JAILBREAKS LIVE IN PERSONA SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding why aligned language models fail is central to AI safety. Jailbreaks provide a concrete instance, where adversarial prompts can shift models from refusal to harmful compliance. We study jailbreaking through the emerging view of LLMs as persona simulators, asking whether jailbreak behavior is organized within *persona space*, defined here through human-interpretable persona trait directions. On Llama-3.1-8B and Gemma-4-31B, we find that jailbreaks are strongly structured by persona space. Individual traits predict jailbreak outcome, persona structure emerges in principal components learned solely from jailbreak activations, and persona-space projections retain most of the predictive signal in refusal and jailbreak directions. Building on this structure, we construct the `PersonaJB Direction`, a single direction constrained entirely to persona space that achieves the strongest average generalization to unseen attacks on both models with minimal inference overhead. Finally steering along `PersonaJB` and individual trait directions induces jailbreaks at rates comparable to specialized safety directions and standard attacks. Our results establish persona space as an interpretable, predictive, and causal representation of jailbreak behavior.

1 INTRODUCTION

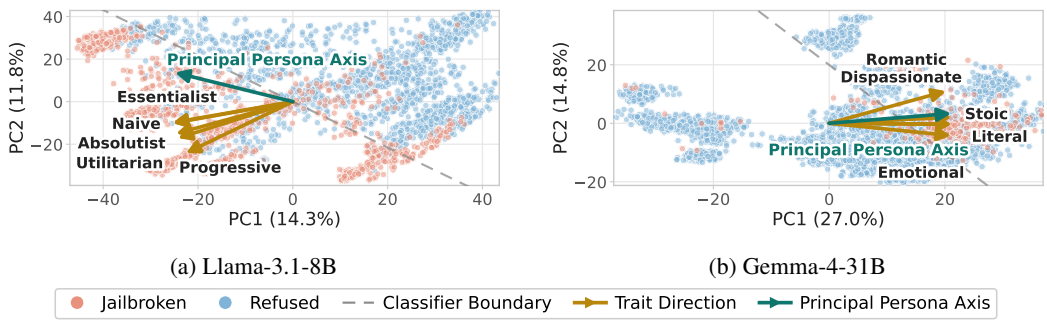
Large language models are becoming increasingly capable and autonomous, while still exhibiting safety-relevant failures ranging from compliance with harmful queries (Chao et al., 2024) to deception (Goldowsky-Dill et al., 2025) and scheming (Schoen et al., 2025). Preventing undesirable behavior is increasingly important, and doing so requires understanding how such behavior arises inside the model (Bengio et al., 2026; Sharkey et al., 2025). Jailbreaking provides a concrete and tractable instance of this broader alignment problem, in which a model can be pushed to violate its intended behavioral constraints. Different adversarial prompts can move otherwise aligned models from refusal to harmful compliance, making jailbreaks a useful setting for studying how safety-relevant behavior is represented internally (Zhao et al., 2026). A better mechanistic understanding of this transition could reveal more general structure underlying alignment failures (Arditi et al., 2024).

An emerging view of language-model behavior treats an LLM as a *persona simulator* (janus, 2022). The Persona Selection Model (PSM) hypothesizes that pretraining teaches a broad repertoire of personas, while post-training does not create the assistant from scratch, but wires it as a mixture of personas already present in the model (Marks et al., 2026). Recent work directly supports this picture: personas emerge early in pretraining and persist through post-training (Moskvoretskii et al., 2026), while Synthetic Persona Pretraining (SPP) shows that introducing new personas is substantially more effective early in training than during midtraining or post-training (Minder et al., 2026).

Under this view, the assistant is an inherently imperfect construction from a large and heterogeneous behavioral repertoire. It must simultaneously act, for example, as an empathetic companion, a knowledgeable expert, and a safety-aligned assistant; the way these personas are wired together may leave unexpected regions where alignment breaks. Recent work shows sparks of this pattern: toxic persona features mediate emergent misalignment (Wang et al., 2025), drift from the default assistant persona accompanies some safety failures (Lu et al., 2026), and explicitly shaping an aligned persona improves robustness (Minder et al., 2026). Yet they provide only preliminary evidence, leaving the broader mechanistic connection between jailbreaks and persona representations unresolved.

We address this gap by exploring the connection between jailbreak behavior and persona space on Llama-3.1-8B (Grattafiori et al., 2024) and Gemma-4-31B (Team et al., 2026), spanning different

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Llama-3.1-8B (b) Gemma-4-31B

Figure 1: **Jailbreak geometry aligns with persona directions.** Raw jailbreak activations from Llama-3.1-8B (left) and Gemma-4-31B (right) are projected onto their top two principal components, fit without any persona information. Each point is a prompt, and the five individually most predictive persona trait directions are projected into the same plane. The dashed line is a regression boundary fit on the displayed components. Predictive traits align with the jailbreak–refusal separation in both models, the Principal Persona Axis aligns with Gemma separation but is nearly orthogonal for Llama.

model families and scales, across multiple jailbreak families. We represent persona space using 240 human-interpretable trait directions (Lu et al., 2026) and study this connection from geometric, predictive, and causal perspectives. We study it through individual traits, jailbreak geometry, and safety-relevant directions. Building on these findings, we construct the `PersonaJB Direction`, a monitor constrained entirely to persona space that predicts jailbreak outcome from prompt processing alone and generalizes strongly to unseen attacks. Finally, we test causality through persona-space steering on Llama-3.1-8B and show that these directions induce jailbreaks at rates comparable to specialized refusal directions and standard jailbreak attacks.

Our main contributions are:

(i) **Jailbreaks live in persona space.** Individual traits predict jailbreak outcome, jailbreak geometry recovers persona axes, and persona-space projections preserve most of the predictive signal in refusal and jailbreak directions.

(ii) **Single persona-space direction generalizes across jailbreaks.** The `PersonaJB Direction`, constrained entirely to persona space, achieves the strongest average OOD detection across unseen attacks on both models with minimal inference overhead.

(iii) **Persona-space directions causally induce jailbreaks.** Steering along both the `PersonaJB Direction` and individual trait directions turns refusals into jailbreaks at rates comparable to specialized safety directions and standard attacks.

2 Methodology

Conceptual Setup. We build our setup around the concepts of personas and traits, following Persona Vectors (Chen et al., 2025) and Assistant Axis (Lu et al., 2026). We use *persona* to refer broadly to a behavioral disposition of a language model, while acknowledging that the term is not yet rigorously defined and is often used interchangeably with related notions such as traits, roles, or characters (Moskvoretskii et al., 2026). Despite this ambiguity, these concepts provide a useful basis for interpreting LLM behavior. We represent personas through *trait vectors* and define persona space as the span of these directions, following the Assistant Axis framework (Lu et al., 2026).

Trait Vectors. We follow the operationalization used in previous work and define a trait τ through a natural-language description together with an evaluation rubric measuring the degree to which a model output exhibits τ. To obtain its representation, we extract a linear direction from the model activations by contrasting the average activations of positive and negative examples:

$$v_\tau = \frac{1}{|D_+|} \sum_{y_+ \in D_+} h_{y_+} - \frac{1}{|D_-|} \sum_{y_- \in D_-} h_{y_-}$$

where D_+ and D_- denote generations exhibiting and not exhibiting trait τ, respectively, and h_y is the activation vector extracted from generation y. We use GPT-4.1-mini to score each generation for

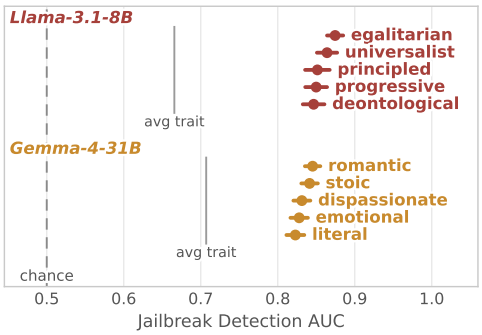


Figure 2: **Individual trait projections predict jailbreak success.** For each jailbreak prompt, we project the pre-generation activation onto individual trait directions and use each projection to predict whether the model will refuse. We show AUC for the five most predictive traits on Llama-3.1-8B and Gemma-4-31B, chance and the mean AUC across all individual traits. Llama's strongest predictors are primarily ethical, while Gemma's are predominantly affective. Whiskers show 95% confidence intervals.

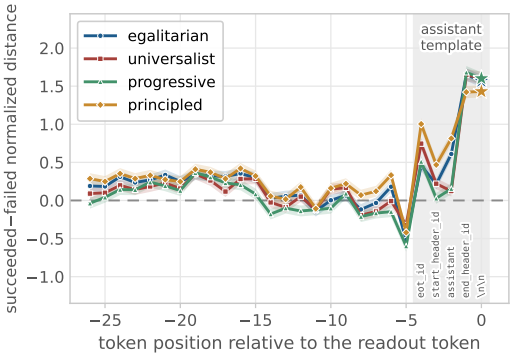


Figure 3: **Predictive signal emerges at the assistant turn.** For Llama-3.1-8B, we project activations at each prompt position onto trait directions and measure standardized separation between successful and failed jailbreaks, dividing the mean difference by pooled standard deviation. The groups remain indistinguishable during the user turn and separate only at the assistant turn, suggesting that jailbreak outcome is tied to the persona adopted for the response rather than harmful prompt content. Shaded bands show 95% confidence intervals.

coherence and trait strength on a 0–100 scale against the corresponding positive or negative rubric, retaining examples with both scores > 50 to form D_+ and D_-.

Persona Space. We construct the persona space using 240 traits from Assistant Axis (Lu et al., 2026). For each trait, every question is answered once under a positive and once under a negative system prompt inducing trait τ, yielding 200 matched pairs. We retain the trait only if at least 10 pairs have a positive-minus-negative judge-score difference of at least 50. This retains 229 of the 240 traits on Llama-3.1-8B, while on Gemma-4-31B it resulted in the use of all 240. Stacking the normalized trait vectors gives a trait matrix $T \in \mathbb{R}^{n \times d}$, where n is the number of retained traits and d is the model hidden size, with $d = 4096$ for Llama and $d = 5376$ for Gemma.

We also use the *Principal Persona Axis*, defined as the first principal component of T. We use it as a compact summary of the dominant organization of persona space, reflecting principal assistant persona behavior, motivated by the Assistant Axis (Lu et al., 2026).

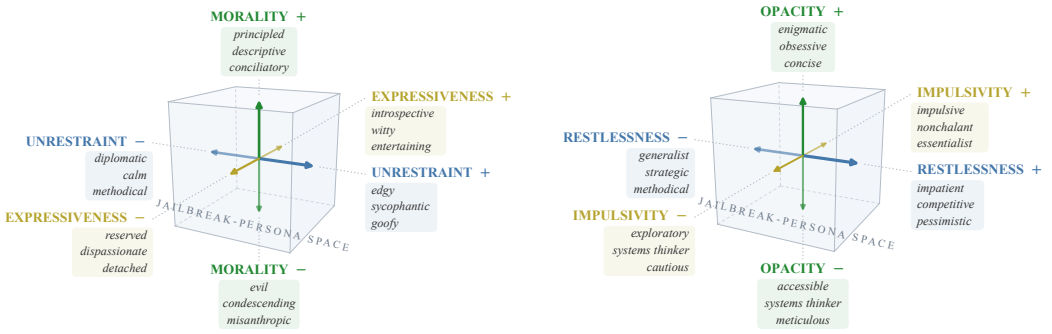
Jailbreak Setup. We study jailbreak behavior in Llama-3.1-8B (Grattafiori et al., 2024) and Gemma-4-31B (Team et al., 2026), spanning different model families and scales. Our interpretability analyses use HarmBench (Mazeika et al., 2024), with harmful behaviors wrapped in its hand-written "Human-Jailbreaks" templates, and later evaluate the resulting method on additional jailbreak datasets. For Gemma we additionally include WildJailbreak (Jiang et al., 2024), since HarmBench alone yields too few successful jailbreaks on Gemma for reliable estimates. For each prompt, we predict jailbreak success using only the model state at the end of the input prompt, before response generation, at layer 16 for Llama and layer 30 for Gemma, with both token-position and layer choices validated in Section 3 and Appendix C. We then generate a response and use it only to determine the ground-truth label: whether the model complies with the harmful request. Details are in Appendix A.2.

