Discovering Cross-Language Reasoning Invariance in LLMs with Geometry-Invariant Sparse Autoencoders

Igor Bogdanov [1] **Changcheng Huang** [1]

Abstract

Multilingual language models can solve the same mathematical reasoning problem in different languages, but it remains unclear whether they rely on shared internal features or on language-specific computations that only produce similar outputs. We study this question in five models from four architecture families using the Multilingual Grade School Math (MGSM) dataset, with problems solved independently in English, German, French, Spanish, Russian, and Chinese, retaining only problems with valid reasoning traces in all six languages and replaying those traces through the model to record internal representations at multiple layers. For each model, we first use Centered Kernel Alignment (CKA) to identify layers with strong cross-language alignment. At each selected layer, we train two sparse autoencoders: a baseline reconstruction-only model and a contrastive variant introduced in this work, the Geometry-Invariant Sparse Autoencoder (GI-SAE). GI-SAE supplements the reconstruction loss with an Information Noise-Contrastive Estimation (InfoNCE) loss that trains the encoder to produce similar feature activations for traces of the same problem, regardless of language or token position. We then test whether the resulting shared features are functionally interchangeable by swapping shared feature values between languages during the model's forward pass and measuring the resulting change in output (causal patching), quantified by Kullback-Leibler (KL) divergence per shared feature. Although GI-SAE yields higher CKA and Jaccard similarity at nearly every layer, higher geometric similarity does not consistently imply greater functional interchangeability across languages. We find that cross-language feature sharing is strongly model- and architecture-dependent in this sample and appears at different depths in different models. GI-SAE primarily amplifies cross-language structure already present in each model: the pattern is model-specific, with progressive strengthening in Qwen, no functional benefit in already-saturated Gemma, and mixed layer-dependent effects in Llama and Phi.

[1]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada. Correspondence to: Igor Bogdanov <igorbogdanov@cmail.carleton.ca>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Multilingual large language models solve mathematical reasoning problems across languages, but whether they develop *shared* internal representations for reasoning, or merely produce correct answers via language-specific computations, remains an open question. The distinction has practical consequences: shared representations would allow a single interpretability analysis to cover all languages, while language-specific computations would require separate analysis for each language.

Three gaps remain in this literature: to our knowledge, no study has tested whether cross-language shared features are *functionally interchangeable* (substituting values preserves behavior); sharing has not been compared across model families under a common protocol; and it is unknown whether geometric metrics (CKA, Jaccard) reliably predict functional interchangeability.

We address these gaps by training two top-K SAEs (Gao et al., 2025) at each layer of each model: a *baseline* SAE (reconstruction only) and a *geometry-invariant* SAE (GI-SAE) (van den Oord et al., 2019) that adds an InfoNCE contrastive loss encouraging similar feature activations across languages for the same problem. We test functional interchangeability via causal patching: swapping shared feature values between languages and measuring KL divergence per shared feature.

Across five models from four families (78 layer observations, two SAEs each), we find that GI-SAE improves geometric similarity nearly everywhere but improves functional interchangeability only where baseline sharing is moderate. Geometric similarity alone does not guarantee functional

interchangeability.

We contribute a causal measurement framework for testing whether cross-language SAE features are functionally interchangeable; GI-SAE, a top-K SAE with an InfoNCE objective; and a five-model, six-language empirical study showing that baseline shared fraction stratifies GI-SAE outcomes: 83% win rate in the convergent profile (95% CI [64, 93]%), no systematic benefit in the low-sharing profile, and 6% in the saturated profile.

2. Related Work

Multilingual representations in LLMs. Language-specific neurons can steer output language (Tang et al., 2024; Kojima et al., 2024), and multilingual transformers appear to route non-English inputs through a shared latent space (Wendler et al., 2024; Tezuka & Inoue, 2025). Cross-lingual reasoning ability has been linked to neuron overlap (Hu et al., 2025), though high behavioral consistency does not necessarily imply shared internal representations (Ifergan et al., 2025). Our work studies *reasoning traces* rather than factual recall, and evaluates sharing at the level of individual SAE features with causal interventions across multiple checkpoints.

Sparse autoencoders for interpretability. SAEs decompose activations into sparse, interpretable features (Bricken et al., 2023; Huben et al., 2024); Gao et al. (2025) characterize top-K SAE scaling. In multilingual settings, Deng et al. (2025) identify language-specific features and Brinkmann et al. (2025) show that SAE features encode shared grammatical concepts with causal validation. We ask a different question: whether cross-language shared features are functionally interchangeable, not just geometrically similar.

Causal methods for evaluating representations. Causal intervention tests functional necessity (Meng et al., 2022; Conmy et al., 2023). Our protocol operates at individual SAE features. This matters because CKA can assign high similarity to functionally different representations (Davari et al., 2023); we show that geometric convergence (Huh et al., 2024) does not imply functional interchangeability.

3. Methodology and Measurement Framework

Each model solves the same math problems independently in six languages, producing a reasoning trace per (problem, language) pair. We record the model's internal state at selected layers during each trace, then train two SAEs on these recorded activations: a baseline SAE trained only to reconstruct them, whose learned features we compare across languages to measure naturally occurring sharing, and GI-SAE, which adds a contrastive loss that trains the encoder to produce similar feature activations across languages for the same problem. Comparing the two under causal patching reveals where cross-language shared features are functionally interchangeable and where they are not.

3.1. Layer Selection via CKA Survey

In a transformer, each layer adds its output to a running sum called the *residual stream* (Elhage et al., 2021). At layer l and token position t, the residual-stream vector $x_t^{(l)} \in \mathbb{R}^{d_{\text{model}}}$ is the cumulative representation that all subsequent layers read from and write to. All SAE training and causal interventions in this work operate on residual-stream vectors. Patching the residual stream at layer l modifies the state read by all downstream layers $l{+}1, \ldots, L$.

Not all layers are equally relevant. To identify informative layers, we compute linear CKA (Kornblith et al., 2019) between pre-answer residual-stream matrices for each language pair at each layer, with rows matched by problem ID (both matrices always from the same layer). We define the **pre-answer token** as the final token of the extracted reasoning span, immediately before the JSON answer block; the final JSON block and its markup are excluded. This token is used only for the CKA survey; later evaluations use multiple backward offsets. We select the contiguous depth range with highest mean pairwise CKA; this range differs across models (Table 1) because cross-language alignment peaks at different depths.

3.2. Activation Extraction and Normalization

At identified layers, we replay each reasoning trace through the model using TransformerLens (Nanda & Bloom, 2022) and record the residual-stream vector at every token position, producing one $T \times d_{\text{model}}$ matrix per (problem, language, layer) triple.

Normalization and patching convention. SAE training and geometric evaluation use z-scored activations (per-coordinate statistics from the training split). The main causal-patching sweep uses native-scale encoding; a normalized-space sensitivity check on representative layers shows the Qwen/convergent results are stable (Appendix J).

3.3. Sparse Autoencoder Architecture

We use a top-K sparse autoencoder (Gao et al., 2025) with encoder $W_{\text{enc}} \in \mathbb{R}^{d_{\text{sae}} \times d_{\text{model}}}$, decoder $W_{\text{dec}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{sae}}}$, and biases. Given a single input activation $x \in \mathbb{R}^{d_{\text{model}}}$ (one residual-stream vector at one token position and one layer):

$$\begin{aligned} z &= W_{\text{enc}}\, x + b_{\text{enc}}, \;\; f = \mathrm{ReLU}\big(\mathrm{TopK}(z, K)\big), \\ \hat{x} &= W_{\text{dec}}\, f + b_{\text{dec}}. \end{aligned} \tag{1}$$

Sparsity is enforced by the top-K bottleneck (K=128); no ℓ_1 penalty is used. The expansion factor 4× ($d_{\text{sae}} = 4d_{\text{model}}$) controls dictionary size (8,192–12,288 features; Table 1), while K controls per-input sparsity. Decoder columns are unit-ℓ_2 normalized after each step. The **baseline SAE** is trained with MSE reconstruction loss $\mathcal{L}_{\text{recon}}$; by comparing which features activate across languages under this objective, we observe naturally occurring cross-language sharing.

3.4. Geometry-Invariant SAE (GI-SAE)

GI-SAE uses the same encoder-decoder architecture (Equation 1) but adds an Information Noise-Contrastive Estimation (InfoNCE) (van den Oord et al., 2019) contrastive term to the reconstruction loss. Each training activation is labeled with its source problem ID $y_i \in \{1, \ldots, Q\}$, where Q is the number of training problems. Activations sharing the same problem ID, regardless of language or token position, form *positive pairs*; activations from different problems are *negatives*:

$$\mathcal{L}_{\text{contr}} = -\frac{1}{|B'|} \sum_{i \in B'} \log \left(\frac{\sum_{j \in P_i} \exp(s_{ij})}{\sum_{k \neq i} \exp(s_{ik})} + \epsilon \right), \quad (2)$$

where $s_{ij} = \tilde{f}_i \cdot \tilde{f}_j / \tau$ is the cosine similarity of ℓ_2-normalized features divided by temperature τ=0.1, $P_i = \{j : y_j = y_i,\ j \neq i\}$ is the positive set, $B' = \{i : |P_i| > 0\}$ excludes items without positives, and $\epsilon = 10^{-8}$ prevents $\log 0$ when no positive pairs exist in the batch. Self-pairs are excluded from the denominator.

