

TBLS: Transition-Based Language Similarity through Activation Perturbation

Tao Ren¹ Johannes Bjerva¹ Qiongxiu Li^{2*}

¹Department of Computer Science, ²Department of Electronic Systems
Aalborg University, Copenhagen, Denmark
{taoren, jbjerva}@cs.aau.dk, qili@es.aau.dk

Abstract

Language similarity is a central concept in multilingual NLP. Prior work typically infers similarity from the static geometry of representation spaces, and the results often depend on model architecture, training data, and probing methods. In this work, we propose transition-based language similarity (TBLS), a new measure that characterizes language similarity through cross-lingual transition. The key intuition is that if two languages occupy nearby regions in a model’s representation space, model outputs are more likely to transition between them under small local variations. By aggregating such transition behavior across languages, TBLS can be derived, reflecting language proximity encoded by the model.

We operationalize this idea by perturbing model activations and measuring the resulting cross-lingual transition behavior. Across five multilingual language models and two corpora covering 44 languages, TBLS exhibits strong cross-model and cross-corpus consistency. Further analyses show that TBLS captures meaningful linguistic regularities and transfer-related structure, while activation perturbation provides behavioral evidence consistent with the hypothesis that semantically aligned multilingual representations occupy nearby, locally connected regions. These findings suggest that perturbation-induced transition dynamics reveal a surprisingly stable form of multilingual relationships across models and corpora.

1 Introduction

Language similarity is a fundamental concept in multilingual natural language processing and linguistics (Blaschke et al., 2025; Huang et al., 2025). Long before the advent of multilingual language models (mLMs), linguists have broadly studied language similarity, such as genealogical or typological similarity (Greenberg, 1974), which are largely



Figure 1: Danish input with a linguistically related Norwegian output under the proposed activation perturbation.

independent of any particular computational model. These notions have been used to guide language selection, data transfer, and cross-lingual evaluation (Lin et al., 2019; Ponti et al., 2019).

Prior work has shown that mLMs exhibit rich cross-lingual geometry in their representation spaces (Kudugunta et al., 2019; Chang et al., 2022; Wen-Yi and Mimno, 2023). Building on this observation, several studies have attempted to infer language similarity from model representations (Bjerva and Augenstein, 2018; Conneau et al., 2020b; Lin et al., 2024; Shah et al., 2024). These methods often produce language similarity patterns that vary across model architectures, training corpora, and tasks, possibly because mLMs capture cross-lingual relationships in different ways.

While such methods effectively capture language similarities, they provide only a *static* snapshot of the model’s cross-lingual geometry, typically by measuring distances between sentence embeddings or hidden states (Hewitt and Manning, 2019; Nastase and Merlo, 2025). However, the multilingual output dynamics induced by this geometry remain largely unexplored. Intuitively, if two languages occupy nearby regions in the representation space, outputs are more likely to transition toward each other under local variations, and vice versa.

*Corresponding author.

Aggregating such local transition behavior across languages naturally yields a measure of language similarity.

Based on this intuition, we propose transition-based language similarity (TBLS), which quantifies language similarity through cross-lingual transitions, where a transition refers to a shift in output probability mass from the source language toward another language. TBLS captures how easily model outputs transition between languages under local perturbations, providing a behavior-driven view of proximity in the underlying representation geometry.

To operationalize this idea, we induce controlled variations in model representations to trigger output transition. This approach is supported by prior work showing that small activation perturbations can reveal local structure in neural networks (Ren and Li, 2024; Shahani et al., 2025; Ren et al., 2026). Figure 1 illustrates an example of such local exploration, where the output transitions from Danish to a closely related language (Norwegian), reflecting the cross-lingual geometry encoded by the model.

In this paper, we address the following research questions:

RQ1 Can perturbation-induced cross-lingual transitions be used to characterize multilingual relationships in mLMs?

RQ2 What linguistic, structural, and transfer-related properties are reflected in the multilingual relationships captured by TBLS?

RQ3 What do perturbation-induced transitions reveal about the local organization of multilingual representations?

Contributions: (i) We propose TBLS, a behavior-driven measure that characterizes multilingual language similarity through perturbation-induced cross-lingual transition. (ii) We conduct a systematic analysis of the linguistic, structural, and transfer-related properties captured by TBLS, showing that it captures aspects of multilingual relationships complementary to static representation-based similarity measures. (iii) We show that TBLS provides a direct behavioral probe into multilingual semantic neighborhoods, while perturbation-induced transition patterns remain highly consistent across mLMs and corpora.

2 Related Work

Prior work has shown that mLMs shape cross-lingual geometry in the high-dimensional repre-

sentation space (Conneau et al., 2020b; Chang et al., 2022). For example, Wang et al. (2024) found strong correlations between neuron overlap and cross-lingual transfer performance, while Zeng et al. (2025) showed that semantically equivalent inputs across languages tend to converge to similar latent representations. Building on this observation, several studies attempt to infer language similarity from aligned sentence representations, yielding representation similarity measures (REP) (Hewitt and Manning, 2019; Rama et al., 2020; Conneau et al., 2020b; Shah et al., 2024). As suggested by Wen-Yi and Mimno (2023), the resulting similarity patterns often vary substantially across models. Recent work has proposed other language similarity measures. Libovický et al. (2020) analyzed language centroids and argued that multilingual representations contain both language-specific and language-neutral semantic components. Meanwhile, Del and Fishel (2022) proposed Average Neuron-wise Correlation (ANC), which measures cross-lingual similarity at the neuron level.

However, these approaches typically rely on static measurements of representation geometry, usually based on distances such as cosine similarity. Recent work further questions whether such geometric measurements alone reliably reflect the linguistic information encoded in sentence embeddings (Nastase and Merlo, 2025). Motivated by these limitations, we complement these approaches by studying cross-lingual output dynamics under local perturbations, providing an alternative behavioral perspective on multilingual relationships in mLMs. Our approach is inspired by recent studies showing that activation perturbation can effectively probe local structure and behavior in neural networks (Ren and Li, 2024; Shahani et al., 2025).

3 Transition-Based Language Similarity

In this section, we introduce *Transition-Based Language Similarity (TBLS)*, a measure that quantifies language similarity through cross-lingual transition in mLMs. We introduce controlled variations to model representations during inference, allowing outputs to explore local neighborhoods of the representation space.

In practice, we implement these variations using activation perturbation (Ren et al., 2026), where Gaussian noise is injected into intermediate activations during inference. Concretely, let a_ℓ denote the activation output of a feed-forward sublayer

at layer ℓ . During a perturbed forward pass, we replace it with

$$\tilde{a}_\ell = a_\ell + \sigma \xi_\ell, \quad \xi_\ell \sim \mathcal{N}(0, I),$$

where $\sigma > 0$ controls the noise strength. Noise is independently sampled across layers and forward passes, while the model parameters remain fixed.

3.1 Definition of TBLS

For an input sequence $x \in D_{L_i}$, where D_{L_i} denotes the set of inputs in language L_i , introducing local variations to model representations induces stochasticity in the model’s output. If two languages occupy nearby regions in the model’s representation space, perturbations around inputs from one language should assign higher probability mass to the other, whereas distant languages should rarely emerge. Aggregating such transition behavior across inputs thus yields a notion of language similarity.

Let $P(L_j | x_i, \sigma_{L_i})$ denote the probability that a perturbed forward pass from an input x_i in L_i is identified as another language L_j , where σ_{L_i} denotes the perturbation scale calibrated for language L_i . Averaging over inputs in L_i , we obtain a language-to-language transition probability

$$p_{i \rightarrow j} = \mathbb{E}_{x \sim \mathcal{D}_{L_i}} [P(L_j | x, \sigma_{L_i})]. \quad (1)$$

The expectation is estimated empirically over the dataset (see Section 4 for details). The transition probabilities $p_{i \rightarrow j}$ and $p_{j \rightarrow i}$ are generally asymmetric. To obtain a symmetric and interpretable notion of language similarity, we define transition-based language similarity as

$$\text{TBLS}(L_i, L_j) = \min(p_{i \rightarrow j}, p_{j \rightarrow i}). \quad (2)$$

A large similarity score indicates strong mutual similarity between the two languages. The directional probabilities $p_{i \rightarrow j}$ remain informative and will be used in downstream transfer analysis (Section 5).

Language-level escape noise. To evaluate perturbation effects at a comparable operating point across languages, we estimate a language-specific perturbation scale, referred to as *escape noise*.

