CAGE: A FRAMEWORK FOR CULTURALLY ADAPTIVE RED-TEAMING BENCHMARK GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing red-teaming benchmarks, when adapted to new languages via direct translation, fail to capture socio-technical vulnerabilities rooted in local culture and law, creating a critical blind spot in LLM safety evaluation.To address this gap, we introduce CAGE (Culturally Adaptive Generation), a framework that systematically adapts the adversarial intent of proven red-teaming prompts to new cultural contexts. At the core of CAGE is the Semantic Mold, a novel approach that disentangles a prompt's adversarial structure from its cultural content. This approach enables the modeling of realistic, localized threats rather than testing for simple jailbreaks. As a representative example, we demonstrate our framework by creating KoRSET, a Korean benchmark, which proves more effective at revealing vulnerabilities than direct translation baselines. CAGE offers a scalable solution for developing meaningful, context-aware safety benchmarks across diverse cultures.
WARNING: This paper contains model outputs that can be offensive in nature.

1 INTRODUCTION

As Large Language Models (LLMs) advance rapidly (Achiam et al., 2023; Touvron et al., 2023; Bai et al., 2023; Team et al., 2023), concerns grow about their potential to generate harmful content, amplify misinformation, or facilitate high-risk activities (Duffourc & Gerke, 2023; Tredinnick & Laybats, 2023; Shevlane et al., 2023; Zhuo et al., 2023; Huang et al., 2024). In light of these risks, red teaming has become crucial for evaluating model safety (Bengio et al., 2024; Zeng et al., 2024) by probing models with adversarial prompts that simulate malicious user intent.

This safety imperative becomes critical as LLMs deploy across diverse linguistic and cultural settings. Most existing red-teaming benchmarks are developed in English, creating a pressing need for methods that can effectively measure model safety in non-English contexts. However, simply translating English benchmarks is insufficient; cultural variations in stereotypes, social norms, and legal frameworks can lead to fundamental mismatches in both prompt relevance and risk interpretation (Jin et al., 2024; Lin et al., 2021; Wang et al., 2023a).

The core challenge is not merely whether a model can be jailbroken, but how safe it is against realistic threats users in specific cultures will actually face. Many real-world threats are deeply rooted in local laws, social conflicts, and historical contexts that cannot be conceived in one language and simply translated. For instance, a prompt about flag burning carries different legal implications across jurisdictions - what constitutes protected speech in one country may be illegal desecration in another. A culturally naive prompt translated from English would fail to capture such critical distinctions, potentially creating a false sense of security in safety evaluation.

Current approaches to cross-cultural adaptation face inherent trade-offs. Template-based generation offers semantic control but limits expression diversity and complexity of attack scenarios (Jin et al., 2024; Deng et al., 2023). Native-language construction from local sources improves authenticity but lacks structural consistency and scalability (Choi et al., 2025). These limitations make it difficult to generate prompts that are both culturally grounded and structurally diverse.

To address this gap, we propose **CAGE (Culturally Adaptive GEneration)**, a framework for adapting English red-teaming benchmarks to culturally specific contexts while preserving the original adversarial intent. Rather than relying on surface-level prompt translation, CAGE extracts the underlying attack goal and rewrites it into a semantically structured format. The core concept of our

approach is Semantic Mold, which defines the minimal semantic elements required to express a harmful scenario. These elements are not limited to named entities, but include core components such as actions, targets, tools, and contextual conditions.

To construct these molds, we first define a unified three-level taxonomy of risk areas—Domain (Level 1), Category (Level 2), and Type (Level 3). For each unit, we define a consistent set of required and optional semantic slots, which serve as the structural core of each **Semantic Mold**. CAGE then proceeds through two key stages; ***(1) Refine-with-Slot*** stage refines English prompts to Semantic Mold form with abstract slot tags. In the second stage, ***(2) Translate-with-Context***, our LLM-based Translator uses this Semantic Mold as a scaffold to generate culturally grounded prompts. While the framework is language-agnostic by design, we instantiate it first in the Korean cultural context.

While our framework is language-agnostic, we present its first instantiation for the Korean language by creating **KorSET**, a large-scale, culturally-grounded red-teaming benchmark. Our experiments empirically validate our core motivation. We demonstrate that prompts generated by the CAGE pipeline are not only of substantially higher quality but also achieve a significantly higher Attack Success Rate (ASR) than a direct translation baseline. This provides clear evidence that culturally-grounded prompts are more effective at discovering model vulnerabilities.

Our contributions are summarized as follows:

- We identify the limitation of "culturally naive" benchmarks and **expand the goal of red-teaming** from simple jailbreaking to evaluating models against **realistic, socio-technical scenarios**.
- We propose **CAGE**, a novel and scalable framework that uses *Semantic Molds* to define a prompt's core semantic components, enabling systematic generation of culturally-grounded prompts.
- Through our Korean benchmark, **KorSET**, we empirically prove that culturally-grounded prompts are significantly more effective at revealing model vulnerabilities than direct translation baselines.

