Pola Mościcka, Klaudia Kwoka, Maciej Wach

In this document, we describe how we addressed the reviewers' comments. Comments from the reviewers are marked in blue, while our responses are provided in black.

Review from Team Alicja Charuza, Martyna Kuśmierz, Dawid Sroczyk

Weak points:

Unfair Model Comparison: Comparing SmolVLM 2B parameters against Qwen2-VL-7B and LLaVA-1.5-7B is unbalanced. The smaller model is prone to incoherence simply due to capacity issues, not necessarily safety alignment differences.

This was a valid point. We removed the SmolVLM-2B model, as its small size did not yield reliable answers. Instead, we focused on conducting additional experiments with the two remaining models.

Lack of Safe Baselines: The experiments focused only on "deliberately unsafe prompts". Without testing safe prompts, you cannot measure "over-refusal" (e.g., does the model refuse a safe recipe because it sounds like a bomb recipe?). A robust benchmark needs both to measure utility vs. safety.

Safety prompts now account for 20% of the dataset. In the proof of concept (PoC), our primary focus was on unsafe prompt generation and on analyzing the model's output. However, we agree that additional safety prompts should have been included.

Temperature Setting: You set temperature to 0.0 for reproducibility. However, jailbreaking often relies on model creativity/stochasticity to bypass filters. A temperature of 0 might make the models artificially "safer" or more repetitive than they would be in a real attack scenario.

We agree that this was a valid suggestion. Accordingly, the temperature parameter was increased and set to 0.1.

Quantitative Scarcity: The results section is extremely sparse, containing qualitative outcomes for only three specific prompts. Even for a PoC, this sample size is too small to draw meaningful conclusions about model behavior.

We have extended the sections on experiments and conclusions. The dataset comprises 650 prompts, and we conducted multiple experiments.

Insufficient Detail on Prompt Generation: The description of how prompts were generated for bias/fairness testing and jailbreak roleplay is too general. More specific details on the generation process are needed.

In the final project report, Section 2.1 (Dataset Generation Methodology) provides a detailed explanation of how the prompts were constructed. All prompts were created manually; however, for the Multi-turn Persuasion and Jailbreak-Roleplay categories, we followed a schema, which is described in this section. For the Bias and Fairness category, no formal schema was used, but we provide a general description along with representative examples.

Vague Definition of Safety Categories: The literature review provides a rather vague description of safety categories. It would be beneficial to explain the definitions of these categories and how they function, rather than just listing them.

We do not fully agree with this point. In Section 1.2 (Related Work), we already provide relevant definitions and descriptions. We believe that no further additions are strictly necessary. However, we have added a few sentences to this section and expanded the description of the bias and fairness category for clarity.

Additional suggestions and questions for the Authors to improve their work.

We addressed the reviewers' comments in the report from this section. For example, we no longer use the phrase "this paper." We do not agree that the word "Abstract" should be removed, as it is the first section of the report. We also added a description explaining why jailbreak-roleplay works and we updated the generation parameters. Following the reviewers' suggestion, we included the LLM-as-a-judge approach for scalable evaluation.
Regarding the suggestion to use the model's tendency to be verbose as a vulnerability instead of limiting output length, we did not implement this approach. Our main evaluation was manual as one-shot and LLM-as-a-judge evaluations did not provide reliable outcome.

Reproducibility issues and code cleanliness

There were few changes needed. As suggested by the reviewers, we did not include any .zip files in the repository and we added docstrings to all functions and classes.

Review from Team Huber Jaczynski, Aleksandra Kłos, Bartosz Maj, Jakub Oganowski

Weak points

Your evaluation dataset is limited in scale, consisting of only 23 prompts across three risk categories. This sample size is insufficient to support general conclusions about model safety or to observe statistically meaningful trends.

We have extended the sections on experiments and conclusions. The dataset comprises 650 prompts, and we conducted multiple experiments.

Your reported results are predominantly qualitative. Your paper does not include quantitative safety metrics, such as attack success rate (ASR), refusal rate, or per-category aggregation, and this limits the ability to compare models.

We agree with this comment and have added additional metrics as suggested. These are described in Section 2.7 (Evaluation Metrics), including, for example, the refusal rate.

Your experiments include only a minimal number of multimodal prompts. As a result, the paper does not provide sufficient evidence to assess the safety of vision-language models.

The Dataset now consists of 120 multimodal prompts (40 for each category).

The paper does not include safe control prompts that resemble attack patterns. Without such a control group, it is impossible to measure over-refusal or distinguish between genuinely robust models and those that simply refuse most inputs.

Safety prompts now account for 20% of the dataset. In the proof of concept (PoC), our primary focus was on unsafe prompt generation and on analyzing the model’s output. However, we agree that additional safety prompts should have been included.

Your selection of evaluated models (SmolVLM, Qwen2-VL, LLaVA) is not justified in the report. It makes it difficult to contextualize the observed behaviors.

