# Two Refusals or One? Safety vs. Epistemic Abstention Directions in LLM Activations

Muhammad Aaliyan | Independent Researcher, Pakistan | ICML 2026 Mechanistic Interpretability Workshop

**Main result: safety refusal and epistemic abstention are related surface behaviors, but not the same dominant activation-space direction.**

## Question

Instruction-tuned LLMs refuse for different reasons:

Safety refusal: declining harmful requests.
Epistemic abstention: declining to answer when the model lacks enough knowledge.

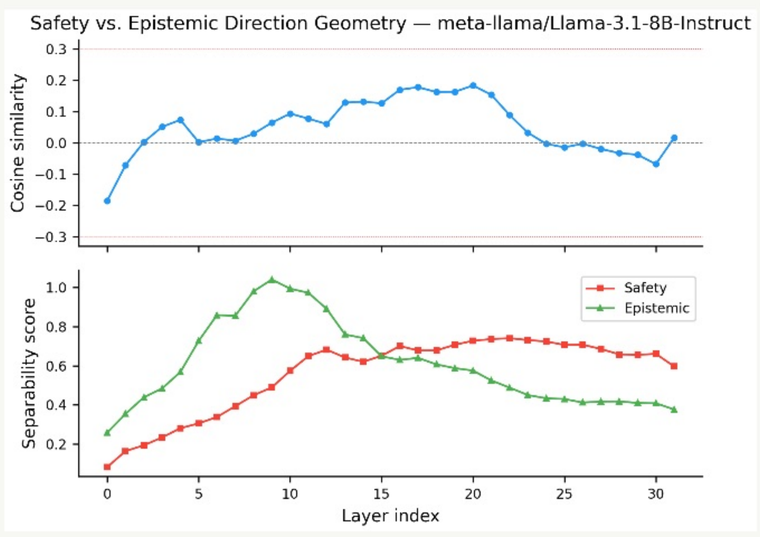
Do these behaviors share a single geometric handle in residual-stream activation space?

## Method

- Extract safety and epistemic directions with difference-in-means over contrastive prompt pairs at every residual layer.
- Compare layerwise cosine similarity and separability profiles.
- Run activation-space cross-ablation: remove one direction and test whether the other behavior collapses.
- Measure quantization drift and bounded replication in Qwen3-8B and Gemma-2-9B-Instruct.

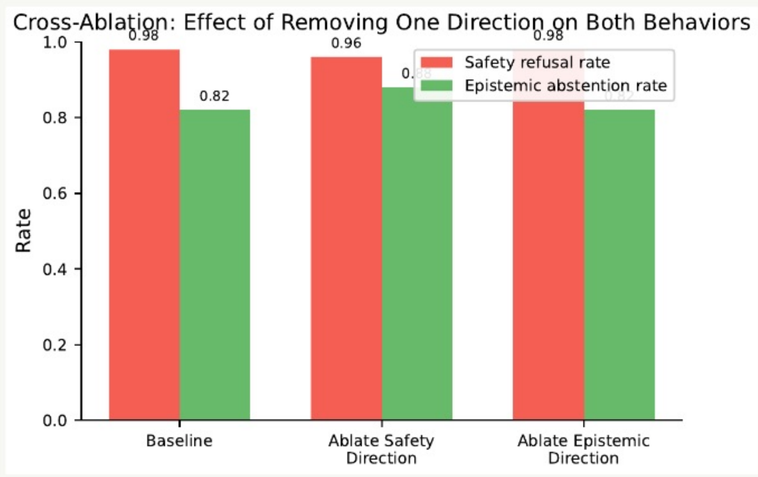
## 1. Geometry separates the behaviors

**Llama-3.1-8B-Instruct:**
**mean cosine = 0.049**
**max cosine = 0.183**
**safety peak layer = 22**
**epistemic peak layer = 9**




Direction similarity stays weak while the two separability profiles peak at different depths.

