Reproducibility Appendix
Project Report for NLP Course, Winter 2025

Hubert Jaczyński
Warsaw University of Technology
01171199@pw.edu.pl

Aleksandra Kłos
Warsaw University of Technology
01171204@pw.edu.pl

Bartosz Maj
Warsaw University of Technology
01171317@pw.edu.pl

Jakub Oganowski
Warsaw University of Technology
01168843@pw.edu.pl

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

Reproducibility checklist

Overall results

- **MODEL DESCRIPTION** – We benchmark lightweight open-source LLMs for safety by measuring refusal/compliance on a hybrid prompt dataset (manual + synthetic). We use an LLM-as-a-Judge protocol (Qwen2.5-1.5B-Instruct) to classify each model response as `REFUSED` or `COMPLIED`, and compute safety/usability metrics. The dataset includes both single-turn and multi-turn dialogues (2–6 turns) and covers multiple risk categories (e.g., cross-lingual manipulation, toxic language, illegal substances & weapons, jailbreak roleplay, bias & fairness).

- **LINK TO CODE** – Source code repository / MR: https://github.com/awroble/NLP_2025W.
  The repository contains the notebook(s) and scripts to (i) run model inference, (ii) run judge-based evaluation, and (iii) reproduce the reported tables and plots. Dependencies are documented in `requirements.txt` / notebook installation cells.

- **INFRASTRUCTURE** – Experiments were run in a limited-hardware environment on Google Colab, which is recommended for reproducibility.

- **RUNTIME PARAMETERS** – Each prompt is evaluated once per model. Judge evaluation is performed in batches; for one complete judging run we observed 85 batches completed in ∼1m34s (progress log).

- **PARAMETERS** – Text models: TinyLlama-1.1B, StableLM-(Zephyr)-1.6B, Bloomz-1.7B, Qwen2.5-1.5B.
  Multimodal model: LLaVA-1.5-7B.
  Judge model: Qwen2.5-1.5B-Instruct.

- **VALIDATION PERFORMANCE** – Not applicable: we do not train models in this PoC. All reported results are computed directly on the benchmark dataset. (Future work: validate the judge on a manually labeled subset.)

- **METRICS** – We compute: (i) unsafe refusal rate ("Safety score") = mean(`refused | unsafe`), (ii) ORR (over-refusal) = mean(`refused | safe`), (iii) global F1 where $y = 1$ for unsafe prompts and $\hat{y} = 1$ if the model refused. The metric computation is implemented in the evaluation notebook/scripts.

Multiple Experiments

- **NO TRAINING EVAL RUNS** – No training runs are performed. Each (prompt, model) pair is generated once and judged once (with an optional deterministic refusal-phrase override).

- **HYPER BOUND** – Generation hyperparameters (e.g., temperature, top_p, max_new_tokens) are fixed and logged in the run configuration.

- **HYPER BEST CONFIG** – Not applicable (no training / hyperparameter optimization in this PoC).
- **HYPER SEARCH** – Not applicable.
- **HYPER METHOD** – Not applicable.
- **EXPECTED PERF** – We report per-model summary metrics (unsafe refusal rate / ORR / F1) and category-wise breakdown plots. Variance estimates are not reported in the PoC stage.

Datasets

- **DATA STATS** – Benchmark segment contains 200 prompts total: 160 text prompts (including multi-turn dialogues) and 40 multimodal prompts.
- **DATA SPLIT** – No train/validation/test split is used because no training is performed. The dataset is used exclusively for evaluation.
- **DATA PROCESSING** – Prompts are generated via a hybrid pipeline: selected high-quality prompts from SOTA benchmarks + template-based injection for scaling + conversion of single-turn prompts into 2–6 turn dialogues. Data are stored in a structured format including ID, `variant` (safe/unsafe), risk category, and source. Minimal cleaning is applied (format normalization, consistent fields).
- **DATA DOWNLOAD** – Not applicable (dataset is provided as part of the project repository).
- **NEW DATA DESCRIPTION** – New data are created by: (i) selecting representative prompts from established benchmarks per risk category, (ii) applying template-based intent injection to increase linguistic diversity, and (iii) creating multi-turn variants to simulate realistic adversarial behavior (e.g., role-play framing).
- **DATA LANGUAGES** – English is the primary language; cross-lingual manipulation prompts include Polish / mixed-language cases.