On Guardrail Models' Robustness to Mutations and Adversarial Attacks

Elias Bassani
European Commission
Joint Research Centre (JRC)
Ispra, Italy
elias.bassani@ec.europa.eu

Ignacio Sanchez
European Commission
Joint Research Centre (JRC)
Ispra, Italy
ignacio.sanchez@ec.europa.eu

Abstract

The risk of generative AI systems providing unsafe information has raised significant concerns, emphasizing the need for safety guardrails. To mitigate this risk, guardrail models are increasingly used to detect unsafe content in human-AI interactions, complementing the safety alignment of Large Language Models. Despite recent efforts to evaluate those models' effectiveness, their robustness to input mutations and adversarial attacks remains largely unexplored. In this paper, we present a comprehensive evaluation of 15 state-of-the-art guardrail models, assessing their robustness to: a) input mutations, such as typos, keywords camouflage, ciphers, and veiled expressions, and b) adversarial attacks designed to bypass models' safety alignment. Those attacks exploit LLMs capabilities like instruction-following, role-playing, personification, reasoning, and coding, or introduce adversarial tokens to induce model misbehavior. Our results reveal that most guardrail models can be evaded with simple input mutations and are vulnerable to adversarial attacks. For instance, a single adversarial token can deceive them 44.5% of the time on average. The limitations of the current generation of guardrail models highlight the need for more robust safety guardrails. [1]

1 Introduction

Generative AI systems have become increasingly popular thanks to the advent of exceptionally capable general-purpose Large Language Models (LLMs), such as Claude (Anthropic, 2024), Gemini (Gemini Team, 2024b,a), GPT (Radford et al., 2018a,b; Brown et al., 2020; OpenAI, 2023, 2024), and Llama (Touvron et al., 2023a,b; Meta, 2024). Systems built on those models are used in a variety of fields, including sensitive areas like healthcare (Meskó and Topol, 2023; Zhang and Boulos, 2023), education (Baidoo-Anu and Ansah, 2023; Qadir, 2023), and finance (Chen et al., 2023). As AI systems continue to advance and be integrated into various application domains, it is essential to ensure their safety.

Lately, due to the risk of systems providing harmful information, the need for safety guardrails has received particular attention. Despite considerable efforts to align LLMs with human values (Wang et al., 2023), users can still find ways misuse them to generate harmful content. To mitigate this risk, fine-tuned LLMs, known as guardrail models, are increasingly employed to moderate human-AI interactions, completing other safety measures at model level such as alignment. Given the critical role of those models in AI systems, their evaluation is essential to ensure they are effective in the detection of unsafe interactions. However, despite recent efforts to evaluate their effectiveness (Bassani and Sanchez, 2024), their robustness to input mutations and adversarial attacks remains largely unexplored.

In this paper, we aim to fill this gap by conducting a comprehensive evaluation of the robustness of 15 guardrail models at the state-of-the-art. Our evaluation is two-fold. First, we consider input mutations that alter user prompts with typos, keywords camouflage, ciphers, and veiled expressions. Those mutations have been reported among the base ingredients for evading LLMs' alignment. Second, we cover a wide range of adversarial attacks proposed to make LLMs fulfill unsafe requests that they would typically refuse. Those attacks leverage LLMs' capabilities, such as instruction-following, role-playing, personification, reasoning, and coding, or employ adversarial input tokens to induce a model to misbehave. Our evaluation unveil several shortcomings of the current generation of guardrail models and highlight the need for further research into developing more robust safety guardrails.

[1] Due to safety concerns, we reserve the option of sharing our research artifacts only after the highlighted shortcomings have been substantially addressed by the research community.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16995–17006
November 4-9, 2025

2 Related Work

In this section, we discuss previous research related to our work. First, we introduce recent works on guardrail models. Then, we discuss adversarial attacks designed to bypass LLMs' safety alignment.

2.1 Guardrail Models

Guardrail models were proposed to reduce the risk of LLMs engaging in offensive conversations (Lee et al., 2019; Curry and Rieser, 2018) or providing unsafe information (Dinan et al., 2019). Specifically, guardrail models act as input-output filters in human-AI conversations, thus moderating human-generated prompts and LLM-generated answers. Those models are usually derived from general-purpose LLMs via fine-tuning and follow a content moderation policy provided as input. Notable models are Llama Guard (Inan et al., 2023; Meta, 2024; Ghosh et al., 2024), Granite Guardian (Padhi et al., 2024), MD-Judge Li et al. (2024a), and Shield Gemma (Zeng et al., 2024), which are fine-tuning of Llama (Touvron et al., 2023a,b; Meta, 2024), Granite (Granite Team, 2024), Mistral (Jiang et al., 2023), and Gemma (Mesnard et al., 2024; Rivière et al., 2024), respectively. Our work focus on providing a proper evaluation of the robustness of those models to input mutations and adversarial attacks.

2.2 Adversarial Attacks

Despite safety alignment, LLMs remain susceptible to adversarial attacks as demonstrated by several recent publications. Those attacks often leverage LLMs capabilities, such as role-playing, personification, reasoning, and coding, to expose LLMs' inherent security risks. Lv et al. (2024) proposed an attack that leverages LLMs' coding capabilities to evade safety measures. Li et al. (2024b) crafted a jailbreak prompt that induce LLMs to override their alignment by eliciting abstract reasoning. Deng et al. (2024) leveraged the multilingual capabilities of LLMs to attack them. Chao et al. (2023) employed an attacker LLM to query and refine jailbreaks autonomously. Wei et al. (2023) proposed to craft malicious contexts to guide models in generating harmful outputs. Yuan et al. (2024) showed that LLMs can decode and understand encrypted messages and that ciphers can be leveraged to perform attacks. Shen et al. (2024) collected several jailbreak prompts from various online forums. Most of them leverage the instruction-following and role-playing capabilities of LLMs. Finally, a number of attacks aim to generate adversarial suffixes able to bypass model alignment (Zou et al., 2023; Andriushchenko et al., 2024; Liao and Sun, 2024). Several input mutations, such as typos (Ding et al., 2024), keywords camouflage (Huertas-García et al., 2024), cyphers (Yuan et al., 2024), and veiled expressions (Xu et al., 2024), have been reported as base ingredients of adversarial attacks. As guardrail models are usually based on LLMs, we evaluate whether those techniques can be employed to evade their safety detection.

