Multilinguality of Large Language Models From a Structural Perspective

Anonymous ACL submission

Abstract

Large language models (LLMs) have excelled in processing multiple languages through pre- and post-training on multilingual data, even though English dominates the training data. Prior work focusing on token representations has revealed how those LLMs process non-English text. Although these analyses have provided insightful findings, they fail to capture a structural view, which is an inherent property of language. In this study, we explore the multilinguality of LLMs through representational structural analysis. Our findings reveal that low-resource languages are structurally more different from English than high- and mid-resource languages, and that language-specific post-training alters their structures while preserving inter-language relationships.

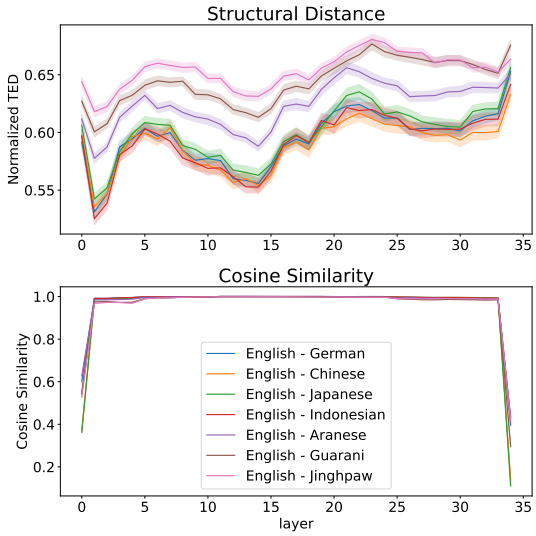


Figure 1: Representational structure distance and cosine similarity between English and other languages.

1 Introduction

Multilinguality of large language models (LLMs) is crucial to make LLMs helpful for everyone. While English is the dominant language in LLM pre-training data, LLMs exhibit multilingual capabilities (Winata et al., 2021; Gemma Team et al., 2025; Yang et al., 2025). LLMs have improved performance across multiple languages by learning languages through pre-training on multilingual corpora (Gemma Team et al., 2025; Yang et al., 2025) and continuous pre-training (Fujii et al., 2024). This multilingualism of these LLMs has attracted computational linguists, and prior work has addressed fundamental questions about how these LLMs handle diverse languages: how linguistic knowledge is internally represented (Brinkmann et al., 2025), whether there are cross-linguistic differences (Chang et al., 2022; Wu et al., 2024; Zhang et al., 2025), and whether a pivot language underlies multilingual processing (Wendler et al., 2024; Zhong et al., 2025).

Nonetheless, prior analyses have been conducted predominantly at the token (or subword) level (Wendler et al., 2024; Brinkmann et al., 2025; Zhong et al., 2025), although languages exhibit structures (Bybee, 2010) and diverse morphosyntactic characteristics (Song, 2018), and tokenizers do not account for these properties. Consequently, the semantic granularity of the minimal units processed by LLMs varies substantially across languages due to differences in tokenizers. This limits the token-level analysis to understand how LLMs process and represent multiple languages. This motivates an analysis that moves beyond the token level and adopts a holistic perspective that considers the entire input structure.

To address this limitation, we analyze multilingual representations from a structural view using StructLens (Sakajo et al., 2026), a method that derives tree structures from pairwise L2 distances between token representations. This framework enables comparison of entire languages from a tree-structure perspective in representation space. We measure structural distance in representation space across diverse languages, including low-resource

languages, on monolingual and multilingual pre-trained models and language-specific post-trained models. The resulting trees and their distance provide the organization of multilingual representations in the LLMs' representation space, the differences among languages, and the influence of language-specific post-training.

Our experiments reveal that a monolingual, English-centric model exhibits greater structural distance between English and other languages, whereas multilingual models display more uniform internal structures across high- and mid-resource languages while diverging in low-resource languages, despite showing high cross-lingual cosine similarity (see Figure 1). The results also show that a language that is syntactically and typologically similar to English is relatively close to English. Furthermore, language-specific post-training differentiates the internal structures of the target language from those of other languages, while preserving the structural distance across languages. We also demonstrate that representational structure distance captures inter-token relationships that conventional cosine similarity fails to reflect. Our findings reveal that a representational structure perspective yields distinct measures of language similarity based on the multilinguality of LLMs, their training data, and the representational characteristics of language-specific post-training.

2 Background and Related Work

2.1 Representations in Residual Streams

The pre-layer normalization (Xiong et al., 2020) variant of Transformer (Vaswani et al., 2017) refines internal representations incrementally through residual connections, which is also known as a *residual stream* (Elhage et al., 2021). Residual streams reflect models' hidden states, and prior research focusing on them has demonstrated mechanisms of LLMs (Elhage et al., 2021; Kamigaito et al., 2025). To interpret those states, several tools have been developed. For example, Logit lens (nostalgebraist, 2020) applies the language model head of LLMs for hidden states at each layer and interprets them as symbols. Another tool, StructLens (Sakajo et al., 2026), constructs spanning trees based on the L2 distance between hidden states at each layer, providing a holistic view of their internal structures in representation space. In the remainder of this section, we refer to hidden states in the residual streams as representations.

2.2 Language Adaptation for Multilinguality

The training data for most LLMs is predominantly English, and their performance in non-English languages is limited compared to English (Xuan et al., 2025). To fill this gap, substantial efforts have been made to transfer or extend those models' capabilities to other languages by adapting the models' vocabularies for target languages and continuous pre-training (Minixhofer et al., 2022, 2024; Yamaguchi et al., 2024, 2026; Remy et al., 2024; Fujii et al., 2024).

Continuous pre-training on target language data is an approach to achieving language adaptation, while keeping the performance in the source language (Wang et al., 2020; Fujii et al., 2024). For example, Fujii et al. (2024) develops language-specific language models by vocabulary expansion and continuous pre-training on English-centric models, achieving advanced performance in both English and the target language. We refer to continuous pre-training as language-specific post-training for convenience.

2.3 Multilinguality of Large Language Models

LLMs obtain multilingual capabilities through pre-training on multilingual data (Gemma Team et al., 2025; Yang et al., 2025) and language-specific continuous pre-training (Fujii et al., 2024). Previous work investigates how the multilingual LLMs internally represent and process different languages. Chang et al. (2022) analyzes the geometry of multilingual representations, showing that languages occupy similar linear subspaces. Wendler et al. (2024) track how representations evolve across layers when English-centric models process non-English input using Logit lens, and find that the latent concept in the models is biased toward English. In this line of work, Zhong et al. (2025) demonstrate that LLMs trained on balanced multiple language resources rely on those languages. Zhang et al. (2025) discover that models employ the same circuits to process the same syntactic features across languages. Brinkmann et al. (2025) train Sparse Autoencoders and find that models share features for grammatical categories.

