AUTOCAP: Towards Automatic Cross-lingual Alignment Planning for Zero-shot Chain-of-Thought

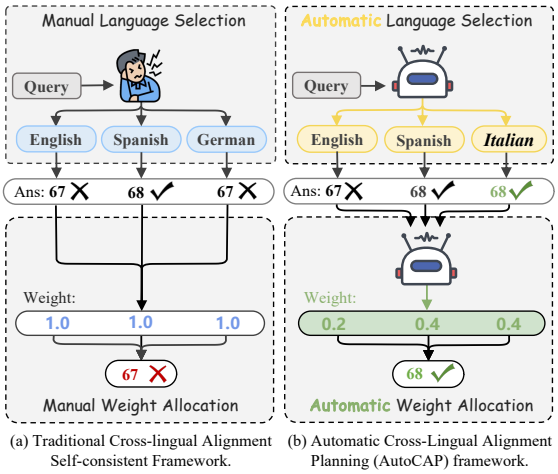
Yongheng Zhang[1*] **Qiguang Chen**[2*] **Min Li**[1] **Wanxiang Che**[2] **Libo Qin**[1†]
[1]School of Computer Science and Engineering, Central South University, China
[2]Research Center for SCIR, Harbin Institute of Technology, Harbin, China
zyhbrz@gmail.com, qgchen@ir.hit.edu.cn, lbqin@csu.edu.cn

Abstract

Cross-lingual chain-of-thought can effectively complete reasoning tasks across languages, which gains increasing attention. Recently, dominant approaches in the literature improve cross-lingual alignment capabilities by integrating reasoning knowledge from different languages. Despite achieving excellent performance, current methods still have two main challenges: (1) *Manual language specification*: They still highly rely on manually selecting the languages to integrate, severely affecting their generalizability; (2) *Static weight allocation*: Current methods simply integrate all languages equally. In fact, different language reasoning paths should have different weights to achieve better complementation and integration. Motivated by this, we introduce an Automatic Cross-lingual Alignment Planning (AUTOCAP) for zero-shot chain-of-thought to address the above challenges. The core of AUTOCAP consists of two components: (1) *Automatic Language Selection Prompting* to guide LLMs to select appropriate languages and (2) *Automatic Weight Allocation Prompting* to automatically allocate alignment weight scores to each reasoning path. Extensive experiments on several benchmarks reveal that AUTOCAP achieves state-of-the-art performance, surpassing previous methods that required manual effort.

1 Introduction

Large language models (LLMs) have achieved substantial advancements (Brown et al., 2020; Chen et al., 2021; Qin et al., 2024a). Particularly noteworthy is the emergence of the Chain-of-Thought (CoT), which has further enhanced the ability of LLMs to handle complex reasoning tasks (Wei et al., 2022; Wang et al., 2022). In addition, as globalization continues to advance, aligning representations across different languages has become an urgent issue (Pires et al., 2019; Mulcaire et al., 2019; Qin et al., 2024b). This has motivated researchers to explore Cross-lingual CoT, aiming to break down language barriers by integrating CoT from different languages (Qin et al., 2023; Chai et al., 2024).



(a) Traditional Cross-lingual Alignment Self-consistent Framework. (b) Automatic Cross-Lingual Alignment Planning (AutoCAP) framework.

Figure 1: Traditional Cross-lingual Self-consistent framework (a) vs. Automatic Cross-lingual Alignment Planner framework (b). Previous approaches require manually specifying the aligned languages (English, Spanish, and German), and assigning the same weights to these languages. In contrast, our framework (AUTOCAP) uses the *Automatic Language Selection* and *Automatic Weight Allocation* to automatically select the most appropriate languages and weights.

Specifically, Shi et al. (2022) introduce a multilingual dataset for mathematical reasoning and propose a method requiring LLMs to use English for CoT prediction and problem-solving. Qin et al. (2023) develop a two-stage framework to aid LLMs in understanding problems across languages through manual language specification and reasoning. Ranaldi and Zanzotto (2023) present a cross-lingual, multi-step reasoning approach with a self-consistent prompting mechanism to enhance reasoning in multiple languages. Huang et al. (2023) introduce cross-lingual thought prompting using a generic template to enhance the reasoning capa-

*Equal Contribution.
† Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 9191–9200

bilities and performance of LLMs on multilingual tasks. In addition, recent cross-lingual CoT research has significantly improved complex reasoning across languages by aligning representations and integrating diverse linguistic reasoning paths (Shi et al., 2022; Tanwar et al., 2023; Qin et al., 2023).

Despite their improved performance, as shown in Figure 1 (a), the current approaches still face two key challenges:

(1) *Manual language specification*: The process continues to depend heavily on the manual selection of languages for integration, which not only expends substantial human effort but also significantly hurts its generalizability;

(2) *Static weight allocation*: The current methods simply integrate all languages equally, leading to sub-optimal performance. Actually, to achieve better integration of knowledge across languages, different language reasoning paths should possess different weights relative to the query.

