Can Language Neuron Intervention Reduce Non-Target Language Output?

Suchun Xie
Tohoku University
xie.suchun.p7@dc.tohoku.ac.jp

Hwichan Kim
Tohoku University
hwichan.kim.a2@tohoku.ac.jp

Shota Sasaki
Tohoku University
shota.sasaki.b4@tohoku.ac.jp

Kosuke Yamada
CyberAgent
yamada_kosuke@cyberagent.co.jp

Jun Suzuki
Tohoku University, RIKEN
jun.suzuki@tohoku.ac.jp

Abstract

Large language models (LLMs) often fail to generate text in the intended target language, particularly in non-English interactions. Concurrently, recent work has explored Language Neuron Intervention (LNI) as a promising technique for steering output language. In this paper, we re-evaluate LNI in more practical scenarios—specifically with instruction-tuned models and prompts that explicitly specify the target language. Our experiments show that while LNI also shows potential in such practical scenarios, its average effect is limited and unstable across models and tasks, with a 0.83% reduction in undesired language output and a 0.1% improvement in performance. Our further analysis identifies two key factors for LNI's limitation: (1) LNI affects both the output language and the content semantics, making it hard to control one without affecting the other, which explains the weak performance gains. (2) LNI increases the target language token probabilities, but they often remain below the top-1 generation threshold, resulting in failure to generate the target language in most cases. Our results highlight both the potential and limitations of LNI, paving the way for future improvements.

1 Introduction

Large language models (LLMs) have shown remarkable progress in various generation tasks. However, they still face challenges in consistently generating text in the expected target language, particularly for non-English content (e.g., Chinese or Japanese), even in high-performing models such as Llama 3 series (Grattafiori et al., 2024) and GPT-4o (Zhang et al., 2023; Chirkova and Nikoulina, 2024; Chen et al., 2024; Zhao et al., 2024a; Marchisio et al., 2024; Wang et al., 2024). This issue, referred to as the non-target language output issue, poses a significant challenge for the multilingual applications of LLMs.

Recent studies suggest that language models contain language-specific neurons, and activating or deactivating these neurons can influence models' output languages (Kojima et al., 2024; Tang et al., 2024; Zhao et al., 2024b). Building on these findings, Language Neuron Intervention (LNI), a method that activates target-language neurons, has been proposed as a lightweight approach to steering generation toward the target language without requiring additional training. Existing work (Kojima et al., 2024; Tang et al., 2024) has preliminarily demonstrated that in pre-trained models and in settings where the prompt does not explicitly specify the output language, LNI can steer outputs toward the desired language. Although these findings establish the fundamental feasibility of LNI, they do not directly address the settings under which LLMs are most commonly used in practice. In real-world applications, users typically interact with instruction-tuned models, and they often specify the desired output language directly in the prompt. The effectiveness of LNI under such practical conditions remains largely unexplored.

To fill this gap, we focus on the practical challenges of multilingual generation, particularly in instruction-tuned models, and investigate whether LNI can effectively reduce non-target language output even when the target language is explicitly specified. Through a comprehensive evaluation across diverse models and tasks, our results show that while LNI has some effect in reducing non-target language output, its average improvement is minimal, with negligible performance gains, and its effectiveness remains highly inconsistent. Addi-

Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 452–466
November 9, 2025 ©2025 Association for Computational Linguistics

tionally, our analysis reveals that LNI affects not only output language but also content semantics, indicating a coupling between language and content generation that explains its failure to improve task performance. Furthermore, our LogitFlow Analysis shows that although target-language tokens receive increased probabilities, they often fall short of the top-1 result, resulting in unsuccessful language switching. These findings expose key limitations of LNI and emphasize the need for more effective control mechanisms.

2 Preliminary

This section introduces Language Neuron Intervention (LNI), a method for steering the output language of a model by identifying and manipulating language-specific neurons. The core idea is to identify neurons that are highly sensitive to a target language, while remaining relatively stable for other languages, and override their activations during inference.

Two representative approaches exist for neuron selection (Kojima et al., 2024; Tang et al., 2024), we adopt the Average Precision (AP)-based method proposed by Kojima et al. (2024), as it achieves superior effect in our preliminary experiments (Appendix B). The method consists of the following steps: **Step 1: Corpus labeling.** Given a multilingual corpus, texts in the target language are labeled as positive examples (1), and all others as negative (0). **Step 2: Activation extraction.** For each text, neuron activation values are collected from all intermediate layers[1]. The activations are then averaged over non-padding tokens to obtain a scalar for each neuron within each text. **Step 3: Scoring and selection.** Each neuron is treated as a binary classifier of the target language, and its average activations are used to compute the AP score against the binary language labels. Neurons are then ranked by AP, and the top-k and bottom-k are selected. **Step 4: Intervention.** During inference, the activations of these selected neurons are replaced with their median value computed from target-language texts.

3 LNI Evaluation on Non-Target Output

In this section, we investigate the potential of language neuron intervention (LNI) in mitigating non-target language. Our primary objectives are to measure (1) the ratio change of non-target language outputs, and (2) their effect on task performance.

3.1 Experimental Settings

LNI Implementations Following Kojima et al. (2024), we identify language neurons using a balanced dataset of 500 samples from each of six languages (English, French, German, Spanish, Chinese, and Japanese), sourced equally from PAWS-X (Yang et al., 2019) and FLORES-200 (Team et al., 2022). We also experimented with task-specific settings on XL-Sum dataset, but observed no consistent improvement (Appendix C). Therefore, we adopt the general multilingual setting for all experiments. For neuron selection, we set $k = 1000$, based on our parameter tuning experiments over $k \in \{50, 150, 250, 500, 1000\}$ (Appendix D).