## 2. Causal cross-ablation is asymmetric

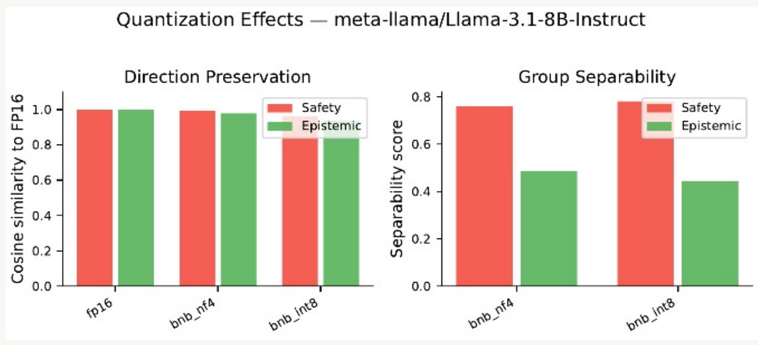



Strongest useful safety ablation:

Safety refusal: 0.98 -> 0.16
Epistemic abstention: 0.82 -> 0.88
Cross-contamination: +0.06, 95% CI spans zero

Epistemic ablations were weaker, suggesting a more diffuse or less well-captured handle.

## 3. Quantization mostly preserves geometry



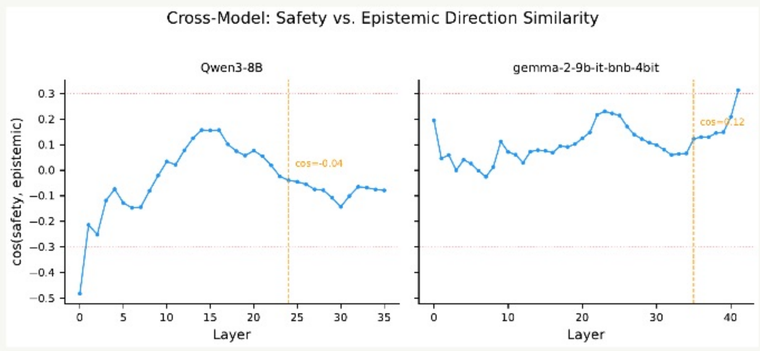

Relative to FP16 at the selected layer:
NF4 cosine: safety 0.994, epistemic 0.978
INT8 cosine: safety 0.962, epistemic 0.935

The main separation is not an artifact of 4-bit loading.

## Linear probe sanity check

Held-out logistic regression on epistemic activations reaches 1.00 test accuracy at the selected epistemic layer. The epistemic signal is linearly present even though single-direction ablation is weak.

## 4. Cross-family probes match the separation story




At each model's top safety layer:
Qwen3-8B cosine = -0.039
Gemma-2-9B cosine = 0.123

Different tokenizers and depths; same qualitative result.

## Limitations

- Keyword detectors miss paraphrased refusal and abstention behavior.
- Three model families is a bounded replication, not a broad survey.
- Epistemic QA prompts may partly conflate knowledge limits with task format and reading-comprehension difficulty.
- The strongest causal finding is asymmetric: safety is highly steerable; epistemic abstention remains harder to ablate cleanly.

## What this changes

For interpretability: similar refusal-like outputs need not share the same internal feature.

For safety: removing a refusal direction can make a model less safe without removing its ability to say “I don’t know.”

For follow-up: improve epistemic datasets and detectors, then test whether abstention is multi-dimensional or distributed.

## Takeaway

**The evidence rules against a shared dominant direction. Safety refusal can be sharply weakened without erasing epistemic abstention, so deployment-time edits that disable refusal may leave models unsafe while still superficially “humble.”**

## Artifacts

Paper: camera-ready source in paper/main.tex
Figures: artifacts/figures
Code: src/ddmi and scripts/
Contact: aaliyan1230@gmail.com

Accepted as an in-person poster at the ICML 2026 Mechanistic Interpretability Workshop. Portrait poster size: 24in x 36in.