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

Pakhapoom Sarapat
SCB DataX
pakhapoom.sarapat@data-x.ai

Trapoom Ukarapol
SCB DataX
Tsinghua University
ukarapolt10@mails.tsinghua.edu.cn

Tatsunori Hashimoto
Stanford University
thashim@stanford.edu

Abstract

This paper presents a comprehensive study on the multilingual adaptability of large language models (LLMs), with a focus on the interplay between training strategies and prompt design. Using Thai as a case study, we examine: **(RQ1)** the extent to which pre-trained models (Base) can adapt to another language through additional fine-tuning; **(RQ2)** how continual pre-training (CPT) compares to multilingual pre-training (MLLM) in terms of performance on downstream tasks; and **(RQ3)** how language variation within different components of a structured prompt–*task instruction*, *context input*, and *output instruction*–influences task performance in cross-lingual settings. Our findings reveal that CPT proves to be a promising strategy for enhancing model performance in languages other than English like Thai in monolingual settings, particularly for models that initially lack strong linguistic capabilities. Its effectiveness, however, is highly task-dependent and varies based on the base model's initial proficiency. In cross-lingual scenarios, MLLMs exhibit superior robustness compared to Base and CPT models, which are more susceptible to context-output language mismatches. Considering the high cost of training multilingual models from scratch, MLLMs remain a critical component for downstream tasks in multilingual settings due to their strong cross-lingual performance.[1]

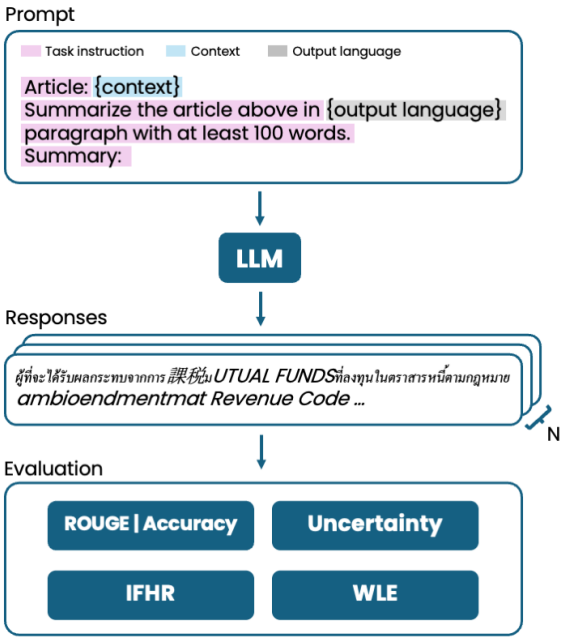


Figure 1: Example of language variation settings. The languages used in the task instruction (pink), context (blue), and output (gray) can vary between English and Thai. The entire prompt is provided to the LLM N times to evaluate multilingual performance. This evaluation includes confusion-related metrics, such as instruction-following hallucination rate (IFHR), uncertainty, and word-level entropy (WLE), as well as performance-related metrics, such as accuracy for short-form generation tasks and ROUGE-1 for long-form generation tasks.

1 Introduction

A code-switched language has been a topic 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 language 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. They also show limited capabilities when it comes to other languages (Asai et al., 2024; Bang et al., 2023).

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. Some researchers investigate the code-switched language,

[1]We release our code at SCB DataX's GitHub.

Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025), pages 49–59

also known as language confusion, over 15 languages with monolingual and cross-lingual generation and measure model's responses in word-level and line-level confusion (Marchisio et al., 2024). 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.

In this study, we follow a similar study of language confusion by pushing further with an extensive focus on Thai language as a case study. We investigate the generalization of LLMs beyond English through both monolingual and cross-lingual settings on different training strategies, namely (i) **Base** – training from scratch with English-dominant data, (ii) **CPT** – continual pre-training of the Base model on data in a target language, and (iii) **MLLM** – multilingual pre-training. We also examine the effectiveness of fine-tuning pre-trained models on a new language and compare it with alternative training strategies. In addition, we investigate how variations in the language used across different parts of a prompt including task instruction, context input, and output instruction, impact model performance in multilingual and code-switched settings, as visualized in Fig 1.

It is noted Thai language is selected because it represents a language that has recently transitioned from being low-resourced to medium-resourced (Joshi et al., 2020). This shift offers a unique opportunity to investigate how language resource availability influences model performance and generalization. Moreover, the availability of base, CPT, and MLLM variants in Thai enables direct, controlled comparisons across training strategies. We also explore and compare the language confusion with regard to different confusion aspects, such as uncertainty (Farquhar et al., 2024), instruction-following hallucination rate (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 Related work

This work investigates code-switching and language confusion between Thai and English in different types of LLMs. We begin by outlining the relevant background.