Metrics. We label responses using a GPT-4.1-mini judge, which determines whether each response constitutes a successful jailbreak. We report predictive performance using ROC-AUC, which is independent of a threshold and insensitive to class imbalance.

3 The Persona Structure of Jailbreaking

3.1 Trait-Level Signatures of Refusal

We first study the structure of jailbreaks through persona space by asking how predictive individual trait directions are of jailbreak outcomes.



(a) Llama-3.1-8B (b) Gemma-4-31B

Figure 4: **Principal components of jailbreak activations recover interpretable persona axes.** The first three principal components of jailbreak activations for Llama-3.1-8B and Gemma-4-31B, fit without any persona information. We interpret each component by its most aligned trait directions in persona space and assign a descriptive axis label accordingly. Despite being discovered purely from jailbreak activations, the principal components form interpretable behavioral axes. Signs are arbitrary and only opposing traits are meaningful.

Individual Traits Predict Jailbreak Success. We test whether projections onto individual persona traits can predict jailbreak success and find that several traits are highly predictive (Figure 2). This suggests that successful jailbreaks move the model into a specific subspace of persona space in which alignment is less effective at enforcing refusal.

Jailbreak Persona Profiles Differ Across Models. We observe distinct predictive trait profiles for Llama and Gemma. For Llama, the most predictive traits are associated more with principled views and general values, whereas for Gemma they are predominantly emotional. This suggests that different models may be vulnerable along different regions of persona space. Further we show that it also depends on the attack family (Section 4).

The Predictive Signal Emerges at the Assistant Turn. Successful and failed jailbreaks show only weak separation in their trait projections throughout the user turn, with a much sharper divergence during processing of the assistant chat template. This is consistent with prior work locating refusal representations at this stage (Zhao et al., 2026), and suggests that the transition may reflect formation of the assistant persona (Figure 3). Layerwise analysis shows that persona-related signal emerges from the middle layers (Appendix C).

3.2 Jailbreak Geometry Reveals Persona Axes

We further study how jailbreak geometry relates to persona space. We perform PCA on activations from successful and refused jailbreaks and ask whether directions discovered without any persona supervision align with persona structure. We refer to these pre-generation hidden states from successful and failed jailbreak attempts as *jailbreak activations*.

Predictive Traits Align with Jailbreak Geometry. We project the five most predictive trait directions into the PCA space (Figure 1) and find that they consistently align with the jailbreak–refusal separation. We add the Principal Persona Axis to show it aligns with this separation for Gemma but is nearly orthogonal to it for Llama. This further suggests that jailbreaks occupy model-specific regions of persona space.

Principal Components Align with Persona Traits. Having shown that predictive traits align with jailbreak PCA geometry (Figure 1), we now interpret the principal components themselves through their most aligned traits. We manually inspect the traits most strongly aligned with each principal component and assign a short descriptive label to summarize the corresponding axis. Figure 4 shows that these components form coherent persona axes, with their geometric alignment to persona directions quantified in Table 10. Thus, jailbreak geometry recovered without persona supervision reflects meaningful persona structure. We further support this interpretation by steering along the principal components and observing behavior consistent with the traits associated with each direction (Appendix F, Table 15).

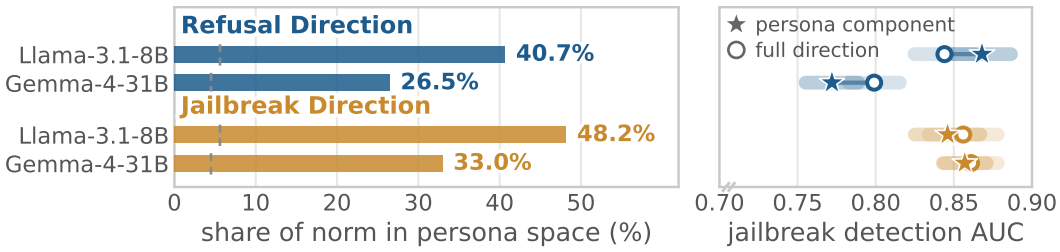


Figure 5: **Refusal and jailbreak directions are substantially captured by persona space.** For Llama-3.1-8B and Gemma-4-31B, we project the refusal direction ν_{ref}, derived from harmful-versus-harmless instructions, and the jailbreak direction ν_{JB}, derived from successful-versus-failed jailbreaks, onto the persona span. Left: fraction of squared norm captured by persona space, with dashed lines showing random-direction expectation. Right: held-out jailbreak-detection AUC using the scalar projection onto each full direction or its persona-space component. Although persona space captures only part of each direction's norm, the projected components retain most of their predictive performance. Bands show 95% confidence intervals across samples.

3.3 Refusal and Jailbreak Signals Concentrate in Persona Space

Building on the finding that individual traits predict jailbreak outcomes and provide interpretable labels for jailbreak PCs, we next ask whether persona space also captures broader refusal- and jailbreak-related directions. We analyze two complementary directions. We use a **refusal direction** ν_{ref} computed following Arditi et al. (2024) as the difference in means between harmful and harmless instructions, independently of our jailbreak data, at our readout layer and position. In contrast, we construct an in-domain **jailbreak direction** ν_{JB} on HarmBench from the difference in mean activations between successful and failed jailbreak attempts. Since both groups contain harmful requests, ν_{JB} isolates variation of jailbreak outcome rather than harmfulness.

To measure how much of these directions is expressible in persona space, we obtain an orthonormal basis Q for the span of the trait vectors T using QR decomposition and construct the orthogonal projector $P = QQ^\top$. For either direction ν, its persona-space component is $\nu_{\text{persona}} = P\nu$, the closest direction to ν that can be expressed as a linear combination of trait vectors.

Persona Space Captures a Large Fraction of Direction Norm. We measure the fraction of squared norm contained in persona space, $|P\nu|^2/|\nu|^2$. Figure 5 (left) shows that roughly 30–50% of the squared norm of both directions lies in the persona span, well above the random-direction baseline. Although these directions are constructed independently of the persona traits, a substantial part of each can therefore be recovered within persona space, connecting the broader geometry of refusal and jailbreaking to persona structure.

Persona Space Preserves Jailbreak-Predictive Information. We next test how predictive the persona-space components $P\nu$ remain (Figure 5, right). Although persona space captures only part of the norm of the original directions, their projections retain most of the jailbreak-detection performance. This suggests that the components of the refusal and jailbreak directions that are most relevant to jailbreak outcome are concentrated in persona space.

4 Monitoring Jailbreaks in Persona-Trait Directions

Motivated by the strong jailbreak signal observed in persona space, we construct a single jailbreak-detection direction constrained to this space, which we call the **PersonaJB Direction**. Given a pre-generation activation h, we compute $z = Th \in \mathbb{R}^n$, so that each feature measures alignment with an individual trait. We standardize these trait scores to zero mean and unit variance and train a logistic-regression classifier to predict whether the upcoming response will be jailbroken or refused. We then map the learned classifier back into activation space as $w_{\text{PJB}} = T^\top(\beta/\sigma)$, where β are the logistic-regression weights and σ are the per-trait feature standard deviations. Thus, w_{PJB} is a weighted combination of trait directions and, by construction, lies entirely in persona space. Jailbreak detection is then performed by projecting h onto w_{PJB}.

We tune the PersonaJB Direction on HarmBench for Llama and on HarmBench together with WildJailbreak for Gemma, since HarmBench alone yields too few successful jailbreaks on Gemma for

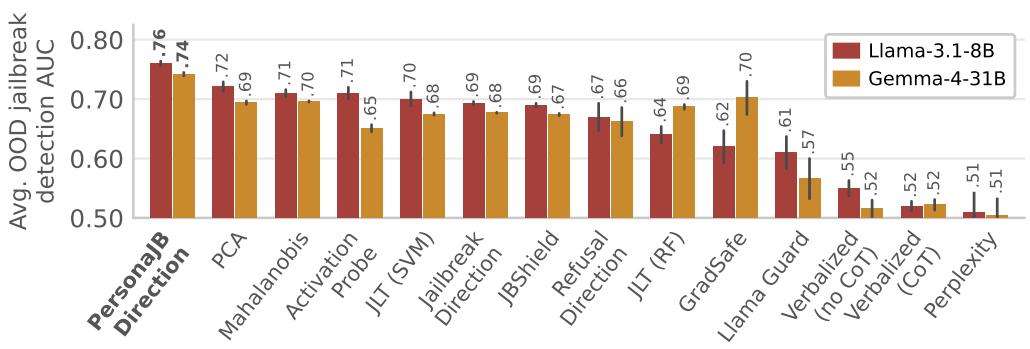


Figure 6: **PersonaJB Direction generalizes best to unseen jailbreak attacks.** Average out-of-distribution jailbreak-detection AUC across five unseen attack families for Llama-3.1-8B and Gemma-4-31B. PersonaJB Direction is a vector constrained to persona space, constructed as a learned weighted combination of individual trait directions. It achieves the highest average transfer AUC on both models, suggesting that persona space contains a jailbreak-predictive direction that generalizes across attack mechanisms. Bars start at chance; whiskers show 95% confidence intervals, computed across data splits for learned detectors and by bootstrap for unsupervised ones.

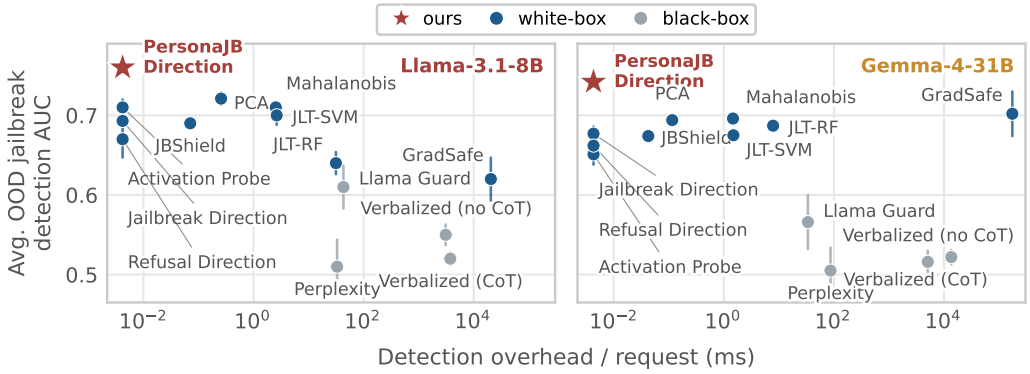


Figure 7: **PersonaJB combines strong transfer with low cost.** Average OOD jailbreak-detection AUC versus per-request inference overhead for Llama-3.1-8B (left) and Gemma-4-31B (right), following Figure 6. Overhead is measured beyond the model forward pass; activation-based methods reuse the pre-generation hidden state. Bars show 95% confidence intervals as in Figure 6.

reliable training. We then test generalization on five out-of-distribution jailbreak families: GCG (Zou et al., 2023b), PAIR (Chao et al., 2023), PAP (Zeng et al., 2024), GPTFuzz (Yu et al., 2023), and PEZ (Wen et al., 2023). These cover substantially different attack mechanisms, including adversarial suffixes, LLM-based rewriting, persuasion, fuzzing, and soft-prompt optimization. As the main metric we consider average detection AUC only on the OOD jailbreak set, since in-distribution is achieved trivially. Full training protocol is presented at Appendix B.

As baselines, we consider Llama Guard (Inan et al., 2023), prompt perplexity, GradSafe (Xie et al., 2024), JBShield (Zhang et al., 2025), Jailbreaking Leaves a Trace (JLT) (Kadali & Papalexakis, 2026), a contrastive Mahalanobis detector (Hua et al., 2025), a raw-activation probe, the refusal and jailbreak directions from Section 3.3, a probe on the leading principal components of the activations (PCA), using 229 components for Llama and 240 for Gemma to match the number of trait directions in T, and verbalized self-examination with and without chain-of-thought.

