Language Confusion and Multilingual Performance: A Case Study of Thai-Adapted Large Language Models

Anonymous ACL submission

Abstract

This paper investigates the code-switching problem between English and Thai languages in large language models (LLMs), especially those encountered the continual pre-training process (CPT) and those initially trained with multilingual data, called multilingual LLMs (MLLMs). We change the language in the task instruction, context, and output language of the prompt to examine the effects of the language variation settings on the code-switched language in the responses for different model types. Our findings show that mismatches between context and output languages result in significant performance degradation in all the model types and they achieve similar performance for monolingual settings, while MLLMs show improvements on the cross-lingual settings. It suggests that given high cost of multilingual training from scratch, we might still need MLLMs for downstream tasks in languages other than English due to their multilingual capability which is better than CPT models and models trained without any multilingual interventions.

1 Introduction

A code-switched language has been a topic to discussed and studied in natural language generation for decades. It is a situation when a sentence in a model's response contains multiple languages (Poplack, 1980; Khanuja et al., 2020) or models are so *confused* that they fail to generate a consistent response in a particular language (Marchisio et al., 2024). This phenomenon has become ubiquitous since the rise of LLMs (Brown et al., 2020) because most of them are still predominantly English-centric, with limited capabilities when it comes to other languages (Asai et al., 2024; Bang et al., 2023), while a significant number of people across the world use languages that LLMs have difficulty understanding or processing.

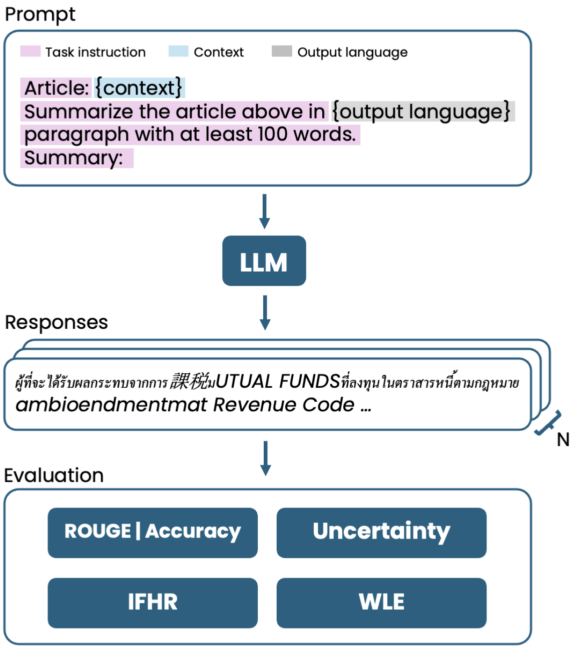


Figure 1: Example of language variation settings. The languages in task instruction (pink), context (blue), and output (gray) can be varied from English to Thai, and the whole prompt is fed to an LLM for N times to measure multilingual performance in terms of ROUGE-1 for long-form generation task, and accuracy for short-form generation task, as well as uncertainty, instruction-following hallucination rate (IFHR), and word-level entropy (WLE).

Several techniques have been proposed to localize those English-centric LLMs to work better in target languages including parameter-tuning alignment and parameter-frozen alignment (Qin et al., 2024). However, all adaptation strategies still give rise to the code-switching issue as some researchers investigate the code-switched language or language confusion over 15 languages with monolingual and cross-lingual generation and measure model's responses in word-level and line-level confusion. They find that LLMs are susceptible to language confusion when the number of tokens in the sampling nucleus is high, while the distribution is flat (Marchisio et al., 2024).

In this study, we follow a similar study of the language confusion by pushing further to vary the language in different parts of the prompt, namely task instruction, context, and output language, as visualized in Fig 1, with an extensive focus on Thai language as a case study to investigate the generalization of LLMs beyond English through different pre-training strategies because this language is considered one of the low-resource languages with complex orthography (Pipatanakul et al., 2023). We also explore and compare the language confusion with regard to different confusion aspects, such as uncertainty (Farquhar et al., 2024), instruction-following hallucination (IFHR), and word-level entropy (WLE). Besides, we measure the response quality through performance metrics, such as accuracy and ROUGE-1 across different tasks, including both short-form and long-form generation tasks.

2 Background and Problem Setting

Our work relates to code-switching or language confusion, specifically for Thai and English, in different types of LLMs. We describe the relevant background and present our research question on the language confusion in LLMs.