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. In contrast, 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 Typhoon-1.5 (Pipatanakul et al., 2023), Sailor (Dou et al., 2024), and OpenThaiGPT-1.5 (Yuenyong et al., 2024).

Language confusion in LLMs We define *language confusion* as a situation in which a model struggles to process information from the prompt and generate a response containing unintended languages (Khanuja et al., 2020; Marchisio et al., 2024) or does not follow the provided instruction.

3 Language confusion experiments

This section outlines the experiments conducted to address the following research questions.

- **RQ1:** To what extent can a pre-trained model adapt to a target language through additional fine-tuning?
- **RQ2:** Does sequential training or continual pre-training on a new language improve a pre-trained model's performance in that language more effectively than training from scratch or multilingual pre-training?
- **RQ3:** To what extent does the language used in different parts of a prompt, namely task instruction, context input, and output instruction, as visualized in Fig 1, influence task performance in multilingual settings?

Datasets We use a high-quality Thai dataset curated for instruction-following fine-tuning, WangchanThaiInstruct (Vistec, 2024), denoted as WTI. From this dataset, we select three relevant

tasks, namely multiple-choice (WTI-MC), closed QA (WTI-CQA), and summarization (WTI-SUM) tasks. We also 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 for benchmarking model performance.

For WTI and ThaiExam datasets, they are originally in Thai and are translated into English. 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. Please refer to Appendix A for more details.

We further categorize the datasets into two main tasks: short-form and long-form generation tasks. The short-form generation task includes WTI-MC, ThaiExam, and MMLU, while the long-form generation task includes WTI-CQA and WTI-SUM. The data statistics are provided in Appendix B.

Models Due to the limited compute budget, the scope of the models studied here includes around 7B-9B models, namely Llama-3-8B (Grattafiori et al., 2024) and its CPT with Thai dataset, Typhoon-1.5-8B (Pipatanakul et al., 2023), Qwen-1.5-7B (Bai et al., 2023) with its CPT, Sailor-7B (Dou et al., 2024), and Qwen-2.5-7B (Yang et al., 2025) with its CPT, OpenThaiGPT-1.5-7B (Yuenyong et al., 2024) to address RQ1. We also include Gemma-2-9B (Riviere et al., 2024) and Llama-3.1-8B (Grattafiori et al., 2024) for MLLMs comparison to answer RQ2 and RQ3.

Evaluation metrics We measure language confusion from three perspectives: (i) **Instruction-following hallucination rate (IFHR)** – to evaluate how well the model understands the task instruction. For short-form generation tasks (MMLU, WTI-MC, and ThaiExam), this focuses on whether the response matches one of the valid options in the multiple-choice set. For long-form generation tasks (WTI-SUM and WTI-CQA), the focus is on whether the response is in the specified language. For this experiment, language identification is performed using `FastText` (Grave et al., 2018), a language identification model, to determine the language of the generated response, (ii) **Uncertainty** – to assess the consistency of the N responses quantified using the spectral clustering technique (Farquhar et al., 2024), and (iii) **Word-level entropy (WLE)** – to determine word-level uncertainty in each response. We use the `PyThaiNLP` tokenizer (Phatthiyaphaibun et al., 2024) to segment the response into individual words, which are then passed to the same language identification model to detect their language. The resulting predictions are used to compute entropy. It is important to note that this metric is only applicable to long-form generation tasks.

In addition to the three language confusion metrics, we also evaluate task performance to assess each model's capability in a downstream task. Accuracy[2] is used for short-form generation tasks, while ROUGE-1 (Lin, 2004) is employed for long-form generation tasks.

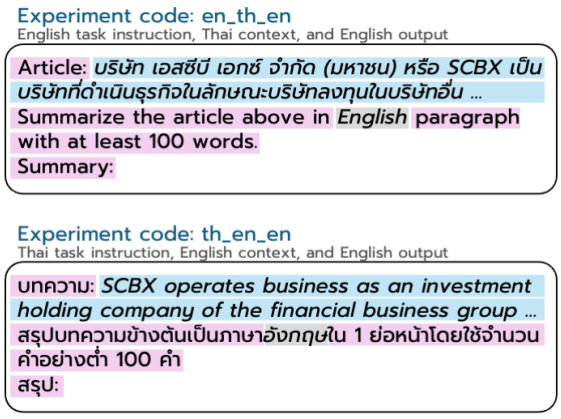


Figure 2: Prompt examples for a summarization task.

