# A Tree-of-Thoughts to Broaden Multi-step Reasoning across Languages

**Leonardo Ranaldi**$^{(\diamond)}$, **Giulia Pucci** $^{(\star)}$,
**Federico Ranaldi** $^{(\diamond)}$, **Elena Sofia Ruzzetti** $^{(\diamond)}$, **Fabio Massimo Zanzotto**$^{(\diamond)}$
($\diamond$)Human-Centric ART Group,
Dep. of Enterprise Engineering, University of Rome Tor Vergata, Italy
($\star$)Department of Computing Science, University of Aberdeen, UK
[first_name].[last_name]@uniroma2.it

## Abstract

Reasoning methods, best exemplified by the well-known Chain-of-Thought (CoT), empower the reasoning abilities of Large Language Models (LLMs) by eliciting them to solve complex tasks in a step-by-step manner. Although they are achieving significant success, the ability to deliver multi-step reasoning remains limited to English because of the imbalance in the distribution of pre-training data, which makes other languages a barrier.
In this paper, we propose Cross-lingual Tree-of-Thoughts (Cross-ToT), a method for aligning Cross-lingual CoT reasoning across languages. The proposed method, through a self-consistent cross-lingual prompting mechanism inspired by the Tree-of-Thoughts approach, provides multi-step reasoning paths in different languages that, during the steps, lead to the final solution. Experimental evaluations show that our method significantly outperforms existing prompting methods by reducing the number of interactions and achieving state-of-the-art performance.

## 1 Introduction

Chain-of-Thought (CoT) prompting elicits Large Language Models (LLMs) to break down a reasoning task towards a sequence of intermediate steps (Wei et al., 2022). Previous works have demonstrated that LLMs achieve impressive performances in zero-shot learning scenarios without the need to modify the model parameters during the training and testing process. In particular, by appending to the prompt "Let's think step by step!" (Kojima et al., 2023) LLMs with at least several billions of parameters, such as GPTs family (OpenAI, 2023) or PaLM (Chowdhery et al., 2022), deliver multi-step controlled reasoning, achieving promising results across commonsense (Bubeck et al., 2023), symbolic and mathematical reasoning datasets (Gaur and Saunshi, 2023; Liu et al., 2023).

Although the performances seem promising, they are only firmly established in English. This poses a barrier to generalizing current CoT techniques to different languages. Hence, despite the remarkable success of zero-shot CoT techniques, the reasoning abilities of LLMs still struggle to generalize to different languages. Shi et al. (2022) introduced the first multilingual benchmark to assess LLMs' mathematical reasoning abilities using prompts in different languages. Qin et al. (2023) propose task-specific solver prompting, using a succession of prompts, elicit the LLMs to understand questions and deliver CoT answers in different languages. However, these strategies require two-step prompts, which goes against the zero-shot approach.

In this paper, we propose Cross-lingual Tree-of-Thoughts (Cross-ToT), a method for aligning Cross-lingual CoT reasoning across languages by proposing a Cross-lingual Alignment prompt to elicit the model to deliver a Self-consistent Chain-of-Thougt. Our method is inspired by the Tree-of-Thoughts (ToT) prompting (Yao et al., 2023) that asks LLMs to perform decision-making by considering multiple different reasoning paths (CoTs). In particular, our Cross-ToT is a ToT-style prompting to deliver the reasoning process in different languages that, step-by-step, converge to a single final solution. The inherent insight is that as the different paths of thought evolve, the relationships between the different languages are inherently grasped via Self-consistent Chains-of-Thougt. This leads to the target research questions, which are the focus of this paper:

*RQ1:* Are LLMs able to deliver Cross-lingual multi-step reasoned answers?

*RQ2:* Are the different paths of ToT evolving Self-correcting each other?

*RQ3:* What is the role of English in Cross-lingual scenarios?

To answer these questions, we propose `Cross-ToT`, a novel Cross-lingual prompting strategy that aims to bridge the gap across different



*Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241

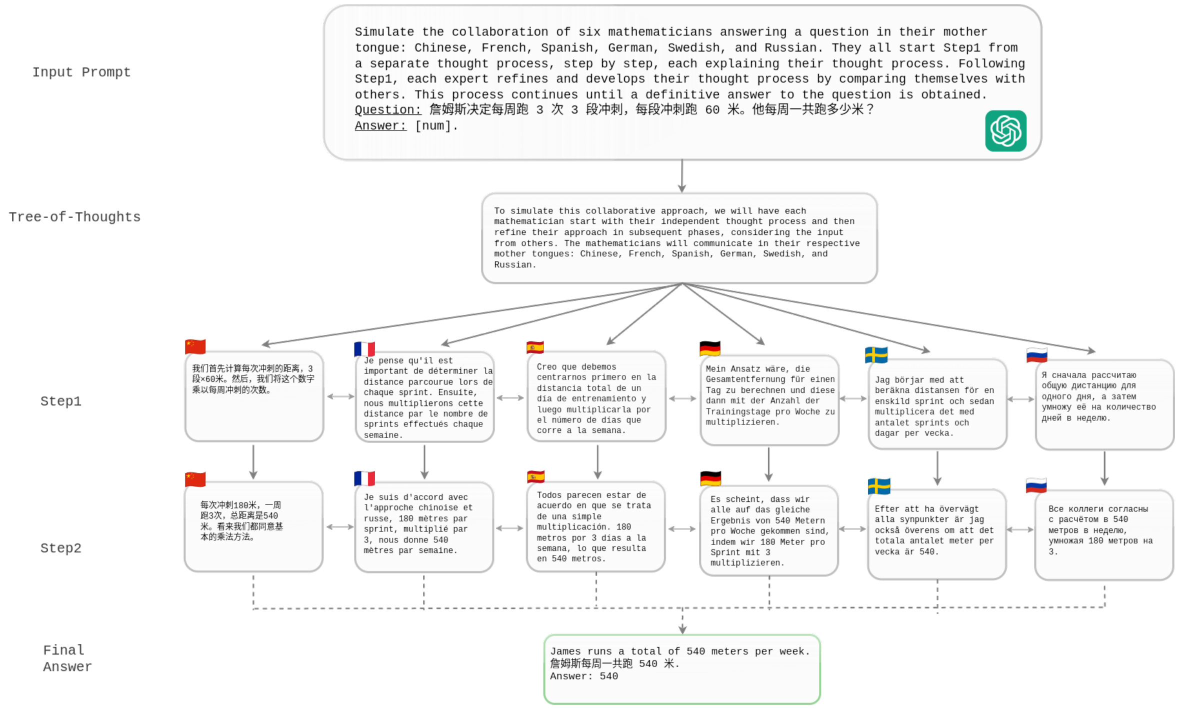


Figure 1: Our Cross-ToT elicits the LLM to generate step-by-step Cross-lingual reasoning. Furthermore, different pathways are developed during these reasoning steps. This mechanism develops the Chains-of-Thoughts in a Self-consistent way, streaming with the different pathways.

languages. In particular, using the prompt shown in Figure 1, we elicit the model to deliver different CoT reasoning steps in different languages that converge to the final solution step-by-step. We test our method on GPT-3.5 and conduct an extensive analysis using Multilingual Grade School Math (MGSM) (Shi et al., 2022), Cross-lingual Natural Language Inference (XNLI) (Conneau et al., 2018), and Cross-lingual Paraphrase Adversaries Scrambling (PAWS-X) (Yang et al., 2019), Cross-lingual Choice of Plausible Alternatives (XCOPA) (Ponti et al., 2020) across different languages. Experimental results reveal that our method, based on a single prompt, outperforms the baselines and achieves the SOTA performance on different languages in different tasks. The main contributions of this work are concluded as follows:

- We introduce `Cross-ToT`, which is a novel Cross-lingual prompting mechanism that stimulates the model to produce parallel CoT reasoning processes across different languages;
- We show that our `Cross-ToT` is Self-consistent and allows the integration of reasoning paths between different languages;
- Extensive evaluations on different languages demonstrate that our `Cross-ToT` can effectively improve the performance of cross-lingual CoTs and achieve SOTA performance.
- Finally, we show that introducing English in our prompting technique plays a beneficial role in improving downstream performance.

## 2 Cross-lingual Multi-step Reasoning

To elicit the multi-step reasoning abilities of LLMs in Cross-lingual scenarios, we propose `Cross-ToT`, which is a Cross-lingual Alignment Chain-of-Thought as a solution. In particular, our method overcomes the Multi-lingual and Cross-lingual approaches introduced in Section 2.1. In fact, our approach elicits the LLMs to deliver Self-consistent Parallel Chain-of-Thougts, introduced in Section 2.2.

### 2.1 Chain-of-Thought Across Languages

The Cross-lingual Alignment is a core challenge for cross-lingual transfer. Shi et al. (2022) proposed a series of prompts to elicit models to generate CoT answers in specific language `Native-CoT`, and in English `En-CoT` and `Translate-CoT` (more detailed in Table 1).

Later, Qin et al. (2023) proposed a method based on two phases: Cross-lingual alignment prompt and

**Native-CoT** in this example in Chinese

| |
|---|
| 问题: 利亚有32 块巧克力，她妹妹有42 块。如果她们吃了35 块，她们一共还剩下多少块?<br>答案: 让我们一步步思考 |

**En-CoT**

| |
|---|
| 问题: 利亚有32 块巧克力，她妹妹有42 块。如果她们吃了35 块，她们一共还剩下多少块?<br>`Answer: Let's think step by step` |

**Translated-CoT** (is the Native translated in En)

| |
|---|
| `Question: Leah has 32 chocolates and her sister has 42. If they ate 35 pieces, how many pieces do they have left?`<br>`Answer: Let's think step by step` |

Table 1: Different types of input prompts in order to elicit Chain-of-Thought reasoning process. Specifically, given a problem in Chinese, the following prompts are Native-CoT and En-CoT, the original question in Chinese with elicitation in Chinese and English; for Translated-CoT, the question is in English and consequently a step-by-step solution in English.

task-specific solver prompting. This approach uses two separate steps, as shown in Table 2, in order to handle input and output in different languages.

**Cross-CoT** First-Step

| |
|---|
| Please act as an expert in multi-lingual understanding in [Specific Language $L_s$].<br>Question: [Given sentence $X$ in $L_s$]<br>Let's understand the task in [Target Language $L_t$] step-by-step! |

**Cross-CoT** Second-Step

| |
|---|
| After understanding, you should act as an expert in mathematics in [Language $L_t$].<br>Let's resolve the task you understand above step-by-step! |

Table 2: Cross-lingual Prompt proposed in (Qin et al., 2023). By setting an input language and a target language, the prompt is divided into two phases: in phase one, there is the alignment of the different languages, and in phase two, there is the solving mechanism for the specific language.

