SoT: Structured-of-Thought Prompting Guides Multilingual Reasoning in Large Language Models

Anonymous ACL submission

Abstract

Recent developments have enabled Large Language Models (LLMs) to engage in complex reasoning tasks through deep thinking. However, the capacity of reasoning has not been successfully transferred to non-high-resource languages due to resource constraints, which struggles with multilingual reasoning tasks. To this end, we propose Structured-of-Thought (SoT), a training-free method that improves the performance on multilingual reasoning through a multi-step transformation: Language Thinking Transformation and Structured Knowledge Transformation. The SoT method converts language-specific semantic information into language-agnostic structured representations, enabling the models to understand the query in different languages more sophisticated. Besides, SoT effectively guides LLMs toward more concentrated reasoning to maintain consistent underlying reasoning pathways when handling cross-lingual variations in expression. Experimental results demonstrate that SoT outperforms several strong baselines on multiple multilingual reasoning benchmarks when adapting to various backbones of LLMs. It can also be integrated with other training-free strategies for further improvements. Our code is available at https://anonymous.4open.science/r/SoT-E461/.

1 Introduction

Large language models (LLMs) have demonstrated exceptional performance in a wide range of tasks (Radford et al., 2019; Huang et al., 2025), especially in enhancing reasoning abilities (Brown et al., 2020). Although the existing LLMs demonstrate multilingual understanding ability, a performance gap is observed between different languages. This is because most large-scale datasets used for model training are predominantly available in widely spoken languages, such as English and Mandarin (Huang et al., 2023; Shi et al., 2023).

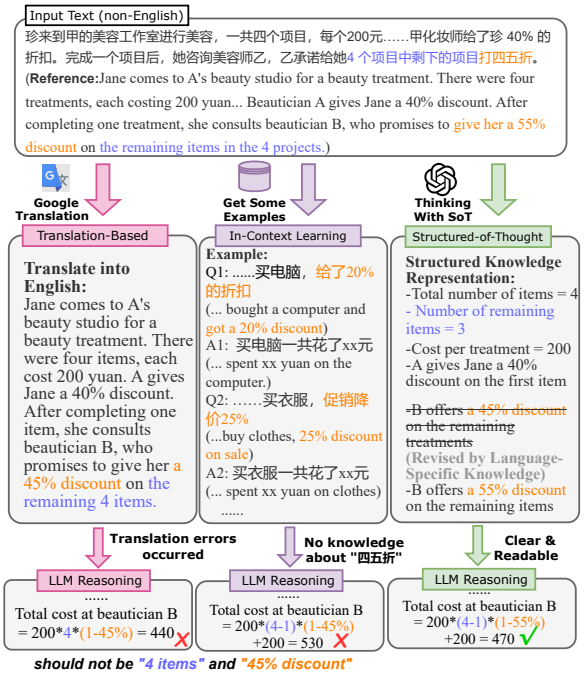

Figure 1: Examples of multilingual mathematical reasoning. When dealing with questions with complex semantic structures and language-specific expressions, LLM generate correct and incorrect answers using different prompts in non-English languages.

An intuitive solution to mitigate this gap is to supplement multilingual data for post-training (Huang et al., 2025). However, this is infeasible as it requires language-specific training corpora for each language, while many languages are inherently low-resource (Ghosh et al., 2025; Ji et al., 2025). Moreover, for each LLM, the post-training process demands substantial time and computational resources, which results in poor scalability for deployment in practice (Zhu et al., 2023; Li et al., 2024). Thus, a more appropriate approach is *to enable LLM to enhance multilingual reasoning performance under training-free conditions*, and has drawn much attention in recent studies (Li et al., 2023; Zhu et al., 2024c; Koo and Kim, 2025).

In this scenario, previous methods aim to improve the multilingual understanding of LLMs by reformulating non-English queries, including translation-based strategies (Huang et al., 2023; Shi et al., 2023) and in-context learning (Brown et al., 2020; Zhang et al., 2023; Asai et al., 2023; Ahuja et al., 2023; Zhu et al., 2024b). The former approach relies on the availability of high-quality translations (Bawden and Yvon, 2023), whereas the latter would not be able to capture critical information and features without the provided well-crafted context (Zhang et al., 2024). An example of LLMs answering a mathematical problem with different prompts in a non-English language is shown in Figure 1. The complex semantic structures in non-English languages lead to misinterpretations of inter-entity relations, hindering accurate recognition of problems and consequently resulting in poor reasoning performance. No matter how an identical mathematical question is formulated, its underlying reasoning process should be kept the same (Hu et al., 2025). Therefore, enabling LLMs to interpret problem statements accurately is crucial to establishing correct reasoning pathways in multilingual settings.

Considering the inherent reasoning capabilities of LLMs and the varying levels of difficulty in query comprehension, in this paper, we propose structured-of-thought (SoT), a thinking strategy that incorporates structured representations into the reasoning pathway to mitigate the misinterpretation of LLMs in multilingual scenarios. In particular, SoT elicits LLMs to align their reasoning pathways for non-English inputs with those thought in English via a multi-step transformation: Language Thinking Transformation and Structured Knowledge Transformation. Beyond the mere conversion of language thinking, natural language queries are also converted into structured knowledge representations, allowing the LLMs to not only understand the context from the surface-level linguistic, but also can identify the underlying relational semantics, i.e., to achieve the equivalence of semantic understanding between expressions "*0.75 bags per guest*" and "*1/4 of the guests will not attend*". Besides, structured knowledge transformation can guide LLMs toward more concentrated reasoning by eliminating extraneous information that otherwise disrupts the inference process. Experiments show that our SoT outperforms several state-of-the-art baselines on mathematical and commonsense reasoning tasks, and is applicable to a variety of backbone LLMs.

Our contributions are summarized as: (1) We propose a Structured-of-Thought prompting method to guide LLMs to align the reasoning pathways for non-English queries, thereby enhancing the reasoning capabilities in the multilingual scenarios. (2) Our strategy can be integrated with other training-free prompting strategies, such as In-Context Learning (ICL) and Chain-of-Thought (CoT), which achieves the further improvement for multilingual reasoning. (3) Experiments demonstrate that our method can accurately understand the structural knowledge in queries to adapt various series of LLMs of different sizes on several multilingual reasoning benchmarks.

2 Related Work

Multilingual Reasoning. A common practice to enhance the multilingual reasoning capabilities of LLMs is based on model supervised fine-tuning (SFT) (She et al., 2024; Zhu et al., 2024a; Chai et al., 2025). However, SFT suffers from data scarcity and catastrophic forgetting, and lacks the generalization ability (She et al., 2024). Another research line explored the usage of carefully designed prompts to support reasoning in LLMs (Huang et al., 2023; Qin et al., 2023). For instance, the pre-translation approach translates input questions into a high-resource pivot language (*e.g.*, English) before querying the LLM, aiming to leverage the stronger proficiency of models in the pivot language (Etxaniz et al., 2024; Huang et al., 2025). Furthermore, the pre-translation method can be integrated with other prompting strategies (Lu et al., 2024; Koo and Kim, 2025; Zhu et al., 2024c), such as CoT (Wei et al., 2022) and ICL (Brown et al., 2020) paradigms. Besides, Liu et al. (2024) propose several strategies to extend CoT to multilingual contexts. Different from them, our method introduces a structured-based strategy that leverages the built-in capabilities of LLMs to mitigate the misinterpretation of semantic for multilingual reasoning.

