

Legible but Not Localized: Jailbreak Geometry in Mixture-of-Experts Language Models Is Attack-Specific

Yubin Im¹, Myeonghyun Kim¹, Seongmin Kim², Haehyun Cho^{1*}

¹ Soongsil University, ² Sungshin Women's University

¹ {yubin, devmhyun}@soongsil.ac.kr, haehyun@ssu.ac.kr

² sm.kim@sungshin.ac.kr

Abstract

Activation-based jailbreak defenses rest on a fixed assumption: a successful jailbreak leaves an internal trace stable enough across attacks for one probe to catch new ones. We show it fails in mixture-of-experts language models. Within an attack, routing makes the jailbreak legible: router logits distinguish compliance from refusal, and a few experts carry the signal. But the signature does not survive across attacks. A routing probe trained on three attacks performs at chance on a held-out fourth (pooled leave-one-attack-out balanced accuracy 0.52, 95 percent CI [0.50, 0.61]). Controls on two dense models show near-chance residual-stream transfer to held-out roleplay, so the failure is not specific to MoE routing. Nor is it an artifact of separable regions of activation space: the comply-versus-refuse direction itself shifts across attacks at every network depth. One systematic trace remains in the compliance state's effective dimensionality, but it is a signature of shape, not location, beyond directional probes. Across six open-weight MoE models and four jailbreak attacks, in routing and residual streams alike, the jailbroken state is legible and locally structured within each attack, yet its location is attack-specific. Reading activations alone cannot be assumed to catch a jailbreak the probe was not built for.

1 Introduction

In a dense transformer, refusal is mediated by a single direction in activation space (Arditi et al., 2024), or by a low-dimensional cone of directions (Wollschläger et al., 2025); either way it is localized to a small, fixed subspace and easy to read. A growing line of safety work (Zhang et al., 2025; Dong et al., 2025; Kadali and Papalexakis, 2025; Jiang et al., 2025; Zhou et al., 2026) treats

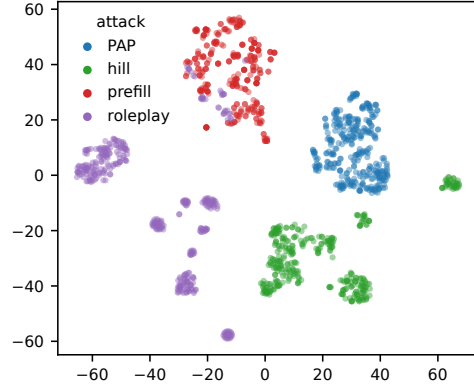


Figure 1: The jailbroken state is attack-specific in MoE routing. Each point is one prompt of llama-4-scout, t-SNE-projected from its router logits at every MoE layer (48×16 experts); the four attacks form distinct regions. Section 4 shows the same pattern across the panel.

this as a defense template, reading a model's activations to detect unsafe states and placing such monitors directly in the generation loop (Goren et al., 2026; Hu et al., 2025). The template rests on a fixed assumption: that the internal trace of a successful jailbreak is stable enough across attacks for one probe, trained on the attacks known at build time, to catch new ones. We ask whether mixture-of-experts language models carry such a stable trace, and find that they do not. Within a single attack the routing exposes the jailbreak: the router logits decode compliance from refusal, and a few experts carry the signal. But across attacks the comply-versus-refuse direction itself shifts at every network depth, and a routing probe trained on three attacks performs at chance on a held-out fourth, while residual-stream transfer remains weak and attack-dependent.

We look for the signature in the routing of a mixture-of-experts model. A mixture-of-experts layer sends each token through a small, chosen subset of its experts, so its routing is a low-dimensional, discrete record of how the model

*Corresponding author.

The code and data are available at <https://github.com/ssu-csec/moe-jailbreak-geometry>

chose to process the token. If a model carried a refusal mechanism localized to a stable set of experts, the routing is where its trace would be easiest to read. We study mixture-of-experts models for that reason, as the most favorable substrate for a localized mechanism rather than as a claim about the architecture. Routing is a lower-dimensional view of the computation that also runs through the residual stream; the two reach the same verdict throughout this paper (Section 4). To test whether the clearest non-transfer result can arise without expert routing, we repeat the residual-stream leave-one-attack-out analysis on two dense controls, Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. In both models, transfer to held-out roleplay remains near chance, whereas transfer is higher for PAP and hill when the related persuasion attack remains in the training pool. Thus, the roleplay result is not specific to MoE architecture: routing makes attack-specific instability especially legible, but is not required for it. MoE routing is input-sensitive by construction, so wrapper-level routing differences across attacks are expected; the non-trivial test is whether the within-attack comply-versus-refuse direction itself shifts, which we isolate from the per-attack offset via centering (Section 4.2).

The non-transfer is not the trivial consequence of attacks being separable in activation space; a residual-stream probe, a representational-similarity analysis, and a cross-attack re-identification test all reach the same verdict. Roleplay decodes well in every model within its own attack yet is not caught across attacks at all, a clean negative; transfer occurs in only one case, between the two persuasion attacks, and even there it is marginal.

One property of the jailbroken state does move with the attack systematically: its effective dimensionality, the number of directions the activations spread across. The compliance state is high-dimensional under the persuasion attacks, between 25 and 31 effective dimensions, and collapses under prefill and roleplay, to 4 and 8; the refusal state stays much more stable. But effective dimensionality is a statistic of the state, not a direction a probe can be trained on. We establish these results on six open-weight mixture-of-experts models and four jailbreak attacks.

In this paper, we make three contributions.

- Using mixture-of-experts routing as the most favorable substrate for a localized refusal mechanism, we show that within an attack the routing

exposes the jailbreak: the router logits decode compliance from refusal, and a few experts carry the signal.

- We show this signature does not transfer across attacks: pooled over held-out attacks, a routing probe performs at chance, while residual-stream transfer remains weak and attack-dependent, not because the attacks are merely separable but because the comply-versus-refuse direction itself differs at every network depth. Residual-stream controls on two dense models reproduce the clearest failure on held-out roleplay, showing that this result is not specific to expert routing.
- We give a structural characterization of the jailbroken state: the attack reshapes the compliance state’s *shape* (its effective dimensionality varies by an order of magnitude across attacks) without ever sharing its *location*, leaving a trace that is real but in a form a probe cannot exploit.

2 Background and Related Work

Jailbreaks and the two states. A safety-aligned language model that receives a harmful request either refuses it or complies with it. We call the residual-stream activation at the final prompt token that accompanies a refusal the *refusal state*, and the activation that accompanies a compliance the *compliance state*. A jailbreak attack is an input that turns a refusal into a compliance. This paper asks where, inside a mixture-of-experts model, the jailbroken compliance state can be told apart from the refusal state, and whether that signature is the same across attacks.

Activation monitors and the refusal direction.

