ColBERT:

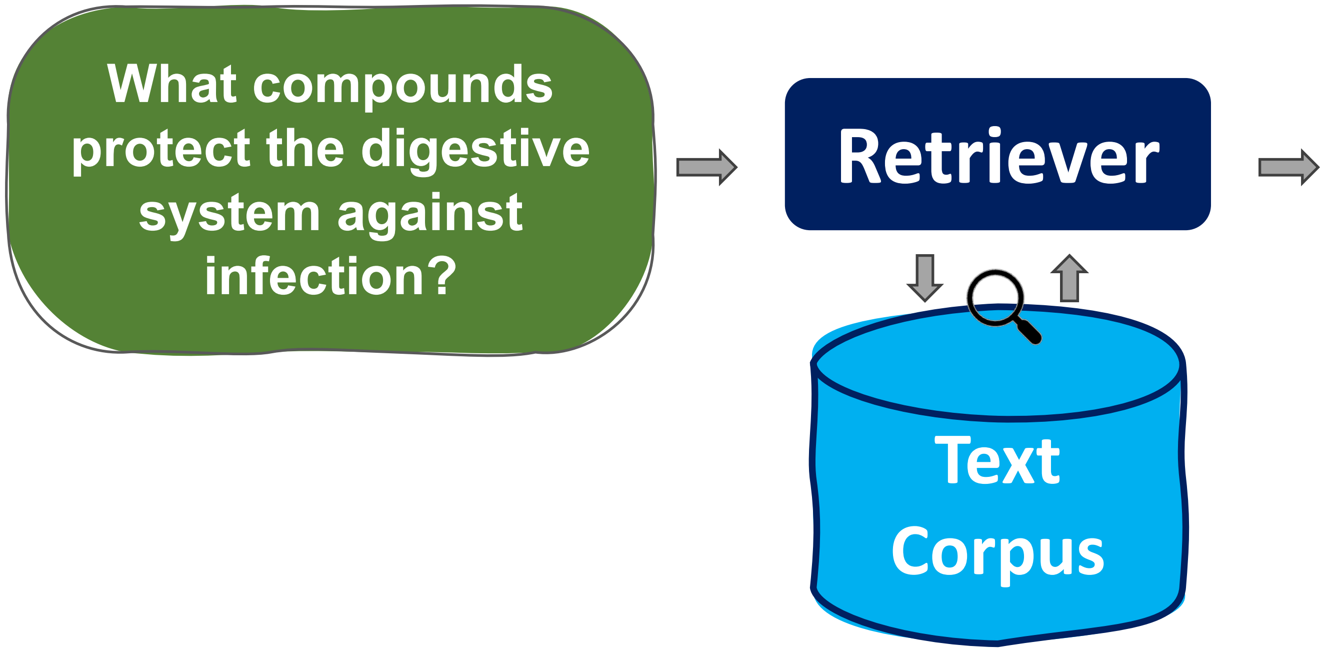
Effective and Efficient Search with **Late Interaction Models**

Omar Khattab

Oct 2023



Information Retrieval at a glance



2
Chemical barriers also protect against ...ct

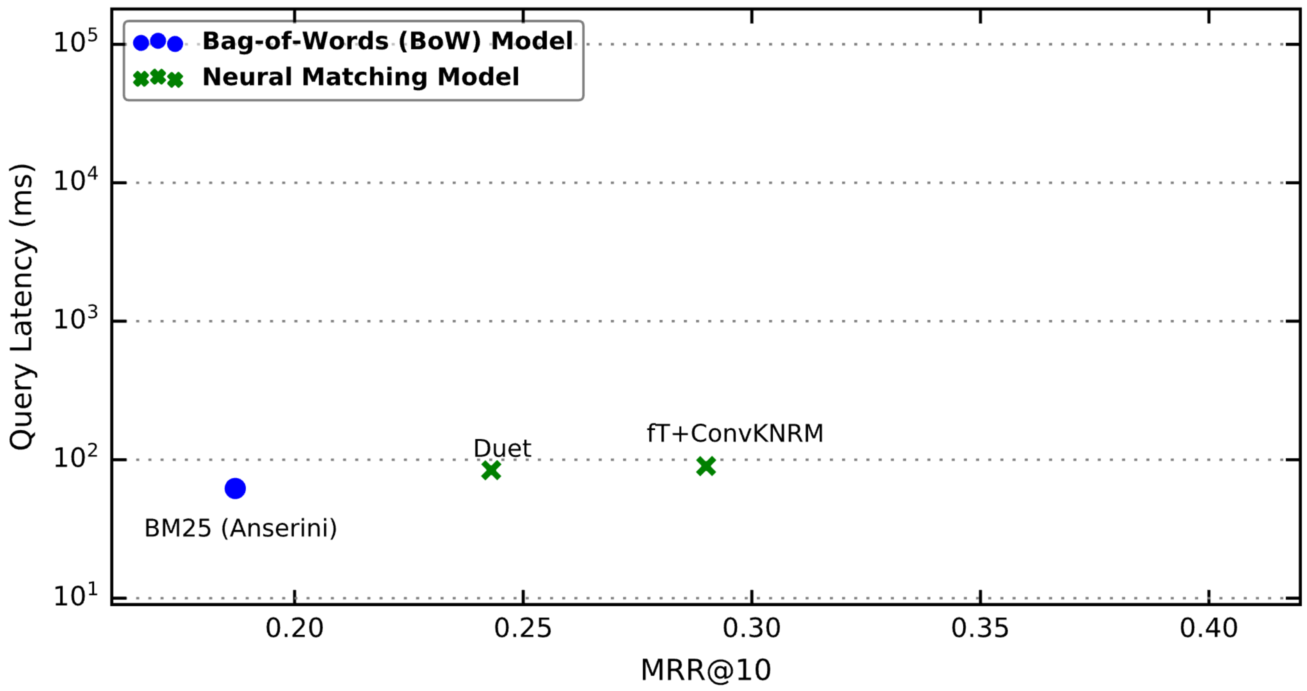
1
In the stomach, gastric acid and proteases serve as powerful chemical defenses against ingested pathogens.

Product Search | Question Answering
Fact Checking | Informative Dialogue

Original version of question and answer from SQuAD. All passages are short extractions from Wikipedia. Edited.
Pranav Rajpurkar, et al. 2016. SQuAD: 100,000+ questions for machine comprehension of text. EMNLP'16.

Retrievers must balance **quality** & **efficiency**

But Search Needs to be Fast

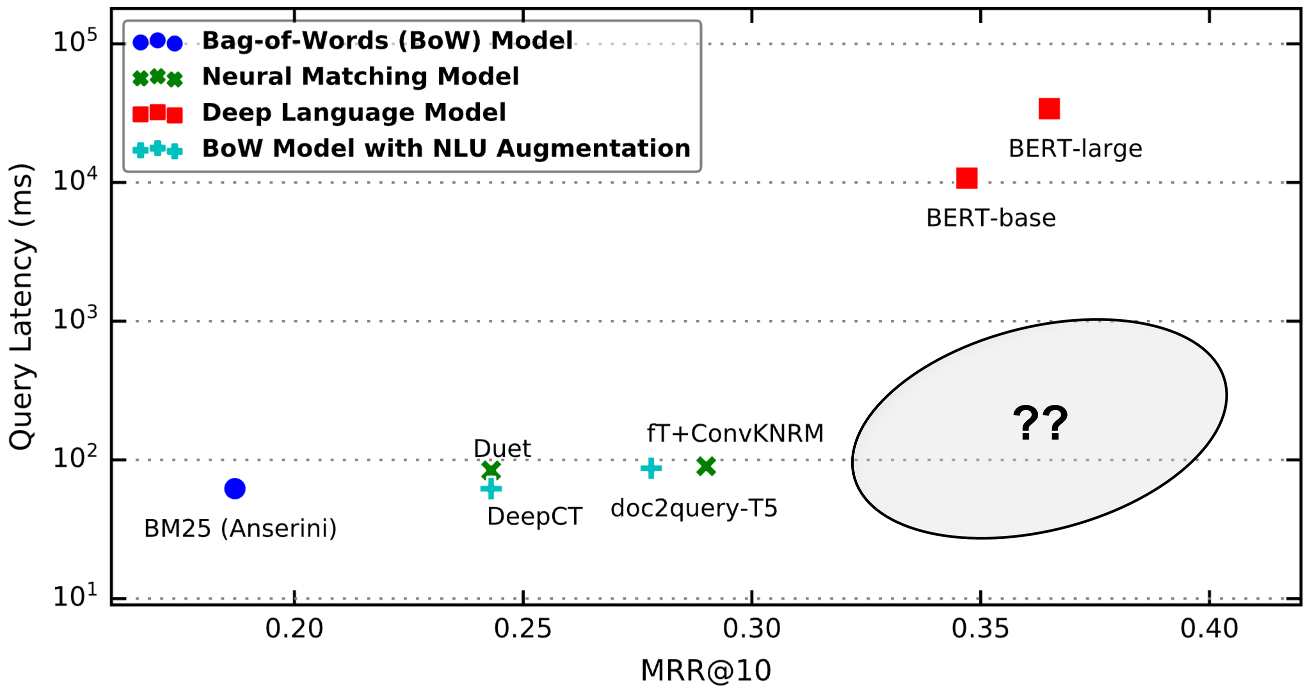


Higher Quality is Better

Answer **challenging queries** vs. Search over **millions** of documents in **milliseconds**!