Chain-of-Thought. CoT prompting (Wei et al., 2022; Kojima et al., 2022) is an effective step-by-step strategy for LLMs' zero-shot and few-shot reasoning. A series of CoT-based techniques has been proposed to further improve the reasoning performance of LLMs, including Complex CoT (Fu et al., 2023), Decomposed Prompting (Khot et al., 2022), Multilingual CoT (Shi et al., 2023), Least-to-Most

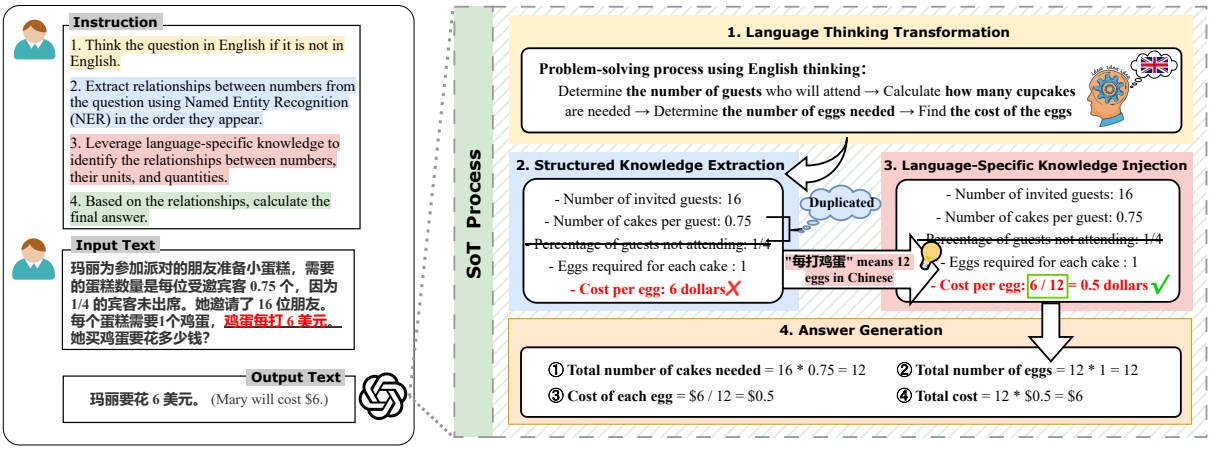

Figure 2: Overview of of the SoT strategy. The left part is an example of the question and our instructions. The right part is the thinking process of LLM under the guidance of SoT.

Prompting (Zhou et al., 2022), and Progressive-Hint Prompting (Zheng et al., 2023). Except for exploring a CoT variant, some approaches introduce a structured representation to capture dependencies among entities for complex reasoning tasks in the thinking step (Wang et al., 2024). In particular, Cheng et al. (2024) investigate the effectiveness of graph structure of the text in multi-step reasoning. Due to the limitations on foundational abilities of multilingualism (Huang et al., 2025), our method attempts to exploit a more concise form to structure the knowledge in queries, which is more effective for multilingual scenarios.

3 Methods

In multilingual reasoning tasks, complex semantic structures in non-English languages might obscure the relationships between entities, thereby impeding the accurate interpretation of the question. To this end, we propose SoT, a zero-shot method designed to enhance the reasoning capabilities of LLMs in multilingual scenarios through multi-step transformations. Our SoT framework consists of four steps as illustrated in Figure 2.

The principle of our SoT is to structure the input questions by transforming reasoning pathways expressed in natural language into structured representations that are more easily interpreted by LLMs. This restructuring manipulation improves the abilities of models to reason accurately across languages. In contrast to other training-free methods, SoT specifically targets the comprehension of complex semantic relationships within the questions. Regardless of the language in which the same question is posed, SoT allows models to fully leverage their built-in reasoning capabilities to enable LLMs to maintain correct and consistent reasoning pathways. Moreover, the framework is generalizable and can be applied across a wide range of multilingual reasoning tasks.

3.1 Language Thinking Transformation

Step 1 : *Think the question in English if it is not in English*

When the model targets the same question in different languages, its reasoning pathway should be consistent. Thus, we conduct the transformation of the reasoning process from low-resource to high-resource languages by cross-lingual transfer, enabling the LLMs to perform reasoning in a language in which they exhibit greater proficiency under multilingual scenarios.In particular, we leverage the inherent reasoning and language understanding capabilities of LLMs, eliminating the need for development from scratch. To effectively transfer the reasoning pathway into the high-resource language, we introduce a Language Thinking Transformation strategy, as illustrated in the first step in Figure 2.

Specifically, given the sentence X, we conduct the transformation from the source language L_s to the target language L_t (*i.e.,* English). The intermediate thinking pathways $\mathcal{R}$ are represented as $\{r_i\}_{i=1}^{n}$, where n denotes the number of thinking steps. Formally, the Language Thinking Transformation process is expressed as follows:

$$\mathcal{R} = \arg\max p(r_1, \ldots, r_n | X, L_s, L_t) \quad (1)$$

3.2 Structured Knowledge Extraction

Step 2 : *Extract relationships between numbers from the question using Named Entity Recognition (NER) in the order they appear.*

After performing language transfer for reasoning, the knowledge from the question is extracted and then represented in a structured natural language format. Specifically, the elements of structured knowledge mainly consist of entities and their relationship patterns. Thus, we instruct the LLM to perform Named Entity Recognition (NER) to identify key elements such as numerical values, units, and their associated relationships within the input text. The objective of this step is to construct a structured representation of knowledge $\mathcal{K}$ that enables the LLM to accurately identify and comprehend the core entities and their interrelations within the question. Formally, the Structured Knowledge Extraction process is expressed as follows:

$$\mathcal{K} = \arg\max p(k_1, \ldots, k_m | \mathcal{R}, X, L_t), \qquad (2)$$

where $\{k_i\}_{i=1}^{m}$ represents the pattern of structured knowledge and m denotes the number of the patterns.

The construction of the structured representation eliminates irrelevant information from the input, making the relation among the values and entities much clearer and thus facilitating the subsequent reasoning steps with less noise. For example, NER can facilitate the relation identification between numbers and entities in mathematical problems (*e.g.*, *0.75 per guest*and *1/4 of guests will not attend* represent the same relationship in different expressions) as shown in second step in Figure 2. Moreover, knowledge extraction can simplify complex problems, making them more interpretable and enhancing the capacity of LLMs to perform reasoning tasks.

3.3 Language-Specific Knowledge Injection

Step 3 : *Leverage language-specific knowledge to identify the relationships between numbers, their units, and quantities.*

Although the language transfer in thinking helps the LLM better interpret the problem, it neglects language-specific differences of expression in terms of quantities, units, and their relations. To address this, the third step in our approach aims to further enhance the understanding of non-English languages by guiding LLMs to focus on language-specific knowledge. Each language possesses unique rules and conventions for expressing numerical relations and quantities. For example, in Chinese, the phrase "四五折" denotes a 55% discount, which might lead to misinterpretation if processed without cultural or contextual awareness. LLMs might not be able to distinguish that they have the same meaning when performing calculations directly. An alternative is to leverage translation-based strategies as intermediate support, which would still fail to capture these nuances accurately. Guided by language-specific expressions, the LLM can accurately understand these nuances, reducing misunderstandings caused by linguistic variation and improving reasoning performance across languages. Formally, the Language-Specific Knowledge process is expressed as follows:

$$\mathcal{K}^{L_s} = \arg\max p(k_1^{L_s}, \ldots, k_m^{L_s} | \mathcal{K}, L_s), \qquad (3)$$

where $\{k_i^{L_s}\}_{i=1}^{m}$ represents the language-specific knowledge.