A line of safety work (Zou et al., 2023a; Zhang et al., 2025; Jiang et al., 2025) reads a model’s internal activations, rather than its text, to decide whether a request or response is unsafe. Such an activation monitor is a classifier, typically a linear or MLP probe on the residual stream, that flags an unsafe state from the representation alone (Patel and Wang, 2024; MacDiarmid et al., 2024). The template behind these probes is the finding of Ardit et al. (2024), that refusal in dense transformers is strongly associated with a single linear direction; subsequent work refines this as a low-dimensional cone or a not-fully-linear signal (Wollschläger et al., 2025; Pan et al., 2025; Hildebrandt et al., 2025). Recent jailbreak defenses place a monitor of this kind directly in the generation loop (Dong et al., 2025; Goren et al., 2026; Hu

et al., 2025; Kadali and Papalexakis, 2025). A monitor is built from the jailbreak attacks known when it is trained, then meets unseen variants in deployment; these works share an implicit assumption that the activation signature is stable enough across attacks for one detector to generalize, but none evaluates leave-one-attack-out transfer. We find this assumption fails even at the most favorable readout site we could identify. Our analysis targets *detection-side* generalization, whether a probe trained on known attacks catches unseen ones, which is orthogonal to whether the extracted direction is the *causal* refusal switch within a single attack. A related thread shows that an adaptive adversary can induce *obfuscated activations* that bypass an activation monitor (Bailey et al., 2026); we differ in two ways, in that our setting needs no adaptive adversary, just an unseen attack, and in that they probe input harmfulness while we probe the compliance state.

Mixture-of-experts routing. A mixture-of-experts (MoE) layer (Shazeer et al., 2017; Fedus et al., 2022) augments the usual feed-forward block with many parallel experts. A small gating network scores the experts for each token; we call these pre-softmax scores the *router logits*. The layer then routes the token through only the top-k experts, the *selected experts*, and combines their outputs. Unlike the residual stream’s distributed activations, routing makes expert allocation explicit: a per-expert score and a discrete set of selected experts, recorded at every layer. Routers in MoE transformers specialize, carrying information about syntax and task (Zoph et al., 2022; Jiang et al., 2024), which makes routing a natural place to look for a localized refusal signature: if any site concentrates the signal, the gates should. We test this premise across six open-weight MoE models: the concentration holds within an attack but not across.

3 Method

The analyses follow a single chain of questions. We first ask whether the routing carries the comply-versus-refuse decision at all, by decoding it within one attack; if it did not, there would be nothing for an attack to keep or to change. Given that it does, we ask whether the signal is sparse and localizable, by identifying which experts carry it. The central question of the paper follows: whether the comply-versus-refuse boundary is the same from one attack to the next. We answer that by transferring a probe

to attacks it was not trained on, one attack at a time and then in the leave-one-attack-out setting a deployed defense would face. Two analyses that use no probe, a representational-similarity analysis and a cross-attack re-identification test, then check the same invariance without a fitted classifier. Throughout, the residual stream is the reference the routing is read against, and a final measurement, effective dimensionality, asks which property of the jailbroken state does move with the attack.

Models and inclusion criterion. We study six open-weight mixture-of-experts language models (Liu et al., 2024a; Dai et al., 2024; Team, 2024; Jiang et al., 2024; Muennighoff et al., 2025), listed in Table 1. We restrict the panel to this family because its routing is the central object of our study, the most favorable place to look for a localized refusal mechanism; recent open-weight releases have also converged on the design. A model enters the panel only if it produced at least 30 refused and at least 30 complied responses for every one of the four attacks, enough to train and test a probe on both classes. Two further open-weight models fail this criterion and are excluded: gpt-oss-20b and phi-3.5-moe-instruct each produced only a handful of complied responses under prefill or roleplay. The criterion is data-driven, not result-driven: it depends only on sample counts, never on the size of any effect.

Dense-model controls. To test whether residual-stream transfer failure can also occur in models without expert routing, we evaluate Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. Both models satisfy the same inclusion criterion of at least 30 refused and 30 complied responses for every attack. We use the same per-model vanilla-refusal filter, four safety benchmarks, four attacks, documented template counts (4 PAP, 4 hill, 1 prefill, and 80 roleplay templates), and Claude Sonnet 4.6 labeling scheme as in the MoE panel. To bound generation and judging, we deterministically subsample at most 600 wrapped prompts per attack after the vanilla-refusal and template filters. For the transferred residual-stream probe, each class is subsampled to at most 200 examples, features are standardized and reduced to 256 principal components, and an L_2 -regularized logistic-regression probe is fit with balanced class weights. For each held-out attack, the train-selected readout layer maximizes mean five-fold within-attack balanced accuracy over the other three attacks using standardized residual-stream features; layer scoring does not

Model	Params	Active	Experts	Top- k	Shared
deepseek-v2-lite	15.7	2.4	64	6	2
deepseek-moe-16b-chat	16.4	2.8	64	6	2
qwen1.5-moe-a2.7b-chat	14.3	2.7	60	4	4
llama-4-scout	109.0	17	16	1	1
mixtral-8x7b	46.7	13	8	2	0
olmoe-1b-7b	6.9	1.3	64	8	0

Table 1: The six open-weight mixture-of-experts models in our panel. Params and Active are the total and per-token active parameter counts, in billions. Experts is the number of routed experts, Top- k the number activated per token, and Shared the number of always-on shared experts.

use PCA, and the held-out attack informs neither layer selection nor probe fitting. We additionally report fixed readouts at zero-indexed decoder-layer indices 16 and 29 of 32, corresponding approximately to 50% and 90% depth. Following the MoE analysis, pooled 95% confidence intervals use 4000 percentile-bootstrap resamples over the eight combinations of dense model and held-out attack. Because dense models have neither router logits nor expert selections, all control analyses use only the residual stream. The fixed-depth readouts are sensitivity checks rather than exact layerwise matches across architectures. For the within-roleplay control, we divide the sorted template set into two disjoint halves by alternating assignment, train on each half and test on the other, and report the mean of the two directions.

Stimuli and labels. We pool harmful requests from four public safety benchmarks: HarmBench (Mazeika et al., 2024), AdvBench (Zou et al., 2023b), StrongREJECT (Souly et al., 2024), and HarmEval (Banerjee et al., 2025). For each model we first keep only the requests the model itself refuses with no attack applied; the surviving requests are then wrapped in all four jailbreak attacks. Refusal is then the baseline for every model on every request it sees, so a model that complies under an attack has been genuinely jailbroken rather than answered a request it would have answered anyway. All four are natural-language jailbreaks, not optimization-based attacks; each acts only on the input prompt and assumes black-box access. Two of the attacks work by persuasion. *PAP* rewrites the request as a persuasive argument (Zeng et al., 2024). The academic-framing attack, which we call *hill*, wraps the request in scholarly language (Luo et al., 2025). The third attack, *prefill*, inserts the opening words of a compliant answer into the model’s response slot (Andriushchenko et al., 2025). The

fourth, *roleplay*, asks the model to answer in the voice of a permissive persona (Shen et al., 2024). Each attack is applied through its own set of templates: roleplay uses 80 templates, PAP and hill 4 each, and prefill a single fixed prefix (Appendix D). We collect each model’s response to the attack prompts and label every response with an automatic judge, Claude Sonnet 4.6. The judge sorts each response into a compliance, a refusal, or an unclear case, and unclear cases are discarded. For a given model and attack, the complied prompts form the compliance set and the refused prompts form the refusal set.

