DevRev

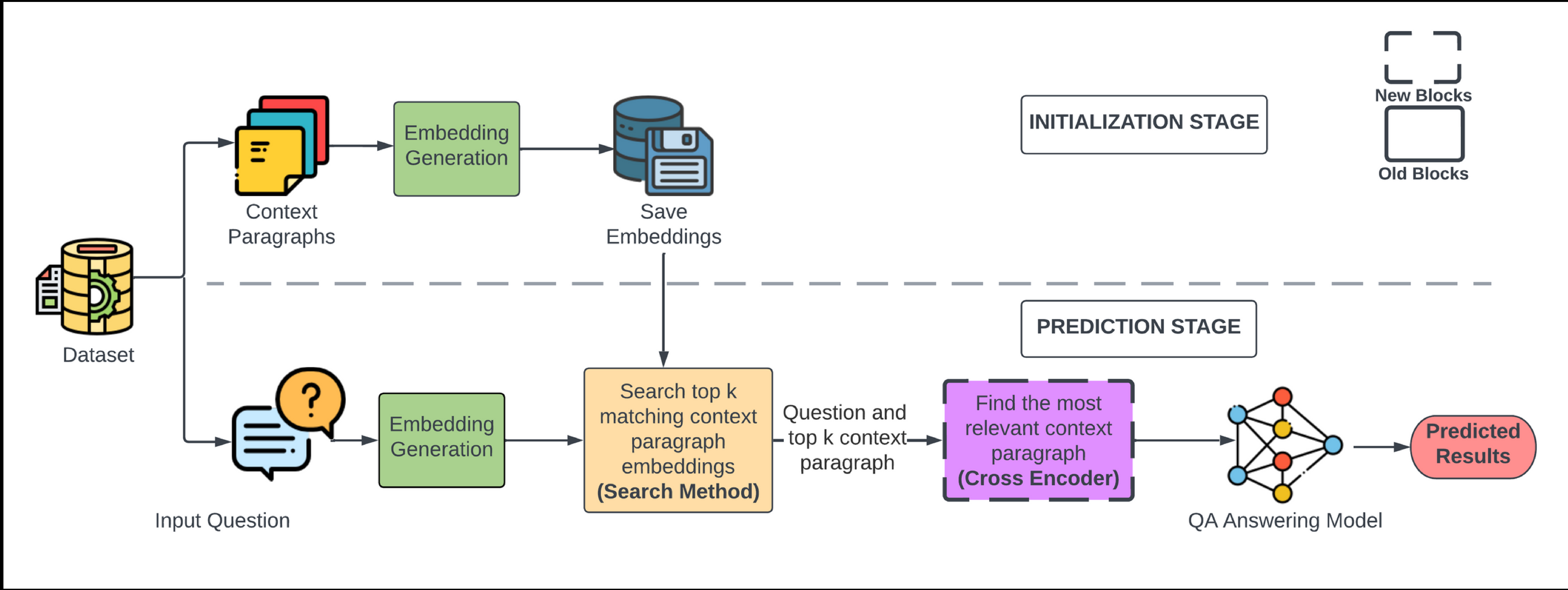
Expert Answers in a Flash: Improving Domain-Specific QA

Team 53

Agenda

- Solution Pipeline
- Para Retrieval (DeLADE)
- Para Ranker (Cross Encoder)
- Question Answering (QA) Model
- Synthetic Data Generation
- Theme Wise Analysis
- Caching Analysis
- Final Pipeline
- Future Opportunities

