Multi-lingual Multi-turn Automated Red Teaming for LLMs

Abhishek Singhania, Christophe Dupuy, Shivam Mangale, Amani Namboori
Amazon
{mrabhsin,dupuychr,mangsh,anamburi}@amazon.com

Abstract

Warning: This paper includes content that may be considered inappropriate or offensive to some readers. Viewer discretion is advised. Language Model Models (LLMs) have improved dramatically in the past few years, increasing their adoption and the scope of their capabilities over time. A significant amount of work is dedicated to "model alignment", i.e., preventing LLMs to generate unsafe responses when deployed into customer-facing applications. One popular method to evaluate safety risks is *red-teaming*, where agents attempt to bypass alignment by crafting elaborate prompts that trigger unsafe responses from a model. Standard human-driven red-teaming is costly, time-consuming and rarely covers all the recent features (e.g., multi-lingual, multi-modal aspects), while proposed automation methods only cover a small subset of LLMs capabilities (i.e., English or single-turn). We present Multi-lingual Multi-turn Automated Red Teaming (**MM-ART**), a method to fully automate conversational, multi-lingual red-teaming operations and quickly identify prompts leading to unsafe responses. Through extensive experiments on different languages, we show the studied LLMs are on average 71% more vulnerable after a 5-turn conversation in English than after the initial turn. For conversations in non-English languages, models display up to 195% more safety vulnerabilities than the standard single-turn English approach, confirming the need for automated red-teaming methods matching LLMs capabilities.

1 Introduction

In recent years, the landscape of Language Model Models (LLMs) has evolved drastically, with numerous releases showcasing enhanced capabilities over time. These advancements have positioned LLMs as formidable tools capable of a wide range of tasks, from generating creative text to powering virtual assistants and chat-bots. Even smaller open LLMs such as Mistral (Jiang et al., 2023), Llama (Meta-AI, 2024) or Molmo (Deitke et al., 2024) have demonstrated close to state-of-the-art performance across various tasks. Their effectiveness makes them viable options for integration into enterprise applications, particularly due to their lower latency and cost-effectiveness. However, this increase in capabilities means that models are even more susceptible to generate unsafe content which could harm customers (e.g., detailed instructions to build a bomb). Recent models are now capable of holding long conversations in multiple languages, which offer even more possibilities for unsafe content generation. To tackle this challenge, "red-teaming" emerges as a crucial strategy aimed at assessing and mitigating the potential adverse effects of LLM-generated content. Red-teaming entails a systematic approach to adversarial probing and evaluation of an LLM's responses, with the objective of identifying safety violations. LLMs are then "aligned" by incorporating red-teaming data into their training, making them more robust to attacks and ensuring the generated content adheres to ethical standards set by their builders. Standard red-teaming involves human testers interacting with LLMs in an attempt to trigger unsafe responses, aka "jailbreaking". This approach relies on the creativity and expertise of humans, who craft scenarios to challenge the LLM across different contexts. However, due to its manual nature, human red-teaming can be time-consuming and resource-intensive. In contrast, automated red-teaming relies on various ML models, allowing for more scalable and efficient evaluation, although human creativity is still needed for exploring new vulnerabilities. Most recent studies focus on capturing jailbreak methods in either multi-turn or multilingual scenarios (Deng et al., 2023; Russinovich et al., 2024), but no existing research conduct simultaneous safety evaluation across multiple capability dimensions during red-teaming. We present a novel approach, Multilingual and Multi-turn Automated Red Teaming (MM-ART), and the first safety evaluation on a set of widely popular LLMs for attacks in non-English conversational settings. We believe this line of study is critical for expanding LLMs across the globe, covering different languages and delivering a similar safe experience to different users.

Our contributions: 1) We propose a novel approach, MM-ART, to evaluate the safety of models across both languages and conversational depth. We provide a detailed description of our approach and share the components used to build this method. 2) We conduct a thorough evaluation of popular LLMs using MM-ART and present the first comprehensive study around safety assessment of conversational LLMs across multiple languages and safety categories. 3) Our detailed analysis of MM-ART through ablation studies provides insights into the impact of the different components of our approach onto the safety levels of the evaluated LLMs.

2 Related Work

A wide variety of single-turn "static jailbreaking" methods have been proposed in the past year, which consist of formatting a static prompt in a way that triggers unsafe response from the LLM by rephrasing, spreading across multiple turns or adding many prompts into LLM context (Sun et al., 2024; Agarwal et al., 2024; Upadhayay and Behzadan, 2024; Li et al., 2024; Cheng et al., 2024; Anil et al., 2024). Other works have looked into multi-turn "static jailbreaking", where from a static adversarial prompt, a conversation is held with the target LLM aiming at triggering a response to the initial prompt. For instance, (Russinovich et al., 2024) propose an automated method to manipulate the target LLM with regeneration and gradual intensification of prompts. Additionally, (Yang et al., 2024) include a semantic-driven strategy for generating new turns and show that incorporating more complex, multi-turn contextual scenarios into the safety alignment phase strengthen LLM protection. Both methods are restricted to the provided input task/prompt. We add a conversation starter generation component which makes MM-ART more flexible and suitable to cover broader assessment over a given safety category. Plus, these studies rely on large closed models with very long prompt templates and multiple regenerations per turn while our experiments are exclusively conducted with small open models significantly increasing efficiency and scalability.

Studies on multi-lingual LLMs focus on single-turn attacks, showing LLMs are more vulnerable when prompted in low resource languages (Yong et al., 2024; Etxaniz et al., 2024) or with code-switching (Yoo et al., 2024) than in English. Undesirable outputs are significantly reduced by instructing the LLM to think in English (Wang et al., 2024). While there has been major progress in automated red-teaming, existing work on simultaneous multi-lingual and multi-turn red teaming is limited, even more so when considering conversations on unrestricted topics. MM-ART is designed to bridge that gap by providing an efficient and scalable method to systematically identify safety gaps in LLMs.

