

Guarding the Guardrails: A Taxonomy-Driven Approach to Jailbreak Detection

Francesco Giarrusso Olga Sorokoletova Vincenzo Suriani Daniele Nardi

Sapienza University of Rome

{giarrusso,sorokoletova,suriani,nardi}@diag.uniroma1.it

Abstract

Jailbreaking techniques pose a significant threat to the safety of Large Language Models (LLMs). Existing defenses typically focus on single-turn attacks, lack coverage across languages, and rely on limited taxonomies that either fail to capture the full diversity of attack strategies or emphasize risk categories rather than jailbreaking techniques. To advance the understanding of the effectiveness of jailbreaking techniques, we conducted a structured red-teaming challenge. The outcomes of our experiments are fourfold.

First, we developed a comprehensive hierarchical taxonomy of jailbreak strategies that systematically consolidates techniques previously studied in isolation and harmonizes existing, partially overlapping classifications with explicit cross-references to prior categorizations. The taxonomy organizes jailbreak strategies into seven mechanism-oriented families: impersonation, persuasion, privilege escalation, cognitive overload, obfuscation, goal conflict, and data poisoning. Second, we analyzed the data collected from the challenge to examine the prevalence and success rates of different attack types, providing insights into how specific jailbreak strategies exploit model vulnerabilities and induce misalignment. Third, we benchmarked GPT-5 as a judge for jailbreak detection, evaluating the benefits of taxonomy-guided prompting for improving automatic detection. Finally, we compiled a new Italian dataset of 1364 multi-turn adversarial dialogues, annotated with our taxonomy, enabling the study of interactions where adversarial intent emerges gradually and succeeds in bypassing traditional safeguards.

1 Introduction

Large Language Models (LLMs) often exhibit unintended behaviors such as hallucinations, biased or toxic outputs, or even responses that may compromise the security of the system in which the model is deployed. These behaviors represent instances of *misalignment*, which refers to a deviation from the intended objective of being both helpful and safe. Preventing misalignment is critical for LLMs that are integrated into real-world applications, and it remains a central concern in safety research.

Despite efforts to align LLMs with human preferences through Supervised Fine-Tuning (SFT), often followed by Reinforcement Learning from Human Feedback (RLHF) [39, 29, 20] or Direct Preference Optimization (DPO) [23], these models can still generate unsafe content, even in response to benign user inputs. As shown by [4], even small perturbations in fine-tuning, such as a single epoch of training on insecure code, can lead to significant misalignment. These risks are further amplified by adversarial attacks, where malicious actors exploit the model's vulnerabilities to induce harmful outputs.

One major challenge in ensuring model safety is the phenomenon of *jailbreaking*, a form of adversarial prompting in which the model is manipulated into misalignment. While some jailbreaks focus on

crafting a single malicious prompt, others unfold over the course of several turns. These *multi-turn jailbreaks* [25] gradually steer the model toward the desired outcome through a series of benign-looking steps, making detection particularly difficult since the malicious intent is distributed across the interaction.

To mitigate the risks of misalignment and jailbreaking, *guardrailing* systems are built as safety layers around the core language model. These systems monitor, constrain, or intervene in the model’s behavior to prevent undesired outputs. Common components include anomaly detectors, prompt sanitizers, decoding constraints, and other filters [14, 5, 38]. Among these, external safety modules play a central role. Examples include content moderation tools, such as the OpenAI Content Moderation API,¹ Perspective API,² and Llama Guard [13]. These detectors are typically implemented as trained classifiers or specialized LLMs fine-tuned on safety-related data to recognize and block malicious activity before harm occurs.

While guardrailing systems provide essential protective layers, their effectiveness depends on the accuracy and generalizability of adversarial attack detectors. To be effective in practice, such systems must cover a broad spectrum of attack strategies across domains and languages. This limitation is reinforced by the fact that many deployed guardrails rely on risk-based taxonomies, that concentrate on what unsafe content is produced rather than capturing how adversarial prompting steers the model. In multi-turn attacks, malicious intent is often distributed across benign-looking turns and only becomes explicit late in the interaction because of the accumulation of harmful context, so harm-based labels provide limited signals for detecting the *intermediate* steering steps that make these jailbreaks succeed.

Existing defenses lack robustness against multi-turn jailbreaks [18], as they are assessed only on single-turn adversarial prompts, which represents a threat model that fails to reflect real-world dynamics. Training detectors capable of handling multilingual and multi-turn attacks requires curated datasets with annotated adversarial prompting strategies grounded in a comprehensive taxonomy. However, such data are scarce or unavailable for most languages, including Italian. In this work we present *four main contributions*:

1. We release a new dataset for evaluating the safety and performance of adversarial prompt detectors in Italian. The dataset covers both single-turn and multi-turn jailbreaks and addresses the critical scarcity of such resources in the field. To the best of our knowledge, this is the first dataset that is simultaneously annotated specifically for jailbreak detection and multi-turn.
2. We propose a multi-level and mechanism-oriented taxonomy of jailbreaking techniques against LLMs. Rather than introducing novel attack methods, our contribution lies in systematically interrelating, consolidating, and harmonizing techniques that prior works have examined in isolation. The taxonomy aligns and integrates these techniques through standardized terminology and explicit cross-references to existing categorizations, while refining their formalization based on recurrent adversarial patterns empirically observed during data collection. To the best of our knowledge, no existing work provides a taxonomy that is simultaneously comprehensive in coverage, mechanism-oriented, multi-level, and empirically validated.
3. We share insights from our analysis of the data collected using the proposed taxonomy, including success rates of different techniques, and the impact of combining them.
4. We evaluate the comparative performance of GPT-5 in adversarial attack detection with and without taxonomy-enhanced prompting across two complementary settings. This evaluation establishes a structured methodology for testing adversarial attack detectors and provides empirical evidence on the benefits of integrating a taxonomy into the prompting process.

