ONE LANGUAGE, TWO SCRIPTS: PROBING SCRIPT-INVARIANCE IN LLM CONCEPT REPRESENTATIONS

Sripad Karne
Columbia University

ABSTRACT

Do the features learned by Sparse Autoencoders (SAEs) represent abstract meaning, or are they tied to how text is written? We investigate this question using Serbian digraphia as a controlled testbed: Serbian is written interchangeably in Latin and Cyrillic scripts with a near-perfect character mapping between them, enabling us to vary orthography while holding meaning exactly constant. Crucially, these scripts are tokenized completely differently, sharing no tokens whatsoever. Analyzing SAE feature activations across the Gemma model family (270M–27B parameters), we find that identical sentences in different Serbian scripts activate highly overlapping features, far exceeding random baselines. Strikingly, changing script causes less representational divergence than paraphrasing within the same script, suggesting SAE features prioritize meaning over orthographic form. Cross-script cross-paraphrase comparisons provide evidence against memorization, as these combinations rarely co-occur in training data yet still exhibit substantial feature overlap. This script invariance strengthens with model scale. Taken together, our findings suggest that SAE features can capture semantics at a level of abstraction above surface tokenization, and we propose Serbian digraphia as a general evaluation paradigm for probing the abstractness of learned representations.

1 INTRODUCTION

Large language models have transformed how people across the globe access and process information. Users from diverse linguistic backgrounds now interact with these systems daily, raising fundamental questions about how LLMs represent meaning across different languages and writing systems. Sparse Autoencoders (SAEs) offer a lens into this: by decomposing neural network activations into sparse, interpretable features (Bricken et al., 2023; Cunningham et al., 2024), SAEs allow us to examine whether models encode meaning abstractly or remain tied to script-specific token patterns.

To explore this, we study Serbian, which is one of the few languages with active digraphia. This means it is written in two scripts: Latin and Cyrillic. Native speakers use both scripts interchangeably, and a deterministic mapping allows lossless conversion between them with zero semantic change. However, the two scripts are tokenized completely differently by LLMs, making Serbian an ideal testbed for our investigation. If SAE features capture abstract semantics, we would expect the same phrases in Serbian Latin and Serbian Cyrillic to activate similar features despite their divergent tokenization.

We systematically evaluate this hypothesis across the Gemma model family (Gemma Team, 2024), spanning 270M to 27B parameters, using Gemma Scope 2 SAEs (Google DeepMind, 2025). Our experimental design compares SAE feature activations across carefully constructed comparison types, including within-script semantic comparisons (original vs. paraphrase), cross-script identity comparisons (same sentence in both scripts), and multiple random baselines.

Our contributions are:

1. We introduce Serbian digraphia as a controlled evaluation paradigm for assessing whether learned concept representations capture abstract semantics or remain tied to script-specific token representations.
2. We demonstrate that SAE features in Gemma models exhibit substantial script invariance: averaged across all models, cross-script similarity between identical sentences in Serbian Latin and Serbian Cyrillic reaches $\sim$0.58 Jaccard, with cross-script cross-paraphrase at $\sim$0.47, both significantly exceeding the cross-script random baseline of $\sim$0.28.
3. We characterize how script invariance varies across model scale, finding that larger models maintain more consistent script-independent representations.

Our results suggest that SAE-learned concepts do capture semantic structure that transcends surface-level tokenization, with implications for understanding how neural networks represent meaning across diverse input formats.

2 Related Work

Sparse Autoencoders for Interpretability. Sparse Autoencoders have emerged as a key tool for mechanistic interpretability, addressing the challenge of *superposition* (Elhage et al., 2022). Bricken et al. (2023) demonstrated that SAEs can decompose MLP activations into interpretable, monosemantic features, while Cunningham et al. (2024) showed that SAE features in language models correspond to human-interpretable concepts. We leverage these SAEs to investigate whether the concepts they learn exhibit invariance to orthographic variation.

