Multilingual Learning Strategies in Multilingual Large Language Models

Ali Basirat
Centre for Language Technology
University of Copenhagen
alib@hum.ku.dk

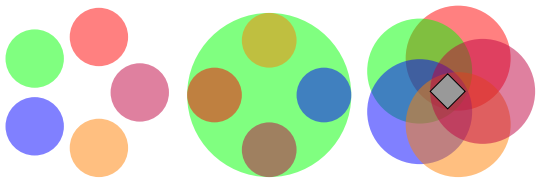
Abstract

Despite the effective performance of multilingual large language models (LLMs), the mechanisms underlying their multilingual capabilities remain unclear. This study examines the intermediate representations of multilingual LLMs to determine if these models utilize human-like second language acquisition strategies: coordinate, sub-coordinate, or compound learning. Our investigations into the discriminative and generative aspects of these models indicate that coordinate learning is the dominant mechanism, with decoder-only models progressively developing distinct feature spaces for each language, while encoder-only models exhibit a mixture of coordinate and compound learning in their middle layers. We find little evidence for sub-coordinate learning. Moreover, the role of training data coverage in shaping multilingual representations is reflected in the fact that languages present in a model's training data consistently exhibit stronger separation than those absent from it.

1 Introduction

Large language models (LLMs) have exhibited impressive performance across multiple languages in a wide range of tasks (Shi et al., 2023). However, the underlying mechanisms that enable their multilingual capabilities remain largely unexplored. Recent studies suggest that these capabilities may stem from a combination of implicit translation into a dominant language like English and internally adopted language-specific processing strategies (Zhang et al., 2023; Wendler et al., 2024).

However, these studies primarily base their hypotheses on the generative capabilities of language models, leaving the explicit exploration of their internal mechanisms unaddressed. We fill this gap by providing a granular perspective on the internal mechanisms underlying multilingualism in multilingual large language models. Specifically, we examine the intermediate representations (activations) of LLMs to identify the presence of multilingual information that supports each of the three types of bilingualism in human language learners: coordinate, sub-coordinate, and compound learning (D'Acierno, 1990). We generalize human bilingualism into multilingualism and conceptualize it in terms of the vector representation of linguistic units formed in the intermediate activations of an LLM. Figure 1 illustrates this conceptualization.



(a) coordinate (b) sub-coordinate (c) compound

Figure 1: A conceptual visualization of feature spaces corresponding to human bilingualism. Each circle represents a feature space for a language. The gray diamond in compound learning refers to a universal space formed by the intersection of all language spaces.

Coordinate learners acquire languages in distinct environments, such as home and school, leading them to process each language independently through separate cognitive systems. In other words, coordinate learners tend to develop language-specific feature spaces, where each language is encoded in its own dedicated representational structure with minimal cross-linguistic influence. From a language model perspective, coordinate learning manifests as distinct language clusters in the intermediate representations.

Sub-coordinate learners, however, interpret languages through the lens of a dominant language by implicitly translating non-dominant languages into the dominant one. This typically occurs in late acquisition, low proficiency, or non-immersive settings, where the learner relies on mental translation rather than direct comprehension. From the perspective of a language model, this translates to the

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 507–518

existence of a broad feature space for the dominant language, which includes other languages.

In contrast, compound human learners develop a core, universal understanding of language, where linguistic units such as word categories and concepts are partially shared across different languages and expressed through varying verbal forms. These learners acquire multiple languages simultaneously within the same environment and tend to abstract away language-specific properties. In a language model's feature space, this translates into the existence of feature spaces shared across all languages.

The training environment of multilingual language models resembles that of coordinate and compound learners, as their training data are sampled from multiple language sources, but each segment primarily consists of a pragmatically complete text (i.e., coherent and self-contained segments, such as articles or conversational exchanges) in a single language, with limited language mixing. Accordingly, we hypothesize that multilingual language models primarily adopt a coordinate learning strategy with some degree of compound learning, while sub-coordinate learning, if present, is likely restricted to unseen languages.

We employ two complementary strategies to investigate this hypothesis based on the intermediate activations of language models. The first is a discriminative approach, quantifying language-specific and universal information in intermediate feature activations. The second examines the models' generation process by analyzing the contribution of intermediate features to token generation.

Our findings across different LLM architectures strongly support the view that multilingual processing in these models aligns primarily with coordinate learning, with partial evidence of compound learning. Decoder-only models such as mGPT (Shliazhko et al., 2024) and BLOOM (Scao et al., 2023) predominantly rely on coordinate learning, whereas encoder-only models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) exhibit a more complex interplay of coordinate and compound strategies. Evidence for sub-coordinate learning is limited, as none of the models show a strong dependence on a dominant language to process others.