In this paper, we introduce an automatic cross-lingual alignment planning (AUTOCAP) framework to address the above challenges. Specifically, as shown in Figure 1 (b), AUTOCAP comprises two key modules: (1) *Automatic Language Selection Prompting* and (2) *Automatic Weight Allocation Prompting*. In more detail, *Automatic Language Selection Prompting* is applied to enable LLMs to automatically select the most accurately aligned languages for reasoning for each query. After selecting the languages, *Automatic Weight Allocation Prompting* is used for automatically allocating an alignment weight score to each language reasoning path. Finally, accurate reasoning answers can be obtained by integrating the CoT reasoning paths across different languages and their corresponding weight scores.

Experimental results on several benchmarks show that AUTOCAP achieves superior performance compared to previous baselines, even surpassing previous manually selected language methods. In addition, extensive analysis demonstrates the strong generalization ability of AUTOCAP.

In summary, our key contributions include:

- We introduce Automatic Cross-lingual Alignment Planning (AUTOCAP), which greatly alleviates the burden of manually selecting languages and weights.
- The core of AUTOCAP comprises *Automatic Language Selection Prompting* and *Automatic Weight Allocation Prompting*, which achieves to automatically select the most appropriate languages and weights for cross-lingual CoT.
- Extensive experiments on several benchmarks demonstrate that AUTOCAP surpassed the previous approaches, achieving state-of-the-art performance and exhibiting strong generalizability.

All the code will be publicly available at https://github.com/BRZ911/AutoCAP.

2 Preliminaries

This section outlines the preliminaries of mono-lingual chain-of-thought (§2.1) and cross-lingual chain-of-thought (§2.2).

2.1 Mono-lingual Chain-of-Thought

Mono-lingual Chain-of-Thought (Wei et al., 2022) triggers LLMs to directly reason step-by-step in the source language to solve tasks. Formally, when presented with a query $\mathcal{Q}$ expressed in the source language L_{src}, the LLMs generate a reasoning path, which can be simplified and represented as follows:

$$\mathcal{R}_{src} = \text{argmax}\ P(\mathcal{R}|\mathcal{Q}, L_{src}), \quad (1)$$

where $\mathcal{R}_{src}$ denotes the generated reasoning path with multiple steps in the source language L_{src}. Following this, the LLMs produce the final results $\mathcal{Y}$, which are obtained by:

$$\mathcal{Y} = \text{argmax}\ P(y|\mathcal{Q}, L_{src}, \mathcal{R}_{src}). \quad (2)$$

2.2 Cross-lingual Chain-of-Thought

For better cross-lingual generalization for CoT in multilingual scenarios, Qin et al. (2023) propose Cross-lingual Chain-of-Thought methods to align multilingual representations explicitly. Formally, given a query $\mathcal{Q}$ in source language L_{src}, experts *manually* select a target language L_{tgt} to serve as an anchor for cross-lingual alignment. The LLMs then generate an alignment $\mathcal{A}$ as follows:

$$\mathcal{A} = \text{argmax}\ P(a|\mathcal{Q}, L_{src} \rightarrow L_{tgt}). \quad (3)$$

Subsequently, the LLMs produce the result $\mathcal{R}$ by:

$$\mathcal{R}_{tgt} = \text{argmax}\ P(\mathcal{R}|\mathcal{A}, L_{tgt}). \quad (4)$$

Finally, the LLMs determine the final results $\mathcal{Y}$ based on the reasoning path $\mathcal{R}_{tgt}$ in L_{tgt} and generated alignment $\mathcal{A}$:

$$\mathcal{Y} = \text{argmax}\ P(y|\mathcal{A}, L_{tgt}, \mathcal{R}_{tgt}). \quad (5)$$

