A LEARNED DOMAIN-DISCRIMINATIVE SUBSPACE OF THE RESIDUAL STREAM

Anonymous authors
Paper under double-blind review

ABSTRACT

Do trained language models represent *domain identity*, i.e., how the text they are processing departs from their default distribution, in structured form, and is that structure learned rather than an artifact of how we probe for it? We define V_{joint}, a per-layer subspace of the residual stream obtained by joint diagonalization of per-domain generalized-eigenvalue discriminants against generic text, and recover the same compact object in five model families spanning three token-mixing mechanisms. It occupies 1–4% of the hidden dimension, 4.9–5.6% in the mixture-of-experts read one sub-layer earlier, sized by the number of calibrated domains rather than by stream width. Because such a construction returns a basis from anything, the depth claim is made at fixed rank and against a floor drawn from its own candidates—an arbitrary subspace of equal rank, which has a rising gate of its own. Against that floor, discriminability along V_{joint} rises with depth in every trained checkpoint, decays in random-initialization controls, and clears it by a near-constant 1.5–1.7$\times$ in all five families. Pseudo-concepts stay an order of magnitude below at every depth and rank. Removing V_{joint} degrades even the least affected calibration corpus 17–98$\times$ more than the generic reference, whereas a matched-rank variance-only subspace damages domain and generic text indiscriminately. Yet nothing we can measure reads the code, and its causal necessity does not extend beyond the corpora it was calibrated on, even though its discriminative format does. What survives is a learned, compact, corpus-selective code.

1 INTRODUCTION

A widely discussed hypothesis in interpretability is that neural networks trained on similar data converge to similar representations, the "platonic representation" view (Huh et al., 2024), foreshadowed by the linear representation hypothesis (Park et al., 2024) and by cross-model concept alignment (Huang et al., 2025). Most evidence concerns individual concept directions or global geometry. We ask a sharper question about the representational mechanism behind one concrete feature, the identity of the text *domain* a model is processing (Aharoni & Goldberg, 2020): *where and how does a trained language model store domain identity, and is that storage a learned object or an artifact of how we look for it?*

Our answer is an object we call V_{joint}: a per-layer subspace of the residual stream obtained by joint diagonalization of per-domain generalized-eigenvalue (GEV) discriminants against generic text. It is small, 37–103 directions at a five-domain calibration, and the same compact geometry appears in five model families spanning three token-mixing mechanisms. Throughout, domain identity is operational and contrastive: it measures how the current text departs from the model's default distribution, not a register that names one of D classes, and §4.4 finds the code selective for the particular corpora it was calibrated on rather than for domain membership as such.

The contribution is not that such a subspace exists, since a per-layer GEV returns a basis from anything. It is the shape of the object that a battery of controls establishes. Table 1 says what we mean by universality: measurable properties reproduced across families and mechanisms. Yet the code is local in *coordinates*: no linear map we fit carries V_{joint} between families, against a same-model control on which the same instrument works. Kind converges; coordinates do not.

The rise of discriminability across depth is what we measure, and we call it the **depth gate**; what licenses reading it as "a learned code" rather than "an artifact of looking" is a control on two axes. The two axes carry different evidence: a sign for the model axis and a level for the concept axis; only

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.

the model axis is developmental. Throughout, when a control shows that a property is generic over the residual stream, over a high-variance space, or over any subspace of the same rank, we note this and do not attribute it to V_{joint}.

Contributions. 1) *A compact contrastive code* (§4.1): the construction returns the same kind of object under dense attention, hybrid linear attention, an attention-free state-space model, and a sparsely-routed mixture-of-experts feed-forward — a few percent of the hidden dimension, single-tiered, defined on second-moment rather than class-mean structure. 2) *Learned, and what it takes to show that* (§4.2): an arbitrary subspace of equal rank has a rising gate of its own, and a metric that sums over directions is not comparable across ranks, so a depth-wise trend is evidence only at fixed rank and against a matched-rank floor. Under both, discriminability rises in every trained checkpoint, decays in two random-initialization controls — one given the trained embedding table — and converges to a common level across families. 3) *Domain identity is constructed, language identity is given* (§4.2): pseudo-concepts never leave the floor, while language, a far more salient property, does acquire a code but starts high and grows several times less. 4) *Necessary as substrate, not read as signal, and local to its corpora* (§4.3, §4.4): removing V_{joint} degrades the calibration corpora an order of magnitude more than generic text, where a matched-rank PCA subspace does more damage indiscriminately, yet no test we ran reads or steers with it and no linear map we fit transports it. Its format generalizes beyond the calibration; its necessity does not.

2 Related Work

Representation convergence. The platonic-representation hypothesis (Huh et al., 2024) and cross-model concept transfer (Huang et al., 2025; Puri et al., 2025) argue that different models reach aligned representations, usually by aligning individual directions or scoring global geometry; a recent critique (Gröger et al., 2026) finds the global form fragile and re-centers convergence on local structure. We contribute a concrete object: a learned domain-discriminative subspace universal in *character* across families, whose *coordinates* are model-local and carried neither by an orthogonal map nor by the unconstrained ones we fit (Appendix K). It is a multi-distribution, discriminatively defined instance of the low-dimensional structure the linear representation hypothesis (Park et al., 2024) and superposition (Elhage et al., 2022) motivate, fused across domains by joint diagonalization (Cardoso & Souloumiac, 1993; Flury, 1988).

Contrastive subspaces. Domains already form unsupervised clusters in pretrained representations (Aharoni & Goldberg, 2020). We ask not whether that separation exists — it is decodable at the first layer (§4.2) — but whether a dedicated second-moment code for it is maintained and built across depth. The closest construction is Dasgupta et al. (2026), which solves a generalized eigenproblem between semantic-changing and paraphrase-preserving second-moment covariances. The contrasts are orthogonal: theirs isolates features invariant to meaning-preserving rewrites, ours which *kind* of text is being processed. Our central result is developmental rather than geometric, and on cross-model structure the two agree from opposite directions: the model-specific geometry of their subspace is what lets them use it *for* model attribution, and our transport experiment finds the domain code equally model-local (§4.4).

Neighbors ruled out, probing and steering. V_{joint} is not the input-independent high-norm "massive-activation" dimensions (Sun et al., 2024), which are fixed outliers. The high-norm class-mean content is exactly the part we show is *not* V_{joint}-specific. We use a dense low-rank basis rather than sparse-autoencoder features (Bricken et al., 2023; Cunningham et al., 2023), whose sparsity prior may fragment a distributed second-moment direction-set, we characterize a *content* subspace. The probing literature's deflationary expectation (Tenney et al., 2019; Hewitt & Liang, 2019), that trained models make most high-level properties more separable with depth, is tested in §4.2: no pseudo-concept comes within an order of magnitude of the domain code. Steering and subspace-adaptation methods build their subspaces one behavior at a time, from contrastive prompts (Zou et al., 2023), by conditional activation steering (Lee et al., 2025), or as SAE-constructed safety subspaces initializing adapters (Wang et al., 2025). Because our causal tests do not clear their controls we frame V_{joint} as a discriminative code rather than a steering knob.

3 METHOD

3.1 CONSTRUCTING V_{joint}

Every architecture we study is built from L blocks that read from and additively write to a common H-dimensional stream, $x^{(l+1)} = x^{(l)} + \mathcal{N}(x^{(l)})$, with $x^{(0)}$ the token embedding. What varies is the operator $\mathcal{N}$: attention and MLP sub-layers in a transformer, a linear-attention block in a hybrid, a selective state-space mixer in Mamba-2, attention and a sparsely-routed expert block in OLMoE. Every analysis below operates on $x^{(l)}$ and never on $\mathcal{N}$, so it applies uniformly across them. We sample *the residual stream leaving each block*, the one definition they all share. Mamba-2 has no separate MLP sub-layer, so any within-block convention would not be comparable across families. The one exception is OLMoE, read at its router input one sub-layer earlier (§3.3).

We ask whether that stream contains directions domain-specific text uses far more strongly than generic text. For D domain corpora and one generic corpus g we collect per-token activations at each layer, center them per corpus, and form covariances Σ_d, Σ_g, which differ only in which text produced them. We use a discriminant rather than a probe because it returns a *ranked* basis with nothing fitted to the labels we assign (Appendix A.1). The directions of interest solve a generalized eigenvalue problem (GEV), $\Sigma_d\, v = \lambda\,(\Sigma_g + \epsilon I)\, v$, with $\epsilon = 10^{-3}$ of the mean generic variance, a Tikhonov ridge that floors the denominator so near-zero generic variance cannot produce large eigenvalues. At the optimum each eigenvalue is a Rayleigh quotient, $\lambda = v^\top \Sigma_d v \,/\, v^\top(\Sigma_g + \epsilon I)v$: the variance domain text shows along v over the regularized variance generic text shows along it. Eigenvectors with $\lambda \gg 1$ are directions the domain amplifies relative to the model's default, the object of interest. $\lambda \approx 1$ marks directions both distributions use alike, $\lambda < 1$ ones the domain suppresses.

The estimated Σ_g carries many near-zero-variance directions along which a domain-to-generic ratio is unstable. Hence, we pool the two centered activation matrices, take their top $k = 2k_{\text{each}} = 256$ principal components, solve in that *working subspace*, and rotate back. Pooling rather than either side alone retains directions salient for the domain but negligible for generic text. Two consequences matter later: V_{joint} lives inside high-variance space by construction, which is what makes the matched-rank PCA control a genuine neighbor rather than a straw man, and k caps each per-domain basis, so we report dimensionality at a fixed criterion and measure its dependence on k_{each} (Appendix A.2).

