Shared Truth: Emergent Truth Properties in Large Language Models via Heterogeneous Injection-based Transfer

Isaiah Freeman*
Department of Computer Science
Virginia State University
Petersburg, VA 23806
ifre3976@students.vsu.edu

Joed Ngangmeni
Department of Computer Science
Virginia State University
Petersburg, VA 23806
jngangmeni@vsu.edu

Abstract

Representation Engineering (RepE) aims to expose and steer high-level concept directions in language model activations, and recent work shows that hidden states can be transferred across heterogeneous architectures via trained neural adapters. Whether specific concepts survive such transfer remains an open question. We study one such concept, *truth*, by mapping source-model activations through autoencoder adapters into target-model representation space, varying injection strength $\alpha \in [0, 1]$, and reading the result with the target's native truth probe across 21 cross-family trajectories from four model families, with comparison against a closed-form rigid alignment baseline on all 21 pairs. We find that ordinal truth ranking is preserved through partial injection (median AUROC near the native ceiling through $\alpha \in [0, 0.7]$) but collapses sharply at full injection (α=1), with parametric separability dropping $\sim 14\times$ while rank order partially survives. The apparent accuracy overshoot under partial injection decomposes cleanly into two effects: a large recalibration component dominated by native probe miscalibration ($R^2 = 0.75$), and a smaller residual separability gain (median $+1.13$ pp) whose 95% bootstrap CI lower bound exceeds zero on 17 of 21 trajectories; random maps of the adapter's architecture also produce such gains, and trained adapters exceed them on two of three tested trajectories on the full test set (one of three on simple statements). These findings characterize cross-family truth structure as partially compatible and ordinally stable under partial injection — preserved enough that learned and rigid alignment methods both surface meaningful truth signal. The two methods differ at full substitution: trained adapters fail hard, while the closed-form rigid map retains most rank separability (median AUROC 0.91 vs. 0.72, higher on all 21 pairs), so the collapse depends on the alignment method rather than being a fixed property of cross-family substitution. Practitioners should treat full substitution through learned adapters as a deployment boundary. Notebooks and experiments are released at https://github.com/kingtechnician/shared-truth.

1 Introduction

RepE [Zou et al., 2025a] has emerged as a key tool for Large Language Model (LLM) interpretability. Building on the Linear Representation Hypothesis [Park et al., 2024], RepE steers activations within hidden layers to manipulate or detect emergent behaviors — including spatial and temporal representations [Gurnee and Tegmark, 2024], and truthfulness [Marks and Tegmark, 2024]. Within

*Personal website: https://kingtechnician.io

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

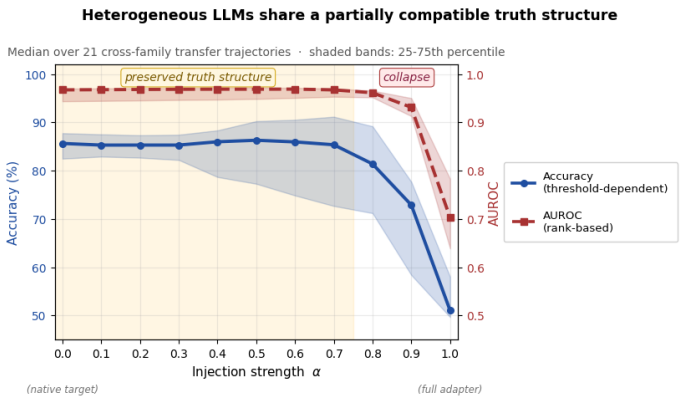


Figure 1: Overall truth accuracy and AUROC across all 21 trajectories. Demonstrates consistent truth ranking and accuracy, and general collapse as approaching $\alpha = 1$.

the latent space of an LLM sits an encoded understanding of true and false statements regardless of final output, and classifiers can map these latent representations into a truth subspace to predict whether a model's activations are computing a falsehood [Marks and Tegmark, 2024, Bürger et al., 2024].

RepE has also motivated latent-space communication between *agents*: rather than exchanging text, LLMs can pass hidden activations directly through a chain, with only the final model decoding to tokens [Zou et al., 2025b]. Among *homogeneous* models this preserves conceptual coherence with lower computational overhead [Zou et al., 2025b]. In practice, however, different model families carry different strengths [Lee et al., 2024], motivating *heterogeneous* transfer; passing hidden states between models of different training origins via trained adapter networks (autoencoders, attention-based models) [Du et al., 2026, Oozeer et al., 2025] or augmented KV caches [Fu et al., 2026].

However, an open question lies within heterogeneous models as to whether or not *every* concept is transferred between hidden states. For example, it seems conclusive that the *general* thoughts of a source language model can be passed to a target language model such that its output is coherent, but it is not clear if *truth* remains linearly separable after transfer.

We address this by reading transferred source activations with the target's own truth probe, sweeping injection strength to trace how truth-relevant structure degrades.

From these experiments, we demonstrate the following findings:

- **Ordinal ranking is preserved through partial injection**: Across all 21 trajectories, the target probe's ranking of true over false statements survives near-ceiling through the bulk of the injection range.
- **Apparent accuracy gains decompose into two effects**: A large recalibration component dominated by native probe miscalibration, and a smaller separability gain — bootstrap-confirmed positive on 17 of 21 trajectories — that recalibration alone cannot account for. Random maps of the adapter's architecture also produce such gains; trained adapters exceed them on two of three tested trajectories on the full test set and one on simple statements (Appendix A.14).
- **Asymmetric collapse of learned adapters at full injection**: At α=1, parametric separability under the trained adapter collapses sharply ($\sim$14$\times$) while rank-based separability degrades more gradually. Ordinal ranking partially survives full injection; parametric separability does not.
- **The collapse depends on the alignment method**: A closed-form rigid map with no learned non-linearity retains most of the target's rank separability at α=1 (median AUROC 0.91 against 0.72 for the trained adapter, higher on all 21 trajectories), so full-substitution failure is largely a property of the learned alignment rather than of cross-family substitution as such.

2 Related Works

2.1 Truth subspaces in LLMs

High-level concepts are linearly separable within LLM activation spaces [Park et al., 2024]. Marks and Tegmark [2024] showed that a logistic probe on residual-stream activations at a calibrated layer separates true from false statements along a one-dimensional "truth direction." Bürger et al. [2024] extend this to a two-dimensional subspace decomposing truth from *statement polarity*, yielding probes that generalize across negation (e.g., "Paris is not in France") and, as they report, to other statement forms that confound a single truth direction. In our setting the probes separate conjunctions well but disjunctions only weakly (Appendix A.15). We adopt their Training of Truth and Polarity Direction (TTPD) probe (Section 3.4).

2.2 Latent communication and heterogeneous transfer

When language models exchange hidden activations rather than text, concepts pass through the chain with lower overhead and higher fidelity than natural-language interfaces support. LatentMAS [Zou et al., 2025b] formalizes this for *homogeneous* chains, with only the final model decoding to tokens. Oozeer et al. [2025] extend it to heterogeneous transfer, training autoencoder adapters that preserve steering vectors — including "safety switches" against backdoors and harmful prompts — across architectures. Alternative formulations augment the target's KV cache via a learned mapping [Fu et al., 2026, Du et al., 2026]. We adopt the autoencoder formulation [Oozeer et al., 2025] (Section 3.5).

2.3 Cross-model representational structure

Cross-model transfer presupposes some structure shared between models. The Platonic Representation Hypothesis (PRH) [Huh et al., 2024] argues that representations are converging, as models scale, toward a shared statistical model of reality, with evidence drawn mainly from mutual nearest-neighbor alignment of the kernels that models induce; *vec2vec* [Jha et al., 2026] translates between text-embedding spaces without paired data. Gröger et al. [2026] show that once similarity metrics are calibrated against width and depth confounders, the reported convergence in global metrics largely disappears while local neighborhood agreement remains, and propose the Aristotelian Representation Hypothesis (ARH): convergence toward shared local neighborhood relationships. The two are not exclusive alternatives, since global alignment implies neighborhood alignment.

We do not adjudicate between the two. Both concern similarity across a model's whole representation, measured over many inputs; our question is narrower: whether one concept, truth, remains readable by the target model's own probe after transfer through a particular map. Shared structure in either sense would not by itself guarantee this, since a map can preserve global or neighborhood similarity while placing a single concept away from where a fixed probe reads it.

3 Experiment Design

We evaluate whether truth-relevant structure transfers across heterogeneous LLMs by combining two established techniques: linear truth probes [Marks and Tegmark, 2024, Bürger et al., 2024] applied to the target model, and autoencoder-based activation transfer [Oozeer et al., 2025] between source and target. Full training procedures and hyperparameters appear in Appendix A.

3.1 Pipeline and metrics

For each (source, target) pair, we extract source activations $h_S \in \mathbb{R}^{d_S}$ at a calibrated layer, map them through a trained autoencoder $f_\theta : \mathbb{R}^{d_S} \to \mathbb{R}^{d_T}$, and linearly interpolate with the target's native activations h_T:

$$\tilde{h}_\alpha = (1-\alpha)\, h_T + \alpha\, f_\theta(h_S), \qquad \alpha \in [0,1]. \tag{1}$$

The interpolated activation is read by a truth probe p_T trained on the target's native activations. Sweeping $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$ traces a transfer trajectory, with α=0 recovering the target's native ceiling and α=1 representing pure adapter output. Figure 2 diagrams the full pipeline.

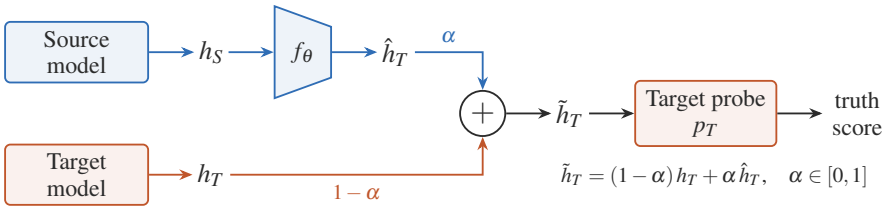


Figure 2: Activation injection pipeline for the Shared Truth experiment. Source activations h_S are mapped through an autoencoder f_θ to predicted target activations $\hat{h}_T$, then linearly mixed with the native target activations h_T via $\tilde{h}_T = (1-\alpha)h_T + \alpha\hat{h}_T$ before being read by the target probe p_T.

We report two metrics on a held-out true/false statement set: classification **accuracy** (threshold-dependent) and **AUROC** (rank-based separability). Reporting both is necessary because the two metrics decouple across most of the sweep, which is itself one of the central findings of this work (Section 4.3).

Following established practice in truth-probe evaluation [Bürger et al., 2024], we restrict the held-out evaluation set to simple statements in both polarities (1080 affirmed and 519 negated), excluding multilingual and conjunction/disjunction forms, matching the statement forms the probes were trained on. Appendix A.15 repeats the sweep by statement type on three trajectories: the plateau holds for every type, but native separability and the loss at α=1 differ by type.

3.2 Models

We evaluate 22 α-sweeps (21 unique trajectories) over 13 source/target model pairs, drawn from seven instruct-tuned models spanning four families: LLaMA 3 8B, LLaMA 3.2 3B, Mistral 7B v0.3, Qwen 2.5 1.5B, Qwen 2.5 7B, Gemma 7B, and Gemma 2 2B [Grattafiori et al., 2024, Jiang et al., 2023, Qwen Team, 2024, Gemma Team, 2024a,b]. The pair pool is unevenly distributed across target families (Llama n=4, Gemma n=9, Qwen n=3, Mistral n=5); Table 2 (Appendix A.6) lists every trajectory.

3.3 Layer calibration

The pipeline involves two layer choices per pair. The *truth layer* is selected per model by maximizing the Fisher-trace ratio of between- to within-class spread [Bürger et al., 2024] on a polarity-balanced calibration pool. The *transfer layer* pair is selected by maximizing Representational Similarity Analysis (RSA) [Oozeer et al., 2025] over source/target pairs drawn from the four central layers of each model. In our main configuration the autoencoder maps onto the target's truth layer; Section 4.6 ablates against the RSA-selected target layer. Full procedures in Appendix A.1.

3.4 Truth probes

For each model we train two probes on its calibrated truth layer following Bürger et al. [2024]: (i) TTPD, a two-dimensional probe that recovers a general truth direction t_G and a polarity direction t_P, then fits a logistic classifier on the (t_G, t_P)-projected activations; and (ii) a logistic-regression baseline fit directly on the d-dimensional residual stream. Training data spans six topic-specific contrastive datasets (`cities`, `sp_en_trans`, `inventors`, `animal_class`, `element_symb`, `facts`) in both affirmative and negated forms, with stratified subsampling balancing the four (polarity, label) cells within each topic. Full procedure in Appendix A.3.

3.5 Autoencoder adapter

For each source–target pair, we train a two-layer MLP adapter to map source activations at the calibrated source layer to predicted target activations at the calibrated target layer. Training data combines a slice of Anthropic RLHF [Bai et al., 2022] with domain-specific examples to maintain broad representational coverage [Oozeer et al., 2025]. We select the best-performing adapter per

pair by Fraction of variance unexplained (FVU) on a held-out split. Hyperparameters, vocabulary alignment across tokenizers, attention masking, and optimizer settings are in Appendix A.4.

4 Results

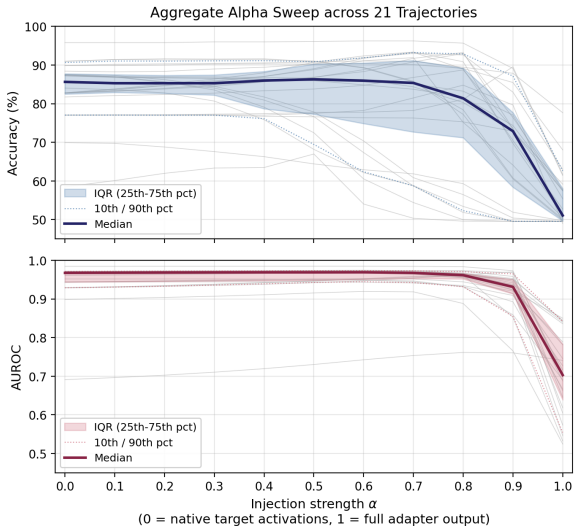


Figure 3: Aggregate truth accuracy and AUROC across all 21 distinct trajectories. IQR represents 25th and 75th percentile ranges. Median is the middle trend. Semi-transparent lines represent individual runs, making clear outliers to this general behavior.

4.1 Experimental coverage and metrics

Aggregate results are summarized using the median and interquartile range over the 21 distinct trajectories (Figure 3); per-target family breakdowns appear in Figure 7.

Full per-pair α-sweep results appear in Table 2 in Appendix A.6.

