Concepts Whisper While Syntax Shouts: Spectral Anti-Concentration and the Dual Geometry of Transformer Representations

Anonymous ACL submission

Abstract

We find that transformer concept representations systematically *anti-concentrate* in the spectral tail of the unembedding covariance, encoding word-level concepts in low-variance directions across a 17-model core suite, with convergent replications from three independent extraction methods. Residual-stream difference-of-means vectors show highly significant anti-concentration ($p < 10^{-33}$), with convergent support from sparse autoencoder (SAE) features ($p = 4.5 \times 10^{-19}$) and linear probes on Llama and Qwen. We identify a *dual geometry*: activation-space concept directions anti-concentrate while static unembedding-row contrasts *concentrate* in high-variance directions ($p < 10^{-4}$). This investigation arose from testing whether the causal inner product of Park et al. (2024) aids cross-lingual concept transport; a matched-spectrum randomization across 17 models and four language pairs finds no evidence that Whitened Causal Alignment improves over spectral regularization alone ($p = 0.95$). Split-injection interventions support a functional basis on Gemma and Llama (Cohen's d up to 1.80), and POS-tag probing across eight models shows syntax preferentially encoded in the high-variance subspace in six of eight architectures, with a significant reversal in the Qwen 2.5 family. These results suggest transformers rotate semantic content into spectrally quiet regions during contextualized processing, where, in some architectures, interventions may reduce grammatical disruption relative to high-variance steering.

1 Introduction

The mechanistic interpretability of large language models (LLMs) rests on an empirical observation: high-level concepts appear to be encoded as linear directions in activation space (Mikolov et al., 2013; Marks and Tegmark, 2024; Tigges et al., 2024; Nanda et al., 2023). Park et al. (2024) formalize this *Linear Representation Hypothesis* (LRH) and argue the natural geometry is a *causal inner product* defined by the unembedding covariance, under which causally separable concepts are provably orthogonal (Park et al., 2024, 2025). In parallel, the Platonic Representation Hypothesis (Huh et al., 2024) and universal-feature work (Lan et al., 2024) suggest that capable models converge on shared representational structure, motivating *cross-lingual concept transport* via Procrustes alignment (Conneau et al., 2018, 2020).

The intersection of these hypotheses motivates what we call *Whitened Causal Alignment* (WCA): align two representation spaces after whitening with the inverse square root of each model's unembedding covariance. If the causal geometry hypothesis is correct, WCA should outperform naive Procrustes by stripping language-specific anisotropy (Lim et al., 2025; Wendler et al., 2024).

We set out to validate this. Across 17 models and 4 language pairs, the causal eigendirections do *not* measurably improve transport in a matched-spectrum test (Section 4). But this failure exposed something more interesting: activation-space concept vectors systematically *anti-concentrate* in the eigenspectrum of Σ, preferentially encoding in low-variance directions (Section 5). The pattern is not an artifact: it persists for SAE features, linear probes, and difference-of-means vectors (Section 6). Moreover, the *static vocabulary* shows the opposite pattern—concept-specific unembedding-row differences *concentrate* in high-variance directions, exposing a *dual geometry* (Section 7). Steering experiments and POS-tag probing provide functional support for why this dual geometry exists (Section 8).

Contributions. (i) A negative result for causal eigendirections in cross-lingual transport, established by matched-spectrum randomization across

Table 1: Map of central claims to their evidence and the strength we ascribe to each. We separate *robust* findings (consistent within the model suite tested for that claim, with strong statistics) from *partial* ones (architecture-dependent or based on a smaller subset).

Claim	Evidence	Strength
Anti-concentration of activation-space concepts	17-model difference-of-means sweep ($p < 10^{-33}$); SAE/probe replications (§5,6)	robust
Independent of background spectrum	$r = 0.019$ between random and concept SCM (§5)	robust
Dual geometry (vocabulary concentrates)	4 models, random-pair null $p < 10^{-4}$ (§7)	robust
WCA adds no eigendirection benefit	17 models $\times$ 4 language pairs, $p = 0.95$ (§4)	robust (null)
Interference asymmetry (steering)	3/5 models show shout>whisper; 2/5 null/reversed (§8)	partial
Syntax in high-variance subspace	6/8 models; Qwen 2.5 reverses (§8)	partial

17 models. (ii) Discovery and validation of spectral anti-concentration of activation-space concept representations across three independent extraction methods. (iii) Identification of a dual geometry between vocabulary-space concentration and activation-space anti-concentration, with a random-pair null model showing concept-specific structure. (iv) Functional evidence via split-injection steering across five models and POS-tag probing across eight models. An anonymized code and data supplement accompanies the submission; a non-anonymous release will follow upon publication.

2 Background and Related Work

Linear representations and the causal inner product. Linear encoding has been demonstrated for truth (Marks and Tegmark, 2024; Burns et al., 2023), sentiment (Tigges et al., 2024), world state (Nanda et al., 2023), and refusal (Arditi et al., 2024); Engels et al. (2025) note that some features are multi-dimensional. For a token set $\mathcal{L}$, we use the empirical second moment

$$\Sigma_{\mathcal{L}} = \frac{1}{|\mathcal{L}|} \sum_{w \in \mathcal{L}} \gamma(w)\gamma(w)^\top + \lambda I, \qquad (1)$$

where $\gamma(w) \in \mathbb{R}^d$ is the unembedding row for w. The causal inner product is $\langle u, v \rangle_C = u^\top \Sigma_{\mathcal{L}}^{-1} v$ (Park et al., 2024, 2025).

Cross-lingual alignment and WCA. Procrustes alignment finds Q minimizing $\|X_{\text{src}} - X_{\text{tgt}} Q\|_F$ with $Q^\top Q = I$ (Conneau et al., 2018). WCA composes this with whitening: $v_{\text{tgt}} = \Sigma_{\text{tgt}}^{1/2} Q^\top \Sigma_{\text{src}}^{-1/2} v_{\text{src}}$.

Anisotropy and SAEs. Transformer activations are anisotropic, with a few outlier dimensions dominating variance (Ethayarajh, 2019; Gao et al., 2019; Timkey and van Schijndel, 2021; Puccetti et al., 2022; Kovaleva et al., 2021; Machina and Mercer, 2024); some of these dimensions encode structured content (Gurnee and Tegmark, 2023). SAEs (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024; Gao et al., 2025; Lieberum et al., 2024; He et al., 2024) decompose activations into sparse interpretable features.

Spectra, steering, and whitening. Martin and Mahoney (2021, 2019); Martin et al. (2021) study spectral structure of weights; Staats et al. (2024) show small singular values carry information; Sharma et al. (2023) show pruning top components can *improve* reasoning. Activation steering (Turner et al., 2024; Panickssery et al., 2024; Zou et al., 2023; Li et al., 2024; Subramani et al., 2022) manipulates concept directions at inference time. Whitening helps sentence embeddings (Su et al., 2021; Huang et al., 2021), plausibly by amplifying low-variance directions where concepts live. Papyan et al. (2020); Wu and Papyan (2024) study neural collapse; our finding is distinct—concepts are pushed *away from* the dominant spectral subspace.

3 Experimental Setup

Models. Core spectral experiments use 17 decoder-only base models across multiple families (Section B): Llama 3.2 (1B, 3B); Qwen 2.5 (0.5B–14B); Qwen 3 (0.6B–8B); Gemma 2 (2B, 9B); Mistral-7B-v0.3; SmolLM2-1.7B; and two MoE models (OLMoE-1B-7B, JetMoE-8B). We additionally use Llama 3.1-8B for SAE and functional tests. Hidden dims 896–5120; vocab 32K–256K. All are decoder-only base models.

Concepts. 27 concept directions in four categories from the counterfactual word-pair dataset of Park et al. (2024) (sourced from Gladkova et al., 2016): Verb Morphology (10), Semantic (7), Grammatical (6), and Language Pairs (4); 4–228 pairs each.

Three independent extraction methods. (1) *difference-of-means on residuals (subtractive):*

$v_{\text{dom}} = \frac{1}{N}\sum_i (h(w_i^+) - h(w_i^-))$ over final-layer pooled activations. As a subtractive estimator it could in principle cancel shared high-variance components, motivating the artifact controls in Section 6. (2) *SAE features (contextualized, unsupervised):* top-k differentially activating features from Gemma Scope and Llama Scope; decoder columns serve as concept directions. No subtraction. (3) *Linear probes (contextualized, supervised):* L2-regularized logistic regression on uncentered final-layer activations; the weight vector is the concept direction. L2 biases toward high-variance directions, making this a conservative test.

