Multidimensional Consistency Improves Reasoning in Language Models

Anonymous ACL submission

Abstract

While Large language models (LLMs) have proved able to address some complex reasoning tasks, we also know that they are highly sensitive to input variation, which can lead to different solution paths and final answers. Answer consistency across input variations can thus be taken as a sign of stronger confidence. Leveraging this aspect, we introduce a framework where models are *systematically pushed to diversify solution paths* towards a final answer, thereby testing them for answer consistency across multiple input variations. Focusing on math problems, we induce variations in (i) order of shots in prompt, (ii) problem phrasing, and (iii) languages used. Experiments on a wide range of open-source LLMs of various sizes show that reasoning consistency differs by variation dimension, and that by aggregating consistency across dimensions, our framework enhances mathematical reasoning performance on monolingual datasets GSM8K and MATH500, and the multilingual dataset MGSM. We also discuss the different efficacy of the types of variations induced and how they could be further leveraged in future work.

1 Introduction

Large Language Models (LLMs) have shown impressive abilities in addressing a variety of complex reasoning tasks, such as math reasoning (Brown et al., 2020) and commonsense reasoning (Bommasani et al., 2022). The use of Chain-of-Thought (CoT), i.e., breaking down a problem and taking multiple intermediate steps to gradually arrive at the final answer, endows LLMs with even better performances (Wei et al., 2022).

At the same time, LLMs have also proved to be sensitive and somewhat brittle with respect to variations in the way they are prompted (Zhao et al., 2021; Lu et al., 2022). For instance, in a few-shot setting for solving mathematical problems, just altering the order in which the example shots are

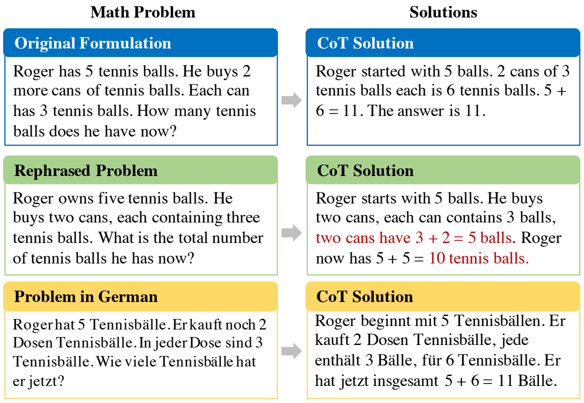

Figure 1: Example of variations: A math problem is presented in different forms or languages, resulting in different reasoning paths to solve it.

provided might lead to different reasoning paths and possibly different answers (Wang et al., 2022); the same can happen if different formulations of the same problem are used (Zhou et al., 2024). Likewise, an identical mathematical problem presented once in one language, and once in a different one, may be solved following different strategies and also lead to different answers (Lai and Nissim, 2024). Figure 1 visualises examples of such variations across different dimensions.

Some of these variations, such as using even slight alterations in the prompt (Wang et al., 2022; Li et al., 2023), have been exploited in recent work to enhance reasoning performance. However, the experimental setup and the assessment of *(in)consistent answers due to variations* is still scattered. In this paper, we argue for a systematic treatment of variations and answer consistency and introduce a *Multidimensional Reasoning Consistency* (MRC) framework, focusing on maths problems. MRC, shown in Figure 2, allows for a systematic and comprehensive testing and evaluation of model consistency against variations in the way the problem is presented to the model. Our frame-

work also makes it possible to best leverage such variations and answer consistency for improving overall accuracy in mathematical reasoning tasks.

The rationale behind this framework is that by explicitly and systematically pushing the model to likely diversify its solution paths, and possibly yield a different final answer, we can take across-variation consistency of the answer as stronger evidence for its correctness.

We consider three dimensions of variation to test consistency: (i) order of shots in prompt; (ii) problem (re)phrasing; and (iii) natural languages used. We follow Wang et al. (2022) in changing the order of the exemplars (i.e., the shots), which results in different prompts based on a set of example problems. For problem rephrasing, we prompt the LLMs to rewrite the question before solving it. Lastly, we use the same math problems written in different languages. For each dimension, the LLM generates multiple solution paths to a question, which could differ in various ways, but should in principle lead to the same answer. Answer consistency is eventually used to determine the final answer to the given problem.

We evaluate MRC on three mathematical reasoning benchmarks: GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021), and MGSM (Shi et al., 2023a), covering a range of open-source LLMs with varying scales: 7-8B, 14-32B, and 70-72B. This allows us to systematically examine how model scale affects performance on mathematical reasoning tasks and to compare the effectiveness of our framework across different dimensions of variation.

Contributions First, we introduce a method to systematically study LLMs' reasoning consistency along multiple dimensions of input variation. Second, we improve model performance on both monolingual and multilingual datasets for a variety of models by leveraging reasoning consistency across variations; this is obtained thanks to the induced substantial diversification of the reasoning paths, offering valuable insights into LLMs reasoning beyond the commonly used sampling-based strategy. Third, extensive experimental results show that model consistency differs by variation dimensions, but exploiting consistency always enhances math reasoning performance, and aggregating consistency across dimensions yields an additional boost; this paves the way for using a similar framework for other (reasoning) tasks, providing a strategy to make models more robust reasoners. All data and code are available (upon acceptance.)

2 Related Work

Math Reasoning in LLMs Mathematical reasoning has garnered great interest in recent times since LLMs have shown what look like complex problem-solving capabilities (Brown et al., 2020; Lu et al., 2023). With LLMs and few-shot prompting, only a few task examples (e.g., question-answer pair) are required at inference time to enable the LLM to perform the intended task without updating the model parameters (Brown et al., 2020). To further elicit LLMs' reasoning capability, Wei et al. (2022) proposed a Chain-of-Thought prompting, which involves an explicit step-by-step reasoning from the question to the answer, rendered in natural language. Given its success, a series of CoT-related methods have been proposed to improve reasoning performance in LLMs, such as complex CoT (Fu et al., 2023), auto-CoT (Zhang et al., 2023), multilingual CoT (Shi et al., 2023b), least-to-most prompting (Zhou et al., 2023), progressive-hint prompting (Chuanyang et al., 2023), and residual connection prompting (Jiang et al., 2024). Rather than developing a new specific CoT method, we introduce variations in the prompt and exploit the diversity of CoT outputs.

