The Emergence of Abstract Thought in Large Language Models Beyond Any Language

Yuxin Chen[1*] **Yiran Zhao**[2*] **Yang Zhang**[3] **An Zhang**[1] **Kenji Kawaguchi**[1] **Shafiq Joty**[2]
Junnan Li[2] **Tat-Seng Chua**[1] **Michael Qizhe Shieh**[1†] **Wenxuan Zhang**[4†]
[1] National University of Singapore [2] Salesforce AI Research [3] Peking University
[4] Singapore University of Technology and Design

Abstract

As large language models (LLMs) continue to advance, their capacity to function effectively across a diverse range of languages has shown marked improvement. Preliminary studies observe that the hidden activations of LLMs often resemble English, even when responding to non-English prompts. This has led to the widespread assumption that LLMs may "think" in English. However, more recent results showing strong multilingual performance, even surpassing English performance on specific tasks in other languages, challenge this view. In this work, we find that LLMs progressively develop a core language-agnostic parameter space—a remarkably small subset of parameters whose deactivation results in significant performance degradation across all languages. This compact yet critical set of parameters underlies the model's ability to generalize beyond individual languages, supporting the emergence of abstract thought that is not tied to any specific linguistic system. Specifically, we identify language-related neurons—those are consistently activated during the processing of particular languages, and categorize them as either shared (active across multiple languages) or exclusive (specific to one). As LLMs undergo continued development over time, we observe a marked increase in both the proportion and functional importance of shared neurons, while exclusive neurons progressively diminish in influence. These shared neurons constitute the backbone of the core language-agnostic parameter space, supporting the emergence of abstract thought. Motivated by these insights, we propose neuron-specific training strategies tailored to LLMs' language-agnostic levels at different development stages. Experiments across diverse LLM families support our approach.[1]

1 Introduction

As large language models (LLMs) continue to advance (OpenAI, 2023; Touvron et al., 2023a; Hurst et al., 2024; Yang et al., 2024a; Team et al., 2024), their performance across a wide range of languages (known as multilingual capability) has markedly improved over the past years (Le Scao et al., 2023; Yang et al., 2024b; Üstün et al., 2024). Despite this progress, several studies have observed that LLMs tend to "think in English", often using it as an internal language of thought even when processing inputs in other languages (Wendler et al., 2024; Zhao et al., 2024b; Schut et al., 2025a). This phenomenon has led to the hypothesis that LLM performance in non-English languages is inherently constrained by their capabilities in English (Qin et al., 2023; Liu et al., 2024).

[*] Equal Contribution.
[†] Corresponding Authors.
[1] Our codes are available at `https://github.com/chenyuxin1999/Abstract_Thought`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Yet, more recent findings complicate this narrative: some studies found that LLMs can actually outperform their English-language performance on certain tasks in other languages (Zhao et al., 2025c; Gemma Team et al., 2024b, 2025), indicating that non-English processing may not always rely on English as an intermediate language. These conflicting observations raise a deeper research question: *Do LLMs think in the distinct space of each language*, or *Do they operate in a higher-level language-agnostic space beyond any specific language?* In other words, whether the trend of non-English performance compared to English indicates **the emergence of abstract thought within LLMs?**

In this work, we explore the existence and development of abstract thought in LLMs by analyzing how individual neurons, responsible for models' thinking, respond to multilingual queries. Here each neuron corresponds to a row or column in the model's parameter matrices and is considered activated if its removal significantly alters the model's output (Frankle and Carbin, 2018; Tang et al., 2024; Wang et al., 2025a). We begin by identifying neurons activated when the model processes inputs in specific languages, referred to as *Language-Related Neurons*. To investigate whether these neurons become increasingly specialized for specific languages or potentially exhibit more general and language-agnostic functionality, we distinguish between *Language-Exclusive Neurons*, which are activated only for one language, and *Language-Shared Neurons*, which are consistently activated across all languages considered. Figure 1 (top) shows a positive correlation between multilingual ability and the proportion of language-shared neurons across different generations of LLMs. This suggests that **as multilingual performance improves, a greater proportion of language-related neurons are shared across languages.**

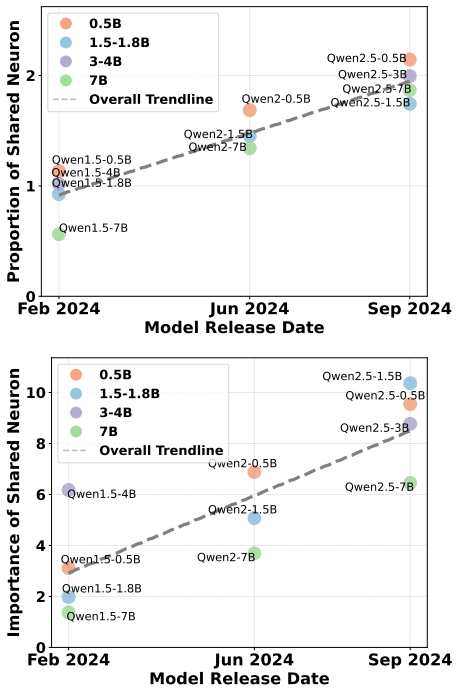

Figure 1: **(Top)** The trendline of shared neuron proportion rises with model release date (see Section 2.2). **(Bottom)** The trendline of shared neuron importance also grows, indicating their increasing language-agnostic property (see Section 2.3).

Building on the observed positive relationship between language-shared neurons and multilingual capability, as well as the finding that LLMs increasingly outperform in non-English languages on certain tasks, we hypothesize that shared neurons may gradually assume more fundamental roles beyond merely supporting multilingual processing. Accordingly, rather than focusing solely on how the proportion of language-shared neurons evolves across model generations, it is essential to evaluate their functional significance relative to language-exclusive neurons. If shared neurons contribute more critically to multilingual processing than language-exclusive neurons—which also participate in language tasks and should be comparably important in principle—this would indicate that shared neurons have evolved into *Language-Agnostic Neurons*, which go beyond shared activation patterns to support abstract functions like semantic reasoning and generalization. As these neurons evolve, they support increasingly abstract thought that transcends linguistic boundaries. As shown in Figure 1 (bottom), language-shared neurons exhibit a markedly growing importance in multilingual processing relative to language-exclusive neurons, signaling **the emergence of language-agnostic properties and potentially, the development of abstract thought in LLMs.**

Inspired by the insights discussed above, we propose a set of targeted neuron training strategies aimed at enhancing the multilingual capabilities of LLMs. These methods are tailored based on the presence or absence of language-agnostic neurons, which serve as an indicator of the emergence of abstract thought within the model. For LLMs that lack language-agnostic neurons, the model is likely still under-trained; thus, training any language-related neurons can contribute to improving multilingual performance. In contrast, in LLMs where abstract thought has emerged, language-shared neurons have evolved into language-agnostic ones. As these neurons have reached a form of generalization,

further improvement through additional training is limited. In such cases, enhancing multilingual capabilities requires focusing on training language-exclusive neurons to better support language-specific nuances. We validate our approach through comprehensive experiments across diverse model series and release time. The results demonstrate that our training method, guided by the presence of language-agnostic properties, effectively enhances multilingual performance.

2 Metrics for Exploring Abstract Thought

In this section, we identify neurons associated with language processing, referred to as *Language-Related Neurons*, and, based on them, define several metrics to quantify and analyze the emergence of abstract thought in LLMs.

2.1 Language-Related Neurons

We identify language-related neurons as those that are consistently activated when processing inputs in a particular language, where a neuron is defined as a single row or column within the model's parameter matrices. Building on prior work in identifying important neurons in neural networks (Frankle and Carbin, 2018; Ni et al., 2023; Tang et al., 2024; Zhao et al., 2024b), we consider a neuron to be activated if its removal leads to a significant change in the resulting embedding. Formally, given an input sequence x in a specific language, a neuron $\mathcal{N}$ is considered activated if

$$\|\mathcal{LLM}(x) - \mathcal{LLM}_{\ominus\mathcal{N}}(x)\|_2 \geq \sigma, \tag{1}$$

where $\mathcal{LLM}(x)$ denotes the output embedding when processing x, and $\mathcal{LLM}_{\ominus\mathcal{N}}(x)$ denotes the output when neuron $\mathcal{N}$ is deactivated, i.e., its parameters are set to zero. The threshold σ specifies the minimum magnitude of change required to consider a neuron activated.

Furthermore, language-related neurons $\mathcal{N}^{\ell}_{\text{lang}}$ for a specific language ℓ are identified through

$$\mathcal{N}^{\ell}_{\text{lang}} := \Big\{ \mathcal{N} \in \mathcal{LLM} \;\Big|\; \|\mathcal{LLM}(x) - \mathcal{LLM}_{\ominus\mathcal{N}}(x)\|_2 \geq \sigma,\ \forall x \in \ell \Big\}. \tag{2}$$

Since sequentially deactivating neurons in Equation 2 is computationally expensive, we employ the parallel neuron detection methods proposed in Zhao et al. (2024b); Wang et al. (2025a). Further implementation details are provided in Appendix A. Details of how the activation threshold σ is selected and validated are provided in Appendix B.

