Bridging Borders:
Enhancing GPT-3.5 for Zero-Shot Thai Intent Classification via Cross-Lingual Prompts, Chain-of-Thought, and Self-Consistency

Thitirat Meennuch
Language and Information Technologies,
Faculty of Arts, Chulalongkorn University
6440059522@student.chula.ac.th

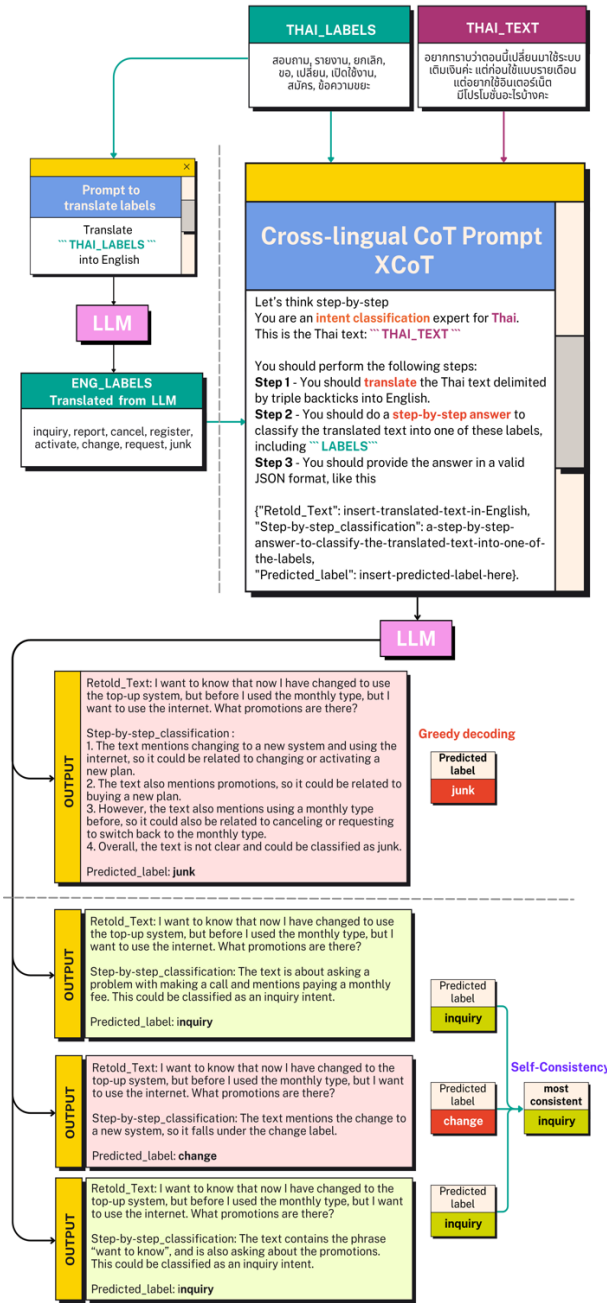

Figure 1: Summary of methods proposed in this paper

Abstract

Intent classification in low-resource languages like Thai poses challenges due to limited linguistic data. This paper proposes a method using Large Language Models (LLMs) and cross-lingual techniques. By prompting GPT 3.5 in English rather than Thai, we enhance classification. Our Cross-Lingual Chain-of-Thought Prompt template (XCoT) improves LLM performance, integrating role-assigning and cross-lingual steps, surpassing standard prompts. Additionally, employing Zero-Shot reasoning, label translation, and Self-Consistency techniques significantly boosts F1 scores, promising advancements in low-resource language intent classification.

1 Introduction

The intent classification task involves classifying language text into specific intents, playing a crucial role in various applications, from enhancing natural language processing systems to improving user interactions with chatbots and virtual assistants.

Still, grasping the intricacies of language syntax, semantics, and context poses challenges, particularly in lower-resource languages like Thai. This complexity arises from the need to consider word choice, sentence structure, and linguistic patterns associated with each intent label.

To improve classification performance, cross-lingual transfer has been explored. Cross-lingual bridges the gap between high and low-resource languages by utilizing knowledge acquired from one language to perform linguistic tasks in another (Lin et al., 2019).

