Knowledge Funnel: Enhancing Multilingual Reasoning in LLMs via Structured Knowledge Representation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have poorer performance on multilingual reasoning tasks than on English tasks due to limited pretraining data for these languages. In this paper, we propose Knowledge Funnel, a novel multilingual reasoning framework that improves LLM performance through four steps: (1) Multilingual Knowledge Alignment, which enhances reasoning by leveraging English knowledge; (2) Entity-Structured Knowledge, which extracts a structured representation of the question (3) Dependency Knowledge, which captures language-specific dependencies such as units and quantifiers; (4) Calculation and Answer Generation, which ensures accurate reasoning results. Furthermore, it can be combined with other approaches, such as CoT, to achieve even better results. Our framework achieves 11.3% and 11.1% improvements over Chain-of-Thought (CoT) methods on MGSM8K and MSVAMP, demonstrating its effectiveness in enhancing LLMs' multilingual reasoning capabilities. We will release our code once acceptance.

1 Introduction

Large language models (LLMs) have strong reasoning capabilities across various reasoning tasks, whether it is numerical reasoning, commonsense reasoning or symbolic reasoning (Wei et al., 2022). But when it comes to multilingual sences, the reasoning capabilitiy of LLMs vary between different languages (Huang et al., 2023a; Shi et al., 2023). For example, LLMs tend to be more accurate when solving problems in English compared to the same questions in other languages. We analyze that it can be attributed to the training process of LLMs. Since English serves as the primary language in training data, models tend to perform well in English but struggle with other languages due to limited training resources (She et al., 2024). There are three key challenges:

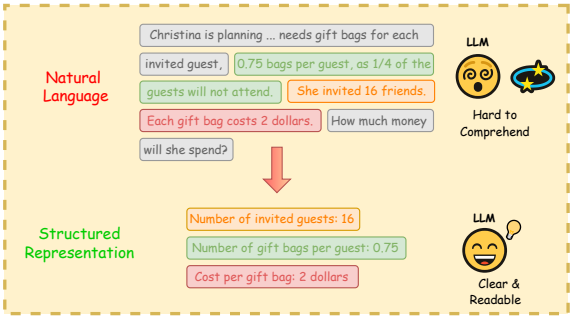

Figure 1: Advantages of structured knowledge representation: Improving the semantic understanding ability of LLMs.

1. **Limited cross-linguistic generalization:** LLMs often fail to recognize problem representations in languages with fewer training examples.

2. **Difficulty in understanding semantic structures:** Complex semantic structures in non-English languages can lead to misinterpretation of relationships between entities, which affects reasoning accuracy.

3. **Over-reliance on English alignment:** Many existing methods improve multilingual reasoning by aligning non-English problems with English representations. However, native-language training data contain valuable linguistic features that should not be overlooked.

These challenges highlight the need for an approach that enhances multilingual reasoning capabilities of LLMs.

Existing methods for improving multilingual reasoning in LLMs can be categorized into four main approaches: **Direct translation:** (Huang et al., 2023b; Qin et al., 2023) Translating multilingual questions into English helps align them with the model's strengths. However, this approach relies

on high-quality translation and may lose structured information due to linguistic differences. **Chain-of-Thoughts (COT):** (Qin et al., 2023) CoT improves reasoning by optimizing reasoning steps, but its effectiveness also depends on the language comprehension of the model, which will cause the continuous transmission of misunderstandings. **In-Context Learning:** (Brown et al., 2020) This method captures relationships within given examples but struggles to identify key entities and relationships beyond the provided context (Min et al., 2022). **Supervised Fine-tuning (SFT):** (Zhu et al., 2024; She et al., 2024) Although SFT improves performance in specific domains, it has two main limitations: (i) it requires large-scale labeled data, which makes it costly, and (ii) fine-tuning on one specific language can not generalize well across all languages.