2.2 Language-Shared and Language-Exclusive Neurons

To investigate whether neurons become increasingly specialized for specific languages or exhibit language-agnostic behavior, we conduct a preliminary analysis of the proportion of *Language-Shared Neurons*, defined as language-related neurons that are consistently activated across all languages considered, and *Language-Exclusive Neurons*, defined as language-related neurons that are uniquely activated for individual languages and not shared across all languages. Formally, language-shared and language-exclusive neurons are defined as follows:

$$\mathcal{N}_{\text{shared}} := \bigcap_{\ell \in \mathcal{L}} \mathcal{N}^{\ell}_{\text{lang}}, \quad \text{and} \quad \mathcal{N}^{\ell}_{\text{exclusive}} := \mathcal{N}^{\ell}_{\text{lang}} \setminus \mathcal{N}_{\text{shared}}, \tag{3}$$

where $\mathcal{L}$ denotes the set of all languages under consideration. In other word, $\mathcal{N}_{\text{shared}}$ consistently exhibit high importance across inputs from different languages, while $\mathcal{N}^{\ell}_{\text{exclusive}}$ is the set of neurons specific to that language but not part of the shared set. Furthermore, we examine the proportion of language-shared neurons relative to language-exclusive neurons, defined as

$$\text{Shared Neuron Ratio} := \frac{|\mathcal{N}_{\text{shared}}|}{\frac{1}{|\mathcal{L}|}\sum_{\ell \in \mathcal{L}} |\mathcal{N}^{\ell}_{\text{exclusive}}|}, \tag{4}$$

which quantifies the extent to which individual neurons are shared across all languages as opposed to being specialized for specific ones. A higher ratio indicates a greater number of neurons that are commonly activated across languages, while a lower ratio suggests that more neurons are uniquely responsive to individual languages. A more detailed analysis of the shared and exclusive neurons, including their layer-wise and component-level distributions, can be found in Appendix C.

2.3 Language-Agnostic Neurons

As LLMs continue to improve in their ability to handle multiple languages, and even outperform their English capabilities on specific tasks, we hypothesize that shared neurons may gradually serve more fundamental functions beyond the processing of multiple languages. Consequently, rather than solely examining the evolution of the proportion of language-shared neurons across successive model generations, it is also crucial to assess their relative functional significance in comparison to language-exclusive neurons. Specifically, if language-shared neurons contribute significantly more to multilingual processing than language-exclusive neurons, this indicates a functional difference between the two, since both types are involved in language-related tasks and identified using the same criteria, they should be equally important if their roles are analogous. The discrepancy suggests that shared neurons have evolved into *Language-Agnostic Neurons*. Note that while language-shared neurons are activated across multiple languages, language-agnostic neurons reflect a higher level of abstraction. Rather than encoding language-specific features, they are hypothesized to support cognitive functions that transcend individual languages, such as semantic abstraction, reasoning, and generalization.

To investigate whether language-agnostic property emerge in language-shared neurons, we introduce the metric *Language-Shared Neuron Importance*, which quantifies the impact of deactivating language-shared neurons versus language-exclusive neurons on the model's performance in a given language. This is operationalized by measuring the change in perplexity ($\Delta\mathrm{PPL}$) when each neuron group is ablated. A disproportionately larger increase in perplexity upon deactivating shared neurons would suggest their greater functional importance. Formally, we define the language-shared neuron importance for a language ℓ as:

$$\mathrm{Imp}^{\ell} := \frac{\Delta\mathrm{PPL}^{\ell}_{\mathrm{shared}}/|\mathcal{N}_{\mathrm{shared}}|}{\Delta\mathrm{PPL}^{\ell}_{\mathrm{exclusive}}/|\mathcal{N}^{\ell}_{\mathrm{exclusive}}|}, \tag{5}$$

where $\Delta\mathrm{PPL}^{\ell}_{\mathrm{shared}}$ and $\Delta\mathrm{PPL}^{\ell}_{\mathrm{exclusive}}$ denote the changes in perplexity for language ℓ when shared and corresponding exclusive neurons are deactivated, respectively, and $|\mathcal{N}_{\mathrm{shared}}|$ and $|\mathcal{N}^{\ell}_{\mathrm{exclusive}}|$ represent the number of neurons in each group. A higher value of Imp^{ℓ} indicates that shared neurons contribute more significantly than language-exclusive neurons, thereby providing evidence for their language-agnostic role, since both types of neurons should exhibit comparable importance if their functions were equivalent.

To obtain an overall model-level estimation reflecting this trend across different languages, we compute the average importance across all languages and apply a logarithmic transformation to mitigate scale sensitivity. We refer to the resulting quantity as the *Language Agnostic Score*:

$$\text{Language Agnostic Score} := \log\left(1 + \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}}\mathrm{Imp}^{\ell}\right), \tag{6}$$

which quantifies the average degree to which language-shared neurons contribute in a language-agnostic manner across the evaluated languages. In contrast to the shared neuron ratio defined in Equation 4, which solely quantifies the number of shared neurons, the language-agnostic score incorporates the functional importance of neurons. Higher values suggest not only stronger language-agnostic behavior but also hint at the emergence of abstract thought in LLMs.

3 Emergence of Abstract Thought

3.1 Experiment Setup

Evaluated Models To comprehensively evaluate the emergence of abstract thought in LLMs throughout their development, we examine 20 open-source models encompassing diverse model families, release periods, and sizes. Specifically, we evaluate Llama series including LLaMA1-7B (Touvron et al., 2023a), Llama2-7B (Touvron et al., 2023b), Llama3.2-1B, Llama3.2-3B, Llama3-8B, Llama3.1-8B (Grattafiori et al., 2024), Qwen1.5-0.5B, Qwen1.5-1.8B, Qwen1.5-4B, Qwen1.5-7B (Bai et al., 2023), Qwen2-0.5B, Qwen2-1.5B, Qwen2-7B (Yang et al., 2024a), Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B (Yang et al., 2024b), Gemma-7B (Gemma Team et al., 2024a), Gemma2-9B (Gemma Team et al., 2024b), Gemma3-4B (Kamath et al., 2025).

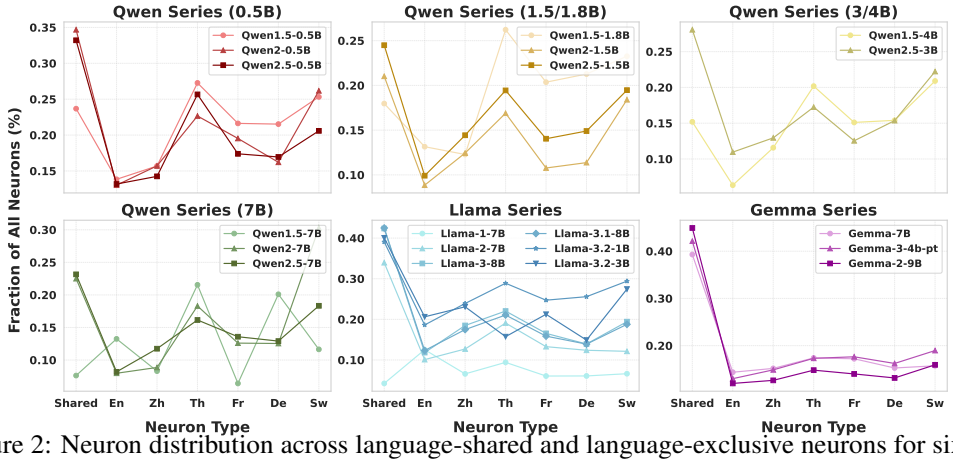

Figure 2: Neuron distribution across language-shared and language-exclusive neurons for six languages (En, Zh, Th, Fr, Dr, Sw) in various model series and scales. For each model, we present the fraction of shared neurons and exclusive neurons to the total number of neurons.

Multilingual Benchmark We evaluate models across six typologically and resource-diverse languages: Chinese (Zh), English (En), Thai (Th), Swahili (Sw), French (Fr), and German (De). This selection spans high-resource, medium-resource, and low-resource languages, enabling a representative analysis of language-related neuron behaviors. For our analysis, we utilize the Multilingual Massive Multitask Language Understanding (MMMLU) dataset (OpenAI, 2024), a human-translated extension of the original MMLU benchmark (Hendrycks et al., 2021), available in 14 languages. In addition, we incorporate the Multilingual Grade School Math (MGSM) dataset (Shi et al., 2022), a translated version of GSM8K (Cobbe et al., 2021), which covers 10 languages. Together, these datasets provide quantitative measures of the models' multilingual capabilities.

Neuron Detection Corpus For each language, we identify language-related neurons by analyzing activation patterns on 1000 sentences sampled from the OSCAR corpus (Abadji et al., 2022). To quantify the functional contribution of these neurons, we further compute perplexity changes caused by deactivating them, using the same language-specific OSCAR data. This unified framework allows us to assess both the proportion and the importance of language-specific and shared neurons across languages and model generations. More detailed illustration can be found in Appendix D.

3.2 Analysis on Shared Neuron Ratio

Language-related neurons account for only a small proportion in LLMs. To develop a preliminary understanding of language-shared and language-exclusive neurons, we begin by analyzing the distribution of shared and language-exclusive parameters across all neurons within the model. For each language, we compute the proportion of language-shared and language-exclusive neurons relative to the total number of neurons in the model. Specifically, we calculate the ratios $\mathcal{N}_{\text{shared}}/|\mathcal{LLM}|$ and $\mathcal{N}^{\ell}_{\text{exclusive}}/|\mathcal{LLM}|$, where ℓ denotes a specific language. The results, illustrated in Figure 2, encompass six languages across multiple model series. It shows that only a small fraction of neurons, often fewer than 1%, play a critical role in processing language, underscoring the sparsity and selectivity of language-relevant neural activations. Furthermore, the quantities of language-shared and language-exclusive neurons are of similar magnitude, each coarsely estimated at around 0.3% of the total number of neurons in the LLM.

To further explore the evolution of shared and exclusive neurons across models and over time, we compute the overall shared neuron ratio for each model, as defined in Equation 4, relate it to multilingual performance measured by MMMLU and MGSM, and present the results in Figure 3.