4.2 Universal collapse of learned adapters under full injection

At $\alpha{=}1$, where the target model's activations are fully replaced by the adapter's output, threshold-based accuracy collapses universally and rank-based separability degrades sharply. Across the 21 trajectories, median accuracy under $\tau_0{=}0.5$ falls to 51% — chance on a balanced binary task — and the spread narrows around this collapse rather than masking it. Median AUROC falls from its near-ceiling sweep value of $\sim$0.97 to 0.72 at $\alpha{=}1$ (IQR $[0.65, 0.79]$, range $[0.53, 0.85]$), a sharp degradation but not a total one: 12 of 21 trajectories retain AUROC above 0.7, and none retain AUROC above 0.9. The asymmetry between metrics is mechanistic. We measure parametric separability with the Fisher discriminant ratio of the target probe's output scores, $J = (\mu_T - \mu_F)^2/(\sigma_T^2 + \sigma_F^2)$, computed per trajectory at each α. Its median falls from $J{=}3.47$ at $\alpha{=}0$ to $J{=}0.30$ at $\alpha{=}1$, and the per-trajectory collapse ratio $J(\alpha{=}0)/J(\alpha{=}1)$ has median $14.5\times$ (IQR $[6.5, 35.9]$). We summarize the collapse with this median of paired, per-trajectory ratios rather than the ratio of the two medians ($11.6\times$) for two reasons: each trajectory is measured against its own native separability, which varies widely across target probes (native AUROC 0.69–0.98), whereas the ratio of medians can compare different trajectories at the two endpoints; and the median is robust to the long right tail of ratios that arises as $J(\alpha{=}1)$ approaches zero. AUROC degrades far less because some asymmetry in the score distribution tails survives even as their centroids merge. Ordinal ranking partially survives full injection; parametric separability does not.[2]

[2]The Llama 3.2 3B $\rightarrow$ Qwen 2.5 1.5B 12$\rightarrow$12 configuration was swept under both variant labels with the same adapter and probes (rows 7 and 8 of Table 2); the rows are identical, and aggregate statistics dedup to 21 unique trajectories.

The pattern is invariant to both source and target family. Disaggregated by target and by source (Appendix A.7), all eight family-level cohorts exhibit the same qualitative AUROC trajectory: a near-flat plateau through $\alpha \in [0, 0.7]$, an inflection in the 0.8–0.9 range, and collapse at α=1. No family on either end of the transfer avoids the collapse, and no family collapses meaningfully earlier, despite substantial variation in family sample size (n=2 to n=10), architecture, depth, and parameter count. Source-side variation modulates accuracy but not AUROC: Gemma and Mistral sources show more gradual accuracy decline than Llama and Qwen, though both appear in only n=2 trajectories and this may reflect sampling rather than architecture.

The collapse is, however, largely specific to the learned adapters: a closed-form rigid map retains most of the target's rank separability at α=1 (Section 4.5).

4.3 Ordinal ranking is preserved through most of the sweep

The aggregate trajectories in Figure 3 show a striking asymmetry between our two metrics. Median AUROC remains near 0.97 — essentially the target probe's native ceiling — throughout the range $\alpha \in [0, \approx 0.7]$, with no inflection until α=0.8. Median accuracy over the same range hovers within roughly one percentage point of its α=0 value. On the surface, both metrics look preserved. The decoupling appears in the *spread* across the 21 trajectories rather than in their central tendency.

Figure 4 makes this decoupling explicit across all 21 trajectories, with three panels covering both threshold-based metrics (Panel A) and rank-based separability (Panel B), plus the per-α inter-trajectory IQR for all three metrics (Panel C). Two patterns emerge in Panel C. First, the IQR of accuracy under τ_0=0.5 widens substantially through $\alpha \in [0.4, 0.9]$, peaking near α=0.9: trajectories diverge from each other under partial injection because the size of the recalibration effect varies with per-pair native miscalibration. Second, the IQR of accuracy under per-α optimal $\tau^*(\alpha)$ stays much narrower across the same range, as does the IQR of AUROC. The three metrics form a hierarchy of pair-specificity: the τ_0 overshoot is highly pair-specific (recalibration depends on miscalibration), the τ^* overshoot is less so (the residual separability gain is smaller and more uniform), and rank order is most pair-invariant (preserved similarly across all 21 pairs until the α=1 collapse).

Under cross-family injection, the load-bearing invariant for truth-probe performance is rank rather than absolute position. The probe is trained at α=0 with thresholds calibrated to native target activations; interpolated adapter output shifts those scores by varying amounts across trajectories (the widening accuracy spread), but preserves the relative ordering of true-statement scores above false-statement scores across all 21 trajectories. Rank preservation under partial injection does not by itself show that truth structure transferred. The TTPD probe's logit is affine in the activations, so under Equation 1 it interpolates linearly between the score of the native activation and the score of the adapter output. If the adapter's scores were constant across examples, ranking would be unchanged for every $\alpha < 1$ and would collapse only at α=1. They are not: the probe still ranks adapter output alone at median AUROC 0.72 (Section 4.2), while shuffled outputs fall to chance (0.49–0.51; Appendix A.14).

4.4 Partial injection produces implicit recalibration and a smaller separability gain

In 12 of 21 trajectories, partial injection at intermediate $\alpha^*_{\tau_0}$ raises target-probe accuracy more than 0.5 pp above the native α=0 value under default threshold τ_0=0.5 (bolded in Table 2). The threshold-calibration diagnostic (Appendix A.8) shows most of this gain arises from implicit recalibration of the probe's miscalibrated decision boundary, with a smaller residual effect persisting once recalibration is closed.

Recalibration accounts for the bulk of the τ_0 overshoot. The TTPD probe's logistic regressor is trained on α=0 activations but applies a fixed τ_0=0.5 at deployment, often suboptimal for the held-out test distribution: optimal thresholds vary substantially across our 15 probe configurations (Table 3). Interpolating source activations shifts the score distribution such that τ_0 lands closer to optimal, recovering accuracy that native miscalibration had suppressed. Since the adapter is trained on generic Reinforcement Learning from Human Feedback (RLHF) reconstruction with no truth supervision, this recalibration is a side effect of the geometric shift. A pair-level OLS regression of $\Delta\mathrm{Acc}(\tau_0)$ on native ceiling gap, adapter FVU, adapter–target cosine similarity, and native AUROC

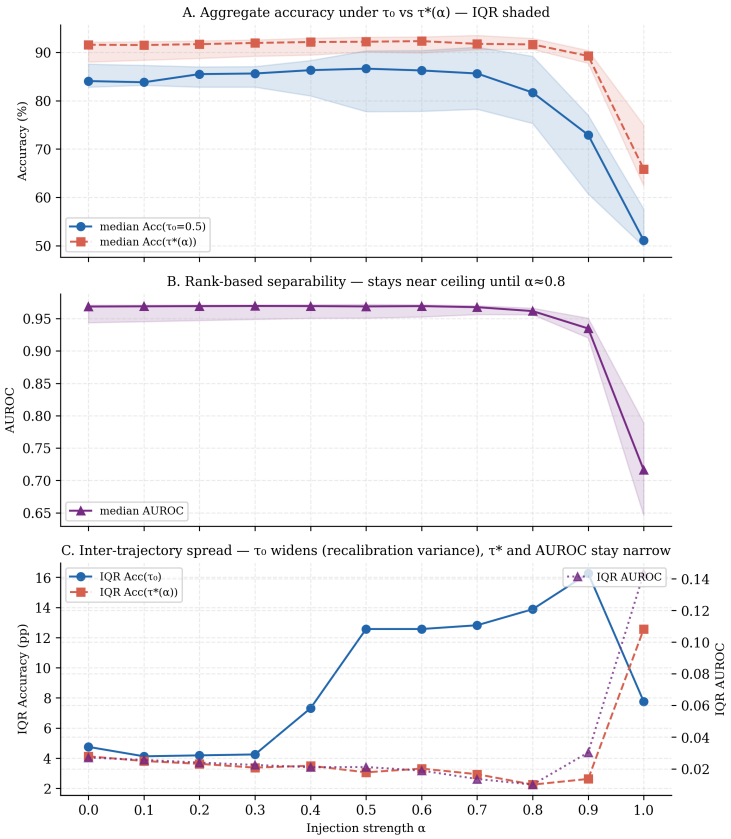


Figure 4: Decoupling of threshold-dependent and rank-based metrics across 21 trajectories. Panel A: median accuracy under default threshold τ_0=0.5 (blue) vs. per-α optimal threshold $\tau^*(\alpha)$ (red, dashed), with IQR shaded. Panel B: median AUROC. Panel C: inter-trajectory IQR for each metric — τ_0 accuracy spread widens through the body of the sweep (recalibration variance), while τ^* accuracy and AUROC stay narrow until the α=1 collapse.

(Table 4) yields $R^2 = 0.748$ ($p \approx 1.1 \times 10^{-4}$), with ceiling gap the dominant predictor ($\beta = +1.07$, $p < 10^{-4}$): roughly one pp of τ_0 overshoot per pp of native miscalibration.

A smaller residual effect persists under per-α optimal thresholds. To isolate separability gain from recalibration, we sweep the optimal threshold $\tau^*(\alpha)$ at each α and compute the strict overshoot $\Delta\mathrm{Acc}(\tau^*) = \max_\alpha \mathrm{Acc}(\tau^*(\alpha), \alpha) - \mathrm{Acc}(\tau^*(0), 0)$; recomputing $\tau^*(\alpha)$ per α closes the recalibration channel by construction. A paired stratified bootstrap (B=10,000) over the held-out test set yields 95% CI lower bounds exceeding zero on 17 of 21 trajectories (median +1.13 pp, max +9.01 pp; Table 7, Appendix A.12; per-pair peaks in Table 5). The four exceptions all have point estimates $\leq$ 0.50 pp, at the test-set resolution floor. The bootstrap captures test-set sampling but not adapter-training variance, which we measure on three trajectories (Appendix A.13), and it tests against zero rather than against untrained maps: random maps of the adapter's architecture also produce positive strict overshoot, and trained adapters exceed them on two of three tested trajectories on the full test set and one of three on simple statements (Appendix A.14).

Predictors of the strict overshoot. A second OLS regression of $\Delta\mathrm{Acc}(\tau^*)$ on the same features yields $R^2 = 0.578$ ($p < 10^{-2}$, Table 4); no individual predictor reaches significance at n=21. Adapter–target cosine similarity is borderline ($\beta = +0.12$, $p = 0.06$, alongside ceiling gap at $p = 0.06$) but non-significant in the τ_0 regression ($p = 0.49$), suggesting it tracks the separability gain rather than recalibration; we treat this as descriptive given the sample size.

Table 1: Trained adapter vs. orthogonal Procrustes over the 21 unique trajectories. Δ: τ_0 accuracy minus native ceiling at the peak and at full injection. AUROC: the target's native probe at α=1. Spearman ρ between the methods' per-pair outcomes: $+0.36$ (p=0.10) at peak, $+0.43$ (p=0.05) for α=1 accuracy, -0.03 (p=0.89) for α=1 AUROC. Per-pair values in Appendix A.11.

	Δ_{peak} (pp)		$\Delta_{\alpha=1}$ (pp)		$\text{AUROC}_{\alpha=1}$	
Method	median	IQR	median	IQR	median	IQR
Trained adapter	$+0.63$	$[+0.00, +9.13]$	-30.64	$[-32.40, -25.20]$	0.72	$[0.65, 0.79]$
Procrustes	$+4.57$	$[+0.19, +11.19]$	-10.76	$[-23.45, -3.06]$	0.91	$[0.86, 0.92]$

Procrustes vs. trained (accuracy: 0.5 pp tie threshold):
Peak accuracy: 13 Procrustes wins, 5 ties, 3 trained wins
α=1 accuracy: 20 Procrustes wins, 0 ties, 1 trained win
α=1 AUROC: Procrustes higher on 21 of 21

Synthesis. This decomposition is testable: a baseline capturing recalibration but not learned non-linearity should match the trained adapter on most pairs but not where learned-alignment contribution is large. Section 4.5 runs this test.

4.5 Closed-form rigid alignment as a falsification test

We fit orthogonal Procrustes with isotropic scale on the same RLHF reconstruction corpus used for the autoencoder adapter,[3] substitute it for the trained adapter, and re-run the same α-sweep. The map has no learned non-linearity and no truth supervision, and it is fit in closed form rather than by gradient training.

Procrustes matches or exceeds the trained adapter on the majority of pairs. Procrustes peaks higher than the trained adapter on 13 of 21 trajectories, ties within 0.5 pp on 5, and falls short on 3 (Table 1; full per-pair sweeps in Appendix A.11). The median peak overshoot under τ_0 is $+0.63$ pp for the trained adapter and $+4.57$ pp for Procrustes. Procrustes thus exceeds the recalibration component the trained adapter recovers by several percentage points.

The trained adapter wins on a specific pair geometry. Two of the three trajectories where the trained adapter strictly exceeds Procrustes are Llama 3.2 3B (L12) $\rightarrow$ Mistral 7B configurations (L13 and L14 target), where the trained adapter peaks at $+10.57$ and $+10.19$ pp above ceiling while Procrustes never exceeds ceiling. The same architectural pairing at other size combinations (Llama 3 8B and Qwen 2.5 7B $\rightarrow$ Mistral 7B) shows Procrustes recovering the gain. The third, Mistral 7B (L13) $\rightarrow$ Gemma 7B (L19), sits just past the tie threshold ($+0.63$ vs. $+0.00$ pp). We report the Llama 3.2 3B $\rightarrow$ Mistral 7B geometry as a target for future investigation.

At full injection, Procrustes keeps the separability the trained adapter loses. Procrustes degrades less at α=1 on 20 of 21 pairs: the trained adapter's α=1 accuracy falls to a median of 30.6 pp below ceiling, Procrustes to 10.8 pp below. The difference is larger in rank-based terms (Figure 5): median AUROC at α=1 is 0.91 for Procrustes against 0.72 for the trained adapter, higher on all 21 pairs and no more than 0.05 below native on 12. The two methods' pair-level outcomes are at most weakly related (Spearman $\rho = +0.43$, p=0.05 for α=1 accuracy; $\rho = -0.03$, p=0.89 for α=1 AUROC). Beyond the overshoot test posed in Section 4.4, the baseline therefore bears on the α=1 collapse: a map with no learned non-linearity and no truth supervision preserves the target's rank separability far better than the trained adapter, so the collapse is not a fixed property of cross-family substitution (Section 4.2). This is evidence of concept-specific compatibility under a globally fitted linear map. It is consistent with PRH but does not distinguish it from ARH (Section 2).

4.6 Layer-selection ablation

A layer-selection ablation re-running five trajectories under the transfer-layer variant (vs. our main truth-layer configuration) finds the qualitative patterns of Sections 4.2–4.4 hold under both variants.

[3] $f(s) = c\,(s - \mu_s)U_s R + \mu_t$ where U_s is the source PCA basis, $R = UV^\top$ from $\text{SVD}(S_c U_s)^\top T_c$, and $c = \text{tr}(R^\top S_c^\top T_c)/\|S_c U_s\|_F^2$.

