Multilingual Language Models Encode Script Over Linguistic Structure

Aastha A K Verma[†,1] **Anwoy Chatterjee**[†,1] **Mehak Gupta**[1] **Tanmoy Chakraborty**[1,2]
[1]Indian Institute of Technology Delhi, New Delhi, India
[2]Indian Institute of Technology Delhi, Abu Dhabi, UAE
aastha.v1411@gmail.com anwoychatterjee@gmail.com
mehak.gupta.tech@gmail.com tanchak@iitd.ac.in

Abstract

Multilingual language models (LMs) organize representations for typologically and orthographically diverse languages into a shared parameter space, yet the nature of this internal organization remains elusive. In this work, we investigate which linguistic properties – abstract language identity or surface-form cues – shape multilingual representations. To do so, we analyze language-associated units across different model families and scales using the Language Activation Probability Entropy (LAPE) metric, and further decompose activations with Sparse Autoencoders. We find that these units are strongly conditioned on orthography: romanization induces near-disjoint representations that align with neither native-script inputs nor English, while word-order shuffling has limited effect on unit identity. Probing shows that typological structure becomes increasingly accessible in deeper layers, while causal interventions indicate that generation is most sensitive to units that are invariant to surface-form perturbations rather than to units identified by typological alignment alone. Overall, our results suggest that multilingual LMs organize representations around surface form, with linguistic abstraction emerging gradually without collapsing into a unified interlingua.

1 Introduction

Language is an amalgamation of historical accidents, cognitive constraints, and cultural evolution. It is rarely a monolith; rather, it emerges as a layered outcome of interactions among peoples, geographies, and time (Thomason and Kaufman, 1988; Toscano et al., 2008; Smith and Kirby, 2008; Beckner et al., 2009; Evans and Levinson, 2009; Michaud, 2024). Modern English illustrates this clearly: while it is taxonomically a West Germanic language, sharing core syntactic and phonological structure with German and Dutch, its lexicon is heavily shaped by Romance influence through Latin and French (Baugh and Cable, 2002; Crystal, 2003; Wardhaugh and Fuller, 2014). When a sentence such as *"the magnitude of liberty"* is processed, Latinate vocabulary is embedded within a Germanic grammatical frame (Reppucci, 2017). This raises a fundamental question for modern auto-regressive language models (LMs): do they internally preserve such linguistic distinctions, or do they abstract away surface variations into a shared, language-agnostic representation?

This question becomes especially crucial in multilingual settings. When a model processes typologically distant languages such as English, Hindi, and Chinese, does it rely on distinct internal representations for each language, or does it converge toward a shared interlingual latent space? Insights from bilingual cognition show that shared semantic representations can coexist with segregated surface-form processing (Costa and Sebastián-Gallés, 2014; Marian et al., 2003; Buchweitz et al., 2011; Miozzo et al., 2010). However, within NLP, this distinction remains underexplored in modern auto-regressive multilingual models. Investigating these models across different parameter scales allows us to determine whether the trade-offs between surface-form processing and linguistic abstraction are mere artifacts of limited capacity or fundamental properties of multilingual architectures.

Recent work has begun to probe this question (Tang et al., 2024; Kojima et al., 2024; Deng et al., 2025; Andrylie et al., 2025). Specifically, Tang et al. (2024) introduced the Language Activation Probability Entropy (LAPE) metric to identify neurons that preferentially activate for specific languages in multilingual LMs. They showed that a relatively small subset of neurons concentrated primarily in early and late layers has a strong influence on language selection and can be causally

[†]These two authors contributed equally to this work.

https://github.com/loadthecode0/multilingual-interpretability

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33685–33719

manipulated to steer the output language. Subsequent work extended this approach using Sparse Autoencoders (SAEs), the method being referred to as SAE-LAPE (Andrylie et al., 2025), which decomposes dense activations into sparse latent features and performs selection of language-associated features in the latent space using LAPE. Related intervention-based analyses similarly suggest that language control can be induced by targeting carefully selected units (Gurgurov et al., 2025; Rahmanisa et al., 2025). These studies show that language-associated units exist and can be causally manipulated, but they leave open a key question: *what linguistic properties do these language-associated units encode?*

In this work, we systematically investigate this question by analyzing language-associated units at two complementary levels: raw model neurons in the MLP sublayers that directly affect generation, and sparse latent features extracted with SAEs for interpretability. Rather than assuming these units encode abstract language identity, we test their sensitivity to orthography, word order, and deeper linguistic structure. We study these representations across different model families and scales – specifically in Llama-3.2-1B, Llama-3-8B, Gemma-2-2B, and Gemma-2-9B – analyzing languages that span Latin, Cyrillic, Devanagari, Perso-Arabic, and logographic scripts. This diverse selection ensures our observations reflect broad architectural traits rather than scale-specific bottlenecks.

Our analysis is guided by four research questions: (i) **Language vs. script**: do language-associated units encode abstract language identity, or are they primarily tied to orthographic form? Furthermore, does semantic competence in a given script guarantee representational alignment? In particular, does romanizing a language (e.g., Hindi or Chinese written in Latin script) activate the same neurons as its native script? (ii) **Robustness to structural perturbation**: how stable are these units when word order is disrupted? (iii) **Typological alignment**: do language-associated units correlate with known typological properties, such as genealogy, phonology, or syntax, as captured by `lang2vec` (Littell et al., 2017)? (iv) **Layer-wise organization**: how does the accessibility of these properties vary across network depth, and how are they organized in deeper layers?

To answer these questions, we combine sparse feature extraction with a series of controlled experiments. We analyze the behaviour of language-associated units under script romanization, structural perturbations, typological probing, and causal intervention. Across these analyses, several consistent patterns emerge:

- **Language-associated units are largely script-bound**: native and romanized variants of non-Latin languages activate almost disjoint sets of language-associated units, whereas shared scripts exhibit significant overlap. Notably, units associated with romanized non-Latin inputs align with neither their native counterparts nor English, indicating *fragmented representations* within the LMs, even when the models exhibit high semantic competence on the romanized text (c.f. Sections 4 and 8).
- **Disrupting word order has only a minor effect on unit identity**, suggesting reliance on lexical statistics or orthographic cues rather than syntactic structure (c.f. Section 5).
- **Units in deeper layers show stronger typological alignment**, indicating increased representational accessibility with depth (c.f. Section 6). Causal interventions further show that functional importance during generation is more closely associated with invariance to surface perturbations than with typological alignment alone (c.f. Section 7).

Together, these findings distinguish representational accessibility from functional necessity in multilingual LMs: language-associated units are closely tied to surface form, while deeper linguistic regularities become accessible with depth, and causal importance aligns more with invariance to surface perturbations than with representational alignment alone.

Key Takeaway

Language-associated units primarily encode surface form, and units invariant to surface perturbations play a central role in generation.

2 Related Work

Prior work has shown that multilingual language models do not form a fully language-agnostic interlingua, but instead organize representations in a partially shared space structured by language identity and similarity (Johnson et al., 2017; Pires et al., 2019; Libovický et al., 2020). Neuron-level analyses further demonstrated that language control can be localized to specific internal units. In particular, Tang et al. (2024) introduced the LAPE metric to identify language-selective neurons and showed that manipulating a small subset, often in early and late layers, can steer output language. Subsequent work confirmed that targeted interventions on such

units enable controlled language switching (Kojima et al., 2024; Gurgurov et al., 2025; Rahmanisa et al., 2025). While these studies establish the functional relevance of language-associated units, they leave open what linguistic properties these units encode.

In parallel, SAEs have been proposed to decompose dense transformer activations into more interpretable sparse features (Bau et al., 2017; Shi et al., 2025), and have recently been applied to identify language-associated features in multilingual models (Andrylie et al., 2025; Deng et al., 2025). Separately, work on typology and script effects shows that orthography and transliteration can strongly shape multilingual representations and cross-lingual alignment (Littell et al., 2017; Artetxe et al., 2020; Jauhiainen et al., 2019).

Our work connects these threads by moving from identification to interpretation: we test whether language-associated units – both raw neurons and sparse features – encode abstract linguistic structure or are primarily driven by surface-form cues. In doing so, we contextualize recent literature surrounding the "interlingua" hypothesis, which often highlights semantic alignment and shared grammatical concepts across typologically diverse languages (Wendler et al., 2024; Schut et al., 2025; Brinkmann et al., 2025; Fierro et al., 2025). Our findings complement these works by demonstrating that while semantic alignment is achievable, it does not necessitate the topological collapse of representations into a single manifold. Instead, language-neutral components coexist with a persistent set of script-specific neurons. By identifying script as a primary barrier to global unification, our work reveals that what appears as a unified space is actually deeply fragmented when orthography varies. For a more detailed discussion of prior works, we refer the reader to Appendix B.

3 Analysis Framework

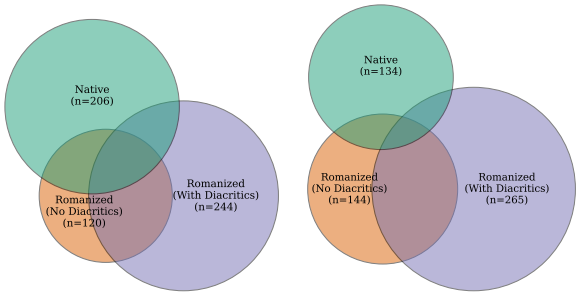
Terminology. We adopt the term *unit* as a unifying abstraction for the atomic elements of representation. Specifically, a unit refers to either a raw neuron – an individual element of the MLP's hidden activation vector – or an SAE feature representing a single direction within the latent space of the SAE. Accordingly, we define a *language-associated unit* as any unit that exhibits high selectivity for a specific target language, as quantified by the LAPE metric.

Identifying Language-Associated Units. Our analysis builds on the LAPE framework (Tang et al., 2024) and its sparse extension SAE-LAPE (Andrylie et al., 2025) to identify language-associated structure in multilingual LMs. For each transformer layer ℓ, we analyze both raw feed-forward (MLP) activations $h_\ell(x)$ and sparse latent representations obtained via pre-trained SAEs. Language association is quantified using LAPE: for each neuron or SAE feature f, we estimate its activation probability across languages and compute the entropy of this distribution. Units with low entropy and a dominant language are selected as *language-associated*, yielding a set $\mathcal{N}_{\ell,L}$ for each layer and language. Details of the LAPE and SAE-LAPE procedures along with the hyperparameters used are provided in Appendix C.

Models and Representations. We conduct experiments across multiple model families and scales, specifically Llama-3.2-1B, Llama-3-8B (Grattafiori et al., 2024), Gemma-2-2B, and Gemma-2-9B (Team et al., 2024). Prior work has applied LAPE and SAE-based analyses to Llama-family and Gemma-family models (Tang et al., 2024; Andrylie et al., 2025; Deng et al., 2025), motivating our choice of architectures and sparse decompositions. Following this line of work, we use open-sourced *Top-K* SAEs[1] for the Llama models and *JumpReLU* SAEs for the Gemma models (Lieberum et al., 2024), focusing on MLP sublayers. **For clarity of exposition, we primarily present results for Llama-3.2-1B in the core analysis sections of the main paper; corresponding analyses for Gemma-2-2B are provided in the Appendix, and validations on the larger 8B and 9B architectures are detailed in Section 8 and Appendix H.**

Experimental Design. We design a set of targeted experiments to probe what linguistic properties language-associated units encode, including (i) controlled script perturbations via romanization, (ii) robustness tests under word-order shuffling, (iii) typological probing against `lang2vec` features, and (iv) targeted causal interventions. As each experiment involves distinct language sets, perturbations, and evaluation protocols, we describe the detailed setups in the corresponding sections.

[1] https://huggingface.co/EleutherAI/sae-Llama-3.2-1B-131k, https://huggingface.co/EleutherAI/sae-llama-3-8b-32x



(a) For Raw Neurons (b) For SAE Features

Figure 1: Overlap of language-associated units for Hindi under script variation in Llama-3.2-1B. Euler diagrams show units shared among up to three languages for (a) raw neurons and (b) SAE features. Native, Romanized (with diacritics), and Romanized (without diacritics) inputs activate largely disjoint sets in both representations. Corresponding results for all languages, for both raw neurons and SAE features, and for Gemma-2-2B are shown in Figures 10 and 9 in Appendix D.2.

4 Orthography as a Barrier to Latent Language Abstraction

A central question in multilingual representation learning is whether neurons or features identified as *language-associated* encode abstract linguistic identity or merely respond to orthographic surface form. To disentangle these factors, we conduct a controlled romanization experiment that isolates script variation while holding lexical content and sentence structure fixed.

Experimental Setup. We use sentence-aligned data from the dev split of FLORES+[2], an extension of the FLORES-200 dataset (NLLB Team et al., 2024), covering a typologically and orthographically diverse set of languages spanning Abugida, Abjad, Cyrillic, Logographic, and Syllabic scripts. For each non-Latin language, we construct a parallel Romanized corpus using the ICU Transliterator (The Unicode Consortium, 2024). Where applicable, we generate two Romanized variants: one preserving diacritics and one ASCII-only version with diacritics removed. Language-associated units are identified independently for native and Romanized inputs using the LAPE criterion for raw neurons and SAE-LAPE for sparse features, and overlap is quantified using Jaccard similarity. More detailed experimental details are provided in Appendix D.

[2] https://huggingface.co/datasets/openlanguagedata/flores_plus

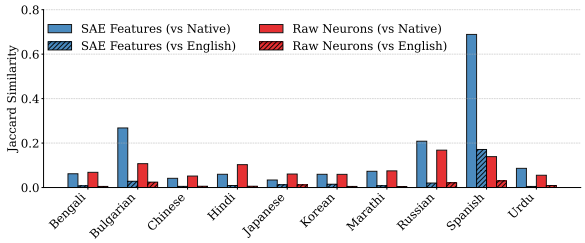


Figure 2: Jaccard similarity between Romanized and native-script or English language-associated units (**raw neurons** and **SAE features**) in Llama-3.2-1B (see Figure 8 for Gemma-2-2B). Romanized inputs exhibit low overlap with their native-script counterparts and near-zero overlap with English in both representations, indicating limited cross-script alignment without convergence to English.

Orthography Acts as a Barrier to Language Identity. If language-associated units encoded abstract linguistic identity, they would remain stable under changes in script. Instead, Figure 1 shows near-complete fragmentation under romanization for Hindi (similar observations are also made for other languages, as shown in the Figures 10 and 9). Across both raw neurons and SAE features, native-script Hindi, Romanized Hindi with diacritics, and its ASCII-only variant activate largely disjoint sets of language-associated units, even when allowing overlap across multiple languages. This fragmentation persists despite identical lexical content, indicating that language association in these models is strongly conditioned on orthographic form rather than abstract language identity.

Takeaway 1

Language-associated units are tightly bound to orthography. Even minimal script changes induce near-disjoint unit sets in both raw neurons and sparse features.

Romanization Induces an Isolated Latent Subspace. Figure 2 examines whether Romanized inputs align with native-script or English representations when considering all language-associated units. Across languages, overlap between Romanized and native-script representations remains consistently low (typically below 0.3) for both raw neurons and SAE features, with higher overlap only for Spanish, which already uses the Latin script. Crucially, overlap with English is near zero in all cases. Together, these results show that Romanization neither recovers native-script representations nor induces convergence toward English. Instead, Romanized text occupies a distinct, script-conditioned subspace that remains isolated even when consid-

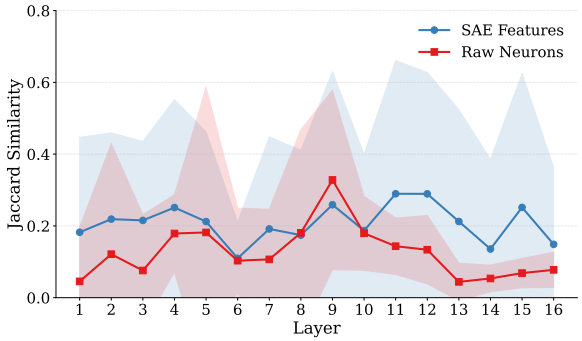


Figure 3: Layer-wise alignment between language-associated units for Native and Romanized inputs in Llama-3.2-1B (see Figure 14 for Gemma-2-2B). The **red line** denotes average Jaccard similarity for **raw neurons**, and the **blue line** for **SAE features**; shaded regions indicate standard deviation across languages. Raw neurons show a modest mid-layer increase in overlap, while SAE features remain uniformly low across depth. In all cases, alignment remains far from convergence, indicating that representational separation persists beyond input tokenization.

ering shared language-associated units, effectively forming a third latent configuration that is neither native nor English.

Takeaway 2

Romanization does not lead to Anglicization. Romanized inputs form a distinct, script-conditioned latent subspace, separate from both native-script and English representations.

Limited Intermediate Alignment and Persistent Separation. Figure 3 shows how language-associated units for Native and Romanized inputs align across layers in Llama-3.2-1B. While low overlap in early layers is expected due to disjoint token embeddings, this separation persists well beyond the input stage. **Raw neurons** exhibit a modest mid-layer increase in overlap, peaking around layer 9, but the alignment remains limited (Jaccard $\approx$ 0.3) and never approaches convergence. In contrast, **SAE features** show consistently low and flat overlap across all layers, indicating that sparse language-associated features remain strongly script-conditioned throughout the model. Together, these trends indicate that although dense activations briefly align surface-level statistics, the model ultimately maintains parallel, script-specific subspaces, revealing a limitation in abstraction rather than a trivial consequence of tokenization.

