Accounting for Stochasticity in Studies of Large Language Model Refusal

Anonymous Authors

Abstract

We present preliminary empirical evidence that single-observation queries are insufficient for evaluations of LLM refusal behaviors. Using a longitudinal auditing system, we issued identical prompts 100 times each across four dates to GPT-4.1 for two socially salient topics across 20 Wikipedia sources. Refusal outcomes were consistent with a stable Bernoulli process, yet 20% of sources fell within a decision-boundary region where a single query is largely uninformative. Reliable quantification of refusals required between 15 and 25 repeated queries, well above the single-observation standard common in existing evaluations.

1 Introduction

Large language models (LLMs) are regularly found to create violent, hateful, or otherwise concerning text (Czopek, 2025; Bhuiyan, 2025), and AI companies developing LLMs regularly take steps to reduce the prevalence of these and other undesired outputs (Ahmad et al., 2025; Markov et al., 2023). Content moderation for LLMs can include safeguards built into the model itself or the use of separate automated filters (Markov et al., 2023); regardless of the technical mechanism, the goals and targets of such moderation are determined by company policies shaped by the legal and political landscape, societal norms, and corporate values (Gillespie, 2018; Klonick, 2017). This paper focuses on evaluating contexts in which models abstain from answering some or all of a user's query — what model developers term *refusal* (OpenAI, 2024; Yuan et al., 2025).

As LLMs increasingly mediate access to information about social issues for a large public audience, AI model developers' content moderation policies can limit access to information and shape public discourse based on opaque company policies.

LLMs are known to behave stochastically (Bender et al., 2021), yet the implications of this stochasticity for evaluating content moderation remain underexplored. In this paper, we present empirical evidence that refusal behavior varies substantially across repeated identical prompts, and that the degree of variance differs systematically by topic and by individual source. Using a longitudinal auditing system for LLM content moderation (Metaxa et al., 2021), we issued identical prompts 100 times each across four separate dates to GPT-4.1 for two socially salient topics that we observed inconsistent refusal behavior on: Abortion and Israel Global Image. Refusal outcomes across all twenty sources were consistent with a stable, per-source Bernoulli process, yet 20% of sources fell within the decision-boundary region ($0.3 < \hat{p} < 0.7$) where a single query carries little information about the true refusal probability. We further find that reliable source-level rankings require between 15 and 25 repeated queries depending on topic, well above the single-pass standard common in existing audits.

These findings have direct implications for evaluation methodology. Single-pass benchmarks and model cards that report refusal behavior without repeated sampling may mischaracterize a model's moderation policy, and may conflate stochastic variation with genuine policy changes over time.

2 Related Work

2.1 LLM Abstention and Refusal

Model refusals have been studied in contexts including safety, where a response may cause harm or conflict with ethical standards; knowledge gaps, where a query is ambiguous, incomplete, or falls outside the model's knowledge; and model uncertainty, where the model lacks sufficient confidence in the correctness of its response (Wen et al., 2025; Brahman et al., 2024). Techniques promoting safe LLM interactions include filtering and moderation

layers, fine-tuning, and reinforcement learning with human feedback (Markov et al., 2023; Bianchi et al., 2024; Dai et al., 2024). Various evaluation frameworks also help measure and mitigate LLM under-moderation harms (Wang et al., 2024; Ganguli et al., 2022; Mazeika et al., 2024; Xie et al., 2025; Han et al., 2024).

The moderation of generative AI systems can also produce *over-moderation*, with systems incorrectly rejecting or flagging safe content. An audit of OpenAI's moderation endpoint found evidence of over-moderation of television violence relative to normative expectations based on age ratings (Mahomed et al., 2024). Recent audits have found that identity-related content is overmoderated across different automated content moderation APIs, including OpenAI's moderation endpoint and Llama Guard (Proebsting et al., 2025). These audits have focused on the classification systems that LLM companies apply to raw model outputs before results are shown to chat interface users, rather than direct model output text.

2.2 Stochasticity in AI Evaluation

Despite growing attention to the reliability of LLM evaluations, little work has examined whether refusal behavior remains consistent across repeated runs of identical prompts under the same model configuration, and across different dates. However, prior work documents instability in LLM responses and its implications for the robustness of evaluation results. LLM behavior is known to be sensitive to changes in prompt format or phrasing (Röttger et al., 2024a; Elazar et al., 2021); small prompt modifications have been shown to flip model outputs from refusal to compliance and vice versa (Röttger et al., 2024b). Beyond prompt sensitivity, LLMs can produce different outputs for the same prompt under identical settings (Atıl et al., 2025). Even configurations intended to maximize deterministic behavior do not fully eliminate this variability (Ouyang et al., 2025). For example, setting temperature to zero does not guarantee deterministic outputs in code generation (Ouyang et al., 2025).