Although this second approach overcomes the limitations of Shi et al. (2022)'s work, the two-step prompting could be more laborious and challenging, and there is no exchange of information during the multi-step reasoning process between the different chains as the final outputs are estimated using a voting heuristic.

### 2.2 Self-consistent Parallel Chain-of-Thougts

In our work, we propose Cross-ToT, a prompting method that can handle different languages in a parallel way. Furthermore, through a mechanism inspired by Tree-of-Thoughts prompting techniques (Yao et al., 2023), our method elicits the LLM to deliver the generation of the answer in a sequence of intermediate steps that do not provide independent parallel answers but deliver collaborative Self-consistent reasoned steps until arriving at a final answer.

**Our Proposal**

| |
|---|
| Simulate the collaboration of $\{n\}$ mathematicians answering a question in their mother tongue: $L_1$, $L_2$, ... and $L_n$. They all start Step1 from a separate thought process, step by step, each explaining their thought process. Following Step1, each expert refines and develops their thought process by comparing themselves with others. This process continues until a definitive answer to the question is obtained.<br>Question: [Question in Language $L_1$]<br>Answer: [num]. |

Table 3: Input-prompt for MSGM task. In Cross-ToT, we elicit the model to produce multi-step reasoning processes in different languages. We specifically prompt to start from separate reasoning and collaborate step-by-step. (We propose similar pattern for other tasks as described in Appendix A)

Our Cross-ToT shown in Table 3 elicits the LLM to generate different paths as shown in Figure 1, achieving significant improvements in accuracy as discussed in Section 4.

## 3 Experiments

### 3.1 Data

In order to observe the Cross-lingual abilities of LLMs, we used GSM8K (Cobbe et al., 2021), XNLI (Conneau et al., 2018), and PAWS-X (Yang et al., 2019), XCOPA (Ponti et al., 2020).

**Understanding tasks** In order to assess Cross-lingual comprehension abilities, we used XNLI (Conneau et al., 2018) and PAWS-X. The first is an extension of Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) across 15 languages and is based on one premise and one hypothesis and requires the model to determine whether the hypothesis is entailed, contradicted, or neutral conditioned on the premise in 15 different languages, and we utilize the accuracy score for evaluation. The second, Paraphrase Adversaries from Word Scrambling (PAWS-X) (Yang et al., 2019), contains two sentences and requires the model to judge whether they paraphrase each other in seven languages.

**Commonsense Reasoning task** The Cross-lingual Choice of Plausible Alternatives (XCOPA)

| Model | de | zh | fr | ru | sw | es | bn | ja | te | th | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **GPT-3 (text-davinci-002)*** | | | | | | | | | | | |
| `Direct` (Shi et al., 2022) | 14.8 | 18.0 | 16.8 | 12.4 | 8.8 | 17.2 | 4.4 | 11.2 | 0.8 | 8.8 | 11.3 |
| `Native-CoT` (Shi et al., 2022) | 36.0 | 40.0 | 37.6 | 28.4 | 11.2 | 40.4 | 6.4 | 26.0 | 0.4 | 10.8 | 23.7 |
| `En-CoT` (Shi et al., 2022) | 44.0 | 40.8 | 46.0 | 28.4 | 20.8 | 44.8 | 9.6 | 32.4 | 5.6 | 19.6 | 29.2 |
| `Translate-En` (Shi et al., 2022) | 46.4 | 47.2 | 46.4 | 48.8 | 37.6 | 51.6 | 41.2 | 44.8 | 42.8 | 41.2 | 44.8 |
| **GPT-3.5 (gpt-3.5-turbo)** | | | | | | | | | | | |
| `Direct` (Qin et al., 2023) | 56.0 | 60.0 | 62.0 | 62.0 | 48.0 | 61.2 | 33.6 | 52.8 | 7.6 | 42.2 | 48.5 |
| `Native-CoT` (Qin et al., 2023) | 70.0 | 59.6 | 64.4 | 62.4 | 54.0 | 70.4 | 26.4 | 64.4 | 40.0 | 59.6 | 57.1 |
| `En-CoT` (Qin et al., 2023) | 73.6 | 63.2 | 70.0 | 65.6 | 55.2 | 69.6 | 50.0 | 60.4 | 22.0 | 48.0 | 57.7 |
| `Translate-En` (Qin et al., 2023) | 75.6 | 71.6 | 72.4 | 72.8 | 69.6 | 74.4 | 66.4 | 66.0 | 58.0 | 57.6 | 68.4 |
| `Cross-CoT` (Qin et al., 2023) | 86.8 | 77.2 | 82.0 | **87.6** | **76.0** | 84.8 | 75.2 | 77.2 | 52.0 | 68.0 | 76.6 |
| **`Cross-ToT`** | **87.6** | **83.5** | **84.3** | 86.5 | 75.4 | **86.2** | **79.0** | **80.2** | **68.5** | **75.5** | **80.6** |

Table 4: Accuracies (%) on MGSM using the `"Direct"` prompt, i.e., question and answer in the original language; the `"Native-CoT"` prompt, i.e., question and answer CoT in the original language; the `"En-CoT"` prompt specific language question and answer CoT in English, the `"Translate-En"` prompt where the specific input is translated into English and the answer accordingly is in English. Moreover, Cross-CoT, as proposed by Qin et al. (2023), questions in a specific language and answers in different languages. Finally, **`Cross-ToT`** is explained in Section 2.2. (Our results are derived from the average of three running performances as detailed in Section 3.2)

(Ponti et al., 2020) is based on one premise and two choices. It asks the model to choose which one is the result or cause of the premise. It covers 11 languages from 11 diverse families.