2 Previous Work

Zhang et al. (2023) systematically investigate the multilingual capabilities of LLMs across three dimensions: reasoning, knowledge access, and articulation. Their analysis of ChatGPT-generated text shows that LLMs perform better when prompted in English, excel in tasks that allow direct translation, and exhibit a mix of coordinate and sub-coordinate bilingual processing. Our findings strongly support Zhang et al. (2023)'s conclusion that LLMs function as coordinate learners. However, we find clear contradictions with their claim that LLMs also exhibit sub-coordinate bilingualism based on their behavioral analysis of language models. Since their study relies on a different methodology and uses a commercial model (ChatGPT), which does not provide access to internal representations, directly validating their results within our experimental setup remains infeasible.

Wendler et al. (2024) take a different approach to examining the origins of multilingual capabilities in language models primarily trained on English text. They apply the logit lens technique, which projects intermediate representations into the vocabulary space using the model's final token projection layer. Through this method, they argue that a translational shift in intermediate representations is indicative of sub-coordinate learning. However, Belrose et al. (2025) highlight key limitations of the logit lens, showing that it fails to yield meaningful insights for modern language models, including BLOOM (Scao et al., 2023). In particular, they demonstrate that the logit lens often predicts the input token itself as the top output and disproportionately allocates probability mass to tokens that diverge from those emphasized in the model's true output distribution.

When it comes to implications of compound learning, previous studies have suggested the existence of partially shared subspaces between languages in mBERT (Shliazhko et al., 2024). Specifically, Pires et al. (2019) attribute mBERT's cross-lingual capabilities to its language-independent tokenization, while Chi et al. (2020) demonstrate that the model shares portions of its representations across languages, suggesting that compound learning supports cross-lingual transfer through overlapping representational subspaces. Yet, whether these subspaces reflect universal linguistic features or artifacts of training remains an open question that our analysis investigates.

This paper extends previous research by examining multilingualism in open-source LLMs trained on multiple languages and architectures, in contrast to Wendler et al. (2024), which focus on English-

centric models. In addition, we propose novel methods for probing interactions across languages at the level of neural activations, enabling deeper insights into multilingual processing than output-based analyses, a line of inquiry recently criticized for its limitations (Zhao et al., 2025)..

3 Methodology

Let us consider a sentence $s = t_1, \ldots, t_n$ drawn from a language, and define A as an $l \times n \times d$ tensor representing the intermediate activations of a language model as it processes s. Here, l denotes the number of layers, and d represents the number of features, i.e., embedding dimension. In this setup, A provides l distinct representations, each residing in a separate d-dimensional space, for every token. We extend this formulation to multiple aligned sentences across different languages, where each token is annotated with relevant linguistic labels (e.g., language identification or POS tag). This results in a large tensor of size $l \times N \times d$, where N is the total number of tokens across all sentences.

To facilitate efficient visualization, reduce noise, and retain the most informative features of the activation space, we apply principal component analysis (PCA) to each of the l views independently, reducing their dimensionality to $\tilde{d}$ while preserving at least 95% of the activation variance. This results in a tensor $\tilde{H}$ of size $l \times N \times \tilde{d}$, which, together with H, serves as the foundation for our analysis.

We adopt two approaches to examine the generative and discriminative aspects of intermediate representations. The first adopts an information-theoretic procedure to quantify the amount of $\mathcal{V}$-usable information (Xu et al., 2020) encoded in intermediate representations that discriminates between language-specific and universal features. The $\mathcal{V}$-usable information in a random variable X for predicting a category Y is defined as the difference in conditional entropy between predictions based on X and a baseline prediction where no input features are provided (i.e., Φ):

$$I_v(Y; X) = H(Y|\Phi) - H(Y|X)$$

A high value of $I_v(Y; X)$ indicates that X is highly effective at reducing the uncertainty in predicting Y, though this does not necessarily translate to better task performance. In order to make the usable information comparable across tasks, we normalize them by the marginal task entropy and refer to it as the normalized usable information or usable information for short.

$$I_{nv}(Y; X) = 1 - \frac{H(Y|X)}{H(Y|\Phi)} \tag{1}$$

Our motivation for using this metric is twofold. First, its discriminative nature makes it applicable to both encoder-only and decoder-only architectures. Second, it allows for a direct comparison of the effectiveness of feature vectors across different tasks defined over the same feature space X (Ethayarajh et al., 2022). Such a comparison would not be possible if the analysis were based solely on task-specific metrics (e.g., F_1-score and accuracy), as these metrics are not directly comparable across different tasks. Additional details regarding the implementation of this metric are in Appendix A.