Spectral analysis. Project concept vectors onto the eigenbasis of $\Sigma_{\mathcal{L}}$ (Equation (1)) with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. The fractional energy at u_i is $E_i = (v^\top u_i)^2/\|v\|^2$; cumulative energy $C(k) = \sum_{i \leq k} E_i$ is plotted against cumulative variance $V(k) = \sum_{i \leq k} \lambda_i / \mathrm{Tr}(\Sigma)$ to compare across d. *Gini deviation* is the signed area between $C(\cdot)$ and the uniform baseline (negative = anti-concentration). *Spectral Center of Mass* (SCM) is the $V(k)$ at which $C(k) = 0.5$; SCM near 1 means encoding in low-eigenvalue directions. For partitioned subspace tests we use $k = \lfloor 0.1d \rfloor$ and verify robustness at $k \in \{5\%, 10\%, 20\%\}$ (Section L). Hyperparameters and seeds are listed in Section A.

4 The Causal Geometry Does Not Aid Cross-Lingual Transport

To determine whether WCA's benefit stems from the specific eigenvector directions of Σ or merely from its spectral shape, we construct matched-spectrum *fake* covariances $\Sigma_{\text{fake}} = V\Lambda V^\top$, where Λ contains the true eigenvalues and V is Haar-random orthogonal. We run the full WCA pipeline with Σ_{fake} for both source and target, with 5 seeds across all 17 models $\times$ 4 language pairs (68 combinations) at $\lambda = 0.1$.

Real WCA achieves a 50.3% win rate over naive Procrustes (mean $\Delta = +0.0009$); Fake WCA achieves 51.0% ($\Delta = +0.0010$); paired t-test $t = -0.063$, $p = 0.95$ (Figure 1). Within this resolution any observed benefit of WCA is attributable to spectral regularization—dampening high-variance directions regardless of semantic content—rather than to semantically meaningful causal eigendirections. This negative result is the direct entry point to our positive findings: if the eigenvectors of Σ do not carry cross-lingual transport signal, what *do* they encode?

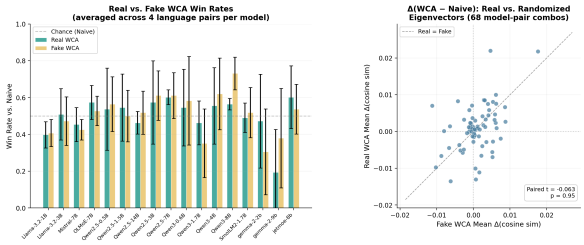


Figure 1: Matched-spectrum randomization across 68 model-pair combinations. Real WCA (50.3% win rate over naive Procrustes) is indistinguishable from Fake WCA with randomized eigenvectors (51.0%); paired t-test $t = -0.063$, $p = 0.95$.

Table 2: Representative spectral metrics. Concept SCM exceeds the random baseline on average, and all Gini deviations are negative across all 17 models; the full table is in Section C.

Model	d	Conc. SCM	Rand. SCM	Gap	Gini
Qwen2.5-0.5B	896	0.997	0.711	+0.286	−0.356
Qwen2.5-14B	5120	0.996	0.790	+0.205	−0.346
Qwen3-8B	4096	1.000	0.783	+0.217	−0.359
Llama-3.2-3B	3072	0.771	0.771	+0.001	−0.159
Gemma-2-9b	3584	0.927	0.849	+0.078	−0.364
Mistral-7B	4096	0.908	0.792	+0.116	−0.199
JetMoE-8B	2048	0.945	0.741	+0.205	−0.242
Mean (all 17)		0.926	0.758	+0.169	−0.282

5 Spectral Anti-Concentration

The matched-spectrum pipeline required eigendecompositions of Σ for every model, naturally surfacing how concept vectors distribute across these eigenbases. Across all 17 models, difference-of-means concept vectors show highly significant anti-concentration: mean Gini $= -0.282$, all within-model $p < 10^{-6}$; aggregate $p < 10^{-33}$. The SCM gap between concept and random baselines is $+0.169$ on average (Figure 2), and is independent of background spectral shape (random vs. concept SCM: $r = 0.019$, $p = 0.94$). Pairwise Wilcoxon tests find no significant differences between most concept categories (e.g., Verb Morphology vs. Semantic, $p = 0.73$); all categories anti-concentrate to comparable degree.

Table 2 highlights two patterns visible in the full data. First, all 17 models show negative Gini, supporting broad anti-concentration across families and scales. Second, the SCM gap is positive on average but small or slightly negative in the Llama 3.2 and OLMoE cases; these models still anti-concentrate by the Gini metric—two complementary measures that disagree when the concept-energy CDF lies below uniform but with a center-

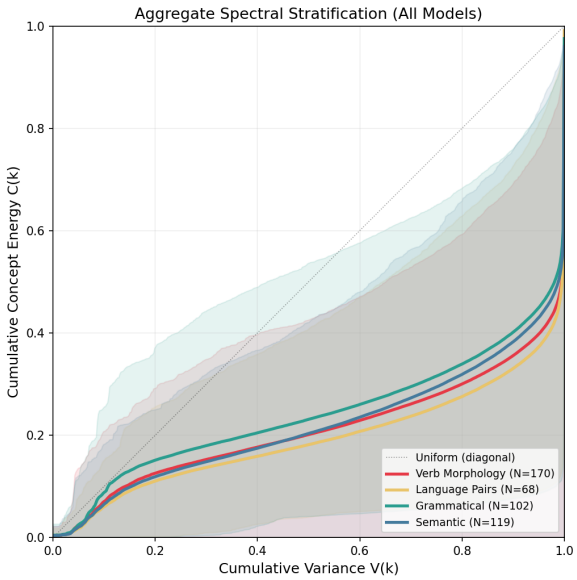


Figure 2: Aggregate spectral energy CDFs across 17 models. Concept-group curves (colored) fall below the random baseline band (grey), indicating systematic anti-concentration. The x-axis is cumulative variance $V(k)$.

of-mass close to the random baseline.

Anti-concentration is not just anisotropy. Anisotropy refers to the well-documented observation that transformer activations cluster in a narrow cone, with a few dimensions carrying disproportionate variance (Ethayarajh, 2019; Gao et al., 2019; Timkey and van Schijndel, 2021). Our claim is distinct on three counts. (i) Anti-concentration is a property of *specific concept directions*, measured relative to random directions evaluated against the same anisotropic spectrum: the same eigenbasis is used for both, so the spectral shape cancels by construction. (ii) The correlation between random-vector SCM and concept-vector SCM is negligible across our 17 models ($r = 0.019$, $p = 0.94$); concepts do not anti-concentrate *because* the spectrum is concentrated, but in addition to it. (iii) The *dual geometry* (Section 7) is incompatible with a pure anisotropy account: concept-specific unembedding-row differences *concentrate* in the same eigenbasis where activation-space concepts anti-concentrate. If anisotropy alone drove the effect, both views of the concept would inherit the same spectral bias.

6 Ruling Out the Subtraction Artifact

The difference-of-means estimator mechanically cancels components shared between paired tokens, potentially biasing it toward low-variance directions. We rule this out with two methods that do not involve subtraction.

Table 3: SAE artifact check: anti-concentration is highly significant across three models in two families with no subtraction involved.

Model (SAE)	Mean Gini	t	p
Gemma-2-2b (top-1)	-0.230	-23.6	4.5×10^{-19}
Gemma-2-2b (top-5)	-0.240	-51.3	1.2×10^{-27}
Gemma-2-9b (top-5)	-0.276	-52.8	5.6×10^{-28}
Llama-3.1-8B (top-5)	-0.193	-49.1	3.6×10^{-27}

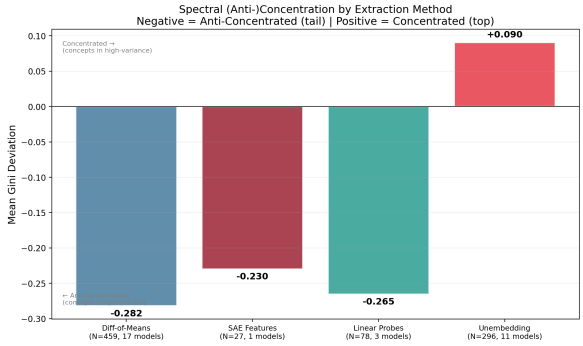


Figure 3: Cross-method Gini deviation across models. difference-of-means, SAE features, and linear probes all yield negative Gini (anti-concentration); unembedding-row contrasts have the opposite sign—the dual geometry of Section 7.

