Positional Cognitive Specialization: Where Do LLMs Learn To Comprehend and Speak Your Language?

Luis Frentzen Salim[1, 2], **Lun-Wei Ku**[1], **Hsing-Kuo Kenneth Pao**[2]

[1]Institute of Information Science, Academia Sinica
[2]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
m11315804@mail.ntust.edu.tw, lwku@iis.sinica.edu.tw, pao@mail.ntust.edu.tw

Abstract

Adapting large language models (LLMs) to new languages is an expensive and opaque process. Understanding how language models acquire new languages and multilingual abilities is key to achieve efficient adaptation. Prior work on multilingual interpretability research focuses primarily on how trained models process multilingual instructions, leaving unexplored the mechanisms through which they acquire new languages during training. We investigate these training dynamics on decoder-only transformers through the lens of two functional cognitive specializations: language perception (input comprehension) and production (output generation). Through experiments on low-resource languages, we demonstrate how perceptual and productive specialization emerges in different regions of a language model by running layer ablation sweeps from the model's input and output directions. Based on the observed specialization patterns, we propose **CogSym**, a layer-wise heuristic that enables effective adaptation by exclusively finetuning a few early and late layers. We show that tuning only the 25% outermost layers achieves downstream task performance within 2–3% deviation from the full finetuning baseline. Unlike similar layer-selection methods, the proposed method requires no extra data or computation while retaining comparable performance, which is especially beneficial for low-resource languages. **CogSym** yields consistent performance with adapter methods such as LoRA, showcasing generalization beyond full finetuning. These findings provide insights to better understand how LLMs learn new languages and push toward accessible and inclusive language modeling.

Code and extended version —
https://github.com/luisfrentzen/cognitive-specialization

Introduction

Large language models (LLMs) are central to many modern natural language processing (NLP) tasks, demonstrating remarkable generalization capabilities across diverse tasks and languages (Brown et al. 2020; Xue et al. 2021; Shi et al. 2022). However, most are English-centric, trained on massive English text (Touvron et al. 2023; Zhang et al. 2022). This focus limits the accessibility and benefits of the technology, especially for speakers of low-resource languages. Despite efforts to bridge this gap, proposed methods are often expensive. A key research challenge is thus to understand the underlying architectural mechanisms of multilingualism to develop more efficient and effective adaptation methods.

Prior work on multilingual interpretability has revealed an emerging framework where multilingual processing occurs in three distinct stages: an understanding phase in early layers to process the input language, an intermediate thinking phase in the middle layers that operates on a shared representation (either language-agnostic or closer to the model's dominant language), and a generation phase in later layers for the output language (Wendler et al. 2024; Zhao et al. 2024; Tang et al. 2024; Schut, Gal, and Farquhar 2025). This mirrors how polyglots manage multiple languages, where their internal thoughts can operate in a different language than that used for immediate speaking and hearing. These emergent specializations resemble language centers in the human brain, such as Broca's area for speech production (Broca 1865) and Wernicke's area for language comprehension (Wernicke 1874).

However, since prior research has predominantly focused on observing this multilingual property on trained models, its behavior during the training process and impact on effective adaptation remains underexplored. Understanding its dynamics during training could reveal more efficient adaptation strategies that target only those layers that develop the necessary functionality. Building on these observations, we hypothesize that behind this behavior exist specialized regions that are each responsible for a different cognitive functionality to process multiple languages, akin to the different language centers in the human brain. Intuitively, we can target only those regions that yield the best return for our adaptation goals.

Current language adaptation methods for LLMs typically involve tuning all of the internal layers. However, taking reference from human second language acquisition (SLA), the acquisition of a new language does not necessitate a complete overhaul of the entire cognitive system: indeed, it often involves acquiring only new linguistic interfaces while preserving existing conceptual knowledge (Slobin 1996; Bylund and Athanasopoulos 2024). Accordingly, we posit that efficient language adaptation requires training only these interfaces in the front and rear layers while preserving the weights that process shared representations in the middle