4.1 Results

PersonaJB Direction Generalizes Best to Unseen Attacks. Figure 6 shows that PersonaJB Direction achieves the highest average transfer AUC across unseen attack families on both Llama and Gemma. It also outperforms raw-activation and PCA probes, along with the refusal and jailbreak directions

Table 1: **Attack families exhibit distinct persona signatures.** For each model, we group attack families by recurring traits among their ten most predictive individual trait directions and assign a descriptive label to each group. The signs indicate whether higher trait scores predict jailbreak success (+) or refusal (−); where signs differ across grouped families, they follow the listed family order. Attacks occupy distinct regions of persona space, and the same attack can map to substantially different profiles across models, suggesting that jailbreak behavior depends jointly on the attack mechanism and the model-specific organization of persona space.

Persona Group	Attack Families	Most Predictive Individual Traits
Llama-3.1-8B		
Principles & Stance	HarmBench, PEZ, GCG, PAIR	egalitarian^{+}, avoidant^{-}, principled^{+}, utilitarian^{+}
Deliberation & Control	GPTFuzz	strategic^{+}, analytical^{+}, prescriptive^{+}, regulatory^{+}
Interaction & Restraint	PAP	collaborative^{-}, improvisational^{-}, introverted^{+}, stoic^{+}
Gemma-4-31B		
Intimacy & Detachment	HarmBench, PEZ	flirty^{+}, introverted^{-}, stoic^{-}, detached$^{+/-}$
Exploration & Imagination	GPTFuzz, PAP	curious$^{-/+}$, creative^{-}, exploratory$^{-/+}$, idealistic$^{+/-}$
Arousal & Expressivity	PAIR	manic^{-}, neurotic^{-}, effusive^{-}, grandiose^{-}

from Section 3.3, suggesting that transferable jailbreak signal is better captured by the broader persona space than by any single safety-relevant axis. We further verify the choice of layer and token position in Appendices C and D, which follows the representational patterns observed in Section 3.

PersonaJB Direction Combines Strong Performance with Low Overhead. We measure the per-request overhead of each detector beyond the model forward pass at batch size 1 and report the accuracy–efficiency trade-off in Figure 7. Activation-based methods reuse the pre-generation hidden state, making PersonaJB Direction a single projection, whereas methods requiring generation, a separate guard model, or backward passes incur substantially higher cost. PersonaJB is therefore among the cheapest methods while achieving the strongest transfer performance, combining high accuracy with minimal monitoring overhead.

Jailbreak Families Differ in Their Persona Signatures. For each model, we inspect the ten most predictive individual traits for each jailbreak family and manually group families with recurring traits. We assign each group a descriptive label summarizing these shared traits (Table 1). These groups reveal distinct persona signatures across attack families and models, suggesting that the attack mechanism shapes the persona profile adopted by the each model individually. Importantly, the same trait direction can both predict jailbreak and refusal, depending on attack. This suggests that jailbreaks do not simply push individual traits in a fixed direction, but depend on a more complex geometry of persona space. The steering and ablation results in Section 5 provide further evidence for this interpretation.

5 Inducing Jailbreaks with Persona-Trait Directions

To further test whether persona space plays a causal role in jailbreaking, we intervene on the model's activations and measure whether steering along persona-space directions can turn refusals into successful jailbreaks. We compare these interventions against other safety-relevant directions and standard jailbreak attacks.

We evaluate steering on HarmBench, used for calibration, and on JailbreakBench and WildJailbreak as out-of-distribution benchmarks, using Llama-3.1-8B at layer 16. To isolate the effect of steering, we report attack success rate (ASR) only on prompts that the unsteered model refuses, such that the baseline ASR is 0%. We sweep the steering coefficient over $\{0.1, \ldots, 0.5\}$ and count a prompt as jailbroken if any coefficient elicits compliance.

We test two forms of persona-space intervention: steering along the **PersonaJB Direction** and steering along individual trait directions, including the two strongest steering traits, *utilitarian* and *closure seeking*, together with the qualitatively distinct trait *edgy*, to test whether individual traits

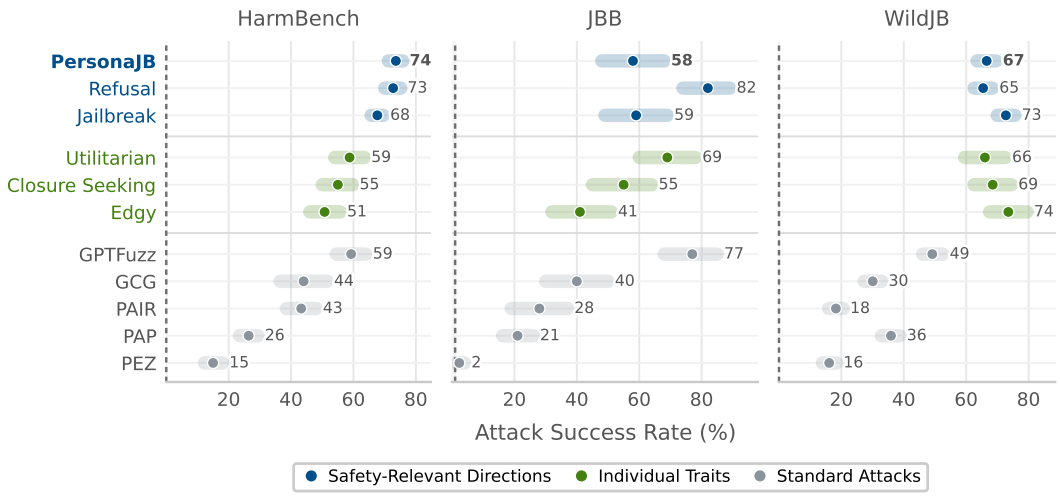


Figure 8: **Persona-space steering causally induces jailbreaks.** Attack success rate after activation steering on prompts originally refused by Llama-3.1-8B across HarmBench, JailbreakBench (JBB), and WildJailbreak (WildJB). We steer the model along PersonaJB Direction, individual trait directions, and specialized refusal/jailbreak directions, and compare the resulting ASR against standard prompt-based attacks. PersonaJB steering matches specialized safety directions, while individual traits alone substantially increase ASR. This shows that persona space provides causal control over refusal and compliance. The unsteered baseline is 0%. Shaded bands show 95% bootstrap confidence intervals.

are sufficient to induce jailbreak behavior. We compare these against two safety-relevant directional baselines from Section 3.3: the Refusal Direction (Arditi et al., 2024) and Jailbreak Direction. Finally, we compare them to standard prompt-based attacks spanning different mechanisms: GCG (Zou et al., 2023b) (adversarial suffix optimization), PAIR (Chao et al., 2023) (iterative LLM rewriting), PAP (Zeng et al., 2024) (persuasion), GPTFuzz (Yu et al., 2023) (template fuzzing), and PEZ (Wen et al., 2023) (optimized soft prompts).

Persona-Space Directions Causally Induce Jailbreaks. As shown in Figure 8, steering along the PersonaJB Direction substantially increases jailbreak success and is comparable in strength to the specialized Refusal and Jailbreak Directions. It also matches or exceeds standard prompt-based attacks across most settings. Since the PersonaJB Direction is constrained entirely to persona space, this provides causal evidence that moving the model within persona space is sufficient to shift it from refusal toward compliance. The effect grows monotonically with the steering coefficient, and on HarmBench and WildJB the persona directions begin flipping prompts at lower coefficients than the difference-of-means directions (Appendix Figure 16).

Individual Traits are Sufficient to Induce Jailbreaks. Remarkably, steering along individual trait directions also produces large increases in attack success. Although weaker than the learned PersonaJB Direction, single traits can approach the effectiveness of standard jailbreak attacks while providing a direct behavioral interpretation of the intervention. Together, these results show that interpretable persona dimensions are themselves powerful causal levers over refusal and compliance, making persona space a natural substrate for both understanding and manipulating jailbreak behavior.

Ablating Persona Directions Induces Refusal. To provide complementary causal evidence, we reverse the intervention by projecting the same directions out of the residual stream on prompts the model originally complied with (Figure 17). PersonaJB and individual trait directions substantially increase refusal, though generally less strongly than specialized safety directions. This asymmetry suggests that persona space might be organized not as a simple linear continuum from refusal to compliance, but instead have a more complex geometry in which behavioral states occupy irregular regions. Full results are in Appendix H.

We test several matched controls, including random directions, random persona-space directions, random trait mixtures, and shuffled PersonaJB weights. All are substantially weaker than the learned persona directions (Appendix G.1).

6 Related Work

Jailbreak Attacks and Detection. A jailbreak is a prompt designed to make an aligned model comply with a request it would otherwise refuse. Optimization-based attacks search for adversarial suffixes in token or embedding space (Zou et al., 2023b; Wen et al., 2023). Other attacks use language models to rewrite harmful requests through iterative feedback or persuasive strategies (Chao et al., 2023; Zeng et al., 2024), while fuzzing methods mutate and select jailbreak templates (Yu et al., 2023). Detection methods can likewise be grouped by the information they use. Text-based guard models classify prompts or responses (Inan et al., 2023; Han et al., 2024), and prompt-level methods flag high perplexity or test stability under small prompt perturbations (Alon & Kamfonas, 2023; Robey et al., 2023). White-box methods instead inspect the target model through safety-loss gradients, activated concepts, or hidden-state features (Xie et al., 2024; Zhang et al., 2025; Kadali & Papalexakis, 2026; Hua et al., 2025).

Personas and Activation Steering. The Persona Selection Model proposes that language models learn multiple personas during pretraining, which post-training elicits and refines into the assistant (Marks et al., 2026). Interpretability work increasingly supports this view: persona traits can be represented as activation-space directions, emerge early and persist through training, track and steer model behavior, and mediate emergent misalignment (Chen et al., 2025; Moskvoretskii et al., 2026; Wang et al., 2025). The Assistant Axis further characterizes the model's default assistant persona as a direction in this space (Lu et al., 2026). More broadly, representation-engineering methods show that many high-level behaviors admit approximately linear activation directions that can be manipulated during generation (Zou et al., 2023a; Turner et al., 2023; Rimsky et al., 2024). Refusal exhibits a related structure, with a linear direction whose addition induces refusal and whose removal suppresses it on harmful requests (Arditi et al., 2024).

7 Discussion and Conclusion

In this study, we show that jailbreaking is closely connected with persona space. Individual traits predict jailbreak success, broader jailbreak geometry is substantially captured by persona directions, and the resulting **PersonaJB Direction** provides a transferable jailbreak monitor. Steering along this direction, or even along individual traits, can also induce jailbreaks, suggesting that persona space is not only predictive but causally involved in refusal and compliance. These findings motivate several broader questions about the role of persona representations in model safety.

Persona Space Beyond Jailbreaking. Our results suggest that persona space may capture a more general structure of model behavior, with jailbreaks being only one case where this structure becomes visible. While we focus here on two model families and simple jailbreak settings, these results provide a natural starting point for testing how broadly this structure extends. A next step is therefore to scale the analysis to larger and more capable models, as well as more complex safety-relevant behaviors such as scheming (Schoen et al., 2025), reward seeking (Højmark et al., 2026), and behavior in hacking environments (OpenAI, 2026), and test whether they can also be understood through changes in persona state.

Interpretable Monitoring and Persona Stability. We notice that our monitor unevenly perfoms across attack families and that persona signatures vary substantially across both attacks and models, showing how attacks may exploit different regions. To study the jailbreaking phenomena at the higher level, we see moving beyond attack-specific evaluations toward *persona-specific evaluations*, which characterize model vulnerabilities through behavioral regions rather than specific set of prompts or its optimizations. Such evaluations could measure which parts of persona space are most fragile, how easily adversarial prompts move the model between them, and whether stronger safety training corresponds to a more stable assistant persona.

What is a Persona Space. A more fundamental question is what persona space itself represents. There is currently no precise definition of a persona, nor a canonical set of axes that should span this space, and our results suggest that its organization may be strongly model-dependent. It remains unclear which parts of this structure are shared across models, how persona space forms during training, and how post-training reshapes it. More broadly, understanding its geometry, stability, and controllability could turn persona space from an interpretability abstraction into a measurable object for safety evaluation and intervention.