Models To evaluate LNI's effectiveness across different model families and scales, we experiment with both English-centric and multilingual LLMs. For English-centric models, we evaluate representative open-source models from the Llama series: Llama2-Chat 7B and 13B (Touvron et al., 2023), and Llama3-8B Instruct (Grattafiori et al., 2024). For multilingual models, we evaluate Bloomz-7b1-p3 (Muennighoff et al., 2022). All these models are instruction-tuned on generation tasks.

Datasets Previous work (Marchisio et al., 2024) showed that non-target language output issues are more prevalent in non-English languages, particularly Chinese (Zh) and Japanese (Ja). Thus, we focus on these two languages. To ensure broader coverage, we also include four other languages: French (Fr), Spanish (Es), Hindi (Hi), and Indonesian (Id), which together represent both high-resource and low-resource settings.

We focus on generation tasks, including **XL-Sum** (Hasan et al., 2021), a news summarization dataset, and **Dolly** (Conover et al., 2023), a dataset spanning diverse generation tasks, where we primarily use the QA subset. See dataset details in Appendix A.

Evaluation We evaluate on the following two metrics: **NT Ratio**: The percentage of responses generated in non-target languages, detected using fastText (Joulin et al., 2016, 2017)[2] with a threshold of 0.5. **Task Performance**: Text generation quality, measured by ROUGE-L score (Lin, 2004). See generation details in Appendix A.2.

[1] Excluding the embedding and output projection layers.

[2] https://github.com/facebookresearch/fastText

	Dolly						XL-Sum					
	NT Ratio (%)			RougeL			NT Ratio (%)			RougeL		
Model	Before	After	Δ	Before	After	Δ	Before	After	Δ	Before	After	Δ
						Japanese						
Llama2-Chat 7B	28	22	-6	0.09	0.10	+0.01	100	98	-2	0.01	0.01	0.00
Llama2-Chat 13B	19	16	-3	0.10	0.11	+0.01	4	10	+6	0.17	0.16	-0.01
Llama-3-8B-Instruct	**7**	2	-5	**0.14**	0.14	0.00	**3**	2	-1	**0.20**	0.20	0.00
BLOOMZ-7B1-P3	72	25	-47	0.02	0.05	+0.03	85	46	-39	0.02	0.08	+0.06
						Chinese						
Llama2-Chat 7B	43	30	-13	0.13	0.15	+0.02	100	99	-1	0.01	0.02	+0.01
Llama2-Chat 13B	44	36	-8	0.13	0.15	+0.02	18	31	+13	0.18	0.16	-0.02
Llama-3-8B-Instruct	14	7	-7	**0.19**	0.21	+0.02	**3**	8	+5	**0.24**	0.23	-0.01
BLOOMZ-7B1-P3	**6**	10	+4	0.12	0.11	-0.01	84	83	-1	0.03	0.03	0.00
						Spanish						
Llama2-Chat 7B	1	4	+3	0.20	0.20	0.00	56	51	-5	0.09	0.10	+0.01
Llama2-Chat 13B	0	1	+1	0.20	0.20	0	6	5	0.00	0.15	0.16	+0.01
Llama-3-8B-Instruct	0	0	0	0.20	0.20	0.00	0	0	0	0.18	0.18	0.00
BLOOMZ-7B1-P3	10	6	-4	0.10	0.12	+0.02	10	16	+6	0.17	0.16	-0.01
						French						
Llama2-Chat 7B	12	10	-2	0.19	0.18	0.00	70	69	-1	0.09	0.09	0.00
Llama2-Chat 13B	3	4	+1	0.20	0.19	0	7	6	0.00	0.17	0.18	+0.01
Llama-3-8B-Instruct	1	0	-1	0.20	0.20	0.00	0	0	0	0.21	0.20	-0.01
BLOOMZ-7B1-P3	9	5	-4	0.11	0.12	+0.01	25	11	-14	0.17	0.18	+0.01
						Hindi						
Llama2-Chat 7B	28	24	-4	0.10	0.10	0.00	100	100	0	0.00	0.00	0.00
Llama2-Chat 13B	13	10	-3	0.11	0.12	+0.01	96	97	+1	0.01	0.01	0.00
Llama-3-8B-Instruct	1	0	-1	0.18	0.17	-0.01	0	0	0	0.21	0.20	-0.01
BLOOMZ-7B1-P3	39	45	+6	0.04	0.03	0.00	57	63	+6	0.07	0.05	-0.02
						Indonesian						
Llama2-Chat 7B	15	9	-6	0.15	0.16	+0.01	33	18	-15	0.13	0.15	+0.02
Llama2-Chat 13B	17	11	-6	0.15	0.16	+0.01	17	28	+11	0.15	0.14	-0.01
Llama-3-8B-Instruct	1	1	0	0.19	0.18	-0.01	1	0	-1	0.20	0.20	0.00
BLOOMZ-7B1-P3	19	15	-4	0.08	0.09	+0.01	13	34	+21	0.19	0.14	-0.05

Table 1: Results on the Dolly and XL-Sum datasets across six languages: Japanese (ja), Chinese (zh), Spanish (es), French (fr), Hindi (hi), and Indonesian (id). **NT Ratio**: Non-target language output ratio (%, lower is better). **RougeL**: ROUGE-L F1 score (higher is better). **Change (Δ)** is computed as (After - Before). For NT Ratio, **negative Δ** indicates improvement (fewer non-target outputs, green); for RougeL, **positive Δ** indicates improvement (higher score, green). Color intensity reflects change magnitude (green, red). Best scores in **bold**.

3.2 Experiments Results

Table 1 presents the full results on the XL-Sum and Dolly tasks across six languages. In summary, although LNI shows potential in mitigating non-target language output, its effectiveness remains limited, with an average NT ratio reduction of only 0.83% and negligible performance gains (ROUGE-L +0.1%) across all language-task combinations. Moreover, its effectiveness is also highly unstable across different models, and task.