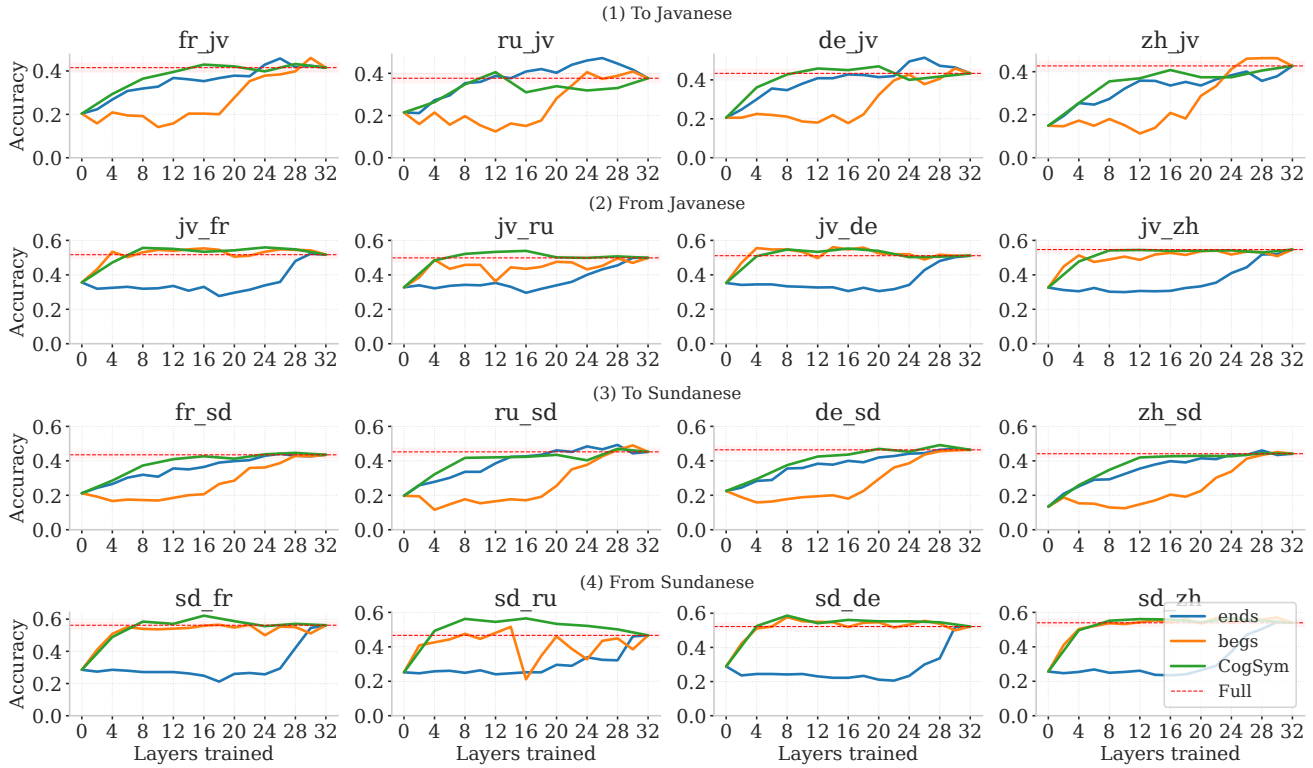

Figure 1: Word-level translation task performance of each ablation sweep, namely front sweep (begs), rear sweep (ends), and both-ends or Cognitive-Symmetric (CogSym) sweep. Plots visualize emerging patterns as each region is expanded.

layers. To test this, we ablate consecutive layer ranges starting from the front and rear ends of the model, sweeping in two-layer intervals. We then probe the model's comprehension and generation abilities by evaluating the trained models on translation tasks between the newly trained language and languages known by the base model. Specifically, we examine tasks that require understanding or comprehension of a target language, and tasks that require generation or production of the target language, e.g., translation from known languages to the target language (generation) and from the target language to known languages (comprehension). We refer to the newly trained or target languages as *unknown* languages and existing languages as *known* languages.

Our primary contributions in this work are threefold:

1. We observe the existence of what we term *perceptual-productive specialization* during the language adaptation process; this is responsible for the different phases of multilingual processing, and is a phenomenon that reflects the language centers of the human brain.
2. We propose a simple yet effective positional heuristic, **CogSym** that is training-method agnostic, proving effective for both full finetuning and parameter efficient finetuning (PEFT) methods like low rank adaptation (LoRA). Moreover, it works without requiring any additional data, which is valuable for low-resource languages. We demonstrate that this heuristic achieves performance comparable to full-model tuning by training as few as the 25% outermost layers in low-resource scenarios.
3. We characterize the key properties of this heuristic, providing novel insight into the effective and resource-efficient adaptation of LLMs for diverse linguistic contexts, especially for low-resource languages.

Related Work

Multilingual Language Models. Modern language models have shown excellent multilingual proficiency. Multilingual models such as mBERT (Devlin et al. 2019), XLM-R (Conneau et al. 2020), mGPT (Lin et al. 2022), PolyLM (Wei et al. 2023), and Aya (Üstün et al. 2024) demonstrate how a single model can effectively represent and process dozens of languages simultaneously. However, they often exhibit a significant performance gap on lower-resource languages due to their skewed pretraining corpora. Training a powerful model from scratch for a low-resource language is also infeasible due to the scarcity of high-quality, large-scale data. To address this, research has focused on various strategies for language adaptation. These approaches typically involve resource-intensive continued pretraining on target-language data or finetuning on specific downstream tasks to transfer the model's existing capabilities to a new linguistic context.

Multilingual Interpretability. A central challenge in the study of LLMs are their black-box nature. The field of interpretability aims to demystify the internal mechanisms of these models to understand how they achieve complex ca-

Figure 2: Downstream task performance of each ablation sweep

pabilities. Interpretability is crucial to developing robust, reliable, and efficient models. Multilingual models are no exception to this.