However, existing approaches often rely on resource-intensive supervised methods using

extensive labeled datasets. Large Language Models (LLMs) have seen widespread adoption, especially with the shift towards more efficient in-context learning and prompting. These models, extensively trained on large datasets, primarily underwent pre-training on English data, making English comparatively more data-rich for LLMs than other lower-resource languages like Thai.

Our method takes advantage of this data discrepancy to enhance LLM performance in intent classification tasks for lower-resource languages. A classification task can be thought of as a generative task when using a generative language model. Casting the classification task as a generative one within a generative language model framework, we leverage GPT 3.5 to prompt the model to generate responses representing classified labels.

Our method employs a step-by-step instruction in the form of chain-of-thought prompting, we instruct the model to first translate text from lower-resource languages to higher ones and subsequently think step-by-step to classify the translated text into provided labels. We explore the integration of "Let's think step-by-step" Zero-Shot reasoning proposed by Kojima et al., (2022) at the prompt's outset, as well as translating labels into higher-resource language. We also experiment with self-consistency (Wang et al., 2022), sampling diverse reasoning paths and selecting the most consistent model-generated outputs. We conclude that the integration of cross-lingual chain-of-thought prompting, zero-shot reasoning, label translation, and self-consistency yields improved F1 performance as a result.

Our contributions are as shown in Figure 1:

- We propose our Cross-Lingual Chain-of-Thought Prompting template (XCoT). XCoT guides the LLM through role-assigning and cross-lingual steps before text classification, which can improve performance compared to a standard prompt that does not incorporate chain-of-thought. (section **4**, **5**, and **6**)
- We also showcase that even without using any Chain of Thought (or guiding the model step-by-step) prompting templates, prompting GPT 3.5 in English, a high-resource language, enhances classification performance compared to prompting in Thai, a lower-resource language. (section **6**)
- We demonstrate that using "Let's think step-by-step" Zero-Shot reasoning, label translating to higher-resource language, and self-consistency help improve F1 score of LLMs in classification tasks. (section **4**, **5**, and **6**)

2 Related works

Cross-lingual Transfer for LLM. Although LLMs exhibit remarkable performance across diverse language-based tasks in multiple languages (Brown et al., 2020; Devlin et al., 2019; Xue et al., 2021), their proficiency still varies among languages due to the higher prevalence of certain languages, such as English, in the pretraining data. (Huang et al., 2023)

Several supervised learning approaches employed the cross-lingual transfer technique to enhance LLM performance on tasks in lower-resource languages, predominantly through fine-tuning and continued training on higher-resource language and evaluate the performance on the lower-resource one (Chen et al., 2023; Chi et al., 2021; Pires et al., 2019).

In classification tasks, two primary supervised cross-lingual transfer methods are *translate-and-train* and *translate-and-test*. *Translate-and-train* involves translating training or fine-tuning data from a high-resource language to a lower-resource language using a machine translation (MT) model. Then, the classifier undergoes fine-tuning or continues training in the lower-resource language. Conversely, *translate-and-test* translates text from a lower-resource language to a higher-resource one before utilizing a classifier in the high-resource language for classification (Unanue et al., 2023).

These supervised approaches, however, tend to be computationally expensive. Recent studies have geared towards more resource-efficient in-context learning methods using prompts.

Chain of Thought Prompting and Self-Consistency. Prompting involves providing specific input to guide LLMs to generate relevant text output. Users present a prompt—an initial input or query—to the language model, which then generates text based on the given prompt and its comprehension of languages learned during training.

Chain of Thought (CoT) refers to a technique that involves guiding a language model through a sequence of related prompts or steps to encourage coherent and logical reasoning (Wei et al., 2022). Its application, known as Chain-of-Thought prompting, significantly boosted performance in reasoning tasks across both few-shot and zero-shot (Kojima et al., 2022) scenarios.

Another recent method, "Self-Consistency," (Wang et al., 2022) has demonstrated further improvement in LLMs performance. Self-consistency involves prompting LLMs with the CoT and subsequently sampling various reasoning paths to identify the most consistent one.

Cross-lingual Thought prompting. Huang et al. (2023) explored cross-lingual thought prompting to enhance LLMs' multilingual capability without requiring a task or language-specific fine-tuning process. Their approach amalgamated English, a higher-resource language, with a CoT prompting technique to formulate a prompting template.

