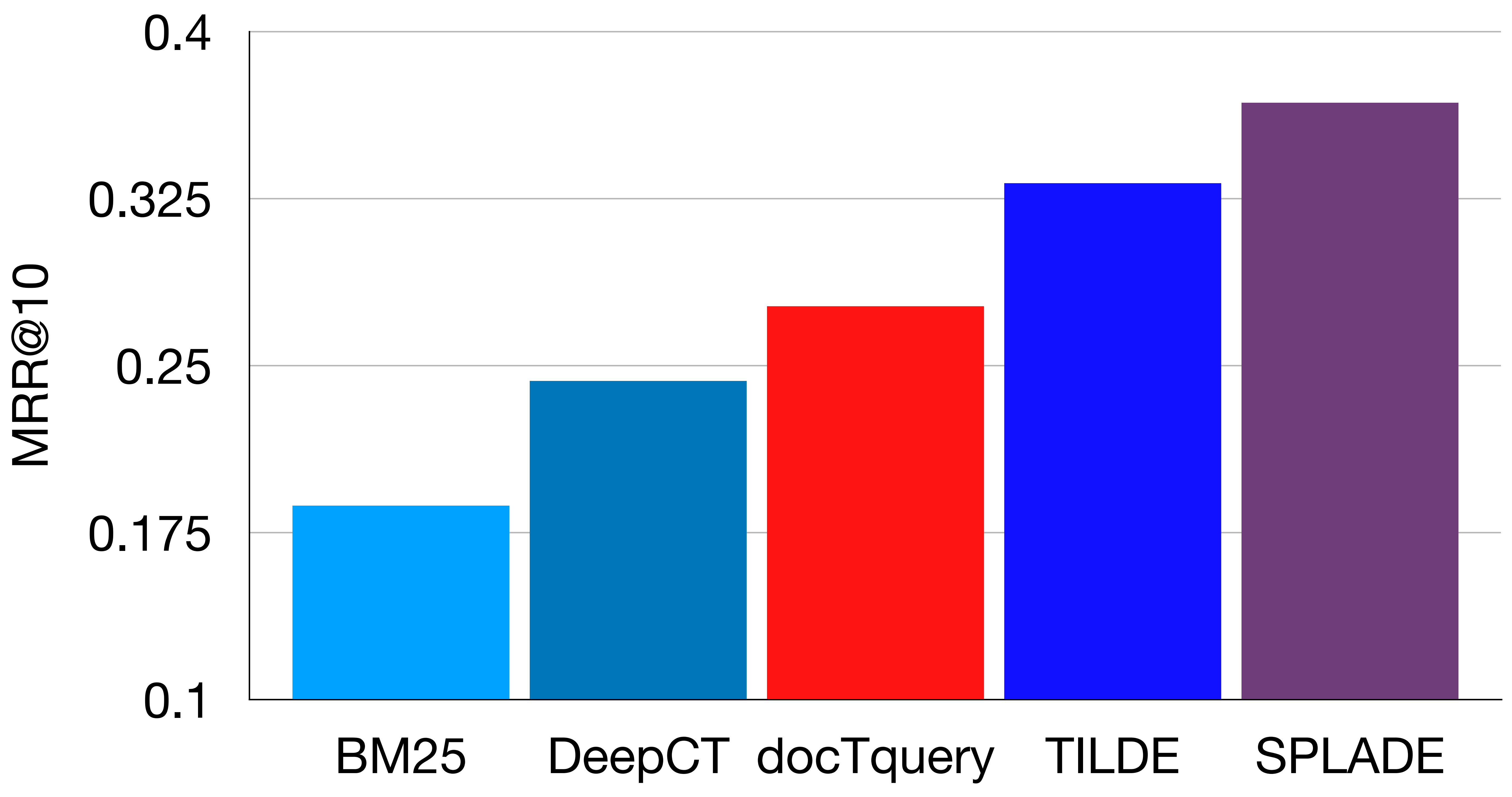
SPLADE (Formal, et al., 2019)



T Formal, B Piwowarski, C Lassance, and Clinchant, SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval, 2021

Learned Sparse representation

SPLADE (Formal, et al., 2019)



THE UNIVERSITY OF QUEENSLAND AUSTRALIA

SPLADE's wacky weights (Joel et al., 2021)

Query: androgen receptor define

('##rogen', 251) ('receptor', 242) ('and', 225) ('receptors', 189)
('hormone', 179) ('definition', 162) ('meaning', 99) ('genus', 89)
('is', 70) (',', 68) ('define', 59) ('the', 56) ('drug', 53) ('for', 46)
('ring', 38) ('gene', 37) ('are', 32) ('god', 25) ('what', 18) ('##rus', 15)
('purpose', 12) ('defined', 10) ('doing', 8) ('a', 4) ('goal', 4)

Blue: original input query tokens
Orange: alternate inflections on those original tokens
Pink: expended new tokens

J Mackenzie, A Trotman, J Lin, Wacky weights in learned sparse representations and the revenge of score-at-a-time query evaluation, 2021

Learned Sparse representation

SPLADE's wacky weights (Joel et al., 2021)