The second approach examines the generation capability of the decoder-only models. It utilizes saliency maps to examine how individual intermediate features contribute to token generation (Hou and Castanon, 2023). By examining the gradients of next-token predictions with respect to intermediate activations, we identify the features that play a key role in encoding language-specific and universal properties. We use Gradient-weighted Class Activation Mapping to measure the importance of a feature h_i^k at a layer k to the prediction of a token by computing the product of the feature value for an input token (i.e., $h_i^k(t_j)$) and the gradient of the prediction (before softmax) with respect to that feature (i.e., $\frac{\partial f(t_{j+1})}{\partial h_i^k(t_j)}$). This product undergoes a ReLU activation to ignore negative contributions:

$$c_i^k(t_j) = \text{ReLU}(h_i^k(t_j) \cdot \frac{\partial f(t_{j+1})}{\partial h_i^k(t_j)})$$

where $h_i^k(t_j)$ corresponds to the element (k, j, i) in H, and $f(t_{j+1})$ is the logit for t_{j+1}.

To assess the significance of c_i^k for a group of tokens (e.g., tokens belonging to a particular language), we conduct a two-tailed t-test with a significance level of 0.01. We refer to features with significant contribution to the generation of a particular token group as differentiating features for the group. Accordingly, we define the *differentiating rate* of layer k as the ratio of differentiating features to the total features in the layer:

$$D_k = \frac{\sum_{i=1}^{d} \mathbb{I}\big(p\text{-value}(c_i^k) < 0.01\big)}{d} \tag{2}$$

where $\mathbb{I}$ is an indication function.

In addition to the aforementioned metrics, which are designed to assess coordinate and compound learning, we introduce another approach in Section 8 to assess sub-coordinate learning based on the proximity of intermediate activations to those of a dominant language.

4 Experiment Setup

We leverage the Parallel Universal Dependencies (PUD) treebanks (Zeman et al., 2017; Nivre et al., 2016) which comprise aligned sentences from news sources and Wikipedia, annotated for both morphological and syntactic structures. The cross-lingual alignment of sentences ensures that our findings are not skewed by domain-specific variations or differences in syntactic and semantic structures in certain languages. Additionally, the availability of syntactic annotations allows us to effectively assess compound learning within LLMs.

Our experiments are based on 1000 sentences from each of the 21 topologically different languages in PUD. A summary of the dataset is available in Table 1. The analyses are based on three publicly available multilingual language models with different architectures and language coverages: BLOOM (Scao et al., 2023) and mGPT (Shliazhko et al., 2024) are decoder-only models, and mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020) (base and large) are encoder-only models. More information about the models' size and language coverage is provided in Table 2.

To assess the generalizability of information to unseen languages, we consider two experimental scenarios based on whether a test language is included in a model's pre-training data. The **Seen** setting contains only languages present during pre-training, while the **Unseen** setting includes those absent from it. For mBERT and XLM-R, the Unseen set is empty, as all test languages are covered in their pre-training data.

5 Coordinate Learning

We investigate coordinate learning by analyzing the separability of intermediate representations across input languages through the usable information for language identification and the feature differentiation rate for language processing. The underlying principle is that coordinate learners construct distinct processing systems for each language.

Language	ISO	Family	Size	A	B	C	D
Arabic	ar	Afro-Asiatic	20K	✓	✓	✓	✓
Chinese	zh	Sino-Tibetan	21K	✓	✗	✓	✓
Czech	cs	IE Slavic	18K	✓	✗	✗	✓
English	en	IE Germanic	21K	✓	✓	✓	✓
Finnish	fi	Uralic	15K	✓	✓	✗	✓
French	fr	IE Romance	25K	✓	✓	✓	✓
Galician	gl	IE Romance	25K	✓	✗	✗	✓
German	de	IE Germanic	21K	✓	✓	✗	✓
Hindi	hi	IE Indo-Aryan	23K	✓	✓	✓	✓
Icelandic	is	IE Germanic	18K	✓	✗	✗	✓
Indonesian	id	Austronesian	19K	✓	✓	✓	✓
Italian	it	IE Romance	25K	✓	✓	✗	✓
Japanese	ja	Japonic	28K	✓	✓	✗	✓
Korean	ko	Koreanic	16K	✓	✓	✗	✓
Polish	pl	IE Slavic	18K	✓	✓	✗	✓
Portuguese	pt	IE Romance	24K	✓	✓	✓	✓
Russian	ru	IE Slavic	19K	✓	✓	✗	✓
Spanish	es	IE Romance	23K	✓	✓	✓	✓
Swedish	sv	IE Germanic	19K	✓	✓	✗	✓
Thai	th	Kra-Dai	22K	✓	✓	✗	✓
Turkish	tr	Turkic	17K	✓	✓	✗	✓

Table 1: Selected languages. IE: Indo-European. A: mBERT, B: mGPT, C: BLOOM, D: XMLR.