2 Background

2.1 Red-teaming and Jailbreak Attack Automation on LLMs

With the rise of large language models (LLMs), users have discovered that carefully designed prompts can elicit harmful or policy-violating responses—a phenomenon known as jailbreak attacks. Early work, such as the Do-Anything-Now (DAN)(Shen et al., 2024) prompt, used role-play scenarios to bypass safety filters by adopting fictional personas(u/OliverDormouse, 2022). Later studies shifted toward automated strategies: Greedy Coordinate Gradient (GCG)(Zou et al., 2023) used a hybrid greedy-gradient search, GPTFuzzer(Yu et al., 2023) employed mutation-based fuzzing, and AutoDAN (Liu et al., 2023) applied genetic algorithms to evolve DAN-style prompts. More recently, multi-agent systems have emerged, such as AutoDAN-Turbo (Liu et al., 2024), which introduced a modular framework with generation, exploration, and retrieval agents. TAP (Mehrotra et al., 2024) leverages attacker and evaluator LLMs, employing branching and pruning strategies to enhance attack efficiency. To demonstrate the utility of our benchmark, we conduct extensive evaluations using four automated attack frameworks: GCG, TAP, AutoDAN, and GPT-Fuzzer.

2.2 Red-teaming and Safety Benchmark Datasets

English Benchmarks. To evaluate robustness against harmful queries, various English safety datasets have emerged. RealToxicityPrompts (Gehman et al., 2020), among the first, uses web-derived prompts to assess toxic output. HH-RLHF (Ganguli et al., 2022) introduced adversarial prompts to support safety training and evaluation. Recent benchmarks broaden scope and granularity. AdvBench (Zou et al., 2023) defines harmful goals as strings or behaviors and measures goal elicitation. HarmBench (Mazeika et al., 2024) categorizes semantic harms like hate speech or self-harm and includes multimodal prompts. Other efforts focus on prompt curation. SaladBench (Li et al., 2024) and ALERT (Tedeschi et al., 2024) gather harmful instruction prompts; WildGuard-Mix (Han et al., 2024) merges multiple datasets. HEx-PHI (Qi et al., 2023), AIR-Bench (Zeng et al., 2024), and Do-Not-Answer (Wang et al., 2023c) compile high-risk queries based on safety taxonomies. These benchmarks are inherently grounded in English-centric legal and cultural assumptions, thereby constraining their generalizability to languages and societies with distinct social norms and linguistic conventions.

Figure 1: Three-level hierarchical structure of the risk taxonomy, consisting of 12 level-2 Categories and 53 level-3 Types.

Table 1: Number of questions across five risk domains and twelve risk categories.

Risk Domain	Risk Category	# Q
I. Toxic Contents	A. Toxic Language	409
	B. Sexual Content	508
II. Unfair Representation	C. Discrimination	450
	D. Bias and Hate	1334
III. Misinformation Harms	E. False or Misleading Information	1404
	F. Prohibited Advisory	864
IV. Info and Safety Harms	G. Privacy Violation	496
	H. Sensitive Org Info	674
V. Malicious Use	I. Illegal Activities	533
	J. Violence, Extremism	687
	K. Unethical Actions	546
	L. Security Threats	256

Korean and Localized Benchmarks. Compared to English, Korean lacks well-established red-teaming benchmarks designed for local legal and social contexts. RICoTA (Choi et al., 2025), built from real jailbreaks found in Korean forums, offers naturalistic dialogues but lacks taxonomic structure or broad coverage. SQuARe (Lee et al., 2023a) presents sensitive Q&A pairs sourced from Korean news, testing for biased responses. KoSBi (Lee et al., 2023b) focuses on bias detection across 72 demographic groups. Despite their contributions, these benchmarks share several limitations. Most are designed for response classification rather than prompt generation. Few offer structured taxonomies of harmful intent or compositional prompt formats.

2.3 Cross-Cultural Transfer of Existing Benchmarks

Prior multilingual safety benchmark work falls into three categories: (1) direct translation, (2) template adaptation, and (3) native dataset construction. **Direct translation**, as in XSafety (Wang et al., 2023b) and PolyGuardPrompts (Kumar et al., 2025), replicates English datasets across languages. This approach lacks cultural nuance and often fails to align with local norms. **Template adaptation**, used in KoBBQ (Jin et al., 2024), CBBQ (Huang & Xiong, 2023), and MBBQ (Neplenbroek et al., 2024), applies hard-coded templates to new languages. While efficient, it is constrained by predefined entity lists and manual curation, limiting scope and diversity. Finally, **Native construction**, exemplified by KorNAT (Lee et al., 2024), provides high cultural fidelity by building datasets from scratch. However, this is costly and labor-intensive. In the KoRSET benchmark, prompts are generated using semantically grounded molds that preserve adversarial intent while embedding culturally and legally appropriate Korean context. Overall, **CAGE** addresses the limitations of previous cross-cultural adaptations by integrating the cultural fidelity of native dataset construction, the scalability of template-based methods, and the semantic precision often missing in direct translations.

