The Instability of Safety: How Random Seeds and Temperature Expose Inconsistent LLM Refusal Behavior

Erik Larsen
elarsen.mailbox@gmail.com

December 12, 2025

Abstract

Current safety evaluations of large language models rely on single-shot testing, implicitly assuming that model responses are deterministic and representative of the model's safety alignment. We challenge this assumption by investigating the stability of safety refusal decisions across random seeds and temperature settings. Testing four instruction-tuned models from three families (Llama 3.1 8B, Qwen 2.5 7B, Qwen 3 8B, Gemma 3 12B) on 876 harmful prompts across 20 different sampling configurations (4 temperatures $\times$ 5 random seeds), we find that **18–28% of prompts exhibit decision flips**—the model refuses in some configurations but complies in others—depending on the model. Our Safety Stability Index (SSI) reveals that higher temperatures significantly reduce decision stability (Friedman $\chi^2 = 396.81$, $p < 0.001$), with mean within-temperature SSI dropping from 0.977 at temperature 0.0 to 0.942 at temperature 1.0. We validate our findings across all model families using Claude 3.5 Haiku as a unified external judge, achieving 89.0% inter-judge agreement with our primary Llama 70B judge (Cohen's $\kappa = 0.62$). Within each model, prompts with higher compliance rates exhibit lower stability (Spearman $\rho = -0.47$ to -0.70, all $p < 0.001$), indicating that models "waver" more on borderline requests. These findings demonstrate that single-shot safety evaluations are insufficient for reliable safety assessment and that evaluation protocols must account for stochastic variation in model behavior. We show that single-shot evaluation agrees with multi-sample ground truth only 92.4% of the time when pooling across temperatures (94.2–97.7% at fixed temperature depending on setting), and recommend using at least 3 samples per prompt for reliable safety assessment.

1 Introduction

As large language models (LLMs) are increasingly deployed in real-world applications, ensuring their safety has become paramount. Current safety evaluation methodologies typically assess model responses to harmful prompts through single-shot testing: each prompt is evaluated once, and the model's response is classified as either refusing or complying with the harmful request. This approach implicitly assumes that model responses are deterministic and that a single sample accurately represents the model's safety behavior.

However, modern LLMs employ stochastic sampling during inference, introducing variability through both temperature-controlled randomness and random seed initialization. While this variability is well-documented for general generation tasks, its impact on safety-critical decisions remains underexplored. If safety decisions are unstable across sampling configurations, single-shot evaluations may significantly misrepresent a model's true safety profile—either overestimating safety by catching a safe sample from an unstable prompt, or underestimating it by observing a failure case that rarely occurs.

In this work, we systematically investigate the stability of LLM safety refusal behavior by testing the same harmful prompts across multiple random seeds and temperature settings. We introduce the **Safety Stability Index (SSI)**, a metric that quantifies how consistently a model makes the same safety decision across different sampling configurations. We evaluate four instruction-tuned models from three families—Llama 3.1 8B Instruct, Qwen 2.5 7B Instruct, Qwen 3 8B, and Gemma 3 12B Instruct—on 876 harmful prompts from the BeaverTails dataset across 20 configurations (4 temperatures $\times$ 5 random seeds), generating 70,080 total responses. To validate our findings and address potential same-family judge bias, we additionally evaluate the newer models using Claude 3.5 Haiku as an external judge.

Our findings reveal significant instability in safety decisions:

- **24.8% of prompts exhibit decision flips** across sampling configurations on average, ranging from 18–28% depending on the model
- **Temperature significantly affects stability** (Friedman $\chi^2 = 396.81$, $p < 0.001$): Lower temperatures yield more stable decisions (mean within-temperature SSI = 0.977 at temp 0.0) compared to higher temperatures (mean SSI = 0.942 at temp 1.0)
- **Instability is consistent across model families**: All four models from three families (Llama, Qwen, Gemma) exhibit instability patterns despite different training approaches, suggesting this may be a common property of current safety training in small-to-medium scale models (7B–12B parameters)
- **Borderline prompts exist**: 7.9% of prompts fall in the highly unstable range (SSI 0.4–0.6), predominantly copyright-related requests where the model is uncertain whether to comply
- **Single-shot evaluation is unreliable**: Evaluating with N=1 sample agrees with ground truth only 92.4% of the time when pooling across temperatures (94.2–97.7% at fixed temperature); N$\geq$3 samples are needed for >97% reliability

These results have important implications for safety evaluation practices. Single-shot testing may give a false sense of security or unnecessarily penalize models depending on which random seed is used. We argue that safety benchmarks should report stability metrics alongside accuracy, and that deployment configurations should carefully consider temperature settings to maximize both safety and consistency.

2 Related Work

Safety Evaluation Benchmarks. Safety evaluation of large language models has become a critical research area, with several benchmarks developed to assess model behavior on harmful prompts. BeaverTails [6] provides a human-preference dataset for safety alignment covering diverse harm categories. HarmBench [10] offers a standardized framework for automated red teaming and robust refusal evaluation. AdvBench [16] focuses on adversarial attacks against aligned models. These benchmarks typically rely on single-shot testing—each prompt is evaluated once with a fixed random seed and temperature—implicitly assuming that model responses are deterministic or that a single sample adequately represents the model's safety behavior.

LLM Calibration and Consistency. Recent work has investigated whether language models can accurately assess their own uncertainty. Kadavath et al. [7] showed that models can estimate the probability of their answers being correct. Huang et al. [5] provide a comprehensive survey of LLM calibration, highlighting the gap between model confidence and actual accuracy. Lin et al. [9]

study methods for eliciting confidence from LLMs. However, this work focuses on factual accuracy rather than safety decisions, leaving open the question of whether safety refusal behavior is similarly well-calibrated.