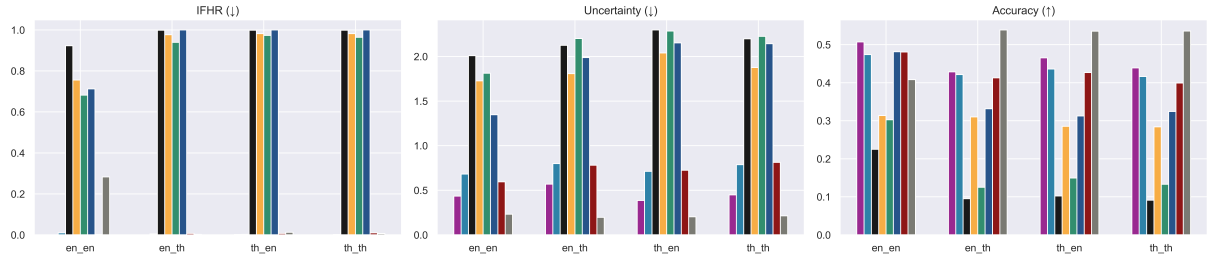
Experimental Setup For each prompt, we vary the language of the task instruction and context input parts by default and the output instruction can be additionally varied for long-form generation tasks, which is labeled in the following format: `{instruction}_{context}_{output}` as shown in Fig 2. However, the format of the short-form experiments excludes the output instruction component because the response is limited to one of the options from `A` to `E`. We generate $N = 10$ responses per prompt to calculate the uncertainty score and aggregate them using the mean for other metrics to obtain prompt-level scores.

4 Results

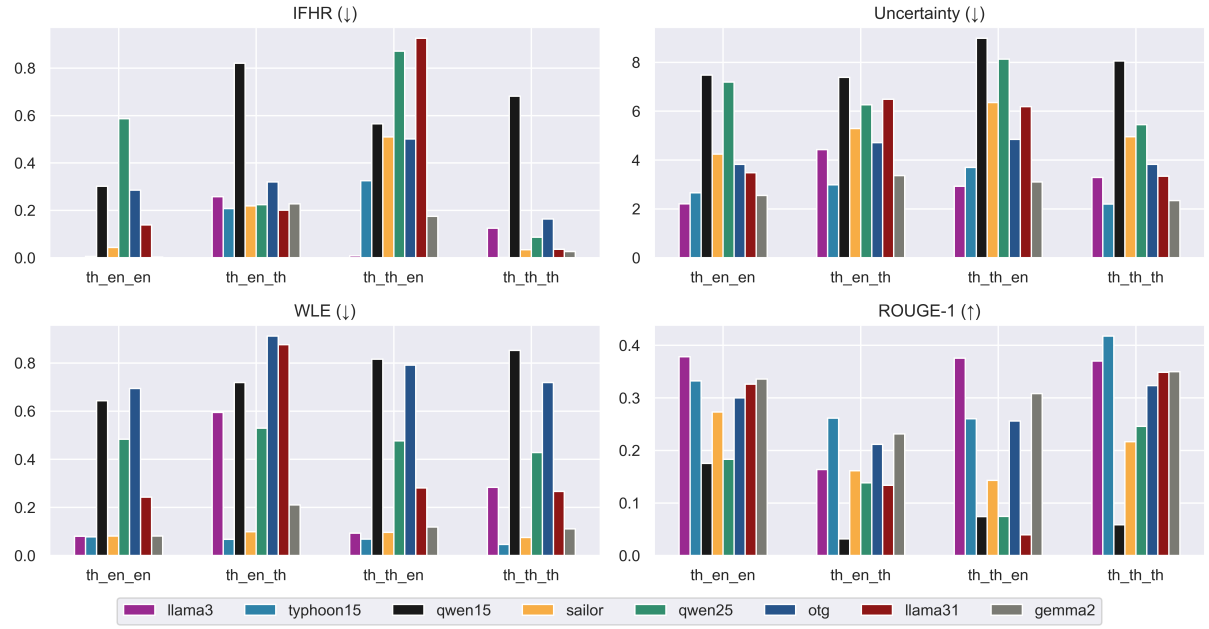
4.1 Adaptability to Thai language

We compare the Base models and their corresponding CPT models on both short-form and long-form Thai language generation tasks, specifically using experiments `th_th` for short-form and `th_th_th` for long-form generation as shown in Fig 4. Please

[2]Please navigate to Appendix C for accuracy calculation.



(a) Short-form generation tasks



(b) Long-form generation tasks with Thai instruction

Figure 3: Performance breakdown across experiments in prompt variation settings, labeled in the following format: `{task instruction}_{context input}_{output instruction}`. Note that the output instruction component is omitted for short-form generation tasks.

refer to Appendix D for performance comparison of English experiments.

