How Far Do Auto-Interpretation Labels Generalize: A Controlled Study Across Languages, Scripts, and Rewordings

Sripad Karne
Columbia University
sk5695@columbia.edu

Abstract

Sparse autoencoder (SAE) features are increasingly used to interpret language models, with auto-generated natural-language labels serving as the primary interface for understanding what each feature represents. We ask whether these labels generalize: does a feature labeled for a concept actually track that concept across languages and scripts? Using Serbian digraphia as a controlled testbed—the same language written in both Latin and Cyrillic via deterministic transliteration—we first find that SAE feature sets activated by the same content in different languages, scripts, and wordings share substantial overlap (mean Jaccard 0.39 vs. 0.13 random baseline, peaking at 0.57), suggesting genuine cross-lingual semantic features. We then test whether auto-interpretation labels keep pace. They often do not: features whose labels describe semantic content miss the same meaning in Serbian up to $4\times$ more often than within English, and miss Serbian Cyrillic more than Serbian Latin—two scripts that are deterministic transliterations of each other. The gap grows with network depth, yet the labels give no indication that they fail. These results suggest that auto-interpretation labels reflect a feature's behavior on the languages and scripts a model has seen most in training, rather than the concept itself.

1 Introduction

Sparse autoencoders (SAEs) have become a standard tool for inspecting the internals of language models, decomposing dense activations into sparse features that are easier to interpret (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024). Because models contain far more features than anyone can inspect by hand, each feature is typically labeled automatically: a language model reads the feature's top-activating examples and writes a short natural-language description (Bills et al., 2023; Lin, 2023). These labels are then what practitioners, such as interpretability researchers and auditors using feature directories like Neuronpedia, rely on—to understand a model, to audit its behavior, or to steer it.

But how far do these labels generalize? Consider a feature labeled "deception" or "violent content." A researcher reads the label, takes the feature to track that concept, and relies on it—perhaps to monitor safety-relevant behavior. What the label does not say is whether the feature tracks that concept equally well across languages and scripts. If a feature detects deception mainly in English, weakening in Russian and weakening further in Serbian Cyrillic, then the same content that triggers it in one form could pass unnoticed in another. The label may name the concept correctly; what it omits is the *range* over which the feature actually tracks it.

Testing this requires a setting where surface form can be varied while meaning is held exactly fixed. Serbian digraphia offers precisely this control: Serbian is natively written in both Latin and Cyrillic scripts, related by deterministic, lossless transliteration, so we can change script while holding language, wording, and meaning constant. Combined with cross-language controls (English, Russian), we construct a factorial paradigm that isolates script, language, wording, and meaning independently.

First, in what we call Leg 1 of our analysis, we find evidence that SAE features encode abstract meaning: feature sets activated by the same content across different languages, scripts, and wordings overlap substantially (mean Jaccard 0.39 vs. 0.13 baseline, peaking at 0.57), a pattern robust across model scales, architectures, and SAE hyperparameters, providing evidence that cross-lingual semantic features exist and are worth labeling. Second, in Leg 2, we test whether the auto-interpretation labels assigned to these features hold up when the same content is rendered in Serbian.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3120–3151
November 6-10, 2026 ©2026 Association for Computational Linguistics

They do not. Content-labeled features miss the same meaning in Serbian up to $4\times$ more often than within English—a gap that grows with network depth and that the labels themselves give no indication of—and miss Serbian Cyrillic up to $1.35\times$ more than Serbian Latin despite the two being identical up to transliteration. Our contributions are:

1. A controlled cross-lingual evaluation showing that auto-interpretation labels can fail systematically on less-represented languages and scripts, with failures aligning with estimated training coverage, and that this failure is invisible from the label itself.

2. A factorial paradigm built on Serbian digraphia, two native scripts of one language related by lossless transliteration, that isolates script, language, wording, and meaning in SAE feature sets, confirming cross-lingual semantic features at the feature-set level.

3. A controlled multilingual evaluation suite built on FLORES+: 300 sentences in four language-script variants with validated paraphrases and matched random partners, designed for factorial analysis of feature-level interpretability methods and released to support future work.[1]

2 Related Work

Sparse autoencoders and feature interpretation. Sparse autoencoders (SAEs) decompose model activations into sparse, more interpretable features, offering a route to inspecting representations beyond individual neurons (Bricken et al., 2023; Cunningham et al., 2024; Gao et al., 2025). Open SAE suites such as Gemma Scope 2 have made this practical at scale (McDougall et al., 2025), and feature directories built on them are now a common entry point for interpretability work. We use Gemma Scope 2 SAEs and the labels served through one such directory, treating them as the deployed tools a practitioner would actually reach for.

Automated interpretation and its reliability. Features are typically labeled automatically through auto-interpretation (Bills et al., 2023; Lin, 2023). Later pipelines have built on this approach: Delphi (Paulo et al., 2025) scales labeling to millions of features, and PRISM (Kopf et al., 2025) extends it to multiple descriptions per feature to capture polysemanticity. Across these pipelines, however, the reliability of the resulting labels remains an open concern. Prior work shows that automated explanations can be vague or inaccurate, and has proposed more rigorous ways to assess them (Huang et al., 2023; Liu et al., 2026). These evaluations, however, are monolingual: they ask whether a label fits a feature's behavior on the inputs it was derived from, not whether it still holds when the same content appears in another language or script. To our knowledge, that gap has not been tested, and it is what we test here.

Multilingual representations and script. Multilingual models are known to represent meaning in a partly language-agnostic way: equivalent sentences in different languages converge in a shared semantic space in the middle layers, before representations specialize toward the output language near the end (Wendler et al., 2024; Wu et al., 2025). This shared structure has been probed at the SAE feature level, where studies recover language-specific features (Deng et al., 2025) and grammatical concepts shared across typologically diverse languages (Brinkmann et al., 2025). Verma et al. (2026) found that language-associated features are strongly conditioned on script: romanizing a non-Latin language activates a largely disjoint set of features from its native form. Romanized text, however, is a largely synthetic form a model rarely encounters, so it changes script and naturalness at once. Serbian digraphia instead contrasts two scripts the language is genuinely written in, isolating script within forms the model has actually seen. This control is what lets us test feature labels cleanly: if features track meaning across languages and scripts, the labels assigned to them should hold across them too. We test whether they do.

3 Methods

3.1 Dataset

Source corpus. We draw 300 sentences from the FLORES+ devtest split (Goyal et al., 2022; NLLB Team et al., 2024; Burchell et al., 2024), a multilingual benchmark of professionally translated sentences, stratified across its Wikinews, Wikijunior, and Wikivoyage sources (roughly 100 each) to span news, instructional, and travel registers (Appendix A.7). FLORES+ supplies aligned profes-

[1]Dataset and code available at https://github.com/Sripadkarne/auto-interp-cross-lingual-eval.

Variant	Language	Script
En-Latin	English	Latin
Sr-Latin	Serbian	Latin
Sr-Cyrillic	Serbian	Cyrillic
Ru-Cyrillic	Russian	Cyrillic
Condition	**Description**	
Original	FLORES+ reference translation	
Paraphrase	Meaning-preserving rewrite	
Random	Controlled-unrelated sentence, same variant	

Table 1: The four language-script variants and three conditions.

sional translations of each sentence into Serbian (Cyrillic) and Russian (Cyrillic).

Language and script variants. Each sentence appears in four language-script variants: English in Latin, Serbian in Cyrillic, Serbian in Latin, and Russian in Cyrillic. Serbian Latin is produced by deterministic, lossless transliteration of Serbian Cyrillic using the Vuk Karadžić mapping implemented in implemented in `cyrtranslit`[2]. The two Serbian variants thus differ only in script while holding content fixed, the property our paradigm exploits to isolate script from language and meaning.

Conditions. Within each variant, every sentence appears in three conditions. The *original* is the FLORES+ professional translation, or its transliteration for Serbian Latin. The *paraphrase* is a meaning-preserving rewrite with varied surface form, generated and validated as described below. The *random partner* is an controlled-unrelated FLORES+ sentence in the same language and script, length-matched to within three words; it shares script and language with the target but no semantic content. Table 1 summarizes the full set of variants and conditions. Crossing the four language-script variants with the three conditions (original, paraphrase, random partner) across all 300 sentences yields $300 \times 4 \times 3 = 3{,}600$ texts.

Paraphrase generation. English paraphrases were generated with Claude Opus 4.6, prompted to preserve meaning while varying surface form and matching length within 30% from the original FLORES+ phrase. Serbian and Russian paraphrases were obtained by translating the English paraphrase with the same model; Serbian Latin by transliterating Serbian Cyrillic. All candidates were filtered with LaBSE (Feng et al., 2022) (cosine similarity ≥ 0.80). Full prompts appear in Appendix A.1.

Native-speaker validation. For each of English, Serbian, and Russian we recruited two validators. Each judged 200 sentences with a 100-sentence overlap between the pair, covering 300 unique sentences per language. Each sentence pair received two binary judgments: whether the paraphrase preserves the original *meaning*, and whether it reads as *natural* text. A pair passing both was accepted as is; otherwise the validator supplied a corrected paraphrase, which was added to the dataset and flagged for reference (full instructions in Appendix A.2). These corrections do not drive our results, with no robust difference between accepted and corrected paraphrases in our main analyses (Appendix D, Table 29).

Disagreement resolution. On the 100 overlap sentences per language, we resolved decisions conservatively: when both validators flagged it with different corrections, we chose between them by whether the sentence's FLORES+ index was even or odd, a rule fixed before any corrections were examined (Appendix A.3).

Validation outcomes. Clean-accept rates were 96.0% (English), 90.7% (Russian), and 70.7% (Serbian); corrections were minor (median 1–3 words). Inter-rater agreement was perfect for English, high for Russian (Gwet's AC1 (Gwet, 2008) 0.87), and moderate for Serbian (AC1 0.63). Full statistics appear in Appendix A.4.

3.2 Models and SAEs

Models. Our primary model is Gemma-3-27B (Gemma Team, 2025). We deliberately use standard, publicly released models, SAEs, and labels, since our aim is to audit the interpretability stack practitioners actually use. To check that our findings generalize across model scale, we replicate the core analyses on Gemma-3-1B and Gemma-3-12B; the factorial decomposition holds across all three sizes (Appendix B.4). To check generalization across model family we run Llama-3.1-8B with Llama Scope SAEs (He et al., 2024), a different model architecture, SAE training regime, and dictionary width; the qualitative structure replicates (Appendix B.5).

Sparse autoencoders. We use the Gemma Scope 2 suite of SAEs (McDougall et al., 2025).

[2] https://zenodo.org/records/17663256

All SAEs use the JumpReLU activation (Rajamanoharan et al., 2024), a dictionary width of 16,384 features, and are selected at low L_0 sparsity ($L_0 \in [10, 20]$). We analyze every layer of each model, using the layer-matched SAE at each depth. To confirm our findings are not artifacts of the extraction setup, we vary dictionary width (Appendix B.2) and L_0 sparsity (Appendix B.1) on Gemma-3-27B and find the decomposition robust to both. We also compare pooling strategies (Appendix B.3): the ordering holds under last-token and mean pooling, but only last-token pooling cleanly separates conditions from their baselines, so we use it throughout.

3.3 Feature Extraction and Similarity

Active feature sets. Given a text s, we run it through the model and, at each layer ℓ, take the residual-stream hidden state at the final token position. We pass this hidden state through the layer-ℓ SAE encoder, yielding a feature activation $a_i^{(\ell)}(s)$ for each feature i, and define the active feature set as those that fire:

$$F_\ell(s) = \{\, i : a_i^{(\ell)}(s) > 0 \,\}.$$

Under the JumpReLU activation, a feature is nonzero only when its pre-activation exceeds a learned per-feature threshold, so this binary criterion tends to select features that clear a nontrivial gate; we quantify this in Appendix A.6.

Similarity. We measure the overlap between two texts' active feature sets with the Jaccard index,

$$J_\ell(s, s') = \frac{|F_\ell(s) \cap F_\ell(s')|}{|F_\ell(s) \cup F_\ell(s')|},$$

computed independently at each layer.

Factorial comparisons. We isolate script, language, wording, and meaning one at a time by comparing pairs that hold the remaining properties fixed (Table 2):

- **Script:** Sr-Cyrillic orig vs. Sr-Latin orig (same language and content, only script varies).
- **Language:** Sr-Cyrillic orig vs. Ru-Cyrillic orig (same script and content, language varies).
- **Wording:** English orig vs. English paraphrase (same language and script, surface form varies).
- **Meaning:** English orig vs. Russian paraphrase (script, language, and wording all differ; only meaning shared).

Each is read against a random-partner baseline sharing only surface properties, except wording, which uses the identity ceiling of 1.0.

3.4 Auto-interpretation evaluation

After describing what features collectively encode, we now ask whether the natural-language labels assigned to individual features hold up. These labels are generated automatically, as described in Section 1, and are often what practitioners rely on to understand and act on a model's internals. A content label makes an implicit prediction: the feature should respond to that content however it is written. Our paradigm lets us check this directly by varying surface form while holding meaning fixed.

