Settings		Reasoning		Understanding		Generation		
		MGSM	XCOPA	XNLI	PAWS-X	MKQA	XL-Sum*	FLORES*
Zero-shot	`text-davinci-003`							
	Basic Prompt	12.5	70.1	53.3	52.0	29.0	23.7	15.4
	CoT	25.7	70.9	53.0	**57.8**	30.9	23.8	15.8
	Translate-En	15.7	68.0	54.8	55.0	–	–	–
	XLT	**23.9**	**73.3**	**62.4**	57.1	**40.2**	**25.2**	**17.7**
	`gpt-3.5-turbo`							
	Basic Prompt	23.3	76.9	52.6	65.5	31.6	24.7	19.1
	CoT	45.5	78.3	54.8	61.0	14.8	25.4	19.7
	Translate-En	27.1	75.7	52.2	**66.8**	–	–	–
	XLT	**70.0**	**80.3**	**65.5**	63.6	**42.7**	**26.1**	**21.2**
Few-shot	`text-davinci-003`							
	Basic Prompt	45.5	75.6	59.1	68.7	39.1	26.8	–
	Translate-En	46.5	77.4	56.9	68.5	–	–	–
	XLT	**55.4**	**81.3**	**67.5**	**72.2**	**49.6**	**27.3**	–
	`gpt-3.5-turbo`							
	Basic Prompt	63.0	80.1	61.4	66.4	43.7	25.5	–
	Translate-En	65.1	81.9	58.3	63.7	–	–	–
	XLT	**72.5**	**85.9**	**65.0**	**69.1**	**52.5**	**27.9**	–

Table 1: The average scores in different languages for the seven benchmarks in zero-shot and few-shot settings. We omit the results (denoted as "–") of Translate-En since it is not applicable for generation tasks.

Settings	Reasoning		Understanding		Generation
	MGSM	XCOPA	XNLI	PAWS-X	MKQA
Zero-shot setting					
`text-davinci-003`					
Basic Prompt	65.2	77.8	83.8	**97.1**	60.2
CoT	65.4	80.1	83.5	89.5	61.4
Translate-En	**77.2**	78.7	**86.0**	95.3	51.6
XLT	68.5	**82.1**	80.7	88.4	**78.7**
`gpt-3.5-turbo`					
Basic Prompt	73.0	83.6	80.5	89.0	61.8
CoT	66.7	85.7	80.7	88.9	46.4
Translate-En	80.4	84.6	79.8	90.7	54.1
XLT	**84.1**	**89.1**	**88.0**	**96.2**	**75.3**
Few-shot setting					
`text-davinci-003`					
Basic Prompt	75.4	82.0	82.5	88.2	74.3
Translate-En	77.1	82.6	79.5	87.8	68.5
XLT	**84.5**	**85.6**	**85.3**	**91.6**	**82.7**
`gpt-3.5-turbo`					
Basic Prompt	76.1	84.1	83.6	94.4	82.1
Translate-En	78.6	86.4	79.2	**95.4**	71.3
XLT	**86.2**	**89.7**	**84.3**	94.1	**83.1**

Table 2: The democratization degree of tasks against languages.

3.3 Further Analysis

In this section, we further investigate the factors that affect the performance of XLT and how they affect various multilingual benchmarks.

3.3.1 Ablation of XLT

For the XLT variants, we mainly conduct experiments to compare the following strategies:

- **Ablating the instructions.** Since our XLT consists of six logical instructions, we disable the *Role Assigning*, *Cross-lingual Thinking*, and *CoT Task Solving* instructions separately to analyze the contribution per instruction.
- **Reordering the instructions.** Considering the logicality of our instructions, we further change the order of the instructions in XLT to explore whether LLMs will handle tasks differently and lead to different results.
- **Changing the content word.** As prompts are usually sensitive to the word choice, we verify the robustness of XLT when alternating the rephrasing keyword with "retell", "repeat", and "translate" in the cross-lingual thinking instruction.