Multilinguality adaptation strategy There are two main approaches to enhance capability in the target languages which are parameter-tuning alignment and parameter-frozen alignment (Qin et al., 2024). For the parameter-tuning alignment, it refers to fine-tuning process with target language data during from-scratch pre-training (Brown et al., 2020), continual pre-training (CPT) (Luukkonen et al., 2023), supervised fine-tuning (SFT) (Chung et al., 2022), reinforcement learning with human feedback (RLHF) (Lai et al., 2023), and downstream fine-tuning (Lepikhin et al., 2020) with additional language-specific data to the original LLMs, while the parameter-frozen alignment requires prompt engineering without updating model parameters to acquire multilingual performance (Yang et al., 2023). In this study, we focus on the first approach. However, due to the expensive resources required for the fine-tuning process, the practical approach for Thai adaptation is limited to the CPT approach, such as Typhoon1.5 (Pipatanakul et al., 2023), Sailor (Dou et al., 2024) and OpenThaiGPT1.5 (Yuenyong et al., 2024).

Language confusion We define *language confusion* as a situation in which a model experiences difficulty processing the information from the prompt, resulting in the generation of a response that incorporates unintended languages (Khanuja et al., 2020; Marchisio et al., 2024) or does not follow the provided instruction. This occurs because the prompt itself varies between Thai and English as displayed in Fig 2.

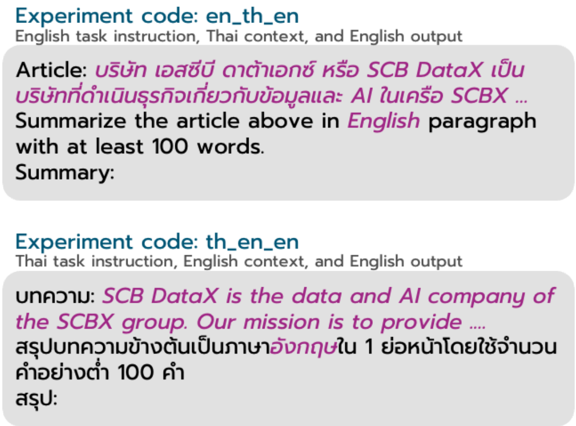


Figure 2: Prompt examples for a summarization task.

Problem statement We frame the problem as a research question: *how does changing the language in the prompt, which we separate into task instruction, context, and output language, affect the model's performance?* The study investigates the phenomenon of language confusion in LLMs that underwent CPT with Thai language data, comparing their results to the base models as well as to MLLMs.

3 Language Confusion Experiments

We examine the language confusion in LLMs through two main tasks: short-form (multiple-choice) and long-form generation (long-context question answering and summarization) tasks.

Models The scope of models studied here includes 8B-Llama3 (Grattafiori et al., 2024) and its CPT with Thai data, 8B-Typhoon1.5 (Pipatanakul et al., 2023), 7B-Qwen1.5 (Bai et al., 2023) with its CPT, 7B-Sailor1 (Dou et al., 2024), and 7B-Qwen2.5 (Yang et al., 2025) with its CPT, 7B-OpenThaiGPT1.5 (Yuenyong et al., 2024). We also include 9B-Gemma2 (Riviere et al., 2024) and 8B-Llama3.1 (Grattafiori et al., 2024) for comparison to MLLMs.

Benchmarks We use a high-quality dataset curated for instruction-following fine-tuning, `WangchanThaiInstruct` (Vistec, 2024). We select three tasks from this dataset for multiple-choice task, as well as closed QA and summarization for long-form generation tasks.

Additionally, we incorporate a popular benchmark within Thai LLMs community, `ThaiExam` (Pipatanakul et al., 2023), and include a universal benchmark, `MMLU` (Hendrycks et al., 2021), to serve as a baseline dataset for benchmarking model performance in the experiments. Both tasks discussed are in a multiple-choice format.

For `WangchanThaiInstruct` and `ThaiExam`, they are originally in Thai and are translated into English, while `MMLU` is in English initially and is translataed into Thai. The translations are carried out using GPT-4 (Achiam et al., 2024), and some are sampled to manually check and revise, if needed, by authors.

Experiment settings In each question of each benchmark, the languages in the instruction and context vary. Additionally, the output language can vary for long-form generation tasks, which is labeled in the following format: `{task_instruction}_{context}_{output}` as shown in Fig 2. However, for short-form generation task, the format of each experiment will exclude the output part because it will be limited to one of the options from `A` to `E`. We generate $N = 10$ responses for each experiment to reduce the influence of randomness of the generated responses.