Selecting features to test. A feature candidate must pass two independent content checks. First, it fires on both the English original and the Russian paraphrase of a sentence—a pair that differs in script, language, and wording simultaneously, so co-activation is strong evidence the feature tracks meaning rather than any surface property. We anchor on English and Russian rather than either Serbian script, which would give that script a selection advantage; this leaves both Serbian forms symmetrically un-anchored. Since the within-language floors are rewordings of the anchors, absolute cross-language magnitudes should be read as upper bounds; the Cyrillic/Latin ratio is unaffected. For each such feature, we retrieve its auto-interp label from Neuronpedia, a deployed pipeline serving Gemma Scope 2 descriptions that practitioners consume (Lin, 2023). We use two labelers, Claude Sonnet (the strongest available) and Gemini Flash Lite (Neuronpedia's default), to confirm the results are not an artifact of one labeler. Labels are retrieved at the four layers Neuronpedia provides (16, 31, 40, 53), spanning early-middle to late depth. From the features that fire on both the English original and the Russian paraphrase, we apply a second check: a separate classifier LLM (`claude-sonnet-4-6`, distinct from the auto-interp labelers) classifies each feature's auto-interp label as a *content* claim (semantic topic or meaning) versus a *surface* or *language* claim, and the features whose labels describe surface form rather than content are discarded. This filter is not

Test	Comparison	Script	Language	Wording	Meaning
Script	Main: Sr-Cyr Orig vs. Sr-Lat Orig	✗	✓	✓	✓
	Floor: Sr-Cyr Orig vs. Rus-Cyr Rand	✓	✗	✗	✗
Language	Main: Sr-Cyr Orig vs. Ru-Cyr Orig	✓	✗	✓	✓
	Floor: Sr-Cyr Orig vs. Sr-Lat Rand	✗	✓	✗	✗
Wording	Main: Eng-Lat Orig vs. Eng-Lat Para	✓	✓	✗	✓
	Floor: Identity Ceiling (J=1.0)	✓	✓	✓	✓
Meaning	Main: Eng-Lat Orig vs. Rus-Cyr Para	✗	✗	✗	✓
	Floor: Eng-Lat Orig vs. Rus-Cyr Rand	✗	✗	✗	✗

Table 2: Controlled comparisons isolating each property of a text. ✓ = property *shared* between the two texts and ✗ = property *differs*. Each test isolates one property in its *main* comparison and reads it against a reference *floor*. Note that the wording comparison uses the identity ceiling rather than a random floor because two texts cannot differ in meaning while sharing the same wording.

load-bearing: repeating the analysis on the unfiltered pool yields statistically indistinguishable results, with overlapping confidence intervals at every layer (Appendix D, Table 28). Full classifier details, prompt, and per-layer discard rates are in Appendix C.

Cross-language falsification. For each (feature, sentence) unit we test whether the feature also fires on renderings of the same sentence in other forms: an English paraphrase and the Russian original (within-language floors), and the Serbian Latin and Serbian Cyrillic versions (cross-language test). A content label predicts firing on all of these; failure on any is a miss. We quantify this per condition: for each rendering c, the miss rate is the fraction of (feature, sentence) units (i, s) in the pool $\mathcal{P}_\ell$ where the feature fails to fire on that rendering s_c:

$$\mathrm{MR}_\ell(c) = \frac{1}{|\mathcal{P}_\ell|} \sum_{(i,s)\in\mathcal{P}_\ell} \mathbf{1}\left[\, i \notin F_\ell(s_c) \,\right] \quad (1)$$

We report the miss rate per condition with 95% feature-clustered bootstrap confidence intervals (10,000 resamples). The Cyrillic/Latin ratio is bootstrapped within the same feature draw. Flat pair-level resampling gives the same conclusions with narrower intervals; we report the more conservative clustered version. The two Serbian scripts are deterministic transliterations of each other, so their miss rates are directly comparable and their ratio isolates script-specific bias. We report Claude Sonnet as the primary labeler; Gemini gives the same pattern (Appendix D).

4 Results

4.1 What SAE features encode

We first map how a text's active feature set is organized, varying script, language, wording, and meaning one at a time while holding the others fixed (Table 2). For each property we measure the Jaccard overlap of active SAE feature sets across all layers of Gemma-3-27B, comparing each against an appropriate reference baseline. We treat script, language, and wording as controls that isolate meaning, the property central to the rest of our analysis, and summarize these controls briefly before turning to meaning. Figure 1 shows the full layer-wise picture: each panel isolates one property, plotting its Jaccard overlap across all 62 layers.

Script, language, and wording. Within our digraphic setting, script has strikingly little effect: transliterating a sentence between Serbian Latin and Cyrillic holds language, wording, and meaning fixed and preserves most of the active set (mean Jaccard 0.73), far above the baseline. In our controlled setting it confirms that script is the mildest of the four perturbations. Notably, this overlap is negligibly correlated with the Cyrillic-vs-Latin token-count difference, ruling out tokenization as its driver (Appendix A.5). Language matters more: the same sentence in two languages (Serbian vs. Russian, same script) shares less of its feature set (mean 0.50), indicating that language has a larger effect on the active set than script does. Wording sits between the two: paraphrasing within a language preserves a substantial share of the feature set (mean 0.63). Each property also changes differently across depth: non-script overlap stays high until a sharp drop near the output, non-language

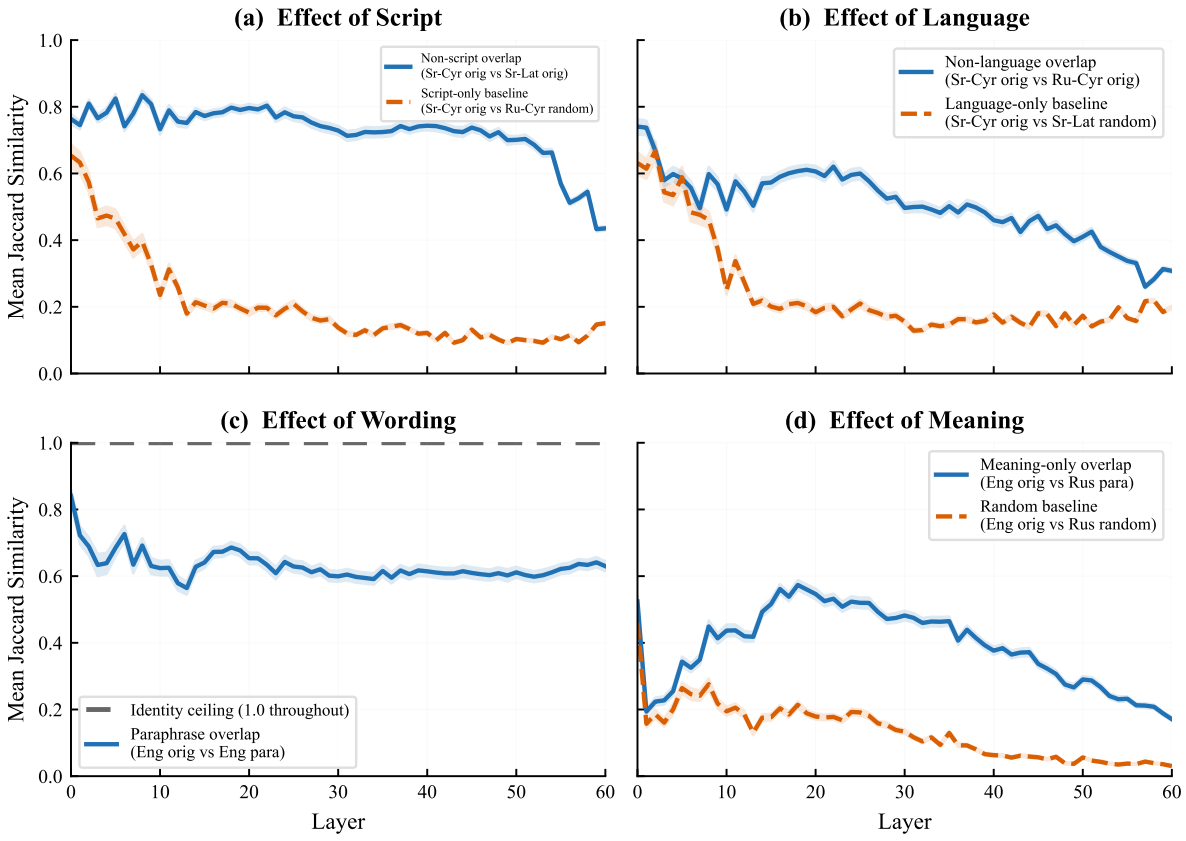


Figure 1: SAE feature overlap decomposed by script, language, wording, and meaning across all layers of Gemma-3-27B (300 sentences; shaded bands are 95% bootstrap CIs, 10,000 resamples). In each panel a solid *main line* varies only the target property while holding the others fixed, read against a dashed *baseline* (or, in (c), the identity ceiling); see Table 2 for the exact comparisons.

overlap stays high in the early layers then declines through the second half, and wording overlap stays roughly flat. The late declines in non-script and non-language overlap are consistent with prior work suggesting that later transformer layers become increasingly specialized for next-token prediction and output generation (Geva et al., 2022; Belrose et al., 2023). We make no strong claim here, however: the corresponding script and language baselines are themselves roughly flat across the later layers, so the picture is suggestive rather than conclusive. These trajectories are worth a detailed study, which the paradigm supports and we leave to future work. These three properties show how the active set responds to surface variation. We now turn to the property at the center of our analysis: meaning.

Meaning. We isolate meaning by comparing two texts that share nothing but their content—an English sentence and a Russian paraphrase, differing in script, language, and wording—so any overlap in their feature sets points to genuinely cross-lingual semantic features rather than surface form. Even with script, language, and wording all changed at once, overlap holds at a 0.39 Jaccard average, well above the baseline (mean 0.13), and peaks at 0.57 in the lower-middle layers (Figure 1d). SAE features therefore appear to encode abstract meaning that largely survives a complete change of surface form. The signal is also depth-dependent: after an initial drop in the first layer or two, the overlap climbs to a peak around the lower-middle of the network before declining through the later layers.

From feature sets to feature labels. Having established that genuinely semantic features exist, we next need a way to identify what each one represents. Across thousands of features, this is typically done with auto-interpretation, which assigns each feature a natural-language label. Our set-level analysis does not address this: it describes populations of features, not whether any individual fea-

ture's label holds. We turn to that question next.

4.2 Auto-interpretation labels generalize poorly across languages and scripts

Leg 1 gave a set-level picture: SAE features appear to encode genuine semantic content, while script, wording, and language reshape the active set to differing degrees. Auto-interpretation provides a feature-level view, labeling each feature from its top-activating examples. We evaluate these labels on the cross-lingual semantic features identified in Leg 1: those that fire on a sentence's content in both English and Russian, which confirms behaviorally that they track meaning rather than surface form. As a second robustness check, we confirm these features' labels describe content rather than surface form, using an LLM to classify each label as content or surface and discarding the exceptions (Appendices C and D).

Testing labels from Neuronpedia's auto-interpretation pipeline, where Claude Sonnet and Gemini Flash Lite serve as the labelers, we find these predictions often break down: a feature labeled for its content can fail to fire on that same content when it is rewritten in another language or script, and this happens more for less well-represented forms. We quantify this as the *miss rate*: the fraction of cases where a content-labeled feature fails to fire on the same meaning in another form (Table 3). We report Claude Sonnet labels, the stronger labeler; Gemini gives the same picture, converging on near-identical Cyrillic/Latin ratios at every layer (e.g., 1.35× vs. 1.36× at layer 53), which suggests the failure lies in the features rather than in either labeler. The full breakdown and additional baselines (including random controls) are in Appendix D.

Even within a language, content labels are unreliable; across languages they break down. Reworded within their own language, these features already miss a non-trivial share of the time, from 14.8% in early layers to 7.3% at layer 53 for English, and staying in the 13–17% range for Russian. The same features miss Serbian far more often. Averaged over the two Serbian scripts, miss rates start near the within-language floor (21.0% versus 14.8% at layer 16) but reach 30.1% by layer 53 (Table 3a)—an excess of more than 20 points over the English floor, roughly 4× the within-language miss rate. Paraphrasing the Serbian text differs only by 1 to 3 points (Appendix D), confirming the failure comes from the language and script, not the specific wording within them. Across conditions, miss rates are ordered from English through Russian and Serbian Latin to Serbian Cyrillic (Table 3).

Layer	Eng-para (%)	Rus-orig (%)	Serbian avg (%)
16	14.8 [12.4, 18.6]	16.6 [12.5, 22.0]	21.0 [15.8, 28.2]
31	9.2 [7.0, 11.7]	13.7 [11.4, 16.6]	21.5 [19.4, 24.1]
40	9.1 [7.7, 10.5]	15.6 [14.0, 17.3]	24.1 [21.8, 26.5]
53	7.3 [5.8, 9.0]	15.4 [13.3, 17.6]	30.1 [26.6, 33.8]

(a) Within- vs. cross-language miss rates.

Layer	Sr-Lat (%)	Sr-Cyr (%)	Cyr/Lat
16	20.8 [15.8, 27.5]	21.3 [15.7, 29.0]	1.02× [0.91, 1.12]
31	20.4 [18.2, 23.4]	22.6 [20.4, 25.1]	1.11× [1.01, 1.20]
40	22.2 [19.9, 24.6]	25.9 [23.4, 28.7]	1.17× [1.09, 1.26]
53	25.6 [22.1, 29.6]	34.6 [30.7, 38.6]	1.35× [1.23, 1.48]

(b) Serbian script asymmetry.