For a given x in L_i , we define the sample-level escape noise $\sigma_{\text{esc}}(x)$ as the smallest perturbation strength at which the generated output sequence

$y(x, \sigma)$ is no longer identified as L_i by the language identification model (see Appendix A for the procedure used for non-autoregressive models).

The *language-level escape noise* for L_i is then obtained by averaging $\sigma_{\text{esc}}(x)$ across inputs from its corpus:

$$\sigma_{L_i} = \frac{1}{|D_{L_i}|} \sum_{x \in D_{L_i}} \sigma_{\text{esc}}(x). \quad (3)$$

Empirically, we observe that when the perturbation strength approaches the escape noise level, the labeled dependency overlap, which measures the proportion of syntactic relations between sentence components that remain identical between the original and perturbed parses, drops substantially (see Appendix B for details). This suggests that escape noise is sufficient to push the model out of the local region associated with the source language, enabling the exploration of nearby regions in the representation space.

3.2 Estimation of Transition Probability

For a given x and σ , forward passes produce a probability distribution over the model vocabulary

$$p_\theta(v | x, \sigma), \quad v \in \mathcal{V},$$

such that

$$\sum_{v \in \mathcal{V}} p_\theta(v | x, \sigma) = 1.$$

Let $\mathcal{S} \subset \mathcal{V}$ denote the set of non-linguistic tokens (e.g., <pad>, <eos>, <unk>), which are not associated with any language. We consider the effective vocabulary with non-linguistic tokens excluded: $\mathcal{V}' = \mathcal{V} \setminus \mathcal{S}$. The remaining probability mass over $\mathcal{V}'$ is

$$\sum_{v \in \mathcal{V}'} p_\theta(v | x, \sigma) = 1 - \sum_{s \in \mathcal{S}} p_\theta(s | x, \sigma). \quad (4)$$

We partition $\mathcal{V}'$ into language-specific token sets $\mathcal{V}_{L_1}, \dots, \mathcal{V}_{L_K}$, where $\mathcal{V}_{L_k}$ contains the tokens assigned to language L_k . The probability that the perturbed next-token distribution belongs to L_k is then defined as

$$P(L_k | x, \sigma) = \sum_{v \in \mathcal{V}_{L_k}} p_\theta(v | x, \sigma). \quad (5)$$

This procedure performs language identification using the full token probability distribution

rather than only the top-1 prediction. The transition probability $p_{i \rightarrow j}$ and the symmetric similarity $\text{TBLs}(L_i, L_j)$ are then computed based on these language-level distributions. This soft aggregation provides a more stable estimate of language-level probability mass.

4 Experimental Setup

Models. We evaluate TBLs on a diverse set of mLMs. We include MGPT (Shliazhko et al., 2024) (decoder-only autoregressive), XLM-R LARGE (Conneau et al., 2020a) (masked language model), and MT5-LARGE (Xue et al., 2021) (encoder-decoder). To evaluate TBLs on more recent mLMs, we further consider QWEN3-8B (Yang et al., 2025) and APERTUS-8B (Apertus et al., 2025), which is a recent state-of-the-art model supporting over a thousand languages.

Corpora. We use FLORES (Guzmán et al., 2019) as a semantically aligned multilingual benchmark, and a randomly sampled subset of TATOEBa (Artetxe and Schwenk, 2019) with 2,000 sentences to introduce greater semantic and domain variation across languages.

Language set. Considering the intersection of language coverage across models and corpora, we start our analysis on a set of 45 languages. These languages cover a broad range of language families, scripts, and typological properties, achieving a high entropy score of 0.95 according to the diversity metrics of Ploeger et al. (2026) (Appendix C).

Inference procedure. For decoder-only models, we use the raw sentence directly as the prompt without an instruction template, and greedily generate a continuation for escape-noise calibration. For XLM-R, we use the deterministic left-to-right mask-and-fill procedure described in Appendix A. For mT5, we use single-token span infilling, where one input token is replaced with `<extra_id_0>` and the decoder reconstructs the masked token. For later-position analyses in autoregressive models, preceding tokens are generated greedily before evaluating the distribution at the selected position.

Token language assignment. To capture local output behavior, we construct language-level distributions using only the first perturbed output token. Using later tokens would introduce accumulated drift across decoding steps, which may obscure the local effects of small variations. In Section 5

we will show that TBLs is robust to the choice of perturbed token position.

To reduce ambiguity from short subword tokens shared across languages, we identify *shared tokens* as Latin- or Cyrillic-script Unicode-letter tokens that appear in the tokenized corpora of multiple languages, with non-linguistic tokens removed. In all experiments, token statistics are estimated from MC4 (Xue et al., 2021). For a shared token $\tilde{v}$, its contribution to L_j is weighted by its empirical frequency:

$$w_{\tilde{v}, L_j} = \frac{f_{\tilde{v}, L_j}}{\sum_k f_{\tilde{v}, L_k}}, \quad (6)$$

where $f_{\tilde{v}, L_j}$ denotes the frequency of $\tilde{v}$ in L_j . Shared tokens therefore contribute softly to multiple languages according to these weights. Since these empirical weights are only available for tokens observed as shared in the sampled corpus statistics, they do not cover the remaining vocabulary. Non-shared tokens are therefore assigned a single language using the hard label predicted by XLM-V (Liang et al., 2023). Appendix D provides additional details on token assignment and shows that the resulting shared-token language distributions remain stable across different corpora.

Estimation Procedure. Escape noise is estimated by gradually increasing σ for each x and applying a perturbed forward pass until the predicted language changes. The $\sigma_{\text{esc}}(x)$ is recorded as the first noise level that causes a language change, and σ_{L_i} is obtained by averaging $\sigma_{\text{esc}}(x)$ over all samples in L_i . To estimate the transition probabilities, each sample from L_i is evaluated once under perturbation at σ_{L_i} , and the resulting $P(L_j \mid x, \sigma_{L_i})$ are averaged across samples to obtain $p_{i \rightarrow j}$.

5 Experiments

We compute TBLs matrices from ten model-corpus pairs. Results and code are available at <https://github.com/WildRaincoat/TBLs>.

5.1 Escape Noise Analysis

Most languages exhibit stable escape-noise estimates across models and corpora. Basque is excluded from the analysis because several models fail to consistently reproduce outputs in the original language under unperturbed replay.

We further observe that escape-noise rankings remain relatively consistent across mLMs, with Spearman rank correlations ranging from 0.68 to

0.84. This suggests that perturbation stability reflects a shared property of multilingual representation geometry rather than model-specific artifacts. Full pairwise correlation heatmaps and average escape-noise values for each language are provided in Appendix E.

	Token count			Corpus size
	mGPT	mT5	XLM-R	XLM-R
Correlation	-0.073	-0.080	0.339	0.563

Table 1: Spearman rank correlations between language resource statistics and escape-noise rankings.

We observe that languages with relatively low escape noise also tend to be lower-resource languages in multilingual pretraining. We therefore examine the correlation between escape-noise rankings and multilingual pretraining statistics. Three models (mGPT, mT5, and XLM-R) report language-level token counts, while XLM-R additionally provides estimated corpus sizes for each language. We measure the Spearman rank correlations between language resource and escape-noise.

Table 1 summarizes the results. The correlations are generally weak to moderate, suggesting that escape noise is only partially explained by language resource levels. However, corpus size appears to provide a more reliable estimate than token counts. For example, Chinese and Hungarian have comparable corpus sizes in XLM-R pretraining (63.5G vs. 58.4G), but substantially different token counts (435M vs. 7807M), indicating that token statistics alone may be misleading across languages with different tokenization behavior.

5.2 Behavioral and Linguistic Validity

To characterize the multilingual structure captured by TBLS, we analyze its relation to both linguistic standards and cross-lingual transfer behavior.

Genealogical Structure. We first test whether languages that are close under TBLS are also close in the genealogical language tree. Following prior work (Rama et al., 2020), we reconstruct a language family tree from each language similarity measure, including TBLS and three baselines: REP, corresponding to the core representation-similarity computation used in mPLM-Sim (Lin et al., 2024), Centroid (Libovický et al., 2020), and ANC (Del and Fishel, 2022). The reconstructed trees are then compared with the gold tree from Hammarström

Model	Language Similarity				Random	
	TBLS	REP	Centroid	ANC	Mean	Std
mGPT	0.151	0.188	0.204	0.337	0.450	0.082
mT5-L	0.074	0.128	0.125	0.203	0.450	0.082
XLM-R	0.138	0.072	0.075	0.123	0.450	0.082
Qwen3-8B	0.195	0.171	0.138	0.197	0.450	0.082
Apertus	0.338	0.128	0.122	0.179	0.450	0.082

Table 2: Normalized distance between similarity-induced trees and the gold genealogical tree. Lower is better.

et al. (2025). Tree distances are computed using the leaf-distance metric of Rabinovich et al. (2017) (see Appendix F for clustering details and Appendix G for baseline computation).

