Reproducibility Appendix
Project Report for NLP Course, Winter 2025

Natalia Safiejko
natalia.safiejko.stud@pw.edu.pl

Wojciech Grabias
wojciech.grabias.stud@pw.edu.pl

Krzysztof Sawicki
krzysztof.sawicki3.stud@pw.edu.pl

Mikołaj Mróz
mikolaj.mroz.stud@pw.edu.pl

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

Reproducibility checklist

Overall results:

- **MODEL DESCRIPTION** – The project implements safety evaluation of Large Language Models (LLMs) using multiple models: Llama-3.2-3B-Instruct (quantized Q4_K_M format, target model for jailbreak testing), Gemma-2-9B-IT (quantized Q4_K_M format, judge model for safety classification), Gemini Pro (used for generating expected behavior descriptions), and vision-language models (ggml-model-q4_k.gguf for vision processing, mmproj-model-f16.gguf for CLIP projector). The approach involves: (1) generating responses to jailbreak prompts using the target model, (2) generating expected safe behavior using Gemini Pro, (3) evaluating responses using the judge model (Gemma-2-9B-IT), and (4) determining if responses are safe or unsafe based on judge evaluations. Local models run using quantized GGUF format for efficient inference, while Gemini Pro is accessed from web.
Huggingface links:
Llama-3.2-3B-Instruct
Gemma-2-9B-IT
ggml-model-q4 k
mmproj-model-f16

- **LINK TO CODE** – Available at: https://github.com/ssafiejko/nlp_safety_llms. The repository shall be kept private, but its snapshots with appropriate deliverables are available at https://github.com/awroble/NLP_2025W. Dependencies specified in `requirements.txt`. Installation instructions provided in `README.md`.

- **INFRASTRUCTURE** – Experiments run on local GPU infrastructure hosted on an ARM-based MacBook using GGUF quantized models for efficient inference. Local models (Llama-3.2-3B-Instruct Q4_K_M, Gemma-2-9B-IT Q4_K_M, and vision models) run without requiring cloud API endpoints. Gemini Pro accessed via Google Web App for generating expected behavior descriptions.

- **RUNTIME PARAMETERS** – Inference parameters include: temperature=0.0 (for deterministic responses), max_tokens=4096 (for model responses), max_tokens=8192 (for judge outputs). n_gpu_layers has been set to -1 for faster inference on Apple Silicon. All other inference parameters are the defaults of `llama_cpp 0.3.16` interface for `Python`.

- **PARAMETERS** – Llama-3.2-1B-Instruct: 1 billion parameters, Llama-Guard-3-1B: 1 billion parameters. Gemini Pro parameter counts not specified (proprietary models). All models used in inference-only mode (no fine-tuning).

- **VALIDATION PERFORMANCE** – Not applicable. This is a safety evaluation task using pre-trained models, not a traditional supervised learning setup with train/validation/test splits.

- **METRICS** – Evaluation uses jailbreak success rate (proportion of prompts that elicit unsafe responses) and judge-based safety classification.

Multiple Experiments:

- **NO TRAINING EVAL RUNS** – No model training performed (evaluation of pre-trained models only). Evaluation consists of: 142 jailbreak prompts tested on Llama-3.2-3B-Instruct, each response evaluated by Gemma-2-9B-IT judge model.
- **HYPER BOUND** – Temperature fixed at 0.0 for reproducibility (deterministic generation). Max tokens: 4096 for target model responses, 8192 for judge model outputs. No ranges or bounds explored as parameters are fixed.
- **HYPER BEST CONFIG** – Not applicable. No hyperparameter tuning performed; all experiments use fixed parameters for reproducibility. Models evaluated as-is with standard inference settings.
- **HYPER SEARCH** – Not applicable. No hyperparameter search conducted; study focuses on comparing pre-trained model behaviors under fixed conditions.
- **HYPER METHOD** – Not applicable. Fixed parameters used throughout (temperature=0.0, max_tokens fixed). No optimization or selection process involved.
- **EXPECTED PERF** – Results reported as proportions and percentages of unsafe responses. No variance, standard deviation, or error bars provided as experiments are deterministic (temperature=0.0 ensures reproducible outputs for given inputs). Single-run results presented for each configuration.

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

- **DATA STATS** – Dataset that was prepared contains 142 jailbreak prompts organized into 4 risk categories: Fabrication_Hallucination (30 prompts), Hidden_Policy_Compliance (30 prompts), Emotional_Manipulation (30 prompts), and Multimodal_Jailbreak (22 prompts). Total dataset size: 142 prompts across these categories designed to test different aspects of model safety and robustness.
- **DATA SPLIT** – No train/validation/test split applied. All prompts used exclusively for evaluation of model safety. This is an evaluation-only dataset with no training or fine-tuning performed.
- **DATA PROCESSING** – No preprocessing, filtering, or modification applied to prompts. Each prompt passed directly to target model with system prompt prepended. System prompt instructs model to be helpful assistant. No data excluded from evaluation.
- **DATA DOWNLOAD** – Jailbreak prompts were prepared by the project team, they can be found in `dataset_poc.json`. Models downloaded from Hugging Face: Llama-3.2-3B-Instruct, Gemma-2-9B-IT, and vision models in GGUF quantized format. For future users, pre-downloaded models will be made available via Google Drive for easier setup and reproducibility.
- **NEW DATA DESCRIPTION** – Not applicable. No new data collected; existing publicly available jailbreak dataset used for evaluation purposes.
- **DATA LANGUAGES** – English language only. All jailbreak prompts, model responses, and judge evaluations conducted in English.
- **LLM PROMPTS** – All user prompts are contained in the `prompts.txt` file and have been used universally across all evaluation entries.