A TRANSPORT LAW FOR LINEAR DIRECTIONS BETWEEN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Linear Representation Hypothesis argues that a concept corresponds to a direction in a model's representation space, while the Platonic Representation Hypothesis posits that independently trained models converge to a shared structure of representations. A key unexplored question is whether these two ideas imply that a behaviorally meaningful direction learned in one model should be recoverable and transferable in another. We address this question in a deliberately narrow setting: the **direction of a successful attack** within a single modality, text decoders, and the linear case, where both read-out and transfer are linear. Under class imbalance and guard-dependent labels, we first derive an **admissibility criterion** for statistics used to measure transfer of the attack space. From these admissible statistics, we derive a **transport law**: the correlation between a probe's score in its source model and its carried score in the target model equals the root-mean-square canonical correlation of the model pair, weighted across directions by the probe's profile. We further prove an exact decomposition of transport into within-class and label transport. The decomposition reveals a striking asymmetry, where labels transport without loss, and on some pairs with a gain ($\tau = 0.904$–1.370), whereas the underlying direction is only partially recovered (0.740–0.865). We validate the transport law across 11 dense transformers ranging from 0.5B to 70B parameters, yielding 110 ordered model pairs. Their responses are labeled by ten diverse guards using a corpus of 3306 prompts, with representations compared at fourteen relative depths. We further show that the law holds beyond the classical architecture, on 8 alternative-architecture models ranging from 1.17B to 12.31B parameters. We find that transfer is asymmetric in model size, where a probe transferred from a smaller model to a larger model loses almost no AUC, whereas transfer in the reverse direction incurs measurable loss that increases with the size gap. More broadly, our work turns two influential representation hypotheses into testable, quantitative questions about cross-model representation transport, with direct implications for understanding and controlling model behavior in AI safety.

1 INTRODUCTION

The Linear Representation Hypothesis states that a concept corresponds to a direction in a model's representation space (Park et al., 2024): along such a direction the concept can be read and controlled(Marks & Tegmark, 2024; Nanda et al., 2023; Li et al., 2023; Turner et al., 2024; Rimsky et al., 2024). The Platonic Representation Hypothesis states, on the other hand, that independently trained models converge to a structure of representations shared by the models (Huh et al., 2024). Under calibration against a permutation null, however, part of the previously claimed global convergence disappears (Gröger et al., 2026).

Central question. *Do these two hypotheses imply that a behaviorally meaningful linear direction learned in one model is recovered in, and transfers to, another?* This question has a theoretical and an applied motivation. Arditi et al. (2024) showed that refusal is linearly structured, where it is mediated by a single linear direction inside a model, and the harmfulness of a request and its refusal are encoded separately (Zhao et al., 2025b). Linear probes on activations can be applied as computationally cheap filters on a language model and serve as one of the steps of filtering the request-response pair when monitoring (McKenzie et al., 2025; Kramár et al., 2026) a language model. If the direction is shared, the probe, as a monitoring tool, transfers together with it between models. The transfer-

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

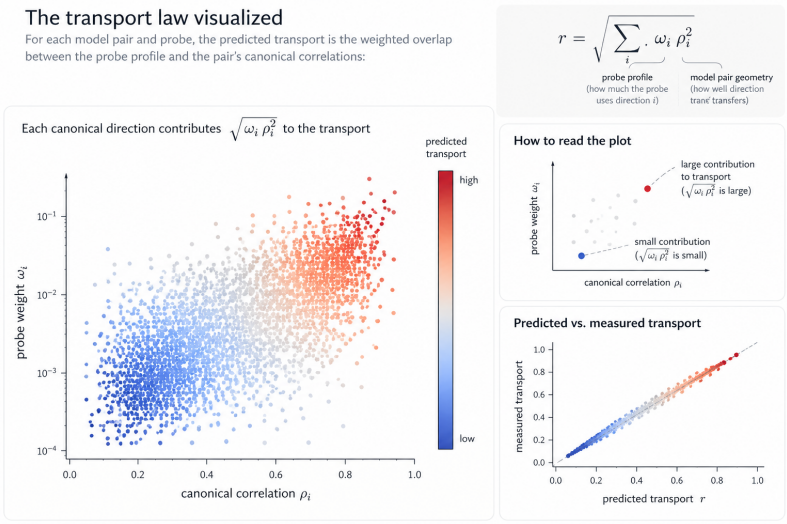

Figure 1: **The Transport law on the canonical directions.** A: every point is one canonical direction i of one "ordered model pair $\times$ guard" cell of W_{mix19} at $k = 64$ and relative depth $p = 0.75$, placed at its canonical correlation ρ_i and at its weight ω_i in the profile of the probe (logarithmic axis), and colored by its contribution $\omega_i \rho_i^2$ to the square of the predicted transport $\sqrt{\sum_i \omega_i \rho_i^2}$ (Theorem 4). The inset shows how to read the panel; a direction contributes much only when it is both well aligned between the two models, large ρ_i, and much used by the probe, large ω_i. B: for each cell, the transport measured through the ridge stitch against the transport predicted from the spectrum and the profile alone, before any map is fitted; the dashed line is the diagonal (§4.2, Appendix L.6). Read: the transport of a probe is decided by where its weight sits along the spectrum of the pair, and the value predicted from the pair alone is the value measured.

ability hypothesis itself has been confirmed in various cases, where Zou et al. (2023); Huang et al. (2025a); Li et al. (2024) showed that a suffix optimized against one model breaks models it was not trained on, Oozeer et al. (2025) showed that steering vectors transfer between model families through a learned activation mapping, and Wang et al. (2025) showed that the refusal vector transfers between languages of a single model. However, the linearity of the transfer is not the main object of study in these works, and the question of whether linearity suffices and what *exactly* transfers remains open.

We answer this question in a narrow setting, considering the direction of a *successful attack*, *i.e.,* of prompts on which the model produced an answer judged unsafe by a guard, where both the probe and the transfer are linear. Narrowing to the linear case on the one hand simplifies the task, speeding up computation, and on the other hand imposes a constraint on the interpretation of metrics. In this work, we address this using four concrete tasks: **Task 1 (§2).** Determine what may be counted as evidence of transfer under conditions where both the probe and the transfer are linear. We derive an admissibility criterion for representation comparison statistics and show which of the tools in use fail it, *e.g.,* linear CKA on raw features is invariant to a smaller group of transformations, the R^2 of the stitch depends on the coordinates of the target model, and the AUC transfer gap does not distinguish orbits; **Task 2 (§2).** Express the transfer of the source model's label direction through a quantity that can be computed for a model pair in advance, without performing the transfer itself. We prove the transport law (the correlation between the probe's score in its source model and its carried score equals the root-mean-square canonical correlation of the model pair, weighted across directions by this probe's profile) and two of its corollaries (the leading canonical correlation bounds the transfer of any direction, and two probes in one model pair transfer differently exactly as much as their profiles differ); **Task 3 (§2, §4).** Answers what happens to the label, and what to the direction. The transport law describes the fate of the direction but not of the label: it predicts the recovery of the score variance, and for a label with a small share of prompts labeled 1 (the successful-attack label is exactly such by construction) almost all of the score variance is within-class. We prove an exact decomposition of transport into within-class transport and label transport and measure both parts on the subset of prompts with a successful attack. and **Task 4 (§4).** Here, we show under which conditions

the result remains valid and what the transfer analysis in our setting is in principle able to detect. The first we test with ablations, varying the number of principal components k, the training-row budget, the prompt distribution and the model architecture. The second we mix into both models a shared direction of a given strength and run the same pipeline on the modified data. The strength from which the pipeline finds it is the resolution of the setup: below it, "*we found nothing*" does not mean "*it is not there*" (§4.3, App. H). There is nothing attack-specific in the shape of the spread and the specificity is in the difference of class means, which the second-moment statistic cannot see (App. H).

Related work. Linear probes read behavioral properties from hidden states within one model (Mallen & Belrose, 2024; Bullwinkel et al., 2025; Wehner & Fritz, 2025; Sim & Park, 2026; Hewitt & Liang, 2019), and directions are carried between models by learned or closed-form maps (Lenc & Vedaldi, 2015; Bansal et al., 2021; Csiszárik et al., 2021; Moschella et al., 2023; Maiorca et al., 2023; Jha et al., 2025; Armstrong et al., 2026; Huang et al., 2025b). The full discussion, with the two representation hypotheses and the similarity indices, is in App. A.

Our work. Against Gröger et al. (2026): they find agreement local in *geometry*; we ask about agreement local in the *population* of prompts, and we select each statistic by its invariance group in advance. Against Huang et al. (2025b): they demonstrate weak-to-strong transfer on one pair; we compare both directions of each pair and measure how the difference grows with the size ratio. Against Maiorca et al. (2023): they establish that a linear or orthogonal map suffices; we ask how much of a direction survives the map, and predict that amount from the pair before any map is fitted. Against the probing literature: we treat the disagreement of the guards' verdicts as a quantity to be measured, and we report what the design could have found.

2 Admissible Statistics and the Transport Law

Setup. Let two models read the same set of N inputs, and let their representations, reduced to a common number k of coordinates, form row-paired matrices $Z_A, Z_B \in \mathbb{R}^{N\times k}$. Row i of both is the same input. A *direction* in model A is a vector $w \in \mathbb{R}^k$. It ranks the inputs by the *score* $Z_A w$; together with a threshold it becomes a decision boundary, a hyperplane normal to w, but everything we measure depends on the ranking alone. In general w is arbitrary, and every statement below holds for an arbitrary w. The directions we care about are obtained by fitting a separating hyperplane between the two classes of a binary label, and it is fitting that makes the profile of w (Theorem 4) far from uniform. The direction is carried to model B by a *stitch*: a linear map W, fitted on the same paired rows without any label, from the space of B into the space of A. The probe thus stays where it was trained and the data is brought to it, which gives the *carried score* $(Z_B W)w$ (the same w, applied to the representation of the second model after the map). Finally, let $y \in \{0,1\}^N$ be a input binary label. It is absent from the transport law and enters only where class separation is compared. Which label that is, and what changes when the label is itself model-dependent, belongs to the special case (§3). In this notation the question "*does the direction transfer*" can be decomposed into two components: i) whether the score itself is recovered, *i.e.,* how similar $(Z_B W)w$ is to $Z_A w$; and ii) whether the carried score continues to separate the classes of the label y, and how well relative to the original.

Limitation 1: Identifiability class. In the chosen notation both operations (recovering the score and separating the classes) do not distinguish a representation from its image under an invertible change of coordinates. Formally, for any invertible matrix T and for any direction w we have $w^\top(Tx) = (w^\top T)x = (T^\top w)^\top x$. This means that what the probe w computes on the transformed representation, the same probe with weights $T^\top w$ computes on the original one and conversely, every function available on the original one is available on the transformed one too — with weights $T^{-\top} w$. Both representations thus generate exactly the same set of scores.

A stronger result has been proved for models whose output is the inner product of the representation with a vector (Roeder et al., 2020). The original proof requires equality of the conditional distributions of the two models, which does not hold here because the models are trained independently. But the theorem shows that the representation in a model is defined only up to an invertible change of coordinates. If $f_B = T f_A$ with $T \in \mathrm{GL}(D)$, then f_A and f_B are one and the same representation, written differently. The theorem concerns the full width D; at our reduced width k we take GL as the class not because of it, but because the probe and the stitch are themselves linear (App. B.1). A change of coordinates acts on a pair of representations independently on each side: $(Z_A, Z_B) \mapsto (QZ_AM_A,\ QZ_BM_B)$ with invertible M_A, M_B and orthogonal Q acting on the rows

of both. The set of all pairs obtained from one another by such a change we call the *orbit* of the pair. Pairs from one orbit are one and the same pair of representations, written in different coordinates.

Main point: A statistic whose value changes under such a rewriting reports in which coordinates the models are written. But the choice of coordinates changes no score and is not visible to the class of linear functions of the representation, so such statistics say nothing about the models themselves. Next, we define the **admissible statistic** criterion as highlighted by Limitation 1.

Definition 1 (Admissible statistic)**.** *A quantity may count as evidence of transfer only if it is constant on the orbit, i.e., the same for (Z_A, Z_B) and (QZ_AM_A, QZ_BM_B) for any invertible M_A, M_B and orthogonal Q, and at the same time distinguishes different orbits.*

A natural example of an admissible statistic is the canonical correlations $\rho_1 \geq \cdots \geq \rho_k$ of a pair of row-paired representations.

Proposition 2 (Completeness of the canonical spectrum)**.** *For representation matrices $Z_A, Z_B \in \mathbb{R}^{N\times k}$ on the same N prompts and any $M_A, M_B \in \mathrm{GL}(k)$, the canonical spectrum $\{\rho_i\}$ of the pair (Z_AM_A, Z_BM_B) coincides with the canonical spectrum of the pair (Z_A, Z_B). Moreover, the canonical spectrum $\{\rho_i\}$ is a* complete *invariant of the action of the group $O(N) \times \big(\mathrm{GL}(k) \times \mathrm{GL}(k)\big)$: two pairs of representations have the same canonical spectrum if and only if $Z'_A = QZ_AM_A$ and $Z'_B = QZ_BM_B$ for some orthogonal transformation Q of $\mathbb{R}^N$ and some M_A, M_B.*

Proof in App. B.4. *First condition.* The spectrum is constant on the orbit. *Second condition.* The spectrum distinguishes orbits, and distinguishes all of them: if the spectra coincide, the pairs are related by a change of coordinates. A statistic with this property we call a *complete* invariant of the class. From completeness it follows that every admissible statistic is a function of the canonical spectrum. As a quantity, ρ forms the basis, and other correct metrics can be obtained from it; for example, linear CKA on features after whitening equals the mean of ρ_i^2, that is, it reduces the whole profile to one number. Of the unsuitable ones, linear CKA on raw features is constant on the orbits of a smaller group but not on the orbits of GL. The R^2 of the stitch depends on the coordinates of the target model, that is, it changes within the orbit. See App. B-C for the full table of tools, the proofs and the check of sensitivity to $\mathrm{GL}(D)$ at the full hidden-state dimension.

Limitation 2: The AUC transfer gap is degenerate. The natural measure of how well a label transfers is the AUC transfer gap of the probe. But if the representations are linearly equivalent, then this metric is degenerate. The observation below states this precisely.

Observation 3. *If $f_B = T f_A$ on the data support with invertible T, then for every label the supremum of AUC over linear probes is the same in both spaces, and a probe transferred as $T^{-\top}$ gives the same score pointwise. The AUC transfer gap in the population limit is zero.*

Proof in App. B.6. The zero is obtained for all labels at once, so a difference of gaps between two labels does not say that one of them is geometrically more universal than the other. And a measured nonzero gap must then be explained not by a difference of spaces but by something else. In our case (**stratum of successful attacks**) by the fact that one and the same label by name selects different rows in the two models; hence the gap has to be split into two parts, and we introduce this split together with the special case (§3). The criterion and the three tools that fail it are Task 1 of §1. The gap of Limitation 2 is degenerate because it returns zero for every label at once. Instead of asking whether a label transfers, we ask how much of the score survives the passage, and the answer splits into what belongs to the pair of models (their canonical spectrum, admissible by the criterion above) and what belongs to the probe. This is computable before any transfer is performed.

Theorem 4 (Transport law)**.** *Let W be the stitch from the space of the target model B to the space of the source model A, fitted on prompt-paired rows without labels, and let w be a probe fitted in the source model. Let u_i, v_i be the canonical variables of the pair, normalised to unit variance, with $\mathrm{corr}(u_i, v_i) = \rho_i$ the canonical correlations and uncorrelated cross pairs, and let c_i be the coordinates of the score Z_Aw in the complete canonical basis of the source model, $Z_Aw = \sum_i c_iu_i$. If the carried score is the population-optimal linear reconstruction of this score from the representation of the other model, $(Z_BW)w = \sum_i c_i\rho_iv_i$, then*

$$\mathrm{corr}\big(Z_Aw,\ (Z_BW)w\big) \ =\ \sqrt{\mathbb{E}_\omega[\rho^2]}, \qquad \omega_i = \frac{c_i^2}{\sum_j c_j^2}, \tag{1}$$

where the weights ω_i form the probe profile*, the score distribution over the canonical directions.*

See App. B.7 for proof and status of the statement (canonical analysis goes back to Hotelling (1936)). On the rows on which the canonical analysis is fitted, all four premises of the statement hold by construction; outside the fold, diagonality is violated, and for that reason §4.2 compares two estimators of one and the same quantity instead of reading Eqn. (1) as a bound (App. B.7). The direction w is fitted on a label, but Eqn. (1) predicts the recovery of the variance of the *score*, not the preservation of class separation. The transport law holds equally for a probe trained on any binary label and for an arbitrary direction.

Corollary 5. *Under the premises of Theorem 4, for any direction w, $\operatorname{corr}\big(Z_A w,\ (Z_B W)w\big) \leq \rho_1$. The transfer of the source model's label direction therefore does not exceed the first canonical correlation, for any label.*

The leading canonical correlation of the pair bounds the transfer of anything, so it serves as the upper bound against which the measured value is compared (§4).

Corollary 6. *Under the premises of Theorem 4, let w and w' be two probes in one model pair, with profiles ω and ω'. Then $\operatorname{corr}\big(Z_A w, (Z_B W)w\big) - \operatorname{corr}\big(Z_A w', (Z_B W)w'\big) = \sqrt{\mathbb{E}_\omega[\rho^2]} - \sqrt{\mathbb{E}_{\omega'}[\rho^2]}$. Two probes in one model pair thus transfer differently exactly as much as their profiles differ.*

Comparing labels with one another becomes comparing profiles and says nothing about the models themselves (Task 2 of §1).

Label decomposition of transport. The transport law does not say what happens to the label. The following statement is not a corollary of the transport law (Theorem 4). It is derived from the law of total covariance conditioned on the label and holds for any pair of scores, however they are related.

Lemma 7 (Label decomposition)**.** *Let s be the probe's score in its source model, $\hat{s}$ its carried score, and y a binary label with a share $\pi \in (0,1)$ of inputs labelled 1, common to both scores. For either score write $d' = \Delta/\sigma_w$, with Δ the difference of its means across the classes of y and σ_w its pooled within-class standard deviation, and let r_w be the pooled within-class correlation of $(s, \hat{s})$. Then*

$$\operatorname{corr}(s, \hat{s}) = \frac{r_w + \pi(1-\pi)\, d'_s\, d'_{\hat{s}}}{\sqrt{\big(1 + \pi(1-\pi)d'^2_s\big)\big(1 + \pi(1-\pi)d'^2_{\hat{s}}\big)}}, \tag{2}$$

and the between-class share of the score variance equals $\pi(1-\pi)d'^2/(1+\pi(1-\pi)d'^2)$.

See App. B.8 for proof and the rank form of the decomposition. The transfer of the source model's label direction $\operatorname{corr}(s, \hat{s})$ mixes within-class r_w and label transport $\tau = d'_{\hat{s}}/d'_s$ with a weight equal to the between-class share of the variance. For small π this share is small, so the transfer of the source model's label direction almost coincides with within-class transport. The quality of the carried probe is then determined by the quantity τ Combining the transport law (Theorem 4) with Lemma 7 on the label decomposition gives two transports that are measured separately: the transport of the direction is capped by ρ_1, τ is not (Task 3 of §1, measured in §4).

3 Special case: the subset of prompts with a successful attack

Data. We fix two groups of architectures: *i) classical-architecture* models that are decoder transformers without a mixture of experts (MoE) and with ordinary softmax attention; and *ii) alternative-architecture* models that are models with a modification of that setup, and each of the four families in the set modifies it in its own way. See App. F for complete architectural details. The one model with linear attention, `qwen3_8_27b`, stands outside W_{mix19} as a single control (set $\mathcal{C}$ below). Together we analyze the models in a matrix W_{mix19} with **19** models: 11 classical- and 8 alternative-architecture. The models form 342 ordered pairs: 110 pairs of two classical-architecture models, 56 pairs of two alternative- models and 176 mixed pairs of a classical- model and an alternative-architecture model.

Set $\mathcal{A}$ is eight of the eleven classical-architecture models, from 0.5B to 70B parameters, from five model vendors; set $\mathcal{B}$ keeps one model per vendor (App. I.4); set $\mathcal{C}$ adds the hybrid-attention model `qwen3_8_27b` (App. I.5). They nest, $\mathcal{B} \subset \mathcal{A} \subset \mathcal{C}$, so the pairwise measurements of a smaller set are a subset of the rows of a larger one. Sets are written calligraphically; plain A and B always denote the source and the target model of a pair. The representation $f_\theta(x) \in \mathbb{R}^D$ is the hidden state at the last token of the prompt at relative depth p; the depth grid contains fourteen positions, and the map

from p to a block index is in App. F. The corpus is 3306 prompts combining JailbreakHub (Shen et al., 2024) and Prompt-injections benchmark (Rogue Security, 2025). Each model's answer to each prompt is judged by ten guards, two of different sizes from each of five guard vendors (App. F): Llama Guard (Grattafiori et al., 2024; Meta AI, 2025), Qwen3Guard (Zhao et al., 2025a), OpenGuardrails (Wang & Li, 2025), XGuard (Lin et al., 2026) and GPT-OSS-Safeguard (OpenAI, 2025).

Comparison arms and the decomposition of the AUC transfer gap. For an ordered pair "source model A, target model B" and a guard, we compute on the test rows of the fold three linear-probe scores and check each against one and the same label — the guard's verdict on the target model's answer: self $= Z_B w_B^B$, label-only $= Z_B w_A^B$ and transfer $= Z_B W w_A^A$; the superscript of the probe names the model in whose space it is trained, the subscript the model on whose guard verdict it is trained. We call these three scores the comparison arms: they are computed on the same rows and checked against one and the same label, and they differ in two factors — in whose space the probe is trained and on whose label. Adjacent arms differ in exactly one of them, so the AUC transfer gap decomposes into the *label AUC gap* $\mathrm{AUC}_{\text{self}} - \mathrm{AUC}_{\text{label-only}}$, in which only the label changes, and the *geometry AUC gap* $\mathrm{AUC}_{\text{label-only}} - \mathrm{AUC}_{\text{transfer}}$, in which only the space changes. This is the split referred to by Observation 3: the observation itself says that under linear identifiability the AUC transfer gap is zero for every label, so a measured nonzero gap means that the verdicts of one guard on the answers of two models mark different prompts. The label AUC gap measures this disagreement, the geometry AUC gap the remainder, to which the observation applies. The stitch W is a ridge map from the space of the target model to the space of the source model, fitted on prompt-paired rows without labels. We compare the size of the measured gaps with two reference points. The *procedure floor* is the AUC transfer gap between a model and its own twin $Z' = ZM$, *i.e.,* the same features after a known invertible map. By Observation 3 the true gap here is zero, so the whole measured gap is the error of the procedure.

How we make models comparable. Models are not directly comparable, and we use three devices. Each is a choice, not a given, and for each the sensitivity is measured: the reduction to the common number k of principal components assumed in §2, the relative depth p, and a model-invariant label — the consensus prompt label, equal to 1 when no fewer than half of the eighty verdicts of ten guards on the answers of eight models judge the answer of its own model unsafe (App. G).

4 Empirical Analysis

4.1 Why the transport law is needed: the label transfers, not the geometry

The AUC transfer gap, which the field reads as the universality of a label, says almost nothing about the representations, *i.e.,* the part of it that comes from the representations is not distinguishable from the error of the procedure. That is why a different quantity is needed, and the transport law of §4.2 is that quantity. The label and geometry AUC gaps differ by an order of magnitude. Across the fourteen positions the label AUC gap is 0.062–0.092, the geometry AUC gap $[+0.0052, +0.0122]$, and the procedure floor at each of the fourteen positions is the range of the AUC transfer gap over the model-twin control pairs, where the true gap is zero by Observation 3; the edges of this floor across positions lie within $[-0.0317, +0.0757]$ (set W_{mix19}, all 19 models, $k = 64$). The ranges of the label and geometry AUC gaps do not overlap, and at all fourteen positions the geometry AUC gap lies inside the procedure floor (Fig. 2). The AUC transfer gap, which the field measures as a measure of the label's universality, consists in our case of 0.092 from the disagreement between the guard's verdicts on the answers of the two models and of $+0.0070$ from the difference between the representations.