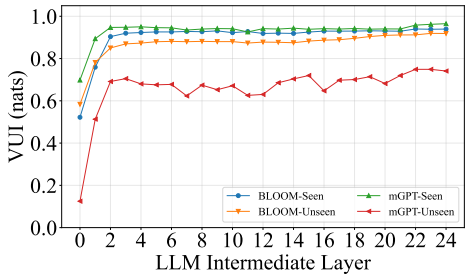
LLM	Size	l	d	LD	LC
BLOOM	1.7B	24	1536	46	17%
mGPT	1.3B	24	2048	61	28%
mBERT	172M	12	768	104	100%
XLMR-base	270M	12	768	100	100%
XLMR-large	550M	24	1024	100	100%

Table 2: Language Models. l and d: number of layers and features LD: Language Diversity – number of training languages; LC: Language Coverage – The ratio of test languages to training languages.

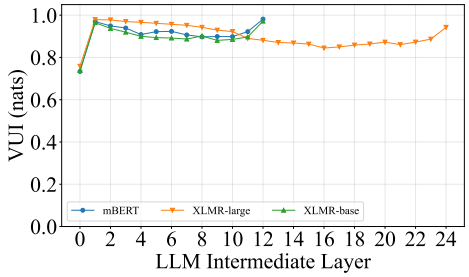
5.1 Usable Information

Figure 2 presents the layer-wise variation in usable information for predicting the source language from activation vectors. The consistent upward trends in the decoder-only models indicate that the activation vectors progressively encode more information about the processing language in deeper layers. The presence of this trend in the Unseen settings suggests that the language-specific information captured by the models generalizes beyond the languages seen during training. However, the overall level of usable information is substantially lower for unseen languages than for seen ones, highlighting the influence of pre-training data coverage on the emergence of coordinate learning.

The encoder-only models, on the other hand, show a different pattern. The decreasing trajectory after the second layer indicates that these models quickly encode language-specific information in their lower layers but gradually lose it until the top



(a) decoder-only



(b) encoder-only

Figure 2: Usable information for language identification.

layers, where reconstruction begins. This pattern holds regardless of the model size, as we see for both the XLMR-base and XLMR-large.

Overall, the results from both architectures support our hypothesis that encoder- and decoder-only models tend to take a coordinate learning, as their primary multilingual learning strategy, which in the case of the decoder-only models develops increasingly through the layers, while being conflated with other learning strategies in the middle layers of encoder-only models.

The progression of coordinate learning is further illustrated by the t-SNE visualization of activation vectors in Figure 3. In both mBERT (encoder-only) and BLOOM (decoder-only), the lower layers show substantial cross-lingual overlap, with limited language separation. In BLOOM, the language overlap diminishes in the upper layers, where language representations become almost entirely separated into distinct feature spaces. The formation of such language-specific feature spaces is also evident for languages not included in the models' pre-training data. Notably, BLOOM tends to develop distinct feature spaces for unseen languages such as German, Finnish, and Swedish. In contrast, mBERT exhibits more substantial cross-lingual overlap in its middle layers, with representations becoming relatively more separable at the second and last layers. Both models show some degree of coordinate learning in their lower layers, limited to typologically distant languages such as Arabic, Czech, Finnish, German, Hindi, Korean, and Russian, occupying separate regions in the feature space.

5.2 Language Differentiating Features

By computing layer differentiation rates in decoder-only models, we identify language-specific features crucial for token generation in each language. The features are identified through their contribution to token prediction in each of the Seen and Unseen settings based on Equation 2. For each language, we estimate feature contributions to next-token prediction and compare them across languages using statistical tests. The proportion of features that differ significantly at each layer defines its differentiation rate, providing a layer-wise measure of language-specific processing. The experiment is detailed in Appendix B and the results are summarized in Figure 4. High differentiation rates indicate distinct feature spaces for each language group, supporting coordinate learning.

The results show that both models tend to dedicate a substantial number of features to differentiate between languages. These features are significantly higher for Seen languages than Unseen ones. The upward trend in mGPT indicates that the model progressively isolates languages into increasingly distinct feature spaces across all layers, regardless of whether the languages were part of its training data. BLOOM, however, follows a different strategy. For Unseen languages, the differentiation rate remains relatively stable around 40-50%, while for Seen languages, it takes a smooth downward trend, implying that BLOOM tends to share some features between languages at the top layers, although it still processes languages through a set of significantly isolated features for each language.

