Group 14

DPR Basemodel vs. Finetuned

Adrian Bamberger, 11701452 — Data selection, Model Training
Valentin Forster, 11823171 — Data Processing, Model Training
Anukun Thurner, 1048105 — Evaluation, Plotting

https://github.com/ValentinForster/AIRGroup14

Research Question

“How does a fine-tuned Dense Passage Retriever trained on German data compare to a pre-trained model for answering German questions?”

Methodology

Download the datasets

- GermanQuAD
- GermanDPR

Data needs to be in specific format

Conversion scripts:

- JSON Conversion to haystack format.py
- Unicode encoding replacement.py

GermanQuAD_train.json (22,4 MB)

GermanDPR_train.json (54 MB)

Evaluation on

- test dataset (respectively)
- train dataset (respectively)
- GermanDPR_test

Datasets

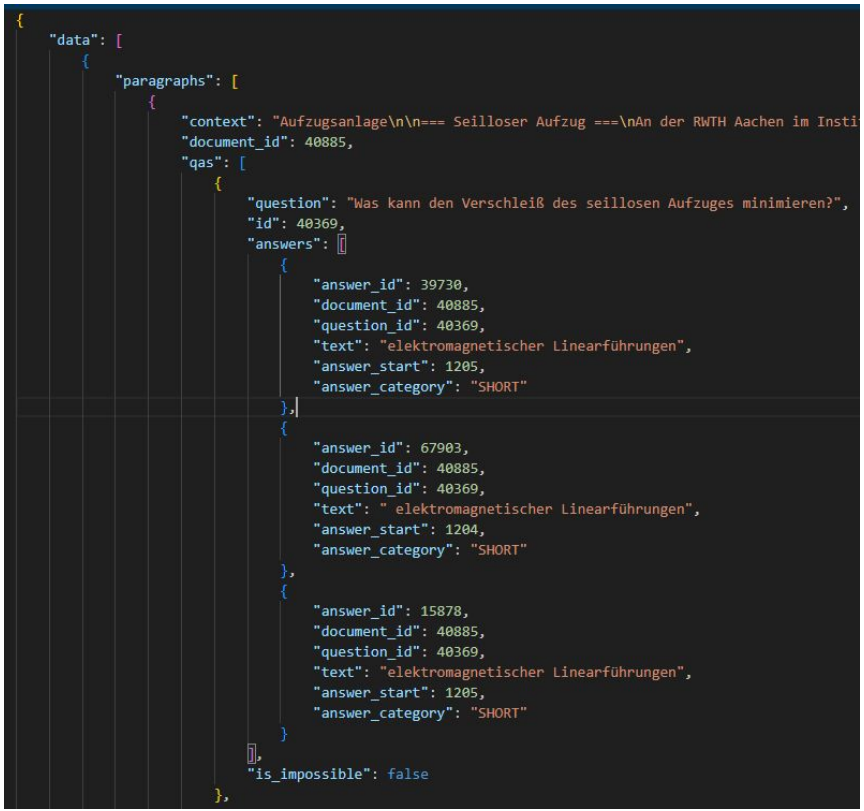
GermanQuAD

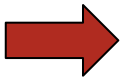
- Question Answering Dataset
- human-labeled German QA dataset
- consisting of 13,722 questions

