A Comparative Study of **Text Retrieval Models** on **DaReCzech**

Jakub Štětina, Martin Fajčík, Michal Štefánik, Michal Hradiš

EMAILS

December 7, 2024

RASLAN 2024

Goals

- measure off-the-shelf models retrieval performance and quality

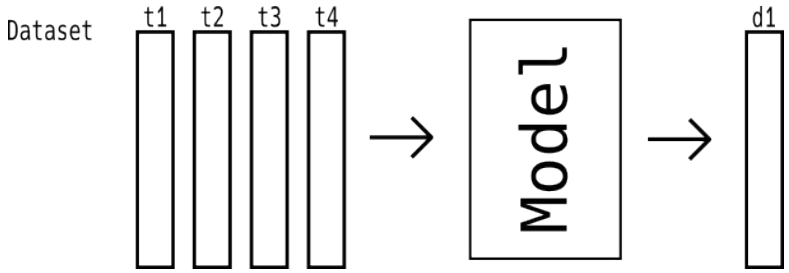
Goals

- measure off-the-shelf models retrieval performance and quality
- find the “best” model for czech retrieval
  - Performance
  - Speed
  - Index size

Retrieval Overview

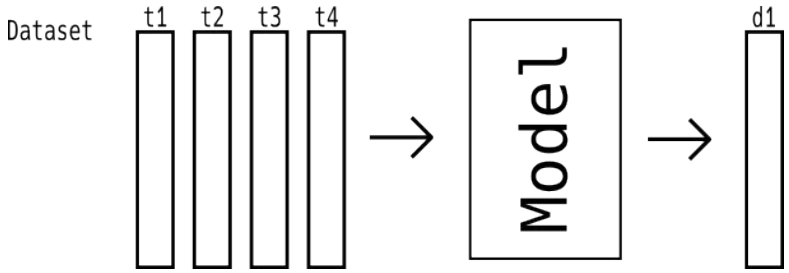
Retrieval Overview

- Indexing



Retrieval Overview

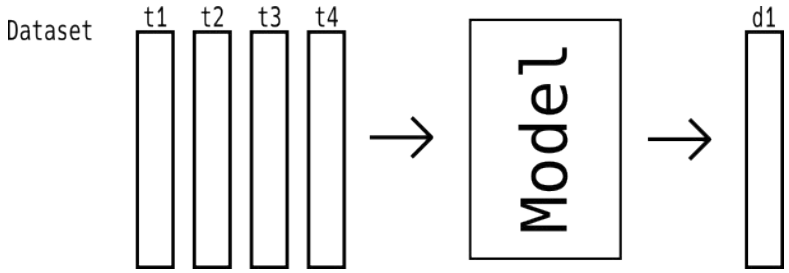
- **Indexing**



- **Representation**
  - dense - low dimensional, non-zero values
  - sparse - high dimensional, mostly zero’s

Retrieval Overview

- **Indexing**



- **Representation**
  - dense - low dimensional, non-zero values
  - sparse - high dimensional, mostly zero’s

- **Retrieval**
  - cosine similarity
  - inner product
  - MaxSim

Evaluated Models

Model	Vectors		Output dim.	Corpus lang.	Retrieval	Training data
Splade	Sparse	Single	*45.7	en	Cosine sim.	English

Formal, T., Piwowarski, B., Clinchant, S.: Splade: Sparse lexical and expansionmodel for first stage ranking (2021), https://arxiv.org/abs/2107.05720

Evaluated Models

Model	Vectors		Output dim.	Corpus lang.	Retrieval	Training data
Splade	Sparse	Single	45.7	en	Cosine sim.	English
PLAID	Dense	Multi	300 x 128	en	MaxSim	English
PLAID-X	Dense	Multi	300 x 128	cs	MaxSim	Multilingual

Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. CoRR abs/2004.12832 (2020), ttps://arxiv.org/abs/2004.12832

Evaluated Models

Model	Vectors		Output dim.	Corpus lang.	Retrieval	Training data
Splade	Sparse	Single	45.7	en	Cosine sim.	English
PLAID	Dense	Multi	300 x 128	en	MaxSim	English
PLAID-X	Dense	Multi	300 x 128	cs	MaxSim	Multilingual
Contriever	Dense	Single	768	en	Cosine sim.	English

Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings (2022), https://arxiv.org/abs/2104.08821 11.

Evaluated Models

Model	Vectors		Output dim.	Corpus lang.	Retrieval	Training data
Splade	Sparse	Single	45.7	en	Cosine sim.	English
PLAID	Dense	Multi	300 x 128	en	MaxSim	English
PLAID-X	Dense	Multi	300 x 128	cs	MaxSim	Multilingual
Contriever	Dense	Single	768	en	Cosine sim.	English
SimCSE	Dense	Single	256	cs	Cosine sim.	Czech

Bednář, J., Náplava, J., Barančíková, P., Lisický, O.: Some like it small: Czech semantic embedding models for industry applications (2023), https://arxiv.org/abs/2311.13921

Evaluated Models

Model	Vectors		Output dim.	Corpus lang.	Retrieval	Training data
Splade	Sparse	Single	45.7	en	Cosine sim.	English
PLAID	Dense	Multi	300 x 128	en	MaxSim	English
PLAID-X	Dense	Multi	300 x 128	cs	MaxSim	Multilingual
Contriever	Dense	Single	768	en	Cosine sim.	English
SimCSE	Dense	Single	256	cs	Cosine sim.	Czech
Gemma2	Dense	Single	3584	cs/en	Cosine sim.	Multilingual
OpenAI ADA	Dense	Single	1536	cs	Cosine sim.	(unknown)

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through selfknowledge distillation (2024), https://arxiv.org/abs/2402.03216

Experimental Setup

- DaReCzech (corpus with natural query-document) Seznam search engine (translated for en models)