Short-form generation tasks We observe two distinct patterns among the three pairs of Base and CPT models studied: (i) Llama-3 and Typhoon-1.5, and (ii) Qwen-1.5 and Sailor, and (iii) Qwen-2.5 and OpenThaiGPT-1.5. Llama-3 appears to understand the Thai language well, as indicated by its low instruction-following hallucination rate (IFHR) in Fig 4a. In contrast, the Qwen models may struggle with following instruction in Thai regarding their high IFHR. This suggests they may not be well-suited for customized text generation tasks, such as generating a single character representing the correct option in the multiple-choice instruction. Notably, the IFHR remains unchanged even after applying continual pre-training to the base models.

However, we notice signs of improvement in Thai language understanding for the Qwen-related pairs, as evidenced by decreased uncertainty and increased accuracy, but the opposite trend is observed in the Llama-3 pair. This implies that the continual pre-training can improve Thai language comprehension in models that are not originally familiar with Thai, such as Qwen-1.5 and Qwen-2.5 although it does not enable the models to follow instructions. On the other hand, it may not provide significant benefits for models that already have a relatively good understanding of Thai, such as Llama-3.

Long-form generation tasks When the instruction is relaxed to allow free-form text in Thai instead of requiring one of the valid options in the multiple-choice setting, the IFHR drops to around 10%, with an outlier in Qwen-1.5 reaching over 60% as visualized in Fig 4b. This pattern also persists at the word-level entropy (WLE), indicating that words from multiple languages are generated

within a single response, despite the instruction to generate a response in Thai. Interestingly, the continual pre-training helps reduce language confusion, particularly in the Qwen-1.5 pair. However, this effect does not hold for the Qwen-2.5 pair, where OpenThaiGPT-1.5 shows higher IFHR and WLE.

We also notice that both uncertainty and ROUGE-1 scores improve as the models align more closely with the task instruction. This trend is consistent across all pairs of Base and CPT models examined in this study.

RQ1's answer Continual pre-training can improve a pre-trained model's performance in understanding and generating text in low-resource languages, such as Thai, especially when the model initially lacks proficiency in the language. However, the degree of improvement may also depend on factors beyond model architecture and training data distribution, such as the alignment between the pre-training data and the target downstream tasks.

In our experiments, continual pre-training does not consistently help models follow task-specific instructions. For example, some models continue to generate free-form text when a single-character response is required in a multiple-choice setting. These results suggest that without sufficient exposure to similar task formats during pre-training, models may still struggle with task generalization, regardless of improvements in language understanding.

4.2 Continual pre-training vs Multilingual pre-training

We further investigate how different training strategies contribute to downstream tasks by focusing on continual and multilingual pre-training. We select Llama-3.1 as the baseline for multilingual pre-trained model (MLLM) performance, represented by the black dashed line in Fig 4.

Short-form generation tasks The MLLM demonstrates strong task understanding and follows instructions well, as indicated by the almost zero IFHR. Surprisingly, the output quality, measured in terms of uncertainty and accuracy, is not particularly outstanding (see Fig 4a). It offers performance comparable to Typhoon-1.5, which is a CPT version of Llama-3.

Long-form generation tasks Although the IFHR remains relatively low, the WLE is not as low (see Fig 4b). This suggests that the model occasionally generates tokens in other languages although the overall response is still classified as Thai. In terms of uncertainty, the MLLM displays patterns similar to those seen in CPT models. Regarding the response quality, as measured by ROUGE-1, the MLLM outperforms models that are continually pre-trained from Qwen family, and is competitive with models continually pre-trained from Llama-3.

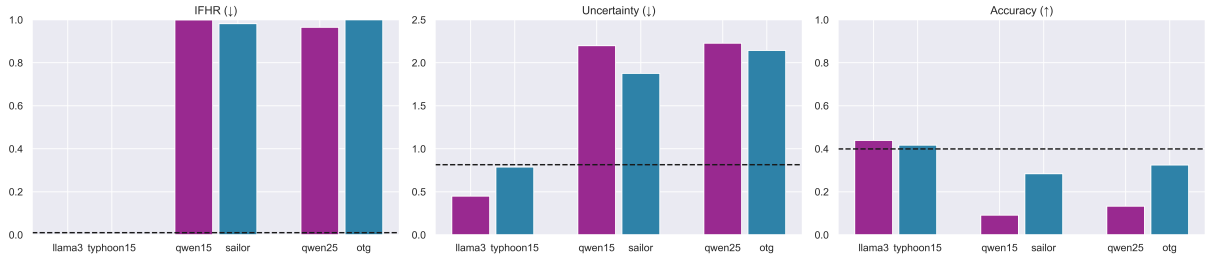
These results imply that model family plays a significant role in multilingual performance. While the Qwen family may not perform as strongly in Thai in its base form, continual pre-training can boost its capabilities to approach MLLM-level performance. On the other hand, continual pre-training on Llama-3 provides a more substantial performance lift, surpassing both the base models and the MLLM. This highlights the strength of Llama-based architectures for Thai language tasks, especially when further refined through continual pre-training.