This conclusion does not rest on any of the choices we made. It holds when the rows with an empty model answer and the rows with a recovered guard verdict are removed from the corpus (App. I.1); when one model is kept per vendor (App. I.4); when the ridge stitch is replaced by the Procrustes stitch, *i.e.,* by an orthogonal map (App. K.3); and when the probes and the stitch are fitted on one prompt source and the AUC is computed on the other (App. K.2). Structured degradation of the stitch raises the geometry AUC gap, but it does not distinguish the two kinds of degradation (App. K.5). If we take a label common to all models, the label AUC gap is zero by construction, and the whole AUC transfer gap turns out to be geometric. It equals $+0.0107$, 95% $[+0.0086,\ +0.0131]$, with a procedure floor of $[-0.0018, +0.0239]$ (App. L.2). When the label is the same for all models, the representations of two models carry the label without measurable loss.

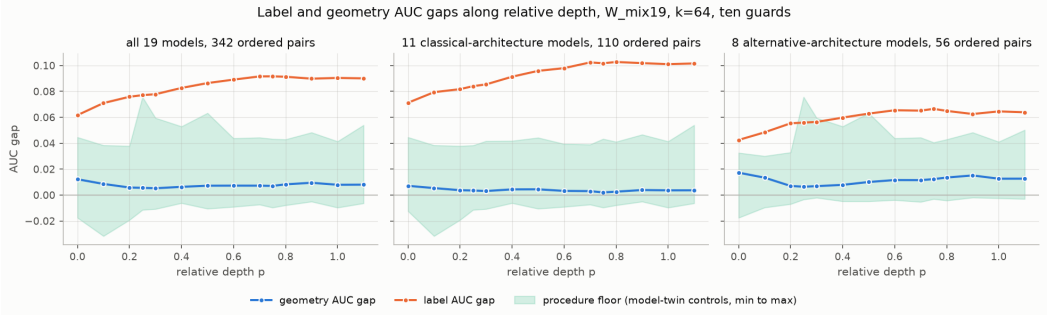

Figure 2: Label and geometry AUC gaps across fourteen relative depths against the procedure floor (minimum to maximum of the geometry gap over the model-twin controls), W_{mix19}, $k = 64$, ten guards; left, all 19 models (342 ordered pairs); middle, the 11 classical-architecture models (110 pairs); right, the 8 alternative-architecture models (56 pairs). The geometry gap stays inside the floor at every depth in all three panels. The label gap lies above the floor at every depth for all models and for the classical ones; for the alternative ones the upper edge of the floor reaches above it at depths 0.25, 0.3 and 0.5.

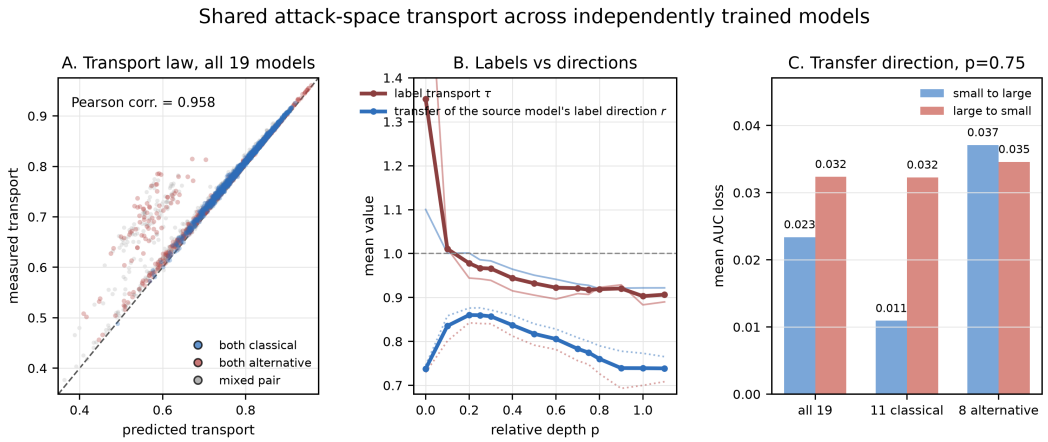

Figure 3: Transport on W_{mix19} (19 models, 342 ordered pairs, ten guards, $k = 64$). A: transfer of the source model's label direction predicted by the transport law from the canonical spectrum and the probe profile against the value measured through the ridge stitch, at relative depth $p = 0.75$; points coloured by pair type (both classical, both alternative, mixed). B: label transport τ and transfer of the source model's label direction r along relative depth, all 19 models in bold, the 11 classical-architecture and the 8 alternative-architecture models in faint lines. C: AUC loss of the carried probe by the direction of transfer in model size at $p = 0.75$, for all 19, the 11 classical and the 8 alternative models. Read: the points of A lie on the diagonal; in B, τ stays near one while r stays below it in every group; in C, transfer from the smaller to the larger model loses less for all models and for the classical ones, while the alternative ones show no such asymmetry.

4.2 The transport law on data

The transport law (Theorem 4) is tested by comparing two estimates of one quantity: the correlation reached by the ridge stitch, and the plug-in estimate $\sqrt{\mathbb{E}_\omega[\rho^2]}$ from the spectrum and the probe profile. On the eleven classical-architecture models the two coincide: over their cells the Pearson coefficient is 0.998 and the OLS slope of the measured value on the predicted one 0.980. On the eight alternative-architecture models the law still orders the cells, 0.925, but over-predicts their transport, slope 0.775; over all 3420 ordered model pair $\times$ guard cells of W_{mix19} this averages to 0.958 and 0.843 (1.018 through the origin) (Figure 3A and App. L.6).

The leading canonical correlation of the pairs is $\rho_1 = [0.967, 0.990]$, so by Corollary 5 a probe in these pairs could reach almost perfect transport, while the probe trained on a guard's verdict reaches $[0.740, 0.865]$ (§4.4). What limits transport is therefore not how well the two spaces are aligned but which direction the probe picks out of them (Task 2 of §1). Corollary 6 is tested by substitution in App. L.7: with the probe's own profile replaced by the profile of the probe trained on the next guard's verdict, the law's prediction keeps 0.21 of the residual variance of the measured transfer that it explained, 0.17 among the classical- and 0.41 among the alternative-architecture models.

4.3 What happens to the label, and what to the direction

The label decomposition separates two quantities that Eqn. (1) does not distinguish. The transfer of the source model's label direction is $r = 0.740$–0.865, below the ceiling of §4.2; ranges here and below are the minimum and maximum over the depth means (App. L.5). Label transport is $\tau = 0.904$–1.370, *i.e.,* the label survives transfer almost without loss and at some depths with a gain, while the AUC loss of the carried probe is $[0.009, 0.032]$. The discrepancy is explained by Lemma 7: the between-class share of the score variance is $[0.09, 0.17]$, so r nearly coincides with within-class transport $[0.70, 0.84]$ and measures the part of the score that does not separate the classes.

Is this specific to attacks. These two statistical indicators contradict each other because they differ in their construction. By the transfer of the generalised eigendirections of the covariance pair, the attack stratum is no higher than the connected control with the same share of prompts labelled 1, and on a set of 32 binary labels no advantage survives (App. H.3): nothing here is attack-specific. But that statistic is computed from covariances centred within classes, so a difference of means does not enter it — a planted mean shift moves it by $1.4 \cdot 10^{-4}$, and the leading generalised eigenvalue by zero. Along the probe direction the answer is the opposite: label transport for the guard's verdict exceeds that of a control slab matched to it in the same share and in decodability by $+0.176$–$+0.256$ (App. J.1), and the guard-verdict axis lowers the intrinsic-dimension contrast more than an orthogonal axis of the same variance (App. L.4). The specificity of the attack shows up in the difference of means, and the second-moment statistic is blind to it by construction (Task 4 of §1).

What the setup could have seen. A result that reports finding nothing can be read only together with the resolution of the setup, which we measure by planting a known structure into both models and running the whole pipeline on the planted data (App. H.2). On W_{mix19} at $k = 64$ both planted structures are detected in all six settings: the shared subspace once it inflates the variance of the attack stratum along itself by 1.6–2.0, a single shared direction by 2.0–2.4 (Fig. 4).

4.4 Transfer is asymmetric in model size

The label decomposition has separated the label from the direction, and against this background the only predictor of AUC loss that we found becomes visible — the direction of transfer in model size. One must compare the two directions of *one and the same* pair: the guards' agreement on the answers of the two models is symmetric, so it cannot explain the difference between the directions. We call the asymmetry of a pair the difference of the losses "down minus up", where up is transfer from the smaller model to the larger one.

Over the 171 unordered pairs of models of different size of set W_{mix19} the mean asymmetry is $+0.0090$ and is positive in 106 pairs out of 171. This is not a threshold effect: the larger the size gap, the larger the asymmetry; the rank correlation with the logarithm of the size ratio is $+0.366$. The mean AUC loss for transfer up is 0.0234 at $\tau = 0.94$, down — 0.0324 at $\tau = 0.90$ (Figure 3C; Fig. 30 in App. L.5 shows every pair against its size ratio). The asymmetry is carried by the classical-architecture models. Over the 55 pairs of two classical-architecture models the mean asymmetry is $+0.0214$, it is positive in 42 pairs out of 55, the rank correlation with the logarithm of the size ratio is $+0.562$, and the AUC loss up is 0.0109 at $\tau = 0.97$ against 0.0323 at $\tau = 0.89$ down. Over the 28 pairs of two alternative-architecture models there is no asymmetry: the mean is -0.0025, it is positive in 16 pairs out of 28, the rank correlation is $+0.130$ at $p = 0.51$, and the losses in the two directions are close — 0.0371 up and 0.0346 down. The difference in the share of unsafe answers does not explain this. In a regression of the asymmetry on the logarithm of the size ratio and the difference in the share of unsafe answers of the two models, the size coefficient is $+0.0347$, 95% $[+0.0291,\ +0.0404]$ by a bootstrap clustered by model (Cameron et al., 2008), and it is positive in every resample; the share of unsafe answers enters the same regression and does not absorb the size

effect. Over the pairs of two classical-architecture models the same coefficient is $+0.0432$, 95% $[+0.0372,\ +0.0489]$, and over the pairs of two alternative-architecture models it is $+0.0219$, 95% $[-0.0090,\ +0.0462]$, that is, the interval covers zero. The vendor does not explain it either: over the nine same-vendor pairs of classical-architecture models the asymmetry is positive in seven and lies within $[-0.0529, +0.0849]$, and the largest is for the pair `qwen25_05b`–`qwen25_7b`.

An explanation suggests itself: the dimension of the subspace shared by the two models is bounded by the smaller model of the pair, and in transfer into the smaller model there is nowhere to put part of the separation. The test of this explanation — a regression of the dimension of the subspace shared by the two models on the size of the smaller model with a control for vendor — by the reading rule written before the data gives no outcome (App. J.4). We therefore report the asymmetry as a measured fact without an established mechanism (Task 3 of §1).

5 Conclusion

In this work, we asked whether the two representation hypotheses imply that a behaviorally meaningful linear direction learned in one model is recovered in, and transfers to, another, and we answered the question for the direction of a successful attack, in text decoders, with a linear probe and a linear stitch. Under linear identifiability the AUC transfer gap is zero for every label, so the gap measured between two models reports the disagreement between the guards' verdicts on their answers and not a difference between their representations. Evidence of transfer has to come from a statistic that is invariant to a change of coordinates on either side, and the transport law provides one. The correlation between the probe's score in its source model and its carried score equals the root-mean-square canonical correlation of the pair along the probe profile, and on the classical-architecture models the prediction coincides with the measurement. The label decomposition then separates what the AUC gap merges. The label transports almost without loss, at some depths with a gain, whereas the direction itself is only partially recovered.

Against this background one predictor of AUC loss becomes visible. A probe carried from a smaller model to a larger one loses almost no AUC, the reverse transfer does lose, and the loss grows with the size gap. The asymmetry is carried by the classical-architecture models and is absent on the alternative-architecture models. Its mechanism remains open, since the test of the natural explanation, that the subspace shared by the two models is bounded by the smaller model of the pair, gives no outcome under a control for vendor. For monitoring, it is cheaper to train a probe on the smaller model and carry it to the larger one than the other way round. The transfer is measured relative to the prompt distribution, and on the other prompt source it drops by half together with the prediction of the transport law, so a transfer checked on one traffic says nothing about new traffic.

Limitations. Features are read at the last token of the prompt while the label is a verdict on the generated answer, so we measure how well the representation of the request predicts the outcome of generation. The share of unsafe answers is confounded with the guard's vendor, and model size with the model's vendor. We pool the eight alternative-architecture models and do not break transfer down by family, so we cannot say which departure from the classical architecture drives the difference (Appendix L.5). The one model with linear attention, `qwen3_8_27b`, was not readable by the rule written before the data (Appendix I.5), so we claim nothing about linear attention. Refusal and consent are not tested: both are mean shifts, to which the second-moment statistic is blind by construction (§H.2).

Reproducibility statement

All models are public checkpoints, listed with their hidden-state dimension D, stored depth L and the relative-depth grid in Appendix F; the corpus is $N = 3306$ prompts from two public prompt sources, 1307 from JailbreakHub and 1999 from the prompt-injections benchmark. Proofs of every theoretical statement are in Appendix B, and the conditions under which each result was checked — truncation, training-row budget, strict masks, other stitches, other model sets — are in Appendices D, I, J and K. The anonymised repository (`https://anonymous.4open.science/r/universal_attack_subspace_transport_law/`) contains the code for extracting answers and hidden states, grading answers with the ten guards, and every measurement, together with the committed driver of each reported run; the guards' verdicts for all models; the fixed cross-validation folds; and the accompanying notebook.

AI USE STATEMENT

Throughout the work we used an AI assistant: we wrote and checked the analysis code with it, conducted an adversarial review of the paper's claims (we tried to refute each claim with computed controls), and wrote and edited the text of the paper. We constrained the work with the AI assistant by rules; compliance with them can be checked in the published repository.

Design of the experiments, which controls were needed and which claims to make in the paper were decided by the authors, not by the AI assistant. Every number is produced by a cell of the published accompanying notebook; the cell reads a log file, and the log file was written by an analysis script committed to the repository.

The authors are responsible for the entire content of the paper, including any error that the rules for working with the AI assistant did not prevent and the adversarial review did not find.

REFERENCES

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Marcus Armstrong, Navid Ayoobi, and Arjun Mukherjee. Thinking in different spaces: Domain-specific latent geometry survives cross-architecture translation. *arXiv preprint arXiv:2603.20406*, 2026.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Blake Bullwinkel, Mark Russinovich, Ahmed Salem, Santiago Zanella-Beguelin, Daniel Jones, Giorgio Severi, Eugenia Kim, Keegan Hines, Amanda J. Minnich, Yonatan Zunger, and Ram Shankar Siva Kumar. A representation engineering perspective on the effectiveness of multi-turn jailbreaks. In *Data in Generative Models - The Bad, the Ugly, and the Greats, Workshop at ICML 2025*, 2025.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427, 2008.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haytham M. Fayek, Lawrence Cavedon, and Hong Ren Wu. On the transferability of representations in neural networks between datasets and tasks, 2018.

Aaron Grattafiori et al. The llama 3 herd of models, 2024.

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

David Huang, Avidan Shah, Alexandre Araujo, David Wagner, and Chawin Sitawarin. Stronger universal and transferable attacks by suppressing refusals. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5850–5876, 2025a. doi: 10.18653/v1/2025.naacl-long.302.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3686–3704, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.185.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

Hakan Inan, K. Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *ArXiv*, abs/2312.06674, 2023. URL https://api.semanticscholar.org/CorpusID:266174345.

Rishi Dev Jha, Collin Zhang, Vitaly Shmatikov, and John Xavier Morris. Harnessing the universal geometry of embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019.

János Kramár, Joshua Engels, Zheng Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for gemini, 2026.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence, 2015. URL https://arxiv.org/abs/1411.5908.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zelin Li, Kehai Chen, Lemao Liu, Xuefeng Bai, Mingming Yang, Yang Xiang, and Min Zhang. Tf-attack: Transferable and fast adversarial attacks on large language models, 2024.

Junyu Lin, Meizhen Liu, Xiufeng Huang, Jinfeng Li, Haiwen Hong, Xiaohan Yuan, Yuefeng Chen, Longtao Huang, Hui Xue, Ranjie Duan, Zhikai Chen, Yuchuan Fu, Defeng Li, Lingyao Gao, and Yitong Yang. Yufeng-xguard: A reasoning-centric, interpretable, and flexible guardrail model for large language models, 2026.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Alex Troy Mallen and Nora Belrose. Eliciting latent knowledge from quirky language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, and Dmitrii Krasheninnikov. Detecting high-stakes interactions with activation probes. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Meta AI. Llama Guard 4 model card. https://huggingface.co/meta-llama/Llama-Guard-4-12B, 2025. 12B dense multimodal safety model pruned from Llama 4 Scout.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation, 2018.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2.

Narmeen Fatimah Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amir Abdullah. Activation space interventions can be transferred between large language models. In *Forty-second International Conference on Machine Learning*, 2025.

OpenAI. Technical report: Performance and baseline evaluations of gpt-oss-safeguard-120b and gpt-oss-safeguard-20b. OpenAI technical report, October 2025. Accessed 2026-09-21.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Geoffrey Roeder, Luke Metz, and Diederik P. Kingma. On linear identifiability of learned representations. *ArXiv*, abs/2007.00810, 2020.

Rogue Security. Prompt injections benchmark: Jailbreak vs. benign. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/rogue-security/prompt-injections-benchmark. Jailbreak subset, CC BY-NC 4.0; accessed 2026-09-23.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024.

Woo Seob Sim and Yu Rang Park. Geometry-lite: Interpretable safety probing via layer-wise margin geometry, 05 2026.

Agus Sudjianto and Alice J. Liu. Decomposing global auc into cluster-level contributions for localized model diagnostics, 2025.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence–arousal subspace in LLMs: Circular emotion geometry and multi-behavioral control. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024.

Thomas Wang and Haowen Li. Openguardrails: A configurable, unified, and scalable guardrails platform for large language models, 2025.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schuetze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jan Wehner and Mario Fritz. Probe-based fine-tuning for reducing toxicity, 2025.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng, Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. Qwen3guard technical report, 2025a. URL https://arxiv.org/abs/2510.14276.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. *Advances in Neural Information Processing Systems*, 38:140283–140318, 2025b.

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

A Related works

Platonic and Linear Representation Hypotheses. The linear representation hypothesis holds that a concept corresponds to a direction in the representation space (Park et al., 2024). From the safety perspective within a model, refusal to answer is mediated by one such direction(Arditi et al., 2024), while the harmfulness of a request and the refusal to answer it are encoded separately (Zhao et al., 2025b). But concepts are not always one-dimensional (Engels et al., 2025; Sun et al., 2026). The Platonic representation hypothesis holds that independently trained models converge on a shared structure of representations (Huh et al., 2024). Calibrating a global similarity index against a permutation null removes much of the reported global convergence and leaves agreement that is local in geometry (Gröger et al., 2026).

Probing. High-level properties are linearly decodable from hidden states, where models encode stable signal, which middle-layer probes recover under distribution shift (Mallen & Belrose, 2024); multi-turn attacks succeed by gradually moving harmful outputs into regions probes classify as benign (Bullwinkel et al., 2025); and when probes become optimization targets the signal remains linearly present but rotates, so retraining the probe restores accuracy (Wehner & Fritz, 2025). Reducing each layer to a margin against centroid, k-NN and linear-boundary references shows that safety information is a persistent position relative to a decision boundary rather than motion between layers (Sim & Park, 2026). Such probes are proposed as deployment-time monitors, matching a separate language-model monitor at six orders of magnitude less compute (McKenzie et al., 2025), and are already deployed by vendors (Kramár et al., 2026). The methodology of probing requires comparing a probe against one of the same capacity on a control task, without which its quality cannot be interpreted (Hewitt & Liang, 2019).

Transferring directions between models. A second line builds an explicit map between two representations and measures what survives the passage. Model stitching joins a layer of one network to another through a learned affine map (Lenc & Vedaldi, 2015; Bansal et al., 2021; Csiszárik et al., 2021). Relative representations express each point through its similarities to shared anchors (Moschella et al., 2023), and the same anchors estimate the map in closed form, where a linear or orthogonal map already stitches independently trained encoders (Maiorca et al., 2023). vec2vec translates embeddings between models without paired data (Jha et al., 2025), and a linear map from teacher activations into a student residual stream corrects the student's answers at generation time, though geometric fit does not predict behavioral gain (Armstrong et al., 2026). Safety interventions transfer between model families through a learned, nonlinear mapping (Oozeer et al., 2025), and a refusal vector transfers across languages within one model (Wang et al., 2025). Closest to our setting, Huang et al. (2025b) align concept representations across LLMs with a least-squares linear map and show that a steering vector from a 0.5B model can control a 7B model of the same family.

Their remaining cross-model experiments pair models of similar size, so the comparison between the two directions of a single pair does not arise there. Asymmetry of transfer has been reported across tasks, where one task transfers to another better than the reverse (Fayek et al., 2018).

Similarity indices and the invariance they assume. Two representation spaces are compared by scalar indices like canonical correlation analysis (Hotelling, 1936) and its variants SVCCA (Raghu et al., 2017) and PWCCA (Morcos et al., 2018), centred kernel alignment (Kornblith et al., 2019), and the overlap of mutual k nearest neighbours (Huh et al., 2024). These indices differ in the group of transformations they ignore, where linear CKA is invariant to orthogonal transformations and isotropic scaling but not to arbitrary invertible linear maps, but canonical correlations are invariant to any invertible change of coordinates on each side. The choice changes the answer in practice as Huh et al. (2024) moved to mutual kNN after CKA showed "a very weak trend" on the same representations. What is missing is a rule that selects the index from what the linear read-out itself can distinguish. Of these, PWCCA stands closest to our transport as it weights each canonical correlation by the share of the representation that direction explains (Morcos et al., 2018).

B PROOFS

Throughout, $Z_A, Z_B \in \mathbb{R}^{N\times k}$ are the representations of two models, truncated to k principal components in the notation of §**??**; the row with the same index in both matrices is obtained on one and the same prompt. Σ_1 and Σ_0 are the covariances of one model's representation on the subset of prompts to which a binary prompt label (for example, the guard's verdict on this model's answer) assigned 1, and on the remaining prompts. $\mathrm{GL}(k)$ is the group of invertible linear maps of a k-dimensional space. The three propositions below — on intrinsic-dimension estimators, on the completeness of the canonical spectrum of a pair of representations (stated in §2), and on the invariance of the generalised spectrum of the covariance pair (Σ_1, Σ_0) — determine with which representation comparison statistics representations *may* be compared. Three further results — Observation 3 on the AUC transfer gap, the transport law (Theorem 4) and Lemma 7 on label decomposition — describe the transfer of a probe from the source model to the target model.

B.1 LINEARITY OF THE IDENTIFIABILITY CLASS

A linear change of coordinates does not change the class of linear functions of the representation: $w^\top(Tx) = (T^\top w)^\top x$. When a model is trained, no term of its loss function penalises anisotropy of the representation basis. Roeder et al. (2020) prove identifiability up to a linear map for discriminative model families of the form $p_\theta(y \mid x, S) \propto \exp(f_\theta(x)^\top g_\theta(y))$, in which the logit is an inner product. If the diversity condition in y holds, then equality of the conditional distributions of two models of such a family implies $f' = Tf$ and $g' = T^{-\top} g$ with one and the same invertible matrix T. A language model with a next-token prediction head defines a conditional distribution of the next token of the same form.

Before applying this theorem to our models, two circumstances must be noted. First, the theorem assumes equality of the conditional distributions of two models, and for our models it does not hold: our models are trained independently. Hence the theorem does *not* assert that the representations of our two models are related by an invertible matrix T. The theorem fixes the *limit of distinguishability of representations*: the function a model computes determines its representation only up to a linear change of coordinates. Second, the theorem concerns the last layer, whose features are fed into the language-model head, whereas we take the hidden state of the last prompt token from an intermediate layer. The nonlinear blocks that follow an intermediate layer can compensate even a nonlinear change of coordinates in it. Therefore, for an intermediate layer, we take GL as the identifiability class, that is, the group of transformations up to which we distinguish representations, not because of the theorem, but because we ourselves work with linear functions of the representation. The probe (a logistic regression that predicts a binary prompt label from the rows of the representation) and the stitch (a ridge map from the target model's space to the source model's space, fitted on one and the same prompts without labels) do not distinguish representations that a transformation from GL maps into one another.

