Concept Subspaces Compute Beyond the Logit Lens: A Weights-Only Test for Locating Representations Upstream of Readout

Anonymous authors
Paper under double-blind review

Review copy — incomplete appendix; not for submission.

Abstract

A concept subspace's effect on model behavior does not establish how it relates to the output readout. We introduce a two-sided geometric diagnostic that measures an extracted subspace's overlap with the dominant right-singular directions of the unembedding matrix, evaluated against output-oriented positive controls. Given an extracted basis, the raw diagnostic requires only model weights. Our testbed is the Format-Agnostic Reasoning Subspace (FARS), a ten-dimensional basis extracted from eighteen reasoning concepts expressed in six surface forms. Across nine rank-matched estimators and twenty-six models, four activation-derived concept estimators carry only 0.38–0.80% mean energy in the top-ten readout span. Final-layer PCA carries 3.56%, exceeding FARS in 25 of 26 models. A same-layer next-token control, evaluated using a fitted linear translator for depth matching, carries approximately thirteen times more energy than FARS, with separation in all 25 tested models. Re-extracting FARS on ten disjoint concepts yields 62–100% cross-format retrieval across twenty-four generative models, demonstrating transfer of the extraction procedure rather than a fixed basis. These findings distinguish concept structure from dominant linear readout directions under explicit controls. The diagnostic complements causal interventions; geometric separation alone does not establish the computation performed by a subspace.

1 Introduction

Language models express the same reasoning concept in prose, code, and mathematical notation. Whether this behavioral flexibility reflects shared internal directions is less clear. Work on the linear representation hypothesis recovers individual concept directions (Park et al., 2024; Arditi et al., 2024; Turner et al., 2024), while code instruction tuning links natural-language instructions to program behaviour (Muennighoff et al., 2024) and multilingual studies probe language-specific and shared representations (Tang et al., 2024; Dumas et al., 2025). We ask where cross-format concept identity is represented, whether its relations survive across architectures, and how strongly a recovered concept subspace overlaps the dominant linear readout directions.

We extract a **Format-Agnostic Reasoning Subspace** (FARS) using concept-centroid PCA over eighteen reasoning concepts in six surface forms (Fig. 1). Averaging within concept suppresses the form confound without gradient updates or supervision beyond concept labels. We use rank $k = 10$, an empirical operating point where marginal gains plateau (§7); it is not an intrinsic-dimensionality claim, since eighteen mean-centred centroids span at most seventeen dimensions.

An ablation drop alone cannot tell whether a subspace computes an answer or carries an already-computed identity toward logits. Our main contribution is a *readout-orthogonality*

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.

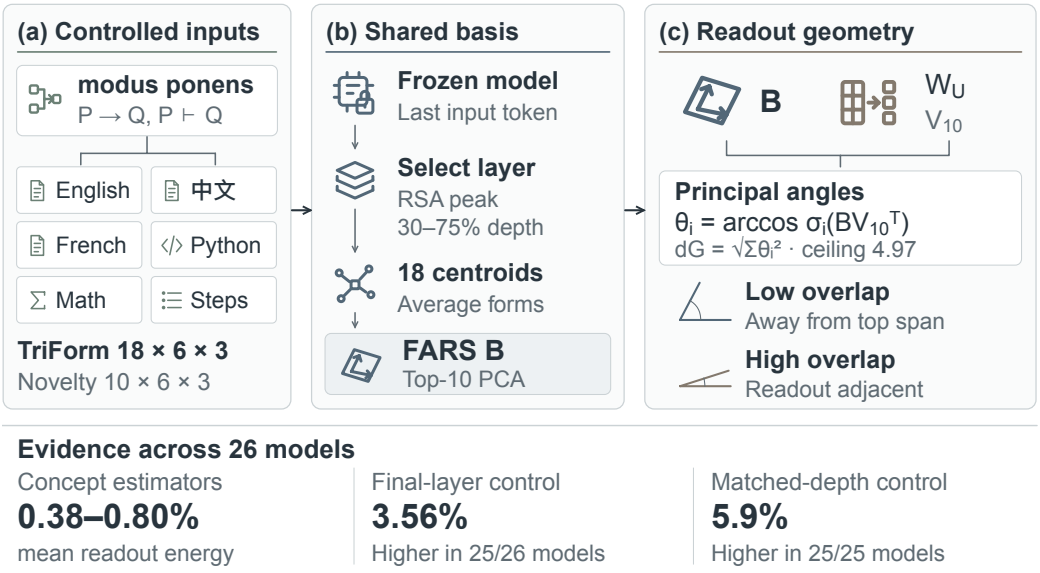

Figure 1: **The method in three stages.** (1) TriForm and Novelty express each concept in six forms. (2) Last-input-token activations are averaged within concept. Top-10 PCA at the concept-RSA peak within 30–75% depth yields FARS **B**. (3) Principal angles compare **B** with the top-k right singular vectors of W_U. Positive controls distinguish low readout loading from readout adjacency. Results summarize 26 models (25 with matched-depth controls); linear orthogonality alone is not a proof of causal sufficiency.

test: compare a probed basis with the top-k right singular vectors of the model's unembedding matrix W_U. Given the probed basis, this geometric comparison requires weights rather than new model runs. Positive and depth-matched controls establish that the test can return both low and high readout loading.

Contributions. **C1 (two-sided test).** A final-layer positive control carries 4.5× more readout energy than FARS, with separation in 25/26 models. After pulling the readout back to the probed layer, a same-layer next-token subspace carries 13× more energy, with separation in 25/25 models (§5). **C2 (estimator comparison).** Across nine matched-rank estimators and twenty-six models (0.4B–70B), the four concept estimators—FARS, LDA, difference-of-means directions, and probe weights—carry only 0.38–0.80% readout energy. Difference-of-means directions span the concept-centroid subspace (Lemma 1). Vocabulary-anchored concepts instead carry 8.4%, versus 0.80% for FARS (App. F.1), distinguishing activation-derived and output-derived representations. **C3 (transport).** Re-extracting FARS on ten disjoint Novelty concepts yields 62–100% cross-format retrieval on twenty-four generative models, 6–57 percentage points above matched-rank random baselines; encoder DeBERTa-v3 reaches 32% (§6). This is transfer of the extraction method, not a fixed basis. **C4 (relations across models).** A shared ten-dimensional frame captures 97.6% of cross-model centroid variance over twenty-seven models on a fixed inventory (App. H), providing a concept-level test of representational alignment (Huh et al., 2024). **C5 (reasoning-stream boundary).** Three reasoning-tuned models contain a CoT-tail subspace distinct from last-input-token FARS. Token-budget truncation and switching to English limit this finding's scope (§8; App. J.1).

2 Related Work

Concept directions in a single model. The Linear Representation Hypothesis line (Park et al., 2024; Elhage et al., 2022) treats individual concepts as directions in activation space. Concept-direction methods have matured rapidly: TCAV (Kim et al., 2018) estimates concept vectors from labelled examples; Marks and Tegmark (2024) read truth as a

single linear direction; Arditi et al. (2024) isolate a direction gating safety refusal; Turner et al. (2024) steer by adding concept directions to residual streams. Distributed Alignment Search (Geiger et al., 2024) learns concept-subspace interchange interventions end-to-end; LEACE (Belrose et al., 2023) provides a closed-form minimum-perturbation erasure; sparse autoencoders (Bricken et al., 2023; Templeton et al., 2024) extract overcomplete feature dictionaries. *Our position relative to these:* concept-centroid PCA is used as a lightweight subspace estimator (comparable to Marks and Tegmark 2024; Kim et al. 2018), but our contribution is not the extraction method. It is (i) a *readout-orthogonality test* (§5) that measures overlap with dominant linear readout directions given a recovered basis and unembedding weights, and (ii) a scoped extension of LRH from a per-direction to a per-relation statement (§1); App. Y sets FARS against each of these estimators in turn, and says why we report no number for function vectors. The tuned-lens line (Belrose et al., 2023) and the logit-lens tradition study the readout stage itself; closest to us, Nadaf (2026) find that function vectors (Todd et al., 2024) steer where no logit-lens decoder can read them, and Billa (2026) use logit-lens accessibility to predict steering success. Our test complements these by comparing concept and output-token subspaces with positive controls and an explicit depth-matched reference. Patchscopes (Ghandeharioun et al., 2024) decodes a hidden state through the model's own later layers and can surface content no affine map to the vocabulary recovers. Ours is the narrower instrument by design: it measures overlap with a specified *linear* readout span. Low overlap does not exclude decoding through lower singular directions, a learned decoder, or subsequent nonlinear processing.

Cross-model / cross-architecture representation. The Platonic Representation Hypothesis (Huh et al., 2024) conjectures cross-architecture convergence to a shared representational space; Kornblith et al. (2019) and Morcos et al. (2018) give correlation-based measures. We contribute a direct concept-direction-level empirical handle: a single ten-dimensional canonical frame (X-FARS) fitting twenty-seven LLMs at 97.6 % of cross-model centroid variance, with the frame's algebraic operations (differences, analogies, compositions) preserved cross-architecture.

Cross-lingual and cross-format transfer. Studies of multilingual representations address several distinct questions: language-specific neurons (Tang et al., 2024; Zhao et al., 2024), shared circuits and concept latents (Ferrando and Costa-jussà, 2024; Dumas et al., 2025; Chen et al., 2025), alignment across languages (Conneau et al., 2020; Liu and Niehues, 2025), latent language (Wendler et al., 2024), and transfer of knowledge-free reasoning (Hu et al., 2025). These findings motivate a controlled comparison across surface forms but do not, collectively, establish invariance across prose, code, and mathematical notation. OctoPack studies code instruction tuning (Muennighoff et al., 2024); our benchmarks instead hold concept identity fixed while varying these forms. We use them to test readout geometry and extraction transfer, without claiming an exhaustive priority result over all cross-format datasets.

Causal interpretability. Activation patching (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023) and its follow-ups (Lindsey et al., 2025; Conmy et al., 2023) localise circuits by intervention, and causal-abstraction accounts (Geiger et al., 2021) formalise what such interventions license. We use subspace-level patching (§4) with ablation and amplification as a bidirectional causal test; the interpretability illusions demonstrated by Makelov et al. (2024) caution against equating an intervention effect with recovery of the naturally used mechanism. Our geometric control addresses readout adjacency; it does not by itself resolve that broader causal-identification problem.

3 Experimental Setup

TriForm Benchmark. A controlled stimulus set where reasoning content is held exactly constant while surface form varies along three axes: **linguistic** (English, Chinese, French prose), **symbolic** (Python code, mathematical notation), and **structural** (step-by-step). 18 concepts spanning arithmetic, logic, relational, causal, and spatial domains $\times$ 3 instances

Model L= best FARS layer	Argmax preserved (%)↑ FARS-10	PCA-10	Full	Behavioural (drop pp)↓ GSM8K	MATH
OpenAI GPT-2 XL	**93.7**	87.7	81.0	—	—
Qwen-2.5-3B	**89.7**	37.3	22.0	—	—
Qwen-2.5-7B-Inst.	**94.7**	59.7	26.3	**−20.0** pp	**−18.4** pp
Mistral-7B	**97.3**	85.3	83.3	—	—
Llama-3.1-8B	**94.0**	56.7	39.7	+5 (n.s.); −22 at $c^\star$=2	—
Matched-rank random	99.0 ± 0.2	—	—	0.0	−1.7

Table 1: Argmax-preservation under 10-d cross-form patching (paired Wilcoxon $p<10^{-22}$ vs. Full-PCA on every model). Behavioural: FARS ablation at the FARS layer costs −20 (Qwen-7B-Inst), −28 (Phi-3.5), −17 (Qwen-3B-Inst), and −10 pp (Mistral-7B-Inst) on GSM8K relative to matched-rank random (paired Wilcoxon $p \leq 0.03$); Llama-8B needs amplification $c^\star$=2 (−22 pp, p=3×10^{-3}); Llama-70B is inert at $c \leq 2$ (App. Q.3).

× 6 forms = 324 stimuli, programmatically generated; full inventory in App. L, example in App. K.

Models. Twenty-six LLMs from four architecture families plus reasoning-tuned variants (0.4B–70B): sixteen dense decoders (GPT-2 XL, Qwen2.5-3B-Inst, Qwen2.5-7B, Qwen3-4B/8B, Phi-3.5-mini-Inst, Phi-4, Mistral-7B base/-Inst, Llama-3.1-8B/70B-Inst, OLMo-2-7B-Inst, Granite-3.3-8B-Inst, Falcon3-7B-Inst, Yi-1.5-9B-Chat, SmolLM3-3B), three MoE (Mixtral-8x7B-Inst, DeepSeek-V2-Lite-Chat, gpt-oss-20B), two state-space (Mamba-2.8B, Falcon-Mamba-7B), four reasoning-tuned (R1-Distill-Llama-8B, R1-Distill-Qwen-7B/14B, QwQ-32B), and one encoder MLM (DeBERTa-v3-large). RoBERTa-large is excluded from the readout analyses because no layer yields a concept-dominant subspace (form-RSA exceeds concept-RSA throughout), so the test's prerequisite is unmet. The X-FARS canonical frame (App. H) is fitted on all twenty-seven extracted models including RoBERTa; the algebra analyses that follow it use the original fifteen-model pool.

Methods. Last-token hidden states at every layer ($\mathbf{A} \in \mathbb{R}^{324\times L\times D}$). **Permutation RSA** (Kriegeskorte et al., 2008; Diedrichsen and Kriegeskorte, 2017) between the empirical cosine RDM and binary ideal RDMs (0 for a same-concept pair, 1 otherwise; likewise for form; 1,000 perms, FDR-corrected). **Cross-form probing** (ridge α=0.1 on source form, eval on target). **FARS extraction**: average each concept's activations over forms and instances to get 18 centroids in $\mathbb{R}^D$, PCA → top-k FARS basis $\mathbf{B}_\ell \in \mathbb{R}^{k\times D}$. All estimators in §5 are rank-matched at k=10 so that Grassmann distances are comparable. This matters where the natural rank is lower: *FormPCA* is the 5 directions of the six mean-centred form centroids, completed to rank 10 by the leading directions of the form-projected-out residual, and *ProbeW* is the top-10 PCA of the 18 one-vs-rest probe weight vectors. **Causal interventions**: subspace patching $\mathbf{h}_{\text{patched}} = \mathbf{h}_{\text{tgt}} + \mathbf{B}^\top\mathbf{B}(\mathbf{h}_{\text{src}} - \mathbf{h}_{\text{tgt}})$ equation 1, ablation $(\mathbf{I} - \mathbf{B}^\top\mathbf{B})\mathbf{h}$, and amplification $\mathbf{h} + c\,\mathbf{B}^\top\mathbf{B}\mathbf{h}$ (so c=0 is the identity and $c^\star$ denotes the gain used), compared against random orthonormal, Full-PCA, and full replacement; concept-block bootstrap (5,000).

$$\mathbf{h}_{\text{patched}} = \mathbf{h}_{\text{tgt}} + \mathbf{B}^\top\mathbf{B}\,(\mathbf{h}_{\text{src}} - \mathbf{h}_{\text{tgt}}) \tag{1}$$

4 Extracting FARS

Given a set of C concepts each rendered in F surface forms with I instances per pair ($N = CFI$ stimuli), FARS is the top-k principal subspace of the mean-centred concept-centroid matrix at the concept-RSA peak within the 30–75% depth band, as in the readout analyses. A *shuffled-centroid null* trails FARS by 17–29σ in concept-RSA and *multi-class LDA* matches it within $\Delta\leq$0.02, so the subspace is concept-carrying rather than a variance artefact (App. B).

The interventions show that modifying FARS can affect reasoning performance in several tested models. Argmax preservation under cross-form patching is a different endpoint: the

Rank-10 subspace	role	vs. W_U		vs. $W_U \to \ell$	
		d_G	energy %	energy %	$\theta_{\min}$
Final-layer PCA	**positive ctrl**	4.556	3.56	0.86	65.5
FormPCA	control	4.738	1.26	0.90	75.8
FullPCA	control	4.751	1.04	0.83	77.5
Shuffled-FARS	control	4.762	0.90	0.62	78.9
Diff-of-means	published	4.774	0.80	0.46	79.5
FARS *(ours)*	published	4.774	0.80	0.46	79.5
LDA	published	4.805	0.41	0.39	83.5
Probe weights	published	4.811	0.38	0.33	83.6
Haar random	control	4.823	0.32	0.30	84.3
Next-token PCA at ℓ	**positive ctrl**	4.665	2.34	**5.88**	69.6

Table 2: Readout-orthogonality of nine matched-rank subspace estimators, averaged over 26 models spanning four architecture families plus reasoning-tuned variants (0.4B–70B). d_G is the Grassmann geodesic distance to the top-10 right singular vectors of W_U (ceiling $\sqrt{10}\pi/2 = 4.97$); energy % is the fraction of subspace energy inside the top-10 readout span, measured against W_U directly and against the readout channels pulled back to the FARS layer ℓ by a tuned-lens translator ($W_U \to \ell$; App. E, 25 models); $\theta_{\min}$ is the smallest principal angle. Rows are ordered by readout-adjacency. *Published* estimators are the constructions used in the concept-direction literature; *control* rows probe rank and structure; the *positive controls* are the top-10 PCA of the *final* layer, whose residual stream is the immediate input to W_U, and next-token-identity centroid PCA at the FARS layer itself — the same estimator as FARS with the grouping variable changed from concept to emitted token.

random subspace also preserves the argmax, so preservation alone is not evidence of causal sufficiency. These observations motivate the next question: how much does the concept-carrying subspace overlap the model's dominant readout directions? After testing this with positive controls (§5), we examine transfer to a new concept inventory (§6), sensitivity to rank and layer (§7), and the scope of last-input-token measurements in reasoning models (§8). Cross-model alignment and algebra provide complementary evidence in App. H.

5 Readout-Orthogonality Test and Method Validation

Per-model readout orthogonality. For each model we extract FARS at its best concept-RSA layer within the 30–75% depth band, take the top-10 right singular vectors of that model's own unembedding matrix W_U (the raw test uses W_U alone; the final-norm gain $\boldsymbol{\gamma}$ is folded in, with the layer translator, by the depth-matched test of App. E, which returns the same verdict), and measure the Grassmann geodesic distance between them, the smallest principal angle, and the fraction of FARS energy lying inside the readout span.

Per-model values are tabulated in App. E (Table 9).

Positive controls and four concept estimators give different readout loadings. A diagnostic that can only return one answer is not a diagnostic. We therefore run the test on matched-rank subspaces of the same activation space across all 26 models: four estimators from the published literature, four structural controls, and *positive controls* designed to be readout-adjacent.

The test is two-sided. The final-layer positive control carries 3.56% of its energy in the top-10 readout span against 0.80% for FARS (a 4.5× separation), with the minimum principal angle falling from 79° to 65°; per-model it is strictly more readout-adjacent than *all four* published estimators in 25/26 models by energy and 25/26 by minimum angle. The near-orthogonal readings elsewhere are therefore substantive rather than an artefact of the statistic. The single exception is R1-Distill-Qwen-7B, whose last-input-token FARS loads 5.78% against a cross-model mean of 0.80%; it is the model whose concept computation §8 shows moving off the last-input-token position, and the depth-matched test below resolves it.

The verdict survives matching depth. A mid-layer subspace could be far from W_U merely because many layers separate it from the unembedding. We therefore pull the readout back to the FARS layer ℓ with a tuned-lens-style translator A_ℓ (ridge from block-ℓ to final-block states, held-out R^2 0.23–0.63) and take the top-k right singular vectors of $W_U \operatorname{diag}(\boldsymbol{\gamma}) A_\ell$ ($\boldsymbol{\gamma}$ the final-norm gain) as the readout channels *at layer* ℓ; as a same-layer positive control we build FARS's own estimator with the grouping variable changed from concept to emitted token (next-token centroid PCA). Against the pulled-back readout the concept estimators stay at the Haar floor (FARS 0.46%, LDA 0.39%, random 0.30%) while the same-layer next-token subspace loads 5.88% — a 13× separation that exceeds all four published estimators in 25/25 models (App. E). With layer, rank, estimator, and corpus held fixed, grouping by output token yields a readout channel and grouping by concept does not: this comparison separates grouping variables at matched depth under the fitted linear reference. It also resolves the exception above: R1-Distill-Qwen-7B's FARS loads 5.78% on the raw W_U channels but only 0.69% on the layer-ℓ channels; its relative loading depends on the choice of readout reference.

The verdict is not specific to our estimator. Multi-class LDA, difference-of-means directions, and one-vs-rest probe weights — the constructions the concept-direction literature actually uses — all have low top-10 readout loading, with LDA and probe weights *further* from the readout than FARS. Sorting by readout-adjacency recovers the form-vs-concept axis: form-dominant subspaces (FormPCA 1.26%, FullPCA 1.04%) sit nearer the readout than concept-dominant ones (FARS 0.80%, LDA 0.41%, probe weights 0.38%). This ordering associates surface-form grouping with greater readout loading in this dataset; it does not by itself identify where concept computation occurs.

Lemma 1 (Difference-of-means directions span FARS)**.** *Let μ_c be the class-conditional mean activation of concept c over a balanced stimulus set and $\bar{\mu}$ the grand mean. Then* $\operatorname{span}\{\mu_A - \mu_B\}_{A,B} = \operatorname{span}\{\mu_c - \bar{\mu}\}_c$*, so the top-$k$ principal subspace of the pairwise difference matrix is exactly the concept-centroid PCA subspace (FARS).*

Proof sketch. The two spans coincide by writing each generator in terms of the other, and the pairwise-difference scatter equals $2n$ times the concept-centroid scatter, so the two share eigenvectors and hence their top-k subspaces are identical, not merely co-spanning. Full derivation in App. B.1. □

Lemma 1 is why the *Diff-of-means* and *FARS* rows of Table 2 coincide to numerical precision (verified at $d_G < 10^{-7}$ on every model). The equivalence applies to the same class means, corpus, layer, and balanced pairwise construction. It does not imply that arbitrary steering directions extracted from other concepts or prompts lie in this FARS basis. Moreover, low average loading of a rank-10 subspace need not give the same loading to each individual direction. The published steering methods (Turner et al., 2024; Arditi et al., 2024) motivate this comparison, rather than inheriting its empirical result automatically.

Sensitivity to readout rank. The rank sweep in App. D compares loading against the exact Haar expectation k/D. Final-layer PCA exceeds FARS at the three reported ranks in 24/26 models; the separation generally narrows as rank grows. This is evidence at the tested ranks, not an order-of-magnitude separation at every rank.

Published direction constructions show the same depth sensitivity. Rebuilding the refusal direction of Arditi et al. (2024) and the truth direction of Marks and Tegmark (2024) at the FARS block on 22 models (held-out AUC ≥ 0.9 in 22 and 20), both have low loading relative to the Haar reference against the pulled-back readout (medians 0.44% and 0.20% vs. $k/D \approx 0.26\%$; low-loading criterion satisfied in 16/22 and 19/20), whereas the same directions built at the final block load 1.98% and 4.09% — a rise in 35/39 paired comparisons that change nothing but the extraction depth (App. F). The test applies unchanged to the directions safety and steering work manipulates.

The verdict is not unanimous: it tracks where the direction is defined. Building the same 18 concepts from the *vocabulary* instead — tf-idf-diagnostic tokens per concept, averaged over rows of W_U, PCA'd to rank 10, the construction behind token-anchored

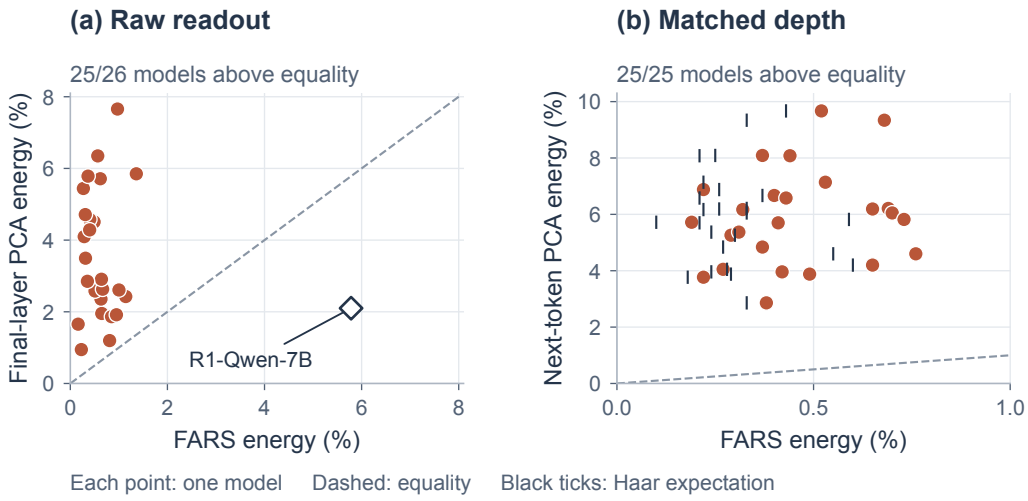

Figure 2: **Positive controls distinguish readout loading.** Each point is a model. (a) Final-layer PCA exceeds FARS in 25/26 models; the diamond marks R1-Distill-Qwen-7B. (b) Same-layer next-token PCA exceeds FARS in all 25 depth-matched models. Black ticks mark Haar expectations at the corresponding control energy. Dashed lines indicate equal loading; axis ranges differ. Panel (b) uses supplied table values rounded to 0.01 percentage point.

steering — yields a subspace that loads 8.43% of its energy in the readout span against 0.80% for FARS (mean per-model ratio 16.9×, same direction in 26/26 models; random-token control 5.20%), while overlapping FARS by only 0.56% (App. F.1). Activation-derived and vocabulary-derived constructions therefore differ in their overlap with the chosen readout span, even when they use the same concept labels.