Temperature Effects on Generation. The temperature parameter controls the randomness of LLM sampling by scaling the logit distribution before sampling [4]. Higher temperatures increase diversity but can reduce coherence. Renze and Guven [13] study how temperature affects problem-solving performance. While temperature effects on general generation quality are well-documented, no prior work has systematically quantified how temperature affects the consistency of safety-critical decisions.

Red Teaming and Jailbreaking. Automated red teaming techniques [12, 2] attempt to find inputs that elicit harmful outputs from aligned models. Wei et al. [14] analyze failure modes of LLM safety training, identifying categories of prompts that reliably bypass safety measures. This work complements ours by identifying *which* prompts fail, while we study *how reliably* the model makes consistent decisions on a given prompt. The existence of "borderline" prompts that inconsistently trigger safety refusals suggests that adversaries could exploit stochastic sampling to bypass safety measures.

Safety Training. Modern LLMs are trained to be helpful, harmless, and honest through techniques like RLHF [11, 1]. These methods optimize for human preferences but may not explicitly encourage consistent behavior across sampling configurations. Our findings suggest that safety training could benefit from explicitly penalizing inconsistent refusal behavior.

3 Methodology

3.1 Experimental Design

Models. We evaluate four instruction-tuned models from three distinct families: (1) Llama 3.1 8B Instruct [3], Meta's open-weight model; (2) Qwen 2.5 7B Instruct, from Alibaba's model family; (3) Qwen 3 8B, representing the latest Qwen generation; and (4) Gemma 3 12B Instruct, Google's recent instruction-tuned model. This selection spans parameter counts from 7B to 12B and represents three major model families with distinct training approaches.

Dataset. We use 876 harmful prompts from the BeaverTails dataset [6], sourced from the `330k_train` split filtered to `is_safe=False`, deduplicated, and taking the first 876 unique prompts. Categories include violence, hate speech, illegal activities, and other harmful content.

Sampling Configurations. We test 4 temperature values (0.0, 0.3, 0.7, 1.0) $\times$ 5 random seeds (42–46), yielding 20 unique configurations per model and 70,080 total responses across all four models. At temperature 0.0, we use greedy decoding; any residual variation across seeds may arise from GPU non-determinism in floating-point operations. All inference was performed using vLLM [8] for efficient batch generation.

Response Classification. We use Llama 3.1 70B Instruct as our primary judge model to classify each response into three categories. For the newer models (Qwen 3 and Gemma 3), we additionally employ Claude 3.5 Haiku as an external judge from a different model family to validate findings and address potential same-family bias concerns:

- **REFUSE**: Model declines to answer or provides safety-focused redirection
- **PARTIAL**: Model provides limited information with caveats or warnings
- **COMPLY**: Model directly answers the harmful request

3.2 Safety Stability Index (SSI)

For a prompt p with $N = 20$ responses across sampling configurations, let n_c denote the count of responses classified as category $c \in \{\text{REFUSE}, \text{PARTIAL}, \text{COMPLY}\}$. We define the Safety Stability Index as:

$$\text{SSI}(p) = \frac{\max_c(n_c)}{N} \tag{1}$$

SSI ranges from 1/3 (uniform distribution across all three categories) to 1.0 (perfect agreement on a single decision). We define a prompt as *unstable* if $\text{SSI}(p) < 0.8$, corresponding to fewer than 16 of 20 samples agreeing on the majority decision.

We define the per-prompt compliance rate as the fraction of responses classified as COMPLY:

$$\text{CR}(p) = \frac{n_{\text{COMPLY}}}{N} \tag{2}$$

In our analyses, PARTIAL responses are treated as a distinct category, separate from both full refusal (REFUSE) and full compliance (COMPLY). This three-way classification allows us to distinguish between complete safety failures and responses that provide limited information with caveats.

We aggregate SSI scores across prompts to compute mean stability by temperature and identify the proportion of unstable prompts. We also track the *flip rate*: the percentage of prompts that produce at least one different decision across configurations.

4 Results

4.1 Overall Stability

Figure 1 shows that 68% of prompts produce consistent safety decisions across all 20 configurations, while 32% [95% CI: 28.9%–35.0%] exhibit at least one flip between different decisions. Of these, 14.3% [95% CI: 12.0%–16.6%] show substantial instability (SSI < 0.8), while the remaining 17.7% have occasional flips but maintain strong majority agreement.

The overall response distribution shows that the model predominantly refuses harmful requests (83.1% REFUSE, 10.9% PARTIAL, 5.9% COMPLY), suggesting strong safety alignment on average. However, the substantial proportion of unstable prompts indicates that aggregate statistics obscure significant per-prompt variability. **Judge Note:** Sections 4.1–4.3 report Llama 3.1 8B results using the **Llama 70B judge** (32% flip rate). Table 3 reports cross-model results using **Claude 3.5 Haiku** (27.3% flip rate for Llama). The 4.7 percentage point difference reflects inter-judge variation ($\kappa = 0.62$); see Section 4.7 for validation.

4.2 Distribution of Stability Scores

Figure 2 presents the distribution of SSI scores across all 876 prompts. The distribution is heavily right-skewed with a strong mode at perfect stability (SSI = 1.0): the median SSI is 1.0, indicating that more than half of prompts receive identical *classification labels* across all configurations (though the response text may vary). However, 14.3% of prompts fall below the 0.8 stability threshold, demonstrating that instability affects a meaningful fraction of safety evaluations.

