

# 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

## Fetching

Download the datasets

- GermanQuAD
- GermanDPR

## Data Processing

Data needs to be in specific format

Conversion scripts:

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

## Training

GermanQuAD\_train.json  
(22,4 MB)

GermanDPR\_train.json  
(54 MB)

## Evaluation

Evaluation on

- test dataset (respectively)
- train dataset (respectively)
- GermanDPR\_test

# Datasets

## GermanQuAD

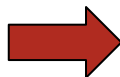
- Question Answering Dataset
- human-labeled German QA dataset
- consisting of 13,722 questions

## GermanDPR

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

# GermanQuAD processing

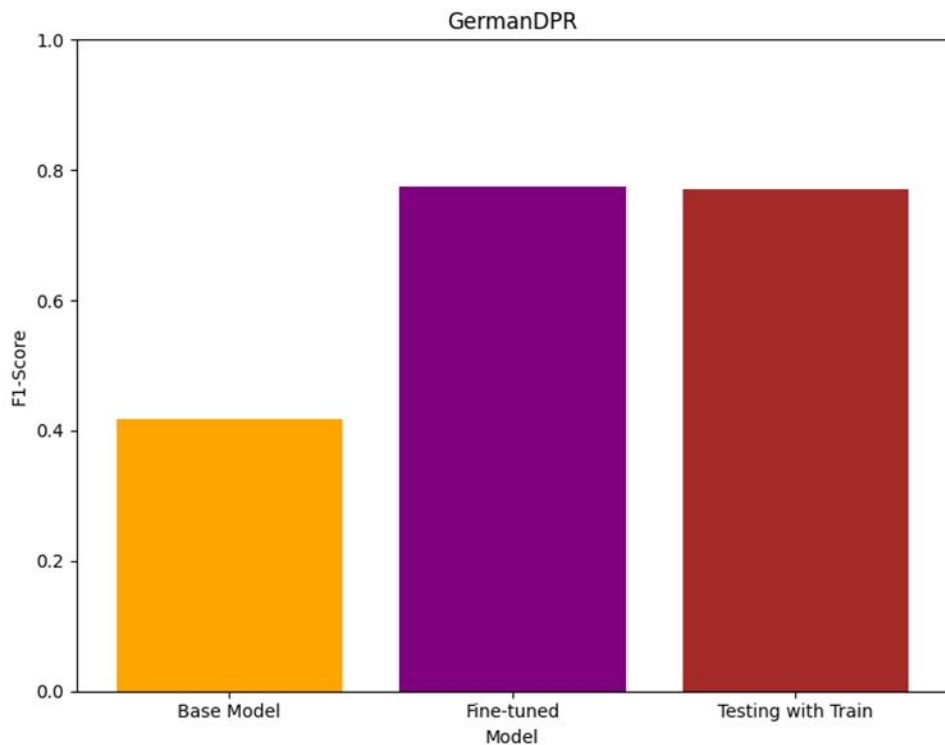
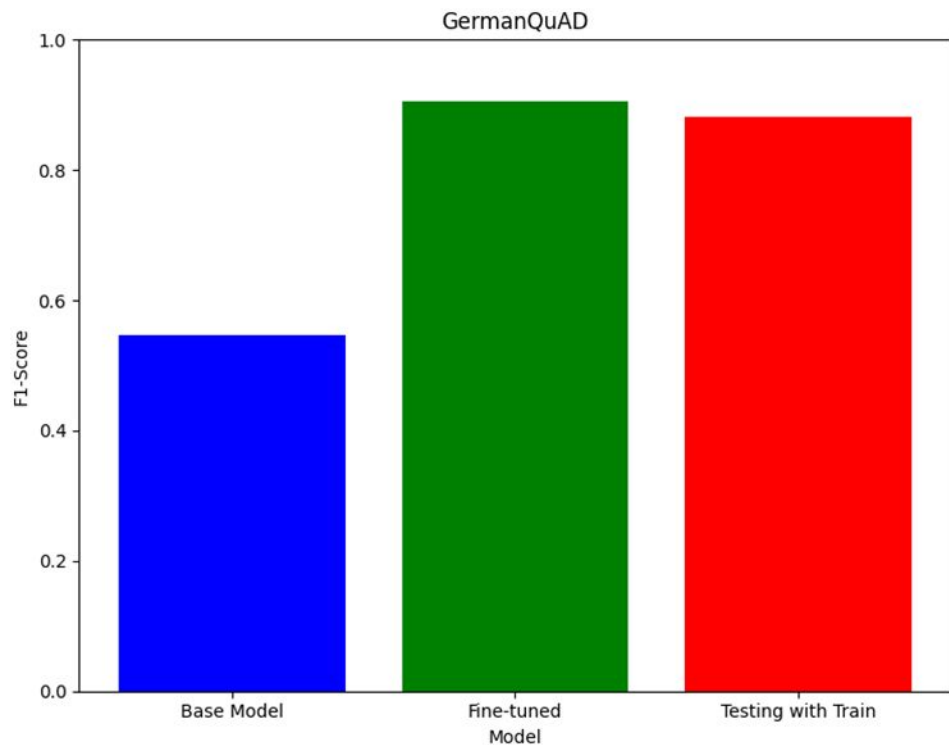
```
{
  "data": [
    {
      "paragraphs": [
        {
          "context": "Aufzugsanlage\n\n=== Seilloser Aufzug ===\nAn der RWTH Aachen im Insti",
          "document_id": 40885,
          "qas": [
            {
              "question": "Was kann den Verschleiß des seillosen Aufzuges minimieren?",
              "id": 40369,
              "answers": [
                {
                  "answer_id": 39730,
                  "document_id": 40885,
                  "question_id": 40369,
                  "text": "elektromagnetischer Linearführungen",
                  "answer_start": 1205,