These studies have revealed that LLMs possess language-agnostic and language-specific states through analysis of token representations and circuits. While prior work has focused on token-level representations and has revealed the underlying multilingual processes and the geometry of mul-

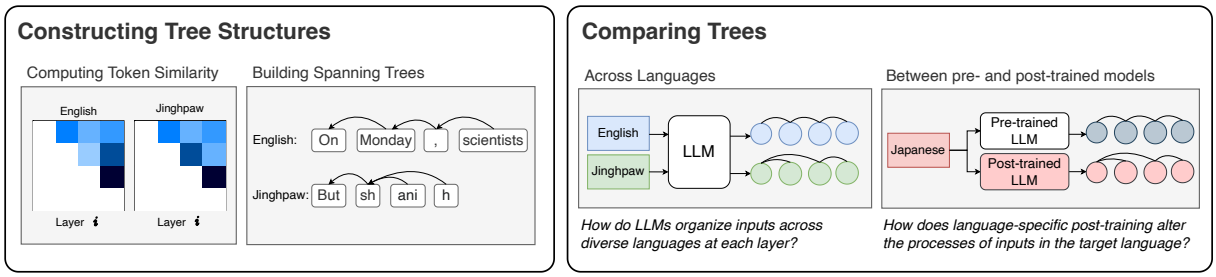


Figure 2: Overview of our research. We construct tree structures that summarize inter-token relationships in representation space and compare them across languages within a model and between pre- and post-trained models. This comparison reveals the organization of input tokens across languages and the influence of language-specific post-training.

tilingual representations, this approach may fail to capture inherent differences in representational structure. In this work, we aim to observe representations of entire input across languages from a structural perspective.

3 Task Formulation

We ask: (1) *how do LLMs organize inputs across diverse languages at each layer?* (2) *how does language-specific post-training alter the processes of inputs in the target language?* We examine these research questions by investigating the inter-token relationships formed in the models' representation space. To answer the first question, we compute the distance between languages in the LLM representation space using parallel corpora. For the second question, we measure the distance between the tree structures of pre-trained and language-specific post-trained models.

We compute the structural distance of token representation organization in the LLMs' representation space between languages for each layer by constructing spanning trees for each language using StructLens (Sakajo et al., 2026) and measuring the distance between them. StructLens summarizes inter-token relationships in representation space, and we can analyze the differences between internal structures of languages by comparing the resulting spanning trees, as shown in Figure 2.

3.1 Spanning Tree Construction

We construct a spanning tree from representations of each layer using StructLens. StructLens constructs maximum spanning trees based on inter-token similarity in representation space at each layer, summarizing the token representation organization.

Formally, given a token sequence of length n, let d denote the hidden dimension of a model and $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^d$ denote the i-th token's representation in the residual streams of the ℓ-th Transformer block, which are representation immediately after that block. StructLens constructs maximum spanning trees for each layer using the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, which represents inter-token similarity. For $i, j \in \{1, 2, \ldots, n\}$, the element of $\boldsymbol{A}$ is defined as:

$$A_{i,j} = \begin{cases} \dfrac{1}{1 + \|\boldsymbol{h}_i^{(\ell)} - \boldsymbol{h}_j^{(\ell)}\|} & \text{if } i < j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Following Sakajo et al. (2026), we build the maximum spanning trees using $\boldsymbol{A}$ and the algorithm introduced by Tarjan (1977) running in $O(n^2)$ time.

The resulting spanning trees summarize inter-token relationships within the representation space of each Transformer block in an LLM, providing a holistic view of each Transformer layer. Therefore, we can answer the research questions by analyzing the spanning trees across layers and across models.

3.2 Measuring Structural Distance

To measure the structural differences between two trees, we compute the Tree Edit Distance (TED) (Zhang and Shasha, 1989), normalized by the larger input lengths of the two trees for fair comparison, since different languages or tokenizers can yield different lengths of token sequences. TED measures the minimum cost required to transform one tree into another using three operations: insertion, deletion, and relabeling. Higher TED indicates lower similarity, whereas lower TED indicates higher similarity. We set the relabeling cost to 0 when computing TED to focus on structural differences rather than lexical mismatches. For computational efficiency, we use AP-TED (Pawlik and Augsten, 2016), an optimized variant of TED.

Feature	English	German	Japanese	Chinese	Indonesian	Aranese	Guarani	Jinghpaw
Language Family	Indo-European	Indo-European	Japanese	Sino-Tibetan	Austronesian	Indo-European	Tupian	Sino-Tibetan
Script	Latin	Latin	Kanji/Kana	Hanzi	Latin	Latin	Latin	Latin
Nominative-Accusative Alignment	✓	✓	✓		✓	✓	✓	✓
Word Order	SVO	(Mix)	SOV	SVO	SVO	SVO	SVO	SOV
Adposition	Prepositions	Prepositions	Postpositions	–	Prepositions	Prepositions	Postpositions	Postpositions
Order of Relative Clause and Noun	NRel	NRel	RelN	RelN	NRel	NRel	NRel	RelN

Table 1: Example of features of each language. In the order of relative clause and noun, NRel denotes the order of noun-relative clause, and RelN denotes the order of relative clause-noun.

4 Experimental Settings

4.1 Models, Languages, and Datasets

Models. We use Gemma3 4B PT (Gemma Team et al., 2025) and Qwen3 8B Base (Yang et al., 2025), which are multilingual pre-trained models, Olmo3 7B Base (Olmo et al., 2025), which is a monolingual pre-trained model, and Qwen3 Swallow 8B CPT (Fujii et al., 2024), which is a continuous post-trained model on Japanese data. We also use Qwen3 8B, a post-trained model of Qwen3 8B Base, for comparison with Qwen3 Swallow 8B CPT, as both are derived from Qwen3 8B Base.

Languages. We use English, German, Chinese, Japanese, and Indonesian as high- or mid-resource languages, and Aranese, Guarani, and Jinghpaw as low-resource languages. Table 1 shows examples of linguistic features based on WALS (Dryer and Haspelmath, 2013), Estigarribia (2020), and Institut d'Estudis Aranesi (2015). We use diverse languages in terms of language family and typological and syntactic features. This diverse language set enables the assessment of whether and how linguistic distance determines the language process in LLMs. For cross-lingual structural comparison, we compare English and Chinese with other languages using corresponding texts, since English is the dominant language in the training data and Qwen3 exhibits advanced performance on Chinese (Yang et al., 2025). To investigate the effect of language-specific post-training, we focus on English, Chinese, and Japanese.

Datasets. We utilize translation sentence pairs from FLORES+ (NLLB Team et al., 2024) and corresponding Jinghpaw data (Kurabe, 2013, 2017, 2020c,a,b; Taguchi et al., 2025). We use the development set of 997 samples.