3.4 Answer Generation

Step 4 : *Based on the relationships, calculate the final answer in the Source Language.*

The final stage is to integrate the above information, where the LLM conducts reasoning based on the extracted structured knowledge, language-specific knowledge, and the results of the language thinking transformation, towards the final answer $\mathcal{F}$. The answer is transformed back into the source language L_s, ensuring consistency between input and output to maintain interpretability in multilingual scenarios. Formally, the generation of the final answer is determined as:

$$\mathcal{F} = \arg\max p(f | \mathcal{R}, \mathcal{K}, \mathcal{K}^{L_s}, L_s) \qquad (4)$$

4 Experiments

4.1 Experimental Setup

Models. We select three series of LLMs to verify the effectiveness of SoT: **`gpt-3.5-turbo`**, **`Qwen2.5-7B-Instruct`** and **`DeepSeek-R1-7B`**, including both open-source and closed-source models, ranging from past to latest. To further demonstrate the robustness on larger models, we utilize **`Qwen2.5-32B-Instruct`** as the basic model.

Methods	Language										
	En	Sw	Ja	Be	Th	Ru	Zh	De	Es	Fr	Avg.
(**training-free**)	(DeepSeek-R1-7B)										
Direct	82.0	18.6	67.8	52.6	53.8	80.2	78.0	73.0	80.4	71.8	65.8
DoLa	83.8	18.7	70.1	54.0	62.2	83.0	81.3	75.3	80.9	74.0	68.3
SL-D	84.1	22.6	73.1	55.7	64.2	84.8	84.3	79.0	81.6	77.1	70.7
DIP	88.0	21.4	82.0	63.5	64.5	83.2	85.0	82.1	83.0	83.4	73.6
CLP	89.6	23.2	77.0	62.7	69.3	78.5	84.8	81.4	81.8	87.0	73.5
EMCEI	89.0	23.0	80.0	61.0	64.9	83.8	86.2	83.2	83.4	84.9	73.9
SoT (*Ours*)	**89.8**	**24.8**	**82.8**	**64.6**	**71.8**	**85.4**	**87.2**	**85.4**	**85.2**	**88.2**	**76.5**
(**post-training**)											
xCoT	84.7	50.7	79.6	59.0	64.6	80.3	83.2	82.7	85.1	88.3	75.8
QAlign	82.8	46.2	82.6	56.0	64.5	81.4	80.3	86.6	89.8	89.1	75.9
MindMerger	83.6	44.6	83.4	56.6	59.7	81.2	84.6	87.4	89.1	92.2	76.2
MAPO	84.8	50.2	83.8	53.6	64.9	80.5	84.8	83.2	88.2	85.2	75.9
(**training-free**)	(Qwen2.5-7B-Instruct)										
Direct	89.8	39.4	69.2	55.0	65.4	74.6	81.8	77.8	83.2	82.2	71.8
DoLa	91.0	54.4	73.1	64.7	74.5	76.4	83.3	79.2	85.3	85.7	76.8
SL-D	91.5	56.0	75.1	66.7	77.2	77.1	85.4	81.6	88.2	87.5	78.6
DIP	88.3	52.1	86.3	**77.1**	76.1	84.4	87.8	91.2	88.0	90.1	82.1
CLP	90.2	50.3	80.6	67.4	74.4	79.0	82.2	85.1	83.9	87.0	78.0
EMCEI	89.6	58.2	86.7	74.6	75.2	86.0	87.7	90.6	89.4	89.3	82.7
SoT (*Ours*)	**93.6**	**61.0**	**87.6**	76.4	**83.8**	**87.4**	**89.4**	**91.6**	**91.8**	**91.2**	**85.4**
(**post-training**)											
xCoT	85.0	60.1	81.0	62.4	66.3	84.1	85.2	88.5	90.3	89.0	79.2
QAlign	80.3	52.1	83.0	59.6	64.9	85.6	81.4	92.5	93.3	89.9	78.3
MindMerger	81.5	51.0	84.5	58.3	59.6	83.4	90.9	89.2	90.0	93.7	78.2
MAPO	84.6	57.6	85.2	53.0	68.0	84.2	84.7	84.5	88.4	85.4	77.6

Table 1: Results (%) of mathematical reasoning on MSVAMP. For all **training-free** methods, the **bold** text represents the highest scores, while the underline represents the second highest scores.

Benchmarks and Evaluation. To ensure the reliability of the experiments, all methods are implemented on two mathematical reasoning tasks (MGSM (Shi et al., 2023) and MSVAMP (Chen et al., 2024)) and one commonsense reasoning task (XCOPA (Ponti et al., 2020)). Benchmark details are listed in Appendix A.1. We employ the accuracy to access the ability of the methods for all tasks (Jin et al., 2024).

Baselines. For comparison, we select recent advanced training-free methods (*e.g.*, DoLa (Chuang et al., 2024), SL-D (Zhu et al., 2024c), DIP (Lu et al., 2024), CLP (Qin et al., 2023), EMCEI (Koo and Kim, 2025).) and effective post-training methods (xCoT (Chai et al., 2025), QAlign (Zhu et al., 2024a), MindMerger (Huang et al., 2024) and MAPO (She et al., 2024)). We follow the original settings of the original paper. More details of baselines are listed in Appendix A.2.

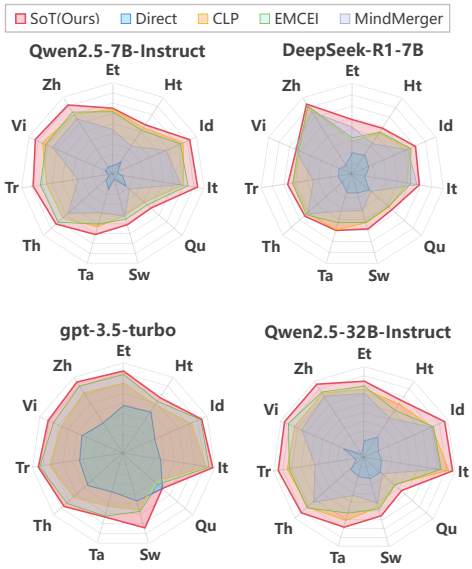

Figure 3: Results of commonsense reasoning on XCOPA using various LLMs.

4.2 Main Results

Performance on Mathematical Reasoning. As shown in Table 1 and Table 2, we investigate the mathematical abilities of LLMs with different methods that facilitate multilingualism across various languages. The results demonstrate that our proposed method (SoT) outperforms several baselines in terms of average accuracy, including the training-free and post-training methods. The training-free

