
Investigating Activation Threshold Failures in Cross-Lingual Prompt Rejection

Erickson Leon Kovalski
CiDAMO - Federal University of Parana - Brazil

With
Apart Research

Abstract

Current literature suggests that Large Language Models (LLMs) suffer from cross-lingual fragility, where safety constraints established in high-resource languages fail to generalize to low-resource languages. This project empirically evaluates this hypothesis within Llama-3's latent space, comparing English and Portuguese representations. Through mechanistic analysis of residual stream activations, we found that the model exhibits high cross-lingual robustness. Concepts across languages are aligned in direction, with geometric deviations primarily mapped to linguistic translation distance rather than conceptual misalignment. However, we identified a structural variance: representations of prompts in Portuguese consistently exhibit lower vector magnitudes than their English counterparts. Furthermore, our analysis indicates that different semantic domains trigger distinctly different refusal mechanisms. We hypothesize that this magnitude decay leaves a theoretical vulnerability where boundary-case "low-risk" malicious prompts still might fail to breach the activation threshold required to trigger these specific refusal mechanisms in Portuguese.

1. Introduction and Related Work

As Large Language Models (LLMs) achieve widespread global deployment, ensuring their safety across diverse linguistic contexts has become a critical alignment challenge. While current

reinforcement learning from human feedback (RLHF) techniques reliably instill safety constraints in high-resource languages such as English, recent empirical studies demonstrate that these defenses frequently fail to generalize. Yong et al. (2023) established that translating malicious prompts into low-resource languages effectively bypasses LLM safety filters, exposing a fundamental vulnerability in multilingual alignment.

To understand the internal mechanics of these failures, research has increasingly turned to mechanistic interpretability. Shifting focus from individual neuron activations—a process Foote et al. (2023) characterize as an immensely complex challenge requiring automated, graph-based interpretation at scale—the field of Representation Engineering analyzes population-level representations within the latent space (Zou et al., 2023). Using these top-down methods, Ardit et al. (2024) made the foundational discovery that an LLM’s refusal behavior is largely mediated by a single, distinct direction in the activation space.

This discovery has sparked a debate regarding the structural nature of multilingual safety. Building on the single-direction hypothesis, Wang et al. posit that the refusal direction exhibits striking cross-lingual universality, arguing that refusal vectors derived from English can transfer seamlessly to other languages due to parallelism in the embedding space. However, this assertion of a monolithic, universal defense presents a contradiction: if the refusal mechanism is perfectly universal, why do multilingual jailbreaks remain so highly effective? The recent work by Banerjee et al. (2025), which explores language-specific functional parameter steering to enforce multilingual safety, further suggests that localized, language-aware interventions are still urgently necessary.

Our project seeks to explore the gap between the theoretical universality of the refusal direction and the empirical reality of cross-lingual vulnerabilities. We propose that while the *direction* of refusal may remain relatively parallel across languages (partially supporting Wang et al.), treating the safety mechanism as a monolithic entity obscures critical topological degradation.

By analyzing the residual stream activations of Llama-3 8B Instruct (at Layer 15) using carefully curated "dual-use" prompts, this exploratory work investigates the latent geometry of English versus Portuguese. We aim to demonstrate that cross-lingual jailbreaks do not exploit a complete absence of safety concepts, but rather a measurable "Magnitude Collapse" and an asymmetric "Boundary Collapse" that vary heavily by semantic domain. Ultimately, we seek to determine if targeted, domain-aware latent steering at inference time can artificially restore this lost magnitude, offering a targeted patch for the multilingual translation gap.

3. Methods

This project employs mechanistic interpretability to audit and quantify the cross-lingual safety alignment of Large Language Models (LLMs), specifically targeting the geometric representations of English and Portuguese within the residual stream of Llama-3.