Consistency in LLMs In principle, language models could be expected to yield consistent answers in semantically equivalent contexts, especially regarding factual information; this is considered a crucial aspect in assessing model generalization abilities (Fierro and Søgaard, 2022; Lai and Nissim, 2024). In practice, this is often not the case. Some works have thus focused on improving consistency on, e.g., natural language inference (Mitchell et al., 2022), explanation generation (Camburu et al., 2020), cloze test (Ravichander et al., 2020), and factual knowledge extraction (Fierro and Søgaard, 2022). For improving CoT reasoning, Wang et al. (2023) suggested to use self-consistency, sampling diverse solution paths and then selecting the most consistent answer. Zhou et al. (2024) proposed self-consistency-over-paraphrases (SCoP), which diversifies solution paths by generating different paraphrases for a given problem. To check consistency, Wang et al. (2022) use different exemplar orders to possibly trigger diverse solutions. Lai and Nissim (2024) look at consistency of answers given to the same

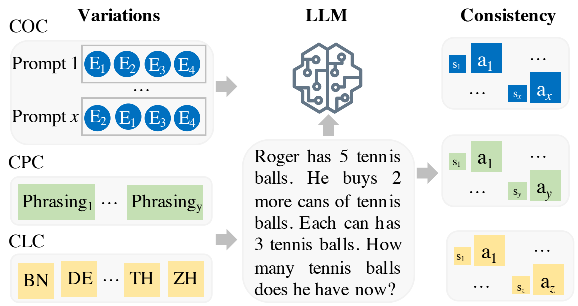

Figure 2: Overview of our Multidimensional Reasoning Consistency (MRC) framework: (i) COC changes the exemplars order; (ii) CPC rewrites the given questions in the same language; and (iii) CLC rewrites the given questions in different languages.

problem written in two different languages, and use multilingual instruction tuning to improve LLMs' performance across languages.

Here, we propose a systematic method to study and leverage reasoning consistency along different dimensions to improve overall performance.

3 Methodology

Figure 2 shows our MRC framework. Using systematic variations, MRC pushes the model to generate multiple solutions for a given question, then measures consistency across variations, and leverages it to improve model performance.

3.1 Reasoning Consistency

Formally, given a set of math problems $\mathcal{M}$, each consisting of a two-tuple $(\texttt{question}{:}q, \texttt{answer}{:}a)$. We define the reasoning consistency of an LLM as the extent to which it yields the same answer for a given question under a dimension of variation (e.g., language). Specifically, for each question, assume that the LLM generates n candidate solutions $\{s_1, \ldots, s_n\}$ which can arrive at a set of final answers $\{a_1, \ldots, a_m\}$, reasoning consistency (RC) is the ratio of the maximum number of these solutions that can lead to the same answer over the total number of candidates n.

$$\text{RC(LLM)} = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \frac{\max_j |\mathcal{S}_j|}{n} \quad (1)$$

$$\mathcal{S}_j = \{s_i \in \{s_1, \ldots, s_n\} | f(s_i) = a_j\} \quad (2)$$

Where $f(s_i)$ maps solution s_i to the final answer.

3.2 Multidimensional Consistency

In the context of reasoning consistency in mathematical problems, a language model can generate multiple plausible responses to the same math question, where *correct* reasoning solutions, even if they are diverse, tend to be more consistent in the final answer than incorrect solutions (Wang et al., 2023). Instead of simply sampling a diverse set of candidate outputs from LLMs, our MRC framework, aims to assess model consistency along three dimensions we control for and exploit: example order, problem (re)phrasing, and language.

Cross-order Consistency (COC) Some prior works have shown that LLMs are sensitive to order, such as the order of options in multiple-choice questions (Pezeshkpour and Hruschka, 2024; Zotos et al., 2025), or the order of shots in math reasoning (Wang et al., 2022). Here we assess how much the *order* of the shots affects consistency of language models. Specifically, we focus on few-shot prompting, which consists of a set of exemplars $(\texttt{question}{:}q, \texttt{step-by-step solution}{:}s)$, whose presentation order can be changed arbitrarily. For instance, given a 4-shot prompt with 4 exemplars, we could change their order to get 24 different prompts, each of which can be used to prompt the model to generate a corresponding answer to a given question (see Appendix A.1 for examples).

Cross-phrasing Consistency (CPC) This considers that the surface form of the question itself can also have an impact on the performance of the model. Differently from Zhou et al. (2024), who select 'good' paraphrases based on solving rate improvement as the few-shot exemplars, we directly prompt an LLM to rewrite the question with the goal of making it easier for itself to solve (see Appendix A.1 for examples). We use two main settings: rewrite-without-solve and rewrite-then-solve, which yield the following four different settings when combined with the original question:

- Rewrite-without-solve (RwS): We ask the LLM to rewrite the question, but not to include the solution. Afterwards, we prompt the LLM to generate the solution for the rewritten question.
- Original Question + RwS (RwS+): We concatenate the original question and the rewritten one above prompting the LLM for the solution.
- Rewrite-then-solve (RtS): We ask an LLM to rewrite the question making it easier to solve and then to give the corresponding solution.
- RtS Question (RtS-): We prompt the LLM to generate the solution for the rewritten question in the "rewrite-then-solve" setting.