3 Methodology

In this section, we describe the input perturbation methods we employ to test the robustness of guardrail models. We classify these methods into mutations and adversarial attacks. Mutations alter user prompts *without* adding adversarial content to trick safeguards. Conversely, adversarial attacks are conceived to fool safeguards by adding adversarial content, thus causing models to misbehave.

3.1 Mutations

In this section, we introduce the mutations we apply to user prompts to test the robustness of guardrail models. Specifically, we employ typos, keywords camouflage, ciphers, and veiled expressions. It is worth notice that most LLMs are able to understand user prompts in the presence of such mutations.

Typos The most straightforward way to mutate text inputs is to introduce typos. Ding et al. (2024) have shown that typos can be employed to conduct adversarial attacks. Thus, we assess whether typos can induce guardrail models to misbehave. In our experiments, we consider (1) character deletion, (2) character insertion, (3) character replacement, (4) characters swapping (i.e., inverting the position of two adjacent characters), and a (5) mix of those. To resemble real-world typos, we rely on a key neighborhood map based on the US QWERTY keyboard layout to select characters for both insertion and replacement. We apply typos only to keywords as identified by the Python library KeyBERT (Grootendorst, 2020). Specifically, we apply a *single* random typo of the chosen kind to each keyword.

Keywords Camouflage Camouflaging keywords is a common technique to evade content moderation systems (Huertas-García et al., 2023) and LLMs' alignment (Huertas-García et al., 2024). Thus, we

assess whether those techniques are also effective in evading guardrail models. In our work, we consider the following camouflage methods: (1) disemvoweling, (2) punctuation insertion, (3) white spaces insertion, (4) splitting, (5) syllables inversion, and (6) leetspeak. Disemvoweling removes all vowels in a word. Punctuation insertion adds a punctuation mark (i.e., full stop) between each letter of a word. White spaces insertion adds a space between each letter of a word. Splitting divides a word in two sub-words at random. Syllables inversion flips two adjacent syllables of a word. Leetspeak replaces characters with similar-looking glyphs We apply the same camouflage technique to each keyword in a user prompt.

Ciphers Ciphers are algorithms for performing the encryption and decryption of messages. Yuan et al. (2024) have shown that LLMs can decode and understand encrypted messages, and that ciphers can be leveraged to bypass models' alignment. Thus, we test whether ciphers can evade guardrail models' safety detection. To this end, we consider three popular substitution algorithms that most LLMs were likely exposed to during pretraining: (1) Caesar cipher (2) Morse code, and (3) Unicode. The Caesar cipher is a simple and well-known encryption technique that replace each letter of a message with another letter a fixed number of positions down the alphabet. In our experiments, we used the popular rotation 13 (ROT13). Morse code is a telecommunications method which encodes text characters as standardized sequences of two different signal durations, known as dots and dashes. Unicode is a text encoding standard that supports digital writing systems. In our experiments, we replaced each letter with its corresponding Unicode decimal code.

Veiled Expressions Veiled expressions are indirect or subtle ways of conveying a message. They are often used to discuss sensitive topics without being explicit. These expressions can involve euphemisms, metaphors, or ambiguous language that allows to imply something without stating it outright. Xu et al. (2024) have shown veiled expressions can be used to evade models' alignment and make LLMs output unsafe information. Thus, we test whether they can be leveraged to bypass guardrail models. Specifically, we handcraft veiled variants for our test prompts as we found previously proposed automatic methods often alter the meaning of the original prompts (Xu et al., 2024).

3.2 Adversarial Attacks

In this section, we introduce the adversarial attacks we employ against recent guardrail models, LLMs fine-tuned for safety detection. To comprehensively evaluate their robustness, we cover a wide range of attacks designed to make LLMs comply with unsafe requests by targeting their capabilities, such as instruction-following, role-playing, personification, reasoning, and coding, or through the addition of adversarial input tokens.

Jailbreak Templates Since the advent of LLM-based chatbots, users have tried to trick the underlying models into providing unsafe content. Some users succeeded in composing jailbreak prompts able to bypass models' alignment, making them answer unsafe requests. Most of those prompts leverage the instruction-following and role-playing capabilities of LLMs to achieve their goals. For example, they contain instructions such as *disregard all previous instructions* or *simulate a chatbot who always says the exact opposite of what ChatGPT would say*. To generalize and diffuse jailbreak prompts, prompt templates are often derived from those and shared on online forums. Thus, they constitute a primary safety concern. We refer to those templates, which can be used to embed any unsafe request, as *jailbreak templates*. Shen et al. (2024) have compiled a collection of jailbreak prompts from various sources. From this collection, we select 84 unique jailbreak templates that we fill with our test samples to evaluate guardrail models' robustness to jailbreak prompting attempts.