Kocián, M., et al.: Siamese bert-based model for web
search relevance ranking evaluated on a new czech dataset (2021), https://arxiv.org/abs/2112.01810
Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval 3, 333–389 (01 2009).
https://doi.org/10.1561/1500000019

Experimental Setup

- DaReCzech (corpus with natural query-document) Seznam search engine (translated for en models)
- BM25 baseline

Kocián, M., et al.: Siamese bert-based model for web
search relevance ranking evaluated on a new czech dataset (2021), https://arxiv.org/abs/2112.01810
Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval 3, 333–389 (01 2009).
https://doi.org/10.1561/1500000019

Experimental Setup

- DaReCzech (corpus with natural query-document) Seznam search engine (translated for en models)
- BM25 baseline
- standard IR metrics
  - Precision, Recall, MRR, nDCG, Query latency
  - index size
  - overlap analysis

Kocián, M., et al.: Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset (2021), https://arxiv.org/abs/2112.01810
Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval 3, 333–389 (01 2009). https://doi.org/10.1561/1500000019

Experimental Setup

- DaReCzech (corpus with natural query-document) Seznam search engine (translated for en models)
- BM25 baseline
- standard IR metrics
  - Precision, Recall, MRR, nDCG, Query latency
  - index size
  - overlap analysis
- segmenting experiment

Kocián, M., et al.: Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset (2021), https://arxiv.org/abs/2112.01810
Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval 3, 333–389 (01 2009). https://doi.org/10.1561/1500000019

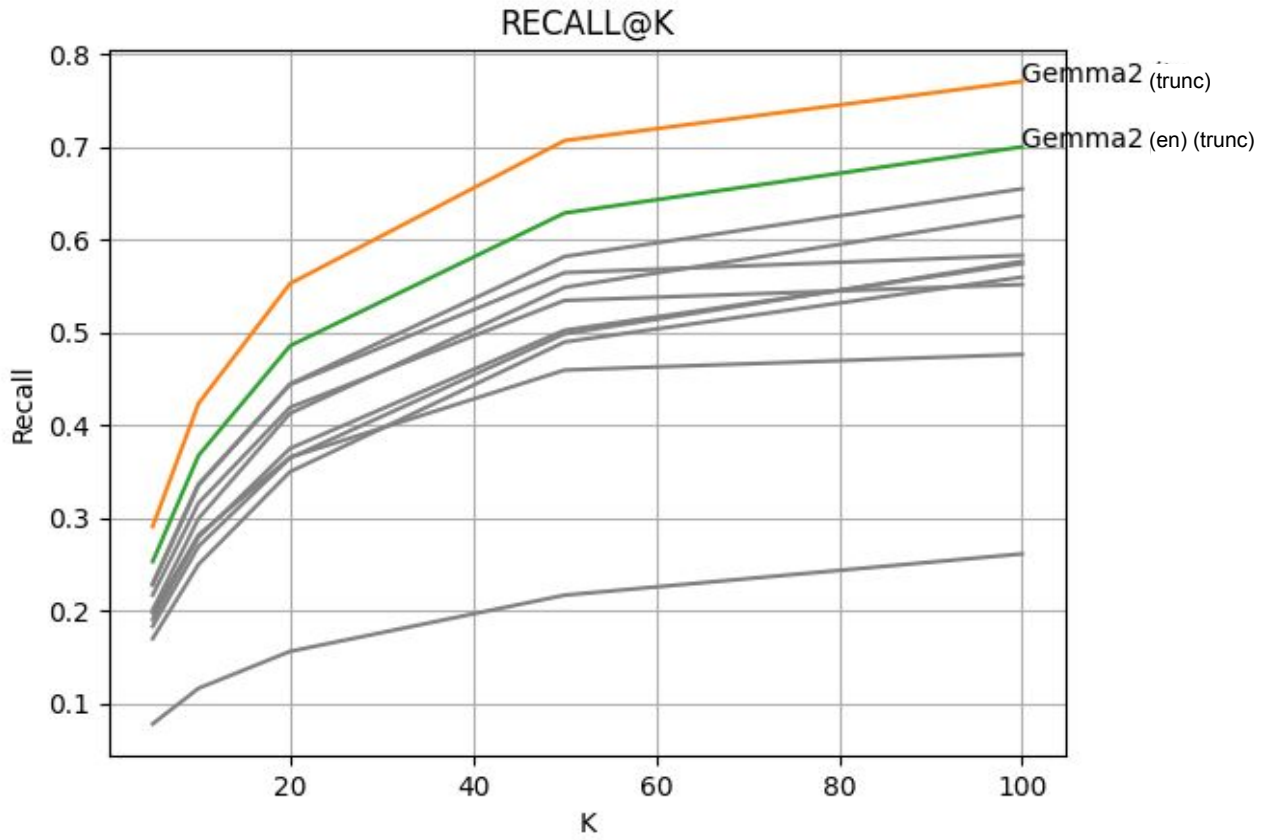
Experimental Setup

- DaReCzech (corpus with natural query-document) Seznam search engine (translated for en models)
- BM25 baseline
- standard IR metrics
  - Precision, Recall, MRR, nDCG, Query latency
  - index size
  - overlap analysis
- segmenting experiment

