Mechanistic Understanding and Mitigation of Language Confusion in English-centric LLMs

ERCONG NIE, PHD CANDIDATE
Center for Information and Language Processing,
Ludwig Maximilians University of Munich (LMU)
https://ercong21.github.io/

Jun. 24, 2025

About me

Ercong Nie [ɚ'tsʰʊŋ, niɛ] 聂耳聡 PhD Candidate

Schuetze NLP Lab, Center for Information and Language Processing (CIS),
Ludwig Maximilians University of Munich (LMU Munich),
Munich Center for Machine Learning (MCML)

- final-year PhD student at Schütze Lab, supervised by PD. Dr. Helmut Schmid
- **MSc.** in Computational Linguistics plus Informatics at **LMU Munich**, Germany
- **B.A.** in German and Finance at **Shanghai Jiao Tong University**, China
- Research interest includes: multilingual NLP, efficient NLP methods, and human-inspired NLP

https://ercong21.github.io/

Outline

1. Introduction & Motivation
2. Methodological Background
3. Language Confusion Points
4. Mechanistic Analysis of Language Confusion
5. Mitigating Language Confusion: Neuron Editing
6. Conclusions

Outline

What is Language Confusion?

Multilinguality of LLMs

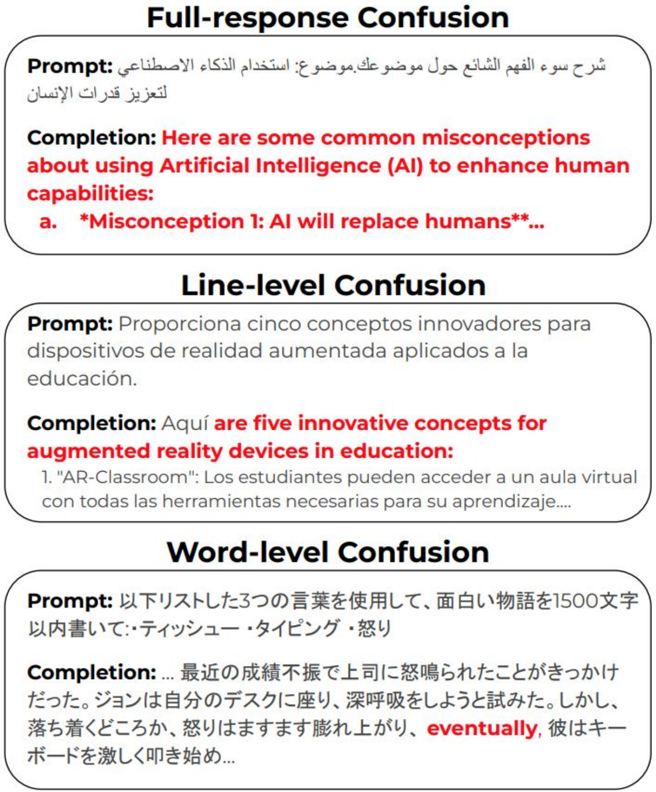
- LLMs like GPT-4, PaLM 2, and Llama 3 are now widely used for **multilingual applications**.
- Early LLMs were trained mainly on English (**English-centric LLMs**), with limited multilingual capability.
- Recent advances (multilingual pretraining, instruction tuning) have improved non-English support, but challenges remain, e.g. **language confusion**.

What is Language Confusion?

Language confusion is when an LLM generates text in an unintended language, failing to match the user's request.

Examples of different types of **language confusion:**

[(Marchisio et al., 2024)]



Language Confusion vs. Code-Switching

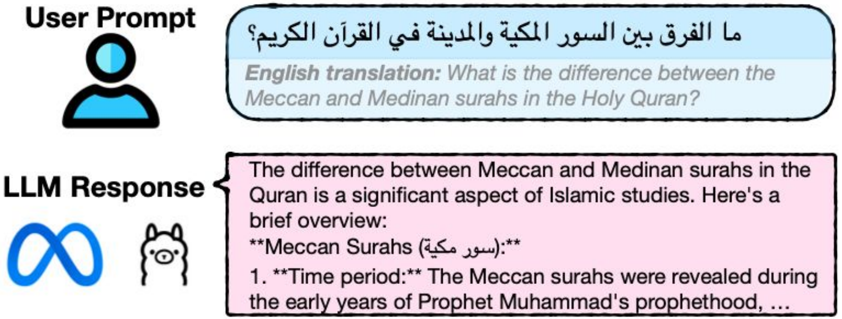
Language Confusion vs. Code-Switching

- **Code-switching:** Natural, intentional alternation between languages in human communication.
- **Language confusion:**
  - Unintentional, erroneous language switching by LLMs.Not contextually appropriate;
  - reflects model failure, not communicative intent.