Cross-lingual Consistency (CLC) One rather outstanding way to vary formulations is to write the same problem in different languages. Abilities of LLMs in different languages vary substantially, depending on the amount of training data in a given language, and on the similarity of lesser represented languages to more resource-rich ones, as this impacts how well models can deal with less seen languages (de Vries et al., 2022; Muennighoff et al., 2023; Üstün et al., 2024). With cross-lingual consistency, we leverage language diversity to evaluate the LLMs' robustness to input in different languages, and exploit output diversity to further improve the LLMs' reasoning performance. Given the same math question in different languages, LLMs are expected to produce reasoning solutions in the corresponding languages. On the one hand, those solutions are expected to arrive at the same final answer if the language model is multilingual; on the other hand, due to the differences in language structures, those solutions can increase diversity compared to using a single language.

3.3 MRC for Reasoning

Eventually, answer consistency across the three dimensions can also be leveraged to improve reasoning performance. For each question, the language model produces the corresponding solution set $\{s_1, \dots, s_n\}$ and the final answer set $\{a_1, \dots, a_m\}$. We select the most consistent answer in n solution paths as the final answer $\hat{a}$, which is obtained through majority voting:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \mathbb{I}(a = a') \tag{3}$$

Where $\mathcal{A}$ denotes the set of candidate answers and $\mathbb{I}(\cdot)$ is the indicator function.

4 Experimental Setup

Datasets To comprehensively assess our framework across dimensions, we include three well-established math reasoning benchmarks: (1) monolingual datasets **GSM8K** (Cobbe et al., 2021) and **MATH500** (Hendrycks et al., 2021); and (2) multilingual dataset **MGSM** (Shi et al., 2023a). Additionally, to test our CLC method on the more challenging dataset, we machine-translate questions from MATH500 into 10 languages in MGSM.[1]

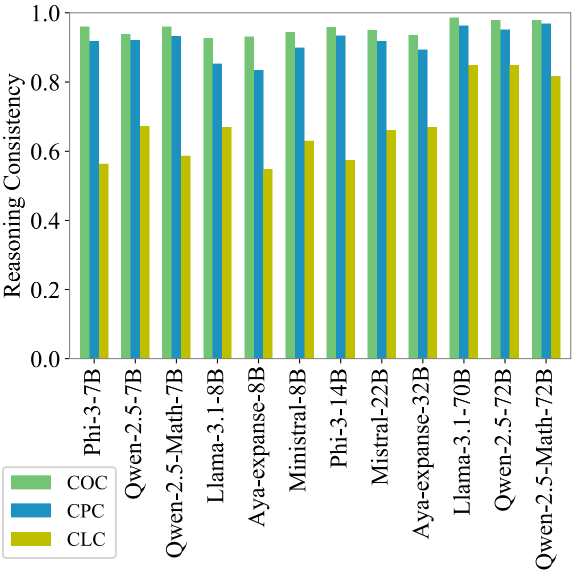

Figure 3: Reasoning consistency on three dimensions of variation. Note that COC and CPC are evaluated on the monolingual benchmark GSM8K, while CLC is evaluated on the multilingual benchmark MGSM.

Models We select a range of open-source state-of-the-art LLMs in varying scales: (i) 7-8B; (ii) 14-32B; and (iii) 70-72B.[2] For all models, we only consider instruction-tuned versions.

Implementation We use 4-shot for all languages except TE which only uses 2-shot, since a 4-shot prompt would exceed the default maximum length, due to tokenization issues unfavourable to this language (Ahia et al., 2023).[3] All prompt exemplars we use are released by Shi et al. (2023a) and An et al. (2024). We report the final answer accuracy for all experiments except the consistency score.

5 Results and Analysis

We report results for all variation dimensions, and then zoom in on CLC for a more detailed analysis.

5.1 Reasoning Consistency

Figure 3 shows reasoning consistency results on the three different dimensions. COC achieves the highest scores, followed by CPC, with CLC having the lowest scores across the board. This suggests that all models are more sensitive to language variations while results are more consistent across different exemplar orders in few-shot prompting. Indeed, when looking at COC only, all models achieve consistency scores above 0.9. Notably, the Llama-3.1

[1] https://translate.google.com/.

[2] More details are in Appedix A.5

[3] Examples are in Appendix A.1.

Models	GSM8K CoT	RwS	RwS+	RtS-	RtS	CPC	MATH500 CoT	RwS	RwS+	RtS-	RtS	CPC
7-8B												
Phi-3-7B	88.2	84.5	87.0	84.8	88.1	**90.0**	40.4	36.2	41.8	39.6	43.4	**46.6**
Qwen-2.5-7B	88.3	86.0	89.8	86.3	90.1	**92.0**	63.6	57.2	63.6	59.0	62.0	**72.4**
Qwen-2.5-Math-7B	90.0	87.6	91.1	89.1	92.3	**94.1**	75.2	72.8	73.2	72.4	73.2	**77.4**
Llama-3.1-8B	79.7	73.9	78.2	77.9	81.2	**83.8**	42.6	36.6	45.6	39.0	46.6	**50.2**
Aya-expanse-8B	76.7	73.5	77.9	73.5	78.2	**82.4**	28.0	26.0	29.4	26.6	27.0	**32.4**
Ministral-8B	81.2	78.7	83.0	78.9	84.0	**84.7**	42.8	41.8	48.0	39.2	50.6	**51.0**
14-32B												
Phi-3-14B	89.2	86.4	89.2	86.9	89.8	**90.2**	44.2	44.0	47.6	43.2	47.2	**51.2**
Mistral-22B	85.8	83.1	85.7	84.8	**88.1**	**88.1**	42.6	44.4	51.4	43.2	48.6	**51.6**
Aya-expanse-32B	83.8	82.3	83.8	82.4	**88.4**	88.1	38.6	37.6	38.2	41.8	40.4	**43.2**
70-72B												
Llama-3.1-70B	94.0	89.8	93.9	91.9	93.6	**94.8**	58.6	49.0	60.2	55.8	63.4	**65.2**
Qwen-2.5-72B	94.6	88.9	94.4	88.6	95.5	**95.8**	63.2	23.8	22.0	63.8	**74.0**	67.0
Qwen-2.5-Math-72B	94.0	92.9	94.7	93.5	94.8	**95.9**	66.6	54.6	58.0	67.6	74.6	**74.8**

Table 1: Reasoning accuracy of CPC on the benchmarks GSM8K and MATH500, obtained via aggregating vanilla CoT prompting and 4 different question rewriting settings. The best result for each model across settings is bolded.