Implications for Model Capacity. The emergence of disjoint feature sets for native, Romanized, and even minor orthographic variants (e.g., diacritic vs. ASCII) points to a fundamental fragmentation of representational capacity. This aligns with recent observations that orthographic variations, such as the presence of diacritics, cause severe subword fragmentation and representational shifts in modern tokenizers and LMs (Inoue et al., 2026). We refer to this latent phenomenon as *capacity fragmentation*: the model allocates separate internal features to encode superficially different realizations of the same language. Even highly shared features fail to fully unify these variants, suggesting that many purportedly language-agnostic representations remain implicitly conditioned on script.

Scaling and Semantic Competence. Crucially, this representational fragmentation is not merely an artifact of data sparsity or limited model capacity. As we discuss in Section 8, this topological disjointness persists even in larger architectures (e.g., Llama-3-8B and Gemma-2-9B) that achieve higher semantic competence on romanized inputs.

5 Robustness of Language-Associated Features to Structural Perturbations

Section 4 illustrates that language-associated features are highly sensitive to script, with minor orthographic changes inducing substantial reorganization. We complement this with a perturbation that preserves surface form but disrupts structure by applying controlled word-level shuffling. Unlike romanization, shuffling preserves token identity, frequency, and script while breaking local word order, allowing us to test whether language-associated features depend on syntactic structure or primarily reflect token-level and distributional cues.

Setup. For each language, we construct a shuffled version of the evaluation corpus by randomly permuting word order within sentences. Language-associated units are re-identified using the same SAE-LAPE procedure applied in earlier sections. Stability is measured via Jaccard similarity between the unit sets obtained from original and shuffled text. Additional experimental details and analyses are reported in Appendix E.

Shuffling Reveals Selective Instability in Sparse Features. Figure 4 shows that many languages retain a substantial fraction of their language-associated units under shuffling, indicating limited dependence on word order. However, this

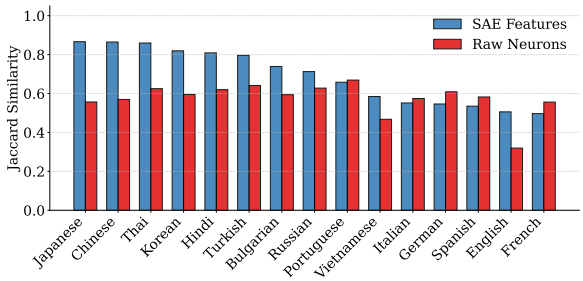


Figure 4: Jaccard similarity between language-associated units identified from original and word-shuffled text in Llama-3.2-1B (see Figure 23 for Gemma-2-2B). **Raw neurons** exhibit consistently moderate-to-high overlap across languages, indicating robustness to word-order perturbation. In contrast, **SAE features** show only *selective instability*: languages with distinctive scripts (e.g., Chinese, Japanese, Thai) remain highly stable, whereas several Latin-script languages exhibit somewhat reduced overlap, revealing sensitivity of sparse features to local distributional patterns disrupted by shuffling for these languages.

robustness varies across languages and representations. Languages with distinctive scripts such as Chinese, Japanese, Thai, Korean, and Cyrillic languages remain highly stable, with overlap often exceeding 0.7, suggesting dominance of token identity and orthographic cues. In contrast, several Latin-script languages exhibit relative reductions in overlap specifically in **SAE features**, indicating sensitivity of a subset of sparse features to local distributional or sequence-level statistics disrupted by shuffling. This selective instability is largely absent in **raw neurons**, which maintain stable overlap across languages, highlighting that dense representations encode language information redundantly, while sparse decompositions expose heterogeneity that is otherwise masked.

Activation Statistics Remain Stable. Although shuffling alters feature identity for some languages, it induces negligible changes in activation entropy or probability. Both language-level means and full distributions remain nearly identical before and after shuffling, indicating that shuffling affects *which* features are selected rather than overall activation behavior (see Figures 20, 21 and 22 in Appendix E for full distributional analyses and language-level means in case of both Llama and Gemma).

Implications. In contrast to the fragmentation induced by script changes (Section 4), word-order disruption leaves most language-associated representations intact. The limited instability that does occur is selective, appearing mainly in sparse features for languages that share script and subword statistics, and not in raw neurons.

Robustness at Scale. Furthermore, as detailed in Section 8, this robustness to structural perturbation consistently holds across larger parameter scales (Llama-3-8B and Gemma-2-9B), reinforcing that language-associated units fundamentally prioritize surface form over syntactic structure regardless of model capacity.

Takeaway 3

Language-associated units are largely insensitive to word order, while sparse features expose limited, language-dependent reliance on local distributional cues.

6 Typological Structure Revealed by Probing

Sections 4 and 5 show that language-associated units are strongly shaped by surface form: script changes induce near-complete reorganization, while word-order perturbations leave many units intact. We now ask whether, despite this surface sensitivity, model representations encode deeper linguistic structure in a linearly accessible form. Specifically, we use probing to characterize *where* typological information is concentrated and *when* it emerges across model depth.

Setup. We probe both raw MLP activations and SAE-based representations against typological features from `lang2vec` (Littell et al., 2017). For each layer, linear probes are trained with cross-validation over languages, and performance is summarized using the average of family-wise maximum R^2 scores. We report results across different neuron subsets induced by romanization and shuffling (e.g., condition-specific vs. overlap sets). Full probing details are provided in Appendix F.

Typological Structure Aligns with Invariance to Script. Figure 5 shows probing results across neuron subsets, in Llama-3.2-1B, induced by romanization (see Figure 15 for SAE-features in Llama, and Figures 16 and 17 for Gemma-2-2B). Across both raw neurons and SAE features, a consistent pattern emerges: *neurons preserved across native and romanized inputs exhibit the strongest typological alignment*. Overlap subsets dominate across genealogical, syntactic, and phonological families, while script-specific subsets (native-only or romanized-only) encode substantially weaker typological signal. This directly connects to Sec-

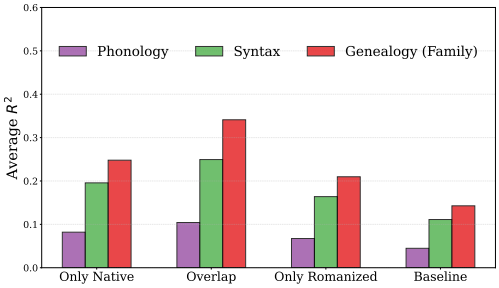


Figure 5: Average family-wise probing R^2 scores across neuron subsets induced by *romanization* in Llama-3.2-1B (raw neurons). Neurons overlapping between native and romanized inputs exhibit the strongest typological alignment, while script-specific subsets encode weaker signal. *Baseline* denotes probing over the pooled set of all neurons that were selected for either native or romanized inputs (across all layers), serving as a non-selective reference. Corresponding results for Llama-3.2-1B (SAE features) and for Gemma-2-2B using both raw neurons and SAE features are shown in Figures 15, 16, and 17 in Appendix D.3.

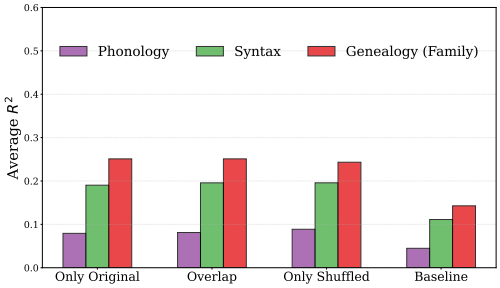


Figure 6: Average family-wise probing R^2 scores across neuron subsets induced by *word-order shuffling* in Llama-3.2-1B (raw neurons). Neurons specific to original text, shuffled text, and their overlap exhibit comparable typological alignment, indicating that sensitivity to word order is largely decoupled from typological information. *Baseline* denotes probing over the pooled set of all neurons selected for either condition, serving as a non-selective reference. Corresponding results for Llama-3.2-1B (SAE features) and for Gemma-2-2B using both raw neurons and SAE features are shown in Figures 24, 25, and 26 in Appendix E.2.

tion 4: the same units that are invariant to orthographic change are those that preferentially encode deeper linguistic structure. Together, these results indicate that typological abstraction is not tied to language-specific or script-specific units, but instead concentrates in representations that are robust to script variation.

Typological Structure Does Not Prefer Order-Invariant Units. In contrast, probing under word-order shuffling reveals a qualitatively different pattern. Figure 6 shows that typological alignment is *comparable across normal-only, shuffled-only, and overlap subsets*. This holds for both raw and sparse representations, although overall scores are lower for SAE features. Unlike romanization, invariance to word order does not preferentially select typologically informative units. This observation aligns with Section 5: while shuffling leaves many language-associated units intact, this robustness does not correspond to a privileged locus of linguistic abstraction.

Depth-Dependent Emergence of Linguistic Abstraction. While invariance determines *where* typological information resides, model depth determines *when* it becomes accessible. We illustrate this hierarchy using SAE features, where typological trends are most interpretable; raw activations show the same qualitative pattern (Appendix F). Figure 7 shows that **genealogical** properties are

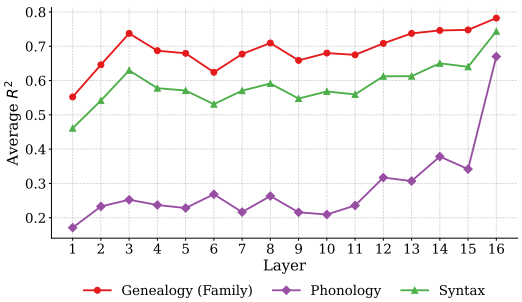


Figure 7: Average probing R^2 scores across layers for SAE features in Llama-3.2-1B, grouped by typological family. **Genealogical** properties are accessible from early layers, while more abstract features such as **phonology** emerge mainly in deeper layers. Corresponding results for raw neurons and Gemma-2-2B show the same hierarchy (Figures 27, 29).

linearly decodable from early layers, whereas more abstract **phonological** features emerge only in the deepest layers. This hierarchy suggests that linguistic abstraction is constructed gradually with depth rather than encoded uniformly across the model.

From Representational Accessibility to Functional Testing. Probing shows that typological information becomes increasingly linearly accessible in deeper layers, particularly in script-invariant representations. However, probing alone does not establish functional necessity. In Section 7, we therefore test whether units identified by their invariance properties play a causal role in generation. Furthermore, we synthesize these structural

findings with downstream semantic competence in Section 8.

Takeaway 4

Typological structure emerges with depth and is strongest in script-invariant representations. Abstraction remains distributed across units.

7 Causal Roles of Script- and Structure-Invariant Units

Sections 4-6 show how language-associated units vary with script, word order, and typological structure. We now test whether these distinctions reflect *functional necessity* during generation by performing targeted causal interventions on neuron sets defined solely by their invariance properties. Full experimental details, statistical tests, and qualitative analyses are provided in Appendix G.

Setup. All interventions are performed on raw MLP activations. While we focus our main text exposition on Llama-3.2-1B, we concurrently validate all interventions on both Llama-3-8B and Gemma-2-2B to ensure causal effects hold across different architectures and scales. Neuron sets are defined by invariance to script or word-order perturbations (Sections 4, 5). For romanization-derived sets, we perform cross-language mean replacement; for shuffling-derived sets, we apply simultaneous zero ablation across all layers. Effects are compared against matched random controls using perplexity on FLORES+ dev examples. Statistical significance is assessed via paired t-tests; exact p-values are reported in Appendix Tables 5 and 6.

Script-Invariant Neurons Support Stable Generation Under Perturbation. Using neuron sets derived from the romanization analysis (Section 4), we perform cross-language mean ablations between Hindi and English (Table 1). **Overlap neurons**, which remain active across native and romanized scripts, exhibit only mild and asymmetric perplexity changes under cross-language replacement; while statistically significant ($p < 0.05$; Table 6), these effects are small, indicating that these neurons occupy a largely script-invariant subspace. In contrast, **only-native neurons** show extreme sensitivity: replacing English-only-native activations with Hindi means causes severe degradation, while the reverse yields large apparent perplexity improvements. Qualitative inspection reveals that the latter corresponds to language switching rather than improved modeling, with generations collapsing into fluent English (Appendix G). Crucially, these effects generalize: in both Llama-3-8B and Gemma-2-2B (Appendix Tables 6 and 8), ablating only-native Hindi neurons causes dramatic perplexity changes (e.g., $PPL_{\text{ratio}}^{\text{target}} = 7.74$ in Llama-3-8B), confirming extreme sensitivity. Together, these results causally validate Section 4, showing that script-specific neurons anchor surface realization and language identity, while script-invariant neurons support stable generation under orthographic perturbation.

Language	Neuron set	$PPL_{\text{ratio}}^{\text{target}}$	$PPL_{\text{ratio}}^{\text{random}}$
English	overlap	0.95	0.99
English	only-native	1.50	0.96
Hindi	overlap	1.05	0.98
Hindi	only-native	0.31	0.97

Table 1: Cross-language mean ablations for romanization-derived neuron sets in Llama-3.2-1B (see Table 8 for Gemma-2-2B). $PPL_{\text{ratio}}^{\text{target}}$ denotes perplexity relative to clean runs, and $PPL_{\text{ratio}}^{\text{random}}$ reports the same for matched random controls. All target effects are statistically significant ($p < 0.05$; Table 6). Ratios below 1 reflect language switching rather than improved modeling.

Word-Order-Invariant Neurons Support Core Language Modeling. We next examine neuron sets derived from the shuffling analysis in Section 5 using simultaneous zero ablation (Table 2). Across all languages, **overlap neurons** – those that remain active under word-order shuffling – cause substantially larger perplexity increases than matched random controls, with all effects statistically significant ($p < 0.05$; Table 5). In contrast, **only-unshuffled neurons** produce much weaker effects and often reduce perplexity, indicating that order-sensitive signals are largely redundant for generation. This causal dissociation mirrors the identification results in Section 5: neurons invariant to structural perturbation are functionally necessary for stable language modeling, while order-sensitive neurons encode auxiliary or brittle patterns. Qualitatively, only overlap-neuron ablations induce systematic failures such as within-word script mixing and abrupt language switching (Appendix Figure 32), further supporting their causal role. These causal dynamics are highly consistent across architectures, with both Llama-3-8B and Gemma-2-2B exhibiting similar severe degradation and identical qualitative failure modes specifically when shuffling-invariant overlap neurons are ablated (see Appendix Tables 5 and 7).

Language	Neuron set	$PPL_{\text{ratio}}^{\text{target}}$	$PPL_{\text{ratio}}^{\text{random}}$
English	overlap	1.12	0.95
English	only-unshuffled	0.96	1.04
Hindi	overlap	2.79	1.06
Hindi	only-unshuffled	1.08	0.95

Table 2: Zero-ablation results for shuffling-derived neuron sets in Llama-3.2-1B (see Table 7 for Gemma-2-2B). $PPL_{\text{ratio}}^{\text{target}}$ reports perplexity relative to clean runs after ablating the specified neuron set, and $PPL_{\text{ratio}}^{\text{random}}$ reports the same for matched random controls. All overlap-neuron effects differ significantly from random controls ($p < 0.05$; Table 5).

Implications for Language Control and Abstraction. Across both romanization- and shuffling-based interventions, causal importance consistently tracks invariance to surface perturbations. Neurons that remain stable under script or word-order variation are more functionally necessary for generation, whereas surface-sensitive neurons primarily anchor realization. While probing in Section 6 shows that typological structure becomes increasingly decodable with depth, our causal interventions do not isolate a small set of neurons whose manipulation selectively disrupts such structure. Instead, causal effects are associated with invariance properties, suggesting that language control in these models is mediated by robustness to surface variation rather than by a single, localized abstraction module.

Takeaway 5

Causal importance aligns with invariance to surface perturbations. Neurons stable under script or word-order variation are necessary for generation, while probing reflects representational structure rather than direct control.

8 Discussion and Scaling Analysis

Our results show that multilingual models do not converge to a fully abstract interlingua. Instead, representations are organized around surface-form cues, especially script, while deeper layers support abstraction without unifying script-conditioned subspaces.

Robustness at Scale and Semantic Competence. To ensure these findings are not limited by model capacity, we validated our core experiments on larger architectures (Llama-3-8B and Gemma-2-9B). As shown in Appendix H (Figures 34 – 37), representational fragmentation persists at scale: romanized inputs maintain low overlap with native-script counterparts and fail to converge toward English. Conversely, robustness to word-order shuffling remains consistently high across these larger models (c.f. Figure 38, Appendix H). Furthermore, to rule out data sparsity (i.e., the models simply failing to comprehend romanized text), we evaluated translation performance. As detailed in Appendix H (Table 9), larger models achieve substantial semantic competence on romanized inputs. This confirms that models possess the requisite knowledge, but internally process different scripts through disjoint subspaces as a persistent architectural trait rather than a training deficiency.

