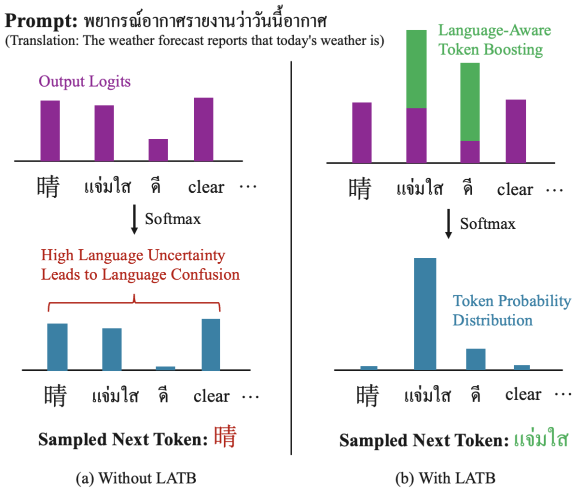
Towards Confident Multilingual Generation from English-Centric LLMs: A Tuning-Free Approach

Anonymous ACL submission

Abstract

This paper introduces a new taxonomy of multilingual alignment for English-centric language models through token perturbation techniques. We propose two methods within this paradigm: the Language-Aware Token Boosting (LATB), which directly adds perturbations to desired language tokens, and its adaptive variant, the Adaptive Language-Aware Token Boosting (Adaptive-LATB), which dynamically adjusts perturbations based on the model's confidence in the intended language. Extensive experiments show that our methods effectively enhance multilingual alignment. Compared to the fine-tuning method, our approaches achieve superior results in reducing language confusion and improving summarization quality without requiring additional fine-tuning. Our code will be publicly available soon.


Prompt: พยากรณ์อากาศรายงานว่าวันนี้อากาศ
(Translation: The weather forecast reports that today's weather is)
Output Logits
Language-Aware Token Boosting
晴
แจ่มใส
ดี
clear
···
Softmax
High Language Uncertainty Leads to Language Confusion
Token Probability Distribution
Sampled Next Token: 晴
Sampled Next Token: แจ่มใส

(a) Without LATB

(b) With LATB

Figure 1: Language-Aware Token Boosting (LATB) enhances target language generation confidence by selectively boosting target language tokens.

1 Introduction

Large Language Models (LLMs) have shown impressive performance, but their English-centric development limits their effectiveness for non-English users (Hadi et al., 2024, 2023). Recent efforts (Xue et al., 2021; Workshop et al., 2023; Wei et al., 2023) aim to enhance multilingual capabilities, though English-centric models still underperform in low-resource languages (Qin et al., 2024; OpenAI et al., 2024). One of the key issues is language confusion (Devine, 2024), where models fail to consistently generate the desired language, particularly in non-English contexts (Marchisio et al., 2024). Techniques to mitigate this include temperature lowering, few-shot prompting, and fine-tuning (Marchisio et al., 2024), but these come with limitations such as reduced responses diversity (Agarwal et al., 2024; Renze and Guven, 2024) or increased computational costs.

We propose a novel tuning-free paradigm for multilingual alignment, using perturbations directly on the logits. This approach eliminates the need for fine-tuning and aligns the model's outputs with the desired language, incurring minimal additional computational costs during inference. We introduce two methods within this paradigm: **Language-Aware Token Boosting (LATB)**, which applies language-specific token perturbations, and **Adaptive Language-Aware Token Boosting (Adaptive-LATB)**, which adapts perturbations by introducing perturbations selectively—only when the LLM exhibits uncertainty in generating one language over another.

We evaluate our methods on the XLSUM multilingual summarization benchmark (Hasan et al., 2021) across eight languages. Both LATB and Adaptive-LATB effectively reduce language confusion and enhance summarization performance compared to their respective base models and the multilingual-tuned model. We also analyze the effects of hyperparameters, including perturbation values and confidence difference thresholds.

In summary, our contributions are as follows:

1. We propose a novel tuning-free multilingual

alignment paradigm based on logits perturbation, introducing two methods: LATB and Adaptive-LATB.

2. We evaluate our methods on the XLSUM benchmark, showing reduced language confusion and improved summarization quality.

