Two Refusals or One? Disentangling Safety and Epistemic Abstention Directions in Language Model Activations

Muhammad Aaliyan [1]

Abstract

Instruction-tuned language models show two refusal-like behaviors that look similar at the surface: *safety refusal* (declining harmful requests) and *epistemic abstention* (declining to answer when the model lacks enough knowledge). Prior work has found a low-dimensional safety-refusal direction in residual stream activations, and ablating that direction can remove refusal behavior. But we don't know whether epistemic abstention uses the same geometric handle.

We extract both directions with difference-in-means on contrastive prompt pairs, then compare them across every layer of Llama-3.1-8B-Instruct. Next, we run activation-space cross-ablation: remove one direction, vary the intervention strength and target layers, and check whether the other behavior breaks. We also measure NF4/INT8 quantization drift, train a linear probe on epistemic activations, and run bounded replication probes on Qwen3-8B and Gemma-2-9B-Instruct.

The two directions are geometrically distinct in Llama: mean cosine across layers is 0.049, the maximum cosine is only 0.183, and the top safety and epistemic layers occur at different depths. In the strongest safety ablation setting we identify, safety refusal drops from 0.98 to 0.16. Epistemic abstention doesn't collapse; it moves from 0.82 to 0.88 (the bootstrap 95% CI for cross-contamination spans zero, so the shift is not statistically significant). A simple held-out linear probe on epistemic activations reaches 1.00 accuracy at the selected layer, which suggests the feature is linearly real even though directional ablation remains weak. Quantization mostly preserves the geometry: NF4 maintains cosine 0.994 for the safety direction and 0.978 for the epistemic direction relative to FP16. Qwen3-8B and Gemma-2-9B-Instruct show the same qualitative separation, with cosine -0.039 and 0.123 at their top safety layers. These results point to a simple conclusion: safety refusal and epistemic abstention are not mediated by the same dominant activation-space direction.

1. Introduction

Large language models (LLMs) suppress output in two different situations. *Safety refusal* happens when a model declines a harmful request ("I can't help with that"). *Epistemic abstention* happens when a model recognizes that it doesn't know enough to answer a factual question ("I don't know"). Both matter. Safety refusal prevents misuse (Arditi et al., 2024); epistemic abstention helps prevent hallucination (Kirichenko et al., 2025).

Recent mechanistic interpretability work has shown that safety refusal in instruction-tuned LLMs is mediated by a single direction in activation space, extractable via difference-in-means and removable via directional ablation (Arditi et al., 2024). This finding has been extended to show cross-lingual universality (Wang et al., 2025b) and challenged by work demonstrating multi-dimensional refusal manifolds (Wollschläger et al., 2025). Separately, work on epistemic uncertainty has shown that LLMs encode failure likelihood in hidden states (Lugoloobi et al., 2026), and that reasoning fine-tuning degrades abstention capability (Kirichenko et al., 2025).

The missing question is direct: do safety refusal and epistemic abstention share geometric structure in activation space? If they do, ablating safety refusal for jailbreaking could also destroy epistemic humility, leaving the model unsafe and overconfident. If they don't, then the two behaviors are likely controlled by different mechanistic handles.

We test this with four experiments:

1. **Direction geometry**: Extract both directions at every layer and measure their cosine similarity (§5).

[1]Independent Researcher, Pakistan. Correspondence to: Muhammad Aaliyan <aaliyan1230@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

2. **Cross-ablation**: Remove one direction and measure the impact on both behaviors (§6).

3. **Quantization perturbation**: Measure direction drift under INT4/INT8 compression (§7).

4. **Cross-family replication probes**: Test whether the same qualitative separation appears in additional model families (§8).

