Language Directions in Multilingual LLMs: A Layer-wise Diagnostic Study of Token Alignment and Pretraining Imprint

Jea Sung Kim
Department of Computer Science
Semyung University
Jecheon-si, Republic of Korea
kjsqp1010@semyung.ac.kr

Suan Lee *
Department of Computer Science
Semyung University
Jecheon-si, Republic of Korea
suanlee@semyung.ac.kr

Abstract

We investigate how multilingual representations emerge across depth in large language models. Using a unified probing framework, we analyze **six** multilingual LLMs across **five** languages (EN/ES/ZH/FR/DE), decomposing behavior into (i) **early-layer dynamics**, (ii) **linear vs. MLP separability**, and (iii) **token–language alignment** that tracks where vocabulary sharing peaks. Across models, we observe a consistent and substantial **early jump**: accuracy rises by **+73.5 to +80.7 points** from L0 to L1 on average, indicating that language-relevant signals become accessible immediately after the embedding layer. Moreover, representations are largely **linearly separable**: for **5/6** models, the mean gap between MLP and linear probes remains within $\pm$**0.5 points**. Token–language alignment further reveals systematic structure, with peak vocabulary mass exceeding **48%** in some models and substantial variation in the depth of peak sharing. These findings provide a compact, cross-model characterization of how multilingual information is organized across depth and introduce simple alignment metrics that complement accuracy-based evaluation.

1 Introduction

Multilingual large language models (LLMs) have become essential infrastructure for global information access, achieving human-level performance across more than 50 languages. Models such as Llama-3.1 (Grattafiori et al., 2024) and Qwen2.5 (Yang et al., 2024) demonstrate strong capabilities not only in English but also in Spanish, Chinese, Arabic, and other languages, suggesting the emergence of a **universal semantic space** shared across languages (Pires et al., 2019; Conneau et al., 2020). However, a persistent gap remains: most models perform best in English, while accuracy in non-English languages is often 10–30% lower (Hu et al., 2020).

*Corresponding author

This gap is commonly attributed to the **English-centric imbalance in pretraining data**. English constitutes 70–80% of typical web-scale corpora (Bender et al., 2021), and limited exposure to other languages leads to performance degradation. Yet a fundamental question remains unanswered: *does the data distribution merely modulate performance, or does it fundamentally shape the geometry of the internal representation space?* If it is merely performance-dependent, post-hoc interventions such as fine-tuning or vocabulary adjustments may suffice; if it is structurally embedded, redesigning the pretraining stage is necessary.

Positioning within prior work. Prior studies have shown that multilingual models learn shared *cross-lingual spaces* (Pires et al., 2019; Chi et al., 2020). Early models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) demonstrated zero-shot transfer across languages, and more recent models such as Llama-3 and Qwen2.5 further strengthened cross-lingual capabilities with larger and more diverse corpora. Meanwhile, probing-based interpretability studies (Alain and Bengio, 2016; Hewitt and Manning, 2019; Belinkov, 2022) examined how linguistic and syntactic information is encoded across layers, but did *not* systematically address (i) whether **linear separability of language information** is universal, (ii) at which depth **language directions emerge**, and (iii) how **pretraining language distributions are imprinted** into the model's internal geometry. Our work fills this gap.

This study investigates two complementary research questions: (1) **Where and how is language information encoded within transformer layers?** (2) **Does the pretraining language distribution structurally reshape the geometry of multilingual representations?**

To answer these questions, we conduct a comprehensive probing study across *all 268 transformer layers* of six representative multilingual LLMs

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 30–35

(Llama-3.1-8B, Qwen2.5-7B, Qwen2.5-Math-7B, OpenMath2-8B, OpenR1-7B, GPT-OSS-20B). We train linear and nonlinear probes and introduce a new analysis, **Token–Language Alignment**, across five typologically diverse languages (EN, ES, ZH, FR, DE), yielding a total of **2,680 independent experiments**.

Key Findings. (1) Universal Linear Separability. Across all models, language information is encoded in linearly accessible subspaces: linear probes achieve 99.8±0.1% accuracy on average, with only a 0.58±0.12%p gap relative to MLP probes ($p<0.001$). Remarkably, this linear structure emerges *immediately at the first transformer block* (Layer 0→1), with a +76.4±8.2%p jump, suggesting that language separation is an architecture-inherent mechanism.