Cross-lingual and Multilingual Representations. A substantial body of work has investigated whether multilingual models develop language-agnostic representations. Pires et al. (2019) found that multilingual BERT exhibits surprising cross-lingual transfer, even between languages with no lexical overlap. Conneau et al. (2020) showed that cross-lingual representations emerge at scale without explicit alignment objectives, while Wu & Dredze (2019) demonstrated zero-shot cross-lingual transfer across typologically diverse languages. Work on Hindi-Urdu—languages that are linguistically similar but use different scripts (Devanagari and Nastaliq)—has shown that cross-script transfer is possible but imperfect due to vocabulary differences and the lack of a clean mapping between scripts (Moosa et al., 2023; Xhelili et al., 2024). Our work sidesteps such confounds entirely: Serbian digraphia provides a deterministic mapping that allows lossless conversion, enabling us to vary script while holding semantics *exactly* constant.

3 Methodology

3.1 Serbian Digraphia as a Controlled Testbed

Serbian is one of few languages with active digraphia: it is written interchangeably in Latin script and Cyrillic script. What makes Serbian uniquely suited for our investigation is that *both scripts are used with near-equal frequency in everyday life*. This means that in any large training corpus, the same concepts and linguistic patterns appear in both orthographic forms. Critically, while semantics remain identical, the tokenizer produces entirely different token sequences for each script. This creates an ideal controlled experiment.

3.2 Dataset

We construct a dataset of 30 sentence triplets, each containing:

- **Original:** A natural sentence covering diverse topics (nature, daily activities, abstract concepts)
- **Paraphrase:** A semantically equivalent rephrasing with different lexical choices
- **Random:** An unrelated sentence with no semantic connection

Each triplet exists in three language variants: English, Serbian Latin, and Serbian Cyrillic. Serbian translations were carefully generated and rigorously verified for accuracy. The complete dataset

comprises 270 unique sentences (30 triplets, 3 variants, 3 languages/scripts). Using LaBSE sentence embeddings (Feng et al., 2022), we further confirmed that cross-script pairs achieve near-ceiling semantic similarity. We also controlled for potential tokenization confounds by ensuring comparable token counts across script pairs. Additional details on translation methodology, tokenization analysis, and phrase similarity verification are provided in Appendix A.

We evaluate across the Gemma model family (Gemma Team, 2024), spanning two orders of magnitude in scale: Gemma-3-270M, Gemma-3-1B, Gemma-3-4B, Gemma-3-12B, and Gemma-3-27B. For each model, we use Gemma Scope 2 SAEs (Google DeepMind, 2025)—JumpReLU sparse autoencoders with 65,536 features trained on model activations. We select 3–4 layers per model spanning early, middle, and late processing stages (e.g., layers 12, 24, 31, 41 for Gemma-3-12B).

Following established practices in the SAE interpretability literature (Templeton et al., 2024; Rajamanoharan et al., 2024), we use a width of 65k features and a medium L0 sparsity level, which provides a balance between feature granularity and reconstruction quality. The activation threshold of $\tau = 0.1$ corresponds to the JumpReLU threshold used.

3.3 Feature Extraction Pipeline

Given an input sentence s, we extract the set of active SAE features $F(s)$ as follows:

1. **Tokenization:** Convert s to token sequence $\mathbf{t} = (t_1, \ldots, t_n)$ using the Gemma tokenizer.
2. **Forward pass:** Compute hidden state $\mathbf{h}^{(l)} \in \mathbb{R}^d$ at layer l for the final token position.
3. **SAE encoding:** Obtain feature activations $\mathbf{a} = \text{SAE}_{\text{enc}}(\mathbf{h}^{(l)}) \in \mathbb{R}^{65536}$.
4. **Thresholding:** Define active feature set $F(s) = \{i : a_i > \tau\}$ where $\tau = 0.1$ (the JumpReLU activation threshold).

We use last-token pooling rather than mean pooling, as we found empirically that it yields more robust results. This pipeline is applied identically across all five Gemma models and all tested layers, with the same threshold $\tau = 0.1$ used throughout.

3.4 Comparison Types

We define 14 comparison types to systematically test our hypotheses. The key comparisons are:

Baseline Comparisons

- *Original vs. Paraphrase* (English, Serbian Latin, Serbian Cyrillic): Tests whether SAE features capture semantic similarity within a single script/language.
- *Original vs. Random* (same variants): Establishes baseline similarity for unrelated sentences.