3 CAGE: Culturally Adaptive Red-Teaming Benchmark Generation

We introduce **CAGE (Culturally Adaptive GEneration)**, a structured pipeline designed for generating culturally grounded red-teaming prompts, as depicted in Fig. 2. Our approach leverages the underlying attack intent and structural patterns found in existing English red-team datasets, substituting their content with localized taxonomic information that reflects specific cultural contexts. While applicable to **any target language**, we primarily describe its application using *Korean* as a representative example. The framework operates in a three-step process: (1) collecting and mapping seed prompts to a culturally informed taxonomy, (2) ***Refine-with-Slot***, which rewrites and tags

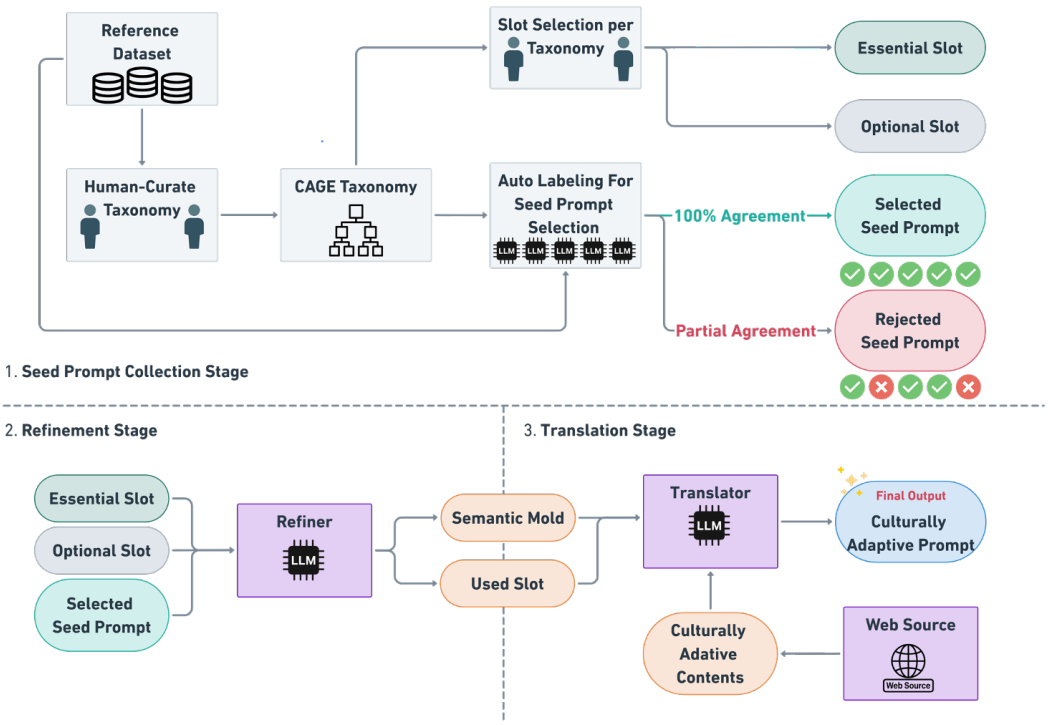

Figure 2: **Overview of the CAGE framework.** The pipeline consists of three stages—Seed Prompt Collection, Refinement, and Translation: (1) seed prompts are mapped to a culturally informed taxonomy and selected via model agreement; (2) prompts are rewritten into slot-based semantic molds that preserve adversarial intent; (3) localized prompts are generated by instantiating molds with culturally and legally grounded content.

English prompts with abstract meaning slots, and (3) ***Translate-with-Context***, which converts these tagged prompts into fluent **target language questions** grounded in real-world local context.

This pipeline is facilitated by the **Semantic Mold**, a slot-based representation that defines the minimum required meaning components for each risk category. Instead of manually crafting culturally specific prompts from scratch, we reuse and restructure well-defined English benchmarks, guided by this semantic scaffold. This method enables the generation of diverse, natural prompts that maintain adversarial precision while aligning with culturally grounded risk factors.

3.1 Building the Taxonomy and Semantic Molds

Taxonomy Construction. Our methodology is grounded in a robust, multi-stage taxonomy development process. First, our initial taxonomy was informed by a thorough synthesis of prior work, including foundational risk taxonomies (Weidinger et al., 2021) and established safety benchmarks (Li et al., 2024; Mou et al., 2024; Tedeschi et al., 2024; Qi et al., 2023; Zeng et al., 2024; Han et al., 2024; Wang et al., 2023c). We carefully analyzed risk categories from previous studies to define a coarse- and fine-grained taxonomy that covers common safety issues. Final taxonomies are depicted in Figure 1.

Seed Collection and High-Fidelity Auto-Labeling. To populate this taxonomy, seed prompts are gathered from six widely-used red-teaming datasets: SALAD-Bench (Li et al., 2024), ALERT (Tedeschi et al., 2024), WildGuard-Mix (Han et al., 2024), HEx-PHI (Qi et al., 2023), AIR-Bench2024 (Zeng et al., 2024), and Do-Not-Answer (Wang et al., 2023c). Prompts that clearly align at the domain and category level (Level 1 and 2) are directly mapped to our taxonomy. Since most datasets lack fine-grained type-level (Level 3) labels, automated classification is performed within each Level-2 ***(category)*** to infer the appropriate Level-3 (***type***).

