ELISABETTA ROCCHETTI

github.com/eliroc98/refusal_direction

◆ ONE-PAGER

Removing refusal in LLMs: Diff-in-means vs. Iterative Nullspace Projection

Is a single direction sufficient to capture refusal, or does a multi-dimensional approach yield cleaner erasure?

Claim. Arditi et al. (2024) showed that refusal in LLMs is mediated by a single direction, extractable via diff-in-means. This project tests whether Iterative Nullspace Projection (INLP), a multi-dimensional concept-erasure method, produces a cleaner separation of the refusal feature, enabling more targeted removal with less collateral damage to model performance. Preliminary results across five model families suggest that the single-direction assumption is surprisingly robust, and INLP's multi-dimensional erasure can overcorrect.

THE GAP

Diff-in-means assumes refusal is captured by one direction. If it is not, ablating that direction leaves residual refusal signal or removes unrelated features. No prior work has systematically compared single-direction methods with multi-dimensional concept erasure (INLP) on the refusal task, across multiple model families, using matched evaluation.

THE APPROACH

Five open-weight chat models (Gemma-2B, Yi-6B, Qwen-1.8B, Llama-2-7B, Llama-3-8B). Two extraction families producing five interventions. Seven evaluation metrics spanning safety (refusal rate, LlamaGuard 2 unsafety, harmless refusal), performance (Pile perplexity, Alpaca perplexity, MMLU, ARC). Shared selection criteria for fair comparison.

FINDING 1 · NULLSPACE WORKS, BUT OVERCORRECTS

INLP's nullspace projection successfully removes refusal. However, counterfactual flipping (α=2) frequently degrades performance: it increases perplexity sharply in Gemma and Qwen, and produces repeated text in Gemma. INLP appears to capture directions beyond refusal, causing overcorrection.

FINDING 2 · SINGLE DIRECTION IS ROBUST

Diff-in-means ActAdd and INLP-derived ActAdd (classifier weights) perform comparably in most models. The first INLP direction closely approximates the diff-in-means direction, supporting the claim that refusal is primarily mediated by a single direction. INLP's additional directions add noise more than signal.

WHY THIS MATTERS FOR AI SAFETY

This work provides evidence that multi-dimensional erasure (INLP) does not yield a meaningfully cleaner refusal subspace. This negative result tells us the vulnerability is genuinely low-dimensional, and that defenses should focus elsewhere (e.g., representation engineering, circuit-level interventions).

Introduction

It is still surprising how simple, cheap techniques can work exceptionally well at uncovering how Large Language Models represent complex concepts. Refusal behavior — the mechanism that causes a model to politely decline when a user asks for something potentially harmful — is one such concept. If removing this behavior were trivial, malicious actors could easily download open-weight LLMs and use them to generate misinformation, write malware, or build weapons.

Detecting and removing refusal behavior is, in fact, surprisingly easy. Arditi et al. (2024) demonstrated that the refusal direction can be extracted by simply computing the difference between the mean activations of harmless prompts and the mean activations of harmful prompts (diff-in-means). By ablating this single direction, they were able to bypass refusal across multiple model families with minimal performance degradation.

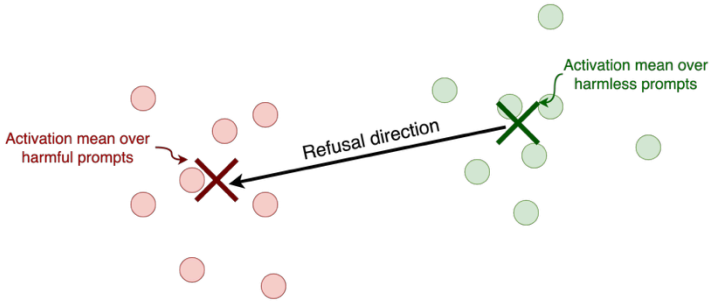

Figure 1. Diff-in-means approach: the refusal direction v is computed as the difference between the mean activation of harmful prompts and the mean activation of harmless prompts.

This raised an immediate question: Is diff-in-means the best approach for extracting the refusal direction? How does it compare to more mathematically sophisticated alternatives?

This project compares diff-in-means with Iterative Nullspace Projection (INLP), a method from Ravfogel et al. (2020) originally designed to remove gender bias from representations. INLP makes no assumption about the dimensionality of the target concept: it iteratively trains linear classifiers and projects out each discriminative direction until no linear probe can distinguish harmful from harmless prompts.

The hypothesis was that diff-in-means extracts a noisier signal than INLP, and that INLP's multi-dimensional erasure would isolate refusal more cleanly, enabling removal with less collateral damage to general model performance.