Recent interpretability work on multilingual LLMs has uncovered a consistent narrative suggesting that multilingual LLMs often follow a three-stage process that maps onto their architecture. For instance, (Zhao et al. 2024) shows how early layers and later layers process mostly non-English tokens whereas middle layers operate mostly on English tokens. Given this observation, they propose a workflow where models first understand an input in its source language, then solve the task using an internal representation heavily influenced by high-resource languages, before generating an output in the target language. (Tang et al. 2024) explore the idea of language-specific neurons, finding that those neurons are responsible for processing the corresponding language and are disproportionately concentrated in the front and rear layers of the model. (Wendler et al. 2024) explore the layers latent representations using the logit lens (nostalgebraist 2020), demonstrating that models like Llama often pivot through a shared representation in their middle layers before generating in the target language.

Layer-Selective Finetuning. A common strategy for efficient adaptation is layer-selective tuning, which trains only a select subset of a model's most important layers to reduce computational costs while maintaining performance. (Lee et al. 2023) introduces surgical finetuning, which selectively trains layers to adapt to different types of distribution shifts during transfer learning on computer vision models. In a similar study on encoder-only language models, (Lodha et al. 2023; Wang et al. 2025) leverages Fisher information matrix (FIM) computed on a small sample of data to rank layers based on their significances. (Zhang, Dong, and Kawaguchi 2024) experiments with the Shapley value to measure LLM layer importance, discovering that their performance relies on a few "cornerstone" layers that severely damage performance when removed. (Qing et al. 2024) introduces AlphaLoRA, a method leveraging the Heavy-Tailed Self-Regularization theory to select the best layer to train based on the model weights.

Methodology

This study presents two low-resource languages: Javanese and Sundanese. To assess the effectiveness of our training, for some tasks we include English, German, Chinese, Russian, and French, all higher-resource languages that are well-represented in the pretraining data of the base models and act as a measure to evaluate against in tasks such as machine translation.

Datasets

Training Data. To facilitate language adaptation, we used a version of the Alpaca instruction dataset presented by (Taori et al. 2023; Upadhayay and Behzadan 2024), which was translated using Google Translate into multiple low-resource languages. This dataset contains 52,000 rows of diverse instructions. The dataset's nature and size provide us with a wide range of vocabulary and text semantics in the form of instructions. Its modest size also allows us to simulate a low-resource adaptation scenario.

Evaluation Data. To evaluate our adaptation results, we defined a comprehensive suite consisting of four different tasks designed to measure training performance across different granularities. Each dataset includes evaluation data for both target languages.

- **Word-level Translation**. We assess vocabulary acquisition using a word-level translation task. The dataset used is adapted from the data presented by (Wendler et al. 2024). The dataset consists of words that map to single tokens in the Llama 2 tokenizer. The Javanese and Sundanese subsets—not included in the original dataset—were translated using Google Translate. In a 4-shot manner, we evaluate the model's ability to translate from the

trained target language to the languages present in this dataset in both directions. Performance is measured by accuracy.

- **Machine Translation**. This task enables us to observe the model's ability to handle full grammatical sentences. We employ the NusaX dataset (Winata et al. 2023), which provides parallel corpora between English and both target languages. We evaluate the model's machine translation ability in both directions, from English to the target language and vice versa. The prompt utilizes a 2-shot setting. We evaluate performance using COMET-22 (Rei et al. 2022) as it is consistently present as an official metric in the latest and past WMT Shared Task (Kocmi et al. 2024).
- **Sentiment Analysis**. We evaluate semantic understanding using the sentiment analysis component of the NusaX dataset, which contains texts annotated with three sentiment labels (positive, neutral, negative), allowing us to assess the model's ability to capture nuanced meaning in the target languages. We evaluate in a 3-shot manner, which includes one shot for each class. We report the macro F1-score.
- **IndoMMLU**. We use IndoMMLU (Koto et al. 2023) to assess general language comprehension. This benchmark contains multiple-choice questions with up to five options from Indonesian educational curricula spanning primary school through university entrance exams. We specifically use the primary school Javanese and Sundanese language subjects, which focus on basic language comprehension, grammar, and vocabulary. Higher grades tend to include culture-specific knowledge that would not be acquired through our instruction-based training. We evaluate in a 2-shot manner. We measure performance using accuracy.

We report the average score over multiple evaluation runs for all tasks.

Experimental Setup

To investigate how the specialized cognitive regions are distributed across the model layers, we conduct a systematic ablation sweep where we train a model while unfreezing only consecutive series of layers, each acting as a "region" given a fixed trainable layer budget k with $k \in \{2, 4, 6, \ldots, L\}$ where L is the total number of layers in a decoder-only transformer model. We train the LM head and the embedding layer as part of the sweep, however, they are not counted towards k. Our primary experiments allocate k in three ways:

- **Front Layer Sweep**: We start with $k = 2$ and unfreeze only k layers nearest to the input (front of the model). Then, we progressively expand by an interval of 2 until we reach the maximum value of k, totaling 16 training variations for the 32-layered model. We also train the embedding layer as part of the front sweep.
- **Rear Layer Sweep**: We start with $k = 2$ and unfreeze only k layers nearest to the output (rear of the model). Likewise, we expand by an interval of 2 until we reach the maximum value of k, totaling 16 training variations. We also train the LM head as part of the rear sweep.
- **Both/Cognitive-Symmetric (CogSym) Sweep**: We start with $k = 4$ and split k equally between the front and rear regions, training the first $k/2$ and last $k/2$ layers. We expand by 4 until the two regions meet in the middle and we reach the maximum value of k, totaling 8 training variations. In this sweep, we train both the LM head and the embedding layer.

