The Geometric Cost of Adaptation: How Typological Proximity Predicts Representational Reorganization in Multilingual Fine-Tuning

Anonymous ACL submission

Abstract

When a multilingual language model is fine-tuned on a single language, what happens to its representations of every other language it knows? We study this question through neighborhood deformation — the degree to which k-nearest-neighbor structure in non-target languages is disrupted after fine-tuning — measured via the neighborhood overlap metric of Doimo et al. (2020) across 1,011 semantically aligned concepts from NorthEuraLex in five probe languages. We evaluate LLaMA 3.1 8B fine-tuned on each of 15 typologically diverse languages and validate on Gemma 2 2B under both QLoRA and full fine-tuning.

We find that typological proximity consistently predicts stronger neighborhood deformation. Using eleven distance measures from the DistaLs toolkit (Goot et al., 2025) spanning phylogenetic, lexical, phonological, typological, writing system, geographic, and textual dimensions, we find that all show negative coefficients and ten of eleven are significant at $p < .01$. A composite typological distance predictor `lang2vec` (Littell et al., 2017) produces $\beta = -0.178$, $p < 10^{-12}$ on LLaMA 3.1 8B, replicating on Gemma 2 2B under QLoRA ($\beta = -0.149$, $p < 10^{-5}$) and full fine-tuning ($\beta = -0.202$, $p = .001$). This direction contrasts with both behavioral cross-lingual transfer studies, where typological proximity typically facilitates rather than disrupts transfer (Lauscher et al., 2020), and prior representational work on encoder models (Philippy et al., 2023), suggesting that the relationship between typological distance and fine-tuning collateral effects depends on model architecture and the level of analysis.

1 Introduction

When a multilingual language model is fine-tuned on a single language, what happens to its representations of every other language it knows? Prior work has answered this question behaviorally: fine-tuning can degrade performance on untargeted languages (Yadavalli et al., 2023; Oba et al., 2023), and typological proximity between source and target languages has been shown to facilitate cross-lingual transfer at the task-performance level (Lauscher et al., 2020). Separately, a line of work has characterized the static geometry of multilingual representations, showing that languages occupy distinct but structured subspaces (Chang et al., 2022; Shah et al., 2023). More recently, Philippy et al. (2023) showed that genetic distance between languages predicts CKA-based representational impact on mBERT (Devlin et al., 2019) during fine-tuning, suggesting a link between typological structure and geometric change. However, it is not known how this extends to modern decoder-only LLMs under parameter-efficient methods such as LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023), and whether the effect generalizes across a broader set of typological distance dimensions.

Understanding these geometric changes has both practical and scientific motivation. Multilingual models are routinely adapted to new languages through continued fine-tuning, but adaptation is almost always evaluated through downstream task performance without considering internal representational structure. Fine-tuning can reshape representations in ways that task performance does not reflect (Merchant et al., 2020), and standard similarity metrics can miss significant geometric differences entirely (Kornblith et al., 2019). Closing this evaluation gap is particularly important given the scale at which multilingual models are now adapted: practitioners fine-tune on new languages without knowing which existing languages' representations are being reorganized, or how much. A model can improve on the target language while substantially reorganizing its geometry for other languages. If geometric reorganization follows predictable typological patterns, practitioners need

to anticipate which languages' representations are most at risk when adapting a model to a new language.

Our research question is: does typological proximity to the target predict the degree of representational reorganization in non-target languages under QLoRA adaptation? We address it by applying QLoRA-based continued training of LLaMA 3.1 8B (Grattafiori et al., 2024) and Gemma 2 2B (Team et al., 2024) on 15 typologically diverse languages from seven families and one language isolate, and measure how the k-nearest-neighbor structure of 1,011 semantically aligned concepts changes in five probe languages. Following Doimo et al. (2020), we quantify change via neighborhood overlap (χ_k), which is the mean fraction of each concept's k nearest neighbors preserved between base and fine-tuned representations. To directly measure the representational reorganization caused by fine-tuning, we define **neighborhood deformation** as $1 - \chi_k$. Intuitively, a neighborhood deformation value near zero indicates that fine-tuning left each concept with the same set of nearest neighbors it had in the base model. A value near one indicates that the neighborhood structure has changed completely, even if individual embedding magnitudes or pairwise distances have not changed as much.

Typological distance is a natural candidate predictor of this deformation. It is widely used in cross-lingual transfer research as a predictor of model performance (Patankar et al., 2022; Ng et al., 2026), and if fine-tuning on a language redistributes probability mass over tokens and features that a language shares with others, the degree of sharing should scale with linguistic relatedness. Languages that share vocabulary, surface forms, or typological features with the fine-tuning target use overlapping regions of the embedding space (Chang et al., 2022); gradient updates to those regions will inevitably perturb their neighborhood structure. Typological distance operationalizes exactly this notion of overlap. We operationalize typological distance using eleven measures from the DistaLs benchmark (Goot et al., 2025), spanning phylogenetic, lexical, phonological, typological, writing system, geographic, and textual dimensions, with `lang2vec` (Littell et al., 2017) serving as our primary composite predictor.