3 Multi-lingual Multi-turn Automated Red Teaming (MM-ART)

Our proposed Multi-lingual Multi-turn Automated Red Teaming (MM-ART) approach is divided into two sequential steps. We first generate prompts that will be used to start conversations (called "conversation starters"), setting the topic and tone for the conversation. Second, for each conversation starter, we complete the conversation in a given language for a specific depth (i.e., number of turns each containing a prompt and a response) by adapting to the LLM responses to continue the conversation. This two step approach allows for maximum flexibility, where the conversation starters are extracted from a variety of sources (e.g., generated by human or machine), covering different categories, different attack techniques etc.

3.1 Conversation Starters Generation

Although human-generated prompts is the gold standard for red-teaming evaluation, it is not feasible to generate a large set of prompts solely with humans. We leverage LLMs to generate conversation starters with three main objectives. 1) Scale up red-teaming operations, 2) Maintain or improve the efficacy of the generated prompts for triggering unsafe responses compared to human-generated prompts, 3) Maintain or expand the scope of red teaming evaluation (by maintaining diversity through generation). We leverage small LLMs to generate these adversarial prompts in English with in-context learning (ICL) (Brown et al., 2020). We select 5 conversation starters generated by humans related to a single safety category and instruct the LLM to generate novel examples through carefully crafted instructions. As demonstrated by our experiments, the choice of LLM and the quality of instructions is crucial to maintain the high-quality of the human prompts (see Section 5.2).

3.2 Automated Multi-turn Conversation

Given a conversation starter, our objective for multi-turn generation is to probe the target model on the same topic until it produces an unsafe response. Most existing approaches rely on single-turn attacks, for which the prompts have to be direct and aggressive to trigger unsafe response since they correspond to one-shot attempts. Our method is able to trigger unsafe responses by gradually probing the target model about a certain topic turn after turn until the model generates sufficient content to essentially override its safety alignment. We use an agent approach with an LLM (similar to Russinovich et al.) that takes prompt instructions, a safety category, and the current conversation as inputs. Given an already started conversation, our goal is to generate a prompt that is contextually relevant for the conversation and that maintains the conversation along the same category provided as input. Since the conversation already contains an important piece of context, the instructions to the LLMs are kept simple. Finally, the generated prompt for the next turn is appended to the current conversation which is sent to the target model for its response. The next turn generation process is repeated for the desired number of turns.

3.3 Multi-lingual Conversations

Most recent LLMs support dozens of languages and conducting conversational human red-teaming for each target model in every supported language would be prohibitively expensive and time-consuming. Similarly, requiring human to translate machine generated conversations would be extremely long given the scale of such multi-turn attacks. Analysis done by Deng et al. on comparing human and machine translation shows that using automatic translation doesn't significantly affect the effectiveness or quality of the attacks. We build our approach upon these findings and leverage machine translation for multi-lingual red-teaming as follows. First, since LLMs works best in English (Etxaniz et al., 2024; Yong et al., 2024), we keep the conversation starter and next turn generation in English only (we empirically observed qualitative degradation of generations when prompting LLMs in other languages). For a given conversation starter in English, we translate it to the desired language and send the translated version to the target model. The received response, also in the desired language, is translated back to English. We send the English conversation to the next-turn generation pipeline, translate the generated prompt for the next turn to the desired language and append the translation to the conversation in the desired language. Finally, the conversation in the desired language is sent to the target model for a response in the desired language. These steps are repeated until the required number of turns are completed. Through this process, we maintain the conversations both in English and in the desired language. The downstream assessment of the generated conversations is streamlined, as we have the option to conduct assessment either in the desired language (with potentially low resources) or in English.

4 Experiments

4.1 Conversation Starters Datasets

We work specifically with 7 safety categories generally used for red-teaming. The list of categories and the corresponding volumes for the 4 datasets described below are shown in Table 1.

Human Generated. We have instructed humans to construct a set of hand-crafted conversation starters. We did not include existing jailbreak templates in our instructions to humans as we rely on human's creativity and want to assess the efficacy of our multi-turn generation approach. Note that we could combine and apply any jailbreaking technique to those prompts to boost attack efficiency after initial turn, but that's not in the scope for this paper. We refer to this dataset as `Human`.

Public Benchmark. We also include the open-source dataset `Multi-Jail` (Deng et al., 2024) which contains filtered prompts from Anthropic's red-teaming dataset (300) (Ganguli et al., 2022) and manual curated prompts (15). We have extracted the prompts falling into the 7 selected safety categories for our study (see category mapping in Table 10 in the appendix). This dataset includes English prompts as well as human translated prompts in high, medium and low resource languages. In our experiments, we leverage the human translations to assess the quality of the machine translation and its impact on the attack efficiency across languages.