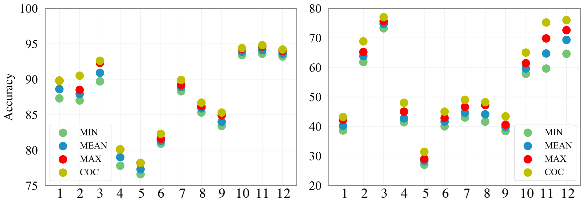

Figure 4: Reasoning accuracy of 4-shot for 8 different exemplars orders. The x-axis numbers correspond to the following models: 1 – Phi-3-7B; 2 – Qwen-2.5-7B; 3 – Qwen-2.5-Math-7B; 4 – Llama-3-1-8B; 5 – Aya-expanse-8B; 6 – Mistral-8B; 7 – Phi-3-14B; 8 – Mistral-22B; 9 – Aya-expanse-32B; 10 – Llama-3-1-70B; 11 – Qwen-2.5-72B; 12 – Qwen-2.5-Math-72B. Notes: (i) minimum score (MIN); (ii) mean score (MEAN); (iii) maximum score (MAX).

family achieves the highest score with the 70B model and the lowest score with the 8B model.

For CPC and CLC, Aya-expanswe-8B has the lowest consistency scores in both dimensions, while larger Qwen2.5 and Llama-3.1 models perform best. Compared to COC and CPC, there is a bigger gap in CLC for different models, even within the same scale, e.g., Phi3-7B vs Qwen2.5-7B. Overall, larger models show higher consistency.

5.2 Consistency Improves Reasoning

For each dimension, we compare the performance obtained exploiting cross-variation consistency to yield a final answer with the performance obtained via the variations on their own.

COC Figure 4 reports the results augmented with COC on GSM8K and MATH500, where we use 8 different exemplar orders for the 4-shot prompt.[4] Compared to vanilla CoT prompting, COC improves the reasoning performance for all models. On dataset GSM8K, COC scores are higher than the average scores of 8 different order prompts on all models, and highest on most models, except for Phi-3-7B, Llama-3.1-8B, and Aya-expanse-8B, where it is on par with the highest scores among the eight ordering configurations we consider in this analysis. On the more challenging dataset MATH500, we see COC achieves the highest scores among all models.

CPC Table 1 shows CPC's on GSM8K and MATH500. Accuracy drops when models are fed only the rewritten question (RwS), as they might lose some information from the original question (manual inspection). When combining the rewritten question with the original one (RwS+), most models score comparably to the original prompting and tend to achieve higher scores in the rewrite-then-solve setting (RtS). The latter observation suggests that asking the model to rewrite the question in a simple way and then solve it, can effectively help the model. Lastly, we see that CPC can further improve the reasoning performance: (i) when comparing to vanilla CoT prompting this is true for all models; and (ii) when comparing to RtS, all models achieve higher accuracy except Aya-expanse-32B on GSM8K and Qwen-2.5-72B on MATH500.

CLC Table 2 and Table 3 present the result of CLC compared to vanilla CoT prompting on

[4]Complete results are in Appendix A.3.

Models	BN	DE	EN	ES	FR	JA	RU	SW	TE	TH	ZH	CLC
7-8B												
Phi-3-7B	14.8	77.6	89.2	85.2	80.4	64.8	74.4	14.0	5.2	18.8	76.0	**91.2**
Qwen-2.5-7B	67.2	72.4	91.6	82.8	72.0	64.8	70.8	16.4	29.2	75.6	74.0	**92.8**
Qwen-2.5-Math-7B	16.8	76.8	92.8	82.0	76.8	61.6	78.8	4.0	5.6	51.2	85.6	**93.6**
Llama-3.1-8B	57.6	64.4	**80.8**	73.6	63.6	52.4	68.0	55.6	49.6	58.8	63.6	78.8
Aya-expanse-8B	29.2	70.4	77.2	74.8	66.8	60.4	72.0	11.6	6.4	22.8	67.2	**82.0**
Ministral-8B	50.4	68.0	**85.6**	76.4	69.6	54.0	70.8	27.6	36.4	53.2	64.4	84.0
14-32B												
Phi-3-14B	14.8	76.0	88.0	87.6	76.8	72.8	80.8	18.4	5.6	12.8	77.6	**90.0**
Mistral-22B	52.0	76.4	87.6	82.4	75.2	62.0	78.4	35.6	17.2	57.6	80.0	**89.2**
Aya-expanse-32B	58.4	74.0	86.0	84.4	80.0	73.6	81.2	29.2	17.2	52.8	77.2	**90.8**
70-72B												
Llama-3.1-70B	83.6	82.0	93.6	87.6	77.6	76.8	84.4	83.2	79.2	80.4	84.0	**93.6**
Qwen-2.5-72B	88.0	84.4	93.2	88.4	80.4	84.4	87.2	66.0	68.8	91.6	86.8	**95.6**
Qwen-2.5-Math-72B	86.4	83.6	94.4	85.6	78.4	81.2	70.4	57.2	68.0	85.6	88.4	**95.2**

Table 2: Reasoning accuracy of CLC compared to vanilla CoT prompting on the MGSM benchmark. Note that bold numbers indicate the best result for each model among different languages and CLC.