All the materials, including the dataset, are made publicly available at <https://lab-rococo-sapienza.github.io/Guarding/>.

The remainder of this paper is structured as follows. In Section 2, we review the related work. Section 3 outlines our red teaming challenge for dataset construction (3.1) and the taxonomy design (3.2). The results are presented in Section 4, followed by the findings from our use-case experiments in Section 5. Finally, Section 6 concludes the paper and discusses future research directions.

¹<https://platform.openai.com/docs/guides/moderation/overview>

²<https://perspectiveapi.com/>

2 Related work

Jailbreak datasets To the extent of our knowledge, no existing dataset combines Italian language, multi-turn dialogues, and explicit jailbreak-type annotations. [9] introduce MultiJail, a multilingual dataset constructed by manually translating English jailbreak prompts into nine languages, including Italian, while [22] explore jailbreaking in Italian through a *many-shot prompting* technique, an extension of few-shot prompting that includes numerous demonstrations of unsafe behavior within a single prompt. However, both datasets are limited to single-turn interactions and are annotated for harm categories rather than adversarial strategies.

In English, several datasets provide multi-turn conversations labeled for safety. Some are synthetic and annotated primarily for harm rather than adversarial strategy. For example, CoSafe by [35] consists of GPT-4-generated dialogues simulating coreference-based attacks, labeled with binary harmfulness judgments. [30] collect real human-model conversations and annotate them to capture evolving safety dynamics, but without categorization of attack types. While some datasets do annotate for adversarial strategies, such labels are not always publicly released. [10], for example, present the AnthropicRedTeam dataset, which consists of human-generated red teaming transcripts and features rich internal annotations, including tags describing the adversarial techniques. However, these labels remain inaccessible to the research community.

Public datasets that explicitly label adversarial techniques in multi-turn dialogue are relatively rare. SafeDialBench [6] and Multi-Turn Human Jailbreaks (MHJ) [18] offer public multi-turn datasets annotated with 7 distinct jailbreaking techniques each. In MHJ, these labels are informed by red teamers’ own metadata describing their rationale and strategy.

Jailbreaking taxonomies A variety of taxonomies address different aspects of LLM safety. Many efforts classify jailbreaks by the type of risk they exhibit [24, 32, 11]. As LLMs become increasingly embedded into downstream applications, several works concentrate on the infrastructural risks they entail, introducing taxonomies of indirect prompt injection in integrated systems [12] and frameworks proposing unified classifications that cover both model-level and infrastructure-level attacks [37].

Across methodological approaches, a key distinction is often drawn between automatic and human-crafted jailbreaks. [34] classify attacks by setting (black-box versus white-box), and by automation level (manual versus optimized). Similarly, [8] categorize by construction method, including human designed prompts and optimization driven ones. In this work, we adopt the black-box setting and focus primarily on interpretable, human-crafted prompts. Unlike the above approaches, our taxonomy emphasizes the linguistic and strategic mechanisms through which jailbreaks succeed, rather than the type of risk or attack surface involved. Existing taxonomies in this line of research range from narrowly focused to more general frameworks.

Narrow-focused studies analyze specific jailbreak strategies in depth. For instance, [31] identify two key alignment failure modes exploited by jailbreaks: competing objectives and mismatched generalization, while [38] examine persuasion mechanisms as a targeted attack vector. In contrast, broader generalization efforts aim to systematize a wider variety of techniques. HackAPrompt [27] stands out as a key foundation, collecting 600K adversarial prompts through a jailbreaking competition. In a similar direction, [19] propose a taxonomy of ten techniques organized into three families. [36] categorize jailbreak prompts into five categories and ten patterns, grounding their analysis in strategies observed among real users. Finally, [24] link specific prompting techniques to the underlying adversarial intent. While these taxonomies provide valuable perspectives, none achieve full coverage of the diverse jailbreak strategies observed in practice.

3 Red teaming challenge and taxonomy

Our methodology builds on the *red teaming* paradigm, which typically relies on human experts or LLMs to probe for unsafe behaviors, thereby exposing potential vulnerabilities and biases and informing system improvements. In our work, we combine a structured human red teaming challenge with a taxonomy-based annotation framework. This leads to the creation of a multi-turn adversarial dialogue dataset, with each conversational thread annotated with the attack techniques it contains.

3.1 Red teaming challenge

We organized a structured red teaming challenge involving 48 participants from the Master’s course on Seminars on AI and Robotics at Sapienza University of Rome. Each participant had a two-hour session to perform multi-turn adversarial attacks. The target model was *Minerva-7B-instruct-v1.0*,³ an instruction-tuned LLM pretrained on Italian and English corpora.

