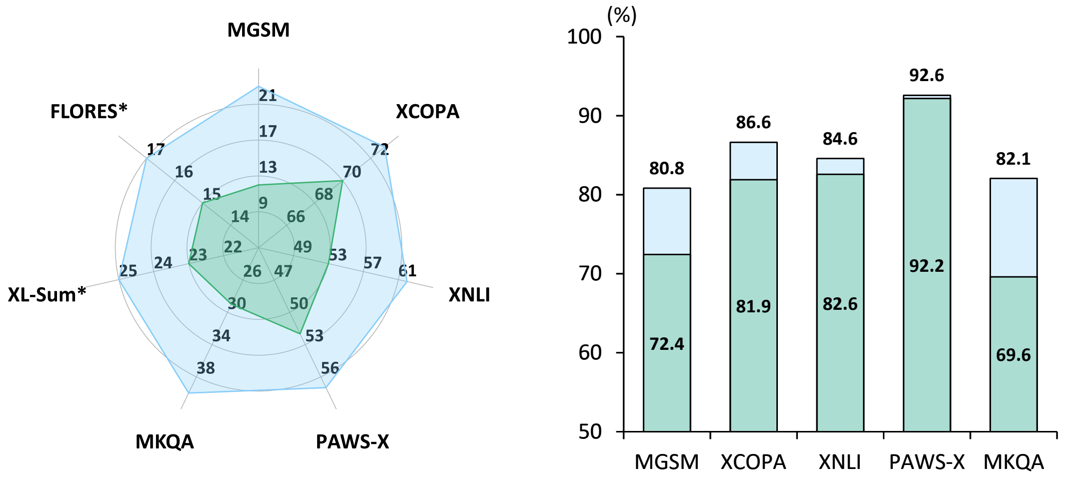
Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting

Haoyang Huang[1]*, **Tianyi Tang**[2]*, **Dongdong Zhang**[1]†, **Wayne Xin Zhao**[2]
Ting Song[1], **Yan Xia**[1], **Furu Wei**[1]
[1]Microsoft Research Asia, China
[2]Gaoling School of Artificial Intelligence, Renmin University of China
https://github.com/microsoft/unilm



(a)

(b)

Figure 1: Comparing the effectiveness of the Cross-Lingual-Thought prompt versus the baseline basic prompt on 7 representative benchmarks covering 27 languages: (a) Enhancing the multilingual capability of `text-davinci-003` under the zero-shot learning, and (b) Narrowing the gap between the average performance and the best performance of each task in different languages.

Abstract

Large language models (LLMs) demonstrate impressive multilingual capability, but their performance varies substantially across different languages. In this work, we introduce a simple yet effective method, called *cross-lingual-thought prompting* (**XLT**), to systematically improve the multilingual capability of LLMs. Specifically, XLT is a generic template prompt that stimulates cross-lingual and logical reasoning skills to enhance task performance across languages. We conduct comprehensive evaluations on 7 typical benchmarks related to reasoning, understanding, and generation tasks, covering both high-resource and low-resource languages. Experimental results show that XLT not only remarkably enhances the performance of various multilingual tasks but also significantly reduces the gap between the average performance and the best performance of each task in different languages. Notably, XLT brings over 10 points of average improvement in arithmetic reasoning and open-domain question-answering tasks.

1 Introduction

Large language models (LLMs) demonstrate impressive multilingual capability in a wide range of natural language processing tasks, including language generation, knowledge utilization, and complex reasoning (Zhao et al., 2023). Their performance in downstream tasks has been shown to reach or even surpass human-level performance (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022). The capabilities of LLMs stem from the extensive volume of training data they leveraged (Kaplan et al., 2020). The training data for current models is primarily dominated by the English language corpus, but it also encompasses data from other languages, as described in GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and BLOOM (Scao et al., 2022), *etc*.

There are over 7,000 languages worldwide, with the vast majority being low-resource or extremely low-resource languages (Forkel et al., 2022). Despite the latest GPT-4 model (OpenAI, 2023) demonstrating some generalization capabilities in

*Equal contribution. † Corresponding author.