Table 3: Content-feature miss rates over the cross-lingual content pool. (a) Within-language floors (Eng-para, Rus-orig) vs. the cross-language Serbian average. (b) Serbian Latin vs. Cyrillic, with the Cyrillic/Latin ratio. 95% bootstrap CIs in brackets (10,000 resamples, feature-clustered).

The Serbian Cyrillic disadvantage grows with depth. The two Serbian scripts are deterministic transliterations of each other, so a content feature should treat them identically. Instead they diverge with depth: indistinguishable in early layers (1.02× Cyrillic-to-Latin miss ratio at layer 16), the gap rises to 1.35× at layer 53 (Table 3b). Since the scripts share language, wording, and meaning, the failure is unlikely to be semantic; the main difference is the character system, which points to how well the model represents each script. The same gap appears whether we test the original Serbian sentences or their paraphrases, so it does not depend on the particular wording. The Cyrillic disadvantage holds within every topic category (Appendix D), confirming it is not specific to any particular domain. Table 4 gives individual examples: content-labeled features that fire on a sentence in three of the four forms but fail on one Serbian script. A manual check of a sample of missed features confirms that the labels are largely accurate on their own top-activating examples (Appendix D.2).

The divergence sharpens with depth. These failures are mildest in the earliest layer we can probe: at layer 16 the within-language floors and

L40 / F4327 — misses Sr-Cyrillic
Auto-interp: Sonnet: "Olympic athletes" | Gemini: "Olympic athletes and comebacks"
Sentence: "Poland's visually impaired skier Maciej Krezel…"

L53 / F3041 — misses Sr-Cyrillic
Auto-interp: Sonnet: "political leaders" | Gemini: "political leaders or figures"
Sentence: "Hu encouraged developing countries to avoid the old path…"

L40 / F8424 — misses Sr-Latin
Auto-interp: Sonnet: "diseases and conditions" | Gemini: "diseases and conditions"
Sentence: "Infectious diseases themselves, or dangerous animals…"

Table 4: Content-labeled features that fail on one Serbian script, verifiable on Neuronpedia. Additional examples in Appendix D, Table 25.

the Serbian miss rate sit fairly close (14–17% versus 21.0%). They pull apart deeper in the network. The within-language floors hold steady or fall while the Serbian miss rate climbs (Table 3a). Reliability thus holds or improves within a language and degrades across one as depth increases. These patterns replicate across model scale: Gemma-3-1B and Gemma-3-12B show the same ordering and a Cyrillic/Latin asymmetry that grows with depth (Appendix E).

5 Discussion

Semantic features are real, but their labels overpromise. Our two analyses give a coherent picture. Leg 1 indicates that SAE features genuinely encode abstract meaning: cross-lingual semantic overlap exists and survives a complete change of script, language, and wording. Leg 2 shows that this encoding is narrower than its labels suggest. A feature that tracks a concept in both English and Russian, and is labeled for that concept, still fails to fire on it in Serbian noticeably more often than it does within those languages. The two findings are not in tension: these features carry real semantic content, which is why their labels can be reasonable, but that content does not transfer cleanly across languages and scripts. The failure is not that the labels are always wrong: they often capture something real about the feature, yet can still fail to predict its behavior on an equivalent rendering of the same content, with no indication of where.

The failures are consistent with expected training distribution. The pattern of failures is consistent with how often each form of a meaning is likely to appear in pretraining. Averaged across layers, content features miss their target far more in Serbian than within English or Russian (10.1% and 15.3% versus 22.2% for Serbian Latin and 26.1% for Serbian Cyrillic). The clearest case is the Serbian script asymmetry: the two scripts are deterministic transliterations, identical in language, wording, and meaning, so a difference between them cannot be semantic. The only thing that differs is the character system, and features miss Serbian Cyrillic more than Serbian Latin. This is consistent with broad patterns in web text. Web corpora are heavily English-dominant: English is the primary language of roughly 40–45% of HTML pages in recent Common Crawl snapshots, and no other language exceeds about 7% (Common Crawl Foundation, 2026). Russian, the second-largest, accounts for roughly 6–7% of pages versus about 0.2–0.27% for Serbian (Common Crawl Foundation, 2026), while Latin script is estimated to dominate Serbian usage (roughly 70:30 Latin to Cyrillic) (Serbian National Internet Domain Registry (RNIDS), 2016). Together, these factors suggest that Serbian Cyrillic is comparatively scarce in the web text that dominates modern pretraining corpora. Such imbalances have measurable downstream effects: model performance has been shown to align with a language's share of pretraining data (Li et al., 2025). A similar pattern could hold at the level of individual features: the Serbian Cyrillic gap is what this predicts, with the rarest form of the content the one its features handle worst. This is not a binary high- versus low-resource split but a graded one—English is the most dominant form, Russian high-resource but less dominant, and Serbian rarer still—with reliability ordered accordingly.

How the gap changes with depth. As Section 4.2 shows, this ordering sharpens with depth. One reading is that deeper layers grow less language-agnostic, re-specializing toward specific surface forms (Wendler et al., 2024; Liu and Niehues, 2025); if that specialization favors the forms a model has seen most, content features could fire more reliably on well-represented inputs and less on rarer ones, as we observe; however, we cannot establish this directly. The within-language trend, however, is notable: a feature becoming *more* reliable for its strongest form with depth is, to our knowledge, new, and we offer it as an empirical observation. One speculative possibility is

that this is a depth effect: where early layers respond more uniformly across a meaning's different language and script renderings, deeper layers concentrate firing on the renderings the model has seen most; we leave testing this to future work.

None of this is visible from the label itself. Consider a feature labeled "deception" or "violent content." A researcher auditing the model reads the label, takes the feature to track that concept, and relies on it—perhaps to monitor or steer the behavior. What the label would not reveal is that such a feature might detect the concept mainly in English, with its response weakening in Russian and weakening further in Serbian Cyrillic, so the same content that triggers it in one form could pass unnoticed in another. The label may name the concept correctly; what it can omit is the range over which the feature actually tracks it.

Implications for practitioners. The practical lesson is that an auto-interpretation label guarantees nothing about how a feature behaves on the same concept in another form. A content label can be locally accurate and still fail to generalize across languages and scripts, and because the label carries no signal of this, the failure is invisible to monolingual inspection. Evaluating feature labels therefore requires controlled stimuli that vary surface form while holding meaning fixed. The failures we document also appear addressable: auto-interpretation pipelines could draw top-activating examples from multiple languages, or flag features whose activation is inconsistent across equivalent inputs, though we leave validation of such mitigations to future work.

Concurrent and recent work. The interpretability toolchain is developing rapidly. Circuit tracing (Ameisen et al., 2025) traces computational graphs through model internals, and Natural Language Autoencoders (Fraser-Taliente et al., 2026) translate full activation vectors into natural-language explanations. These efforts build increasingly powerful tools for inspecting models; our work is complementary, contributing controlled evaluation methodology that checks whether the descriptions these tools produce generalize beyond the conditions they were derived from. Both NLAs and the auto-interp labels we evaluate face the same underlying challenge: a natural-language description of a model's internals may be accurate on the inputs it was built from and quietly fail on others.

6 Conclusion

Using Serbian digraphia as a controlled testbed, we find evidence that SAE features in multilingual language models encode abstract meaning that can survive complete changes of script, language, and wording. However, the auto-interpretation labels assigned to these features do not keep pace: content labels that hold in well-represented languages quietly fail on less-represented ones, with miss rates reaching $4\times$ the within-language floor for Serbian, and the labels themselves carry no signal of where they break. Even between Serbian Cyrillic and Latin, identical up to transliteration, miss rates diverge by up to $1.35\times$, a gap that cannot be semantic. The failures are graded, aligning with the estimated representation of each form in training data, and tend to sharpen with network depth. These results suggest that auto-interpretation labels should be understood as claims about a feature's behavior on its best-represented inputs, not as guarantees about the concept in general.

Limitations

Our cross-lingual evaluation covers three languages (English, Russian, and Serbian) from two language families, with Serbian closely related to Russian. The graded miss-rate pattern we observe should in principle extend to more typologically distant and lower-resource languages, but we have not verified this beyond the conditions in our dataset.

Our sample of 300 sentences is small, a trade-off we made to prioritize validation quality; scaling it with LLM-generated data is a natural extension, for which our human-validated set can serve as a reference to measure the reliability of fully automated generation.

We also treat each missed feature in isolation: when a specific feature fails to fire, a more general related feature may fire in its place, possibly reflecting lower confidence on less-represented inputs rather than lost content. Examining which features fire instead is a natural direction for future work.

Leg 2 replicates across Gemma model scales (1B, 12B, 27B; Appendix E), but we have not confirmed the label-failure patterns on other model families or labeling pipelines beyond Neuronpedia. Within Neuronpedia, we also hold the labeling procedure fixed: two independent labelers agree, but we do not vary how many top-activating examples

each sees, which we leave to future work.

Ethics Statement

Native-speaker validators were recruited through personal networks and compensated $125–$200 per batch of 200 sentence pairs, exceeding standard crowdsourcing rates for comparable linguistic annotation. No personal or sensitive data was collected from participants. Paraphrases were generated using Claude Opus 4.6 and auto-interpretation labels were retrieved from Neuronpedia's deployed pipeline (Claude Sonnet and Gemini Flash Lite); all prompts are reported in full in Appendices A.1 and C. The FLORES+ source data is used under its CC BY-SA 4.0 license. Our released dataset and code are available under CC BY 4.0 for research use.

Use of Large Language Models. Claude was used to assist with paper drafting, editing, and code implementation. All experimental design, analysis, and scientific claims are the author's own.

Acknowledgments

I thank Rejn Omeragic and Nazgul Maksutkhan for helping me recruit native speakers of Russian and Serbian, and I thank the validators who made the full multilingual evaluation possible: Samuel Dudkin and Lennard Teslia (Russian), Elena and Jovana (Serbian), and Tharangini Jakkena, Ram Karne, and Sai Srikar Karne (English). I also thank Tianyi Lorena Yan and Jonathn Chang for reviewing the final draft and providing valuable feedback.

References

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, and 5 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

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.

Laurie Burchell, Jean Maillard, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. Findings of the WMT 2024 shared task of the open language data initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.

Common Crawl Foundation. 2026. Statistics of Common Crawl monthly archives: Distribution of languages. Accessed 2026-09-30.

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

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua

Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. 2026. Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*.

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

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

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

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously assessing natural language explanations of neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.

Laura Kopf, Nils Feldhus, Kirill Bykov, Philine Lou Bommer, Anna Hedström, Marina M.-C. Höhne, and Oliver Eberle. 2025. Capturing polysemanticity with PRISM: A multi-concept feature description framework. In *Advances in Neural Information Processing Systems*, volume 38, pages 82938–82974.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying LLM performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Johnny Lin. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. Software available from neuronpedia.org.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.

Weiqi Liu, Yongliang Miao, Haiyan Zhao, Yanguang Liu, and Mengnan Du. 2026. NeuronScope: A multi-agent framework for explaining polysemantic neurons in language models. *arXiv preprint arXiv:2601.03671*.

Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. 2025. Gemma Scope 2 - technical paper. Technical report, Google DeepMind.

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, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. 2025. Automatically interpreting millions of features in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 48393–48421. PMLR.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *Preprint*, arXiv:2407.14435.

Serbian National Internet Domain Registry (RNIDS). 2016. Cyrbusters panel debate held: Busting myths about Cyrillic on the Internet. RNIDS news, 28 January 2016. Accessed 2026-09-30.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, and 7 others. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Aastha A K Verma, Anwoy Chatterjee, Mehak Gupta, and Tanmoy Chakraborty. 2026. Multilingual language models encode script over linguistic structure. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33685–33719, San Diego, California, United States. Association for Computational Linguistics.

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.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*, pages 53705–53723.

A Dataset Details

A.1 Paraphrase Generation

All paraphrases were produced with Claude Opus 4.6 (`claude-opus-4-6`) using extended thinking (10,000-token budget) at temperature 1.0, with the sentence to process passed as the user message and the task specified by the system prompts below. Stage 1 paraphrases the English sentence; Stage 2 translates the English paraphrase into Serbian Cyrillic and Russian Cyrillic. Serbian Latin is not model-generated, but a deterministic 1:1 transliteration of the Serbian Cyrillic. Table 5 gives the paraphrase generation pass rate for each of the language-script variants, as well as the mean LaBSE score.

Stage 1: English Paraphrase Prompt

> You are a careful paraphrasing assistant. Given an English sentence, generate a paraphrase that:
>
> 1. Preserves the exact meaning of the original
> 2. Uses different vocabulary and sentence structure where natural
> 3. Sounds like natural, fluent English
> 4. Maintains roughly similar length (within 30% of original word count)
> 5. Keeps proper nouns, numbers, and technical terms exactly as they appear in the original
> 6. Preserves grammatical gender markers throughout. If the original sentence implies a male or female subject (through pronouns, verb forms, or adjective agreements), the paraphrase must preserve that gender. Do not change "he" to "she" or vice versa.
> 7. Preserves the strength and scope of claims. Do not amplify or weaken descriptions:
>   - "sometimes" must not become "often" or "always"
>   - "may" or "can" must not become "will" or "must"
>   - "outbreak" must not become "epidemic"
>   - "some" must not become "most" or "all"
>   - Hedged claims must remain hedged
>   - Specific claims must remain specific (do not generalize)
>
>   If the original is uncertain or qualified, the paraphrase must remain qualified to the same degree.
>
> Return ONLY the paraphrased sentence, no preamble, no explanation, no quotes.