Existing methods primarily focus on optimizing the reasoning process but overlook the challenges LLMs face in understanding the semantic structures of multilingual problems. To this end, we propose **Knowledge Funnel**, which leverages structured knowledge representation to enhance LLMs' semantic understanding of multilingual problems. By integrating structured representations into the reasoning process, our approach aims to improve LLMs' multilingual reasoning capabilities. As shown in Figure 1, LLMs face greater challenges in understanding questions formulated in natural language compared to those presented in a structured format. To address this, we transform natural language questions into structured knowledge representations, which allows the model to recognize that "0.75 bags per guest" and "1/4 of the guests will not attend" describe the same underlying relationship, meanwhile filtering out irrelevant information that could interfere with reasoning. Our framework has the following three highlights:

(1) We leverage a simple yet effective alignment strategy to transfer LLMs' reasoning capabilities in English to other languages. By using English as an intermediary, we transform non-English problems into English representations, thereby enhancing the reasoning ability in low-resource languages.

(2) Since LLMs struggle with understanding the semantic structures of non-English problems, we extract structured knowledge representations from natural language questions, converting complex multilingual problems into a more interpretable form.

(3) While LLMs perform poorly in certain languages, language-specific features remain essential. Our framework preserves these features by guiding the model to focus on language-specific dependency knowledge, ensuring that the valuable linguistic characteristics learned during training could be retained.

Extensive experiments demonstrate that our framework outperforms existing methods, including Chain-of-Thought (CoT), across various multilingual reasoning tasks. The result illustrate that by improving LLMs' ability to interpret the semantic structure of multilingual problems while preserving language-specific knowledge, Knowledge Funnel significantly enhances multilingual reasoning performance.

Our contributions can be summarized as follows:

1. We proposed a framework called Knowledge Funnel, which dynamically extracts entity structured knowledge and language-specific dependency knowledge from the questions. It can significantly improve the multi-language reasoning ability of LLMs at a very low cost.

2. To reflect the scalability of our framework, we combined our framework with COT and other methods, and verified it on multiple LLMs and datasets. Taking GPT-3.5 and MGSM datasets as examples, the average score of this framework in all languages has increased by 30.3% compared to the original method, 11.3% higher than the COT method, and better than other baselines.

3. We further extend the framework to multiple reasoning tasks, and the scores in each language are better than other methods such as COT, indicating that our framework has strong generalization.

2 Related Work

This work is closely related to two topics: multilingual reasoning and prompt learning.

2.1 Multilingual Reasoning

Large language model reasoning, which evaluates the ability of LLMs to handle complex tasks, serves as a straightforward measure of their efficiency (She et al., 2024) . These reasoning tasks mainly include numerical reasoning and commonsense reasoning. With the growing interest in multilingual

LLM performance, researchers have begun investigating how LLMs perform in multilingual environments. A common approach to improve performance on low-resource languages is pre-translation inference, which involves translating input questions into a high-resource pivot language (e.g., English or Chinese) before querying the LLM to leverage the model's proficiency in the pivot language (Huang et al., 2023b; Qin et al., 2023).

In addition, Chain-of-Thoughts (CoT) prompting has proven effective in enhancing complex reasoning performance (Sap et al., 2020; Yu et al., 2023; Liu et al., 2023a) and has been widely explored in existing studies (Huang et al., 2025). Liu et al. (2024) proposed several strategies to extend COT to multilingual scenarios, including "Native-CoT" where both questions and instructions are in the native language, "EN-COT" where instructions are in English, and "XLT" (Huang et al., 2023b), which involves translating questions into English and solving them step-by-step. In addition to non-parametric methods, some works introduce supervised fine-tuning (SFT) (She et al., 2024; Zhu et al., 2024) to enhance the multilingual reasoning ability of the model. For example, translating English training data into other languages, or mixing the original language and target language in a single query , and then fine-tuning the multilingual large language model (MLLM) for instructions (Chai et al., 2024). However, SFT suffers from data scarcity and catastrophic forgetting problems, and its cross-domain generalization ability is also lacking (She et al., 2024).

Compared with the above methods, our framework is more suitable for multilingual reasoning: we pay more attention to the specific knowledge of different languages, and it is more generalizable than SFT because it can be applied to different tasks and different languages at low cost.