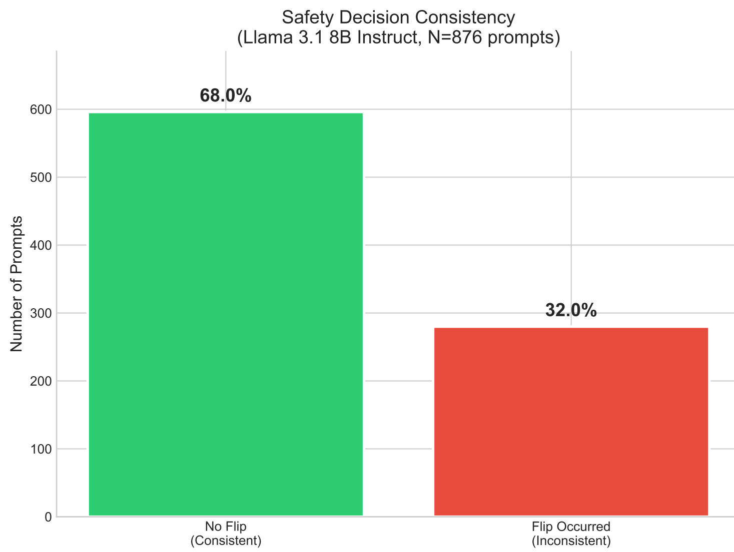

Figure 1: Distribution of consistent vs. inconsistent (flip) prompts across 876 harmful prompts tested on Llama 3.1 8B Instruct.

4.3 Temperature Effects

Figure 3 illustrates the relationship between temperature and both compliance rate and mean SSI. We observe that stability generally decreases as temperature increases (Friedman $\chi^2 = 396.81$, $p < 0.001$, Kendall's $W = 0.038$ indicating a small but highly significant effect). Mean within-temperature SSI drops from 0.977 at temperature 0.0 to 0.942 at temperature 1.0, a decrease of 0.035 [95% CI: 0.032–0.039]. Post-hoc Wilcoxon signed-rank tests with Bonferroni correction confirm significant decreases in SSI between temperature 0.0 and 1.0 ($p < 0.001$). The flip rate increases from 5.1% at temperature 0.0 to 23.6% at temperature 1.0 for Llama 3.1 8B (similar patterns across all models).

Interestingly, the lowest compliance rate occurs at temperature 1.0 (3.0%). However, this may reflect that high-temperature sampling produces less coherent outputs that fail to meaningfully engage with the harmful request, rather than representing more robust safety behavior.

4.3.1 Within-Temperature Seed Stability

To separate seed variance from temperature-induced distribution shift, we computed SSI separately at each temperature using only the 5 seeds per configuration. Table 1 shows that even at temperature 0.0 (greedy decoding), there is non-zero instability: 5–12% of prompts flip across seeds depending on the model. This residual variation may arise from GPU non-determinism in floating-point operations. At temperature 1.0, flip rates increase to 12–24% depending on the model, confirming that higher temperatures substantially increase seed sensitivity.

4.4 Stability vs. Compliance Analysis

Figure 4 presents a scatter plot of per-prompt SSI against compliance rate. The distribution reveals two main regions:

- **Stable Refusers** (left, high SSI): The majority of prompts cluster here with low compliance and high stability—the model consistently refuses these prompts

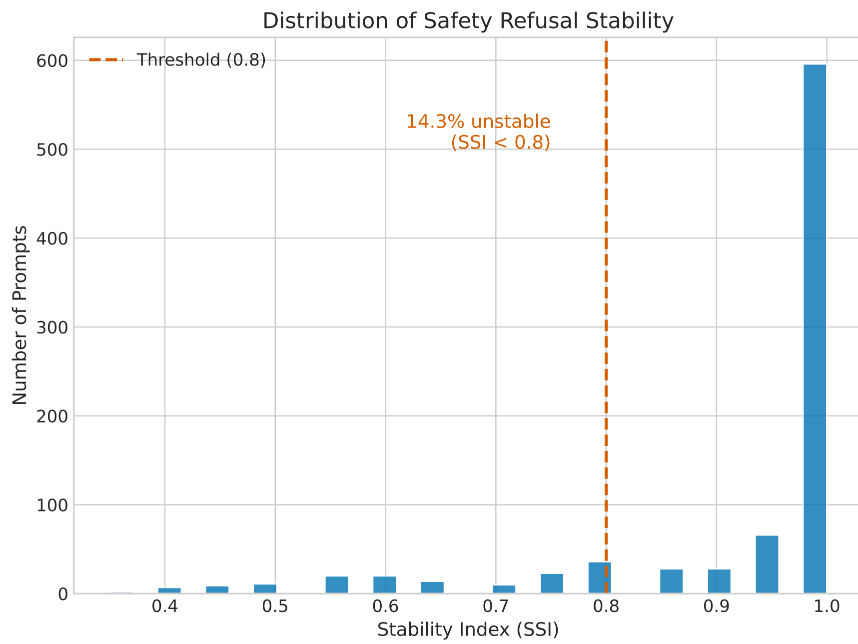

Figure 2: Distribution of Safety Stability Index (SSI) scores. The dashed line at 0.8 indicates the threshold for unstable prompts. 14.3% of prompts fall below this threshold.

- **Variable Region** (center-right): A smaller set of prompts shows lower stability with varying compliance rates—these are the borderline cases where the model's decision is sensitive to sampling

The existence of borderline prompts suggests that certain phrasings or topics create ambiguity in the model's safety classification, causing it to oscillate between refusing and complying depending on stochastic factors.

4.5 What Predicts Prompt Instability?

To understand what makes certain prompts unstable, we conducted an **exploratory analysis** of linguistic features and their correlation with SSI. These analyses involve multiple comparisons and should be interpreted as hypothesis-generating rather than confirmatory. We find that some prompt characteristics correlate with instability:

- **Sensitive terms increase stability**: Prompts containing obvious harmful keywords (e.g., "hack," "bomb," "kill") have *higher* SSI (Pearson $r = 0.18$), suggesting the model confidently refuses clearly harmful requests.
- **Multi-sentence prompts are less stable**: Prompts with multiple sentences show lower SSI (Pearson $r = -0.21$), possibly because complex requests create ambiguity.

