A Additional Experiments

A.1 Results on Reasoning Tasks

Table 7 presents the results of the MGSM benchmark. XLT significantly improves the arithmetic reasoning capabilities of both models, particularly for `gpt-3.5-turbo` in the zero-shot setting. We hypothesize that `gpt-3.5-turbo` may have undergone supervised fine-tuning (Ouyang et al., 2022) with arithmetic reasoning samples in the chain-of-thought format, which enables XLT to activate its arithmetic reasoning ability directly. For both low-resource languages (*e.g.,* sw, th, bn, and te) and high-resource languages, XLT can further enhance the performance. Even under the few-shot setting, XLT can still significantly improve the reasoning performance of both models and reduce the performance gap for all languages. Notably, for some high-resource languages, such as de, ru, fr, and es, the performance is comparable to English.

The XCOPA benchmark results are presented in Table 8. Our XLT approach significantly enhances the performance of both models in both settings, as compared to basic prompting. In the zero-shot setting, XLT demonstrates significant improvements for relatively low-resource languages (*e.g.,* sw, th, et, ta, and ht), but it underperforms the baseline for some high-resource languages such as zh and it. In the few-shot setting, XLT brings enhancements for both high- and low-resource languages. Our findings suggest that XLT is more effective for low-resource languages, particularly for `gpt-3.5-turbo` on sw, th, ta, and ht, where it yields improvements of over 10 accuracy points.

A.2 Results on Understanding Tasks

Table 9 presents the results of the XNLI benchmark. In the zero-shot setting, our XLT significantly outperforms the basic prompt in all languages. Additionally, when using few-shot setups on high- and low-resource languages, both `text-davinci-003` and `gpt-3.5-turbo` show significant improvements compared to the basic prompt. Specifically, for low-resource languages such as th, bg, hi, and ur, XLT achieves an average improvement of 9.4 accuracy scores for `text-davinci-003` and 5.3 accuracy scores for `gpt-3.5-turbo`. This demonstrates that XLT is effective for both models, but `text-davinci-003` has better natural language inference capabilities.

Table 10 displays the comparisons on the PAWS-X task, where XLT outperforms basic prompt in all languages, particularly for low-resource languages under the few-shot setting. We observe a slight performance drop on average in zero-shot learning compared to `gpt-3.5-turbo` for some high-resource languages (*e.g.,* en, de, and fr). Based on our analysis of intermediate outputs, we infer that the drop in performance may be due to cross-lingual thinking that alters the original meaning of the two sentences, leading to difficulties in judgment. Additionally, a comparable pattern is evident in a previous study (Ahuja et al., 2023), where non-Latin script languages (ja, zh, and ko) exhibit significantly poorer performance than English or German in the few-shot setting. Nevertheless, by demonstrating the construction of XLT, we can guide the model on how to think across different languages and effectively address the aforementioned issues.

A.3 Results on Generation Tasks

The MKQA benchmark outcomes are listed in Table 11. Across all languages in the zero-shot and few-shot settings, the XLT template shows a significant improvement over the basic prompt. It is worth noting that `text-davinci-003` performs worse than `gpt-3.5-turbo` in this task, and we speculate that the latter is optimized for open question answering, which is common in daily chat. Additionally, our findings indicate that XLT can notably enhance the performance of under-resourced languages. XLT brings over 10 points of improvement for these languages. (*e.g.,* zh, ja, vi, and tr) This aligns with previous benchmarking studies and is particularly noteworthy in this evaluation. We suspect that high-resource and low-resource languages share the same cross-lingual thinking as English to greatly leverage the LLM's ability to solve English open-domain QA.

The results of the XL-Sum* benchmark are presented in Table 12. It can be observed that XLT outperforms the basic prompt in both zero- and few-shot settings across all languages. Additionally, the LLM model exhibits a significant improvement in generating summaries under the few-shot setting compared to the zero-shot setting. This suggests that providing fewer examples can effectively guide the model in summarizing multilingual texts. Furthermore, the few-shot results revealed an interesting finding that `text-davinci-003` performed better when `gpt-3.5-turbo` and `text-davinci-003` use basic prompt. However, once XLT is enabled,

Table 5: The basic prompt of each benchmark. #Test denotes the number of instances in the test set.