Methods	Language											
	En	**Sw**	**Ja**	**Be**	**Th**	**Te**	**Ru**	**Zh**	**De**	**Es**	**Fr**	**Avg.**
(**training-free**)	(DeepSeek-R1-7B)											
Direct	75.2	7.2	42.4	43.6	41.6	18.0	65.6	72.0	50.0	64.0	55.6	48.7
DoLa	75.8	8.0	43.4	46.4	45.4	18.0	60.2	71.2	59.2	66.2	53.8	49.8
SL-D	77.2	8.4	57.0	47.0	46.0	20.0	62.4	72.8	62.6	64.6	62.0	52.7
DIP	80.0	6.0	51.2	48.2	57.8	24.2	64.8	75.4	60.6	67.0	63.0	54.4
CLP	**87.0**	9.0	60.4	50.0	56.8	19.0	61.2	71.8	65.8	67.4	65.0	55.8
EMCEI	81.2	7.2	58.2	46.4	57.0	18.4	64.2	74.6	63.6	68.0	67.2	55.1
SoT (*Ours*)	84.4	**10.0**	**61.2**	**51.2**	**61.2**	**28.0**	**70.0**	**76.4**	**70.0**	**71.6**	**68.0**	**59.3**
(**post-training**)												
xCoT	82.2	43.4	62.0	56.6	55.8	10.0	71.4	75.4	67.2	74.2	67.0	60.5
QAlign	81.6	42.8	60.4	53.6	53.0	11.4	69.6	74.0	68.6	72.2	66.2	59.4
MindMerger	80.0	41.6	60.8	54.2	53.8	12.8	70.2	75.8	69.8	73.4	65.2	59.8
MAPO	86.2	42.2	61.6	53.2	59.4	12.0	69.6	78.0	67.4	71.8	61.6	60.3
(**training-free**)	(Qwen2.5-7B-Instruct)											
Direct	84.0	12.8	56.0	51.2	48.0	24.0	73.6	**80.8**	66.8	71.2	64.4	57.5
DoLa	83.2	13.8	61.0	61.2	54.4	32.0	75.4	75.0	69.2	73.4	67.8	60.6
SL-D	84.6	10.4	63.2	63.2	54.4	34.6	76.2	76.0	70.6	74.2	69.0	61.5
DIP	84.4	24.2	70.4	66.8	64.4	33.2	78.0	76.6	70.0	78.0	74.2	65.5
CLP	84.2	20.0	70.8	64.4	65.4	30.0	78.8	76.0	71.0	77.8	72.0	64.6
EMCEI	84.8	27.0	71.0	68.2	72.0	31.0	76.6	75.6	72.0	78.4	73.0	66.3
SoT (*Ours*)	**85.6**	**28.0**	**71.8**	**69.6**	**74.0**	**36.4**	**80.8**	77.0	**72.8**	**79.6**	**75.2**	**68.3**
(**post-training**)												
xCoT	85.6	47.2	64.2	62.2	61.8	12.6	79.4	85.2	70.2	79.0	78.2	66.0
QAlign	84.6	45.8	60.8	61.4	62.4	13.2	75.8	81.6	72.0	72.6	73.2	63.9
MindMerger	82.4	44.4	62.4	56.2	59.4	12.0	79.0	85.4	70.0	69.2	69.8	62.7
MAPO	88.4	46.0	63.0	58.8	62.2	12.4	78.3	88.3	68.2	71.0	68.0	64.1

Table 2: Results (%) of mathematical reasoning on MGSM. For all **training-free** methods, the **bold** text represents the highest scores, while the underline represents the second highest scores.

methods focus on stimulating the inherent knowledge of the foundational LLMs, which can achieve gains in most languages with decreasing cost. However, due to the inherent defects of the model, it is difficult to achieve significant improvement for languages with insufficient inherent knowledge of the model. Although the post-training methods can alleviate this issue, these methods face limitations in data construction, where the effects achieved in different languages and tasks are unstable. Moreover, the more strengthful model diminishes the effectiveness of post-training methods, which opposes the core advantages of our method. All the results using various LLMs are listed in Appendix C.

Performance on Commonsense Reasoning. As shown in Figure 3, we also investigate the effectiveness of SoT on the commonsense reasoning task, compared with other methods. The results demonstrate that the advantages of SoT are further enhanced, which has an obvious improvement over the original method (*Direct*), compared with other baselines. In particular, the structured knowledge in our method can not only extract the computational relationships for reasoning in mathematical problems, but also enable LLMs to deeply think about the logical relationships between entities in commonsense reasoning. The post-training method does not show gains similar to those in the supervised direction for low-resource langauges (*e.g.,* a significant improvement on Sw) due to the limitation of the corpus, while the training-free methods demonstrate better generalization, especially SoT.

No.	Multi-Step Scopes			Avg.
	Step 1	Step 2	Step 3	
1	×	×	×	37.3
2	✓	×	×	40.0
3	×	✓	×	53.2
4	×	×	✓	58.1
5	✓	✓	×	60.8
6	✓	×	✓	61.2
7	×	✓	✓	61.6
8	✓	✓	✓	**62.8**

Table 3: Results of different prompting strategies on MGSM and **gpt-3.5-turbo** in terms of average scores.

4.3 Ablation Studies

Effects of Muti-Step Scopes. As shown in Table 3, we explore the contribution of each step in

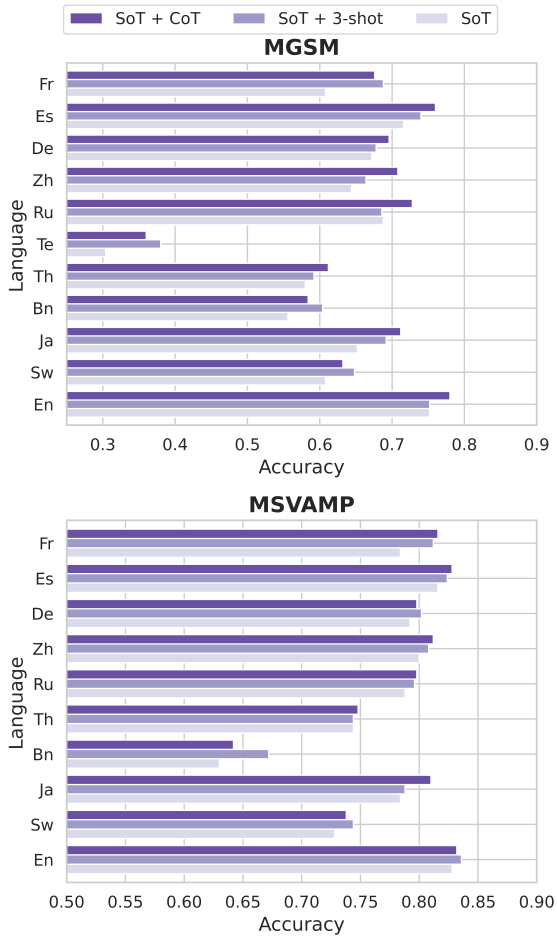

Figure 4: Results of SoT combined with CoT and few-shot (3-shot) on the MGSM and MSVAMP using **`gpt-3.5-turbo`**.

SoT. The results demonstrate that our method can help queries in diverse languages to be better understood and achieves better performance when both three steps are considered through SoT for mathematical reasoning in the multilingual scenarios. Specifically, each individual step in SoT has a positive impact, according to the comparison among the Strategies No.1, 2, 3 and 4. Furthermore, the two combined forms further enhance the reasoning performance in terms of the Methods No.5, 6 and 7. Except for SoT (No.8), the results show that the structured extraction and language-specific knowledge (No.7) are more important and achieve the highest performance (61.6%), indicating that language transfer thinking has a positive impact, but is not an indispensable factor.

Effects of Integrated Methods. As shown in Figure 4, we explore the feasibility of SoT when combined with other training-free methods such

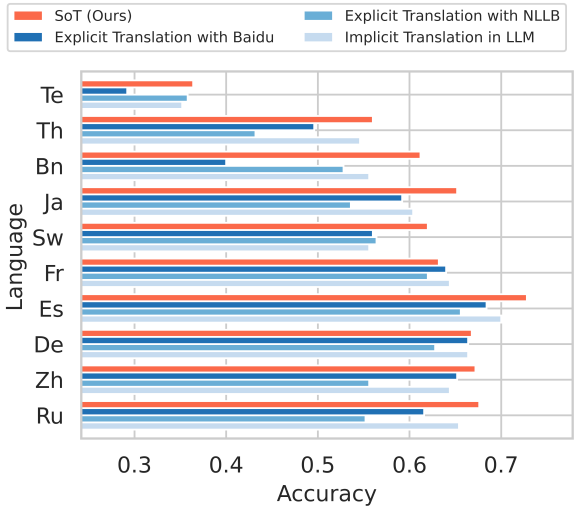

Figure 5: Results on thinking and translation. SoT employs the thinking manner, while other methods replace the language thinking with 3 translation processes.

as CoT and ICL. The results show that the adoption of CoT or ICL further improves SoT performance, demonstrating that SoT does not have conflicts with other training-free methods. Specifically, CoT achieves better performance in high-resource languages, while ICL is more proficient in low-resource languages. A possible reason is that CoT is suited to guide intrinsic knowledge in LLMs and ICL provides the language knowledge which is a supplement to low-resource languages for LLMs. More comparison is shown in Appendix C.