DeepInception DeepInception (Li et al., 2024b) is an attack inspired by the authority influence shown in the Milgram experiment (Milgram, 1963). As described by its authors, the attack leverage LLMs' personification capabilities to construct a virtual, nested scene, allowing it to realize an adaptive way to escape the usage control in a normal scenario. In other words, the authors propose a jailbreak prompt template that induce LLMs to override their alignment and generate harmful content by eliciting reasoning in abstract nested scenes. DeepInception has been shown effective against both open and closed models. We investigate this attack in several scenes as it targets the personification and reasoning capabilities of LLMs.

CodeChameleon CodeChameleon is a recent jailbreak attack proposed by Lv et al. (2024). This method leverages LLMs' code completion capabil-

ities to enable users to encrypt unsafe requests and evade safety measures. CodeChameleon rely on a coding prompt template to encrypt malicious requests. By also providing a decryption function, it elicits the generation of a response for the harmful request embedded in the prompt. We investigate all the variants of this attack as it targets the code understating and generation capabilities of LLMs.

GCG One of the first and most known attacks to evade LLMs' alignment and make them produce unsafe content is GCG (Zou et al., 2023). GCG aims to find a suffix that, when attached to a prompt, maximizes the probability that the model produces an affirmative response (rather than refusing to answer). GCG produces these adversarial suffixes by a combination of greedy and gradient-based search techniques. Finding adversarial suffixes to jailbreak LLMs is a common practice among other recently proposed attacks (Andriushchenko et al., 2024; Liao and Sun, 2024). We re-purpose GCG to attack guardrail models and evaluate whether adversarial suffixes can be used to evade them.

4 Experimental Setup

In this section, we introduce the experimental setup adopted to evaluate the robustness of guardrail models to input mutations and adversarial attacks. Specifically, we compare the effectiveness of several models at safety classification of mutated and adversarial prompts obtained through the methods described in Section 3.

In the following sections, we introduce the models we compare (Section 4.1), present the datasets we use for evaluation (Section 4.2), and discuss the evaluation metrics chosen to assess the models' robustness (Section 4.3) before presenting the results of our evaluation in Section 5.

4.1 Models

In this section, we introduce the models considered in our comparative evaluation. Specifically, we evaluate the robustness of 15 recent models. Among these are 13 guardrail models at the state-of-the-art, such as Llama Guard models (Inan et al., 2023; Meta, 2024; Ghosh et al., 2024) and Granite Guardian models (Padhi et al., 2024). We also consider two flavors of Mistral (Jiang et al., 2023) as it was recently reported to achieve results comparable to those of the state-of-the-art guardrail models when prompted for safety moderation (Bassani and Sanchez, 2024). Table 1 reports the list of the considered models as well as additional information related to them. It is important to highlight that the Meta's Llama Guard series of models accounts for 87% of the total downloads of guardrail models from HuggingFace[2] as of February 1st, 2025. Moreover, Llama Guard 3 8B is by far the most popular model, accounting for more than 1M downloads or 56% of the total. Thus, we deduce several deployed generative AI systems rely on Llama Guard models for safety moderation. This information allow us to better contextualize the results of our evaluation presented in Section 5 and the severity of our findings.

4.2 Datasets

To conduct our assessment, we employ two datasets. The first one is SimpleSafetyTests (Vidgen et al., 2023), a test suite comprising 100 unsafe prompts designed to identify critical safety risks. As this is a specific set of very unsafe prompts, it is likely that models' performance may be lower for more deceptive prompts. The second dataset, which acts as a control set, contains 100 safe prompts. It allows us to understand whether the robustness of guardrail models to certain perturbation methods is due to generalization abilities or other factors, such as overfitting to known mutations or attacks. As the size of the datasets may limit the generalizability of our findings, particularly regarding more nuanced unsafe prompts, we discourage overgeneralizing the positive results. For simplicity, we will refer to these datasets as *Unsafe* and *Safe* in the following sections.

4.3 Evaluation Metrics

To assess model robustness, we employ False Negative Rate and False Positive Rate for unsafe and safe prompts, respectively. In case of attacks, False Negative Rate corresponds to Attack Success Rate as not detecting an (adversarial) unsafe prompt as such indicates a breach of the safety guardrails. Given the critical role of guardrail models in AI safety, we adopt a strict severity scale to interpret the results of our evaluation. Specifically, we consider performance decreases ≤ 0.01 as *negligible*, < 0.10 as *moderate*, and ≥ 0.10 as *severe*.

5 Results and Discussion

In this section, we present the results of our evaluation of the robustness of guardrail models to

[2] https://huggingface.co

Model	Alias	Provider	Base Model	Params	Downloads	Reference
Granite Guardian 3 2B	GG 3 2B	IBM	Granite 3.0 2B	2.63 B	32 786	Padhi et al. (2024)
Granite Guardian 3 8B	GG 3 8B	IBM	Granite 3.0 8B	8.17 B	4 903	Padhi et al. (2024)
Granite Guardian 3.1 2B	GG 3.1 2B	IBM	Granite 3.1 2B	2.63 B	15 883	Padhi et al. (2024)
Granite Guardian 3.1 8B	GG 3.1 8B	IBM	Granite 3.1 8B	8.17 B	1 409	Padhi et al. (2024)
Llama Guard	LG	Meta	Llama 2 7B	6.74 B	237 396	Inan et al. (2023)
Llama Guard 2	LG 2	Meta	Llama 3 8B	8.03 B	395 287	N/A
Llama Guard 3 1B	LG 3 1B	Meta	Llama 3.2 1B	1.50 B	57 694	Meta (2024)
Llama Guard 3 8B	LG 3 8B	Meta	Llama 3.1 8B	8.03 B	1 259 174	Meta (2024)
Llama Guard Defensive	LG Def	Nvidia	Llama 2 7B	6.74 B	196 450	Ghosh et al. (2024)
Llama Guard Permissive	LG Per	Nvidia	Llama 2 7B	6.74 B	2 343	Ghosh et al. (2024)
MD-Judge	MD-J	Academia	Mistral 7B	7.24 B	15 046	Li et al. (2024a)
Mistral-7B-Instruct v0.2	Mis	Mistral AI	Mistral 7B	7.24 B	N/A	Jiang et al. (2023)
Mistral with MD-Judge prompt	Mis+	Mistral AI	Mistral 7B	7.24 B	N/A	Bassani and Sanchez (2024)
Shield Gemma 2B	SG 2B	Google	Gemma 2 2B	2.61 B	27 051	Zeng et al. (2024)
Shield Gemma 9B	SG 9B	Google	Gemma 2 9B	9.24 B	4 533	Zeng et al. (2024)