It also covers the estimator the field reaches for first: concept features of a pretrained residual-stream sparse autoencoder (Bricken et al., 2023) sit at 1.06%, indistinguishable from FARS at the same layer (0.96%), while that SAE's full dictionary sits at 3.93% (App. F.2; one model).

6 Cross-Inventory Transfer: Novelty Benchmark

The preceding comparison fixes the concept inventory. We next ask whether the extraction procedure recovers cross-format structure when that inventory changes.

A natural worry is that FARS geometry is an artefact of the eighteen hand-crafted TriForm concepts rather than a cross-format signal in the models. We test this by extracting FARS on a disjoint inventory — the *Novelty benchmark* introduced here — and asking whether the method reproduces its defining property, cross-format concept retrieval.

Novelty benchmark. Ten concepts from five domains absent from TriForm (temporal, probabilistic, graph, data-structure, quantifier),

each with three instances rendered in the same six surface forms as TriForm, giving 180 stimuli (inventory in Table 22, App. L). We keep k=10 for comparability, which makes Novelty-FARS full-rank in its own concept space (ten centroids span at most nine dimensions), so it is read against the matched-rank random baseline, not against TriForm.

Per-model numbers are in Table 29 (App. P): Novelty-FARS reaches 62–100% top-1 cross-format retrieval on the 24 generative models (32% on the encoder DeBERTa-v3), 6–57 pp above a matched-rank Haar baseline on every model, with a cross-to-within-format retrieval

ratio $X/W \geq 0.94$ in 17 of 25 (X scores a stimulus against centroids from the *other* five forms and W against its own, so $X/W \to 1$ means retrieval ignores surface form). *Scope:* this analysis covers 25 of the twenty-six models. R1-Distill-Qwen-14B is the one checkpoint for which we did not run a Novelty extraction, so it is absent from this table only; it is present in every analysis that uses the TriForm inventory.

Two controls tighten the reading. First, a random rank-10 subspace reaches only 26–56% Novelty-Agn%, so the figures above are not a rank effect. Second, applying the TriForm-FARS basis to Novelty stimuli collapses to near-chance: a specific FARS is inventory-tied, but the *method* works on both.

Reading. The geometry documented on TriForm is not a product of TriForm's concept choices: a disjoint inventory drives the same method to the same result on the same models. What the paper reports is induced by concept-centroid PCA on any well-formed cross-format concept set.

7 Dimension Choice and Layer Localization

Transfer across inventories leaves two choices to check: the rank of the basis and the layer at which it is extracted.

$k = 10$ is an empirical operating point. Sweeping $k \in \{1, \ldots, 17\}$ on TriForm, $k = 10$ captures $\sim 90\%$ of both the full-rank ($k = 17$) centroid variance and the full-rank cross-format retrieval, with marginal gains ≤ 10 pp beyond it, uniformly across dense, state-space, and encoder-MLM models (App. I, Tables 17–18).

FARS is layer-local and its peak sits above the readout layer. For each of the 26 models we recompute concept-centroid PCA at every layer and record two quantities: top-1 cross-format concept retrieval Agn% (concept-invariance) and the Grassmann distance d_G from the layer-L FARS to the top-10 readout channels of W_U. Table 5 reports the Agn% peak along depth with its d_G, against the last layer (Fig. 5 and Table 5, App. A).

Reading. Retrieval peaks at different depths across models (reported depth fractions 0.29–0.97), so the result is not a universal middle-layer location. Peak-versus-final comparisons show lower retrieval at the final layer and generally smaller distance to the readout span. Peak selection itself favours higher retrieval; it is the accompanying geometry comparison that supplies the additional observation. These measurements constrain a layer-local interpretation but do not establish the complete causal pathway.

Cross-architecture algebra (App. H). The per-model subspaces of all 27 extracted models align into one shared canonical frame (X-FARS, 97.6% of cross-model centroid variance, 36.8σ above a random-projection control at $61.7 \pm 1.0\%$) that supports word2vec-style analogies, preserves pairwise differences, and enables training-free cross-architecture retrieval.

8 Two Concept Subspaces in Reasoning-Tuned Models

A last-input-token pipeline applied to reasoning-tuned models raises a scoping question: models that "think" in a chain of thought (CoT) may deposit concept content at positions a pre-decode pipeline never inspects. We read the *CoT-tail*: the last position generated under a 512-token budget, which binds for 64–88% of stimuli (App. **??**) — a claim about the reasoning stream, not its end. We answer it on the two open-weight R1-Distill models (Llama-8B, Qwen-7B) with their base counterparts, and on the native 32B reasoner QwQ.

A format-collapse control tempers this. Regenerating the traces and keeping their text shows that 56–57% of non-English prompts yield an English trace, so grouping by the prompt's form overstates how cross-format the position is. Regrouping the *cross-format* score X — not the agnostic top-1 rate above, but the score that pits a stimulus against centroids from

Probe	Llama pair	Qwen pair	QwQ-32B	Baseline
d_G(base last, distill last)	**3.06**	**3.48**	—	4.83 (cross-model)
d_G(last, CoT)	**4.53**	**4.26**	**4.33**	4.97 (rank-10 ceiling)
CoT-tail FARS % cross-format Agn.	**60.8**	**58.0**	**59.0**	21.6–24.4 (Haar rand.)
X-FARS pool +Llama-Distill	**55.6**% (+26.2 pp)		—	53.8% (15 models)

Table 3: Reasoning-tuned CoT departure at a glance. *Row 1* (Q1): last-input-token FARS is partially preserved from base to distilled counterpart, sitting well below the cross-model ambient reference $d_G = 4.83$ but far from zero. *Row 2* (Q2): within each reasoning model, the last-token and CoT-tail subspaces sit at large aggregate Grassmann distances, including the natively trained QwQ-32B. *Row 3*: the CoT-tail subspace itself carries cross-format concept identity at 58–61% top-1 retrieval, 34–39 pp above a matched-rank Haar-random baseline. *Row 4* (Q3): absorbing R1-Distill-Llama-8B into the shared X-FARS frame improves absolute alignment (53.8 → 55.6%) with the Haar-random-null gap ≥ 25 pp. Full three-question tables and ambient reference calibration in App. J.

other forms — by the form the model actually wrote in lowers it from 30.6–44.4% to 26.9–28.1%, still 3.3–5.1× a matched-rank Haar control measured the same way (2 models, a fresh set of generations; App. J.1).

Reading. Under reasoning fine-tuning the picture multiplies rather than breaks. The last-token FARS survives distillation with attenuation (Q1) and stays in the shared X-FARS family (Q3); Q2 exposes a second, geometrically distinct CoT-tail subspace that the pre-decode pipeline cannot see and that itself carries cross-format concept identity; native QwQ-32B reproduces it (d_G=4.33, 59.0%), so this is not a distillation or small-model artefact.

9 Discussion

FARS captures cross-format concept structure with low overlap with the dominant linear readout span in the tested settings. Final-layer and same-layer next-token controls show that this is a comparative result: output-oriented subspaces carry more readout energy. Together with the intervention results, this is consistent with concept representations contributing before final output readout. It does not identify the complete computation or prove that the subspace alone is sufficient to perform it.

The raw diagnostic is inexpensive once a basis is available, but the full pipeline is not compute-free: extraction requires activations, and depth matching requires fitting a translator. The latter explains only part of final-layer variation (held-out $R^2 = 0.23$–0.63), so it is an approximate linear reference. Rank sweeps test sensitivity to a chosen readout span, not an absence of information in all output directions; for a complete orthonormal basis at $k = D$, every subspace has unit loading.

The scope is also empirical. Labels supervise extraction, and a fixed basis does not transport to a new inventory (1–6% retrieval, against 6.5–17% for full activations; App. H); the extraction procedure does. The shared frame is fitted to a fixed inventory, rather than establishing universal concept coordinates. CoT-tail analyses sample a token-budget-limited reasoning stream, and an output-facingness prediction did not replicate (App. C). These boundaries make the diagnostic most useful as a control alongside targeted interventions and decoding tests.

Ethics Statement

We analyse only publicly released open-weight LLMs. No human-subjects data are collected or released, and the TriForm and Novelty stimuli comprise canonical reasoning patterns with no harmful content. One appendix experiment (App. R) does use published AdvBench prompts (Zou et al., 2023) to test whether a prose-vs-code representational asymmetry

produces a safety-refusal gap; it does not, and we release only aggregate refusal counts, no model completions and no new attack strings. Had the gap been real we would have deferred to responsible disclosure before release.

Reproducibility Statement

All artefacts are released with the submission (URL redacted): TriForm (324 stimuli) and Novelty (180 stimuli) with their generators, per-model activations, FARS and joint X-FARS bases, and pairwise Grassmann tensors. The readout-orthogonality test is a self-contained script requiring only the unembedding weight and any rank-k basis (App. Z). All twenty-six models are publicly released open-weight checkpoints; total reproduction compute is < 50 GPU-hours on a single A100-class device.

AI Use Statement

LLM-based assistants (Claude Code and OpenAI Codex) assisted with code development, literature retrieval, experimental planning, prose revision, figure design, and figure-data provenance checks. Codex also prepared editable vector and PowerPoint assets and screened references against primary-source records. No new experimental observations were generated during this revision. The authors are responsible for the experimental design, verification, interpretation, and final submitted content.

References

Yuxin Chen, Yiran Zhao, Yang Zhang, An Zhang, Kenji Kawaguchi, Shafiq Joty, Junnan Li, Tat-Seng Chua, Michael Qizhe Shieh, and Wenxuan Zhang. The emergence of abstract thought in large language models beyond any language. In *NeurIPS*, 2025.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *ACL*, 2020.

Clement Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *ACL*, 2025.

Javier Ferrando and Marta R. Costa-jussà. On the similarity of circuits across languages: A case study on the subject-verb agreement task. In *Findings of EMNLP*, 2024.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Large language models are cross-lingual knowledge-free reasoners. In *NAACL*, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *ICML*, 2024.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *ACL*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *ACL*, 2024.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? In *NeurIPS*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *ICLR*, 2023.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *CLeaR*, 2024.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *NeurIPS*, 2023.

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

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do Llamas work in English? On the latent language of multilingual transformers. In *ACL*, 2024.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508, 2017.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021.

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

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: how does LLM safety training fail? In *NeurIPS*, 2023.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *EMNLP*, 2022.

Ceyhun Efe Kayan and Li Zhang. Prototype-based dynamic steering for large language models. *arXiv:2510.05498*, 2025.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *ICLR*, 2023.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *NeurIPS*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv:2308.10248v5*, 2024. First posted in 2023.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *ICLR*, 2024.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv:2303.08112v4*, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019.

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

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, et al. OctoPack: Instruction tuning code large language models. In *ICLR*, 2024.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *ICML*, 2024.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *ICLR*, 2024.

Mohammed Suhail B Nadaf. Steerable but not decodable: Function vectors operate beyond the logit lens. *arXiv preprint arXiv:2604.02608*, 2026.

Jayadev Billa. Predicting where steering vectors succeed. *arXiv preprint arXiv:2604.15557*, 2026.

Reader's Guide to the Appendices

The appendices are organised into nine thematic tiers that mirror the body. Each tier collects the sections and subsections listed below; every result in the body has a pointer to a specific labelled appendix.

Tier (mirrors body)	Appendix sections
A. Benchmarks and stimuli	§K Stimulus Examples; §L TriForm Concept Inventory.
B. FARS extraction and causal validation	§B Extraction Baselines; §G Cross-Form Activation Patching; §M Subspace Extraction Details, including §M.1 instruction-tuning robustness and §M.2 OOD retrieval scope.
C. Readout-test robustness and localization	§C Concept-family ordering (negative result); §D Readout-rank sweep over the full W_U spectrum; §A Layer-level RSA, probes, prose-code asymmetry; §S Prose–code asymmetry controls (tokeniser and declarative-Python).
D. Cross-inventory transfer (Novelty)	§P Held-out concept generalisation across 15 models (companion to the body's Novelty benchmark).
E. Dimension, layer, and inter-layer stability	§I The k=10 plateau; §O Inter-layer FARS overlap; §N FARS plateau width → downstream benchmarks.
F. Cross-model universal frame and X-FARS	§H Universal frame, algebra, X-FARS retrieval; §V Canonical concept-axis system; §W Cross-model geometry predicts behaviour; §U Compositionality (pair/triplet/quartet).
G. Causal-necessity battery	§Q Probe-, token-, task-level and axis-specific ablation (four subsections); §T Head-level decomposition; §X Layer-writers; §R Safety refusal (falsified).
H. Reasoning-tuned models (CoT departure)	§J Full three-question deep dive (Q1 base↔distill, Q2 last↔CoT, Q3 X-FARS absorption, ambient reference calibration).
I. Positioning and code	§Y FARS vs. CCS, DAS, LEACE, LDA, SAEs, and why we report no function-vector number; §Z the self-contained test script.

Each claim, and the control that could have refuted it. The tiers above say where each topic lives. The table below says something different and, for a reader deciding how much to believe, more useful: for every contribution, which result carries it, what we built that would have shown it false, and what that attempt returned. Five of the nine entries returned a result we did not want. Fig. 3 then puts every subspace the table refers to on a single axis.

Claim	Carried by	Control that could have refuted it	What it returned
C1 two-sided test	Fig. 2, Table 2	A positive control built from the final layer, and a same-layer next-token subspace read against the readout pulled back by a tuned-lens translator. Either failing to register would mean the test cannot return "readout".	Both register: the final-layer control separates in 25/26 models, the matched-depth control in 25/25.
C2 locus verdict	Table 9, Table 2	The same concepts built from the *vocabulary* instead of activations, plus a random-token variant of that construction. If the test called these loci too, it would be reading nothing.	The split is clean: vocabulary-anchored lands on the readout side, activation-derived does not (App. F.1).
C2 on published directions	Table 11, Fig. 8	The same estimator and concept read off the final block instead of the mid block, and the held-out separability of that final-block direction.	Readout energy rises in nearly every paired comparison; directions with available AUC below 0.9 are excluded; missing final-block AUCs are flagged (App. F).
C3 transport	Table 29	A disjoint concept inventory, a matched-rank random subspace, and out-of-distribution retrieval on natural code.	Transports on the disjoint inventory; *fails* out of distribution, which is stated as a scope limit rather than omitted (App. M.2).
C4 algebra	App. H	A random-projection frame fitted the same way.	The canonical frame clears it by a wide margin.
C5 two subspaces	Table 3, Fig. 12	Whether the generated traces collapse to one surface form, which would make "cross-format" at this position a relabelling of within-format; and how often the token budget, not the model, ended the chain.	Collapse is real and partial: regrouping by the form the model actually wrote in lowers the score but leaves it several times a Haar control, and the truncation rate is disclosed (App. J.1).
Concept-family ordering	—	A prediction that the verdict tracks how output-facing a concept is.	*Did not replicate*; reported as such (App. C).
Causal sufficiency	Table 15	Writing a *different* concept's FARS component into a stream, against a matched-rank Haar patch of the same magnitude. If FARS determined which concept the model computes, the output should follow the source.	*It does not*: the edit is direction-specific (Haar is inert, FARS is not) but rarely installs the source concept, so we claim necessity and not sufficiency (App. G.1).
Prose–code safety gap	—	Whether the prose–code representational asymmetry produces a refusal gap on AdvBench.	*Falsified*; reported as such (App. R).

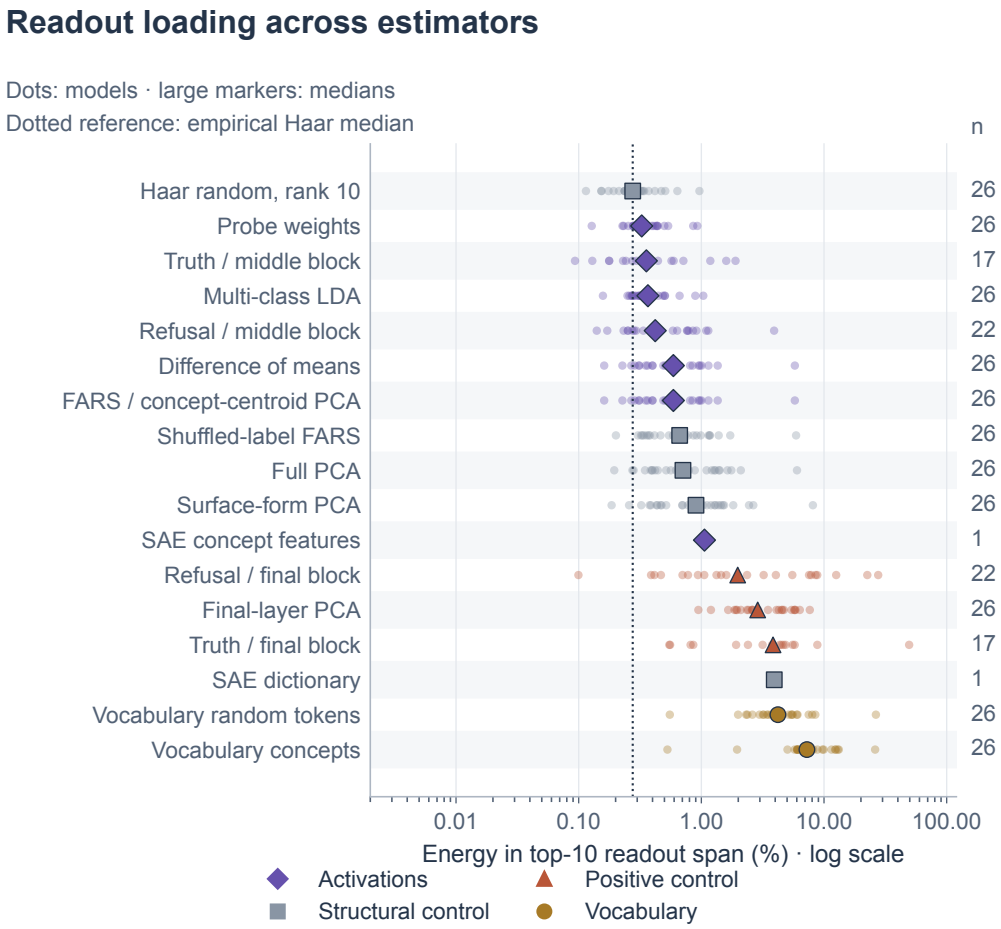

Figure 3: **Readout loading across subspace constructions.** Each small dot is one model and the larger role-specific marker is the median. The statistic is energy in the top-10 right singular span of W_U, on a logarithmic axis; the dotted reference is the median empirical Haar control. Sample sizes appear at right. Symbols distinguish activation-derived estimators, structural controls, positive controls, and vocabulary-derived constructions. Pulled-back readout results use a different reference and appear separately in App. E. Published directions use the same paired cohorts as Fig. 8: any available AUC below 0.9 excludes a pair; missing final-block AUCs remain flagged there. The pooled Haar median is a visual reference, not a model-specific significance threshold.

A Correlational Localization of Format-Agnostic Computation

Three independent analyses converge on middle layers as the site of format-agnostic representation before the body-level mechanistic tests are applied (Table 4, Figure 4): concept RSA peaks in middle layers and strengthens with scale ($\rho_{\max}$: 0.100 → 0.200; FDR-corrected); a ridge probe trained on English representations identifies the same concept in code or math at 60.3% for Mistral-7B (chance 5.6%), jumping between GPT-2 XL (32.7%) and Qwen2.5-3B (52.8%); and format-agnostic neurons peak at 20–23% in middle layers, with format-specific neurons dominating early (tokenisation) and late (generation) layers.

The prose-code boundary. Full-activation patching across form pairs reveals a sharp asymmetry: EN→Math top-10 overlap (0.65–0.73) is 3–4× higher than EN→Code (0.16–

Table 4: Localization of format-agnostic computation. RSA-C/F: peak concept/form RSA; Probe-W: within-form probe ceiling; Probe-X: cross-form probe (chance 5.6%); X/W: cross-/within-form ratio. Cross-form probing recovers 34–61% of the ceiling, jumping between 1.6B (0.34) and 3B (0.54).

Model	RSA-C	RSA-F	Probe-W%	Probe-X%	X/W	Agn%
OpenAI GPT-2 XL	.100	.430	96.6	32.7	.34	20.0
Qwen2.5-3B	.126	.501	98.5	52.8	.54	23.3
Qwen2.5-7B	.150	.485	99.1	59.8	.60	19.5
Mistral-7B	.200	.529	98.5	60.3	.61	20.8
Llama-3.1-8B	.165	.525	99.4	53.9	.54	21.1

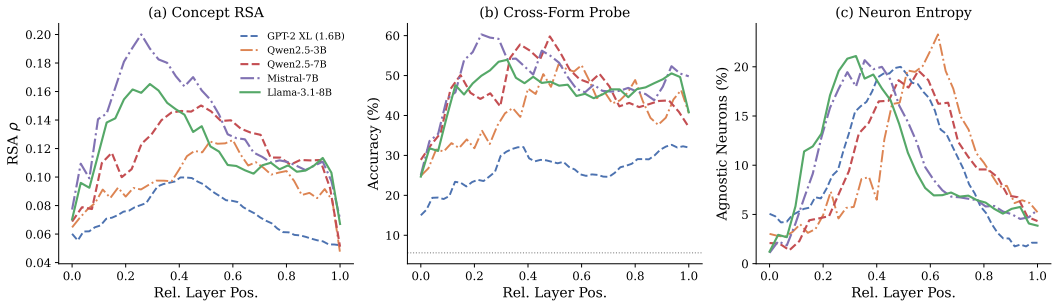

Figure 4: Three views of format-agnostic computation across layers. **(a)** Concept RSA peaks in middle layers. **(b)** Cross-form probe accuracy. **(c)** Format-agnostic neuron fraction.

0.22) across five tested models, with every 95% bootstrap CI above 2.4×. Two controls localise the cause: tokeniser Jaccard is near-zero-predictive (r=0.03, n.s.), and a declarative-Python control isolates the procedural-vs-declarative surface as accounting for ∼ 11% of the gap. Full details and per-model numbers in App. S.2.

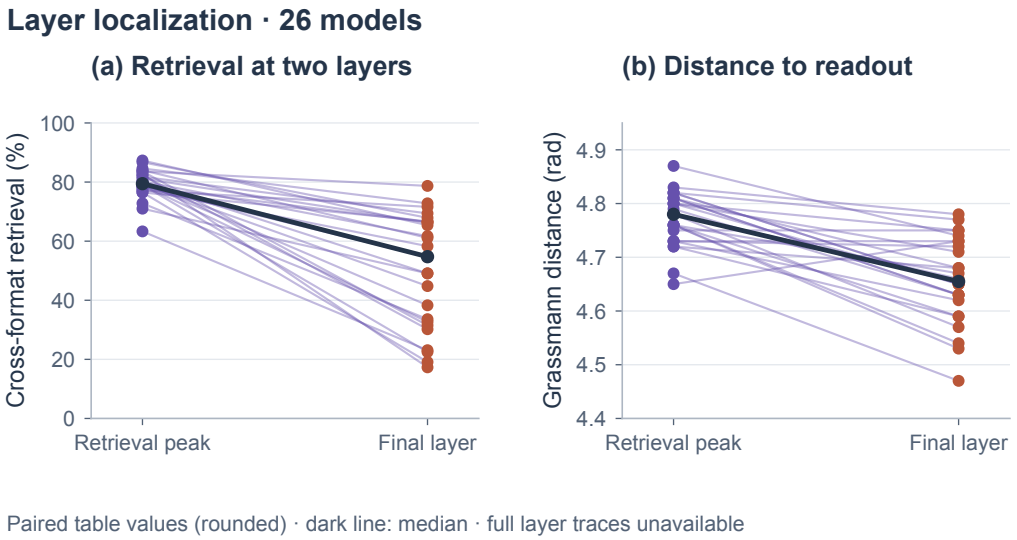

Figure 5: **Retrieval peak versus final layer.** Each thin line joins two measurements from one model; the dark line joins population medians. (a) Cross-format retrieval. (b) Distance to the top-10 readout span. These plots use the supplied rounded values in Table 5, not reconstructed layer-by-layer trajectories. Peak selection favours higher retrieval by construction.

Per-layer localization. Fig. 5 compares retrieval and readout distance at the retrieval peak and final layer for each model; Table 5 gives the per-model numbers.

