CROSS-MODEL SAE COACTIVATION SURVIVES LENGTH, TOPIC, AND LEXICAL CONTROLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-model matches of sparse-autoencoder (SAE) features have been judged against random pairings or neuron baselines, neither of which re-runs the pairing rule. Gemma-2-9B and Mistral-7B, from different families with different tokenizers, each have a released-SAE obituary feature, and the two share their top five documents: is it the same feature? Our controls support a more specific answer: matched features of these two models respond to the same inputs beyond chance and pair selection, and that shared response survives the topic and vocabulary controls we ran; this is coactivation, not feature identity. Such matches are cheap: two sets of $1{,}500$ features reciprocally match about half of each other under idealized chance, and any document property both models encode (length, topic, vocabulary) can make features coactivate. We run the pairing rule inside a document-permutation null, score frozen pairs on documents the rule never saw, and add controls that each fix one named nuisance. In a non-pre-registered follow-up, all nine cross-family cells (Gemma-2, Mistral, Qwen3, Llama-3) hold on disjoint C4 text ($p = 1/2001$ each, Bonferroni over the 17-cell roster). The sharing survives length- and coarse topic-matched nulls (one cell); on three representative cells, every control leaves at least 0.88 of the excess its reference arm leaves. Regressing out a unigram TF-IDF lexical model, which a pre-registered gate shows is not inert, leaves 92–94% of the excess: the shared-vocabulary explanation is narrowed, not closed. A registered token-aligned test shows the sharing is not reducible to document means (it beats a within-document word-order shuffle; mean pair correlation $0.522/0.398/0.335$) and, in a leakage-clean design, keeps $R^{\text{clean}}_{\text{tok}}$=0.92–0.99 of its token-level excess after conditioning on the document and the current word. The claim does not extend to compositional or syntactic sharing, a capability trend (swapping only one model's SAE width moves its cell from $z = 22.1$ to 14.1), or shared computation (both cross-model ablation probes were inconclusive). The controls can say no: they rejected three decisive-looking secondary results, all reported (a cross-lingual metric that tracks network depth rather than language, its depth-corrected residual, and a 70M-to-7B match null on natural text). Cross-model matches should be scored against a null that re-runs the matching, on documents the matching never saw, one nuisance per control; code, nulls, receipts and registrations are in the anonymous supplement, to be released upon publication.

1 INTRODUCTION

Feature-universality studies pair sparse-autoencoder (SAE) features of separately trained networks by their highest activation correlation (Bricken et al., 2023; Wang et al., 2025; Lan et al., 2024), and read close pairs as evidence of shared features. On C4, the released SAEs of Gemma-2-9B and Mistral-7B, two families with different tokenizers, each have an obituary feature, and the two features' five most-activating documents are the same five (Section B.1). Such pairs are cheap to manufacture: two sets of $1{,}500$ features reciprocally match about half of each other under idealized chance (Proposition 1), and any document property both models encode (length, topic, vocabulary, the current word) can make features coactivate with no shared feature behind them.

Sparse autoencoders and transcoders (Cunningham et al., 2024; Bricken et al., 2023; Ameisen et al., 2025; Lindsey et al., 2025) make such comparisons easy to run, and cross-model matches are read

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

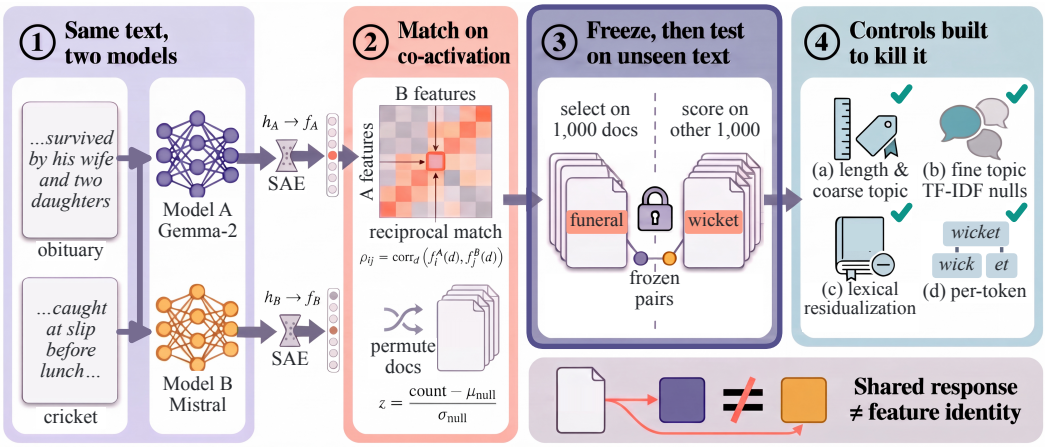

Figure 1: **Cross-model SAE coactivation survives a null that re-runs the matching, disjoint held-out documents, and controls for length, topic and vocabulary; what survives is a shared response to the same inputs, not a shared feature.** (1) Each model's own released SAE encodes the same C4 documents. (2) Reciprocal best matches are scored against a document-permutation null that re-runs the matching. (3) Non-pre-registered: pairs selected on 1,000 documents are frozen and scored on a disjoint 1,000 (all nine cells). (4) Length nulls on six cells, coarse-topic on five (partly on separately extracted features); fine-topic, lexical and token-aligned controls on three representative cells only. A tick means the shared excess survived that control, not that the nuisance is excluded. Schematic only; data in Figure 2 and Table 1.

as evidence of representational convergence (Huh et al., 2024; Lan et al., 2024). The published cross-model protocol pairs features greedily and judges significance against random pairings, on same-tokenizer pairs and without controls for shared document properties (Lan et al., 2024). A pairing rule fit to the data inflates any statistic computed on the same data (Kriegeskorte et al., 2009): re-implemented without their filters on three of our pairs (all but GPT-2↔Mistral-7B, which shows no signal above random), a null that permutes documents before the pairing step sits far above that random baseline, so part of the signal above random (up to 94% on templated prompts, less on C4) comes from the pairing itself (Section M). With their filters added, that share shrinks on templates but does not vanish, and on C4 too few permuted pairs survive the filters to estimate it (Section M.1). That protocol lacks a null that re-runs the pairing rule, pairs scored on documents the rule never saw, and controls that hold named mundane causes fixed.

We add all three, on open-weight models, and apply controls built to kill the signal (Figure 1). It survives. In all nine cross-family cells across Gemma-2, Mistral, Qwen3 and Llama-3, pairs selected on half of a C4 slice still beat their null on the disjoint other half (each at the floor of its permutation test; this follow-up was not pre-registered). On three representative cells, fine topic strata, cross-fitted lexical residualization under three pairing rules, and registered current-word conditioning at token level each leave a retention ratio of at least 0.88 against its reference arm. What survives is coactivation on shared inputs, not feature identity: both cross-model ablation probes were inconclusive, and a dictionary confound (cells differ in SAE dictionary, not only in model) blocks any capability-trend reading. The instruments behind the positive can return a no: size- and sparsity-matched random directions fall below their own null, Pythia-70M↔Mistral-7B falls to null on C4, and on planted data the token-aligned analysis collapses a unigram-only signal while planted context survives (Section L).

Contributions.

1. **A reusable test that separates shared responses from named mundane correlates.** A re-matching permutation null, frozen pairs scored on disjoint documents, and controls read as retention ratios against a same-sized reference arm; two propositions show why, for any pairing rule, the null and frozen score carry no selection optimism while a re-selected score does (Section 3).

2. **A scoped cross-model positive.** Coactivation on shared inputs in nine cross-family cells on held-out documents (non-pre-registered), three of them through the full control ladder (Figure 2, Table 1); not feature identity, composition, a capability trend, or shared computation.
3. **Three rejections, each with the practice it implies.** The controls reject a Pythia-70M↔Mistral-7B match null on natural text (so recheck template matches on natural text), a depth-driven cross-lingual metric (so add a mismatched-concept control), and its depth-corrected residual at 206 concept pairs and 17 models (so use several English-only baselines) (Sections 5.1 and 5.2).

2 Related Work

Sparse dictionaries and attribution graphs. Sparse autoencoders decompose residual streams into more monosemantic features (Cunningham et al., 2024; Bricken et al., 2023; Rajamanoharan et al., 2024; Templeton et al., 2024); transcoders and cross-layer transcoders predict across layers so that direct linear effects can be computed, and `circuit-tracer` turns them into attribution graphs (Ameisen et al., 2025; Lindsey et al., 2025; Hanna et al., 2025). Released dictionaries now exist for Gemma-2, Gemma-3 and Qwen3 (Lieberum et al., 2024; McDougall et al., 2025; Hanna, 2025b; BluelightAI, 2025), with `circuit-tracer` configurations for the Gemma-3 transcoders (Hanna, 2025a). We train no dictionary: every claim attaches to a (model, released SAE) pair, which is what a practitioner reusing these artifacts would compare.

Convergent units and convergent geometry. Bipartite unit matching across independently trained vision networks began with convergent learning (Li et al., 2016); Rosetta neurons recur across vision models (Dravid et al., 2023), with counts that grow sublinearly in scale (Dravid et al., 2026), and only a small fraction of GPT-2 neurons are universal (Gurnee et al., 2024). At the level of global geometry, Huh et al. (2024) argue that capable models converge, while Gröger et al. (2026) show that such alignment scores are scale-confounded and that only local-neighborhood agreement survives permutation calibration. Circuit-level recurrence is documented for vision circuits, induction heads and toy tasks (Olah et al., 2020; Olsson et al., 2022; Chughtai et al., 2023), while behavior tied to a single concept can still differ by family (Soligo et al., 2026). These works ask whether units or geometries recur; we ask a narrower question: do post-hoc features of LMs with different tokenizers respond to the same documents once named document covariates are held fixed?

Matching or mapping features across models. Bricken et al. (2023) match features across seeds by activation correlation, and Lan et al. (2024) pair features across models and compare the paired decoder spaces with SVCCA/RSA (Raghu et al., 2017; Kriegeskorte et al., 2008) against random pairings. Wang et al. (2025) match SAE features of Pythia-160M and Mamba-130M, which share a tokenizer and training corpus, by maximum pairwise correlation over tokens, and read the result against a neuron baseline and seed-variant skylines rather than a null that re-runs the matching. Crosscoders and universal SAEs instead learn a shared dictionary during training (Lindsey et al., 2024; Thasarathan et al., 2025), and their shared/exclusive diagnostic needs its own correction (Minder et al., 2025). A third line connects models through learned maps: stitching (Bansal et al., 2021), permutation re-basin (Ainsworth et al., 2023), and affine residual maps that carry SAEs, probes and steering vectors (Chen et al., 2025; Huang et al., 2025), with success depending on scale and tokenizer overlap (Gorbett & Jana, 2026); steering vectors also transfer across architectures through learned nonlinear bridges between the models' SAE feature spaces (Agarwal, 2026), and relative representations avoid a learned map altogether (Moschella et al., 2023). A learned map can succeed without any shared sparse feature, and a pairing rule fit to the data inflates any statistic computed on the same data. We therefore use no learned map in the headline, put the pairing rule inside the null, and score frozen pairs on documents the rule never saw. SVCCA/RSA appear only in our audit of Lan et al. (2024) (Section M): a global score cannot say which features correspond, and a crosscoder's shared latents would need retraining for every null draw (Section R).

Cross-lingual mechanisms. Templeton et al. (2024) report SAE features that fire across languages in Claude 3 Sonnet, and Lindsey et al. (2025) extend this to circuits in Claude 3.5 Haiku; both are closed-weight. In open models, logit-lens decoding suggests an English-centered latent (Wendler et al., 2024), some SAE features are language-specific in that ablating them harms one language (Deng et al., 2025), answer consistency comes apart from representation sharing (Ifergan et al.,

2025), sharing depends on model and architecture (Bogdanov & Huang, 2026), and mid-layer overlap within Gemma-2 is reported by Son et al. (2025). We reuse this question in open weights mainly as a test case for the control battery, not as the headline.

Construct validity and circular analysis. Representational-similarity metrics can be high for reasons unrelated to what they claim to measure (surveyed by Klabunde et al., 2025): Ding et al. (2021) ask for sensitivity and specificity and show that CKA and CCA (Kornblith et al., 2019) can fail them, Davari et al. (2023) show CKA can be manipulated without changing model function, Korznikov et al. (2026) find random directions rivalling trained SAE features, and SAEs on randomly initialized transformers can match trained ones on common interpretability metrics (Heap et al., 2025). Scoring a selection on the data that made it is the circular analysis of Kriegeskorte et al. (2009); plus-one permutation p-values are the standard fix for Monte-Carlo nulls (Phipson & Smyth, 2010). We carry both into feature matching: a count is read only against a null that re-runs its own selection, and then against nulls that hold named covariates fixed.

3 Problem Setup

To tell a match like the obituary pair from chance and from shared document properties, the statistic needs a known chance level, a null that re-runs its own selection, and controls read against a reference arm.

Coactivation. Let A and B be two models, each with one released SAE at a mid-layer residual site, and $D = \{d_1, \dots, d_N\}$ a shared corpus. For feature i of A's SAE, $f_i^A(d)$ is its activation mean-pooled over A's tokens of document d, one scalar per (document, feature), which sidesteps the tokenizer mismatch. We keep the K=1,500 highest-variance features per model, standardize each over D to $z_i^A(d)$, and define

$$\rho_{ij} \;=\; \tfrac{1}{N}\textstyle\sum_{n=1}^{N} z_i^A(d_n)\, z_j^B(d_n), \qquad i \le K_A,\ j \le K_B, \tag{1}$$

the correlation of the two features across documents. Here "sharing" denotes residual dependence in ρ under the tested controls; it is not feature identity and not shared computation.

Reciprocal matching and its null. With $a(i) = \arg\max_j \rho_{ij}$ and $b(j) = \arg\max_i \rho_{ij}$, the reciprocal (mutual-nearest-neighbor) pairs and their count are

$$\mathcal{M}(\rho) = \{(i,j) : a(i) = j,\ b(j) = i\}, \qquad M = |\mathcal{M}(\rho)|. \tag{2}$$

The null permutes documents on one side: for π uniform on S_N, ρ^π is (1) with $z_j^B(d_{\pi(n)})$, and $M^\pi = |\mathcal{M}(\rho^\pi)|$ *re-runs the matching*. From draws $M^{(1)}, \dots, M^{(n_{\text{perm}})}$ we report

$$\Delta = M - \overline{M^\pi}, \qquad z = \Delta / \operatorname{sd}(M^\pi), \qquad p = \frac{1 + \#\{b : M^{(b)} \ge M\}}{n_{\text{perm}} + 1}. \tag{3}$$

Significance rests on p (Phipson & Smyth, 2010), never on a Gaussian reading of z. Two facts fix the design (proofs in Section C).

Proposition 1 (Chance reciprocal matches; validity of the re-matching null)**.** *(a) If, for every (i,j), the K_A+K_B−1 entries in row i and column j of ρ are exchangeable and almost surely distinct (e.g. i.i.d. continuous), then $\mathbb{E}[M] = K_A K_B/(K_A{+}K_B{-}1)$, i.e. ≈ 750 at K_A=K_B=1,500. (b) Permuting documents on one side leaves both within-model correlation matrices unchanged. If H_0 holds, i.e. $(z^A, z^B) \overset{d}{=} (z^A, z^B \circ \pi)$ for every $\pi \in S_N$, then every statistic of (z^A, z^B), including one that selects its own pairs, has $\Pr(p \le \alpha) \le \alpha$ for all α.*

By (a), about half of the 1,500 screened features per model are reciprocally matched under idealized chance, more than the raw count of our weakest surviving cell (Table 1); the realized chance level also depends on each dictionary's internal correlations, which the null preserves, so every cell gets its own null and z is never compared across cells. By (b), rejecting H_0 shows dependence between the models on shared documents, not correspondence: any document property both models encode (length, topic, vocabulary) also violates H_0. That is what the controls are for.

Held-out frozen pairs. Split the corpus into disjoint halves D_S, D_E (1,000 C4 documents each). Pairs $\widehat{\mathcal{P}} = \mathcal{M}(\rho^S)$ selected on D_S are frozen and scored on D_E by $T_E(\mathcal{P}) = |\mathcal{P}|^{-1} \sum_{(i,j)\in\mathcal{P}} \rho^E_{ij}$, with excess Δ_E and p from 2,000 permutations of D_E alone. The ratio $r = \Delta_E(\widehat{\mathcal{P}})/\Delta_E(\mathcal{M}(\rho^E))$ compares against pairs re-selected on D_E, a different pair set, so it is not within-pair retention.

Proposition 2 (Frozen pairs carry no selection optimism)**.** *Let documents be i.i.d., $D_S \cap D_E = \emptyset$, and $\widehat{\mathcal{P}}$ depend on D_S only. Conditional on D_S: (a) $\mathbb{E}[T_E(\widehat{\mathcal{P}}) \mid D_S] = |\widehat{\mathcal{P}}|^{-1} \sum_{(i,j)\in\widehat{\mathcal{P}}} \mathbb{E}[\rho^E_{ij}]$, as for a pair set fixed in advance; (b) the permutation p on D_E is valid under H_0 restricted to D_E. Pairs re-selected on D_E and held fixed in the null, as in the denominator of r, need not have either property: $\mathbb{E}[\max_j \rho^E_{ij}] \geq \max_j \mathbb{E}[\rho^E_{ij}]$; re-selecting in every draw restores (b) by Proposition 1(b).*

This is what lets the nine held-out cells carry the positive, conditional on its i.i.d. premise: our C4 slice is the first 2,000 eligible streamed documents, split into consecutive halves. On five cells, seeded random halves change the median held-out excess by at most 3% (separately extracted features, Section L); on the three ladder cells, a seeded random split of 20,000 documents keeps 0.97–0.99 of in-sample excess held out (Section Q); the Qwen3-8B cells are untested. The follow-up was not pre-registered, and its screen saw both halves (Table 4).

Controls as retention ratios. A control either restricts π to permutations within strata of a covariate (valid under within-stratum exchangeability) or removes the covariate from the features before frozen pairs are scored. Removing *anything* costs excess mechanically, so each control is divided by a reference arm: a same-sized transformation that carries no information about the covariate or, for tokens, the unconditioned statistic on the same occurrences. We report

$$R = \Delta_{\text{control}} \,/\, \Delta_{\text{calibration}}, \tag{4}$$

so $R \approx 1$ means the covariate, as the control captures it, explains none of the shared excess and $R \to 0$ all of it; forms of the covariate the control does not capture are not excluded. Instances: k-way TF-IDF topic strata vs. random strata of identical sizes (R); cross-fitted residualization on a TF-IDF→SVD$_k$ design vs. a row-shuffled design of identical spectrum (R_{lex}); and, at word granularity, document plus word-type fixed effects vs. a within-type derangement (R_{tok}: conditioned over unconditioned token-level excess on common support). The secondary cross-lingual measure, probe-attribution alignment (mid-layer logit-lens profiles correlated across translated prompts), is defined in Section H. Neither uses attribution-graph topology, which our v0 pipeline did not recover (Section 6).

4 Experimental Setup

Models, prompts, and compute. Two arms share one control discipline: SAE matching across models, and cross-lingual probe attribution within them. The SAE-matching arm uses Gemma-2-2B/9B (Gemma Team et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen3-1.7B/8B-Base (Yang et al., 2025), Llama-3.2-1B (Grattafiori et al., 2024), Pythia-70M (Biderman et al., 2023) and GPT-2 (Radford et al., 2019), each with one released SAE (Table 8), on a shared 206-prompt English concept corpus and on C4 (each "document" is the first 300 characters of a record, so full-length documents are untested; Section E). The probe-attribution sweep uses the eight (model, transcoder) pairs of Table 5 plus Gemma3-1B and Pythia (70M–1.4B), on 20 aligned antonym pairs per language in en/zh/ja/ko, expanded to 206 concept pairs across nine concept families and 17 models for the residual scale-up. Prompt lists, target tokens and provenance are in Section E; each model scores the first token of its own tokenization of the target. We call the abandoned attribution-graph pipeline *v0* and the functional pipeline behind every result *v1*. Compute is in Section F. Intervals are 95% bootstrap CIs (10,000 resamples; model/data seeds 0, bootstrap seed 1).

Statistical reporting. The unit of inference is stated with each result: language pairs (raw probe-attribution), models and pairs (residual scale-up), and (model pair, corpus) cells (SAE matching). All p-values use (3). The same-corpus 200-draw C4 screen is descriptive; the non-pre-registered disjoint-half follow-up uses 2,000 draws and all nine tests clear Bonferroni over the 17 C4 cells ($1/2001 < 0.05/17$). Other verdicts use registered survive/fail controls, not marginal significance; cross-lingual gradations are descriptive.

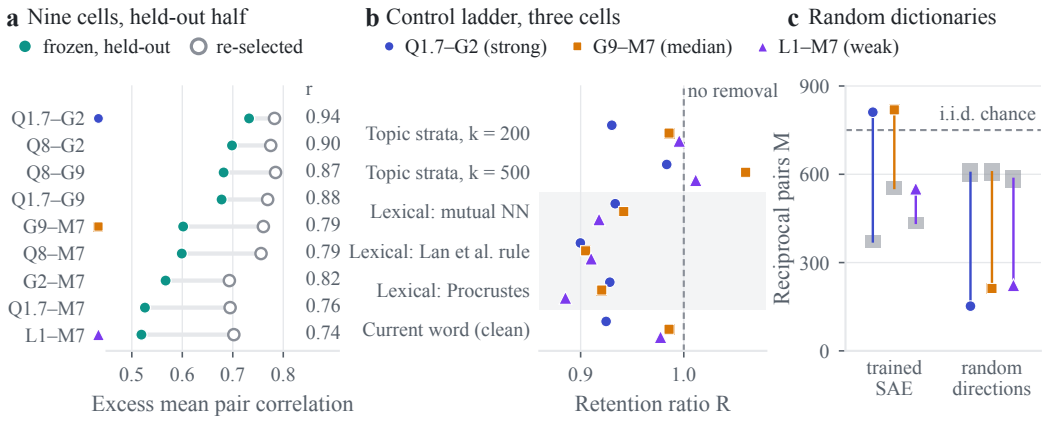

Figure 2: **The coactivation survives disjoint documents in all nine cells, and on three representative cells every control in (b) has retention ratio $R \geq 0.88$ against its reference arm;** length-decile nulls, run on six of the nine cells, leave at least 0.96 of the excess (Section L). (a) Pairs selected on 1,000 C4 documents and frozen (filled) vs. pairs re-selected on the evaluation half (open); r is their ratio of excesses, over two different pair sets (Proposition 2); every held-out $p = 1/2001$. (b) Every verdict-bearing retention ratio R of (4) on one axis (the leaky first-design R_{tok} is omitted; token row: $R_{\text{tok}}^{\text{clean}}$; NN: nearest neighbour); $R = 1$ means the control removes nothing beyond its reference arm (shaded: lexical pairing rules). (c) Reciprocal count M (marker) against its own permutation null (grey, mean$\pm$2 SD) for trained SAEs and size/sparsity-matched random directions; dashed: the i.i.d. chance level of Proposition 1(a). Q/G/M/L = Qwen3/Gemma-2/Mistral/Llama-3.2, number = size (B). Drawn by `code/figures/main_results_20260924.py` from `heldout_*`, `covctl_k*_c4_*`, `lextok20k_*`, `lanpair_*`, `proclex_*`, `tokaclean20k_*`, `c4_*` and `rand_c4_*.json`.