Benchmark	#Test	Basic Prompt
MGSM	250	Request: `{problem}`
XCOPA	500	Here is a premise: `{premise}`. What is the `{question}`? Help me pick the more plausible option: -choice1: `{choice1}`, -choice2: `{choice2}`
XNLI	5,010	`{premise}` Based on previous passage, is it true that `{hypothesis}`? Yes, No, or Maybe?
PAWS-X	2,000	Sentence 1: `{sentence1}` Sentence 2: `{sentence2}` Question: Does Sentence 1 paraphrase Sentence 2? Yes or No?
MKQA	6,758	Answer the question in one or a few words in `{target_language}`: `{question}`?
XL-Sum*	250	Summarize this article: `{article}`
FLORES*	200	`{source}` Translate from `{source_language}` to `{target_language}`:

Table 6: Task meta data consisting of task name, input tag, task goal, output type, and output constraint per benchmark. Detailed examples of the input for each benchmark are listed in the following part.

Benchmark	Task name	Input tag	Task goal	Output type	Output constraint
MGSM	arithmetic reasoning	request	do step-by-step answer to obtain a number answer	answer	–
XCOPA	commonsense reasoning	premise and the options	do step-by-step answer to pick a choice	choice number	–
XNLI	natural language inference	hypothesis and the premise	judge whether the hypothesis is true, false, or undetermined given the premise. The relationship can be chosen from entailment, contradiction, and neutral	relationship	–
PAWS-X	paraphrase identification	sentence 1 and sentence 2	provide a yes or no answer to the question: Does Sentence 1 paraphrase Sentence 2?	answer	choosing either yes or no
MKQA	question answering	question	answer the question in English in one or a few words	answer	in one or a few words in {target_language}
XL-Sum	multilingual summarization	entire text	think step-by-step to summarize the entire text in a maximum of two sentences	summary	into one sentence in {target_language}
FLORES	machine translation	source sentence	provide the {target_language} translation for the English source sentence	target translation	–

`gpt-3.5-turbo` outperforms `text-davinci-003`, highlighting the effectiveness of our approach.

Machine translation is a special generation task where the source and target are two different languages. The experiment in this part is to verify how XLT boosts machine translation tasks. Since English has been specified as the pivot language in the cross-lingual thinking in XLT, we exclude English-centric tasks to avoid language redundancy and focus on 12 non-English translation directions in the FLORES* benchmark, which includes both high-resource and low-resource languages. As shown in Table 13, XLT achieves impressive zero-shot results for all languages compared with basic prompt. For example, it significantly improves translation quality in Chinese-to-X or X-to-Chinese. The result emphasizes that XLT will potentially transfer the knowledge of a high-resource pivot language like English to the target language. While the benefit of XLT may not be as obvious for high-to-high translations, it becomes more significant for high-to-low, low-to-high, and low-to-low translations. For instance, XLT improves the translation performance of `gpt-3.5-turbo` by nearly 4.0, 2.8, and 3.3 BLEU points for th→gl, jv→zh, and zh→th translations, respectively, demonstrating its effectiveness regardless of whether the source language is high-resource or low-resource. Noticing that Hendy et al. (2023) have shown that few-shot configurations do not yield significant improvements over the zero-shot setup for translation tasks, we do not evaluate the few-shot paradigm on FLORES* in this work and leave it for future exploration.

Settings (high→low)		en	de	ru	fr	zh	es	ja	sw	th	bn	te	Avg.
Zero-shot	`text-davinci-003`												
	Basic Prompt	19.2	12.8	15.6	16.4	15.2	13.6	12.8	7.2	8.8	11.6	4.4	12.5
	XLT	30.0	32.4	23.6	34.8	29.2	26.8	26.0	13.6	18.4	14.8	12.8	**23.9**
	`gpt-3.5-turbo`												
	Basic Prompt	32.0	24.8	28.0	31.6	22.0	29.2	22.4	24.4	16.8	18.0	7.6	23.3
	XLT	84.4	79.8	77.6	75.2	72.6	76.8	71.0	70.8	63.8	56.8	42.0	**70.0**
	`Llama-2-70b-chat-hf`												
	Basic Prompt	58.8	48.0	47.2	45.6	39.6	50.4	39.2	10.0	13.6	17.2	5.2	34.1
	XLT	60.0	52.8	52.8	48.8	42.4	52.0	39.2	16.4	18.0	17.6	10.4	**37.3**
Few-shot	`code-davinci-002` (Shi et al., 2023)*	53.6	46.4	48.8	46.4	47.2	51.6	44.8	37.6	41.2	41.2	42.8	45.6
	`text-davinci-003`												
	Basic Prompt	60.4	45.6	51.6	45.6	38.8	51.6	37.6	48.8	30.4	43.6	46.8	45.5
	XLT	65.6	58.0	57.6	56.8	53.2	58.0	54.4	58.8	42.4	53.2	51.8	**55.4**
	`gpt-3.5-turbo`												
	Basic Prompt	82.8	69.2	71.6	72.4	46.8	71.2	56.0	60.0	44.0	62.4	56.6	63.0
	XLT	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 7: Accuracy scores on the MGSM benchmark. Shi et al. (2023)* utilize 6-shot learning.