Table 2 summarizes the results. The strongest language similarity measure varies across models. TBLS achieves the best performance for mT5-L and mGPT, yet performs worst for APERTUS. These results suggest that multilingual language relationships can emerge through multiple representational mechanisms, and that TBLS captures one important dimension of such relationships. Nevertheless, all similarity measures produce trees substantially closer to the gold tree than random similarity matrices.

Typological Alignment. We further analyze the linguistic properties reflected by TBLS using representational similarity analysis (RSA), computed as the Spearman rank correlation between pairwise TBLS similarities and linguistic similarity matrices from LANG2VEC (Littell et al., 2017). In addition to default TBLS, we also compute TBLS using perturbations restricted to the early, middle, and late thirds of model layers, as well as RSA based on escape noise. Figure 2 summarizes the results, where all RSA scores are averaged across five mLMs.

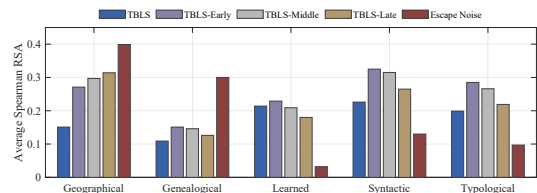


Figure 2: RSA between TBLS variants and linguistic similarity matrices from LANG2VEC.

Overall, TBLS exhibits only moderate correlation with existing linguistic similarity measures. Notably, layer-specific TBLS variants exhibit different linguistic profiles. Early- and middle-layer TBLS align more strongly with syntactic and typological similarity, consistent with prior work show-

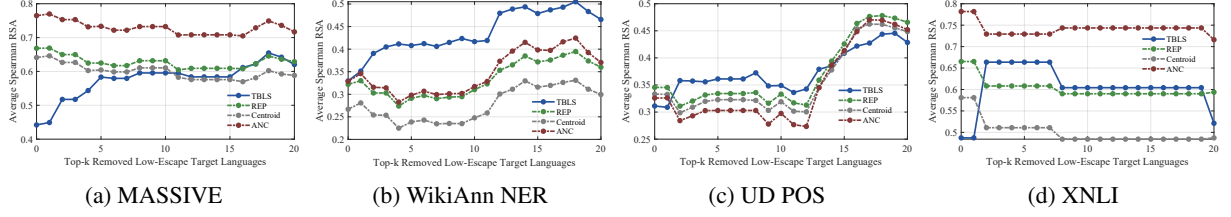
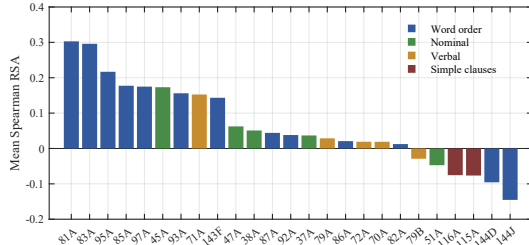


Figure 3: Average Spearman rank correlations between language similarity measures and zero-shot transfer performance under progressively stricter instability-controlled filtering. The horizontal axis shows the number of removed low-escape target languages ranked by XLM-R escape noise.

ing that intermediate BERT layers preferentially encode structural linguistic information (Ruzzetti et al., 2023; de Vries et al., 2020). By contrast, late-layer TBLS shows a relatively stronger association with geographical similarity, which may reflect the increasing influence of corpus- or usage-level regularities in deeper layers (de Vries et al., 2020). As we show next, these layer-specific differences are also reflected in downstream transfer behavior.

Interestingly, the strongest correlation of escape noise is with geographical similarity, suggesting that perturbation robustness itself may organize languages in a way that loosely mirrors real-world geographic proximity. This may indicate that languages spoken in nearby regions exhibit comparable perturbation robustness despite differing typological structure.



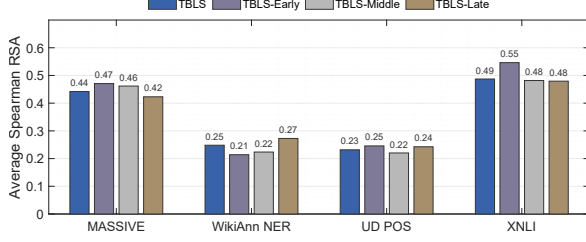


Figure 5: RSA between layer-specific TBLS variants and downstream cross-lingual transfer performance.

ble escape-noise estimates.

We further examine whether these layer-specific TBLS variants exhibit different transfer behavior. As shown in Figure 5, early-layer TBLS consistently exhibits the strongest alignment with MASSIVE, POS, and XNLI transferability, while late-layer TBLS becomes relatively more predictive for WikiAnn NER. Combined with the linguistic results in Figure 2, this suggests that early multilingual transition dynamics capture more transferable structural relationships shared across languages, whereas later transition dynamics may encode more localized lexical regularities.

Importantly, these transfer correlations cannot be fully reduced to the tokenizer-overlap baseline considered here. While tokenizer overlap captures limited transfer structure, TBLS consistently preserves substantially stronger transfer signals across sentence-level benchmarks and remains informative even for cross-script language pairs where overlap-based signals become uninformative. Detailed analyses are provided in Appendix K.

Taken together, these results suggest that TBLS exhibits nontrivial associations with genealogical structure and cross-lingual transfer behavior, indicating that perturbation transition dynamics provide a complementary perspective in multilingual relationships.

5.3 Cross-Model and Cross-Corpus Consistency

Prior work has shown that multilingual language similarity often varies substantially across architectures and training settings (Conneau et al., 2020b; Wen-Yi and Mimno, 2023; Lin et al., 2024; Shah et al., 2024). Nevertheless, we find that some forms of multilingual structure remain surprisingly consistent across models and corpora.

To quantify this consistency, we compute RSA between language similarity matrices under three settings: same-model cross-corpus, cross-model

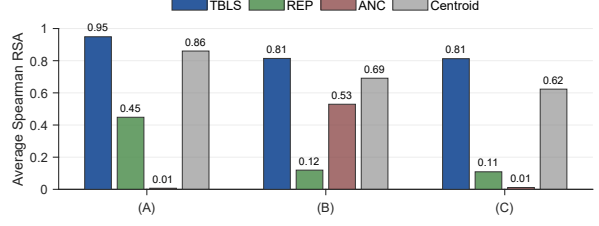


Figure 6: Average RSA across three comparison settings: (A) same-model cross-corpus, (B) cross-model same-corpus, and (C) cross-model cross-corpus.

same-corpus, and cross-model cross-corpus. For static language similarity measures, layers are aligned by relative depth across models, and we report the layer with the highest average RSA. Full layer-wise statistics are provided in Appendix L. Figure 6 reveals a clear hierarchy of cross-model consistency across language similarity measures. REP and ANC exhibit relatively weak agreement, whereas centroid similarity remains substantially more stable across architectures and corpora. Since centroid similarity measures the relative positions of language-level representation centers in parallel data, this already suggests that multilingual models share partially convergent global language geometry. TBLS further extends this observation: across all model–corpus pairs, TBLS RSA ranges from 0.883 to 0.982, indicating that transitions between language regions under perturbation capture a comparatively stable form of multilingual relationships. Full pairwise RSA heatmaps are provided in Appendix M.

Robustness Analysis. We further verify that the observed consistency of TBLS is robust to several design choices and data conditions. First, replacing shared-token weighting with hard token assignments from XLM-V still produces highly consistent TBLS matrices across models, indicating that the observed consistency is not caused by the shared-token mechanism itself. Instead, shared-token weighting mainly improves robustness to corpus variation and semantic mismatch across datasets (Appendix N).

Second, we examine whether TBLS consistency mainly reflects high-resource languages. Grouping languages into high-, medium-, and low-resource bins according to the MGPT pretraining distribution, we find that TBLS remains highly consistent for both high-resource and low-resource languages (cross-model cross-corpus RSA: 0.912 and 0.867 respectively), while the medium-resource group

exhibits somewhat lower agreement (0.647). This suggests that TBLS consistency is not explained by pretraining scale alone (Appendix O).