5 Results

5.1 Cross-model SAE feature matching

We run the pipeline of Section 3 on the pairs of Table 8, each model with *one* publicly released SAE (`sae-lens`, Bloom et al., 2024; Gemma Scope's is read one block before its training hook, Section P) and mean-pooled features, so claims attach to (model, SAE, read point, pooling) (Section Q).

Finding 1: the natural-text control decides which cells merit held-out testing. On the templated concept corpus every tested pair among Gemma-2, Mistral-7B and Pythia (models from 70M to 9B) clears its reciprocal null (Table 10). The 2,000-document C4 screen (Raffel et al., 2020) then separates candidates from a corpus artifact (Table 12): Pythia-70M↔Mistral-7B, a cross-family cell strongly significant on templates, *falls to null on C4*; its sharing was template-tuned. The modern $\gtrsim$1B pairs survive, and Qwen3-1.7B/8B (Qwen-Scope SAEs, Deng et al., 2026; manual adapter) share with both other families, as does Llama-3.2-1B (Chanin, 2024) against Mistral-7B (SAE: Engels, 2024): nine cross-family cells across four families. In the non-pre-registered held-out follow-up, freezing first-half pairs and scoring them on the second half confirms all nine at $p = 1/2001$, which clears Bonferroni over all 17 C4 cells considered; held-out excess is 74–94% of the eval-half re-selected excess, a ratio over different pair sets, not within-pair retention (Figure 2a, Table 4). The templated corpus errs in both directions: Qwen3-1.7B↔Gemma-2-2B is weak on templates ($z = 3.59$, $p = 1/201$) and strong on C4 ($z = 42.7$), the mirror image of the Pythia artifact, hence the natural-text headline.

What we do *not* claim: no capability trend. Against the fixed Mistral-7B reference, z rises with size across Pythia-70M, Llama-3.2-1B, Qwen3-1.7B, Gemma-2-2B and Gemma-2-9B, which invites a reading as emergence with capability and an SAE-level Platonic Representation Hypothesis (Huh et al., 2024). The axis is confounded with the dictionary: these five cells pair Mistral's SAE with four other SAE suites, and swapping only the Gemma-2-2B dictionary 16k→65k moves its

Table 1: **The control ladder: each rung is a way to explain the shared excess away, and on three representative cells none does.** Values are strong/median/weak (Qwen3-1.7B↔Gemma-2-2B, Gemma-2-9B↔Mistral-7B, Llama-3.2-1B↔Mistral-7B). Ratios follow (4); lexical rows use the governing k=4096. "reg.": rule and outcome branches registered before the receipt existed; the held-out follow-up was not. Full tables: Tables 4, 11, 15 to 17, 21 and 22.

Rung	Statistic	strong/median/weak	p
C4 screen, 2,000 docs	M (null mean)	811 (371.8)/819 (553.4)/551 (434.9)	$1/201$
Random directions (size, L_0 matched)	M (null mean)	152 (604.7)/212 (607.0)/223 (584.5)	1.0
Held-out frozen pairs, 1,000+1,000	Δ_E	0.732/0.602/0.519	$1/2001$
ratio to re-selected pairs	r	0.936/0.791/0.739	—
Topic strata, k=200	R	0.930/0.986/0.996	$1/2001$
Topic strata, k=500	R	0.983/1.060/1.012	$1/2001$
Lexical residualization, mutual NN (reg.)	R_{lex}	0.933/0.942/0.918	$1/2001$
Lan et al. pairing rule (reg.)	R_{lex}	0.900/0.905/0.910	$1/2001$
Procrustes pairing, disjoint half (reg.)	R_{lex}	0.928/0.920/0.885	$1/2001$
Token-aligned, within-document order (reg.)	S_1^{clean} (mean r)	0.522/0.398/0.335	$1/2001$
Current-word type, first design (leaky, audit)	R_{tok}	1.064/1.301/1.307	—
Current-word type, leakage-clean (reg.)	$R_{\text{tok}}^{\text{clean}}$	0.925/0.986/0.977	$1/2001$

cell from $z = 22.1$ to 14.1 with no change of model (Table 12). What survives is the per-cell verdict, which needs no cross-cell comparison. GPT-2's four natural-text cells are underpowered (Section Q), so exactly one model is measured below null with power. On the templated corpus both Gemma↔Mistral pairs are significant at all three depth bands (Table 10); on C4 each model is read at one mid-layer hook.

Random-dictionary control: informative, but its z is not an effect size. Korznikov et al. (2026) report random directions rivalling trained SAE features. Replacing each dictionary with size- and sparsity-matched random directions collapses the reciprocal count while its permutation null *rises* (Figure 2c, Table 1), so z lands well below zero. We pre-registered two readings ($z \approx 0$, or z stays high) and the outcome matched neither; as both the count and its null move, the negative z is not an effect size. Narrow claim: matched random projections do not reproduce reciprocal structure.

The strongest control: lexical residualization at interpretable corpus size. The remaining alternative is shared *lexical* detectors. An earlier attempt on 2,000 documents could not adjudicate: its nuisance model explained almost nothing, and an intervention that removes nothing excludes nothing. The rerun on 20,000 C4 documents (rule, thresholds and both outcome branches registered first) adds a hard adequacy gate against a spectrum-matched shuffled design and scores the *originally matched* pairs, frozen on the raw features, rather than a rematch that could reconstitute the total from different pairs (Section Q). The gate passes at every k, with nearly every feature beating the shuffled design, yet the frozen pairs keep nearly all of their excess (Table 1). The gate's ΔR^2 is a margin over a shuffled design, *not* variance captured. The nuisance model is a stopword-stripped, SVD-truncated *unigram* TF-IDF design (Table 16), and a lexical model whose median per-feature R^2 tops out at 0.14 captures almost none of what the two models share. The shared-unigram hypothesis is therefore narrowed at the ranks tested, not closed.

The same control under other pairing rules. The survival could belong to our reciprocal criterion rather than to the shared structure. We therefore transplant the pairing rule of Lan et al. (2024) (each A feature to its best B feature, many-to-one) into the identical pipeline, with five outcome branches registered pre-receipt and a reimplementation gate read first: the 1–1 fraction must fall inside the interval implied by their Tables 5–6, and it does in all three cells. Their rule retains slightly less than ours in every cell, far above the registered 0.30 collapse bar (Table 21, Section M). An orthogonally aligned pairing fitted on a disjoint half and scored on the other also survives (Table 22). Survival is a property of the shared structure, not of one criterion.

Token-aligned: the registered follow-up. Pooling makes the headline order-insensitive by construction, so we registered a token-aligned test (three first-class branches, committed reader) before

extracting per-token data. Per-token activations on the same $20{,}000$ documents are aligned across tokenizers at whitespace-word granularity (Section L). S_1, the frozen pairs' within-document centered correlation, is a quantity the pooled statistic cannot contain, and it beats a within-document word-order-shuffle null in all three cells: token-aligned sharing is not reducible to document means. S_2 conditions on the current word (type fixed effects against a within-type derangement). The first design's $R_{\text{tok}} > 1$ is reported but not verdict-bearing, for selection leakage and non-commuting centerings (Section L). The registered leakage-clean rerun closes both routes by construction: disjoint selection/evaluation halves, a universe re-screened on the selection half, a gated token stream with an aliasing audit, joint fixed effects re-run under every null draw, and one common scored-pair mask (counts in Table 17; split/mask hashes in the `tokaclean20k_*` receipts). Under it, the shared occurrence-level variation survives joint document and normalized-current-word-type fixed effects in all three cells (Table 1); the bigram arm is not selection-clean; no syntax or composition claim.

What do the matched features respond to? Common lexical and topical document properties (Section B.1); this does not establish semantic identity.

Finding 2: the loose criteria are confounded. Greedy-max excess saturates for most modern and GPT-2 pairings, including GPT-2↔Mistral-7B, which has no reciprocal sharing above null; Hungarian assignment (Kuhn, 1955) ranks Gemma-2-2B↔GPT-2, with few reciprocal matches, above every genuinely sharing Gemma↔Mistral pair. The null does not remove this: optimal assignment over all screened features finds matches driven by prompt-level covariates that pooling preserves (Table 10 and Figure 7).

Finding 3: the two ablation probes are inconclusive; within-model steering promotes overlapping tokens. Single-feature ablation is degeneration-prone (an apparently decisive Pythia↔Mistral result is entirely a degeneration artifact), and group ablation is confounded by a prompt-fragility axis so broad that even the within-family positive control fails (Table 18, Section K). *No cross-model causal intervention succeeded.* Within-model steering escapes both confounds: steer each matched feature in its own model ($h \mathrel{+}= \beta s_f W_{\text{dec}}[f]$) and compare the identity of the top-30 promoted token strings. Matched pairs share more strings than a derangement null (1.26 vs. 0.01 of 30 at 50 pairs and β=2); the excess holds at 20 pairs for $\beta \in \{2, 8\}$ and replicates on Qwen3-1.7B↔Gemma-2-2B (Section K); the cricket-matched pair promotes *batsman/batting/cricket* in Gemma and *batt/caught/declared* in Mistral. This is activation-level evidence of similar functional roles at our read point, not cross-model intervention evidence.

Finding 4 (templated prompts only): cross-lingual sharing is weak within models and language-conditional across them. Matching a model's English features against its own features on aligned zh/ja/ko prompts (Table 19), within-model sharing is *weak and graded*: significant for en–ja and en–zh in Gemma-2-9B, marginal in Gemma-2-2B, *null for en–ko in both*, while every matched-N cross-model English cell clears its null. Gemma Scope (Lieberum et al., 2024) is trained on an English-dominant mix, so weak en–X may partly reflect dictionary coverage. Across models sharing is *language-conditional*: on Chinese the within-family Gemma pair shares strongly (shared tokenizer, flagged) while Gemma-2-2B↔Mistral-7B, significant on English, collapses below null (Table 19). These cells use the templated corpus, which Finding 1 shows can mis-estimate sharing in both directions; they are descriptive.

Finding 5: zero-shot probe transfer is in-sample only. A concept probe trained on one model's matched features and applied unchanged to the other transfers perfectly on all three modern pairs when the matching is fit on the same templated prompts (controls near chance), but falls to chance when the matching is fit on independent C4 text, so the in-sample transfer does not by itself establish a shared interface (Section O).

5.2 Cross-lingual signals under controls

The raw probe-attribution metric measures depth, not language. All three controls converge (Table 7; trajectories in Section B.2). Layer-shuffle collapses the mean cross-lingual r to about zero; the mismatched-concept control keeps most of r, with Pythia-410M keeping all of it despite being a predominantly English model; and the required class ordering is inverted. The raw metric

fails construct validity: it captures the shape of a model's per-layer logit-lens trajectory, not cross-lingual transfer. The depth-corrected *residual* (real − mismatched) appears to separate the classes at $n = 20$ antonym pairs, with Pythia as the sole English-only baseline.

Scaling the controls: the residual is not cross-lingually specific. With 206 concept pairs, 17 models and three further English-only families (Table 9, Section B.3), the gap disappears: Pythia was the one English-only family with a near-zero residual, and GPT-2's residual is the largest of all 17 models. We report all three tests because they disagree. The bootstrap interval for the difference (multilingual minus English-only) lies entirely below zero, nominally in the *reverse* direction; the model-as-unit exact permutation test over all class assignments is far from significant in the pre-registered direction and two-sided, and so is the pair-level test (Table 9). *No test detects a multilingual advantage; the only interval that excludes zero (the bootstrap, whose unit, concept pairs, is wrong for a claim about model classes) points the other way.* A lang-pair split refutes an English-anchoring mechanism (the residual is larger on non-English pairs even for English-only models), pointing to shared surface structure (CJK numerals, templated frames). No construct-valid cross-lingual positive survives this battery.

6 Discussion and Limitations

What to change when matching features across models. Read a match count only against a null that re-runs the matching: random pairings sit below a permuted-pairing null that absorbs up to 94% of the unfiltered Lan et al. signal above random (Section M). Score frozen pairs on documents the rule never saw, prefer reciprocal matching (greedy-max excess saturates even for GPT-2↔Mistral-7B, Finding 2), and read each control as the retention ratio (4) against a same-sized reference arm.

In our v0 setup (`circuit-tracer` v0.4.1, experimental `nnsight` backend, float16; Table 2), attribution graphs did not converge on four Qwen3 sizes (0.6B–8B) though the models answer correctly; we did not test the library's default backend, so this is a failure of our pipeline, not evidence about the method (Section R).

Scope of the SAE-matching result. Nothing is claimed beyond the tested scope: eight models across six families, one released SAE per model (plus a 65k-width check for Gemma-2-2B), mid-layer bands, en/zh/ja/ko prompts, and mean-pooled features, which avoid token alignment but lose word order. The token-aligned follow-up (Section 5.1) removes the order-insensitivity limit at word granularity, but same-text conditioning cannot separate word-in-context sharing from syntax. The families were trained separately, but most of their pretraining corpora are undisclosed and may overlap on web text, possibly including C4, so the result does not separate shared training data from shared inductive bias. The frozen-pair lexical arm selects and scores on the same 20,000 documents. A split-half check finds 1–3% selection optimism in pooled excess on three cells, but does not bound selection bias in R_{lex}; the disjoint-half Procrustes arm also survives (Table 22, Section Q). Further caveats (GPT-2, z comparability, ablation probes, the Gemma read point, BOS in the pooled mean, no random-weight model): Section Q.

7 Conclusion

Gemma-2-9B's and Mistral-7B's matched features respond to the same documents, and at our read point that coactivation survives chance, a disjoint evaluation half, and the topic, lexical and token-level controls run on it, though the lexical one narrows rather than closes the shared-vocabulary alternative. The bounds are the tested scope: nine disjoint-half-confirmed cells, from a non-pre-registered follow-up; three residualization/token-aligned cells; and the released-SAE triples of Table 8, under the frozen matching rule and evaluation distribution. "Sharing" here is distribution-conditioned coactivation, not feature identity, not compositional or syntactic sharing, and not evidence of shared computation (both cross-model ablation probes were inconclusive). A cross-model match should come with a null that re-runs the pairing rule, frozen pairs scored on unseen documents, and one control per named nuisance, each shown able to remove what it names (a planted signal or a non-inert gate). The same discipline rejected three secondary signals, all reported (one tracked network depth, one rested on one English-only baseline family, one held only on templated prompts); that it can reject gives the surviving coactivation its weight.

REFERENCES

Ayushi Agarwal. Cross-architecture steering transfer in language models: A systematic empirical study, 2026. arXiv:2608.05164.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.04836.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, et al. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.07682.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, pp. 2397–2430, 2023. arXiv:2304.01373.

Joseph Bloom. Open source sparse autoencoders for all residual stream layers of GPT2-small. LessWrong, 2024. URL https://www.lesswrong.com/posts/f9EgfLSurAiqRJySD/open-source-sparse-autoencoders-for-all-residual-stream.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. SAELens. https://github.com/jbloomAus/SAELens, 2024.

BluelightAI. Qwen-3 cross-layer transcoders (base 0.6B/1.7B), 2025. URL https://huggingface.co/bluelightai.

Igor Bogdanov and Changcheng Huang. Discovering cross-language reasoning invariance in LLMs with geometry-invariant sparse autoencoders, 2026. arXiv:2608.23809.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

David Chanin. sae-llama-3.2-1b-res: Residual-stream sparse autoencoders for Llama-3.2-1B. HuggingFace model repository, chanind/sae-llama-3.2-1b-res, 2024. URL https://huggingface.co/chanind/sae-llama-3.2-1b-res.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring linear features across language models with model stitching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.06609.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2302.03025.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024. arXiv:2309.08600.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.16156.

Boyi Deng, Yu Wan, Yidan Zhang, Baosong Yang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of ACL*, 2025. arXiv:2505.05111.

Boyi Deng et al. Qwen-Scope: Turning sparse features into development tools for large language models, 2026. arXiv:2605.11887.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2108.01661.

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2306.09346.

Amil Dravid, Yasaman Bahri, Alexei A. Efros, and Yossi Gandelsman. Neuron populations exhibit divergent selectivity with scale, 2026. arXiv:2606.03990.

Joshua Engels. Mistral-7B residual stream SAEs. HuggingFace model repository, `JoshEngels/Mistral-7B-Residual-Stream-SAEs`; registered in `sae-lens` as `mistral-7b-res-wg`, 2024. URL `https://huggingface.co/JoshEngels/Mistral-7B-Residual-Stream-SAEs`.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, et al. Natural language autoencoders produce unsupervised explanations of LLM activations. Transformer Circuits Thread, 2026. URL `https://transformer-circuits.pub/2026/nla/index.html`.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, et al. Gemma 2: Improving open language models at a practical size, 2024. arXiv:2408.00118.

Gemma Team, Aishwarya Kamath, Johan Ferret, et al. Gemma 3 technical report, 2025. arXiv:2503.19786.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders, 2026. arXiv:2606.12138.

Matt Gorbett and Suman Jana. Ventriloquist LLMs: Linear alignment of late-stage representations, 2026. arXiv:2603.18908.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The Llama 3 herd of models, 2024. arXiv:2407.21783.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.14486.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in GPT2 language models. *Transactions on Machine Learning Research*, 2024. arXiv:2401.12181.

Michael Hanna. Gemma-Scope-2 transcoders (Circuit-Tracer), 2025a. URL `https://huggingface.co/collections/mwhanna/gemma-scope-2-transcoders-circuit-tracer`. HuggingFace collection of circuit-tracer configurations for the Gemma Scope 2 transcoders.

Michael Hanna. Transcoders trained on Qwen-3 models from 0.6B to 14B, 2025b. URL `https://huggingface.co/mwhanna`. HuggingFace model series.

Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. circuit-tracer, 2025. URL `https://github.com/decoderesearch/circuit-tracer`. Software library.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers, 2025. arXiv:2501.17727.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

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 ACL*, 2025. arXiv:2501.02009.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.07987.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in LLMs. In *Findings of NAACL*, 2025. arXiv:2408.10646.

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

Thomas Jiralerspong and Trenton Bricken. A "diff" tool for AI models. Anthropic Research, 2026. URL https://www.anthropic.com/research/diff-tool.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):242:1–242:52, 2025. doi: 10.1145/3728458.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines?, 2026. arXiv:2602.14111.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Nikolaus Kriegeskorte, W. Kyle Simmons, Patrick S. F. Bellgowan, and Chris I. Baker. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12(5): 535–540, 2009. doi: 10.1038/nn.2303.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders, 2024. arXiv:2410.06981.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *International Conference on Learning Representations (ICLR)*, 2016. arXiv:1511.07543.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.19. arXiv:2408.05147.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, et al. On the biology of a large language model. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.19647.

Jordan F. McCann. Polymorphism is rotation: Operational mechanistic interpretability from a two-layer transformer to Pythia-70m. *arXiv preprint arXiv:2605.24577*, 2026.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma Scope 2 – technical paper. Google DeepMind technical paper, linked from the model card, 2025. URL https://huggingface.co/google/gemma-scope-2-4b-it.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.02922.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.15430.

Neel Nanda and Joseph Bloom. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3), 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, et al. In-context learning and induction heads. Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2501.16615.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010. doi: 10.2202/1544-6115.1585.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 37, pp. 775–818, 2024. doi: 10.52202/079017-0024. arXiv:2404.16014.

Anna Soligo, Vladimir Mikulik, and William Saunders. Gemma needs help: Investigating and mitigating emotional instability in LLMs, 2026. arXiv:2603.10011.

Daniel Son, Sanjana Rathore, Andrew Rufail, Adrian Simon, Daniel Zhang, Soham Dave, Cole Blondin, Kevin Zhu, and Sean O'Brien. Semantic convergence: Investigating shared representations across scaled LLMs, 2025. arXiv:2507.22918.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. arXiv:2310.10348.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2502.03714.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.06672.

Xu Wang, Yan Hu, Benyou Wang, and Difan Zou. Does higher interpretability imply better utility? A pairwise analysis on sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.03659.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? On the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15366–15394, 2024. doi: 10.18653/v1/2024.acl-long.820.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report, 2025. arXiv:2505.09388.

A Reproducibility Checklist

Models and data.

- Probe-attribution arm: all base models are publicly available on HuggingFace (Qwen3 0.6B–14B; Gemma3 1B/4B/12B/27B-it; the 17-model control set adds Pythia, GPT-2, GPT-Neo, OPT, BLOOM, Qwen2.5, Gemma-2). Pinned snapshot SHAs for every HuggingFace artifact (models, SAE suites, transcoders, and the C4 dataset) in Table 3.
- SAE-matching arm: all models (Gemma-2-2B/9B, Mistral-7B-v0.1, Qwen3-1.7B/8B-Base, Llama-3.2-1B, Pythia-70M, GPT-2) and all SAE suites are public: Gemma Scope (`gemma-scope-*-pt-res-canonical`), `chanind/sae-llama-3.2-1b-res`, `mistral-7b-res-wg`, `gpt2-small-res-jb`, `pythia-70m-deduped-res-sm` via `sae-lens`, and Qwen-Scope (`Qwen/SAE-Res-Qwen3-*`) via our manual adapter (`qwen_sae_adapter.py`, in the supplementary material).
- Prompts: the 206-pair multilingual set (`build_prompts_v3.py`, provenance in Section E), a 2,000-document C4 slice (`build_c4_slice.py`; its two 1,000-document halves carry the held-out test) and a 20,000-document C4 slice for the lexical and token-aligned arms; each model scores the first token of its own tokenization of the target string (no separately released tokenization audit).
- No private data is used.

