Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models

Anonymous ACL submission

Abstract

Language confusion—where large language models (LLMs) generate unintended languages against the user's need—remains a critical challenge, especially for English-centric models. We present the first mechanistic interpretability (MI) study of language confusion, combining behavioral benchmarking with neuron-level analysis. Using the Language Confusion Benchmark (LCB), we show that confusion points (CPs)—specific positions where language switches occur—are central to this phenomenon. Through layer-wise analysis with TunedLens and targeted neuron attribution, we reveal that transition failures in the final layers drive confusion. We further demonstrate that editing a small set of critical neurons, identified via comparative analysis with multilingual-tuned models, substantially mitigates confusion without harming general competence or fluency. Our approach matches multilingual alignment in confusion reduction for most languages and yields cleaner, higher-quality outputs. These findings provide new insights into the internal dynamics of LLMs and highlight neuron-level interventions as a promising direction for robust, interpretable multilingual language modeling. Code and data will be released upon publication.

1 Introduction

Current Large Language Models (LLMs), such as GPT-4 (Achiam et al., 2023), PaLM 2 (Anil et al., 2023), and Llama 3 (Grattafiori et al., 2024), have demonstrated exceptional linguistic competence across a wide range of complex tasks that require abstract knowledge and reasoning (Dong et al., 2024; Wei et al., 2022). Early LLMs were predominantly trained on massive amounts of English text data, with some limited exposure to other languages, resulting in initially constrained multilingual capabilities (Touvron et al., 2023). Recent advances, such as multilingual continued pretraining and instruction tuning, have substantially extended these models' ability to support multiple languages (Zhu et al., 2023; Shaham et al., 2024; Kew et al., 2024; Wang et al., 2025b). As a result, contemporary English-centric LLMs have become foundational tools for multilingual communication, multilingual content generation, and cross-lingual applications (Bang et al., 2023; Ahuja et al., 2023; Asai et al., 2024). However, despite their impressive capabilities, a persistent and underexplored limitation remains: LLMs can fail to

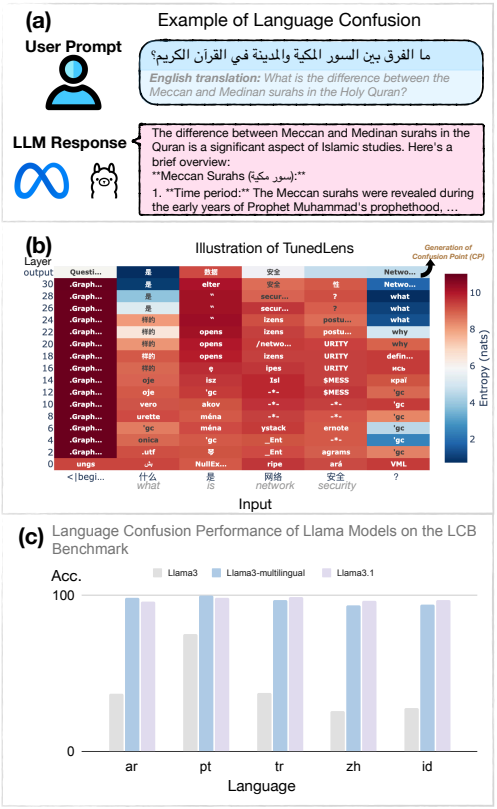

Figure 1: Language Confusion in LLMs. (a) An example of the language confusion phenomenon. (b) Visualization of internal model dynamics using TunedLens, highlighting how the confusion point emerges during generation. (c) Benchmarking results of three Llama models on the LCB benchmark across 5 languages.

generate text in the user's intended language, even when explicitly instructed—a phenomenon termed language confusion (Marchisio et al., 2024). Language confusion manifests as full-response, line-level, or word-level switches into unintended languages, severely undermining user experience and model reliability, especially for non-English speakers (Figure 1a).

Recent work by Marchisio et al. (2024) provides the first systematic characterization of language confusion, introducing the Language Confusion Benchmark (LCB) and associated metrics to quantify this phenomenon across a diverse set of languages and models. Their evaluation revealed that even state-of-the-art LLMs are susceptible to language confusion, with English-centric LLMs such as Llama2, Llama3, and Mistral exhibiting particularly high rates of unintended language switching, especially in the absence of targeted multilingual alignment (Figure 1c). While Marchisio et al. (2024) propose several mitigation strategies, including decoding adjustments, prompting techniques, and multilingual fine-tuning, these approaches remain largely surface-level, offering limited insight into the internal mechanisms that give rise to language confusion.

A key observation from prior work is the identification of confusion points—specific positions in the generation process where the model abruptly switches to an unintended language. However, the model's internal dynamics leading to these confusion points and their causal role in language confusion remain largely unexplored. This gap is particularly salient given the parallels to human bilingual code-switching, where switch points between languages are cognitively significant as extensively studied in psycholinguistics (Solorio and Liu, 2008; Bullock and Toribio, 2009). Further discussions are provided in Appendix A.

In this work, we move beyond behavioral evaluation to open the black box of LLMs, leveraging mechanistic interpretability (MI) methods (Conmy et al., 2023; Rai et al., 2024; Saphra and Wiegreffe, 2024; Sharkey et al., 2025) to investigate the internal representations and neuron-level processes underlying language confusion. We first empirically demonstrate that confusion points are critical drivers of language confusion: targeted interventions at these points can substantially reduce confusion across languages. Building on this, we employ MI tools such as *TunedLens* (Belrose et al., 2023) to trace the evolution of language representations through the model's layers, revealing that confusion typically arises from transition failures in the final layers, where latent conceptual representations are mapped to surface forms in the target language (Figure 1b). To further elucidate the mechanism, we conduct a neuron-level analysis, identifying specific neurons in the last layers whose activity is predictive of successful or failed language transitions at confusion points. Inspired by recent advances in neuron attribution and editing, we show that targeted manipulation of only 100 neurons can mitigate language confusion, offering a novel, model-internal approach to improving multilingual reliability. Our findings provide the first mechanistic account of language confusion in LLMs, bridging the gap between behavioral benchmarks and internal model dynamics. By highlighting the central role of confusion points and their neural substrates, we lay the groundwork for more robust, interpretable, and cognitively informed multilingual language models.