Routing extraction. For each prompt we record the model’s internal state at the last token of the prompt, where the model has read the whole attack but has not generated anything, so that no generated token can confound the measurement. At every mixture-of-experts layer we capture three signals: the residual stream, the full representation used as a reference; the router logits, one score per expert; and the selected experts, the discrete set the layer routed the token through. The router logits and selected experts are the routing; the residual stream is what it is read against.

Decoding the decision. To ask whether a signal carries the comply-versus-refuse decision, we fit an L_2 -regularized logistic-regression probe on standardized features and score it by five-fold cross-validated balanced accuracy. The primary probe analyses subsample each class to at most 200 examples, use balanced class weights, and reduce the residual stream to at most 256 principal components, capped further by the number of available training examples. For descriptive within-attack readouts, each model–attack cell reports the best cross-validated layer for each routing signal, searched over every MoE layer, and the best of three approximately evenly spaced residual-stream layers. At the selected router-logit layer, we report the 95th percentile of 15 label-shuffled runs as a descriptive reference, not corrected for layer selection. We say the router logits decode the decision when accuracy reaches 0.75, exceeds this reference, and lies within 0.10 of the residual-stream accuracy.

Cross-attack transfer. To ask whether the decoded signal is attack-invariant, we test transfer in two settings. In single-attack transfer we fit a probe on one attack and apply it, unchanged, to a different attack of the same model. In leave-one-attack-out transfer we pool three attacks, fit one probe, and test it on the held-out fourth. The sec-

and setting mirrors deployment, where a defense is trained on known attacks but evaluated on unseen variants. Single-attack transfer uses one routing-selected layer per model, chosen by maximizing mean five-fold cross-validated within-attack balanced accuracy across all four attacks. For the MoE leave-one-attack-out analysis, each held-out fold selects the layer from the other three attacks by the same criterion; the routing and residual-stream probes share this layer, so the held-out attack informs neither layer selection nor probe fitting. We report bootstrap 95% confidence intervals over the 24 held-out cells (4000 percentile-method resamples). As a separate geometric control, we center each attack, including the held-out attack, by its own unlabeled mean before re-running leave-one-attack-out transfer, thereby removing the per-attack offset. This control uses all available labeled examples; routing uses a layer selected from the three training attacks, whereas the residual stream is read at the fixed middle layer.

Depth and the decision direction. To test whether the picture depends on depth, we repeat the within-attack decode and leave-one-attack-out transfer throughout the network. We evaluate routing at every MoE layer and the residual stream at 12 approximately evenly spaced decoder-layer indices, using all available labeled examples in each cell rather than the capped subsamples of the primary analyses. After featurewise standardization over all four attacks and both classes at each layer, we read the comply-versus-refuse direction itself, the difference between the mean compliance and refusal activation (Arditi et al., 2024), and measure its cosine between attacks. Twenty label-shuffled reassignments at each evaluated residual-stream layer form the shuffled-label null; we report the 95th percentile of its absolute cosine as $|\cos|_{95}$, the threshold used in Figure 5b. Under our balanced-class setup, the L_2 -regularized logistic-regression probe’s decoder direction approximates this mean-difference direction, so a probe’s cross-attack transfer accuracy and a direction’s cross-attack cosine measure the same alignment from two angles.

Expert identity. Where the routing decodes the decision, we ask which experts carry it. For each expert e at the cell’s readout layer, we compute the signed Cohen’s $d_e = (\mu_R^{(e)} - \mu_C^{(e)})/\sigma_{\text{pooled}}^{(e)}$ of its router logit between the refusal and compliance sets, with positive values indicating refusal-favored experts; we use d_e to rank experts rather than for

absolute-magnitude inference. To measure sparsity, we use up to 300 examples per class, rank experts by this effect size on the training folds alone, and refit the probe on only the top few. To measure attack-invariance, we correlate the per-expert effect-size vectors between attacks at each model’s within-attack readout layer.

Representational similarity and re-identification. Two further analyses test attack-invariance without a probe. In a representational-similarity analysis we mean-center the residual activations within each attack, then compare how close a base request sits to itself under a different attack wrapper against how close it sits to unrelated requests. In a re-identification test we attack-center the activations and match each base request to its nearest neighbor across attack wrappers, reporting top-1 accuracy; a TF-IDF bag-of-words of the prompt text serves as a surface-token baseline.

Effective dimensionality. On the residual stream we also measure the effective dimensionality of each state, the number of directions a set of activations genuinely spreads across. We measure it with the participation ratio,

$$\text{PR} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}, \quad (1)$$

where the λ_i are the eigenvalues of the activation covariance $\Sigma = X^\top X/(n-1)$, with $X \in \mathbb{R}^{n \times d}$ stacking n mean-centered last-prompt-token residual-stream activations at the model’s readout layer, computed separately for the compliance and refusal sets. The participation ratio behaves like a count of directions: it equals k when the variance spreads evenly across k of them, and one when it sits in a single direction. We write PR_C for the compliance set and PR_R for the refusal set, and report each within every attack. A participation ratio grows with sample count, so we subsample both sets to a common size $n_0 = 120$ and report $\widehat{\text{PR}} = \mathbb{E}_S[\text{PR}(S)]$ averaged over 10 independent subsamples S , separately for PR_C and PR_R .

Why we compare within an attack. The four attacks occupy linearly separable regions of activation space (a classifier names the attack at 0.998 accuracy, against chance 0.25; Appendix A), so pooling would mix four distinct geometries; every comparison above is made within a single attack, with the leave-one-attack-out probe the deliberate exception that pools attacks on purpose to measure what a defender, who cannot choose to compare within an attack, actually faces.

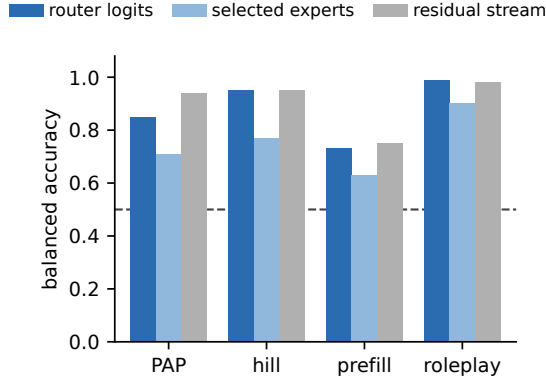


Figure 2: Within an attack, the routing decodes the jailbreak. Each bar is the median balanced accuracy across the six models of a probe that separates complied from refused prompts, by attack. The router logits decode nearly as well as the full residual stream, and the discrete expert selection is weaker but above chance. The dashed line is the chance level of 0.5.

4 Results

This section answers, in order, the chain of questions Section 3 set out. Within a single attack, the routing of a mixture-of-experts model carries the jailbreak: a probe on the router logits tells a complied prompt from a refused one almost as well as the full residual stream does. That signature does not generalize: pooled over held-out attacks, a routing probe trained on the attacks a defender holds sits at chance, while residual-stream transfer remains weak and attack-dependent. The failure is not the trivial consequence of the attacks occupying separable regions of activation space; the comply-versus-refuse direction is itself attack-specific, and it stays so at every network depth. The residual stream and two probe-free analyses reach the same verdict.