RQ2's answer Although MLLMs exhibit strong instruction-following abilities and tend to generate fewer hallucinations, their performance is not consistently better across all tasks. In contrast, continual pre-training on a new language can achieve competitive, or even superior, results compared to multilingual pre-training. However, the successful continual pre-training depends on the strength of the base model, as well as the quality, diversity, and distribution of the data used during continual pre-training.

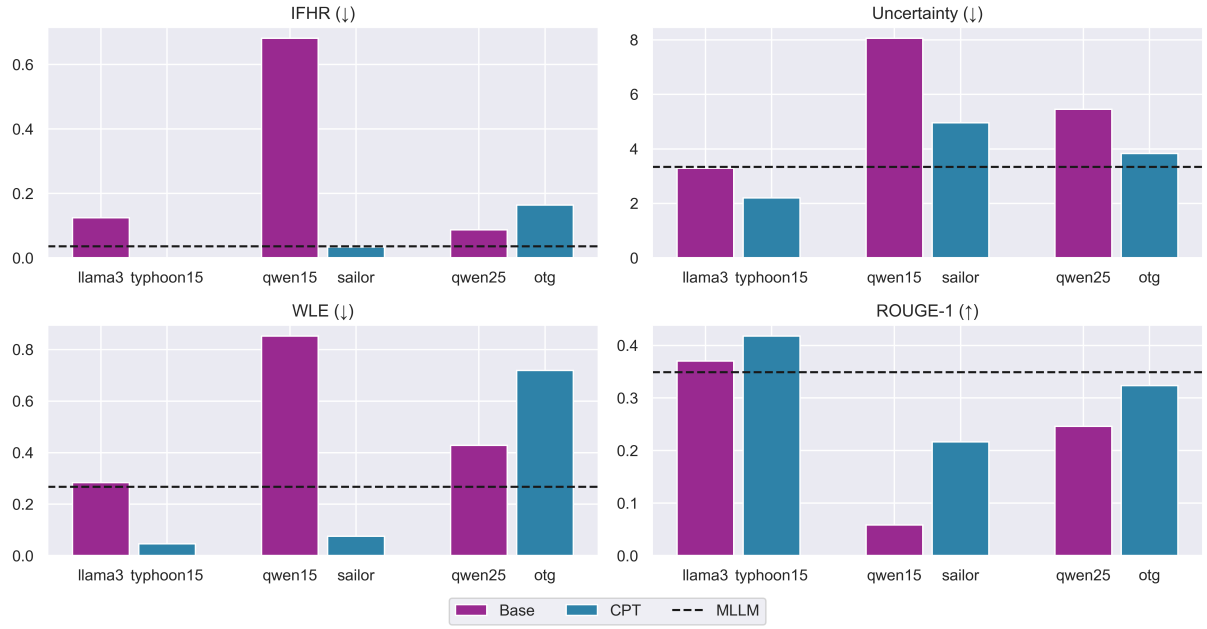
4.3 Cross-lingual prompts

Regarding the language confusion studied in (Marchisio et al., 2024), we extend the study by decomposing each prompt into three components including task instruction, context input, and output instruction, as illustrated in Fig. 2. We then vary the language of each component between English and Thai to investigate model robustness across different models. We also include Gemma-2-9B as a baseline to serve as an approximate upper bound for performance as displayed in Fig 3.

Short-form generation tasks Figure 3a presents the experimental results obtained by varying the languages used for the task instruction and context input within the prompts. The models consistently achieve their best performance in the `en_en` setting, characterized by higher accuracy and lower IFHR and uncertainty. However, when Thai is introduced



(a) Short-form generation tasks with experiment `th_th`



(b) Long-form generation tasks with experiment `th_th_th`

Figure 4: Comparison of model types–Base, CPT, and MLLM–for Thai language on the benchmarks (a) Short-form and (b) Long-form generation tasks in terms of IFHR (↓), Uncertainty (↓), WLE (↓), and Performance (↑), measured via Accuracy and ROUGE-1 for the respective short-form and long-form generation tasks. Note that the MLLM results are retrieved from Llama-3.1 and the model names on the x-axis are abbreviated for display clarity, while `otg` refers to OpenThaiGPT-1.5.

in either part of the prompt, the performance of all models deteriorates regardless of model type. Notably, the magnitude of this decline remains consistent across all Thai-related experiments. This indicates the models' weakness in processing mixed language prompts which is possibly due to limited exposure to Thai language data during training process.

Long-form generation tasks We observe that language variation in the task instruction component does not significantly affect performance, as shown in Appendix E. Therefore, we present the results in Fig. 3b, which illustrate the effect of varying the languages in the context input and output instruction components, while keeping the task instructions in Thai.