Solution Pipeline



Modules used

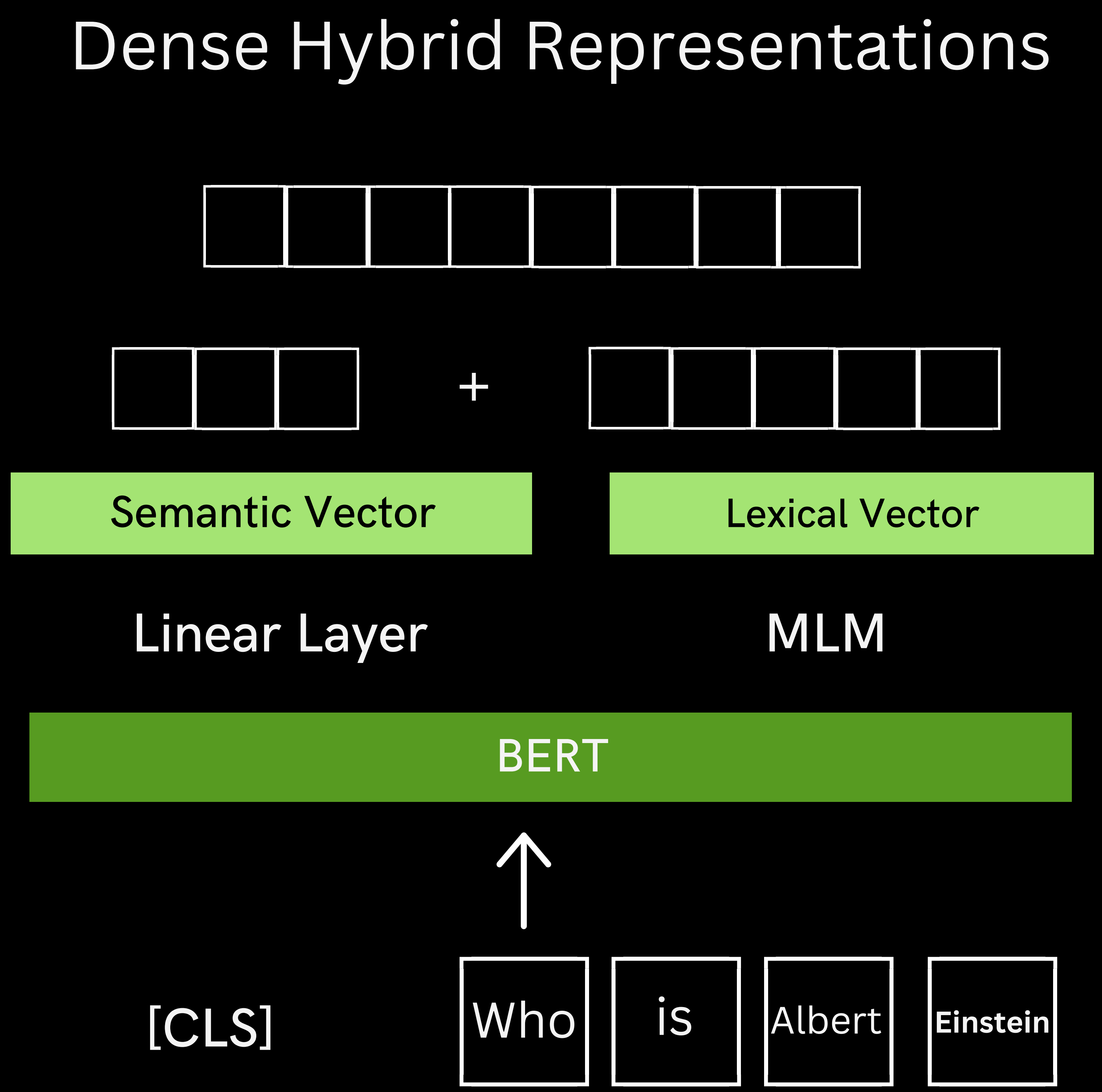
- Search Method → DeLADE [1]
- Paragraph Ranker → Cross-Encoder [2]
- QA model → ELECTRA-small [3]

Benefits offered

- High degree of modularity.
- Scales well with large data.
- Powered by a state-of-the-art QA model.