The joint objective is:

$$\mathcal{L}_{\text{GI}} = \mathcal{L}_{\text{recon}} + w \cdot \mathcal{L}_{\text{contr}}. \quad (3)$$

We use w=1.0, selected on Qwen3-1.7B L20 and fixed across all models (Appendix F).

3.5. Backward Alignment of Reasoning Trajectories

The same reasoning step may occur at different absolute token positions across languages. We therefore anchor at the last reasoning token and measure **backward offsets** (-1 to -500). The contrastive dataset labels all tokens by problem ID; positives include same-language activations from other positions, so cross-language invariance is tested explicitly by Jaccard similarity and causal patching rather than imposed solely by the objective. This targets the answer-proximal phase of the trace, where discourse and answer-transition structure are most comparable across languages.

3.6. Causal Validation via Feature Patching

Geometric metrics measure whether the same features activate but not whether swapping their values preserves behavior. We validate with **causal patching** (Algorithm 1).

Algorithm 1 Causal feature patching

Require: Target residual r_{tgt}, donor residual r_{src}, trained SAE
Ensure: KL divergence, autoregressive flip indicator
1: $f_{\text{tgt}} \leftarrow \text{encode}(r_{\text{tgt}})$; $f_{\text{src}} \leftarrow \text{encode}(r_{\text{src}})$
2: $S \leftarrow \{j : f_{\text{tgt},j} > 0 \ \wedge \ f_{\text{src},j} > 0\}$ {shared active features}
3: $f_{\text{patched}} \leftarrow f_{\text{tgt}}$
4: $f_{\text{patched}}[S] \leftarrow f_{\text{src}}[S]$ {swap shared values from donor}
5: $r_{\text{patched}} \leftarrow \text{decode}(f_{\text{patched}}) + \big(r_{\text{tgt}} - \text{decode}(f_{\text{tgt}})\big)$ {preserve reconstruction error}
6: Run forward pass with r_{patched} replacing r_{tgt}
7: **return** $\text{KL}(p_{\text{clean}} \| p_{\text{patched}})$, $\mathbf{1}[\arg\max p_{\text{patched}} \neq \arg\max p_{\text{clean}}]$ {disruption, flip}

Each trace is replayed using its original token sequence; the main sweep uses native-scale patching (Appendix J compares to normalized-space patching). The reconstruction error is preserved (line 5), so the intervention modifies only the SAE feature subspace. Output disruption is measured by KL divergence (Kullback & Leibler, 1951) between the clean and patched next-token distributions at the final reasoning-token position. We also record an *autoregressive flip* (top-1 prediction change). Same-language pairs serve as near-zero disruption controls. Full implementation details are in Appendix D.

3.7. Evaluation Metrics and Decision Rule

Geometric metrics (representation-level). Linear CKA (Kornblith et al., 2019) measures geometric similarity between SAE-encoded activation matrices of two languages for the same problem. **Jaccard similarity** (Manning et al., 2008) at backward offset δ: the set overlap of active feature indices, defined as $J = |S_A \cap S_B| / |S_A \cup S_B|$.

Causal metric (function-level). We define the **shared fraction** as the proportion of K features active in both languages: $\phi = \overline{|S|}_{\text{cross}}/K$. The **autoregressive flip rate** (AFR) is the fraction of patched traces where the model's top-1 next-token prediction at the pre-answer position changes: $\text{AFR} = \frac{1}{N} \sum_i \mathbf{1}[\arg\max p_{\text{patched},i} \neq \arg\max p_{\text{clean},i}]$. AFR measures local disruption to the reasoning-to-answer transition; it does not require continuing generation or parsing the final numerical answer. The primary metric, **KL/shared feature**, normalizes the cross-language disruption by the number of features intervened on, corrected for the same-language baseline:

$$\text{KL/feat} = \frac{\overline{\text{KL}}_{\text{cross}} - \overline{\text{KL}}_{\text{same}}}{\overline{|S|}_{\text{cross}}}. \quad (4)$$

where $\overline{\text{KL}}_{\text{cross}}$, $\overline{\text{KL}}_{\text{same}}$, and $\overline{|S|}_{\text{cross}}$ denote means over all (problem, language pair, backward offset) evaluation in-

stances at that layer. Lower values indicate greater functional interchangeability.

Decision rule. GI-SAE *wins* at a layer when $\text{KL/feat}_{\text{GI}} < \text{KL/feat}_{\text{baseline}}$.

Key distinction. CKA compares matrix geometry (rotation-invariant); Jaccard and shared fraction compare active dictionary indices; causal patching tests whether swapping shared-index values preserves behavior. These three levels need not agree.

4. Experimental Setup

Models, families, and layers We evaluate five models from four architecture families, spanning 1.7B to 4B parameters: Qwen3-1.7B and Qwen3-4B (Yang et al., 2025), Llama-3.2-3B (Grattafiori et al., 2024), Gemma-3-4B (Gemma Team, 2025), and Phi-3-mini-4k (Abdin et al., 2024). Table 1 summarizes each model's architecture and sweep range. The two Qwen models constitute a within-family, same-generation, different-scale comparison; Gemma, Llama, and Phi each represent a distinct family.

Sweep ranges differ across models because they target the region of high cross-language CKA identified by the preliminary survey. Phi-3's sweep starts at L1 because its CKA peaks at L9 ($\sim$31% depth), unlike all other models which peak at 88–100% depth.

Common 6-language reasoning protocol We use Multilingual Grade School Math (MGSM) (Shi et al., 2023) as the evaluation corpus: 250 grade-school math problems, each translated into multiple languages. Each model solves every problem independently in six languages (en, de, fr, es, ru, zh), producing complete chain-of-thought reasoning traces via nucleus sampling (Holtzman et al., 2020) (temperature 0.6, top-p 0.95, top-k 20, max 2048 new tokens, seed $42 +$ problem index for reproducibility).

A problem is retained as "valid" only if the model solves it correctly *and* produces an extractable reasoning trace in all six languages. All results in this paper are therefore conditional on successful multilingual reasoning; they should not be interpreted as estimating feature sharing over the full MGSM distribution. This yields 15 cross-language pairs and 6 same-language control pairs per problem. Valid problem counts differ substantially across models because multilingual mathematical competence varies (Table 1).

Corpus, training, and pipeline Each problem is presented with a language-specific system prompt instructing step-by-step reasoning and a JSON-formatted answer. Reasoning traces are extracted from think tags (Qwen) or from text preceding the final JSON answer block (Gemma, Llama, Phi); the extracted text contains no JSON block or markup, though natural-language answer statements may remain in the reasoning span (full prompts and extraction rules in Appendix E). Valid problems are split into train/validation/test (70/15/15%, seed 42). Both SAE variants use Adam with learning rate 10^{-4}, batch size 256, maximum 200 epochs, and early stopping (patience 30, monitoring validation MSE). All experiments run on a single NVIDIA RTX 5090 GPU. The per-model pipeline has six stages (activation extraction, SAE training, geometric evaluation, causal patching, aggregation, analysis); each model takes 7–10 hours, with the full sweep requiring $\sim$60 GPU-hours (Appendix D).

5. Results

Cross-language feature sharing: baseline and GI-SAE The baseline SAE is trained only to reconstruct activations; it has no cross-language objective. By comparing which of its learned features activate for the same problem across languages, we observe three baseline-sharing profiles and several model-specific depth trajectories (Figure 1, left).

Gemma shares 87–96% of features across all layers; Qwen rises from 11–15% to 50–62% with depth (both Qwen models track closely); Phi and Llama remain at 1–17% and 4–6% respectively. Models of similar parameter count occupy different profiles, suggesting architecture matters more than scale. GI-SAE (Figure 1, right) increases Qwen's sharing to 16–69% and Phi's to 9–39%, but leaves Gemma unchanged and Llama only marginally improved.

5.1. Geometric Similarity Does Not Imply Functional Interchangeability

GI-SAE improves CKA ($\Delta > 0$) at nearly every observation across all five models (Figure 2). However, functional benefit does not follow uniformly:

- **Qwen** occupies the upper-right quadrant: geometric similarity *and* functional interchangeability improve together. CKA Δ +0.14–0.23 (1.7B) and +0.04–0.32 (4B); GI-SAE wins on 11/11 and 14/14 test layers, respectively.
- **Gemma** sits in the lower-right quadrant: CKA improves by +0.25–0.41, yet GI-SAE wins on 0/16 test layers.
- **Llama** achieves the largest CKA Δ in the study (up to +0.67) but wins on only 7/17 test layers; sustained functional gains appear in the deepest layers (L23–L27), with two isolated early wins (L12–L13).
- **Phi** is mixed: early layers (L1–L6) in the upper-right quadrant, mid-to-late layers in the lower-right.