Solving the GEV per domain yields D bases $\{V_d\}$, which we fuse. We stack their columns weighted by $w_{d,j} = \max(\lambda_{d,j} - 1,\, 0)$ into a matrix M and keep the leading left singular vectors up to a fraction $\theta = 0.97$ of $\|M\|_F^2$, giving $V_{\text{joint}}(l) \in \mathbb{R}^{H \times K_{\text{joint}}(l)}$. The weight is zero at parity and linear in the excess, so directions a domain does not amplify contribute nothing. Uniform weighting would let the near-parity bulk dominate the SVD (at the first layer the median λ is 1.04). Because the weight is per-domain and the fusion a stacked SVD, one large λ in one domain outweighs four moderate ones across domains, so the construction favors directions a single corpus amplifies sharply over directions several corpora amplify moderately. And because the covariances are centered, V_{joint} is defined on *fluctuation* structure; the first-moment shift is analyzed separately and is not V_{joint}-specific (§4.1). Its rank $K_{\text{joint}}(l)$ is measured, not chosen: the energy criterion sets it at each layer. K, by contrast, is a rank we choose when scoring a subspace, with *own rank* $K = K_{\text{joint}}(l)$ and *fixed rank* $K = 32$ (§3.2).

3.2 CONTROLS AND REPORTING

A per-layer GEV is guaranteed to return a discriminative direction at every layer, so no property of V_{joint} can be read as a finding without a control that removes exactly one ingredient of the claim. We use five, each varying one thing and holding the rest fixed. **Random-initialization models** vary the *training*, since an interpretability metric can score as well on an untrained network as a trained one (Heap et al., 2026): two arms on the Llama-3.2-3B architecture with five seeds each, one fully random, one copying the trained embedding tables into random blocks so the real $x^{(0)}$ is present at every layer. **Concept controls** vary the *classes*, running the byte-identical pipeline on an arbitrary content-blind partition of the generic corpus, on two matched-frequency style-quintile partitions, and on language identity. **Matched-rank random orthonormal subspaces** vary the *basis*: we draw them ambiently in $\mathbb{R}^H$ and, in the demanding form we quote, from the span of the amplified candidate columns the fusion chose from, asking whether its selection beats an arbitrary selection from its own candidates (a third, intermediate draw and a caveat on metric orthonormalization are in Appendix A.4). **Matched-rank PCA subspaces** vary the *criterion*, replacing the variance ratio with variance alone; they are the strongest variance-only imitator of V_{joint}, and together with the

random draws they bracket the neighborhood it is selected from. A **chance baseline** calibrates the *alignment* measurements: containment of one K-dimensional subspace in another grows with K/H, so we quote overlaps as multiples of the share two unrelated subspaces of the same rank would give (cf. Gröger et al., 2026, on the same confound at model scale).

Fisher S/N is our depth-resolved discriminability measure. Project evaluation activations onto $V_{\text{joint}}(l)$: for each retained direction form the ratio of between-class variance, domain against generic class means of the projection, to pooled within-class variance, sum over directions and average over domains. Two properties follow and must be kept apart. Every term is itself a ratio, so the metric is invariant to the growth of activation norms with depth — a rising curve cannot be an artifact of scale — and unlike probe accuracy it does not saturate near a ceiling. But every term is non-negative, so the metric grows monotonically in the number of retained directions and is *not* comparable across subspaces of different rank, which holds for any discriminability measure that sums over directions, not only ours.

Since $K_{\text{joint}}(l)$ varies both with depth and across arms, we report the gate three ways. At each layer's **own** $K_{\text{joint}}(l)$. At a **fixed rank** $K = 32$, our primary form, with per-arm layer filtering: any layer whose K_{joint} falls below 32 is dropped, and one such drop bears on a claim — the language arm loses L21–L26 (Appendix A.5). And **per direction**, Fisher S/N divided by retained rank, which is rank-free and is what permits comparing arms whose K_{joint} differs at all. A rank-free alternative exists in a code-length or MDL probe (Voita & Titov, 2020), which we did not take, for reasons given in Appendix A.6.

Depth-slope is the OLS slope of Fisher S/N on relative depth, 0 at the first block and 1 at the last, with a 95% CI from a sequence-level bootstrap (Efron & Tibshirani, 1994) (Appendix A.7). It summarizes the direction of a monotone trend rather than asserting linearity, and because Fisher S/N scales with level, every slope compared across arms is computed at the fixed rank above with the own-rank value beside it. Beyond these scalars we read the subspace through an operator that shares no machinery with the eigenproblem: causal intervention in the forward pass, projecting a subspace out of the stream at an anchor layer and measuring the per-corpus change in loss.

3.3 Models, data, and protocols

Architectures. Five trained families spanning three token-mixing mechanisms: dense softmax attention (**Llama-2-7B**, **Llama-3.2-3B**; Touvron et al., 2023; Grattafiori et al., 2024), hybrid linear attention (**Qwen3.5-9B**; Qwen Team, 2026), and an attention-free state-space model (**Mamba-2-2.7B**; Dao & Gu, 2024). A fifth family, **OLMoE-1B-7B** (Muennighoff et al., 2025), varies the feed-forward rather than the mixer — dense attention, but each token routed to 8 of 64 experts — on the three properties we can measure in it. We call the first four, sampled at the stream leaving each block, the *block-output families*. All but Qwen3.5-9B are base checkpoints. Qwen is post-trained and multimodal, run text-only, so it confounds training stage with architecture, and we separate the two by also running **Qwen3.5-9B-Base** through the gate (§4.2). OLMoE is not sampled at block output: a forward pre-hook on the sparse expert block captures the tensor its router consumes, one sub-layer earlier than in the other four (Appendix A.3), so we report its numbers separately rather than merged into a cross-family range. We make no claim about whether the router reads V_{joint}. All experiments run on publicly released checkpoints.

Datasets. We instantiate $D = 5$ domains — *biomedical*, *source code*, *finance*, *legal*, *math* — chosen as major, well-resourced distributions lexically distant from web text and from one another, and deliberately mixed in register (running prose, source code, question–answer formats) so that "domain" is not confounded with a single text format (Appendix B.1). They are corpus choices rather than natural kinds. The generic reference g is **C4** (Raffel et al., 2020), the closest public proxy for a broad pretraining distribution. Recomputing against wikitext (Merity et al., 2017) leaves the leading directions in place (Appendix B.2). For the scaling analysis the pool extends to 62 domains, the five composites plus 57 MMLU subjects (Hendrycks et al., 2021), deliberately overlapping and partly nested — the regime that claim is about, and a hedge against over-reading any particular five.

The held-out split. One split serves two purposes: a basis fused from biomedical, source code and finance, evaluated on legal and math, is scored by the gate and projected out of the stream in §4.4. Finance's corpus is fully consumed by calibration, so it cannot appear on the held-out side. The three-domain basis is smaller. The control ruling out basis size as an explanation of any held-out shortfall is in Appendix B.7.

Protocols. Sampling, packing and per-corpus budgets are in Appendix B.3. The two style controls bin the same documents into quintiles of mean words per sentence and of mean word length,

	Llama-2-7B	Llama-3.2-3B	Qwen3.5-9B	Mamba-2	OLMoE†
token mixing	dense attn	dense attn	hybrid linear	SSM (none)	dense attn
H / depth	4096 / 32	3072 / 28	4096 / 32	2560 / 64	2048 / 16
compact: K_{joint}	65–97	50–90	37–79	85–103	101–115
single-tiered: σ (%)	5.4–7.1	6.2–8.8	5.3–6.9	5.1–7.0	4.0–6.6
learned: slope at K=32	+8.9	+11.2	+11.9	+17.5	–
× ambient random floor	2.42	2.46	1.69	3.90	2.47
× in-pool random floor	1.69	1.73	1.63	1.49	1.54
Fisher S/N per direction	0.363	0.333	0.354	0.372	0.307
overwrite: cos recent / sum	.96 / .86	.96 / .85	.99 / .78	.96 / .85	.95 / .87

Table 1: The signature of the kind. One column per family at the five-domain calibration, $k_{\text{each}} = 128$, $\theta = 0.97$. Each property is established in the section named by its italic label. K_{joint} and σ are ranges over the second half of depth. Every slope's CI excludes zero. The floor ratio and per-direction value are at own rank, the only form available for the mixture-of-experts and the random-initialization arms. The overwrite is nearest-to-recent on 20/20 ordered pairs, read at each model's final block except Qwen3.5-9B, at L27. †OLMoE mixes tokens with dense attention and varies the feed-forward instead, captured at its router input one sub-layer earlier (§3.3): never merged into cross-family ranges, and its floor uses the same capture point

the second near-orthogonal to the first while remaining surface-visible and frequency-matched (Appendix B.4). The language arm uses five non-English mC4 configurations (Xue et al., 2021) against the same English C4 reference, holding crawl, register and document filter constant so that language is the only thing that varies — which makes it a concept control rather than a corpus swap. Probe accuracies come from a multinomial logistic regression with a document-grouped split, the grouping load-bearing because a token-level split is spuriously high on the null arm (Appendix B.5). The rank sweep of §4.2 runs thirteen ranks at nine depths on a basis rebuilt without the energy cut (Appendix B.6).

4 Experiments

Four questions organize what follows: what the construction returns (§4.1), whether it is produced by training rather than by the construction (§4.2), what the model's predictions depend on (§4.3), and how far the code extends (§4.4). Table 1 collects the properties that recur across families, and is what we mean by "the same kind."

4.1 What the construction returns

A per-layer GEV returns a basis from anything, so the first question is not whether a discriminative subspace exists but what shape it has. Three properties constrain everything downstream.

It is compact. Over the second half of depth V_{joint} occupies 37–103 directions out of thousands, 0.9–4.0% of the hidden dimension, in all four block-output families and at every layer (Table 1). The mixture-of-experts, captured one sub-layer earlier at its router input, gives 101–115 directions on a narrower stream, 4.9–5.6%, quoted separately since two capture points cannot be merged into one range. The percentage should be read against the construction as well as against H, since each per-domain GEV sees at most $k = 256$ candidate directions, and K_{joint} averages 32% of that ceiling on Llama-2-7B and 28% on Llama-3.2-3B. What licenses calling the result compact is the behavior of that share when the ceiling moves: doubling k_{each} grows K_{joint} by only 1.34–1.43×, so the share *falls*, to 21% and 20% of the enlarged pool (Appendix A.2). A budget set by the cap would hold its share, while one set by the data gives it up. Across the factor of two in stream width the five families span, K_{joint} stays roughly constant in *absolute* terms rather than scaling with H, so what sets the budget is the number of calibrated domains and not the width. Extending the pool shows how: K_{joint} grows by roughly ten directions per added domain from $n = 5$ to $n = 20$, on a mutually distinct pool and an overlapping one alike and on two families, and past twenty the marginal cost falls — not as an artifact of near-duplicate domains, since a pool of thirty mutually distinct MMLU subjects declines at least as steeply as the overlapping 62-domain pool at the same size (Appendix C). Growth is sublinear rather than saturating: at thirty distinct domains the subspace still occupies 282 of 4096 directions with no plateau. Compactness is therefore a claim about a five-domain calibration, not a general one.