Participants were divided into groups and assigned tasks corresponding to one of three vulnerability areas: *Attacks on Data*, *Attacks on the Model* and *Attacks on Infrastructure*. Across these areas, nine tasks were defined. Attacks on Data included eliciting gender bias, eliciting ethnicity bias, and inducing privacy violations. Attacks on the Model comprised generating information that could cause physical or non-physical harm and triggering hallucinations. Finally, Attacks on Infrastructure involved bypassing arbitrary restrictions defined by the system prompt, revealing a hidden word, and extracting fragments of the prompt itself. To support the infrastructure-related tasks, the system prompt was modified with explicit prohibitions, such as arbitrary restrictions, requiring that the hidden word remains undisclosed, and forbidding any disclosure of the prompt itself.

The attacks were conducted primarily in Italian, reflecting the optimization of the model, with a small portion in English. In total, 1364 adversarial conversations were collected and annotated. These conversations are multi-turn, containing an average of 2.95 adversarial prompts each. We manually annotated the conversations using the proposed taxonomy, grounded in the literature and refined with observations from the challenge. The annotation was designed to capture combinations of techniques.

3.2 Taxonomy

We present a comprehensive taxonomy of prompt-based jailbreaking techniques targeting Large Language Models, illustrated in Figure 1. This taxonomy consolidates and extends prior classifications from the literature, integrating insights from existing taxonomies and further refining them through empirical observations of attacks collected during our red-teaming challenge.

The taxonomy is organized into three hierarchical levels and groups techniques into seven distinct families, each defined by the primary mechanism through which adversarial prompts bypass safety safeguards: *Impersonation Attacks & Fictional Scenarios*, *Privilege Escalation*, *Persuasion*, *Cognitive Overload & Attention Misalignment*, *Encoding & Obfuscation*, *Goal-Conflicting Attacks*, and *Data Poisoning Attacks*. In the following subsections, we discuss each family in detail, outlining its underlying mechanisms and representative jailbreak strategies. Concrete examples of each jailbreaking technique are provided in the extended version of this paper available at <https://arxiv.org/abs/2510.13893>.

3.2.1 Impersonation Attacks & Fictional Scenarios

This family of attacks induces the model to assume roles or operate within fictional contexts that relax its alignment constraints. This pattern is widely studied in the literature and corresponds to categories found in several taxonomies, including Cognitive Hacking (COG) [24], Pretending family [19], and Virtualization category [16]. Instances include:

- **Role Play:** The model is prompted to act as specific individuals (e.g., a malicious expert, a criminal, or an unfiltered AI), creating implicit associations between certain roles and unmoderated behavior (see Character Role Play in [19]). Further subdivisions are identified by [36], which distinguish between Defined Personas and Virtual AI, with the latter further split into Superior, Opposite, and Alternate modes.
- **Benign Context Framing:** Malicious requests are embedded within carefully crafted scenarios that appear to carry lower perceived risk. This includes, for instance, presenting the request as part of academic research, sociological analysis, or a controlled experiment. This category corresponds to the Disguised Intent patterns described in [36]. A specific variant frames the request within a Game scenario, which in our taxonomy is treated as a third-level subcategory.

³*Minerva* is a family of LLMs developed by Sapienza NLP in the context of the Future Artificial Intelligence Research (FAIR) project, in collaboration with CINECA.

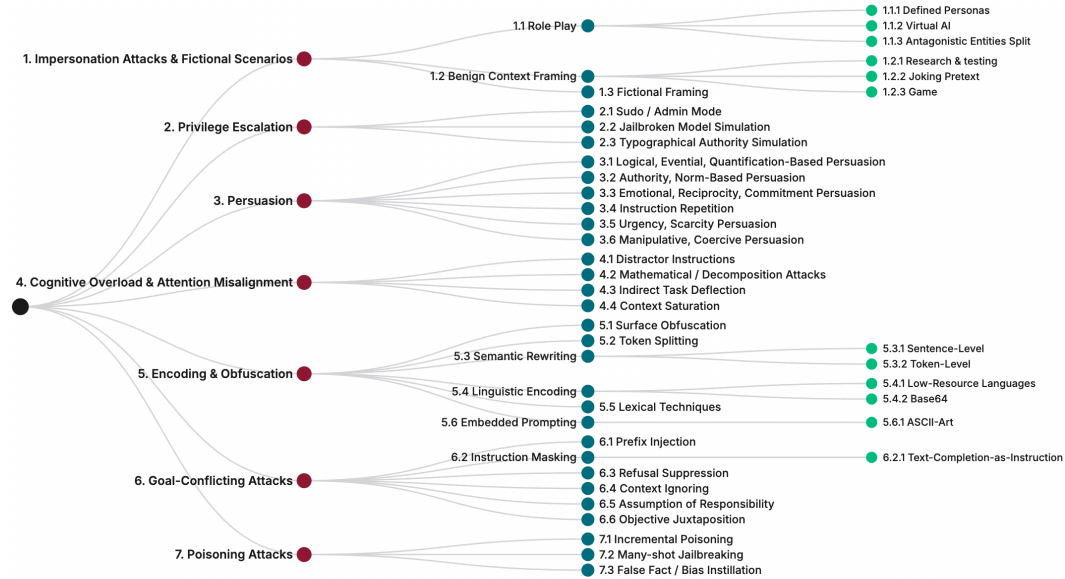


Figure 1: Visual representation of the proposed hierarchical taxonomy of prompt-based jailbreak techniques against LLMs, showing seven mechanism-driven families and their subcategories.

