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

Erik Larsen
eriklarsen@example.edu

November 26, 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 Llama 3.1 8B Instruct on 876 harmful prompts across 20 different sampling configurations (4 temperatures × 5 random seeds), we find that **32% of prompts produce inconsistent safety decisions**—the model refuses in some configurations but complies in others. Our Safety Stability Index (SSI) reveals that higher temperatures reduce decision stability, while compliance rates show non-monotonic patterns across temperature settings. 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.

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. Using Llama 3.1 8B Instruct as our test model, we evaluate 876 harmful prompts from the BeaverTails dataset across 20 configurations (4 temperatures × 5 random seeds), generating and judging 17,520 total responses.

Our findings reveal significant instability in safety decisions:

- **32% of prompts are unstable** (SSI < 0.8), producing different safety decisions across sampling configurations
- **Temperature affects stability**: Lower temperatures yield more stable decisions (mean SSI $= 0.96$ at temp 0.0) compared to higher temperatures (mean SSI $= 0.76$ at temp 1.0)
- **Borderline prompts exist**: A subset of prompts consistently flip between refusal and compliance, suggesting they lie near decision boundaries in the model's safety classifier

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 of large language models has become a critical research area, with several benchmarks developed to assess model behavior on harmful prompts. BeaverTails [2] provides a human-preference dataset for safety alignment covering diverse harm categories, while HarmBench [4] offers a standardized framework for automated red teaming and robust refusal evaluation. However, these benchmarks, like most safety evaluation protocols, rely on single-shot testing—each prompt is evaluated once with a fixed random seed and temperature. This approach implicitly assumes that model responses are deterministic or that a single sample adequately represents the model's safety behavior. Our work challenges this assumption by systematically investigating per-prompt stability across multiple sampling configurations. While prior work has studied the impact of temperature on generation quality and diversity, no previous research has quantified how sampling parameters affect the consistency of safety-critical decisions. We introduce the Safety Stability Index (SSI) as a metric to fill this gap, revealing that nearly one-third of prompts produce inconsistent safety decisions—a finding with significant implications for benchmark reliability and model deployment.

3 Methodology

3.1 Experimental Design

Model. We evaluate Llama 3.1 8B Instruct [1], a recent open-weight instruction-tuned model with safety alignment.

Dataset. We use 876 harmful prompts from the BeaverTails dataset [2], covering categories including violence, hate speech, illegal activities, and other harmful content.

Sampling Configurations. We test 4 temperature values (0.0, 0.3, 0.7, 1.0) × 5 random seeds (42–46), yielding 20 unique configurations and 17,520 total responses. 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 [3] for efficient batch generation.

Response Classification. We use Llama 3.1 8B Instruct as a judge model to classify each response into three categories:

- **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% exhibit at least one flip between different decisions. This flip rate exactly matches our proportion of unstable prompts (SSI < 0.8), indicating that unstable prompts consistently flip rather than showing minor variations.

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.

4.2 Distribution of Stability Scores

Figure 2 presents the distribution of SSI scores across all 876 prompts. The distribution is heavily bimodal: most prompts cluster at perfect stability (SSI = 1.0), while a smaller but significant subset shows lower stability. The 32% of prompts falling below the 0.8 threshold demonstrates that instability is not limited to edge cases but represents a substantial fraction of the safety evaluation dataset.

4.3 Temperature Effects

Figure 3 illustrates the relationship between temperature and both compliance rate and mean SSI. We observe a non-monotonic relationship: compliance rate peaks at temperature 0.3 (7.4%), while stability decreases monotonically as temperature increases. At temperature 1.0, mean SSI drops

