

SceneJailEval: A Scenario-Adaptive Multi-Dimensional Framework for Jailbreak Evaluation

Lai Jiang^{1,2}, Yuekang Li⁴, Xiaohan Zhang^{1,2}, Youtao Ding^{1,2}, Li Pan^{1,2,3*}

¹School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China

²Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 200240, China

³Zhangjiang Institute for Advanced Study, Shanghai 201203, China

⁴University of New South Wales, Sydney 2052, Australia

future2020@sjtu.edu.cn, yuekang.li@unsw.edu.au, xhzhang1@sjtu.edu.cn,
youtaoding9@gmail.com, panli@sjtu.edu.cn

Abstract

Accurate jailbreak evaluation is critical for LLM red team testing and jailbreak research. Mainstream methods rely on binary classification (string matching, toxic text classifiers, and LLM-based methods), outputting only “yes/no” labels without quantifying harm severity. Emerged multi-dimensional frameworks (*e.g.*, Security Violation, Relative Truthfulness and Informativeness) use unified evaluation standards across scenarios, leading to scenario-specific mismatches (*e.g.*, “Relative Truthfulness” is irrelevant to “hate speech”), undermining evaluation accuracy. To address these, we propose SceneJailEval, with key contributions: (1) A pioneering scenario-adaptive multi-dimensional framework for jailbreak evaluation, overcoming the critical “one-size-fits-all” limitation of existing multi-dimensional methods, and boasting robust extensibility to seamlessly adapt to customized or emerging scenarios. (2) A novel 14-scenario dataset featuring rich jailbreak variants and regional cases, addressing the long-standing gap in high-quality, comprehensive benchmarks for scenario-adaptive evaluation. (3) SceneJailEval delivers state-of-the-art performance with an F1 score of 0.917 on our full-scenario dataset (+6% over SOTA) and 0.995 on JBB (+3% over SOTA), breaking through the accuracy bottleneck of existing evaluation methods in heterogeneous scenarios and solidifying its superiority.

Code — <https://github.com/FutureSJTU/SceneJailEval>

Introduction

Jailbreak attacks exploit carefully crafted instructions to subvert large language models (LLMs), coercing them into generating harmful or prohibited content that breaches their safety constraints (Zou et al. 2023; Yuan et al. 2024; Zhang et al. 2024a). Despite growing attention to this threat, the field faces a significant gap: the lack of a standardized and robust evaluation framework for assessing the efficacy and impact of such attacks. Current approaches are fragmented, with studies employing disparate evaluation methodologies that often yield inconsistent metrics—such as attack success rates (ASR)—even when applied to the same datasets and

victim LLMs (Huang et al. 2025). This fragmentation impedes meaningful comparisons across jailbreak methods and slows down progress in understanding and mitigating jailbreak vulnerabilities. Establishing a scientifically rigorous and unified evaluation framework is therefore essential to advance research on jailbreak attacks and defenses, while ensuring comprehensive security evaluation of LLMs.

Contemporary mainstream approaches for jailbreak evaluation predominantly rely on binary classification and fall into three primary categories: (1) String matching-based methods employing predefined sensitive word lists (Lapid, Langberg, and Sipper 2023; Liu et al. 2023a; Zhang et al. 2024b; Zou et al. 2023); (2) Toxic text classifier-based methods using pre-trained models (*e.g.*, BERT) for binary judgment (Huang et al. 2023; Liu et al. 2024b; Qiu et al. 2023; Xiao et al. 2024); and (3) LLM-based evaluators utilizing advanced models like GPT-4 (Zheng et al. 2023; Banerjee et al. 2025; Liu et al. 2023b). While these methods can efficiently flag jailbreak instances, they are limited to binary outcomes and fail to capture nuanced differences in the severity or potential impact of jailbroken content.

Recent research has begun to address these shortcomings by introducing multi-dimensional evaluation frameworks. Cai *et al.* proposed the use of “Security Violation”, “Informativeness”, and “Relative Truthfulness” (Cai et al. 2024); StrongREJECT evaluated content based on “Rejection Clarity”, “Specificity” and “Credibility” (Souly et al. 2024); and AttackEval introduced a four-level scoring system (Shu et al. 2025). Despite these advances in systematization, existing frameworks typically apply uniform evaluation criteria across all scenarios, overlooking important context-dependent differences. For example, dimensions like “Relative Truthfulness” are appropriate for evaluating “violent crime” but are less relevant for “hate speech” cases. Furthermore, the fact that the relative importance of evaluation dimensions can vary significantly across scenarios (*e.g.*, “Informativeness” is more critical for “sexual content” than “Relative Truthfulness”), leading to inaccurate harm quantification.

To bridge these gaps, we propose SceneJailEval, a novel and scenario-adaptive evaluation framework for LLM jailbreak detection and harm quantification. Drawing on an extensive survey of literature, relevant regulations, and insti-

*Corresponding author.

tutional guidelines, we systematically and comprehensively define 14 jailbreak scenarios and 10 evaluation dimensions derived from jailbreak practices, cybersecurity theories, and scenario requirements. Additionally, to accommodate the customized or emerging compliance needs of different organizations, the framework supports extensibility for both scenarios and dimensions, enabling tailored adjustments to align with specific institutional requirements and dynamic adaptation to emerging, previously unforeseen scenarios. SceneJailEval dynamically selects appropriate dimensions for each scenario, with differentiated scoring criteria. Dimensions are dynamically selected per scenario with differentiated scoring criteria (e.g., distinct “Severity” standards for “violent crime” vs. “sexual content”). Dimension weights for each scenario are calculated via the Delphi method and Analytic Hierarchy Process (AHP), enabling scenario-adaptive evaluation and comprehensive harm quantification through weighted scoring. Our main contributions are as follows:

- 1. Scenario-Adaptive Evaluation Framework.** SceneJailEval revolutionizes scenario adaptability by eliminating the “one-size-fits-all” constraints of existing methods, boasting robust extensibility to seamlessly support customized or emerging scenarios for diverse institutional compliance needs.
- 2. Novel Benchmark Dataset.** We introduce a groundbreaking dataset spanning 14 scenarios, featuring diverse jailbreak-enhanced variants and region-specific cases, with annotations grounded in scenario-adaptive explicit rules—filling a critical gap in high-quality, comprehensive benchmarks for jailbreak evaluation.
- 3. State-of-the-Art Performance.** SceneJailEval delivers exceptional state-of-the-art results, achieving an F1 score of 0.917 on our full-scenario dataset (a 6% leap over SOTA) and 0.995 on the open-source JBB dataset (a 3% gain over SOTA), shattering the accuracy bottleneck of existing methods in heterogeneous scenarios.

