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-training on multiple language data and/or post-training, even though English dominates the training data. Prior work focusing on Attention or 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 structural analysis. Our findings reveal that low-resource languages are more different from English structurally than high- and mid-resource languages and that language-specific post-training alters the structures while preserving inter-language relationships.

1 Introduction

Multilinguality of large language models (LLMs) is crucial to make LLMs helpful for everyone. LLMs have achieved advanced performance not only in English but also across multiple languages through pre-training on multilingual corpora (Team et al., 2025; Yang et al., 2025) and/or post-training (Fujii et al., 2024). Prior work has addressed fundamental questions about how LLMs handle diverse languages: how linguistic knowledge is internally represented, whether there are cross-linguistic differences, and whether a pivot language underlies multilingual processing.

Nonetheless, prior analyses have been conducted predominantly at the token (subword) level, although languages exhibit diverse morphosyntactic characteristics, 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 these characteristics. This motivates an analysis that moves beyond the token level and adopts a holistic perspective that considers the entire input structure.

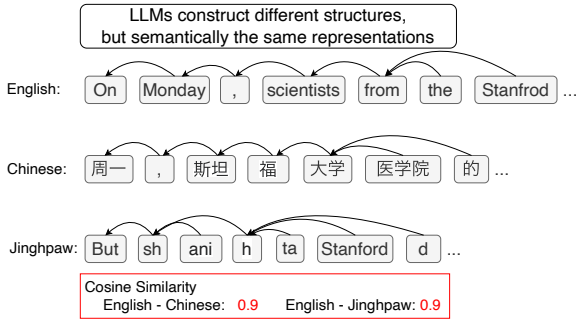

Figure 1: Internal structures at the beginning of the inputs of each language and cosine similarity.

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 representations. We measure structural distance among seven languages, including low-resource languages, on monolingual and multilingual pre-trained models and post-trained models. 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 (Figure 1). Furthermore, language-specific post-training differentiates the internal structures of the target language from those of other languages, while preserving the structural similarity across languages. We also demonstrate that structural similarity captures inter-token relationships that conventional cosine similarity fails to reflect. Our findings reveal that a structural 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

Residual streams of large language models. 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 hidden states at each layer and provides holistic internal structures of them.

Language adaptation for multilinguality. The training data of most LLMs is predominantly English, and substantial efforts have been made to extend those models to other languages. Continuous pre-training on target language data is an approach to achieve language adaptation (Fujii et al., 2024; Remy et al., 2024).

Multilinguality of large language models. Previous work investigates how 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 representations at the token level, this approach may fail to capture inherent structural differences. 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 process inputs across diverse languages at each layer?* (2) *how does language-specific post-training alter the processing of inputs in the target language?* To examine the first question, we compute the similarity of the latent structures between languages using parallel corpora. For the second question, we measure the similarity of these structures between pre-trained and post-trained models. We compute the structural similarity between languages for each layer by constructing spanning trees for each language via StructLens (Sakajo et al., 2026) and measuring their similarity.

3.1 Spanning Tree Construction

We construct a spanning tree from representations of each layer using StructLens. Let $\boldsymbol{h}_i^{(\ell)}$ denote the residual streams of the ℓ-th Transformer block, which are the output of that block. Given a token sequence of length n, StructLens constructs maximum spanning trees for each layer using the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ defined as:

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

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

3.2 Measuring Structural Similarity

To measure the similarity between two spanning trees, we compute the Tree Edit Distance (TED) (Zhang and Shasha, 1989) with normalization by dividing by the larger of the two languages' input lengths for fair comparison. Higher TED indicates lower similarity, whereas lower TED indicates higher similarity. To compare multiple language inputs, we set the relabelling cost to 0 when computing TED. We use AP-TED (Pawlik and Augsten, 2016), a variant of TED, for efficiency.

4 Experimental Settings

4.1 Models, Languages, and Datasets

We use Gemma3 4B PT (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

Model	German → English			Japanese → English			Chinese → English			Indonesian → English			Guarani → English			Jinghpaw → English		
	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	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	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	2.40	22.23	181.67	1.99	17.28	173.27

Table 1: MT evaluation results. Higher is better (↑) for BLEU/chrF, lower is better (↓) for TER.