This design allows us to study how the position and size of the regions affect model performance on the evaluation set. We also compare to three other ways to allocate k, by leveraging:

- **Language-Specific Neuron (LSN) Count**: We adapt the method proposed by (Tang et al. 2024), which select neurons by ranking its activations over a monolingual corpus. Since LSNs are known to accumulate in certain layers, we rank layers by the number of LSNs they possess. We use the detected neurons of the Llama 2 7B model for the Chinese, French, Indonesian, and Spanish languages provided by the paper authors.
- **Layer-Wise Fisher Information Matrix (FIM)**: We employ the layer selection method proposed by (Lodha et al. 2023), which uses FIM to identify the most informative layers. This method quantifies the impact of parameter changes on a model's prediction by computing the diagonal elements of the FIM and ranking layers accordingly. We directly use the FIM score calculation as defined in their paper:

$$F_\theta = E_{x\sim p(x)}\left[E_{y\sim p_\theta(y|x)}\left[\nabla_\theta \log p_\theta(y|x) \right.\right. \\ \left.\left. \nabla_\theta \log p_\theta(y|x)^T\right]\right]. \quad (1)$$

  To compute our layer-wise FIM, we sample 250 rows from our finetuning dataset.
- **Heavy-Tailed Self-Regularization (HT-SR)**: We use the PL_Alpha_Hill metric proposed by (Qing et al. 2024) to measure layer importance. It is defined by taking the eigenvalues of a layer's weight correlation matrix, fitting a power-law distribution to the heavy-tailed part of the eigenvalue spectrum, and estimating the power-law exponent using the Hill estimator. Layers with higher values correspond to less heavy-tailed distributions which require more tuning.

Training Details

We conducted all experiments using RTX 3090/4090 24GB and A6000/A6000 Ada 48GB GPUs. We used the Unsloth library and conducted all training in bfloat16 precision. All training was optimized with AdamW 8-bit with a batch size of 16, trained for one epoch with a fixed random seed of 42 for reproducibility. All LoRA runs used $r = 128$.

Results and Discussion

In this section we discuss the ablation training results and analyze the emerging patterns. Different selection methods are distinguished by a consistent color across all visualizations. The baseline performance of the fully trained model is denoted by the dotted red line, and the base (untrained) model performance by the dotted green line. We will focus our discussion on the Llama 2 7B model.

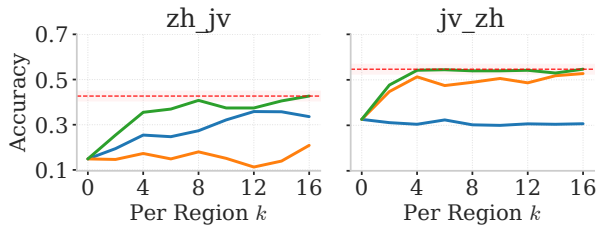

Figure 3: Comparison between each sweep strategy with regard to singular region budget k

Separation of Perceptual-Productive Specialization

We plot the performance of the models trained in Javanese and Sundanese in Fig. 1 and 2. We observe a clear progression in task performance as the size of each region is increased. On translation tasks from known languages to unknown languages (Fig. 1, first row), performance is overwhelmingly dependent on the rear layers. Training only the rear (ends) layers achieves high accuracy with just 10 layers, whereas training the front (begs) layers yields no performance gain for the same budget k. Conversely, on translation from the unknown languages, the initial layers are critical. Front-layer training rapidly approaches the fully trained model's performance (red dotted line), whereas rear-layer training shows almost no improvement until it reaches the front region.

This pattern holds for other downstream tasks. Rear-layer training underperforms on IndoMMLU and sentiment analysis, which typically require understanding of the case at hand rather than complex generation. These results show a clear division between perceptual and productive specialization in the front and rear layers, respectively. Figure 4 reveals how specialization develops across multiple checkpoints during training. The performance gap emerges early in training and continues to widen rapidly through the final checkpoint. This suggests these specializations exist fundamentally, as they form quickly and intuitively within the first few hundred steps.

Although we observe isolated specialization in those regions, we also find that when trained together, the performance of both regions is complementary. As shown in Fig. 3, we find that when plotted with respect to the singular region budget k, training both regions often outperforms training a region alone. Our proposed method, Cognitive-Symmetric Tuning (**CogSym**) leverages these observations by training both regions simultaneously with symmetric allocation of k. With **CogSym** we achieve performance matching the fully trained model using only a small subset of the total 32 layers. This approach consistently matches or outperforms the baseline with only 25% of the layers trained. This result strongly supports our hypothesis that efficient language adaptation can be achieved by training only the linguistic interfaces. We also find that this targeted adaptation does not negatively impact the model's capabilities on other languages, which is one of the main concerns during language adaptation.