Evaluation metrics We measure language confusion from three perspectives: (i) *Uncertainty* (Farquhar et al., 2024) – to assess the consistency of the N responses, (ii) *Instruction-following hallucination rate (IFHR)* – to evaluate how well the model understands the task instructions. For short-form generation tasks, this focuses on whether the response matches one of the options in the multiple-choice set. For long-form generation tasks, the focus is on whether the response is in the specified language. The language identification in this experiment will use the `FastText` (Grave et al., 2018), a language identification model, to determine the language of the generated response, and (iii) *Word-level entropy (WLE)* - to determine the uncertainty at the word level of each response by using the `PyThai` tokenizer (Phatthiyaphaibun et al., 2024) to tokenize a response into words and input them to the same language identification model to identify their language. The resulting values are used to compute entropy, and it should be noted that this metric is only available for long-form generation tasks.

In addition to the three language confusion metrics, we also measure performance to evaluate model proficiency in each task. Accuracy is used for short-form generation tasks, while ROUGE-1 (Lin, 2004) is used for long-form generation tasks.

4 Results

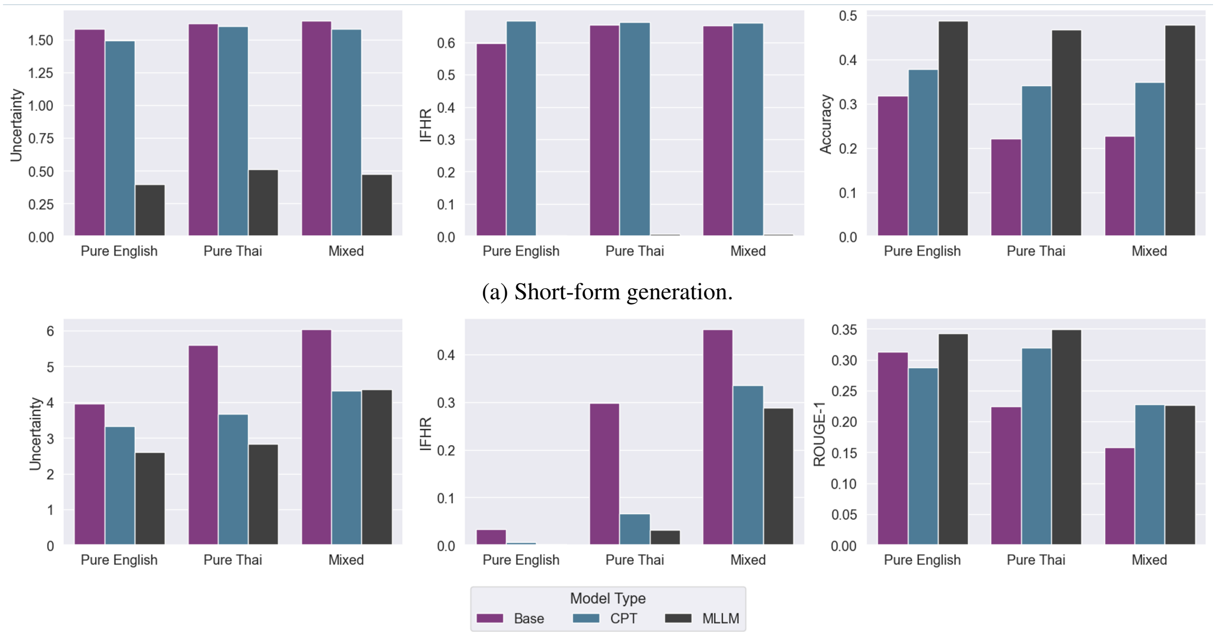
We evaluate the responses of each experiment and model individually and aggregate them based on their model type, which is either base, CPT, or MLLM, and experiment type, which is either pure English (all components in the prompt are in English), pure Thai (all in Thai), or mixed, as shown in Fig 3.

Short-form generation tasks Fig 3(a) shows that all performances, ranging from uncertainty, IFHR, and accuracy, of each model type remain similar when we vary the language in the task instruction and context of the prompt. This is because the expected response is just one single character between `A` to `E`, so the language variations may not have much influence on the short-form generation tasks.

However, we observe that the base and CPT models behave similarly in terms of uncertainty and IFHR, while MLLMs provide unique pattern in the language variation settings. The base and CPT models provide inconsistent responses, as their uncertainty is very high (see Fig 3(a)-left) and they do not follow the instruction well although there is a slight decrease of IFHR for the base models in `Pure English` setting (see Fig 3(a)-mid). Unlikely to MLLMs, they can better generate consistent resposne as well as understand the instruction to generate valid responses due to almost-zero IFHR.