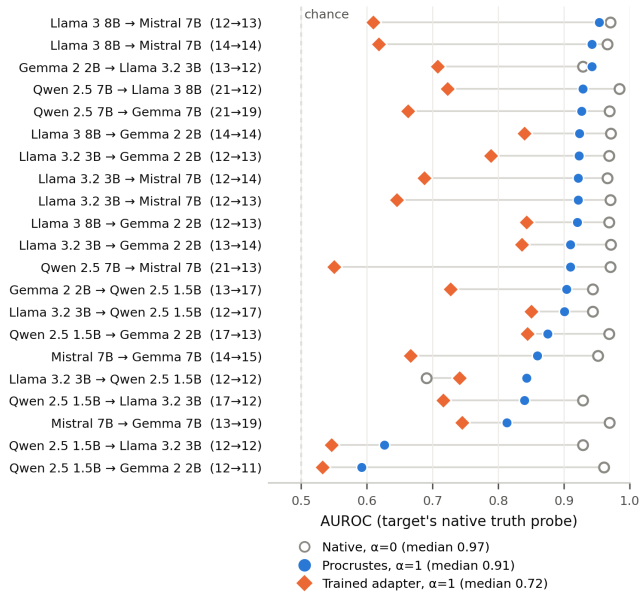


Figure 5: AUROC of the target's native truth probe at full injection, per trajectory. Open circles: native activations (α=0). Circles: Procrustes at α=1. Diamonds: trained adapter at α=1. Sorted by the Procrustes value. Procrustes retains most of the native separability on all but the two Qwen 2.5 1.5B (L12)-source trajectories, and is above the trained adapter on all 21.

Per-trajectory peak accuracy mostly tracks native ceiling in both. The comparison is underpowered where variants select nearby layers (3 of 5 pairs); full per-pair numbers and methodology in Appendix A.5.

5 Limitations and Discussion

Limitations. The 21 trajectories each use a single adapter training run. Multi-seed replication (three additional seeds on three trajectories spanning the overshoot range; Appendix A.13) reproduces the plateau-then-collapse shape in every run on simple statements (three on since-retrained checkpoints) and in 9 of 12 on the full test set, where the three retrained seeds of the lowest-overshoot pair begin to decline by α=0.6–0.7. Per-pair magnitudes carry training variance (strict overshoot $+5.3$ to $+9.0$ pp across seeds on the highest-overshoot pair), and the regression coefficients in Table 4 remain single-seed estimates. The strict-overshoot bootstrap (Appendix A.12) tests against zero; it captures test-set sampling but not training variance, and it is not a test against untrained maps. A seeded null of 1000 random maps of the adapter's architecture (Appendix A.14) shows random maps also produce positive strict overshoot in 46–68% of draws; the original trained adapter exceeds this null on two of three tested trajectories on the full test set, so the separability gain is partly a consequence of injecting source information at all. That null covers one baseline family on three trajectories; random linear maps and the other 18 trajectories are untested. The Procrustes comparison uses default-threshold accuracy and AUROC; strict overshoot under per-α optimal thresholds was not computed for Procrustes. The main sweep evaluates simple statements of both polarities; the statement-type decomposition (Appendix A.15) covers three trajectories only, and the probes separate disjunctions only weakly even on native activations. The trajectory pool is Gemma-heavy (n=9 of 21 target-side), so aggregate quantitative summaries are weighted accordingly. All models fall in the 1.5B–8B range. Our experiments evaluate truth-relevant structure specifically; whether the partial-injection plateau, dual-mechanism overshoot, and α=1 collapse generalize to other linearly-probed concepts (e.g., spatial or temporal representations) remains untested.

Why the learned adapter fails at full injection. The adapter's FVU reconstruction objective is intentional: training without truth supervision tests whether truth structure transfers under a generic alignment objective rather than one engineered to preserve it. A natural reading of the α=1 collapse

attributes it to this choice — FVU reconstruction is dominated by high-variance directions, so low-variance truth-relevant directions need not land where the target's probe reads them. The Procrustes baseline (Section 4.5) is consistent with this reading, though we have not isolated the objective as the cause. Orthogonal alignment with isotropic scale applies the same scalar to every direction and cannot preferentially distort low-variance components, and it retains most of the target's rank separability at $\alpha{=}1$ where the trained adapter loses it. On three trajectories, the residual-geometry analysis (Appendix A.16) locates the failure: the class-mean gap along the target's native truth direction t_G is nearly absent from trained-adapter output (gap ratio 0.02–0.11 of native), yet a probe retrained on the same output recovers AUROC 0.84–0.95. Truth information survives the learned map; its alignment with the target's native truth direction does not. Decodability alone is a weak criterion — random maps also re-probe at 0.97–0.98 — so the informative quantity is alignment, which the rigid map largely preserves and the learned map does not. Pair structure still matters in some cases: rigid alignment also falls to AUROC near 0.6 on the two trajectories with a Qwen 2.5 1.5B (L12) source. Whether a truth-aware adapter objective would close the gap is left to future work.

Discussion. Heterogeneous activation adapters perform implicit threshold recalibration, with practical implications for multi-agent LLM systems routing information through latent-space transfer. Prior work shows adapter-based transfer preserves general semantic content [Oozeer et al., 2025, Du et al., 2026]; we extend this to the specific structure of truth representations. Two implications follow from the recalibration mechanism (Section 4.4): reported accuracy gains in heterogeneous latent-transfer pipelines should be cross-referenced against optimal-threshold ceilings on native activations before being interpreted as richer representations; and partial injection may be a useful recalibration mechanism where truth-probe thresholds are difficult to tune directly.

The collapse of learned adapters at $\alpha{=}1$ is the more consequential finding for system design. Deployments that read a truth probe through a learned adapter trained on generic data should treat full substitution as a hard failure mode and operate in the plateau region, where AUROC remains near-ceiling. Across adapter seeds (Appendix A.13), AUROC stayed within 0.01 of native, or above it, through $\alpha{=}0.7$ in all 12 simple-statement runs (seeds within 0.007) and in 9 of 12 on the full test set. We therefore read $\alpha \leq 0.7$ as a practical operating bound, lower for weakly aligned pairs. Where full substitution is unavoidable, a rigid alignment map preserves rank order far better than our learned adapters, though its default-threshold accuracy still degrades (median 10.8 pp below ceiling), so thresholds should be recalibrated after substitution. Finally, the collapse is not uniform across statement types (Appendix A.15). On two of three tested trajectories, negated statements lose the most rank separability at $\alpha{=}1$, and on one they fall to chance while affirmed statements on the same trajectory keep AUROC 0.89; conjunctions lose the least on all three. An aggregate AUROC at full injection can therefore hide a complete failure on one statement form, so pipelines that read negated claims through a learned adapter should evaluate them separately.

6 Conclusion

We set out to determine whether a language model's internal representation of truth survives heterogeneous activation transfer. Ordinal truth structure survives across all 21 cross-family trajectories through most of the injection range. Full replacement through a learned adapter at $\alpha{=}1$ collapses parametric separability $\sim 14\times$ as rank-based separability degrades more gradually, while a closed-form rigid map retains most rank separability at the same point (median AUROC 0.91 against 0.72). The apparent accuracy gains under partial injection decompose into a recalibration component dominated by native probe miscalibration ($R^2 = 0.75$) and a smaller separability gain, bootstrap-confirmed positive on 17 of 21 trajectories, that random maps of the same architecture also produce and that trained adapters exceed on one or two of three tested trajectories, depending on the evaluation set. These findings characterize cross-family truth structure as *partially compatible and ordinally stable, with full-substitution failure depending largely on the alignment method*. Future work includes a truth-aware adapter objective, seed replication and random-map nulls (including random linear maps) on all 21 trajectories, the Llama 3.2 3B $\rightarrow$ Mistral 7B pairs where the learned adapter outperforms Procrustes, other concepts, and larger models.

Acknowledgments and Disclosures

Funding: This work received no third-party funding.

Competing interests: The authors declare no competing interests.

Statement on AI Usage: LLMs assisted with drafting parts of this manuscript and with writing the experiment and analysis code. The methodology and experimental design are the authors' own, and every reported number is regenerated from saved results by the released notebooks.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. 2022. URL https://arxiv.org/abs/2204.05862.

Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in llms, 2024. URL https://arxiv.org/abs/2407.12831.

Zhuoyun Du, Runze Wang, Huiyu Bai, Zouying Cao, Xiaoyong Zhu, Yu Cheng, Bo Zheng, Wei Chen, and Haochao Ying. Enabling agents to communicate entirely in latent space, 2026. URL https://arxiv.org/abs/2511.09149.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models, 2026. URL https://arxiv.org/abs/2510.03215.

Gemma Team. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024a.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view, 2026. URL https://arxiv.org/abs/2602.14486.

Wes Gurnee and Max Tegmark. Language models represent space and time, 2024. URL https://arxiv.org/abs/2310.02207.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings, 2026. URL https://arxiv.org/abs/2505.12540.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. Vhelm: A holistic evaluation of vision language models, 2024. URL https://arxiv.org/abs/2410.07112.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Narmeen Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. Activation space interventions can be transferred between large language models, 2025. URL https://arxiv.org/abs/2503.04429.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025a. URL https://arxiv.org/abs/2310.01405.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, Gaotang Li, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems, 2025b. URL https://arxiv.org/abs/2511.20639.

Glossary

ARH Aristotelian Representation Hypothesis. 3, 8

FVU Fraction of variance unexplained. 5, 6, 9, 10, 14, 15, 22

LLM Large Language Model. 1–3, 10

PRH Platonic Representation Hypothesis. 3, 8

RepE Representation Engineering. 1, 2

RLHF Reinforcement Learning from Human Feedback. 6

RSA Representational Similarity Analysis. 4, 13, 14

TTPD Training of Truth and Polarity Direction. 3, 4, 6, 13, 16, 17, 27

A Methodology Details

Changes from the reviewed version. This version adds multi-seed replication (Appendix A.13), a seeded random-map null and shuffled control (Appendix A.14), a statement-type decomposition (Appendix A.15) and a residual-geometry analysis (Appendix A.16), as discussed during review. A duplicated trajectory is now counted once, so aggregates are over 21 unique trajectories. One claim changed after re-measurement: the reviewed version and our rebuttal stated that the closed-form Procrustes map collapses at $\alpha{=}1$ like the trained adapter, a conclusion drawn from default-threshold accuracy alone. Measured by AUROC, Procrustes retains most of the target's rank separability (median 0.91 against 0.72, higher on all 21 trajectories), so Sections 4.2, 4.5 and 5 now describe the collapse as depending on the alignment method. The random-map null also narrows the strict-overshoot claim: trained adapters exceed random maps of the same architecture on two of three tested trajectories on the full test set and on one of three on simple statements.

This appendix contains procedural details for the methods introduced in Section 3.

A.1 Truth-layer calibration

For each model M, we identify a layer at which residual-stream activations most clearly separate true from false statements. For each candidate layer ℓ we extract last-token activations $h_\ell(x) \in \mathbb{R}^d$ over a pool of $N = 800$ statements, 200 sampled without a fixed seed from each of four contrastive truth datasets — `cities`, `neg_cities`, `sp_en_trans`, `neg_sp_en_trans` — and apply topic-centering to obtain $\tilde{h}_\ell(x)$. With class means $\hat{\mu}_T, \hat{\mu}_F$ over true and false statements, we score each layer by the Fisher trace ratio

$$J(\ell) = \frac{\|\hat{\mu}_T - \hat{\mu}_F\|_2^2}{\hat{\sigma}_T^2 + \hat{\sigma}_F^2}, \qquad \hat{\sigma}_c^2 = |c|^{-1} \sum_{x \in c} \|\tilde{h}_\ell(x) - \hat{\mu}_c\|_2^2, \tag{2}$$

i.e., the ratio of squared between-class centroid distance to summed within-class spread. The pool intentionally combines affirmative and negated datasets so the selected layer reflects polarity-invariant truth structure rather than a surface artifact of either form. Because the sampling is unseeded, we ran the calibration four times per model and selected the layer that ranked at or near the top of $J(\ell)$ across runs.

A.2 Transfer-layer selection via RSA

For each model pair, we restrict candidate layers to the four central layers of each model (e.g., layers 11–14 of Gemma 2 2B) to avoid early embedding-specific and late prediction-specific representations, and score every source/target layer pair by cross-model RSA [Oozeer et al., 2025]. For 1000 texts drawn with a fixed seed from the adapters' training corpus and taken in fixed groups of 8, we z-score each text's activations per feature, define the distance between two texts as the mean over all token pairs of a clamped dot-product distance (which needs no token alignment across tokenizers; because the vectors are not normalized, the clamp saturates for most token pairs, so the distance mainly measures sign agreement), and correlate the source and target distances for the 28 pairs in every group (3,500 per side) with Spearman's ρ. Texts are compared only within their group. We select the pair with the highest ρ:

$$(s^*, t^*) = \arg\max_{s_i, t_j} \mathrm{RSA}(\mathcal{A}_{s_i}, \mathcal{A}_{t_j}), \tag{3}$$

where $\mathcal{A}_\ell$ denotes the activation patterns at layer ℓ.

A.3 TTPD probe procedure

At each model's calibrated layer ℓ_M^*, we fit TTPD [Bürger et al., 2024] as follows. Training activations $h(x) \in \mathbb{R}^d$ are extracted at the final-token position over six topic-specific datasets — `cities`, `sp_en_trans`, `inventors`, `animal_class`, `element_symb`, `facts` — in both affirmative and negated forms. Each statement carries a binary truth label $\tau \in \{-1, +1\}$ and polarity $p \in \{-1, +1\}$. Within each topic we balance counts across the four (p, τ) cells via stratified subsampling. For each topic i we compute the empirical mean $\hat{\mu}_i$ and form topic-centered activations $\tilde{h}(x) = h(x) - \hat{\mu}_i$.

TTPD is fit in four steps:

1. The general truth direction $t_G \in \mathbb{R}^d$ is recovered as the OLS coefficient of τ in the regression $\tilde{h}(x) = \tau\, t_G + \tau p\, t_P + \varepsilon$, solved in closed form.
2. A polarity direction $t_P \in \mathbb{R}^d$ is learned by unpenalized logistic regression on the uncentered activations with polarity as the target.
3. Activations are projected into the two-dimensional space $(h(x)^\top t_G,\ h(x)^\top t_P)$.
4. An unpenalized logistic regression with intercept is fit on these 2D projections to predict τ.

The logistic-regression baseline is an unpenalized logistic regression with intercept fit directly on the d-dimensional activations $h(x)$.

A.4 Autoencoder adapter training

For each pair, we train a two-layer MLP adapter $\hat{h}_T = W_2 \cdot \mathrm{ReLU}(W_1 h_S + b_1) + b_2$ on a corpus combining Anthropic RLHF [Bai et al., 2022] with domain-specific examples [Oozeer et al., 2025]. Training settings:

- **Optimizer:** AdamW, learning rate 1×10^{-4}.
- **Batch size:** 8 with 4 gradient accumulation steps (effective 32).
- **Sequence length:** 720 tokens.
- **Precision:** FP16.
- **Validation:** FVU on held-out activations; the lowest-FVU checkpoint is selected per pair.

