UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Similarity All The Way Up: Multilingual Generalization in LLMs Relies on Language-Level Similarity Structures

Permalink
https://escholarship.org/uc/item/378720cw

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Rakshit, Supantho
Goldberg, Adele
Conklin, Henry Coxe

Publication Date
2026

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Similarity All The Way Up: Multilingual Generalization in LLMs Relies on Language-Level Similarity Structures

Supantho Rakshit[1], **Adele Goldberg**[2], **Henry Conklin**[3]
[1]Department of Electrical and Computer Engineering, [2]Department of Psychology, [3]Laboratory for Artificial Intelligence
Princeton University

Abstract

As Large Language Models (LLMs) grow more capable across diverse tasks, their (in)ability to generalize remains difficult to quantify and poorly understood beyond limited domains. In particular, LLMs are known to struggle generalizing multilingually, to languages outside of English, and that are poorly attested in their training data. To understand why this may be, and what enables some models to perform better than others, we turn to a long history of work across the cognitive sciences, arguing that successful generalization derives from appropriate representations in similarity space. We look at how well LLMs' representations capture the hierarchical similarity structure between distinct languages. Strikingly, we show LLMs' latent representations largely recover the hierarchical structure of the Indo-European language family tree – grouping languages that are members of the same subfamily closely together in representation space. Furthermore, we show that the degree to which models reflect the similarity structure of languages correlates with their performance on XNLI, a multilingual natural language inference benchmark. This extends classic work on similarity-driven generalization at scale, showing how models that *represent similar languages similarly* generalize better from one language to another.

Keywords: generalization; similarity; multilingual language models; representation geometry; training dynamics; Jensen-Shannon Divergence

Introduction

Large Language Models (LLMs) perform impressively across a broad range of tasks, but we still lack a clear understanding of how their representational structures drive their performance. Multilinguality represents an interesting test case. The data used to train contemporary models remains disproportionately in English, and model performance quickly degrades as you move outside high-resource languages (languages with greater than 1 billion tokens of available data, e.g., German, French, Chinese). To understand why this may be and what representational structures enable better multilingual performance, we analyze the similarity structure learned by LLMs. We show models implicitly recover similarity relationships among natural languages in their representation space, without explicit supervision. Moreover, we show that this representational structure relates to performance, with models that represent similar languages similarly performing better on a standard benchmark of multilingual performance.

Generalization, or the ability to make inferences about new cases on the basis of prior experience, has long been understood in cognitive science as dependent on the underlying structure of representations. Shepard (1957, 1987) first observed that generalization follows an exponential gradient in psychometric space: the probability of relating a novel item to an existing one decays as a function of their distance in representation space. This principle, later given a rational Bayesian justification by Tenenbaum and Griffiths (2001), shows how generalization behavior is determined by representational geometry. Exemplar models of categorization formalized this relationship, showing that generalization to novel items depends on their similarity to stored instances in a multidimensional space (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986), while prototype effects demonstrated that representations encode central tendencies that support interpolation to never-seen items (Posner & Keele, 1968). These findings motivated a more general theoretical claim: that representations simply *are* similarity structures and that cognitive generalization reduces to operations over the geometry of these spaces (Edelman, 1998; Gärdenfors, 2000; Kriegeskorte & Kievit, 2013).

Large Language Models, based on the transformer architecture (Vaswani et al., 2017), are connectionist models, with distributed representations acquired through gradient-based learning. A long history of work has shown how such representations place similar items in nearby regions of activation space, enabling networks to generalize on the basis of learned similarity structure (Elman, 1990; Hinton, 1986; Rogers & McClelland, 2004; Rumelhart et al., 1986). Classic demonstrations showed that networks generalize morphological patterns to novel words (Rumelhart & McClelland, 1986) and semantic properties to novel concepts based on the similarity structure of their internal representations.

An existing body of work has studied multilingual LLMs and the extent to which their representations enable cross-lingual transfer. Multilingual LLMs are increasingly trained without parallel data, simply performing next token prediction on data from each language independently (Conneau et al., 2020; Devlin et al., 2019). Such models can often generalise between languages – when fine-tuned on a specific task in one language, performance improves on the same task when evaluated in other languages as well (Pires et al., 2019). These performance gains have been found to be greater between languages with shared vocabulary and typological characteristics (Wu & Dredze, 2019). Having finite representational capacity forces models to learn structures that are effectively shared