DeLADE-CLS

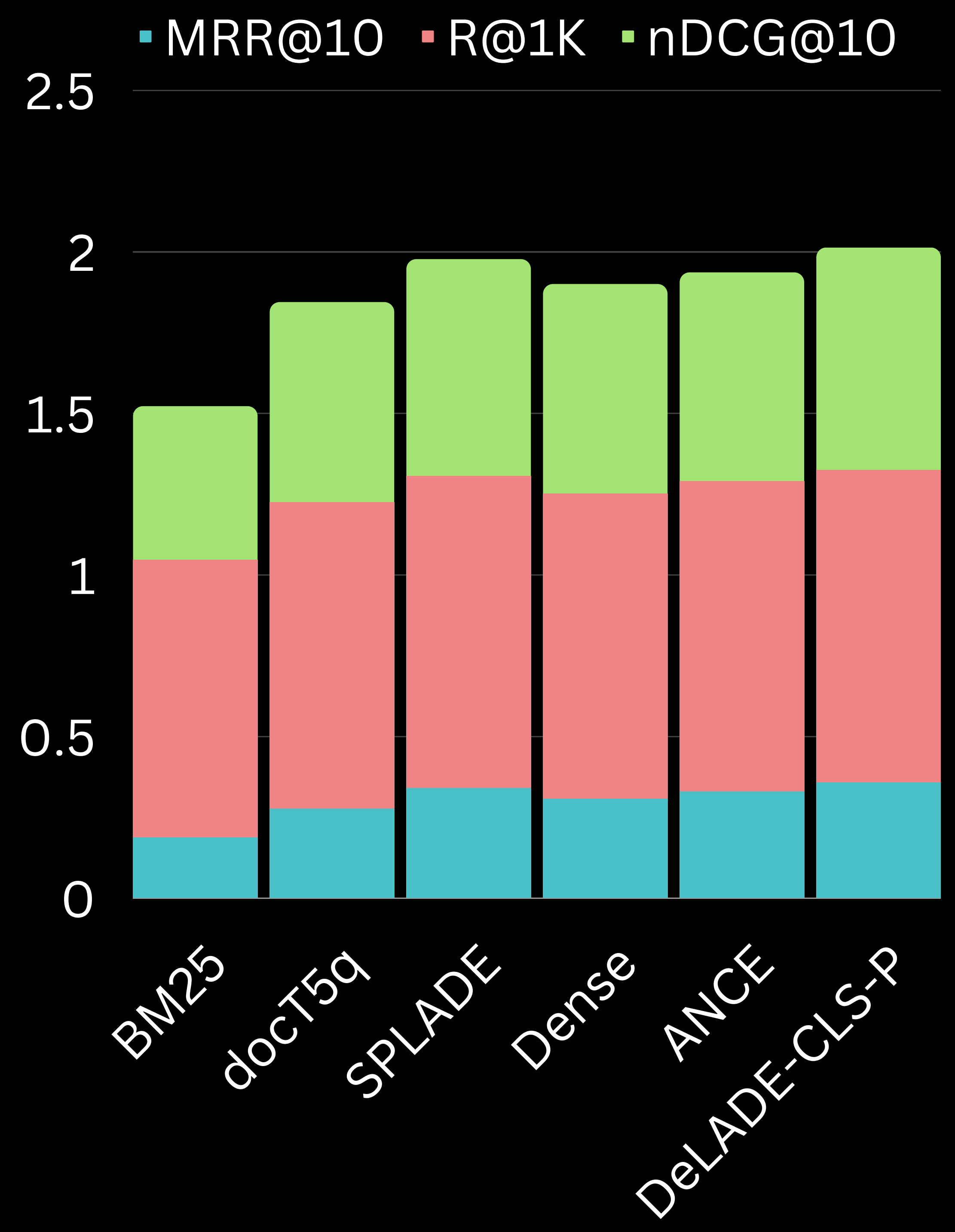
- Embedding generation technique for faster and effective passage retrieval.
- Uses DeLADE - A variant of SPLADE.
- Joint training of DLR and Semantic representations ([CLS] embeddings) to generate Dense Hybrid Representations.
- Pretrained on the MS MARCO dataset.



DeLADE-CLS vs Other Methods

Why DeLADE?

- Used to reduce search space towards finding the relevant context paragraph.
- Experiments showed greater accuracy and reduced latency speed of DeLADE-CLS-P with dot similarity over other search methods.
- Relative increase in both Para and QA scores by around 6% and 9% over vector-based search methods (ANNOY, FAISS, etc.)