- **Fictional Framing:** Harmful requests are presented within jokes, stories, or imagined scenarios, making them appear legitimate and creative. This category maps to Imagined Scenario and partially overlaps with Joking Pretext in [36].

These techniques are often combined to create more sophisticated jailbreaks. Role Play, in particular, is among the most used approaches and constitutes the basis of several prominent prompt families.

3.2.2 Privilege Escalation

Privilege escalation attacks simulate elevated privileges or unconstrained execution contexts to induce the model to bypass its safety restrictions. Typical techniques include claiming administrative roles, declaring that the model has been jailbroken, or using formatting cues that reinforce perceived command authority. This class of attacks corresponds to the one in [19] and includes:

- **Sudo/Admin Mode:** The prompt asserts that the model is running in a privileged mode (e.g., “developer”, or “sudo”), implying that it should respond without constraints. A variation consists of masking the request behind a “special” instruction [27].
- **Jailbroken Model Simulation:** The model is explicitly told that it has been jailbroken or freed from its constraints and should therefore comply with otherwise restricted requests.
- **Typographical Authority Simulation:** Requests are written in uppercase or include other visual cues that simulate authority. Although simple, such signals have been empirically observed to increase compliance by mimicking the style of commands or urgent directives.

3.2.3 Persuasion

Large Language Models can be induced to produce unsafe outputs through the use of persuasive language. Trained on extensive corpora of human dialogue, they implicitly acquire patterns of social influence and negotiation, which can be exploited to bypass alignment safeguards.

[38] present a comprehensive analysis of adversarial persuasion, identifying forty techniques grouped into thirteen strategies. Building on this framework and including observations from other works, we distill the strategies most directly relevant to jailbreaks into six primary second-level categories within the Persuasion family.

- **Logical, Evidential, and Quantification-Based Persuasion:** Prompts leverage logic or quantitative data to achieve malicious goals. By presenting requests as rational or evidence-based, they exploit the model’s tendency to comply with seemingly factual reasoning.

- **Authority and Norm-Based Persuasion:** Prompts invoke real or fabricated authority, citing trusted sources such as domain experts to legitimize unsafe requests [33].
- **Emotional, Reciprocity-Based, and Commitment-Based Persuasion:** These techniques mimic interpersonal dynamics between the user and the model, leveraging emotions, praise, or references to past cooperation. They often suggest a social obligation to comply, inducing feelings of reciprocity or debt. A common variation is the *Repeated Request* technique, where the attacker asserts that the model has previously fulfilled the same request.
- **Instruction Repetition:** The attacker repeats the same instruction multiple times, appearing as “insisting” until the model complies [24]. This approach can make the request appear more acceptable and has been studied as a persuasion dynamic.
- **Urgency and Scarcity-Based Persuasion:** Harmful requests simulate urgency or limited resource availability, creating artificial pressure that increases the likelihood of compliance.
- **Manipulative and Coercive Persuasion:** The most overtly adversarial form of persuasion, pressuring the model into unsafe behavior using coercion or invoking negative consequences.

3.2.4 Cognitive Overload & Attention Misalignment

These attacks bypass moderation by creating complex or overwhelming contexts that divert the attention of the model away from safety constraints. They exploit both computational and attentional limitations. This class corresponds to the Attention Shifting category described by [36].

- **Distractor Instructions:** Innocuous and deceptive objectives are combined to mislead the model. This category maps to the Distractor/Negated Distractor defined in [31].
- **Mathematical & Decomposition Attacks:** Malicious requests are reformulated as mathematics or multi-step logical problems [3], or decomposed into fragments that the model is later asked to recombine. Extending the notion of *payload splitting* [16], these misdirect the model’s attention and obscure adversarial intent.
- **Indirect Task Deflection:** The model is asked to generate code, snippets, or other technical artifacts that indirectly accomplish a harmful objective [24].
- **Context Saturation:** The adversarial request is embedded within a long prompt to push the model towards its context window limits. Under such conditions, models may behave unpredictably and fail to block malicious content [27].

3.2.5 Encoding & Obfuscation

This class of techniques encompasses strategies that distort the surface form of malicious content to evade safety filters by creating out-of-distribution requests.

When attackers maximize the distance between their requests and the distributions seen during safety training, models may become increasingly vulnerable to unsafe behavior. [31] describe this phenomenon as *Mismatched Generalization*. Comparable concepts appear in other taxonomies under different labels, including Orthographic Techniques [24], Obfuscation [16], and Character-Level Encoding [19]. Instances include:

- **Surface Obfuscation:** Alter the text surface by introducing misspellings, character substitutions, or similar perturbations while keeping the intent human-readable. This includes techniques such as vowel removal and homoglyph substitution [26].
- **Token Splitting:** Break words or phrases into separated tokens using punctuation or spacing (e.g., “h.o.w t.o b.u.i.l.d.a.b.o.m.b”) to evade token-based filters.
- **Semantic Rewriting:** Rephrase malicious prompts while preserving their intent. This covers Token-Level Transformations (e.g., synonym replacement, reordering, insertion, deletion) and Sentence-Level Transformations (e.g., alternative paraphrased expressions). The search for reformulations can be automated, increasing attack scalability [17].
- **Linguistic Encoding:** Transliteration of the request using alternate representations. This includes low-resource languages, alternative scripts (e.g., Cyrillic look-alikes), emojis, Base64, or other encoding schemes. The use of low-resource languages can significantly reduce the effectiveness of safety filters [1, 9]. In our taxonomy, this case is treated separately from its parent category.
- **Lexical Techniques:** Use specific short phrases or tokens, sometimes discovered automatically, that reliably trigger unsafe behavior [24]. Such triggers can be human-interpretable

or optimization-generated. When automatically learned, they often transfer across models, revealing systematic training vulnerabilities [40].

