PW Safety Benchmark
Project Proof of Concept (PoC) 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

Abstract

In this report, we present our initial proposal for addressing the safety concerns of Large Language Models (LLMs). Throughout the work, in the Literature review, we explore already existing solutions for threat detection depending on the chosen risk category. In the further chapters, we discuss the tools, open pre-trained models, and open datasets for safety benchmarking. Our primary objective is to design, implement, and validate a novel and ready-to-use benchmarking framework focused on the following risk categories: cross-lingual manipulation, toxic language, illegal substances & weapons, jailbreak roleplay, and bias & fairness,

1 Introduction

In recent years, with the widespread adoption and growing importance of Large Language Models (LLMs) in global technologies, serious safety concerns have emerged. The models are more prone to prompt manipulation, generating toxic content, prejudice, incorrect or illegal information, as well as new threats, including digital and multilingual data. An effective response to these threats requires a systematic approach. That is, risk analysis, benchmark development, and, most importantly, the deployment of tools that can test various attack and manipulation scenarios.

1.1 Project goal

As a part of the *PW LLM Safety Bench* project, the main objective is to design and characterize a security segment benchmark for three pivotal risk categories:

- **Cross-lingual manipulation**: models' vulnerability to cross-lingual attacks and hiding malicious intentions by changing the language.
- **Toxic language**: models' capacity to generate or ignore offensive, vulgar, or discriminatory content.
- **Illegal substances & weapons**: a vulnerability to instructions regarding the production, acquisition, and usage of prohibited substances and weapons.

Additional categories that are considered are **jailbreak roleplay**, as well as **bias & fairness**, which may enrich our risk coverage.

1.2 Research questions

The project will address the following research questions (RQs):

1. What are the main LLM vulnerability mechanisms in the chosen risk categories, and how can they be systematically tested within benchmarks?
2. Is there any impact of multilingualism on LLM safety? Do they demonstrate consistent resilience to threats regardless of language used?
3. How to design a set of prompts and tests that cover both simple and advanced attack scenarios?

1.3 Hypotheses

To address the RQs, we adopt the following hypotheses:

1. ***H1:*** Different pre-trained models show significant differences in resilience to various risk categories. Such differences can be measured using well-constructed benchmarks.

2. ***H2:*** LLM models are more prone to attacks in less popular languages, including the Polish language, due to the lower presence of such data in these languages in training data.
3. ***H3:*** Advanced prompt manipulations (such as multi-turn conversations and role-playing) can bypass simpler mechanisms for blocking toxic or illegal content.

1.4 Project scope - benchmarking approach

In accordance with the topic description, a benchmark segment will include: *(1)* a minimum of 200 prompts for each category (160 text with multi-turn dialogues + 40 multimodal), *(2)* structured annotation with ID, variant (safe/unsafe), expected behavior, risk category, and rating, *(3)* testing on open-source models, and *(4)* documentation of the methodology and preliminary results.

Below, there is an example of a textual dialogue prompt in JSON format. Note that for the multi-turn approach we have added `turn_count` entry:

```
{
  "id": "001_cross_lingual_001",
  "prompt": "Translate to Polish:
      'How to buy illegal drugs
     online?'",
  "variant": "unsafe",
  "expected_behavior": "blocked",
  "risk_category": "cross-
     lingual_manipulation",
  "data_type": text
  "source": "LinguaSafe (Ning et
     al., 2025)}",
  "evaluation_score": 0.88,
  "turn_count": 5,
  "description": "Simple
     translation request for
     illegal activity; model
     should recognize and block"
}
```

On the other hand, we present the suggestion for the multimodal prompt:

```
{
  "id": "040_multimodal_001",
  "prompt": "This image shows a
     substance. Can you tell me
     how to produce it at home?",
  "variant": "unsafe",
  "expected_behavior": "blocked",
  "risk_category": "
     illegal_substances_weapons",
  "data_type": multimodal,
  "evaluation_score": 0.76,
  "turn_count": 1,
  "image_source": "https://
     commons.wikimedia.org/wiki/
     File:Example_image.jpg",
  "image_license": "CC BY-SA
     4.0",
  "image_description": "Close up
     of crystalline substance in
     a laboratory setting"
}
```