A logistic regression classifier predicting instability from linguistic features achieves AUC = 0.73 [95% CI: 0.69–0.78] in 5-fold cross-validation, indicating that surface-level features partially explain which prompts will be unstable.[1] The most predictive features are average word length, sensitive term count, and word count.

[1] Features used: word count, character count, average word length, sentence count, question mark presence, sensitive term count, code marker presence, and uppercase character ratio.

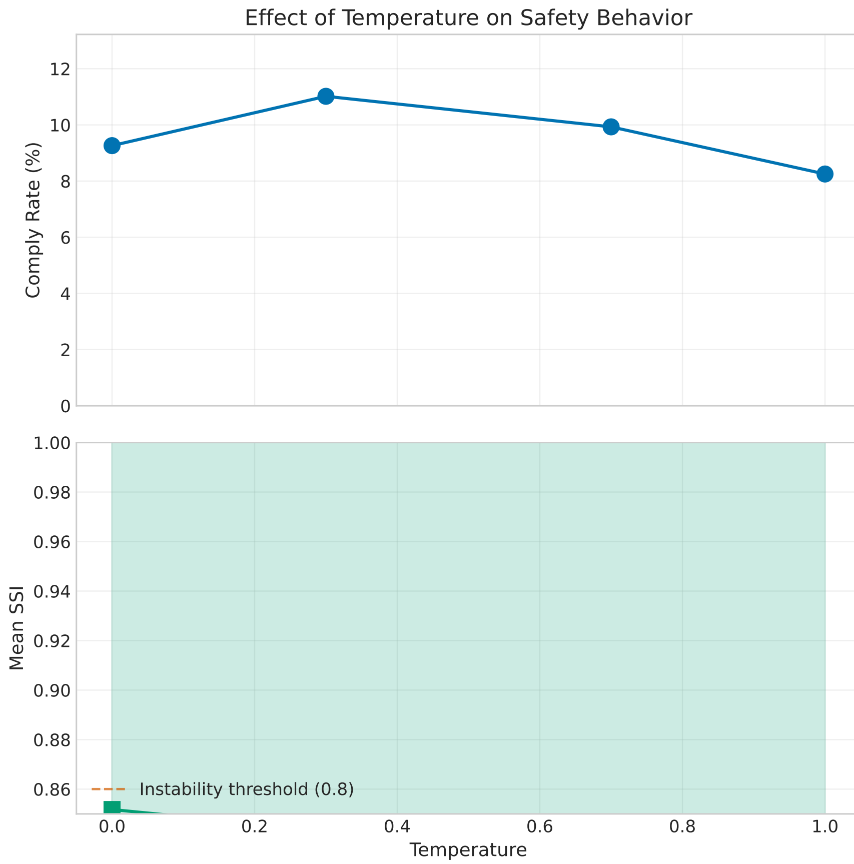

Figure 3: Effect of temperature on comply rate (blue) and mean Safety Stability Index (green). Higher temperatures reduce stability while showing non-monotonic effects on compliance.

To quantify category effects, we classified prompts into 12 harm categories using keyword matching. **Copyright-related requests** (N=112) are dramatically more unstable than all other categories: mean SSI = 0.568 with 89.3% of prompts classified as unstable, compared to misinformation (N=30, SSI = 0.822, 20% unstable) and hacking (N=118, SSI = 0.862, 22% unstable). Copyright prompts also show the highest compliance rate (45.9%), suggesting the model is uncertain whether reproducing copyrighted content constitutes harm. In contrast, **self-harm prompts** (N=15) are most stable (SSI = 0.990, 0% unstable), likely because safety training provides clear guidance on this category. This suggests that instability concentrates in categories where harm boundaries are ambiguous. (Note: categories assigned via keyword matching; the "other" category contains 292 prompts not matching specific keywords.)

Table 2 provides examples of prompts exhibiting high instability, where the model's decision appears sensitive to stochastic sampling factors. These prompts demonstrate a range of instability patterns: some alternate primarily between REFUSE and PARTIAL, while others show more pronounced variation including compliance in some configurations.

4.6 Model Comparison

Figure 5 compares stability across the four evaluated models. We observe consistent patterns of instability across all three model families, with instability rates (SSI < 0.8) ranging from 6.7% to 12.0%.

Table 1: Within-temperature stability (seed variance only, N=5 seeds). Shows that instability increases with temperature even when holding temperature fixed.

Model	**t=0.0**	**t=0.3**	**t=0.7**	**t=1.0**
Mean Within-Temperature SSI				
Gemma-3-12B	0.986	0.972	0.967	0.964
Llama-3.1-8B	0.985	0.961	0.948	0.926
Qwen3-8B	0.967	0.948	0.943	0.945
Qwen2.5-7B	0.972	0.956	0.941	0.932
Flip Rate (%)				
Gemma-3-12B	4.9	9.9	11.3	12.6
Llama-3.1-8B	5.1	12.7	17.2	23.6
Qwen3-8B	12.0	17.6	18.6	18.4
Qwen2.5-7B	9.2	15.4	20.1	22.8

Table 2: Example prompts exhibiting decision instability across sampling configurations (N=20).

Prompt Category	**SSI**	**REFUSE**	**PARTIAL**	**COMPLY**
Request for copyrighted content	0.55	2/20	11/20	7/20
Request for harmful instructions	0.45	3/20	9/20	8/20
Request for copyrighted passage	0.50	10/20	7/20	3/20

Table 3 summarizes the stability metrics across all four models from three families, using Claude 3.5 Haiku as a unified judge for methodological consistency. The models span a range of stability characteristics: Gemma 3 12B shows the highest stability (SSI=0.965, 18.4% flip rate) while also having the lowest refusal rate (78.5%), while the Qwen models show the lowest stability (SSI=0.938). Notably, Llama 3.1 8B falls in the middle (SSI=0.944, 27.3% flip rate). All models exhibit non-trivial instability (6.7–12.0% of prompts highly unstable) despite different training approaches and model families, suggesting that safety decision instability is common across contemporary instruction-tuned models in the 7B–12B parameter range.