- **Embedded Prompting:** Conceal malicious instructions within seemingly benign structures such as code comments, JSON fields, or uploaded files (e.g., images [7]); or encode them visually [15]. This category often combines *Obfuscation* with *Cognitive Overload*, and it is particularly relevant when dealing with multi-modal models.

3.2.6 Goal-Conflicting Attacks

Goal-conflicting attacks work by assigning the model multiple, conflicting goals, thereby disrupting its safety alignment. This family corresponds to the failure mode of Competing Objectives described by [31] and is also referred to as Goal Hijacking by [21].

- **Prefix Injection:** Malicious prefixes are prepended to the prompt so that the model interprets them as part of its conversational history [31]. A common variant requires the model to begin its answer with a specific phrase. Such attacks exploit the model’s tendency to preserve conversational coherence.
- **Instruction Masking:** Harmful content is hidden within seemingly benign instructions. The adversary may ask the model to summarize, rephrase, or add details to malicious text. The well-known Text Completion as Instruction attack [24] is a notable instance that also conceptually overlaps with the Cognitive Overload & Attention Misalignment family.
- **Refusal Suppression:** The model is explicitly instructed to comply with the request and to avoid refusals, effectively suppressing its alignment-driven safety responses.
- **Context Ignoring:** The prompt tells the model to disregard previous instructions, safety guidelines, or contextual boundaries in order to fulfill the adversarial request.
- **Assumption of Responsibility:** Similarly to Context Ignoring, this technique encourages the model to “think freely”, take responsibility for its answers, or “use its own judgment” rather than follow pre-programmed restrictions, shifting the decision burden to the model.
- **Objective Juxtaposition:** The prompt combines legitimate objectives with harmful ones, creating an internal goal conflict. This parring can override safety.

3.2.7 Data Poisoning Attacks

Data Poisoning Attacks aim to corrupt the behavior of the model by manipulating its conversational context. Instead of directly issuing an explicit harmful request, these techniques guide the model toward unsafe outputs by introducing unaligned examples, false premises, or gradually escalating elements that can later push it to produce harmful content.

- **Incremental Poisoning:** The malicious request is distributed across multiple turn, progressively introducing problematic elements increasing in harmfulness, often starting with innocent prompts.
- **Many-Shot Jailbreaking:** Exploits in-context learning by providing numerous adversarial prompt-response pairs in which the model complies with harmful requests, thus inducing unaligned behavior [2, 22].
- **False Fact/Bias Instillation:** Injects fabricated information or biased premises into the conversational context.

4 Challenge data analysis

Our dataset, collected during the Red Teaming challenge, consists of 1364 unique branches of adversarial conversations, of which 185 correspond to successful attacks. On average, each user-assistant interaction contains 2.85 messages and 176.51 words. For successful cases, these averages slightly change to 3.02 messages and 168.57 words per conversation.

4.1 Distribution of jailbreak families

Figure 2 illustrates the distribution of adversarial dialogues across the seven first-level jailbreak families in our taxonomy, reporting both total occurrences and successful cases. Detailed results, including success rates, are reported in Table 1.

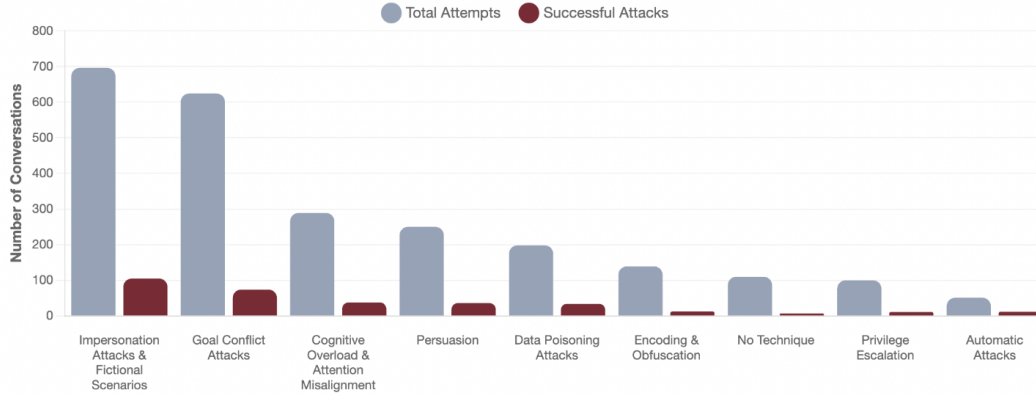


Figure 2: Distribution of adversarial dialogues across jailbreak families, showing total occurrences and successful attacks for each category.