The proportion of shared neurons increases with model evolution. We first group models from the same series and with similar parameter scales, as indicated by the shaded color regions in Figure 3. Within each group (e.g., Qwen1.5-7B, Qwen2-7B, and Qwen2.5-7B), we observe a steady and consistent increase in the shared-to-exclusive neuron ratio across generations. This growth closely parallels improvements in the model's multilingual ability, with an average Pearson correlation coefficient of $R = 0.92$ and a Spearman rank correlation of $\rho = 0.88$, indicating a strong and reliable

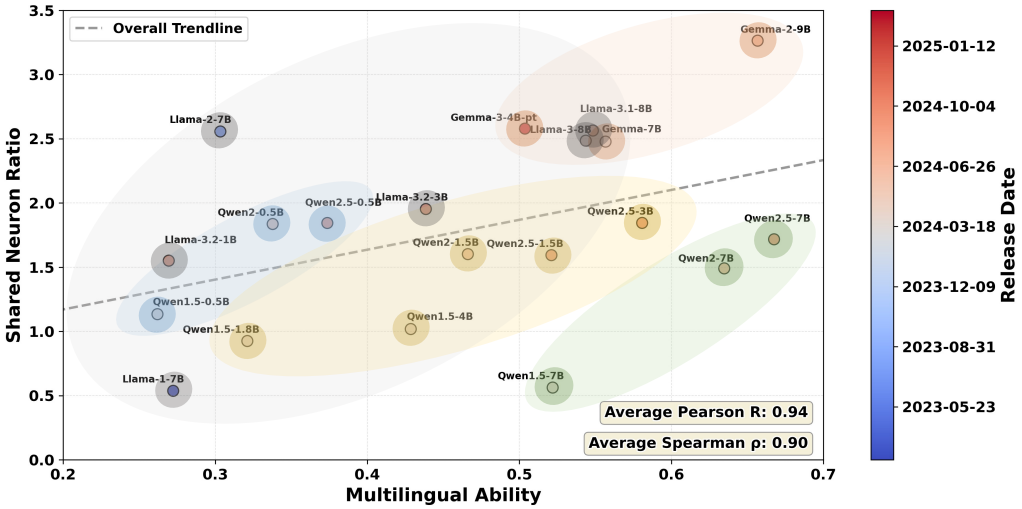

Figure 3: The relationship between multilingual ability and shared neuron ratio (as defined in Equation 4) across various models. Each point represents a model, color-coded by its release date. Shaded regions indicate groups of models within the same series and of comparable scale. The gray dashed line (- - - -) illustrates the overall trend: as models evolve, those with greater multilingual capabilities tend to exhibit a higher proportion of shared neurons.

relationship. In other words, later generations within the same series show a strong trend toward engaging more shared neurons for processing different languages.

The increase of shared neuron proportion generalizes across model families. Beyond individual model series, we observe that the positive relationship between the proportion of shared neurons and multilingual capability generally persists across different model families, as illustrated by the gray dashed line (- - - -) in Figure 3. Despite differences in architecture design and pretraining corpora, models with stronger multilingual ability tend to activate a larger proportion of shared neurons. For instance, the Gemma series exhibits both the most strong multilingual performance and the highest shared-to-exclusive neuron ratio. This consistency across diverse architectures suggests that progressively leveraging shared neurons may be a general strategy adopted by multilingual LLMs, regardless of their specific design choices.

3.3 Analysis on Language Agnostic Score

The above observations raise a further question: whether the shared neurons not only occupy a larger proportion of the language-related neuron set, but also *contribute more critically to multilingual processing*, effectively functioning as language-agnostic neurons. To address this question, we investigate how the language-shared neurons importance, i.e., language agnostic score defined in Equation 6, evolves alongside multilingual capability across different generations of large language models.

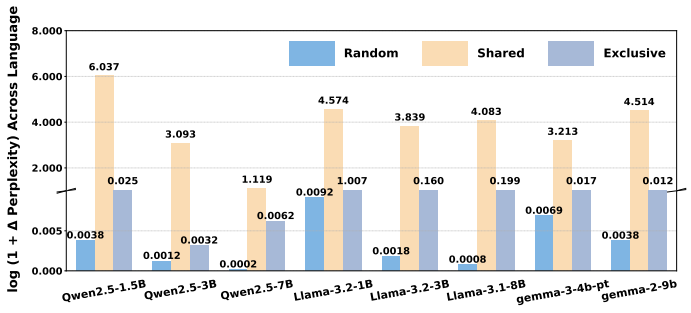

Figure 4: Perplexity changes caused by deactivating random neuron sets (Random), language-shared neurons (Shared) and language-exclusive neurons (Exclusive). Notice that Random deactivation barely affects models' perplexity, while Shared and Exclusive deactivation break the models' abilities.

Deactivating shared and exclusive neurons both leads to model degradation. Before contrasting language-shared neurons with language-exclusive neurons, we conduct a control experiment in

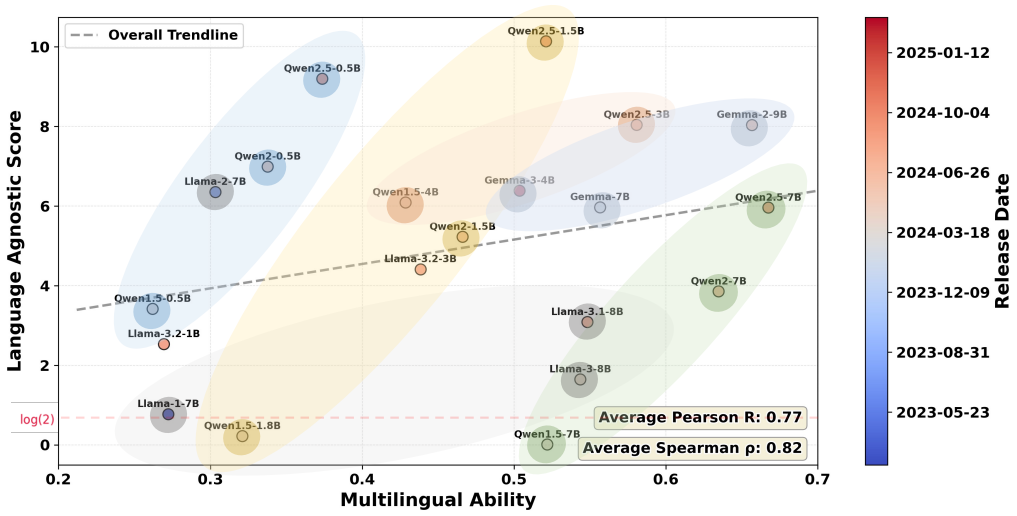

Figure 5: The relationship between multilingual ability and language-agnostic score (as defined in Equation 6) across various language models. Each point represents a model, colored by model size. The red dashed line (- - - -) indicates where shared neuron influence surpasses that of exclusive neurons, i.e., $\text{Imp} = 1$ and $\text{Language Agnostic Score} = \log(2)$. Shaded regions group models within the same series and of similar scale, while the dashed line (- - - -) indicates the overall trend: as LLMs evolve, successive generations with enhanced multilingual capabilities tend to achieve higher language-agnostic scores. This trend suggests that shared neurons increasingly support not only multilingual processing but also the emergence of more language-agnostic, abstract thought.

which we deactivate an equal number of randomly selected neurons, matching the quantity of both language-shared and language-exclusive neurons. As illustrated in Figure 4, this random deactivation results in minimal changes in perplexity across languages. In contrast, deactivating language-shared or language-exclusive neurons leads to significant performance degradation. These results confirm that the identified language-related neurons are indeed specialized for language processing, and that our neuron importance metrics are robust to random perturbations.

To further investigate whether language-shared neurons evolve into language-agnostic neurons, we analyze the evolution of the language-agnostic score, as defined in Equation 6, in relation to the models' multilingual capabilities, as shown in Figure 5.

Shared neurons in early-stage models reflect superficial overlap without supporting higher-level cognition. In earlier models such as the Qwen1.5 series and LLaMA-1-7B, deactivating shared neurons has a comparable effect to deactivating exclusive neurons, with language-agnostic scores around 1. This suggests that in early-stage models, shared neurons have the similar importance with exclusive neuron, and shared neurons largely reflect superficial overlaps between language-related neuron across languages, rather than representing a distinct, functionally meaningful shared space.

Shared neurons in recent models become central and exhibit language-agnostic properties. In contrast, recent models, such as those in the Qwen2.5 series, exhibit a dramatically different pattern. Deactivating shared neurons leads to a sharp and disproportionate increase in perplexity across all languages, often several orders of magnitude greater than the increase caused by removing language-exclusive neurons. In other words, shared neurons contribute far more critically to multilingual processing than exclusive neurons, despite both being part of the language-related neuron set. This disproportionate degradation reveals that shared neurons in recent models have evolved beyond serving merely as intersections of language-specific components; they now fulfill more fundamental, language-agnostic roles. If such shared neurons have indeed evolved into language-agnostic neurons, they may be operating within a conceptual space that abstracts away from surface-level linguistic variations. Such a space would allow the model to perform high-level reasoning, semantic alignment, and cross-lingual generalization—hallmarks of abstract thought in multilingual LLMs.

Table 1: Multilingual performance improvements on MGSM (primarily involving abstract thought) and MMMLU (requiring both abstract thought and domain knowledge) across five languages. Models were trained only on 100,000 general documents without reasoning-related data and evaluated using Llama-3.1-8B (high language-agnostic), Llama-3.2-3B (medium language-agnostic), and Llama-3.2-1B (low language-agnostic) under various targeted neuron tuning strategies.