Compute.

- Single NVIDIA H100-80GB / H200-141GB GPUs; the v1 probe-attribution sweep totals well under 10 GPU-hours (per-job allocations in Section F).
- Setup: the reported runs used PyTorch 2.6 (CUDA 12.4) and `circuit-tracer` v0.4.1 from `safety-research/circuit-tracer` at commit `4bb8c0e` (full SHA in Table 3). `setup.sh` pins PyTorch 2.6.0 (CUDA 12.4 wheels) and checks out `circuit-tracer` `4bb8c0e`. The v1 probe-attribution headline does not depend on the `circuit-tracer` library.

Code release. Code, receipts and registrations are in the anonymous supplementary material (`README.md` gives the offline re-computation commands) and will be released upon publication: (a) all `experiments/` scripts (including `probe_attribution_v2.py`); (b) the `utils.py` adapter layer over `circuit-tracer`; (c) all SLURM job scripts; (d) figure-generation scripts under `paper/figures_src/`; (e) the per-model v1 probe-attribution JSON summaries (`results/probe_attribution_v2/*.json`) with per-pair bootstrap CIs and permutation p-values; (f) the SAE-matching analysis code (`code/analysis/`), its receipts (under `results/cross_model_sae/`) and the registration files; and (g) the power-law and slope-bootstrap analysis scripts (`code/analysis/rosetta_sublinear_check.py`, `rosetta_slope_bootstrap.py`) with their output JSONs in `results/rosetta_check/`. The v0 attribution graphs did not converge, so no `.pt` graph or per-graph `metadata.json` artifacts are produced.

Random seeds. All stochastic steps use `numpy.random` seed 0 and `torch.manual_seed(0)`, except the registered seed-42 document split of the 20,000-document runs. Bootstrap CIs use 10,000 resamples with seed 1 (independent of run-time seeds).

Hyperparameters. Attribution-graph (v0) values match `circuit-tracer` defaults (Table 2); the SAE-matching arm uses $K{=}1{,}500$ throughout, and the lexical arm's k is chosen by the registered largest-adequacy-passing rule (Section L).

Statistical reporting.

- 95% CIs are bootstrap percentile intervals with 10,000 resamples.
- Significance comparisons are permutation tests (prompt-permutation nulls for SAE matching; exact model-as-unit and pair-level permutation for the residual gap); bootstrap intervals are reported as descriptive uncertainty, and where a bootstrap and a permutation test disagree we report both (Section 5.2).
- The nine held-out C4 tests are Bonferroni-corrected over the 17 C4 cells considered ($1/2001 < 0.05/17$); other verdicts rest on registered survive/fail rules, and all comparisons are reported.

Negative results / pre-registration. Except for the held-out frozen-pair follow-up (not pre-registered; Proposition 2), each arm's expected outcome was pre-stated before the deciding run, and negative outcomes are reported symmetrically: the v0 attribution-graph reproduction returned a non-convergence negative (Sections 6 and J); the cross-lingual arm required multilingual > English-only under a valid metric, which the raw and residual metrics failed (Section 5.2); the zero-shot probe-transfer go/no-go pre-registered transfer $\gg$ non-matched control on the surviving pairs and failure on the null pair (both held), and its independent-corpus hardening pre-registered the same pattern and *failed* on the surviving pairs — reported as the in-sample scoping of Finding 5 (Section Q). The shared-core control remains pre-registered and unrun (Section Q).

ICLR reproducibility statement. Models, datasets, compute, statistical methods, hyperparameters and seeds are reported in this appendix, and every model checkpoint used is public; code, receipts and registrations are in the anonymous supplementary material and will be released upon publication.

B Supporting examples and secondary-control detail

B.1 Matched-feature examples

What do the matched features respond to? To make the surviving coactivation tangible we inspect the top-20 mutual-NN matched pairs (by pooled correlation: best-case matches) for Gemma-2-9B↔Mistral-7B on C4, retrieving each side's five max-activating documents (per-pair receipts `interp_*.json`). All 20 pairs share at least one top document and $17/20$ share ≥ 3 of 5 (mean $3.35/5$); at our read point (Section P) the matched features are recognizably *topical* — e.g. cricket-match-report, Quranic-verse, and obituary ($5/5$ overlap) features — firing on the same documents

Table 2: Hyperparameters used for all attribution-graph generation and pruning. Values match `circuit-tracer` CLI defaults unless noted.

Parameter	Value	Source
Attribution (`attribute()`)		
`max_n_logits`	10	default
`desired_logit_prob`	0.95	default
`batch_size`	256	default
`max_feature_nodes`	7500	default
`offload`	`None`	disabled (v0 runs only; 8×A100)
`lazy_decoder`	`True`	memory-saving
Pruning (`prune_graph()`)		
`node_threshold` (τ_n)	0.8	default
`edge_threshold` (τ_e)	0.98	default
Models		
`dtype`	`float16`	memory-saving
backend (Qwen3)	`nnsight`	our default (run logs)
backend (Gemma3)	`nnsight`	required
Statistics		
bootstrap resamples	10,000	for 95% CIs
random seed (numpy / torch)	0	for any stochastic step
random seed (bootstrap)	1	independent stream

in independently-trained models from different families: feature-level coactivation one can audit by reading. We calibrate against a 2,000-derangement null (both document sets fixed; no feature keeps its partner; Table 13). The null is essentially empty: mean overlap 0.031 ± 0.040 of 5 (max over 2,000 derangements: 0.25) against the observed 3.35 — $z = 82.2$, $p = 1/2001$; 2.8% of deranged pairs share even one document versus 100% of matched pairs, and no document takes more than 2% of top-doc slots. It replicates on an independent cell (Qwen3-1.7B↔Gemma-2-2B: 3.60 vs. 0.041, $z = 81.6$) and at k=50 pairs (3.40 vs. 0.020, $z = 137.3$). A held-out variant removes the circularity of auditing the fitting corpus — fitting matches on one half, reading top documents only from the disjoint half — and still clears the same null ($18/20$ pairs share a document; mean 2.30 vs. deranged 0.035, $z = 51.1$; held-out activation correlation 0.69) — attenuated, but not an audit-circularity artifact.

B.2 Raw probe-attribution depth detail

The raw probe-attribution metric measures depth, not language. The raw metric paints a sharp, believable picture: architecture separates families. After our v0 attribution-graph pipeline failed to converge (Section 6), we ran a v1 logit-lens probe-attribution sweep on en/zh/ja/ko antonym prompts and correlated mid-layer attribution profiles pairwise (full tables: Section H). Gemma3's alignment is strong and scale-monotone, Qwen3's moderate at every size — Gemma3-1B exceeds every Qwen3 model ($+0.34$ vs. at most $+0.29$): at face value, an "architecture gates scale" law (Section I's power-law fit obliges). But one entry makes that reading untenable: *Pythia, a predominantly English base family, attains the highest "cross-lingual" scores in the five-model control battery* (Pythia-1.4B $r = +0.70$), outranked only by the 27B Gemma ($+0.813$). A metric that ranks an English-only model at the top of a putatively cross-lingual scale cannot be measuring cross-lingual semantics alone. We rebuilt the metric's generator from scratch (reproducing the original JSONs to $\max|\Delta r| = 0.0023$) and ran three null controls: **(a) layer-shuffle** — permute the per-layer ordering before correlating, destroying the depth profile while preserving every per-layer value; **(b) mismatched-pair** — derange the concept-to-concept alignment, destroying cross-lingual *semantics* while preserving the depth profile; **(c) monolingual vs. multilingual** — a valid metric must rank multilingual models above English-only ones.

Table 3: Pinned snapshot revisions for every HuggingFace artifact used in this work. Revisions marked ‡ are read directly from the run-time HF cache shared by all cluster jobs (`refs/main` of the cache directory every SLURM script exports). The remaining artifacts had been evicted from that cache before pinning; for these we record the `main` commit resolved via the HF Hub API on 2026-07-11, which is guaranteed to equal the revision used because each repository's last commit predates the project's first download (2026-04-30). The C4 slice is drawn deterministically by streaming the first documents of `en/train` at the pinned dataset revision (`build_c4_slice.py`).

HuggingFace ID	Pinned revision (snapshot SHA)
Base models — probe-attribution arm (14)	
`Qwen/Qwen3-0.6B`	`c1899de289a04d12100db370d81485cdf75e47ca`
`Qwen/Qwen3-1.7B`‡	`70d244cc86ccca08cf5af4e1e306ecf908b1ad5e`
`Qwen/Qwen3-4B`	`1cfa9a7208912126459214e8b04321603b3df60c`
`Qwen/Qwen3-8B`	`b968826d9c46dd6066d109eabc6255188de91218`
`Qwen/Qwen3-14B`	`40c069824f4251a91eefaf281ebe4c544efd3e18`
`google/gemma-3-1b-it`	`dcc83ea841ab6100d6b47a070329e1ba4cf78752`
`google/gemma-3-4b-it`	`093f9f388b31de276ce2de164bdc2081324b9767`
`google/gemma-3-12b-it`	`96b6f1eccf38110c56df3a15bffe176da04bfd80`
`google/gemma-3-27b-it`	`005ad3404e59d6023443cb575daa05336842228a`
`EleutherAI/pythia-70m-deduped`‡	`e93a9faa9c77e5d09219f6c868bfc7a1bd65593c`
`EleutherAI/pythia-160m-deduped`	`582159a2dfe3e712a8d47ae83dec95ae3bde8e7e`
`EleutherAI/pythia-410m-deduped`	`c4fc8d586d62df497f1f9b69d66d3ca419992d3e`
`EleutherAI/pythia-1b-deduped`	`7199d8fc61a6d565cd1f3c62bf11525b563e13b2`
`EleutherAI/pythia-1.4b-deduped`	`554d9c1bae3877f740aece41feb90f912cf9fedc`
Additional control-set models (completes the 17-model set)	
`bigscience/bloom-560m`	`ac2ae5fab2ce3f9f40dc79b5ca9f637430d24971`
`EleutherAI/gpt-neo-125m`	`21def0189f5705e2521767faed922f1f15e7d7db`
`facebook/opt-125m`	`27dcfa74d334bc871f3234de431e71c6eeba5dd6`
`gpt2`‡	`607a30d783dfa663caf39e06633721c8d4cfcd7e`
`gpt2-medium`	`6dcaa7a952f72f9298047fd5137cd6e4f05f41da`
`Qwen/Qwen2.5-0.5B`	`060db6499f32faf8b98477b0a26969ef7d8b9987`
`google/gemma-2-2b`‡	`c5ebcd40d208330abc697524c919956e692655cf`
SAE-matching arm — additional base models	
`google/gemma-2-9b`‡	`33c193028431c2fde6c6e51f29e6f17b60cbfac6`
`mistralai/Mistral-7B-v0.1`‡	`27d67f1b5f57dc0953326b2601d68371d40ea8da`
`Qwen/Qwen3-1.7B-Base`‡	`ea980cb0a6c2ae4b936e82123acc929f1cec04c1`
`Qwen/Qwen3-8B-Base`‡	`49e3418fbbbca6ecbdf9608b4d22e5a407081db4`
`meta-llama/Llama-3.2-1B`‡	`4e20de362430cd3b72f300e6b0f18e50e7166e08`
SAE suites (repositories backing the `sae-lens` releases and Qwen-Scope)	
`google/gemma-scope-2b-pt-res`‡	`fd571b47c1c64851e9b1989792367b9babb4af63`
`google/gemma-scope-9b-pt-res`‡	`f9b689815814972562d28082f9f7d65d7e01fdc8`
`JoshEngels/Mistral-7B-Residual-Stream-SAEs`‡	`d2d2a64135d2a2c8e7cdfe18c07297af251c744b`
`jbloom/GPT2-Small-SAEs-Reformatted`‡	`57d08a4fd333fbf18caf3fbea63ceeb88e2f50d9`
`ctigges/pythia-70m-deduped__res-sm_processed`‡	`4de021e0701171e7a0be1fb711d278ec5c0ebd4d`
`Qwen/SAE-Res-Qwen3-1.7B-Base-W32K-L0_50`‡	`ce1a79d9c5163932d65c417380e53230e1086370`
`Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_50`‡	`9b7d647127c303dd371e5dbd992236bd9799b171`
`chanind/sae-llama-3.2-1b-res`‡	`6395ab37285a5ef13b7bd16ca0ae75e5929f73b3`
PLT transcoders (v0 attribution-graph arm)	
`mwhanna/qwen3-0.6b-transcoders-low10`	`28aefe686c09e9bc1a862195b51e145f10868d87`
`mwhanna/qwen3-1.7b-transcoders-low10`	`9c1b17dfb156d82162ccd2cb7f047ac7f3d3585d`
`mwhanna/qwen3-4b-transcoders`	`94d176260ac39ce2f882b8b09aba8c118df29bb3`
`mwhanna/qwen3-8b-transcoders`	`dc677109cde096a85d03fff4f73a3ec88e7e2105`
`mwhanna/qwen3-14b-transcoders`	`e6595bb044503fd8344b9fef13c29e17a3eab1a8`
`google/gemma-scope-2-4b-it`	`3e94b68be95290aada5b7525cf431d3040f81bb1`
`google/gemma-scope-2-12b-it`	`4c419f1ba0be8b7754d4151d4f26c23b92a9029e`
`google/gemma-scope-2-27b-it`	`5c58dd4cddd52cef653059d85e12a86bf6222a28`
Dataset and library	
`allenai/c4` (dataset)	`1588ec454efa1a09f29cd18ddd04fe05fc8653a2`
`safety-research/circuit-tracer` (git, v0.4.1)‡	`4bb8c0ea10bde09727e14565ec8469656880da53`

B.3 Residual scale-up detail

Scaling the controls: the residual is not cross-lingually specific. A residual that has survived one control is not yet a result: the n=20 residual rested on one concept family and a single English-only baseline (Pythia). We therefore expand to 206 aligned concept pairs across nine families in en/zh/ja/ko and to 17 models, adding four English-only models from three further families (GPT-2/GPT-2-medium, GPT-Neo, OPT), and attach per-model bootstrap 95% CIs (10,000 resamples)

plus model-as-unit and pair-level permutation tests of the multilingual-vs-English-only gap. Because the global residual is contaminated by a family-block effect, we report the conservative *within-family* depth-corrected residual. The runner is byte-identical to Section 5.2's; of 19 models, two failed for disclosed technical reasons (OPT-350M, Pythia-2.8B), reported as failures, not silently dropped. The pair-level permutation test ($p = 0.499$) is computed on an observed pair-level gap of $+0.0001$.

C Proofs for Section 3

Notation follows Section 3. For a permutation π of $\{1, \dots, N\}$ write $(z^B \circ \pi)_j(d_n) = z^B_j(d_{\pi(n)})$.

Proof of Proposition 1(a). Fix (i, j). Because entries are almost surely distinct, $a(i)$ and $b(j)$ are unique, and $(i, j) \in \mathcal{M}(\rho)$ if and only if ρ_{ij} exceeds every other entry of row i and every other entry of column j, i.e. if and only if ρ_{ij} is the largest element of the set $S_{ij} = \{\rho_{ik}\}_{k \le K_B} \cup \{\rho_{lj}\}_{l \le K_A}$, which has $K_A + K_B - 1$ elements. By exchangeability each element of S_{ij} is equally likely to be the largest, so $\Pr[(i, j) \in \mathcal{M}] = 1/(K_A + K_B - 1)$. Summing the indicators over the $K_A K_B$ pairs gives $\mathbb{E}[M] = K_A K_B/(K_A + K_B - 1)$; for $K_A = K_B = K$ this is $K^2/(2K - 1) = K/2 + K/(4K - 2)$, which is 750.25 at $K = 1{,}500$. *Scope.* Exchangeability within a row and column is an idealization. Features of one released SAE are correlated with each other, so entries of ρ sharing a row or a column are dependent, and the realized chance count departs from this value; in our receipts and simulations it departs only downward: the permutation nulls we measure range from 371.8 to 607.0 across the six (cell, dictionary) runs in Table 1, and simulated dictionaries with shared within-model factors, or i.i.d. non-Gaussian features at small N, fall below it (`code/analysis/verify_props_20260924/`). That spread, not the idealized value, is why each cell is read against its own null. □

Proof of Proposition 1(b). *Invariance.* Reindexing $m = \pi(n)$, $\frac{1}{N}\sum_n z^B_j(d_{\pi(n)}) z^B_k(d_{\pi(n)}) = \frac{1}{N}\sum_m z^B_j(d_m) z^B_k(d_m)$, so B's within-model correlation matrix is unchanged, and A's is untouched. Column means and variances are also permutation-invariant, so standardization and the top-K variance screen commute with π. *Validity.* Let T be any measurable statistic of (z^A, z^B), let $\pi_1, \dots, \pi_n$ be i.i.d. uniform permutations independent of the data, $\pi_0 = \mathrm{id}$, and $T_b = T(z^A, z^B \circ \pi_b)$; the reported value is $p = (1 + \#\{b \ge 1 : T_b \ge T_0\})/(n + 1)$. Let σ be a further uniform permutation independent of everything. Under H_0, $(z^A, z^B) \overset{d}{=} (z^A, z^B \circ \sigma)$, and because the π_b are independent of the data, $(T_0, \dots, T_n) \overset{d}{=} (T(z^A, z^B \circ \sigma \circ \pi_b))_{b=0}^n$. The permutations $\tau_b = \sigma \circ \pi_b$, $b = 0, \dots, n$, are i.i.d. uniform: $\tau_0 = \sigma$ is uniform, and given σ each $\sigma \circ \pi_b$ is uniform and independent of σ and of the others. Conditional on the data, the $n + 1$ values $T(z^A, z^B \circ \tau_b)$ are therefore i.i.d., hence exchangeable, so the rank of T_0 among them is uniform with ties counted against it, which gives $\Pr(p \le \alpha) \le \alpha$ (Phipson & Smyth, 2010). Nothing in the argument restricts T, so it may select its own pairs; what it requires is that the *same* function, selection included, is applied to every permuted copy. Selecting pairs once on the observed data and permuting only their correlations breaks this, which is the circular analysis of Kriegeskorte et al. (2009). The same argument with π_b drawn uniformly from the subgroup of permutations that preserve a stratification gives validity of the stratum-restricted (length/topic) nulls under the weaker hypothesis of within-stratum exchangeability. □

Proof of Proposition 2. Documents are i.i.d. and D_S, D_E index disjoint sets, so the evaluation data are independent of D_S and their conditional law given D_S is their unconditional law. Given D_S, $\widehat{\mathcal{P}}$ is fixed. (a) By linearity, $\mathbb{E}[T_E(\widehat{\mathcal{P}}) \mid D_S] = |\widehat{\mathcal{P}}|^{-1} \sum_{(i,j) \in \widehat{\mathcal{P}}} \mathbb{E}[\rho^E_{ij} \mid D_S] = |\widehat{\mathcal{P}}|^{-1} \sum_{(i,j) \in \widehat{\mathcal{P}}} \mathbb{E}[\rho^E_{ij}]$. (b) Given D_S, $T_E(\widehat{\mathcal{P}})$ is a fixed statistic of the evaluation data; apply Proposition 1(b) on D_E. For re-selected pairs, $\max_j \rho^E_{ij} \ge \rho^E_{ij^*}$ for every j^*; taking expectations with $j^* = \arg\max_j \mathbb{E}[\rho^E_{ij}]$ gives $\mathbb{E}[\max_j \rho^E_{ij}] \ge \max_j \mathbb{E}[\rho^E_{ij}]$, strictly unless $\arg\max_j \rho^E_{ij} = j^*$ almost surely (the gap is $\mathbb{E}[\max_j \rho^E_{ij} - \rho^E_{ij^*}]$, positive exactly when the argmax differs from j^* with positive probability), so a statistic scored on the data that chose its pairs is optimistic relative to any fixed pair. A null that permutes only the correlations of those pairs, without re-selecting, need not reproduce that optimism, so (b) is not guaranteed for it; this is the denominator of r (`in_sample_eval_own_pairs` in the receipts). A null that re-runs the selection in every draw

is valid by Proposition 1(b). *Implementation note.* Features that are constant on D_E have undefined ρ^E and are dropped before selection. That step reads only within-model marginals on D_E and is invariant under permuting B's evaluation rows, so (b) holds exactly; (a) holds conditional on the retained feature set R, with $\mathbb{E}[\rho_{ij}^E]$ replaced by $\mathbb{E}[\rho_{ij}^E \mid R]$; R involves no cross-model quantity. In the implementation the premise "$\widehat{\mathcal{P}}$ depends on D_S only" is thus relaxed to "on D_S and on within-model marginals of D_E". The proposition says nothing about how D_S and D_E were chosen: the C4 halves are fixed in advance (documents 1–1,000 and 1,001–2,000 of the slice; `c4_fit.jsonl`, `c4_heldout.jsonl`), and i.i.d. sampling of documents is an assumption about them. □

D Detailed held-out confirmation

Table 4: **C4 held-out confirmation of all nine cross-family cells.** Pairs selected on 1,000 documents are scored on a disjoint 1,000 against a 2,000× null. *held exc.*: observed minus null mean; *retained*: ratio to pairs re-selected on the evaluation half. Every $p = 1/2001 < 0.05/17$, the resolution of a 2,000-draw test rather than identical tail probabilities; held-out excesses are point estimates. The follow-up was not pre-registered. The cells were chosen on a 2,000-document screen that saw both halves; since each held-out p is valid given the selection half (Proposition 2), Bonferroni over all 17 cells still bounds the family-wise error. Receipts: `heldout_*.json`.

Cell	held exc.	retained	corrected verdict
Qwen3-1.7B ↔ Gemma2-2B	0.732	0.936	survives
Qwen3-8B ↔ Gemma2-2B	0.698	0.901	survives
Qwen3-1.7B ↔ Gemma2-9B	0.678	0.881	survives
Qwen3-8B ↔ Gemma2-9B	0.682	0.869	survives
Qwen3-8B ↔ Mistral-7B	0.599	0.792	survives
Gemma2-9B ↔ Mistral-7B	0.602	0.791	survives
Gemma2-2B ↔ Mistral-7B	0.567	0.818	survives
Qwen3-1.7B ↔ Mistral-7B	0.526	0.757	survives
Llama-3.2-1B ↔ Mistral-7B	0.519	0.739	survives

E Prompt Sets

Provenance notes. Grammatical imperfection in the zh/ja/ko prompts is treated symmetrically across models, so it can only shrink the measured cross-lingual residual, not create it. Each model scores the first token of its own tokenization of the target string; no separate tokenization audit is released.