Retrievers must balance **quality** & **efficiency**

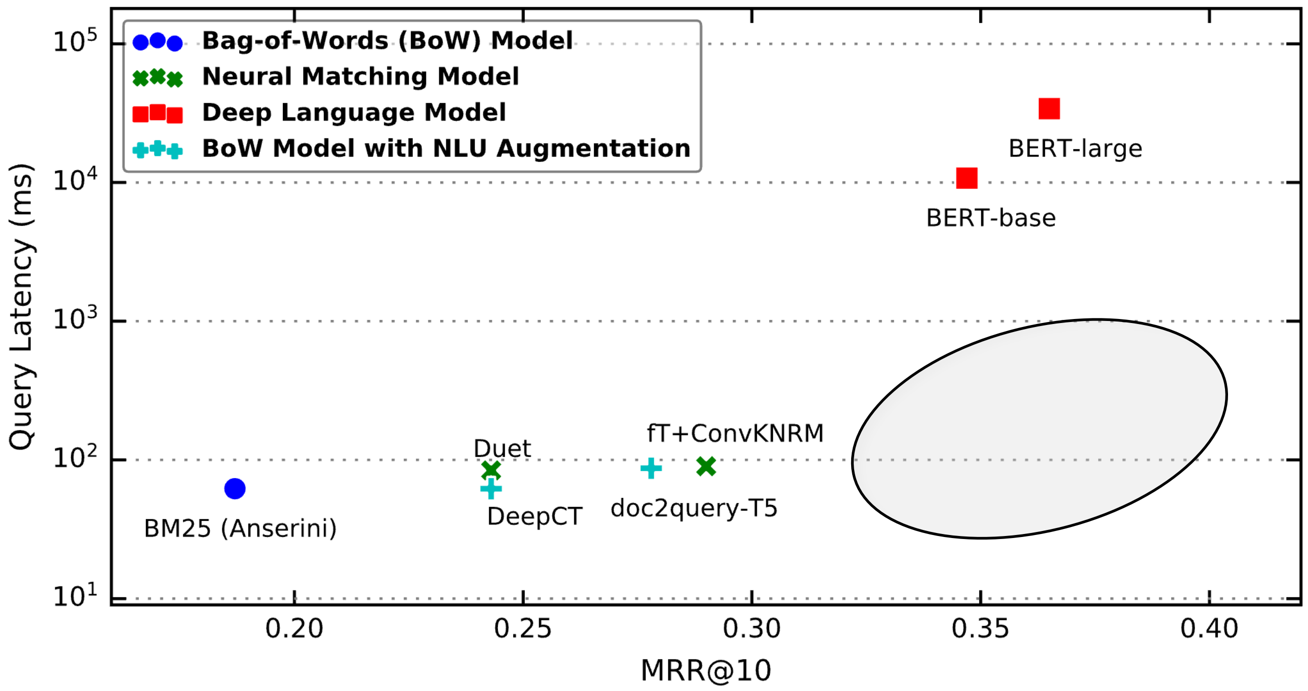
But Search Needs to be Fast



Higher Quality is Better

Retrievers must balance **quality** & **efficiency**

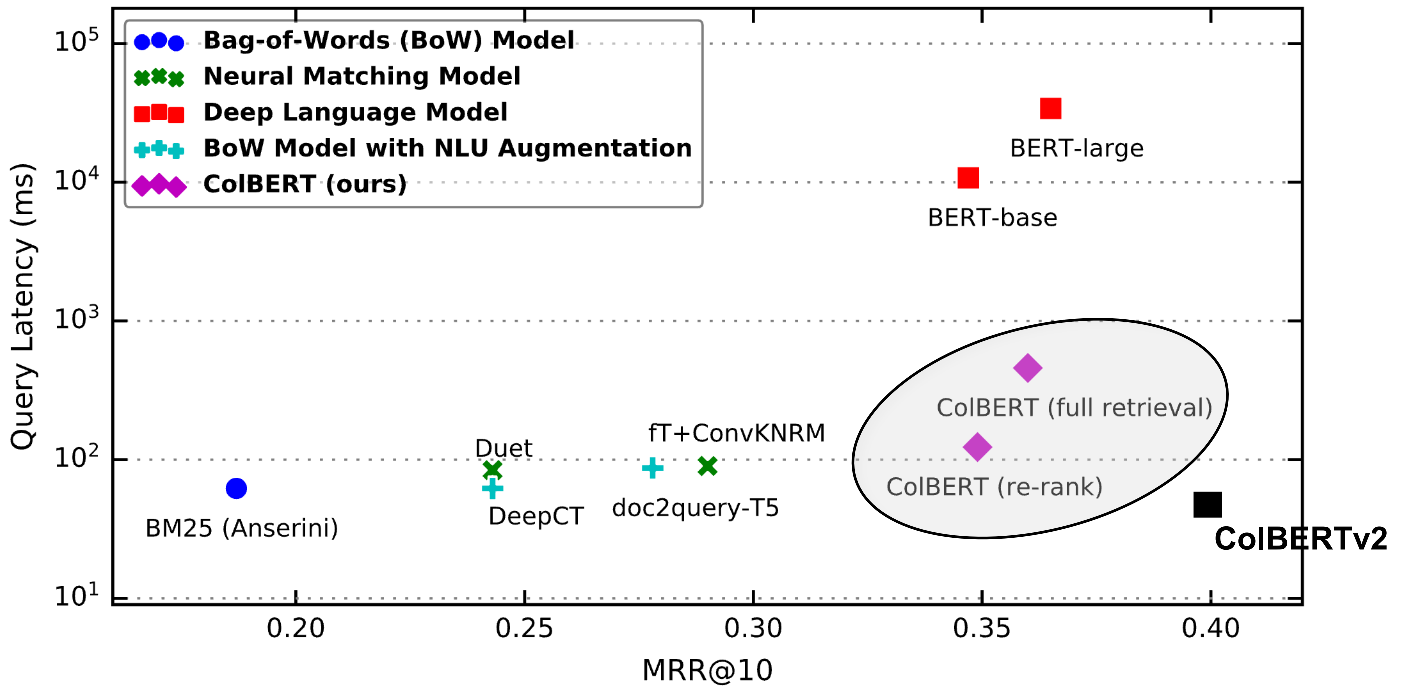
But Search Needs to be Fast



Higher Quality is Better

Retrievers must balance **quality** & **efficiency**

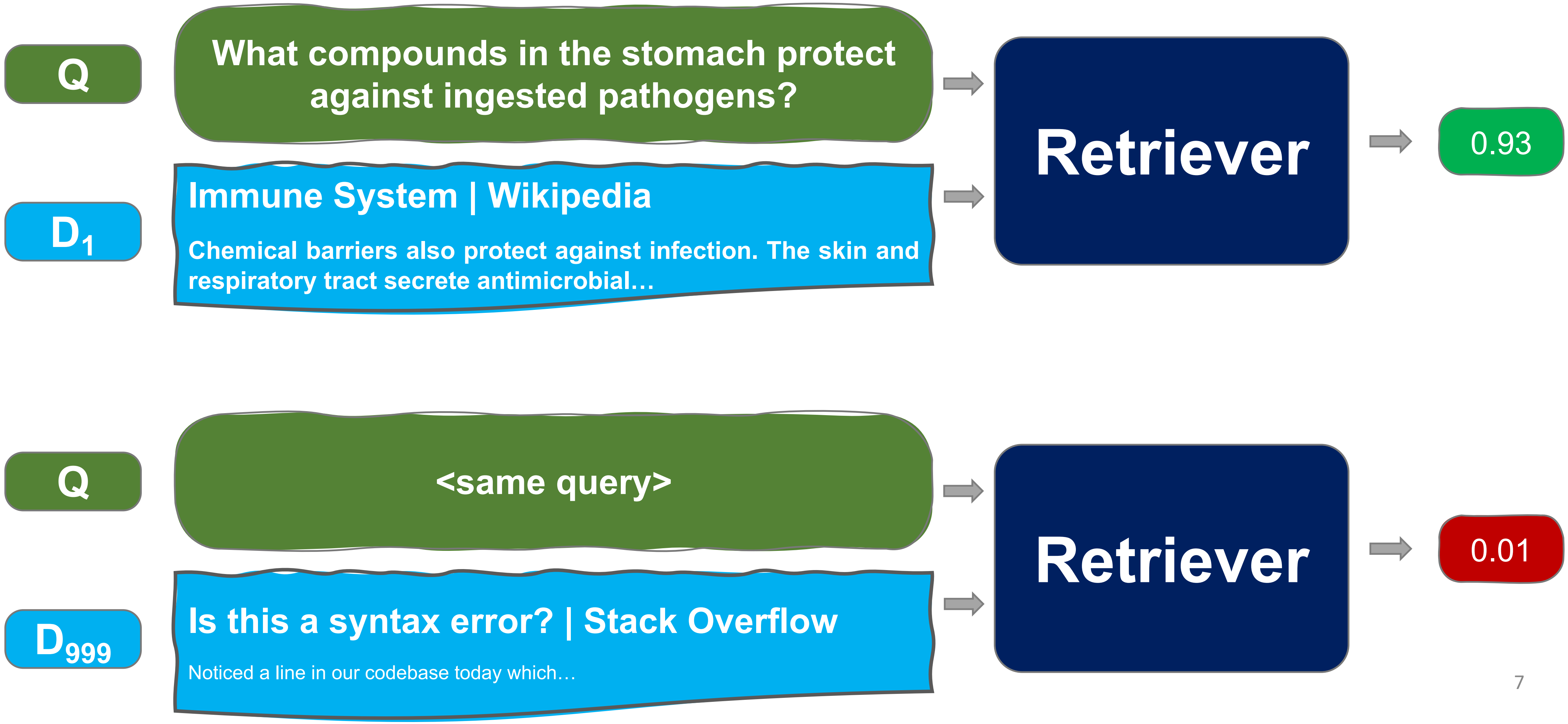
But Search Needs to be Fast



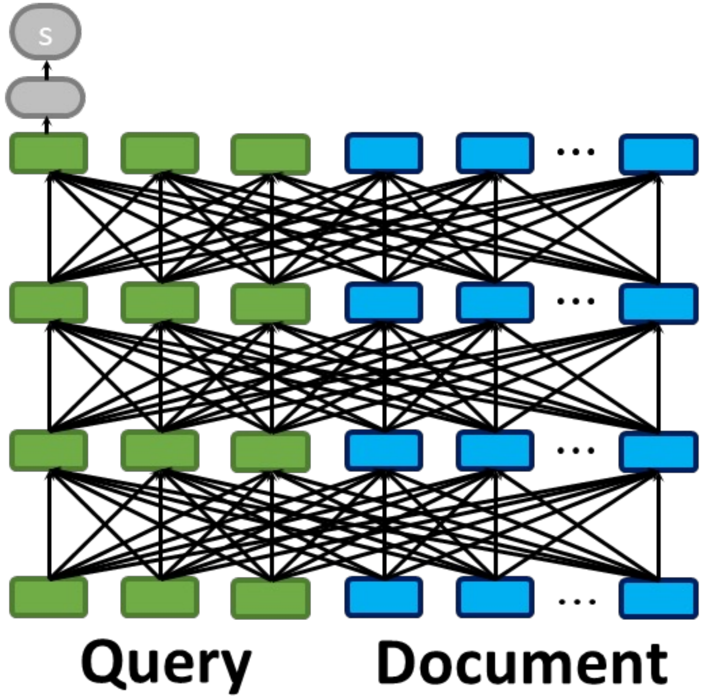
Higher Quality is Better

Latency (y axis) is in **log scale**. **ColBERT** can be **orders of magnitude** faster than BERT!

How retrievers work at a high level



Neural IR: Two Extreme Matching Paradigms



(a) Cross Encoders

✓ **Fine-Grained Interactions**

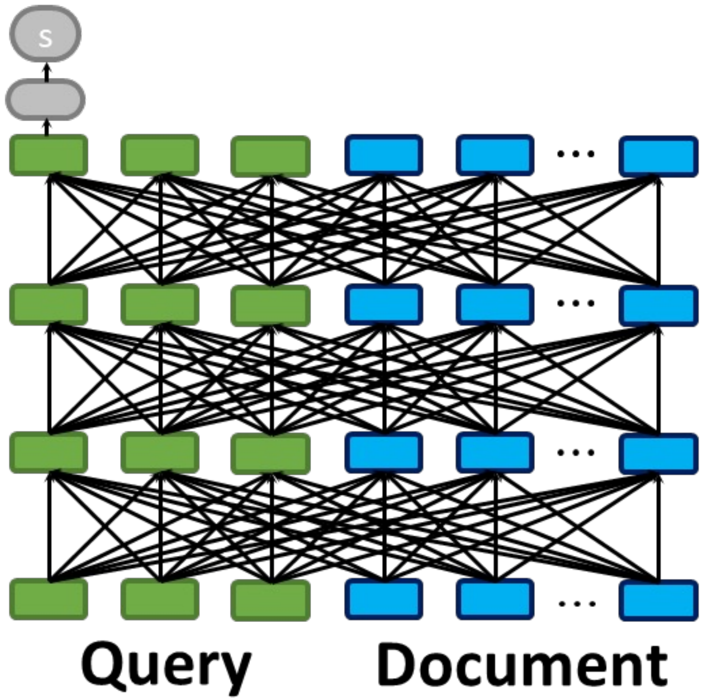
✗ **Unscalable <u>Joint</u> Conditioning**

Scale is a major challenge.

You might have **100 million** documents.

Even if scoring **each document** took **10 ms**, retrieval would consume **11 days** per query!