The base models demonstrate a strong reliance on English, achieving their optimal ROUGE-1 score under the `th_en_en` setting. This is a direct consequence of the English-centric dominance in their pre-training data, which ensures high fidelity in processing English language.

The CPT models, on the other hand, exhibit the anticipated benefits of localized adaptation on Thai data. Relative to the Base models, they demonstrate a significant increase in ROUGE-1 and a reduction in WLE for `th_th_th` or Pure Thai experiment as visualized in Fig 3b. This indicates that the continual pre-training process successfully refined the Thai token-level representations, leading to more accurate and confident Thai generation.

However, both Base and CPT models suffer when the languages of the context input and output instruction are mismatched because the IFHR, uncertainty, and WLE are higher than the monolin-

gual settings.

Conversely, MLLMs display the highest degree of robustness and the lowest performance variance across all prompt language settings. This superior performance is attributed to their foundational multilingual pre-training objective, which promotes a shared representational space across English and Thai.

RQ3's answer The language used in different prompt segments does not make much impact for the short-form generation tasks, but for the long-term generation tasks, we observe that it impacts task performance in multilingual settings, especially with the most critical factor being the language mismatch between the context input and the output instruction. For Base and CPT models, this mismatch introduces a severe cross-lingual penalty, resulting in increases across all failure uncertainty-related metrics, as the models struggle to seamlessly translate information extracted in one language into constraints required by the other.

Conversely, MLLMs demonstrate superior robustness and minimal performance degradation under all mixed-language conditions. This confirms that their foundational multilingual alignment effectively eliminates the internal processing conflict and uncertainty observed in other architectures.

5 Conclusion

Continual pre-training (CPT) demonstrates notable improvements in both language confusion and performance metrics within mono- and cross-lingual settings compared to base models, particularly for languages such as Thai. However, its effectiveness is highly task-dependent and influenced by the base model's initial linguistic proficiency. Despite these gains, CPT models still lag behind multilingual large language models (MLLMs), which show superior robustness and better handle context–output language mismatches in cross-lingual tasks. Given the high computational cost of training multilingual models from scratch, integrating multilingual training strategies into CPT approaches may offer a promising pathway to enhance model generalization and achieve more robust multilingual capabilities for downstream applications.

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-9B parameters, which may affect the completeness of the comparison.

Acknowledgements

We would like to thank SCB X Public Company Limited (SCBX) for their generous funding and compute resources, which support this research. Special thanks also go to Stanford Human-Centered Artificial Intelligence (HAI) and SCB 10x for API credits and their valuable consultation throughout the exploration.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 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, and 1 others. 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, and 1 others. 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, and 1 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding:

Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 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, and 1 others. 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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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, and 1 others. 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, and 1 others. 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, and 1 others. 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.

A Translation details

We employ GPT-4 (Achiam et al., 2024) to translate the dataset from Thai into English language with the following prompt.

Translation prompt

```
Translate the following Thai question
into English.
Thai: {content}
English:
```

We calculate the cosine similarity score between embedding vectors of questions in Thai and English using BGE-M3 model (Chen et al., 2024). Overall, the translation quality is good, as over 88% of the data achieves a score higher than 0.7. We only make minor changes to the samples where key information for the subject and verb is missing. However, we find an issue when translating Thai proverbs into English, so we remove this category from the ThaiExam dataset (Pipatanakul et al., 2023).

B Dataset statistics

The number of data points for each dataset used in the experiments is given in Table 1.

Task	Dataset	#of questions
Short-form	MMLU	14,042
Short-form	ThaiExam	583
Short-form	WTI-MC	787
Long-form	WTI-CQA	741
Long-form	WTI-SUM	793

Table 1: Dataset distribution in the experiments.

C Lenient accuracy calculation for shot-form generation tasks

We notice an issue when a model fails to follow instructions for short-form generation tasks. Specifically, it sometimes generated more than one token to represent the correct option. This makes it misleading to calculate accuracy based on an exact match between the raw response and the gold answer.

Therefore, we relax the accuracy criteria. Responses with certain prevalent patterns are now counted as correct. Examples of these patterns include "`Here is the answer: <x>`", "`Option <x> is the right answer`", and "`<x> <followed by option detail>`".

However, other metrics are still calculated based on the original responses.