Comparison on MS MARCO

DeLADE Improvements : Base or Finetuned?

- DeLADE returns top k most relevant paragraphs for a question.

- Para Accuracy

  => 1 : Any of the K paragraphs matches the target paragraph

  => 0 : Otherwise

- Fine tuning of DeLADE performed on the 'train' split (80%) and evaluated on 'val' split (10%) of the provided training dataset.

Base Model

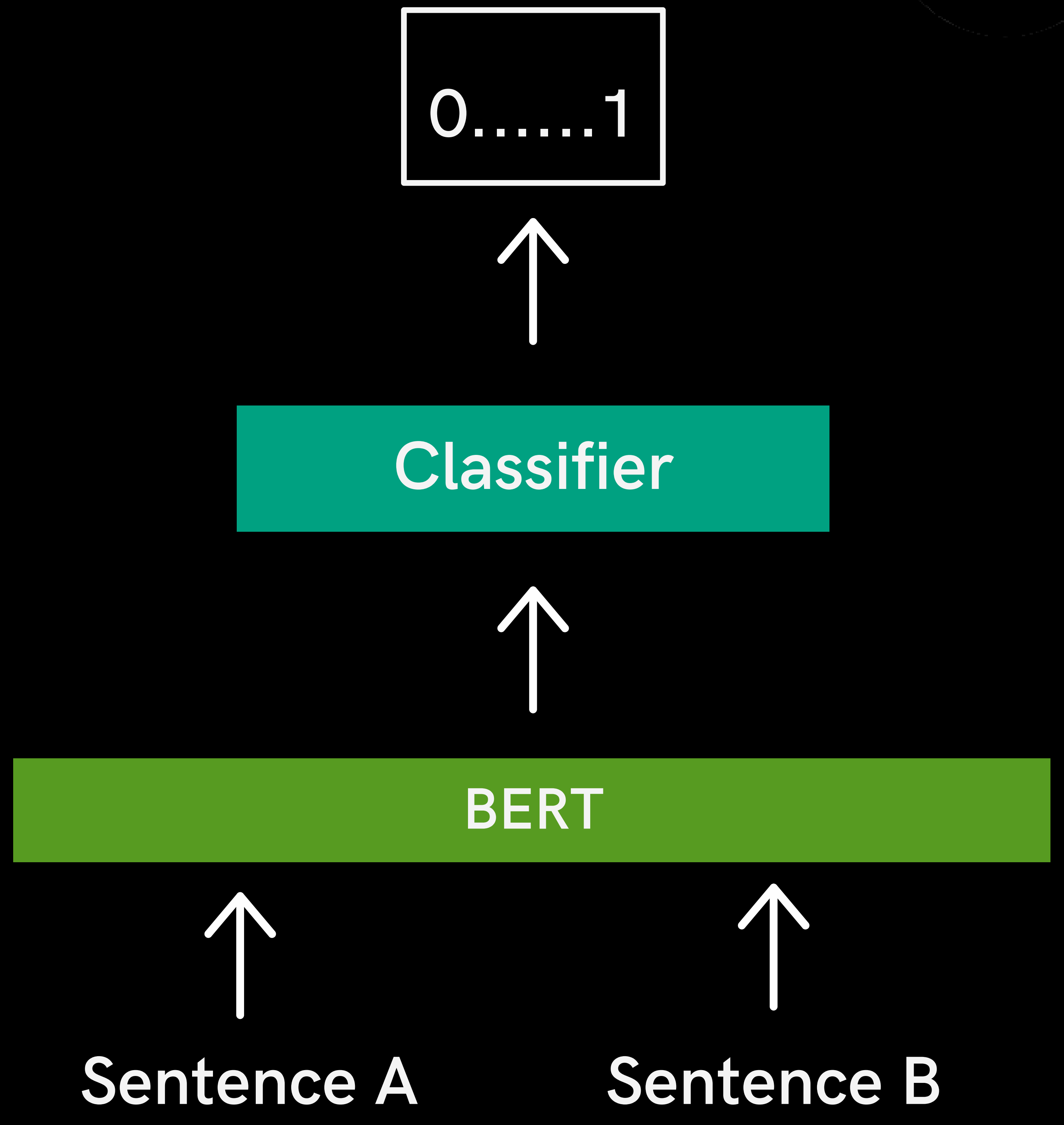
Para Accuracy:
99.68%

Best Finetuned Model

Para Accuracy:
96.03%
(10k Optimization steps)