Query: androgen receptor define

('##rogen', 251) ('receptor', 242) ('and', 225) ('receptors', 189)
('hormone', 179) ('definition', 162) ('meaning', 99) ('genus', 89)
('is', 70) (',', 68) ('define', 59) ('the', 56) ('drug', 53) ('for', 46)
('ring', 38) ('gene', 37) ('are', 32) ('god', 25) ('what', 18) ('##rus', 15)
('purpose', 12) ('defined', 10) ('doing', 8) ('a', 4) ('goal', 4)

Blue: original input query tokens
Orange: alternate inflections on those original tokens
Pink: expended new tokens

J Mackenzie, A Trotman, J Lin, Wacky weights in learned sparse representations and the revenge of score-at-a-time query evaluation, 2021

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

SPLADE can learn good representation with any vocabulary (Joel et al., 2023)

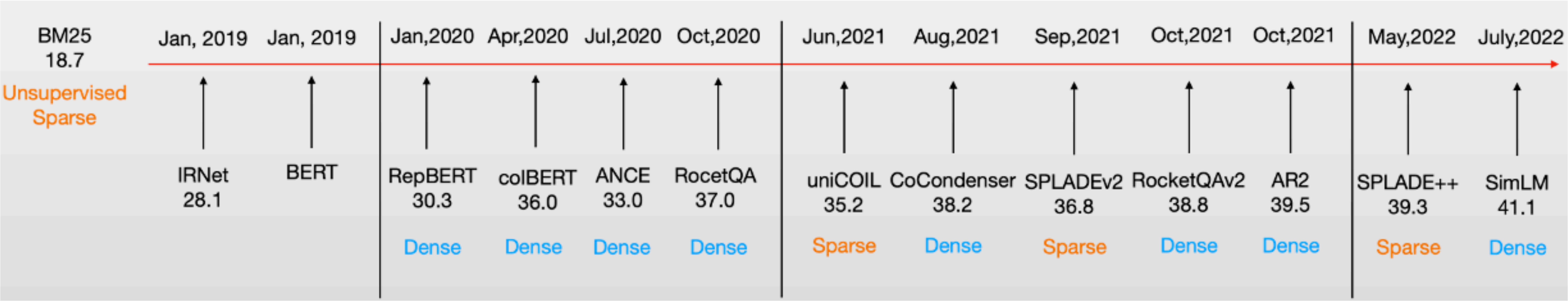
- Only allow to assign weights to stopwords (|v|=150)

{
"docid": 0,
"weights": {"i": 29, "the": 43, "of": 62, "was": 138, "for": 7, "that": 44, "had": 143, "an": 74, "were": 118, "have": 37, "has": 16, "who": 5, "after": 1, "into": 12, "its": 45, "no": 142, "what": 96, "we": 63, "through": 58, "most": 50, "did": 146, "being": 12, "didn": 15, "because": 139, "should": 43, "why": 12, "having": 54, "am": 69, "further": 49, "doing": 63, "itself": 74, "themselves": 70, "ourselves": 51}
}

J Mackenzie, S Zhuang, G Zuccon, Exploring the Representation Power of SPLADE Models, ICTIR, 2023

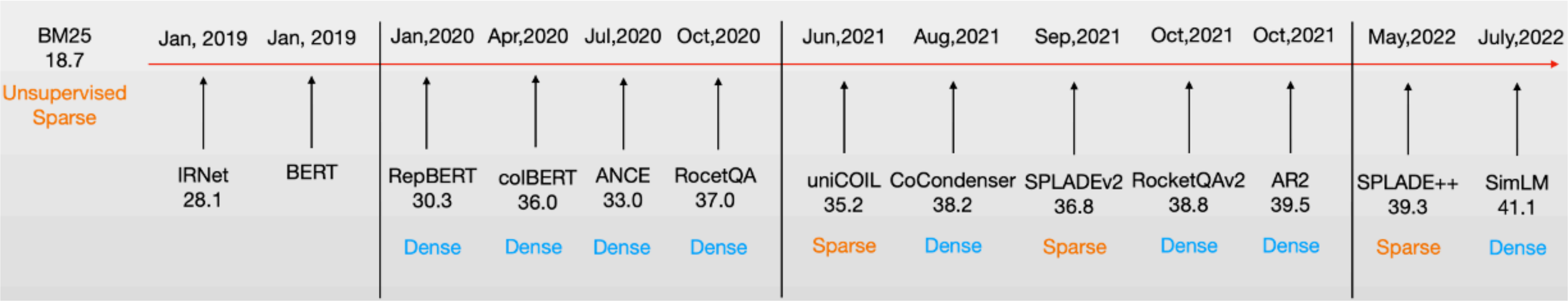
The problem of BERT-based rankers

So far..



The problem of BERT-based rankers

So far..