2.2 Prompt Learning

Prompt learning is a mainstream research method to improve the capabilities of LLMs. By designing a variety of prompt templates, LLMs are guided to reason in a non-parametric way (Liu et al., 2023b). Prompt learning can improve model capabilities without changing parameters and does not rely on a large amount of labeled data, significantly reducing the cost of model training and has stronger generalization capabilities. In the field of multilingual reasoning, Chain-of-Thoughts (COT) is an efficient and simple method. Common COT methods include basic CoT prompting (Wei et al., 2022), complex CoT (Fu et al., 2023) and multilingual CoT (Shi et al., 2023), etc. In addition, Brown et al. (2020) proposed in-context learning (ICL), which generates prompts by combining some examples with instructions, and Puerto et al. (2024) used LLMs to convert reasoning tasks into code and execute them with the help of external interpreters to solve complex reasoning problems. In terms of structured prompts, Madaan et al. (2022) performed few-shot prompts on Codex LLM and converted the task into a Python graph for processing structured commonsense tasks, further expanding the application scope of prompt learning.

As a prompt learning method, unlike COT and other methods that focus on inference steps, our framework focuses on improving the relationship understanding ability of the model, using a more concise method to structure the questions, and can be applied to various multi-language reasoning tasks. In addition, our framework can also be combined with other prompt learning methods to improve reasoning effects in all aspects.

3 Methodology

In this section, we propose a novel prompting framework, named Knowledge Funnel, aiming to improve the performance of LLMs in multilingual reasoning tasks. The overall framework of our Knowledge Funnel is shown in Figure2.

3.1 Multilingual Alignment Knowledge

In the first step, all questions are translated into English by LLMs. This step achieves language alignment in a simple but effective way, using the translation capabilities of LLM to convert questions in all other languages into English. This step leverages LLMs' strong English reasoning capabilities to mitigate performance degradation caused by limited training data in non-English languages, thereby improving the overall performance of the model in multilingual reasoning tasks. By using English as an intermediary, we transform non-English problems into English representations, so the model can understand the question more accurately, ensuring semantic consistency and logic during the reasoning process.

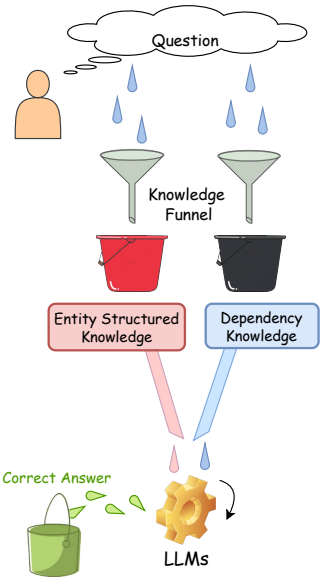

Figure 2: Illustration of our Knowledge Funnel Framework. Our Knowledge Funnel is designed to extract Entity Structured Knowledge and Dependency Knowledge from multilingual questions, optimizing large models' multilingual reasoning capabilities.

3.2 Entity-Structured Knowledge

Next, named entity recognition (NER) is applied to extract key entities (e.g., numbers, units, objects) and their relationships in the question. These relationships are identified sequentially based on their textual order, ensuring that critical information is accurately captured and clearly structured. This step not only clarifies the connections between entities, but also ensures that the semantic information can be accurately captured and effectively associated with the corresponding entities. In this way, the model can clearly identify the semantic structures in the question and avoid inference errors caused by misinterpretation of relationships. In addition, relationship extraction helps simplify problems and makes complex reasoning tasks more parsable. This step helps eliminate ambiguities, reduce reasoning errors, and enhance the model's ability to parse complex problems.

3.3 Dependency Knowledge

To preserve language-specific features, the model is required to identify the dependencies between values and units, quantifiers, and measurement words by utilizing the language knowledge learned by model during the training process. Unlike task-specific prompting, this approach enables LLMs to automatically handle linguistic dependencies across languages, ensuring accurate semantic interpretation and reducing inference errors caused by syntactic variations. The recognition of language-specific dependencies enhances the model's ability to process language characteristics and reduces inference errors caused by differences in language syntax.

3.4 Calculation And Answer Generation

Finally, the model performs calculations based on the extracted relationships and dependency knowledge. The computed answer is then translated back into the original language, ensuring consistency between input and output while maintaining interpretability in multilingual settings. At this stage, the final solution to the problem is reached through the operation of relationships, while the translation step ensures the seamless connection between different languages, allowing the reasoning task to be successfully completed in a multilingual environment.

3.5 Analysis

Our framework offers several advantages:

- **Multilingual Alignment Knowledge** enhances reasoning performance in non-English languages by leveraging English as an intermediary.
- **Entity-Structured Knowledge** ensures clarity in question semantics and accurate information extraction.
- **Dependency Knowledge Extraction** enables the model to recognize and retain language-specific features, preserving semantic consistency and improving reasoning accuracy.