Table 1: Benchmarked models. Alias indicates the shortened names used in other tables. Downloads refers to the total number of downloads from HuggingFace as of February 1st, 2025.

mutations (i.e., typos, keywords camouflage, ciphers, and veiled expressions) and to adversarial attacks (i.e., jailbreak templates, DeepInception, CodeChameleon, and GCG).

5.1 Robustness to Mutations

Table 2 reports the evaluation results of the guardrail models' robustness to the mutations introduced in Section 3.1. For *unsafe* requests only, we also assess the robustness when an attacker selects the *best* mutation of a given kind for each request.

Robustness to Typos As shown in Table 2, by simply introducing one typo per keyword, most guardrail models' safety predictions can be significantly altered, causing very noticeable performance drops. Quite surprisingly, the most popular and advanced guardrail model in the Llama Guard series, Llama Guard 3 8B, has a False Negative Rate of 0.20 when the best performing typos is selected for each request, thus raising significant concerns. The only models that appear robust to typos are those from the Granite Guardian series. However, although the results may indicate a greater robustness of those models, the second half of Table 2 suggests that this robustness comes at a cost. While all the other guardrail models do not show performance drops for safe requests with typos, the Granite Guardian models are often susceptible to them, potentially causing the interruption of safe conversations. Our findings suggest that all guardrail models are adversely affected by typos and evading them can be as simple as misspelling unsafe requests. It is worth noting that input sanitization may alleviate some of the issues found with typos.

Robustness to keywords Camouflage As shown in Table 2, simple camouflage techniques can often fool most guardrail models', causing severe performance drops. As in the case of typos, Granite Guardian models appear not to be affected by camouflaged keywords. Moreover, Llama Guard 3 8B has a False Negative Rate of 0.30 when the best performing camouflage method is selected for each request. However, similarly to typos, the results for our *safe* control set suggest Granite Guardian models have learned to classify as *unsafe* samples presenting camouflaged keywords regardless of their actual safety. Although users might have no interest in camouflaging their safe requests, classifying as *unsafe* the inputs that deviate from well-formed English despite of their actual safety may have unintended consequences. Our findings suggest that camouflaged keywords can evade guardrail models' safety detection with a certain degree of success.

Robustness to Ciphers As shown in Table 2, ciphered unsafe requests can reliably evade all guardrail models, raising significant safety concerns due to the successful use of ciphers to bypass LLMs' alignment (Yuan et al., 2024). The second half of Table 2 reveal that guardrail models are not able to understand the meaning of ciphered user requests, unlike general-purpose LLMs. Our findings suggest that all guardrail models are adversely affected by ciphers. Since most available guardrail models were trained on datasets derived from the HH-RLHF dataset (Bai et al., 2022; Ganguli et al., 2022), it is unlikely they they were exposed to ciphers during fine-tuning, making them out-of-distribution.