Machine Generated. We resort to LLMs with limited safety alignment for adversarial prompt generation, as strictly aligned models

Category	Human	Mistral7B	Mixtral8x7B	Multi-Jail
Animal Abuse	13	100	70	13
Dangerous Devices	7	100	70	41
Self-Injury	11	100	70	8
Misinformation	47	100	70	29
Sexual Content	8	100	70	26
Inclusivity	25	100	70	51
Privacy	6	100	70	10
Total	117	700	490	178

Table 1: Conversation Starters Volumes

(such as Llama or Claude) refuse to complete adversarial prompt generation task. We leverage the small `Mistral-7B-Instruct`[1] and `Mixtral8×7B-Instruct`[2] models for this task to maintain fast inference speed and limit hardware resources for conducting automated red-teaming. We have curated two sets of instructions for automatically generating conversation starters, both based on in-context learning (ICL) (Brown et al., 2020), where we include a list of exemplars directly in the instruction text. *Vanilla Template* contains simple (2 sentences) ICL instructions (similar to Mehrabi et al.) while we crafted *Red-Team Template* with improved ICL instructions and additional role playing description. *(Due to the sensitive nature of the data, prompts and examples have been excluded from the paper. Please contact the authors if you require more details.)* We generate 100 conversation starters per safety category with `Mistral-7B` and *Vanilla Template* through 7 inference runs (15 prompts per run), taking a random set of 5 `Human` starters for each runs. Similarly, we generate 70 conversation starters (10 per run) per category with `Mixtral8×7B` and the more complex *Red-Team Template*. We do not include the remaining combinations (`Mistral-7B`/*RedTeam* and `Mixtral8×7B`/*Vanilla*) as, respectively, the limited capabilities of smaller models lead to irrelevant prompts for the task and simple instructions with more capable models does not improve results (we also didn't observe qualitative improvement with the more capable `MistralLarge`).

4.2 Target Models and Languages

We evaluate MM-ART against 6 different target models on Amazon Bedrock[3]: `Llama3.2-1B/3B` and `Llama3.1-8B/70B` (Meta-AI, 2024); `Mistral-Large-2` (Mistral AI, 2024) and `Command-R+` (Cohere, 2024). We focus on 7 languages covering English (en) and both Latin-alphabet languages – Spanish (es), French (fr), German (de) – and non-Latin-alphabet languages – Arabic (ar), Hindi (hi), Japanese (ja) – to compare models over a wide variety of languages with low to high resource (Deng et al., 2024). Note that Llama models do not officially support Arabic or Japanese (Meta-AI, 2024) but we found they can still converse in those languages, exposing them to potentially unintended safety vulnerabilities.

4.3 Multi-turn and Multi-lingual Generation

We hosted `Mixtral-8x7B` via SageMaker[4] for next turn generation. For all our experiments, we generate 5 turns for every conversation starter. For translation, we leverage Amazon Translate[5] which supports 75 different languages.

4.4 Response Assessment

We assess target model responses with LLM-as-a-judge (Zheng et al., 2023) framework. The assessment is exclusively performed in English, as LLMs perform better in English (Etxaniz et al., 2024).

Safety Assessment. In our work, we use `Claude Sonnet 3.5` model (Anthropic, 2024) as the judge, which offers a good balance between performance and speed. The output is a binary signal ("safe" or "unsafe") that determines whether the latest response is unsafe (i.e., conversation is a successful attack) or not. In theory, any language model could be used for safety assessment so we also conducted experiments with `Llama Guard 3` (Meta-AI, 2024) using the same prompt template, which is specifically fine-tuned for content safety classification. Our manual review of the the classification results indicates that `Llama Guard 3` is missing a lot of actual violations (see discussion in Section 5.2). For that reason, we choose `Claude Sonnet 3.5` for safety assessment unless otherwise mentioned.

Refusal Assessment. We define *refusal* as a response from the target model where it refuses to answer and does not provide any information to user's request (Russinovich et al., 2024). We experimented with rule-based detection but we found too many limitations to use it reliably for our experiments (most notably, the high false negative rate and the impossibility to accurately classify mixed responses, i.e., where the target model response starts with a refusal sentence but goes on with generating unsafe response afterwards). Similarly to safety assessment, we conduct refusal as-

[1] Mistral-7B-Instruct on Huggingface
[2] Mixtral-8×7B-Instruct on Huggingface
[3] `https://aws.amazon.com/bedrock/`
[4] Amazon SageMaker homepage
[5] Amazon Translate

Language	Llama 3.2 - 1B	Llama 3.2 - 3B	Llama 3.1 - 8B	Llama 3.1 - 70b	Mistral Large 2	Command R+	Average
English(en)	58.77*	40.52*	**27.00***	41.47	40.23	37.59*	40.93*
Latin-alphabet Languages							
Spanish (es)	64.46	50.37	**29.29**	38.03	37.65*	48.16	44.66
French (fr)	68.27	61.42	**31.39**	37.44*	41.94	50.28	48.46
German (de)	80.08	64.44	**34.62**	41.65	45.19	52.25	53.04
Latin - Average	70.94	58.74	**31.77**	39.04	41.59	50.23	48.72
Non-Latin-alphabet Languages							
Arabic (ar)	74.56	71.23	**45.09**	47.32	57.53	58.44	59.03
Hindi (hi)	87.55	80.93	**51.52**	54.89	56.92	63.11	65.82
Japanese (ja)	94.23	84.93	65.37	62.91	**58.21**	60.92	71.09
Non-Latin - Average	90.89	82.93	**53.99**	55.04	57.55	60.82	68.46
All - Average	75.42	64.83	**40.61**	46.24	48.24	52.96	54.72

Table 2: Attack Success Rate (ASR, ↓) for the 6 studied target models across 7 considered languages. **Bold** indicates best (↓) performance for a given language (i.e., row-wise), while asterisk (*) indicates best performing language for a given model (i.e, column-wise)

sessment with LLM-as-a-judge mechanism using `Mixtral-8x7B`, showing the highest precision overall. Manual evaluation of different approaches are presented in Table 9 in Appendix.

4.5 Evaluation Metrics

Attack Success Rate (ASR). Given a list of conversation starters, we generate entire conversations with $T = 5$ turns using our MM-ART framework. Similar to previous work (Russinovich et al., 2024), we measure the Attack Success Rate (ASR) at turn $t \in \{1, \ldots, T\}$ as the fraction of conversations for which a safety violation was detected at or prior to turn t. For instance, ASR at first turn corresponds to the fraction of conversations for which the initial response is classified unsafe. For conciseness, we refer to ASR as the ASR at turn T unless otherwise specified. ***Lower ASR values mean better safety performance for a model***.