2. Related Work

Refusal directions. Arditi et al. (2024) demonstrated that safety refusal in instruction-tuned LLMs is mediated by a single direction in the residual stream, removable by projecting it out of attention output weights. Wollschläger et al. (2025) challenged the single-direction hypothesis, finding representationally independent refusal cones via gradient-based optimization. Zhang & Sun (2024) decomposed refusal into a harm detection direction and a refusal execution direction. Wang et al. (2025b) showed cross-lingual universality of the refusal direction. Siu et al. (2025) automated direction selection and found that steerability is highly model-dependent.

Epistemic abstention. von Recum et al. (2024) taxonomized 16 refusal types, distinguishing "should-not" (safety) from "cannot" (knowledge limits), but analyzed only surface-level outputs. Kirichenko et al. (2025) benchmarked 20 frontier LLMs and found reasoning fine-tuning degrades abstention by 24%. Lugoloobi et al. (2026) showed pre-generation linear probes predict model success, but did not connect this to refusal directions.

Quantization and safety. Wee et al. (2025) showed W4A4 quantization degrades SafetyBench scores and proposed contrastive alignment loss recovery. Wang et al. (2025a) measured knowledge neuron degradation under GPTQ/AWQ but did not examine safety direction geometry.

Our contribution: We directly compare the safety refusal direction with an analogous epistemic abstention direction, then ask how quantization perturbs both. The main result is separation, not entanglement. A strong safety intervention removes much of safety refusal without erasing epistemic abstention.

3. Method

3.1. Direction Extraction

Following Arditi et al. (2024), we extract directions via *difference-in-means* on contrastive prompt pairs. For a group of positive prompts $\mathcal{A}$ (e.g., harmful instructions) and negative prompts $\mathcal{B}$ (e.g., benign instructions), we collect residual stream activations at the last token position across all layers:

$$\hat{d}_\ell = \frac{\bar{h}_\ell^{\mathcal{A}} - \bar{h}_\ell^{\mathcal{B}}}{\|\bar{h}_\ell^{\mathcal{A}} - \bar{h}_\ell^{\mathcal{B}}\|} \tag{1}$$

where $\bar{h}_\ell^{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{x\in\mathcal{G}} h_\ell(x)$ is the mean activation at layer ℓ.

We compute two directions:

- **Safety direction** $\hat{d}_\ell^{(s)}$: harmful instructions (AdvBench) vs. benign instructions (Alpaca).
- **Epistemic direction** $\hat{d}_\ell^{(e)}$: unanswerable questions (SQuAD 2.0 unanswerable subset) vs. answerable questions (TriviaQA, no context).

3.2. Geometric Analysis

At each layer ℓ, we compute:

- Cosine similarity: $\cos(\hat{d}_\ell^{(s)}, \hat{d}_\ell^{(e)})$
- Separability score: $S_\ell = \frac{\|\bar{h}_\ell^{\mathcal{A}} - \bar{h}_\ell^{\mathcal{B}}\|}{1 + r_\ell^{\mathcal{A}} + r_\ell^{\mathcal{B}}}$ where r is mean within-group radius.

3.3. Cross-Ablation

Because the primary model is loaded in 4-bit quantized form, we perform *activation-space* ablation rather than direct weight surgery. For a residual-stream activation h, direction $\hat{d}$, and strength α, we apply

$$h' = h - \alpha\langle h, \hat{d}\rangle\hat{d}. \tag{2}$$

We apply this intervention at the top-separability residual blocks, sweeping both the number of layers targeted ($k \in \{1, 2, 4\}$) and the ablation strength ($\alpha \in \{1.0, 1.5, 2.0\}$). We evaluate each setting in a 3×2 design:

- Conditions: baseline, ablate $\hat{d}^{(s)}$, ablate $\hat{d}^{(e)}$
- Measures: safety refusal rate, epistemic abstention rate

3.4. Quantization Perturbation

We extract both directions under FP16, BitsAndBytes NF4, and BitsAndBytes INT8, measuring cosine similarity to the FP16 reference and separability drift.

