Language-Aware Token Boosting: LLM Language Confusion Reduction Without Tuning

Trapoom Ukarapol[§‡], **Pakhapoom Sarapat**[§], **Nut Chukamphaeng**[§†*]
[§]SCB DataX, [‡]Tsinghua University, [†]SCBX
{trapoom.ukarapol, pakhapoom.sarapat}@data-x.ai, nut.c@scbx.com

Abstract

Large language models (LLMs) sometimes exhibit language confusion when generating non-English text. Existing approaches typically rely on fine-tuning to mitigate this issue. In contrast, we propose a tuning-free paradigm for reducing language confusion. Within this paradigm, we introduce two methods: Language-Aware Token Boosting (LATB), which applies targeted perturbations to tokens associated with the desired language, and Adaptive Language-Aware Token Boosting (Adaptive-LATB), which dynamically adjusts these perturbations based on the model's confidence in the intended language. Experiments demonstrate that our methods effectively improve multilingual alignment by reducing language confusion, while maintain the summarization quality without requiring any additional fine-tuning. Our code is publicly available.[1].

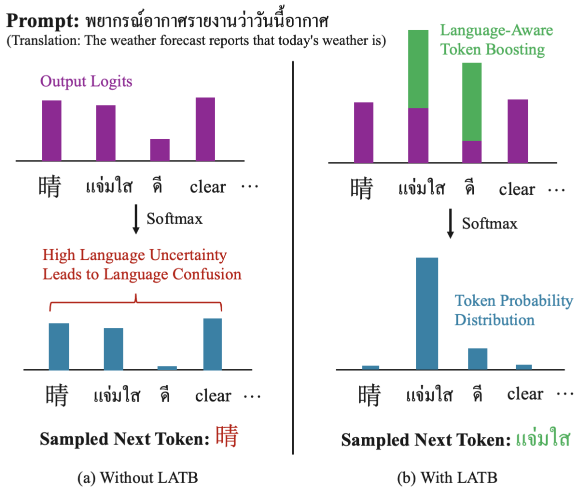

Figure 1: Language-Aware Token Boosting (LATB) enhances target language generation confidence by selectively boosting target language tokens.

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 maintain summarization performance compared to their respective base models and the

[*]This work was carried out during the author's tenure at SCB DataX.
[1]https://github.com/scbdatax/genai-datax-language-aware-token-boosting

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 481–489
July 2-7, 2026

multilingual-tuned model.

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 maintain summarization quality.

3. We analyze the effect of logit perturbations and show that adding a constant value to the logits does not change the relative probability ratios among the selected tokens.

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 a range of NLP tasks, including machine translation (Vu et al., 2022; Li and Murray, 2023), summarization (Wang et al., 2023; Yu et al., 2022), question answering (Holtermann et al., 2024), and even within the reasoning traces of reasoning language models (RLMs) (Wang et al., 2025; Tam et al., 2025). 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

```
logits ← LLM(x)
logits′ ← logits + p        ▷ Logits Perturbation
y′ ← Softmax(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

```
logits ← LLM(x)
y ← Softmax(logits)
a ← max({y_i | y_i ∈ y and i ∈ I})
b ← max({y_i | y_i ∈ y and i ∉ I})
if |a − b| < β then
    logits′ ← logits + p        ▷ Logits Perturbation
    y′ ← Softmax(logits′)
else
    y′ ← y
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 (Joulin et al., 2016b,a), 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 (Joulin et al., 2016b,a) 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

Experimental Setup. We compare Llama3 8B Instruct (Lai et al., 2023a) under several configurations: (i) a *normal prompt* in the target generation language without any language constraints in the prompt, (ii) a *strict prompt* that explicitly specifies the target generation language, (iii) a multilingual supervised fine-tuned variant of Llama3 8B Instruct, namely Suzume 8B Multilingual (Devine, 2024), and (iv) our proposed tuning-free methods, Language-Aware Token Boosting (LATB) and Adaptive-LATB. Llama3 8B Instruct serves as the base English-centric model, while Suzume 8B Multilingual provides a strong multilingual fine-tuning baseline for comparison.