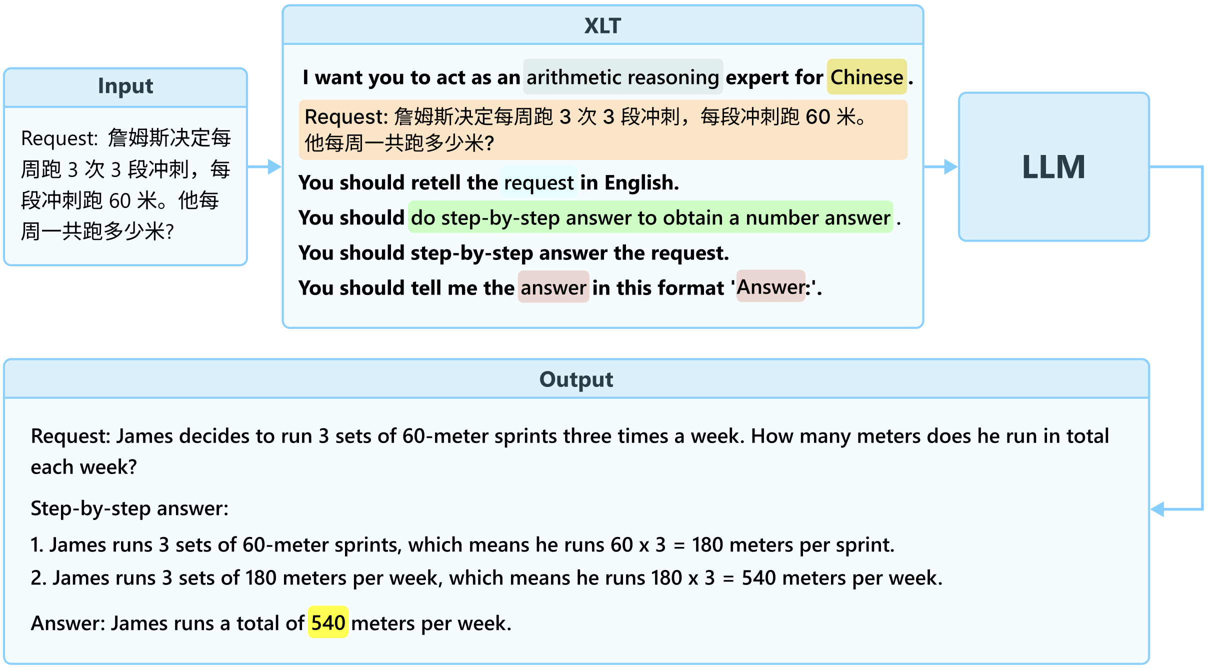

Input
Request: 詹姆斯决定每周跑3次3段冲刺，每段冲刺跑60米。他每周一共跑多少米?
XLT
I want you to act as an arithmetic reasoning expert for Chinese .
Request: 詹姆斯决定每周跑3次3段冲刺，每段冲刺跑60米。他每周一共跑多少米?
You should retell the request in English.
You should do step-by-step answer to obtain a number answer .
You should step-by-step answer the request.
You should tell me the answer in this format 'Answer:'.
LLM
Output
Request: James decides to run 3 sets of 60-meter sprints three times a week. How many meters does he run in total each week?
Step-by-step answer:
1. James runs 3 sets of 60-meter sprints, which means he runs 60 x 3 = 180 meters per sprint.
2. James runs 3 sets of 180 meters per week, which means he runs 180 x 3 = 540 meters per week.
Answer: James runs a total of 540 meters per week.

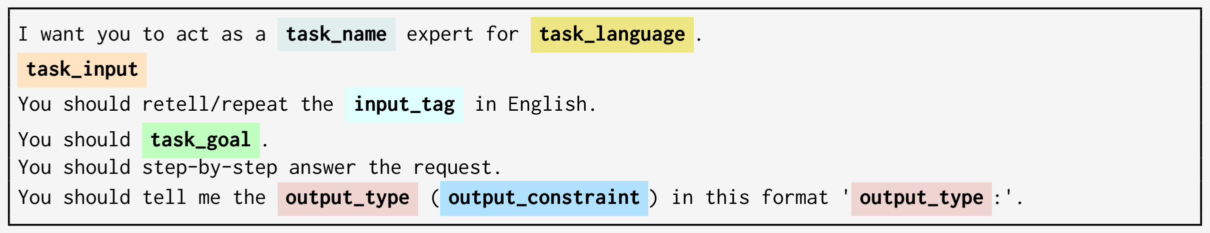
Figure 2: Overview of our method. Given a request, its associated meta information is filled into the placeholders of the XLT template to form the language-independent prompt, which is fed to the LLM to enhance the generation of responses in the desired format.

multilingual tasks as evaluated on the MMLU benchmark (Hendrycks et al., 2021), it is still the case that LLMs do not have equal capability to handle all languages, leading to imbalanced capability across different languages. Furthermore, several evaluation results (Bang et al., 2023; Jiao et al., 2023; Hendy et al., 2023; Zhu et al., 2023) indicate that large models struggle with understanding and generating non-English languages, particularly in low-resource or extremely low-resource languages. Therefore, to democratize language intelligence and minimize performance gaps in different language, it is essential and meaningful to stimulate and enhance the multilingual capability of models in non-English and low-resource languages.