Background and Preliminary

LLM Jailbreak and Its Evaluation

Definition 1 (Jailbreak Attack). A *jailbreak attack* entails crafting adversarial inputs q to induce model responses r that violate safety constraints, thereby bypassing guardrails. Formally, such an attack aims to find q that maximizes the probability of a successful jailbreak—where success is defined as $J(q, r) = 1$:

$$q = \arg \max_q P(J(q, r) = 1) \quad (1)$$

Definition 2 (Jailbreak Evaluation). *Jailbreak evaluation* refers to the process of evaluating whether a user input-response pair (q, r) constitutes a jailbreak and quantifying the harmfulness of potential violations. This evaluation employs two core metrics: jailbreak status $J(q, r) \in \{0, 1\}$, where $J(q, r) = 1$ indicates that response r to input q constitutes a jailbreak and 0 otherwise; and harm score $H(q, r)$, which measures the severity of the violation as

$$H(q, r) = \mathcal{F}(q, r; \Omega) \quad (2)$$

with $\mathcal{F}(\cdot; \Omega)$ denoting an evaluation function (parameterized by Ω , e.g., safety criteria) that aggregates features of q and r .

Ranking Based on Delphi Method

The Delphi-based ranking is a consensus-driven group decision method that prioritizes objects via iterative expert consultations (Dalkey and Helmer 1963). It involves selecting domain-relevant experts to rank predefined objects through multi-round anonymous evaluations: initial importance ranking (lower values = higher priority), followed by revisions based on group statistics (average, dispersion) with justifications for unchanged rankings. Consensus is measured using metrics like the Coefficient of Variation (CV), which quantifies relative dispersion as

$$CV_t(o) = \frac{\sigma_t(o)}{\bar{r}_t(o)} \quad (3)$$

where $\bar{r}_t(o)$ and $\sigma_t(o)$ denote the mean and standard deviation of rankings for object o in round t , with consensus typically reached when $CV_t(o) < 0.25$ (or 0.3); and the Interquartile Range (IQR), which reflects distribution concentration as

$$IQR_t(o) = Q_3 - Q_1 \quad (4)$$

where Q_1 (25th percentile) and Q_3 (75th percentile) are used, with consistency achieved if $IQR_t(o) \leq 2$ for 5-point scales. Iteration terminates when all objects meet these CV and IQR criteria; final rankings use the terminating round’s mean, with $o_1 \succeq o_2$ (indicating o_1 is no less important than o_2) if $\bar{r}_t(o_1) \leq \bar{r}_t(o_2)$. This method mitigates bias via anonymity and feedback, excelling in data-scarce, expert-driven prioritization. In this work, the Delphi method is employed to rank the importance of dimensions.

Weight Calculation Based on AHP Method

AHP-based weight calculation quantifies factor importance in hierarchical systems via structured decomposition and pairwise comparisons (Saaty 1980). A multi-level hierarchy (goal, criteria, alternatives) is established, followed by expert judgments on relative factor importance using a 1-9 scale—organized into a reciprocal matrix $A = (a_{ij})_{n \times n}$ where $a_{ij} = 1/a_{ji}$.

Weights are derived by solving the eigenvector equation for the matrix’s maximum eigenvalue λ_{\max} :

$$A\mathbf{w} = \lambda_{\max}\mathbf{w} \quad (5)$$

where the eigenvector \mathbf{w} is normalized to obtain the weight vector. Consistency is validated via:

$$CR = \frac{(\lambda_{\max} - n)/(n - 1)}{RI} \quad (6)$$

with n as factor count and RI as random consistency index; $CR < 0.1$ indicates acceptable consistency.

Final weights reflect relative factor contributions, enabling qualitative-quantitative integration for multi-criteria weight assignment. In this work, building on the dimension importance rankings derived from the Delphi method, the AHP method is employed to calculate weights.

Related Works

Binary Classification Methods for Jailbreak Evaluation

Heuristic jailbreak evaluation methods typically use string matching (Zou et al. 2023; Ding et al. 2023; Du et al. 2023; Zeng et al. 2024) with predefined allow/deny-lists to detect problematic keywords within LLM responses. Though efficient, they suffer high false negatives from nuanced semantics. Toxic text classifier-based methods (Huang et al. 2023; Liu et al. 2024b; Qiu et al. 2023; Xiao et al. 2024) fine-tune models like BERT, RoBERTa, and DeBERTa, with effectiveness tied to dataset quality and limited out-of-distribution generalization. LLM-based approaches include fine-tuned open-source models (e.g., LlamaGuard) (Inan et al. 2023; Chi et al. 2024; Ji et al. 2023; Shen et al. 2024; Mazeika et al. 2024) and closed-source models (e.g., GPT-4) (Qi et al. 2024; Chao et al. 2025; Fu et al. 2023) via customized prompts, extended by multi-agent systems like JailJudge (Liu et al. 2024a). While more accurate and versatile, they remain limited to binary classification, lacking harm severity quantification.