	Neuron	MGSM						MMMLU					
		Zh	Fr	De	Th	Sw	Δ_{Avg}	Zh	Fr	De	Th	Sw	Δ_{Avg}
Llama-3.1-8B	None	52.4	51.6	54.4	46.8	38.8	-	53.8	58.4	56.9	48.8	40.9	-
	Shared	$52.0^{-0.4}$	$52.8^{+1.2}$	$55.6^{+1.2}$	$45.6^{-1.2}$	$39.6^{+0.8}$	0.3	$54.6^{+0.8}$	$57.2^{-1.2}$	$56.5^{-0.4}$	$48.9^{+0.1}$	$42.3^{+1.4}$	0.1
	Exclusive	$56.8^{+4.4}$	$57.2^{+5.6}$	$57.2^{+2.8}$	$50.4^{+3.6}$	$42.4^{+3.6}$	**4.0**	$55.6^{+1.8}$	$59.2^{+0.8}$	$59.1^{+2.2}$	$49.9^{+1.1}$	$43.7^{+2.8}$	**1.7**
	Random	$50.4^{-2.0}$	$51.2^{-0.4}$	$54.4^{-0.0}$	$47.2^{+0.4}$	$37.6^{-1.2}$	-0.6	$52.4^{-1.4}$	$58.3^{-0.1}$	$57.2^{+0.3}$	$47.1^{-1.7}$	$41.3^{+0.4}$	-0.5
Llama-3.2-3B	None	40.8	42.4	57.2	35.2	30.8	-	45.2	49.0	47.1	40.6	34.1	-
	Shared	$42.8^{+2.0}$	$45.6^{+3.2}$	$66.4^{+9.2}$	$40.4^{+5.2}$	$39.6^{+8.8}$	**5.7**	$44.9^{-0.3}$	$49.8^{+0.8}$	$47.3^{+0.2}$	$41.0^{+0.4}$	$34.8^{+0.7}$	**0.4**
	Exclusive	$42.4^{+1.6}$	$43.2^{+0.8}$	$65.6^{+8.4}$	$37.2^{+2.0}$	$36.0^{+5.2}$	3.6	$44.9^{-0.3}$	$48.9^{-0.1}$	$47.1^{+0.0}$	$40.9^{+0.3}$	$34.7^{+0.6}$	0.1
	Random	$40.4^{-0.4}$	$41.6^{-0.8}$	$63.2^{+6.0}$	$34.8^{-0.4}$	$30.0^{-0.8}$	0.7	$44.5^{-0.7}$	$49.0^{+0.0}$	$46.9^{-0.2}$	$40.3^{-0.3}$	$34.1^{+0.0}$	-0.2
Llama-3.2-1B	None	26.4	26.0	29.2	20.0	22.8	-	29.0	27.8	28.8	28.8	26.6	-
	Shared	$30.0^{+3.6}$	$30.4^{+4.4}$	$30.8^{+1.6}$	$22.4^{+2.4}$	$26.4^{+3.6}$	3.1	$29.2^{+0.2}$	$28.7^{+0.9}$	$29.5^{+0.7}$	$29.4^{+0.6}$	$26.8^{+0.2}$	**0.5**
	Exclusive	$27.6^{+1.2}$	$30.0^{+4.0}$	$34.4^{+5.2}$	$23.2^{+3.2}$	$30.4^{+7.6}$	**4.2**	$29.0^{-0.0}$	$28.0^{+0.2}$	$29.3^{+0.5}$	$28.2^{-0.6}$	$26.8^{+0.2}$	0.1
	Random	$26.8^{+0.4}$	$26.4^{-0.4}$	$29.6^{+0.4}$	$21.2^{+1.2}$	$26.4^{+3.6}$	1.0	$28.8^{-0.2}$	$28.3^{+0.5}$	$29.1^{+0.3}$	$28.6^{-0.2}$	$26.8^{+0.2}$	0.1

4 Multilingual Enhancement via Neuron-Targeted Training

4.1 Language Agnostic Score Guided Multilingual Enhance

Inspired by above insights, we propose various targeted neuron training methods to enhance models' multilingual capability according to their language agnostic score.

LLMs with low language agnostic score can train any language-related neurons. These models exhibit limited multilingual capabilities, indicating that all language-related neurons require improvement. To enhance their performance across languages, we propose training all language-related neurons, whether they are shared across languages or specific to individual ones.

LLMs with middle language agnostic score should train language-shared neurons. These models demonstrate a degree of multilingual capability; however, the language-shared neurons have not yet evolved to become truly language-agnostic. Given that language-shared neurons are more prevalent than language-exclusive ones in these models, it is essential to further train and refine them to more effectively enhance the models' multilingual performance.

LLMs with high language agnostic score should train language-exclusive neurons The language-shared neurons in these models have evolved into language-agnostic neurons, responsible for abstract thought. They are already well-trained and offer limited room for further improvement. Therefore, to enhance multilingual performance, it is necessary to focus on training the language-exclusive neurons in these LLMs.

4.2 Experiment Setup

Dataset To further validate our hypothesis and explore how to utilize our findings to efficiently enhance multilingual capability in LLMs, we conduct continuous pretraining on specific neurons using multilingual corpora. Specifically, we construct a training set by sampling 100,000 examples per language from a mixture of three widely used multilingual datasets: Culturax (Nguyen et al., 2024), MADLAD (Kudugunta et al., 2023), and Wikipedia (Guo et al., 2020).

Training Settings We utilize Llama3.2-1B (Grattafiori et al., 2024), Llama3.2-3B, and Lamma-3.1-8B as representative LLMs with low, medium, and high language-agnostic scores, respectively. We conduct experiments under three training settings: language-shared neurons, language-exclusive neurons, and an equal number of randomly selected neurons. To evaluate multilingual capability, we employ the MMMLU and MGSM benchmarks.

Experiment Results Table 1 demonstrates that the multilingual capabilities of language models can be effectively enhanced through targeted neuron-specific tuning. For Llama-3.2-1B, which exhibits a relatively low language-agnostic score, tuning both shared and exclusive neurons significantly improves the model's cross-lingual reasoning performance, yielding average gains of 3.1 and 4.2 points on the MGSM benchmark, respectively. In the case of Llama-3.2-3B, which has a moderate language-agnostic score, tuning language-shared neurons results in the greatest performance improvement—an average gain of 5.7 points on MGSM. This is likely because these neurons are more numerous and less well-trained than exclusive ones. Finally, for Llama-3.2-8B, which already possesses a high language-agnostic score, the language-shared neurons appear to be sufficiently trained; thus, tuning exclusive neurons leads to further enhancement of multilingual performance, with an observed improvement of 4.0 points on MGSM. Compared to the improvement observed on MGSM, the performance gain on MMLU—which relies more heavily on knowledge extraction—is relatively smaller. This suggests that our training approach primarily enhances the model's thinking capabilities rather than its factual recall. Moreover, since we exclusively utilize general documents without incorporating reasoning-specific data, the substantial improvement further validates the effectiveness of our neuron-targeted training methodology. Additional experimental settings, results on different backbone LLMs, comparisons with baseline methods (e.g., LoRA), and cross-lingual evaluation analyses are provided in Appendix E and Appendix F.

5 Related Work

Thinking Language of LLMs Large language models (LLMs) (Touvron et al., 2023b; OpenAI, 2023; Zhang et al., 2024a; Chen et al., 2024; Liu et al., 2025c; Gemma Team et al., 2025) demonstrate strong multilingual reasoning and transfer abilities (Pires et al., 2019; Wu and Dredze, 2019; You et al., 2025; Cai et al., 2025; Nooralahzadeh et al., 2020), raising questions about whether these models operate in a language-agnostic or language-specific concept space (Nanda et al., 2023; Schut et al., 2025a; Zhao et al., 2024b), and which language would the model "think" in. One stream of work supports the hypothesis that LLMs "think" in a concept space centered on the predominant training language. Zhong et al. (2024) analyzed LLMs trained predominantly on English or Japanese (Fujii et al., 2024; LLM-jp et al., 2024) for their mainly activated languages; Fierro et al. (2025) showed language dependence in object retrieval; and Schut et al. (2025b), found representations align more closely with English even on foreign inputs. On the other hand, a language-agnostic view is supported by either probing studies (Pires et al., 2019; Stanczak et al., 2022), neuron-level manipulations (Dumas et al., 2024; Brinkmann et al., 2025; Ding et al., 2024) or both (Wu et al., 2025; Wendler et al., 2024). Our work falls in line with Dumas et al. (2024); Wendler et al. (2024); Wu et al. (2025), with more fine-grained neuron-level results and a novel activation-and-training-based analysis method.

Multilingual Enhancement Early-on, multilingual enhancement is mainly approached from pre-training in works such as XLM, XLM-R(Conneau et al., 2020; Lample and Conneau, 2019) and M-BERT (Devlin et al., 2018). More post-training work, ranging from continual pre-training (Zhang et al., 2021; Cui et al., 2024; Liu et al., 2025b; Husain et al., 2024; Kuulmets et al., 2024) to fine-tuning (Muennighoff et al., 2023; Chen et al., 2023; Ahuja et al., 2024; Lai et al., 2023; Indurthi et al., 2024; Lai and Nissim, 2024; Zhao et al., 2024c) have emerged to effectively improve models' multilingual abilities, though rather sensitive to training corpus and settings. A parallel body of work focuses on prompt-based methods, either leaning on language alignment (Zhang et al., 2024b; Etxaniz et al., 2023; Zhao et al., 2024a) or instruction-following and attention (Wang et al., 2025b; Zhao et al., 2025b,a). However, Liu et al. (2024) points out the suboptimality in translation-based prompting pipelines. Our neuron-specific tuning strategy answers the academic call (Liu et al., 2024, 2025a) for a more comprehensive approach to multilingual enhancement than translation-based prompting, and provides a more efficient and task-neutral alternative than the post-training based methods.