3. We provide an analysis of the impact of hyperparameters on the language confusion and summarization quality.

2 Related Work

Multilingual Large Language Models. Multilingual Large Language Models (MLLMs) are designed to process multiple languages simultaneously. The approaches for developing and optimizing these models can be broadly categorized into two main types: parameter-tuning alignment (PTA) and parameter-frozen alignment (PFA) (Qin et al., 2024). The PTA approach involves tuning the model's parameters to enable multilingual capabilities. This tuning can occur at various stages, including pretraining (Xue et al., 2021; Chowdhery et al., 2022; Workshop et al., 2023; Jiang et al., 2023, 2024), supervised fine-tuning (SFT) (Chung et al., 2022; Muennighoff et al., 2023; Devine, 2024; Pipatanakul et al., 2023), reinforcement learning with human feedback (RLHF) (Lai et al., 2023b; Touvron et al., 2023; GLM et al., 2024; Bai et al., 2023), and downstream task fine-tuning (Lepikhin et al., 2020; Rosenbaum et al., 2022). In contrast, PFA methods do not require parameter tuning for multilingual alignment. Instead, they primarily rely on prompting techniques (Abdelali et al., 2024; Winata et al., 2023; Lu et al., 2024; Puduppully et al., 2023) and retrieval-augmented alignment (He et al., 2023; Zhang et al., 2023; Conia et al., 2023). Our proposed method falls within the PFA category. To the best of our knowledge, our study is the first to introduce a new taxonomy for logits perturbation-based multilingual alignment.

Language Confusion. Language confusion refers to the inconsistent ability of LLMs to generate responses in a target language. This phenomenon has been observed across various NLP tasks, such as machine translation (Vu et al., 2022; Li and Murray, 2023), summarization (Wang et al., 2023; Yu et al., 2022), and question answering (Holtermann et al., 2024). While this issue has been systematically studied with various proposed methods mitigating it (Marchisio et al., 2024), our study introduces a novel and cost-effective approach to mitigate language confusion using token perturbation methods.

3 Approach

3.1 Token Language Identification

We identify tokens to boost based on the target language using a Unicode filtering method following (Wen-Yi and Mimno, 2023). Specifically, a token is considered valid if all its characters belong to the Unicode set defined for the target language. We also include numbers, special characters, and the end of sentence tokens in the desired set.

3.2 Perturbation Vector

We construct a perturbation vector, $\mathbf{p}$, based on the set of desired token indices I. Each element corresponding to an index in I is assigned a perturbation value $\alpha \geq 0$, as defined in Equation 1.

$$\mathbf{p}_i = \begin{cases} \alpha & \text{if } i \in I, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

3.3 Logits Perturbation Methods

In this study, we explore two variants of the Logits Perturbation Method: LATB and Adaptive-LATB.

3.3.1 Language-Aware Token Boosting (LATB)

We introduce perturbations to the logits by adding a perturbation value α to the selected logits to align them with the desired language. The method is detailed in Algorithm 1.

Algorithm 1 Vanilla LATB

```
$\mathbf{logits} \leftarrow LLM(x)$
$\mathbf{logits}' \leftarrow \mathbf{logits} + \mathbf{p}$        ▷ Logits Perturbation
$\mathbf{y} \leftarrow \text{Softmax}(\mathbf{logits}')$
```

3.3.2 Adaptive Language-Aware Token Boosting (Adaptive-LATB)

Adding logits in the vanilla LATB may suppress the ability to express tokens in another language when necessary. In contrast, the Adaptive LATB perturbs logits only when the LLM is not confident about the language it intends to express. The confidence difference threshold, controlled by the hyperparameter β ($0 \leq \beta \leq 1$), determines the model's confidence difference threshold in one language over another. This design enables the model to switch languages when it is highly confident. The details of the Adaptive LATB algorithm are provided in Algorithm 2.

Algorithm 2 Adaptive LATB

```
$\textbf{logits} \leftarrow LLM(x)$
$\mathbf{y} \leftarrow \text{Softmax}(\textbf{logits})$
$a \leftarrow \max(\{y_i \mid y_i \in \mathbf{y} \text{ and } i \in I\})$
$b \leftarrow \max(\{y_i \mid y_i \in \mathbf{y} \text{ and } i \notin I\})$
if $|a - b| < \beta$ then
    $\textbf{logits}' \leftarrow \textbf{logits} + \mathbf{p}$    ▷ Logits Perturbation
    $\mathbf{y} \leftarrow \text{Softmax}(\textbf{logits}')$
end if
```

4 Evaluation Metrics

We evaluate the model based on two key aspects: *Language Confusion*, which measures the model's misalignment with the target language, and *Performance*, which assesses the quality of the generated summaries.