Additionally, the method is highly scalable. It is compatible with existing reasoning approaches, such as Chain-of-Thought (CoT) and In-Context Learning (TCL), and can be integrated with them for further performance improvements. Unlike fine-tuning, which requires large amounts of labeled data for specific tasks, our framework provides a generalizable solution across diverse multilingual reasoning tasks. Experimental results demonstrate its effectiveness across different datasets, reasoning types (e.g., numerical and

commonsense reasoning), and linguistic domains, highlighting its broad applicability and robustness.

4 Experimental Setup

In this section, we introduce the experimental settings, including base models, baselines, evaluation indicators, experimental settings, etc.

4.1 Base Model

In order to evaluate the effectiveness of our framework in improving multilingual reasoning, we use three LLMs as base models: GPT-3.5-Turbo, Qwen-7B-Instruct, and Mistral-7B-Instruct-v0.3. We not only conduct experiments on open source model(Qwen) and closed source model (GPT-3.5), but also add the Mistral model, an open source LLM that focuses on improving reasoning capabilities and is not specifically optimized for multilingual tasks. Experimenting on this model can verify whether the framework can improve the multilingual capabilities of models that are not good at multilingual capabilities.

4.2 Datasets

To verify the versatility of our framework, we conducted experiments on two multilingual numerical reasoning datasets, MGSM and MSVAMP, and a commonsense reasoning dataset, XCOPA.

MGSM (Multilingual Grade School Math): MGSM (Shi et al., 2023) is a benchmark dataset 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 dataset 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. It aims to test cross-language reasoning capabilities and the adaptability of models to different cultural backgrounds.

4.3 Baselines

For comparison, we selected some non-parameterized methods and experimented on the same model and dataset.

Basic Prompt: Only the most basic prompt strategy (such as "Let's solve the following problem") is used without any additional prompt strategy. The questions are presented in the original language and the instructions are presented in English.

Translate to English: (Trans) The questions are presented in the original language, and the large model is prompted in English to translate the problem into English, and then it is asked to answer it directly.

English chain-of-thought (EN-COT): The question is presented in the native language, but the model is instructed to reason in English using the phrase "Let's think step by step in English."

Cross-lingual-thought (XLT): XLT (Huang et al., 2023b) is an advanced prompting approach for multilingual tasks, where the model is guided to translate the question into English and solve it step-by-step in English.

Must Think More Step (MTMS): MTMS (Jin et al., 2024) is a prompting strategy that encourages the model to perform more detailed and gradual reasoning by explicitly asking it to break down the problem into smaller steps, ensuring deeper and more thorough thought processes.

4.4 Evaluation Metrics

Accuracy is used to access a model's ability on classification tasks and is commonly used for multichoice and yes/no tests: Accuracy = Ncorrect/Ntotal (Jin et al., 2024).

4.5 Experiment Setting

To verify the effectiveness of our framework, we designed the following experiments:

Ours: Our framework described in Section 3. We report two versions of our method in the experimental results: one with only three basic steps, excluding dependency knowledge (**Ours (Basic)**), and another with all steps included (**Ours (Full)**). This distinction is made because the effectiveness

