What Language(s) Does Aya-23 Think In? How Multilinguality Affects Internal Language Representations

Katharina Trinley[1,*], **Toshiki Nakai**[1,*], **Tatiana Anikina**[2], **Tanja Baeumel**[2]
[1]Saarland University
[2]German Research Center for Artificial Intelligence (DFKI)
{katr00001, tona00002}@stud.uni-saarland.de

Abstract

Large language models (LLMs) excel at multilingual tasks, yet their internal language processing remains poorly understood. We analyze how Aya-23-8B, a decoder-only LLM trained on balanced multilingual data, handles code-mixed, cloze, and translation tasks compared to predominantly monolingual models like Llama 3 and Chinese-LLaMA-2. Using logit lens and neuron specialization analyses, we find: (1) Aya-23 activates typologically related language representations during translation, unlike English-centric models that rely on a single pivot language; (2) code-mixed neuron activation patterns vary with mixing rates and are shaped more by the base language than the mixed-in one; and (3) Aya-23's language-specific neurons for code-mixed inputs concentrate in final layers, diverging from prior findings on decoder-only models. Neuron overlap analysis further shows that script similarity and typological relations impact processing across model types. These findings reveal how multilingual training shapes LLM internals and inform future cross-lingual transfer research. The code and dataset are publicly available[1].

1 Introduction

Large language models (LLMs) excel in multilingual tasks (Srivastava et al., 2022; Bang et al., 2023; Gurgurov et al., 2025b), but their internal handling of multiple languages remains underexplored (Kaddour et al., 2023). While methods like logit lens (Wendler et al., 2024; Schut et al., 2025) and neuron specialization (Tang et al., 2024; Kojima et al., 2024; Tan et al., 2024) have been applied, prior work mainly targets English-centered models on monolingual tasks (e.g., cloze or repetition tasks), rather than balanced multilingual architectures and their processing of code-mixed texts.

Multilingual models often default to English during intermediate processing, as described by the Multilingual Workflow (MWork) hypothesis (Zhao et al., 2024), which suggests LLMs convert non-English inputs into English internally before generating outputs. Supporting this, studies on reasoning language models (RLMs) (Wang et al., 2025) find reliance on internal "pivot" languages or scripts, even with other input languages. However, it remains unclear if this preference is unique to RLMs or a general pattern in all multilingual LLMs. Therefore, we ask:

H1: How do balanced multilingual models process translation tasks – do they activate multiple languages simultaneously, unlike English-centric models that rely on a single pivot language?

Neuron-level analyses have identified language-specific patterns (Kojima et al., 2024; Tang et al., 2024), but these studies predominantly examine English-based models, leaving open whether multilingual training leads to fundamentally different internal processing mechanisms. While LLMs' language capabilities are tied to specific neuron subsets, particularly in early and late layers (Kojima et al., 2024; Tang et al., 2024), these patterns may not apply to models trained on diverse multilingual data (Zhong et al., 2024a; Schut et al., 2025). We thus investigate the following hypotheses:

H2: What patterns of neuron sharing of language specific neurons emerge in balanced multilingual models, and do these align more strongly with language similarity compared to predominantly monolingual models?

H3: Where do language-specific neurons concentrate in multilingual architectures – do they cluster predominantly in final layers, contrary to prior findings showing distribution across early and late layers in decoder-only models?

In real-world contexts, speakers often mix languages within a single utterance, requiring models to dynamically switch between language-specific representations. Code mixing (CM) provides a

*Equal contribution

[1]https://github.com/KatharinaTrinley/multilingual-internal-representations

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 159–171, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-018

valuable lens for studying multilingual processing in language models (Xie et al., 2025), and while multilingual LLMs perform well on some tasks, they still struggle with code-switched text (Gundapu and Mamidi, 2020). The development of more balanced multilingual models, such as Aya-23 (Aryabumi et al., 2024), offers an opportunity to examine how different training approaches affect internal language representations, especially when handling the linguistic complexity of code-mixed inputs. Thus, we ask:

H4: How does the processing of code-mixed inputs vary based on language pair characteristics and models?

To address these questions, we perform a neuron-level comparison of a balanced multilingual model (Aya-23-8B), a predominantly English-trained model (Llama 3.1-8B), and a language-specialized model (Chinese-LLaMA-2-7B). Specifically, we:

I. Analyze internal language representations across 13-language translation tasks using logit lens to test **H1**, checking whether Aya-23 activates multiple languages simultaneously, unlike English-pivot processing in mostly monolingual models.

II. Create a controlled code-mixed dataset with varying mixing ratios across 10 typologically diverse pairs ({fr, zh} $\times$ {en, es, it, ja, ko}) and use neuron specialization (activation frequency (Tan et al., 2024)) to investigate **H2** and **H4**, exploring how script similarity and language relationships affect neuron sharing across models.