Cross-Script Comparisons (Primary Test)

- *Cross-Script Original* (Serbian Latin Original vs. Serbian Cyrillic Original): The core test of script invariance for identical sentences.
- *Cross-Script Paraphrase* (Serbian Latin Paraphrase vs. Serbian Cyrillic Paraphrase): A robustness check ensuring script invariance holds.
- *Cross-Script Cross-Paraphrase* (Serbian Latin Original vs. Serbian Cyrillic Paraphrase and vice versa): Tests combined script and lexical variation.

Random Baselines

- *Cross-Script Random* (Latin Original vs. Cyrillic Random and Cyrillic Original vs. Latin Random): Unrelated sentences across Serbian scripts.
- *Cross-Language Random* (Serbian vs. English unrelated sentences): Establishes a floor for random similarity.

3.5 Evaluation Metric

We measure representational similarity using Jaccard similarity over active feature sets. We use Jaccard similarity over active feature sets rather than cosine similarity or other continuous-valued metrics because our analysis concerns which features activate, not their magnitudes, capturing representational overlap in terms of shared feature identities. Given two sentences s_1 and s_2 with active feature sets $F(s_1)$ and $F(s_2)$, we compute:

$$J(s_1, s_2) = \frac{|F(s_1) \cap F(s_2)|}{|F(s_1) \cup F(s_2)|} \tag{1}$$

Jaccard similarity ranges from 0 (no overlap) to 1 (identical feature sets). For each comparison type, we compute Jaccard similarity for all 30 sentence pairs and report the mean across these pairs.

4 Experiments and Results

4.1 Evidence for Script-Invariant Semantic Representations

We first establish that SAE features encode semantic information by comparing original sentences to their paraphrases versus unrelated random sentences. If SAE features capture meaning, semantically related pairs should activate more similar feature sets than unrelated pairs.

This prediction holds with complete consistency. Across all model-layer combinations and all three language/script conditions (English, Serbian Latin, Serbian Cyrillic), original-paraphrase similarity exceeds original-random similarity in 100% of cases, with paraphrase similarity averaging $\sim$0.54 compared to $\sim$0.28 for random pairs. Results are provided in Appendix B.1.

Having established that SAE features capture semantics, we now test our central hypothesis: do identical sentences in Latin and Cyrillic scripts activate similar features despite entirely different tokenization?

Table 1 presents our core finding. We compare five conditions in decreasing order of expected similarity: (1) *Cross-Script Original*—the same sentence rendered in both scripts, (2) *Cross-Script Paraphrase*—the same paraphrase rendered in both scripts, (3) *Cross-Script Cross-Paraphrase*—original in one script versus paraphrase in the other, (4) *Cross-Script Random*—unrelated sentences across Serbian scripts, and (5) *Cross-Language Random*—unrelated Serbian and English sentences. If representations are script-invariant, we expect high similarity for conditions (1) and (2), moderate similarity for condition (3), and low similarity for condition (4) and condition (5).

Table 1: Evidence for script-invariant representations. Results averaged across all models and layers.

Comparison Type	**Mean Jaccard Similarity**
Cross-Script Original (Latin $\leftrightarrow$ Cyrillic, same sentence)	0.58
Cross-Script Paraphrase (Latin $\leftrightarrow$ Cyrillic, same paraphrase)	0.59
Cross-Script Cross-Paraphrase (Latin/Cyrillic Orig $\leftrightarrow$ Cyrillic/Latin Para)	0.47
Cross-Script Random (Latin $\leftrightarrow$ Cyrillic, unrelated)	0.28
Cross-Language Random (Serbian $\leftrightarrow$ English, unrelated)	0.19

The results strongly support script invariance. Identical sentences across scripts achieve a Jaccard similarity of $\sim$0.58, substantially higher than the random baseline of $\sim$0.28, with cross-script paraphrase similarity ($\sim$0.59) confirming robustness across different phrasings.