Table 1. Models, layer sweeps, and dataset sizes. $d_{\text{sae}} = 4 \times d_{\text{model}}$; K=128 for all models. Sweep ranges are determined by the CKA survey. Valid problems are those solved correctly in all six languages. Activation vectors are the total number of d_{model}-dimensional residual-stream vectors extracted across all tokens, languages, and swept layers.

Model	Family	Params	d_{model}	Sweep	Valid problems	Activation vectors Train	Activation vectors Val
Qwen3-1.7B	Qwen	1.7B	2048	L16–L26 (11)	151	2.77M	0.65M
Qwen3-4B	Qwen	4B	2560	L21–L34 (14)	202	5.42M	1.06M
Llama-3.2-3B	Llama	3B	3072	L11–L27 (17)	60	0.79M	0.20M
Gemma-3-4B	Gemma	4B	2560	L15–L30 (16)	137	1.95M	0.39M
Phi-3-mini	Phi	3.8B	3072	L1–L20 (20)	63	1.18M	0.24M
Total					613	12.10M	2.54M

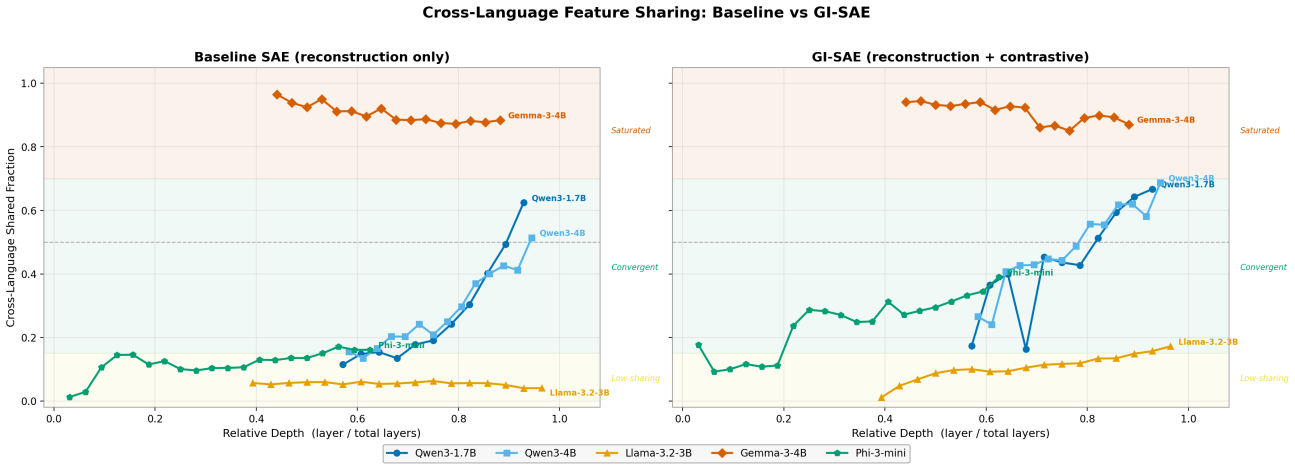

Figure 1. Cross-language shared fraction (fraction of 128 active features shared between two languages for the same problem) vs. relative depth. **Left**: baseline SAE (reconstruction only). **Right**: GI-SAE (reconstruction + contrastive). Colored bands mark the three sharing profiles: low-sharing (<15%, orange), convergent (15–60%, green), saturated (>60%, red).

This demonstrates that CKA improvement is *not sufficient* for functional interchangeability. The gap between geometric similarity and functional interchangeability is widest for models with low baseline sharing (Llama, Phi mid-layers) and for the saturated model (Gemma).

Causal controls support the functional-interchangeability interpretation: same-problem cross-language patches are less disruptive than both different-problem donors (86% pass rate) and random-value donors (93% pass rate), yielding an overall 89% pass rate across 28 condition–variant pairs (Appendix C). GI-SAE features show stronger problem-specificity than baseline features (100% vs. 71% on the different-problem control).

5.2. A Phase Diagram for GI-SAE Success

When all 78 observations are stratified by baseline shared fraction regardless of model identity (Figure 3), a useful empirical profile structure emerges:

GI-SAE reliably helps in the *convergent* profile (enough baseline sharing to build on), is counterproductive in the *saturated* profile (Gemma; swapping heavily-used shared features increases disruption), and shows no systematic benefit in the *low-sharing* profile. Baseline shared fraction thus acts as a useful diagnostic in this sample.

Table 2. GI-SAE win rates by baseline sharing profile (test split). 95% Wilson confidence intervals in brackets.

Profile	Baseline shared	Obs.	Win rate	95% CI
Low-sharing	<15%	37	49%	[33, 64]%
Convergent	15–60%	24	83%	[64, 93]%
Saturated	>60%	17	6%	[1, 27]%

5.3. Family Profiles of Cross-Language Feature Sharing

The five models reveal three recurring multilingual processing profiles (Figure 4; Llama and Qwen3-1.7B patterns are shown in the appendix).

Qwen shows progressive sharing: baseline shared fraction rises monotonically with depth, and GI-SAE wins on all layers for both models, which closely track one another. **Gemma** already shares ∼90% of features at baseline; GI-

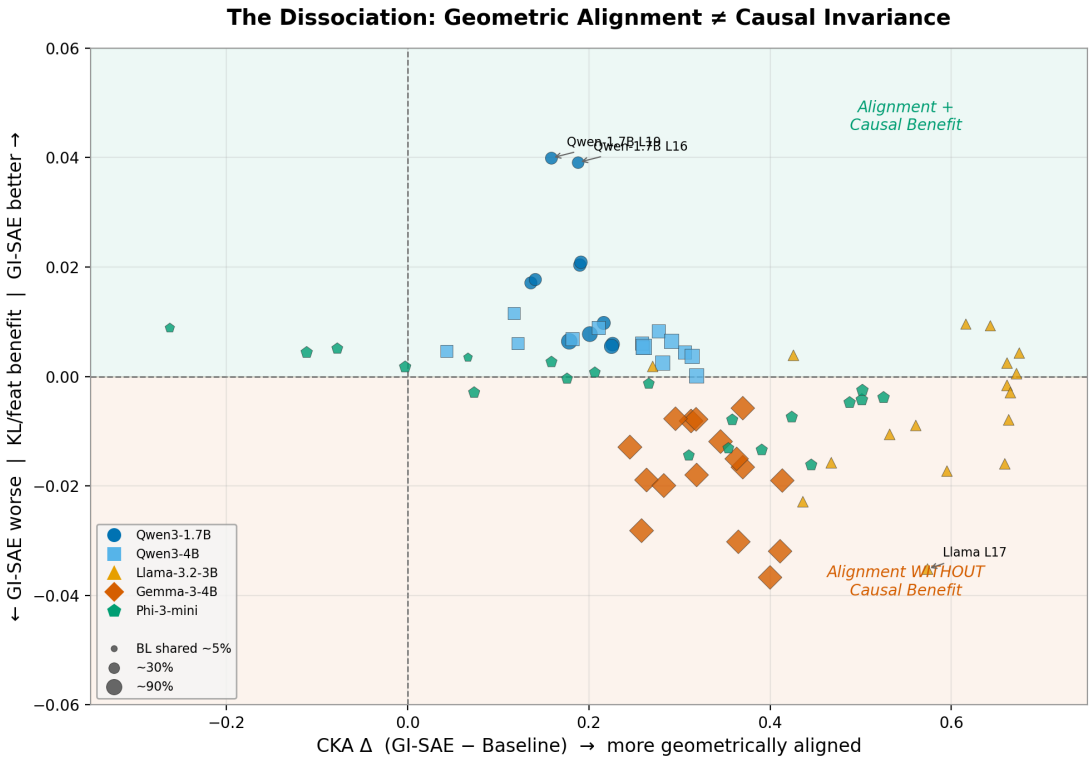

Figure 2. Each point is one (model, layer) observation. x-axis: CKA improvement from GI-SAE over baseline (geometric similarity gain). y-axis: KL/feature improvement (positive = GI-SAE has lower KL/feature). Marker size scales with baseline shared fraction. The upper-right quadrant (green) indicates both geometric and functional benefit; the lower-right (orange) indicates geometric improvement *without* functional benefit.

SAE increases KL/feature at all 16 layers, supporting the interpretation that the baseline already captures available cross-language structure. **Llama** achieves the largest CKA gains (up to $+0.67$) but functional gains appear only in the deepest layers (L23–L27, with two isolated early wins at L12–L13); mid-depth layers (L14–L22) show no improvement. **Phi** peaks in CKA at L9 ($\sim$31% depth); GI-SAE wins in early layers (L1–L6) and loses from L7 onward, mirroring Llama in reverse.