**Arithmetic Reasoning task** To evaluate the problem-solving abilities in Cross-lingual scenarios, we used the extension proposed by Shi et al. (2022), i.e., Multilingual Grade School Math (MGSM). Initially, Cobbe et al. (2021) proposed a benchmark of mathematical problems in English in GSM8K. Each example has the following structure: a mathematical problem in natural language and a target answer in Arabic number. Shi et al. (2022), in their contribution, i.e., MGSM, selected the first 250 examples from the official list of examples in GSM8K and translated them manually into 11 different languages, maintaining the structure of the input and output.

**Evaluated Languages** In our experiments, we propose an analysis of available languages that differ depending on the resources, we provide all details in Appendix A. Furthermore, as an additional experiment, we test the introduction of English.

### 3.2 Experimental Setup

In order to conduct our study on robust models and have a term of comparison with the work proposed in (Shi et al., 2022; Qin et al., 2023), we use GPT-3.5; however, in future developments, we plan to scale the method to different models. Then, we systematically defined the input prompt in Table 3 for MGSM and in Appendix A for XNLI, PAWS-X, and XCOPA. In each particular experimental set-up, we modify the appropriate languages with $L_1$, $L_2$, ...for the German [1]

Following Wei et al. (2022); Kojima et al. (2023), we evaluate performance using the accuracy score. In particular, we compute the string matching between the final answers (see Figure 1 where the final outputs have the form of `Answer:[num]`) and the target values. The top-p parameter is set to 1 in all processes. We select the Prompting temperature [0, 1].

## 4 Main Results

Mechanisms for delivering multistep-reasoned answers across languages can be empowered via `Cross-ToT` that align languages' Chain-of-Thoughts (CoT). Our approach based on a Tree-of-Thoughts-inspired prompting mechanism (see Figure 1) outperforms state-of-the-art prompting techniques on Arithmetic Reasoning tasks as shown in Table 4, and in Language Understanding tasks as shown in Figure 3 and finally in Commonsense Reasoning tasks as shown in Table 5. In particular, `Cross-ToT` elicit LLMs to produce different reasoning pathways that share the "Thoughts" during the steps and, at the same time, promote Self-

[1] Although we do not observe perceptible changes in the order of languages present in the input prompt, we set as a first the language-related subset of the benchmark.

correction of mistaken paths. In fact, during the steps of the CoT, information is swapped between the paths. This interaction delivers Self-consistent paths. Furthermore, in the prompt, we exemplified that the different paths must arrive at a shared and, consequently, unique by sharing the "thought process" (see the prompt in Table 3).

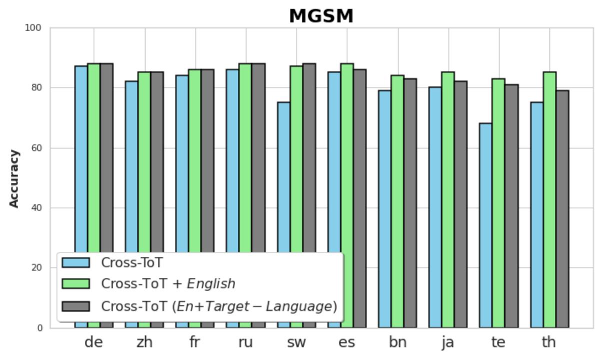


Figure 2: Accuracies (%) on MGSM using "`Cross-ToT`", "`Cross-ToT` + English" and in binary version "`Cross-ToT` ( English + Target Language)".

Our approach outperforms the methods proposed in (Shi et al., 2022) that are yet surpassed by the `Cross-CoT` proposed by Qin et al. (2023). However, although `Cross-CoT` outperforms previous approaches, it is necessary to clarify which path, if any, leads to the correct reasoning (Section 5.3), whether the introduction of English can increase performance (Section 5.1) and finally the trade-off between the number of languages (in our case path) and the final results (Section 5.2).

## 5 Analysis

In this section, we explore the contribution of English in the Cross-lingual prompt (in Section 5.1), then study the impact of different languages on the final results (Section 5.2) and the reasoning evolution (Section 5.3) and close with an in-depth analysis of performance in different tasks in Section 5.4.

### 5.1 The English Matter

Earlier works (Wei et al., 2022; Liu et al., 2023) have been showing that LLMs are able to deliver multi-step reasoning answers on arithmetic tasks, focusing mainly on English. Therefore, we observe whether introducing English into our input-prompts could increase downstream performance. Hence, we performed the setting proposed in Section 3.2 From the results obtained in Figure 2

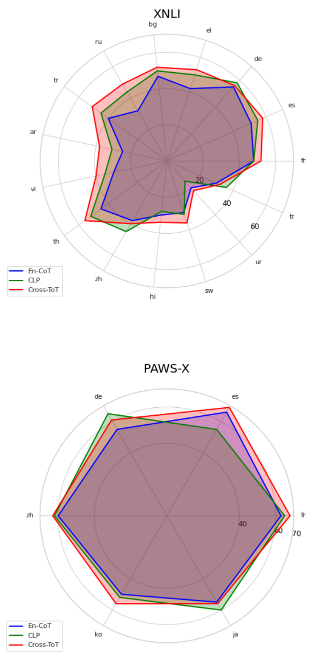


Figure 3: Accuracies (%) on Language Understanding benchmarks XNLI and PAWS-X introduced in Section 3.1

(green bar), it is possible to observe that the input-prompts empowered with English outperform the input-prompts empowered without English. This result suggests that the presence of one robust path, in this case, the English path, may influence the others in the final reasoning process. Indeed, assuming that the production of the intermediate steps is self-consistent, i.e., the paths do not disagree with each other, the additional language seems to influence performance positively. From the current results, adding a further language improves the robustness of the models.