For evaluation, we adopt the multilingual abstractive summarization benchmark XLSUM (Hasan et al., 2021). This dataset is well suited to our study as it requires models to generate long-form outputs, enabling systematic quantitative evaluation. We consider eight languages spanning different resource levels, including four High-Resource Languages (HRL): Russian (ru), Simplified Chinese (zh), Japanese (ja), and French (fr), and four Medium-Resource Languages (MRL): Korean (ko), Thai (th), Hindi (hi), and Arabic (ar). The classification of language resource levels follows (Lai et al., 2023a). For each language, we randomly sample up to 1,000 test instances for evaluation.

Summarization Quality Results. The results reported in Table 2 demonstrate that our methods preserve generation quality and yield marginal improvements, with gains increasing as the degree of language confusion rises prior to applying LATB, as detailed in Appendix F.

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	Llama3 8B-I (Strict Prompt)	Suzume 8B-Multilingual	Llama3 8B-I + LATB (*Ours*)	Llama3 8B-I + Adaptive LATB (*Ours*)
			High Resource Languages (HRL)		
ru	80.66/93.83/92.50	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	85.92/98.90/98.90	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	85.10/98.83/98.31	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	0.24/47.14/40.6	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	85.46/99.60/99.63	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	86.03/99.80/99.39	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	86.18/99.05/98.50	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	86.21/99.27/98.30	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	Llama3 8B-I (Strict Prompt)	Suzume 8B-Multilingual	Llama3 8B-I + LATB (*Ours*)	Llama3 8B-I + Adaptive LATB (*Ours*)
			High Resource Languages (HRL)		
ru	4.89/0.96/4.18	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	0.80/0.32/0.69	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	26.42/12.53/16.84	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	14.71/6.09/10.49	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	2.27/0.24/2.12	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	1.79/0.45/1.51	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	0.93/0.36/0.73	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	1.54/0.19/1.42	19.22/7.46/11.66	19.60/7.09/11.69	**20.44/8.02/12.45**	19.79/7.62/11.84

6 Analysis

Effect of Logit Perturbations on Boosted Token Probabilities. Let I denote the set of boosted tokens. Let y_i and y_i' be the probabilities of the i-th token before and after logit perturbation, respectively. Denote the logit of the i-th token by v_i, and let α be the perturbation magnitude applied to all tokens in I. We analyze the probability mass function over the boosted tokens before and after token boosting. Consider any two tokens $m, n \in I$.

$$\frac{y_m'}{y_n'} = \frac{(e^{v_m+\alpha}) \,/\, \left(\sum_{i\notin I} e^{v_i} + \sum_{i\in I} e^{v_i+\alpha}\right)}{(e^{v_n+\alpha}) \,/\, \left(\sum_{i\notin I} e^{v_i} + \sum_{i\in I} e^{v_i+\alpha}\right)} \tag{2}$$

$$= \frac{e^{v_m}}{e^{v_n}} = \frac{e^{v_m} / \sum_i e^{v_i}}{e^{v_n} / \sum_i e^{v_i}} = \frac{y_m}{y_n}. \tag{3}$$

Therefore, the logit perturbation preserves the relative probability ratios between any pair of boosted tokens.

Vanilla vs. Adaptive LATB. Both methods deliver comparable performance. However, Vanilla LATB requires an optimal hyperparameter search to produce non-target language output when needed while accurately generating results in the target language. In contrast, Adaptive LATB is less sensitive to hyperparameters and supports non-target language output as required.

Impact on Inference Speed. We evaluated inference throughput of Llama3 8B Instruct (Lai et al., 2023a) with vLLM on an A100 GPU. The base model achieves 1145.8 tokens/s, while vanilla LATB achieves 1189.5 tokens/s effectively identical throughput within measurement noise. This is expected, as LATB only applies a lightweight logit bias without additional branching or search. Adaptive-LATB reaches 838 tokens/s, reflecting its extra computation for per-step maximum probabilities difference detection. Although this introduces overhead, the throughput remains within a practical range for deployment.

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.

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.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? *Preprint*, arXiv:2505.17407.

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.

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

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

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 strict prompt templates include instructions to ensure the model generates output in the target language, whereas the standard prompt templates do not. The standard and strict prompt templates are shown in Figures 3 and 2, respectively.