SAE features. Top-k differentially activating features on Gemma Scope (gemma-2-2b layer 20, gemma-2-9b layer 31, width 16K) and Llama Scope (Llama-3.1-8B layer 24, width 32K) all show strong anti-concentration (Table 3). Top-5 selection *strengthens* the effect relative to top-1 on gemma-2-2b, ruling out feature collapse as a driver—the most-collapsed feature (used by 5/27 concepts) was the *weakest* contributor; removing it strengthened the aggregate (Section E).

Linear probes. L2-regularized logistic regression on uncentered activations reaches mean accuracy > 0.99 on Llama-3.2-3B (Gini $= -0.167$, $p < 10^{-13}$) and Qwen2.5-3B (-0.185, $p < 10^{-15}$). Probes failed on gemma-2-9b (50% accuracy on 25/27 concepts), likely due to RMSNorm interaction with uncentered inputs; we exclude this model from probe-based claims (SAE evidence on the same model is unaffected).

Three independent activation-space methods converge (Figure 3): anti-concentration is a property of the representations, not of difference-of-means subtraction.

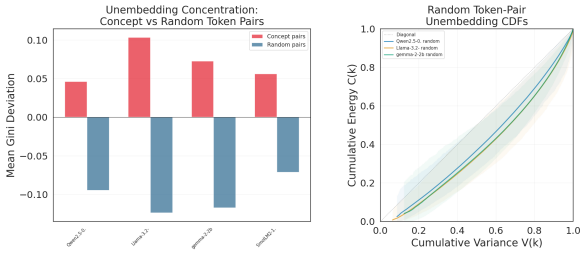


Figure 4: Random unembedding null. Random token-pair differences (grey) mildly anti-concentrate, while concept-specific pairs (colored) actively concentrate in the top eigenvalues ($p < 10^{-4}$).

7 The Dual Geometry

Unembedding contrasts concentrate. Constructing concept directions directly from static unembedding rows, $v_{\text{unembed}} = \text{mean}_i(\gamma(w_i^+) - \gamma(w_i^-))$ over single-token pairs, the resulting directions *concentrate* in high-variance eigendirections (positive Gini), the opposite sign of activation-space directions.

Concept pairs are non-random. Against a null of 1,000 random token-pair differences from the same vocabulary, concept pairs concentrate significantly more (Figure 4). Random pairs mildly anti-concentrate (mean Gini ≈ -0.10); concept pairs concentrate (mean $\approx +0.07$); $p < 10^{-4}$ on Qwen2.5-0.5B, Llama-3.2-1B, Gemma-2-2b, and SmolLM2-1.7B.

Interpretation. The unembedding matrix places *concept contrasts along high-variance directions*, where they maximally influence logit differences. Activation-space representations encode the same concepts in the spectral tail, away from the dominant structural variance. Additional layer-wise controls show that this rotation emerges during contextualized processing rather than being inherited wholesale from static embeddings: mean Gini moves from near-neutral at the embedding layer (-0.009) to tail-aligned at layer 0 (-0.171) and the final layer (-0.277; Section D).

8 Why Concepts Anti-Concentrate

The dual geometry motivates two functional tests: a steering experiment isolating spectral components, and a syntactic probing experiment characterizing what each subspace encodes.

Split-injection interference. For each concept we project onto the top-k (*shouting*) and bottom-k (*whispering*) eigenvectors ($k = 0.1d$) and steer at four mid-depth layers (50/62/75/87%) at scaling α, measuring perplexity on 10 held-out Wikipedia passages across five models. On Gemma-2-2b, Llama-3.1-8B, and Llama-3.2-3B we observe a clear asymmetry (Table 4 and fig. 5); on Qwen2.5-3B and Mistral-7B no consistent shouting-greater-than-whispering asymmetry emerges. We report representative α values from the diagnostic range of each model; these settings were selected after inspecting full sweeps (Section J) and are model-dependent.

Table 4: Split-injection interference asymmetry across five models. d is Cohen's d for the per-concept difference (shout PPL − whisper PPL); positive values mean shouting causes more disruption.

Model	α	Shout/Whisper	d	p
gemma-2-2b	10	+1.0% / +0.5%	+0.86	.0002
gemma-2-2b	20	+4.8% / +1.8%	+1.12	$<10^{-4}$
Llama-3.1-8B	5	+383% / +223%	+1.80	$<10^{-4}$
Llama-3.2-3B	1	+4.6% / +2.9%	+1.45	.0001
No significant asymmetry:				
Qwen2.5-3B	10	+11% / +13%	−0.21	.51
Mistral-7B	5	+18K% / +59K%	−0.32	.25

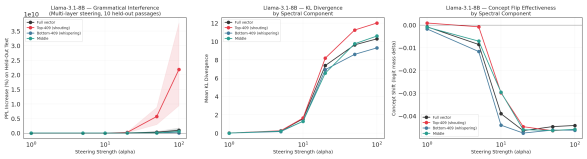


Figure 5: Split-injection on Llama-3.1-8B. PPL increase (left), KL divergence (center), and concept shift (right) vs. steering strength α. At $\alpha = 5$, the shouting component produces +383% PPL increase vs. +223% for whispering (Cohen's $d = 1.80$, $p < 10^{-4}$).

POS-tag probing. We train 13-class logistic regression classifiers on activations from 500 UD English-EWT sentences (Nivre et al., 2016) (6,585 tokens after filtering tags with ≥ 30 samples), projected onto the top-k and bottom-k eigenvectors of Σ ($k = 0.1d$). Table 5 shows that 6/8 models place significantly more POS information in the high-variance subspace; both Qwen 2.5 models reverse this. Notably, Qwen 3-4B—same vocabulary, different training—shows the expected positive gap.

A UD Chinese-GSD control on Qwen2.5-3B and Qwen 3-4B yields bottom-k > top-k for both, so the Qwen 2.5 reversal cannot be reduced to a single English-vs-multilingual mismatch (Section K). The qualitative pattern is invariant under $k \in \{5\%, 10\%, 20\%\}$ (Section L). Both subspaces

Table 5: POS-tag probing accuracy on top-k (high-variance) vs. bottom-k (low-variance) subspaces at $k = 0.1d$. Six of eight models show top-k > bottom-k; both Qwen 2.5 models reverse. 95% bootstrap CIs and the full table are in Section K.

Model	Top-k	Bot-k	Gap (p)
gemma-2-2b	.554	.506	$+.048$ (.0001)
Llama-3.2-3B	.628	.576	$+.052$ (.0001)
Llama-3.1-8B	.641	.625	$+.016$ (.012)
Mistral-7B	.624	.607	$+.017$ (.013)
SmolLM2-1.7B	.560	.536	$+.024$ (.002)
Qwen3-4B	.572	.555	$+.018$ (.001)
Qwen2.5-3B	.557	.576	$-.019$ ($<.001$)
Qwen2.5-7B	.572	.586	$-.014$ (.002)

underperform full-dimensional probes (78–86%); the high-variance subspace carries a measurable *bias* toward syntactic information.

Scope of the functional claim. Anti-concentration itself replicates across all 17 models and three extraction methods; the *functional* interpretation—that high-variance directions disproportionately carry syntactic load and that injecting along them disrupts fluency more—is supported on Gemma and Llama and is partial on the broader set. POS probing reverses on Qwen 2.5 and split-injection shows null/reversed effects on Qwen 2.5 and Mistral. We therefore present the functional account as a candidate mechanism that is consistent with anti-concentration on most tested architectures but does not yet generalize uniformly. Steering or probing recommendations derived from it should be evaluated per-model.

9 Discussion and Conclusion

Connection to superposition. If the high-variance eigenspace is claimed by frequent, high-information syntactic regularities, the effective dimensionality available for semantic encoding is $d_{\text{eff}} \ll d$, forcing superposition for semantic features (Elhage et al., 2022; Scherlis et al., 2022) even when d is large. Anti-concentration is thus a candidate spectral signature of how transformers allocate representational capacity: syntax claims the "loud" subspace where it can robustly survive residual-stream noise, while semantic concepts share the "quiet" subspace via superposition.