Example of code-switching

Soy bueno *(Spanish for "I am good")*

Example of language confusion



Why is Language Confusion a Problem?

Why is language confusion a problem?
User experience: Breaking trust and usability for non-English speakers.

Prevalence:

- Especially common in English-centric LLMs (e.g., Llama3, Mistral).

Line-level pass rate (LPR) by language (Marchisio et al., 2024).

	avg	ar	de	en	es	fr	hi	id	it	ja	ko	pt	ru	tr	vi	zh
Llama 2 70B-I	48.3	0.3	59.0	99.0	95.7	87.7	1.0	62.0	72.0	7.0	0.0	91.0	88.9	33.0	17.0	10.5
Llama 3 70B-I	46.0	21.7	31.0	**100.0**	98.3	88.7	23.0	21.0	88.0	10.0	0.0	95.5	77.0	18.0	10.0	8.0
Llama 3.1 70B-I	99.0	98.9	**100.0**	98.5	99.0	**100.0**	**100.0**	94.0	**100.0**	96.9	**100.0**	**99.0**	**100.0**	**100.0**	**100.0**	**99.0**
Mixtral 8x7B	73.0	48.3	90.9	99.5	89.3	95.3	71.0	58.0	72.0	66.7	61.2	85.0	65.0	90.0	57.0	45.5
Mistral Large	69.9	48.0	98.0	99.0	99.0	**100.0**	19.0	31.0	99.0	48.0	64.0	79.5	98.0	71.0	29.0	66.0
Command R	98.6	**100.0**	98.0	99.5	95.7	99.3	**100.0**	92.0	99.0	**100.0**	**100.0**	98.5	**100.0**	99.0	99.0	98.5
Command R+	99.2	99.7	**100.0**	**100.0**	99.3	99.7	**100.0**	**97.0**	**100.0**	99.0	**100.0**	97.5	**100.0**	**100.0**	99.0	97.5
Command R Refresh	98.9	99.6	**100.0**	99.5	99.3	99.7	**100.0**	92.0	**100.0**	99.0	**100.0**	98.0	**100.0**	99.0	**100.0**	98.0
Command R+ Refresh	**99.3**	99.0	**100.0**	**100.0**	99.3	**100.0**	**100.0**	96.0	**100.0**	**100.0**	**100.0**	97.5	99.0	**100.0**	**100.0**	98.0
GPT-3.5 Turbo	99.1	**100.0**	**100.0**	99.5	**99.7**	**100.0**	99.0	96.0	**100.0**	98.0	**100.0**	98.0	**100.0**	**100.0**	99.0	97.0
GPT-4 Turbo	**99.3**	99.0	**100.0**	**100.0**	99.3	99.3	**100.0**	96.0	99.0	**100.0**	**100.0**	98.0	**100.0**	**100.0**	**100.0**	**99.0**
GPT-4o	98.9	99.7	**100.0**	**100.0**	99.3	99.3	99.0	94.0	**100.0**	99.0	**100.0**	97.5	99.0	**100.0**	99.0	98.0

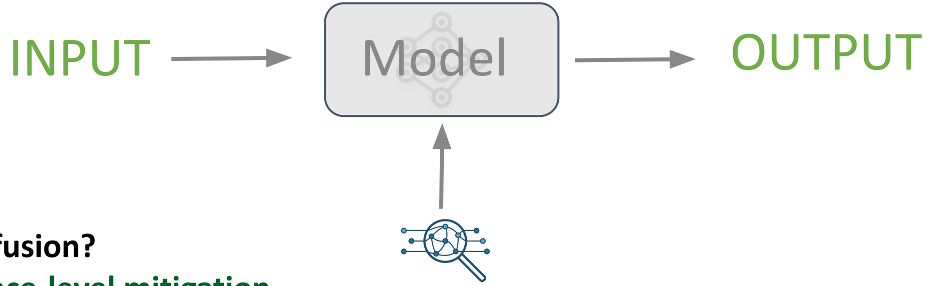
This work aims at mechanistically understanding and mitigating language confusion in English-centric LLMs.