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

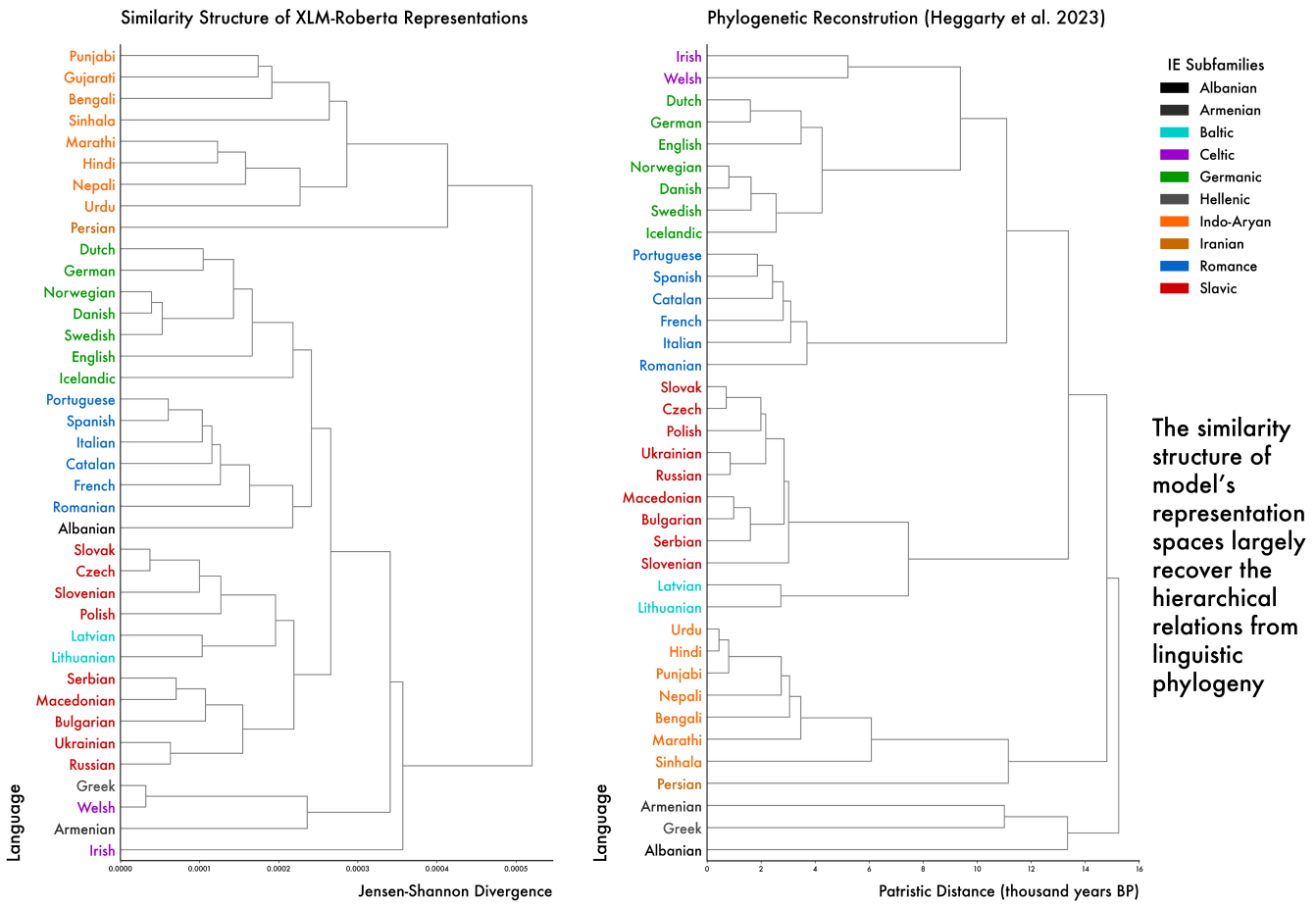

Figure 1: **LLMs Implicitly Recover Relations Between Languages:** Similarity structure among 38 Indo-European languages, as derived from XLM-Roberta's representations (Left) and reconstructed from Heggarty et al. (2023) (Right). Colors represent phylogenetic language subfamilies, with a legend in the top right.

across languages, with models converging to representations that are not entirely specific to a single language, instead aligning related semantic concepts cross-linguistically (Conneau et al., 2020). Previous work using mBERT representations of multilingual word lists finds that phylogenetic distance is the strongest predictor of pair-wise representational distances for 100 languages (Rama et al., 2020). Here, instead of looking at the pairwise case, we look at whether LLMs recover the full set of similarity structures across languages, exemplified in our case by the Indo-European family tree. Our analysis systematically compares twelve LLMs with varying architectures, aggregating representations over a full set of documents in each language rather than focusing on concept-level word embeddings. Further, we predict that models that represent *similar languages similarly* will generalise better across languages. If the similarity structure of a representation space defines its generalization behaviour (Tenenbaum & Griffiths, 2001), then representations that reflect hierarchical relations among languages, beyond mutually interpretable languages like German and Dutch, should generalise best. That is, models that group linguistic subfamilies (e.g. Germanic, Slavic, Romance) closely together should perform better at tasks that require cross-linguistic transfer.