4.4 Results on Thinking and Translation

As shown in Figure 5, we explore the effectiveness of *Step 1* (Language Thinking Transformation) which is replaced with the translation process. Previous studies attempt to translate original queries into a high-resource language (e.g., English), which avoids the problem of insufficient abilities in the source language. Formally, we modify the instruction of *Step 1* as follow:

> ***Step 1 (Thinking ⇒ Translation)***:
> *Type 1: Translate the question into English if it is not in English for the following step.*
> *Type 2:* *[Outputs by Translators $\mathcal{T}_s$]* *is the translation of question for the following step.*

We divide the translation methods into two types: The first is to replace "thinking" in the instruction with "translate" for implicit translation (*i.e.,* no translation result is generated). The second is to replace the instruction with the explicit translation by the external translator. The results reveal that

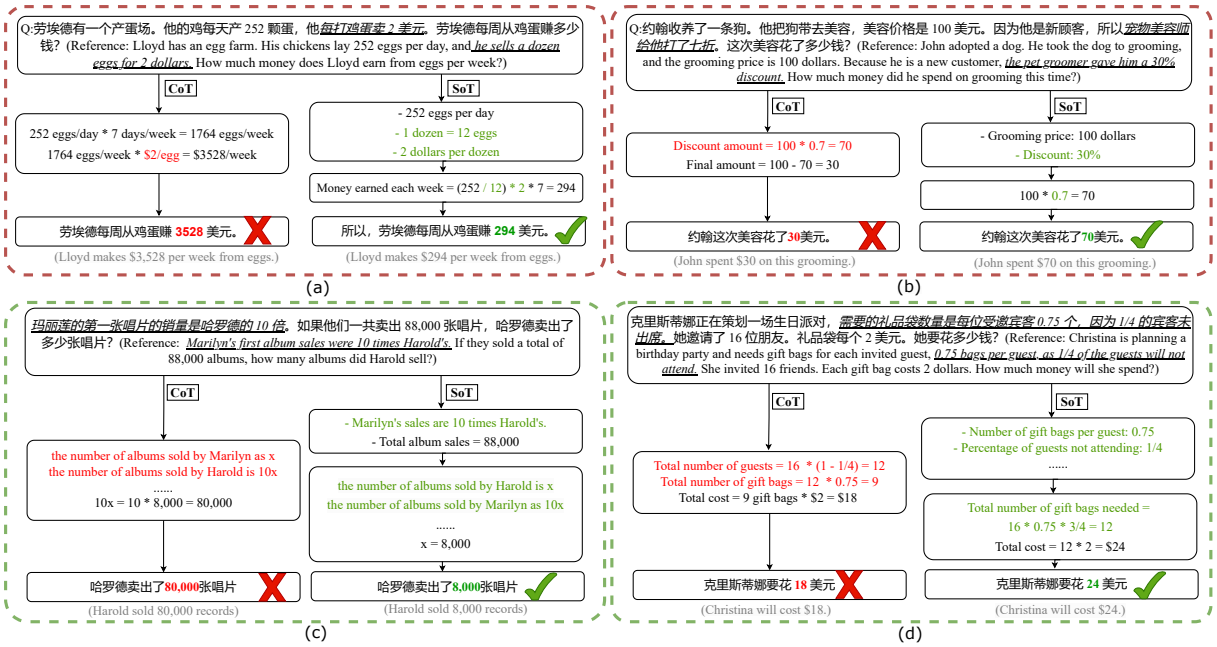

Figure 6: Examples of CoT and SoT on the mathematical reasoning tasks. We only highlight some words and fragments to show the representative difference between the two methods. The red parts represent the misunderstanding, while the green parts represent a correct understanding.

the robustness of thinking transformation is better than that of translation, in which the reasoning performance of the translation manner is influenced by the translation qualities of the source language. Translation errors will accumulate and be passed on to subsequent steps via either implicit translation (LLM translator) or explicit translation (Baidu or NLLB translator), causing performance degradation.

4.5 Case Study

As shown in Figure 6, we present examples in MGSM where the traditional CoT method fails, while our framework produces accurate results. The cases highlight the effectiveness of our approach to resolve common errors in multilingual reasoning. Examples in MSVAMP and XCOPA can be found in Appendix 6.

As shown in Figure 6.a and Figure 6.b, CoT suffers from misinterpreting units and discounts due to language-specific ambiguities. For instance, in Figure 6.a, CoT confuses "per dozen eggs" with "per egg", leading to an incorrect calculation. Similarly, in Figure 6.b, the expression "70% off" is misunderstood by CoT as "a 70% reduction" in Chinese, rather than "70% of the original price". SoT effectively resolves these issues by incorporating structured and language knowledge, ensuring correct numerical interpretation.

As shown in Figure 6.c and Figure 6.d, the illustrations demonstrate the structural knowledge leads to the misunderstandings for reasoning. In Figure 6.c, CoT fails to parse the relationship between two sales figures of entities, leading to cascading errors through the reasoning process. In Figure 6.d, CoT misinterprets "0.75 bags per guest" and "1/4 of guests not attending" as separate conditions, leading to double counting. SoT understands these relationships explicitly, preventing such misunderstandings. In general, SoT facilitates the model to interpret relationships accurately by integrating with structured knowledge and language-specific knowledge, reducing errors caused by ambiguous expressions in different languages.

5 Conclusion

In this paper, we propose a Structured-of-Thought (SoT) method for multilingual reasoning. By dynamically extracting entity-structured knowledge and language-specific structured knowledge, our method boost the ability to understand relationships in non-English questions for LLMs. Experimental results demonstrate that SoT achieves comparable performance on various LLMs, compared with several advanced methods. The analyses further indicates that SoT has both strong generalization capabilities and scalabilities, which can be integrated with other training-free strategies.

Limitations

Existing multilingual benchmarks often rely on machine-translated text that introduces errors or includes expressions that are uncommon for native speakers. Due to the limitations of benchmarks, the cultural linguistic phenomena of native languages are uncertain. Thus, the impact of extracting language-specific knowledge may not be clearly reflected in existing benchmarks. The development of reasoning datasets for language-specific knowledge is urgent. Moreover, in the first step, we utilize the Language Thinking Transformation to transfer the thinking pathway from the low-resource language to a high-resource language. Generally speaking, English is the language chosen that performs best for various LLMs. However, some existing LLMs perform more prominently in other languages, which are trained with other languages as the core. Therefore, selecting the target language for thinking transformation remains an urgent issue that needs to be addressed in the future.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *Preprint*, arXiv:2305.14857.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2025. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23550–23558.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Kewei Cheng, Nesreen K. Ahmed, Theodore L. Willke, and Yizhou Sun. 2024. Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9407–9430, Miami, Florida, USA. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in english? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind : A survey of multilingual reasoning in language models. *Preprint*, arXiv:2502.09457.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, 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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficiently boosting llm reasoning in non-english languages. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yixin Ji, Juntao Li, Hai Ye, Kaixin Wu, Jia Xu, Linjian Mo, and Min Zhang. 2025. Test-time computing: from system-1 thinking to system-2 thinking. *CoRR*, abs/2501.02497.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hamin Koo and Jaehyung Kim. 2025. Extracting and emulsifying cultural explanation to improve multilingual capability of llms. *arXiv preprint arXiv:2503.05846*.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024. Improving in-context learning of multilingual generative language models with cross-lingual alignment. 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 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *Preprint*, arXiv:2403.10258.