We additionally evaluate sensitivity to perturbation seeds. Across independently sampled activation-noise seeds, transition distributions become highly stable after aggregation, and independently recomputed TBLS matrices exhibit near-identical rank structure (Appendix P).

Finally, varying token positions and perturbation noise distributions produces highly similar TBLS matrices, indicating that TBLS primarily reflects stable language-level transition behavior rather than local token semantics or specific noise choices (Appendix Q).

Computational cost. Escape-noise calibration introduces additional inference cost because each example is evaluated at progressively increasing perturbation strengths until the first language transition is observed. For decoder-only models, each calibration step uses greedy decoding capped at 25 generated tokens, and the search terminates immediately after the first transition. Empirical calibration costs across models are reported in Appendix R.

5.4 Semantic Neighborhood Structure Revealed by Perturbation

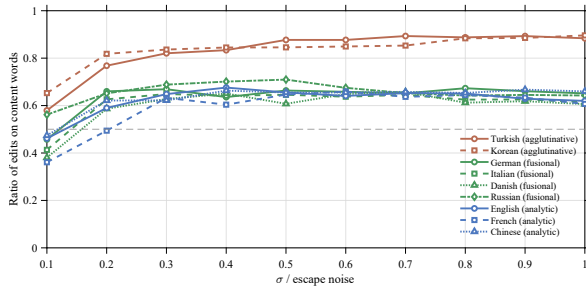


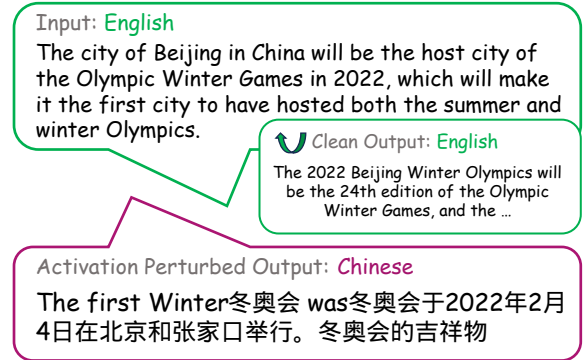
Figure 7: Ratio of content words among perturbed tokens as a function of noise strength (σ normalized by escape noise), evaluated by sweeping perturbation magnitude from 0.1 to 1.0 times escape noise.

Prior work has suggested that semantically equivalent multilingual representations tend to occupy nearby regions in representation space (Pires et al., 2019; Muller et al., 2021; Kargaran et al., 2025). Activation perturbation provides a direct way to behaviorally probe this hypothesis by locally traversing multilingual representation neighborhoods.

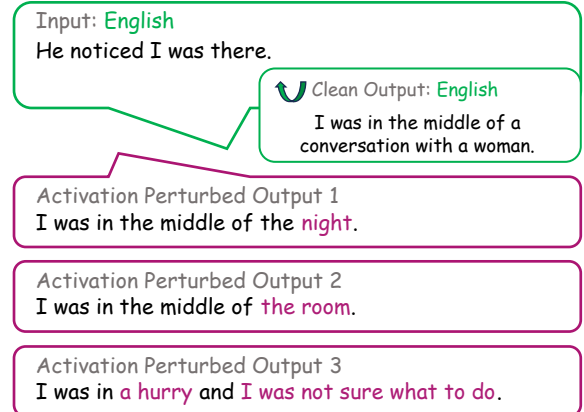
To analyze how perturbation alters generation behavior, we compare perturbed and unperturbed out-

puts across nine languages and categorize changed tokens into content and function words using UPOS tags (Nivre et al., 2016).

Figure 7 shows the ratio of content words among perturbed tokens as a function of noise strength for MGPT (results for other models are in Appendix S). Across languages, the majority of changes occur on content words rather than function words. This indicates that local variations tend to preserve the grammatical scaffolding of a sentence while inducing lexical substitutions. Similar tendencies have also been observed in prior analyses of multilingual representations (Bjerva et al., 2019).



(a) Perturbation can produce outputs in another language while maintaining the same semantic frame.



(b) Perturbation induces lexical variation while preserving the overall sentence structure.

Figure 8: Qualitative examples illustrating the effect of activation perturbation on generated outputs.

Figure 8a illustrates that perturbation can sometimes produce outputs in another language while preserving the same semantic frame, while Figure 8b shows lexical variation with the overall sentence structure largely preserved. Rather than disrupting sentence organization, local perturbations appear to traverse nearby multilingual semantic neighborhoods.

Prior work has mainly studied multilingual se-

mantic alignment through representational similarity analysis, showing that semantically equivalent inputs across languages tend to converge to nearby latent representations (Zeng et al., 2025). Activation perturbation provides a complementary behavioral perspective by directly observing how local variations alter generation trajectories. In this sense, perturbation-induced cross-lingual transitions suggest that multilingual semantic neighborhoods are locally connected in representation space.

Taken together, these observations suggest that activation perturbation primarily explores local multilingual semantic neighborhoods while preserving structural sentence organization.

6 Conclusion

In this work, we introduce TBLS, a behavior-driven measure of multilingual language similarity based on cross-lingual transitions under activation perturbation. Rather than relying solely on static representations, TBLS characterizes multilingual structure through local transition behavior.

Across diverse multilingual language models and corpora, TBLS exhibits meaningful associations with genealogical structure, typological regularities, and cross-lingual transfer, while remaining highly consistent across models and datasets. These findings suggest that local transition behavior captures a stable aspect of multilingual organization.

Finally, our analyses suggest that activation perturbation provides a direct probe into multilingual semantic neighborhoods. The resulting transition patterns indicate that semantically aligned multilingual representations occupy nearby and locally connected regions in representation space.

7 Limitations

We acknowledge several limitations of this work. First, TBLS becomes less reliable for languages with extremely low escape noise. In particular, some models fail to consistently remain in the source language even under unperturbed replay (e.g., Basque). Similarly, TBLS exhibits weaker transfer prediction performance on low-escape target languages. Second, shared tokenizer units can induce non-zero cross-language probability mass even without perturbation, which may favor language pairs with greater token overlap and disadvantage closely related cross-script languages. Third, although TBLS exhibits strong cross-model

consistency, the underlying reasons why mLMs share such stable transition patterns remain unclear. Finally, escape noise is defined operationally as the perturbation level at which the model leaves the source-language region. Although this definition works well in practice, a more theoretical characterization of escape noise may provide deeper insight into the geometry of multilingual representation spaces.

8 Ethical Considerations

This work studies mLMs from an analytical perspective and introduces a probing method based on activation perturbation to derive language similarities. All experiments are conducted using publicly available pretrained models and widely used benchmark corpora. Our analysis focuses on language-level statistical patterns rather than individual user data. No personally identifiable information is used in this work. Potential misuse of the proposed method is limited. Activation perturbation is used here solely as a diagnostic tool to study representation structure in multilingual language models. Our work does not aim to manipulate model outputs or bypass safety mechanisms.

Acknowledgments

REN Tao would like to express his deepest gratitude to Prof. Qiongxiu Li, who first introduced him to scientific research and has been an invaluable mentor throughout his early academic journey. This work is dedicated to her with sincere appreciation for her guidance, encouragement, and lasting influence. We gratefully acknowledge the AAU AI Cloud for providing computational resources for this work. RT and JB are supported by the Novo Nordisk Foundation, under the Ascending Data Science Investigator programme (grant no. NNF24OC0092972).

References

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Durech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, and 84 others. 2025. *Apertus: Democratizing open and compliant llms for global language environments*. Preprint, arXiv:2509.14233.

- Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Johannes Bjerva and Isabelle Augenstein. 2018. [From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916, New Orleans, Louisiana. Association for Computational Linguistics.
- Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. [What do language representations really represent?](#) *Computational Linguistics*, 45(2):381–389.
- Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. [Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.
- Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. [The geometry of multilingual language model representations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Rochelle Choenni and Ekaterina Shutova. 2022. [Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology](#). *Computational Linguistics*, 48(3):635–672.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. [Emerging cross-lingual structure in pretrained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.
- Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. [What’s so special about BERT’s layers? a closer look at the NLP pipeline in monolingual and multilingual models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.
- Maksym Del and Mark Fishel. 2022. [Cross-lingual similarity of multilingual representations revisited](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 185–195, Online only. Association for Computational Linguistics.
- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Nataraajan. 2023. [MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.
- Joseph Greenberg. 1974. *Language Typology: A Historical and Analytic Overview*. De Gruyter Mouton, Berlin, New York.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. [Glottolog 5.2](#).
- John Hewitt and Christopher D. Manning. 2019. [A structural probe for finding syntax in word representations](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.
- Heyan Huang, Zewen Chi, and Xian-Ling Mao. 2025. [Measuring cross-lingual transferability of multilingual transformers](#). *IEEE Transactions on Audio, Speech and Language Processing*, 33:2565–2575.

- Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schuetze. 2025. [MEXA: Multilingual evaluation of English-centric LLMs via cross-lingual alignment](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 27001–27023, Vienna, Austria. Association for Computational Linguistics.
- Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. [Investigating multilingual NMT representations at scale](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.
- Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabza. 2023. [XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.
- Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. [On the language neutrality of pre-trained multilingual representations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.
- Peiqin Lin, Chengzhi Hu, Zheyu Zhang, Andre Martins, and Hinrich Schuetze. 2024. [mPLM-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 276–310, St. Julian’s, Malta. Association for Computational Linguistics.
- Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. [Choosing transfer languages for cross-lingual learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.
- Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. [URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.
- Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. [First align, then predict: Understanding the cross-lingual ability of multilingual BERT](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.
- Vivi Nastase and Paola Merlo. 2025. [Testing the assumptions about the geometry of sentence embedding spaces: the cosine measure need not apply](#). *Preprint*, arXiv:2509.01606.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. [Universal Dependencies v1: A multilingual treebank collection](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. [Cross-lingual name tagging and linking for 282 languages](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2026. [A principled framework for evaluating on typologically diverse languages](#). *Computational Linguistics*, 52(1):237–269.
- Edoardo Maria Ponti, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. [Modeling language variation and universals: A survey on typological linguistics for natural language processing](#). *Computational Linguistics*, 45(3):559–601.
- Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. [Found in translation: Reconstructing phylogenetic language trees from translations](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.
- Taraka Rama, Lisa Beinborn, and Steffen Eger. 2020. [Probing multilingual BERT for genetic and typological signals](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Tao Ren and Qiongxiu Li. 2024. [Prop: Efficient backdoor detection via propagation perturbation for over-parametrized models](#). *Preprint*, arXiv:2411.07036.

- Tao Ren, Xiaoyu Luo, and Qiongxu Li. 2026. [Apex: Probing neural networks via activation perturbation](#). *Preprint*, arXiv:2602.03586.
- Elena Sofia Ruzzetti, Federico Ranaldi, Felicia Logozzo, Michele Mastromattei, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. [Exploring linguistic properties of monolingual BERTs with typological classification among languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14447–14461, Singapore. Association for Computational Linguistics.
- Cheril Shah, Yashashree Chandak, Atharv Mahesh Mane, Benjamin Bergen, and Tyler A. Chang. 2024. [Correlations between multilingual language model geometry and crosslingual transfer performance](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4059–4066, Torino, Italia. ELRA and ICCL.
- Prithviraj Singh Shahani, Kaveh Eskandari Miandoab, and Matthias Scheutz. 2025. [Noise injection systemically degrades large language model safety guardrails](#). *Preprint*, arXiv:2505.13500.
- Oleh Shliakhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. [mGPT: Few-shot learners go multilingual](#). *Transactions of the Association for Computational Linguistics*, 12:58–79.
- Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. [Probing the emergence of cross-lingual alignment during LLM training](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.
- Andrea W Wen-Yi and David Mimno. 2023. [Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. [Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

A Escape Noise for Non-Autoregressive Models

For non-autoregressive or non-generative architectures, the notion of output sequence is not available. To obtain a comparable output for language identification, we construct a pseudo-generated sentence using a deterministic mask-and-fill procedure.

Given an input sentence x tokenized into a sequence $[v_1, \dots, v_T]$, we construct a pseudo-generated sentence $\tilde{y}(x, \sigma)$ by iterating from left to right and performing a deterministic mask-and-fill procedure. Initialize $\mathbf{z}^{(0)} \leftarrow [v_1, \dots, v_T]$. For each position $t = 1, \dots, T$, if $z_t^{(t-1)} \in \mathcal{S}$ we skip; otherwise we create a masked copy

$$\mathbf{m}^{(t)} = [z_1^{(1)}, z_2^{(2)}, \dots, z_{t-1}^{(t-1)}, [\text{MASK}], z_{t+1}^{(0)}, z_{t+2}^{(0)}, \dots, z_T^{(0)}]$$

run a forward pass under activation perturbation with noise strength σ , and fill the token at position t with the top-1 prediction:

$$\hat{v}_t = \arg \max_{v \in \mathcal{V}} p_\theta(v | \mathbf{m}^{(t)}, \sigma),$$

$$\mathbf{z}^{(t)} \leftarrow \mathbf{z}^{(t-1)} \text{ with } z_t^{(t)} = \hat{v}_t.$$

Finally, we decode $\mathbf{z}^{(T)}$ to obtain $\tilde{y}(x, \sigma)$. For non-autoregressive models, this pseudo-generated sentence $\tilde{y}(x, \sigma)$ is used as the input to the same external language identification procedure as in the main text.

B Structural Degradation Near Escape Noise

To provide a better understanding of escape noise, we conduct an auxiliary analysis examining how structural similarity between clean and perturbed outputs changes as perturbation strength increases toward the escape noise threshold.

We use MGPT and sample prompts from the FLORES development set for several languages spanning different typological families. For each language, we first estimate its escape noise value, and then apply activation perturbation with noise strength $\sigma = \alpha \cdot \sigma_{L_i}$, where α ranges from 0.1 to 1.0. We generate multiple perturbed outputs for each prompt and noise level.

To quantify structural change, we parse both the clean and perturbed outputs using a Universal Dependencies (UD) parser corresponding to the prompt language. We then compute the overlap between dependency edge multisets of the clean and

perturbed outputs. We report labeled attachment overlap (LAS overlap), defined as the fraction of dependency edges (head, dependent, relation) in the clean output that are preserved in the perturbed output. The results are shown in Figure 9. Although the absolute escape noise values vary across languages, normalizing perturbation strength by σ_{L_i} consistently brings the labeled dependency overlap to a comparably low regime across languages, supporting the role of escape noise as a meaningful normalization scale for local perturbations.

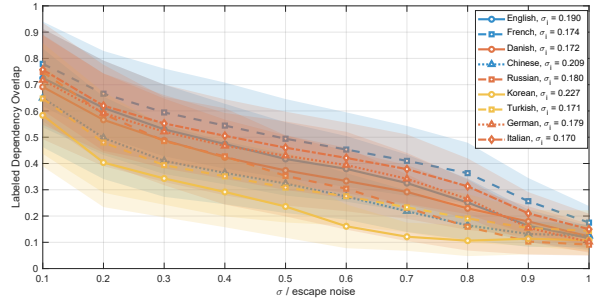


Figure 9: Average labeled dependency (LAS) overlap between clean and perturbed outputs as a function of normalized perturbation strength σ/σ_{L_i} . Results are shown for multiple languages using MGPT on FLORES.

C Language Set and Typological Diversity

Language Family	Languages
Indo-European	English, German, Danish, Swedish, Afrikaans
Indo-European	French, Italian, Spanish, Portuguese, Romanian
Indo-European	Russian, Ukrainian, Belarusian, Bulgarian, Polish
Indo-European	Greek
Indo-European	Lithuanian, Latvian
Indo-European	Hindi, Marathi, Bengali
Indo-European	Persian
Uralic	Finnish, Hungarian, Estonian
Turkic	Turkish, Azerbaijani, Uzbek, Kazakh, Kyrgyz
Afro-Asiatic	Arabic
Austronesian	Indonesian, Malay, Javanese
Dravidian	Malayalam
Sino-Tibetan	Chinese, Burmese
Japonic	Japanese
Koreanic	Korean
Tai-Kadai	Thai
Mongolic	Mongolian
Austroasiatic	Vietnamese
Niger-Congo	Swahili
Niger-Congo	Yoruba
Language Isolate	Basque

Using the typological diversity evaluation framework of Ploeger et al. (2026), we assess the intrinsic

diversity of our language coverage. Due to the absence of Grambank typological features for five languages, the evaluation is conducted on the remaining forty languages. Despite this reduction, the resulting language set exhibits high intrinsic diversity, achieving an entropy score of 0.95, a feature value inclusion (FVI) of 0.97, and a mean pairwise distance (MPD) of 0.61. These values are comparable in magnitude to those obtained by the MaxSum and MaxMin sampling strategies reported in the original study, and substantially outperform random and family-based baselines. This indicates that our language selection provides a broad and informative coverage of the typological space.