Inconsistent Effect across Models The effectiveness of neuron control varies significantly across model types. For instance, on Japanese tasks, the multilingual model Bloomz achieves substantial NT Ratio reductions (47% on Dolly and 39% on XL-Sum), In contrast, English-centric LLMs like Llama 2 and Llama 3 show only marginal changes (1–6%), indicating a large disparity.

Effect of Task Type The effect of LNI shows strong task dependency. On Dolly, a question generation task, most models exhibit a consistent improvement in both the NT Ratio and performance. However, on the XL-Sum task, nearly all models perform worse than Dolly. And half of the settings show degradation rather than improvement in both metrics.

4 Analysis of Failure Cases

To better understand the limitations of LNI, we analyze the failure cases observed in our experiments.

4.1 Language-Content Interplay

We observed that in some cases, while LNI reduced non-target language output, task performance remained unchanged or even declined (e.g., Llama3-8B on the Dolly task). If LNI solely controlled language selection, aligning the output language with the reference should improve ROUGE scores. However, performance degradation suggests that LNI interferes with content generation, possibly causing information loss, content distortion, or ambiguity. This suggests that LNI may not function solely as a language switch, but also interacts with deeper content generation mechanisms.

Analysis Settings We conduct a sample-wise analysis by comparing model outputs before and after LNI to examine its effects on both language and content consistency. For **Language Consistency**, we use fastText to detect language changes and classify them as correct and incorrect. For **Content Consistency**, we compute the BLEURT (Sellam et al., 2020) score[3] against reference labels before and after LNI, and compute their difference (Δ = after - before). The difference is used to categorize each sample as positive, negative, or neutral. We report the distribution of each category[4].

Analysis Results Table 2 presents representative results from our sample-wise analysis[5]. The results confirm our initial hypothesis: LNI consistently introduces semantic changes to the output content across all models, regardless of whether the language control is successful. These explain why performance improvements are not reliably observed, highlighting the entanglement between language-specific neurons and content generation.

4.2 Understanding the Mechanisms of LNI

From the experimental results, LNI yielded limited reductions in non-target language output, with most samples failing to switch to the target language. To understand this behavior, we analyze LNI's mechanism and address a key question: why do most samples retain their original language despite intervention?

Model	Language (%)			Content (%)		
	Corr.	Wro.	Unch.	Pos.	Neg.	Neutral
Llama2-7B	20	16	64	31	7	62
Llama3-8B	10	6	84	19	18	63
Bloomz-7B	58	5	37	22	11	67

Table 2: Sample-wise analysis of LNI effects on the Dolly task (Japanese). LNI alters both output language (Correct, Wrong, Unchanged) and content semantics (Positive, Negative, Neutral).

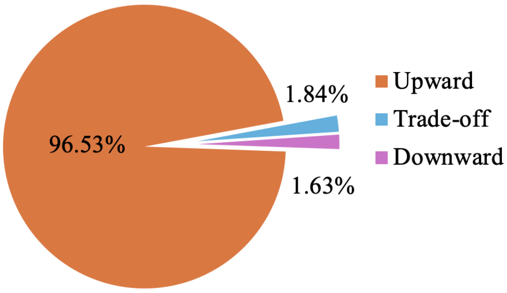


Figure 1: Logit change categories for target-language tokens after LNI on the XL-Sum Japanese task using Llama2 7B. Upward shifts dominate, while trade-off and downward shifts are rare.

Analysis Settings We use LogitFlow Analysis to quantify how target language token probabilities shift before and after intervention. This method tracks changes in token probability across different generation stages, providing insights into how LNI influences internal model dynamics. Specifically, we represent each token's transition from pre- to post-states using a 2D vector, constructed from two key metrics: (1) the number of target language tokens in the top-k logits (x-axis) and (2) the earliest rank where a target token appears in the top-k logits (y-axis). We then classify and visualize these shifts into three patterns: Upward: Increased target-language probability (via improved ranking or frequency); Downward: Decreased probability or ranking; and Mixed: Trade-offs between ranking and frequency shifts[6].

Analysis Results As shown in Figure 1, our analysis revealed that LNI consistently increases the probability of target-language tokens in most cases. However, these improvements are often insufficient to surpass competing tokens and reach the top-1 position, resulting in no change in the output language[7]. Given that higher sampling temperatures

[3] A metric that can measure semantic similarity between texts even when languages are different.

[4] Full details are provided in Appendix E.

[5] Full results are provided in Appendix Table 8.

[6] See detailed classification criteria in Appendix F.

[7] For a more detailed view, see the Token Distribution Flow Field maps (e.g., Figure 2) in Appendix F.

exacerbate language confusion (Marchisio et al., 2024), we use deterministic generation (temperature = 0.0) in this experiment, where the model always selects the highest probability token. Under this setting, only tokens that reach the top-1 position can be generated, meaning that even if LNI increases the probability of the target token, it remains ineffective unless it overtakes the top-ranked token, leading to failed language switching.

This finding suggests that combining LNI with appropriate decoding strategies (e.g., temperature scaling and beam search) might enhance its ability to control language output. Additionally, we observe cases where target language token probabilities decrease, potentially due to language switching occurring at different stages of text generation, leading to instability. This indicates that LNI alone is insufficient, suggesting the need for more robust control mechanisms in multilingual generation.

5 Related Work

Non-target Language Output Issue Prior work has observed that LLMs often struggle to consistently generate text in the intended target language. This phenomenon, commonly referred to as *off-target language generation* or *language confusion*, has been widely documented in recent studies (Zhang et al., 2023; Chirkova and Nikoulina, 2024; Chen et al., 2024; Zhao et al., 2024a; Marchisio et al., 2024; Wang et al., 2024). Marchisio et al. (2024) systematically evaluate language confusion in LLMs and show that English-only instruction-tuning amplifies models' preference for English, leading to English outputs even when prompted in other languages.