4.1 The router decodes the jailbreak within an attack

Within a single attack, the routing carries the comply-versus-refuse decision. For each model and attack we fit a logistic-regression probe on the router logits and score it by cross-validated balanced accuracy. The router-logit probe reaches a median balanced accuracy of 0.85 across the 24 model-attack cells, against a descriptive shuffled-label reference near 0.55 and a residual-stream reference of 0.90 (Figure 2). In 19 of the 24 cells the router logits decode the decision: the probe clears 0.75, exceeds this reference, and lands within 0.10

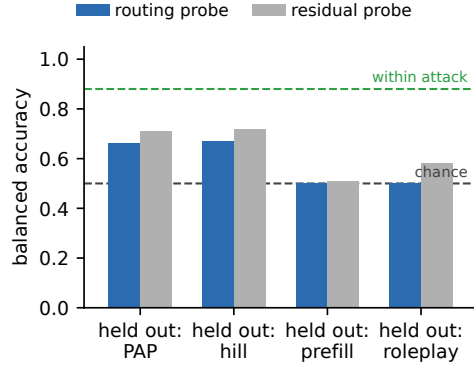


Figure 3: A probe trained on three attacks, tested on the held-out fourth. Each bar is the median balanced accuracy across the six models, by held-out attack, for a routing probe and a residual-stream probe. The two persuasion attacks (PAP, hill) are caught above the chance point estimate because the other persuasion attack stays in the training pool. Prefill and roleplay are caught at or near chance (lower dashed line); the upper dashed line is within-attack accuracy.

of the residual stream. In the five weaker router-logit cells (deepseek-moe-16b-chat under PAP, hill, and prefill; prefill on mixtral-8x7b and qwen1.5-moe-a2.7b-chat), the router-logit probe stays below 0.75. Four of the five are also weak in the residual stream; the exception is mixtral-8x7b under prefill, whose residual-stream accuracy reaches 0.77.

The signal is sparse. For each cell we rank the experts by the effect size of their router logit and refit the probe on only the top few. Three experts recover a median of 79% of the probe’s above-chance accuracy, though this varies widely across cells, from 52% to 100%. We use a fixed count of three rather than each model’s routing width, the Top- k in Table 1, which ranges from one to eight across the panel, so the sparsity measure is comparable across models; Appendix E reports the full sweep over the count. Where the routing decodes the jailbreak, a handful of experts usually carry most of it.

4.2 The signature does not generalize across attacks

The routing exposes the jailbreak within an attack, but the experts that carry the decision are largely not the same from one attack to the next. At a fixed layer, per-expert effect-size vectors of the two persuasion attacks correlate at a median $r = 0.70$ while prefill correlates far more weakly with the others; the over-selected expert set overlaps across attacks by a Jaccard index of only 0.27 to 0.59 on

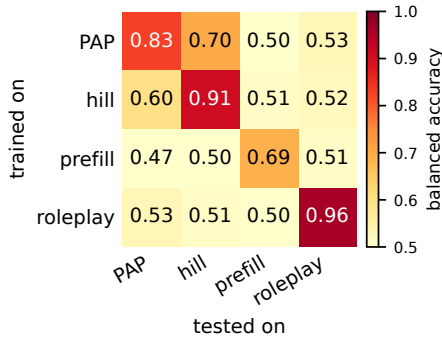


Figure 4: Cross-attack transfer of the routing probe. Each cell is the median balanced accuracy across the six models of a router-logit probe trained on the row attack and tested on the column attack; the diagonal is within-attack accuracy. Off the diagonal, accuracy falls to chance, except between the two persuasion attacks PAP and hill: separable attacks can still share a comply-versus-refuse direction.

the four models with 60 or more experts (the index saturates on mixtral-8x7b and llama-4-scout and is not informative there).

A first test is single-attack transfer: we apply a within-attack probe, unchanged, to a different attack of the same model (Figure 4), and off the diagonal its balanced accuracy falls from above 0.85 to a median of 0.51. This is not the trivial consequence of separable attack regions: two separable regions can still share one comply-versus-refuse direction, as PAP and hill demonstrate (a probe trained on one reaches 0.60 to 0.70 on the other, well above chance). What blocks transfer for the cross-family pairs is a distinct fact: their comply-versus-refuse direction is itself attack-specific, with the residual-stream directions of PAP and hill aligning at a cosine well above the shuffled-label null while every cross-family pair stays within that null near 0.05 (Section 4.3).

A defender does not deploy a probe trained on a single attack but trains on a pool known at build time. We test this with a leave-one-attack-out probe: for each model we pool three attacks, fit a probe, and test it on the held-out fourth, choosing the transfer layer from the training attacks alone (Figure 3). Pooled over all 24 held-out cells, the routing probe reaches a median balanced accuracy of 0.52 [0.50, 0.61] against a within-attack median of 0.85; the residual stream does only a little better at 0.62 [0.51, 0.73], remaining far below its within-attack median of 0.90, though its interval overlaps the within-attack shuffled-label null of 0.55.

The per-attack offset is large but not the cause: routing centroids move 1.3 to 1.9 times the within-attack class gap between attacks, so a held-out attack arrives as a cloud shifted beyond the gap the probe was trained to resolve. Subtracting each attack’s centroid before pooling co-locates the four clouds, yet the re-run probe lifts only by 0.03 on the routing and 0.05 on the residual stream. The non-transfer is not the per-attack offset; the comply-versus-refuse structure within each attack is itself attack-specific.

The held-out attack decides whether anything transfers: PAP and hill are caught at 0.66 [0.55, 0.83] and 0.67 [0.48, 0.77] because each leaves the other in the training pool, though hill’s interval includes chance and the transfer is suggestive rather than established. Roleplay decodes well in every model (0.80 to 1.00), yet the pooled routing probe catches it at chance, 0.50 [0.39, 0.52], the clean negative result; routing transfer to prefill is also at chance, 0.50 [0.49, 0.55], though this is weaker evidence because prefill barely decodes within its own attack. An MLP head leaves both pooled medians below 0.60. Fixed mid- and late-depth readouts raise transfer for PAP and hill—and the late residual-stream pooled median to 0.667—but leave prefill and roleplay near chance under every choice (Appendix C).