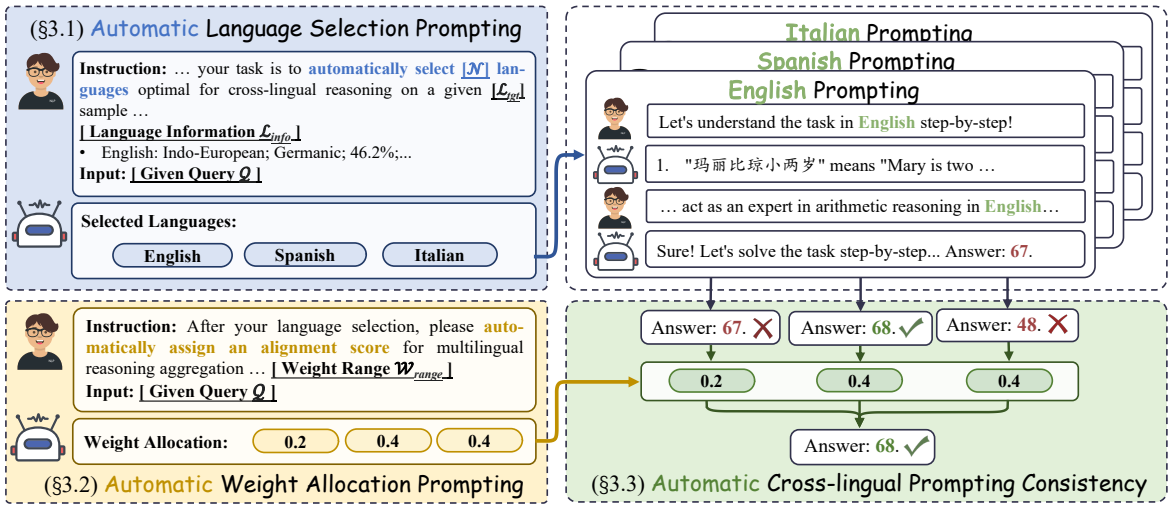


Figure 2: The overall workflow of AUTOCAP, which consist of *Automatic Language Selection Prompting* and *Automatic Weight Allocation Prompting*.

3 Automatic Cross-lingual Alignment Planning

In this section, we introduce an Automatic Cross-lingual Alignment Planning (AUTOCAP) framework to automatically select the most appropriate languages and weights for cross-lingual CoT, which consists of two main components: (1) *Automatic Language Selection Prompting* and (2) *Automatic Weight Allocation Prompting*.

3.1 Automatic Language Selection Prompting

To address the significant challenge of manual language selection, as shown in Figure 2, we propose *Automatic Language Selection Prompting* (ALSP) to autonomously and intelligently choose the most suitable languages and further utilize the cross-lingual capabilities of LLMs. Specifically, the prompt content is as follows:

> **Instruction:** ... your task is to select [Selected Number $\mathcal{N}$] languages optimal for cross-lingual reasoning on a given [Source Language $\mathcal{L}_{src}$] sample...
> [Language Information $\mathcal{L}_{info}$]
> **Input:** [Given Query $\mathcal{Q}$]

Specifically, ALSP directs LLMs to select the [Selected Number $\mathcal{N}$] of languages by analyzing the [Given Query $\mathcal{Q}$], respective [Source Language $\mathcal{L}_{src}$], and a comprehensive list of potential target language information [Language Information $\mathcal{L}_{info}$]. The language selection process can be formulated as:

$$\mathcal{L}'_{tgt} = \underset{\mathcal{L}}{\operatorname{argmax}} \sum_{i=1}^{\mathcal{N}} P(\mathcal{L}^i_{tgt} | \mathcal{Q}, \mathcal{L}_{src}, \mathcal{L}^i_{info}), \quad (6)$$

where $\mathcal{L}'_{tgt} = \{\mathcal{L}^{i'}_{tgt}\}_{i=1}^{\mathcal{N}}$ represents the final set of chosen target languages, and $\mathcal{L}^i_{info} \in \mathcal{L}_{info}$ encompasses respective language family, language branch, the proportion of available pre-training data to facilitate informed decision-making.

3.2 Automatic Weight Allocation Prompting

After selecting the target language, as shown in Figure 2, we further introduce *Automatic Weight Allocation Prompting* (AWAP). Specifically, our carefully designed prompts for guiding LLMs to automatically allocate weight to each language inference path are as follows:

> **Instruction:** After your language selection, please assign an alignment score for multilingual reasoning aggregation...
> [Weight Range $\mathcal{W}_{range}$]
> **Input:** [Given Query $\mathcal{Q}$]

Specifically, this process dynamically allocates weights from [Weight Range $\mathcal{W}_{range}$] to languages based on their relevance to the [Given Query $\mathcal{Q}$], enhancing performance of LLMs by aligning the [Source Language $\mathcal{L}_{src}$] to target language $\mathcal{L}^{i'}_{tgt}$ generated from last turn