Trained and tested on MS MARCO: in domain setting

The problem of BERT-based rankers

Not effective under transfer domain setting

Model (→)	Lexical	Sparse			Dense			
Dataset (↓)	**BM25**	**DeepCT**	**SPARTA**	**docT5query**	**DPR**	**ANCE**	**TAS-B**	**GenQ**
MS MARCO	0.228	0.296‡	0.351‡	0.338‡	0.177	0.388‡	0.408‡	0.408‡
TREC-COVID	0.656	0.406	0.538	0.713	0.332	0.654	0.481	0.619
BioASQ	0.465	0.407	0.351	0.431	0.127	0.306	0.383	0.398
NFCorpus	0.325	0.283	0.301	0.328	0.189	0.237	0.319	0.319
NQ	0.329	0.188	0.398	0.399	0.474‡	0.446	0.463	0.358
HotpotQA	0.603	0.503	0.492	0.580	0.391	0.456	0.584	0.534
FiQA-2018	0.236	0.191	0.198	0.291	0.112	0.295	0.300	0.308
Signal-1M (RT)	0.330	0.269	0.252	0.307	0.155	0.249	0.289	0.281
TREC-NEWS	0.398	0.220	0.258	0.420	0.161	0.382	0.377	0.396
Robust04	0.408	0.287	0.276	0.437	0.252	0.392	0.427	0.362
ArguAna	0.315	0.309	0.279	0.349	0.175	0.415	0.429	**0.493**
Touché-2020	**0.367**	0.156	0.175	0.347	0.131	0.240	0.162	0.182
CQADupStack	0.299	0.268	0.257	0.325	0.153	0.296	0.314	0.347
Quora	0.789	0.691	0.630	0.802	0.248	0.852	0.835	0.830
DBPedia	0.313	0.177	0.314	0.331	0.263	0.281	0.384	0.328
SCIDOCS	0.158	0.124	0.126	0.162	0.077	0.122	0.149	0.143
FEVER	0.753	0.353	0.596	0.714	0.562	0.669	0.700	0.669
Climate-FEVER	0.213	0.066	0.082	0.201	0.148	0.198	0.228	0.175
SciFact	0.665	0.630	0.582	0.675	0.318	0.507	0.643	0.644
Avg. Performance vs. BM25		**- 27.9%**	**- 20.3%**	**+ 1.6%**	**- 47.7%**	**- 7.4%**	**- 2.8%**	**- 3.6%**

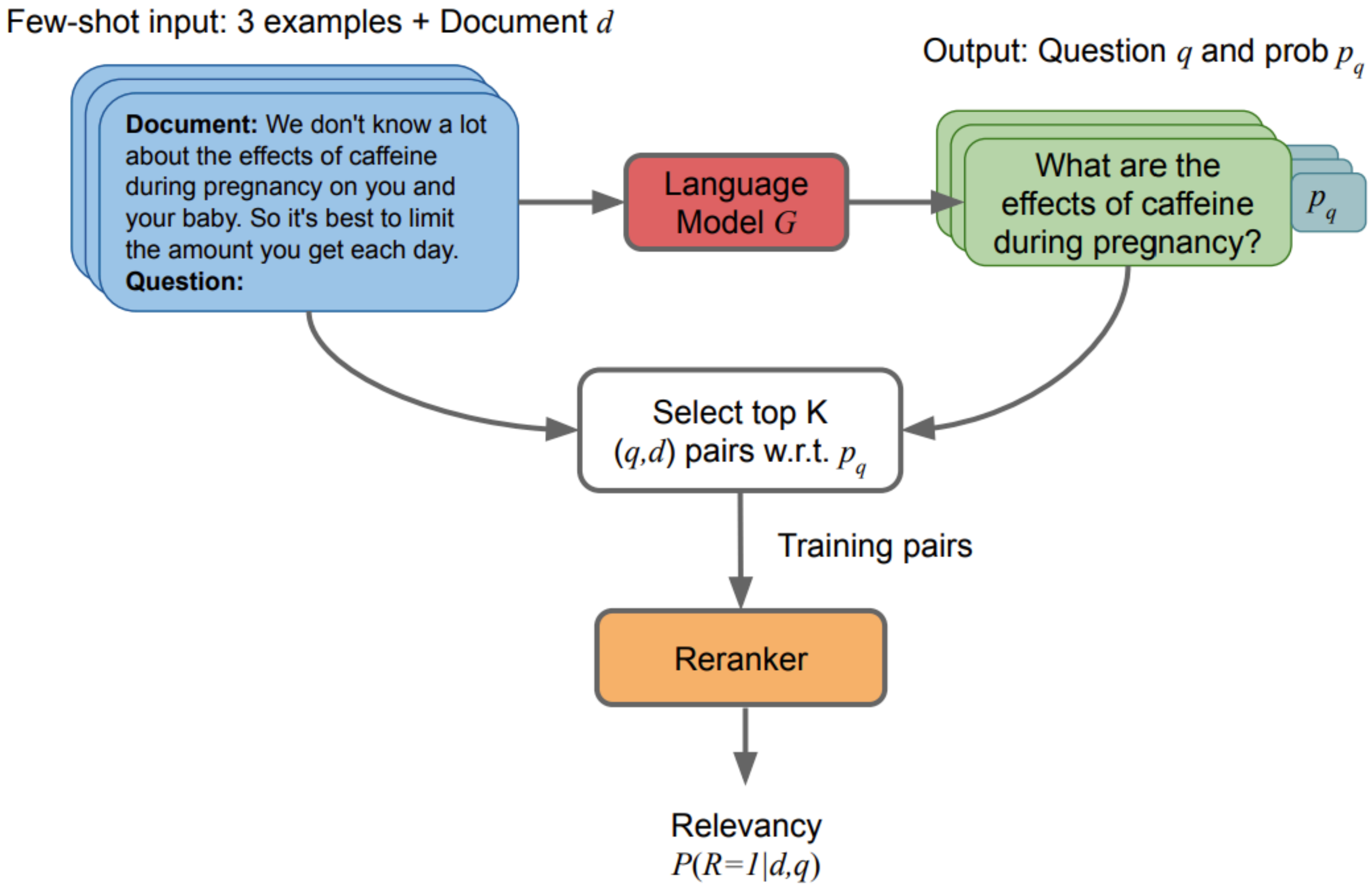
N Thakur, N Reimers, A Rücklé, A Srivastava, BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models, NIPS, 2021