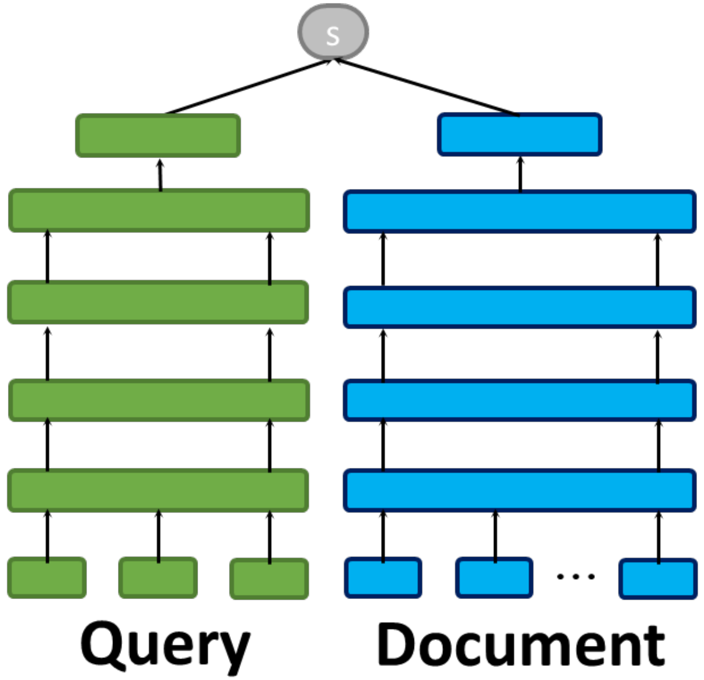
Neural IR: Two Extreme Matching Paradigms



(a) Cross Encoders

✓ **Fine-Grained Interactions**

✗ **Unscalable Joint Conditioning**

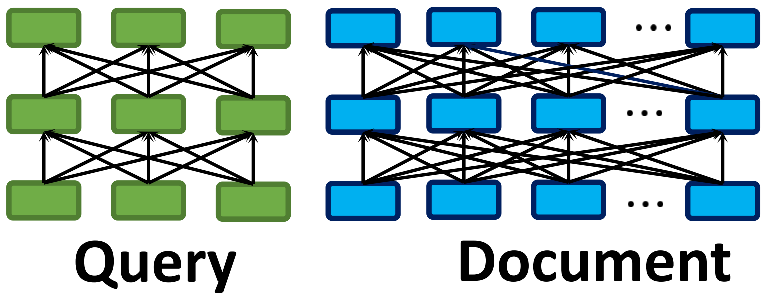


(b) Single-Vector Representations

✓ **Independent, Dense Encoding**

✗ **Coarse-Grained Representation**

ColBERT: Late Interaction

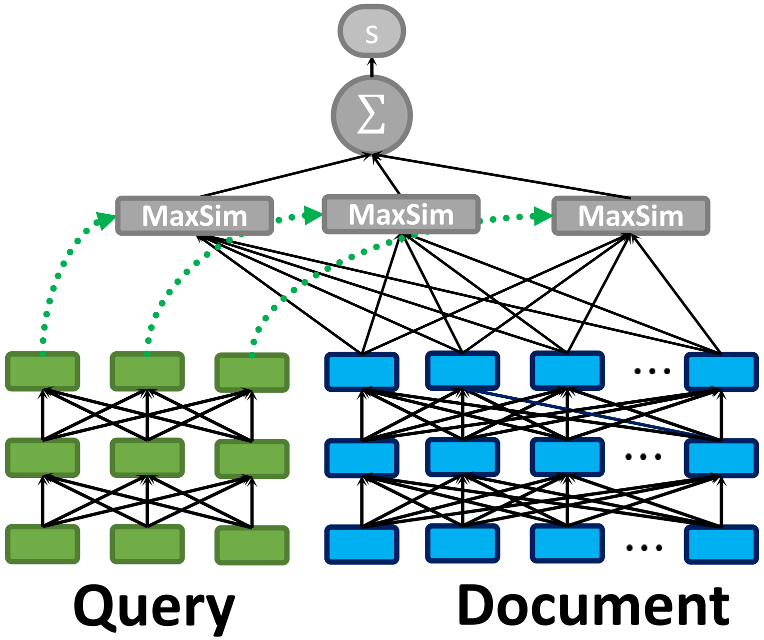


(c) Late Interaction

- ✓ Independent Encoding
- ✓ Fine-Grained Representations
- ✓ Scalable Nearest-Neighbor Search

Omar Khattab and Matei Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR 2020.

ColBERT: Late Interaction



(c) Late Interaction

- ✔ Independent Encoding
- ✔ Fine-Grained Representations
- ✔ Scalable Nearest-Neighbor Search

End-to-End Retrieval over Wikipedia (21M passages) takes **70ms**.

Omar Khattab and Matei Zaharia. "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT." SIGIR 2020.

Late Interaction: Real Example of Matching

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was released […] on August 8, 1986

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was released […] on August 8, 1986

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was released […] on August 8, 1986

when did the transformers cartoon series come out?

[…] the animated […] The Transformers […] […] It was released […] on August 8, 1986

So, how can ColBERT do interaction at scale?

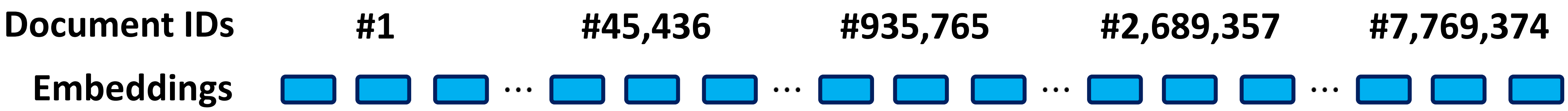
Key Idea:
Retrieval just needs the top-K results.
Only score documents that are actually promising.

Even if we have **100 million** documents, let's score only the most promising **10 thousand**.

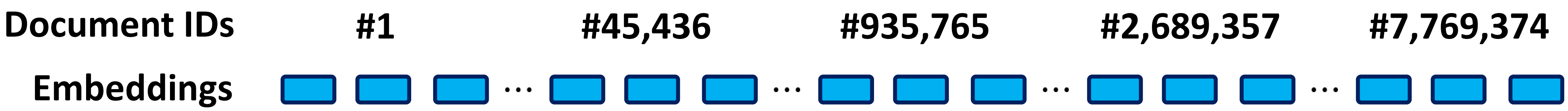
ColBERT: End-to-End Retrieval

Document IDs	**#1**	**#45,436**	**#935,765**	**#2,689,357**	**#7,769,374**

ColBERT: End-to-End Retrieval

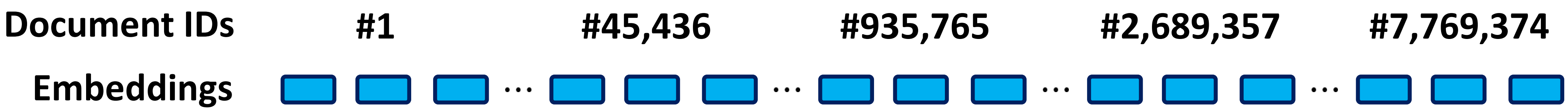


ColBERT: End-to-End Retrieval

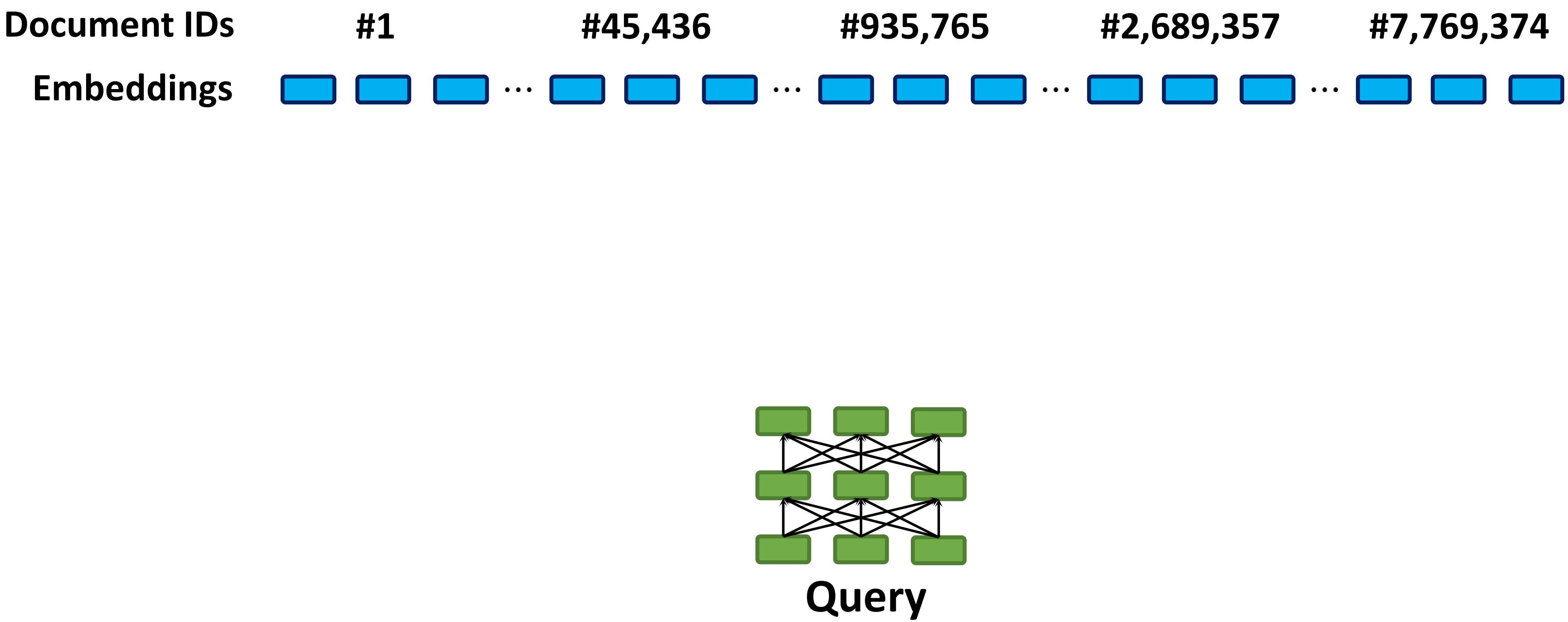


Indexed for fast vector-similarity search.
We use Facebook's faiss.