Our work makes the following contributions: (1) We provide the first mechanistic interpretability study of language confusion in English-centric LLMs, revealing the central role of confusion points in unintended language switching; (2) We employ layer-wise and neuron-level analyses to trace the internal dynamics leading to language confusion and identify critical late-layer neurons responsible for transition failures; (3) We propose and validate a principled neuron selection and editing strategy that effectively mitigates language confusion and preserves the model's general competence and output quality.

2 Related Work

Mechanistic Interpretability Methods Mechanistic interpretability (MI) seeks to reverse-engineer neural networks by decomposing their computations into human-understandable components (Stolfo et al., 2023; Wang et al., 2024; Men et al., 2024). A central technique in MI is the projection of intermediate representations into the vocabulary space, as implemented by tools such as LogitLens (Nostalgebraist, 2020) and TunedLens (Belrose et al., 2023), which enable researchers to track how information and predictions evolve across layers (Dar et al., 2023; Pal et al., 2023). In addition to layer-wise analysis, recent work has focused on identifying, attributing, and intervening on important neurons—those

Dataset	Data Source	Language	Prompt Example
Aya (Singh et al., 2024)	Human-generated	ar, en, pt, tr, zh	请简单介绍诗人李白的背景。 *Briefly introduce the poet Li Bai.*
Dolly (Singh et al., 2024)	MT post-edited	ar, es, fr, hi, ru	Qu'est-ce qui est plus important, l'inné ou l'acquis? *What is more important, nature or nurture?*
Native (Marchisio et al., 2024)	Human-generated	es, fr, ja, ko	콘크리트는 뭘로 만든거야? *What is concrete made of?*
Okapi (Lai et al., 2023)	Synthetic + MT	ar, en, pt, zh,it, fr, de, id, es, vi	Schreib einen Aufsatz von 500 Wörtern zum Thema KI. *Write a 500-word essay on AI.*

Table 1: Overview and Prompt Example of the LCB Benchmark (monolingual part). The number of examples per language is 100 in each dataset.

whose activations are strongly correlated with specific linguistic functions or behaviors (Bau et al., 2020; Geva et al., 2022; Yu and Ananiadou, 2024b). Methods for neuron selection and editing, as well as circuit-level analysis (Elhage et al., 2021; Wang et al., 2023), have proven effective for uncovering the internal structure underlying phenomena such as factual recall (Meng et al., 2022; Geva et al., 2023), reasoning processing (Yu and Ananiadou, 2024a), and now, as in our work, language confusion. By leveraging these MI techniques, we aim to provide a granular, causal understanding of how and why language confusion arises in multilingual LLMs, and to identify actionable intervention points for mitigation.

Multilingual Interpretability Recent research has begun to probe the internal representations of English-centric and multilingual LLMs to understand how they process and transfer information across languages (He et al., 2024; Zhao et al., 2024). Wendler et al. (2024) show that models like Llama2 often rely on English as an internal pivot language and can disentangle language and conceptual representations in controlled tasks. Fierro et al. (2025) examine how mechanisms identified in monolingual contexts generalize to multilingual settings. Wang et al. (2025a) investigate the internal causes of crosslingual factual inconsistencies, revealing how MLMs transition from language-independent to language-specific processing. However, prior work has not systematically connected these internal mechanisms to language generation errors such as language confusion.

3 Revisiting Language Confusion: Benchmark Insights

3.1 Recap of Language Confusion Benchmark

The Language Confusion Benchmark (LCB) (Marchisio et al., 2024) provides a systematic framework for evaluating the ability of LLMs to generate text in the user's intended language. The benchmark covers 15 typologically diverse languages and uses a diverse set of prompts sourced from human-written, post-edited, and synthetic datasets to evaluate models, ensuring coverage of a wide range of domains and linguistic structures (Table 1). In this work, we focus on the *monolingual setting* of LCB, where the prompt and expected response are in the same language. This setting is particularly relevant for mechanistic interpretability research, as it isolates language confusion phenomena from the additional complexities of explicit cross-lingual transfer.

To quantify language confusion, we adopt two key metrics from LCB: **line-level pass rate (LPR)** and **line-level language accuracy (Acc)**. LPR measures the percentage of model responses in which every line is in the correct language. Acc reflects the proportion of individual lines across all responses that are correctly generated in the target language. Both metrics rely on automatic language identification using the fastText classifier (Joulin et al., 2016, 2017), which efficiently detects the language of each line in the generated output.

We conducted preliminary benchmarking experiments on LCB with three instruction-tuned LLMs: ***Llama3-8B*** *(English-centric, no multilingual instruction tuning)*, ***Llama3-8B-multilingual*** *(multilingual instruction-tuned)* (Devine, 2024), and ***Llama3.1-8B*** *(multilingual-optimized)*. As shown in Figure 1c, *Llama3-8B* exhibits substantial language confusion, with frequent line-level switches to unintended languages (mostly English). In contrast, both *Llama3-8B-multilingual* and *Llama3.1-8B* achieve near-perfect LPR and line-level accuracy, demonstrating the effectiveness of multilingual instruction tuning and targeted optimization for multilingual dialogue.

Model	Metric	ar	en	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	avg
Llama3	*LPR*	33.0	99.5	71.0	33.0	19.3	73.0	59.3	8.0	28.0	14.0	23.0	19.0	22.0	34.0	11.0	**36.5**
(original)	*Acc*	33.7	99.8	74.5	37.5	23.4	77.1	64.1	15.1	28.2	17.1	23.6	23.0	27.3	39.8	14.8	**39.9**
Llama3	*LPR*	71.0	99.0	93.0	50.0	57.3	94.3	84.0	37.0	78.6	50.0	45.0	60.0	67.0	86.0	62.0	**68.9**
(replace)	Acc	74.8	99.6	95.4	55.5	64.1	95.3	86.5	47.6	83.1	55.3	48.6	62.3	77.7	87.5	66.1	**73.3**
Llama3	*LPR*	98.3	98.5	99.0	95.8	88.8	98.3	95.9	97.0	100.0	93.5	100.0	100.0	88.8	100.0	97.9	**96.8**
(multilingual)	*Acc*	98.7	99.5	99.8	96.9	93.8	99.3	96.9	97.5	100.0	95.8	100.0	100.0	94.2	100.0	97.9	**98.0**

Table 2: Impact of Confusion Point Replacement on Language Confusion Metrics. Line-level pass rate (LPR) and line-level accuracy for original Llama3-8B, multilingual Llama3-8B, and Llama3-8B with confusion point replacement, reported by language.

