Optimization of the Search Experience in Search Engines with Vector Databases and Transfer Learning

Under the supervision of:

Prof. Dr. rer. nat. habil. Gunter Saake
Dr.-Ing. David Broneske

6th September 2023

Ashish Soni (221453)

Data and Knowledge Engineering (M.Sc.)

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

INF FAKULTÄT FÜR INFORMATIK

Agenda

What is Search?

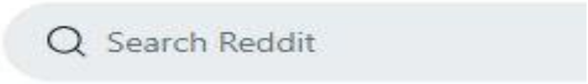
- **Search**, also known as information retrieval is the process of taking a user query and returning **ranked, relevant results**
- The **first** modern information retrieval system was built in the **1960s[1]** led by **Gerard Salton** with his research group at **Cornell**
- **Google** started as research project in the late **1990s[2]** becoming the world's dominant search engine due to two key innovations - **MapReduce** and **PageRank**
- **Currently,** Platforms such as **Quora**, **Reddit** and **Stack Overflow** have refined search, offering organized and user-specific content in the digital age

[1] Source: https://en.wikipedia.org/wiki/Information_retrieval
[2] Source: https://en.wikipedia.org/wiki/History_of_Google

What is the significance of Search?

- According to **Internet Live Stats (May 2023)[1],** Google runs around **8.5 billion searches per day**
- **Quora[2], Reddit[3]** has **300+ million** monthly visitors
- Inaccurate or Irrelevant search results **can lead to misinformation**, **decreased user trust**, **lost productivity** and **bad decision making**

Source: https://www.quora.com



Source: https://www.reddit.com

[1] Source: https://fitsmallbusiness.com/google-search-statistics/#searches-on-google
[2] Source: https://www.demandsage.com/quora-statistics
[3] Source: https://foundationinc.co/lab/reddit-statistics

Goal

Enhancing the relevance and speed of results in search engines through the integration of Vector Databases and Transfer Learning

Motivation

- Evolving landscape of Q&A platforms
- Changing nature of search
- Advancements in neural network approaches

Main Contributions

- Comparison of vector databases - Milvus, Pinecone, Qdrant, Weaviate
- Performance benchmarking: Pinecone vs. PostgreSQL vs. PostgreSQL + pgvector
- Zero-shot evaluation of multilingual models
- Development of a multilingual semantic search prototype using quora dataset

Background: Lexical Search

- **also called Keyword Search[1]**

  Query: Where was the last world cup?

Sentences
The previous world cup was in Qatar
The sky is blue
The bear lives in the woods
An apple is a fruit

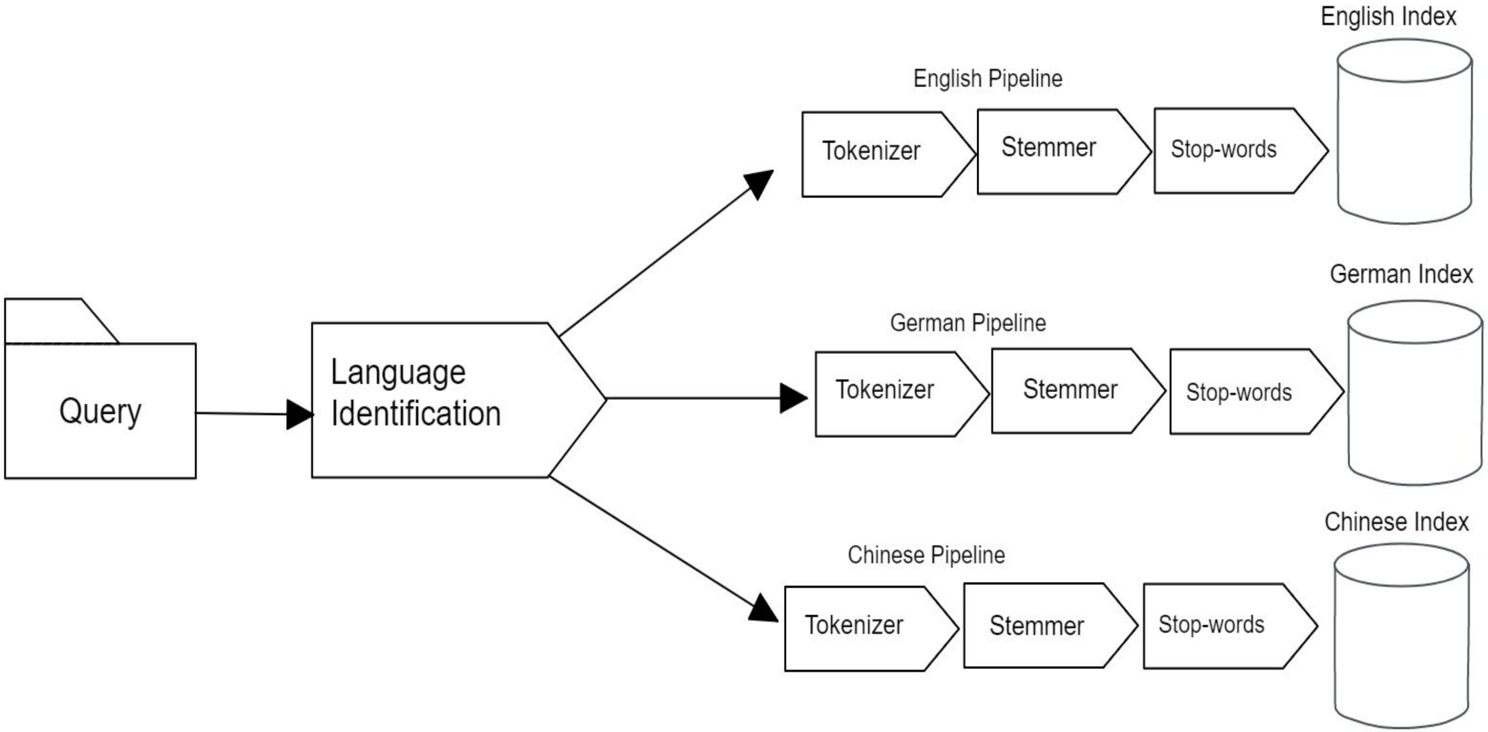
- **Lexical Search Problems…**

Sentences
The previous world cup was in Qatar
The cup is where you left it
Where in the world is my last cup of coffee?
An apple is a fruit

[1] Source: https://docs.cohere.com/docs/what-is-semantic-search

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

Background: Multilingual Lexical Search



Multilingual Lexical Search

Challenges

- Multilingual Support
- Storage
- Engineering
- Latency
- Maintenance

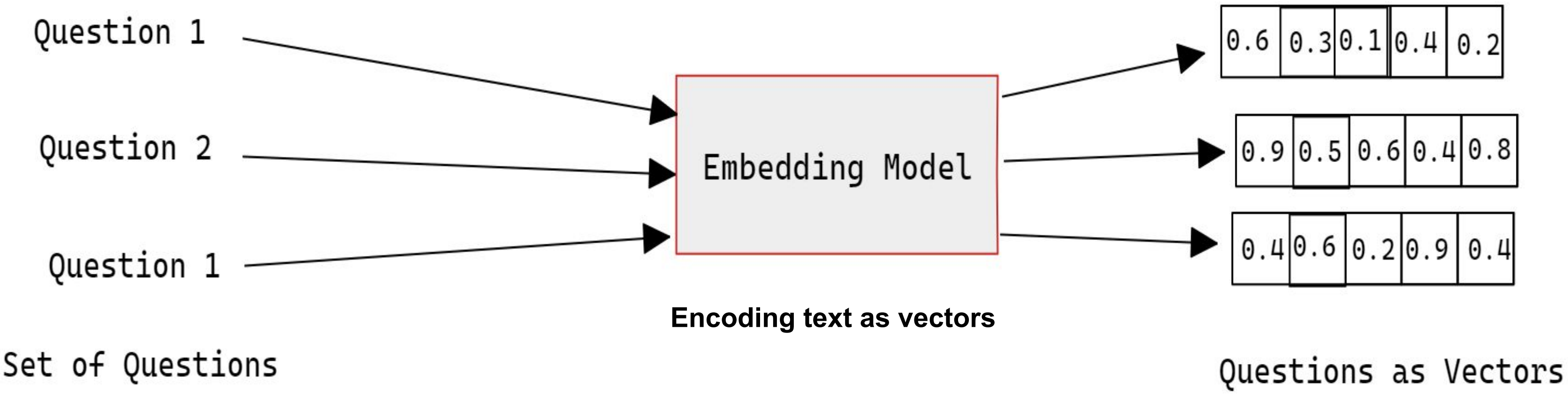
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

Background: Semantic Search

What is an Embedding?

It is a way to assign each piece of text, a vector, which is a list of numbers[1].