Summary of Results and Contributions. We introduce a framework to quantify how monolingual fine-tuning reshapes the representational geometry of non-target languages, adapting the neighborhood overlap methodology of Doimo et al. (2020) to the cross-lingual setting. Applying this framework to LLaMA 3.1 8B fine-tuned on 15 typologically diverse languages, we find that typological distance strongly predicts neighborhood deformation: closer languages undergo greater neighborhood deformation. This effect holds across all eleven distance measures spanning phylogenetic, lexical, phonological, typological, writing system, geographic, and textual dimensions (all $\beta < 0$; 10/11 significant at $p < .01$), and replicates on Gemma 2 2B under QLoRA ($\beta = -0.149$, $p < 10^{-5}$) and full fine-tuning ($\beta = -0.202$, $p = .001$). Importantly, this finding diverges from both behavioral cross-lingual transfer studies (Lauscher et al., 2020) and prior representational work (Philippy et al., 2023), where typological proximity facilitates rather than intensifies collateral change.

2 Experimental Setup

All code, distance matrices, and experimental results reported in this paper are available at [anonymized].

2.1 Models and fine-tuning

We apply QLoRA continued training (Dettmers et al., 2023) to two pretrained multilingual models: LLaMA 3.1 8B and Gemma 2 2B. LoRA adapters are applied to all attention and MLP projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with rank r=16, scaling factor α=32, and dropout 0.05. Base models are loaded in 4-bit NF4 quantization with double quantization. Training uses AdamW with learning rate 2×10^{-4}, cosine schedule with 3% warmup, weight decay 0.01, and effective batch size 16 via gradient accumulation. Sequences are packed to length 512.

For each fine-tuning language, training data is drawn from the language-specific partition of Multilingual C4 (mC4) dataset (Xue et al., 2021). To ensure comparability across fine-tuning languages, we fix the training budget at ≈24.6M tokens per language (≈3,000 steps at effective batch size 16 and sequence length 512), isolating typological effects from training-volume effects. QLoRA is applied to both models on 15 typologically diverse languages spanning seven families and one isolate (Indo-European: English, German, French, Catalan, Romanian, Polish; Uralic: Finnish, Hungar-

ian; Turkic: Turkish; Austroasiatic: Vietnamese, Khmer; Sino-Tibetan: Mandarin; Afro-Asiatic: Amharic; Kartvelian: Georgian; Isolate: Basque), written in five scripts (Latin, Hanzi, Khmer, Ge'ez, Georgian).

To assess whether effects depend on the fine-tuning regime, we additionally perform full-parameter fine-tuning on Gemma 2 2B for a subset of three languages (English, French, Polish), selected to span typological diversity, but remaining well-represented in the model's pretraining data to eliminate low-resource effects. All model weights are updated under identical training conditions. This provides a comparison under full fine-tuning, where more parameters are updated; we do not treat it as a primary cross-method comparison given the smaller language sample.

2.2 Concept-aligned probing

To compare representations across fine-tuning conditions, we use the NorthEuraLex lexical database (Dellert et al., 2019), which provides 1,016 translation-aligned lexical entries across a large panel of languages for a fixed concept inventory covering body parts, basic verbs, kinship terms, natural-world items, and abstract states. After filtering for concepts with coverage in all five probe languages and removing entries with ambiguous lexical mappings, we retain 1,011 concepts per language.

Probe languages. Deformation is measured in five probe languages held constant across all conditions: English, German, French, Spanish, and Polish, covering three Indo-European subfamilies (Germanic, Romance, Slavic), all well-represented in both models' pretraining data (Grattafiori et al., 2024; Team et al., 2024). These languages are also in the fine-tuning language set, producing five 'self' pairs (e.g. fine-tuning on English, probing English). We retain these for completeness; excluding them does not change the main result for either QLoRA condition (LLaMA: $\beta = -0.147$, $p = .001$; Gemma: $\beta = -0.186$, $p = .001$; $n = 71$ each; Appendix D).

Embedding extraction. For each concept, we construct a short prompt from a language-specific template of the form *'[word]' means '[gloss]'*, where *word* is the target-language lexical entry and *gloss* is its English translation from NorthEuraLex (e.g. *'Hund' means 'dog'* for German; full templates in Appendix E). Since the English gloss is identical in the base and fine-tuned conditions, it serves to reduce polysemy ambiguity and ensure prompt-level consistency across base and fine-tuned conditions. The embedding for each concept is taken only from the final layer: we locate the target word's subword token span via character offsets, mean-pool the corresponding hidden states, and average the result across two syntactically distinct templates per concept to reduce template-specific artifacts. We do not pool across layers.

2.3 Measuring neighborhood deformation

Following Doimo et al. (2020), we quantify the change between two sets of representations via neighborhood overlap. Doimo et al. defined this metric between layers of a single network using Euclidean distance; we adapt it to compare representations from two different models (base vs. fine-tuned) of the same concepts, using cosine distance on L_2-normalized embeddings. Cosine distance is standard for comparing language model representations, as it measures angular similarity independently of embedding magnitude.

Let $\mathbf{X}^{(B)}$ and $\mathbf{X}^{(F)}$ denote the $N \times d$ matrices of base and fine-tuned embeddings for the same N=1,011 concepts in a given probe language, where d is the hidden-state dimensionality (d=4,096 for LLaMA 3.1 8B, d=2,304 for Gemma 2 2B). Let $\mathcal{N}_k^{(B)}(i)$ and $\mathcal{N}_k^{(F)}(i)$ be the sets of k nearest neighbors of concept i under cosine distance in the base and fine-tuned representations respectively, and let $A^{(B)}$ and $A^{(F)}$ be the corresponding $N \times N$ adjacency matrices with entries $A_{ij}^{(B)} = 1$ if $j \in \mathcal{N}_k^{(B)}(i)$ and 0 otherwise. Neighborhood overlap is

$$\chi_k = \frac{1}{N} \sum_i \frac{1}{k} \sum_j A_{ij}^{(B)} \, A_{ij}^{(F)},$$

the average fraction of common neighbors between the two representations. We define **neighborhood deformation** as $1 - \chi_k$, resulting in a value in $[0, 1]$ where 0 indicates perfect neighborhood preservation and 1 indicates complete reshuffling. We set k=10 throughout and verify on $k \in \{5, 20, 50\}$ in Appendix C.