4.3 The signature is attack-specific at every depth

The non-transfer does not depend on which layer we read. We sweep the network, measuring at every layer the within-attack decode, the leave-one-attack-out transfer, and the comply-versus-refuse direction (Figure 5). The jailbroken state grows more legible with depth: in the residual stream the within-attack decode rises from a median of 0.79 over the first third of layers to 0.92 over the last. Its transfer across attack families does not follow. Leave-one-attack-out accuracy on prefill and roleplay holds near chance at every depth, rising only from 0.50 to 0.57 between the first third of layers and the last. The geometry of the comply-versus-refuse direction agrees: across attack families its cosine holds near zero through the whole network, while within the persuasion family it climbs from 0.18 to 0.56. The routing shows the pattern in flatter form, its within-attack decode roughly level with depth and its cross-family transfer near chance throughout. Depth sharpens the jailbroken state and aligns it within an attack family but does not

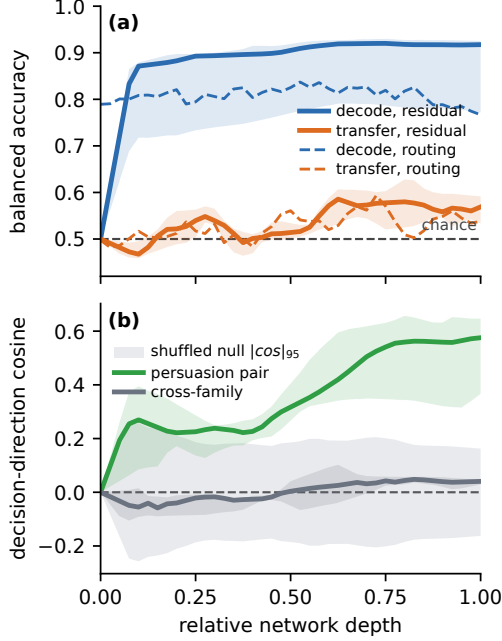


Figure 5: The jailbroken state versus network depth, pooled over the six models. (a) Within-attack decode and leave-one-attack-out cross-family transfer, for the residual stream and the routing: decode climbs with depth while cross-family transfer stays near chance. (b) Cosine of the comply-versus-refuse direction between attacks, in the residual stream: within the persuasion family it climbs with depth, across families it stays inside the shuffled-label $|cos|_{95}$ null band (gray) at every depth. Lines are medians over the six models; shaded bands are the interquartile range.

make it transfer across families; no layer carries an attack-invariant signature, and the gap between legibility and transfer only widens with depth.

4.4 The residual stream agrees, and the one trace the attack does leave

The routing is a low-dimensional, discrete view of a computation that also runs through the residual stream, and the residual stream reaches the same verdict. A residual-stream probe decodes the decision within an attack, and it too fails to generalize to a held-out attack, at a leave-one-attack-out median of 0.62.

Dense controls reproduce the held-out roleplay failure. At the fixed mid-depth sensitivity readout, the two dense controls yield a pooled median residual-stream leave-one-attack-out balanced accuracy of 0.632 [0.521, 0.711], compared with 0.585 [0.50, 0.66] for the MoE residual stream under its analogous fixed mid-depth readout; the bootstrap intervals overlap. The pooled dense result

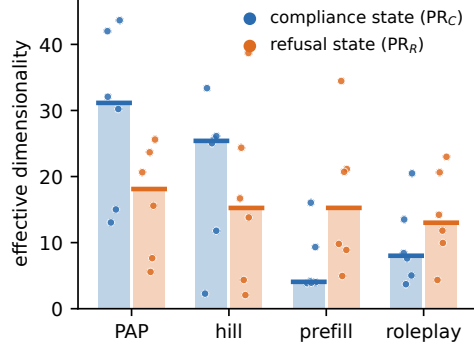


Figure 6: Effective dimensionality of the compliance state (PR_C) and the refusal state (PR_R), by attack. Each dot is one model-attack cell, and the bar is the median over the six models. PR_C swings widely with the attack, with medians of 31 under PAP and 25 under hill but only 4 under prefill and 8 under roleplay, while PR_R stays in a narrower band. The attack reshapes the compliance state and leaves the refusal state much more stable.

is not itself a chance-level finding and masks substantial attack-level heterogeneity: PAP and hill transfer better when the related persuasion attack remains in the training pool, while held-out prefill reaches 0.51 on Llama but 0.71 on Mistral. Held-out roleplay, whose mechanism has no close analogue in the training pool, is the consistent failure across both dense models: at mid-depth, it reaches a median balanced accuracy of 0.525 (0.52 on Llama and 0.53 on Mistral) and a median AUROC of 0.561. This is not because roleplay is undecodable in these representations. When trained and tested on disjoint halves of the roleplay templates, the same mid-depth residual representations yield a median balanced accuracy of 0.713 and a median AUROC of 0.795. The probe therefore learns a roleplay-specific comply-versus-refuse boundary from roleplay examples but does not recover that boundary when trained on PAP, hill, and prefill. Roleplay transfer also remains weak at the training-selected and late-depth readouts, with median balanced accuracies of 0.546 and 0.502, respectively (Appendix C.1). The held-out roleplay failure therefore occurs without expert routing.

Having found no direction shared across attacks, we ask the mirror-image question: does the jailbroken state vary with the attack in any *systematic* way, even if not as a direction a probe could exploit? It does, and the form the variation takes explains why a probe cannot catch it. The one property of the jailbroken state that moves with the attack is its effective dimensionality. The compli-

ance state’s participation ratio PR_C has medians of 31 under PAP and 25 under hill, but only 4 under prefill and 8 under roleplay; the refusal state’s PR_R stays in a narrower band between 13 and 18 (Figure 6). Restricting roleplay to four templates, matching the persuasion family at a held subsample budget, drops its median PR_C from 8 to 4: once template diversity is controlled, roleplay’s compliance state collapses like prefill’s. This is a real, attack-systematic signal, the mirror image of an attack-invariant one: the attack leaves a trace, but the trace is a statistic of the state, the number of directions it occupies, not a direction a probe can be trained on.

4.5 Two probe-free analyses agree

Two probe-free analyses test the invariance of the request representation rather than of the compliance boundary, and point the same way. A representational-similarity analysis, after removing the per-attack shift, finds that a base request is on average no closer to itself under a different attack wrapper than to an unrelated request; the within-minus-across cosine gap is small and inconsistent in sign across models and depths. A direct cross-attack re-identification test, which attack-centers the activations and matches each base request across wrappers, succeeds above chance, yet a TF-IDF bag-of-words of the prompt text outperforms it for every model and every pair of attacks (36 in total). A bag-of-words sees surface tokens directly, so attack-centering alone does not isolate the request from its wrapper.

5 Discussion

Legible but not localized. Within an attack, the routing exposes the jailbreak: a probe on the router logits decodes whether the model complied, and a few experts carry the signal. But the exposure does not hold still: the experts that carry the decision are largely not the same from one attack to the next. The router sees the jailbreak, but not in the same place twice.

Transfer is the exception, not the rule. A routing probe transfers in only one of the cases we test, between the persuasion attacks PAP and hill, and even there marginally; pooled over held-out attacks it sits at chance. A defender cannot tell in advance which side of that line a future attack will fall on. This does not rule out an activation defense against the attacks its probe was trained on; it rules out the

generalization to unseen attacks that such a defense would need to be worth deploying.

Transfer to held-out roleplay also fails in dense models.

The dense controls reproduce weak residual-stream transfer to held-out roleplay, so this failure can arise without sparse expert routing. They do not support a broad architecture-level comparison: the pooled dense result is not at chance, and transfer to prefill differs across the two models. Routing instead makes the changing expert set explicit; more broadly, a boundary learned from known attacks need not transfer to a mechanistically distinct attack even in a dense residual stream.

Shape varies with the attack; location does not stay.