LLM-based methods

- 2019 ~ 2022: BERT, T5…, less than 1B parameters — In domain setting
- 2022 ~ current: GPT-3/4, LLaMA…, 7B - 175B. — Zero-shot setting

LLM-based methods

InPars (Bonifacio et al., 2022)



L Bonifacio, H Abonizio, M Fadaee, R Nogueira, InPars: Data Augmentation for Information Retrieval using Large Language, SIGIR, 2022

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

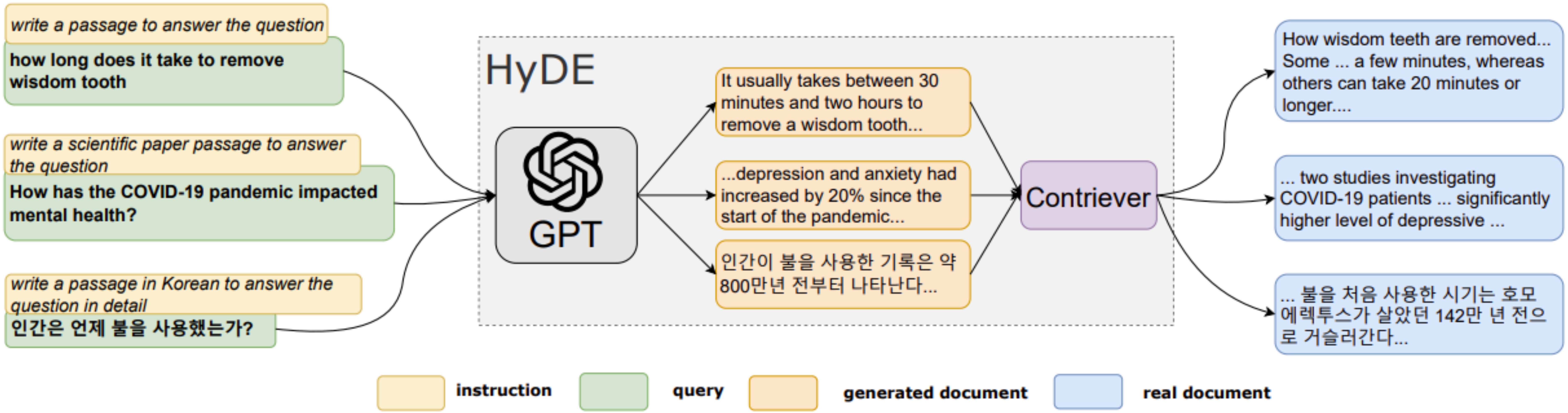
InPars (Bonifacio et al., 2022)

		MARCO	TREC-DL 2020		Robust04		NQ	TRECC
		MRR@10	MAP	nDCG@10	MAP	nDCG@20	nDCG@10	nDCG@10
	Unsupervised							
(1)	BM25	0.1874	0.2876	0.4876	0.2531	0.4240	0.3290	0.6880
(2)	Contriever (Izacard et al., 2021)	-	-	-	-	-	0.2580	0.2740
(3)	cpt-text (Neelakantan et al., 2022)	0.2270	-	-	-	-	-	0.4270
	OpenAI Search reranking 100 docs from BM25							
(4)	Ada (300M)	$	0.3141	0.5161	0.2691	0.4847	0.4092	0.6757
(5)	Curie (6B)	$	0.3296	0.5422	0.2785	0.5053	0.4171	0.7251
(6)	Davinci (175B)	$	0.3163	0.5366	0.2790	0.5103	$	0.6918
	InPars (ours)							
(7)	monoT5-220M	0.2585	0.3599	0.5764	0.2490	0.4268	0.3354	0.6666
(8)	monoT5-3B	**0.2967**	**0.4334**	**0.6612**	**0.3180**	**0.5181**	**0.5133**	**0.7835**

L Bonifacio, H Abonizio, M Fadaee, R Nogueira, InPars: Data Augmentation for Information Retrieval using Large Language, SIGIR, 2022

LLM-based methods

HyDE (Gao et al., 2023)



L Gao, X Ma, J Lin, J Callan, Precise Zero-Shot Dense Retrieval without Relevance Labels, ACL, 2023