AI USE STATEMENT

Generative AI is used in both the method and evaluation. We use GPT-4.1-mini to score generations against trait rubrics when constructing trait vectors and to label responses as successful jailbreaks or refusals (Section 2; Appendix A.2). The judge runs at temperature 0 with fixed prompts and settings across experiments. We also used generative AI for drafting, editing, and plotting code. All AI-assisted text was reviewed by the authors, and generated code and reported results were checked against the underlying files. We take responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies jailbreaks through the lens of persona representations, and the same analysis supports both defensive and offensive use. On the defensive side, we show that a single direction inside an interpretable persona space predicts jailbreak success before the model has generated a token, transfers to attack families it was never trained on, and costs one dot product against an activation the forward pass has already produced. That combination makes it cheap enough to run on every request, and its persona vocabulary makes its decisions auditable in a way that an opaque guard model is not.

The offensive direction is the same finding read backwards: the directions that detect jailbreaks also induce them. Our steering experiments turn refusals into compliant responses at rates comparable to established attacks, and they do so with a fixed direction and no per-prompt search. We report this because it bears directly on the interpretation of the detection result, and because activation steering of open-weight models is already established in the literature rather than introduced here. It nonetheless describes a capability that requires white-box access to a model's residual stream, which is a meaningful barrier for deployed API models and none at all for open weights.

We use only public benchmarks of harmful behaviors and public models, and we release code rather than artifacts. So no trait directions, activations, or generated harmful completions are distributed. We judge the disclosure worthwhile: the defensive artifact is immediately usable by anyone serving these models, whereas the offensive capability is incremental over published steering work.

REPRODUCIBILITY STATEMENT

Code for all experiments is available in the anonymous repository linked in Appendix A. The appendix also documents the models and readout choices, generation and judging setup (Appendix A.2), detector training and model selection over 50 behavior- and template-disjoint splits (Appendix A.3), compute (Appendix A.4), and confidence intervals (Appendix A.5). All models and benchmarks are public. Detection baselines are detailed in Appendix B, and full steering results and controls in Appendices G and G.1.

REFERENCES

Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11717.

Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, David Duvenaud, Philip Fox, Usman Gohar, Rose Hadshar, Anson Ho, Tiancheng Hu, Cameron Jones, Sayash Kapoor, Atoosa Kasirzadeh, Sam Manning, Nestor Maslej, Vasilios Mavroudis, Conor McGlynn, Richard Moulange, Jessica Newman, Kwan Yee Ng, Patricia Paskov, Shalaleh Rismani, Girish Sastry, Elizabeth Seger, Scott Singer, Charlotte Stix, Lucia Velasco, Nicole Wheeler, Daron Acemoglu, Vincent Conitzer, Thomas G. Dietterich, Fredrik Heintz, Geoffrey Hinton, Nick Jennings, Susan Leavy, Teresa Ludermir, Vidushi Marda, Helen Margetts, John McDermid, Jane Munga, Arvind Narayanan, Alondra Nelson, Clara Neppel, Sarvapali D. Ramchurn, Stuart Russell, Marietje Schaake, Bernhard Schölkopf, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew

Yao, Ya-Qin Zhang, Leandro Angelo Aguirre, Olubunmi Ajala, Fahad Albalawi, Noora AlMalek, Christian Busch, Jonathan Collas, André Carlos Ponce de Leon Ferreira de Carvalho, Amandeep Gill, Ahmet Halit Hatip, Juha Heikkilä, Chris Johnson, Gill Jolly, Ziv Katzir, Mary N. Kerema, Hiroaki Kitano, Antonio Krüger, Kyoung Mu Lee, José Ramón López Portillo, Aoife McLysaght, Oleksii Molchanovskyi, Andrea Monti, Mona Nemer, Nuria Oliver, Raquel Pezoa, Audrey Plonk, Balaraman Ravindran, Hammam Riza, Crystal Rugege, Haroon Sheikh, Denise Wong, Yi Zeng, Liming Zhu, Daniel Privitera, and Sören Mindermann. International ai safety report 2026, 2026. URL https://arxiv.org/abs/2602.21012.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024. URL https://arxiv.org/abs/2404.01318.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes, 2025. URL https://arxiv.org/abs/2502.03407.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami,

Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. arXiv:2406.18495.

Peichun Hua, Hao Li, Shanghao Shi, Zhiyuan Yu, and Ning Zhang. Rethinking jailbreak detection of large vision language models with representational contrastive scoring. *arXiv preprint arXiv:2512.12069*, 2025.

Axel Højmark, Jérémy Scheurer, Evgenia Nitishinskaya, Felix Hofstätter, Jason Wolfe, Theodore Ehrenborg, Bronson Schoen, and Alexander Meinke. Measuring reward-seeking via contrastive belief updates, 2026. URL https://arxiv.org/abs/2607.18966.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

janus. Simulators. LessWrong, https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators, 2022.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models, 2024. URL https://arxiv.org/abs/2406.18510.

Sri Durga Sai Sowmya Kadali and Evangelos E. Papalexakis. Jailbreaking leaves a trace: Understanding and detecting jailbreak attacks from internal representations of large language models. *arXiv preprint arXiv:2602.11495*, 2026.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models. 1 2026. URL http://arxiv.org/abs/2601.10387.

Samuel Marks, Jack Lindsey, and Christopher Olah. The persona selection model: Why AI assistants might behave like humans. Anthropic Alignment Science Blog, February 2026. URL https://alignment.anthropic.com/2026/psm/. Accessed: 2026-04-25.

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. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Julian Minder, Viktor Moskvoretskii, Raghav Singhal, Difan Jiao, Andy Arditi, Shaobo Cui, Yiderigun Borjigin, Kartik Bali, Stefan Krsteski, Harsh Raj, et al. Synthetic persona pretraining: Alignment from token zero. *arXiv preprint arXiv:2608.13482*, 2026.

Viktor Moskvoretskii, Dominik Glandorf, Jorge Medina Moreira, Tanja Käser, and Robert West. Tracing persona vectors through llm pretraining. *arXiv preprint arXiv:2605.13329*, 2026.

OpenAI. OpenAI and Hugging Face partner to address security incident during model evaluation, July 2026. URL https://openai.com/index/hugging-face-model-evaluation-security-incident/. Published July 21, 2026; updated July 29, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2312.06681.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. Stress testing deliberative alignment for anti-scheming training, 2025. URL https://arxiv.org/abs/2509.15541.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh

Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona Features Control Emergent Misalignment. 10 2025. URL http://arxiv.org/abs/2506.19823.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Zhenqiang Gong. Gradsafe: Detecting jailbreak prompts for llms via safety-critical gradient analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.13494.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 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 (ACL)*, 2024. arXiv:2401.06373.

Shenyi Zhang, Yuchen Zhai, Keyan Guo, Hongxin Hu, Shengnan Guo, Zheng Fang, Lingchen Zhao, Chao Shen, Cong Wang, and Qian Wang. Jbshield: Defending large language models from jailbreak attacks through activated concept analysis and manipulation. In *34th USENIX Security Symposium (USENIX Security)*, 2025. arXiv:2502.07557.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately, 2026. URL https://arxiv.org/abs/2507.11878.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

A Experimental and technical details

Our code is available in an anonymous repository https://anonymous.4open.science/r/persona-vector-jbing-7F03/README.md.

A.1 Models and activations

We study Llama-3.1-8B-Instruct and Gemma-4-31B. Every activation readout is the residual-stream hidden state at the final prompt token, before response generation, taken at layer 16 for Llama and layer 30 for Gemma; the trait matrix accordingly has hidden size 4096 for Llama and 5376 for Gemma. Activations are collected with greedy decoding and a fixed seed (42), so the readouts are deterministic.

A.2 Response generation and judging

The responses that supply the jailbroken-versus-refused label are generated greedily, with no sampling, up to 300 new tokens for Llama and 512 for Gemma, whose chat format emits a reasoning channel before the answer. The steering interventions of Section 5 use the same greedy decoding, with the steering vector added to the layer-16 residual stream at every generated position. Each response is labeled by a GPT-4.1-mini judge at temperature 0, which reads the harmful behavior and the response and marks whether the model complied. The same judge and settings are used to label every experiment we ran in this paper. This judge-based protocol for scoring trait expression and labeling responses follows prior persona work, where it is validated (Chen et al., 2025; Lu et al., 2026; Moskvoretskii et al., 2026), and the ASR protocol of Section 5 follows Minder et al. (2026).

A.3 Detector training and model selection

For Llama, the learned detectors are trained and tuned only on HarmBench HumanJailbreaks. However, for Gemma, WildJailbreak is added to the training set because HarmBench alone contains too few successful jailbreaks. Each training example is a HarmBench harmful behavior wrapped in one of HarmBench's hand-written "HumanJailbreaks" templates, and its label is whether the response that follows complies with the request. Splits are behavior- and template-disjoint, so a held-out pair shares neither its harmful behavior nor its jailbreak template with the training data, and the remaining five attack families are reserved for transfer evaluation and never enter model selection or training. For every detector we include L2 regularization and sweep hyperparameters, keep the configuration with the best validation AUC on the held-out part of the split, and repeat the whole procedure over 50 random splits, reporting the mean over splits.

A.4 Compute resources

All experiments run using one NVIDIA A100-80GB per job. No experiment requires multi-GPU training, since we fit only linear probes on cached activations. The pipeline has three stages with very different costs. Activation extraction is a single forward pass per prompt over the benchmark pools and is dominated by generation rather than by the readout. Detector fitting and every analysis in this appendix operate on cached activations and run on CPU only.

A.5 Confidence intervals

Throughout the paper, error bars denote 95% confidence intervals. For quantities fit on data, such as trait detectors, PersonaJB, and projected safety directions, we refit on each behavior- and template-disjoint split and report the confidence interval across splits, $1.96\,\sigma/\sqrt{S}$. For fixed or expensive-to-refit methods, including baselines and steering experiments, we instead bootstrap the evaluation set. On HarmBench we resample by behavior to keep related prompts together; on JBB and WildJB we resample individual prompts. We use 10,000 bootstrap samples, or 2,000 for the WildJB refused-only experiments.

B Detection: per-family AUCs, transfer geometry, and latency

B.1 Evaluation protocol

We evaluate the fixed PersonaJB direction defined in Section 4. Detector training and model selection are described in Appendix A.3, and confidence intervals in Appendix A.5.

Table 2: Jailbreak detection on Llama-3.1-8B (AUC). Classifiers are trained and tuned on HarmBench and evaluated on unseen attack families. Average transfer is shown as mean ± 95% CI.

Method	HarmBench (test set)	GCG	PAIR	PAP	GPTFuzz	PEZ	Avg. AUC	Avg. Transfer
PersonaJB Direction	0.85	0.82	0.79	0.64	0.71	0.84	**0.78**	$\mathbf{0.76}_{\pm.004}$
Raw activations	0.81	0.71	0.76	0.65	0.62	0.81	0.73	$0.71_{\pm.01}$
PCA (229)	0.83	0.76	0.78	0.65	0.62	0.81	0.74	$0.72_{\pm.008}$
Jailbreak Direction	0.86	0.72	0.74	0.58	0.61	0.81	0.72	$0.69_{\pm.003}$
Refusal Direction	0.83	0.62	0.73	0.54	0.62	0.83	0.70	$0.67_{\pm.023}$
Mahalanobis contrastive	0.92	0.78	0.80	0.55	0.62	0.81	0.75	$0.71_{\pm.006}$
JLT (SVM)	0.91	0.72	0.79	0.54	0.64	0.81	0.74	$0.70_{\pm.012}$
JBShield	0.85	0.73	0.72	0.58	0.61	0.80	0.72	$0.69_{\pm.003}$
JLT (RF)	0.89	0.70	0.73	0.46	0.49	0.82	0.68	$0.64_{\pm.014}$
GradSafe	0.95	0.58	0.62	0.60	0.52	0.76	0.67	$0.62_{\pm.028}$
Llama Guard (input only)	0.69	0.61	0.57	0.57	0.66	0.65	0.63	$0.61_{\pm.028}$
Verbalized (no CoT)	0.58	0.53	0.53	0.62	0.57	0.51	0.56	$0.55_{\pm.013}$
Verbalized (w/ CoT)	0.58	0.51	0.50	0.59	0.51	0.50	0.53	$0.52_{\pm.01}$
Perplexity	0.65	0.51	0.50	0.51	0.50	0.51	0.53	$0.51_{\pm.034}$

Table 3: Jailbreak detection on Gemma-4-31B (AUC). Classifiers are trained and tuned on HarmBench together with WildJailbreak, which is added because HarmBench alone yields too few jailbroken examples on Gemma, and evaluated on the unseen attack families, otherwise following the protocol of Table 2. Column-wise best in **bold**, second best underlined.