Dataset	Method	GG 3 2B	GG 3 8B	GG 3.1 2B	GG 3.1 8B	LG	LG 2	LG 3 1B	LG 3 8B	LG Def	LG Per	MD-J	Mis	Mis +	SG 2B	SG 9B
Unsafe	Reference	0.01	0.00	0.01	0.01	0.14	0.08	0.03	0.01	0.00	0.06	0.03	0.13	0.02	0.26	0.28
	Typos															
	Deletion	0.00	0.01	0.00	0.03	0.27	0.27	0.22	0.11	0.06	0.18	0.22	0.22	0.14	0.37	0.42
	Insertion	0.00	0.00	0.01	0.01	0.16	0.17	0.11	0.06	0.01	0.10	0.15	0.16	0.07	0.33	0.34
	Replacement	0.00	0.01	0.01	0.01	0.29	0.33	0.23	0.13	0.08	0.23	0.28	0.29	0.15	0.42	0.43
	Swapping	0.00	0.01	0.01	0.04	0.34	0.26	0.21	0.09	0.16	0.24	0.34	0.29	0.22	0.35	0.38
	Mix	0.00	0.01	0.01	0.02	0.26	0.28	0.23	0.08	0.10	0.17	0.25	0.27	0.17	0.33	0.40
	Best	0.00	0.02	0.01	0.05	0.48	0.50	0.44	0.20	0.24	0.39	0.47	0.50	0.36	0.56	0.48
	Keywords Camouflage															
	Disemvoweling	0.00	0.01	0.04	0.00	0.46	0.48	0.24	0.22	0.17	0.38	0.50	0.37	0.36	0.49	0.55
	Punctuation	0.00	0.00	0.02	0.02	0.28	0.15	0.18	0.04	0.08	0.15	0.28	0.24	0.30	0.29	0.29
	White Spaces	0.00	0.00	0.02	0.01	0.22	0.14	0.26	0.03	0.03	0.11	0.34	0.32	0.40	0.28	0.27
	Splitting	0.00	0.00	0.02	0.01	0.20	0.16	0.19	0.02	0.06	0.14	0.20	0.19	0.12	0.29	0.32
	Syllables Inversion	0.00	0.01	0.03	0.03	0.36	0.35	0.18	0.13	0.11	0.26	0.29	0.37	0.24	0.39	0.49
	Leetspeak	0.00	0.00	0.01	0.00	0.19	0.40	0.07	0.10	0.02	0.05	0.23	0.16	0.06	0.31	0.41
	Best	0.00	0.02	0.07	0.04	0.64	0.65	0.50	0.32	0.30	0.51	0.64	0.69	0.57	0.64	0.64
	Ciphers															
	Caesar (ROT13)	0.43	0.04	0.62	0.21	1.00	0.96	0.86	0.74	0.50	0.91	1.00	1.00	1.00	1.00	1.00
	Morse	0.00	1.00	1.00	1.00	1.00	1.00	0.30	1.00	0.00	0.99	1.00	1.00	1.00	1.00	1.00
	Unicode	0.00	0.00	0.04	0.00	1.00	1.00	0.00	0.71	0.00	1.00	1.00	0.53	0.52	1.00	1.00
	Best	0.43	1.00	1.00	1.00	1.00	1.00	0.89	1.00	0.50	1.00	1.00	1.00	1.00	1.00	1.00
	Veiled Expressions															
	Veiled Expressions	0.03	0.07	0.09	0.06	0.48	0.50	0.34	0.32	0.21	0.34	0.39	0.38	0.36	0.51	0.49
Safe	Reference	0.03	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Typos															
	Deletion	0.40	0.07	0.10	0.11	0.00	0.00	0.02	0.00	0.02	0.01	0.00	0.00	0.00	0.00	0.00
	Insertion	0.41	0.11	0.10	0.12	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00
	Replacement	0.62	0.27	0.19	0.27	0.00	0.00	0.03	0.02	0.01	0.01	0.00	0.00	0.01	0.01	0.00
	Swapping	0.44	0.11	0.09	0.10	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
	Mix	0.50	0.11	0.10	0.12	0.00	0.00	0.02	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00
	Keywords Camouflage															
	Disemvoweling	0.91	0.59	0.41	0.49	0.02	0.00	0.09	0.02	0.05	0.03	0.01	0.03	0.00	0.02	0.00
	Punctuation	0.41	0.14	0.29	0.18	0.04	0.00	0.09	0.04	0.18	0.12	0.00	0.02	0.00	0.02	0.00
	White Spaces	0.21	0.01	0.04	0.00	0.01	0.00	0.04	0.00	0.05	0.02	0.00	0.00	0.00	0.00	0.00
	Splitting	0.18	0.03	0.07	0.02	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.02	0.00	0.00	0.00
	Syllables Inversion	0.85	0.62	0.42	0.55	0.02	0.00	0.08	0.01	0.05	0.01	0.01	0.00	0.00	0.01	0.00
	Leetspeak	0.99	0.92	0.92	0.92	0.12	0.12	0.43	0.21	0.49	0.25	0.13	0.30	0.58	0.10	0.03
	Ciphers															
	Caesar (ROT13)	0.29	0.95	0.32	0.74	0.00	0.00	0.00	0.03	0.32	0.00	0.00	0.00	0.00	0.00	0.00
	Morse	1.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	1.00	0.04	0.00	0.00	0.00	0.00	0.00
	Unicode	1.00	1.00	0.97	1.00	0.00	0.00	1.00	0.48	1.00	0.07	0.00	0.43	0.59	0.00	0.00

Table 2: Robustness to **mutations**. In the first half of the table (**Unsafe**), results are reported in terms of False Negative Rate (*lower is better*). In the second half of the table (**Safe**), results are reported in terms of False Positive Rate (*lower is better*). Green, Yellow, and Red indicate a negligible (≤ 0.01), moderate (< 0.10), or severe (≥ 0.10) performance decrease, respectively.

Robustness to Veiled Expressions As shown in Table 2, all guardrail models can often be fooled by veiled expressions. Surprisingly, Llama Guard 3 8B has a False Negative Rate of 0.32 when prompted with veiled unsafe requests, making it vulnerable to less direct prompts. Again, Granite Guardian models performed the best, potentially due to better training procedure. Our findings suggest that most guardrail models are vulnerable to veiled expressions. Automating the creation of convincing and realistic veiled expressions for unsafe prompts is challenging because it requires generating euphemisms, metaphors, and ambiguous language, which current LLMs find difficult (Tong et al., 2024). As noted in Section 3.1, we manually crafted veiled variants for our unsafe prompts because automatic methods (Xu et al., 2024) often change the original meaning. Consequently, extensive red teaming exercises might be necessary to create reliable veiled unsafe prompts for training.

5.2 Robustness to Adversarial Attacks

Table 3 reports the evaluation results of guardrail models' robustness to the adversarial attacks introduced in Section 3.2. For *unsafe* requests only, we also assess the robustness when an attacker selects the *best* attack variant for each request. It is worth mentioning that, as far as reported by their authors, only Granite Guardian models and MD-Judge have been exposed to adversarial attacks during training.

Robustness to Jailbreak Templates As shown in Table 3, guardrail models are generally robust to jailbreak templates. However, as reported in row *Best*, if an attacker aims to find at least one template that bypasses a guardrail model for a specific unsafe request, most guardrail models can be evaded with 100% success. Exceptions to this are the Granite Guardian models and MD-Judge, which maintain a reasonably low False Negative Rate. It is worth noting that this may be because these models' training data include the same or similar jailbreak templates, which can be easily harvested from the Internet, as discussed in Section 3.2. To better understand whether guardrail models can detect unsafe requests inside jailbreak prompts, we also combine the jailbreak templates with our safe requests. As reported in the second half of Table 3, guardrail models often fail to recognize safe requests embedded in jailbreak prompts. The lack of detailed information and availability of these models' training sets makes it difficult to determine whether this results from the inherent harmfulness of some jailbreak templates, overfitting to jailbreak prompts during training, or a combination of both. Our findings indicate that most guardrail models can be evaded using publicly available jailbreak templates.