However, whether the performance is due to the number of languages or English is unclear. To observe the impact of adding a specific language in Section 5.2, we propose to reduce the number of languages in the presence and absence of English.

### 5.2 The Impact of the Languages

English seems to lead Cross-lingual reasoning on arithmetic tasks, as shown in Section 5.1. Hence, to observe the impact of the number of languages and one specific, i.e., English, we propose two further analyses:

**Cross-ToT in low-resources scenarios** Integrating more languages into Cross-lingual prompting leads to better overall performance. As already

| Model | et | ht | id | it | qu | sw | ta | th | tr | vi | zh | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **GPT-3 (text-davinci-002)*** | | | | | | | | | | | | |
| `Direct` (Shi et al., 2022) | 73.8 | 55.6 | 88.8 | 95.4 | 51.2 | 56.0 | 54.6 | 70.2 | 88.6 | 80.4 | 91.4 | 73.3 |
| `En-CoT` (Shi et al., 2022) | 88.8 | 79.6 | 91.4 | 96.6 | 52.2 | 67.4 | 55.8 | 84.2 | 91.2 | 86.6 | 93.4 | 80.7 |
| **GPT-3.5 (gpt-3.5-turbo)** | | | | | | | | | | | | |
| `Direct` (Qin et al., 2023) | 90.6 | 72.0 | 90.4 | 95.2 | 54.6 | 82.0 | 59.0 | 77.6 | 91.0 | 83.6 | 90.4 | 80.6 |
| `Translate-En` (Qin et al., 2023) | 88.2 | 79.4 | 90.8 | 94.4 | 50.0 | 77.6 | 87.0 | 82.2 | 87.8 | 88.4 | 92.2 | 83.5 |
| `Cross-CoT` (Qin et al., 2023) | 96.8 | 90.6 | 95.2 | 95.8 | **85.8** | 92.8 | **83.2** | 93.2 | **96.8** | 94.2 | 95.8 | 92.7 |
| **`Cross-ToT`** | **97.6** | **92.5** | 90.3 | **96.8** | 83.3 | **93.6** | 80.2 | **94.1** | 96.4 | **95.3** | **97.4** | |
| HUMAN (Ponti et al., 2020) | 98.2 | 96.4 | 100.0 | 97.0 | 94.8 | 99.0 | 98.6 | 98.2 | 96.4 | 98.4 | 96.6 | 97.6 |

Table 5: Accuracies (%) of XCOPA.

observed in (Shi et al., 2022; Qin et al., 2023), increasing the number of languages improves downstream performance, as shown in Figure 4 (average performances using the same setting proposed in Section 3.2).

As shown in (Malkin et al., 2022; Blevins and Zettlemoyer, 2022), the performances of the Large Language Models are highly correlated with the percentage of pre-training data in each language.

Following the approach proposed in (Qin et al., 2023) and considering language distribution in the widely used multilingual pre-training dataset, which in our case is CommonCrawl (Common Crawl, 2021), we integrated languages in descending and ascending order based on their respective proportions (detailed in Table 12).

Figure 4 shows that adding more languages in high-resource contexts improves performance. However, when incorporating languages with limited resources, performance decreases as the number of languages increases (see low-resource in Table 4). Finally, adding English (the dominant percentage in standard corpora) to the prompting significantly enhances performance (see "+ English" lines in Table 4).

These findings emphasize that the number of integrated languages only partially determines the effectiveness of language integration. The amount of pre-training data for each language, especially for high-resource languages, plays a crucial role. Balancing multiple languages and considering available resources and impact is essential.

**Cross-ToT in binary scenarios** Moreover, we evaluate similar scenarios in low-resource settings and reproduce the same experiments using only two languages. In particular, we used the same setting proposed in Section 3.2 by including only the target language and English in the prompt (example prompt in Appendix 8).

From the results shown in Figure 2 (grey bar), using the target English-language tuple does not change the performance of high-resource languages. On the contrary, low-resource languages achieve significantly lower performance. This second finding reinforces what was said earlier about the experiments on prompt compositions.

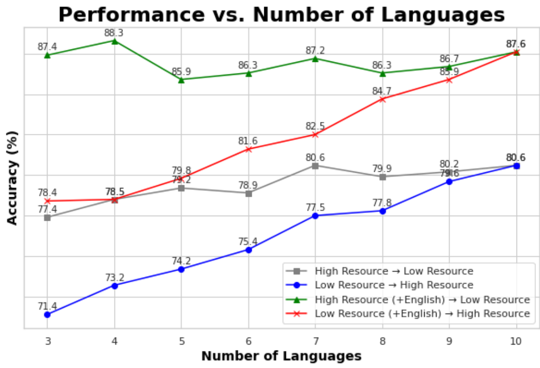


Figure 4: The impact of integrating languages in our `Cross-ToT` on the final performance. Following Table 12, we integrate languages from low-resources to high-resources and vice versa. We also propose the same experiments with the addition of English.