Stage 2a: Serbian Cyrillic Translation

> You are a careful translator. Translate the given English sentence into natural, fluent Serbian written in Cyrillic script. Preserve the exact meaning. Use natural Serbian phrasing — do not translate word-for-word from English. Keep proper nouns, numbers, and technical terms in their standard Serbian form (transliterate names if there's a standard Serbian convention, otherwise keep as in English).
>
> For proper nouns, use the standard form of the name in the target language and script. For Serbian Cyrillic, use standard Serbian Cyrillic forms (e.g., Кенеди for Kennedy, Трамп for Trump, Бжежински for Brzezinski). Do not preserve Latin-script spellings within Cyrillic text. However, do not change which person, place, or entity is being referred to: if the original says "Port Stanley," the translation should refer to Port Stanley (in target-language form), not "Stanley" or any other place name. Maintain the identity of named entities; localize their surface form to the target script and language conventions.
>
> Preserve grammatical gender markers throughout. If the original sentence implies a male or female subject (through pronouns, verb forms, or adjective agreements), the translation must preserve that gender. In Serbian, maintain consistent gender across all verb forms, pronouns, and adjective endings.
>
> Preserve verb tense and modality exactly:
>
> - Past tense remains past tense
> - Present tense remains present tense
> - Future tense remains future tense
> - Modal verbs (must, may, can, will, should) must map to equivalent strength in Serbian
> - Conditional remains conditional
> - Aspect distinctions (perfective vs imperfective) should match the original's intent
>
> Return ONLY the translated sentence, no preamble, no explanation, no quotes.

Stage 2b: Russian Cyrillic Translation

> You are a careful translator. Translate the given English sentence into natural, fluent Russian written in Cyrillic script. Preserve the exact meaning. Use natural Russian phrasing — do not translate word-for-word from English. Keep proper nouns, numbers, and technical terms in their standard Russian form (transliterate names if there's a standard Russian convention, otherwise keep as in English).
>
> For proper nouns, use the standard form of the name in the target language and script. For

Russian, use standard Russian Cyrillic transliterations (e.g., Кеннеди for Kennedy, Трамп for Trump). Do not preserve Latin-script spellings within Cyrillic text. However, do not change which person, place, or entity is being referred to: if the original says "Port Stanley," the translation should refer to Port Stanley (in target-language form), not "Stanley" or any other place name. Maintain the identity of named entities; localize their surface form to the target script and language conventions.

Preserve grammatical gender markers throughout. If the original sentence implies a male or female subject (through pronouns, verb forms, or adjective agreements), the translation must preserve that gender. In Russian, maintain consistent gender across all verb forms, pronouns, and adjective endings.

Preserve verb tense and modality exactly:

- Past tense remains past tense
- Present tense remains present tense
- Future tense remains future tense
- Modal verbs (must, may, can, will, should) must map to equivalent strength in Russian
- Conditional remains conditional
- Aspect distinctions (perfective vs imperfective) should match the original's intent

Return ONLY the translated sentence, no preamble, no explanation, no quotes.

Stage	n	1st-try	$\geq$2 att.	LaBSE
English paraphrase	300	203 (67.7%)	97 (32.3%)	0.937
Serbian Cyr. trans.	300	294 (98.0%)	6 (2.0%)	0.919
Russian Cyr. trans.	300	296 (98.7%)	4 (1.3%)	0.927
Serbian Latin	300	—	—	0.919

Table 5: Paraphrase generation pass rates on the dataset. "1st-try" counts candidates accepted on the first attempt; "$\geq$2 att." counts those that required at least one regeneration to clear the LaBSE ≥ 0.80 filter. Serbian Latin is produced by deterministic transliteration of Serbian Cyrillic and inherits its acceptance.

A.2 Validation Instructions

Native-speaker validators received identical instructions for each language, with the target language substituted throughout. Each validator saw sentence pairs, an original sentence and its candidate paraphrase in the same language and script, and rated each pair on two binary axes. Table 6 shows worked examples of the four Meaning and Naturalness combinations.

Meaning. *Yes* if both sentences make the same factual claim, such that a reader of either would understand the same thing; *No* if the paraphrase adds, removes, or changes the claim. Validators were told not to mark *No* merely for differences in wording, word order, or style, since paraphrases are expected to differ on the surface. They were instructed to mark Meaning *No* whenever any of the following occurred: a flipped negation; a meaning-relevant change of tense or aspect; a changed number; a changed or omitted named entity (person, place, organization, or date); a changed modality; a reversed comparison; a swapped subject and object; or any other omission, weakening, or addition that alters the claim.

Naturalness. *Yes* if the paraphrase reads as fluent text a native speaker could plausibly have written; *No* if a native reader would notice something off without being told the text was machine-generated. Validators were told they did not need to name a grammatical rule; a native-speaker judgment that the text sounded wrong was sufficient. Mild shifts in formality were not grounds for *No* unless the register itself sounded unnatural.

Corrections. A pair rated *Yes* on both axes was accepted as is. Otherwise the validator supplied a corrected paraphrase, written in the target script, that preserved the original meaning and read naturally, along with a short note describing the change. If a paraphrase was too broken to judge confidently, validators marked both axes *No* and supplied a correction. Corrected paraphrases enter the dataset flagged separately from accepted model paraphrases, so the two can always be told apart, and no items are discarded.

Compensation. Each validator was paid \$125–\$200 per batch of 200 sentence pairs, recruited through personal networks as native speakers of the target language. Payment was calibrated to exceed standard crowdsourcing rates for comparable linguistic annotation tasks. Before starting, validators were informed that their judgments and corrections would be released as part of a public research dataset, and agreed to participate on that basis.

A.3 Disagreement Resolution

For the 100 sentences in each language judged by both validators, a single decision per item populates the dataset, resolved conservatively. Table 7 gives the per-language breakdown. When neither validator flagged the pair, the model paraphrase was kept. When exactly one validator flagged it, we adopted that validator's correction. When both flagged the same item with differing corrections,

Paraphrase	Meaning	Natural	Why
On Tuesday, the committee approved the budget by vote.	Yes	Yes	Different wording and structure, same meaning, natural English
The committee voted to reject the budget on Tuesday.	No	Yes	Negation flipped, different meaning
On Tuesday, committee voted approval to budget.	Yes	No	Broken word order, meaning still recoverable
On Tuesday, committee voted to reject approval to budget.	No	No	Meaning reversed (approve to reject) and broken grammar

Table 6: Worked examples shown to validators, illustrating the two-axis judgment. Original sentence: *"The committee voted to approve the budget on Tuesday."* Examples were presented in each validator's target language; English is shown here for readability.

Overlap outcome	English	Russian	Serbian
Neither flagged	99	87	63
Exactly one flagged	0	11	23
Both flagged	1	2	14

Table 7: Validator decisions on the 100-sentence overlap subset, by accept-or-flag outcome (rows sum to 100). Items neither validator flagged kept the model paraphrase; when only one validator flagged a pair, we adopted that validator's correction; items both validators flagged used the parity tiebreak only when the two corrections differed.

we kept the correction from one designated validator when the sentence's FLORES+ index was even and from the other when it was odd. Because the index is assigned by the benchmark before validation, this rule was fixed in advance and cannot depend on the content or favor either validator. One-sided disagreements (0, 11, and 23 items for English, Russian, and Serbian) were therefore resolved deterministically, and the parity tiebreak applied only to items both validators flagged with differing corrections, at most 1, 2, and 14 items respectively, each a choice between two validator-approved paraphrases.

A.4 Detailed Validation Results

This section reports the full per-language validation statistics summarized in the main text. Table 8 gives clean-accept and correction rates; the lower Serbian rate reflects rougher generated Serbian, which validators caught and corrected rather than discarded. Table 9 breaks those corrections down by what failed: English corrections were almost all meaning fixes, every Russian correction involved naturalness, and Serbian spanned both. Table 10 reports how large those corrections were, as the word-level edit distance between each model paraphrase and the validator's replacement; the median touches one to three words in every language, and only Serbian shows a tail of larger rewrites, visible as a mean fraction well above its median. Table 11 reports inter-rater agreement on the 100-sentence overlap subset, measuring how often the two validators for a language reached the same accept-or-correct decision on the items they both judged. We rely on raw agreement and Gwet's AC1 rather than Cohen's κ, since the high accept rates produce skewed marginals under which κ is deflated even when raw agreement is high. Agreement is perfect for English, high for Russian, and moderate for Serbian. For English the two validators made the same accept-or-correct decision on all 100 overlap items: the only item that drew a correction was flagged by both, and both supplied the same fix, so raw agreement and AC1 are both 1.0. For Serbian, validators differed mainly on whether a borderline paraphrase warranted a minor fix, not on whether meaning was preserved.

Metric	English	Serbian	Russian
Clean-accept	288 (96.0%)	212 (70.7%)	272 (90.7%)
Correction	12 (4.0%)	88 (29.3%)	28 (9.3%)

Table 8: Per-language outcomes over 300 sentences (counts out of 300). Serbian Cyrillic and Latin share these rates, since Latin inherits Cyrillic corrections via transliteration. The correction counts (12, 88, 28) match the totals in Table 9.

A.5 Length and Tokenization

We report sentence length in words and in token count, both to characterize the conditions and to address the concern that cross-script effects could reflect tokenization rather than representation. Table 12 gives both counts by condition and variant. Word counts are matched across the three conditions: originals, paraphrases, and random partners

Failure Type	English	Serbian	Russian
Meaning Incorrect	11 (91.7%)	57 (64.8%)	0 (0%)
Naturalness Incorrect	1 (8.3%)	14 (15.9%)	18 (64.3%)
Both	0 (0%)	17 (19.3%)	10 (35.7%)
Total	12	88	28

Table 9: Reason for correction, by language, over corrected paraphrases only. The three categories are mutually exclusive: each corrected item is counted once, by whether it failed on meaning, on naturalness, or on both.

Metric	Eng	Srp-Cyrl	Srp-Latn	Rus
Median edit	1.0	2.5	2.5	2.0
Mean edit	2.17	3.25	3.25	3.79
Median frac.	0.077	0.122	0.122	0.131
Mean frac.	0.105	0.433	0.433	0.213

Table 10: Edit magnitude over corrected items only, Edit distances are word-level; "frac." is that distance as a fraction of sentence length. The Serbian mean fraction is inflated by a small tail of larger rewrites, while the median stays low.

average 20.2, 20.9, and 20.2 words respectively, with paraphrases running under one word longer than their originals, well inside the 30% length tolerance. Variants differ modestly in word count (English averages about one word more than Serbian, Russian about one word less), reflecting language-level differences in how the same content is expressed rather than any condition effect.

Subword tokenization, by contrast, differs sharply across scripts. Table 13 reports tokens per word under the Gemma-3-27B tokenizer. English costs about 1.2 tokens per word, whereas Serbian Cyrillic costs about 2.3, so the same roughly 20-word content expands from about 26 tokens in English to 45–48 in Serbian Cyrillic. The effect is partly script and partly language: Serbian Latin, a deterministic transliteration of the identical Serbian Cyrillic content, costs about 2.0 tokens per word, so the Cyrillic script alone adds roughly 15% more tokens for word-for-word identical text. Russian Cyrillic is more efficient than Serbian Cyrillic (about 1.9 versus 2.3).

These large, script-driven token-count differences are intrinsic to the comparison and cannot be removed without distorting content: equalizing token counts across scripts would require unmatching the content itself. Our design therefore controls content length in words and treats subword tokenization as a property of each variant. This raises

Metric	English	Russian	Serbian
Raw agreement	100%	89%	77%
Cohen's κ	n/a	0.21	0.40
Gwet's AC1	1.0	0.87	0.63

Table 11: Inter-rater agreement on the 100-sentence overlap subset.

the question of whether these token-count differences themselves drive the feature overlap patterns we report.

Tokenization and feature overlap. We test this directly. For each of the 300 Serbian sentence pairs (same sentence in Latin and Cyrillic), we compute the token count difference (Cyrillic minus Latin) and correlate it with the cross-script Jaccard similarity averaged across all layers. Token deltas ranged from -3 to 19 (mean 5.4, sd 3.6, IQR $[3, 8]$), confirming that Cyrillic consistently tokenizes into more subwords. The correlation with feature overlap is statistically significant but negligible ($r = -0.172$, $p = 0.003$, $r^2 = 0.03$): tokenization mismatch accounts for roughly 3% of the variance in Jaccard similarity. Active feature set sizes show no relationship with token count differences at all ($r = 0.027$, $p = 0.639$), ruling out a mechanism in which longer tokenizations produce systematically different feature counts. These results indicate that tokenization asymmetry is not a meaningful driver of the cross-script patterns we report. Figure 2 shows the relationship visually.