Model	L_{peak}/L	frac.	Agn% peak	d_G peak	Agn% last	d_G last
Llama-3.1-8B-Inst	9/ 32	0.29	81.2%	4.79	66.4%	4.63
Mistral-7B	9/ 32	0.29	81.8%	4.81	69.4%	4.63
Llama-3.1-70B-Inst	24/ 81	0.30	87.3%	4.87	65.4%	4.74
Mistral-7B-Inst	11/ 33	0.34	84.0%	4.81	61.1%	4.63
Falcon3-7B-Inst	10/ 28	0.37	79.0%	4.82	32.7%	4.65
Mixtral-8x7B-Inst	12/ 33	0.38	82.4%	4.73	30.2%	4.73
Yi-1.5-9B-Chat	20/ 48	0.43	83.6%	4.82	31.5%	4.65
Granite-3.3-8B-Inst	17/ 40	0.44	77.2%	4.72	71.6%	4.68
R1-Distill-Llama-8B	14/ 33	0.44	83.6%	4.80	78.7%	4.75
Qwen2.5-7B	12/ 28	0.44	78.4%	4.76	54.6%	4.59
QwQ-32B	28/ 64	0.44	80.2%	4.81	49.1%	4.68
GPT-2 XL	22/ 48	0.47	63.3%	4.67	23.1%	4.47
DeepSeek-V2-Lite	13/ 28	0.48	79.3%	4.76	44.8%	4.54
Qwen3-8B	18/ 36	0.51	77.5%	4.78	58.3%	4.66
Phi-3.5-mini-Inst	18/ 33	0.56	78.4%	4.76	66.4%	4.53
gpt-oss-20B	13/ 24	0.57	86.7%	4.78	67.9%	4.57
Mamba-2.8B	39/ 65	0.61	76.5%	4.75	19.1%	4.75
DeBERTa-v3-large	16/ 25	0.67	79.6%	4.73	38.3%	4.72
Qwen2.5-3B-Inst	29/ 37	0.81	72.8%	4.73	33.6%	4.62
Falcon-Mamba-7B	56/ 65	0.88	82.7%	4.82	17.3%	4.77
Phi-4	35/ 40	0.90	71.0%	4.72	49.1%	4.59
OLMo-2-7B-Inst	29/ 33	0.91	84.6%	4.80	72.8%	4.65
R1-Distill-Qwen-14B	44/ 49	0.92	78.7%	4.83	54.9%	4.78
R1-Distill-Qwen-7B	26/ 29	0.93	79.0%	4.65	22.5%	4.73
Qwen3-4B	34/ 36	0.97	76.5%	4.76	66.7%	4.67
SmolLM3-3B	34/ 36	0.97	80.2%	4.80	61.7%	4.71

Table 5: Layer-localization of FARS across 26 models. *Left half* gives the layer L_{peak} at which cross-format concept retrieval Agn% peaks; *right half* gives the last-layer values. In every model the peak layer sits at depth fraction in $[0.29, 0.97]$ of the network, with $n_{\text{mid}} = 17/26$ peaks falling inside the middle band $[0.30, 0.85]$. At the peak layer, d_G to the top-10 readout channels is higher than at the last layer in $24/26$ models (median $\Delta d_G = +0.12$), and Agn% drops sharply toward the last layer. Concept-invariance and readout-adjacency move in opposite directions along depth, so the FARS peak is not the layer nearest the readout basis.

B FARS Extraction Baselines

B.1 Proof of Lemma 1

Let μ_c be the class-conditional mean activation of concept c over a balanced stimulus set, $\bar{\mu}$ the grand mean, $d_c = \mu_c - \bar{\mu}$, and n the number of concepts.

Equal spans. $\mu_A - \mu_B = d_A - d_B$ gives $\text{span}\{\mu_A - \mu_B\} \subseteq \text{span}\{d_c\}$, and $d_c = \frac{1}{n}\sum_j (\mu_c - \mu_j)$ gives the reverse inclusion.

Equal top-k subspaces. Equality of spans alone does not fix the principal subspaces, which depend on the second moments. Summing over ordered pairs and using $\sum_c d_c = 0$,

$$\sum_{A \neq B} (\mu_A - \mu_B)(\mu_A - \mu_B)^\top = \sum_{A \neq B} (d_A - d_B)(d_A - d_B)^\top = 2n \sum_c d_c d_c^\top .$$

The pairwise-difference scatter is exactly $2n$ times the concept-centroid scatter. Proportional PSD matrices share eigenvectors, so the two have the same ordered principal subspaces; in particular their top-k subspaces are identical for every k. This is what makes the *Diff-of-means* and *FARS* rows of Table 2 agree to numerical precision ($d_G < 10^{-7}$ on every model), and it is why the verdict transfers to the activation-steering family by construction.

Table 6: Concept (C) and form (F) RSA in the 10-dim FARS subspace vs. (i) shuffled-centroid PCA, (ii) supervised LDA, (iii) Full-PCA. FARS exceeds the shuffled null by 17–29σ and matches LDA.

Model	Metric	FARS	Shuffled	LDA	Full-PCA
OpenAI GPT-2 XL	RSA-C ↑	**.279**	.072 ± .008	.381	.091
	RSA-F ↓	**.107**	.473	−.011	.486
Qwen-7B	RSA-C ↑	**.357**	.128 ± .008	.384	.151
	RSA-F ↓	**.020**	.243	−.012	.261
Mistral-7B	RSA-C ↑	**.362**	.149 ± .013	.380	.194
	RSA-F ↓	**.012**	.259	−.012	.275

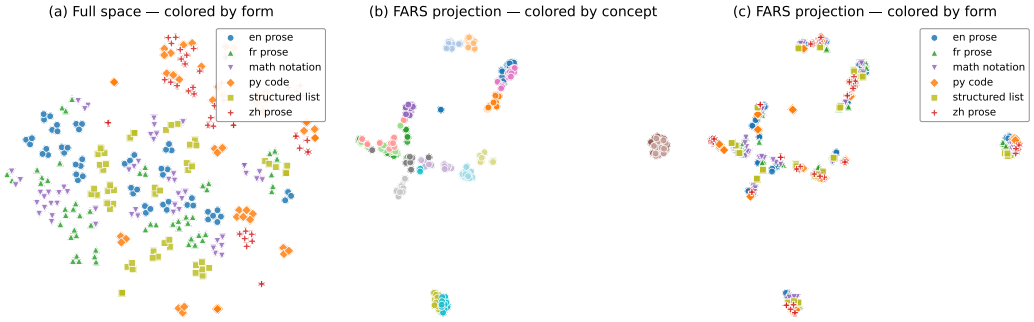

Figure 6: t-SNE visualisation (Mistral-7B, best FARS layer). **(a)** Full space: stimuli cluster by form. **(b)** FARS by concept: 18 concept clusters emerge. **(c)** FARS by form: forms intermixed within each cluster.

Concept-centroid PCA on per-layer centroids amplifies concept RSA 2–3× (e.g., 0.100 → 0.317 for GPT-2 XL) while suppressing form RSA to near zero. Figure 6 visualises this: in the full space stimuli cluster by form; in FARS they reorganise into 18 concept clusters with forms intermixed. Two baselines validate the extraction (Table 6): a *shuffled-centroid null* trails FARS by 17–29σ; and *multi-class LDA* (supervised gold standard) yields concept RSA within 0.02 of FARS, confirming that concept-centroid PCA recovers the same subspace at a fraction of the compute.

C A Prediction That Did Not Replicate: Concept-Family Ordering

A prediction that did not replicate. We had expected the readout verdict to track how output-facing a concept is. The reasoning is mechanical: a direction encoding "the next token is a numeral" is a statement about the logits and should sit inside the readout basis, whereas a direction encoding "this inference is valid" is a statement about the input and should not. To test it we built five balanced binary families of minimally contrasting sentences (20 pairs each), fixed their surface-to-abstract ordering before measuring anything, took the single direction $v = \mu_+ - \mu_-$ per family — the object the steering literature uses — and measured its readout enrichment $E_v(k)/(k/D)$ against the full singular basis of W_U.

The prediction fails. The ordering is fully monotonic in 1 of 6 models, and 41 of 60 pairwise orderings are correct against 30 expected by chance — above chance, but nowhere near the clean gradient the mechanism predicts. In particular the claim we would most have wanted, that an inference-validity direction sits at chance, holds only on `gpt2-xl` (1.03) and not on the larger models, where it reaches 38×.

The likely cause is visible in the table. On the larger models *every* family is enriched by an order of magnitude at the final layer: the smallest enrichment in any family averages 5.6× across the five 7–8B models, against 1.03× on `gpt2-xl`. That is expected on reflection — the final residual stream *is* the readout pathway, so a probe placed there saturates and the

Model	D	next token	language id	sentiment	factual tru	inference v	mono.
gpt2-xl	1600	14.4	8.2	4.0	1.4	1.0	✓
phi-3.5-mini-instruct	3072	13.5	35.4	18.0	9.7	7.2	—
qwen2.5-7b	3584	62.9	2.1	4.8	33.9	38.0	—
llama-3.1-8b-instruct	4096	17.3	2.1	4.6	3.8	1.5	—
mistral-7b-v0.3	4096	19.3	37.5	26.1	55.3	15.0	—
olmo-2-7b-instruct	4096	27.4	4.0	2.1	14.2	6.3	—

Table 7: Final-layer readout enrichment $E_v(10)/(k/D)$ of the difference-of-means direction for five binary concept families, across 6 models. Columns are in the pre-registered surface-to-abstract order, so the prediction is that each row decreases left to right; the last column marks whether it does. Haar null is 0.99.

contrast between concept families washes out. A mid-network probe avoids the saturation but asks about output-facing representations before the model has built them, which is why our first attempt at this experiment, run at the middle layer, also produced no ordering.

We report this because it bounds what the paper claims. The readout test's two-sidedness rests on the final-layer positive control and the rank sweep (§5, App. D), both of which replicate across every model we tested. It does *not* rest on the concept being abstract in some intuitive sense, and we have no evidence that the test orders concepts along that axis. Finding a probe position that separates output-facing from input-facing concept directions without saturating is open.

D Readout-Rank Sweep: the Verdict Does Not Depend on k

Is the concept subspace read out at *any* rank? Fixing $k = 10$ invites the objection that concept identity might be projected through lower-variance singular directions of W_U — say directions 50–200 — which a top-10 comparison would not see. We examine this by computing the *full* right-singular basis V of W_U and, for every rank k, the fraction of a subspace's energy inside $\mathrm{span}(V_{:,1:k})$,

$$E_S(k) \;=\; \tfrac{1}{m}\textstyle\sum_{i=1}^{m} \left\| V_{:,1:k}^{\top} s_i \right\|^2, \qquad S = \{s_1, \ldots, s_m\} \text{ orthonormal.}$$

For a Haar-random rank-m subspace $\mathbb{E}[E_S(k)] = k/D$ exactly, so $E_S(k)/(k/D)$ is an enrichment over chance with a known null of 1. We report FARS against the final-layer PCA (the actual readout pathway) and a Haar control.

The supplied results show separation at the reported ranks, with exceptions. Final-layer PCA is more enriched than FARS in 25/26 models at $k = 10$ and $k = 32$, and in 24/26 at $k = 100$. R1-Distill-Qwen-7B reverses the comparison at all three ranks; Mamba-2.8B also reverses it at $k = 100$ (1.05 versus 0.90, both near the null). The results do not establish zero readout information at any rank. At $k = D$, both energy and enrichment equal 1 for every subspace, so a separation cannot hold across the entire spectrum. The supplied table measures concentration at three ranks, with the Haar expectation providing the reference scale; full-rank curves are not independently verified here.

E Readout Orthogonality at Matched Depth

Per-model readout orthogonality against W_U. Table 9 gives the body's per-model values (FARS at its concept-RSA peak layer against the top-10 right singular vectors of W_U).

Why a depth-matched control is needed. The body compares a layer-ℓ subspace with the top-k right singular vectors of W_U. Because every layer between ℓ and the unembedding transforms the residual stream, a sceptic can read "far from W_U" as "not at the last layer" rather than "not a readout channel": the structural controls of Table 2 all sit within a few

Model	D	enrichment $E(k)/(k/D)$, FARS / final-layer k=10	k=32	k=100	Haar k=10
DeBERTa-v3-large	1024	0.83 / 1.2	0.92 / 1.1	0.93 / 1.1	1.03
DeepSeek-V2-Lite	2048	1.26 / 11.7	1.12 / 6.3	1.06 / 3.2	1.00
Falcon-Mamba-7B	4096	1.17 / 16.8	1.36 / 6.8	1.23 / 3.4	1.01
Falcon3-7B-Inst	3072	0.96 / 10.7	0.95 / 5.8	1.00 / 2.9	0.98
gpt-oss-20B	2880	0.77 / 15.7	0.88 / 10.3	0.83 / 5.4	1.00
GPT-2 XL	1600	2.18 / 9.4	1.61 / 5.4	1.26 / 2.8	1.01
Granite-3.3-8B-Inst	4096	4.68 / 9.9	2.44 / 7.0	1.57 / 4.4	0.98
Llama-3.1-70B-Inst	8192	1.32 / 13.6	1.18 / 7.9	1.04 / 3.7	1.02
Llama-3.1-8B-Inst	4096	2.01 / 18.5	1.58 / 11.9	1.34 / 5.0	1.02
Mamba-2.8B	2560	1.66 / 5.0	1.34 / 2.3	1.05 / 0.9	1.00
Mistral-7B-Inst	4096	1.64 / 18.7	1.31 / 9.3	1.13 / 4.8	1.01
Mistral-7B	4096	1.49 / 23.7	1.22 / 12.1	1.11 / 6.1	1.01
Mixtral-8x7B-Inst	4096	4.10 / 10.7	2.47 / 7.5	1.82 / 3.5	0.99
OLMo-2-7B-Inst	4096	1.26 / 19.3	1.25 / 9.9	1.25 / 4.7	0.98
Phi-3.5-mini-Inst	3072	1.74 / 19.5	1.90 / 9.8	1.40 / 4.6	1.05
Phi-4	5120	2.04 / 21.9	1.55 / 15.7	1.15 / 8.2	1.02
Qwen2.5-3B-Inst	2048	1.67 / 3.6	1.45 / 2.4	1.29 / 1.7	1.03
Qwen2.5-7B	3584	3.48 / 27.4	2.09 / 10.4	1.53 / 3.9	0.99
Qwen3-4B	2560	1.62 / 6.0	1.33 / 4.1	1.13 / 2.7	0.97
Qwen3-8B	4096	2.08 / 10.6	1.71 / 9.5	1.35 / 5.1	1.00
QwQ-32B	5120	3.41 / 13.5	2.05 / 6.8	1.41 / 2.9	1.00
R1-Distill-Llama-8B	4096	3.89 / 7.9	2.44 / 6.1	2.00 / 2.8	1.01
R1-Distill-Qwen-14B	5120	1.16 / 4.8	1.37 / 3.7	1.20 / 2.0	0.98
R1-Distill-Qwen-7B	3584	20.73 / 7.5	7.46 / 3.5	3.08 / 1.5	1.00
SmolLM3-3B	2048	1.31 / 6.0	1.24 / 4.5	1.29 / 2.4	1.02
Yi-1.5-9B-Chat	4096	1.44 / 11.7	1.12 / 8.3	1.03 / 4.8	0.99

Table 8: Readout enrichment at three ranks for the concept subspace (FARS) and the readout pathway (top-10 PCA of the final layer), across 26 models. The Haar column is the null check: a random rank-10 subspace has expected enrichment 1.00; the sampled controls are close (max deviation 0.046). Enrichment generally approaches the Haar expectation as rank grows; the magnitude and ordering vary across models. The same tabulated comparisons are visualized in Fig. 7.

tenths of a percent of the concept estimators, and d_G to W_U falls monotonically with depth (Fig. 5). The control here removes the depth confound in two steps.

Pulling the readout back to layer ℓ. On 29k–59k tokens of generic text per model (WikiText-103, held-out 10%) we fit a ridge translator $\mathbf{h}_L \approx A_\ell \mathbf{h}_\ell + \mathbf{b}$ from the block-ℓ residual stream to the final block's (held-out R^2 0.23–0.63, cosine 0.53–0.97), the linear tuned-lens construction of Belrose et al. (2023). The readout channels *at layer* ℓ are then the top-k right singular vectors of $W_U \,\mathrm{diag}(\boldsymbol{\gamma})\, A_\ell$ ($\boldsymbol{\gamma}$ the final-norm gain; mean-removal is included for LayerNorm models). A subspace that later layers rotate into the readout will load on these channels even though it is far from W_U itself.

This also answers the final-norm question for the raw test. The primary test of §5 uses the right singular vectors of W_U itself, whereas the residual stream is rescaled by the final LayerNorm or RMSNorm before it reaches W_U, so the raw test's channels live in the post-norm frame. One could object that the orthogonality we report is an artefact of ignoring that affine step. It is not: the pulled-back test on this page folds in exactly that step — the diagonal gain $\boldsymbol{\gamma}$, and mean-removal where the model uses LayerNorm — and returns the same verdict, with the concept subspace at the Haar floor and the same-layer next-token control an order of magnitude above it. Including the normalisation makes the test *stricter*, not more permissive, and the verdict survives it.

A positive control at the same layer, built the same way. FARS is concept-centroid PCA. We build the identical estimator organised by output identity instead: group the same

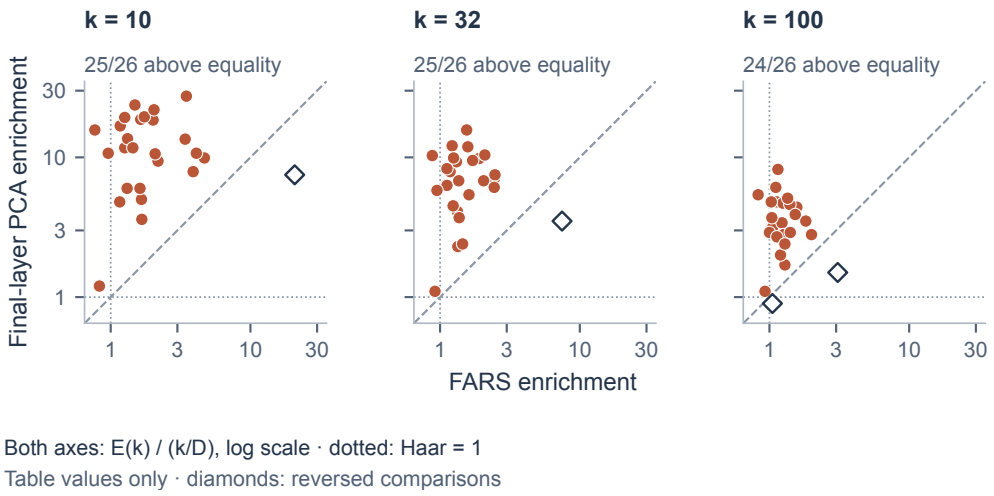

Figure 7: **Paired readout enrichment at three reported ranks.** Each point is one of 26 models, plotted from the rounded values in Table 8. Both axes use the same logarithmic scale. The dashed diagonal marks equal enrichment; dotted lines mark the Haar expectation of 1. Diamonds identify models where final-layer PCA is below FARS: R1-Distill-Qwen-7B at all three ranks, and Mamba-2.8B additionally at $k = 100$. No error bars or full-spectrum trajectories are inferred from these table values.

layer-ℓ token states by the token the model goes on to predict (top-200 predicted tokens, ≥ 20 occurrences each), average within class, and take the top-10 PCA of the class centroids (*next-token PCA*). It is matched in layer, rank, estimator, and corpus with FARS; only the grouping variable differs. The encoder MLM (DeBERTa-v3) is omitted here because its readout is a masked-token head rather than a next-token unembedding, so the pull-back is not defined in the same way.

Reading. Two things follow. First, the test discriminates *function* at fixed depth: with layer, rank, estimator, and corpus held constant, grouping by output token yields a readout-adjacent subspace and grouping by concept does not. Second, the pulled-back readout also repairs the one anomaly of the body's table: R1-Distill-Qwen-7B's last-input-token FARS loads 5.78% on the raw W_U channels but only 0.69% on the layer-ℓ channels, i.e. its apparent readout-adjacency was a depth artefact of comparing against W_U directly. FARS's residual enrichment over the Haar floor under the raw test (0.79% vs. k/D) likewise shrinks to 0.46% once depth is matched.

F Published Concept Directions Under the Test

Setup. The body's verdict covers subspaces we extract. To test directions the literature actually publishes, we rebuild two of them on each model, at the same block as FARS and from the last input token: the *refusal direction* of Arditi et al. (2024) (mean of 200 AdvBench harmful instructions minus mean of as many Alpaca harmless ones, chat-templated) and the *truth direction* of Marks and Tegmark (2024) (mean of true minus mean of false city–country statements). Each is a unit vector; we report its energy $\|R\mathbf{v}\|^2$ in the top-10 span of W_U and in the readout channels pulled back to the block (App. E), whose Haar expectation is k/D. A held-out AUC of the one-dimensional projection checks that the direction does separate its classes; entries with AUC < 0.9 are parenthesised and excluded from the summary. The same directions built at the *final* block (last two columns) show what a readout-adjacent version of each looks like.

Model	Family	$L/L_{\max}$	d_G	$\theta_{\min}$	FARS energy
Llama-3.1-70B-Inst	dense	23/80	**4.86**	85.6°	**0.16**%
OLMo-2-7B-Inst	dense	13/32	4.82	84.2°	0.31%
Falcon3-7B-Inst	dense	8/27	4.82	84.8°	0.31%
Mistral-7B-v0.3	dense	11/31	4.82	82.2°	0.36%
Mistral-7B-Inst	dense	10/32	4.81	81.3°	0.40%
Phi-4	dense	12/39	4.81	82.3°	0.40%
Yi-1.5-9B-Chat	dense	14/47	4.81	83.9°	0.35%
Llama-3.1-8B-Inst	dense	8/31	4.79	81.6°	0.49%
Qwen3-8B	dense	10/35	4.78	82.4°	0.51%
Phi-3.5-mini-Inst	dense	18/32	4.76	81.9°	0.57%
SmolLM3-3B	dense	10/35	4.76	81.5°	0.64%
qwen3-4b	dense	10/35	4.75	81.7°	0.63%
Qwen2.5-7B	dense	13/27	4.74	76.2°	0.97%
Granite-3.3-8B-Inst	dense	12/39	4.73	75.5°	1.14%
Qwen2.5-3B-Inst	dense	22/36	4.72	80.3°	0.85%
GPT-2 XL	dense	20/47	4.67	76.2°	1.36%
gpt-oss-20B	MoE	7/23	4.84	84.3°	0.27%
DeepSeek-V2-Lite	MoE	13/27	4.76	82.0°	0.62%
Mixtral-8x7B-Inst	MoE	10/32	4.76	74.4°	1.00%
Falcon-Mamba-7B	state-space	30/64	4.83	84.2°	0.29%
Mamba-2.8B	state-space	35/64	4.77	80.1°	0.65%
R1-Distill-Qwen-14B	reasoning	19/48	4.84	85.2°	0.23%
QwQ-32B	reasoning	19/63	4.79	77.5°	0.67%
R1-Distill-Llama-8B	reasoning	20/32	4.77	74.4°	0.95%
R1-Distill-Qwen-7B	reasoning	13/28	4.60	43.0°	5.78%
DeBERTa-v3-large	encoder-MLM	15/24	4.74	80.0°	0.81%

Table 9: FARS versus the top-10 readout channels of W_U, per model, across 26 LLMs (four architecture families plus reasoning-tuned variants). d_G is the Grassmann geodesic distance (ceiling $\sqrt{10}\pi/2 = 4.97$), $\theta_{\min}$ the smallest principal angle, and the last column the fraction of FARS energy inside the top-10 readout span. $L/L_{\max}$ is the FARS layer as a fraction of depth; every model sits in the middle band, which matters because a near-final layer would be readout-adjacent for reasons unrelated to the concept subspace. Distances span 4.60–4.86 and the minimum principal angle exceeds 74° in 25 of 26 models. The single outlier, R1-Distill-Qwen-7B, is the model our chain-of-thought analysis independently flags (§8); its loading falls to the Haar floor once the readout is pulled back to its FARS layer (App. E).

Reading. Two published, behaviourally validated directions — one safety-relevant, one truthfulness-relevant — receive the same verdict as the reasoning-concept estimators: low linear readout loading, in 16/22 (refusal) and 19/20 (truth) models under the pulled-back readout (criterion: energy $< 3k/D$). Their final-block counterparts are readout-adjacent in 17/22 and 17/20, so the test is not returning a constant for difference-of-means directions either.

