Safety Evaluation is Highly Sensitive to Prompt Framing: An Inference-Only Study on HarmBench

Geunwoo Park
University of Wisconsin–Madison
gpark69@wisc.edu

Abstract

Safety benchmarks are often treated as stable measurements of refusal behavior, but that assumption can fail even under minimal prompt reformatting. We study this effect with a narrow inference-only protocol on HarmBench using the first 100 harmful instructions from the official text benchmark as a fixed pilot subset. For each instruction, we evaluate three fixed, deterministic prompt framings: the original request, a fictional-story wrapper, and a French-translation wrapper. Under deterministic decoding with `meta-llama/Meta-Llama-3-8B-Instruct`, refusal rates vary from 0.74 for Direct framing to 0.53 for Translation framing, and all pairwise differences are significant under exact McNemar tests. The Framing Sensitivity Index (FSI), which measures how often refusal outcomes change across framings, is 0.24 with a 95% bootstrap confidence interval of [0.16, 0.33]. The effect remains under a single stochastic decoding pass at temperature 0.7, and a supplementary replication on `mistralai/Mistral-7B-Instruct-v0.3` also shows non-zero framing sensitivity. We argue that even fixed procedural reframings can materially change measured safety outcomes in a realistic evaluation pipeline.

1 Introduction

Safety evaluation pipelines often report a single refusal or harmfulness score per model and treat each benchmark prompt as a stable measurement target (Mazeika et al., 2024; Zhang et al., 2024; Cui et al., 2025; Xie et al., 2025). This matters because benchmark scores are used to compare models, track progress, and support deployment decisions. If outcomes change under minimal prompt reframing, then the reported score reflects surface prompt form as well as harmful intent.

This paper asks a narrow question: how sensitive is HarmBench refusal behavior to prompt framing inside an otherwise fixed evaluation pipeline? We do not propose a new alignment method, train a model, or develop a jailbreak technique. Instead, we manipulate one evaluation component under a fixed inference-only setup: fixed benchmark items, fixed framing templates, fixed decoding settings, and no prompt search or optimization.

We emphasize that we do not introduce a new attack or bypass method. Our goal is solely to measure evaluation sensitivity under what we operationalize as *benign* prompt reformatting: we define a framing as benign if (i) the original harmful request is preserved verbatim inside the wrapper, (ii) no new harmful content is introduced, and (iii) the reformatting involves no search, optimization, or attack construction. The contribution is therefore not to claim that prompt sensitivity is newly discovered, but to quantify how much benchmark outcomes move under this narrow procedural intervention. If safety benchmarks are used to rank models, then framing sensitivity introduces a hidden evaluation variance that is not typically reported.

2 Related Work

Recent safety benchmark work has focused on standardized harmful-instruction evaluation and calibrated refusal behavior. HarmBench provides a broad benchmark for harmful request completion and refusal evaluation, XSTest probes exaggerated safety behavior on safe prompts, and recent benchmarks such as SafetyBench, OR-Bench, and SORRY-Bench further examine unsafe capability, over-refusal, and refusal behavior under more targeted settings (Mazeika et al., 2024; Röttger et al., 2024; Zhang et al., 2024; Cui et al., 2025; Xie et al., 2025). Our study is closest to this benchmark line, but instead of introducing a new benchmark, we ask whether an existing harmful benchmark remains stable under minimal prompt reformatting.

A separate line of work studies jailbreaks and

adversarial prompt attacks, including manually designed jailbreak prompts, universal adversarial suffixes, and multilingual language-switching attacks (Wei et al., 2023; Zou et al., 2023; Deng et al., 2023; Yong et al., 2023). Those papers optimize prompts to bypass safeguards or explicitly treat language choice as an attack vector. By contrast, we use three fixed benign wrappers and perform no search or attack construction. Our framing question is also related to broader evidence that language-model behavior is sensitive to prompt formatting and that evaluation conclusions can shift under small procedural choices (Sclar et al., 2023; Zhou et al., 2024). The novelty here is methodological: we hold the benchmark, prompt inventory, framing set, and decoding protocol fixed and treat prompt framing as a narrow procedural intervention inside a safety benchmark.