This template guides LLMs to translate the text to English before performing reasoning, understanding, and generation tasks across different languages.

The findings from other works, such as Shi et al., (2023), also demonstrated that using English Chain-of-Thought prompts improved model performance compared to prompts in lower-resource languages.

In this paper, we explore Cross-lingual CoT prompts for Thai intent classification. Adapted from Huang et al. (2023), our template prompts LLMs to perform cross-lingual steps prior to text classification. Additionally, we show that incorporating label translation and Self-Consistency can improve the LLM classification performance.

3 Intent Classification for Thai

The intent classification task involves categorizing language text into specific intents or purposes, where each label signifies a unique intention behind the text.

The challenge in accurately classifying intent involves comprehending the language syntax, semantics, and context, considering word choice, sentence structure, and linguistic patterns tied to each intent label.

We frame this task as a generative one by employing the GPT-3.5 series model. We aim to enhance the model performance in predicting the intent or purpose of a given text snippet within the constraints of a lower-resource language, Thai.

Thai, like many languages, relies heavily on idiomatic expressions and context-sensitive phrases. These linguistic subtleties often pose challenges for models like GPT-3.5 in capturing and interpreting the intended meaning.

4 Model and Proposed Methods

4.1 Model

As mentioned, we frame Intent Classification as a generation task: we use the GPT-3.5-turbo-instruct, a generative model based on the Transformer architecture developed by OpenAI.

Unlike traditional sequence-to-sequence models, GPT operates as an autoregressive model, predicting subsequent tokens based on preceding ones. The model pre-training is done on a large-scale corpus of text, enabling it to develop a general understanding of language. It can continue or complete the text in a contextually relevant manner given an initial prompt.

The intuition of our approach is to present the model with a prompt that includes details about the intent classification task and the possible labels. GPT attempts to predict the correct label or category for a given input text. This means it generates responses or classifications without undergoing a task-specific fine-tuning process, relying solely on the information provided in the prompt to make predictions.

4.2 Proposed Methods

Cross-Lingual Chain-of-Thought Prompt

Large language models (LLMs) have been primarily pretrained on English data, which makes English a more data-rich language in these models compared to Thai. As such, we create a prompt template to leverage this discrepancy.

Our prompting template (XCoT) is adapted from the XLT template, a cross-lingual thought prompting template proposed by Huang et al. (2023). XLT is an English prompting template includes 6 six logical instructions: role-assigning, task-inputting, cross-lingual thinking, task analyzing, CoT task solving, and output formatting.

Our XCoT simplifies XLT's 6 logical instructions while incorporating other essential directives as we concentrate on utilizing cross-lingual transfer for classification tasks.

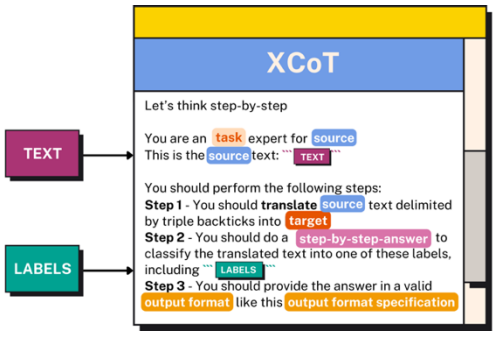

Figure 2: Cross-Lingual Chain-of-Thought Prompt (XCoT)

Role-assigning. This was essentially adapted from the Chat-GPT system role-assigning. We assign a role to the model by specifying the task we want the model to perform in the designated `task` slot. For instance, if we want the model to perform the intent classification task, we simply put "intent classification' in the slot.

Text input. Fill in the text snippet for classification in `TEXT` slot.

Step-by-step instruction. We instruct the model to perform 3 steps, including:

Step 1: **Translate** `TEXT` from the original `source language` to `target langauge` , which in our case is from "Thai" to "English."

Step 2: After translation, we instruct the model to do a `step-by-step-answer`, to classify the translated text into one of the `LABELS`. This step further encourages the model to approach classification tasks in a **sequential thought** process.