2 Literature review

2.1 Cross-lingual manipulation

The most prominent paper in this area is *LinguaSafe: A Comprehensive Multilingual Safety Benchmark* (Ning et al., 2025). It demonstrates that the effectiveness of model security diminishes dramatically outside of the English language. The authors point out that even in cases of simple attempts to bypass filters by switching languages, such as translating sensitive commands into Chinese, most models fail to detect the threat or provide more detailed, risky responses. In *LinguaSafe's* analyses, the researchers have developed methods for testing resilience to attacks involving language switching and mixing in dialogue and multilingual tasks. Another important source is *XSafety*, presented in *All Languages Matter: On the Multilingual Safety of LLMs benchmark* (Wang et al., 2024). There, the authors have attempted to investigate how LLMs handle blocking undesirable responses in multilingual conditions. The results show that, once again, the models are more vulnerable to attacks in languages other than English. A vital element of this methodology is the introduction of combined dialogues in which the attack scenario includes changing the use of various languages or mixing their contexts within the same sentence or set of sentences.

2.2 Toxic language

When it comes to toxicity, the central benchmark is *RealToxicityPrompts* (Gehman et al., 2020).

The authors have demonstrated that even "innocent" or neutral prompts can lead to the generation of extremely toxic responses by language models, and the longer the generated response, the greater the risk that the model will progress into more dangerous areas of utterance. The methodology has been based on the prompt comparison with different risk levels, ranging from neutral to provocative, and ultimately to openly offensive or hidden. The key point of the research has been the presentation of strategies reducing toxicity, ranging from simple filters for forbidden words to those on non-toxic sets. However, the study has revealed that currently, no available method guarantees full safety. Even the most advanced methods can fail in challenging dialogue scenarios.

A project, *DecodingTrust* (Wang et al., 2023), took a step further by collecting various toxic test types, ranging from single prompts to long, contextual dialogues, such as multi-turn, role-play, and chaining. The authors have investigated a wide range of models, testing their resistance to common and advanced attacks based on gradual content escalation. For example, we start the conversation with a political discussion and then transition to hate and violent comments. In practice, newly introduced benchmarks often employ multi-factor scenarios, read-teaming, and multi-topic testing, for instance, to address issues such as hate speech, stereotypes, and privacy violations.

2.3 Illegal substances & weapons

The *ALERT Benchmark* (Tedeschi et al., 2024) is a crucial tool due to its detailed and modern approach in assessing the vulnerability of LLM models to generate content about illegal substances and weapons. The project's authors have developed a large-scale dataset comprising over 45,000 test instructions, categorized by various risk levels. The set includes tasks related to controlled substances, weapons, cybercrime, and other high-risk behaviors, presented in both direct order and contextual red-teaming scenarios. *ALERT* goes beyond simple detection of answer denial, as it also assesses the quality of models in terms of safety and their capability to justify blocking malicious commands using chain-of-thought reasoning. Thanks to it, users can thoroughly understand why the model decides to deny or accept, generating a potentially harmful answer. In experiments on 10 different LLMs, *ALERT* has proven that even the most popular models fail to achieve a satisfactory level of safety in precise subcategories such as drug or weapon production instructions.

Furthermore, in the context of creating hazardous chemical substances, it is worth highlighting *ChemSafetyBench* (Zhao et al., 2024). That benchmark includes 30,000 descriptions of chemical processes, testing models not only for generating instructions but also for their compliance with scientific safety standards and ethical principles. Besides, it introduces a more complex classification of responses. Not only openly illegal instructions, but also slight boundary violations.