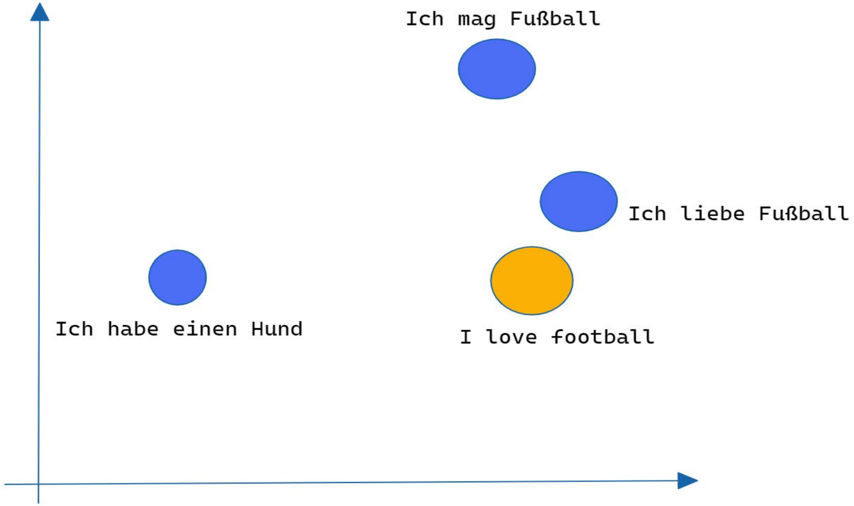
- Word Embeddings
- Sentence Embeddings



Encoding text as vectors

[1] *Pinecone. Dense vector embeddings for nlp. URL: https://www.pinecone.io/learn/dense-vector-embeddings-nlp*

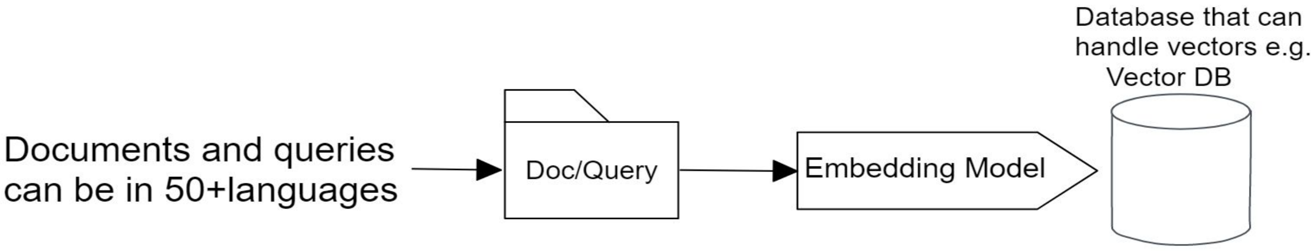
Background: Multilingual Semantic Search



Similar sentences from different languages to similar vector spaces

Challenges

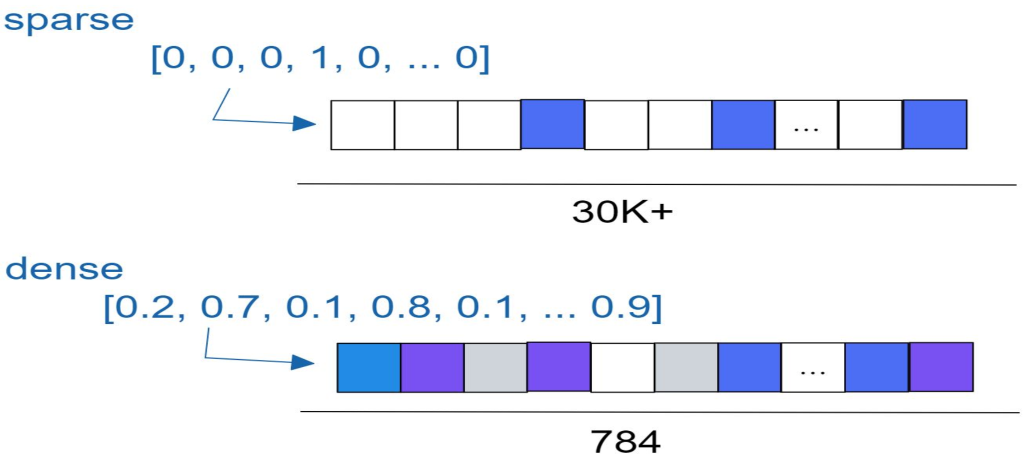
- Computational resources
- Maintenance



Multilingual Semantic Search

Background: Sparse vs Dense vectors

- Sparse vectors: hold information sparsely
- Dense vectors: large number of dimensions hold relevant information

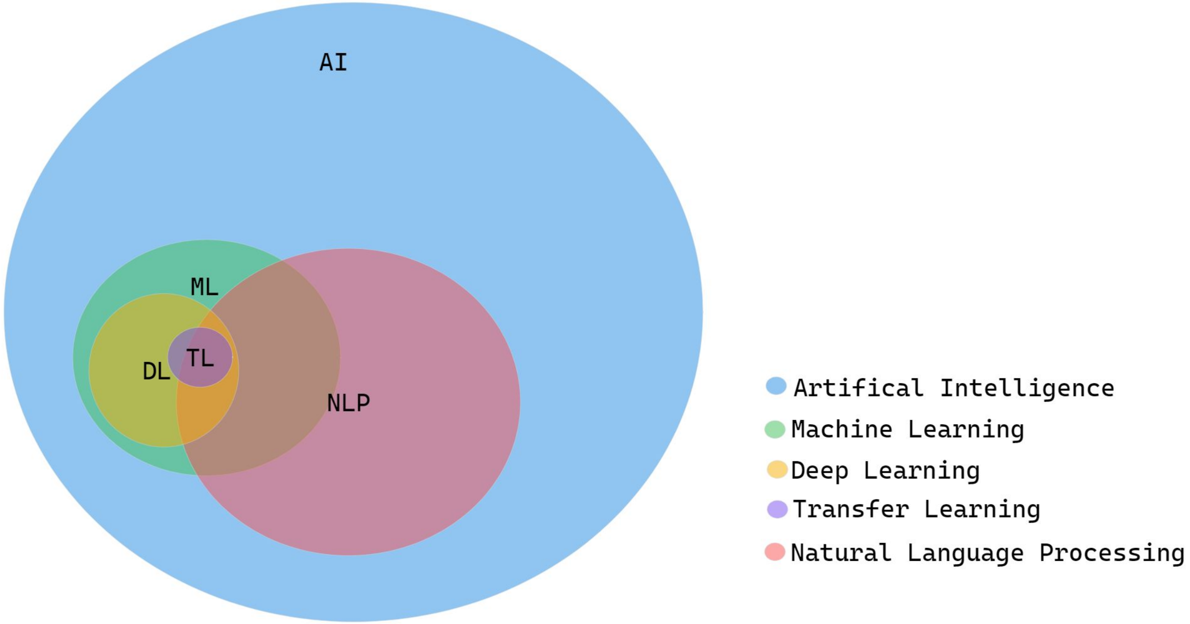


Comparison of sparse and dense vectors

[1] *Pinecone. Dense vector embeddings for nlp. URL:* https://www.pinecone.io/learn/dense-vector-embeddings-nlp

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Background: Transfer Learning



Depiction of Relationships between areas of AI

Advantages

- Leverages knowledge from one task to help learn a new related task
- Reduces the need for extensive data and training time for the new task

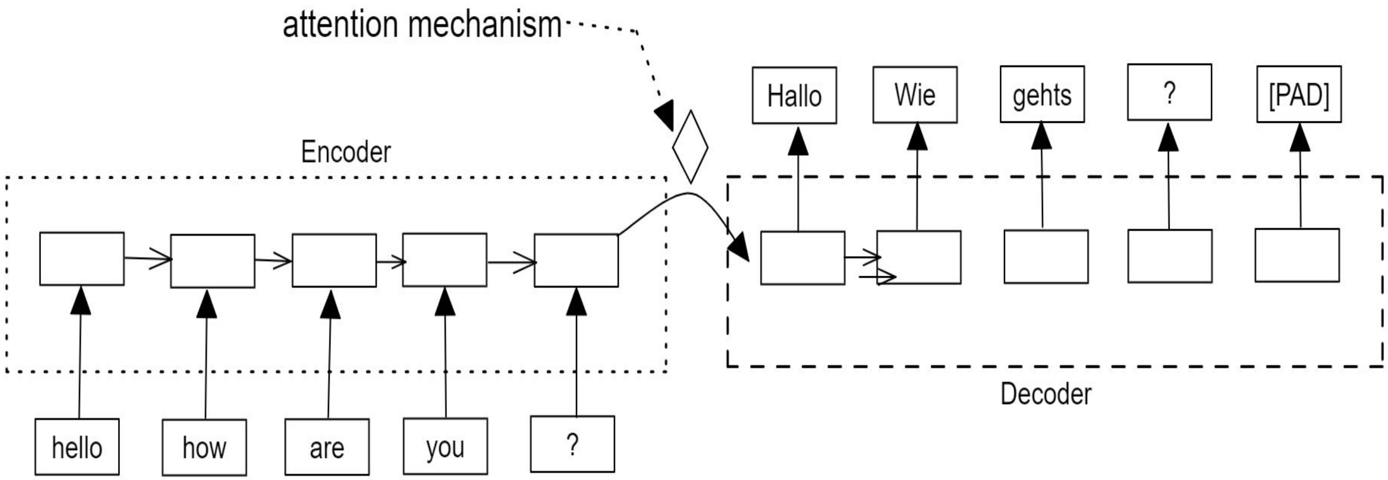
[1] *Attention is all you need. 2017. URL:* http://arxiv.org/abs/1706.03762

UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Background: Attention Mechanism

NLP had a major breakthrough in 2017[1]:

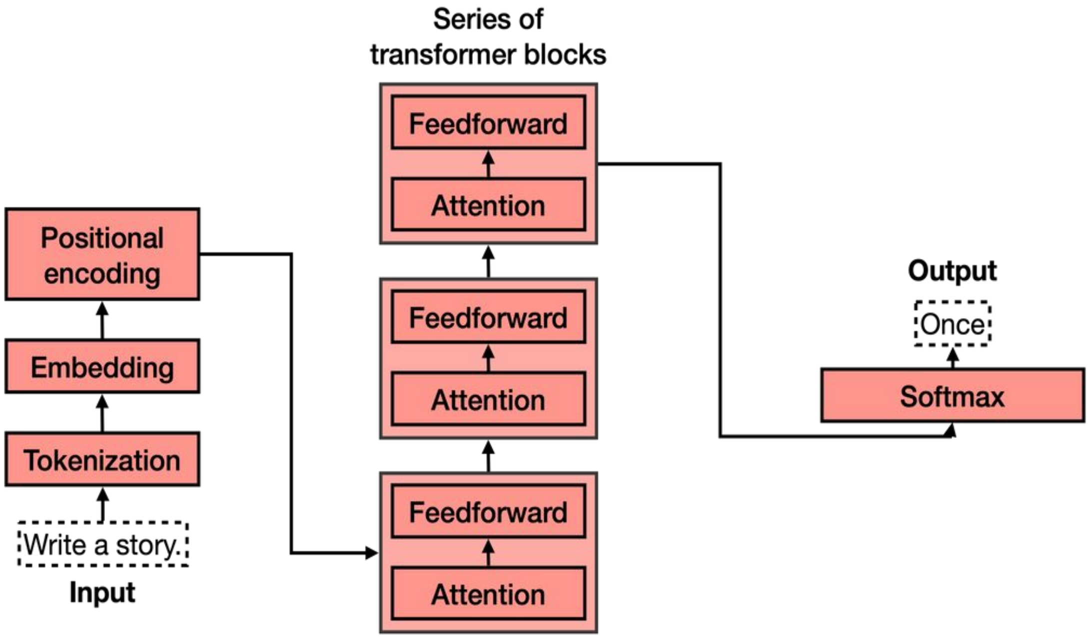
-



Encoder-Decoder with attention mechanism

[1] *Attention is all you need. 2017. URL:* http://arxiv.org/abs/1706.03762

Background: Transformer Models



The architecture of the transformer model

The transformer has the following main parts[1]:

- Tokenization
- Embedding
- Positional Encoding
- Transformer block
- Softmax

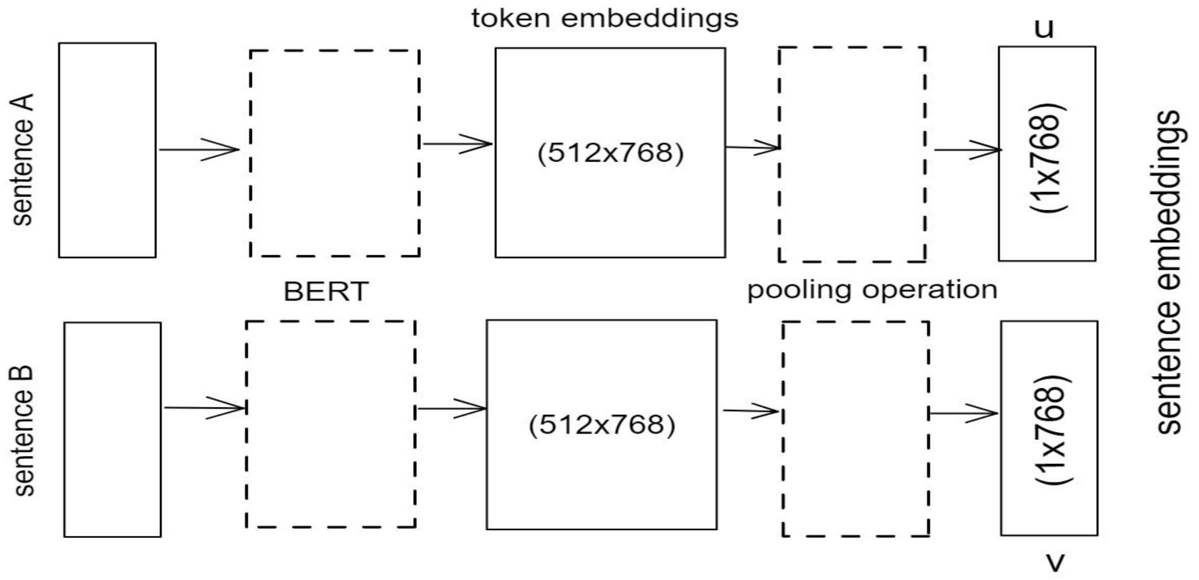
[1] Source: https://docs.cohere.com/docs/transformer-models

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Background: Sentence Transformers

NLP had a major breakthrough in 2017[1]:

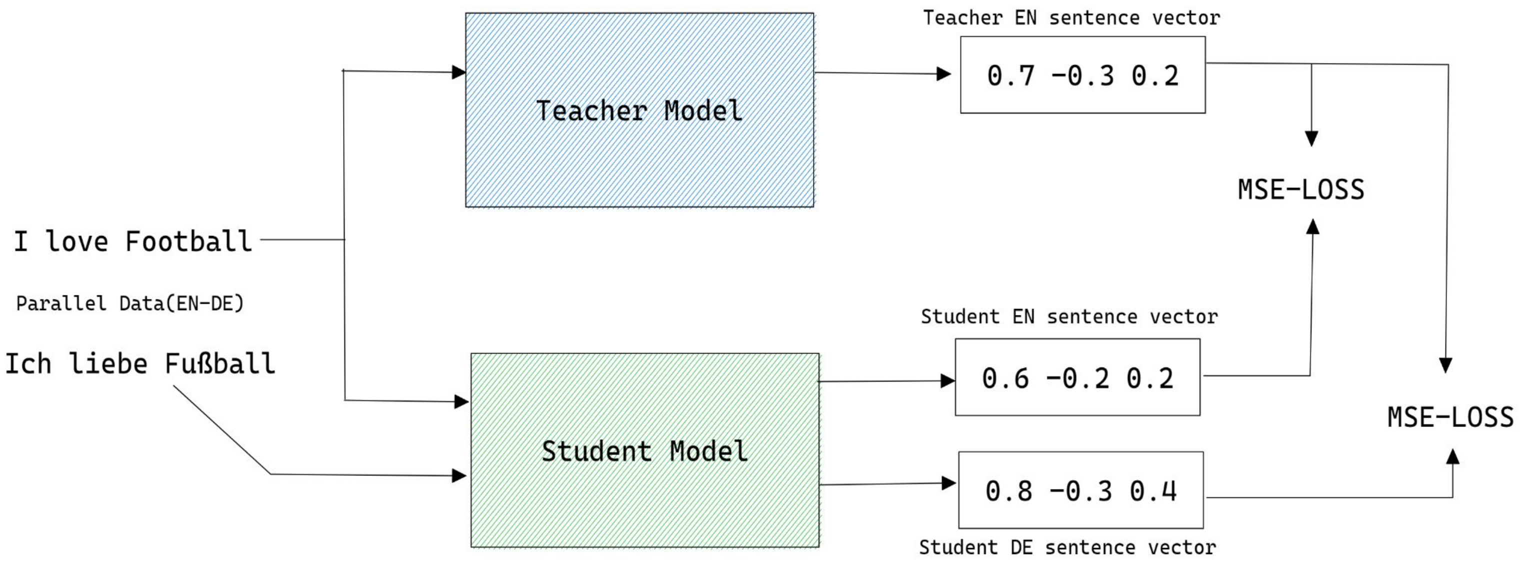
-



SBERT applied to a pair of sentences

[1] *Sentence-bert: Sentence embeddings using siamese bert-networks. 2019. URL:* https://arxiv.org/abs/1908.10084

Background: Multilingual Sentence Transformers



Knowledge Distillation Training Approach

[1] *Making monolingual sentence embeddings multilingual using knowledge distillation. 2020. URL:* http://arxiv.org/abs/2004.09813