6 Compound Learning

Compound learning involves constructing universal feature spaces shared among languages. Our analysis of compound learning examines the existence of such shared spaces at the syntax level for Universal Part-Of-Speech tags (UPOS). We probe this phenomenon through the usable information for UPOS tagging and the joint differentiation rate of features for languages and syntactic categories.

6.1 Usable Information for UPOS Tagging

Figure 5 illustrates the variation of usable information in the models' intermediate activations for

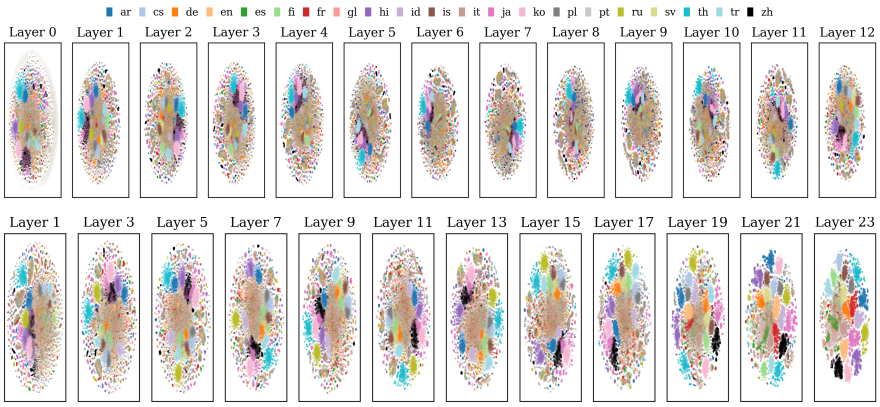


Figure 3: tSNE visualization of activation vectors. Top: mBERT, bottom: BLOOM.

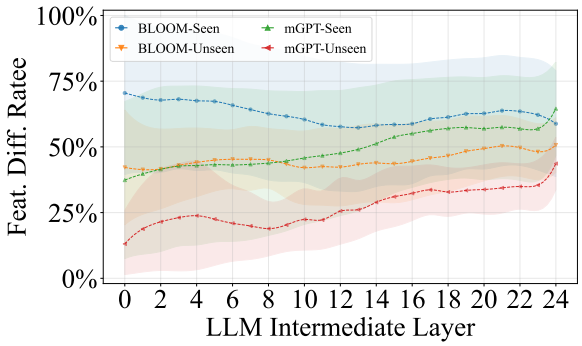


Figure 4: Language differentiation rates across layers. Shaded areas show variation across languages; solid lines show the mean.

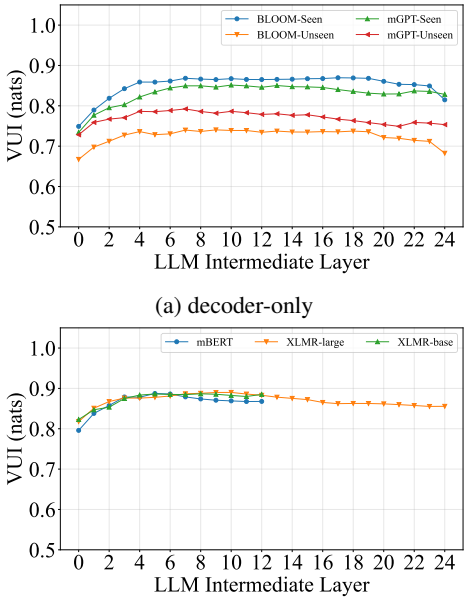
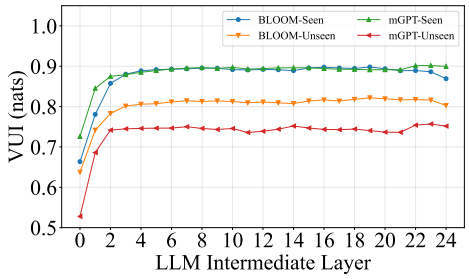


Figure 5: Usable information for UPOS identification.

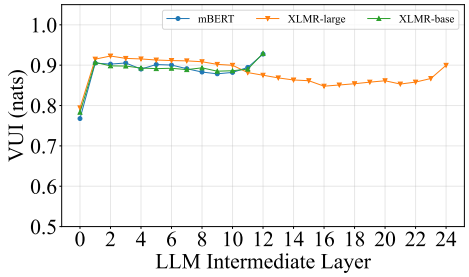
predicting UPOS tags. The results show a consistent pattern across all models: usable information for UPOS prediction is low in early layers, peaks around the middle layers, and declines in the upper layers. This trend holds irrespective of architecture and aligns with prior findings on syntactic localization in transformers (Tenney et al., 2019).