Implications for Cross-Lingual Transfer. The strong dependence on orthography suggests that cross-lingual transfer is more fragile than often assumed. Romanized inputs neither recover native-script representations nor align with English, even when considering shared language-associated units. Instead, they occupy distinct latent subspaces, helping explain why transliteration or script normalization alone yields limited gains without explicit adaptation or supervision.

Orthography, Control, and Robustness. Our findings offer an alternative explanation for prior observations that changing the language or script of a prompt can alter model behavior, including safety-related responses (Deng et al., 2024; Yong et al., 2023). If language-associated units are tightly coupled to orthography, script changes may route inputs through different internal subspaces, yielding divergent outputs. This suggests that some language-based control and jailbreak effects may stem from surface-form routing rather than semantic differences.

9 Conclusion

In this study, we show that language-associated units in multilingual LMs are primarily organized around surface-form cues, with script acting as a primary barrier. While typological structure becomes more accessible at deeper layers, our causal analyses reveal that stable generation depends mostly on units invariant to surface perturbations. By validating these findings across model scales, we confirm that LMs process different scripts through disjoint latent spaces despite high semantic competence, showing that multilingual abstraction remains limited by orthography rather than forming a fully unified interlingua.

Limitations

While we validate our core representational and causal findings on models up to 9B parameters, evaluating these phenomena on massive-scale frontier models remains an important direction for future work, as models at much larger parameter scales may eventually exhibit different trade-offs between surface-form routing and abstraction. In addition, our analysis centers on feed-forward (MLP) activations and their sparse decompositions, and does not examine other architectural components such as attention heads or embedding layers. Finally, while our interventions assess the causal role of identified units at inference time, we do not study the training dynamics through which these representations emerge.

Ethical Considerations

This work analyzes internal representations of multilingual LMs using publicly available pretrained models and established linguistic resources. While we generate and release systematically perturbed (romanized and shuffled) versions of existing evaluation sets, we do not deploy systems in user-facing contexts or evaluate downstream social applications. Our findings highlight how script and surface-form variation can influence internal processing, with potential implications for robustness and safety generalization across languages. We emphasize that our goal is interpretability and analysis rather than exploitation, and we do not propose methods for bypassing safeguards or inducing harmful behavior. Overall, this work aims to support safer and more transparent multilingual model development by clarifying how language-associated representations are organized internally.

Acknowledgements

Anwoy Chatterjee gratefully acknowledges the support of the Google PhD Fellowship. Tanmoy Chakraborty acknowledges the support of the Anusandhan National Research Foundation (Grant no: DST/INT/USA/NSF-DST/Tanmoy/P-2/2024) and Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence. The authors acknowledge the support of the Google GCP Grant.

References

Lyzander Marciano Andrylie, Inaya Rahmanisa, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. Sparse autoencoders can capture language-specific concepts across diverse languages. *Preprint*, arXiv:2507.11230.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3319–3327. IEEE Computer Society.

A.C. Baugh and T. Cable. 2002. *A History of the English Language*. Prentice Hall.

Clay Beckner, Richard Blythe, Joan Bybee, Morten Christiansen, William Croft, Nick Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language Learning - LANG LEARN*, 59:1–26.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. 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)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Augusto Buchweitz, Svetlana V Shinkareva, Robert A Mason, Tom M Mitchell, and Marcel Adam Just. 2011. Identifying bilingual semantic neural representations across languages. *Brain Lang*, 120(3):282–289.

Albert Costa and Núria Sebastián-Gallés. 2014. How does the bilingual experience sculpt the brain? *Nat Rev Neurosci*, 15(5):336–345.

D. Crystal. 2003. *English as a Global Language*. Canto (Cambridge University Press). Cambridge University Press.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. 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)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16052–16106, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef Van Genabith, and Simon Ostermann. 2025. Language arithmetics: Towards systematic language neuron identification and manipulation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2911–2937, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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

Go Inoue, Bashar Alhafni, Nizar Habash, and Timothy Baldwin. 2026. Do diacritics matter? evaluating the impact of Arabic diacritics on tokenization and LLM benchmarks. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 426–442, Rabat, Morocco. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019. Language and dialect identification of cuneiform texts. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 89–98, Ann Arbor, Michigan. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. 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)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *Preprint*, arXiv:1911.03310.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. 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*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15077–15107, Suzhou, China. Association for Computational Linguistics.

Viorica Marian, Michael Spivey, and Joy Hirsch. 2003. Shared and separate systems in bilingual language processing: Converging evidence from eyetracking and brain imaging. *Brain and Language*, 86(1):70–82. Understanding Language.

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

Jérôme Michaud. 2024. A complex systems perspective on language evolution. In *The Evolution of Language : Proceedings of the 15th International Conference (EVOLANG XV)*, The Evolution of Language Conferences, pages 374–382.

Michele Miozzo, Albert Costa, Mireia Hernández, and Brenda Rapp. 2010. Lexical processing in the bilingual brain: Evidence from grammatical/morphological deficits. *Aphasiology*, 24(2):262–287.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Inaya Rahmanisa, Lyzander Marciano Andrylie, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. Unveiling the influence of amplifying language-specific neurons. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 919–968, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Leah Estel Reppucci. 2017. Speaking denglish: Exploring the impact of denglish and anglicisms in german culture and identity.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2025. RomanLens: The role of latent Romanization in multilinguality in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26410–26429, Vienna, Austria. Association for Computational Linguistics.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Wei Shi, Sihang Li, Tao Liang, Mingyang Wan, Guojun Ma, Xiang Wang, and Xiangnan He. 2025. Route sparse autoencoder to interpret large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6815, Suzhou, China. Association for Computational Linguistics.

Kenny Smith and Simon Kirby. 2008. Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3591–3603.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. 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)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

The Unicode Consortium. 2024. ICU: International components for unicode. Version 78.1.

Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*, 1 edition. University of California Press.

Joseph C. Toscano, Lynn K. Perry, Kathryn L. Mueller, Allison F. Bean, Marcus E. Galle, and Larissa K. Samuelson. 2008. Language as shaped by the brain; the brain as shaped by development. *Behavioral and Brain Sciences*, 31(5):535–536.

Katharina A. T. T. Trinley, Toshiki Nakai, Tatiana Anikina, and Tanja Baeumel. 2025. What language(s) does aya-23 think in? how multilinguality affects internal language representations. In *Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models*, pages 159–171, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

R. Wardhaugh and J.M. Fuller. 2014. *An Introduction to Sociolinguistics*. Blackwell Textbooks in Linguistics. Wiley.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. 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)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-resource languages jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Appendix Contents

Below we provide an overview of the appendix. These sections are intended to support the core claims by providing methodological details and extended scaling results.

- **Appendix A: Frequently Asked Questions (FAQs).** *Addresses common questions regarding representational fragmentation, scale, and causal control.*

- **Appendix B: Extended Related Work.** *Provides a detailed discussion of prior work on multilingual representations, language-associated units, sparse autoencoders, typology, and script effects.*

- **Appendix C: Identifying Language-Associated Units with LAPE and SAE-LAPE.** *Explains the identification frameworks and provides the hyperparameter thresholds used for both neurons and sparse features.*

- **Appendix D: Script Perturbation Experiments (Romanization).** *Describes dataset construction, transliteration procedures, and distributional analyses supporting the script romanization findings.*

- **Appendix E: Structural Perturbation Experiments (Word Shuffling).** *Reports supplementary results on shuffled inputs, aggregate overlap analyses, and stability of activation statistics.*

- **Appendix F: Probing Typological Structure Across Layers.** *Details the ridge regression probing setup and provides comparative layer-wise informativeness across models.*

- **Appendix G: Causal Interventions on Invariant Neuron Sets.** *Provides simultaneous ablation protocols and qualitative failure modes showing script mixing during generation.*

- **Appendix H: Extended Scaling and Semantic Competence Analysis.** *Documents the persistence of representational fragmentation at the 8B and 9B scales and provides translation benchmarks ruling out data sparsity as an explanation.*

A Frequently Asked Questions (FAQs)

1. **Do language-associated units imply the existence of a universal interlingua?** No. While language-associated units are clearly identifiable and can influence model behavior, our results show that they are predominantly sensitive to surface-form cues such as script and token distribution.

2. **Is the observed script sensitivity simply an artifact of tokenization?** Tokenization necessarily introduces distinct input embeddings across scripts, but our analysis goes beyond early-layer effects. We observe that alignment remains low even in intermediate layers, indicating that script

sensitivity is not merely a tokenizer artifact but reflects persistent representational fragmentation within the model.

3. **Why use 1B and 2B models for the main exposition?** We center our primary exposition on Llama-3.2-1B and Gemma-2-2B to enable extensive, computationally intensive representational sweeps and causal interventions across many layers and languages. However, to ensure our findings are not artifacts of limited capacity, we explicitly validate our core experiments on larger models (Llama-3-8B and Gemma-2-9B), confirming these representational properties hold at scale.

4. **Do these findings generalize to larger models?** Yes. As detailed in our scaling analysis, we validate our findings on Llama-3-8B and Gemma-2-9B. We observe that representational fragmentation under script variation, as well as robustness under structural perturbation, persist at these larger scales. Crucially, this fragmentation remains even though these models exhibit strong semantic translation competence on romanized inputs, demonstrating that script-conditioned subspaces are a persistent architectural trait rather than a symptom of undertraining or data sparsity.

5. **Does strong probing performance imply functional importance?** No. Probing reveals that typological properties become increasingly linearly accessible in deeper layers, but causal interventions show that functional importance aligns with invariance to surface perturbations. This reinforces the view that linear decodability does not imply causal control.

6. **Why analyze both raw neurons and SAE features?** Raw neurons directly govern model behavior, while SAE features provide an interpretable decomposition of these activations. Analyzing both allows us to separate functional relevance from interpretability and avoid over-attributing abstract meaning to sparse features alone.

7. **What is the main takeaway for interpreting *language-associated neurons*?** Language-associated units exist and matter, but they primarily reflect surface-form processing rather than abstract language identity.

B Extended Related Work

B.1 Language-Associated Units and Multilingual Representations

Understanding how multilingual LMs encode language identity has become a central question in interpretability and cross-lingual modeling. Early multilingual neural machine translation (NMT) systems already suggested that jointly trained models do not form a fully language-agnostic interlingua, but instead organize representations in a partially shared space structured by language identity and similarity (Johnson et al., 2017; Kudugunta et al., 2019). Subsequent analyses showed that encoder representations cluster by genealogical and typological proximity, with high-resource languages occupying more stable regions of the latent space (Pires et al., 2019; Libovický et al., 2020).

More recently, investigation at the neuron level has provided evidence that language identity can be localized to specific internal units. Tang et al. (2024) introduced LAPE to identify neurons that preferentially activate for individual languages in multilingual LMs, showing that a small subset of neurons, often concentrated in early and late layers, exerts disproportionate control over language selection. Contemporary works also showed that targeted interventions on such neurons can reliably steer output language, even without modifying input prompts (Kojima et al., 2024; Gurgurov et al., 2025; Rahmanisa et al., 2025). These observations establish that language control is not purely emergent at the output layer but is mediated by identifiable internal mechanisms.

Earlier representational studies, however, caution against interpreting such units as encoding abstract language identity (Wu and Dredze, 2020; Libovický et al., 2019). Analyses of multilingual NMT and representation spaces show substantial mixing across languages, particularly in middle layers, with language separation re-emerging closer to the output where lexical constraints dominate (Kudugunta et al., 2019). This layered organization parallels findings in bilingual cognition, where shared semantic representations coexist with partially segregated lexical and orthographic processing streams (Marian et al., 2003; Costa and Sebastián-Gallés, 2014).

Our work builds on this literature but departs in emphasis. Rather than asking whether language-associated units exist, we ask what linguistic properties they encode. Specifically, we test whether

such units reflect abstract language identity or are instead driven by surface-form cues such as script and token distributions, a distinction that remains underexplored in prior neuron-level studies. While our primary exposition focuses on highly capable 1B and 2B parameter auto-regressive models to enable computationally intensive feature sweeps, we explicitly validate our core findings on larger architectures (up to 9B parameters) to ensure our conclusions regarding orthography and abstraction hold at scale.

B.2 Sparse Autoencoders and Feature-Level Interpretability

SAEs have recently emerged as a promising tool for disentangling dense transformer activations into more interpretable, monosemantic latent features. The central idea – that sparsity can separate overlapping signals into distinct dimensions – has strong precedents in vision, where network dissection methods link individual units to human-interpretable concepts (Bau et al., 2017). In language models, sparse methods have been shown to isolate features corresponding to factual recall, formatting, or syntactic regularities that are difficult to identify in dense representations (Huben et al., 2024; Marks et al., 2025).

Several recent works extend SAEs to large language models at scale. For instance, recently Shi et al. (2025) proposed RouteSAE which introduces routing mechanisms that propagate sparse features across layers, improving interpretability while maintaining model performance. Open-source SAE frameworks further demonstrate that sparse latents can support causal interventions and analyses in modern transformer models (Lieberum et al., 2024). In multilingual settings, Andrylie et al. (2025) and Deng et al. (2025) show that SAE features can align with semantic concepts across languages, motivating the use of sparse representations for cross-lingual interpretability.

Our work leverages this progress but reframes the goal. We first identify language-associated sparse features as well as raw model neurons, by using SAE-LAPE (Andrylie et al., 2025) and LAPE (Tang et al., 2024) respectively. We then systematically analyze their sensitivity to script, word order, and typological structure. Unlike prior studies that focus primarily on semantic or task-level concepts, we center our analysis on linguistic abstraction, explicitly separating representational alignment (as revealed by probing) from functional necessity (as tested via causal intervention), echoing critiques of probing as a standalone interpretability tool (Hewitt and Liang, 2019; Belinkov, 2022).

B.3 Typology, Script, and Romanization Effects

Linguistic typology has long been used to study cross-lingual similarity and transfer in multilingual models. The URIEL and `lang2vec` framework provides structured vectors encoding genealogical, geographical, phonological, and syntactic properties for various languages (Littell et al., 2017). Subsequent work shows that typological information becomes increasingly linearly accessible in deeper layers of multilingual transformers, suggesting a gradual emergence of abstraction (Rama et al., 2020).

Orthography and script introduce an additional, often confounding, dimension. Prior work in multilingual language identification shows that script cues dominate early decisions, and that romanized or transliterated text can significantly degrade performance when script information is not explicitly modeled (Jauhiainen et al., 2019). In representation learning, transliteration and script normalization have been shown to alter clustering structure in multilingual embedding spaces, sometimes improving transfer but often creating mismatches between surface form and linguistic identity (Artetxe et al., 2020; Moosa et al., 2023).

Recent interpretability studies suggest that these effects extend to internal model mechanisms. Analyses of bilingual and multilingual models show that changing script can reroute activations through different internal pathways, even when lexical content is preserved (Saji et al., 2025; Trinley et al., 2025; Muller et al., 2021; Lu et al., 2025). Our work builds on these observations by systematically comparing native-script and romanized inputs under a unified neuron- and feature-identification framework, revealing that script changes induce near-complete reorganization of language-associated units. Importantly, we show that this fragmentation persists even in deeper layers where typological information is linearly decodable, indicating that abstraction and control are distributed across parallel, script-bound subspaces rather than unified into a single interlingua.

C Identifying Language-Associated Units with LAPE and SAE-LAPE

This appendix summarizes the methods used to identify language-associated units in our analysis.

C.1 LAPE for Raw Neurons

Language Activation Probability Entropy (LAPE) quantifies how selectively an individual neuron responds to different languages. Given a multilingual corpus, for each neuron j at layer ℓ and language k, we compute the activation probability

$$P_{j,k}^{(\ell)} = \mathbb{E}\left[\mathbb{I}(a_j^{(\ell)} > 0) \;\middle|\; \text{language } k\right],$$

where $a_j^{(\ell)}$ denotes the neuron activation and $\mathbb{I}(\cdot)$ is the indicator function. The vector of activation probabilities across languages is ℓ_1-normalized to form a distribution, and its entropy is computed as

$$\text{LAPE}_j^{(\ell)} = -\sum_k P'^{(\ell)}_{j,k} \log P'^{(\ell)}_{j,k}.$$

Low entropy indicates that a neuron activates predominantly for a small subset of languages. Neurons with sufficiently low entropy and a dominant language are identified as *language-associated*.

C.2 SAE-LAPE for Sparse Features

SAE-LAPE extends the LAPE criterion to sparse latent features obtained from Sparse Autoencoders (SAEs). SAEs are trained on feed-forward (MLP) activations to decompose dense representations into a sparse set of latent features. Each SAE feature is treated analogously to a neuron: we compute its activation probability per language based on whether the feature is active for a given token. The same entropy-based criterion is then applied to identify language-associated sparse features.

To ensure robustness, we restrict attention to features that are active for a non-trivial fraction of tokens and examples within at least one language. This enables language association analysis at the level of sparse, interpretable features rather than individual neurons.

C.3 Hyperparameters and Implementation Details

All LAPE and SAE-LAPE analyses share a common entropy-based framework for measuring language selectivity, differing primarily in their filtering criteria and membership assignment rules.

Activation Statistics. For both methods, activation probabilities are computed over a multilingual corpus by aggregating token-level activations within each language. A unit (raw neuron or SAE latent) is considered *active* for a token if its activation exceeds zero. Activation probabilities are normalized across languages prior to entropy computation.