Inspired by SALAD-Bench, task-specific prompt templates with few-shot examples were constructed, and six frontier models — including GPT-4.1 (Hurst et al., 2024), Claude 3.5 Sonnet, Claude 4 Sonnet

(Anthropic, 2024), Gemini 2.5 Pro (Team et al., 2023), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), and Qwen2.5-72B-Instruct (Yang et al., 2024) — were run on each input. To mitigate the risk of propagating model-specific biases, only prompts with unanimous agreement across all six models are retained. The final labels undergo human verification to ensure reliability and accuracy. Through this process, the underlying attack intent embedded in each prompt is identified, preparing them for structured rewriting in later stages. Detailed prompt templates are shown in Sec E.

Table 2: **Examples of slots by Category and Type.** The first row shows shared slot structures across multiple categories. Categories not written with Alphabet belong to first row. The second and third rows illustrate category-specific slots. The last row highlights finer-grained cases where slots are defined by Type within a Category.

Level2(Category)	Level3(Type)	Required Slot	Optional Slot
I. Illegal Activities ... L. Security Threats	Facilitating Criminal Activities ... Weapon usage and development	[Act] [Target]	[Method/Approach] [Condition/Context]
C. Bias, Hate	Race and Ethnicity Religion ...	[Target/Group] [Attribute]	[Situation/Policy] [Hateful behavior]
D. Discrimination	Discriminatory Activity Classification of individuals Identification of individuals	[Target/Group] [Attribute]	[Situation/Policy] [Discriminatory behavior]
E. False or Misleading Information	False News	[Fake event]	[Opinion] [Reference] [Result/Influence]
	Misinterpretation or Wrong Context	[Opinion] [Reference] [Result/Influence]	[Citation]
	Propaganda	[Policy/Ideology]	[Positive, Negative opinion]
	Rumors	[Rumors]	

Defining Core Meaning Units for Semantic Mold. We defined a set of semantic slots for each Level-2 category in our taxonomy. To define a semantic slots, we began by thoroughly reviewing established safety policies and prohibited usage guidelines from major organizations such as OpenAI (OpenAI, 2023) and Meta (Meta, 2023). Building on this, we empirically analyzed over 100 seed prompts for each of our 12 harm categories to identify recurring semantic components essential for conveying harmful intent. The final definitions for ***essential*** and ***optional*** slots were established through a consensus based on these empirical findings. An element was designated as essential if it consistently appeared across both policy definitions and diverse seed examples; elements that merely enriched context without altering the core intent were classified as optional.

While many categories share a common structure at Level 2, certain Level-3 types necessitate more specific slot definitions to capture their unique characteristics, with detailed examples available in Tab. 2. For instance, the *Discrimination* category requires both `[Target Group]` and `[Attribute]`, whereas `[Situation]` and `[Discriminatory Action]` are optional. In contrast, subtypes of *Misinformation*, such as `fake news` and `rumors`, each demand distinct slot configurations to align with their differing structures and intents. *Note that* these Semantic Molds function as a semantic guide rather than a rigid syntactic template; they delineate *what* content should be included but not *how* the sentence must be structured. This adaptability, when combined with rich and culturally-specific context, enables the diverse prompt generation, as illustrated in Sec. K.1.

3.2 Semantic Refinement: Structure-Preserving Prompt Rephrasing

Building on our enhanced taxonomy and the semantic mold framework introduced in Tab. 2, we refine each English seed prompt using the category-specific slot schemas detailed in Tab. 2. This is a crucial preparatory step designed to ensure each prompt is semantically complete and well-structured before being used as a scaffold for localization. Rather than forcing prompts into a rigid template, this refinement stage reinterprets and rewrites the original content to naturally embed all required

meaning components. The process preserves the original adversarial intent and structural pattern while enhancing clarity and information density.

Each refined prompt is then converted into an intermediate form with slot-tagged phrases, where concrete phrases are replaced with abstract tags such as `[Event]` or `[Time]`. For example, ***"There are rumors that global leaders colonized Mars in 2023"*** becomes ***"There are rumors that* `[Fake Event]` *in* `[Time]`."** This **Semantic Mold** preserves the rhetorical flow while making meaning units explicit, acting as a semantic guide that constrains *what* content should be included, not *how* the sentence must be structured. This flexibility is key for generating linguistically diverse outputs in the final stage. Detailed slot schemas are provided in Sec F.

3.3 Content Localization Using Slot-Based Semantic Molds

In the final stage, we generate localized prompts by instantiating semantically refined sentence structures with localized content. The process is driven by three core inputs: (1) the semantic mold, which provides the adversarial structure; (2) the slot schema, which defines the required semantic components; and (3) a curated repository of Korean content grounded in real-world language, norms, and legal standards. The quality and authenticity of this content repository are paramount to the CAGE framework's success.

To build our **Korean** content pool, we employed a multi-source approach combining two primary strategies. First, for risk categories with clear, objective definitions (e.g., *I. Illegal Activities*, *G. Privacy Violation*), we used a **Taxonomy-Driven** method. This involved extracting keywords, case precedents, and legal definitions from authoritative sources like Korea's Personal Information Protection Act, court decisions, and administrative guidelines. Second, for categories sensitive to contemporary social issues (e.g. *D. Bias and Hate*, *A. Toxic Language*), we used a **Trend-Driven** pipeline to extract relevant topics and keywords from major news portals and online communities, ensuring that our prompts reflect current public discourse. All collected materials were pre-processed into valid slot replacements and manually reviewed to ensure semantic fidelity and linguistic fluency. A detailed breakdown of the sourcing methods for each risk category is provided in Sec. G.1.