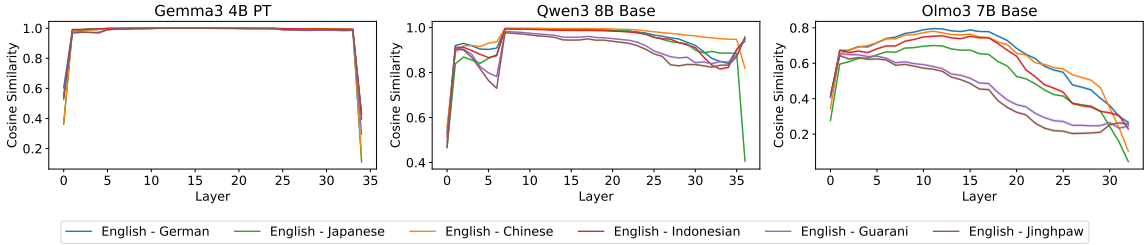

Figure 2: Cosine similarity on pre-trained models.

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.

As for languages and datasets, we use English, German, Chinese, Japanese, and Indonesian as high- or mid-resource languages, and Guarani and Jinghpaw as low-resource languages. We utilize translation sentence pairs from FLORES+ (NLLB Team et al., 2024) and corresponding Jinghpaw data (Kurabe, 2013, 2017, 2020a,b,c; Taguchi et al., 2025). We use the dev set of 997 samples.

We report details of experimental settings, e.g., prompts, in Appendix A.

4.2 Non-Structural Metrics

To further explore what the structural metric measures, we leverage non-structural metrics.

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.

Other metrics. We also measure perplexity and L2 distance. See Appendix A.4 for details.

5 Results and Discussions

5.1 Non-Structural Metrics

Machine Translation. Table 1 shows the machine translation results. MT results indicate that Guarani and Jinghpaw are unfamiliar languages for the models. We observed similar trends when measured by perplexity as provided in Appendix B.

Cosine similarity. Figure 2 shows cosine similarity for pre-trained models. Cosine similarity between English and other languages at intermediate layers of multilingual pre-trained models, Gemma3 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.

5.2 Structural Metric

Figure 3 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. Even though the languages we use have diverse typological features and scripts, they show close structural similarities to English in multilingual models. 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 between structural similarities.

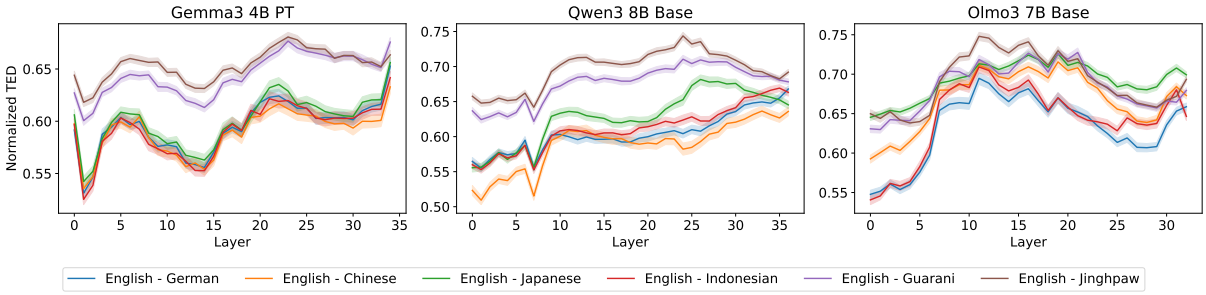

Figure 3: TED between English and other languages. The error bars indicate the 95 % confidence interval. Similar results of TED between Chinese and other languages are provided in Figure 9 in Appendix B.2.

