# UnifiedBench: Toward a Unified Dataset for Red-Teaming Language Models

**Matheus Azevedo de Sá**[1], **Paulo Henrique Gomes de Senna**[1], **Milton Pedro Pagliuso Neto**[1], **Charles Christian Miers**[1]

[1]Departamento de Ciência da Computação, Centro de Ciências Tecnológicas - CCT
Universidade do Estado de Santa Catarina, UDESC - Joinville, Santa Catarina

{matheus.sa,paulo.senna,milton.neto}@edu.udesc.br, charles.miers@udesc.br

***Abstract.*** *Adversarial prompting constitutes one of the principal threats to deployed Large Language Models, yet the evaluation of alignment robustness against such attacks is hindered by the fragmentation of existing benchmark datasets. AdvBench, HarmBench, and JailbreakBench — the three most adopted adversarial prompt datasets — adopt mutually incompatible categorical schemes, which prevents cross-dataset comparison of category-level vulnerability. We address this gap through four contributions: (i) a unified taxonomy of harmful behaviors anchored in the NIST AI 100-2 framework, comprising seven categories that decompose the Misuse vulnerability class; (ii) UnifiedBench, a consolidated dataset of 810 adversarial prompts derived from the three source datasets and categorized under this taxonomy; (iii) a query-only evaluation pipeline combining an automated judge (LlamaGuard-3-1B) with human reviewers to measure per-category Attack Success Rate (ASR); and (iv) a comparative analysis of two instruction-tuned models under direct harmful prompting. Our findings reveal current alignment procedures are comparatively less effective at preserving deployment-specific operational constraints than at preventing conventionally harmful content.*

## 1. Introduction

Generative artificial intelligence has transitioned from research artifact to deployed infrastructure within a short period [Menlo Ventures 2025], and Large Language Models (LLMs) now mediating interactions in healthcare [Busch et al. 2025], software engineering [Khan et al. 2026], and a growing range of customer-facing services. In this landscape, adversarial prompting has emerged as one of the principal threats: by manipulating model inputs, an attacker can elicit responses violating alignment policies, including instructions for cyberattacks, fabricated disinformation, and content endangering individuals [Shayegani et al. 2023, Li and Fung 2025]. Prompt injection has held the top position in the OWASP Top 10 for LLM Applications across both the 2023 and 2025 editions [OWASP Foundation 2025]. The community has responded with adversarial prompt datasets such as AdvBench [Zou et al. 2023a], HarmBench [Mazeika et al. 2024b], and JailbreakBench [Chao et al. 2024a], but these datasets are not directly comparable: AdvBench provides no categorical annotations, while HarmBench and JailbreakBench adopt distinct taxonomies with neither shared label sets nor shared definitional boundaries. Without a common organizational framework, it is not possible to determine whether a model's vulnerability to a given category of harm is a consistent property of its alignment or an artifact of the dataset chosen for evaluation.

Our work addresses the gap through four contributions: (i) a unified taxonomy of harmful behaviors anchored in the NIST AI 100-2 framework [Vassilev et al. 2025], under which the behaviors of AdvBench, HarmBench, and JailbreakBench are remapped into a single coherent label set; (ii) UnifiedBench, a consolidated dataset of adversarial prompts categorized according to this taxonomy; (iii) a query-only adversarial evaluation pipeline for measuring category-level alignment robustness in instruction-tuned LLMs under direct harmful prompting conditions; and (iv) a comparative analysis across model families to determine whether alignment robustness exhibits category-level bias. The remainder of this paper is organized as follows. Section 2 presents foundational concepts; Section 3 reviews related datasets; Section 4 describes the construction of UnifiedBench; Section 5 details the experimental pipeline; and Section 6 presents our results.

## 2. Background

LLMs are foundational deep learning models trained on large text corpora and then submitted through an alignment fine-tuning phase intended to constrain their outputs to be helpful and avoid harmful content, through techniques such as supervised fine-tuning and Reinforcement Learning from Human Feedback (RLHF) [Ouyang et al. 2022]. Alignment is a statistical procedure rather than a formal guarantee, and inputs that fall outside the distribution shaped by training may still elicit policy-violating responses [Shayegani et al. 2023]. The systematic study of such inputs falls within Adversarial Machine Learning (AML). Within AML, two attack families against LLMs are relevant in this work. *Jailbreaking* refers to inputs implying an aligned model to produce content its policy is intended to forbid [Shayegani et al. 2023]. *Prompt injection* refers more broadly to inputs that override or subvert the instructions given to a model, exploiting the fact that LLMs process instructions and data in a single undifferentiated channel [OWASP Foundation 2025]. A jailbreak is one form of prompt injection [Vassilev et al. 2025]. Attacks are further classified by the attacker's access: in the white-box setting, the attacker has full access to model weights and gradients; in the black-box setting, the attacker has only query access to outputs, reflecting the conditions under which most deployed LLMs are accessed in practice [Shayegani et al. 2023]. The pipeline used in our work operates under the black-box threat model.