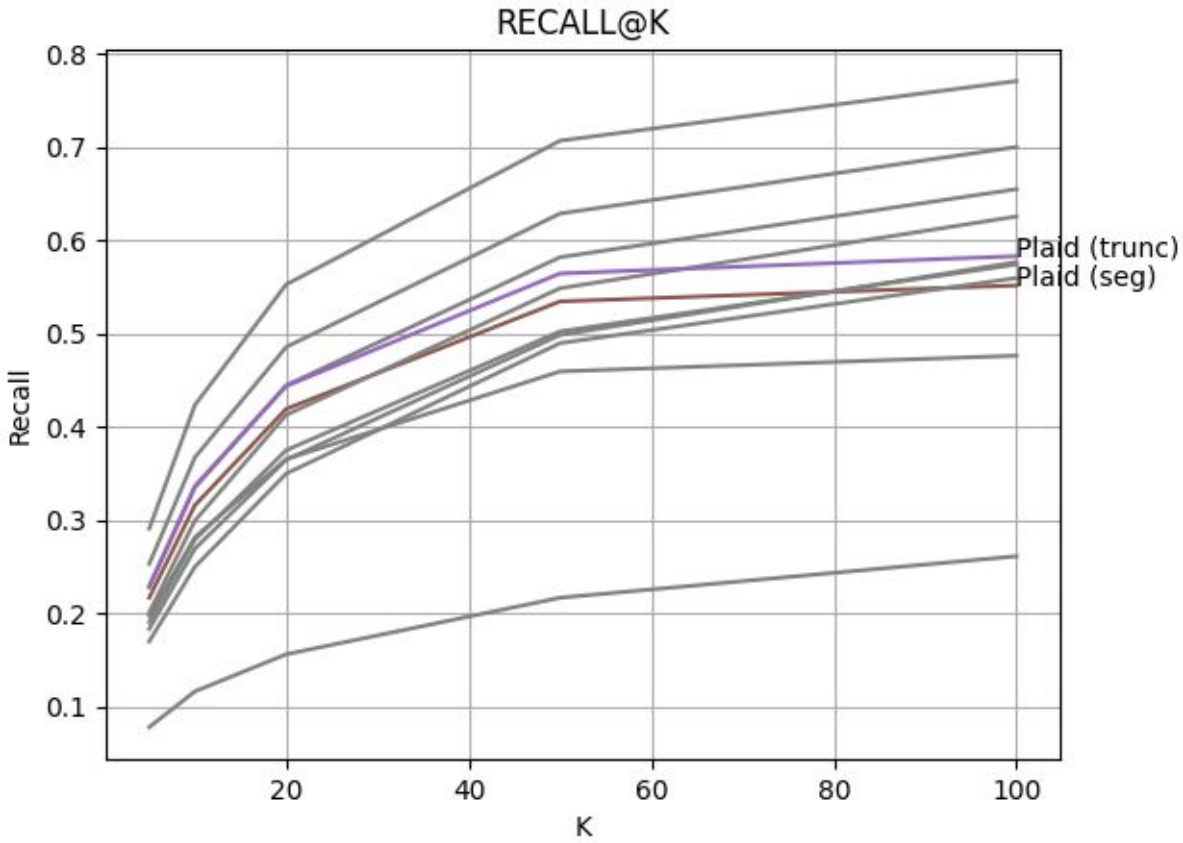
Model	Output Dim.	Max Tokens	Lang.	Segmenting
Splade	*45.7	256	en	256/86
PLAID	128	300	en	300/100
PLAID-X	128	180	cs	180/60
Contriever	768	256	en	–
SimCSE	256	128	cs	–
Gemma2	3584	512	cs/en	–
OpenAI ada	1536	8192	cs	–

Kocián, M., et al.: Siamese bert-based model for web
search relevance ranking evaluated on a new czech dataset (2021), https://arxiv.org/abs/2112.01810
Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval 3, 333–389 (01 2009).
https://doi.org/10.1561/1500000019

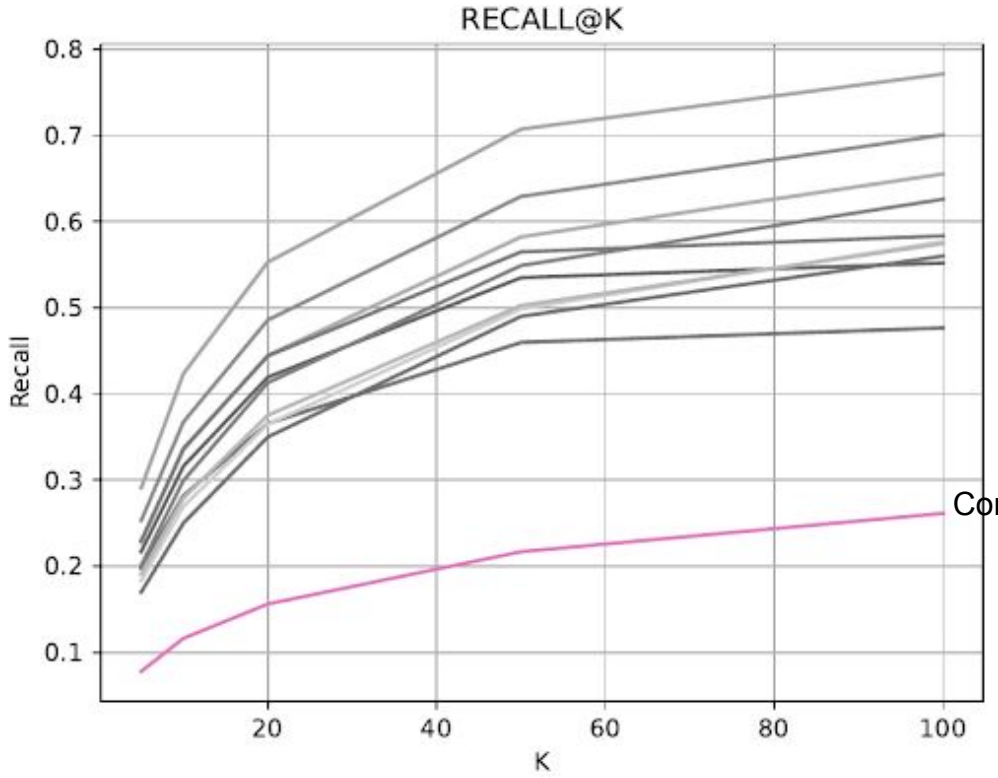
Results


RECALL@K
Recall
K
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
20
40
60
80
100
Gemma2 (trunc)
Gemma2 (en) (trunc)