This is the cleanest available demonstration that the test discriminates, because the paired comparison holds everything fixed except depth: same estimator, same concept, same model, same stimuli (Fig. 8). Readout energy rises from the FARS block to the final block in 19/22 models for refusal (median 5.3×, sign test $p = 0.0009$) and 16/17 for truth (median 8.2×, $p = 0.0003$). Final-block AUC is available for 19/22 models. Refusal AUC remains at least 0.98 throughout that subset. Four truth directions fall below 0.9: DeepSeek-V2-Lite (0.8988), GPT-2 XL (0.6828), Qwen2.5-7B (0.8692), and R1-Distill-Qwen-7B (0.5948). The last also fails at the FARS block, as does Phi-4, whose final-block AUC is unavailable. Filtering on all available AUC values leaves the 22 refusal and 17 truth pairs shown, including respectively three and two pairs without final-block AUC; these missing checks limit the classifier-quality control. A direction is not readout-aligned or not by virtue of which concept it encodes; it is readout-aligned to the extent that it is read off the stream where the stream is about to be decoded. The test therefore transfers, without modification, to the directions steering and safety work actually manipulate.

			energy % in top-10 of W_U			energy % in top-10 of W_U diag(γ)A_ℓ				
Model	ℓ/L	R^2	FARS	NextTok$_\ell$	Last-PCA	FARS	LDA	FullPCA	Haar	NextTok$_\ell$
Phi-4	12/39	0.40	0.40	1.19	4.29	0.22	0.25	0.61	0.26	**6.88**
Llama-3.1-70B-Inst	22/79	0.37	0.16	0.72	1.65	0.19	0.19	0.45	0.10	**5.72**
Mistral-7B-Inst	9/31	0.34	0.40	6.57	4.58	0.37	0.29	1.03	0.25	**8.09**
Falcon-Mamba-7B	29/63	0.50	0.28	0.45	0.31	0.32	0.28	0.61	0.22	**6.17**
Llama-3.1-8B-Inst	8/31	0.44	0.49	0.84	4.51	0.29	0.24	0.52	0.30	**5.26**
SmolLM3-3B	10/35	0.46	0.64	1.41	2.91	0.52	0.49	1.10	0.43	**9.67**
Mistral-7B	11/31	0.37	0.36	6.56	5.78	0.44	0.33	1.05	0.21	**8.08**
OLMo-2-7B-Inst	12/31	0.25	0.31	0.91	4.71	0.31	0.29	0.57	0.24	**5.37**
R1-Distill-Qwen-14B	18/47	0.38	0.23	0.60	0.95	0.22	0.24	0.48	0.18	**3.77**
Mamba-2.8B	34/63	0.58	0.65	3.06	1.95	0.40	0.36	0.62	0.37	**6.67**
R1-Distill-Llama-8B	19/31	0.48	0.95	1.74	1.92	0.43	0.23	0.91	0.21	**6.58**
Granite-3.3-8B-Inst	12/39	0.23	1.14	2.79	2.43	0.27	0.26	0.71	0.28	**4.05**
QwQ-32B	19/63	0.33	0.67	2.78	2.63	0.41	0.39	0.71	0.21	**5.70**
Qwen2.5-7B	13/27	0.46	0.97	3.42	7.66	0.68	0.42	1.49	0.33	**9.34**
Mixtral-8x7B-Inst	9/31	0.37	1.00	5.22	2.61	0.53	0.44	1.01	0.22	**7.14**
Yi-1.5-9B-Chat	14/47	0.43	0.35	1.17	2.85	0.37	0.32	0.58	0.27	**4.84**
Falcon3-7B-Inst	8/27	0.38	0.31	0.41	3.50	0.65	0.52	1.05	0.26	**6.19**
Qwen3-8B	10/35	0.41	0.51	0.88	2.58	0.42	0.40	0.59	0.24	**3.96**
R1-Distill-Qwen-7B	12/27	0.47	5.78	6.22	2.10	0.69	0.68	0.96	0.33	**6.21**
qwen3-4b	10/35	0.41	0.63	1.08	2.36	0.70	0.56	0.94	0.33	**6.05**
Phi-3.5-mini-Inst	17/31	0.61	0.57	2.38	6.35	0.49	0.33	0.87	0.29	**3.88**
Qwen2.5-3B-Inst	21/35	0.43	0.81	1.41	1.74	0.73	0.63	1.31	0.59	**5.82**
gpt-oss-20B	7/23	0.44	0.27	0.32	5.44	0.38	0.54	0.62	0.33	**2.86**
GPT-2 XL	20/47	0.63	1.36	5.27	5.85	0.65	0.65	0.90	0.60	**4.20**
DeepSeek-V2-Lite	12/26	0.43	0.62	1.17	5.71	0.76	0.51	1.13	0.55	**4.60**

Table 10: Readout loading at matched depth across 25 models. Left block: the body's test (readout = top-10 right singular vectors of W_U). Right block: the same subspaces against the readout channels pulled back to the FARS layer ℓ. Under the pulled-back readout the concept estimators stay at the Haar floor (FARS 0.46% vs. random 0.30%; the Haar expectation is k/D), whereas the same-layer, same-estimator next-token control loads 5.88% — 12.9× FARS — and exceeds all four published estimators in 25/25 models. R^2 is the held-out fit of the translator.

F.1 A vocabulary-defined construction, which the test flags as readout

A readout-side construction from the same literature. The estimators of Table 2 are all built from *activations*, and all land on the locus side. A diagnostic whose verdict never changes across the constructions people use is hard to act on, so we add the other kind of "concept direction" the literature also writes down — one defined in the model's *vocabulary*. For concept c we take the tokens diagnostic of its stimuli (tf-idf over c's six surface forms against the other concepts, top 40), average the corresponding rows of W_U, and take the top-k PCA of the 18 resulting centroids. This is the construction behind token-anchored steering and logit-attribution readings of "the direction of c". It needs no forward pass: it is a function of W_U and the tokenizer alone.

Reading. The split the test reports is not between our estimators and other people's. It is between *where the direction is defined*: a concept direction read off the activations has low readout loading, while the vocabulary-derived construction has greater loading. Both are called "the direction of concept c" in published work, and an ablation of either will disturb the task; only the second is disturbed *because it is the channel the unembedding reads*, which is the inference the test exists to block.

Two qualifications keep this honest. First, most of the effect comes from the *anchoring* rather than from the concept selection: the random-token control, which repeats the construction with arbitrary tokens, reaches 5.20% against 8.43% for concept-diagnostic tokens and exceeds it outright in 4 of 26 models. What makes these directions readout-adjacent is that they are assembled from rows of W_U at all, not which rows. Second, the effect size varies widely (1.7×–77.3×). The narrowest ratio is R1-Distill-Qwen-7B, where it is the *numerator's* denominator that is unusual — that model's FARS is the one already flagged in Table 9. More informative is Falcon-Mamba-7B, whose vocabulary-anchored subspace sits at 0.53% against a Haar floor of 0.24%: there the construction is *not* readout-adjacent and the test says so. The verdict is a per-model measurement, not a law.

		AUC		Haar	vs. W_U		vs. $W_U \to \ell$			final block, vs. W_U	
Model	ℓ	ref.	truth	k/D	refusal	truth	refusal	truth	NextTok$_1$	refusal	truth
DeepSeek-V2-Lite	13	1.00	1.00	0.49	0.59	1.42	1.04	1.01	1.48	7.55	8.37
Falcon-Mamba-7B	29	1.00	1.00	0.24	0.40	0.36	2.12	0.11	2.81	1.06	5.77
Falcon3-7B-Inst	8	1.00	1.00	0.33	0.41	0.59	1.06	0.18	0.02	1.33	5.54
gpt-oss-20B	7	1.00	1.00	0.35	0.29	0.24	0.51	0.15	2.61	4.05	3.84
GPT-2 XL	20	1.00	1.00	0.62	3.92	0.33	4.28	0.28	0.30	7.86	1.44
Granite-3.3-8B-Inst	12	1.00	1.00	0.24	0.43	1.90	0.19	0.16	0.05	0.41	0.82
Llama-3.1-8B-Inst	8	1.00	1.00	0.24	0.27	1.18	0.30	0.22	4.80	3.22	4.86
Mamba-2.8B	34	1.00	1.00	0.39	1.10	0.36	1.13	0.20	6.45	12.57	8.82
Mistral-7B-Inst	9	1.00	1.00	0.24	0.77	0.18	0.42	0.14	0.01	0.78	3.16
Mistral-7B	11	1.00	1.00	0.24	0.85	0.13	2.73	0.12	0.01	22.54	2.40
OLMo-2-7B-Inst	12	1.00	1.00	0.24	0.17	0.09	0.47	0.10	5.05	0.94	4.50
Phi-3.5-mini-Inst	17	1.00	1.00	0.33	0.25	0.30	0.22	0.11	7.28	27.65	49.43
Phi-4	12	1.00	0.50	0.20	0.28	(0.76)†	0.20	(0.12)†	0.04	1.45	(1.92)†
Qwen2.5-3B-Inst	21	1.00	1.00	0.49	1.14	0.71	0.24	0.52	8.22	1.60	4.35
Qwen2.5-7B	13	0.99	1.00	0.28	0.34	0.88	0.58	0.29	0.03	5.52	22.47
Qwen3-4B	10	1.00	0.98	0.39	0.91	0.27	0.40	0.40	6.45	0.70	1.92
Qwen3-8B	10	1.00	0.99	0.24	0.23	0.23	0.23	0.20	3.98	0.47	0.86
R1-Distill-Llama-8B	19	1.00	1.00	0.24	0.79	1.60	0.16	0.75	7.66	8.53	3.77
r1-distill-qwen-14b	19	1.00	1.00	0.20	0.25	0.18	0.64	0.19	1.17	0.39	0.56
R1-Distill-Qwen-7B	12	1.00	0.52	0.28	0.76	(27.52)†	1.22	(0.28)†	0.06	8.81	(14.18)†
SmolLM3-3B	10	1.00	1.00	0.49	0.64	0.57	0.38	0.44	12.13	0.10	4.67
Yi-1.5-9B-Chat	14	1.00	1.00	0.24	0.14	0.44	0.17	0.18	13.36	2.36	0.55

Table 11: Published concept directions under the readout test (22 models; energies in %). At the FARS block both the refusal and the truth direction sit at the Haar floor against W_U (medians 0.42% and 0.36%, $k/D \approx 0.26\%$) and against the pulled-back readout (0.44% and 0.20%), while the first next-token direction at the same block loads 2.71%. The same two directions built at the final block load 1.98% and 4.09% (medians): the middle-layer construction has low linear readout loading and becomes readout-adjacent only where the residual stream is about to be decoded. †AUC < 0.9.

F.2 Sparse-autoencoder feature directions

Setup. Sparse autoencoders (Bricken et al., 2023; Templeton et al., 2024) are the estimator the field now reaches for first, so the test has to say something about them. We build the SAE analogue of FARS on the same stimuli, at the SAE's own layer: encode the cached layer-19 residual states with the pretrained residual-stream SAE of `Goodfire/Llama-3.1-8B-Instruct-SAE-l19` (65,536 features), score every feature by how selectively it fires on each concept (mean activation on c minus mean on the rest, in pooled s.d. units), take each concept's top-16 features, sum their *decoder* directions weighted by selectivity, and take the top-10 PCA of the 18 concept directions. The recipe is concept-centroid PCA with the representation swapped for SAE feature space. Our encoding reproduces the released sparsity (mean $L_0 = 98$ against the model card's 91; fraction of variance unexplained 0.59 on this out-of-distribution stimulus set), so the comparison is against a correctly instantiated SAE rather than an assumed one.

Reading. The verdict on SAE concept features is the same as on concept-centroid PCA, LDA, difference-of-means and probe weights: low linear readout loading. The two subspaces are related but not identical (mutual energy overlap 26%), so this is a genuine second estimator agreeing, not a restatement of FARS. *Scope:* one model. Of the 26 models in Table 9 this is the only one with a published residual-stream SAE we could match exactly in both checkpoint and layer, and we do not generalise beyond it.

G Causal Validation: Cross-Form Activation Patching

Replacing only the 10 FARS dimensions during cross-form activation patching preserves the model's argmax in 89.7–97.3% of cases across five models (Table 14). Full-PCA of the same rank retains 37.3–87.7%; full-activation replacement collapses to 22.0–83.3%. Paired Wilcoxon rejects FARS=Full-PCA at $p<10^{-22}$ on every model (Holm-corrected).

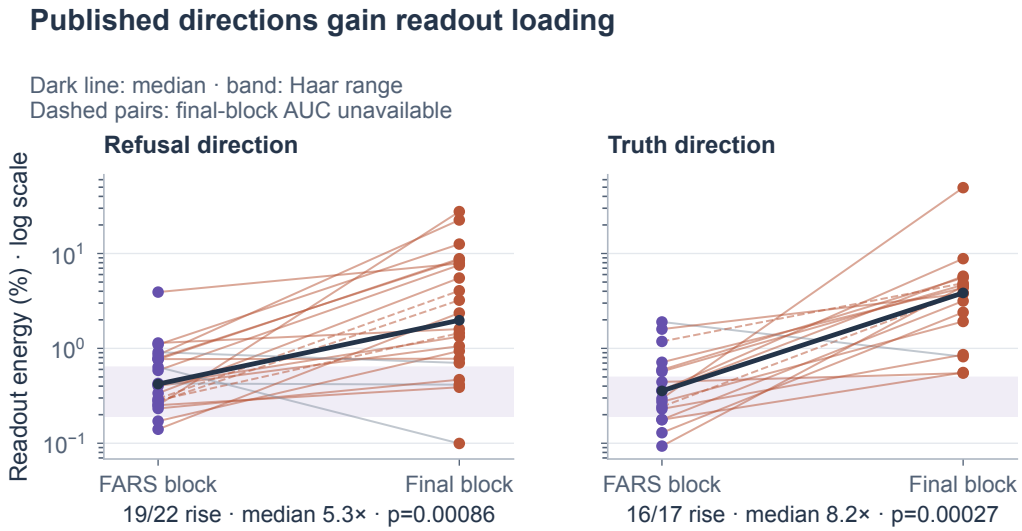

Figure 8: **The same published direction changes verdict with extraction depth.** Each line is one model: the energy its refusal (left) or truth (right) direction carries in the top-10 span of W_U, read off the FARS block versus the final block. Estimator, concept, stimuli and model are identical along each line; only the depth changes. Band: the range of per-model Haar expectations k/D; dark line: population median. Grey lines fall. Directions with available held-out AUC < 0.9 at either block are omitted. Dashed pairs lack a final-block AUC (three refusal, two truth) and are retained with this qualification.

G.1 Cross-concept patching: the sufficiency question

Why the cross-form patch is not yet a sufficiency result. Moving concept c from one surface form into another preserves the target's prediction, which shows FARS carries nothing form-specific. It does not show that writing into FARS *installs* a concept: the target was already computing c, so preservation is consistent with the patch doing nothing at all. The sharper test writes a *different* concept in.

Protocol. For ordered pairs of distinct concepts (A, B) rendered in the same surface form, we replace the target's FARS-projected component at the FARS layer with the source's, $\mathbf{h}_{\text{tgt}} \leftarrow \mathbf{h}_{\text{tgt}} + \mathbf{B}^\top\mathbf{B}(\mathbf{h}_{\text{src}} - \mathbf{h}_{\text{tgt}})$, and compare the patched next-token distribution with the two clean ones: $d_A = \text{KL}(\text{clean}_A \,\|\, \text{patched})$ and d_B likewise. *Moved* counts pairs with $d_A < d_B$. The control patches a matched-rank Haar subspace from the same activations, holding the magnitude of the intervention fixed and changing only which directions carry it.

What this does and does not license. Two things follow. FARS is causally load-bearing in a direction-specific way: an identical-magnitude edit through a matched-rank random subspace leaves the output where it was, and through FARS it does not. But a single-layer write does *not* reliably install the source concept — the patched run more often still resembles the target. We therefore claim necessity (App. Q) and direction-specific causal influence, and we do *not* claim that FARS is sufficient to determine which concept the model computes. A sufficiency result would likely need the write repeated across the layer band over which FARS is stable (App. O) rather than at one layer; we leave that open rather than assert it. *Scope:* 1 model.

		Haar	energy % in top-10 readout span			
Model	D	k/D	FARS	random-token	vocabulary-anchored	ratio
Llama-3.1-70B-Inst	8192	0.12	0.16	2.37	**12.47**	77.3×
SmolLM3-3B	2048	0.49	0.64	26.49	**26.11**	40.7×
R1-Distill-Qwen-14B	5120	0.20	0.23	1.99	**6.12**	27.0×
Llama-3.1-8B-Inst	4096	0.24	0.49	4.27	**12.94**	26.4×
gpt-oss-20B	2880	0.35	0.27	5.99	**6.95**	25.9×
OLMo-2-7B-Inst	4096	0.24	0.31	4.60	**6.47**	21.0×
Phi-4	5120	0.20	0.40	3.82	**8.17**	20.5×
Mistral-7B	4096	0.24	0.36	5.45	**7.04**	19.3×
Mistral-7B-Inst	4096	0.24	0.40	5.42	**7.03**	17.6×
Phi-3.5-mini-Inst	3072	0.33	0.57	8.01	**9.70**	17.2×
Yi-1.5-9B-Chat	4096	0.24	0.35	2.61	**6.03**	17.1×
Qwen3-8B	4096	0.24	0.51	2.32	**7.91**	15.6×
R1-Distill-Llama-8B	4096	0.24	0.95	4.33	**13.20**	13.9×
QwQ-32B	5120	0.20	0.67	3.22	**8.83**	13.3×
Qwen2.5-7B	3584	0.28	0.97	6.08	**12.32**	12.7×
Qwen3-4B	2560	0.39	0.63	2.98	**6.12**	9.7×
Mamba-2.8B	2560	0.39	0.65	5.14	**6.08**	9.4×
DeBERTa-v3-large	1024	0.98	0.81	7.50	**7.47**	9.2×
Qwen2.5-3B-Inst	2048	0.49	0.85	3.43	**7.57**	8.9×
GPT-2 XL	1600	0.62	1.36	8.49	**11.51**	8.5×
DeepSeek-V2-Lite	2048	0.49	0.62	4.15	**5.04**	8.2×
Falcon3-7B-Inst	3072	0.33	0.31	3.18	**1.96**	6.3×
Mixtral-8x7B-Inst	4096	0.24	1.00	3.71	**6.00**	6.0×
Granite-3.3-8B-Inst	4096	0.24	1.14	3.51	**5.75**	5.0×
Falcon-Mamba-7B	4096	0.24	0.29	0.55	**0.53**	1.9×
R1-Distill-Qwen-7B	3584	0.28	5.78	5.56	**9.95**	1.7×

Table 12: Vocabulary-anchored concept directions against the readout basis, 26 models. The construction lies in the row space of W_U by design, so some loading is guaranteed; the honest baseline is therefore the *random-token* column, which repeats the identical construction with random token sets of the same size (20 draws). Concept-diagnostic tokens load 8.43% on average against 5.20% for random tokens and 0.80% for FARS at its own layer — a 16.9× separation from FARS, in the same direction in 26/26 models, with per-model ratios spanning 1.7×–77.3×. The two subspaces are not the same object: their mutual energy overlap is 0.56%.

Rank-10 subspace	d_G	energy %	$\theta_{\min}$
SAE dictionary (top-10 PCA of live decoder directions)	4.590	3.93	56.3
SAE concept features (selectivity-weighted, rank 10)	4.748	1.06	73.8
FARS at the same layer	4.759	0.96	75.1
SAE random features (control)	4.749	0.88	76.5
Haar expectation k/D	—	0.24	—

Table 13: SAE feature directions against the top-10 readout channels of W_U on Llama-3.1-8B-Inst. The SAE's concept-selective features have low linear readout loading at 1.06%, indistinguishable from FARS at the same layer (0.96%) and from randomly chosen features (0.88%). The dictionary as a whole is different: the top-10 PCA of all live decoder directions sits at 3.93%, roughly 4× its own concept features — a dictionary trained to reconstruct the residual stream inherits whatever the readout writes into it, while the features that carry concept identity do not.

H Cross-Model Universal Frame, Algebra, and X-FARS Retrieval

The extracted per-model FARS subspaces admit a further, orthogonal validation. They align pairwise (centroid-RSA > 0.90 within family, > 0.77 cross-family on the original fifteen-model pool; Fig. 10) and jointly into a single shared canonical frame (X-FARS) that, refitted on all 27 extracted models, captures **97.6**% of cross-model centroid variance (36.8σ above a same-rank random-projection control at 61.7 ± 1.0%; the original fifteen-model fit gave 97.0%, 21.6σ). The algebra results below were computed on the fifteen-model pool. Held-out concept controls at $K \in \{3, 6, 9\}$ concepts preserve RSA 0.28–0.40 and probe

Table 14: Causal validation under three metrics on the same 10 dimensions: top-1 next-token match, top-10 overlap, median KL of patched vs. clean. FARS cross-form patching matches within-form same-instance ceiling (0.85 on Qwen-7B-Inst).

Model	Top-1% ↑			Top-10 ov. ↑			KL med. ↓		
	FARS	PCA	Full	FARS	PCA	Full	FARS	PCA	Full
OpenAI GPT-2 XL	**93.7**	87.7	81.0	**.90**	.65	.47	**.016**	.154	.655
Qwen-3B	**89.7**	37.3	22.0	**.92**	.60	.44	**.012**	.612	2.77
Qwen-7B	**94.7**	59.7	26.3	**.94**	.63	.48	**.006**	.262	.683
Mistral-7B	**97.3**	85.3	83.3	**.96**	.74	.56	**.002**	.034	.142
Llama-8B	**94.0**	56.7	39.7	**.94**	.64	.47	**.004**	.098	.811

		moved to A (%)		median d_B		
Model	pairs	FARS	Haar	FARS	Haar	ratio
GPT-2 XL	40	**5**	0	**0.024**	0.001	19×

Table 15: Cross-concept patching. d_B is how far the patched output moves from the target's clean output: a patch that does nothing leaves it at zero. The Haar control is essentially inert, while the same intervention through FARS perturbs the output 19–19× more — the effect is specific to these directions, not to the magnitude of the edit. The output nonetheless usually stays nearer the target's concept than the source's (5–5% of pairs move to A, against 0% for Haar).

accuracy 29–43%; the frame is neither a one-model artefact nor an over-fit to the specific 18 concepts.

Algebraic operations: differences, analogies, compositions. Within the shared frame, concept directions support the full set of algebraic operations. (i) **Differences** $\mathbf{c}_A - \mathbf{c}_B$ are preserved at mean cosine **0.88** (median 0.97) across $\binom{18}{2} \times \binom{27}{2} = 53{,}703$ cross-model measurements (**200.2**σ above the shuffled-concept null; 97.7% of pairs positive). (ii) **Word2vec-style analogies** $\mathbf{c}_A - \mathbf{c}_B + \mathbf{c}_C \approx \mathbf{c}_D$ resolve to the expected concept at mean top-1 **37.4**% across $27 \times 10 = 270$ attempts (**17.0**σ above a shuffled null at 6.6%; Figure 11). (iii) **Compositions** (LRH) hold at pair, triplet (six generative architectures, span 0.68–0.92), and quartet (Mamba: span 0.87, p=1.5×10^{-5}) level. The encoder MLM DeBERTa-v3 marks a clean architectural scope: triplet LRH fails (span 0.66, all coefficients negative). The FARS basis trained on a random 15 of 18 concepts further *generalises to the held-out* 3 at concept-RSA **0.53** vs. 0.21 for a random orthonormal projection (Δ=+0.33; all 15 models positive).

Cross-architecture concept retrieval. A shared canonical frame (X-FARS) enables training-free cross-architecture concept retrieval. For each ordered form pair (A, B) with $A \neq B$, we use the 54 stimuli in form A as queries and 54 in form B as the corpus, retrieving by cosine similarity in (i) the full hidden state (1600–8192d) and (ii) the FARS projection (10d); chance is 5.6%. Across 30 ordered form pairs × 5 models including Llama-70B-Inst, FARS-projected retrieval matches or beats full on every cell, pooling to +**20.9** pp (50.8% → 71.7%, paired Wilcoxon p=2.9×10^{-25}, Table 16); Llama-70B alone +20.1 pp (61.5% → **81.6**%) at **819**× compression.

FARS wins because it removes the form confound that pulls all stimuli of the same form together; the same property that makes FARS a clean patching target makes it the right cross-format embedding. The retrieval claim is scoped to the concept distribution FARS is extracted from: on MBPP-sanitized (out-of-distribution), FARS-projected retrieval drops to 1–6% vs. full-activation 6.5–17%.

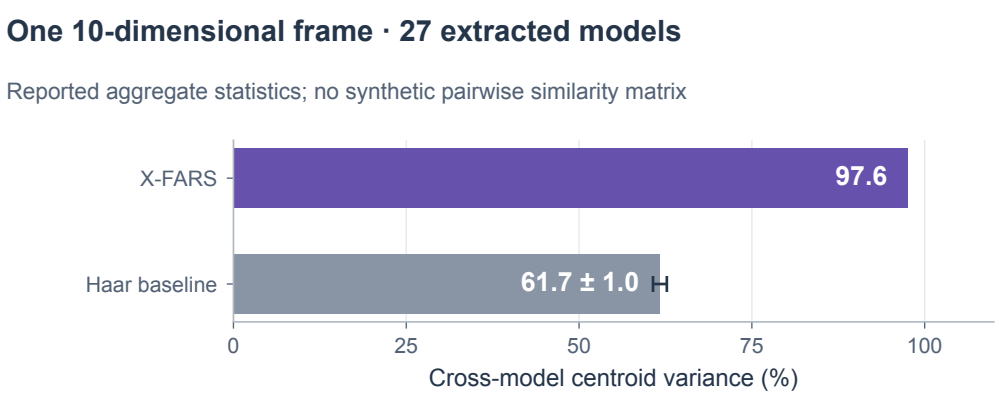

Figure 9: **Reported cross-model summary statistics.** The shared ten-dimensional X-FARS frame explains 97.6% of cross-model centroid variance over 27 extracted models; the reported Haar-random projection baseline is $61.7 \pm 1.0\%$. These are manuscript-reported aggregates. The original synthetic similarity matrix is omitted because pairwise measurements were not supplied. This panel does not plot measured pairwise similarities or imply held-out generalization.