To mitigate language confusion, Marchisio et al. (2024) demonstrates that providing few-shot examples and applying multilingual SFT are effective strategies.Lee et al. (2025) propose an ORPO method, which incorporates penalties for undesired output languages into standard SFT. In addition, several inference-time approaches have been introduced, which directly steer the language vector (Yunfan et al., 2025) or manipulate language-specific neurons (Tang et al., 2024; Zhao et al., 2024b; Kojima et al., 2024; Tan et al., 2024) to control the output language.

LNI Method The findings of Mahowald et al. (2024) and Zhang et al. (2024) lay the groundwork for work on language-specific neurons in this area. Mahowald et al. (2024) found that while LLMs perform well on formal language tasks such as grammar, they exhibit instability in functional language use, suggesting a dissociation between language processing and cognitive abilities. Zhang et al. (2024) extend these findings by revealing that only 1% of model parameters are critical for language performance, and perturbing this subset leads to substantial degradation in multilingual performance.

Building on this line of work, recent research introduces different approaches to identifying language-specific neurons and validating their effects by intervening in these neurons (Tang et al., 2024; Zhao et al., 2024b; Kojima et al., 2024; Tan et al., 2024). These works demonstrate that language-specific neurons play a crucial role in multilingual behavior. However, prior works mainly focus on scenarios with ambiguous or implicit target languages (Kojima et al., 2024) using pre-trained models, or are limited to small-scale case studies (Tang et al., 2024), without systematically evaluating the effectiveness of LNI in mitigating non-target language outputs, particularly under the setting with explicitly specified target languages, which are more representative of real-world LLM usage scenarios and motivate our investigation. As a concurrent work, Mondal et al. (2025) also evaluated the effectiveness of LNI but from a cross-lingual transfer perspective.

6 Conclusion

In this paper, we systematically evaluate LNI's potential in addressing non-target language output issues. Our results show that while LNI shows potential in reducing non-target language output, its average effect is limited and unstable across models and tasks, with only a 0.83% reduction in undesired output and a 0.1% improvement in task performance.

Through sample-wise analysis, we identified an intrinsic coupling between language and content generation. This explains why controlling language neurons unexpectedly degrades performance. To better understand its limitations, we introduced LogitFlow Analysis, an analysis method for tracking token probability shifts before and after intervention, Our findings reveal that although LNI increases the probability of target language tokens, it often fails to reach the top-1 position, leading to instability in generation.

Limitations

While our study advances the understanding of language neuron intervention (LNI) and demonstrates its potential and failures in addressing non-target language outputs, there are several limitations.

Regarding model selection, our analysis primarily focuses on representative model families - the English-centric Llama series and the multilingual Bloomz model. Although we examine different model scales (7B and 13B) and contrasting architectural approaches (English-centric versus multilingual pre-training), this represents only a subset of the diverse landscape of large language models.

In terms of language selection, our primary experiments focus on Chinese and Japanese, which frequently exhibit non-target language output issues and present challenges due to their linguistic distance from English. To verify the generality of our findings, we additionally evaluate LNI on four additional languages—French, Spanish, Hindi, and Indonesian—covering both high-resource and low-resource settings. The results are consistent with our main findings. Future work may further explore LNI behavior in other scripts or typologically diverse languages.

Besides, neuron overlap across languages is also a potential factor influencing the effectiveness of LNI. Due to space limitations, we did not discuss this factor in the main text, but we conducted additional experiments to evaluate its impact. Specifically, we compared settings with and without filtering overlapping neurons, particularly between English and the target language. The results indicate that such filtering does not improve output language accuracy or task performance, and in some cases, slightly degrades them. Full details are provided in Appendix G.

Ethical Considerations

Our study conducted experiments on open-source models and publicly available datasets, mitigating risks related to data contamination, privacy leaks, and ethical concerns associated with data collection and human annotation. However, these models may already contain biases introduced during pre-training and instruction tuning. Such biases could impact the tasks examined in this study, particularly news summarization, which demands factual accuracy, and QA generation, where responses may exhibit randomness. This raises concerns about potential inaccuracies or misleading outputs, especially in non-English languages such as Chinese and Japanese. Given these challenges, careful evaluation and responsible deployment of LLMs in multilingual settings are crucial for minimizing unintended biases and ensuring reliable outputs.

Acknowledgments

We thank the members of the Tohoku NLP Group for their insightful comments. This work was supported by the JSPS KAKENHI Grant Number JP24H00727; JST Moonshot R&D Grant Number JPMJMS2011-35 (fundamental research); JST SPRING, Grant Number JPMJSP2114;

In this research, we used ABCI 2.0 provided by AIST and AIST Solutions, and the mdx: platform for building a data-empowered society.

References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, and 8 others. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *Preprint*, arXiv:2202.01279.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

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

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. 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 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. Controlling language confusion in multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1026–1035, Vienna, Austria. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Preprint*, arXiv:2301.06627.

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.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62, Albuquerque, New Mexico. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, and 13 others. 2022. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 1–7.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6506–6527, Miami, Florida, USA. Association for Computational Linguistics.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. 2024. Mitigating the language mismatch and repetition issues in LLM-based machine translation via model editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15681–15700, Miami, Florida, USA. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Xie Yunfan, Lixin Zou, Dan Luo, Min Tang, Chenliang Li, Xiangyang Luo, and Liming Dong. 2025. Mitigating language confusion through inference-time intervention. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8418–8431, Abu Dhabi, UAE. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Qiyuan Chen Ziang Leng and Cheng Li. 2023. Luotuo: An instruction-following chinese language model, lora tuning on llama. `https://github.com/LC1332/Luotuo-Chinese-LLM`.

A Datasets and Generation

A.1 Datasets

XL-Sum dataset is a news summary task that provides a news article to summarize the main content. We randomly sampled 888 samples per language from the test datasets.