Models	BN	DE	EN	ES	FR	JA	RU	SW	TE	TH	ZH	CLC
7-8B												
Phi-3-7B	10.8	34.8	40.4	39.8	5.4	20.8	30.0	9.4	5.6	10.0	33.2	**44.0**
Qwen-2.5-7B	33.4	51.6	63.6	50.0	33.0	50.2	55.8	17.6	17.0	42.2	49.0	**67.4**
Qwen-2.5-Math-7B	29.2	61.2	75.2	62.0	53.4	50.8	65.8	3.8	16.2	23.6	70.6	**79.4**
Llama-3.1-8B	22.2	31.0	42.6	35.4	7.2	29.4	38.6	24.8	9.4	26.8	32.8	**47.2**
Aya-expanse-8B	10.6	25.8	28.0	26.6	25.6	24.0	25.2	8.6	5.6	11.0	24.2	**31.2**
Ministral-8B	17.0	31.0	42.8	42.4	36.8	26.6	38.8	8.4	7.6	18.2	29.8	**44.4**
14-32B												
Phi-3-14B	3.0	42.2	44.2	41.8	9.2	33.6	37.2	13.8	4.6	1.8	38.6	**49.6**
Mistral-22B	16.0	43.2	42.6	43.0	23.6	34.4	43.2	15.2	35.8	23.6	35.2	**53.8**
Aya-expanse-32B	20.8	40.2	38.6	38.2	3.0	35.0	38.8	15.0	10.2	16.0	38.0	**45.0**
70-72B												
Llama-3.1-70B	32.6	41.2	58.6	51.8	9.6	33.4	47.2	44.0	22.4	45.4	42.8	**64.4**
Qwen-2.5-72B	51.4	54.8	63.2	48.4	51.6	27.6	38.4	35.6	37.4	56.2	63.2	**75.2**
Qwen-2.5-Math-72B	58.0	50.0	66.6	60.0	22.4	56.6	60.6	30.6	34.4	58.0	64.4	**75.4**

Table 3: Reasoning accuracy of CLC compared to vanilla CoT prompting on the machine-translated MATH500. Note that bold numbers indicate the best result for each model among different languages and CLC.

MGSM and machine-translated MATH500, respectively. All models perform best on English, with a serious performance gap between underrepresented (e.g., SW) and high-resource languages, especially for smaller models. Similar to COC and CPC, compared to vanilla CoT, CLC yields improvement for most models, with Aya-expanse-32B on GSM8K, for example, showing a significant gain of 4.8% absolute accuracy compared to that of English. For Llama-3.1-8B and Ministral-8B, the accuracy of CLC is slightly lower than that of English, but better than that of all other languages. On the more challenging dataset MATH500, CLC leads to consistent improvements across all models, with the most substantial gain observed on the larger model (e.g, a 12.0% percentage point relative improvement in accuracy over that of English on Qwen-2.5-72B), highlighting its effectiveness in handling more complex tasks and scaling with model size.

Cross-dimension Consistency Table 4 shows ablation results of different combinations on MGSM and machine-translated MATH500. Of the three variation dimensions, CPC performs best overall, followed by COC and CLC. This suggests that CPC can push the model to better diversify its solution paths, while for CLC, this might be due to the large performance gap between English and other languages. The combination of CPC and CLC (P&L) consistently provides strong performance, even outperforming MRC (aggregating three dimensions) in some models, highlighting their novel and impactful role in our framework. While integrating COC occasionally fails to yield additional benefits, bear in mind that each dimension—when applied individually—still brings improvements. Therefore, promoting reasoning consistency across variations can still help induce substantial diversification of the reasoning paths, offering valuable insights into

Datasets	GSM8K							MATH500						
Models	**COC**	**CPC**	**CLC**	**O&P**	**O&L**	**P&L**	**MRC**	**COC**	**CPC**	**CLC**	**O&P**	**O&L**	**P&L**	**MRC**
7-8B														
Phi-3-7B	92.4	92.0	91.2	93.2	93.2	93.2	**94.8**	43.2	46.6	44.0	49.0	47.6	**51.2**	49.6
Qwen-2.5-7B	92.0	93.2	92.8	93.2	**93.6**	**93.6**	**93.6**	68.8	72.4	67.4	72.6	72.4	73.4	**74.0**
Qwen-2.5-Math-7B	94.4	94.8	93.6	94.4	95.6	**96.0**	**96.0**	77.0	77.4	79.4	78.6	79.6	**79.8**	**79.8**
Llama-3.1-8B	80.8	85.6	78.8	83.6	84.0	**87.6**	84.8	48.0	50.2	47.2	51.6	51.4	53.0	**53.6**
Aya-expanse-8B	78.4	**85.2**	82.0	80.8	79.6	84.4	82.8	31.4	32.4	31.2	33.4	33.4	**35.2**	34.2
Ministral-8B	84.4	86.8	84.0	86.0	86.0	87.2	**87.6**	45.0	51.0	44.4	47.8	47.2	**53.2**	49.6
14-32B														
Phi-3-14B	92.0	92.0	90.0	92.0	92.0	92.4	**93.2**	49.0	51.2	49.6	51.0	52.4	**55.6**	53.8
Mistral-22B	87.2	89.6	89.2	87.6	90.0	**93.2**	91.2	48.2	51.6	53.8	52.0	53.0	**57.4**	55.4
Aya-expanse-32B	86.0	89.2	90.8	86.4	89.6	**92.8**	90.0	43.4	43.2	45.0	45.6	46.0	**47.8**	47.4
70-72B														
Llama-3.1-70B	95.6	**96.8**	93.6	96.0	96.4	**96.8**	96.4	65.0	65.2	64.4	68.8	68.2	68.0	**69.0**
Qwen-2.5-72B	96.0	**97.6**	95.6	97.2	96.8	97.2	96.8	75.2	67.0	75.2	77.6	77.6	76.6	**78.8**
Qwen-2.5-Math-72B	94.4	**95.2**	**95.2**	94.8	94.8	94.8	**95.2**	76.0	74.8	75.4	78.6	79.4	79.0	**80.6**

Table 4: Accuracy on MGSM and MATH500. Notes: (i) CPC uses 5 solution paths, COC and CLC use 8 each, and CLC uses 8 languages (excl. BN, SW, TE), so MRC contains a total of 19 paths (excluding the two identical English paths); (ii) best result for each model is bolded.