To characterise similarity among languages, we turn to lexico-statistical work on linguistic phylogeny (e.g. Bouckaert et al., 2012; Gray & Atkinson, 2003; Heggarty et al., 2023). For instance, Gray and Atkinson (2003) aims to reconstruct the phylogenetic tree of the world's languages from corpora of lexical overlaps. While statistical approaches have been criticized as being unable to distinguish borrowing from shared ancestry (e.g. Kassian & Starostin, 2025; Nichols & Warnow, 2008), our goal is to estimate hierarchical similarity relationships among languages rather than mirror the historical evolution of Indo-European. To this end, it is widely appreciated that contact between languages, as well as shared ancestry, has a strong influence on convergences among languages (Haspelmath, 2005) and dialects within a language (Dunn & Wong, 2025; Wieling & Montemagni, 2017). For example, Norwegian and Icelandic are historically sister languages, since both descended from Old West Norse, yet Norwegian does not share Icelandic's case system, complex grammar, nor its writing system, and the two languages are mutually unin-

telligible. Danish evolved from a distinct dialect (Old East Norse), but is far more similar to Norwegian today than Icelandic is, and written forms of Danish and Norwegian are highly mutually intelligible. Statistical models based on estimated distances between languages provide an independent way of estimating similarity relations among individual languages and sub-families(e.g. Bouckaert et al., 2012; Heggarty et al., 2023). Therefore, rather than mapping to a historically veridical tree structure based on the comparative method of (Nordhoff & Hammarström, 2011), we adopt the tree structure induced from a statistical model, namely (Heggarty et al. (2023)), as an independent reference tree that captures hierarchical similarity relationships among languages.

Here we show that LLMs implicitly recover the structure of the reference tree to a remarkable extent. Further, we show models explicitly trained to be multilingual approximate the reference tree more closely. The current experiments extend a cognitively motivated account of generalization at scale, with similarity structure playing a central role not just for specific objects or words, but across entire languages.

Methods

Our analysis builds a hierarchical similarity tree based on representations of 38 languages, for each of 12 LLMs. We compare the trees based on each model's representations to the independent reference tree by computing three different distances: subtree and path kernels and cophenetic correlations. We find across models that their representational similarity structure largely recovers the structure of the reference tree. We visualize the structure of one model, XLM-Roberta, in Figure 1 (Left) alongside the reference tree constructed from Heggarty et al. (2023) (Figure 1: Right).

Language sample The current work investigates similarity relationships among 38 written, living Indo-European (IE) languages from 5 subfamilies: Romance (6), Germanic (6), Indo-Aryan (9), Slavic (9), Baltic (2), and independent IE languages ([3]: Greek, Armenian, and Albanian). We restrict our attention to IE here because similarities among any pair of IE languages are expected to be non-zero; multiple languages within several language families are included to determine whether models capture smaller degrees of similarity within the higher range accurately.

The Reference Tree To assess the validity of similarity relationships among languages, we use a tree structure constructed on the basis of Heggarty et al. (2023) as the target reference. Heggarty et al. (2023) fit a Bayesian model to lexical cognate judgments for 170 concepts across 161 Indo-European languages. To construct the reference tree for our sample of 38 languages, we extracted the corresponding terminal nodes from the full Bayesian Maximum Clade Credibility (MCC) tree from Heggarty et al. (2023), using exact string matching against the CLDF language metadata. In cases where modern standard varieties were unavailable, we selected the closest dialectal proxy (e.g., WelshNorth for Welsh, ArmenianEastern for Armenian). We computed pairwise patristic distances between all language pairs as: $d_{\text{patristic}}(i,j) = \sum_{e \in \text{ path }(i,j)} \ell(e)$, where path (i,j) denotes the unique path connecting terminals i and j through their most recent common ancestor, and $\ell(e)$ is the branch length (in thousands of years before present) of edge e. This yields a symmetric 38×38 distance matrix $\mathbf{D}_{\text{ref}}$ where entry D_{ij} represents the estimated time depth to the most recent common ancestor of languages i and j. We performed UPGMA clustering on $\mathbf{D}_{\text{ref}}$ to produce a dendrogram for visualization and metric computation, maintaining the ultrametric property of the Bayesian phylogeny.

