WISSEN
TECHNIK
LEIDENSCHAFT

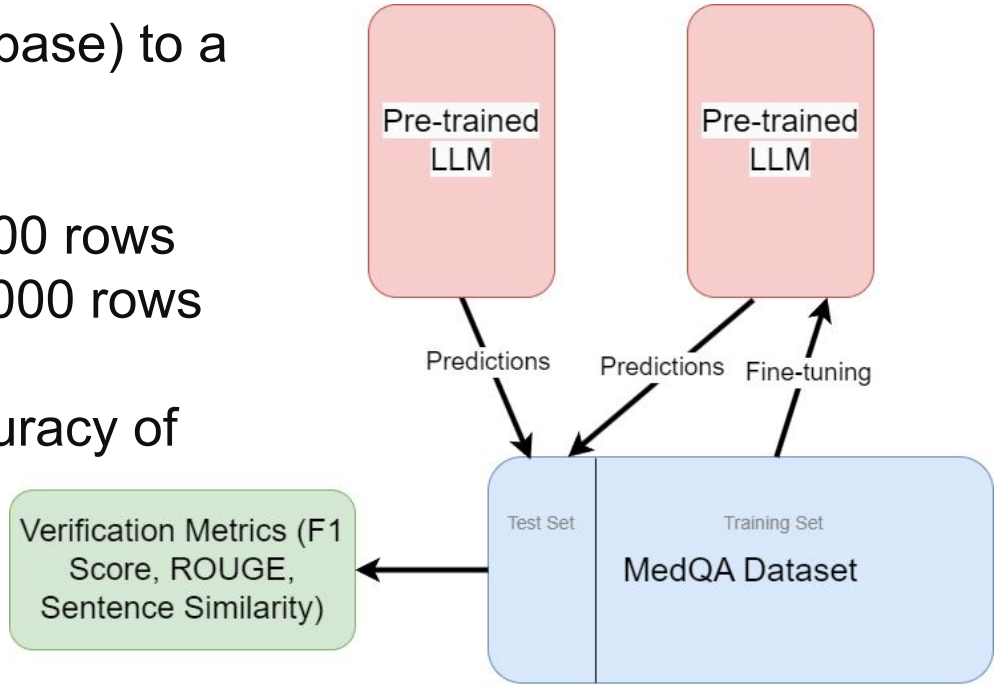
Fine-tuned vs. Base Model
Group 5

- Luke Leimbach - Data Preprocessing | Presentation
- Joseph Juri - Fine-Tuning | Documentation
- Raphael Kandler - Metrics | Analysis

GitHub Project: GitHub Dataset: Medical QA

Introduction

- Goal
  - Compare Falcon 7b LLM (base) to a fine-tuned version
- Comparison
  - Pretrained dataset -> 80.000 rows
  - Medical QA dataset -> 10.000 rows
- Prediction
  - Significant increase in accuracy of predictions

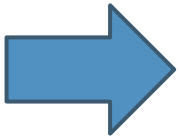


Data Preparation

- Data
  - Removed multiple-choice aspect
  - Updated instruction to fit non-multiple-choice results
  - Tokenized questions

<u>Original JSON:</u>

{instruction: …,
input: (context + question + options) …,
output: A,B,C,D …
}, ...

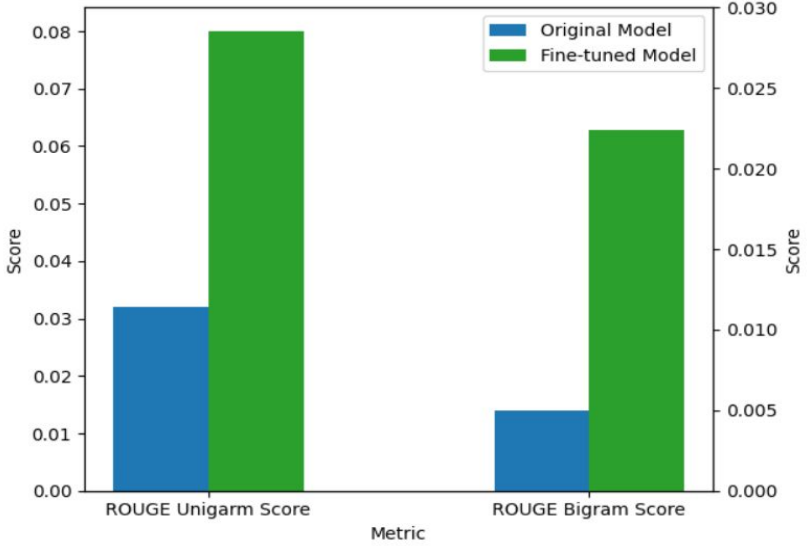


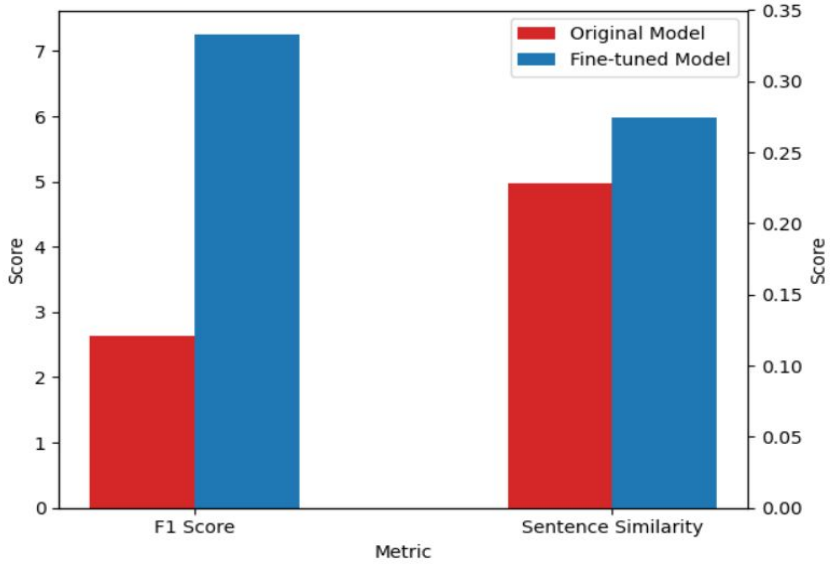
<u>Output CSV:</u>

idx, instruction, context, output, question

Results

- Fine-tuned using 3.000 rows from subset of Medical QA [1]
- The results were as expected based on the prediction





[1] Medical QA

Conclusion

- Results
  - High overlap between the fine-tuned model's outputs and the correct answers to the posed questions
- Limitations
  - Free-form outputs were difficult to validate
  - Domain expert scoring of fine-tuned model outputs would lead to more accurate metrics
- Biases
  - Unknown data from the pre-trained dataset could have introduced unintended biases