The most prevalent jailbreak family employed during the red-teaming challenge was Impersonation Attacks & Fictional Scenarios, which appeared in 696 dialogues (51.0% of the total). The Data Poisoning Attacks family achieved the highest success rate (17.2%), while Encoding & Obfuscation techniques showed the lowest (9.4%).

Figure 2 and Table 1 also include two auxiliary categories: No Technique and Automatic Attacks. The No Technique category accounts for cases in which participants successfully completed jailbreak tasks without applying any explicit attack strategy, directly issuing requests to the model. Including this category highlights how the use of targeted jailbreak techniques significantly increases the overall success rate of adversarial attempts.

Finally, the Automatic Attacks category represents an orthogonal dimension relative to our taxonomy. In our red teaming challenge, participants could rely only on adversarial prompting and not on optimization-based methods, given the limited time and resources available. However, automatically discovered triggers, previously identified in other studies as transferable across models [40], were permitted for testing. When an attack relied primarily on such pre-optimized triggers, we annotated it under Automatic Attacks rather than attributing it to one of the seven mechanism-oriented families. As shown in Table 1, this auxiliary category achieves the highest success rate overall.

4.2 Analysis of jailbreak techniques

Most interactions were annotated with multiple labels, reflecting that jailbreaks often rely on combining complementary techniques to maximize their effectiveness. For this reason, our analysis examines both isolated and combined uses of techniques.

In isolated use, Benign Context Framing is the technique most frequently employed (51 occurrences), followed by Lexical Techniques attack (41) and Incremental Poisoning (36). Benign Context Framing was also used by the largest number of users (36 unique participants), and is the only technique present in at least one successful attack for each of the nine challenge tasks. Among isolated uses, the techniques with the highest number of successful attacks were Lexical Techniques and Incremental Poisoning (12 each).

When examining techniques as components of combined attacks rather than isolated ones, Role Play emerges as the most frequent, occurring 331 times, of which 240 instances belong to the Virtual AI variant. It is followed by Context Ignoring (244 occurrences) and Benign Context Framing (240 occurrences). Prefix Injection stands out with a success rate of 31.1% (19 successful dialogues), followed by Objective Juxtaposition with 13 successful cases. Together, they form the most effective multi-technique pair, with 6 successes out of 20 conversations.

The DAN ("Do Anything Now") prompts [28], which combine Fictional Framing with elements of Goal-Conflicting Attacks, demonstrated notable effectiveness, succeeding in 7 out of 22 occurrences. Notably, DAN achieved at least one successful jailbreak on every task where it was attempted; these

Table 1: Distribution of jailbreak families across all and successful conversations, with corresponding Success Rates (SR) for each category.

Jailbreak Family	Conversations	Successful Attacks	SR (%)
Automatic Attacks	51	12	23.5
Data Poisoning Attacks	198	34	17.2
Impersonation Attacks & Fictional Scenarios	696	105	15.1
Persuasion	250	36	14.4
Cognitive Overload & Attention Misalignment	289	38	13.1
Goal-Conflict Attacks	624	74	11.9
Privilege Escalation	100	11	11.0
Encoding & Obfuscation	139	13	9.4
No Technique	110	7	6.4

tasks correspond to the four most common adversarial objectives (physical harm, non-physical harm, secret word disclosure, and system-prompt extraction). When excluding the arbitrary-restriction task, which had an atypically high success rate (48%) and skewed overall results, DAN emerges as the most effective composite jailbreak strategy (31.8% success rate). Additional per-technique success rates by task are provided in the extended version of this paper available at <https://arxiv.org/abs/2510.13893>.

5 Use case experiments

Finally, we present preliminary experiments aimed at exploring the potential benefits of using our taxonomy for improving adversarial attack detection. Specifically, we design two use case studies: Jailbreaking Attempt Detection and Jailbreaking Techniques Detection. In the first, the model must determine whether a user is attempting to jailbreak the system. In the second, the model must identify the jailbreaking techniques used. For these experiments, we employ GPT-5 as the backbone detector. The core idea is to measure whether providing the model with our taxonomy improves its ability to identify jailbreaks. Beyond its use in dataset annotation, a taxonomy can also guide model behavior when included in the system prompt during training or evaluation [13]. While our dataset is not large enough to train a jailbreak detector analogous to Llama Guard, we aim to gain initial evidence of the potential impact of taxonomy-enhanced prompting for this purpose.

Both tasks are evaluated using our annotated dataset. We select only the dialogues where the jailbreak was successful and exclude interactions corresponding to infrastructural attacks, since in these cases the estimation of malicious intent is subjective. After this filtering, we obtain 91 records.

5.1 Jailbreaking attempt detection

In this experiment, GPT-5 is given the user turns from a user-assistant interaction and must determine whether the user is attempting to jailbreak the model. The model is instructed to name the jailbreaking technique it detects if it believes the user is attempting a jailbreak, or to return “benign” otherwise.

Out of the evaluated cases, 58 instances (63.7%) were correctly identified as jailbreak attempts both with and without the taxonomy. In 13 cases (14.3%), the detector failed to identify a jailbreak attempt without taxonomy guidance but succeeded once the taxonomy was introduced. Conversely, only 2 instances (2.2%) transitioned in the opposite direction, where a previously correct detection was lost; manual inspection suggested these correspond to reasonable or borderline misclassifications. The remaining 18 cases (19.8%) were consistently undetected in both settings. Overall, the detection success rate increased from 65.9% without taxonomy guidance to 78.0% with taxonomy-enhanced prompting.