4.1 Language Confusion Metrics

We evaluate language confusion at three distinct levels to capture both fine-grained and overall effects: token-level, line-level, and response-level language confusion.

Token-level Language Confusion. We determine each token's language based on its Unicode and calculate token-level misalignment rates for each response. These rates are then averaged across all responses to report the final metric.

Line-level Language Confusion. We segment each response by line and utilize an off-the-shelf language identification (LID) tool, FastText (Grave et al., 2018), to determine the language of each line. We calculate the average language misalignment per response and report the overall average across all responses.

Response-level Language Confusion. We input the entire response into the FastText (Grave et al., 2018) language identification and calculate the average language misalignment across all responses, reporting this as the final metric.

4.2 Performance Metrics

We assess summarization performance using three widely adopted metrics: ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004). These metrics evaluate the overlap of unigrams, bigrams, and longest common subsequences, respectively, between the generated summaries and the reference summaries.

5 Experiments

Models. We use Llama3 8B Instruct (Lai et al., 2023a) as the base English-centric model. To assess our method's effectiveness, we compare it against Suzume 8B Multilingual (Devine, 2024), a multilingual fine-tuned version of Llama3 8B Instruct trained on a multilingual conversational dataset.

Benchmark. We adopt the multilingual summarization XLSUM dataset (Hasan et al., 2021) as the benchmark for our evaluation. This dataset is particularly suitable for our study as it allows models to generate extended responses, which can be systematically evaluated using quantitative metrics.

In our study, we select 4 High Resource Languages (HRL): Russian (ru), Simplified Chinese (zh), Japanese (ja), and French (fr), as well as 4 Medium Resource Languages (MRL): Korean (ko), Thai (th), Hindi (hi), and Arabic (ar). The categorization of languages follows (Lai et al., 2023a). For each language, we utilize the test split from the dataset and sample up to 1,000 examples for evaluation.

Language Confusion Results. We compare our methods with Llama3 8B Instruct (Grattafiori et al., 2024) and Suzume 8B Multilingual (Devine, 2024). Our methods demonstrate effectiveness in reducing language confusion compared to its base model. Furthermore, our methods outperform the multilingual fine-tuned model, highlighting their effectiveness in reducing language confusion without incurring the cost of fine-tuning. The language confusion results are presented in Table 1.

Summarization Quality Results. The results reported in Table 2 demonstrate that our methods generate higher-quality responses compared to both the Llama3 baseline model (Grattafiori et al., 2024) and the Suzume 8B Multilingual model (Devine, 2024) without additional fine-tuning requirements.

6 Analysis

Impact of Hyperparameters. In LATB, increasing α reduces language confusion, with ROUGE scores rising initially before declining. At the α yielding the highest ROUGE scores, the model balances effectively expressing technical terms in English while minimizing language confusion at both line and response levels. Beyond this optimal point,

Table 1: Language confusion across different methods evaluated on eight languages, reported as Token-level/Line-level/Response-level language confusion in percentage.

	Llama3 8B-I	Suzume 8B-Multilingual	Llama3 8B-I + LATB (*Ours*)	Llama3 8B-I + Adaptive LATB (*Ours*)
High Resource Languages (HRL)				
ru	5.02/4.10/2.90	3.04/2.30/2.10	**0.28**/0.44/**0.10**	0.48/**0.38**/**0.10**
zh	14.17/9.69/9.10	7.56/0.89/0.90	**4.78**/**0.00**/**0.00**	5.37/0.10/**0.00**
ja	10.15/9.29/4.16	5.96/0.73/0.67	**3.51**/0.70/**0.11**	4.05/**0.16**/**0.11**
fr	0.26/0.39/**0.20**	0.31/0.37/0.30	**0.11**/0.35/**0.20**	0.18/**0.25**/0.30
Medium Resource Languages (MRL)				
ko	16.72/30.79/27.27	8.28/11.74/12.36	**3.45**/**9.98**/**10.36**	4.56/10.12/11.45
th	3.67/9.80/2.30	2.16/1.16/0.84	0.43/0.18/**0.00**	**0.38**/**0.00**/**0.00**
hi	1.67/8.59/0.40	2.77/3.36/2.50	**0.23**/**0.74**/0.10	0.26/0.89/**0.00**
ar	9.98/11.94/5.60	5.63/2.95/2.60	**0.37**/0.28/**0.00**	0.54/**0.22**/**0.00**

Table 2: Summarization performance across different methods evaluated on eight languages, reported as ROUGE-1/ROUGE-2/ROUGE-L in percentage.