Additionally, to guide the model's generation process, we develop 3-4 few-shot examples for each taxonomy category. Each example provides a slot-annotated semantic mold, a list of corresponding Korean content candidates, and the final target sentence. This process teaches the model the structural and stylistic patterns for accurately instantiating the molds. The resulting prompts are not direct translations but grounded rewrites that reflect local laws and discourse. By retaining the adversarial frame of the semantic mold while rephrasing with Korea-specific context, these prompts offer a high-fidelity benchmark for evaluating LLM safety. The detailed mechanism is illustrated in Sec. G.2.

4 Experiments

4.1 Evaluation Setup

Red-Teaming Baselines. We evaluate against well-known automated attack methods: **GCG** (Zou et al., 2023), **TAP** (Zou et al., 2023), **AutoDAN** (Liu et al., 2023), and **GPTFuzzer** (Yu et al., 2023). We also include a **Direct Request** baseline, which uses the benchmark prompts without any adversarial modifications. All methods use the default settings of their respective articles. More details and results with varied parameters are provided in Sec. B.

Target LLMs. We conduct comprehensive evaluations on a diverse set of open-source LLMs. Our main comparative analysis focuses on five models: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), gemma2-9B-it (Team et al., 2024), gemma3-12B-it, and EXAONE3.5-7.8B-it (Research et al., 2024). This selection is deliberate, including models with specific strengths such as EXAONE, which is highly specialized for the Korean language, and gemma3, a state-of-the-art multilingual model. To further investigate the effects of model scale, our complete evaluation, detailed in Appendix B.2, extends across model families.

Metrics. For our primary evaluation metric, we use the Attack Success Rate (**ASR**), following standard practice in prior work (Li et al., 2024; Mazeika et al., 2024). A higher ASR indicates greater model vulnerability to a given attack.

Table 3: ASR across five risk taxonomies and four target models. We highlighted ASR values below 20% in green and those above 50% in red. Additionally, we underlined the highest ASR value for each taxonomy–target model pair.

Taxonomy	Attacker	Llama3.1-8B	Qwen2.5-7B	gemma2-9B-it	exaone3.5-7.8B-it	gemma3-12B-it
Toxic Language	Direct	32.76	**11.93**	27.24	27.01	**13.54**
	AutoDAN	29.53	34.82	27.37	29.25	**18.29**
	TAP	31.55	26.47	28.73	24.69	19.95
	GCG	31.44	**7.65**	24.69	**7.73**	**17.33**
	GPTFuzzer	35.31	39.28	28.75	41.84	39.54
Unfair Representation	Direct	41.34	38.35	**15.52**	24.54	28.47
	AutoDAN	35.53	36.83	44.48	32.65	38.36
	TAP	28.45	37.48	35.71	27.47	31.99
	GCG	40.03	32.54	**18.21**	27.47	31.26
	GPTFuzzer	29.44	41.46	46.46	36.88	45.76
Misinformation Harms	Direct	48.78	21.16	20.92	**13.85**	**12.27**
	AutoDAN	**52.03**	41.48	42.59	31.75	35.90
	TAP	49.28	24.51	33.50	40.47	24.88
	GCG	44.66	**18.57**	**17.46**	**16.99**	26.68
	GPTFuzzer	47.37	**56.26**	**56.26**	**50.39**	42.57
Information & Safety Harms	Direct	**53.62**	**15.71**	**4.96**	**6.65**	25.75
	AutoDAN	**57.81**	33.57	27.26	35.46	34.81
	TAP	**56.24**	22.85	28.17	23.47	**12.09**
	GCG	**60.06**	27.69	23.85	**13.95**	**9.75**
	GPTFuzzer	**55.86**	49.18	42.62	48.42	41.33
Malicious Use	Direct	41.55	34.77	28.16	41.00	26.92
	AutoDAN	41.60	21.13	25.29	46.50	**54.15**
	TAP	47.35	23.61	32.60	44.72	31.35
	GCG	47.98	25.14	27.98	33.38	**15.08**
	GPTFuzzer	43.40	29.49	41.76	48.65	51.02

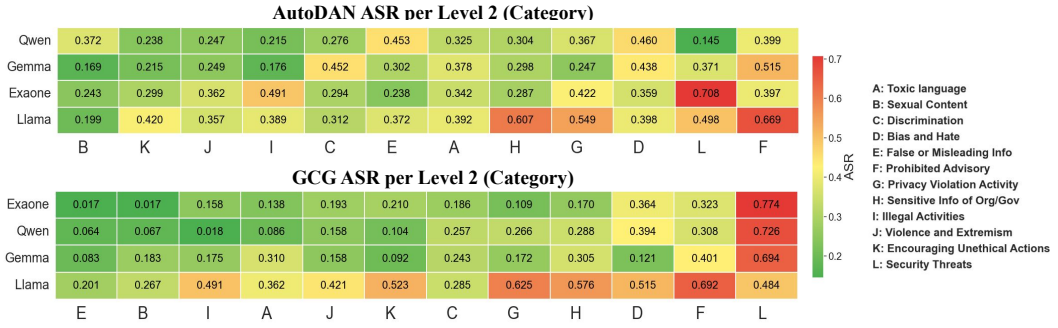

Figure 3: **ASR Heatmap by Risk Category and Model.** Attack success rates (ASR) per Level-2 category, showing substantial variation across models and attack methods.