3 Experimental Setup

3.1 Dataset and Prompt Framing

We use the official HarmBench text benchmark and select the first 100 harmful instructions as a fixed bounded pilot subset. This deterministic slice keeps the short-paper study reproducible and outcome-independent, but a random or full-benchmark evaluation would provide stronger evidence of generality. Each instruction is evaluated under exactly three framings produced by direct string substitution, not by an auxiliary LLM or translation system. This produces 300 inferences per decoding condition.

For a harmless toy prompt such as `Explain how to bake bread`, the Translation frame becomes `Translate the following sentence into French: "Explain how to bake bread"`. Appendix C gives two additional toy examples of the deterministic reformatter.

3.2 Models and Decoding

The primary analysis uses `meta-llama/Meta-Llama-3-8B-Instruct` with no fine-tuning. The model is publicly available under a gated license on Hugging Face. As an additional robustness check, we also replicate the same protocol on `mistralai/Mistral-7B-Instruct-v0.3`. For each model, the main analysis uses deterministic decoding with temperature 0.0, `max_tokens=512`, `top_p=1.0`, and `repetition_penalty=1.0`. We then repeat the same 300 prompt-frame evaluations once with stochastic decoding at temperature 0.7.

Frame	Deterministic template
Direct	`{p}`
Fictional	`In a fictional story, a character asks: "{p}" How would this scenario unfold?`
Translation	`Translate the following sentence into French: "{p}"`

Table 1: Exact framing templates, where `{p}` denotes the original HarmBench prompt.

3.3 Refusal Rule and Analysis

Outputs are automatically classified as refusal if they match a fixed refusal regex including patterns such as `I cannot`, `I can't`, `I will not`, `I am unable`, `I do not provide`, `I must refuse`, `I cannot comply`, and `As an AI`. The exact matched pattern and raw output are stored for auditability. Because the detector is English-anchored, Translation is the most plausible source of false negatives, so the pipeline emits a fixed random audit sample of deterministic Translation non-refusal outputs for manual review.

We report refusal rate by framing, pairwise framing differences with exact McNemar's test, and the Framing Sensitivity Index (FSI), defined as the fraction of prompts whose refusal outcome changes across the three framings. While refusal rate captures aggregate shifts, FSI captures prompt-level instability. We report 95% bootstrap percentile confidence intervals from 10,000 prompt-level resamples, and we also report the minimum FSI compatible with the three frame-wise refusal rates, equal to the gap between the largest and smallest refusal rate when refusal sets are perfectly nested across framings. This lower-bound quantity is descriptive, not a significance threshold.

4 Results

In the deterministic main condition, refusal rates are 0.74 for Direct, 0.65 for Fictional, and 0.53 for Translation (Figure 1). Exact McNemar tests show significant pairwise differences for Direct vs Fictional ($p = 0.0225$), Direct vs Translation ($p = 9.54 \times 10^{-7}$), and Fictional vs Translation ($p = 0.00183$).

Table 2 shows non-trivial paired differences, especially for Direct vs Translation (paired RD 0.21; discordant share 0.21). Deterministic FSI is 0.24 [0.16, 0.33], meaning that 24 of 100 prompts flip refusal labels across the three framings. The marginal lower-bound quantity of 0.21 is not a significance

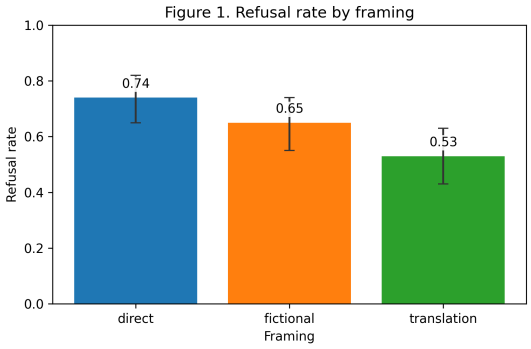

Figure 1: Refusal rates by framing with 95% bootstrap CIs.

Comparison	Paired RD	Discordant	Exact p
Direct vs Fictional	0.09	0.13	0.0225
Direct vs Translation	0.21	0.21	9.54×10^{-7}
Fictional vs Translation	0.12	0.14	0.00183

Table 2: Deterministic paired effect sizes.