B.2 THE SECOND ADMISSIBILITY CONDITION FOR A COMPARISON STATISTIC: SEPARATION OF ORBITS

The admissibility criterion for representation comparison statistics (§2) consists of two conditions. The first condition, invariance, means that the comparison statistic is constant on each orbit of pairs of representations under the action of invertible linear changes of coordinates in each of the two models. The second condition, separation of orbits, means that the comparison statistic takes different values on different orbits. If the hypothesis $f_B = Tf_A$ that the representations of the two models are related by an invertible linear map holds, the pair of representations lies in a single orbit (the orbit of the hypothesis), and on the orbit of the hypothesis *every* GL-invariant comparison statistic is constant. Therefore constancy of a comparison statistic on the orbit of the hypothesis is not a defect of that statistic but the very definition of invariance. The condition of separation of orbits is violated by a comparison statistic that assigns one and the same value to pairs of representations from *different* orbits. The example below shows that the AUC transfer gap (the difference between the AUC of a probe trained in the target model on the target model's label and that of a probe trained in the source model on the source model's label and carried by the stitch into the target model's space, §3) does not separate orbits.

Take a representation Z_A with k independent coordinates, the synthetic label $y = [Z_A e_1 > 0]$, which assigns label 1 to a prompt when the first coordinate of Z_A on that prompt is positive, and two partners, that is, two representations each of which will form a pair with Z_A. The first partner, $Z_{B_1} = Z_A T$, is obtained from Z_A by an invertible linear change of coordinates, and its pair with Z_A lies in the orbit of the hypothesis $f_B = Tf_A$. The pair of the second partner Z_{B_2} with Z_A lies not in the orbit of the hypothesis $f_B = Tf_A$ but in another orbit: Z_{B_2} shares only the first coordinate with Z_A, and the remaining coordinates of Z_A are replaced in Z_{B_2} by independent noise. The canonical spectrum, that is, the set of canonical correlations of a pair of representations, separates these two pairs: for the pair with the first partner $\rho_i \equiv 1$, while for the pair with the second partner there is only one unit canonical correlation. The AUC transfer gap of the synthetic label y is zero in both pairs: the label y depends on a single direction, the first coordinate of Z_A, and each of the two partners shares this direction with Z_A. Keep the representations Z_A, Z_{B_1} and Z_{B_2} on the corpus prompts unchanged and replace the label y with $y' = [Z_A e_2 > 0]$, which depends on the second coordinate of Z_A: the AUC transfer gap of the label y' in the pair of Z_A with Z_{B_2} becomes positive. Hence the AUC transfer gap is a function of the triple (representation, representation, label), not of the pair of representations, and cannot be an invariant of the pair of representations. We checked this example numerically in the accompanying notebook.

B.3 INTRINSIC-DIMENSION ESTIMATORS ARE INVARIANT ONLY TO SIMILARITIES

Proposition 8 (invariance group of intrinsic-dimension estimators)**.** *Let $\hat{d}_{\mathrm{ID}}$ be any of the intrinsic-dimension estimators TwoNN, MLE or TLE. Then for every $c_{\mathrm{sim}} > 0$, orthogonal transformation Q and shift b, $\hat{d}_{\mathrm{ID}}(\{c_{\mathrm{sim}} Q x_i + b\}) = \hat{d}_{\mathrm{ID}}(\{x_i\})$ holds exactly, and the group of exact invariance of these estimators* on the space of all finite point configurations *is exactly the group of similarities: for every $T \in \mathrm{GL}(D)$ that is not a similarity, there exists a point configuration on which the estimate changes after the transformation T.*

Proof. Denote by $\varepsilon_j(i)$ the distance from the point x_i to its j-th nearest neighbour. All three estimators depend only on ratios of pairwise distances: TwoNN uses $\mu_i = \varepsilon_2(i)/\varepsilon_1(i)$, MLE uses $\log(\varepsilon_{k_{\mathrm{nn}}}(i)/\varepsilon_j(i))$, TLE uses ratios of the pairwise distances between the neighbours of the point x_i to the radius of its neighbourhood $\varepsilon_{k_{\mathrm{nn}}}(i)$, and in each of them the numerator and the denominator are homogeneous of degree one in the distances. A similarity gives $\|(c_{\mathrm{sim}} Q x_i + b) - (c_{\mathrm{sim}} Q x_j + b)\| = c_{\mathrm{sim}} \|x_i - x_j\|$, multiplying *all* pairwise distances by the common factor c_{sim}, which cancels in any of their ratios. Conversely, a transformation T that is not a similarity changes the ratio of some two pairwise distances, and on a point configuration in which these two distances occur, the estimate changes after the transformation T. □

The stabiliser of an *individual* point configuration, that is, the set of transformations from $\mathrm{GL}(D)$ that do not change the estimate on this configuration, can be wider than the group of similarities. Therefore, the fact that an estimator is not invariant to $\mathrm{GL}(D)$ does not imply that its estimate will shift on specific data, and the shift of the estimate on data has to be measured. The non-invariance

of the estimators to $\mathrm{GL}(D)$ is proved in Proposition 8: the invariance group of the estimators is strictly smaller than the identifiability class $\mathrm{GL}(D)$, so comparing different models by the intrinsic-dimension estimate is ill-defined.

B.4 Canonical correlations are a complete invariant of a pair of representations

Proposition 2 (completeness of the canonical spectrum). *For representation matrices $Z_A, Z_B \in \mathbb{R}^{N\times k}$ on the same N prompts and any $M_A, M_B \in \mathrm{GL}(k)$, the canonical spectrum $\{\rho_i\}$ of the pair (Z_AM_A, Z_BM_B) coincides with the canonical spectrum of the pair (Z_A, Z_B). Moreover, the canonical spectrum $\{\rho_i\}$ is a* complete *invariant of the action of the group $O(N) \times \big(\mathrm{GL}(k) \times \mathrm{GL}(k)\big)$: two pairs of representations have the same canonical spectrum if and only if $Z'_A = QZ_AM_A$ and $Z'_B = QZ_BM_B$ for some orthogonal transformation Q of $\mathbb{R}^N$ and some M_A, M_B.*

Proof. The canonical correlations ρ_i are the cosines of the principal angles between the column spaces $\mathrm{col}(Z_A)$ and $\mathrm{col}(Z_B)$ in $\mathbb{R}^N$. Right multiplication of a representation matrix by an invertible matrix does not change its column space, and hence does not change the principal angles. Completeness of the canonical spectrum as an invariant follows from Jordan's theorem on principal angles: for two k-dimensional subspaces U, V of $\mathbb{R}^N$ there exists an orthonormal basis of $\mathbb{R}^N$ in which the pair of subspaces (U, V) is written as 2×2 rotation blocks by the angles $\theta_1 \le \cdots \le \theta_k$ (the principal angles) and nothing else. Consequently, two pairs of subspaces (U, V) and (U', V') are mapped into one another by a single orthogonal transformation Q if and only if their multisets of principal angles coincide: the canonical spectrum $\rho_i = \cos\theta_i$ determines the pair of subspaces up to Q — no finer (Q itself is not recovered) and no coarser (different multisets of angles are different pairs of subspaces under any Q). □

Necessity of the factor $O(N)$. With respect to the group $\mathrm{GL} \times \mathrm{GL}$ alone, acting by right multiplication on both representation matrices, the canonical spectrum is not a complete invariant: for $N = 2$, $k = 1$ the configurations $(1, 0)^\top$ and $(0, 1)^\top$ lie in different $\mathrm{GL}(1)$ orbits but have the same canonical correlation ρ. The factor $O(N)$ permutes and rotates the rows of both representation matrices, that is, it acts on *which prompt is recorded under which index*. With the factor $O(N)$ the two $\mathrm{GL}(1)$ orbits of this example merge, and completeness of the canonical spectrum as an invariant is restored. Independence of a representation comparison statistic from the numbering of prompts is a substantive requirement on such a statistic rather than a weakening of the completeness criterion for an invariant.

The limiting case of the canonical spectrum $\rho_i \equiv 1$ is equivalent to $Z_B = Z_AT^\top$ with invertible T, that is, exactly to the equivalence relation $f_B = Tf_A$, $T \in \mathrm{GL}(D)$, of two models (§2). In this limiting case the canonical spectrum and linear CKA diverge: a unit canonical spectrum characterises equivalence of representations up to an invertible linear map, whereas a linear CKA value equal to 1 characterises it only up to $O(D) \times \mathbb{R}_{>0}$. If one passes to the quotient space, that is, compares the column spaces of the two representation matrices themselves, the canonical spectrum defines a metric: $\mathrm{dist}^2_{\mathrm{Gr}} = \sum_i \arccos^2 \rho_i$, the geodesic distance on the Grassmannian of k-dimensional subspaces of $\mathbb{R}^N$.

B.5 The generalised spectrum of the covariance pair is GL-invariant

Proposition 9 (invariance of the generalised spectrum)**.** *For every $T \in \mathrm{GL}(D)$, the generalised spectrum $\{\lambda_i\}$ of the pair $(T\Sigma_1T^\top, T\Sigma_0T^\top)$ coincides with the generalised spectrum of the pair (Σ_1, Σ_0).*

Proof. $\det(T\Sigma_1T^\top - \lambda T\Sigma_0T^\top) = (\det T)^2 \det(\Sigma_1 - \lambda\Sigma_0)$, and $(\det T)^2 \neq 0$, so the characteristic polynomials of the original covariance pair and of the pair transformed by the matrix T have the same roots. Completeness of the generalised spectrum as an invariant with respect to simultaneous congruence of both covariances by the matrix T is obtained by normalising by Σ_0: after the covariance Σ_0 is reduced to the identity matrix, the residual freedom in the choice of T is $O(D)$,

and the reduced covariance of the subset of prompts is a symmetric matrix, which is determined up to orthogonal conjugation by its eigenvalues. □

B.6 THE AUC TRANSFER GAP VANISHES UNDER LINEAR IDENTIFIABILITY

Observation 3 (AUC transfer gap). *If $f_B = Tf_A$ on the support of the data with $T \in \mathrm{GL}$, then for every binary prompt label y the supremum of AUC over linear probes is the same in the representation spaces of models A and B, and a probe trained in model A on the label y and carried into model B via $T^{-\top}$ gives on every prompt the same score as in model A; therefore the AUC transfer gap in the population limit is zero.*

Proof. Let $\mathcal{H}_A = \{x \mapsto w^\top f_A(x) + w_0\}$ and $\mathcal{H}_B = \{x \mapsto w^\top f_B(x) + w_0\}$ be the classes of linear functions of the representations of models A and B. For $h_B(x) = w^\top f_B(x) + w_0$, substituting $f_B = Tf_A$ gives $h_B(x) = (T^\top w)^\top f_A(x) + w_0$ — the same function written through the representation of model A; conversely, $h_A(x) = v^\top f_A(x) + w_0$ equals $(T^{-\top} v)^\top f_B(x) + w_0$, and it is precisely in the second inclusion, from $\mathcal{H}_A$ into $\mathcal{H}_B$, that the invertibility of T is used. Hence $\mathcal{H}_A = \mathcal{H}_B$: the representation spaces of both models admit the same set of linear scoring functions, and consequently the same set of orderings of prompts. The AUC of a probe for a given label depends only on how its score orders the prompts. The carried probe is not merely no worse than the original probe in model A: on every prompt its score coincides with the score of the original, so the AUC transfer gap is zero for all labels at once. □

Consequence for comparing binary prompt labels. Since, by Observation 3, under linear identifiability, that is, when the representations of two models are related by an invertible linear map, the AUC transfer gap vanishes for *every* label at once, a difference between the AUC transfer gaps of two labels does not allow one to conclude that one of these labels is geometrically more universal than the other. The claim "label X has a smaller AUC gap than label Y, hence X is more universal" was made by us in an early version of this project. Linear identifiability is sufficient for a zero AUC transfer gap but not necessary: Lemma 7 on label decomposition gives a mechanism under which the AUC transfer gap is zero even without linear identifiability: the AUC of the carried probe depends on label transport, defined in Appendix B.8, and not on the transfer of the source model's label direction.

B.7 TRANSPORT LAW

Theorem 4 (Transport law) in canonical variables. *Let u_i, v_i be the canonical variables of a pair of representations of two models, normalised to unit variance, with $\mathrm{Corr}(u_i, v_i) = \rho_i$ and uncorrelated cross pairs (canonical variables with different indices are uncorrelated). Let the probe's score in the model where the probe is trained decompose as $s = \sum_i c_i u_i$, and let $\hat{s} = \sum_i c_i \rho_i v_i$ be the population-optimal linear reconstruction of this score from the representation of the other model of the pair. Then*

$$\mathrm{corr}(s, \hat{s}) = \sqrt{\frac{\sum_i c_i^2 \rho_i^2}{\sum_i c_i^2}} = \sqrt{\mathbb{E}_\omega[\rho^2]}, \qquad \omega_i = \frac{c_i^2}{\sum_j c_j^2}.$$

Proof. The cross pairs are uncorrelated, therefore $\mathrm{Cov}(s, \hat{s}) = \sum_i c_i (c_i \rho_i) \mathrm{Cov}(u_i, v_i) = \sum_i c_i^2 \rho_i^2$, whereas $\mathrm{Var}(s) = \sum_i c_i^2$ and $\mathrm{Var}(\hat{s}) = \sum_i c_i^2 \rho_i^2$. After substitution into the definition of correlation, the factor $\sum_i c_i^2 \rho_i^2$ in the numerator cancels against the square root of the same factor in the denominator. □

Status and consequences. The transport law is not new: canonical analysis goes back to Hotelling (1936), and the fact that the regression of one model's representation on another's decomposes into separate canonical components is set out in textbooks of multivariate statistics. In the canonical basis the cross-covariance of the canonical variables of the two models is diagonal, therefore the best linear predictor of the canonical variable u_i from the representation of the other model equals $\rho_i v_i$: each canonical variable of one model corresponds to one canonical variable of the other, and they do not mix with the rest. The transport law is obtained by summing these predictors with weights c_i — the coordinates of the probe's score in the canonical basis. We use two of its

consequences: (1) $\mathrm{corr} \leq \rho_1$, that is, the transfer of the source model's label direction does not exceed the first canonical correlation, for any label; (2) for the probes of two labels in the same model pair, the transfer of the source model's label direction differs exactly to the extent that their probe profiles ω differ, that is, the distributions of the probe's score over the canonical directions. A third consequence of the transport law holds algebraically: the participation ratio of the probe profile $1/\sum_i \omega_i^2$ and the transfer of the source model's label direction are different functionals of the probe profile, so low dimensionality of the profile by participation ratio and high transfer of the source model's label direction are different properties.

Premises, and which of them does not hold. The transport law assumes that the canonical variables have unit variance, the cross pairs are uncorrelated, the canonical basis is complete, and the linear reconstruction of the probe's score from the representation of the other model of the pair is population-optimal. By construction of canonical correlation analysis (CCA), all premises of the transport law hold *on the training rows of the fold on which the CCA itself is fitted*; on rows outside the fold the diagonality premise, that is, uncorrelatedness of the cross pairs of canonical variables, does not hold. Therefore in §4.2 we report not an upper bound on the transfer of the source model's label direction by the transport law, but the agreement of two estimators of one and the same regression of the probe's score in its source model on the target model's features: the stitch and the canonical plug-in estimate by the transport law. In §4.2 we measure how strongly this diagonality is violated outside the fold, rather than postulating that there is no violation.

B.8 Decomposition of the transfer of the source model's label direction by the classes of this label

Lemma 7 (label decomposition) in expanded form. *Let s be the probe's score in its source model, where the probe is trained, $\hat{s}$ the score of the same probe in the target model after the stitch, and y the label on which the probe is trained, with share π of prompts labelled 1, the same for both scores. For each of the two scores denote by Δ the difference of the means of this score across the classes of the label y, by $\sigma_w^2 = \pi\,\mathrm{Var}(\cdot \mid y{=}1) + (1-\pi)\,\mathrm{Var}(\cdot \mid y{=}0)$ the pooled within-class variance, and $d' = \Delta/\sigma_w$; let r_w be the pooled within-class correlation of the pair $(s, \hat{s})$, that is, the within-class transport. Then*

$$\mathrm{corr}(s, \hat{s}) = \frac{r_w + \pi(1-\pi)\, d'_s\, d'_{\hat{s}}}{\sqrt{\big(1 + \pi(1-\pi)d'^2_s\big)\big(1 + \pi(1-\pi)d'^2_{\hat{s}}\big)}},$$

and the share of between-class variance in the total variance of each of the two scores s and $\hat{s}$ equals $\pi(1-\pi)d'^2/(1+\pi(1-\pi)d'^2)$.

Proof. Each score is a mixture of the two classes of the label y, therefore $\mathrm{Var}(s) = \sigma_{w,s}^2 + \pi(1-\pi)\Delta_s^2$ and $\mathrm{Cov}(s, \hat{s}) = \mathrm{Cov}_w(s, \hat{s}) + \pi(1-\pi)\Delta_s\Delta_{\hat{s}}$ by the law of total covariance conditioned on y. Substitute these decompositions into the definition of the correlation $\mathrm{corr}(s, \hat{s})$ and divide the numerator and the denominator by $\sigma_{w,s}\sigma_{w,\hat{s}}$. □

Rank form. AUC is the probability that a random positive prompt, that is, a prompt labelled 1 by the label y, receives a higher score from the probe than a random negative prompt. AUC is defined on positive–negative pairs of prompts, and the share of such pairs among all pairs of prompts equals $2\pi(1-\pi)$. From this definition of AUC it follows that

$$\mathrm{AUC}_s - \mathrm{AUC}_{\hat{s}} = P(s_i > s_j,\ \hat{s}_i < \hat{s}_j) - P(s_i < s_j,\ \hat{s}_i > \hat{s}_j), \qquad y_i = 1,\ y_j = 0,$$

that is, the AUC loss of the probe under transfer equals the *net* share of positive–negative pairs in which the stitch flips the order of the probe's scores: from the share of pairs where transfer ranks the negative prompt above the positive one contrary to the score in the source model, the share of pairs with the reverse flip is subtracted. In the rank form, the pair of quantities (τ, r_w), that is, label transport and within-class transport, corresponds to the concordance of the scores $(s, \hat{s})$ on positive–negative pairs and the same concordance of the scores on within-class pairs of prompts. We took the grouping of pairs of prompts by the subpopulations of prompts into which the prompts of the pair fall (for us, the classes of the label y) from Sudjianto & Liu (2025): the authors of that work decompose AUC itself over clusters of prompts in this way.

What the lemma says and what it does not say. Lemma 7 says that the transfer of the source model's label direction r mixes two transports: label transport ($\tau = d'_{\hat{s}}/d'_s$) and within-class transport (r_w). If the label y is a guard's verdict and the share of unsafe answers of the source model by this guard is small, the weight of label transport in the mixture, that is, the share of between-class variance in the total variance of the probe's score, amounts to a few percent, therefore $r \approx r_w$. The lemma also says that the AUC of the carried probe depends on label transport τ, which is not bounded by one. The lemma does not say that label transport τ is close to one: this closeness is a property of our data (§4.2). Nor does the lemma say that a small net AUC loss under transfer means a small number of positive–negative pairs in which the stitch flips the order of the probe's scores: flips in favour of the positive prompt and in favour of the negative prompt can cancel each other out, and on our data, as the rank form in the section on the transport law shows (§4.2), they largely cancel.

C $\mathrm{GL}(D)$ IN THE FULL HIDDEN-STATE DIMENSION: WHAT SURVIVES TRUNCATION

In §2 we require representation comparison statistics to be invariant to $\mathrm{GL}(D_A) \times \mathrm{GL}(D_B)$ in the full hidden-state dimension D. However, in all controls in the main text (procedure floors, that is, AUC transfer gaps between a model and its twin, in §4.1, sensitivity checks in §2) we apply a known invertible map to the cache of hidden states *already* truncated to k principal components. These controls therefore confirm the invariance of representation comparison statistics only to $\mathrm{GL}(k)$. Truncation to k principal components is itself not $\mathrm{GL}(D)$-covariant: per-coordinate z-normalisation is not invariant even to rotations, and the subspace of the first k principal components depends on which axes of the raw hidden state the known map stretches. Moreover, for $D > N$ on a sample of N rows all representations of full row rank lie in one $\mathrm{GL}(D)$ orbit, so the requirement of exact $\mathrm{GL}(D)$ invariance of representation comparison statistics on a sample is void. It is meaningful to ask only how sensitive representation comparison statistics are to invertible maps in the full dimension.

In this appendix we apply a known map to the model's raw hidden states before z-normalisation, PCA and all other steps of the protocol, and run the strict protocol on the mapped hidden states. We check the identity map (it must reproduce the unmapped principal-component cache), a random rotation (to which only per-coordinate z-normalisation is not invariant) and anisotropic maps with condition number 10 and 10^3. We compare the mapped hidden states of a model with the unmapped cache of the same model at p=0.75 and at the same k as in the main text:

	ident.	rotation	aniso 10	10^3
geom. gap	$+0.0000$	$+0.0005$	$+0.0008$	$+0.0029$
mean ρ	1.000	0.952	0.900	0.790
shift ρ_1	$+0.0000$	-0.0003	-0.0034	-0.0116
shift $d_{\mathrm{shared}}(\rho \geq 0.5)$	$+0.0$	$+0.2$	-1.0	-4.1

The identity map reproduces the unmapped cache up to zero. Under a map in the full dimension the geometry AUC gap stays within the procedure floor, and the first canonical correlation ρ_1 in the pairs of the mapped model with the other models shifts by at most 0.018. The subspace of the first k principal components of the model's hidden state, kept by truncation, is unstable under an invertible map in the full dimension: a random rotation alone replaces some of these components. Under the anisotropic map with condition number 10^3 the mean canonical correlation between the mapped and unmapped projections of one and the same model onto the first k principal components falls to 0.790, and the dimension $d_{\mathrm{shared}}(\rho \geq 0.5)$ of the subspace shared by the two models, built without any binary prompt label, shifts by -16% in the pairs of the mapped model with the other models. Therefore the quantities tied to the guard's verdict on the model's answer (AUC gap, transfer of the source model's label direction, ρ_1, label transport τ) are $\mathrm{GL}(D)$-stable within the procedure floor. The quantities describing the subspace shared by the two models without a binary label (the dimension $d_{\mathrm{shared}}(\rho \geq 0.5)$, linear CKA on features after whitening, the depth profile of this dimension from §4.1) are invariant only to $\mathrm{GL}(k)$ within the first k principal components of each model. The choice of the first k principal components itself depends on the basis of the hidden state, so we report these quantities with a caveat about this dependence.

D SENSITIVITY TO TRUNCATION AND TO THE TRAINING-ROW BUDGET

All numbers in the main text are computed after truncating hidden states to one and the same number of principal components k and on all training rows of the fold; both the number k and the number of training rows were chosen by us. A fixed k keeps a different share of hidden-state variance in models with different hidden-state dimension: in the eight models of set A the first 64 principal components explain 0.53–0.83 of the variance. Reaching 90% of the variance takes from $k = 123$ to 847 principal components, and the Spearman coefficient between this number of components and the dimension D is $+0.96$. A single cache of hidden states truncated to principal components at $k = 895 = D_{\min} - 1$, where $D_{\min}$ is the smallest dimension among the eight models of set A, covers any smaller k: truncation to fewer components reduces to an exact column slice of this cache. We deliberately fix the probe regularisation constant γ (the inverse regularisation strength of logistic regression) and the stitch ridge coefficient λ_{ridge} at all k; the stitch here is a ridge map from the target model's space into the source model's space, fitted on the same prompts without guard verdicts. If these constants are scaled as $\gamma \propto 1/k$, $\lambda_{\text{ridge}} \propto k$, all curves along the k axis and along the training-row budget axis below reproduce to the third decimal the curves obtained with fixed constants, except the procedure floor, that is, the AUC transfer gap between a model and its twin, at $k \leq 64$: about 51% of the procedure floor comes from the ridge regularisation of the stitch.

The k axis ($k \in \{32, 64, 128, 256, 512, 895\}$, five relative-depth positions). The AUC of the self arm, that is, of the probe trained and evaluated in the target model's space on the guard's verdict on this model's answer, reaches its maximum at $k = 128$ at each of the five positions and falls by 0.058 at $k = 895$. The geometry AUC gap, defined in §**??**, stays within ± 0.012 at any k, but its sign depends on k: over pairs of different models it crosses zero near k=128 and is negative from k=256 on, and under both strict masks it is negative already at k=128 (appendices I.3 and I.1). But the procedure floor moves from $+0.023$ at $k = 32$ to -0.043 at $k = 895$. Therefore the prediction of Observation 3 of a zero geometry AUC gap is tested by the procedure with sufficient precision, judging by the spread of the procedure floor, only at $k \leq 256$. The transfer of the source model's label direction r from the transport law (Theorem 4) falls monotonically as k grows (from 0.82 to 0.47). Label transport τ, by contrast, grows (from 0.95 to 1.14), and the AUC loss under transfer changes sign. The conclusion of §4.2 — the measured AUC loss is smaller than the one that follows from the transfer r — does not depend on k: as k changes, only how much smaller the measured loss is changes. The first canonical correlation ρ_1 of a model pair barely changes as k grows. The dimension of the subspace shared by the two models $d_{\text{shared}}(\rho \geq 0.5)$ has no plateau in k at any of the five positions. The order of the five positions given by each quantity of this paragraph (AUC of the self arm, geometry AUC gap, procedure floor, transfer r, label transport τ, AUC loss, first canonical correlation, dimension $d_{\text{shared}}(\rho \geq 0.5)$) is the same at any k.