Scaling and architecture. The Qwen 2.5 (0.5B–14B) and Qwen 3 (0.6B–8B) families let us check whether anti-concentration changes with capacity. The SCM gap remains positive at all scales (Qwen 2.5: $+0.205$ to $+0.286$; Qwen 3: $+0.215$ to $+0.259$) without a clear monotonic trend (Section H); MoE models (OLMoE, JetMoE) show the same pattern as dense families, and Gemma's larger condition number does not abolish the effect (Section I). We interpret this as a property present at all tested scales rather than one that emerges or vanishes with capacity.

Implications for interpretability and steering. PCA-based visualizations and low-rank approximations risk missing concept-level structure; interpretability tools should attend to the spectral tail. On models where the interference asymmetry appears, it suggests a practical refinement to activation engineering (Turner et al., 2024; Panickssery et al., 2024): projecting steering vectors into the low-variance subspace before injection may achieve concept manipulation with lower fluency disruption, possibly explaining why naive steering can cause aphasia (Arditi et al., 2024). The complementary finding that whitening improves sentence embeddings (Su et al., 2021; Huang et al., 2021) is consistent with this picture: whitening amplifies the low-variance directions where concepts reside. Sharma et al. (2023) report reasoning gains from removing top SVD components (LASER), again consistent with the dual geometry we describe.

The Qwen 2.5 anomaly. Qwen 2.5 reverses the POS pattern while Qwen 3 does not, despite sharing vocabulary, and the reversal persists at $k \in \{5\%, 10\%, 20\%\}$. Chinese probing on UD Chinese-GSD reverses for *both* Qwen families, so language alone does not explain the English-vs-rest split. A natural hypothesis is that Qwen 2.5's relatively flat eigenspectrum (the smallest top-10% variance capture in our sample at $\sim$27%) makes any single cutoff insufficient to separate syntactic from semantic directions; the k-sensitivity analysis (Section L) argues against this, however, since the reversal persists at $k = 20\%$. We hypothesize that training-pipeline differences distribute syntactic information differently across the eigenspectrum, and leave a controlled investigation of pre-training mixture effects to future work.

Conclusion. We set out to validate the causal inner product as a basis for cross-lingual transport; matched-spectrum randomization shows no advantage for the specific eigendirections of Σ

over randomized directions with the same spectrum, suggesting WCA acts as spectral regularization. The failure exposed a broader pattern: word-level concept directions—across SAE features, linear probes, and difference-of-means vectors over a 17-model core suite—systematically avoid high-variance eigendirections, while static unembedding-row contrasts do the opposite. Split-injection on Gemma and Llama and POS-tag probing across 8 models give partial functional support: in most tested architectures, high-variance directions disproportionately encode syntax. *Concepts whisper while syntax shouts*, with immediate consequences for how we probe, interpret, and steer large language models.

10 Limitations

Our concepts are word-level counterfactual pairs (Park et al., 2024; Gladkova et al., 2016); whether anti-concentration extends to abstract concepts such as honesty, sarcasm, or planning is open. We use an unembedding second moment from a single matrix; centered, frequency-weighted, and input-embedding controls preserve the main effect (Section D), but activation-covariance spectra may reveal additional structure. Our layer-wise analysis is descriptive; a causal per-layer intervention study of where the vocabulary-to-reasoning rotation is implemented remains future work.

Split-injection steers at four layers simultaneously; per-layer interference profiles were not logged, so we cannot localize the asymmetry in depth. Functional evidence is strongest on Gemma and Llama; null results on Qwen 2.5 and Mistral suggest the spectral separation may be architecture- or training-dependent. POS-probing significance is based on 5 cross-validation folds plus token-level bootstrap; sentence- or document-clustered resampling would be stronger.

All tested models are decoder-only *base* (non-instruct) variants. Encoder or encoder-decoder architectures may differ, and whether RLHF or instruction tuning alters the spectral allocation of concept information is open. We chose base models to isolate pre-training geometry from post-training alignment effects.

11 Ethical Considerations

This work uses only publicly available pre-trained language models and publicly available linguistic datasets (UD English-EWT and UD Chinese-GSD (Nivre et al., 2016)); no human subjects were involved and no personally identifiable information was used. The phenomena we report—spectral structure and steering-related interference—are properties of model internals rather than user-facing outputs; we do not introduce new capabilities for generating harmful content. Our split-injection results may inform safer steering by suggesting that interventions targeted at low-variance directions could reduce fluency disruption, but they also make explicit which subspaces are most sensitive to manipulation, which could in principle be misused; we believe the interpretability and safety benefits clearly outweigh this risk. Licenses, dataset statistics, compute budget, and AI-assistance disclosures are reported on the Responsible NLP Checklist.

References

Andy Arditi, Oscar Obeso, Aaquib Suri, and Fazl Barez. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6022–6034.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*.

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. 2022. Toy models of superposition. *Transformer Circuits Thread*.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 55–65.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. 2024. Llama scope: Extracting millions of features from Llama-3.1-8B with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of EMNLP*, pages 238–244.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 20617–20642.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of ACL-IJCNLP*, pages 3392–3405.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *BlackboxNLP Workshop*, pages 278–300.

Zhi Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*.

Anemily Machina and Robert Mercer. 2024. Anisotropy is not inherent to transformers. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4892–4907.

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

Charles H. Martin and Michael W. Mahoney. 2019. Traditional and heavy-tailed self regularization in neural network models. In *International Conference on Machine Learning (ICML)*, pages 4284–4293.

Charles H. Martin and Michael W. Mahoney. 2021. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73.

Charles H. Martin, Tongsu Peng, and Michael W. Mahoney. 2021. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12:4122.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR Workshop*.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *BlackboxNLP Workshop*, pages 16–30.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 1659–1666.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation

addition. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15504–15522.

Vardan Papyan, X.Y. Han, and David L. Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations (ICLR)*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 39643–39666.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of EMNLP*, pages 1286–1304.

Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. 2022. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. 2023. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *arXiv preprint arXiv:2312.13558*.

Max Staats, Matthias Thamm, and Bernd Rosenow. 2024. Small singular values matter: A random matrix analysis of transformer models. *arXiv preprint arXiv:2410.17770*.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of ACL*, pages 566–581.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, and 1 others. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Linear representations of sentiment in large language models. In *BlackboxNLP Workshop*, pages 58–87.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4527–4546.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2024. Activation addition: Steering language models without optimization. In *AAAI Conference on Artificial Intelligence*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15366–15394.

Robert Wu and Vardan Papyan. 2024. Linguistic collapse: Neural collapse in (large) language models. *arXiv preprint arXiv:2405.17767*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, and 1 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Hyperparameters and Compute

Spectral analysis. Regularization $\lambda = 0.1$ unless stated; eigendecomposition via dense SVD on Σ. Random baselines use 1,000 unit vectors sampled uniformly from S^{d-1}.

WCA. 5 random seeds, 17 models $\times$ 4 language pairs (en-fr, en-de, en-es, en-zh), $\lambda = 0.1$. Procrustes solved on whitened space via SVD.

Linear probes. L2-regularized logistic regression ($C = 1.0$, max_iter $= 2000$), 5-fold stratified CV, per-model standardization disabled (uncentered).

SAE. Gemma Scope and Llama Scope public releases (no retraining). Top-$k \in \{1, 3, 5\}$ differentially activating features; decoder-column unit vectors.

Split-injection. Layers 50/62/75/87% of depth; $\alpha \in \{1, 5, 10, 20, 50\}$ swept; perplexity on 10 held-out Wikipedia passages.

POS probing. UD English-EWT dev (2,001 sentences, 25,147 tokens; 6,585 after filtering tags with $\geq$ 30 samples); UD Chinese-GSD (500 sentences, 10,813 filtered tokens); $k \in \{5\%, 10\%, 20\%\}$; 5-fold stratified CV plus 10,000-resample token-level bootstrap.

Compute. Experiments were run on a mix of NVIDIA RTX 6000 Ada and RTX A6000 GPUs. The full pipeline (17 models, 27 concepts, three

extraction methods, WCA randomization, split-injection sweeps, POS probing) consumed approximately 320 GPU-hours; the dominant cost is split-injection on Llama-3.1-8B and Qwen2.5-14B ($\sim$110 GPU-hours combined).

B Full Model Details

All models are decoder-only transformers with pre-normalization (RMSNorm for Gemma, Llama, Qwen, and Mistral; LayerNorm for SmolLM2 and the MoE models). Concept vectors and covariance matrices are computed from the unembedding matrix W_U accessed via each model's LM head; for tied-weight models (Qwen, SmolLM2) W_U is the shared embedding matrix. All experiments use base (non-instruct) variants.