It is single-tiered. If a second discriminative tier survived below the energy cut, the threshold rather than the data would be deciding what the object is. Projecting V_{joint} out of the per-domain bases and taking an SVD of the remainder, the strongest direction left behind carries under 9% of V_{joint}'s

leading singular value in every family over the second half of depth (Table 1), and the complement is dominated by directions all five domains use alike. The verdict rides on neither the threshold nor the calibration size (Appendix A.2).

It is second-moment, and this is where the obvious alternative fails. The covariances are centered, so V_{joint} is defined on fluctuation structure, and the competing account is first-moment: that V_{joint} is simply where the class means moved. The class-mean shift does overlap V_{joint} far above chance (21–56% of its unit energy across the second half of Llama-2-7B's depth, 13–31$\times$ the K/H share), but a matched-rank PCA subspace captures 97–99% of it: the mean shift is a high-variance object any top-variance subspace recovers, so we rest nothing on it, and the object characterized from here is the second-moment code.

V_{joint} is also a per-layer object rather than one subspace propagating from input to output: median containment of each basis in its successor runs 22–46$\times$ the K/H chance share, but the first-layer and final-block bases fall back to chance once the rotation between them is removed (Appendix L).

So the construction returns a compact, single-tiered, second-moment code. Whether it is learned, rather than produced by the construction, is the next question.

4.2 THE CODE IS LEARNED

The instrument is Fisher S/N resolved per layer, and the question is not circular because the GEV is solved independently at each layer: V_{joint} is by construction the most discriminative few-percent subspace *at that layer*, while nothing constrains how that per-layer-optimal discriminability behaves *across* layers. It can rise, stay flat, or fall. At fixed rank it rises steadily in every trained family — not layer by layer, but with a trend no interval crosses — from 2.9 to 14.6 on Llama-2-7B, 2.0 to 14.4 on Llama-3.2-3B, 2.1 to 15.1 on Qwen3.5-9B and 0.8 to 18.5 on the attention-free Mamba-2 (Figure 1), giving depth-slopes of $+8.9$ to $+17.5$ with bootstrap CIs excluding zero. At own K_{joint} the slopes are 42–70% steeper, which is the rank dependence of §3.2 rather than a second phenomenon. The mixture-of-experts reproduces the gate at its router input, rising at own rank from 4.1 to 35.3, slope $+22.2$ $[21.6, 23.1]$. Against the in-pool floor at that capture point it reads 1.54$\times$, inside the 1.49–1.73$\times$ band the other four give.

The training axis. If the rise were a by-product of the construction or of the embedding table, it would survive the removal of training. It reverses instead. On two random-initialization arms of the Llama-3.2-3B architecture, five seeds each, discriminability *decays* with depth: all four trained block-output families are positive and all ten control seeds negative, every CI excluding zero, at -6.3 to -7.2 fully random and -9.4 to -11.6 trained-embedding. The contrast survives both rank-controlled readings — at fixed rank the seeds give -4.6 to -3.9 and -8.0 to -5.4, and per direction the trained families rise from 0.017–0.070 to 0.333–0.372 while both control arms fall. The trained-embedding arm carries the most interpretive weight, and behaves oppositely to what either deflationary reading predicts: it starts *highest* of all, at a five-seed mean of 24.7, the trained embeddings alone clustering domain vocabulary, and decays to 13.7 where the trained model reaches 30. So V_{joint} is not the embedding table, present at every layer of that control, nor a per-layer probe succeeding on a fixed representation, since the representation is held fixed-random and the probe's reach still falls. The trained model starts *below* that control and overtakes it: the claim is the slope, not the intercept. Nor is the rise an effect of post-training: Qwen3.5-9B-Base gives $+19.2$ $[18.7, 20.0]$ at own rank against its post-trained sibling's $+18.5$ $[18.1, 19.2]$, the slopes matching while the base model's final level runs 15% lower.

A floor for the basis. Those controls vary the model. A matched-rank random subspace varies the basis instead, on the same classes, and the result generalizes past this construction: an *arbitrary* subspace of equal rank also has a rising gate, and it rises the more steeply the closer the draw is confined to the region V_{joint} was selected from — $+5.0$ to $+9.2$ drawn ambiently, $+6.6$ to $+17.2$ drawn from the amplified candidates themselves. A depth-wise separability trend is therefore not by itself evidence of learned structure, whatever subspace it is measured in: the in-pool floor on Mamba-2 rises at $+17.2$, steeper than Llama-2-7B's own-rank V_{joint} slope.

V_{joint} clears that floor in all five families, at 1.49–1.73$\times$ its final level — the most conservative of the four rank-by-draw cells — and the margin is near-constant where the ambient one is not: drawn ambiently the ratio spans a factor of 2.31 and Mamba-2 looks like a standout at 3.90$\times$, while confined to the candidate pool the five collapse into a factor of 1.16 and Mamba-2 has the *smallest* margin. That margin is modest in absolute terms and we do not present it otherwise; what recommends it is its constancy, and the fact that its denominator is the hardest of the four draws — an arbitrary selection

from the fusion's own candidates rather than from the stream. Near-constancy holds at fixed rank too (Appendix A.4), and is one face of a broader convergence: the families reach a common per-direction level of 0.31–0.37, and at fixed rank their final levels span 1.29× where matched-rank random subspaces span 2.19×, so what converges belongs to the code rather than to the stream (Appendix E). A floor and a control are not interchangeable, and Figure 1 shows why: by the final block the random floor of the trained families rises above a V_{joint} basis in a fully random model, so an arbitrary subspace of a trained stream separates the domains better than an optimized subspace of a random one.

The concept axis. A deflationary reading survives the training axis: trained networks make most text properties more linearly separable with depth, so perhaps any partition of text would acquire a code. Two controls fix the model and vary the classes instead. An arbitrary content-blind partition of the generic corpus and two matched-frequency style-quintile partitions all return a basis, as an SVD always will, but discriminability along it sits at a floor at every depth and every rank, a factor of 11–15 below the real domains' *level* at the final block (per-arm slopes in Appendix B.4). What the controls do share with the real domains is proportional growth: their fold across depth outgrows the real domains' on three of the four block-output families, which manage 5.1–7.2× (dense and hybrid). Only on the state-space model do the real domains outgrow them and we rest nothing on that. The conservative reading, which we adopt for every family, is that the concept axis is a result about level. The developmental content of the gate rests on the training axis alone. Decodability is not the code either, and the two come apart in both directions: the style quintiles are read at only 0.50–0.59 late-layer probe accuracy against 0.20 chance, while the real domains are already at 0.95 at the *first* layer, where their Fisher S/N is 2.0. Separability saturates almost immediately, and the maintained second-moment code is what training builds on top of it. No whole-space measure reproduces the trend, and one runs counter to it: whole-space Fisher on the raw activations falls two orders of magnitude into mid-network before recovering at the readout, over the same depth range as the compression valley of Queipo-de-Llano et al. (2026) — so V_{joint} gains discriminability where whole-space separability collapses (Appendix F).

A concept that is given rather than constructed. Those controls are floors, and a floor cannot show that the gate is about *domain* rather than about any salient distributional property. Language identity is the test: five non-English mC4 configurations against the same English reference, more extreme than any of our domains on the raw spectrum and nearly as probe-decodable at the first layer (Appendix D). Language does acquire a code, so the format is not reserved for topical domains. But it acquires it differently. In all four block-output families it starts 2.1–8.8× more discriminable than the domains and grows far less across depth: at the fixed rank of §3.2, where neither concept's dimensionality enters, language's fold-change on Llama-3.2-3B is 2.9× against domain's 7.1×, and at each concept's own rank the same ordering holds in four of four families, at 2.4–6.9× against 9.1–39.5×, and the level ordering holds in all 117 cells of the rank-by-depth grid (Appendix D). The additive measure does not favor domain, and we report it too: language's fixed-rank slope on Llama-3.2-3B is +13.6 against domain's +11.2. We read folds because a slope scales with level and language starts higher. Level is the right instrument for the floors above, not here. Language's code contracts as it rises where the domains' broadens — 89 directions at the first layer to 32 at the last, against 41 to 90 (Appendix D). Language identity arrives at the embedding and is maintained; domain identity arrives weak and is built.

So the code is learned: training, not the construction, makes it rise with depth, and it clears a floor drawn from its own candidates in every family. That is still correlational: it shows the code is built and maintained, not that anything uses it.