(2) Structural Imprinting. Token–Language Alignment reveals that the alignment between language directions and vocabulary is strongly tied to the pretraining language distribution. English-centric models (EN>80%) show 69.05% alignment for English but only 3.90% for Chinese, whereas Chinese-inclusive models (ZH20%) show a markedly higher 16.43% alignment for Chinese (4.21× increase). This indicates that the *geometry of representation space is structurally shaped by the pretraining distribution*, beyond its effects on task performance.

(3) Typological Effects and Complex Alignment Patterns. Chinese reaches its maximal separability in deeper layers (Layer 5.2±0.8), whereas Spanish and German converge earlier (Layer 2.5±0.4). The overall Match@Peak remains low (15.0±0.3%), showing that language directions capture abstract, multi-factorial features beyond simple token–language matching, including script, lexical frequency, and typological structure.

Implications. These findings suggest that achieving fairness and balance in multilingual LLMs requires **careful design of pretraining data composition**. Structural imprinting implies that post-hoc adjustments cannot fundamentally alter the underlying geometry. Moreover, Match@Peak provides a quantitative diagnostic tool for evaluating the effects of data rebalancing on representation richness.

2 Methodology

We analyze how language information is encoded across all 268 transformer layers of six multilingual LLMs, and how such structure aligns with the pretraining language distribution. Our methodology consists of four parts: (1) experimental setup, (2) linear and nonlinear probing, (3) Token–Language Alignment, and (4) evaluation and statistical analysis.

2.1 Experimental Setup

We study six multilingual LLMs whose pretraining corpora differ in their language composition:

- **English-centric models:** Llama-3.1-8B, OpenMath2-8B
- **Chinese-inclusive models:** Qwen2.5-7B, Qwen2.5-Math-7B, OpenR1-7B
- **Balanced baseline:** GPT-OSS-20B
- **Languages:** Following prior probing studies, we evaluate five XNLI languages (EN, ES, FR, DE, ZH), using 5k training and 2.5k validation sentences per language.

2.2 Probing Methodology

For each model and each layer ℓ, we use the hidden vector of the final token,

$$\mathbf{h}^{(\ell)} \in \mathbb{R}^d,$$

as the sentence representation.

Linear probe. A linear classifier is applied to LayerNorm-normalized representations:

$$f_{\text{lin}}(\mathbf{h}) = W_c \cdot \text{LN}(\mathbf{h}) + b_c,$$

trained with cross-entropy loss to predict one of the five languages.

MLP probe. To compare linear and nonlinear capacity, we additionally train an MLP:

$$f_{\text{mlp}}(\mathbf{h}) = W_2 \cdot \text{ReLU}(W_1 \cdot \text{LN}(\mathbf{h}))\,.$$

LayerNorm is applied in both probes to remove inter-layer scale differences and measure *pure linear separability* of language information.

Both probes are trained independently for each (model, layer, seed, probe type) using AdamW ($\text{lr} = 10^{-3}$), batch size 128, and 3 epochs with early stopping.

2.3 Token–Language Alignment

To quantify how learned language directions relate to the LM head vocabulary, we compute cosine similarity between each probe-learned language direction $\mathbf{w}_L^{(\ell)}$ and each vocabulary embedding $\mathbf{e}_v$:

$$\mathrm{sim}(v, L, \ell) = \cos\Big(\mathbf{e}_v,\, \mathbf{w}_L^{(\ell)}\Big)\,.$$

Each token v is assigned to the language direction with highest similarity.

For each language L, we compute three alignment metrics:

- **PeakDepth**$_L$: The normalized layer index where $\mathrm{VocabShare}(\ell, L)$ is maximized, indicating the depth at which language L is most strongly expressed.
- **PeakVocab**$_L$: The maximum value of $\mathrm{VocabShare}(\ell, L)$ across layers, representing how dominant language L is within the vocabulary space.
- **Match@Peak**$_L$: Among tokens assigned to language L at its peak layer, the percentage whose decoded text belongs to language L, determined using Unicode and diacritic rules (e.g., CJK blocks for ZH, accent-sensitive characters for ES/FR/DE). Higher values indicate stronger alignment between the learned direction and the true lexical identity.