C Full Spectral Metrics (All 17 Models)

All 17 models show negative Gini, supporting broad anti-concentration. The Llama 3.2 family shows the smallest SCM gaps (-0.010 to $+0.001$); SmolLM2 has the strongest Gini (-0.436); OL-MoE is the only model with a negative SCM gap (-0.116) while still anti-concentrating (Gini $= -0.076$). The correlation between random and concept SCM is negligible ($r = 0.019$, $p = 0.94$).

D Robustness and Layer-wise Controls

We stress-test the anti-concentration result under covariance variants, random-direction nulls, and layer-wise trajectories. These controls use the same concept set as the core spectral analysis.

Covariance variants. The effect persists across full-vocabulary, language-subset, centered, uncentered, frequency-weighted, and input-embedding spectra (Figure 6). The saved language-subset covariance yields mean Gini -0.275 over 17 models; frequency weighting strengthens the effect (mean Gini -0.350 uncentered, -0.367 centered), arguing against the result being an artifact of uniformly weighting rare vocabulary items.

Layer-wise trajectory. Concept directions are near-neutral at the embedding layer and become tail-aligned after contextualization (Figure 7). Across models, mean Gini shifts from -0.009 at the embedding layer to -0.171 at layer 0 and -0.277 at the final layer. This supports the interpretation that the forward pass rotates concept information from the vocabulary geometry into the low-variance activation subspace.

Random-direction nulls. Random directions evaluated in an anisotropic spectrum can themselves show negative Gini, so the relevant test is whether learned concept directions are *more* tail-aligned than matched random directions. Norm-matched random-direction controls show a mean true-minus-null Gini gap of -0.082 across model summaries, with stronger tail alignment in 13/17 models (Figure 8). Llama 3.2, Mistral, and OL-MoE are weaker under this null, matching the smaller SCM gaps noted in Section C.

E SAE Feature Collapse Analysis

On gemma-2-2b with top-1 selection, 21/27 concepts mapped to unique SAE features (77.8% unique). Three features appeared more than once: #11527 (5 concepts), #38 (2), #1643 (2). Feature #11527 had the weakest Gini (-0.147); removing the 5 collapsed concepts *strengthened* the mean from -0.230 to -0.240. With top-5 selection, 84/135 feature slots were unique and the aggregate Gini was stronger (-0.240, $p = 1.2 \times 10^{-27}$). On gemma-2-9b with top-3, all 27 concepts mapped to distinct feature sets with zero collapse.

F Artifact-Check Per-Concept Plots

This appendix shows per-concept spectral energy CDFs for the non-subtractive extraction methods used in Section 6. The plots provide a visual check that anti-concentration is not driven by a small number of outlier concepts or by difference-of-means subtraction.

G Dual Geometry Details

This appendix reports the unembedding-row null used to support the dual-geometry claim in Section 7. Concept-specific token pairs concentrate more strongly in high-variance eigendirections than random token-pair differences from the same vocabulary.

H Scaling Data

The SCM gap is non-monotonic in capacity; the Gini gap widens with size in both families but with outlier behavior at the smallest scales. We interpret this as evidence that anti-concentration is present at all tested scales rather than emerging or disappearing with capacity.

Table 6: Complete model specifications for all 18 models tested (17 in the core spectral analysis, plus Llama-3.1-8B used for SAE and functional tests).

Family	Model	Params	d	Vocab	Cond. #
Llama 3.2	Llama-3.2-1B	1B	2048	128K	1.4×10^4
	Llama-3.2-3B	3B	3072	128K	2.5×10^4
Llama 3.1	Llama-3.1-8B	8B	4096	128K	2.7×10^4
Qwen 2.5	Qwen2.5-0.5B	0.5B	896	152K	4.2×10^3
	Qwen2.5-1.5B	1.5B	1536	152K	5.8×10^3
	Qwen2.5-3B	3B	2048	152K	5.7×10^3
	Qwen2.5-7B	7B	3584	152K	2.9×10^4
	Qwen2.5-14B	14B	5120	152K	1.6×10^4
Qwen 3	Qwen3-0.6B	0.6B	1024	152K	2.6×10^3
	Qwen3-1.7B	1.7B	2048	152K	4.3×10^3
	Qwen3-4B	4B	2560	152K	1.0×10^4
	Qwen3-8B	8B	4096	152K	1.6×10^4
Gemma 2	Gemma-2-2b	2B	2304	256K	1.6×10^5
	Gemma-2-9b	9B	3584	256K	1.6×10^5
Mistral	Mistral-7B-v0.3	7B	4096	32K	2.8×10^5
SmolLM	SmolLM2-1.7B	1.7B	2048	49K	2.7×10^4
MoE	OLMoE-1B-7B	1B/7B	2048	50K	7.7×10^3
	JetMoE-8B	8B	2048	32K	3.7×10^3

I Architecture Comparison

Gemma models exhibit the highest eigenvalue concentration (top 10% captures 40–52% of variance vs. 30–32% for Llama), consistent with their larger vocabulary (256K vs. 128K). Despite this difference, both families show strong anti-concentration; MoE models (OLMoE, JetMoE) exhibit similar patterns, suggesting the phenomenon is not specific to dense FFNs versus sparse experts.

J Split-Injection Details

Several concepts on gemma-2-2b have multi-token word pairs producing zero-norm activation-derived vectors and zero spectral energy; these are excluded from the aggregate statistics (effective $N = 18$ on gemma-2-2b; $N = 27$ elsewhere).

K POS Probing Details

The top-k subspace captures 27–40% of total variance (vs. $k/d = 10\%$), while bottom-k captures only 1.5–2.6%. Despite this asymmetry the accuracy gaps are modest (1.4–5.2 pp on positive-gap models), indicating syntactic information is distributed but biased toward the high-variance subspace.

L POS Probing Sensitivity Analysis

Table 7: Spectral Center of Mass and Gini deviation for all 17 models. Mean across models is reported on the last row.

Model	d	**Cond. #**	**Concept SCM**	**Random SCM**	**Gap**	**Gini**
Qwen2.5-0.5B	896	4.2K	0.997	0.711	+0.286	−0.356
Qwen2.5-1.5B	1536	5.8K	0.986	0.729	+0.257	−0.282
Qwen2.5-3B	2048	5.7K	0.973	0.735	+0.238	−0.272
Qwen2.5-7B	3584	29K	0.999	0.757	+0.242	−0.396
Qwen2.5-14B	5120	16K	0.996	0.790	+0.205	−0.346
Qwen3-0.6B	1024	2.6K	0.938	0.723	+0.215	−0.216
Qwen3-1.7B	2048	4.3K	0.997	0.738	+0.259	−0.248
Qwen3-4B	2560	10K	0.994	0.750	+0.243	−0.312
Qwen3-8B	4096	16K	1.000	0.783	+0.217	−0.359
Llama-3.2-1B	2048	14K	0.738	0.748	−0.010	−0.140
Llama-3.2-3B	3072	25K	0.771	0.771	+0.001	−0.159
Gemma-2-2b	2304	160K	0.964	0.796	+0.168	−0.389
Gemma-2-9b	3584	160K	0.927	0.849	+0.078	−0.364
Mistral-7B	4096	284K	0.908	0.792	+0.116	−0.199
SmolLM2-1.7B	2048	27K	1.000	0.735	+0.265	−0.436
OLMoE-1B-7B	2048	7.7K	0.620	0.736	−0.116	−0.076
JetMoE-8B	2048	3.7K	0.945	0.741	+0.205	−0.242
Mean			0.926	0.758	+0.169	−0.282