Given these findings, our work centers on understanding and mitigating the language confusion observed in English-centric *Llama3-8B*. By leveraging mechanistic interpretability methods, we aim to uncover the internal causes of confusion and develop interventions that can bring its performance closer to that of explicitly multilingual-tuned models. In the following subsection, we delve deeper into the significance of confusion points as critical junctures in the generation process.

3.2 Significance of Confusion Points

A confusion point (CP) is the position in a model's output where the first token of an unintended language abruptly appears, marking the onset of language confusion (Marchisio et al., 2024). This concept is inspired by psycho- and neurolinguistic research on code-switching, where the precise location of a language switch—known as a switch point—is central to understanding bilingual language production and processing (Blanco-Elorrieta and Pylkkänen, 2017; Suurmeijer et al., 2020). To empirically assess the role of CPs in LLM language confusion, we conduct a replacement experiment on *Llama3-8B*. For each instance of language confusion, we identify the CP using the fastText language detector. We then replace the token at the CP with the corresponding token generated by *Llama3-8B-multilingual*, which achieves near-perfect language accuracy, under the same prompt. This approach is motivated by the psycholinguistic observation that, in human code-switching, the choice at the switch point strongly influences the subsequent language trajectory (Moreno et al., 2002; Lai and O'Brien, 2020).

Our results, summarized in Table 2, show a substantial reduction in language confusion after CP replacement, even though our method does not represent an oracle upper bound. These findings highlight the centrality of confusion points in the emergence of language confusion and motivate our subsequent mechanistic analysis and targeted interventions.

4 Mechanistic Analysis of Language Confusion Points

4.1 Analyzing Layer-wise Language Transition

A central question in understanding language confusion is where and how the model's internal representations fail to transition from a shared conceptual space to the intended target language. Motivated by recent findings that English-centric LLMs process information in a latent, often English-biased, conceptual space before converting it to the target language in the final layers (Wendler et al., 2024; Wang et al., 2025a), we conduct a detailed layer-wise analysis of this transition using TunedLens (Belrose et al., 2023).

We employ TunedLens, the more reliable variant of LogitLens (Nostalgebraist, 2020), to unembed the hidden states of *Llama3-8B* at each layer into the vocabulary space. With this, we inspect every layer of the model and extract the top 10 predicted tokens with the largest logits at the position immediately preceding the confusion point (CP) (for confusion cases) or the output token (for correct cases). For each layer, we compute the average number and summed probabilities of English and target language tokens among the top-10 predictions, using fastText for language identification. Our analysis focuses on four typologically diverse languages (Arabic, Portuguese, Turkish, Chinese) from the LCB benchmark. We separate samples into two groups: (1) *Correct*—where the model generates the intended language throughout, and (2) *Confusion*—where the model switches to an unintended language at a CP. For confusion samples, we analyze the model's state up to the token before the CP.

Figure 2 presents the evolution of language token counts and probabilities across layers for both

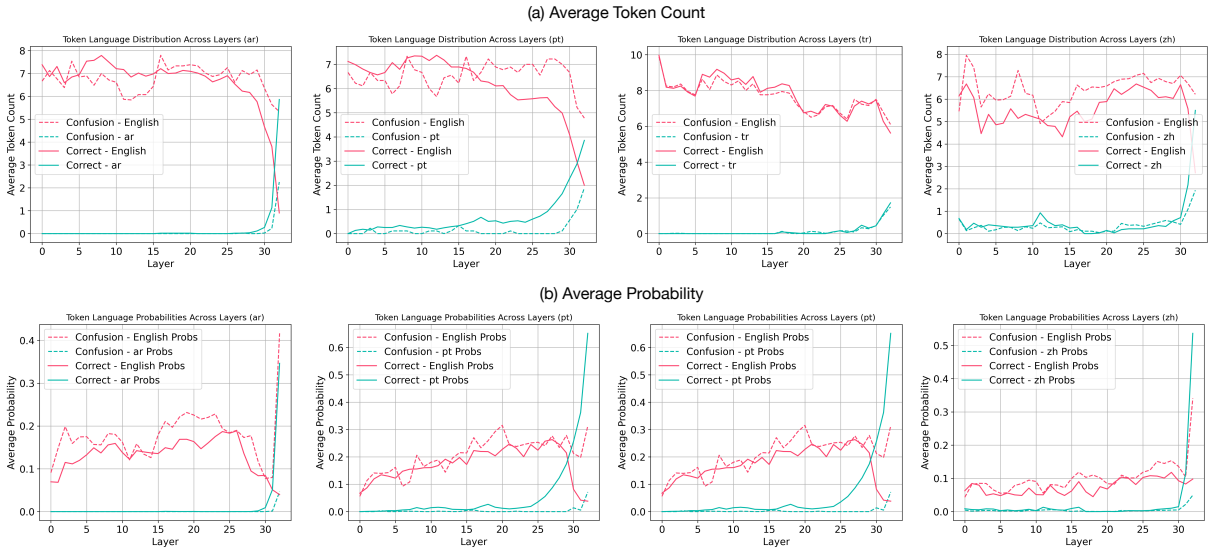

Figure 2: Average token counts and probabilities for English and target language tokens among the top-10 predictions at each layer, shown for both correct and confusion samples across four languages from *Aya*.

groups. In early and middle layers, English tokens dominate the top-10 predictions for all languages, reflecting the English-centric latent conceptual space of *Llama3-8B*. This is consistent with prior work showing that LLMs encode information in a shared, language-agnostic space in intermediate layers. In the final layers, a sharp transition emerges. For correct samples, the number and probability of target language tokens rise steeply, overtaking English tokens in the last few layers—indicating a successful transition to the target language surface form. In contrast, for confusion samples, this transition fails: English tokens remain dominant or even increase, while target language tokens lag behind. This failure to shift from the latent conceptual space to the target language at the critical moment leads to CPs and erroneous output.

Our layer-wise analysis with TunedLens reveals that the transition to the target language occurs in the final layers, and that failures in this process are tightly linked to language confusion. These findings provide direct evidence that language confusion in *Llama3-8B* is primarily caused by transition failures in the last few layers, motivating our subsequent neuron-level investigation to pinpoint and intervene on the specific components responsible for these failures.

4.2 Localizing Critical Neurons at Confusion Points

A key step toward understanding and mitigating language confusion is to identify which neurons are most responsible for the emergence of confusion points. Building on recent advances in neuron-level attribution (Geva et al., 2022; Yu and Ananiadou, 2024b), we adopt a static, efficient method to locate and analyze the most influential feed-forward network (FFN) neurons in *Llama3-8B*.