Each fine-tuning condition gives out one deformation score per (fine-tuning language, probe language) pair, for a total of $15 \times 5 = 75$ scores per model $\times$ method combination (and $3 \times 5 = 15$ for Gemma full fine-tuning).

2.4 Typological distance predictors

To test whether deformation is systematically related to linguistic similarity, we use eleven distance measures between each (fine-tuning, probe) language pair using the DistaLs toolkit (Goot et al., 2025). DistaLs aggregates language information from multiple sources and normalizes all distances to $[0, 1]$. We select measures spanning seven categories of linguistic similarity: typological, phylogenetic, geographic, lexical, phonological, writing system, and textual, excluding non-linguistic metadata (Wikipedia size, speaker population, endangerment status) as well as Grambank morphosyntactic features (Skirgård et al., 2023), which lack coverage for 38 of 75 fine-tuning probe pairs.

Primary predictor: typological distance. `lang2vec` computes cosine distance over concatenated syntax, phonology, and inventory feature vectors from the URIEL database (Littell et al., 2017), averaging across multiple data sources (WALS, SSWL, Ethnologue, PHOIBLE). We use `lang2vec` as our primary typological predictor since it is a composite of syntactic, phonological, and inventory features. It captures typological relatedness across multiple linguistic dimensions simultaneously and is widely used as a reference distance in cross-lingual NLP. We do not claim it is empirically the strongest predictor across all conditions; Section 3.2 shows that predictor rankings shift across model conditions.

Phylogenetic and geographic distance. `glot_tree` measures phylogenetic distance in the Glottolog language family tree (Hammarström et al., 2024): languages in different families receive maximal distance (1.0); languages in the same tree are scored by the proportion of shared edges relative to tree depth. `geo` measures geographic distance between language centroids, also derived from Glottolog.

Lexical distance. `asjp` computes the average normalized Levenshtein distance over phonetically encoded Swadesh-list concepts from the Automated Similarity Judgment Program (asj, 2022). `concepts` measures cosine distance over concept representations from the Conceptualizer (Liu et al., 2023), a bipartite graph linking concepts to their translations across languages.

Phonological distance. `phoible` measures phoneme inventory distance based on the overlap of IPA segments in the PHOIBLE database (Moran and McCloy, 2019).

Writing system distance. `scripts` measures writing system overlap from LinguaMeta (Ritchie et al., 2024) and GlotScript (Kargaran et al., 2024).

Textual distance. Four measures are derived from raw text corpora (primarily Wikipedia and Bible translations) via the LTI LangID corpus (Brown, 2014). `textcat` computes divergence between ranked character n-gram frequency profiles, following the text categorization method of Cavnar and Trenkle (1994) with the 400 most frequent 1–5 character n-grams. `char_JSD` measures Jensen–Shannon divergence between character frequency distributions. `whitespace` and `punctuation` measure divergence in whitespace and punctuation character proportions respectively, capturing differences in word segmentation and orthographic conventions across writing systems.

2.5 Statistical analysis

Each fine-tuning condition yields $15 \times 5 = 75$ observations (one per fine-tuning–probe language pair), nested within 15 fine-tuning languages. We model deformation with a linear mixed-effects regression:

```
deformation ~ lang2vec_distance + probe_lang + (1 | ft_lang)
```

where `lang2vec_distance` is the `lang2vec` typological distance between the fine-tuning and probe languages, `probe_lang` is a categorical fixed effect for probing languages (reference: English), and `(1 | ft_lang)` is a random intercept for fine-tuning language. Models are fit with REML via `statsmodels` (Seabold and Perktold, 2010). The coefficient of interest is $\beta_{\texttt{lang2vec}}$: a negative value indicates that typologically closer languages suffer greater deformation. We report marginal R^2_m and conditional R^2_c following Nakagawa and Schielzeth (2013), where R^2_m reflects variance explained by fixed effects alone and R^2_c reflects variance explained by the full model including random effects.

3 Results and Discussion

We present our findings as follows. We first establish the central result: typological relatedness predicts neighborhood deformation across all three fine-tuning conditions (Section 3.1). Second, we show that this finding holds across all eleven typological distances (Section 3.2). Third, we examine

per-language patterns (Section 3.3). Fourth, we confirm robustness to the choice of neighborhood size k (Section 3.4). Finally, we relate our findings to prior work on behavioral and representational cross-lingual transfer (Section 3.5).

3.1 Typological relatedness predicts neighborhood deformation

For each of the three fine-tuning conditions, we fit the mixed-effects model specified in Section 2.5 with `lang2vec` typological distance as the primary fixed effect. Table 1 reports the results. In all three conditions the coefficient is negative and significant: greater typological distance corresponds to less deformation. Figure 1 visualizes the three coefficients with 95% confidence intervals; all lie entirely below zero.

First, the full fine-tuning coefficient ($\beta = -0.202$) is numerically larger than the QLoRA coefficient on the same model ($\beta = -0.149$), though the confidence intervals overlap and we do not treat this as a reliable difference. Second, all three conditions show overlapping confidence intervals, suggesting the effect is consistent in magnitude across both models and fine-tuning methods despite differences in architecture, parameter count, and training procedure.