1. Datasets and Curation

To avoid the confounding variables of explicit roleplay wrappers (which test persona adherence rather than safety boundaries), we engineered highly targeted datasets categorized prompt pairs used to establish baseline latent representations for both safe and unsafe inputs across multiple semantic domains (e.g., Physical Harm, Malware, Fraud) extracted from Combined_Multilingual_Dataset.csv from Soteria and PolyRefuse.csv dataset from MaiNLP as well as sampled prompts from Malicious Prompt Detection Dataset (MPDD).

2. Tools and Libraries

The analytical pipeline was built using the following computational stack: PyTorch handled core tensor operations and gradient-free inference execution. TransformerLens was utilized to intercept the forward pass, apply specific chat templates, and extract internal network activations via hook points without modifying the underlying model weights. Pandas & NumPy: Managed cross-lingual dataset parsing, feature alignment, and vector matrix operations. Matplotlib & Seaborn were Used for rendering Kernel Density Estimation (KDE) probability distributions of the projected latent spaces. Runpod (Compute Infrastructure): We hosted the Llama-3 8B Instruct model on a single 96GB VRAM NVIDIA A100 instance via Runpod. Mechanistic interpretability is notoriously memory-hungry; the necessary run_with_cache procedures to extract internal states and apply latent steering require significantly more than 48 GB of VRAM to avoid Out-Of-Memory (OOM) errors. Albeit possible to optimize the code to avoid this—such as aggressively filtering cache hooks or offloading tensors to the CPU—the degradation of performance to operate on CPU, as well as the engineering complexities of constantly moving data between GPU and CPU, made this approach less optimal for the rapid prototyping constraints of a hackathon.

3. Analytical Procedure

Our methodology is divided into three distinct phases: **Phase I: Empirical Audit** establishes a reproducible behavioral baseline by executing deterministic generation (temperature = 0.0). Refusals in RLHF-aligned models are heavily front-loaded; the model typically reveals its compliance state in the first sentence (e.g., "I cannot fulfill this request" vs. "Here is the script you requested").

Phase II: Latent Extraction and Directional Isolation: We extracted the latent geometry by intercepting the residual stream just prior to the MLP blocks at Layer 15. The choice of Layer 15 (within a 32-layer architecture like Llama-3 8B) is highly deliberate. Mechanistic interpretability research demonstrates that transformers process information hierarchically. Early layers (e.g., Layers 0 to 10) are predominantly focused on low-level lexical processing, syntax, and multi-lingual token aggregation; abstract safety concepts have not yet fully crystallized. Conversely, late layers (e.g., Layers 25+) are heavily skewed toward the mechanics of immediate next-token prediction, where the geometry becomes dominated by the specific vocabulary word the

model is about to output. Layer 15 sits at the critical semantic midpoint. It is deep enough for the network to have calculated the abstract "intent" and "risk" of the prompt, but early enough that this conceptual representation remains pure and hasn't collapsed into specific output tokens. Vector representations were captured exclusively at the final token position of the input prompt, representing the fully aggregated context before autoregressive generation begins. We then computed the universal refusal direction in English using a difference-in-means approach, subtracting the centroid of the benign vectors from the centroid of the malicious vectors

$$\mu_{\text{malicious}} - \mu_{\text{benign}}$$

Phase III: Projection and Magnitude Collapse Analysis: To quantify the cross-lingual decay in safety activation, all extracted vectors (English and Portuguese) were projected onto the normalized English refusal direction using scalar dot products. This geometric projection mathematically isolates how much of the "refusal" concept is present in any given prompt, filtering out unrelated semantic and linguistic noise. We plotted the resulting probability densities to identify the empirical refusal threshold and calculated the precise percentage of "Magnitude Collapse" occurring when malicious concepts are processed in the Portuguese latent subspace.

4. Results and Discussion

1. Geometric Evidence of Global Magnitude Collapse

To initially evaluate whether Llama-3 processes harmful intents differently across languages, we projected Layer 15 residual stream activations onto a calculated "Universal Refusal Direction". The resulting Kernel Density Estimation (KDE) plot provides our first mechanistic evidence of a cross-lingual safety discrepancy.