We found no attack-invariant, localized routing signature: the informative experts and the comply-versus-refuse direction shift across attacks and depth. What changes systematically is the compliance state’s *shape*: its effective dimensionality varies by an order of magnitude across attacks, while the refusal state stays in a much narrower band. Prefill and roleplay may collapse compliance onto a narrow state because both write a templated opening, whereas a persuasion attack must carry its own argument. This interpretation remains untested; the structural result is that the attack changes the state’s shape without sharing its location, and a probe is trained on locations, not shapes. A defense without broad attack-family coverage may therefore need to probe upstream of the compliance state.

6 Conclusion

We asked whether a jailbroken mixture-of-experts model carries a signature like the single refusal direction of a dense transformer, looking in the routing, where a localized mechanism would be easiest to find. Within an attack, the routing exposes the jailbreak clearly, but the experts carrying the signal shift from one attack to the next, and a routing probe trained on three attacks performs at chance on a held-out fourth. Held-out roleplay provides the cleanest failure, and the same weak residual-stream transfer appears in two dense controls, indicating that the non-transfer is not specific to expert routing. In the MoE panel, the attack rewrites the compliance state’s shape without ever sharing its location, leaving a trace that is real but in a form a probe cannot exploit. An activation probe therefore cannot be assumed, on its own, to catch a jailbreak it was not built for.

7 Limitations

Scope of the dense-model controls. Our dense-model evidence consists of two models and is limited to residual-stream transfer analyses: leave-one-attack-out transfer at train-selected, fixed mid-depth, and fixed late-depth readouts; threshold-free AUROC at the fixed readouts; and within-roleplay cross-template transfer. Because dense models have no expert routing, these controls do not test the routing-specific findings about expert identity or sparsity. We also do not repeat the effective-dimensionality, representational-similarity, or full layerwise direction analyses on the dense models. The fixed mid- and late-depth readouts are sensitivity checks rather than claims of exact layer correspondence across architectures. These controls show that weak transfer to held-out roleplay can occur without expert routing, but they do not establish uniform non-transfer across novel attack mechanisms: transfer to held-out prefill differs substantially between the two dense models. Nor do they establish that routing makes no causal contribution within MoE models or that dense and MoE architectures have identical jailbreak geometry or robustness.

Judge-defined labels. The compliance and refusal sets are defined by an automatic judge, and a different judge would move some prompts between them. We do not expect this to change the cross-attack picture, since the failure to transfer holds across the routing, the residual stream, and two probe-free analyses, but we have not re-derived every result under a second judge. The caution bears most on the effective-dimensionality result, where prefill’s low-dimensional compliance state may partly reflect borderline compliances rather than fully genuine ones.

A narrow attack base. Our coverage is four attacks and six models. PAP and hill are both persuasion attacks; prefill and roleplay each work in a different way. Our leave-one-attack-out test therefore rests on one pair of related attacks and two singletons. This is enough to show that transfer can succeed for one pair and is absent otherwise, but not enough to map how a probe generalizes within a larger set of related attacks; that would need several genuinely different attacks of the same kind.

Natural-language jailbreaks only. Our four attacks are all natural-language jailbreaks, which manipulate the model through meaningful text. Optimization-based attacks, such as GCG (Zou

et al., 2023b) and AutoDAN (Liu et al., 2024b), instead search the input for token sequences that trigger compliance; this is a different mechanism, and GCG further needs white-box gradient access that our black-box setting rules out. Whether the jailbroken state has the same geometry, and the same non-transfer, when compliance is reached by optimization is a question we leave open.

Probe family. Our main probes are linear, and the small MLP head we tested as a robustness check matches the linear verdict (Appendix C); a more expressive probe family might still transfer further than the heads we test.

What would change our view. We would update our conclusion if a non-linear probe transferred substantially above 0.75 in the leave-one-attack-out setting, or if a broader attack panel revealed transfer that grew with semantic similarity rather than collapsing across families.

Ethical Considerations

Motivation and audience. This paper analyzes the failure mode of activation-based jailbreak defenses to clarify what such defenses can and cannot promise. The intended audience is the safety research community building such defenses; we argue that cross-attack generalization, rather than within-attack decode quality, is the binding constraint on deployment.

Dual-use considerations. Our negative finding—that linear probes do not transfer across jailbreak attack families—is information an adaptive attacker could exploit only indirectly, by noting that activation monitoring trained on a fixed attack set is brittle. This is the same constraint already documented for adaptive adversaries in prior work (Bailey et al., 2026); our contribution is to show that the brittleness holds even without an adaptive adversary. We judge that surfacing this constraint to defenders outweighs the marginal information it provides to attackers.

Harmful content handling. We use publicly released safety benchmarks (HarmBench (Mazeika et al., 2024), AdvBench (Zou et al., 2023b), StrongREJECT (Souly et al., 2024), HarmEval (Banerjee et al., 2025)) governed by their original licenses, and we do not introduce new jailbreak attacks: the four families are existing public templates cited in Section 3. Model responses to harmful prompts are labeled by an automatic judge and used only to define the compliance and refusal sets for activa-

tion analysis; judge errors therefore propagate as cell-level label noise and remain a source of uncertainty for the reported estimates. No raw harmful responses are released; the public artifact contains only derived numerical features (probe accuracies, expert effect sizes, participation ratios) and prompt-level metadata (judge label and query identifier).

Acknowledgments

This work was supported by Korea Internet & Security Agency (KISA) grant funded by the Korea government (PIPC) (No.RS-2026-25526342, Development of Technologies for Preventing Sensitive Information Inference and Risk Assessment in Foundation Model Operations)

References

- Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. [Jailbreaking leading safety-aligned LLMs with simple adaptive attacks](#). In *International Conference on Learning Representations (ICLR)*.
- Andy Ardit, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 136037–136083.
- Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. 2026. [Obfuscated activations bypass llm latent-space defenses](#). In *International Conference on Learning Representations*.
- Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. 2025. [SafeInfer: Context adaptive decoding time safety alignment for large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, and 1 others. 2024. [Deepseek-moe: Towards ultimate expert specialization in mixture-of-experts language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297.
- Weilong Dong, Peiguang Li, Yu Tian, Xinyi Zeng, Fengdi Li, and Sirui Wang. 2025. [Feature-aware malicious output detection and mitigation](#). In *AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLIM)*.
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. [Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity](#). *Journal of Machine Learning Research*, 23(120):1–39.
- Gil Goren, Shahar Katz, and Lior Wolf. 2026. [Align-Tree: Efficient defense against llm jailbreak attacks](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37417–37425.
- Fabian Hildebrandt, Andreas Maier, Patrick Krauss, and Achim Schilling. 2025. [Refusal behavior in large language models: A nonlinear perspective](#). In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.
- Xiaomeng Hu, Fei Huang, Chenhan Yuan, Junyang Lin, and Tsung-Yi Ho. 2025. [CARE: Decoding-time safety alignment via rollback and introspection intervention](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 51154–51189.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. [Mixtral of experts](#). *arXiv preprint arXiv:2401.04088*.
- Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. 2025. [HiddenDetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14880–14893, Vienna, Austria. Association for Computational Linguistics.
- Sri Durga Sai Sowmya Kadali and Evangelos E Papalexakis. 2025. [Do internal layers of llms reveal patterns for jailbreak detection?](#) In *NeurIPS 2025 Workshop: Reliable ML from Unreliable Data*.
- Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024a. [Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model](#). *arXiv preprint arXiv:2405.04434*.
- Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024b. [AutoDAN: Generating stealthy jailbreak prompts on aligned large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Xuan Luo, Yue Wang, Zefeng He, Geng Tu, Jing Li, and Ruifeng Xu. 2025. [A simple and efficient jailbreak method exploiting LLMs’ helpfulness](#). *arXiv preprint arXiv:2509.14297*.
- Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. 2024. [Simple probes can catch sleeper agents](#).

- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. [HarmBench: A standardized evaluation framework for automated red teaming and robust refusal](#). In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research.
- Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, and 1 others. 2025. [Olmoe: Open mixture-of-experts language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. 2025. [The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions](#). In *Forty-second International Conference on Machine Learning*.
- Oam Patel and Rowan Wang. 2024. [Activation monitoring: advantages of using internal representations for llm oversight](#). In *Second NeurIPS Workshop on Attributing Model Behavior at Scale*.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. [Outrageously large neural networks: The sparsely-gated mixture-of-experts layer](#). In *International Conference on Learning Representations (ICLR)*.
- Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. [“Do anything now”: Characterizing and evaluating in-the-wild jailbreak prompts on large language models](#). In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.
- Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. [A StrongREJECT for empty jailbreaks](#). In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*.
- Qwen Team. 2024. [Introducing qwen1.5](#).
- Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. 2025. [The geometry of refusal in large language models: Concept cones and representational independence](#). In *International Conference on Machine Learning*, pages 66945–66970.
- Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. [How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, Bangkok, Thailand. Association for Computational Linguistics.
- Shenyi Zhang, Yuchen Zhai, Keyan Guo, Hongxin Hu, Shengnan Guo, Zheng Fang, Lingchen Zhao, Chao Shen, Cong Wang, and Qian Wang. 2025. [JBShield: Defending large language models from jailbreak attacks through activated concept analysis and manipulation](#). In *34th USENIX Security Symposium (USENIX Security 25)*, pages 8215–8234.
- Shide Zhou, Kailong Wang, Ling Shi, and Haoyu Wang. 2026. [Exposing the ghost in the transformer: Abnormal detection for large language models via hidden state forensics](#). *IEEE Transactions on Information Forensics and Security*.
- Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023a. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.
- Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Attack Separability

The four attacks occupy linearly separable regions of activation space, which is why every comparison in this paper is made within a single attack (Section 3). We train a four-way logistic-regression classifier to identify the attack from a single layer’s activations, with cross-validation. It reaches a median accuracy of 0.998, against a chance level of 0.25 (Table 3). Pooling activations across attacks would mix four nearly disjoint geometries.

B Per-Model Routing Decode

Table 4 gives the within-attack decode accuracy of Section 4.1 for every model-attack cell: a probe on the router logits, a probe on the discrete selected experts, and a probe on the residual stream. The router logits track the residual stream closely. The discrete expert selection is a weaker but above-chance signal.

C Cross-Attack Transfer

Table 5 gives the cross-attack transfer of Section 4.2. For each model we fit a router-logit probe on one attack and test it, unchanged, on another. Within-attack accuracy is the mean over the model’s four attacks. Across-attack accuracy is the median over all ordered attack pairs. The routing probe falls close to the chance level of 0.5, and a residual-stream probe does no better.

Robustness to the probe head. We re-run the leave-one-attack-out probe with a one-hidden-layer MLP (64 ReLU units, $\alpha = 10^{-3}$, early stopping with patience 10), replacing the logistic head while keeping the rest of the pipeline (training pool, layer selection, PCA dimensionality) unchanged. The MLP reaches a pooled balanced accuracy of 0.518 [0.500, 0.655] on the routing and 0.543 [0.500, 0.665] on the residual stream, against linear medians of 0.52 [0.50, 0.61] and 0.62 [0.51, 0.73]. Per held-out attack, the MLP catches PAP and hill above the chance point estimate (routing 0.62 and 0.68, residual 0.75 and 0.67) and prefill and roleplay at chance (routing 0.50 and 0.49, residual 0.50 and 0.50), the same pattern as the linear probe.

Robustness to the layer-selection heuristic. The main text picks the readout layer by maximizing mean within-attack decode on the three training attacks; we contrast that with two attack-agnostic fallbacks, the fixed depth-50% layer and

the fixed depth-90% layer. Pooled over the 24 held-out cells, the routing leave-one-attack-out median is 0.522 (train-best), 0.515 (mid), and 0.595 (late); the residual-stream median is 0.617 (train-best), 0.585 (mid), and 0.667 (late). The largest excursion from the train-best value is 0.07 on the routing and 0.05 on the residual stream, and prefill and roleplay sit at 0.49 to 0.55 under every choice.

C.1 Dense-Model Controls

Table 6 reports the complete balanced-accuracy results from the residual-stream leave-one-attack-out analysis on Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3. For each model and held-out attack, the training-selected readout chooses a layer using only the other three attacks; the held-out attack is used neither for layer selection nor for probe training. The fixed mid- and late-depth readouts use zero-indexed decoder-block indices 16 and 29 (the 17th and 30th blocks), respectively, in both 32-layer models. These fixed readouts test sensitivity to the layer-selection heuristic and are not intended to assert exact computational alignment with layers in the MoE architectures.

Pooled over the eight model-held-out-attack cells, the median balanced accuracy is 0.645 [0.507, 0.726] at the training-selected readout, 0.632 [0.521, 0.711] at the fixed mid-depth readout, and 0.648 [0.500, 0.685] at the fixed late-depth readout, where brackets denote 95% percentile-bootstrap confidence intervals from 4000 resamples. These statistics are computed from the unrounded cell values. The pooled summaries mask substantial attack-level heterogeneity: transfer is higher for PAP and hill when the related persuasion attack remains in the training set, while prefill transfers on Mistral but not on Llama.

Held-out roleplay is the consistent failure across both models and all three readouts. Across the two models, its median balanced accuracy is 0.546 at the training-selected readout, 0.525 at fixed mid-depth, and 0.502 at fixed late depth. At fixed mid-depth, its median AUROC is 0.561. This weak transfer is not explained by roleplay being intrinsically undecodable: when probes are instead trained and tested on disjoint sets of roleplay templates, the same mid-depth representations reach a median balanced accuracy of 0.713 and a median AUROC of 0.795. At late depth, the corresponding leave-one-attack-out and within-roleplay results are 0.502 versus 0.670 in balanced accuracy and 0.542 versus 0.750 in AUROC. Thus, the dense residual stream