Model	bn	de	es	fr	ja	ru	sw	te	th	zh	AVG
GPT-3† (Brown et al., 2020)											
`Direct` (Shi et al., 2022)	4.4	14.8	17.2	16.8	11.2	12.4	8.8	0.8	8.8	18.0	11.3
`Native-CoT` (Shi et al., 2022)	6.4	36.0	40.4	37.6	26.0	28.4	11.2	0.4	10.8	40.0	23.7
`En-CoT` (Shi et al., 2022)	9.6	44.0	44.8	46.0	32.4	28.4	20.8	5.6	19.6	40.8	29.2
`Translate-En` (Shi et al., 2022)	41.2	46.4	51.6	46.4	44.8	48.8	37.6	42.8	41.2	47.2	44.8
PaLM† (Chowdhery et al., 2023)											
`Direct` (Shi et al., 2022)	17.2	18.8	20.0	19.6	16.0	22.0	15.6	17.6	16.8	19.2	18.3
`Native-CoT` (Shi et al., 2022)	46.0	49.2	56.8	46.4	40.0	48.4	35.2	45.6	52.8	46.8	48.7
`En-CoT` (Shi et al., 2022)	46.4	53.6	58.0	51.2	49.6	55.6	44.4	46.8	49.6	46.0	50.1
`Translate-En` (Shi et al., 2022)	53.2	57.2	60.0	55.2	50.0	59.6	51.2	49.6	50.8	55.6	54.2
GPT-3.5 (OpenAI, 2022)											
`Direct`‡ (Qin et al., 2023)	33.6	56.0	61.2	62.0	52.8	62.0	48.0	7.6	42.4	60.0	48.6
`Native-CoT`‡ (Qin et al., 2023)	26.4	70.0	70.4	64.4	52.8	62.4	54.0	10.4	40.0	59.6	51.0
`En-CoT`‡ (Qin et al., 2023)	50.0	73.6	69.6	70.0	60.4	65.6	55.2	22.0	48.0	63.2	57.8
`Translate-En`‡ (Qin et al., 2023)	66.4	75.6	74.4	72.4	66.0	72.8	69.6	**58.0**	57.6	71.6	68.4
`CLP`‡ (Qin et al., 2023)	64.8	80.0	82.4	79.2	69.2	81.6	74.8	38.8	62.0	73.6	70.6
`CLSP`‡ (Qin et al., 2023)	72.4	86.0	84.0	82.0	76.4	86.8	76.8	50.0	65.2	75.2	75.5
`Cross-ToT`‡ (Ranaldi and Zanzotto, 2023)	-	87.6	86.2	84.3	-	86.5	75.4	-	-	83.5	-
AUTOCAP	**76.0**	**88.0**	**86.8**	**84.4**	**79.6**	**88.0**	**78.4**	52.0	**69.2**	**84.0**	**78.6**

Table 1: Accuracy (%) on MGSM. "`Direct`" prompt refers to directly asking and answering in the original language. "`Native-CoT`" prompt denotes answering with CoT in the native language. "`En-CoT`" prompt refers to mandating the use of CoT in English. "`Translate-En`" prompt signifies translating the query into English and then responding in English. The result with † represents 6-shot sample prompt sourced from Shi et al. (2022). The result with ‡ indicates that it comes from Qin et al. (2023) and Ranaldi and Zanzotto (2023). For a fair comparison, for the integration method, we used 6 languages for integration.

more effectively. Formally, the automatic weight for each language can be obtained by:

$$\mathcal{W}_i^{'} = \underset{w \in \mathcal{W}_{range}}{\operatorname{argmax}} \, p(w|\mathcal{Q}, \mathcal{L}_{src} \rightarrow \mathcal{L}_{tgt}^{i'}, \mathcal{L}_{info}^{i}), \quad (7)$$

where $\mathcal{W}_i^{'}$ represents the cross-lingual alignment weight for the i-th target language.

3.3 Automatic Cross-lingual Prompting Consistency

By automatically determining the relevant language and its associated weight, our framework further adapts *Automatic Cross-lingual Prompting Consistency* to more effectively merge multilingual alignments, leading to improved consistency across languages. Following Equation 4, we collect a set of generated results $\mathbf{R}$. The formulation of the final integrated result $\hat{\mathcal{R}}$ is presented as follows:

$$\hat{\mathcal{R}} = \underset{\mathcal{R} \in \mathbf{R}}{\operatorname{argmax}} \sum_{i=1}^{\mathcal{N}} \sum_{r \in \mathbf{R}} \mathcal{W}_i^{'} \cdot \mathbb{1}(\mathcal{R} = r), \quad (8)$$

where $\mathcal{R}$ and r both denote a reasoning outcome generated based on a specific formula in the target language $\mathcal{L}_{tgt}^{i'}$ from the generated result set $\mathbf{R}$, and $\mathcal{W}_i^{'}$ represents the weight assigned to that language. Additionally, $\mathbb{1}(X)$ is the indicator function, which returns 1 if X is true and 0 if it is false.

4 Experiments

4.1 Dataset and Baseline

Following Wei et al. (2022); Qin et al. (2023), we assess the performance of AUTOCAP on the widely utilized multilingual mathematical reasoning dataset MGSM (Shi et al., 2022) and select three representative LLMs as backbones, namely PaLM (Chowdhery et al., 2023), GPT3 (Brown et al., 2020), and GPT-3.5 (OpenAI, 2022). The top-p and temperature parameters in all processes are selected within the range of [0, 1].

4.2 Main Results

We follow previous work (Wei et al., 2022; Qin et al., 2023) to adapt accuracy (Acc.) as the metric to evaluate the performance. The main results are shown in Table 1.