LLM-based methods

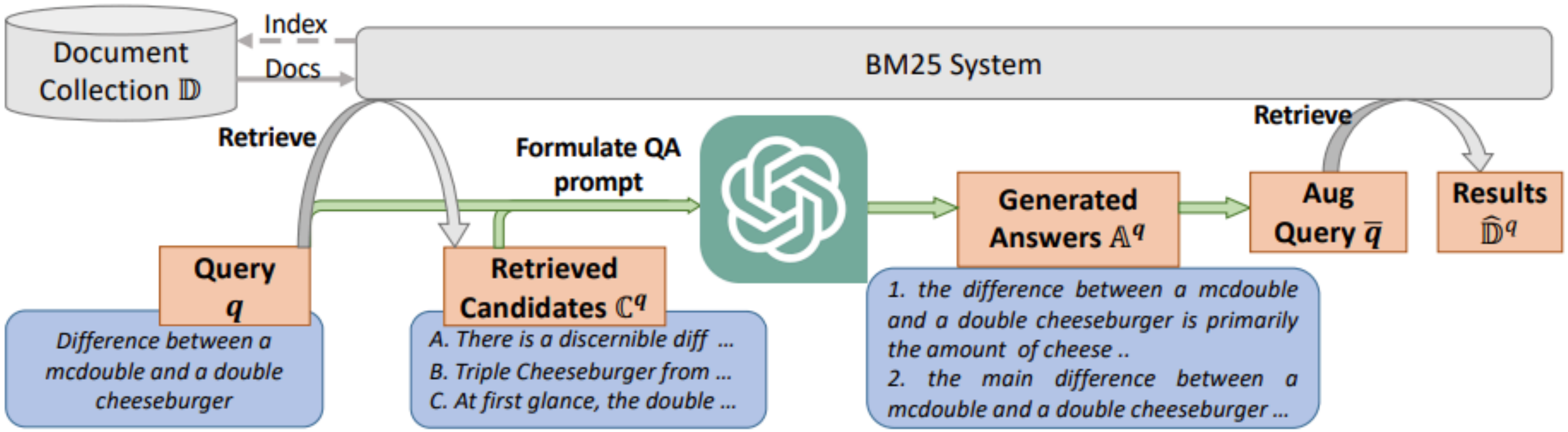
HyDE (Gao et al., 2023)

	Scifact	Arguana	Trec-Covid	FiQA	DBPedia	TREC-NEWS
			nDCG@10			
w/o relevance judgement						
BM25	67.9	39.7	**59.5**	23.6	31.8	39.5
Contriever	64.9	37.9	27.3	24.5	29.2	34.8
HyDE	**69.1**	**46.6**	59.3	**27.3**	**36.8**	**44.0**

L Gao, X Ma, J Lin, J Callan, Precise Zero-Shot Dense Retrieval without Relevance Labels, ACL, 2023

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

LameR (Shen et all., 2023)



T Shen, G Long, X Geng, C Tao, T Zhou, D Jiang, Large Language Models are Strong Zero-Shot Retriever, 2023

LLM-based methods

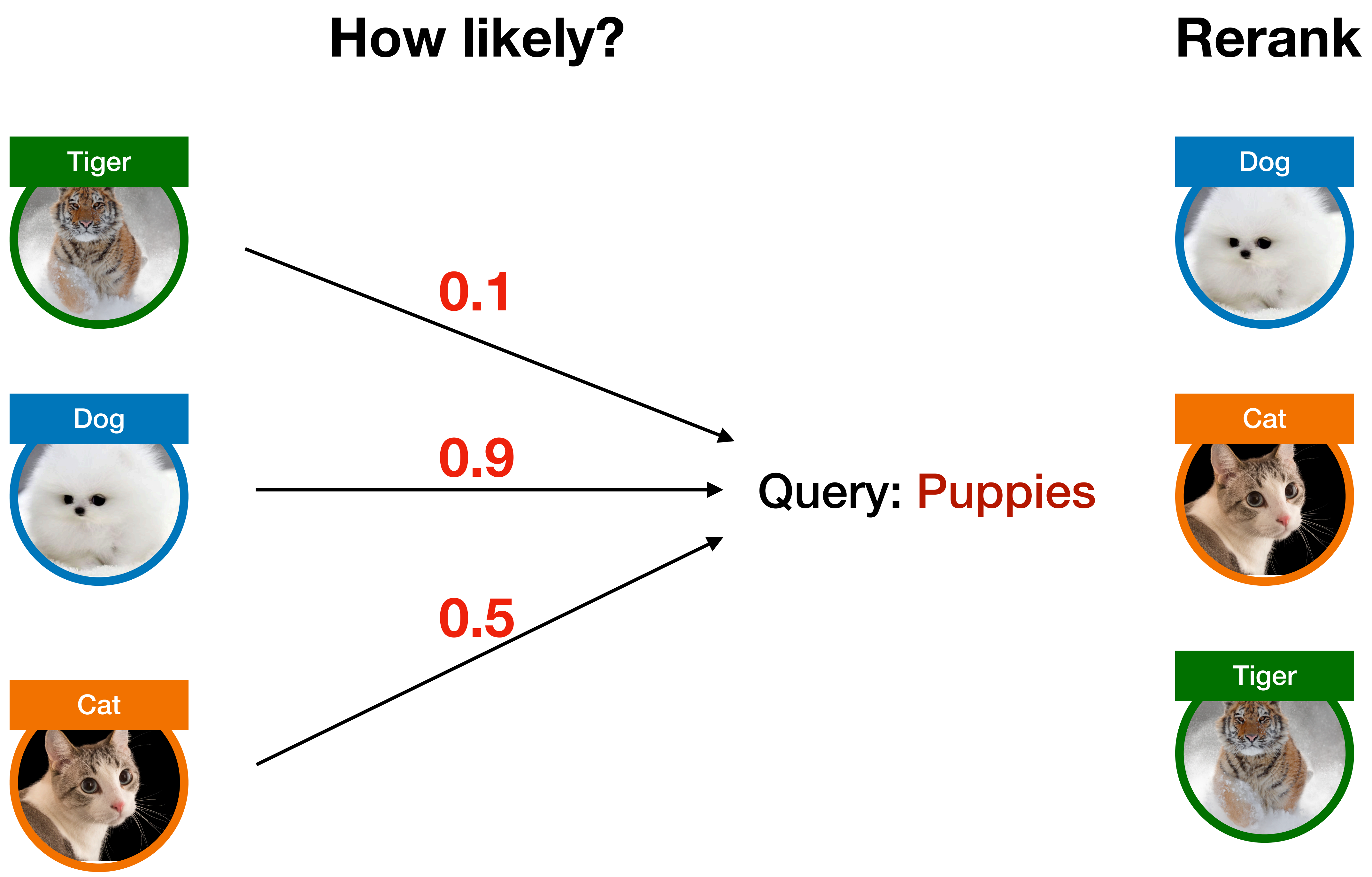
LameR (Shen et all., 2023)