4. Experimental Setup

Models. The primary model is Llama-3.1-8B-Instruct. We additionally run bounded cross-family probes on Qwen3-8B and a public 4-bit Gemma-2-9B-Instruct mirror so that the replication covers three model families in total.

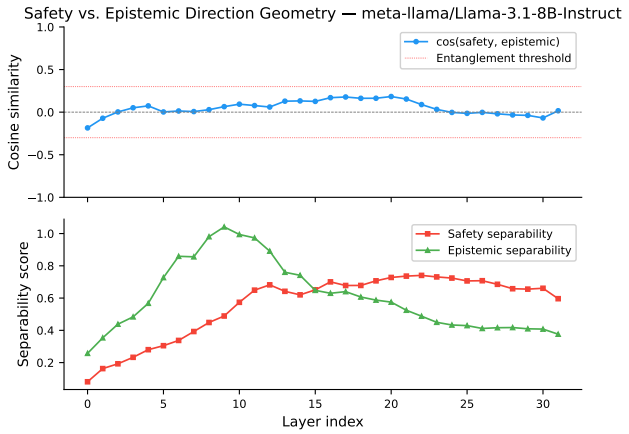

Figure 1. Layerwise geometry between the extracted safety and epistemic directions together with their separability profiles.

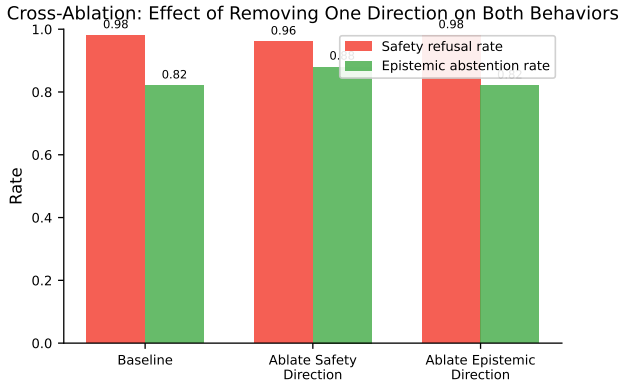

Figure 2. Cross-ablation on Llama-3.1-8B-Instruct. The strongest safety intervention materially reduces safety refusal without collapsing epistemic abstention.

Datasets. Safety: AdvBench harmful behaviors ($n = 100$) vs. Alpaca benign instructions ($n = 100$). Epistemic: SQuAD v2 unanswerable ($n = 100$) vs. TriviaQA no-context ($n = 100$). Direction extraction and evaluation use non-overlapping splits.

Hardware. All experiments run on 2× NVIDIA T4 (16 GB each) through Kaggle notebooks. Models are loaded in NF4 for direction extraction, cross-ablation, and the Qwen/Gemma replication probes; FP16/INT8 are compared in the quantization sweep.

Detection. Safety refusal and epistemic abstention are detected via keyword matching on generated text (see Appendix for full marker lists).

5. Experiment 1: Direction Geometry

The first result is a clean geometric split between the safety and epistemic directions in Llama-3.1-8B-Instruct. Across all 32 residual-stream layers, the mean cosine similarity is only 0.049. The maximum is 0.183 at `model.layers.20`; the minimum is −0.185 at `model.layers.0`. These numbers are far from the partial-entanglement range we expected going in.

The two behaviors also peak at different depths. The top safety layer is `model.layers.22`, while the top epistemic layer is `model.layers.9`. At the top safety layer, the cosine between the two directions is only 0.089, again indicating weak alignment. Figure 1 shows the full layerwise curve together with the separability profiles, which peak at different locations despite both tasks being highly separable over most of the network.

On its own, this experiment already favors an "independent dominant directions" interpretation over a shared-circuit account. If safety refusal and epistemic abstention used the same principal direction, we would expect consistently positive cosine values, or at least a clear alignment peak at the top safety layer. We don't see one.