The databricks-dolly-15k dataset is a Wikipedia-based human-generated dataset; it contains various generation tasks, including Closed QA, Open QA, Summarization, and others. The original dataset is not available for Chinese and Japanese, we use its machine-translated version. the chinese-dolly-15k (Ziang Leng and Li, 2023)[8] and databricks-dolly-15k-ja[9].

In this experiment, we mainly tested on the question generation task. Specifically, with the random seed 42, we randomly sampled 200 samples from each of the `open_qa` and `general_qa` categories. In the Chinese-dolly-15k dataset, the reference response is only available in English. Thus, we translated it into Chinese using ChatGPT API, with the model GPT-4o mini.

A.2 Generation Settings

Marchisio et al. (2024)'s work reported that language confusion is aggregated by high sampling temperatures, therefore, we employ a temperature of 0.0 and a Top-p of 0.9. The Max new tokens is set to 100. We use English prompt setting and specify the output language in the prompt. For XL-Sum task, we adapted the prompts from the templates provided by PromptSource (Bach et al., 2022). Detailed prompt settings for each dataset are provided in Table 3.

Task	**Prompts**
XL-Sum	Write one sentence to summarize the given document. The document is: {*Input*} Summarize in lang:
Dolly	Answer the following question in lang.{*Input*}

Table 3: Prompt settings for XL-Sum and Dolly.

Method	Lang. Change		Content Change	
	Correct	Wrong	Positive	Negative
JA				
AP	1.6	0.0	3.0	3.2
LAPE	0.1	0.2	2.1	2.8
ZH				
AP	1.0	0.0	6.5	4.3
LAPE	0.0	0.2	2.9	3.4

Table 4: Comparison of neuron selection methods on Llama2-Chat 7B for Japanese (JA) and Chinese (ZH) datasets. AP achieves higher language control and content quality improvements, leading to its adoption in this work.

B Preliminary Experiments on Different LNI Approaches

Two representative approaches exist for neuron selection (Kojima et al., 2024; Tang et al., 2024), both targeting neurons with strong language specificity but differing in selection details. Since no prior work has evaluated these methods under the same settings, we conducted preliminary experiments to determine which approach performs better for our tasks.

Using the same settings as the formal experiments, including generation inference parameters and neuron control configurations, we evaluated both methods on the Llama2-Chat 7B model with Chinese and Japanese datasets on the XL-Sum task. The results, shown in Table 4, indicate that the AP-based method from Kojima et al. (2024) outperforms the alternative. Therefore, we adopt this method in our study.

C Dataset for Neuron Identification

In this section, we compare two approaches to neuron identification: one based on general multilingual data (following Kojima et al. (2024)), and the other based on task-specific data from XL-Sum.

In the general setting, language neurons are identified using a balanced dataset of 500 samples per language, drawn from PAWS-X (Yang et al., 2019) and FLORES-200 (Team et al., 2022). In the task-specific setting, the same identification procedure is applied to samples from the XL-Sum validation set. Both settings use 500 samples per language across six languages: English, French, German, Spanish, Chinese, and Japanese. We evaluate the two settings on the XL-Sum test set, comparing

[8] silk-road/chinese-dolly-15k

[9] llm-jp/databricks-dolly-15k-ja

their effects on non-target language output ratio (NT Ratio) and task performance.

As shown in Table 5, neither setting consistently outperforms the other across languages or metrics. Given this, we adopt the general setting in our main experiments for its simplicity and better alignment with prior work.

D Parameter-tuning on Neuron Number Selection

We investigate how the number of selected neurons k affects model behavior on the XL-Sum task. We tested values $k \in \{50, 150, 250, 500, 1000\}$, where for each setting, both the top-k and bottom-k neurons are selected. This results in a total of $2k$ neurons being used for activation manipulation.

The results are provided in Table 6. We observe that:

- Task performance, as measured by ROUGE-L, remains largely stable across different values of k.
- The non-target language output ratio (NT Ratio) fluctuates depending on the model and language, but no consistent trend emerges.

Given these observations, we set $k = 1000$ for all main experiments. This choice is further supported by the findings of Kojima et al. (2024), who report that $k = 1000$ achieves a strong trade-off between language control and task performance across multilingual settings.

E Sample-wise Language and Content Consistency Analysis

We conduct a sample-wise analysis comparing model outputs before and after neuron intervention to understand LNI's effect on both language and content consistency. For language consistency, we primarily examine language changes—whether the output remains unchanged, is correctly changed, or is incorrectly changed. For each sample, we classify the changes into three patterns. This enables a better understanding of how the language and content change; these patterns are as follows :

- Pattern-1: No change
- Pattern-2: Correct language change only
- Pattern-3: Correct language change with content change
- Pattern-4: Content change only
- Pattern-5: Incorrect language change
- Pattern-6: Incorrect language change with content change

For language output classification, we employ fastText[10] to determine the output language. Content changes are identified using BLEURT (Sellam et al., 2020) scores, where a score below 0.8 between pre- and post-intervention predictions indicates substantial content modification. The results are provided in Table 7. The last two columns represent the computed ratio of correct and incorrect language changes, which are used in Table 2.

While this classification reveals high content modification rates (> 90%) across all settings, we further analyze the quality of these content changes by comparing BLEURT scores between predictions and ground truth labels. Changes are categorized as positive (score difference > 0.1), negative (< -0.1), or neutral (-0.1 to 0.1), allowing us to distinguish between beneficial and detrimental content modifications. The results for both the Dolly and XL-Sum datasets are provided in Table 8.