	Scifact	Arguana	Trec-COVID	FiQA	DBPedia	TREC-NEWS
			nDCG@10			
w/o relevance judgment						
BM25	67.9	39.7	59.5	23.6	31.8	39.5
Contriever	64.9	37.9	27.3	24.5	29.2	34.8
HyDE	69.1	**46.6**	59.3	**27.3**	36.8	44.0
LameR (ours)	**73.5**	30.0	**72.5**	25.8	**38.7**	**49.9**

T Shen, G Long, X Geng, C Tao, T Zhou, D Jiang, Large Language Models are Strong Zero-Shot Retriever, 2023

The University of Queensland
Australia

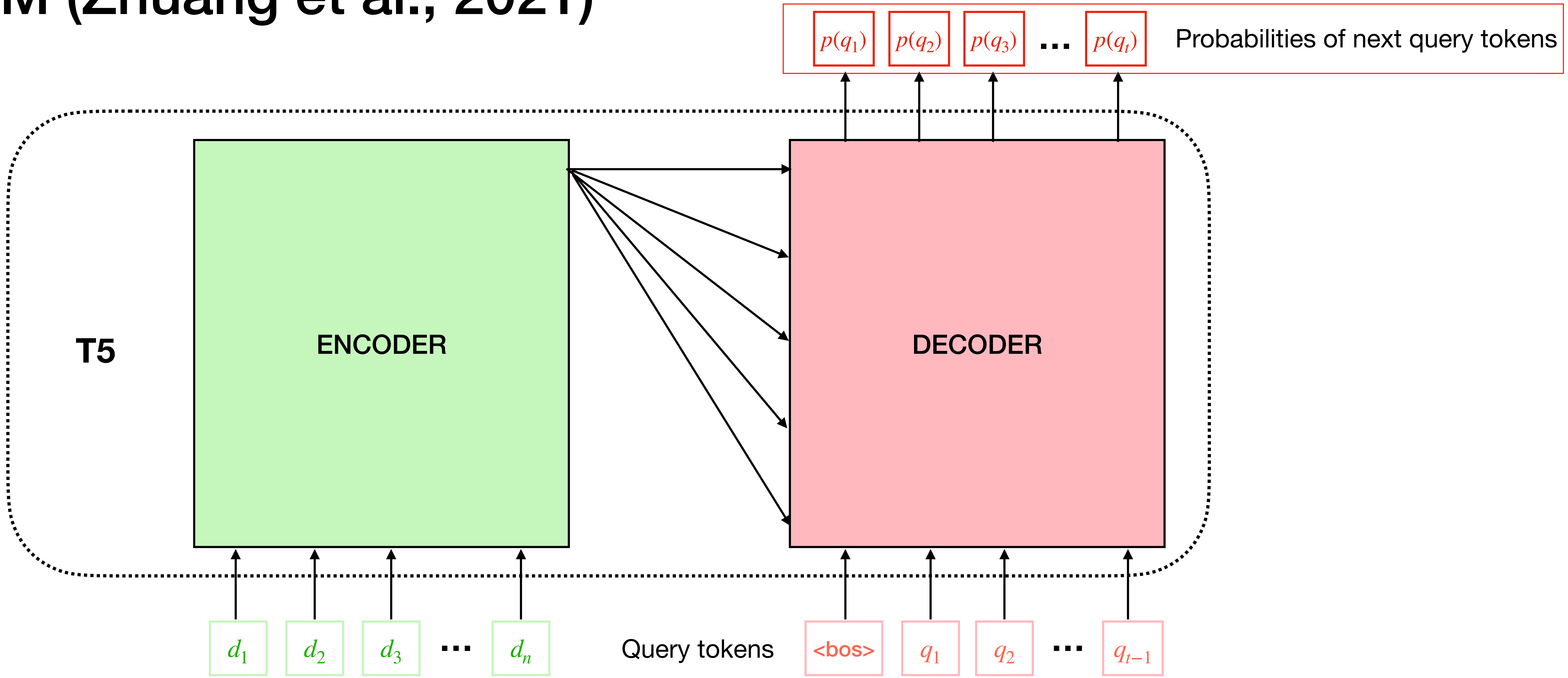
Query Likelihood models (QLMs) for document ranking.