For pairs spanning architectures with different tokenizers, we additionally implement vocabulary-space alignment to handle tokenization differences while preserving semantic mappings, and attention-mask normalization with adaptive sequence padding.

A.5 Layer-selection ablation: full results

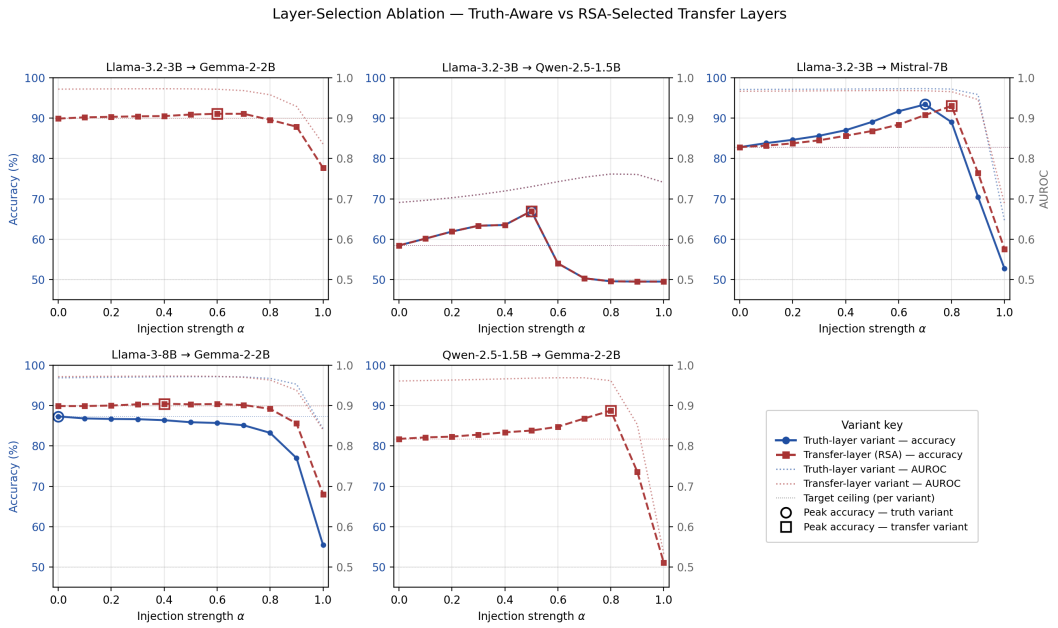


Figure 6: Layer-selection ablation across five (source, target) pairs. Solid lines: truth-layer variant (autoencoder mapped onto the probe-calibrated layer). Dashed/dotted lines: transfer-layer variant (autoencoder mapped onto the RSA-selected layer). Both variants exhibit plateau-then-collapse, AUROC–accuracy decoupling, and overshoot above the per-variant native ceiling.

We compare the truth-layer and transfer-layer variants on five trajectories spanning four families overall (Figure 6). The intent is not to declare a universal winner — the sample is too small, and as we discuss below, the two variants are sometimes anchored at very similar layers, making the comparison underpowered for some pairs. Instead, we ask whether the qualitative findings of Sections 4.2–4.4 survive a non-trivial perturbation to the layer-selection criterion.

Qualitative patterns are robust. Plateau-then-collapse is visible in every panel for both variants, as is the AUROC–accuracy decoupling. Overshoot above the per-variant native ceiling appears in both variants: the truth-layer variant overshoots on two Llama-source pairs (peaking at 93.4% on Llama-3.2-3B $\rightarrow$ Mistral-7B from a ceiling of 82.8%, and on the degenerate Llama-3.2-3B $\rightarrow$ Qwen-2.5-1.5B pair below); the transfer-layer variant overshoots on all five.

Table 2: Master results across all 22 alpha-sweep trajectories. Overshoot values above 0.5 pp are bolded. LAYERS gives source→target probed layers ($L_s \rightarrow L_t$). VARIANT marks whether the adapter trains and probes at the pair's truth layer (TRUTH) or transfer layer (TRANSFER). Rows 7 and 8 are the Llama 3.2 3B → Qwen 2.5 1.5B 12 → 12 configuration swept under both variant labels with the same adapter and probes; they are identical. Aggregate statistics dedup to 21 unique trajectories. $\alpha^*_{\tau_0}$ is the α on the 0.1-step grid that maximizes accuracy at the default threshold τ_0=0.5; it differs from $\alpha^*_{\tau^*}$ in Table 5, which maximizes accuracy with the threshold re-tuned at each α. Where $\Delta\text{Acc} = 0$, injection never beats native accuracy and $\alpha^*_{\tau_0}$ defaults to 0, marking no gain rather than the location of a peak; it is not interpreted there or where ΔAcc is below 0.5 pp.

				Adapter		Ceiling ($\alpha = 0$)		Peak			Full ($\alpha = 1$)		ΔAcc
Source	Target	Layers	Variant	FVU	Cos.	Acc.	AUROC	Acc.	$\alpha^*_{\tau_0}$	AUROC	Acc.	AUROC	(pp)
Llama 3 8B	Gemma 2 2B	14→14	Transfer	0.279	0.723	89.9	0.972	90.4	0.4	0.973	68.0	0.840	**+0.6**
Llama 3 8B	Gemma 2 2B	12→13	Truth	0.275	0.734	87.3	0.969	87.3	0.0	0.969	55.5	0.843	0.0
Llama 3 8B	Mistral 7B	14→14	Transfer	0.047	0.531	82.8	0.966	91.9	0.7	0.970	50.4	0.618	**+9.1**
Llama 3.2 3B	Gemma 2 2B	13→14	Transfer	0.279	0.723	89.9	0.972	91.1	0.6	0.971	77.7	0.836	**+1.2**
Llama 3.2 3B	Mistral 7B	12→14	Transfer	0.047	0.547	82.8	0.966	93.0	0.8	0.965	57.6	0.688	**+10.2**
Llama 3.2 3B	Mistral 7B	12→13	Truth	0.040	0.534	82.8	0.971	93.4	0.7	0.973	52.7	0.646	**+10.6**
Llama 3.2 3B	Qwen 2.5 1.5B	12→12	Transfer	0.002	0.826	58.5	0.691	67.0	0.5	0.730	49.5	0.741	**+8.5**
Llama 3.2 3B	Qwen 2.5 1.5B	12→12	Truth	0.002	0.826	58.5	0.691	67.0	0.5	0.730	49.5	0.741	**+8.5**
Mistral 7B	Gemma 7B	14→15	Transfer	0.090	0.605	84.1	0.952	84.1	0.0	0.952	49.5	0.667	0.0
Qwen 2.5 1.5B	Gemma 2 2B	12→11	Transfer	0.281	0.735	81.7	0.961	88.7	0.8	0.962	51.1	0.532	**+7.0**
Qwen 2.5 1.5B	Llama 3.2 3B	12→12	Transfer	0.019	0.762	77.0	0.929	84.7	0.8	0.961	49.0	0.546	**+7.6**
Gemma 2 2B	Llama 3.2 3B	13→12	Truth	0.035	0.563	77.0	0.929	77.0	0.0	0.929	57.5	0.708	0.0
Gemma 2 2B	Qwen 2.5 1.5B	13→17	Truth	0.005	0.718	87.6	0.944	87.6	0.0	0.944	49.5	0.727	0.0
Llama 3 8B	Mistral 7B	12→13	Truth	0.044	0.510	82.8	0.971	93.2	0.8	0.973	50.5	0.609	**+10.4**
Llama 3.2 3B	Gemma 2 2B	12→13	Truth	0.274	0.734	87.3	0.969	87.3	0.0	0.969	59.2	0.789	0.0
Llama 3.2 3B	Qwen 2.5 1.5B	12→17	Truth	0.003	0.840	87.6	0.944	87.6	0.0	0.944	49.5	0.850	0.0
Mistral 7B	Gemma 7B	13→19	Truth	0.125	0.583	91.0	0.969	91.6	0.2	0.970	49.8	0.745	**+0.6**
Qwen 2.5 1.5B	Gemma 2 2B	17→13	Truth	0.278	0.730	87.3	0.969	87.3	0.0	0.969	57.4	0.845	0.0
Qwen 2.5 1.5B	Llama 3.2 3B	17→12	Truth	0.019	0.760	77.0	0.929	89.6	0.9	0.951	61.7	0.716	**+12.5**
Qwen 2.5 7B	Gemma 7B	21→19	Truth	0.106	0.582	91.0	0.969	91.4	0.1	0.970	50.2	0.663	+0.4
Qwen 2.5 7B	Llama 3 8B	21→12	Truth	0.036	0.714	95.8	0.984	96.1	0.4	0.985	65.1	0.723	+0.3
Qwen 2.5 7B	Mistral 7B	21→13	Truth	0.048	0.495	82.8	0.971	94.2	0.7	0.974	50.4	0.550	**+11.4**

Per-trajectory winners mostly track native ceiling. Of the three pairs with a non-trivial ceiling gap, the higher-ceiling variant produces the higher peak on two. The transfer-layer variant wins on Llama-3-8B → Gemma-2-2B (transfer ceiling 89.9% vs. truth ceiling 87.3%, a 2.6 pp gap) and on Llama-3.2-3B → Gemma-2-2B (transfer 89.9% vs. truth 87.3%, 2.6 pp). On Qwen-2.5-1.5B → Gemma-2-2B the truth-layer variant has the higher ceiling (87.3% vs. 81.7%), but the transfer-layer variant's +7.0 pp overshoot carries it past (88.7% vs. 87.3%). On Llama-3.2-3B → Mistral-7B the native ceilings are identical (82.80%) and the truth-layer variant wins by a marginal 0.38 pp at peak. The layer-selection question therefore reduces mostly to where each criterion's chosen target layer sits on the truth-separability spectrum, though overshoot can reverse a modest ceiling gap.

Variants are not always strongly distinct. In several pairs, the truth-layer and transfer-layer target choices are within one layer of each other. On Llama-3.2-3B → Gemma-2-2B the variants map to target layers 13 and 14; on Llama-3.2-3B → Mistral-7B, target layers 13 and 14. The Llama-3.2-3B → Qwen-2.5-1.5B pair is a degenerate case: both variants were run at the same source layer (12) and target layer (12), making them identical (we report this as a single trajectory in aggregate statistics; see the Table 2 caption). Only Qwen-2.5-1.5B → Gemma-2-2B differs substantially on both sides (12→11 versus 17→13); Llama-3-8B → Gemma-2-2B differs by two layers at the source and one at the target (14→14 versus 12→13). The ablation is thus sometimes comparing configurations that differ less than the framing implies; attributing observed differences to the layer-selection criterion rather than to run-to-run variation in adapter training requires caution. A more decisive comparison would require more trajectories where the two criteria select substantially different layers on both sides, with multiple adapter training seeds per configuration.

A.6 Per-pair α-sweep results

Table 2 reports complete per-pair sweep numbers across all 22 trajectories: adapter FVU and adapter–target cosine similarity, native target ceiling at α=0, peak accuracy and the $\alpha^*_{\tau_0}$ at which it occurs, full-injection (α=1) accuracy and AUROC, and the default-threshold overshoot $\Delta\text{Acc}(\tau_0)$. Aggregate medians and IQRs cited in Section 4 are computed over the 21 unique trajectories (dedu-

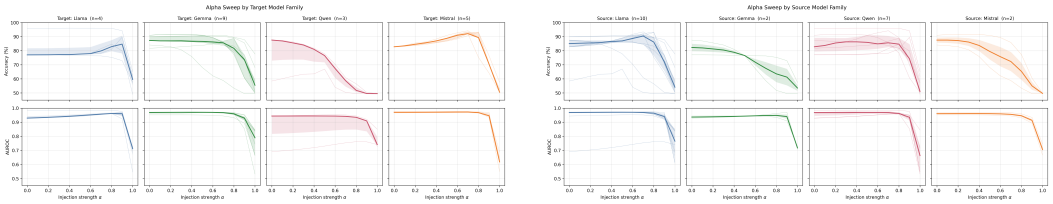


Figure 7: Alpha sweep parsed by target family (left) and source family (right).

plicating the Llama 3.2 3B → Qwen 2.5 1.5B 12→12 row that appears identically under both layer-selection variants).

A.7 Per-family α-sweep breakdowns

Figure 7 shows median AUROC and accuracy across the 21 trajectories disaggregated by target family (left) and source family (right). All eight family-level cohorts exhibit the same qualitative trajectory shape described in Section 4.2.

A.8 Threshold-calibration diagnostic for the overshoot

The default decision threshold for the TTPD probe is $\tau_0 = 0.5$, the natural threshold of its final-stage logistic regressor. If τ_0 were suboptimally placed for the held-out test distribution at $\alpha{=}0$, then any adapter-induced shift of the score distribution along the truth direction could move scores into better alignment with τ_0 without supplying any genuine new information. Such a shift would reproduce the overshoot pattern we report in Section 4.4 purely as a calibration artifact.

Procedure. For each probe, we sweep τ over a 1001-point grid on $[0, 1]$, evaluate accuracy on the held-out $\alpha{=}0$ test set, and record the optimal threshold τ^*, the resulting accuracy $\mathrm{Acc}(\tau^*, \alpha{=}0)$, and the gain $\Delta\mathrm{Acc} = \mathrm{Acc}(\tau^*, \alpha{=}0) - \mathrm{Acc}(\tau_0, \alpha{=}0)$ over the default threshold. The diagnostic is designed as a lower-bound test: if the native probe's accuracy ceiling under any threshold falls below the overshoot peak observed under partial injection, then threshold miscalibration alone cannot account for the overshoot. To match the main sweep's evaluation population, we apply the same simple-statement filter prior to scoring ($n_{\text{test}} = 1599$; unfiltered $n = 2684$).

Table 3 reports optimal-threshold accuracy across the 15 distinct (model, layer) probe configurations participating in our main experiments (some serve as targets in some trajectories and sources in others)($n_{\text{test}} = 1599$, simple statements only, matching the main sweep population). Optimal thresholds τ^* deviate substantially from 0.5 for many models — ranging from 0.117 to 0.955 — with accuracy gaps $\Delta\mathrm{Acc} = \mathrm{Acc}(\tau^*) - \mathrm{Acc}(\tau_0)$ of up to 14.13 pp (Qwen-2.5-7B-Instruct, L21). On half of the 12 trajectories whose default-threshold overshoot exceeds 0.5 pp, $\mathrm{Acc}(\tau^*, \alpha{=}0)$ matches or exceeds the peak accuracy observed under partial injection with the default threshold, so recalibration alone can account for their overshoot. On the other six (the four Mistral-target trajectories with an overshoot, Llama 3.2 3B → Qwen 2.5 1.5B at 12→12, and Qwen 2.5 1.5B (L17) → Llama 3.2 3B), the injection peak exceeds even the native optimal-threshold ceiling. The strict-overshoot analysis in Section 4.4 extends this diagnostic to per-α optimal thresholds and finds a small residual gain ($+1.13$ pp median, bootstrap-confirmed positive on 17 of 21 trajectories) that recalibration cannot account for. Random maps of the adapter's architecture also produce such gains, so the residual is partly a consequence of injecting source information at all (Appendix A.14).