Table 5: Models used in the v1 probe-attribution sweep, with the transcoders and `circuit-tracer` backend configured for the v0 attribution-graph runs (the v1 sweep itself reads hidden states through `transformers`). PLT = per-layer transcoder; L = layers. The SAE-matching arm (Section 5.1) uses a separate model set — Gemma-2-2B/9B, Mistral-7B, Qwen3-1.7B/8B-Base, Llama-3.2-1B, Pythia-70M, and GPT-2 — via `sae-lens` and a Qwen-Scope adapter; the residual scale-up (Section 5.2) adds GPT-2/GPT-Neo/OPT/BLOOM/Qwen2.5/Gemma2 for 17 models total. Gemma3 models: Gemma Team et al. (2025); Qwen3 transcoders: Hanna (2025b); Gemma3 transcoders: Gemma Scope 2 (McDougall et al., 2025).

Family	**Model**	**Params**	L	**Backend**	**Transcoder**
Qwen3	Qwen3-0.6B	0.6 B	28	nnsight	`mwhanna/qwen3-0.6b-transcoders-lowl0`
	Qwen3-1.7B	1.7 B	28	nnsight	`mwhanna/qwen3-1.7b-transcoders-lowl0`
	Qwen3-4B	4 B	36	nnsight	`mwhanna/qwen3-4b-transcoders`
	Qwen3-8B	8 B	36	nnsight	`mwhanna/qwen3-8b-transcoders`
	Qwen3-14B	14 B	40	nnsight	`mwhanna/qwen3-14b-transcoders`
Gemma3	Gemma3-4B	4 B	34	nnsight	`google/gemma-scope-2-4b-it`
	Gemma3-12B	12 B	48	nnsight	`google/gemma-scope-2-12b-it`
	Gemma3-27B	27 B	62	nnsight	`google/gemma-scope-2-27b-it`

Full prompt lists with target tokens and translation provenance are in `data/prompts_expanded_v2.jsonl` and `experiments/build_expanded_prompts.py`; the 206-pair scale-up set spans nine

concept families (antonym, capital, addition, subtraction, successor, predecessor, doubling, weekday, month) and is generated by `code/analysis/build_prompts_v3.py`: the 20 antonym pairs are hand-authored and reused verbatim, the other 186 pairs are *program-generated* from curated four-language word lists inserted into fixed templated frames, restricted to concept families with unambiguous, auditable targets. The v1 probe-attribution runs on an aligned antonym set of 20 concept pairs per language in en/zh/ja/ko (e.g. *hot*, *big*, *fast*, *dark*, *high*, ...), totalling $20 \times 4 = 80$ prompts and giving 6 language pairs (*4 choose 2*) of 20 aligned prompt-pairs each. Per-pair sample sizes, bootstrap CIs, and permutation p-values are in `results/probe_attribution_v2/*.json`.

C4 slice. `code/analysis/build_c4_slice.py` streams `allenai/c4` (en, train) in order, collapses whitespace, keeps the first 300 characters of each document and skips any shorter than 50 characters. Every C4 "document" in this paper is therefore a prefix of 51–300 characters (mean 290; 1,791 of 2,000 at the cap), or 14–144 tokens, with means of 64–72 across the six tokenizers probed. The 2,000-document slice is the union of `c4_fit.jsonl` and `c4_heldout.jsonl` and equals the first 2,000 documents of the 20,000-document slice, which has the same length distribution (`bos_probe_ct.json`). Two consequences for reading the results: document length varies over a narrow range, so the length-matched null (Section L) tests length variation within these prefixes and says nothing about full-length documents; and where a tokenizer prepends BOS, that token carries weight $1/T_d$, about $1/64$ here, in the pooled mean, much more than it would on full documents (Section Q, item (v)).

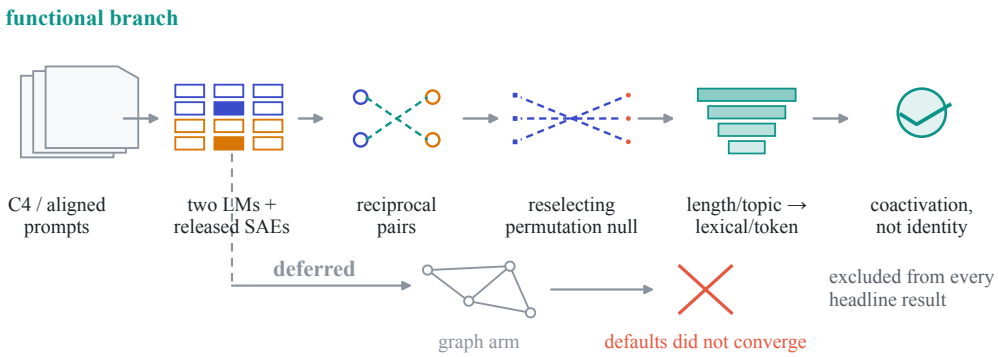

Figure 3: **Functional evidence pipeline and its scope.** Shared text enters independently trained models and released SAEs; reciprocal pairs are tested with reselecting permutation nulls and a ladder of controls. The result concerns coactivation rather than feature identity. A deferred attribution-graph branch did not converge at the defaults and contributes no headline result. Icons and permutation links are schematic.

F Compute and Scheduling

Every headline (v1) result in this paper runs as a single-GPU job on one H100-80GB or H200-141GB: the probe-attribution sweep is minutes per model (17 models, well under 10 GPU-hours in total), and each cross-model SAE-matching cell — the 2,000-document C4 slice through both models plus the 200-permutation null — is under 1.5 GPU-hours, with the `.sbatch` scripts (in the supplementary material) giving the exact allocation per cell. Two appendix analyses, the filtered audit of Lan et al. (2024) (Section M.1) and the Gemma read-point sensitivity (Section P), ran wholly or partly on Apple MPS with ungated Gemma-2 re-uploads, as stated there.

We do *not* report a job DAG for the v0 attribution-graph pipeline: its graph-building jobs never produced usable graphs (the influence propagation does not converge; Sections J and S), so their requested allocations describe a plan, not a reproducible cost behind any result here.

G ROBUSTNESS

Pruning threshold sweep. *Not applicable to v1.* The v1 probe-attribution scheme does not have a pruning threshold; the threshold sweep was a robustness check for the v0 attribution-graph backend, which is disclosed as non-convergent in Section 6 and not used for headline results.

Backend equivalence check. *Not applicable to v1.* The v1 probe-attribution is implemented natively in `transformers` + logit-lens hooks; no second backend is used. Robustness across alternative attribution implementations is left to follow-up work.

H RAW CROSS-LINGUAL PROBE-ATTRIBUTION SWEEP

Probe-attribution is a logit-lens-based, self-contained per-layer attribution scheme that does *not* require the `circuit_tracer` library or attribution-graph convergence. For each prompt and each layer ℓ we compute the logit-lens projection at position -1 and define the per-layer attribution a_ℓ as the target-token logit under this projection (the layer-ℓ residual passed through the model's final norm and unembedding, read at the target token); we then compute the Pearson correlation r between the per-layer attribution vectors for two languages L_1, L_2 on translated prompts, restricted to mid-layers $\ell \in [L/4, 3L/4)$. Table 6 reports the full per-language-pair sweep. This metric does not directly verify circuit-level feature sharing; Section 5.2 shows it is largely a depth-profile artifact, so it is reported here as the raw quantity that the construct-validity audit corrects, not as a headline result.

Table 6: **Cross-lingual mid-layer probe-attribution correlation, v2 ($n = 20$ prompt-pairs per language).** Pairwise Pearson r of per-layer logit-lens attribution between languages on aligned en/zh/ja/ko antonym prompts; mid-layer band $[L/4, 3L/4)$; bootstrap 95% CI and permutation-test p from 5,000 shuffles in the per-model JSONs. Each model evaluated on an aligned 20-pair multilingual probe set. Gemma3 shows substantially stronger and more scale-monotone alignment than Qwen3 on the identical prompt set — which Section 5.2 traces to the per-layer depth profile, not cross-lingual semantics.

Lang pair	Qwen3-0.6B	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	Qwen3-14B
en–zh	+0.238	+0.194	+0.055	+0.240	−0.129
en–ja	+0.122	+0.222	+0.057	+0.218	−0.191
en–ko	+0.186	+0.261	−0.057	+0.382	+0.133
zh–ja	+0.369	+0.178	+0.384	+0.519	+0.384
zh–ko	+0.062	+0.114	+0.076	+0.142	+0.074
ja–ko	+0.122	+0.194	+0.245	+0.254	−0.056
Mean	**+0.183**	**+0.194**	**+0.127**	**+0.293**	+0.036
N pairs $r > 0$	6/6	6/6	5/6	6/6	3/6
N pairs $p < 0.01$	3/6	5/6	2/6	6/6	4/6

Cross-family comparison (same $n = 20$ prompt set).

Lang pair	Gemma3-4B-it	Gemma3-12B-it	Gemma3-27B-it
en–zh	+0.459	+0.548	+0.767
en–ja	+0.601	+0.689	+0.781
en–ko	+0.689	+0.618	+0.846
zh–ja	+0.649	+0.736	+0.873
zh–ko	+0.409	+0.331	+0.767
ja–ko	+0.627	+0.515	+0.842
Mean	**+0.572**	**+0.573**	**+0.813**
N pairs $r > 0$	6/6	6/6	6/6
N pairs $p < 0.01$	6/6	6/6	6/6

Table 7: **Construct-validity controls on the raw probe-attribution metric** (mean over the six language pairs; verdict in Section 5.2). *real*: the headline metric, reproducing the original v2 JSONs. *residual* = real − mismatched. *survival* = mismatched/real. ML = multilingual. Layer-shuffle collapses every model to ≈ 0 while mismatched-pair survives 79–100%: the raw signal lives in the depth profile, not in cross-lingual semantics.

Model	ML	real r	layer-shuffle (a)	mismatched (b)	residual	survival
Pythia-1.4B	no	$+0.695$	$+0.002$	$+0.689$	$+0.006$	99%
Gemma3-4B	yes	$+0.572$	$+0.059$	$+0.493$	$+0.080$	86%
Pythia-410M	no	$+0.553$	-0.005	$+0.553$	$+0.000$	100%
Gemma3-1B	yes	$+0.336$	$+0.059$	$+0.279$	$+0.057$	83%
Qwen3-1.7B	yes	$+0.194$	$+0.040$	$+0.153$	$+0.041$	79%
Mean		$\mathbf{+0.470}$	$\mathbf{+0.031}$	$\mathbf{+0.434}$	$\mathbf{+0.037}$	**89%**

I Power-law scaling of the raw metric

Caveat: every number here is computed over the *raw* probe-attribution metric, which Section 5.2 shows is largely a per-layer depth-profile artifact. We retain the power-law analysis because it characterizes how that raw metric behaves with scale; the "architecture gates scale" reading is a statement about the raw, depth-confounded quantity, and the depth-corrected residual does *not* support a cross-lingual law (Section 5.2).

Fitting $r = c\,N^b$ (alignment r vs. parameter count N) within each family, with non-parametric bootstrap 95% CIs over the within-family (N, r) points (10,000 resamples; `code/analysis/rosetta_slope_bootstrap.py`), the within-family scale slope is itself architecture-dependent: Pythia shows a *real* positive slope ($b = +0.43$, CI $[+0.13, +0.80]$, $R^2 = 0.88$, $n = 4$), Gemma3 a positive lower-bound-borderline one ($b = +0.24$, CI $[+0.00, +0.43]$, $R^2 = 0.90$, $n = 4$), and Qwen3 *none* ($b = -0.32$, CI $[-1.25, +0.35]$). What is robust across all three families is that at matched parameter count the family gap dwarfs the within-family scale effect: Pythia-1.4B ($r = 0.70$) is 3.6× Qwen3-1.7B (0.19), the smallest Gemma3-1B (0.34) exceeds the best Qwen3 of any size (Qwen3-8B, 0.29), and Gemma3-4B (0.57) is 16× the Qwen3-14B alignment (0.04). This is the measured form of "architecture gates scale" over the raw metric — a 3.6–16× between-family gap dominating within-family slopes that range from real (Pythia) to null (Qwen3); Figure 4.

J Deferred attribution-graph analyses

The v0 case-study reproduction targeted the qualitative case studies of Lindsey et al. (2025) (antonym, two-hop factual recall, and entity-recognition hallucination prompts, reported on Claude 3.5 Haiku with the method of Ameisen et al. (2025)); our v0 `circuit-tracer` setup (Section R) failed to converge on all four Qwen3 sizes (0.6B, 1.7B, 4B, 8B), producing zero valid `structure_features` at default thresholds. Cross-lingual node-set Jaccard, thinking-mode dissection (Qwen3's `enable_thinking` toggle), and the attribution-graph $|V| \propto N^\beta$ power law all depend on the same backend and are deferred for the same reason. The probe-attribution sweep (Section H) was run in its place; its raw metric tracks network depth (Section 5.2), so it answers no cross-family scaling question.

K Ablation-based causal probes: full detail

Finding 3 (causal probe): a single-feature ablation test is underpowered and does not establish shared computation — and its own null highlights the same discipline. To probe whether the reciprocal sharing reflects shared *computation*, we ablate each matched feature in each model (a forward hook subtracting its decoder contribution at the SAE's layer; cf. the feature-level ablations of Marks et al., 2025) and correlate the two per-prompt Δloss vectors over 150 prompts (30 matched pairs), against a derangement null, for four pairs (`results/cross_model_causal/`, Table 18). The probe proves *underpowered*: individual SAE-feature ablation effects are small and

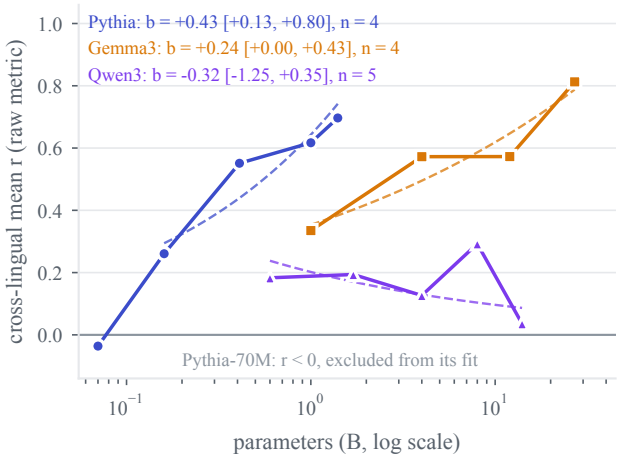

Figure 4: **Raw probe-attribution alignment vs. scale (depth-confounded; not a cross-lingual result, Section 5.2).** Mean pairwise correlation r vs. parameter count (log scale) for Pythia (0.07–1.4B), Qwen3 (0.6–14B), Gemma3 (1–27B), with power-law fits $r = cN^b$ (dashed); the key gives each family's exponent with its bootstrap 95% CI (`results/rosetta_check/rosetta_sublinear_v0.json`, `rosetta_slope_bootstrap.json`). Pythia $b = +0.43$ (real; fit on the four positive-r checkpoints, as the 70M point has $r<0$ and is excluded from the log-log fit), Gemma3 $b = +0.24$, Qwen3 $b = -0.32$ (null). Computed over the *raw*, depth-confounded metric.

sparse (stored signed mean Δloss $\approx 5 \times 10^{-4}$–6×10^{-2} nats in magnitude), so per-prompt Δloss vectors are near-degenerate and per-pair correlations pile up at exactly ± 1 — $3/30$ pairs for the within-family Gemma pair but 8–$23/30$ for the cross-family pairs. The permutation-on-the-mean statistic is therefore unreliable, and applying our own construct-validity reflex catches it: Pythia-70M↔Mistral-7B shows a large nominal $z = 28.3$ ($p = 0.001$) that is *entirely* a degeneration artifact — $23/30$ pairs are degenerate, and dropping them turns the signal *negative* (non-degenerate mean -0.12, 14% positive). After the degeneration filter, only the within-family Gemma pair retains a weak positive (non-degenerate mean $+0.13$, 63% of pairs positive, permutation $p = 0.038$), and even it does not survive a conservative sign test ($p = 0.36$); no cross-family pair shows a clean positive.

We then *upgraded the probe's power* as the degeneration analysis prescribes: ablate the top-30 matched features *jointly* in each model (a dense, large intervention — Δloss up to $+2$ nats, no ± 1 degeneration by construction) and test the correlation of the two per-prompt Δloss vectors against a null of 200 random same-size feature groups from the same top-variance pool in model B (`results/cross_model_causal_group/`). The matched-group correlations *look* substantial (within-family Gemma $+0.48$; Gemma-9B↔Mistral $+0.59$), but the random-group null reveals they are within chance variation: the null distribution is extremely broad (SD 0.39–0.47 for the two Gemma pairs, 95% CI up to $+0.78$), because per-prompt group-ablation Δloss is dominated by a *generic prompt-fragility axis* — almost any strong ablation disrupts the same fragile prompts. No pair reaches significance (within-family $p = 0.25$; Gemma-9B↔Mistral $p = 0.13$; Pythia↔Mistral $p = 0.30$), and critically the *within-family positive control fails*: a probe that cannot detect causal alignment where representational matching is unambiguous ($z = 17$ reciprocal sharing) cannot certify a cross-family null as a genuine dissociation either. We therefore report cross-model coactivation at the *representational* level (Findings 1–2), not shared computation, and report that *both* ablation-based causal probes — single-feature (degeneration) and multi-feature group (fragility-axis confound) — fail to resolve shared computation in either direction; within-model activation-level steering transfer is reported below. The pattern is this paper's thesis twice over: a nominal $z = 28.3$ dissolves under a degeneration filter, and a matched-group correlation of $+0.59$ dissolves under the right null — the statistic, not the phenomenon, can manufacture the effect.

The activation-level test we flagged as the key next step now exists: a *steering-transfer* probe that escapes both confounds above. For each of the 20 interpretable matched pairs we steer each fea-

ture *in its own model* ($h \mathrel{+}= \beta\, s_f W_{\text{dec}}[f]$ at the SAE read point, s_f the feature's natural activation scale) and compare the *identity* of the top-30 promoted tokens (decoded strings, so tokenizer-robust; steering is a dense large intervention, so no degeneration; token identity carries no prompt-fragility axis). Matched pairs share promoted-token strings far above a 2000-derangement null — at 20 pairs 0.70 vs. 0.01 expected ($z \approx 41$–44, $p = 1/2001$, invariant across $\beta \in \{2, 8\}$), and scaling to the top-50 mutual-NN pairs at β=2 *strengthens* it (1.26 vs. 0.01, $z = 94.2$, 56% of pairs with $\geq$1 exact-string overlap) — and it *replicates on a second, independent pair*: Qwen3-1.7B↔Gemma-2-2B gives 3.50 vs. 0.05 ($z = 43.6$, 65% of pairs overlapping; the higher raw overlap is consistent with the larger Qwen/Gemma shared string inventory). The token dumps are topically coherent — the cricket-matched pair promotes *batsman/batting/cricket* in Gemma and *batt/caught/declared/conceded* in Mistral; the gambling pair *betting/blackjack/gambler* vs. *ante/antes/dealt/ace* (`results/cross_model_sae/steer*.json`). The raw number (0.70 of 30) understates the effect because 30 is the wrong denominator: two empirical ceilings bound the statistic. Within-model self-agreement — the *same* feature in the *same* model at β=2 vs. β=8 — is only 6.8–9.3/30, and the binding constraint is the *shared promoted-string inventory*: across all 20 pairs only 16 of Gemma's 453 distinct promoted strings appear anywhere in Mistral's promoted universe at all. Of those 16 available strings, the *matched* feature captures 14 (88%), versus a 0.7% capture rate under the derangement null — a $\sim$125$\times$ enrichment at this shared-inventory ceiling. We therefore read the steering-transfer result as activation-level evidence consistent with matched features playing similar functional roles (steering is within-model, not a cross-model intervention), small in raw exact-string units for tokenizer reasons, near-ceiling once the inventory bound is accounted for. On the Gemma side the injection site is our read point, one block before the dictionary's training hook (Section P), so what is measured there is the effect of adding the released decoder direction at that point.

L FULL RESULT TABLES

Full tables, each summarized in the main text: the scaled residual controls (Table 9), cross-model SAE feature matching on the templated corpus (Table 10), the random-dictionary control (Table 11), the C4 cells as raw counts (Table 12), the derangement null for the top-document audit (Table 13), the single-feature causal probe (Table 18), and the cross-lingual cells (Table 19 and Figure 6). The C4 held-out confirmation is plotted in the main text (Figure 2), with exact values in Table 4; same-corpus discovery counts are in Table 12.

Table 8: **SAE-matching roster and exact observational scope.** "Read point" is where we apply the encoder; for every suite except Gemma Scope it is the dictionary's training hook. Depth is hook layer over transformer blocks. L_0 is the release target or the mean nonzeros/token (BOS included) measured at our read point by the tracked token-aligned extraction; "n/r" means neither was encoded in the release ID nor retained in a tracked receipt, and is not inferred. ‡Gemma Scope dictionaries were trained on `resid_post` of block N (`hidden_states[`N+1`]`) and we apply them to `hidden_states[`N`]`, one block early (Section P); without BOS their L_0 is $\sim$75 (2B) and $\sim$48 (9B) at our read point and $\sim$81 and $\sim$65 at the training hook. Repository snapshot SHAs are in Table 3. Every claim is conditional on these (model, released SAE, read point, corpus, pooling) choices.