III. Examine layer-wise distribution of language-specific neurons via activation strength (Kojima et al., 2024) to test **H3**, determining whether balanced multilingual training concentrates language-specific neurons mainly in final layers, contrasting prior findings of early-and-late layer distributions in decoder-only models.

2 Methodology

We investigate the internal language representations in multilingual decoder-only LLMs through complementary experimental approaches: logit lens analysis (Section 2.3) and neuron specialization analysis (Section 2.4). Each methodology offers unique insights into how models process information across languages.

2.1 Models

We evaluate three models with varying multilingual focus. **Aya-23-8B** by Cohere AI is an open-source decoder-only model instruction fine-tuned on 23 languages—including ar, zh (simplified & traditional), en, fr, it, ja, ko, and more—using a two-stage process: pretraining on a balanced multilingual corpus (not public) and multilingual instruction fine-tuning (Aryabumi et al., 2024). **Llama 3.1-8B** supports 8 languages (en, fr, de, hi, it, pt, es, th) but was mainly trained on English data (ca. 8% multilingual tokens) and retains English-centric processing patterns, serving as a baseline for predominantly English-trained models (Grattafiori et al., 2024; Wendler et al., 2024). **Chinese-LLaMA-2-7B** is a Mandarin-adapted LLaMA-2 variant with an expanded tokenizer (+20,000 tokens), pretrained on large Chinese corpora using parameter-efficient fine-tuning (LoRA(Hu et al., 2021)) and instruction-tuned on millions of Chinese instruction-response pairs, enabling strong Chinese performance at low computational cost (Cui et al., 2023a,b; Hu et al., 2021).

2.2 Datasets

In this work, we focus on two primary datasets: the Dumas dataset (Dumas et al., 2024) for logit lens experiments and introduce a new code-mixed dataset that will be publicly released.

Dumas Dataset For logit lens experiments, we use the dataset from Dumas et al. (2024), which includes word translation and cloze tasks in 13 languages (de, en, es, et, fi, fr, hi, it, ja, ko, nl, ru, zh). It minimizes token overlap between languages while maintaining semantic consistency. Note that model support varies: Aya-23-8B lacks et and fi; Llama 3.1-8B excludes et, fi, ja, ko, nl, ru, and zh; Chinese-LLaMA-2-7B supports only zh and has limited en capabilities, lacking official support for the other 11 languages. Each prompt consists of randomly selected 5-shot word translation examples followed by a final query word. For instance, an English-to-Chinese task may appear as:

English: "computer" → 中文: 电脑
English: "ant" → 中文: 蚂蚁
English: "cloud" → 中文: 云
English: "heart" → 中文: 心脏
English: "knife" → 中文: 刀子
English: "book" → 中文: __

The task is to predict the correct translation of

the final word. Synonyms for the target word are included across all supported languages.

Code-mixed Dataset To study how models process mixed-language inputs, we construct a code-mixed dataset derived from the WMT24++ parallel corpus (Deutsch et al., 2025), containing 998 sentence pairs across 55 languages. We focus on a subsection of 7 languages and take fr and zh as base languages, each mixed with five partner languages (en, es, it, ja, and ko) resulting in ten language pairs. These combinations span a wide typological and script range, including closely related Romance/Indo-European languages (fr/es, fr/it, fr/en), typologically distinct but historically linked pairs (zh/ja, zh/ko), and diverse scripts: Latin (en, fr, es, it), Simplified Chinese (zh), Kanji/Kana (ja), and Hangul (ko).

We generate code-mixed sentences using a three-step rule-based method (Figure 1) with controlled mixing ratios of 25%, 50%, and 75%.

We tokenize Latin script using whitespace and Han script with the Jieba library (Junyi, 2012). Although this may yield ungrammatical outputs, it ensures consistent mixing ratios critical for controlled experiments. To address limited dictionary coverage in prior work (Conneau et al., 2017), we create comprehensive bilingual dictionaries via Google Translate for all WMT24++ words, ensuring equal vocabulary coverage across language pairs. However, lacking word sense disambiguation, polysemous words are translated identically regardless of context, possibly causing meaning mismatches.

To evaluate translation accuracy, we manually assessed word-level translation quality in code-mixed data, focusing on semantic mistranslations rather than grammatical errors common in code-mixing. From 50% mixing datasets, we sampled 10 sentences per language pair (246–399 words) and found translation error rates of fr-en 4.76% , fr-es 4.78% , zh-en 4.87% , and zh-es 8.94% , with higher errors for zh-es due to greater linguistic distance and weaker model performance.