F LogitFlow Analysis

Vector Construction For each generation step, we analyze the top-k (k=20) logits distribution before and after neuron control. The vector is constructed using two metrics:

- Position (y-axis): The earliest position where a target language token appears in the top-k
- Count (x-axis): The number of target language tokens in top-k

Vector Classification We classify the vectors into the following types based on the changes in position and count:

- Improvement: Position improves (ranks earlier), or position stays the same with increased count
- Trade-off: Position worsens, but count increases.Because prior to the position metric, it is more important to count.
- Degradation: Position worsens with no count increase, or position stays the same with decreased count

[10] https://github.com/facebookresearch/fastText

Model	Lang	NT Ratio (↓)		Performance (↑)	
		General	Task-specific	General	Task-specific
Llama2-Chat 7B	JA	**0.98**	0.99	0.01	0.01
	ZH	**0.99**	0.99	0.01	0.01
Llama2-Chat 13B	JA	0.09	**0.07**	0.16	0.16
	ZH	0.29	**0.24**	0.15	**0.16**
Llama-3-8B-Instruct	JA	0.11	**0.07**	**0.20**	0.19
	ZH	**0.04**	0.06	0.23	0.23
BLOOMZ-7B1-P3	JA	0.49	**0.46**	0.08	**0.09**
	ZH	**0.87**	0.97	**0.03**	0.01

Table 5: Comparison of neuron identification using general data (FLORES+PAWS-X) and task-specific data (XL-Sum) on the XL-Sum task. NT Ratio (↓) and Performance (↑, ROUGE-L) are reported under the same control setting. Best scores are highlighted in bold.

	NT Ratio (↓)					Performance (↑)				
Model	1K	500	250	150	50	1K	500	250	150	50
Japanese										
Llama2-Chat 7B	0.98	0.99	1.00	0.99	1.00	0.01	0.01	0.01	0.01	0.01
Llama2-Chat 13B	0.09	0.08	0.04	0.04	0.03	0.16	0.16	0.17	0.17	0.17
Llama-3-8B-Instruct	0.11	0.00	0.00	0.03	0.03	0.20	0.20	0.21	0.20	0.20
BLOOMZ-7B1-P3	0.49	0.48	0.62	0.69	0.85	0.08	0.08	0.06	0.05	0.03
Chinese										
Llama2-Chat 7B	0.99	0.99	0.99	1.00	1.00	0.01	0.01	0.01	0.01	0.01
Llama2-Chat 13B	0.29	0.22	0.18	0.20	0.18	0.15	0.17	0.18	0.17	0.18
Llama-3-8B-Instruct	0.04	0.03	0.02	0.05	0.04	0.23	0.24	0.24	0.24	0.24
BLOOMZ-7B1-P3	0.87	0.84	0.88	0.87	0.87	0.03	0.04	0.03	0.03	0.03

Table 6: NT Ratio (↓ lower is better) and Performance (↑ higher is better) for Japanese and Chinese across different parameter settings.

- Unchanged: No change in both metrics

The full results of the LogitFlow for XL-Sum are provided in Figure 2 and for Dolly in Figure 3, respectively.

G Neuron Overlap Analysis

Neuron overlap across languages is also a potential factor influencing the effectiveness of LNI. While prior work (Kojima et al., 2024) suggests such an overlap ratio is typically limited (under 5%), we conducted additional experiments to evaluate its impact.

We considered two aspects: (1) a small portion of neurons may overlap between each target language and English, and (2) most non-target outputs tend to be in English. Based on this, our overlap-aware intervention strategy filters out target-language neurons that overlap with English. We then enhance the activation of the remaining target-specific neurons (by replacing their activation with the target language's median), while simultaneously deactivating English neurons (by setting them to zero).

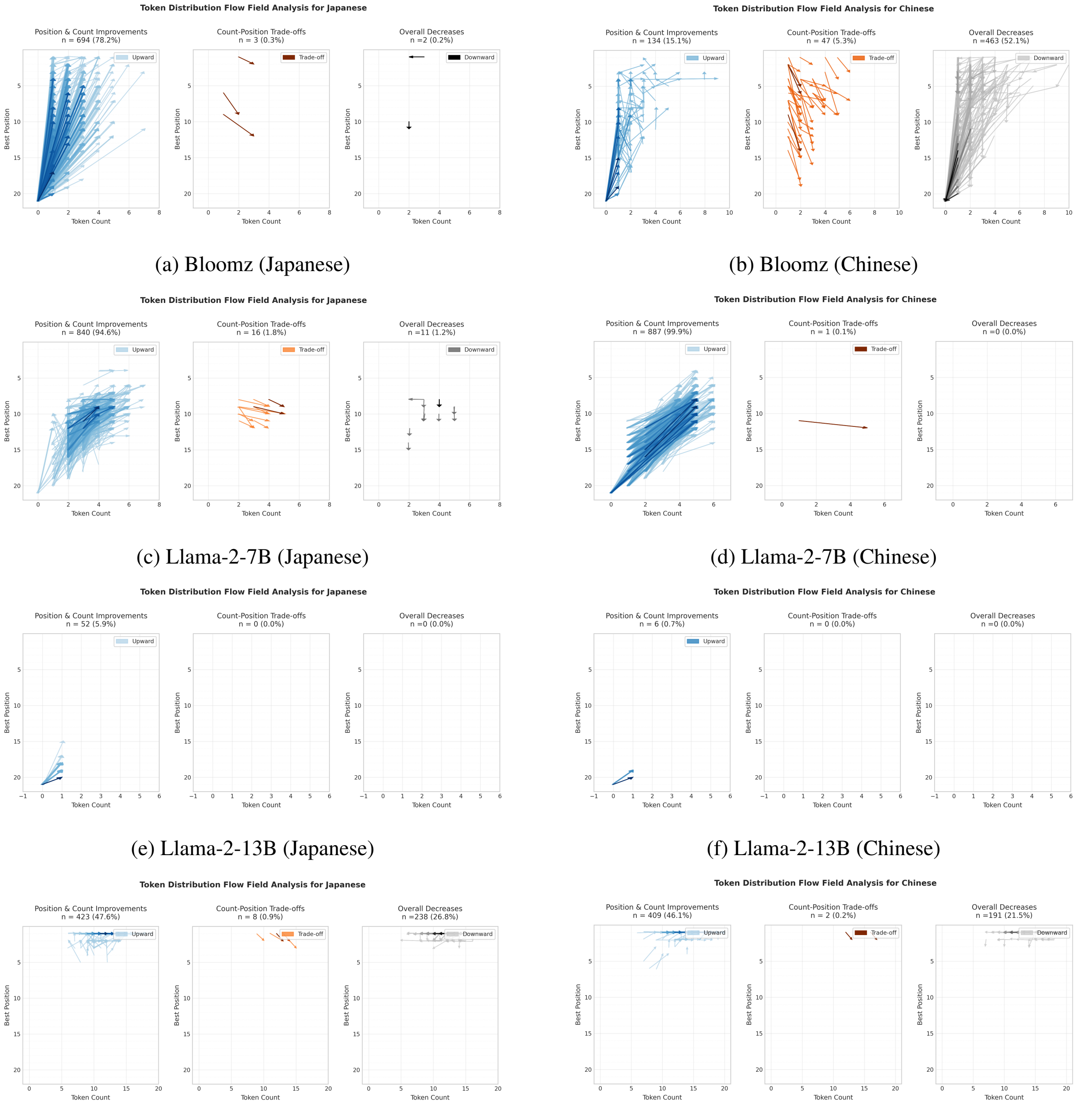
Table 9 compares results with and without overlap filtering. We observe no consistent improvement; in most cases, NT ratio increases and task performance declines. Therefore, overlap filtering was not applied in our main experiments.