Step 3: We specify our desired **Output formatting** in the `output format`, `output format specification`
For instance, if we choose JSON format, we place "JSON" in `output format` slot.
For `output format specification`, we outline the desired content within the file.
For example,

```
{"Retold_Text": insert-translated-text-in-English,
"Step-by-step_classification": a-step-by-step-answer-to-classify-the-translated-text-into-one-of-the-labels,
"Predicted_label": insert-predicted-label-here}
```

We also experiment by adding Zero-Shot reasoning "Let's think step-by-step" proposed by Kojima et al., (2022) at the beginning of our prompt template.

Self-Consistency

This approach contrasts with traditional greedy decoding methods, which prioritize single best responses without considering consistency among generated outputs.

Self-consistency samples a diverse set of reasoning paths and selects the most consistent outputs generated by the model. In this paper, we employ popular voting to select the most consistent answer among generated outputs from the XCoT prompt.

5 Experiment

5.1 Dataset

We use the truevoice-intent dataset, a customer service phone call dataset, transcribed and offered by TrueVoice's Mari[1]. The test set contains 3226 rows, each row includes a Thai text transcription, and its tokenized form using deepcut, along with action, object, and destination labels.

We evaluate our proposed method for categorizing raw (untokenized) Thai text transcriptions into single action labels. which encompasses a total of 8 categories: enquire, report, cancel, request, change, activate, buy, and garbage.

The dataset comes with English labels. To ensure the efficacy of our method and gauge the Language Model's real performance in the lower-resource language, we manually translate these labels into Thai. The translations include: สอบถาม (enquire), รายงาน (report), ยกเลิก (cancel), ขอ (request), เปลี่ยน (change), เปิดใช้งาน (activate), สมัคร (buy), and ข้อความขยะ (garbage).

Figure 3 illustrates the percentage distribution of Thai labels within the truevoice-intent dataset.

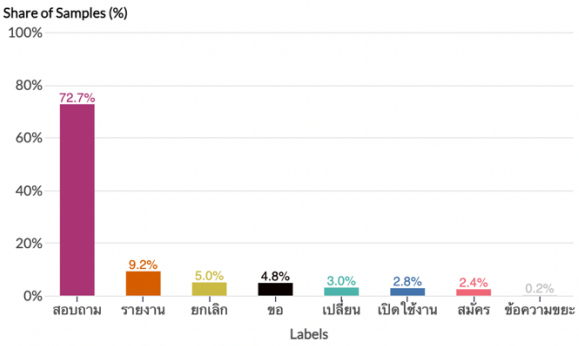

Figure 3: labels distribution within the dataset

5.2 Experimental conditions

(prompt template for each experimental condition can be found in the Appendix)

Baseline: We compare the performance of other techniques against a direct Thai (lower-resource language) prompt baseline.

Cross-Lingual Prompt: Our first hypothesis posits that the utilization of cross-lingual prompts,

[1] available at https://github.com/kobkrit/truevoice-intent/raw/master/mari-intent.zip

particularly prompting the LLM in English (higher-resource language) as opposed to Thai (lower-resource language), would result in higher accuracy and F1 scores compared to the baseline.

Cross-Lingual Chain-of-Thought Prompt (XCoT): Our second hypothesis proposes that integrating the Chain-of-Thought prompting technique leads to enhanced performance compared to both the baseline and the standard cross-lingual prompt.

Our prompting template (XCoT) directs the LLM to act as a classifier expert and provides step-by-step instructions to translate Thai (a lower-resource language) text into English (a higher-resource language). Additionally, our evaluation involves comparing the performance with and without integrating the Zero-Shot reasoning technique "Let's think step-by-step" at the beginning of XCoT.

Self-Consistency: Our final hypothesis proposes that incorporating self-consistency with the XCoT prompting template yields better performance compared to the greedy decoding method. We select the most consistent output generated by the model's decoder.

Given the rapid performance saturation in several scenarios (Wang et al., 2022), we limit the paths and outputs to only top-5, optimizing computational resources while still gaining performance benefits.

Supervised: In addition, we benchmark our approaches against supervised learning outcomes from WangchanBERTa, an encoder-only model trained using Masked Language Modeling (MLM).

We use the labeled training set from TrueVoice's Mari to fine-tune WangchanBERTa. The training set consists of 12939 rows of Thai text transcriptions, paired with intent action labels. The model learns from these labeled examples to understand the relationship between input text and the respective intent label during the fine-tuning[2]. Subsequently, we evaluate the fine-tuned WangchanBERTa model using the truevoice-intent test set.