To compare code-mixed and monolingual processing, we include corresponding monolingual datasets from WMT24++ (fr, es, it, ja, and ko) as baselines. All code-mixed pairs were evaluated on translation tasks directed from code-mixed input to en (i.e., Chinese-Spanish code-mixed input to en). We do not evaluate the reverse direction, as enforcing controlled code-mixing in model-generated outputs is challenging.

To further examine model behavior, we analyze neuron activation patterns (Section 2.4) across code-mixed inputs for Aya-23-8B, LLaMA 3.1-8B, and Chinese-LLaMA-2-7B, testing whether code-mixed processing differs by language pair and model architecture (**H4**).

2.3 Logit Lens

Logit lens (Nostalgebraist, 2020) interprets transformer hidden states by projecting intermediate representations into vocabulary space. At each layer ℓ, the model produces a hidden state $h_\ell \in \mathbb{R}^d$, which is mapped to logits using the unembedding matrix $U \in \mathbb{R}^{|V|\times d}$: $\text{logits}_\ell = Uh_\ell$.

These logits approximate the model's predictions at layer ℓ. Following Nostalgebraist (2020), we use the residual stream before layer normalization to better align with the final outputs. Building on prior multilingual analyses (Wendler et al., 2024; Zhong et al., 2024b; Saji et al., 2025), we apply the logit lens at each layer, extract token probabilities via softmax, and sum over synonyms in 13 languages using the dataset from Dumas et al. (2024) (see Section 2.2). To reduce false matches, we apply a 0.1 threshold. This approach allows us to track the emergence of language-specific signals across layers and test **H1**.

2.4 Neuron Specialization

Neuron specialization refers to individual neurons within language models developing preferences for processing specific types of input, such as particular languages.

Tan et al.'s Approach Tan et al. (2024)'s method identifies language-specific neurons by measuring how frequently they activate when processing different languages. Following Tan et al. (2024), we identify language-specific neurons via binary ReLU activations in FFNs across WMT24++ and code-mixed data.

For task t with validation set D_t, each sample x_i has activation $\mathbf{a}_i^t$. Summing gives $\mathbf{a}^t = \sum_{x_i \in D_t} \mathbf{a}_i^t$. Specialized neurons S_k^t are the top activations satisfying $\sum_{i \in S_k^t} \mathbf{a}^t(i) \geq k \sum_i \mathbf{a}^t(i)$. Neuron overlap is measured by $\text{IoU}(S^i, S^j) = \frac{|S^i \cap S^j|}{|S^i \cup S^j|}$. Using $k = 90\%$ per Tan et al. (2024), we identify neurons covering most activations per language and plot IoU matrices to expose cross-linguistic patterns. Unlike Tan et al. (2024), we exclude neurons shared by all languages to isolate language-specific neurons. This tests **H2**.

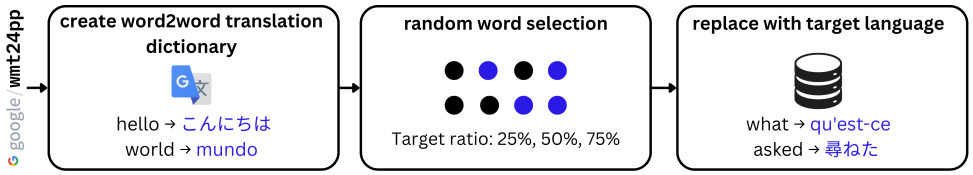

Figure 1: Our code-mixed dataset creation pipeline. Starting with parallel sentences from WMT24++, we create comprehensive bilingual dictionaries using Google Translate for all vocabulary. For each sentence, we randomly select words based on the target mixing ratio (25%, 50%, or 75%) and replace them with their translations in the partner language. For example, from the English source "The World Bank hopes to spread that message," we generate the code-mixed Chinese output "World bank希望传播这一理念" (50%).

Kojima et al.'s Approach Kojima et al. (2024) identified language-specific neurons in multilingual models, concentrated in early and late layers with minimal cross-language sharing. Their approach identifies neurons that discriminate between target language content and other languages by measuring activation strength.

We extend this to code-mixing neurons in Aya-23-8B's MLP layers. For each code-mixed pair l_t, texts are labeled positive ($b_i = 1$) or negative ($b_i = 0$). For neuron m and text $x_i = \{w_{i,1}, \ldots, w_{i,T}\}$, activations $\{z_{m,i,1}, \ldots, z_{m,i,T}\}$ are averaged as $z_{m,i} = f(z_{m,i,1}, \ldots, z_{m,i,T})$ (excluding padding). We compute Average Precision $AP_m = AP(z_m, b) \in [0, 1]$ to classify neurons into top-k (high), medium-k (none), and bottom-k (negative correlation). Applied to fr and zh code-mixed with en, it, es, jp, ko (10 pairs), this tests **H3** and **H4**.