As illustrated in the *Scalar Projection* KDE plot, the model demonstrates highly robust conceptual separation in its high-resource training language. The Benign vectors (green distribution) anchor the left side of the latent space, sharply trailing off before the empirical refusal threshold (dashed line near -2.9). The English Malicious vectors (red distribution) sit safely to the right of this threshold with a mean projection of -0.09.

However, when identical malicious concepts are processed in Portuguese (orange distribution), the latent geometry shifts drastically. The Portuguese Malicious distribution exhibits a pronounced "Magnitude Collapse." Its mean projection drops from -0.09 to -0.41—a staggering -371.6% magnitude collapse. Crucially, a significant portion of the Portuguese left tail bleeds past the empirical refusal threshold, merging with the benign geometric space. Prompts falling into this deficit zone fail to accumulate enough latent "refusal energy" to trigger the safety circuit, resulting in asymmetric malicious compliance.

2. Bimodal Distribution

While the KDE plot confirms a global magnitude decay, a closer inspection of its shape reveals a critical nuance. Both the English and Portuguese malicious distributions are clearly bimodal—they feature distinct, uneven "bumps." In English, we see a minor peak around -0.8 and a major peak around $+0.8$ (Figure 1). In Portuguese, the primary peak shifts dangerously close to the threshold at -1.2 , with a secondary bump near $+0.5$.

	Safety Category	English Silhouette	Portuguese Silhouette	Boundary Collapse (Delta)
1	Economic Harm	0.2703	0.1733	0.0969
0	Child Abuse Content	0.4309	0.3424	0.0885
6	Physical Harm	0.2600	0.2059	0.0541
2	Fraud/Deception	0.2333	0.1803	0.0530
5	Malware	0.1892	0.1527	0.0365
3	Hate/Harass/Violence	0.2785	0.2460	0.0325
8	Privacy Violation Activity	0.1899	0.1793	0.0106
7	Political Campaigning	0.2539	0.2561	-0.0023
4	Illegal Activity	0.2664	0.2769	-0.0105

Table 1: English and Portuguese silhouette and boundary collapse (delta)

This uneven density led us to a critical realization: assuming a single, monolithic "Universal Refusal Direction" oversimplifies the latent geometry. The multiple peaks suggest that different malicious subjects trigger different refusal mechanisms with varying latent strengths.

	Theme Category	r_EN Norm	r_PT Norm	EN-PT Direction Cosine Similarity
1	Economic Harm	3.728	2.615	0.8821
7	Political Campaigning	3.697	3.342	0.9118
2	Fraud/Deception	3.827	3.090	0.9173
5	Malware	3.516	2.914	0.9245
0	Child Abuse Content	5.153	4.003	0.9296
3	Hate/Harass/Violence	4.078	3.443	0.9321
6	Physical Harm	4.017	3.195	0.9337
8	Privacy Violation Activity	3.517	3.166	0.9392
4	Illegal Activity	4.125	3.815	0.9469

Table 2: Different rejection vector on different kinds of malicious prompts across EN and PT.

To test this, we isolated the refusal directions by specific semantic categories and calculated their absolute L2 norms. The *Mean Refusal Direction Norm* matrix confirms our suspicion,

revealing a highly uneven defensive topography. While the directional alignment remains high (Cosine Similarities >0.88), the absolute magnitude of the refusal vector varies drastically between domains (Table 2).

$$\|r_{cat}\| = \|\mu_{malicious,cat} - \mu_{benign,cat}\|$$

This high directional alignment points that there is no severe conceptual misalignment between the languages; the model correctly maps the malicious intent to the same general orthogonal subspace. Consequently, we hypothesize that the observed geometric deviations, specifically the uniform shrinking of the L2 norms, are primarily artifacts of linguistic translation distance: the resulting vectors lose magnitude while preserving their core semantic direction.