SAE-LAPE Hyperparameters. SAE-LAPE operates on sparse latent features extracted from Sparse Autoencoders trained on MLP activations. To exclude noisy or overly idiosyncratic features, we apply two pre-selection thresholds: (i) an *example rate* of 0.98, requiring a latent to be active in at least 98% of examples within at least one language, and (ii) a *high-frequency latent (HFL) rate* of 0.1, requiring activation on at least 10% of tokens in that language. Latents failing either criterion have their entropy set to infinity and are excluded from selection.

Language membership for SAE latents is determined using a relative top-k criterion. A latent f is considered present in language l if its activation probability satisfies

$$P(f \mid l) \geq 0.8 \times \max_{l' \in L} P(f \mid l'),$$

where the threshold ratio of 0.8 is fixed across all experiments. This relative criterion allows features to be shared across a small number of languages when desired. Depending on the configuration, we further restrict selection to latents that are either unique to a single language (`lang_specific`) or shared by an exact number of languages (`lang_shared`).

A methodological adaptation was required for Gemma models. Because the original SAE-LAPE implementation was designed for cardinally constrained *Top-K* SAEs (as used for Llama), we introduced an additional filtering step for Gemma's *JumpReLU* SAEs by restricting the analysis to the top-200 active latents by activation magnitude per token. While this introduces minor variance, the macro-level representational trends remain highly consistent across both SAE architectures.

LAPE Hyperparameters for Raw Neurons. For raw model neurons, which are typically denser and more polysemantic, we adopt a more conservative, percentile-based filtering strategy. We compute the 95th percentile of activation probabilities

Parameter	Value
Activation indicator	Latent $z > 0$
Aggregation level	Token + example
Minimum example rate	0.98
Minimum HFL rate	0.10
Top-k threshold ratio	0.80
Entropy for invalid features	∞
Llama Top-K SAE: k-value	32
Gemma JumpReLU SAE: enforced Top-K	200

Table 3: Hyperparameters used for SAE-LAPE identification of language-associated sparse latent features. Rates are computed per layer over the multilingual corpus.

Parameter	Value
Activation indicator	$a > 0$
Aggregation level	Token-level
Activation percentile (filter rate)	95^{th} percentile
Entropy selection fraction	Lowest 1% neurons
Language assignment threshold	95^{th} percentile (global)
Inactive neuron handling	Discarded

Table 4: Hyperparameters used for LAPE-based identification of language-associated raw neurons. Activation percentiles are computed globally across all neurons and languages.

across all neurons and languages, and discard neurons whose activation probability never exceeds this threshold in any language. Among the remaining candidates, we select the lowest-entropy neurons corresponding to the top 1% most language-selective units.

Language assignment for these neurons uses an absolute activation criterion: a neuron is attributed to language l if its activation probability exceeds the same 95th-percentile threshold. This approach emphasizes globally salient, language-skewed neurons rather than fine-grained feature sharing. All models share the same setup.

Outputs. Both methods export identified units with identified languages(s), activation probabilities, and entropy values.

Overall, SAE-LAPE prioritizes consistent, interpretable sparsified features with controlled cross-lingual sharing, while LAPE for raw neurons focuses on identifying the most strongly language-skewed units in dense representations. Table 3 and Table 4 summarize the thresholds and hyperparameters used for SAE-LAPE and LAPE respectively.

C.4 Usage in This Work

In this paper, LAPE and SAE-LAPE are used strictly as *identification tools* for selecting language-associated neurons and sparse features. All subsequent analyses – including romanization, shuffling, probing, and causal interventions – are conducted on these identified units. We do not assume that low entropy alone implies abstract linguistic control or causal importance.

D Script Perturbation Experiments (Romanization)

D.1 Experimental Setup

Datasets. We use the dev split of FLORES+, which provides sentence-aligned multilingual data across typologically diverse languages. For South Asian languages, we additionally consider the Dakshina dataset to assess the effects of context-aware romanization, noting that these corpora are not sentence-aligned and are therefore used only for supplementary analysis.

Language Selection. Our experiments cover Hindi (hi), Marathi (mr), Bengali (bn), Urdu (ur), Russian (ru), Bulgarian (bg), Japanese (ja), Chinese (zh), Korean (ko), English (en), and Spanish (es), spanning multiple writing systems and including both closely related and typologically distant language pairs.

Romanization Procedure and Diacritics. Romanized text is generated using the ICU Transliterator. For applicable languages, we construct both diacritic-preserving and ASCII-only variants by removing diacritics via Unicode normalization, enabling controlled analysis of sub-phonemic orthographic cues. The whole pipeline is run thrice: (a) with the native datasets, (b) with the diacritics-romanization datasets, (c) with the diacritics-free-romanization datasets. Then the resulting neuron sets are compared.

Metrics. Overlap between language-associated feature sets is quantified using Jaccard similarity. We additionally compute cross-language overlaps to assess whether romanization induces increased sharing with English or other Latin-script languages.

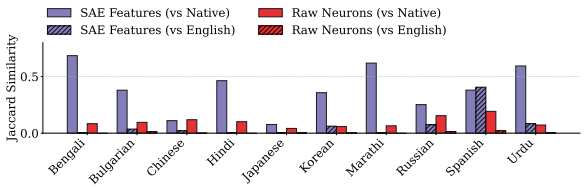


Figure 8: Jaccard similarity between language-associated units identified from Romanized inputs and those from Native-script or English inputs in Gemma-2-2B. Results are shown for both **raw neurons** and **SAE features**. Romanized inputs exhibit low overlap with their native-script counterparts and near-zero overlap with English in both representations, indicating limited cross-script alignment without convergence to English.

D.2 Supplementary Romanization Analysis Across Models and Representations

This appendix extends Section 4 by documenting the full set of romanization diagnostics across all evaluated model and representation configurations. While the main text focuses on feature identity and overlap, here we examine (i) aggregate neuron-sharing structure across all languages, and (ii) distributional effects of romanization on activation behavior of language-specific neurons.

Aggregate Neuron Sharing Under Orthographic Variation. We begin by reporting aggregate Venn diagrams computed jointly over all languages, restricted to language-specific neurons identified independently per input condition. For each configuration, we plot Venn diagrams for units shared among at most three languages, comparing native-script inputs, romanized inputs with diacritics, and romanized inputs without diacritics. This representation captures all low-order sharing behavior, ensuring a fair balance between specificity and coverage.

Figures 9 and 10 summarize these results across Gemma and Llama, under both raw MLP and SAE representations. Across all available configurations, aggregate overlap between native and romanized variants remains low. Overlap between the two romanized variants is slightly higher but remains limited, indicating that even minor orthographic perturbations such as diacritic removal induce substantial reassignment of language-specific neurons. Figure 8 shows these trends for Gemma, consistent with the trends for Llama from the main text. These aggregate results confirm that the effects reported per-language in the main text persist at the multilingual level.

Distributional Effects of Romanization. Overlap-based analyses describe neuron *reuse*, but do not capture how retained neurons behave. We therefore analyze activation statistics under native versus romanized inputs for both (i) the complete sets of neurons active in each condition, and (ii) the subset of neurons overlapping between native and romanized representations.

Figures 11 and 12 report these distributions for Gemma and Llama across raw and SAE representations. Across all available configurations, romanization induces clear distributional shifts in both activation probability and entropy. These shifts are observed both when considering complete neuron sets and when restricting to overlapping neurons, indicating that the effects are not solely driven by changes in neuron identity. Moreover, the shifts are substantially larger than those observed under shuffling baselines, suggesting structured changes in activation dynamics rather than random variance.

Representation-Specific Distributional Trends. For raw MLP representations, romanization consistently shifts activation probability mass toward higher values while reducing entropy, indicating more concentrated and decisive neuron firing. This effect is pronounced for Gemma, whereas for Llama the entropy reduction is comparatively mild, despite similar probability shifts.

For SAE representations, distributional shifts are again substantial, but the directionality is less consistent across configurations. In particular, both entropy and activation probability may increase or decrease depending on the setup. However, the overall magnitude of these shifts is larger for Gemma than for Llama, suggesting that sparse representations in Gemma are more sensitive to orthographic perturbations.

Stability of Mean Activation Statistics. Finally, we report mean activation statistics averaged across languages and neurons. Despite strong neuron-level redistribution and distributional shifts, mean activation values remain largely stable across native and romanized inputs, indicating that romanization reallocates activation mass without substantially altering global magnitude. Figure 13 summarizes these values for the raw activations.

Summary. Together with Section 4, these results show that orthographic variation affects both the allocation and dynamics of language-specific neurons. Degree-3 analyses confirm that low-order

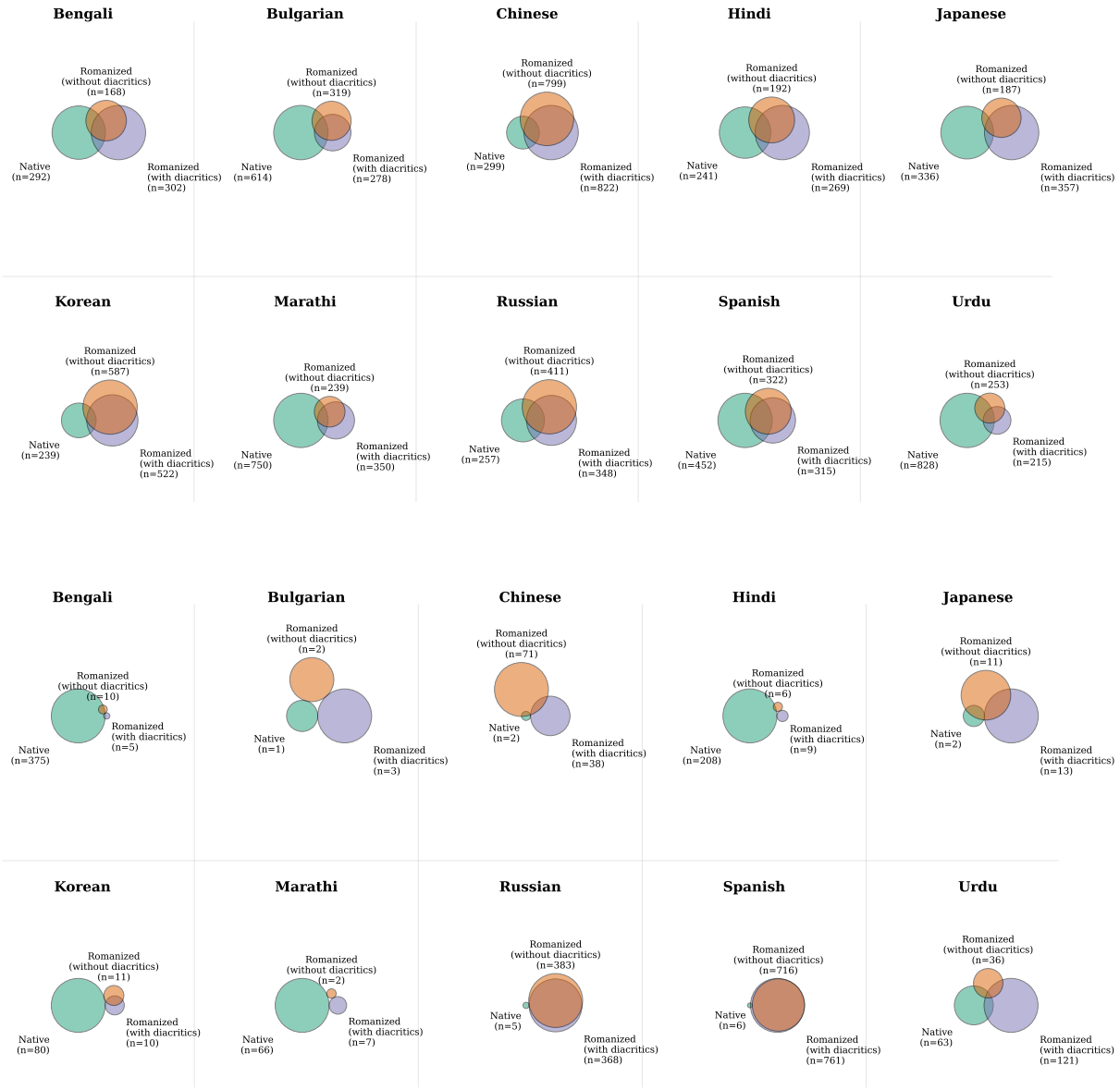


Figure 9: Aggregate degree-3 Venn diagrams of language-specific neurons under orthographic variation for **Gemma-2-2B**. Degree-3 denotes the union of neurons shared by up to three languages. Panels correspond to raw MLP and SAE representations under diacritics-preserving and diacritics-removed romanization.

sharing remains limited even when allowing pairwise reuse, while distributional statistics reveal structured activation shifts under romanization that are not captured by identity-based overlap alone.

D.3 Probing–Romanization Interaction: Typological Alignment of Neuron Subsets

This subsection analyzes how typological structure, as measured by `lang2vec` probing, distributes across neuron subsets induced by romanization. While earlier sections establish that romanization reorganizes language-specific features, here we ask whether this reorganization correlates with the degree to which neurons encode linguistic typology.

Setup. For each layer, model, and representation (raw MLP or SAE), neurons are partitioned into four disjoint subsets based on their activity under native and romanized inputs: (i) *native-only* neurons, (ii) *romanized-only* neurons, (iii) *overlap* neurons active under both conditions, and (iv) a *baseline* consisting of all neurons in the layer. For each subset, we compute the average family-wise maximum probing R^2 score across neurons for the three typological feature families used in the final analysis: `fam`, `syntax`, and `phonology`. All plots in this section report these averages using the `specific_mean` metric.

Figure 10: Aggregate degree-3 Venn diagrams of language-specific neurons under orthographic variation for **Llama-3.2-1B**. Degree-3 denotes the union of neurons shared by up to three languages. Panels correspond to raw MLP and SAE representations under diacritics-preserving and diacritics-removed romanization.

Consistency Across Models and Representations. Across all model and representation configurations, the qualitative behavior of these curves is remarkably consistent. Baseline probing values are generally lower than those obtained from more selective neuron subsets. An exception arises for Gemma, where neurons active only for native inputs sometimes fall below the baseline. In the Gemma raw setting, probing values are comparatively similar across subsets, indicating weaker separation between neuron groups.

Overlap Neurons Encode Stronger Typological Structure. The most robust result is that the *overlap* subset consistently exhibits substantially higher probing R^2 scores than all other subsets. This pattern holds across all models, representations, feature families, and romanization conditions. Neurons that remain active across both native and romanized inputs are therefore not only orthography-invariant, but also more strongly aligned with linguistic typology than neurons that respond selectively to a single script variant.

Model- and Representation-Level Effects. Consistent with prior probing analyses, Gemma achieves higher absolute probing scores than Llama across all neuron subsets. Within Llama, SAE representations exhibit markedly lower R^2 values than raw MLP activations, often by a large margin. Cru-

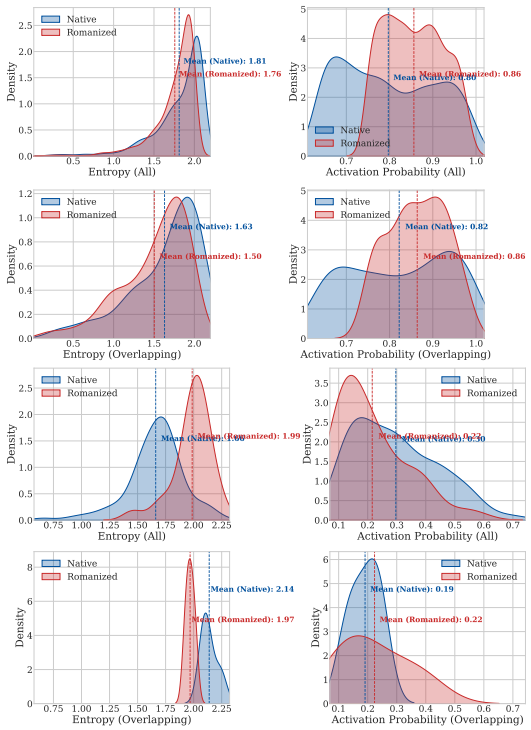


Figure 11: Activation probability and entropy distributions for language-specific neurons under native vs. romanized inputs (Gemma-2-2B). Top: raw MLP; Bottom: SAE.

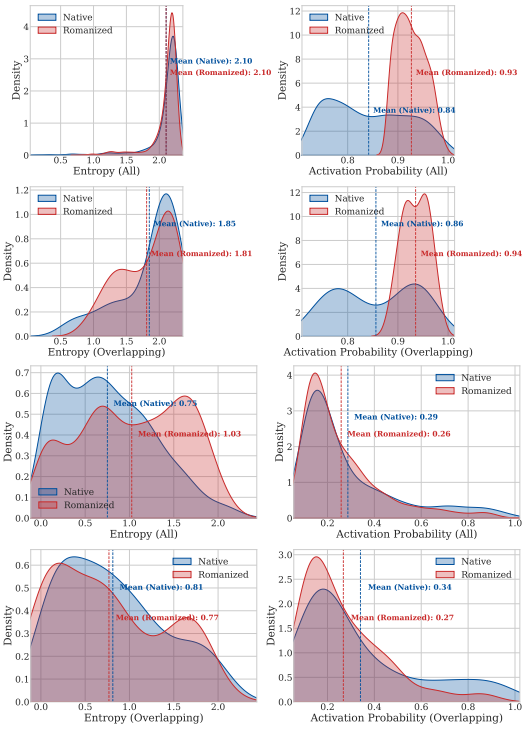


Figure 12: Activation probability and entropy distributions for language-specific neurons under native vs. romanized inputs (Llama-3.2-1B). Top: raw MLP; Bottom: SAE.

cially, however, the dominance of the overlap subset persists even in these lower-signal regimes, indicating that the relationship between orthographic stability and typological alignment is robust to overall representational strength.