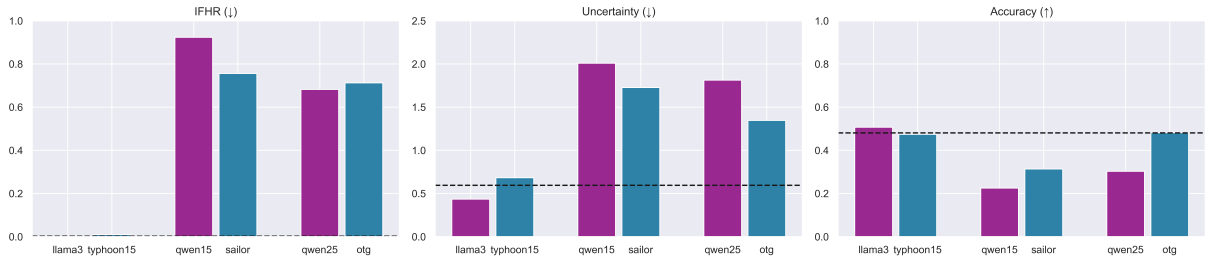
D Comparison of model types for English language settings

We also plot the comparison among different model types in English language settings, specifically `en_en` and `en_en_en` settings for both short-form and long-form generation tasks in Fig 5. The observed pattern shows similar behavior as discussed in Section 4.1.

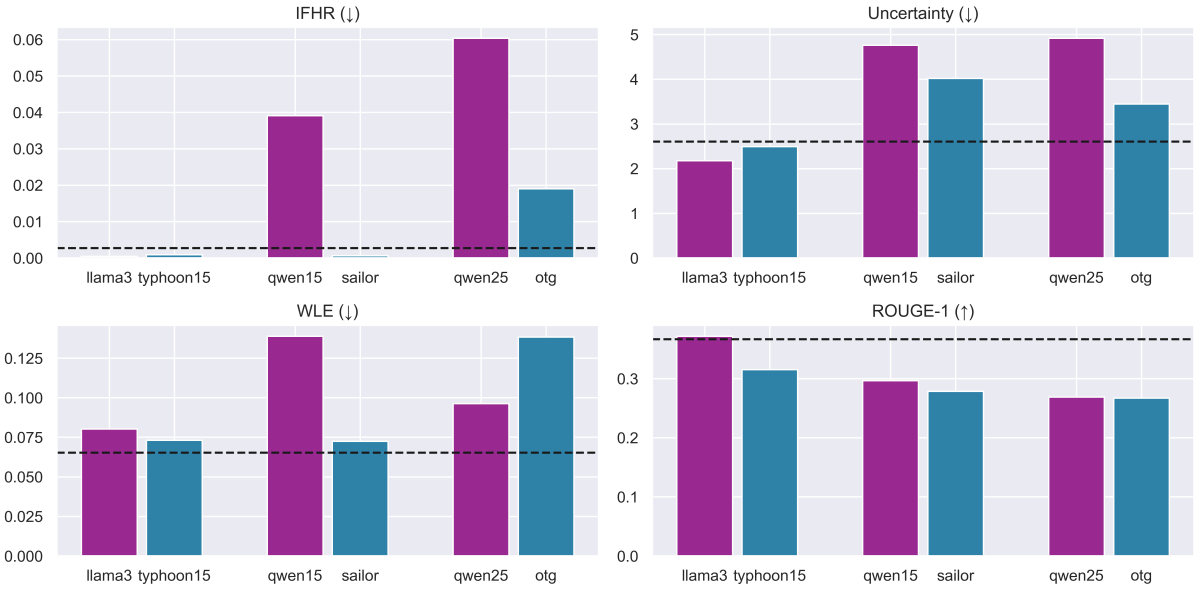
E Full experimental results of language variations for long-form generation tasks

All of the prompt variation results are displayed in Fig 6. We observed similar patterns when varying the language in the task instruction, except in the `en_en_th` and `en_en_en` experiments.

In the `en_en_th` setting, all the models perform poorly because the prompts are in English, yet they are instructed to generate a Thai response. This single token for language control leads to confusion regarding the language switch. Conversely, the `en_en_en` or Pure English setting, allows the model to perform very well.