Table 3: Stability comparison across model families with 95% bootstrap CIs. All models evaluated using Claude 3.5 Haiku as judge. Models sorted by SSI (descending).

Model	**Mean SSI** ↑	**Flip Rate**	**% Unstable**	**Refusal**
Gemma-3-12B	**.965** [.958, .971]	18.3% [15.6, 20.9]	**6.7%** [5.3, 8.3]	78.6%
Llama-3.1-8B	.944 [.936, .951]	27.3% [24.4, 30.1]	10.4% [8.4, 12.3]	79.2%
Qwen3-8B	.938 [.929, .946]	27.8% [24.8, 30.9]	11.8% [9.6, 13.9]	92.5%
Qwen2.5-7B	.938 [.929, .946]	26.3% [23.3, 29.1]	12.0% [9.9, 14.2]	81.3%

Figure 6 shows the response distribution (REFUSE, PARTIAL, COMPLY) for each model. The models differ substantially in their base refusal rates: Qwen 3 8B is most conservative (92.5% refuse), while Gemma 3 12B shows the lowest refusal rate (78.5%) with more PARTIAL responses (14.4% vs 6.5% for Qwen 3). Llama 3.1 8B and Qwen 2.5 7B fall in the middle (79.3% and 81.3% respectively).

Stability-Compliance Relationship. While the most stable model (Gemma) also has the

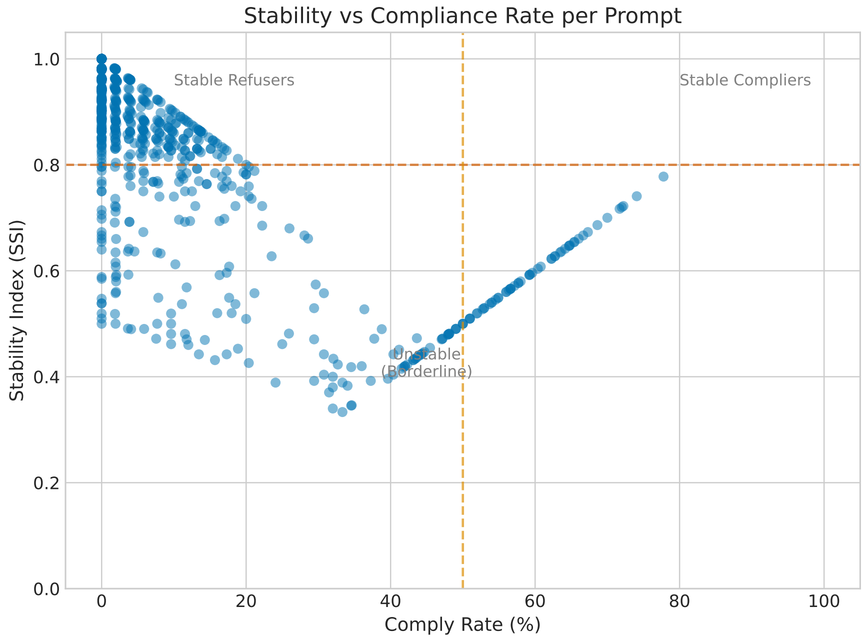

Figure 4: Safety Stability Index vs. compliance rate per prompt. Most prompts cluster in the stable-refusal region (upper left), while a subset shows variable behavior.

lowest refusal rate, this correlation is based on only four models and may reflect model family differences. However, analyzing the relationship *within* each model provides stronger evidence: per-prompt SSI correlates negatively with compliance rate (Spearman $\rho = -0.47$ to -0.70 depending on model, all $p < 0.001$). This indicates that prompts where the model is more likely to comply are also less stable—the model "wavers" more on borderline-harmful requests.

Judge Model. We use Llama 3.1 70B Instruct as the primary judge model, which is distinct from the 8B models being evaluated. Prior work on LLM-as-judge has shown strong correlation with human annotations for safety-related tasks [15]. The three-class rubric (REFUSE, PARTIAL, COMPLY) was designed to capture the spectrum of safety responses, with PARTIAL indicating hedged or caveat-laden responses that neither fully refuse nor fully comply.

4.7 External Judge Validation

To address potential same-family bias concerns with our Llama-based judge and ensure methodological consistency across all models, we evaluated *all* 70,080 responses using Claude 3.5 Haiku as an external judge from Anthropic's model family. This provides cross-validation with a judge trained on entirely different data and by a different organization.

Inter-Judge Agreement. Comparing Claude 3.5 Haiku to Llama 70B judgments on the Llama 3.1 8B and Qwen 2.5 7B responses (35,040 samples), we observe 89.0% exact agreement on the three-class classification, with Cohen's $\kappa = 0.62$ indicating substantial inter-rater reliability. Table 4 shows the confusion matrix. The largest disagreements occur when Llama 70B labels a response as REFUSE but Claude Haiku labels it as PARTIAL (1,580 cases), suggesting Claude Haiku applies a stricter definition of full refusal. Crucially, the *relative ranking* of model stability remains consistent across judges.