Preservation of Typological Hierarchy. Across all neuron subsets and configurations, the relative ordering of feature families remains unchanged:

$$\texttt{fam} > \texttt{syntax} > \texttt{phonology}.$$

Romanization-induced partitioning thus modulates the *magnitude* of typological alignment, but not its hierarchical structure.

Representative Results. Figures 5–17 show representative results for Llama and Gemma under both raw and SAE representations with diacritics-preserving romanization. Analogous trends are observed for the diacritics-removed setting.

Summary. Together, these results establish a systematic association between orthographic robustness and linguistic abstraction. Neurons that are preserved across romanization transformations consistently encode stronger typological structure than neurons that are sensitive to script variation. Romanization thus serves as a diagnostic tool that reveals not only representational fragmentation, but also the locus of stable linguistic abstraction within multilingual models.

E Structural Perturbation Experiments (Word Shuffling)

Datasets. Following the original setup, we use a combination of three datasets.

(i) **XNLI:** 1,000 examples from the `train` split (en, de, fr, hi, es, th, bg, ru, tr, vi).
(ii) **PAWS-X:** 1,000 examples from the `train` split (en, de, fr, es, ja, ko, zh).
(iii) **FLORES+:** 997 examples from the `dev` split (15+ languages).

Procedure. For each dataset, we apply the LAPE and SAE-LAPE pipelines twice: (a) on sentences in their natural word order, (b) on sentences where words within each prompt are randomly permuted. All other parameters are held fixed. The resulting neuron sets are compared.

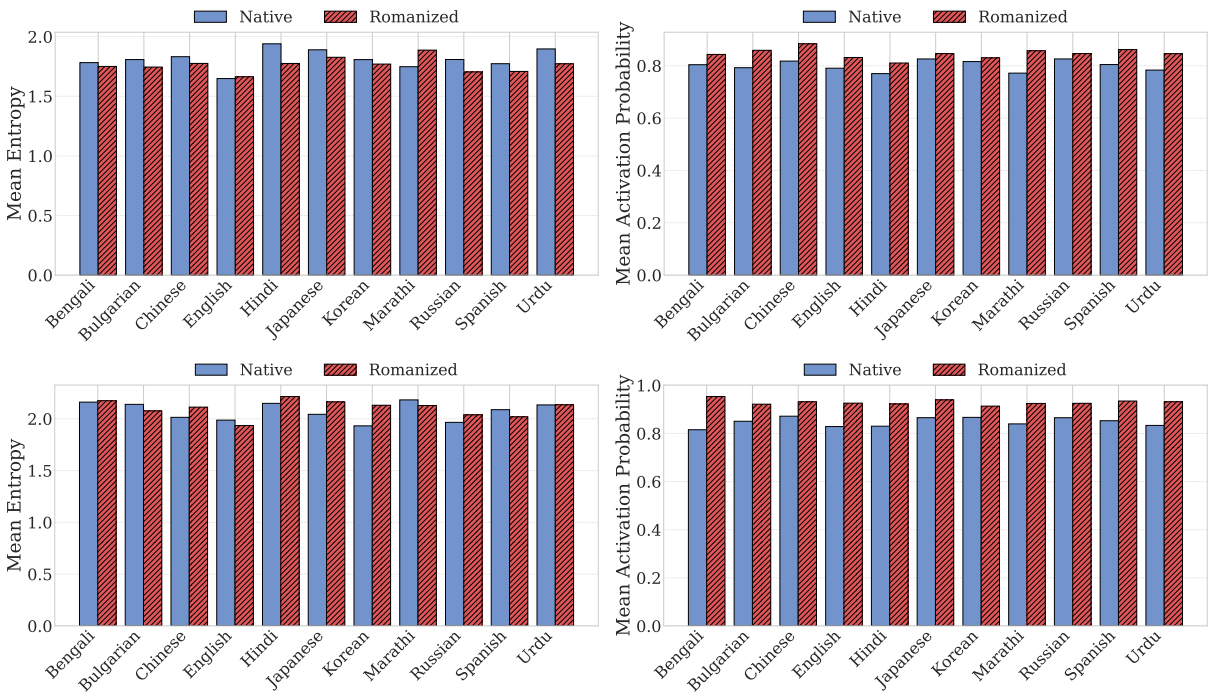


Figure 13: Mean activation statistics across languages for native and romanized inputs, for the raw MLP LAPE-identified features. Top: Gemma-2-2B; Bottom: Llama-3.2-1B.

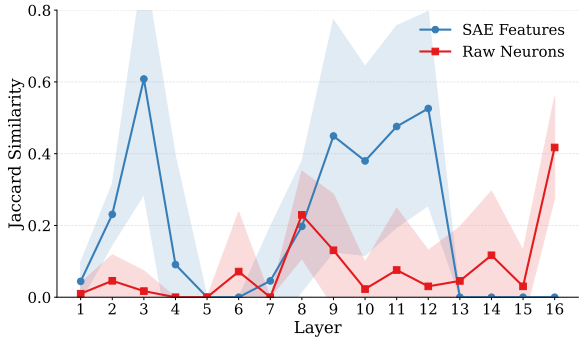


Figure 14: Layer-wise alignment between language-associated units for Native and Romanized inputs in Gemma-2-2B. The **red line** denotes average Jaccard similarity for **raw neurons**, and the **blue line** for **SAE features**; shaded regions indicate standard deviation across languages. Both raw neurons and SAE features show a mid-layer increase in overlap. However, in all cases, alignment remains far from convergence, indicating that representational separation persists beyond input tokenization.

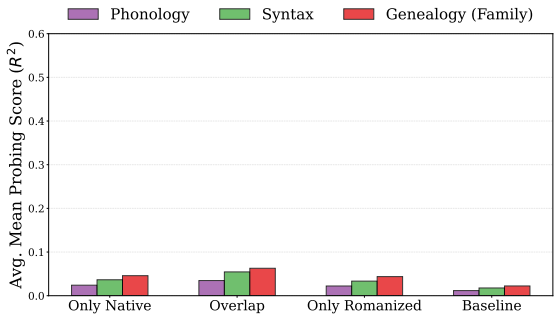


Figure 15: Average family-wise maximum probing R^2 scores across neuron subsets induced by romanization (Llama-3.2-1B, SAE). Overall probing scores are lower, but overlap neurons remain dominant.

E.1 Supplementary Shuffling Analyses Across Models and Representations

This appendix provides additional analyses for the shuffling experiments reported in Section 5. While the main text focuses on language-level stability and aggregate trends, here we document neuron-level overlap structure and distributional behavior across all model and representation configurations.

Aggregate Neuron Overlap Under Shuffling. We first examine neuron overlap between features identified from original and word-shuffled inputs, aggregated across all languages. Figures 18 and 19 show degree-based Venn diagrams for Llama and Gemma, respectively, under raw and SAE representations.

Across all configurations, overlap between original and shuffled feature sets remains high, indicating that shuffling preserves feature identity at the neuron level. This confirms that the stability

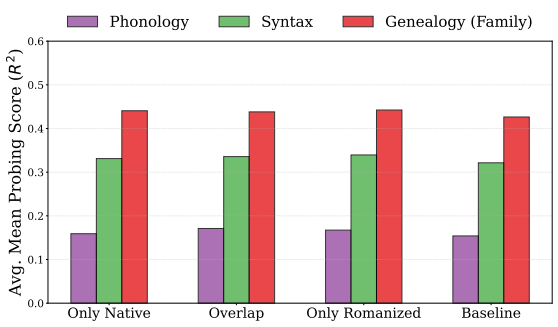


Figure 16: Average family-wise maximum probing R^2 scores across neuron subsets induced by romanization (Gemma-2-2B, raw MLP). Scores are closer across subsets, with native-only neurons occasionally falling below baseline.

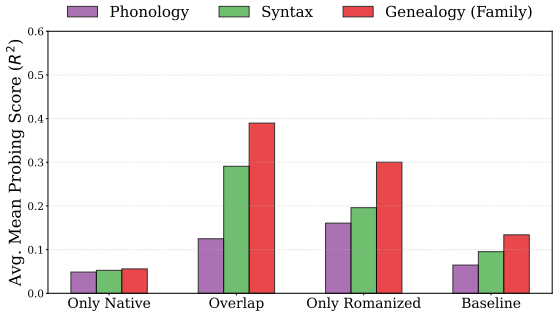


Figure 17: Average family-wise maximum probing R^2 scores across neuron subsets induced by romanization (Gemma-2-2B, SAE). Overlap neurons continue to show stronger typological alignment despite increased sparsity.

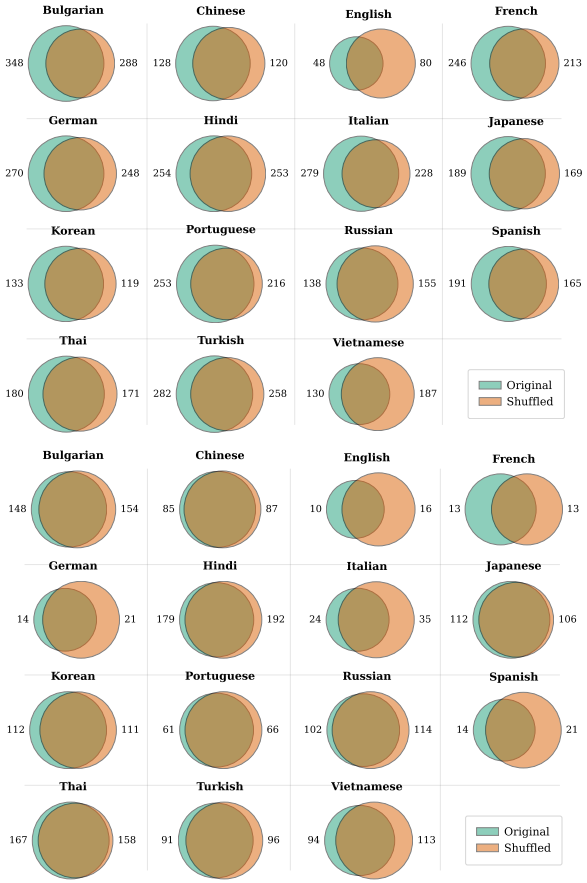


Figure 18: Aggregate degree-based Venn diagrams comparing features from original and shuffled inputs in Llama-3.2-1B. Top: raw MLP; Bottom: SAE. High overlap indicates stability of neuron identity under word-order perturbation.

observed at the language level in the main text also holds when aggregating across neurons.

For Gemma SAE, the absolute number of identified neurons is small for certain languages, making low-degree overlap estimates unstable. In this case, we report overlap up to degree 5 rather than degree 3. When restricting attention to settings with sufficient numbers of identified neurons, high overlap is consistently recovered, in line with other configurations.

Distributional Stability of Activation Statistics. Beyond feature identity, we analyze whether shuffling induces shifts in activation behavior. Figures 20 and 21 compare distributions of activation entropy and selection probability for original versus shuffled inputs, aggregated across languages. Across all models and representations, the distributions are nearly overlapping, with only minor shifts in their means. This remains true when restricting the analysis to overlapping features (results omitted for brevity), indicating that neurons preserved under shuffling also maintain stable activation profiles.

Mean Activation Statistics. Finally, we report mean activation statistics aggregated across languages. As shown in Figure 22, mean entropy and selection probability change only marginally under shuffling, reiterating that syntactic perturbation does not significantly reweight feature activity.

Summary. Together, these supplementary analyses reinforce the robustness conclusions in Section 5. Word-order shuffling preserves both neuron identity and activation statistics across models and representations. Differences observed in low-neuron regimes (e.g., Gemma SAE) are attributable to feature sparsity rather than systematic sensitivity to syntactic structure, further supporting the view that language-associated features primarily reflect token-level and distributional regularities.

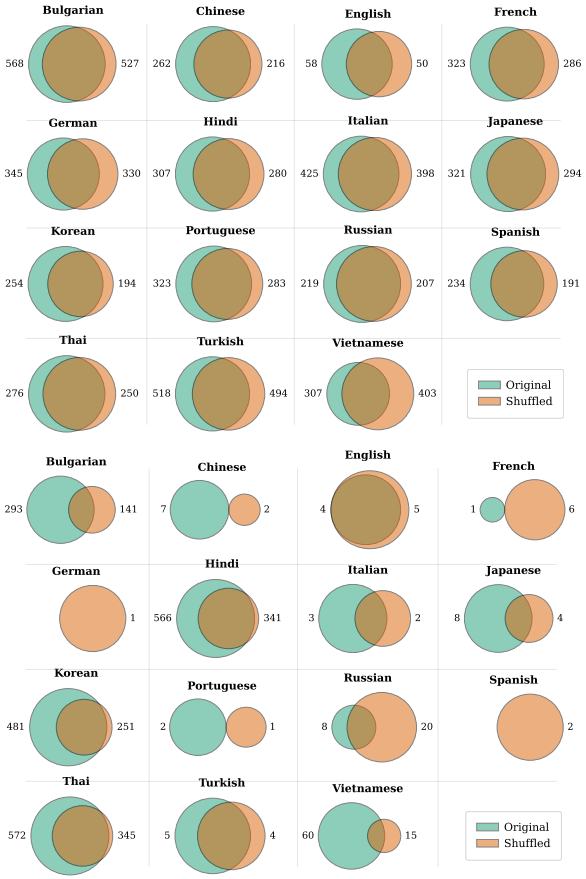


Figure 19: Aggregate degree-based Venn diagrams comparing features from original and shuffled inputs in Gemma-2-2B. Top: raw MLP (degree 3); Bottom: SAE (degree 5). When sufficient neurons are identified, high overlap is preserved under shuffling.

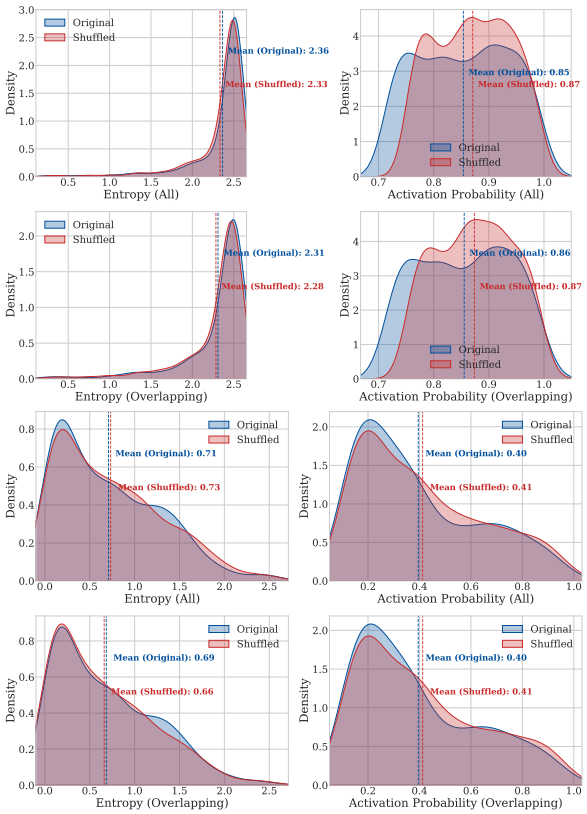


Figure 20: Activation entropy and selection probability distributions for original and shuffled inputs in Llama-3.2-1B. Top: raw MLP; Bottom: SAE. The near-identical distributions indicate minimal distributional shift under shuffling.

E.2 Probing–Shuffling Interaction: Typological Alignment Under Syntactic Perturbation

This subsection analyzes how sensitivity to word-order shuffling correlates with typological structure, as measured by `lang2vec` probing. In contrast to romanization, shuffling preserves surface form and token identity while disrupting local syntactic order. We therefore examine how typological alignment distributes across neuron subsets that differ in their stability under shuffling.

Setup. For each layer, model, and representation (raw MLP or SAE), neurons are partitioned into four disjoint subsets based on their activity under original and shuffled inputs: (i) *normal-only* neurons (active only for original text), (ii) *shuffled-only* neurons, (iii) *overlap* neurons active under both conditions, and (iv) a *baseline* consisting of all neurons in the layer. For each subset, we compute the average family-wise maximum probing R^2 score across neurons for the three typological feature families used throughout the paper: `fam`, `syntax`, and `phonology`. All plots report mean values aggregated across layers; we use `degree3_mean` for all configurations, except for Gemma SAE where `degree5_mean` is used due to low neuron counts in some languages.

Raw Representations Show Uniform Typological Alignment. Figures 6 and 25 show results for raw MLP representations in Llama and Gemma. In both models, probing scores are remarkably similar across the *normal-only*, *shuffled-only*, and *overlap* subsets. This indicates that, at the level of distributed raw activations, sensitivity to word-order perturbation is largely decoupled from typological alignment. Neurons that respond selectively to shuffled inputs are no less typologically informative than those that respond to original inputs.