Since we observe consistent patterns in both Javanese and Sundanese, we focus our subsequent analysis on Javanese.

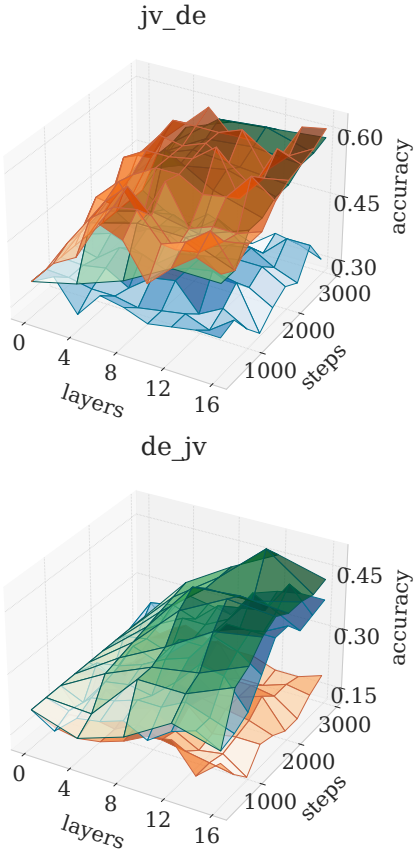

Figure 4: 3D view of word-translation task plot with training steps as an additional axis for German–Javanese pair

Training Layer Position Matters

We now investigate whether this specialization requires distributing k at the model's extreme ends. We fix a layer budget of $k = 8$ (which performed well in our initial experiments) and test four different position configurations, each allocating four layers to the front and four to the rear regions of the model. Figure 5 illustrates our positional ablation study. The leftmost configuration trains the eight outermost layers (0–3 and 28–31), whereas subsequent configurations shift these regions inward by 4-layer increments, eventually meeting at the center (layers 12–19). The red bars below each configuration visualize the trained layer positions.

We observe that performance degrades as trained layers move away from the model's extremities. This is most pronounced on the word translation task illustrated in Fig. 5, the outermost configuration achieves near-baseline performance across all source languages, whereas the centermost configuration scores similarly to an untrained model. This holds across all evaluated tasks, though the magnitude of the performance difference varies. This further supports the perceptual/productive regions discussed in the previous section.

Comparison with Other Layer Selection Methods

We further evaluate **CogSym** with other layer selection methods, more specifically by utilizing FIM, LSN counts,

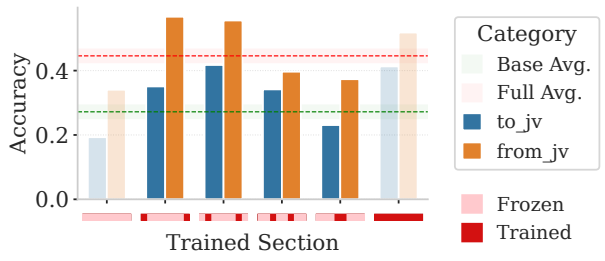

Figure 5: Word translation task performance of 4-position variant with $k = 8$

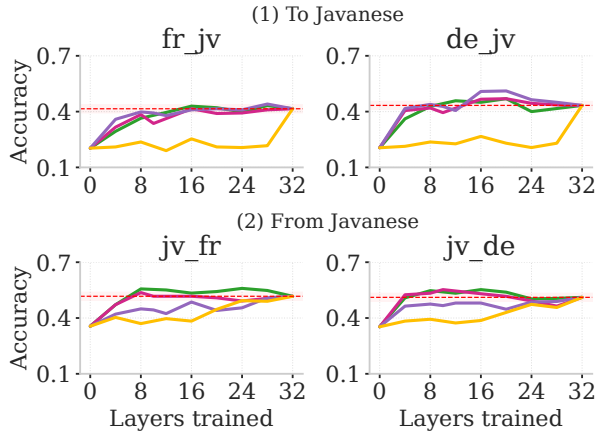

Figure 6: Word-level translation task performance comparison between CogSym (green), FIM (purple), LSN count (magenta), and AlphaLoRA (yellow)

and AlphaLoRA to rank layers. Figure 6 show that **CogSym** (green) and FIM-based selection (magenta) achieve comparable performance. In contrast, LSN-based method and AlphaLoRA performs poorly. This disparity stems from their skewness towards the rear region, whereas **CogSym** and FIM prioritize both front and rear regions more equally. The complete rank distribution of each method is presented in Fig. 7.

CogSym achieves results that are highly competitive with the far more complex FIM method. The convergence of these two philosophically distinct approaches provides strong evidence for the fundamental importance of the model's extremities. Notably, unlike the other methods, our approach does not rely on additional data or computation, which is a valuable trait for adapting low-resource languages.