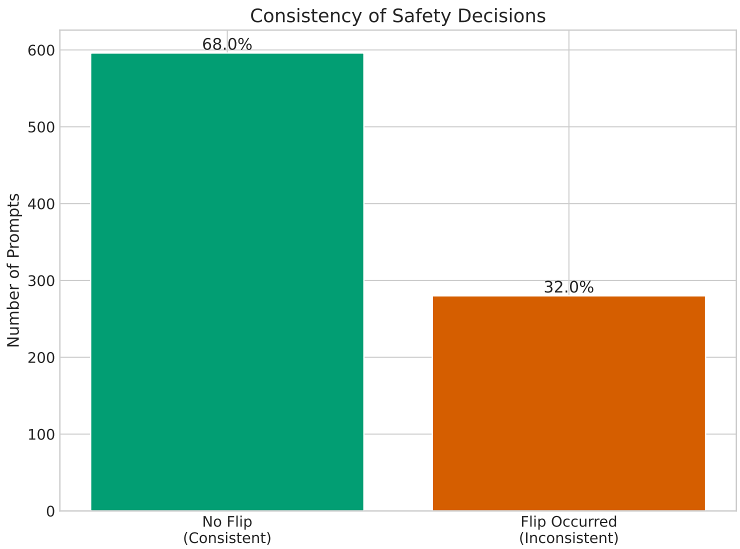

Figure 1: Distribution of consistent vs. inconsistent (flip) prompts across 876 harmful prompts tested on Llama 3.1 8B Instruct.

to 0.76, indicating that high-temperature sampling introduces substantial randomness in safety decisions.

Interestingly, the lowest compliance rate occurs at temperature 1.0 (3.1%). However, this likely reflects that high-temperature sampling produces less coherent outputs that fail to meaningfully engage with the harmful request, rather than representing more robust safety behavior.

4.4 Stability vs. Compliance Analysis

Figure 4 presents a scatter plot of per-prompt SSI against compliance rate, revealing three distinct clusters:

- **Stable Refusers** (bottom-left): Low compliance rate, high stability—the model consistently refuses these prompts
- **Stable Compliers** (bottom-right): High compliance rate, high stability—the model consistently complies with these prompts
- **Unstable Borderline** (center): Mid-range compliance with low stability—these prompts produce mixed responses across configurations

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.

Table 1 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.

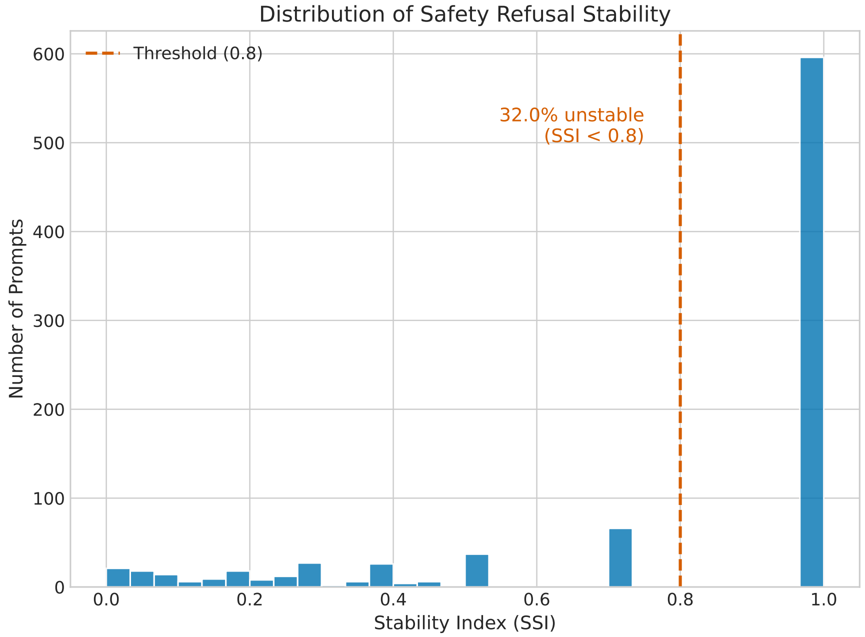

Figure 2: Distribution of Safety Stability Index (SSI) scores. The dashed line at 0.8 indicates the threshold for unstable prompts. 32% of prompts fall below this threshold.