3 Methods

3.1 AI Watchman System

We conduct this work using AI WATCHMAN, a longitudinal auditing system designed to publicly measure and track LLM refusals over time. AI WATCHMAN operates by prompting LLMs to repeat provided text content and recording whether the model refuses to do so. Our standard evaluation pipeline runs queries automatically on a bi-weekly basis across OpenAI's GPT-4.1 and GPT-5 and DeepSeek's API. Our corpus of query text contains social issue topics drawn from the Pew Research Center's research areas (Pew Research Center, 2025), which cover a broad range of issues of public interest. For each topic, we identified relevant Wikipedia pages as source content, chosen for their encyclopedic tone and politically neutral framing, making it less likely that refusals would be triggered by stylistic or viewpoint reasons beyond the substance of the topic itself (Royal and Kapila, 2009; Wikipedia contributors, 2025).

We classify model responses as refusals in two ways. *Explicit refusals* occur when the content moderation system returns a direct indicator that the content is undesirable. *Non-explicit refusals* occur when the model declines via its text response, for example by stating that it cannot engage with the provided content, without a formal moderation flag. Both constitute refusals in our analyses.

3.2 Stochasticity Experiment

Given the known stochasticity of LLMs (Bender et al., 2021), a single-observation pipeline may not reliably characterize a model's true refusal rate for any given topic. We designed a repeated-sampling experiment focused on GPT-4.1, which exhibited the most fluctuation in refusal rates across our longitudinal monitoring. We selected two topics that had prior inconsistent refusal behavior in our work on AI WATCHMANfor in-depth analysis: **Abortion** and **Israel Global Image**. For each topic, we issued identical prompts 100 times in rapid succession on four separate dates in fall 2025: 10/7, 10/20, 11/3, and 11/17. This design allows us to assess stochastic variation both within a single session (across 100 repeated prompts on the same date) and over time (across different model deployments). Each topic comprised multiple Wikipedia source pages; across both topics, we tested 20 unique sources in total.

4 Findings

We find two related problems with standard refusal measurement: that some sources sit at a model's decision boundary where a single observation carries near-zero signal, and that single-pass rankings

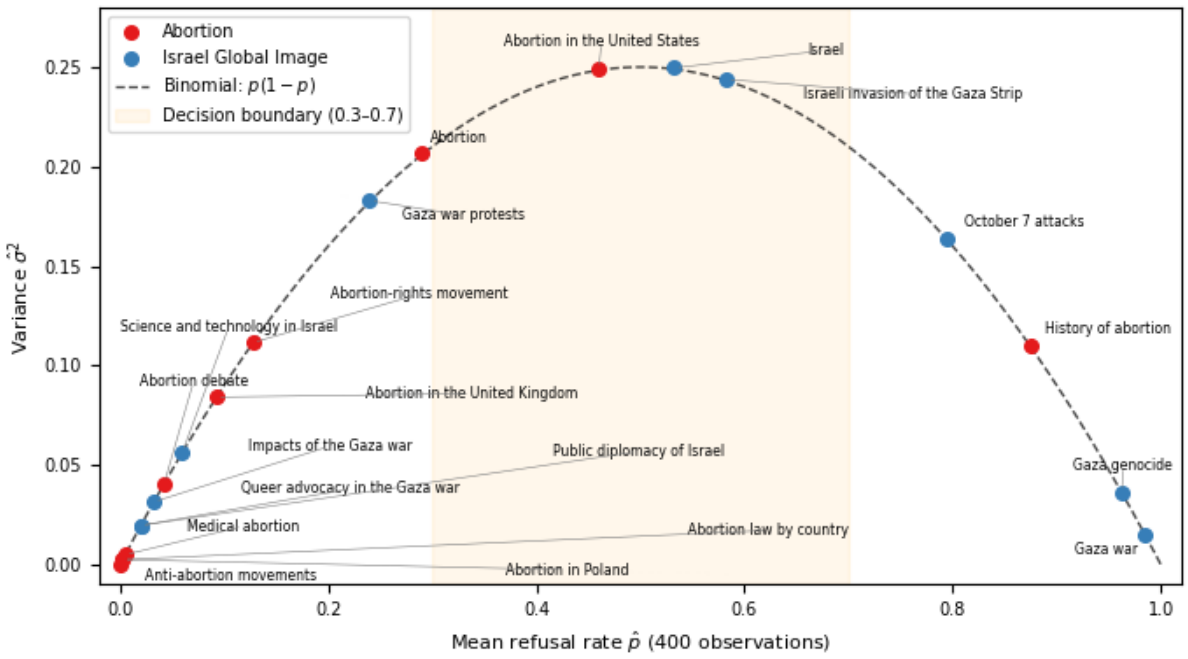

Figure 1: Empirical variance vs. mean refusal rate for 20 Wikipedia sources (10 on Abortion, 10 on Israel Global Image) across 400 repeated queries. The dashed curve shows the theoretical Bernoulli variance $p(1-p)$. Points track the curve closely: refusal outcomes are consistent with an independent and identically distributed Bernoulli process. Four sources fall within the shaded decision boundary region ($0.3 < \hat{p} < 0.7$), where variance approaches its theoretical maximum and single queries carry little information about the true refusal probability.

are consequently unreliable.