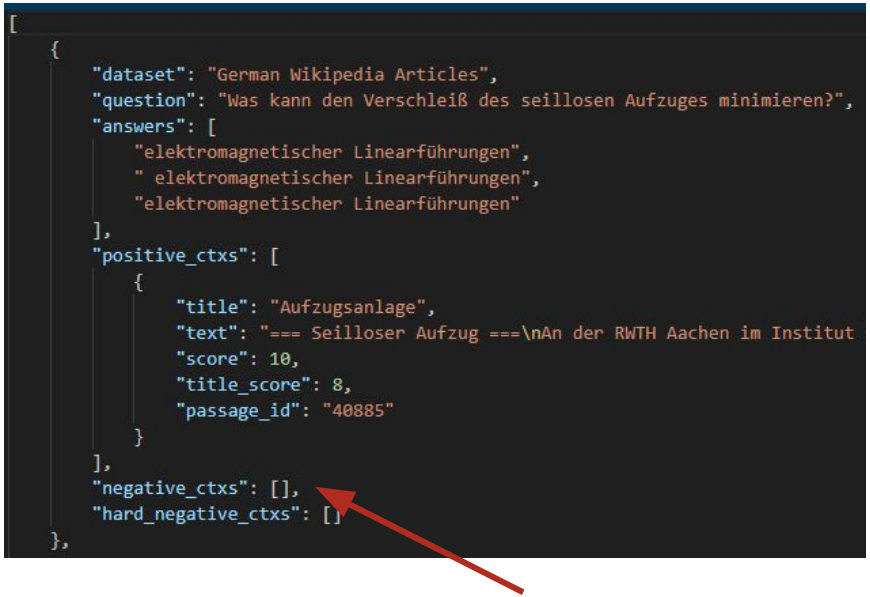
GermanDPR

- Passage Retrieval Dataset
- GermanQuAD as a starting point + hard negatives from a dump of the full German Wikipedia