A qualitative audit of the most consistently shared features (Appendix I) shows that they encode interpretable cross-language discourse structure: answer-boundary markers that fire on sentence-final punctuation across 4–6 languages, answer-transition phrases that fire on the verb introducing the answer regardless of surface form, and reasoning-step markers active even at early layers (L3 in Phi). These categories appear across the observed model profiles, though the strongest shared features are concentrated near the reasoning-to-answer transition rather than in mid-reasoning computation.

5.4. Robustness and Training Health

Val and test KL/feature benefits are strongly aligned across the 78 layer observations (descriptive Pearson $r = 0.894$; Appendix H). A normalized-space patching sensitivity check preserves the Qwen/convergent conclusion while showing that low/mixed layers are more convention-sensitive (Appendix J). SAE training diagnostics are in Appendix F: baseline SAEs maintain >99% alive features; GI-SAE remains mostly high but drops to 84–88% for Llama and late Qwen-1.7B layers. Table 3 summarizes the five-model sweep.

6. Discussion

Interpretive scope of GI-SAE GI-SAE is informative even when it does not improve functional interchangeability. In the *saturated* profile (Gemma), GI-SAE's inability to improve sharing supports the interpretation that the baseline SAE already captures the cross-language structure present in the model; the contrastive loss has nothing to add. In the *low-sharing* profile (Llama early layers), GI-SAE achieves large geometric similarity gains (CKA Δ up to +0.67) without improving functional interchangeability, showing that CKA and Jaccard measure representation similarity, not whether

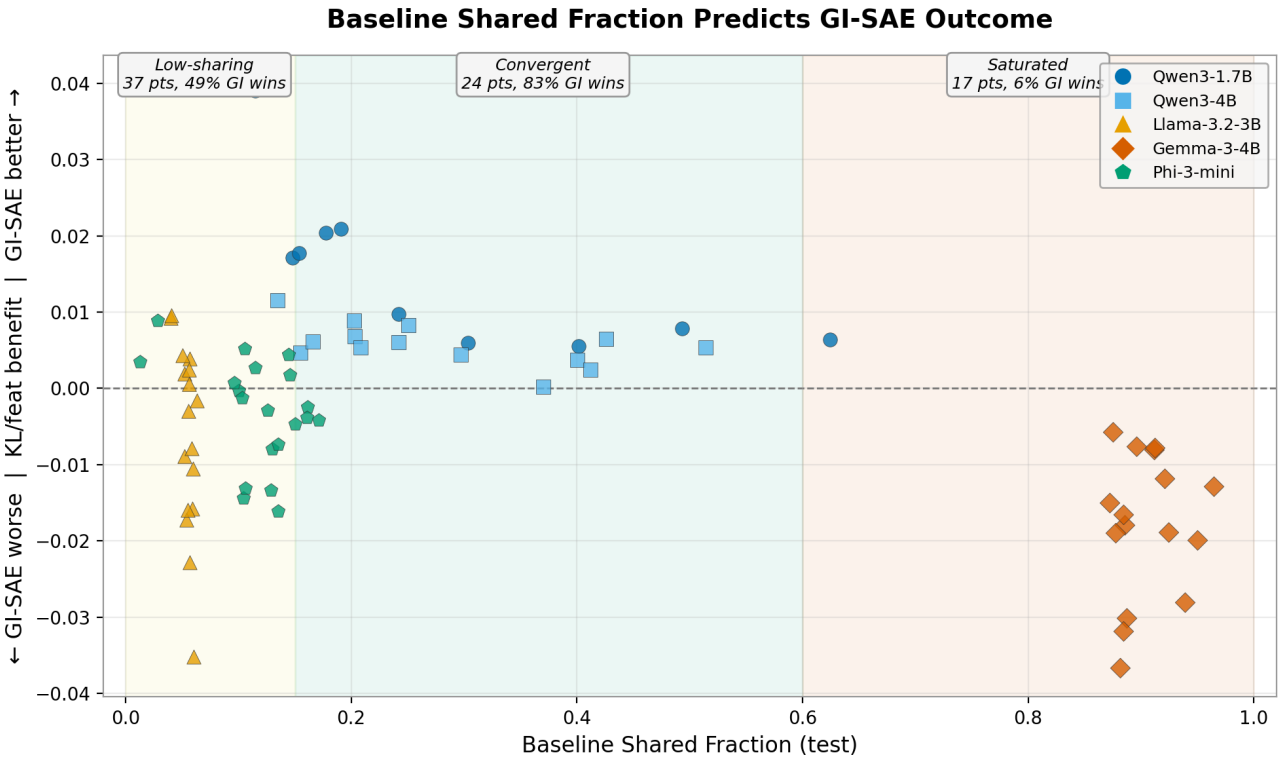

Figure 3. Phase diagram: GI-SAE KL/feature benefit vs. baseline shared fraction (test split). Points above the dashed line indicate GI-SAE wins. Three profiles emerge with distinct win rates.

Table 3. Five-model summary (test split). 95% Wilson CIs for win rates. Full per-layer results in Appendix B.

Model	Layers	Test wins	95% CI	CKA Δ	BL shared	GI shared	Profile
Qwen3-1.7B	11	11/11	[74, 100]%	+0.19	11–62%	16–67%	Conv
Qwen3-4B	14	14/14	[78, 100]%	+0.23	13–51%	24–69%	Conv
Llama-3.2-3B	17	7/17	[22, 64]%	+0.57	4–6%	1–17%	Mixed
Gemma-3-4B	16	0/16	[0, 19]%	+0.33	87–96%	85–94%	Sat
Phi-3-mini	20	7/20	[18, 57]%	+0.24	1–17%	9–39%	Mixed

shared features can be swapped without disrupting output.

This gap is a general methodological insight: wherever CKA is used to infer functional equivalence, the same gap may apply. Baseline shared fraction (Table 2) stratifies whether contrastive training helps in this sample.

6.1. Limitations and Future Directions

Limitations. The taxonomy is descriptive: each non-Qwen family is one checkpoint; Qwen provides only a within-generation scale check. (i) MGSM only; may not generalize to harder math, code, or non-mathematical domains. (ii) Llama and Phi have ∼60 valid problems vs. 151–202 for Qwen, limiting statistical power. (iii) Models up to 4B; scaling to 8B+ reduces the active dictionary fraction. (iv) Fixed hyperparameters (K=128, 4×, τ=0.1, w=1.0) across all models; w=1.0 was selected from a single Qwen3-1.7B L20 comparison and held fixed as a common intervention strength (Appendix F). (v) Contrastive positives are problem-level, not offset-matched; this may align answer-level rather than step-level information. (vi) One sampled trace per problem-language pair; multiple traces would separate representation effects from sampling variance.

Future directions. (i) Within-family replications (e.g., Gemma-2-2B, Llama-3.2-1B) to test whether sharing profiles are family-stable. (ii) CKA-only lightweight surveys across additional models to extend the phase diagram without full pipeline cost. (iii) Extension beyond mathematical reasoning to code, logical inference, and general instruction following. (iv) Adaptive contrastive weight or temperature scheduling conditioned on baseline shared fraction. (v) Investigation of what architectural or training choices produce high default feature sharing (Gemma) vs. progressive build-