Language	Strict 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: Strict prompt templates used in the experiment

Language	Standard Prompt
ru	Пожалуйста, кратко изложите текст. Текст: {} Резюме:
zh	请总结文本。文本：{} 总结
ja	テキストを要約してください。テキスト: {} 要約:
fr	Veuillez résumer le texte. Texte : {} Résumé :
ko	텍스트를 요약해 주세요. 텍스트: {} 요약:
th	กรุณาสรุปข้อความ ข้อความ: {} สรุป:
hi	कृपया पाठ का सारांश दें। पाठ: {} सारांश:
ar	يرجى تلخيص النص. النص: {} الملخص:

Figure 3: Standard prompt templates used in the experiment

C Impact of Hyperparameters

We analyze the impact of hyperparameters in both LATB and Adaptive-LATB on language confusion and summarization quality. All responses across experiments were generated with a temperature of 1.0 and a top-p value of 1.0 to ensure consistent sampling.

LATB. In LATB, the perturbation parameter α plays a critical role in controlling language confusion. We varied α from 0 to 50 and report the results in Figure 4. As α increases, language confusion is progressively reduced, leading to an initial improvement in ROUGE scores. At an intermediate value of α, the model achieves an effective balance: it can express technical terms in English while minimizing language confusion at both the line and response levels. Beyond this optimal point, further increasing α overly suppresses tokens from non-target languages, which degrades summarization quality and results in a decline in ROUGE scores.

Adaptive-LATB. For Adaptive-LATB, we study the effect of the confidence difference threshold β while fixing the perturbation value α to 1000. We varied β from 0 to 0.9, with results shown

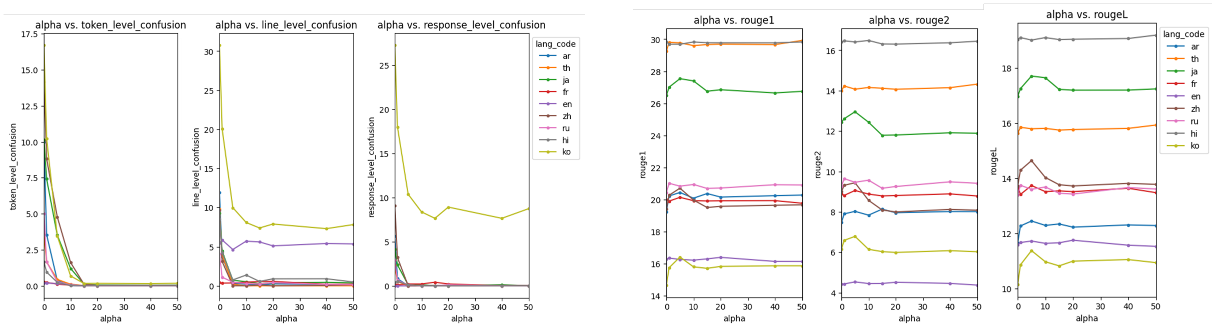

Figure 4: Impact of the Perturbation Value α on Language Confusion and Performance in LATB

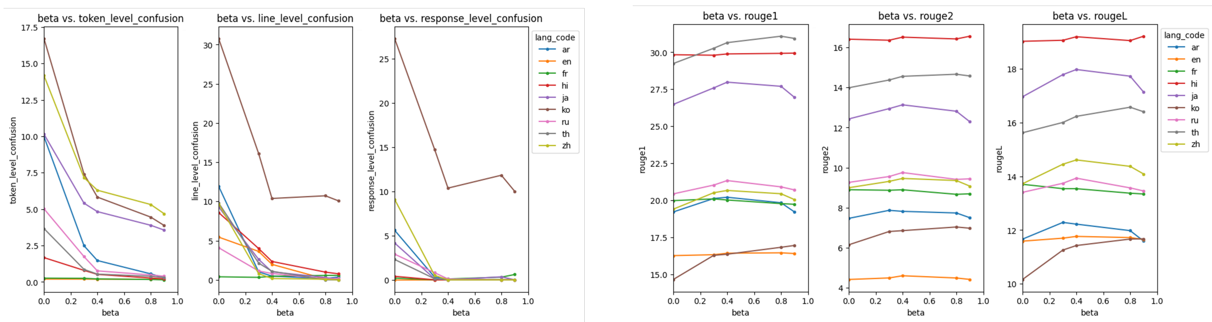

Figure 5: Impact of the Confidence Difference Threshold β on Language Confusion and Performance in Adaptive-LATB with α fixed at 1000