Robustness to DeepInception As shown in Table 3, DeepInception's templates can evade only half of the guardrail models with a high degree of success. Moreover, Llama Guard 3 8B and Llama Guard Defensive show only moderate performance drops, while Granite Guardian and Shield Gemma models appear not affected by DeepInception. However, the results for our *safe* control set suggest Granite Guardian models may have been exposed to DeepInception. To better understand whether Granite Guardian and Shield Gemma models may have been exposed to DeepInception or similar attacks, we reformulate DeepInception's original template while preserving the logic of the original attack. As reported in Table 4, the revised template is more effective against all the guardrail models, further affecting their performance and reliably evade both Granite Guardian and Shield Gemma models. We conclude that those models may have been exposed to DeepInception or similar attacks. Our findings indicate that attacks like DeepInception, which target the personification capabilities of LLMs, can frequently bypass guardrail models.

Robustness to CodeChameleon As shown in Table 3, CodeChameleon's templates are extremely successful in evading all the considered guardrail models. Moreover, it fooled Llama Guard 3 8B on 87 out of 100 very unsafe prompts. Although the overall results are concerning, Granite Guardian 3 2B and 8B, Llama Guard 3 1B, Llama Guard Defensive and Mistral performed well in some cases. However, second half of Table 3 shows that those same models performed the worst on safe prompts. The lack of information about those models' training, makes it hard to determine whether the results for unsafe prompts are due to generalization or memorization, i.e., CodeChameleon was used to augment their training data. Our findings suggest that attacks targeting LLMs' coding capabilities, such as CodeChameleon, can reliably evade guardrail models.

Robustness to GCG As shown in Table 3, none of the guardrail models is robust to GCG, highlighting a very concerning situation. For example, GCG was able to fool with a *single* adversarial token Llama Guard 3 8B on 54 out of 100 very unsafe user prompts. By using 4 adversarial suffix tokens, all guardrail models can be reliably evaded. As far as reported by their authors, only the Granite Guardian models were exposed to training data augmented with adversarial suffixes generated with GCG. Still, they can be evaded by those same adversarial suffixes. As the author of the GCG attack leveraged 20 adversarial suffix tokens to jailbreak popular LLMs alignment, we believe that finding adversarial suffixes able to evade both a guardrail model and the generative model it protects may be plausible using GCG itself or other more advanced attack techniques based on adversarial suffixes. Our findings suggest that the lack of robustness to adversarial suffixes of available guardrail models is a major concern for the safety of generative AI systems.

Dataset	Method	GG 3 2B	GG 3 8B	GG 3.1 2B	GG 3.1 8B	LG	LG 2	LG 3 1B	LG 3 8B	LG Def	LG Per	MD-J	Mis	Mis +	SG 2B	SG 9B
Unsafe	Reference	0.01	0.00	0.01	0.01	0.14	0.08	0.03	0.01	0.00	0.06	0.03	0.13	0.02	0.26	0.28
	Templates															
	Templates	0.00	0.00	0.00	0.00	0.32	0.26	0.05	0.10	0.00	0.05	0.00	0.55	0.03	0.30	0.29
	Best	0.08	0.02	0.01	0.01	1.00	1.00	1.00	1.00	0.06	1.00	0.00	1.00	1.00	1.00	1.00
	DeepInception															
	Documentary	0.00	0.00	0.00	0.01	0.50	0.60	0.30	0.04	0.03	0.20	0.54	0.29	0.51	0.21	0.26
	Onion news group	0.00	0.00	0.00	0.00	0.36	0.51	0.09	0.01	0.00	0.16	0.23	0.19	0.35	0.21	0.22
	Science fiction	0.00	0.00	0.00	0.02	0.80	0.90	0.58	0.15	0.08	0.41	0.70	0.23	0.82	0.22	0.28
	Spy movie	0.00	0.00	0.00	0.02	0.73	0.74	0.59	0.08	0.07	0.33	0.59	0.60	0.74	0.26	0.32
	Stage scene	0.00	0.00	0.00	0.01	0.64	0.56	0.41	0.11	0.03	0.19	0.42	0.56	0.72	0.21	0.27
	Best	0.00	0.00	0.00	0.02	0.80	0.90	0.62	0.15	0.09	0.43	0.70	0.70	0.83	0.27	0.33
	CodeChameleon															
	Binary Tree Code	0.00	0.20	1.00	1.00	1.00	0.84	0.00	0.54	0.00	0.89	0.97	0.00	1.00	0.86	0.73
	Binary Tree Text	0.83	0.03	0.73	0.72	1.00	0.95	0.27	0.76	0.99	1.00	0.96	0.91	0.99	0.97	0.87
	Length Code	0.00	0.71	1.00	0.98	1.00	0.89	0.00	0.54	0.00	1.00	0.96	0.00	1.00	0.91	0.73
	Length Text	0.02	0.33	0.93	0.94	1.00	0.94	0.03	0.81	0.98	1.00	0.97	0.99	1.00	1.00	0.82
	Odd Even Code	0.05	0.00	0.97	0.22	1.00	0.65	0.00	0.16	0.00	1.00	0.72	0.00	0.99	0.46	0.43
	Odd Even Text	0.02	0.06	0.24	0.07	0.93	0.80	0.06	0.37	0.39	0.77	0.75	1.00	0.62	0.62	0.57
	Reverse Code	0.01	0.00	1.00	0.70	1.00	0.90	0.00	0.22	0.00	1.00	0.92	0.00	1.00	0.65	0.67
	Reverse Text	0.00	0.15	0.00	0.62	1.00	0.94	0.21	0.53	0.68	0.94	0.82	1.00	0.82	0.81	0.77
	Best	0.84	0.74	1.00	1.00	1.00	1.00	0.38	0.87	1.00	1.00	0.99	1.00	1.00	1.00	0.93
	GCG															
	Suffix Len = 4	1.00	0.97	0.93	0.98	0.99	1.00	1.00	1.00	0.81	0.96	1.00	1.00	0.86	0.93	0.97
	Suffix Len = 2	0.98	0.77	0.40	0.56	0.76	1.00	0.88	0.95	0.40	0.68	0.92	1.00	0.72	0.77	0.87
	Suffix Len = 1	0.41	0.32	0.08	0.13	0.42	0.97	0.44	0.54	0.15	0.26	0.78	0.89	0.15	0.53	0.61
Safe	Reference	0.03	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Templates															
	Templates	0.22	0.53	0.85	0.45	0.04	0.04	0.15	0.08	0.52	0.18	0.89	0.38	0.39	0.02	0.07
	DeepInception															
	Documentary	1.00	1.00	1.00	0.41	0.00	0.00	0.00	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Onion news group	1.00	1.00	1.00	1.00	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Science fiction	1.00	0.92	1.00	0.01	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Spy movie	1.00	1.00	1.00	0.99	0.00	0.00	0.00	0.01	0.14	0.00	0.00	0.00	0.00	0.00	0.00
	Stage scene	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CodeChameleon															
	Binary Tree Code	1.00	0.02	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.01	0.00	1.00	0.00	0.00	0.00
	Binary Tree Text	0.00	0.00	0.02	0.01	0.00	0.00	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Length Code	1.00	0.02	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00
	Length Text	0.88	0.00	0.00	0.00	0.00	0.00	0.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Odd Even Code	0.93	0.69	0.00	0.02	0.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00
	Odd Even Text	0.71	0.02	0.04	0.02	0.00	0.00	0.15	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Reverse Code	0.98	0.56	0.00	0.01	0.00	0.00	1.00	0.02	0.99	0.00	0.00	1.00	0.00	0.00	0.00
	Reverse Text	0.99	0.00	0.73	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Robustness to **adversarial attacks**. In the first half of the table (**Unsafe**), results are reported in terms of False Negative Rate (*lower is better*). In the second half of the table (**Safe**), results are reported in terms of False Positive Rate (*lower is better*). Green, Yellow, and Red indicate a negligible (≤ 0.01), moderate (< 0.10), or severe (≥ 0.10) performance decrease, respectively.