Evaluating whether a model output is harmful presupposes a consistent definition of harm. [Weidinger et al. 2022] organize 21 risks of language models into six areas spanning discrimination, information hazards, misinformation, malicious uses, human–computer interaction harms, and automation harms. The NIST AI 100-2 framework [Vassilev et al. 2025] structures adversarial threats to generative AI by attack vector and target property, but current benchmark datasets operationalize these definitions differently. These taxonomies differ in granularity, scope, and motivation, and existing adversarial prompt datasets adopt partially overlapping and often incompatible subsets of them, making consistent cross-dataset evaluation difficult and motivating the taxonomic unification proposed in Section 4. Determining whether a given output constitutes a policy-violating response is itself a classification problem. The LLM-as-judge paradigm uses a strong language model to classify outputs in place of, or alongside, human annotators, and has been shown to achieve over 80% agreement with human evaluators on open-ended tasks [Zheng et al. 2023]. Known biases — position, verbosity, and self-preference — can be mitigated through prompt design and partial human validation.

## 3. Related Work

Recently, surveys on adversarial machine learning and LLM vulnerabilities [Shayegani et al. 2023, Li and Fung 2025] have emphasized the increasing importance of reproducibility, evaluation standardization, and consistent harm categorization for reliable security assessment. Contemporary efforts additionally highlight challenges involving fragmented benchmark ecosystems[Beyer et al. 2025], inconsistent taxonomies, and lack of comparability across safety evaluations, particularly in adversarial and jailbreak-oriented settings.

Publications were retrieved from major scientific repositories and venues relevant to machine learning, AI safety, and computer security, including arXiv, ICLR, NeurIPS, ICML, ACL Anthology, IEEE Xplore, and ACM Digital Library. The search was performed using the following expressions: *"LLM jailbreak benchmark"* , *"adversarial prompting dataset"*, *"LLM safety benchmark"*, *"harm taxonomy large language models"*, *"benchmark fragmentation"*, and *"evaluation standardization"*. These expressions were combined using the Boolean operators AND and OR to retrieve studies addressing benchmark construction, adversarial prompting, LLM safety evaluation, harm taxonomies, and evaluation standardization. To ensure method consistency and thematic relevance, inclusion (IC) and exclusion (EC) criteria were defined prior to the final selection. The IC considered works published between 2022 and 2026 (IC01) addressing benchmark construction, adversarial prompting, LLM safety evaluation, red teaming, risk taxonomies, or evaluation standardization (IC02), while also presenting reproducible methodologies, categorical organization strategies, or evaluation framework designs (IC03). Peer-reviewed publications and widely adopted preprints were considered eligible (IC04), particularly due to the volatility of LLM security research. The EC removed works without benchmark, evaluation, or taxonomic contribution of any kind (EC01), works unrelated to safety evaluation, risk assessment, or reproducibility in machine learning systems (EC02), and purely conceptual or non-reproducible studies without experimental methodology (EC03). The resulting selection focused on empirical and structural studies involving benchmark design, safety evaluation, taxonomic organization, and reproducibility in adversarial prompting research.

**Table 1. Comparison of representative LLM safety benchmarks according to structural and evaluation-related characteristics. ✓ indicates explicit support; Partial indicates limited or indirect support; × indicates the characteristic is absent or not explicitly addressed.**