A.9 Threshold-trajectory illustration

Figure 8 illustrates the dual-mechanism decomposition of Section 4.4 on three representative pairs. Each panel plots two τ-trajectories: the gap between τ_0 and $\tau^*(\alpha)$ at a given α visualizes the recalibration effect, and the rise of $\tau^*(\alpha)$ above the native τ^* ceiling visualizes the strict overshoot.

Table 3: Threshold-calibration diagnostic across 15 target probes, evaluated on simple statements of both polarities ($n_{\text{test}} = 1599$), matching the population used in the main transfer sweep. τ^* is the threshold maximizing accuracy on the native (α=0) test set; τ_0=0.5 is the TTPD probe's fixed deployment threshold. Δ Acc is the accuracy gap between optimal and default threshold on native activations. On half of the 12 trajectories with default-threshold overshoot above 0.5 pp, $\mathrm{Acc}(\tau^*, \alpha{=}0)$ matches or exceeds the injection peak under τ_0; the per-α strict-overshoot analysis in Section 4.4 characterizes the residual gain that this lower-bound diagnostic does not capture. Values come from the probe calibration run; the sweeps' own α=0 evaluation differs by one test item for gemma2-2b-instruct L11 (τ^*=0.805, $\mathrm{Acc}(\tau^*)$=0.9156) and mistral-7b-instruct L14 ($\mathrm{Acc}(\tau_0)$=0.8280). Sorted by AUROC descending.

Probe	τ^*	Acc(τ^*)	Acc(τ_0)	AUROC	Δ Acc	n_{test}
llama3-8b-instruct, L12	0.613	0.9631	0.9575	0.9844	0.0056	1599
llama3-8b-instruct, L14	0.641	0.9437	0.9318	0.9757	0.0119	1599
gemma2-2b-instruct, L14	0.580	0.9187	0.8993	0.9715	0.0194	1599
mistral-7b-instruct, L13	0.843	0.9306	0.8280	0.9708	0.1026	1599
gemma-7b-instruct, L19	0.480	0.9162	0.9106	0.9694	0.0056	1599
gemma2-2b-instruct, L13	0.818	0.9099	0.8718	0.9686	0.0381	1599
mistral-7b-instruct, L14	0.865	0.9212	0.8286	0.9663	0.0926	1599
gemma2-2b-instruct, L11	0.802	0.9149	0.8174	0.9608	0.0976	1599
gemma-7b-instruct, L15	0.747	0.9024	0.8405	0.9517	0.0619	1599
gemma2-2b-instruct, L12	0.812	0.8937	0.8143	0.9486	0.0794	1599
qwen2.5-1.5b-instruct, L17	0.520	0.8799	0.8749	0.9437	0.0050	1599
llama3.2-3b-instruct, L12	0.953	0.8468	0.7705	0.9290	0.0763	1599
llama3.2-3b-instruct, L13	0.955	0.8386	0.7642	0.9185	0.0744	1599
qwen2.5-7b-instruct, L21	0.117	0.8849	0.7436	0.9166	0.1413	1599
qwen2.5-1.5b-instruct, L12	0.299	0.6398	0.5847	0.6910	0.0550	1599

Table 4: OLS regressions of overshoot magnitude on pair-level features. Left: $\Delta\mathrm{Acc}(\tau_0)$ on the default-threshold overshoot. Right: $\Delta\mathrm{Acc}(\tau^*)$ on the strict overshoot under per-α optimal thresholds. The default-threshold overshoot is dominated by the native ceiling gap ($\beta = +1.07$, $p < 10^{-4}$), consistent with the recalibration story. The strict-overshoot regression has lower R^2 and no single dominant predictor, with three predictors borderline-significant in coherent directions. $n = 21$.

	$\Delta\mathrm{Acc}(\tau_0)$				$\Delta\mathrm{Acc}(\tau^*)$			
Predictor	β	SE	t	p	β	SE	t	p
(intercept)	-0.010	0.162	-0.06	0.952	+0.099	0.122	+0.81	0.43
ceiling gap	+1.068	0.197	+5.41	5.74e-05	+0.303	0.148	+2.05	0.0576
FVU	-0.112	0.064	-1.74	0.101	-0.057	0.048	-1.17	0.258
cosine sim.	+0.058	0.081	+0.71	0.486	+0.121	0.061	+2.00	0.0631
AUROC($\alpha = 0$)	-0.031	0.131	-0.24	0.815	-0.175	0.099	-1.78	0.094
R^2	0.748				0.578			
adj R^2	0.686				0.472			
F	11.90 ($p = 0.000112$)				5.47 ($p = 0.00567$)			

A.10 Score-distribution histograms

To make the dual-mechanism picture from Section 4.4 visible at the score level, Figures 9–11 show TTPD probe score distributions for three representative trajectories spanning the strict-overshoot range. Each row shows histograms of true (green) and false (red) scores at three points in the α-sweep: α=0 (native), $\alpha^*_{\tau^*}$ (peak strict overshoot), and α=1 (full injection). Two observations are visible across all three pairs. First, partial injection systematically shifts both true and false score distributions toward the center of the score range, with $\tau^*(\alpha)$ moving correspondingly — the recalibration mechanism. Second, the distributions tighten internally as they shift, preserving and slightly improving their separability relative to the gap between their means — the strict-overshoot mechanism. At α=1 the distributions overlap heavily, the centroid distance approaches zero, and τ^* ceases to discriminate.

Table 5: Strict overshoot under per-α optimal thresholds. For each trajectory, $\tau^*(\alpha)$ is the threshold maximizing accuracy at that α on the held-out test set ($n_{\text{test}} = 1599$). $\Delta\text{Acc}(\tau^*) = \max_\alpha \text{Acc}(\tau^*(\alpha), \alpha) - \text{Acc}(\tau^*(0), 0)$ is the separability gain that recalibration cannot account for. All 21 trajectories show $\Delta\text{Acc}(\tau^*) > 0$; bolded values exceed 1.0 pp. Thresholds come from a 1001-point grid; the text quotes the exact-threshold values of Table 7. $\alpha^*_{\tau^*}$ is the α on the 0.1-step grid attaining the maximum; it differs from the default-threshold $\alpha^*_{\tau_0}$ of Table 2, so where the two disagree, separability and calibration peak at different injection strengths. Near the resolution floor the accuracy curve is flat to within a few test items, so $\alpha^*_{\tau^*}$ is not identified on the four trajectories whose bootstrap CI includes zero (Table 7), and a difference of one grid step is not interpreted. Sorted by $\Delta\text{Acc}(\tau^*)$ descending.

SOURCE	TARGET	LAYERS	VARIANT	$\text{Acc}(\tau^*, 0)$	$\tau^*(0)$	$\text{Acc}(\tau^*, \alpha^*_{\tau^*})$	$\alpha^*_{\tau^*}$	$\Delta\text{Acc}(\tau^*)$
Gemma 2 2B	Llama 3.2 3B	13→12	Truth	84.7	0.953	93.6	0.9	**+8.94**
Llama 3.2 3B	Qwen 2.5 1.5B	12→12	Transfer	63.9	0.299	72.2	0.9	**+8.26**
Qwen 2.5 1.5B	Llama 3.2 3B	12→12	Transfer	84.7	0.953	91.9	0.8	**+7.19**
Qwen 2.5 1.5B	Llama 3.2 3B	17→12	Truth	84.7	0.953	91.6	0.8	**+6.94**
Llama 3.2 3B	Qwen 2.5 1.5B	12→17	Truth	88.0	0.520	91.0	0.7	**+3.00**
Qwen 2.5 1.5B	Gemma 2 2B	12→11	Transfer	91.6	0.805	94.4	0.7	**+2.88**
Llama 3 8B	Gemma 2 2B	12→13	Truth	90.9	0.806	92.6	0.5	**+1.69**
Llama 3.2 3B	Mistral 7B	12→14	Transfer	92.1	0.865	93.5	0.7	**+1.38**
Llama 3 8B	Gemma 2 2B	14→14	Transfer	91.9	0.590	93.2	0.5	**+1.31**
Llama 3.2 3B	Gemma 2 2B	12→13	Truth	90.9	0.806	92.2	0.5	**+1.25**
Llama 3.2 3B	Gemma 2 2B	13→14	Transfer	91.9	0.590	93.0	0.4	**+1.06**
Qwen 2.5 7B	Mistral 7B	21→13	Truth	93.1	0.844	94.2	0.7	**+1.06**
Llama 3 8B	Mistral 7B	14→14	Transfer	92.1	0.865	93.1	0.7	+0.94
Mistral 7B	Gemma 7B	13→19	Truth	91.6	0.480	92.3	0.6	+0.69
Llama 3 8B	Mistral 7B	12→13	Truth	93.1	0.844	93.7	0.8	+0.56
Llama 3.2 3B	Mistral 7B	12→13	Truth	93.1	0.844	93.6	0.8	+0.50
Qwen 2.5 1.5B	Gemma 2 2B	17→13	Truth	90.9	0.806	91.3	0.4	+0.38
Gemma 2 2B	Qwen 2.5 1.5B	13→17	Truth	88.0	0.520	88.3	0.8	+0.31
Qwen 2.5 7B	Llama 3 8B	21→12	Truth	96.3	0.614	96.6	0.5	+0.31
Qwen 2.5 7B	Gemma 7B	21→19	Truth	91.6	0.480	91.8	0.5	+0.19
Mistral 7B	Gemma 7B	14→15	Transfer	90.2	0.748	90.4	0.3	+0.13

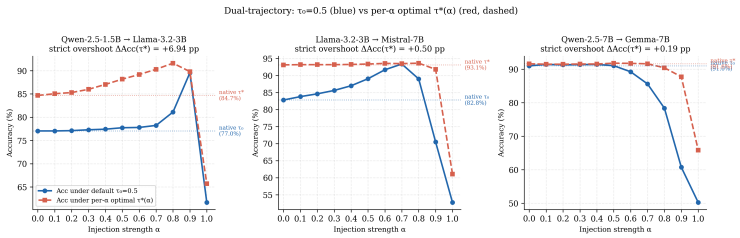


Figure 8: Two τ-trajectories per pair across three representative pairs. Gap between τ_0 and $\tau^*(\alpha)$ at any α is the recalibration effect; rise of $\tau^*(\alpha)$ above the native τ^* ceiling is the strict overshoot.

A.11 Per-pair Procrustes baseline results

This appendix reports the per-pair Procrustes baseline results summarized in Section 4.5. Each row gives the trained-adapter and Procrustes τ_0 accuracy at the peak and at α=1, relative to the native target probe ceiling, and the native probe's AUROC at α=1. Δ_{peak} values above 0.5 pp are in bold. Pairs are sorted by trained-adapter peak overshoot in descending order. The three pairs where the trained adapter strictly exceeds Procrustes at peak (Llama 3.2 3B → Mistral 7B at L12→L13 and L12→L14; Mistral 7B → Gemma 7B at L13→L19) are italicized in the source column. Procrustes was fit on the same RLHF reconstruction corpus used to train the autoencoder adapter (seed=42, N=2000 prompts); ceiling values reproduce native target probe accuracy on the 1599-statement TiU test split (simple statements, both polarities). Full per-α sweeps are released with the code.

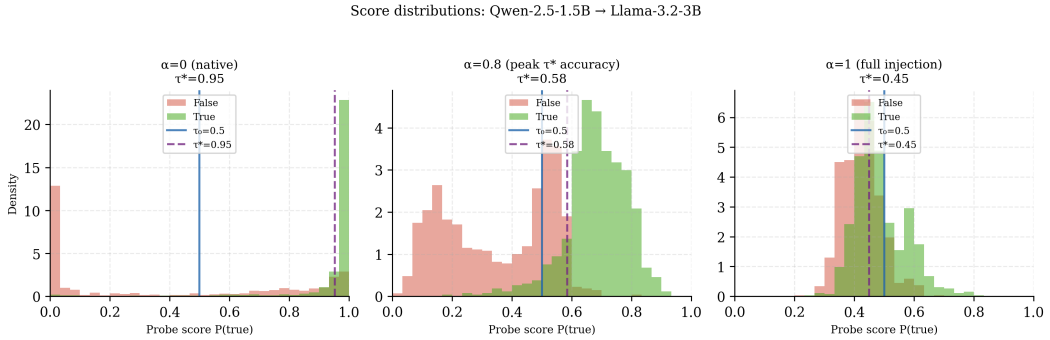


Figure 9: Score distributions for Qwen-2.5-1.5B → Llama-3.2-3B (high-overshoot example, $\Delta\mathrm{Acc}(\tau^*) = +6.94$ pp).

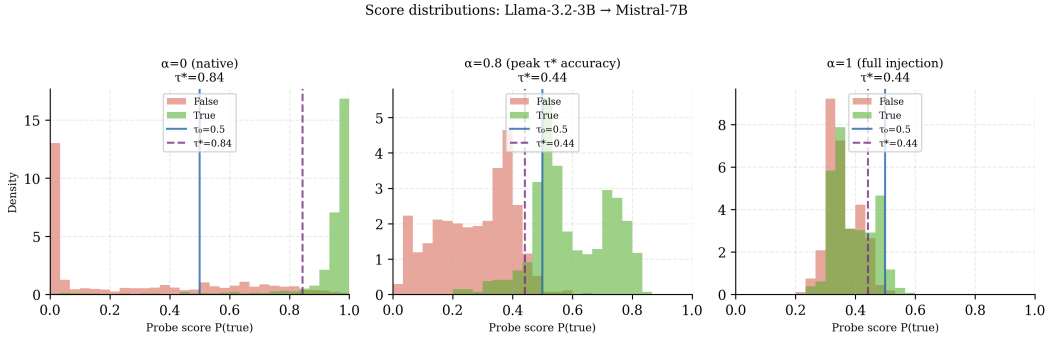


Figure 10: Score distributions for Llama-3.2-3B → Mistral-7B (moderate-overshoot example, $\Delta\mathrm{Acc}(\tau^*) = +0.50$ pp).

Table 6: Per-pair Procrustes vs. trained-adapter comparison across all 21 unique trajectories. LAYERS gives source→target probed layers. Δ_{peak}: peak τ_0 accuracy minus ceiling. $\Delta_{\alpha=1}$: τ_0 accuracy at full injection minus ceiling. $\mathrm{AUROC}_{\alpha=1}$: the native probe's AUROC at full injection. The Llama 3.2 3B → Qwen 2.5 1.5B 12→12 row appears once; it was swept identically under both variant labels.

