Bachelorarbeit

Evaluierung von Retrieval-Augmented Generation und Fine-Tuning in der Wissensrepräsentation: Ein Vergleich mit Großen Sprachmodellen

Evaluation results

Results for each model and variants.

How to visualize results

Use requirements_eval.txt for venv.
All plots and table scripts to visualize data are located here.
Answers.html can load multiple results and display question, answers, references for each model and variant.

How to train model?

Use requirements.txt for venv.
Add Model to models list, set model_index, save.
Then `py train_ll.py`.
Trained models are saved here.

How to evaluate model?

Use requirements.txt for venv.
for basemodel "Qwen/Qwen2-0.5B-Instruct"

```
python evaluate_llm.py --modelname "<full_model_name>"
```

for basemodel with rag

```
python evaluate_llm.py --modelname "<full_model_name>" --use_rag
```

for ft model

```
python evaluate_llm.py --modelname "<full_model_name>" --use_ft
```

for ft model with rag

```
python evaluate_llm.py --modelname "<full_model_name>" --use_ft --use_rag
```

if using ft model, local ft model will be loaded

gpu stats cli

```
watch -n 0.5 nvidia-smi
```