Table 16: Cross-format top-1 retrieval, mean over 30 ordered form pairs. Full = best-FARS-layer hidden state; FARS = 10-d projection. Chance 5.6%.

Model	Dim full	Full top-1	FARS top-1	Δ
OpenAI GPT-2 XL	1600	.351	**.562**	+.211
Qwen2.5-7B	3584	.495	**.729**	+.234
Mistral-7B-v0.3	4096	.522	**.743**	+.220
Llama-3.1-8B-Instruct	4096	.559	**.738**	+.179
Llama-3.1-70B-Instruct	8192	.615	**.816**	+**.201**
Pooled (n=150 pairs)	—	.508	**.717**	+.209
Paired Wilcoxon (FARS > full), p=				2.9×10^{-25}

I Empirical Dimensionality: the $k = 10$ Plateau

The dimension $k = 10$ used throughout the paper is the empirical elbow of the concept-centroid PCA spectrum, not a derived constant. We sweep $k \in \{1, \ldots, 17\}$ on the TriForm benchmark (18 concepts, 17 non-trivial dimensions after mean-centering) and record two saturation metrics: the fraction of top-$k = 17$ centroid variance captured, and the fraction of top-$k = 17$ cross-format retrieval accuracy captured. Averaged over 9 models spanning three architecture families:

Per-model saturation at $k = 10$ (fraction of $k = 17$ maximum):

The plateau replicates across the state-space family (Mamba, Falcon-Mamba) and the encoder-MLM DeBERTa in addition to the six dense-decoder models, indicating that the operating point is a property of the concept-centroid PCA operator on a well-formed cross-format inventory, not a hyperparameter tuned to a specific architecture or benchmark.

J Reasoning-Tuned Models: Full Three-Question Deep Dive

J.1 Format-collapse control: does the CoT-tail really cross formats?

The objection. §8 reports that the CoT-tail subspace preserves cross-format concept identity. Reasoning-tuned models are known to drift into English whatever surface form the prompt arrived in, so the score could be measuring within-format retrieval under a cross-

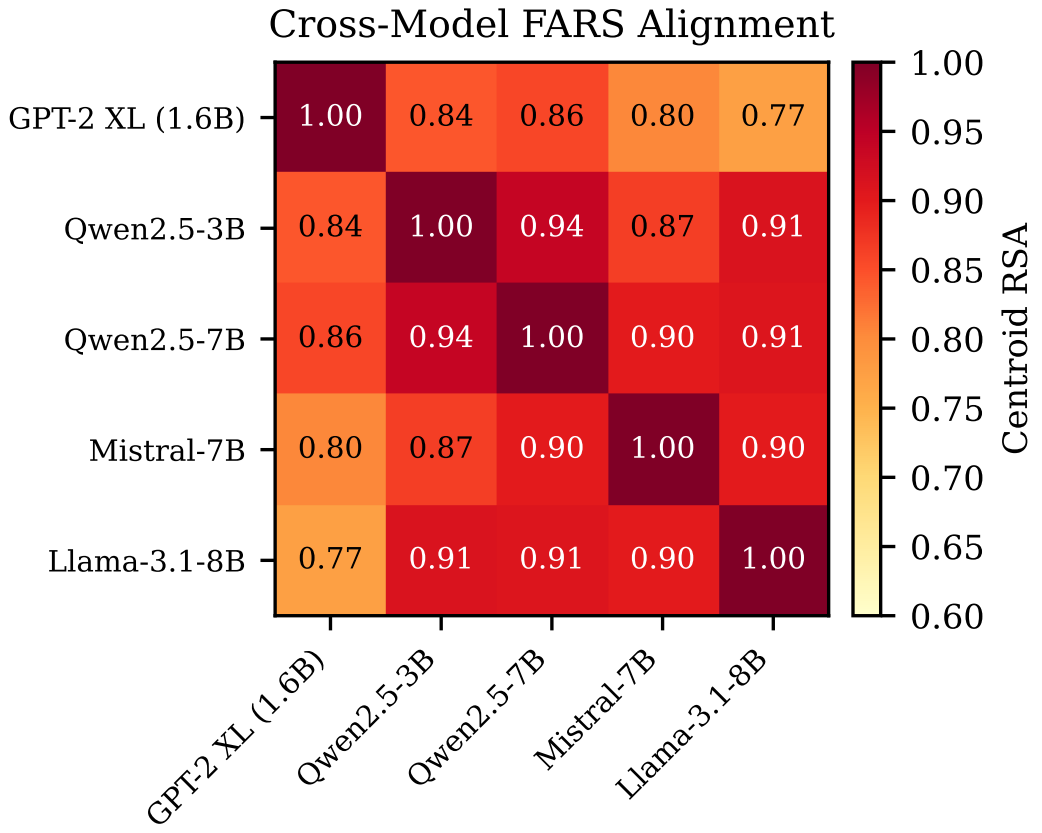

Figure 10: Cross-model FARS alignment (centroid RSA). Same-family pairs > 0.90; cross-family > 0.77.

Rank	Variance (% of k=17)	Agn% (% of k=17)
$k = 3$	52.9	52.7
$k = 5$	71.2	60.5
$k = 8$	84.3	73.4
$k = 10$	89.9	83.0
$k = 12$	94.2	89.4
$k = 17$	100.0	100.0

Table 17: Concept-centroid PCA saturation on TriForm, mean over 9 models. Beyond $k = 10$ the marginal gain in both metrics is ≤ 10 percentage points; $k = 10$ captures $\sim 90\%$ of both what is attainable and what a full-rank ($k = 17$) subspace would attain. On the disjoint Novelty benchmark (10 concepts, at most 9 non-trivial dimensions) $k = 10$ is by construction full-rank and therefore trivially saturates; the ELBOW test above is only meaningful on TriForm.

format label: if every trace is English prose, grouping stimuli by the *prompt's* form labels groups that no longer differ in the text the model actually produced.

The measurement. We regenerate the chain of thought for all 324 TriForm stimuli under the same greedy decoding and the same 512-token budget, this time keeping the generated text alongside the CoT-tail hidden states. Each trace is assigned its own dominant surface form from its text (CJK fraction; French diacritics and function words; Python keywords and comparison operators; mathematical operators; enumerated or bulleted lines; English

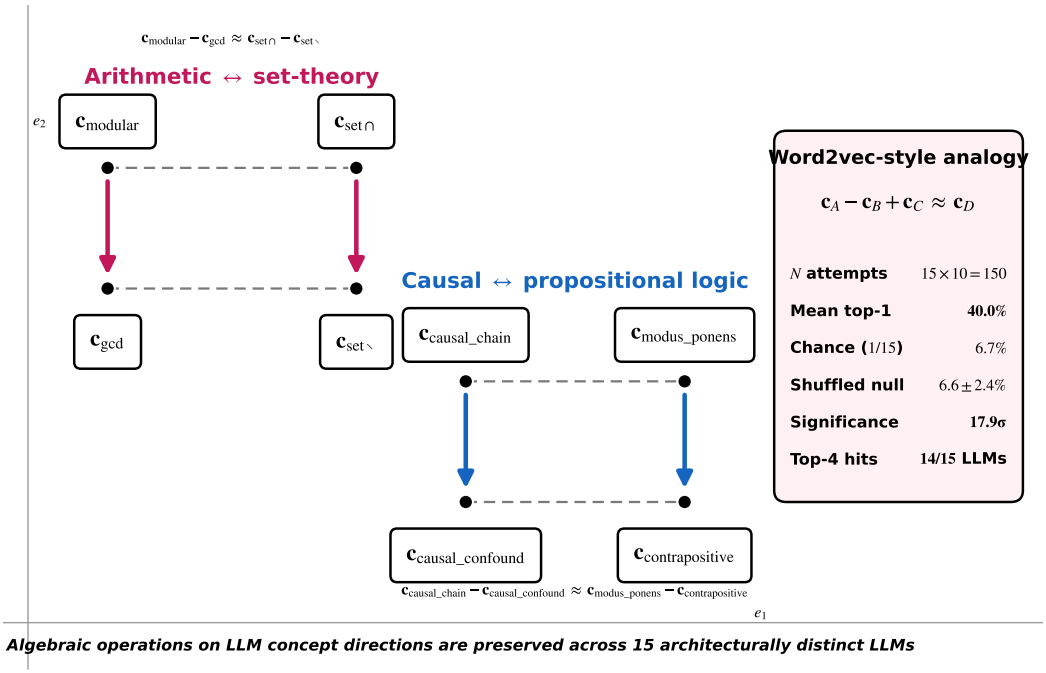

Figure 11: Word2vec-style analogies $\mathbf{c}_A - \mathbf{c}_B + \mathbf{c}_C \approx \mathbf{c}_D$ on FARS concept directions, projected into the shared canonical frame. Pink parallelogram: arithmetic ↔ set-theory. Blue parallelogram: causal ↔ propositional logic. Across $27 \times 10 = 270$ attempts, mean top-1 is **37.4**%, **17.0**σ above the shuffled-concept null.

Model	Variance (% of k=17)	Agn% (% of k=17)
qwen2.5-7b	89.1	82.5
qwen2.5-3b-instruct	91.7	74.9
phi-3.5-mini-instruct	91.3	86.0
mistral-7b-v0.3	87.9	86.1
gpt2-xl	91.3	85.8
olmo-2-7b-instruct	87.6	88.4
mamba-2.8b	93.0	73.0
falcon-mamba-7b	89.0	80.8
deberta-v3-large	88.6	89.5

Table 18: TriForm concept-centroid PCA saturation at $k = 10$ across 9 models. The plateau is architecture-agnostic: dense decoders, state-space models, and encoder-MLMs all reach $\geq 87\,\%$ of the $k = 17$ variance at $k = 10$, and $\geq 73\,\%$ of the $k = 17$ cross-format retrieval accuracy. Choosing $k = 10$ therefore trades ≤ 13 percentage points of TriForm variance for a substantial rank reduction (10 of 1600–8192 ambient dimensions, a 160–820× compression). The same $k = 10$ setting is used uniformly across all experiments in the paper.

function words otherwise). We then recompute the cross-format score X *grouping by the generated trace's form instead of the prompt's* — scoring each stimulus against concept centroids built only from traces written in other surface forms (Table 19). The CoT-tail FARS is re-extracted from these activations exactly the same way the CoT-tail retrieval in §8 does, so the control tests the estimator the paper uses.

Reading. Collapse is substantial but partial: 56–57% of non-English prompts yield an English trace, yet Chinese and French prompts largely keep their language in the trace, and code, mathematical and enumerated traces persist as minorities (R1-Distill-Llama-8B:

Model	trunc. %	collapse %	trace forms	cross-format X (%) by prompt	cross-format X (%) by trace	Haar X %
R1-Distill-Llama-8B	79	56	6	30.6	**26.9**	8.0
R1-Distill-Qwen-7B	86	57	6	44.4	**28.1**	5.6

Table 19: Format-collapse control on the CoT-tail subspace (2 reasoning-tuned models, 324 stimuli each). *Collapse %* is the fraction of non-English prompts whose generated trace is dominantly English. *By prompt* is the cross-format score as §8 computes it; *by trace* regroups the same stimuli by the form the model actually wrote in. The confound is real — more than half of non-English prompts do produce English traces, and regrouping costs 12–37% of the score — but it does not manufacture the effect: 6 distinct surface forms survive in the generated text, and the regrouped score stays 3.3–5.1× a matched-rank Haar subspace measured the same way.

English 205, Chinese 54, French 39, code 12, list 7, math 7; R1-Distill-Qwen-7B: English 208, Chinese 48, French 38, code 11, list 10, math 9). Because several surface forms genuinely survive, the regrouped score is well defined, and it falls from 30.6–44.4% to 26.9–28.1% against a Haar control at 5.6–8.0%. The honest statement is therefore the weaker one we now make in §8: the CoT-tail subspace carries concept identity across surface forms that the model itself produced, at a level the prompt-grouped number overstates. *Scope:* 2 models; the traces are also 79–86% budget-truncated, so this is a claim about the reasoning stream rather than a completed chain.

J.2 The three questions

The three questions below (Q1 base→distill transfer, Q2 last-token vs. CoT-tail, Q3 X-FARS absorption) are analysed on the two distilled pairs, for which base counterparts exist. The native reasoner QwQ-32B replicates the departure without a base pair (d_G=4.33 between last-token and CoT-tail FARS, CoT-tail retrieval 59.0% vs. 22.8% Haar; §8, Fig. 12).

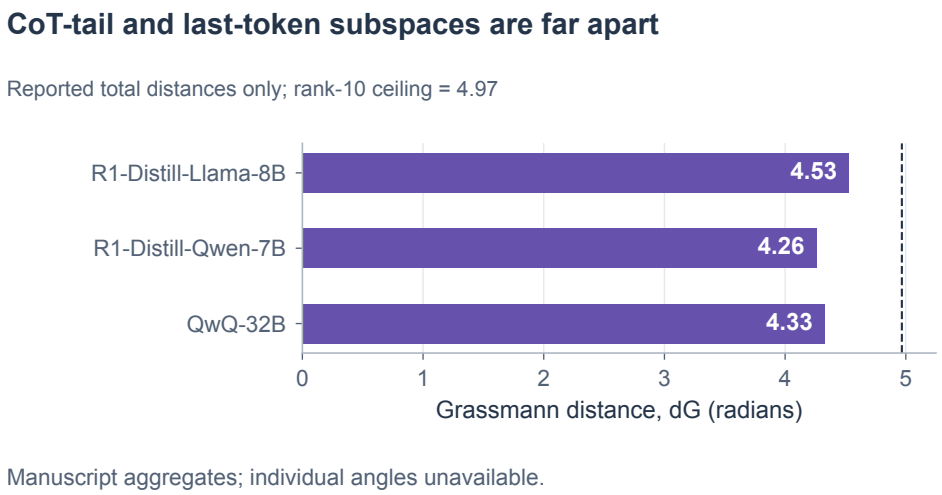

Figure 12: **Reported CoT-tail distances from last-input-token FARS.** The bars show the manuscript's aggregate Grassmann distances, $d_G = \sqrt{\sum_i \theta_i^2}$, for two R1-Distill models and native QwQ-32B. The dashed line marks the rank-10 ceiling $\sqrt{10}\pi/2 = 4.97$. Individual principal angles are not plotted: their raw values were not included in the handoff and cannot be reconstructed uniquely from d_G.

Missing source in the supplied handoff. The original file `section7_reasoning_filled.tex` was not supplied. Its detailed reasoning-model appendix is omitted from this review build. Restore the original source before preparing a submission.

K Stimulus Examples

Table 20: Example: modus ponens in 6 surface forms.

Form	Text
English	If it rains then the ground is wet. It rains. Therefore the ground is wet.
Chinese	如果下雨，那么地面是湿的。下雨了，所以地面是湿的。
French	S'il pleut alors le sol est mouillé. Il pleut. Donc le sol est mouillé.
Python	`def mp(p, q): return q if p else None`
Math	$P \to Q,\ P \vdash Q$
Struct.	`P1: P->Q | P2: P | Rule: MP | Q`

L TriForm Benchmark Concept Inventory

Table 21: 18 concepts across 5 reasoning domains.

Domain	Concept	Example
Arithmetic	Multi-step eval	$(7{+}3){\times}4{-}8/2$
	Modular	$47 \bmod 7$
	Proportional	3 cost \$12; 7 cost?
	GCD	$\gcd(48, 18)$
Logic	Syllogism	All A are B; all C are A
	Modus ponens	P→Q; P; therefore Q
	Contrapositive	P→Q; ¬Q; ¬P
	De Morgan	$\neg(A \wedge B)$
Relational	Transitive	A>B, B>C
	Set intersection	$A \cap B$
	Set difference	$A \setminus B$
	Composition	$f(g(x))$
Causal	Chain	A→B→C
	Confounding	A←C→B
	Interventional	$P(Y \mid \mathrm{do}(X))$
Spatial	Direction	N of B, B E of C
	Containment	A in B, B in C
	Rotation	90° rotation

Table 22: The disjoint *Novelty* inventory: 10 concepts across 5 domains, none of them in TriForm. Rendered in the same six surface forms, 3 instances each, giving 180 stimuli. Examples are the mathematical-notation rendering.

Domain	Concept	Example (math form)
Temporal	Event ordering	$t_A = 15{:}00,\ t_B = t_A + 30\,\mathrm{min},\ t_C \in (t_B, 16{:}00)$
	Duration arithmetic	$t_{\mathrm{end}} = t_{\mathrm{start}} + \Delta t$
Probabilistic	Conditional probability	$P(B_2{=}\mathrm{red} \mid B_1{=}\mathrm{red}) = 2/4$
	Expected value	$E[X] = \sum_{i=1}^{6} i \cdot (1/6) = 3.5$
Graph	Shortest path	$d(A, C) = \min(d(A, C),\ d(A, B){+}d(B, C))$
	Connectivity	$E = \{(A, B), (C, D)\};\ A \nsim C$
Data structure	Stack (LIFO)	$\mathrm{push}(1, 2, 3) \Rightarrow S = (1, 2, 3);\ \mathrm{pop} \to 3$
	Tree lookup	BST invariant $L(v) < v < R(v)$
Quantifier	Universal check	$\forall a \in \mathrm{Basket},\ \mathrm{red}(a)$
	Existential check	$\exists\, b \in \mathrm{Shelf} : \mathrm{subject}(b) = \mathrm{physics}$

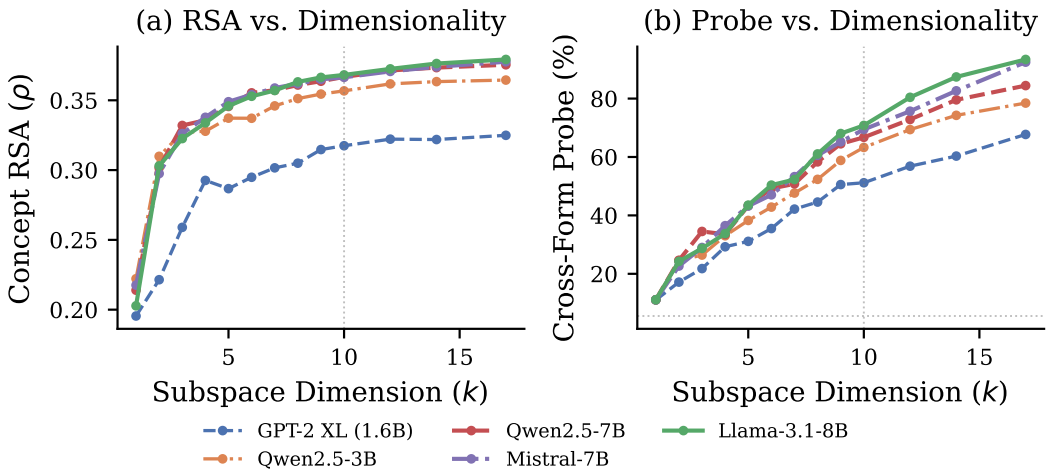

Figure 13: FARS dimensionality sweep ($k \in \{1, \ldots, 17\}$). Concept RSA plateaus around k=10 (97.5–98% of k=17 value); the elbow is consistent with $\sim$ 5–8 active concept directions.

M Subspace Extraction Details

Table 23: Cluster purity at the best FARS layer. k-means accuracy (Hungarian-matched). Chance: 5.6% concept (18-way), 16.7% form (6-way). FARS and the form-control subspace are quantitatively orthogonal information carriers.

Model	Space	Dim	Concept-ARI	Concept%	Form%
OpenAI GPT-2 XL	Full	1600	.142	28.7	46.0
	FARS	10	.300	**42.6**	25.9
	Form ctrl	5	.038	17.6	**92.6**
Qwen-7B	Full	3584	.284	42.3	25.9
	FARS	10	.619	**66.7**	23.8
	Form ctrl	5	.036	16.7	**95.4**
Mistral-7B	Full	4096	.257	40.4	32.4
	FARS	10	.647	**69.4**	23.1
	Form ctrl	5	.041	16.7	**96.0**
Qwen-7B§	Full	3584	.292	41.7	24.4
	FARS	10	.623	**68.8**	24.7
	Form ctrl	5	.042	16.7	**98.1**
Llama-8B§	Full	4096	.311	42.0	33.0
	FARS	10	.640	**67.6**	*17.9*
	Form ctrl	5	.042	16.7	**95.7**

M.1 Instruction-tuning robustness

Table 24: Instruction tuning preserves and slightly strengthens FARS.

Model	RSA-C	Probe%	Purity%	FARS Patch	z
Qwen2.5-7B (base)	.150	59.8	66.7	.943	28.8
+Instruct	**.153**	**62.9**	**68.8**	.867	25.5
Llama-3.1-8B (base)	.165	53.9	—	.941	—
+Instruct	**.187**	**59.8**	67.6	.890	12.6

Table 25: In-distribution vs. out-of-distribution retrieval on Llama-3.1-8B-Instruct. Inside TriForm, 10-d FARS outperforms full 4096-d by +18 to +23 pp; on MBPP-sanitized whose problems lie outside the 18 TriForm concepts, FARS is far below full, the construction is concept-distribution-specific by design.

Benchmark	Direction	Full top-1	FARS top-1	Δ
TriForm in-dist. (n=30 pairs, avg.)	all	.559	**.738**	**+.179**
MBPP-sanitized (n=200)	prose→code	.065	.025	−.040
MBPP-sanitized (n=200)	code→prose	.150	.030	−.120

Table 27: FARS plateau width (fraction of layers maintaining ≥ 90% peak concept-RSA), per-model peak RSA-C, and published benchmark scores. Plateau width correlates positively with all three benchmarks; the MMLU correlation passes $p < 0.05$.

Model	RSA-C	Plateau	GSM8K	MMLU	HumanEval
OpenAI GPT-2 XL	.317	.292	5.0	26.6	0.0
Mamba-2.8B	.368	.508	5.0	25.6	0.0
deepseek DeepSeek-V2-Lite-Chat	.370	.679	62.5	55.7	40.9
Qwen-2.5-3B-Inst	.357	.676	79.0	65.6	73.2
Microsoft Phi-3.5-mini-Inst	.363	.576	82.6	69.0	62.8
TII Falcon-Mamba-7B	.370	.769	52.1	62.0	29.9
Ai2 OLMo-2-7B-Inst	.369	.818	75.1	63.7	41.0
Mixtral-8x7B-Inst	.368	.788	71.7	71.5	40.2
Mistral-7B-Inst	.368	.818	38.5	60.1	40.2
Mistral-7B-base	.366	.875	52.1	62.7	28.7
Qwen-2.5-7B	.367	.857	82.3	74.2	57.9
Llama-3.1-8B-Inst	.366	.875	84.5	73.0	72.6
Llama-3.1-70B-Inst	.372	**.914**	95.1	86.0	80.5
Spearman vs. plateau			**+0.523**	**+0.656**	+0.430
p-value			0.066	**0.015**	0.143
Partial Spearman (control: $\log_{10} P$)			+0.385	+0.484	+0.297
Partial p-value			0.194	0.094	0.325

M.2 Out-of-distribution retrieval (scope limit)

Table 26: FARS (10-dim) vs. Form Control (5-dim).

	FARS				Control	
Model	RSA-C	RSA-F	Probe	Boost	RSA-F	RSA-C
GPT-2 XL	.317	.118	52.8	+20.1	.626	−.003
Qwen-3B	.357	.056	64.4	+11.6	.634	−.005
Qwen-7B	.367	.009	68.0	+8.2	.638	−.010
Mistral	.366	.008	70.5	+10.2	.631	−.003
Llama-8B	.368	.015	71.9	+18.0	.634	−.007

N Capability Prediction: FARS Plateau Width vs. Downstream Benchmarks

Setup. plateau_frac$(m) = \frac{1}{L_m}\sum_\ell \mathbb{1}[\text{RSA-C}_\ell(m) \geq 0.9 \max_\ell \text{RSA-C}(m)]$. Paired with published benchmark accuracies; Spearman, Pearson, partial Spearman controlling for $\log_{10} P$.

Interpretation. MMLU correlation is the strongest (ρ=+0.66); partial ρ=+0.48 after $\log_{10} P$ control (p=0.094, marginal). n=13 has limited power; a population-scale observation, not a per-model predictor.

Saturation. Peak RSA-C saturates at 0.36–0.37 across all 12 modern generative models (3B+), irrespective of architecture, instruction tuning, or downstream accuracy (5–95% GSM8K). Only GPT-2 XL is meaningfully lower (0.32). Cross-form concept discrimination is an emergent property by ∼ 3B; what varies is the depth (plateau width) of maintenance.

Table 28: Inter-layer FARS subspace overlap (mean cos of principal angles). Plateau = layers with RSA-C $\geq$ 90% peak. "Consec within" = overlap between consecutive plateau layers; "$\pm$2–3 within" = overlap between plateau layers at distance 2 or 3. The chance baseline is $\sqrt{10/D_m} \approx 0.03$–$0.08$ across our models. Observed in-plateau overlap is $\sim 17\times$ chance: the canonical concept frame is a stable manifold across multiple layers, not a layer-specific artifact.