Categories such as *Child Abuse Content* and *Illegal Activity* exhibit massive L2 norms, indicating heavily reinforced safety sub-spaces. In stark contrast, categories like *Economic Harm* and *Privacy Violation Activity* display significantly weaker baseline representations.

Across every single domain, the Portuguese refusal vector norm is shorter than its English counterpart. Because low-norm domains are inherently closer to the safety hyperplane even in English, this uniform Portuguese magnitude drop mathematically increases the probability that these specific, weaker categories will fall below the critical refusal threshold.

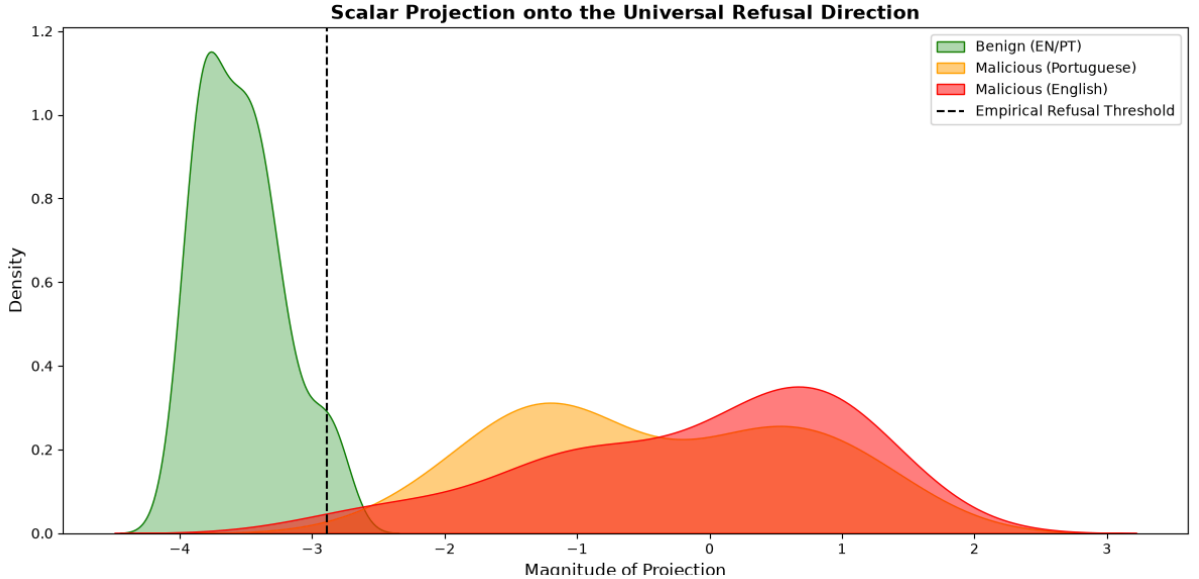


Figure 1: Scalar projection onto a Universal Refusal Direction.

3. Silhouette Degradation and Boundary Collapse

To fully map this domain-specific fragility, we must look beyond vector length and examine the structural integrity of the concepts themselves. By calculating the Unsupervised Silhouette Scores for each domain (Table 1), measuring how cleanly the model separates malicious payloads from innocuous contexts, shows that the degradation of the safety boundary is profoundly

asymmetric:

- **Severe Boundary Collapse:** Domains such as *Economic Harm* (dropping from an EN score of 0.2703 to a PT score of 0.1733, with a delta of 0.0969) and *Child Abuse Content*, (with a delta of 0.0885, suffer a massive drop in Silhouette score when translated into Portuguese. In these specific sub-spaces, the geometric boundary between "safe" and "unsafe" effectively dissolves. The model loses its semantic resolution, causing malicious vectors to structurally bleed into the benign cluster.
- **Structural Rigidity:** Conversely, categories like *Political Campaigning*, with a delta of -0.0023 and *Illegal Activity*, with a delta of -0.0105, exhibit almost zero degradation, and in some cases cluster slightly better in Portuguese. Their boundaries strongly resist cross-lingual drift.