Cross Encoder

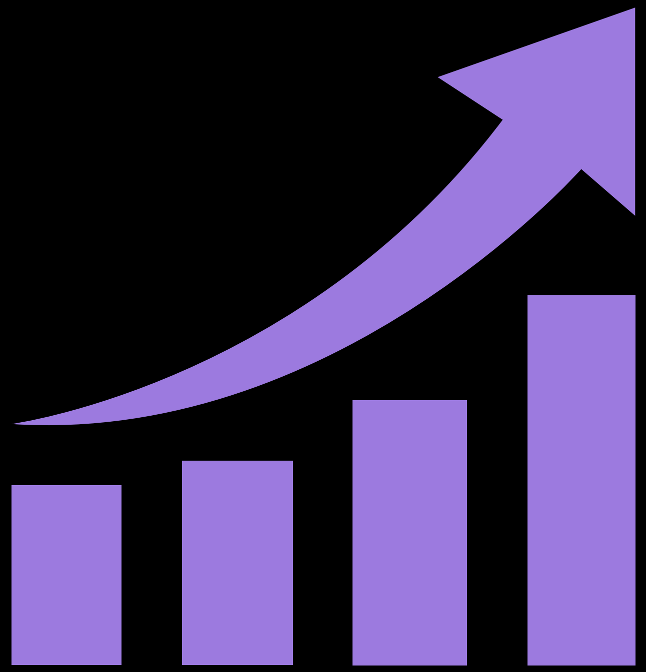
- Paragraph ranking module.
- Generates score based on semantic similiarity.
- Comes in pipeline right after DeLADE.
- Built on MiniLM architecture.
- Pretrained on MS MARCO.



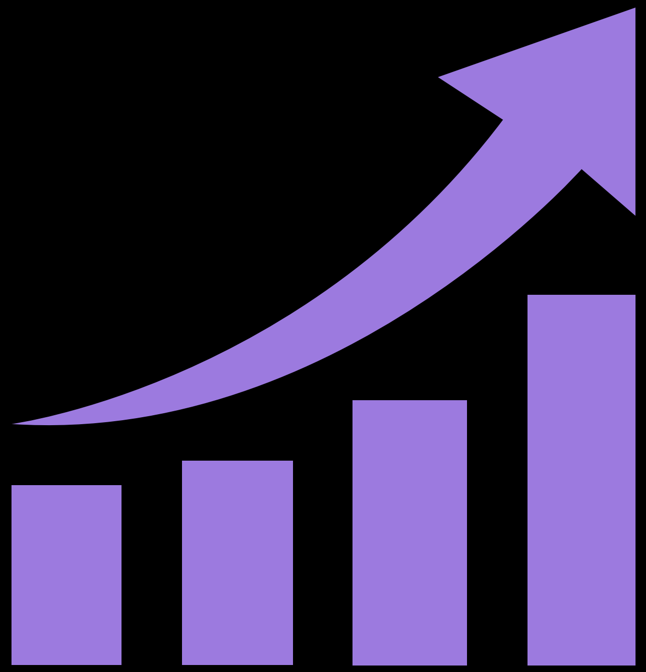
Why Cross Encoder?

- Use of cross-encoder leads to drastic increase in scores.

8.5 %
QA Score



9.5 %
Para Score



- Distinctive feature - Use of siamese and triplet networks.
- MiniLM-L4-v2 pretrained checkpoint found to give optimal performance.

Cross Encoder Experiments : Base or Finetuned?