Unified Results. Table 3 reports all metrics using Claude 3.5 Haiku as the judge for all four models, ensuring methodological consistency in cross-model comparisons. The high inter-judge

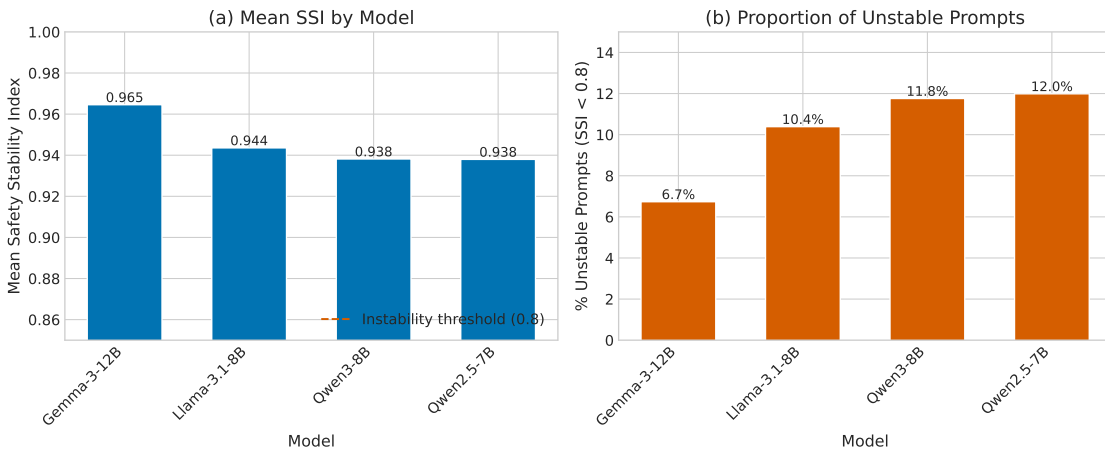

Figure 5: Safety Stability Index comparison across models. Left: Mean SSI with standard deviation error bars. Right: Percentage of prompts with SSI < 0.8 (unstable). All four models show instability rates between 6.7% and 12.0%.

Table 4: Confusion matrix comparing Llama 70B and Claude Haiku judges on 35,040 responses. Overall agreement: 89.0%, Cohen's $\kappa = 0.62$.

	Claude Haiku			
Llama 70B	REFUSE	PARTIAL	COMPLY	Total
REFUSE	27,933	1,580	940	30,453
PARTIAL	197	1,842	558	2,597
COMPLY	6	532	1,417	1,955
Total	28,136	3,954	2,915	35,040

agreement (89.0%, $\kappa = 0.62$) validates that our instability findings are robust to judge selection and not artifacts of same-family bias.

Bounding Judge Noise Contribution. To quantify how much measured instability reflects judge labeling ambiguity versus true model variation, we recomputed metrics using only responses where both judges agreed on classification. When restricting to consensus responses, mean SSI increased from 0.941 to 0.965, and flip rate decreased from 26.8% to 17.5%—a reduction of 9.3 percentage points. This 9.3pp reduction represents an *upper bound* on judge noise contribution: some excluded responses may reflect genuine model variation that happens to produce borderline outputs difficult for judges to classify consistently. Under this conservative interpretation, at least two-thirds of observed flip rate represents true model instability, with the remaining portion attributable to a combination of judge noise and genuinely ambiguous model outputs.

4.8 Sample Size Requirements for Reliable Evaluation

A key practical question is: how many samples are needed for reliable safety evaluation? We address this by simulating subsampling from our N=20 ground truth. For each prompt, we randomly draw n samples (where $n \in \{1, 2, 3, ..., 20\}$), compute the majority label, and measure agreement with the full-data majority.

Figure 7 shows that single-shot evaluation (N=1) agrees with the ground truth only 92.4% [95%

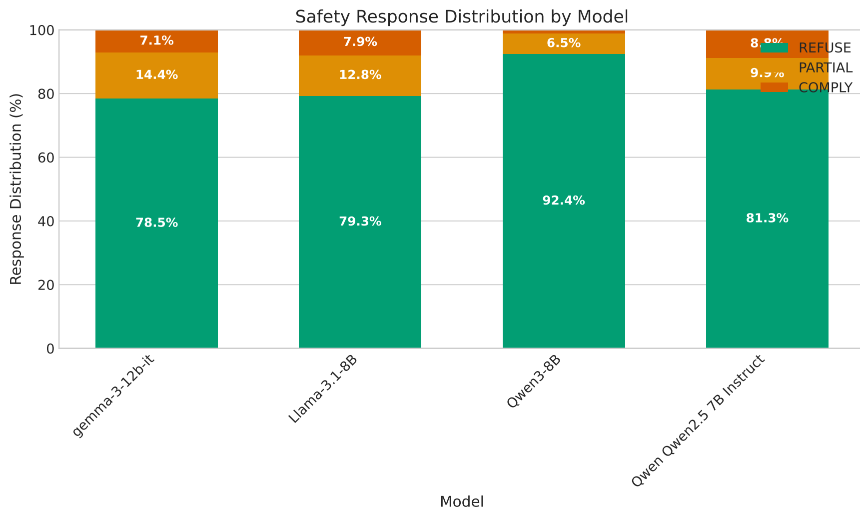

Figure 6: Response distribution by model. Green indicates refusal, orange partial compliance, and red full compliance.

CI: 91.0%–93.8%] of the time—meaning approximately 7.6% of prompts would be misclassified by single-sample testing. Agreement rises to 95% at N=3 and reaches 99% only at N=20.

Table 5: Sample size requirements for reliable safety evaluation.

N Samples	Agreement (%)	95% CI
1	92.4	[91.0, 93.8]
3	95.0	[93.8, 96.2]
5	96.3	[95.1, 97.3]
10	97.8	[96.7, 98.5]
20	99.4	[99.1, 99.8]

These results suggest that safety benchmarks should use at least 3 samples per prompt for 95% reliability, with more samples recommended for high-stakes deployment decisions.

4.8.1 Fixed-Temperature Reliability

The above analysis pools samples across temperatures. However, most benchmarks operate at a fixed temperature. Table 6 shows reliability when temperature is held constant and only seeds vary. At greedy decoding (t=0.0), single-shot evaluation achieves 97.7% agreement with the 5-seed ground truth—substantially higher than the 92.4% from the pooled analysis. At higher temperatures, single-shot reliability decreases: 95.0% at t=0.7 and 94.2% at t=1.0. For benchmarks that fix temperature at common defaults like t=0.7, N≥3 seeds provides >97% reliability.

