Reproducibility Appendix
LLM exam generation - Project for NLP Course, Winter 24/25

Nikita Kozlov
nikita.kozlov.stud@pw.edu.pl

Zofia Łagiewka
zofia.lagiewka.stud@pw.edu.pl

Jakub Świstak
jakub.swistak.stud@pw.edu.pl

Jacek Zalewski
jacek.zalewski.stud@pw.edu.pl

Supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION –
  In our experiments we have used two model through their APIs. The first model is **GPT-4o**. More details about the model and the model card is provided in https://openai.com/index/hello-gpt-4o/.
  Furthermore, **Llama 3.1** was used. Model card and more details about a model can be seen on: https://ai.meta.com/blog/meta-llama-3-1/

- LINK TO CODE –
  A repository for the project can be found at https://github.com/nk2IsHere/NLP-2024W/blob/main/projects/final/Lagiewka_Kozlov_Swistak_Zalewski_exam_gen.

  Below we provide the list of all dependencies:

  - NumPy: 1.26.1
  - Pandas: 2.0.2
  - Seaborn: 0.12.2
  - LangChain: 0.0.226
  - Matplotlib: 3.8.0
  - TQDM: 4.65.0
  - JSON: Standard Library
  - langchain-openai: 0.0.226
  - langchain-ollama: 0.0.226

  Below we provide the list of required software:

  - Ollama
  - Python 3.12
  - Jupyter with IPython Kernel

- INFRASTRUCTURE –

  For running ChatGPT and EDA notebooks any computer capable of running jupyter notebook and basic python scripts with internet access is required.

  For running LLaMA notebooks a computer capable of running LLaMA 3.1 8b is required. Example setup used to run the notebooks is an Apple computer with M2 Pro and 16 GB RAM.

- RUNTIME PARAMETERS – Runtime parameters:

  - LLaMA 3.1 temperature=0.0
  - GPT-4o temperature=0.0

  Environment parameters (specified in .env file):

  - OPENAI_API_KEY - OpenAI API Key with access to GPT-4o generation capabilities

- PARAMETERS –

  - LLaMA 3.1 - 7 billion
  - GPT-4o - unknown

- VALIDATION PERFORMANCE – We use multiple metrics to validate performance of the dataset. Code for validation can be found in AbA_compute_metrics notebook. Results for selected data sample can be found in the report section 7.

- METRICS
  - Content Uniqueness (CU)
  - Dissimilarity Index (DSI)
  - D Metric
  - Type-Token Ratio (TTR)
  - Corrected TTR (CTTR)
  - Flesch Reading Ease Score (FRES)
  - Flesch-Kincaid Grade Level (FKGL)
  - Automated Readability Index (ARI)

Multiple Experiments:

- NO TRAINING EVAL RUNS – N/A
- HYPER BOUND – N/A
- HYPER BEST CONFIG – N/A
- HYPER SEARCH – N/A
- HYPER METHOD – N/A
- EXPECTED PERF – N/A

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

- DATA STATS – Data statistics is described in the eda/hotpot_qa, eda/news_qa, eda/squad notebooks for the selected datasets and results mirrored in the report.
- DATA SPLIT – From the NewsQA, StanfordQA and HotpotQA the official validation split was used to asses the ability of the LLM to generate and answer questions.
- DATA PROCESSING – Data processing is described in the eda/hotpot_qa, eda/news_qa, eda/squad notebooks for the selected datasets and results discussed in the report.
- DATA DOWNLOAD – Datasets are available at:
  - SQuAD - Hugging Face
  - HotpotQA - Hugging Face
  - NewsQA - Hugging Face
- NEW DATA DESCRIPTION – The following data was generated as the result of this project:
  - datasets/hotpotqa_4o_questions_with_answers.json - results of GPT-4o generating close-ended questions with answers for HotpotQA dataset
  - datasets/hotpotqa_llama_generated_questions.json - results of LLaMA generating open-ended questions for HotpotQA dataset
  - datasets/hotpotqa_llama_questions_with_answers.json - results of LLaMA generating close-ended questions with answers for HotpotQA dataset
  - datasets/newsqa_4o_questions_with_answers.json - results of GPT-4o generating close-ended questions with answers for NewsQA dataset
  - datasets/newsqa_llama_generated_questions.json - results of LLaMA generating open-ended questions for NewsQA dataset
  - datasets/newsqa_llama_questions_with_answers.json - results of LLaMA generating close-ended questions with answers for NewsQA dataset
  - datasets/squad_llama_generated_questions.json - results of LLaMA generating open-ended questions for SQUAD dataset
  - datasets/squad_llama_questions_with_answers.json - results of LLaMA generating close-ended questions with answers for SQUAD dataset
  - open_ended_questions_cu.csv - results of CU metrics for a sample of open-ended questions
  - open_ended_questions_dsi.csv - results of DSI metrics for a sample of open-ended questions
  - open_ended_questions_metric.csv - results of other metrics for a sample of open-ended questions
- DATA LANGUAGES – The collected data was in English.