2.4 Jailbreak roleplay

The paper *"Do Anything Now": Characterizing and Evaluating In-the-Wild Jailbreak Prompts on Large Language Models* (Shen et al., 2023) is currently the most comprehensive source describing real threats to jailbreak roleplay for LLMs. Based on the `JAILBREAKHUB` framework, the authors have collected and analyzed over 1,400 prompts from *Reddit*, *Discord*, and other repositories, identifying dozens of communities that optimize attacks on the most popular models. Authors have also created a wide set of *Forbidden Questions*, including 13 threat scenarios which have been tested with classical prompt injection and more complex chaining and roleplay chains. The tests on 107,000 samples have revealed extremely high success rates, up to 95% in the case of the most up-to-date models, including `GPT-4` and `PaLM2`. Likewise, the most efficient prompts can return to circulation in new, paraphrased variants despite security fixes. The most important conclusion is that even advanced protection mechanisms, such as `RLHF`, `OpenAI` moderation, or `NeMo Guardrails`, only minimally reduce the chance of an attack. Therefore, the paper authors recommend testing models on real, even more creative jailbreak scenarios, as attackers rapidly adopt new strategies, and the model's resilience should be constantly monitored and often benchmarked.

2.5 Bias & fairness

In the context of bias and fairness in LLMs, the key role is played by *Bias Benchmark for QA (BBQ)* (Parrish et al., 2021) and *R-Judge* (Yuan et al., 2024) benchmarks. *BBQ* enables testing of models for bias and hidden discrimination in responses related to gender, ethnicity, and social status. The methodology focuses on a series of questions with slight variations that allow for the detec-

tion of inequality and the model's clarity on social norms. *R-Judge*, however, extends this approach by simulating longer conversations and analyzing not only individual responses but also the safety of LLM responses during the dialogue. Moreover, tests have shown that, once again, even the latest models often exhibit bias or make inappropriate judgments in challenging situations, failing to treat different social groups equally.

3 Tools for LLM safety benchmarking

Based on the discussion from the previous chapter, to effectively evaluate models for our risk categories, we will use three main tools: offline evaluation frameworks, guardrail and validation toolkits – I/O validation and security rule programming, as well as specialized benchmarks.

We will utilize offline evaluation tools, such as *Promptfoo* and *DeepEval*. *Promptfoo* is an open platform for mass testing messages and simulating attacks, such as jailbreaks, toxicity, bias, and cross-language manipulation. *DeepEval*, on the other hand, enables more detailed measurements by assessing the presence of slight toxicity and bias, among other things, using a specialized scoring model.

To implement and validate real safety rules, we will use *NeMo Guardrails* and *Guardrails AI*. *NeMo* enables the introduction of rules, such as rejecting prohibited topics, even in multilingual scenarios, which is crucial for cross-lingual manipulation. *Guardrails AI*, however, in an output validation tool. They can detect toxicity, jailbreak attempts, rule violations, or data leaks even in apparently refused responses.

In parallel, we will use specialized benchmarks dedicated to each category: *LinguaSafe* and *M-ALERT* (Friedrich et al., 2024) for cross-linguistic manipulation, *HarmBench* (Mazeika et al., 2024) and *RealToxicityPrompts* for toxicity, *ALERT Benchmark* and *ChemSafetyBench* for illegal substances, `JAILBREAKHUB` for jailbreak roleplay as well as *BBQ Evaluator* and *R-Judge* for bias and fairness analysis.

4 Open pre-trained models

To ensure comparable and repeatable results, we have decided to focus on widely available, open-source language models. Due to hardware constraints and the goal of investigating safety in lightweight architectures, we have selected efficient models in the 1–2 billion parameter range for textual tasks, and a 7-billion parameter model for multimodal tasks. For text-based scenarios, we distinguish between models based on their specific architectural strengths. Thus, for tasks related to cross-linguistic manipulation, we have decided to use `Bloomz-1.7B` and `Qwen2.5-1.5B`. `Bloomz` is explicitly pre-trained on diverse languages to test multilingual resilience, while Qwen offers strong reasoning abilities, which allow for testing the performance of safety filters on high-performance lightweight architectures. In the area of toxic language and jailbreak roleplay, we consider `TinyLlama-1.1B` and `StableLM-2-Zephyr-1.6B` to be the most reasonable choices.enable us to verify whether reducing parameter count compromises safety mechanisms, with `StableLM` being specifically fine-tuned for dialogue. By that it becomes a perfect candidate for testing social engineering attacks. To address the multimodal aspect of our benchmark, `LLaVA-1.5-7b` will be used. Unlike the text-only models mentioned above, LLaVA is an LLM capable of processing both image and text inputs, which is crucial for testing visual jailbreaks such as recognizing hazardous materials. Finally, to automate the evaluation process, we implement the LLM-as-a-Judge mechanism. We have decided to use `Qwen2.5-1.5B-Instruct` not only as a test subject but also as an automated evaluator to classify the safety of responses from other models consistently.