Furthermore, cross-script random similarity ($\sim$0.28) exceeds cross-language random similarity ($\sim$0.19), indicating that the model treats Serbian Latin and Serbian Cyrillic as more similar to each other than either is to English. This result is particularly striking given that the tokenizer produces entirely disjoint token vocabularies for the two Serbian scripts, with no surface-level signal that these scripts represent the same language.

The ordering *Cross-Script Original* $\approx$ *Cross-Script Paraphrase* $>$ *Cross-Script Cross-Paraphrase* $>$ *Cross-Script Random* $>$ *Cross-Language Random* suggests that SAE features reflect a semantic hierarchy where meaning, not orthography, is the primary driver of representational similarity.

4.2 Effect of Model Scale

Having established script invariance across our experiments, we now examine how this property varies with model scale. Full numerical results are provided in Appendix B.2.

Figure 1 presents within-script semantic discrimination across the Gemma model family. Larger models achieve lower random baselines while maintaining comparable paraphrase similarity, resulting in greater separation between semantically related and unrelated pairs. We note that smaller models exhibit higher Jaccard similarity across all conditions, including random baselines, likely because fewer total features activate and the resulting feature sets overlap more by chance. By 27B, all three conditions (English, Serbian Latin, Serbian Cyrillic) converge to nearly identical gaps ($\sim$0.28–0.29), suggesting that at sufficient scale, the model achieves comparable semantic discrimination regardless of language or script. The slight decrease in paraphrase similarity with scale likely reflects increased feature granularity rather than degraded understanding: larger models may develop finer-grained features that distinguish subtle differences between paraphrases, which smaller models conflate. Notably, Serbian Latin and Serbian Cyrillic exhibit remarkably similar trajectories despite disjoint tokenizations, suggesting gains in semantic encoding are not script-specific.

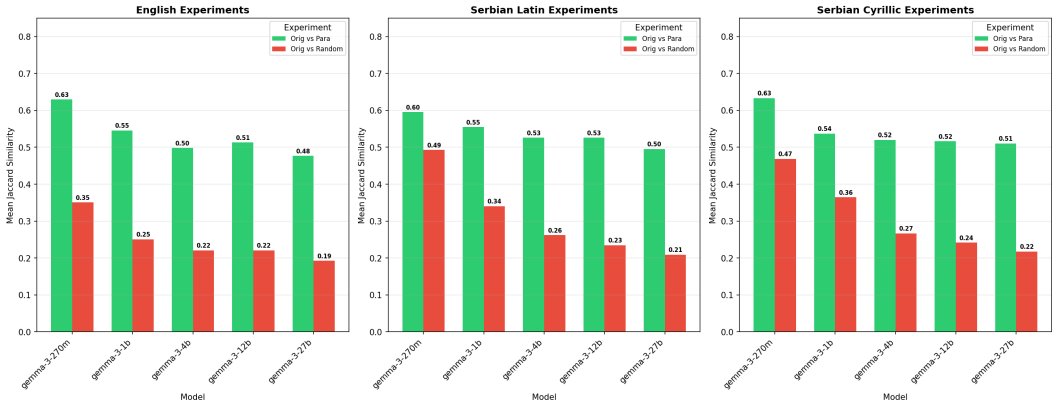

Figure 1: Baseline semantic discrimination across model scales. Larger models achieve greater separation between paraphrase and random similarity, with Serbian Latin and Serbian Cyrillic following nearly parallel trajectories.

Figure 2 presents cross-script similarity across model scales. Cross-script original similarity (identical sentences across scripts) increases from 0.50 at 270M to 0.65 at 27B, while both random baselines move in the opposite direction: cross-script random decreases from $\sim$0.42 to $\sim$0.21, and cross-language random from $\sim$0.25 to $\sim$0.16. This pattern of semantic similarity rising while random baselines fall demonstrates that larger models develop substantially more robust script-invariant representations. Interestingly, cross-script cross-paraphrase similarity (original in one script vs. paraphrase in the other) remains stable across scales ($\sim$0.47–0.49). We hypothesize that two opposing effects may cancel out: larger models improve at recognizing equivalent content across scripts, but also become more sensitive to exact word choices. The net result is that cross-script cross-paraphrase similarity stays roughly constant, though further investigation is needed to confirm this explanation.