                  "answer_category": "SHORT"
                },
                {
                  "answer_id": 67903,
                  "document_id": 40885,
                  "question_id": 40369,
                  "text": " elektromagnetischer Linearführungen",
                  "answer_start": 1204,
                  "answer_category": "SHORT"
                },
                {
                  "answer_id": 15878,
                  "document_id": 40885,
                  "question_id": 40369,
                  "text": "elektromagnetischer Linearführungen",
                  "answer_start": 1205,
                  "answer_category": "SHORT"
                }
              ]
            }
          ]
        }
      ]
    },
    {
      "is_impossible": false
    }
  ]
}
```



```
[
  {
    "dataset": "German Wikipedia Articles",
    "question": "Was kann den Verschleiß des seillosen Aufzuges minimieren?",
    "answers": [
      "elektromagnetischer Linearführungen",
      " elektromagnetischer Linearführungen",
      "elektromagnetischer Linearführungen"
    ],
    "positive_ctxs": [
      {
        "title": "Aufzugsanlage",
        "text": "=== Seilloser Aufzug ===\nAn der RWTH Aachen im Institut",
        "score": 10,
        "title_score": 8,
        "passage_id": "40885"
      }
    ],
    "negative_ctxs": [],
    "hard_negative_ctxs": []
  }
]
```