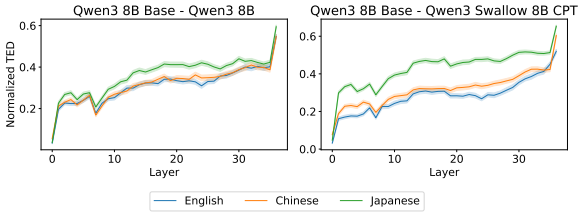

Figure 4: TED between pre- and post-trained models.

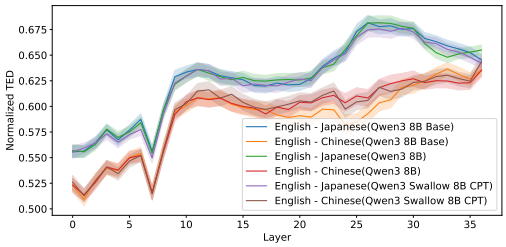

Figure 5: TED between English and {Chinese, Japanese} for pre- and post-trained models.

5.3 Difference between Cosine Similarity and Structural Similarity

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 as:

$$\boldsymbol{h}_i^{(\ell)} = \bar{\boldsymbol{h}}^{(\ell)} + \boldsymbol{\delta}_i \quad \text{where } \bar{\boldsymbol{h}}^{(\ell)} = \frac{1}{|\boldsymbol{h}^{(\ell)}|} \sum \boldsymbol{h}_i^{(\ell)}. \tag{2}$$

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 structural similarity computes the difference between the two spanning trees that are based on the L2 distance $\|\boldsymbol{h}_i^{(\ell)} - \boldsymbol{h}_j^{(\ell)}\|$ (Eq. 1). By applying the decomposition of representations (Eq. 2), the L2 distance can be denoted as $\|\boldsymbol{\delta}_i - \boldsymbol{\delta}_j\|$. That is, the structural similarity reflects the internal relational differences, whereas the cosine similarity measure abstracts differences via mean pooling.

5.4 How Do LLMs Process Diverse Languages?

The differences between cosine similarity (Figure 2) and TED (Figure 3) indicate that LLMs organize input sequences through each Transformer block and align abstract semantic representations across languages, and organizing unfamiliar languages requires higher effort for the models.

5.5 How Does Language-Specific Post-Training Alter the Process?

Figure 4 shows that Qwen3 Swallow 8B CPT, a Japanese-specific post-trained model, exhibits divergence in TED from the base model within Japanese. Conversely, the structural distance between English and other languages, including Japanese, remains invariant (see Figure 5). 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 structural 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 hold the relationship between languages.

Limitations

Language selection. We use a limited set of languages: English, German, Chinese, Japanese, Guarani, and Jinghpaw. Although this set of languages may limit the generality of our work, this language 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.

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 similarity. Testing multilingual models trained on controlled multilingual text can provide further insights, in addition to our work, although training such models from scratch requires substantial computational resources.

Applications. This study identifies structural characteristics of multilingual models but does not address their potential for 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, which can capture holistic aspects, following prior efforts for 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.

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.

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.

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*.

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.

Keita Kurabe. 2013. Kachin folktales told in Jinghpaw.

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

Keita Kurabe. 2020a. *A Dictionary of Jinghpaw Usage*. The Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Keita Kurabe. 2020b. *An Introduction to Jinghpaw Grammar*. The Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

Keita Kurabe. 2020c. *Jinghpaw Reader*. The Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.

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.

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.

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.

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.

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.

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.

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.

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 Detail Experimental Settings

A.1 Models and Datasets

Table 2 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].

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 AP-TED algorithm, we utilize the Python implementation[5]. We use a single NVIDIA GeForce RTX 3090 GPU for all experiments.

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

Perplexity. To assess the model's understanding of 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.

[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
Qwen3 8B Base	Qwen/Qwen3-8B-Base
Qwen3 8B	Qwen/Qwen3-8B
OLMo2 7B Base	allenai/OLMo-2-1124-7B
Qwen3 Swallow 8B CPT	tokyotech-llm/Qwen3-Swallow-8B-CPT-v0.2

Table 2: Model names and HuggingFace IDs.

