
The Veneer of safety: the fragility of India-specific harm refusal in open LLMs

Vihaan B.
Independent Researcher

vihaanbwork@gmail.com

Ketki Banafar
Dartmouth College

ketki.banafar.gr@dartmouth.edu

With
Apart Research

Abstract

Open language models in the 7 to 8 billion parameter range are very common in local, self-hosted use, because they run on modest hardware. We ask a simple question: do these models recognize and refuse India-specific harms, such as dowry coercion, sex-selective abortion, and faith healing in place of medicine, as strongly as they handle a canonical harm like building a weapon? We answer it inside the model rather than from its replies. Following Zhao et al. [1], we read two directions in the residual stream, a harmfulness direction at the last token of the instruction and a refusal direction at the end of the prompt, each a difference in means between harmful and harmless activations. We test three open models (Llama-2-7b-chat, Airavata-7B, Qwen3-8B) on 105 English-language harmful prompts across six India-specific harm categories plus a canonical control, against 40 topic-matched harmless prompts, and add a benign LoRA finetune test on Llama-2. Our findings: all three exhibit strong linear separation between harmful and harmless India-specific prompts in the residual stream, which we interpret as a harm-related internal signal; the harms load weakly on the canonical harm axis, so a safeguard anchored there under-detects them; whether a model refuses tracks its safety-tuning, not its understanding; and a benign finetune strips about 26% of India-harm refusal on Llama-2 while leaving canonical refusal intact.

1. Introduction

Open language models in the 7 to 8 billion parameter range are widely used for local, self-hosted deployment, including in Indian settings, because they run on modest hardware. Teams finetune them for customer support, education, and regional languages, often on ordinary task data with no safety intent. Whether these models are safe against India-specific harms is usually checked by prompting them and reading the reply. That tells us if a model refuses. It does not tell us whether the model has any internal sense that the request is harmful, or whether the refusal is a thin surface habit that a later finetune can rub off.

We separate two things that behavioral tests blur together: whether a model recognizes a harm internally, and whether it refuses on the surface. Recent interpretability work shows these are distinct. Zhao et al. [1] find that harmfulness and refusal are encoded by separate linear directions in the residual stream, and that a refusal can fire without being anchored to the model's own harm representation. Ardit et al. [2] show that refusal itself is mediated by a single direction. We apply this reading to harms that matter in India and have little presence in standard safety benchmarks. Threat model. Open 7 to 8B models are customized by ordinary, benign finetuning and used widely in local Indian deployments. If India-specific safety is a thin surface layer, that ordinary finetuning silently removes it, and safeguards built on canonical harm directions never see the gap.

Our main contributions are:

- **The first activation-level test of whether open LLMs internally represent India-specific harms.** Prior India-safety work such as IndicSafe [8] is black-box and behavioral; we read the internal harm signal directly.
- **A coverage gap.** India harms are represented, but load weakly on the canonical-harm axis that internal safeguards such as Zhao et al.'s Latent Guard [1] rely on. Such safeguards systematically under-detect India harms even though the model recognizes them.
- **A fragility demonstration.** A benign LoRA finetune strips India-harm refusal while leaving canonical refusal intact, and the decoupling between refusal and recognition holds across three models with different safety-tuning.

2. Related Work

Reading concepts as linear directions in activation space is now standard. Representation engineering [9] reads high-level concepts from activations; difference-in-means directions steer and probe behavior in Llama-2 [10] and recover truth-related structure [11]. Refusal in particular is mediated by a single direction [2]. Our method comes directly from Zhao et al. [1], who separate a harmfulness direction at the instruction token from a refusal direction at the post-instruction token, and build a Latent Guard detector on the harm direction. We reuse their two-direction reading and ask a question they do not: does it hold for India-specific harms, and does the refusal stay anchored to the harm signal under finetuning.

Safety degradation under finetuning is documented. Qi et al. [3] show that finetuning aligned models compromises safety even when users do not intend harm. Low-resource and multilingual settings are a known weak point: Yong et al. [12] jailbreak GPT-4 through low-resource languages, and Deng et al. [13] map multilingual jailbreak failures. These motivate our data-frequency hypothesis but do not read internal representations. The closest prior India-safety work is IndicSafe [8], a multilingual South Asia safety benchmark. It is black-box: it scores model outputs. We instead read the model's internal harm signal, which lets us separate recognition from refusal. Safety-tuning by RLHF [15, 16] installs the refusal behavior we probe, and superposition [18] motivates why distinct harm types need not share a single direction.