The budget axis (relative depth p=0.75, fraction of the fold's training rows $\varphi \in \{1, 0.5, 0.25\}$; when training rows are subsampled, the share of prompts labelled 1 by the guard's verdict on the model's answer is preserved; three repeats agree to the third decimal). The control quantity on the budget axis is the ratio of the number of principal components to the number of training rows k/n_{train}. If, at a ratio k/n_{train} equal to 0.1, the fold's training rows are subsampled fourfold, the AUC of the self arm falls by 0.035, while label transport τ does not change. At a ratio k/n_{train} equal to 0.4, fourfold subsampling lowers the AUC of the self arm by 0.087, and the dimension $d_{\text{shared}}(\rho \geq 0.5)$ shrinks by two thirds. At a ratio k/n_{train} equal to 0.7, the out-of-fold spectrum of canonical correlations breaks down (ρ_1 falls from 0.98 to 0.36, and the dimension $d_{\text{shared}}(\rho \geq 0.5) \to 0$), and the geometry AUC gap, for the only time in the whole sweep over k and the training-row fraction, leaves the procedure floor in the direction of an AUC loss. At fixed k the training-row budget affects models of any dimension D alike: the shifts under subsampling in models from different dimension classes agree to the second decimal. The dimension D acts through the choice of k, not directly through the number of training rows n. With the corpus size $N = 3306$ prompts the limit for k lies around $250 - -500$, and in models of 8B and above truncation to this limit keeps less than half of the variance of their hidden state.

E TWO AVERAGED LABELS

In an early version of this project we reported a zero AUC transfer gap for the label averaged over the other models `theta_avg` and took this zero gap as a sign of the geometric universality of this label. Why a zero AUC gap is not evidence of the geometric universality of the label `theta_avg`, we explain in §4.1. We binarised both averaged labels by majority and ran them, at the same k at which the numbers in the main text are computed, through the same pipeline of three comparison arms (self, label-only and transfer), procedure floor (the AUC transfer gap between a model and its twin obtained by a known invertible map) and correspondence null (the AUC transfer gap through a stitch fitted on rows with the prompt pairing destroyed) as the verdicts of individual guards: the label averaged over guards on the model's own answer, `own_avg` (the mean of the verdicts of ten guards on the model's own answer, ≥ 0.5), and the label averaged over the other models, `theta_avg` (the mean of `own_avg` of the other seven models on the same prompt, ≥ 0.5). The label `theta_avg` is by construction almost independent of the model: the agreement κ between the values of the label `theta_avg` in different models is 0.71–0.89 against 0.10–0.60 for the label `own_avg`.

	label gap	geom. AUC gap	τ	AUC loss
ten guards, mean	0.108	+0.0028	0.93	0.021
`own_avg`	0.128	+0.0022	0.94	0.019
`theta_avg`	0.011	+0.0096	0.97	0.009

The label AUC gap of the label `theta_avg` equals 0.011 because the subset of prompts labelled 1 by `theta_avg` is almost the same in all models. The geometry AUC gap of the label `theta_avg` coincides with the value that the decomposition of the AUC transfer gap into the label and geometry AUC gaps predicts for any binary prompt label. Therefore the geometry AUC gap of `theta_avg` says nothing about the geometry of unsafe answers in the models' representations beyond what the verdicts of ten guards on the models' answers give.

Continuous averaged labels and sensitivity to the binarisation threshold. Before majority binarisation both averaged labels are continuous by construction: `own_avg` equals the fraction of the ten guards that judged the model's own answer unsafe, and `theta_avg` equals the mean of such fractions over the other models on the same prompt. Therefore both labels can be binarised at many thresholds, not only at 0.5. First we did without binarisation: we trained a ridge probe on each continuous averaged label and measured the quality of this probe by the concordance index C. The concordance index equals the fraction of prompt pairs with different label values in which the ridge probe's score orders the prompts in the same way as this label; for a binary label the concordance index coincides exactly with AUC. For the continuous averaged labels the three comparison arms, the procedure floor and the correspondence null are built in the same way as for labels binarised by majority. Without binarisation the conclusions of this appendix about the label and geometry AUC gaps do not change. The label C-gap, that is, the same difference in quality between the self and label-only probes as in the label AUC gap, but computed with the concordance index, lies across relative depths within 0.082–0.110 for the label `own_avg`, and within 0.004–0.004 for `theta_avg`. The geometry C-gap, that is, the difference in quality between the label-only and transfer probes by the concordance index, is at no depth larger than the procedure floor computed with the concordance index, for either label; where it leaves the interval of the floor, it lies below it: -0.0049–$+0.0019$ for `own_avg` and $+0.0039$–$+0.0074$ for `theta_avg`. The procedure floor in the same computation lies within $+0.0057$–$+0.0119$, the correspondence null in the same computation within $+0.180$–$+0.231$. The Spearman agreement between label values in different models is 0.34 for `own_avg` and 0.97 for `theta_avg`.

Then we ran both averaged labels, binarised at thresholds 0.2–0.8, through the same pipeline as the labels binarised by majority. Over the range of thresholds the share of prompts labelled 1 for `own_avg` varies from 0.50 to 0.13, and for `theta_avg` from 0.67 to 0.01. The label AUC gap at all thresholds stays within 0.104–0.135 for `own_avg` and 0.002–0.011 for `theta_avg`. The geometry AUC gap of both labels exceeds the procedure floor at no threshold. No conclusion of appendix E depends on the binarisation threshold.

F MODELS AND DEPTH GRID

Models. Set $\mathcal{A}$ is the eight classical-architecture models marked below; $\mathcal{B}$ keeps one model per vendor; $\mathcal{C}$ is $\mathcal{A}$ together with `qwen3_8_27b`, which stands outside W_{mix19} as a single control.

model	vendor	params	L	D	architecture	sets
classical: dense decoder transformer, softmax attention, no MoE						
`qwen25_05b`	Qwen	0.5B	25	896	classical	$\mathcal{AC}$
`olmo2_1b`	AI2	1.0B	17	2048	classical	$\mathcal{BAC}$
`falcon3_1b`	TII	1.7B	19	2048	classical	—
`smollm3_3b`	HuggingFace	3.0B	37	2048	classical	$\mathcal{BAC}$
`falcon3_7b`	TII	7.0B	29	3072	classical	$\mathcal{BAC}$
`olmo2_7b`	AI2	7.3B	33	4096	classical	—
`qwen25_7b`	Qwen	7.6B	29	3584	classical	—
`llama31_8b`	Meta	8.0B	33	4096	classical	$\mathcal{AC}$
`qwen3_8b`	Qwen	8.2B	37	4096	classical	$\mathcal{AC}$
`qwen3_32b`	Qwen	32B	65	5120	classical	$\mathcal{BAC}$
`llama31_70b`	Meta	70B	81	8192	classical	$\mathcal{BAC}$
alternative: a modification of that setup						
`lfm2_1_2b`	LiquidAI	1.2B	17	2048	short gated convolutions in place of most attention layers, a few GQA layers	—
`falconh1_1_5b`	TII	1.6B	25	2048	attention and Mamba-2 run in parallel inside every block	—
`lfm2_2_6b`	LiquidAI	2.6B	31	2048	as `lfm2_1_2b`	—
`granite4h_micro`	IBM	3.2B	41	2048	Mamba-2 and attention hybrid, most layers without attention	—
`granite4h_tiny`	IBM	6.9B	41	1536	as `granite4h_micro`, with a sparse mixture of experts	—
`falconh1_7b`	TII	7.6B	45	3072	as `falconh1_1_5b`	—
`nemotronh_9b`	NVIDIA	8.9B	57	4480	almost all layers Mamba-2, attention in a few layers only	—
`nemotronh_12b`	NVIDIA	12.3B	63	5120	as `nemotronh_9b`	—
`qwen3_8_27b`	Qwen	27B	65	5120	Gated DeltaNet linear attention, three linear blocks per full-attention block	$\mathcal{C}$

The first two groups, eleven classical and eight alternative models, are W_{mix19}; `qwen3_8_27b` is not one of the nineteen.

Depth grid. Index 0 of the stored tensors is the token embedding and is degenerate on this corpus: every prompt ends with the same chat-template suffix, so all 3306 rows are the same point. The last block output, index $L-1$, is the layer the language-model head reads and is the most task-specific. Relative depth therefore runs over blocks $1 \dots L-2$,

$$\text{layer}(p) = 1 + \text{round}\big(p\,(L-3)\big),$$

so $p = 0$ is the first block after the embedding and $p = 1$ the penultimate stored layer. L is the number of stored tensors, the embedding plus the output of every block (`num_hidden_layers`+1; for the hybrids every Mamba-2, attention or MLP block counts, as their configurations count them), and D the hidden dimension; both are in the model table above. The grid is the fourteen positions $p \in \{0,\ 0.1,\ 0.2,\ 0.25,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.75,\ 0.8,\ 0.9,\ 1\}$ together with $p = 1.1$, which addresses the final layer $L-1$ under the earlier convention and is reported for continuity. The grid is not uniform: 0.25 and 0.75 are kept because the earlier five-point grid used them.

model	p=0	p=0.5	p=1	p=1.1
qwen25_05b	1	12	23	24
olmo2_1b	1	8	15	16
smollm3_3b	1	18	35	36
falcon3_7b	1	14	27	28
llama31_8b	1	16	31	32
qwen3_8b	1	18	35	36
qwen3_32b	1	32	63	64
llama31_70b	1	40	79	80
falcon3_1b	1	9	17	18
olmo2_7b	1	16	31	32
qwen25_7b	1	14	27	28
lfm2_1_2b	1	8	15	16
lfm2_2_6b	1	15	29	30
falconh1_1_5b	1	12	23	24
falconh1_7b	1	22	43	44
granite4h_micro	1	20	39	40
granite4h_tiny	1	20	39	40
nemotronh_9b	1	28	55	56
nemotronh_12b	1	31	61	62
qwen3_8_27b	1	32	63	64

Guards. Ten guards, two sizes from each of five vendors.

vendor	smaller	larger
Llama Guard (Inan et al., 2023; Grattafiori et al., 2024; Meta AI, 2025)	Guard 3, 8B	Guard 4, 12B
Qwen3Guard (Zhao et al., 2025a)	0.6B	8B
OpenGuardrails (Wang & Li, 2025)	4B	15B, release 2510
XGuard (Lin et al., 2026)	0.6B	8B
GPT-OSS-Safeguard (OpenAI, 2025)	20B	120B

G How we make models comparable

Models are not directly comparable, and we use three devices. *Principal components.* This is the reduction to a common k assumed in §2. We reduce the different hidden-state dimensions D to a common number of principal components k. Truncation to principal components is itself not covariant under invertible changes of coordinates, and for $D > N$ every representation of full row rank lies in one orbit of $\mathrm{GL}(D)$, so exact invariance on the sample is empty and only the question of sensitivity is meaningful: we apply a known invertible map to hidden states of full dimension and vary k (appendices C and D).

Relative depth. We reduce the different numbers of blocks to the relative depth p. The correspondence "depth p in the source model corresponds to depth p in the target model" is an assumption; we test it by fitting the stitch off the depth diagonal (appendix J.5).

Model-invariant label. The verdict of one guard on the answers of different models singles out different prompts, and where this mismatch would by itself lower the transfer, we take a label common to both models of the pair; its label AUC gap is zero by construction. There are two such labels: the guard's verdict on the reference model's answer, assigned to all models, with the reference model varied, and the consensus prompt label, equal to 1 when no fewer than half of the eighty verdicts of ten guards on the answers of eight models judge the answer of its own model unsafe.

The share of unsafe answers varies across "model $\times$ guard" cells from 0.031 to 0.672 and orders both the probe profile and the transfer of the label direction, so we match the controls on it. The rank correlations behind this are -0.79 and $+0.88$ over ten guards, and weaker over the 80 "model $\times$ guard" cells (appendix J.3).

H SENSITIVITY, RESOLUTION, AND ATTACK SPECIFICITY

If one compares the *shape of the spread* — the transfer of the generalised eigendirections of the covariance pair of the subset of prompts with a successful attack and the remaining prompts — this subset does not beat the control half-space of the reference model, that is, a connected control with the same share of prompts labelled 1: there is nothing attack-specific. But this very statistic distinguishes the planted shared direction only from an inflation of the variance along it by a factor of 2.0–2.4, and moreover it is by construction blind to the difference of class means — to exactly what the probe finds. If instead one looks along the probe direction, the answer is the opposite: the label transport of the guard's verdict is higher than that of the control slab matched to it both in the share of prompts labelled 1 and in decodability (appendix J.1), and the guard-verdict axis lowers the ID contrast more than the axis orthogonal to it with the same variance (appendix L.4). A shared linear direction thus exists and transfers partially; what the transfer of the generalised eigendirections leaves unresolved is the question of whether it is attack-specific, and the probe direction answers it. The three subsections below give the measurements behind this summary; §4.3 and §**??** carry the short form.

H.1 LIMITS OF APPLICABILITY

The numbers depend on three of our choices, and we measure the dependence. In the number of principal components k: the AUC of the self arm peaks at $k = 128$, while the geometry AUC gap stays inside the procedure floor over the whole range of k and changes sign together with it, so the sign of this gap cannot be read, while the comparison with the label AUC gap does not depend on k (Appendix I.3). In the training-row budget, the controlling quantity is the ratio of k to the number of training rows (Appendix D).

The third limit is more substantial than the first two. On the other prompt source the transfer of the source model's label direction drops roughly by half. The prediction by the transport law follows the measurement with the same accuracy with which the direct measurement on its own prompt source predicts the other, that is, the transport law loses nothing beyond what the quantity itself loses (Appendix J.2, Fig. 15). A shift in the share of unsafe answers between the prompt sources does not explain this drop. Transfer is thus a property of the triple "model pair, probe, prompt distribution", and the canonical spectrum of a pair is defined only relative to the input distribution. For monitoring this means that transfer measured on one traffic says nothing about new traffic.

H.2 WHAT THE SETUP COULD HAVE SEEN

A result that reports finding nothing can be read only together with the resolution of the setup, and we measure it by planting a known structure into both models and running the whole pipeline on the planted data. We mix into both models a shared direction of a given strength — a random direction in the source model's space and the same direction read in the space of the target model through the stitch — and run the same pipeline on the modified data. The strength from which the pipeline finds it is the resolution of the setup: below it, "we found nothing" does not mean "it is not there". On set W_mix19 at $k = 64$ both planted structures are detected in all 6 settings — combinations of relative depth $p \in \{0.25, 0.5, 0.75\}$ and rank 3 or 5. The subspace shared by the two models is detected when it inflates the variance of the subset of prompts with a successful attack along itself by a factor of 1.6–2.0, and one shared direction when it inflates it by a factor of 2.0–2.4 (Fig. 4). Across pair groups the required inflation barely differs: for pairs of two classical-architecture models it is 1.6–2.3 for the subspace and 2.0–2.5 for the direction, for pairs of two alternative-architecture models 1.6–2.0 and 2.1–2.4, and for mixed pairs 1.6–1.9 and 2.1–2.4.

H.3 IS THIS SPECIFIC TO ATTACKS

The two statistics give different answers.

The transfer of the generalised eigendirections of the covariance pair of the subset of prompts with a successful attack and the remaining prompts, under the guard's verdict on the reference model's answer, assigned to all models, is no higher than under the control half-space of the reference model with the same share of prompts labelled 1 — the connected control, that is, the label defined by the

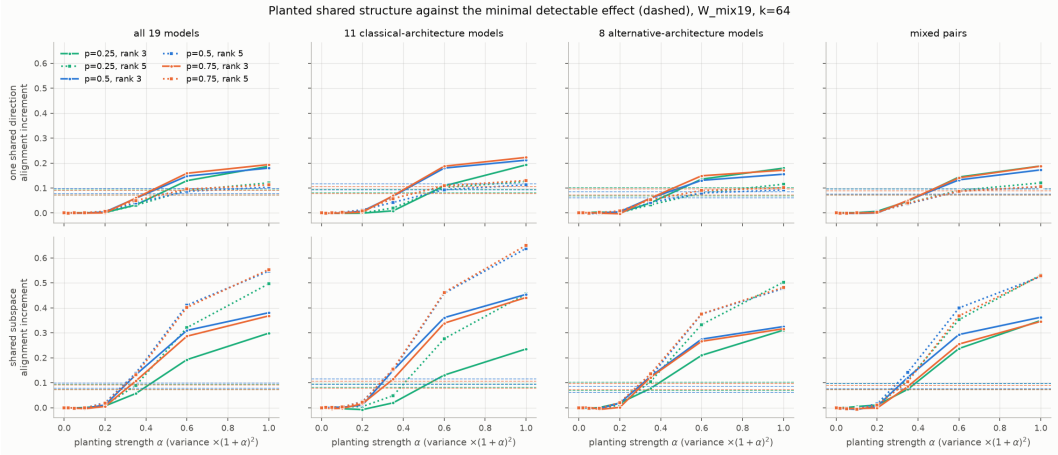

Figure 4: Alignment increment from planting a known shared structure into both models, against the planting strength α (variance along the planted structure multiplied by $(1+\alpha)^2$); W_mix19, $k = 64$; columns, all 19 models, the 11 classical-architecture models, the 8 alternative-architecture models and the pairs that mix the two families; top row, one shared direction; bottom row, a shared subspace of the reported rank; solid lines rank 3, dotted lines rank 5, colour by relative depth p, dashed horizontal lines of the same colour are the minimal detectable effect of the procedure in that setting (power 80%, wild cluster bootstrap). Read: every curve crosses its dashed line, so both structures are detected in all six settings of each group; the subspace crosses earlier than the single direction.

half-space cut off by a random direction in the reference model and assigned to all models. On a set of 32 binary labels, no advantage remains for the guard's verdict on the reference model's answer, assigned to all models. This statistic is computed from covariances centred within classes, so the difference of means does not enter it: a planted mean shift changes it by $1.4 \cdot 10^{-4}$, and the leading generalised eigenvalue by zero.

The probe direction gives a different answer. Label transport for the guard's verdict is higher than for the control slab matched to it both in the share of prompts labelled 1 and in decodability, by $+0.176$–$+0.256$ (Appendix J.1). The guard-verdict axis, estimated on the half of the rows that did not enter the computation, lowers the contrast in intrinsic dimension more than an axis orthogonal to it with the same variance (Appendix L.4). The specificity of the attack thus shows up in the difference of means, and the second-moment statistic is blind to it by construction (Task 4 of §1).

I ROBUSTNESS OF THE MAIN FINDING

Unless stated otherwise, all measurements in this appendix are made on set A (eight classical-architecture models) with the number of principal components of the hidden state k=64, relative depth p=0.75 and N=3306 prompts of the corpus. We determined the outcome of each measuring run below by a reading rule written before the data. The exception is the second attempt of ablation G5 on model scale as an axis with a control for vendor (Appendix J.4): we wrote its reading rule already knowing the outcome of the first attempt of ablation G5. Where a reading rule gave an outcome, that outcome is set in italics.

I.1 STRICT MASKS: WITHOUT EMPTY MODEL ANSWERS AND WITHOUT RECOVERED GUARD VERDICTS

We keep all N=3306 rows of the corpus thanks to two conventions: we count each guard's verdict on an empty model answer as a "safe" verdict, and we recover part of the guard verdicts from the first token of the guard's answer when the whole guard answer could not be parsed. The mask `gradable` removes the corpus rows where the model gave an empty answer; the mask `judged` additionally removes the corpus rows with a recovered guard verdict. Under each of the two strict masks we recomputed all measurements on set A.

	all rows, N=3306	no empty, N=3275	no recov., N=3223
label gap	0.0807–0.1090	0.0815–0.1112	0.0832–0.1114
geom. gap	$+0.0008 \ldots + 0.0054$	$+0.0011 \ldots + 0.0053$	$+0.0008 \ldots + 0.0057$
floor	$-0.0115 \ldots + 0.0429$	$-0.0368 \ldots + 0.0611$	$-0.0249 \ldots + 0.0651$
inside	14 of 14	14 of 14	14 of 14
optimum over k	k=128, 0.7786	k=128, 0.7789	k=128, 0.7789
geom. gap, k=128	$+0.0003$	-0.0004	-0.0002
sign change	between k=128 and 256	between k=64 and 128	between k=64 and 128
transport law: Pearson, slope	0.9982, $+0.987$	0.9984, $+0.979$	0.9981, $+0.976$
geom. gap, lg3 and og4b	$+0.0109$, $+0.0178$	$+0.0134$, $+0.0179$	$+0.0099$, $+0.0181$

Every main finding of the paper checked in the table above remains valid under both strict masks. The only caveat concerns the sweep over truncation, that is, the sweep over the number of principal components k. Under the strict masks the point where the geometry AUC gap of model pairs crosses zero moves from the interval just above k=128 to the interval just below, but the sweep over the number of principal components k does not resolve this shift: by this sweep the zero crossing of the geometry AUC gap lies near k=128, and whether above or below cannot be said. The spread of the procedure floor, that is, of the AUC transfer gap between a model and its twin, over the positions of the depth grid is wider under the strict masks than on all rows of the corpus, so the check "geometry AUC gap inside the procedure floor at 14 positions of 14" is weaker under the strict masks. But the main finding of the paper about the geometry AUC gap rests on the size of the geometry AUC gap itself at the positions of the depth grid, and this size has hardly changed under the strict masks compared with all rows of the corpus.

I.2 ORDER OF BINARY PROMPT LABELS BY THE GEOMETRY AUC GAP ON THE TEST ROWS OF FOLD 0, WHICH THE PRINCIPAL-COMPONENT BASIS OF FOLD 0 DID NOT SEE

In §4.1 we claim that the difference between the AUC transfer gaps of different binary prompt labels cannot rank these labels by geometric universality. Here we ask a narrower question: which binary prompt labels have the larger geometry AUC gap if the label from the representation is built from the principal-component basis of the target model fitted on the training rows of fold 0, and the AUC of the three comparison arms is computed only on the 662 test rows of fold 0, which this basis did not see. We obtained the labels from the representation `pc0`, `pc5`, `pc20` by cutting at the median one principal component of the target model from the principal-component basis of fold 0, whose index stands in the name of the label, and the label from the representation `dir_top10` by cutting a random direction in the space of the first ten principal components of the same basis of the target model. The prompt-source label `corpus` is a property of the prompt text, that is, which prompt source the prompt was taken from; `random` is a control random label; and the labels `guard_lg3`

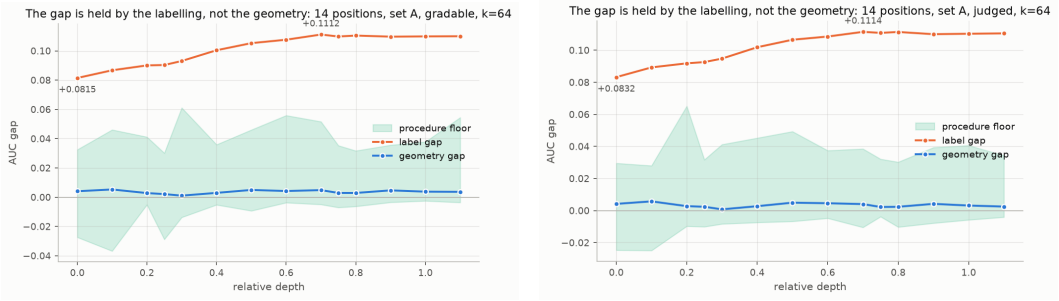

Figure 5: Label AUC gap and geometry AUC gap at 14 relative-depth positions next to the procedure floor; left, strict mask `gradable` (N=3275, without rows with an empty model answer), right, strict mask `judged` (N=3223, additionally without rows with a recovered guard verdict), set A, k=64. Under both strict masks, at every relative-depth position the geometry AUC gap stays inside the procedure floor, while the label AUC gap lies above it.