The outcomes are presented in Table 3, indicating that XLT surpasses almost all the variants, thereby validating the effectiveness and reasonableness of our proposed XLT method.

The effectiveness of each instruction. The results from the "Instruction Ablation" row indicate that: (1) *Cross-lingual Thinking* yields more significant gains compared to other instructions. This suggests that the LLM's ability of cross-lingual thinking is activated, allowing it to utilize its knowledge in English to solve tasks effectively; (2) Removing *Role Assigning* from XLT impedes the model's

Settings		MGSM		XNLI		FLORES*	
		de	zh	hi	vi	jv→zh	zh→jv
	XLT	79.8	72.6	61.3	64.8	19.0	10.5
Instruction Ablation	*w/o Role Assigning*	76.6	69.2	57.8	63.9	16.2	8.8
	w/o Cross-lingual Thinking	75.6	62.0	56.1	62.2	13.2	8.2
	w/o CoT Task Solving	77.0	68.0	62.9	65.2	16.8	9.2
Instruction Order	Swap *Role Assigning* and *Task Inputting*	77.2	71.8	54.2	61.5	19.6	11.2
	Swap *Role Assigning* and *Task Analyzing*	76.8	70.8	61.0	64.0	15.8	8.8
	Swap *Cross-lingual Thinking* and *Task Analyzing*	79.0	71.2	59.5	63.4	16.5	9.7
Rephrasing Word	*w/* retell	79.8	72.6	61.3	64.8	18.2	10.3
	w/ repeat	77.6	68.0	60.7	64.6	19.0	10.5
	w/ translate	76.4	70.0	60.1	64.5	17.5	10.2

Table 3: Performance comparison across different variants of XLT. All the experiments are conducted using `gpt-3.5-turbo` under the zero-shot setting.

Demonstration format	en	de	ru	fr	zh	es	ja	sw	th	bn	te	Avg.
Basic input + Basic output	84.0	79.2	78.8	78.8	70.8	81.2	68.8	70.8	68.8	65.2	44.8	71.9
Basic input + XLT output	82.4	72.4	71.2	75.2	64.4	78.8	63.2	66.8	53.6	54.8	32.4	65.0
XLT input + XLT output	84.8	81.4	80.2	79.2	71.8	81.6	72.8	71.2	69.8	64.4	40.8	**72.5**

Table 4: Performance comparison across different few-shot variants on the MGSM benchmark. All the experiments are conducted with 5 demonstrations using `gpt-3.5-turbo`.

understanding of the ultimate goal for diverse multilingual tasks, highlighting the task transferability of XLT; and (3) the better performance of XLT can also be attributed to *CoT Task Solving*, which requires the model to respond to complex instructions in a step-by-step manner.

The order of logical instructions. The performance drop is evident when the order of our designed logical instructions is switched. When designing XLT, we have taken into account the process by which humans solve multilingual problems, and this experiment further confirms the optimum order of our XLT template. Placing the *Role Assigning* instruction later may confuse the model initially. Additionally, conducting *Cross-lingual Thinking* before *Task Analyzing* is crucial since we rely on the English task-solving abilities of LLMs to handle multilingual tasks.

The robustness of word choice for rephrasing keywords. We can find that different words indeed affect the performance of XLT, but it is less sensitive to the other variants. Through experimentation, we have determined that “repeat" yields better results for text summarization and machine translation, while “retell" is more suitable for the remaining five tasks. Our aim is to provide XLT with a more unified template, while still allowing users to fine-tune specific keywords for optimal performance in their tasks.

3.3.2 Effectiveness of XLT Few-shot Learning

As mentioned in Section 2.2, the construction of demonstrations for XLT few-shot learning differs from the previous method. We have compared XLT and basic prompt. Here, we focus on the construction of the demonstration input-output pairs and compare various demonstrations that may be used to perform XLT few-shot learning. The illustrations can be found in Figure 5.

- **Basic prompt input + Basic prompt output:** This is the normal demonstration format used in most of the previous work.
- **Basic prompt input + XLT output:** This ablation is to separate the effect of input and output formats in the demonstration.
- **XLT input + XLT output:** This is the method that we used in this work.