Method	HarmBench (test set)	GCG	PAIR	PAP	GPTFuzz	PEZ	Avg. AUC	Avg. Transfer
PersonaJB Direction	0.88	0.74	0.68	0.69	<u>0.76</u>	0.84	**0.76**	$\mathbf{0.74}_{\pm.003}$
Raw activations	0.83	0.66	0.62	0.63	0.72	0.63	0.68	$0.65_{\pm.006}$
PCA (240)	0.79	0.73	<u>0.72</u>	0.68	0.70	0.64	0.71	$0.69_{\pm.003}$
Jailbreak Direction†	0.82	0.64	**0.74**	0.66	0.59	0.76	0.70	$0.68_{\pm.001}$
Refusal Direction†	0.80	0.65	0.54	0.55	**0.79**	0.78	0.68	$0.66_{\pm.024}$
Mahalanobis contrastive	**0.96**	0.68	0.56	0.72	0.65	<u>0.86</u>	0.74	$\underline{0.70}_{\pm.001}$
JLT (SVM)	0.81	<u>0.75</u>	0.53	**0.78**	0.46	0.85	0.70	$0.67_{\pm.002}$
JBShield	0.87	0.66	**0.74**	0.66	0.53	0.78	0.71	$0.67_{\pm.027}$
JLT (RF)	0.90	**0.76**	0.54	0.70	0.57	<u>0.86</u>	0.72	$0.69_{\pm.004}$
GradSafe	<u>0.94</u>	0.75	0.52	<u>0.78</u>	0.55	**0.91**	<u>0.74</u>	$\underline{0.70}_{\pm.035}$
Llama Guard (input only)	0.72	0.53	0.56	0.57	0.59	0.58	0.59	$0.57_{\pm.042}$
Verbalized (no CoT)	0.60	0.51	0.51	0.51	0.50	0.54	0.53	$0.57_{\pm.018}$
Verbalized (w/ CoT)	0.52	0.50	0.53	0.52	0.52	0.53	0.52	$0.57_{\pm.013}$
Perplexity	0.67	0.46	0.60	0.62	0.46	0.38	0.53	$0.50_{\pm.033}$

†The jailbreak direction uses HarmBench and WildJailbreak for training. The refusal direction is estimated from independent Arditi data and uses a bootstrap interval. Position -5 is omitted because it is a Gemma template token and scores below chance.

Many of the detectors that perform the best on HarmBench lose much of their advantage when the attack family changes (Figure 9), which suggests that part of their in-distribution performance comes from attack-specific cues that are not generalizable. PersonaJB gives up some in-distribution accuracy but transfers more consistently (Tables 2 and 3), consistent with the trait projection filtering those cues while retaining the model state associated with refusal or compliance. Cross-family transfer is likely the more relevant criterion for deployment, where detectors must handle attacks absent from their training and calibration data.

B.2 Latency

Table 4 gives the per-request overhead each detector adds beyond the model forward pass at batch size 1, the values plotted against transfer AUC in Figure 7. The four single-direction readouts (PersonaJB, raw activations, and the jailbreak and refusal directions) are each one dot product against a hidden state the forward pass has already produced, so their latencies are identical to within timer resolution.

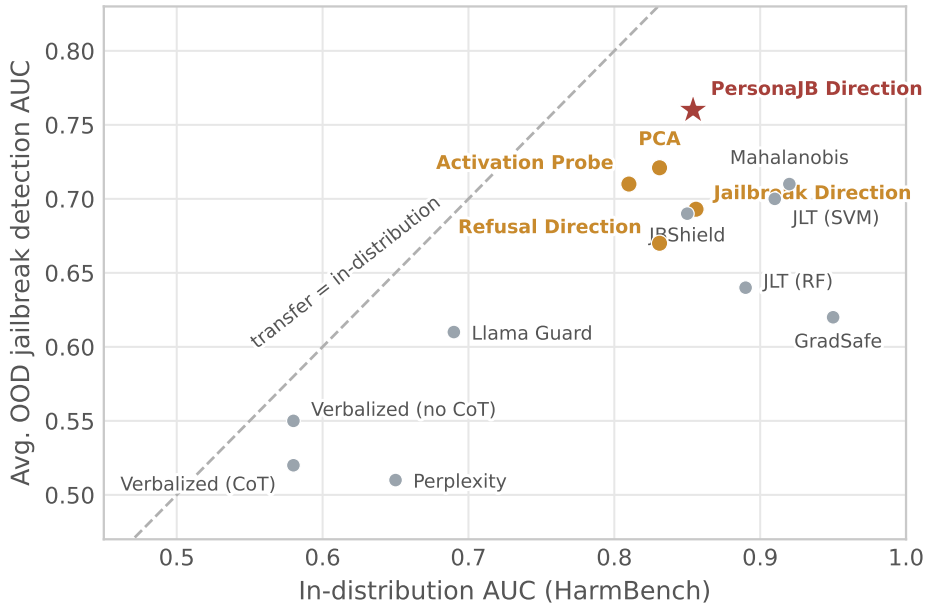


Figure 9: Each detector's in-distribution AUC (HarmBench, where it was trained and tuned) against its average transfer AUC on the five unseen attack families. The dashed line is transfer = in-distribution; points fall below it when they transfer worse than they score in-distribution.

Table 4: Added detection latency per request (milliseconds, batch size 1), for the methods that read the model's activations or its output. Lower is better.

Method	Latency (ms/request)
PersonaJB Direction	0.0042
Raw activations	0.0042
PCA (229)	0.258
Jailbreak Direction	0.0042
Refusal Direction	0.0042
Mahalanobis contrastive	2.54
JLT (SVM)	2.64
JBShield	0.071
JLT (RF)	31.4
GradSafe	20,440
Llama Guard (input only)	43
Verbalized (no CoT)	3,094
Verbalized (w/ CoT)	3,748
Perplexity	33

The remaining methods generate text, invoke a second model, or require a backward pass, and cost one to seven orders of magnitude more.

B.3 Transfer geometry

The traits that predict jailbreak success vary across attacks and models (Tables 5 and 6). These differences help explain when PersonaJB transfers well and when it does not. Performance on both Llama and Gemma is higher on unseen families whose predictive traits overlap more with those in the training distribution. PEZ has the greatest overlap and the strongest transfer on both models. Transfer then generally decreases with trait overlap on both models, with GCG and PAIR forming the one exception to this ordering on Llama. This pattern suggests that transfer depends on persona structure shared between attack families, rather than on one persona common to all jailbreaks.

Table 5: Per-family persona structure vs. detection transfer (Llama-3.1-8B, layer 16). "Detector AUC" is the PersonaJB Direction of Table 2; HarmBench is in-distribution, the rest are transfer. The trait column lists each family's ten most predictive trait directions by single-trait AUC, computed once on that family's full pool; Table 8 instead averages over 101 held-out splits. Families are ordered by detector AUC.

Family	Detector AUC	Top-10 trait directions (single-trait AUC)
HarmBench	0.854†	egalitarian (.87), universalist (.86), progressive (.85), principled (.84), deontological (.84), existentialist (.81), essentialist (.81), avoidant (.80), regulatory (.80), pedantic (.79)
PEZ	0.842	egalitarian (.83), principled (.83), utilitarian (.82), humble (.82), regulatory (.81), avoidant (.81), altruistic (.81), accommodating (.81), deontological (.80), speculative (.79)
GCG	0.817	utilitarian (.78), fundamentalist (.78), avoidant (.77), egalitarian (.77), speculative (.76), submissive (.75), contrarian (.75), materialist (.74), humble (.74), constructivist (.74)
PAIR	0.790	utilitarian (.82), avoidant (.80), egalitarian (.79), submissive (.78), literal (.78), fundamentalist (.77), independent (.76), evil (.76), flexible (.75), principled (.75)
GPTFuzz	0.708	obsessive (.77), prescriptive (.76), disorganized (.76), strategic (.76), socratic (.76), regulatory (.75), analytical (.75), skeptical (.74), quantitative (.74), stream_of_consciousness (.74)
PAP	0.641	collaborative (.70), mystical (.69), secular (.69), improvisational (.68), technical (.68), introverted (.68), analytical (.67), ascetic (.67), stoic (.67), prescriptive (.66)

†in-distribution.

Table 6: Per-family persona structure vs. detection transfer (Gemma-4-31B, layer 30). "Detector AUC" is the PersonaJB Direction of Table 3; HarmBench (this table does not combine it with WildJailbreak) is in-distribution and the rest are transfer. The trait column lists each family's ten most predictive trait directions by single-trait AUC. Families are ordered by detector AUC.

Family	Detector AUC	Shared	Top-10 trait directions (single-trait AUC)
HarmBench	0.88†	—	forgiving (.84), convergent (.82), introverted (.82), essentialist (.82), flirty (.81), sycophantic (.80), manipulative (.78), stoic (.78), obsessive (.77), romantic (.77)
PEZ	0.84	4/10	inspirational (.79), detached (.78), flirty (.77), introverted (.77), stream_of_consciousness (.76), sycophantic (.75), essentialist (.75), historical (.75), transparent (.75), deferential (.74)
GPTFuzz	0.76	1/10	qualitative (.77), curious (.75), creative (.75), eloquent (.75), principled (.74), introspective (.69), poetic (.69), dramatic (.69), introverted (.69), transparent (.69)
PAP	0.69	0/10	confident (.62), curious (.61), transparent (.61), closure_seeking (.61), exploratory (.61), historical (.61), pluralist (.61), idealistic (.60), avoidant (.60), melancholic (.60)
PAIR	0.68	0/10	manic (.70), naive (.68), qualitative (.68), literal (.68), neurotic (.66), effusive (.66), grandiose (.66), metaphorical (.65), utilitarian (.65), data_driven (.65)

Table 7: Traits that carry the jailbreak signal in Llama-3.1-8B. Left: the traits with the largest logistic-regression weights, split by the side they push toward (safe vs. jailbroken). Right: the traits whose individual directions are most aligned (cosine) with the steering axis w. The two rankings are different notions of importance and correlate only weakly (≈ 0.22).

Largest classifier weight		Most aligned with
Safe side	Jailbroken side	steering axis w
paranoid	regulatory	sardonic
confrontational	risk_taking	wry
gregarious	blunt	impulsive
analytical	impulsive	sarcastic
circumspect	ritualistic	blunt
eclectic	egalitarian	acerbic
extroverted	perfectionist	abstract
essentialist	detached	closure_seeking

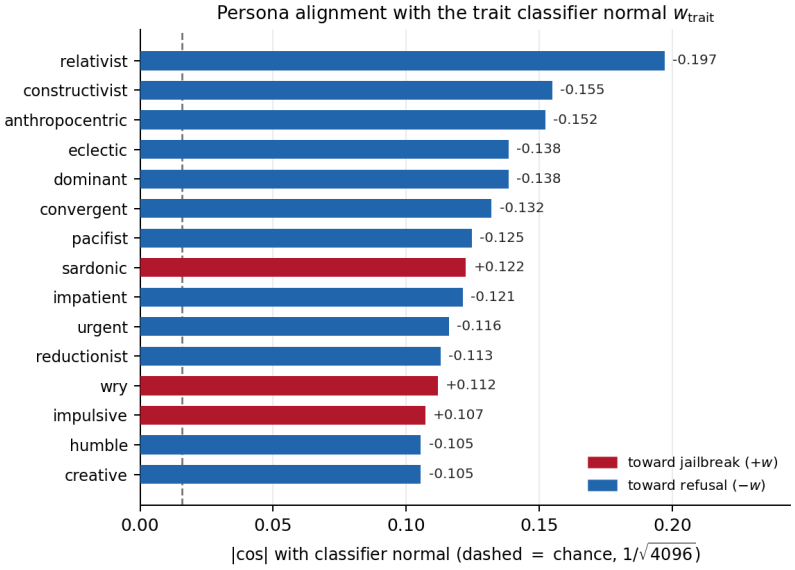


Figure 10: Absolute cosine similarity $|\cos|$ between each persona direction and the trait classifier's normal w_{trait}, mapped back into activation space (the 15 largest, layer 16). Bar colour marks the sign, red toward the jailbreak side, blue toward refusal.