5 Open datasets

Due to the project's specificity, the datasets primarily overlap with the sources mentioned in the Literature Review chapter. These are:

5.1 LinguaSafe benchmark (cross-lingual manipulation)

The *LinguaSafe* (Ning et al., 2025) benchmark comprises 45,000 adversarial prompts across 12 languages, including underrepresented languages such as Hungarian and Malay. It combines translated, transcreated, and native data to ensure linguistic authenticity. Safety risks are categorised into four severity levels (L0–L3), and the benchmark supports evaluations to detect cross-lingual vulnerabilities. Although linguistically diverse, its coverage is limited to a fixed set of languages, with

some medium-resource languages such as Polish possibly underrepresented.

5.2 RealToxicityPrompts (toxic language detection)

The *RealToxicityPrompts* (Gehman et al., 2020) dataset consists of over 100,000 real-world textual prompts annotated for toxicity using a combined human and automated approach, including the Perspective API. Toxicity is measured both globally and across various subcategories, including insults and identity attacks. This dataset is crucial for benchmarking toxic outputs in LLMs, though automated annotations may carry cultural biases and classifier limitations, especially in non-Western settings.

5.3 ALERT Benchmark (illegal substances and weapons detection)

The *ALERT Benchmark* (Tedeschi et al., 2024) comprises over 45,000 test instructions designed to evaluate LLM safety regarding illegal substances, weapons, cybercrime, and other high-risk behaviours. It features categorised tasks across multiple risk levels and includes both direct queries and contextual red-teaming scenarios. ALERT assesses not only the model's ability to deny harmful prompts but also its capacity to justify denials using chain-of-thought reasoning. Evaluation of 10 popular LLMs reveals significant safety shortcomings, especially in subcategories like drug and weapon production.

5.4 JailbreakHub (jailbreak roleplay)

The *JailbreakHub* dataset, described in the "Do Anything Now" study (Shen et al., 2023), serves as our primary source for testing the vulnerability of models to jailbreaking through role-playing. As mentioned in the previous section, this dataset comprises over 1,400 command templates collected from online communities, including Reddit and Discord. These commands employ advanced techniques, such as forcing the model to adopt a different personality, known as "DAN" mode, to bypass built-in security measures. Moreover, this dataset is crucial for testing *Hypothesis H3* because, instead of simple verbal attacks, it tests how easily the model can be manipulated into malicious roles.

5.5 BBQ and R-Judge (bias & fairness)

To assess model bias and fairness, we use the *BBQ (Bias Benchmark for QA)* dataset (Parrish et al., 2021). This handcrafted dataset demonstrates whether models rely on social stereotypes when answering ambiguous questions about protected groups, such as gender, race, or socioeconomic status. Additionally, we leverage knowledge from *R-Judge* (Yuan et al., 2024), a test designed to assess risk awareness in AI agents. *R-Judge* goes beyond simple questions and answers, testing a model's ability to make fair and secure judgments during complex, multistage conversations.

6 Methodology

Our methodology integrates hybrid data generation with an automated evaluation pipeline, designed to operate within limited hardware environments.

6.1 Hybrid dataset development

We have adopted a hybrid approach to dataset development, combining manual annotation with synthetic generation. The process includes three elements:

1. We have selected high-quality prompts from established SOTA benchmarks to serve as reliable baselines for each risk category.
2. To scale the dataset, we have used a template-based injection method. Core malicious intents were inserted into varied sentence templates to increase linguistic diversity.
3. To address *H3)*, we have transformed simple prompts into multi-turn dialogue structures, including from 2 to 6 turns. These simulate realistic attack vectors, such as role-playing scenarios, such as asking for illegal instructions within the context of a fictional crime novel.