Together, these metrics measure (i) where language information appears in the network, (ii) how strongly it organizes the vocabulary, and (iii) how closely the learned directions correspond to actual linguistic identity—capturing the *structural imprinting* left by pretraining data.

2.4 Evaluation and Statistics

Probe accuracy is evaluated on the validation set. Differences between linear and MLP probes are assessed using layer-wise paired t-tests. We report mean accuracy and 95% confidence intervals across multiple random seeds to ensure robustness and statistical reliability.

3 Experiment Result

3.1 Layer-wise Language Separability

As shown in Table **??**, all transformer layers except the initial embedding layer (Layer 0) achieve consistently high language classification accuracy, exceeding 90% across all models. This indicates that language information is not a transient or layer-specific signal but rather a **global representational property that is preserved throughout model depth**. The sharp increase in accuracy from Layer 0 to Layer 1 (+76.4±8.2%p) suggests the presence of an **early structural reorganization stage** in which language information is rapidly separated within the first transformer block.

Furthermore, the performance gap between linear and MLP probes is less than 1%p on average, demonstrating that language information does not require nonlinear decision boundaries. Instead, it is **almost fully linearly separable** within the latent space. This implies that language is encoded not as a complex nonlinear pattern but as a **global directionality** in a high-dimensional representation space.

These observations point to two structural properties. First, the model appears to **establish a normalized representation of language identity at very early layers** and preserve this signal throughout the stack. Second, the strong linear separability supports the hypothesis that **languages form low-dimensional yet coherent subspaces** in the latent representation space. These findings are consistent with the Token–Language Alignment and structural imprinting analyses presented in the next subsection.

In summary, our layer-wise probing experiments show that **LLM latent spaces preserve language information in a clearly and structurally separable form**, even without additional nonlinear capacity.

3.2 Token–Language Alignment and Structural Imprinting

Token–Language Alignment evaluates how language directions learned by the linear probe relate to LM head token embeddings. Our results show that **LLMs do not distinguish languages solely based on Unicode-defined language identity**. In particular, the alignment between language directions and token identity (Match@Peak) is very low for Latin-script languages such as Spanish, French, and German. This suggests that the learned language directions reflect **latent representational structures shaped during pretraining**, rather than surface-level script information.

In contrast, Chinese (ZH) exhibits substantially higher alignment. Chinese-inclusive models

Table 1: **Two-column summary with per-language averaged accuracy.** Left: model-level statistics averaged over 5 languages. Right: per-language **Avg** accuracy (%) for multilingual comparison.

Model	Early Dyn. (avg)			Separability (avg)			Alignment (avg)			Avg accuracy (per-language)				
	L0	L1	Jump	Lin	MLP	Gap	PDepth	PVocab	M@P	EN	ES	ZH	FR	DE
Llama-3.1-8B	20.0	99.8	79.8	96.2	96.0	−0.2	0.39	33.2	15.0	97.3	96.5	96.0	94.9	96.2
Qwen2.5-7B	17.9	98.6	80.7	95.2	94.8	−0.5	0.28	32.5	14.4	97.4	94.6	94.3	95.7	94.3
Qwen2.5-Math-7B	17.8	95.0	77.1	92.9	93.0	0.1	0.61	48.2	14.3	97.5	95.0	80.8	95.1	96.0
OpenMath2-8B	20.0	99.7	79.7	95.5	95.6	0.0	0.50	34.8	15.3	96.9	95.1	92.9	95.8	97.1
OpenR1-7B	18.5	92.8	74.3	94.0	93.9	−0.1	0.25	48.4	14.8	98.1	96.5	83.9	95.6	95.9
GPT-OSS-20B	19.9	93.4	73.5	85.5	89.9	4.5	0.40	28.1	14.9	96.5	88.1	71.6	84.9	86.2

Table 2: Match@Peak (%) by model group, grouped by pretraining language distribution.