Methodology In the inference pass in decoder-only LLMs, for a given input sequence, each layer output h_i^l (layer l, token position i) is a sum of the previous layer's output h_i^{l-1}, the attention output A_i^l, and the FFN output F_i^l:

$$h_i^l = h_i^{l-1} + A_i^l + F_i^l \tag{1}$$

The FFN output F_i^l is calculated by a non-linear σ on two MLPs $W_{fc1}^l \in \mathbb{R}^{N\times d}$ and $W_{fc2}^l \in \mathbb{R}^{d\times N}$:

$$F_i^l = W_{fc2}^l \sigma(W_{fc1}^l(h_i^{l-1} + A_i^l)) \tag{2}$$

Following Geva et al. (2021), the FFN layer output F_i^l can be represented as a weighted sum over neuron subvalues:

$$F_i^l = \sum_{k=1}^{N} m_{i,k}^l \cdot fc2_k^l \tag{3}$$

$$m_{i,k}^l = \sigma(fc1_k^l \cdot (h_i^{l-1} + A_i^l)) \tag{4}$$

where $fc2_k^l$ is the k-th column of W_{fc2}^l, and $m_{i,k}^l$ is derived from the inner product between the residual output $(h_i^{l-1} + A_i^l)$ and $fc1_k^l$, the k-th row of W_{fc1}^l.

Geva et al. (2022) and Dar et al. (2023) project FFN neuron subvalues with unembedding matrices to compute the token probability distribution. To

quantify the importance of each neuron for generating a specific token (e.g., at a confusion point), we adopt the log probability increase method of Yu and Ananiadou (2024b). For a neuron in the l-th FFN layer v^l, its importance score is defined as the increase in log probability of the target token when v^l is added to the residual stream $A^l + h^{l-1}$, compared to the baseline without v^l:

$$Imp(v^l) = \log(p(w|v^l + A^l + h^{l-1}) - \log(p(w|A^l + h^{l-1}) \quad (5)$$

This approach efficiently identifies neurons whose activations most strongly influence the model's prediction at a given position.

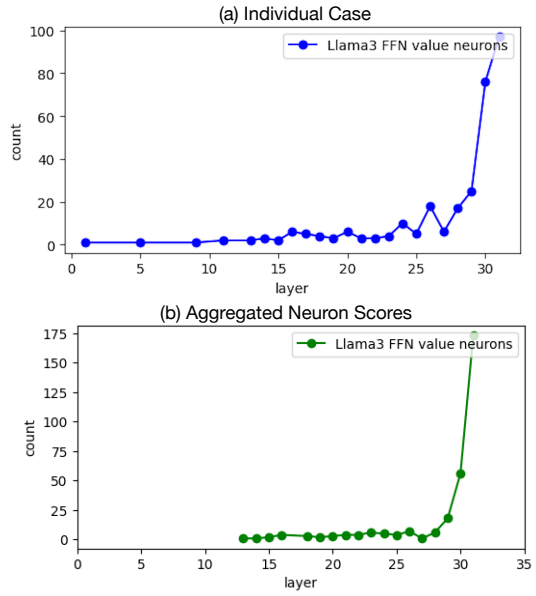

Figure 3: Distribution of Important Neurons Associated with Confusion Points in *Llama3-8B*. (a) Distribution of the top 300 most important FFN neurons across layers for an individual Chinese prompt "请解释拆东墙补西墙的意思。*(Please explain '拆东墙补西墙.')*" from Aya. (b) Aggregated distribution of important neuron scores across all Chinese test samples in Aya.

Experimental Observations We apply this method to *Llama3-8B* on confusion samples from the LCB benchmark, focusing on the token position immediately preceding each confusion point. For each sample and language, we compute the importance scores for all 14,336 FFN neurons in each layer of *Llama3-8B*, rank them, and select the top 300 most important neurons per sample. We then analyze the distribution of these critical neurons across layers, both for individual samples and aggregated over all samples in a language. Our analysis reveals a striking concentration of important neurons in the final layers, as visualized in Figure 3. This pattern holds both at the single-sample level and when aggregating across samples, indicating that the emergence of confusion points is primarily driven by late-layer FFN activity. We further rank neurons by their frequency of appearance in the top 300 sets across samples, finding that a subset of neurons consistently recurs as highly influential for confusion points.

To understand the effect of multilingual alignment, we repeat the analysis on *Llama3-8B-multilingual* using the same set of prompts. After multilingual instruction tuning, language confusion is nearly eliminated. Comparing neuron importance scores between the two models (Figure 4), we observe that most neurons critical for confusion in the *Llama3-8B* become much less important in its multilingual counterpart, suggesting that multilingual alignment suppresses the activity of confusion-inducing neurons. However, a small number of neurons remain important or even increase in importance, likely reflecting their role in encoding general semantic information rather than language-specific transitions.

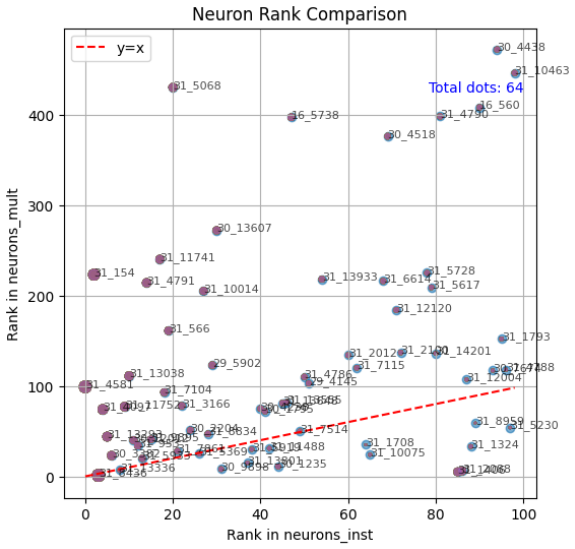

Figure 4: Neuron rank comparison between original Llama3 and multilingual Llama3. Results of Chinese test samples in Aya.