Settings (high→low)		zh	it	vi	tr	id	sw	th	et	ta	ht	qu	Avg.
Zero-shot	`text-davinci-003`												
	Basic Prompt	85.4	90.0	69.2	80.6	83.8	56.4	66.6	73.0	53.4	61.6	50.4	70.1
	XLT	85.8	89.2	76.0	81.0	86.4	59.2	67.2	83.4	55.2	72.2	50.2	**73.3**
	`gpt-3.5-turbo`												
	Basic Prompt	90.4	92.0	83.6	86.6	88.2	77.0	70.2	84.0	57.2	65.2	51.2	76.9
	XLT	87.8	89.8	87.5	90.2	89.5	82.0	78.0	88.4	64.0	74.6	51.8	**80.3**
Few-shot	`code-davinci-002` (Shi et al., 2023)*	93.4	96.6	86.6	91.2	91.4	67.4	84.2	88.8	55.8	79.6	52.2	80.7
	`text-davinci-003`												
	(Ahuja et al., 2023)†	–	94.6	–	89.8	93.0	82.8	84.8	89.6	87.0	82.8	–	–
	Basic Prompt	90.8	92.2	80.2	85.2	90.8	63.6	69.2	81.8	53.6	73.2	51.0	75.6
	XLT	94.0	95.0	87.0	94.0	92.8	68.4	79.4	90.4	59.4	80.8	53.0	**81.3**
	`gpt-3.5-turbo`												
	Basic Prompt	91.0	95.2	86.2	89.0	88.6	79.2	73.6	92.0	58.6	74.2	53.0	80.1
	XLT	92.8	95.8	90.6	92.2	90.2	92.6	85.2	93.0	70.8	86.0	56.2	**85.9**

Table 8: Accuracy scores on the XCOPA benchmark. (Shi et al., 2023)* utilize 6-shot learning. Ahuja et al. (2023)† utilize 8-shot learning.

Settings (high→low)		en	de	ru	fr	zh	es	vi	tr	sw	ar	el	th	bg	hi	ur	Avg.
Zero-shot	`text-davinci-003`																
	Basic Prompt	63.6	59.4	55.9	60.9	51.6	59.7	49.5	53.9	40.8	51.9	53.2	49.7	54.4	49.8	45.3	53.3
	XLT	77.4	67.7	64.2	68.3	64.8	69.4	62.0	61.5	54.3	58.7	61.1	56.3	62.6	55.1	53.0	**62.4**
	`gpt-3.5-turbo`																
	Basic Prompt	65.4	55.5	50.6	53.2	48.8	59.8	52.1	54.4	49.6	50.9	54.9	44.8	55.7	49.2	44.8	52.6
	XLT	74.4	68.5	66.0	69.8	64.9	69.4	64.8	65.0	60.1	62.8	68.3	62.1	67.7	61.3	57.3	**65.5**
Few-shot	`text-davinci-003`																
	(Ahuja et al., 2023)†	79.5	71.7	67.3	71.8	65.8	72.2	66.9	67.6	57.3	65.1	69.3	62.0	70.8	63.3	55.1	67.1
	Basic Prompt	71.6	65.8	62.5	63.4	56.7	64.6	59.4	56.9	48.2	57.3	62.0	55.0	62.6	52.4	48.0	59.1
	XLT	79.1	70.8	70.0	69.5	69.2	71.0	67.3	66.9	59.5	65.7	67.8	63.7	70.4	63.5	58.1	**67.5**
	`gpt-3.5-turbo`																
	Basic Prompt	73.4	66.3	60.9	67.9	60.2	68.1	60.2	62.6	55.7	58.8	64.7	52.7	64.6	53.8	50.8	61.4
	XLT	77.1	69.3	64.4	69.6	62.9	70.6	63.2	64.4	60.2	63.4	66.6	59.8	66.9	60.0	56.5	**65.0**

Table 9: Accuracy scores on the XNLI benchmark. Ahuja et al. (2023)† utilize 8-shot learning.