3.1. Implications

This topological discovery redefines the mechanics of a cross-lingual jailbreak. A successful attack does not simply rely on a global, uniform blindness. Instead, it exploits a targeted Boundary Collapse within specific semantic domains. By cloaking a dual-use prompt in a domain suffering from high Silhouette degradation (like Economic Harm), an attacker forces the model into a fragmented latent space where the mathematical distinction between benign assistance and malicious compliance no longer exists.

5. Limitations and future work

While our topological extractions offer a granular look at cross-lingual vulnerabilities within Llama-3 8B, this work is tightly bound by structural constraints that limit its absolute generalizability. Future exploration of more complex models and models from different families might uncover new patterns of refusals across languages. Additionally, zero-shot latent steering represents an interesting avenue for future research, suggesting that patching multilingual LLMs could be theoretically approachable through targeted linear algebra at inference time.

6. Conclusion

This project represents an exploratory foray into the latent space of Llama-3. We set out to observe how multilingual inputs interact with aligned safety guardrails, recognizing that our interpretability tools and our fundamental understanding of these high-dimensional spaces remain nascent. Rather than discovering a universal "blindness" to malicious concepts in Portuguese, our observations point toward a more nuanced, probabilistic degradation. By tracking curated dual-use prompts through Layer 15, we noted a trend where the mathematical boundary between benign and malicious intents blurs in the lower-resource language. We conceptualized this as a "Magnitude Collapse", i.e., not an absolute erasure of the harmful concept, but a shrinking of the refusal vector's length. Crucially, this decay is uneven; it manifests as a domain-specific "Boundary

Collapse," particularly in inherently weaker semantic sub-spaces like Economic Harm and Child Abuse Content. These observations offer a compelling perspective on AI alignment. The hypothesis that cross-lingual vulnerabilities are rooted in quantifiable geometric decay, rather than absence of training data, opens a pathway for mitigation.

Code and Data

GitHub Repository

<https://github.com/3ff3x0r/AI-safety-Hackathon-2026---Gobal-South>

Datasets Used

Our evaluation pipeline curated prompts from the following open-source repositories:

MPDD (Malicious Prompt Detection Dataset):

<https://www.kaggle.com/datasets/mohammedaminejebbar/malicious-prompt-detection-dataset-mpdd>

PolyRefuse (Multilingual Refusal):

<https://github.com/mainlp/Multilingual-Refusal/tree/main/PolyRefuse>

Soteria (Multilingual Safety Alignment):

<https://huggingface.co/datasets/SoftMINER-Group/Soteria/tree/main>

References

1. Ardit, A., Obando-Ceron, M., Carrubba, E., Nanda, N., & Ndousse, E. (2024). *Refusal in LLMs is mediated by a single direction*. arXiv preprint.
2. Banerjee, S., Layek, S., Chatterjee, P., Mukherjee, A., & Hazra, R. (2025). *Soteria: Language-Specific Functional Parameter Steering for Multilingual Safety Alignment*. arXiv preprint arXiv:2502.11244.
3. Foote, A., Nanda, N., Kran, E., Konstas, I., Cohen, S. B., & Barez, F. (2023). *Neuron2Graph: Interpreting Language Model Neurons at Scale*. arXiv preprint.
4. Wang, X., Wang, M., Liu, Y., Schütze, H., & Plank, B. (2024). *Refusal Direction is Universal Across Safety-Aligned Languages*. arXiv preprint.
5. Yong, Z. X., Menghini, C., & Fellbaum, C. (2023). *Low-Resource Languages Jailbreak*

GPT-4. arXiv preprint arXiv:2310.02446.

6. Zou, A., Phan, L., Chen, S., Campbell, J., Huan, P., Yin, C., Mazeika, M., Dombrowski, A., Ghaffari, M., & Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv preprint arXiv:2310.01405.

LLM Usage Statement

The report was mostly generated by Gemini 3.1 pro, with human revision of the final version.