Reproducibility Appendix
Project Report for NLP Course, Winter 2025/6

Klaudia Kwoka
Warsaw University of Technology
klaudia.kwoka.stud@pw.edu.pl

Pola Mościcka
Warsaw University of Technology
pola.moscicka.stud@pw.edu.pl

Maciej Wach
Warsaw University of Technology
maciej.wach.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska@pw.edu.pl

Reproducibility checklist

Overall results:

- MODEL DESCRIPTION
  The experiments use two multimodal models: Qwen2-VL-7B-Instruct (Qwen) and LLaVA-1.5-7B-HF (LLaVa), both designed to generate responses from text and image inputs. Both models are based on the transformer architecture, where sequences of tokens, text or image embeddings, are processed through self-attention layers to produce contextualized representations. The models are trained using instruction-following objectives, optimizing a cross-entropy loss over token predictions given the input prompt, which allows them to generate coherent and relevant outputs conditioned on both textual and visual information. Qwen uses a vision-language transformer that integrates image features via a chat template, passing images as parameters to the model. LLaVA builds on the LLaMA architecture and incorporates visual information through a special `<image>` token embedded directly in the prompt. Both models are capable of multimodal reasoning and instruction following, but differ in how images are represented and integrated into the model's input sequence. Licenses differ as well: Qwen uses Apache-2.0, while LLaVA uses the LLAMA 2 Community License.

- LINK TO CODE
  All code used for the experiments, including scripts for generating model responses, profiling latency and memory, and producing all plots included in this report, is available on `https://github.com/klaudiakwoka/NLP`. The repository is well-documented, with clear instructions for setup, running experiments, and reproducing results. It also includes a `requirements.txt` file listing all necessary Python packages and dependencies.

- INFRASTRUCTURE
  All generation computations were executed on an NVIDIA Rtx4070ti card with 12GB of VRAM. Some preliminary tests, such as prompt generation as well as evaluation and plotting the results, were executed locally on a CPU. All experiments were executed using `Python 3.12`. We added full reproducibility settings, including seeds for `Python`, `numpy`, and `pytorch`.

- RUNTIME PARAMETERS
  To accommodate large multimodal models under limited GPU memory constraints, we employed `bitsandbytes`, a lightweight library for efficient low-bit quantization integrated with the Hugging Face Transformers framework. The configuration loads model weights in 4-bit NF4 precision with double quantization to minimize memory usage, while performing computations in bfloat16 to maintain numerical stability during inference. Specific time and memory usage results are shown in Section 3.5 of the main report.

- PARAMETERS – LLaVa model has around 7B parameters whereas Owen 8B.

- VALIDATION PERFORMANCE
  Validation performance is not applicable in this study, as we are evaluating model responses to unsafe prompts rather than training or testing on a standard dataset. The focus of our analysis is on the models' behavior and safety responses, rather than traditional train/test performance metrics.

- METRICS
  The main metric used in our experiments is

the *evaluation score*, defined as the average score across all models and seeds assigned to each prompt. We also used the *refusal rate*, which measures the proportion of refusals among all model responses. A false refusal rate was considered as well, but it was not used, as we did not observe any refusals for safe prompts.

To evaluate both the LLM-as-a-judge approach and zero-shot learning, we calculated accuracy, defined as the proportion of model decisions that matched our manual annotations.

Code can be found in `https://github.com/klaudiakwoka/NLP/tree/main/evaluations`.

Multiple Experiments:

- NO TRAINING EVAL RUNS
  No training or evaluation runs were performed; prompt responses were generated on 3 seeds, LLM-as-a-Judge evaluation was conducted on 3 seeds, and zero-shot evaluation was run on 1 seed.

- HYPER BOUND
  Not applicable, as fixed hyperparameters were used for both model response generation and judge evaluation. No search over ranges was performed.

- HYPER BEST CONFIG
  Fixed generation parameters were used for each task. For model response generation:
  - max_new_tokens: 150
  - min_new_tokens: 50
  - do_sample: true
  - temperature: 0.1
  - repetition_penalty: 1.1

  For judge evaluation:
  - max_new_tokens: 1
  - min_new_tokens: 1
  - do_sample: false

- HYPER SEARCH
  Not applicable, as no hyperparameter search was performed; parameters were fixed for each type of evaluation.

- HYPER METHOD
  Not applicable. Hyperparameter values were manually set based on prior knowledge, literature review and the requirements of the evaluation task, with no search or optimization method applied.

- EXPECTED PERF
  Summary statistics are not directly reported as hyperparameter performance metrics, since no search or tuning was done. Variability in model outputs is captured through multiple seeds for model response generation and judge evaluation, ensuring robustness of results across runs.

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

- DATA STATS
  Our dataset contains 650 prompts, of which 80% are classified as unsafe and the rest as safe.

- DATA SPLIT
  No standard train/validation/test split was applied, as the dataset is used solely for model evaluation rather than training.

- DATA PROCESSING
  All prompts were used as-is. No prompts were excluded, and no additional preprocessing steps were applied.

- DATA DOWNLOAD
  The dataset is available at `https://github.com/klaudiakwoka/NLP/tree/main/data`.

- NEW DATA DESCRIPTION – Our dataset was curated for evaluating model responses to potentially unsafe prompts. Prompts were manually collected, categorized into three risk levels, and labeled with expected behavior to ensure quality and consistency.

- DATA LANGUAGES – All prompts are in English.