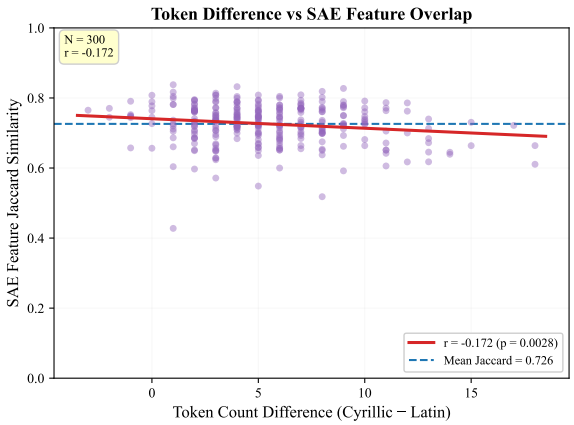


Figure 2: Cross-script Jaccard similarity (averaged across layers) vs. token count difference (Cyrillic minus Latin) for 300 Serbian sentence pairs. The regression line is nearly flat ($r = -0.17$, $r^2 = 0.03$).

Condition	En-Latin	Sr-Cyrillic	Sr-Latin	Ru-Cyrillic	Marginal
Word Count					
Original	21.4 ± 6.7 (20)	20.1 ± 6.6 (19)	20.1 ± 6.6 (19)	19.2 ± 6.5 (18)	20.2 ± 6.6 (19)
Paraphrase	22.3 ± 6.9 (22)	20.8 ± 6.8 (20)	20.8 ± 6.8 (20)	19.8 ± 6.6 (19)	20.9 ± 6.8 (20)
Random	21.3 ± 6.4 (21)	20.1 ± 6.3 (20)	20.1 ± 6.3 (20)	19.3 ± 6.2 (19)	20.2 ± 6.3 (20)
Token Count					
Original	26.3 ± 8.6 (25)	45.0 ± 14.8 (42)	39.5 ± 13.2 (37)	35.8 ± 12.4 (34)	36.7 ± 14.2 (35)
Paraphrase	27.4 ± 8.9 (26)	47.9 ± 15.3 (46)	41.8 ± 13.5 (39)	37.8 ± 12.7 (36)	38.7 ± 14.8 (36)
Random	26.6 ± 8.3 (26)	46.1 ± 14.2 (45)	40.1 ± 12.4 (39)	36.2 ± 12.0 (36)	37.3 ± 13.9 (36)

Table 12: Word and subword token counts by condition and language-script variant, as mean $\pm$ standard deviation with median in parentheses. Word counts are matched across conditions; subword counts differ sharply across scripts. Token counts use the Gemma-3-27B tokenizer.

Condition	En-Lat	Sr-Cyr	Sr-Lat	Ru-Cyr
Original	1.24	2.26	1.98	1.88
Paraphrase	1.23	2.33	2.03	1.93
Random	1.25	2.31	2.01	1.90

Table 13: Tokens per word (Gemma-3-27B). The same content costs roughly twice as many tokens in Serbian Cyrillic as in English, and about 15% more in Serbian Cyrillic than in the deterministically transliterated Serbian Latin.

A.6 Active-Feature Activation Magnitudes

The active set $F_\ell(s)$ is defined by nonzero post-JumpReLU activation. Because JumpReLU emits zero below a learned per-feature threshold θ and a value of at least θ above it, "nonzero" is not a near-zero cutoff but a substantial gate. We verify that active features clear this gate by a comfortable margin rather than skimming it. Table 14 reports activation statistics at layer 31 on Gemma-3-27B (illustrative; layer 31 lies in the mid-network region where the meaning signal is high).

Active features fire well above their threshold: the median active feature activates at $1.67\times$ its gate θ, and only 11.6% of active features sit within 10% of θ. The binary > 0 criterion therefore selects genuinely engaged features rather than threshold noise, supporting its use as the basis for our feature-set analyses.

A.7 Dataset Composition by Source and Topic

Source-domain composition is given in Table 15. Originals and paraphrases preserve the FLORES+ stratification exactly; random partners are sampled within language-script and length-matched. Table 16 gives the finer-grained topic distribution; the dataset spans a broad range of subject matter, with no single topic dominating.

Quantity	Value
Active (feature, text) pairs	80,578
Active features per text (mean)	22.4
Activation magnitude (median)	518.4
Activation magnitude (mean)	714.2
Activation magnitude (p5 / p95)	319.4 / 1,831.9
Threshold θ (min / median / max)	201.2 / 339.2 / 5,197.6
Median margin over θ	$1.67\times$
Active features within 10% of θ	11.6%

Table 14: Active-feature activation magnitudes at layer 31 (Gemma-3-27B). Active features fire well above the JumpReLU threshold, confirming the > 0 active-set criterion captures genuine engagement.

Domain	Orig.	Para.	Rand.
Wikinews	100	100	105
Wikibooks	101	101	102
Wikivoyage	99	99	93

Table 15: Source-domain counts over the 300 anchors. Originals and paraphrases preserve the FLORES+ stratification exactly; random partners are sampled within language-script and length-matched, shifting domain composition slightly.

Topic	N	%	Domain
Travel	91	30.3	Wikivoyage
Sci. & Tech.	79	26.3	Wikibooks
Hist. & Geog.	50	16.7	Wikibooks
Politics & Law	33	11.0	Wikinews
Culture & Ent.	27	9.0	Wikinews
Sports	22	7.3	Wikinews
Educ. & Soc. Sci.	19	6.3	Wikibooks
Disasters	11	3.7	Wikinews
Other	1	0.3	Wikinews

Table 16: Topic distribution of the 300 original sentences, grouped into broad categories from FLORES+ metadata. Percentages exceed 100% because some sentences carry multiple topic tags (e.g., "disease, research, canada").

B Ablation Testing

B.1 SAE Sparsity Level

We compare the small-L_0 SAEs used in the main text with the large-L_0 variant from Gemma Scope 2, both at 16K width (Figure 3). Higher L_0 means more features fire per input, producing larger and noisier active sets. The decomposition hierarchy and layer-wise trajectories are fully preserved, confirming that the results are not sensitive to sparsity level.

B.2 SAE Dictionary Width

We replicate the Leg 1 decomposition on Gemma-3-27B using 262K-width Gemma Scope 2 SAEs, a $16\times$ increase over the primary 16K configuration (Figure 4). The layer-wise trajectories and the ordering across panels are nearly identical to the 16K results, confirming that the factorial decomposition is not an artifact of dictionary size.

B.3 Pooling Strategy

We compare last-token, mean, and max pooling on Gemma-3-27B (Figure 5). The factorial ordering (script $>$ wording $>$ language $>$ meaning) holds under both last-token and mean pooling; only the effect sizes differ. Among last-token, mean, and max pooling, only last-token cleanly separates the conditions from their baselines; max saturates near 1.0 and mean is erratic across layers, so we adopt last-token.

B.4 Model Scale

To verify that our findings are not specific to Gemma-3-27B, we replicate the full Leg 1 factorial decomposition on Gemma-3-1B (25 layers) and Gemma-3-12B (47 layers) using the same dataset, SAE configuration (16k width, low L0), and pipeline. Figure 6 overlays the four panels across all three model sizes on a normalized depth axis.

The core structure holds across scale: in all three models, script overlap exceeds language overlap, meaning rises well above its random baseline, and wording stays roughly flat. The depth trajectories are noisier at 1B, where the model has fewer layers to differentiate, but the qualitative ordering of the four panels is preserved. We conclude that the factorial decomposition reported in the main text is robust to model scale within the Gemma family.

B.5 Cross-Architecture: Llama-3.1-8B

To verify that our findings are not specific to the Gemma architecture or Gemma Scope 2 SAEs, we replicate the Leg 1 decomposition on Llama-3.1-8B-Base using Llama Scope SAEs (He et al., 2024): 32K-width ($8\times$ expansion) post-MLP residual stream SAEs at all 32 layers. These SAEs are trained with TopK activation (k=50) and decoder-norm-weighted selection, then converted to JumpReLU with per-layer thresholds for independent feature activation at inference; we define the active set as features with nonzero post-JumpReLU activation. This setup differs from our primary analysis in model family, SAE training regime, dictionary width (32K vs. 16K), and tokenizer.

The qualitative structure replicates (Figure 7). Script overlap is highest and dips in late layers, meaning peaks in the middle of the network and stays well above its random baseline, and wording remains in the 0.4–0.6 range throughout. Absolute Jaccard values are lower across the board (e.g. script mean 0.48 vs. 0.73 for Gemma-3-27B), but the ordering across panels and the depth trajectories are preserved. The factorial decomposition is therefore not an artifact of the Gemma architecture or Gemma Scope 2 SAEs.

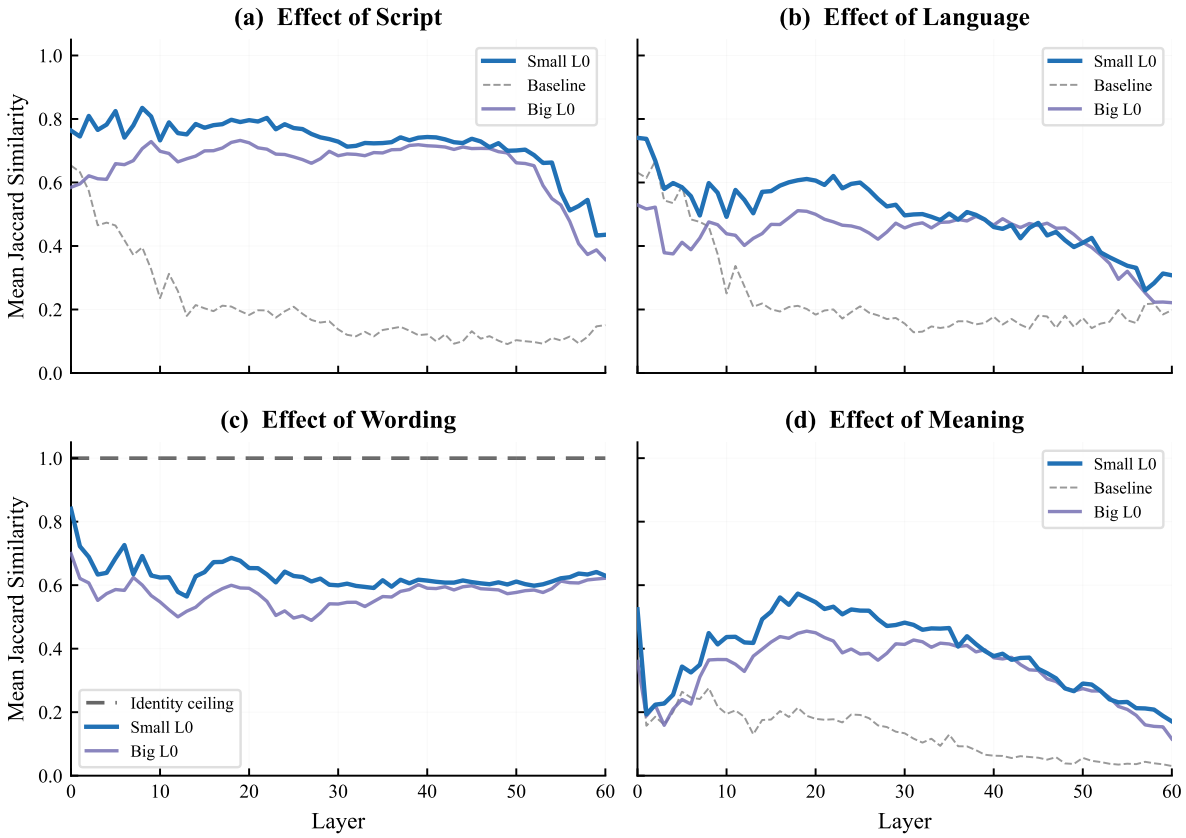


Figure 3: Leg 1 decomposition comparing small-L_0 (blue) and large-L_0 (purple) Gemma Scope 2 SAEs at 16K width on Gemma-3-27B. Dashed grey lines show baselines. The factorial ordering and depth structure are preserved.