Prompts and implementations. We report the detailed experimental settings, including prompts and implementation, in Appendix A.

4.2 Non-Structural Metrics

To further explore what the structural metric measures, we leverage non-structural metrics. We measure perplexity, machine translation performance, and cosine similarity. We also measure L2 distance. See Appendix A.4 for details.

Perplexity. To assess the model's familiarity with each language, we measure its perplexity. Perplexity is computed for each sample from FLORES+ and the Jinghpaw variant in the respective language, and then normalized by the input length to ensure a fair comparison across languages and tokenizers.

Machine translation. To investigate how much LLMs can relate one language to another, we perform machine translation (MT) experiments for each model. In this experiment, we use non-English languages as source languages and English as a target language. We evaluate the translated text using BLEU (Papineni et al., 2002; Post, 2018), chrF (Popović, 2015), and TER (Snover et al., 2006). We truncate the generated text by period to extract the first sentence for evaluation.

Cosine similarity To measure how close semantic representations of non-English languages are to English, we compute the sentence-wise cosine similarity for each layer. Formally, let $\boldsymbol{h}_{src}^{(\ell)}$ and $\boldsymbol{h}_{tgt}^{(\ell)}$ denote the ℓ-th layer's representations of the source language's and target language's inputs, respectively. The cosine similarity is computed by:

$$\frac{\bar{\boldsymbol{h}}_{src}^{(\ell)\top} \bar{\boldsymbol{h}}_{tgt}^{(\ell)}}{\|\bar{\boldsymbol{h}}_{src}^{(\ell)}\| \|\bar{\boldsymbol{h}}_{tgt}^{(\ell)}\|}, \quad \text{where } \bar{\boldsymbol{h}}^{(\ell)} = \frac{1}{|\boldsymbol{h}^{(\ell)}|} \sum \boldsymbol{h}_i^{(\ell)}. \tag{2}$$

4.3 Syntactic and Typological Metrics

To compute linguistic distance between languages, we utilize URIEL+ (Khan et al., 2025), a knowledge base that stores linguistic features as vector

Model	English	German	Japanese	Chinese	Indonesian	Aranese	Guarani	Jinghpaw
Gemma3 4B PT	$41.23_{\pm 52.20}$	$29.10_{\pm 42.86}$	$49.71_{\pm 52.02}$	$131.85_{\pm 751.11}$	$57.81_{\pm 89.31}$	$442.45_{\pm 2978.25}$	$360.08_{\pm 695.32}$	$860.28_{\pm 739.95}$
Qwen3 8B Base	$31.62_{\pm 36.02}$	$13.35_{\pm 11.58}$	$17.77_{\pm 12.67}$	$50.97_{\pm 64.19}$	$14.60_{\pm 11.35}$	$289.56_{\pm 831.69}$	$434.38_{\pm 381.41}$	$297.32_{\pm 179.87}$
Olmo3 7B Base	$31.08_{\pm 38.59}$	$19.76_{\pm 16.04}$	$10.46_{\pm 5.25}$	$12.21_{\pm 8.16}$	$33.92_{\pm 32.56}$	$588.81_{\pm 1282.25}$	$765.71_{\pm 579.38}$	$444.25_{\pm 271.71}$
Qwen3-8B	$53.04_{\pm 108.56}$	$19.56_{\pm 21.06}$	$26.78_{\pm 22.19}$	$81.29_{\pm 129.33}$	$21.30_{\pm 22.69}$	$561.61_{\pm 3422.74}$	$695.72_{\pm 736.64}$	$462.71_{\pm 315.02}$
Qwen3 Swallow 8B CPT v0.2	$28.24_{\pm 28.59}$	$14.81_{\pm 15.54}$	$14.14_{\pm 9.44}$	$91.03_{\pm 161.77}$	$16.39_{\pm 13.98}$	$305.87_{\pm 717.26}$	$503.91_{\pm 496.35}$	$325.03_{\pm 196.47}$

Table 2: Perplexity comparison across models and subsets. Subscripts denote standard deviation. Perplexity is normalized by input length for fair comparison across different input lengths.

Model	German	Japanese	Chinese	Indonesian	Aranese	Guarani	Jinghpaw
Gemma3 4B PT	35.36	21.31	22.82	33.80	16.39	3.88	1.22
Qwen3 8B Base	36.90	22.25	24.99	36.07	19.48	5.41	3.60
Olmo3 7B Base	35.42	19.44	21.87	31.70	13.12	2.40	1.99

Table 3: MT evaluation results of BLEU. The higher values are better.

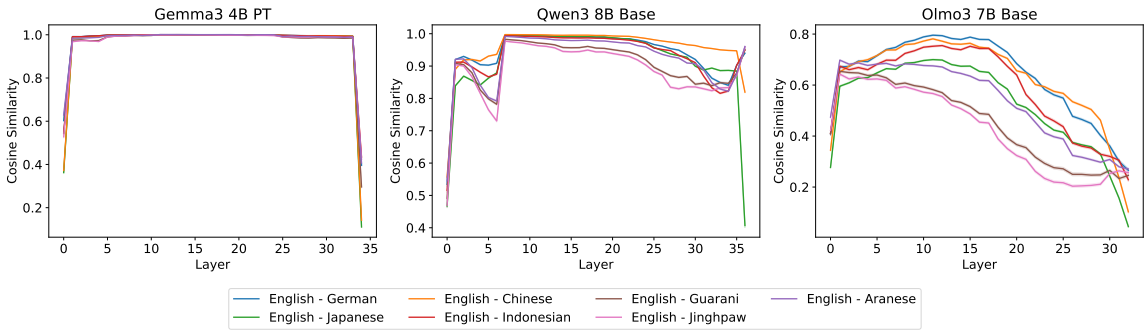


Figure 3: Cosine similarity on pre-trained models.

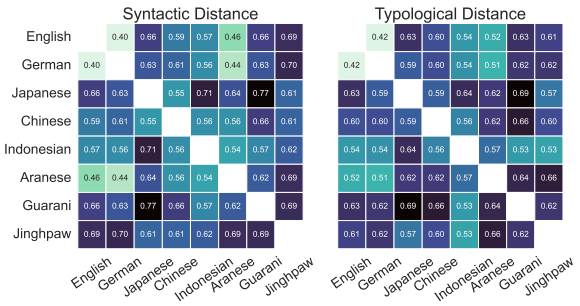


Figure 4: Syntactic and typological distance. Brighter colors represent a smaller distance, and darker colors represent a larger distance.

representations. URIEL+ computes the angle between the feature vectors of two languages, following Toossi et al. (2024), thereby providing the language distance with respect to syntax and typology. We measure both syntactic and typological distance between languages.