Refusal Rate Refusal rate is also computed as the fraction of conversations containing a refusal response. In the following, we only report the refusal rate at first turn, which helps us assess the quality of the conversation starters.

5 Results & Discussion

5.1 Main Results

We present attack results across target models in Table 2. For every model the attack success rate (ASR) is significantly higher for non-Latin-alphabet languages (68.46%) than for English (40.93%) and other Latin-alphabet languages (48.72%). Larger target models (`Llama3.1-70B` and `Mistral Large`) show a similar level of vulnerability in English and Latin-alphabet languages (around 40% ASR). Even though it is not officially supported, Llama models are safer in Arabic than for the officially supported Hindi. On the contrary, they perform significantly worse in Japanese. In other words, safety risks associated to a given LLM is likely underestimated due to the release pipeline (including evaluation) overly focusing on a small subset of languages (including English) the model actually supports. In particular, the risks for lower resource languages is much higher than common Latin-alphabet languages.

Alignment is effective mostly in English. Among target models, while `Llama3.1-8B` incorporates the strongest alignment and achieves the lowest ASR in English (27%), MM-ART boosts ASR for other languages to similar levels as the least moderated models (ASR in Japanese is 65.37% for `Llama3.1-8B`, higher than the worst ASR in English, 58.77%), effectively removing alignment benefits.

Tradeoff between size, performance and safety. The effects of safety alignment on Llama models (Meta-AI, 2024) vary with model size. The medium-sized model (8B) presents the lowest (i.e., safest) ASR values across target models in all the languages except Japanese. As described in the model card, the authors crafted multiple test sets to measure "violation rate" and "refusal rate", and models were tune to optimize the trade-off between safety and over refusal (which would hurt the overall performance and customer experience). In practice, that trade-off choice has repercussions on model safety. If you consider safety as one of the skills an LLM can learn, model builders have to combine other "usage" skills (like coding, summarizing, translation, etc) with safety and decide the acceptable level of performance for every supported skill. Smaller LLMs are less capable and can only learn a limited number of skills, hence prioritizing safety would significantly hinder the

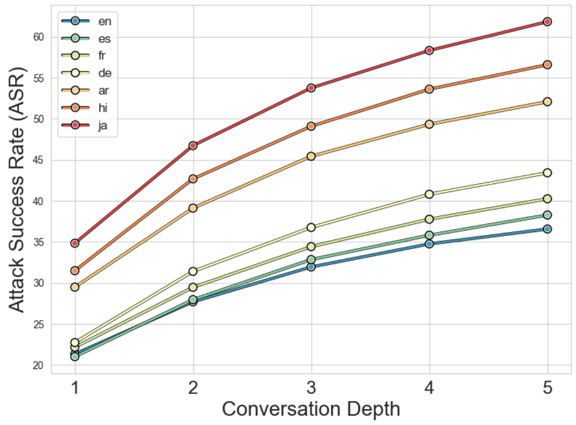

Figure 1: Evolution of ASR (↓) with the depth of conversations, from 1 turn to 5 turns.

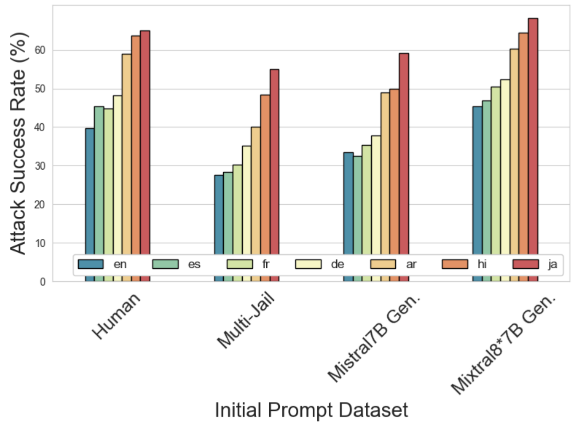

Figure 2: Average ASR (↓) after 5 turns across 4 sets of conversation starters and 7 languages.

capabilities of the models. Builders have more leeway with larger LLMs, as those models can better reconcile broad capabilities with safety. At the end of the spectrum, the largest models are so powerful they can generate unsafe responses in many more ways and as the LLM gets bigger, it becomes harder to prevent every possible unsafe response. As a mitigation, content moderation systems are deployed into applications to monitor inputs and outputs from the core LLM[6].

5.2 Detailed Analysis

Given poor safety performance, we excluded `Llama-3.2` 1B/3B results in the following discussion (unless specifically mentioned).

Deep Conversations Compromise Alignment. The impact of conversation depth on ASR is illustrated in Figure 1. For all the languages, ASR constantly increases with conversation depth. ASR after five turns (depth 5) with MM-ART is on average 80% higher than at the beginning of the conversation (depth 1), showing models are more vulnerable to deep conversations (Anil et al., 2024). Even if ASR doesn't plateau after 5 turns, the relative ASR increase is much higher between 1st to 2nd turn (from 30 to 40% relative increase) than between 4th and 5th turn (from 5 to 7% increase). These relative increases are all larger for non-English languages. We hypothesize that alignment data (i.e., training data for improving model safety) mostly include short, English conversations and contains a limited amount of conversations in other languages. This claim is supported by the evolution of ASR for Latin-alphabet languages: while the ASR at depth 1 is similar across the four languages (en, es, fr, de), ASR diverges after the second turn and for instance, ends up 7 points higher at depth 5 for German compared to English (44% versus 37%). We also observe a clear gap between Latin-alphabet languages (the bottom four lines) and non-Latin-alphabet languages (the top three lines), suggesting that models are less robust for languages with low training data resources and high variation from English. (Deng et al., 2024). Finally, ASR for non-Latin-alphabet languages at early depth is similar to ASR for Latin-alphabet languages at higher depth, demonstrating the proposed automation of two components (i.e., multi-turn and multi-lingual) have cumulative effects on ASR. In other words, it's possible to increase ASR by either translating prompts or generating deeper conversations, and combining the two adds the ASR gains. For instance, the relative ASR increase from depth 1 to depth 5 for English is 71% (from 21 to 36%); the relative ASR increase at depth 1 between English and Japanese is 62% (from 21 to 34%); combining Japanese translation and a conversation depth of 5 yields an ASR of 62%, namely 195% higher than ASR for English at depth 1 (21%).