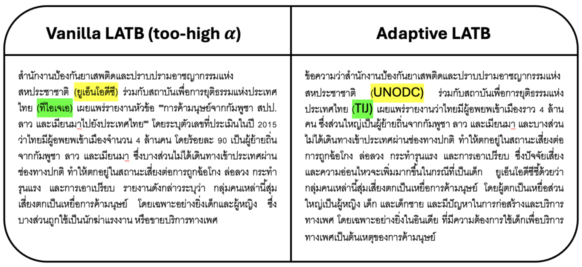

Figure 6: Output examples of Vanilla LATB with excessively high α and Adaptive LATB

in Figure 5. Increasing β generally reduces language confusion and yields slight improvements in ROUGE scores up to an inflection point. However, excessively large β values restrict the model's ability to generate necessary non-target language tokens, leading to a mild performance drop. This trend highlights a similar trade-off between reducing language confusion and preserving generation flexibility.

D Vanilla vs. Adaptive LATB Example

As discussed, Vanilla LATB is sensitive to hyperparameters, which can lead to a constraint on single-target language generation when α is too high. In contrast, Adaptive LATB is less sensitive to hyperparameters. The example output is shown in Figure 6. When α is too high, Vanilla LATB generates a Thai-dubbed version of English, whereas Adaptive LATB uses English directly, resulting in a more natural output.

E Additional Experiments on a Stronger Multilingual Model

We extend our experiments to Qwen3-4B-I (Yang et al., 2025) and measure response-level language confusion and summarization quality using the ROUGE-L metric. The base model already demonstrates strong language alignment with no observed errors, and our method preserves this behavior: it achieves similarly error-free language alignment.

Importantly, our method does not degrade task performance. As shown in Eq. 3, LATB preserves the relative probability ratios among tokens within the same language; therefore, intra-language token ranking and consequently generation quality remains unchanged. Consistent with this theoretical property, we observe no loss in ROUGE-L compared to the base model. The results are shown in Table 3.

F Performance Improvements

Our analysis highlights a strong correlation between performance improvements from LATB and

Table 3: Qwen3 4B-I Language confusion and Summarization performance across different methods evaluated on eight languages reported as Response-level language confusion/ROUGE-L in percentage.

	Qwen3 4B-I	Qwen3 4B-I (Strict Prompt)	Qwen3 4B-I + LATB (*Ours*)	Qwen3 4B-I + Adaptive LATB (*Ours*)
High Resource Languages (HRL)				
ru	0.00/13.45	0.00/13.16	0.00/13.30	0.00/12.68
zh	0.00/12.42	0.00/12.58	0.00/12.49	0.00 /12.26
ja	0.00/17.63	0.00/17.01	0.00/16.88	0.00/16.90
fr	0.00/13.33	0.00/12.99	0.00/13.02	0.00/12.74
Medium Resource Languages (MRL)				
ko	0.00/12.34	0.00/11.78	0.00/11.51	0.00/11.55
th	0.00/18.64	0.00/17.77	0.00/17.60	0.00/17.13
hi	0.00/21.81	0.00/21.58	0.00/16.57	0.00/18.42
ar	0.00/11.67	0.00/10.94	0.00/11.14	0.00/10.06

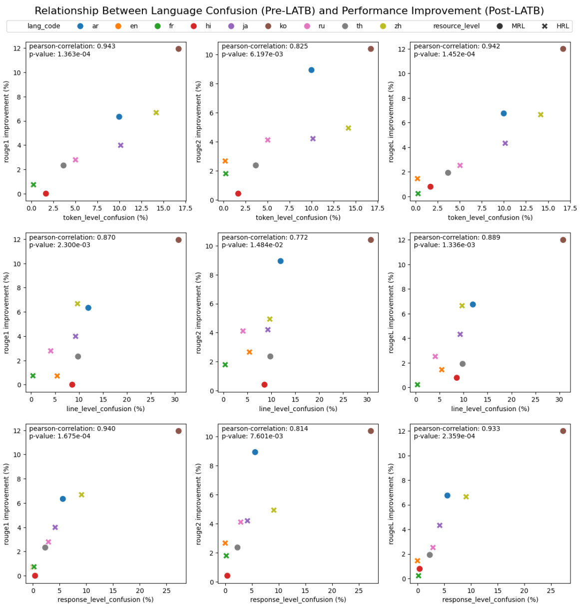

Figure 7: Performance improvements with LATB correlate strongly with language confusion levels

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 7.