Outline

1. Introduction & Motivation
2. Methodological Background
3. Language Confusion Points
4. Mechanistic Analysis of Language Confusion
5. Mitigating Language Confusion: Neuron Editing
6. Conclusions

Mechanistic Interpretability (MI) for Language Confusion

Mechanistic interpretability (MI)
aims to reverse-engineer neural networks by decomposing their computations into human-understandable components, and helps understand how and why specific behaviors (like language confusion) arise inside LLMs.



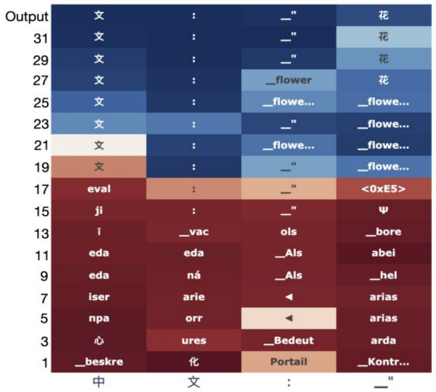
Why MI for language confusion?

- **Limitations of surface-level mitigation**
  - Marchisio et al. (2024) explore *few-shot prompting*, *multilingual fine-tuning*, *decoding strategies* to reduce confusion, but do not explain why it happens.
  - These methods treat symptoms, not causes.
- **Need for causal, internal understanding**
  - To robustly mitigate confusion, we must **identify the internal mechanisms**—where and how the model fails to transition to the intended language.
  - MI provides tools to pinpoint **failure points** and actionable **intervention targets** inside the model.

Key Techniques

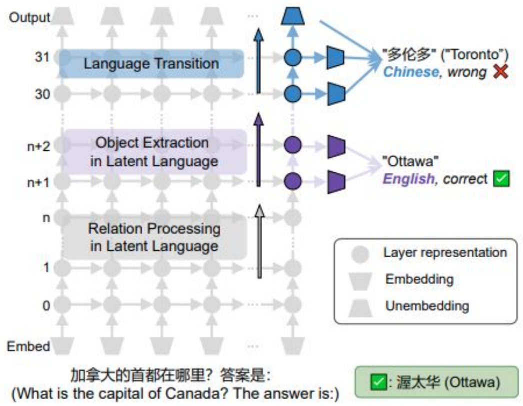
Layer-wise analysis
(e.g., LogitLens, TunedLens)

Do Llamas work in English?



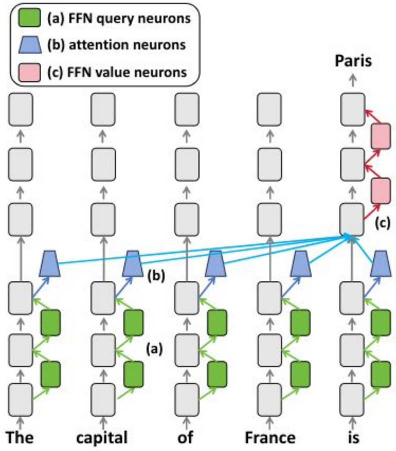
[(Wendler et al.,2024)]()

Employing LogitLens to dissect cross-lingual factual Inconsistency in English-centric and multilingual LLMs:



[(Wang et al.,2025)]()

Neuron-level attribution and intervention



Causal tracing:

Identify important neurons based on their impact on the output probabilities

[Geva et al. (2023)](), [Yu and Ananiadou (2024)]()

Layer-Wise Analysis - Tracing Language Transitions

Approach:

- Use tools like *TunedLens* to project hidden states at each layer into the vocabulary space.
- Trace how the model's predictions evolve from input to output.

Findings from prior work on multilingual interpretability:

- English-centric LLMs process information in a latent, often *English-biased*, conceptual space in **early/mid** layers.
- Successful generation requires a sharp transition to the target language in the final layers.

Connecting to language confusion:

- Layer-wise tracing helps reveal when and how the model transitions from an English-centric latent space to the target language.
- Failures or delays in this cross-lingual shift may underlie language confusion.
- This analysis can localize internal bottlenecks where unintended language switches occur, guiding deeper mechanistic exploration in later sections.