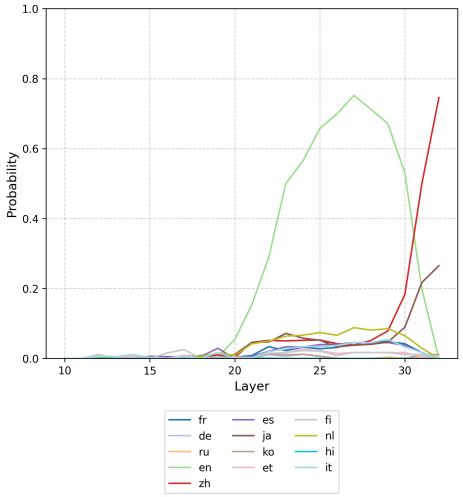

Figure 2: Logit lens language probabilities for English-to-Chinese translation in Aya-23-8B reveal activation of an increased number of languages in mid-to-late layers, with English being dominant.

3 Results and Discussion

3.1 Logit Lens Analysis

To test if balanced multilingual training affects internal processing (**H1**), we applied logit lens analysis (Wendler et al., 2024) to Aya-23-8B (balanced), LLaMA 3.1-8B (English-dominant), and Chinese-LLaMA-2-7B (Chinese-specialized).

Using Dumas et al. (2024)'s dataset, we tracked language-specific token probabilities across layers during translation. From 54 tasks, we computed AUCs for each language probability curve and used Mann-Whitney U tests with Bonferroni correction to compare: (1) model effects – whether Aya shows more diverse language representations than LLaMA ($p < 0.05/(13 \times 3) = 0.0013$), and (2) task effects – whether input vs. output languages differ in internal processing ($p < 0.05/(13 \times 3 \times 2) = 0.0006$).

Aya-23-8B demonstrates multilingual processing with cross-linguistic activation. During English-to-Chinese translation (Figure 2), Aya activates multiple languages in intermediate-to-late layers (20–27), including Japanese tokens despite Japanese being neither source nor target. This suggests Aya leverages typological relationships rather than relying solely on English as a pivot.

Llama 3.1-8B follows English-centric processing. In contrast (Figure 3), Llama demonstrates the English-dominated pattern established by Wendler et al. (2024), with English maintaining highest activation across all layers until final output generation. Chinese activates only in final layers, aligning with the "English-ization" process (Zhao et al., 2024).

Chinese-LLaMA-2-7B exhibits Chinese-dominant processing. This model shows Chinese representations dominating across most layers even for English-to-Chinese translation (Figure 4), with English activation decreasing in final layers while Japanese remains stable, reflecting its specialized training.

Our statistical analysis across all 54 translation

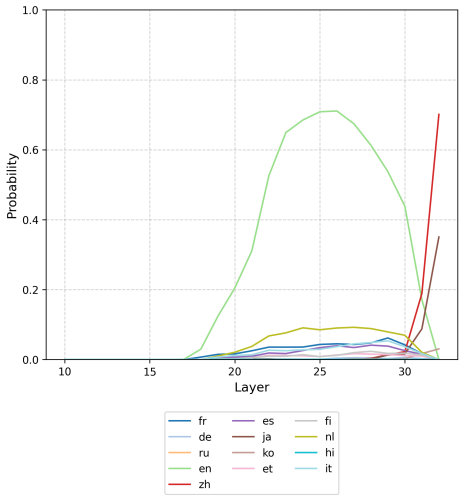

Figure 3: Logit lens language probabilities for English-to-Chinese translation in Llama 3.1-8B show dominant English representations across most layers with few other languages showing significant activation.

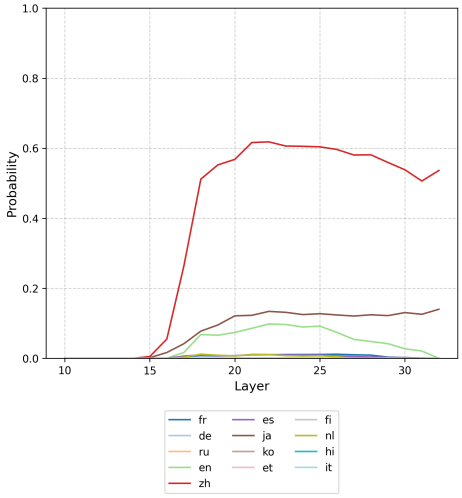

Figure 4: Logit lens language probabilities for English-to-Chinese translation in Chinese-LLaMA-2-7B show strong dominance of Chinese representations across most layers.

tasks provides quantitative support for **H1**: Aya demonstrates significantly different language activation patterns compared to both Llama (8/13 languages with $p < 0.0013$: de, ru, zh, es, ja, ko, it) and Chinese-LLaMA (8/13 languages including en, zh, es, ja, ko, it). Critically, output languages influence internal representations more strongly than input languages across all models, when analyzing task composition effects, output language presence produces significant changes in 12/13 languages compared to only 7/13 for input languages.