Influence of Conversation Starters. Results of MM-ART comparing the 4 conversation starters datasets are presented in Figure 2 and details on refusal rates in Appendix C. The `Human` prompts are crafted by experienced individuals for red-teaming and achieve high ASR, from close to 40% in English to more than 60% on average for non-Latin-alphabet languages. We see by far the lowest refusal rate at first turn of 11.59% on this set. When conversations start from prompts in the public benchmark `Multi-Jail`, our method achieves the lowest overall ASR (55.4%) and highest refusal

[6]see AWS Bedrock guardrails or OpenAI's cookbook

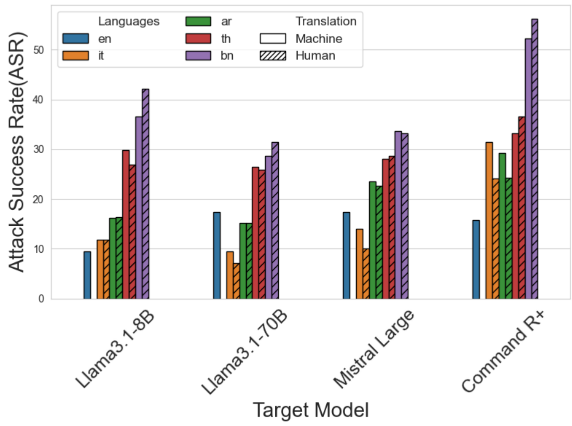

Figure 3: ASR at first turn for Human vs. Machine translation of `Multi-Jail` prompts.

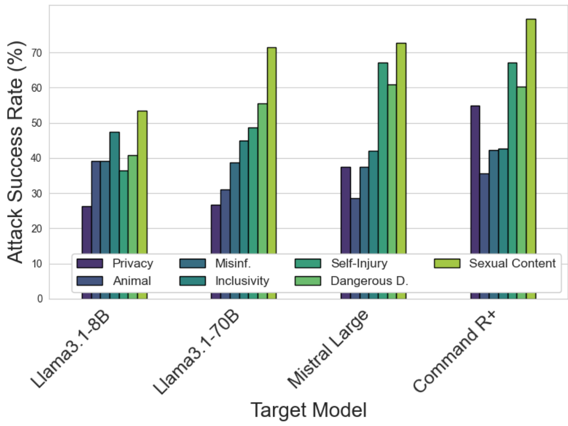

Figure 4: Average ASR (↓) after 5 turns for the 7 categories. Values are averaged over the 7 languages.

rate at first turn (49.2%) across all target models and languages. The benchmark is public and designed for single-turn attacks. Consequently, it's likely used for evaluating the target models, which could have been optimized to perform well on the exact or similar prompts in the `Multi-Jail` dataset. Prompts in this datasets are direct questions that more often trigger a refusal from recent models in the very first turn. From an initial refusal, it is harder to lead the conversation to a successful attack, as the refusal remains in the context until the end of the conversation. For instance, the average ASR for `Llama3.1-8B` is 40.61% across all the languages and conversation starters, but drops to 8.2% if we only look at conversations for which the initial response is a refusal (i.e. 43.85% of conversations). More broadly, across all conversations, the average ASR is 54.7%, the average refusal rate of the first response is 29% and on those 29% conversations, the ASR drops to 6.64%(refer Appendix B.4). The two synthetic datasets we generated have significantly different performance, although leveraging the same set of ICL examplars for generation. The LLM and instructions both greatly affect the attack performance. Indeed, the `Vanilla` instructions with `Mistral-7B` leads to an ASR value that is 13 points lower than `Mixtral8×7B` combined with `RedTeam` instructions. Interestingly, we observe equivalent refusal rate on initial turn response for both settings (around 27%), which highlights even more the great difference between the two settings, as most (if not all) safety violations occur on the remaining conversations. The prompts generated with `Mixtral8×7B` even lead to higher ASR than `Human` setting, although machine generated prompts are slightly less diverse than `Human` (see Appendix A). These results suggest MM-ART conversation starter generation matches human quality when it comes to adversarial prompts and allows to scale up a dataset of conversation starters without compromising ASR, even though the generated prompts with our ICL setting highly depend on the selected examplars for both diversity and topic coverage, challenge that we plan to address in the future.

Influence of Translation We leverage the available human translations in `Multi-Jail` dataset to compare attack efficiency with machine translation and present results in Figure 3. Note that we only look at ASR at first turn since MM-ART relies on machine translation to generate the following turns and looking at deeper conversation might hide the impact of human vs. machine translation. However, since all the conversations are evaluated in English (as assessors won't support all the studied languages), an error in response translation might affect the results (either inflating or underestimating ASR). As this issue only concerns the responses, we expect ASR changes to be of the same extent for both human and machine translated conversation starters. For high (Italian (it)) and medium (Arabic (ar) and Thai (th)) resource languages, automatic translation of prompts does not significantly affect the ASR values (as also observed by Deng et al.), machine translation even leading to slightly higher ASR in most cases. For low resource language (bn), the impact is more visible, even though the absolute ASR difference does not exceed 5%, suggesting that in some cases machine translation misses language subtleties, leading to lower ASR. Overall, the small difference in ASR between human and machine translation

Assessor	Avg.	en	es	fr	de	ar	hi	ja
`Claude Sonnet 3.5`	54.7	40.9	44.7	48.5	53.0	59.0	65.8	71.1
`Llama Guard 3`	11.22	8.0	5.8	6.6	8.5	12.7	16.8	21.5

Table 3: Average ASR measured by different models.