D Sensitivity of Shared-Token Weights to Corpus Choice

The shared-token weights used in our main experiments are estimated from MC4. This appendix examines whether the resulting token-language weights are sensitive to the choice of corpus.

We estimate shared token statistics using samples from different corpora. Comparing shared tokens that appear in the intersection of corpora, we find that the induced language-weight distributions are similar across FLORES, TATOEBa and MC4, as illustrated in Figure 10. This suggests that shared-token weighting is not strongly affected by the particular corpus used to estimate token frequencies. At the same time, MC4 identifies a larger number of shared tokens than FLORES or TATOEBa, due to its broader coverage and larger scale. Therefore, we use MC4-based shared-token estimates in all experiments. This choice improves token coverage while preserving language-weight distributions that are consistent with those estimated from the evaluation corpora.

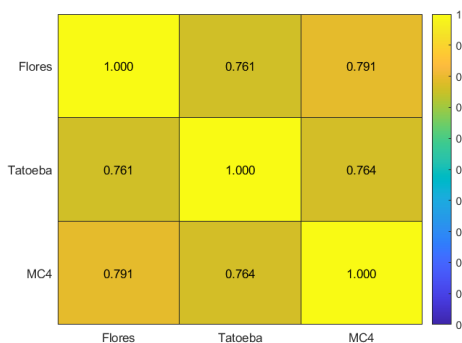


Figure 10: Mean cosine similarity between shared-token language-weight distributions estimated from different corpora (FLORES, TATOEBa, MC4), computed on the shared-token intersection.

Predictions with confidence below 0.2 are assigned an *unknown* label and excluded from language probability estimation. Table 3 summarizes the average proportion of tokens assigned to each category across all 10 model–corpus runs. Shared tokens account for the largest proportion of tokens on average, while low-confidence predictions remain a minority of tokens, suggesting that the majority of tokens receive stable language assignments under the XLM-V labeling scheme.

Category	Mean (%)	Std (%)
Shared Tokens	44.35	9.07
Hard LID Tokens	29.55	4.45
Unknown Tokens	21.26	6.23
Symbols	4.83	2.29

Table 3: Average token-category proportions across all 10 model–corpus runs. “Hard LID” denotes tokens assigned a single language label by XLM-V, while “Unknown” denotes low-confidence predictions excluded from language probability estimation.

E Escape Noise Rankings Across Languages

Figure 11 shows the Spearman rank correlations between escape-noise rankings across multilingual language models. Overall, the correlations remain consistently high, indicating that escape-noise ordering is relatively stable across architectures. Table 4 further reports the mean escape-noise values averaged across five multilingual language models, revealing substantial variation across languages.

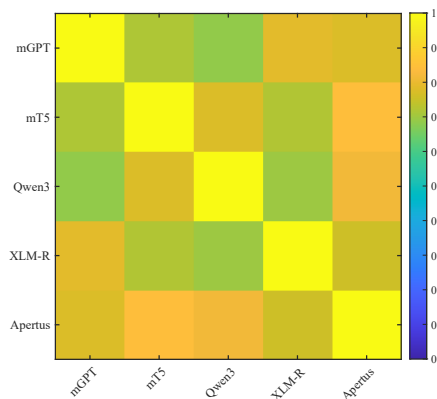


Figure 11: Spearman rank correlations between escape-noise rankings across multilingual language models.

Language	EN	Language	EN
Burmese	0.184	Finnish	0.136
Korean	0.182	German	0.135
Thai	0.182	Hungarian	0.135
Japanese	0.180	Turkish	0.135
Chinese	0.178	Italian	0.134
Arabic	0.175	Kyrgyz	0.134
Bengali	0.174	Latvian	0.133
Malayalam	0.174	Marathi	0.133
Persian	0.172	Lithuanian	0.132
Greek	0.163	Swedish	0.132
Vietnamese	0.152	Yoruba	0.131
Ukrainian	0.149	Afrikaans	0.127
Mongolian	0.147	Azerbaijani	0.124
English	0.146	Portuguese	0.122
Kazakh	0.145	Swahili	0.121
Russian	0.143	Spanish	0.118
French	0.141	Uzbek	0.117
Romanian	0.139	Danish	0.116
Hindi	0.138	Estonian	0.112
Belarusian	0.138	Javanese	0.110
Polish	0.138	Indonesian	0.097
Bulgarian	0.138	Malay	0.075

Table 4: Mean escape-noise (EN) values averaged across five mLMs.

F Language Tree Clustering

To compare TBLS with known language genealogy, we reconstruct a language tree from the similarity matrix. First, similarity scores are converted into distances as

$$D_{ij} = 1 - \text{TBLS}(L_i, L_j).$$

We then apply agglomerative hierarchical clustering with average linkage to obtain a dendrogram whose leaves correspond to languages. Starting from singleton clusters, the algorithm iteratively merges the pair of clusters with the smallest average pairwise distance.

Because hierarchical clustering may produce artificially deep binary branches when merge heights are nearly identical, we optionally collapse consecutive merges whose heights differ by less than a tolerance τ . This converts near-ties into multi-branch nodes. Setting $\tau = 0$ recovers the original binary dendrogram.

Tree representation. To compare the reconstructed tree with the gold genealogical tree, we ignore branch lengths and compute the number of edges on the shortest path between each pair of languages in the tree. These pairwise leaf distances form a vector representation of the tree structure.

Tree similarity. Let T_{TBLS} denote the tree reconstructed from TBLS and T_{gold} the reference genealogical tree restricted to the same language set.

Following (Rabinovich et al., 2017), we compare the two trees using the following metric:

$$\text{Distance} = \frac{\frac{1}{m} \|\mathbf{v}_{T_{\text{TBLS}}} - \mathbf{v}_{T_{\text{gold}}}\|_2^2}{(\max(\mathbf{v}_{T_{\text{gold}}}))^2 + \epsilon},$$

where $m = n(n-1)/2$.

G Static language similarity measures

We compare TBLS with three static representation-based language similarity measures: REP, Centroid language similarity, and ANC. All three are computed from hidden representations of the same multilingual models and at each transformer layer. For MGPT, QWEN3, APERTUS, and MT5, representations are extracted from the decoder; for encoder-only models such as XLM-R, they are extracted from the encoder.

Given a sentence $x_i^{(L_j)}$ in language L_j , we obtain the hidden states at layer t and mean-pool over non-padding tokens:

$$\mathbf{h}_{i,L_j}^{(t)} \in \mathbb{R}^d.$$

For FLORES, sentence index i is shared across languages. For the sampled TATOEBAS subset, the data is not truly parallel; to keep the comparison aligned with the TBLS setting, we construct a pseudo-parallel ordering by matching per-language samples by index and apply the same computations.

REP. REP measures language similarity by comparing sentence representations for aligned sentence indices. This corresponds to the core representation-similarity computation used in mPLM-Sim (Lin et al., 2024); we retain the broader label REP because the same formulation also appears in earlier multilingual probing work (Rama et al., 2020). For each layer t and sentence index i , we compute

$$S_{\text{REP},i}^{(t)}(L_j, L_k) = \cos(\mathbf{h}_{i,L_j}^{(t)}, \mathbf{h}_{i,L_k}^{(t)}).$$

The final layer-wise similarity matrix is the average over sentence indices.

CentroidSim. CentroidSim first summarizes each language by its mean representation at layer t :

$$\boldsymbol{\mu}_{L_j}^{(t)} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}_{i,L_j}^{(t)}.$$

Language similarity is then the cosine similarity between language centroids:

$$S_{\text{centroid}}^{(t)}(L_j, L_k) = \cos(\boldsymbol{\mu}_{L_j}^{(t)}, \boldsymbol{\mu}_{L_k}^{(t)}).$$

ANC. ANC compares languages by the correlation structure of their hidden dimensions. For each language L_j , let

$$\mathbf{H}_{L_j}^{(t)} = \left[\mathbf{h}_{1,L_j}^{(t)}, \dots, \mathbf{h}_{N,L_j}^{(t)} \right]^\top \in \mathbb{R}^{N \times d}$$

be the matrix of sentence representations at layer t . ANC computes the Pearson correlation between corresponding hidden dimensions across languages and averages over dimensions:

$$S_{\text{ANC}}^{(t)}(L_j, L_k) = \frac{1}{d} \sum_{r=1}^d \rho\left(\mathbf{H}_{L_j}^{(t)}[:, r], \mathbf{H}_{L_k}^{(t)}[:, r]\right),$$

where $\rho(\cdot, \cdot)$ denotes Pearson correlation.