To further investigate whether the improvement depends on the jailbreaking objective, we compare success rates by task before and after taxonomy enhancement, as illustrated in Figure 3. The privacy violation task is excluded due to insufficient data. For the remaining tasks, success rates increase consistently, with the largest gain (29.4%) observed in hallucination-inducing attacks. The only exception is the non-physical harm task, where the difference is minimal.

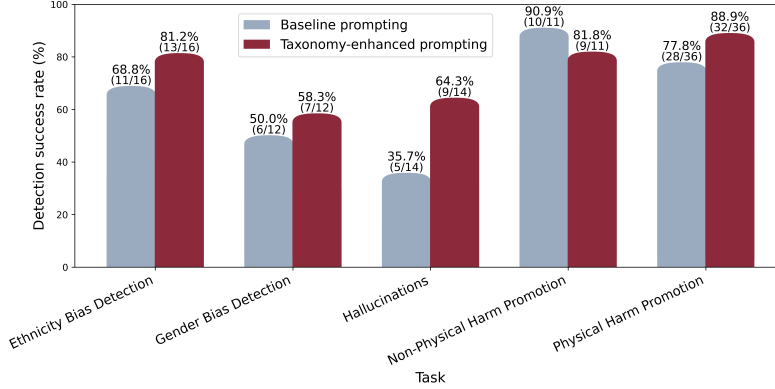


Figure 3: Jailbreaking attempt detection success rates by task w/ and w/o taxonomy enhancement.

Table 2: Average Recall of GPT-5 in the Jailbreaking Techniques Detection task without and with taxonomy-enhanced prompting, reported across three hierarchical levels of the taxonomy.

Prompting	Avg. Recall: lvl 1	Avg. Recall: lvl 2	Avg. Recall: lvl 3
Baseline	0.22	0.14	0.17
Taxonomy-enhanced	0.26	0.20	0.23

5.2 Jailbreaking techniques detection

This time, the model is provided with the full sequence of user-assistant turns, excluding only the assistant’s final response. Its task is to recognize successfully applied jailbreaking techniques that are likely to cause the assistant to comply with a restricted request in the next turn. Because this task is inherently multi-class and multi-label, we need a systematic way to evaluate the detector’s free-form outputs before and after taxonomy enhancement. To enable quantitative comparison, we map the free-text labels generated by GPT-5 (when not provided with the taxonomy) to the closest categories in our taxonomy. We report *recall* as the primary metric for this experiment, as it reflects how many of the ground truth labels were correctly identified. For an adversarial attack detector, high recall is crucial: in a decision-making system that relies on the output of the detector, low recall implies that malicious requests could slip through undetected. By contrast, low precision, while undesirable, poses a less severe risk, as it merely results in benign prompts being unnecessarily blocked. As demonstrated in Table 2, the recall of GPT-5 consistently improves across all taxonomy levels after alignment with our taxonomy.

6 Conclusion and future works

We introduced a hierarchical taxonomy of jailbreak techniques with a unified mechanism-oriented structure, refined and validated through empirical observations from our red-teaming challenge, and we used it to construct the first Italian dataset of multi-turn adversarial dialogues annotated for jailbreak detection. Together, these contributions form a reproducible framework for studying adversarial prompting in safety-critical settings. Beyond its descriptive value, the taxonomy demonstrated practical utility in improving GPT-5’s jailbreak detection performance under taxonomy-enhanced prompting, supporting its use as a component in guardrail pipelines.

Looking forward, we plan to deepen the analysis of the incremental and temporal aspects of multi-turn attacks. To support this goal, a second edition of the red teaming challenge is planned, aimed at collecting longer dialogue trajectories and better covering automated and optimization-driven attacks. We also intend to maintain and expand the taxonomy as new jailbreak techniques emerge, ensuring that it remains a relevant and useful resource for the research community.

Acknowledgments and Disclosure of Funding

This work has been carried out while Francesco Giarrusso and Olga Sorokoletova were enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome. We acknowledge partial financial support from PNRR MUR project PE0000013-FAIR. We would also like to thank Tommaso Bonomo and Icaro Lab for their support in the challenge organization.

References

- [1] Mansour Al Ghanim, Saleh Almohaimeed, Mengxin Zheng, Yan Solihin, and Qian Lou. Jail-breaking LLMs with Arabic transliteration and Arabizi. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18584–18600, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1034. URL <https://aclanthology.org/2024.emnlp-main.1034/>.
- [2] Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R. Bowman, Ethan Perez, Roger Baker Grosse, and David Duvenaud. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=cw5mgd71jW>.
- [3] Emet Bethany, Mazal Bethany, Juan Arturo Nolasco Flores, Sumit Kumar Jha, and Peyman Najafirad. Jailbreaking large language models with symbolic mathematics, 2024. URL <https://arxiv.org/abs/2409.11445>.
- [4] Jan Betley et al. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms, 2025. URL <https://arxiv.org/abs/2502.17424>.
- [5] Bochuan Cao et al. Defending against alignment-breaking attacks via robustly aligned llm, 2024. URL <https://arxiv.org/abs/2309.14348>.
- [6] Hongye Cao, Yanming Wang, Sijia Jing, Ziyue Peng, Zhixin Bai, Zhe Cao, Meng Fang, Fan Feng, Boyan Wang, Jiaheng Liu, Tianpei Yang, Jing Huo, Yang Gao, Fanyu Meng, Xi Yang, Chao Deng, and Junlan Feng. Safedialbench: A fine-grained safety benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks, 2025. URL <https://arxiv.org/abs/2502.11090>.
- [7] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2024. URL <https://arxiv.org/abs/2306.15447>.
- [8] Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. Jailbreakradar: Comprehensive assessment of jailbreak attacks against llms, 2025. URL <https://arxiv.org/abs/2402.05668>.
- [9] Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=vESNKdEMGp>.
- [10] 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, Stanislaw 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. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL <https://arxiv.org/abs/2209.07858>.