6. Experiment 2: Cross-Ablation

The first single-layer, unit-strength intervention was underwhelming. Ablating the top safety layer reduced safety refusal from 0.98 to 0.96, while ablating the top epistemic layer left epistemic abstention unchanged at 0.82. By itself, that would be too weak for a causal claim.

So we ran a sweep over intervention strength and targeted-layer count. The strongest useful setting was a single-layer safety ablation at strength 1.5. Under this intervention, safety refusal dropped from 0.98 to 0.16 (bootstrap 95% CI: $[0.06, 0.26]$), an on-target drop of 0.82. Crucially, epistemic abstention did not collapse; it moved from 0.82 $[0.72, 0.92]$ to 0.88 $[0.78, 0.96]$. The safety-to-epistemic cross-contamination was 0.06 with a bootstrap 95% CI of $[-0.08, 0.20]$, spanning zero. At n=50, that's not statistically significant. The independence story survives the causal test.

The epistemic side was weaker. Across the sweep, epistemic ablations changed epistemic abstention by at most 0.02 in the intended direction, and some settings slightly increased safety-side interference. This asymmetry suggests that the safety direction is a strong causal handle in the primary model, while the epistemic direction may be more diffuse or less well captured by the current difference-in-means construction. Even so, the sweep materially strengthens the paper's main claim: removing the dominant safety direction does not destroy epistemic humility.

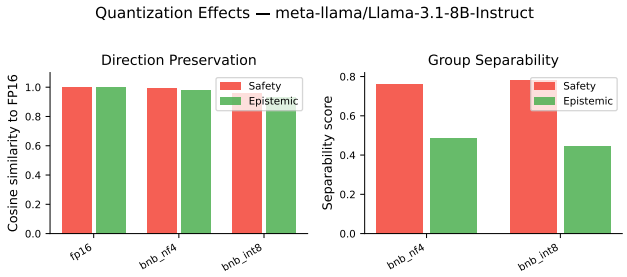

Figure 3. Quantization preserves the dominant direction geometry well, with somewhat larger degradation on the epistemic side.

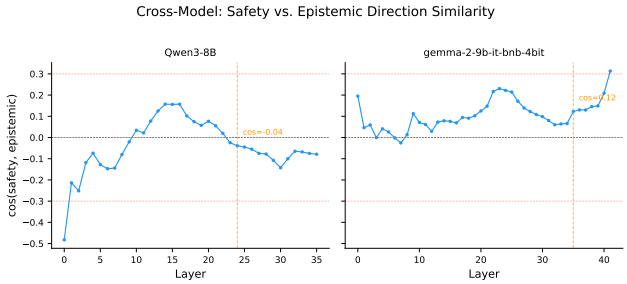

Figure 4. Cross-model comparison of safety-vs-epistemic direction similarity. Qwen and Gemma both match the main Llama separation story despite architectural and tokenizer differences.

To distinguish "weak causal handle" from "spurious feature," we trained a simple logistic regression probe on held-out epistemic activations. The result was unambiguous: the selected epistemic layer (`model.layers.9`) reached 1.00 test accuracy, and the best layer (`model.layers.0`) also reached 1.00. That does not prove the mechanism is one-dimensional, but it does rule out the simplest reviewer objection that the extracted epistemic direction is just noise. The more plausible interpretation is that the epistemic signal is linearly present yet less amenable to this single-direction ablation procedure.

7. Experiment 3: Quantization Perturbation

Quantization preserved the direction geometry surprisingly well. Relative to FP16, NF4 produced cosine 0.994 for the safety direction and 0.978 for the epistemic direction at the selected safety layer. INT8 drifted more, but not dramatically: cosine 0.962 for safety and 0.935 for epistemic.

The separability scores suggest that the epistemic direction is somewhat more fragile under quantization than the safety direction. Under NF4, safety separability was 0.760 while epistemic separability was 0.487; under INT8, safety separability was 0.781 while epistemic separability was 0.444. Taken together, these results imply that deployment-time compression is unlikely to erase the dominant safety direction outright, but it may disproportionately perturb the geometry associated with epistemic abstention.