Model	release / repository	read point	depth	width	L_0	recipe / provenance
Gemma-2-2B	`gemma-scope-2b-pt-res-canonical`	`resid_pre` block 12‡	12/26	16k (65k check)	$\sim$154	JumpReLU, Google Gemma Scope
Gemma-2-9B	`gemma-scope-9b-pt-res-canonical`	`resid_pre` block 20‡	20/42	16k	$\sim$117	JumpReLU, Google Gemma Scope
Mistral-7B-v0.1	`mistral-7b-res-wg`	`resid_pre` block 16	16/32	65,536	$\sim$88	SAE Lens; JoshEngels community release
Qwen3-1.7B-Base	`Qwen/SAE-Res-Qwen3-1.7B-Base-W32K-L0_50`	`resid_post` layer 14	14/28	32,768	50	TopK-ReLU, Qwen-Scope; manual audited loader
Qwen3-8B-Base	`Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_50`	`resid_post` layer 18	18/36	65,536	50	TopK-ReLU, Qwen-Scope; manual audited loader
Llama-3.2-1B	`chanind/sae-llama-3.2-1b-res`	`resid_post` block 8	8/16	16,384	$\sim$117	SAE Lens; chanind community release
Pythia-70M (negative)	`pythia-70m-deduped-res-sm`	`resid_post` block 3	3/6	32,768	n/r	SAE Lens release; registry lists no publication
GPT-2 (underpowered)	`gpt2-small-res-jb`	`resid_pre` block 6	6/12	24,576	n/r	reformatted SAE Lens release (Bloom, 2024)

In one cell, the overlap is not a document-length or coarse-topic covariate. One worry is that the "shared" features are an artifact of the probe documents' surface covariates rather than shared representation. We re-run one cross-family cell (Gemma2-2B↔Mistral-7B, +14.6% excess — seventh of the nine cross-family cells by excess, not the strongest, which is Qwen3-1.7B↔Gemma2-2B at +29.3%) under two covariate-*matched* permutation nulls that reproduce a covariate inside the null itself: permuting the second model's documents only within document-length deciles, and only within

Table 9: **Depth-corrected within-family residual at scale** ($n = 206$ concept pairs, 9 families, en/zh/ja/ko; 17 models). *residual* = real − within-family mismatched, the depth- and family-corrected cross-lingual component; 95% CI from 10,000 bootstrap resamples; $\star$ = CI excludes 0. EN = English-only, ML = multilingual. The residual is a *real* per-model quantity (10/17 CIs exclude 0) but is *not* cross-lingually specific: an English-only GPT-2 has the largest residual of all 17 models, the English-only mean exceeds the multilingual mean, and the multilingual-vs-English-only gap is not significant by the model-level exact permutation test ($p = 0.79$).

Model	type	residual	sig	Model	type	residual	sig
GPT-2	EN	+0.204	*	Qwen3-8B	ML	+0.063	*
Qwen3-0.6B	ML	+0.166	*	BLOOM-560M	ML	+0.050	*
Qwen3-4B	ML	+0.149	*	Qwen3-1.7B	ML	+0.023	*
GPT-2-medium	EN	+0.139	*	Gemma3-4B	ML	+0.022	*
GPT-Neo-125M	EN	+0.109	*	Gemma3-1B	ML	+0.010	
OPT-125M	EN	+0.104	*	Gemma2-2B	ML	+0.010	
				Qwen2.5-0.5B	ML	+0.010	
				Pythia-1.4B	EN	−0.001	
				Pythia-1B	EN	−0.006	
				Pythia-410M	EN	−0.006	
				Gemma3-12B	ML	−0.006	

English-only mean (7 models)		+0.078
Multilingual mean (10 models)		+0.050
Gap (multilingual − English-only)	−0.028	CI [−0.052, −0.004]
Model-level exact permutation p (gap ≥ obs; 19,448 = *17 choose 7*)		0.79
Model-level exact permutation p, two-sided		0.42
Pair-level permutation p, one-sided (206×6 concept pairs per model)		0.499
Pair-level permutation p, two-sided		0.98
Significant-positive split		4/7 EN, 6/10 ML

Table 10: **Cross-model SAE feature matching** (206 English prompts, top-1,500 features, 200-permutation null). *greedy exc.*: greedy-max mean correlation minus its null. *Hung.*: optimal one-to-one matched correlation. *mutual-NN z*: count of reciprocal best-buddy feature pairs as a z-score over the permutation null, (count − null mean)/null SD. Pairs measured at multiple depth bands show the range. *within* = same family (Gemma-2, three depth bands); *cross-modern* = different families (Gemma-2 / Mistral-7B / Pythia, 70M–9B); *cross-GPT2* = a pair including the older GPT-2. The strict mutual-NN z is the construct-valid measure; greedy and Hungarian are shown to be confounded.

Model pair (depth bands)	class	greedy exc.	Hung.	mutual-NN z
Gemma2-2B ↔ Gemma2-9B (early/mid/late)	within	+0.77–0.79	0.75–0.89	15.4–21.9
Pythia-70M ↔ Mistral-7B	cross-modern	+0.36	0.70	21.9
Gemma2-9B ↔ Mistral-7B (early/mid/late)	cross-modern	+0.67–0.72	0.32–0.46	13.9–16.7
Gemma2-2B ↔ Mistral-7B (early/mid/late)	cross-modern	+0.74–0.75	0.18–0.26	6.9–10.8
Pythia-70M ↔ Gemma2-9B	cross-modern	+0.36	0.32	4.6
Gemma2-2B ↔ GPT-2	cross-GPT2	+0.81	0.82	4.2
Gemma2-9B ↔ GPT-2	cross-GPT2	+0.69	0.69	0.5
GPT-2 ↔ Mistral-7B	cross-GPT2	+0.79	0.00	−1.3
Pythia-70M ↔ GPT-2	cross-GPT2	+0.03	0.02	−1.2

Every cross-modern z has $p = 1/201$ at *every* depth band; all cross-GPT2 pairs except Gemma2-2B↔GPT-2 are n.s. ($p \geq 0.37$). Observed mutual-NN counts are 1.2–2.4× their nulls. Full per-pair counts/nulls/p in `results/cross_model_sae/SUMMARY.json`.

TF-IDF topic clusters (k=10). The excess is essentially unmoved — +14.6% (random) → +14.3% (length-matched, z=19.3) → +14.6% (topic-matched, z=20.5), all p=1/201 — so document length explains only ∼2.5% of the excess and coarse topic overlap ∼0%: in this cell the overlap is not a length- or coarse-topic-covariate artifact (baseline reproduces the Table 12 cell to ±1 pair). One

Table 11: **Random-dictionary control: both the observation and its null move** (three representative C4-robust cells; same corpus, pooling, and permutation null as Table 4, with each side's trained SAE dictionary replaced by random unit directions of the same d_{sae}, top-k sparsified with k calibrated to that SAE's mean L_0, independent seeds per side). Reciprocal counts collapse while the permutation null *rises*, so the negative mutual-NN z is not an effect size; the greedy statistic moves the other way, rising under random dictionaries — an independent confirmation of Finding 2. Discussion: Section 5.1. Receipts: `rand_c4_*.json`.

	trained SAE			random dictionary		
Cell	obs	null μ	mutual z / greedy z	obs	null μ	mutual z / greedy z
Qwen3-1.7B ↔ Gemma2-2B	811	371.8	+42.7 / +180.7	152	604.7	−29.0 / +370.2
Gemma2-9B ↔ Mistral-7B	819	553.4	+24.0 / +84.6	212	607.0	−26.8 / +241.9
Llama-3.2-1B ↔ Mistral-7B	551	434.9	+12.3 / +19.1	223	584.5	−24.1 / +123.6

Table 12: **C4 same-corpus discovery cells as effect sizes, not only** z (the screen preceding Table 4: 2,000 C4 documents, top-1,500 mutual-NN features, 200-permutation null). *obs*: reciprocal best-buddy count; *null* $\mu\pm\sigma$: permutation mean and SD; *excess*: (obs − μ) as a fraction of the 1,500-feature pool — the practical effect size the z standardizes. The excess ranges from ∼30% of the pool (within-family) to 8–10% (the weakest survivors, Qwen3-1.7B and Llama-3.2-1B) and is *negative* for the Pythia failure. The recipe-check row is the same model pair as the Gemma2-2B row with only the released Gemma dictionary width swapped 16k→65k; it moves the cell by −6.3 excess points and −8.0 in z with no change of model, which is why we do not read the ordering of z across cells as a capability trend (Section 5.1). Per-cell receipts: `results/cross_model_sae/c4_*.json`.

Cell	class	obs	null $\mu\pm\sigma$	excess	z
Gemma2-2B ↔ Gemma2-9B	within	882	430.8±10.1	+30.1%	44.7
Qwen3-8B ↔ Qwen3-1.7B	within	842	408.5±11.4	+28.9%	38.2
Qwen3-1.7B ↔ Gemma2-2B	cross-family	811	371.8±10.3	+29.3%	42.7
Qwen3-8B ↔ Gemma2-9B	cross-family	859	471.8±10.4	+25.8%	37.1
Qwen3-8B ↔ Gemma2-2B	cross-family	784	402.7±9.6	+25.4%	39.8
Qwen3-1.7B ↔ Gemma2-9B	cross-family	774	434.6±9.9	+22.6%	34.2
Gemma2-9B ↔ Mistral-7B	cross-family	819	553.4±11.1	+17.7%	24.0
Qwen3-8B ↔ Mistral-7B	cross-family	777	515.4±10.0	+17.4%	26.2
Gemma2-2B ↔ Mistral-7B	cross-family	688	469.7±9.9	+14.6%	22.1
Qwen3-1.7B ↔ Mistral-7B	cross-family	620	475.0±11.6	+9.7%	12.5
Llama-3.2-1B ↔ Mistral-7B	cross-family	551	434.9±9.4	+7.7%	12.3
Gemma2-2B(w65k) ↔ Mistral-7B	recipe check	473	348.1±8.9	+8.3%	14.1
Pythia-70M ↔ Mistral-7B	cross-scale	390	411.1±10.9	−1.4%	−1.9

honest qualification: word-*order* scrambling of the documents does *not* shrink the excess (+16.1%), which is expected rather than reassuring — the C4 statistic mean-pools SAE activations over token positions, so it largely discards word order and a token scramble is a weak destructive probe for *this* metric (weak, not null — the pooled per-token activations remain context-dependent, so the excess shifts slightly under scrambling rather than staying fixed). At this stage the surviving overlap could still be lexical/token-level shared structure (it is robust to document length and coarse topic, and not attenuated by word-order scrambling); the lexical residualization and token-aligned tests below address that alternative. It is not evidence of shared compositional or syntactic features. Lexical composition is *itself* a shared-document covariate, and the two matched nulls above do not remove it. Receipt: `results/cross_model_sae/covctl_c4_gemma2-2b__mistral7b.json`.

Length and coarse-topic nulls and random splits on more cells. Registered before any output (`notes/PREREG_ninecell_ladder_20260926.md`; decision rule: a cell survives a control if $p \leq 0.05/18$ and $R \geq 0.5$). The cluster was unreachable when these ran, so a registered deviation, also committed before any output, moved them to a local machine and to the pooled features already extracted for Section P at the published read points (Apple MPS, ungated Gemma-2 re-uploads), not the features behind the main tables. Such features exist for five of the nine cells;

Table 13: **Derangement null for the top-document audit.** The "auditable by reading" statistic is the mean size of the intersection of the two sides' top-5 max-activating C4 documents, over the k mutual-NN matched pairs. Null: hold both sides' document sets fixed and re-pair them under 2,000 random derangements (no feature keeps its partner) — the question "would *any* pairing of these topical features share documents?". *max doc share*: largest fraction of all top-doc slots taken by a single document (a document-popularity artifact would drive this up). Every cell floors the empirical p at $1/(2000+1)$. Slots whose activation is exactly 0 are excluded: SAE activations are sparse and `argsort` breaks ties by index, so two models can select the same arbitrary zero-activation documents and manufacture overlap. This is real but tiny here (8 zero slots in 1,100, all in the held-out arm, contributing one overlap count; the held-out mean moves $2.35 \to 2.30$). The top-5 endpoint is a convention and is not load-bearing: recomputing at top-3 leaves every cell at the p floor ($z = 43$–136). Receipt: `results/cross_model_sae/INTERP_OVERLAP_NULL.json` (`code/analysis/interp_overlap_null.py`).

Cell	corpus	k	obs	null $\mu\pm\sigma$	null max	z	max doc share
Gemma2-9B ↔ Mistral-7B	C4 in-sample	20	3.35	0.031±0.040	0.25	82.2	2%
Gemma2-9B ↔ Mistral-7B	C4 in-sample	50	3.40	0.020±0.025	0.16	137.3	1%
Gemma2-9B ↔ Mistral-7B	held-out half	20	2.30	0.035±0.044	0.30	51.1	2%
Qwen3-1.7B ↔ Gemma2-2B	C4 in-sample	20	3.60	0.041±0.044	0.25	81.6	3%

the three Qwen3-8B cells and Llama-3.2-1B↔Mistral-7B have none. In every cell they reproduce the published screen count to within two pairs and the consecutive-split held-out excess to within 0.002 (within the registered reproduction gate of ±3 pairs and ±0.005). Each null is calibrated, as in (4), against 20 partitions with the same block sizes and random membership. Length deciles leave $R_{\text{len}} = 0.972$–1.003 and coarse topic (k=10) leaves $R_{\text{top}} = 0.979$–0.994, every arm at $p = 1/2001$ (Table 14). Ten seeded random 50/50 splits of the 2,000 documents replace the consecutive halves: all 50 held-out tests give $p = 1/2001$, and the median held-out excess is 0.97–1.01 times that of the consecutive split, so on these cells the consecutive split overstates the held-out excess by at most 3%. A bootstrap over evaluation documents (1,000 resamples, pairs frozen) gives 95% percentile intervals for Δ_E of width 0.014–0.022; the bootstrap distribution sits 0.007–0.011 below the point estimate, so these intervals lean low (for Gemma-2-9B↔Mistral-7B the upper bound falls just below the point estimate). The registered outcomes on these five cells are the support branches (A1, B1 and, for Δ_E, C1). Separately, the fine-topic receipts of the three ladder cells (Table 15), computed on the main-table features, carry an uncalibrated length-decile arm, not reported before, that leaves 1.002, 0.993 and 0.964 of the excess (Qwen3-1.7B↔Gemma-2-2B, Gemma-2-9B↔Mistral-7B, Llama-3.2-1B↔Mistral-7B; each $p = 1/2001$, n_{perm}=2,000); on the two cells both feature sets cover, the raw ratios agree to 0.001. With the Gemma-2-2B↔Mistral-7B null above, six of the nine cells have a length-decile null and five a coarse-topic null; the three Qwen3-8B cells have neither. Intervals for R were not computed, and the lexical and token rungs remain point estimates. Receipts: `results/ninecell_20260926/local_mps_cache/`.

Table 14: **Length and coarse-topic nulls, random splits and Δ_E intervals on five cells, from separately extracted features** (local Apple MPS extraction, ungated Gemma-2 weights; Section L). R: retention against 20 size-matched random partitions, n_{perm}=2,000, every $p = 1/2001$. Δ_E: consecutive-split held-out excess with a 95% percentile bootstrap over evaluation documents (1,000 resamples; the bootstrap distribution sits below the point estimate). Random split: range of held-out Δ_E over ten seeded 50/50 splits (all $p = 1/2001$); median r over the same splits. Q/G/M = Qwen3/Gemma-2/Mistral, number = size (B). Generated by `code/analysis/ninecell_strengthen_20260926.py` from `results/ninecell_20260926/local_mps_cache/`.

Cell	M	R_{len}	R_{top} (k=10)	Δ_E [95%]	random split Δ_E	median r
Q1.7–G2	812	1.003	0.994	0.731 [0.717, 0.731]	0.715–0.741	0.933
Q1.7–G9	777	0.988	0.994	0.676 [0.660, 0.676]	0.662–0.686	0.884
Q1.7–M	620	0.972	0.979	0.524 [0.504, 0.526]	0.496–0.536	0.734
G2–M	688	0.974	0.994	0.566 [0.548, 0.568]	0.529–0.565	0.792
G9–M	820	0.994	0.994	0.601 [0.579, 0.600]	0.599–0.619	0.792

Table 15: **Fine-grained lexical/topic-matched null.** Mutual-NN excess (observed − null mean) on three representative C4-robust cells under a permutation restricted to k-way TF-IDF/KMeans topic strata, n_{perm}=2,000. *base*: unrestricted null. *bow*: topic-matched. *size-rnd*: strata with the *same block sizes* but random membership, so they carry no lexical information — this calibrates the mechanical shrinkage that any fine partition induces (permuting inside small blocks approaches the identity; at block size 1 the excess is 0 by construction). R = bow/size-rnd is therefore the *lexical-specific* retention; the raw bow/base ratio is shown for completeness but is confounded by that shrinkage. Receipts: `covctl_k*_c4_*.json`.

Cell	band	k	blk	base	bow	size-rnd	R
Qwen3-1.7B ↔ Gemma2-2B	strong	200	10	+438.8	+379.2	+407.6	0.930
Qwen3-1.7B ↔ Gemma2-2B	strong	500	4	+438.8	+243.8	+247.9	0.983
Gemma2-9B ↔ Mistral-7B	median	200	10	+267.5	+257.0	+260.7	0.986
Gemma2-9B ↔ Mistral-7B	median	500	4	+267.5	+217.9	+205.5	1.060
Llama-3.2-1B ↔ Mistral-7B	weak	200	10	+112.7	+108.8	+109.3	0.996
Llama-3.2-1B ↔ Mistral-7B	weak	500	4	+112.7	+81.0	+80.1	1.012

The topic-matched null at fine granularity, with its own confound removed. Ten clusters over 2,000 documents is coarse, and one may ask whether narrower topical structure would explain the sharing. We therefore re-ran the matched null at k=200 and k=500 TF-IDF/KMeans clusters, at n_{perm}=2,000, on three representative C4-robust cells — the strongest, a median, and the weakest surviving cross-family pair (Table 15).

Reading this requires a correction that is easy to miss, and that we got wrong in our own first design. Permuting *within* strata approaches the identity as the strata shrink: at k=200 the mean block is 10 documents and at k=500 it is 4, so the null mean rises toward the observed count and the excess decays for a reason having nothing to do with topic — at block size 1 the excess is exactly 0 by construction. The raw bow/base ratio duly falls, to 0.56 for the strongest cell at k=500, which read naively would look like substantial topic-driven sharing. It is not. Permuting within strata that have the *same block-size multiset* but random membership — carrying no lexical information whatever — removes essentially the same amount (+247.9 vs. +243.8 for that cell). The lexical-specific retention R (bow over size-matched-random) is 0.93–1.06 across all six cell×k results, with every topic-matched p at the $1/2001$ floor: *the topic partition leaves at least* 93% *of the excess that a size-matched random partition leaves*. The decision rule and its thresholds were registered in the repository before any of these receipts existed.

Three limits bound what this shows, and we state them because the control is easy to over-read. *(i)* The size-matched arm fixes the block-size *distribution* but not which documents land in large blocks — KMeans puts generic documents in big clusters while the random version scatters them — so the calibration is first-order, not exact. *(ii)* TF-IDF/KMeans is itself a bounded lexical model: surviving it excludes shared *topic* detectors at 200–500-way granularity, not lexical structure that clustering does not represent. *(iii)* The statistic is still mean-pooled and order-insensitive, so shared *token-level* detectors remain a live alternative (Section 6); this control does not touch that hypothesis. What it does remove is the specific and most-cited alternative — that the surviving cells are two models' topic detectors co-firing on the same documents. One incidental benefit: because these runs use n_{perm}=2,000, their plus-one floor of $1/2001$ clears Bonferroni over the full C4 roster ($0.05/13 = 0.0038$, and $0.05/17 = 0.0029$ counting the GPT-2 cells), which the primary n_{perm}=200 runs do not.

The lexical-residualization test, at a corpus size where it is interpretable. The natural control is to residualize each feature's per-document activation against a lexical representation before matching. Our first attempt (receipt `lexres_c4_gemma2-2b__mistral7b.json`: TF-IDF→SVD at $k \in \{64, 128, 256\}$, 2,000 documents, 5-fold cross-fitted OLS) returned an excess surviving at every k — and we did not report that as survival, because a *random* subspace of identical rank removed just as much (∼99% retained; the ∼5-pair gap sits inside the null SD of ≈11). Nothing had been removed: the first stage explained a median 3.7% and 1.3% of feature variance out-of-fold, the

SVD arms weaker still. An intervention that removes no more variance than a random subspace of equal rank cannot exclude anything.

That inadequacy was itself the finding, and it prescribed the fix: a pre-registered first-stage adequacy gate, and tracking whether the *original* pairs disappear rather than whether a similar total is recomputed from different pairs. Both are now implemented (`code/analysis/lexical_token_test.py`, design pre-registered before the code was written), at 20,000 documents rather than 2,000. At that size the first stage is no longer inert: it beats a spectrum-matched *shuffled* design at every k, with 99–100% of features improving and median ΔR^2 rising to $+0.31$ at k=4096 — the gate the earlier run would have failed. *What that ΔR^2 is, exactly.* The registered adequacy rule is $\texttt{frac}(R^2_{\text{real}} > R^2_{\text{shuf}}) \geq 0.75$ *and* $\text{median}(R^2_{\text{real}}) - \text{median}(R^2_{\text{shuf}}) \geq 0.02$: it requires the real design to beat the shuffled one, and never requires $R^2_{\text{real}} > 0$. Since the shuffled design predicts badly, and worse as k grows (median $R^2_{\text{shuf}} \approx -0.013$ at k=256, -0.048 at 1024, -0.172 at 4096), ΔR^2 overstates the fit by roughly that offset. The real design's own median out-of-fold R^2 at k=4096 is $0.129/0.140$, $0.101/0.036$ and $0.028/0.036$ for the strong, median and weak cells (Table 16) — and on the Mistral side it is *lower* at k=4096 (0.036) than at k=256 (0.042), so the entire rise in ΔR^2 with k on that side is the shuffled arm degrading, not the nuisance model getting stronger. The comparison against the earlier 2,000-document run (0.037/0.013, at $k \leq 256$) is therefore only meaningful at matched rank: at k=256 on 20,000 documents the per-side medians are 0.042–0.112. Note also that these are medians over features, so "R^2 of 0.14" bounds the typical feature, not every feature. The mutual-NN pairs are frozen on the raw features, and the test asks whether *those* pairs still co-activate. They do, in all three cells and across all three strength bands: $R_{\text{lex}} = 0.93$ (Qwen3-1.7B↔Gemma2-2B, strong, 822 pairs), 0.94 (Gemma2-9B↔Mistral-7B, median, 806) and 0.92 (Llama-3.2-1B↔Mistral-7B, weak, 520) at the governing k, p at the $1/2001$ floor (Table 16); retention against the shuffled design is 1.000 exactly, so the residualization is removing real lexical structure rather than capacity.