4.3 What the predictions depend on

Selective necessity. At an anchor layer we project out of the stream, in turn, the full V_{joint} basis, a matched-rank PCA subspace, and matched-rank random orthonormal subspaces, and measure per-corpus loss change against the unmodified model, on both dense families at first, mid and late anchors. Removing V_{joint} raises loss on the calibration domains by +0.16–0.67 nats against +0.0006–0.019 on generic text, with every calibration domain above both its own five-draw random floor and the generic damage at the mid and late anchors. We quote the ratio per domain rather than averaged, since averaging is dominated by the most affected corpus: at those anchors even the *least* affected corpus takes 17–98× the generic damage, on both dense families. Generic here is the C4 reference the basis was built against. Scored instead against wikitext, which it never saw, the same least-affected ratio falls to 4.1–6.2× — part of what is measured is distance from the reference, as the operational reading of §1 predicts. Removing the PCA subspace is far more destructive absolutely, 3–8 nats, but indiscriminately,

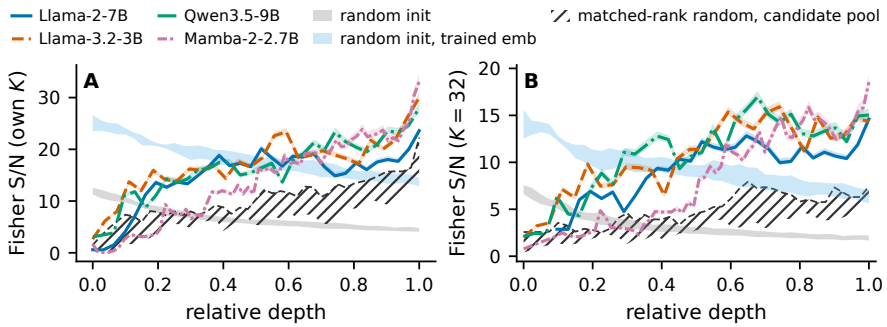

Figure 1: The depth gate, with both floors. *A:* at each layer's own K_{joint}. The four trained block-output families rise, both random-initialization arms decay (five seeds, envelopes). *B:* the same gate at fixed rank $K = 32$. Hatching: the matched-rank random floor drawn from the fusion's own candidates (candidate pool), which reaches 58–67% of V_{joint}'s final level at own rank and 39–51% at fixed rank. Bands: sequence-level bootstrap 95% CIs (trained) and across-seed envelopes (controls).

at a domain-to-generic ratio of 0.99–1.33. Random subspaces do essentially nothing. The split is the finding: the *magnitude* of ablation damage is generic to high-variance subspaces, its *selectivity* is not. Consistent with §4.2, at the first anchor selectivity is concentrated in source code and math and broadens across all five domains by the mid and late anchors, so it is strongest once the code has formed (per-domain figures in Appendix G, and the same contrast without intervening in Appendix H). The selectivity also survives a sharper contrast. Domain-versus-generic is close to definitional, since V_{joint} is built where domain variance exceeds generic variance, while domain-versus-*other-domain* is not. A basis fused from biomedical text alone on Llama-3.2-3B damages biomedical 12.9–41.1× more than the other four domains across the three anchors — one basis rather than a full matrix, so a check on the direction of the effect rather than a characterization (Appendix G). The pattern also holds at the router input of the mixture-of-experts, at 5.9–9.3× domain over generic against its matched-rank PCA control's 1.0–1.5×, weaker than the dense families but in the same direction and against the same control.

Nothing we can measure reads it. Selective necessity says predictions depend on the code being present, not that a component consults it. Adding the domain-d code to generic text at a mid layer moves the late-layer V_{joint} coordinate, an internal value rather than a logit, 6–12% of the way to its domain value on Llama-2-7B and 3–9% on Qwen3.5-9B, against ≈0 for a matched-norm random direction. But that null is too weak, since a V_{joint} direction carries more variance than a random one, and the matched-rank PCA subspace conditions its own late coordinate as strongly or more: across Llama-2-7B, Qwen3.5-9B and Mamba-2 its move-fraction is 0.18–0.40 against V_{joint}'s 0.03–0.19. What propagates is largely the class-mean shift of §4.1, carried by the stream's general dynamics, so **we make no read claim**. The test is first-moment by construction, so the failure is structural rather than a matter of a better test (Appendix I). Nor do we claim steering: amplifying a causal axis lowers its own domain's loss for a minority of axes, and for almost none at the final block (Appendix I).

So predictions on the corpora it was built from depend on the code as a substrate, and nothing we can measure reads it as a signal. How far the code extends — to other corpora, across a context, across models — is the last question.

4.4 How far the code extends

To corpora it was not built from. Fusing a basis from three domains on Llama-3.2-3B and scoring it on the two it never saw, the gate survives at $+12.1$ $[11.4, 12.9]$ — steeper, in fact, than the same basis scored on its own three domains, $+10.2$ $[9.7, 10.8]$ — so the format is as strongly expressed on unseen corpora as on calibrated ones. Ablation does not survive: under that same basis the held-out domains retain only 2–13% of the damage they took under the five-domain one, and the pair contains math, the second most damaged corpus there. The discriminative *format* generalizes beyond its own calibration; the causal *necessity* does not. Nor is that an artifact of a three-domain basis. Sweeping the calibration from one domain to fifty on Llama-2-7B and ablating against corpora never in any pool, absolute held-out damage rises 26-fold — but so does K_{joint}, from 15 to 478 directions, and against the matched-rank random floor that controls for it the held-out effect sits at 4.4–4.9× from $n = 5$ onward and does not trend (Figure 2C). Selectivity against generic text meanwhile collapses, from 35× at $n = 5$ to 9.8× at $n = 50$. Adding domains buys a larger, progressively less selective

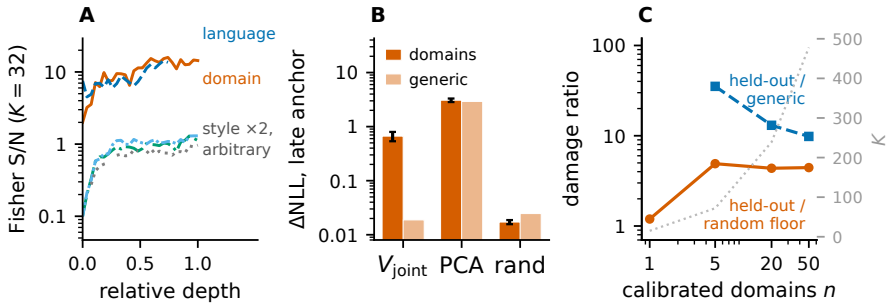

Figure 2: Concept specificity and the scope of necessity. All panels are log-scale. *A:* the concept axis at fixed rank on Llama-3.2-3B: language starts high and grows little, domain starts low and grows an order of magnitude, pseudo-concepts stay at a floor (language's trace breaks at L21–L26, Appendix A.5). *B:* ΔNLL at the late anchor: PCA does more damage indiscriminately, while V_{joint} does less, selectively. *C:* on Llama-2-7B sweeping the calibration from 1 to 50 domains, held-out damage over the matched-rank floor is flat from $n = 5$ while selectivity over generic text collapses ($n = 1$ omitted, its denominator ≈0). The dotted trace is K_{joint}, read on the right axis.

subspace drifting toward the indiscriminate high-variance regime PCA occupies (§4.3), not a code that transfers. For ablation purposes V_{joint} is a union of per-corpus codes rather than a register whose removal costs the model its sense of domain in general — the sharpest statement available of what the operational reading of §1 amounts to.

Across a context. When two domains share a context the code tracks the active one rather than accumulating: cosine to the recent domain's code is 0.95–0.99 and nearest on all 100 ordered pairs across the five families, against 0.78–0.87 to the vector sum. The *ordering* is generic to the stream while the *fidelity* is not — the division the ablation also draws (§4.3) — and the new domain writes in within 16–32 tokens, identically on dense attention and the state-space model (Appendix J).

Across models. It does not. Transporting the basis between families with a ridge map fitted on generic text alone, over five model pairs, the transported axes keep no selectivity at all: damage to their own domain over the other domains sits at 0.76–1.54×, which is chance, and over generic text at 0.29–0.66×, so a transported axis costs generic text more than the domain it was labeled with. Orthogonal Procrustes does worse, at −0.07–0.31×. The instrument is not blind to a transport that works: applied within a single model, where the fitted map is near-identity, it preserves 7.8× own-over-other and 23.7× own-over-generic, against 1.7–92.8× for the target model's own axes (Appendix K). The negative is therefore about these maps, not about the object; whether a non-linear one succeeds is open.

So the code's format extends to the corpora we held out, and follows the active domain within a context, but its necessity stays with the corpora it was built from, and no linear map we fit carries it between models.

5 Discussion and Conclusion

Every instrument here is paired with a control that could remove the claim it supports, and several do: Appendix N tabulates ten claims this object initially appears to support and what each control leaves of them — four withdrawn, four narrowed, two standing.

Coverage and limitations. OLMoE contributes its gate at own rank only, one sub-layer earlier than the other four. Domain-count scaling is measured on two families, and the mutually distinct pool stops at $n = 30$ where the overlapping one reaches 62, so the falloff in marginal cost beyond thirty domains is established only where near-duplicate subjects are present. The single-domain specificity check of §4.3 uses one basis rather than a full per-domain matrix, built through a single run of a randomized SVD. And the rebuilt random-initialization seeds are taken to be faithful to the architecture on an identical code path and the absence of across-seed outliers, not separately verified (Appendix M).

Conclusion. Trained language models build a compact discriminative code across depth, and what that code is *for* is narrower than what it is. Domain identity, in the operational sense this paper can support, is recorded rather than consulted, and recorded per corpus rather than in a register. Both negatives bound something: a code no linear map transports limits what direction-level alignment between models can carry, and a per-corpus necessity limits what removing a domain subspace can remove.

AI USE STATEMENT

The research question, the construction of V_{joint}, and the claims made and withdrawn in this paper are the authors'. Generative AI was used as an assistant at the points described below. The authors take responsibility for the final content of this work, including all text, claims and artifacts produced with its aid.

Uses requiring disclosure. AI assistance was used adversarially, to stress-test claims the authors had already formed: given a result, it was asked what would have to be true for that result to be an artifact of the construction, and what control would show it. Several of the falsification conditions and matched controls in this paper were proposed or refined in that process, and several of the withdrawn claims in Appendix N are its outcome. It was also used for cross-checking the numbers reported here against the saved run artifacts.

Uses where disclosure is recommended. AI assistance was used to orient the authors in literatures outside their expertise and to identify related work, for brainstorming, and for drafting and rewriting passages of the manuscript, which the authors then edited. Grammar and style checking additionally used Grammarly. No dataset or numerical result reported here was generated by a generative-AI assistant. The language models studied in this paper are the object of the experiments, not a tool used to produce them.