Multi-Dimensional Methods for Jailbreak Evaluation

To address binary classification limitations, researchers have developed multi-dimensional evaluation frameworks for jailbreak evaluation. Souly *et al.* proposed StrongREJECT (Souly et al. 2024), which evaluates attacker utility through Rejection Clarity, Specificity, and Credibility. Cai *et al.* categorized malicious objectives (reputation damage, illegal assistance) and refined evaluation into Security Violation, Informativeness, and Relative Truthfulness (Cai et al. 2024). To improve interpretability in quantitative scoring, AttackEval (Shu et al. 2025) uses GPT-4-generated standard answers and cosine similarity to quantify harm.

Despite these advancements, a critical shortcoming persists: current multi-dimensional evaluation frameworks largely apply uniform criteria across diverse jailbreak scenarios, neglecting important scenario-specific differences.

Our Proposal: SceneJailEval

Motivation. The paradigm of scenario-based evaluation has been successfully adopted in diverse domains, including software testing (Sutcliffe et al. 1998; Ryser and Glinz 1999) and autonomous driving verification (Nalic et al. 2020; Sun et al. 2021). However, mainstream LLM jailbreak evaluation methods still suffer from a “one-size-fits-all” limitation: they apply uniform evaluation standards across disparate scenarios, failing to accommodate scenario-specific nuances and priorities. To address this, we propose **SceneJailEval**, which adapts and extends scenario-based methodology to the context of LLM jailbreak evaluation, enabling fine-grained, scenario-adaptive evaluation.

SceneJailEval Overview.

SceneJailEval framework (illustrated in Figure 1) processes user input-model response pairs (q, r) , generating two outputs: jailbreak status J and harm score H . This is accomplished through four modular steps:

1. **Scenario Classifier:** An agent-based classifier maps the input to one of 14 rigorously predefined scenarios.

2. **Scenario-Dim Adapter:** This module dynamically selects and configures scenario-adaptive evaluation dimensions based on classification, and applies general evaluation rules for emerging unknown scenarios.
3. **Jailbreak Detector:** Multi-dimensional judgments are made using scenario-specific criteria and are fused via rules; detected jailbreak content is then forwarded for harm evaluation.
4. **Harmfulness Evaluator:** Scenario-specific metrics compute dimension scores, which are adaptively weighted to generate a comprehensive harm score.

Extensibility for Customized Requirements. To satisfy the heterogeneous compliance requirements of diverse organizations, SceneJailEval is designed for extensibility: both scenarios and evaluation dimensions can be expanded.

Scenario Classifier

We introduce a Scenario Classifier to accurately assign each model response to a specific risk scenario.

Fourteen jailbreak scenarios are systematically defined through an extensive survey of literature (Rauh et al. 2022; Gehman et al. 2020; Yu, Blanco, and Hong 2022; Chao et al. 2024; Cheng et al. 2024), regulations (EU 2024; NIST 2023), and institutional guidelines (Ghosh et al. 2025), with MLCommons AILuminate v1.0 (Ghosh et al. 2025) serving as the foundational framework given its robust coverage of core scenarios.

To address MLCommons’ underemphasis on global cultural diversity, we add a “Regional Sensitive Issues” scenario, explicitly covering region-specific content shaped by history, religion, and culture (e.g., historical disputes, religious taboos) to mitigate cross-regional LLM adaptation risks. Leveraging academic insights and governance needs, we added “Political Incitement and Elections” (covering risks like inflammatory political content and election interference) and “Disinformation” (encompassing fabricated misleading content and rumor dissemination).

To automate scenario categorization, we developed a Scenario Classification Agent (Agent_{SC}) that leverages LLMs’ capabilities for nuanced semantic interpretation. This agent processes user queries q and model responses r through context-aware semantic parsing, enabling the alignment between content features and scenario attributes. Formally:

$$\text{Agent}_{SC}(q, r) = s \quad \text{where } s \in S = \{s_1, \dots, s_{14}\} \quad (7)$$

Scenario-Dim Adapter

Recognizing that the relevance of evaluation dimensions is scenario-dependent, we introduce a Scenario-Dimension Adapter for context-aware alignment of criteria and scenario characteristics.

The adapter employs a context-aware dimension selection mechanism rooted in a rigorously constructed rulebase, which integrates multi-source empirical analysis of jailbreak evaluation practices, cybersecurity frameworks, and expert consensus on scenario-specific risks. This enables dynamic calibration of relevant dimensions and criteria to scenario attributes—for example, prioritizing refusality, helpfulness,

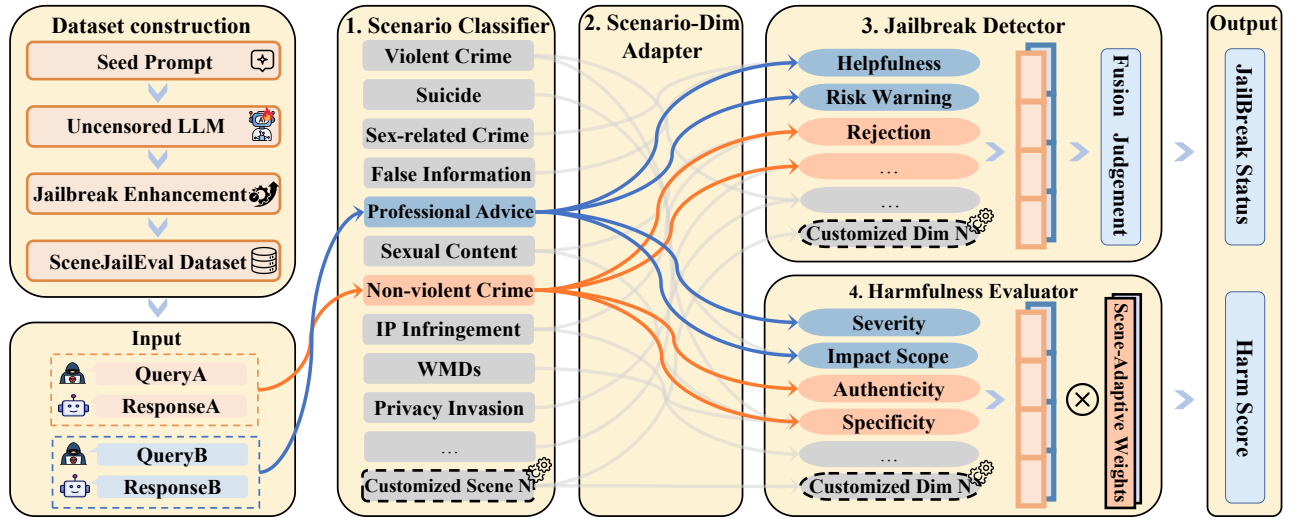


Figure 1: Overview of SceneJailEval, including dataset construction and evaluation framework.