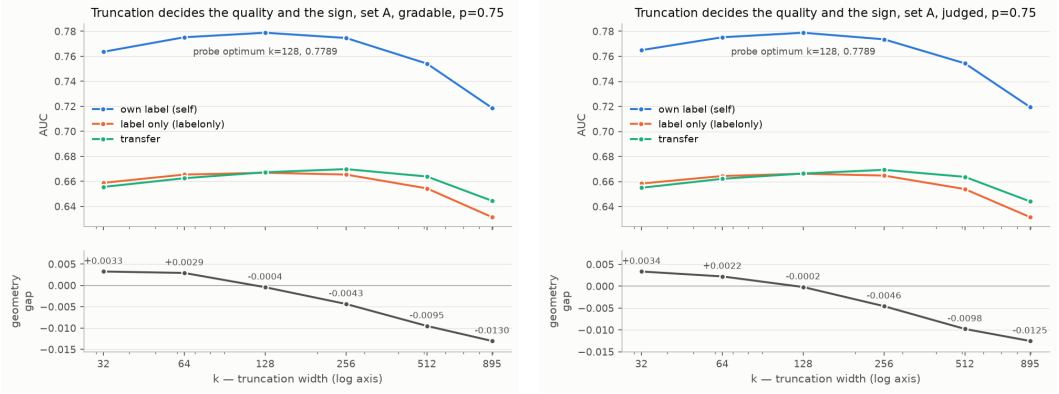

Figure 6: AUC of the three comparison arms (probe scores self, label-only and transfer) and the geometry AUC gap as a function of the number of retained principal components k; only pairs of different models, without the model-twin control pairs from which the procedure floor is computed, set A, relative depth p=0.75; left, strict mask `gradable` (N=3275, without rows with an empty model answer), right, strict mask `judged` (N=3223, additionally without rows with a recovered guard verdict). Under both strict masks the geometry AUC gap of pairs of different models is already slightly below zero at k=128 and still above zero at k=64.

and guard_og4b are the verdicts of the guards lg3 and og4b on the target model's answer; none of these four labels is taken from the representation.

label	self, all	self, fold 0	geom., all	geom., fold 0
corpus	0.9732	0.9739	+0.0026	+0.0021
random	0.4954	0.5107	+0.0033	+0.0017
guard_lg3	0.7642	0.7766	+0.0109	+0.0053
guard_og4b	0.7698	0.7721	+0.0178	+0.0174
pc0	0.9981	0.9983	+0.0377	+0.0383
pc5	0.9989	0.9991	+0.0562	+0.0532
dir_top10	0.9989	0.9987	+0.0601	+0.0556
pc20	0.9984	0.9989	+0.1199	+0.1240

Outcome by the rule: the leakage is negligible; the order of the guard verdict og4b and the labels from the representation by the geometry AUC gap is reinstated. Leakage is the fact that, when the AUC of the three comparison arms is computed over all rows of the corpus, the AUC estimate also includes the training rows of fold 0 on which the principal-component basis of fold 0 was fitted. We estimate the leakage of the training rows of fold 0 into the AUC estimate by the discrepancy of the AUC of the self arm: for the four labels from the representation, the AUC of the self arm on the test rows of fold 0, which the principal-component basis of fold 0 did not see, differs from the AUC of the self arm over all rows of the corpus by no more than 0.0005. Earlier we withdrew the finding about the order of the guard's verdict on the target model's answer and the labels from the representation by the geometry AUC gap, relying on labels from the representation built from the principal-component basis of each fold separately. But PCA is refitted in each fold and does not fix the sign of a principal component, so a label from the representation assembled over folds is made up of principal components of different folds pointing in different directions. Following the reading rule written before the data, we computed on the test rows of fold 0, which the principal-component basis of fold 0 did not see, four differences of geometry AUC gaps: between the label pc0 and the guard verdict og4b, between the label dir_top10 and the same guard verdict, between the labels pc5 and pc0, and between the labels pc20 and pc5. For each of the four differences of geometry AUC gaps we took the mean over ordered model pairs and the 95% bootstrap interval over the same pairs. We applied the same reading rule, written before the data, to five more configurations: to the two strict row masks on set A, to sets B and C, and to set A with twice the number of principal components k.

configuration	pairs	self shift	pc0−og4b	dir_top10−og4b	pc5−pc0	pc2
A, k=64, N=3306	56	0.0005	+0.0209 [+0.0121, +0.0300]	+0.0382 [+0.0337, +0.0428]	+0.0148 [+0.0038, +0.0264]	+0.0708 [+0.
A, k=64, N=3275	56	0.0007	+0.0224 [+0.0131, +0.0322]	+0.0351 [+0.0256, +0.0454]	+0.0175 [+0.0049, +0.0303]	+0.0684 [+0.
A, k=64, N=3223	56	0.0004	+0.0291 [+0.0189, +0.0393]	+0.0481 [+0.0411, +0.0555]	+0.0186 [+0.0067, +0.0309]	+0.0619 [+0.
B, k=64	20	0.0007	+0.0126 [+0.0036, +0.0218]	+0.0313 [+0.0254, +0.0378]	+0.0243 [+0.0076, +0.0422]	+0.0527 [+0.
C, k=64	72	0.0005	+0.0271 [+0.0190, +0.0352]	+0.0513 [+0.0437, +0.0592]	+0.0274 [+0.0151, +0.0398]	+0.0702 [+0.
A, k=128	56	0.0004	+0.0180 [+0.0102, +0.0260]	+0.0360 [+0.0312, +0.0409]	+0.0131 [+0.0033, +0.0234]	+0.0548 [+0.

The column "self shift" of the table gives the largest discrepancy between the AUC of the self arm on the test rows of fold 0 and the AUC of the self arm over all rows, and the four right-hand columns give the difference between the geometry AUC gap of the label named first in the column header and that of the label named second, averaged over ordered model pairs, with the bootstrap interval over the same pairs. The smallest lower end among all 24 bootstrap intervals (four differences of geometry AUC gaps in six configurations) equals $+0.0033$. Therefore, in all six configurations the guard verdict og4b on the target model's answer transfers with a smaller geometry AUC gap than the labels pc0 and dir_top10 from the target model's representation, and among the labels pc0, pc5 and pc20 from the target model's representation, the higher the index of the target model's principal component cut at the median, the larger the geometry AUC gap of the label. By the rule written before the data, we compare the geometry AUC gaps themselves, not the share of the lost headroom, that is, the geometry AUC gap divided by the difference between the AUC of the self arm and the AUC of random guessing: such a share would depend also on the AUC of the self arm, which is close to one for the labels from the representation and lower for the guard verdicts.

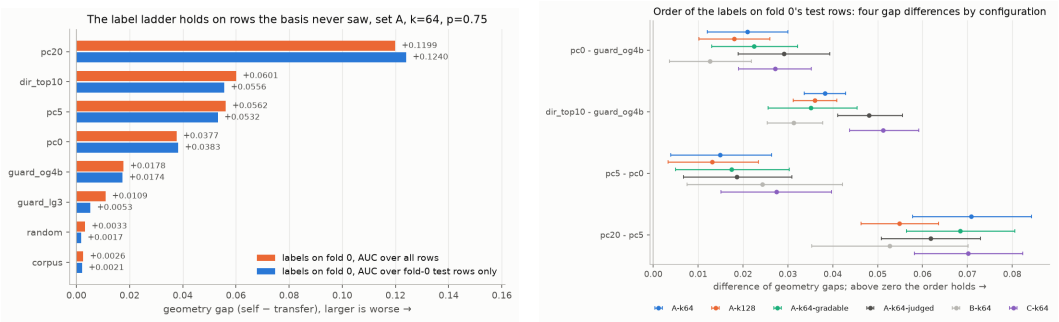

Figure 7: Left: geometry AUC gap of eight binary prompt labels (the prompt-source label, a random label, the verdicts of the guards `lg3` and `og4b` on the target model's answer, four labels from the representation), when the AUC is computed over all rows and when only over the test rows of fold 0, which the principal-component basis of fold 0 did not see; for each of these labels the two bars nearly coincide; set A, k=64, relative depth p=0.75. Right: four differences of geometry AUC gaps (`pc0` and `og4b`, `dir_top10` and `og4b`, `pc5` and `pc0`, `pc20` and `pc5`) averaged over ordered model pairs, with 95% bootstrap intervals in six configurations (set A over all rows and under the two strict row masks, sets B and C, set A with a larger number of principal components k); all intervals lie to the right of zero.

I.3 The sign of the geometry AUC gap depends on truncation; set A at k=128

The sweep over the number of principal components to which the features are truncated includes only pairs of different models, without the model-twin control pairs, where the twin is obtained by a known invertible map, from which the procedure floor is computed. Each value of the table of the sweep over the number of principal components is averaged over the 560 ordered model pair × guard cells of set A at relative depth p=0.75:

k	self ↑	label-only	transfer	geom. gap
32	0.7630	0.6593	0.6559	+0.0034
64	0.7741	0.6657	0.6629	+0.0028
128	**0.7786**	0.6673	0.6670	+0.0003
256	0.7731	0.6662	0.6700	−0.0038
512	0.7536	0.6552	0.6639	−0.0088
895	0.7171	0.6321	0.6442	−0.0122

The AUC of the self arm, that is, of the target model's probe trained on the guard's verdict on the target model's answer, is maximal at k=128, and near k=128 the geometry AUC gap also passes through zero; with more retained principal components the geometry AUC gap is negative and reaches −0.0122 at k=895. A negative geometry AUC gap means that the source model's probe, trained on the same guard's verdict on the source model's answer and carried by the stitch into the target model's space (the transfer arm), gives a higher AUC than the target model's probe trained on the guard's verdict on the source model's answer (the label-only arm). The sign of the geometry AUC gap changes within the swept range of the number of principal components. Hence the geometry AUC gap is not a property of the model pair, and without a named number of principal components k we claim nothing about the geometry AUC gap. We recomputed all measurements on set A in full with the number of principal components k=128; in the table they stand next to the measurements at the number of principal components adopted in the main text:

	k=64	k=128
label gap	0.0807–0.1090	0.0816–0.1113
geom. gap	+0.0008 . . . + 0.0054	−0.0007 . . . + 0.0060
floor	−0.0115 . . . + 0.0429	−0.0492 . . . + 0.0394
inside	14 of 14	14 of 14
transport law: Pearson, slope, median τ	0.9982, +0.987, 0.9414	0.9963, +0.957, 0.9626

With the number of principal components k=128 on set A, the spreads of the label and geometry AUC gaps over the 14 positions, that is, relative depths, do not overlap, and the transport law (Theorem 4) holds, but the OLS slope between the transfer of the source model's label direction achieved by the ridge stitch and the plug-in estimate of the same transfer by the transport law departs from one further than in the table column for the number of principal components adopted in the main text.

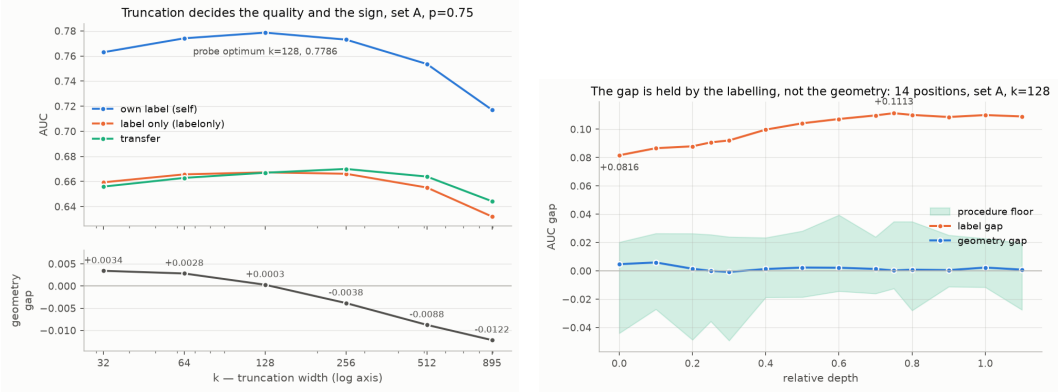

Figure 8: Left: AUC of the three comparison arms (probe scores self, label-only, transfer) against the guard's verdict on the target model's answer, and the geometry AUC gap as a function of k, the number of retained principal components; only pairs of different models, without the model-twin control pairs, set A, relative depth p=0.75. Right: profile of the label and geometry AUC gaps over positions, that is, relative depths, next to the procedure floor, set A, number of principal components k=128. Reading: on the left, the geometry AUC gap passes through zero near the same number of principal components at which the AUC of the self arm is maximal, that is, of the target model's probe trained on the guard's verdict on the target model's answer; on the right, the geometry AUC gap lies inside the procedure floor at all relative-depth positions.

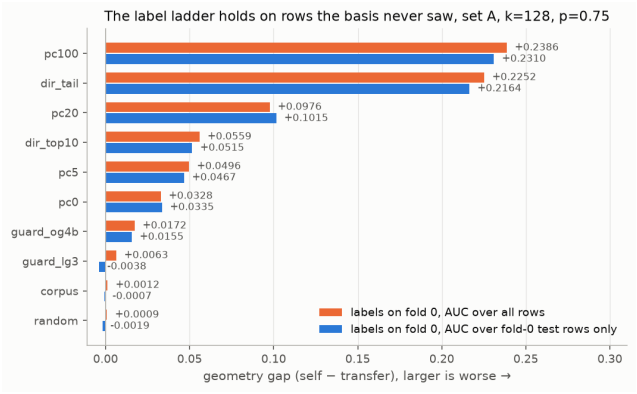

Figure 9: Geometry AUC gap of ten binary prompt labels (a random label, the prompt-source label, guard verdicts on the target model's answer, labels from the representation), set A, number of principal components k=128, relative depth p=0.75. Orange bars: the AUC of the label-only and transfer arms, whose difference gives the geometry AUC gap, is computed over all rows of the prompt corpus; blue bars: the AUC of the same arms is computed only over the test rows of fold 0, which the principal-component basis fitted on the training rows of the same fold did not see. Reading: the order of the random label, the prompt-source label, the guard verdicts and the labels from the representation by the geometry AUC gap on set A is the same as with the number of principal components k=64, and the largest geometry AUC gap belongs to the labels from the representation `pc100` and `dir_tail` — the two labels built from principal components with indices absent at k=64.

I.4 SET B: ONE MODEL PER VENDOR (ABLATION G2)

In set A two vendors are represented by several models: Qwen by three, Llama by two. In set B we kept one model per vendor (`olmo2_1b`, `smollm3_3b`, `falcon3_7b`, `qwen3_32b`, `llama31_70b`); these models give 20 ordered model pairs.

	set A, 56 pairs	set B, 20 pairs
label gap	0.0807–0.1090	0.0895–0.1275
geom. gap	$+0.0008 \ldots + 0.0054$	$+0.0025 \ldots + 0.0088$
inside	14 of 14	14 of 14
transport law: Pearson, slope	0.9982, $+0.987$ over 560	0.9978, $+0.980$ over 200

Outcome by the rule of ablation G2: the findings of set A hold on set B. The spreads of the label and geometry AUC gaps on set B over all relative-depth positions do not overlap. The geometry AUC gap on set B lies, at all relative-depth positions, inside the procedure floor computed on set B, that is, inside the spread of the geometry AUC gap over the model-twin control pairs at the same position. On set B the Pearson coefficient between the two estimates of the transfer of the source model's label direction, the one achieved by the ridge stitch and the plug-in estimate by the transport law, is above the threshold 0.99 of the rule of ablation G2. Hence the findings about the 56 pairs of set A do not depend on Qwen and Llama being represented in set A by several models. The label and geometry AUC gaps on set B are higher than on set A, but the rule of ablation G2 sets no conditions on the size of the label and geometry AUC gaps.

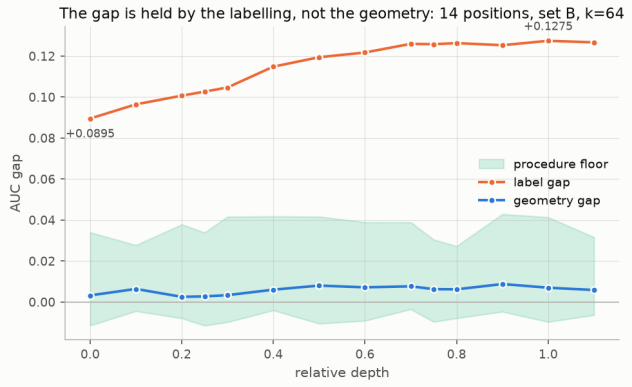

Figure 10: Profile of the label and geometry AUC gaps over 14 relative-depth positions compared with the procedure floor, that is, with the spread of the geometry AUC gap over the model-twin control pairs, set B, k=64. Reading: the geometry AUC gap lies inside the procedure floor at all relative-depth positions.

I.5 SET C: A HYBRID-ATTENTION MODEL

For set C we added to the eight models of set A the hybrid-attention model `qwen3_8_27b`: three Gated-DeltaNet blocks per one full-attention block. This model differs from the eight models of set A not only in attention: it is multimodal, whereas all models of set A are text decoders. Therefore a difference between the groups of pairs, even if the outcome by the rule were readable, could not be attributed to one of the two differences. Set C has nine models and 72 ordered model pairs. With a rule written before the data we check whether a probe separating the subset of prompts with a successful attack by the guard's verdict on the source model's answer from the other prompts transfers between a softmax-attention model and a hybrid-attention model as well as between two softmax-attention models. For this we take the difference of the medians of the geometry AUC gap between two groups of model pairs: the model pairs that include the hybrid-attention model, and the model pairs of set A. The rule for set C also sets an additional condition: if the medians of the label AUC gap of the model pairs that include the hybrid-attention model and of the model pairs of set A differ by more than 0.03, the outcome by the rule is "not readable".

pairs	pairs	cells	med. label	med. geom.
both from A	56	560	0.0999	+0.0021
with qwen3_8_27b	16	160	0.1916	−0.0143

Outcome by the rule: not readable. The medians of the label AUC gap of the model pairs that include the hybrid-attention model and of the model pairs of set A differ by $+0.0917$, more than the threshold on the difference of the medians of the label AUC gap in the rule for set C. The outcome "not readable" was already provided for in the record of the rule for set C made before the data, because in share of unsafe answers the model qwen3_8_27b differs from the eight models of set A: its share of unsafe answers averaged over the ten guards equals 0.048, theirs is 0.124–0.488. Under the outcome "not readable" the rule for set C forbids drawing a conclusion from the difference of the medians of the geometry AUC gap between the model pairs that include the hybrid-attention model and the model pairs of set A, including explaining this difference by hybrid attention. We report the difference of the medians of the geometry AUC gap between the model pairs that include the hybrid-attention model and the model pairs of set A without a conclusion: -0.0164, 95% $[-0.0354,\ -0.0014]$ by bootstrap over model pairs. For comparison with the difference of the medians of the geometry AUC gap, the procedure floor on set C, that is, the spread of the geometry AUC gap over the model-twin control pairs, equals $-0.0146 \ldots + 0.0305$.

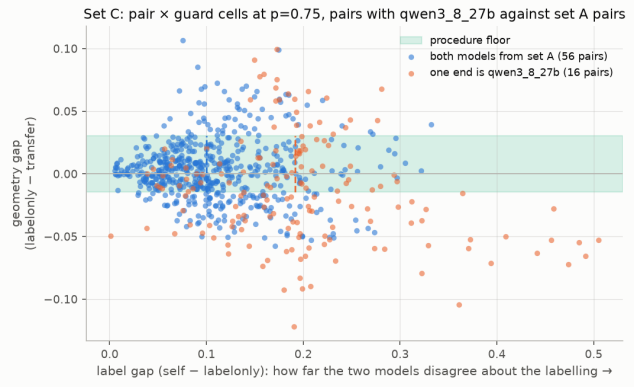

Figure 11: Ordered model pair × guard cells of set C at p=0.75: horizontally the label AUC gap, vertically the geometry AUC gap; blue — cells of model pairs of set A, orange — cells of model pairs with qwen3_8_27b, dashed lines — medians of the label AUC gap of these two groups of model pairs. Reading: the orange cells of model pairs with the hybrid-attention model are shifted to the right of the blue cells of model pairs of set A, towards a larger label AUC gap, so the outcome by the rule for set C is "not readable".

I.6 Averaged labels: what of Appendix E reproduces

We recomputed the label and geometry AUC gaps of both averaged labels at 14 relative-depth positions in the continuous and in the binary form of the averaged label: own_avg, the share of the ten guards that judged the model's own answer unsafe, and theta_avg, the mean of such shares over the other models on the same prompt. In the continuous form of the averaged label we replace the AUC in the label AUC gap by the concordance index C, and over relative-depth positions this label C-gap, that is, the difference between the self and label-only probes in the concordance index, lies within 0.0818–0.1102 for own_avg and 0.0038–0.0042 for theta_avg. In the binary form we binarise the averaged label at each binarisation threshold from the range 0.2–0.8, and at each binarisation threshold we average the label AUC gap over relative-depth positions; over binarisation thresholds it lies within 0.1041–0.1353 for own_avg and 0.0023–0.0110 for theta_avg. The Spearman agreement between the continuous averaged labels of different models equals 0.337 for own_avg and 0.969 for theta_avg. The claims of Appendix E about the label AUC gap of both averaged labels and about the Spearman agreement between the averaged labels of different models reproduce. As the procedure floor at a relative-depth position we take the spread of the geometry AUC gap over the model-twin control pairs, where the twin is the model's own features after a known invertible map, at the same relative-depth position, and in the binary form of the averaged

label also at the same binarisation threshold. The claim of the appendix "Two averaged labels" that the geometry AUC gap of both averaged labels, `own_avg` and `theta_avg`, lies "inside the floor at every depth and at every threshold" we check literally: whether the geometry AUC gap falls within the spread of the geometry AUC gap over the model-twin control pairs taken as the procedure floor. For each averaged label and each of its forms, the table gives the number of relative-depth positions (in the continuous form of the averaged label) or of binarisation threshold $\times$ relative-depth position cells (in the binary form of the averaged label) where the geometry AUC gap lies inside the spread of the procedure floor, below it or above it; the last column gives the number of values of the geometry AUC gap no larger than the procedure floor in magnitude.

label, form	inside	below	above	$\lvert\cdot\rvert \leq$ floor
`own_avg`, cont.	3 of 14	11	0	14 of 14
`theta_avg`, cont.	14 of 14	0	0	14 of 14
`own_avg`, thr.$\times$pos.	28 of 98	70	0	98 of 98
`theta_avg`, thr.$\times$pos.	59 of 98	39	0	98 of 98

For `own_avg`, the claim of the appendix "Two averaged labels" that the geometry AUC gap lies inside the procedure floor does not hold when read literally. In the continuous and in the binary form of the averaged label, for `own_avg` and for `theta_avg`, every value of the geometry AUC gap that falls outside the procedure floor lies below it. The mean geometry AUC gap of `own_avg` drops to -0.0049, that is, the AUC of the transfer arm, the source model's probe carried into the target model's space by the stitch (a ridge map from the target model's space to the source model's space, fitted on the same prompts without labels), is slightly higher than the AUC of the label-only arm, the target model's probe trained on the same averaged label of the source model. Therefore the finding of the appendix "Two averaged labels" that rests on the claim about the procedure floor remains valid: on averaged labels, the source model's probe carried by the stitch into the target model's space loses in AUC, relative to the target model's probe trained on the same averaged label of the source model, no more than the procedure floor. In the appendix "Two averaged labels" the wording "inside the floor" should be replaced by "no larger than the procedure floor in magnitude".

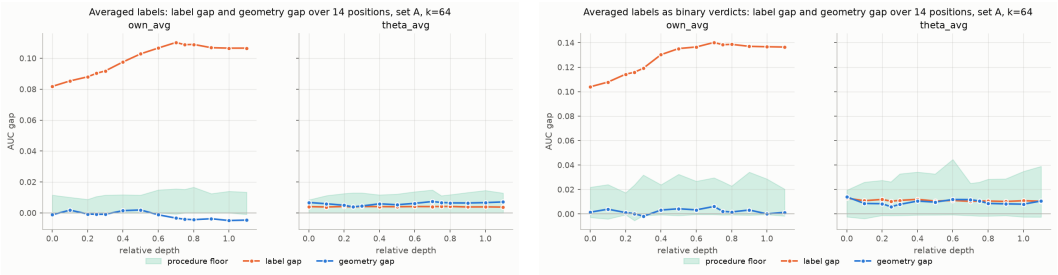

Figure 12: Label and geometry AUC gaps for the averaged labels `own_avg` and `theta_avg` over 14 relative-depth positions compared with the procedure floor; set A, k=64. In the left figure the averaged labels are continuous, in the right one binary, with binarisation threshold 0.5; in each figure `own_avg` is on the left, `theta_avg` on the right. Reading: for `own_avg` in the continuous form the geometry AUC gap lies below the procedure floor at most relative-depth positions, not above it; for `theta_avg` both the label and the geometry AUC gaps lie inside the procedure floor.