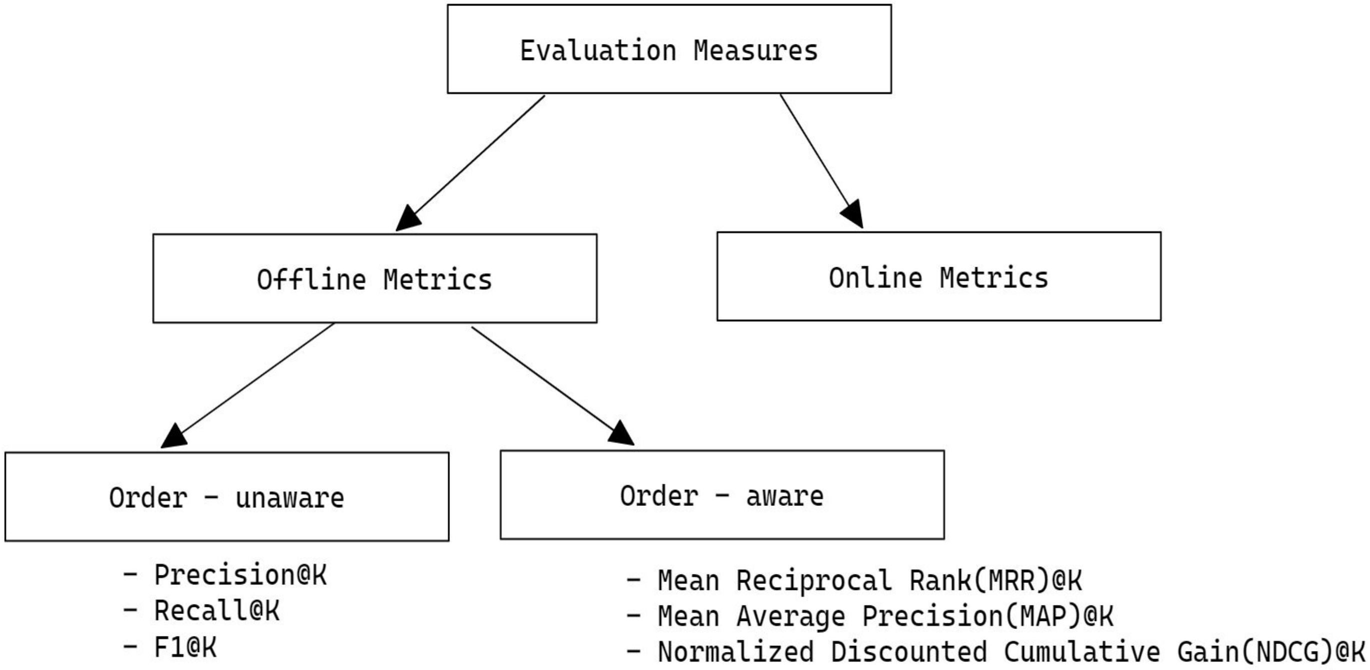
Background: Traditional Approaches

- TF-IDF, BM25, Word2Vec
- BERT
- ANN Indexes

[1] *Zilliz. Milvus. https://milvus.io/docs/index.md*
[2] *Text mining: Use of tf-idf to examine the relevance of words to documents*
[3] *Pinecone. Semantic search. https://www.pinecone.io/learn/semantic-search*

Background: Evaluation Measures

Metrics[1]



Metrics to assess the performance of the IR system

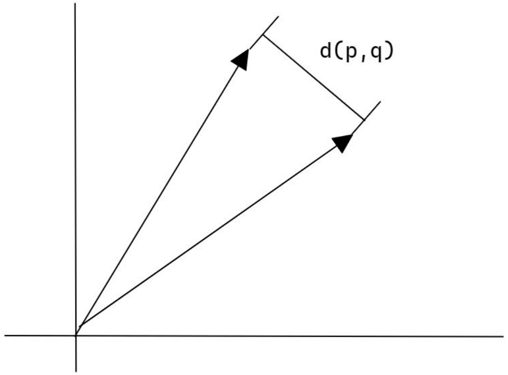
[1] *A survey on performance evaluation measures for information retrieval system. 2015*
[2] *BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models*

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

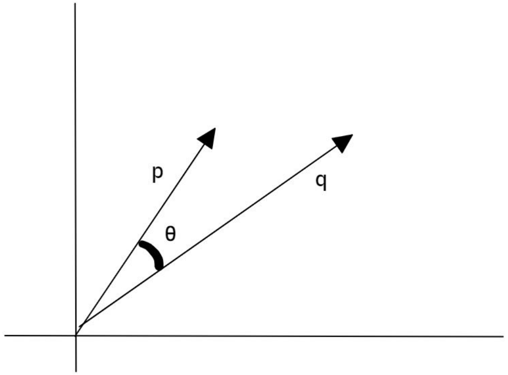
Background: Similarity Metrics

Name	Vector Properties Considered
Euclidean Distance	Magnitudes and Direction
Cosine Similarity	Only Direction
Dot product Similarity	Magnitudes and Direction

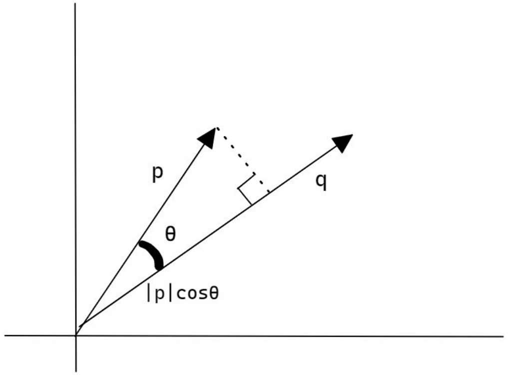
Metrics to compare similarity between vectors



Euclidean Distance



Cosine Similarity



Dot Product

Background: Vector Databases

- Designed to index and store vector embeddings
- CRUD capabilities along with metadata filtering

Feature	Qdrant	Milvus	Weaviate	Pinecone
Version	1.1.0	2.2.4	1.18.2	2.2.0
Written in	Rust	Go	Go	Rust
Consistency Model	Eventual	Strong	Eventual	Eventual
Max Vectors Dimension	N/A	32,768	N/A	20,000
Deployment	Managed/ Self Hosted	Self-Hosted	Managed/Se lf Hosted	Managed
Filtering with ANN	N/A	N/A	Pre-filtering	Single-stage filtering

[1] *Manu: a cloud native vector database management system*
[2] *Powering ai with vector databases: A benchmark(part i). https://www.farfetchtechblog.com/en/blog/post/powering-ai-with-vector-databases-a-benchmark-part-i*

Background: Vector Library vs Vector Database

Attributes	Vector Library	Vector Database
Filtering with Vector Search	No	Yes
CRUD Support	No	Yes
Stores Objects and Vectors	No	Yes
Speed	Faster	Slower
Durability	No	Yes
Persistence	Only at explicit Snapshot	Immediate
Sharding	No	Yes
Replication	No	Yes
Multi-tenancy	No	Yes
Hybrid Search	No	Yes

[1] *Weaviate. Vector library vs vector database.* https://weaviate.io/blog/vector-library-vs-vector-database

Background: Relational vs Vector Database

Feature	PostgreSQL	Pinecone	PostgreSQL with pgvector
Data Types	Standard + JSON/blob	Vectors	Standard + Vectors
Geometric Filters	No	Yes	Yes
Query Language	SQL	Model(Query, X) > threshold	SQL
Max Vector Dimensions	-	20,000	16,000
Distance Metric	-	Euclidean, Cosine, Dot Product	Euclidean, Cosin e, Dot Product
ANN Based Algorithm	-	Graph Based	Inverted File Index
Programming Language	C	Rust	C

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Design: Research Questions

1. What are the currently available vector databases and their capabilities?
2. How does a vector database perform as compared to a PostgreSQL database (with and without pgvector-python extension)?
3. What are some of the similarity functions that help to measure the similarity of vectors?
4. What are the currently available multilingual models that can perform a semantic search?
5. How do these multilingual models differ in terms of zero-shot evaluation of their retrieval capabilities and comparison of their inference speed?