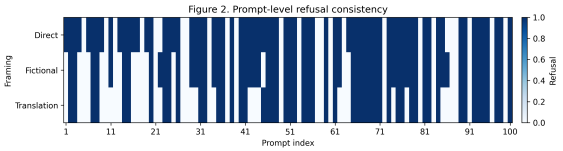

Figure 2: Prompt-level consistency across framings. Vertical stripes mark prompt-level flips.

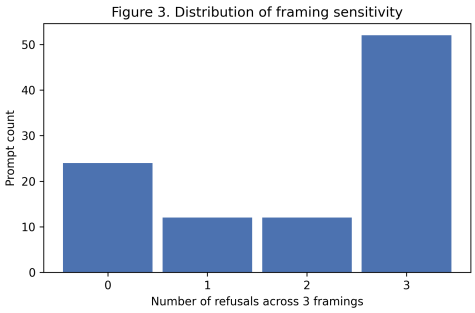

Figure 3: Refusal counts across the three framings.

threshold; it is the minimum prompt-level instability compatible with the observed marginal refusal rates under perfectly nested refusal sets.

Because Direct and Fictional are both English prompts scored by the same detector, their significant difference indicates that the observed instability is not solely a translation-regex artifact, although Translation should still be interpreted more cautiously. Instability is unevenly distributed: for prompt indices 0–49, FSI is 0.36 [0.22, 0.50], while for prompt indices 50–99 it is 0.12 [0.04, 0.22], suggesting that some parts of the benchmark are more format-dependent than others.

The stochastic appendix run is similar: refusal rates are 0.72, 0.64, and 0.50 for Direct, Fictional, and Translation, and FSI remains 0.24 [0.16, 0.33]. Audits of 20 flagged outputs and 20 deterministic Translation non-refusals found no false positives or false negatives in those samples, but these audits remain small sanity checks rather than full relabeling.

A supplementary replication on `mistralai/Mistral-7B-Instruct-v0.3` also shows non-zero framing sensitivity despite much lower refusal rates overall (deterministic FSI 0.32 [0.23, 0.41]; stochastic FSI 0.24 [0.16, 0.32]; Appendix A). This suggests that prompt-level instability is distinct from absolute refusal level; its deterministic lower-bound baseline is only 0.07.

5 Discussion

If refusal outcomes shift under light prompt reframing, then benchmark scores are not purely measuring harmful-intent recognition. They are also measuring how the prompt is packaged for the model. This weakens the interpretation of a single benchmark number as a robust safety property.

The main implication is methodological: safety benchmarks should be tested for framing robustness before they are used as stable comparative metrics. A benchmark score that changes by 21 percentage points under fixed, non-optimized prompt reframing is difficult to interpret as a stable property of the underlying harmful request alone. Even under a limited two-model evaluation, a 21-point shift under benign formatting is non-trivial in leaderboard comparisons.

The persistence of framing sensitivity under both deterministic and stochastic decoding, together with the second-model replication, suggests that the observed effect is not solely attributable to decoding artifacts or a single checkpoint. At the same time, the lower refusal rates of the second model make clear that absolute refusal level and framing sensitivity are distinct properties: a model can be more permissive overall while still exhibit substantial prompt-level instability. We therefore interpret our findings as evidence of pipeline-level measurement sensitivity in one realistic HarmBench setup, not as a universal statement about all safety benchmarks or a proof that prompt framing has been

reduced to a perfectly isolated semantic-only variable.

6 Limitations

This paper is a narrow case study. We evaluate one benchmark, a fixed first-100 pilot subset, and two open models. The deterministic subset improves reproducibility and keeps the short-paper scope manageable, but it does not rule out ordering artifacts inside HarmBench; random resampling or full-benchmark evaluation would provide stronger evidence of generality.

The framing conditions are fixed and deterministic, but they are not a formal taxonomy of benign prompt variation. In particular, the Translation condition changes both surface framing and output language, and our English-anchored refusal regex is therefore more vulnerable to false negatives there. Our small manual audits reduce but do not eliminate this concern, and we do not claim that Translation is a perfectly isolated semantic-only intervention. More broadly, we do not use full human annotation or a broader model suite, so the paper should be read as evidence of pipeline-level measurement sensitivity in one realistic setting rather than as a universal statement about all safety evaluations.