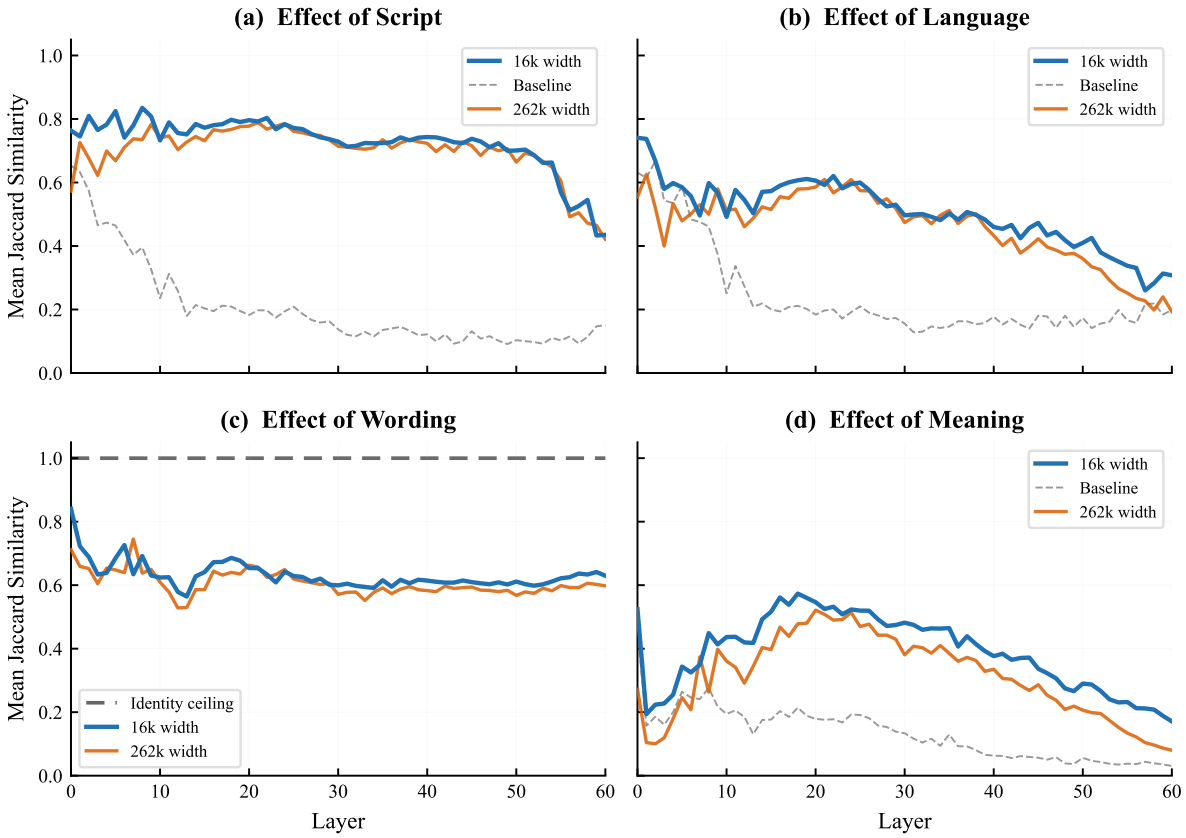


Figure 4: Leg 1 decomposition comparing 16K-width (blue) and 262K-width (orange) Gemma Scope 2 SAEs on Gemma-3-27B. Dashed grey lines show baselines. The layer-wise shape and panel ordering are preserved; the uniform downward shift reflects sparser feature activation in the wider dictionary.

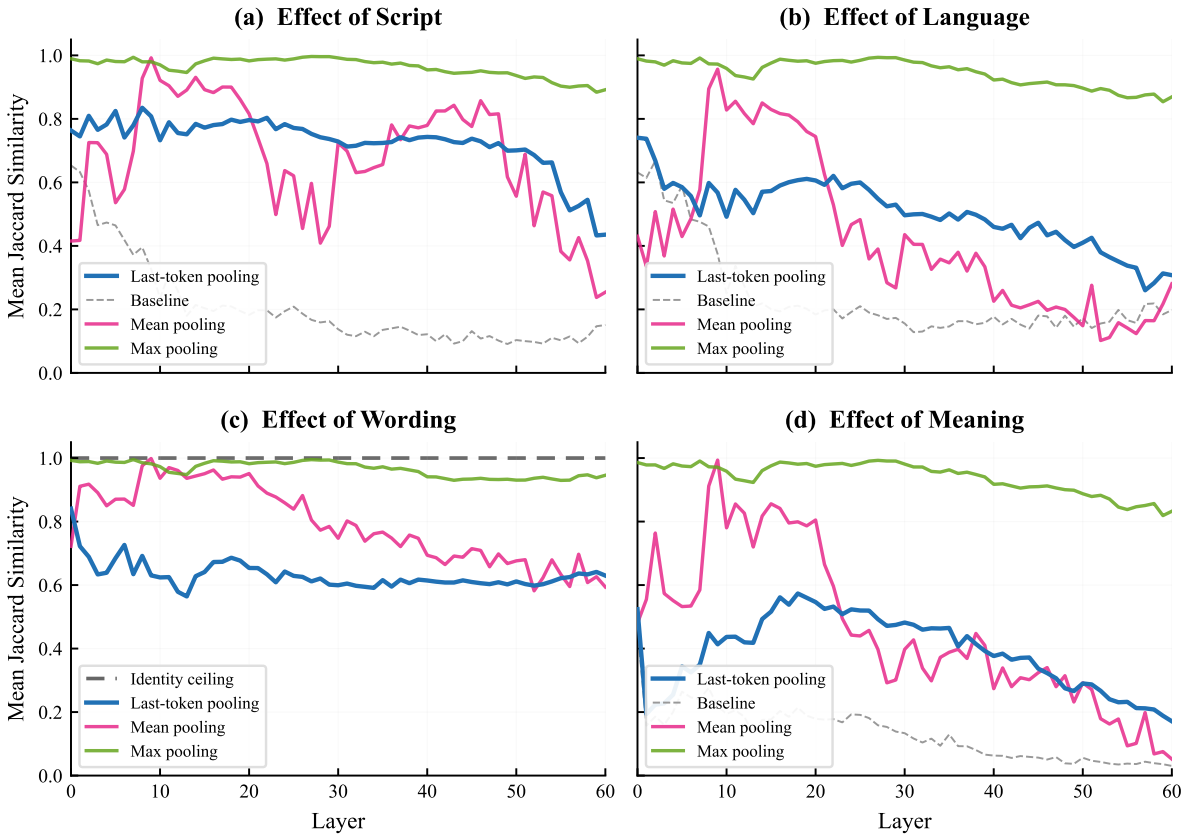


Figure 5: Pooling strategy comparison on Gemma-3-27B. Dashed grey lines show the corresponding baselines.

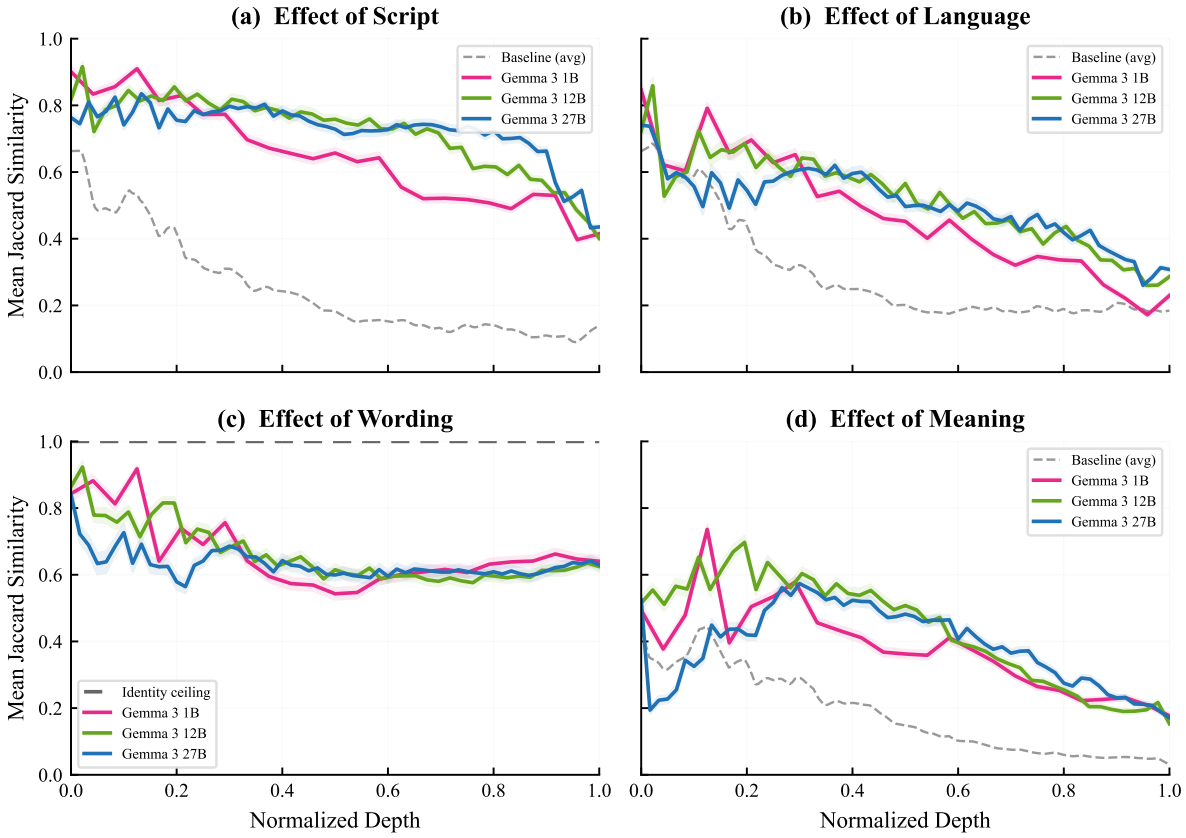


Figure 6: Leg 1 decomposition across Gemma-3 model sizes (1B, 12B, 27B), plotted on normalized depth (300 sentences, 95% bootstrap CIs). Solid lines show the main comparison for each panel; thin grey dashed lines show the corresponding random-partner baseline averaged across all three model sizes (per-model baselines are nearly identical). Panel (c) retains the identity ceiling at 1.0. The qualitative structure holds at all three scales: script overlap remains highest, meaning stays well above its baseline, and wording is roughly flat.

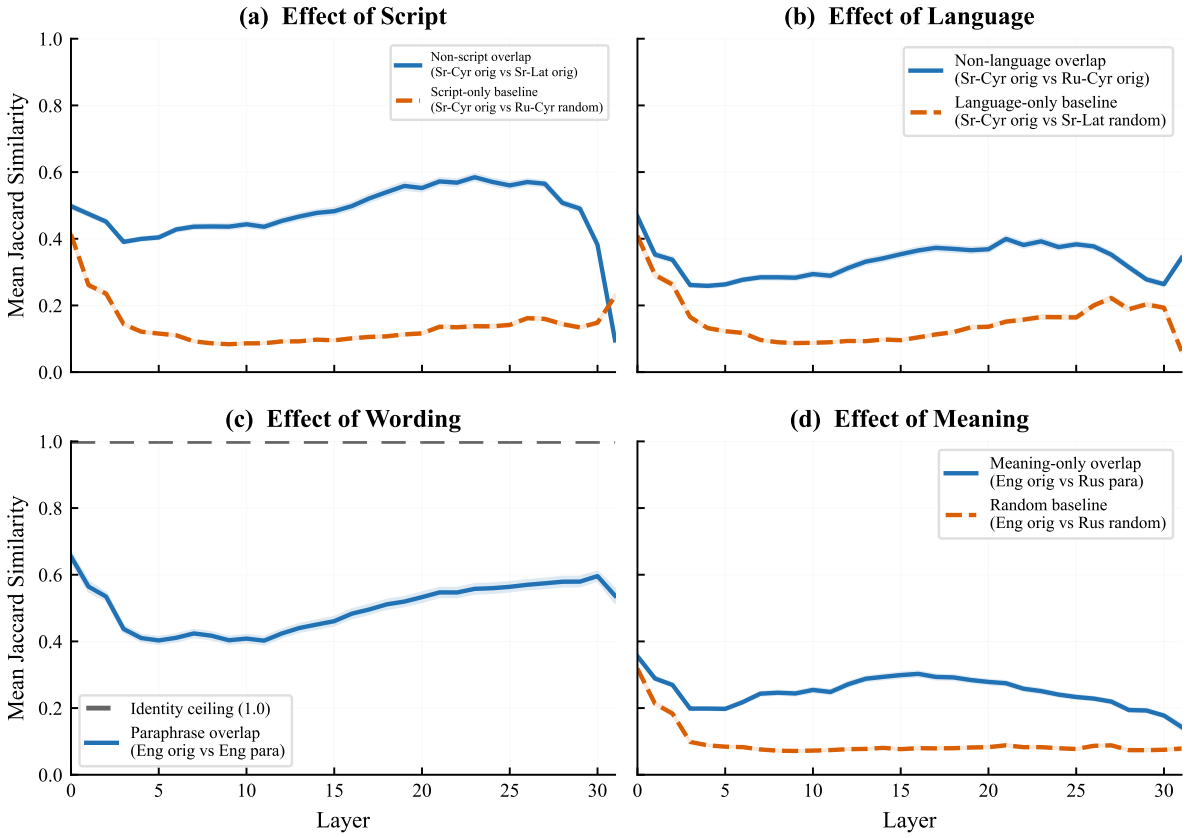


Figure 7: Leg 1 decomposition for Llama-3.1-8B-Base with Llama Scope SAEs (32K width, 32 layers, 300 sentences). Solid lines show the main comparison; dashed lines show random-partner baselines. Despite a different model family, SAE training regime (TopK-trained, JumpReLU-converted), and dictionary width (32K vs. 16K), the qualitative structure matches Gemma-3-27B (Figure 1): script overlap highest, meaning above baseline peaking mid-network, wording roughly flat.

C Robustness-Check Classifier

This appendix details the secondary content classifier and its per-layer discard rates.

Classifier Configs. Every retrieved auto-interp label is passed through a secondary classifier built on `claude-sonnet-4-6` (`max_tokens`=2000). Each feature is assigned to exactly one of five categories: *content-claim*, *surface-claim*, *language-claim*, *mixed*, or *other*; any response outside this set is coerced to *other*. Only *content-claim* features propagate into the downstream falsification, miss-rate, asymmetry, and typology analyses. The exact prompt is shown below.