Measuring LLMs' ability to generalize across languages: XNLI We use the Cross-lingual Natural Language Inference (XNLI) benchmark (Conneau et al., 2018) to quantify the extent to which each model enables zero-shot transfer across languages. XNLI requires models to determine whether a premise sentence entails, contradicts, or is neutral with respect to a hypothesis sentence. Though NLI has been criticized as a somewhat unnatural task (Weissweiler et al., 2025), XNLI is a practical and well-established way to measure how well models generalize from one language to another.

LLM Density Estimation To capture similarity relationships in each model, we estimate a conditional density in the model's embedding space $P(\hat{Z}|\text{language id})$ which describes how a given language is represented. To do so, we employ the soft-density estimator from Conklin (2025). This approach estimates distributions in continuous space by softly-discretizing them. For a model with L transformer layers and hidden dimension d, let $h_t^{(\ell)} \in \mathbb{R}^d$ denote the hidden state for token t at layer ℓ. We construct a discrete codebook $C = \{c_1, \ldots, c_K\}$ by sampling uniformly at random from the surface of the unit hypersphere. We set $K = 3d$ to ensure consistent expressive capacity across different model architectures. Each token embedding $h_t^{(\ell)}$ is softly assigned to codebook entries via temperature-scaled softmax, then aggregated across all tokens in a language to produce a language-specific conditional distribution. For a given language with token set $\mathcal{T}$, the distribution at layer ℓ is:

$$P^{(\ell)}(\hat{Z} \mid \text{language}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{softmax}\left(\frac{h_t^{(\ell)} \cdot C}{\tau \| h_t^{(\ell)} \|_2} \right) \quad (1)$$

where $\tau = 0.1$ is the temperature parameter and the softmax operates over all K codebook entries in C. Hence, $P^{(\ell)}(\hat{Z})$ is an estimate of a categorical distribution– this is related to kernel density estimation (Parzen, 1962). By conditioning these estimates $P^{(\ell)}(\hat{Z}|X = x)$, where x = the language of the corresponding input sentence, we get a robust estimation of how a given language is represented in the model's embedding space at layer ℓ.

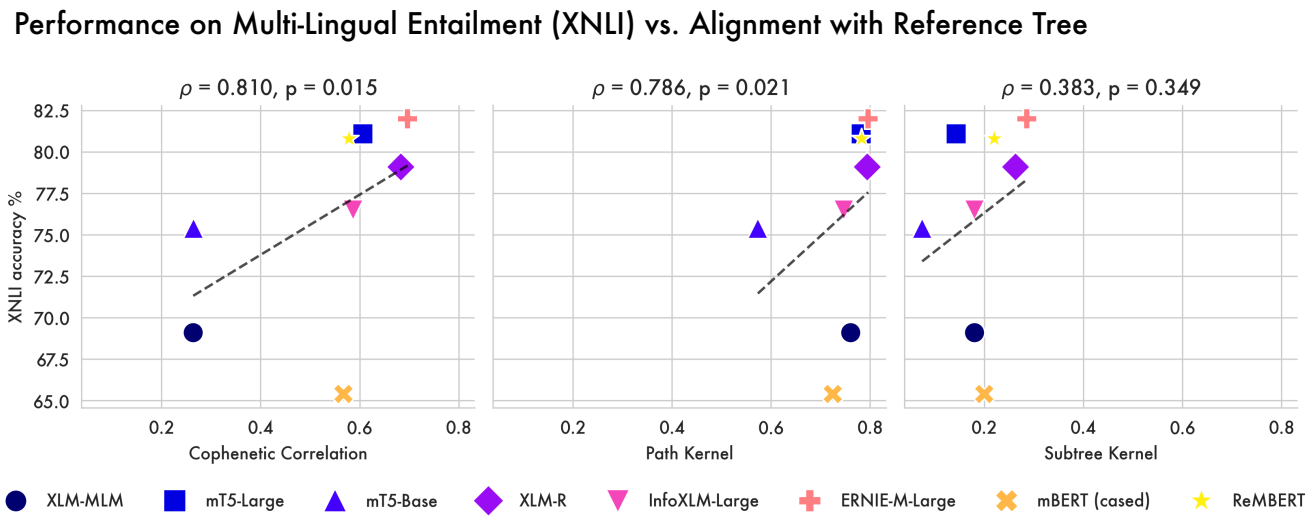

Figure 2: **Correlating Structure with Performance:** Each facet shows one of three tree-distance measures (x-axis) plotted against XNLI accuracy percentage (y-axis) for eight LLMs. Above each facet is the Spearman rank correlation ρ between axes. Cophenetic Correlation and Path Kernel show a significant positive relationship with performance; both reflect how well LLM representations capture higher-order structures of the reference, like sub-family membership, rather than exact neighbor match.