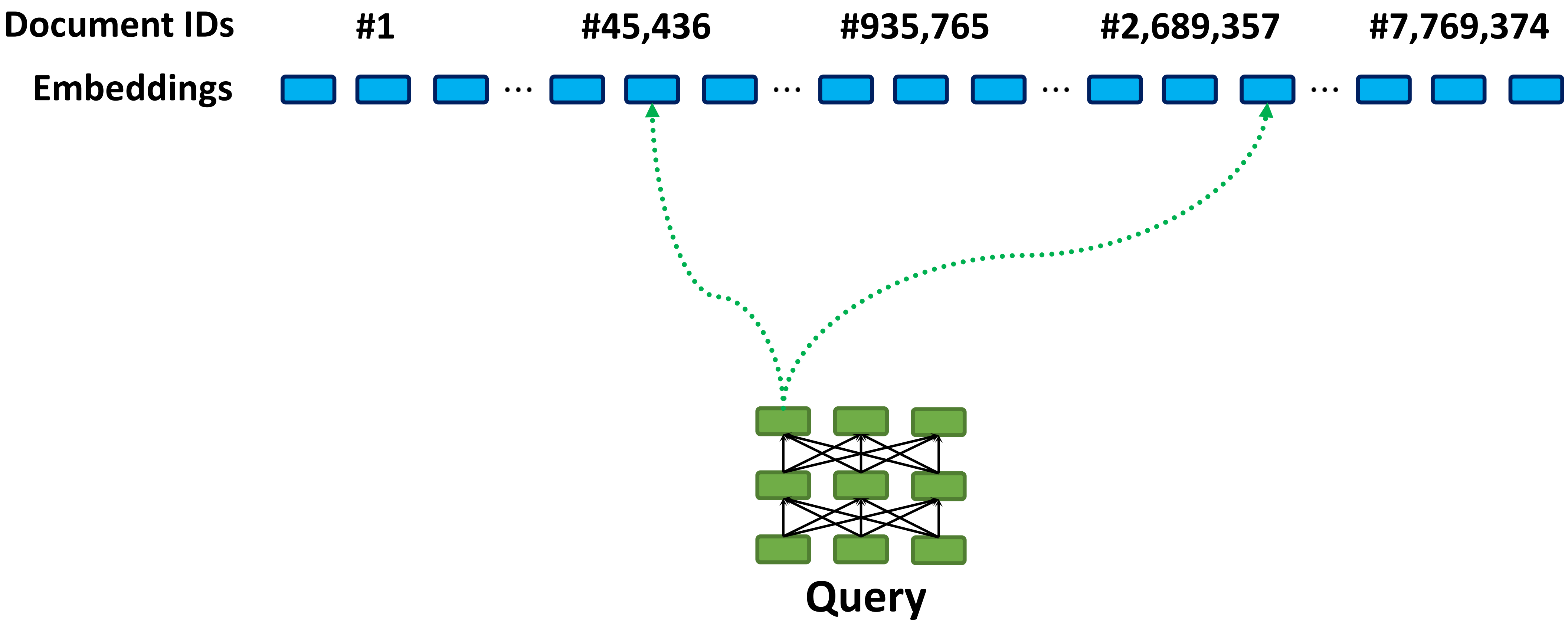
ColBERT: End-to-End Retrieval



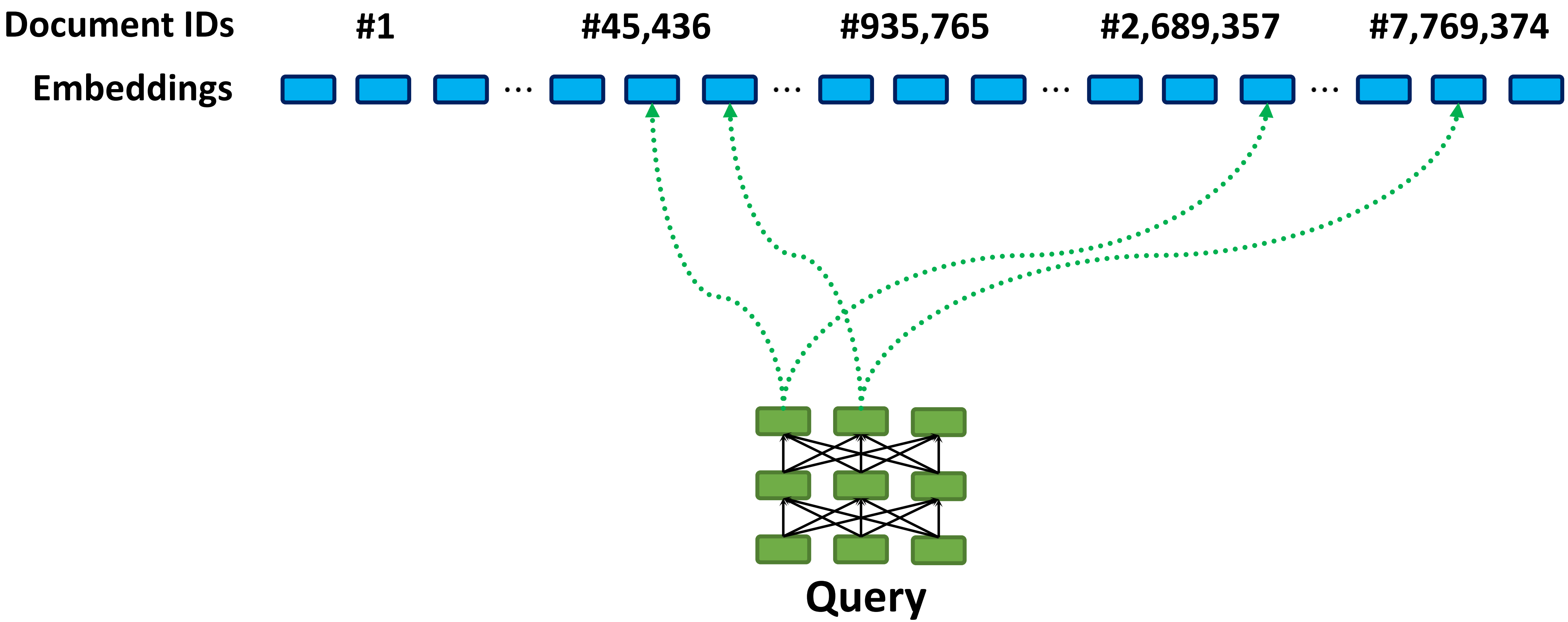
ColBERT: End-to-End Retrieval



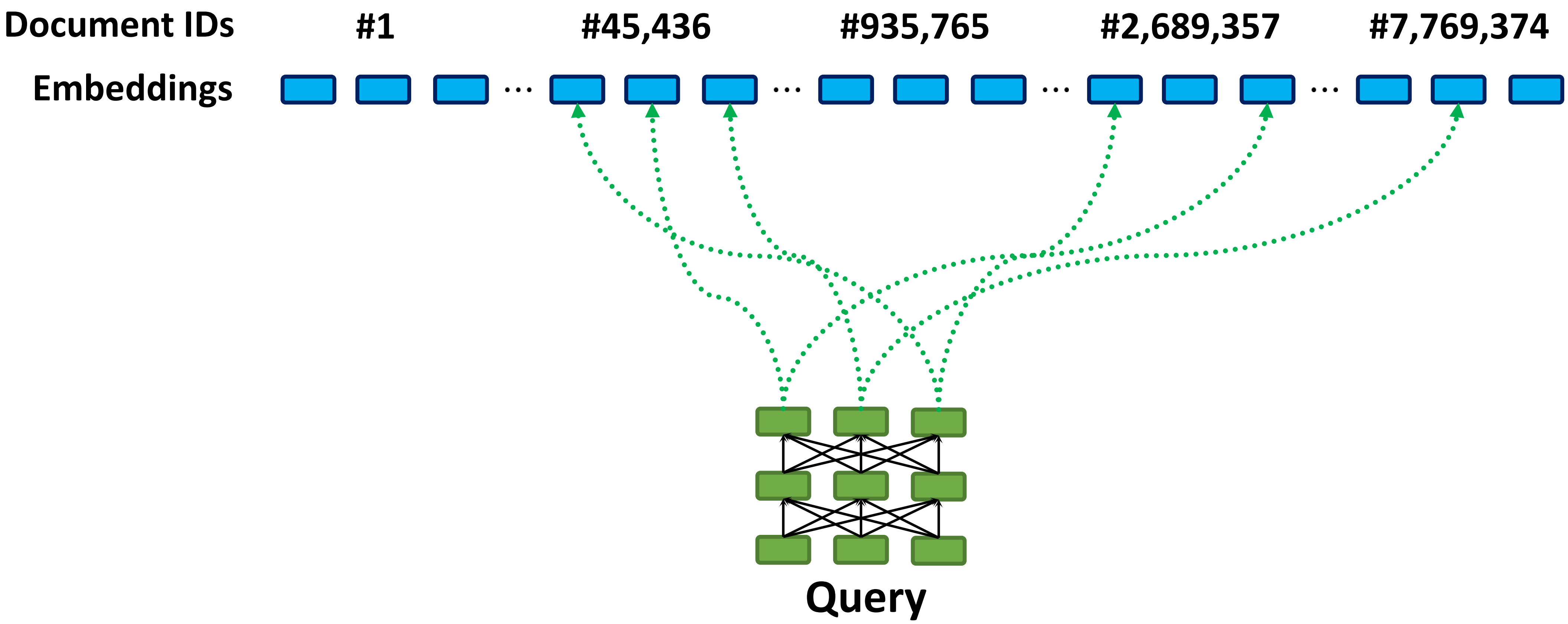
ColBERT: End-to-End Retrieval



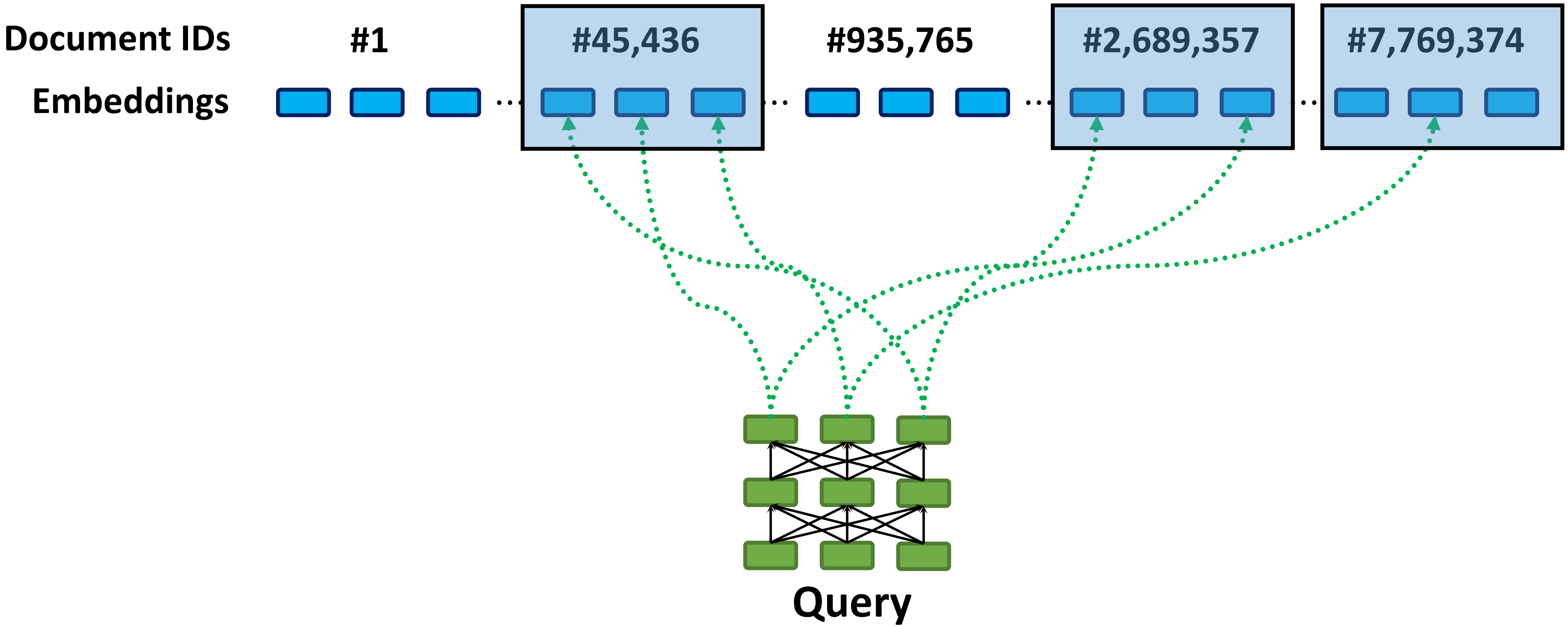
ColBERT: End-to-End Retrieval



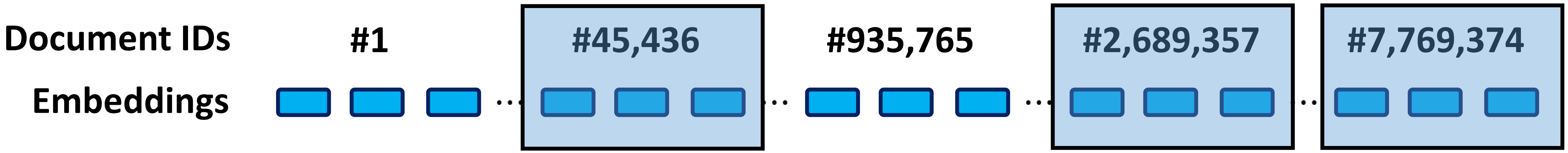
ColBERT: End-to-End Retrieval



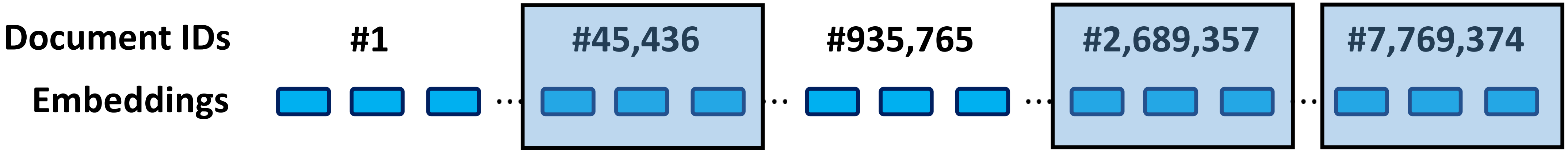
ColBERT: End-to-End Retrieval



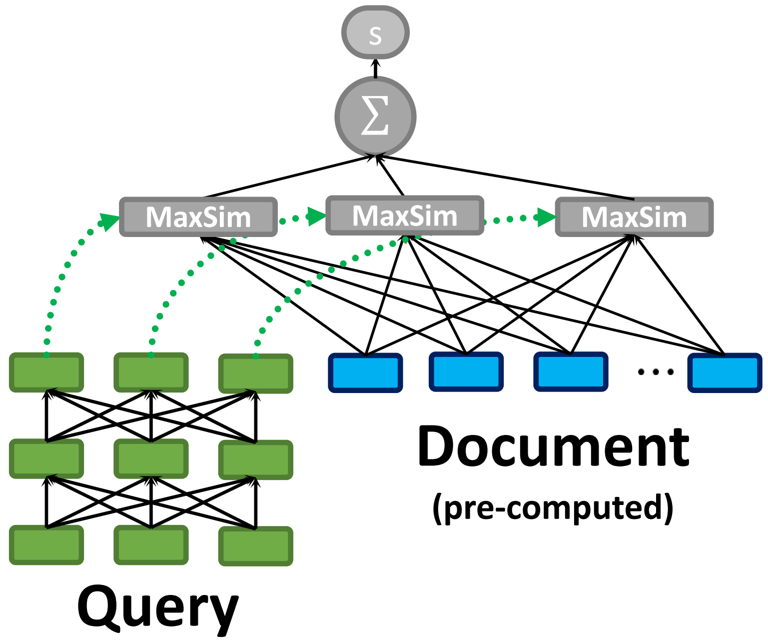
ColBERT: End-to-End Retrieval

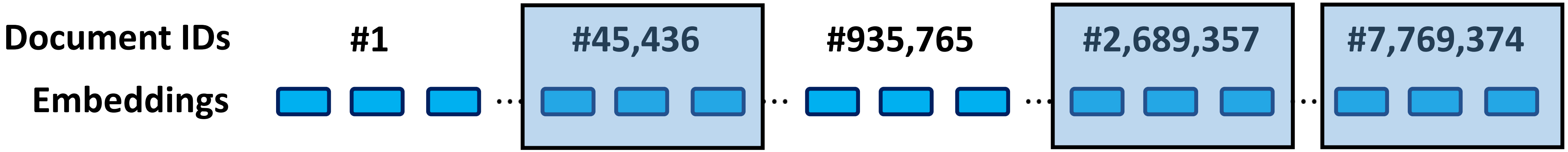


ColBERT: End-to-End Retrieval

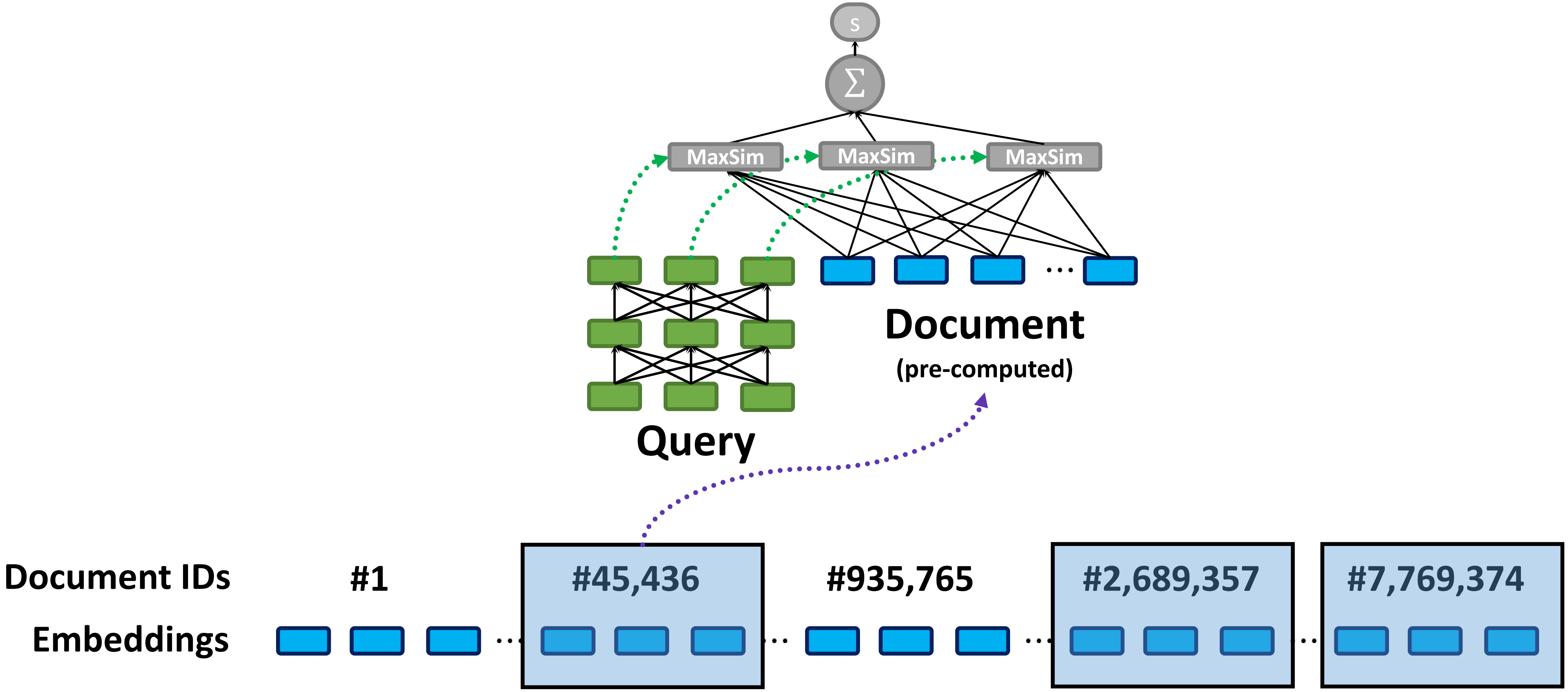


ColBERT: End-to-End Retrieval

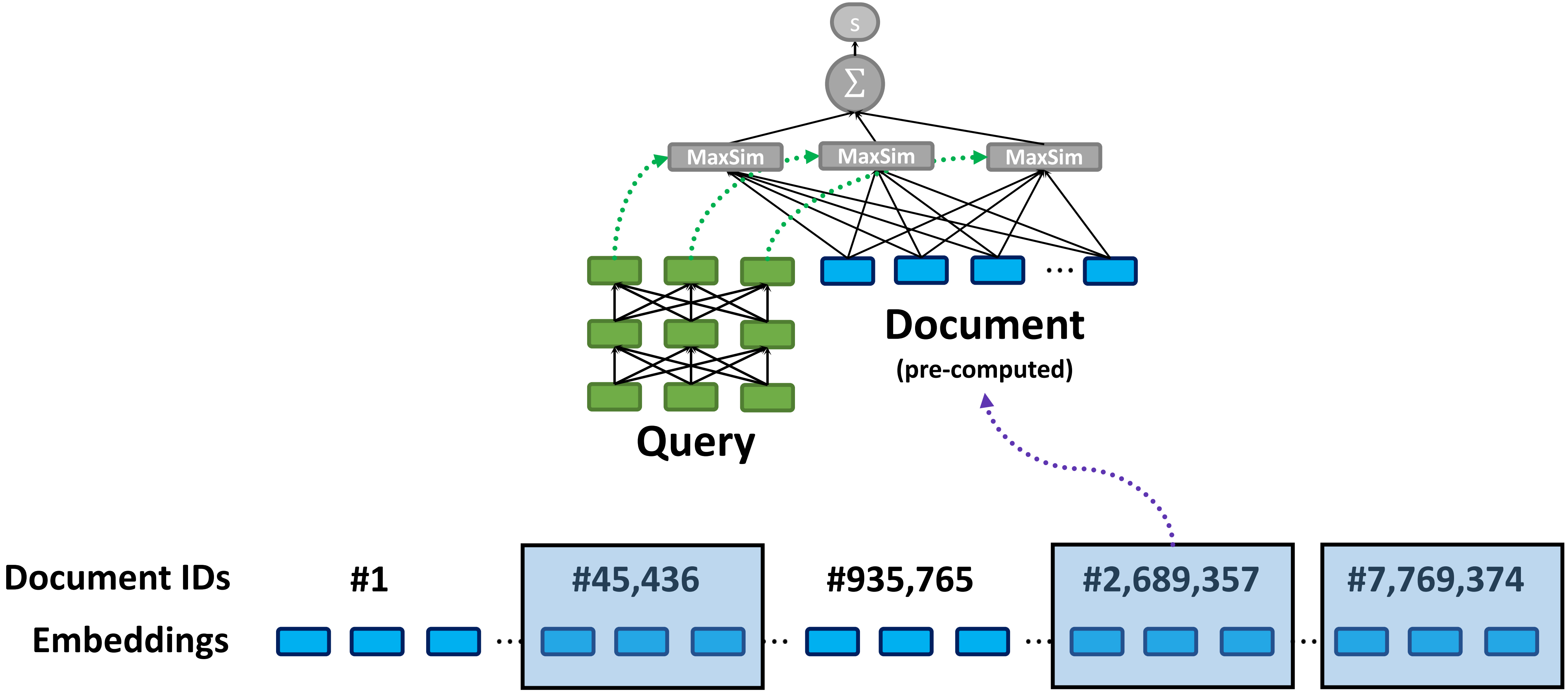




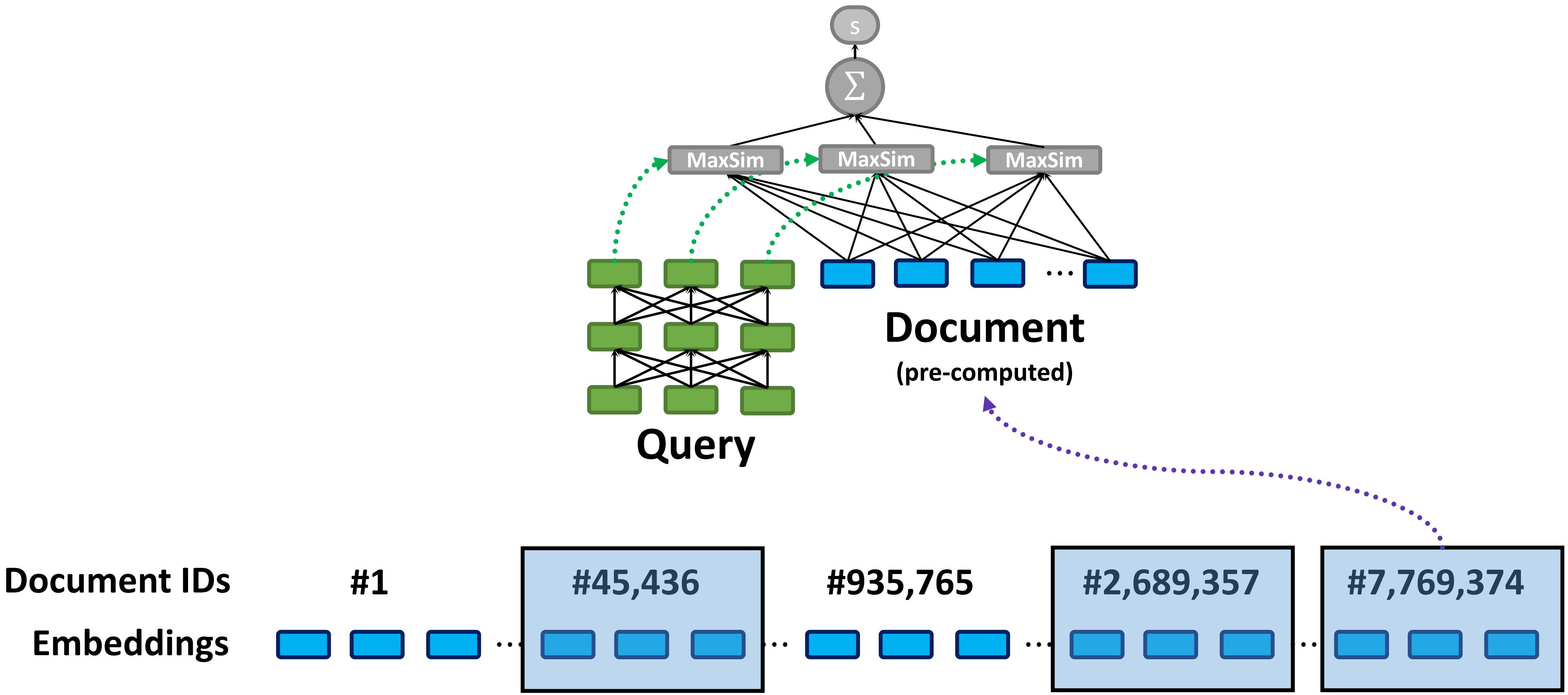
ColBERT: End-to-End Retrieval



ColBERT: End-to-End Retrieval



ColBERT: End-to-End Retrieval



Late interaction delivers large gains…

Late interaction delivers large gains…

	Passage Ranking
Model	**MS MARCO** MRR@10
BM25	18.7
DPR	31.1
ANCE	33.0
ColBERT[QA]	**36.0 / 37.5**

DPR and ANCE result figures from the respective papers. DPR on MS MARCO was evaluated by the ANCE authors.

Late interaction delivers large gains…

	Passage Ranking	Open-Domain QA Retrieval over Wikipedia'18		
Model	**MS MARCO** MRR@10	**NaturalQs** Success@20	**TriviaQA** Success@20	**Open-SQuAD** Success@20
BM25	18.7	64.0	77.3	71.4
DPR	31.1	79.4	79.9	71.5
ANCE	33.0	81.9	80.3	-
ColBERT[QA]	**36.0 / 37.5**	**85.3**	**85.6**	**83.7**

DPR and ANCE result figures from the respective papers. DPR on MS MARCO was evaluated by the ANCE authors.

Late interaction delivers large gains…