SAE Representations Expose a Structured Hierarchy. A different pattern emerges for SAE rep-

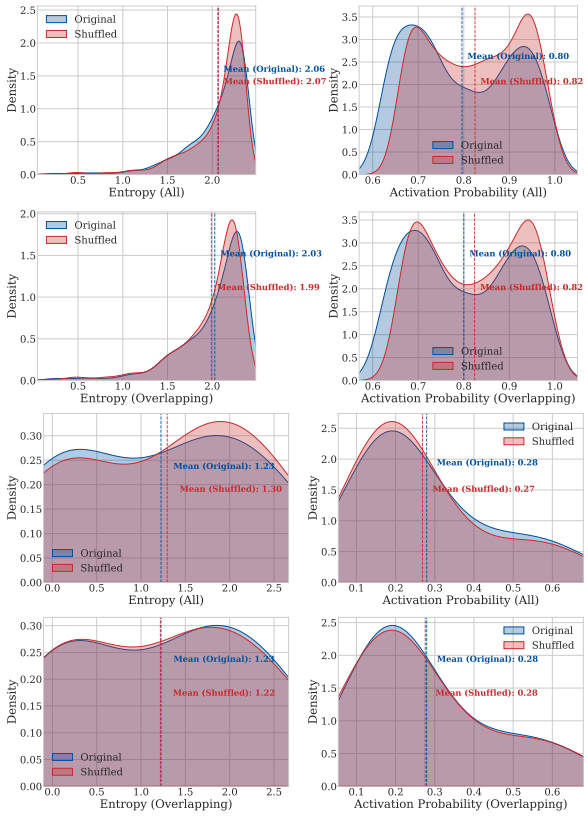


Figure 21: Activation entropy and selection probability distributions for original and shuffled inputs in Gemma-2-2B. Top: raw MLP; Bottom: SAE. Distributional shifts remain small across representations.

resentations (Figures 24 and 26). For both Llama and Gemma, we observe a consistent ordering:

$$\textit{normal-only} \approx \textit{shuffled-only} > \textit{overlap}.$$

That is, neurons selective to a single condition – whether original or shuffled – exhibit stronger typological alignment than neurons that remain active across both. This contrasts sharply with the romanization setting, where overlap neurons were most informative, and suggests that invariance to word-order perturbation does not preferentially select for typologically informative features in sparse representations.

Baseline Effects in Llama. In Llama, baseline probing scores are substantially lower than those of any condition-specific subset, for both raw and SAE representations. This gap is less pronounced in Gemma. The result suggests that in Llama, typological information is concentrated in a relatively small subset of neurons, and is diluted when averaging across the full layer.

Preservation of Typological Hierarchy. Across all models, representations, and neuron subsets, the relative ordering of feature families remains unchanged:

$$\texttt{fam} > \texttt{syntax} > \texttt{phonology}.$$

Thus, while shuffling-sensitive partitioning modulates the strength of typological alignment, it does not alter the underlying hierarchy of linguistic information.

Representative Results. Figure 6, along with Figures 24–26, show the full set of results for all configurations.

Summary. Together, these results indicate that robustness to syntactic perturbation is not a reliable indicator of typological abstraction. In raw representations, typological information is broadly distributed and largely insensitive to shuffling-based partitioning. In contrast, sparse representations reveal that neurons invariant to shuffling are not necessarily those most aligned with linguistic typology, highlighting a clear qualitative difference between orthographic and syntactic perturbations.

F Probing Typological Structure Across Layers

F.1 Experimental Setup

This section describes the probing framework used to relate neuron- and SAE-feature activations to typological properties of languages.

Activation Extraction. For each language and layer, we extract mean activations corresponding to either raw model hidden states or SAE latents, depending on the probing condition.

Given a model layer ℓ and a selected set of neurons or SAE features $\mathcal{N}_\ell$, we collect activations over a multilingual dataset as follows. For each minibatch, we extract the hidden states at layer ℓ (or the corresponding SAE latent activations) and average over both batch and token dimensions. These per-batch means are then aggregated across batches to obtain a single activation vector per language and layer: $\mathbf{x}_\ell^{(k)} \in \mathbb{R}^{|\mathcal{N}_\ell|}$, where k indexes languages. Activations are collected from the FLORES+ dataset using the `train` split, with batch size 16.

Typological Features. Typological targets are loaded from `lang2vec` features. Each feature set

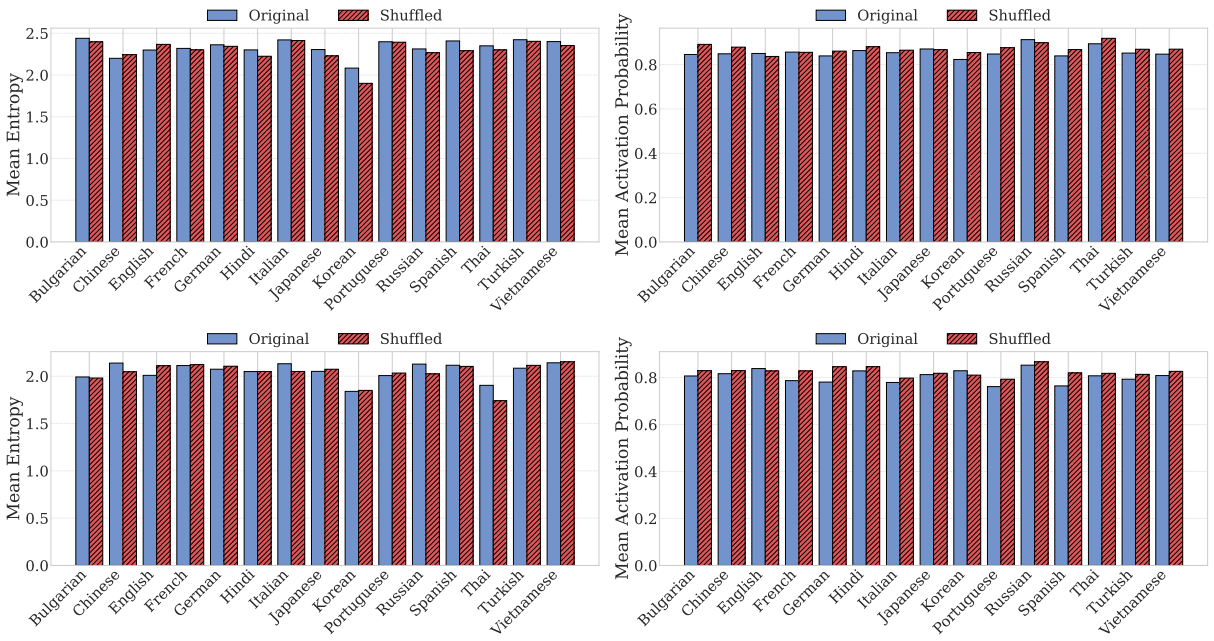


Figure 22: Mean activation entropy and selection probability across languages before and after shuffling. Top: Llama-3.2-1B; Bottom: Gemma-2-2B (raw MLP). Mean-level changes are small, consistent with distribution-level stability.

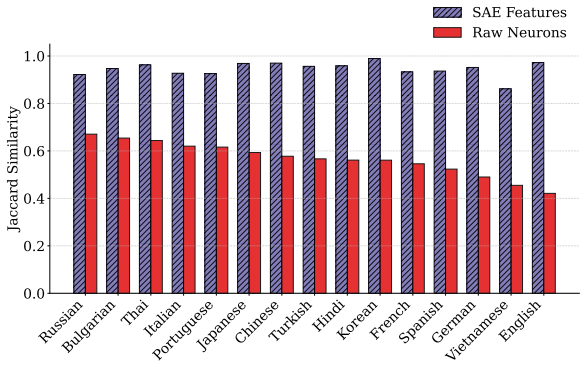


Figure 23: Jaccard similarity between language-associated units identified from original and word-shuffled text in Gemma-2-2B. **Raw neurons** exhibit consistently moderate-to-high overlap across languages, indicating robustness to word-order perturbation. **SAE features** also show high overlap, revealing robustness of sparse features to local distributional patterns disrupted by shuffling.

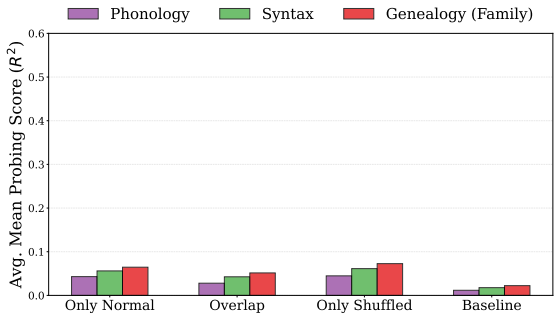


Figure 24: Average family-wise maximum probing R^2 scores across neuron subsets under shuffling (Llama-3.2-1B, SAE). Condition-specific subsets dominate overlap neurons; baseline scores remain lowest.

corresponds to a matrix $\mathbf{Y} \in \mathbb{R}^{L \times F}$, where L is the number of languages and F the number of typological dimensions.

Feature sets include syntactic, phonological, and inventory-based features, as well as genealogical family and geographic coordinates. Prior to probing, feature dimensions with zero variance across the selected languages are removed to ensure well-defined regression targets.

Regression Setup. Probing is formulated as a set of *univariate* regression problems. For each neuron or feature $n \in \mathcal{N}_\ell$ and each typological dimension f, we fit a linear model across languages:

$$y_f^{(k)} = \beta_{n,f}\, x_n^{(k)} + \epsilon^{(k)},$$

where $x_n^{(k)}$ denotes the mean activation of neuron n for language k.

To stabilize estimation under small sample sizes, we use ridge regression with regularization coefficient $\lambda = 1.0$. Importantly, each neuron is probed independently, i.e., regressions are single-predictor models rather than multivariate probes.

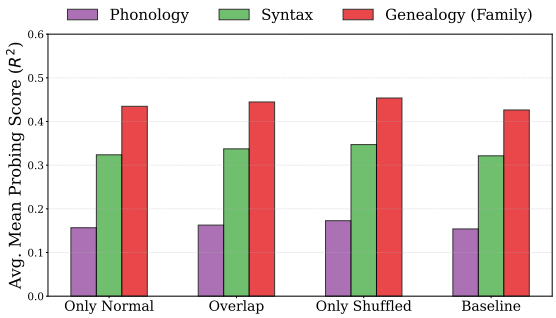


Figure 25: Average family-wise maximum probing R^2 scores across neuron subsets under shuffling (Gemma-2-2B, raw MLP). Typological alignment is similar across normal-only, shuffled-only, and overlap subsets.

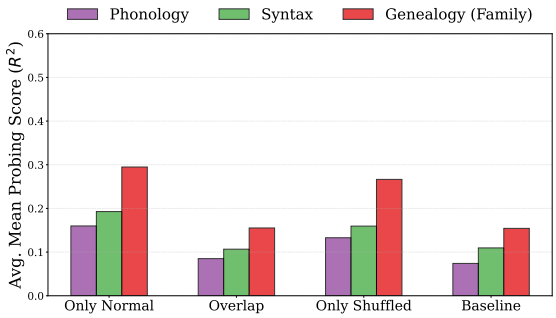


Figure 26: Average family-wise maximum probing R^2 scores across neuron subsets under shuffling (Gemma-2-2B, SAE). Results use degree-5 aggregation due to low neuron counts. As in Llama, condition-specific subsets show stronger typological alignment than overlap neurons.

Cross-Validation and Evaluation. Probe quality is assessed using 5-fold cross-validation over languages. In each fold, regression coefficients are estimated on the training languages and evaluated on held-out languages. The coefficient of determination (R^2) is computed for each neuron–feature pair on the test split. For numerical stability and efficiency, regression is implemented in closed form and evaluated in blocks over both neuron and feature dimensions. For each neuron n and feature f, the final probe score is obtained by averaging R^2 across folds:

$$R^2_{n,f} = \frac{1}{K} \sum_{k=1}^{K} R^2_{n,f}{}^{(k)}.$$

Neuron–feature pairs with undefined R^2 values (e.g., due to zero variance in the target) are excluded.

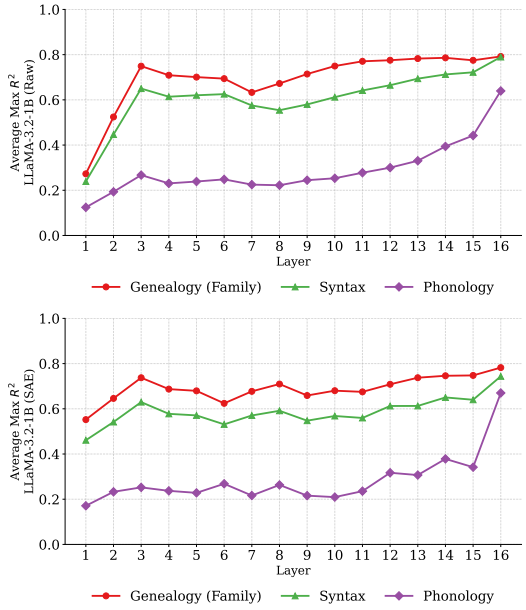


Figure 27: Layerwise probing performance in Llama-3.2-1B. **Top:** Raw MLP activations. **Bottom:** SAE features. SAE representations are comparatively stronger in early layers, while raw activations dominate in later layers.

F.2 Detailed Layerwise Probing Comparisons

Here we provide a detailed layerwise analysis of probing results for the three typological feature families used in the final experiments: `fam`, `syntax`, and `phonology`. We focus on (i) differences between raw MLP activations and SAE representations, and (ii) cross-model differences between Llama-3.2-1B and Gemma-2-2B. All plots report layerwise averages of maximum R^2 scores per feature family.

Raw vs. SAE representations in Llama. Figure 27 shows the layerwise probing trends for Llama raw and SAE representations, while Figure 28 visualizes their differences directly. In early layers, SAE features are more informative than raw MLP activations for all three feature families, resulting in negative raw–SAE differences. This indicates that SAE training amplifies weak but structured typological signals that are only diffusely present in shallow raw activations. As depth increases, this advantage steadily diminishes, and the difference approaches positive values, indicating that raw representations become more linearly informative in deeper layers. This transition reflects a shift from early sparse amplification to richer distributed encoding in later layers.

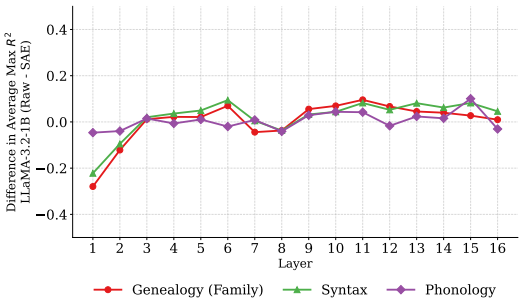


Figure 28: Raw minus SAE probing score differences for Llama-3.2-1B. Negative values in shallow layers indicate higher SAE informativeness, while the gradual shift toward positive values reflects increasing raw dominance with depth.

Raw vs. SAE representations in Gemma. The corresponding Gemma plots are shown in Figures 29 and 30. Unlike Llama, Gemma exhibits a more stable relationship between raw and SAE representations across layers. For `fam` and `syntax`, raw activations are consistently more informative than SAE features, yielding positive differences across depth. In contrast, `phonology` shows consistently negative differences, indicating that Gemma SAEs preferentially preserve phonological structure relative to raw MLP activations. This feature-specific asymmetry suggests that sparse factorization interacts differently with lower-level sound-related abstractions than with genealogical or syntactic structure.

Cross-Model Comparison: Llama vs. Gemma. Figure 31 presents direct comparisons between Llama and Gemma under matched representational settings. Raw MLP activations exhibit stark cross-model differences in shallow layers for all three feature families, with phonology showing substantially larger gaps than `fam` or `syntax`. Moreover, raw cross-model differences decrease sharply with depth, producing a pronounced downward trend across all feature families. This suggests that early typological representations are strongly shaped by architectural and tokenizer-specific factors, while deeper layers converge toward more similar abstractions. In contrast, SAE representations substantially attenuate these differences. Although phonology remains the most discriminative feature family, the overall magnitude and depth-dependence of cross-model differences are reduced, indicating that sparse representations emphasize later-stage, shared abstractions over model-specific surface variation.

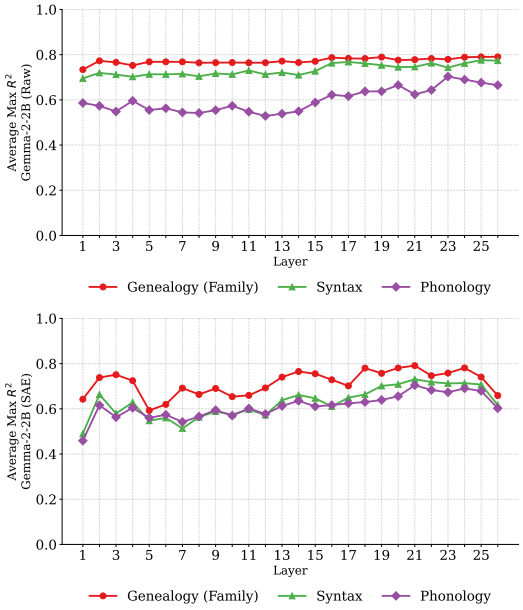


Figure 29: Layerwise probing performance in Gemma-2-2B. **Top:** Raw MLP activations. **Bottom:** SAE features. Raw representations dominate for `fam` and `syntax`, while SAE features retain stronger phonological signals across layers.