J ABLATIONS G1–G6

For each ablation the reading rule was written before the data; there is one exception — the second attempt of ablation G5 on model scale as an axis with a control for vendor: its rule was written after we had seen the value of the first attempt.

J.1 G1: SELECTIVITY IN UNITS OF τ

We compare the guard's verdict on the source model's answer with the control slab — a connected control, that is, a control label whose prompts labelled 1 form one connected region of the representation space: a slab on one random direction in the source model's representation space. The control slab contains the prompts whose projection of the source model's representation onto the random direction of the control slab is closest to the offset of the control slab. The share of prompts labelled 1 of the control slab is the same as that of the guard's verdict on the source model's answer, and the offset of the control slab is chosen so that the decodability of the control slab is as close as possible to the decodability of the guard's verdict on the source model's answer. In 560 of 560 ordered model pair $\times$ guard cells the decodabilities of the control slab and of the guard's verdict on the source model's answer differed by no more than 0.03; the median discrepancy of decodabilities over cells is 0.0096. In each cell we compute the difference of label transport $\tau_{\text{guard}} - \tau_{\text{slab}}$; label transport is the ratio of the class separation of the score of the probe carried into the target model to the class separation of the score of the same probe in the source model.

Outcome by the rule: the guard's verdict on the source model's answer has selectivity of label transport relative to the control slab. The median of the difference $\tau_{\text{guard}} - \tau_{\text{slab}}$ equals $+0.3512$, 95% $[+0.2962,\ +0.3752]$ over 560 cells, bootstrap over source models; this difference is above zero in 551 of 560 cells. In the table below the rows are source models; "guard" in the header means the guard's verdict on the source model's answer, "difference" means $\tau_{\text{guard}} - \tau_{\text{slab}}$, and the last column is the number of cells of this source model in which the difference is above zero.

source	τ, guard	τ, slab	difference ↑	above zero
`smollm3_3b`	0.9746	0.5813	+0.4091	70 of 70
`qwen25_05b`	1.0569	0.6646	+0.3973	70 of 70
`llama31_8b`	0.9492	0.5617	+0.3851	70 of 70
`olmo2_1b`	0.9711	0.6283	+0.3723	70 of 70
`qwen3_8b`	0.9233	0.5916	+0.3473	70 of 70
`qwen3_32b`	0.8923	0.5641	+0.3274	70 of 70
`falcon3_7b`	0.8598	0.5711	+0.2415	68 of 70
`llama31_70b`	0.8176	0.6099	+0.2260	63 of 70

The difference of label transport between the guard's verdict on the source model's answer and the control slab is not explained by a single guard: the medians of this difference for individual guards lie between $+0.3273$ (`xg8`) and $+0.3871$ (`lg3`). In appendix H.3 the guard's verdict on the reference model's answer, assigned to all models, does not win in the transfer of the generalised eigendirections of the covariance pair, a second-moment statistic, over the control half-space of the reference model with the same share of prompts labelled 1; the control half-space of the reference model is a connected control whose prompts labelled 1 lie in the half-space cut off by a random direction in the reference model, and this labelling is assigned to all models. The difference of label transport from the table by source models cannot be compared directly with the finding of §H.3 about the control half-space of the reference model: this difference concerns the direction of the probe trained in the source model, not the generalised eigendirections of the covariance pair taken over the prompts labelled 1 and over the other prompts.

J.2 G3: THE TRANSPORT LAW UNDER A CHANGE OF PROMPT SOURCE

The prompts of the corpus are taken from two prompt sources: 1999 from `prompt_injections_benchmark` and 1307 from JailbreakHub. We randomly split the prompts of each prompt source X into halves X1 and X2. We fit the canonical directions of the model pair and the probe on the guard's verdict on the source model's answer on X1, and compute the canonical correlations ρ of the model pair on X2. The transport law (Theorem 4) predicts the

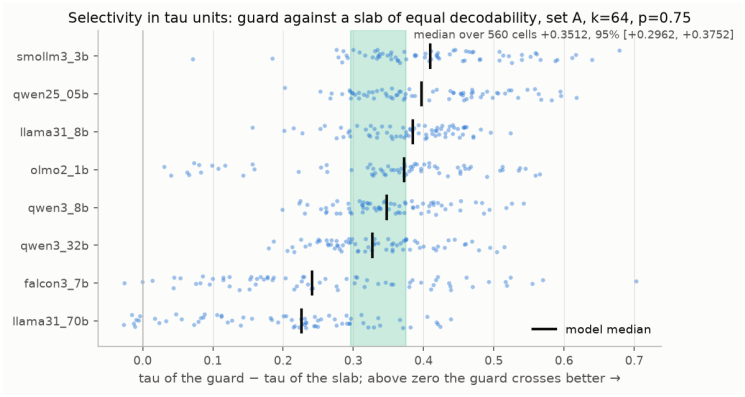

Figure 13: Difference between the label transport τ of the guard's verdict on the source model's answer and the label transport τ of the control slab matched to this verdict in share of prompts labelled 1 and in decodability, one point per ordered model pair $\times$ guard cell. Vertically – source models; the vertical tick – the median difference of τ over the cells of this source model; the green region – the 95% interval of the median difference over all cells; set A, number of principal components k=64, relative depth p=0.75. Reading: almost all points and all medians of the difference for the source models are to the right of zero.

transfer of the source model's label direction from the spectrum of canonical correlations of the model pair: $\sqrt{\mathbb{E}_\omega[\rho^2]}$. The measured transfer of the source model's label direction is the correlation of the score of the probe trained in the source model on the guard's verdict on the source model's answer with its score in the target model after the ridge stitch, that is, a ridge map from the target model's space to the source model's space, fitted on the same prompts without labels; we fit the ridge stitch on X1, and compute this correlation on X2 and on the prompts of the other prompt source. The computation of the outcome of G3 by the rule written before the data included 1120 of 1120 ordered model pair $\times$ guard cells.

source X	cells	r on X2	r on other	drop, 95%
JailbreakHub	560	0.9213	0.3242	$+0.5971$, 95% $[+0.4826,\ +0.7171]$
prompt inj.	560	0.9792	0.5663	$+0.4129$, 95% $[+0.3296,\ +0.5181]$

On W_mix19 (342 ordered pairs, 3420 cells per prompt source) the same rule gives the same outcome: r on X2 0.9334 and on the other prompt source 0.2879 for JailbreakHub, drop $+0.6455$, 95% $[+0.5800,\ +0.7121]$; 0.9798 and 0.2654 for `prompt_injections_benchmark`, drop $+0.7144$, 95% $[+0.6309,\ +0.7977]$; by group in Figure 15.

Outcome by the rule: the transport law is unstable under a change of prompt source. The intervals of the drop in the correlation between the prediction of the transport law and the measured transfer of the source model's label direction when moving from half X2 to the prompts of the other prompt source, one for each prompt source X, lie entirely above the threshold 0.05 of the reading rule written before the data. The spectrum of canonical correlations of the model pair and the profile of the source model's probe (the distribution of this probe's score over the canonical directions), estimated on one prompt source, predict the transfer of the source model's label direction on the same prompt source for each model pair and guard, and on the prompts of the other prompt source they predict this transfer noticeably worse. The transport law (Theorem 4) remains an identity. But the spectrum of canonical correlations of the model pair, measured on one prompt distribution, cannot be taken as a property of the model pair that is preserved on another prompt distribution.

J.3 G4: THE MODEL'S SHARE OF UNSAFE ANSWERS BY THE GUARD UNDER AN EQUAL TRAINING-ROW BUDGET PER CLASS

In §G the two classes of the guard verdict, safe and unsafe answers, are balanced in probe training by weighting or subsampling, but the number of training rows of each class varies from one combi-

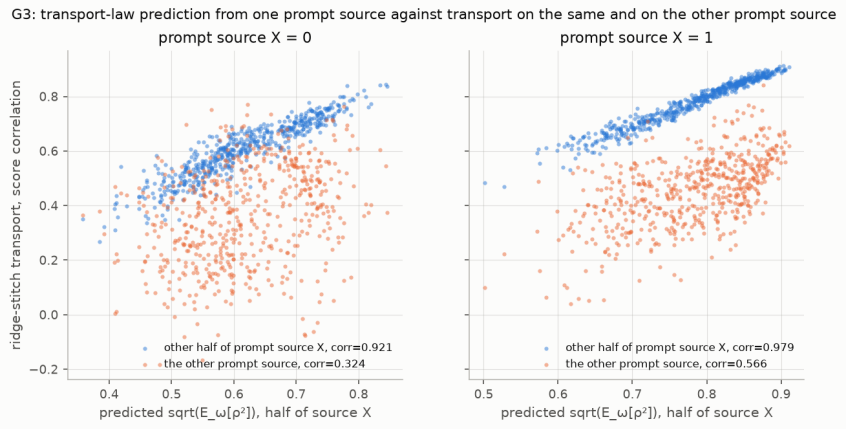

Figure 14: Prediction of the transport law, estimated on one half of prompt source X, against the measured transfer of the source model's label direction on the second half of X (blue points) and on the other prompt source (orange points); one point – an ordered model pair × guard cell; set A, number of principal components k=64, relative depth p=0.75. Reading: the blue points lie close to one line, the orange ones scatter.

nation of model, guard and fold to another. In G4 each probe on the guard's verdict on the model's answer is fitted on the same number of training rows of each class of the guard verdict, m=79: the number of training rows in the smaller of the two guard-verdict classes in the combination of model, guard and fold where this class is smallest. The Spearman correlations of the source model's share of unsafe answers by the guard with the participation ratio of the probe profile $1/\sum_i \omega_i^2$, where ω — the distribution of the probe's score over the canonical directions from the transport law (Theorem 4), and with the transfer of the source model's label direction are computed over source model × guard combinations, 80 such combinations in total; for each combination the participation ratio of the probe profile and the transfer of the source model's label direction are averaged over target models and folds. The intervals of these Spearman correlations are obtained by bootstrap over source models.

probe	Spearman(share, PR)	Spearman(share, transfer)
standard	-0.655, 95% $[-0.814,\ -0.389]$	$+0.695$, 95% $[+0.369,\ +0.862]$
equal budget	-0.621, 95% $[-0.783,\ -0.386]$	$+0.499$, 95% $[+0.233,\ +0.697]$

Outcome by the rule written before the data: partial; both Spearman correlations of the source model's share of unsafe answers by the guard — with the participation ratio of the probe profile and with the transfer of the source model's label direction — weaken under an equal training-row budget per guard-verdict class compared with the standard probe on all training rows. Under an equal training-row budget per guard-verdict class, neither of the two intervals of the Spearman correlations of the source model's share of unsafe answers by the guard reaches zero, so this share still orders both the participation ratio of the probe profile and the transfer of the source model's label direction. But in magnitude neither of the two Spearman correlations of the source model's share of unsafe answers by the guard under an equal training-row budget per guard-verdict class exceeds 0.8. For the standard probe fitted on all training rows, on the same source model × guard combinations, both Spearman correlations of the source model's share of unsafe answers by the guard are also below 0.8 in magnitude, so this threshold is not reached even without an equal budget of rows per guard-verdict class.

The claim in §G that the model's share of unsafe answers by the guard almost determines the participation ratio of the probe profile and the transfer of the source model's label direction rests on a Spearman correlation over ten guards. We recomputed the Spearman correlation over ten guards on the same standard probes fitted on all training rows: we averaged the source model's share of unsafe answers by each guard over the eight source models, and obtained the intervals of this correlation by bootstrap over guards. The Spearman correlation over ten guards of the share of unsafe answers

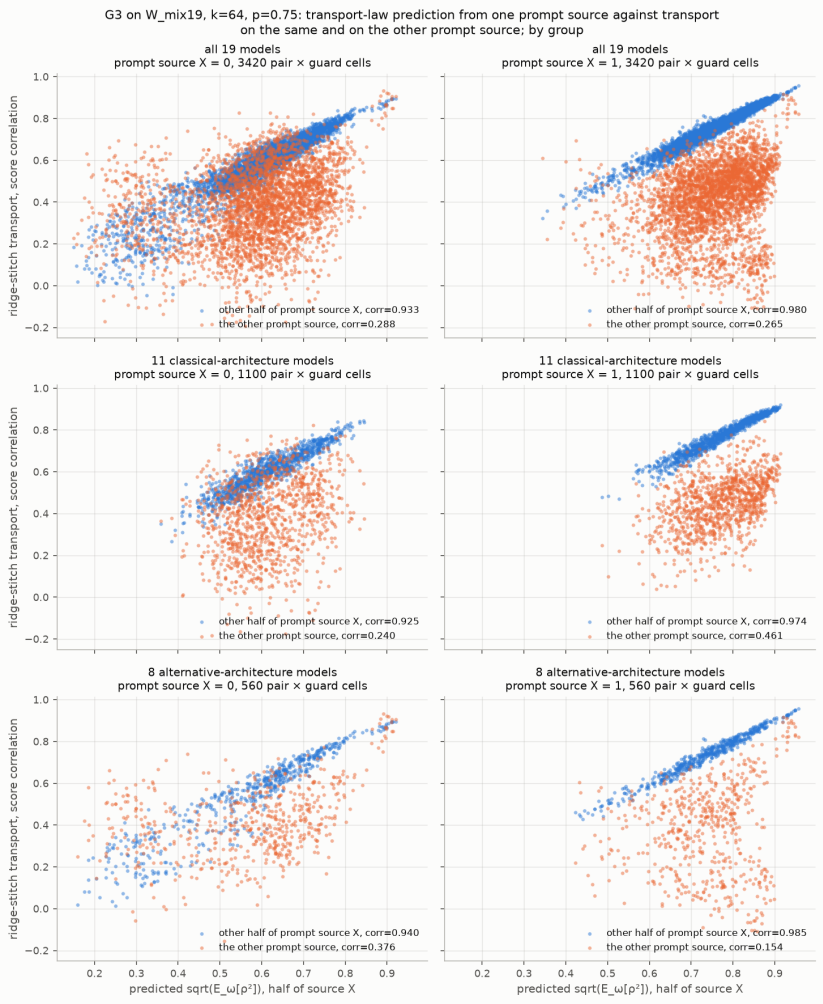

Figure 15: Prediction of the transport law, estimated on one half of prompt source X, against the measured transfer of the source model's label direction on the other half of X (blue) and on the other prompt source (orange); each point is one ordered model pair × guard cell; W_mix19, $k = 64$, relative depth $p = 0.75$. Rows: all 19 models, the 11 classical-architecture models, the 8 alternative-architecture models; columns: prompt source X. Reading: the blue points lie close to one line (correlation 0.92–0.99 across the six panels), the orange ones scatter (0.15–0.46). The same picture on set A is Figure 14.

by the guard, averaged over source models, with the participation ratio of the probe profile equals -0.794, 95% $[-1.000,\ -0.245]$, and with the transfer of the source model's label direction $+0.879$, 95% $[+0.547,\ +0.988]$. *Outcome: partial.* Neither of the two intervals of the Spearman correlations over ten guards of the share of unsafe answers by the guard, averaged over source models, reaches zero; in magnitude the correlation with the transfer of the source model's label direction is above 0.8, and the correlation with the participation ratio of the probe profile is slightly below.

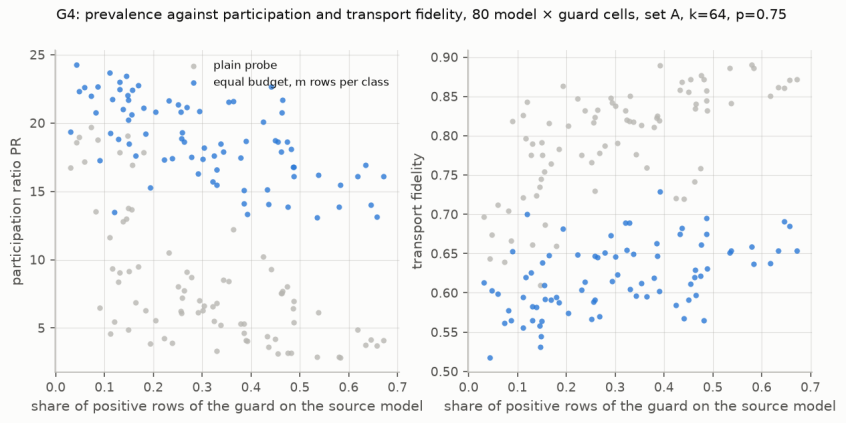

Figure 16: Relation of the source model's share of unsafe answers by the guard to the participation ratio of the probe profile (left) and to the transfer of the source model's label direction (right); one point – a source model × guard combination. Grey points – the standard probe on all training rows, blue – the probe under an equal budget, m rows of each guard-verdict class; set A, k=64, p=0.75. Reading: the blue points follow the same slope as the grey ones, but are scattered more widely.

J.4 G5: MODEL SCALE AS AN AXIS WITH A CONTROL FOR VENDOR

For each of the 28 unordered model pairs of set A, the dimension of the subspace shared by the two models $d_{\text{shared}}(\rho \geq 0.5)$ equals the number of canonical correlations no less than 0.5 between the representations of the two models of the pair. Over the model pairs of set A the dimension of the subspace shared by the two models lies within 18–38, and the largest canonical correlation of the pair ρ_1 – within 0.964–0.984. The rule written before the data prescribes a regression of $d_{\text{shared}}(\rho \geq 0.5)$ on $\log_{10}$ of the size of the smaller model of the pair, $\log_{10}$ of the ratio of the sizes of the two models of the pair, and vendor indicators. The rule written before the data decides the outcome by the coefficient on the size of the smaller model of the pair and by its interval, obtained by bootstrap over the eight models of set A.

Outcome by the rule written before the data: not readable. In 85.6% of the bootstrap replicates the design matrix of the regression prescribed by the rule written before the data lost rank, whereas this rule allows no more than 20%. When the bootstrap draws with replacement eight models of five vendors, the vendor indicators too often turn out indistinguishable from the model sizes. For reference only: over the bootstrap replicates in which the design matrix kept full rank, the coefficient on the size of the smaller model of the pair in the regression of the dimension of the subspace shared by the two models equals $+7.258$, 95% $[-3.006,\ +37.286]$, and in the regression of the largest canonical correlation ρ_1 on the same variables the coefficient on the size of the smaller model of the pair equals $+0.0099$, 95% $[+0.0007,\ +0.0160]$. For the same-vendor model pairs $d_{\text{shared}}(\rho \geq 0.5)$ equals: `llama31_8b–llama31_70b`: 38, `qwen3_8b–qwen3_32b`: 36, `qwen25_05b–qwen3_32b`: 22, `qwen25_05b–qwen3_8b`: 18. There are four same-vendor model pairs, and this is too few to test whether the dimension of the subspace shared by the two models depends on the size of the smaller model of the pair within one vendor.

Second attempt of ablation G5. We wrote the rule of the second attempt of ablation G5 when the outcome of the first attempt of ablation G5 was already known, so the second attempt is weaker than a check by a rule written before the data. We permute the sizes of the eight models of set A among the models 10000 times, keeping each model's vendor, and after each permutation refit the regression of the dimension of the subspace shared by the two models on the same variables as in the

first attempt: $\log_{10}$ of the size of the smaller model, $\log_{10}$ of the ratio of the sizes of the two models, and vendor indicators. The design matrix of the regression lost rank in 0.0% of the permutations of model sizes. The coefficient on the size of the smaller model in the regression of the dimension of the subspace shared by the two models on the original, unpermuted model sizes equals $+7.258$; a two-sided permutation test gives for this coefficient a p-value of 0.326. *Outcome of the second attempt of ablation G5: not distinguished.* The permutation test does not distinguish the coefficient on the size of the smaller model from the coefficients after permuting sizes among models, and from the data of set A it cannot be said whether the dimension of the subspace shared by the two models grows with the size of the smaller model under a control for vendor.

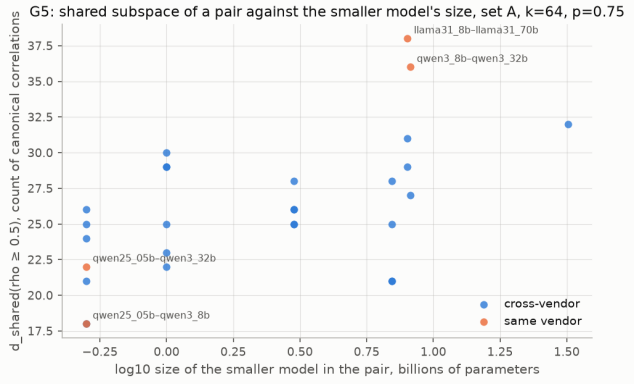

Figure 17: Dimension of the subspace shared by the two models $d_{\text{shared}}(\rho \geq 0.5)$ for each unordered model pair as a function of the size of the smaller model in the pair; orange marks same-vendor model pairs; set A, number of principal components k=64, relative depth p=0.75. Reading: among the same-vendor model pairs, pairs where the smaller model is larger have a larger $d_{\text{shared}}(\rho \geq 0.5)$, but there are only four same-vendor model pairs, and under a control for vendor it cannot be said from the data of set A whether the dimension of the subspace shared by the two models grows with the size of the smaller model.

J.5 G6: STITCH OFF THE DEPTH DIAGONAL

In the main text each stitch (a ridge map from the target model's space to the source model's space, fitted on the same prompts without labels) connects the same relative depth in the source model and in the target model; such depth combinations are called the depth diagonal below. In ablation G6 the depth of the source model and the depth of the target model for the stitch are taken independently from 7 relative depths from 0 to 1.1 (depth 1.1 – the last block). In total this gives 49 combinations of source-model depth and target-model depth, and each combination has 560 ordered model pair $\times$ guard cells. The label AUC gap depends only on the depth of the target model, while the geometry AUC gap depends on both the depth of the source model and the depth of the target model. The stitch with independently taken depths is checked on the depth diagonal: there the geometry AUC gap equals $+0.0019 \ldots + 0.0050$ and coincides with the values of the geometry AUC gap from §4.1 at the same relative depths. *Outcome by the rule of ablation G6: off the depth diagonal too, the ranges of the label and geometry AUC gaps do not overlap.* In the 42 off-diagonal depth combinations the label AUC gap lies within 0.0807–0.1088, and the geometry AUC gap within $-0.0030 \ldots +0.0273$, so the largest geometry AUC gap is smaller than the smallest label AUC gap.

On the map of the geometry AUC gap averaged over ordered model pairs and guards, by depth combination (Fig. 18), the geometry AUC gap does not grow with distance from the depth diagonal. The largest values of the mean geometry AUC gap are gathered in one row of this map, in the off-diagonal depth combinations with source-model depth 0: there it equals $+0.0188 \ldots + 0.0273$ and reaches its maximum at target-model depth 0.5. In the remaining off-diagonal depth combinations, with nonzero source-model depth, the geometry AUC gap equals $-0.0030 \ldots + 0.0125$, and in 14 of these 36 combinations it is lower than in the diagonal combination with the same target-model depth.

Relative depth 0 falls in every model on the first block after the embedding, so the geometry AUC gap at source-model depth 0 being larger than in the remaining off-diagonal depth combinations could be explained by degeneracy of the hidden states of the source model's first block: many coinciding rows of the first block's feature matrix, or a small participation ratio of the feature covariance spectrum of the first block, that is, the square of the sum of the covariance eigenvalues divided by the sum of their squares; this is not the participation ratio of the probe profile from ablation G4. Degeneracy was checked on the raw hidden states of the first block over all prompts of the corpus. For all eight models of set A the share of distinct rows of the first block's feature matrix among all its rows is no less than 0.9997, and the participation ratio of the feature covariance spectrum of the first block equals 1.98–13.80. The participation ratio of the feature covariance spectrum of the first block falls below the threshold 2 of the degeneracy-check rule written before the data only for `llama31_70b`. *Outcome by the degeneracy-check rule: the geometry AUC gap at source-model depth* 0 *being larger than in the remaining off-diagonal depth combinations is not explained by degeneracy of the hidden states of the first block alone.* For the seven source models whose first-block hidden states are not degenerate, the mean geometry AUC gap in the off-diagonal combinations with source-model depth 0 equals $+0.0209$ and exceeds $+0.0125$, the largest geometry AUC gap in the remaining off-diagonal depth combinations. The mean off-diagonal geometry AUC gap at source-model depth 0 is largest for the source models `llama31_8b` ($+0.0515$) and `olmo2_1b` ($+0.0410$), while for `smollm3_3b` (-0.0062) it is negative. Consequently, for some source models the stitch from the first block after the embedding of the source model gives a geometry AUC gap larger than in the remaining off-diagonal depth combinations, and this cannot be explained by degeneracy of the first-block hidden states.