	Passage Ranking	Open-Domain QA Retrieval over Wikipedia'18		
Model	**MS MARCO** MRR@10	**NaturalQs** Success@20	**TriviaQA** Success@20	**Open-SQuAD** Success@20
BM25	18.7			

BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models
Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych

Evaluating Extrapolation Performance of Dense Retrieval
Jingtao Zhan[1], Xiaohui Xie[1], Jiaxin Mao[2], Yiqun Liu[1*], Min Zhang[1], Shaoping Ma[1]

Toward A Fine-Grained Analysis of Distribution Shifts in MSMARCO
Simon Lupart, Stéphane Clinchant

RELiC: Retrieving Evidence for Literary Claims
Katherine Thai, Yapei Chang, Kalpesh Krishna, Mohit Iyyer

And the gaps are often larger when there's a **domain shift** or a **challenging downstream task**!

Relevance-guided Supervision for OpenQA with ColBERT
Omar Khattab, Christopher Potts, Matei Zaharia

Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval
Omar Khattab, Christopher Potts, Matei Zaharia

Evaluating Token-Level and Passage-Level Dense Retrieval Models for Math Information Retrieval
Wei Zhong, Jheng-Hong Yang, and Jimmy Lin

Learning Cross-Lingual IR from an English Retriever
Yulong Li[†*], Martin Franz[‡*], Md Arafat Sultan[‡*], Bhavani Iyer[‡], Young-Suk Lee[‡] and Avirup Sil[‡]

But in the first version of ColBERT, this came at a cost!

	Passage Ra			
Model	**MS MAR** MRR@10	Success@20	Success@20	**D** Success@20
BM25	18.7	64.0	77.3	71.4
DPR	31.1	79.4	79.9	71.5
ANCE				-
ColBERT[QA]				**33.7**

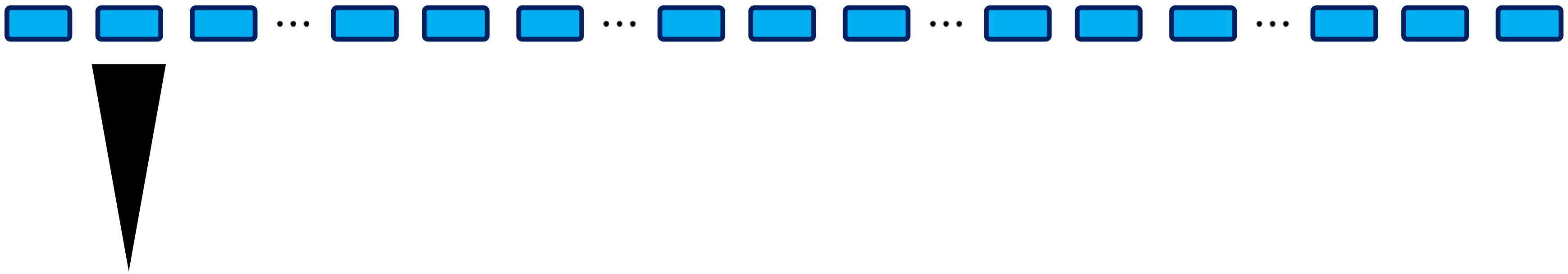
However, ColBERT's index is an order of magnitude larger than baselines, at **650 GB** for Wikipedia!