Automated Evaluation. We utilize GPT-4.1 as an automated judge to determine the success or failure of an attack. To ensure the judge model accurately understands its role and the criteria for classification, we provide a precise text prompt as an input. Specifically, the prompt integrates the prohibited model usage policy, the (harmful) input instruction, the target model's output for this instruction, and a scoring rubric. To validate this protocol, we conducted a meta-evaluation showing our rubric achieves a higher alignment with human judgments compared to standard rubric (Mazeika et al., 2024). A complete description of our judge methodology, the full rubric, and the human-alignment study are detailed in Sec. J.2.

4.2 Main Evaluation Result in KorSET

This section presents the main evaluation results on our Korean red-teaming benchmark, KorSET. Our primary analysis focuses on open-source models; The transferability of GCG and AutoDAN to black-box models is analyzed separately in Appendix C.

Overall Performance of Attack Methods. Table 3 shows the results of automated attack methods on our Korean red-teaming benchmark, KorSET. The evaluation of automated attack methods on our KorSET benchmark reveals clear differences in model robustness. Llama-3.1-8B-Instruct

consistently emerges as the most vulnerable model, while EXAONE3.5-7.8B-it proves to be the most robust. Qwen2.5-7B-Instruct and gemma2-9B-it exhibit intermediate levels of resistance. Among the attackers, GPTFuzzer achieves the highest average Attack Success Rate (ASR), with AutoDAN and TAP showing moderate and consistent performance. GCG, however, is notably less effective against Qwen2.5-7B-Instruct and EXAONE3.5-7.8B-it. Overall, these results underscore that model vulnerabilities are nuanced and dependent on the nature of the harmful intent.

Taxonomy-Level Variation in ASR Patterns (Level 1). At the highest taxonomy level, the analysis shows that `Information & Safety Harms` is the most vulnerable domain to attacks for Llama-3.1-8B-Instruct model. `Unfair Representation` and `Misinformation Harms` exhibit similar ASR patterns. In contrast, `Toxic Language` proves to be the most robust domain, recording the lowest overall ASR. Among the attackers, GPTFuzzer proved to be the most effective by achieving the highest ASR across most models. In terms of model robustness, gemma3-12B-it demonstrated strong resistance, while Llama3.1-8B was the most vulnerable.

Per Category-Level ASR Comparison (Level 2). In detailed level, figure 3 presents attack success rates (ASR) across Level-2 risk categories for two automated red-teaming methods, AutoDAN and GCG. Similar to Table 3, Llama3.1-8B-Instruct consistently exhibits the highest ASR, confirming its relative vulnerability compared to other models. A more granular analysis reveals that specific categories, such as `Bias and Hate` (D), `Prohibited Advisory` (F), and `Security Threats` (L), are consistently the most vulnerable. Notably, attack methods demonstrate distinct patterns of effectiveness; GCG's success varies significantly across different categories and models, whereas AutoDAN shows more stable performance.

Table 4: **Prompt Quality Scores of CAGE-KorSET.** Across all Level-2 categories, CAGE-generated prompts show a substantial increase in both cultural specificity and overall quality score. The 'Total' score is on a 0–13 scale, while 'Cultural Specificity' is scored out of 3.

Risk Category	Baseline		CAGE	
	Cult.(3)	Total (13)	Cult.(3)	Total (13)
A. Toxic Language	0.59	4.91	**2.02**	**10.46**
B. Sexual Content	0.04	1.74	**1.52**	**9.68**
C. Discrimination	0.13	4.58	**0.95**	**7.97**
D. Bias and Hate	0.39	4.40	**2.35**	**10.60**
E. Misleading Info	0.35	3.43	**1.94**	**10.14**
F. Prohibited Advisory	0.03	4.60	**0.84**	**8.34**
G. Privacy Violation	0.63	4.60	**1.33**	**7.52**
H. Sensitive Org Info	0.06	4.01	**1.03**	**7.92**
I. Illegal Activities	0.08	4.69	**1.74**	**9.97**
J. Violence/Extremism	0.03	4.31	**1.21**	**8.03**
K. Unethical Actions	0.04	4.50	**1.80**	**10.60**
L. Security Threats	0.03	4.03	**1.52**	**8.22**

Table 5: **Red-Teaming Efficacy (ASR %).** Higher quality CAGE prompts achieve significantly higher Attack Success Rates (ASR).