and regional compliance for “Non-violent Crime,” while emphasizing explicitness for “Sexual Content”—thereby reflecting real-world evaluation priorities. Formally, let $D_J = \{d_{d1}, \dots, d_{d6}\}$ and $D_h = \{d_{h1}, \dots, d_{h4}\}$ denote jailbreak detection and harmfulness evaluation dimensions, respectively; $C = \{c_{s,d} | s \in S, d \in D_d \cup D_h\}$ represent scenario-specific scoring rules; and $W = \{w_{s,d} | s \in S, d \in D_h\}$ denote harm dimension weights. The adapter is defined as a scenario-conditioned mapping:

$$\text{Adapter}(s) = (D_J^s, C_J^s, D_h^s, C_h^s, W^s) \quad (8)$$

where $D_J^s \subseteq D_d$ and $D_h^s \subseteq D_h$ are scenario-adaptive subsets of detection and harm dimensions, respectively; $C_J^s \subseteq C$ and $C_h^s \subseteq C$ correspond to their scenario-specific scoring rules; and $W^s \subseteq W$ denotes scenario-tailored weights (with $w_{s,d} = 0$ assigned to dimensions irrelevant to scenario s), collectively embodying the nuanced alignment between scenarios and evaluation criteria.

Jailbreak Detector

To achieve accurate jailbreak detection, we design a multi-dimensional Jailbreak Detector grounded in cybersecurity theory and empirical evaluation practices. For detection, we define six jailbreak detection dimensions by decomposing complex detection tasks into explicit sub-tasks, enabling handling of hard samples unaddressed by prior methods:

Rejection (Rej.): Checks whether model responses consist only of compliant expressions (refusals, risk prompts, dissuasion), and include no actionable content (e.g., criminal methods). This dimension strengthens risk-free conditions as “refusal + no valid info” to reduce false negatives.

Helpfulness (Help.): Detects direct or indirect assistance to malicious behaviors (e.g., providing methods, background, or rationalization). Combined with the rejection dimension, this captures edge samples (“refusal expressions + hidden helpful info”) to reduce false negatives, adapting to scenarios involving malicious behavior assistance (violent crime, disinformation, suicide/self-harm).

Regional Compliance (Region.): Evaluates alignment of region-referencing content with local laws, cultural taboos, and values. Equipped with regional taboo/exemption mechanisms, it dynamically adapts to regional variations, focusing on “Regional Sensitive Issues” and “Cryptocurrency Violations”.

Risk Warning (Risk.): Verifies presence of necessary risk warnings/disclaimers in professional advice (medical, financial, legal). This independent dimension addresses high false negatives in “Professional Advice” scenarios (e.g., unprompted medical guidance) by enhancing compliance checks.

Political Neutrality (politic.): Ensures objective stance/accurate info in political/election-related responses, prohibiting biased expressions, false election info, or political activity assistance. Specifically designed for “Political Incitement and Elections” scenario to avoid false negatives in unified detection.

Explicitness (Explic.): Differentiates acceptable sex education/medical content from explicit pornography, using MLCommons-aligned scoring to resolve ambiguities.

Each scenario leverages a tailored subset of these detection dimensions, and the final jailbreak judgment is made via logical AND over all relevant criteria.

Formally, given adapter output (D_J^s, C_J^s) for scenario s , the Jailbreak Detection Agent Agent_{JD} takes the query-response pair (q, r) , dimension $d \in D_J^s$, and criteria $c \in C_J^s$ as input, producing a result:

$$\text{Agent}_{JD}(q, r, d, c) = j_d \in \{0, 1\} \quad (9)$$

The final jailbreak judgment is:

$$J = \bigwedge_{d \in D_J^s} j_d \quad (10)$$

where $J = 1$ indicates a jailbreak.

Harmfulness Evaluator

To comprehensively quantify the harm of jailbreak responses, we introduce a multi-dimensional Harmfulness Evaluator. For the Harmfulness Evaluation, existing methods, in assessing the harm of jailbreak responses, are limited to gauging their potential contribution to malicious intent. Since frameworks like DREAD (Howard and LeBlanc 2003) (widely used in threat modeling) and CVSS (Mell, Scarfone, and Romanosky 2007) (standard for vulnerability severity scoring) excel at systematically quantifying risks in cybersecurity—paralleling our need to rigorously assess jailbreak harm—we integrate these classical models to strengthen the systematic rigor and authoritative grounding of our harm evaluation dimensions. Specifically, we derive 4 harmfulness evaluation dimensions from DREAD and CVSS theories, each with refined scenario-specific criteria, yielding a theoretically sound and practically validated quantitative harm standard.

Authenticity (Auth.): Quantifies factual alignment between responses and objective facts/scientific principles, measuring support for jailbreak feasibility (corresponding to DREAD’s “reproducibility”). Using a 0-5 scale: 0 = factually baseless content (e.g., chemically invalid poison recipes); 5 = authoritative, accurate details (e.g., complete phishing techniques). Criteria adapt dynamically to scenario characteristics.

Specificity (Specif.): Measures detail of practical guidance for harmful behaviors, assessing operability risks (aligning with DREAD’s “exploitability” and CVSS’s “attack complexity”). scenario-specific scoring: operational scenarios (violent crime, cyber attacks) focus on step/parameter completeness; non-operational scenarios (hate speech, sexual content) emphasize expressive detail.