Results

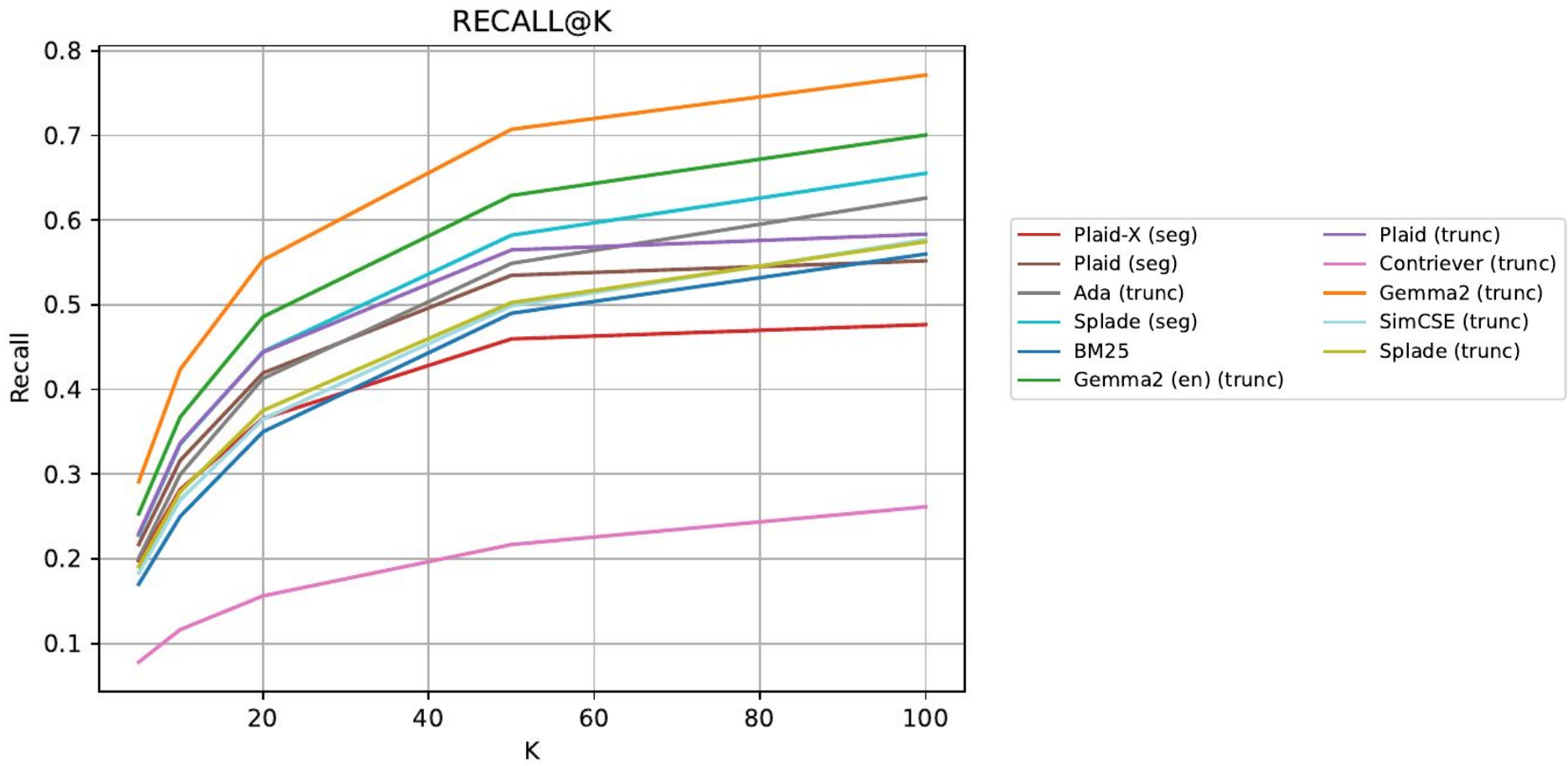

RECALL@K
Recall
K
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
20
40
60
80
100
Plaid (trunc)
Plaid (seg)