From the results, we observe that AUTOCAP attains superior performance compared to all baseline models by achieving state-of-the-art performance, even surpassing the ensemble methods `CLSP` and

Model	bn	de	es	fr	ja	ru	sw	te	th	zh	AVG
AUTOCAP	76.0	88.0	86.8	84.4	79.6	88.0	78.4	52.0	69.2	84.0	78.6
w/o AWAP	68.8	84.8	86.8	82.0	79.2	88.4	78.8	49.6	66.0	80.0	76.4 *(-2.2)*
w/o ALSP	60.0	80.4	84.8	81.2	73.6	85.6	64.8	45.2	66.8	78.0	72.0 *(-6.6)*
w/o AWAP & ALSP	58.0	79.6	84.8	80.8	70.4	84.0	64.4	44.0	66.8	78.0	71.1 *(-7.5)*

Table 2: Ablation experiment on GPT3.5. "*w/o ALWP*" refers to removing *Automatic Language Selection Prompting* (ALWP). "*w/o ALWP*" refers to removing *Automatic Weight Allocation Prompting* (ALSP). "*w/o ALWP & ALSP*" refers to removing both ALWP and ALSP.

`Cross-ToT`, which manually select reasoning languages. Specifically, AUTOCAP achieves an average accuracy improvement of over 3.1%, outperforming `CLSP` across all tested languages. This demonstrates that AUTOCAP, while implementing the automatic selection of reasoning languages, can better elicit the cross-lingual CoT reasoning capabilities of LLMs.

4.3 Analysis

To gain deeper insights into our approach, we explored the following research questions:
(1) *Are all modules effective for* AUTOCAP*?*
(2) *Can the interactive capabilities of LLMs enhance the performance of* AUTOCAP*?*
(3) *Can* AUTOCAP *work well on fewer languages?*
(4) *Why can* AUTOCAP *works?*
(5) *Can* AUTOCAP *generalize to other open-source models?*
(6) *Can* AUTOCAP *generalize to other benchmarks?*
(7) *What is the intuition behind* AUTOCAP*?*

4.3.1 Answer 1: All Modules in AUTOCAP are Effective for AUTOCAP

In this section, we explore whether *Automatic Language Selection Prompting* and *Automatic Weight Allocation Prompting* are effective.

Automatic Language Selection Prompting is effective. To analyze the effectiveness for *Automatic Language Selection Prompting* (ALSP), we removed the ALSP and randomly selected six languages for all query data instead. As indicated in Table 2 (*w/o ALSP*), there is a significant decline in reasoning performance across all languages. In particular, there was an average accuracy reduction of 6.6%. It indicates that ALSP can effectively select the more optimal languages for alignment, which significantly improves the process of bridging linguistic gaps, directly contributing to a notable enhancement in the performance of multilingual CoT.

Automatic Weight Allocation Prompting is also crucial for performance enhancement. To verify the impact of excluding the *Automatic Weight Allocation Prompting* (AWAP) from our AUTOCAP. By default, we adopt the setting of Qin et al. (2023) and set the weight of all languages to 1 by default. As presented in Table 2 (*w/o AWAP*), it demonstrates a notable reduction in cross-lingual CoT performance, with an average decrease of 2.2%. The absence of specific weightings for each language resulted in lesser coherence when merging the outputs from multilingual reasoning, adversely affecting the reasoning process of overall effectiveness. This performance decline highlights the critical role of the AWAP. Specifically, AWAP accurately allocates weights and facilitates detailed planning to enhance the degree of cross-lingual alignment, thereby improving the precision of multilingual reasoning.

The combination of AWAP and ALSP brings a higher improvement. To validate the functionality and effectiveness of the combination of the *Automatic Weight Allocation Prompting* and *Automatic Language Selection Prompting*, we conducted an experiment where both modules were simultaneously removed. As shown in Table 2 (*w/o AWAP & ALSP*), we observe a significant decline in performance compared to the AUTOCAP, with a decrease of 7.5%. This decline was also evident when compared to the individual ablations of the ALSP and AWAP modules. The absence of both modules resulted in a substantial decrease in accuracy in multilingual reasoning, underscoring the importance of language selection and weighting for better cross-lingual reasoning.

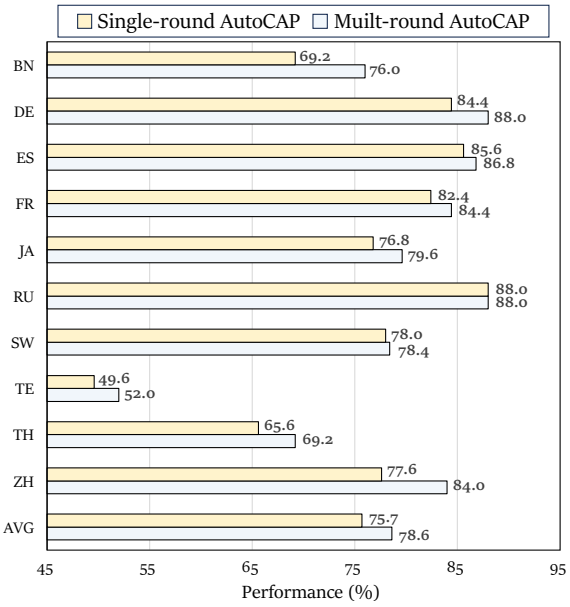


Figure 3: The accuracy of single-round AUTOCAP compared to multi-round AUTOCAP. AVG stands for average performance score.