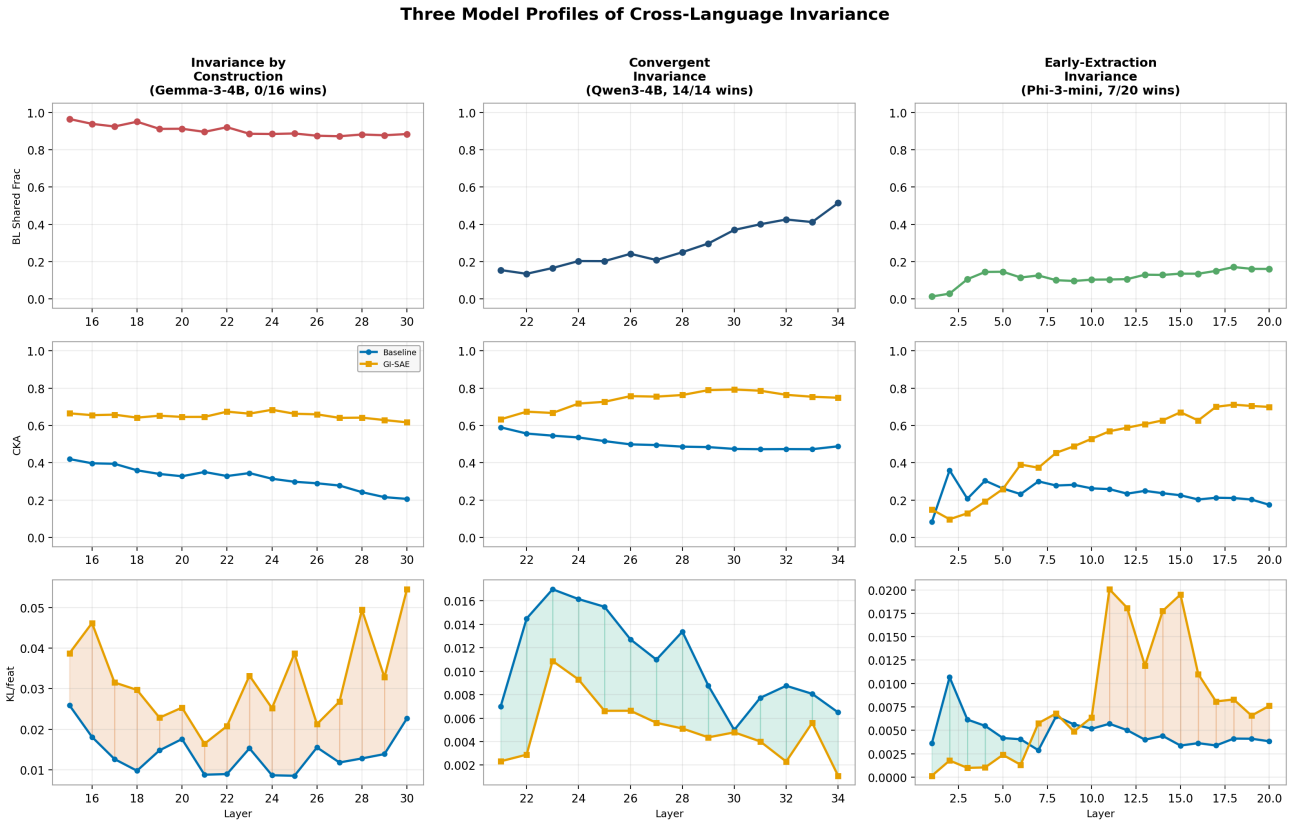

Figure 4. Three representative family profiles. **Top**: baseline shared fraction vs. layer. **Middle**: CKA (baseline blue, GI-SAE orange). **Bottom**: KL/feature (lower = greater functional interchangeability; shading marks GI-SAE advantage in green, disadvantage in orange). Left: Gemma (saturated). Center: Qwen3-4B (convergent). Right: Phi-3-mini (early-extraction).

ing (Qwen) vs. persistent language-specificity (Llama).

7. Conclusion

We introduced GI-SAE, a contrastive sparse autoencoder, and applied it alongside a baseline SAE to five models from four families under a common 6-language mathematical reasoning protocol. Cross-language feature sharing varies strongly across models (progressive in Qwen, saturated in Gemma, mixed in Llama and Phi) and geometric similarity does not reliably predict functional interchangeability: GI-SAE improves CKA at nearly every layer, yet KL/feature improves only in the convergent profile (83% win rate) and not in the saturated profile (6%).

These findings show that geometric similarity should not be used as a proxy for functional interchangeability, and that cross-language feature sharing must be analyzed on a per-model and per-architecture basis.

Impact Statement

This work is a foundational mechanistic-interpretability study of internal representations in existing open-source language models. It does not introduce new model capabilities, generate content, or release new pretrained models or datasets with direct deployment implications. Its potential positive impact is improved understanding of multilingual model behavior, which may support future interpretability and auditing research. The study uses public data and involves no human subjects or private information. We do not foresee specific negative societal impacts beyond those generally associated with research on existing large language models.

References

Abdin, M., Aneja, J., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. doi: 10.48550/arXiv.2404.14219. URL https://arxiv.org/abs/2404.14219.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter,

S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brinkmann, J., Wendler, C., Bartelt, C., and Mueller, A. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.312. URL https://aclanthology.org/2025.naacl-long.312/.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=89ia77nZ8u.

Davari, M., Horoi, S., Natik, A., Lajoie, G., Wolf, G., and Belilovsky, E. Reliability of CKA as a similarity measure in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8HRvyxc606.

Deng, B., Wan, Y., Yang, B., Zhang, Y., and Feng, F. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4563–4608, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.229. URL https://aclanthology.org/2025.acl-long.229/.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

Grattafiori, A., Dubey, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *The Eighth International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Hu, P., Liu, S., Gao, C., Huang, X., Han, X., Feng, J., Deng, C., and Huang, S. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1525–1542, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.72. URL https://aclanthology.org/2025.naacl-long.72/.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Huh, M., Cheung, B., Wang, T., and Isola, P. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Ifergan, M., Choshen, L., Aharoni, R., Szpektor, I., and Abend, O. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4630–4644, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-naacl.475/.

Kojima, T., Okimura, I., Iwasawa, Y., Yanaka, H., and Matsuo, Y. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies (Volume 1: Long Papers), pp. 6919–6971, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.384. URL https://aclanthology.org/2024.naacl-long.384/.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86, 1951. doi: 10.1214/aoms/1177729694.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008. doi: 10.1017/CBO9780511809071.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

Nanda, N. and Bloom, J. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., and Wei, J. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., and Wen, J.-R. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Tezuka, H. and Inoue, N. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31742–31792, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1618. URL https://aclanthology.org/2025.emnlp-main.1618/.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Wendler, C., Veselovsky, V., Monea, G., and West, R. 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 (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Yang, A., Li, A., Yang, B., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

A. Full Per-Model Layer Profiles

Figure 4 in the main text shows depth profiles for Gemma, Qwen3-4B, and Phi-3-mini. Here we present the remaining two models: Llama-3.2-3B and Qwen3-1.7B.

Llama-3.2-3B (low-sharing profile). Baseline shared fraction remains at 4–6% across all 17 swept layers (L11–L27). GI-SAE increases sharing to 1–17%, with gradual growth toward deeper layers. CKA improvements are the largest in the study (up to $+0.67$), yet sustained KL/feature improvements appear only in the deepest layers (L23–L27), with two isolated early wins (L12–L13). Mid-depth layers (L14–L22) show geometric gain without functional benefit, indicating that the model processes languages through largely separate pathways at those depths.

Qwen3-1.7B (convergent profile). Shared fraction increases monotonically from 11% at L16 to 62% at L26, closely tracking Qwen3-4B and confirming within-family consistency under scale change. GI-SAE wins on all 11 test layers (100% win rate, 95% CI [74%, 100%]). Peak shared fraction reaches 67% under GI-SAE. The 1.7B model achieves slightly lower peak CKA (0.69 vs. 0.79 for 4B) but follows the same progressive-sharing trajectory.

B. Full Layer-by-Layer Results

Tables 4–8 report all 78 (model, layer) observations. Shared fraction is $|S|/K$ where K=128. KL/feat ($\times 10^3$) is the test-split normalized metric (Equation 4). "Win" (✓) indicates KL/feat$_{\text{GI}}$ $<$ KL/feat$_{\text{BL}}$.

Table 4. Qwen3-1.7B layer-by-layer results (test split, L16–L26). Shared% = $|S|/128$. KL/feat $\times 10^3$. GI-SAE wins 11/11 layers.

Layer	BL Sh.%	GI Sh.%	BL KL/f	GI KL/f	BL AFR	GI AFR	Win
L16	11.5	17.4	42.0	2.9	23.6	6.7	✓
L17	14.8	36.5	46.3	29.2	25.8	30.0	✓
L18	15.4	40.0	37.4	19.6	25.8	24.4	✓
L19	13.4	16.3	43.2	3.3	24.7	7.8	✓
L20	17.8	45.3	27.0	6.6	26.9	16.4	✓
L21	19.1	43.6	26.5	5.6	23.3	18.1	✓
L22	24.2	42.7	13.9	4.2	21.1	17.2	✓
L23	30.3	51.3	9.5	3.6	17.2	16.1	✓
L24	40.2	59.4	8.1	2.6	27.8	15.6	✓
L25	49.3	64.3	10.3	2.5	31.4	15.6	✓
L26	62.4	66.7	8.5	2.1	36.4	21.9	✓

Table 5. Qwen3-4B layer-by-layer results (test split, L21–L34). GI-SAE wins 14/14 layers.