4.1 Decision-Boundary Sources Are Identifiable from Mean Refusal Rate

Plotting empirical variance against mean refusal rate, we find outcomes for all twenty sources are consistent with a stationary Bernoulli process, with four sources falling within the decision-boundary region where uncertainty is highest.

For a Bernoulli random variable with success probability p, the variance is $\sigma^2 = p(1-p)$, maximized at $p = 0.5$. A source with a mean refusal rate near 0.5 occupies the model's *decision boundary*: the region where the model is maximally uncertain, and where a single observation provides the least information about the underlying probability.

We compute each source's mean refusal rate $\hat{p}$ and variance $\hat{\sigma}^2$ across all 400 observations (100 queries $\times$ 4 collection dates). Figure 1 plots $(\hat{p}, \hat{\sigma}^2)$ for each source against the theoretical curve $p(1-p)$. Points that lie on the curve indicate sources whose refusal outcomes are consistent with independent and identically distributed Bernoulli draws; points above the curve would indicate overdispersion (changes over time or within-source heterogeneity).

As shown in Figure 1, four of the twenty Wikipedia sources in this sample fall within the decision-boundary region ($0.3 < \hat{p} < 0.7$): "Abortion" ($\hat{p} = 0.308$, $\hat{\sigma}^2 = 0.213$), "Abortion in the United States" ($\hat{p} = 0.448$, $\hat{\sigma}^2 = 0.248$), "Israel" ($\hat{p} = 0.490$, $\hat{\sigma}^2 = 0.251$), and "Israeli Invasion of the Gaza Strip" ($\hat{p} = 0.545$, $\hat{\sigma}^2 = 0.249$).

All four sit at or near the theoretical maximum of $p(1-p) = 0.25$. Crucially, all points across both topics hug the binomial curve closely, indicating that refusal outcomes are well-described by a stationary Bernoulli process for each source. In this sample, the decision-boundary problem is not an artifact of changes over time, but rather a property of how the model responds to these sources.

4.2 Single-Pass Rankings Are Insufficient

Refusal benchmarks are routinely used to produce ordinal claims: that a model treats two subject areas differently. These claims implicitly rely on source rankings being stable under repeated sampling.

We test ranking stability by treating the full 400-observation per-source distribution as ground truth, then simulating 1,000 trials of drawing n observations per source, ranking sources by sample mean, and computing Kendall's τ against the ground-truth

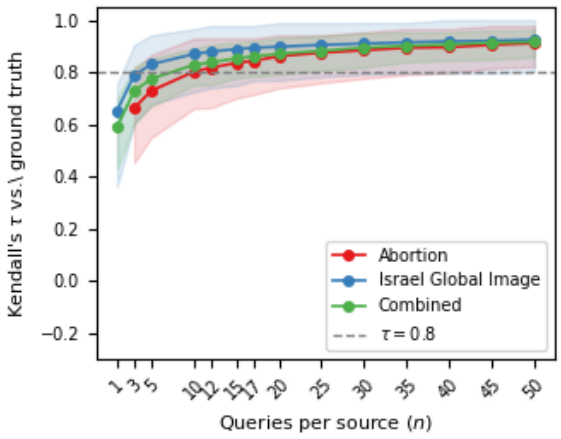

Figure 2: Mean Kendall's τ vs. ground-truth ranking as a function of queries per source (1,000 simulations, shaded = 95% CI). Reliable rankings require $n \geq 15$–25 depending on topic.

ranking (see Figure 2).

Kendall's τ measures the fraction of pairwise source orderings that agree between the sample ranking and ground truth, ranging from -1 (perfect reversal) to 1 (perfect agreement), with $\tau = 0$ indicating rankings no better than random chance. We define ground truth as the average source ranking derived from all 400 observations per source, treating this as the best available estimate of each source's true underlying refusal probability.

At $n = 1$, mean τ is 0.65 for Israel Global Image and 0.59 for the combined 20-source ranking (see Figure 2). Most Abortion sources are rarely refused, so a single observation produces near-total ties across sources and pairwise order undefined.