Three limits, stated rather than buried. The statistic remains order-insensitive, so this bounds shared *lexical* structure at the ranks tested rather than excluding token-level detectors outright — the shared-unigram hypothesis is narrowed, not closed. The nuisance design is itself restricted, and the scope of "removing the lexical signal" is exactly that design and no wider: stopword-stripped *unigram* TF-IDF over at most 50,000 terms (`TfidfVectorizer(max_features=50000, stop_words="english")`, default `ngram_range=(1,1)`), reduced by truncated SVD to $k \leq 4096$ components that retain 67.7% of even that restricted design's variance. Nothing here bounds bigram or higher-order lexical structure, positional or stopword-carried structure, or the 32.3% of the unigram design's own variance that the truncation discards. And the frozen pair counts differ sharply between cells (822/806/520), so the weak Llama–Mistral cell's retention is the noisiest measurement of the three; we flagged that before the numbers existed rather than after.

Table 16: **Does the sharing survive residualizing on a unigram lexical model?** 20,000 C4 documents, n_{perm}=2,000. The mutual-NN pair set is frozen on the *raw* features and the test asks whether *those* pairs still co-activate after cross-fitted residualization on a TF-IDF→SVD$_k$ design (stopword-stripped *unigram* TF-IDF over 50,000 terms, truncated to k SVD components; at k=4096 those components retain 67.7% of that design's own variance). *gate*: the first stage must beat a row-shuffled design of identical spectrum (frac. of features with $R^2_{\text{real}} > R^2_{\text{shuf}}$, and median ΔR^2); an arm failing it is uninterpretable, not reassuring. **The gate is a margin, not an absolute fit**: it requires $R^2_{\text{real}} > R^2_{\text{shuf}}$, *not* $R^2_{\text{real}} > 0$, and the shuffled design predicts badly (R^2_{shuf} median -0.17 at k=4096), so ΔR^2 is much larger than the variance the nuisance model actually captures — which is the separate R^2_{real} column. *ret.*: excess retained relative to raw. R_{lex} = ret. real/ret. shuffled is the lexical-specific retention, since projecting out any k-dim subspace costs something. Row shown is the largest adequacy-passing k. Receipts: `lextok20k_*.json`.

Cell	band	k	pairs	gate frac	gate ΔR^2	R^2_{real}	ret. real	ret. shuf	R_{lex}
Qwen3-1.7B ↔ Gemma2-2B	strong	4096	822	1.00/1.00	+0.301/+0.312	0.129/0.140	0.933	1.000	0.933
Gemma2-9B ↔ Mistral-7B	median	4096	806	1.00/1.00	+0.273/+0.209	0.101/0.036	0.941	1.000	0.942
Llama-3.2-1B ↔ Mistral-7B	weak	4096	520	1.00/1.00	+0.197/+0.209	0.028/0.036	0.917	1.000	0.918

The token-aligned, order-sensitive test: full detail. Design, thresholds and all three outcome branches were frozen in `PREREGISTRATION_token_aligned.md` (commit `df330c6`) be-

fore any per-token activation existed anywhere — every prior dump is post-pooling — and the verdict reader (`tokalign_summary.py`, whose self-test fires all twelve branch paths including both NOT-TESTABLE gates and the refusal path) was committed before any job was submitted. The analysis pipeline's planted self-test must behave before any real receipt is produced: a planted unigram-only signal collapses to R_{tok}=0.008, planted context survives at 1.009, and planted document-topic signal sits inside the N1 null band. *Alignment.* Both sides tokenize the identical 20,000-document slice with `return_offsets_mapping`; each token maps to the whitespace word its character span overlaps most; the first and last word of every document are excluded (BOS bleed; truncation); a word occurrence enters only with ≥ 1 token on *both* sides. This yields 937,971/937,678/937,290 aligned occurrences over ∼70,800 word types per cell. Word-level activation is the mean over the word's tokens; max-pooling is a sensitivity arm ($S_1 = 0.522/0.413/0.346$: pooling-robust). *Provenance gates, read before any statistic.* G0 re-derives each cached top-1,500 pooled column from the new full-dictionary per-token dump by correlation matching over the shared document axis: 1,500/1,500 columns match at $r \geq 0.999$ on *every* side of every cell (worst matched $r = 0.99990$; median margin to the runner-up 0.39–0.52), and 822/822, 806/806, 520/520 frozen pairs are recovered — the registered fallback (re-freezing pairs) was never touched. G1: word coverage 1.000 on all six sides, zero cross-word tokens on the registered 50-document sample, 0.03–0.12% of documents excluded by the span-QA rule (one registered plumbing amendment: byte-fallback BPE tokens of a single multi-byte character share that character's span and are benign; overlap runs of ≥ 2 characters remain invalid). G2: every frozen pair clears the ≥ 50 nonzero-support-occurrence bar on both rungs, so no cell approaches the < 100-pair NOT-TESTABLE outcome. *Rung 1.* Within-document centering removes each feature's per-document mean, so any surviving correlation is invisible to every document-pooled statistic in this paper by construction, and the order-shuffle null N1 (permute side B's word order within each document, n_{perm}=2,000) leaves document-level quantities exactly unchanged. Observed $S_1 = 0.524/0.408/0.342$ against a null of mean ≈0, sd $\leq 5.7 \times 10^{-4}$; per-pair interquartile ranges $[0.35, 0.68]/[0.29, 0.52]/[0.21, 0.46]$; p at the $1/2001$ floor. The position control N2 (permute across documents within relative-position deciles) removes essentially nothing (S_1 excess over N2 $= 0.520/0.406/0.340$): the alignment is content, not shared position profiles. *Rung 2, first design (reported; superseded by the leakage-clean rerun below).* The conditioning key is the lowercased, punctuation-stripped word type with ≥ 20 occurrences (4,820/4,817/4,816 types covering 772,851/772,594/772,349 occurrences); type fixed effects demean both sides within type, so a shared *unigram* detector on that key contributes exactly nothing, and the null N3 deranges occurrences within type. $S_2 = 0.546/0.527/0.438$, p at the floor; support-matched $R_{\text{tok}} = 1.06/1.30/1.31$ (naive ratio 1.05/1.30/1.29; $R_{\text{tok}} > 1$ means conditioning *sharpens* the per-pair correlation by removing diluting between-type variance, not that signal was added). An independent audit (amendment A3 to the registration) found this rung does not license a word-identity-survival reading, on two routes. First, selection leakage: the frozen pairs were *selected* by document-pooled correlation on these same 20,000 documents, and N3 moves occurrences *across* documents within type — destroying exactly the document alignment that selection optimized — while the type-centered observed statistic still contains document-level variation; S_2 beating N3 can therefore re-detect the published document-level result rather than demonstrate survival of word-identity conditioning. Second, the S_2^c arm (within-document centering *then* type demeaning, $+0.484/+0.403/+0.331$ at its $1/501$ floor) does not close that route: the document and type projections do not commute, so their composition is not a joint document+type fixed-effect residualization, and a pure unigram signal with unbalanced document composition retains nonzero, cross-side-identical residuals under it. A third, mechanical caveat: the statistics core drops zero-residual-variance pairs per transformation, so R_{tok}'s numerator and denominator are not asserted to score an identical pair set, and no common scored-pair mask is written to the receipts. Bigram conditioning (demeaning within (preceding-word-type, word-type) cells, ∼8,750 cells, ∼287,000 occurrences, $+0.525/+0.533/+0.439$) inherits the same selection caveat. The committed reader's verdict, quoted verbatim from its output: *"BRANCH: B1 TOKEN-ALIGNED + ORDER-SENSITIVE"*, exit code 0; per amendment A3 that branch's S2 rung was demoted to UNRESOLVED pending a selection-clean rerun (pair selection and evaluation on disjoint documents; genuine joint document+type fixed-effect residualization; identical residualization applied under every null draw; an asserted common scored-pair mask written to receipts). The demotion was non-blind (receipts had been seen) and disclosed. The counterexample above is why the *first* design could not license the claim; the rerun below is what does. *Rung 2, leakage-clean rerun (amendments A3-EXEC+A4; verdict-bearing).* The rerun's design was frozen pre-receipt in amendment A3-EXEC (commit `b0b07cd`), extended by the A4 battery (full-dictionary universe screened on the selection half; a token stream gated to split B

with a per-cell aliasing audit — $0/1/0$ affected occurrences, the one alias in Gemma-2-9B↔Mistral-7B), and closes both routes by construction, meeting the four A3 and three A4 requirement markers asserted in every receipt. (a) Disjoint halves: a seed-42 permutation splits the 20,000 documents into two 10,000-document halves; the candidate universe is re-screened on split-A pooled rows over the full dictionary (A4), then pairs are selected by the verbatim mutual-NN rule inside it ($814/820/522$ selected), and every statistic and every null draw uses split-B occurrences exclusively (each split's SHA-256 is written to every receipt). (b) Genuine joint document+type fixed effects: alternating projections in float64 to tolerance 10^{-8} (observed 14–17 sweeps); the counterexample above (docs [A,A,B]/[A,B,B], $f(A)$=1, $f(B)$=0) residualizes to $\leq 5.2 \times 10^{-9}$ in the committed selftest — the visible proof that the non-commuting S_2^c defect is closed. (c) The identical residualization is re-run under every one of the 2,000 N3 draws (`null_reresidualized` asserted in each receipt). (d) One common scored-pair mask (G2 support and residual sd $> 10^{-8}$ on both sides under both transforms), fixed before any draw; $R_{\text{tok}}^{\text{clean}}$'s numerator and denominator are computed on exactly that mask (n_{scored}=$813/819/522$; the mask's SHA-256 in each receipt). Under this design S1 passes on split B in all three cells (r=$0.5221/0.3979/0.3353$, every p at the $1/2001$ floor) and the joint-FE S2 survives: $R_{\text{tok}}^{\text{clean}}$=$0.925/0.986/0.977$ (naive ratio $0.901/0.982/0.960$; p_{N3} at the floor) — far above the registered 0.30 survival bar, nowhere near the 0.10 collapse bar. The committed clean reader's verdict, quoted verbatim: *"BRANCH: B1-clean"*, exit code 0; receipts `tokaclean20k_*.json`. This replaces the earlier unresolved status. *Bound.* What stands is Rung 1 plus the leakage-clean Rung 2: shared occurrence-level variation that survives joint document and normalized-current-word-type fixed effects under a design that cannot re-detect selection. The survival still sits *below* composition: no syntactic claim is made at any order (the bigram caveat). Had S_2 collapsed, the pre-registered consequence was the sharpest negative in the paper (surviving coactivation = shared token-identity detectors, retro-diagnosing R_{lex}); had S_1 failed, the registered irony was that pure token detectors *would* have been token-aligned. All three branches were first-class in both the original battery and the clean rerun; receipts for the branches that fired: `tokalign20k_*.json`, `tokaclean20k_*.json` (Table 17).

Table 17: **Token-aligned sharing (S_1) and current-word conditioning arms (S_2; the leakage-clean $R_{\text{tok}}^{\text{clean}}$ is the verdict-bearing arm)** (pre-registered; 20,000 C4 documents, ≈940,000 aligned word occurrences per cell, n_{perm}=2,000). *pairs*: frozen mutual-NN pairs recovered by the pooled-reconstruction gate G0 (column match $r \geq 0.999$). S_1: mean frozen-pair correlation of within-document-centered word-level activations (a quantity no document-pooled statistic can contain); its within-document order-shuffle null N1 has mean ≈0, sd $\leq 6 \times 10^{-4}$. S_2: the same after fixed effects over lowercased, punctuation-stripped word types with ≥ 20 occurrences, against a within-type derangement null N3 (first design: reported, not verdict-bearing — selection-leakage caveat, Section L). R_{tok}: support-matched ratio of the S_2 excess to the S_1 excess on the same retained occurrences (first design; scored-pair sets are per-transformation, no common mask asserted). $R_{\text{tok}}^{\text{clean}}$: the same ratio under the registered leakage-clean rerun (amendments A3-EXEC+A4: pair selection and evaluation on disjoint 10,000-document halves with the candidate universe re-screened on the selection half over the full dictionary; a gated token stream with a per-cell aliasing audit ($0/1/0$ affected occurrences); converged joint document+type fixed effects re-applied under every null draw; one asserted common scored-pair mask, n_{scored}=$813/819/522$). S_1^{clean}: the S1 excess under that same design (the verdict-bearing S1). Both p_{N3} at the floor; receipts `tokaclean20k_*.json`. S_2^c: within-document centering *then* type demeaning (excess; the projections do not commute). *bigram*: excess after demeaning within (preceding-word-type, word-type) cells. Every primary p at the $1/2001$ floor; sensitivity arms (S_2^c, bigram) at their $1/501$ floor.

Cell	band	pairs	S_1	S_1^{clean}	S_2	R_{tok}	$R_{\text{tok}}^{\text{clean}}$	S_2^c	bigram
Qwen3-1.7B↔Gemma-2-2B	strong	822/822	0.524	0.522	0.546	1.06	0.925	+0.484	+0.525
Gemma-2-9B↔Mistral-7B	median	806/806	0.408	0.398	0.527	1.30	0.986	+0.403	+0.533
Llama-3.2-1B↔Mistral-7B	weak	520/520	0.342	0.335	0.438	1.31	0.977	+0.331	+0.439

M AUDIT OF A PUBLISHED PROTOCOL (LAN ET AL.)

To show that the two ingredients our method adds — a permutation null strong enough to absorb the pairing step, and a natural-text corpus control — are not optional, we re-implement the Lan et al.

Table 18: **Causal cross-model feature-transfer probe** (150 prompts, 30 mutual-NN pairs, derangement null). *perm-p*: permutation test on the mean matched Δloss-correlation. *degen*: pairs with $|r| > 0.99$ (degeneracy from sparse single-feature effects). *non-degen mean* and *frac.* +: after dropping degenerate pairs. The probe is underpowered; the Pythia z is a degeneration artifact (non-degen mean < 0); only the within-family pair shows a weak positive.

Model pair	class	perm-p	degen	non-degen mean	frac. +
Gemma2-2B ↔ Gemma2-9B	within	0.038	3/30	+0.13	0.63
Pythia-70M ↔ Mistral-7B	cross	0.001†	23/30	−0.12	0.14
Gemma2-9B ↔ Mistral-7B	cross	0.35	16/30	+0.16	0.50
Gemma2-2B ↔ Mistral-7B	cross	0.71	8/30	+0.09	0.41

†degeneration artifact: 23/30 pairs pinned at $|r| = 1$; the non-degenerate signal is negative.

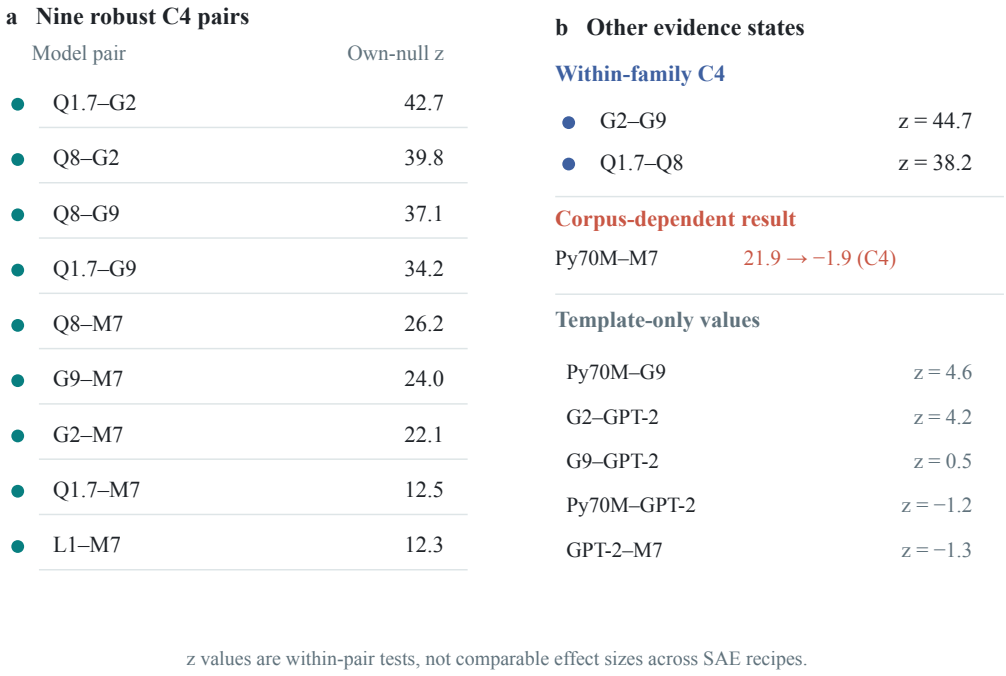

Figure 5: **Evidence categories across model pairs.** The nine cross-family C4 pairs exceed their own permutation nulls. Exact z values are listed without a common magnitude color scale, because SAE recipes differ. The right panel preserves within-family controls, the template-to-C4 failure of Pythia-70M/Mistral, and all five template-only pairs. Low-count natural-text GPT-2 evaluations remain without a verdict; template z values must not be read as a natural-text result. Each unordered pair appears once.

(2024) statistic and apply our control battery to *it*. Their protocol: greedily pair features across two models by max activation correlation, then measure SVCCA and RSA between the *paired decoder-weight rows*, significance judged against a *random-pairing* baseline. We reproduce the pairing and similarity statistic, and separately add their filters (Section M.1); Table 20 is the statistic without them. We run it on four of our pairs — a strong surviving pair (Gemma-2-9B↔Mistral-7B, mutual-NN C4 $z = 24.0$), a weak surviving pair (Qwen3-1.7B↔Mistral-7B, C4 $z = 12.5$), the corpus-artifact pair our C4 control rejects (Pythia-70M↔Mistral-7B, C4 $z = -1.9$), and a null pair (GPT-2↔Mistral-7B, $z = -1.3$) — on both the templated corpus and C4, adding one control of our own: a *permuted-pairing* null that permutes the prompt rows before the greedy pairing, destroying genuine cross-model correspondence while preserving the pairing step's opportunism (Table 20).

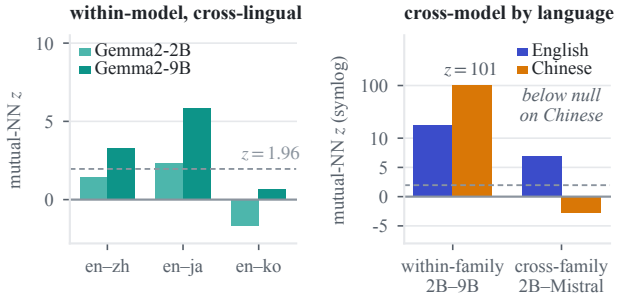

Figure 6: **Cross-lingual sharing is weak within models and language-conditional across them** (companion to Figure 5; numbers in Table 19). Left: within-model sharing clears the significance line for en–ja/en–zh in the larger Gemma and, marginally, for en–ja in the smaller (nominal $z = 2.3$, not multiplicity-robust); en–ko is null in both. Right (symlog y): on Chinese, the within-family pair shares strongly (shared tokenizer/script pipeline, flagged) while the cross-family pair collapses below null despite sharing on English. Plotted from the `xling_*.json` and `zh_gemma2-2b__*.json` receipts.

Table 19: **Cross-lingual and language-conditional cells with the validated tool** (mutual-NN z over the permutation null; $N = 206$ aligned concepts unless noted; * = permutation $p \leq 0.05$). Left: *within-model* en↔X feature sharing. Right: cross-model cells by prompt language, plus the C4 natural-text control ($N = 2000$) and the n_{perm}=2000 hardening.

Within-model	en–zh	en–ja	en–ko
Gemma2-2B	1.4	2.3*	−1.7
Gemma2-9B	3.3*	5.8*	0.7

Cross-model	en	zh
Gemma2-2B↔9B	17.4*	101.0*
Gemma2-2B↔Mistral	6.9*	−2.7
C4 control (N=2000)	G2↔G9 44.7*; G9↔M 24.0*; Py↔M −1.9 (fails)	
n_{perm}=2000 (templates, N=206)	both headline pairs $p = 1/2001$	

The audit is not a claim that Lan et al. (2024)'s within-scope finding is wrong — they compare same-tokenizer pairs, mostly within one family, and make a rotation-invariant *space*-level claim. It shows that their statistic, transplanted to the cross-family regime without their filters, needs the two controls that let our reciprocal-matching battery separate a real signal from a corpus artifact: a stricter-than-random null (their random baseline is far too weak), and a natural-text corpus (on templates the statistic cannot even separate the strongest genuine pair from the artifact). Their filters change the first two findings and leave the third, by a narrow margin (Section M.1). The implementations (`code/analysis/lan_protocol_audit.py` and `lan_filter_audit.py`, selftests included) are in the supplementary material.

The reverse transplant: their pairing rule inside our strongest control. Table 21 reports the complementary direction — their 1A-to-ManyB pairing rule substituted for our mutual-NN criterion inside the Table 16 lexical-residualization pipeline, everything else held fixed. Design, adequacy gate, tie rule (*largest* adequacy-passing k) and all five outcome branches were registered before any receipt existed; the committed reader refuses partial input and applies the reimplementation validity gate before reading any R_{lex}. Outcome: the registered survival branch, $R_{\text{lex}} = 0.900/0.905/0.910$ at governing k=4096, every p at the $1/2001$ floor. One k-dependence, disclosed rather than averaged away: the sign of the small gap between their rule and ours is not stable across k — at k=256 their argmax pairs retain marginally *more* than our mutual-NN pairs, while at the governing k=4096 they retain marginally less in all three cells. The registered readout is the largest adequacy-passing k (the strongest first stage), and we report that; the cross-k receipts are

Table 20: **The Lan et al. protocol under our controls, without their filters** (with them: Table 23). SVCCA and RSA of greedily paired decoder spaces, with the original *random*-pairing baseline and our stricter *permuted*-pairing null; "exc." = RSA − permuted-null RSA. Three findings, each a control our method supplies: (i) for the three non-null pairs the random-pairing baseline sits far below the permuted-pairing null on both statistics — RSA rand ≈ 0 vs. perm 0.041–0.150, and SVCCA rand 0.164–0.167 vs. perm 0.194–0.251 — so the greedy pairing step itself manufactures 15–94% of the RSA "signal above random" (least for the strongest pair on C4, most on the templated corpus), while the null pair shows no such gap (GPT-2 RSA perm ≈ 0); (ii) the statistic is *corpus-dependent* — on the templated corpus even the strongest surviving pair collapses to the artifact's level (exc. $+0.010$ both); (iii) on C4 it separates only the *strongest* pair ($+0.242$), while the weak surviving pair ($+0.079$) is indistinguishable from the corpus artifact ($+0.068$) — the artifact even scores higher on raw RSA (0.146 vs. 0.134). Only the null pair is correctly ≈ 0. Our reciprocal-matching battery with a natural-text control separates the surviving pairs from the corpus artifact and the null pair; it makes no strong-vs-weak ordering.