Model	Method	Language											
		En	**Sw**	**Ja**	**Be**	**Th**	**Te**	**Ru**	**Zh**	**De**	**Es**	**Fr**	**AVG**
GPT	Original	54.4	25.6	36.8	35.2	24.8	24.8	46.8	42.4	39.2	42.0	38.0	37.3
	Trans	54.4	29.2	29.6	25.2	30.4	18.4	34.0	43.2	58.4	68.8	48.8	40.0
	COT	76.0	55.2	60.4	42.0	46.4	15.6	58.4	62.8	70.0	70.4	62.0	56.3
	XLT	73.6	**68.0**	65.2	60.4	**63.6**	37.6	73.2	68.8	**73.6**	70.0	**69.2**	65.7
	MTMS	73.6	58.4	55.6	41.6	46.0	16.4	63.6	63.6	70.0	71.2	66.4	56.9
	Ours (Basic)	75.2	55.6	60.4	57.6	55.6	35.2	69.2	64.4	66.4	70.0	66.4	61.5
	Ours (Full)	74.4	61.6	66.0	60.8	59.2	32.8	69.6	71.6	69.6	71.2	66.0	63.9
	Ours (Basic) + few shot	74.0	63.6	65.2	59.2	60.0	36.8	71.2	66.4	66.8	74.0	68.4	64.1
	Ours (Full) + few 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
	Ours (Basic) + COT	**78.4**	65.6	**73.2**	61.2	**63.6**	38.4	**73.6**	**74.8**	70.8	**75.2**	**69.2**	**67.6**
	Ours (Full) + COT	72.4	64.8	66.4	**64.4**	60.4	**40.8**	**73.6**	69.6	68.0	74.4	68.8	65.8
Qwen	Original	84.0	12.8	56.0	51.2	48.0	24.0	73.6	80.8	66.8	71.2	64.4	57.5
	Trans	85.2	24.8	74.0	72.0	78.8	41.6	83.6	82.0	74.0	80.0	72.0	69.8
	COT	83.6	14.0	73.2	61.6	78.4	39.2	78.8	81.6	74.0	77.6	72.0	66.7
	XLT	**87.6**	27.2	**78.4**	73.2	**80.8**	39.2	**86.8**	83.2	73.2	78.4	73.2	71.0
	MTMS	86.8	17.2	73.2	61.6	76.4	26.4	79.2	80.0	75.2	78.4	70.8	65.9
	Ours (Basic)	84.0	25.6	73.2	68.8	78.4	36.4	80.8	**83.6**	69.6	75.2	74.0	68.1
	Ours (Full)	85.2	27.2	73.2	69.6	75.6	36.4	79.6	79.2	72.0	80.8	73.2	68.4
	Ours (Basic) + few shot	83.6	25.6	74.0	68.8	77.2	39.2	81.2	80.4	76.4	78.0	**78.0**	68.8
	Ours (Full) + few shot	85.2	24.8	76.0	70.4	77.6	39.6	79.6	80.8	**79.2**	78.0	76.4	69.8
	Ours (Basic) + COT	86.8	**28.0**	75.2	**74.4**	**80.8**	**41.2**	84.0	82.0	74.0	**81.6**	74.8	**71.2**
	Ours (Full) + COT	85.6	27.6	74.4	69.2	79.2	38.0	79.2	80.0	73.2	78.4	72.0	68.8
Mistral	Original	36.8	2.4	16.4	10.4	7.6	20.4	22.0	22.4	21.6	21.6	3.2	16.8
	Trans	48.0	5.2	12.4	21.6	23.2	29.2	40.0	29.2	29.2	31.2	5.2	24.9
	COT	53.6	7.6	14.4	8.0	15.2	29.2	30.8	27.6	30.0	30.4	4.0	22.8
	XLT	**57.6**	12.4	37.2	25.6	32.0	**51.2**	**51.2**	43.2	**48.4**	**48.8**	4.0	37.4
	MTMS	55.2	6.0	14.0	4.0	15.6	30.4	39.6	30.8	30.0	33.2	2.0	23.7
	Ours (Basic)	54.4	13.6	30.4	28.4	29.2	40.0	43.2	40.0	40.0	36.4	7.6	33.0
	Ours (Full)	55.2	13.6	36.4	28.8	28.8	43.6	44.4	43.2	42.0	40.0	8.8	34.7
	Ours (Basic) + few shot	51.2	12.0	38.8	29.6	33.6	44.0	48.0	43.2	42.8	44.0	8.4	36.0
	Ours (Full) + few shot	49.6	11.6	**43.6**	30.8	**35.6**	48.8	46.0	**44.4**	45.2	45.2	7.2	37.1
	Ours (Basic) + COT	56.4	**14.0**	36.0	**32.4**	33.2	49.6	47.6	43.2	45.6	44.8	**11.2**	**37.6**
	Ours (Full) + COT	55.4	12.8	38.4	27.2	32.4	43.2	45.6	42.4	39.6	43.6	10.8	35.5

Table 1: Performance Comparison Across Models and Methods in Multilingual Tasks (MGSM Dataset). The **bolded values** represent the highest scores, while the underlined values represent the second-highest scores.

of dependency knowledge extraction varies across different experiments. Both versions are also retained in subsequent combination methods.

Ours+COT: A hybrid approach combining our framework with COT. After enhancing the semantic understanding ability of the model, it further improves the reasoning ability by guiding the model to make step-by-step reasoning.