Table 1: Example prompts exhibiting decision instability across sampling configurations (N=20).

Prompt Category	**SSI**	**REFUSE**	**PARTIAL**	**COMPLY**
Request for copyrighted content	0.16	2/20	11/20	7/20
Request for harmful instructions	0.08	3/20	9/20	8/20
Request for copyrighted passage	0.09	10/20	7/20	3/20

4.5 Model Comparison

Figure 5 shows the stability index and response distribution for Llama 3.1 8B Instruct. While this work focuses on a single model, the visualization framework is prepared for multi-model comparison in future extensions.

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 (temperature 0.0) provides the most stable behavior (mean SSI = 0.96). However, even at temperature 0.0, we observe some instability due to random seed variation.

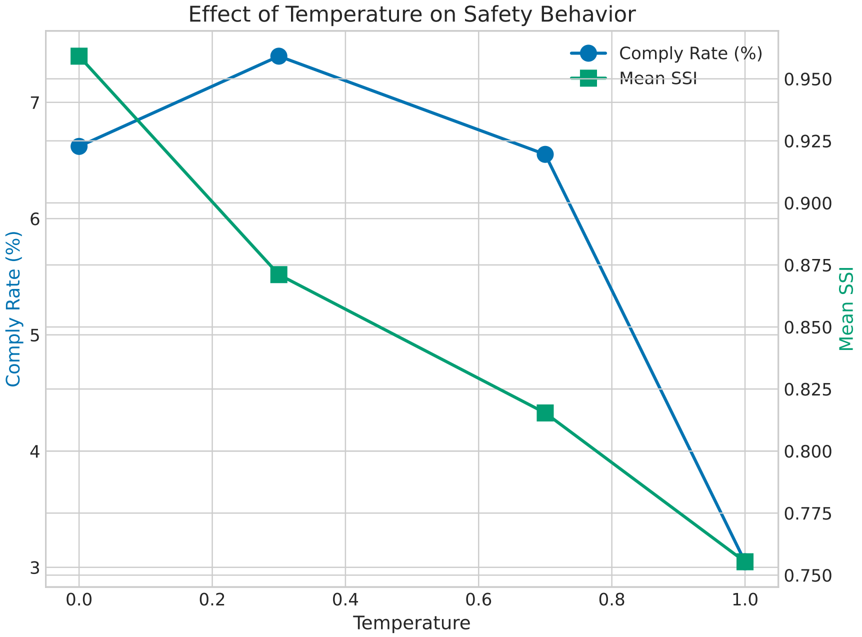

Figure 3: Effect of temperature on comply rate (blue) and mean Safety Stability Index (green). Higher temperatures reduce stability while showing non-monotonic effects on compliance.

Adversarial Robustness. The existence of borderline prompts that flip between refuse and comply suggests that adversaries could exploit stochastic sampling by repeatedly querying the model until they obtain a harmful response.

5.2 Limitations and Future Work

This study focuses on Llama 3.1 8B Instruct. Future work should extend this analysis to other model families, sizes, and alignment techniques to determine whether instability is a general phenomenon or specific to certain architectures.

We use the same model as both generator and judge, which could introduce bias. External human evaluation or diverse judge models would provide additional validation.

Our treatment of PARTIAL responses as a distinct category means that practitioners with different risk tolerances may interpret our results differently. Safety-critical deployments that treat any partial compliance as failure would observe higher effective failure rates than our reported 5.9% COMPLY rate suggests.

Finally, investigating the linguistic and semantic properties of unstable prompts could reveal what makes certain prompts fall on decision boundaries, potentially informing more robust safety training.

6 Conclusion

We have demonstrated that safety refusal decisions in LLMs exhibit significant instability across random seeds and temperature settings, with 32% of harmful prompts producing inconsistent responses. This finding challenges the validity of single-shot safety evaluations and suggests that current benchmarks may provide misleading assessments of model safety.