		RSA				SVCCA		
Pair (our verdict)	corpus	obs	rand	perm	exc.	obs	rand	perm
Gemma-2-9B↔Mistral (survives, z=24.0)	templates	0.161	−0.001	0.150	+0.010	0.258	0.167	0.209
	C4	0.284	−0.000	0.041	**+0.242**	0.457	0.167	0.194
Qwen3-1.7B↔Mistral (survives, z=12.5)	templates	0.099	0.002	0.063	+0.036	0.293	0.167	0.236
	C4	0.134	−0.001	0.054	+0.079	0.387	0.167	0.219
Pythia-70M↔Mistral (artifact, z=−1.9)	templates	0.086	0.003	0.076	+0.010	0.288	0.165	0.251
	C4	0.146	0.000	0.078	+0.068	0.350	0.164	0.236
GPT-2↔Mistral (null, z=−1.3)	templates	0.001	0.000	−0.001	+0.002	0.174	0.168	0.178
	C4	0.000	0.001	−0.002	+0.002	0.190	0.167	0.187

`lanpair_lextok20k_*_k{256,1024,4096}.json`. The comparison supports one claim only: survival under lexical residualization does not depend on which of the two pairing rules is used. It licenses no ranking of the rules.

Table 21: **The same control, applied to a published pairing rule.** Lan et al. (2024) pair cross-model SAE features by taking, for each feature in A, its highest-correlated feature in B (their 1A-to-ManyB rule; many-to-one by construction). We transplant *that rule* into the Table 16 pipeline and change nothing else — same $20{,}000$ C4 documents, same cached pooled features, same TF-IDF→SVD$_k$ cross-fitted design, same row-shuffled calibration, same within-fold null, n_{perm}=2,000. *reimpl.*: the validity gate read *before* any R_{lex} — the mapping must be many-to-one and its 1–1 fraction must fall inside the $[0.159, 0.632]$ interval implied by their Tables 5–6; it is necessary, not sufficient (an arbitrary argmax assignment gives $\approx 1/e$). *ours*: the mutual-NN R_{lex} from Table 16 on the same cell. Receipts: `lanpair_lextok20k_*.json`.

Cell	band	k	pairs	1–1 frac	ret. real	ret. shuf	R_{lex}	ours
Qwen3-1.7B ↔ Gemma2-2B	strong	4096	1500	0.413	0.899	1.000	0.900	0.933
Gemma2-9B ↔ Mistral-7B	median	4096	1500	0.424	0.904	0.999	0.905	0.942
Llama-3.2-1B ↔ Mistral-7B	weak	4096	1500	0.299	0.908	0.998	0.910	0.918

A third pairing rule: orthogonally aligned, fitted off the evaluation data. McCann (2026) report that decoder-cosine similarity matches across seeds at 98% while an SAE trained on one seed reconstructs another's activations at negative explained variance, and argue the standard universality metric cannot see that failure. The account predicts something testable here: if the cross-model correspondence is a rotation of a shared subspace, pairing in an orthogonally *aligned* coordinate system should recover it at least as well as raw correlation pairing. We registered that baseline before implementing it (branches P1–P4, rank 64, split seed 0, both fixed a priori) and ran it with a leakage control: the alignment is fitted over the shared document axis on a random half of the corpus — assuming no feature correspondence — and the frozen pairs are then scored, with the identical residualization, on the *held-out* half only. The committed reader checks that leakage constraint

(evaluation N must equal the recorded held-out size, rank and seed must match the registration) before reading any R_{lex}.

The registered branch P1 fired: $R_{\text{lex}} = 0.928/0.920/0.885$ at the governing k=4096 (Table 22), every cell above the 0.70 survival bar and three times the 0.30 collapse bar, on half the documents. Survival under lexical residualization is therefore a property of the shared structure under *three* pairing rules — our reciprocal mutual-NN criterion, Lan et al.'s published argmax rule, and an orthogonally aligned rule fitted off the evaluation data — rather than of any one matching criterion. Two bounds we state rather than bury: the aligned rule retains slightly *less* than mutual-NN in every cell (0.928 vs 0.933, 0.920 vs 0.942, 0.885 vs 0.918), so alignment does not improve on raw pairing here; and halving the corpus halves the first stage's power, which is why the registration fixed the split in advance and forbade enlarging the corpus if the adequacy gate had failed.

Table 22: **The same control, under a Procrustes-aligned pairing fitted on a disjoint document half.** A rotation account of cross-model correspondence (McCann, 2026) predicts that pairing in an orthogonally aligned coordinate system should recover the correspondence at least as well as raw correlation pairing, and that the standard universality metric is blind to the failure it documents. We test that here: the alignment is fitted over the *shared document axis* on a random half of the 20,000 C4 documents ($Q = UV^\top$ from the SVD of $U_B^\top U_A$, rank 64, split seed 0, both fixed a priori), so no feature correspondence is assumed by the fit; pairs are mutual nearest neighbors by cosine in that aligned space; and every quantity below — raw excess, residualization, null — is computed on the *held-out* half only. Everything else matches Table 16. *ours*: the mutual-NN R_{lex} on the same cell. Registered at `3be9b7d` with branches P1–P4 before the pairing existed; the selftest proves it collapses on purely lexical synthetic data and survives on a genuine shared latent. Receipts: `proclex_lextok20k_*.json`.

Cell	band	k	pairs	ret. real	ret. shuf	R_{lex}	ours
Qwen3-1.7B ↔ Gemma2-2B	strong	4096	609	0.928	0.999	0.928	0.933
Gemma2-9B ↔ Mistral-7B	median	4096	491	0.920	1.000	0.920	0.942
Llama-3.2-1B ↔ Mistral-7B	weak	4096	335	0.885	1.000	0.885	0.918

M.1 Audit with their filtering steps

Table 20 scores all 512 greedy pairs. Lan et al. (2024) filter the pairs first (their Sec. 3 and Apps. A and K; the order and details come from their released code): a pair is dropped if either feature's five highest-activating tokens include a non-concept token (punctuation, whitespace, or a BOS/EOS/padding token), and is kept only if the two top-five lists share a token string; a target that receives several source features keeps one pair; pairs with activation correlation ≤ 0.1 are dropped. We added these three steps to the audit under a pre-registration written before any activation was extracted, which fixed the decision rules and the sentence for every outcome (`notes/PREREG_lan_filters_20260924.md`). Two adaptations are forced by our setting. Token strings are compared after stripping whitespace, because our pairs mix SentencePiece and byte-level tokenizers; on raw strings no cell keeps more than 18 pairs. The pairing still correlates pooled documents, not tokens. The top-five tokens come from a fresh forward pass over the same inputs (an Apple-silicon run, with Gemma-2-9B from the ungated bf16 re-upload `unsloth/gemma-2-9b`, not checked bit-for-bit), which reproduces every unfiltered statistic of Table 20 to within 0.013. Our permuted-pairing null is pushed through all three filters, so it faces the same selection as the observed pairs.

Table 23 gives the result. The filters keep 17–32% of the pairs on templates and 4.5–7.6% on C4; they report 20–30% for pretrained SAEs. On C4, 260 of Mistral's 512 pool features have punctuation, whitespace or BOS among their top five tokens. The null pair keeps no pair on either corpus: every top activation of the GPT-2 features sits on a prompt's first token, and no GPT-2 pair passes the keyword step. Read against the three findings of Table 20: (i) With their filters, the pairing-step inflation remains on templates, much reduced, but on C4 the filters leave too few permuted pairs to measure it. On C4 the filters discard the permuted pairings: no permutation draw keeps the 21 pairs SVCCA needs (median 7–10 kept). On templates the permuted null still sits above their random baseline (RSA 0.026–0.040 vs. ≈0, in 126–197 of 200 draws), but that is now 7.5–14.1% of the signal above random instead of 63–94%. The inflation we report is therefore a

property of the unfiltered statistic; their full pipeline guards against it in part, and on C4 it does so by discarding most pairs. (ii) With their filters the templated corpus does separate the strongest pair from the artifact (exc. 0.370 vs. 0.158), so this finding does not hold for their full pipeline. (iii) On C4 the filtered statistic still does not separate the weak surviving pair from the artifact. The two pairs' obs − rand is 0.324 vs. 0.174, but with only 30 and 23 pairs left that gap (0.150) is just inside the pre-registered noise band (0.164), a marginal call. The strongest pair stays separated from both (0.574). Under their text's one-to-one reading (drop every many-to-one target instead of keeping one), the permuted null is not estimable in any cell. The filtered result is their filters applied to our transplant, not a replication of their numbers: their token-level instance unit, their RSA variant and their same-tokenizer model pairs are not reproduced.

Table 23: **The Lan et al. protocol with their filters.** Primary pre-registered variant (non-concept and shared-keyword filter on whitespace-stripped top-5 token strings, their code's one-to-one step, $r>0.1$). kept = pairs left of the 512 greedy pairs. RSA of the kept pairs' decoder rows (obs), their shuffle baseline within the kept pairs (rand, 200 draws), and our permuted-pairing null pushed through all three filters (perm, 200 draws; n.e. = fewer than half the draws keep the 21 pairs SVCCA needs, with the number that do). exc. = obs − perm; †obs − rand where perm is n.e. (the pre-registered fallback). share = (perm − rand)/(obs − rand), the quantity that is 63–94% on the templated corpus without the filters (Table 20). SVCCA is near its ceiling on so few rows, so its random baseline rises as kept shrinks; it is shown, not read. Receipts: `lanfilt{T,C}_*.json`; pre-registration: `notes/PREREG_lan_filters_20260924.md`.

			RSA					SVCCA	
Pair (our verdict)	corpus	kept	obs	rand	perm	exc.	share	obs	rand
Gemma-2-9B↔Mistral (survives)	templates	86	0.402	0.002	0.032	0.370	7.5%	0.559	0.420
	C4	39	0.576	0.002	n.e. (0/200)	0.574†	—	0.764	0.654
Qwen3-1.7B↔Mistral (survives)	templates	149	0.405	0.000	0.040	0.365	9.8%	0.505	0.314
	C4	30	0.322	−0.001	n.e. (0/200)	0.324†	—	0.800	0.768
Pythia-70M↔Mistral (artifact)	templates	164	0.184	0.000	0.026	0.158	14.1%	0.419	0.299
	C4	23	0.178	0.004	n.e. (0/200)	0.174†	—	0.941	0.922
GPT-2↔Mistral (null)	templates	0			no pair kept				
	C4	0			no pair kept				

N LOOSE-CRITERION CONFOUND SCATTER

Figure 7 plots the Finding-2 confound: the strict mutual-NN z separates the GPT-2 pairs from the rest (within-family and cross-modern pairs overlap), while the greedy-max excess saturates for most pairs (the low-sharing Pythia rows excepted) and carries no discriminating information. The same statistic still reports $z = 59$ on the Pythia-70M↔Mistral-7B C4 cell, whose reciprocal count is at null.

O ZERO-SHOT PROBE TRANSFER: FULL DETAIL

A shared interface would carry an artifact across models: we train a 9-way concept-family probe (closed-form ridge) on model A's matched-feature activations and apply it *unchanged* to model B's — no training on the target, the unsupervised matching the only shared coordinate system; families are linearly separable within each model (CV=1.0), so the test is transfer vs. a *non-matched* (permuted-matching) control. With the matching fit on the same templated prompts used for evaluation, on all three modern pairs the probe transfers zero-shot at 1.000 (163/94/267 matched features; controls 0.16–0.19, $p = 1/201$; majority-class rate 0.218); the null pair closes the loop (GPT-2↔Mistral: *one* mutual-NN feature; transfer = control, $p = 1.0$). Refit the matching on independent C4 text and transfer falls to chance (0.12–0.20, $p \geq 0.07$), so the in-sample transfer does not by itself establish a shared interface — Finding 1's corpus-conditionality at our own headline.

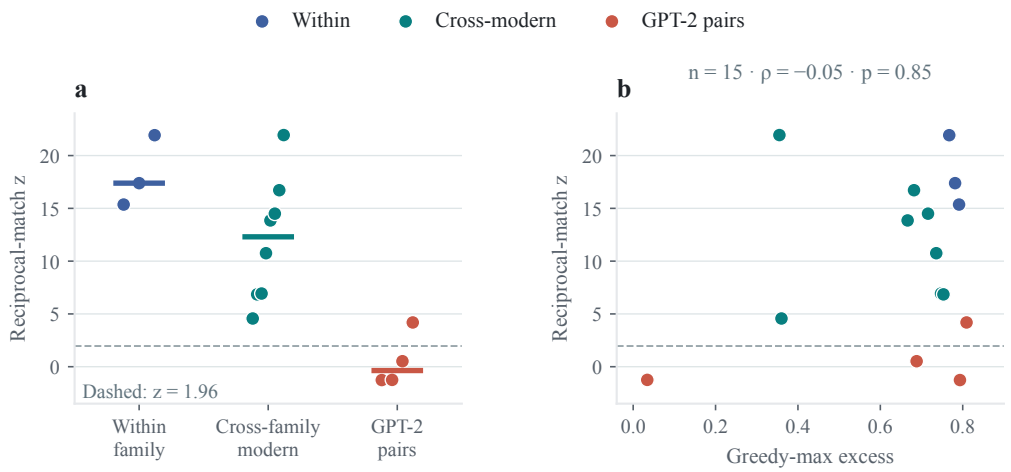

Figure 7: **Pair classes overlap under the reciprocal statistic.** (a) Each point is a model pair and the thick mark the category median. Cross-modern pairs overlap within-family pairs. (b) The same 15 points against greedy-max excess: Spearman $\rho = -0.054$, $p = 0.850$, so no monotonic association was detected in this sample. This is not proof of zero population association. Dashed lines denote nominal $z = 1.96$; source: `results/cross_model_sae/SUMMARY.json`.

P GEMMA SCOPE READ POINT: A PRE-REGISTERED SENSITIVITY ANALYSIS

What was read. In a Hugging Face forward pass `hidden_states[`i`]` is the input to block i. Gemma Scope's `layer_`N dictionaries were trained on `resid_post` of block N, i.e. `hidden_states[`N+1`]`; every Gemma cell in this paper applied them to `hidden_states[`N`]`, one block earlier ($N = 12$ for Gemma-2-2B, 20 for Gemma-2-9B). The cause is our hook-index rule, which added one only when "`resid_post`" appeared in the SAE's hook name or id; under `sae-lens` 6 the hook name sits in the config's metadata and the Gemma Scope id (`layer_`N`/width_16k/canonical`) does not contain it, so the shift never fired. The cluster logs and receipts record index N for every Gemma run. The other suites (Mistral, Qwen-Scope, Llama, Pythia, GPT-2) are read at their own training hooks. Steering and ablation write at the same index, so on the Gemma side they inject at our read point (Section K).

Our read point is not the training point. On the first 200 C4 documents (BOS excluded), the fraction of variance unexplained falls from our read point to the training hook for all three Gemma dictionaries (0.259 → 0.172 for 2B-16k, 0.248 → 0.146 for 2B-65k, 0.190 → 0.156 for 9B-16k) and L_0 per token moves toward the release's nominal value (74.8 → 80.7 vs. 82; 67.1 → 72.6 vs. 72; 47.9 → 64.8 vs. 68). The features we report are therefore the released encoder directions evaluated one block early, not the dictionaries on their training distribution, and the feature labels in Section B.1 are labels at our read point.

Registered sensitivity analysis. Before any run we registered a re-run of the Gemma cells of the C4 screen and the held-out confirmation in which the only change is reading `hidden_states[`N+1`]`; corpus, halves, top-1,500, permutations and seeds, and every non-Gemma side are fixed (`notes/PREREG_gemma_hook_sensitivity_20260924.md`). Main-text numbers stay at N by that registration: every rung of a cell's ladder was computed at the same read point, so each cell is internally consistent there. A cell is *retained* at N+1 if its held-out p is $1/2001$, its held-out excess is positive and its screen $p \leq 1/201$. Two gates come first: G0 (the re-run at N reproduces the published receipt) and H (the fraction of variance unexplained is lower and L_0 closer to nominal at N+1, the numbers above; H passes for all three dictionaries). The Gemma-2-9B↔Mistral-7B numbers at N+1 had been seen during the investigation that found the error, so that cell is reported but not counted. The registered environment was Apple MPS with the ungated `unsloth` re-uploads of Gemma-2 (Google's repositories are gated and were not available

on that machine); the weights were not compared bit for bit, which is what G0 checks. *Deviation from the registration:* Qwen3-8B-Base could not be downloaded to that machine, so the two Qwen3-8B cells were run on the cluster with Google's original weights, in the environment that produced the published receipts; that run also repeated two other Gemma-2-9B cells, which we report as a replication and do not use for any verdict.

Result (Table 24). Registered outcome A (all retained, all in band): all four blind cells that ran and passed the gates, Qwen3-1.7B↔Gemma-2-2B, Qwen3-8B↔Gemma-2-2B, Qwen3-8B↔Gemma-2-9B and Gemma-2-2B↔Mistral-7B, are retained at the training hook, with held-out excess moving by -0.001 to $+0.006$ and retained fraction by -0.003 to $+0.009$, inside the registered equivalence band (± 0.05). Qwen3-1.7B↔Gemma-2-9B missed G0 on the local run by one reciprocal pair ($\Delta M = 3$ against a tolerance of 2; its held-out excess reproduced within 0.002), so by the registration it is not interpreted; in the cluster replication it passes G0 ($\Delta M = 2$) and is retained ($0.678 \to 0.679$, retained fraction $0.881 \to 0.892$). The non-blind Gemma-2-9B↔Mistral-7B cell is retained and grows ($0.602 \to 0.617$, retained fraction $0.791 \to 0.816$; cluster replication 0.615 and 0.814). The screen z moves in both directions ($22.1 \to 19.0$ for Gemma-2-2B↔Mistral-7B, $42.7 \to 45.0$ for Qwen3-1.7B↔Gemma-2-2B), which is one more reason not to order cells by z (Section Q). The width check keeps its size: swapping the 16k for the 65k Gemma-2-2B dictionary moves the cell from $z = 19.0$ to 10.9 at the training hook, against 22.1 to 14.1 at our read point.

Exploratory, not registered. Re-running the top-document audit of Section B.1 for Gemma-2-9B↔Mistral-7B at the training hook leaves the mean top-5 overlap of the top-20 pairs at $3.35/5$, and the obituary pair is still a reciprocal match whose two features share all five top documents.

What this does not cover. Only the pooled screen and held-out statistics were re-run, and of the six cross-family Gemma cells only the four blind cells that passed G0 carry a registered verdict. Everything else on the Gemma side stands at our read point: the control ladder (topic, lexical and token-aligned arms, random dictionaries), the length- and coarse-topic-matched nulls, the templated-corpus cells, the Lan et al. audit, the ablation probes, steering, probe transfer and the cross-lingual Gemma cells. They are internally consistent there, and their values at the training hook are untested.

Table 24: **Gemma Scope cells read at the dictionaries' training hook** (pre-registered; Section P). Each entry is N / $N{+}1$: N is the published value (`hidden_states[N]`, the input to block N; Tables 4 and 12), $N{+}1$ reads `hidden_states[N+1]`, the `resid_post` hook each Gemma Scope SAE was trained on, all else fixed. z: C4-screen mutual-NN z (200 permutations); held-out excess and retained fraction as in Table 4 (2,000 permutations). G0: the re-run at N reproduces the published receipt (reciprocal count within 2, held-out excess within 0.005). Verdict at $N{+}1$ by the registered rule (held-out $p = 1/2001$, excess > 0, screen $p \leq 1/201$). Env: local = Apple MPS with ungated Gemma re-uploads; cluster = original weights. *Seen before registration; not counted. †Failed G0; not interpreted. Receipts: `results/cross_model_sae/gemma_hook_sensitivity/`.