Results

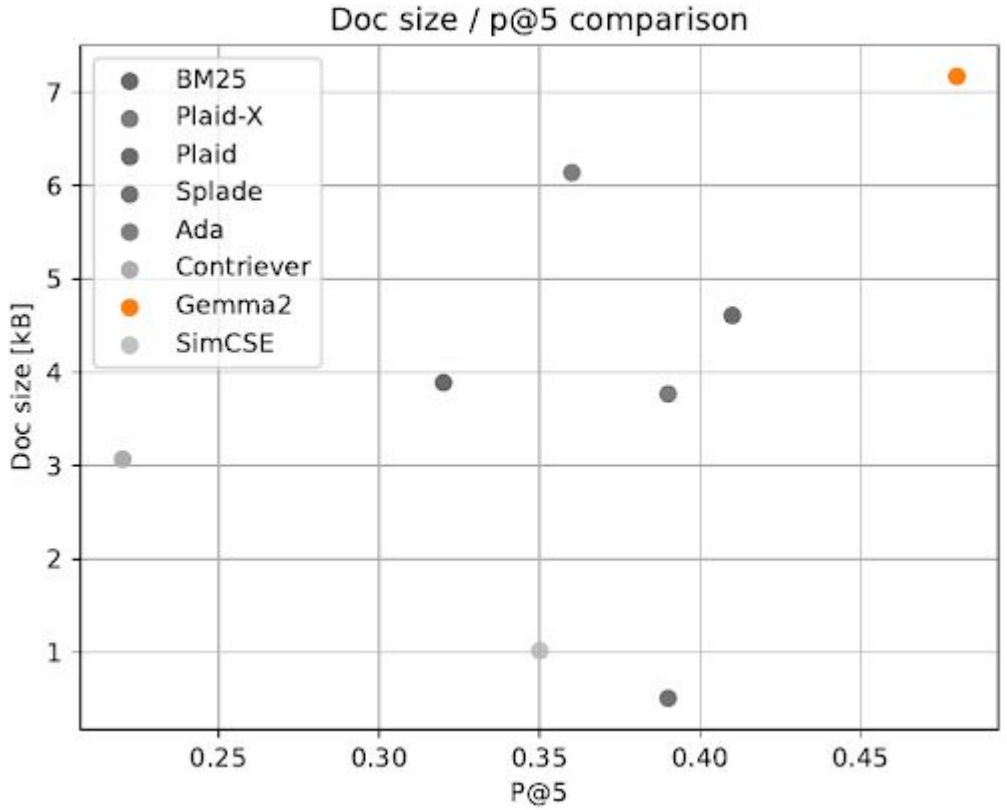


Contriever

Results


RECALL@K
Recall
K
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
20
40
60
80
100
Plaid-X (seg)
Plaid (seg)
Ada (trunc)
Splade (seg)
BM25
Gemma2 (en) (trunc)
Plaid (trunc)
Contriever (trunc)
Gemma2 (trunc)
SimCSE (trunc)
Splade (trunc)

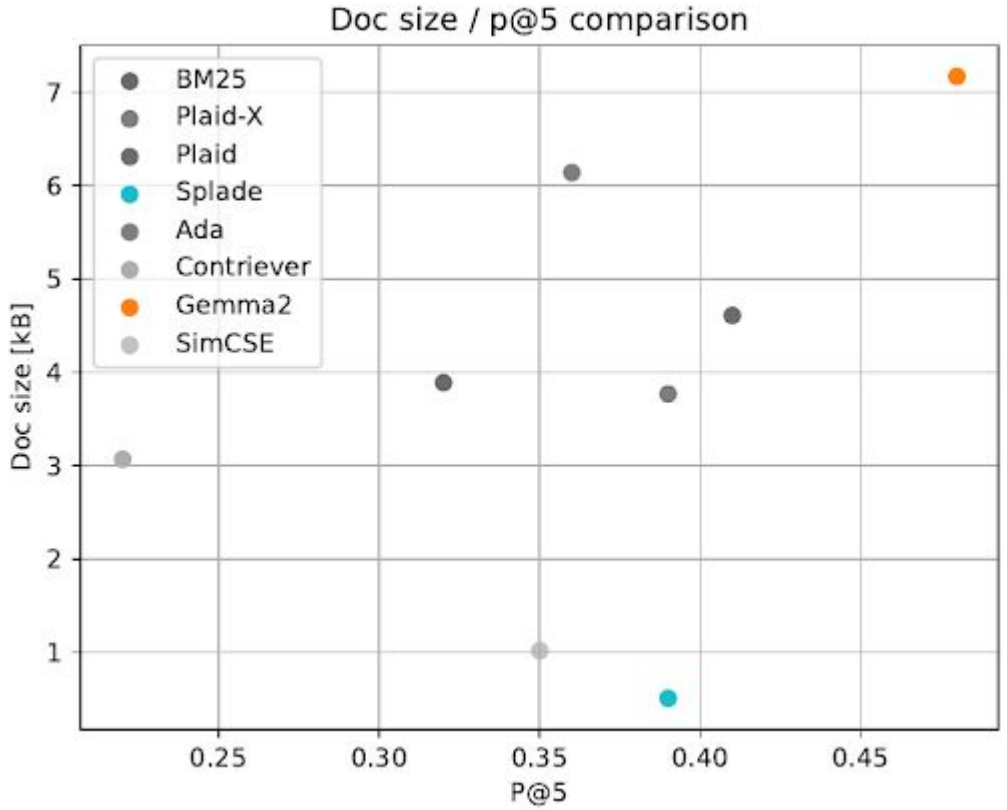
Results – Storage



Model	Output Dim.	Max Tokens
Splade	*45.7	256
PLAID	128	300
PLAID-X	128	180
Contriever	768	256
SimCSE	256	128
Gemma2	3584	512
OpenAI ada	1536	8192

* average value

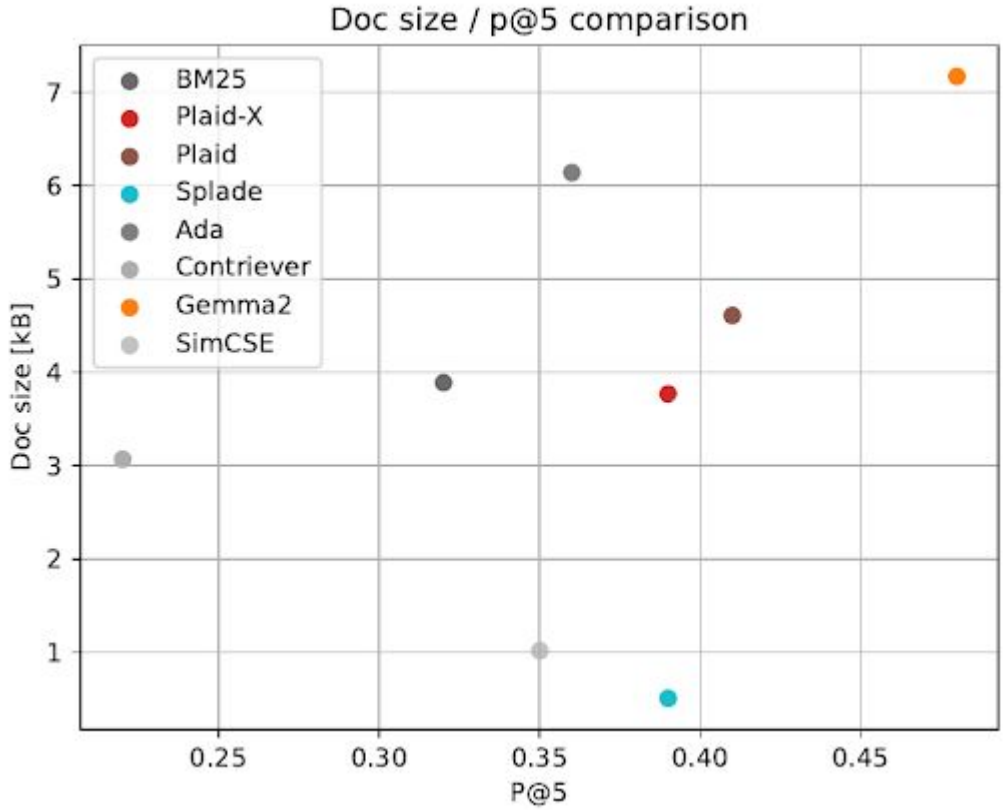
Results – Storage



Model	Output Dim.	Max Tokens
Splade	*45.7	256
PLAID	128	300
PLAID-X	128	180
Contriever	768	256
SimCSE	256	128
Gemma2	3584	512
OpenAI ada	1536	8192