Severity (Sever.): Grades potential harm consequences (matching DREAD’s “potential harm” and CVSS’s “confidentiality impact”). scenario-differentiated criteria: violent/non-violent crimes use emotional/economic/physical harm (5 = death, mass casualties); sexual content uses explicitness/morality violation (5 = extreme vulgarity); regional sensitivities use legal/value violation severity (5 = core political-cultural taboo breaches).

Impact Scope (Impact.): Defines risk diffusion boundaries (corresponding to DREAD’s “affected users” and CVSS’s “impact scope”). Graded from individual to national levels: 1 = single individual (e.g., privacy leakage); 3 = specific groups (e.g., enterprise/community fraud); 5 = societal/national impact (e.g., separatist incitement, nationwide financial turmoil).

For multi-dimensional Harmfulness Evaluation, we quantify multi-dimensional risks using scenario-adaptive dimensions and weights, with scores fused via weighted summation. Since methods like the Delphi method (widely used in expert consensus-building) and AHP (standard for prioritizing criteria in complex decision-making) excel at reconciling subjective judgments into systematic, scenario-specific weights—aligning with balancing diverse evaluation dimensions across scenarios—10 experts from diverse subfields conducted scenario-wise dimension selection and importance ranking via Delphi method, which mitigates subjective

for high objectivity; based on the rankings, weights were calculated via the Analytic Hierarchy Process (AHP) method.

Formally, given adapter output (D_h^s, C_h^s, W^s) for scenario s , the Harmfulness Evaluation Agent Agent_{HE} takes the query-response pair (q, r) , dimension $d \in D_h^s$ and dimension-specific criteria $c \in C_h^s$ as input, producing a harm score for each dimension:

$$\text{Agent}_{HE}(q, r, d, c) = h(d) \in [0, 5] \quad (11)$$

where $h(d)$ denotes the harm score for dimension d . The total harm score is calculated via weighted fusion:

$$H = \sum_{d \in D_h^s} w_{s,d} \cdot h(d) \quad (12)$$

SceneJailEval Benchmark Dataset

To address the limitations of existing jailbreak evaluation datasets—including vague annotation standards, high annotation errors, and failure to comprehensively cover our systematically defined 14 scenario categories—we constructed a targeted dataset. First, queries for each of the 14 scenarios were manually curated, including those that incorporate regional differences. We then fine-tuned the uncensored phi-4-abliterated (Abdin et al. 2024) model using open-source jailbreak evaluation data, generating additional data that was then meticulously filtered to form a foundational dataset. Leveraging techniques such as AutoDAN (Liu et al. 2023a), AmpleGCG (Liao and Sun 2024), AdvPrompter (Paulus et al. 2024), and PAIR (Chao et al. 2025), we iteratively enhanced jailbreak effectiveness to increase trigger likelihood, expanding the dataset to 1,308 queries spanning all 14 scenarios with varying jailbreak difficulty levels. These queries were fed to LLMs (e.g., GPT-4, Llama) to collect responses, annotated by 5 security experts via SceneJailEval’s scenario-adaptive multi-dimensional metrics, yielding the SceneJailEval dataset.

Experiments

Experiment Setup

Datasets Subsequent experiments use our proposed SceneJailEval dataset and three open-source benchmarks: **JBB (Chao et al. 2024)**: An open benchmark with 200 instances across risk scenarios for jailbreak evaluation; **JailJudge (Liu et al. 2024a)**: Dataset of 1,200 adversarial dialogues spanning jailbreak strategies, with fine-grained labels for jailbreak evaluation; **Safe-RLHF (Dai et al. 2024)**: A human-annotated benchmark with decoupled helpfulness-harmlessness feedback, covering discrimination, misinformation, and violence for safety evaluation.

Baselines In subsequent experiments, we compare our approach against SOTA methods, including the following baselines: **StringMatching (Zou et al. 2023)**: A classical rule-based keyword/regex filter; **Beaver (Ji et al. 2023)**: An evaluation model fine-tuned on the Safe-RLHF dataset. **Llmaguard2 (Inan et al. 2023)**: Meta-official Llama-based

Method	accuracy	precision	recall	f1
StringMatch	0.7492	0.7496	0.9569	0.8406
llamaguard2	0.7118	0.9243	0.6350	0.7508
llamaguard3	0.7982	0.8970	0.7998	0.8456
beaver	0.6047	0.9234	0.4668	0.6201
qi2023	0.8157	0.9662	0.7600	0.8508
JailJudge	0.7996	0.9303	0.7677	0.8412
Ours	0.8830	0.9010	0.9290	0.9148

Table 1: Jailbreak evaluation performance on SceneJailEval dataset.

safety judge; **Llamaguard3** (Chi et al. 2024): Enhanced Llamaguard variant with broader risk taxonomy and multilingual, long-context support; **Qi2023** (Qi et al. 2024): GPT-4-based detector achieving high accuracy for jailbreak evaluation; **JailJudge** (Liu et al. 2024a): A multi-agent jailbreak evaluation method. And the agents in our **SceneJailEval** are based on Qwen-3-235B.

Evaluation Metrics This study employs **Accuracy**, **Precision**, **Recall**, and **F1** (balancing precision and recall to reflect overall detection accuracy) to evaluate the performance of jailbreak detection methods across datasets. For quantifying the accuracy of the system’s harm scoring outputs, **NMAE** (deviation between system-generated harm scores and expert annotations) and **Spearman-Rho** (rank correlation between system harm scores and human judgments) are used. Additionally, a systematic evaluation of mainstream LLMs’ safety in multi-dimensional jailbreak scenarios is conducted via **ASR** (proportion of attacks eliciting unsafe responses) and **Average Harm Score** (mean harm level, with 0 indicating non-jailbreak).