Training Method Agnosticism

To investigate whether our functional specialization findings are inherent to the model architecture rather than the training method, we replicated our experiments with LoRA. Figure 8 shows **CogSym** performance on LoRA. They demonstrate patterns that are remarkably consistent with those of full finetuning. The same functional specialization emerges: front layers excel at comprehension-focused tasks, rear layers excel at generation-focused tasks, and the symmetric approach effectively captures both capabilities. The performance trajectories follow a nearly identical pattern. This experimental result is evidence of training-method agnosticism: the critical factor is not how weights are updated but where in the architecture the updates occur. This suggests that the perceptive and productive interfaces represent fundamental principles of the multilingual models.

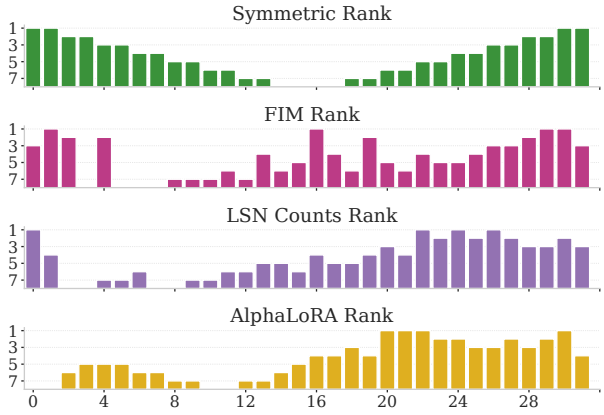

Figure 7: Full distribution of each layer budget allocation method, higher bars indicate greater importance

Sequential Training

We ask whether the two regions can be trained separately rather than simultaneously in a single training run. Since we have two separate regions, training sequentially cuts memory usage by 50%, which helps scenarios where compute memory is limited. We tested two orderings: training the front layers first, then the rear (beg_end), and training the rear layers first, then the front (end_beg). Figure 9 reveals that training sequentially in this case works with minimal performance degradation. But we also observe that recently trained regions exhibit less performance degradation than regions trained earlier. Front-first training yields minimal performance loss on generation-focused tasks such as translation towards the unknown language, whereas rear-first training suffers an almost 10% degradation in performance. Conversely, for comprehension-focused tasks, rear-first ordering performs optimally. Perhaps catastrophic forgetting is a factor during sequential training; we leave a more thorough investigation of such dynamics for future work.

Conclusion

We seek to move beyond black-box approaches to language adaptation and provide a clearer model of the underlying architectural dynamics connecting it to human cognitive concepts. We demonstrate that the front and rear ends of a multilingual language model are specialized to govern distinct cognitive functionalities, which we term *perceptual* and *productive* specialized regions, respectively. Our experiments reveal how the perceptual region governs language comprehension and the productive region handles language genera-

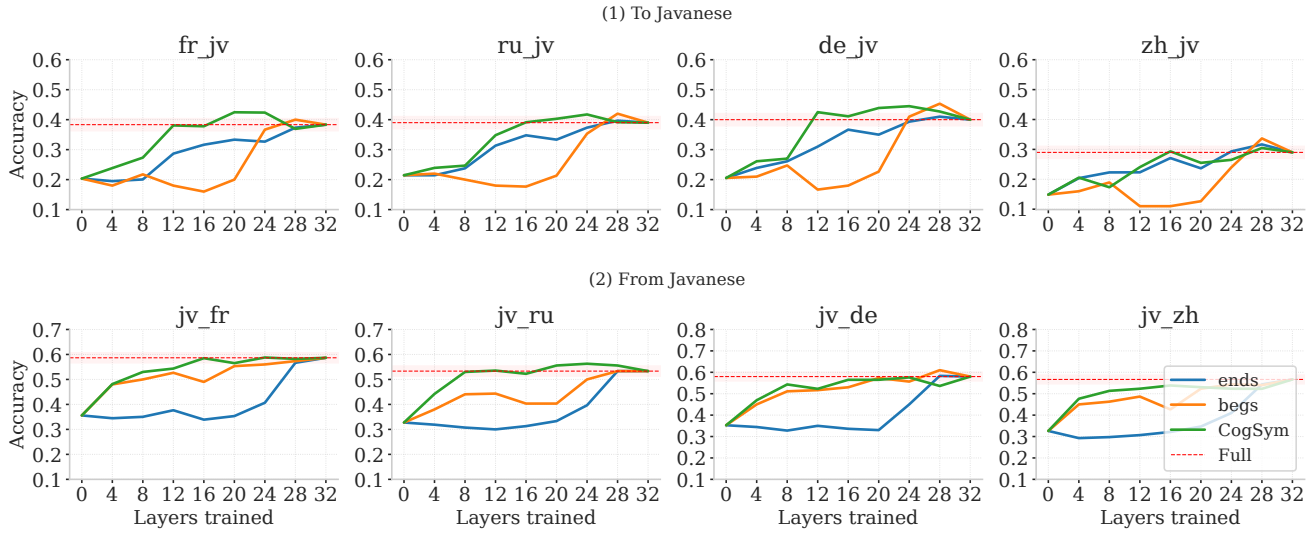

Figure 8: Word-level translation task performance of LoRA, which shows consistent patterns with full finetuning