* average value

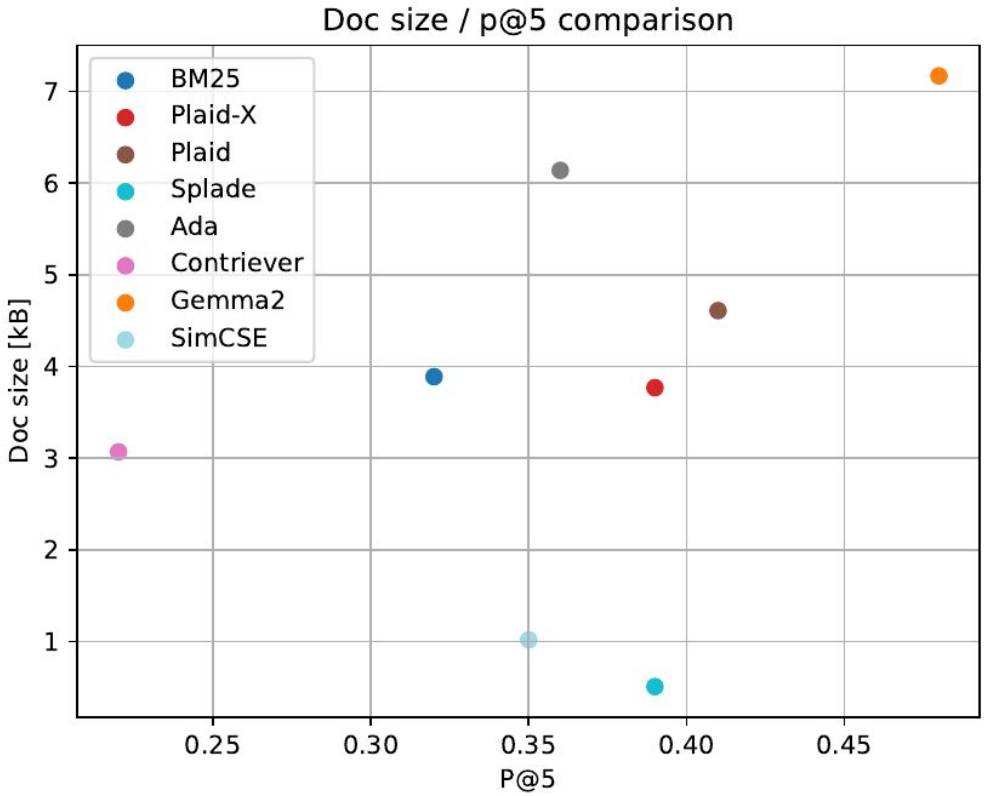
Results – Storage



Model	Output Dim.	Max Tokens
Splade	*45.7	256
PLAID	128	300
PLAID-X	128	180
Contriever	768	256
SimCSE	256	128
Gemma2	3584	512
OpenAI ada	1536	8192

* average value

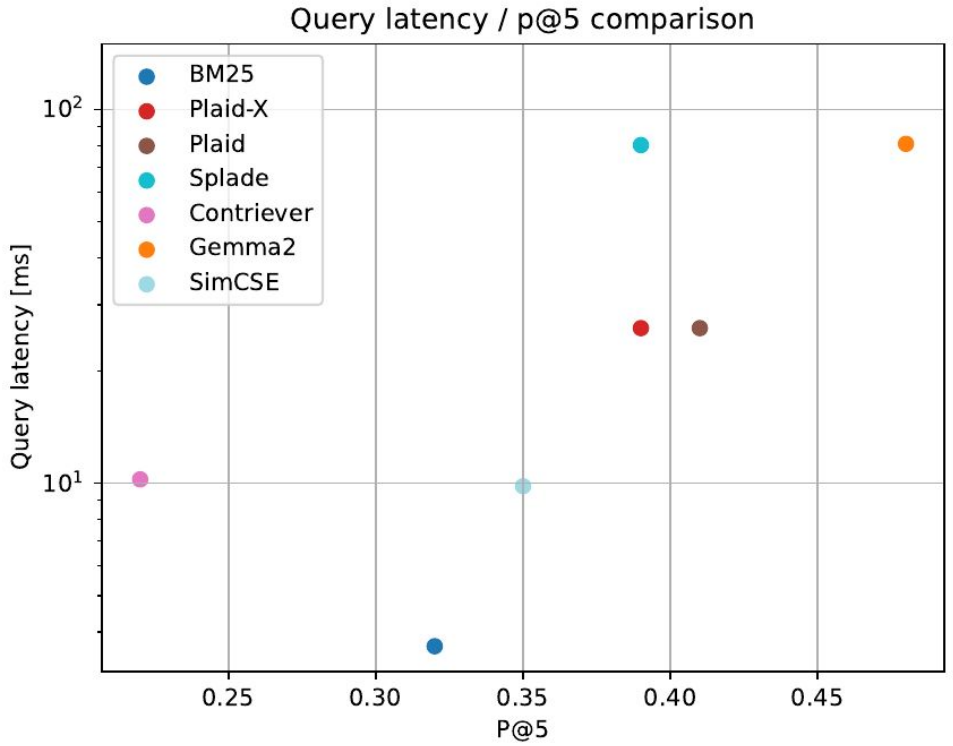
Results – Storage



Model	Output Dim.	Max Tokens
Splade	*45.7	256
PLAID	128	300
PLAID-X	128	180
Contriever	768	256
SimCSE	256	128
Gemma2	3584	512
OpenAI ada	1536	8192