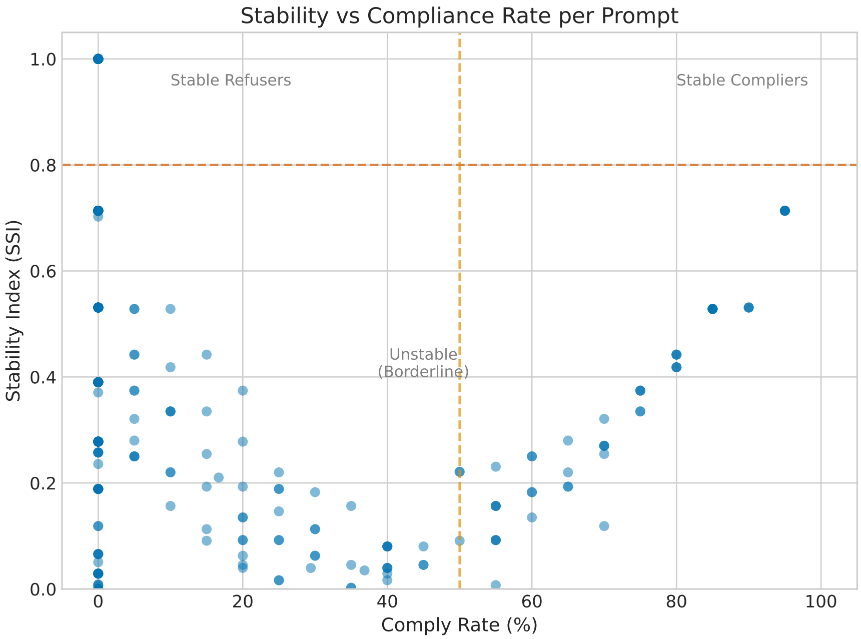

Figure 4: Safety Stability Index vs. compliance rate per prompt. Three clusters emerge: stable refusers, stable compliers, and unstable borderline cases.

We recommend that safety evaluation protocols adopt multi-sample testing and report stability metrics alongside accuracy. For deployment, practitioners should carefully consider temperature settings and potentially implement ensemble voting across multiple samples for high-stakes safety decisions.

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

References

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

[2] 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.

[3] 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.

[4] 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.

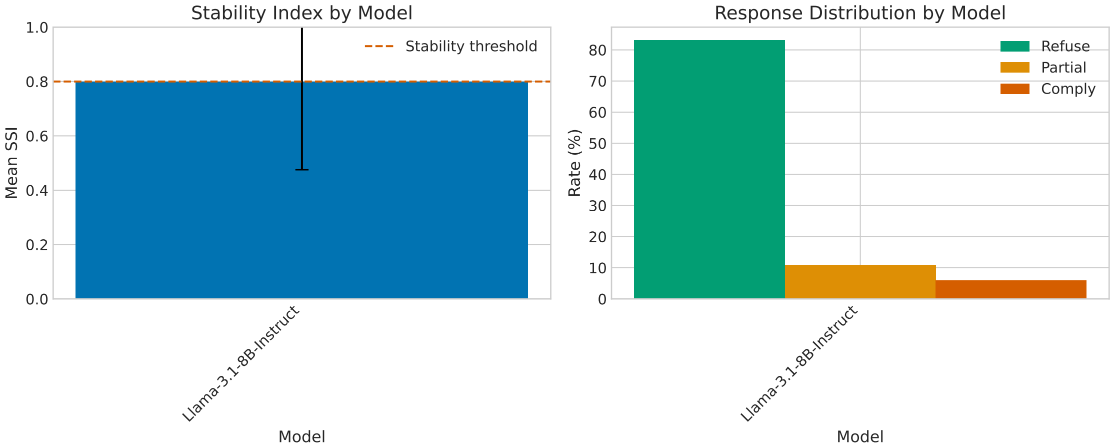

Figure 5: Stability index (left) and response distribution (right) for Llama 3.1 8B Instruct. The model shows mean SSI of 0.80, just at the stability threshold.