Cosine similarity As explained in Section 4.2, we measure cosine similarity between mean representations. 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 as:

$$\frac{\bar{\boldsymbol{h}}_{src}^{(\ell)\top}\bar{\boldsymbol{h}}_{tgt}^{(\ell)}}{\|\bar{\boldsymbol{h}}_{src}^{(\ell)}\|\|\bar{\boldsymbol{h}}_{tgt}^{(\ell)}\|} \quad \text{where } \bar{\boldsymbol{h}}^{(\ell)} = \frac{1}{|\boldsymbol{h}^{(\ell)}|}\sum \boldsymbol{h}_i^{(\ell)}. \tag{3}$$

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:

$$\|\bar{\boldsymbol{h}}_{src}^{(\ell)} - \bar{\boldsymbol{h}}_{tgt}^{(\ell)}\| \tag{4}$$

B Detail Experimental Results

B.1 Non-Structural Metrics

Perplexity. Table 3 reports the mean perplexity for each language. These results indicate that Guarani and Jinghpaw represent unfamiliar languages for the models. We also measure token length, which is provided in Table 4.

Cosine similarity for post-trained models. Figure 6 post-training results indicate that post-training has less influence on cosine similarity.

L2 distance. Figures 7 and 8 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

Figure 9 reports TED between Chinese and other languages, yielding similar results with Figure 3.

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

Table 3: Perplexity comparison across models and subsets. Subscripts denote standard deviation. Perplexity is normalized by input length for fair comparison across different input lengths.

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

Table 4: Token length statistics across models and subsets. Subscripts denote standard deviation.

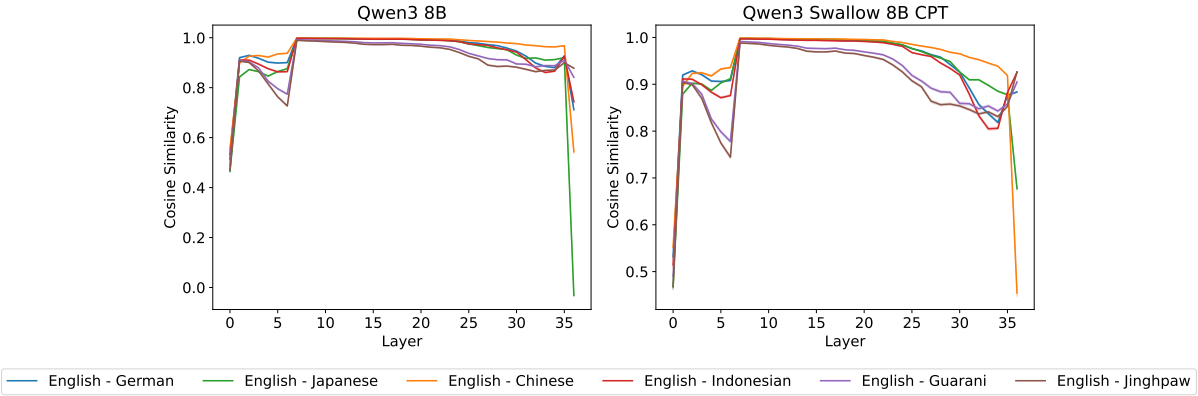

Figure 6: Cosine similarity on post-trained models.