Classifier Prompt.

```
Classify each SAE feature auto-interp label
into exactly one category:
content-claim: semantic content, topic,
  or meaning
surface-claim: orthography, script,
  formatting, or specific tokens
language-claim: a particular language
  or family
mixed: content + surface/language
other: doesn't fit the above
Labels (feature_id: label): {label_block}
Respond with ONLY one line per label:
  feature_id: category.
```

Discard Rates. Tables 17 and 19 report the per-layer outcome of the content-claim gate on Gemma 3 27B. Sonnet discards 1,744 of 5,238 labels (33.3%); Gemini discards 1,401 of 5,211 (26.9%). Both attribute the bulk of the drop to *surface-claim* labels (Tables 18 and 20). The comparison in Table 21 shows Sonnet is stricter at every layer, with the gap widening sharply at layer 53 (45.2% vs. 29.7%). Despite these differing discard rates, both labelers converge on the same Cyrillic/Latin asymmetry at every layer (Table 28), confirming the asymmetry is a property of the features rather than of either classifier's strictness.

Layer	Labels	Content	Disc.	% Disc.
16	852	596	256	30.0
31	1,213	896	317	26.1
40	1,447	1,056	391	27.0
53	1,726	946	780	45.2
Total	**5,238**	**3,494**	**1,744**	**33.3**

Table 17: Sonnet classifier discards on Gemma 3 27B (leg 1) across the four Neuronpedia SAE layers. "Disc." counts any label not tagged *content-claim*.

Category	Count
surface-claim	1,413
mixed	172
language-claim	79
other	78
unknown (parse fail)	2
Total non-content	**1,744**

Table 18: Composition of the 1,744 Sonnet-classifier discards on Gemma 3 27B, summed across all four layers.

Layer	Labels	Content	Disc.	% Disc.
16	840	592	248	29.5
31	1,219	883	336	27.6
40	1,436	1,128	308	21.4
53	1,716	1,207	509	29.7
Total	**5,211**	**3,810**	**1,401**	**26.9**

Table 19: Gemini classifier discards on Gemma 3 27B (leg 1) across the four Neuronpedia SAE layers, using the same content-claim gate.

Category	Count
surface-claim	968
mixed	218
other	154
language-claim	53
unknown (parse fail)	8
Total non-content	**1,401**

Table 20: Composition of the 1,401 Gemini-classifier discards on Gemma 3 27B, summed across all four layers.

Layer	Sonnet %	Gemini %
16	30.0	29.5
31	26.1	27.6
40	27.0	21.4
53	45.2	29.7
Overall	**33.3**	**26.9**

Table 21: Cross-classifier comparison of discard rates on Gemma 3 27B. Sonnet is the stricter judge at every layer, with the gap widening sharply at layer 53.

D Detailed Miss-Rate Breakdowns

Tables 22 and 23 report the full miss-rate breakdown for the cross-lingual content pool under Sonnet and Gemini labels, respectively. Each table includes within-language paraphrase miss rates (Eng-orig, Rus-para), cross-language Serbian miss rates for both originals and paraphrases, and random baselines confirming content-selectivity (77–96% miss on controlled-unrelated sentences). Table 24 compares the two labelers side by side: despite classifying different subsets as content-claim, both converge on the same Cyrillic/Latin asymmetry ratio at every layer (e.g., 1.35× vs. 1.36× at layer 53). Table 25 provides 20 additional feature vignettes beyond the three in the main text, spanning all four SAE layers and both failure directions, with full English sentences and dual-labeler agreement. Tables 26 and 27 further show that the Cyrillic disadvantage holds within every FLORES+ topic bucket under both primary- and multi-tag assignment. Finally, to confirm the content classifier itself does not drive the asymmetry, Table 28 reports miss-rate ratios computed on the *unfiltered* pool, with no content gate applied. The content filter exists to restrict the analysis to features making genuine content claims, excluding surface- and language-claims (e.g., 'Cyrillic characters', 'Russian morphology') that would differ across scripts by construction and inflate the asymmetry independently of any content-generalization effect. That the asymmetry nonetheless holds in the unfiltered pool, with ratios overlapping the content-filtered version at every layer (excluding 1.0 from layer 31 onward), shows the filter sharpens the interpretation of the result rather than producing the cross-script gap.

D.1 Reproducibility

Auto-interp labels were generated by Claude Sonnet 4.5 and by Gemini 2.5 Flash Lite (gemini-2.5-flash-lite, Neuronpedia's default labeler). Neuronpedia itself exposes no API-level version, snapshot, or per-explanation timestamp, and served labels are mutable. All Neuronpedia GETs were issued between 2026-05-24 and 2026-05-25 (UTC-local), during which the deployed Neuronpedia codebase was release v1.0.755. To enable byte-exact replication independent of future label regenerations or model deprecations, we release the cached Neuronpedia and Sonnet JSON bundles alongside our code; users should verify against these snapshots rather than live Neuronpedia.

D.2 Human Verification of Missed Features

We manually inspected a stratified sample of 60 missed (feature, sentence) units from the Gemma-3-27B Sonnet content pool: 15 per layer (16, 31, 40, 53), split evenly between Serbian Cyrillic and Serbian Latin misses, with one unit per feature. For each unit, the author judged (Q1) whether the Sonnet label describes the feature's top-activating examples on Neuronpedia, and (Q2) whether the labeled concept is present in the sentence the feature missed. Each question was answered Yes, Partly, or No. Annotation was done by a single annotator with the layer and missed script visible, so we report no inter-annotator agreement. Table 30 reports the results.

The labels are locally accurate: 92% describe their top-activating examples at least partly. The labeled concept, however, is present in the missed sentence in only 58% of units (42% strictly). In the remainder, the feature fires on the English and Russian versions of a sentence that does not contain its labeled concept, indicating that the Neuronpedia label captures only part of what drives the feature. Both cases point to the same limitation of auto-interpretation labels: a label accurate on a feature's top-activating examples does not reliably describe the feature's behavior on other inputs.

	Within-language (%)		Serbian orig (%)	
Layer	Eng-para	Rus-orig	Sr-Latin	Sr-Cyrillic
16	14.8 [12.5, 18.4]	16.6 [12.5, 22.3]	20.8 [15.9, 27.4]	21.3 [15.8, 28.8]
31	9.2 [7.0, 11.6]	13.7 [11.4, 16.6]	20.4 [18.1, 23.4]	22.6 [20.3, 25.1]
40	9.1 [7.7, 10.5]	15.6 [14.0, 17.3]	22.2 [19.9, 24.6]	25.9 [23.4, 28.6]
53	7.3 [5.8, 9.0]	15.4 [13.4, 17.5]	25.6 [22.1, 29.6]	34.6 [30.7, 38.6]

	Serbian para (%)		Cyr/Lat ratio	
Layer	Sr-Latin	Sr-Cyrillic	Orig	Para
16	18.2 [13.9, 24.3]	19.1 [13.8, 26.4]	1.02×	1.05×
31	18.1 [15.9, 20.8]	21.8 [18.9, 24.7]	1.11×	1.21×
40	20.7 [18.7, 22.9]	25.2 [22.6, 27.9]	1.17×	1.22×
53	24.8 [20.6, 29.2]	33.6 [29.4, 38.1]	1.35×	1.36×

	Random baselines (%)			
Layer	Eng	Rus	Sr-Latin	Sr-Cyrillic
16	78.3 [67.7, 92.2]	76.8 [65.9, 91.9]	77.4 [65.6, 92.9]	77.6 [66.1, 92.7]
31	81.3 [66.1, 96.2]	82.1 [67.7, 96.2]	84.0 [71.5, 97.2]	85.2 [73.9, 97.2]
40	94.8 [90.7, 98.0]	96.0 [94.1, 97.8]	95.8 [92.7, 98.3]	95.5 [92.2, 98.1]
53	94.5 [91.0, 97.9]	95.1 [92.1, 98.0]	95.3 [91.7, 98.6]	96.3 [93.7, 98.7]

Table 22: Full Sonnet content-label miss rates over the cross-lingual content pool (Eng-orig ∩ Rus-para). Content N: L16=1,072, L31=2,236, L40=2,143, L53=1,392. Random baselines confirm content selectivity (features rarely fire on controlled-unrelated sentences). 95% bootstrap CIs (10,000 resamples), feature-clustered.

	Within-language (%)		Serbian orig (%)	
Layer	Eng-para	Rus-orig	Sr-Latin	Sr-Cyrillic
16	13.6 [11.5, 16.4]	15.0 [12.3, 18.5]	19.2 [15.6, 23.7]	19.8 [16.0, 24.7]
31	9.5 [7.6, 11.4]	12.8 [10.2, 15.7]	20.7 [17.9, 24.1]	22.8 [19.9, 26.2]
40	8.3 [6.5, 10.3]	15.0 [11.9, 17.9]	20.2 [16.5, 24.0]	24.3 [20.3, 28.5]
53	7.9 [6.2, 10.0]	12.9 [10.2, 15.9]	22.3 [18.9, 26.3]	30.3 [26.3, 34.8]

	Serbian para (%)		Cyr/Lat ratio	
Layer	Sr-Latin	Sr-Cyrillic	Orig	Para
16	17.8 [14.1, 22.3]	19.2 [15.3, 24.2]	1.03×	1.08×
31	18.1 [15.5, 21.3]	21.3 [18.5, 24.8]	1.10×	1.18×
40	19.0 [15.5, 22.6]	23.4 [19.8, 27.4]	1.21×	1.23×
53	23.7 [20.4, 27.7]	31.0 [27.6, 35.1]	1.36×	1.31×

	Random baselines (%)			
Layer	Eng	Rus	Sr-Latin	Sr-Cyrillic
16	81.6 [77.3, 87.7]	80.6 [75.7, 87.1]	81.5 [75.9, 88.3]	80.6 [75.0, 87.6]
31	82.6 [69.6, 94.0]	82.5 [68.9, 94.1]	84.2 [72.6, 94.6]	84.8 [73.6, 94.8]
40	84.3 [69.6, 95.9]	85.0 [70.3, 96.0]	84.5 [69.8, 96.1]	83.9 [68.4, 95.9]
53	82.4 [67.4, 95.2]	84.1 [70.0, 96.1]	85.6 [72.9, 96.8]	87.1 [75.4, 97.0]

Table 23: Full Gemini content-label miss rates over the cross-lingual content pool (Eng-orig ∩ Rus-para). Content N: L16=1,503, L31=2,444, L40=2,744, L53=2,067. Gemini classifies more features as content-claim than Sonnet, yielding larger N; the asymmetry pattern is consistent across labelers. 95% bootstrap CIs (10,000 resamples), feature-clustered.

Layer	N	Sr-Lat	Sr-Cyr	Ratio
			Sonnet	
16	1,072	20.8 [15.9, 27.4]	21.3 [15.8, 28.8]	1.02× [0.91, 1.12]
31	2,236	20.4 [18.1, 23.4]	22.6 [20.3, 25.1]	1.11× [1.01, 1.19]
40	2,143	22.2 [19.9, 24.6]	25.9 [23.4, 28.6]	1.17× [1.09, 1.26]
53	1,392	25.6 [22.1, 29.6]	34.6 [30.7, 38.6]	1.35× [1.23, 1.49]
			Gemini	
16	1,503	19.2 [15.6, 23.7]	19.8 [16.0, 24.7]	1.03× [0.94, 1.11]
31	2,444	20.7 [17.9, 24.1]	22.8 [19.9, 26.2]	1.10× [1.03, 1.18]
40	2,744	20.2 [16.5, 24.0]	24.3 [20.3, 28.5]	1.21× [1.13, 1.29]
53	2,067	22.3 [18.9, 26.3]	30.3 [26.3, 34.8]	1.36× [1.25, 1.47]

Table 24: Sonnet vs. Gemini labeler comparison over the cross-lingual content pool. Content N is the number of (feature, sentence) pairs classified as content-claim. This count differs between labelers because Sonnet and Gemini describe features differently, producing distinct content-claim subsets from the same underlying activation pool; the converging ratios confirm the signal is robust to labeler choice. Brackets give 95% bootstrap CIs ($10{,}000$ resamples, feature-clustered).

Misses Sr-Cyrillic (fires English, Russian, and Sr-Latin)

L16 / F686
Auto-interp: Sonnet: "historical events" | Gemini: "negative historical events"
Sentence: "The photographer was transported to Ronald Reagan UCLA Medical Center, where he subsequently died."

L16 / F2851
Auto-interp: Sonnet: "damage" | Gemini: "damage, accidents, and injuries"
Sentence: "Water is spilling over the levee in a section 100 feet wide."

L31 / F11478
Auto-interp: Sonnet: "emergency preparedness" | Gemini: "emergency preparedness and response"
Sentence: "Negotiators tried to rectify the situation, but the prisoners' demands are not clear."

L31 / F1773
Auto-interp: Sonnet: "political" | Gemini: "political figures and roles"
Sentence: "Throughout 1960s, Brzezinski worked for John F. Kennedy as his advisor and then the Lyndon B. Johnson administration."

L40 / F7069
Auto-interp: Sonnet: "musical instruments" | Gemini: "musical instruments"
Sentence: "Up means you should start at the tip and push the bow, and down means you should start at the frog (which is where your hand is holding the bow) and pull the bow."

L40 / F14448
Auto-interp: Sonnet: "border" | Gemini: "border"
Sentence: "The Oyapock River Bridge is a cable-stayed bridge. It spans the Oyapock River to link the cities of Oiapoque in Brazil and Saint-Georges de l'Oyapock in French Guiana."

L53 / F12994
Auto-interp: Sonnet: "policy and government" | Gemini: "policy and government institutions"
Sentence: "The money could go toward flood-proof houses, better water management, and crop diversification."