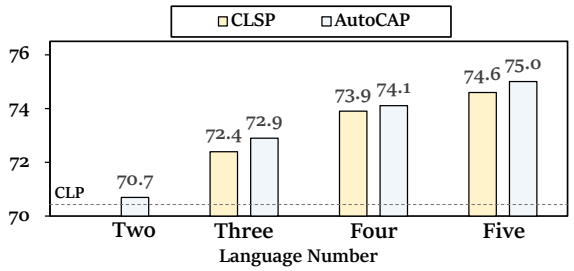


Figure 4: Performance comparison results of CLSP (Qin et al., 2023) and AUTOCAP on different languages.

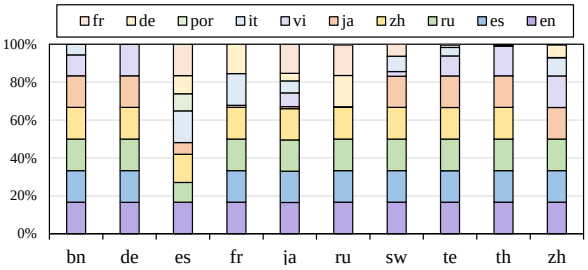


Figure 5: Languages for cross-lingual reasoning and their proportions.

4.3.2 Answer 2: The interactive feature of LLMs boosts AUTOCAP performance.

To investigate the influence of the interactive capabilities of LLMs on AUTOCAP, we differentiate between single-round prompting and multi-round prompting. Specifically, in the single-round prompting approach, we instruct LLMs to simultaneously select reasoning languages and allocate weights. Conversely, in the multi-round prompting approach (AUTOCAP), in the first round, LLMs try to select the language, and then LLMs are required to allocate weight in the second round.

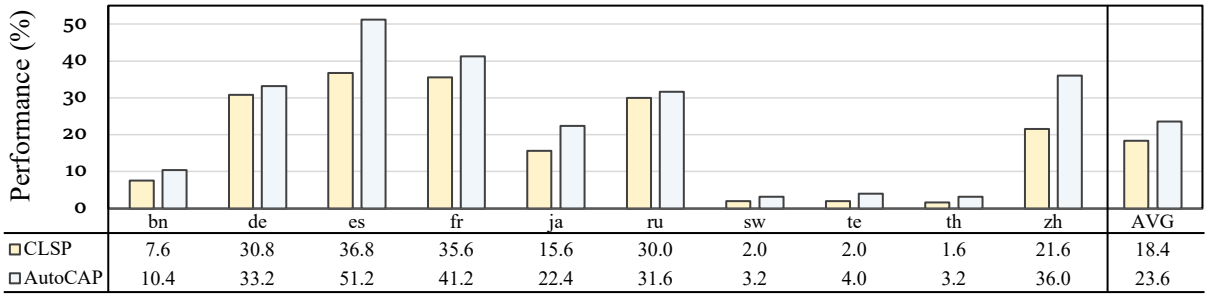
As illustrated in Figure 3, the average performance of single-round interactions exhibited a decrease of 2.9%, which indicates that leveraging the interactive capabilities of LLMs can significantly enhance the performance of cross-lingual CoT and the capability of language planning.

4.3.3 Answer 3: AUTOCAP also achieves positive results on fewer languages

To showcase the efficacy of our method using a limited number of languages, with a varied number of languages, specifically ranging from three to five. As shown in Figure 4, a positive trend can be observed: as the number of languages incorporated increases, there is a corresponding enhancement in model performance. This pattern underscores the scalability and robustness of our approach in processing multilingual inputs. Further, to explore the effectiveness of AUTOCAP, we compared our results with the state-of-the-art method, CLSP (Qin et al., 2023), which uses manually selected languages and a voting mechanism to merge answers from multiple languages. In detail, as shown in Figure 4, when evaluated across linguistic settings encompassing three, four, and five languages, our model consistently outperformed the CLSP framework, registering an average performance uplift of 0.4%. Such findings demonstrate the efficiency of our approach across varying language counts, which highlights the potential of our method in leveraging cross-lingual planning automatically to improve cross-lingual CoT capabilities.

4.3.4 Answer 4: AUTOCAP selects more diverse languages for better result

In investigating the capacity of AUTOCAP to encompass a broader linguistic spectrum, this section counts the variety and distribution of reasoning languages selected by AUTOCAP. The statistical results are shown in the Figure 5, in its reasoning processes, AUTOCAP incorporated a minimum of seven and a maximum of ten distinct languages. This demonstrates a significant enhancement in linguistic diversity when contrasted with the conventional approach of static language selection, underscoring the superior adaptability and breadth of languages facilitated by the autonomous choices made by LLMs.



	bn	de	es	fr	ja	ru	sw	te	th	zh	AVG
CLSP	7.6	30.8	36.8	35.6	15.6	30.0	2.0	2.0	1.6	21.6	18.4
AutoCAP	10.4	33.2	51.2	41.2	22.4	31.6	3.2	4.0	3.2	36.0	23.6

Figure 6: Comparison of the performance between AUTOCAP and CLSP on the open-source LLM Mistral (Jiang et al., 2023).