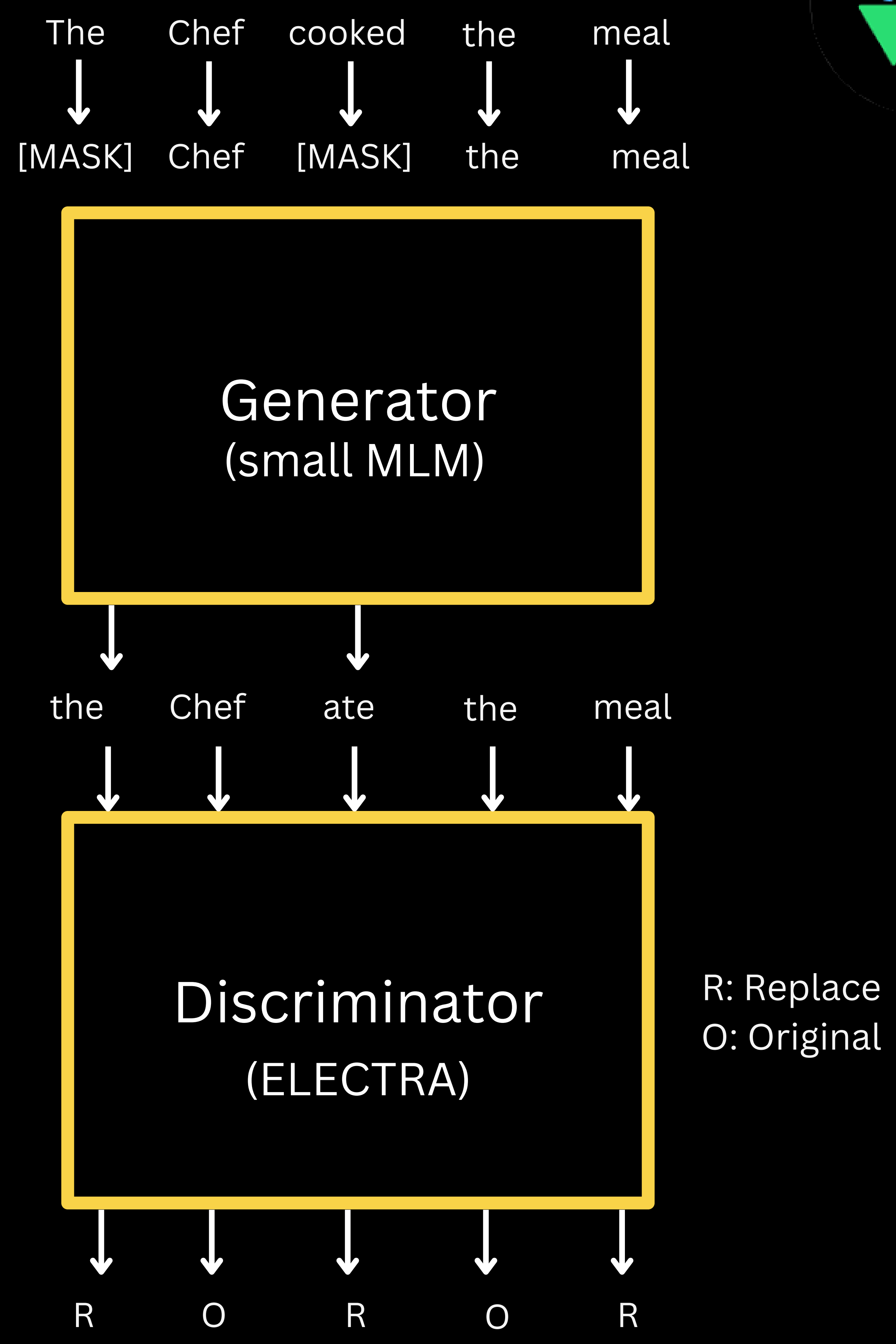
- Fine tuning of cross encoder performed on the 'train' split of the provided dataset, with results on the 'val' split.

- Most optimal performance - finetuned till 17 epochs.

Base Model	Best Finetuned Model
Para Accuracy: 90.3%	Para Accuracy: 95.1%

ELECTRA-small

- Based on replaced token detection.
- Requires less compute than MLM based learning.
- Has three variants - small, base and large, with the latter returning SoTA results on SQuAD v2.
- We use pretrained ELECTRA-small finetuned on SQuAD v2.



Why ELECTRA-small ?