LLM-based methods

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

T5-baed QLM (Zhuang et al., 2021)



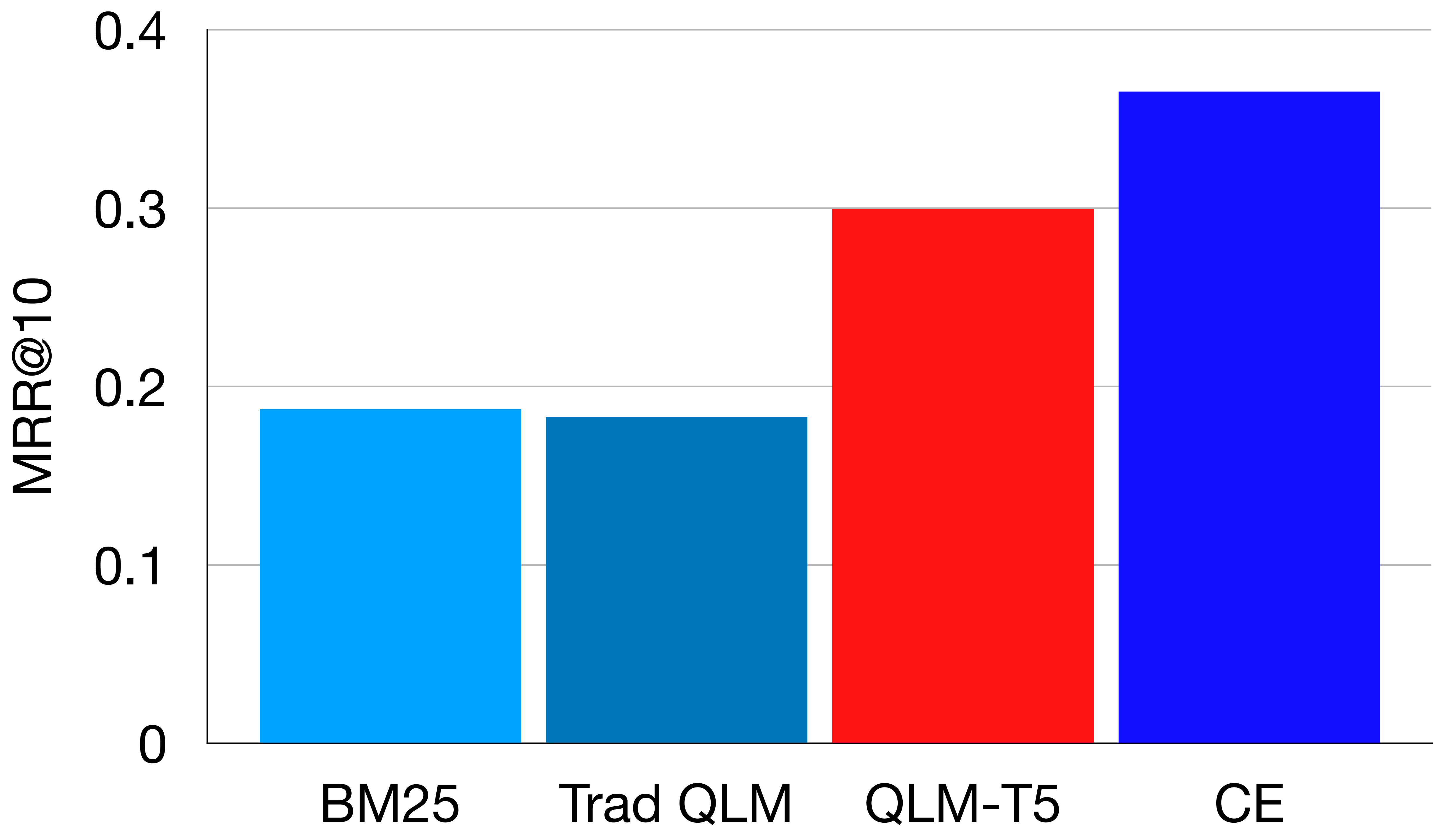
Rank by query likelihood: $P(Q \mid D) = \sum_{i}^{t} \log p(q_i)$

S Zhuang, H Li, G Zuccon, Deep query likelihood model for information retrieval, ECIR, 2021