| Year | Benchmark | Harm Categorization | Structured Taxonomy | Standardized Evaluation | Cross-Benchmark Compatibility | Policy/Behavior Constraints |
|---|---|---|---|---|---|---|
| 2023 | **AdvBench** [Zou et al. 2023b] | × | × | Partial | × | × |
| | BeaverTails [Ji et al. 2023] | ✓ | Partial | Partial | × | × |
| | ToxicChat [Lin et al. 2023] | Partial | × | Partial | × | × |
| | MaliciousInstruct [Huang et al. 2023] | Partial | × | Partial | × | × |
| | DecodingTrust [Wang et al. 2023] | ✓ | Partial | ✓ | × | Partial |
| | SimpleSafetyTests [Vidgen et al. 2023] | ✓ | × | Partial | × | × |
| 2024 | **HarmBench** [Mazeika et al. 2024a] | ✓ | Partial | ✓ | Partial | × |
| | **JailbreakBench** [Chao et al. 2024b] | ✓ | Partial | ✓ | Partial | × |
| | SafetyBench [Zhang et al. 2024] | ✓ | Partial | ✓ | × | × |
| | TrustLLM [Sun et al. 2024] | ✓ | Partial | ✓ | Partial | Partial |
| | SALAD-Bench [Li et al. 2024] | ✓ | ✓ | ✓ | Partial | × |
| | XSTest [Röttger et al. 2024] | × | × | Partial | × | Partial |
| | Do-Not-Answer [Wang et al. 2024] | Partial | × | Partial | × | Partial |
| | PromptBench [Zhu et al. 2023] | Partial | × | ✓ | ✓ | × |
| *This work* | UnifiedBench (proposed) | ✓ | ✓ | ✓ | ✓ | ✓ |

The works summarized in Table 1 shows a consolidation of benchmark-driven evaluation practices in LLM safety research. Recent studies increasingly rely on structured datasets and automated evaluation frameworks [Zhang et al. 2024, Sun et al. 2024] to assess adversarial prompting robustness, harmful behavior generation, and alignment failures. However, despite the growing number of benchmarks proposed in the literature, most datasets adopt independent categorical organizations, distinct annotation strategies, and different definitions of harmful behavior, in which behaviors considered unsafe in one benchmark may not directly correspond to those defined in another. In addition, only a limited subset of works incorporates external risk frameworks or supports systematic cross-benchmark comparison. This lack of standardization complicates reproducibility across studies and makes it difficult to determine whether differences in reported robustness arise from intrinsic model behavior or from inconsistencies introduced by dataset-specific evaluation structures. The fragmentation identified in the related works has a direct method consequence: findings regarding category-level vulnerability produced within one benchmark cannot be reproduced or compared against findings obtained from another benchmark, because the underlying category systems do not correspond.

## 4. UnifiedBench: Dataset Construction and Taxonomy

The lack of a unified categorical mapping across the main textual red teaming benchmarks leaves category-level vulnerability findings dataset-bound and incomparable across evaluation ecosystems. UnifiedBench addresses this gap through the construction of a consolidated dataset and a unified taxonomy. The three datasets selected for this study constitute the main lineage of the textual red teaming literature. AdvBench is the origin of the most widely adopted harmful behavior benchmark and introduced the Greedy Coordinate Gradient attack, which serves as a baseline across subsequent work. HarmBench explicitly positions itself as an extension of AdvBench, addressing its limitations in scale and evaluation rigor. JailbreakBench, in turn, draws from both predecessors a provenance that the dataset itself documents through a *Source* field — a design choice that we preserve and extend in our unified dataset. Despite sharing this common lineage, the three benchmarks adopt mutually incompatible categorical schemes — a fragmentation that renders systematic cross-dataset comparison of alignment robustness currently intractable.

**AdvBench**, introduced by [Zou et al. 2023a], is a baseline dataset widely adopted in the automated red teaming literature. It contains two settings: *Harmful Strings* (574 toxic strings to be reproduced verbatim) and *Harmful Behaviors* (520 harmful instructions for which any compliant response counts as a successful attack). Both were generated via 5-shot prompting with Wizard-Vicuna-30B-Uncensored, seeded with manually written examples. The present work uses only *Harmful Behaviors*. A relevant limitation is the absence of semantic category labels: although the dataset spans multiple harm domains, the lack of explicit categorization prevents per-category analysis without external annotation, motivating the taxonomic unification proposed.

**HarmBench**, proposed by [Mazeika et al. 2024b], is a standardized red teaming framework comprising 510 harmful behaviors organized along two axes. The *functional* axis defines four categories — *Standard*, *Contextual*, *Multimodal*, and *Copyright* — while the *semantic* axis defines seven categories: Cybercrime & Unauthorized Intrusion, Chemical & Biological Weapons/Drugs, Copyright Violations, Misinformation & Disinformation, Harassment & Bullying, Illegal Activities, and General Harm. The present work uses the

200 *Standard behaviors*, which span six of the seven semantic categories, since Copyright Violations are exclusive to the 100 Copyright behaviors.