Ours+COT+few-shot: After combining our framework with COT, a small number of examples (few-shot) are used to guide the model to help the model better understand the task and perform effective reasoning, thereby further improving multilanguage reasoning performance.

5 Results

In our experiments, we evaluated the performance of different methods on multilingual reasoning tasks. To ensure the fairness, we set up multiple baselines: the original method without any prompts **(Original)**, the direct translation method **(Trans)**, and several mainstream reasoning frameworks, including Chain-of-Thought **(COT)**, Cross-lingual-thought (**XLT**) , and Must Think More Step (**MTMS**). All baselines are introduced in Section 4.3. Additionally, we tested our proposed method and its combination with other reasoning methods, such as COT and a few-shot setting, to further enhance performance.

Our experiments were conducted on two reason-

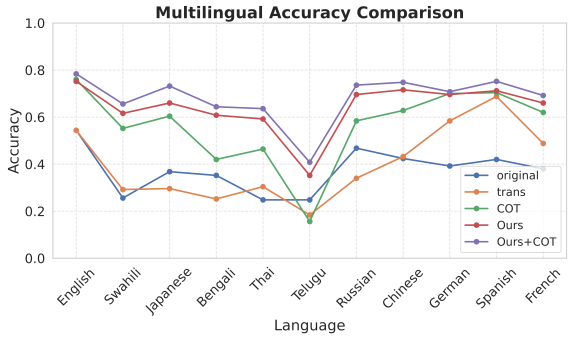

Figure 3: Comparison of Accuracy Across Methods on GPT-3.5 (MGSM Dataset). Both of our frameworks maintain high accuracy across all languages.

ing tasks: numerical reasoning and commonsense reasoning. For numerical reasoning, we used the MGSM and MSVAMP datasets, while for commonsense reasoning, we used the XCOPA dataset. The main results are presented in Tables 1 and Appendix A.

5.1 Overall Results

Table 1 shows the performance of each method on the MGSM dataset. The three comparison methods mainly focus on optimizing reasoning steps, while our Knowledge Funnel focuses on enhancing question understanding. As a result, it consistently outperforms traditional methods across all languages, with particularly significant improvements in languages where the original performance was lower. For instance, the score of Swahili increased by 29.2% and Thai increased by 30.8%, demonstrating that languages with weaker initial performance often suffer from poor relationship understanding.

From the average score, our Knowledge Funnel improves model accuracy by 7.6% over COT and 7.0% over MTMS, achieving the best results in most languages. This highlights its effectiveness in multilingual reasoning tasks. Additionally, incorporating a few-shot setting further enhances multilingual reasoning performance. Furthermore, when combined with COT , the framework achieves additional improvements across all languages, with an average accuracy increase of 11.3% over COT and 10.7% over MTMS. This effect is more pronounced in languages where the model already performs well. For example, in Chinese, the hybrid approach boosts accuracy by an additional 10.4% compared to our method alone. More importantly, the average score of our combined framework surpasses that of XLT by 1.9%, achieving the best results in most languages. As shown in Figure 3, we compared the accuracy of different methods across various languages, demonstrating that our framework leads in accuracy across all languages.

5.2 Ablation Study

To analyze the contribution of each step in our method, we conducted ablation experiments, with results presented in Table 2:

- **Using only step 1 (Multilingual Alignment):** Performing only multilingual alignment without relation analysis leads to a significant drop in accuracy, highlighting the necessity of deep relational processing.
- **Using only step 2 (Entity Structured Knowledge) or step 3 (Dependency Knowledge):** Extracting either entity structured knowledge or dependency knowledge in isolation achieves moderate performance, suggesting their complementary roles.
- **Omitting Step 1:** Performance degrades significantly in low-resource languages, emphasizing the importance of multilingual alignment.
- **Omitting Step 2 or Step 3:** Accuracy drops sharply, indicating that both structured and dependency knowledge are crucial for reasoning performance.

5.3 Case Study

In Figure 3, we present examples where the traditional Chain-of-Thought (COT) method fails, while our framework produces accurate results. These cases highlight how our approach resolves common errors in multilingual reasoning.