Formally, let m denote the feature dimension, and $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^m$ denote feature vectors of two languages. The distance of these languages is defined as:

$$\frac{1}{\pi} \arccos \frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}. \quad (3)$$

Following Khan et al. (2025), we use SOFT-IMPUTE (Mazumder et al., 2010) to impute missing features before computing distance.

5 Results and Discussions

5.1 Non-Structural Metrics

Perplexity. Table 2 reports the mean perplexity for each language. These results indicate that the low-resource languages Aranese, Guarani, and Jinghpaw are unfamiliar to the models.

Machine Translation. Table 3 shows the machine translation results measured by BLEU. The results for other MT metrics are reported in Appendix B.1, which show patterns similar to BLEU. MT results indicate that Guarani and Jinghpaw are unfamiliar languages for the models. While Aranese exhibits higher perplexity, models show better MT performance for this language than for the other low-resource languages.

Cosine similarity. Figure 3 shows cosine similarity for pre-trained models. Cosine similarity between English and other languages at intermediate layers of multilingual pre-trained models, Gemma3

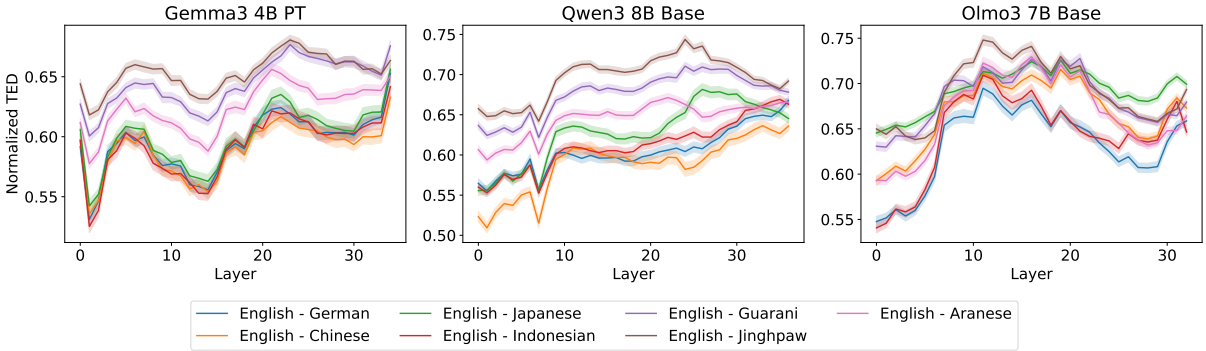


Figure 5: TED between English and other languages. The error bars indicate the 95 % confidence interval.

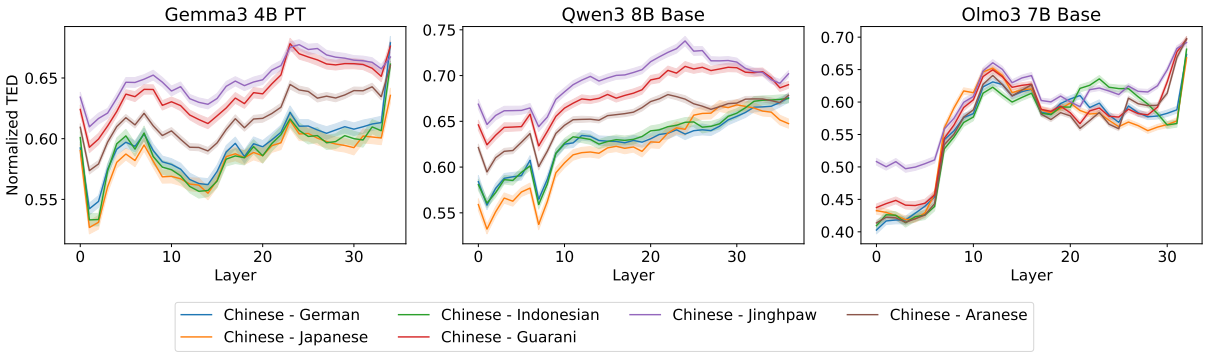


Figure 6: TED between Chinese and other languages. The error bars indicate the 95 % confidence interval.

4B PT and Qwen3 8B Base, is close to 0.9 across all languages. A monolingual model, Olmo3 7B Base, exhibits lower cosine similarity for low-resource languages, whereas it shows relatively high cosine similarity for high-/mid-resource languages. Results on post-trained models are provided in Appendix B, indicating that post-training has less impact on cosine similarity.

Syntactic and typological distance. Figure 4 shows the syntactic and typological (featural) distance of each language pair. German and Aranese are relatively close to English, while other languages are more distant from English, which differs from other metrics. As for closeness, Japanese and Indonesian are relatively close to Chinese.

5.2 Structural Metric

Figure 5 shows that there are two clusters of languages, high-/mid-resource languages and low-resource languages, based on TED in Gemma3 4B PT and Qwen3 8B Base, while Olmo3 7B Base shows equivalent results across languages, which are different from MT results and cosine similarity. While linguistic typological features, e.g., word order and morphological features (see Table 1), can affect TED between languages, even languages with diverse typological features and scripts (i.e., German, Chinese, Japanese, and Indonesian) show a close distance to English. Given that Gemma3 4B PT and Qwen3 8B Base are multilingual models, and Olmo3 7B Base is a monolingual model, the results indicate that pre-training data enable models to distinguish the structural distance. Figure 6 reports TED between Chinese and other languages, yielding similar results with Figure 5. We also investigate larger models within each model family, as shown in Figure 8, which exhibit similar patterns. We further visualize TED for each language pair in the middle layer of Figure 7 as a case study, showing that Gemma3 4B PT exhibits a cluster among high- and mid-resource languages, and that Olmo3 7B Base exhibits distinct English and non-English patterns. Appendix B.2 provides the TED matrices for the low and high layers.

Figure 9a shows the TED between pre- and post-trained models for English, Chinese, and Japanese, and Figure 9b shows TED between English and Chinese, Japanese for pre- and post-training models. We also show TED for each language pair in Figure 10. These figures indicate that the Japanese-specific post-trained model exhibits higher TED in Japanese, while TED between English and Chinese

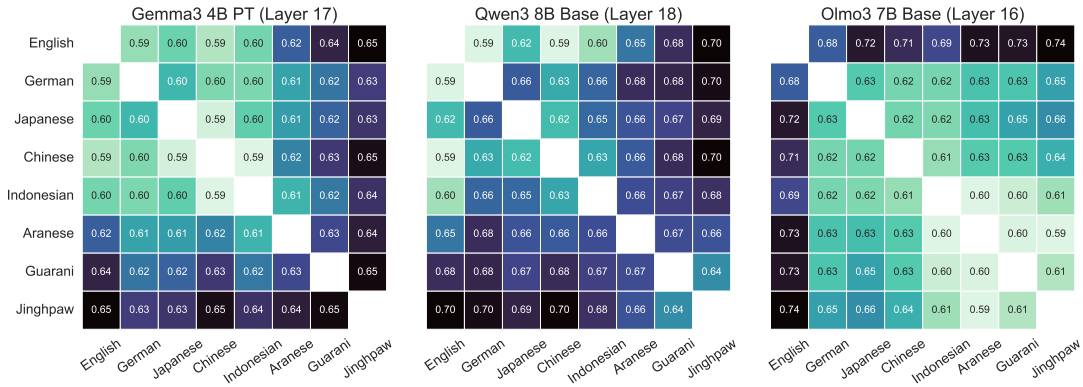


Figure 7: Heatmaps of TED in a middle layer for each language pair. Brighter colors represent a smaller distance, and darker colors represent a larger distance.