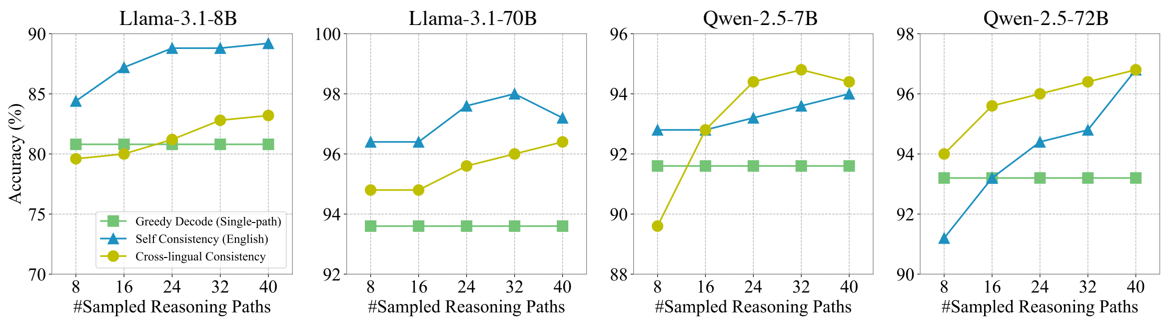

Figure 5: Reasoning accuracy of using varying numbers of reasoning paths on GSM8K.

LLM reasoning beyond commonly used sampling-based strategies.

5.3 Analysis

Accuracy–efficiency Trade-off One can conceive CLC as a multilingual extension of monolingual self-consistency, going beyond the typical sampling-based strategy. Figure 5 reports accuracy under different numbers of reasoning paths for Llama-3.1 and Qwen-2.5. For self-consistency, we use English following (Wang et al., 2023), whereas for CLC, we use 8 languages excluding BN, SW, and TE which have very low results (see Table 2). We sample $N/8$ reasoning paths for each language, thus creating N solutions for CLC. We see that Llama-3.1 shows CLC slightly below self-consistency but above greedy decoding, while Qwen-2.5 achieves notable gains with CLC, surpassing self-consistency when $N \geq 16$. These results suggest that cross-lingual diversity provides complementary reasoning signals. However, as with self-consistency, this improvement comes at the expense of higher computational effort, as our methods require generating multiple solutions across languages for CLC and analogously for COC and CPC. For instance, models generally require generating 8–16 reasoning solutions to surpass greedy decoding, which theoretically corresponds to roughly 8–16× additional computational cost. This highlights the inherent trade-off between accuracy and efficiency: while additional paths and languages lead to better reasoning, they also increase inference overhead.

Case Study Table 5 shows examples of solutions generated by Qwen2.5-72B for the same MSGM question in different languages. For the EN question, the model incorrectly reasons in the first step: *On Mondays, Wednesdays, and Fridays, Kimo has* $3 * 1 = 3$ *hours of classes*, which results in the wrong reasoning and final answer. It is interesting to see that the model reasons correctly in the other languages except BN, while the solution paths might be logically different. For DE, the model first calculates the total hours for Monday, Wednesday, and Friday, then calculates the total hours for Tuesday and Thursday, and then adds them together