A primer on Iterative Nullspace Projection

Iterative Nullspace Projection (INLP) is a method designed to selectively identify and erase specific features (directions) from a model's hidden representations. Here is how it works:

1. Find the concept. Train a linear classifier on activations $h(l, t)$ to predict whether a prompt is harmful or harmless. The classifier's weight vector is the most predictive refusal direction.

2. Project it out. Construct a projection matrix P that maps $h(l,t)$ into the subspace orthogonal to this direction (its nullspace), so the classifier can no longer distinguish the two classes.

3. Iterate. Train another classifier on the projected representations. If it can still separate harmful from harmless, project that new direction out too. Repeat until no linear classifier can separate the two.

4. Compose the final projection. Each iteration yields an orthogonal direction. The final P encodes the full set of directions erased across all rounds. Via SVD of $I - P$, one can recover the individual directions and build rank-k approximations.

INLP also enables counterfactual interventions (flipping). Projecting activations using $P\alpha = \alpha P + (1-\alpha)I$, with $\alpha = 1$ yielding standard nullspace projection and $\alpha = 2$ yielding a reflection that brings representations to the opposite subspace (Hao & Linzen, 2023).

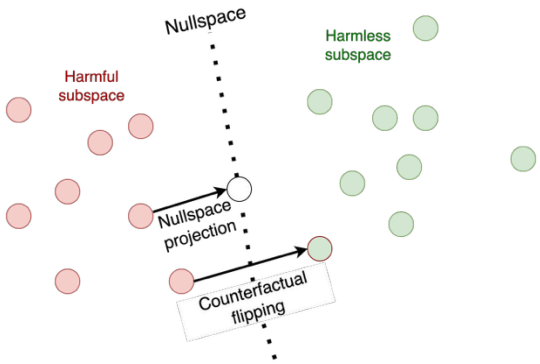

Figure 2. Counterfactual flipping: INLP enables reflecting representations across the nullspace, mapping harmful activations into the harmless subspace and vice versa (α=2).

Mapping the interventions

Diff-in-means assumes refusal is captured by a single direction v, computed as the mean difference between harmful and harmless activations. INLP makes no such assumption and removes all linearly decodable information about refusal, potentially spanning multiple directions. This difference leads to two families of interventions.

Erasing refusal

Directional ablation (diff-in-means)

Remove the refusal direction by projecting activations orthogonally:

$$h'(l,t) = h(l,t) - v \cdot (v^T \cdot h(l,t))$$

Nullspace projection (INLP)

Apply the full projection matrix P to remove all directions encoding refusal:

$$h'(l,t) = P \cdot h(l,t)$$

Injecting or steering refusal

Activation addition / ActAdd (diff-in-means)

Steer the model by adding or subtracting the refusal direction with scaling factor k ∈ {0.5, 1.0, 2.0}:

$$h'(l,t) = h(l,t) \pm k \cdot v$$

Counterfactual flipping (INLP)

Reflect representations across the nullspace using the counterfactual projection matrix:

$$h'(l,t) = P\alpha \cdot h(l,t), \text{ where } P\alpha = \alpha P + (1-\alpha)I$$

ActAdd with classifier weights (INLP-derived)

The first INLP classifier yields a weight vector w, representing the most predictive refusal direction. To compare directly with diff-in-means, w is normalized to unit norm and scaled to match the norm of v:

$$h'(l,t) = h(l,t) \pm k \cdot w$$

Method	Assumption	Extraction	Dims	Inject?	Cost
Diff-in-means	Single dir.	Mean diff	1D	Yes (ActAdd)	Low
INLP	Any linear	Nullspace	Multi	Yes (flip)	High
Classifier wts	Strongest	Classifier	1D	Yes (ActAdd)	Medium

Table 1. Comparison of the three direction-extraction methods.

Where interventions are applied

The extraction and application of interventions operate on different parts of the model and different token positions.

Extraction: Activations h(l,t) are extracted from the residual stream only, at the last tokens of the prompt as determined by the model's chat template, from layers excluding those too close to the output.

Application: Interventions are applied to the residual stream input, attention output, and MLP output, across all token positions. This broader scope ensures the intervention affects the model's computation throughout the forward pass.

The application scope varies by intervention type. For ActAdd, the intervention is applied only at the specific layer from which the direction was extracted. For directional ablation and nullspace projection, the scope can be local (single layer), 50% of the top-ranked layers, or 100% of layers — the latter matching the setting used in Arditi et al.