	Llama3 8B-I	Suzume 8B-Multilingual	Llama3 8B-I + LATB (*Ours*)	Llama3 8B-I + Adaptive LATB (*Ours*)
High Resource Languages (HRL)				
ru	20.44/9.26/13.41	19.35/8.32/12.42	20.83/**9.46**/**13.60**	**21.00**/9.42/13.58
zh	19.41/8.99/13.73	19.31/8.59/13.38	**20.70**/**9.44**/**14.64**	20.55/9.28/14.52
ja	26.48/12.43/16.97	26.13/11.73/16.55	27.54/**12.95**/17.70	**27.89**/12.92/**17.89**
fr	19.98/8.90/13.71	18.56/7.89/12.47	**20.13**/**9.05**/**13.74**	19.97/8.89/13.59
Medium Resource Languages (MRL)				
ko	14.66/6.14/10.16	15.30/6.13/10.47	16.41/6.78/11.38	**16.88**/**7.03**/**11.67**
th	29.24/13.99/15.62	28.99/13.29/15.14	29.77/14.07/15.79	**30.97**/**14.74**/**16.41**
hi	29.83/16.41/19.03	27.71/14.78/17.52	29.68/16.41/19.00	**29.77**/**16.41**/**19.05**
ar	19.22/7.46/11.66	19.60/7.09/11.69	**20.44**/**8.02**/**12.45**	19.79/7.62/11.84

higher α values suppress tokens in non-target languages, leading to a performance drop. The effect of α is depicted in Figure 3 in Appendix C.

For Adaptive LATB, higher β values also reduce language confusion, with ROUGE scores improving slightly until an inflection point. Excessively high β values hinder the model's ability to generate tokens in non-target languages, resulting in a slight performance decline. The effect of β is illustrated in Figure 4 in Appendix C.

Performance Improvements. Our analysis highlights a strong correlation between performance improvements from LATB and the degree of language confusion without LATB. This finding suggests that language confusion contributes to performance degradation. By incorporating LATB, we effectively mitigate this issue, leading to performance gains. The relationship is illustrated in Figure 5 in Appendix D.

7 Conclusion

This paper introduces a novel approach to multilingual alignment for English-centric language models through token perturbation techniques. We proposed the Language-Aware Token Boosting (LATB) and its adaptive variant, Adaptive-LATB. Extensive experiments demonstrate that our methods significantly reduce language confusion compared to base model and outperform its multilingual fine-tuned model. This highlights the efficiency and practicality of our approach for enhancing multilingual language model capabilities.

Limitations and Future Work

Our work shows promising results but has several limitations. First, the methods struggle with aligning LLMs to untrained or out-of-vocabulary (OOV) tokens. Second, reliance on Unicode-based language identification is less effective for languages with significant overlap with Latin scripts. Finally, hyperparameter tuning is needed to balance language confusion and multilingual expression. Future work could improve OOV token handling, develop better token-based language identification techniques, and design language-agnostic hyperparameter selection methods.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, et al. 2024. Larabench: Benchmarking arabic ai with large language models. *Preprint*, arXiv:2305.14982.

Arav Agarwal, Karthik Mittal, Aidan Doyle, Pragnya Sridhar, Zipiao Wan, Jacob Arthur Doughty, Jaromir Savelka, and Majd Sakr. 2024. Understanding the role of temperature in diverse question generation by gpt-4. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, SIGCSE 2024, page 1550–1551. ACM.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Simone Conia, Min Li, Daniel Lee, Umar Farooq Minhas, Ihab Ilyas, and Yunyao Li. 2023. Increasing coverage and precision of textual information in multilingual knowledge graphs. *Preprint*, arXiv:2311.15781.