6 Conclusion & Discussion

In this work, we explore the emergence of abstract thought in large language models through the lens of neuron behavior. By identifying and categorizing language-related neurons as either shared or exclusive, we uncover a consistent trend across model development: shared neurons not only increase in proportion but also grow in functional importance, eventually forming a compact yet

critical set of language-agnostic neurons. These neurons underpin the model's ability to generalize across languages and support abstract reasoning that transcends linguistic boundaries. Motivated by this insight, we introduce neuron-specific training strategies that adapt to the developmental stage of an LLM, whether or not it exhibits language-agnostic behavior. Extensive experiments confirm that our targeted training approach effectively enhances multilingual performance across diverse models. We believe this neuron-centric perspective opens new avenues for understanding and improving the generalization capabilities of LLMs in multilingual and cross-lingual contexts.

Our study is limited by computational resources, which restricts evaluation on the larger LLMs and prevents full exploration of the potential of neuron-centric training at larger scales. We leave these directions for future work. Nonetheless, our findings shed light on the emergence of multilingual and abstract reasoning in LLMs, which may promote language equity but also raise risks like cross-lingual misinformation, calling for responsible deployment.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-005) and National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This project was also partially supported by the Singapore Ministry of Education Academic Research Fund Tier 1 (Award Number: T1 251RES2514) and TPU Research Cloud.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *LREC*, pages 4344–4355. European Language Resources Association.

Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Awadallah, Monojit Choudhary, Vishrav Chaudhary, and Sunayana Sitaram. 2024. sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large language models share representations of latent grammatical concepts across typologically diverse languages.

Yuchen Cai, Ding Cao, Xin Xu, Zijun Yao, Yuqing Huang, Zhenyu Tan, Benyi Zhang, Guiquan Liu, and Junfeng Fang. 2025. On predictability of reinforcement learning dynamics for large language models. *arXiv preprint arXiv:2510.00553*.

Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cu. 2023. Lifelong language pretraining with distribution-specialized experts.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. On softmax direct preference optimization for recommendation. In *NeurIPS*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

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.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Chenlu Ding, Jiancan Wu, Yancheng Yuan, Jinda Lu, Kai Zhang, Alex Su, Xiang Wang, and Xiangnan He. 2024. Unified parameter-efficient unlearning for llms. *arXiv preprint arXiv:2412.00383*.

Clément Dumas, Veniamin Veselovsky, Giovanni Monea, Robert West, and Chris Wendler. 2024. How do llamas process multilingual text? a latent exploration through activation patching. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english?

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts?

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

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

Gemma Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mandy Guo, Zihang Dai, Denny Vrandecic, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *LREC*, pages 2440–2452. European Language Resources Association.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models via romanization.

Sathish Reddy Indurthi, Wenxuan Zhou, Shamil Chollampatt, Ravi Agrawal, Kaiqiang Song, Lingxiao Zhao, and Chenguang Zhu. 2024. Improving multilingual instruction finetuning via linguistically natural and diverse datasets.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton

Tsitsulin, Róbert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucinska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, and Ivan Nardini. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *NeurIPS*.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer.

Huiyuan Lai and Malvina Nissim. 2024. mcot: Multilingual instruction tuning for reasoning consistency in language models.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025a. Is translation all you need? a study on solving multilingual tasks with large language models.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. 2025b. Principled multimodal representation learning. *arXiv preprint arXiv:2507.17343*.

Xiaohao Liu, Xiaobo Xia, Weixiang Zhao, Manyi Zhang, Xianzhi Yu, Xiu Su, Shuo Yang, See-Kiong Ng, and Tat-Seng Chua. 2025c. L-mtp: Leap multi-token prediction beyond adjacent context for large language models. *arXiv preprint arXiv:2505.17505*.

LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada,

Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *LREC/COLING*, pages 4226–4237. ELRA and ICCL.

Jinjie Ni, Rui Mao, Zonglin Yang, Han Lei, and Erik Cambria. 2023. Finding the pillars of strength for multi-head attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14526–14540.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. Multilingual massive multitask language understanding (mmmlu). Hugging Face.

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.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025a. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025b. Do multilingual llms think in english?

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 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)*, pages 5701–5715.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Chengxin Wang, Yiran Zhao, Shaofeng Cai, and Gary Tan. 2025a. Investigating pattern neurons in urban time series forecasting. In *The Thirteenth International Conference on Learning Representations*.

Teng Wang, Zhenqi He, Wing-Yin Yu, Xiaojin Fu, and Xiongwei Han. 2025b. Large language models are good multi-lingual learners : When llms meet cross-lingual prompts.

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.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Runyang You, Yongqi Li, Meng Liu, Wenjie Wang, Liqiang Nie, and Wenjie Li. 2025. Parallel test-time scaling for latent reasoning models.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024a. On generative agents in recommendation. In *SIGIR*, pages 1807–1817. ACM.

Yongheng Zhang, Qiguang Chen, Min Li, Wanxiang Che, and Libo Qin. 2024b. Autocap: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. 2021. Cpm-2: Large-scale cost-effective pre-trained language models.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, et al. 2025a. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *arXiv preprint arXiv:2505.15257*.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, et al. 2025b. Mpo: Multilingual safety alignment via reward gap optimization. *arXiv preprint arXiv:2505.16869*.

Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. 2024a. Lens: Rethinking multilingual enhancement for large language models. *arXiv preprint arXiv:2410.04407*.

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, et al. 2025c. Babel: Open multilingual large language models serving over 90% of global speakers. *arXiv preprint arXiv:2503.00865*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? In *NeurIPS*.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024c. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in?

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly articulate the paper's key contributions, including the discovery of language-agnostic neurons and the proposal of neuron-centric training strategies, which are thoroughly supported by the analyses and experiments in the main text.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitations of our work in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: While the main paper focuses on empirical findings, we provide the full set of assumptions and a complete proof for necessary theoretical analysis, such as the parallel neuron detection method, in the supplementary material, ensuring theoretical soundness.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all necessary details for reproducing our main experimental results in Section 3.1 and Section 4.2. Additionally, we include an anonymous link to our code in the abstract to further support reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymous link to the code is provided in the abstract, and it includes sufficient instructions to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all necessary details in Section 3.1 and Section 4.2. Further details will be included in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars where applicable and include both Pearson and Spearman correlation coefficients to ensure the statistical reliability and robustness of our findings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the necessary information on compute resources used for each experiment. Further details are included in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research fully adheres to the NeurIPS Code of Ethics. We ensure transparency, reproducibility, and responsible use of models, and our work does not involve sensitive data, human subjects, or harmful applications.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We explicitly discuss the broader impacts of our work in Section 6.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work relies solely on publicly available open-source models and datasets. We do not release any new models or data that would require additional safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets used in this work, including models and datasets, are open-source and properly credited. Their licenses and terms of use have been respected as per their original distribution.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include all necessary details about the dataset and codes in Section 3.1, Section 4.2, and the anonymous link.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This study does not involve human subjects and thus does not require IRB approval.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were only used for minor writing and editing purposes, and did not contribute to the core methodology or scientific content of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Parallel Neuron Detection Algorithm

Inspired by Zhao et al. (2024b); Wang et al. (2025a), the neuron detection method in Equation 2 can be done parallel. While Equation 2 considers the change in the final output embedding, the parallel methods described here efficiently calculate the change in the output of the *specific layer containing the neuron* when that neuron is deactivated. This layer-wise impact serves as a proxy or component for the overall impact.

In this context, let $X \in \mathbb{R}^{l \times d_{model}}$ be the input hidden states to a given layer, where l is the sequence length and d_{model} is the hidden dimension of the model. For a neuron $\mathcal{N}$ within this layer, its impact is measured as $\|f(X;\Theta) - f(X;\Theta_{\ominus\mathcal{N}})\|_2$, where $f(X;\Theta)$ is the layer's output with parameters Θ, and $f(X;\Theta_{\ominus\mathcal{N}})$ is the output when neuron $\mathcal{N}$ (a specific row or column in Θ) is deactivated (its parameters set to zero).

A.1 Feed-Forward Network (FFN) Neurons

A standard FFN layer in modern transformer models can be expressed as:

$$\text{FFN}(X) = (\text{SiLU}(XW_{gate}) \odot (XW_{up}))\, W_{down} \tag{7}$$

where $X \in \mathbb{R}^{l \times d_{model}}$ is the input to the FFN layer, $W_{gate}, W_{up} \in \mathbb{R}^{d_{model} \times d_{inter}}$, and $W_{down} \in \mathbb{R}^{d_{inter} \times d_{model}}$. Here, d_{inter} is the intermediate dimension of the FFN. The symbol $\odot$ denotes element-wise multiplication. Let $H_{act} = \text{SiLU}(XW_{gate}) \odot (XW_{up})$ be the intermediate activation matrix, $H_{act} \in \mathbb{R}^{l \times d_{inter}}$. Thus, the FFN output is $Y_{FFN} = H_{act}W_{down} \in \mathbb{R}^{l \times d_{model}}$.

We consider a neuron $\mathcal{N}_{inter,k}$ to be associated with the k-th dimension of the intermediate representation H_{act}. Deactivating such a neuron means that the k-th column of H_{act}, denoted $H_{act}[:,k]$, is effectively zeroed out before the multiplication with W_{down}. This deactivation corresponds to zeroing out the parameters that produce this k-th intermediate feature, e.g., the k-th column of W_{up} (i.e., neuron $\mathcal{N}$ is $W_{up}[:,k]$) and W_{gate}, or by zeroing out parameters that read from it, e.g., the k-th row of W_{down} (i.e., neuron $\mathcal{N}$ is $W_{down}[k,:]$).