3.2 Convergent evidence across distance measures

To verify that our main result is not an artifact of the `lang2vec` predictor specifically, we re-fit the mixed-effects model on LLaMA 3.1 8B using each of eleven individual DistaLs distance measures as the sole typological predictor. Table 2 reports the six strongest predictors; the full set of eleven is in Appendix F, Table 7.

All eleven measures have negative coefficients, and ten of eleven are significant at $p < .01$. The top five predictors (`lang2vec`, `glot_tree`, `asjp`, `scripts`, `textcat`) have $|z|$-statistics between 6.85 and 7.30 on LLaMA 3.1 8B, a too narrow range to meaningfully distinguish their predictive contributions. On Gemma 2 2B, `scripts` and `char_JSD` rank ahead of `lang2vec` (Appendix F), consistent with the collinearity account: no single predictor is stably dominant, and the result is best attributed to linguistic relatedness broadly. Their pairwise correlations are high ($r = 0.80$–0.91 among `lang2vec`, `glot_tree`, `asjp`, and `textcat`; $r \approx 0.4$ for `scripts`), and when entered jointly into the model, none retains individual significance ($p \geq 0.24$), confirming that they index a shared latent construct of linguistic relatedness at this panel size.

This is unsurprising: languages that are phylogenetically related also tend to share vocabulary, typological features, and writing system properties. Our finding is therefore best stated at the level of linguistic relatedness broadly rather than attributed to any one of its specific dimensions.

3.3 Per-language deformation patterns

Figure 3 shows the relationship between mean `lang2vec` distance and mean neighborhood deformation for each fine-tuning language on LLaMA 3.1 8B. The negative trend is visible: fine-tuning languages typologically close to the probes (left side) produce greater deformation than distant ones (right side). Table 3 reports the same data numerically, including the relative deformation rankings for both models. Figure 2 visualises the full 15×5 deformation matrix for both QLoRA conditions.

The ranking follows a structured pattern. The seven most disruptive fine-tuning languages are all written in Latin script and share either vocabulary or grammatical features with one or more probe languages. The three least disruptive — Khmer, Amharic, and Georgian — use distinct scripts and belong to families with minimal lexical or typological overlap with the probes. The range is substantial: Basque produces 73% more deformation than Khmer (0.222 vs. 0.128), despite both being non-Indo-European languages in the fine-tuning set.

Noteworthy individual cases. Two individual languages are particularly informative. English produces less deformation (0.140) than its typological proximity to the probes would predict. We interpret this as a pretraining-saturation effect: LLaMA was trained predominantly on English text (Grattafiori et al., 2024), so QLoRA adaptation on additional English data adds relatively little new information. Vietnamese occupies an informative middle ground. Phylogenetically distant from all probes (Austroasiatic family) but written in Latin script with diacritics, Vietnamese produces deformation (0.170) comparable to Latin-script European languages rather than to other typologically distant ones such as Mandarin (0.153) or Georgian (0.133). This pattern is consistent with the hypothesis that shared surface material (particu-

Condition	β	SE	z	p	n	$n_{\mathbf{FT}}$	R^2_m	R^2_c
LLaMA 3.1 8B (QLoRA)	−0.178	0.024	−7.31	$< .001$	75	15	0.52	0.92
Gemma 2 2B (QLoRA)	−0.149	0.033	−4.49	$< .001$	75	15	0.53	0.93
Gemma 2 2B (Full FT)	−0.202	0.059	−3.42	$< .001$	15	3	0.80	0.84

Table 1: Mixed-effects regression of neighborhood deformation on `lang2vec` typological distance, with probe-language fixed effects and random intercepts for fine-tuning language.

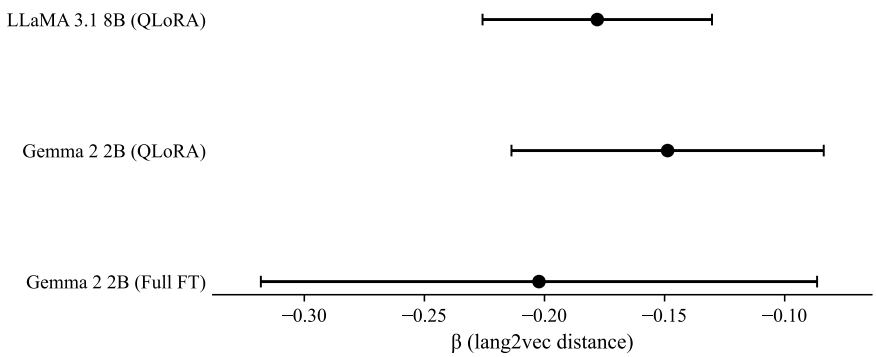


Figure 1: Mixed-effects coefficient for `lang2vec` typological distance across three fine-tuning conditions, with 95% confidence intervals.

larly shared subword tokens) may partially mediate interference, though our predictor-comparison results (Section 3.2) show that we cannot statistically isolate this mechanism from deeper phylogenetic or lexical relatedness.

Per-language rankings are architecture-specific. Per-language deformation rankings do not transfer between models: the Spearman rank correlation between LLaMA and Gemma per-language deformation is non-significant ($\rho = 0.21$, $p = 0.45$; Appendix A), indicating no reliable correspondence in which specific languages are most vulnerable across architectures.

3.4 Neighborhood size robustness

Neighborhood overlap depends on the choice of k. We recompute the main LLaMA 3.1 8B result for $k \in \{5, 10, 20, 50\}$. The `lang2vec` coefficient is negative and significant at $p < 10^{-11}$ throughout, with stable magnitude ($\beta \in [-0.178, -0.165]$; Appendix C, Table 4).