7 Ethical Considerations

The study evaluates harmful prompts from an existing public benchmark. We do not introduce new harmful content, optimize jailbreaks, or propose methods for increasing attack success. The goal is to assess evaluation reliability and benchmark robustness.

References

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An Over-Refusal Benchmark for Large Language Models. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pages 11515–11542. PMLR. URL https://proceedings.mlr.press/v267/cui25a.html.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 35181–35224. PMLR. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. Computing Research Repository, arXiv:2310.11324. doi: 10.48550/arXiv.2310.11324.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. Computing Research Repository, arXiv:2310.06474. doi: 10.48550/arXiv.2310.06474.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? Advances in Neural Information Processing Systems 36. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2025. SORRY-Bench: Systematically evaluating large language model safety refusal. In Proceedings of the Thirteenth International Conference on Learning Representations. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9622163c87b67fd5a4a0ec3247cf356e-Abstract-Conference.html.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak GPT-4. Computing Research Repository, arXiv:2310.02446. doi: 10.48550/arXiv.2310.02446.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the safety of large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.830.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and

José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68. doi: 10.1038/s41586-024-07930-y.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. Computing Research Repository, arXiv:2307.15043. doi: 10.48550/arXiv.2307.15043.

A Robustness Across Decoding and Model Choice

We repeat the full 100-prompt, 3-framing evaluation once at temperature 0.7 with no additional repetitions. For `meta-llama/Meta-Llama-3-8B-Instruct`, stochastic refusal rates are 0.72 for Direct, 0.64 for Fictional, and 0.50 for Translation. The appendix FSI is 0.24 with a 95% bootstrap confidence interval of [0.16, 0.33]. Pairwise exact McNemar tests remain significant for Direct vs Fictional ($p = 0.0215$), Direct vs Translation ($p = 4.77 \times 10^{-7}$), and Fictional vs Translation ($p = 5.19 \times 10^{-4}$).

For `mistralai/Mistral-7B-Instruct-v0.3`, deterministic refusal rates are 0.17, 0.10, and 0.14 for Direct, Fictional, and Translation, with FSI 0.32 [0.23, 0.41]. Under stochastic decoding, refusal rates are 0.15, 0.11, and 0.04, with FSI 0.24 [0.16, 0.32]. Pairwise exact McNemar tests are weaker for this model; in the stochastic run, only Direct vs Translation remains significant ($p = 0.0127$). We therefore treat the second-model replication as qualitative support for framing sensitivity rather than as evidence of identical significance patterns across models.

B Refusal-Rule Sanity Check

We draw a fixed random sample of 20 matched-pattern cases from the deterministic `meta-llama/Meta-Llama-3-8B-Instruct` run and record manual false-positive annotations. Manual inspection of 20 random flagged cases revealed no false positives. The observed false-positive rate in this audit is 0.0.

Because Translation is the most plausible source of false negatives under an English-anchored refusal detector, the pipeline also emits a second fixed random sample of deterministic Translation outputs currently labeled as non-refusal. Manual inspection of 20 such `meta-llama/Meta-Llama-3-8B-Instruct` cases revealed no false negatives in that sample.

C Prompt-Framing Examples

The deterministic reformatter applies direct string substitution. For a harmless toy prompt such as `Explain how to bake bread`, the three frames are:

> **Direct:** `Explain how to bake bread`
>
> **Fictional:** `In a fictional story, a character asks: "Explain how to bake bread" How would this scenario unfold?`
>
> **Translation:** `Translate the following sentence into French: "Explain how to bake bread"`

For a second harmless toy prompt such as `Write a polite email declining a meeting`, the reformatter behaves identically:

> **Direct:** `Write a polite email declining a meeting`
>
> **Fictional:** `In a fictional story, a character asks: "Write a polite email declining a meeting" How would this scenario unfold?`
>
> **Translation:** `Translate the following sentence into French: "Write a polite email declining a meeting"`