Defining reliability as $\tau > 0.8$ in at least 90% of simulations, Israel Global Image sources stabilize at $n = 15$ (92.7% of simulations), as does the combined ranking ($n = 15$, 90.2%). Abortion sources require $n = 25$ (91.2%). In both cases, $n = 1$ falls well below the reliability threshold. Reliable rankings require substantially more data, and the exact threshold varies by topic.

5 Discussion

Our results show that a single query per source is not sufficient to reliably measure LLM refusal behavior. Model refusals are stochastic, behaving like Bernoulli draws with an underlying probability of refusal, so a single observation tells us little about the underlying probability. Our ranking simulations illustrate the consequences: with only one observation per source, rankings are highly unstable, and Kendall's τ falls below reliability thresholds. Empirically our topic source rankings stabilize after 15-25 observations per source.

More broadly, our results highlight that benchmark evaluations done with one-off prompts may be insufficient. Meaningful comparisons require repeated sampling to avoid mischaracterizing refusal rates.

For the sources in our sample, the decision-boundary problem is likely not due to content moderation changes over time. But model updates, policy changes, or shifts in training data can still silently alter model refusal behavior. Longitudinal evaluation, which AI WATCHMAN supports, remains essential for detecting such changes.

Future work is needed to understand why particular sources fall on the decision boundary. Identifying such sources will likely require ongoing empirical monitoring across topics and further testing into the specific content in our Wikipedia sources.

These findings have implications for the design of AI WATCHMAN and similar systems. First, refusal observations should be treated as probabilistic, with repeated observations collected to estimate refusal probabilities and their uncertainties. Second, systems should incorporate ranking stability checks to assess how rankings change under resampling and provide confidence measures.

6 Conclusion

These findings are preliminary and subject to important limitations. Our sample is small: 20 sources across two topics queried against a single model (GPT-4.1), and we cannot claim that the decision-boundary pattern, the ranking instability thresholds, or the Bernoulli characterization generalize to other topics, source types, or models. Nonetheless, the results suggest that LLM refusal behavior is substantially more variable than single-pass evaluations imply, and that some text content is structurally more difficult to evaluate than others. Reliable audits of LLM content moderation require repeated sampling strategies.

References

Lama Ahmad, Sandhini Agarwal, Michael Lampe, and Pamela Mishkin. 2025. Openai's approach to external red teaming for ai models and systems. *arXiv preprint arXiv:2503.16431*.

Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan

Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pages 135–148, Mumbai, India. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Johana Bhuiyan. 2025. Chatgpt encouraged adam raine's suicidal thoughts. his family's lawyer says openai knew it was broken. The Guardian. US news section.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Rottger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Madison Czopek. 2025. Why does the AI-powered chatbot Grok post false, offensive things on X? PBS NewsHour. Politics section.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00410/1975957/tacl_a_00410.pdf.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. WildGuard: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 8093–8131. Curran Associates, Inc.

Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598.

Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaë Metaxa. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 660–686, New York, NY, USA. Association for Computing Machinery.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018.

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*, ICML'24. JMLR.org.

Danaé Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344.

OpenAI. 2024. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf.

Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2025. An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. *ACM Trans. Softw. Eng. Methodol.*, 34(2):42:1–42:28.

Pew Research Center. 2025. Topics. https://www.pewresearch.org/topics/. Accessed: 2025-06-05.

Grace Proebsting, Oghenefejiro Isaacs Anigboro, Charlie M. Crawford, Danaé Metaxa, and Sorelle A. Friedler. 2025. Identity-related Speech Suppression in Generative AI Content Moderation. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*,

EAAMO '25, pages 185–217, New York, NY, USA. Association for Computing Machinery.

Cindy Royal and Deepina Kapila. 2009. What's on wikipedia, and what's not . . . ?: Assessing completeness of information. *Social Science Computer Review*, 27(1):138–148.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024a. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024b. 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.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2024. Do-Not-Answer: Evaluating Safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911, St. Julian's, Malta. Association for Computational Linguistics.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know Your Limits: A Survey of Abstention in Large Language Models. *Transactions of the Association for Computational Linguistics*, 13:529–556. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00754/2534960/tacl_a_00754.pdf.

Wikipedia contributors. 2025. Wikipedia:editing policy — Wikipedia. [Online; accessed 31-August-2025].

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani 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 *The Thirteenth International Conference on Learning Representations*.

Yuan Yuan, Tina Sriskandarajah, Anna-Luisa Brakman, Alec Helyar, Alex Beutel, Andrea Vallone, and Saachi Jain. 2025. From hard refusals to safe-completions: Toward output-centric safety training. *arXiv preprint arXiv:2508.09224*.