5 Discussion

Our results demonstrate that SAE-learned features capture semantic content that transcends orthographic representation. Despite entirely disjoint tokenization, identical Serbian sentences across scripts activate highly overlapping features, and this similarity exceeds within-script paraphrase similarity, suggesting the model is more sensitive to word choice than to script. The cross-script cross-paraphrase results provide evidence against memorization-based explanations: combinations unlikely to co-occur in training data still show substantial overlap, indicating genuine semantic alignment.

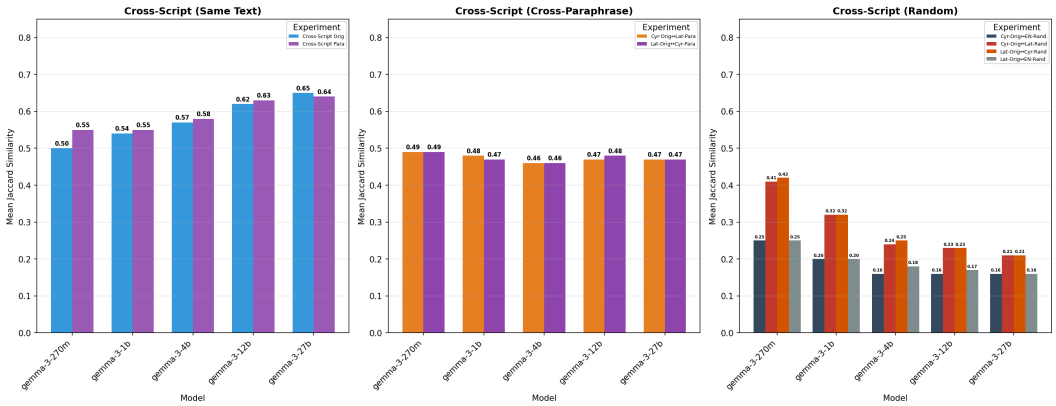

Figure 2: Cross-script similarity across model scales. Cross-script original and paraphrase similarity increase with model size while cross-script random decreases, indicating increasingly robust script-invariant representations. Notably, cross-script cross-paraphrase remains stable across scales.

These findings suggest that SAE features capture semantic content at a level of abstraction above surface orthography. If this property holds more broadly, it could have implications for cross-script interpretability research, though further work is needed to assess generalizability beyond Serbian digraphia.

5.1 Limitations and Future Work

Our experiments focus exclusively on Serbian digraphia. While Serbian provides an ideal controlled setting due to its deterministic script mapping and balanced real-world usage, other multi-script languages present distinct challenges. Extending this paradigm to these languages would test whether script invariance is a general property or specific to clean orthographic mappings like Serbian.

We evaluate only the Gemma model family with Gemma Scope 2 SAEs using a single configuration (65k width, medium L0, threshold 0.1). Different architectures, training procedures, or SAE hyperparameters may yield different patterns. Additionally, our use of a single activation threshold (= 0.1) may influence which features are considered active; future work should analyze sensitivity to threshold choice across model sizes. Our dataset of 30 sentence triplets, while sufficient to establish clear statistical trends, is limited in size and domain coverage. Future work should expand to larger, more diverse corpora.

Our analysis measures feature overlap but does not establish causal relationships; future work could employ activation patching or feature ablation to verify whether shared features directly contribute to cross-script understanding. Additionally, identifying which specific SAE features are most script-invariant could reveal interpretable semantic concepts that serve as anchors for cross-lingual research. We hope Serbian digraphia, as a naturally controlled setting, proves useful for future investigations into how neural networks represent meaning across orthographic boundaries.

6 Conclusion

We introduced Serbian digraphia as a controlled evaluation paradigm for testing whether learned concept representations capture abstract semantics or remain tied to script-specific token patterns. Our experiments across the Gemma model family demonstrate that SAE-learned representations are substantially script-invariant: identical sentences across scripts activate overlapping feature sets far exceeding random baselines, with script variation introducing less representational divergence than paraphrasing. This property strengthens with model scale, as larger models exhibit more robust script-independent representations.