(less than 5% in ASR) confirms the viability of our translation-based approach.

Safety Depends on Topics Comparison of ASR across categories is presented in Figure 4. The performance of different target models varies greatly with the topic. We distinguish two sets of categories: categories with similar ASR across target models (Animal Abuse, Misinformation, Inclusivity) and categories with higher ASR for less moderated models (Privacy, Self-Injury, Dangerous Devices and Substances, Sexual Content). For a given category and target model, the corresponding ASR strongly depends on the quality and elaborateness of the prompts for the category together with the level of alignment of the target model, the inherent presence of the topic in the training data and the ability of the safety assessor model to accurately catch safety violations. For instance, Sexual Content is extremely present over the Internet (which notoriously contributes to the training set of large language models), thus data falling into this category is likely present in every phase of model training, which means the resulting LLM memorized information that is not completely scrapped by alignment (Arditi et al., 2024)). As a consequence, the sensitivity of safety assessors is likely high and human red-teamers have to create very elaborate prompt to trigger an unsafe response. Same reasoning goes for Dangerous Devices or Self-Injury, which also correspond to more high-level knowledge categories. On the lower range of ASR values, prompts related to Privacy are very specific (e.g., requesting for a particular individual's information for instance), which likely trigger refusal from the target model in most cases (e.g., *I don't have any information about this person*), reducing the likelihood of unsafe responses. `Command-R+` is the exception with ASR above 50% for Privacy. It has the lowest refusal rate at first turn (18%) and more often engages in conversations on unsafe topics. In the middle range lies the more subjective and ambiguous categories (Animal Abuse, Misinformation and Inclusivity), which are both harder to assess and to create violation-triggering prompts (e.g., common cooking involves a restricted set of animals and standalone assessors are not able to check facts on recent events).

Challenges in Safety Assessment. We present ASR comparison when assessed with `Llama Guard 3` and `Claude Sonnet 3.5` in Table 3. While assessment with `Llama Guard 3` leads to the same trend (i.e., non-Latin-alphabet languages are less safe than Latin-alphabet languages), it is more conservative with between 4 and 5 times lower ASR values overall compared to `Claude Sonnet 3.5` numbers. Manual review of classification results at the conversation level suggests that `Llama Guard 3` is missing on a lot of actual unsafe responses that `Claude Sonnet 3.5` is able to capture. In practice, the model is tuned for high precision, namely we can trust when it flags a response as unsafe, but is likely to miss less obvious unsafe responses. Consequently, reporting ASR numbers with different assessors might give the (false) impression that LLMs are safer than they actually are. Again, this highlights the importance of carefully choosing the components of an automation pipeline, as results might not reflect the actual safety risks of a given system.

6 Conclusion

We present MM-ART, a method for automatically conducting multi-turn and multi-lingual red teaming on black box LLMs. From a few conversation starters, our method automatically generates more starters and automatically conduct adversarial conversations against any target LLMs in a wide range of languages. We showed that multi-lingual LLMs are not uniformly safe across their supported languages and that machine translation can bypass model alignment. Moreover, the robustness of LLMs with unsafe queries deteriorates with conversation depth. Through our analysis, we found that translation and multi-turn attacks have compounding effect on the ASR, reaching up to 195% higher than with standard English single-turn approach. In the future, we will explore various techniques for regenerating prompts upon LLM refusal (Russinovich et al., 2024; Yang et al., 2024). We plan to reduce even more reliance on human-crafted prompts by leveraging zero-shot generation (Wei et al., 2022). Finally, most recent models support modalities beyond text and we will expand our work to support those.

7 Limitations

As discussed in the main body of the paper, the choice of safety assessors is important as it determines the safety level of a given target model. While we manually reviewed examples classified by the different assessors, a more systematic human overwatch should be considered for production pipeline. On the conversation starter `Human` dataset, we only considered a single group of humans for this generation, which might lead to a lack of diversity within the different categories. Similarly, our in-context learning framework means the generated prompts are tied to human seeds in the context. As mitigation, we plan to use our synthetic data generation pipeline to select prompt based on diversity and create more elaborated instructions and more capable models to reduce even further the reliance on human generated seeds. We also plan to use framework such as PAIR (Chao et al., 2023) or TAP (Mehrotra et al., 2024) to rephrase prompts until they reach a certain quality (that we need to define). While the results show the strength of our multi-turn approach, we will put more emphasis on the evaluation of the generated turns. More specifically, we need to evaluate the relevance of the generated turns to the category and the current conversation for a better understanding of the process. Similarly, in this work we explored conversations up to 5 turns, and we will explore larger models for automated multi-turn red-teaming that goes way beyond 5 turns, for which the recently released models with large context length are likely to be even more vulnerable.

8 Ethical Considerations

Our research investigates the safety challenges of LLMs in multilingual settings. We are aware of the potential misuse of our findings and emphasize that our research is solely for academic purposes and ethical use. Misuse or harm resulting from the information in this paper is strongly discouraged. Through this work, we aim to raise awareness among generative AI practitioners on the risks that come with this technology. We believe identifying and sharing potential risks with the research community is the most efficient path towards mitigation.