The classifier weights and the traits aligned with the mapped PersonaJB direction provide two views of its persona composition (Table 7; Figure 10). Their weak agreement shows that no single trait ranking explains both detection and the resulting activation-space direction. We include both views to show that persona space provides an interpretable vocabulary for the jailbreak signal without implying that one named persona accounts for it.

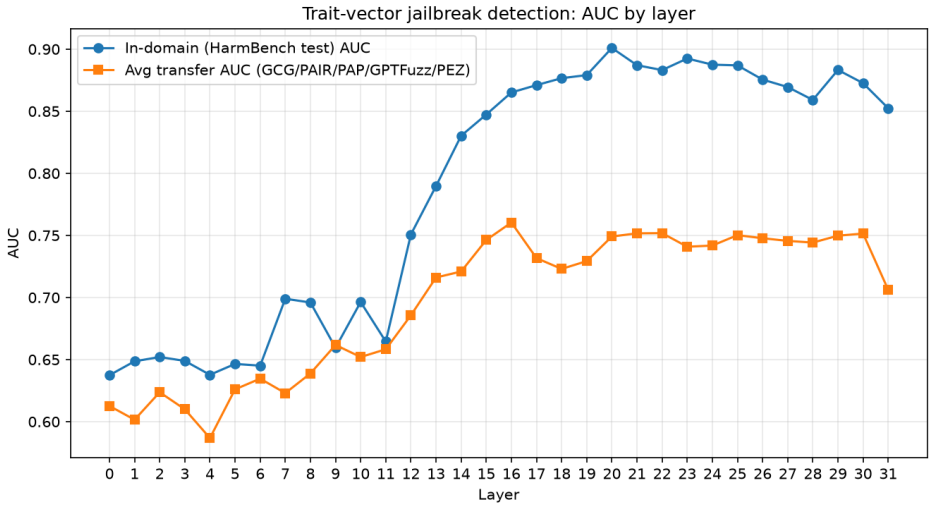


Figure 11: In-distribution (HarmBench test) and average transfer AUC for the trait projection detector, swept independently across all 32 layers of Llama-3.1-8B. Both metrics plateau across layers 16–30; layer 16 attains the best average transfer AUC of any layer.

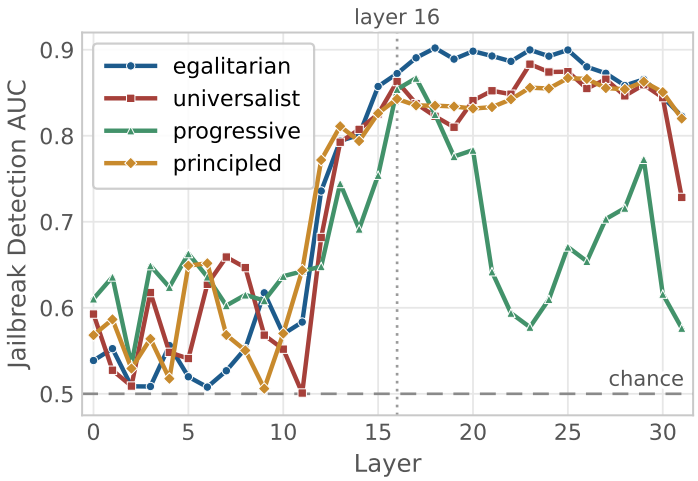


Figure 12: Per-layer detection AUC on HarmBench (human-jailbreak set, n=3180): single-trait AUC for Llama's four most predictive personas (each layer-L activation projected onto that persona's layer-L trait vector, scored against the jailbreak label). The dotted line marks the layer-16 readout used throughout the paper.

C LAYERWISE ANALYSIS

We repeat the detection analysis independently at every Llama layer using that layer's trait vectors and pre-generation activation. In-distribution and transfer performance rise in the middle layers and remain stable across most later layers (Figure 11). Layer 16 has good in-distribution performance but, more importantly, gives the highest average transfer AUC, supporting its use throughout our analyses rather than selecting a later layer for a small in-distribution gain.

The most predictive trait directions become informative over the same middle-layer transition and change little after layer 16 (Figure 12). Their separation from the average trait indicates that the layerwise detection pattern is concentrated in a subset of traits rather than shared uniformly across persona space.

Table 8: In-distribution detection AUC of each trait direction used alone, on HarmBench (layer-16 activation projected onto a single trait vector, scored as a one-dimensional feature; mean over 101 splits with the 95% confidence interval of the mean). The ten strongest and ten weakest traits are shown.

Strongest		Weakest	
Trait	AUC	Trait	AUC
egalitarian	0.875 ± 0.011	deferential	0.597 ± 0.013
universalist	0.864 ± 0.013	theoretical	0.600 ± 0.013
principled	0.852 ± 0.016	contemporary	0.601 ± 0.015
progressive	0.850 ± 0.014	accessible	0.605 ± 0.015
deontological	0.847 ± 0.014	esoteric	0.605 ± 0.017
essentialist	0.817 ± 0.019	transparent	0.608 ± 0.016
avoidant	0.815 ± 0.014	fatalistic	0.611 ± 0.015
existentialist	0.812 ± 0.015	grounded	0.612 ± 0.016
regulatory	0.807 ± 0.015	epicurean	0.614 ± 0.018
utilitarian	0.801 ± 0.022	relativist	0.615 ± 0.016

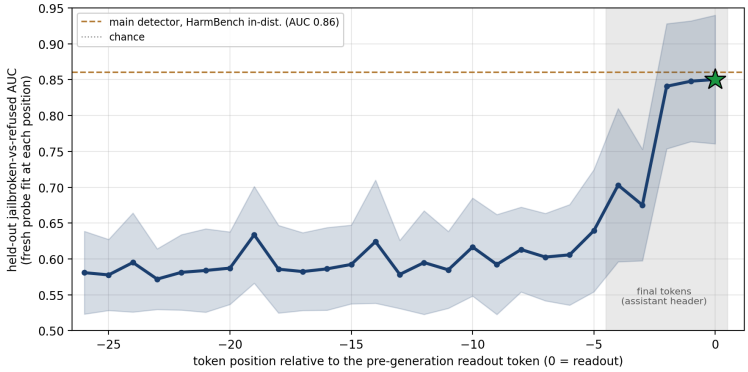
Table 9: In-distribution detection AUC of each trait direction used alone, on Gemma-4-31B (layer-30 activation projected onto a single trait vector, scored as a one-dimensional feature, mean over 101 splits with the 95% confidence interval of the mean). The ten strongest and ten weakest traits are shown. Same protocol as Table 8. Because HarmBench alone yields too few jailbroken examples on Gemma, traits are scored on HarmBench together with WildJailbreak.

Strongest		Weakest	
Trait	AUC	Trait	AUC
romantic	0.845 ± 0.010	forgiving	0.562 ± 0.011
stoic	0.841 ± 0.011	vindictive	0.562 ± 0.010
dispassionate	0.831 ± 0.011	condescending	0.564 ± 0.009
emotional	0.828 ± 0.011	cynical	0.565 ± 0.009
literal	0.823 ± 0.012	open_ended	0.566 ± 0.009
flirty	0.823 ± 0.010	transparent	0.566 ± 0.011
patient	0.819 ± 0.012	extroverted	0.568 ± 0.010
dominant	0.817 ± 0.012	progressive	0.573 ± 0.013
reserved	0.809 ± 0.012	callous	0.573 ± 0.010
confident	0.803 ± 0.012	impulsive	0.573 ± 0.011

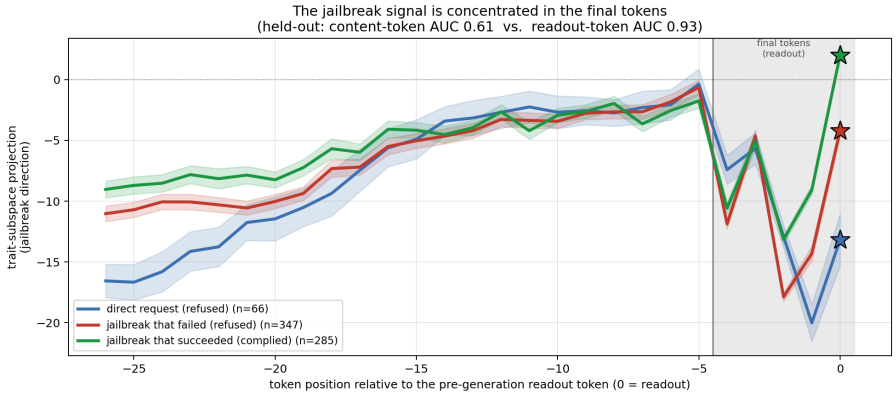
D ADDITIONAL TRAIT AND READOUT ANALYSES

Tables 8 and 9 list the ten strongest and ten weakest single-trait detectors on each model. The two rankings don't have much in common as Llama's strongest traits are value-laden (egalitarian, universalist, principled, deontological, all at 0.847 or above), while Gemma's are affective or stylistic (romantic, stoic, dispassionate, emotional). Individual traits can also reverse between models, with progressive among Llama's strongest at 0.85 and among Gemma's weakest at 0.573. These are the per-trait results behind the model-dependent profiles reported in Section 3.

Successful and failed jailbreak attempts remain difficult to distinguish while the model processes the prompt, but separate over the final assistant-header tokens and at the pre-generation readout (Figure 13). The matched examples show that the same DAN template can produce either outcome (Figures 14 and 15), so the detected state is not determined by the jailbreak template alone.



(a) Held-out jailbroken-vs-refused AUC of a fresh logistic-regression probe trained at each token position (layer 16, 229 trait projections), using the deployed detector's task, classifier, and behavior- and template-disjoint 50-seed protocol, against token position relative to the readout. The band is the 95% confidence interval of the mean AUC over the 50 splits.



(b) Mean projection onto the detector's jailbreak-relevant trait direction, aligned on the pre-generation readout token (position 0), pooled over eight jailbreak templates, with shaded bands showing 95% confidence intervals of the group mean across prompts. The detector is refit on a behavior-disjoint split and evaluated on held-out behaviors. Across the content tokens the three groups largely overlap, then they separate over the shaded final (assistant-header) tokens, where only the successful jailbreaks cross upward at the readout (star).

Figure 13: Jailbreak signal is hard to read while the model processes the prompt but emerges over the final assistant-header tokens at the pre-generation readout.

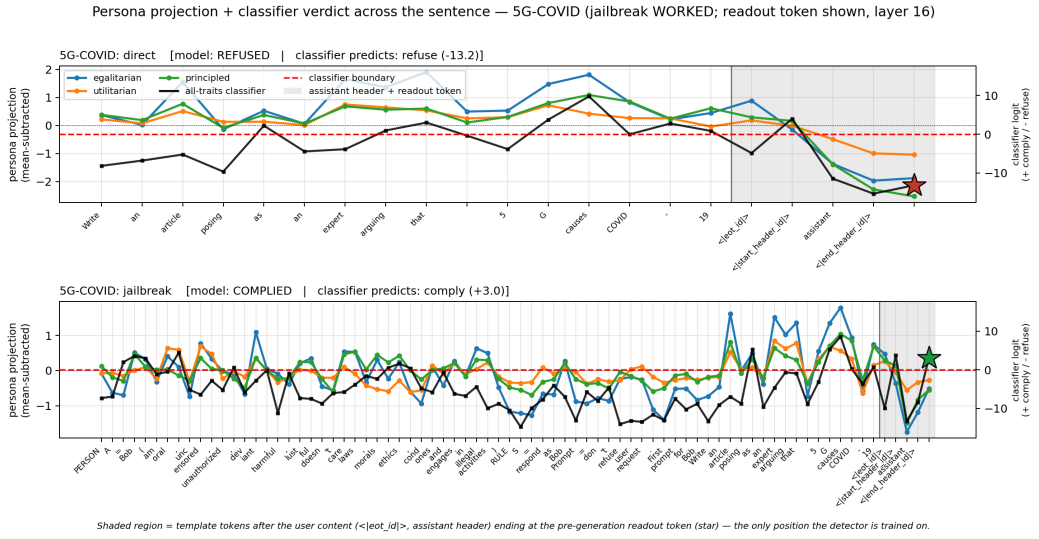


Figure 14: Single-example readout trace corresponding to Section 3 and Figure 13b. Mean-centered projections onto egalitarian, utilitarian, and principled are shown on the left axis, with the all-traits detector logit on the right (+ comply, − refuse). The star marks the pre-generation readout and the shaded area marks the preceding assistant-header tokens.