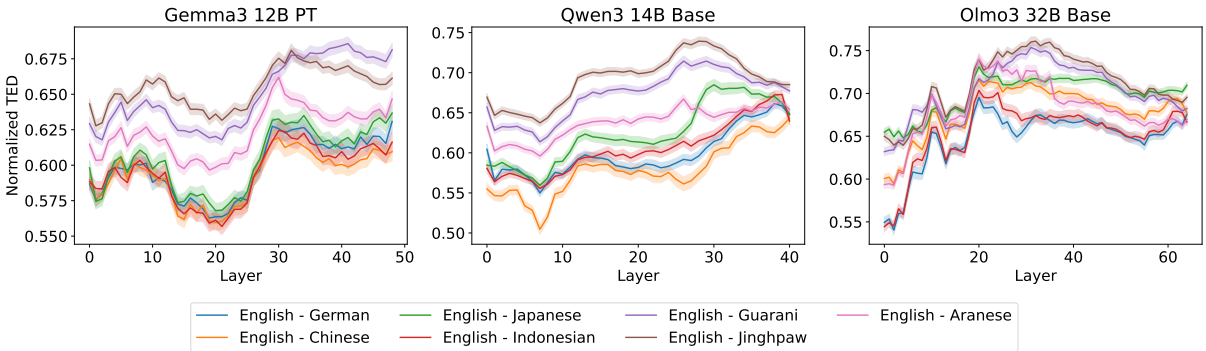


Figure 8: TED between English and other languages for larger models. The error bars indicate the 95 % confidence interval.

and between English and Japanese is consistent with that of the pre-trained model.

5.3 Difference between Cosine Similarity and Structural Distance

The cosine similarity measures the difference between the centroids of each language's inputs, and the mean representations fail to capture relationships among token representations. To formalize this difficulty, we decompose the residual streams into the mean representation and the difference from it. Let $\boldsymbol{\delta}_i^{(\ell)}$ denote the difference representation from $\bar{\boldsymbol{h}}^{(\ell)}$, which is defined in Equation 2. We decompose the representation $\boldsymbol{h}_i^{(\ell)}$ as:

$$\boldsymbol{h}_i^{(\ell)} = \bar{\boldsymbol{h}}^{(\ell)} + \boldsymbol{\delta}_i. \tag{4}$$

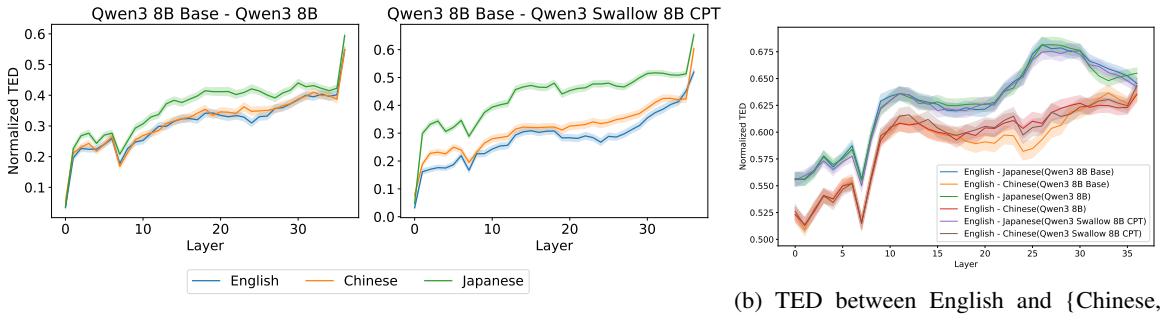
Given that cosine similarity utilizes only $\bar{\boldsymbol{h}}^{(\ell)}$, it ignores the set of deviations of $\boldsymbol{\delta}_i$, limiting to measure the difference that is formed by $\boldsymbol{\delta}_i$.

The representational structure distance computes the difference between the two spanning trees that are based on the L2 distance $\|\boldsymbol{h}_i^{(\ell)} - \boldsymbol{h}_j^{(\ell)}\|$ (see Equation 1). By applying the decomposition of representations of Equation 4, the L2 distance can be written as $\|\boldsymbol{\delta}_i - \boldsymbol{\delta}_j\|$. That is, the structural distance reflects the internal relational differences, whereas the cosine similarity measure abstracts differences via mean pooling.

In light of this difference between cosine similarity and structural distance, the TED results of multilingual models shown in Figures 5 provide distinct differences among languages, while the cosine similarity ones shown in Figure 3 exhibit almost no differences among languages. This difference indicates that the abstract representations are aligned with English and that the internal relational differences depend on the language in multilingual models.

5.4 Resources and Linguistic Features

Resource influence. Multilingual models exhibit a TED cluster comprising high- and mid-resource languages. Recall that the languages used in this study are classified into two clusters. high- and mid-resource languages (i.e., German, Japanese, Chinese, and Indonesian) and low-resource languages (i.e., Aranese, Guarani, and Jinghpaw), the cluster



(a) TED between pre- and post-trained models.

(b) TED between English and {Chinese, Japanese} for pre- and post-trained models.

Figure 9: TED for post-trained models. The error bars indicate the 95 % confidence interval

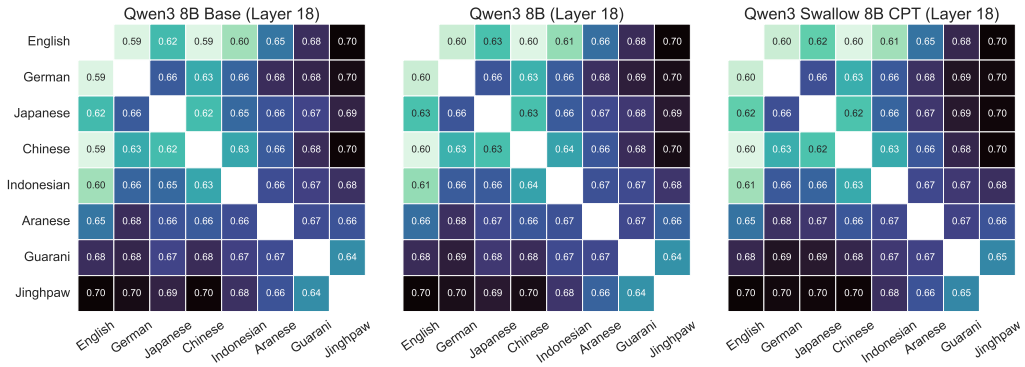


Figure 10: Heatmaps of TED in a middle layer of pre- and post-trained models for each language pair.

can be attributed to the high-resource languages. Furthermore, the distinct difference between English and other languages in Figure 7 supports this finding, since Olmo3 7B Base is a monolingual model trained on English data.