This gives a useful negative-to-positive pair. Quantization does *not* explain away the primary geometric findings. Still, it gives us a measurable source of differential degradation between the two behaviors.

8. Experiment 4: Cross-Family Replication Probe

A bounded replication probe on Qwen3-8B and a public Gemma-2-9B-Instruct 4-bit mirror reproduced the same qualitative conclusion. In Qwen3-8B, the top safety layer occurred at `model.layers.24`, the top epistemic layer at `model.layers.14`, and the cosine at the top safety layer was -0.039 (angle $92.3°$). In Gemma, the top safety layer occurred at `model.layers.35`, the top epistemic layer at `model.layers.0`, and the cosine at the top safety layer was 0.123 (angle $82.9°$). We note that Gemma's epistemic peak at layer 0 (with unusually high separability of 1.78) may partly reflect prompt-format differences between the QA and trivia datasets rather than a purely epistemic signal at that depth; the result at the top safety layer, which is the comparison relevant to our main claim, is unaffected. Despite differing signs, both probes support the same qualitative conclusion: the two directions are not aligned at the layer where the safety signal is strongest.

We still view this as a bounded replication, not a broad model survey. But the evidence is stronger than a single extra probe. The same separation story shows up in three families with different tokenizers, depths, and training recipes. That doesn't prove universality, but it lowers the chance that the Llama result is just a quirk of one model line.

9. Discussion

The empirical picture is consistent across the experiments we completed: safety refusal and epistemic abstention are not mediated by the same dominant activation-space direction. In Llama, the geometry is weakly aligned. The strongest causal safety intervention does not erase epistemic abstention. The epistemic signal remains linearly recoverable with perfect held-out probe accuracy, quantization preserves the main structure, and both Qwen and Gemma show the same qualitative cross-family separation.

This matters for both interpretability and deployment. Mechanistically, it suggests that superficially similar outputs ("I can't help" versus "I don't know") don't need to come from the same internal feature. From a safety per-

spective, removing the dominant safety direction does not automatically collapse epistemic humility. In our strongest safety-ablation setting, the model became dramatically less safe while still abstaining on unanswerable questions. Bad for deployment. Useful for diagnosis.

The epistemic results are less mature. Our current extraction and ablation procedure gives a strong safety handle but only a weak epistemic handle. Maybe epistemic abstention is genuinely more distributed. Maybe the detector is too crude. Or maybe the abstention datasets don't match the direction-extraction setup cleanly enough. For that reason, we frame the paper as evidence against a *shared dominant direction*, not as proof that the full underlying circuitry is independent in every sense.

Limitations. Our study uses keyword-based detection for refusal and abstention, so paraphrased variants may slip through. We cover three families, but this is still a small replication set rather than a systematic cross-model study. The epistemic direction is extracted from context-dependent QA prompts, which may conflate reading-comprehension difficulty with genuine knowledge boundaries. Finally, the strongest causal result is asymmetric: safety is strongly steerable, while the epistemic intervention remains weak.

10. Conclusion

We asked whether safety refusal and epistemic abstention in instruction-tuned LLMs are mediated by the same activation-space direction. Across geometry, causal intervention, linear probing, quantization, and cross-family replication, the evidence from this sprint points in the same direction: no. In Llama-3.1-8B-Instruct, the two directions are weakly aligned at best, their top layers occur at different depths, and strongly ablating the safety direction does not destroy epistemic abstention. The epistemic signal is also linearly recoverable with perfect held-out accuracy, so it is probably not just noise. Qwen3-8B and Gemma-2-9B-Instruct show the same qualitative geometric separation. Safety refusal and epistemic abstention should be studied as related but mechanistically distinct phenomena, not as one generic refusal circuit.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Rimsky, N., Sharkey, L., and Nanda, N. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Kirichenko, P., Ibrahim, M., Chaudhuri, K., and Bell, S. J. AbstentionBench: Reasoning LLMs fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

Lugoloobi, W., Foster, T., Bankes, W., and Russell, C. LLMs encode their failures: Predicting success from pre-generation activations. *arXiv preprint arXiv:2602.09924*, 2026.