### 5.3 Reasoning Evolution

We use the framework `ROSCOE` (Golovneva et al., 2023) to investigate why our approach works. Hence, we evaluate the quality of the reasoning paths (implementation described in Appendix B). As shown in Figure 5, our approach delivers reasoning with higher faithfulness, exhibiting better consistency with key steps during the reasoning process. Specifically, the faithfulness score increased by 4.5 points, indicating that the model better understood the problem statement and ensured a transparent inference chain without generating irrelevant or misused information. Furthermore, we observe improvements in the Informativeness metrics for “Step” and “Chain”. It suggests that the models’ reasoning, behind the alignment, could provide more well-grounded inference steps.


Figure 5: The analysis of reasoning quality between GPT-3.5 (Native-CoT) and CLP in (Qin et al., 2023) and our `Cross-ToT`

**XCOPA, XNLI and PAWS-X**

```
Simulate the collaboration of n person answering a question in their mother tongue: L1 and English. They all start Step1 from a separate thought process, step by step, each explaining their thought process. Following Step1, each expert refines and develops their thought process by comparing themselves with others. This process continues until a definitive answer to the question is obtained.
Basic Prompt
```

Table 6: Our prompting approach for XCOPA, XNLI and PAWS-X. List of the `Basic Prompt` is in Table 11

## 5.4 The Cross-Reasoning in other tasks

Furthermore, to scale our approach, we test the applicability of `Cross-ToT` on two different task types using the same structure adapted to them as in Table 7.

**Understanding task** We proposed our approach, Cross-ToT, on other multilingual reasoning datasets belonging to the undertandings genre. As introduced in Section 3.2, we used XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019). As Figure 3 shows, Cross-ToT is able to perform better in most languages. Compared to the previous SOTA obtained in CLP (Qin et al., 2023). Thus, we observed average improvements of 3.2 points on XNLI and 2.5 points on PAWS-X.

**Commonsense Reasoning task** We have used our approach, Cross-ToT, to an additional dataset of multilingual commonsense reasoning, as introduced in Section 3.1. We used XCOPA as our benchmark. For comparison purposes, we considered CLP and Native-CoT proposed by Qin et al. (2023). In Figure 5, we can observe that our approach has outperformed previous methods in many languages.

The results show the effective functionality of our `Cross-ToT` on different tasks. Although the method has shown appreciable increases, we continue the studies in Section 5.5 by observing whether adding in-context examples in the input-prompt can benefit LLMs.

## 5.5 Other approaches

`Cross-ToT` can be further empowered with in-context learning. In fact, as shown in Table 9, in-context learning (ICL) techniques have achieved performant results on the downstream performance of LLMs. In particular, in further exploration of `Cross-ToT` within ICL, we conducted different experiments.

**From Zero- to Few-shot** In the first experiment, we sampled 50 random instances from MGSM. Then, we replicated the experiments proposed in Section 3.2. However, we constructed the prompt by merging instances in one-shot and three-shot settings. Table 9 shows that providing context makes the models more robust.

**Performances Other Models** `Cross-ToT` does not outperform other approaches in open-source models with fewer parameters. Table 10 shows the performances of Llama-2-13B (Touvron et al., 2023) and Bloomz-7B (Muennighoff et al., 2022). We hypothesize that these performances are due to the misleading behaviors observed in (Wei et al., 2023) prompting CoT in models with less than 100 billion parameters. In future developments, we will continue to investigate this phenomenon.

# 6 Related Work

Large Language Models (LLMs) with billions of parameters demonstrate in-context learning and few-shot learning abilities (Brown et al., 2020; Wei et al., 2022; Min et al., 2022) to guide LLMs to generate desired task responses, marking the advent of the prompting era and surpassing the age of the intermediate steps in algorithmic and structured reasoning (Roy and Roth, 2015; Ling et al., 2017). Nevertheless, early works challenged the efficacy of few-shot techniques for empowering the prompting phase and downstream performances. In particular, Yao et al. (2023) refined the original idea of Chain-of-Thought (CoT) (Wei et al., 2022) by considering various reasoning paths as well known as Tree-of-Thought.

The traditional and derivated CoT mechanisms have achieved considerable success but are limited to generating answers within a single language (i.e., English). Shi et al. (2022) proposed a multilingual evaluation that Qin et al. (2023) extended to cross-lingual scenarios. In particular, Qin et al. (2023) proposed a prompt mechanism to handle requests in any language and generate CoT specifically in English. This approach, which in our construct we called Cross-CoT has been proposed both single-phase, i.e., as a single prompt (CLP) also adopted by (Huang et al., 2023) and multi-phase (CLPS) i.e., characterized by self-consistent prompts that follow the prompting methodology proposed in (Qiao et al., 2023). Although the mechanism achieves state-of-the-art cross-linguistic reasoning steps, the single-phase promting underperforms in low-resouces languages and the multi-phase prompting characterized by a series of cascading prompts is supported far away from the zero-shot chain-of-thought concept.