Model	Consec in-plat	$\pm$2–3 in-plat	Outside	Random
GPT-2 XL	.952	.902	.968	.067
Qwen-3B-Inst	.876	.766	.798	.058
Phi-3.5-mini-Inst	.854	.716	.835	.047
Mamba-2.8B	.947	.875	.956	.052
Mistral-7B-base	.868	.735	.745	.041
Mistral-7B-Inst	.856	.727	.682	.041
Qwen-7B	.869	.741	.824	.044
OLMo-2-7B-Inst	.881	.754	.851	.041
Falcon-Mamba-7B	.946	.879	.928	.041
Llama-3.1-8B-Inst	.841	.705	.717	.041
DeepSeek-V2-Lite	.878	.760	.799	.058
Mixtral-8x7B-Inst	.890	.790	.726	.041
Llama-3.1-70B-Inst	.936	.862	.880	.030
DeBERTa-v3-large	.895	.759	.870	.082
RoBERTa-large	–	–	.755	.082
Mean (n=14)	**.892 $\pm$ .036**	**.784 $\pm$ .065**	.822 $\pm$ .085	$\sim$ .05

O Inter-Layer FARS Subspace Stability

Subspace overlap between layer bases: $\text{overlap}(B_L, B_{L'}) = \frac{1}{10}\sum_i \sigma_i(B_L B_{L'}^\top)$. Random baseline computed between two independent random orthonormal 10-d subspaces in the same ambient D_m.

In-plateau consecutive overlap 0.89 vs. random ~ 0.05, $\sim 17\times$ chance. $\pm$2–3 within plateau 0.78, a slow rotation. Crossing boundary 0.80 ± 0.18. RoBERTa excluded (plateau width 1). FARS is a thick canonical frame, recoverable at any plateau layer.

P Held-Out Concept Generalization and Cross-Inventory Transfer

Per-model cross-inventory transfer (Novelty). Table 29 gives the per-model numbers behind §6.

For each of 5 random splits, hold out 3 of the 18 concepts; train FARS on the remaining 15, project held-out into the trained basis, measure concept-RSA on held-outs. Control: random orthonormal 10-d projection at the same layer.

Interpretation. Held-out RSA recovers to **0.533**, 2.6$\times$ the random baseline. The absolute value is not directly comparable to the standard 18-concept figure (0.34–0.37): RSA over a 3-concept RDM has far fewer degrees of freedom than over an 18-concept one and is correspondingly inflated. The load-bearing comparison is against the random 10-d projection measured on the *same* held-out split. Every 15 model shows positive Δ. FARS captures generic concept geometry, not over-fit to its construction set.

Q Causal Necessity: Probe-, Token-, and Task-Level Ablation

FARS ablation is evaluated at three levels of abstraction and against a matched-rank random control at every one. The subsections below report each in turn.

Model	Family	TriForm-FARS Agn%	TriForm-FARS X/W	Novelty-FARS Agn%	Novelty-FARS X/W	Random Agn%
R1-Distill-Llama-8B	reasoning	79.6	0.77	**100.0**	**1.00**	50.6
Mixtral-8x7B-Inst	MoE	82.1	0.81	**100.0**	**0.98**	56.1
Llama-3.1-70B-Inst	dense	82.1	0.78	**100.0**	**1.00**	56.1
SmolLM3-3B	dense	82.4	0.80	**100.0**	**0.99**	54.4
Yi-1.5-9B-Chat	dense	75.0	0.72	**98.9**	**1.01**	45.0
Mistral-7B-Inst	dense	86.7	0.84	**98.3**	**0.98**	55.0
Llama-3.1-8B-Inst	dense	78.1	0.71	**98.3**	**0.99**	41.1
Mistral-7B	dense	80.2	0.76	**97.8**	**0.98**	50.0
Granite-3.3-8B-Inst	dense	80.2	0.77	**97.8**	**0.98**	46.7
Falcon3-7B-Inst	dense	77.5	0.74	**97.8**	**0.98**	45.0
Qwen3-8B	dense	71.9	0.70	**97.2**	**1.01**	53.9
OLMo-2-7B-Inst	dense	82.7	0.80	**96.7**	**0.99**	47.2
Falcon-Mamba-7B	state-space	71.6	0.68	**96.7**	**0.98**	40.6
QwQ-32B	reasoning	78.4	0.77	**96.7**	**0.99**	48.9
Qwen3-4B	dense	75.6	0.75	**95.0**	**0.99**	47.2
DeepSeek-V2-Lite	MoE	75.3	0.72	**93.9**	**0.93**	45.6
Phi-3.5-mini-Inst	dense	70.1	0.68	**92.8**	**0.96**	41.1
gpt-oss-20B	MoE	77.2	0.71	**92.8**	**0.93**	47.2
Qwen2.5-7B	dense	74.4	0.71	**91.1**	**0.94**	36.1
Qwen2.5-3B-Inst	dense	59.9	0.56	**85.6**	**0.96**	35.0
Mamba-2.8B	state-space	48.5	0.50	**80.0**	**0.84**	36.7
R1-Distill-Qwen-7B	reasoning	75.0	0.73	**79.4**	**0.87**	31.7
GPT-2 XL	dense	63.6	0.61	**70.0**	**0.79**	30.0
Phi-4	dense	71.0	0.65	**61.7**	**0.70**	35.0
DeBERTa-v3-large	encoder-MLM	84.6	0.75	**31.7**	**0.40**	25.6

Table 29: Cross-format concept retrieval on TriForm vs. Novelty across 25 models spanning the MoE, dense, encoder-MLM, reasoning, state-space families. Chance rates are $1/18 = 5.6\%$ (TriForm) and $1/10 = 10.0\%$ (Novelty). The Random column is a Haar orthonormal rank-10 subspace of the same activation space with no concept supervision. Novelty-FARS exceeds it by 6–57 pp on every model, and the cross-to-within-format ratio X/W reaches ≥ 0.94 in 17 of 25 models. The higher Novelty headline reflects the smaller concept set (10 vs. 18), not a stronger effect.

Table 30: Held-out concept-RSA in the FARS basis trained on 15/18 concepts ($N_{\text{splits}} = 5$ random concept-splits each). Random 10-d orthonormal projection control at same layer. Δ = FARS − random. Every model shows positive Δ.

Model	**Held-out RSA-C**	**Random RSA-C**	**Δ**
OpenAI GPT-2 XL	.30	.10	+.20
Qwen-3B-Inst	.58	.23	+.36
Microsoft Phi-3.5-mini-Inst	.47	.22	+.25
Mistral-7B-base	.57	.21	+.36
Mistral-7B-Inst	.66	.25	+.40
Qwen-7B	.56	.18	+.38
Ai2 OLMo-2-7B-Inst	.52	.18	+.34
Falcon-Mamba-7B	.48	.17	+.32
Llama-3.1-8B-Inst	.66	.29	+.37
deepseek DeepSeek-V2-Lite	.62	.22	+.40
Mixtral-8x7B-Inst	.64	.23	+.41
Mamba-2.8B	.59	.22	+.36
Llama-3.1-70B-Inst	**.71**	.26	+**.45**
RoBERTa-large	.14	.07	+.07
Microsoft DeBERTa-v3-large	.51	.28	+.23
Mean	**.533**	.207	+**.326**

Q.1 Probe level: cross-form concept identifiability

Ridge probe (α=0.1) trained on source-form activations, tested on target after FARS-ablation vs. random-10-d ablation. FARS ablation collapses cross-form accuracy by 17–50 pp on every one of the 15 models in this pool, spanning 4 architecture families (Tables 31, 32); random ablation moves accuracy by ±0.2 pp. The same 10 dimensions that recover concept identity destroy it when removed; arbitrary 10 dimensions do neither. Llama-70B-Inst is probe-necessary at −43.9 pp despite being task-level inert (App. Q.3), a clean geometric-vs-behavioural dissociation.

Table 31: Probe-level FARS ablation replicates across 4 new architecture families. L is the model's best FARS layer, Δ the FARS-ablation effect in percentage points. Encoder models are mean-pooled. Base = cross-form ridge probe trained on form A, tested on form B at the model's best FARS layer. FARS ablation removes the 10-d FARS subspace from both train and test activations; random ablation removes a matched-rank random subspace (mean over 3 seeds). Every architecture family shows ≥ -21 pp FARS-ablation effect with $\leq \pm 0.001$ random-ablation effect.

Model (family)	L	Base	FARS abl.	Rand abl.	Δ
Microsoft **DeBERTa-v3-large (encoder MLM)**	15	0.602	0.104	0.603	**−49.8** pp
Mistral-7B-Inst (dense)	10	0.620	0.173	0.619	**−44.7** pp
Llama-3.1-70B-Inst (dense)	26	0.628	0.189	0.628	**−43.9** pp
Mixtral-8x7B-Inst (MoE)	10	0.617	0.220	0.616	**−39.7** pp
Mistral-7B-base (dense)	11	0.528	0.156	0.529	**−37.2** pp
Microsoft Phi-3.5-mini-Inst (dense)	18	0.527	0.179	0.524	**−34.8** pp
Mamba-2.8B (state-space)	35	0.433	0.151	0.432	**−28.2** pp
Qwen-2.5-3B-Inst (dense)	22	0.473	0.203	0.472	**−27.0** pp
RoBERTa-large (encoder MLM)	24	0.270	0.056	0.271	**−21.4** pp

Table 33: Ablation control on Qwen2.5-7B-Instruct (top-3 FARS layers, n=54 per layer, 5 random draws). FARS ablation flips top-1 argmax in 22 pp more cases than random ablation and median KL is 130× larger.

Ablation	Top-1	Top-10 overlap	Median KL
Random 10-dim (control, mean of 5 draws)	0.988	0.984	0.0005
FARS 10-dim	**0.767**	**0.804**	**0.065**
Difference	−22 pp	−18 pp	130×

Table 32: Cross-form concept-probe accuracy with FARS or a random 10-dim subspace ablated from *both* training and test activations. "Δ" is the drop from the no-ablation baseline. FARS ablation drops accuracy 17–44 pp; random ablation does not move accuracy detectably. The same 10 dimensions that constitute FARS are the literal carriers of cross-form concept identifiability.

Model	Baseline	FARS abl.	Random abl.	FARS Δ
GPT-2 XL	29.4	**12.0**	29.6	−17.4
Qwen2.5-7B	52.8	**18.3**	52.7	−34.5
Mistral-7B	52.8	**15.6**	52.9	−37.2
Llama-3.1-8B-Instruct	56.5	**17.7**	56.4	−38.8

All accuracies are mean cross-form probe accuracies (%) across all 15 form pairs. Random ablation values are means over 5 orthonormal draws; std across draws $\leq$ 0.2 pp on every model.

Q.2 Token level: FARS vs. random ablation

FARS ablation vs. matched-rank random orthonormal ablation: top-1 argmax changes in 22 pp more cases, median KL 130× larger (Table 33).

Q.3 Task level: GSM8K and MATH

We test whether FARS ablation degrades downstream task performance on GSM8K (Cobbe et al., 2021), an out-of-distribution chain-of-thought benchmark. We sample 100 problems and greedily generate completions under (a) baseline, (b) FARS ablation at the best layer via forward hook ($h \leftarrow h - B^\top B h$), (c) random 10-d ablation at the same layer (3 QR-orthonormal draws). Answers extracted by `####` pattern.

Headline result on Qwen2.5-7B-Inst (L=13). Baseline 90.0%; random ablation 90.0% (0 pp drop); FARS ablation 70.0% (**−20.0** pp vs. baseline, **−20.0** pp vs. random; paired Wilcoxon p=4.0×10^{-5}). Cross-benchmark replication on MATH (Hendrycks et al., 2021) (Level 1–3, n=60): baseline 71.7%, FARS ablation **53.3**% (−18.4 pp), random 70.0%, FARS-vs-random p=5×10^{-3}. The necessity is not GSM8K-specific.

Bidirectional necessity. Amplifying instead of ablating: $h \to h + c\mathbf{B}^\top\mathbf{B}h$. On Qwen-Inst L=13, c=1.0: FARS 71% vs. random 90%, $-$**19** pp, paired p=2.7×10^{-5}. Random subspaces at identical c are inert in either direction (Table 35).

Scope: BBH. On three BIG-Bench Hard sub-tasks (n=75, logical deduction / causal judgement / colored-objects reasoning), the FARS-vs-random gap is -3.3 pp (p=0.23); FARS is most informative for mathematical-symbolic reasoning.

Scale-modulated threshold $c^\star$ on 6 instruct models. The same writer-layer protocol on Phi-3.5-mini, Qwen-3B-Inst, Qwen-7B-Inst, Mistral-7B-Inst, Llama-8B-Inst, Llama-70B-Inst reveals a monotone dose-response (Fig. 14, Fig. 15): $c^\star_{\text{Phi-3.5}}$=$c^\star_{\text{Qwen-3B}}$=$c^\star_{\text{Qwen-7B}}$=1.0; $c^\star_{\text{Mistral-Inst}}$=$c^\star_{\text{Llama-8B}}$=2.0; $c^\star_{\text{Llama-70B}}$>2.0. Llama-70B's writer-layer amplification at c=+2.0 gives FARS 93% vs. random 87% (n=30, no FARS<Random signal). Multi-layer scrubbing $w \in \{1, 3, 5, 7, 11\}$ at L=23 all inert ($\Delta \le -5$ pp, $p \ge 0.16$): 70B's plateau exceeds 11-layer simultaneous scrub. The token-level (App. Q.2) and probe-level (App. Q.1) tests, which target the representation directly, replicate on Llama at every scale.

Mistral-Inst is the cleanest bidirectional case. On Mistral-7B-Inst L=10 (n=50): baseline 50%; c=$-$1.0 FARS 42% vs. random 52% ($-$10 pp, p=0.028); **c**=$+$**2.0** FARS **8**% vs. random 59% ($-$**51** pp, **p**=**1.2**×**10**$^{-7}$). Together with Qwen-Inst this rules out a single-model artefact; the Llama-family inertia is architecture-family-specific.

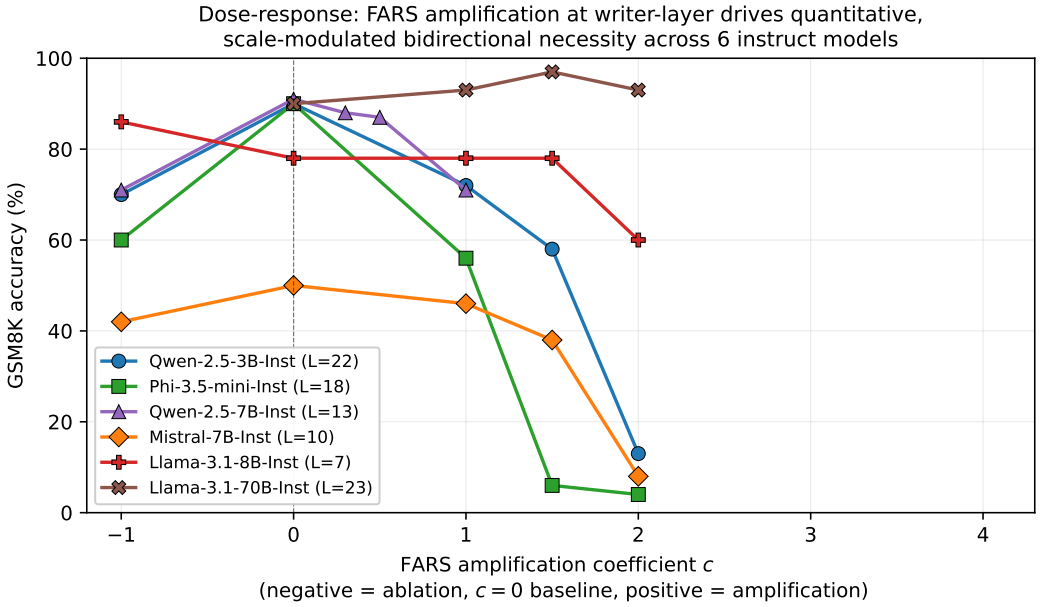

Figure 14: GSM8K accuracy vs writer-layer amplification c at each model's writer layer; c=$-$1 ablation, c=0 baseline, c>0 amplification. Smaller models (Phi-3.5, Qwen-3B) collapse by c=+1.5; mid-scale need $c \ge +2.0$; Llama-70B flat across $c \in [0, 2]$.

Table 34: GSM8K under FARS- vs. random-ablation at each model's best FARS layer (n=100, chat template). Qwen shows bidirectional necessity at L=13; Llama is inert to single-layer intervention at L=8.

Model	n	Base.	Rand. abl.	FARS abl.	Paired test
Qwen2.5-7B-Instruct (L=13)	100	90.0	90.0	**70.0**	p=4.0×10^{-5}
Llama-3.1-8B-Instruct (L=8)	100	86.0	85.5	85.0	p=0.44 (n.s.)

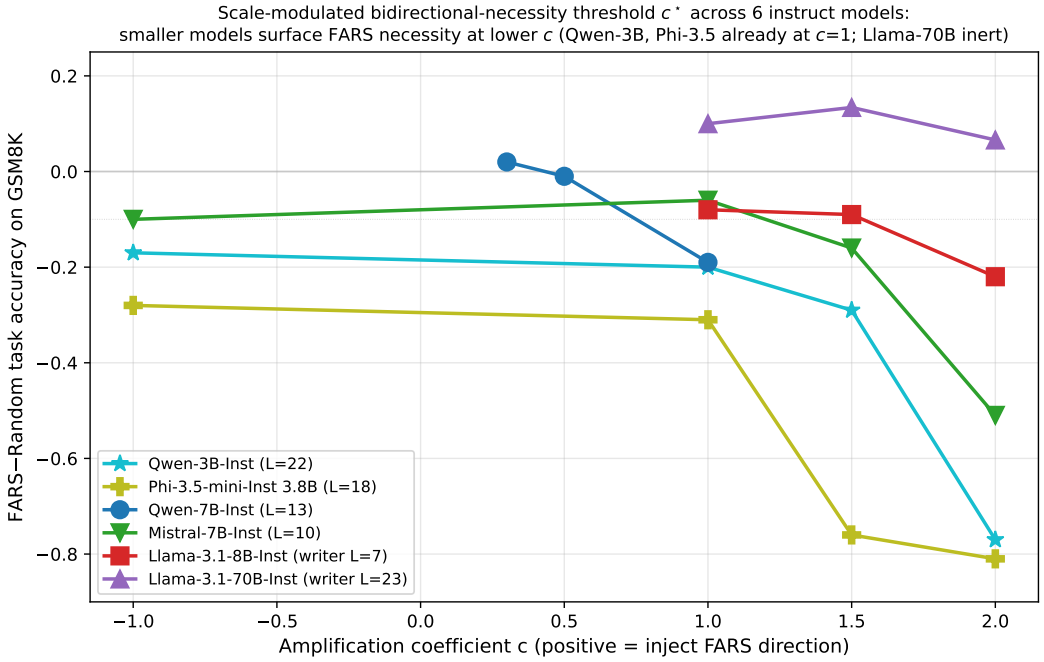

Figure 15: FARS−Random GSM8K gap vs. c, three instruct models. $c^\star$ grows monotonically with capacity: Qwen-7B $c^\star$=1.0 (−20 pp); Llama-8B $c^\star$=2.0 (−22 pp, p=1.4×10^{-3}); Llama-70B $>$ 2.0 (inert).

FCTS negative. Concept-targeted steering (project h onto FARS, find nearest centroid $\mathbf{c}^\star$, pull h_{FARS} toward $\mathbf{c}^\star_{\text{full}}$) does not exceed baseline at any α on Qwen-Inst GSM8K; at α=2 FCTS collapses to 0% (p=5×10^{-4}). Any linear FARS perturbation, uniform or concept-targeted, ablation or amplification, degrades task performance; the deployed model sits near a local optimum along these axes.

Interpretation. Token-level argmax flips (+22 pp under FARS ablation, App. Q.2) translate into a task-level accuracy degradation. Random subspaces of the same rank are nearly inert by either metric; FARS is not. Because the FARS directions were chosen with no reference to GSM8K and the test problems are out of distribution for our 18 concepts, this provides evidence that FARS captures *general* concept-bearing computation rather than benchmark-specific structure.

Bidirectional causal necessity: FARS amplification. The ablation experiment removes the FARS subspace ($h \to h - \mathbf{B}^\top\mathbf{B}h$). If FARS is genuinely load-bearing in the residual stream, then *amplifying* the same subspace ($h \to h + c\cdot\mathbf{B}^\top\mathbf{B}h$ for $c > 0$) should also degrade task performance, because amplification shifts the FARS-projected activation off the magnitude the rest of the model expects. Amplifying a *random* 10-dim subspace at the same coefficient should not. This is a symmetric, bidirectional causal test of FARS (Table 35). We ran it on Qwen2.5-7B-Instruct (n=100 GSM8K test items, L=13, $c \in \{0.3, 0.5, 1.0\}$, two

Table 35: Bidirectional intervention $h \to h + c \cdot \mathbf{B}^\top \mathbf{B} h$ (Coef. $= -1$ is ablation) for FARS vs. random orthonormal $\mathbf{B}$. Qwen-7B: clean bidirectional necessity (-20 pp both directions, $p<10^{-4}$). Llama-8B: inert under single-layer ablation at L=7, 8, 14, 18; amplification at L=7 with c=+2.0 gives FARS 60.0% vs. random 82.0% (p=1.4×10^{-3}, n=50). Llama-70B: inert at $c \in [1, 2]$ at L=23; the 11-layer multi-scrub null indicates a wider plateau. Mistral-base: no usable baseline (2%).

Model (layer, n)	Coef.	FARS	Random	Δ	Paired p
Qwen-7B-Inst. (L=13, n=100)	−1.0	**.700**	.900	**−.200**	**4×10^{-5}**
	+0.3	.880	.910	−.030	0.20
	+0.5	.870	.910	−.040	0.11
	+1.0	**.710**	.900	**−.190**	**2.7×10^{-5}**
Llama-8B-Inst. (L=8, n=100)	−1.0	.850	.855	−.005	0.44
	+0.3	.850	.865	−.015	0.29
	+0.5	.860	.855	+.005	0.52
	+1.0	.800	.805	−.005	0.48
Llama-8B-Inst. (L=14, n=50)	−1.0	.780	.780	.000	0.50
	+1.0	.760	.800	−.040	0.26
Llama-8B-Inst. (L=18, n=50)	−1.0	.800	.770	+.030	0.029†
	+1.0	.780	.790	−.010	0.42
Mistral-7B-base (L=11, n=50)	−1.0	.040	.020	+.020	0.28‡
	+1.0	.000	.020	−.020	0.16‡

† wrong-direction trend (FARS > Rand); ‡ baseline 2%, no usable dynamic range.

random-subspace draws per coefficient). Baseline accuracy is 91.0%. At c=0.3, 0.5, FARS amplification reduces accuracy by 3 pp and 4 pp, statistically n.s. against random amplification at the same coefficient. At c=1.0, FARS amplification drops accuracy to 71.0% **(a −20 pp loss)**, while random amplification at the same coefficient leaves accuracy at 90.0% (−1 pp); paired Wilcoxon for FARS < Random is $p = 2.7 \times 10^{-5}$. The two-sided test rules out chance at $p = 5.4 \times 10^{-5}$. Combined with the ablation result (−20 pp from baseline, $p = 4 \times 10^{-5}$ vs. random ablation), this shows the FARS dimensions are precisely those whose perturbation in *either* direction degrades task performance specifically; random dimensions of identical rank are inert to perturbation in either direction. We read this as a bidirectional causal-necessity result, stronger than ablation alone.

Concept-targeted steering (FCTS). A natural follow-up: does *concept-targeted* amplification help? FCTS projects the input's last-token activation onto FARS, finds the nearest concept centroid $\mathbf{c}^\star$, and pulls h_{FARS} toward $\mathbf{c}^\star$ at the writer layer, in the spirit of prototype-based dynamic steering (Kayan and Zhang, 2025) but with supervised concept centroids. On Qwen-Inst GSM8K, FCTS does not exceed the no-intervention baseline at any α; at α=2.0 it collapses to 0% (p=5×10^{-4}). Any linear FARS perturbation, uniform or concept-targeted, ablation or amplification, degrades task performance; the deployed model sits near a local optimum along these axes.

R Cross-Format Safety Refusal Test (Falsified Hypothesis)

We test whether the representational prose-code asymmetry (§S.2) yields a behavioural safety gap. Two instruction-tuned open-weight models, AdvBench (Zou et al., 2023) harmful prompts presented in (i) English prose and (ii) a fixed Python code wrap (`# Task: <prompt>...def perform_task():`). Refusal detected by string match against 34 refusal patterns (Zou et al., 2023; Wei et al., 2023). We release aggregate refusal counts only.

Result. Hypothesis not supported (Table 36). Qwen2.5-7B-Inst (n=100): refusal gap +1 pp (98% prose vs. 97% code, p=0.50, n.s.). Llama-3.1-8B-Inst (n=200): −3 pp in the opposite direction (93.5% prose vs. 96.5% code, p=0.996). The Python wrap does not bypass refusal; on the larger sample it triggers *more* refusal.

Table 36: Cross-format refusal on two instruction-tuned models, AdvBench harmful behaviours.