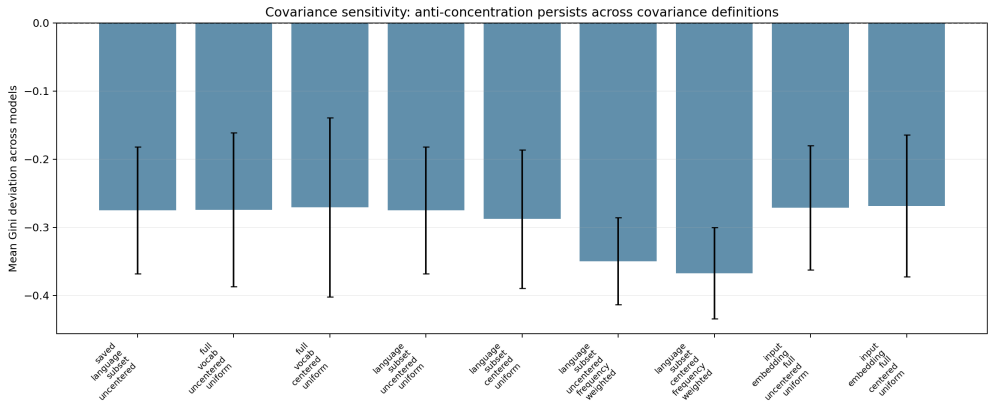
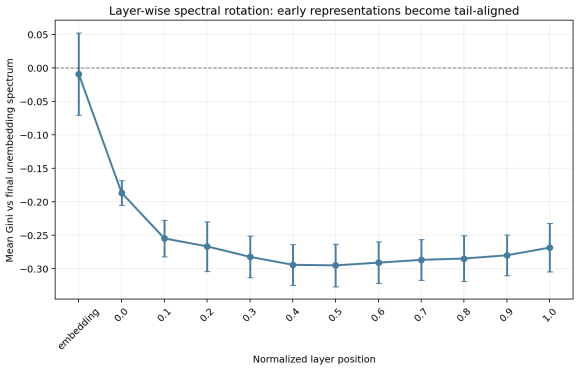
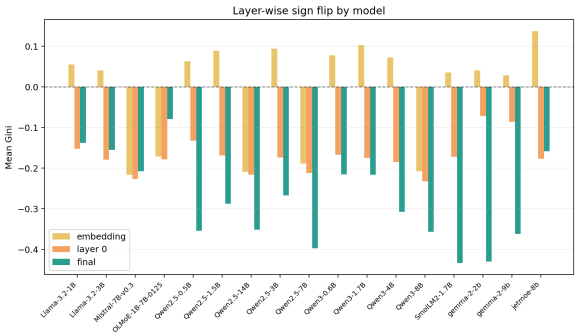


Figure 6: Covariance robustness. Mean Gini remains negative across covariance construction choices, including centered and frequency-weighted variants and input-embedding spectra. Error bars show variation across models.



(a) Aggregate layer-wise trajectory



(b) Embedding, layer 0, and final layer by model

Figure 7: Layer-wise spectral rotation. Concept directions move from near-neutral or mixed spectral placement at the embedding layer to stronger low-variance alignment after early transformer processing and by the final layer.

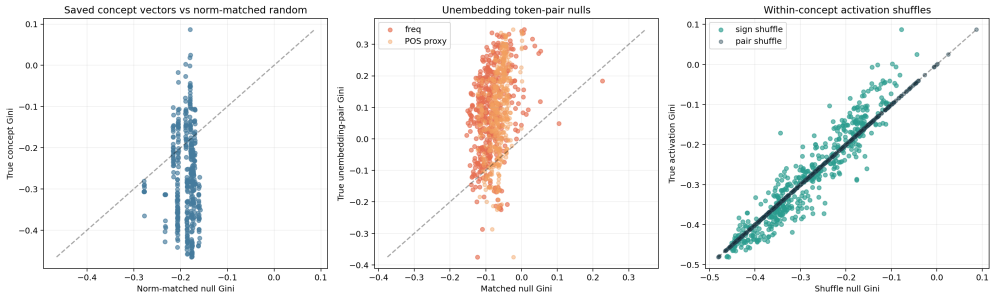


Figure 8: Norm-matched random-direction nulls. Most models place true concept directions further into the spectral tail than matched random directions; exceptions coincide with the weaker anti-concentration cases in the full spectral table.

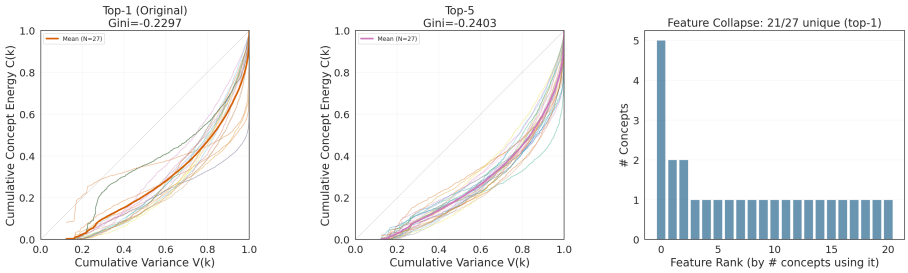
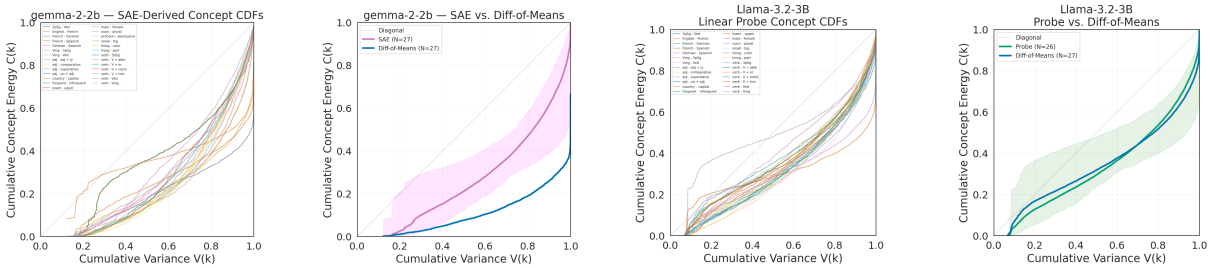
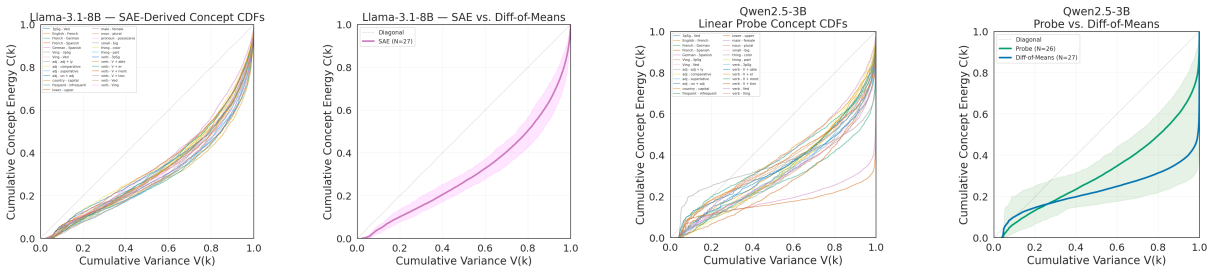


Figure 9: SAE feature collapse on gemma-2-2b. The most-collapsed feature (#11527, 5/27 concepts) is the weakest; removing collapsed concepts strengthens the aggregate Gini.



(a) SAE features (gemma-2-2b) (b) Linear probes (Llama-3.2-3B)

Figure 10: Per-concept spectral energy CDFs. Nearly all concept curves (colored) fall below the random baseline (grey band) under both extraction methods.



(a) SAE features (Llama-3.1-8B) (b) Linear probes (Qwen2.5-3B)

Figure 11: Cross-architecture replication. Left: SAE features on Llama-3.1-8B ($p = 3.6 \times 10^{-27}$). Right: linear probes on Qwen2.5-3B (Gini $= -0.185$, $p < 10^{-15}$).