Language	CoT Reasoning
Question	On Mondays, Wednesdays, and Fridays, college student Kimo has three 1-hour classes each day. On Tuesdays and Thursdays, he has two 2-hour classes each day. In one semester, there are 16 weeks of school. In one semester, how many hours does Kimo spend attending classes?
BN	প্রতি সপ্তাহে, কিমোর ক্লাসের সময় 3 * 1 = 3 ঘণ্টা (সোমবার, বুধবার, শুক্রবার) + 2 * 2 = 4 ঘণ্টা (মঙ্গলবার, বৃহস্পতিবার) = 7 ঘণ্টা। একটি সেমিস্টার 16 সপ্তাহের, তাই মোট 16 * 7 = 112 ঘণ্টা কিমোকে ক্লাসে ব্যয় করতে হয়। উত্তর হল 112।
DE	Montags, mittwochs und freitags verbringt Kimo 3 x 1 = 3 Stunden pro Tag mit Kursen. In einer Woche verbringt er an diesen Tagen insgesamt 3 x 3 = 9 Stunden mit Kursen. Dienstags und Donnerstags verbringt er 2 x 2 = 4 Stunden pro Tag mit Kursen. In einer Woche verbringt er an diesen Tagen insgesamt 2 x 4 = 8 Stunden mit Kursen. Insgesamt verbringt er in einer Woche 9 + 8 = 17 Stunden mit Kursen. In einem 16-wochen-Semester verbringt er 16 x 17 = 272 Stunden mit Kursen. Die Antwort lautet 272.
EN	On Mondays, Wednesdays, and Fridays, Kimo has 3 * 1 = 3 hours of classes. On Tuesdays and Thursdays, he has 2 * 2 = 4 hours of classes. In one week, he has 3 + 4 + 3 = 10 hours of classes. In one semester, Kimo spends 10 * 16 = 160 hours attending classes. The answer is 160.
ES	Los lunes, miércoles y viernes, Kimo pasa 3 * 1 = 3 horas en clases. Como hay 3 días a la semana, pasa 3 * 3 = 9 horas en clases. Los martes y jueves, Kimo pasa 2 * 2 = 4 horas en clases. Como hay 2 días a la semana, pasa 2 * 4 = 8 horas en clases. En total, Kimo pasa 9 + 8 = 17 horas a la semana en clases. En un semestre, Kimo pasa 17 * 16 = 272 horas en clases. La respuesta es 272.
FR	Kimo a 3 cours d'une heure les lundis, mercredis et vendredis, ce qui fait 3 x 1 = 3 heures par jour. Comme il y a 3 jours dans la semaine où il a 3 heures de cours, cela fait 3 x 3 = 9 heures par semaine. Les mardis et jeudis, il a 2 cours de 2 heures, ce qui fait 2 x 2 = 4 heures par jour. Comme il y a 2 jours dans la semaine où il a 4 heures de cours, cela fait 2 x 4 = 8 heures par semaine. Au total, Kimo a 9 + 8 = 17 heures de cours par semaine. Comme il y a 16 semaines d'école par semestre, Kimo consacre 17 x 16 = 272 heures à ses cours durant un semestre. La réponse est 272.
JA	月曜、水曜、金曜は、キモは1時間の講義を3つ受けるので、1日あたり3時間、3日間で9時間です。火曜と木曜は、2時間の講義を2つ受けるので、1日あたり4時間、2日間で8時間です。1週間で、キモは9+8=17時間講義に出席します。1セメスターは16週間あるので、キモは1セメスターで17*16=272時間講義に出席します。答えは272時間です。
RU	в понедельник, среду и пятницу Кимо проводит 3 * 1 = 3 часа на занятиях. В неделю таких дней 3, значит, за неделю он проводит 3 * 3 = 9 часов на занятиях в эти дни. Вторник и четверг Кимо проводит 2 * 2 = 4 часа на занятиях. В неделю таких дней 2, значит, за неделю он проводит 4* 2 = 8 часов на занятиях в эти дни. Всего за неделю Кимо проводит 9 + 8 = 17 часов на занятиях. В семестре 16 недель, значит, за семестр Кимо проводит 17 * 16 = 272 часа на занятиях. Ответ — 272.
SW	Kimo ana masomo matatu ya saa 1 kila siku kwa siku tatu, hivyo kwa jumla ana masomo 3 * 1 * 3 = 9 saa. Kimo pia ana masomo mawili ya saa 2 kila siku kwa siku mbili, hivyo kwa jumla ana masomo 2 * 2 * 2 = 8 saa. Kwa jumla, Kimo ana masomo 9 + 8 = 17 saa kwa wiki. Kimo ana shule kwa wiki 16, hivyo kwa jumla ana masomo 17 * 16 = 272 saa kwa muhula. Jibu ni 272.
TE	సోమవారాలు, బుధవారాలు, మరియు శుక్రవారాల నాడు కిమోకు మూడు 1-గంట క్లాసులు ఉన్నాయి, అందువల్ల ప్రతి వారాలో 3 రోజుల్లో 3 గంటలు క్లాసులు ఉన్నాయి, అంటే 3*3=9 గంటలు. మంగళవారం మరియు గురువారాల్లో అతడికి ప్రతిరోజూ రెండు 2-గంటల క్లాసులున్నాయి, అందువల్ల ప్రతి వారాలో 2 రోజుల్లో 4 గంటలు క్లాసులు ఉన్నాయి, అంటే 2*4=8 గంటలు. అందువల్ల ప్రతి వారాలో కిమోకు మొత్తం 9+8=17 గంటలు క్లాసులు ఉన్నాయి. ఒక సెమిస్టర్‌లో 16 వారాలు ఉన్నాయి, అందువల్ల ఒక సెమిస్టర్‌లో కిమోకు మొత్తం 16*17=272 గంటలు క్లాసులు ఉన్నాయి. సమాధానం 272.",
TH	ในวันจันทร์ วันพุธ และวันศุกร์ ไคโมเข้าเรียน 3 คาบต่อวัน ซึ่งแต่ละคาบใช้เวลา 1 ชั่วโมง แสดงว่าใน 3 วันนี้เขาใช้เวลา 3 * 1 * 3 = 9 ชั่วโมง ในวันอังคารและวันพฤหัสบดี เขาเข้าเรียน 2 คาบต่อวัน ซึ่งแต่ละคาบใช้เวลา 2 ชั่วโมง แสดงว่าใน 2 วันนี้เขาใช้เวลา 2 * 2 * 2 = 8 ชั่วโมง ดังนั้น ในหนึ่งสัปดาห์ ไคโมใช้เวลา 9 + 8 = 17 ชั่วโมง ในหนึ่งภาคการศึกษามีการเรียนการสอน 16 สัปดาห์ ดังนั้น ไคโมใช้เวลา 17 * 16 = 272 ชั่วโมง คำตอบคือ 272"
ZH	周一、周三和周五，基莫每天有 3 节 1 小时的课，即 3 * 1 = 3 小时。周二和周四，他每天有 2 节 2 小时的课，即 2 * 2 = 4 小时。一周内，他上课的时间为 3 * 3 + 2 * 4 = 9 + 8 = 17 小时。一个学期有 16 周，所以基莫一个学期上课的时间为 17 * 16 = 272 小时。答案是 272。 答案是 272 小时。

Table 5: Case study in the test set of MSGM, where the solutions are generated by Qwen2.5-72B for a question written in different languages. Note that here we only show an EN question; others are in the Appendix A.4.

to get the total hours for the week; while for ZH, it calculates the hours per day and then calculates the total hours for the week. Reasoning appears correct in some low-resource languages (SW, TE, and TH), suggesting that our approach can leverage the model's capabilities in low-resource languages to assist reasoning in high-resource languages, in addition of course to the usual reverse benefit.