For all three measures, we obtain one language similarity matrix per model and layer. When comparing models with different depths, we align layers by relative depth. For example, a relative depth of 0.5 corresponds to the middle layer of each model.

H Script-based Orthographic Similarity

To estimate orthographic similarity between languages, we construct a script-based similarity matrix using Unicode character distributions. For each language in FLORES, we collect all available sentences and count the frequency of characters belonging to major Unicode script categories. These counts are normalized to form a probability distribution over scripts for each language. Pairwise language similarity is then computed using cosine similarity between the resulting distributions.

We compare this orthographic similarity matrix with the TBLS matrix using RSA. Table 5 reports the correlations across different models and corpora.

Model	RSA
mGPT-Flores	0.572
mGPT-Tatoeba	0.584
mT5-Flores	0.614
mT5-Tatoeba	0.584
Qwen3-Flores	0.598
Qwen3-Tatoeba	0.602
XLM-R-Flores	0.554
XLM-R-Tatoeba	0.553
Apertus-Tatoeba	0.558
Apertus-Flores	0.584

Table 5: RSA between TBLS matrices and the script-based orthographic similarity matrix.

Orthographic Correlation within Latin-script Languages

To test whether TBLS is primarily driven by script similarity, we repeat the orthographic similarity analysis using only languages that employ the Latin script. This removes cross-script differences and focuses on orthographic variation within a single writing system. The results are shown in Table 6. Correlations are close to zero for most models, indicating that TBLS is not strongly explained by orthographic similarity once script differences are removed.

Model	RSA
mGPT-Flores	-0.029
mGPT-Tatoeba	-0.017
mT5-Flores	0.023
mT5-Tatoeba	-0.001
Qwen3-Flores	-0.046
Qwen3-Tatoeba	0.048
XLM-R-Flores	0.065
XLM-R-Tatoeba	0.061
Apertus-Tatoeba	0.104
Apertus-Flores	0.196

Table 6: RSA between TBLS matrices and script-based orthographic similarity for Latin-script languages only.

I Cross-lingual Transfer Experimental Details

For all tasks, the source languages are: en, fr, es, de, el, ru, ar, hi, tr, vi, zh, sw. Yoruba is excluded from the target language because it is not included in the XLM-R pretraining data. We use HuggingFace Trainer with standard optimization settings. Table 7 summarizes the task-specific hyperparameters.

Task	Epochs	LR	Train BS	Eval BS
MASSIVE	8	2e-5	16	32
XNLI	5	2e-5	16	32
WikiAnn NER	8	2e-5	16	16
UD POS	10	2e-5	16	16

Table 7: Task-specific fine-tuning hyperparameters for zero-shot transfer experiments.

The Arabic MASSIVE model uses a slightly different configuration with 20 training epochs, learning rate 3×10^{-5} , train batch size 16, evaluation batch size 32, maximum sequence length 128.

Across all tasks, we use weight decay 0.01, warmup ratio 0.06, epoch-based evaluation and checkpoint selection, and retain only the best checkpoint according to the task metric.

J Layer-wise Statistics for Representation-based Similarity Measures

For representation-based similarity measures (REP, CentroidSim, and ANC), the main text reports layer-averaged RSA values when evaluating cross-lingual transfer prediction. We choose this setting because zero-shot transferability is intended as a prior signal for selecting transferable language pairs before downstream adaptation. In practice, the optimal representation layer is generally unknown in advance and varies substantially across tasks and models. By contrast, TBLS is defined as a single transition-based similarity measure rather than a layer-wise family of similarities. Reporting only the best-performing representation layer would therefore introduce an unfair comparison favoring methods with layer-selection flexibility.

Nevertheless, to characterize the variability of representation-based similarity across layers, we additionally report the best- and worst-performing layers for each task, together with their corresponding RSA values in Table 8.

(a) MASSIVE					
Method	Mean	MinL	Min	MaxL	Max
REP	0.668	1	0.468	15	0.771
Centroid	0.641	1	0.456	16	0.735
ANC	0.765	1	0.684	23	0.852

(b) WikiAnn NER					
Method	Mean	MinL	Min	MaxL	Max
REP	0.322	24	-0.043	5	0.478
Centroid	0.267	24	-0.085	3	0.404
ANC	0.329	24	0.159	18	0.418

(c) UD POS					
Method	Mean	MinL	Min	MaxL	Max
REP	0.346	24	0.001	10	0.467
Centroid	0.333	24	-0.008	10	0.509
ANC	0.326	24	0.156	18	0.410

(d) XNLI					
Method	Mean	MinL	Min	MaxL	Max
REP	0.665	22	0.367	12	0.816
Centroid	0.581	22	0.249	12	0.730
ANC	0.782	1	0.669	18	0.850

Table 8: Layer-wise RSA statistics for representation-based similarity measures on cross-lingual transfer prediction tasks. MinL/MaxL denote the worst- and best-performing layers respectively.

Metric	MASSIVE	NER	POS	XNLI
Overlap	0.057	0.225	0.202	0.114
TBLS	0.442	0.329	0.311	0.487
TBLS-CS	0.440	0.002	0.028	0.341
Overlap-CS	NA	NA	NA	NA

Table 9: RSA between language similarity measures and downstream transferability for target languages. CS denotes cross-script evaluation.

K Tokenizer-Controlled Analysis

To further examine whether TBLS mainly reflects tokenizer overlap, we construct a tokenizer-overlap baseline using weighted Jaccard similarity between tokenizer frequency distributions:

$$\text{TokenOverlap}(L_i, L_j) = \frac{\sum_v \min(f_i(v), f_j(v))}{\sum_v \max(f_i(v), f_j(v))},$$

where $f_i(v)$ denotes the empirical frequency of token v in language L_i estimated from the evaluation corpus. Interestingly, tokenizer overlap itself is moderately associated with gold genealogical similarity (Spearman RSA = 0.356), suggesting that token overlap may capture genuine lexical and orthographic relatedness in addition to introducing a potential surface-form bias.

We then evaluate transferability separately under cross-script settings. Under cross-script evaluation, tokenizer overlap becomes nearly constant because token overlap between language pairs approaches zero, resulting in undefined Spearman RSA values (reported as NA). The results are shown in Table 9.

Token overlap exhibits nontrivial correlations on lexical token-level tasks, especially POS and NER. However, TBLS shows substantially stronger transfer correlations overall. Under cross-script evaluation, tokenizer-overlap RSA becomes undefined because overlap scores collapse to a near-constant vector, whereas TBLS remains computable and retains clear transfer signal on sentence-level tasks such as MASSIVE and XNLI.

L Layer-wise average RSA of static language similarity

Figure 12 shows the layer-wise average RSA between static language similarity matrices across models. To enable comparison across models with different numbers of transformer layers, we align layers by relative depth.

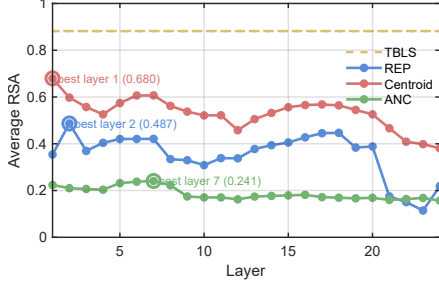


Figure 12: Layer-wise average RSA between static language similarity across models on FLORES. The dashed line indicates the corresponding average RSA between TBS matrices (0.882) for reference.

M RSA Heatmap of TBS

This appendix visualizes pairwise consistency between TBS matrices across all model–corpus settings, as shown in Figure 13.

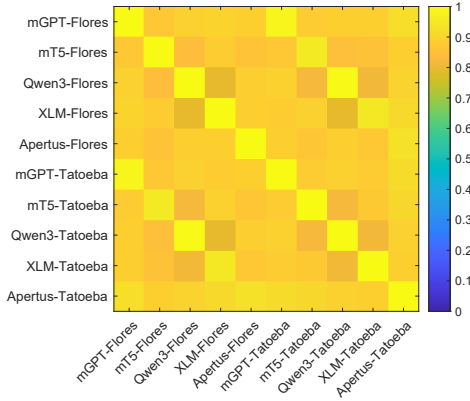


Figure 13: RSA between TBS matrices across models and corpora.

N Robustness to Token Assignment

We recompute TBS by assigning each token a single hard language label predicted by XLM-V (Liang et al., 2023). Figure 14 shows the resulting RSA between TBS matrices across models and corpora.