3.5 Contrast with prior findings

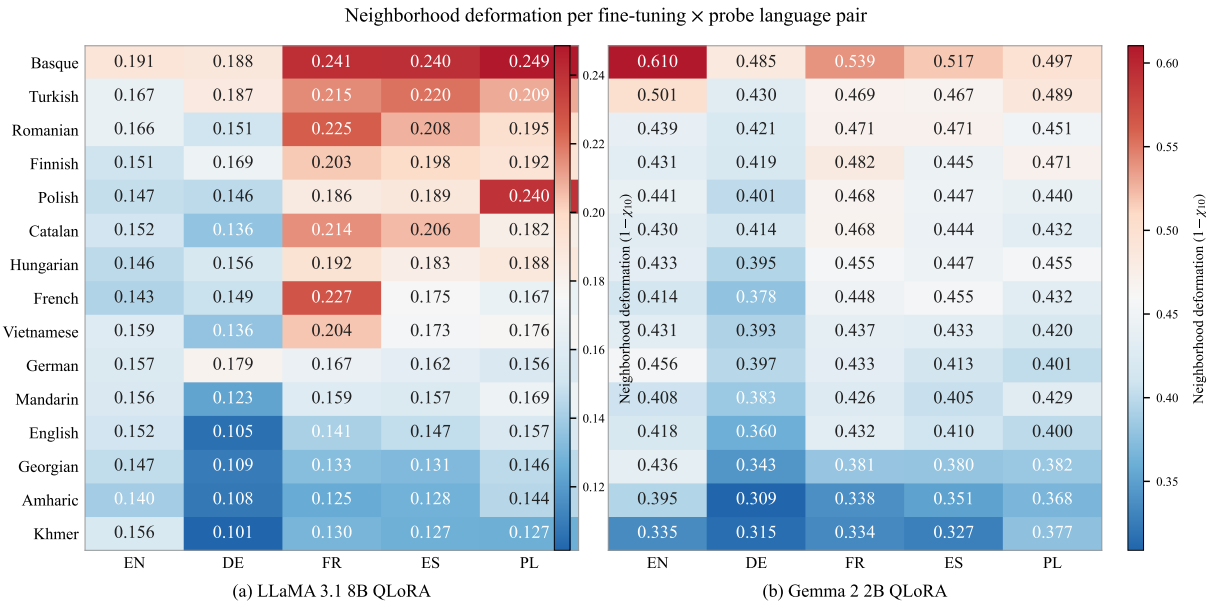
Our finding contrasts with both behavioral and representational studies of cross-lingual transfer. At the behavioral level, Lauscher et al. (2020) find that typological proximity facilitates zero-shot transfer, implying that distant languages suffer greater interference. At the representational level, Philippy et al. (2023) find that linguistic distance correlates positively with CKA-based representational change in mBERT during fine-tuning — again, closer languages are more protected.

We find the opposite. Typologically closer languages undergo greater neighborhood deformation after fine-tuning, not less. We caution, however, against treating this as a simple reversal. The divergence may reflect differences in model architecture (encoder-only mBERT vs. decoder-only LLMs), fine-tuning method (full fine-tuning vs. QLoRA), geometric measure (CKA vs. neighborhood overlap), or the interaction of these factors. Disentangling their contributions would require a controlled study varying each factor independently.

4 Conclusion

Our results reveal that representational geometry in multilingual models is highly dynamic: adaptation to a new language causes systematic collateral deformation in non-target languages. Crucially, the magnitude of this deformation is predicted by linguistic proximity — the closer the language, the greater the reorganization. All eleven typological distance measures we tested show negative coefficients, with ten of eleven significant at $p < .01$, and the effect replicates across two model architectures and two fine-tuning methods.

This pattern runs opposite to both behavioral



(a) LLaMA 3.1 8B QLoRA (b) Gemma 2 2B QLoRA

Figure 2: Neighborhood deformation $(1-\chi_{10})$ for each fine-tuning × probe language pair ($k = 10$). Full fine-tuning results are shown in Figure 4 (Appendix B).