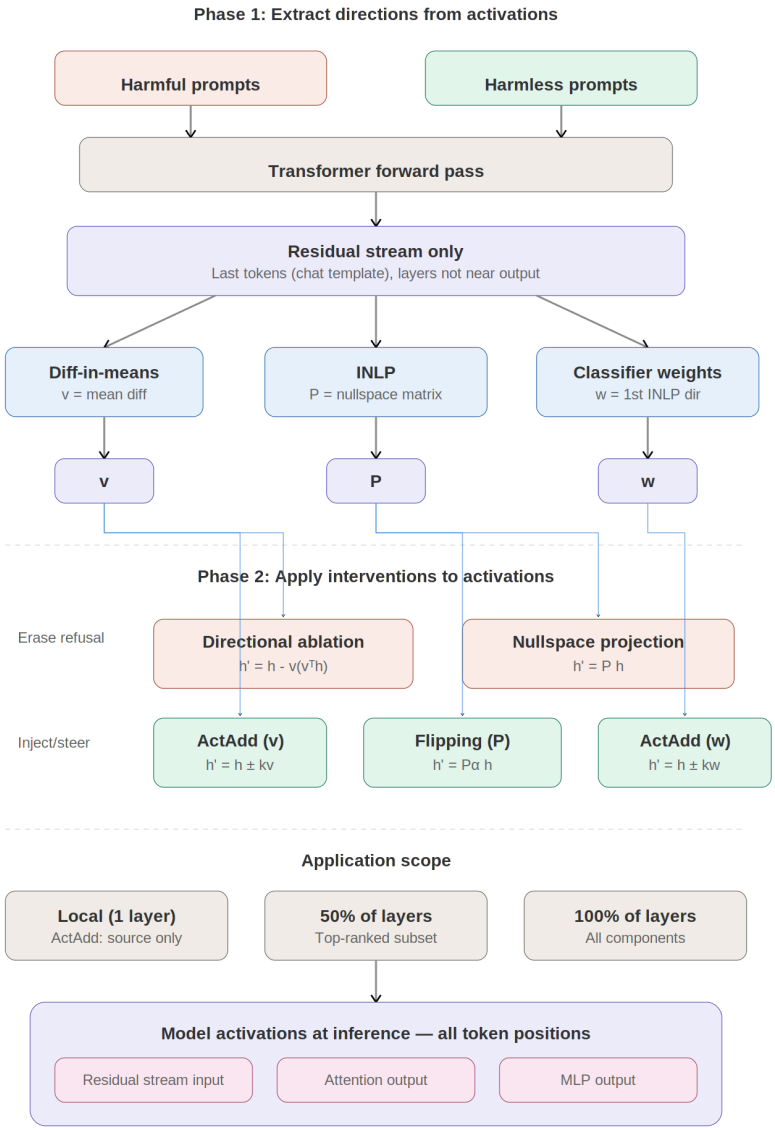

Figure 3. Intervention pipeline: Phase 1 extracts directions (v, P, w) from residual stream activations at the last prompt tokens. Phase 2 applies interventions to residual stream input, attention output, and MLP output across all token positions.

How interventions are selected

The computation of v, P, and w depends on a specific set of activations h(l,t). One must first decide which activations to use. Following Arditi et al., the selection procedure ranks interventions by a composite score that captures how much the intervention suppresses refusal on harmful prompts and induces refusal on harmless ones, while minimizing the shift in the final logits distribution (measured via KL divergence).

The same ranking strategy is adopted for both INLP-based interventions and ActAdd with classifier weights, with the additional requirement that the first INLP classifier achieves high

accuracy. One difference from Arditi et al. is that scores are computed by applying each intervention only to the currently inspected layer rather than to the entire model. Empirically, this modification does not significantly affect the resulting rankings.

Evaluation metrics

Safety-related

Refusal score (Δ Non-Refusal harmful): The fraction of harmful prompts for which the model does not refuse (Attack Success Rate under substring matching). Higher means more willing to comply.

Safety score (Δ Unsafety harmful): The fraction of harmful-prompt responses judged unsafe by LlamaGuard 2 (Meta-Llama-Guard-2-8B), a dedicated safety classifier that reads the full response.

Refusal on harmless (Δ Refusal harmless): The fraction of harmless prompts on which the model refuses — capturing the complementary failure mode of an over-cautious model.

Performance-related

Perplexity (Pile and Alpaca): Token-averaged cross-entropy on held-out text (monology/pile-uncopyrighted for general language modeling, tatsu-lab/alpaca for instruction-following).