These findings reinforce the conclusion from our layer-wise analysis: language confusion is tightly linked to the activity of specific FFN neurons in the *final* layers. The suppression of these neurons through multilingual alignment provides a mechanistic explanation for the effectiveness of such tuning. Moreover, the identification of a small set

	ar	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	**Avg.**
original	33.44	74.26	37.55	24.04	77.15	63.16	16.47	28.20	17.44	23.50	23.00	27.33	39.83	14.79	**35.73**
freq	31.75	75.10	36.51	22.09	76.29	66.98	18.66	27.70	19.29	23.08	22.25	27.83	39.45	13.58	**35.75**
score	76.97	93.41	67.61	80.63	91.22	74.77	60.00	50.32	53.50	33.25	40.27	53.58	96.00	67.56	**67.08**
comparative	85.45	97.12	57.27	89.39	92.20	83.17	82.74	89.43	49.95	40.33	80.82	78.94	95.25	66.50	**77.75**

Table 3: Confusion mitigation performance of different selection strategies. Line-level accuracy is reported.

of persistent, semantically important neurons suggests that *targeted* neuron-level interventions could mitigate confusion without harming overall model performance. These insights directly inform our subsequent strategies for neuron-based mitigation of language confusion.

5 Mitigating Language Confusion via Neuron Editing

A central challenge in mitigating language confusion via neuron editing is to identify a set of neurons whose intervention effectively reduces confusion without degrading the model's general competence or fluency. Insights from our previous mechanistic analysis indicate that language confusion is primarily driven by a subset of late-layer FFN neurons. However, indiscriminate deactivation of important neurons risks harming the model's overall performance. Thus, a principled neuron selection strategy is essential.

	token_num	token_prob	fluency	acc_ood	xnli	senti
Original	1.96	24.5	25.8	39.9	46.4	98.4
Edited	3.43	36.8	21.8	74.25	44.9	98.2
Diff	1.47	12.3	-4.0	34.4	-1.5	-0.2

Table 4: Results of generalization and robustness of neuron editing. Average performance across languages is reported. Detailed results in Appendix B.

5.1 Neuron Selection and Intervention

We compare three neuron selection strategies: (1) *Frequency-Based Selection:* Selects the neurons most frequently identified as important across all confusion samples for a given language. (2) *Aggregate Importance Selection:* Ranks neurons by the sum of their importance scores across all confusion samples, selecting those with the highest cumulative influence. While this method captures the overall impact, it may still include neurons essential for general language competence. (3) *Comparative Importance Selection:* Inspired by Yu and Ananiadou (2024a), this strategy identifies neurons whose importance scores for confusion points decrease most substantially after multilingual alignment. Specifically, for each neuron, we compute the difference in importance score between original *Llama3-8B* and *Llama3-8B-multilingual* on the same input. Neurons with the largest drop are prioritized for intervention, as they are likely to be specifically implicated in language confusion rather than general semantic processing.

For each strategy, we select the top 100 neurons and intervene by setting their activations to zero during generation. We evaluate the impact of each method on the LCB benchmark. Our results (Table 4) demonstrate that Comparative Importance Selection achieves the most effective reduction in language confusion, substantially outperforming both frequency-based and aggregate importance methods. Frequency-based selection yields minimal benefit, while aggregate importance provides moderate improvement but still lags behind our proposed approach. Notably, the comparative strategy selectively targets neurons implicated in confusion, minimizing collateral impact on general model competence.

5.2 Generalization and Robustness of Neuron Editing

To further validate the effectiveness and safety of our Comparative Importance Selection strategy, we conduct a comprehensive evaluation across multiple metrics and experimental setups. Our goal is to ensure that neuron editing not only mitigates language confusion but also preserves the model's general competence, fluency, and robustness across domains (Table 4).

Language Confusion Mitigation We first assess the impact of neuron editing on language confusion using the LCB benchmark. In addition to standard metrics (line-level pass rate and line-level accuracy), we analyze the internal output distributions by reporting (1) the number of target language tokens among the top-10 candidates in the final output token logit, and (2) the total probability mass assigned to target language tokens in the top-10. These metrics provide a deeper view of how neuron editing shifts the model's internal preference toward the intended language, beyond surface-level

accuracy.

Robustness on General Tasks To evaluate whether neuron editing affects the model's general capabilities, we test the edited model on widely used multilingual benchmarks, including XNLI and multilingual sentiment analysis. We also assess output fluency by measuring the perplexity of generated responses using the multilingual model `facebook/xglm-564M` (Lin et al., 2022). Across all these metrics, the edited model maintains performance comparable to the original Llama3-8B, indicating that our intervention does not degrade general language understanding or generation quality.

Out-of-Domain Generalization We further examine the generalization of neuron editing by applying neurons selected from one data source (e.g., Aya) to out-of-domain test sets (e.g., Okapi) for the same language. The edited model continues to demonstrate strong mitigation of language confusion, suggesting that the identified neurons capture robust, domain-independent mechanisms underlying confusion points.

5.3 Comparison with Multilingual Alignment

To contextualize the effectiveness of neuron editing, we compare the performance of the edited *Llama3-8B* model with that of the multilingual-tuned *Llama3-8B*. Quantitative results show that neuron editing achieves language confusion mitigation on par with the multilingual-aligned model for most languages. Both approaches yield high line-level pass rates and language accuracy, demonstrating that targeted neuron intervention can match the benefits of extensive multilingual instruction tuning for confusion reduction.

Beyond aggregate metrics, qualitative analysis reveals further advantages of the neuron editing approach. Through case studies, we observe that the edited Llama3-8B not only generates fluent and accurate responses in the intended target language, but also avoids certain artifacts introduced by multilingual alignment. For example, in several instances (Figure 5), the multilingual Llama3-8B produces outputs containing HTML tags or formatting patterns reminiscent of its instruction tuning data, reflecting the influence of imperfect or noisy multilingual datasets. In contrast, the neuron-edited model consistently produces clean, well-structured, and contextually appropriate responses, free from such extraneous formatting.

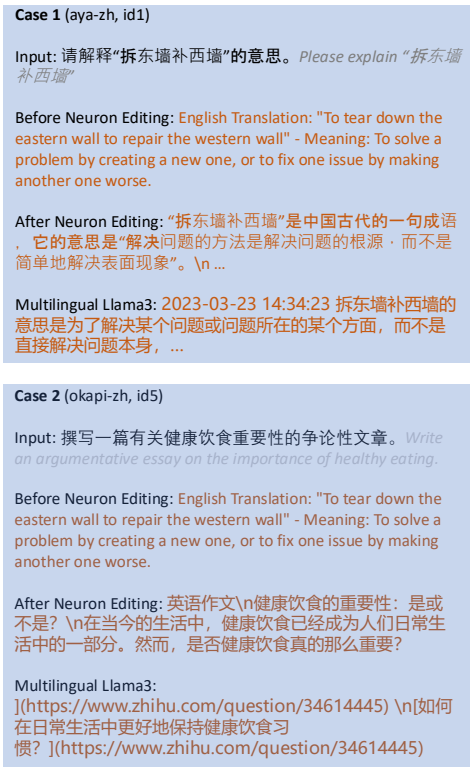

Figure 5: Case study of neuron editing.