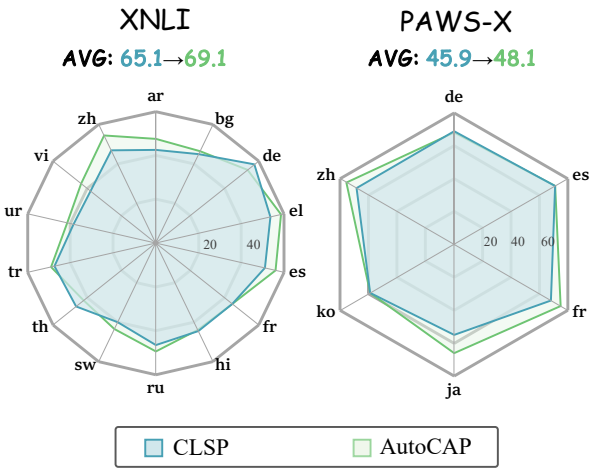


Figure 7: Accuracy on the XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019) datasets.

4.3.5 Answer 5: AUTOCAP also works on other LLM

To further validate the scalability and universality of AUTOCAP, we adapt AUTOCAP on other open-source LLMs. The experimental outcomes on LLM Mistral (Jiang et al., 2023) are depicted in Figure 6, demonstrating the optimal capability of AUTOCAP on open-source LLMs. With an average performance improvement of at least 5.2% over CLSP, these results further attest to the broad applicability of AUTOCAP.

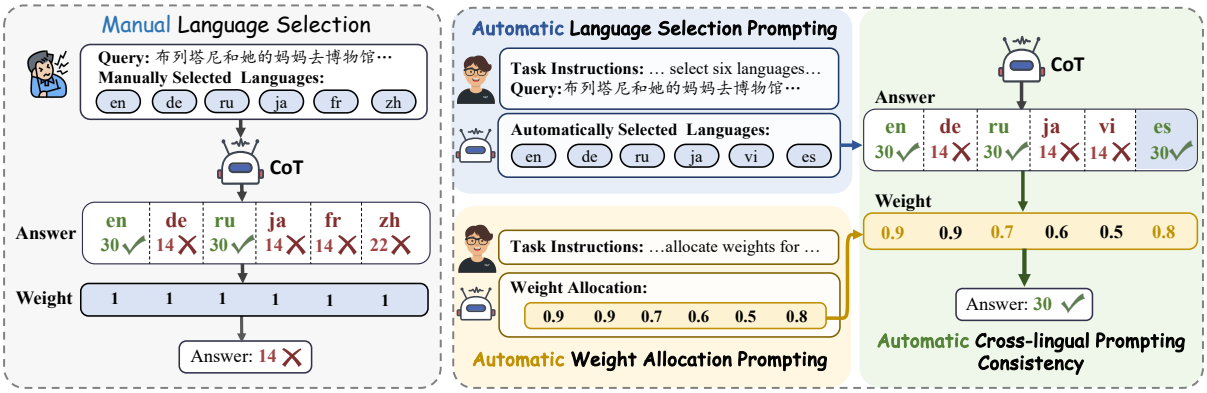
4.3.6 Answer 6: AUTOCAP exhibits strong generalization on other benchmarks.

To further explore the effectiveness of AUTOCAP on other tasks, following Qin et al. (2023), we conducted experiments on two multilingual datasets, XNLI (Conneau et al., 2018) and PAWS-X (Yang et al., 2019). The results, as shown in Figure 7, indicate that AUTOCAP achieved better performance than CLSP on both datasets, with an average improvement of 4.0% on XNLI and 2.2% on PAWS-X. And it surpasses the performance of all languages compared with CLSP. These effectively illustrate the generalization of AUTOCAP on different cross-lingual COT tasks.

4.3.7 Answer 7: Qualitative analysis

To provide a more intuitive understanding of our method, we present a distinct case for qualitative analysis in this section. As shown in Figure 8 (a), CLSP conducts reasoning in six manually selected languages. The reasoning results are correct in English (en) and Russian (ru), but incorrect in German (de), Japanese (ja), French (fr), and Chinese (zh). CLSP treats each reasoning path equally, integrating different paths solely through a voting mechanism, which unfortunately led to the incorrect answer "14".

Conversely, as depicted in Figure 8 (b), AUTOCAP automatically selects six reasoning languages based on the query during the first interaction round. Although reasoning in German (de), Japanese (ja), and Vietnamese (vi) lead to incorrect answers, English (en), Russian (ru), and Spanish (es) produce correct ones. In the second round, it assigns respective weight scores to each language. By aggregating these weighted scores, our method successfully circumvents the incorrect reasoning, ultimately selecting the correct answer "30". These cases demonstrate the effectiveness and intuitiveness of our method. Specifically, our AUTOCAP is capable of performing automatic cross-lingual planning on both languages and respective weights. Such advanced planning effectively decreases the cross-lingual alignment difficulties.