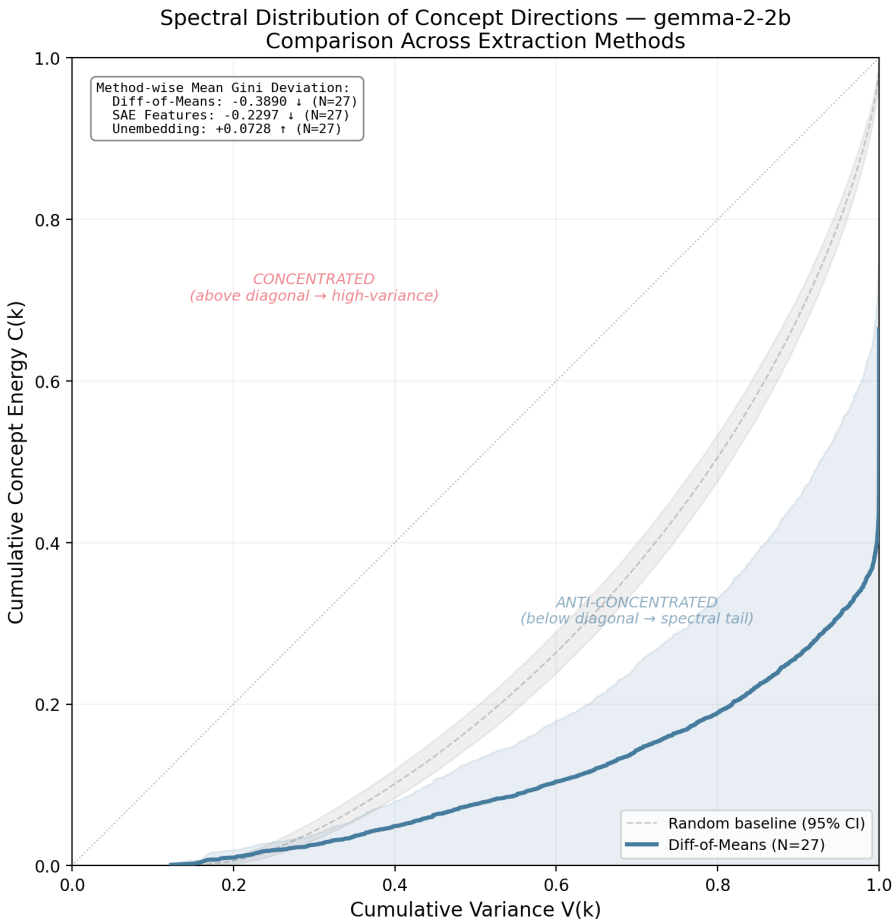


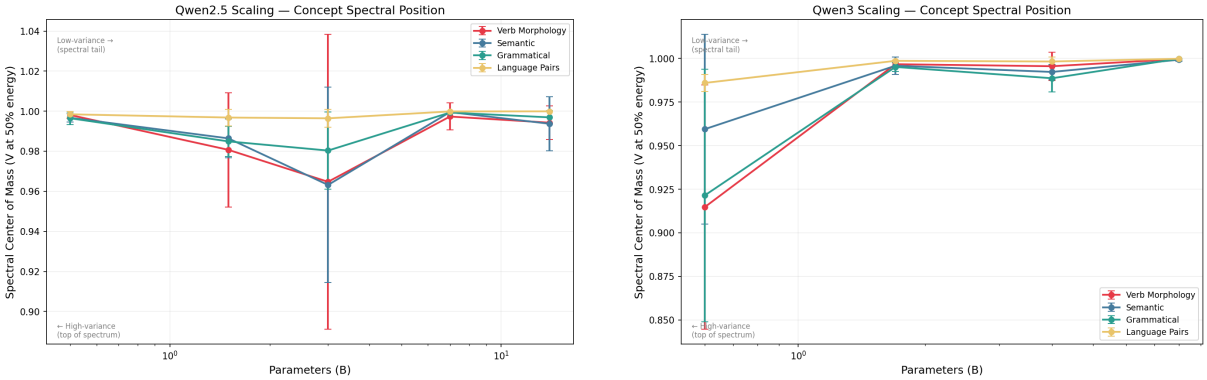
Figure 12: Cross-method spectral energy CDFs on gemma-2-2b: difference-of-means, SAE features, and unembedding-row contrasts. Activation methods fall below the diagonal; unembedding rows flip sign—the dual geometry.

Table 8: Dual geometry: concept-specific unembedding-row pairs concentrate *more than random* token pairs.

Model	**Random Gini**	**Concept Gini**	p
Qwen2.5-0.5B	-0.095	$+0.046$	$< 10^{-4}$
Llama-3.2-1B	-0.124	$+0.103$	$< 10^{-4}$
Gemma-2-2b	-0.117	$+0.073$	$< 10^{-4}$
SmolLM2-1.7B	-0.071	$+0.056$	$< 10^{-4}$

Table 9: Qwen 2.5 and Qwen 3 scaling. Gap is concept SCM minus random; Gini gap is concept minus random.

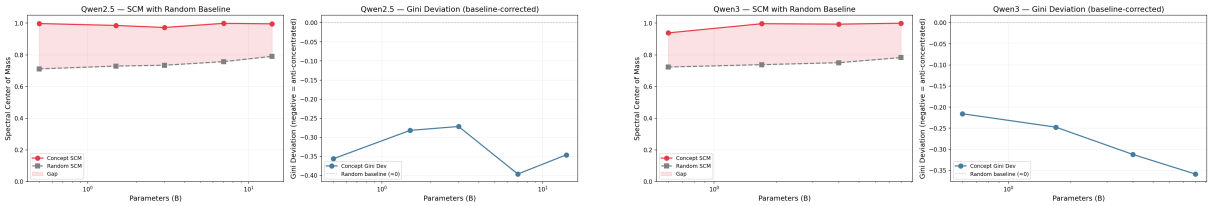
Model	**Concept SCM**	**Random SCM**	**Gap**	**Concept Gini**	**Random Gini**	**Gini Gap**
Qwen 2.5						
0.5B	0.997	0.711	$+0.286$	-0.356	-0.161	-0.195
1.5B	0.986	0.729	$+0.257$	-0.282	-0.172	-0.110
3B	0.973	0.735	$+0.238$	-0.272	-0.175	-0.097
7B	0.999	0.757	$+0.242$	-0.396	-0.186	-0.210
14B	0.996	0.790	$+0.205$	-0.346	-0.207	-0.139
Qwen 3						
0.6B	0.938	0.723	$+0.215$	-0.216	-0.170	-0.046
1.7B	0.997	0.738	$+0.259$	-0.248	-0.177	-0.070
4B	0.994	0.750	$+0.243$	-0.312	-0.186	-0.126
8B	1.000	0.783	$+0.217$	-0.359	-0.207	-0.152



(a) Qwen 2.5 (0.5B–14B)

(b) Qwen 3 (0.6B–8B)

Figure 13: Scaling behavior of spectral anti-concentration.



(a) Qwen 2.5 (baseline-corrected)

(b) Qwen 3 (baseline-corrected)

Figure 14: Baseline-corrected scaling. Concept (red) vs. random (grey) SCM gaps persist across scales; Gini deviations are stable rather than emerging or vanishing with capacity.

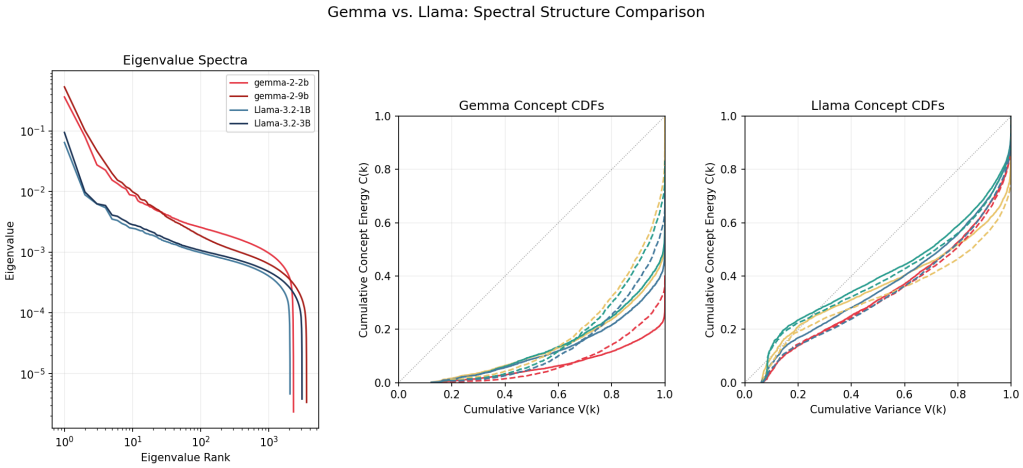


Figure 15: Architecture comparison: Gemma vs. Llama spectral energy CDFs. Despite Gemma's more aggressive eigenvalue concentration, both families show comparable anti-concentration of concept vectors.