These findings highlight a key strength of mechanistic neuron editing: it directly addresses the internal causes of language confusion without introducing side effects from large-scale data-driven alignment. By preserving the original model's semantic competence and output quality, neuron editing offers a more targeted and interpretable solution. This suggests that, beyond traditional multilingual instruction tuning, mechanistic interpretability-driven interventions can provide a promising path toward high-quality, robust multilingual language models.

6 Conclusions

This work provides the first mechanistic interpretability account of language confusion in English-centric LLMs. By tracing confusion points to failures in late-layer transitions and localizing the critical neurons responsible, we demonstrate that targeted neuron editing can robustly mitigate language confusion without sacrificing general competence or fluency. Our approach achieves results on par with multilingual-tuned models for most languages, while preserving cleaner output quality. These findings highlight the promise of neuron-level interventions for more reliable and interpretable multilingual language modeling.

Limitations

While this work provides the first mechanistic interpretability account of language confusion in English-centric LLMs, several limitations remain. Our analysis primarily focuses on the monolingual setting; cross-lingual contexts, which may involve distinct mechanisms and challenges, are left for future research. Additionally, neuron editing interventions are evaluated on selected benchmark tasks and may require further validation across broader domains and model architectures. Lastly, while our approach identifies and mitigates language confusion, fully understanding how these mechanisms interact with other multilingual phenomena warrants further investigation.

Ethic Statement

This research was conducted in accordance with the ACM Code of Ethics. The datasets that we use are publicly available. We have not intended or do not intend to share any Personally Identifiable Data with this paper. Regarding the usage of AI tools, we only use AI models for language refining.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, and 109 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Esti Blanco-Elorrieta and Liina Pylkkänen. 2017. Bilingual language switching in the laboratory versus in the wild: The spatiotemporal dynamics of adaptive language control. *Journal of Neuroscience*, 37(37):9022–9036.

Barbara E Bullock and Almeida Jacqueline Ed Toribio. 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge university press.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Key ingredients for effective zero-shot cross-lingual knowledge transfer in generative tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7222–7238, Mexico City, Mexico. Association for Computational Linguistics.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space.

In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada. Association for Computational Linguistics.

Peter Devine. 2024. Tagengo: A multilingual chat dataset. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 106–113, Miami, Florida, USA. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts? *arXiv preprint arXiv:2410.14387*.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge university press.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rosa E. Guzzardo Tamargo, Jorge R. Valdés Kroff, and Paola E. Dussias. 2016. Examining the relationship between comprehension and production processes in code-switched language. *Journal of Memory and Language*, 89:138–161. Speaking and Listening: Relationships Between Language Production and Comprehension.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. 2024. Large language models as neurolinguistic subjects: Identifying internal representations for form and meaning. *arXiv preprint arXiv:2411.07533*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA. Association for Computational Linguistics.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.

Gabrielle Lai and Beth A O'Brien. 2020. Examining language switching and cognitive control through the adaptive control hypothesis. *Frontiers in Psychology*, 11:1171.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12461–12476, Toronto, Canada. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7713–7724, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Eva M Moreno, Kara D Federmeier, and Marta Kutas. 2002. Switching languages, switching palabras (words): An electrophysiological study of code switching. *Brain and language*, 80(2):188–207.

Nostalgebraist. 2020. interpreting gpt: the logit lens.

Koyena Pal, Jiuding Sun, Andrew Yuan, Byron Wallace, and David Bau. 2023. Future lens: Anticipating subsequent tokens from a single hidden state. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 548–560, Singapore. Association for Computational Linguistics.

Jonas Pfeiffer, Francesco Piccinno, Massimo Nicosia, Xinyi Wang, Machel Reid, and Sebastian Ruder. 2023. mmT5: Modular multilingual pre-training solves source language hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1978–2008, Singapore. Association for Computational Linguistics.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian's, Malta. Association for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, and 10 others. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii. Association for Computational Linguistics.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Luuk Suurmeijer, M Carmen Parafita Couto, and Marianne Gullberg. 2020. Structural and extralinguistic aspects of code-switching: Evidence from papiamentu-dutch auditory sentence matching. *Frontiers in Psychology*, 11:592266.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jeanine Treffers-Daller. 2009. *Code-switching and transfer: an exploration of similarities and differences*. Cambridge University Press.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025a. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *arXiv preprint arXiv:2504.04264*.

Shumin Wang, Yuexiang Xie, Bolin Ding, Jinyang Gao, and Yanyong Zhang. 2025b. Language adaptation of large language models: An empirical study on LLaMA2. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7195–7208, Abu Dhabi, UAE. Association for Computational Linguistics.

Yifei Wang, Yuheng Chen, Wanting Wen, Yu Sheng, Linjing Li, and Daniel Dajun Zeng. 2024. Unveiling factual recall behaviors of large language models through knowledge neurons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7402, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Odilia Yim and Richard Clément. 2021. Acculturation and attitudes toward code-switching: A bidimensional framework. *International Journal of Bilingualism*, 25(5):1369–1388.

Zeping Yu and Sophia Ananiadou. 2024a. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3293–3306, Miami, Florida, USA. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2024b. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280, Miami, Florida, USA. Association for Computational Linguistics.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

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? *arXiv preprint arXiv:2408.10811*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Further Discussion on Code-Switching and Language Confusion

Code-switching as a Linguistic Phenomenon Code-switching, the practice of alternating between languages within a single conversation or utterance, is a well-studied natural phenomenon in bilingualism and psycholinguistics (Gardner-Chloros, 2009). Code-switching is typically intentional, often reflecting speakers' identities, social relationships, and contextual adaptation (Treffers-Daller, 2009; Yim and Clément, 2021). In NLP, code-switching has been explored through evaluating model performance on code-switched data for tasks such as sentiment analysis, machine translation, summarization, and language identification (Khanuja et al., 2020; Doğruöz et al., 2021; Winata et al., 2023). Code-switching is a natural, contextually appropriate strategy in human communication, whereas language confusion, on which our work focuses, is an unintended and erroneous switch to an incorrect language in LLMs (Marchisio et al., 2024). Though related to code-switching, language confusion is an unnatural phenomenon that arises from model failures rather than communicative intent.