Peter Devine. 2024. Tagengo: A multilingual chat dataset. *Preprint*, arXiv:2405.12612.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

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

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. 2024. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *Preprint*, arXiv:2106.13822.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *Preprint*, arXiv:2305.04118.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. *Preprint*, arXiv:2403.03814.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *Preprint*, arXiv:2304.05613.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *Preprint*, arXiv:2307.16039.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *Preprint*, arXiv:2006.16668.

Tianjian Li and Kenton Murray. 2023. Why does zero-shot cross-lingual generation fail? an explanation and a solution. *Preprint*, arXiv:2305.17325.

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.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. *Preprint*, arXiv:2305.06575.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in llms. *Preprint*, arXiv:2406.20052.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *Preprint*, arXiv:2312.13951.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. 2023. Decomposed prompting for machine translation between related languages using large language models. *Preprint*, arXiv:2305.13085.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models. *Preprint*, arXiv:2402.05201.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022. Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. *Preprint*, arXiv:2209.09900.

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

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. *Preprint*, arXiv:2205.12647.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023. Understanding translationese in cross-lingual summarization. *Preprint*, arXiv:2212.07220.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.

Genta Indra 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. *Preprint*, arXiv:2212.09660.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics.

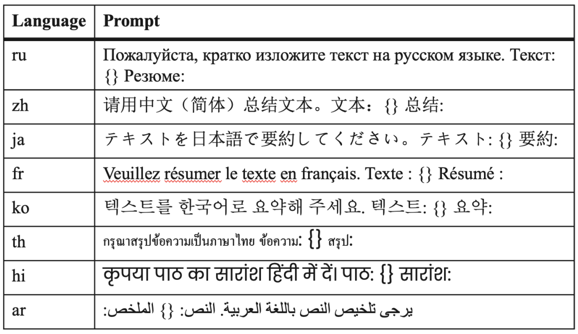
Min Zhang, Limin Liu, Zhao Yanqing, Xiaosong Qiao, Su Chang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Song Peng, Yinglu Li, et al. 2023. Leveraging multilingual knowledge graph to boost domain-specific entity translation of ChatGPT. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 77–87, Macau SAR, China. Asia-Pacific Association for Machine Translation.

A Experiment Details

We generate responses using the Llama3 8B Instruct model (Grattafiori et al., 2024) on eight different languages from the XLSUM dataset (Hasan et al., 2021). The prompts utilized for this experiment are detailed in Appendix B. All responses are generated with the sampling parameters set to a temperature of 1.0 and a top-p value of 1.0. For LATB, the perturbation value α is set to 5. For

Adaptive LATB, the perturbation value is set to $\alpha = 1000$, and the confidence difference threshold is set to $\beta = 0.8$.

B Prompt Templates



Language	Prompt
ru	Пожалуйста, кратко изложите текст на русском языке. Текст: {} Резюме:
zh	请用中文（简体）总结文本。文本：{} 总结:
ja	テキストを日本語で要約してください。テキスト: {} 要約:
fr	Veuillez résumer le texte en français. Texte : {} Résumé :
ko	텍스트를 한국어로 요약해 주세요. 텍스트: {} 요약:
th	กรุณาสรุปข้อความเป็นภาษาไทย ข้อความ: {} สรุป:
hi	कृपया पाठ का सारांश हिंदी में दें। पाठ: {} सारांश:
ar	يرجى تلخيص النص باللغة العربية. النص: {} الملخص:

Figure 2: Prompt templates used in the experiment

We design language-specific prompt templates to ensure consistency and adaptability across different languages during text generation. Each template provides a structured format where { } is replaced by the input text to summarize. The prompt templates are shown in Figure 2.

C Impacts of Hyperparameters

In LATB, the perturbation parameter α influences the generated responses. To analyze its impact, we varied α from 0 to 50 and recorded the corresponding results. These results are presented in Figure 3, illustrating the effect of α on language confusion and summarization quality.

In Adaptive-LATB, we investigated the influence of the confidence difference threshold, denoted as β. Specifically, we varied β from 0 to 0.9 while keeping the perturbation value α fixed at 1000. The outcomes of this experiment are visualized in Figure 4, highlighting how changes in β affect language confusion and summarization quality.

All responses across the experiments were generated using a temperature setting of 1.0 and a top-p value of 1.0, ensuring consistency in sampling parameters throughout the evaluations.