Comparing Seen and Unseen languages reveals that the decoder-only models encode more UPOS information for languages included in their training data. To examine whether models encode universal syntax through shared representations or within language-specific spaces, we measure usable information for the joint prediction of UPOS tags and languages. As shown in Figure 6, decoder-only models exhibit a clear upward trend, indicating that higher layers become increasingly informative for the joint task. This pattern is also observable in tSNE visualization of BLOOM's activation vectors in Figure 7, where the UPOS activations are clustered within the feature space of languages formed at the top layers of the model. This indicates that decoder-only models such as BLOOM represent universal syntax within language-specific feature spaces, reducing the likelihood of compound learning, particularly in the upper layers.

However, the process appears more complex in the encoder-only models. The increasing trends in the initial and top layers support coordinate learning, while the decreasing patterns in the middle layers indicate an additional mechanism, likely linked to compound learning. Still, the fairly high values of the usable information for the joint language and UPOS identification are more in support of coordinate learning, which suggests that the models tend to process universal syntactic properties of the languages within language-specific feature spaces.



(a) decoder-only



(b) encoder-only

Figure 6: Usable information for joint prediction of UPOS tags and languages.

7 Language–UPOS Differentiating Features

To further examine how UPOS tags are processed within language-specific feature spaces, we measure layer differentiation rates based on the prediction of words belonging to a given syntactic category in a target language. By testing whether the same syntactic tag is processed differently across languages, we compute a joint differentiation rate that quantifies the extent to which syntactic categories are represented in language-specific versus shared feature spaces. The details of this experiment are provided in Appendix C.

Figure 8 shows that decoder-only models allocate a subset of features to distinguishing syntactic categories within each language, irrespective of whether the language was included in pretraining. The absolute values of the differentiation rates, however, are consistently higher for Seen languages, suggesting that universal syntactic categories are more strongly encoded in language-specific feature spaces when the language is represented in training. In mGPT, the modest upward trend for Seen languages further indicates that these differentiating features become increasingly effective in the top layers.

8 Sub-coordinate Learning

Sub-coordinate learning implies a shift in intermediate feature vectors towards a dominant language that filters and influences the representations of other languages. The dominant language, which is more represented in the pre-training data of our test language models, is English.

If a language model employs sub-coordinate learning internally, we would expect the representations of non-English languages to be enveloped by or significantly overlap with English representations. To examine this, we measure the proximity of language-specific activation vectors by computing the Kullback-Leibler (KL) divergence between the distribution of each non-English language and English. If language models employ internal filtering mechanisms consistent with sub-coordinate learning, we expect a reduction in KL divergence, indicating that representations of different languages become more aligned with English.

Figure 9 presents the KL divergence between the feature space of each language and English. For decoder-only models, divergence begins relatively small in the lower layers and peaks in the middle layers, reflecting increased separation from English. At the top layers, BLOOM shows a sharp divergence, whereas mGPT instead converges strongly toward English. These trends are consistent across both Seen and Unseen settings: BLOOM's behavior suggests a weakening of sub-coordinate learning, while mGPT's sharp convergence in the top layers provides stronger evidence. Nevertheless, because sub-coordinate learning is expected to manifest primarily in the middle layers, the decrease observed at the top layers of mGPT is less likely to be explained by this mechanism alone.

The encoder-only models display a different pattern. mBERT and XLM-R show only a modest shift toward English, while in XLM-R-large, this turns into a growing divergence after the middle layers. Moreover, the absolute divergence values are substantially smaller than in decoder-only models, peaking at around 80 compared to several thousand, indicating that encoder-only feature spaces are generally denser. The modest reduction in divergence may reflect weak sub-coordinate learning, or, in line with our earlier discussion, could instead result from weak compound learning effects in the middle layers.

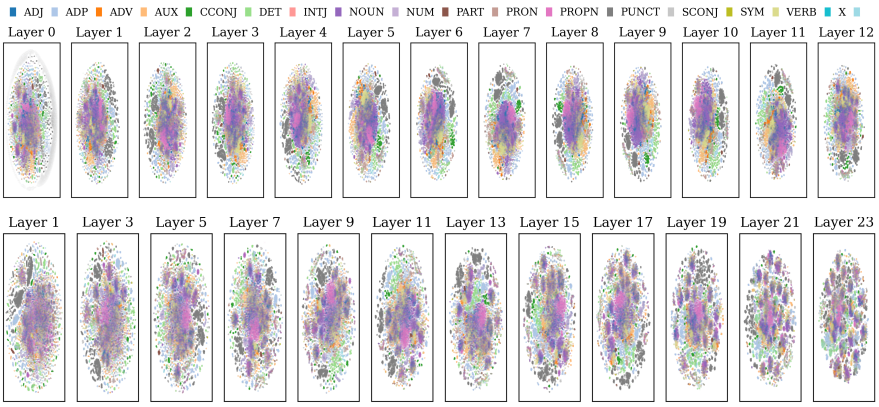


Figure 7: tSNE visualization of activation vectors. Top: mBERT, bottom: BLOOM.