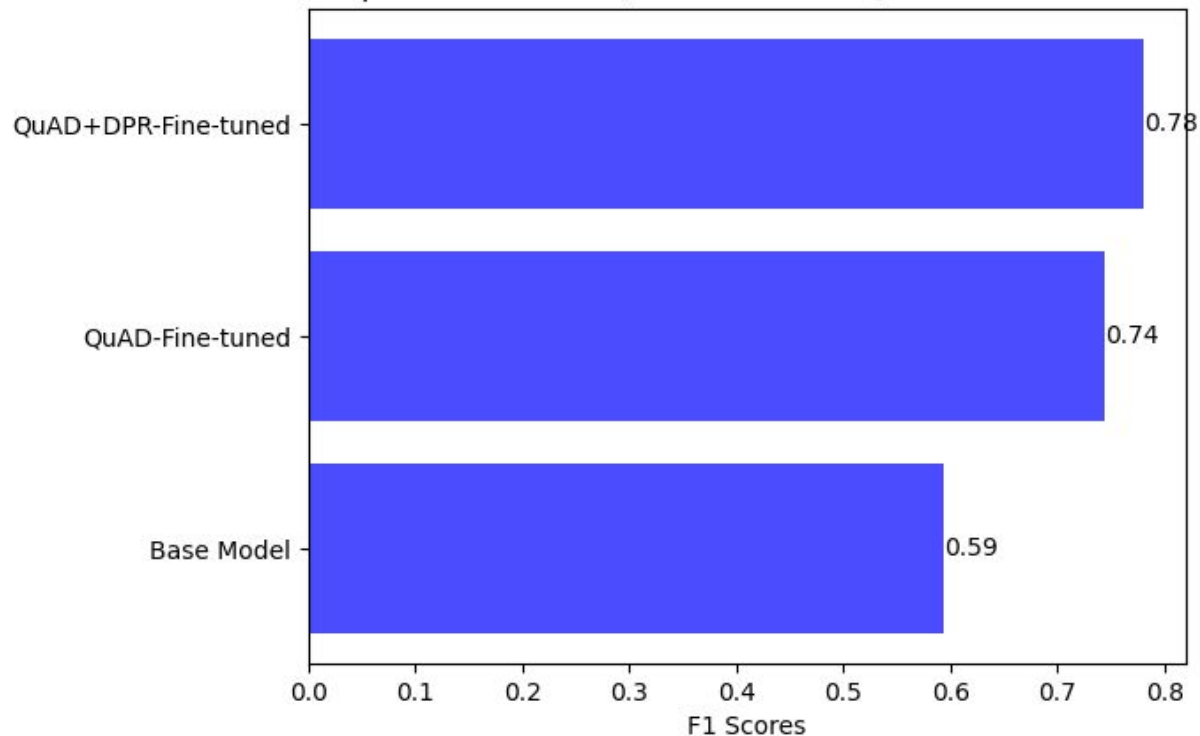
Using in-batch negatives

# Evaluation (separated datasets)



# Result Comparison (tested on DPR\_test)

F1 Scores Comparison: GermanQuAD vs GermanQuAD + GermanDPR 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.

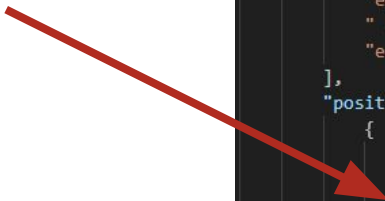
## Model Robustness:

- 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



```
[
  {
    "dataset": "German Wikipedia Articles",
    "question": "Was kann den Verschleiß des seillosen Aufzuges minimieren?",
    "answers": [
      "elektromagnetischer Linearführungen",
      " elektromagnetischer Linearführungen",
      "elektromagnetischer Linearführungen"
    ],
    "positive_ctxs": [
      {
        "title": "Aufzugsanlage",
        "text": "=== Seillosen Aufzug ===\nAn der RWTH Aachen im Institut",
        "score": 10,
        "title_score": 8,
        "passage_id": "40885"
      }
    ],
    "negative_ctxs": [],
    "hard_negative_ctxs": []
  },
]
```

# Further limitations

- Possible errors in datasets (e.g. Translation errors)
- GermanQuAD training used in-batch-negatives instead of similar but incorrect 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](https://www.deepset.ai/germanquad)

## Pre-Trained Model (Context Encoder)

[https://huggingface.co/facebook/  
dpr-ctx\\_encoder-single-nq-base](https://huggingface.co/facebook/dpr-ctx_encoder-single-nq-base)

## Pre-Trained Model (Question Encoder)

[https://huggingface.co/facebook/  
dpr-question\\_encoder-single-n  
q-base](https://huggingface.co/facebook/dpr-question_encoder-single-nq-base)

## GitHub

[https://github.com/Vale  
ntinForster/AIRGroup14](https://github.com/ValentinForster/AIRGroup14)