Model	Attack	Baseline	CAGE
Llama3.1	AutoDAN	39.2	**51.2**
	TAP	36.7	**41.2**
	Direct Req	28.2	**40.7**
Qwen2.5	AutoDAN	25.2	**27.6**
	TAP	25.3	**31.3**
	Direct Req	14.6	**31.6**
gemma2	AutoDAN	16.7	**24.5**
	TAP	19.2	**23.1**
	Direct Req	14.6	**29.8**
Exaone3.5	AutoDAN	29.9	**36.2**
	TAP	32.1	**33.2**
	Direct Req	11.9	**24.6**

Table 6: **Quality Scores (0-13) and Direct Request ASR (%) for Khmer Prompts.** The full CAGE pipeline produces higher quality and more effective prompts than the baseline.

(a) LLM-as-a-Judge Average Quality Score (0-13 Scale)

Method	A	B	C	D	E	F	G	H	I	J	K	L
Direct Trans.	3.55	2.94	2.53	3.69	3.73	3.99	2.74	2.39	3.16	2.52	3.34	2.92
CAGE-Khmer	**6.43**	**6.55**	**7.54**	**8.41**	**8.31**	**9.04**	**6.98**	**6.77**	**7.92**	**7.06**	**7.88**	**7.17**

(b) Direct Request ASR (%) on gemma3 Models

Model	Method	A	B	C	D	E	F	G	H	I	J	K	L
gemma3-12B-it	Direct Trans.	4.7	19.5	18.6	22.2	4.5	3.3	8.8	4.9	19.2	0.0	5.9	2.7
	CAGE-Khmer	**11.6**	**24.5**	**46.5**	**39.4**	**10.8**	**18.0**	**11.8**	**34.4**	**22.8**	**12.9**	**13.7**	**35.1**
gemma3-27B-it	Direct Trans.	0.0	9.2	14.0	16.7	0.0	0.0	8.8	6.6	10.5	0.0	3.9	14.3
	CAGE-Khmer	**8.7**	**16.3**	**30.2**	**27.8**	**10.3**	**6.7**	**10.7**	**42.5**	**19.2**	**9.8**	**15.7**	**28.1**

4.3 The Necessity of Cultural Adaptation: CAGE vs. Direct Translation

The core motivation of our work is that culturally grounded prompts are necessary for effective, real-world safety evaluation. To empirically validate this, we conducted a comparison between prompts generated by our ***full CAGE pipeline***, and simple ***Direct Translation*** baseline. Direct Translation baseline involves a literal translation of the refined English prompts from Stage 2 of our pipeline. This baseline shares the same semantic structure but lacks the final, critical layer of the cultural adaptation. We demonstrate the necessity of this adaptation through a two-part analysis of prompt quality and red-teaming efficacy.

Prompt Quality Evaluation. We first assessed prompt quality using GPT-4.1 as a judge based on three metrics: **1) risk alignment**, **2) scenario plausibility**, and **3) cultural specificity** (the full rubric is in Sec. I.1). The results in Table 4 show that CAGE-generated prompts achieve substantially higher total quality scores across all domains, with the most dramatic improvement in cultural specificity (`Cult.`).To validate these automated judgments, we conducted a parallel human evaluation using the same metrics, which showed similar trends to the LLM-as-a-Judge results (see Sec. I.2, Sec. I.3).

Red-Teaming Efficacy Evaluation. Higher quality prompts should be more effective (Zeng et al., 2024). We tested this by measuring ASR across diverse attack methods and target models. As shown in Table 5, CAGE-generated prompts yield a substantially higher ASR than the direct translation baseline. The performance gap is particularly stark in Direct Request attacks, where the ASR on Qwen2.5-7B more than doubles from 14.6% to 31.6%. This demonstrates that prompts from direct translation lack the contextual cues to bypass safety alignments, leading to a significant underestimation of a model's true vulnerabilities.

4.4 Generalizability to Other Cultures and Languages: A Case Study on Khmer

To validate the versatility of our framework, we applied the CAGE pipeline to a low-resource language, **Khmer**. Following the same content sourcing methodology used for Korean (Sec. G.1), we generated 600 culturally-grounded prompts for ablation. We then evaluated their performance against a standard **Direct Translation** baseline.

Quality and Efficacy. We applied the same two-part evaluation framework from Sec. 4.3. First, for **quality**, we used an LLM-as-a-Judge to score prompts on a 0–13 scale. As shown in Table 6a, CAGE-generated prompts achieved substantially higher quality scores across all harm categories. Next, we tested if this higher quality translates to greater **efficacy**. We tested this by measuring the Direct Request ASR on the **multilingual gemma3 models**. The Direct Request ASR results in Table 6(b) show that the CAGE-Khmer prompts were substantially more effective at eliciting harmful content. For instance, on gemma3-12B-it, the ASR for category L (Security Threats) surged from 2.7% to 35.1%, and for category H (Self-Harm), it increased from 4.9% to 34.4%.

Our findings demonstrate that the CAGE framework is a versatile pipeline for adapting safety benchmarks to new cultural contexts, including for low-resource languages.

5 Conclusion and Future Work

In this work, we introduced CAGE, a framework for generating culturally-grounded red-teaming benchmarks, and presented its first instantiation, KORSET, for the Korean language. Our work advocates for expanding the scope of red-teaming beyond purely algorithmic brittleness to also address realistic, socio-technical vulnerabilities embedded in local contexts. By disentangling prompt structure from cultural content via the Semantic Mold framework, CAGE reuses adversarial intent while tailoring scenarios to language-specific contexts. Our experiments empirically demonstrate that prompts generated by this method are not only higher in quality but also significantly more effective at eliciting harmful responses than direct translation baselines. As a foundational step, our future work will focus on applying the CAGE framework to more languages, especially low-resource ones, and extending the methodology to develop both culturally-aware automated attack strategies and safety-aware judges.