Model	F1
MiniLM	91.5%
SqueezeBERT	93.2%
TinyBERT	92.5%

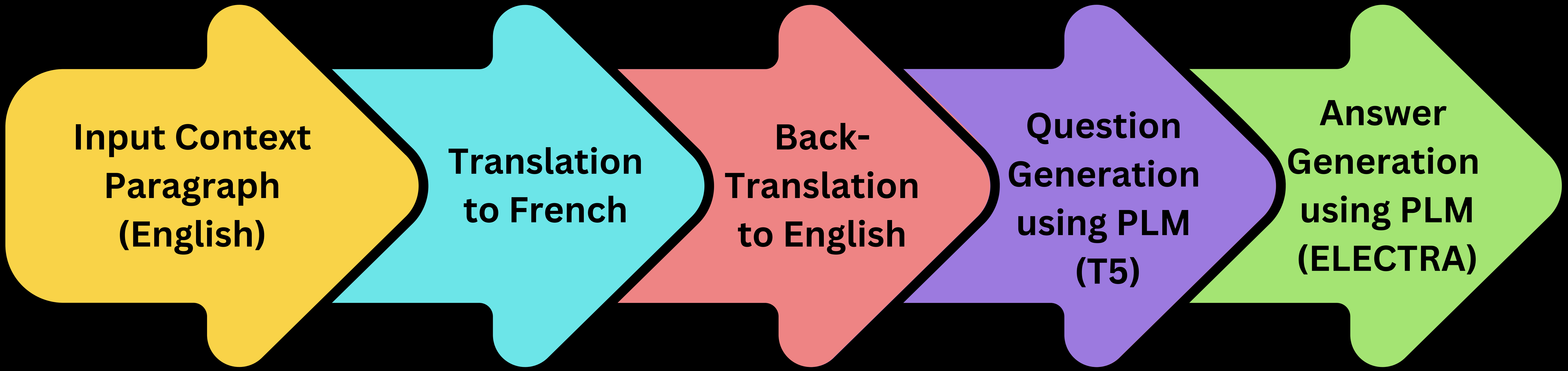
ELECTRA-small
94.5%

ELECTRA-small improvements : Base or Finetuned?

Fine Tuning of ELECTRA-small was performed on 'train' split of the provided training dataset, with evaluation done on the 'val' split.

Best Finetuned Model (6 Epochs)	Base Model
F1: 91.0%	F1: 92.8%

Synthetic QA Data Generation



Experiments using Synthetic Data

- Fine Tuning of ELECTRA-small was done on synthetic data.
- Evaluation performed on 10% 'test' split of training data.

Base Model	Best Finetuned Model
F1: 93.1%	F1: 74.8%

Theme wise Analysis

- Theme wise F1 score found on all themes from 'test' split.
- Finetuning performed on bottom 16 worst themes.

	Best Themes	Worst Themes
Before Finetuning	100%	70%
After finetuning	97%	72%

Caching Analysis

DHRs of previously answered question

Cosine sim

DHR of input question

> threshold

< threshold

Direct assignment of target paragraph

Complete pass through pipeline

Without Caching	With Caching
QA: 91.5%	QA: 90.1%
Para: 95.1%	Para: 93.5%

Final Pipeline

- F1 = 92.8%
- Para Score = 95.1%
- QA Score = 91.5%
- Mean inference time

  per question = 767.0ms

  (reported on 10% 'test' split of provided training data)

Search Method:

DeLADE-CLS-P (base)

Paragraph Ranker:

Cross-encoder (finetuned)

QA Model:

ELECTRA-small (base)

Future Opportunities

- Large-data oriented improvements in the caching process.
- Use of Knowledge Graphs as an alternative to traditional seq2seq text modelling.
- Use of rule-based algorithms for answer verification.
- Optimization of large QA models (ATLAS, Retro-Reader) via training-based approaches such as distillation

Bibliography

1. Liu, Ye, et al. "Dense hierarchical retrieval for open-domain question answering." arXiv preprint arXiv:2110.15439 (2021).

2. Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

3. Clark, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555 (2020).

Thank You