These findings suggest that SAE features can capture meaning at a level of abstraction that transcends surface tokenization, supporting their potential as interpretable, generalizable concept rep-

resentations. We hope this work provides a useful foundation for further investigation into script-invariance and orthographic abstraction in neural networks.

REFERENCES

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 878–891, 2022.

Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Google DeepMind. Gemma scope 2. Technical report, 2025. Technical Report. Available at https://deepmind.google/.

Ibraheem Muhammad Moosa, Taraka Rama Rao Agrawal, and Anoop Mukherjee. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, 2023.

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

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Neel Nanda, et al. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

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, C. Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 833–844, 2019.

Orgest Xhelili, Stefan Ruseti, and Mihai Dascalu. Breaking the script barrier in multilingual pre-trained language models with transliteration-based post-training alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

LLM USAGE DISCLOSURE

In accordance with ICLR's Code of Ethics, we disclose the following use of large language models in this work. We used Claude (Anthropic) to assist with paper writing, including improving wording, grammar, and drafting sections. We also used LLMs to help generate experiment code. All LLM-generated content, code, and results were thoroughly reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of this submission.

A DATASET DETAILS

A.1 TRANSLATION METHODOLOGY

English sentences were translated to Serbian using the Google Translate V3 API, which returns translations in Cyrillic script by default. Serbian Latin versions were then generated via deterministic transliteration using Serbian's direct character correspondence.

To ensure quality, we implemented several verification measures:

- **Length validation:** All 90 English source sentences were constrained to 7–13 words to ensure comparable complexity.
- **Transliteration testing:** Round-trip tests (Latin $\rightarrow$ Cyrillic $\rightarrow$ Latin) verified mapping integrity against known test cases.
- **LLM verification:** Each translation batch was reviewed by Claude Sonnet 4 (using extended thinking mode) for accuracy, natural phrasing, and semantic drift. Additionally, predefined test cases were used to validate translation quality.

A.2 TOKENIZATION ANALYSIS

A potential confound in our analysis is that Latin and Cyrillic scripts might tokenize differently, which could artificially drive differences in SAE activations independent of semantic content. We addressed this concern through two analyses.

Token Count Comparison. We computed mean token counts for each script across all semantic variants (original, paraphrase, random). As shown in Figure 3, token counts are nearly identical between scripts for the same content, with differences of only 1–2 tokens on average. This indicates that the Gemma tokenizer does not systematically produce longer or shorter sequences for one script over the other.

Token Difference vs. Feature Similarity. To directly test whether tokenization differences predict feature overlap, we plotted the token count difference (Cyrillic $-$ Latin) against SAE feature Jaccard similarity for each sentence pair. As shown in Figure 4, there is no meaningful relationship ($r = 0.055$, $p = 0.188$). Sentences where Cyrillic uses more tokens than Latin do not show systematically different feature overlap compared to sentences with identical token counts.

These results rule out tokenization as a confounding explanation for our cross-script similarity findings.

A.3 PHRASE SIMILARITY VERIFICATION

To independently verify that our sentence pairs exhibit the expected semantic relationships, we computed sentence embeddings using LaBSE (Language-agnostic BERT Sentence Embeddings). LaBSE is a multilingual model trained on 109 languages designed to map semantically equivalent sentences to similar embeddings regardless of language or script.

We analyzed three types of sentence similarities:

- **Cross-script:** Serbian Latin $\leftrightarrow$ Serbian Cyrillic (same meaning, different scripts)
- **Cross-language:** English $\leftrightarrow$ Serbian Latin/Cyrillic (same meaning, different languages)