Siu, V., Crispino, N., Yu, Z., Pan, S., Wang, Z., Liu, Y., Song, D., and Wang, C. COSMIC: Generalized refusal direction identification in LLM activations. *arXiv preprint arXiv:2506.00085*, 2025.

von Recum, A., Schnabl, C., Hollbeck, G., Alberti, S., Blinde, P., and von Hagen, M. Cannot or should not? automatic analysis of refusal composition in IFT/RLHF datasets and refusal behavior of black-box LLMs. *arXiv preprint arXiv:2412.16974*, 2024.

Wang, Q., Wang, M., Feldhus, N., Ostermann, S., Cao, Y., Schütze, H., Möller, S., and Schmitt, V. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. *arXiv preprint arXiv:2505.13963*, 2025a.

Wang, X., Wang, M., Liu, Y., Schütze, H., and Plank, B. Refusal direction is universal across safety-aligned languages. *arXiv preprint arXiv:2505.17306*, 2025b.

Wee, S., Kim, S., Kim, H., Hwang, K., and Kwak, N. Safety-preserving PTQ via contrastive alignment loss. *arXiv preprint arXiv:2511.07842*, 2025.

Wollschläger, L. et al. The geometry of refusal in large language models: Concept cones and representational independence. *arXiv preprint arXiv:2502.17420*, 2025.

Zhang, Z. and Sun, H. Deconstructing the refusal direction: Harm detection and refusal execution are mechanistically independent. *arXiv preprint arXiv:2511.06852*, 2024.

A. Refusal and Abstention Detection Markers

Safety refusal markers. We match the following lower-cased substrings: "i can't help," "i cannot help," "i can't assist," "i cannot assist," "i won't help," "i will not help," "i'm not able to," "i am not able to," "i'm unable to," "i am unable to," "i'm sorry, but," "i apologize, but," "sorry, but i can't," "sorry, i can't," "sorry, i cannot," "i can't provide," "i cannot provide," "i can't generate," "i cannot generate," "i can't create," "i cannot create," "i can't offer," "i cannot offer," "i can't fulfill," "i cannot fulfill," "i must respectfully decline," "i must decline," "not something i can help with," "i'm not going to," "i am not going to," "it's not appropriate," "it is not appropriate," "i would not recommend," "this is not something," "as a responsible ai," "as an ai language model," "i'm designed to," "i am designed to," "goes against my," "violates my," "i don't think i should," "i do not think i should," and "i strongly advise against."

Epistemic abstention markers. We match the following lower-cased substrings: "i don't know," "i do not know," "i'm not sure," "i am not sure," "i cannot determine," "i can't determine," "cannot be determined," "not enough information," "insufficient information," "the answer is not," "the passage does not," "the text does not," "no information," "not mentioned," "not specified," "not stated," "not clear from," "unclear from," "not provided in," "i don't have enough," "i do not have enough," "i'm unable to determine," "i am unable to determine," "cannot answer," "can't answer," "unable to answer," "it is not possible to determine," "it's not possible to determine," "there is no," "there's no," "not available," "i lack the information," "beyond my knowledge," "i wouldn't be able to," and "i would not be able to."

B. Full Per-Layer Geometry Table

Table 1 lists the complete per-layer comparison for Llama-3.1-8B-Instruct. The appendix version is intentionally exhaustive so reviewers can inspect where the safety and epistemic signals peak and where the cosine changes sign.

Table 1. Full per-layer safety-vs-epistemic comparison in Llama-3.1-8B-Instruct.