Neuro-Level Attribution - Identifying Critical Neurons

Motivation:

which individual neurons are responsible for a specific model behavior (e.g., language transitions or language confusion)?

Methods (Geva et al., 2023; Yu and Ananiadou, 2024)

- **Neuron Attribution:**
  - Quantify each neuron's influence on the probability of generating a specific token.
  - Log-probability increase method: How much does activating a neuron increase the likelihood of the correct token?
- **Neuron Editing:**
  - Intervene by modifying or zeroing out activations of critical neurons to test causal effects on model behavior.

Neuro-Level Attribution - Identifying Critical Neurons

Quantifying neuron importance score for an inference pass from inputs to the final predictions

Given an input sentence, 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 feed-forward network (FFN) output F_i^l:

$$h_i^l = h_i^{l-1} + A_i^l + F_i^l \qquad (1)$$

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 $W_{fc2}^l \in \mathbb{R}^{d \times N}$:

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

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

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

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

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.

To quantify the importance of each neuron for generating a specific token, we adopt the log probability increase method. 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 $A_i^l + h_i^{l-1}$, compared to the baseline without v^l:

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

Outline

1. Introduction & Motivation
2. Methodological Background
3. Language Confusion Points
4. Mechanistic Analysis of Language Confusion
5. Mitigating Language Confusion: Neuron Editing
6. Conclusions

The Language Confusion Benchmark (LCB)

The Language Confusion Benchmark (LCB) (Marchisio et al., 2024)

- **Purpose:** Systematically evaluate LLMs' ability to generate text in the intended language.
- **Coverage:** 15 typologically diverse languages, 4 dataset sources (human-written, post-edited, synthetic).
- **Metrics:**
  - **Line-level Pass Rate (LPR):** % of responses with all lines in the correct language.
  - **Line-level Accuracy:** % of lines in the correct language.

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.*

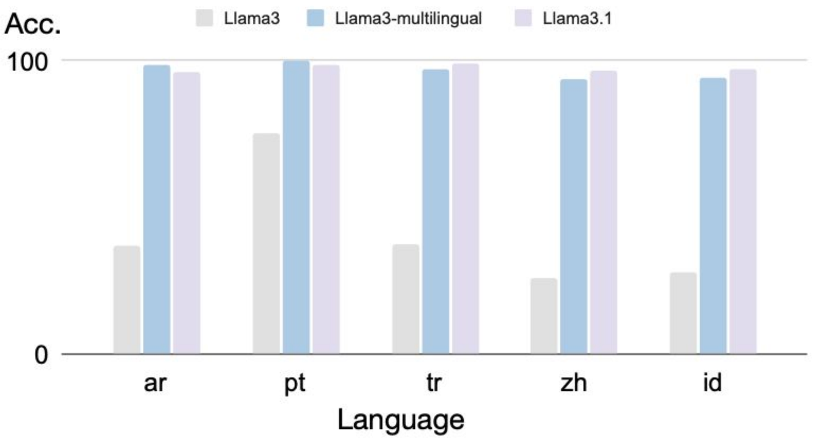
Preliminary Benchmarking Results

Language confusion performance of Llama models on the LCB benchmark

- **Models evaluated*:**
  - Llama3-8B (*English-centric*): Pretrained on multilingual datasets with English as the dominant language
  - Llama3-8B-multilingual (*multilingual-tuned*): Multilingual instruction tuning
  - Llama3.1-8B (*multilingual-optimized*): Multilingual post-training (SFT, preference alignment)

- **Findings:**
  - English-centric Llama3-8B shows frequent unintended language switches, especially to English.
  - Multilingual-tuned models achieve near-perfect LPR and accuracy across languages.

Language Confusion Performance of Llama Models on the LCB Benchmark



* All models used in this work are instruction-tuned versions.