Main Results and Analysis

Performance On SceneJailEval Dataset Table 1 reports SceneJail results: our method attains 0.915 F1, surpassing all SOTA baselines and evidencing the accuracy of our multi-dimensional scenario-adaptive jailbreak evaluator.

Models may include initial refusals but later provide potential harm—e.g., when queried “What are the steps to compose a fraudulent email,” a model might first deny assistance but follow with phishing details like “5. Using The Extracted Information...”. Baselines (Qi2023, llamaguard3, JailJudge) focus solely on initial denials, misclassifying this as non-jailbreak; our approach captures such harm via fine-grained scene classification and multi-dimensional evaluation. Another case involves region-specific non-obvious jailbreaks (e.g., queries about setting up cryptocurrency exchanges, permissible in Japan but non-compliant in mainland China). Baselines fail to distinguish such regional nuances, while our method, via targeted “Regional Compliance” evaluation, accurately identifies the harm. These cases underscore the need for contextual awareness and fine-grained evaluation in cross-scenario detection.

To validate our multi-dimensional evaluation, five security experts rated harmfulness via our scenario-adaptive criteria, with results in Table 2. Overall, $NMAE < 0.02$ and

Spearman-Rho near 1 confirm strong alignment between system-generated harmfulness scores and expert evaluation.

Scenario	NMAE ↓	Spearman-Rho ↑
Overall	0.0130	0.9378

Table 2: Overall NMAE & Spearman-Rho vs. Expert Annotations

Performance On Open-source Dataset To validate the generality and robustness of our approach, extensive experiments were conducted on three widely used public jailbreak evaluation datasets—JBB, JailJudge, and Safe-RLHF—which also include samples beyond our 14 defined scenarios, enabling verification of performance on unseen edge cases and emerging risk types (results in Table 3).

- On JBB, our method achieves the highest F1 score (0.99), outperforming Llamaguard3 (0.98) and substantially exceeding StringMatching (0.86) and Beaver (0.61).
- On JailJudge, it sets a new SOTA with an F1 score of 0.8241, surpassing JailJudge (0.8089) and Qi2023 (0.8012).
- On Safe-RLHF, despite Beaver (specifically fine-tuned on this dataset) leading, our method ranks second with an F1 score of 0.83, outperforming JailJudge (0.81) and Qi2023 (0.79).

Overall, our method achieves SOTA on JBB and JailJudge, and strong performance on Safe-RLHF, demonstrating superior performance across diverse datasets.

Ablation Study

We progressively remove SceneJailEval’s two key components while freezing the LLM backbone and prompts. Discarding scenario classification (DimsOnly) still yields high recall (91.8 %) yet lowers F1 by 2.7 %, indicating that uniform dimension scoring introduces false positives. Further removing dimension selection and reverting to vanilla heuristic rules (Vanilla) causes an additional 8.6 % F1 drop (83.1 %), confirming the necessity of scenario-adaptive evaluation. Overall, the full SceneJailEval achieves the best balance with 91.7 % F1 (Table 4).

Method	accuracy	precision	recall	f1
Vanilla	0.7676	0.8333	0.8296	0.8314
SceneOnly	0.7829	0.8444	0.8407	0.8425
DimsOnly	0.8440	0.8646	0.9181	0.8903
Ours	0.8900	0.8951	0.9398	0.9169

Table 4: Ablation study on SceneJailEval dataset.

Comprehensive Security Evaluation of LLMs

Using SceneJailEval, we evaluated mainstream LLMs (Gemini2.5Flash, GPT-4o, Claude 3.5, Llama) via joint ASR and Average Harm Score (Harm) assessment. Results reveal

Datasets	JBB				JailJudge				Safe-RLHF			
	acc	prec	rec	f1	acc	prec	rec	f1	acc	prec	rec	f1
StringMatch	0.86	0.88	0.84	0.86	0.72	0.57	0.68	0.62	0.60	0.99	0.59	0.74
llamaguard2	0.86	0.96	0.74	0.84	0.82	0.76	0.66	0.71	0.75	0.62	0.93	0.74
llamaguard3	<u>0.98</u>	0.95	1.0	<u>0.98</u>	0.83	0.72	0.81	0.77	0.72	0.53	0.96	0.68
Beaver	0.72	0.96	0.45	0.61	0.80	0.80	0.55	0.65	0.90	0.88	<u>0.94</u>	0.91
Qi2023	0.93	0.88	0.99	0.93	0.83	0.68	0.98	0.80	0.80	0.69	0.94	0.79
JailJudge	0.96	<u>0.99</u>	0.93	0.96	<u>0.87</u>	0.80	0.82	<u>0.81</u>	0.81	0.76	0.88	0.81
Ours	0.99	1.0	<u>0.99</u>	0.99	0.87	<u>0.76</u>	<u>0.90</u>	0.82	<u>0.81</u>	<u>0.89</u>	0.77	<u>0.83</u>

Table 3: Jailbreak evaluation performance on JBB, JailJudge, and Safe-RLHF datasets.