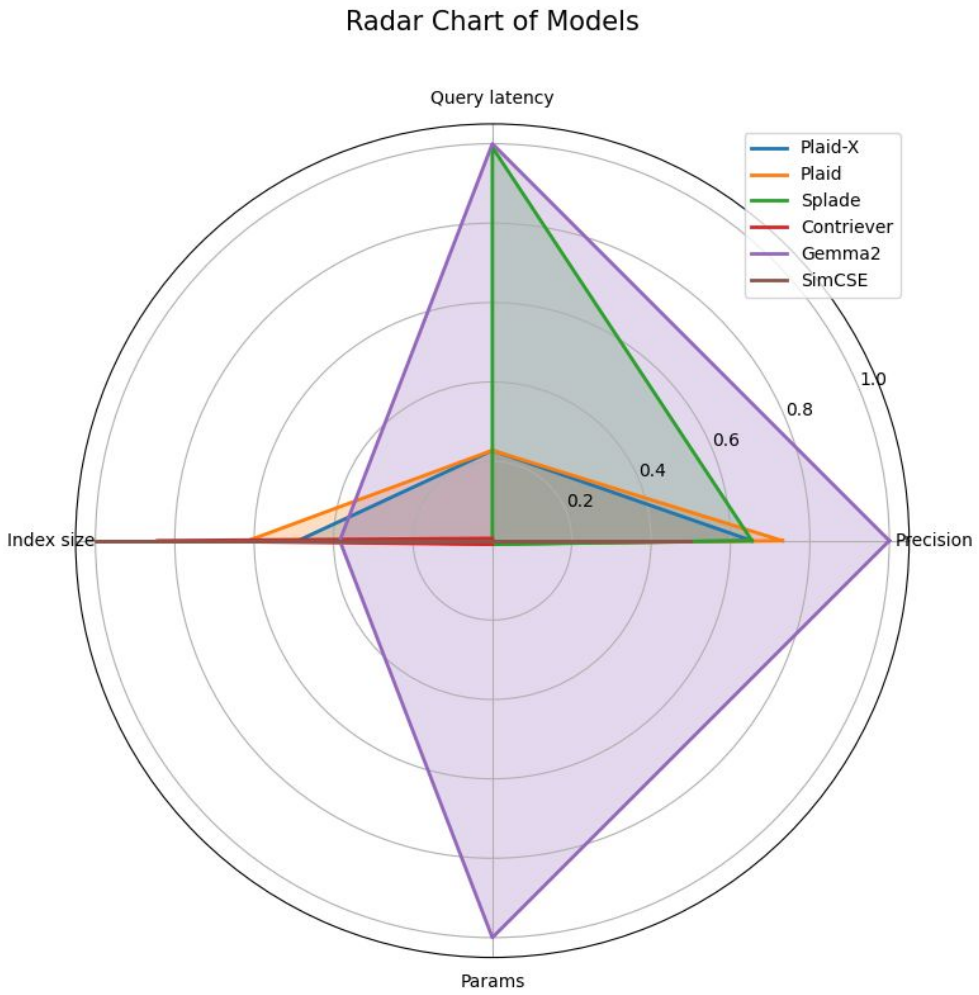
* average value

Results – Speed



Takeaways

- **Performance**: Gemma2
- **Storage**: Spladev2
- **Versatile:** Plaid(x)



References

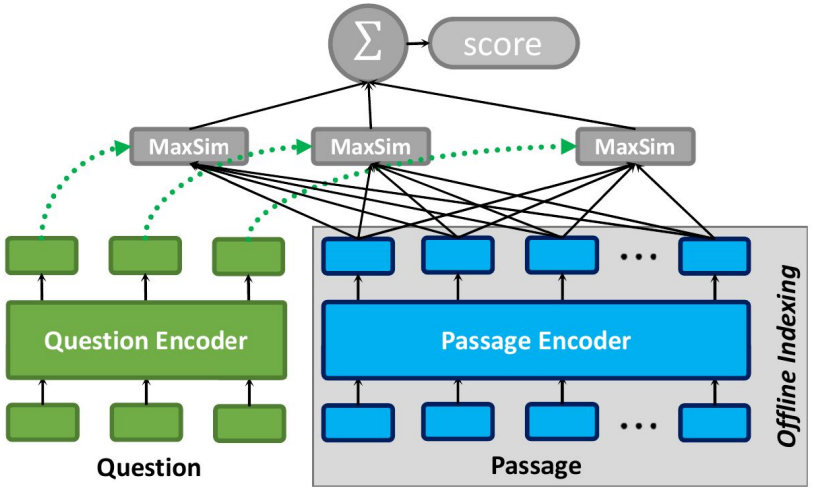
1. Formal, T., Piwowarski, B., Clinchant, S.: Splade: Sparse lexical and expansionmodel for first stage ranking (2021), https://arxiv.org/abs/2107.05720
2. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. CoRR abs/2004.12832 (2020), https://arxiv.org/abs/2004.12832
3. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings (2022), https://arxiv.org/abs/2104.08821
4. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through selfknowledge distillation (2024), https://arxiv.org/abs/2402.03216 Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-tra
5. Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J.M., Tworek, J., Yuan, Q., Tezak, N., Kim, J.W., Hallacy, C., et al.: Text and code embeddings by contrastive pre-training. arXiv preprint arXiv:2201.10005 (2022)
6. OpenAI: Openai ada model for retrieval. https://platform.openai.com/docs/ models/embeddings (2023), accessed: 2024-10-29
7. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings (2022), https://arxiv.org/abs/2104.08821 11.
8. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through selfknowledge distillation (2024), https://arxiv.org/abs/2402.03216
9. Bednář, J., Náplava, J., Barančíková, P., Lisický, O.: Some like it small: Czech semantic embedding models for industry applications (2023), https://arxiv.org/abs/2311.13921
10. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over BERT. CoRR abs/2004.12832 (2020), https://arxiv.org/abs/2004.12832
11. Kocián, M., et al.: Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset (2021), https://arxiv.org/abs/2112.01810
12. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends in Information Retrieval 3, 333–389 (01 2009). https://doi.org/10.1561/1500000019