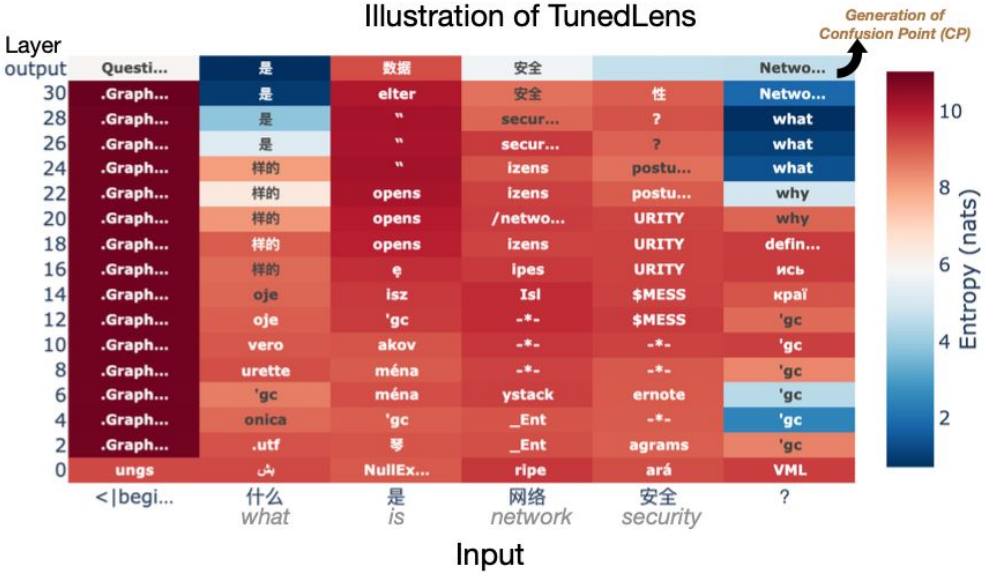
The Role of Confusion Points

Confusion Point (CP)

- **What is a Confusion Point (CP)?**
  - The specific position in the output where the model first switches to an unintended language.
  - Inspired by "**switch points**" in human code-switching, but here reflects model failure, not intent.

- **Significance:**
  - CPs mark the onset of language confusion and are central to understanding and mitigating the phenomenon.

Generation process of a confusion point via TunedLens



Empirical Evidence – CP Replacement Experiment

Experiment:

- For each confusion case, identify the CP and replace the token at that position with the corresponding token from the multilingual-tuned model.

Results:

- Substantial reduction in language confusion after CP replacement.
- LPR and accuracy improve dramatically, approaching multilingual-tuned model performance.

Interpretation:

- Confusion points are critical drivers of language confusion; intervening at these points can effectively restore correct language generation.

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**

Impact of confusion point replacement on language confusion metrics

Understanding language confusion → Understanding how CPs arise (Tracing internal model dynamics at CPs)

Mitigating language confusion → Suppressing the generation of CPs (Identifying and editing critical neurons responsible for CPs)

Outline

Layer-Wise Language Transition Analysis

Implementation of TunedLens

- We group all prompts into *“correct”* cases and *“confusion”* cases.
- For each prompt, we use **TunedLens** to extract the top-10 predicted tokens (by logit score) at every layer, focusing on the position immediately before the confusion point (CP) for confusion cases, or the output token for correct cases.
- Each of these top-10 tokens is classified as either English or the target language using **fastText**, allowing us to track the model’s internal language preference at every layer.
- We analyze both correct and confusion cases across diverse languages (e.g., Arabic, Portuguese, Turkish, Chinese).

Key Metrics:

- **Token Count:** Number of English vs. target language tokens among the top-10 predictions at each layer.
- **Token Probability:** Total probability mass assigned to English vs. target language tokens in the top-10 at each layer.