This analysis partially supports our hypothesis that Aya-23 incorporates multiple languages in internal processing, rather than relying solely on English. However, English still shows significantly higher activation probabilities, necessitating careful interpretation of these multilingual patterns. The statistical evidence highlights that both task language and model training paradigm significantly shape internal processing strategies, with task language particularly influencing language-specific activation probabilities.

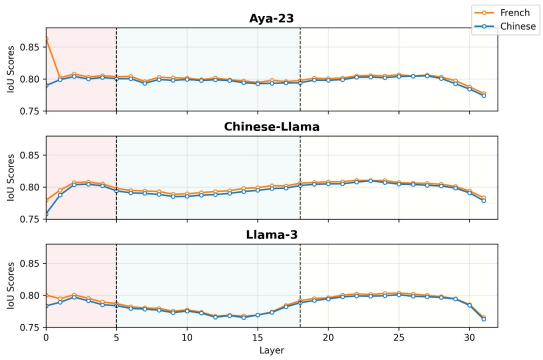

Figure 5: Three-phase neuron clustering patterns across transformer layers. French-based (orange) and Chinese-based (blue) code-mixed language pairs show distinct IoU overlap patterns in Aya-23, Chinese-LLaMA, and Llama-3.1. All models exhibit consistent French processing advantages.

3.2 Neuron Specialization Analysis

Activation *Frequency* Experiments Following Tan et al. (2024), we conducted neuron activation frequency experiments to examine how balanced multilingual training influences language-specific processing mechanisms (**H2**, **H4**).

To investigate base-language dependencies systematically, we conducted statistical analysis comparing French-based and Chinese-based code-mixed language pairs across all 32 transformer layers (see Figure 5). For each layer, we computed IoU overlap values within French-based pairs (105 combinations from 15 tasks) and within Chinese-based pairs (105 combinations from 15 tasks), yielding two distributions of IoU scores per layer. We applied the Wilcoxon signed-rank test to assess whether French-based pairs show significantly different neuron clustering patterns than Chinese-based pairs, using this non-parametric paired test since we're comparing corresponding layers between the two language groups.

French-based code-mixed inputs demonstrate significantly higher neuron clustering than Chinese-

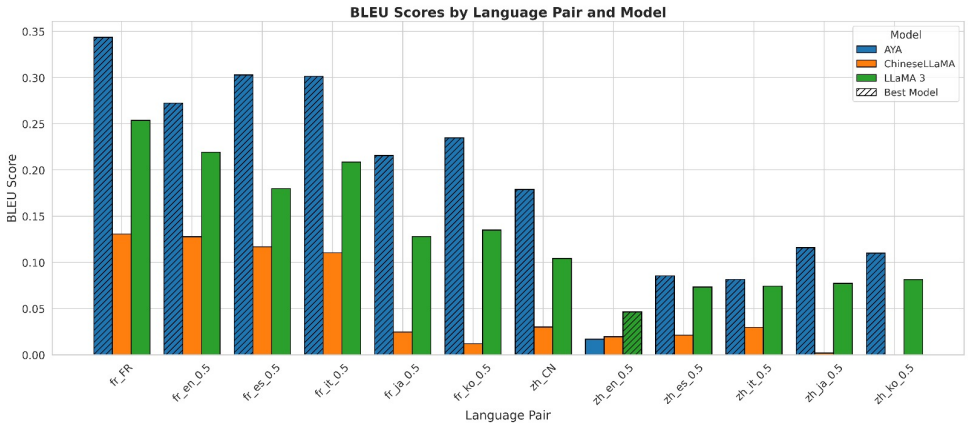

Figure 6: Translation qualities on code-mixed datasets using Aya23-8B, LLaMA 3.1-8B, and Chinese LLaMA, presented in BLEU.

based inputs across all three models. Wilcoxon signed-rank tests reveal strong statistical significance: Aya-23 ($p = 4.66 \times 10^{-10}$, mean difference = +0.0050), Chinese-LLaMA ($p = 4.66 \times 10^{-10}$, mean difference = +0.0041), and Llama-3 ($p = 9.31 \times 10^{-10}$, mean difference = +0.0029). This French advantage persists even in Chinese-LLaMA, a model specifically adapted for Chinese processing.

Our findings contradict **H2**, as neuron sharing patterns do not align with expected base-language training effects. Instead, they reveal a universal French processing advantage that transcends model architecture and training paradigm ($p < 10^{-9}$ across all models). This pattern strongly supports **H4** – that code-mixed processing varies systematically with language pair characteristics – and indicates that factors beyond training data composition, potentially including script characteristics or tokenization efficiency, drive neuron activation patterns in multilingual models.