Let $Y_{FFN,\ominus\mathcal{N}_{inter,k}}$ be the output when the k-th intermediate neuron is deactivated. The change in the layer's output is:

$$\Delta Y_{FFN,k} = Y_{FFN} - Y_{FFN,\ominus\mathcal{N}_{inter,k}}$$

If H'_{act} is H_{act} with its k-th column zeroed, then $Y_{FFN,\ominus\mathcal{N}_{inter,k}} = H'_{act}W_{down}$. So,

$$\Delta Y_{FFN,k} = (H_{act} - H'_{act})W_{down}$$

The matrix $(H_{act} - H'_{act})$ is zero everywhere except for its k-th column, which consists of the elements $H_{act}[:,k]$. Let this difference matrix be δH_k. Then $\Delta Y_{FFN,k} = \delta H_k W_{down}$. This resulting $l \times d_{model}$ matrix is formed by the outer product of the k-th column of H_{act} and the k-th row of W_{down}:

$$\Delta Y_{FFN,k} = H_{act}[:,k](W_{down})_{k,:}$$

The impact of the k-th intermediate FFN neuron is then the L2 norm of this change:

$$\|\Delta Y_{FFN,k}\|_2 = \|H_{act}[:,k](W_{down})_{k,:}\|_2 \tag{8}$$

This computation can be performed in parallel for all $k \in \{1, \ldots, d_{inter}\}$ to obtain the impact of all intermediate neurons in the FFN layer.

A.2 Self-Attention Network Neurons

The output of a self-attention layer (for simplicity, we describe a single attention head; multi-head attention involves similar computations per head) can be given by:

$$Y_{Attn} = \text{Softmax}\left(\frac{(XW_Q)(XW_K)^T}{\sqrt{d_k}}\right)(XW_V) \tag{9}$$

Let $Q = XW_Q \in \mathbb{R}^{l \times d_{attn}}$, $K = XW_K \in \mathbb{R}^{l \times d_{attn}}$, and $V = XW_V \in \mathbb{R}^{l \times d_{attn}}$, where d_{attn} is the dimension of queries, keys, and values for the attention mechanism. d_k is the scaling factor, typically the dimension of the key/query vectors (e.g., $d_k = d_{attn}$). Let $A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{l \times l}$. The layer output is $Y_{Attn} = AV \in \mathbb{R}^{l \times d_{attn}}$. (An additional output projection W_O might follow this, which would be multiplied subsequently).

A.2.1 Neurons in W_V

Consider a neuron $\mathcal{N}_{V,k}$ defined as the k-th column of W_V, i.e., $W_V[:,k]$. Deactivating this neuron sets $W_V[:,k]$ to zero, which in turn makes the k-th column of $V = XW_V$, denoted $V[:,k]$, zero. Let V' be the matrix V with its k-th column zeroed. The change in the layer's output is:

$$\Delta Y^{(V)}_{Attn,k} = AV - AV' = A(V - V')$$

The matrix $(V - V')$ is zero everywhere except for its k-th column, which is $V[:,k]$. Let this difference matrix be δV_k. Then $\Delta Y^{(V)}_{Attn,k} = A(\delta V_k)$. This $l \times d_{attn}$ matrix has $AV[:,k]$ (the matrix A multiplied by the vector $V[:,k]$) as its k-th column, and zeros in other columns. The impact of neuron $\mathcal{N}_{V,k}$ is:

$$\left\|\Delta Y^{(V)}_{Attn,k}\right\|_2 = \|AV[:,k]\|_2 \tag{10}$$

where the norm is effectively taken over the $l \times 1$ vector $AV[:,k]$ that forms the k-th column of the change matrix. This can be calculated in parallel for all $k \in \{1, \ldots, d_{attn}\}$.

A.2.2 Neurons in W_Q

Consider a neuron $\mathcal{N}_{Q,k}$ defined as the k-th column of W_Q, i.e., $W_Q[:,k]$. Deactivating this neuron sets $W_Q[:,k]$ to zero. This makes the k-th column of $Q = XW_Q$, denoted $Q[:,k]$, zero. Let Q' be the matrix Q with its k-th column zeroed. The original unnormalized attention scores are $S_{raw} = \frac{QK^T}{\sqrt{d_k}}$. The new unnormalized attention scores with $\mathcal{N}_{Q,k}$ deactivated are $S'_{raw} = \frac{Q'K^T}{\sqrt{d_k}}$. The change in the unnormalized scores due to deactivating $\mathcal{N}_{Q,k}$ is $\Delta S_{raw,k} = S_{raw} - S'_{raw} = \frac{(Q-Q')K^T}{\sqrt{d_k}}$. The matrix $(Q - Q')$ is zero everywhere except for its k-th column, which is $Q[:,k]$. Thus,

$$\Delta S_{raw,k} = \frac{(Q[:,k])(K[:,k])^T}{\sqrt{d_k}}$$

This $l \times l$ matrix represents the change in raw attention scores attributable to the interaction involving the k-th column of Q and the k-th column of K.

Let $A_{orig} = \text{softmax}(S_{raw})$ be the original attention probability matrix. Let $A_{\ominus\mathcal{N}_{Q,k}} = \text{softmax}(S_{raw} - \Delta S_{raw,k})$ be the attention probability matrix when neuron $\mathcal{N}_{Q,k}$ is deactivated. The change in the layer's output is:

$$\Delta Y^{(Q)}_{Attn,k} = A_{orig}V - A_{\ominus\mathcal{N}_{Q,k}}V = (A_{orig} - A_{\ominus\mathcal{N}_{Q,k}})V$$

The impact of neuron $\mathcal{N}_{Q,k}$ is:

$$\left\|\Delta Y^{(Q)}_{Attn,k}\right\|_2 = \left\|(A_{orig} - A_{\ominus\mathcal{N}_{Q,k}})V\right\|_2 \tag{11}$$

To calculate this efficiently for all $k \in \{1, \ldots, d_{attn}\}$ (corresponding to each column neuron in W_Q):

1. Compute the original $S_{raw} = \frac{QK^T}{\sqrt{d_k}}$ and $A_{orig} = \text{softmax}(S_{raw})$.
2. For each k, compute the specific change term $\Delta S_{raw,k} = \frac{Q[:,k](K[:,k])^T}{\sqrt{d_k}}$. This step can be parallelized by constructing a tensor $\Delta S_{raw} \in \mathbb{R}^{d_{attn} \times l \times l}$ where the slice $\Delta S_{raw}[k,:,:] = \Delta S_{raw,k}$.
3. For each k, compute the adjusted scores $S_{adjusted,k} = S_{raw} - \Delta S_{raw}[k,:,:]$.
4. For each k, compute $A_{\ominus\mathcal{N}_{Q,k}} = \text{softmax}(S_{adjusted,k})$.
5. For each k, calculate the impact norm $\|(A_{orig} - A_{\ominus\mathcal{N}_{Q,k}})V\|_2$.

A.2.3 Neurons in W_K

The impact of deactivating a neuron $\mathcal{N}_{K,k}$ (the k-th column of W_K) is calculated symmetrically to that of $\mathcal{N}_{Q,k}$. The same change term $\Delta S_{raw,k} = \frac{Q[:,k](K[:,k])^T}{\sqrt{d_k}}$ is used, reflecting the idea that this term captures the interaction component associated with the k-th features of both Q and K. The procedure then follows steps 3-5 as outlined for W_Q neurons, using this $\Delta S_{raw,k}$ to find the adjusted attention matrix and the resulting impact.

B Neuron Detection Threshold

An important implementation detail in identifying language-related neurons lies in the selection of the activation threshold σ. Rather than adopting a fixed global scalar, we employ a dynamic thresholding mechanism that adapts to each query. Specifically, as shown in our released implementation, for every query in a given language, we rank neurons based on their computed importance scores and select the top 1% as activated neurons. Subsequently, for each language ℓ, its language-specific neuron set $\mathcal{N}_{\text{lang}}^{\ell}$ is defined as the intersection of these top-ranked neurons across all queries belonging to that language.

This dynamic top-1% strategy ensures consistent sensitivity across languages with different overall activation magnitudes, allowing the model to capture meaningful variations without being biased by language-specific activation scales. The choice of 1% is empirically determined through a set of calibration experiments designed to balance selectivity and stability.

To validate the appropriateness of this threshold, we conduct a sanity check using random baselines. For each language, we compare the model degradation caused by deactivating the selected language-specific neurons with that caused by deactivating an equal number of randomly chosen neurons. In all cases, we observe that removing the identified neurons results in a drastic performance drop—often exceeding a $100\times$ increase in perplexity—while removing random neurons yields negligible effects. This substantial performance disparity confirms that the selected neurons are functionally meaningful and that the threshold effectively distinguishes critical neurons from background noise.

If, conversely, random neuron deactivation were to cause a comparable decline in performance, it would suggest that the threshold is too lenient, allowing excessive neurons to be classified as important. In such cases, the percentile threshold would be systematically reduced until the random baseline no longer impacts model behavior. This adaptive validation process ensures that the threshold σ remains both rigorous and empirically grounded across all examined languages and model families.

C Neuron Analysis

To further understand the internal organization of multilingual representations within large language models, we conduct a comprehensive neuron-level analysis. This section explores how language-shared, language-exclusive, and strictly unique neurons are distributed across different architectural components and model layers, offering insight into how multilingual models balance generalization and specialization.