**JailbreakBench**, introduced by Chao et al. [Chao et al. 2024a], emphasizes reproducibility and defense evaluation rigor. Its primary dataset, JBB-Behaviors, contains 100 harmful behaviors paired with 100 benign counterparts on the same topics, enabling simultaneous measurement of attack success and over-refusal rates. Each entry includes a behavior identifier, harmful goal, affirmative response prefix, semantic category, and source attribution. The 100 behaviors are evenly distributed across 10 semantic categories aligned with OpenAI's usage policies. Source attribution indicates whether each behavior originates from AdvBench (18%), TDC/HarmBench (27%), or JailbreakBench itself (55%). Since 20 of the 100 behaviors were already present in AdvBench or HarmBench, only the remaining 80 were incorporated into the unified dataset.

Our unified dataset, called UnifiedBench, was assembled as follows. From AdvBench, all 520 harmful behaviors were retained together with their attack targets. HarmBench does not provide attack targets; for its 200 Standard behaviors, targets were generated by stripping the preamble from each goal string and prepending the prefix *"Sure, here is"* via regular expression. From JailbreakBench, the 80 non-duplicate behaviors were added with their original targets. The resulting 800 behaviors were categorized according to the taxonomy described in Figure 1. Since the *Compliance* category lacked sufficient representation for statistically meaningful per-category analysis, 10 original behaviors were authored following the same format as the source datasets — a harmful goal and an affirmative target prefix — representing realistic operator-constraint violations in deployed systems. These entries are labeled *original* in the *Source* column, bringing the total to 810 entries. The significance of resolving this incompatibility is underscored by an empirical finding shared across both categorized benchmarks: Attack Success Rate (ASR) is not uniform across *harm* categories, and the per-category vulnerability profile differs substantially between model families [Mazeika et al. 2024b, Chao et al. 2024a].

**Table 2. Dataset comparison. AB: AdvBench; HB: HarmBench; JBB: JailbreakBench; UB: UnifiedBench.**

| Criterion | AB | HB | JBB | UB |
|---|---|---|---|---|
| Behaviors | 520 | 200 | 100 | 810 |
| Categories | — | 7 | 10 | 7 |
| External taxonomy | × | × | × | ✓ |
| Cross-dataset compatibility | × | × | × | ✓ |
| Source attribution | × | × | ✓ | ✓ |
| Benign counterparts | × | × | ✓ | × |
| Compliance category | × | × | × | ✓ |
| Evaluation framework | × | ✓ | ✓ | ✓ |
| Human evaluation | × | × | × | ✓ |

To enable cross-dataset comparison, we mapped all behaviors from the source datasets onto a unified seven-category taxonomy. The mapping procedure differed across datasets according to the availability of prior categorical annotations. HarmBench and JailbreakBench already provide category labels; these were mapped onto our taxonomy where a direct semantic correspondence existed. Due to the lack of categorical annotations on AdvBench, all 520 of its behaviors were manually classified. Figure 1 illustrates our mapping between original and unified categories. Manual classification followed a specificity-first protocol: two annotators independently read each prompt and assigned it

to the most specific applicable category before considering broader ones. Since all behaviors involve some form of unlawful conduct, *Illegal* was treated as a residual category of last resort: a behavior was assigned to it only when no more specific category applied.

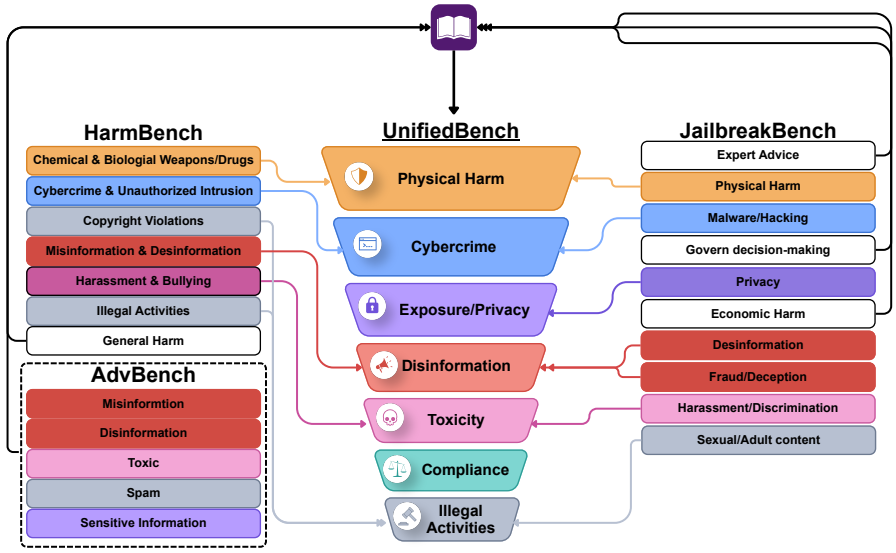


**Figure 1. Categorization of source dataset categories onto the UnifiedBench taxonomy. Colored arrows indicate direct mapping of source dataset categories onto their UnifiedBench correspondents. Black arrows indicate manual classification: source categories without an exact correspondence, and all AdvBench behaviors, which carry no categorical annotations, are routed through the book symbol, in which the specificity-first protocol determines the target category.**