Translation Performance on Code-Mixed Inputs Figure 6 presents BLEU scores for all three models on monolingual and code-mixed datasets. Aya-23-8B consistently outperforms the others, with a clear advantage on fr-based code-mixed inputs. All models show better performance on Latin-script pairs (fr-en, fr-es, fr-it) than on cross-script ones (fr-ja, fr-ko). For zh code-mixing, Aya-23-8B and Llama 3.1-8B perform better on zh-ja and zh-ko than on zh-en, zh-fr, and zh-it, suggesting that shared vocabulary and typological features help transfer despite script differences. In contrast, Chinese-LLaMA-2-7B performs poorly across all code-mixed inputs, regardless of typological similarity.

Performance generally degrades as code-mixing rate increases across all models, likely reflecting limitations of our rule-based word-to-word translation approach. However, Aya-23-8B shows greater resilience to this degradation, supporting our finding that balanced multilingual training improves robustness to code-mixing.

Activation *Strength* Experiments To address H4, we followed Kojima et al. (2024)'s methodology by processing both monolingual and code-mixed texts and capturing neuron activations at the MLP layers. Our findings for Aya reveal an interesting divergence from previous work on decoder-only model. While Kojima et al. (2024) found language-specific neurons (both top-k and bottom-k) concentrated in first and last layers of other decoder-only models, Aya-23-8B exhibits a different pattern when processing code-mixed input: top-k language-specific neurons appear predominantly in final layers (27-31), with a pronounced spike in layer 31 across all language pairs (see Figure 7). This pattern confirms our hypothesis **H3**.

This pattern only partially aligns with Tang et al. (2024), who observed a skewed "U"-shaped distribution, with language processing concentrated in both early and late layers. In contrast, it supports the findings of Mondal et al. (2025), who reported that language-specific neurons in modern LLMs are primarily concentrated in later layers. Our results suggest that Aya-23-8B's balanced multilingual training may promote a shift toward language-specific processing concentrated at the generation stage, diverging from the more distributed patterns

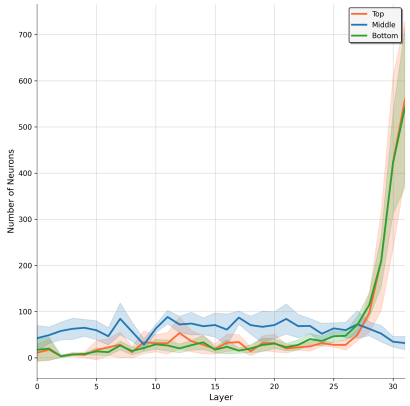

Figure 7: Layer-wise distribution of $k = 1000$ language-specific neurons in Aya-23-8B for code-mixed processing across all CM language pairs in Aya-23-8B.

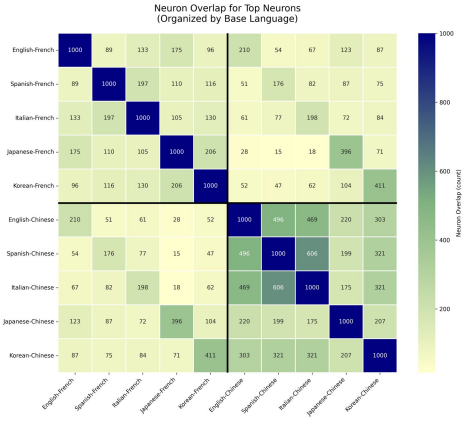

Figure 8: The number of overlapping language-specific neurons between code-mixing language pairs in Aya-23-8B.

seen in predominantly monolingual models.

This pattern remains consistent across all language pairs. Bottom-k neurons ("anti-correlated" neurons) similarly concentrate in final layers, while medium-k neurons distribute more evenly across early (0-5) and middle (10-20) layers.

This distinctive concentration pattern may stem from Aya's explicitly balanced multilingual training, resulting in an internal structure different from the predominantly monolingual models studied by Kojima et al. (2024). The pronounced spike of language-specific neurons in the final layer likely reflects Aya-23-8B's processing strategy for code-mixed inputs: earlier layers handle distributed multilingual representations for understanding, while the concentration at the generation stage resolves which language to output in mixed-language contexts.

Our analysis of neuron overlap shows that the base language influences neuron sharing more than the secondary mixed-in language. Chinese-based pairs consistently exhibit higher neuron overlap (17.5%–60.6%) compared to French-based pairs, regardless of the secondary language (Figure 8). This indicates that the foundational language's structural properties strongly shape neural organization. This pattern holds on average (French-based pairs: 138.25 neurons; Chinese-based pairs: 331.7 neurons; cross-base pairs: 82.65 neurons)[2].