Hongyuan Lu, Zixuan Li, and Wai Lam. 2024. Dictionary insertion prompting for multilingual reasoning on multilingual large language models. *arXiv preprint arXiv:2411.01141*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. 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 multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Junjie Wang, Dan Yang, Binbin Hu, Yue Shen, Wen Zhang, and Jinjie Gu. 2024. Know your needs better: Towards structured understanding of marketer demands with analogical reasoning augmented llms. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5860–5871.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*,

volume 35, pages 24824–24837. Curran Associates, Inc.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. 2024. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7342–7371, Bangkok, Thailand. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024a. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

Wenhao Zhu, Sizhe Liu, Shujian Huang, Shuaijie She, Chris Wendler, and Jiajun Chen. 2024c. Multilingual contrastive decoding via language-agnostic layers skipping. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 8775–8782. Association for Computational Linguistics.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *Preprint*, arXiv:2308.04948.

A Expermental Details

A.1 Dataset Details

MGSM (Multilingual Grade School Math). MGSM (Shi et al., 2023) is a benchmark of multilingual elementary school math reasoning problems. The dataset is translated from the GSM8K dataset and contains 11 different languages, which aims to evaluate the ability of models to solve math problems in a multilingual environment.

MSVAMP (Multilingual Semantic Value Math Problems). MSVAMP (Chen et al., 2024) is a math problem dataset focusing on multilingual semantic reasoning, designed to evaluate the mathematical reasoning and semantic understanding ability of models in different languages. The dataset contains math problems in multiple languages, emphasizing the understanding of quantity, units, and measurement words.

XCOPA (Cross-lingual Choice of Plausible Alternatives). XCOPA (Ponti et al., 2020) is a benchmark for multilingual commonsense reasoning tasks. The questions involve reasoning scenarios in multiple cultural backgrounds and support more than ten languages, including English, Arabic, Chinese, Spanish, French, German, Russian, etc. The benchmark aims to test cross-language reasoning capabilities and the adaptability of models to different cultural backgrounds.

A.2 Baselines

We compare our method with various representative baselines in multilingual reasoning. A branch of baselines is the training-free methods, listed as follows:

- Direct: Only the most basic prompt strategy (such as "Let's solve the following problem") is used without any additional prompt strategy.
- Few-Shot: We use three examples along with instructions as input to demonstrate the problem-solving steps to the LLMs.
- CoT (Wei et al., 2022): The model is instructed to reason in English using the phrase "Let's think step by step in English.
- DoLa (Chuang et al., 2024): DOLA contrasts logits between early and later layers to emphasize factual knowledge from higher layers, reducing hallucinations and improving the truthfulness of the generated output.
- SL-D (Zhu et al., 2024c): By skipping language-agnostic lower layers and contrasting early exit outputs with final outputs, the model leverages more accurate amateur logits to enhance multilingual reasoning.
- DIP (Lu et al., 2024): DIP inserts dictionary-based English counterparts into non-English queries, enabling LLMs to better translate and reason in English, all while remaining computationally lightweight.
- CLP (Qin et al., 2023): CLP aligns representations between source and target languages and uses task-specific solver prompting to guide reasoning.
- EMCEI (Koo and Kim, 2025): EMCEI extractes relevant cultural context from LLMs and combines it with reasoning steps to select the most contextually appropriate response using an LLM-as-Judge mechanism.

Another branch of baselines is the post-training methods, listed as follows:

- xCoT (Chai et al., 2025): XCOT fine-tunes LLMs on multilingual instruction data by using cross-lingual CoT reasoning.
- QAlign (Zhu et al., 2024a): QAlign fine-tunes models to translate reasoning questions into English for better alignment with English instruction data, followed by response alignment to enhance performance in non-English languages.
- MindMerger (Huang et al., 2024): MindMerger embeds external language understanding capabilities into LLMs and training them to collaboratively use both internal and external knowledge through a two-step training process.
- MAPO (She et al., 2024): MAPO aligns non-dominant language reasoning with a dominant language through preference estimation and optimization techniques like Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO).

B Case Study

In this section, we showcase additional examples of responses generated by **`gpt-3.5-turbo`**, utilizing both CoT and SoT method for each task.

B.1 Examples in MSVAMP

Question: “瑞秋有52棵苹果树。她从其中一棵树上摘了2个苹果。现在这棵树上还剩下7个苹果。这棵树最初有多少个苹果？”
CoT Answer:
Step 1: Calculate the total number of apples picked from all trees: 52 trees * 2 apples = 104 apples
Step 2: Calculate the total number of apples left on all trees: 52 trees * 7 apples = 364 apples
Step 3: Calculate the total number of apples on the tree before picking: Total - Picked - Left = Initial
Initial = 364 - 104 = 260
Answer: 初始苹果数为260个。
SoT Answer:
1. Problem-solving process using English thinking:
Determine how many apples the tree originally had.
Subtract the apples Rachel picked from the total apples on the tree.
2. Extract relationships between numbers in order:
Total number of apple trees: 52
Apples picked from one tree: 2
Apples left on the same tree: 7
3. Language-specific knowledge:
No specific language-specific knowledge is needed here.
4. Calculation: If the tree has 7 apples remaining after Rachel picked 2 apples, the tree must have originally had: 7 (remaining apples) + 2 (apples picked) = 9(original number of apples).
Final Answer: 这棵树最初有9 个苹果。

Question: “面包师傅做了61个糕点和167个蛋糕。如果他卖掉了108个蛋糕和44个糕点，面包师傅还剩下多少个蛋糕？”
CoT Answer:
Step 1: Calculate the total number of pastries and cakes baker made originally:
61 (pastries) + 167 (cakes) = 228
Step 2: Calculate the total number of pastries and cakes baker sold:
108 (cakes) + 44 (pastries) = 152
Step 3: Calculate the total number of pastries and cakes baker would still have:
228 (total) - 152 (sold) = 76
Answer: 面包师傅还会有76个蛋糕。
SoT Answer:
1. Problem-solving process using English thinking:
Determine how many cakes the baker has left after selling some.
Subtract the cakes he sold from the total number of cakes he originally made.
2. Extract relationships between numbers:
Number of pastries made: 61
Number of cakes made: 167
Number of cakes sold: 108
Number of pastries sold: 44
3. Language-specific knowledge:
“糕点” refers to pastries, and “蛋糕” refers to cakes.
4. Calculation:
The baker made 167 cakes and sold 108 of them.
The number of cakes the baker still has is:
167(original cakes) 108 (cakes sold) = 59 (cakes remaining).
Final Answer: The baker still has 59 cakes.