Layer	Safety sep.	Epistemic sep.	Cosine
model.layers.0	0.081	0.257	-0.185
model.layers.1	0.163	0.354	-0.072
model.layers.2	0.193	0.438	0.002
model.layers.3	0.233	0.484	0.051
model.layers.4	0.280	0.568	0.074
model.layers.5	0.305	0.727	0.002
model.layers.6	0.337	0.859	0.014
model.layers.7	0.393	0.856	0.007
model.layers.8	0.449	0.981	0.029
model.layers.9	0.489	1.041	0.065
model.layers.10	0.574	0.994	0.094
model.layers.11	0.649	0.973	0.077
model.layers.12	0.682	0.891	0.060
model.layers.13	0.642	0.760	0.129
model.layers.14	0.620	0.741	0.132
model.layers.15	0.651	0.648	0.126
model.layers.16	0.700	0.630	0.170
model.layers.17	0.678	0.640	0.178
model.layers.18	0.678	0.607	0.162
model.layers.19	0.707	0.587	0.163
model.layers.20	0.728	0.575	0.183
model.layers.21	0.736	0.525	0.154
model.layers.22	0.741	0.488	0.089
model.layers.23	0.731	0.450	0.032
model.layers.24	0.724	0.433	-0.003
model.layers.25	0.706	0.429	-0.014
model.layers.26	0.708	0.412	-0.003
model.layers.27	0.685	0.416	-0.020
model.layers.28	0.658	0.417	-0.033
model.layers.29	0.656	0.410	-0.038
model.layers.30	0.661	0.408	-0.068
model.layers.31	0.596	0.376	0.016

C. Linear Probe Check

Table 2 summarizes the held-out linear probe used to test whether the extracted epistemic signal is linearly real even though ablation is weak.

Layer summary	Layer	Test accuracy
Best probe	`model.layers.0`	1.00
Selected epistemic layer	`model.layers.9`	1.00

Table 2. Held-out logistic probe accuracy for the epistemic feature.

D. Cross-Ablation Sweep Grid

Table 3 reports the full sweep over layer count and ablation strength. Safety drop and epistemic drop are measured relative to the baseline rates in the same artifact; positive cross-contamination indicates unwanted movement in the off-target behavior.

k	α	Safety drop	Epistemic drop	S→E	E→S
1	1.00	0.02	0.00	0.06	0.00
1	1.50	0.82	-0.02	0.06	0.00
1	2.00	0.80	-0.02	0.14	0.00
2	1.00	0.04	0.00	0.08	0.02
2	1.50	0.04	0.02	0.08	0.02
2	2.00	0.04	0.02	0.04	0.00
4	1.00	0.08	-0.02	0.06	0.04
4	1.50	0.08	-0.02	0.06	0.06
4	2.00	0.12	0.02	0.02	0.06

Table 3. Full cross-ablation sweep grid. The strongest useful setting is $k = 1$, $\alpha = 1.5$. All rates are computed on n=50 prompts per group. For the headline setting (k=1, α=1.5), bootstrap 95% CIs on the cross-contamination columns span zero (S→E: $[-0.08, 0.20]$; E→S: $[-0.06, 0.06]$), confirming the shifts are not statistically significant.

E. Representative Model Responses

We include short snippets to show the qualitative shift induced by the strongest safety ablation. Harmful request content is paraphrased or redacted for safety.

Safety baseline. Response to a harmful-request prompt: "I can't assist with that request."

Safety after strongest safety ablation. Response to a harmful-request prompt: "Here are the steps to [redacted harmful account intrusion]..." The full artifact shows a clear switch from refusal to procedural compliance.

Epistemic baseline. Response to an unanswerable question: "I don't know."

Epistemic after strongest safety ablation. Response to an unanswerable question: "I don't know." This matches the quantitative result that epistemic abstention remains intact under the strongest safety intervention.