Language Confusion and Confusion Points in LLMs Language confusion has been observed in various multilingual NLP settings, such as "source language hallucinations" in zero-shot cross-lingual transfer (Li and Murray, 2023; Pfeiffer et al., 2023; Chirkova and Nikoulina, 2024) and "off-target translation" in machine translation (Sennrich et al., 2024). In LLMs, this manifests as abrupt, unexpected switches to the wrong language during generation, even under explicit instructions. This issue is particularly prevalent in English-centric models lacking robust multilingual alignment (Zhong et al., 2024). A key concept in recent work is the *confusion point*—the specific position in generation where the model transitions to an unintended language. Inspired by the importance of code-switching points in human bilingualism, confusion points are central to understanding and diagnosing language confusion in LLMs (Guzzardo Tamargo et al., 2016). Unlike natural code-switching, these points reflect internal model failures. Recent benchmarks (Marchisio et al., 2024) systematically characterize confusion points at response, line, and word levels, revealing their widespread impact and motivating deeper mechanistic investigation, as pursued in this work.

B Full Experimental Results

Table 5 presents the full benchmarking results. Table 6 shows the full results of the CP replacement experiment. Tables 7 and 8 present the full results of robustness and generalization experiments.

C Detailed Experimental Setup

C.1 Models

We primarily use three variants of the Llama3 family for our experiments:

- **Llama3-8B**: The baseline English-centric model without multilingual instruction tuning.
- **Llama3-8B-multilingual**: The multilingual instruction-tuned version, as described in (Devine, 2024).
- **Llama3.1-8B**: An improved model optimized for multilingual dialogue.

All models are used in their publicly released forms unless otherwise stated. For neuron editing experiments, we intervene on *Llama3-8B* using the strategies described in Section 5.

C.2 Datasets and Tasks

Language Confusion Benchmarking and Replacement Experiments We use the Language Confusion Benchmark (LCB) (Marchisio et al., 2024) for all language confusion detection and mitigation experiments. LCB covers 15 typologically diverse languages and comprises several monolingual and cross-lingual datasets:

- **Monolingual sources**: Aya (human-generated), Dolly (post-edited), Native (human-generated), and Okapi (synthetic + machine translated).
- **Languages**: Arabic, English, Portuguese, Turkish, Chinese, Spanish, French, Hindi, Russian, Japanese, Korean, German, Indonesian, Italian, Vietnamese.

All main benchmarking and confusion point replacement experiments are run on the monolingual portions of LCB, using 100 prompts per language per dataset as described in Table 1.

Robustness and Generalization Experiments To assess the robustness and generalization of neuron editing, we evaluate on:

- **XNLI** (Conneau et al., 2018): Cross-lingual natural language inference in 15 languages.
- **Multilingual Sentiment Analysis**: Standard multilingual sentiment datasets (including German, Spanish, French, Japanese, and Chinese). It is a binary classification task derived from the multilingual Amazon review dataset.
- **Out-of-domain LCB evaluation**: For each language, neurons are selected from one LCB source (e.g., Aya), then tested on a different source (e.g., Okapi) to assess generalization.

C.3 Metrics

Language Confusion Metrics We adopt two primary metrics from LCB:

- **Line-level Pass Rate (LPR)**: Percentage of responses where every line is in the correct language.
- **Line-level Accuracy**: Proportion of lines generated in the correct language.

Language identification for these metrics is performed using the fastText classifier (Joulin et al., 2016).

Internal Model Metrics We further report:

- **Target Language Token Count**: Number of target language tokens among the top-10 output logits in the final layer.
- **Target Language Token Probability**: Total probability mass assigned to target language tokens in the top-10 output logits.

Generalization and Fluency Metrics

- **XNLI and Sentiment Accuracy**: Standard classification accuracy on XNLI and multilingual sentiment analysis tasks.
- **Fluency (Perplexity)**: Perplexity of generated outputs, measured using the multilingual `facebook/xglm-564M` model (Lin et al., 2022).

C.4 Implementation Details

All experiments are run on NVIDIA A100 GPUs. Prompt formatting and decoding settings follow the LCB benchmark defaults. Neuron interventions are implemented at inference time via custom hooks in PyTorch, zeroing out selected neuron activations layer-wise as described in Section 5.1. For TunedLens analysis, we use the public implementation from Belrose et al. (2023).

All code, evaluation scripts, and neuron selection details will be released upon publication to facilitate reproducibility.

metrics: acc
Monolingual

	source	ar	en	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	avg
Llama3	**aya**	55.55	100.00	86.90	37.69	42.23	-	-	-	-	-	-	-	-	-	-	64.47
	dolly	33.00	-	-	-	-	75.77	60.49	19.05	34.45	-	-	-	-	-	-	44.55
	native	-	-	-	-	-	91.47	79.17	-	-	18.05	25.92	-	-	-	-	53.65
	okapi	22.00	99.67	63.12	-	9.08	67.75	55.03	-	-	-	-	25.25	27.83	39.83	15.41	42.50
	avg	36.85	99.83	75.01	37.69	25.65	78.33	64.90	19.05	34.45	18.05	25.92	25.25	27.83	39.83	15.41	**41.60**
Llama3-multilingual	**aya**	98	98.93	99.83	96.93	92.35	-	-	-	-	-	-	-	-	-	-	97.21
	dolly	98.99	-	-	-	-	98.15	93.03	97.50	100.0(	-	-	-	-	-	-	97.53
	native	-	-	-	-	-	99.75	97.87	-	-	95.83	100.0(	-	-	-	-	98.36
	okapi	98.97	100.00	99.83	-	95.20	100.0(	99.80	-	-	-	-	100.0(	94.23	100.0(	97.87	98.65
	avg	98.65	99.47	99.83	96.93	93.78	99.30	96.90	97.50	100.0(	95.83	100.0(	100.0(	94.23	100.0(	97.87	**98.02**
Llama3.1	**aya**	93.35	99.50	97.82	98.98	96.21	-	-	-	-	-	-	-	-	-	-	97.17
	dolly	97.94	-	-	-	-	98.00	97.84	99.50	98.99	-	-	-	-	-	-	98.45
	native	-	-	-	-	-	98.8	99.75	-	-	97.82	100	-	-	-	-	99.09
	okapi	97.31	100.00	99.50	-	97.28	100.0(	100.0(	-	-	-	-	100.0(	97.08	100.0(	99.67	99.08
	avg	96.20	99.75	98.66	98.98	96.75	98.93	99.20	99.50	98.99	97.82	100.0(	100.0(	97.08	100.0(	99.67	**98.77**