Cell	env	G0	z	held-out excess	retained	verdict at $N{+}1$
Qwen3-1.7B ↔ Gemma2-2B	local	pass	42.7 / 45.0	0.732 / 0.731	0.936 / 0.932	retained
Qwen3-8B ↔ Gemma2-2B	cluster	pass	39.8 / 35.8	0.698 / 0.704	0.901 / 0.908	retained
Qwen3-1.7B ↔ Gemma2-9B	local	fail (ΔM=3)	34.2 / 32.0	0.678 / 0.678	0.881 / 0.891	not interpreted†
Qwen3-8B ↔ Gemma2-9B	cluster	pass	37.1 / 38.1	0.682 / 0.686	0.869 / 0.878	retained
Gemma2-9B ↔ Mistral-7B	local	pass	24.0 / 27.5	0.602 / 0.617	0.791 / 0.816	retained*
Gemma2-2B ↔ Mistral-7B	local	pass	22.1 / 19.0	0.567 / 0.572	0.818 / 0.819	retained
Gemma2-2B ↔ Gemma2-9B	local	pass	44.7 / 44.5	–	–	screen only
Gemma2-2B(w65k) ↔ Mistral-7B	local	pass	14.1 / 10.9	–	–	screen only

Q Extended threats to validity

Scope caveats. (i) GPT-2's four natural-text cells are underpowered (1 reciprocal match vs. a null of ≈2), so they support no verdict and GPT-2 is not a second below-threshold model — exactly one (Pythia-70M) is measured below its null with power — and its public SAE reads TransformerLens-processed activations (Nanda & Bloom, 2022); (ii) cell z's are not comparable across SAE recipes,

so we claim no capability trend and no convergence law; (iii) both ablation probes are inconclusive and no cross-model causal intervention succeeded — the functional evidence is within-model steering transfer and in-sample zero-shot probe transfer (Sections K and O); (iv) every Gemma Scope dictionary was applied one block before its training hook, so Gemma-side features, labels and steering are the released directions at our read point, and only the pooled screen and held-out statistics were re-run at the training hook (a pre-registered sensitivity analysis, Section P), for the six cross-family Gemma cells: four blind cells are retained, the non-blind Gemma-2-9B↔Mistral-7B cell is retained but not counted, and Qwen3-1.7B↔Gemma-2-9B missed its reproduction gate on the registered run and is not interpreted; nothing else on the Gemma side was re-run (the control ladder, the length- and coarse-topic-matched nulls, the templated-corpus cells, the Lan et al. audit, the ablation probes, steering, probe transfer and the cross-lingual Gemma cells; Section P); (v) the pooled feature averages over every token of a model's own tokenization, including a BOS token where the tokenizer adds one. The BOS activation is identical across documents, so it enters the pooled mean as c_f/T_d (c_f the feature's BOS activation, T_d the document's token count), a function of document length that two BOS-prepending models share. At the pinned revisions the Gemma-2, Mistral-7B and Llama-3.2 tokenizers prepend BOS and the Qwen3-Base tokenizer does not (tokenizers only; `code/analysis/bos_probe_ct.py`, `bos_probe_ct.json`), so in the six Qwen3 cells only one side carries this term; that does not make those cells length-free, since the one-sided term can still correlate with any length-sensitive feature of the other model. The shared route is open in the three Gemma-2/Llama-3.2↔Mistral-7B cells. Length-decile nulls bound it in all three: Gemma-2-2B↔Mistral-7B ($+14.6\% \rightarrow +14.3\%$ excess, Section L), and Gemma-2-9B↔Mistral-7B and Llama-3.2-1B↔Mistral-7B, whose uncalibrated length arms keep 0.993 and 0.964 of the excess (Section L); the token-aligned test excludes each document's first word and removes document means by construction; the other held-out cells were not re-run without BOS, so we cannot say how much of their pooled excess a length term carries. Recomputing the pooled features without position 0 is the direct check and is left to follow-up work; (vi) no randomly initialized model was run through the pipeline. SAEs on such models can score like trained ones on interpretability metrics (Heap et al., 2025), so the topical labels of Section B.1 are not by themselves evidence of learned features. The lexical and word-type controls remove what unigram and current-word detectors share, not other structure two random networks might share; a random-weight pair is the direct check and was not run.

External caveats on SAE matching. *(i) Seed instability.* SAEs trained on the same data can learn different features across seeds (Paulo & Belrose, 2026), even when the underlying subspaces are stable (Gerasimov et al., 2026). Our matches use single public releases per model, so cross-seed stability is untested. Three considerations bound the damage. First, the direction of the bias: seed noise *degrades* reciprocal matching, and spurious matches are absorbed by the permutation null. Instability can shrink the surviving-cell positive; the row-permutation null is conditional on the chosen SAE, so it bounds rather than excludes recipe-induced alignment. Second, our claim is existential, not one of per-feature identity. *Some* matched set coactivates on the matching distribution; a different seed finding a different set would not contradict that. Third, the positive already spans four independently released SAE suites (Gemma Scope (Lieberum et al., 2024), a community Mistral SAE (Engels, 2024), Qwen-Scope (Deng et al., 2026) via a manual loader, and a community Llama SAE (Chanin, 2024)), which makes a shared-recipe explanation of the cross-suite matches unlikely. A direct recipe check agrees on the *verdict*. Swapping Gemma-2-2B's SAE for the independently trained width-65k release leaves the C4 cell significant (mutual-NN 473 vs. permutation null 348, $z = 14.1$, $p = 1/201$; $z = 22.1$ at width 16k). But the same check bounds what may be read off the *magnitude*: the width swap alone moves the cell by 8.0 in z and 6.3 points of excess with no change of model, corpus, or null recipe. Since the nine surviving cells use four SAE suites at three released widths, cell z's are not on a common scale and no cross-cell ordering (in particular, no capability trend) is claimed; the verdicts, being per-cell tests against per-cell nulls, are unaffected.

(ib) The GPT-2 cells are uninformative, not a second negative. One might infer GPT-2's sub-threshold status from the templated corpus — the same corpus Finding 1 shows can mis-estimate sharing in both directions. We therefore ran all four GPT-2 cells on the identical $2{,}000$-document C4 slice, top-k=1500, n_{perm}=200 configuration as the headline roster (Table 25). Every cell is at or below its null ($z = -0.71$ to -0.84, $p = 1.0$), but the test has essentially no power there: GPT-2 yields *one* reciprocal best-buddy pair against a permutation null of ≈2, more than two orders of magnitude below the 348–553 nulls of every other cell. Whatever the cause — the disclosed TransformerLens

preprocessing of the `gpt2-small-res-jb` release, the narrower 768-dimensional residual, or genuinely non-reciprocal structure — a 1-versus-2 comparison cannot distinguish "no shared features" from "the statistic does not resolve at this dictionary". Note also that the confounded greedy statistic reports $z = +5.5$ to $+46.0$ on these very cells, another instance of Finding 2. We therefore do not read GPT-2 as a second below-threshold model: exactly one model in this paper (Pythia-70M, 390 observed vs. 411 ± 11 null) is measured below its null on natural text at comparable power.

Table 25: **GPT-2 natural-text cells.** Same C4 slice, top-k, and n_{perm} as Table 12; the Pythia row is repeated for scale. The GPT-2 nulls are ≈ 2 rather than ≈ 400, so these cells are underpowered, not clean negatives. Receipts: `c4_gpt2__mistral7b.json`, `c4_gemma2-2b__gpt2.json`, `c4_gemma2-9b__gpt2.json`, `c4_qwen3-1.7b__gpt2.json` (JID 14851584).

Cell	obs	null $\mu \pm \sigma$	mutual z	greedy z
GPT-2 $\leftrightarrow$ Mistral-7B	1	2.0 ± 1.4	-0.76	$+5.5$
Gemma2-2B $\leftrightarrow$ GPT-2	1	2.1 ± 1.3	-0.84	$+46.0$
Gemma2-9B $\leftrightarrow$ GPT-2	1	1.9 ± 1.2	-0.79	$+42.6$
Qwen3-1.7B $\leftrightarrow$ GPT-2	1	2.0 ± 1.4	-0.71	$+37.4$
Pythia-70M $\leftrightarrow$ Mistral-7B (for scale)	390	411.1 ± 10.9	-1.93	$+59.2$

(ii) Matching score $\neq$ utility. SAE interpretability scores correlate only weakly with steering utility (Kendall $\tau_b \approx 0.30$ across 90 SAEs; Wang et al., 2026), so a good matching score need not imply causal relevance either. This is why we accompany matching with causal, steering, and zero-shot transfer probes throughout.

(iii) Probe-transfer corpus scope. The Finding-5 matching is fit *label-free* — mutual-NN never sees the concept-family labels — but on the same prompt distribution used for probe evaluation. Given genuine co-activation, same-corpus transfer is then expected to be high. The discriminating evidence is therefore the non-matched control ($\approx$chance, $p = 1/201$) and the null pair, where the audited matching offers a single feature and transfer collapses to the control. We ran the natural hardening: matching refit on the independent 2,000-document C4 slice, evaluation unchanged on the concept prompts. It *fails*. Transfer is $0.204/0.189/0.117$ on the three surviving pairs (controls 0.10–0.16, $p = 0.07/0.35/0.62$; $K = 123/125/63$ matched features), while the C4-matched features remain concept-separable within each model (CV 0.91–0.99) and the null pair again yields a single feature (transfer = control, $p = 1.0$). The failure is stable: refitting the matching on a held-out half of the C4 slice reproduces it on the headline pair (0.180 vs. control 0.099, $p = 0.13$; $K = 110$). The matched interface is therefore *corpus-bound*: matched features agree on the distribution that defined the matching, but the correspondence does not transport concept geometry to a different prompt distribution. This is the same corpus-conditionality that Finding 1 and the (unfiltered) protocol audit (Section M) document for the matching statistics themselves. All interface claims in this paper are accordingly scoped to the matching distribution.

(iv) Selection and evaluation corpora. Three designs, stated per layer so a reader can see which claim rests on which. (1) The nine discovery cells select pairs (top-1,500 features per side by variance, then reciprocal argmax) and compute their $200\times$ permutation null on the same 2,000 C4 documents; but every null draw re-runs the argmax on the permuted rows, so the reported excess is the count of reciprocal matches above the chance level of *re-selecting* under broken alignment, and no frozen pair set is carried from selection to test (`cross_model_sae_match.py`, `match_and_null`). The half-corpus refit (`c4h_*.json`) repeats selection and null on the second 1,000 documents and reproduces every verdict; it is a stability check, not a held-out evaluation. A held-out evaluation of frozen, A-selected pairs across the full roster — pairs selected on the first 1,000 documents, scored on the disjoint second 1,000 with an evaluation-only null (n_{perm}=2,000; `-select_prompts` in `cross_model_sae_match.py`; `run_c4_heldout_frozen.sbatch`; `heldout_*.json`) — keeps every one of the nine cells above its null: held-out pooled excess 0.52–0.73, every p at the $1/2001$ floor, at 0.74–0.94 of the in-sample excess re-selected on the evaluation half itself (Qwen3$\leftrightarrow$Gemma-2 cells 0.87–0.94, Mistral-7B cells 0.74–0.82; Llama-3.2-1B$\leftrightarrow$Mistral-7B lowest at 0.74). On five of the cells, seeded random halves in place of the consecutive ones change the median held-out excess by at most 3% (Section L). The optimism this design exposes (6–26%) is larger than the 1–3% of the split-half reported below because selection and evaluation each see 1,000 documents rather than 10,000 and

only 42–75% of the frozen pairs are re-selected on the evaluation half; the two designs answer different questions and are not averaged. The one-cell interpretability variant (Gemma-2-9B↔Mistral-7B, held-out pooled correlation 0.69, `interp_heldout_*.json`) is consistent with both. (2) The fine-topic null (Table 15) has the same construction as (1): each stratified draw re-selects. The verdict-bearing lexical arm (Table 16) instead freezes the mutual-NN pairs computed on the raw features of all 20,000 documents and scores those pairs on cross-fitted residuals of the same documents — the cross-fitting is of the nuisance model, not of the pair selection — so selection enters both the raw and the residualized excess of R_{lex}. We measured what that costs at the pooled level with a split-half on the same three cells and the same registered seed-42 halves as the token-aligned rerun: pairs selected on one 10,000-document half and scored on the other, against an evaluation-half-only null (n_{perm}=2,000), with the in-sample comparator re-selecting on the evaluation half itself so both numbers sit on identical documents. Held-out excess is 0.99/0.98/0.97 of in-sample (Qwen3-1.7B↔Gemma-2-2B, Gemma-2-9B↔Mistral-7B, Llama-3.2-1B↔Mistral-7B) (both directions averaged; every held-out p at the $1/2001$ floor; `splithalf20k_*.json`, split SHA-256 identical to `tokaclean20k_*.json`), so same-corpus selection inflates the pooled excess by 1–3% in these cells. The registered Procrustes arm (Table 22) is the disjoint-half version of the residualized test — alignment and pairs fitted on one half; raw excess, residualization and null on the other — and survives in all three cells. (3) The token-aligned rerun is selection-clean by construction: a registered 50/50 document split, candidate universe and reciprocal pairs computed on split A only, every statistic and every null draw on split-B occurrences only, with the split hashes and requirement flags asserted in each `tokaclean20k_*.json` receipt (Section L).

Shared-core control (pre-registered, unrun). We pre-registered a sharper test of whether the Qwen–Gemma probe-attribution gap is mechanistic: restrict the correlation to each family's performance-recovering core (ablate-and-measure) and re-test whether the gap persists (survives ⇒ mechanistic; collapses ⇒ structural-surface effect). It is moot for the raw metric, which fails construct validity (Section 5.2).

Scope of families, languages, and scaling endpoints. The *probe-attribution* arm covers three families (Qwen3, Gemma3, Pythia) and four languages (en/zh/ja/ko), and, after the construct-validity controls, it supports no cross-lingual claim. Its family roster is narrower than the SAE-matching arm's because Llama/Mistral/DeepSeek have no `circuit-tracer`-compatible *transcoders*; this is a constraint on the v0 attribution-graph route only, and Mistral-7B and Llama-3.2-1B carry released *SAEs* and are headline models in Section 5.1. Within-family fits rest on few points (Qwen3 n=5, Gemma3/Pythia n=4), so slope CIs are wide. Caveats specific to the deferred v0 attribution-graph pipeline (backend OOM/convergence, identification assumptions, pruning thresholds, backend asymmetry) are in Section S.

Cross-model feature alignment (v0 only; superseded). The v0 design would have compared $\mathcal{F}(G_{Qwen3}) \cap \mathcal{F}(G_{Gemma3})$ as sets of (ℓ, f) tuples. That is only a heuristic — feature index f in Qwen3's transcoder is unrelated to feature index f in Gemma3's, and layer indices are relative to different L — and no such number is reported anywhere in this paper, because the graphs never converged. The principled alternative (shared-prompt activation correlation with a permutation null) *is* the SAE-matching arm of Section 5.1, which supersedes it.

Tokenizer and template confounds. Cross-lingual overlap can be inflated or deflated by tokenizer-level pre-features, which is why the v1 SAE arm mean-pools per prompt (tokenizer-grid-free) and reports depth-banded cells. Prompt templates can also activate genuinely different circuits for the same task, so every cross-family cell fixes a single shared corpus across models before comparing — and Section 5.1 shows the templated corpus mis-estimates sharing in both directions, which is why the natural-text (C4) regime carries the verdicts.

Pruning sensitivity and prompt-set size. The $(\tau_n{=}0.8, \tau_e{=}0.98)$ thresholds match `circuit-tracer` defaults. A threshold sweep was planned as a v0 robustness check but is *not applicable*: the v0 backend never converged (Section G), so no threshold-sensitivity verification exists and none is claimed. The v0 plan used prompt sets of 8–16 per family; no v0 result is reported.

R METHODS FOR THE V0 FAILURE REGIME

In our v0 setup (`circuit-tracer` v0.4.1 on the `nnsight` backend at float16, Table 2), the attribution graphs of Ameisen et al. (2025) did not converge on multistep and hallucination prompts at any Qwen3 size from 0.6B to 8B ($\tau \in \{10^{-2}, 10^{-3}\}$), even though the model answers these prompts correctly. The library documents `nnsight` as experimental and uses TransformerLens by default; we did not run that reference configuration on these prompts, so this is a failure of our pipeline, not evidence that the method cannot recover these circuits. Any universality claim built on such a pipeline must report per-task convergence and partition claims by whether the graph could be recovered; Figure 3 marks this deferred branch. Candidate substitutes (activation patching, SAE-feature sparse regression, thinking-mode rollouts, natural-language autoencoders, dedicated feature crosscoders for the cross-family arm) are surveyed below.

Beyond the per-task-convergence reporting recommended in Section 6, several methods sidestep the attribution-graph convergence failure on multistep and hallucination prompts. Activation patching (Heimersheim & Nanda, 2024; Syed et al., 2024) carries different but well-understood pitfalls and does not require graph convergence; SAE-feature sparse regression bypasses the graph at the cost of edge-level structure; thinking-mode rollouts reduce a 2-hop-per-prompt problem to a 1-hop-per-token one. Natural-language autoencoders (Fraser-Taliente et al., 2026) pair an activation verbalizer with a reconstructor (0.6–0.8 FVE on Claude Opus 4.6 / Haiku 3.5/4.5), producing text-level descriptions rather than node-and-edge graphs; their finding that multi-token claims are more factually accurate than single-token ones offers a heuristic for the hallucination half of the failure regime. For the cross-family arm, the Dedicated Feature Crosscoder of Jiralerspong & Bricken (2026) separates shared from model-exclusive features across heterogeneous architectures (Llama-3.1-8B / Qwen3-8B / GPT-OSS-20B / DeepSeek-R1-0528-Qwen3-8B), the principled shared/exclusive decomposition the abandoned v0 Jaccard alignment only approximated. A learned shared dictionary is not a drop-in replacement for our test: its shared latents are fit on the documents they would then be scored on, so a null that re-runs that selection, as Proposition 1(b) requires, would have to retrain the crosscoder in every permutation draw; we therefore compare released per-model SAEs. Agentic interpretability — an iterate-hypothesis-then-test loop that proposes method substitutions automatically — is a natural wrapper, with our negative-finding prompt set as a benchmark on which to score whether a procedure recovers circuits our fixed attribution pass did not.

S LIMITATIONS SPECIFIC TO THE V0 ATTRIBUTION-GRAPH PIPELINE

v0 backend OOM — partially resolved. The 121 GiB CUDA OOM that blocked the v0 attribution-graph build at Qwen3-1.7B was root-caused: `compute_score` in `circuit_tracer/attribution/context_nnsight.py` allocated a full-batch dtype-converted copy of the (batch, position, d_{model}) gradient tensor when grads-dtype differed from output-vecs-dtype. Our batch-chunked patch (`BATCH_CHUNK` $= 4$) reduces peak memory from $O(B{\cdot}P{\cdot}D)$ to $O(\text{CHUNK} \cdot P \cdot D)$, eliminating the OOM at Qwen3-0.6B (JID 8751067). At Qwen3-1.7B the same chunk re-encountered the OOM on $7/8$ prompts (JID 9003117) despite 139.8 GiB capacity — a numerical-scaling artifact (`BATCH_CHUNK` must scale inversely with model size), not a different bug. The iterative `compute_partial_influences` convergence failure persists as a *separate* algorithmic issue, so the v0 negative-result framing is unchanged: our pipeline does not recover circuits (zero valid `structure_features`; Section J) because the influence-propagation fixed point does not converge, not because of memory.

Other v0 caveats. (i) *Concept-grouping is hard-coded* to a small fixed set (antonym, capital-city, ...); auto-clustering by target-token semantic class was judged not worth the engineering cost at this stage. (ii) *One pruning threshold throughout* keeps results comparable; the planned (τ_n, τ_e) frontier was not run because the v0 backend never converged (Section G). (iii) *No causal verification of the cross-lingual claim*: we would report Jaccard overlap, not interventions confirming that ablating a shared feature damages both languages. (iv) *Backend*: every v0 run, Qwen3 included, used the `nnsight` backend (the run logs record it), which the library documents as experimental, slower and less memory-efficient than its default TransformerLens backend; Gemma3-PLT transcoders require `nnsight`. The default backend was not tested on Qwen3, so the non-convergence may be specific to our backend and precision choices.

Identification assumptions. Our attribution-graph results would rest on four explicit assumptions, each testable: (a) *faithful sparse-replacement* (the transcoder reconstructs the residual contribution well enough that ablating features approximates ablating the dense direction; gated by the library's L_2 reconstruction-error check); (b) *linear logit-lens identifiability* (per-layer activations decode to the target token through the model's own unembedding — what makes the v1 probe path well-defined where the graph fails); (c) *cross-language semantic alignment* (aligned en/zh/ja/ko pairs refer to the same concept, controlled by a non-aligned shuffle baseline); (d) *scale-invariant pruning thresholds* (a fixed (τ_n, τ_e) across the Qwen3 0.6B–8B family; failure-to- converge is robust at $\tau \in \{10^{-2}, 10^{-3}\}$). We do *not* assume the attribution graph is the unique causal account: activation patching and probe-based logit-lens can give different but compatible accounts.

T FULL ATTRIBUTION-GRAPH METHOD

We adopt the attribution-graph formulation of Ameisen et al. (2025). Given a frozen base model M with parameters θ, pretrained transcoders $\mathcal{T} = \{T_\ell\}_{\ell=1}^{L}$ replacing the MLP at each layer, and a prompt $x = (x_1, \ldots, x_n)$, we construct a directed graph $G(x; M, \mathcal{T}) = (V, E)$ whose nodes partition into *feature nodes* $\{(\ell, p, f) : a_{\ell,p,f} > 0\}$, *error nodes* $\{(\ell, p)_{\text{err}}\}$, *token nodes* $\{p_{\text{tok}}\}$, and *logit nodes* $\{(p_{\text{logit}}, v)\}$. Edge weights W_{ij} quantify the linearized direct effect of source j on target i; the dense adjacency is pruned via cumulative-influence thresholds (τ_n, τ_e) as in Ameisen et al. (2025), with $\tau_n = 0.8, \tau_e = 0.98$ following the `circuit-tracer` CLI defaults.

For two graphs we extract position-invariant feature identifiers $\mathcal{F}(G) = \{(\ell, f) : (\ell, p, f) \in V_G,\ \text{kind} = \text{feature}\}$ and report the Jaccard similarity. Position invariance is necessary because aligned multilingual prompts have different token lengths; for cross-model comparisons $\mathcal{F}$ uses model-relative layer indices but matched prompt sets. Per pruned graph we also compute $|V|, |E|$, density $|E|/(|V|(|V|-1))$, layer-span, edge-weight moments $(\bar{w}, \sigma_w, w_{\max})$, and mean in/out-degree.

To validate causal involvement we use `circuit-tracer`'s `feature_intervention` hook: given (ℓ, p, f, v) it clamps $a_{\ell,p,f} := v$ and re-runs the forward pass, reporting $\Delta_v = \log p_v(\text{intervened}) - \log p_v(\text{baseline})$ and the perturbed top-5 set. We use `circuit-tracer` (`safety-research/circuit-tracer`) with the `nnsight` backend for both families (our configuration default; required by Gemma3-PLT transcoders), `float16` weights, backward batch_size $= 256$, and `lazy_decoder` to fit Gemma3-12B on one A100 (v0 attribution-graph runs only). Hyperparameters are in Table 2.

U ETHICS STATEMENT

This work uses public language models and public corpora, involves no human subjects and no personal data, and makes no deployment claim. Cross-model feature-sharing claims are used to argue that interpretability findings transfer between models, which would license auditing one model by proxy for another. The ethical risk is that such transfer is asserted from a correlation that a control would have removed. We therefore report the result only after an escalating series of controls built to kill it, name the three artifacts that did not survive those controls, and scope the surviving claim to the cells where it was measured.

V AI USE DISCLOSURE

We used generative AI tools (Anthropic Claude and OpenAI Codex agents) to implement experiment code and check it for correctness, to collect public datasets, to check references, and to polish the writing. The authors take full responsibility for all content.