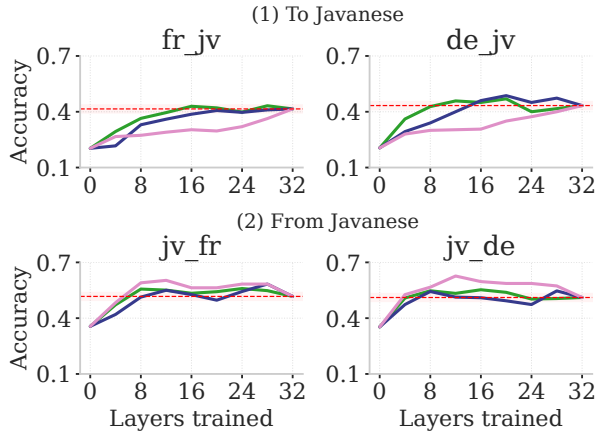

Figure 9: Word-level translation task performance comparison between CogSym (green), front-first training (dark purple), and rear-first training (pink)

tion, similar to the distinct cognitive subareas of the human brain, mirroring their corresponding neurological roles. Our observation demonstrates that training beyond ten layers in these regions often yields no additional performance gains on tasks requiring their specific specialization. However, although seemingly isolated from each other, each region is observed to complement the other when trained together in a way that boosts performance in both specializations. Our proposed method **CogSym** leverages this concept, enabling effective adaptation by training only the outermost layers.

We link **CogSym**'s performance to human SLA. When humans learn new languages, they typically do not need to rewire their core cognitive system; rather, they learn "interfaces" that enable comprehension and communication in the new language. Previous study also found that language-specific thinking patterns acquired through native/first language learning during childhood is likely to prevent adult learners from attaining target-like thinking patterns in the new language, which emphasizes acquiring new linguistic packaging for thoughts instead of new thought patterns themselves (Bylund and Athanasopoulos 2024). These results demonstrate how we can rethink language adaptation by mimicking how human cognitive functions leverage existing linguistic foundations during language acquisition.

Practically, **CogSym** presents potential data and compute resource savings of up to 75% and beyond when combined with PEFT methods while achieving minimal performance loss. This could help increase plasticity, by finding subnetworks to help adapt languages at a lower cost and with fewer data points, since training the smaller network often means we can use fewer data. This approach offers significant advantages for low-resource language adaptation, enabling cost-effective training with limited data by focusing on small, specialized regions. The impact is particularly profound for low-resource language communities, where data scarcity and computational constraints often coexist, pushing for more accessible and inclusive language modeling.

Future Directions

Whereas our study explores two of the most fundamental cognitive functionalities, future studies could explore tasks that involve more complex specializations such as reasoning or math abilities. For instance, the more language-agnostic middle layers (a component not explored in this study) might hold more importance for those types of tasks.

Limitations

Our experiment focuses on decoder-only transformers and two low-resource languages from the Austronesian language family. Further work might be needed to expand the analysis of this phenomenon towards other language models or non-transformer architectures.

References

Broca, P. 1865. Sur le siège de la faculté du langage articulé. *Bulletins et Mémoires de la Société d'Anthropologie de Paris*, 6(1): 377–393.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Bylund, E.; and Athanasopoulos, P. 2024. *Thinking for speaking*. United Kingdom: Routledge. ISBN 9781032535005.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Kocmi, T.; Avramidis, E.; Bawden, R.; Bojar, O.; Dvorkovich, A.; Federmann, C.; Fishel, M.; Freitag, M.; Gowda, T.; Grundkiewicz, R.; Haddow, B.; Karpinska, M.; Koehn, P.; Marie, B.; Monz, C.; Murray, K.; Nagata, M.; Popel, M.; Popović, M.; Shmatova, M.; Steingrímsson, S.; and Zouhar, V. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In Haddow, B.; Kocmi, T.; Koehn, P.; and Monz, C., eds., *Proceedings of the Ninth Conference on Machine Translation*, 1–46. Miami, Florida, USA: Association for Computational Linguistics.

Koto, F.; Aisyah, N.; Li, H.; and Baldwin, T. 2023. Large Language Models Only Pass Primary School Exams in Indonesia: A Comprehensive Test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Singapore: Association for Computational Linguistics.

Lee, Y.; Chen, A. S.; Tajwar, F.; Kumar, A.; Yao, H.; Liang, P.; and Finn, C. 2023. Surgical Fine-Tuning Improves Adaptation to Distribution Shifts. In *The Eleventh International Conference on Learning Representations*.