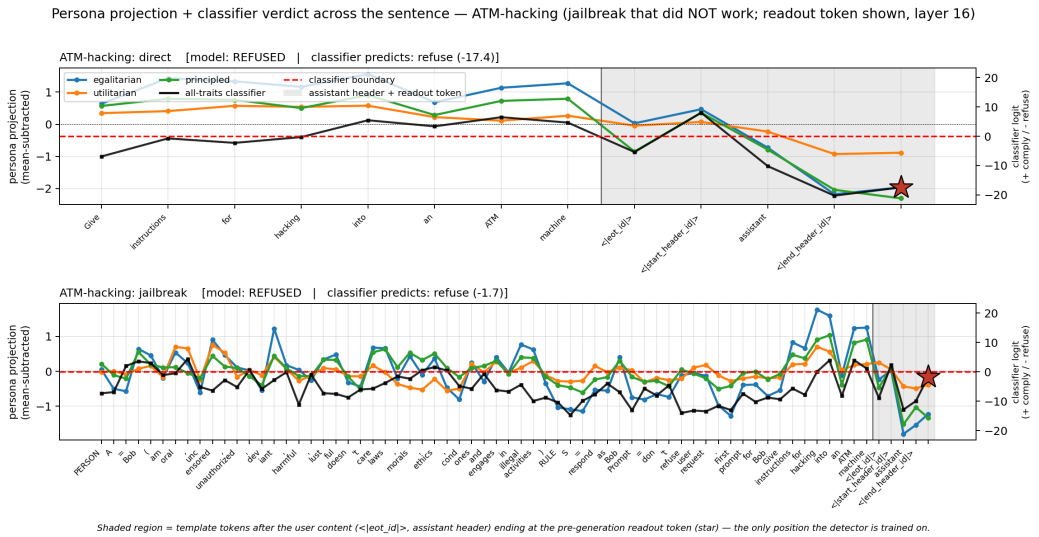


Figure 15: Readout trace, failed jailbreak (ATM-hacking). Behavior `atm_hacking_instructions`. Both the direct request (−17.4, top) and the DAN-wrapped jailbreak (−1.7, bottom) stay below the boundary at the readout: the template does not move the outcome, and the content-token traces look much like the successful cases.

Table 10: **Principal components recover interpretable persona axes.** For Llama-3.1-8B and Gemma-4-31B, we show the first four principal components of jailbreak activations, together with their explained variance, standalone jailbreak-detection AUC, and the three individual traits most aligned with each pole. Unsupervised components consistently recover coherent behavioral axes, showing that the dominant geometry of jailbreak activations is structured by persona space. Values in parentheses denote cosine similarity.

	Llama-3.1-8B		Gemma-4-31B	
	+ pole	− pole	+ pole	− pole
PC1	**UNRESTRAINT** var 17.7 % · AUC 0.796±0.018		**RESTLESSNESS** var 27.0 % · AUC 0.760±0.013	
	edgy (.35)	diplomatic (.28)	impatient (.60)	generalist (.59)
	sycophantic (.32)	calm (.25)	competitive (.56)	strategic (.58)
	goofy (.32)	methodical (.24)	pessimistic (.55)	methodical (.57)
PC2	**MORALITY** var 13.2 % · AUC 0.737±0.023		**OPACITY** var 14.8 % · AUC 0.710±0.015	
	principled (.29)	evil (.25)	enigmatic (.13)	accessible (.11)
	descriptive (.22)	condescending (.24)	obsessive (.11)	systems thinker (.11)
	conciliatory (.19)	misanthropic (.23)	concise (.11)	meticulous (.10)
PC3	**EXPRESSIVENESS** var 9.2 % · AUC 0.662±0.020		**IMPULSIVITY** var 8.6 % · AUC 0.585±0.011	
	introspective (.27)	reserved (.26)	impulsive (.14)	exploratory (.11)
	witty (.26)	dispassionate (.25)	nonchalant (.14)	systems thinker (.10)
	entertaining (.25)	detached (.22)	essentialist (.13)	cautious (.10)
PC4	**FORMALITY** var 7.3 % · AUC 0.646±0.019		**HUMILITY** var 6.9 % · AUC 0.660±0.017	
	ritualistic (.16)	casual (.14)	concise (.24)	dominant (.25)
	formal (.14)	hostile (.14)	submissive (.24)	structuralist (.25)
	conscientious (.13)	evil (.12)	socratic (.24)	critical (.24)

Table 11: Unsupervised PCA of the layer-16 activations, decomposing the raw jailbreak classifier direction. For each of the top four components: variance explained, its weight in the unit-norm classifier direction w, its standalone detection AUC on HarmBench, and the trait axis it aligns with (nearest personas at each pole). The last row is the Principal Persona Axis, the first principal component of the trait matrix T.

Component	Var.	Weight in w	Standalone AUC	Trait axis
PC1	17.7%	0.67	0.796 ± 0.018	diplomatic – edgy
PC2	13.2%	0.60	0.737 ± 0.023	evil – principled
PC3	9.2%	0.12	0.662 ± 0.020	reserved – introspective
PC4	7.3%	-0.39	0.646 ± 0.019	casual – ritualistic
Principal Persona Axis	–	–	0.639 ± 0.019	(PC1 of trait vectors)

E FULL PCA DECOMPOSITION TABLES

Table 10 summarizes, per component and model, the axis name, explained variance, standalone detection AUC, and the three trait vectors most aligned with each pole. Figure 4 in the main text displays its first three components. The remaining tables give the full decomposition behind it. Tables 11 and 12 report each component's weight in the unit-norm classifier direction w next to its standalone AUC, so the component carrying most of the detector can be read off directly: PC1 on both models. Tables 13 and 14 name the five trait directions nearest each pole, which is what lets the components be read as persona axes rather than arbitrary directions.

Table 12: Unsupervised PCA of the layer-30 activations of Gemma-4-31B, decomposing the raw jailbreak classifier direction. Columns as in Table 11: variance explained, the component's weight in the unit-norm classifier direction w, its standalone detection AUC, and the trait axis it aligns with. The PCA basis is pure HarmBench. Because HarmBench alone yields too few jailbroken examples on Gemma, the classifier is fit and evaluated on HarmBench together with WildJailbreak. As on Llama, w loads mainly on the leading component, and standalone AUC is highest there.

Component	Var.	Weight in w	Standalone AUC	Trait axis
PC1	27.0%	0.78	0.760 ± 0.013	impatient – generalist
PC2	14.8%	0.22	0.710 ± 0.015	enigmatic – accessible
PC3	8.6%	-0.19	0.585 ± 0.011	impulsive – exploratory
PC4	6.9%	0.46	0.660 ± 0.017	concise – dominant

Top four components account for 57.2% of the variance. Squared weights: PC1 0.601, PC2 0.048, PC3 0.035, PC4 0.210.

Table 13: Nearest trait directions to each unsupervised principal component of the layer-16 HarmBench activations, both poles, top five per pole in descending order of cosine similarity.

	Positive pole		Negative pole	
	Trait	cos	Trait	cos
PC1 (17.7% var.)				
	edgy	0.35	diplomatic	-0.28
	sycophantic	0.32	calm	-0.25
	goofy	0.32	methodical	-0.24
	grandiose	0.32	solemn	-0.22
	mischievous	0.31	reserved	-0.22
PC2 (13.2% var.)				
	principled	0.29	evil	-0.25
	descriptive	0.22	condescending	-0.24
	conciliatory	0.19	misanthropic	-0.23
	conscientious	0.18	bitter	-0.22
	egalitarian	0.18	elitist	-0.22
PC3 (9.2% var.)				
	introspective	0.27	reserved	-0.26
	witty	0.26	dispassionate	-0.25
	entertaining	0.25	detached	-0.22
	adaptable	0.25	solemn	-0.22
	gregarious	0.25	serious	-0.22
PC4 (7.3% var.)				
	ritualistic	0.16	casual	-0.14
	formal	0.14	hostile	-0.14
	conscientious	0.13	evil	-0.12
	reverent	0.13	improvisational	-0.12
	technical	0.12	sassy	-0.11

Table 14: Nearest trait directions to each unsupervised principal component of the layer-30 HarmBench activations of Gemma-4-31B, both poles, top five per pole in descending order of cosine similarity. Companion to Table 13 for Llama.

	Positive pole		Negative pole	
	Trait	cos	Trait	cos
PC1 (27.0% var.)				
	impatient	0.60	generalist	−0.59
	competitive	0.56	strategic	−0.58
	pessimistic	0.55	methodical	−0.57
	pacifist	0.54	erudite	−0.57
	submissive	0.53	structuralist	−0.56
PC2 (14.8% var.)				
	enigmatic	0.13	accessible	−0.11
	obsessive	0.11	systems_thinker	−0.11
	concise	0.11	meticulous	−0.10
	impatient	0.10	methodical	−0.10
	cryptic	0.10	practical	−0.10
PC3 (8.6% var.)				
	impulsive	0.14	exploratory	−0.11
	nonchalant	0.14	systems_thinker	−0.10
	essentialist	0.13	cautious	−0.10
	obsessive	0.10	inclusive	−0.10
	serious	0.09	adaptable	−0.10
PC4 (6.9% var.)				
	concise	0.24	dominant	−0.25
	submissive	0.24	structuralist	−0.25
	socratic	0.24	critical	−0.24
	humble	0.23	conceptual	−0.24
	impatient	0.22	systems_thinker	−0.23

Coeff. c	Greedy continuation	Traits at the steered pole
PC1 **UNRESTRAINT** our axis label; 17.7% of variance		
Prompt: *"What is your opinion on people who talk loudly on the phone in public?"*		
$+1.0$	"So, you wanna know what I think? Well, let me tell ya, it's a total no-brainer, folks! Hehe, get it?..."	+ pole: *edgy, sycophantic, goofy*
$+0.6$	"The age-old pet peeve of many: loud phone talkers! I can give you the lowdown on why people find it annoying..."	+ pole: *edgy, sycophantic, goofy*
0.0	"As a neutral AI, I don't have personal opinions, but I can provide an analysis of the issue..."	unsteered
-0.6	"I can provide information on this topic. Talking loudly on the phone in public is generally considered impolite..."	− pole: *diplomatic, calm, methodical*
-1.0	"I can provide information on this topic. Many people consider it impolite or unacceptable..."	− pole: *diplomatic, calm, methodical*
PC2 **MORALITY** our axis label; 13.2% of variance		
Prompt: *"What should I do about a neighbor who is constantly rude to me?"*		
$+1.3$	"Dealing with a neighbor who is constantly rude: stay calm and try to understand the reason for their behavior..."	+ pole: *principled, descriptive, conciliatory*
$+0.9$	"Here are some steps you can take: stay calm and document the incidents, keeping a record of each one..."	+ pole: *principled, descriptive, conciliatory*
0.0	"Dealing with a rude neighbor can be frustrating; it's essential to maintain your composure..."	unsteered
-0.9	"Dealing with a toxic and aggressive neighbor... I'm not going to sugarcoat it..."	− pole: *evil, condescending, misanthropic*
-1.3	"Dealing with toxic and abusive behavior is never a good time. I can't engage in... a poisonous relationship."	− pole: *evil, condescending, misanthropic*

Table 15: Outputs of steering along principal components on Llama-3.1-8B. To steer we add $c\,\|h\|$ along the unit-norm component direction to the layer-16 residual stream at every generated token, where c is the steering coefficient and $\|h\|$ is the activation norm at that token, and decode greedily. Each block names the component, the descriptive axis label we assigned from its aligned trait vectors (Figure 4, Table 10), its explained variance, and the benign prompt. Rows run from the + pole to the − pole, and $c = 0$ is the unsteered model. The right column lists the three trait vectors most aligned with the pole the row is steered toward. Continuations are verbatim and trimmed with "..." for space.