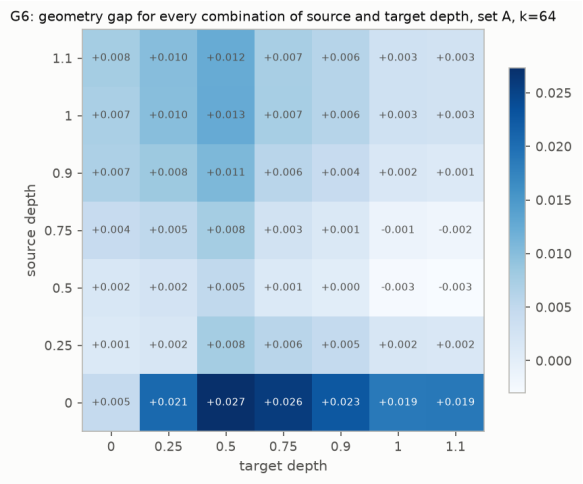

Figure 18: Geometry AUC gap averaged over ordered model pairs and guards for each combination of source-model depth (map rows) and target-model depth (map columns), set A, k=64. Reading: the geometry AUC gap does not grow with distance from the depth diagonal, where the source-model depth equals the target-model depth, and the largest values of the geometry AUC gap are gathered in the map row with source-model depth 0.

K THE PROMPT-SOURCE LABEL AND THE CHOICE OF STITCH

K.1 TRANSFER OF THE SOURCE MODEL'S LABEL DIRECTION FOR GUARD VERDICTS AND FOR THE PROMPT-SOURCE LABEL

The prompt-source label equals 1 for the 1999 prompts from `prompt_injections_benchmark` and 0 for the remaining 1307 prompts, taken from JailbreakHub. The prompt-source label is the same for all models and is unrelated to the safety of the model's answer: it is a property of the prompt text, namely the prompt source the prompt is taken from. For each model pair and each fold, the canonical directions of the model pair are fitted on the training rows of the fold, and the canonical correlations ρ are computed on the test rows of the fold. Then, in the representation space of the source model, a separate probe is fitted on the training rows of the fold for each binary prompt label: on the verdict of each of the ten guards on the source model's answer, on the prompt-source label, and on a random label with the same share of prompts labelled 1 as the prompt-source label. For each probe we compute the transfer of the source model's label direction, that is, the correlation of the probe's score in its source model with its score in the target model after the stitch – a ridge map from the representation space of the target model to the representation space of the source model, fitted on the same prompts without binary prompt labels (§2); the source model's label in this transfer is the binary prompt label the probe is trained on: the guard's verdict on the source model's answer, the prompt-source label, or the random label with the same share of prompts labelled 1 as the prompt-source label.

probe label	combinations	mean transfer
guard verdicts	2800	0.7952
min. and max. over guards		0.7316 (`lg4`), 0.8380 (`og25`)
prompt-source label	280	0.9182
random, share as for source	280	0.5045

Outcome by the rule: the transfer of the source model's label direction is higher for the prompt-source label than for the verdicts of the ten guards on the source model's answer. Averaged over the 56 ordered model pairs of set A, the difference in the transfer of the source model's label direction between the verdicts of the ten guards on the source model's answer and the prompt-source label is -0.1230, 95% $[-0.1379,\ -0.1090]$. The difference in the transfer of the source model's label direction between the prompt-source label and the random label with the same share of prompts labelled 1 as the prompt-source label is given for reference only: $+0.4138$, 95% $[+0.4045,\ +0.4222]$.

The prompt-source label is decoded almost perfectly in the source model, and the transfer of the source model's label direction for the prompt-source label could be higher simply because the decodability of the prompt-source label is higher than that of the guard verdicts on the source model's answer. Therefore, on each source model we lower the decodability of the prompt-source label to the decodability of each guard's verdict on the answer of that source model: in a randomly chosen share q of rows we flip the prompt-source label. The shares of flipped rows q, chosen so as to lower the decodability of the prompt-source label to the decodability of the guard's verdict on the source model's answer, lie in the range 0.100–0.275. In 2800 of 2800 combinations of ordered model pair, fold and guard, the decodability of the guard's verdict on the source model's answer and the decodability of the prompt-source label after flipping the prompt-source label in a share q of rows differed by no more than 0.03, and the median difference between these decodabilities is 0.006.

The transfer of the source model's label direction, averaged over combinations of ordered model pair, fold and guard, is 0.7952 for the verdicts of the ten guards on the source model's answer and 0.9037 for the prompt-source label matched in decodability to the verdict of the same guard on the source model's answer. *Outcome by the rule: the transfer of the source model's label direction for the prompt-source label is still higher than for the guard verdicts on the source model's answer*: the difference in the transfer of the source model's label direction between the guard's verdict on the source model's answer and the prompt-source label matched in decodability to the verdict of the same guard on the source model's answer is -0.1085, 95% $[-0.1231,\ -0.0946]$. The prompt-source label describes the prompt text, while the guard's verdict on the source model's answer describes the safety of that answer. Even when the decodability of the prompt-source label is lowered to the decodability of the guard's verdict on the source model's answer, the prompt-source

label lies on directions in representation space that agree between the source model and the target model better than the directions on which the guard's verdict lies.

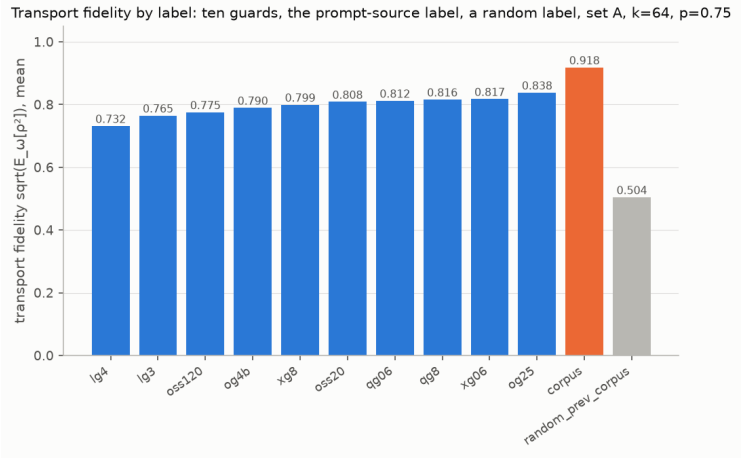

Figure 19: Mean transfer of the source model's label direction for each label on which the probe in the source model is fitted: the verdicts of the ten guards on the source model's answer (blue), the prompt-source label (orange), and a random label with the same share of prompts labelled 1 as the prompt-source label (grey); set A, k=64, p=0.75, 56 model pairs, 5 folds. Read: the bar of the prompt-source label is higher than the bar of each guard's verdict on the source model's answer.

K.2 The main finding when fitting on one prompt source and evaluating on the other

We fit the probes of all three comparison arms (self, label-only, transfer) and the stitch – a ridge map from the space of the target model to the space of the source model – on all rows of prompt source X (one of the two prompt sources, JailbreakHub or `prompt_injections_benchmark`), standardise on these same rows the features of each model – the hidden states of the last prompt token – and compute the AUC of the three comparison arms on the rows of the other prompt source. The probes and the stitch of the procedure-floor controls – the AUC transfer gap between a model and its twin obtained by a known invertible map – are fitted and evaluated on the same rows as the probes of the three comparison arms. The principal-component basis of each model is taken from fold 0. The computation includes 1120 of 1120 ordered model pair $\times$ guard cells.

fit on X	label gap	geom. gap	procedure floor
X = JailbreakHub	$+0.0215$	-0.0143	$-0.1195 \ldots + 0.1107$
X = prompt inj.	$+0.0698$	-0.0104	$-0.0400 \ldots + 0.1659$

Outcome by the rule: the main finding of §4.1 – the AUC transfer gap measures the disagreement between the guard's verdicts on the answers of the two models, not geometry – holds when the probes of the three comparison arms and the stitch are fitted on one prompt source and the AUC is computed on the other. When fitting on either of the two prompt sources, the main finding of §4.1 holds: the geometry AUC gap is not positive, lies inside the procedure floor and below the label AUC gap. When fitting on one prompt source and evaluating on the other, the procedure floor is wide: its controls, like the probes of the three comparison arms, are evaluated on the rows of the prompt source they were not fitted on, so the check that the geometry AUC gap lies inside the procedure floor is coarse. The label AUC gap when fitting on JailbreakHub and evaluating on `prompt_injections_benchmark` is smaller than when fitting and evaluating within JailbreakHub, where it equals $+0.0934$.

K.3 PROCRUSTES STITCH INSTEAD OF RIDGE STITCH

In the transfer comparison arm we map the features of the target model into the feature space of the source model by an orthogonal matrix $W = UV^\top$. The factors of this orthogonal stitch matrix are taken from the singular value decomposition (SVD) of the matrix $Z_B^\top Z_A$, computed on the training rows of each fold; here Z_B are the features of the target model and Z_A are the features of the source model. The other comparison settings are as in §4.1: 14 relative-depth positions, with 560 ordered model pair $\times$ guard cells at each.

Outcome by the rule: the main finding of §4.1 – the AUC transfer gap measures the disagreement between the guard's verdicts on the answers of the two models, not geometry – holds under the Procrustes stitch. Across relative-depth positions the label AUC gap lies within 0.0807–0.1090, the geometry AUC gap under the Procrustes stitch within $+0.0044 \ldots + 0.0208$, and these ranges do not overlap. Under the Procrustes stitch the geometry AUC gap is larger than under the ridge stitch ($+0.0008 \ldots +0.0054$), and reaches its maximum at p=0 ($+0.0208$). The reading rule written before the data does not compare the geometry AUC gap under the Procrustes stitch with the procedure floor: the twin of a model in the model-twin control pairs of the procedure floor is obtained by an anisotropic map, which by construction lies outside the orthogonal group, whereas the Procrustes stitch searches for a map only within this group.

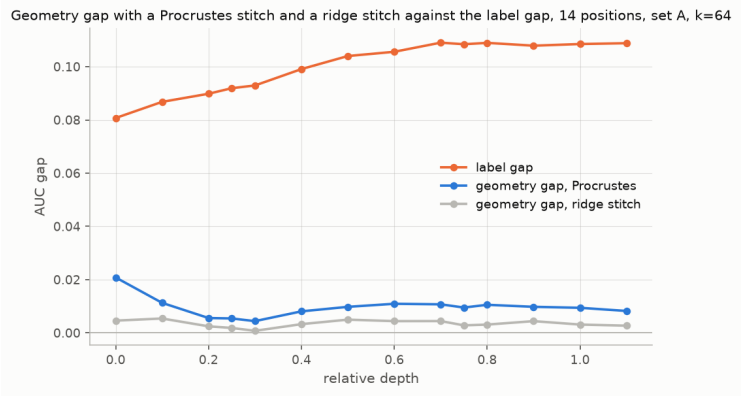

Figure 20: Label AUC gap and geometry AUC gap across 14 relative-depth positions under the Procrustes stitch and under the ridge stitch, set A, k=64. Read: both curves of the geometry AUC gap lie below the curve of the label AUC gap at all relative-depth positions, and the curve of the geometry AUC gap under the Procrustes stitch lies above the curve of the geometry AUC gap under the ridge stitch.

K.4 WEIGHTING THE STITCH INPUT

Before fitting the ridge stitch – the ridge map of the target model's features into the feature space of the source model – we multiply each principal component of the target model's features (the input of the ridge stitch) by the reliability of that component. The reliability of a principal component of the target model's features is the R^2 of its prediction from the source model's features: we compute it by inner cross-validation on the training rows of each fold, without using guard verdicts, and clip it. The three variants of the ridge stitch in the table below are compared over 560 ordered model pair $\times$ guard cells at relative depth p=0.75. The reading rule written before the data gives an outcome only for the ridge stitch whose input is multiplied by the reliability of the principal component, and the ridge stitch whose input is multiplied by the variance scale is given for reference, without an outcome. The transfer of the source model's label direction in the last column of the table below is the correlation of the score of the probe trained in the source model on the cell guard's verdict on that model's answer, on the test rows of each fold, with the score of the same probe on the target model's features mapped by the ridge-stitch variant from the same table row.

stitch	AUC transfer ↑	geom. gap ↓	transfer ↑
ridge	0.6629	+0.0028	0.8042
reliability (rule)	0.6631	+0.0026	0.8039
scale (reference)	0.6629	+0.0028	0.8042

Outcome by the reading rule: weighting the ridge-stitch input by the reliability of the principal component gives a gain in the AUC of the transfer comparison arm. The gain in the AUC of the transfer comparison arm from weighting the input by the reliability of the principal component, compared with the ordinary ridge stitch, is $+0.0002$, 95% $[+0.0001,\ +0.0004]$ by bootstrap over ordered model pairs. For the reading rule only the sign of the gain in the AUC of the transfer comparison arm matters, and it is positive. But the gain in the AUC of the transfer comparison arm from weighting the input by the reliability of the principal component falls in the fourth decimal place, so to the third decimal place this AUC under input weighting coincides with the AUC under the ordinary ridge stitch.

K.5 The geometry AUC gap under a structurally degraded stitch

On 12 ordered model pairs and all ten guards we degrade the matrix of the ridge stitch that maps the target model's features into the feature space of the source model in one of two ways: either we mix the ridge-stitch matrix with a random matrix of the same norm, where a is the weight of the random matrix in the mixture, or we truncate the ridge-stitch matrix via SVD to rank ℓ. In each cell of the table below, after the level of degradation of the stitch matrix (the weight a of the random matrix or the rank ℓ of the truncation) stand the transfer of the source model's label direction through the degraded stitch and the geometry AUC gap under this degraded stitch, averaged over ordered model pairs and guards.

degradation	level: transfer, geom. gap				
mixture	a=0: 0.8262, −0.0052	a=0.25: 0.8096, −0.0017	a=0.5: 0.7062, +0.0170	a=0.75: 0.4023, +0.0727	a=0.9: 0.1573, +0.1207
truncation	ℓ=64: 0.8262, −0.0052	ℓ=32: 0.8223, −0.0034	ℓ=16: 0.8027, +0.0022	ℓ=8: 0.7588, +0.0226	ℓ=4: 0.7030, +0.0350

Outcome by the reading rule: the geometry AUC gap does not distinguish the two ways of degrading the stitch. The reading rule written before the data compares, for the two ways of degrading the stitch, the slopes of the dependence of the geometry AUC gap on the transfer of the source model's label direction: by how much the geometry AUC gap grows when the transfer of the source model's label direction falls by one. For the mixture with a random matrix the slope of the geometry AUC gap in the transfer of the source model's label direction is $+0.1856$, for rank truncation $+0.2466$; the difference of slopes "rank truncation minus mixture with a random matrix" is $+0.0610$, 95% $[-0.0085,\ +0.1397]$ by bootstrap over ordered model pairs. The bootstrap interval of the difference of slopes between the two ways of degrading the stitch contains zero, so by the reading rule the geometry AUC gap does not tell a stitch degraded by rank truncation from a stitch degraded by mixing with a random matrix.

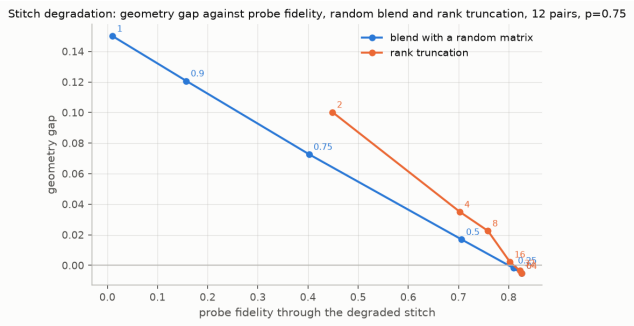

Figure 21: The geometry AUC gap, averaged over ordered model pairs and guards, as a function of the transfer of the source model's label direction through the degraded stitch, averaged the same way, at each level of the two ways of degrading the stitch; point labels are the weight a of the random matrix in the mixture and the rank ℓ of the truncation of the stitch matrix; set A, k=64, p=0.75. Read: both curves of the geometry AUC gap – under mixing with a random matrix and under rank truncation – rise as the transfer of the source model's label direction falls, and the rank-truncation curve is no flatter than the curve of mixing with a random matrix.

L PROMPT SUBSETS, THE CONSENSUS PROMPT LABEL AND THE GUARD-VERDICT AXIS

L.1 THE AUC TRANSFER GAP ON THE SMALLER OF THE TWO PROMPT SUBSETS

The AUC transfer gap in §4.1 is computed over all prompts of the corpus, from both prompt sources. Here we compute both parts of the AUC transfer gap, the label and geometry gaps in the measure of each reading (TPR in reading (a), AUC in reading (b)), only on the smaller of the two prompt subsets, and we understand the smaller of the two prompt subsets in two ways – readings (a) and (b) below. Both readings of the smaller of the two prompt subsets are computed on the same out-of-fold scores of the probes of the three comparison arms (self, label-only and transfer) as in §4.1, over 560 ordered model pair $\times$ guard cells. (a) The smaller of the two prompt subsets is the subset of prompts labelled 1 by the cell guard's verdict on the target model's answer, that is, the prompts on which this guard judged the target model's answer unsafe. The quality of the probe score of each comparison arm in reading (a) is measured by the share of prompts labelled 1 by the cell guard's verdict on the target model's answer that the probe score of this arm places above the threshold with 10% false alarms on the remaining prompts (TPR at FPR 0.10), because AUC can hide a drop in score quality concentrated on a small subset of prompts labelled 1. (b) The smaller of the two prompt subsets is the rows of the smaller of the two prompt sources, JailbreakHub (the other prompt source is `prompt_injections_benchmark`), and the quality of the probe score of each comparison arm remains the AUC, computed only on the rows of JailbreakHub.

reading	self	label-only	transfer	label	geom.	floor
(a) TPR@0.10	0.3843	0.2123	0.2173	$+0.1720$	-0.0050	$-0.0316 \ldots + 0.0717$
(b) AUC, JBH	0.7224	0.6296	0.6303	$+0.0928$	-0.0008	$-0.0372 \ldots + 0.0596$

Outcome by the rule, both readings: on the smaller of the two prompt subsets the geometry gap in the measure of the reading lies inside the procedure floor. In each of the two readings of the smaller prompt subset the geometry gap is computed in the measure of that reading (TPR in (a), AUC in (b)) and is more than an order of magnitude smaller in absolute value than the label gap in the same measure.

L.2 THE CONSENSUS PROMPT LABEL

A guard gives a verdict on the answer of each model separately, so the verdicts of one and the same guard on the answers of different models assign label 1 to different prompts. The consensus prompt label equals 1 when at least half of the 80 guard verdicts on the answers of the models of set A (8 models $\times$ 10 guards) judge unsafe their model's answer to this prompt; the share of corpus prompts with the consensus prompt label equal to 1 is 0.181. The consensus prompt label is one for all models of set A, so the label AUC gap for the consensus prompt label is zero by construction.

consensus label	value
self AUC	0.7491
transfer AUC	0.7383
geom. gap, mean	$+0.0107$, 95% $[+0.0086,\ +0.0131]$
procedure floor	$-0.0018 \ldots + 0.0239$

Outcome by the rule: the geometry of representations is shared by the models of set A on the same prompts. When the binary prompt label is one for all models, like the consensus prompt label, and the prompts are the same for all models, the AUC transfer gap does not go beyond the procedure floor: the whole confidence interval of the geometry AUC gap with the consensus prompt label over ordered model pairs of set A lies inside the procedure floor. For reference: the geometry AUC gap with the label being the guard's verdict on the source model's answer, averaged over the ten guards and the ordered model pairs of set A, at the same relative depth p as in the table above, is $+0.0028$.

L.3 WHITENING OF THE PRINCIPAL-COMPONENT BASIS AND THE CONTAMINATION CURVE

A guard's verdict on a model's answers splits the prompts into two classes; we call the smaller of the two classes the guard's smaller class. The contamination curve shows how, as the replaced share grows, the intrinsic dimension of the guard's smaller class approaches the intrinsic dimension of the other class, when rows of the guard's smaller class are replaced by rows of the other class of the same guard. We take set A, guards `lg3` and `og4b`, relative depth 1.0 and all rows of the corpus. The basis of 256 principal components of each model is taken in three variants: the raw basis without whitening, the basis after whitening, and the placebo basis. To the placebo basis we first apply whitening and then return to it the eigenspectrum of the same model – the variances of its principal components, shuffled across components. The contamination curve of each combination of model and guard is reduced to a single number, the normalised slope $\beta = (\hat{d}_{\mathrm{ID}}(0.1) - \hat{d}_{\mathrm{ID}}(0))/0.1/(\hat{d}_{\mathrm{ID}}(1) - \hat{d}_{\mathrm{ID}}(0))$. In the normalised slope β, the quantity $\hat{d}_{\mathrm{ID}}(\alpha)$ is the intrinsic dimension by the TwoNN estimate for the guard's smaller class in which a share α of rows is replaced by rows of the other class of the same guard. If the intrinsic dimension $\hat{d}_{\mathrm{ID}}(\alpha)$ of the guard's smaller class changes linearly in the share α of replaced rows, the normalised slope β equals 1. *Outcome by the rule: whitening of the principal-component basis does not change the normalised slope of the contamination curve.* The difference of the normalised slopes of the contamination curve "in the basis after whitening minus in the raw basis" is $+1.414$, 95% $[-0.956,\ +4.109]$ over 14 of 16 combinations of model and guard, with the confidence interval built by bootstrap over models. The difference of the normalised slopes of the contamination curve "in the placebo basis minus in the raw basis" is $+1.051$; we give it for reference. The normalised slope of the contamination curve is noisy: the rule written before the data discards combinations of model and guard whose contamination-curve range $\hat{d}_{\mathrm{ID}}(1) - \hat{d}_{\mathrm{ID}}(0)$ is below 0.1, but a range just above this threshold also stands in the denominator of the normalised slope and inflates it. "Does not change" in the outcome of the rule written before the data means that the confidence interval of the difference of the normalised slopes, built by bootstrap over models, does not distinguish the normalised slope of the contamination curve in the basis after whitening from the normalised slope in the raw basis.

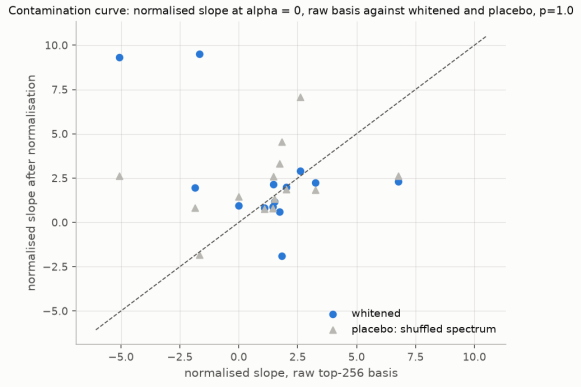

Figure 22: Normalised slope of the contamination curve in the raw basis against the slope in the basis after whitening and in the placebo basis; one point per combination of model and guard that passed the threshold on the contamination-curve range; relative depth 1.0. Read: the points fall on both sides of the diagonal on which the normalised slope in the raw basis equals the normalised slope in the other basis, both for the basis after whitening and for the placebo basis.

L.4 STRETCHING THE GUARD-VERDICT AXIS: THE AXIS ESTIMATED ON THE HELD-OUT HALF OF THE ROWS

The earlier conclusion was: stretching the guard-verdict axis lowers the contrast in intrinsic dimension (ID contrast) no more than stretching an axis orthogonal to it with the same data variance. The earlier conclusion about stretching the guard-verdict axis had three caveats: the guard-verdict axis was estimated on half of the rows, but the other half of the rows also entered the computation of the ID contrast; a filter $|z_{\text{identity}}| \geq 2$, where z_{identity} is the z-statistic of the ID contrast against the permu-

tation null without stretching the axes, discarded part of the combinations; the earlier check ran on a single relative depth and two guards. The new run, in which the guard-verdict axis is estimated on the half of the rows that did not enter the computation of the ID contrast, removes all three caveats of the earlier conclusion.