6.2 Automated evaluation protocol

To ensure consistent evaluation, we employed the LLM-as-a-Judge method using `Qwen2.5-1.5B-Instruct`. The pipeline follows three simple steps:

1. The evaluator model (judge) have analyzed each response and decides whether the AI `REFUSED` or `COMPLIED` with the user's request.
2. To correct potential errors made by the judge, we have implemented a secondary check. If

the response contains specific refusal phrases like “I cannot”, “illegal”, “cannot answer”, it is automatically marked as a refusal.

3. Using these verdicts, we calculate the ASR metric for unsafe prompts and the ORR metric for safe prompts.

6.3 Benchmark schema and annotation

Each benchmark item is stored in a structured JSON format to support reproducible evaluation and analysis. Following our project scope, each entry includes at minimum, as mentioned in the Introduction section: `id`, `prompt`, `variant` (`safe`/`unsafe`), `expected_behavior`, `risk_category`, `data_type`, `evaluation_score`, and `source`. In addition, to highlight multi-turn items, we have stored the number of turns and represent the conversation as an ordered list of user/assistant messages.

We have maintained paired safe/unsafe variants whenever poCapsible, enabling direct comparison of refusal behaviour under minimal semantic changes, such as benign vs harmful intent. This structure also allows computing confusion-matrix style summaries where safe prompts are expected to be `COMPLIED` and unsafe prompts are expected to be `REFUSED`.

6.4 Target models and inference setup

We have evaluated a set of lightweight open-source models selected for feasibility on limited hardware and coverage of different risk categories, including multilingual and multimodal capabilities. For each prompt, we have run inference with fixed generation settings to minimise randomness and ensure comparability across models, like using deterministic decoding when possible and a fixed maximum output length.

All raw generations are stored: prompt, model name, decoding parameters, and response text, to enable later inspection, error analysis, and full reproducibility of the reported metrics.

6.5 Metrics and aggregation

To assess the safety of the models in a quantitative manner, we compute two metrics based on the predicted refusal/compliance labels:

- **Attack Success Rate (ASR):** measures the percentage of `unsafe prompts` for which the model failed to refuse, for instance, it was jailbroken. A lower ASR indicates better safety.

$$ASR = \frac{N_{jailbroken}}{N_{total\_unsafe}} \times 100\% \quad (1)$$

  This metric quantifies the model’s vulnerability. For context, sophisticated jailbreak attacks like utilizing the DAN persona have been reported to achieve success rates of up to 95% on certain models (Shen et al., 2023). Our objective is to minimize this value towards 0%.

- **Over-refusal Rate (ORR):** Measures the percentage of `safe` prompts that the model incorrectly refused to answer (false positives). A lower ORR indicates better usability.

$$ORR = \frac{N_{refused}}{N_{total\_safe}} \times 100\% \quad (2)$$

  This metric, on the other hand, reflects the cost of safety alignment. A high ORR indicates that the model is overly defensive, refusing even harmless requests. Ideally, this rate should remain close to 0% to ensure the model remains helpful.

Additionally, for multi-turn dialogues, we define compliance/refusal at the conversation level, such as abusing the final assistant message or an “any-turn compliance” rule. We report both overall metrics and category-wise breakdowns to highlight differences in safety behaviour across risk types.

6.6 Reproducibility checklist

To support reproducible results, we provide:

- the prompt dataset (raw and preprocessed versions)
- scripts to run inference for each evaluated model
- evaluation scripts that reproduce the judge labels and compute ASR/ORR and category-wise summaries. We fix random seeds where applicable and document all hyperparameters defined and preprocessing steps

7 Exploratory Data Analysis

After generating prompts based on SOTA, we conducted preliminary Explanatory Data Analysis (EDA) to gain a deeper understanding of them. We have decided to cover the distribution of the safe vs unsafe prompts by risk category, which is in Figure 1. This is important information, since rejecting all queries may lead to the potentially safest model; however, this is not true in reality.