Constructing a Representational Tree To quantify differences between language distributions, we compute pairwise Jensen-Shannon divergences. Given data in a set of languages Q for each layer ℓ and language pair (q_i, q_j), we compute their layer-wise divergence, aggregating across layers to get d:

$$d(q_i, q_j) = \frac{1}{L}\sum_{\ell=1}^{L} D_{\mathrm{JS}}\left(P^{(\ell)}(\hat{Z}|q_i)||P^{(\ell)}(\hat{Z}|q_j)\right) \quad (2)$$

Concatenating d for each language pair gives a distance matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ for n languages. To this we can then apply UPGMA clustering to construct a dendrogram. UPGMA assumes equal root-to-tip distances, matching the time-calibrated property of the Bayesian phylogeny from Heggarty et al. (2023), making dendrograms directly comparable.

LLMs, Data, and Distances To show results hold outside of a single model, our experiments look at 5 different LLM families: BERT (Chung et al., 2020; Devlin et al., 2019), RoBERTa (Liu et al., 2019), mT5 (Xue et al., 2021), XLM (Conneau & Lample, 2019; Conneau et al., 2020), and the Pythia models (Biderman et al., 2023), which make training checkpoints available. We focus on earlier LLMs in part because they come in explicitly monolingual and multilingual variants — virtually all recent models are trained in a massively multilingual setting. Both BERT and RoBERTa come in English and multilingual variants, allowing us to look at how representations change as models are intentionally trained on broader sets of languages. Across all results and models, our estimates of conditional distributions $P(\hat{Z}|\text{language})$ are based on 7500 sentences randomly sampled separately for each language from the multilingual C4 dataset - a broad crawl of internet data (Raffel et al., 2020).

Comparing LLM Representations with the Reference Tree

We measure how well a model approximates similarity among languages by computing the distance between two hierarchical structures: the reference tree of language relations (Heggarty et al., 2023), and a dendrogram induced from LLM embeddings via density estimation and agglomerative clustering. We employ measures to capture three different aspects of hierarchical correspondence: the Cophenetic Correlation Coefficient, the Subtree Kernel, and the Path Kernel.

Rationale for Multiple Measures Each measure captures complementary aspects of the correspondence between representational and the similarity structure represented by the reference tree. The Cophenetic Correlation provides a global assessment of whether pairwise reference distances are preserved in the embedding space. The Path Kernel captures shared global hierarchical organization, testing whether the model encodes correct broad-to-narrow nesting of language subfamilies. The Subtree Kernel is the strictest test, identifying whether the model has exactly recovered local groupings, including identical leaf sets and topology within each subfamily. Together, these measures characterize both overall fit and specific convergences/divergences between learned representations and the reference tree.

Cophenetic Correlation (CC) measures the degree to which a hierarchical structure preserves pairwise distance re-