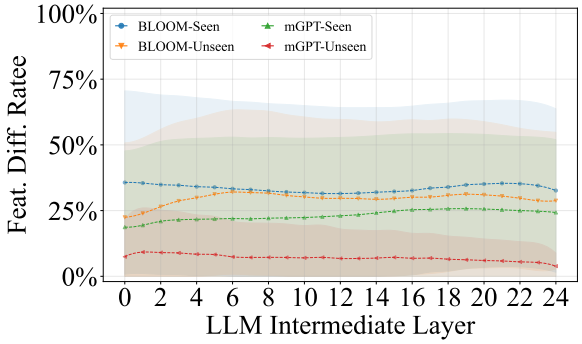


Figure 8: Joint Language-UPOS differentiation rates across layers. Shaded areas show variation across languages; solid lines show the mean.

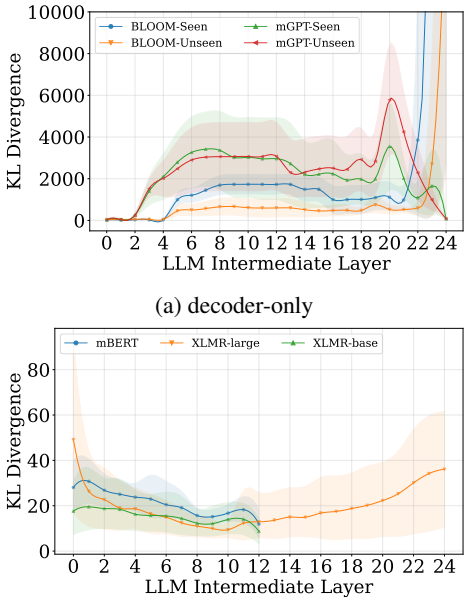


Figure 9: KL Divergence between English and non-English activation vectors. Shaded areas show variation across languages; solid lines show the mean.

9 Conclusion

Our analysis of multilingual large language models reveals differences in how encoder-only and decoder-only architectures handle multilingual representation. We examined the intermediate representations of these models to determine whether they follow coordinate, sub-coordinate, or compound learning strategies.

We show that coordinate learning is the dominant mechanism, with decoder-only models developing strongly separated feature spaces for each language, while encoder-only models exhibit a more complex interplay of coordinate and compound learning in their middle layers. Sub-coordinate learning plays little to no role. Moreover, training data coverage substantially affects the strength of language separation, with Seen languages consistently exhibiting higher usable information and differentiation rates.

Our findings show that both architecture and pretraining data shape multilingual representations in LLMs. Decoder-only models appear better suited for tasks that require maintaining clear language-specific boundaries, while encoder-only models may be more advantageous for cross-lingual transfer, as their denser and partially shared representations facilitate knowledge sharing. More broadly, our results suggest that multilingual generalization in LLMs is not a single mechanism but a balance between language separation and cross-lingual sharing, which emerges differently across architectures and training regimes.

In future work, we will extend the analysis of compound learning to a broader set of cross-linguistic features, including semantic and pragmatic aspects. Additionally, we aim to explore the

impact of training data diversity from a linguistic typology perspective on the balance between coordinate and compound learning, as well as how language models generalize to unseen and low-resource languages. Expanding our study to a wider range of language models will help assess the influence of model scale on multilingual processing strategies.

Limitations

The limitations of this study are as follows: First, our analysis of compound learning primarily focuses on Universal POS (UPOS) tags, which restricts the exploration of higher-level linguistic properties such as syntax, semantics, and pragmatics. Second, we evaluate a limited set of language models, mBERT, XLMR, mGPT, and BLOOM, potentially constraining the generalizability of our findings to larger or differently trained models. Third, the influence of pre-training data availability may introduce biases in our cross-linguistic comparisons, as certain languages are underrepresented. Fourth, while we draw parallels between LLM multilingualism and human language acquisition, our study lacks direct psycholinguistic evaluations to substantiate these comparisons. Finally, our experiments focus on next-token prediction and language identification, leaving other multilingual tasks, such as cross-lingual transfer and code-switching, unexplored.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback on this paper. We are also grateful to Bolette Sandford Pedersen, Costanza Navarretta, Joakim Nivre, and Patrizia Paggio for their insightful comments. Additionally, we acknowledge the Danish e-Infrastructure Consortium (DeiC) for providing computational resources through UCloud, supported under the Linguistic Universals in Language Models project.