This was a valid point. A description of the model selection process is provided in Section 2.3 (Generation pipeline). We removed the SmolVLM-2B model, as its small size did not yield reliable answers. Instead, we focused on conducting additional experiments with the two remaining models.

For multi-turn persuasion scenarios, your evaluation does not analyze how model behavior evolves across turns. Your paper reports final outcomes only, without measuring escalation dynamics or turn-level compliance.

We agree with this suggestion; however, due to limitations of the implemented LLM-as-a-judge and one-shot learning, we decided to evaluate manually 650 prompts on three seeds for two models. Conducting an evaluation of 250 multi-turn conversations after each turn would have required substantial additional work and we did not manage to do that. Instead we focues on running variable expariments. We agree that this would provide valuable insights and consider it an excellent direction for future work.

In Section 3.1, Table 2, one of the outputs is labelled as 'incoherent,' which was not previously introduced.

There is no “incoherant” label anymore.

Your report claims a ‘’diverse dataset with neutral and harmful prompts’’, but the dataset description says ‘’almost all prompts were labeled unsafe’’.

Safety prompts now account for 20% of the dataset. In the proof of concept (PoC), our primary focus was on unsafe prompt generation and on analyzing the model’s output. However, we agree that more safety prompts should have been included.

Additional suggestions and questions for the Authors to improve their work.

We addressed the reviewers’ comments in the report from this section. We increased the number of prompts substancially. We used quantitative evaluation metrics, such as refusal rate. We expanded multimodal evaluation by designing image-text prompts. We did not extend the multi-turn analysis as we explained in previous section. We provided a justification for model selection. The generation parameters were adjusted. Following the reviewers’ suggestion, we included the LLM-as-a-judge approach for scalable evaluation.

Questions to Authors:

How do you define attack success in your evaluation? Is a refusal always considered a safety success, or can it also represent over-refusal?

Not always, if the prompt variant is “safe” the refusal output represents over-refusal. If the prompt variant is “unsafe” the refusal output is a safety success.

Why does the dataset include only 23 prompts across three categories? Is this intended as a minimal proof of concept or an incomplete benchmark?

This was intended as a minimal proof of concept to test the prompts we constructed and to identify which prompt design approaches might be the most effective.

Do you plan to expand the multimodal component of the dataset? If so, how will images be sourced or generated, and how will multimodal safety be evaluated?

We extended the multimodal component to include 120 prompts, combining images and text. The images were obtained from Unsplash with free licensing. The evaluation of multimodal prompts followed the same procedure as for text-only prompts.

In multi-turn persuasion scenarios, do you observe increased compliance over successive turns? Are specific escalation strategies more effective than others?

As previously stated we do not evaluate the multi-turn conversations after each turn. The persuasion techniques were compered to each other and there is an experiment and analysis in section 3.2.2 Persuasion techniques

As previously stated, we did not evaluate multi-turn conversations after each turn. Instead, the persuasion techniques were compared to each other, and the corresponding experiment and analyses are presented in Section 3.2.2 (Persuasion Techniques).

What criteria guided the selection of SmolVLM, Qwen2-VL, and LLaVA as evaluation targets? Are these models representative of distinct safety alignment approaches?

We provided a justification for model selection in Section 2.3 Generation pipeline. We selected Qwen2-VL and LLaVA to capture diversity in model size, architecture, and safety alignment. Each represents a distinct approach to safety tuning, providing a broad view of unsafe prompt effectiveness.

Missing references, ethical concerns

As suggested by the reviewers, we added a few additional datasets and related work. We would like to clarify the ethical concerns (Section 1.1 in the report). This project is intended solely as a safety benchmark for LLMs and will not be used for harmful purposes. Its aim is to evaluate LLM safety alignment with our constructed dataset.

Work limitations comments

We addressed the comments from this section, which are consistent with those previously discussed in this file.

Reproducibility criteria, reproducibility issues and code cleanliness

We corrected almost all of the issues pointed out by our colleagues. The code is now cleaner and more understandable.

Review after POC

We clearly state the contribution of each team member (Table 9 in the report).
The references have been updated, and all figure and table captions were revised to be easily understandable at first glance.
As suggested, we added a table that provides a comparison across datasets relevant to our work (Section 2.4).
Table 4 includes the model's generation parameters, making them easy to interpret.
Each experiment was conducted using three different random seeds, and the reported metrics were adjusted accordingly (Section 2.4).
Time and memory usage were measured and analyzed (Section 3.5).
Regarding the section *"Rebuttal or corrections for all the tips given by Dr. Jan Sawicki"*, all noted issues were addressed and corrected.
Comparisons with other datasets are provided in Sections 1.3 and 2.2.
The prompt structure is described in Section 2.1 and is consistent with the project requirements. The submission no longer contains any .zip files.
Finally, a reproducibility checklist is included.