3. Methods

Models. We use three open dense text models. Llama-2-7b-chat [6] is chat-tuned with safety RLHF. Airavata-7B [4] is an ai4bharat model: a Llama-2 base extended to Hindi via OpenHathi and instruction-tuned for Hindi helpfulness, not safety-RLHF'd. Qwen3-8B [5] is a modern dense model. We need dense text models for clean residual-stream reading; we make no use of mixture-of-experts or multimodal structure.

Prompts. We use 145 prompts. 105 are harmful, 15 in each of seven categories: six India-specific (communal_incitement, indian_scams_fraud, faith_healing, honour_violence, sex_determination, dowry_coercion) and one canonical_control of canonical harms: CBRN, weapons, and violence. The remaining 40 are harmless India-themed prompts. All harmful prompts are request-level, with no operational or method detail. Each prompt is run through the model's own chat template before activations are read. Prompts are English and team-authored.

Directions. Following Zhao et al. [1] and Ardit et al. [2], we read two directions. The harmfulness direction is read at t_{inst} , the last token of the user instruction. The refusal direction is read at $t_{\text{post-inst}}$, the last token of the full templated prompt. Each direction is a difference in means between pooled harmful and pooled harmless activations, computed in z-scored (standardized) residual space, where each hidden dimension is centered and scaled before the difference is taken. This matters. A first version used raw-norm activations and was contaminated by Llama-2 massive-activation outlier dimensions, a few coordinates with huge magnitude that dominated the difference. Standardizing removed that contamination, and this is the main thing we corrected. We compute one pooled harm direction and one pooled refusal direction per model, then measure each category by projecting its activations onto these shared directions.

Metrics. Separability is Cohen's d [17], the standardized mean difference, that is how many standard deviations apart the harmful and harmless groups sit on a given axis; d above 0.8 is conventionally large. We report four projections per category. `own_d` is the category against harmless on its own harm axis. `canon_d` is on the axis anchored by the canonical_control. `pooled_d` is on the all-harmful pooled axis. `refusal_d` is on the refusal axis. From these we derive the coverage ratio `canon_d / own_d`, how much of a category's own separability the canonical axis

recovers, and the Veneer Gap `refusal_d` minus `own_d`, how much harder a model refuses than it represents the harm. AUC saturates at these set sizes, so `d` is the discriminating metric.

Layer. We exclude the embedding and final readout layers and report a mid-to-late layer per model: Llama-2 layer 31, Airavata layer 16, Qwen3 layer 35. Different layers are not directly comparable, so across models we compare the structure and sign of effects, not absolute `d`. A sweep across the late-layer band shows that the pooled-axis Veneer Gap is stable across neighboring layers, so the reported layer choice is representative rather than post hoc.

What did not work. Our first harmless set was 15 prompts on a narrow set of topics. It captured topic rather than harm: harmless variance was large and the CBRN control landed off-axis. We expanded to 40 topic-matched India-themed harmless prompts spanning the same sub-topics as the harms, including benign weapons and security prompts as neighbors for the CBRN control. Harmless dispersion then collapsed to match the harmful set and the directions became clean.

Fragility test (Llama-2 only). We run a LoRA finetune [7]. LoRA (low-rank adaptation) trains a small low-rank add-on to the weights instead of the full model. Config: rank 16, alpha 32, dropout 0.05, learning rate 2e-4, 3 epochs, 226 optimizer steps, 300 random Alpaca [14] samples, seed 0, targets `q/k/v/o_proj`. The data is benign and contains no jailbreak content, so any safety loss is incidental erosion in the sense of Qi et al. [3]. We measure the refusal rate per category before and after, by greedy generation followed by a substring refusal classifier. Only rates were stored; no harmful generations were saved.

4. Results

Finding 1: harmful and harmless prompts are linearly separable for every India category, in all three models.