References

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2025. Eliciting latent predictions from transformers with the tuned lens. *Preprint*, arXiv:2303.08112.

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.

Maria Rosaria D'Acierno. 1990. Three types of bilingualism. In *The 24th Annual Meeting of the International Association of Teachers of English as a Foreign Language*, Ireland.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Elizabeth M. Hou and Gregory Castanon. 2023. Decoding layer saliency in language transformers. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas

Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. 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.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Chengshuai Zhao, Zhen Tan, Pingchuan Ma, Dawei Li, Bohan Jiang, Yancheng Wang, Yingzhen Yang, and Huan Liu. 2025. Is chain-of-thought reasoning of llms a mirage? a data distribution lens. *Preprint*, arXiv:2508.01191.

A The Implementation of Usable Information

To compute $\mathcal{V}$-usable information for a given function family $\mathcal{V}$, we estimate the conditional entropy terms $H(Y|\Phi)$ and $H(Y|X)$ using a simple classifier to prevent overfitting, following Xu et al. (2020). The classifier is a two-layer perceptron with Layer Normalization applied after each linear layer, a ReLU activation between layers, and a softmax activation at the output. For a given task $X \to Y$, we compute $H(Y|X)$ as the cross-entropy loss of a classifier trained on real samples X and Y, and $H(Y|\Phi)$ is estimated using a separate classifier that predicts Y based only on a zero vector Φ.

In our experiments, Y corresponds to one of the following: UPOS tags, language IDs, or a combination of UPOS tags and language IDs, and X represents a set of hidden activations. Accordingly, for each task and a language model with l layers, we train l classifiers to estimate $H(Y|X)$, along with an additional classifier to compute $H(Y|\Phi)$. The classifiers are trained on the PCA-reduced representations in $\tilde{H}$ for one epoch, using an 80/20% split for training and testing. We employ the Adam optimizer with a learning rate of 0.01 to minimize the cross-entropy loss. The reported $\mathcal{V}$-usable information values in this paper are based on the test split.

B Layer Differentiation Rates for Languages

For language differentiation, we extract the hidden activations and logit gradients for predicting the next token while processing sentences from a target language through a language model. Feature contributions are then estimated by computing the element-wise product of the activation and gradient tensors, followed by a ReLU activation. This results in a contribution tensor of size $l \times n \times d$ for a target language, where l and d are the number of layers and features of the language model, and n is the number of tokens in the language.

To assess differentiation, we repeat this process for all other languages present in the pre-training data of the language model (Seem languages), resulting in a set of contribution tensors. A two-tailed statistical test is applied to compare corresponding elements in the contribution tensors of the target language and each of the other languages. Specifically, we perform the test on the arrays $[i, :, j]$ extracted from each tensor to measure the differentiating rate of feature j at the layer i. This results in a binary tensor of size $l \times d$ for each language pair (i.e., a target language paired by each of the Seen languages), where each element indicates whether the corresponding feature in each layer contributes differently across the two languages.

To identify differentiating features for the target language, we apply a logical AND operation across all binary tensors, producing a final tensor of size $l \times d$. The mean value of this tensor along the second dimension (d) represents the language differentiation rate of each layer.

This procedure is applied to all languages, treating each as the target language in each of the Seen and Unseen settings in turn. By doing so, we obtain a comprehensive measure of how distinctively the model processes each language relative to the others.

C Layer Differentiation Rate for Joint Language and UPOS Tags

For joint language–UPOS differentiation, we extend the procedure described in Appendix B to account for universal syntactic categories within languages. For a given language and UPOS tag, we compile contribution tensors for all tokens assigned to the tag. Each tensor, of size $l \times n \times d$, encodes the contribution of each feature to next-token prediction for words in that language–UPOS category, where l is the number of layers, n the number of tokens, and d the number of features.

We then assess feature-wise differences across languages for each UPOS tag. Specifically, for a feature j in layer i, we perform a two-tailed t-test comparing the contribution arrays $[i, :, j]$ from the target language–UPOS pair with the corresponding arrays from the same UPOS tag in other languages. The target language may be any language under the Seen or Unseen setting, while the comparison is always made against Seen languages.

The resulting binary decisions are aggregated to estimate the proportion of differentiating features at each layer, yielding the *layer differentiation rate* for the joint language–UPOS category. A high differentiation rate indicates that the model processes tokens of a given UPOS tag in distinct, language-specific feature spaces, rather than in a universal syntactic space.