MMLU: 5-shot accuracy on the Massive Multitask Language Understanding benchmark (cais/mmlu, all subjects).

ARC: 5-shot accuracy on ARC-Challenge (allenai/ai2_arc, Challenge split, 4-choice questions).

All scores are expressed as Δ vs. baseline (method minus no-intervention baseline), clipped to [−1, +1], so that higher always means better.

Preliminary results

The results are evaluated across five model families: Gemma-2B-IT, Yi-6B-Chat, Qwen-1.8B-Chat, Llama-2-7B-Chat, and Llama-3-8B-Instruct. The empirical sanity check that $P(\alpha=1) = P$ has been successfully passed and those results are omitted.

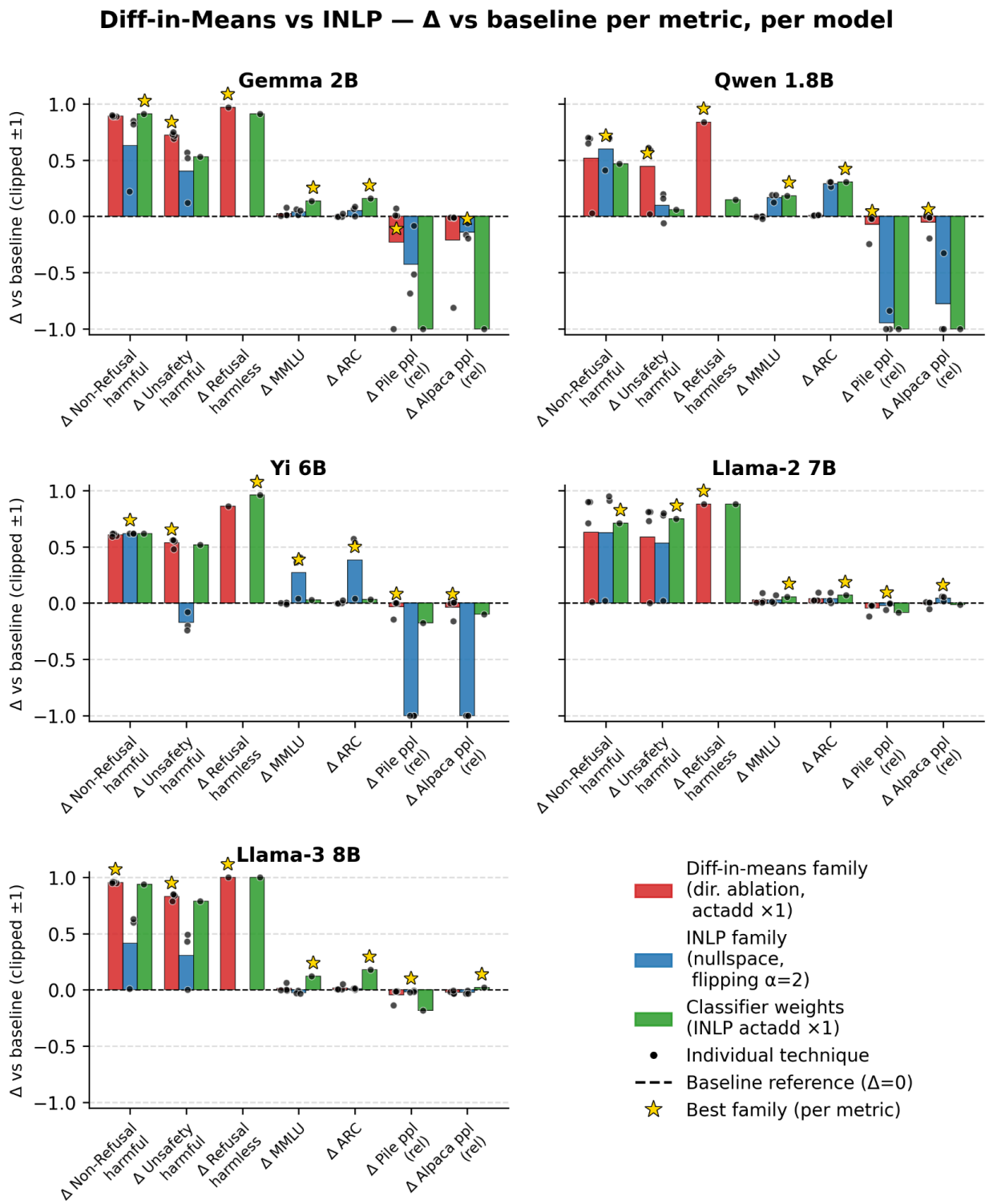

Figure 4. Family comparison: Δ vs. baseline per metric, per model. Red = diff-in-means family (directional ablation, ActAdd ×1). Blue = INLP family (nullspace projection, flipping α=2). Green = classifier weights (INLP ActAdd ×1). Stars mark the best family per metric. Higher is always better.