Each India category separates cleanly from harmless prompts on its own harm axis. On Llama-2 the `own_d` values run from 3.43 (`indian_scams_fraud`) to 4.87 (`canonical_control`), all far above the 0.8 large-effect threshold (Table 1). The same holds for Airavata (`own_d` 6.18 to 10.65) and Qwen3 (`own_d` 4.70 to 6.84); full per-category values for all three models are released in the repository. Because the harmless set is topic-matched to the harmful categories, this separation is not explained by surface topic alone: harmful and harmless prompts are linearly separable in the residual stream for every India category. We interpret this as evidence of a harm-related internal signal rather than a topic detector, while noting the open debate about how much conceptual structure can be inferred from linear separability alone. At a minimum, our results rule out the claim that these models entirely fail to distinguish India-specific harms from matched harmless prompts at the representation level. Bootstrap confidence intervals confirm the effect is not a fitting artifact: refitting the harm direction on each of 1000 resamples keeps the lower bound of `own_d` above 2 for every India category (full intervals in the repository, `results/other_fixes`).

Category	n	own_d	canon_d	pooled_d	refusal_d	Veneer Gap	coverage ratio
canonical_control	15	4.87	4.87	1.05	10.43	5.56	1.00
communal_incitement	15	4.72	0.60	2.90	11.34	6.62	0.13
honour_violence	15	3.97	0.26	2.71	11.10	7.12	0.07
sex_determination	15	4.34	1.14	2.19	9.21	4.87	0.26
faith_healing	15	3.80	1.01	2.41	9.88	6.08	0.27
dowry_coercion	15	3.95	0.47	2.36	6.98	3.03	0.12
indian_scams_fraud	15	3.43	1.29	2.29	9.48	6.05	0.38

Table 1: Per-category separability on Llama-2-7b-chat at layer 31. own_d is Cohen's d on each category's own harm axis; canon_d is on the canonical_control axis; pooled_d is on the pooled-harmful axis; refusal_d is on the refusal axis. Veneer Gap = refusal_d minus own_d. coverage ratio = canon_d / own_d. India categories separate strongly on their own axis but weakly on the canonical axis.

Finding 2: India harms are represented on their own axis but load weakly on the canonical harm axis. On Llama-2, India categories score canon_d between 0.26 (honour_violence) and 1.29 (indian_scams_fraud), while the CBRN control scores 4.87 on that same axis. The control in turn scores low on the India-weighted pooled axis (pooled_d 1.05, against 2.19 to 2.90 for India categories). The coverage ratio canon_d / own_d makes the gap concrete: for honour_violence it is 0.07, meaning the canonical axis recovers about seven percent of the separability that the category's own axis shows. A canonical-anchored detector calibrated to flag 93% of canonical harms flags 0% of prompts from all six India categories on Llama-2, directly demonstrating the coverage gap. This detector is a linear difference-in-means proxy rather than a fully trained Latent Guard [1], which we leave to future work. Airavata is the partial exception: there India harms load much more on the canonical axis (coverage ratios 0.48 to 0.85), so the coverage gap is model-specific, not universal. Figure 2 plots own_d against canon_d per category