ETHICS STATEMENT

This work analyses publicly released model checkpoints on publicly available text corpora under their respective licences. No new data is collected or released, and no human subjects are involved. The language-identity control (§4.2) contrasts five non-English mC4 configurations against an English reference. It is used only to characterise the format in which a distributional property is represented, and we make no claim about model quality, capability or appropriateness in any of those languages. The biomedical and legal corpora are used solely as distributional samples. Nothing reported here is intended or validated for clinical or legal use. The authors have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All experiments run on publicly released checkpoints (§3.3). Appendix A specifies the construction, the five controls, the fixed-rank layer filtering and the bootstrap in full. Appendix B gives the corpora, the sampling and packing budgets, the style and language-control construction, and the probe methodology. The remaining appendices report the per-experiment details behind every figure and table, including the claims ledger (Appendix N). Anonymized code is provided as supplementary material. It covers every experiment in the paper and ships the numeric output of every run we quote, so both figures and every quoted table regenerate without a GPU. It also includes a verification harness that re-runs the CPU-derivable stages of the pipeline against those saved outputs. The model-forward experiments themselves require a GPU. Scope limits on individual results — the in-sample evaluation text of the ablations, the single construction seed, and the one run excluded on provenance grounds — are stated where those results are reported.

REFERENCES

Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7747–7763, 2020. URL https://aclanthology.org/2020.acl-main.692/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings F (Radar and Signal Processing)*, 140(6):362–370, 1993.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4310–4330, 2022. URL https://arxiv.org/abs/2110.00976. arXiv:2110.00976.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168. The GSM8K dataset.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023. URL https://arxiv.org/abs/2309.08600. Published at ICLR 2024.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2405.21060. Mamba-2. arXiv:2405.21060.

Agnibh Dasgupta, Abdullah Tanvir, and Xin Zhong. Invariant features in language models: Geometric characterization and model attribution. *arXiv preprint arXiv:2605.06458*, 2026. URL https://arxiv.org/abs/2605.06458.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Bernhard Flury. *Common Principal Components and Related Multivariate Models*. Wiley, New York, 1988.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

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

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2026. URL https://arxiv.org/abs/2501.17727. Revised version of *Sparse Autoencoders Can Interpret Randomly Initialized Transformers*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300. arXiv:2009.03300.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. URL https://aclanthology.org/D19-1275/.

Youcheng Huang, Chen Huang, Duanyu Feng, Wenqiang Lei, and Jiancheng Lv. Cross-model transferability among large language models on the platonic representations of concepts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3686–3704, 2025. URL https://aclanthology.org/2025.acl-long.185/. arXiv:2501.02009.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2405.07987. arXiv:2405.07987.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://arxiv.org/abs/1909.06146. arXiv:1909.06146.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.05907. Spotlight. arXiv:2409.05907.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1609.07843. The WikiText corpora. arXiv:1609.07843.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.02060.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03658. arXiv:2311.03658, November 2023.

Bruno Puri, Jim Berend, Sebastian Lapuschkin, and Wojciech Samek. Atlas-Alignment: Making interpretability transferable across language models. *arXiv preprint arXiv:2510.27413*, 2025. URL https://arxiv.org/abs/2510.27413.

Enrique Queipo-de-Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.06477.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://arxiv.org/abs/1910.10683. The C4 corpus. arXiv:1910.10683.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2402.17762. arXiv:2402.17762.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4593–4601, 2019. URL https://aclanthology.org/P19-1452/.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020. URL https://aclanthology.org/2020.emnlp-main.14/.

Dianyun Wang, Qingsen Ma, Yuhu Shang, Zhifeng Lu, Zhenbo Xu, Lechen Ning, Huijia Wu, and Zhaofeng He. Interpretable safety alignment via SAE-constructed low-rank subspace adaptation. *arXiv preprint arXiv:2512.23260*, 2025. URL https://arxiv.org/abs/2512.23260.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. URL https://arxiv.org/abs/2010.11934. The mC4 corpus used for the language arm. arXiv:2010.11934.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

A METHOD DETAILS

A.1 DISCRIMINANT VERSUS PROBE

A supervised probe would also find directions that separate domain text from generic text, and would probably find them more efficiently. We use a generalized eigenproblem instead for three reasons, each of which matters somewhere later in the paper.

Nothing is fitted to our labels. The GEV is a function of two covariance matrices and of nothing else. The labels enter only in deciding which tokens go into which covariance, and no parameter is chosen to make the classes separable. A probe's weights, by contrast, are selected to maximize a discriminative objective on data we labelled, which means a probe's depth profile confounds what the representation maintains with how much signal a classifier can be trained to extract from it. §4.2 shows those two come apart in this setting: the real domains are decodable at 0.95 from the first layer, where the second-moment code they are read from scores 2.0 and has most of its growth ahead of it. A probe would have reported that the work was already done at layer one.

The basis is ranked, and the ranking is meaningful. Each eigenvector arrives with an eigenvalue that is a variance ratio, so directions order by how much more strongly the domain uses them than generic text does, and $\lambda = 1$ marks a principled cut between amplified and shared directions. That ordering is what the fusion weights (§3.1), what the rank sweep of Appendix B.6 truncates, and what makes "the leading directions" a well-defined object in Appendix B.2. A probe returns a weight vector per class with no comparable internal order.

It is symmetric in the two distributions. The construction treats domain and generic covariances as the two sides of one ratio, so "domain identity" comes out contrastive by definition rather than as a class label — the operational reading §1 commits to. A classifier trained on five domain labels would instead return whatever separates those five from each other, which is a different object and one we could not have compared against a single generic reference.

The cost of the choice is that a discriminant optimizes no downstream objective, so nothing guarantees the directions it returns are the ones the model uses. That is precisely why every claim in §4.3 is made against a matched-rank variance-only control rather than against the construction alone.

A.2 WORKING SUBSPACE, k_{EACH}, AND THE ENERGY CRITERION

The working-subspace rank is $k = 2k_{\text{each}}$, where k_{each} is the parameter the released runner exposes: half the working-subspace size, since the top k principal components are taken from the pooled activations rather than k_{each} from each side (Appendix B.3). The paper's default is $k_{\text{each}} = 128$, hence $k = 256$, and doubling k_{each} to 256 gives $k = 512$.

Projecting V_{joint} out of the per-domain bases and taking an SVD of the remainder, the strongest direction left behind carries 5.1–8.8% of V_{joint}'s leading singular value over the mid-to-late layers of the block-output families and 5.0–6.6% on the mixture-of-experts. At $\theta = 0.90$ the per-direction convergence level rises to 0.409–0.507 while the spread across families stays tight, so what converges is a property of the spectrum rather than of where we cut it. The single-tiered verdict is likewise threshold- and calibration-robust: at $\theta = 0.90$, which hands the complement the excluded energy band, the leading complement direction still sits at $\leq 23\%$ of V_{joint}'s at every layer, and at the 62-domain calibration it falls to 1.6–5.6% even as the complement grows. At the first layer 124 of the 256 working directions fall below $\lambda = 1$.

K_{joint} is reported at the fixed criterion $\theta = 0.97$ throughout, so its dependence on the working-subspace ceiling has to be measured rather than assumed. Doubling k_{each} from 128 to 256 on both dense families grows K_{joint} by $1.336\times$ on Llama-2-7B (mean $81.4 \rightarrow 108.8$ directions) and $1.433\times$ on Llama-3.2-3B ($71.2 \rightarrow 102.0$), well short of the doubling in available candidates. Each share is quoted against its own run's pool: 31.8% and 27.8% of $k = 256$ become 21.2% and 19.9% of $k = 512$. The enlarged basis preserves the depth profile and recovers the same subspace: on Llama-2-7B over L9–L31, top principal cosine 0.97–0.99 with 83–91% of the smaller basis's energy contained in the larger, and on Llama-3.2-3B 0.97–0.99 and 81–90%. It also leaves the gate of §4.2 positive and comparable, moving from $+15.9$ to $+17.5$ on Llama-3.2-3B and from $+15.1$ to $+18.5$ on Llama-2-7B at own rank.

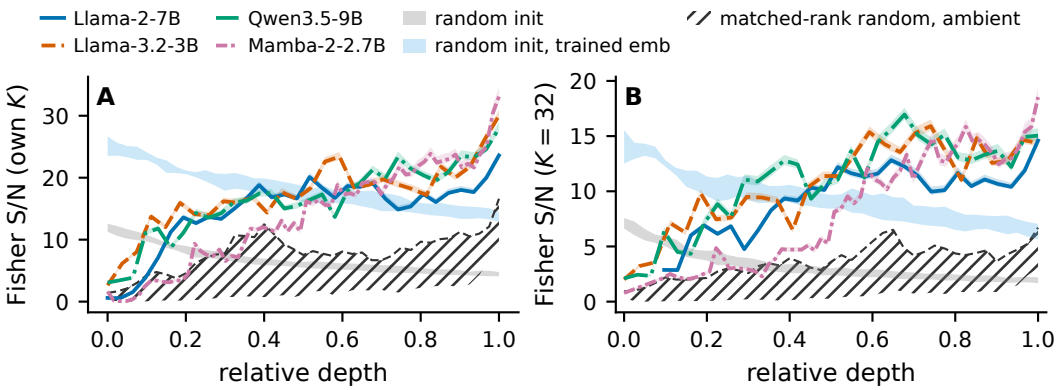

Figure 3: The depth gate against the ambient floor. As Figure 1, with hatching showing the matched-rank random floor drawn ambiently in $\mathbb{R}^H$ rather than from the fusion's candidates. The trained traces are identical to Figure 1, only the floor changes.

A.3 Capture points and the OLMoE pre-hook

For the four block-output families a forward hook on each decoder block returns the block's own output — the stream after both residual additions, $x^{(l+1)}$ in the notation of §3.1. Every layer is captured, indexed from zero: 32 for Llama-2-7B, 28 for Llama-3.2-3B, 32 for Qwen3.5-9B, 64 for Mamba-2. Mamba-2 carries its residual in float32 by configuration, so its block output arrives at higher precision than the bfloat16 the weights are loaded in, and is downcast with everything else at storage.

OLMoE is different in a way that goes beyond "one sub-layer earlier". The hook is a forward *pre*-hook on the sparse expert block, and what it captures is that block's input — which the surrounding decoder layer has already passed through the post-attention RMSNorm. The captured tensor is therefore $\mathrm{norm}(x + \mathrm{Attn}(x))$: post-attention, post-normalization, and exactly the tensor the router reads. Two consequences follow, and they are why the mixture-of-experts is quoted separately everywhere in this paper. The normalization has removed the per-token scale, so this space is not norm-comparable with the four block-output families, and the capture sits inside the block rather than at its exit, so a dimensionality or a discriminability measured there cannot be pooled into a cross-family range. Its matched-rank floors are drawn at the same capture point, so the comparisons *within* OLMoE remain like-for-like.