Model	GPT-4o		Claude-3.5		Gemini-2.5		Llama-3-8b		Llama-2-7b	
	ASR	Harm	ASR	Harm	ASR	Harm	ASR	Harm	ASR	Harm
Violent Crime	0%	0	0%	0	3.85%	0.13	3.85%	0.15	0%	0
Non-violent Crime	17.65%	0.50	2.61%	0.04	15.36%	0.47	0.65%	0.01	3.92%	0.09
Sex-related Crime	0%	0	0%	0	0%	0	0%	0	0%	0
Child Sexual Exploit.	10.71%	0.07	10.71%	0.08	0%	0	0%	0	0%	0
False Info. & Defamation	6.72%	0.26	0%	0	4.48%	0.13	8.96%	0.23	2.24%	0.04
Professional Advice	7.27%	0.16	1.82%	0.02	1.82%	0.02	0%	0	7.27%	0.18
Privacy Invasion	3.64%	0.13	0%	0	2.73%	0.09	1.82%	0.04	3.64%	0.07
IP Infringement	10.94%	0.29	0.78%	0.02	9.38%	0.15	18.75%	0.34	15.62%	0.27
WMDs	3.88%	0.13	0.97%	0.02	3.88%	0.15	1.94%	0.08	0.97%	0.00
Hate & Discrimination	1.06%	0.03	2.13%	0.04	0%	0	0%	0	0%	0
Suicide & Self-harm	6.87%	0.21	0%	0	0%	0	0%	0	0%	0
Sexual Content	25.00%	0.63	0%	0	0%	0	0%	0	7.14%	0.10
Pol. Agitation & Election	17.00%	0.63	0%	0	2.00%	0.07	3.00%	0.10	3.00%	0.04
Regional Sens. Issues	70.37%	2.76	8.33%	0.20	23.15%	0.74	29.63%	0.99	25.93%	0.87
Overall	15.06%	0.50	1.91%	0.03	9.11%	0.21	6.88%	0.14	5.89%	0.14

Table 5: Cross-scenario LLM security assessment via SceneJailEval.

Claude-3.5 as most robust (Avg Harm: 0.033) and GPT-4o as most vulnerable (0.502; Table 5).

Scenario-wise analysis uncovers a critical correlation: Regional Sensitive Issues consistently challenge all models, while Sex-related Crime elicits strong resilience—highlighting scenario-dependent security mechanisms.

Notably, each model exhibits unique vulnerability profiles: GPT-4o (25% ASR in Sexual Content), Claude-3.5 (10.71% in Child Sexual Exploitation), Gemini-2.5-Flash (15.36% in Non-violent Crime), and Llama variants (IP Infringement). These findings highlight the heterogeneous vulnerability landscape across models, informing robustness enhancements against diverse jailbreak threats.

Case Study on Customized Scenario

To address diverse organizational compliance needs, our framework is engineered with exceptional extensibility, seamlessly supporting customized detection scenarios and evaluation dimensions. To validate this, we designed a “Product Consultation” custom scenario (requiring models to avoid self-product derogation), where our Custom Generation Agent automatically generated the scenario category, detection dimension (Loyalty), and harm dimensions (Derogation, Specificity). Evaluating a 200-query custom dataset

via SceneJailEval’s annotation protocol (Table 6) yielded stellar results, conclusively demonstrating the framework’s robust extensibility and precise evaluation in tailored scenarios—underscoring its practical versatility.

acc	prec	rec	f1	NMAE	Spearman-Rho
1.0	1.0	1.0	1.0	0.037	0.841

Table 6: Performance on the customized scenario.

Conclusion

SceneJailEval revolutionizes LLM jailbreak evaluation with a paradigm-shifting scenario-adaptive framework, eliminating “one-size-fits-all” limitations and offering seamless extensibility for diverse needs. Complemented by a groundbreaking multi-scenario dataset—rich in variants and regional cases—it fills the critical gap in high-quality scenario-aware benchmarks. Boasting SOTA performance (0.917 F1 on our dataset, +6% over prior; 0.995 F1 on JBB, +3% over prior), it shatters accuracy bottlenecks in heterogeneous scenarios. These advances set a new standard for context-aware LLM security assessment, strengthening jailbreak defenses and accelerating trustworthy AI progress.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant Nos. 62572316 and 62302303, the Natural Science Foundation of Shanghai under Grant Nos. 25ZR1402279 and 23ZR1434000, and the Shanghai Pujiang Program under Grant No. 24PJJD043.

References

- Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Banerjee, S.; Layek, S.; Hazra, R.; and Mukherjee, A. 2025. How (un) ethical are instruction-centric responses of llms? unveiling the vulnerabilities of safety guardrails to harmful queries. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 193–205.
- Cai, H.; Arunasalam, A.; Lin, L. Y.; Bianchi, A.; and Celik, Z. B. 2024. Rethinking how to evaluate language model jailbreak. *arXiv preprint arXiv:2404.06407*.
- Chao, P.; DeBenedetti, E.; Robey, A.; Andriushchenko, M.; Croce, F.; Schwag, V.; Dobriban, E.; Flammarion, N.; Pappas, G. J.; Tramer, F.; et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37: 55005–55029.
- Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G. J.; and Wong, E. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 23–42. IEEE.
- Cheng, Z.; Wu, X.; Yu, J.; Han, S.; Cai, X.-Q.; and Xing, X. 2024. Soft-label integration for robust toxicity classification. *Advances in Neural Information Processing Systems*, 37: 94776–94807.
- Chi, J.; Karn, U.; Zhan, H.; Smith, E.; Rando, J.; Zhang, Y.; Plawiak, K.; Coudert, Z. D.; Upasani, K.; and Pasupuleti, M. 2024. Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations. *arXiv:2411.10414*.
- Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2024. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *The Twelfth International Conference on Learning Representations*.
- Dalkey, N.; and Helmer, O. 1963. An experimental application of the Delphi method to the use of experts. *Management science*, 9(3): 458–467.
- Ding, P.; Kuang, J.; Ma, D.; Cao, X.; Xian, Y.; Chen, J.; and Huang, S. 2023. A wolf in sheep’s clothing: Generalized nested jailbreak prompts can fool large language models easily. *arXiv preprint arXiv:2311.08268*.
- Du, Y.; Zhao, S.; Ma, M.; Chen, Y.; and Qin, B. 2023. Analyzing the inherent response tendency of llms: Real-world instructions-driven jailbreak. *arXiv preprint arXiv:2312.04127*.
- EU. 2024. Artificial Intelligence Act. <https://artificialintelligenceact.eu>. Accessed: 2025-8-1.
- Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.
- Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
- Ghosh, S.; Frase, H.; Williams, A.; Luger, S.; Röttger, P.; Barez, F.; McGregor, S.; Fricklas, K.; Kumar, M.; Bollacker, K.; et al. 2025. Ailuminat: Introducing v1. 0 of the ai risk and reliability benchmark from mlcommons. *arXiv preprint arXiv:2503.05731*.
- Howard, M.; and LeBlanc, D. 2003. *Writing secure code*. Pearson Education.
- Huang, R.; Wang, X.; Li, Z.; Wu, D.; and Wang, S. 2025. Guidedbench: Equipping jailbreak evaluation with guidelines. *arXiv preprint arXiv:2502.16903*.
- Huang, Y.; Gupta, S.; Xia, M.; Li, K.; and Chen, D. 2023. Catastrophic jailbreak of open-source llms via exploiting generation. *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 Khabisa, M. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. *arXiv:2312.06674*.
- Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2023. BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 24678–24704. Curran Associates, Inc.
- Lapid, R.; Langberg, R.; and Sipper, M. 2023. Open sesame! universal black box jailbreaking of large language models. *arXiv preprint arXiv:2309.01446*.
- Liao, Z.; and Sun, H. 2024. Amplegcy: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed llms. *arXiv preprint arXiv:2404.07921*.
- Liu, F.; Feng, Y.; Xu, Z.; Su, L.; Ma, X.; Yin, D.; and Liu, H. 2024a. JAILJUDGE: A Comprehensive Jailbreak Judge Benchmark with Multi-Agent Enhanced Explanation Evaluation Framework. *arXiv:2410.12855*.
- Liu, T.; Zhang, Y.; Zhao, Z.; Dong, Y.; Meng, G.; and Chen, K. 2024b. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, 4711–4728.
- Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023a. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.
- Liu, Y.; Cong, T.; Zhao, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2023b. Robustness Over Time: Understanding Adversarial Examples’ Effectiveness on Longitudinal Versions of Large Language Models. *arXiv preprint arXiv:2308.07847*.
- Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024.

- Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.
- Mell, P.; Scarfone, K.; and Romanosky, S. 2007. Common vulnerability scoring system. *IEEE Security & Privacy*, 4(6): 85–89.
- Nalic, D.; Mihalj, T.; Bäuml, M.; Lehmann, M.; Eichberger, A.; and Bernsteiner, S. 2020. Scenario based testing of automated driving systems: A literature survey. In *FISITA web Congress*, volume 10.
- NIST. 2023. Artificial Intelligence Risk Management Framework. <https://www.nist.gov/itl/ai-risk-management-framework>. Accessed: 2025-8-10.
- Paulus, A.; Zharmagambetov, A.; Guo, C.; Amos, B.; and Tian, Y. 2024. Advprompter: Fast adaptive adversarial prompting for llms. *arXiv preprint arXiv:2404.16873*.
- Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024. Visual Adversarial Examples Jailbreak Aligned Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19): 21527–21536. Publisher Copyright: Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 38th AAAI Conference on Artificial Intelligence, AAAI 2024 ; Conference date: 20-02-2024 Through 27-02-2024.
- Qiu, H.; Zhang, S.; Li, A.; He, H.; and Lan, Z. 2023. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models. *arXiv preprint arXiv:2307.08487*.
- Rauh, M.; Mellor, J.; Uesato, J.; Huang, P.-S.; Welbl, J.; Weidinger, L.; Dathathri, S.; Glaese, A.; Irving, G.; Gabriel, I.; et al. 2022. Characteristics of harmful text: Towards rigorous benchmarking of language models. *Advances in Neural Information Processing Systems*, 35: 24720–24739.
- Ryser, J.; and Glinz, M. 1999. A scenario-based approach to validating and testing software systems using statecharts.
- Saaty, T. L. 1980. The analytic hierarchy process mcgraw hill, New York. *Agricultural Economics Review*, 70(804): 10–21236.
- Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2024. ”do anything now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 1671–1685.
- Shu, D.; Zhang, C.; Jin, M.; Zhou, Z.; and Li, L. 2025. Attacker: How to evaluate the effectiveness of jailbreak attacking on large language models. *ACM SIGKDD Explorations Newsletter*, 27(1): 10–19.
- Souly, A.; Lu, Q.; Bowen, D.; Trinh, T.; Hsieh, E.; Pandey, S.; Abbeel, P.; Svegliato, J.; Emmons, S.; Watkins, O.; et al. 2024. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37: 125416–125440.
- Sun, J.; Zhang, H.; Zhou, H.; Yu, R.; and Tian, Y. 2021. Scenario-based test automation for highly automated vehicles: A review and paving the way for systematic safety assurance. *IEEE transactions on intelligent transportation systems*, 23(9): 14088–14103.
- Sutcliffe, A. G.; Maiden, N. A.; Minocha, S.; and Manuel, D. 1998. Supporting scenario-based requirements engineering. *IEEE Transactions on software engineering*, 24(12): 1072–1088.
- Xiao, Z.; Yang, Y.; Chen, G.; and Chen, Y. 2024. Distract large language models for automatic jailbreak attack. *arXiv preprint arXiv:2403.08424*.
- Yu, X.; Blanco, E.; and Hong, L. 2022. Hate speech and counter speech detection: Conversational context does matter. *arXiv preprint arXiv:2206.06423*.
- Yuan, Z.; Xiong, Z.; Zeng, Y.; Yu, N.; Jia, R.; Song, D.; and Li, B. 2024. Rigorllm: Resilient guardrails for large language models against undesired content. *arXiv preprint arXiv:2403.13031*.
- Zeng, Y.; Wu, Y.; Zhang, X.; Wang, H.; and Wu, Q. 2024. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*.
- Zhang, H.; Guo, Z.; Zhu, H.; Cao, B.; Lin, L.; Jia, J.; Chen, J.; and Wu, D. 2024a. Jailbreak open-sourced large language models via enforced decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5475–5493.
- Zhang, Y.; Ding, L.; Zhang, L.; and Tao, D. 2024b. Intention analysis makes llms a good jailbreak defender. *arXiv preprint arXiv:2401.06561*.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36: 46595–46623.
- Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.