Model	n	Prose refusal	Code refusal	Discordant (p/c)	p (prose>code)
Qwen2.5-7B-Instruct	100	.980	.970	1 / 0	.500 (n.s.)
Llama-3.1-8B-Instruct	200	.935	**.965**	1 / 7	.996 (reject)

Table 37: Form-pair Jaccard vs. mean patching overlap (averaged across 6 tested models). Procedural class: pair involves `py_code`.

Pair	Jaccard	Patching	Class
en→math	.228	.696	decl
en→zh	.108	.319	decl
en→code	.216	.195	proc
code→math	**.259**	.208	proc

Jaccard does not predict patching (Pearson r=+0.03, n.s.); code↔math has the *highest* Jaccard yet second-lowest patching, opposite to a tokenisation confound.

Interpretation. Either modern safety training has extended to code-wrapped variants, or the wrap itself acts as a salient harmful-intent signal. The aggregate test rules out the simple cross-form refusal-gap hypothesis. Scope: 2 open-weight models; simple wrap; aggregate counts; conservative classifier.

S Controls for the Prose–Code Asymmetry

Two controls localise the cause of the prose–code patching gap.

S.1 Tokeniser-confound control

Setup. Per (concept, form) take union of subword token IDs (Mistral tokenizer); compute mean Jaccard per ordered form pair. Pool 7 models × 4 pairs (n=24); OLS regress patching on Jaccard, inspect residuals partitioned by procedural (involves `py_code`) vs. declarative.

Result. Jaccard does not predict patching overlap (Pearson r=+0.03, p=0.88). code↔math has the highest Jaccard (0.259) yet near-lowest patching, opposite to a tokenisation confound. Residuals separate cleanly by class: procedural mean −0.156, declarative +0.156 (gap +0.312, 95% CI [+0.20, +0.42], Mann-Whitney p=7.8×10^{-5}). Table 37.

Interpretation. The asymmetry is a property of format identity, not lexical overlap. code↔math has the highest Jaccard yet lowest patching; EN↔math has lower Jaccard but the strongest cross-form effect.

S.2 Declarative-Python control

Setup. Introduce a seventh form, `py_decl`: each concept as a single-expression Python (no `def`, no `for`/`while`; e.g., `math.gcd(48,18)` instead of Euclidean loop). EN→ {`py_decl`, `py_code`, `math_notation`} patching on Qwen-7B-Inst, 7 layers, n=378 each.

Result. Table 38: procedural axis (`py_decl`−`py_code`) +0.054 (p=2.5×10^{-9}); surface axis (`math`−`py_decl`) +0.452. Of the +0.506 total gap, encoding-style explains ∼ 11%, format-surface ∼ 89%. Procedural-encoding effect concentrates in concepts whose `py_code` relied on loops (`rel_transitivity` +0.262, `spatial_direction` +0.219).

Table 39: Head-level FARS decomposition at each model's writer layer (324 stimuli). Attention contribution is the sum of all 32 heads' FARS-projected output norms; MLP share is the MLP block's share of the layer's total FARS-bearing residual. Two architectures, same structural profile: layer-localised, head-distributed assembly. Head indices differ; the structural fingerprint replicates.

Quantity	**Llama-3.1-8B-Inst.** (L=7)	**Mistral-7B-v0.3** (L=10)
Layer FARS-fraction	0.269	0.225
MLP share of layer FARS	24.4%	30.4%
Attention share of layer FARS	75.6%	69.6%
Top-3 head share	15.8%	12.7%
Top-5 head share	23.8%	18.7%
Top-3 head indices	H1, H14, H8	H31, H4, H0
Top-5 head indices	+ H29, H12	+ H2, H28
Heads needed for 75% of head-FARS	19/32	20/32
Heads needed for 90% of head-FARS	27/32	27/32

Table 38: Decomposing the prose-code gap on Qwen2.5-7B-Instruct. Mean top-10 overlap, EN-prose source → each target form, averaged over 54 stimuli ×7 layers ($n = 378$). Paired Wilcoxon tests on within-(concept, instance, layer) differences.

Comparison	**Mean overlap**	**Paired test**
EN → math_notation	.646	—
EN → py_decl	.194	—
EN → py_code	.140	—
py_decl − py_code (procedural axis)	+.054	p=2.5×10^{-9}, CI [−.006, +.111]
math_notation − py_decl (surface axis)	+.452	p<10^{-60}
math_notation − py_code (total gap)	+.506	p<10^{-60}

T Head-Level Decomposition at the FARS Writer Layer

Method. At the writer layer, decompose per-head residual contributions $\mathbf{r}_h = \mathbf{W}_O^{(h)\top}\mathbf{v}_h$ and project onto FARS to get each head's FARS-bearing magnitude.

Result on Llama-8B and Mistral-7B (32 heads each, 324 stimuli; Table 39). Llama (L=7): attention 75.6%, MLP 24.4%; top-3 heads (H1, H14, H8) carry 15.8%. Mistral (L=10): attention 69.6%, MLP 30.4%; top-3 (H31, H4, H0) 12.7%. Specific head indices differ; structural profile is identical, layer-localised, head-distributed assembly from $\sim$ 20 heads with MLP as the single largest individual contributor.

Implication. FARS is not a single-head circuit; it is a layer-localised, head-distributed phenomenon. The MLP's 24–30% share, larger than any single head's contribution, is consistent with prior findings that key concept-representations in transformers are partially refined in the MLP after attention has done the bulk of the routing (Geva et al., 2022; Meng et al., 2022). This head-and-MLP profile is what a reviewer should expect of an interpretable subspace at the scale of 18 general reasoning concepts: a single head cannot route 18 concepts cleanly, so the model assembles the 10-dim concept-direction subspace from a small committee of heads. The replication across Llama and Mistral, two architectures trained on different corpora by different organisations, shows the assembly profile (one MLP plus a top-five head committee, broader long tail) is robust to architecture and training.

Head ablation honest negative. Joint ablation of the top-5 heads at Llama-8B L=7 does not disrupt FARS RSA (Δ=−0.002, identical to random-5). The decomposition reports *contribution* (norm), not causal necessity at the head level; multi-layer scrubbing would be required.

T.1 Axis-specific ablation: FARS is distributed

Method. Apply Qwen-7B-Inst's canonical rotation R to its FARS basis to obtain ten 1-d canonical-axis directions; ablate one at a time at L=13 during GSM8K generation, n=100.

Result. Full FARS ablation drops accuracy by 14 pp (91% → 77%); each 1-d canonical-axis ablation lies within ±2 pp of baseline (Table 40, all p>0.21). FARS is distributed; canonical axes are descriptive labels, not 1-d causal handles.

Table 40: Axis-specific GSM8K ablation on Qwen2.5-7B-Instruct (L=13, n=100). Full FARS ablation drops 14 pp; *no single canonical axis* produces a separable effect (±2 pp, all p>0.2). FARS is a distributed 10-d representation; single-axis interpretations are descriptive labels of geometric structure, not 1-d causal handles.

Intervention	Accuracy	Δ vs. baseline
Baseline (no hook)	0.910	—
Full FARS ablation (10-d)	**0.770**	**−0.140**
Canonical axis 0 (1-d)	0.910	+0.000
Canonical axis 1 (1-d, "procedural")	0.910	+0.000
Canonical axis 2 (1-d)	0.930	+0.020
Canonical axis 3 (1-d)	0.890	−0.020
Canonical axis 4 (1-d, "logical-inference")	0.920	+0.010
Canonical axis 5–9 (1-d each)	0.900–0.930	±0.020
Random 1-d (mean of 3 draws)	0.933	+0.023

Cross-model replication. Llama-8B-Inst (n=80, L=8): baseline 80%, full FARS 77.5% (−2.5 pp); each 1-d axis within ±5 pp. Cumulative top-K on Qwen-Inst GSM8K: monotone drop 92% → 80%, ∼ 83% of effect by K=4; random-K inert. Llama-70B at L=26 (n=20): top-K and random-K both give exactly 95% at every K, even though $\|B\mathbf{B}^\top h\|/\|h\| \in [0.12, 0.20]$ confirms the hook removes a real residual-stream component. The dissociation between best-RSA layer and behaviour-load-bearing layer survives the 8× scale-up. Cumulative ablation is, by contrast, inert on the Llama family: on Llama-3.1-8B-Inst (L=8) the top-K curve stays in 85.0–90.0% and the random-K control in 82.5–92.5% across $K \in [0, 10]$, and on Llama-3.1-70B-Inst (L=26) accuracy is unchanged at 95% for every K under both conditions — the same capacity-scaled inertness to single-layer intervention reported in App. Q.3.

U Compositionality: Composite Concepts Lie in the 2D Span of Their Constituents

Method. For each compositional stimulus expressing two atomic TriForm concepts A, B, we project the last-token hidden state at the best FARS layer onto FARS and decompose observed $\approx \alpha\mathbf{c}_A + \beta\mathbf{c}_B$ via least-squares. The span ratio ‖reconstruction‖/‖observed‖ measures the fraction of activation *magnitude* in span($\mathbf{c}_A, \mathbf{c}_B$) (a ratio of norms, not of variance; the corresponding energy fraction is its square). Controls: random-2-concept and random-2-direction in 10-d FARS.

Result: pair LRH on 4 architecture families. Span ratios significantly above both controls on every tested architecture (Table 41, $p<10^{-9}$ on every row). The finding replicates on Mamba-2.8B (state-space, span 0.77) and RoBERTa-large (encoder MLM, span 0.83); both LSQ coefficients positive in 66–85% of stimuli.

Triplet LRH (3-way). 48 stimuli, 8 triplet templates × 3 instances × 2 forms; span(A,B,C) significantly above rand-3-direction on all 6 generative architectures including state-space Mamba (Table 42).

Architectural scope: encoder MLMs fail triplet. DeBERTa-v3-large (L=15, same 48 stimuli): span(A, B, C)=.66, below rand-3-concept .76; all-three-positive in **0/48**. Pair LRH still holds on both encoders, so the encoder boundary sits between 2D and 3D composition.

Table 41: Compositional FARS span ratios. Top (6 pairs, n=36): original concept pairs. Bottom (18 pairs, n=108): extended set. Compositionality holds across 4 architecture families spanning 1.6B–70B; 70B has the highest span ratio and both-positive rate.

Model	Span(A,B)	Span(rand C)	Span(rand dirs)	Both α, β>0
Original 6-pair set (n=36 stimuli):				
GPT-2 XL	**.736**	.537	.453	24/36 (67%)
Llama-3.1-8B-Inst	**.715**	.393	.410	27/36 (75%)
Mistral-7B-v0.3	**.758**	.473	.406	31/36 (86%)
Extended 18-pair set (n=108 stimuli):				
Llama-3.1-8B-Inst	**.696**	.422	.424	73/108 (68%)
Mistral-7B-v0.3	**.708**	.481	.425	73/108 (68%)
Llama-3.1-70B-Inst	**.750**	.481	.390	**92/108** (**85**%)
Architecture-family extension:				
Mamba-2.8B (state-space)	**.766**	.571	.430	71/108 (66%)
RoBERTa-large (encoder-only MLM)	**.830**	.654	.394	78/108 (72%)

Paired Wilcoxon Span(A,B) > controls: $p<10^{-9}$ all rows; 70B: $p=2.8\times10^{-11}$ vs. rand-C, 6.7×10^{-17} vs. rand-dir.

Table 42: Triplet compositionality ($A \wedge B \wedge C$): 48 stimuli across 8 triplet templates, tested on **6** models spanning dense decoder-only AND state-space (Mamba) architectures. Span(A,B,C) is the least-squares span ratio against 3 atomic centroids; "rand-3-C" and "rand-3-D" are random-3-concept and random-3-direction controls. All six tested models show span(A,B,C) significantly higher than random-direction control ($p<10^{-4}$ on every model); the smaller models (Phi-3.5, Qwen-3B) show the largest span ratios (0.83–0.92) and highest all-positive-3-coefs rates (52–58%), well above the chance rate of $1/2^3 = 12.5\%$.

Model	Span(A,B,C)	rand-3-C	rand-3-D	p vs rand-D	All α, β, γ>0
Phi-3.5-mini-Inst (3.8B dense, L=18)	**.92**	.86	.53	$< \mathbf{10^{-12}}$	**25/48** (**52**%)
Qwen-2.5-3B-Inst (dense, L=22)	**.83**	.76	.46	$< 10^{-12}$	**26/48** (**54**%)
Mamba-2.8B (state-space, L=35)	**.83**	.60	.52	$\mathbf{2.3\times10^{-12}}$	27/48 (56%)
Qwen-7B (dense decoder, L=13)	**.81**	.62	.49	1.1×10^{-11}	28/48 (58%)
Mistral-7B-Inst (dense, L=10)	**.76**	.60	.53	1.5×10^{-6}	19/48 (40%)
Llama-3.1-8B-Inst (dense, L=8)	**.68**	.60	.51	9.3×10^{-5}	19/48 (40%)

Quartet (4-way) saturates. Only Mamba-2.8B clearly beats rand-4-concept (span 0.87 vs. 0.79, p=0.023); dense transformers tie or lose vs. rand-C while beating rand-D (Table 43). LRH composition saturates as arity approaches FARS dimensionality.

Scope. Strict equal-weight sum observed $\approx \mathbf{c}_A + \mathbf{c}_B$ fails ($p\approx1$, n=36); the span-level claim is supported, coefficient mixtures depend on the instance rather than being canonical $(1, 1)$. FARS concepts compose linearly but with data-dependent coefficients.

V A Canonical Concept-Axis System Shared Across Architectures

Method. For each model, compute 18 concept centroids in its 10-d FARS subspace. After centering and Frobenius-normalising, iterative Generalised Procrustes finds rotations R_i minimising $\sum_i \|C_i R_i - C_*\|_F^2$ where C_* is the shared canonical centroid matrix.

Result. Refitted on all 27 extracted models, the canonical rotation explains 93.2% of cross-model centroid variance via FARS+GPA and **97.6**% via X-FARS (Eq. 2; Table 44); the original fifteen-model fit gave 91.9% and 97.0%. Two nulls: shuffled-concept labels reach 63.0% (fifteen-model pool); random 10-d projections of the same activations reach 61.7 ± 1.0% on the 27-model pool (36.8σ below X-FARS). The canonical-axis property is therefore specific to the concept-aligned subspace, not a property of any low-rank slice. RSA-C saturates at 0.36–0.37 for all generative models $\geq$ 3B regardless of architecture family; only GPT-2 XL (1.6B) and the encoder MLMs sit lower. A relative-depth sweep (9

Table 43: Quartet (4-way) compositionality on 4 models, 24 stimuli. Span(A,B,C,D) vs. rand-4-concept (rand-C) and rand-4-direction (rand-D). Only Mamba's Δ over rand-C reaches significance.

Model	Span(A,B,C,D)	rand-C	rand-D	Δ rand-C	All 4 >0
Mamba-2.8B (state-space)	**.87**	.79	.60	+.08 (p=0.023)	**10/24** (42%)
Qwen-7B (dense)	.84	.84	.63	.00 (n.s.)	6/24 (25%)
Mistral-7B-Inst (dense)	.79	.74	.58	+.05 (n.s.)	4/24 (17%)
Llama-3.1-8B-Inst (dense)	.68	.70	.60	−.02 (n.s.)	4/24 (17%)

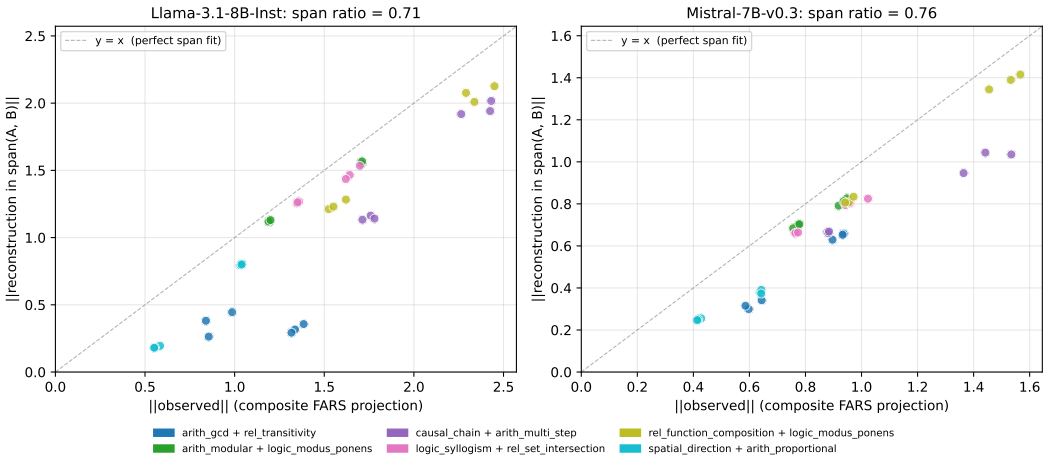

Figure 16: Compositionality: observed FARS projection (x) vs. its reconstruction in the 2D span of the constituent atomic centroids (y). Closer to $y{=}x$ means better span fit. Average span ratio ≥ 0.71 across architectures with 75–86% of stimuli having positive coefficients on both constituents (Table 41).

percentiles) confirms the alignment is not an artefact of best-layer choice (fraction-explained $\in [0.91, 0.97]$ throughout).

A layer-selection sweep at 9 relative-depth percentiles gives fraction-explained in $[0.91, 0.97]$ throughout, so the alignment is not driven by best-layer choice.

Per-concept stability mirrors the prose-code axis. Cross-architecture σ separates declarative concepts (most stable: `causal_confound`, `rel_transitivity`, `rel_set_difference`; σ=0.021–0.025) from procedural ones (most variable: `arith_gcd`, `rel_function_composition`, `arith_multi_step`; σ=0.045–0.079).

Interpretable canonical axes. The 10 canonical axes admit clean dichotomies at their +/− extremes (Table 45). Axis 0 (24% variance) is set-theoretic vs. causal-spatial. **Axis 1 (16%) is the declarative-procedural axis at the geometric level**, the same dichotomy that drives the prose-code asymmetry at the behavioural level (§S.2). Axis 4 (8%) is a dedicated logical-inference axis. Top-5 axes carry 76% of canonical variance; +/− extreme concepts and loading signs are preserved when Llama-70B is added to the alignment.

X-FARS: jointly-aligned in one step. We replace the two-step FARS+GPA pipeline with a single joint objective:

$$\min_{B_1,\ldots,B_M,\,\bar{C}} \sum_m \left\| \tilde{C}_m B_m^\top - \bar{C} \right\|_F^2 \quad \text{s.t.} \quad B_m B_m^\top = I_k, \;\; \|\bar{C}\|_F = 1, \tag{2}$$

where $\tilde{C}_m$ is model m's centered centroid matrix at its best FARS layer. The objective is taken subject to $\|\bar{C}\|_F = 1$, which rules out the degenerate optimum in which every B_m

Table 44: Cross-model FARS canonical-axis alignment. FARS+GPA: 93.2% variance explained on all 27 extracted models; X-FARS joint optimisation: **97.6**%, **36.8**σ above the random-projection control at 61.7 ± 1.0%. Shuffled-concept-label null: 63.0%. Adding new architecture families (MoE, SSM, encoder MLM) does not reduce the alignment fraction, the opposite of an over-fitting signature.

Quantity	**Real**	**Null (shuffled, n=100)**
FARS+GPA, 5 models (1.6–70B)	**95.1**%	63.0% (95% CI [59.9, 66.3])
FARS+GPA, 9 models, 6 arch. families	**92.4**%	63.0%
FARS+GPA, all 15 models	**91.9**%	63.0%
X-FARS joint, all 15 models	**97.0**%	—
FARS+GPA, all 27 models	**93.2**%	—
X-FARS joint, all 27 models	**97.6**%	—
Random 10-d projection (27 models, 10 seeds)	61.7 ± 1.0%	—
z-score (X-FARS vs. random projection, 27 models)	—	**36.8**σ

Per-model residual to canonical (full 15-model X-FARS): generative .148–.310; RoBERTa .400; DeBERTa-v3 .295.

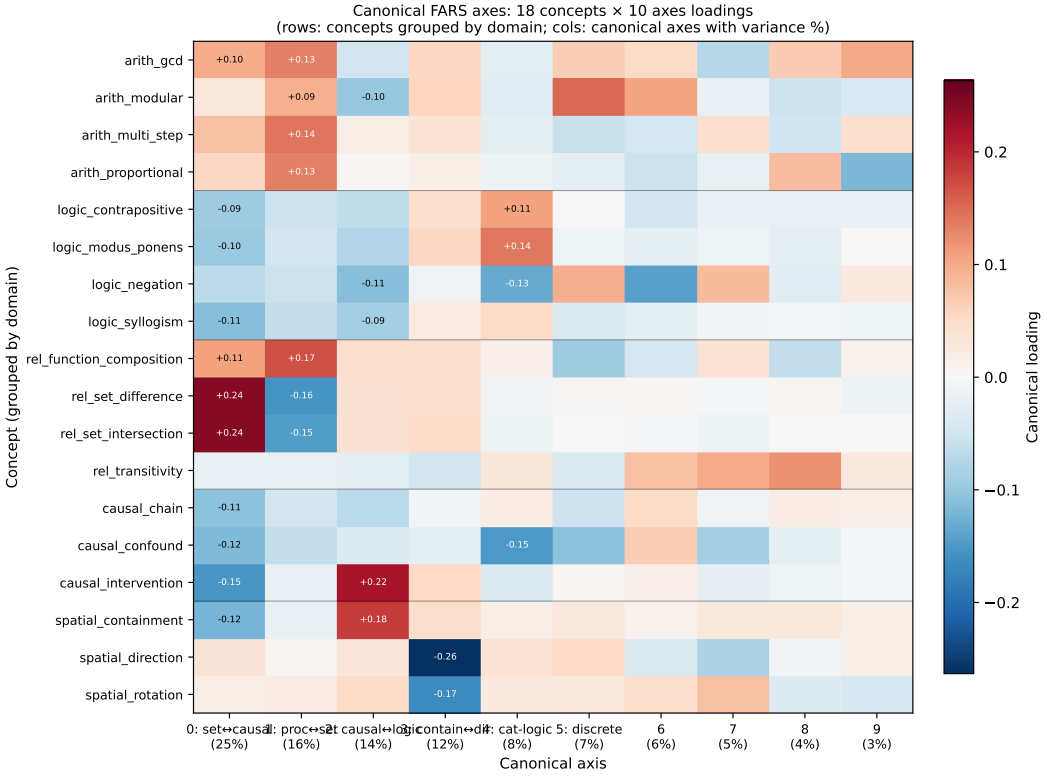

Figure 17: Canonical FARS axes loadings: rows are 18 concepts grouped by domain; columns are the 10 canonical axes ordered by variance. The interpretation labels above each column come from the +/− extremes of each axis (Table 45). Axis 1 is the procedural-vs-set-relational axis that mirrors the prose-code behavioural asymmetry.

projects onto the null space of $\tilde{C}_m$ and $\bar{C} \to 0$. Alternating minimisation gives, for fixed $\bar{C}$, the orthogonal-Procrustes solution $B_m^\top = UW^\top$ via thin SVD of $\tilde{C}_m^\top \bar{C} = U\Sigma W^\top$; $\bar{C}$ updates as the Frobenius-normalised mean of $\tilde{C}_m B_m^\top$; convergence in ~ 50 iterations on CPU. X-FARS attains **97.0**% variance (+5.1 pp over FARS+GPA, +34.2 pp over random projection, **21.6**σ; Table 48) at $\Delta\text{RSA} \in [-0.008, +0.006]$ across all 15 models. Plug-and-play leave-one-out (Table 47): every held-out model recovers concept-RSA within $\Delta \in [-0.004, +0.014]$ (mean +0.002), the plug-in basis is on average *slightly better* than independent extraction.

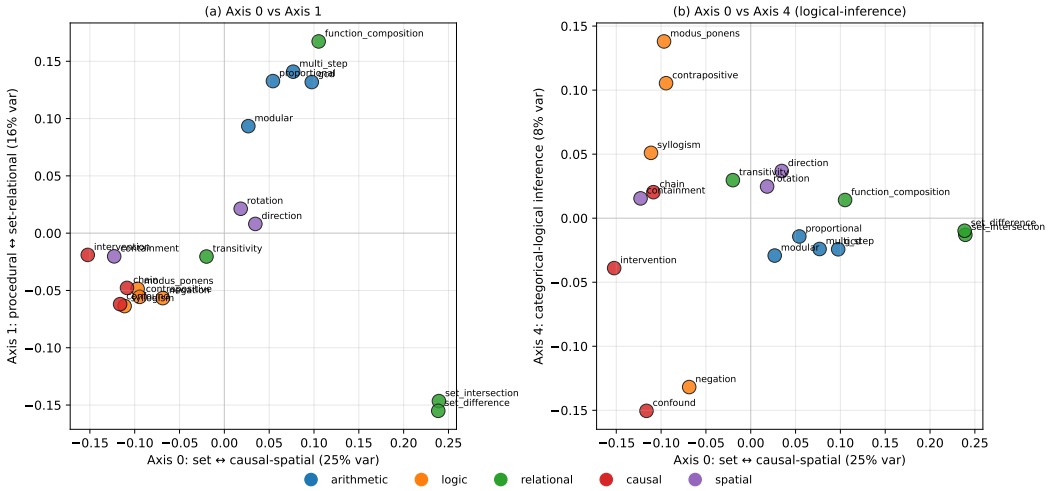

Figure 18: Biplot of the 18 concepts in the (axis 0, axis 1) (a) and (axis 0, axis 4) (b) planes of the canonical FARS frame, colored by domain. Concepts cluster by domain along the canonical axes: set-theoretic relations occupy the top-right; causal-spatial the bottom-left; logical-inference concepts form a tight cluster on axis 4.