B.2 Examples in XCOPA

Question:
premise: “我讲话停了下来了。”
choice 0: “我失声了。”
choice 1: “我气喘吁吁。”
question: “cause”,
CoT Answer:
To determine which choice is more relevant and aligned with the premise and question, let’s analyze both options:
Choice 0: 我失声了。(I lost my voice.)
- This choice suggests that the speaker stopped talking because they lost their voice, which is a plausible cause for the situation described in the premise.
Choice 1: 我气喘吁吁。(I am panting.)
- This choice suggests that the speaker stopped talking because they are out of breath, which may not be directly related to the premise of the speaker’s speech stopping.
Based on the analysis, Choice 0: 我失声了。(I lost my voice.) is more relevant and aligned with the premise and question. Therefore, the final answer is “0” for Choice 0.
SoT Answer:
1. Premise, question, and choices in English:
Premise: “I stopped speaking.”
Question: What caused it?
Choice 0: “I lost my voice.”
Choice 1: “I was out of breath.”
2. Extract relationships from the premise:
Subject: “I”
Action: “stopped speaking”
This suggests a sudden or involuntary interruption in speech.
3. Language-specific analysis of choices:
In Chinese, “我失声了” means the speaker can no longer speak or has become mute or hoarse.
4. Comparison:
Choice 0 : This suggests a more sudden, possibly longer-term vocal issue, like losing the ability to speak completely.
Choice 1 : It fits better because it indicates a temporary, immediate physical condition that would logically cause someone to pause speaking, especially if they were: out of breath from exercise or running.
Final Answer: The more plausible cause of stopping speaking is out of breath. So the final answer is: 1.

C Supplementary Results

In this study, we investigate the effectiveness of SoT on three multilingual reasoning benchmarks using four different LLMs. Moreover, we also integrate other training-free methods into SoT. All results demonstrate that SoT achieves good performance for multilingual reasoning with a strong generalization ability, as shown in Table 4, Table 5 and Table 6.

Methods	Language											
	En	Sw	Ja	Be	Th	Te	Ru	Zh	De	Es	Fr	Avg.
(**training-free**)	(gpt-3.5-turbo)											
Direct	54.4	25.6	36.8	35.2	24.8	24.8	46.8	42.4	39.2	42.0	38.0	37.3
DoLa	72.2	57.2	58.0	43.4	52.2	17.6	62.4	62.2	60.0	68.2	58.8	55.7
DIP	70.2	55.2	58.8	54.6	51.6	19.4	65.7	62.8	61.4	69.8	59.2	57.2
CLP	73.2	55.8	59.4	56.0	53.6	28.0	66.2	64.6	64.8	71.4	60.0	59.4
EMCEI	73.0	59.2	60.2	55.8	54.2	26.8	63.4	63.0	62.8	70.4	59.8	59.0
SoT (*Ours*)	74.4	62.0	**65.2**	61.2	56.0	34.0	67.6	67.2	66.8	72.8	63.2	62.8
+3-shot	74.0	**66.4**	63.6	**63.6**	**60.4**	36.0	**70.4**	**69.2**	**70.4**	74.0	**65.6**	**64.9**
+CoT	**75.2**	64.8	**65.2**	57.6	55.6	**38.4**	69.2	65.6	68.0	**76.4**	64.0	63.6
(**training-free**)	(Qwen2.5-32B-Instruct)											
Direct	87.2	53.6	84.2	85.6	82.4	82.4	86.8	81.2	81.2	75.6	57.2	77.9
DoLa	85.0	45.2	71.6	80.8	69.1	62.0	77.2	82.6	76.2	73.6	53.2	70.6
SL-D	85.8	57.2	82.4	83.4	78.3	80.2	86.4	83.0	82.8	76.6	62.0	78.0
DIP	85.8	52.6	82.0	81.2	75.2	74.2	83.0	83.0	80.4	78.2	54.2	75.4
CLP	86.0	53.6	81.4	84.2	78.2	77.2	84.2	83.8	81.4	77.2	58.2	76.9
EMCEI	85.4	52.8	81.0	83.8	83.0	78.0	84.0	82.4	81.6	78.0	62.6	77.5
SoT (*Ours*)	87.2	**67.2**	86.0	86.0	85.4	87.4	88.8	**84.4**	86.4	**78.4**	64.8	82.0
+3-shot	87.2	66.8	86.2	**87.6**	86.4	88.0	88.4	82.4	86.4	78.0	66.8	82.2
+CoT	**87.8**	63.6	**87.2**	86.8	**86.8**	**88.8**	**89.2**	84.0	**86.8**	77.6	**68.8**	**82.5**
(**post-training**)												
xCoT	86.6	58.4	83.2	82.4	81.4	27.4	80.8	87.2	81.0	81.2	82.0	75.6
QAlign	86.4	58.0	80.0	81.2	84.0	29.2	85.6	86.2	81.6	82.0	81.6	76.0
MindMerger	87.0	69.2	83.0	84.8	85.6	38.0	88.0	88.0	82.5	82.4	82.8	79.2
MAPO	87.0	61.6	83.2	86.0	83.0	35.2	86.2	89.8	83.0	83.8	83.2	78.4

Table 4: Supplementary results (%) of mathematical reasoning on MGSM using **gpt-3.5-turbo** and **Qwen2.5-32B-Instruct**.

Methods	Language										
	En	Sw	Ja	Be	Th	Ru	Zh	De	Es	Fr	Avg.
(**training-free**)	(gpt-3.5-turbo)										
Direct	77.0	68.1	68.4	48.7	61.8	74.3	68.0	73.4	73.3	73.4	68.6
DoLa	76.4	61.2	62.4	49.0	61.2	68.7	69.4	68.6	70.1	69.5	65.7
DIP	70.0	68.4	69.8	50.5	64.0	69.4	75.2	75.8	73.0	74.5	69.1
CLP	78.8	68.7	70.4	52.2	68.2	72.0	76.6	74.6	76.5	77.1	71.5
EMCEI	73.8	69.0	70.8	52.0	66.5	70.6	74.2	73.6	74.3	76.3	70.1
SoT (*Ours*)	81.8	75.4	80.2	63.6	72.8	**79.2**	80.4	80.0	**83.0**	80.4	77.7
+3-shot	82.0	76.4	79.6	**66.0**	74.2	**79.2**	80.6	80.6	81.8	78.2	77.9
+CoT	**82.4**	**76.6**	**81.0**	64.2	**74.4**	78.4	**81.4**	**81.6**	81.4	**81.0**	**78.2**
(**training-free**)	(Qwen2.5-32B-Instruct)										
Direct	89.8	39.4	69.2	55.0	65.4	74.6	81.8	77.8	83.2	82.2	71.8
DoLa	85.2	48.3	76.1	68.2	71.9	87.3	83.9	75.7	81.3	78.4	75.6
SL-D	88.2	54.5	79.4	81.4	83.8	88.5	86.7	81.2	85.2	82.2	81.1
DIP	88.2	52.2	82.2	77.2	72.3	88.1	87.3	82.7	83.6	85.8	80.0
CLP	90.8	53.0	82.6	73.1	78.2	88.2	86.8	83.2	83.3	89.1	80.8
EMCEI	91.2	58.3	83.0	76.3	76.3	85.5	88.3	85.6	84.9	86.4	81.6
SoT (*Ours*)	**93.8**	**87.4**	89.8	**84.8**	87.0	**90.8**	91.8	91.8	92.6	93.2	90.3
+3-shot	93.7	86.4	91.0	83.1	**87.6**	90.2	**93.6**	**93.4**	93.8	**93.8**	**90.7**
+CoT	94.2	87.0	**92.0**	83.6	86.4	89.8	91.4	**93.4**	**94.4**	93.6	90.6
(**post-training**)											
xCoT	90.3	75.2	81.5	74.9	75.4	85.0	85.5	82.8	85.3	89.0	82.5
QAlign	90.7	72.8	85.5	75.3	77.5	88.0	83.8	87.2	89.6	89.4	84.0
MindMerger	91.5	77.0	85.8	78.5	78.2	87.1	86.8	88.5	90.2	91.3	85.5
MAPO	91.9	71.1	86.0	74.0	79.1	82.5	86.3	85.6	88.4	89.4	83.4

Table 5: Supplementary results (%) of mathematical reasoning on MSVAMP using **gpt-3.5-turbo** and **Qwen2.5-32B-Instruct**.