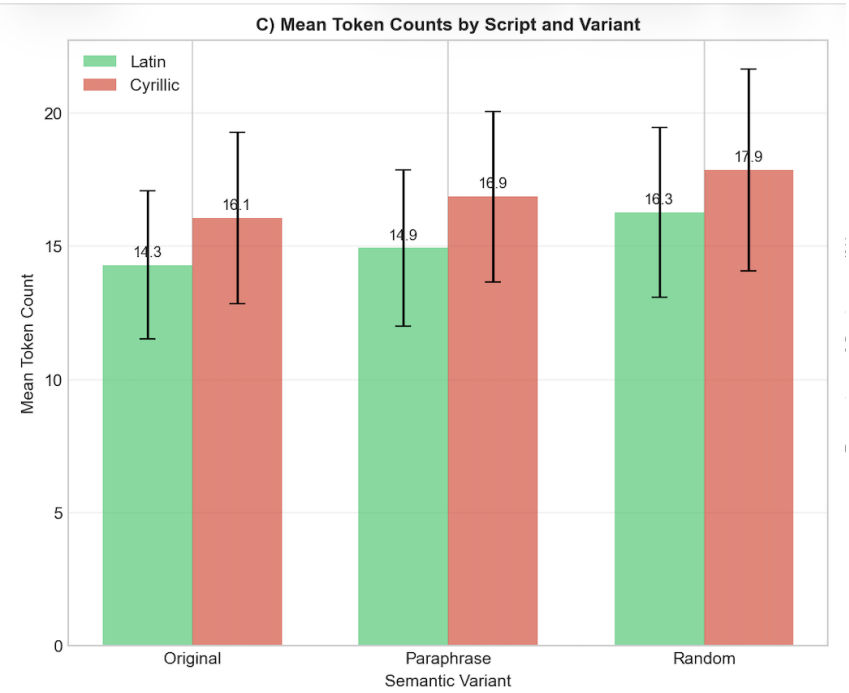

Figure 3: Mean token counts by script and sentence type.

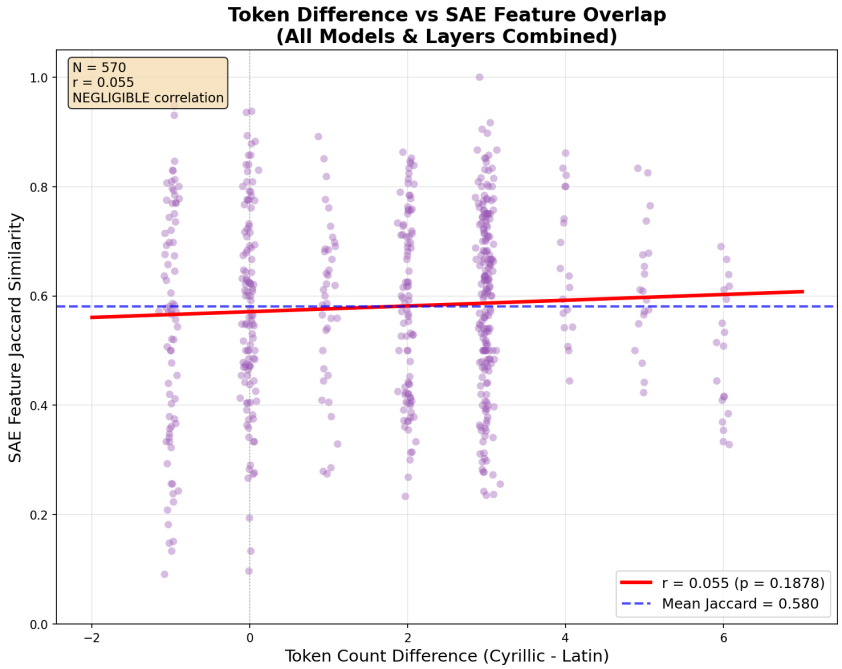

Figure 4: Token count difference vs. SAE feature Jaccard similarity.

- **Paraphrase:** Original sentences vs. semantically equivalent paraphrases (same script and meaning, different wording)

As a control, we also computed random sentence similarity within scripts (semantically unrelated pairs) to establish a baseline.