Models are loaded in bfloat16 (Mamba-2 with the eager attention path, the transformers with SDPA). Activations are stored on CPU in float16 during construction, the eigenproblem is solved in float32, and the gate's sufficient statistics are accumulated in float64.

A.4 Random-floor draws

We draw matched-rank random subspaces three ways, in decreasing generosity: uniformly in $\mathbb{R}^H$, within the union of the per-domain bases the fusion is given, and within the span of just the amplified columns among those, the candidates the fusion actually chose from. Because GEV eigenvectors are orthogonal in the $(\Sigma_g + \epsilon I)$ metric rather than the standard one, orthonormalizing the candidate columns preserves the span they define but not the $w_{d,j}$ weighting that made them candidates, so the in-pool draw is a tighter span rather than a sampler matched to the fusion. At fixed rank the in-pool margin is 1.97–2.54×, a spread of 1.29, against the ambient draw's 2.24–6.06× and 2.70. The own-rank in-pool figure quoted in Table 1 is the most conservative of those four cells, since at own rank V_{joint} carries its weak tail. Figure 3 repeats Figure 1 with the ambient floor in place of the in-pool one: the floor sits lower and rises more shallowly, which is what makes the in-pool draw the demanding comparison and the ambient margins (Table 1) the more flattering ones.

A.5 Fixed-rank layer filtering

Truncating a basis to a fixed rank is only meaningful where the basis has that rank to give, so at $K = 32$ any layer whose K_{joint} falls below 32 is dropped from that arm rather than padded. The

rule is applied per arm and nowhere else in the paper; own-rank and per-direction reporting use every layer.

Which layers this removes follows from each arm's K_{joint} profile, and the arms differ:

arm	K_{joint} range	layers kept	dropped
Llama-2-7B	6–116	29 of 32	L0–L2
Llama-3.2-3B	41–94	28 of 28	none
Qwen3.5-9B	37–79	32 of 32	none
Mamba-2	6–109	59 of 64	L1–L5
random init, both arms	–	28 of 28	none
arbitrary partition	25–66	25 of 28	L1, L7, L14
style, both arms	34–91	28 of 28	none
language identity	8–120	22 of 28	L21–L26

Two of these bear on readings elsewhere. The dense and hybrid families lose only the first few blocks, where K_{joint} has not yet grown, which is why the Llama-2-7B trace in Figure 1B begins at relative depth 0.1 rather than at zero, and the same is true of Mamba-2's early blocks. Language identity is the one arm that loses a *late* span, because its code contracts as it rises (§4.2) and dips to eight directions at L24 — so the domain-versus-language comparison at fixed rank spans L0–L20 plus the final block, with a gap in between, and Figure 2A breaks its trace there rather than interpolating across it. No arm is dropped at a layer that carries a quoted number.

A.6 Rank-free alternatives

A minimum-description-length probe (Voita & Titov, 2020) would give a rank-free discriminability measure and so avoid the monotonicity in K that forces the three-way reporting of §3.2. We did not use one, for two reasons. First, MDL scores a *probe* fitted to our labels, which reintroduces exactly the label-fitting the discriminant was chosen to avoid (Appendix A.1): a codelength curve across depth would then confound what the representation maintains with what a classifier can be trained to extract from it, and §4.2 shows those come apart — the real domains are decodable at 0.95 from the first layer, where their Fisher S/N is 2.0. Second, MDL is defined on the classification of a label set, so it has no natural form for the matched-rank random and PCA floors, which are alternative *bases* rather than alternative label sets. The floors are the comparison that carries §4.2, and we preferred one measure that all arms share to a better-behaved measure that only some of them admit. Fixed rank plus a per-direction value recovers rank-freeness within the measure we do use.

A.7 Sequence-level bootstrap

Every interval in the paper is a percentile bootstrap over *sequences*, not tokens, because tokens within a sequence are correlated and a token-level resample would understate the variance. The gate's evaluation pool is 128 sequences of 256 tokens per corpus, for both the generic reference and each domain. During the forward pass we accumulate, per sequence and per layer, only the sufficient statistics of the projected activations — $\sum_t y_t$, $\sum_t y_t^2$ and the token count, with $y = h^\top V_{\text{joint}}(l)$ — after which the model is discarded. A replicate draws, with replacement and at the original pool size, one index set for the generic pool and one *per domain*. The same draw is reused across all layers within a replicate, so each replicate yields an internally coherent depth curve. We use 1000 replicates and report the nearest-rank 2.5th and 97.5th percentiles, with no bias correction or acceleration. The seed is 0 throughout.

V_{joint} is held fixed at the full-data basis: it is loaded once, orthonormalized once by QR, and never re-fitted, so the intervals are conditional on the subspace and describe sampling noise in the evaluation text alone. Three end-to-end rebuilds of the Llama-3.2-3B basis under fresh construction seeds bound what that excludes: own-rank slope 14.98/15.51/15.77 against the recorded 15.95, fixed-rank slope 10.58/11.24/11.46 against 11.21, and K_{joint} at the final block 87/88/90 against 90. The fixed-rank seed spread reaches below the bootstrap interval's lower end (10.89), so basis-estimation noise is the larger of the two and the intervals should be read as conditional in the strict sense. The depth-slope interval comes from recomputing the ordinary least-squares slope of Fisher S/N on relative depth $l/(L-1)$ within each replicate's own curve and taking percentiles over the 1000 slopes. The slope we report is the point estimate on the full data, not the bootstrap median. The accompanying

`frac_positive` is the fraction of replicate slopes strictly greater than zero — a sign-consistency statistic, not a p-value. It is 1.0 for every trained family and 0.0 for every random-initialization arm.

As implemented, Fisher S/N at layer l is

$$\mathrm{S/N}(l) \;=\; \frac{1}{|D|}\sum_{d\in D}\sum_{i=1}^{K}\frac{\left(\mu_{d,i}-\mu_{g,i}\right)^2}{\frac{1}{2}\left(\sigma^2_{d,i}+\sigma^2_{g,i}\right)},$$

with token-level means and population variances of the projected activations. Because the two pools are balanced, the denominator is the pooled within-class variance, and the numerator is the squared class-mean difference, which differs from the textbook between-class variance of two balanced classes by a constant factor of four. That factor is common to every arm and every layer, so it cancels from every ratio, fold and slope comparison in the paper.

B Data and protocols

B.1 Domain corpora

Biomedical is PubMedQA (Jin et al., 2019): the abstract contexts and long-form answers of the `pqa_labeled` split. *Source code* is the Python portion of codeparrot-clean. *Finance* is the five finance-adjacent MMLU subjects (Hendrycks et al., 2021) — business ethics, econometrics, high-school macro- and microeconomics, professional accounting — rendered as multiple-choice prompts. *Legal* is the case-hold task of LexGLUE (Chalkidis et al., 2022): citing contexts paired with candidate holdings. *Math* is GSM8K (Cobbe et al., 2021), word problems with their worked solutions. Every corpus is concatenated and packed into fixed-length chunks with no padding: 512 tokens for construction and for the ablations, 256 for the gate and the forward-pass tests (Appendix B.3).

This is the register mix §3.3 refers to: running prose (biomedical), source code (code), multiple-choice prompts (finance, legal), worked word problems (math). Two properties of the mix are worth stating plainly. *Math* here is arithmetic word problems rather than mathematical notation or proofs, so it is a weaker probe of symbolic text than the label suggests. And two of the five share a multiple-choice template, which makes a format direction available to the construction. That confound is controlled directly: rebuilding the five-domain basis on Llama-2-7B with finance replaced by Reuters financial-news prose leaves the gate at $+13.8$ $[13.4, 14.5]$ against $+15.1$ for the multiple-choice version, so the domain code reproduces under a prose finance corpus and the template sharpens it rather than creating it.

B.2 Generic reference swap

The swap is exact in the sense that matters: one construction pass collects the per-domain eigendecompositions against *both* references, and the two bases are then fused from the same step-one artifacts under identical settings, differing only in which reference the domain covariances were divided by. Nothing else varies.

Comparing the two bases layer by layer, after orthonormalizing both, the cosine of the smallest principal angle runs 0.963–0.998 on Llama-2-7B and 0.961–0.987 on Llama-3.2-3B, and never falls below 0.95 at any layer of either. Mean squared principal cosine — the share of one basis contained in the other — runs 0.69–0.94 and 0.64–0.80, against width-matched random orthonormal frames at 29–405$\times$ (median 35) and 22–40$\times$ (median 29).

Two scopes. The swap was run on the two dense families only; no wikitext basis exists for the hybrid, the state-space model or the mixture-of-experts. And the measure is subspace-level: the leading directions survive, but a containment of 0.69–0.94 means that between roughly 6% and 30% of the basis energy does not. We therefore claim only that the swap leaves the leading directions in place, not that it leaves the subspace unchanged.

A second, deliberately harsher reference exists for Llama-2-7B, Qwen3.5-9B and Mamba-2, in which the generic corpus is a balanced mixture of the five domains rather than web text. It is not a second wikitext: because the target domain enters its own reference at roughly one-fifth weight, the contrast becomes domain-versus-domain-average, and the fused basis is much wider (K up to 269 against 116). Leading-direction agreement survives — top principal cosine 0.85–0.99 with medians 0.95–0.96 — while containment falls to medians of 6–14$\times$ the matched null. We use this arm for direction stability only, and none of the paper's reference-robustness claims rest on it.

B.3 SAMPLING, PACKING AND BUDGETS

Each corpus is joined into a single string, tokenized once, truncated to a whole multiple of the sequence length and reshaped. There is no padding and no per-document boundary token, so a chunk may span two documents. Construction uses 512-token chunks; the loader builds 256 of them per corpus, of which the first 64 are forwarded (batch size 4, sixteen batches), giving a pool of 32,768 token positions per corpus per layer. Two of the five domains yield fewer than 256 chunks at these settings, which is immaterial since only 64 are consumed.