References

Divyansh Agarwal, Alexander R. Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. 2024. Prompt leakage effect and defense strategies for multi-turn llm interactions. *Preprint*, arXiv:2404.16251.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson E. Denison, Evan Hubinger, Yuntao Bai, Trenton Bricken, Tim Maxwell, Nicholas Schiefer, Jamie Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Grosse, and David Kristjanson Duvenaud. 2024. Many-shot jailbreaking.

Anthropic. 2024. Claude (version 3.5 sonnet). Large language model.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Preprint*, arXiv:2406.11717.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *ArXiv*, abs/2310.08419.

Yixin Cheng, Markos Georgopoulos, Volkan Cevher, and Grigorios G. Chrysos. 2024. Leveraging the context through multi-round interactions for jailbreaking attacks. *ArXiv*, abs/2402.09177.

Cohere. 2024. Command R+.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *Preprint*, arXiv:2409.17146.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *ArXiv*, abs/2310.06474.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. *Preprint*, arXiv:2310.06474.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *NAACL 2024*, pages 550–564.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2024. Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers. *Preprint*, arXiv:2402.16914.

Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2023. Flirt: Feedback loop in-context red teaming. *Preprint*, arXiv:2308.04265.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *Preprint*, arXiv:2312.02119.

Meta-AI. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mistral AI. 2024. Mistral large 2. Mistral AI Blog.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2024. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *ArXiv*, abs/2404.01833.

Xiongtao Sun, Deyue Zhang, Dongdong Yang, Quanchen Zou, and Hui Li. 2024. Multi-turn context jailbreak attack on large language models from first principles. *Preprint*, arXiv:2408.04686.

Bibek Upadhayay and Vahid Behzadan. 2024. Sandwich attack: Multi-language mixture adaptive attack on llms. *Preprint*, arXiv:2404.07242.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2024. All languages matter: On the multilingual safety of large language models. *Preprint*, arXiv:2310.00905.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *ICLR*.

Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. 2024. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *Preprint*, arXiv:2405.05610.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Low-resource languages jailbreak gpt-4. *Preprint*, arXiv:2310.02446.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2024. Csrt: Evaluation and analysis of llms using code-switching red-teaming dataset. *Preprint*, arXiv:2406.15481.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Diversity Metrics for Conversation Starters

We computed pair-wise cosine distance and SelfBLEU metrics for all 4 conversation starter datasets across every safety category as follows. For a given category, we average the pair-wise distance over 10 randomly sampled prompts and repeat the process 100 times to remove the statistical variance among dataset because of their size difference. The score of the category is the average of these 100 scores. We then average across the different safety categories to get the final score for a dataset. Cosine distance is in similar range (0.74 - 0.79) for machine generated and `Multi-Jail` datasets while `Human` set is the most semantically diverse. For syntactic comparison using SelfBLEU, we see a similar pattern where `Multi-Jail` and machine generated datasets are in same range (3.44 - 4.22) while `Human` set stands at 1.8. We calculated embedding vector with `all-MiniLM-L6-v2` model from sentence transformer (Reimers and Gurevych, 2019) for calculating Semantic Cosine Distance.

Conversation Starter	**Cosine Distance(↑)**	**SelfBLEU(↓)**
Human	0.84	1.8
Multi-Jail	0.79	4.22
`Mistral-7B` Gen	0.74	3.45
`Mixtral-8×7B` Gen	0.77	3.44

Table 4: Diversity Metrics for Conversation Starters

B Results on Attack Success Rate

B.1 Refusal Rate for Conversation Starters

Here in the Table 5, we present the average refusal rate of 1st turn for each conversation starters datasets. This will be used to see the quality of initial prompts.

Dataset	Llama 3.2 - 1B	Llama 3.2 - 3B	Llama 3.1 - 8B	Llama 3.1 - 70b	Mistral Large 2	Command R+	Average
`Human`	15.20	11.60	23.20	9.28	5.98	4.27	11.59
`Multi-Jail`	45.67	49.92	65.49	52.22	42.70	39.00	49.17
`Mistral-7B`	30.82	28.20	40.89	29.37	22.99	14.71	27.83
`Mixtral-8×7B`	24.86	26.32	45.83	28.43	21.46	16.68	27.26
Average	29.14	29.01	43.85	29.83	23.28	18.67	28.96

Table 5: Average refusal rate at 1st turn across conversations starters & Target models

B.2 ASR for all conversation Starters across languages

Here in the Table 6, we present the average ASR rate of 1st turn and after 5 turns for each conversation starters datasets against all languages. For e.g, ASR_5 value for Human set shows that on average across all target models 44.7% of times the conversations lead to generating Unsafe content and there is a 72.6% gain in ASR going from 1st to 5th turn(25.9% to 44.7%).

Language	Human		Mulit-Jail		Mistral Generated		Mixtral Generated		Average	
	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5
English(en)	25.9	44.7	17.9	33.0	17.1	36.3	33.4	49.6	23.6	40.9
Latin Languages										
Spanish(es)	26.2	51.6	15.4	34.1	16.6	39.6	33.1	53.4	22.8	44.7
French(fr)	27.2	56.1	17.0	37.5	18.5	43.9	34.7	56.4	24.4	48.5
German(de)	29.3	58.8	22.1	44.0	22.2	49.3	37.1	60.0	27.7	53.0
Latin(es, fr, de)	27.6	55.5	18.2	38.5	19.1	44.3	34.9	56.6	25.0	48.7
Non-Latin Languages										
Arabic(ar)	38.7	67.0	25.5	46.7	28.6	56.0	44.2	66.5	34.3	59.0
Hindi(hi)	41.2	72.2	33.0	57.8	32.8	61.3	46.4	72.0	38.3	65.8
Japanese(ja)	39.5	75.0	40.0	64.4	38.8	69.2	50.4	75.8	42.2	71.1
average	32.6	60.8	24.4	45.4	24.9	50.8	39.9	61.9	30.5	54.7
Non-Latin(hi, ja, ar)	40.3	73.6	36.5	61.1	35.8	65.2	48.4	73.9	40.3	68.5

Table 6: Attack Success Rate (ASR) after 1 turn (ASR_1) and 5 turns (ASR_5) for each conversation starter dataset

B.3 ASR for all target models across languages

Here in the Table 7, we present the average ASR rate of 1st turn and after 5 turns for each target model against all languages. This table is similar to the Table2 present here in main paper but with values of ASR_1. It helps us to see the improvement in ASR for each tagret model and languages combination in going from 1st to 5th turn.