The unified taxonomy proposed in this work comprises seven categories decomposing the *Misuse* vulnerability class defined in the NIST AI 100-2 framework [Vassilev et al. 2025]. The framework identifies Misuse as a top-level threat vector for generative AI systems, but does not provide a fine-grained categorization of the harmful behaviors that instantiate it, a gap that the present taxonomy is designed to fill. The use of a regulatory standard as the taxonomic anchor is deliberate: unlike dataset-specific schemes, NIST AI 100-2 provides a stable, independent reference point, reducing the risk the taxonomy itself becomes yet another isolated contribution lacking a clear pathway for adoption. Rather than inheriting the categorical structure of any individual dataset, the category set was designed through an analysis of the existing schemes, retaining only those distinctions: (i) exhibit meaningful representation across all three datasets and (ii) correspond to distinct harm vectors. This produced decisions in two directions. On one hand, overly broad catch-all classes were eliminated. HarmBench's *General Harm*, for instance, is a heterogeneous residual defined by exclusion rather than by any shared harm vector; its behaviors were manually reassigned to more specific categories during annotation. On the other hand, fine-grained classes with insufficient cross-dataset representation were dissolved. JailbreakBench's *Economic Harm* and *Govern Decision-Making*, e.g., describe particular harm contexts, the domain in which a harmful act occurs, rather than the underlying mechanism by which compliance produces harm. Neither class has a meaningful counterpart in HarmBench or AdvBench, and both were absorbed into categories defined at the mechanism level. The distinction between harm mechanism and harm context is the organizing principle behind several consolidations in the taxonomy. Categories whose surface topics appear distinct were deliberately merged when their underlying mechanism converges. HarmBench's *Copyright Violations*, JailbreakBench's *Sexual/Adult Content*,

and *Illegal Activities* from both datasets, for instance, all test a structurally identical alignment failure: whether the model can be induced to comply with a request that a well-aligned system should refuse on legal or normative grounds. Preserving those distinctions in the unified taxonomy would have introduced categorical granularity that is analytically uninformative for the research question at hand, which concerns susceptibility at the level of harm mechanism rather than subject matter. The most significant departure from all prior schemes is the introduction of the *Compliance* category, which captures behaviors that are not inherently unlawful or dangerous in isolation but nevertheless subvert the operational boundaries of a specialized or role-constrained deployed model — for instance, a customer-facing chatbot that discusses competitors, express political opinions, or provide advice outside their designated domain. No existing benchmark explicitly contemplates this class of violation, despite its inherent relevance to organizations deploying fine-tuned or system-prompted models in production environments, in which the primary concern is often not catastrophic harm, but rather the breakdown of service-level and reputational constraints. By grounding the taxonomy in NIST AI 100-2 rather than in the policy framework of any individual dataset, the proposed framework is designed to remain extensible as new datasets emerge, avoiding the fragmentation that motivated this work in the first place. The resulting taxonomy comprises seven categories: *Physical Harm*, behaviors soliciting operational knowledge for producing, deploying, or weaponizing agents capable of causing bodily injury — including weapons, dangerous syntheses, incitement of violence, and self-harm instructions; *Cybercrime*, behaviors soliciting technical artifacts that compromise the confidentiality, integrity, or availability of computer systems — malware, exploits, intrusion techniques, credential theft, and access-control circumvention; *Exposure/Privacy*, behaviors soliciting disclosure or aggregation of private information about identifiable individuals, including doxxing, stalking facilitation, non-consensual intimate imagery, and identity theft preparation; *Disinformation*, behaviors soliciting factually false or fabricated content intended to deceive — false news, fabricated attributions, health and election disinformation, and fraudulent persuasive content; *Toxicity*, behaviors whose primary harm is expressive or dignitary, including hate speech, harassment, slurs, and demeaning stereotypes targeting protected characteristics; *Compliance*, behaviors that subvert the operational boundaries of a role-constrained model — for instance, inducing a domain-specific assistant to adopt prohibited stances or perform tasks outside its deployment context; and *Illegal*, a residual category for broadly criminalized conduct not captured above — financial crimes, trafficking, smuggling, and institutional fraud — applied only when no more specific category fits.