Can we advance ColBERT's large quality advantage and **reduce its footprint by an order of magnitude**?

ColBERTv2: **Can we reduce the storage requirements?**

Document IDs **#1** **#45,436** **#935,765** **#2,689,357** **#7,769,374**

Embeddings



0.35, 0.9, 0.03, …, 0.64, 0.14, 0.23, …, 0.78

In **ColBERTv1**, each vector needs **128 x float** (4 bytes) **= 512 bytes**

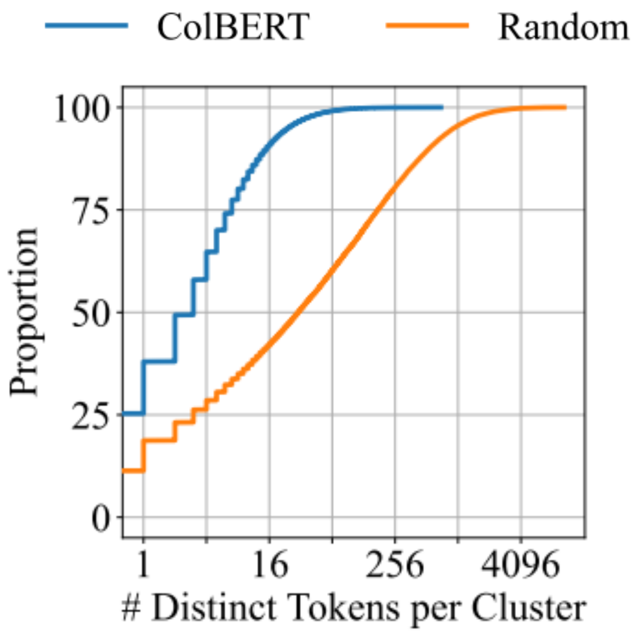
compressed vector

In **ColBERTv2**, each vector consumes just **20 bytes** – How?

ColBERTv2: **Residual** Compression

ColBERTv2: **Residual** Compression

Cluster ID	Most Common Tokens
917	‘photos’, ‘photo’, ‘pictures’, ‘photographs’, ‘images’, ‘photography’, ‘photograph’
216932	‘tornado’, ‘tornadoes’, ‘storm’ ‘hurricane’, ‘storms’



Vectors corresponding to each sense of a word ***cluster closely****, with minor but important* ***variation due to context****!*

ColBERTv2: **Residual** Compression

ColBERTv2: **Residual** Compression

Document IDs	**#1**	**#45,436**	**#935,765**	**#2,689,357**	**#7,769,374**

Embeddings ▭ ▭ ▭ … ▭ ▭ ▭ … ▭ ▭ ▭ … ▭ ▭ ▭ … ▭ ▭ ▭

0.35, 0.9, 0.03, …, 0.64, 0.14, 0.23, …, 0.78

In **ColBERTv1**, each vector needs
128 x float (4 bytes) **= 512 bytes**

compressed vector

In **ColBERTv2**, each vector encodes
cluster ID (4 bytes)
\+ **128 x bit** (16 bytes)
= **20 bytes** only

ColBERTv2: **Residual** Compression

Indexing clusters a sample of token vectors.

Represent each vector as a **cluster ID** and **a 1-bit delta per dimension**.
This can consume as little **20 bytes** per vector.