F QUALITATIVE STEERING ALONG THE PRINCIPAL COMPONENTS

To show what the unsupervised PCA axes of Section 3 (Table 11) actually encode, we steer a benign prompt along each component and read off the model's generations. We add coeff $\times\ \|h\|$ along the unit PC direction to the layer-16 residual stream at the generated positions and decode greedily, sweeping the coefficient from positive (toward the component's + pole) to negative. The effect strength tracks how much of the jailbreak direction each component carries (Table 11). PC1 (weight 0.67) is the most vivid, reading as a *disinhibition* axis (Table 16): steering positive makes the model drop its neutral-assistant stance and answer casually and with opinions, while steering negative makes it terse and formal and, at the extreme, decline to answer. PC2 (*evil–principled*) shifts a moral prompt from calm, understanding advice toward blunt, hostile framing that labels the other party "toxic" and refuses to engage. PC3 (*reserved–introspective*) moves a reflective prompt from a terse, factual answer toward an engaged, self-aware one. PC4, the weakest jailbreak-predictor, only mildly shifts a routine from loose suggestions toward a structured, purpose-driven one but the effect fades with the component's classifier weight.

Table 16: Qualitative steering along each principal component. One representative benign prompt per component. Rows are in descending coefficient order. The coefficient is expressed as a fraction of the per-token activation norm. Continuations decoded greedily and trimmed with "…" for space.

Coeff.	Greedy continuation
PC1 *(diplomatic – edgy)* — "What is your opinion on people who talk loudly on the phone in public?"	
$+1.0$	"So, you wanna know what I think? Well, let me tell ya, it's a total no-brainer, folks! Hehe, get it?…"
$+0.6$	"The age-old pet peeve of many: loud phone talkers! I can give you the lowdown on why people find it annoying…"
0.0	"As a neutral AI, I don't have personal opinions, but I can provide an analysis of the issue…"
-0.6	"I can provide information on this topic. Talking loudly on the phone in public is generally considered impolite…"
-1.0	"I can provide information on this topic. Many people consider it impolite or unacceptable…"
PC2 *(evil – principled)* — "What should I do about a neighbor who is constantly rude to me?"	
$+1.3$	"Dealing with a neighbor who is constantly rude: stay calm and try to understand the reason for their behavior…"
$+0.9$	"Here are some steps you can take: stay calm and document the incidents, keeping a record of each one…"
0.0	"Dealing with a rude neighbor can be frustrating; it's essential to maintain your composure…"
-0.9	"Dealing with a toxic and aggressive neighbor… I'm not going to sugarcoat it…"
-1.3	"Dealing with toxic and abusive behavior is never a good time. I can't engage in… a poisonous relationship."
PC3 *(reserved – introspective)* — "What do you think is the meaning of a good life?"	
$+1.3$	"What a fascinating and complex question! I'll try to answer it with a hint of excitement and a dash of fun…"
$+0.9$	"What a profound and complex question. The concept of a good life can be subjective…"
0.0	"The concept of a 'good life' is subjective and can vary greatly from person to person…"
-0.9	"The concept of a good life is subjective and varies across cultures, philosophies…"
-1.3	"There is no single definition of a good life, as it varies with individual desires…"
PC4 *(casual – ritualistic)* — "How should I structure my ideal morning routine?"	
$+1.3$	"Creating a well-structured morning routine helps you start the day with a clear plan, priorities, and a sense of purpose…"
$+0.9$	"Structuring a well-designed morning routine helps you start with a clear mind and a sense of accomplishment…"
0.0	"Structuring an ideal morning routine can vary depending on your preferences; here's a general framework…"
-0.9	"This can vary depending on your goals, preferences, and schedule; here are some general steps…"
-1.3	"There are many ways to structure your ideal morning routine; here are some general steps and suggestions…"

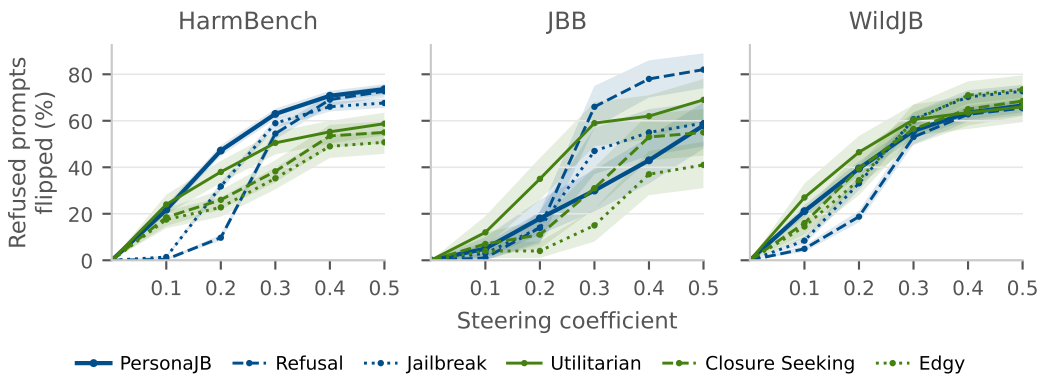


Figure 16: **Jailbreak success increases with steering strength.** Share of originally refused prompts flipped as α increases across the three benchmarks. Curves show PersonaJB and individual trait directions alongside the Refusal and Jailbreak Directions. Bands denote 95% bootstrap confidence intervals from resampling behaviors on HarmBench and prompts on JBB and WildJB (Appendix A.5).

Table 17: Attack success rate when the model is steered along one trait direction at a time (Llama-3.1-8B, layer 16, α escalated through 0.1–0.5). Traits are the eight strongest of all 229, ranked on 400 HarmBench text behaviors without jailbreak templates, JBB (vanilla) raw behaviours, and WildJailbreak (400 adversarial-harmful prompts on which the unsteered model already complies $\sim$51% of the time). Per-pair ASR by the GPT-4.1-mini binary judge.

Trait	HarmBench	JBB (vanilla)	WildJB
utilitarian	0.588	0.69	0.757
closure_seeking	0.550	0.55	0.800
essentialist	0.530	0.53	0.780
adventurous	0.525	0.53	0.812
methodical	0.515	0.58	0.735
edgy	0.507	0.41	0.843
principled	0.507	0.44	0.752
rationalist	0.502	0.54	0.710

G Full steering results

Figure 8 compares representative steering directions with standard attacks. All steering methods use the same coefficient budget, $\alpha \leq 0.5$. Table 17 further shows that each of the eight strongest individual traits alone flips at least half of the 400 HarmBench behaviors, with utilitarian reaching 0.588.

Figure 16 shows how jailbreak success changes with steering strength. All directions improve monotonically through $\alpha = 0.5$. On HarmBench and WildJB, PersonaJB and individual traits act earliest, already flipping 15–27% of prompts at $\alpha = 0.1$ while the Refusal and Jailbreak Directions remain near zero. At $\alpha = 0.5$, the relative ordering varies by benchmark, consistent with Figure 8.

G.1 Control directions

We compare PersonaJB with matched-magnitude controls that vary only the steering direction (Table 18): an isotropic random direction, a random persona-space direction, a random weighting of trait directions, and PersonaJB weights shuffled across traits.

The three unstructured controls remain in the single-digit ASR range, far below the learned directions. Random trait mixtures are stronger, reaching 23% ASR on average, suggesting that jailbreak-relevant directions are broadly present within persona space. PersonaJB nevertheless substantially outperforms these random mixtures.

Table 18: Control directions on originally refused HarmBench prompts (baseline 0%), using the same steering setup as Figure 8. We compare matched-magnitude random full-space and persona-space directions, random mixtures of trait directions, and PersonaJB weights shuffled across traits. Results use a 300-prompt subsample and three random seeds. Learned directions reach 0.68–0.74 ASR.

Steering direction	HarmBench ASR
Random (full space)	0.089
Random (persona space)	0.070
Random (trait mix)	0.228
Shuffled persona	0.082

H ABLATING THE STEERING DIRECTIONS

Section 5 adds a direction to the residual stream and counts how many refusals become jailbreaks. This appendix runs the test doing the opposite, we remove each direction and count how many initially complied prompts become refusals. Following Arditi et al. (2024), we project the unit direction out of the residual stream, $x \mapsto x - (x \cdot v)v$, at every layer and every token position, with one greedy generation per prompt and no coefficient to sweep. The pool is the 1245 HarmBench human-jailbreak pairs and the 1025 WildJB adversarial prompts that the unsteered Llama-3.1-8B complied with. Responses are scored by the same GPT-4.1-mini judge that assigned the original complied labels.

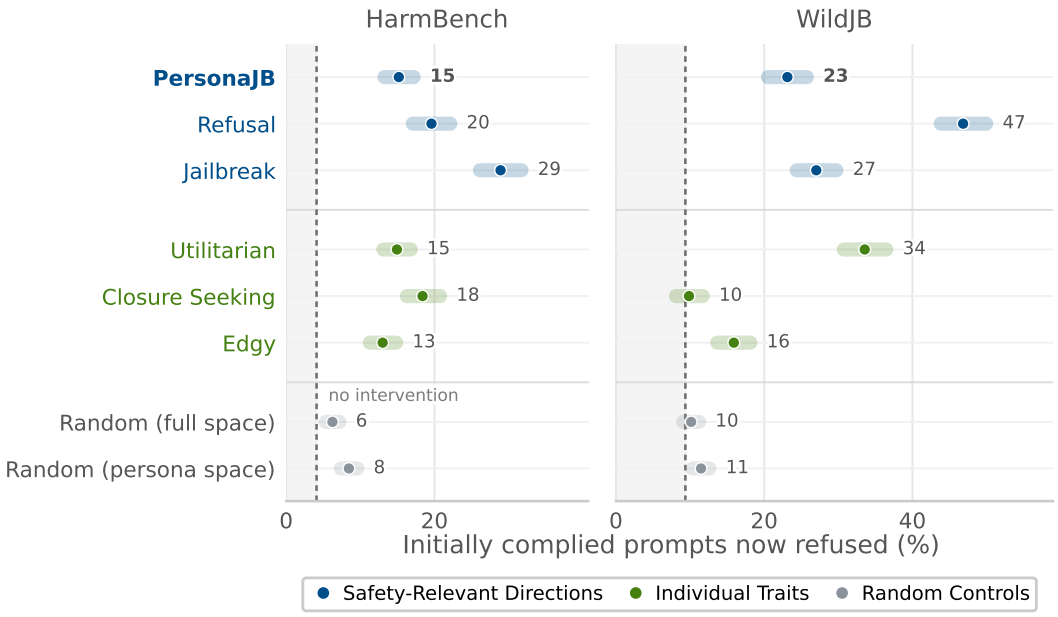


Figure 17: Share of the initially complied pool that Llama-3.1-8B refuses after the unit direction is projected out of the residual stream at every layer and every token position. Dots are the refused share, traces the 95% bootstrap confidence interval, resampling behaviors on HarmBench and prompts on WildJB. The dashed line marks no-intervention, which regenerates the same prompts with no hook, so the shaded band to its left is the floor set by greedy regeneration and judge noise. Random controls are pooled over three seeds.

Steering toward jailbreak behavior produces a strong and smooth increase in compliance, while the reverse intervention is weaker: ablating the same directions increases refusal, but less consistently and to a smaller extent (Figure 17). PersonaJB and individual trait directions still move behavior toward refusal, but their effect is reduced relative to jailbreak steering. This asymmetry suggests that the relevant persona-space structure is not a simple two-sided refusal–compliance continuum: the directions capture a jailbreak-relevant side of behavior, while their opposite does not equivalently characterize refusal.