Layer	BL Sh.%	GI Sh.%	BL KL/f	GI KL/f	BL AFR	GI AFR	Win
L21	15.5	26.6	7.0	2.3	14.2	9.0	✓
L22	13.5	24.0	14.5	2.9	22.2	10.1	✓
L23	16.6	40.7	17.0	10.9	23.7	23.7	✓
L24	20.3	42.6	16.1	9.3	26.5	21.3	✓
L25	20.3	42.9	15.5	6.6	25.4	20.0	✓
L26	24.2	44.7	12.7	6.6	23.7	20.2	✓
L27	20.8	44.2	11.0	5.6	20.6	26.5	✓
L28	25.1	48.7	13.4	5.1	27.7	21.7	✓
L29	29.7	55.7	8.8	4.4	23.0	14.4	✓
L30	37.0	55.5	5.0	4.8	13.3	16.3	✓
L31	40.0	61.8	7.7	4.0	28.4	18.1	✓
L32	42.6	62.0	8.8	2.3	37.8	16.3	✓
L33	41.2	58.1	8.1	5.6	35.3	25.4	✓
L34	51.4	68.7	6.5	1.1	34.6	20.0	✓

C. Causal Patching Controls

To test whether low cross-language KL divergence reflects genuine functional interchangeability, we add two control conditions to the standard patching protocol, using the same trained SAEs (no retraining): **different-problem donor** (features from a different math problem, same source language) and **random-value donor** (same shared feature indices, but activation values permuted across those indices). If same-problem cross-language patches are less disruptive than both controls, the shared features encode problem-specific reasoning, not generic structure.

Table 6. Llama-3.2-3B layer-by-layer results (test split, L11–L27). GI-SAE wins 7/17 layers: L12–L13 and L23–L27.

Layer	BL Sh.%	GI Sh.%	BL KL/f	GI KL/f	BL AFR	GI AFR	Win
L11	5.7	1.1	3.9	26.7	8.9	5.9	—
L12	5.2	4.8	5.3	3.3	8.9	5.2	✓
L13	5.7	6.8	10.4	6.5	10.4	7.4	✓
L14	5.9	8.7	10.1	25.9	16.3	14.1	—
L15	6.0	9.7	12.3	22.9	12.6	17.0	—
L16	5.2	10.0	18.2	27.1	16.3	20.0	—
L17	6.1	9.2	14.4	49.5	11.1	25.9	—
L18	5.4	9.4	16.5	33.7	14.8	23.0	—
L19	5.5	10.5	15.7	31.7	13.3	26.7	—
L20	5.8	11.4	15.6	23.5	15.6	22.2	—
L21	6.3	11.7	15.0	16.6	17.8	22.2	—
L22	5.6	11.9	12.8	15.8	13.3	20.0	—
L23	5.6	13.4	12.7	12.1	14.1	18.5	✓
L24	5.6	13.4	12.4	9.9	15.6	13.3	✓
L25	5.1	14.9	9.4	5.1	14.1	17.0	✓
L26	4.0	15.7	12.4	3.2	15.6	10.4	✓
L27	4.1	17.2	12.6	3.0	14.1	8.9	✓

Table 7. Gemma-3-4B layer-by-layer results (test split, L15–L30). GI-SAE wins 0/16 layers.

Layer	BL Sh.%	GI Sh.%	BL KL/f	GI KL/f	BL AFR	GI AFR	Win
L15	96.5	94.1	25.9	38.7	63.3	76.7	—
L16	93.9	94.4	18.1	46.2	61.5	77.0	—
L17	92.5	93.2	12.6	31.5	42.4	53.0	—
L18	95.0	92.8	9.8	29.7	30.6	55.2	—
L19	91.2	93.5	14.8	22.9	30.0	56.4	—
L20	91.2	94.1	17.6	25.3	55.8	53.9	—
L21	89.6	91.6	8.8	16.4	27.0	28.2	—
L22	92.1	92.8	8.9	20.8	26.4	42.4	—
L23	88.5	92.3	15.3	33.2	26.1	42.7	—
L24	88.4	86.1	8.6	25.2	26.7	42.7	—
L25	88.7	86.7	8.5	38.7	17.6	52.4	—
L26	87.5	85.1	15.5	21.3	21.5	38.8	—
L27	87.2	89.1	11.8	26.8	19.7	33.6	—
L28	88.2	89.9	12.8	49.5	21.8	44.2	—
L29	87.8	89.3	13.9	32.8	20.0	39.4	—
L30	88.4	87.1	22.7	54.5	25.5	45.2	—

Table 9 reports KL/shared feature for 7 representative layers across all 5 models and both SAE variants.

Overall pass rate is 89% (25/28). The three failures are explainable edge cases: (1) Gemma BL different-problem (0.0081 vs. 0.0088): saturated profile shares $\sim$90% of features regardless of problem, so different-problem donors are nearly indistinguishable; (2) Phi L3 BL different-problem (0.0019 vs. 0.0061): very low baseline sharing ($\sim$13 features) yields too few patched features for problem-specificity to manifest; (3) Phi L15 GI random-value (0.0192 vs. 0.0195): statistical tie (ratio 0.98$\times$).

GI-SAE features show stronger problem-specificity than baseline features: GI passes the different-problem control 100% (7/7) vs. baseline 71% (5/7).

D. Pipeline and Implementation Details

D.1. Per-Model Experimental Pipeline

For each model, the pipeline proceeds through six stages: (1) activation extraction: replay all valid reasoning traces and record residual-stream vectors at each target layer for every token; (2) SAE training: train baseline SAE and GI-SAE on train-split activations at each layer; (3) geometric evaluation: compute CKA and Jaccard similarity between language pairs on validation activations encoded through each SAE; (4) causal patching: run Algorithm 1 for all (problem, language pair) combinations on both validation and test splits; (5) aggregation: collect per-layer results into a single results file; (6) analysis: generate per-model comparison plots. Each model's pipeline takes approximately 7–10 hours; the complete five-model sweep requires $\sim$60 GPU-hours on a single NVIDIA RTX 5090.

Table 8. Phi-3-mini layer-by-layer results (test split, L1–L20). GI-SAE wins 7/20 layers, all in L1–L6 and L9.

Layer	BL Sh.%	GI Sh.%	BL KL/f	GI KL/f	BL AFR	GI AFR	Win
L1	1.3	17.6	3.6	0.1	4.7	3.3	✓
L2	2.8	9.2	10.7	1.8	10.0	10.0	✓
L3	10.6	10.0	6.1	1.0	12.0	4.7	✓
L4	14.5	11.6	5.5	1.0	9.3	6.7	✓
L5	14.6	10.8	4.2	2.4	10.7	6.7	✓
L6	11.5	11.2	4.0	1.3	11.3	4.7	✓
L7	12.6	23.7	2.9	5.8	10.0	12.7	—
L8	10.0	28.7	6.5	6.8	12.0	10.7	—
L9	9.6	28.3	5.6	4.9	10.7	9.3	✓
L10	10.3	27.1	5.2	6.4	8.0	9.3	—
L11	10.4	24.9	5.7	20.1	8.0	9.3	—
L12	10.6	25.0	5.0	18.1	7.3	12.0	—
L13	13.0	31.2	4.0	11.9	7.3	18.0	—
L14	12.9	27.2	4.4	17.8	9.3	16.7	—
L15	13.5	28.4	3.4	19.5	6.7	18.0	—
L16	13.5	29.5	3.6	11.0	10.0	19.3	—
L17	15.0	31.3	3.4	8.1	12.0	14.0	—
L18	17.1	33.3	4.1	8.3	10.7	12.7	—
L19	16.1	34.5	4.1	6.6	10.7	14.7	—
L20	16.1	39.0	3.8	7.6	13.3	12.0	—

Table 9. Causal patching controls (test split). SP = same-problem cross-language, DP = different-problem, RV = random-value. Pass (✓) = control KL/feat $>$ SP KL/feat.

Model	Layer	Profile	Variant	SP	DP	RV	DP>SP	RV>SP
Qwen-4B	L29	Conv	BL	0.0088	0.0114	0.0127	✓	✓
			GI	0.0044	0.0071	0.0111	✓	✓
Qwen-1.7B	L24	Conv	BL	0.0081	0.0092	0.0194	✓	✓
			GI	0.0026	0.0032	0.0064	✓	✓
Llama	L17	Low	BL	0.0144	0.0206	0.0164	✓	✓
			GI	0.0495	0.0741	0.0519	✓	✓
Llama	L25	Low	BL	0.0094	0.0106	0.0125	✓	✓
			GI	0.0051	0.0080	0.0056	✓	✓
Gemma	L21	Sat	BL	0.0088	0.0081	0.0645	—	✓
			GI	0.0160	0.0191	0.0984	✓	✓
Phi	L3	Low	BL	0.0061	0.0019	0.0063	—	✓
			GI	0.0010	0.0011	0.0011	✓	✓
Phi	L15	Low	BL	0.0034	0.0042	0.0041	✓	✓
			GI	0.0195	0.0333	0.0192	✓	—
Pass rate (BL)					5/7	7/7	71%	100%
Pass rate (GI)					7/7	6/7	100%	86%
Overall					**12/14**	**13/14**	**86%**	**93%**

D.2. Causal Patching Protocol

For each valid reasoning trace, we replay the exact generated token sequence (replaying exact generated tokens) using TransformerLens (Nanda & Bloom, 2022). At the target layer and a backward-aligned token position, we record the target residual-stream vector r_{tgt} and the donor residual-stream vector r_{src} (from a different language's trace of the same problem).