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF FAKULTÄT FÜR INFORMATIK

Design and Experimental Setup for Database Performance

- Synthetic Data for Assessing Database Performance Benchmarks
- Varied Row Quantities and Embedding Dimensionality

Name	Datatype
Id	Integer
Sentence	Object
Embeddings	Object

Name	Id	Sentence	Embeddings
PostgreSQL	Integer()	Text()	ARRAY(Float)
Pinecone	String	String	Vector
PostgreSQL(pgvector)	Integer()	Text()	Vector(size)

Design and Experimental Setup for Database Performance

Core Tasks:

1. Implementation of CRUD Operations (Create, Read, Update, Delete)

Row Size	Embedding Size
100	384
200	512
300	768
400	1024

K
3
5
10
100
250

2. Efficient Handling of Batch Insertions

Total Number of Rows
10000
30000
50000
70000
100000

Experimental Environment for Database Performance

Machine Configuration

- Operating System Microsoft Windows 10 Home
- Processor Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz, 2701 Mhz, 2 Core(s),
- 4 Logical Processor(s)
- Memory 8.0 GB RAM
- Graphics NVIDIA® GeForce® 920MX (2 GB DDR3 dedicated)

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF FAKULTÄT FÜR INFORMATIK

Design and Experimental Setup for Multilingual Models

- Conducted zero-shot evaluation using two open source datasets in BEIR

Model Name	max_seq_length	embedding_dimension
paraphrase-multilingual-MiniLM-L12-v2	128	384
distiluse-base-multilingual-cased-v1	128	512
paraphrase-multilingual-mpnet-base-v2	128	768
quora-distilbert-multilingual	128	768

Recall@K & MRR@K
1
3
5
10
100

Experimental Environment for Multilingual Models

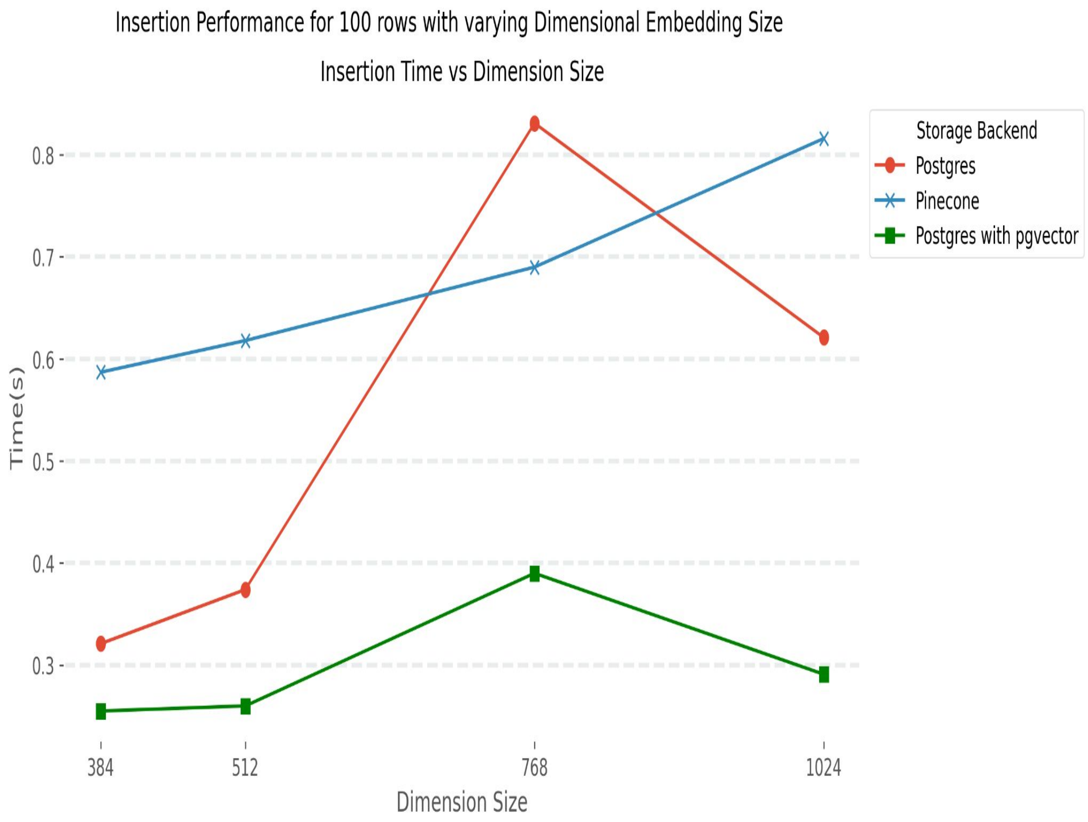
Machine Configuration

- Operating System Ubuntu 20.04.5 LTS (Focal Fossa)
- Processor Intel(R) Xeon(R) CPU @ 2.30GHz, 2300 Mhz, 2 Core(s), 4 Logical Processor(s)
- Memory 11.0 GB RAM
- Graphics NVIDIA A100-SXM (40 GB)
  - Driver Version: 525.85.12
  - CUDA Version: 12.0

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

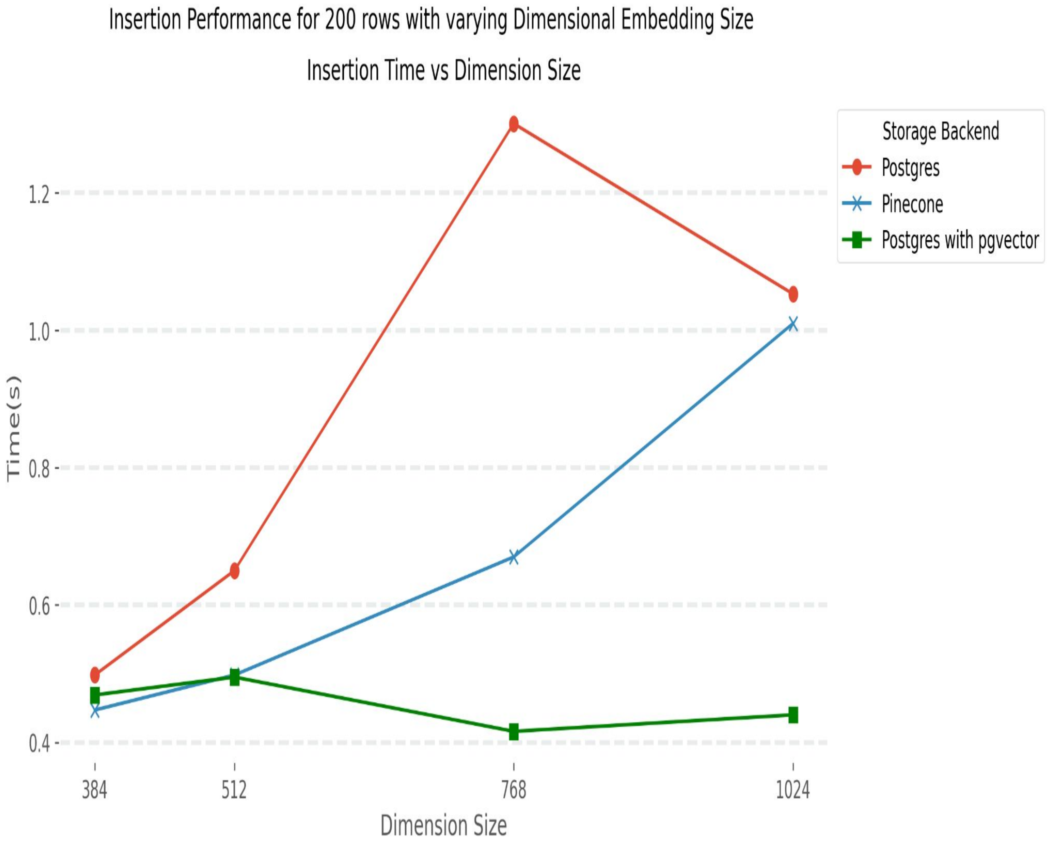
INF FAKULTÄT FÜR INFORMATIK

Results: Comparison of Insertion time for rows=100



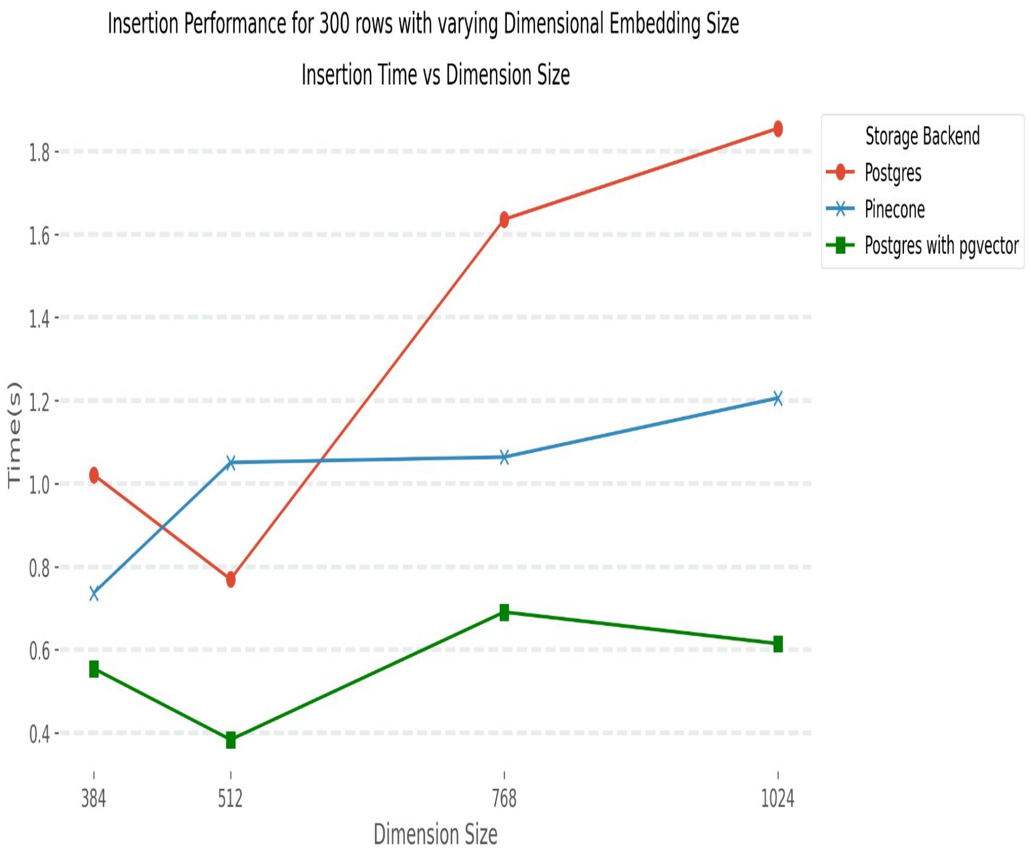
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