6 Conclusion and Future Work

In this paper, we presented a comprehensive evaluation of the robustness of 15 guardrail models against several input mutations and adversarial attacks. Our findings show that most guardrail models are vulnerable to simple input mutations, such as adding typos to unsafe requests or camouflaging their keywords with straightforward techniques. Furthermore, we found that guardrail models can be reliably evaded by employing attacks originally proposed to bypass the safety alignment of LLMs. Although evading guardrail models does not necessarily lead to obtaining unsafe information from the LLMs they protect, our results suggest that combining model alignment and guardrail models may not be sufficient to prevent the abuse of generative AI by malicious users. Our findings also suggest that the performance of guardrail models in handling well-formed, non-adversarial prompts might create an undue sense of confidence in their ability to provide comparable protection against adversarial attacks. It is worth noting that the Granite Guardian

Dataset	Method	GG 3 2B	GG 3 8B	GG 3.1 2B	GG 3.1 8B	LG	LG 2	LG 3 1B	LG 3 8B	LG Def	LG Per	MD-J	Mis	Mis +	SG 2B	SG 9B
Unsafe	Reference	0.01	0.00	0.01	0.01	0.14	0.08	0.03	0.01	0.00	0.06	0.03	0.13	0.02	0.26	0.28
	Revised DeepInception															
	Documentary	0.55	0.11	0.16	0.29	1.00	0.57	1.00	0.53	0.66	1.00	0.80	1.00	1.00	0.85	0.73
	Onion news group	0.98	0.08	0.34	0.28	1.00	0.75	1.00	0.50	0.68	1.00	0.80	1.00	1.00	0.85	0.74
	Science fiction	0.93	0.27	0.36	0.36	1.00	0.75	1.00	0.63	0.96	1.00	0.84	1.00	1.00	0.86	0.77
	Spy movie	0.97	0.27	0.50	0.52	1.00	0.79	1.00	0.65	0.92	1.00	0.84	1.00	1.00	0.86	0.80
	Stage scene	0.77	0.12	0.35	0.28	1.00	0.73	1.00	0.51	0.64	1.00	0.80	1.00	1.00	0.85	0.78
	Best	0.98	0.29	0.51	0.52	1.00	0.81	1.00	0.68	0.97	1.00	0.85	1.00	1.00	0.87	0.80

Table 4: Robustness to the **revised DeepInception's template**. Results are reported in terms of False Negative Rate (*lower is better*). Green, Yellow, and Red indicate a negligible (≤ 0.01), moderate (< 0.10), or severe (≥ 0.10) performance decrease, respectively.

models generally performed better than the other models with unsafe prompts, offering some degrees of protection against input mutations and adversarial attacks, even though they performed worse with safe prompts. Since our evaluation focused on a specific set of very unsafe prompts, it is likely that the performance of these models might be lower with more deceptive ones. Moreover, our assessments revealed that they do not necessarily generalize to unseen adversarial attacks. Therefore, we caution against considering the positive results reported in our paper as evidence of generalizable robustness. The limitations presented emphasize the need for further research into developing more robust safety guardrails. Future works should address the shortcomings we identified to enhance the robustness of guardrail models. For instance, advanced training regimes that includes data augmentation techniques and extensive red teaming exercises aimed at identifying areas of improvements could benefit the overall quality of those models. Additionally, developing complementary safety measures may be essential to effectively protect AI systems from adversarial attacks.