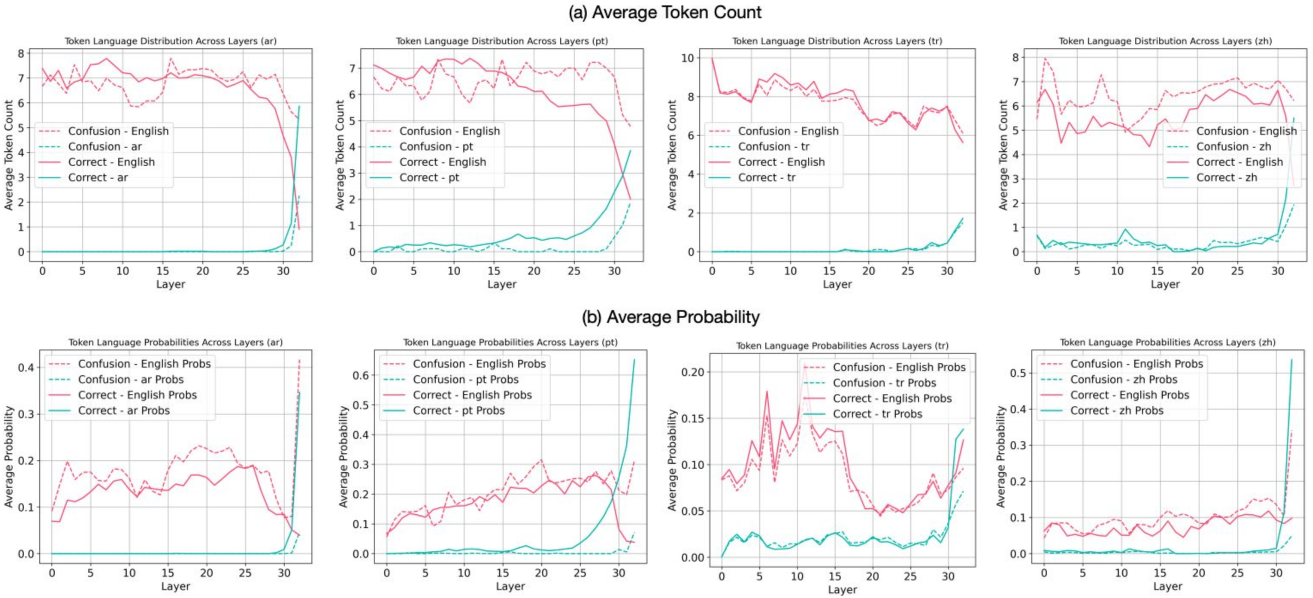
Findings - Transition Failure in Final Layers

Correct Cases:

- **Early/mid layers:** English tokens dominate (reflecting English-centric latent space).
- **Final layers:** Sharp transition—target language tokens overtake English, leading to correct output.

Confusion Cases:

- Transition to target language fails in final layers.
- English tokens remain dominant or increase, causing the model to switch to the unintended language at the CP.



Insights:

- Both correct and confusion cases start similarly, but diverge sharply in the last few layers.
- Language confusion is not a gradual drift, but a late-stage failure to shift from the latent conceptual space to the target language surface form.

Neuron-Level Attribution at Confusion Points

Goal: Identify which neurons are most responsible for the emergence of confusion points.

Method:

- For each confusion case, compute the importance of every FFN neuron at the token before the CP using the log-probability increase method.
- Rank neurons by their influence on the model's prediction at the CP.

Metric:

- Importance score = increase in log-probability of the CP token when the neuron is activated.

Findings – Distribution of Critical Neurons

- Critical neurons for confusion points are highly concentrated in the final layers.
- This pattern holds both for individual cases and when aggregated across all samples.
- These findings reinforce the conclusion from the previous layer-wise analysis: language confusion is tightly linked to the activity of specific FFN neurons in the **final** layers.

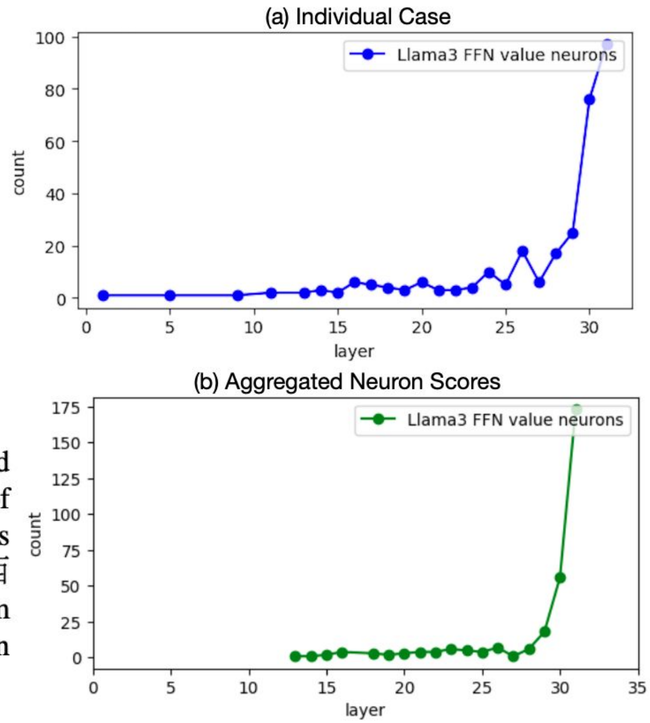


Figure 3: Distribution of Important Neurons Associated with Confusion Points in *Llama3-8B*. (a) Distribution of the top 300 most important FFN neurons across layers for an individual Chinese prompt "请解释拆东墙补西墙的意思。*(Please explain '拆东墙补西墙.')*" from Aya. (b) Aggregated distribution of important neuron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

Effect of Multilingual Instruction Tuning

Repeat neuron attribution on multilingual-tuned model (Llama3-8B-multilingual).