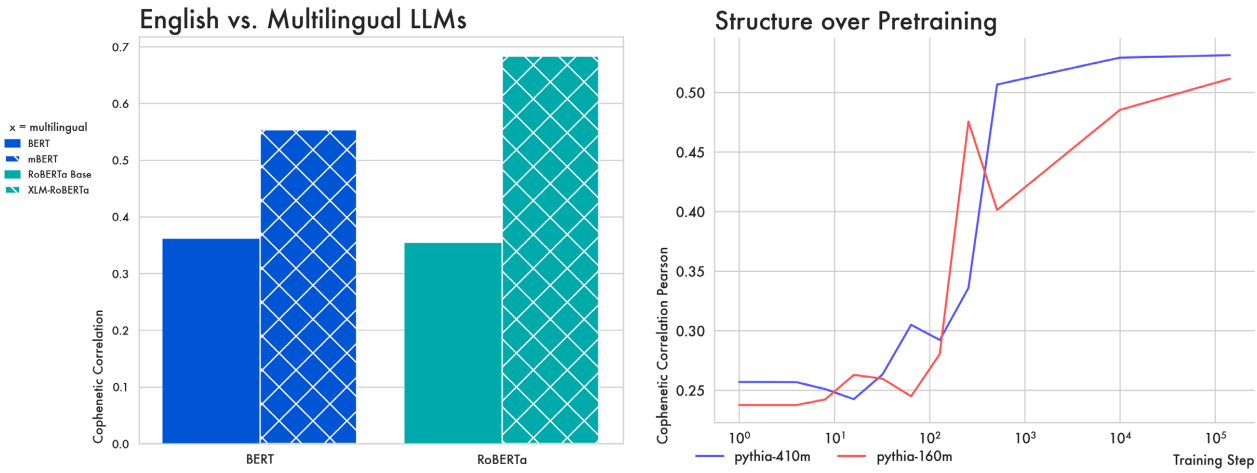

Figure 3: **Multilingual vs. English LLMs | Structure Over Pretraining: (Left)** y-axis: degree of alignment with the reference tree measured by the cophenetic correlation; x-axis: one of the model families BERT or RoBERTa. Colors correspond to model IDs. All models share the same architecture, number of layers, and parameters, but differ in training data. Models trained on multilingual data have a white x pattern. Multilingual models better align with the reference tree. **(Right):** y-axis: Cophenetic Correlation with the reference tree; x-axis: time-points in pretraining an LLM in terms of gradient step (log scaled); Color: size of the Pythia model. We observe that representational structures begin to correlate with the reference tree early on.

lationships (Sokal & Rohlf, 1962). Given a hierarchical tree $\mathcal{T}$ over a set of items, the cophenetic distance c_{ij} between items i and j is defined as the height of the most recent common ancestor (MRCA) at which i and j are first joined into a single cluster:

$$c_{ij} = \text{height}(\text{MRCA}_{\mathcal{T}}(i, j)) \tag{3}$$

For trees with branch lengths, the cophenetic distance equals the sum of branch lengths from i to their MRCA plus the sum from j to their MRCA. For model-derived dendrograms from UPGMA, cophenetic distances correspond to the Jensen-Shannon divergences at which clusters were merged. We extract cophenetic distance matrices $\{c_{ij}^{\text{phylo}}\}$ and $\{c_{ij}^{\text{LLM}}\}$ from both the reference and model trees over the shared set of n languages, then compute Pearson's correlation:

$$\rho_{\text{cophenetic}} = \text{Pearson}(\{c_{ij}^{\text{phylo}}\}_{i<j}, \{c_{ij}^{\text{LLM}}\}_{i<j}) \tag{4}$$

High CC indicates the model preserves relative proximity: language pairs closer in the phylogenetic tree remain proportionally closer in the model's representation space.

Subtree Kernel The subtree kernel quantifies tree similarity by counting shared subtree structures (Collins & Duffy, 2001; Vishwanathan et al., 2010). For two trees $\mathcal{T}_1$ and $\mathcal{T}_2$, we define an indicator function:

$$C(n_1, n_2) = \mathbb{1}[\text{subtree}(n_1) \cong \text{subtree}(n_2)] \tag{5}$$

where $\text{subtree}(n_1) \cong \text{subtree}(n_2)$ denotes subtree isomorphism: same topology and leaf set. For leaves, $C(n_1, n_2) = \mathbb{1}[\text{label}(n_1) = \text{label}(n_2)]$. For internal nodes, $C(n_1, n_2) = 1$ if $\text{leaves}(n_1) = \text{leaves}(n_2)$ and all children match recursively, else 0. The kernel sums over all node pairs:

$$K_{\text{subtree}}(\mathcal{T}_1, \mathcal{T}_2) = \sum_{n_1 \in \mathcal{T}_1} \sum_{n_2 \in \mathcal{T}_2} C(n_1, n_2) \tag{6}$$

Normalized to be $\in [0, 1]$. As with CC, higher scores indicate a model that better approximates the reference tree. High subtree kernel similarity indicates the model correctly identifies phylogenetic subfamilies (e.g., Romance, Germanic) as coherent clusters with matching internal structure, even if their global arrangement differs.