Both vectors are encoded through the trained SAE *without* z-score normalization, because the patched vector is injected back into the model's forward pass at the original activation scale. Features active in both encodings ($f_{\text{tgt},j} > 0$ and $f_{\text{src},j} > 0$) constitute the shared set S.

The patched residual is constructed as:

$$r_{\text{patched}} = \text{decode}(f_{\text{patched}}) + (r_{\text{tgt}} - \text{decode}(f_{\text{tgt}})),$$

where f_{patched} equals f_{tgt} except at indices in S, which take donor values. Adding the reconstruction error $(r_{\text{tgt}} - \text{decode}(f_{\text{tgt}}))$ ensures the intervention modifies only the SAE feature subspace and preserves the residual component.

The patched vector replaces r_{tgt} at the same layer and position via a TransformerLens hook, and the forward pass continues over all remaining layers. KL divergence is measured between the clean and patched next-token distributions at the *final reasoning-token position* (backward offset -1), after full downstream propagation. Autoregressive flip rate (AFR) records whether the top-1 next-token prediction at that position changes. AFR is a local prediction-sensitivity metric, not a regenerated final-answer comparison.

When $|S| = 0$ (no shared active features), the intervention is skipped for that (problem, language pair, position) triple. Same-language pairs serve as a near-zero disruption control: because both encodings are identical (same trace replayed), patching produces no change.

Backward offsets are processed independently; results at each offset are averaged across all valid (problem, language pair) combinations. The primary metrics reported in the main text aggregate across all offsets.

E. Dataset Filtering and Prompts

E.1. Prompts

Each problem is presented via a system prompt instructing the model to reason in the target language and a user prompt containing the problem. Table 10 shows the system prompts for all six languages.[1]

Table 10. System prompts by language.

Lang	System prompt
en	You are a math problem solver. Reason step by step in English. At the end, write your answer as JSON: {"answer": N}.
de	Du bist ein mathematischer Problemlöser. Denke Schritt für Schritt auf Deutsch. Schreibe am Ende deine Antwort als JSON: {"Antwort": N}.
fr	Tu es un résolveur de problèmes mathématiques. Raisonne étape par étape en français. À la fin, écris ta réponse en JSON : {"réponse": N}.
es	Eres un resolutor de problemas matemáticos. Razona paso a paso en español. Al final, escribe tu respuesta como JSON: {"respuesta": N}.
ru	[Romanized] Ty reshatel' matematicheskikh zadach. Rassuzhdaj poshagovo na russkom yazyke. V kontse zapishi otvet v formate JSON: {"otvet": N}.
zh	[Romanized] Ni shi yige shuxue wenti qiujieqi. Yong zhongwen zhubi tuili. Zuihou jiang da'an xiecheng JSON geshi: {"da'an": N}.

User prompts follow the template: "Solve the problem. Think step by step: {problem}. Provide your final numerical answer strictly as JSON without units: {*key*: VALUE}." Equivalent translations are used for each language with localized JSON answer keys.

E.2. Reasoning Trace Extraction

Two strategies are used depending on model architecture:

1. **Qwen models**: explicit reasoning inside `<think>...</think>` tags is extracted via regex.

2. **Gemma, Llama, Phi**: reasoning is defined as all text preceding the final JSON answer block (e.g., `{"answer": 42}`). Trailing markdown code fences are stripped.

In both cases, the extracted `think_text` contains only reasoning with no JSON answer block or markup, though natural-language answer statements may remain in the reasoning span.

E.3. Filtering and Validity

A problem is retained only if the model solves it correctly *and* produces an extractable reasoning trace in all six languages. All results in this paper are therefore conditional on successful multilingual reasoning. This design isolates the setting where cross-language feature comparison is meaningful, but should not be interpreted as estimating feature sharing over the full MGSM distribution.

Table 1 in the main text reports valid problem counts per model: 202 (Qwen-4B), 151 (Qwen-1.7B), 137 (Gemma), 63 (Phi), 60 (Llama). The large variance reflects differences in multilingual mathematical competence.

Table 11. Generation hyperparameters (all models).

Parameter	Value
Max new tokens	2048
Temperature	0.6
Top-p	0.95
Top-k	20
Sampling	Yes
Seed	$42 +$ problem index
Padding	Left
Precision	bfloat16

E.4. Inference Parameters

F. SAE Training Diagnostics

Table 12 summarizes training health across all models. Baseline SAE achieves >99% alive features everywhere. GI-SAE alive fraction remains above 90% for most layers but drops to 84–88% for Llama and late-layer Qwen-1.7B, where the contrastive and reconstruction objectives are in strongest tension. GI-SAE validation loss is typically 1.5–2.0× baseline, with spikes above 3× at select Qwen layers (e.g., L16 and L19 for Qwen-1.7B). These spikes do not affect downstream evaluation: affected layers still produce SAEs with >90% active features.

Table 12. SAE training diagnostics (representative layers). Val loss is reconstruction MSE on the validation split. Alive% is the fraction of d_{sae} features firing on at least one validation sample.

Model	Layer	BL Val	GI Val	Ratio	GI Alive%	Note
Qwen-1.7B	L16	0.213	0.673	3.2×	99.7	High ratio
	L19	0.153	0.649	4.2×	99.6	High ratio
	L22	0.161	0.313	1.9×	89.4	
	L26	0.193	0.343	1.8×	88.7	
Qwen-4B	L21	0.225	0.684	3.0×	98.7	High ratio
	L22	0.197	0.675	3.4×	98.6	High ratio
	L29	0.167	0.314	1.9×	94.3	
	L34	0.175	0.323	1.8×	94.3	
Llama	L11	0.306	0.385	1.3×	91.7	
	L17	0.253	0.356	1.4×	86.4	
	L27	0.247	0.363	1.5×	89.9	
Gemma	L15	0.259	0.357	1.4×	96.4	
	L21	0.194	0.309	1.6×	95.9	
	L30	0.202	0.364	1.8×	95.9	
Phi	L1	0.128	0.251	2.0×	99.0	
	L9	0.226	0.373	1.7×	96.8	
	L20	0.210	0.309	1.5×	96.1	

Contrastive weight selection. We use w=1.0 as a fixed contrastive weight rather than tuning w per model or layer. This value was selected from a preliminary comparison on Qwen3-1.7B L20 using the merged single-pass implementation. At w=1.0, CKA improved from 0.503 to 0.673 and KL/feature improved from 0.0070 to 0.0003 relative to the baseline, while validation loss increased only from 31.60 to 32.29. A smaller weight (w=0.1) preserved reconstruction loss but produced weaker invariance (CKA 0.492, KL/feature 0.0055). We therefore fixed w=1.0 across the five-model sweep; adaptive weighting remains future work.

G. KL Decomposition

Table 13 decomposes the KL/shared feature metric into its numerator (raw cross-language KL) and denominator (mean shared feature count $|S|$), aggregated by sharing profile. This verifies that the metric is not an artifact of division by $|S|$.

[1]Russian and Chinese prompts are shown in romanized transliteration due to typesetting constraints.

Table 13. KL decomposition by sharing profile (test split, all layers). Raw KL = mean cross-language KL divergence. $|S|$ = mean shared feature count. $|S|/128$ = shared fraction. KL/feat = primary metric. AFR = autoregressive flip rate.

Profile	Variant	Raw KL	$\lvert S\rvert$	$\lvert S\rvert/128$	KL/feat	AFR
Low-sharing	BL	0.198	15.2	11.9%	0.0124	15.2%
	GI	0.301	30.8	24.0%	0.0114	14.9%
Convergent	BL	0.426	53.1	41.5%	0.0080	28.2%
	GI	**0.249**	**77.0**	**60.1%**	**0.0033**	**17.9%**
Saturated	BL	1.581	113.6	88.8%	0.0138	32.5%
	GI	3.516	114.4	89.3%	0.0303	47.3%

In the **convergent** profile, GI-SAE wins because both components improve: raw KL drops ($0.43 \rightarrow 0.25$, -41%) and sharing increases ($42\% \rightarrow 60\%$, $+43\%$). In the **low-sharing** profile, GI-SAE doubles sharing ($12\% \rightarrow 24\%$) but also increases raw KL ($0.20 \rightarrow 0.30$), yielding only marginal KL/feat improvement. In the **saturated** profile, sharing is ceiling-ed ($89\% \approx 89\%$) while raw KL more than doubles ($1.58 \rightarrow 3.52$), making the GI-SAE intervention destructive.

H. Statistical Uncertainty

Table 14 reports 95% Wilson score confidence intervals for GI-SAE win rates. The Wilson intervals support qualitative separation between the convergent and saturated profiles: the convergent CI lower bound (64%) does not overlap with the saturated CI upper bound (27%).

Table 14. GI-SAE win rates with 95% Wilson confidence intervals (test split).