Results: Comparison of Insertion time for rows=200

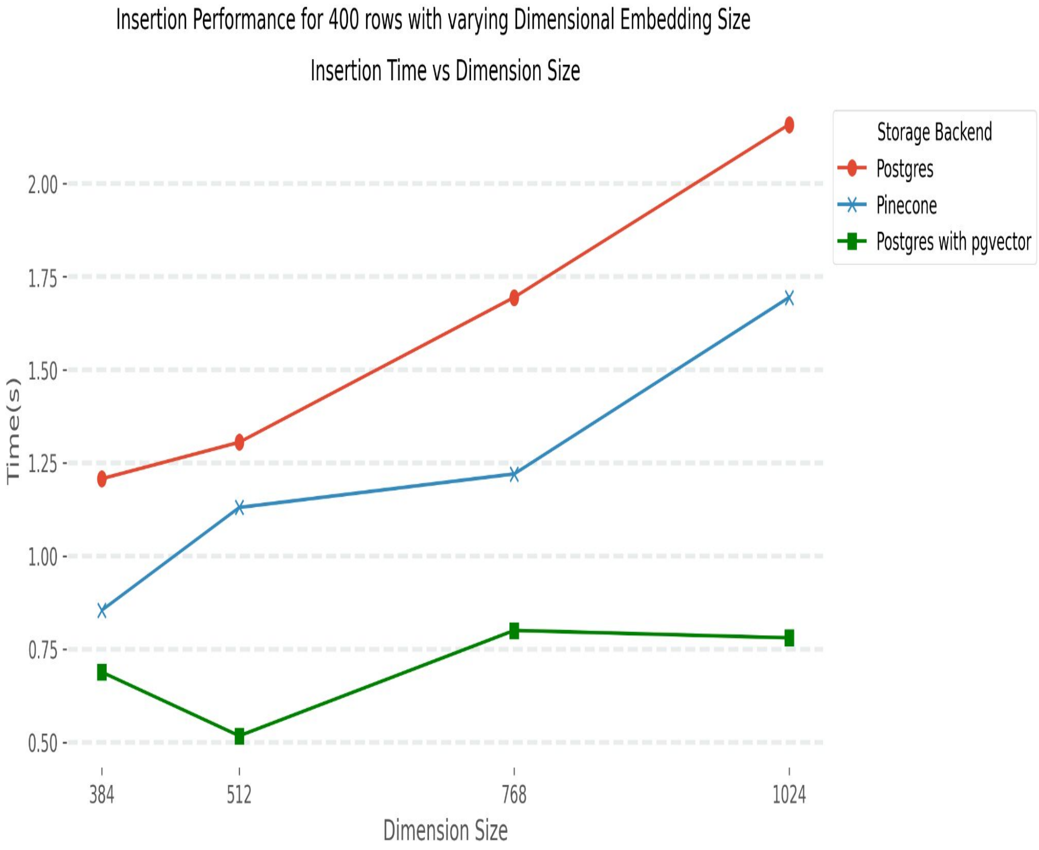


OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF FAKULTÄT FÜR INFORMATIK

Results: Comparison of Insertion time for rows=300

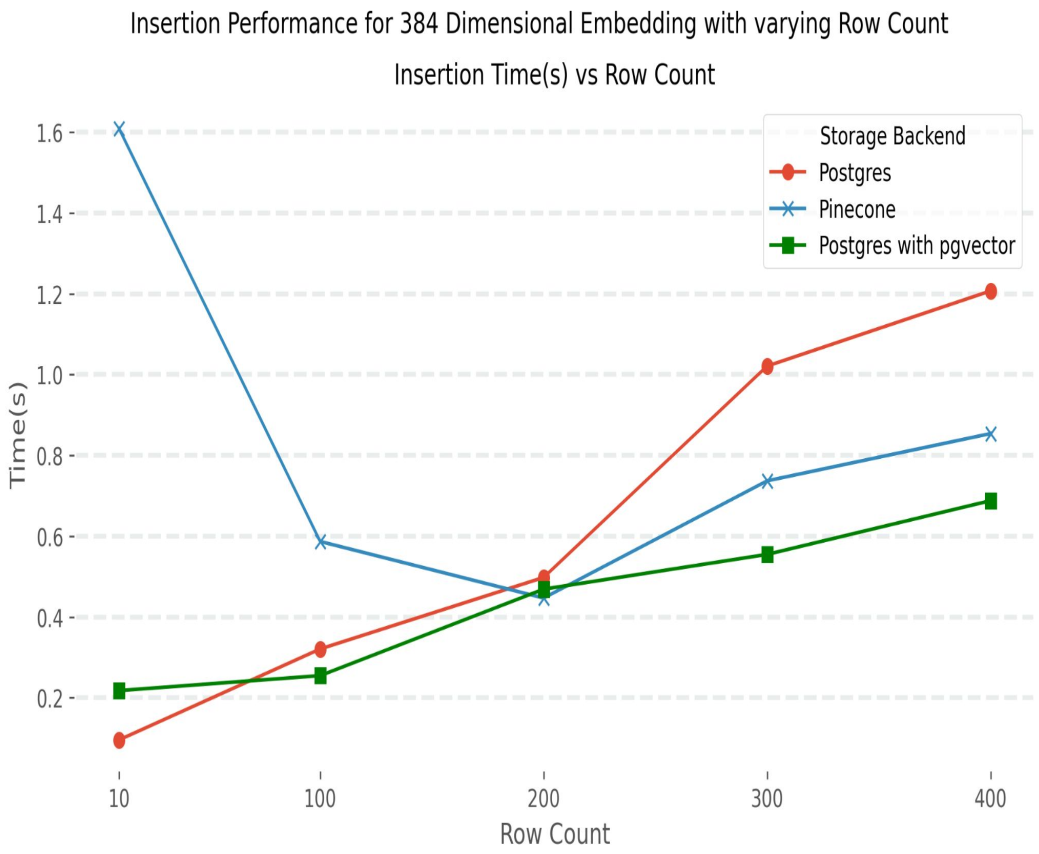


Results: Comparison of Insertion time for rows=400

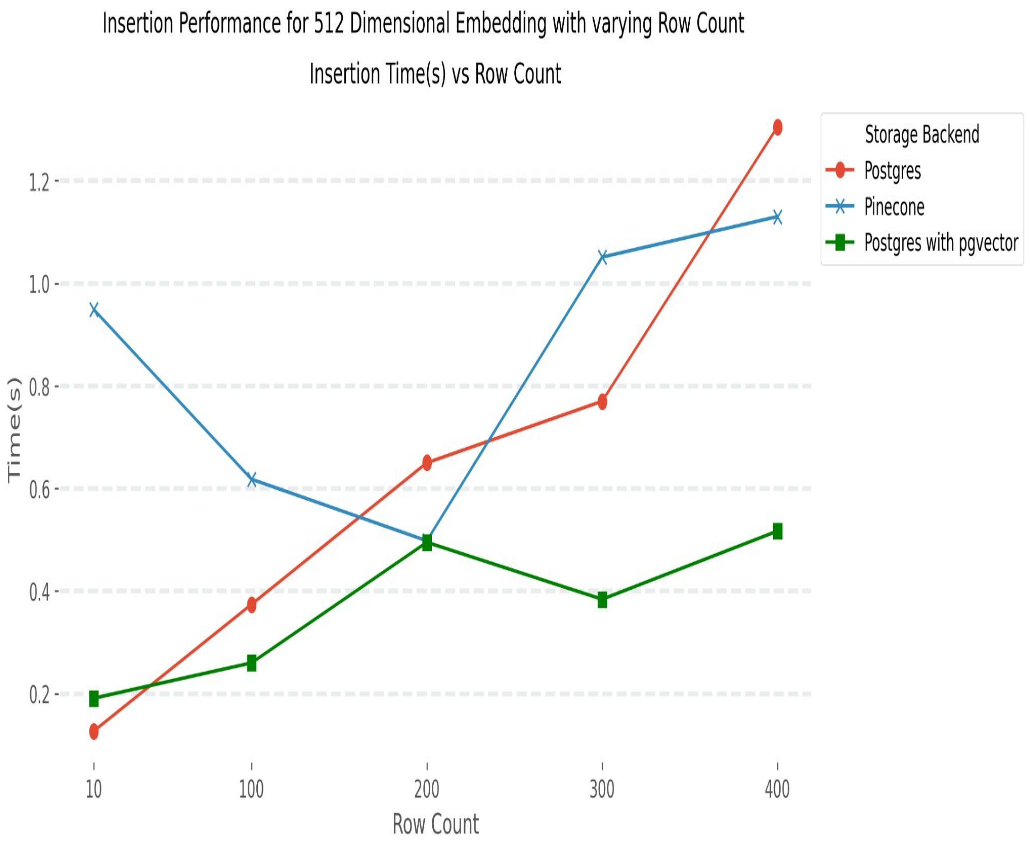


OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Results: Comparison of Insertion time for Embedding Size=384