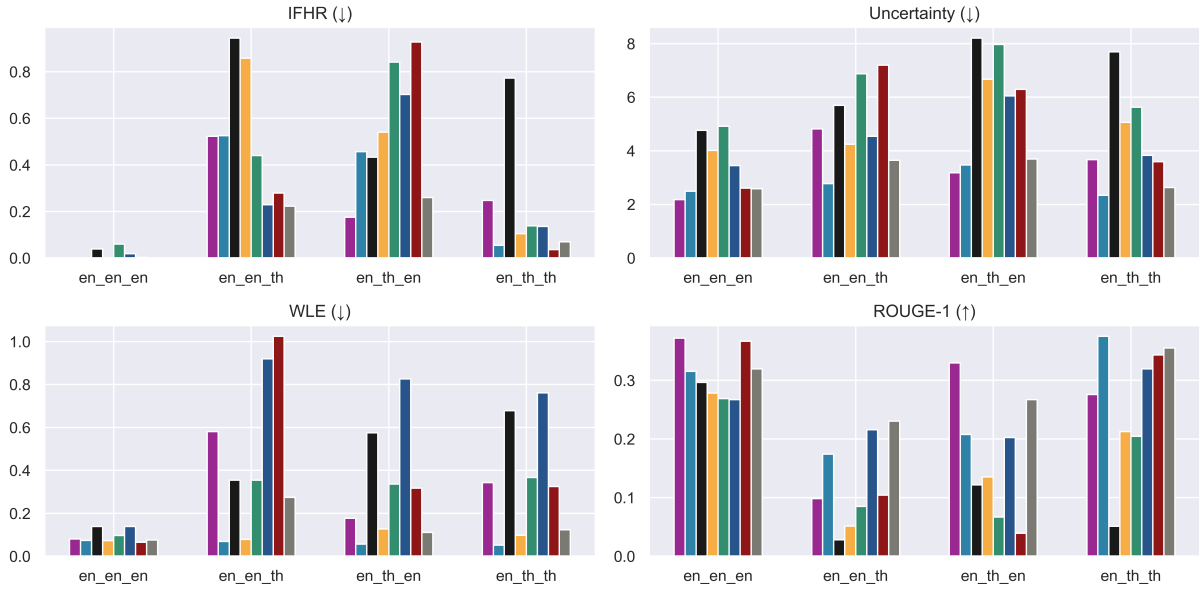


(a) Short-form generation tasks with experiment `en_en`

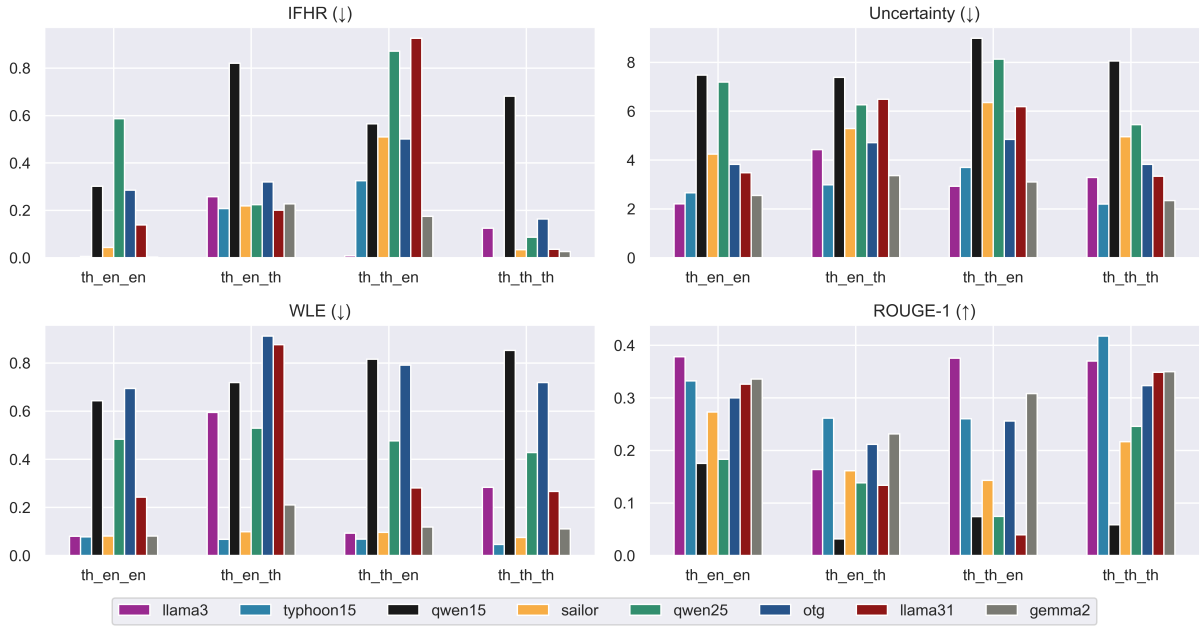


(b) Long-form generation tasks with experiment `en_en_en`

Figure 5: Comparison of model types for English language on the benchmarks (a) Short-form and (b) Long-form generation tasks in terms of IFHR (↓), Uncertainty (↓), WLE (↓), and Performance (↑), measured via Accuracy and ROUGE-1 for the respective short-form and long-form generation tasks. Note that the MLLM results are retrieved from Llama 3.1 and the model names on the x-axis are abbreviated for display clarity, while `otg` refers to OpenThaiGPT 1.5.



(a) English task instruction



(b) Thai task instruction

Figure 6: Performance breakdown across experiments in prompt variation settings, labeled in the following format: `{task instruction}_{context input}_{output instruction}`.