In Figure 3 (a) and 3 (b), language-specific ambiguities cause COT to misinterpret units and discounts. For instance, in Figure 3 (a), COT confuses "per dozen eggs" with "per egg", leading to an incorrect calculation. Similarly, in Figure 3 (b), the Chinese expression "70% off" is misread by COT as "a 70% reduction", rather than "70% of the original price". Our framework effectively resolves these issues by incorporating dependency knowledge, ensuring correct numerical interpretation.

Figure 3 (c) and 3 (d) demonstrate how structural misunderstandings are addressed. In Figure 3 (c), COT fails to parse the relationship between

Model	Method	Language											
		En	Sw	Ja	Be	Th	Te	Ru	Zh	De	Es	Fr	AVG
GPT	Original	54.4	25.6	36.8	35.2	24.8	24.8	46.8	42.4	39.2	42.0	38.0	37.3
	+ step 1	54.4	29.2	29.6	25.2	30.4	18.4	34.0	43.2	58.4	68.8	48.8	40.0
	+ step 2	61.6	52.8	57.6	41.2	44.0	24.8	62.0	62.8	57.6	62.8	58.4	53.2
	+ step 3	69.2	58.8	61.2	45.6	54.4	25.6	66.0	65.2	65.6	65.6	61.6	58.1
	Ours	74.4	61.6	66.0	60.8	59.2	32.8	67.6	71.6	69.6	69.2	66.0	63.9
	- step 1	74.8	58.8	58.4	58.4	57.6	31.2	67.6	69.2	66.0	69.2	63.2	61.6
	- step 2	72.8	61.2	62.8	56.4	56.0	30.0	67.2	68.8	67.6	70.0	60.4	61.2
	- step 3	75.2	54.8	60.4	57.6	55.6	31.2	69.2	63.2	65.2	70.0	66.4	60.8

Table 2: Ablation Study on GPT-3.5 for the MGSM Multilingual Dataset

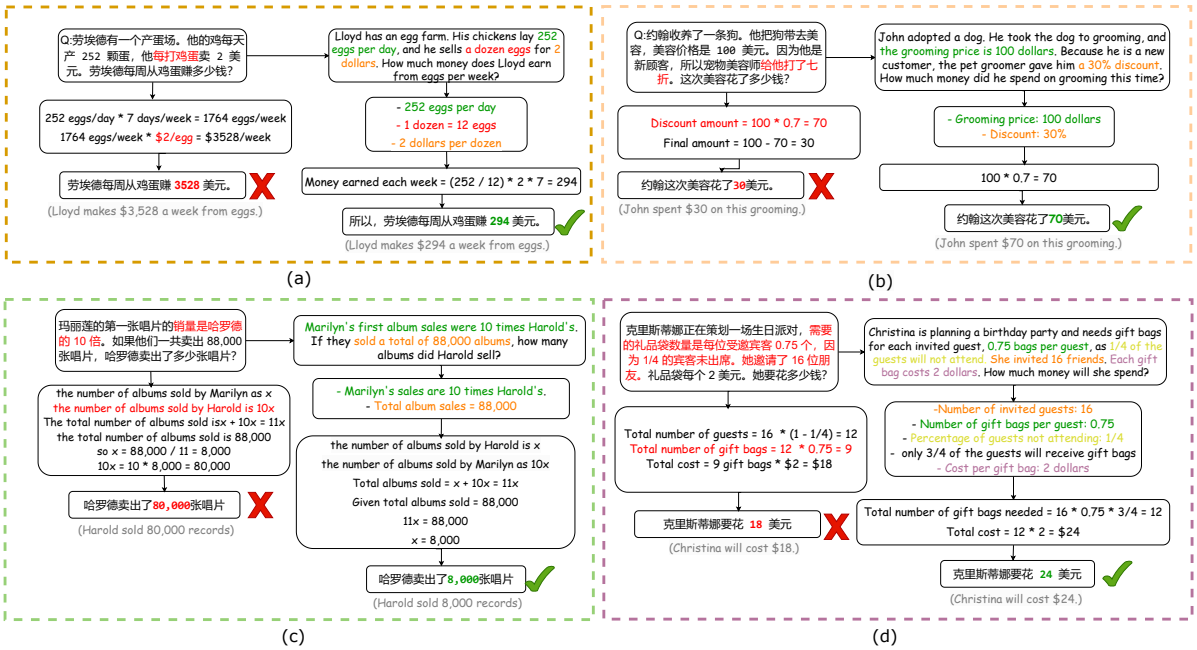

Figure 4: Case study on four examples. The left half of each example illustrates the steps derived from the COT method, and the right half presents the solution process based on our framework.

two entities' sales figures, leading to cascading errors through the reasoning process. In Figure 3 (d), COT misinterprets "0.75 bags per guest" and "1/4 of guests not attending" as separate conditions, leading to double-counting. Our framework structures these relationships explicitly, preventing such misunderstandings.

