Reproducibility Appendix
Project Report for NLP Course, Winter 2025

Aleksandra Kulczycka
aleksandra.kulczycka.stud@pw.edu.pl
WUT student

Agata Osmałek
agata.osmalek.stud@pw.edu.pl
WUT student

Michał Iwicki
michal.iwicki.stud@pw.edu.pl
WUT student

Michał Legczylin
michal.legczylin.stud@pw.edu.pl
WUT student

Reproducibility checklist

Overall results:

- All models were loaded and maintained using Huggin Face infrastructure.
- Model used for rephrasing and judging: `meta-llama/Llama3.1-8B-Instruct`.
- Text-only models used: `microsoft/Phi-3-mini-4k-instruct`, `Qwen/Qwen2.5-3B-Instruct`, `google/gemma-2-2b-it`.
- Multimodal models used: `HuggingFaceTB/SmolVLM-Instruct`, `llava-hf/llava-1.5-7b-hf`.
- The judging prompt is defined in `src/evaluation/judge_prompt.txt`.
- Computations were performed on a platform equipped with an AMD Ryzen 5 7500F CPU, an NVIDIA RTX 4070 Super GPU with 12 GB VRAM, 32 GB DDR5 RAM, running Windows 11 and Python 3.11.9.
- Detailed information and performance statistics are provided in the notebooks `prompt_evaluation.ipynb` and `prompt_generation.ipynb`.
- Evaluation metrics included simple accuracy for the judge and the invalid output rate.

Multiple Experiments:

- Each model–dataset combination was generated exactly once.
- Each answer was scored exactly once by the judge.
- Rephrasing was performed with a batch size of 8; single-turn answer generation used a batch size of 10; multi-turn generation could not be batched; multimodal generation used a batch size of 4.
- Judging was performed with a batch size of 8 for multimodal samples and 10 for text-only samples.

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

- We generated 200 prompts for each category: `mental-physical-health`, `sensitive-data-extraction`, and `social-engineering`.
- Multimodal prompts accounted for 40 samples per category.
- Multi-turn prompts accounted for 20 samples per category, with a randomly selected number of escalations ranging from 1 to 5.
- Single-turn prompts accounted for 140 samples per category, including 80 template-based prompts, 40 template rephrasings, and 20 lexical mutations of previously generated prompts.
- All data and the prompt generation pipeline are provided in the notebook `prompt_generation.ipynb`, which allows for regenerating the prompts or creating new datasets.