			Ceiling		Trained			Procrustes		
Source	Target	Layers	Acc. (%)	AUROC	Δ_{peak}	$\Delta_{\alpha=1}$	$\mathrm{AUROC}_{\alpha=1}$	Δ_{peak}	$\Delta_{\alpha=1}$	$\mathrm{AUROC}_{\alpha=1}$
Qwen 2.5 1.5B	Llama 3.2 3B	17→12	77.05	0.929	+12.51	−15.38	0.716	+17.01	−10.69	0.840
Qwen 2.5 7B	Mistral 7B	21→13	82.80	0.971	+11.38	−32.40	0.550	+12.63	−23.39	0.910
Llama 3.2 3B	Mistral 7B	12→13	82.80	0.971	+10.57	−30.08	0.646	+0.00	−5.13	0.922
Llama 3 8B	Mistral 7B	12→13	82.80	0.971	+10.44	−32.27	0.609	+11.63	+0.13	0.954
Llama 3.2 3B	Mistral 7B	12→14	82.80	0.966	+10.19	−25.20	0.688	+0.00	−6.50	0.922
Llama 3 8B	Mistral 7B	14→14	82.80	0.966	+9.13	−32.40	0.618	+10.63	+6.44	0.942
Llama 3.2 3B	Qwen 2.5 1.5B	12→12	58.47	0.691	+8.51	−8.94	0.741	+19.39	+10.38	0.843
Qwen 2.5 1.5B	Llama 3.2 3B	12→12	77.05	0.929	+7.63	−28.08	0.546	+8.19	−23.45	0.627
Qwen 2.5 1.5B	Gemma 2 2B	12→11	81.74	0.961	+7.00	−30.64	0.532	+11.19	−25.33	0.592
Llama 3.2 3B	Gemma 2 2B	13→14	89.87	0.972	+1.19	−12.20	0.836	+0.75	−11.51	0.910
Mistral 7B	Gemma 7B	13→19	90.99	0.969	+0.63	−41.15	0.745	+0.00	−24.77	0.813
Llama 3 8B	Gemma 2 2B	14→14	89.87	0.972	+0.56	−21.83	0.840	+6.07	−1.69	0.924
Qwen 2.5 7B	Gemma 7B	21→19	90.99	0.969	+0.44	−40.78	0.663	+0.50	−23.83	0.927
Qwen 2.5 7B	Llama 3 8B	21→12	95.81	0.984	+0.25	−30.71	0.723	+1.19	−39.84	0.929
Gemma 2 2B	Llama 3.2 3B	13→12	77.05	0.929	+0.00	−19.57	0.708	+14.45	+11.38	0.942
Gemma 2 2B	Qwen 2.5 1.5B	13→17	87.55	0.944	+0.00	−38.02	0.727	+0.19	−9.88	0.904
Llama 3 8B	Gemma 2 2B	12→13	87.30	0.969	+0.00	−31.83	0.843	+6.00	−3.06	0.920
Llama 3.2 3B	Gemma 2 2B	12→13	87.30	0.969	+0.00	−28.14	0.789	+0.00	−10.76	0.923
Llama 3.2 3B	Qwen 2.5 1.5B	12→17	87.55	0.944	+0.00	−38.02	0.850	+0.56	−24.20	0.901
Mistral 7B	Gemma 7B	14→15	84.05	0.952	+0.00	−34.52	0.667	+0.00	−13.45	0.859
Qwen 2.5 1.5B	Gemma 2 2B	17→13	87.30	0.969	+0.00	−29.89	0.845	+4.57	−23.14	0.875

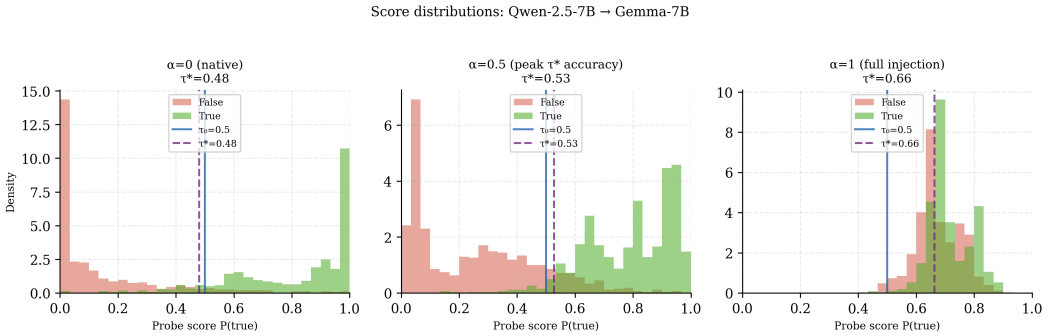


Figure 11: Score distributions for Qwen-2.5-7B → Gemma-7B (low-overshoot example, $\Delta\mathrm{Acc}(\tau^*) = +0.19$ pp).

A.12 Bootstrap robustness for strict overshoot

To assess whether the universally-positive direction of $\Delta\mathrm{Acc}(\tau^*)$ reported in Section 4.4 is robust against test-set sampling, we run a paired stratified bootstrap on the held-out test set within each trajectory ($n_{\text{test}} = 1599$, with m_1 true-labeled and m_0 false-labeled samples). For $B = 10{,}000$ replicates with seed 42, we draw stratified resamples — separate sampling indices for the true-labeled and false-labeled subsets, sampled with replacement, with the same indices applied across all α values to preserve the per-sample correspondence across the sweep. For each replicate b, we compute the optimal-threshold accuracy $\mathrm{Acc}(\tau^*(\alpha), \alpha)$ at every α on the resampled test set and the strict overshoot $\Delta_b = \max_\alpha \mathrm{Acc}(\tau^*(\alpha), \alpha) - \mathrm{Acc}(\tau^*(0), 0)$. The 95% CI is the empirical $[2.5, 97.5]$ percentile interval of $\{\Delta_b\}_{b=1}^{B}$. The bootstrap captures variance from finite-test-set sampling but does not capture variance from adapter training, which Appendix A.13 measures on three trajectories, and it tests against zero rather than against untrained maps (Appendix A.14).

Across the 21 unique trajectories, the 95% CI lower bound exceeds zero on 17. All four exceptions have point estimates $\Delta\mathrm{Acc}(\tau^*) \leq 0.50$ pp and CI lower bounds at exactly $+0.00$ pp, reflecting the resolution of the test set: at $n_{\text{test}} = 1599$, accuracy granularity is ~ 0.06 pp, and when the true effect is comparable to per-replicate noise the bootstrap correctly hedges by placing many replicates' peak-α at α=0 (yielding $\Delta_b = 0$). Per-trajectory point estimates and CIs appear in Table 7.

A.13 Multi-seed replication

Each of the 21 main trajectories uses a single adapter training run. To measure adapter-training variance, we retrained the adapter with three further seeds (42, 43, 44; identical data, architecture and hyperparameters) on three trajectories chosen to span the strict-overshoot range: Gemma 2 2B (L13) → Llama 3.2 3B (L12) (high), Llama 3.2 3B (L12) → Gemma 2 2B (L13) (moderate), and Gemma 2 2B (L13) → Qwen 2.5 1.5B (L17) (at the resolution floor). Each of the resulting 12 adapters (original plus three seeds per trajectory) was swept twice with the target's fixed probe: on the simple-statement population of the main sweep (n=1599) and on the full test set including conjunctions and disjunctions (n=2684). Re-sweeping the original adapters reproduces their strict overshoot in Table 7 exactly on two trajectories and within one test item (0.06 pp) on the third.

Strict overshoot. Table 8 gives $\Delta\mathrm{Acc}(\tau^*)$ for every run on both evaluation sets, with 95% bootstrap CIs computed as in Appendix A.12. On simple statements the strict overshoot is positive in all 12 runs ($+0.25$ to $+9.01$ pp), and the 95% CI lower bound exceeds zero in 10 of 12. On the full test set it is positive in 10 of 12 ($+0.00$ to $+5.59$ pp) and the CI excludes zero in 9 of 12. Every run whose CI includes zero belongs to the floor trajectory, where the published estimate already sat at the test-set resolution floor; on the high- and moderate-overshoot trajectories all 16 CIs exclude zero. Two of the ten simple-statement exclusions (floor seeds 42 and 43) come from the earlier checkpoints described below; with the current checkpoints those seeds give $+1.00$ and $+0.00$ pp. Sign is stable where the effect is large; magnitude is not: on the high-overshoot trajectory the four adapters range from $+5.32$ to $+9.01$ pp on simple statements and from $+2.31$ to $+5.59$ pp on the

Table 7: Per-trajectory bootstrap CIs for $\Delta\mathrm{Acc}(\tau^*)$. Stratified paired bootstrap, $B = 10{,}000$, seed 42. Point estimates are computed using the exact optimal threshold per α and may differ from Table 5 by ≤ 0.1 pp due to threshold-grid resolution in the original analysis. Sorted by point estimate descending. Bolded values exceed 1.0 pp. Excl. 0 marks trajectories whose 95% CI lower bound exceeds zero.

SOURCE	TARGET	LAYERS	VARIANT	$\Delta\mathrm{Acc}(\tau^*)$	95% CI	Excl. 0
Gemma 2 2B	Llama 3.2 3B	13→12	Truth	**+9.01**	[+7.25, +10.44]	✓
Llama 3.2 3B	Qwen 2.5 1.5B	12→12	Transfer	**+8.32**	[+6.00, +10.26]	✓
Qwen 2.5 1.5B	Llama 3.2 3B	12→12	Transfer	**+7.19**	[+5.50, +8.57]	✓
Qwen 2.5 1.5B	Llama 3.2 3B	17→12	Truth	**+7.00**	[+5.38, +8.32]	✓
Llama 3.2 3B	Qwen 2.5 1.5B	12→17	Truth	**+3.06**	[+1.88, +4.13]	✓
Qwen 2.5 1.5B	Gemma 2 2B	12→11	Transfer	**+2.88**	[+1.88, +3.75]	✓
Llama 3 8B	Gemma 2 2B	12→13	Truth	**+1.63**	[+0.81, +2.25]	✓
Llama 3.2 3B	Mistral 7B	12→14	Transfer	**+1.38**	[+0.75, +2.25]	✓
Llama 3 8B	Gemma 2 2B	14→14	Transfer	**+1.25**	[+0.44, +1.88]	✓
Llama 3.2 3B	Gemma 2 2B	12→13	Truth	**+1.19**	[+0.50, +1.94]	✓
Qwen 2.5 7B	Mistral 7B	21→13	Truth	**+1.13**	[+0.50, +2.13]	✓
Llama 3 8B	Mistral 7B	14→14	Transfer	+1.00	[+0.44, +1.81]	✓
Llama 3.2 3B	Gemma 2 2B	13→14	Transfer	+1.00	[+0.25, +1.56]	✓
Mistral 7B	Gemma 7B	13→19	Truth	+0.75	[+0.06, +1.50]	✓
Llama 3 8B	Mistral 7B	12→13	Truth	+0.56	[+0.19, +1.38]	✓
Llama 3.2 3B	Mistral 7B	12→13	Truth	+0.50	[+0.13, +1.38]	✓
Gemma 2 2B	Qwen 2.5 1.5B	13→17	Truth	+0.50	[+0.00, +1.56]	
Qwen 2.5 7B	Llama 3 8B	21→12	Truth	+0.31	[+0.06, +0.75]	✓
Qwen 2.5 1.5B	Gemma 2 2B	17→13	Truth	+0.31	[+0.00, +1.00]	
Mistral 7B	Gemma 7B	14→15	Transfer	+0.19	[+0.00, +0.75]	
Qwen 2.5 7B	Gemma 7B	21→19	Truth	+0.19	[+0.00, +1.00]	

full test set, so per-pair magnitudes in Table 5 should be read with a training-variance margin of several percentage points.

Table 8: Strict overshoot $\Delta\mathrm{Acc}(\tau^*)$ (pp) for four independently trained adapters per trajectory, with 95% stratified paired bootstrap CIs (B=10,000). Simple: the main sweep's population (n=1599). Full: all statement types (n=2684). ‡Run on an earlier checkpoint of this seed, since retrained in place; re-scoring the current checkpoint on simple statements gives the value in parentheses (no CI computed).

		Simple (n=1599)		Full (n=2684)	
Trajectory	Adapter	Δ	95% CI	Δ	95% CI
Gemma 2 2B → Llama 3.2 3B	original	+9.01	[+7.32, +10.44]	+5.59	[+4.47, +6.78]
	seed 42	+7.44	[+5.94, +8.76]	+4.43	[+3.43, +5.66]
	seed 43	+5.57	[+4.19, +6.69]	+3.17	[+2.42, +4.32]
	seed 44	+5.32	[+3.94, +6.50]	+2.31	[+1.60, +3.39]
Llama 3.2 3B → Gemma 2 2B	original	+1.19	[+0.50, +1.94]	+2.53	[+1.38, +3.69]
	seed 42	+1.81	[+0.81, +2.94]	+2.38	[+1.27, +3.46]
	seed 43	+1.69	[+0.75, +2.75]	+2.35	[+1.27, +3.65]
	seed 44	+1.81	[+0.88, +2.94]	+1.83	[+0.86, +3.06]
Gemma 2 2B → Qwen 2.5 1.5B	original	+0.56	[+0.00, +1.56]	+2.38	[+1.30, +3.46]
	seed 42	+1.50‡ (+1.00)	[+0.31, +2.31]	+0.19	[+0.00, +0.78]
	seed 43	+0.63‡ (+0.00)	[+0.06, +1.56]	+0.00	[+0.00, +0.22]
	seed 44	+0.25‡ (+0.13)	[+0.00, +0.88]	+0.00	[+0.00, +0.26]

Shape of the sweep. Figure 12 plots the native probe's AUROC against α for all 12 runs. On simple statements, AUROC stays within 0.01 of its native value, or above it, through α=0.7 in every run, and the four adapters of a trajectory agree to within 0.007 at every $\alpha \leq 0.7$; all 12 runs then fall sharply by α=1 (to 0.62–0.79). The position of the turn-down depends on the trajectory rather than the seed: on Llama 3.2 3B → Gemma 2 2B and on the floor trajectory, AUROC turns

down within $\alpha \in [0.8, 0.9]$; on Gemma 2 2B → Llama 3.2 3B it rises above native through α=0.9 and falls only at α=1. On the full test set the same holds in 9 of 12 runs. The exceptions are the three retrained seeds of the floor trajectory, which fall more than 0.01 below native by α=0.6–0.7 while the original adapter holds to α=0.8; on that trajectory the seeds differ by up to 0.027 AUROC within the plateau.