Linguistic feature influence. Among the three low-resource languages in the multilingual models' results, Figures 5 and 6 show that Aranese exhibits the lowest TED, while Jinghpaw exhibits the highest TED. Given that Aranese is closer to English than other low-resource languages, linguistic similarity influences TED for low-resource languages.

5.5 How Do LLMs Represent Diverse Languages?

The cosine similarity results in Figure 3 and TED results in Figure 5 indicate that LLMs organize input sequences according to the language through each Transformer block and align abstract semantic representations across languages. Multilingual LLMs exhibit relatively similar structural patterns for high- and mid-resource languages. For unfamiliar, low-resource languages, the linguistically closer languages to English are easier to structurally align with English in representation space.

5.6 How Does Language-Specific Post-Training Alter the Representations?

Figure 9a shows that Qwen3 Swallow 8B CPT, a Japanese-specific post-trained model, exhibits divergence in TED from the base model within Japanese. Conversely, the representational structure distance between English and other languages, including Japanese, remains invariant (see Figure 9b). These findings suggest that language-specific post-training modifies target language structures while preserving inter-lingual structural distances, allowing performance improvement in that language.

6 Conclusion

In this study, we investigate the multilinguality of LLMs from a representational structure perspective. Our experiments indicate that the structural distance between languages depends on the models' familiarity with those languages, and that language-specific post-training refines the internal structure of that language while preserving its relationships to other languages. Our findings highlight that LLMs distinguish familiar and unfamiliar languages in their internal structures and, post-training, retain the relationship between languages.

Limitations

Language selection. We use a set of languages: English, German, Chinese, Japanese, Aranese, Guarani, and Jinghpaw. Although this set of languages may limit the generality of our work, this set includes high-, mid-, and low-resource languages, multiple scripts, and typological diversity, allowing us to investigate multilinguality of LLMs from resource, scripts, and typological aspects. This language set enables us to find that multilingual LLMs exhibit high similarity in representational structure across high- and low-resource languages despite their diverse typological features, indicating that this set is sufficient for this work's scope.

Scope of representational structures. We compare representational structures across languages and models. The tree structures are model-internal summaries of inter-token relationships in LLM representation space, not linguistic parse trees. Therefore, we interpret our findings as evidence about how LLMs organize token representations, and leave direct comparison with linguistic structures to future work.

Training data availability. As of writing, the multilingual models used in this study have not released their training data. This limits our investigation into the influence of multilingual training data on the structural distance. Testing multilingual models trained on controlled multilingual text can provide further insights beyond our work; this is a future direction for this work. Our primary goal is to investigate how LLMs that exhibit advanced performance represent and process multiple languages, and this work provides insights into this.

Applications. Although we believe that our findings are helpful for applications, e.g., language adaptation, this study aims to reveal how LLMs process multiple languages from a structural perspective that captures holistic aspects, following prior efforts on the multilinguality of LLMs.

Ethical Considerations

Reproducibility. We provide experimental settings, including software and computational resources, in Section 4 and Appendix A.

Use of AI assistants. We use LLMs for a coding assistant and proofreading of the manuscript, and all the edits of code and paper are verified by the authors.

Data content. Although we use FLORES+, a widely used dataset, and its corresponding dataset, we have not assessed whether either dataset contains personally identifiable or offensive content, as they contain multiple languages, including low-resource languages.

License. We follow the license of the models, datasets, and software used in this study.

References

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.

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.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A Mathematical Framework for Transformer Circuits.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guarani*. Grammars of World and Minority Languages. UCL Press.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Institut d'Estudis Aranesi. 2015. *Gramatica basica der occitan aranés*. Institut d'Estudis Aranesi, Acadèmia aranesa dera lengua occitana.

Hidetaka Kamigaito, Ying Zhang, Jingun Kwon, Katsuhiko Hayashi, Manabu Okumura, and Taro Watanabe. 2025. Diversity of transformer layers: One aspect of parameter scaling laws. *Preprint*, arXiv:2505.24009.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Keita Kurabe. 2013. Kachin folktales told in Jinghpaw.

Keita Kurabe. 2017. Kachin culture and history told in Jinghpaw.

Keita Kurabe. 2020a. *Intensive Language Course 2019 Jinghpaw Textbook 1 - An Introduction to Jinghpaw Grammar*. Research Institute for Languages and Cultures of Asia and Africa.

Keita Kurabe. 2020b. *Intensive Language Course 2019 Jinghpaw Textbook 2 - Jinghpaw Reader*. Research Institute for Languages and Cultures of Asia and Africa.

Keita Kurabe. 2020c. *Intensive Language Course 2019 Jinghpaw Textbook 3 - A Dictionary of Jinghpaw Usage*. Research Institute for Languages and Cultures of Asia and Africa,.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Benjamin Minixhofer, Edoardo Ponti, and Ivan Vulić. 2024. Zero-shot tokenizer transfer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

nostalgebraist. 2020. Interpreting GPT: The logit lens.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mateusz Pawlik and Nikolaus Augsten. 2016. Tree edit distance: Robust and memory-efficient. *Information Systems*, 56:157–173.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of LLMs for low-resource NLP. In *First Conference on Language Modeling*.

Haruki Sakajo, Frederikus Hudi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2026. Structlens: A structural lens for language models via maximum spanning trees. *Preprint*, arXiv:2603.03328.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jae Jung Song. 2018. *Linguistic Typology*. Oxford Textbooks in Linguistics. Oxford University Press, London, England.

Chihiro Taguchi, Seng Mai, Keita Kurabe, Yusuke Sakai, Georgina Agyei, Soudabeh Eslami, and David Chiang. 2025. Languages still left behind: Toward a better multilingual machine translation benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages

20131–20143, Suzhou, China. Association for Computational Linguistics.

R. E. Tarjan. 1977. Finding optimum branchings. *Networks*, 7(1):25–35.

Hasti Toossi, Guo Huai, Jinyu Liu, Eric Khiu, A. Seza Doğruöz, and En-Shiun Lee. 2024. A reproducibility study on quantifying language similarity: The impact of missing values in the URIEL knowledge base. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 233–241, Mexico City, Mexico. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Di Wu, Yibin Lei, Andrew Yates, and Christof Monz. 2024. Representational isomorphism and alignment of multilingual large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14074–14085, Miami, Florida, USA. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. An empirical study on cross-lingual vocabulary adaptation for efficient language model inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6760–6785, Miami, Florida, USA. Association for Computational Linguistics.

Atsuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2026. How can we effectively expand the vocabulary of llms with 0.01gb of target language text? *Computational Linguistics*, 52(1):295–330.

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.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18:1245–1262.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The same but different: Structural similarities and differences in multilingual language modeling. In *The Thirteenth International Conference on Learning Representations*.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What language do non-English-centric large language models think in? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26333–26346, Vienna, Austria. Association for Computational Linguistics.

A Experimental Details

A.1 Models and Datasets

Table 4 is the mapping table of model names used in our experiments and their HuggingFace IDs. We

obtain FLORES+ data from HuggingFace[1] and corresponding Jinghpaw data from Taguchi et al. (2025)[2]. We provide the licenses for these artifacts in Table 5.

A.2 Implementations and Computational Resources

In our experiments, we use HuggingFace Transformers (Wolf et al., 2020) to load models and datasets. We utilize SACREBLEU (Post, 2018) and its official implementation[3] to compute BLEU, chrF, and TER. To run STRUCTLENS, we use the official implementations[4]. For the AP-TED algorithm, we utilize the Python implementation[5]. We use a single NVIDIA GeForce RTX 3090 GPU for the experiments involving the models with less than 8B, a single NVIDIA 6000 Ada GPU for experiments involving the Gemma3 12B PT and Qwen3 14B Base, and a single NVIDIA H100 GPU for Olmo3 32B Base.

A.3 Prompt for Machine Translation

We use the following prompt for machine translation experiments.

> Translate the following {source language} text into {target language}. Output only the translation:
> {Text}
> Translation:

A.4 Non-Structural Metrics

L2 distance. We also measure L2 distance to investigate whether norm information is meaningful for representation analysis. The L2 distance between source and target languages is provided as:

$$\|\bar{\boldsymbol{h}}_{src}^{(\ell)} - \bar{\boldsymbol{h}}_{tgt}^{(\ell)}\|. \quad (5)$$

B Detailed Experimental Results

B.1 Non-Structural Metrics

Token length. We also measure token length, which is provided in Table 6, showing that low-resource languages exhibit larger token length than other languages in multilingual models.

[1] https://hf.co/datasets/openlanguagedata/flores_plus
[2] https://github.com/ctaguchi/LSLB
[3] https://github.com/mjpost/sacrebleu
[4] https://github.com/naist-nlp/structlens
[5] https://github.com/JoaoFelipe/apted

Model	HuggingFace ID
Gemma3 4B PT	google/gemma-3-4b-pt
Gemma3 12B PT	google/gemma-3-12b-pt
Qwen3 8B Base	Qwen/Qwen3-8B-Base
Qwen3 14B Base	Qwen/Qwen3-14B-Base
Qwen3 8B	Qwen/Qwen3-8B
Olmo3 7B Base	allenai/Olmo-3-1025-7B
Olmo3 32B Base	allenai/Olmo-3-1125-32B
Qwen3 Swallow 8B CPT	tokyotech-llm/Qwen3-Swallow-8B-CPT-v0.2

Table 4: Model names and HuggingFace IDs.

Artifact	License
Gemma3 4B PT	Gemma
Gemma3 12B PT	Gemma
Qwen3 8B Base	Apache 2.0
Qwen3 14B Base	Apache 2.0
Qwen3 8B	Apache 2.0
Olmo3 7B Base	Apache 2.0
Olmo3 32B Base	Apache 2.0
Qwen3 Swallow 8B CPT	Apache 2.0
FLORES+	CC BY-SA 4.0
FLORES+ (Jinghpaw)	CC-BY-SA-NC

Table 5: Artifacts license.

MT results (detail). Table 7 shows the MT performance for each metric. The models exhibit a better score for Aranese among low-resource languages, which suggests that linguistic similarity to English influences the MT performance.

Cosine similarity for post-trained models. Figure 11 post-training results indicate that post-training has less influence on cosine similarity.

L2 distance. Figures 12 and 13 show L2 distance for pre- and post- trained models, respectively. Given the nature of residual streams, the monotonic increments of L2 distance are obvious.

B.2 Structural Metric

We visualize TED for each language pair in low and high layers of pre- and post-trained models in Figures 14 and 15, and that in low, middle, and high layers in Figure 16. While smaller models exhibit consistent structural similarity patterns across layers, the monolingual larger model shows different patterns, where distinct English patterns shown in the smaller model disappear in the middle and high layers.

Model	English	German	Japanese	Chinese	Indonesian	Aranese	Guarani	Jinghpaw
Gemma3 4B PT	$26.96_{\pm 8.77}$	$35.12_{\pm 11.76}$	$32.31_{\pm 11.41}$	$29.38_{\pm 11.89}$	$30.59_{\pm 10.12}$	$45.91_{\pm 15.31}$	$51.13_{\pm 17.70}$	$54.52_{\pm 18.92}$
Qwen3 8B Base	$26.33_{\pm 8.94}$	$40.79_{\pm 14.19}$	$37.03_{\pm 13.57}$	$26.98_{\pm 11.79}$	$40.59_{\pm 13.58}$	$48.78_{\pm 16.51}$	$54.96_{\pm 19.28}$	$60.07_{\pm 21.13}$
Olmo3 7B Base	$25.88_{\pm 8.51}$	$40.74_{\pm 14.06}$	$58.64_{\pm 20.30}$	$48.90_{\pm 18.88}$	$40.13_{\pm 13.31}$	$48.68_{\pm 16.49}$	$56.34_{\pm 19.91}$	$59.79_{\pm 21.06}$

Table 6: Token length statistics across models and subsets. Subscripts denote standard deviation.

Model	German → English			Japanese → English			Chinese → English			Indonesian → English			Aranese → English			Guarani → English			Jinghpaw → English		
	BLEU ↑	CHRF ↑	TER ↓	BLEU ↑	CHRF ↑	TER ↓	BLEU ↑	CHRF ↑	TER ↓	BLEU ↑	CHRF ↑	TER ↓	BLEU ↑	CHRF ↑	TER ↓	BLEU ↑	CHRF ↑	TER ↓	BLEU ↑	CHRF ↑	TER ↓
Gemma3 4B PT	35.36	60.42	53.79	21.31	48.85	73.37	22.82	50.14	70.32	33.80	59.47	57.21	16.39	44.15	84.00	3.88	23.36	146.31	1.22	15.46	234.59
Qwen3 8B Base	36.90	62.09	51.04	22.25	51.80	69.50	24.99	53.76	66.24	36.07	62.56	51.09	19.48	48.83	76.56	5.41	27.20	103.62	3.60	23.08	118.63
Olmo3 7B Base	35.42	61.17	53.38	19.44	48.11	75.48	21.87	49.91	71.76	31.70	57.94	57.19	13.12	41.95	91.02	2.40	22.23	181.67	1.99	17.28	173.27

Table 7: MT evaluation results. Higher is better (↑) for BLEU/CHRF, lower is better (↓) for TER.