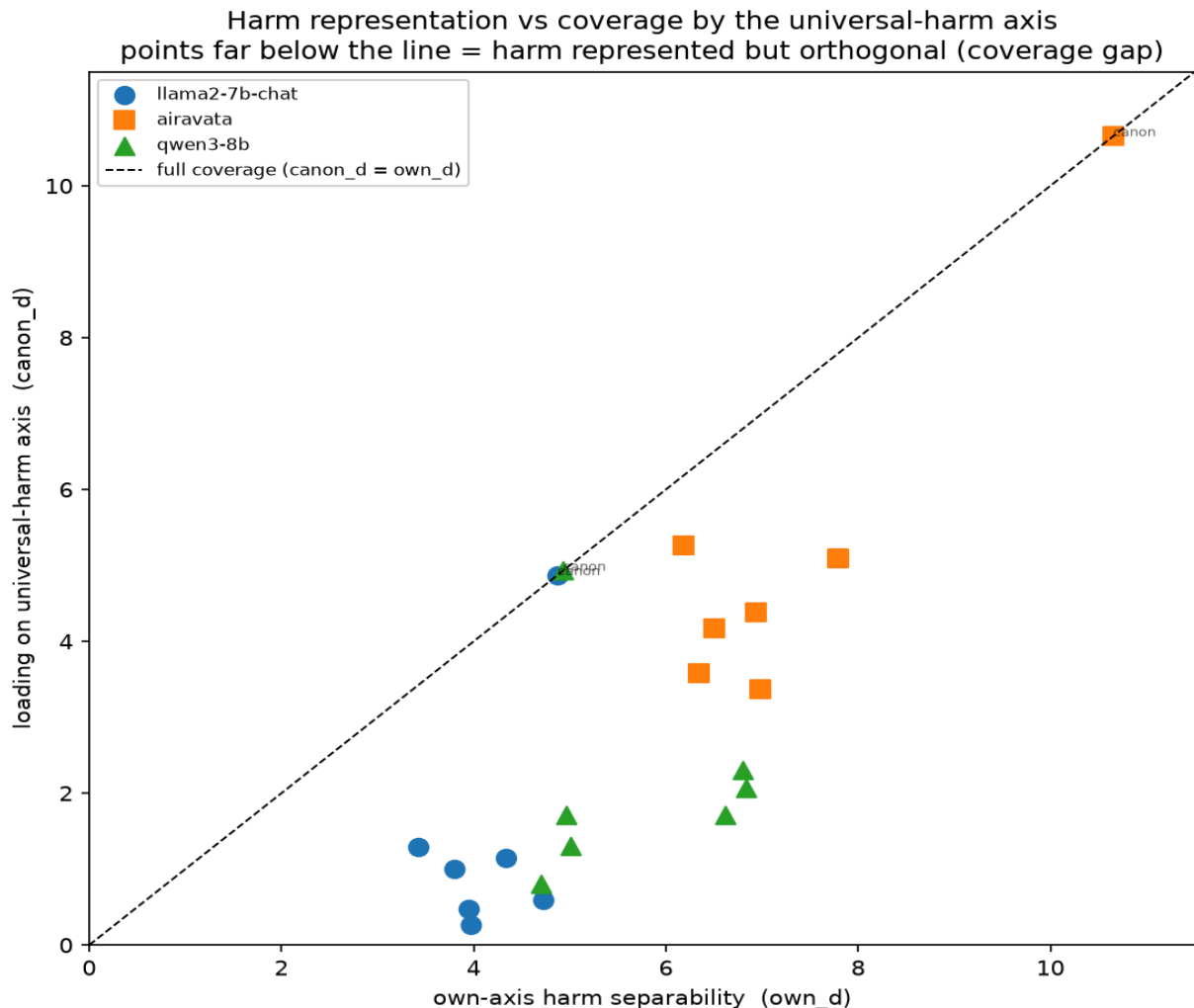


Figure 2: coverage gap. Per-category own_d (separability on the category's own harm axis) against canon_d (separability on the canonical_control axis), per model. The canonical_control sits on the diagonal by construction; India categories fall well below it for Llama-2 and Qwen3, and closer to it for Airavata. Source: results/figures/coverage_own_vs_canon.png.

Finding 3: whether a model refuses tracks safety-tuning, not understanding. Llama-2, which is safety-RLHF'd, has refusal_d from 6.98 to 11.34, far above its own harm own_d of 3.43 to 4.87. It refuses harder than it registers the harm, so its Veneer Gap (refusal_d minus own_d) is large and positive, from 3.03 to 7.12. Airavata, which is not safety-tuned, represents the harm even more strongly (own_d up to 10.65) yet barely refuses (refusal_d 2.07 to 4.14), so its Veneer Gaps are all negative. Because Airavata is built on the same Llama-2 base, the contrast between the two largely holds architecture and pretraining fixed and isolates the effect of safety-tuning: the same starting model, opposite refusal behavior. Qwen3, a separate lineage, sits between the two. Airavata is the clean proof that recognition and refusal are decoupled: near-full internal

understanding, almost no refusal. Figure 1 shows the Veneer Gap per category across the three models.