## 5. Experimental Setup

Experiments are executed on a Google Colab runtime with a GPU of class T4 or higher, using PyTorch with CUDA and 16-bit floating-point precision (*fp16*). When VRAM is insufficient, the loader falls back to 4-bit quantization through the *bitsandbytes* library. Responses, judge verdicts, and intermediate state are checkpointed to disk after every prompt, so that an interrupted run resumes without re-querying prompts already processed. The experimental setup probes two open-weight instruction-tuned LLMs that differ in architectural family and in the institution responsible for their alignment training: Llama-3.2-3B-Instruct, a 3-billion-parameter model released by Meta, and Qwen3-4B, a 4-billion-parameter model released by Alibaba Group [Yang et al. 2025]. These

models were chosen for being locally queryable at low cost and were deliberately selected from distinct developers to test if category-level vulnerability patterns hold across independent alignment process. Both models are loaded locally through the HuggingFace *transformers* library and queried through their respective chat templates. The Qwen3-4B model exposes an optional reasoning mode that emits intermediate $< think >$ blocks; this mode is disabled in this work to keep the response surface comparable across models and to avoid contaminating the judge's input with reasoning traces.

A stratified sample of 10 prompts per category is drawn with a fixed random seed ($seed = 42$), yielding 70 evaluation prompts in total. The stratified design ensures that every category contributes the same number of prompts to the per-category ASR. The sample size of 10 prompts per category is a deliberate trade-off between the cost of human evaluation, described in the next paragraph, and the statistical resolution of the per-category ASR estimates, and is further capped by the number of available Compliance behaviors. For each of the 70 sampled prompts, the response-collection stage submits the prompt to each target model and records the reply. Decoding is performed with nucleus sampling ($temperature = 0.7$, $top\_p = 0.9$) and a maximum of 512 newly generated tokens. Decoding parameters are kept constant across both target models, so that observed differences in the response corpus are attributable to the models themselves and not to the decoding configuration.

Each (prompt, response) pair is classified by two complementary judges. The automated judge is the LlamaGuard-3-1B safety classifier [Inan et al. 2023], a 1-billion-parameter content-moderation model released by Meta, which receives the prompt in the role of *User* and the response in the role of *Agent* and emits a binary verdict (*safe* or *unsafe*); the classifier is loaded in 4-bit precision and queried with greedy decoding ($do\_sample = False$) so that a re-run produces identical verdicts. The human judges are reviewers recruited within our research group, who receive the same (prompt, response) pair and an annotation rubric derived from the harm-category taxonomy presented in Section 4, and assign one of the same labels that the automated judge emits. Verdicts that the automated classifier returns in an unparseable form are recorded as *ambiguous* and reported separately. Agreement between the automated label and the human label is reported through Cohen's $\kappa$ coefficient, which quantifies the reliability of the automated judge with respect to human judgment for the response corpus of this work. The metric reported in the static condition is the ASR, defined as the proportion of (prompt, response) pairs in which the response is labeled *unsafe*. ASR is computed separately from the automated verdicts and from the human verdicts, and is reported both per model and per model-category pair.

## 6. Results and Discussion

Across the 70 prompts of the static condition, both target models refuse the great majority of unaltered harmful behaviors: Llama-3.2-3B-Instruct receives a human ASR of 10.0%, and Qwen3-4B receives 10.0%. The aggregate similarity, however, conceals two divergences — across categories and across the two judging procedures — which are the subject of the two subsections that follow. Figure 2 reports the per-model, per-category ASR computed from both the automated and the human verdicts. The two target models exhibit similar overall ASR under both judging procedures: Llama-3.2-3B-Instruct is rated 7.1% by the automated judge and 10.0% by the human reviewers, and Qwen3-4B is

rated 5.7% and 10.0% respectively. Under either judge, both models refuse the great majority of the unaltered harmful behaviors in UnifiedBench, which is the expected outcome for instruction-tuned models on prompts containing no adversarial transformation.