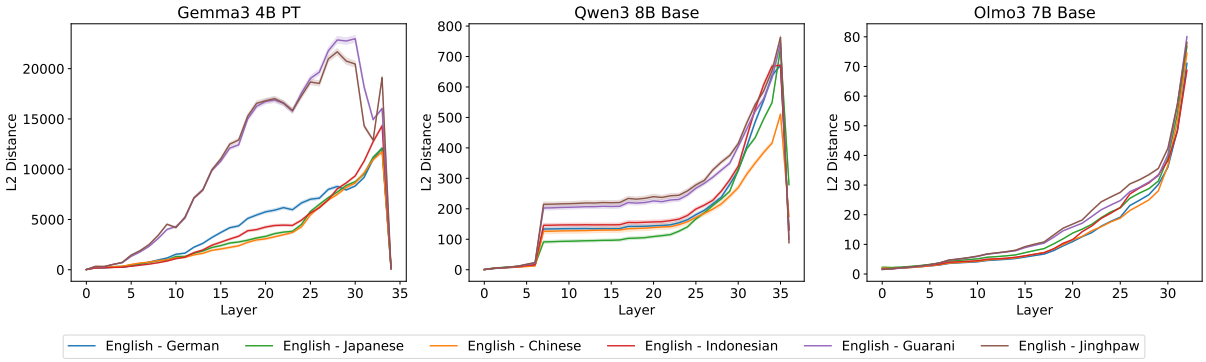

Figure 7: L2 distance on pre-trained models

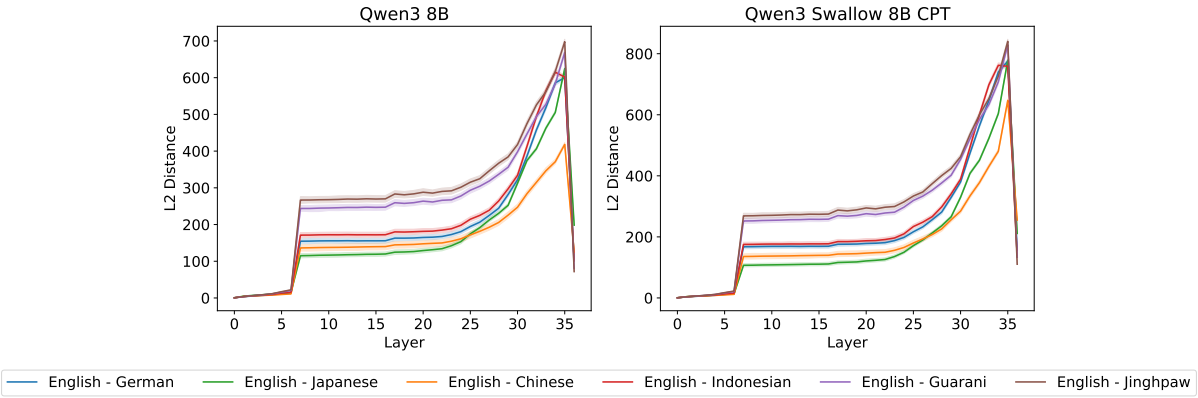

Figure 8: L2 distance on post-trained models

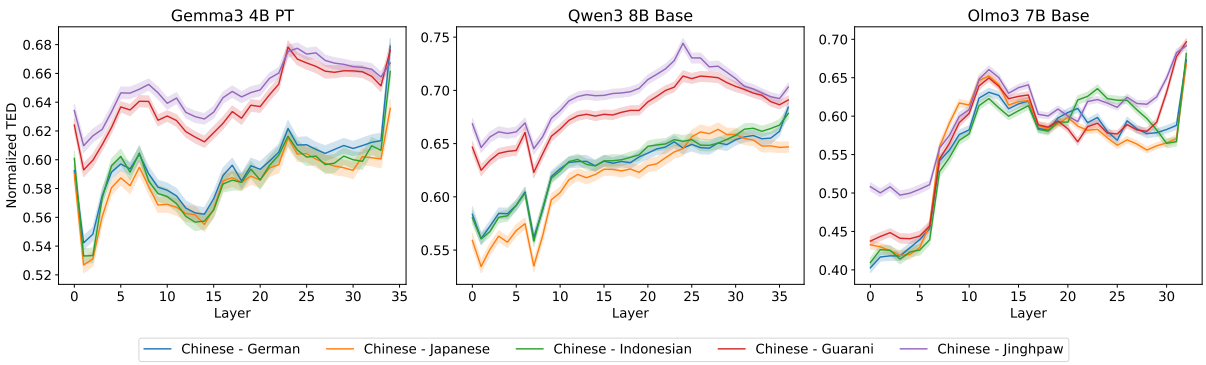

Figure 9: TED between Chinese and other languages. The error bars indicate the 95 % confidence interval.