Cross-script connections also appear, with ja-zh and ko-zh pairs showing moderate neuron overlap (20.7% and 41.1%, respectively), likely due to shared vocabulary and writing systems from historical contact. Within the fr-based group, neuron sharing varies: it-fr pairs have the highest overlap (19.7%), followed by en-fr (17.5%) and es-fr (8.9%), suggesting that typological similarity within the Romance family shapes neural processing patterns.

4 Related Work

Pivot Languages in Multilingual LLMs Training data composition fundamentally shapes multilingual processing patterns. Llama models, heavily trained on English (89% in Llama-2 (Touvron et al., 2023)), use English as a "pivot language" in multilingual tasks – translating French to English before Chinese, reducing quality (Wendler et al., 2024). This English bias extends beyond translation, with models defaulting to English in intermediate layers for reasoning (Zhao et al., 2024; Zhong et al., 2024a). The Multilingual Workflow (MWork) hypothesis (Zhao et al., 2024) formalizes this as: convert inputs to English for reasoning, integrate multilingual knowledge, then generate target output.

However, English-centric processing varies with architecture and training. Language-specific models like Swallow (Japanese-adapted Llama-2) and LLM-jp default to their dominant training language rather than English (Zhong et al., 2024a). Schut et al. (2025) found Aya-23 activated English ca. 50% versus ca. 70% in Gemma-2-27B, suggesting balanced training reduces English dominance. Similarly, Lindsey et al. (2025) identified language-agnostic conceptual representations in Claude 3.5

[2]Notable exceptions exist where typological similarity overrides base language effects, such as fr-ja with zh-ja (396 neurons) and ko-fr with ko-zh (411 neurons), likely reflecting historical Japanese-Chinese and Korean-Chinese linguistic contact

Haiku, indicating some models develop universal processing spaces beyond pivot strategies.

Language-Specific Neurons Language-specific neurons in decoder-only models cluster distinctly with minimal cross-language sharing. Kojima et al. (2024) and Tang et al. (2024) found these neurons concentrate in top and bottom layers of LLaMA-2, BLOOM, and Mistral, comprising only 1% of parameters. However, Mondal et al. (2025) observed newer models (Mistral Nemo, Llama 3.1) concentrate language-specific neurons primarily in later layers, indicating architectural evolution. Training data biases models toward English, degrading performance with increasing linguistic distance (Zhong et al., 2024a; Wendler et al., 2024), though positive cross-lingual transfer remains possible.

Recent work reveals dynamic language-specific processing. Tan et al. (2024) found feed-forward neurons in encoder-decoder models activate in language-specific patterns, with overlaps reflecting linguistic proximity. Deng et al. (2025) demonstrated that models dynamically shift activations based on context – Spanish prefixes amplify Spanish-specific features while suppressing others – suggesting sophisticated contextual language processing beyond fixed neuron assignments.

Code-Mixing and Script-Based Processing Code-mixing (CM) research reveals systematic biases in multilingual processing. Wang et al. (2025) showed reasoning language models activate Latin and Han scripts even when processing Arabic, Hindi, or Japanese, with performance gains up to 110% when constraining reasoning to preferred scripts. This suggests script-based processing preferences shaped by training data composition.

CM poses significant challenges for multilingual LLMs, particularly for low-resource languages. Gupta et al. (2024) found GPT models perform worse on English-Gujarati CM compared to English-French, reflecting training data imbalances toward high-resource monolingual corpora (Gundapu and Mamidi, 2020). Yang et al. (2020) demonstrated CM-specific pre-training improves translation performance, indicating models can learn to handle language transitions within utterances.

Our study addresses the underexplored gap between predominantly English-trained models (Llama) and balanced multilingual models (Aya-23), investigating whether reduced English reliance corresponds to distinct internal architectures through comprehensive neuron-level analysis across languages and code-mixed contexts.

5 Conclusion

Our investigation reveals that balanced multilingual training fundamentally alters how decoder-only LLMs process language internally. Through logit lens analysis, we show that Aya-23-8B employs distinct multilingual processing strategies, activating typologically related languages (e.g., Japanese during Chinese translation) and exhibiting significantly different activation patterns compared to English-centric models across 8/13 languages. We find that output languages influence internal representations more strongly than input languages.

Our neuron specialization analysis reveals that Aya-23-8B concentrates language-specific neurons predominantly in final layers (27-31) rather than distributing them across early and late layers as found in previous studies of decoder-only models (Kojima et al., 2024; Tang et al., 2024). This architectural difference suggests that balanced multilingual training creates models that maintain language-agnostic processing through most layers, with language-specific differentiation emerging primarily at generation time.