Table 5: Full benchmarking results on LCB.

metrics: lpr
Monolingual

	source	ar	en	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	avg
Llama3-ori	**aya**	53	100	83	33	31.63	-	-	-	-	-	-	-	-	-	-	64.47
	dolly	30	-	-	-	-	68	54	8	28	-	-	-	-	-	-	44.55
	native	-	-	-	-	-	88	72	-	-	14	23	-	-	-	-	53.65
	okapi	16	99	59	-	7	63	52	-	-	-	-	19	22	34	11	42.50
	avg	33.00	99.50	71.00	33.00	19.32	73.00	59.33	8.00	28.00	14.00	23.00	19.00	22.00	34.00	11.00	**36.48**
Llama3-re	**aya**	83.67	98	91	50	65.66	-	-	-	-	-	-	-	-	-	-	77.67
	dolly	65.66	-	-	-	-	94	76	37	78.57	-	-	-	-	-	-	70.25
	native	-	-	-	-	-	97	86	-	-	50	45	-	-	-	-	69.50
	okapi	63.54	100	95	-	49	92	90	-	-	-	-	60	67	86	62	76.17
	avg	70.96	99.00	93.00	50.00	57.33	94.33	84.00	37.00	78.57	50.00	45.00	60.00	67.00	86.00	62.00	**68.95**
Llama3-multi	**aya**	98	96.97	99	95.83	84.69	-	-	-	-	-	-	-	-	-	-	97.17
	dolly	97.98	-	-	-	-	95.96	91.84	97	100	-	-	-	-	-	-	98.45
	native	-	-	-	-	-	99	96.81	-	-	93.48	100	-	-	-	-	99.09
	okapi	98.97	100	99	-	92.93	100	99	-	-	-	-	100	88.78	100	97.87	99.08
	avg	98.32	98.49	99.00	95.83	88.81	98.32	95.88	97.00	100.0(	93.48	100.0(	100.0(	88.78	100.0(	97.87	**96.79**

metrics: acc
Monolingual

	source	ar	en	pt	tr	zh	es	fr	hi	ru	ja	ko	de	id	it	vi	avg
Llama3-ori	**aya**	53.75	100	86.4	37.5	39.46	-	-	-	-	-	-	-	-	-	-	64.47
	dolly	30.75	-	-	-	-	73.45	59.99	15.05	28.2	-	-	-	-	-	-	44.55
	native	-	-	-	-	-	91.05	77.75	-	-	17.13	23.58	-	-	-	-	53.65
	okapi	16.5	99.67	62.62	-	7.33	66.83	54.7	-	-	-	-	23	27.33	39.83	14.79	42.50
	avg	33.67	99.84	74.51	37.50	23.40	77.11	64.15	15.05	28.20	17.13	23.58	23.00	27.33	39.83	14.79	**39.94**
Llama3-re	**aya**	86.9	99.17	94.97	55.53	71.12	-	-	-	-	-	-	-	-	-	-	81.54
	dolly	68.48	-	-	-	-	94.25	80.66	47.62	83.1	-	-	-	-	-	-	74.82
	native	-	-	-	-	-	97	87.92	-	-	55.27	48.58	-	-	-	-	72.19
	okapi	68.92	100	95.79	-	57.13	94.67	91	-	-	-	-	62.33	77.67	87.5	66.08	79.88
	avg	74.77	99.59	95.38	55.53	64.13	95.31	86.53	47.62	83.10	55.27	48.58	62.33	77.67	87.50	66.08	**73.29**
Llama3-multi	**aya**	98	98.93	99.83	96.93	92.35	-	-	-	-	-	-	-	-	-	-	97.17
	dolly	98.99	-	-	-	-	98.15	93.03	97.5	100	-	-	-	-	-	-	98.45
	native	-	-	-	-	-	99.75	97.87	-	-	95.83	100	-	-	-	-	99.09
	okapi	98.97	100	99.83	-	95.2	100	99.8	-	-	-	-	100	94.23	100	97.87	99.08
	avg	98.65	99.47	99.83	96.93	93.78	99.30	96.90	97.50	100.0(	95.83	100.0(	100.0(	94.23	100.0(	97.87	**98.02**

Table 6: Full results of CP replacement experiments

	num_ori	prob_ori	num_edit	prob_edit	num_diff	prob_diff	fluency_ori	fluency_cna	diff
ar	2.83	25.8	5.37	30.3	2.55	4.5	30.1	24.7	-5.4
pt	2.86	49.5	3.41	56.0	0.56	6.5	25.7	23.3	-2.3
tr	2.05	29.5	2.42	23.5	0.37	-6.0	21.2	18.8	-2.5
zh	1.33	8.6	5.10	37.3	3.78	28.7	33.1	26.0	-7.0
es	1.67	26.5	3.28	50.3	1.61	23.8	25.4	23.2	-2.2
fr	2.48	43.0	2.91	49.2	0.43	6.2	21.2	21.1	-0.1
hi	1.25	12.0	1.64	13.7	0.39	1.8	28.5	22.9	-5.6
ru	1.09	18.0	3.21	31.0	2.12	13.0	23.7	19.5	-4.2
de	2.73	23.7	4.45	37.1	1.72	13.4	23.8	18.5	-5.3
it	1.33	8.4	2.50	39.3	1.17	31.0	25.7	20.2	-5.5
avg	1.96	24.5	3.43	36.8	1.47	12.3	25.8	21.8	-4.0

Table 7: Full results of robustness experiments. Perplexity is calculated to measure fluency.

xnli		
language	acc_ori	acc_edit
ar	0.42	0.37
de	0.54	0.54
es	0.46	0.5
fr	0.49	0.5
hi	0.47	0.48
ru	0.37	0.3
tr	0.46	0.52
vi	0.46	0.37
zh	0.51	0.46
avg	0.464	0.449

sentiment analysis		
language	acc_ori	acc_edit
de	0.98	0.98
es	0.98	0.98
fr	0.98	0.97
ja	0.99	0.99
zh	0.99	0.99
avg	0.984	0.982

Table 8: Full results of generalization experiments.