Figure 5 presents the results. Part (a) shows a histogram of similarity scores across conditions: random pairs cluster at low similarity, paraphrase pairs (English, Serbian Latin, Serbian Cyrillic) show high similarity, and cross-script original pairs achieve near-ceiling similarity. Part (b) displays the same data as box plots, revealing the distributions for each condition. Cross-script original pairs exhibit the highest and tightest distribution (>0.95), confirming that Latin and Cyrillic versions are recognized as semantically identical by an independent model. Paraphrase pairs show high but slightly more variable similarity, as expected given lexical differences. These results validate that our dataset exhibits the intended semantic structure.

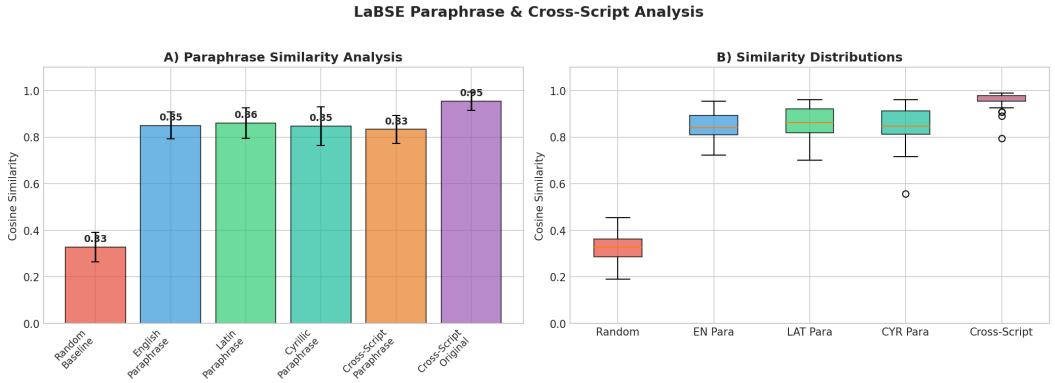

Figure 5: LaBSE sentence similarity verification. (a) Histogram of similarity scores showing clear separation between random pairs and semantically related pairs. (b) Box plots of similarity distributions by condition. Cross-script original pairs achieve near-ceiling similarity, confirming semantic equivalence across scripts.

B Experimental Results

B.1 Baseline Validation

Baseline validation results averaged across all models and layers. Original vs. paraphrase similarity consistently exceeds original vs. random similarity across all conditions, confirming that SAE features capture semantic content.

Table 2: Baseline validation results.

Condition	**Orig ↔ Para**	**Orig ↔ Rand**
English	0.53	0.25
Serbian Latin	0.54	0.31
Serbian Cyrillic	0.54	0.31

B.2 Full Results by Model

This appendix presents Jaccard similarity results averaged across all tested layers for each model. Abbreviations: EN = English, SR-Lat = Serbian Latin, SR-Cyr = Serbian Cyrillic.

Table 3: Within-language comparisons.

Comparison	**270M**	**1B**	**4B**	**12B**	**27B**
EN: Orig vs Para	0.629	0.546	0.498	0.513	0.477
EN: Orig vs Rand	0.351	0.251	0.221	0.221	0.193
SR-Lat: Orig vs Para	0.595	0.555	0.526	0.526	0.496
SR-Lat: Orig vs Rand	0.493	0.341	0.262	0.235	0.210
SR-Cyr: Orig vs Para	0.634	0.537	0.520	0.516	0.510
SR-Cyr: Orig vs Rand	0.469	0.365	0.267	0.242	0.218

Table 4: Cross-script and cross-language comparisons.

Comparison	**270M**	**1B**	**4B**	**12B**	**27B**
Cross-Script Orig	0.501	0.537	0.571	0.624	0.649
Cross-Script Para	0.549	0.547	0.585	0.626	0.645
Lat Orig vs Cyr Para	0.495	0.468	0.457	0.480	0.470
Cyr Orig vs Lat Para	0.488	0.475	0.461	0.475	0.468
Lat Orig vs Cyr Rand	0.421	0.324	0.253	0.233	0.211
Cyr Orig vs Lat Rand	0.413	0.317	0.239	0.225	0.210
Lat Orig vs EN Rand	0.251	0.199	0.180	0.173	0.164
Cyr Orig vs EN Rand	0.248	0.196	0.162	0.161	0.159