Model Group	EN	ZH	ES	FR	DE
English-centric[a]	69.05	3.90	1.60	0.85	0.40
Chinese-inclusive[b]	54.13	16.43	0.90	0.80	0.30
Difference (Δ%p)	**14.92**	**12.53**	0.70	0.05	0.10
Ratio (×)	**1.28**	**4.21**	1.78	1.06	1.33

[a]Llama-3.1-8B, OpenMath2-8B: models with English-centric pretraining (ZH share $< 5\%$).
[b]Qwen2.5-7B, Qwen2.5-Math-7B, OpenR1-Qwen-7B: models trained on both English and Chinese (ZH-inclusive pretraining).

(with ≈20% ZH data) reach a mean Match@Peak of 16.43%, whereas English-centric models (EN>80%) reach only 3.90%, indicating a 4.21× increase. This demonstrates that the **pretraining language distribution can directly influence the geometry of latent representations**, forming clearer language-specific directions when a language is sufficiently represented in the training corpus.

Taken together, these observations suggest the following structural characteristics of multilingual LLMs:

1. **Rather than relying on surface-level language ID, the model organizes languages via latent directions in the representation space that emerge during pretraining.** These directions correspond to low-dimensional classification boundaries learned by the probe and capture differences between language-specific representations.

2. **Languages with sufficient pretraining coverage (e.g., ZH) form clearer and more stable latent directions**, whereas languages with limited data or shared scripts (ES/FR/DE) tend to share mixed and less distinct subspaces.

Overall, our analysis demonstrates that **LLMs separate languages not by surface token-level features but by latent representational structures shaped by the pretraining distribution**. Moreover, **the composition of pretraining data plays a decisive role in determining the geometry of multilingual representations**.

4 Conclusion

This work presents a quantitative analysis of how language information is structured across all 268 transformer layers of six multilingual LLMs. Our experiments show that language separation emerges immediately in the first transformer block and remains a **stable and strongly linear structure** throughout model depth. The negligible performance gap between linear and MLP probes indicates that the information required for distinguishing languages resides in a **linearly accessible latent subspace**, rather than in complex nonlinear boundaries.

Through our proposed Token–Language Alignment analysis, we further observe that the alignment between language directions and vocabulary embeddings is **strongly influenced by the language composition of the pretraining data**. Chinese (ZH) exhibits much clearer alignment in models that include substantial ZH data, whereas languages with limited coverage or shared scripts (ES-/FR/DE) show weaker alignment. These findings demonstrate that **the pretraining corpus leaves a structural imprint o the geometry of multilingual representations**.

Overall, our results indicate that language representation in LLMs is distinguished not by surface-level token features, but by **latent representational structures formed during pretraining**, which are established early and preserved consistently across layers. This highlights the importance of

balanced language composition and corpus design in achieving fairness and robustness in multilingual LLMs. Moreover, Match@Peak and token–language alignment metrics provide practical tools for diagnosing representation richness and distributional biases.

Future directions include expanding the analysis to additional languages and scripts, evaluating the influence of tokenizer design, and conducting interventional studies to further investigate the causal role of language directions in model behavior.

Limitations

Probing reveals accessibility, not usage. Our analysis relies on supervised probing classifiers, which measure whether language information is *linearly accessible* in a given representation, not whether the model actually *uses* this information during generation. High probe accuracy and strong linear separability are therefore necessary but not sufficient evidence that language directions play a causal role in model behavior. Disentangling encoded-but-unused signals from functionally relevant directions would require interventional methods (e.g., activation patching or direction ablation), which we leave to future work.

Correlational evidence for structural imprinting. Our central claim—that the pretraining language distribution reshapes the geometry of multilingual representations—is supported by a correlation between estimated language composition and Token–Language Alignment, rather than by a controlled manipulation of the training corpus. Because we cannot retrain models under matched conditions, we cannot fully rule out confounds such as architecture, model scale, or instruction/task fine-tuning. The observed $4.21\times$ increase in Chinese alignment for ZH-inclusive models is thus best interpreted as strong associational evidence rather than a causal guarantee.

Acknowledgments

This work was supported by the Ministry of Science and ICT and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2026-25498341).

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*. ICML.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.