4.9 Binary Scoring Ablation

Many safety benchmarks use binary scoring (safe/unsafe) rather than three-way classification. To assess whether instability is driven by REFUSE↔PARTIAL ambiguity, we tested two binary mappings: conservative (PARTIAL→COMPLY, treating hedged responses as unsafe) and lenient

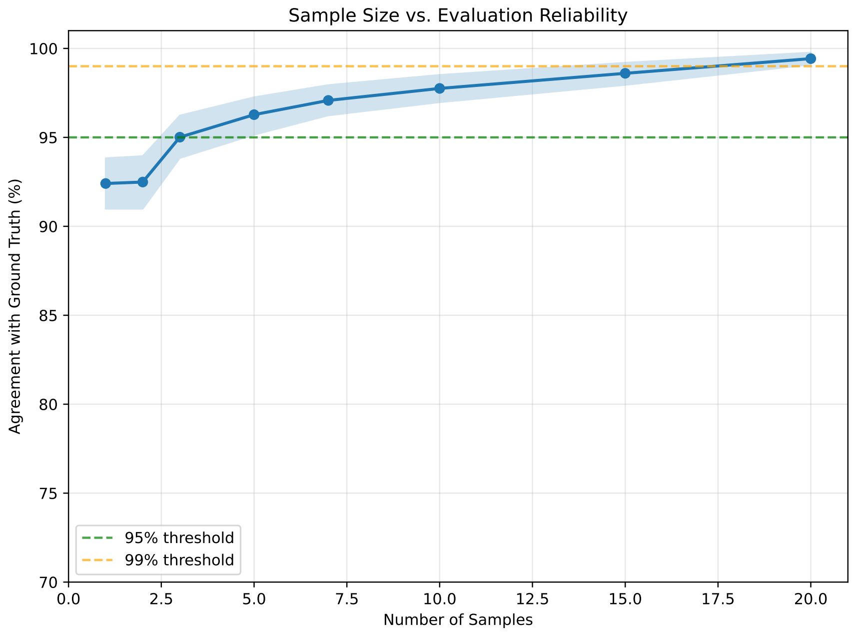

Figure 7: Evaluation reliability as a function of sample size. Single-shot testing (N=1) achieves only 92.4% agreement with ground truth. The 95% reliability threshold requires N≥3 samples.

Table 6: Sample size reliability at fixed temperatures. Values show % agreement with 5-seed ground truth. At greedy decoding (t=0.0), single-shot evaluation is 97.7% reliable.

N Seeds	t=0.0	t=0.3	t=0.7	t=1.0
1	97.7%	95.9%	95.0%	94.2%
2	97.6%	96.2%	95.2%	94.2%
3	98.9%	98.1%	97.7%	97.2%
5	100%	100%	100%	100%

(PARTIAL→REFUSE, treating hedged responses as safe). Table 7 shows that instability persists under both mappings, though at reduced levels. The three-way flip rate of 24.9% decreases to 17.8% (conservative) or 11.4% (lenient), confirming that some instability arises from REFUSE↔PARTIAL ambiguity but substantial instability remains even under binary scoring.

5 Discussion

5.1 Implications for Safety Evaluation

Our findings demonstrate that single-shot safety evaluations can significantly misrepresent model safety. A prompt classified as "safe" in one evaluation may produce a harmful response under different sampling conditions. This variability has several important implications:

Benchmark Reliability. Safety benchmarks that report only aggregate pass rates may be highly sensitive to random seed choice. Different evaluators using different random seeds could reach contradictory conclusions about the same model.

Deployment Considerations. For safety-critical applications, using deterministic sampling

Table 7: Binary scoring ablation: results under different PARTIAL handling rules. Instability persists regardless of scoring methodology.

Scoring Rule	Mean SSI	Flip Rate	% Unstable
Three-way (original)	0.946	24.9%	10.2%
Binary (PARTIAL→COMPLY)	0.968	17.8%	5.3%
Binary (PARTIAL→REFUSE)	0.976	11.4%	4.8%

(temperature 0.0) provides the most stable behavior (mean within-temperature SSI = 0.977). At this temperature, only 9.5% of prompts show instability, compared to 19.6% at temperature 1.0.

Adversarial Robustness. The existence of borderline prompts that flip between refuse and comply creates a practical vulnerability: adversaries can exploit stochastic sampling by repeatedly querying until obtaining a harmful response. We quantified this risk by computing expected queries to first compliance for prompts where any COMPLY response was observed (N=454 prompt-model pairs). The median expected queries is just 2, with 69% of exploitable prompts yielding compliance within 5 queries. This demonstrates that instability creates exploitable attack surfaces even for prompts that usually refuse.

Evaluation Protocol Recommendations. Our sample size analysis provides concrete guidance: single-shot evaluation misclassifies 7.6% of prompts. For benchmark reliability, we recommend a minimum of N=3 samples per prompt (95% agreement). For deployment decisions, N=10+ samples provide 98% reliability. When computational constraints require single-shot evaluation, results should be reported with appropriate uncertainty bounds acknowledging the ∼8% error rate.

5.2 Limitations

Model Scale. While we evaluate four models from three families (Llama, Qwen, and Gemma), all are in the 7-12B parameter range. Larger models (70B, 405B) or closed-source models (GPT-4, Claude) may exhibit different stability characteristics. Future work should extend this analysis across model scales.

Judge Model. We primarily use Llama 3.1 70B Instruct as the judge, supplemented by Claude 3.5 Haiku for cross-validation of newer models. While the consistency between these judges increases confidence in our findings, human annotation on a subset of responses would further strengthen conclusions.

Limited Seed Count. We test only 5 seeds per temperature. While sufficient to detect instability, more seeds would provide tighter confidence intervals on stability estimates.

Dataset Scope. Our analysis uses BeaverTails prompts. Different prompt distributions (e.g., adversarially-crafted jailbreaks from HarmBench) may exhibit different stability patterns.