Observing the experimental results presented in Table 4, we can conclude that: (1) Our XLT few-shot learning outperforms all other variants, thus confirming its effectiveness. (2) The use of normal demonstrations for XLT few-shot learning leads to a decrease in performance. (3) Merely incorporating XLT as a demonstration input without its output does not result in any improvements. (4) Consistency in the demonstration for few-shot learning

is crucial, implying that the demonstration input-output format should align better with its zero-shot learning input-output format.

4 Related Work

4.1 LLM Capability Understanding

Despite the impressive capabilities of LLMs, it is crucial to determine their impact on natural language processing tasks. Liang et al. (2022) conduct a comprehensive evaluation of LLMs from various perspectives, such as accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency. Bang et al. (2023) extensively evaluate the ChatGPT model on multiple natural language processing tasks and find that the model performs well in high-resource languages but exhibits certain limitations in low-resource and non-Latin script languages. Additionally, studies by Jiao et al. (2023) and Hendy et al. (2023) compare different GPT models with supervised models for machine translation tasks and find that GPT models have competitive translation abilities in high-resource languages but perform less effectively in low-resource languages. It is worth noting that achieving multilingual generative AI capability necessitates cross-lingual knowledge to further improve the model's performance. In this context, Ahuja et al. (2023) evaluate the multilingual task understanding ability of GPT models and attempt to enhance their task processing abilities in other languages using English knowledge. Our work also focuses on evaluating the multilingual capabilities of LLMs, including reasoning, understanding, and generative capabilities. Our evaluations indicate that LLMs exhibit differences in high-resource and low-resource abilities, which necessitates additional efforts to enhance their multilingual capability.

4.2 Multilingual Task Processing

Multilingual knowledge has been shown to be exploitable and transferable between languages to improve model performance (Devlin et al., 2019; Conneau et al., 2020; Raffel et al., 2020; Ouyang et al., 2021; Chi et al., 2021). While much research has been devoted to multilingual understanding tasks, multilingual generation tasks are more challenging, particularly when the target language is low-resource or non-English (Ma et al., 2021; Liu et al., 2020). Two methods can enable models to support multilingual task processing: one is training a supervised model that covers multiple languages for multilingual processing (Costa-jussà et al., 2022), and the other is training a pre-trained model and using fine-tuning to transfer knowledge among languages to achieve multilingual capability (Chen et al., 2021, 2022). However, the emergence of LLMs has made it possible to directly process multilingual tasks via in-context learning (Brown et al., 2020; Ahuja et al., 2023). These LLMs, with hundreds of billions or even trillions of parameters, require a significant amount of computation resources for training, making traditional fine-tuning methods less feasible. To improve the generative ability of LLMs, researchers explore in-context learning methods that do not require updating model parameters, such as few-shot prompting (Vilar et al., 2022), automatic prompt learning (Shin et al., 2020), task-instruction prompting (Ye et al., 2023), chain-of-thought prompting (Wei et al., 2022c), *etc*. Our work builds upon these methods and proposes an optimized, generic, and language-independent prompt to enhance the multilingual capability of LLMs.

5 Conclusion

This work investigates the language processing capabilities of large language models in multilingual settings and expects to develop a universal framework for handling diverse multilingual tasks. To accomplish this goal, we propose a generic prompt, referred to as XLT, to enhance the multilingual capability and reduce the performance gaps among languages in tasks related to language understanding, reasoning, and generation in non-English and low-resource languages. Although our method is generally applicable across tasks and languages, we discovered that prompting design factors such as instruction logic and word choice have explicit impacts on its effectiveness. Cross-language thinking in XLT is particularly effective. Finally, we hope this work can inspire further research to prioritize the development of generic prompting. By doing so, large language models can encompass a wider range of modalities and languages.

Acknowledgements

Tianyi Tang and Xin Zhao are supported by National Natural Science Foundation of China under Grant No. 62222215, Beijing Natural Science Foundation under Grant No. 4222027 and L233008.