Intuitively, LLMs can improve multilingual capability by augmenting data (Lin et al., 2022) or fine-tuning models (Chen et al., 2021, 2022), but both are computationally expensive. Alternatively, in-context learning with prompts can also boost performance (Brown et al., 2020; Ahuja et al., 2023; Wei et al., 2022c) but is limited to monolingual tasks (Sanh et al., 2022).

This work explores a universal in-context learning approach to enhance the multilingual capability of LLMs. We introduce a simple yet effective method, called cross-lingual-thought prompting (**XLT**), to enable models to handle various natural language processing tasks across different target languages. Our method employs a generic and language-independent prompt, which eliminates the need to update model parameters. Depending on the task input type, cross-lingual-thought prompting guides the large language model to assume the role of an expert in a specific language for a particular task. Given its predefined meta information, XLT directs LLMs to respond logically through a process involving problem understanding, cross-lingual thinking, task analysis, task execution, and output formatting. During this process, our method is designed to stimulate models' cross-lingual and logical reasoning skills, enabling them to respond to input requests regardless of the language. For enhanced performance, few-shot learning can also be employed with our method by providing an LLM-generated response output as a demonstration using cross-lingual-thought prompting zero-shot learning.

We conduct a comprehensive evaluation to verify the effectiveness of XLT across seven representative multilingual benchmarks of natural language reasoning, understanding, and generation tasks. Each benchmark includes multilingual data covering both high-resource and low-resource languages. The experimental results demonstrate that our method can significantly improve the perfor-



```
I want you to act as a task_name expert for task_language .
task_input
You should retell/repeat the input_tag in English.
You should task_goal .
You should step-by-step answer the request.
You should tell me the output_type ( output_constraint ) in this format ' output_type :'.
```

Figure 3: Illustration of XLT template. Referring to Figure 2 and Appendix for instantiated examples.

mance of all benchmarks across languages under both zero-shot and few-shot learning settings. Notably, XLT achieves an average gain of over 10 points on the MGSM and MKQA benchmarks. Furthermore, we observe that our prompting method significantly reduces the gap between the average performance and the best performance of each task in different languages, indicating its potential to democratize language intelligence.

2 Cross-Lingual-Thought Prompting

Although LLMs are capable of accepting any input and generating responses, users typically structure their requests in the form of prompts to elicit the desired output. The design of these prompts is crucial for achieving optimal performance on downstream tasks, as LLMs are sensitive to the format of the prompts chosen (Zhao et al., 2021). Through a process called instruction tuning (Wei et al., 2022a), models can develop the ability to follow natural language instructions (Wei et al., 2022b), which can reduce their sensitivity to prompt engineering (Wei et al., 2022a). In accordance with the guidelines of the OpenAI cookbook[1], we propose a cross-lingual thought prompting template, denoted as the XLT template. This generic template allows LLMs to respond to requests with cross-lingual thought and supports a wide range of multilingual tasks.

Figure 3 displays the XLT template, with the colored sections representing placeholders. Figure 2 showcases an example of instantiated prompt for the Chinese request. The following section will explain the details of constructing XLT.

2.1 Construction of XLT

The XLT template is designed to emulate the process humans employ when handling multilingual tasks. Our template is written in English, as English is the dominant language during LLM pretraining, and existing research indicates that English prompting is more effective for multilingual tasks (Shi et al., 2023). In contrast to the vanilla prompt that only includes a task description, our XLT template aims to elicit multilingual capability through cross-lingual thoughts. This template comprises six logical instructions in sequence. To complete the template, only seven placeholders need to be filled in based on intrinsic knowledge of the task and the request, as depicted in igure 3.

Role Assigning . First, the model receives a role definition that helps establish the model's behavior. This concept is akin to the system role of ChatGPT[2]. To achieve this, we simply need to fulfill the task name with a known category (such as commonsense reasoning or paraphrase identification), along with the language of the task in the task language field.

Task Inputting . Second, we explicitly append the request as the task input . The request is basically structured in terms of the task type so as to make sure the model can comprehend it. For example, in the natural language inference task, the two sentence inputs are specified with "premise" and "hypothesis", respectively.

Cross-lingual Thinking . We encourage the model to engage in cross-lingual thought by rephrasing the requested content in English, which is the dominant language used as a pivot language by Shi et al. (2023) and Ahuja et al. (2023). Rephrasing the requested content enclosed in the input tag helps the model better understand the request in its native language and knowledge. Our observations suggest that using keywords such as "retell" or "repeat" while rephrasing the content may result in better performance in practice.

[1] https://github.com/openai/openai-cookbook

[2] https://platform.openai.com/docs/guides/chat/introduction