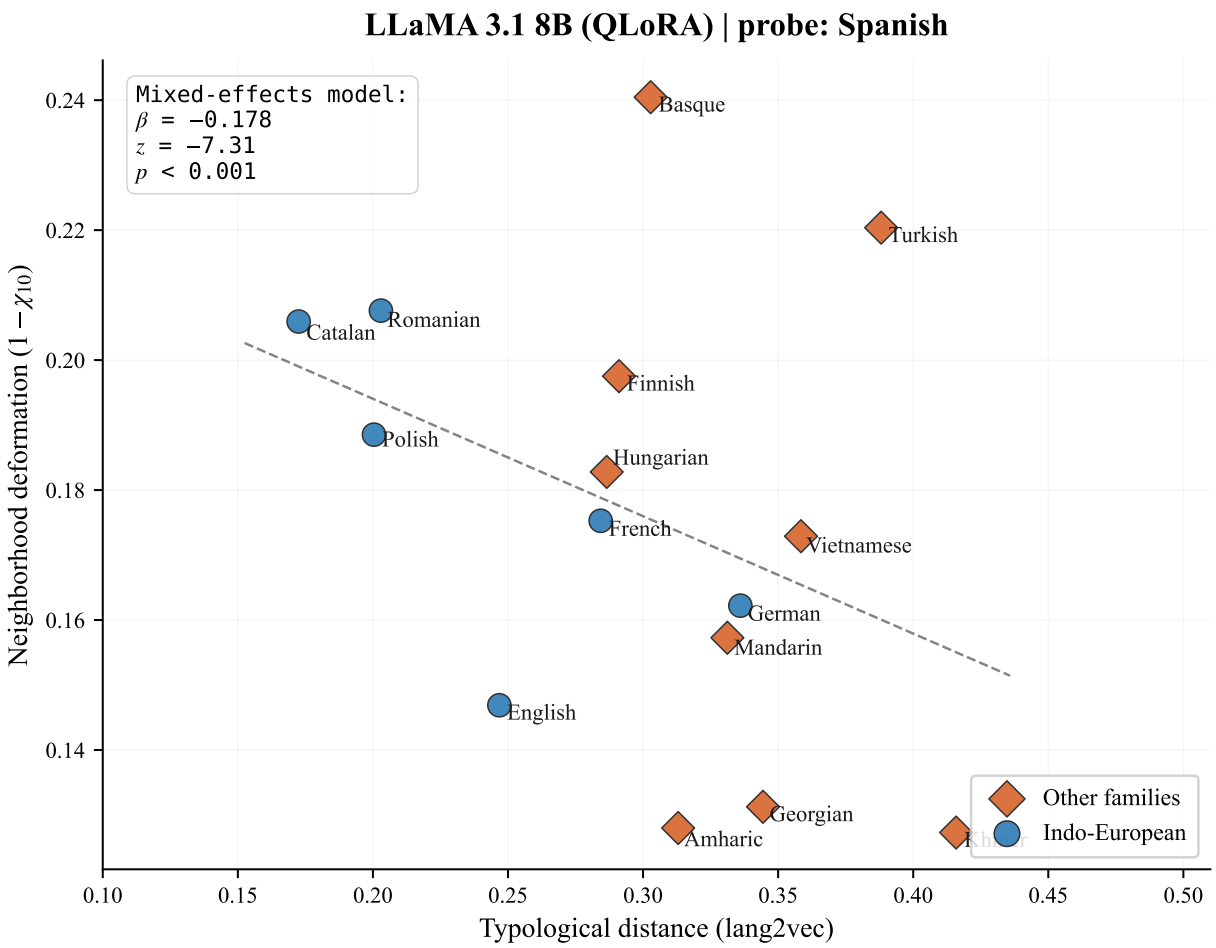


Figure 3: Neighborhood deformation vs. `lang2vec` typological distance for each of 15 fine-tuning languages, probed in Spanish (LLaMA 3.1 8B, QLoRA). Each point is one fine-tuning language; marker shape and color indicate language family.

Predictor	Category	β (SE)	$\lvert z \rvert$	p
lang2vec	Typological	-0.178 (0.024)	7.30	$< 10^{-12}$
glot_tree	Phylogenetic	-0.054 (0.008)	6.98	$< 10^{-12}$
asjp	Lexical	-0.106 (0.015)	6.96	$< 10^{-12}$
scripts	Writing system	-0.056 (0.008)	6.95	$< 10^{-12}$
textcat	Textual	-0.070 (0.010)	6.85	$< 10^{-12}$
phoible	Phonological	-0.068 (0.010)	6.71	$< 10^{-12}$

Table 2: Top six individual typological predictors ranked by $|z|$-statistic in the mixed-effects model (LLaMA 3.1 8B QLoRA, $n = 75$). All eleven predictors tested show negative coefficients; ten of eleven have $p < .01$ (Appendix F). The six strongest span typological, phylogenetic, lexical, writing system, phonological, and textual dimensions. lang2vec (our primary predictor) achieves the highest $|z|$ on LLaMA 3.1 8B; predictor rankings differ on Gemma.

studies of cross-lingual transfer (Lauscher et al., 2020) and prior representational work on encoder models (Philippy et al., 2023), which find that typological proximity facilitates rather than intensifies cross-lingual effects. The dissociation suggests that task performance is an incomplete diagnostic for what fine-tuning does to a multilingual model's internal structure, and that geometric and behavioral interference are driven by different dynamics.

These findings have a practical implication: when adapting a model to a new language, the representations most at risk are those of its closest linguistic neighbors — the languages typically assumed to benefit most from shared structure, not suffer from it. Monitoring representational change in typologically related languages, not just task performance in the target language, may be necessary for safe multilingual deployment.

What drives this representational interference at the mechanistic level remains open. Our collinearity analysis shows that phylogenetic, lexical, phonological, writing system, and textual distance dimensions are empirically inseparable in a 15-language panel. Disentangling them would require either a panel designed to break predictor correlations or direct experimental manipulation of tokenizer design and pretraining composition. The neighborhood deformation framework introduced here provides a concrete diagnostic for that investigation: it can be applied to any setting where base and fine-tuned representations of aligned concepts are available.

Limitations

Our study points at a strong correlation but does not identify a causal mechanism. The strong collinearity among typological predictors (Section 3.2) means the proximate cause remains unresolved, and disentangling these requires either a larger panel designed to break predictor correlations or direct experimental manipulation.

Four further scope limitations apply. First, probe languages were selected to be well-represented in both models' pretraining data; deformation patterns in lower-resource or non-Latin-script probe languages remain unknown. Second, we train at a fixed budget of ≈24.6M tokens, representing a single point on the adaptation curve. Third, we measure deformation in a single layer (the last hidden state); prior work suggests fine-tuning affects upper layers more than lower ones (Merchant et al., 2020; Doimo et al., 2020), and layer-wise patterns may differ from what we report. Fourth, we do not measure behavioral interference directly; our claim of a reversal from behavioral findings rests on contrasting our geometric results against prior work rather than measuring both in a single framework.