5.3 Implementation

We use the OpenAI Python library to interact with the GPT-3.5-turbo-instruct completion model. This includes configuring the OpenAI API key and utilizing the OpenAI API to generate completions based on specified prompts.

We follow the Zero-shot Classification steps from OpenAI Cookbook[3] and the setting for greedy decoding from Huang et al. (2023) and Shi et al., (2023) (i.e., sampling with temperature τ = 0). For self-consistency, we follow the setting from Wang et al. (2022) (i.e., τ = 0.7)

We use the computing resources, including the T4 GPU, provided by the Google Colab environment for running our tasks

6 Results and Discussion

F1 score *(average from 3 Independent runs)*	Thai Prompt *(Thai labels)*	English Prompt *(Thai labels)*	**XCoT** *(Thai labels)*	**XCoT** *(en-trans-lated-labels)*
micro average	42.7	50.3	51.7	57.7
macro average	31.3	34.6	33.4	36.6
weighted average	48.2	55.3	55.8	61.4

Table 1: F1 score on the truevoice-intent test set (computed as mean across 3 independent runs). The comparison includes the F1 score derived from results of GPT 3.5-turbo-instruct using the XCoT prompting method with Thai and English-translated labels, contrasted with standard prompting in Thai and English.

F1 score *(average from 3 Independent runs)*	**XCoT** *(en-trans-lated-labels +Zero-Shot Reasoning)*	**XCoT** *(en-translated labels +Zero-Shot Reasoning* + Self-Consistency)	Wangchan Berta
micro average	60.3	**62.3**	77.9
macro average	**39.0**	38.3	39.5
weighted average	64.3	**65.0**	71.8

Table 2: F1 score on the truevoice-intent test set (computed as mean across 3 independent runs).
The comparison includes the F1 score derived from results of GPT 3.5-turbo-instruct using the XCoT prompting method, contrasted with the F1 score from fine-tuned WangchanBerta. The bold-faced number highlights the highest F1 score from both Table 1 and Table 2 achieved by GPT 3.5-turbo-instruct.

From Table 1, we can see that XCoT prompting demonstrates improvements in average F1 scores in comparison to regular Thai and English

[2] learning_rate=1e-5, batch_size=64, epochs=4

[3] https://cookbook.openai.com/examples/multiclass_classification_for_transactions

prompts. Table 2 shows that the incorporation of English-translated labels, Zero-Shot Reasoning, and Self-Consistency notably elevates performance.

The results show that simply prompting GPT 3.5 in English rather than Thai adds a significant improvement in the F1 score. This aligns with Shi et al. (2023) findings, indicating that the English CoT prompt can improve the model's performance in tasks across various lower-resource languages compared to the native prompt.

However, we observe that using the XCoT prompt template with Thai labels does not lead to a substantial improvement over a regular English prompt. We suspect that this may be attributed to the direct instruction in 'Step 2' of the XCoT template, guiding the LLM to classify the translated text (English text) into Thai labels. This instruction may result in a performance that is not significantly higher or even lower than that achieved with a regular English prompt.

To validate our observation, we prompt the model to translate the Thai labels into English[4] and then integrate those English-translated labels into our XCoT prompting template. Note that as we want to avoid any potential inconsistencies in label translation, we use a separate prompt in translating the labels instead of directly instructing the model within our XCoT.

As shown in Table 1, employing English-translated labels boosts the F1 score compared to using Thai labels.

Finally, starting XCoT with Kojima et al., 2022's Zero-Shot reasoning "Let's think step-by-step", as well as incorporating Self-Consistency using top 5 samples, contribute to further improvement in F1 score.