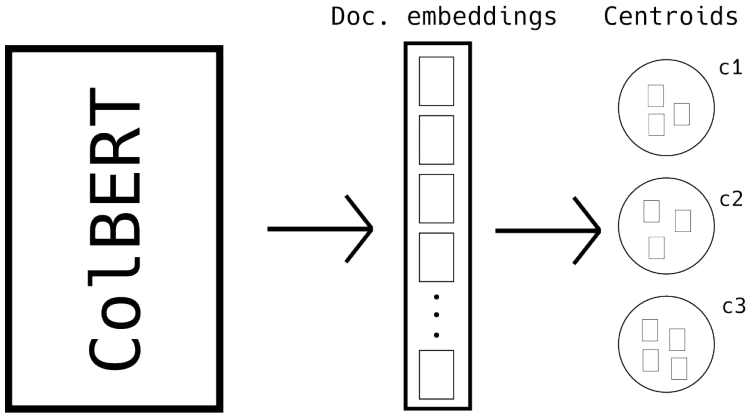
Thank you
for your attention.

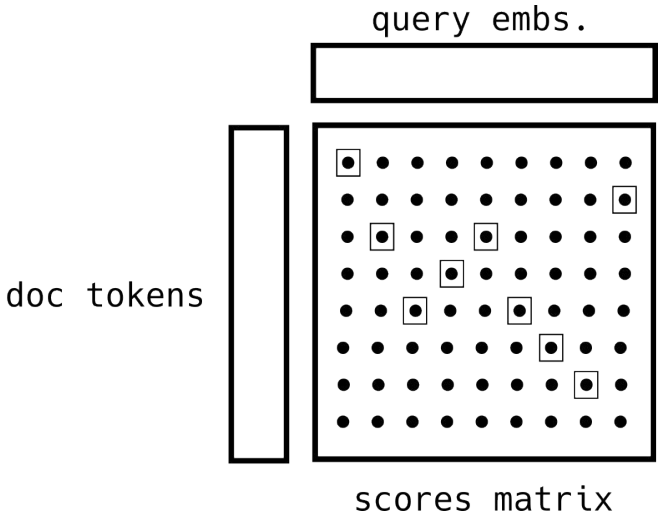
PLAID & PLAIDX

(ColbertV2.0) (plaidx-xlmr-large-mlir-neuclir)

- Dense, multi-vector, english/x-lingual,
- Uses XLM-R for cross-lingual encodings; trained on Chinese, Persian, and Russian.





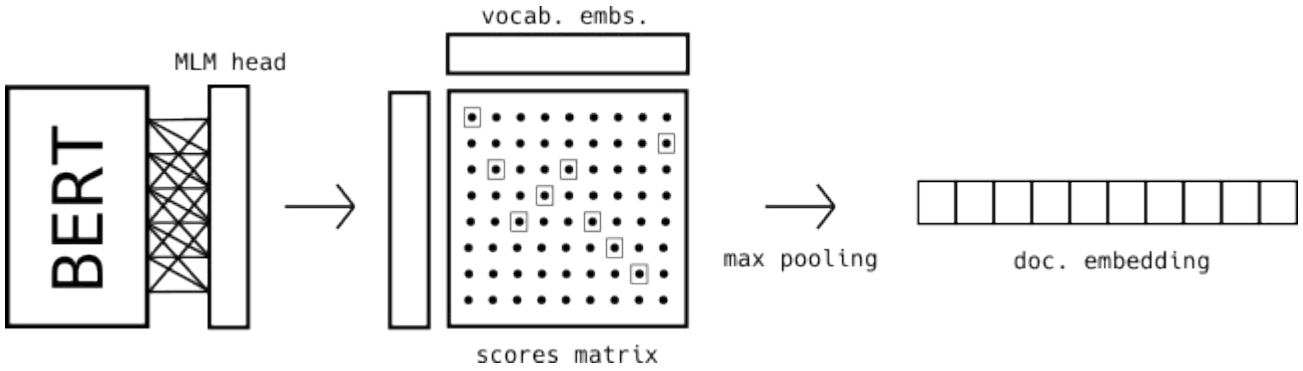


$$S_{q,d} = \sum_{i=1}^{N} \max_{j=1}^{M} Q_i \cdot D_j^T$$

SpladeV2

(splade-cocondenser-ensemledistil)

- English, sparse, single-vector.
- Sparse vocabulary-sized embeddings
- Uses FLOPS regularizer and logarithmic sparsity for efficient representation.



Contriever

(contriever-msmarco)

- English, dense, single-vector.
- Self-supervised contrastive learning for query and document representations.
- Generates positive pairs via window cropping and token deletion; negatives via MoCo sampling.
- Fine-tuned on the English MS MARCO dataset for enhanced performance.

SimCSE

(simcse-dist-mpnet-paracrawl-cs-en)

- Czech, dense, single-vector.
- Trained using contrastive learning
- Pre-trained on a Czech dataset from Seznam.cz
- Optimized for DaReCzech performance.

Gemma2

(BGE Multilingual Gemma2)

- Multilingual, dense, single-vector.
- Fine-tuned with a contrastive objective on diverse languages.
- Based on Google’s Gemma-2-9B model

OpenAI ada

(Text-embedding-ada-002)

- English, dense, single-vector.
- Closed-source