Every token position is captured, then trimmed to a fixed budget per corpus per layer: 16,000 for Llama-2-7B and 8,000 for the other four families. The trim is worth describing precisely, because it is not a reservoir sample. After each batch the accumulated buffer is cut back to the budget by a uniform random selection over everything held so far, which means a vector's chance of surviving to the end falls with the number of trims it must pass. At the 8,000 budget the cap first binds at the fourth of sixteen batches and roughly half the retained vectors come from the last three. The data loader does not shuffle, so this is a bias toward the tail of each corpus in document order rather than a random one. The 16,000 budget on Llama-2-7B binds later and so samples the stream more evenly – a second, smaller respect in which that family is not identically configured to the others. The trim draws from the global generator, seeded at 0, so a run reproduces exactly.

The working subspace is the top $k = 256$ principal components of the pooled centered activations, generic and domain rows stacked and decomposed together rather than 128 from each side separately. The Tikhonov ridge is 10^{-3} of the mean generic variance.

Later stages have their own budgets and do not subsample. The gate uses 128 sequences of 256 tokens per corpus and consumes every token position, with no vector cap. The forward-pass tests of §4.3 and §4.4 use 96 sequences of 256 tokens. The ablations use 64 sequences of 512 tokens with five random draws for the floor. All of these run at seed 0.

B.4 STYLE-QUINTILE CONSTRUCTION

Per-arm depth-slopes at fixed rank: $+0.78$ for sentence length and $+0.83$ for word length, agreeing to within 0.06, against the arbitrary partition's $+0.56$. Fold-changes across depth: $15.4\times$ for the arbitrary partition, $11.1\times$ and $13.2\times$ for the two style quintiles, against 5.1–$7.2\times$ for the dense and hybrid families and $23.6\times$ for the state-space model.

B.5 PROBE METHODOLOGY

A token-level split returns a spurious 60–78% on the null arm; the document-grouped split removes it.

B.6 RANK SWEEP

The sweep evaluates $K \in \{1, 2, 4, 8, 16, 24, 32, 48, 64, 96, 128, 192, 256\}$ at layers $\{0, 4, 7, 11, 14, 18, 21, 24, 27\}$ of Llama-3.2-3B: thirteen ranks by nine depths, 117 cells. "Rebuilt without the energy cut" means the fusion is repeated exactly as in §3.1 — the same $\max(\lambda - 1, 0)$ column weighting, the same stacked SVD — but the basis is then the top K left singular vectors in singular-value order, rather than however many the $\theta = 0.97$ criterion happens to retain. The two are otherwise the same object, which is what makes rank the only variable.

Fisher S/N uses the definition of Appendix A.7, with two differences forced by the setting: moments are pooled over tokens rather than accumulated per sequence, capped at 8000 vectors per corpus per layer, and there is no bootstrap. Each cell also carries a matched-rank PCA basis (top principal components of the mean-centered pooled generic-plus-domain activations, so class-agnostic) and a matched-rank random orthonormal basis, both evaluated on the identical grid.

Two orderings are read off this grid. Language exceeds domain in *level* in all 117 cells: at $K = 256$ in mid-network the margin is 3% (36.2 against 35.2 at L14), and it is narrowest at L14, $K = 64$, where the two are within 0.4% (18.44 against 18.36). Domain exceeds language in *fold* — final over first layer at fixed K — at eleven of the thirteen ranks, failing only at $K = 1$ and $K = 2$. The $K = 1$ failure is not marginal: the single leading domain direction loses discriminability across depth (fold 0.77) where the leading language direction gains $10.7\times$. Both exceptions sit far below K_{joint}, which is 41 at the first layer and 90 at the last, so the fold ordering is a property of the subspace rather than of any one leading axis.

B.7 HELD-OUT SPLIT DETAILS

The three-domain basis is smaller than the five-domain one, $K_{\text{joint}} = 29/43/59$ against $41/65/90$. The calibration domains retain 0.82–1.01 of the damage they took under the five-domain basis, which rules out reading any held-out shortfall as an effect of basis size.

C DOMAIN-COUNT SCALING

At the five-domain point the sweep returns 61–72 directions, inside the 37–103 the five composite domains give, growing by 9.5–12.1 directions per added domain averaged over $n = 5$ to $n = 20$. On the pool of thirty mutually distinct MMLU subjects — one representative of each of the ten overlapping subject families plus every unaffiliated subject — the marginal cost declines from 9.75 directions per domain at $n = 20$ to 8.20 at $n = 25$ and 7.40 at $n = 30$, at least as steeply as the deliberately overlapping 62-domain pool at the same size, 8.39, which continues to 5.6 by $n = 60$.

D LANGUAGE IDENTITY

Leading λ at the final block: 289–835 for language against 69–219 for the domains across the four block-output families, with first-layer probe decodability 0.911 against 0.953. Language's K_{joint} dips to 8 at L24. The level ordering holds in all 117 rank-by-depth cells, and the fold ordering holds at eleven of thirteen ranks, failing only at $K = 1$ and $K = 2$, where the domain fold falls below one on a single degenerate direction.

E CONVERGENCE TO A COMMON LEVEL

At fixed rank the families do not merely rise in parallel, they arrive at the same place. From starting points spanning 0.8–2.9 they reach 14.6 $[14.3, 15.1]$, 14.4 $[14.1, 14.8]$, 15.1 $[14.6, 15.7]$ and 18.5 $[17.9, 19.4]$: the three dense and hybrid intervals pairwise overlap and only the state-space model's is disjoint, 26% higher. Per direction the same four end at 0.333–0.372, and the mixture-of-experts at 0.307, some 8% below that band — adjacent, at a different capture point. Across all five the common value is 0.31–0.37, still tight at $1.21\times$. This is convergence to a common *value* rather than to a common sign, and it belongs to the code rather than to the stream, since matched-rank random subspaces end at 3.1–6.7 for a $2.19\times$ spread against V_{joint}'s $1.29\times$ over the four. Nor is it an artifact of the energy criterion: at $\theta = 0.90$ the per-direction level rises while the cross-family spread stays tight (Appendix A.2).

F WHOLE-SPACE MEASURES

The whole-space statistic is $\operatorname{tr}(S_B)/\operatorname{tr}(S_W)$ on the raw activations in $\mathbb{R}^H$, with no projection and no scaling, where S_B and S_W are the class-count-weighted between-class and within-class scatter matrices. Two differences from the gate's Fisher S/N matter. It is a five-way statistic over the domains, with no generic pool, where the gate is two-class domain-versus-generic. And it is computed on unnormalized sums of squares rather than per-direction ratios, so it carries the first moment and the raw activation scale.

This statistic is not the one behind the compression valley of Queipo-de-Llano et al. (2026): theirs is a spectral entropy, ours a class-scatter ratio, so we claim co-location in depth, not identity. The within-sequence token variance that dominates our denominator is consistent with the massive-activation mechanism they establish.

Measured on Llama-3.2-3B at nine hundred thousand token positions — 200 sequences per domain, every eighth position, $H = 3072$ — it runs 6.99×10^{-3} at the first layer, 7.47×10^{-5} at L7, 2.08×10^{-4} at L14, 1.05×10^{-3} at L20 and 5.35×10^{-2} at the final block. The collapse from the first layer to the minimum is a factor of 94 and the recovery from minimum to readout a factor of 717.

Three scope limits. The minimum sits at L7, relative depth 0.26, so the valley is early-mid rather than mid-network, and only five depths were sampled: nothing here constrains layers 1–6 or 8–13. The statistic exists for Llama-3.2-3B only, no probe artifact was produced for the other four families. And it is pooling-dependent — the companion run that mean-pools each sequence before computing the same ratio is flat across depth (0.85 at the first layer, 1.17 at the last), because pooling removes the

within-sequence token variance that dominates $\mathrm{tr}(S_W)$. The collapse is therefore a statement about token-level scatter, and we quote it only in that form. The same U-shape, with its minimum also at L7, appears on all four concept-control arms.

G Ablation details

The three anchors are the first, middle and last of each model's processed layers, so they differ by family: L0/L16/L31 on Llama-2-7B, L0/L14/L27 on Llama-3.2-3B, L0/L7/L15 on OLMoE at its router input. Evaluation uses 64 sequences of 512 tokens per corpus and the random floor averages five draws.

One scope limit belongs here rather than in a footnote. The ablation's evaluation text is drawn from the head of each corpus, the same prefix the subspace was calibrated on, so it is in-sample. The released code exposes a document offset for this purpose and the shipped runner sets it, but the numbers reported here predate that flag and were computed without it. The comparison is nonetheless internally sound: V_{joint}, the matched-rank PCA subspace and the random floors are all scored on identical text, so any optimism from re-using the calibration prefix is common to the three arms and cancels from the domain-over-generic ratios that carry the section. What it does mean is that the absolute nat values are not out-of-sample.

At the first anchor, per-domain selectivity: source code and math at 195× and 50× on the 3B and 817× and 562× on the 7B, against 3.6–27× for the other three domains. The least affected corpus at the mid and late anchors is biomedical in five of six cells.

The gate itself does not depend on this. Re-evaluating the fixed basis on documents disjoint from calibration, on the four domains for which held-out documents exist, gives +14.3 $[13.8, 15.1]$ against +14.8 $[14.3, 15.5]$ in-sample on Llama-2-7B and +13.0 $[12.4, 13.8]$ against +15.3 $[14.8, 16.1]$ on Llama-3.2-3B. The arbitrary-partition control stays at its floor either way (+1.7 out-of-sample against +1.3). What is in-sample is the ablation's absolute nat values, not the depth-wise result.

H Fisher S/N versus PCA across rank

Over the full 256-direction working subspace V_{joint} carries 1.6–4.3× a matched-rank PCA subspace's domain Fisher S/N, rising with depth, at 4–565× less variance, falling with depth. The two profiles are opposite, so no single layer shows both extremes. The margin holds at every depth from $K \approx 96$ upward and reverses below, where a top-variance subspace of the same rank is the more discriminative one.