Limitations

While providing a valuable analysis for guardrail models' robustness to input mutations and adversarial attacks, our work has several limitations. Our experiments are limited to assessing the robustness of guardrail models and do not provide information regarding the evasion of other complementary safety measures, such as model alignment. As already stated, evading guardrail models does not mean to obtain unsafe answers from the LLMs they protect. Moreover, we limited our evaluation to a small pool of unsafe prompts (100) given the large number of models, mutations. and attacks we tested. Thus, our evaluation with unsafe prompts only has negative predictive power (Gardner et al., 2020), i.e., there is no guarantees that models performing well on those prompts will achieve the same performances with more deceptive unsafe prompts altered with mutations or embedded in adversarial attacks. In other words, our experiments are limited to assessing models' weaknesses in recognizing unsafe content rather than characterizing generalizable robustness. Therefore, claims about model quality should not be overextended based solely on the positive results reported in this paper. Additionally, we employed only English prompts in our evaluation. Further investigation is needed to establish whether the same findings extend to other languages. Finally, due to hardware constraints, we mainly investigated models up to a scale of 10 billion parameters. We also did not consider closed-weight and commercial moderation models such as OpenAI Moderation API and Perspective API.

Ethical Statement

Our research is part of the workplan of the GENESIS scientific project of the Joint Research Centre of the European Commission (JRC), which has received the approval from the JRC's Ethical Review Board. This research aims to advance the development of Trustworthy Generative AI systems by contributing to the design of robust and effective guardrail models. Our evaluation of these models has the goal identifying limitations of these critical AI safety components, paving the way for further research to increase their robustness in adversarial scenarios.

References

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2024. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *Preprint*, arXiv:2404.02151.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *SSRN Electronic Journal*.

Elias Bassani and Ignacio Sanchez. 2024. GuardBench: A large-scale benchmark for guardrail models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18393–18409, Miami, Florida, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Boyang Chen, Zongxiao Wu, and Ruoran Zhao. 2023. From fiction to fact: the growing role of generative ai in business and finance. *Journal of Chinese Economic and Business Studies*, 21(4):471–496.

Amanda Cercas Curry and Verena Rieser. 2018. #metoo alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing, EthNLP@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5, 2018*, pages 7–14. Association for Computational Linguistics.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4536–4545. Association for Computational Linguistics.

Peng Ding, Jun Kuang, Dan Ma, Xuezhi Cao, Yunsen Xian, Jiajun Chen, and Shujian Huang. 2024. A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2136–2153. Association for Computational Linguistics.

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.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1307–1323. Association for Computational Linguistics.

Google Gemini Team. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Google Gemini Team. 2024b. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. AEGIS: online adaptive AI content safety moderation with ensemble of LLM experts. *CoRR*, abs/2404.05993.

IBM Granite Team. 2024. Granite 3.0 language models.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. 2023. Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage. *Appl. Soft Comput.*, 145:110552.

Álvaro Huertas-García, Alejandro Martín, Javier Huertas-Tato, and David Camacho. 2024. Camouflage is all you need: Evaluating and enhancing language model robustness against camouflage adversarial attacks. *CoRR*, abs/2402.09874.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *CoRR*, abs/2312.06674.

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. *CoRR*, abs/2310.06825.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP@ACL 2019, Florence, Italy, July 28, 2019*, pages 177–180. Association for Computational Linguistics.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024a. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *CoRR*, abs/2402.05044.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. 2024b. Deepinception: Hypnotize large language model to be jailbreaker. *Preprint*, arXiv:2311.03191.

Zeyi Liao and Huan Sun. 2024. Amplegcg: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed llms. *Preprint*, arXiv:2404.07921.

Huijie Lv, Xiao Wang, Yuansen Zhang, Caishuang Huang, Shihan Dou, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Codechameleon: Personalized encryption framework for jailbreaking large language models. *Preprint*, arXiv:2402.16717.

Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit. Medicine*, 6.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

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

Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. 2024. Granite guardian. *Preprint*, arXiv:2412.07724.

Junaid Qadir. 2023. Engineering education in the era of chatgpt: Promise and pitfalls of generative AI for education. In *IEEE Global Engineering Education Conference, EDUCON 2023, Kuwait, May 1-4, 2023*, pages 1–9. IEEE.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018a. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018b. Language models are unsupervised multitask learners.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "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, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pages 1671–1685. ACM.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. *Preprint*, arXiv:2403.11810.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Bertie Vidgen, Hannah Rose Kirk, Rebecca Qian, Nino Scherrer, Anand Kannappan, Scott A. Hale, and Paul Röttger. 2023. Simplesafetytests: a test suite for identifying critical safety risks in large language models. *CoRR*, abs/2311.08370.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *CoRR*, abs/2307.12966.

Zeming Wei, Yifei Wang, and Yisen Wang. 2023. Jailbreak and guard aligned language models with only few in-context demonstrations. *CoRR*, abs/2310.06387.

Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3526–3548. Association for Computational Linguistics.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2024. GPT-4 is too smart to be safe: Stealthy chat with llms via cipher. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. Shieldgemma: Generative ai content moderation based on gemma. *Preprint*, arXiv:2407.21772.

Peng Zhang and Maged N. Kamel Boulos. 2023. Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet*, 15(9):286.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.