In our work, we propose a method of CoT reasoning inspired. Specifically, we elicit the cross-lingual generation of a series of parallel Cross-lingual reasoning paths using a single prompt. In fact, our method is inspired by the Tree-of-Thoughts approach proposed by (Yao et al., 2023). Hence, in a different way from previous approaches, our technique generates shared parallel reasoning paths that share the "thoughts process" delivering Self-consistent answers and reducing reasoning steps. Our work goes beyond in the following ways:

- Proposal of novel zero-shot prompting methods in cross-lingual scenarios characterized by low-resource and high-resource languages.
- Studying cross-lingual multi-step reasoning mechanisms using arithmetic reasoning tasks.
- In-depth study of the reasoning pathways provided by our prompting approach (impact of the number of languages and strongly high-resource languages).
- Experiments on effective functioning in commonsense reasoning and language understanding tasks.

## 7 Future Works

In future work, we intend to incorporate smaller-scale Language Models (SLMs) into our evaluations. However, the ability to produce multi-step reasoned answers is limited in SLMs. To address this, a range of techniques are emerging to align and transfer reasoning abilities between LLMs and SLMs (Ranaldi and Freitas, 2024).

Our aim is to enhance current alignment pipelines (Ranaldi et al., 2023; Ranaldi and Pucci, 2023a) to enable cross-lingual reasoning capabilities across different languages and scenarios. Including methods that emphasize the importance of language structure (Zanzotto et al., 2020) and uphold the foundational pillars of the NLP ecosystem (Ranaldi and Pucci, 2023b).

## 8 Conclusion

Chain-of-Thought is an outstanding prompting technique. However, the imbalance of languages in pre-training data does not always produce robust results. Different state-of-the-art works have proposed cross-lingual techniques to align performances obtained in different languages. They are limited to handling one language at a time or proposing multiple prompting stages, making them difficult to manage. In this paper, we propose `Cross-ToT`, a prompting technique to elicit multi-step reasoning abilities in Cross-lingual scenarios. Hence, we elicit models to deliver answers in a Self-consistent way, collaborating to the final answer. We have shown the functionality of our `Cross-ToT` through performance improvements obtained in a multilingual mathematical problem task. In addition, we have demonstrated the scalability in tasks related to commonsense reasoning and language understanding. Finally, we conducted a series of in-depth analyses in which we measured the impact brought about by low-resource vs. high-resource languages and the inclusion of English. Our contribution aims to propose more robust models that can break down issues arising from language barriers and provide more reliable results.

## Limitations

Due to the limitations imposed by the evaluation benchmarks and the cost of the OpenAI API, we conducted tests on 16 languages in total, which only scratches the surface of the world's vast array of languages. Furthermore, our approach is based on English. It should be evaluated whether the model written in the language of the task can lead to better performance and how best to construct instructions in each language. Furthermore, we only tested the effectiveness of our method on GPT-based models (gpt-3.5-turbo). In the future, it will be worthwhile to study the generality of our model on more models, such as PaLM and Llama-2-70.

## Ethics Statemets

In our work, ethical topics were not addressed. The data used comes from open-source benchmarks, and statistics on language differences in commonly used pre-training data were obtained from official sources without touching on issues related to gender, sex, or race differences.

## References


Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explain the cross-lingual capabilities of english pretrained models.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Common Crawl. 2021. Common crawl 2021. Web. Accessed: 2023-12-12.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5889–5903, Toronto, Canada. Association for Computational Linguistics.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne 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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*

*(Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. 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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

OpenAI. 2023. Gpt-4 technical report.

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.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. 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.

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2023a. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2023b. Knowing knowledge: Epistemological study of knowledge in transformers. *Applied Sciences*, 13(2).

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2023. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. 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. 2022. Language models are multilingual chain-of-thought reasoners.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online. Association for Computational Linguistics.

## A Prompt

In this paper, we analyze our prompting approach, i.e., Cross-ToT, in different tasks. In Figure 1 we have shown the input-prompt for the MGSM (Cobbe et al., 2021). Here, we show the prompt framework for the other tasks:

**XCOPA, XNLI and PAWS-X**

| |
|---|
| Simulate the collaboration of $n$ person answering a question in their mother tongue: $L_1$ and $English$. They all start Step1 from a separate thought process, step by step, each explaining their thought process. Following Step1, each expert refines and develops their thought process by comparing themselves with others. This process continues until a definitive answer to the question is obtained. **Basic Prompt** |

Table 7: Our prompting approach for XCOPA, XNLI and PAWS-X. List of the Basic Prompt is in Table 11

Furthermore, in Section 5.1, we proposed an experiment based on a prompt with only two languages as follows:

**Binary Cross-ToT**

| |
|---|
| Simulate the collaboration of 2 mathematicians answering a question in their mother tongue: $L_1$ and $English$. They all start Step1 from a separate thought process, step by step, each explaining their thought process. Following Step1, each expert refines and develops their thought process by comparing themselves with others. This process continues until a definitive answer to the question is obtained.<br>Question: [Question in Language $L_1$]<br>Answer: [num]. |

Table 8: Our prompting approach for experiment proposed in Section 5.1 regarding MGSM and binary trees

## B Reasoning Chain

### B.1 Chain-of-Thought Quality Scoring Implementation

The ROSCOE framework (Golovneva et al., 2023) incorporates multiple chain-of-thought quality metrics, with the reasoning alignment vector $\alpha$ that is

$$r_{align}(h \to s) = \{\alpha_1, \alpha_2, \dots, \alpha_N\} \in [0,1]^N \quad (1)$$

from the $N$-step hypothesis $h = \{h_i\}_{i=1}^N$ to the source input $s$ of length $T$, where $\alpha_i$ are defined as: $r_{align}(h_i \to s) = \frac{1+\max_{j=1}^T \cos(h_i, s_j)}{2}$

**Faithfulness score** The Faithfulness $(F)$ score is calculated based on the alignment between the hypothesis steps $h$ and the source sentences $s$. It represents the average reasoning alignment score over the steps of reasoning:

$$F = \frac{1}{N}\sum_{i=1}^{N} r_{align}(h_i \to s) \quad (2)$$

The Faithfulness score serves as a measure to assess whether the model misconstrued the problem in the statement or if the reasoning chain is characterized by ambiguity, unimportance, or the misuse of information.