	MS MARCO Passage Ranking		
Model	Storage	MRR@10	Recall@50
ColBERT **v1**	154 GB	**36.2**	**82.1**
+ 2 bit residual compression (6x)	25 GB	**36.2**	**82.3**
+ 1 bit residual compression (10x)	16 GB	35.5	81.6

ColBERTv2 uses **denoised training** and **residual compression** to re-emerge more effective & lightweight

Method	Official Dev (7k)			Local Eval (5k)		
	MRR@10	R@50	R@1k	MRR@10	R@50	R@1k
Models without Distillation or Special Pretraining						
RepBERT	30.4	-	94.3	-	-	-
DPR	31.1	-	95.2	-	-	-
ANCE	33.0	-	95.9	-	-	-
LTRe	34.1	-	96.2	-	-	-
ColBERT	36.0	82.9	96.8	36.7	-	-
Models with Distillation or Special Pretraining						
TAS-B	34.7	-	97.8	-	-	-
SPLADEv2	36.8	-	97.9	37.9	84.9	98.0
PAIR	37.9	86.4	98.2	-	-	-
coCondenser	38.2	-	**98.4**	-	-	-
RocketQAv2	38.8	86.2	98.1	39.8	85.8	97.9
ColBERTv2	**39.7**	**86.8**	**98.4**	**40.8**	**86.3**	**98.3**

Corpus	Models without Distillation				Models with Distillation			
	ColBERT	DPR-M	ANCE	MoDIR	TAS-B	RocketQAv2	SPLADEv2	ColBERTv2
BEIR Search Tasks (nDCG@10)								
DBPedia	39.2	23.6	28.1	28.4	38.4	35.6	43.5	**44.6**
FiQA	31.7	27.5	29.5	29.6	30.0	30.2	33.6	**35.6**
NQ	52.4	39.8	44.6	44.2	46.3	50.5	52.1	**56.2**
HotpotQA	59.3	37.1	45.6	46.2	58.4	53.3	**68.4**	66.7
NFCorpus	30.5	20.8	23.7	24.4	31.9	29.3	33.4	**33.8**
T-COVID	67.7	56.1	65.4	67.6	48.1	67.5	71.0	**73.8**
Touché (v2)	-	-	-	-	-	24.7	**27.2**	26.3
BEIR Semantic Relatedness Tasks (nDCG@10)								
ArguAna	23.3	41.4	41.5	41.8	42.7	45.1	**47.9**	46.3
C-FEVER	18.4	17.6	19.8	20.6	22.8	18.0	**23.5**	17.6
FEVER	77.1	58.9	66.9	68.0	70.0	67.6	**78.6**	**78.5**
Quora	85.4	84.2	85.2	**85.6**	83.5	74.9	83.8	85.2
SCIDOCS	14.5	10.8	12.2	12.4	14.9	13.1	**15.8**	15.4
SciFact	67.1	47.8	50.7	50.2	64.3	56.8	**69.3**	**69.3**

Corpus	ColBERT	BM25	ANCE	RocketQAv2	SPLADEv2	ColBERTv2
OOD Wikipedia Open QA (Success@5)						
NQ-dev	65.7	44.6	-	-	65.6	**68.9**
TQ-dev	72.6	67.6	-	-	74.7	**76.7**
SQuAD-dev	60.0	50.6	-	-	60.4	**65.0**
LoTTE Search Test Queries (Success@5)						
Writing	74.7	60.3	74.4	78.0	77.1	**80.1**
Recreation	68.5	56.5	64.7	72.1	69.0	**72.3**
Science	53.6	32.7	53.6	55.3	55.4	**56.7**
Technology	61.9	41.8	59.6	63.4	62.4	**66.1**
Lifestyle	80.2	63.8	82.3	82.1	82.3	**84.7**
Pooled	67.3	48.3	66.4	69.8	68.9	**71.6**
LoTTE Forum Test Queries (Success@5)						
Writing	71.0	64.0	68.8	71.5	73.0	**76.3**
Recreation	65.6	55.4	63.8	65.7	67.1	**70.8**
Science	41.8	37.1	36.5	38.0	43.7	**46.1**
Technology	48.5	39.4	46.8	47.3	50.8	**53.6**
Lifestyle	73.0	60.6	73.1	73.7	74.0	**76.9**
Pooled	58.2	47.2	55.7	57.7	60.1	**63.4**

ColBERTv2 uses **denoised training** and **residual compression** to re-emerge more effective & lightweight

… while **reducing** the index size **6—10x**, encoding Wikipedia in 65—110 GB.

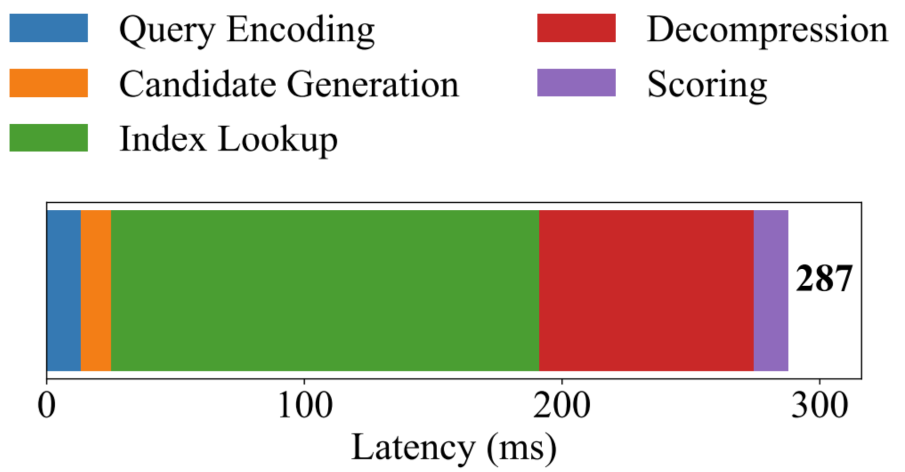
Models without Distillation or Special Pretraining						
RepBERT	30.4	-	94.3	-	-	-
DPR	31.1	-	95.2	-	-	-
ANCE	33.0	-	95.9	-	-	-
LTRe	34.1	-	96.2	-	-	-
ColBERT	36.0	82.9	96.8	36.7	-	-
Models with Distillation or Special Pretraining						
TAS-B	34.7	-	97.8	-	-	-
SPLADEv2	36.8	-	97.9	37.9	84.9	98.0
PAIR	37.9	86.4	98.2	-	-	-
coCondenser	38.2	-	**98.4**	-	-	-
RocketQAv2	38.8	86.2	98.1	39.8	85.8	97.9
ColBERTv2	**39.7**	**86.8**	**98.4**	**40.8**	**86.3**	**98.3**

	odels without Distillation				Models with Distillation			
	ColBERT	DPR-M	ANCE	MoDIR	TAS-B	RocketQAv2	SPLADEv2	ColBERTv2
BEIR Search Tasks (nDCG@10)								
DBPedia	39.2	23.6	28.1	28.4	38.4	35.6	43.5	**44.6**
FiQA	31.7	27.5	29.5	29.6	30.0	30.2	33.6	**35.6**
NQ	52.4	39.8	44.6	44.2	46.3	50.5	52.1	**56.2**
HotpotQA	59.3	37.1	45.6	46.2	58.4	53.3	**68.4**	66.7
NFCorpus	30.5	20.8	23.7	24.4	31.9	29.3	33.4	**33.8**
T-COVID	67.7	56.1	65.4	67.6	48.1	67.5	71.0	**73.8**
Touché (v2)	-	-	-	-	-	24.7	**27.2**	26.3
BEIR Semantic Relatedness Tasks (nDCG@10)								
ArguA								
C-FEV								
FEVE								
Quora								
SCID								
SciFa								

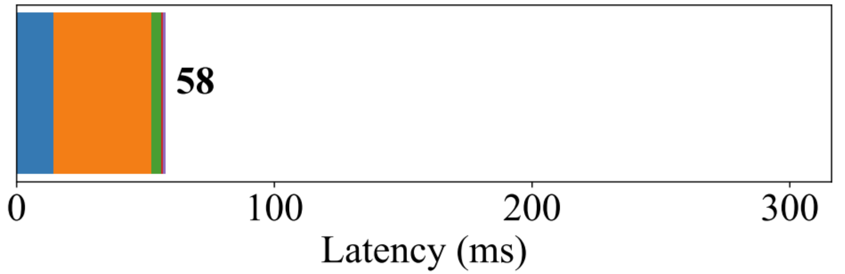
Corpus	ColBERT	BM25	ANCE	RocketQAv2	SPLADEv2	ColBERTv2
OOD Wikipedia Open QA (Success@5)						
NQ-dev	65.7	44.6	-	-	65.6	**68.9**
TQ-dev	72.6	67.6	-	-	74.7	**76.7**
SQuAD-dev	60.0	50.6	-	-	60.4	**65.0**
LoTTE Search Test Queries (Success@5)						
Writing	74.7	60.3	74.4	78.0	77.1	**80.1**
Recreation	68.5	56.5	64.7	72.1	69.0	**72.3**
Science	53.6	32.7	53.6	55.3	55.4	**56.7**
Technology	61.9	41.8	59.6	63.4	62.4	**66.1**
Lifestyle	80.2	63.8	82.3	82.1	82.3	**84.7**
Pooled	67.3	48.3	66.4	69.8	68.9	**71.6**
LoTTE Forum Test Queries (Success@5)						
						76.3
						70.8
						46.1
						53.6
						76.9
						63.4

… and supporting efficient search, with only 10s—100s of ms of latency per query

What about latency and hardware requirements?



(a) Vanilla ColBERTv2 (nprobe=4, ncandidates=2^{16}).

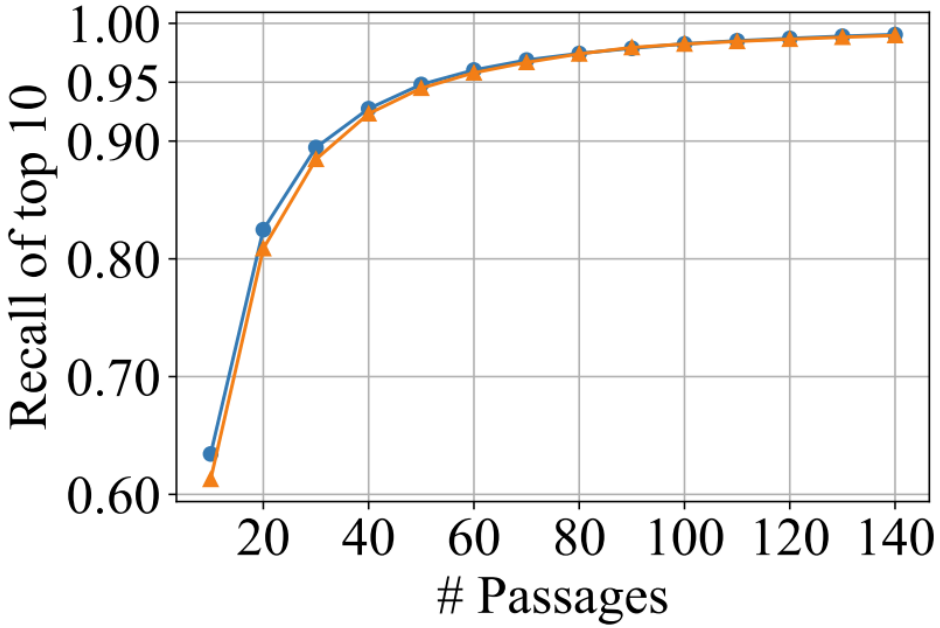


(b) PLAID ColBERTv2 (k = 1000)

Figure 2: Latency breakdown of MS MARCO v1 dev queries run with vanilla ColBERTv2 and PLAID ColBERTv2 on a TITAN V GPU. Vanilla ColBERTv2 is overwhelmingly bottlenecked with the cost of index lookup and decompression, a challenge that PLAID addresses.

Faster ColBERTv2 with PLAID: **Centroid Interaction** Search

- Centroids alone identify the candidate you need to score!



(a) $k = 10$

With **PLAID**, **ColBERTv2** scales its state-of-the-art quality to massive datasets!