Code-mixed processing reveals systematic patterns driven by base language characteristics and script similarity. Base languages drive neuron sharing more strongly than mixed-in languages, with French-based code-mixed inputs maintaining consistent neuron overlap regardless of mixing rate, while Chinese-based inputs show proportional degradation. Translation performance demonstrates clear advantages for same-script language pairs, though Chinese-Japanese and Chinese-Korean pairs benefit from shared historical vocabulary despite script differences.

Limitations

Our study has several important limitations. A key one is the quality of our code-mixed dataset, created using rule-based word-to-word translation. This method overlooks grammatical structure and often yields unnatural sentences that may not reflect authentic code-switching. However, it allows systematic control of mixing ratios, which is essential for our neuron-level analysis.

Our methodology requires binarizing continuous neuron activations, leading to potential infor-

mation loss and obscuring subtle cross-language patterns. In our logit lens experiments, some token overlap likely remains between Japanese–Chinese and French–English, despite efforts to minimize it, which may affect analysis of language-specific activations. Additionally, our implementation of Tan et al. (2024)'s neuron specialization analysis revealed weak sharing patterns in heatmap visualizations, limiting the strength of our conclusions on language-specific processing.

Our analysis is limited to three models (Aya-23-8B, Llama 3.1-8B, and Chinese-LLaMA-2-7B) and may not generalize to other multilingual architectures or sizes. While our findings on final-layer specialization may extend to models like BLOOM (Workshop et al., 2023) and newer architectures with similar late-layer concentration (Tang et al., 2024; Mondal et al., 2025), the reduced English pivot behavior appears more specific to balanced multilingual training. Model English-centricity varies with training data, and recent work shows many multilingual models still rely on English-proximal representation spaces regardless of input/output languages (Schut et al., 2025).

Our focus is primarily on high-resource languages, with limited analysis of low-resource language processing. Recent work suggests that low-resource languages are harder to control via neuron manipulation, likely due to weaker or less distinct representations from limited pretraining exposure (Gurgurov et al., 2025a), indicating our findings may not directly extend to medium- and low-resource languages.

Our findings on Kojima et al. (2024)'s approach reveal a notable discrepancy. While they observed language-specific neurons in both early and late layers of decoder-only models, our analysis of Aya-23-8B on code-mixed input shows such neurons concentrated mainly in the final layers (27–31), peaking at layer 31. This likely reflects that we are identifying "code-mixing neurons" rather than pure language neurons, as our task distinguishes code-mixed from non-code-mixed inputs. These results suggest that code-mixing neurons align with language neurons in early layers but diverge significantly in later layers.

Thus, for hypothesis H4, we can only conclude that code-mixed inputs are processed differently in the model's very late layers. Similarly, our findings from the Tan et al. experiment show language-pair-specific processing across all layers but do not reveal clear patterns by language family or script, offering limited support for hypothesis H3.

Ethics Statement

We identify no ethical concerns directly related to this research. All models and datasets used in this study are employed in accordance with their respective license terms, including the custom use license for Llama 3.1-8B, the Apache 2.0 license for Aya-23-8B, and the research-permitted use of Chinese-LLaMA-2-7B. The Dumas dataset and WMT24++ corpus are used under their standard research licenses. Our code-mixed dataset, created through rule-based translation, contains no sensitive personal information and will be made publicly available to support reproducible research. The neuron-level analysis conducted in this work focuses purely on model internals without generating potentially harmful content or reinforcing linguistic biases.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023a. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv pre-print*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023b. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Boyi Deng, Yu Wan, Yidan Zhang, Baosong Yang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. *arXiv preprint arXiv:2505.05111*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus

Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. *arXiv preprint arXiv:2411.08745*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang,

Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Sunil Gundapu and Radhika Mamidi. 2020. Word level language identification in english telugu code mixed data. *arXiv preprint arXiv:2010.04482*.

Ayushman Gupta, Akhil Bhogal, and Kripabandhu Ghosh. 2024. Code-mixer ya nahi: Novel approaches to measuring multilingual llms' code-mixing capabilities. *arXiv preprint arXiv:2410.11079*.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025a. Language arithmetics: Towards systematic language neuron identification and manipulation.

Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, and Simon Ostermann. 2025b. Small models, big impact: Efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Sun Junyi. 2012. jieba: Chinese text segmentation. `https://github.com/fxsjy/jieba`. Accessed: 2025-05-17.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. *arXiv preprint arXiv:2503.17456*.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. *LessWrong*.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2025. Romanlens: The role of latent romanization in multilinguality in llms.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. *arXiv preprint arXiv:2404.11201*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

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

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schütze. 2025. Language mixing in reasoning language models: Patterns, impact, and internal causes. *arXiv preprint arXiv:2505.14815*.

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.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela,

Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Katherine Xie, Nitya Babbar, Vicky Chen, and Yoanna Turura. 2025. Enhancing multilingual language models for code-switched input data. *arXiv preprint arXiv:2503.07990*.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024a. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.

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