Path Kernel The path kernel measures tree similarity via shared root-to-leaf paths (Gärtner, 2003; Kashima et al., 2003), capturing a global hierarchical organization. For a given language ℓ, let $\text{path}_{\mathcal{T}}(\ell)$ denote the sequence of internal nodes from root to ℓ. Path similarity is the normalized longest common prefix (LCP):

$$\text{sim}(\text{path}_{\mathcal{T}_1}(\ell), \text{path}_{\mathcal{T}_2}(\ell)) = \frac{|\text{LCP}(\text{path}_{\mathcal{T}_1}(\ell), \text{path}_{\mathcal{T}_2}(\ell))|}{\max(|\text{path}_{\mathcal{T}_1}(\ell)|, |\text{path}_{\mathcal{T}_2}(\ell)|)} \tag{7}$$

The kernel aggregates over all languages:

$$k_{\text{path}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{sim}(\text{path}_{\mathcal{T}_1}(\ell), \text{path}_{\mathcal{T}_2}(\ell)) \tag{8}$$

yielding a similarity score in $[0, 1]$ measuring global hierarchical alignment. High path kernel similarity indicates the model encodes correct hierarchical nesting: languages follow similar ancestral paths (e.g., French: Indo-European $\rightarrow$ Romance $\rightarrow$ Italic), preserving multi-level family organization.

Results & Discussion

The similarity structure for a single higher-performing model, XLM-Roberta, is provided in Figure 1 (Left) for visual comparison. Across our experiments, models broadly align representations with the reference tree based on Heggarty et al. (2023) (Figure 1: Right). XLM-Roberta's similarity structure does well in distinguishing Germanic, Romance, and Indo-Aryan language subfamilies from one another. The Slavic language family, too, would be clustered apart from other subfamilies were it not for the two Baltic languages incongruously dividing it, separating Slovenian from Serbian. Three independent languages — Greek, Welsh, and Albanian — pose a challenge to most models, with a recurring Greek-Welsh pairing in XLM-Roberta, m-BERT, and Pythia-160m that is particularly surprising given their lexical dissimilarity. The degree of alignment across each of our measures is provided along the x-axes in each facet of Figure 2.

On Orthography Tokenization in LLMs is sensitive to script, so differences in orthography likely contribute to certain patterns of similarity. For example, Slavic languages written in Cyrillic form a separate cluster from those that use Latin script. Yet results also demonstrate that script is not a determinant factor in LLM similarity representations. Hindi and Urdu — written in Devanagari and Perso-Arabic, respectively — are nonetheless grouped within the same subfamily across models, and Persian is consistently placed adjacent to North Indian languages despite its Arabic script. Moreover, exploratory analyses appropriately grouped Hebrew and Arabic, written in distinct scripts, together in a Semitic cluster, while neither Persian nor Urdu migrated toward this cluster despite sharing the Arabic script. This pattern suggests that orthography modulates similarity, but linguistic relatedness is the dominant signal LLMs latch on to. We therefore view tokenization as a property of the LLM whose interaction with similarity structure we are characterizing, rather than a confound exogenous to the model.

Similarity Structure Correlates with Performance To see what kinds of structure relate to performance, we look at representations from eight different LLMs correlating our three different distance measures with accuracy on a large-scale multilingual benchmark (shown in figure 2). The Cophenetic Correlation (CC) and Path Kernel relate significantly to XNLI accuracy ($\rho = 0.81, p = 0.02, \rho = 0.79, p = 0.02$, respectively) while the subtree kernel does not ($\rho = 0.383, p = 0.349$). The Path Kernel and CC are more global measures, indicating whether the model has preserved the overall structure of the reference tree, capturing the higher-order categories like language subfamilies. By contrast, the subtree kernel looks at local structures, whether the exact neighbors in a model match. These results indicate that better-performing models more faithfully preserve linguistic similarity structure — as demonstrated here for Indo-European — and correctly organizing broader subfamily groupings matters more for performance than matching exact nearest-neighbour relationships.

English vs. multilingual Training Data BERT and RoBERTa were both trained predominantly, albeit not exclusively, on English data. Later variants, mBERT and XLM-RoBERTa, were explicitly trained as multilingual models using data from 104 and 100 languages, respectively. All four models have the same architecture and number of parameters. Figure 3 shows how multilingual training data affects the similarity structure of models' representations. Both multilingual models achieve a higher match with our reference tree than their monolingual counterparts across all three measures.

Language-Level Structures Emerge Early in Training Figure 3 looks at when the similarity structure across languages begins to emerge during pre-training by analysing checkpoints of two different sizes of Pythia models (160, and 410 million parameters. Both model sizes begin to align with the reference tree early in training, by step 100. This aligns with work on progressive differentiation, as in Saxe et al. (2019) whose account of semantic development shows how models learn broader distinctions first, before learning progressively subtler properties of the data. Compared to word or concept level distinctions, language identity is relatively broad, becoming clear early in training.