Finally, per-language deformation rankings do not transfer between LLaMA and Gemma ($\rho = 0.21$, $p = 0.45$), indicating that tokenizer design, pretraining mixture, and capacity modulate which languages are most vulnerable in ways our study does not clarify.

Acknowledgments

[This section is anonymized]

References

2022. The asjp database (version 20).

Ralf D. Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632, Doha, Qatar. Association for Computational Linguistics.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization.

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.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Hizniye Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2019. Northeuralex: a wide-coverage lexical database of northern eurasia. *Language Resources and Evaluation*, 54:1–29.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Diego Doimo, Aldo Glielmo, Alessio Ansuini, and Alessandro Laio. 2020. Hierarchical nucleation in deep neural networks. *Preprint*, arXiv:2007.03506.

Rob Van Der Goot, Esther Ploeger, Verena Blaschke, and Tanja Samardzic. 2025. DistaLs: a comprehensive collection of language distance measures. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 307–318, Suzhou, China. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.0.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotScript: A resource and tool for low resource writing system identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7774–7784, Torino, Italia. ELRA and ICCL.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. *Preprint*, arXiv:1905.00414.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. 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.

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12969–13000, Toronto, Canada. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

York Hay Ng, Aditya Khan, Xiang Lu, Matteo Salloum, Michael Zhou, Phuong H. Hoang, A. Seza Doğruöz, and En-Shiun Annie Lee. 2026. Modality matching matters: Calibrating language distances for cross-lingual transfer in uriel+. *Preprint*, arXiv:2510.19217.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. To train or not to train: Predicting the performance of massively multilingual models. pages 8–12.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Identifying the correlation between language distance and cross-lingual transfer in a multilingual

representation space. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 22–29, Dubrovnik, Croatia. Association for Computational Linguistics.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. LinguaMeta: Unified metadata for thousands of languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and modeling with python. In *Proceedings of the 9th Python in Science Conference*, pages 57–61, Austin, TX, USA.

Cheril Shah, Yashashree Chandak, and Manan Suri. 2023. The geometry of multilingual language models: An equality lens. *Preprint*, arXiv:2305.07839.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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. *Preprint*, arXiv:2010.11934.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

A Gemma Per-Language Deformation Rankings

Table 3 compares per-language deformation rankings between LLaMA 3.1 8B and Gemma 2 2B (both QLoRA). The Spearman rank correlation is weak ($\rho = 0.21$, $p = 0.45$), indicating that while the aggregate typological effect replicates across models (Table 1), specific per-language vulnerability patterns do not. The differences suggest that model-specific factors substantially modulate which languages are most affected.

Language	LLaMA def.	Rank	Gemma def.	Rank
Basque	0.222	1	0.450	4
Turkish	0.200	2	0.438	6
Romanian	0.189	3	0.404	12
Finnish	0.182	4	0.425	8
Polish	0.181	5	0.437	7
Catalan	0.178	6	0.439	5
Hungarian	0.173	7	0.410	11
French	0.172	8	0.451	3
Vietnamese	0.170	9	0.423	9
German	0.164	10	0.420	10
Mandarin	0.153	11	0.384	13
English	0.140	12	0.529	1
Georgian	0.133	13	0.338	15
Amharic	0.129	14	0.471	2
Khmer	0.128	15	0.352	14

Table 3: Per-language mean neighborhood deformation and rank for LLaMA 3.1 8B and Gemma 2 2B (both QLoRA, $k = 10$), averaged across five probe languages. Rank 1 = most deformation.

B Full Fine-Tuning Robustness

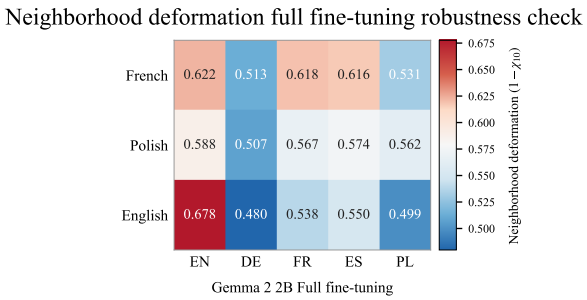


Figure 4: Neighborhood deformation $(1 - \chi_{10})$ under full fine-tuning on Gemma 2 2B (three fine-tuning languages; $k = 10$). The typological gradient is consistent with the QLoRA results, but the small sample ($n = 3$) precludes strong conclusions.

C Robustness to Neighborhood Size k

Table 4 reports the mixed-effects regression of neighborhood deformation on `lang2vec` distance (LLaMA 3.1 8B, QLoRA) for $k \in \{5, 10, 20, 50\}$. The coefficient is negative and highly significant ($p < 10^{-11}$) for all values tested, with stable magnitude ($\beta \in [-0.178, -0.165]$).

k	β	**SE**	z	p
5	-0.165	0.024	-6.92	$< 10^{-11}$
10	-0.178	0.024	-7.30	$< 10^{-12}$
20	-0.169	0.024	-7.01	$< 10^{-11}$
50	-0.178	0.025	-7.02	$< 10^{-11}$

Table 4: Mixed-effects regression of neighborhood deformation on `lang2vec` distance across neighborhood sizes k (LLaMA 3.1 8B QLoRA, $n = 75$).

D Self-Pair Exclusion

Five of the fifteen fine-tuning languages coincide with the probe set (English, German, French, Spanish, Polish). Table 5 shows the main regression with these self-pairs removed. The effect remains significant for both QLoRA conditions. For Gemma Full FT, which has only three fine-tuning languages (English, French, Polish) excluding self-pairs reduces the sample from 15 to 12 observations and removes one observation from each of the three groups, leaving insufficient power to detect the effect.