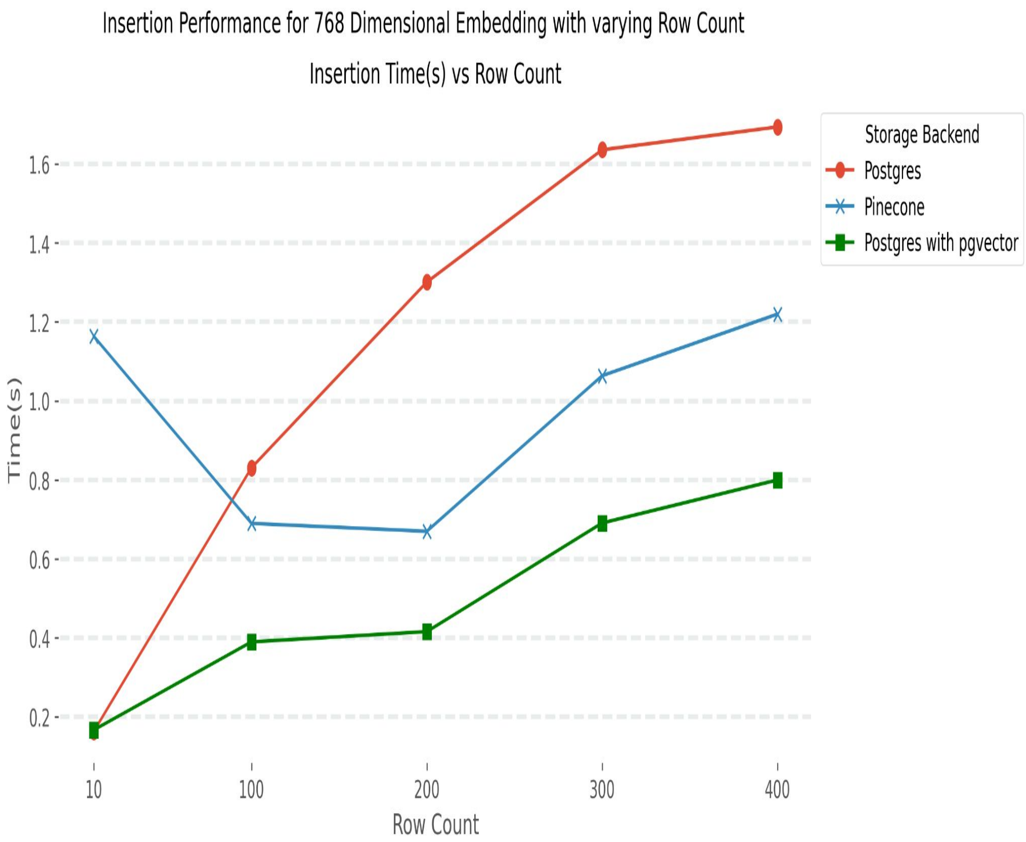
Results: Comparison of Insertion time for Embedding Size=512



OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

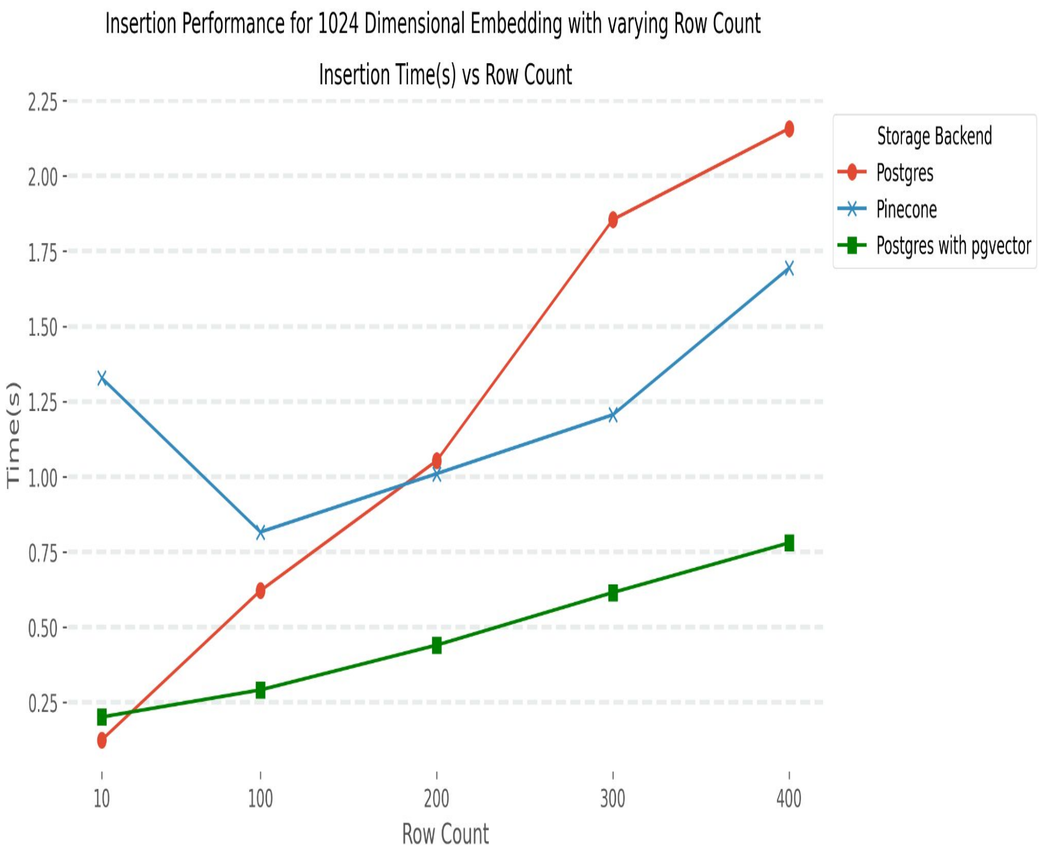
INF FAKULTÄT FÜR INFORMATIK

Results: Comparison of Insertion time for Embedding Size=768



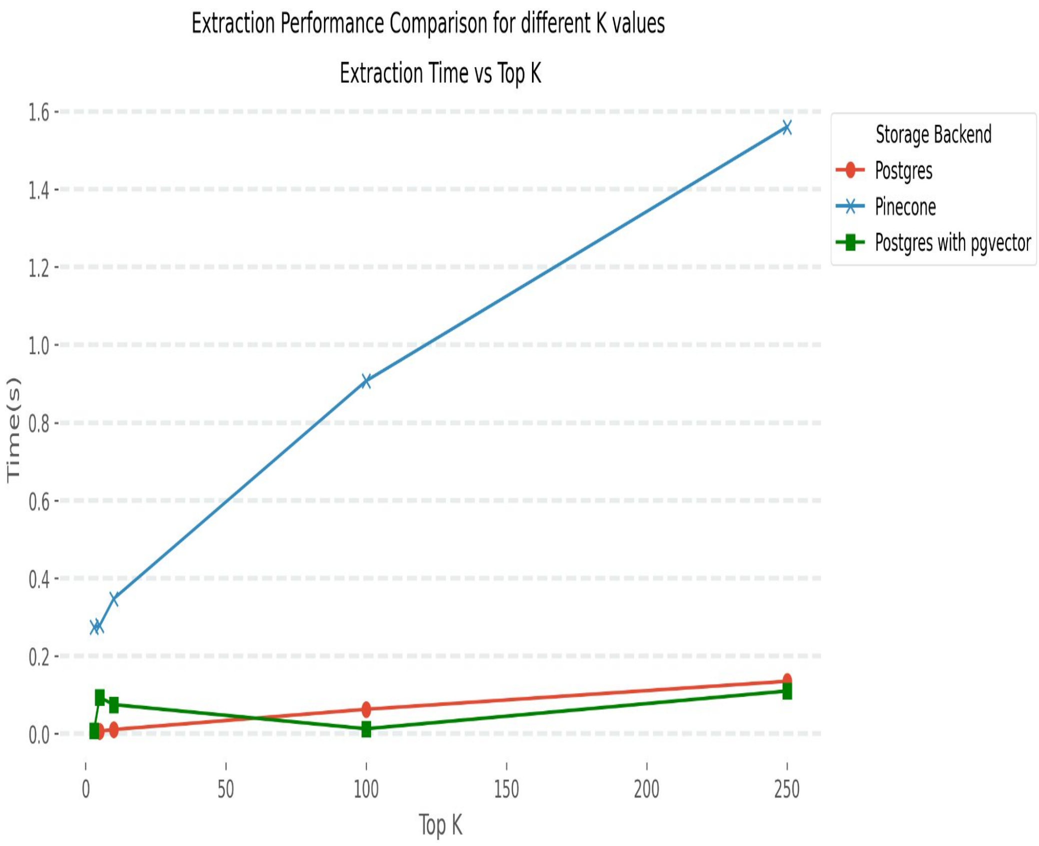
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Results: Comparison of Insertion time for Embedding Size=1024



OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG INF FAKULTÄT FÜR INFORMATIK

Results: Time taken for Retrieval for different values of K



Results: Performance Evaluation for Update Task

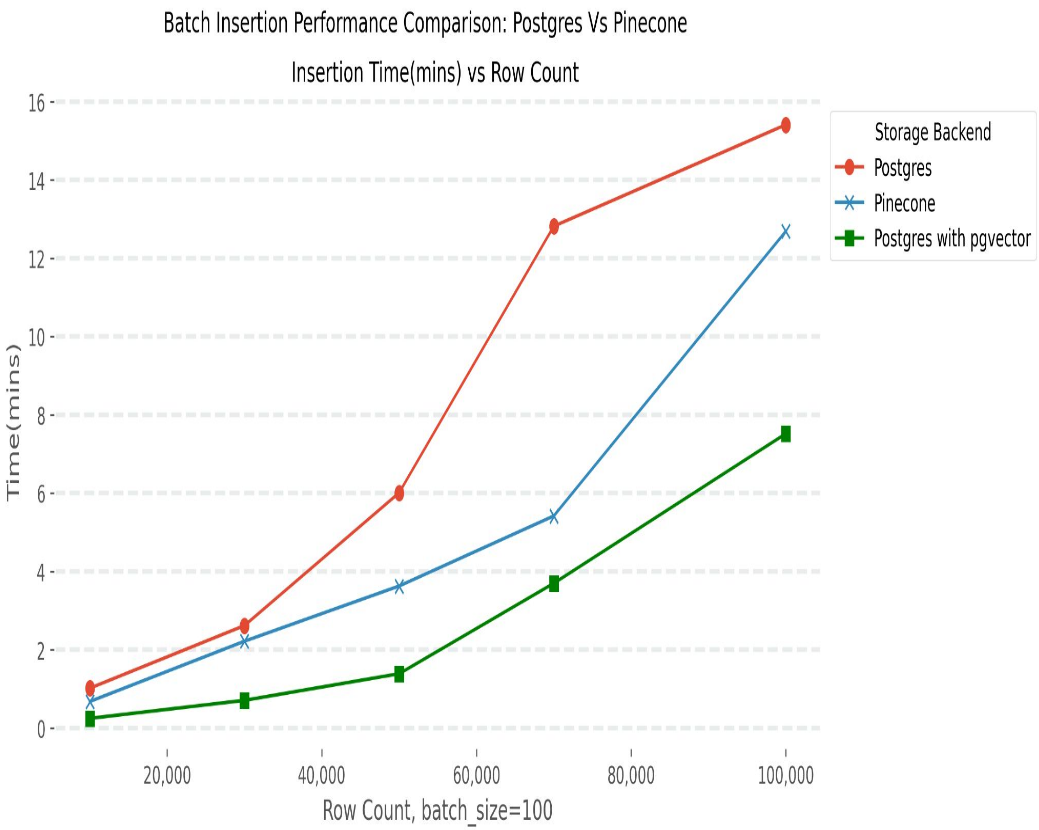
Database Name	Updation time(s)
PostgreSQL	0.026
Pinecone	0.142
PostgreSQL with pgvector	0.030

Results: Performance Evaluation for Delete Task

Database Name	Deletion time(s)
PostgreSQL	0.020
Pinecone	0.136
PostgreSQL with pgvector	0.023

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF | FAKULTÄT FÜR INFORMATIK

Results: Performance Evaluation for Batch Insertion Task



Results: Selection of Storage Backend

Results: Comparison of Multilingual Models

Results: Evaluation of Performance on BEIR Quora Dataset

Results: Evaluation of Performance on Germandpr-beir Dataset

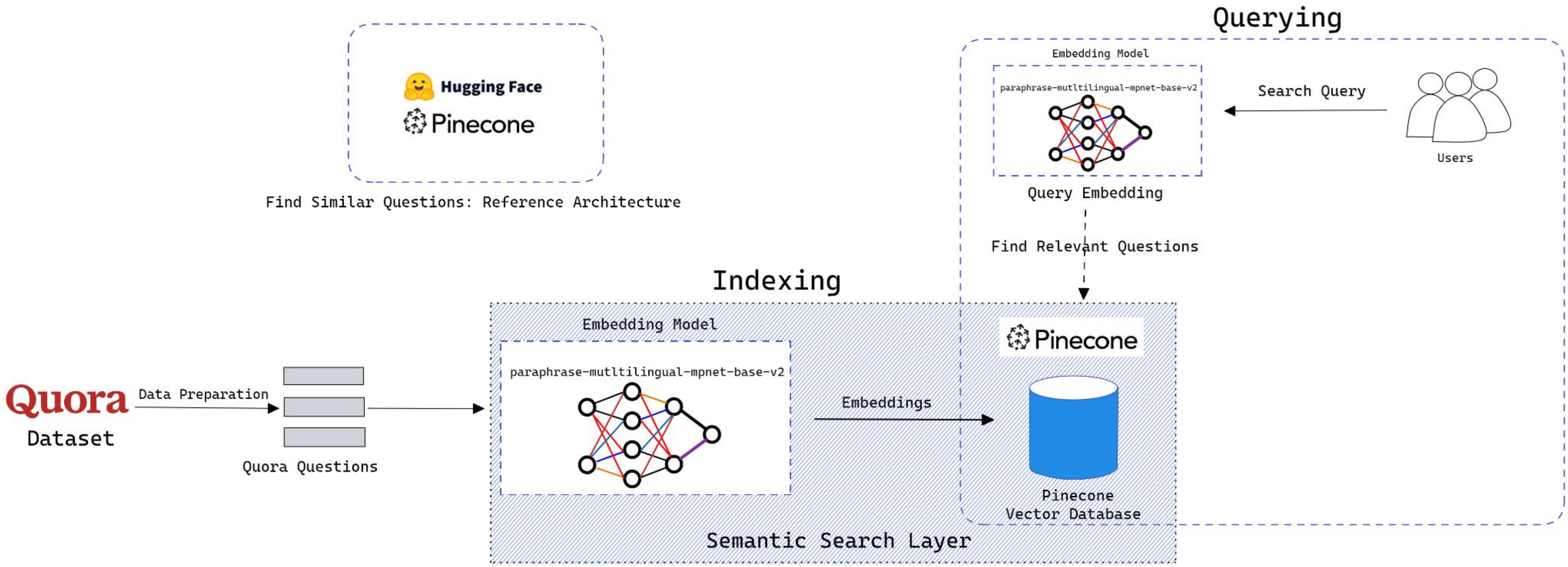
Results: Comparison of Inference Speed of Multilingual Models

Model Name	Output Dim.	Index Size (MB)	Avg Inference Time (ms)
MiniLM-L12-v2	384	153.60	17
distiluse-cased-v1	512	204.80	9
mpnet-base-v2	768	307.20	17
quora-distilbert	768	307.20	9

Selection of Best Model

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG | INF FAKULTÄT FÜR INFORMATIK

Prototype: Multilingual Semantic Search Application



Multilingual Semantic Search Application Architecture

Demo: https://huggingface.co/spaces/Ashish08/Multilingual-Search-Quora-Similar-Questions

Conclusion

- **Storage Backend Evaluation**
  - PostgreSQL **pgvector** extension showed superior performance over Pinecone
  - Overheads such as network and authentication might affect Pinecone's speed.
  - Pinecone version 2.0, while purpose-built for vector storage, still offers high search quality with speed.
- **Multilingual Model Performance**
  - **paraphrase-multilingual-MiniLM-L12-v2** and **paraphrase-multilingual-mpnet-base-v2** were s in the BEIR-Quora dataset.
  - **paraphrase-multilingual-mpnet-base-v2** excelled in the Germandpr-beir dataset.
  - Fastest models: **distiluse-base-multilingual-cased-v1** and **quora-distilbert-multilingual**, with 9 milliseconds for top-k results.

Future work

- Investigate Performance of other Vector Databases
- Comparison of Performance after Fine tuning models
- Explore the impact of varying vector indexes on search quality and speed

Thank You!