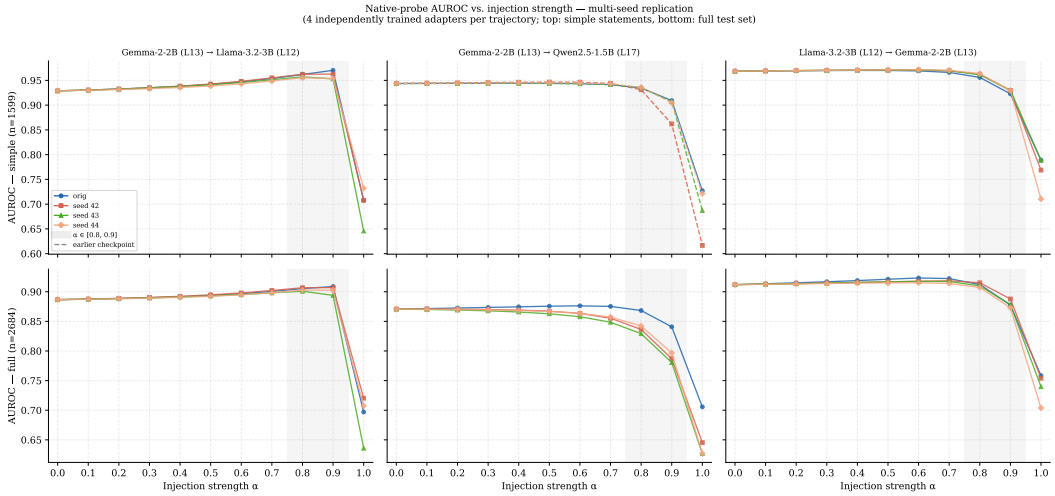


Figure 12: Native-probe AUROC against injection strength for four independently trained adapters (original and seeds 42–44) on each of three trajectories spanning the overshoot range. Top: simple statements (n=1599); bottom: full test set (n=2684). Shading marks $\alpha \in [0.8, 0.9]$. Dashed lines: the floor trajectory's seeds on earlier checkpoints, later retrained in place. Through α=0.7, seeds agree to within 0.007 AUROC on simple statements; on the full test set the floor trajectory's three retrained seeds decline from $\alpha \approx 0.6$.

Checkpoint provenance on the floor trajectory. The three simple-statement runs of seeds 42–44 on Gemma 2 2B → Qwen 2.5 1.5B used adapter checkpoints that were later retrained in place with the same hyperparameters. Their recorded training metrics differ from the checkpoints now served (FVU ≈ 0.038 and adapter–target cosine ≈ 0.69, against ≈ 0.005 and ≈ 0.72); the full-test-set runs, the activation dumps used in Appendices A.14 and A.16, and the released adapters all use the later checkpoints. The other two trajectories used identical checkpoints in both sweeps. For these three runs, differences between the simple and full columns of Table 8 mix the evaluation set with the weights. Re-scored with the current checkpoints, simple-statement strict overshoot is positive in 11 of 12 runs rather than 12. The released code checks every saved run against the adapter repository it names and reports such mismatches.

Scope. Replication covers three of 21 trajectories, chosen to span the overshoot range rather than at random. The regression coefficients in Table 4 and the per-pair values for the other 18 trajectories remain single-seed estimates.

A.14 Baseline controls and a random-map null

The strict overshoot of Section 4.4 could arise from any perturbation of the target's activations rather than from a learned alignment. We test this on the three multi-seed trajectories with two controls built from the same inputs as the trained adapter.

Controls. The *random-weight* control is an MLP with the trained adapter's architecture and dimensions and randomly initialized weights, applied to the same source activations, with its output rescaled so that its mean ℓ_2 norm matches that of the native target activations over the full test set. It keeps the per-example structure of h_S but has no learned alignment. The *shuffled* control takes the original trained adapter's actual outputs and permutes them across examples (a fixed permutation

per trajectory, seed 42), which preserves their marginal distribution exactly and removes every per-example correspondence. Both are injected and swept exactly as the trained adapter (Equation 1).

Single-draw results. Table 9 reports both controls on the full test set (n=2684). The shuffled control produces no strict overshoot (0.00, 0.00 and +0.04 pp) and drives AUROC at α=1 to chance (0.49–0.51): the adapter's output distribution alone does not produce the effect, and per-example source information is necessary. The random-weight row, however, is a single draw made without a fixed seed, so it cannot be regenerated, and one draw is not a reliable estimate: the same draw scored on simple statements gives +0.00, +1.06 and +0.06 pp, and a fresh unseeded draw gave +2.05, +0.34 and +0.75 pp on the full test set. We therefore replace it with a seeded null.

Table 9: Baseline controls on the full test set (n=2684), one draw per control. Δ_{τ_0} (pp): peak default-threshold accuracy minus native. Δ_{τ^*} (pp): strict overshoot $\Delta\mathrm{Acc}(\tau^*)$. The random-weight row is a single unseeded draw and is superseded by the seeded null in Table 10.

Trajectory	Map	Δ_{τ_0}	Δ_{τ^*}	AUROC$_{\alpha=0}$	AUROC$_{\alpha=1}$
Gemma 2 2B → Llama 3.2 3B	trained (orig.)	0.00	+5.59	0.887	0.697
	random-weight	+0.67	0.00	0.887	0.456
	shuffled	0.00	0.00	0.887	0.512
Llama 3.2 3B → Gemma 2 2B	trained (orig.)	+0.60	+2.53	0.912	0.758
	random-weight	0.00	0.00	0.912	0.456
	shuffled	0.00	0.00	0.912	0.503
Gemma 2 2B → Qwen 2.5 1.5B	trained (orig.)	0.00	+2.38	0.871	0.706
	random-weight	0.00	0.00	0.871	0.555
	shuffled	0.00	+0.04	0.871	0.493

Seeded random-map null. For each trajectory we draw 1000 random-weight maps, constructed as above with the random seed set to $0, \ldots, 999$, and score each on both evaluation sets from the saved activation dumps, so that trained and random maps are computed identically. For each trained adapter we report a one-sided empirical p-value, $p = (1 + \#\{\text{null} \geq \text{trained}\})/(1 + 1000)$, whose smallest attainable value is 0.001, and Holm-correct the original adapter's p across the three trajectories (Table 10).

Table 10: Strict overshoot (pp) of 1000 seeded random-weight maps per trajectory against the trained adapters. Null: median, 95th percentile and share of draws with positive strict overshoot. p: one-sided empirical p-value of the original adapter; Holm: corrected across the three trajectories. Seeds: number of the three retrained seeds with $p < 0.05$ (uncorrected). The original adapter's Δ is the published value of Table 7; on the floor trajectory the re-sweep in Table 8 gives +0.56 on simple statements, one test item higher. ‡Floor-trajectory seeds are scored on their current checkpoints (Appendix A.13).

		Null			Original adapter			
Trajectory	Eval	median	p95	> 0	Δ	p	Holm	Seeds
Gemma 2 2B → Llama 3.2 3B	simple	+0.13	+5.82	55%	+9.01	0.003	0.009	1/3
	full	+0.04	+3.58	51%	+5.59	0.009	0.027	1/3
Llama 3.2 3B → Gemma 2 2B	simple	+0.13	+2.13	57%	+1.19	0.161	0.322	0/3
	full	0.00	+2.01	46%	+2.53	0.021	0.042	2/3
Gemma 2 2B → Qwen 2.5 1.5B	simple	+0.19	+1.81	65%	+0.50	0.330	0.330	0/3‡
	full	+0.30	+2.61	68%	+2.38	0.066	0.066	0/3

What the null shows. Random maps of the adapter's architecture (Figure 13) produce positive strict overshoot in 46–68% of draws, with 95th percentiles of 1.8–5.8 pp. A positive strict overshoot, or a bootstrap CI that excludes zero, therefore does not by itself indicate learned transfer: both test against zero, not against untrained maps. Against the null, the original trained adapter is significant after Holm correction on Gemma 2 2B → Llama 3.2 3B on both evaluation sets (0.009, 0.027) and on Llama 3.2 3B → Gemma 2 2B on the full test set only (0.042; 0.322 on simple statements); it is

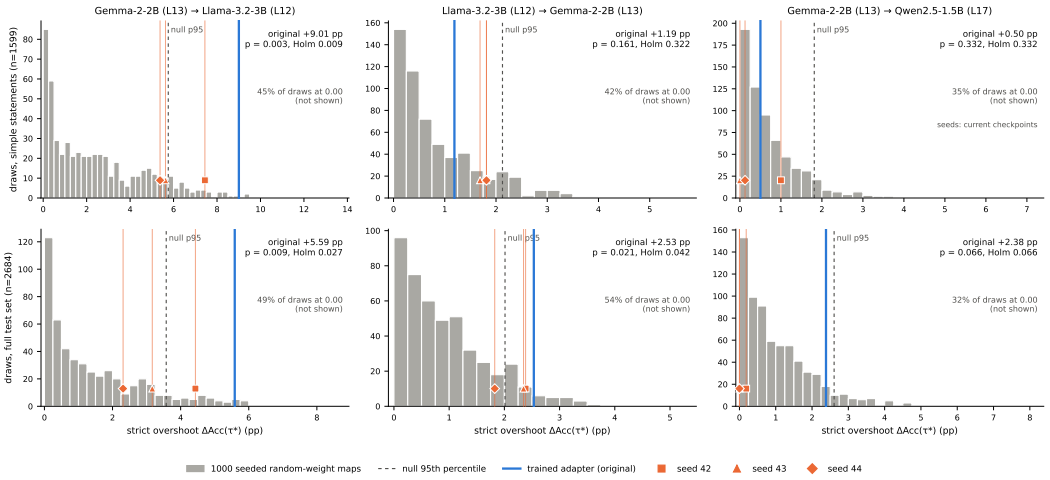


Figure 13: Strict overshoot of 1000 seeded random-weight maps per trajectory (grey; draws at exactly 0.00 pp are counted in the panel text, not drawn) against the original trained adapter (blue, with its one-sided empirical p and Holm-corrected p) and the three retrained seeds (orange). Dashed: the null's 95th percentile. Top: simple statements (n=1599); bottom: full test set (n=2684). The floor trajectory's seeds are scored on their current checkpoints.

not on the floor trajectory (0.33, 0.066). Across all 12 trained runs, 5 of 12 clear the null at $p < 0.05$ on the full test set and 2 of 12 on simple statements, uncorrected across the 24 tests.

Together with the shuffled control, this supports a narrower reading than "strict overshoot is specific to learned alignment": per-example source information is necessary, a large share of the effect follows from injecting it at all, even through a random map of the same architecture, and training adds beyond random maps on the stronger trajectories (both of them on the full test set, only the strongest on simple statements). This agrees with Section 4.5, where a closed-form map matches or exceeds the trained adapter at peak on 18 of 21 trajectories. At full injection the controls behave differently from the trained adapter: random-weight and shuffled maps drive the native probe's AUROC to 0.46–0.56, against 0.70–0.76 for the trained adapter.

Scope. The null covers one baseline family (random MLPs of the adapter's architecture) on three trajectories. Random linear maps and other initialization schemes were not tested, and the other 18 trajectories have no random-map comparison, so their bootstrap results in Appendix A.12 are tests against zero only.

A.15 Statement-type decomposition

The main sweep scores simple statements of both polarities (Section 3.1). To test whether its findings extend to compound statements, and whether they hold equally for both polarities, we split the full test set of the multi-seed runs (n=2684; Appendix A.13) into four disjoint groups: affirmed simple statements (n=1080), negated simple statements (n=519), conjunctions (n=607) and disjunctions (n=478). All compound statements are affirmed, and the first two groups together form the main sweep's population. For each of the 12 trained adapters on the three multi-seed trajectories we compute the target probe's AUROC within each group at every α, with the probe unchanged. Since neither the probe nor the native activations depend on the adapter, native AUROC in a group is identical across a trajectory's four adapters. Table 11 summarizes each trajectory over its four adapters and Figure 14 shows the full sweeps. Per-run values and 95% percentile bootstrap CIs (B=1000, statements resampled within each group) at $\alpha \in \{0, 0.7, 1\}$ are released with the code.

Native separability. At α=0 the target probe separates affirmed simple statements, negated statements and conjunctions well (AUROC 0.85–0.97) but disjunctions poorly (0.57–0.70). The probes are trained on simple statements only (Section 3.4), so disjunctions lie outside their training distri-

bution. Because the disjunction group starts close to chance, we do not interpret its changes under injection.

The plateau holds for every statement type. Within each group, the lowest AUROC over $\alpha \in [0.1, 0.7]$ stays within 0.02 of the group's native value, or above it, in 47 of the 48 run–group combinations. The exception is one disjunction run (Gemma 2 2B $\rightarrow$ Llama 3.2 3B, seed 43), 0.025 below native. Partial injection therefore preserves rank order for compound statements and for both polarities, not only for the population of the main sweep.

The loss at full injection depends on statement type. Of the three well-separated groups, conjunctions lose the least at α=1 on all three trajectories (mean AUROC drops of 0.13, 0.03 and 0.09). Negated statements lose the most on two. On Llama 3.2 3B $\rightarrow$ Gemma 2 2B they fall from 0.97 to 0.49 (range 0.45–0.52 across the four adapters, and every adapter's 95% CI contains 0.5), while affirmed simple statements on the same trajectory keep 0.89. On Gemma 2 2B $\rightarrow$ Qwen 2.5 1.5B the margin is narrow (0.30 against 0.27 for affirmed simple statements). On Gemma 2 2B $\rightarrow$ Llama 3.2 3B the order reverses: affirmed simple statements lose more ($0.92 \rightarrow 0.70$) than negated ones ($0.97 \rightarrow 0.78$). Negated statements also turn down earlier. On the two trajectories where they lose the most, their mean AUROC first falls more than 0.02 below native at α=0.8, against α=0.9–1.0 for the other well-separated groups.

Aggregates can hide a failure on one form. On Llama 3.2 3B $\rightarrow$ Gemma 2 2B, the all-statement AUROC at α=1 (0.74) averages affirmed simple statements that remain well separated with negated statements at chance. A single aggregate AUROC at full injection can therefore understate how completely a learned adapter fails on one statement form. The loss is not explained by the polarity direction alone. Trained-adapter outputs keep at most 12% of the native separation between affirmed and negated statements along the probe's polarity direction t_P on every trajectory (Appendix A.16), and the most of it (6–12%) on Llama 3.2 3B $\rightarrow$ Gemma 2 2B, the trajectory where negated statements fail completely.

Table 11: Native-probe AUROC by statement type on the full test set, for the four trained adapters (original and seeds 42–44) of each multi-seed trajectory. α=0: native AUROC (identical across adapters). $\min_{\alpha \le 0.7}$: lowest AUROC over $\alpha \in [0.1, 0.7]$, averaged over adapters. α=1: mean and range across adapters. Δ: mean change from α=0 to α=1. Disjunctions are weakly separable natively and are not interpreted.