(a) Cross-lingual Self-consistent Prompting (CLSP) (b) Automatic Cross-lingual Alignment Planning (AutoCAP)

Figure 8: Case Study. Figure (a) illustrates the CLSP process, which integrates multiple reasoning answers through the manual selection of reasoning languages and a voting mechanism. Due to the predominance of the incorrect answer "14" in the results, this incorrect answer was selected. In contrast, Figure (b) demonstrates the ability of AUTOCAP to automatically select aligned reasoning languages, with the correct inference answer "30" achieving a higher proportion than CLSP. By combining automatically allocated weights, the correct answer was selected.

5 Related Work

The evolution of LLMs has profoundly propelled the progress in the field of artificial intelligence (Brown et al., 2020; Schaeffer et al., 2024; Yao et al., 2024; Touvron et al., 2023; Tang et al., 2023). In particular, CoT has provided a new perspective for solving complex problems and enhancing the reasoning capabilities of models by directing LLMs through a step-by-step problem-solving process.(Wei et al., 2022; Kojima et al., 2022; Feng et al., 2024; Zhang et al., 2023, 2022).

Current research primarily focuses on English, yet given the existence of over 7,000 languages worldwide, addressing critical challenges like reasoning and generation in minority languages has become an urgent necessity (Qin et al., 2023; Shi et al., 2022; Lin et al., 2021; Huang et al., 2023). Recognizing this gap, recent research has increasingly focused on transcending vanilla CoT to explore its cross-lingual dimensions. In this vein, Shi et al. (2022) pioneer the introduction of the multilingual dataset specifically curated for mathematical reasoning. They advocate for a novel cross-lingual methodology that mandates LLMs to employ English in forecasting CoT sequences, translating the problem into English, and subsequently resolving it via English-based CoT paradigms. Ranaldi and Zanzotto (2023) unveil a self-consistent prompting mechanism within their cross-lingual, multi-step reasoning strategy, significantly augmenting reasoning capabilities across various languages. Tanwar et al. (2023) suggest the integration of exemplars showcasing semantic congruence between source and target languages within the prompt context, aiming to facilitate seamless reasoning transitions across linguistic divides. Moreover, Qin et al. (2023) propose a self-consistent prompting, initially involving the manual selection of the reasoning language, followed by employing a voting mechanism to determine the final reasoning answer. This approach has yielded excellent results in the efficacy of cross-lingual prompting.

In comparison to the previous research, our work focuses on two key aspects. Firstly, we introduce an automatic language selection mechanism, enabling our system to choose the most accurately aligned reasoning languages for each query. Additionally, we develop an automatic weight allocation that effectively integrates the answers provided by various reasoning paths. To our knowledge, this is the first work to automatically select the reasoning languages and assign weights to reasoning paths.

6 Conclusion

In this paper, we present the Automatic Cross-lingual Alignment Planner (AUTOCAP), a novel framework designed for enhancing zero-shot cross-lingual CoT reasoning. AUTOCAP is comprised of two key components: *Automatic Language Selection Prompting* and *Automatic Weight Allocation Prompting*. AUTOCAP achieves to automatically select suitable languages and allocate weights to various reasoning paths in cross-lingual CoT, respectively. Extensive experiments demonstrate that AUTOCAP achieves superior performance, outperforming existing methods in cross-lingual CoT.

7 Limitations

This work achieves automatic language selection and weight allocation for cross-lingual CoT, reducing manual workload significantly and marking a meaningful first step towards automatic resolution of multilingual tasks. In the future, we can apply our framework to multi-agent systems, enabling agents to incorporate automatic selection of reasoning languages and powerful tools to tackle challenging real-world problems. In addition, we can consider the multi-lingual multi-modal CoT scenario by injecting the cross-modal CoT ability (Chen et al., 2024).

8 Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) via grants 62306342, 62236004 and 62206078. This work was also sponsored by the Excellent Young Scientists Fund in Hunan Province (2024JJ4070). We are grateful for resources from the High Performance Computing Center of Central South University.

References

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *ArXiv preprint*, abs/2401.07037.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *ArXiv preprint*, abs/2405.16473.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2024. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *ArXiv preprint*, abs/2112.10668.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3912–3918, Minneapolis, Minnesota. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024a. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

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 *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024b. Multilingual large language model: A survey of resources, taxonomy and frontiers. *ArXiv preprint*, abs/2404.04925.

Leonardo Ranaldi and Fabio Massimo Zanzotto. 2023. Empowering multi-step reasoning across languages via tree-of-thoughts. *ArXiv preprint*, abs/2311.08097.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *ArXiv preprint*, abs/2210.03057.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *ArXiv preprint*, abs/2305.14825.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv preprint*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *ArXiv preprint*, abs/2210.03493.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *ArXiv preprint*, abs/2302.00923.