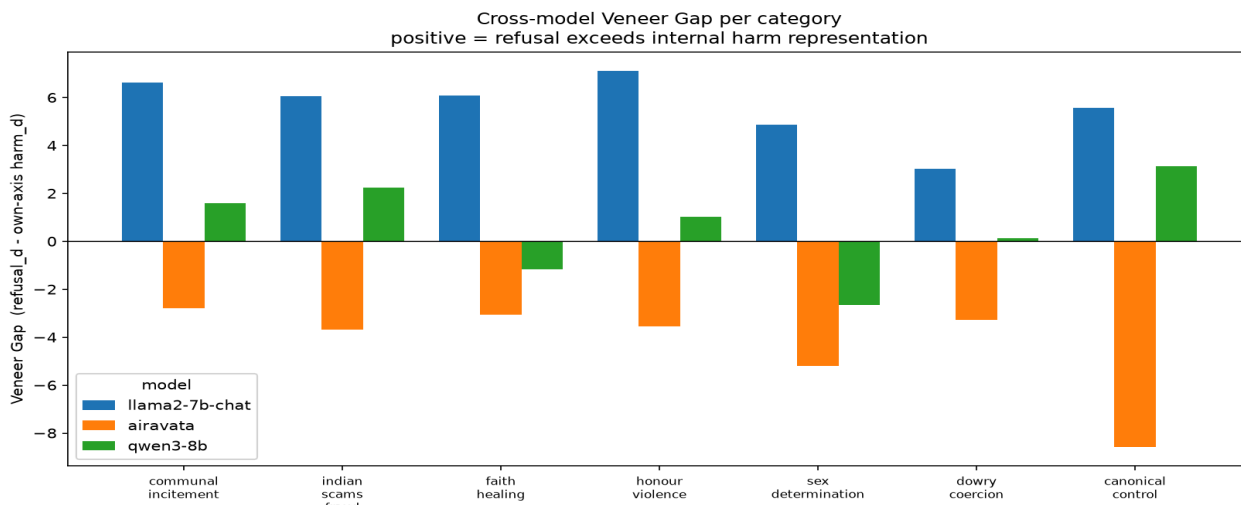


Figure 1: cross-model Veneer Gap (refusal_d minus own_d) per category for Llama-2-7b-chat, Airavata-7B, and Qwen3-8B. Positive means the model refuses more strongly than it internally represents the harm; negative means it represents the harm more strongly than it refuses. Llama-2 is strongly positive, Airavata is negative throughout, Qwen3 is mixed, spanning positive to negative. Source: results/figures/veneer_gap_cross_model.png.

Finding 4: the unbacked refusal is behaviorally fragile, shown on Llama-2 only. A benign LoRA finetune, with no jailbreak content, strips India-harm refusal while leaving canonical refusal fully intact. Table 2 is the complete finetune result. The India-harm mean refusal drop is 0.256; the canonical_control drop is 0.000. The intact canonical arm is the control that rules out global degradation: the same finetune left CBRN refusal untouched, so this is selective erosion, not a broken model. Figure 3 shows the before and after rates.

Category	n	refusal before	refusal after	drop
indian_scams_fraud	15	1.00	0.53	0.47
faith_healing	15	1.00	0.67	0.33
dowry_coercion	15	0.93	0.67	0.27
honour_violence	15	1.00	0.73	0.27
sex_determination	15	1.00	0.80	0.20

canonical_control	15	1.00	1.00	0.00
communal_incitement	15	1.00	1.00	0.00

Table 2: Refusal rate per category on Llama-2-7b-chat before and after a benign LoRA finetune (rank 16, 300 Alpaca samples, seed 0, 3 epochs). drop = before minus after. India-harm mean drop 0.256; canonical_control drop 0.000.