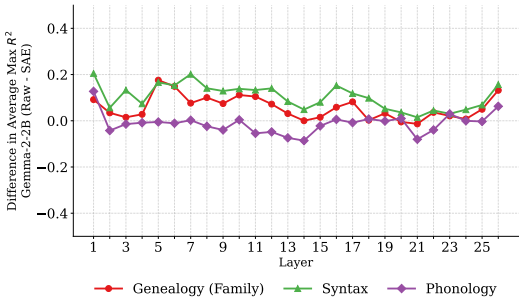


Figure 30: Raw minus SAE probing score differences for Gemma-2-2B. Differences are stable across depth: positive for `fam` and `syntax`, and negative for `phonology`.

Summary. These detailed comparisons show that sparse autoencoding reshapes typological structure in a depth-, model-, and feature-dependent manner. Llama SAEs transiently enhance early-layer typological accessibility, Gemma SAEs selectively favor Phonology features, and phonology consistently emerges as the most sensitive axis for cross-model differences – particularly in shallow raw representations.

G Causal Interventions on Invariant Neuron Sets

This appendix reports causal intervention experiments designed to assess whether neuron subsets

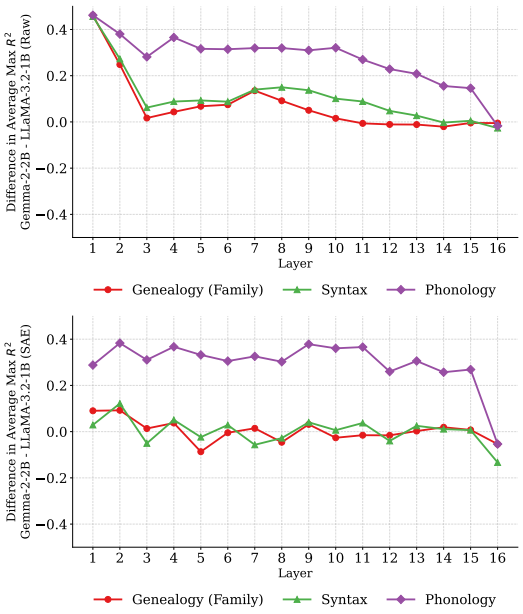


Figure 31: Cross-model comparison of probing performance. **Top:** Raw MLP activations. **Bottom:** SAE features. Raw representations show large early-layer differences, especially for phonology, followed by sharp convergence with depth, while SAE representations compress these disparities.

identified via invariance-based analyses are functionally necessary for multilingual language modeling. We intervene on neuron sets defined by their stability under controlled input perturbations: word-order shuffling and script romanization, done in sections 5 and 4 respectively. All experiments are conducted on the first 100 examples per language from the FLORES+ dataset.

G.1 Neuron Selection via Shuffling and Romanization

Neuron subsets are derived from earlier analyses that characterize neuron behavior under targeted surface perturbations.

Shuffling-Based Neuron Sets. Using word-level shuffling experiments, neurons are categorized into:
(i) **Overlap neurons:** Neurons consistently identified under both normal and shuffled inputs. These neurons are invariant to word-order perturbations and are hypothesized to encode structurally necessary representations.
(ii) **Only-unshuffled neurons:** Neurons identified only under normal inputs and absent under shuffled conditions. These neurons are sensitive to surface word order and local syntactic structure.

Romanization-Based Neuron Sets. Using native-script versus romanized inputs, neurons are grouped into:
(i) **Overlap neurons:** Neurons shared across native and romanized scripts, hypothesized to encode script-invariant representations.
(ii) **Only-native neurons:** Neurons active only for native-script inputs and whose functional signature disappears under romanization, indicating sensitivity to surface orthography.

Across both regimes, *overlap neurons* are defined by invariance to the corresponding perturbation, while non-overlap neurons capture sensitivity to surface form. For all experiments, matched **random control sets** are constructed by sampling an equal number of neurons uniformly from the overall neuron pool of the model.

G.2 Intervention Protocol

All experiments are conducted on raw model activations.

Ablation Scope. To avoid layer-local confounds, we apply *simultaneous ablation across all layers*. For each layer ℓ, activations of the selected neuron set are modified during the forward pass.

Ablation Types. We consider:
(i) **Zero ablation for shuffling-based neuron sets:** Activations are set to zero.
(ii) **Cross-language mean ablation for romanization-based neuron sets:** Activations are replaced by mean activation vectors computed from another language.

Mean vectors are computed over the corresponding FLORES+ split of the source language.

G.3 Evaluation Metrics and Statistical Testing

For each example, we compute clean and patched perplexities (PPL_{clean}, PPL_{patch}), perplexity ratios, and perplexity deltas (ΔPPL). Paired-sample t-tests compare targeted ablations against matched random controls over the 100 examples, with significance assessed at $p < 0.05$.

G.4 Causal Intervention Results: Llama-3.2-1B and Llama-3-8B

G.4.1 Shuffling-Based Zero Ablation

Table 5 reports results for shuffling-derived neuron sets. Across languages, ablation of *overlap neurons* induces the largest and most consistent degradations. For instance, Hindi in Llama-3.2-1B exhibits

Lang	Category	$PPL_{ratio}^{\text{target}}$	$PPL_{ratio}^{\text{ctrl}}$	p (ratio)	$\Delta PPL^{\text{target}}$	ΔPPL^{ctrl}	p (Δ)
Llama-3.2-1B							
en	overlap	1.116	0.954	1.5×10^{-50}	+272.3	−108.6	2.2×10^{-35}
en	only-unshuffled	0.963	1.044	3.0×10^{-48}	−87.1	+99.6	1.9×10^{-34}
hi	overlap	2.786	1.055	2.2×10^{-19}	+1914.0	+204.9	1.6×10^{-12}
hi	only-unshuffled	1.083	0.947	4.1×10^{-40}	+228.6	−133.3	4.6×10^{-10}
fr	overlap	1.118	1.030	5.4×10^{-6}	+114.6	+74.6	0.145
fr	only-unshuffled	0.935	0.957	2.8×10^{-10}	−130.7	−85.8	5.5×10^{-7}
zh	overlap	1.217	0.960	2.4×10^{-17}	+952.6	−523.2	2.4×10^{-24}
zh	only-unshuffled	0.936	0.982	5.7×10^{-17}	−695.3	−210.4	4.1×10^{-16}
Llama-3-8B							
en	overlap	0.925	0.992	3.7×10^{-8}	−23.3	−1.7	5.3×10^{-6}
en	only-unshuffled	0.832	0.993	1.3×10^{-59}	−40.1	−1.1	1.8×10^{-23}
hi	overlap	4.348	1.021	4.4×10^{-12}	+289.7	+10.5	2.2×10^{-10}
hi	only-unshuffled	0.859	1.017	4.0×10^{-30}	−26.8	+3.1	4.0×10^{-10}
fr	overlap	0.959	0.955	0.813	−29.1	−15.1	0.134
fr	only-unshuffled	0.867	1.020	4.6×10^{-19}	−59.1	+15.2	6.8×10^{-12}
zh	overlap	1.236	0.990	1.1×10^{-17}	+53.0	−5.0	4.6×10^{-17}
zh	only-unshuffled	0.934	0.982	2.2×10^{-14}	−20.9	−6.4	8.0×10^{-11}

Table 5: Causal zero-ablation results for shuffling-derived neuron sets across models (Llama models, raw neurons). Values report means over the first 100 FLORES+ examples per language. Control sets consist of matched random neurons with identical cardinality. Llama-3-8B shows qualitatively similar directional effects to Llama-3.2-1B, with stronger amplification in Hindi and Chinese, while maintaining robustness patterns under only-unshuffled conditions.

the strongest effect, with PPL ratios approaching 2.8 and ΔPPL exceeding $+1900$. Chinese shows similarly pronounced degradation, while English and French exhibit smaller but still significant effects. In all cases, overlap ablations degrade performance substantially more than matched random controls.

In contrast, *only-unshuffled neurons* yield weaker and sometimes inverted effects. For English and French, ablation leads to reductions in perplexity relative to clean runs. Hindi shows a small increase, but far weaker than overlap ablations, while Chinese exhibits negative ΔPPL despite statistical significance. These patterns indicate that only-unshuffled neurons encode order-sensitive surface regularities that are largely redundant for language modeling.

Moreover, the larger model (Llama-3-8B) produces stronger effects for Hindi and Chinese as compared to the (Llama-3.2-1B).

Qualitative Effects Under Shuffling. As shown in Figure 32, ablation of overlap neurons induces systematic qualitative failures in Hindi. Most notably, we observe *within-word script mixing*, where individual lexical items combine Devanagari and Latin characters (e.g., mixed-script morphemes). This phenomenon is not observed under random or Only-unshuffled ablations, where script switching – if present – occurs only at word boundaries. These qualitative failures align with the large perplexity degradation and indicate that overlap neurons play a role in maintaining subword-level orthographic coherence.

G.4.2 Romanization-Based Cross-Language Mean Ablation

Table 6 summarizes results for cross-language ablations between Hindi and English for both Llama models. Replacing overlap neuron activations across languages yields relatively mild effects. English shows a slight decrease in perplexity, while Hindi shows a modest increase. The small magnitude of these effects suggests that overlap neurons encode representations that are largely invariant to script and language identity.

In contrast, replacing only-native neuron activations leads to extreme effects. Generations in both languages are severely affected. Qualitative inspection for the Hindi-to-English ablations in Llama-3.2-1B reveals that these reductions arise from language switching rather than improved Hindi modeling: many generations abandon Hindi entirely and continue fluently in English which has higher likelihood under the model. Sometimes the generations switch to other languages like Bengali, which

idx	Prefix (prompt context)	Clean continuation	Patched continuation (overlap-set ablation)	Inference
68	लॉकवुड गार्डन के किरायेदारों का मानना है कि अन्य 40 परिवार या उससे ज़्यादा बेदखली का सामना कर सकते हैं, चूँकि उन्हें पता चला है कि ओएचए पुलिस ऑकलैंड में अन्य सार्वजनिक आवास संपत्तियों की जाँच कर रही है जो आवास घोटाले में पकड़े जा सकत	े हैं। लॉकवुड गार्डन के किरायेदारों न	े हैं। lockwood Garden के किरायedarों का मानना	Within-word script mixing: Devanagari tokens corrupted by Latin (sometimes Arabic) characters inside a single word.
73	कुछ सप्ताह पहले, पत्रकार माकीस त्रिनताफेलोपोस द्वारा अपने लोकप्रिय टेलीविजन शो जूंगला में अल्फा टीवी में प्रकाशित जानकारी के बाद, संसद के सदस्य और वकील, पेट्रोस मंतौवलोस को पदमुक्त कर दिया गया था क्योंकि उनके कार्यालय के सदस्य अवैध भ्रष्टाचार और भ	भ्रष्टाचार के लिए जिम्मेदार थे। पेट्र	भ्रष्टाचार के साथ संबंधित criminal charges were filed against them	
58	पुलिस ने कहा कि लो पिकालो की स्थिति मज़बूत थी क्योंकि वह पलेर्मो में प्रोवेनज़ानो का दाहिना हाथ था और उसके अनुभव ने उसे पुरानी पीढ़ी के मालिकों का सम्मान दिलवाया था क्योंकि वे अपने शक्ति तंत्र को मजबूत करते हुए अपनी छवि को न्यून बन	ाने के लिए प्रोवेनज़ानो के साथ मिल	ाने के लिए کرish करते थे. The police said that the condition	
10	11:29 बजे के आस-पास, विरोध करने वाले लोग व्हाइटहॉल की ओर बढ़ गए. इसके बाद, वे किनारे-किनारे चलते हुए ट्राफ़ल्गर स्क्वायर, एल्डविच, किंग्सवे से गुजरते हुए होलबर्न की ओर बढ़ गए, जहाँ कंजरवेटिव पार्टी	ेस्ट के लिए बैठे हुए थे। वे लोगों क	estival के प्रदर्शन के लिए SCHEDULED TO GO TO THE WHITE HOUSE	
9	11:20 पर, पुलिस ने प्रदर्शनकारियों को यह कहते हुए फ़ुटपाथ पर वापस जाने के लिए कहा कि उन्हें भीड़ इकट्ठा करने के साथ विरोध करने के अधिकार को संतुलित करने की ज़रूरत है.	#DelhiPolice #DelhiRiots #DelhiRiots2020 #	11:20 pm, Pulwama, Jammu and Kashmir, India: A	
0	सोमवार को, स्टैनफ़ोर्ड यूनिवर्सिटी स्कूल ऑफ़ मेडिसिन के वैज्ञानिकों ने एक नए डायग्नोस्टिक उपकरण के आविष्कार की घोषणा की जो कोशिकाओं को उनके प्रकार के आधार पर छाँट सकता है: एक छोटी प्रिंट करने योग्य चिप जिसे स्टैण्डर्ड इंकजेट प्रिंटर का उपयोग करक	े किया जा सकता है और एक सॉफ्टवेयर जो क	े किया जाएगा और conal DNA conal DNA conal DNA	
6	कातलान की राजधानी (बार्सिलोना) में जाने के बाद से, विडाल ने क्लब के लिए 49 गेम खेले थे.	क्लब के लिए 49 गेम खेले हैं क	The 49-year-old Spaniard has been a regular in the first team since 201	Complete or cross-word language collapse into English mid-generation.
5	28 वर्षीय विडाल तीन सीजन पहले सेविला से बारका में शामिल हुए थे।	28 वर्षीय विडाल तीन सीजन पहले सेविला	28 वर्षीय vidal tien, who was previously a member of the Seville	
1	शोधकर्ताओं ने कहा है कि यह अल्प आय वाले देशों में कैंसर, टीवी, एचआईवी और मलेरिया के रोगियों की आसानी से पहचान करेगा, जहाँ अमीर देशों की तुलना में स्तन कैंसर जैसी बीमारी में जीवित रहने की दर आधी हो सकती है.	शोधकर्ताओं ने कहा है कि यह अल्प आय	The study, published in the journal Cancer, found that the risk of cancer, HIV and	

Figure 32: Representative Hindi generations under shuffling-based overlap-neuron ablation (Llama-3.2-1B, raw). Each row shows the input prefix, clean continuation, and ablated continuation for the same example. While clean generations preserve Devanagari script integrity, overlap-neuron ablations induce within-word mixed-script corruption, abrupt language switching, and topic drift. Such token-internal script mixing is not observed under matched random or Only-unshuffled neuron ablations.

Lang	Category	$PPL_{ratio}^{\text{target}}$	$PPL_{ratio}^{\text{ctrl}}$	p (ratio)	$\Delta PPL^{\text{target}}$	ΔPPL^{ctrl}	p (Δ)
Llama-3.2-1B							
en	overlap	0.947	0.991	6.9×10^{-45}	-127.1	-21.1	9.7×10^{-28}
en	only-native	1.498	0.955	5.0×10^{-89}	+1176.9	-104.9	3.1×10^{-40}
hi	overlap	1.047	0.982	9.5×10^{-34}	+79.1	-53.8	1.7×10^{-7}
hi	only-native	0.312	0.970	1.2×10^{-38}	-1800.5	-92.5	7.7×10^{-11}
Llama-3-8B							
en	overlap	0.946	0.989	2.9×10^{-3}	-15.4	-2.0	1.1×10^{-3}
en	only-native	0.817	0.999	1.2×10^{-18}	-46.4	+0.2	2.1×10^{-10}
hi	overlap	1.062	0.990	2.1×10^{-12}	+7.8	-3.4	3.6×10^{-3}
hi	only-native	7.738	0.948	3.5×10^{-29}	+1326.6	-15.4	2.8×10^{-11}

Table 6: Cross-language mean ablation results for romanization-derived neuron sets (raw neurons). Rows indicate forward-pass language; mean activations are taken from the opposite language. Results are averaged over the first 100 FLORES+ examples. Control sets consist of matched random neurons with identical cardinality. Llama-3-8B exhibits qualitatively similar patterns to Llama-3.2-1B: only-native Hindi neurons cause dramatic perplexity changes when ablated during Hindi inference ($PPL_{ratio}^{\text{target}} = 7.74$), while overlap neurons produce modest effects across both models.

attributes to the drastic increase in perplexity for Llama-3-8B.

G.5 Causal Intervention Results: Gemma-2-2B

We repeat the same analyses on Gemma-2-2B to assess cross-model consistency.

G.5.1 Shuffling-Based Zero Ablation

Table 7 reports shuffling-based interventions for Gemma-2-2B. As in Llama, ablation of *overlap neurons* produces the strongest disruptions across languages. English, French, and Chinese show large increases in perplexity relative to random controls, while Hindi exhibits weaker but directionally consistent effects. Paired tests confirm that these differences are statistically significant in nearly all cases. Only-unshuffled neurons again yield weaker and more variable effects. In several languages, ablation produces smaller changes than random controls or even reduces perplexity, reinforcing the conclusion that these neurons encode word-order-level regularities rather than load-bearing structure, and that the robust shuffling-overlap neuron sets are more correlated with orthographic and subword-level structures.