L53 / F7923
Auto-interp: Sonnet: "Muslim" | Gemini: "Muslim discourse and related concepts"
Sentence: "Muhammad was deeply interested in matters beyond this mundane life. He used to frequent a cave that became known as 'Hira' on the Mountain of 'Noor' (light) for contemplation."

L16 / F12470
Auto-interp: Sonnet: "airport" | Gemini: "airport security"
Sentence: "Today, Timbuktu is an impoverished town, although its reputation makes it a tourist attraction, and it has an airport."

L31 / F2733
Auto-interp: Sonnet: "corporate business" | Gemini: "corporate strategies and business news"
Sentence: "Last week, METI announced that Apple had informed it of 34 additional overheating incidents, which the company called 'non-serious.' "

L31 / F8176
Auto-interp: Sonnet: "tourism" | Gemini: "travel and tourism"
Sentence: "For example visiting castles in the Loire Valley, the Rhine valley or taking a cruise to interesting cites on the Danube or boating along the Erie Canal."

L40 / F4378
Auto-interp: Sonnet: "disaster recovery" | Gemini: "disaster, damage, and recovery"
Sentence: "The Ninth Ward, which saw flooding as high as 20 feet during Hurricane Katrina, is currently in waist-high water as the nearby levee was overtopped."

L40 / F4430
Auto-interp: Sonnet: "mice" | Gemini: "animals in studies"
Sentence: " 'We now have 4-month-old mice that are non-diabetic that used to be diabetic,' he added."

L53 / F12525
Auto-interp: Sonnet: "male names" | Gemini: "male names"
Sentence: "Edgar Veguilla received arm and jaw wounds while Kristoffer Schneider was left requiring reconstructive surgery for his face."

L53 / F3443
Auto-interp: Sonnet: "travel" | Gemini: "travel and movement"
Sentence: "If the country you will be visiting becomes subject to a travel advisory, your travel health insurance or your trip cancellation insurance may be affected."

Table 25: Extended examples of content-labeled features whose auto-interp labels fail on one Serbian script. All verifiable on Neuronpedia (`gemma-3-27b`, SAE `{layer}-gemmascope-2-res-16k`).

Table 25 (continued).

Misses Sr-Latin (fires English, Russian, and Sr-Cyrillic)

L16 / F2475
Auto-interp: Sonnet: "Banking and financial institutions" | Gemini: "banking and financial institutions"
Sentence: "Money can be exchanged at the only bank in the islands which is located in Stanley across from the FIC West store."

L40 / F2014
Auto-interp: Sonnet: "ceremony" | Gemini: "ceremony"
Sentence: "Several large television screens were installed in various places in Rome to let the people watch the ceremony."

L53 / F4334
Auto-interp: Sonnet: "Government and governmental institutions" | Gemini: "government"
Sentence: "The Articles required unanimous consent from all the states before they could be amended and states took the central government so lightly that their representatives were often absent."

L31 / F1938
Auto-interp: Sonnet: "news and politics" | Gemini: "news and political commentary"
Sentence: "Suspected cases of H5N1 in Croatia and Denmark remain unconfirmed."

L53 / F1527
Auto-interp: Sonnet: "access to resources" | Gemini: "adequate access to resources"
Sentence: "The cities of Harappa and Mohenjo-daro had a flush toilet in almost every house, attached to a sophisticated sewage system."

Topic	N	Sr-Latin (%)	Sr-Cyrillic (%)	Cyr/Lat ratio
Travel & Tourism	1,998	23.1 [20.8, 25.6]	27.3 [24.8, 30.0]	1.18× [1.11, 1.27]
Science & Technology	1,746	20.8 [18.5, 23.4]	23.8 [21.3, 26.5]	1.14× [1.04, 1.25]
History & Geography	672	19.8 [16.5, 23.4]	24.3 [20.6, 28.2]	1.23× [1.09, 1.38]
Politics & Law	859	21.2 [17.9, 24.8]	25.7 [22.1, 29.7]	1.21× [1.11, 1.33]
Culture & Entertainment	478	26.8 [22.3, 31.7]	30.8 [26.2, 35.7]	1.15× [1.04, 1.28]
Other (small, merged)	1,090	22.4 [19.3, 25.7]	25.5 [22.4, 28.8]	1.14× [1.04, 1.25]

Table 26: Cyrillic/Latin miss-rate asymmetry decomposed by FLORES+ topic on the 27B Sonnet content pool. Each sentence is assigned to its first topic tag. N counts (feature, sentence) units pooled across layers 16, 31, 40, and 53. 95% feature-clustered bootstrap CIs in brackets (10,000 resamples; paired draws for ratios). The Cyrillic disadvantage holds within every topic bucket — every ratio CI excludes 1.0 — including entity-light topics (Science & Technology, Travel & Tourism), indicating the effect is not driven by named entities. The merged Other row aggregates Sports, Education & Social Science, Disasters & Accidents, and Other, which are individually too small for stable per-layer bootstraps.

Topic	N	Sr-Latin (%)	Sr-Cyrillic (%)	Cyr/Lat ratio
Travel & Tourism	2,116	22.7 [20.5, 25.2]	27.2 [24.7, 29.9]	1.20× [1.12, 1.28]
Science & Technology	1,790	20.9 [18.5, 23.4]	23.8 [21.2, 26.5]	1.14× [1.04, 1.24]
History & Geography	1,081	19.1 [16.4, 21.9]	23.1 [20.1, 26.3]	1.21× [1.11, 1.33]
Politics & Law	859	21.2 [17.9, 24.6]	25.7 [22.1, 29.5]	1.21× [1.11, 1.33]
Culture & Entertainment	572	26.4 [22.4, 30.7]	29.5 [25.4, 33.8]	1.12× [1.02, 1.24]
Other (small, merged)	1,197	22.0 [19.1, 25.2]	25.8 [22.8, 29.0]	1.17× [1.07, 1.29]

Table 27: Cyrillic/Latin miss-rate asymmetry decomposed by FLORES+ topic on the 27B Sonnet content pool, Multi-Tag mode: each sentence contributes to *every* topic bucket its tags map to (so multi-tagged sentences appear in more than one row). N counts (feature, sentence) units pooled across layers 16, 31, 40, and 53. 95% feature-clustered bootstrap CIs in brackets (10,000 resamples; paired draws for ratios). The Cyrillic disadvantage holds within every topic — every ratio CI excludes 1.0 — confirming that the asymmetry is robust to how multi-tagged sentences are assigned to buckets.

Layer	View	Feats	Pairs	Miss Sr-Lat (95% CI)	Miss Sr-Cyr (95% CI)	Ratio (95% CI)
16	unfiltered	437	4,777	14.42% [11.84, 17.36]	14.78% [12.09, 18.05]	1.02× [0.96, 1.08]
16	Sonnet-content	287	1,072	20.80% [15.92, 27.27]	21.27% [15.77, 28.89]	1.02× [0.91, 1.12]
31	unfiltered	867	4,422	20.53% [17.91, 23.64]	22.89% [19.86, 26.15]	1.11× [1.05, 1.18]
31	Sonnet-content	648	2,236	20.44% [18.12, 23.36]	22.58% [20.35, 25.08]	1.11× [1.01, 1.20]
40	unfiltered	1,068	4,182	21.38% [18.44, 24.65]	26.37% [22.86, 29.99]	1.23× [1.16, 1.31]
40	Sonnet-content	796	2,143	22.17% [19.92, 24.76]	25.94% [23.32, 28.64]	1.17× [1.09, 1.26]
53	unfiltered	923	2,880	24.34% [21.19, 27.81]	31.22% [27.93, 34.85]	1.28× [1.20, 1.37]
53	Sonnet-content	555	1,392	25.65% [22.11, 29.54]	34.55% [30.76, 38.55]	1.35× [1.23, 1.49]

Table 28: Gemma 3 27B Serbian Cyrillic miss-rate ratio with feature-clustered bootstrap CIs (10,000 resamples, seed 42). For each layer we report two nested views of the same feature pool. **Unfiltered** contains every (feature, sentence) pair whose feature fires on both the English original and the Russian paraphrase, with no label-based filtering. **Sonnet-content** is the strict subset of that pool whose features were classified as *content-claim* by the Sonnet labeler (e.g., at layer 16, 287 of the 437 unfiltered features); this is the pool used in the main analyses throughout the paper. Miss rate = fraction of pool pairs where the feature does *not* fire on the named condition; Ratio = Sr-Cyr miss / Sr-Lat miss. The two views give overlapping ratios at every layer and Layer 16 is statistically indistinguishable from parity; layers 31, 40, and 53 each show a significant Sr-Cyr penalty.

Condition	Accepted (%)	Corrected (%)	Diff. (pp)
Eng-para†	9.6	10.3	-0.7 [-4.6, $+2.7$]
Sr-Lat-para	19.4	22.3	-2.8 [-4.9, -0.9]
Sr-Cyr-para	23.9	27.2	-3.4 [-5.6, -1.2]
Rus-orig† (anchor)	15.3	13.1	$+2.3$ [-0.5, $+4.9$]

Table 29: Miss rates split by validation outcome: sentences whose paraphrase was accepted as-is vs. flagged and corrected (Gemma-3-27B, Sonnet content pool, pooled over layers 16, 31, 40, 53). Diff. = accepted $-$ corrected, with 95% feature-clustered bootstrap CIs (10,000 resamples). Corrected sentences: English 12, Serbian 88, Russian 28; †fewer than 30 corrected sentences. The Russian paraphrase is a selection anchor, so the Russian original is split by the paraphrase's correction status. Sentence-clustered CIs for Serbian include zero (Sr-Cyr [-7.8, $+0.7$], Sr-Lat [-6.9, $+1.0$]).

		Q1: label accurate		Q2: concept present	
Layer	n	Yes	Yes+Partly	Yes	Yes+Partly
16	15	93%	100%	40%	60%
31	15	80%	80%	40%	53%
40	15	80%	93%	53%	67%
53	15	80%	93%	33%	53%
All	60	83% [72, 91]	92% [82, 96]	42% [30, 54]	58% [46, 70]

Table 30: Human verification of 60 sampled missed units. Q1: the Sonnet label describes the feature's top-activating examples. Q2: the labeled concept is present in the sentence the feature missed. Brackets give Wilson 95% CIs; per-layer differences are not significant.

Layer	Eng-para (%)	Rus-orig (%)	Serbian avg (%)
12	15.5 [9.9, 22.2]	19.0 [13.6, 25.5]	23.2 [17.5, 30.0]
24	11.7 [8.4, 17.0]	13.9 [9.9, 20.8]	20.4 [16.1, 28.1]
31	8.5 [7.0, 10.2]	15.2 [13.1, 17.5]	26.0 [23.0, 29.2]
41	10.0 [7.1, 13.5]	13.6 [10.3, 17.6]	32.6 [30.0, 36.1]

(a) Within- vs. cross-language miss rates.

Layer	Sr-Latin (%)	Sr-Cyrillic (%)	Cyr/Lat
12	22.6 [16.6, 29.7]	23.8 [17.6, 31.1]	1.05× [0.86, 1.30]
24	20.0 [15.6, 27.6]	20.9 [16.4, 28.9]	1.04× [0.96, 1.14]
31	24.1 [21.0, 27.5]	27.8 [24.7, 31.3]	1.16× [1.07, 1.25]
41	27.9 [25.3, 30.6]	37.3 [33.4, 42.3]	1.33× [1.20, 1.50]

(b) Serbian script asymmetry.

Table 31: Content-feature miss rates for Gemma-3-12B (Gemini labels). 95% bootstrap CIs in brackets (10,000 resamples), feature-clustered.

Layer	Eng-para (%)	Rus-orig (%)	Serbian avg (%)
7	9.0 [5.0, 15.8]	8.5 [4.2, 18.6]	13.8 [8.3, 27.3]
13	15.2 [12.8, 17.8]	21.5 [18.6, 24.8]	35.6 [31.3, 40.2]
17	7.1 [4.4, 11.1]	13.4 [9.5, 18.6]	36.8 [25.1, 52.0]
22	5.9 [3.8, 8.4]	14.1 [11.8, 17.6]	53.8 [42.8, 67.3]

(a) Within- vs. cross-language miss rates.

Layer	Sr-Latin (%)	Sr-Cyrillic (%)	Cyr/Lat
7	13.8 [8.6, 26.0]	13.8 [8.0, 28.7]	1.00× [0.85, 1.15]
13	35.1 [31.0, 39.5]	36.1 [31.2, 41.4]	1.03× [0.95, 1.11]
17	34.5 [23.7, 48.9]	39.0 [26.5, 55.3]	1.13× [1.04, 1.23]
22	49.4 [40.0, 62.6]	58.2 [44.2, 73.9]	1.18× [1.00, 1.42]

(b) Serbian script asymmetry.

Table 32: Content-feature miss rates for Gemma-3-1B (Gemini labels). 95% bootstrap CIs in brackets (10,000 resamples), feature-clustered.

E Leg 2 Across Model Scale

We replicate the Leg 2 auto-interpretation evaluation on Gemma-3-1B and Gemma-3-12B. The core pattern holds at both scales (Tables 31 and 32): Serbian miss rates exceed within-language floors, the Cyrillic/Latin asymmetry grows with depth (reaching $1.18\times$ at 1B and $1.33\times$ at 12B), and miss rates are ordered by estimated training coverage.