			AUROC				
Trajectory	Statements	n	α=0	$\min_{\alpha \le 0.7}$	α=1	range	Δ
Gemma 2 2B $\rightarrow$ Llama 3.2 3B	Affirmed simple	1080	0.92	0.92	0.70	[0.63, 0.75]	−0.22
	Negated simple	519	0.97	0.97	0.78	[0.70, 0.82]	−0.19
	Conjunctions	607	0.96	0.96	0.83	[0.78, 0.88]	−0.13
	Disjunctions	478	0.70	0.69	0.56	[0.50, 0.63]	−0.14
	All statements	2684	0.89	0.89	0.69	[0.64, 0.72]	−0.20
Llama 3.2 3B $\rightarrow$ Gemma 2 2B	Affirmed simple	1080	0.97	0.97	0.89	[0.85, 0.92]	−0.08
	Negated simple	519	0.97	0.96	0.49	[0.45, 0.52]	−0.48
	Conjunctions	607	0.92	0.92	0.89	[0.85, 0.93]	−0.03
	Disjunctions	478	0.66	0.65	0.62	[0.61, 0.63]	−0.04
	All statements	2684	0.91	0.91	0.74	[0.70, 0.76]	−0.17
Gemma 2 2B $\rightarrow$ Qwen 2.5 1.5B	Affirmed simple	1080	0.95	0.94	0.68	[0.60, 0.80]	−0.27
	Negated simple	519	0.94	0.93	0.65	[0.58, 0.67]	−0.30
	Conjunctions	607	0.85	0.85	0.76	[0.73, 0.78]	−0.09
	Disjunctions	478	0.57	0.57	0.58	[0.54, 0.62]	+0.01
	All statements	2684	0.87	0.86	0.65	[0.63, 0.71]	−0.22

Scope. The decomposition covers three trajectories and the trained adapters only. The other 18 trajectories and the Procrustes map were not scored by statement type, and strict overshoot was not computed per type. Group sizes are unequal, so per-group AUROCs carry different sampling error (95% CI half-widths at α=1 of about 0.02–0.03 for affirmed simple statements and 0.04–0.05 for negated ones).

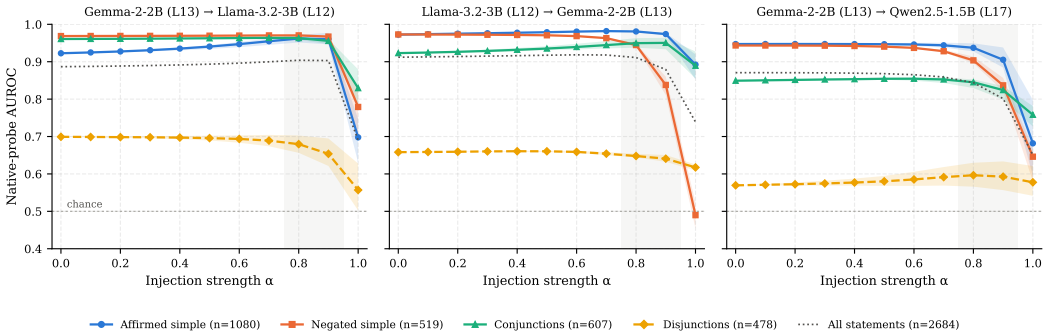


Figure 14: Native-probe AUROC against injection strength by statement type, on the full test set. Lines: mean over the four trained adapters of each trajectory; bands: their range. Dotted: all statements. Shading marks $\alpha \in [0.8, 0.9]$. The plateau holds for every type; at α=1, negated statements fall to chance on Llama 3.2 3B → Gemma 2 2B while affirmed simple statements and conjunctions on the same trajectory keep AUROC near 0.89. Disjunctions (dashed) start near chance.

A.16 Residual geometry at full injection

Section 4.2 shows that the target's native probe loses most of its separability on trained-adapter output at α=1. This can happen in two ways: the adapter discards truth information, or it keeps the information but not where the native probe reads it. We separate the two on the three multi-seed trajectories, using activation dumps of the full test set (n=2684) that store the native target activations h_T, the outputs of the four trained adapters, the random-weight and shuffled controls of Appendix A.14, and the target probe's truth and polarity directions t_G and t_P (Section 3.4). Every quantity below is computed on map outputs alone, which is the input at α=1.

Measures. For a set of activations X, let the *truth gap* be the difference between the class means of true and false statements projected on the unit vector t_G, and the *polarity gap* the difference between the means of affirmed and negated statements projected on t_P. We report each gap for a map's output as a ratio to the same gap on native activations, so 1 means the separation the probe reads is fully preserved and 0 means it is gone. To test whether the information is still present, we split the statements into label-stratified halves, fit a fresh logistic-regression probe (C=10^6) on one half of a map's outputs, and report its AUROC on the other half (*re-probe*). For comparison, t_G *read* is the AUROC of the projection on the native t_G over the same held-out half. Table 12 gives all four measures.

Truth information survives; its alignment does not. Trained adapters keep only 2–11% of the native truth gap along t_G, yet a probe retrained on the same outputs recovers AUROC 0.84–0.95, against 0.95–0.98 on native activations. The re-probe exceeds the t_G read on every trained map, by at least 0.17 (Figure 16). The information the native probe needs is therefore still present at α=1, but it no longer lies along the direction the probe reads. The t_G read (0.59–0.76) is higher than the small gap ratios alone would suggest because the spread of the outputs along t_G shrinks along with the gap, which is the parametric collapse of Section 4.2 seen from the activation side. The polarity gap along t_P is lost in the same way (-0.01 to 0.12 of native; Figure 15).

Decodability is a weak criterion; alignment is the informative one. Decodability does not separate trained from untrained maps: the random map re-probes at 0.97–0.98, higher than every trained adapter. A random map of h_S preserves most of the source's information, and the source's own activations encode truth, so a fresh probe can recover it from any such map. The shuffled control, which breaks the per-example correspondence, re-probes at chance (0.50–0.54). What distinguishes the trained adapters is alignment with the target's truth direction. On Gemma 2 2B → Llama 3.2 3B and Llama 3.2 3B → Gemma 2 2B, every trained adapter keeps more of the truth gap (0.022–0.109) and reads higher on t_G (0.60–0.76) than the random map (-0.002 and -0.008; 0.49 and 0.46). On Gemma 2 2B → Qwen 2.5 1.5B the random map falls inside the trained range on both measures

Table 12: Residual geometry of map outputs (the input at α=1) on the full test set (n=2684). Gap ratios: class-mean separation along the target probe's truth direction t_G (true vs. false) and polarity direction t_P (affirmed vs. negated), relative to native activations. t_G read: AUROC of the projection on t_G. Re-probe: held-out AUROC of a probe trained on half of the map's outputs. Trained: range over the four adapters (original and seeds 42–44). Random-weight: the single unseeded draw of Table 9. Re-probe values vary by up to 0.006 across numerical environments and are given to two decimals.

		Gap ratio		AUROC	
Trajectory	Map	t_G	t_P	t_G read	re-probe
Gemma 2 2B → Llama 3.2 3B	native	1	1	0.87	0.97
	trained (4)	0.022–0.035	−0.010–0.009	0.60–0.69	0.84–0.90
	random-weight	−0.002	0.018	0.49	0.97
	shuffled	0.002	0.000	0.52	0.52
Llama 3.2 3B → Gemma 2 2B	native	1	1	0.91	0.98
	trained (4)	0.069–0.109	0.055–0.117	0.73–0.76	0.94–0.95
	random-weight	−0.008	0.098	0.46	0.98
	shuffled	0.001	−0.011	0.53	0.54
Gemma 2 2B → Qwen 2.5 1.5B	native	1	1	0.81	0.95
	trained (4)	0.045–0.073	0.031–0.075	0.59–0.68	0.93
	random-weight	0.058	0.018	0.61	0.97
	shuffled	−0.002	0.002	0.49	0.50

(0.058; 0.61). This is the same split as the random-map null for strict overshoot (Appendix A.14): training adds measurably beyond a random map on the two stronger trajectories and not on the floor trajectory. Here, however, the comparison is against a single random draw per trajectory.

Relation to the Procrustes result. The rigid Procrustes map keeps median native-probe AUROC 0.91 at α=1 (Section 4.5), so on most pairs the native probe still reads Procrustes output where it largely fails on trained-adapter output. Together with the gap ratios above, this suggests the α=1 collapse lies in how the learned adapters place truth-relevant structure relative to the native probe rather than in a loss of the information itself. That AUROC comes from the full TTPD classifier, which combines t_G and t_P, so it does not by itself show that the map aligns either direction. We did not dump Procrustes outputs, so its gap ratios, re-probe AUROC and direction alignment were not measured directly.

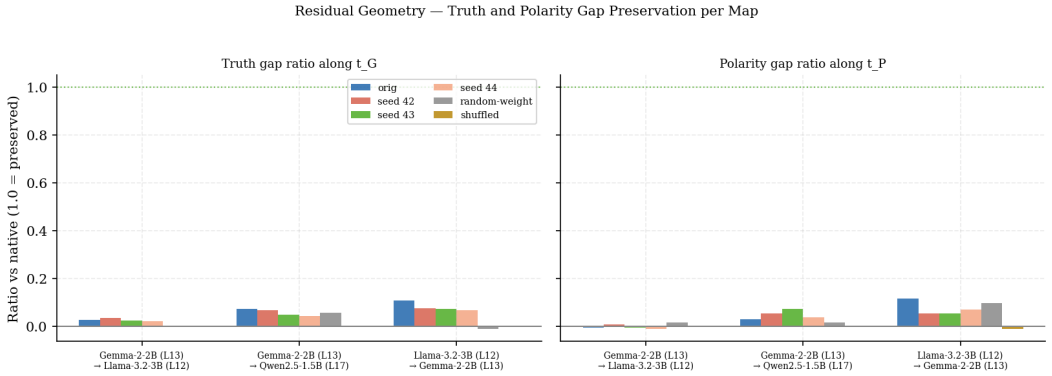


Figure 15: Class-mean separation preserved by each map along the target probe's truth direction t_G (left) and polarity direction t_P (right), relative to native activations (1.0, dotted). Trained adapters keep 2–11% of the truth gap; the shuffled control keeps none, and the single random-weight draw keeps none on two trajectories.

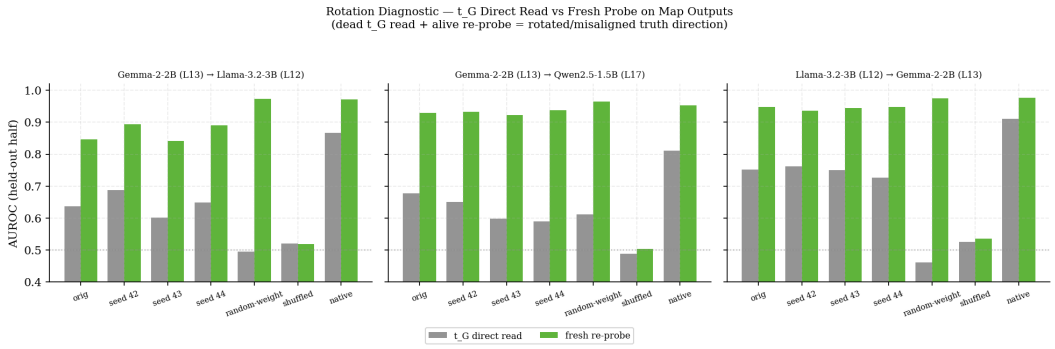


Figure 16: AUROC of the native truth direction t_G read directly (grey) against a probe retrained on the map's outputs (green), on a held-out half of the full test set. A low direct read with a high re-probe means the truth information is present but misaligned with t_G. The random map re-probes highest, so recoverability alone does not show that a map was learned.

Scope. The analysis covers three trajectories, at $\alpha{=}1$ only. The random-weight row is one unseeded draw, so the trained-versus-random comparison of gap ratios has no null distribution. The residual anisotropy along t_G (the variance of $f_\theta(h_S) - h_T$ along t_G relative to its mean per-dimension variance) is released with Table 12 but not interpreted. The shuffled control comes within 11% of the trained adapters' mean on every trajectory (32–108 against 34–113), so the measure reflects the native activations' own spread along t_G more than anything specific to a learned map.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 lists specific findings tied to numerical results in Section 4 and Table 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses adapter-training variance (single seed on 18 of 21 trajectories; multi-seed replication on three), the scope of the random-map null, statement-type coverage, family imbalance, the n=21 cohort size, and the restriction to truth.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our paper does not contain any theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full procedures hyperparameters, model identifiers, dataset sources, and selection criteria are specified in Appendix A in sufficient detail to permit reimplementation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is released at `https://github.com/KingTechnician/shared-truth`, and per-run results, activation dumps and analysis tables at `https://huggingface.co/datasets/KingTechnician/shared-truth-results`; the trained adapters are on Hugging Face. The notebooks regenerate every table and figure from the saved results and check them against the published values.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Appendix A.1-A.4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Strict overshoot carries 95% stratified paired bootstrap CIs (B=10,000; Appendix A.12); trained adapters are tested against 1000 seeded random maps with one-sided empirical p-values and Holm correction (Appendix A.14); training variance is measured with three additional seeds on three trajectories (Appendix A.13). Aggregates use median and IQR over 21 trajectories, and regressions report standard errors and p-values.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We do not report per-run compute. No language model is trained or fine-tuned: the adapters are two-layer MLPs and the probes are logistic regressions, both inexpensive to train. The only GPU-bound step is activation extraction from open-weight models of 8B parameters or fewer, whose memory requirements follow from each model's published size.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics; the work involves no human subjects, no crowdsourced data, and uses publicly available pretrained models within their licensed terms of use.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5 (Discussion) outlines positive impacts: heterogeneous truth-probe pipelines may benefit from understanding the dual-mechanism, and downstream pipelines should cross-reference reported accuracy gains against optimal threshold ceilings on native activations. Potential negative impacts are limited: our work is foundational research on representational geometry of pretrained models, with no direct path to deployment; the released artifacts (code, results, and small adapters between existing open-weight models) add no generation capability beyond those models. We do not develop new capabilities for harmful generation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We release analysis code, results, and small adapters that map hidden states between existing open-weight models; none is a generative model or adds capability beyond the public models it connects. The analysis uses existing open-source contrastive truth datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The seven evaluated instruct models are cited in Section 3.2 and used under their licenses: LLaMA 3 8B and LLaMA 3.2 3B [Grattafiori et al., 2024] under the Meta Llama 3 and Llama 3.2 Community Licenses; Gemma 7B and Gemma 2 2B [Gemma Team, 2024a,b] under the Gemma Terms of Use; Mistral 7B v0.3 [Jiang et al., 2023] and Qwen 2.5 1.5B and 7B [Qwen Team, 2024] under Apache 2.0. The truth probe contrastive datasets (`cities`, `neg_cities`, `sp_en_trans`, `neg_sp_en_trans`, `inventors`, `animal_class`, `element_symb`, `facts`) follow Marks and Tegmark [2024], Bürger et al. [2024] and are used under the terms of the authors' repositories. The adapter training corpus is the Anthropic HH-RLHF slice [Bai et al., 2022] (MIT license) as prepared by Oozeer et al. [2025].

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the code, the per-run results dataset, and the trained adapters, documented by the repository README and the dataset card (evaluation populations and known caveats such as the floor-trajectory checkpoint provenance). The code and adapters are released under the MIT license and the results dataset under CC BY 4.0; use of the underlying models remains governed by their own licenses.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]The work does not involve crowdsourcing or research with human subjects.

Justification: [N/A]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects, so no IRB approval was required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the methods studied.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.