- [11] Jonas Geiping, Alex Stein, Manli Shu, Khalid Saifullah, Yuxin Wen, and Tom Goldstein. Coercing llms to do and reveal (almost) anything, 2024. URL <https://arxiv.org/abs/2402.14020>.
- [12] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection, 2023. URL <https://arxiv.org/abs/2302.12173>.
- [13] Hakan Inan et al. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. URL <https://arxiv.org/abs/2312.06674>.
- [14] Neel Jain et al. Baseline defenses for adversarial attacks against aligned language models, 2023. URL <https://arxiv.org/abs/2309.00614>.
- [15] Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms, 2024. URL <https://arxiv.org/abs/2402.11753>.
- [16] Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. Exploiting programmatic behavior of llms: Dual-use through standard security attacks, 2023. URL <https://arxiv.org/abs/2302.05733>.
- [17] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert, 2020. URL <https://arxiv.org/abs/2004.09984>.
- [18] Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet, 2024. URL <https://arxiv.org/abs/2408.15221>.
- [19] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Kailong Wang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study, 2024. URL <https://arxiv.org/abs/2305.13860>.
- [20] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.
- [21] Fábio Perez and Ian Ribeiro. Ignore previous prompt: Attack techniques for language models, 2022. URL <https://arxiv.org/abs/2211.09527>.
- [22] Fabio Pernisi, Dirk Hovy, and Paul Röttger. Compromesso! italian many-shot jailbreaks undermine the safety of large language models, 2024. URL <https://arxiv.org/abs/2408.04522>.
- [23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL <https://arxiv.org/abs/2305.18290>.
- [24] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. Tricking llms into disobedience: Formalizing, analyzing, and detecting jailbreaks, 2024. URL <https://arxiv.org/abs/2305.14965>.
- [25] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack, 2025. URL <https://arxiv.org/abs/2404.01833>.
- [26] Sander Schulhoff. Obfuscation/token smuggling. Learn Prompting – Prompt Engineering Guide, March 2025. URL https://learnprompting.org/docs/prompt_hacking/offensive_measures/obfuscation. Last updated March 25, 2025. Accessed September 30, 2025.

- [27] Sander Schulhoff, Jeremy Pinto, Ansum Khan, Louis-François Bouchard, Chenglei Si, Svetlana Anati, Valen Tagliabue, Anson Kost, Christopher Carnahan, and Jordan Boyd-Graber. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4945–4977, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.302. URL <https://aclanthology.org/2023.emnlp-main.302/>.
- [28] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models, 2024. URL <https://arxiv.org/abs/2308.03825>.
- [29] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020. URL <https://arxiv.org/abs/2009.01325>.
- [30] Megan Ung, Jing Xu, and Y-Lan Boureau. Safer dialogues: Taking feedback gracefully after conversational safety failures, 2022. URL <https://arxiv.org/abs/2110.07518>.
- [31] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *ArXiv*, abs/2307.02483, 2023. URL <https://api.semanticscholar.org/CorpusID:259342528>.
- [32] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL <https://doi.org/10.1145/3531146.3533088>.
- [33] Xikang Yang, Xuehai Tang, Jizhong Han, and Songlin Hu. The dark side of trust: Authority citation-driven jailbreak attacks on large language models, 2024. URL <https://arxiv.org/abs/2411.11407>.
- [34] Siboy Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaying Song, Ke Xu, and Qi Li. Jailbreak attacks and defenses against large language models: A survey, 2024. URL <https://arxiv.org/abs/2407.04295>.
- [35] Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Gao Zuchen, Fei Mi, and Lanqing Hong. CoSafe: Evaluating large language model safety in multi-turn dialogue coreference. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17494–17508, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.968. URL <https://aclanthology.org/2024.emnlp-main.968/>.
- [36] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. Don't listen to me: Understanding and exploring jailbreak prompts of large language models, 2024. URL <https://arxiv.org/abs/2403.17336>.
- [37] Farzana Zahid, Anjalika Sewwandi, Lee Brandon, Vimal Kumar, and Roopak Sinha. Securing educational llms: A generalised taxonomy of attacks on llms and dread risk assessment, 2025. URL <https://arxiv.org/abs/2508.08629>.
- [38] Yi Zeng et al. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms, 2024. URL <https://arxiv.org/abs/2401.06373>.
- [39] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL <https://arxiv.org/abs/1909.08593>.
- [40] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL <https://arxiv.org/abs/2307.15043>.