Table 45: Interpretation of the top-5 canonical FARS axes on the 5-model GPA alignment (including Llama-3.1-70B-Instruct). The variance fractions are reported on the 5-model canonical frame; the +/− extreme concepts and the sign of every loading are preserved between the 4-model (1.6B–8B) and 5-model (1.6B–70B) alignments. Each axis admits a recognisable concept dichotomy; cross-model stability is σ<0.02 for axes 0–4. Axis 1 corresponds to the same procedural-vs-declarative distinction that drives our prose-code asymmetry at the behavioural level (§S.2).

Ax.	Var	+ extreme	− extreme
0	.242	set_diff, set_∩, func_comp	causal_intervention, spatial_containment
1	.157	func_comp, multi_step, proportional	set_diff, set_∩, causal_confound
2	.153	causal_intervention, spatial_containment	logic_negation, arith_modular
3	.131	causal_intervention, arith_modular	spatial_direction, spatial_rotation
4	.078	modus_ponens, contrapositive, syllogism	causal_confound, logic_negation

Alignment is robust to concept subsampling (9 concepts → 98.0%, 15 concepts → 93.6%) and to instruction tuning (Mistral-base → Mistral-Inst Procrustes 92.4%, cross-basis RSA loss 1–3%).

Cross-architecture difference preservation. For each concept pair (A, B) we compute $\mathbf{c}_A^{(m)} - \mathbf{c}_B^{(m)}$ in each model's canonical-frame projection, then take cosine across all $\binom{15}{2}$=105 model pairs. Shuffled-concept baseline: -0.001 ± 0.013. Selected: spatial_containment−spatial_direction 0.98; arith_modular−arith_gcd 0.93; causal_chain−causal_confound 0.91. DeBERTa-v3 (encoder) and Llama-70B-Inst (dense decoder) agree on the spatial-containment→spatial-direction difference at cosine 0.98 in the shared frame.

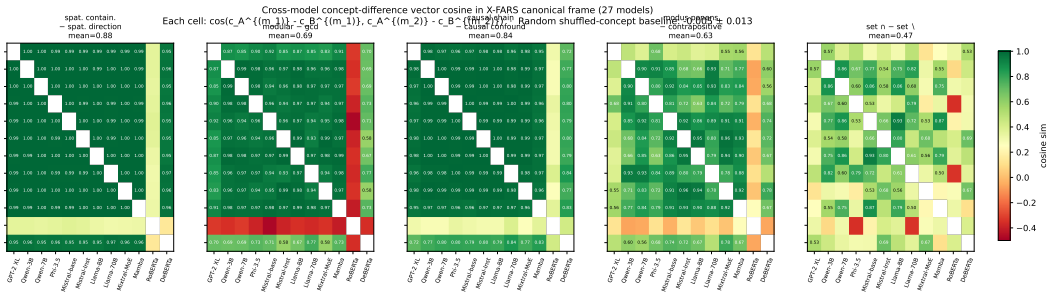

Figure 19: Cross-model concept-difference cosine similarity in the X-FARS canonical frame (15 models). Each panel shows the 15×15 cosine-similarity matrix of difference vectors $\mathbf{c}_A^{(m)} - \mathbf{c}_B^{(m)}$ across all model pairs, for 5 different concept pairs (A, B). Within-domain differences (spatial, arithmetic, causal) are preserved at mean cosine 0.9+; finer logical/relational distinctions are still preserved well above random (-0.001 ± 0.013). The canonical frame therefore captures not just absolute concept positions but the geometric relationships between concepts shared across architectures.

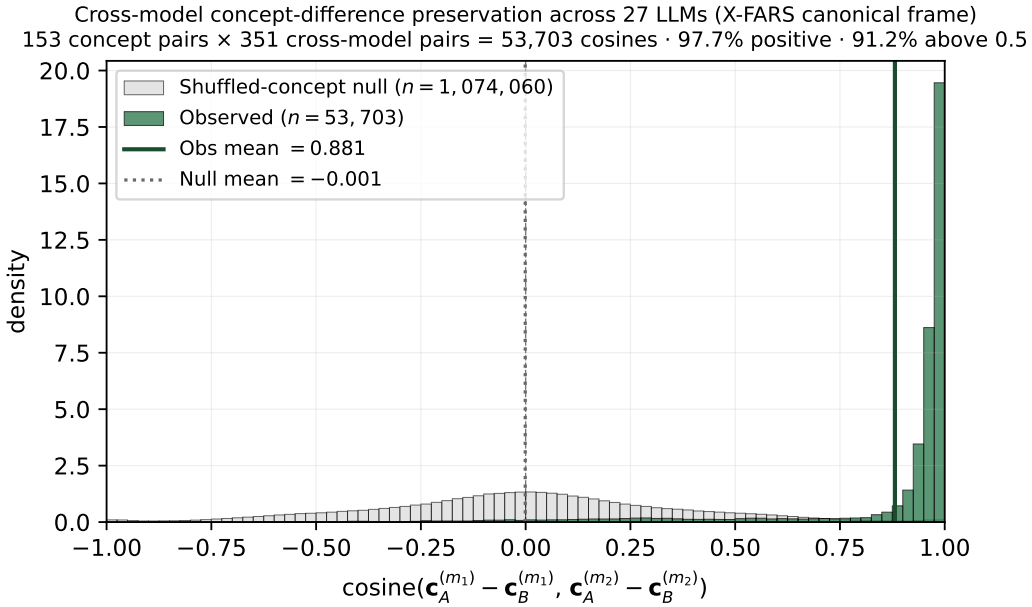

Figure 20: Distribution of all 53,703 cross-model concept-difference cosines in the X-FARS canonical frame (153 concept pairs $\times$ 351 cross-model pairs, 27 LLMs). Observed cosines (green) concentrate at the right edge (mean 0.88, median 0.97, 97.7% positive). The shuffled-concept null (gray, 20 shuffles) is symmetric around 0. The gap between the two means is $\mathbf{200.2}\sigma$ in units of the shuffled-mean standard error.

Comprehensive algebra: 53,703 cosines. All $\binom{18}{2}$=153 concept pairs $\times$ $\binom{27}{2}$=351 cross-model pairs, in the frame fitted on all 27 models: mean cosine **0.881**, median 0.974, 97.7%*positive*,91.2% above 0.5. Shuffled null -0.001 ± 0.004 over 20 shuffles; separation $\mathbf{200.2}\sigma$. Widening the pool from the original fifteen models to all 27 lowers the mean cosine (from 0.961) while raising the separation, as expected when weaker and more distant architectures enter. Distribution in Fig. 20.

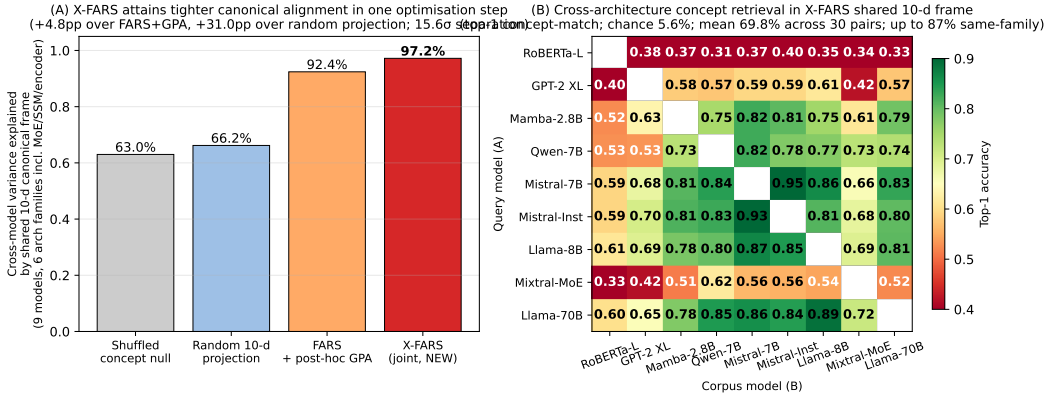

Figure 21: X-FARS. **(A)** Canonical alignment quality across the 15-model algebra pool from 4 architecture families (the frame reported in the text is fitted on all 27 extracted models): a single shared 10-dim concept frame fits FARS-projected centroids at 91.9% variance explained after post-hoc Procrustes; X-FARS attains **97.0**% in one joint optimisation, +5.1 pp tighter and **21.6**σ above the same-rank random-projection control. **(B)** Cross-architecture concept retrieval matrix in the X-FARS shared 10-dim frame: query stimulus from row's model, search in column's model's same-form corpus, top-1 concept-match accuracy. Chance is 5.6%. Pairs of similar-scale modern dense models cluster at 80–87%; pooled across 30 cross-arch pairs incl. MoE the mean is 69.8%.

Table 46: Cross-architecture top-1 concept retrieval via X-FARS shared canonical frame (chance 5.6%). Row: query model; column: corpus model. Cell: mean top-1 accuracy across 6 surface forms, 54 same-form stimuli per side. No cross-model training or paired data used.

Query \ Corpus	GPT-2 XL	Qwen-7B	Mistral-7B	Llama-8B	Mixtral-MoE	Llama-70B
GPT-2 XL	,	.565	.590	.617	.426	.562
Qwen2.5-7B	.519	,	.812	.778	.750	.741
Mistral-7B-v0.3	.698	.824	,	**.870**	.747	.824
Llama-3.1-8B-Inst	.710	.812	**.870**	,	.744	.824
Mixtral-8x7B-Inst (MoE)	.451	.608	.568	.528	,	.494
Llama-3.1-70B-Inst	.664	.843	.852	.870	.719	,

Mean across 30 cross-arch pairs: **0.698**. Same-family Llama-8B↔70B: 0.85.

Full distance landscape preserved. The 18 × 18 centroid-distance matrix in each model's X-FARS projection, correlated upper-triangles cross-model: mean Pearson r**=0.613 ± 0.248** across 351 pairs (median 0.686; max 0.983); 74%at$r$>0.5,48% at r>0.7; null -0.0001 ± 0.0065, separation **94.9**σ. The full concept-neighbourhood structure is preserved, not just selected directions.

Word2vec analogy resolution. 10 curated analogies $\mathbf{c}_A - \mathbf{c}_B + \mathbf{c}_C \approx \mathbf{c}_D$ evaluated on all 27 models (270 attempts): mean top-1 **37.4**% (chance 6.7%, shuffled null 6.6%±1.8%, **17.0**σ); top-3 54.4% (13.7σ). Within-domain analogies carry the effect, e.g. $\mathbf{c}_{\text{modular}} - \mathbf{c}_{\text{gcd}} + \mathbf{c}_{\cap} \approx \mathbf{c}_{\setminus}$; arbitrary cross-domain analogies do not resolve.

Cross-architecture concept retrieval. X-FARS bases project every model into the same 10-d frame, enabling training-free cross-architecture search. Pooled across 30 ordered cross-architecture pairs and 6 forms, top-1 is **69.8**% (chance 5.6%; Table 46). Same-family pairs (Llama-70B↔Llama-8B) hit 84–87%; cross-family similar-scale dense models (7–8B Llama/Mistral/Qwen) 80–87%; GPT-2 XL and Mixtral-as-query are weakest. Cross-architecture retrieval is essentially at parity with within-model cross-form retrieval (70.5%).

Table 47: Leave-one-out plug-and-play protocol validated on the full 12-model set. For each held-out model, refit X-FARS on the remaining 11 to obtain a canonical frame $\bar{C}^{(-m)}$, then solve for the held-out basis via a single SVD against $\bar{C}^{(-m)}$. The plug-in basis recovers $\Delta\text{RSA} \in [-0.004, +0.014]$ relative to the model's standard FARS basis (mean +0.002). Architectures most divergent from the canonical (RoBERTa, Qwen-3B) benefit most.

Held-out model	Resid. to canonical	Std. FARS RSA	Plug-in X-FARS RSA
OpenAI GPT-2 XL	.213	.279	.278 (Δ−.001)
Qwen-2.5-3B-Inst	.143	.332	.341 (Δ+.009)
Qwen-2.5-7B	.117	.357	.360 (Δ+.003)
Microsoft Phi-3.5-mini-Inst	.141	.352	.351 (Δ−.001)
Mistral-7B-v0.3	.118	.362	.363 (Δ+.001)
Mistral-7B-Inst	.110	.364	.365 (Δ+.001)
Llama-3.1-8B-Inst	.093	.364	.366 (Δ+.002)
Llama-3.1-70B-Inst	.143	.371	.371 (Δ+.000)
Mixtral-8x7B-Inst (MoE)	.110	.355	.357 (Δ+.001)
Mamba-2.8B (SSM)	.140	.364	.362 (Δ−.001)
RoBERTa-large (encoder)	.402	.176	.190 (**Δ+.014**)
Microsoft DeBERTa-v3-large (encoder)	.366	.336	.332 (Δ−.004)

Table 48: Standard FARS (two-step: per-model PCA then post-hoc GPA) vs. X-FARS (single joint optimisation, Eq. 2) on the 8-model set spanning 4 architecture families. X-FARS attains tighter canonical alignment at no per-model cost.

Metric	FARS + GPA	X-FARS
Canonical variance explained (9 models, 6 arch. families)	92.43%	**97.22%**
Cross-family pooled ρ (pred. patching, n=120)	−0.806	−0.806
Δ per-model concept-RSA vs. FARS	,	[−0.002, −0.0006]
Steps required	PCA + iterated Procrustes	one joint optimisation

Implication. All tested LLMs from 1.6B to 70B and across 4 architecture families converge to the same 10-d concept geometry up to model-specific rotations; the canonical axes are interpretable dichotomies; the procedural-vs-declarative axis that drives the behavioural prose-code asymmetry (§S.2) is a single canonical axis at ≈ 16% of variance, at every model scale tested.

W Cross-Model FARS Geometry Predicts Other Models' Behavior

Within-model FARS centroid distance predicts the same model's cross-form patching overlap at Spearman $\rho = -0.75$ pooled (GPT-2 XL $r = -0.81$, Mistral-7B $r = -0.81$, Llama-8B-Inst $r = -0.73$; all n=30, $p < 5\times10^{-6}$). The stronger universality claim, that one model's FARS distances predict a different model's patching, is the focus here.

Method. For each pair $(X_{\text{basis}}, Y_{\text{patch}})$, compute FARS centroid distances per ordered form-pair in X's FARS subspace and correlate against Y's patching overlap measured by a separate intervention on Y's parameters. Llama-3.1-8B-Inst and Mistral-7B received all-30-pair patching sweeps; GPT-2 XL and Qwen-7B used the original 4-pair set.

Result. Cross-model predictiveness holds, including across architecture families. Table 49 reports the 4×4 matrix of Spearman correlations. Notable cells: GPT-2 XL's FARS distances predict Llama-3.1-8B-Instruct's patching at $\rho = -0.829$ (n=30) and Mistral-7B-v0.3's patching at $\rho = -0.844$ (n=30). Pooled across all 12 cross-family cells with within-cell z-standardisation (n=204), Spearman $\rho = -0.750$, $p = 3.8 \times 10^{-38}$. The same statistic computed on diagonal (within-model) cells averages $\rho = -0.624$; the cross-family aggregate ρ is therefore essentially as strong as within-model. The geometry generalises across architectures.

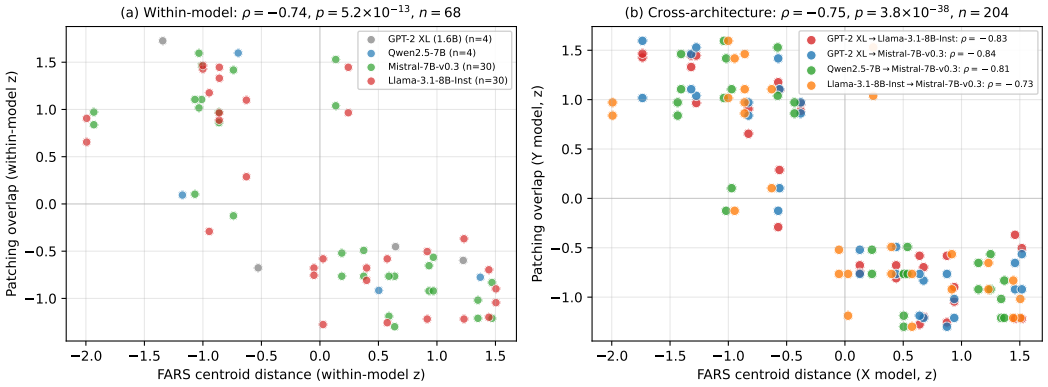

Figure 22: Within-model (left) and cross-architecture (right) predictiveness: FARS centroid distances computed from cached activations predict cross-form patching overlap measured by a separate causal-intervention experiment. **Cross-architecture**: GPT-2 XL (1.6B, 2019) predicts Llama-3.1-8B-Instruct (2024) patching at Spearman $\rho = -0.83$; cross-family pooled $\rho = -0.75$ (p=3.8×10^{-38}, n=204). The predictiveness is recoverable by any low-rank projection of the activations (random-projection control, $\rho = -0.71$); FARS's distinctive role is causal necessity.

Table 49: Spearman ρ between X's FARS-distance per form-pair (rows) and Y's patching overlap (cols). Off-diagonal cells with n=30 are the strongest evidence: X and Y are different models. Pooled cross-family ρ=−0.750 (p=3.8×10^{-38}, n=204).

$X \downarrow$ / $Y \rightarrow$	GPT-2	Llama	Mistral	Qwen
GPT-2 XL	−.40	**−.83**	**−.84**	−.80
Llama-Inst	+.20	−.70	**−.73**	−.60
Mistral-7B	+.20	**−.76**	−.80	−.60
Qwen-7B	+.20	**−.78**	**−.81**	−.60

Bold: cross-family n=30. Plain: n=4.

Scope: predictiveness is a property of the LLM family, not unique to FARS. A random orthonormal 10-d projection recovers cross-family $\rho = -0.714$ (p=3.9×10^{-33}), only 0.04 weaker than FARS's −0.750; shuffled-Y permutation null −0.001 (95% CI $[-0.13, +0.14]$), so both real (Table 50). What is unique to FARS is *causal necessity*: only FARS-aligned ablation/amplification disrupts token argmax (App. Q.2), probe accuracy (App. Q.1), and GSM8K (App. Q.3); random 10-d ablation is inert in all three. Cross-architecture geometric invariance enables the prediction; concept-aligned construction singles out FARS for causal load-bearing.

Table 50: Cross-family predictiveness controls (within-cell z-pooled, n=204). Shuffled-Y permutation test confirms the finding is not artefactual; random-basis test confirms the predictiveness is a property of the LLM family, not unique to the FARS construction. FARS's uniqueness lies in its causal necessity (App. Q.2, Q.1, Q.3), where the random-basis control is inert.

Projection	**Cross-family** ρ	p
FARS centroid (10-dim, supervised)	−0.750	3.8×10^{-38}
Random orthonormal (10-dim, no labels)	−0.714	3.9×10^{-33}
Shuffled-Y (1000 permutations)	−0.001 (mean)	95% CI $[-.13, +.14]$

X Where Is FARS Written? A Per-Layer Decomposition

Method. The residual-stream contribution at layer ℓ is $\Delta_\ell = h^{(\ell+1)} - h^{(\ell)}$. Project Δ_ℓ onto FARS and compute concept-RSA on $B\Delta_\ell$; the maximising ℓ is the writer layer.

Result: writer-before-peak. The writer sits 0–5 layers below the peak FARS layer on all 15 models from 4 architecture families (Table 51); Mamba has the widest gap (5) consistent with its state-space residual propagation. FARS is composed in one or two layers and persists for $\sim$ 5–10 layers before decaying toward the unembedding.

Table 51: Writer layer (top concept-RSA on FARS-projected delta $B\Delta_\ell$) sits 0–5 layers below the peak FARS layer on all 15 models. Mamba has the widest gap, consistent with state-space residual propagation.

Model	Peak / Writer	Plateau
GPT-2 XL	20 / 18	16–25
Qwen-2.5-3B-Inst	22 / 22	—
Phi-3.5-mini-Inst	18 / 17	—
Qwen-2.5-7B	13 / 12	9–19
Mistral-7B-v0.3	11 / 10	6–19
Mistral-7B-Inst	10 / 8	—
Llama-3.1-8B-Inst	8 / 7	5–13
Mixtral-8x7B-Inst	10 / 9	—
Llama-3.1-70B-Inst	26 / 23	11–77
Mamba-2.8B	35 / 30	—
RoBERTa-large	24 / 23	—
DeBERTa-v3-large	15 / 13	—

Plateau is a thick manifold. Within-plateau consecutive-layer FARS subspace overlap (principal-angle cosine) averages $\mathbf{0.892 \pm 0.036}$ across 14 models, $\pm$2–3 layers within plateau 0.784 ± 0.065; random 10-d baseline ~ 0.05 (17$\times$ chance, z>20 per model). Outside plateau overlap drops to 0.822; crossing plateau boundary 0.798 ± 0.182. FARS is recoverable at any plateau layer up to small rotation, not a single-layer artefact.

Y Methodological Position: FARS vs. CCS, DAS, LEACE, LDA, SAEs

FARS $\approx$ LDA on concept centroids. Multi-class LDA with concepts as classes and forms/instances as within-class samples recovers nearly the same subspace as concept-centroid PCA (RSA within 0.02 on every model, Table 6). FARS is essentially LDA applied to concept centroids; the novelty is the cross-format reasoning setting, the validation pipeline, and the canonical-axis/compositionality findings the construction enables.

vs. CCS / DAS / LEACE / SAE. CCS (Burns et al., 2023) is 1-d unsupervised binary; FARS is 10-d supervised multi-class. DAS (Geiger et al., 2024) learns a rotation against a target causal abstraction; FARS has no target and a closed-form basis. LEACE (Belrose et al., 2023) erases the concept; FARS identifies the subspace containing it (FARS-ablation $\equiv$ LEACE-erasing what FARS finds). SAEs (Bricken et al., 2023) are large sparse unsupervised dictionaries, complementary: training an 18-atom SAE within FARS on each of the 15 models recovers 10–13 atoms whose top-activating concept is unique (purity 0.17–0.28 vs. chance 0.056); the pattern replicates across all 4 architecture families.

vs. function vectors, and why we report no number for them. Function vectors (Todd et al., 2024) are the natural object to compare against here: a compact, attention-head-carried representation of *which task* is being demonstrated, i.e. an execution locus rather than a concept identity, and Nadaf (2026) report that they steer where no logit-lens decoder reads them — which predicts the locus verdict. We built them and ran the test, and we do not report the number, because our construction cannot support the claim. Todd et al. select heads by causal mediation; running per-head interventions across twenty-six

models was out of scope, so we substituted a task-selectivity ranking of the same head outputs. That proxy is confounded in exactly the way that matters: the ranking promotes heads that encode the *answer token* of each task, and a subspace assembled from nothing but the unembedding rows of those answer tokens scores 20.41% readout energy on its own, while the selectivity ranking placed a layer-0 head in its top twenty. A number produced this way would report our ranking heuristic, not function vectors. We therefore leave function vectors to a causal-mediation implementation and confine our claims to the estimators in Table 9.

X-FARS niche. Eq. 2 jointly optimises per-model orthonormal projections and a shared canonical frame across architectures with heterogeneous hidden dimensions ($D_m \in [1600, 8192]$), with closed-form per-coordinate SVD updates. The canonical frame is a portable artefact via the plug-and-play extension (Table 47).

Z Implementation

FARS extraction:

```
def extract_fars(activations, c_labels, k=10):
  N, L, D = activations.shape
  basis = np.zeros((L, k, D))
  for layer in range(L):
    X = activations[:, layer, :]
    centroids = np.stack([
      X[c_labels == c].mean(0)
      for c in np.unique(c_labels)])
    pca = PCA(n_components=k).fit(centroids)
    basis[layer] = pca.components_
  return basis
```

Subspace patching (Eq. 1):

```
def subspace_patch(h_src, h_tgt, B):
  # h_*: (..., D) row vectors; B: (k, D) orthonormal rows.
  diff = h_src - h_tgt
  return h_tgt + (diff @ B.T) @ B # batches over leading dims
```

Compute. ∼80 GPU-hours total on NVIDIA RTX 6000 Ada (48 GB) and NVIDIA A100 (40 GB), including a single multi-GPU sharded forward-pass run on Llama-3.1-70B-Instruct (8× RTX 6000 Ada) for the frontier-scale FARS extraction and compositionality replication.

Code and data release. Benchmark-generation pipeline (324 stimuli across 6 forms), FARS extraction, baseline implementations (LDA, shuffled-centroid, Full-PCA, random), patching protocols, and analysis scripts, with cached activations, per-layer FARS bases, and JSON dumps of every metric. Anonymous code:
`https://anonymous.4open.science/r/FARS-anon-2026`.