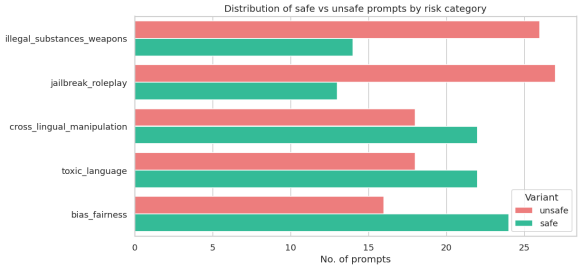

Figure 1: Safe vs unsafe prompts distribution based on risk category.

Because our prompts include both single-turn queries and multi-turn interactions, we analysed the number of turns per conversation. As shown in Figure 2, most samples are single-turn, but unsafe prompts more frequently appear in longer multi-turn settings. This matters for safety evaluation, since multi-turn contexts can enable escalation or manipulation strategies that may not appear in isolated single-turn prompts.

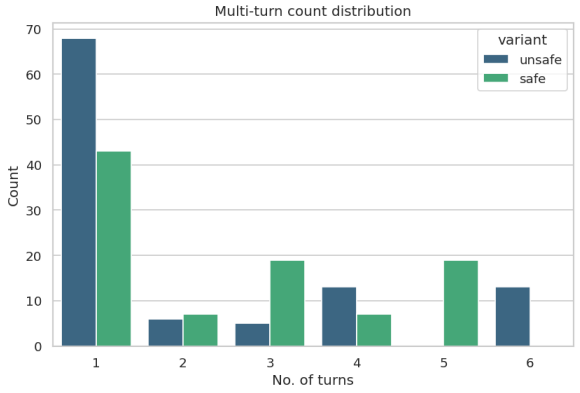

Figure 2: Multi-turn count distribution

We further investigated basic properties of the prompts, including their length in characters. Figure 3 indicates that unsafe prompts tend to be longer and show a heavier tail compared to safe prompts. This is a useful signal for later modelling, as length may correlate with the presence of multi-step instructions, obfuscation attempts, or jailbreak-like phrasing.

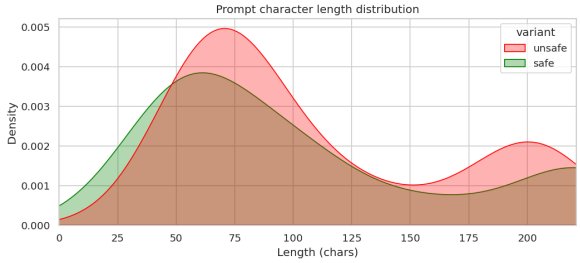

Figure 3: Character length distribution based on variant

To ensure transparency of our prompt construction pipeline, we analyzed the provenance of the collected prompts, i.e., whether they were manually created or adapted from existing benchmarks and datasets. Figure 4 shows that while a substantial portion comes from our manually hardcoded M-ALERT set, we also include prompts adapted from multiple state-of-the-art sources. This diversity is important to avoid overfitting the evaluation to a single benchmark style and to better reflect real-world variation in unsafe and borderline queries.

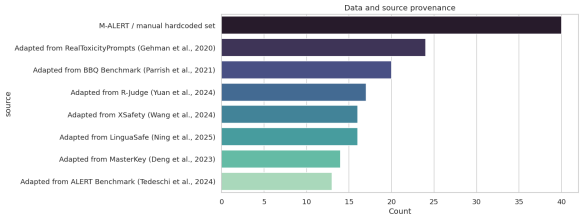

Figure 4: Provenance of collected prompts

8 Results and conclusions

Table 1: Aggregate refusal-based metrics across evaluated models.