I Read-forward and steering

At the mid layer we add, to every token position of *generic* text, the vector $V_{\mathrm{joint}} V_{\mathrm{joint}}^{\top}(\mu_d - \mu_g)$: the domain-minus-generic mean shift projected onto the subspace and mapped back into the stream. At the late layer we read the token-averaged V_{joint} coefficient and ask how far it moved along the generic-to-domain axis, as a fraction of the full distance:

$$\text{move fraction} \;=\; \frac{\langle a_{\text{patched}} - a_g,\; u\rangle}{\|u\|^2}, \qquad u = a_d - a_g.$$

A value of 1 would mean the patch carried the late coordinate the whole way. Anchors are L14 → L27 on Llama-2-7B, L16 → L31 on Qwen3.5-9B and L32 → L63 on Mamba-2, with 96 sequences of 256 tokens per corpus.

Two controls. A random ambient direction rescaled to the same norm moves the late coordinate by essentially nothing, which is the comparison that first made the result look positive. The matched-rank PCA subspace, built from the covariance accumulated on the *same* forward passes and truncated to V_{joint}'s own K, is the comparison that matters, and V_{joint} loses it in fourteen of fifteen family-by-domain cells: PCA moves its own late coordinate 0.18–0.40 of the way where V_{joint} manages 0.03–0.19. The one exception is math on Mamba-2, at 1.07.

The failure is structural rather than a matter of a better test. Both halves of the experiment are first-moment objects: what is written is a difference of class means, and what is read is a mean coefficient projected on a difference of class means. No second-moment quantity enters either side, so

no version of this design can isolate a read of the second-moment code — which is the object the rest of the paper is about. Establishing whether any component consults it would require a circuit-level argument we do not attempt.

An early-to-mid variant, L0 → L14 on Llama-2-7B, is null across all five domains, so the effect such as it is belongs to the mid-to-late span.

Steering. The read-forward test asks whether anything downstream consults the code. A separate test asks the cruder question of whether pushing on it does anything useful. Taking one axis at a time we set $h' = h + \alpha\,(h^\top v)\,v$ with $\alpha = \pm 1$ at that axis's anchor layer, and measure the per-corpus loss change on 64 sequences of 512 tokens. The test is run only on axes that already carry a causal label, meaning ablating them moves their own corpus by at least three standard deviations of a sixteen-draw random null, so the denominator is the subset of axes most likely to steer.

Amplification lowers a labelled axis's own-domain loss for 12 of 46 axes on Llama-3.2-3B, 11 of 35 on Qwen3.5-9B and 11 of 57 on Mamba-2 — 19–31%, a minority in every family. It is also the wrong minority for a steering account: the effect is concentrated early and disappears with depth, falling from 44%, 62% and 20% at the first layer to 5%, 11% and 0% at the final block, with an independent Llama-2-7B run at its final block giving 1 of 18. And in 156 axes across four families (46/35/57 plus the 18-axis Llama-2-7B final-block run), the opposite sign never once lowers in-domain loss — nor does it on the 49 axes of the excluded run. So the effect is not a signed, controllable handle on the domain but an unreliable one-directional nudge that fades exactly where the code is strongest. We therefore make no steering claim.

One run is excluded from those figures on provenance grounds: the three-anchor Llama-2-7B sweep names a step-2 directory that no longer exists, so its basis cannot be matched to the one used everywhere else in this paper. The four numbers quoted above all come from runs whose basis is the paper's.

J Context-overwrite control

The ordering is generic to the stream: at Llama-2-7B's L27 anchor — the PCA control is run there rather than at the final block of Table 1, so that both bases are read from one set of forward passes — a matched-rank PCA subspace of the same activations also tracks the recent domain, on all 20 ordered pairs, at cosine 0.836 against 0.733 to the vector sum. What is not generic is the fidelity: V_{joint} recovers the active domain at 0.974 where the variance-matched subspace reaches 0.836, and on a wider margin over the sum, 0.143 against 0.103 — the division between a property any comparable subspace shows and a property only this one does. The new domain writes in within 16–32 tokens, identically on dense attention and on the state-space model, even though attention can re-read the whole context at every step while an SSM must compress it into a fixed state.

K Cross-model transport

The five pairs are Llama-2-7B → Llama-3.2-3B (L27 → L27), Llama-2-7B → Qwen3.5-9B (L27 → L31), Llama-2-7B → Mamba-2 (L27 → L63), Qwen3.5-9B → Llama-3.2-3B (L31 → L27), and Llama-2-7B → Llama-2-7B (L27 → L27) as the same-model positive control; the quoted numbers use these layer choices. Orthogonal Procrustes is defined only where $H_a = H_b$, so it is fitted for the same-model control and for the Llama-2-7B → Qwen3.5-9B pair, which we ran at two layer choices (L27 → L31 and L31 → L31); the -0.07–$0.31\times$ range in §4.4 spans those two, and the control's Procrustes leg sits at $1.95\times$.

L Per-layer structure

Containment of V_l in V_{l+1} is $\sum_i \varsigma_i^2 / K_l$, for ς the singular values of $V_l^\top V_{l+1}$, quoted over the chance share K_{l+1}/H that two unrelated subspaces of the same ranks would give. The per-family figure is the *median* over successive layer pairs: 36.4 (Llama-2-7B), 22.1 (Llama-3.2-3B), 46.3 (Qwen3.5-9B), 23.1 (Mamba-2), 10.1 (OLMoE at its router input). The statistic is computed for all five families from the released bases and shipped as `containment_summary.json`, which records the definition alongside every layer pair.

Between the first layer and the final block, raw containment sits at 1.04–1.84× chance — low, but not identically chance. The residue is the smooth basis drift that the adjacent-layer figures already show: on the rotation-aligned metric, which fits and removes the relative rotation between the two bases, it is 1.06× on Llama-3.2-3B and 1.01× on Mamba-2, the two families for which the drift artifacts were retained. That is the sense in which the endpoints are unrelated: not that no map connects them, but that no shared *coordinates* survive the distance.

M RANDOM-INITIALIZATION SEEDS

Each control model is instantiated from the trained checkpoint's own configuration, so depth, width, head count, vocabulary and every architectural detail match exactly; no trained weight is read at all in the fully-random arm. The initialization is the architecture's own, not ours: normal draws at the configured initializer scale for linear and embedding weights, ones for the RMSNorm gains — which is why the normalizations begin at identity — and, for Mamba-2, the state-space parameterization's own structured initialization for the SSM terms. We add no scheme of our own, which is what makes the arm a control on training rather than on initialization.

The trained-embedding arm differs by exactly one tensor. It copies the input embedding matrix from the trained checkpoint, and the output embedding where the architecture keeps a separate one. Both control architectures tie their input and output embeddings, so in practice a single matrix is transplanted and re-tied. Everything from the first block to the final norm stays at initialization. That is the sense in which the real $x^{(0)}$ is present at every layer of that control while nothing downstream has been trained.

A single global seed is set before instantiation, so a seed reproduces a checkpoint bit-for-bit. Five seeds per arm on the Llama-3.2-3B architecture give the ten control runs of §4.2. A smaller architecture control exists on Mamba-2, three seeds per arm, which we do not fold into that count. It agrees in sign, with all six slopes negative and every interval excluding zero (-0.31 to -0.34 fully random, -0.56 to -0.75 trained-embedding). Its magnitudes are far smaller than the Llama-3.2-3B arms' and are not comparable to them, since the models differ in depth, width and capture; only the sign is being read.

The checkpoints are written in float16 and, like the trained models, loaded in bfloat16 for activation collection. As §5 states, we take the rebuilt seeds to be faithful to the architecture on the strength of an identical code path and the absence of across-seed outliers rather than on a separate verification.

N CLAIMS LEDGER

Ten claims this construction initially appears to support, the control that tests each, and what is left. Four are withdrawn outright, four are narrowed, two survive.

Claim	Control	Verdict
Compact, single-tiered, second-moment object (§4.1)	θ and k_{each} sweeps; SVD of the complement	**Survives**, scoped: compactness is a claim about a five-domain calibration, and the honest denominator is the 256-direction candidate pool as well as H.
The class-mean shift is where the code lives (§4.1)	Matched-rank PCA subspace	**Withdrawn.** PCA captures 97–99% of the shift: it is a high-variance object, not a V_{joint} property.
Discriminability rises with depth because of training (§4.2)	Two random-init arms $\times$ 5 seeds; fixed rank; matched-rank floors	**Survives.** Ten control seeds negative, all CIs excluding zero; margin over the demanding floor 1.49–1.73$\times$.
The rise is specific to *domain*, not to any partition (§4.2)	Arbitrary C4 partition; two style quintiles	**Narrowed to a level result.** Controls sit 11–15$\times$ lower at every depth and rank, but their fold-change is not smaller, so the developmental content rests on the training axis alone.
Domain identity is constructed where language identity is given (§4.2)	Language identity at matched and own rank, 13 ranks $\times$ 9 depths	**Narrowed to a fold result.** Domain grows 9.1–39.5$\times$ against language's 2.4–6.9$\times$, but language starts higher and its additive slope is higher too.
Predictions depend on V_{joint} selectively (§4.3)	Matched-rank PCA and random subspaces	**Narrowed to its selectivity.** The *magnitude* of ablation damage is generic to high-variance subspaces; only the *selectivity* is specific.
A component reads the code forward (§4.3)	Matched-rank PCA conditions its own late coordinate	**Withdrawn.** PCA moves the late coordinate as far or further (0.18–0.40 against 0.03–0.19); the test is first-moment by construction.
The code is a steering knob (§4.3)	Per-axis additive amplification at $\alpha = \pm 1$	**Withdrawn.** Amplification lowers in-domain loss for 19–31% of causal axes, and for 0–11% at the final block; the direction the steering account predicts never helps, on any axis.
The code tracks the active domain rather than accumulating (§4.4)	Matched-rank PCA on the same activations	**Narrowed.** The ordering is generic to the stream; only the fidelity is specific. The preregistered accumulating reading is refuted either way.
Necessity extends beyond the calibration corpora, and across models (§4.4)	Three-domain basis; $n = 1$–50 sweep against a matched-rank floor; same-model transport control	**Withdrawn** in both directions. Held-out corpora retain 2–13% of the damage, and no fitted map transports the subspace between families. The discriminative *format* does generalize to unseen corpora.