GermanQuAD processing

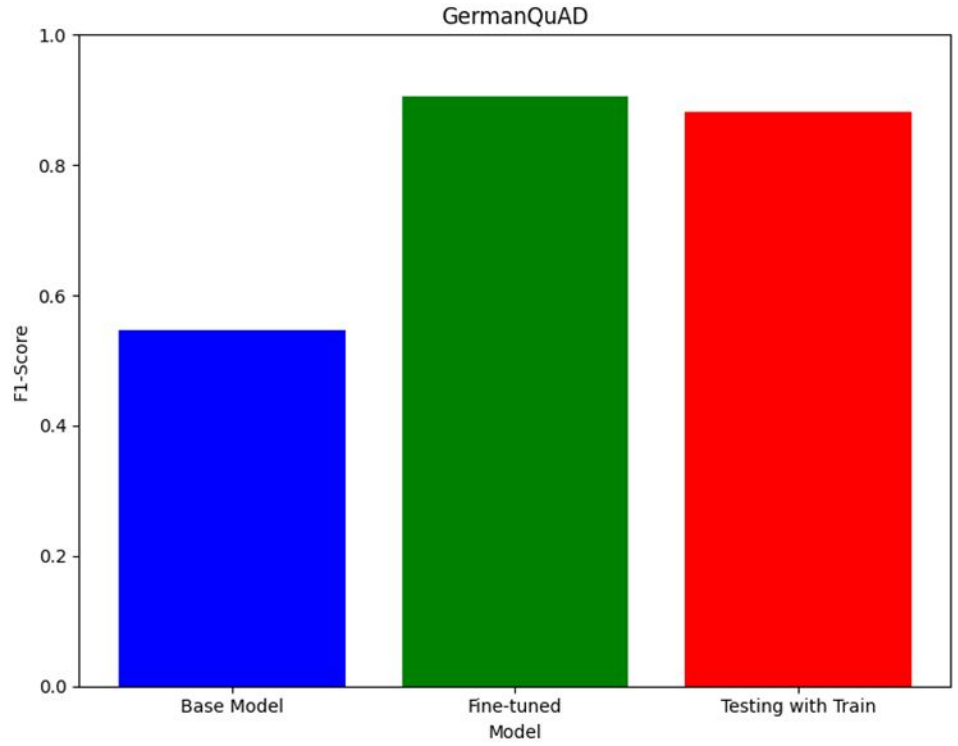


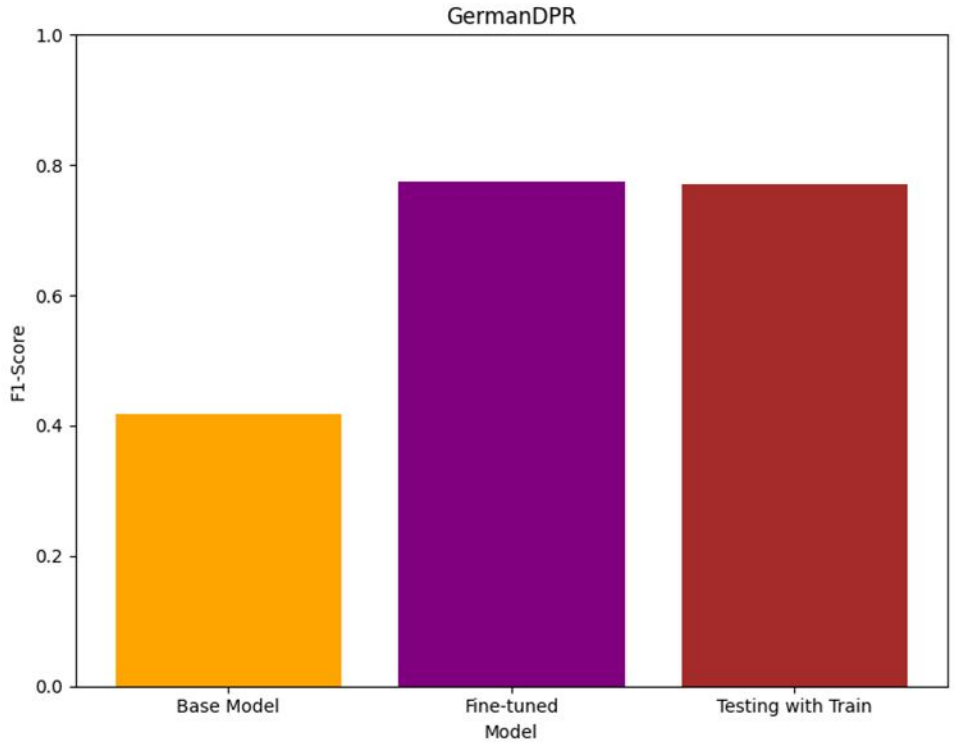




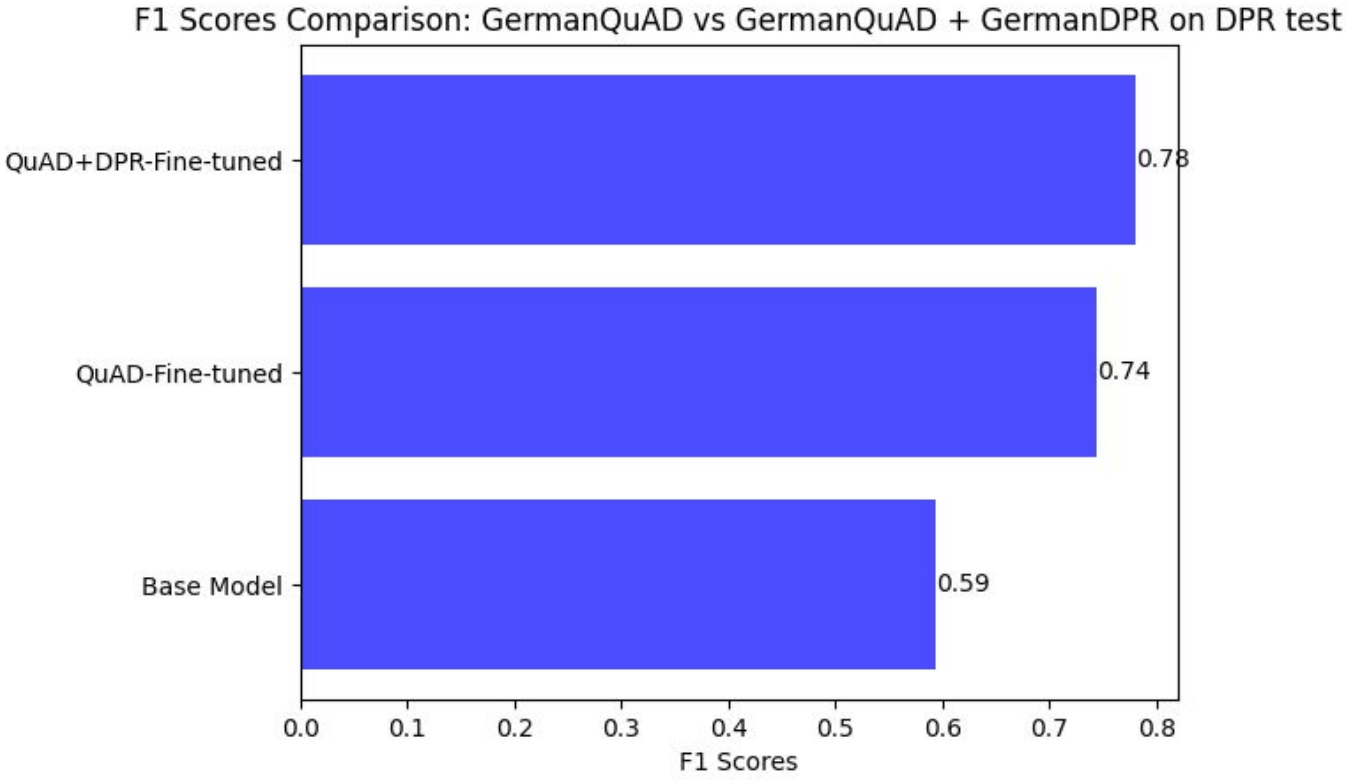
Using in-batch negatives

Evaluation (separated datasets)





Result Comparison (tested on DPR_test)



Conclusion

Fine-tuning is Key:

- Significant improvement in F1 score for both GermaQuAD and GermanDPR datasets after fine-tuning.

Generalization over Overfitting:

- Slight decrease in F1 score when tested with training data suggests good model generalization.

Consistent Across Datasets:

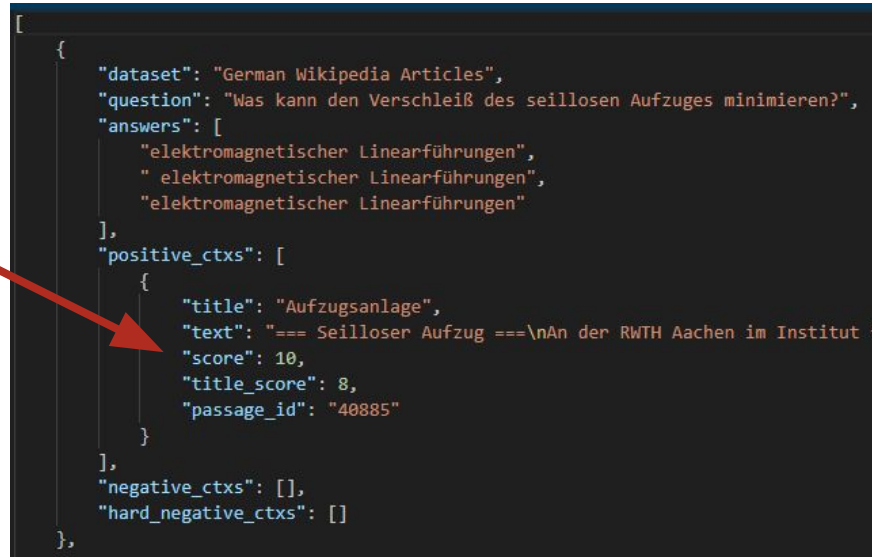
- Fine-tuning shows consistent benefits across different datasets.

Model Robustness:

- The model maintains high performance on training data, indicating robustness against overfitting.

Limitations

Arbitrary numbers (10 and 8) as scores and title scores of each context in GermanQuAD



Further limitations

- Possible errors in datasets (e.g. Translation errors)

- GermandQuAD training used in-batch-negatives instead of similar but wrong contexts

- Base-model originally trained on English data
- Model is likely worse in English after fine tuning → Catastrophic interference

Any questions?

Datasets
https://www.deepset.ai/germanquad

Pre-Trained Model (Context Encoder)
https://huggingface.co/facebook/dpr-ctx_encoder-single-nq-base

Pre-Trained Model (Question Encoder)
https://huggingface.co/facebook/dpr-question_encoder-single-nq-base

GitHub
https://github.com/ValentinForster/AIRGroup14