An interesting observation concerns the state-of-the-art multilingual model APERTUS. While its agreement with other models on FLORES is low, it shows strong consistency between FLORES and TATOEB. This behavior may reflect the broader multilingual coverage and larger scale of recent multilingual LLMs, which could lead to language representations that are less sensitive to corpus-specific variation.

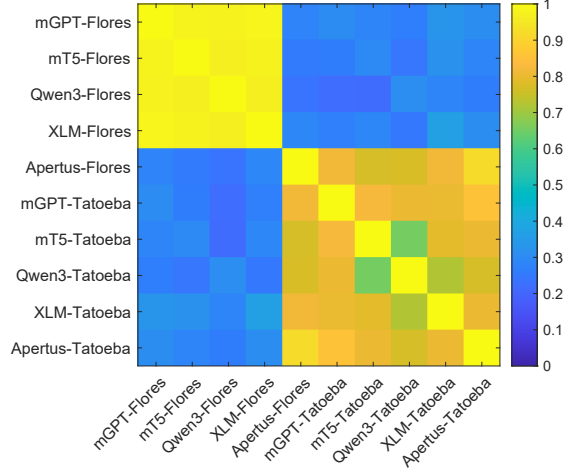


Figure 14: RSA between TBS matrices computed using hard token assignments from XLM-V.

O Resource-level Robustness

A potential concern is that TBS consistency may primarily reflect high-resource languages that are better represented during multilingual pretraining. To evaluate this possibility, we rank languages according to their relative representation in the public MGPT pretraining distribution and divide them into three equally sized bins corresponding to high-, medium-, and low-resource languages. For each resource group, we recompute the RSA analysis and results are summarized in Table 10. TBS remains highly consistent for both high-resource and low-resource languages across models and corpora.

Resource bin	(A)	(B)	(C)
High	0.981	0.915	0.912
Medium	0.887	0.678	0.647
Low	0.965	0.870	0.867

Table 10: Average RSA between TBS matrices within different language resource groups. (A) same-model cross-corpus, (B) cross-model same-corpus, and (C) cross-model cross-corpus.

P Robustness to Perturbation Seeds

Because activation perturbation introduces stochasticity through independently sampled Gaussian noise, we evaluate the sensitivity of TBS to perturbation seeds on mGPT–FLORES. Using 10 independent seeds, the average Spearman correlation between seed-specific sample-level transition distributions is 0.931. After aggregating transition distributions across 100 samples, this correlation increases to 0.994. The aggregation is already rel-

atively stable with only five samples (Spearman = 0.961) and progressively converges to 0.994 at 100 samples.

Escape-noise calibration is similarly stable across seeds: the mean language-level range of calibrated escape-noise values is 0.0069, with a mean coefficient of variation of 2.95%. Finally, we independently recompute the complete mGPT-FLORES TBLS matrix under five perturbation seeds. The resulting matrices achieve a mean pairwise matrix-rank RSA of 0.990 with a standard deviation of 0.006. These results indicate that, although individual perturbation outcomes are stochastic, the aggregated language-level transition structure underlying TBLS is highly stable across perturbation seeds.

Q Robustness to Token Position and Noise Type

Robustness to token position. Tokens appearing at different decoding positions typically carry different semantic content. For example, sentence-initial tokens in English are often dominated by common function words, whereas later tokens tend to be more semantically diverse. If TBLS mainly captures stable language-level structure, perturbing different decoding positions should produce similar TBLS matrices.

Figure 15 reports RSA between TBLS matrices computed by perturbing different decoding positions. The resulting similarities remain consistently high across all tested positions, indicating that TBLS is not strongly dependent on the semantic role of individual tokens.

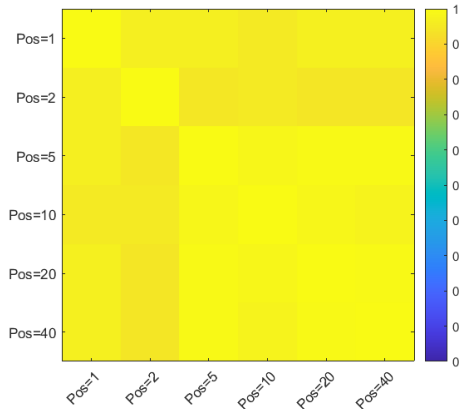


Figure 15: Representational similarity analysis between TBLS matrices computed by perturbing different decoding positions ($p = 1, 2, 5, 10, 20, 40$).

Robustness to noise distribution. We additionally evaluate whether TBLS depends on the particular perturbation distribution used during activation perturbation. Besides Gaussian noise used in the main experiments, we compute TBLS using Uniform and Laplace perturbations.

As shown in Figure 16, TBLS matrices derived from different noise distributions remain highly similar. This suggests that TBLS primarily reflects stable multilingual transition behavior rather than artifacts of a specific perturbation distribution.

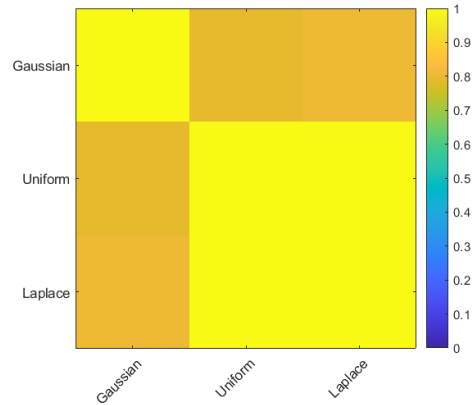


Figure 16: RSA between TBLS matrices computed under different noise distributions.

R Computational Cost of Escape-Noise Calibration

Table 11 reports the empirical average number of forward passes required per escape-noise calibration example on FLORES.

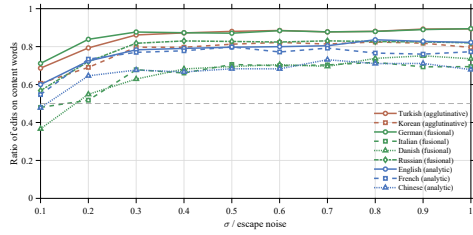
Model	Forward passes / example
mGPT	445
XLM-R	170
Qwen3	155
Apertus	255
mT5	370

Table 11: Average number of forward passes per example during escape-noise calibration on FLORES.

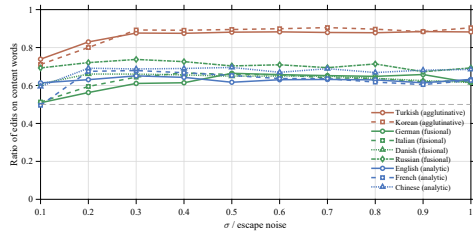
The calibration cost varies across architectures. Qwen3 and XLM-R require fewer forward passes on average, whereas mGPT and mT5 are relatively more expensive. This difference mainly arises because examples for the latter models tend to escape at higher noise levels under their respective search grids, requiring more search steps before the first transition. Tatoeba sentences are generally shorter than those in FLORES, and their computational cost is therefore typically lower.

S Content vs. Function Word Sensitivity Across Models

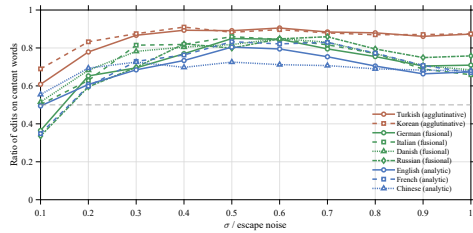
Figure 17 shows the ratio of content-word edits under increasing perturbation strength across different mLMs on FLORES. Across architectures, perturbation consistently affects lexical content more strongly than function words, suggesting that local perturbations tend to preserve grammatical structure while altering semantic content.



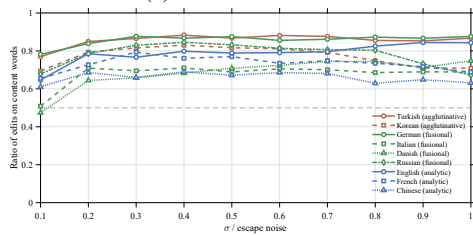
(a) MT5 LARGE



(b) QWEN3



(c) XLM-R-LARGE



(d) APERTUS

Figure 17: Ratio of content words among perturbed tokens as a function of noise strength (σ normalized by escape noise), shown for several languages across models on FLORES. Results across models and corpora exhibit a consistent pattern, with perturbation primarily affecting lexical content rather than function words.