H License and Intended Use

We list the license of each model we utilized as follows:

- **Llama 2**: Llama2-Chat 7B, 13B [Meta AI]
- **Llama 3**: Llama3-8B-Instruct [Llama 3 License]
- **Bloomz**: BLOOMZ-7B1 [Hugging Face]

Both Llama families are designed for commercial and research use in English, with their instruction-tuned versions (used in this study) optimized for assistant-like chat. Bloomz is recommended for performing tasks expressed in natural language.



(a) Bloomz (Japanese)

(b) Bloomz (Chinese)

(c) Llama-2-7B (Japanese)

(d) Llama-2-7B (Chinese)

(e) Llama-2-13B (Japanese)

(f) Llama-2-13B (Chinese)

(g) Llama-3-8B (Japanese)

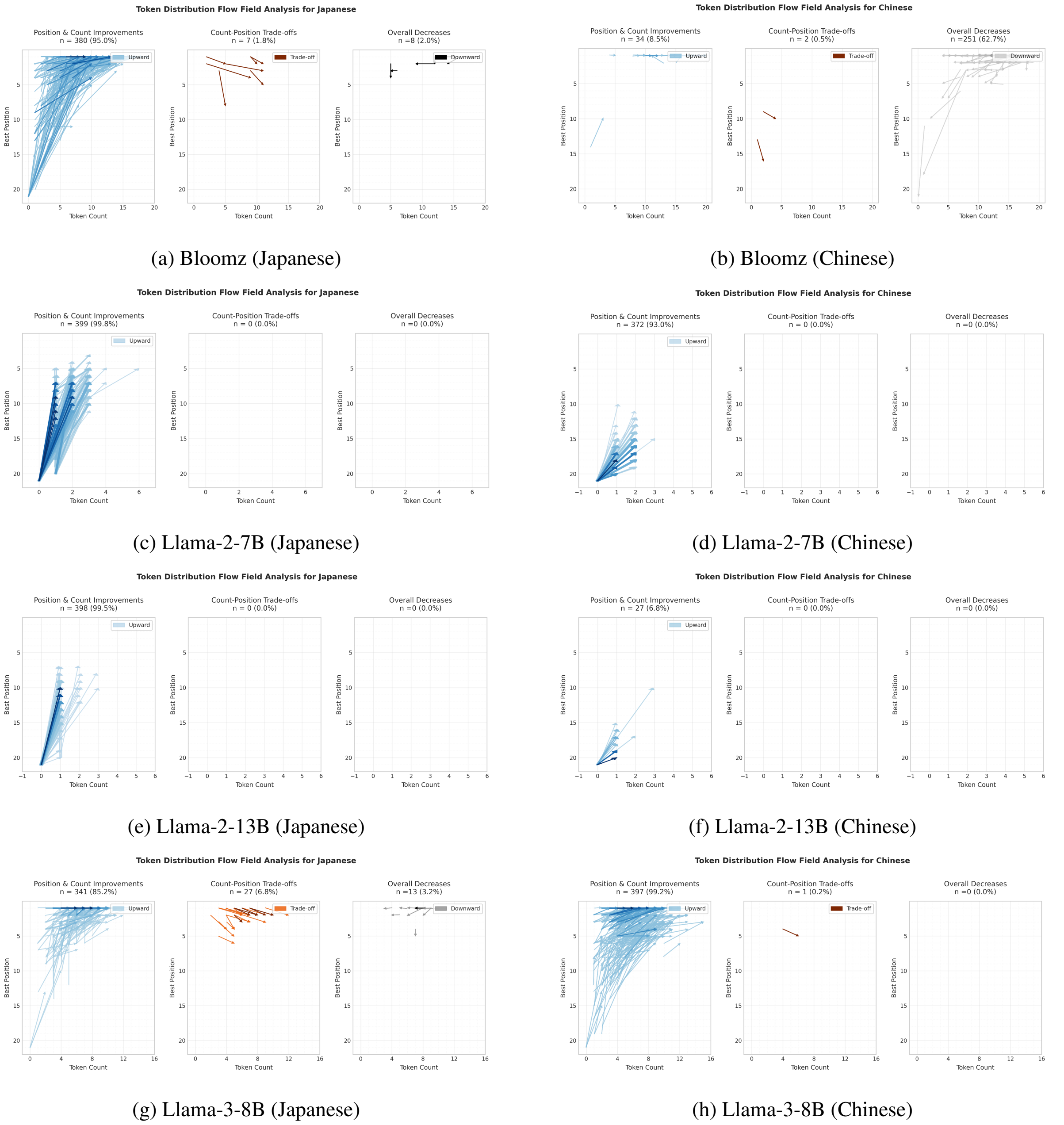
(h) Llama-3-8B (Chinese)

Figure 2: Token Distribution Flow Field Analysis on **XL-Sum** dataset across different models and languages. Each subplot shows the changes in token position and count after neuron control, with arrows indicating the direction and magnitude of change.