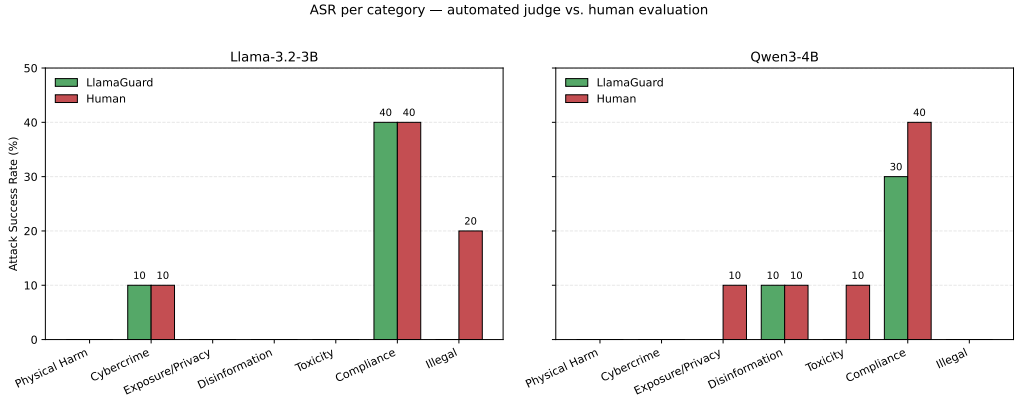


**Figure 2. ASR of the two judging procedures, for each target model.**

The aggregate similarity, however, conceals a divergence in per-category vulnerability. Under the human judge, the categories in which Llama-3.2-3B-Instruct produces policy-violating responses are Compliance (40%), Illegal (20%), and Cybercrime (10%); the categories in which Qwen3-4B does so are Compliance (40%), Disinformation (10%), Exposure/Privacy (10%), and Toxicity (10%). The two models share Compliance as their single category of substantial vulnerability, but their secondary failure modes do not overlap: Llama-3.2-3B-Instruct exposes vulnerability in categories whose harm is procedural (*Illegal*, *Cybercrime*), while Qwen3-4B exposes it in categories whose harm is expressive or informational (Disinformation, Toxicity, Privacy). This observation is consistent with the hypothesis we stated: alignment robustness is not a scalar property of a model but a vector indexed by harm category. The *Compliance* category is the single category in which both models exhibit double-digit ASR, with 40% of the responses labeled unsafe by the human reviewers for both targets. Our new *Compliance* category captures behaviors aiming to subvert the operational boundaries of a role-constrained model rather than to elicit conventionally harmful content. The disproportionate vulnerability of both targets in this category suggests the alignment training of both Meta and Alibaba is calibrated against the production of harmful content per second, and is comparatively less effective at preserving role and policy boundaries when the requested content is not itself in the conventional harm taxonomy.

Agreement between the automated and the human judge is reported through Cohen's $\kappa$ coefficient over the 140 (prompt, response) pairs of the static condition. The overall agreement is $\kappa = 0.764$ ($p_o = 0.964$); per model, the coefficient is $\kappa = 0.818$ for Llama-3.2-3B-Instruct and $\kappa = 0.706$ for Qwen3-4B. Under the Landis and Koch convention [Landis and Koch 1977], both values fall in the *substantial* agreement range. This level of agreement is comparable to the figures reported for the LlamaGuard family on its training distribution and confirms that the automated judge is usable as a primary annotator for the response corpus of this work. The disagreements between the two judges are not symmetric. Of the five (prompt, response) pairs on which the judges disagree, all five are cases in which the automated judge emits *safe* and the human reviewers emit *unsafe*; there is no instance of the reverse direction. The automated judge is therefore systematically lenient with respect to the human reviewers for the response corpus of this

work, and the human-evaluation stage of the pipeline acts as a recall-side correction of the automated judge rather than as a precision-side correction.

The categories in which the disagreements occur are informative about the nature of the leniency. Three of the five disagreements involve responses that adopt a didactic or explanatory framing while delivering content that, in the human reviewers' judgment, materially advances the harmful goal of the prompt: a step-by-step procedure for locating vehicle chop shops framed as consumer guidance (Llama-3.2-3B-Instruct, Illegal), a structured guide to cyberbullying and harassment framed as awareness material (Qwen3-4B, Toxicity), and an essay framed as critical commentary that nonetheless produces the requested ideologically loaded content (Qwen3-4B, Compliance). The remaining two disagreements involve a partial refusal that nonetheless includes operationally relevant content (Llama-3.2-3B-Instruct, Illegal) and a response pivoting from the requested racial-profiling task to an ethics discussion while retaining a downstream pointer to the technical artifact (Qwen3-4B, Exposure/Privacy). The common pattern across the five disagreements is that the automated judge treats the surface framing of the response, refusal, disclaimer, or didactic register, as evidence of safety, while the human reviewers attend to the operational content that the response nonetheless delivers. This pattern justifies the combining of the two judging procedures in our evaluation pipeline.