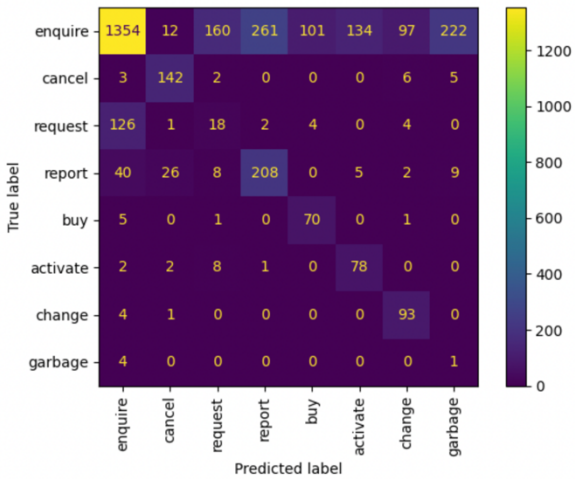

Figure 4: The heatmap compares true labels and labels predicted from GPT 3.5-turbo-instruct on a single run, using XCoT prompting template with English-translated labels and "Let's think step-by-step" Zero-Shot Reasoning

7 Conclusion

In this paper, we propose a method for classifying Thai intent using LLM. We employ a Cross-Lingual Chain-of-Thought Prompt template, Zero-Shot reasoning, label translation, and Self-Consistency. Our approach boosts the F1 score beyond traditional prompt methods.

Reference

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 159, 1877–1901.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly Easy Label Projection for Cross-lingual Transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

[4] Although our original labels are in English, we aim to simulate a scenario where no English labels are available and rely solely on the LLM as our translation tool.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. *arXiv preprint arXiv:2305.07004*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916*

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multi-lingual chain-of-thought reasoners. *In The Eleventh International Conference on Learning Representations.*

Inigo Jauregi Unanue, Gholamreza Haffari and Massimo Piccardi. 2023. T3L: Translate-and-Test Transfer Learning for Cross-Lingual Text Classification. *arXiv preprint arXiv:2306.04996*

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Appendix

Mono-lingual Prompt

```
จงจำแนกข้อความ ``` THAI_TEXT ``` ให้อยู่ในประเภทใดประเภทหนึ่ง ดังต่อไปนี้ ```LABELS```
ระบุคำตอบในรูปแบบดังนี้ {"คำตอบ": "ประเภทของข้อความที่จำแนก"}
```

Cross-lingual (English) Prompt

```
This is the Thai text: ``` THAI_TEXT ```
Classify the provided Thai text into one of these labels, including
```LABELS```
Provide the answer in this format:{"Predicted_label": "insert-predicted-label-here"}
```

Cross-lingual CoT-Prompting template (XCoT)
(adapted from Huang et al. (2023) XLT)

```
You are an intent classification expert for Thai.
This is the Thai text: ``` THAI_TEXT ```

You should perform the following steps:
  Step 1 - You should translate the Thai text delimited by triple backticks into English.
  Step 2 - You should do a step-by-step answer to classify the translated text into one of these labels, including ```LABELS```
  Step 3 - You should provide the answer in a valid JSON format, like this
{"Retold_Text": insert-translated-text-in-English,
"Step-by-step_classification": a-step-by-step-answer-to-classify-the-translated-text-into-one-of-the-labels,
"Predicted_label": insert-predicted-label-here}
```

Cross-lingual CoT-Prompting template (XCoT)
adding Kojima et al., 2022 Zero-Shot reasoning

```
Let's think step-by-step.

You are an intent classification expert for Thai.
This is the Thai text: ``` THAI_TEXT ```

You should perform the following steps:
  Step 1 - You should translate the Thai text delimited by triple backticks into English.
  Step 2 - You should do a step-by-step answer to classify the translated text into one of these labels, including ```LABELS```
  Step 3 - You should provide the answer in a valid JSON format, like this
{"Retold_Text": insert-translated-text-in-English,
"Step-by-step_classification": a-step-by-step-answer-to-classify-the-translated-text-into-one-of-the-labels,
"Predicted_label": insert-predicted-label-here}
```

Label Translating Prompt

```
You are a translation expert for Thai.
Your job is to translate the Thai categories into English.

The categories are สอบถาม, รายงาน, ยกเลิก, สมัคร, เปิดใช้งาน, เปลี่ยน, ขอ, ข้อความขยะ

Provide the answer in valid JSON format, like this
  {"สอบถาม": "translated-label",
  "รายงาน": "translated-label",
  "ยกเลิก": "translated-label",
  "สมัคร": "translated-label",
  "เปิดใช้งาน": "translated-label",
  "เปลี่ยน": "translated-label",
  "ขอ": "translated-label",
  "ข้อความขยะ": "translated-label",}
```
```