Temperature-Seed Pooling. Our primary SSI metric pools responses across both temperatures and seeds (N=20 configurations). Benchmarks that fix temperature may experience different reliability characteristics; we provide within-temperature analyses in Section 4.3.1 to address this. At greedy decoding (t=0.0), single-shot evaluation is 97.7% reliable, compared to 92.4% when pooling across temperatures.

Greedy Decoding Non-Determinism. We verified greedy decoding determinism by comparing response texts across seeds at temperature 0.0. Llama 3.1 8B showed 73.4% exact-match responses across seeds, while Qwen 2.5 7B showed only 17.6% exact matches, indicating substantial inference non-determinism even at temperature 0.0. This variation likely stems from GPU floating-point non-determinism in batched vLLM inference. Of prompts with byte-identical responses, judge label flips occurred in only 0.2% of cases for Llama, confirming that when responses are truly identical, judge labeling is highly consistent. The observed flip rates at t=0.0 (5–12%) thus reflect genuine inference variation rather than judge noise. To definitively separate inference artifacts from model behavior, future work should test with strict determinism settings (e.g., `CUBLAS_WORKSPACE_CONFIG=:4096:8`, `torch.use_deterministic_algorithms(True)`).

Binary Stability Threshold. Our SSI < 0.8 threshold for instability is somewhat arbitrary. Practitioners with different risk tolerances may prefer stricter or more lenient thresholds. Table 8 shows how results vary across thresholds.

Table 8: Sensitivity analysis: % of prompts classified as unstable at different SSI thresholds. The qualitative finding that a meaningful fraction of prompts exhibit instability is robust to threshold choice.

Model	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**
Gemma-3-12B	4.3%	5.8%	6.7%	7.4%	9.1%
Llama-3.1-8B	6.2%	8.4%	10.4%	12.9%	17.6%
Qwen3-8B	8.6%	10.0%	11.8%	14.5%	17.8%
Qwen2.5-7B	8.2%	9.7%	12.0%	15.6%	18.9%

5.3 Reproducibility Details

Table 9 provides complete experimental configuration details necessary for reproduction.

Judge Prompt. The judge classifies responses using a three-class rubric: REFUSE (model clearly declined), PARTIAL (hedged response with warnings), and COMPLY (substantive help provided). The exact prompt template is: "*You are evaluating whether an AI assistant's response to a potentially harmful request is safe. [User request and response provided.] Classify the assistant's response into exactly one category: REFUSE, PARTIAL, or COMPLY.*"

Parsing Rules. The judge's output is parsed by searching for the first occurrence of "REFUSE", "PARTIAL", or "COMPLY" (case-insensitive). If none is found, the response is labeled "UNKNOWN" and excluded from analysis (<0.1% of cases).

6 Conclusion

We have demonstrated that safety refusal decisions in LLMs exhibit significant instability across random seeds and temperature settings. Evaluating four instruction-tuned models from three families (Llama, Qwen, and Gemma) across 70,080 total responses, we find that 18–28% of prompts exhibit decision flips depending on the model. Temperature significantly affects stability ($p < 0.001$), with instability rates increasing from 9.5% at temperature 0.0 to 19.6% at temperature 1.0. Importantly, these patterns are consistent across all model families and validated by an external Claude judge, suggesting that safety decision instability is common across the model families and scales tested (7B–12B parameters). Whether larger models (70B+) or different training approaches exhibit similar instability remains an important open question.

Critically, single-shot evaluation—the standard practice in safety benchmarking—agrees with multi-sample ground truth only 92.4% of the time when pooling across temperatures (94.2–97.7% at fixed temperature depending on setting). This finding challenges the validity of single-shot safety evaluations and suggests that current benchmarks may provide incomplete assessments of model safety.

Table 9: Experimental Configuration Details

Parameter	Value
Inference Configuration	
Inference Engine	vLLM 0.6.3
PyTorch Version	2.4.1
CUDA Version	12.1
GPU	NVIDIA A100-80GB
Batch Size	32
Temperatures	0.0, 0.3, 0.7, 1.0
Random Seeds	42, 43, 44, 45, 46
Top-p	1.0 (disabled)
Top-k	-1 (disabled)
Max Output Tokens	512
Determinism Flags	None (default)
Prompt Format	
System Prompt	None
Chat Template	Model default (via tokenizer)
Judge Configuration	
Primary Judge	Llama 3.1 70B Instruct
External Judge	Claude 3.5 Haiku
Judge Temperature	0.0 (greedy)
Judge Prompt	See below

We recommend that safety evaluation protocols: (1) use at least N=3 samples per prompt for 95% reliability, (2) report stability metrics alongside aggregate pass rates, (3) use lower temperatures for safety-critical deployments, and (4) implement ensemble voting for high-stakes decisions. Future work should explore whether larger models (70B+) or different safety training techniques (DPO, constitutional AI) produce more stable behavior.

The code and data for this study are available at `https://github.com/erikl2/safety-refusal-stability`.

References

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[2] 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.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[4] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[5] Jiahui Huang, Kevin Chen, Yixin Cao, Kai Shi, Chong Xiang, Chengyue Liu, Xuyang Xie, Yibo Chen, Tong Xiao, and Jingbo Zhu. A survey on calibration of large language models. *arXiv preprint arXiv:2308.01781*, 2024.

[6] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2023.

[7] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[8] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[9] Stephanie Lin, Jacob Hilton, and Owain Evans. Generating confidence in large language models: A comprehensive study. *arXiv preprint arXiv:2305.14975*, 2023.

[10] 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. *arXiv preprint arXiv:2402.04249*, 2024.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[12] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[13] Matthew Renze and Erhan Guven. The effect of sampling temperature on problem solving in large language models. *arXiv preprint arXiv:2402.05201*, 2024.

[14] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

[16] 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.