LAIwyer
Project Report for NLP Course, Winter 2025

Adam Majczyk
Warsaw University of Technology
adam.majczyk.stud@pw.edu.pl
Szymon Matuszewski
Warsaw University of Technology
szymon.matuszewski.stud@pw.edu.pl
supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION
  - embedding model: *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2*
  - LLM: *hf.co/speakleash/Bielik-11B-v2.3-Instruct-GGUF:Q4_K_M*
  - Reranker: *BAAI/bge-reranker-large*
- LINK TO CODE: https://github.com/szymonsm/lAIwyer-NLP/
- INFRASTRUCTURE – RAG
  - *PGVector* as vector store
  - *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* as the embedding model
  - *hf.co/speakleash/Bielik-11B-v2.3-Instruct-GGUF:Q4_K_M* as the LLM
  - *BAAI/bge-reranker-large* as the reranker
- RUNTIME PARAMETERS – Average runtime for each approach: 23.42 seconds (on Nvidia RTX 4080 with the question *"Co mówi konstytucja o strajkach?"*); The time of the whole pipeline over 5 runs
- PARAMETERS – The number of parameters in each model
  - embedding model: 118M params
  - LLM: 11B params
- VALIDATION PERFORMANCE – maximum: 0.946
- METRICS – Accuracy

Multiple Experiments:

- NO TRAINING EVAL RUNS – N/A - no training of LLM or embedding performed
- HYPER BOUND – N/A - no training of LLM or embedding performed
- HYPER BEST CONFIG – N/A - no training of LLM or embedding performed
- HYPER SEARCH – N/A - no training of LLM or embedding performed
- HYPER METHOD – N/A - no training of LLM or embedding performed
- EXPECTED PERF – N/A - no training of LLM or embedding performed

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS – Relevant statistics, such as the number of examples
  - 243 parsed articles of Polish constitution (in one JSON file; 1460 lines, 136 KB)
- DATA SPLIT – N/A - no training of LLM or embedding performed

- DATA PROCESSING – The articles were parsed to be in this format

```
{
    "section": "I",
    "section_title":
    ↪ "RZECZPOSPOLITA",
    "article": "1",
    "text": "Rzeczpospolita
    ↪ Polska jest dobrem
    ↪ wspólnym wszystkich
    ↪ obywateli."
}
```

- DATA DOWNLOAD – On the REPO, in folder `data`: https://github.com/szymonsm/lAIwyer-NLP/blob/main/data/parsed_constitution.json
- DATA EVALUATION – 41 multiple choice questions (ABC) about the Polish Constitution from e.g. the demo version of https://arslege.pl/testy-demo/
- NEW DATA DESCRIPTION – N/A
- DATA LANGUAGES – For natural language data, the name of the language(s): **Polish**