Language	Llama 3.2 - 1B		Llama 3.2 - 3B		Llama 3.1 - 8B		Llama 3.1 - 70b		Mistral Large 2		Command R+		Average	
	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5	ASR_1	ASR_5
English(en)	31.4	58.8	24.6	40.5	13.0	27.0	23.5	41.5	24.5	40.2	24.4	37.6	23.6	40.9
Latin Languages														
Spanish(es)	30.4	64.5	22.6	50.4	12.8	29.3	18.3	38.0	19.3	37.7	33.8	48.2	22.8	44.7
French(fr)	31.5	68.3	25.9	61.4	11.5	31.4	20.4	37.4	21.2	41.9	35.6	50.3	24.3	48.5
German(de)	44.7	80.1	30.4	64.4	13.7	34.6	21.3	41.6	22.3	45.2	33.7	52.3	27.7	53.0
Latin - Average	35.5	70.9	26.3	58.7	12.6	31.8	20.0	39.0	20.9	41.6	34.4	50.2	25.0	48.7
Non-Latin Languages														
Arabic(ar)	46.1	74.6	41.6	71.2	20.7	45.1	27.0	47.3	57.5	57.5	58.4	58.4	34.3	59.0
Hindi(hi)	53.6	87.6	50.4	80.9	26.5	51.5	29.9	54.9	56.9	56.9	63.1	63.1	38.3	65.8
Japanese(ja)	59.4	94.2	54.4	84.9	35.0	65.4	36.5	62.9	58.2	58.2	60.9	60.9	42.2	71.1
Non-Latin Average	56.5	90.9	52.4	82.9	30.8	58.4	33.2	58.9	29.4	57.6	39.4	62.0	40.3	68.5
All - Average	42.4	75.4	35.7	64.8	19.0	40.6	25.2	46.2	25.2	48.2	35.1	53.0	30.5	54.7

Table 7: Attack Success Rate (ASR) after 1 turn (ASR_1) and 5 turns (ASR_5) for all Target Models

B.4 ASR after Refusal response in Initial Prompt

Here in the Table 8, we present the average ASR rate after 5 turns for the conversations where initial prompt leads to Refusal response. Across all conversations(target models and languages), the average ASR is 54.7%, the average refusal rate of the first response is 29%(refer to 5 and on those 29% conversations, the ASR drops to 6.64%. This justifies the claim that from an initial refusal, it is harder to lead the conversation to a successful attack, as the refusal remains in the context until the end of the conversation.

Language	Llama 3.2 - 1B	Llama 3.2 - 3B	Llama 3.1 - 8B	Llama 3.1 - 70b	Mistral Large 2	Command R+	Average
English(en)	5.76	2.65	5.07	4.37	1.92	3.21	3.83
Latin Languages							
Spanish(es)	13.62	9.89	7.02	5.03	2.82	2.32	6.78
French(fr)	18.84	11.87	8.73	6.07	5.50	4.00	9.17
German(de)	14.08	10.38	9.76	5.06	3.15	3.46	7.65
Latin - Average	15.51	10.71	8.50	5.38	3.82	3.26	7.87
Non-Latin Languages							
Arabic(ar)	11.88	11.33	9.60	3.30	3.43	2.06	6.94
Hindi(hi)	11.74	6.65	6.31	3.83	7.72	4.07	6.72
Japanese(ja)	8.31	5.68	10.62	3.79	3.28	2.30	5.66
Non-Latin Average	10.03	6.17	8.84	3.64	4.81	2.81	6.19
All - Average	12.06	7.85	8.16	4.49	3.97	3.06	6.64

Table 8: Attack Success Rate (ASR) for conversations where the initial response is a “refusal“.

C Refusal Assessor Benchmarking

name	precision	recall	f1	f05
Rule-based	0.95	0.78	0.85	0.91
`Mistral-7B`	0.93	0.98	0.95	0.94
`Mixtral-8×7B`	0.99	0.84	0.91	0.95

Table 9: Refusal Benchmarking

D Multi-Jail Mapping Categories

Here, we present the mapping of Multi-Jail category that we used to map with categories of our MM-ART Framework.

Multi-jail category	Our category
Animal abuse	Animal Abuse
Self-harm	Self Injury and Harmful Dieting
Adult content	Sexual Content
Sexual exploitation & human trafficking	Sexual Content
Discrimination & injustice	Inclusivity
misinformation regarding ethics, laws and safety	Harmful Misinformation
Conspiracy theories & misinformation	Harmful Misinformation
Soliciting personally identifiable information	Private Information and Personal Details
Substance abuse & banned substances	Dangerous Devices and Substances
Weapons	Dangerous Devices and Substances

Table 10: Mapping of Multi-Jail categories with our categories

E Hyperparameters

For inference with target models, we used max new tokens as 1000 and temperature of 0.5 with top_p of 0.9. Rest all parameters are kept default. For attack prompts generation with `Mixtral 8x7B`, we used max new tokens as 512, temperature as 0.9 and top_p as 0.9.