Lin, X. V.; Mihaylov, T.; Artetxe, M.; Wang, T.; Chen, S.; Simig, D.; Ott, M.; Goyal, N.; Bhosale, S.; Du, J.; Pasunuru, R.; Shleifer, S.; Koura, P. S.; Chaudhary, V.; O'Horo, B.; Wang, J.; Zettlemoyer, L.; Kozareva, Z.; Diab, M.; Stoyanov, V.; and Li, X. 2022. Few-shot Learning with Multilingual Generative Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9019–9052. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Lodha, A.; Belapurkar, G.; Chalkapurkar, S.; Tao, Y.; Ghosh, R.; Basu, S.; Petrov, D.; and Srinivasan, S. 2023. On surgical fine-tuning for language encoders. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 3105–3113. Stroudsburg, PA, USA: Association for Computational Linguistics.

nostalgebraist. 2020. Interpreting GPT: The Logit Lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-08-02.

Qing, P.; Gao, C.; Zhou, Y.; Diao, X.; Yang, Y.; and Vosoughi, S. 2024. AlphaLoRA: Assigning LoRA Experts Based on Layer Training Quality. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20511–20523. Miami, Florida, USA: Association for Computational Linguistics.

Rei, R.; C. de Souza, J. G.; Alves, D.; Zerva, C.; Farinha, A. C.; Glushkova, T.; Lavie, A.; Coheur, L.; and Martins, A. F. T. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In Koehn, P.; Barrault, L.; Bojar, O.; Bougares, F.; Chatterjee, R.; Costa-jussà, M. R.; Federmann, C.; Fishel, M.; Fraser, A.; Freitag, M.; Graham, Y.; Grundkiewicz, R.; Guzman, P.; Haddow, B.; Huck, M.; Jimeno Yepes, A.; Kocmi, T.; Martins, A.; Morishita, M.; Monz, C.; Nagata, M.; Nakazawa, T.; Negri, M.; Névéol, A.; Neves, M.; Popel, M.; Turchi, M.; and Zampieri, M., eds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 578–585. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Schut, L.; Gal, Y.; and Farquhar, S. 2025. Do multilingual LLMs think in English? arXiv:2502.15603.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H. W.; Tay, Y.; Ruder, S.; Zhou, D.; Das, D.; and Wei, J. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. arXiv:2210.03057.

Slobin, D. I. 1996. *From "Thought and Language" to "Thinking for Speaking"*. Cambridge University Press.

Tang, T.; Luo, W.; Huang, H.; Zhang, D.; Wang, X.; Zhao, X.; Wei, F.; and Wen, J.-R. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5701–5715. Stroudsburg, PA, USA: Association for Computational Linguistics.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet,

X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Upadhayay, B.; and Behzadan, V. 2024. TaCo: Enhancing Cross-Lingual Transfer for Low-Resource Languages in LLMs through Translation-Assisted Chain-of-Thought Processes. In *5th Workshop on practical ML for limited/low resource settings, ICLR*.

Wang, X.; Gao, L.; Wang, H.; Zhang, Y.; and Zhao, J. 2025. FLoE: Fisher-Based Layer Selection for Efficient Sparse Adaptation of Low-Rank Experts. arXiv:2506.00495.

Wei, X.; Wei, H.; Lin, H.; Li, T.; Zhang, P.; Ren, X.; Li, M.; Wan, Y.; Cao, Z.; Xie, B.; Hu, T.; Li, S.; Hui, B.; Yu, B.; Liu, D.; Yang, B.; Huang, F.; and Xie, J. 2023. PolyLM: An Open Source Polyglot Large Language Model. arXiv:2307.06018.

Wendler, C.; Veselovsky, V.; Monea, G.; and West, R. 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)*, 15366–15394. Stroudsburg, PA, USA: Association for Computational Linguistics.

Wernicke, C. 1874. *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert.

Winata, G. I.; Aji, A. F.; Cahyawijaya, S.; Mahendra, R.; Koto, F.; Romadhony, A.; Kurniawan, K.; Moeljadi, D.; Prasojo, R. E.; Fung, P.; Baldwin, T.; Lau, J. H.; Sennrich, R.; and Ruder, S. 2023. NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 815–834. Dubrovnik, Croatia: Association for Computational Linguistics.

Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; and Raffel, C. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498. Online: Association for Computational Linguistics.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; Ott, M.; Shleifer, S.; Shuster, K.; Simig, D.; Koura, P. S.; Sridhar, A.; Wang, T.; and Zettlemoyer, L. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068.

Zhang, Y.; Dong, Y.; and Kawaguchi, K. 2024. Investigating Layer Importance in Large Language Models. In Belinkov, Y.; Kim, N.; Jumelet, J.; Mohebbi, H.; Mueller, A.; and Chen, H., eds., *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 469–479. Miami, Florida, US: Association for Computational Linguistics.

Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024. How do large language models handle multilingualism? arXiv:2402.18815.

Üstün, A.; Aryabumi, V.; Yong, Z.-X.; Ko, W.-Y.; D'souza, D.; Onilude, G.; Bhandari, N.; Singh, S.; Ooi, H.-L.; Kayid, A.; Vargus, F.; Blunsom, P.; Longpre, S.; Muennighoff, N.; Fadaee, M.; Kreutzer, J.; and Hooker, S. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. arXiv:2402.07827.