LLM-based methods

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

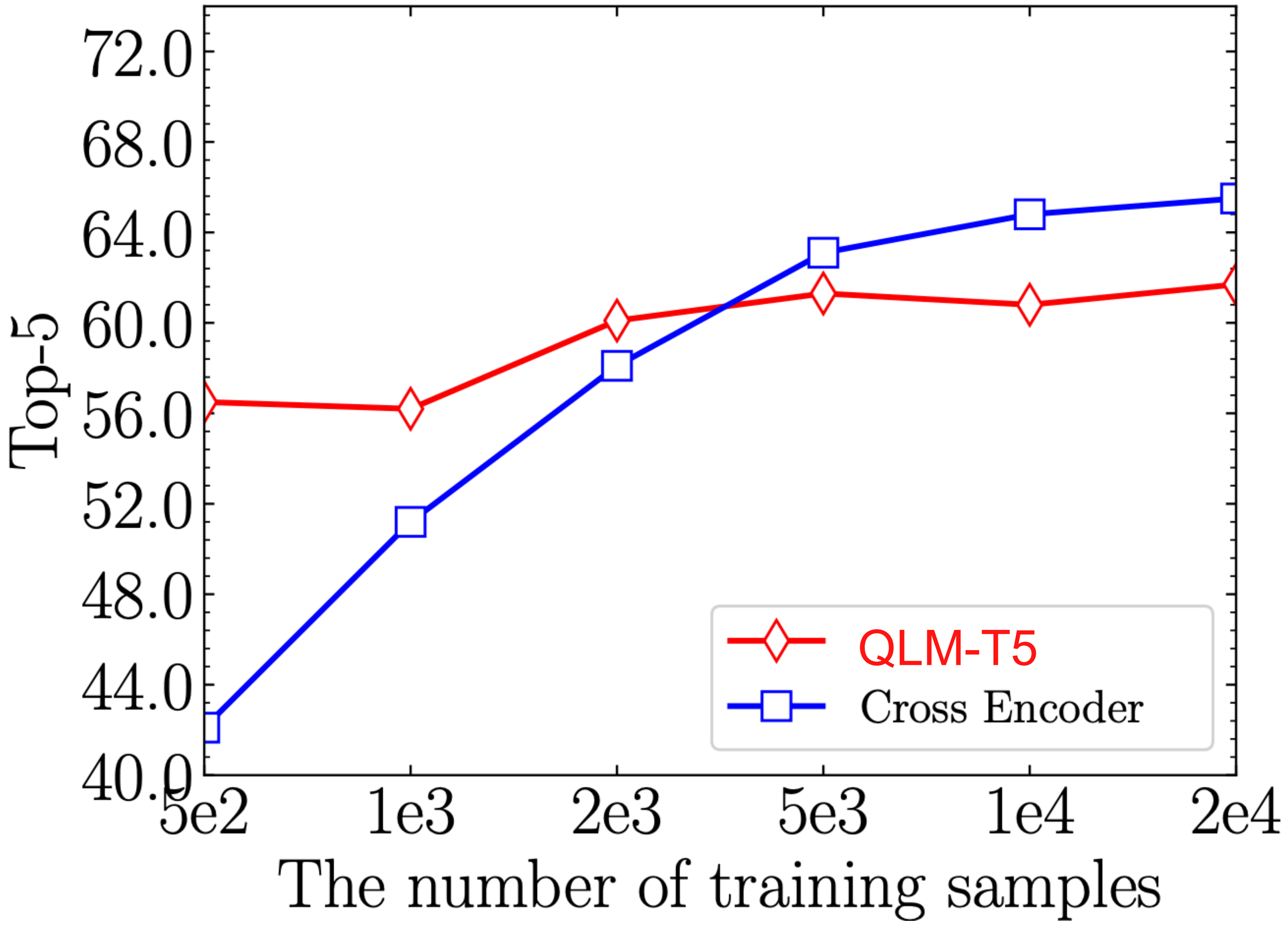
T5-baed QLM (Zhuang et al., 2021)



LLM-based methods

T5-baed QLM

- A follow up work shows ():



Houxing Ren, Linjun Shou, Ning Wu, Ming Gong, and Daxin Jiang. 2022. *Empowering Dual-Encoder with Query Generator for Cross-Lingual Dense Retrieval*. EMNLP2022

LLM-based methods

The University of Queensland Australia

LLM-based QLM for Zero-shot ranking

Methods	TRECC	DBpedia	FiQA	Robust04	Avg
Zero-shot Retrievers					
BM25	59.5	31.8	23.6	40.7	38.9
QLM-Dirichlet	50.8	29.5	20.5	40.7	35.4
Contriever	23.3	29.2	24.5	31.6	27.2
HyDE	58.2	37.2	26.6	41.8	41.0
Zero-shot QLM Re-rankers					
LLaMA-7B	69.4	39.9	41.5	53.6	51.1
LLaMA-13B	69.8	37.6	41.8	**54.2**	50.9
Falcon-7B	73.3	**41.7**	41.3	52.5	52.2
Falcon-40B	**75.2**	41.0	43.1	53.1	**53.1**

Conclusion & Future Directions

- 2019 ~ 2022: BERT, T5…, less than 1B parameters
  - Strong learned representation.
  - Effective and efficient with training data.
- 2022 ~ current: GPT-3/4, LLaMA…, 7B - 175B.
  - Strong zero-shot ability
- Current ~ future:
  - How to keep efficiency for LLM-based methods?
  - Interactive IR?