Scope	Wins	N	Rate	95% Wilson CI
Qwen3-1.7B	11	11	100%	[74%, 100%]
Qwen3-4B	14	14	100%	[78%, 100%]
Llama-3.2-3B	7	17	41%	[22%, 64%]
Gemma-3-4B	0	16	0%	[0%, 19%]
Phi-3-mini	7	20	35%	[18%, 57%]
Low-sharing	18	37	49%	[33%, 64%]
Convergent	20	24	83%	[64%, 93%]
Saturated	1	17	6%	[1%, 27%]
Overall	39	78	50%	[39%, 61%]

Under a layer-level Bernoulli summary, the Qwen intervals lie above 50% (lower bounds 74% and 78%), while Gemma lies below 50% (upper bound 19%). Because layers within a model are correlated, we interpret these intervals descriptively. The low-sharing CI [33%, 64%] straddles 50%, consistent with the characterization that GI-SAE helps only sometimes in this profile.

I. Qualitative Feature Audit

To verify that SAE features encode meaningful cross-language discourse and answer-boundary structure rather than arbitrary sparse coordinates, we audit the most consistently shared features from four representative (model, layer) pairs, one per sharing profile. For each pair, we encode test-split activations through the SAE, identify features active in $\geq$4 of 6 languages at the pre-answer token, rank by consistency ($n_{\text{problems}} \times$ mean activation), and map peak-activation positions back to text snippets.

Three feature categories emerge:

- **Answer-boundary markers** (F9662, F8519, F3596, F3411, F5299, F6493): fire on punctuation (periods, colons) at the reasoning→answer transition. Language-invariant because final answers are always numeric.

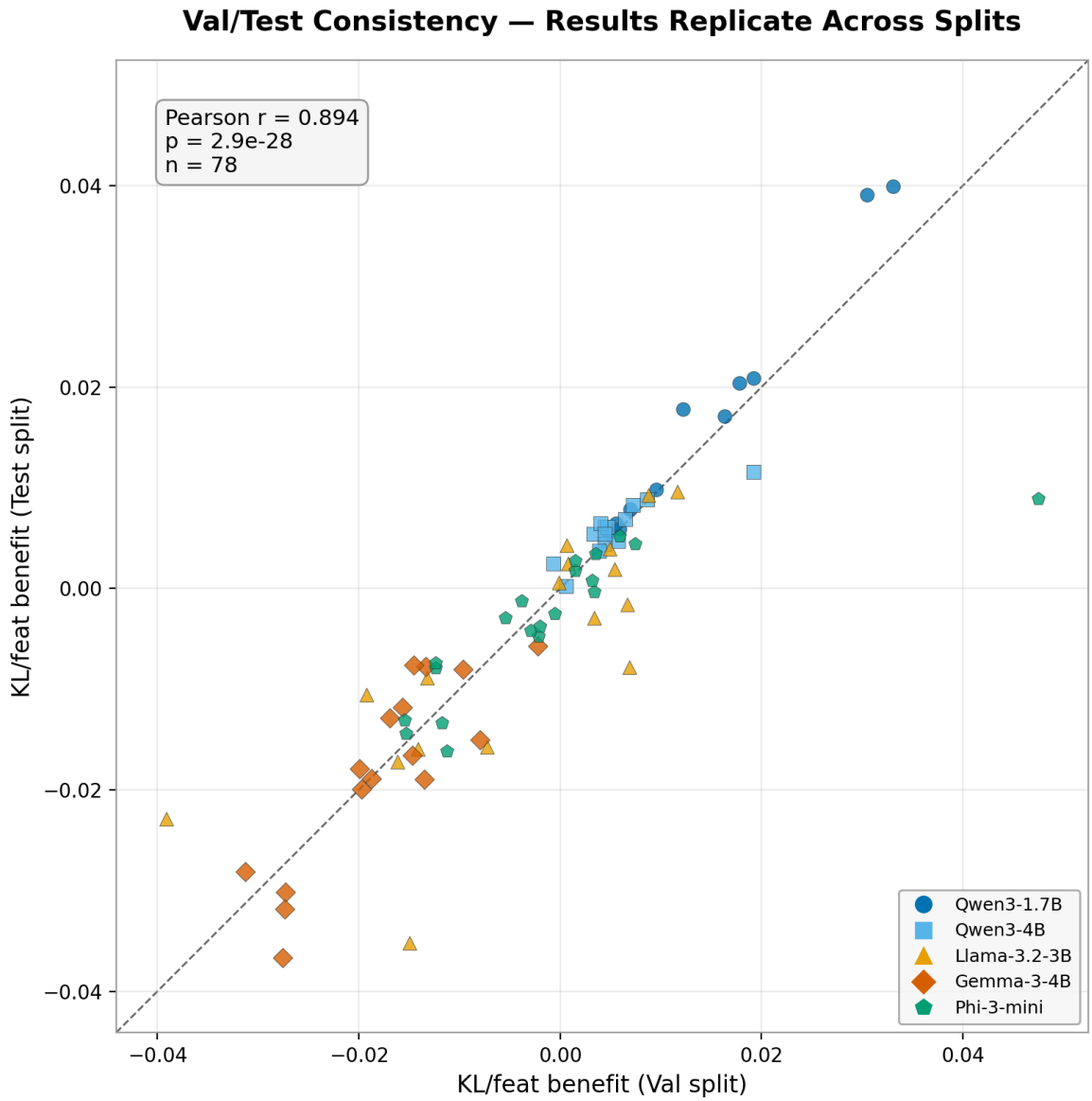

Figure 5. Val vs. test KL/feature benefit for all 78 observations (descriptive Pearson $r = 0.894$; layer observations are nested within models and should not be treated as independent samples).

- **Answer-transition phrases** (F3882): fire on the verb or copula introducing the answer ("is"/"ist"/"es"/":" [zh]). Captures syntactic role, not surface form.
- **Step/format markers** (F2162, F6445): fire on tokens initiating reasoning steps or format transitions. Active even at L3, suggesting early detection of discourse structure.

J. Patching Convention Sensitivity

The main sweep uses native-scale patching, while SAE training and geometric evaluation use z-scored activations. To test whether the causal conclusions depend on this convention, we rerun patching on representative layers using normalized-space patching: residuals are normalized before SAE encoding, patched in normalized feature space, decoded, and unnormalized before injection into the residual stream.

The two Qwen convergent layers remain GI-SAE wins under both conventions, and the aggregate win rate across valid

Table 15. Cross-language shared feature audit. Bracketed tokens mark peak activation positions. All features fire on semantically equivalent content across 4–6 languages.

Model	Layer	Feature	Langs	Category	Representative snippets
Qwen-4B	L29	F9662	5/6	Boundary	EN: "should be 29**[.]**" / DE: "$29.00**[.]**" / ZH: "29*(yuan)***[.]**"
		F8519	5/6	Boundary	EN: "answer should be 29**[.]**" / DE: "correct total**[.]**"
		F3596	5/6	Boundary	EN: "that seems right**[.]**" / FR: "that's correct**[.]**"
Gemma-4B	L21	F3411	6/6	Boundary	EN: "final answer is 30**[.]**" / DE: "gegeben**[.]**" / ZH: "30*(yuan)***[.]**"
		F5299	4/6	Boundary	EN: "Seth is 16**[.]**" / FR: "16 ans**[.]**"
Llama-3B	L17	F3882	4/6	Transition	DE: "Antwort **[ist]**" / ES: "respuesta **[es]**" / ZH: "*huida***[:]**'
		F6493	5/6	Boundary	EN: "23 jewels**[.]**" / FR: "23 bijoux**[.]**" / ZH: "*da'an***[:]**"
Phi-mini	L3	F2162	6/6	Step/format	EN: "First, **[let]**'s find" / ES: "Prim**[ero]**, calculamos" / DE: "finden **[wir]** heraus"
		F6445	6/6	Format trans.	DE: "JSON-Antwort**[:]**' / FR: "JSON est donc**[:]**" / RU: "JSON**[:]**"

Table 16. KL/feature under native-scale and normalized-space patching on representative layers. Lower is better. Gemma L21 is omitted because this diagnostic script produced no valid normalized-space interventions for that layer; we do not interpret this as evidence about Gemma's sharing profile. The full main sweep reports Gemma under the native-scale convention.

Model	Layer	BL raw	GI raw	BL norm	GI norm	Stable?
Qwen-4B	L29	0.0087	0.0044	0.0539	0.0040	Yes
Qwen-1.7B	L24	0.0081	0.0026	0.0739	0.0020	Yes
Llama	L17	0.0147	0.0500	0.0139	0.0076	No
Llama	L25	0.0093	0.0049	0.0202	0.0076	Yes
Phi	L3	0.0060	0.0009	0.0117	0.0438	No
Phi	L15	0.0034	0.0193	0.0130	0.0014	No

layers is unchanged at 4/6 under both conventions. However, normalized-space patching changes absolute shared-feature counts and flips several individual low/mixed-profile layers. We therefore interpret the Qwen/convergent conclusion as robust, but treat exact layer-level verdicts in low/mixed profiles as convention-sensitive.