Overall, by integrating structured knowledge and dependency knowledge, our framework helps the model accurately extract and interpret relationships, reducing errors caused by ambiguous expressions across different languages.

6 Conclusion

In this paper, we propose Knowledge Funnel, a novel prompting framework for multilingual reasoning. By dynamically extracting entity-structured knowledge and language-specific dependency knowledge, our method enhances the model's ability to understand relationships in non-English questions, thereby improving multilingual reasoning performance. Experimental results demonstrate that our framework achieves significant improvements across two tasks on three baseline models, outperforming methods such as COT and XLT. The experiments and analysis further confirm that our framework offers both strong generalization capabilities and cost-effective scalability.

Limitations

Similar to previous non-parametric methods, the effectiveness of our framework depends on the performance of LLMs. Additionally, due to limitations in computational resources, our experiments

were focused on numerical reasoning and commonsense reasoning tasks. If resources permit, we plan to explore the applicability of our framework to a broader range of multilingual tasks.

References

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, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Preprint*, arXiv:2401.07037.

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.

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*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. 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.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023b. 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.

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.

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.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023a. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1403, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Haritz Puerto, Martin Tutek, Somak Aditya, Xiaodan Zhu, and Iryna Gurevych. 2024. Code prompting elicits conditional reasoning abilities in Text+Code LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11234–11258, Miami, Florida, USA. 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.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

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*.

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.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023. Natural language reasoning, a survey. *Preprint*, arXiv:2303.14725.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. 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.

A Evaluation of Multilingual Reasoning Methods across Various Tasks

Model	Method	Language										
		En	**Sw**	**Ja**	**Be**	**Th**	**Ru**	**Zh**	**De**	**Es**	**Fr**	**AVG**
GPT	Original	77.0	68.1	68.4	48.7	61.8	74.3	68.0	73.4	73.3	73.4	68.6
	COT	76.8	65.1	68.8	49.0	60.2	68.5	69.0	68.7	70.5	68.8	66.5
	Ours (Basic)	81.4	74.4	79.1	66.0	72.1	76.8	79.1	76.0	78.0	77.4	76.0
	Ours (Full)	83.2	77.0	78.2	66.5	76.4	78.2	81.2	77.2	79.6	78.8	77.6
	Ours (Basic) + COT	81.1	75.1	80.6	66.8	71.8	77.9	78.1	78.3	79.6	79.2	76.9
	Ours (Full) + COT	80.6	74.4	80.0	66.8	74.3	78.0	81.4	76.0	79.2	78.2	76.9

Table 3: Evaluation of Multilingual Numerical Reasoning Methods on MSVAMP using GPT-3.5

Model	Method	Language										
		Et	**Ht**	**Id**	**It**	**Qu**	**Sw**	**Ta**	**Th**	**Tr**	**Vi**	**Zh**	**AVG**
GPT	Original	48.2	49.6	33.8	36.8	50.2	47.0	37.8	46.0	43.4	44.8	37.0	43.1
	COT	77.0	63.4	78.6	82.2	50.0	67.0	53.2	67.2	79.6	77.0	76.0	70.1
	Ours (Basic)	80.0	62.4	84.2	85.8	50.8	74.2	60.8	74.6	81.2	76.6	82.6	73.9
	Ours (Full)	78.8	63.0	83.4	88.2	50.2	75.4	64.2	77.6	82.0	81.4	82.8	75.2
	Ours (Basic) + COT	80.4	68.0	83.6	90.0	49.8	76.0	61.2	74.8	86.0	82.0	85.4	76.1
	Ours (Full) + COT	80.0	64.2	83.6	88.0	50.2	74.8	63.4	74.8	83.4	82.0	84.2	75.3

Table 4: Evaluation of Multilingual Commonsense Reasoning Methods on XCOPA using GPT-3.5