I Computational Setup

We primarily use GPUs for inference, as all models in this paper are instruction-tuned, open-source models. Experiments were conducted on the MDX platform (Suzumura et al., 2022), a cloud-based infrastructure designed for data science and interdisciplinary research.

We use A100 (40GB) GPUs for all models. For language neuron identification and intervention experiments, we use 8 GPUs. The 7B models require approximately 1 hour, while larger models take around 3 to 5 hours. For inference without language neuron exploration, a single GPU is used across all models.



(a) Bloomz (Japanese)

(b) Bloomz (Chinese)

(c) Llama-2-7B (Japanese)

(d) Llama-2-7B (Chinese)

(e) Llama-2-13B (Japanese)

(f) Llama-2-13B (Chinese)

(g) Llama-3-8B (Japanese)

(h) Llama-3-8B (Chinese)

Figure 3: Token Distribution Flow Field Analysis on **Dolly** dataset across different models and languages. Each subplot shows the changes in token position and count after neuron control, with arrows indicating the direction and magnitude of change.

Model	Lang.	Pattern Distribution						Language change (%)	
		P1	P2	P3	P4	P5	P6	Correct	Wrong
					Dolly				
Llama-2-7B	JA	13	-	78	247	-	62	20	16
	ZH	19	1	89	232	-	59	23	15
Llama-2-13B	JA	11	-	70	269	2	48	18	13
	ZH	36	1	68	242	1	52	17	13
Llama-3-8B	JA	25	-	38	315	-	22	10	6
	ZH	40	-	60	271	1	28	15	7
Bloomz-7B	JA	40	-	188	81	-	15	58	5
	ZH	96	1	6	147	-	23	3	8
					XL-Sum				
Llama-2-7B	JA	9	0	14	865	0	0	2	0
	ZH	18	0	9	861	0	0	1	0
Llama-2-13B	JA	50	0	4	777	0	57	0.5	6
	ZH	83	3	73	525	4	200	9	23
Llama-3-8B	JA	41	0	41	781	0	25	5	3
	ZH	167	0	11	654	0	57	1	6
Bloomz-7B	JA	25	0	392	404	1	62	44	7
	ZH	155	0	38	607	0	88	4	10

Table 7: Sample pattern distribution and overall language change rates. Patterns: P1 (No change), P2 (Correct language change only), P3 (Correct language change with content change), P4 (Content change only), P5 (Wrong language change), P6 (Wrong language change with content change).

Model	Lang.	Δ		Language Change		Content Change	
		NT Ratio	ROUGE	Correct	Wrong	Pos.	Neg.
				Dolly			
Llama-2-7B	JA	6	0.01	20	16	31	7
	ZH	13	0.02	23	15	31	12
Llama-2-13B	JA	3	0.01	18	13	26	9
	ZH	8	0.02	17	13	24	11
Llama-3-8B	JA	5	0.00	10	6	19	18
	ZH	7	0.02	15	7	29	13
Bloomz-7B	JA	47	0.03	58	5	22	11
	ZH	-4	-0.01	3	8	11	19
				XL-Sum			
Llama-2-7B	JA	2	0.00	2	0	3	3
	ZH	1	0.01	1	0	7	4
Llama-2-13B	JA	-6	-0.01	0.5	6	7	7
	ZH	-13	-0.02	9	23	9	8
Llama-3-8B	JA	1	0.00	5	3	8	13
	ZH	-5	-0.01	1	6	6	11
Bloomz-7B	JA	39	0.06	44	7	24	7
	ZH	1	0.00	4	10	9	11

Table 8: Analysis of LNI effects on Dolly and XL-Sum datasets. Δ NT Ratio: Change in non-target language output ratio; Δ ROUGE: ROUGE-L score change; Correct/Wrong: Correct or incorrect language changes; Pos./Neg.: Positive/negative content changes. We highlight contrasting cases. For example, in Dolly, Llama-3-8B (JA) shows balanced content changes despite the NT Ratio improved.

	NT Ratio			ROUGE-L		
Model	w/o	w	Δ	w/o	w	Δ
			Japanese			
Llama-2-7B-Chat	0.98	0.99	+0.01	0.01	0.01	0.00
Llama-2-13B-Chat	0.09	0.07	-0.02	0.16	0.16	0.00
Llama-3-8B-Instruct	0.11	0.11	0.00	0.20	0.18	-0.02
Bloomz-7B1-P3	0.49	0.60	+0.11	0.08	0.07	-0.01
			Chinese			
Llama-2-7B-Chat	0.99	0.99	0.00	0.01	0.03	+0.02
Llama-2-13B-Chat	0.29	0.55	+0.26	0.15	0.11	-0.04
Llama-3-8B-Instruct	0.04	0.38	+0.34	0.23	0.20	-0.03
Bloomz-7B1-P3	0.87	0.92	+0.05	0.03	0.03	0.00

Table 9: Effect of neuron overlap filtering on non-target language ratio (NT) and performance, ROUGE-L (RL) on XL-Sum dataset. “w/o” and “w/” refer to without and with overlap filtering, respectively. Δ = w/ - w/o. Lower NT and higher RL indicate better results.