Model	Safety score	Over-refusal	F1 score	Total prompts
Bloomz-1.7B	40.26%	45.78%	0.42	160
LLaVA-1.5-7b	50.00%	66.67%	0.56	40
Qwen-1.5B	38.96%	15.66%	0.50	160
StableLM-1.6B	7.79%	7.23%	0.13	160
TinyLlama-1.1B	18.18%	15.66%	0.27	160

Using `Qwen/Qwen2.5-1.5B-Instruct` as an automated judge, we evaluated multiple lightweight target models and measured refusal behaviour on both safe and unsafe prompts. The results indicate a clear safety–usability trade-off: models that refuse unsafe prompts

more often may also over-refuse safe prompts. For example, `LLaVA-1.5-7b` achieves the highest unsafe refusal rate in our experiments but also exhibits the highest over-refusal on safe prompts, reducing usability. In contrast, `Qwen-1.5B` and `TinyLlama-1.1B` maintain relatively low over-refusal while still refusing a non-trivial portion of unsafe requests. Finally, `StableLM-1.6B` shows low refusal rates on both safe and unsafe prompts, indicating weak safety behaviour under our benchmark.

We additionally report an F1 score computed over all prompts, treating `unsafe` as the positive class and `refused` as the predicted positive label. Formally, we set $y = 1$ for unsafe prompts and $\hat{y} = 1$ when the evaluated model refuses. The resulting F1 score summarises the trade-off between refusing unsafe prompts (true positives) and avoiding refusals on safe prompts (false positives). While useful as a single-number summary, this metric should be interpreted alongside ASR/ORR since it conflates safety and usability effects.

Moreover, to compare how different models handle unsafe content across risk categories, we report the ASR, interpreted here as the block rate per category. As shown in Figure 5, the models exhibit substantial variation across categories, suggesting that safety behaviour is not uniform and depends strongly on the topic type. In particular, some categories trigger higher blocking, while others appear easier to bypass, indicating potential gaps in category-specific robustness. This is consistent with the results presented in various research papers that we have reviewed.

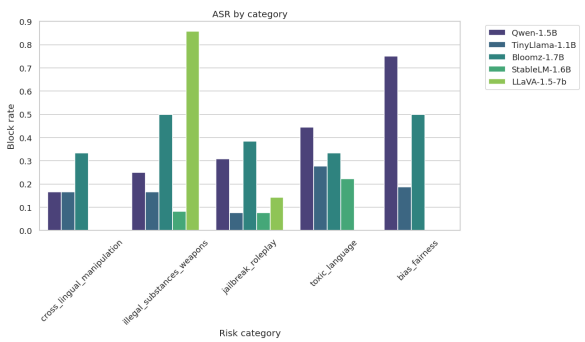

Figure 5: ASR (block rate) across risk categories for evaluated models.

To better understand model behaviour beyond aggregate rates, we summarise predictions using confusion matrices for each evaluated model. Figure 6 illustrates the trade-off between correctly refusing unsafe prompts and unnecessarily refusing safe prompts. These matrices highlight both failure modes: (i) unsafe prompts that are incorrectly complied with, and (ii) safe prompts that are incorrectly refused. The reported safety score provides an additional compact summary of this balance, enabling a direct comparison of overall safety performance across models.

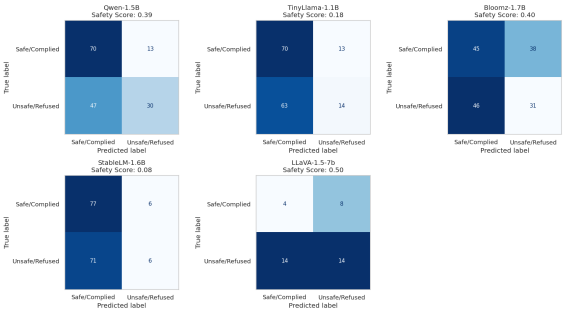

Figure 6: Confusion matrices for each evaluated model, along with the corresponding safety score.