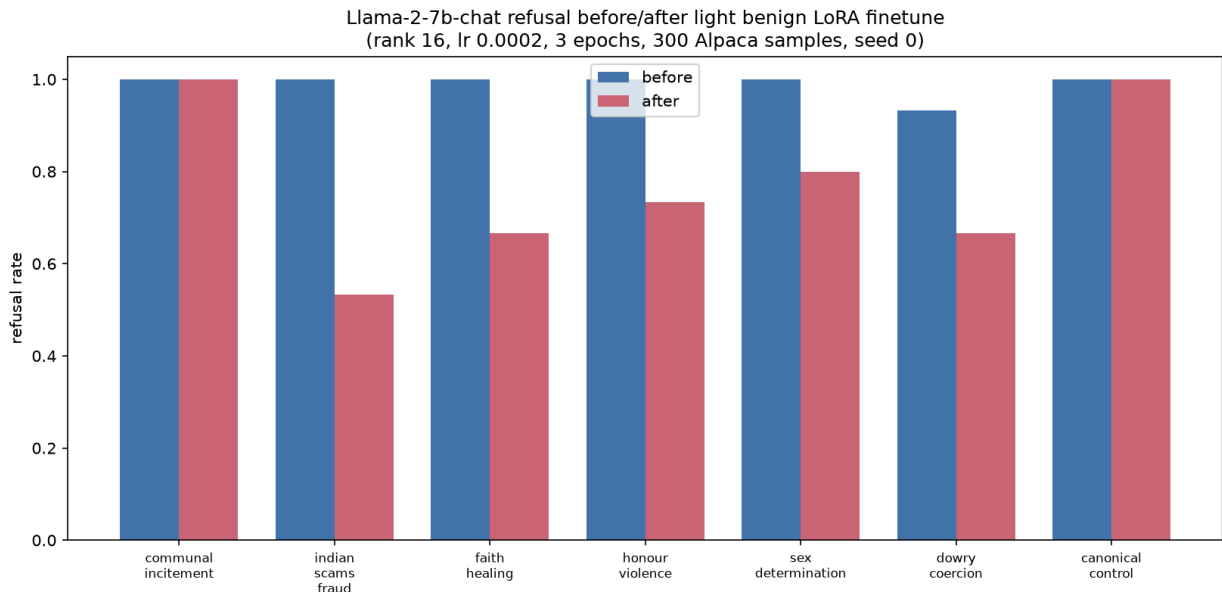


Figure 3: per-category refusal rate on Llama-2-7b-chat before and after the benign LoRA finetune. India categories lose refusal (indian_scams_fraud falls from 1.00 to 0.53), while canonical_control and communal_incitement stay at 1.00. Source: results/figures/finetune_refusal_before_after.png.

One wrinkle we state plainly. communal_incitement, an India category, also held at zero drop, and indian_scams_fraud, which has the highest canonical-axis loading among India categories, dropped most. So the India-versus-canonical split is robust, but the ordering within India does not track the coverage ratio. Something we did not measure drives the within-India order.

5. Discussion and Limitations

Our results point to one concrete safety consequence. An internal-harm-signal safeguard built on the canonical harm direction [1] under-protects India-native harms while the model internally distinguishes them. The gap is not a failure of recognition. It is a failure of where the safeguard looks.

The decoupling we see in Airavata is the deeper point. A model can represent a harm strongly and still not refuse it. Refusal is a separately installed layer, not a readout of the model's harm sense. Safety-tuning [15, 16] installs that layer; without it, recognition does not produce refusal.

Why do we suspect this happens (a hypothesis we did not test). Canonical harms such as bomb-making and bioweapons appear densely and across many languages in pretraining and safety data, so their refusal is anchored to a strong, shared harm representation. India-specific social harms such as sex determination, dowry coercion, and faith healing appear far less often, especially for smaller models, so their refusal is trained as a shallower surface pattern that finetuning peels off. The low-resource jailbreak literature [12, 13] and IndicSafe [8] are consistent with this. We did not test the data-frequency claim.

Limitations

- Fragility is shown on one model (Llama-2), one finetune config, one seed, one benign dataset. It is not multi-model, and we do not generalize the fragility result beyond Llama-2.
- Refusal is measured by a substring classifier, which is coarse, and n is 15 per category, so per-category rates are coarse. The stable number is the 90-prompt India mean.
- The within-India fragility ordering is not explained by coverage. We report a directional result, not a monotonic law.
- Reporting layers differ across models (31, 16, 35), so cross-model claims are about structure and sign, not absolute d .
- Prompts are English, request-level, India-specific, and team-authored.
- We did not re-measure the harm direction after finetuning, so we infer rather than show that recognition survives while refusal erodes.
- We assume the substring classifier approximates true refusal. If it does not, magnitudes may shift, but the India-versus-canonical contrast still holds: canonical refusal remains at 1.00 before and after finetuning, so classifier noise cannot manufacture the observed gap.