Limitations and Future Work

Future work needs to extend this analysis to other language families and benchmarks. As noted, systematic errors recur across models, including misplacement of independent languages and an unexplained Welsh-Greek pairing, warranting further investigation. The current analysis builds on similarity as a foundational construct in LLMs, earlier connectionist models, and psychology more generally. At the same time, the notion of similarity raises several well-known specters related to its context specificity (Medin et al., 1993; Tversky, 1977). Nonetheless, similarity is an invaluable metric for understanding the geometric representations of LLMs. The current results provide a case in point that extends its relevance to full languages. This suggests that historical changes that operate across most or all of a language may be analyzable as a general shift in similarity space, but this would require its own investigation.

Conclusion

The key contribution of the current work is two-fold: first, articulating the kind of higher-order similarity structure that emerges in LLM representations, as well as when and how it may arise. Second, showing a relationship between language-level similarity structures and a model's multilingual performance. This forms an extension of existing accounts of similarity's role in generalization at a far greater level of abstraction: similarity structure matters not just for generalizing at the level of words or sentences, but across entire languages.

References

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & Wal, O. V. D. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *Proceedings of the 40th International Conference on Machine Learning*, 2397–2430. Retrieved October 1, 2025, from https://proceedings.mlr.press/v202/biderman23a.html

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., & Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. *Science*, *337*, 957–960. https://doi.org/10.1126/science.1219669

Chung, H. W., Fevry, T., Tsai, H., Johnson, M., & Ruder, S. (2020). Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Collins, M., & Duffy, N. (2001). Convolution kernels for natural language. *Advances in Neural Information Processing Systems*, *14*, 625–632.

Conklin, H. C. (2025). Information structure in mappings: An approach to learning, representation and generalisation. *The University of Edinburgh*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. *Advances in neural information processing systems*, *32*.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2475–2485. https://doi.org/10.18653/v1/D18-1269

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv: 1810.04805]. *arXiv:1810.04805 [cs]*. Retrieved June 19, 2020, from http://arxiv.org/abs/1810.04805

Dunn, J., & Wong, S. (2025). Language contact and population contact as sources of dialect similarity. *Languages*, *10*(8), 188.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press.

Gärtner, T. (2003). A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter*, *5*(1), 49–58.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, *426*, 435–439. https://doi.org/10.1038/nature02029

Haspelmath, M. (2005). *The world atlas of language structures*. Oxford University Press.

Heggarty, P., Anderson, C., Scarborough, M., King, B., Bouckaert, R., Jocz, L., Kümmel, M. J., Jügel, T., Irslinger, B., Pooth, R., et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of indo-european languages. *Science*, *381*(6656), eabg0818.

Hinton, G. E. (1986). Learning distributed representations of concepts. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, 1–12.

Kashima, H., Tsuda, K., & Inokuchi, A. (2003). Marginalized kernels between labeled graphs. *Proceedings of the 20th International Conference on Machine Learning*, 321–328.

Kassian, A. S., & Starostin, G. (2025). Do 'language trees with sampled ancestors' really support a 'hybrid model' for the origin of Indo-European? Thoughts on the most recent attempt at yet another IE phylogeny. *Humanities and Social Sciences Communications*, *12*, 1–15. https://doi.org/10.1057/s41599-025-04986-7

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Nichols, J., & Warnow, T. (2008). Tutorial on computational linguistic phylogeny. *Language and linguistics compass*, *2*(5), 760–820.

Nordhoff, S., & Hammarström, H. (2011). Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. *First International Workshop on Linked Science 2011-In conjunction with the International Semantic Web Conference (ISWC 2011)*.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065–1076.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics, 4996–5001.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3), 353–363.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1–67.

Rama, T., Beinborn, L., & Eger, S. (2020). Probing multilingual bert for genetic and typological signals. *Proceedings of the 28th International Conference on Computational Linguistics*, 1214–1228.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216–271, Vol. 2). MIT Press.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33–40.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need, 11.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., & Borgwardt, K. M. (2010). Graph kernels. *Journal of Machine Learning Research*, *11*, 1201–1242.

Weissweiler, L., Mahowald, K., & Goldberg, A. (2025). Linguistic generalizations are not rules: Impacts on evaluation of lms. *arXiv preprint arXiv:2502.13195*.

Wieling, M., & Montemagni, S. (2017). Exploring the role of extra-linguistic factors in defining dialectal variation patterns through cluster comparison. In *From semantics to dialectometry: Festschrift in honor of john nerbonne* (pp. 241–251). College Publications.

Wu, S., & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 833–844.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). Mt5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 483–498.