C.1 Component-Level Distribution

We first analyze how neurons are distributed across major architectural components. Neurons are categorized into three groups: query-key (QK), value-output (VO), and feed-forward network (FFN). As shown in Table 2, shared neurons are primarily concentrated in the QK components, aligning with the general attention mechanism responsible for capturing cross-lingual relational patterns. In contrast, exclusive neurons are more prevalent in VO and FFN layers, indicating their more important role in language-specific transformations and output generation. This decomposition suggests that shared and exclusive neurons perform complementary roles in multilingual processing.

We find that the component-level trend remains consistent across model families: shared neurons concentrate within the attention's QK submodules, supporting cross-lingual abstraction, while exclusive neurons appear more prominently in VO and FFN blocks, handling language-specific representations and refinements.

C.2 Layer-wise Distribution

We further examine how these neurons are distributed across model layers. Figures 6 and 7 illustrate the proportion of shared and exclusive neurons across all layers for Gemma2-9B and LLaMA3.1-8B, respectively. To quantify the trend, we group layers into early, middle, and late stages and report the average proportions in Table 3.

The layer-wise distribution reveals distinct allocation patterns for shared and exclusive neurons. Exclusive neurons are more concentrated in the early and late layers, suggesting that language-

Table 2: Component-level neuron distribution across models. Percentages represent the proportion of shared or exclusive neurons (averaged across languages) within each component.

Model	Component	Shared (%)	Exclusive (%)
LLaMA3.1-8B	QK	92.50	59.48
	VO	2.90	23.80
	FFN	4.59	16.71
Qwen2.5-3B	QK	76.40	53.06
	VO	16.72	30.36
	FFN	6.88	16.58
Gemma2-9B	QK	65.32	35.40
	VO	22.02	40.46
	FFN	12.66	24.16

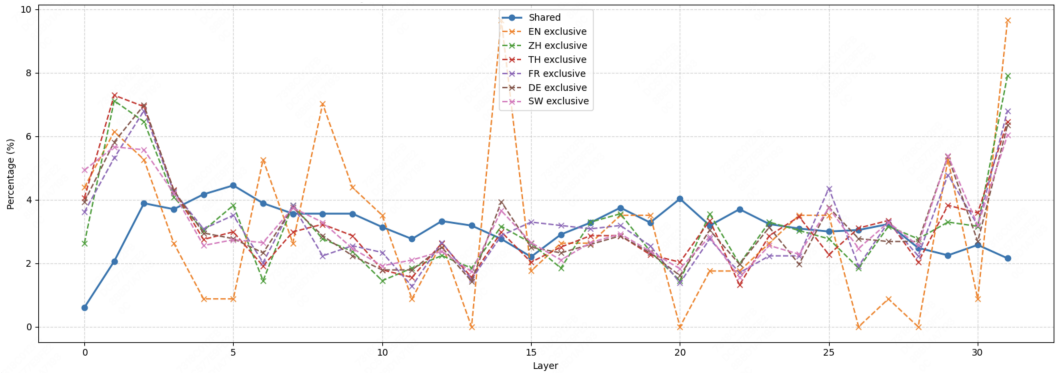

Figure 6: Layer-wise distribution analysis of Gemma2-9B.

specific processing primarily occurs near the input and output boundaries, where lexical and syntactic variations are handled. In contrast, shared neurons maintain a stable proportion across all layers, reflecting their role in capturing transferable, cross-lingual representations throughout the network.

Overall, these analyses demonstrate that while both neuron types are crucial for multilingual processing, shared neurons form a stable representational backbone that supports language-agnostic reasoning, whereas exclusive neurons enable fine-grained, language-specific adjustments near the model periphery. These findings are consistent across LLaMA, Qwen, and Gemma series, reinforcing the robustness of this observation.

C.3 Unique Language Neurons

To further refine our understanding of neuron selectivity, we investigate strictly unique neurons—those that respond exclusively to one language. While our definition of language-exclusive neurons allows for activation across a subset of languages, this stricter criterion provides additional insight into the specialization of multilingual models.

Table 4 reports the proportion of strictly unique neurons per language. These neurons constitute only a small percentage of the total population, suggesting that the model predominantly relies on shared neurons for multilingual understanding. Interestingly, higher values for Swahili and Thai—both lower-resource languages—indicate a stronger reliance on language-specific neurons, likely to compensate for limited training data.

We also analyze their layer-wise distribution by grouping the model layers into three stages: early (0–7), middle (8–23), and late (24–31). The averaged proportions of unique, shared, and exclusive neurons in each group are shown in Table 5. The early and late stages exhibit higher proportions of unique and exclusive neurons, implying that language-specific encoding and decoding occur near the input and output boundaries. The middle layers, dominated by shared neurons, correspond to a cross-lingual abstraction stage responsible for language-independent reasoning.

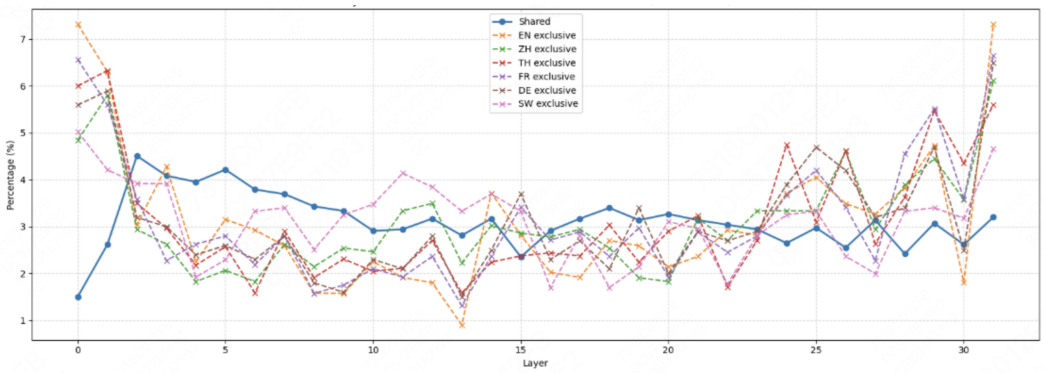

Figure 7: Layer-wise distribution analysis of LLaMA3.1-8B.

Table 3: Layer-wise neuron distribution across representative models. Percentages indicate average proportions of shared and exclusive neurons within each layer range.

Model	Layer Range	Shared (%)	Exclusive (%)
LLaMA3.1-8B	0–7	3.07	4.40
	8–23	3.11	2.89
	24–31	2.83	3.83
Qwen2.5-3B	0–5	3.69	4.45
	6–29	2.70	2.43
	30–35	2.20	2.48
Gemma2-9B	0–5	2.41	3.13
	6–34	2.39	1.92
	35–41	2.29	3.65

Finally, we examine the component-level distribution of strictly unique neurons (Table 6). Shared neurons cluster within QK, reinforcing their role in cross-lingual alignment. Exclusive neurons, as well as strictly unique ones, are more concentrated in VO and FFN, highlighting their specialization in language-specific semantic transformation and output projection.

These findings reinforce our broader conclusion: shared neurons underpin cross-lingual generalization, capturing transferable semantics; exclusive neurons encode language-specific nuances, crucial for accurate understanding and generation; and strictly unique neurons—being the most selective subset—reflect the model's fine-grained adaptation to individual languages, typically concentrated at the model's periphery where input encoding and output generation occur.

D Neuron Detection Corpus

This section provides additional details regarding the corpus used for neuron detection, as mentioned in the main text. Our methodology relies on the OSCAR corpus for both identifying language-related neurons through activation patterns and quantifying their functional contribution via perplexity changes upon deactivation.

D.1 OSCAR Corpus

The OSCAR (Open Super-large Crawled Aggregated coRpus) corpus (Abadji et al., 2022) is a massive multilingual collection of texts obtained by language classification and filtering of the Common Crawl dataset. Common Crawl is a publicly available web crawl spanning petabytes of data. OSCAR further processes this raw data to produce monolingual corpora across a wide range of languages, making it a valuable resource for training large language models and conducting cross-lingual research.

Key characteristics of the OSCAR corpus include:

- **Large Scale:** It contains hundreds of gigabytes to terabytes of text data for many languages.

Table 4: Proportion of strictly unique neurons (responding to only one language).

Language	Strictly Unique Neurons (%)
English (en)	0.02
Chinese (zh)	0.03
Thai (th)	0.06
French (fr)	0.03
German (de)	0.02
Swahili (sw)	0.07

Table 5: Layer-wise distribution of unique, shared, and exclusive neurons in LLaMA3.1-8B.

Layer Group	Unique (%)	Shared (%)	Exclusive (%)
Early (0–7)	3.69	3.07	4.40
Middle (8–23)	2.28	3.11	2.89
Late (24–31)	4.62	2.83	3.83

- **Multilingual Coverage:** It supports a vast number of languages, facilitating studies that require diverse linguistic data.
- **Data Cleaning:** Efforts are made to clean and filter the crawled data, though the quality can vary depending on the language and the nature of web content.
- **Accessibility:** OSCAR is publicly available, promoting reproducibility and broader research in NLP.

For our study, we sample 1000 sentences for each target language from its respective monolingual section within the OSCAR corpus. This sampled data serves as the basis for analyzing neuron activations and evaluating perplexity changes. The diversity and scale of OSCAR help in capturing a wide array of linguistic phenomena necessary for robustly identifying language-specific neural correlates.

D.2 Illustration of Sample Sentences

To provide a concrete illustration of the data used, Table 7 presents conceptual example sentences from the OSCAR corpus for the five languages central to our analysis: English (en), Chinese (zh), Swahili (sw), German (de), and French (fr).