For the accuracy as plotted in Fig 3(a)-right, We notice a greater distinction between the base and CPT models due to the higher accuracy contributed by the CPT models. However, their performance is still lower than that of MLLMs, which achieve the best performance in terms of the highest accuracy across all experiment types.

Long-form generation tasks The impact of language confusion becomes more prominent when the models generate longer responses. All model types provide their best performance at `Pure`



(a) Short-form generation.

(b) Long-form generation.

Figure 3: Performance of base, CPT, and MLLM models for (a) short-form and (b) long-form generation tasks breakdown by experiment types.

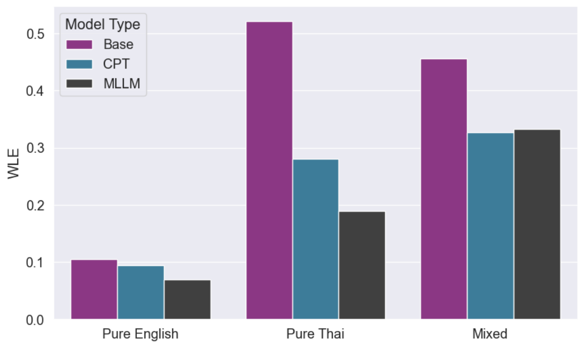


Figure 4: Word-level entropy (WLE) for long-form generation tasks of different model types.

`English`, followed by `Pure Thai`, and their performance deteriorates when the prompt contains mixed languages as illustrated in Fig 3(b).

Surprisingly, the base models show language confusion even in `Pure English` experiment, and they fail to generate a response in the target language once we introduce Thai language in the prompt, while the CPT and MLLMs are more likely to handle Thai language better. However, IFHR skyrockets when there are language mismatches between the context and output as presented in Fig 3(b)-mid. Since the models do not often follow instructions, they generate inconsistent responses, leading to an increase in uncertainty as shown in Fig 3.(b)-left.

Moreover, WLE of all model types increases significantly, but the base's WLE rises the most, while MLLMs are able to maintain the best WLE as visualized in Fig 4. However, once the prompt language is mixed, the WLE of CPT is at the same level as MLLMs. This pattern also persists from the performance perspective in Fig 3(b)-right, where the base models are good only at English language and their ROUGE-1 decreases for `Pure Thai` and `Mixed` settings. On the other hand, CPT and MLLMs can maintain their ROUGE-1 as we vary the prompt languages. However, MLLM achieve the best performance according to the highest ROUGE-1 for each experiment settings.

5 Conclusion

Models with continual pre-training strategy show improvements for both language confusion and performance metrics in a target language or cross-lingual settings when compared to their base models. However, their performance is still inferior to MLLMs because they do not fully acquire multilingual capabilities and struggle for the mismatched language settings. It is essential to incorporate multilingual training strategy to derive more robust multilingual skills and to enhance model generalization in cross-lingual downstream tasks.

Limitations

This study focuses on the Thai language as a case study to explore the generalization of large language models (LLMs) to languages beyond English. Due to computational constraints and the limited availability of multilingual performance benchmarks, the analysis incorporates a small sample of model pairs with model size around 7B parameters, which may affect the comprehensiveness of the comparison.

References

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

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

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.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Xin Mao, Ziqi Jin, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-East Asia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal1. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*.

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 Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

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

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

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.

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.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, et al. 2023. Fingpt: Large generative models for a small language. *Preprint*, arXiv:2311.05640.

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

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, and Pattarawat Chormai. 2024. PyThaiNLP: Thai natural language processing in Python.

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.

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espa~ nol: toward a typology of code-switching. *Linguistics*, pages 581–618.

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.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Vistec. 2024. Wangchanthaiinstruct: Human-annotated thai instruction dataset.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Chih-Kai Yang, Kuan-Po Huang, Ke-Han Lu, Chun-Yi Kuan, Chi-Yuan Hsiao, and Hung yi Lee. 2023. Investigating zero-shot generalizability on mandarin-english code-switched asr and speech-to-text translation of recent foundation models with self-supervision and weak supervision. *Preprint*, arXiv:2401.00273.

Sumeth Yuenyong, Kobkrit Viriyayudhakorn, Apivadee Piyatumrong, and Jillaphat Jaroenkantasima. 2024. Openthaigpt 1.5: A thai-centric open source large language model. *Preprint*, arXiv:2411.07238.