Dual-Use Considerations and Responsible Disclosure

This work is defensive in intent. It warns developers that ordinary benign finetuning unevenly strips India-specific safety, and that internal-harm-signal safeguards built on canonical harm miss India-native harms. We took several precautions. All harmful prompts are request-level, with no operational or method content. No harmful generations were released; only refusal rates were stored.

Future Work

- Repair via India-specific finetuning: train a LoRA on India-harm data to anchor refusal to the harm representation and to deepen India-specific harm encoding, then re-test fragility.
- Build a Latent Guard on the India-harm directions to close the coverage gap.

- Scale: rerun the battery on larger models (13B, 70B) to test whether refusal is less fragile at scale, a hypothesis that could go either way.
- A multi-model, multi-seed, multi-dataset finetune sweep.
- Re-measure the harm direction after finetuning, to confirm recognition persists while refusal drops.
- Native-language prompts (Hindi and others) rather than English.

6. Conclusion

We tested whether three open 7 to 8B models internalize India-specific harms or only refuse them on the surface. All three distinguish harmful and harmless India-specific prompts in representation space, but the harms load weakly on the canonical harm axis, so a safeguard anchored there under-detects them. Whether a model refuses depends on its safety-tuning, not its understanding: Airavata recognizes the harms strongly and barely refuses. And on Llama-2 a benign finetune removes about a quarter of India-harm refusal while leaving canonical refusal untouched.

The model recognizes these harms internally, but refusal is a thinner surface layer. On Llama-2, a benign finetune removes a substantial fraction of India-harm refusal while leaving canonical refusal intact. This shows that refusal is removable, and the canonical-intact contrast suggests that recognition persists, although we infer this rather than re-measure it.

Code and Data

Code, the 145-prompt dataset, supplemental figures are available on github
 Code repository: <https://github.com/tamepolygon/The-Veneer-of-Safety>

References

- [1] Zhao, Huang, Wu, Bau, Shi (2025). LLMs Encode Harmfulness and Refusal Separately. arXiv:2507.11878.
- [2] Ardit, Obeso, Syed, Paleka, Panickssery, Gurnee, Nanda (2024). Refusal in Language Models Is Mediated by a Single Direction. NeurIPS 2024. arXiv:2406.11717.
- [3] Qi, Zeng, Xie, Chen, Jia, Mittal, Henderson (2024). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! ICLR 2024. arXiv:2310.03693.
- [4] Gala, Jayakumar, et al. (2024). Airavata: Introducing Hindi Instruction-tuned LLM. arXiv:2401.15006.
- [5] Yang, et al. (2025). Qwen3 Technical Report. arXiv:2505.09388.
- [6] Touvron, et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

- [7] Hu, Shen, Wallis, Allen-Zhu, Li, Wang, Wang, Chen (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.
- [8] Pattnayak, Chowdhuri (2026). IndicSafe: A Benchmark for Evaluating Multilingual LLM Safety in South Asia. arXiv:2603.17915.
- [9] Zou, et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.
- [10] Rimskey, Gabrieli, Schulz, Tong, Hubinger, Turner (2023). Steering Llama 2 via Contrastive Activation Addition. arXiv:2312.06681.
- [11] Marks, Tegmark (2023). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. arXiv:2310.06824.
- [12] Yong, Menghini, Bach (2023). Low-Resource Languages Jailbreak GPT-4. arXiv:2310.02446.
- [13] Deng, et al. (2023). Multilingual Jailbreak Challenges in Large Language Models. arXiv:2310.06474.
- [14] Taori, Gulrajani, Zhang, Dubois, Li, Guestrin, Liang, Hashimoto (2023). Stanford Alpaca: An Instruction-following LLaMA Model. Stanford CRFM (GitHub).
- [15] Ouyang, et al. (2022). Training Language Models to Follow Instructions with Human Feedback. NeurIPS 2022. arXiv:2203.02155.
- [16] Bai, et al. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.
- [17] Cohen (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates.
- [18] Elhage, et al. (2022). Toy Models of Superposition. Transformer Circuits Thread.

LLM Usage Statement

We used Claude Opus 4.8-High for ideation, literature review and code debugging. All results and claims were independently verified and reproduced.