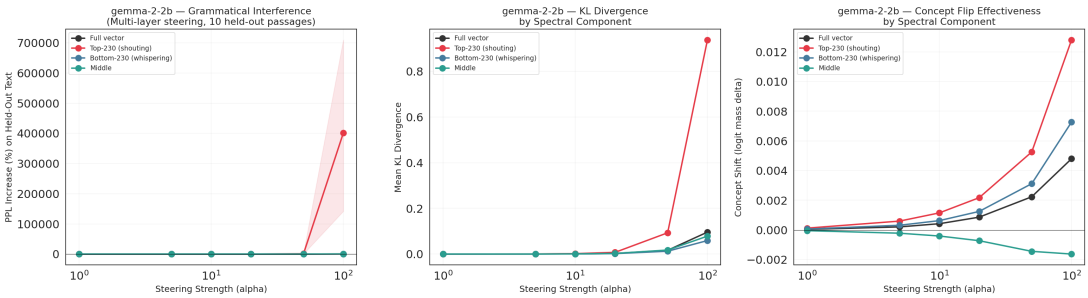
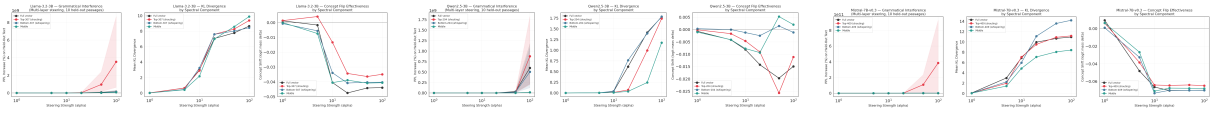


Figure 16: Split-injection on gemma-2-2b: PPL increase, KL divergence, and concept shift vs. α. The shouting component causes consistently more PPL degradation than the whispering component.

Table 10: Full split-injection results across five models. Bold rows mark representative α values used in the main-text summary; remaining rows show saturation or failure regimes from the full sweep.

Model	α	Shout PPL	Whisper PPL	d	p
gemma-2-2b	**10**	+1.0%	+0.5%	+0.86	**.0002**
	20	+4.8%	+1.8%	+1.12	$< 10^{-4}$
	50	+982%	+14%	+0.43	.13
Llama-3.1-8B	**5**	+383%	+223%	+1.80	$< 10^{-4}$
	10	+230K%	+90K%	+0.39	.18
	20	+238M%	+16M%	+0.93	.002
Llama-3.2-3B	**1**	+4.6%	+2.9%	+1.45	**.0001**
	5	+779%	+225%	+0.66	.024
Qwen2.5-3B	10	+11%	+13%	−0.21	.51
	20	+82%	+197K%	−0.61	.035
Mistral-7B	5	+18K%	+59K%	−0.32	.25
	20	+119M%	+839M%	−0.27	.34



(a) Llama-3.2-3B (b) Qwen2.5-3B (c) Mistral-7B-v0.3

Figure 17: Remaining split-injection panels. Llama-3.2-3B shows asymmetry at low α; Qwen2.5-3B and Mistral-7B do not—on Qwen2.5-3B, whispering causes *more* disruption at $\alpha \geq 20$.

Table 11: Full POS probing results. Variance columns: fraction of total eigenspectrum variance captured by each subspace. Random-k is the mean accuracy over 5 random subspaces of the same dimensionality.

Model	d	k	Top-k var.	Bot-k var.	Top-k	Bot-k	Rand-k	Full	Gap
gemma-2-2b	2304	230	39.8%	1.6%	.554	.506	.526	.813	+.048
Llama-3.2-3B	3072	307	32.0%	2.1%	.628	.576	.584	.835	+.052
Llama-3.1-8B	4096	409	27.5%	1.8%	.641	.625	.602	.833	+.016
Mistral-7B	4096	409	31.3%	1.5%	.624	.607	.606	.859	+.017
SmolLM2-1.7B	2048	204	27.4%	2.3%	.560	.536	.537	.781	+.024
Qwen3-4B	2560	256	28.4%	2.4%	.572	.555	.562	.826	+.018
Qwen2.5-3B	2048	204	27.0%	2.6%	.557	.576	.555	.787	−.019
Qwen2.5-7B	3584	358	26.2%	1.9%	.572	.586	.575	.813	−.014

Table 12: Chinese UD probing on Qwen models. Both show bottom-k > top-k, so the Qwen 2.5 English reversal is not explained solely by applying an English-derived covariance to a multilingual model.

Model	k	Top-k	Bot-k	Rand-k	Full	Gap	p
Qwen2.5-3B	204	.411	.462	.427	.692	−.051	.0001
Qwen3-4B	256	.432	.446	.449	.708	−.013	.014

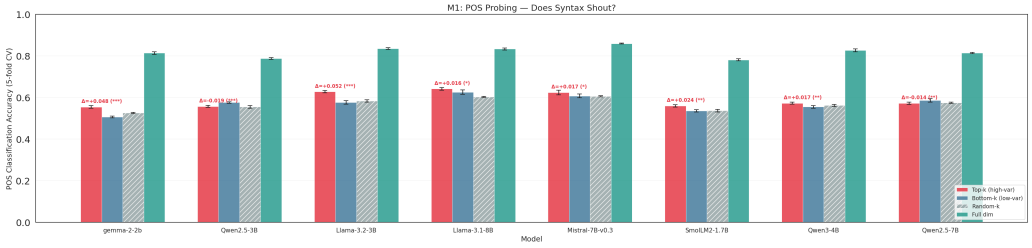


Figure 18: POS-tag probing across 8 models. Top-k (red) consistently outperforms bottom-k (blue) except on Qwen 2.5. Full-dimensional probes (green) serve as upper bounds (78–86%).

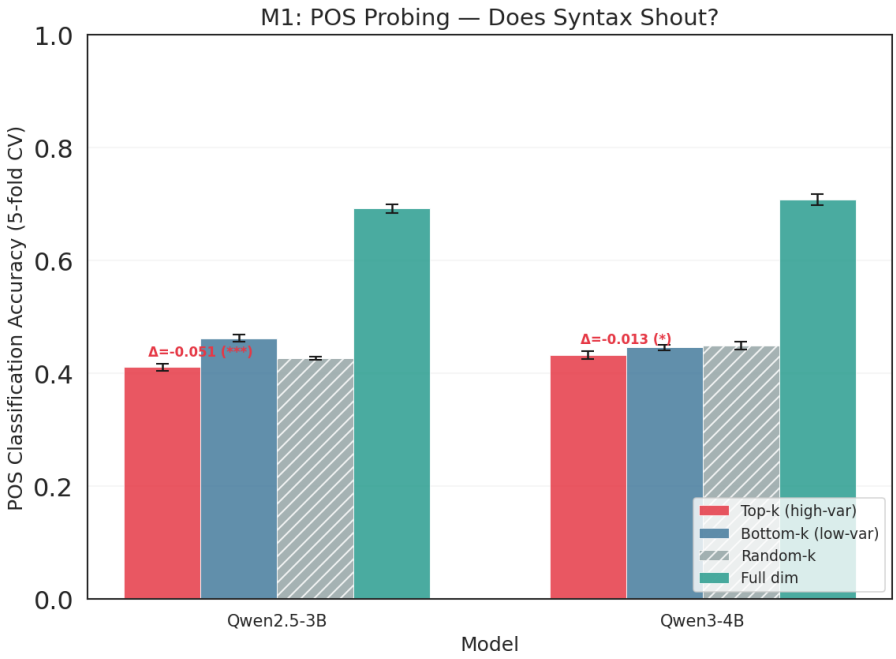


Figure 19: Chinese POS probing on Qwen models: both show bottom-k > top-k.

Table 13: POS-tag probing across spectral cutoffs $k \in \{5\%, 10\%, 20\%\}$: Top-$k$ and Bottom-k accuracies, Gap, and paired t-test p. The qualitative direction of the gap is invariant across cutoffs for all eight models.

	$k = 5\%$				$k = 10\%$				$k = 20\%$			
Model	**Top**	**Bot**	**Gap**	p	**Top**	**Bot**	**Gap**	p	**Top**	**Bot**	**Gap**	p
gemma-2-2b	.494	.442	$+.051$	.0005	.554	.506	$+.048$	.0001	.624	.592	$+.031$	.0006
Llama-3.2-3B	.565	.501	$+.064$	.0008	.628	.576	$+.052$	.0001	.688	.659	$+.028$	.0075
Llama-3.1-8B	.579	.566	$+.013$	.0634	.641	.625	$+.016$	.0121	.703	.690	$+.013$	.0653
Mistral-7B-v0.3	.566	.554	$+.013$	.0506	.624	.607	$+.017$	.0125	.681	.653	$+.028$	.0009
SmolLM2-1.7B	.493	.482	$+.011$	.0503	.560	.536	$+.024$	.0023	.604	.582	$+.023$	.0001
Qwen3-4B	.518	.495	$+.023$	.0052	.572	.555	$+.018$	.0012	.629	.619	$+.010$	.0482
Qwen2.5-3B	.506	.517	$-.011$	.0344	.557	.576	$-.019$	$< .0001$	.611	.631	$-.021$	.0004
Qwen2.5-7B	.516	.525	$-.009$	.1515	.572	.586	$-.014$	.0015	.623	.638	$-.014$	.0129