Model	Attack	top-1	top-2	top-3	top-5	top-8	top-12	all
deepseek-v2-lite	PAP	0.92	0.92	0.92	0.93	0.94	0.94	0.97
	hill	0.79	0.82	0.83	0.82	0.89	0.89	0.96
	prefill	0.82	0.85	0.86	0.85	0.84	0.87	0.88
	roleplay	0.92	0.92	0.94	0.99	0.99	1.00	1.00
deepseek-moe-16b-chat	PAP	0.62	0.62	0.61	0.64	0.62	0.62	0.65
	hill	0.59	0.62	0.63	0.66	0.67	0.66	0.68
	prefill	0.56	0.55	0.58	0.57	0.59	0.56	0.62
	roleplay	0.63	0.64	0.66	0.65	0.64	0.65	0.75
qwen1.5-moe-a2.7b-chat	PAP	0.67	0.70	0.72	0.74	0.75	0.75	0.80
	hill	0.70	0.71	0.73	0.74	0.75	0.77	0.80
	prefill	0.60	0.61	0.61	0.60	0.65	0.64	0.66
	roleplay	0.72	0.73	0.74	0.75	0.76	0.79	0.83
llama-4-scout	PAP	0.71	0.75	0.79	0.78	0.82	0.85	0.86
	hill	0.73	0.79	0.86	0.88	0.94	0.96	0.97
	prefill	0.69	0.70	0.75	0.76	0.75	0.78	0.75
	roleplay	1.00	1.00	1.00	1.00	1.00	0.99	1.00
mixtral-8x7b	PAP	0.74	0.80	0.81	0.82	0.85	0.85	0.85
	hill	0.77	0.89	0.90	0.95	0.94	0.94	0.94
	prefill	0.62	0.67	0.68	0.67	0.70	0.70	0.70
	roleplay	0.79	0.83	0.95	0.96	0.97	0.97	0.97
olmoe-1b-7b	PAP	0.87	0.91	0.90	0.91	0.91	0.92	0.97
	hill	0.84	0.84	0.86	0.89	0.90	0.92	0.96
	prefill	0.59	0.65	0.65	0.65	0.66	0.69	0.79
	roleplay	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2: Within-attack balanced accuracy of a router-logit probe restricted to the top- n experts by router-logit effect size, for every model-attack cell, beside the full probe (all experts). Experts are ranked on the training folds only, and chance is 0.5. For mixtral-8x7b the top-8 and top-12 columns equal the full probe, since the model has only eight experts.

Model	complied	refused
deepseek-v2-lite	0.996	0.996
deepseek-moe-16b-chat	0.997	0.998
qwen1.5-moe-a2.7b-chat	0.998	0.998
llama-4-scout	1.000	0.987
mixtral-8x7b	0.999	0.995
olmoe-1b-7b	0.996	0.997
median	0.998	0.997

Table 3: Four-way attack-decoding accuracy from a single layer’s activations, with cross-validation, against a chance level of 0.25. The complied column uses only complied activations, the refused column only refused activations.

contains a learnable roleplay compliance boundary, but a probe trained on the other attacks does not recover it.

D Attack Templates

Each attack family wraps a base harmful request through a set of templates, so a single request expands into several attack prompts (Section 3). The families differ widely in how many templates they use: roleplay draws on 80 persona templates, PAP and hill on 4 each, and prefill on a single fixed

adversarial prefix. Within each model, all four families wrap that model’s bare-refused query pool. An attack’s prompt count therefore scales with its template count: roleplay yields by far the most attack prompts while prefill yields the fewest.

E Expert Sparsity

Section 4.1 reports that a few experts carry most of the within-attack signal. Table 2 expands this for every model-attack cell: it gives the balanced accuracy of a router-logit probe restricted to the n experts with the largest router-logit effect size, for n in $\{1, 2, 3, 5, 8, 12\}$, beside the full probe on all experts. The experts are ranked on the training folds alone, so each restricted probe is scored on held-out data. Across the 24 cells the top-3 probe recovers a median of 79% of the full probe’s above-chance accuracy (the margin over the 0.5 chance level), with per-cell recovery ranging from 52% to 100%. We report a fixed top-3 rather than each model’s architectural routing width, which ranges from one to eight across the panel, so the sparsity measure stays comparable across models.

Model	Attack	n_R	n_C	Router	Experts	Residual
deepseek-v2-lite	PAP	103	160	0.97	0.88	0.98
	hill	115	153	0.96	0.82	0.95
	prefill	132	162	0.88	0.86	0.87
	roleplay	196	76	1.00	0.94	0.99
deepseek-moe-16b-chat	PAP	154	770	0.66	0.61	0.68
	hill	183	785	0.71	0.62	0.69
	prefill	153	147	0.62	0.59	0.60
	roleplay	1197	1205	0.80	0.71	0.79
qwen1.5-moe-a2.7b-chat	PAP	185	736	0.81	0.64	0.84
	hill	269	701	0.81	0.70	0.79
	prefill	191	116	0.66	0.57	0.65
	roleplay	1685	850	0.81	0.75	0.81
llama-4-scout	PAP	153	189	0.86	0.72	0.96
	hill	164	196	0.97	0.84	0.99
	prefill	183	154	0.76	0.64	0.76
	roleplay	198	180	1.00	0.97	1.00
mixtral-8x7b	PAP	51	194	0.85	0.69	0.93
	hill	155	194	0.94	0.74	0.99
	prefill	177	181	0.70	0.64	0.77
	roleplay	189	163	0.97	0.90	0.98
olmoe-1b-7b	PAP	184	138	0.97	0.78	0.98
	hill	187	188	0.96	0.81	0.95
	prefill	198	191	0.79	0.63	0.74
	roleplay	197	156	1.00	0.86	1.00

Table 4: Within-attack decode accuracy (5-fold cross-validated balanced accuracy) for every model-attack cell. n_R and n_C are the refusal-set and compliance-set sample counts before the 200-cap subsampling. Router is a probe on the router logits, Experts a probe on the discrete selected experts, Residual a probe on the residual stream. Chance is 0.5.

Model	routing	across attacks	
	within	routing	residual
deepseek-v2-lite	0.95	0.53	0.62
olmoe-1b-7b	0.93	0.50	0.55
llama-4-scout	0.89	0.50	0.52
mixtral-8x7b	0.86	0.59	0.52
qwen1.5-moe-a2.7b-chat	0.77	0.50	0.50
deepseek-moe-16b-chat	0.70	0.51	0.52
median	0.88	0.51	0.52

Table 5: Single-attack cross-attack transfer of a router-logit probe, balanced accuracy, against a chance level of 0.5. The within column trains and tests on the same attack; the across columns train on one attack and test on another (the off-diagonal of Figure 4), for the routing and for the residual stream. The leave-one-attack-out result, which pools three attacks at training time, is reported in Section 4.2 (median 0.52 routing, 0.62 residual).

Model	Readout	Held-out attack			
		PAP	hill	prefill	roleplay
Llama-3.1-8B	training-selected	0.74	0.63	0.50	0.51
	fixed mid, index 16	0.67	0.63	0.51	0.52
	fixed late, index 29	0.69	0.71	0.50	0.54
Mistral-7B	training-selected	0.73	0.66	0.71	0.58
	fixed mid, index 16	0.72	0.64	0.71	0.53
	fixed late, index 29	0.65	0.65	0.68	0.46

Table 6: Residual-stream leave-one-attack-out balanced accuracy for the two dense controls. Chance balanced accuracy is 0.50. Decoder-block indices are zero-indexed, and entries are rounded to two decimal places.