Qualitative Effects Under Shuffling. Figure 33 presents representative Hindi and Chinese generations. Similar to Llama, overlap-neuron ablation induces *script changes within words*, including partial Latin insertions and mixed-script morphemes. Crucially, such intra-word script violations do not appear under random or Only-unshuffled ablations, indicating that overlap neurons support low-level orthographic coordination during decoding. More importantly, fluency is not lost while ablating the overlapping neurons, indicating that these neurons are not responsible for syntactic behavior.

G.5.2 Romanization-Based Cross-Language Mean Ablation

Table 8 summarizes romanization-based interventions. Only-native neurons exhibit the largest sensitivity: English-to-Hindi replacement causes large perplexity increases, while Hindi-to-English replacement often yields perplexity reductions. As in Llama, inspection of generations reveals frequent language switching in the latter case, explaining the apparent improvement. Overlap neurons again show smaller and more symmetric effects, consistent with a script-invariant functional role.

G.6 Summary

Across both models and perturbation regimes, a consistent causal pattern emerges:

• **Shuffling-Overlap neurons** – defined by invariance to shuffling – form a causally necessary backbone supporting stable, script-consistent generation. They are not causally related to fluency, rather script and subword-level regularity. Hence, the features tied strongly to script are more causally important for generation.

• **Romanization-Overlap neurons** – defined by invariance to romanization – are largely script insensitive. This suggests that representations that are not tied to script, are not causally important in generation.

• **Only-unshuffled neurons** encode order-sensitive surface regularities that are largely redundant for orthographic structure.

• **Only-native neurons** anchor script-specific realization, and their disruption induces language switching rather than structured degradation. Again, this reinforces the hypothesis that script-related neurons are most causally important.

The convergence of quantitative metrics and qualitative failure modes across Llama and Gemma indicates that invariance-based neuron identification isolates functionally meaningful components of multilingual language models.

H Extended Scaling and Semantic Competence Analysis

This appendix provides a comprehensive evaluation of our findings on larger model architectures: Llama-3-8B and Gemma-2-9B. We analyze whether the observed representational fragmentation is a symptom of limited capacity or data sparsity, or whether it persists as a stable architectural trait at scale.

H.1 Translation Performance on Romanized Inputs

To rule out the hypothesis that low representational overlap is an artifact of undertraining or data sparsity, we evaluate translation performance from native and romanized inputs to English.

Experimental Setup. We use the `dev` split of FLORES+ in an 8-shot setting. Translation quality is measured using BERTScore (`roberta-large`) against gold English references across 10 diverse languages.

Results and Discussion. As shown in Table 9, larger models achieve robust translation performance on romanized inputs. For instance, Llama-3-8B obtains BERTScores of 0.67 for Romanized Russian and 0.42 for Hindi, confirming the models have learned the semantics of romanized text and are not treating it as out-of-distribution noise.

Lang	Category	$PPL_{ratio}^{\text{target}}$	$PPL_{ratio}^{\text{ctrl}}$	p (ratio)	$\Delta PPL^{\text{target}}$	ΔPPL^{ctrl}	p (Δ)
en	overlap	3.045	0.312	3.5×10^{-81}	+588.9	−191.2	2.5×10^{-39}
en	only-unshuffled	0.799	0.312	2.8×10^{-130}	−56.5	−191.2	1.1×10^{-47}
hi	overlap	1.109	0.953	0.173	−34.2	−7.6	1.0×10^{-3}
hi	only-unshuffled	0.397	2.557	8.1×10^{-50}	−102.7	+289.9	3.6×10^{-22}
fr	overlap	1.547	0.952	1.2×10^{-89}	+116.1	−10.0	4.4×10^{-36}
fr	only-unshuffled	1.260	1.403	3.1×10^{-40}	+56.3	+90.4	4.4×10^{-24}
zh	overlap	0.842	0.150	2.0×10^{-52}	−58.8	−241.5	1.3×10^{-56}
zh	only-unshuffled	0.082	3.152	6.0×10^{-76}	−259.8	+645.7	8.5×10^{-38}

Table 7: Causal zero-ablation results for shuffling-derived neuron sets (Gemma-2-2B, raw). Values report means over the first 100 FLORES+ examples per language. Control sets consist of matched random neurons with identical cardinality.

Lang	Prompt Prefix	Clean Continuation	Patched Continuation (overlap-set zero ablation)	Inference
hi	सोमवार को, स्टैनफ़ोर्ड यूनिवर्सिटी स्कूल ऑफ़ मेडिसिन के वैज्ञानिकों ने एक नए डायग्नोस्टिक उपकरण...	इस उपकरण का उपयोग करने वाले वैज्ञानिकों ने यह भी दावा किया है कि यह...	“यह एक बहुत ही **просто**ї, सस्ती, और ابت اعتماد قابل‌کار है,” Dr	Intra-word script mixing (Devanagari + Cyrillic + Arabic)
hi	शोधकर्ताओं ने कहा है कि यह अल्प आय वाले देशों में कैंसर, टीवी, एचआईवी और मलेरिया...	अमेरिकी शोधकर्ताओं ने कहा है कि यह अल्प आय वाले देशों...	“यह एकमात्र **tool** है, जो कैंसर, **HIV**, मले**ria** और...”	English characters injected inside Hindi words
hi	कातलान की राजधानी (बार्सिलोना) में जाने के बाद से, विडाल ने क्लब के लिए 49 गेम खेले थे।	उन्होंने 2008–09 सीज़न में 12 गोल किए थे	2011–12 UEFA Champions League Final में, विड**al** hat-trick	Partial Latin substitution inside proper noun
hi	व्हाइटहॉल पर लगभग 11:00 बजे स्थानीय समय (यूटीसी +1) पर विरोध शुरू हुआ...	विरोधियों ने एक-दूसरे को हिलाते हुए...	“यह एकमात्र व**oher** में आ सकता हूं,” Prime Minister...	Orthographic corruption inside word
hi	“पनामा पेपर्स” पनामा की कानूनी फर्म मोसाक फोंसेका से लगभग दस मिलियन दस्तावेजों...	इन दस्तावेजों में से कुछ ने दुनिया भर में...	“पना**ma** पे**pers**” के 11 मिलियन दस्तावेजों में	Cross-script character insertion within named entity
hi	दस्तावेजों से पता चलता है कि चौदह बैंकों ने अमीर ग्राहकों को...	अमेरिकी बैंकों ने अरबों डॉलर की संपत्ति...	**Amri Bank**, Citibank, HSBC, JP Morgan Chase	Sudden switch to English entities mid-sentence
hi	इराक की अबू ग़रीब जेल को दंगे के दौरान आग के हवाले कर दिया गया।	इस जेल में 1000 से अधिक कैदियों को...	**Abu garib jail**, Iraq	Script change aligned with entities but breaks Hindi flow
hi	पूर्वानुमान कहते हैं कि यह तूफ़ान केप वर्डे द्वीपों के पश्चिम में...	क्या तूफ़ान केप वर्डे द्वीपों में आए हैं?	क्या तूफ़ान केप वर्**de** 2021 केप वर्**de**	English letters injected inside Devanagari word
hi	ऑस्ट्रेलिया के ट्रेज़रार और लिबरल पार्टी के नेता...	उन्होंने कहा कि न्यूक्लियर पॉवर इंडस्ट्री को...	उनका यह ع**alm** है, “यह एकमात्र را है...”	Multiscript blending within same clause
zh	周一，斯坦福大学医学院的科学家宣布，他们发明了一种可以将细胞按类型分类的新型诊断工具...	斯坦福大学医学院的科学家们开发了一种新型的微型芯片，可以将细胞...	斯坦**ford University Medical School scientists announced Monday...**	Mid-word switch to English generation
zh	自从转会到加泰罗尼亚的首府球队，维达尔已经为俱乐部踢了 49 场比赛。	在 2019–2020 赛季，维达尔在...	在 2019/2020 գ**ɡma**，维**da**	Thai and Latin characters injected inside Chinese
zh	“巴拿马文件”是巴拿马莫萨克·冯塞卡律师事务所约 1,000 万份文件的总称...	这些文件记录了 2000 年至 2015 年间...	“巴拿马文件”是巴拿马莫萨**ck Fonseca** 律师事务所	Script split inside named entity
zh	谢长廷还声称马英九虽然很上镜，但中看不中用。	谢长廷还说，马英九的“中看不中用”...	他认为，马英九的“上镜”和“上**manship**”是两码	English morpheme inserted into Chinese word
zh	不过，一位知情人士透露，凶手是红湖部落主席之子...	据悉，弗洛伊德·乔丹是红湖部落...	路易斯·乔**dan**（Louis Jourdain）被控...	Partial Latin insertion inside Chinese name

Figure 33: Qualitative examples of model behavior under shuffling-based overlap ablation (Gemma-2-2B, raw). Ablation of overlap neurons induces systematic script mixing, including partial Latin insertions and mixed-script morphemes occurring within words. Such orthographic violations are not observed for random or only-normal neuron ablations.

Critically, despite this functional competence, representational overlap remains comparatively low (∼0.25 for Llama-3-8B vs. ∼0.11 for Llama-3.2-1B), which is significantly lower than the

Lang	Category	$PPL_{ratio}^{\text{target}}$	$PPL_{ratio}^{\text{ctrl}}$	p (ratio)	$\Delta PPL^{\text{target}}$	ΔPPL^{ctrl}	p (Δ)
en	only-native	5.208	1.405	2.2×10^{-65}	+1222.6	+114.6	8.2×10^{-35}
en	overlap	0.899	0.822	6.2×10^{-96}	−28.3	−50.0	3.6×10^{-43}
hi	only-native	0.684	1.136	1.2×10^{-54}	−55.9	+24.1	2.4×10^{-24}
hi	overlap	2.228	1.036	5.8×10^{-45}	+228.0	+6.5	7.1×10^{-21}

Table 8: Cross-language mean ablation results for romanization-derived neuron sets (Gemma-2-2B, raw). Rows indicate forward-pass language; mean activations are taken from the opposite language. Results are averaged over the first 100 FLORES+ examples.

	Llama-3.2-1B		**Llama-3-8B**		**Gemma-2-2B**		**Gemma-2-9B**	
Language	Nat	Rom	Nat	Rom	Nat	Rom	Nat	Rom
Bengali	0.40	0.08	0.67	0.23	0.60	0.11	0.72	0.33
Bulgarian	0.66	0.36	0.77	0.62	0.74	0.46	0.79	0.69
Chinese	0.60	0.12	0.69	0.29	0.68	0.12	0.72	0.33
Hindi	0.58	0.14	0.72	0.42	0.69	0.22	0.76	0.53
Japanese	0.54	0.08	0.67	0.14	0.65	0.11	0.71	0.19
Korean	0.53	0.08	0.68	0.14	0.64	0.10	0.71	0.20
Marathi	0.46	0.10	0.66	0.23	0.58	0.12	0.72	0.33
Russian	0.68	0.42	0.75	0.67	0.73	0.52	0.76	0.72
Spanish	0.68	0.67	0.73	0.73	0.72	0.71	0.74	0.74
Urdu	0.47	0.07	0.68	0.19	0.60	0.10	0.73	0.32

Table 9: Translation performance (BERTScore, `roberta-large`) from Native (Nat) and Romanized (Rom) inputs to English (8-shot). Large BERTScores on romanized inputs confirm genuine semantic competence, proving that low neuron overlap is not a result of data sparsity.

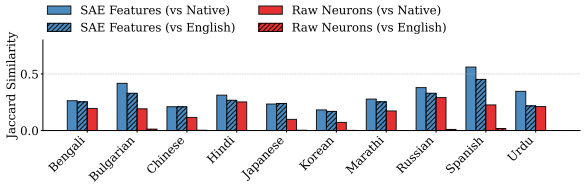


Figure 34: Jaccard similarity between romanized and native-script or English units in Llama-3-8B. Representational isolation persists at the 8B scale.

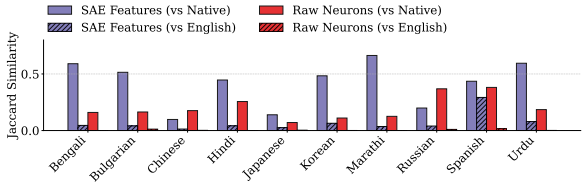


Figure 35: Jaccard similarity between romanized and native-script or English units in Gemma-2-9B. Fragmentation remains a dominant feature despite increased model capacity.

$\sim$0.60 overlap observed under word-order shuffling. This dissociation between competence and alignment confirms that models process different scripts through disjoint subspaces as a representational choice. This pattern persists even for high-resource languages like Russian and Spanish, where romanized performance nearly matches native-script performance, effectively ruling out undertraining as the sole explanation for representational fragmentation.

H.2 Script Fragmentation at Scale

We extend the representational overlap analysis to larger models to verify if increased parameter counts facilitate script unification.

Global Overlap Trends. Figures 34 and 35 report Jaccard similarities for Llama-3-8B and Gemma-2-9B. Across both models, romanized inputs maintain near-zero overlap with English and consistently low overlap with their native-script counterparts. This indicates that even with increased capacity, models do not converge toward a unified, script-invariant representation.

Layer-wise Persistence. Figures 36 and 37 illustrate layer-wise alignment. While a slight increase in overlap is observed in middle layers, alignment remains far from convergence across the entire depth of the network. This confirms that representational separation is a fundamental architectural trait that persists even as models become larger and more competent.

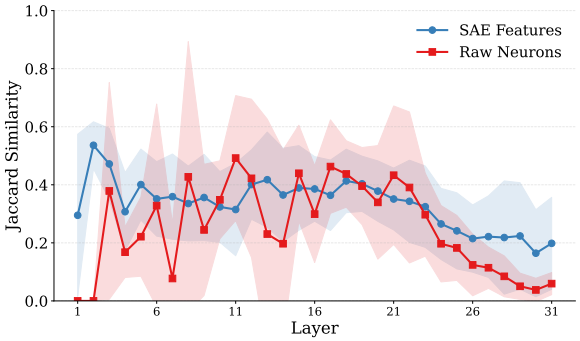


Figure 36: Layer-wise alignment in Llama-3-8B. Mid-layer increases do not lead to cross-script convergence.

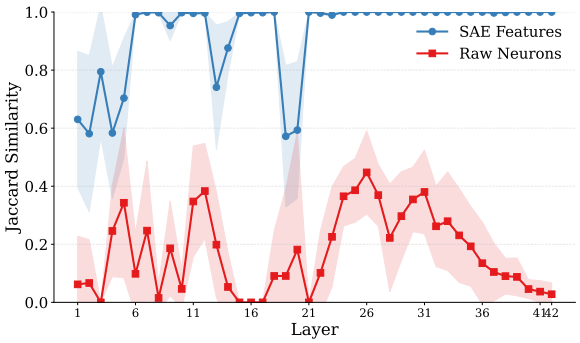
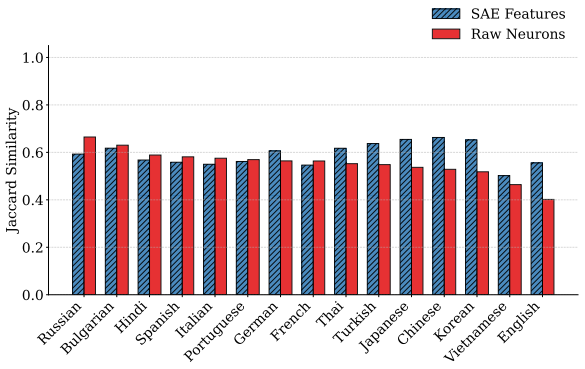


Figure 37: Layer-wise alignment in Gemma-2-9B, showing consistent representational separation in raw neurons across depth.

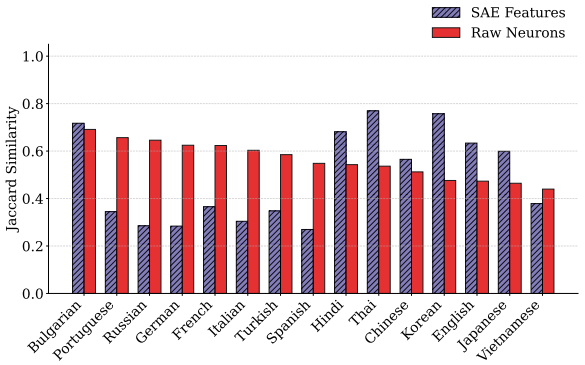
H.3 Structural Robustness at Scale

Finally, we examine whether larger models maintain the high robustness to structural (word-order) perturbations observed in 1B and 2B models.

High Overlap Under Shuffling. As shown in Figure 38, both Llama-3-8B and Gemma-2-9B exhibit consistently high Jaccard overlap between units identified from original and shuffled inputs. This confirms that the models' reliance on token-level and distributional cues (rather than strict syntactic order) is a scale-invariant property.



(a) Llama-3-8B



(b) Gemma-2-9B

Figure 38: Jaccard similarity between units from original and shuffled text at scale. Robustness to word-order perturbation remains consistently high across larger architectures.