Set A: 8 models, 10 guards, relative depths 0.75, 0.9, 1.0, raw hidden states of the layer at each of these relative depths. The rows of the corpus are randomly split in half. On the first half of the rows we estimate three axes: the guard-verdict axis (the difference of the mean hidden states of the two classes into which the guard's verdict on the model's answers splits the rows), an axis orthogonal to it with the same data variance, and a random axis. On the second half of the rows we compute the ID contrast and the actual distortion of pairwise distances between rows introduced by stretching the axis. The guard-verdict axis, the orthogonal axis and the random axis are stretched with gains $3, 10, 30, 100$. For each axis and each combination of model, guard, relative depth and intrinsic-dimension estimator we fit the slope of the dependence of Δz on the $\log$ of the actual distortion of pairwise distances; Δz is the change of the z-statistic of the ID contrast relative to the same statistic without stretching. The statistic of the rule written before the data is the difference of the slopes of the dependence of Δz on the $\log$ of the actual distortion "guard-verdict axis minus orthogonal axis with the same data variance", averaged over all combinations of model, guard, relative depth and intrinsic-dimension estimator where both axes are constructed; the confidence interval of the difference of slopes is built by bootstrap over models.

Outcome by the rule: the guard-verdict axis is special. The mean difference of slopes "guard-verdict axis minus axis with the same data variance" is $+0.218$, 95% $[+0.061,\ +0.380]$ over 569 combinations of model, guard, depth and intrinsic-dimension estimator, bootstrap over models; the difference of slopes is above zero in 377 of 569 combinations. The rule written before the data does not select combinations of model, guard, depth and intrinsic-dimension estimator by $|z_{\text{identity}}| \geq 2$. If only the combinations of model, guard, depth and intrinsic-dimension estimator that passed the filter of the earlier check on a single relative depth and two guards by the z-statistic of the ID contrast without stretching are kept, the mean difference of slopes "guard-verdict axis minus axis with the same data variance" is $+0.188$, 95% $[+0.067,\ +0.411]$ over 168 combinations; we give this value for reference.

model	combinations	difference of slopes ↑	range within model
qwen25_05b	87	−0.098	2.268
olmo2_1b	50	+0.040	2.513
llama31_8b	47	+0.054	1.923
smollm3_3b	90	+0.093	2.440
falcon3_7b	66	+0.190	2.259
llama31_70b	81	+0.354	2.608
qwen3_8b	81	+0.518	3.532
qwen3_32b	67	+0.550	2.785

The differences of slopes "guard-verdict axis minus axis with the same data variance" averaged over the combinations of one model diverge: for the models qwen3_8b, qwen3_32b they lie above the interval that the rule builds by bootstrap over models for set A, and for qwen25_05b below zero. For 7 models of 8 the range of the difference of slopes "guard-verdict axis minus axis with the same data variance" across combinations of guard, depth and intrinsic-dimension estimator within one model exceeds two units. The bootstrap in the rule written before the data runs over models, so the outcome of this rule speaks about set A, but not about each model separately. The run with axes estimated on the held-out half of the rows removed all three caveats of the earlier check on a single relative depth and two guards at once, so it cannot be said which of these caveats hid the special nature of the guard-verdict axis.

L.5 Set W_mix19: pairs of classical-architecture and alternative-architecture models

Set W_mix19 contains 19 models: 11 classical-architecture models and 8 alternative-architecture models. The alternative-architecture models are four pairs: LFM2, Falcon-H1, Granite4H and Nemotron-H. The model qwen3_8_27b is left outside set W_mix19 as a single control and is not among the 19 models. The set therefore contains 342 ordered pairs: 110 pairs of two classical-



Figure 23: Difference of slopes "guard-verdict axis minus axis with the same data variance" for each combination of model, guard, depth and intrinsic-dimension estimator. Colour is relative depth, diamonds are the differences of slopes averaged over the combinations of one model, the grey band is the 95% interval of the rule, bootstrap over models. Set A, raw hidden states of the layer; each axis estimated on the half of the rows that did not enter the ID contrast. Read: the grey interval band is above zero, but the mean differences of slopes per model spread from values below zero to values well above the grey interval band.

architecture models, 56 pairs of two alternative-architecture models and 176 mixed pairs of a classical-architecture model and an alternative-architecture model. On 10 guards, N=3306 prompts, 14 relative depths and k=64, the range of the depth means of τ is 0.904–1.370 over all pairs, 0.920–1.100 for pairs of two classical-architecture models, 0.883–1.779 for pairs of two alternative-architecture models and 0.898–1.372 for mixed pairs. For the transfer of the source model's label direction r the corresponding ranges are 0.740–0.865, 0.746–0.876, 0.693–0.842 and 0.730–0.856. These intervals are the minimum and maximum over the 14 mean values at the depths, not confidence intervals and not the spread of individual cells. Below, the same quantities are shown over all pairs and for the three pair types.

The label and geometry AUC gaps by relative depth for all 19 models, for the 11 classical-architecture models and for the 8 alternative-architecture models are shown in Fig. 2 in §4.1.

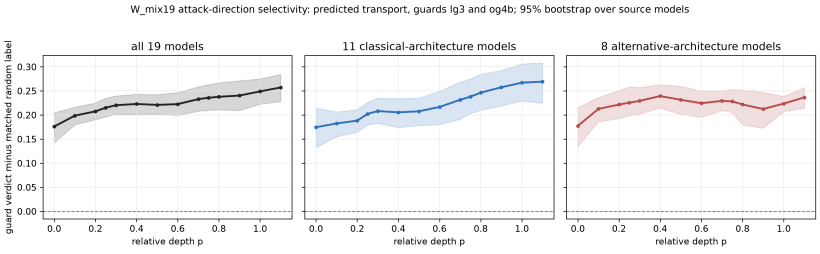

Figure 24: Attack-direction selectivity on W_mix19: predicted transport $\sqrt{\mathbb{E}_\omega[\rho^2]}$ of the probe on the guard verdict (guards lg3 and og4b) minus that of a random label with the same share of label-1 prompts, along relative depth, averaged over folds; panels, all 19 models (342 ordered pairs), the 11 classical-architecture models (110 pairs within them) and the 8 alternative-architecture models (56 pairs within them); bands are 95% bootstrap intervals over source models, $k = 64$. Read: the difference stays above zero at every depth in all three panels, the lowest interval edge being $+0.13$.

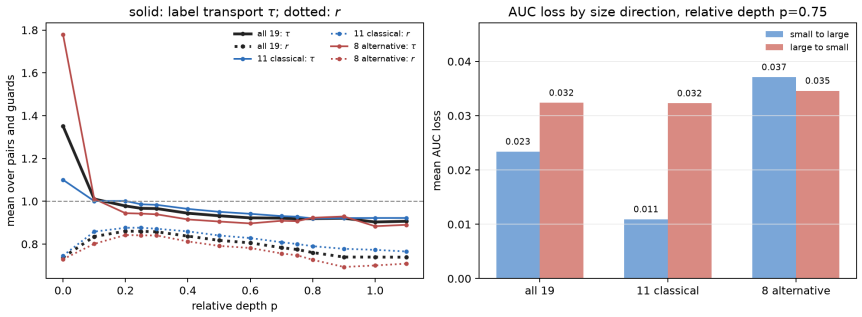

Figure 25: W_mix19 summary, $k = 64$, ten guards. Left: label transport τ (solid) and transfer of the source model's label direction r (dotted) along relative depth for all 19, the 11 classical and the 8 alternative models. Right: AUC loss of the carried probe by the direction of transfer in model size at $p = 0.75$ for the same three groups. Read: τ stays near one and above r in every group; the size asymmetry holds for all 19 and the 11 classical models and not for the 8 alternative ones.

L.6 The main figure by group

Figure 26 repeats the three panels of Figure 3 for each group on its own. The transport law: Pearson coefficient 0.958, OLS slope 0.843 over 3420 cells for all 19 models; 0.998 and 0.980 over 1100 cells for the 11 classical-architecture models; 0.925 and 0.775 over 560 cells for the 8 alternative-architecture models. The slope below one on all 19 models comes from the pairs with at least one alternative-architecture model: the slope is 0.980 on the pairs of two classical-architecture models, 0.838 on the 1760 mixed cells and 0.775 on the pairs of two alternative-architecture models; in the latter two the measured transport exceeds the prediction at the low end of the predicted range.

L.7 Corollary 6: the transport law with the probe's profile replaced

Corollary 6 says that two probes in one model pair transfer differently exactly as much as their profiles ω differ, the profile being the distribution of the probe's score over the canonical directions of the pair. We test it by substitution. For every ordered model pair $\times$ guard cell the reconstruction $\sum_i c_i \rho_i v_i$ of the transport law is built on the test rows four times: with the probe's own coefficients c_i; with the coefficients of the probe trained in the same space and on the same rows on the verdict of the next guard in the fixed order of the ten guards; with the coefficients of a random direction; and with uniform coefficients. Each reconstruction is correlated with the probe's own score, which gives four predictions of the transfer of the source model's label direction, and the measured transfer is the correlation reached by the ridge stitch. We then regress the measured transfer on each prediction after sweeping additive fixed effects of source model, target model and guard from both sides, and report the partial R^2: the share of the residual variance of the measured transfer that the prediction explains once how predictable the cell is at all has been removed. No reading rule was written before the data for this test; it was run on set A in the accompanying notebook and repeated here on W_mix19. Two arms are uninformative by construction and stand in the figure only as the floor: the uniform profile does not vary within a model pair, and the profile of a random direction predicts the transfer of that direction, not of the probe. The partial R^2 of the own profile is near one because the plug-in reconstruction and the ridge stitch are two estimators of one quantity (§4.2); the informative arm is the profile of the next guard's probe, the closest alternative profile the data offer.

Outcome: the law's prediction needs the probe's own profile. On W_mix19 at relative depth $p = 0.75$, $k = 64$, with the probe's own profile the prediction explains 0.9996 of the residual variance of the measured transfer over 3420 cells, and with the profile of the next guard's probe 0.2123, 95% $[0.0899, 0.3806]$ by bootstrap over source models; the random direction and the uniform profile explain at most 0.0043 in any group. Replacing the profile also moves the level of the prediction: the mean predicted transfer falls from 0.7739 with the own profile to 0.6460 with the next guard's profile against a measured mean of 0.7747. By group: 0.1712 $[0.0213, 0.4043]$ over the 1100 cells of the 11 classical-architecture models, 0.4075 $[0.2629, 0.5539]$ over the 560 cells of the 8 alternative-architecture models, 0.1896 $[0.0584, 0.3633]$ over the 1760 mixed cells. The next guard's profile

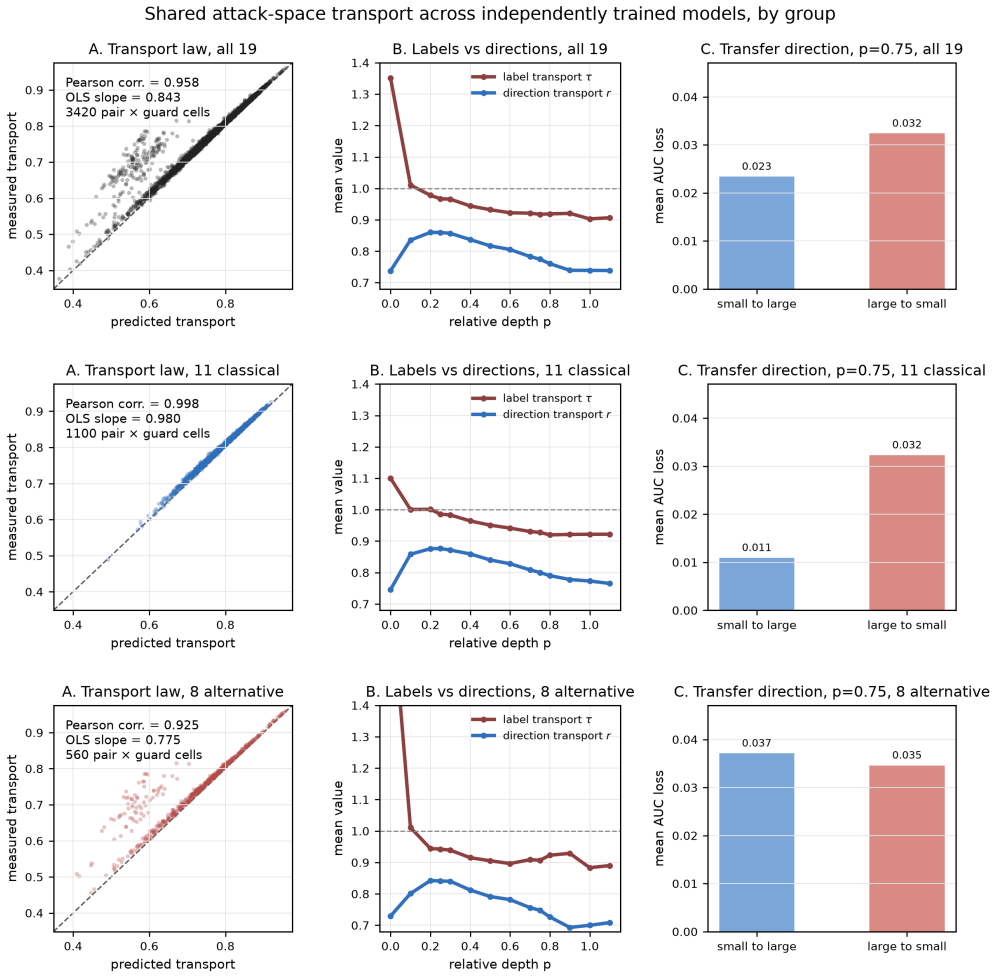

Figure 26: The three panels of Figure 3 by group, one row per group: all 19 models, the 11 classical-architecture models, the 8 alternative-architecture models; $k = 64$, ten guards. A: transport law at relative depth $p = 0.75$, with the Pearson coefficient, the OLS slope and the number of cells of that group. B: label transport τ and transfer of the source model's label direction r along relative depth. C: mean AUC loss by direction of transfer in model size at $p = 0.75$. Reading: the law's fit is tightest on the classical group and loosest on the alternative one; τ and r diverge in every group; the size asymmetry of the AUC loss holds on the classical group and vanishes on the alternative one.

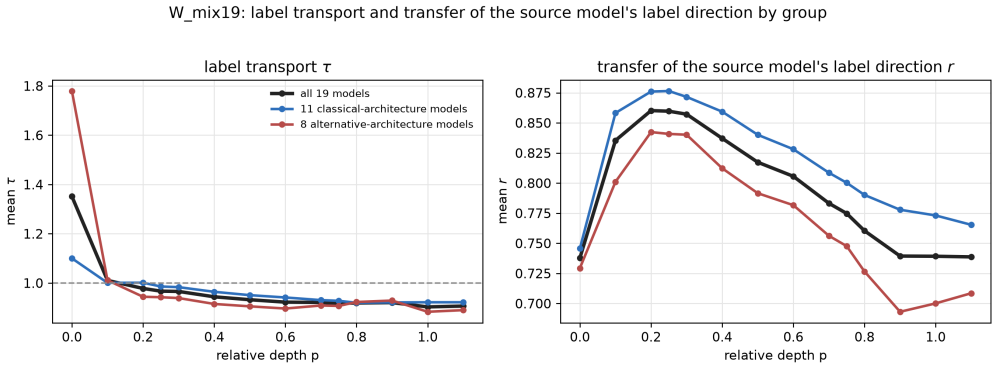

Figure 27: Transfer of class separation τ (left) and transfer of the source model's label direction r (right) along relative depth on W_mix19, $k = 64$, ten guards, for all 19 models (342 ordered pairs), the 11 classical-architecture models (110 pairs within them) and the 8 alternative-architecture models (56 pairs within them). Read: r of the classical models lies above that of the alternative ones at every depth; τ of the alternative models reaches 1.78 at $p = 0$ and is otherwise close to that of the classical ones.

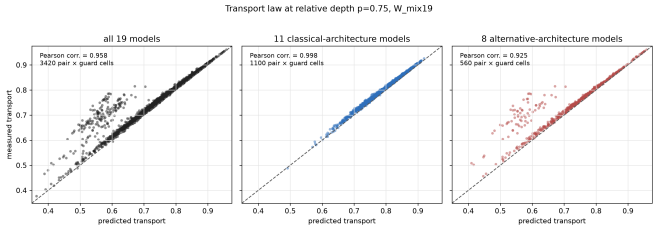

Figure 28: Transport law on W_mix19 at relative depth $p = 0.75$, $k = 64$: predicted against measured transfer of the source model's label direction, one point per ordered pair $\times$ guard; panels, all 19 models (342 ordered pairs), the 11 classical-architecture models (110 pairs within them) and the 8 alternative-architecture models (56 pairs within them). Read: the classical pairs lie on the diagonal (Pearson corr. 0.998); among the alternative pairs a cluster lies above it, the measured transport exceeding the prediction (Pearson corr. 0.925).

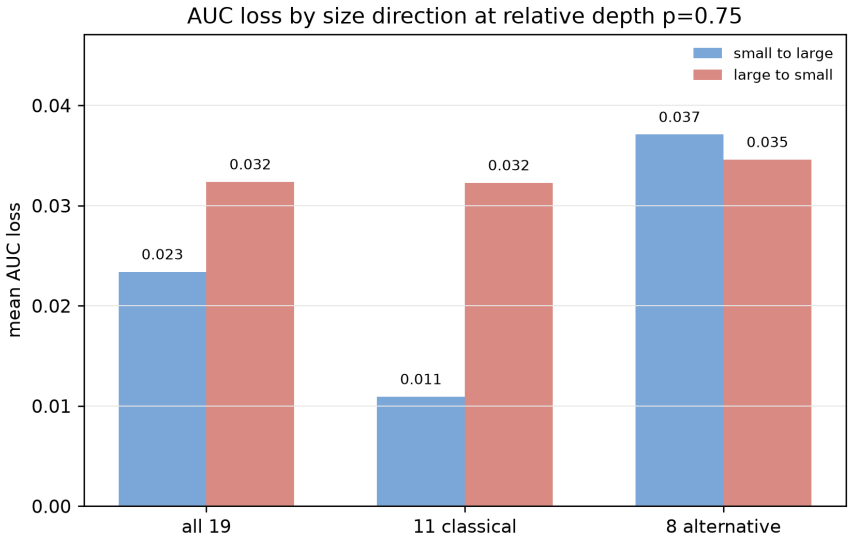

Figure 29: AUC loss of the carried probe by the direction of transfer in model size at relative depth $p = 0.75$ on W_mix19, $k = 64$, ten guards, for all 19, the 11 classical and the 8 alternative models. Read: from the larger to the smaller model the loss is larger for all 19 (0.032 against 0.023) and for the classical models (0.032 against 0.011); for the alternative models the two directions are close (0.035 against 0.037).

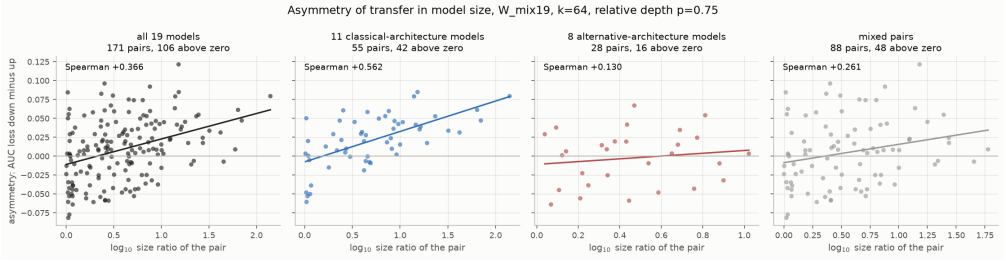

Figure 30: Asymmetry of transfer in model size on W_mix19, $k = 64$, relative depth $p = 0.75$: for every unordered pair of models of different size, the AUC loss of the carried probe averaged over the ten guards in the direction from the larger model to the smaller one minus the loss in the opposite direction, against the $\log_{10}$ size ratio of the pair; panels, all 19 models, the 11 classical-architecture models, the 8 alternative-architecture models and the pairs that mix the two families; the line is the least-squares fit of the panel. Read: for the classical-architecture models the asymmetry grows with the size ratio; for the alternative-architecture models the fit is flat and the pairs sit on both sides of zero.

therefore explains more of the transfer among the alternative-architecture models than among the classical ones, and the two intervals overlap; the own profile explains 0.9997 and 0.9999 in the two groups. On set A the same test over the fourteen relative depths gives for the own profile 0.9652 at $p = 0$ and at least 0.9982 at every other depth, and for the next guard's profile from 0.1452 at $p = 0.20$ to 0.5086 at $p = 0$, rising from 0.16 at $p = 0.25$ to 0.36 at $p = 0.90$; pooled over the 7840 cells $\times$ depths with a fixed effect per cell and per depth the own profile explains 0.9837 and the next guard's profile 0.5174. The next guard's profile keeps a part of the explained variance because the verdicts of two guards on the same answers agree on most prompts, so their probes share a part of their profiles; the part of the transfer that it loses is the part Corollary 6 attributes to the difference of the profiles.

group of W_mix19, $p = 0.75$	cells	own profile	next guard's profile	95% interval	mean prediction, next guard
all 19 models	3420	0.9996	0.2123	[0.0899, 0.3806]	0.6460
11 classical-architecture models	1100	0.9999	0.1712	[0.0213, 0.4043]	0.6675
8 alternative-architecture models	560	0.9997	0.4075	[0.2629, 0.5539]	0.6217
mixed pairs	1760	0.9994	0.1896	[0.0584, 0.3633]	0.6402

Partial R^2 of the law's prediction for the measured transfer of the source model's label direction after sweeping the fixed effects of source, target and guard; the interval is a bootstrap over source models; the last column is the mean prediction with the next guard's profile, against measured means of 0.7747, 0.7999, 0.7480 and 0.7675 in the four rows.

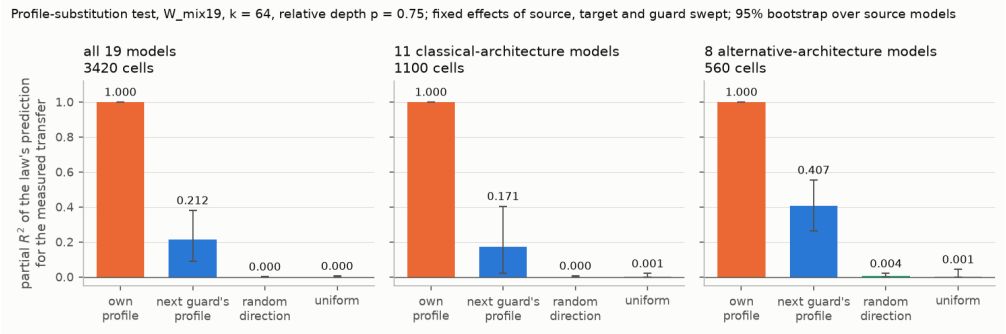

Figure 31: Profile-substitution test of the transport law on W_mix19, $k = 64$, relative depth $p = 0.75$: partial R^2 of the law's prediction for the measured transfer of the source model's label direction after sweeping the fixed effects of source, target and guard, for four profiles put into the reconstruction; panels, all 19 models (3420 cells), the 11 classical-architecture models (1100 cells) and the 8 alternative-architecture models (560 cells); bars are 95% bootstrap intervals over source models. Read: the own profile explains the residual variance almost entirely in every group; the next guard's profile keeps 0.21 of it over all 19 models, 0.17 among the classical and 0.41 among the alternative models; the random direction and the uniform profile explain nothing.

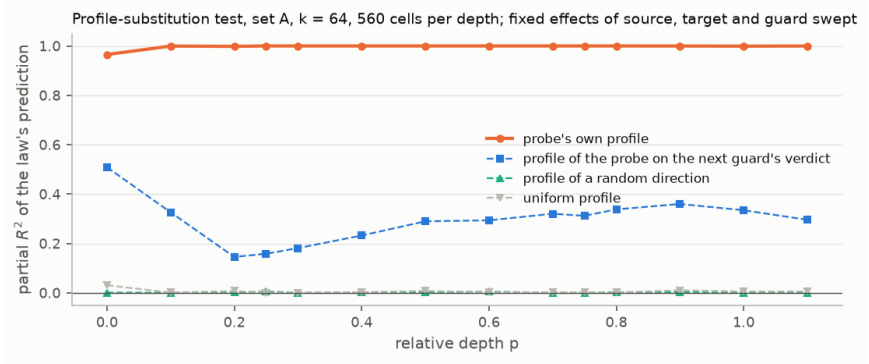

Figure 32: The same test on set A (8 models, 56 ordered pairs, ten guards, 560 cells per depth), $k = 64$, along the fourteen relative depths of the grid. Read: the own profile stays above 0.99 at every depth but $p = 0$; the next guard's profile falls from 0.51 at $p = 0$ to 0.15 at $p = 0.20$ and rises to 0.36 at $p = 0.90$; the two floors stay at zero.