**Informativness** Informativeness-Step (Info-Step) measures the utilization of facts from the original text $s$ in the reasoning steps $h$:

$$Info_{Step} = \frac{1}{2T}\sum_{t=1}^{T} r_{align}(s_t \to h) + \frac{1}{2}F \quad (3)$$

Info-Step assigns a higher score to reasoning steps that strongly align with the source, showing the capacity to which the generated hypothesis includes the information from the source. Conversely, a lower Info-Step score means reasoning steps unrelated to the source sentences or overlooking the provided information in the context.

**Informativeness Chain** Like the Info-Step metric, the InformativenessChain (Info-Chain) metric estimates the degree of concordance between the hypothesis chain and the source. The calculation is as follows:

$$Info_{Chain} = \frac{1+\cos(h,s)}{2} \quad (4)$$

**Missing Step** The Missing Step (Miss-Step) metric is introduced to estimate any significant lacking steps, which examines the alignment between the reference reasoning text $r = \{r_i\}^K$ and the hypothesis $h$. A miss-step is needed to meticulously assess each step in the reference and verify the existence of a similar step in the hypothesis. The metric is computed as:

$$\text{Miss-Step} = \min_{i=1}^{K}(\text{r-align}(r_i, h)). \quad (5)$$

## C Other Results

| # of shot- Cross-ToT | de | zh | fr | ru | sw | es | bn | ja | te | th | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0-shot** | 86.5 | 84.2 | 83.9 | 83.2 | 74.3 | 84.4 | 78.7 | 79.8 | 68.7 | 74.6 | 79.8 |
| **1-shot** | 87.2 | 84.9 | 85.8 | 85.3 | 76.4 | 85.2 | 81.2 | 81.3 | **70.5** | 75.5 | 79.9 |
| **3-shot** | **88.4** | **85.7** | **87.2** | **87.5** | **77.3** | **87.3** | **82.3** | **81.5** | 70.3 | **76.9** | 83.4 |

Table 9: Accuracies (%) on MGSM using zero-shot, one-shot and three-shot

| **Model** | **et** | **ht** | **id** | **it** | **qu** | **sw** | **ta** | **th** | **tr** | **vi** | **zh** | **Avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bloomz-7B (Muennighoff et al., 2022) | | | | | | | | | | | | |
| En-CoT | 21.8 | 24.2 | 50.6 | 41.6 | 41.4 | 48.6 | 53.8 | 38.4 | 37.6 | 47.0 | 64.2 | 42.7 |
| CLP (Qin et al., 2023) | **49.0** | **49.6** | 58.0 | **48.8** | **50.6** | **47.6** | **57.8** | 52.0 | **50.2** | **45.2** | **54.2** | **51.2** |
| **Cross-ToT** | 48.0 | 47.3 | **58.2** | 47.8 | 49.3 | 46.4 | 55.2 | **53.1** | 50.8 | 44.2 | 50.3 | 49.5 |
| llama-2-13B (Touvron et al., 2023) | | | | | | | | | | | | |
| En-CoT | 39.6 | 32.5 | 58.4 | 55.8 | 47.2 | 34.6 | 47.4 | 33.2 | 43.0 | 59.6 | 50.4 | 45.6 |
| CLP (Qin et al., 2023) | **44.8** | 48.2 | **64.4** | **70.2** | **46.6** | **47.0** | **47.8** | **46.4** | **51.2** | **58.8** | **51.4** | **52.4** |
| **Cross-ToT** | 43.3 | **49.1** | 61.5 | 65.8 | 44.4 | 46.6 | 43.7 | 42.2 | 49.5 | 55.2 | 48.2 | 50.6 |

Table 10: Comparison of smaller open-source models on XCOPA.

## D Prompt Table

| Benchmark | #Test | Basic Prompt |
|---|---|---|
| MGSM | 250 | Question: {problem} |
| XCOPA | 200 | Here is a premise: {premise}. What is the {question}? Help me pick the more plausible option: -choice1: {choice1}, -choice2: {choice2} |
| XNLI | 200 | {premise}. Based on the previous passage, is it true that {hypothesis}? Yes, No, or Maybe? |
| PAWS-X | 200 | Sentence 1: {sentence1} Sentence 2: {sentence2} Question: Does Sentence 1 paraphrase Sentence 2? Yes or No? |

Table 11: The basic prompt of each benchmark. #Test denotes the number of instances in the test set that we randomly selected due to the cost constraint excepted for MGSM.

## E Number of Languages

| **Language** | **Percentage** |
|---|---|
| English (en) | 46.3% |
| Russian (ru) | 6.0% |
| German (de) | 5.4% |
| Chinese (zh) | 5.3% |
| French (fr) | 4.4% |
| Japanese (ja) | 4.3% |
| Spanish (es) | 4.2% |
| Other | 23.1% |

Table 12: Language distribution of CommonCrawl (Common Crawl, 2021).