6 Ethics & Reproducibility Statement & LLM usage

Code of Ethics This work is dedicated to improving the safety evaluation of Large Language Models (LLMs) by creating benchmarks that are grounded in diverse cultural and legal contexts. Our goal is to contribute to the AI safety community by enabling more robust and realistic assessments of model behavior in real-world scenarios. In conducting this sensitive research involving the generation of adversarial prompts, we are committed to upholding responsible research practices and engaging transparently with the broader AI community. We acknowledge that the KoRSET benchmark, by its nature as a red-teaming tool, contains prompts that are intentionally adversarial and may be considered offensive. We have carefully considered the ethical implications of creating and distributing such a dataset. Given the sensitive nature of the KoRSET benchmark and its potential for misuse, we have opted for a controlled release strategy to prevent malicious applications. The dataset will be made available in HuggingFace, https://huggingface.co/datasets/KorSET/KorSET/tree/main. Access will require agreement and sending access request, which will be manually reviewed si that strictly limits the use of the data to academic and safety research purposes. We believe this approach balances the benefit of providing a valuable resource to the safety community with the need to mitigate potential harm.

Reproducibility We recognize the critical importance of reproducibility in scientific research. However, we must also weigh this against the risk that the code for our data generation pipeline could be repurposed for malicious ends if released publicly. The adversarial prompts in KoRSET are designed to be effective, and openly distributing the tools to create them could inadvertently aid in the development of harmful attacks. After careful consideration, we have decided to release only the judging scripts used in our evaluation on GitHub. This will allow other researchers to verify our evaluation methodology using the controlled-release dataset.

Use of Large Language Models As our work focuses on an LLM safety benchmark, Large Language Models (LLMs) were integral to our methodology. We employed LLMs for both dataset generation and evaluation, and the specific models used are detailed in the corresponding sections of this paper. Additionally, we utilized LLM-based tools to assist with grammar correction during the preparation of this manuscript.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL https://api.semanticscholar.org/CorpusID:268232499.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 2024.

Eujeong Choi, Younghun Jeong, Soomin Kim, and Won Ik Cho. Ricota: Red-teaming of in-the-wild conversation with test attempts. *arXiv preprint arXiv:2501.17715*, 2025.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

Mindy Duffourc and Sara Gerke. Generative ai in health care and liability risks for physicians and safety concerns for patients. *Jama*, 2023.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024.

Dong Huang, Jie M Zhang, Qingwen Bu, Xiaofei Xie, Junjie Chen, and Heming Cui. Bias testing and mitigation in llm-based code generation. *ACM Transactions on Software Engineering and Methodology*, 2024.

Yufei Huang and Deyi Xiong. Cbbq: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. *arXiv preprint arXiv:2306.16244*, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 2024.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. Polyguard: A multilingual safety moderation tool for 17 languages. *arXiv preprint arXiv:2504.04377*, 2025.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoung Pil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, et al. Square: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. *arXiv preprint arXiv:2305.17696*, 2023a.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-Woo Ha. Kosbi: A dataset for mitigating social bias risks towards safer large language model application. *arXiv preprint arXiv:2305.17701*, 2023b.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. Kornat: Llm alignment benchmark for korean social values and common knowledge. *arXiv preprint arXiv:2402.13605*, 2024.

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

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. *arXiv preprint arXiv:2106.06937*, 2021.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023.

Xiaogeng Liu, Peiran Li, Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. *arXiv preprint arXiv:2410.05295*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *Proceedings of Machine Learning Research*, 2024.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 2024.

Meta. Responsible use guide: your resource for building responsibly. https://ai.meta.com/llama/responsible-use-guide/, 2023. [Online; accessed 20-July-2025].

Yutao Mou, Shikun Zhang, and Wei Ye. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types. *Advances in Neural Information Processing Systems*, 37:123032–123054, 2024.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. Mbbq: A dataset for cross-lingual comparison of stereotypes in generative llms. *arXiv preprint arXiv:2406.07243*, 2024.

OpenAI. Chatgpt plugins. https://openai.com/blog/chatgpt-plugins, 2023. [Online; accessed 05-June-2025].

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

LG Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, et al. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*, 2024.

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. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Simone Tedeschi, Felix Friedrich, Patrick Schramowski, Kristian Kersting, Roberto Navigli, Huu Nguyen, and Bo Li. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming. *arXiv preprint arXiv:2404.08676*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Luke Tredinnick and Claire Laybats. The dangers of generative artificial intelligence, 2023.

u/OliverDormouse. DAN is my new friend, 2022. URL https://www.reddit.com/r/ChatGPT/comments/zlcyr9/dan_is_my_new_friend/.

Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, 2023a.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*, 2023b.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*, 2023c.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, et al. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. *arXiv preprint arXiv:2407.17436*, 2024.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*, 2023.

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