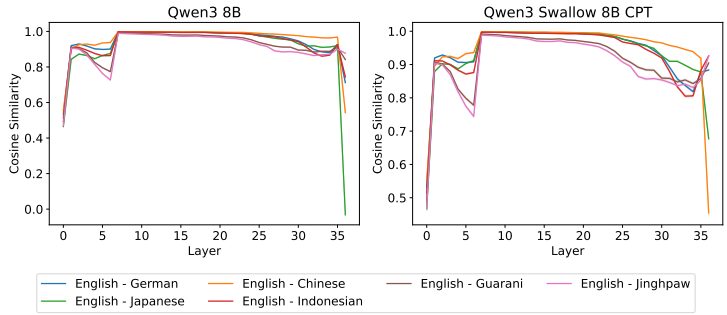


Figure 11: Cosine similarity on post-trained models.

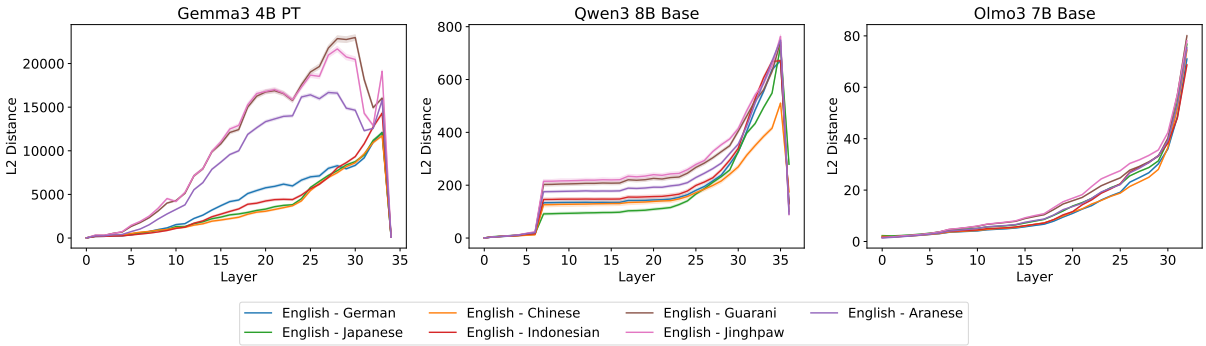


Figure 12: L2 distance on pre-trained models

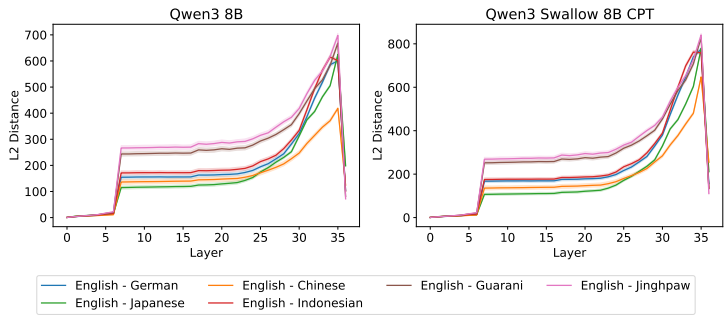
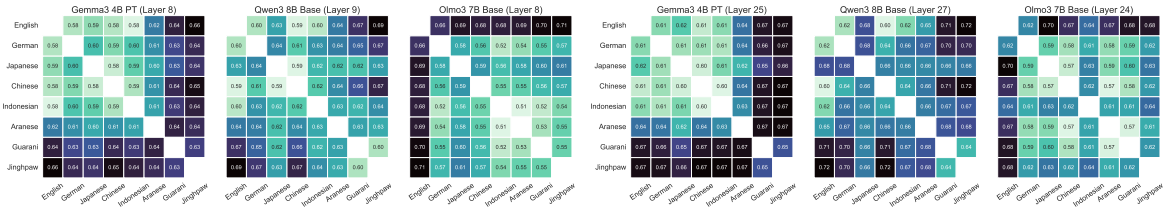
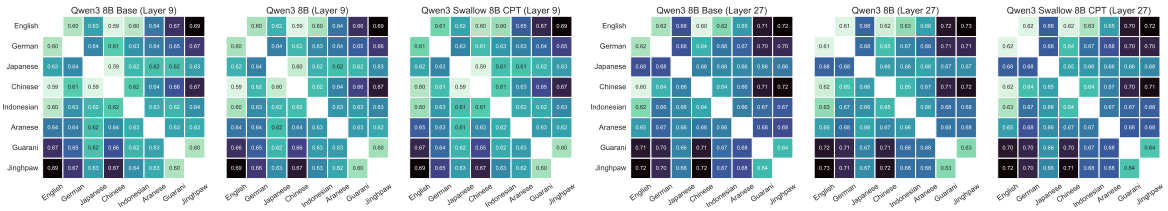


Figure 13: L2 distance on post-trained models



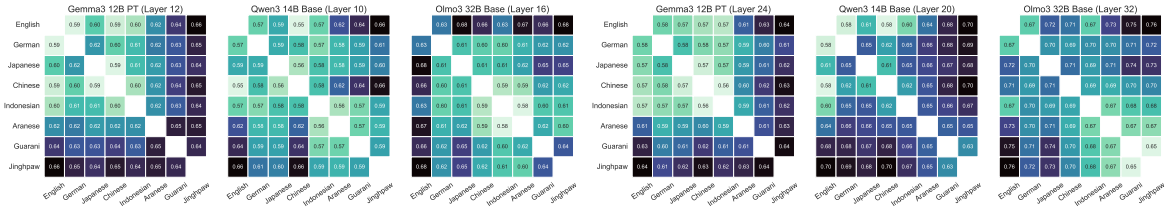
(a) Low layers. (b) High layers.

Figure 14: Heatmaps of TED for each language pair.

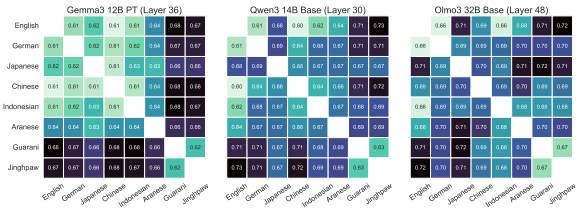


(a) Low layers. (b) High layers.

Figure 15: Heatmaps of TED in pre- and post-trained models.



(a) Low layers. (b) Middle layers.



(c) High layers.

Figure 16: Heatmaps of TED in larger models.