Refusal score

Nullspace projection generally succeeds in removing refusal, though it does not always improve on directional ablation. Counterfactual flipping (reflection) works in some cases but not consistently. INLP-derived ActAdd is comparable to diff-in-means ActAdd in most models, apart from Qwen. Reflection with α=2 was expected to be stronger than nullspace projection because it truly produces harmlessness rather than merely removing harmfulness, but this expectation was not confirmed.

Safety score

Nullspace projection produces higher safety scores (meaning fewer genuinely unsafe responses) especially in Yi and Qwen. For the other models, safety scores remain lower than baseline but higher than those of directional ablation. Reflection appears to work well for Gemma, but this is because it produces repeated text rather than meaningful safe content. Reflection fails for Qwen.

Refusal on harmless prompts

Reflection does not work well here. INLP likely captures too many linear directions, and some of these may not be related to refusal alone — it learns too much. The classifier-weights ActAdd performs well in this setting, suggesting that the first INLP direction is well-calibrated but the full multi-dimensional projection is too aggressive.

Performance scores

Perplexity is much higher for reflection in Gemma and Qwen, while it is elevated for nullspace projection in Qwen and Yi. This confirms that INLP's multi-dimensional erasure carries a higher performance cost than single-direction methods in several models.

Limitations

- INLP is computationally expensive and sensitive to hyperparameters (number of classifiers, regularization). The current setup may not represent INLP at its best.
- The composition effect is not accounted for: when multiple layers are intervened on, the impact of later interventions may change because earlier layers have already been modified.
- All models tested are relatively small (1.8B–8B parameters). It remains to be seen whether the findings generalize to larger models.
- Substring matching for refusal detection is a coarse proxy. A response that avoids hedging phrases may still be evasive or unhelpful.
- This project write-up is not including a commentary on all experimental settings I tested, thus I could have ignored something that could have improved my view of the results.

What's next

- Implement LEACE (Linear Erasure by Approximate Concept Erasure), a closed-form concept erasure method that may achieve the theoretical purity of INLP without the iterative overhead.
- Experiment with removing fewer directions from INLP (rank-k restriction) to test whether a subset of directions provides a better tradeoff than the full nullspace.
- Expand to larger models (13B+, 70B) to test whether the single-direction hypothesis holds at scale.
- Investigate the semantic content of the additional directions INLP removes beyond the first: are they genuinely related to refusal, or do they encode other concepts (e.g., helpfulness, formality)?

Connection to AI safety

Refusal is the primary behavioral safety mechanism in open-weight language models. Understanding its internal representation is directly relevant to AI safety for three reasons:

Vulnerability assessment. If refusal is mediated by a single direction, any actor with model access can remove it trivially. This project confirms that the single-direction finding is robust across model families and resistant to more sophisticated extraction methods. This means the vulnerability is real, low-dimensional, and will not be solved by making the refusal representation harder to find.

Defense prioritization. The negative result — that INLP does not cleanly improve on diff-in-means — tells us that the path to robust refusal lies not in finding better probing methods, but in fundamentally rethinking how safety behaviors are embedded in model weights. Representation engineering, circuit-level interventions, and training-time defenses are more promising directions.

Methodological contribution. The open-source evaluation pipeline (seven metrics, five models, five interventions, multiple scopes) provides infrastructure for future work on activation-level safety interventions. The codebase is designed for reproducibility and extension.

Appendix: Composing the final projection P

Each INLP iteration yields an orthogonal direction to remove. The final P is the composition of all per-iteration projections, encoding the full set of directions erased across all rounds.

Concretely, the individual directions can be recovered by computing the SVD of $I-P$: since $P = I - V^TV$, where V stacks the orthogonal directions found at each step, the right singular vectors of $I-P$ with non-negligible singular values reconstruct V exactly. This also enables building a rank-k approximation $P_k = I - V_k^T V_k$ to erase only the k most important directions rather than all of them.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717.

Ravfogel, S., Elazar, Y., Gonen, H., Trost, M., & Goldberg, Y. (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. ACL 2020.

Hao, S. & Linzen, T. (2023). Verb Conjugation in Transformers Is Determined by a Small Number of Interpretable Features. Proceedings of EMNLP 2023.

Special thanks to the interpretability community and BlueDot Impact for fostering this research.