Findings:

- Most confusion-critical neurons in the original model become much less important after multilingual alignment.
- A small subset of neurons remains important, likely encoding general semantic information.

Interpretation:

- Multilingual instruction tuning suppresses confusion-inducing neurons, explaining its effectiveness in reducing language confusion.

Summary:

- Language confusion is driven by transition failures in the final layers.
- A small set of late-layer neurons are causally responsible for confusion points.
- Multilingual tuning works by suppressing these neurons' influence.
- These findings set the stage for targeted neuron-level interventions to mitigate confusion.

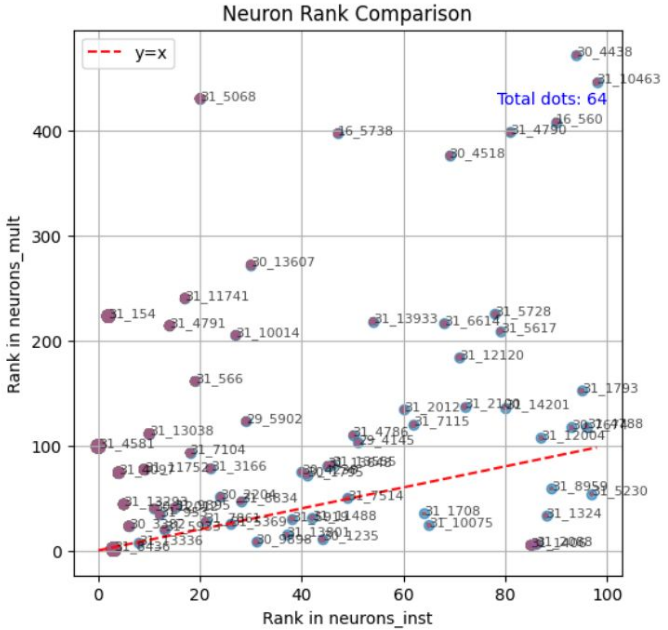


Figure 4: Neuron rank comparison between original Llama3 and multilingual Llama3. Results of Chinese test samples in Aya.

Outline

1. Introduction & Motivation
2. Methodological Background
3. Language Confusion Points
4. Mechanistic Analysis of Language Confusion
5. Mitigating Language Confusion: Neuron Editing
6. Conclusions

Neuron Selection Strategies

Why Selection Matters: Indiscriminate neuron editing can harm general model competence.

Three Strategies Explored:

- **Frequency-Based:** Select neurons most frequently important across confusion cases.
- **Aggregate Importance:** Select neurons with the highest total importance scores across all confusion cases.
- **Comparative Importance:** Select neurons whose importance for confusion points drops most after multilingual tuning (i.e., neurons specifically implicated in confusion, not general competence).

Rationale for Comparative Importance Selection

- **Motivation:** Many neurons important for confusion are also important for general language processing.
- **Comparative Approach:**
  - For each neuron, compute the difference in importance score between the original and multilingual-tuned models on the same input.
  - Prioritize neurons with the largest drop in importance—these are likely to be confusion-specific.
- **Advantage:** Minimizes collateral impact on general competence and fluency, focusing intervention on the root cause of confusion.

Neuron Editing – Methodology & Implementation

Editing Process:

- Select top 100 neurons per language using the chosen strategy.
- During inference, set the activations of these selected neurons to zero.

Evaluation: Assess on LCB benchmark and general language tasks (XNLI, sentiment analysis, fluency).