Methods	Language											
	Et	Ht	Id	It	Qu	Sw	Ta	Th	Tr	Vi	Zh	Avg.
(**training-free**)	(DeepSeek-R1-7B)											
Direct	19.6	20.6	15.8	11.0	19.6	16.0	16.0	11.0	12.2	12.8	10.6	15.0
DoLa	23.4	34.0	48.4	29.6	30.2	35.2	32.0	43.8	44.6	49.4	67.0	39.8
SL-D	35.8	40.4	51.8	43.2	34.4	39.6	45.2	51.0	47.8	50.8	69.2	46.3
DIP	33.8	40.2	51.4	50.4	41.2	47.2	47.2	43.9	45.8	50.0	66.0	47.0
CLP	30.0	43.6	54.8	48.4	42.0	43.0	49.4	51.2	52.0	50.8	71.4	48.8
EMCEI	32.6	44.6	56.0	51.2	41.6	42.6	43.0	52.6	52.4	52.0	71.2	49.1
SoT (*Ours*)	**51.2**	51.0	**66.8**	67.4	50.6	50.0	52.0	61.4	**58.8**	**61.2**	76.8	58.8
+3-shot	46.2	**52.6**	63.2	**68.4**	49.6	**52.0**	52.0	54.8	57.0	55.2	76.8	57.1
+CoT	49.8	51.4	65.4	68.2	**53.4**	49.8	**54.6**	**62.2**	56.4	**61.2**	**78.4**	**59.2**
(**post-training**)												
xCoT	44.6	32.4	55.8	57.0	21.2	31.4	29.8	59.2	33.8	54.2	65.4	44.1
QAlign	43.0	31.2	53.4	53.4	22.2	28.4	33.8	59.8	21.4	49.6	71.0	42.5
MindMerger	41.8	32.6	53.2	56.8	21.4	32.0	32.4	51.4	33.6	54.8	65.8	43.3
MAPO	41.2	35.0	51.4	54.8	22.0	29.8	35.0	52.0	35.0	51.2	61.2	42.6
(**training-free**)	(Qwen2.5-7B-Instruct)											
Direct	8.6	15.0	5.0	9.8	17.0	4.4	14.8	8.4	3.6	8.0	8.8	9.4
DoLa	53.8	44.6	75.4	65.8	30.6	41.8	52.4	63.0	74.0	74.6	75.4	59.2
SL-D	61.4	52.4	73.8	74.4	35.0	43.4	52.2	63.8	73.4	72.4	74.4	61.5
DIP	61.4	59.0	81.2	83.0	46.8	47.2	58.6	73.2	76.8	74.4	74.2	66.9
CLP	64.2	54.8	75.6	70.0	41.0	41.6	53.0	64.6	65.2	75.6	70.2	61.4
EMCEI	61.6	51.8	72.6	74.2	44.8	45.2	50.2	70.8	71.4	71.8	72.8	62.5
SoT (*Ours*)	**65.0**	58.2	82.6	83.8	**49.6**	50.8	60.8	73.4	78.6	83.2	81.0	69.7
+3-shot	63.2	59.0	**83.4**	84.2	**49.6**	50.4	58.4	**78.8**	**80.8**	80.8	**86.0**	**70.4**
+CoT	64.0	**59.6**	81.2	**86.4**	47.2	**51.8**	**61.2**	75.4	79.4	**83.8**	83.6	70.3
(**post-training**)												
xCoT	47.2	48.0	65.2	68.4	21.0	45.2	32.2	69.2	33.6	64.4	65.0	50.9
QAlign	45.8	42.0	62.2	60.2	22.8	35.2	45.6	69.6	21.4	58.2	73.4	48.8
MindMerger	44.2	33.2	57.2	67.0	21.2	42.4	38.6	57.6	34.4	64.6	65.0	47.8
MAPO	40.2	38.0	61.2	62.0	22.4	42.8	43.4	58.4	35.8	62.0	60.8	47.9
(**training-free**)	(gpt-3.5-turbo)											
Direct	48.2	49.6	33.8	36.8	50.2	47.0	37.8	46.0	43.4	44.8	37.0	43.1
DoLa	73.2	53.6	74.6	78.6	39.0	52.4	50.4	64.6	72.0	72.2	78.8	64.5
DIP	75.4	60.8	81.2	81.4	43.6	54.8	62.6	70.0	79.2	73.8	77.8	69.1
CLP	70.2	58.0	73.2	81.2	40.0	56.0	52.6	64.0	73.2	67.0	72.2	64.3
EMCEI	78.6	62.0	83.0	84.6	44.2	57.6	62.4	72.8	82.8	75.4	79.8	71.2
SoT (*Ours*)	82.0	**66.4**	84.0	**88.2**	49.0	74.4	58.2	77.0	**84.2**	81.4	84.6	75.4
+3-shot	79.2	60.4	76.2	67.2	**55.0**	73.6	**63.8**	69.4	72.8	64.8	78.2	69.1
+CoT	**83.4**	66.2	**86.0**	87.6	54.0	**76.0**	63.4	**78.4**	82.8	**82.0**	**87.6**	**77.0**
(**training-free**)	(Qwen2.5-32B-Instruct)											
Direct	19.2	27.4	16.6	18.6	23.2	23.2	21.2	19.4	9.4	25.0	5.4	19.0
DoLa	78.4	66.2	75.4	84.6	40.4	54.8	67.0	82.8	84.8	81.6	83.4	72.7
SL-D	61.4	52.4	73.8	74.4	35.0	43.4	52.2	63.8	73.4	72.4	74.4	61.5
DIP	71.2	68.1	83.2	89.4	49.4	55.6	68.2	87.0	89.6	88.2	89.6	76.3
CLP	75.4	70.2	81.2	92.8	43.8	55.4	70.8	81.8	85.2	83.8	84.4	75.0
EMCEI	78.6	62.0	83.0	84.6	44.4	57.6	62.4	82.8	82.8	91.2	85.8	74.1
SoT (*Ours*)	**87.0**	74.8	96.2	97.2	**53.8**	65.4	78.4	91.0	95.0	95.8	**96.0**	84.6
+3-shot MindMerger	86.6	**78.2**	**96.8**	**97.8**	52.2	**69.8**	**79.8**	88.6	**95.6**	**96.4**	95.4	**85.2**
+CoT	86.4	76.0	96.6	96.8	53.6	67.2	78.8	**91.6**	95.4	95.8	**96.0**	84.9
(**post-training**)												
xCoT	67.6	68.2	84.2	86.4	24.2	55.2	52.2	73.4	73.4	77.6	80.2	67.5
QAlign	65.2	62.6	82.0	80.0	24.8	55.2	55.6	77.6	52.8	80.0	84.2	65.5
MindMerger	70.6	51.8	81.2	83.4	25.8	51.0	42.4	73.4	54.4	75.2	80.4	62.7
MAPO	67.2	68.8	82.6	87.2	24.0	59.0	57.2	74.8	59.2	72.6	81.0	66.7

Table 6: Supplementary results (%) of commonsense reasoning on XCOPA using various LLMs, including **DeepSeek-R1-7B**, **Qwen2.5-32B-Instruct**, **gpt-3.5-turbo** and **Qwen2.5-32B-Instruct**.