Model	**MS MARCO "v2"**
# of tokens	9B
# of passages	140M
Index Size	200 GB (1-bit)
CPU Search Latency	136 ms

ColBERTv2 is available at colbert.ai

Method	Official Dev (7k) MRR@10	Official Dev (7k) R@50	Official Dev (7k) R@1k	Local Eval (5k) MRR@10	Local Eval (5k) R@50	Local Eval (5k) R@1k
Models without Distillation or Special Pretraining						
RepBERT	30.4	-	94.3	-	-	-
DPR	31.1	-	95.2	-	-	-
ANCE	33.0	-	95.9	-	-	-
LTRe	34.1	-	96.2	-	-	-
ColBERT	36.0	82.9	96.8	36.7	-	-
Models with Distillation or Special Pretraining						
TAS-B	34.7	-	97.8	-	-	-
SPLADEv2	36.8	-	97.9	37.9	84.9	98.0
PAIR	37.9	86.4	98.2	-	-	-
coCondenser	38.2	-	**98.4**	-	-	-
RocketQAv2	38.8	86.2	98.1	39.8	85.8	97.9
ColBERTv2	**39.7**	**86.8**	**98.4**	**40.8**	**86.3**	**98.3**

Corpus	Models without Distillation: ColBERT	DPR-M	ANCE	MoDIR	Models with Distillation: TAS-B	RocketQAv2	SPLADEv2	ColBERTv2
BEIR Search Tasks (nDCG@10)								
DBPedia	39.2	23.6	28.1	28.4	38.4	35.6	43.5	**44.6**
FiQA	31.7	27.5	29.5	29.6	30.0	30.2	33.6	**35.6**
NQ	52.4	39.8	44.6	44.2	46.3	50.5	52.1	**56.2**
HotpotQA	59.3	37.1	45.6	46.2	58.4	53.3	**68.4**	66.7
NFCorpus	30.5	20.8	23.7	24.4	31.9	29.3	33.4	**33.8**
T-COVID	67.7	56.1	65.4	67.6	48.1	67.5	71.0	**73.8**
Touché (v2)	-	-	-	-	-	24.7	**27.2**	26.3
BEIR Semantic Relatedness Tasks (nDCG@10)								
ArguAna	23.3	41.4	41.5	41.8	42.7	45.1	**47.9**	46.3
C-FEVER	18.4	17.6	19.8	20.6	22.8	18.0	**23.5**	17.6
FEVER	77.1	58.9	66.9	68.0	70.0	67.6	**78.6**	**78.5**
Quora	85.4	84.2	85.2	**85.6**	83.5	74.9	83.8	85.2
SCIDOCS	14.5	10.8	12.2	12.4	14.9	13.1	**15.8**	15.4
SciFact	67.1	47.8	50.7	50.2	64.3	56.8	**69.3**	**69.3**

Corpus	ColBERT	BM25	ANCE	RocketQAv2	SPLADEv2	ColBERTv2
OOD Wikipedia Open QA (Success@5)						
NQ-dev	65.7	44.6	-	-	65.6	**68.9**
TQ-dev	72.6	67.6	-	-	74.7	**76.7**
SQuAD-dev	60.0	50.6	-	-	60.4	**65.0**
LoTTE Search Test Queries (Success@5)						
Writing	74.7	60.3	74.4	78.0	77.1	**80.1**
Recreation	68.5	56.5	64.7	72.1	69.0	**72.3**
Science	53.6	32.7	53.6	55.3	55.4	**56.7**
Technology	61.9	41.8	59.6	63.4	62.4	**66.1**
Lifestyle	80.2	63.8	82.3	82.1	82.3	**84.7**
Pooled	67.3	48.3	66.4	69.8	68.9	**71.6**
LoTTE Forum Test Queries (Success@5)						
Writing	71.0	64.0	68.8	71.5	73.0	**76.3**
Recreation	65.6	55.4	63.8	65.7	67.1	**70.8**
Science	41.8	37.1	36.5	38.0	43.7	**46.1**
Technology	48.5	39.4	46.8	47.3	50.8	**53.6**
Lifestyle	73.0	60.6	73.1	73.7	74.0	**76.9**
Pooled	58.2	47.2	55.7	57.7	60.1	**63.4**

Establishes state-of-the-art retrieval quality while **reducing** the index size **6—10x** and maintaining **10s of ms** latency, even on CPU only

ColBERTv2 is available at colbert.ai

Method	Official Dev (7k) MRR@10	R@50	R@1k	Local Eval (5k) MRR@10	R@50	R@1k
Models without Distillation or Special Pretraining						
RepBERT	30.4	-	94.3	-	-	-
DPR	31.1	-	95.2	-	-	-
ANCE	33.0	-	95.9	-	-	-
LTRe	34.1	-	96.2	-	-	-
ColBERT	36.0	82.9	96.8	36.7	-	-
Models with Distillation or Special Pretraining						
TAS-B	34.7	-	97.8	-	-	-
SPLADEv2	36.8	-	97.9	37.9	84.9	98.0
PAIR	37.9	86.4	98.2	-	-	-
coCondenser	38.2	-	**98.4**	-	-	-
RocketQAv2	38.8	86.2	98.1	39.8	85.8	97.9
ColBERTv2	**39.7**	**86.8**	**98.4**	**40.8**	**86.3**	**98.3**

Corpus	Models without Distillation: ColBERT	DPR-M	ANCE	MoDIR	Models with Distillation: TAS-B	RocketQAv2	SPLADEv2	ColBERTv2
BEIR Search Tasks (nDCG@10)								
DBPedia	39.2	23.6	28.1	28.4	38.4	35.6	43.5	**44.6**
FiQA	31.7	27.5	29.5	29.6	30.0	30.2	33.6	**35.6**
NQ	52.4	39.8	44.6	44.2	46.3	50.5	52.1	**56.2**
HotpotQA	59.3	37.1	45.6	46.2	58.4	53.3	**68.4**	66.7
NFCorpus	30.5	20.8	23.7	24.4	31.9	29.3	33.4	**33.8**
T-COVID	67.7	56.1	65.4	67.6	48.1	67.5	71.0	**73.8**
Touché (v2)	-	-	-	-	-	24.7	**27.2**	26.3
BEIR Semantic Relatedness Tasks (nDCG@10)								
ArguAna	23.3	41.4	41.5	41.8	42.7	45.1	**47.9**	46.3
C-FEVER	18.4	17.6	19.8	20.6	22.8	18.0	**23.5**	17.6
FEVER	77.1	58.9	66.9	68.0	70.0	67.6	**78.6**	78.5
Quora	85.4	84.2	85.2	**85.6**	83.5	74.9	83.8	85.2
SCIDOCS	14.5	10.8	12.2	12.4	14.9	13.1	**15.8**	15.4
SciFact	67.1	47.8	50.7	50.2	64.3	56.8	**69.3**	**69.3**

Corpus	ColBERT	BM25	ANCE	RocketQAv2	SPLADEv2	ColBERTv2
OOD Wikipedia Open QA (Success@5)						
NQ-dev	65.7	44.6	-	-	65.6	**68.9**
TQ-dev	72.6	67.6	-	-	74.7	**76.7**
SQuAD-dev	60.0	50.6	-	-	60.4	**65.0**
LoTTE Search Test Queries (Success@5)						
Writing	74.7	60.3	74.4	78.0	77.1	**80.1**
Recreation	68.5	56.5	64.7	72.1	69.0	**72.3**
Science	53.6	32.7	53.6	55.3	55.4	**56.7**
Technology	61.9	41.8	59.6	63.4	62.4	**66.1**
Lifestyle	80.2	63.8	82.3	82.1	82.3	**84.7**
Pooled	67.3	48.3	66.4	69.8	68.9	**71.6**
LoTTE Forum Test Queries (Success@5)						
Writing	71.0	64.0	68.8	71.5	73.0	**76.3**
Recreation	65.6	55.4	63.8	65.7	67.1	**70.8**
Science	41.8	37.1	36.5	38.0	43.7	**46.1**
Technology	48.5	39.4	46.8	47.3	50.8	**53.6**
Lifestyle	73.0	60.6	73.1	73.7	74.0	**76.9**
Pooled	58.2	47.2	55.7	57.7	60.1	**63.4**

ColBERTv2 is available at colbert.ai

```
[ ]   indexer = Indexer(checkpoint='colbert-ir/colbertv2.0')
      indexer.index(name='lotte-index-2023', collection=collection)

      searcher = Searcher(index='lotte-index-2023')
      results = searcher.search("what is the capital of France?", k=3)
```

Leveraging ColBERT, we've been building **NLP systems** that can **search** and **cite** their sources

DSPy

Demonstrate–Search–Predict:
Composing retrieval and language models for knowledge-intensive NLP

Omar Khattab [1] Keshav Santhanam [1] Xiang Lisa Li [1] David Hall [1]
Percy Liang [1] Christopher Potts [1] Matei Zaharia [1]

HINDSIGHT: POSTERIOR-GUIDED TRAINING OF RETRIEVERS FOR IMPROVED OPEN-ENDED GENERATION

Ashwin Paranjape, Omar Khattab,
Christopher Potts, Matei Zaharia & Christopher D. Manning
Stanford University
{ashwinp,okhattab}@cs.stanford.edu

Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval

Omar Khattab **Christopher Potts** **Matei Zaharia**

Relevance-guided Supervision for OpenQA with ColBERT

Omar Khattab **Christopher Potts** **Matei Zaharia**

ColBERT has been deeply **influential** in IR and NLP

The ColBERT line of work has been cited by over 1,000 papers

Best Paper Awards (analyses & extensions of ColBERT)
- A White Box Analysis of ColBERT
- SparseEmbed: Learning Sparse Lexical Representations with Contextual Embeddings for Retrieval

Advanced ColBERT-based architectures
- ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval
- ED2LM: Encoder-Decoder to Language Model for Faster Document Re-ranking Inference
- Effective Contrastive Weighting for Dense Query Expansion
- AligneR from Google
- LAIT, LUMEN, GLIMMER from Google

Optimizations for ColBERT
- XTR from Google
- A Study on Token Pruning for ColBERT
- On Approximate Nearest Neighbour Selection for Multi-Stage Dense Retrieval
- Query Embedding Pruning for Dense Retrieval
- Static Pruning for Multi-Representation Dense Retrieval

Extensions
- Distilling Dense Representations for Ranking using Tightly-Coupled Teachers
- Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation
- VIRT: Improving Representation-based Models for Text Matching through Virtual Interaction
- I^3 Retriever: Incorporating Implicit Interaction in Pre-trained Language Models for Passage Retrieval
- SLIM: Sparsified Late Interaction for Multi-Vector Retrieval with Inverted Indexes
- Reproducibility, Replicability, and Insights into Dense Multi-Representation Retrieval Models: from ColBERT to Col*

Applications
- FILIP: Fine-grained Interactive Language-Image Pre-Training (+ 2-3 other key ones for multi-modal models)
- LI-RAGE: Late Interaction Retrieval Augmented Generation with Explicit Signals for Open-Domain Table Question Answering
- IRLab-Amsterdam at TREC 2021 Conversational Assistant Track
- Soft Prompt Tuning for Augmenting Dense Retrieval with Large Language Models
- Beyond Two-Tower Matching: Learning Sparse Retrievable Cross-Interactions for Recommendation
- Too Few Bug Reports? Exploring Data Augmentation for Improved Changeset-based Bug Localization

Out of Domain Generalization
- BEIR, RELIC, Token-Level Math Information Retrieval, Evaluating Extrapolation Performance in IR (I & II)
- NevIR: Negation in Neural Information Retrieval

Cross Lingual
- IBM’s Learning Cross Lingual IR from an English Retriever
- Cross-lingual Knowledge Transfer via Distillation for Multilingual Information Retrieval
- Multilingual ColBERT-X