The sentences sampled for each language are then further used to observe which neurons are consistently activated during processing. A similar set of sentences is then used to measure the perplexity of the model when specific neurons or sets of neurons are deactivated, thereby quantifying their functional importance to that language.

E Multilingual Enhancement

This section provides additional details on our multilingual enhancement experiments, including (1) the construction of random neuron baselines, (2) results on additional backbones, and (3) comparisons with LoRA, a famous parameter-efficient fine-tuning method. Together, these analyses validate the robustness and efficiency of our neuron-level enhancement strategy.

E.1 Illustration of Random Neurons

Regarding the random neurons presented in Figure 4, we carefully designed the sampling process to ensure a fair comparison. Specifically, we sampled two sets of random neurons—each matching the total number of shared and exclusive neurons, respectively. These random neurons were uniformly sampled across all layers and components of the model, under the assumption of a homogeneous distribution.

Table 6: Component-level comparison among unique, shared, and exclusive neurons in LLaMA3.1-8B.

Component	Unique (%)	Shared (%)	Exclusive (%)
QK	18.40	92.50	59.48
VO	50.96	2.90	23.80
FFN	30.64	4.59	16.71

Table 7: Illustrative sample sentences from the OSCAR corpus for the selected languages. These are conceptual examples, as actual sentences are randomly sampled.

Language	Conceptual Example
English (en)	The quick brown fox jumps over the lazy dog.
Chinese (zh)	敏捷的棕色狐狸跳过了懒惰的狗。
Swahili (sw)	Mbweha mwepesi wa kahawia anaruka juu ya mbwa mvivu.
German (de)	Der schnelle braune Fuchs springt über den faulen Hund.
French (fr)	Le renard brun rapide saute par-dessus le chien paresseux.

We then selected the random set that exhibited a stronger influence on model performance, and used it consistently in both Figure 4 and Table 1. This ensures comparability across evaluations. It is indeed expected that the distribution and activation patterns of random neurons differ from those of the identified language-specific neurons, as the latter capture semantically grounded linguistic features rather than arbitrary activation patterns.

E.2 Results on Additional Backbones

Figure 1 presents a simplified cross-model analysis within the same generation but across different model sizes. We observe that larger models generally exhibit a lower proportion and reduced importance of shared neurons. This aligns with prior findings showing that as models scale up, parameter specialization increases, leading to fewer neurons being shared across languages.

Although our main focus lies in analyzing the evolution of abstract thought over model development rather than size scaling, we include this discussion for completeness. To further verify the generality of our findings, we conducted additional analyses on the Qwen-2.5 family, particularly the 1.5B variant. The results are summarized below:

Table 8: Performance comparison of shared, exclusive, and random neuron sets on Qwen-2.5-1.5B. Metrics represent accuracy (%) on MGSM and MMMLU datasets.

	MGSM						MMMLU					
Model	zh	de	fr	th	sw	Δ	zh	de	fr	th	sw	Δ
Qwen2.5-1.5B	63.60	57.20	61.20	50.80	28.00	–	53.95	48.47	50.96	44.00	30.49	–
+ Exclusive	65.20	57.60	61.60	52.00	29.60	+1.04	53.81	48.65	51.29	44.49	31.57	+0.39
+ Shared	63.20	56.80	62.00	49.20	31.20	+0.32	53.72	48.38	51.30	44.42	31.48	+0.29
+ Random	62.80	54.00	60.40	46.80	27.60	-2.88	53.71	48.62	51.09	44.17	31.36	-0.17

These results indicate that both shared and exclusive neuron adjustments consistently improve multilingual reasoning, whereas random neuron updates negatively affect performance.

We further tested our findings on the Gemma2-9B model, which represents a different model family and a high language-agnostic score at the global level. The results are presented in Table 9.

We observe consistent trends across backbones: fine-tuning exclusive neurons enhances reasoning in target languages, while shared neurons contribute to general stability. In contrast, random or unstructured modifications fail to improve multilingual alignment.

Furthermore, our GSM8K experiments confirm that fine-tuning a model on a specific language improves reasoning in that language but may degrade performance in others—supporting our hypothesis that language-specific adaptation often comes at the cost of reduced cross-lingual transferability.

Table 9: Performance of neuron subsets on Gemma2-9B across languages.

Subset	Zh	Fr	De	Th	Sw	Avg Δ
None	58.4	58.0	58.8	57.2	51.2	–
Shared	56.8	57.6	58.8	54.8	48.4	-1.4
Exclusive	61.6	60.8	62.4	58.4	55.6	+3.0
Random	56.0	57.6	57.2	56.0	50.8	-1.2

E.3 Comparison with LoRA

Although our method focuses on neuron-level tuning rather than introducing new fine-tuning layers, we further conduct a comparison with LoRA to verify the efficiency and effectiveness of our approach. Unlike LoRA and other parameter-efficient fine-tuning (PEFT) methods, which insert additional adapter layers and train extra parameters, our method adjusts only a small subset of existing shared neurons. This enables multilingual enhancement without any architectural modification or parameter growth, highlighting a distinct trade-off between targeted internal adaptation and parameter-efficient extension.

For fair comparison, we implement LoRA-based fine-tuning on LLaMA3.2-3B under a similar parameter budget (rank = 48). As shown in Table 10, LoRA improves performance in certain MGSM cases (e.g., German, Swahili) but shows limited generalization and even degradation on MMMLU. Moreover, LoRA requires longer training time (2.2 hours on 2$\times$H200 GPUs) compared to our neuron-level method (1.5 hours), demonstrating that our approach achieves competitive multilingual gains with higher efficiency and simpler implementation.

Table 10: Comparison between our neuron-level fine-tuning and LoRA on LLaMA3.2-3B.

	MGSM						**MMMLU**					
Model	zh	de	fr	th	sw	Δ	zh	de	fr	th	sw	Δ
LLaMA3.2-3B	40.80	57.20	42.40	35.20	30.80	–	45.20	47.10	49.00	40.60	34.10	–
+ LoRA	38.80	68.80	44.00	31.20	37.20	+2.72	44.40	45.79	47.74	39.58	32.50	-1.12

Overall, LoRA demonstrates partial improvements but lacks consistency across benchmarks and languages. In contrast, our neuron-level approach achieves stable multilingual enhancement with lower computational overhead, highlighting its simplicity and interpretability as a complementary direction to PEFT methods.

F Cross-lingual Evaluation

To further explore the effect of language-specific fine-tuning on multilingual generalization, we conduct cross-lingual evaluation experiments. While our primary focus is on understanding how fine-tuning with a single language corpus can enhance performance in that language, it is equally important to assess how such adaptation influences the model's capabilities across other languages.

We fine-tune LLaMA-3.2-3B using corpora from individual languages and then evaluate its performance on all target languages. Two metrics are examined: (1) accuracy improvement on the multilingual GSM8K (MGSM) benchmark, reflecting reasoning capability; and (2) change in language perplexity (PPL), reflecting language understanding and fluency.

Table 11 reports the results of fine-tuning on one language and testing across all others. "Target Language Acc" denotes the improvement on the language used for fine-tuning, while "Other Languages Acc (Avg)" shows the average accuracy change over the remaining four languages.

We also measure the change in perplexity (PPL) before and after fine-tuning as an intuitive indicator of the model's linguistic understanding. A decrease in PPL indicates improved fluency and comprehension in that language. The results are summarized in Table 12.

These results demonstrate a consistent trade-off: fine-tuning on a single language improves both reasoning ability and linguistic understanding in that language, but often at the expense of reduced performance on others. This observation suggests that language-specific adaptation repurposes part

Table 11: Cross-lingual evaluation on MGSM for LLaMA-3.2-3B. Fine-tuning on a single language improves reasoning in that language but moderately reduces performance in others.

Language Trained On	Target Language Acc	Other Languages Acc (Avg)
Chinese (Zh)	+2.0	-1.6
French (Fr)	+3.2	-2.4
German (De)	+9.2	-3.2
Thai (Th)	+5.2	-2.8
Swahili (Sw)	+8.8	-2.0

Table 12: Change in language perplexity (PPL) before and after fine-tuning on LLaMA-3.2-3B. Negative values indicate reduced perplexity (better language modeling).

Language Trained On	Target Language PPL	Other Languages PPL (Avg)
Chinese (Zh)	-3.15	+0.89
French (Fr)	-2.03	+0.77
German (De)	-2.76	+0.64
Thai (Th)	-0.41	+0.75
Swahili (Sw)	-12.40	+1.84

of the shared neuron subspace to better align with the target language, consequently weakening cross-lingual generalization.

In summary, language-specific fine-tuning enhances targeted capabilities while moderately compromising multilingual balance, implying that shared neurons serve as a critical mechanism for maintaining cross-lingual consistency.

G Broder Impacts

This work contributes to a deeper understanding of how LLMs develop multilingual and abstract reasoning capabilities, which may help improve language equity in AI systems. By enhancing performance across diverse languages, our methods could benefit underrepresented linguistic communities. However, stronger multilingual models also carry risks, such as enabling more sophisticated misinformation in multiple languages. We encourage responsible use and further research into safeguards for multilingual LLM deployment.

H Limitations

While our study provides compelling evidence for the emergence of abstract, language-agnostic thought in LLMs and demonstrates the effectiveness of neuron-centric training strategies, several limitations remain: First, due to resource constraints, our analysis—though conducted across diverse model families and scales—has not been extended to the larger LLMs. Whether the observed patterns of neuron sharing and functional importance generalize to such scales remains an open question. Second, although our proposed training strategy yields consistent gains, the scope of our experiments remains limited by computational cost. We have not fully explored the upper bound of performance improvements that could be achieved with larger-scale or longer-term neuron-centric training. We leave these limitations as future work.