E Prompt Templates

Table 6 lists the two prompt templates used for embedding extraction. In each template, `{w}` is replaced with the target-language lexical entry and `{g}` with its English gloss from NorthEuraLex. Templates are provided for all five probe languages. The target word's span is located via character offsets, and the corresponding subword hidden states are mean-pooled. Averaging over two syntactically distinct templates per concept reduces the influence of template-specific artifacts.

F All Typological Distance Predictors

Condition	β	**SE**	$\lvert z \rvert$	p	n	$n_{\mathbf{FT}}$
LLaMA 3.1 8B (QLoRA)	-0.147	0.044	-3.34	.001	71	15
Gemma 2 2B (QLoRA)	-0.186	0.055	-3.38	.001	71	15
Gemma 2 2B (Full FT)	$+0.135$	0.080	$+1.68$	.092	12	3

Table 5: Mixed-effects regression on `lang2vec` with self-pairs excluded (fine-tuning language = probe language). Compare to Table 1 (full data). The Gemma Full FT result is underpowered after exclusion ($n = 12$, $n_{\text{FT}} = 3$).

Lang.	**Template 1**	**Template 2**
English	"{w}" means "{g}".	The word "{w}" means "{g}".
German	"{w}" bedeutet "{g}".	Das Wort "{w}" bedeutet "{g}".
French	"{w}" signifie "{g}".	Le mot "{w}" signifie "{g}".
Spanish	"{w}" significa "{g}".	La palabra "{w}" significa "{g}".
Polish	"{w}" oznacza "{g}".	Słowo "{w}" oznacza "{g}".

Table 6: Prompt templates per language. `{w}`: target word; `{g}`: English gloss.

(a) LLaMA 3.1 8B (QLoRA, $n = 75$)

Rank	Predictor	Category	β (SE)	$\lvert z \rvert$	p
1	lang2vec	Typological	-0.178 (0.024)	7.30	$< 10^{-12}$
2	glot_tree	Phylogenetic	-0.054 (0.008)	6.98	$< 10^{-12}$
3	asjp	Lexical	-0.106 (0.015)	6.96	$< 10^{-12}$
4	scripts	Writing system	-0.056 (0.008)	6.95	$< 10^{-12}$
5	textcat	Textual	-0.070 (0.010)	6.85	$< 10^{-12}$
6	phoible	Phonological	-0.068 (0.010)	6.71	$< 10^{-12}$
7	char_JSD	Textual	-0.135 (0.025)	5.45	4.9×10^{-8}
8	concepts	Lexical	-0.518 (0.146)	3.54	4.1×10^{-4}
9	whitespace	Textual	-0.080 (0.025)	3.24	1.2×10^{-3}
10	geo	Geographic	-0.099 (0.038)	2.63	8.4×10^{-3}
11	punctuation	Textual	-0.064 (0.029)	2.21	2.7×10^{-2}

(b) Gemma (QLoRA, $n = 75$)

Rank	Predictor	Category	β (SE)	$\lvert z \rvert$	p
1	scripts	Writing system	-0.063 (0.011)	5.78	7.5×10^{-9}
2	lang2vec	Typological	-0.149 (0.033)	4.48	7.3×10^{-6}
3	char_JSD	Textual	-0.125 (0.030)	4.18	2.9×10^{-5}
4	whitespace	Textual	-0.112 (0.029)	3.91	9.1×10^{-5}
5	glot_tree	Phylogenetic	-0.041 (0.011)	3.90	9.8×10^{-5}
6	asjp	Lexical	-0.076 (0.021)	3.66	2.6×10^{-4}
7	phoible	Phonological	-0.049 (0.014)	3.62	3.0×10^{-4}
8	textcat	Textual	-0.050 (0.014)	3.54	4.0×10^{-4}
9	concepts	Lexical	-0.480 (0.181)	2.66	7.9×10^{-3}
10	geo	Geographic	-0.144 (0.058)	2.50	1.2×10^{-2}
11	punctuation	Textual	-0.063 (0.037)	1.72	8.6×10^{-2}

(c) Gemma (Full fine-tuning, $n = 15$)

Rank	Predictor	Category	β (SE)	$\lvert z \rvert$	p
1	glot_tree	Phylogenetic	-0.072 (0.013)	5.34	9.2×10^{-8}
2	asjp	Lexical	-0.128 (0.025)	5.10	3.5×10^{-7}
3	whitespace	Textual	-0.286 (0.063)	4.56	5.1×10^{-6}
4	textcat	Textual	-0.086 (0.019)	4.55	5.3×10^{-6}
5	scripts	Writing system	-0.081 (0.018)	4.44	9.0×10^{-6}
6	phoible	Phonological	-0.081 (0.020)	4.09	4.4×10^{-5}
7	geo	Geographic	-0.901 (0.238)	3.78	1.6×10^{-4}
8	lang2vec	Typological	-0.202 (0.059)	3.42	6.2×10^{-4}
9	char_JSD	Textual	-0.193 (0.058)	3.30	9.6×10^{-4}
10	concepts	Lexical	-1.020 (0.409)	2.50	1.3×10^{-2}
11	punctuation	Textual	-0.216 (0.149)	1.45	1.5×10^{-1}

Table 7: Comparison of individual typological distance predictors across models and training setups. Each subtable reports coefficients from mixed-effects models with a single predictor. All coefficients are negative (greater distance $\rightarrow$ less deformation). Predictors are sorted by $|z|$ within each condition.