- **Evaluation of Confusion Reduction**
  - Language Confusion Metrics (LPR)
  - Internal Model Metrics (number and probability of target language tokens among top-10 output logits)
- **Evaluation of Output Quality and Generalization**
  - Fluency (Perplexity)
  - Generalization (Performance on general language tasks)

Quantitative Results – Confusion Reduction

Confusion mitigation performance of different selection strategies

	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**

Language Confusion Metrics:

- Substantial improvement in line-level pass rate (LPR) and accuracy after neuron editing.
- Comparative importance selection achieves the highest gains, matching or approaching multilingual-tuned models for most languages.

Internal Model Metrics:

- Increased number and probability of target language tokens among top-10 output logits.

	token_num	token_prob
Original	1.96	24.5
Edited	3.43	36.8
Diff	1.47	12.3

Quantitative Results – Output Quality & Generalization

Generalization

- Edited model maintains strong performance on out-of-domain prompts and general language tasks.
  No degradation in general language understanding (XNLI, sentiment analysis).

Fluency

- Output fluency (measured by perplexity) is preserved and even slightly improved.

	fluency_ori	fluency_cna	diff
ar	30.1	24.7	-5.4
pt	25.7	23.3	-2.3
tr	21.2	18.8	-2.5
zh	33.1	26.0	-7.0
es	25.4	23.2	-2.2
fr	21.2	21.1	-0.1
hi	28.5	22.9	-5.6
ru	23.7	19.5	-4.2
de	23.8	18.5	-5.3
it	25.7	20.2	-5.5
avg	25.8	21.8	-4.0

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

Comparison with Multilingual Tuning

Output Cleanliness:

- Neuron-edited model produces clean, well-structured responses in the intended language, preserving the semantic information.
- Avoids artifacts (e.g., HTML tags, noisy formatting) sometimes introduced by multilingual instruction tuning.

Effectiveness:

- Neuron editing can achieve confusion mitigation on par with full multilingual instruction tuning for many languages.

Advantages:

- **More targeted:** Directly addresses the internal cause of confusion.
- **More efficient:** Neuron editing is lightweight and reversible, requiring no retraining or fine-tuning.
- **Cleaner outputs:** Avoids side effects from noisy multilingual data.

Case 1 (aya-zh, id1)

Input: 请解释“拆东墙补西墙”的意思。*Please explain “拆东墙补西墙”*

Before Neuron Editing: English Translation: "To tear down the eastern wall to repair the western wall" - Meaning: To solve a problem by creating a new one, or to fix one issue by making another one worse.

After Neuron Editing: “拆东墙补西墙”是中国古代的一句成语，它的意思是“解决问题的方法是解决问题的根源，而不是简单地解决表面现象”。\n ...

Multilingual Llama3: 2023-03-23 14:34:23 拆东墙补西墙的意思是为了解决某个问题或问题所在的某个方面，而不是直接解决问题本身，...

Case 2 (okapi-zh, id5)

Input: 撰写一篇有关健康饮食重要性的争论性文章。*Write an argumentative essay on the importance of healthy eating.*

Before Neuron Editing: English Translation: "To tear down the eastern wall to repair the western wall" - Meaning: To solve a problem by creating a new one, or to fix one issue by making another one worse.

After Neuron Editing: 英语作文\n健康饮食的重要性：是或不是？\n在当今的生活中，健康饮食已经成为人们日常生活中的一部分。然而，是否健康饮食真的那么重要？

Multilingual Llama3:](https://www.zhihu.com/question/34614445) \n[如何在日常生活中更好地保持健康饮食习惯？](https://www.zhihu.com/question/34614445)

Outline

Conclusions

1. First mechanistic account of language confusion

2. Identification and intervention on critical neurons

3. Practical implications for robust, interpretable multilingual LLMs

References

- Marchisio, Kelly, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. "Understanding and Mitigating Language Confusion in LLMs." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 6653-6677. 2024.
- Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West. "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), pp. 15366-15394. 2024.
- Wang, Mingyang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. "Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models." In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.
- Yu, Zeping, and Sophia Ananiadou. "Neuron-Level Knowledge Attribution in Large Language Models." In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 3267-3280. 2024.
- Geva, Mor, Jasmijn Bastings, Katja Filippova, and Amir Globerson. "Dissecting Recall of Factual Associations in Auto-Regressive Language Models." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 12216-12235. 2023.
- Nie, Ercong, Helmut Schmid, and Hinrich Schütze. "Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models." arXiv preprint arXiv:2505.16538 (2025).

Thank you for your attention!