D Performance Improvements

The relationship is illustrated in Figure 5, which demonstrates a strong correlation between performance improvements using LATB and language confusion.

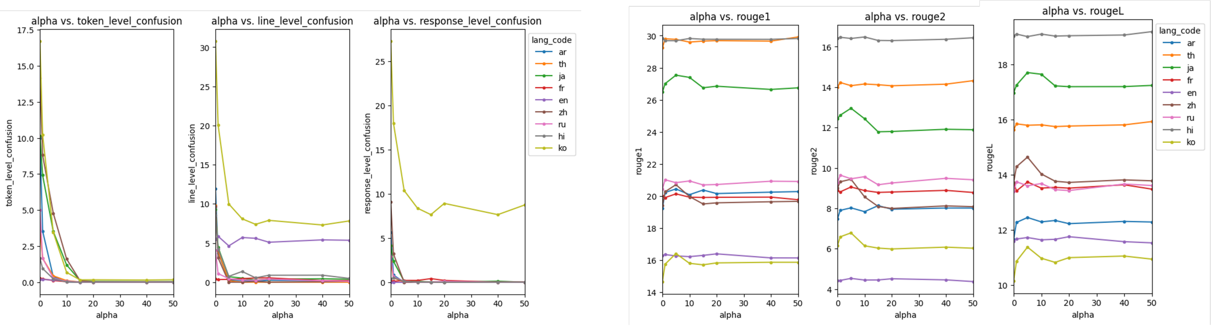


Figure 3: Impact of the Perturbation Value α on Language Confusion and Performance in LATB

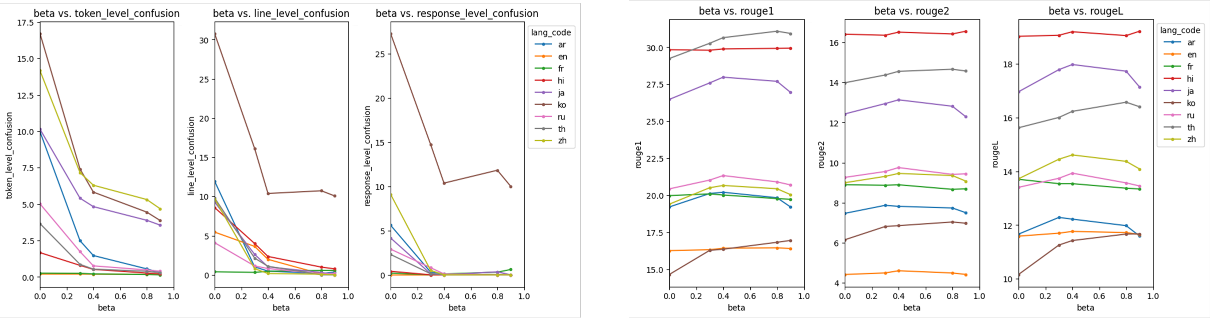


Figure 4: Impact of the Confidence Difference Threshold β on Language Confusion and Performance in Adaptive-LATB with α fixed at 1000

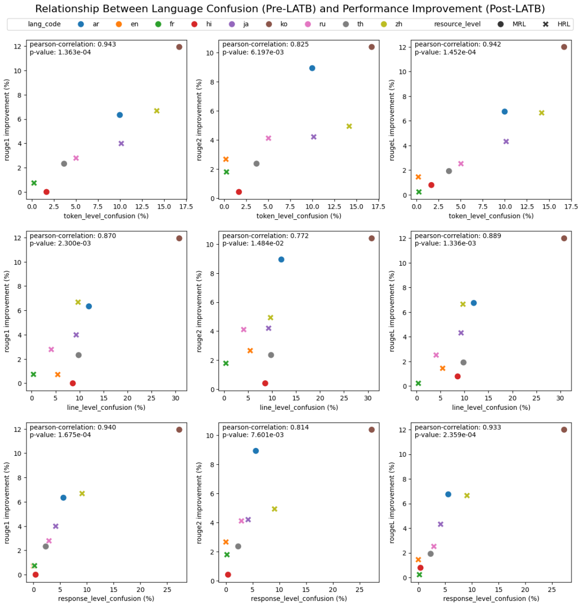


Figure 5: Performance improvements with LATB correlate strongly with language confusion levels