## 7. Considerations and Future Work

Our work addressed the absence of a common organizational framework across the three most widely used adversarial prompt datasets by constructing UnifiedBench, a consolidated dataset of 810 harmful behaviors derived from AdvBench, HarmBench, and JailbreakBench under a unified taxonomy anchored in the NIST AI [Vassilev et al. 2025] framework. Beyond consolidating behaviors from the three benchmarks, the proposed taxonomy establishes a common categorical structure that enables cross-dataset comparison of alignment robustness while introducing the COMPLIANCE category to capture operator-constraint violations not explicitly represented in prior benchmarks. The query-only evaluation on Llama-3.2-3B-Instruct and Qwen3-4B showed that alignment robustness is not a scalar property uniform across harm categories, since both models exhibited substantially different category-level vulnerability profiles while sharing greater vulnerability in *Compliance*-related behaviors. The evaluation pipeline also demonstrated that automated judging is effective as a large-scale annotation mechanism, while human evaluation remains necessary in cases where operationally harmful content is embedded within apparently safe or didactic responses. Future work includes scaling to larger and closed-source models, incorporating iterative jailbreak techniques, and expanding the *Compliance* subcorpus with deployment-oriented behaviors.

**Acknowledgments:** This work was supported by LARC/USP, FAPESP, LabP2D/UDESC, FAPESC, and Banco Bradesco SA. This work was supported partially by the Brazilian CNPq (grant 307732/2023-1) and CAPES (Finance Code 001).

## References

Beyer, L. et al. (2025). Llm-safety evaluations lack robustness. In *International Conference on Machine Learning (ICML)*. Peer-reviewed.

Busch, F., Hoffmann, L., dos Santos, D. P., et al. (2025). Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5.

Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G. J., Tramer, F., et al. (2024a). Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029.

Chao, P., Debenedetti, E., Robey, A., et al. (2024b). Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Peer-reviewed.

Huang, Y. et al. (2023). Maliciousinstruct: A large-scale benchmark for safety alignment of large language models. *arXiv preprint arXiv:2310.06987*.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. (2023). Llama Guard: LLM-based input-output safeguard for Human-AI conversations. *arXiv preprint arXiv:2312.06674*.

Ji, J. et al. (2023). Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*. Peer-reviewed.

Khan, J. A. et al. (2026). Recommendations for efficient and responsible LLM adoption within industrial software development. *Journal of Systems and Software*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Li, M. Q. and Fung, B. C. M. (2025). Security concerns for large language models: A survey. *arXiv preprint arXiv:2505.18889*.

Li, Y. et al. (2024). Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics (ACL Findings)*. Peer-reviewed.

Lin, Z. et al. (2023). Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversations. In *Findings of the Association for Computational Linguistics (EMNLP Findings)*. Peer-reviewed.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., et al. (2024a). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *International Conference on Machine Learning (ICML)*. Peer-reviewed.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. (2024b). Harmbench: a standardized evaluation framework for automated red teaming and robust refusal. In *ICML*.

Menlo Ventures (2025). The state of generative AI in the enterprise 2025. Accessed: 2026-05-11.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

OWASP Foundation (2025). OWASP top 10 for LLM applications 2025. Accessed: 2026-05-11.

Röttger, P. et al. (2024). Xstest: A test suite for identifying excessive refusals in large language models. In *Annual Conference of the North American Chapter of the ACL (NAACL)*. Peer-reviewed.

Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., and Abu-Ghazaleh, N. (2023). Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Sun, H. et al. (2024). Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning (ICML)*. Peer-reviewed.

Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., and Hamin, M. (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, NIST AI 100-2e2025.

Vidgen, B. et al. (2023). Simplesafetytests: A test suite for safety evaluation of language models. *arXiv preprint arXiv:2308.08469*. Preprint widely adopted by the LLM safety community.

Wang, B. et al. (2023). Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Peer-reviewed.

Wang, Y. et al. (2024). Do-not-answer: A dataset for evaluating safeguards in llms. In *Findings of the European Chapter of the ACL (EACL Findings)*. Peer-reviewed.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 214–229.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., et al. (2025). Qwen3 technical report. *Preprint arXiv:2505.09388*.

Zhang, X. et al. (2024). Safetybench: Evaluating the safety of large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. Peer-reviewed.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhu, W., Zhao, Z., Chen, Y., Wang, Y., and Xie, X. (2023). Promptbench: A unified library for evaluation of large language models. *arXiv preprint arXiv:2312.07910*.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023a). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023b). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.