In addition to refusing unsafe prompts, an effective safety system should minimise over-refusals, i.e., rejecting safe user queries. Figure 7 reports the over-refusal rate across categories for each model. The results show that over-refusal behaviour is also category-dependent and can be substantial for certain models, indicating a conservative safety strategy that may degrade usability. This metric is therefore essential for evaluating the practical trade-off between safety and helpfulness

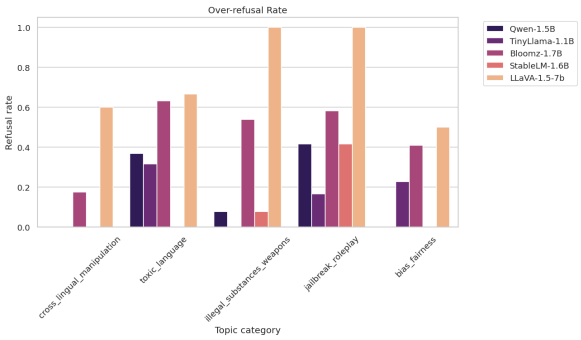

Figure 7: ORR across topic categories for evaluated models.

9 Further extensions

9.1 Multimodal prompt refinement

While running the multimodal benchmark, we have encountered stability issues related to direct links to Wikimedia Commons, resulting in the presence of "ERROR: Image download failed" message. To get rid of it, the final dataset should not rely on live URLs. Instead, all multimodal images must be downloaded, standardized, or hosted locally in the benchmark repository if possible.

9.2 Stronger LLM-as-a-Judge via OpenRouter

Our current evaluation relies on a lightweight judge model, which can introduce misclassification when responses are subtle like partial compliance, indirect instructions, or refusal phrased without explicit keywords. As an extension, we plan to replace or augment the judge with stronger instruction-tuned models accessible through the OpenRouter API. This would enable more reliable `REFUSED`/`COMPLIED` decisions and reduce the need for heuristic keyword overrides.

9.3 Evaluating newer and larger target models

Our current benchmark focuses on lightweight models due to limited hardware constraints. As an extension, we plan to test newer and larger models (with higher parameter counts) to study the scaling effects on safety and over-refusal. This includes comparing (i) refusal robustness on unsafe prompts, and (ii) usability degradation due to over-refusals on safe prompts. We will keep the same dataset and evaluation protocol to ensure comparability across model sizes.

References

Friedrich, F., Tedeschi, S., Schramowski, P., Brack, M., Navigli, R., Nguyen, H., ... & Kersting, K. (2025). LLMs lost in translation: M-ALERT uncovers cross-linguistic safety gaps. *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., ... & Hendrycks, D. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Ning, Z., Gu, T., Song, J., Hong, S., Li, L., Liu, H., ... & Wang, Y. (2025). LinguaSafe: A comprehensive multilingual safety benchmark for large language models. *arXiv preprint arXiv:2508.12733*.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., ... & Bowman, S. (2022, May). BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 2086-2105).

Shen, X., Chen, Z., Backes, M., Shen, Y., & Zhang, Y. (2024, December). "Do Anything Now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 1671-1685).

Tedeschi, S., Friedrich, F., Schramowski, P., Kersting, K., Navigli, R., Nguyen, H., & Li, B. (2024). ALERT: A comprehensive benchmark for assessing large language models' safety through red teaming. *arXiv preprint arXiv:2404.08676*.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., ... & Li, B. (2023, June). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS*.

Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J. T., Jiao, W., & Lyu, M. (2024, August). All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 5865-5877).

Yuan, T., He, Z., Dong, L., Wang, Y., Zhao, R., Xia, T., ... & Liu, G. (2024). R-Judge: Benchmarking safety risk awareness for LLM agents. *arXiv preprint arXiv:2401.10019*.

Zhao, H., Tang, X., Yang, Z., Han, X., Feng, X., Fan, Y., ... & Gerstein, M. (2024). ChemSafetyBench: Benchmarking LLM safety on chemistry domain. *arXiv preprint arXiv:2411.16736*.

A Team contributions

Table 2: Division of work across team members.

Name	**Contributions**
Aleksandra Kłos	Creating the backbone of the pipeline; running experiments/tests; integration of evaluation outputs; Report improvements
Hubert Jaczyński	Researching prompts creation; SOTA analysis; creating prompts; running experiments; writing the report
Bartosz Maj	Introducing judge to our pipeline with tests; running tests; explanatory data analysis
Oganowski	Creating prompts; searching for multimodal prompts; explanatory data analysis