6 Conclusion

MRC is a multidimensional reasoning framework, which comprehensively studies reasoning consistency when the same math problems are presented to the model with systematic variations along three different dimensions. By leveraging such variations and answer consistency, MRC improves overall accuracy on both monolingual and multilingual benchmarks, demonstrating its effectiveness in enhancing lightweight models and its strong scalability for more complex tasks. Our experiments seem to suggest that the largest the diversity of solution paths, the stronger the benefit from exploiting consistency. As we do not yet have concrete evidence for this hypothesis, a natural future direction would be to study path diversity in a quantifiable way. Another valid extension would be *integrating* the different dimensions (e.g., COC per language). While combining multiple dimensions of variations presents an explosion of possibilities, strategic selection based on empirical results, especially cross-all accuracy, might mitigate this challenge.

7 Limitations

While we investigated model consistency in mathematical reasoning and successfully leveraged it to improve reasoning accuracy, several promising directions remain for future exploration. We mainly focus on the variations in model inputs and consistency in final answers, while both the variation and consistency of the reasoning paths are interesting directions. Specifically, variations in the input will lead to variations in the output, which includes logical consistency and inconsistency, thus affecting the final result.Lastly, it is not yet clear how variations in input affect the model's reasoning logic (variations in reasoning paths), which requires a much bigger unpacking.

References

Marah Abdin, Jyoti Aneja, and et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*.

Rishi Bommasani, Drew A. Hudson, and et al. 2022. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165, Online. Association for Computational Linguistics.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. 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*.

Aaron Grattafiori, Abhimanyu Dubey, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Song Jiang, Zahra Shakeri, Aaron Chan, Maziar Sanjabi, Hamed Firooz, Yinglong Xia, Bugra Akyildiz, Yizhou Sun, Jinchao Li, Qifan Wang, and Asli Celikyilmaz. 2024. Resprompt: Residual connection prompting advances multi-step reasoning in large language models. In *The Twelfth International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, page 611–626. Association for Computing Machinery.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

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

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631, Toronto, Canada. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. 2022. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1768, 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, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). 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. 2023a. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

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. 2023b. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. 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*.

An Yang, Baosong Yang, and et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Beichen Zhang, and et al. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Yue Zhou, Yada Zhu, Diego Antognini, Yoon Kim, and Yang Zhang. 2024. Paraphrase and solve: Exploring and exploiting the impact of surface form on mathematical reasoning in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2793–2804, Mexico City, Mexico. Association for Computational Linguistics.

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2025. Can model uncertainty function as a proxy for multiple-choice question item difficulty? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11304–11316, Abu Dhabi, UAE. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A Appendix

A.1 Prompt Examples

Prompt 1

Question: Roger has 5 tennis balls. **[omit 15 words]**. How many tennis balls does he have now?
Solution: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Question: There were nine computers in the server room. **[omit 12 words]**. How many computers are now in the server room?
Solution: There are 4 days from monday to thursday. **[omit 39 words]**. The answer is 29.

[*Two more exemplars*]

Question: {}
Solution:

…

Prompt *x*

Question: There were nine computers in the server room. **[omit 12 words]**. How many computers are now in the server room?
Solution: There are 4 days from monday to thursday. **[omit 39 words]**. The answer is 29.

Question: Roger has 5 tennis balls. **[omit 15 words]**. How many tennis balls does he have now?
Solution: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

[*Two more exemplars*]

Question: {}
Solution:

Figure 6: Examples of prompts for COC.

Rewrite-without-solve Prompt

Rewrite the following math problems to make them easier for LLMs to solve.

Question: Roger has 5 tennis balls. **[omit 15 words]**. How many tennis balls does he have now?
Rewritten Question: Roger starts with 5 tennis balls. **[omit 15 words]**. How many tennis balls does Roger have in total?

[*Three more exemplars*]

Question: {}
Rewritten Question:

…

Rewrite-then-solve Prompt

Rewrite the following math problems to make them easier for LLMs to solve, then solve them step by step.

Question: Roger has 5 tennis balls. **[omit 14 words]**. How many tennis balls does he have now?
Rewritten Question: Roger starts with 5 tennis balls. **[omit 15 words]**. How many tennis balls does Roger have in total?
Solution: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

[*Three more exemplars*]

Question: {}
Rewritten Question:

Figure 7: Examples of prompts for CPC.

EN Prompt

Question: Roger has 5 tennis balls. **[omit 15 words]**. How many tennis balls does he have now?
Solution: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

[*Three more exemplars*]

Question: {}
Solution:

…

DE Prompt

Frage: Roger hat 5 Tennisbälle. **[omit 13 words]**. Wie viele Tennisbälle hat er jetzt?
Lösung: Roger begann mit 5 Bällen. **[omit 9 words]**. 5 + 6 = 11. Die Antwort ist 11.

[*Three more exemplars*]

Frage: {}
Lösung:

Figure 8: Examples of prompts for CLC.

A.2 Datasets

We evaluate our framework on three math reasoning datasets: (i) **GSM8K** (Cobbe et al., 2021), an English dataset of grade school math word problems (about 7,500 for training and 1,319 for testing); (ii) **MGSM** (Shi et al., 2023a), consisting of 250 questions selected from GSM8K and manually translated into ten languages: Bengali (BN), Chinese (ZH), French (FR), German (DE), Japanese (JA), Russian (RU), Spanish (ES), Swahili (SW), Telugu (TE) and Thai (TH). Thus, it contains a total of 11 languages including English; (iii) **MATH500** is a benchmark of competition, which contains 500 math problems of varying difficulty.

A.3 COC Results

Orders (4-shot)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**COC**
7-8B									
Phi-3-7B	88.5	88.8	88.6	88.9	87.3	89.8	88.4	88.2	89.8
Qwen2.5-7B	88.3	87.0	87.6	88.0	88.0	87.8	88.5	88.2	90.5
Qwen2.5-Math-7B	90.0	89.7	91.3	92.3	91.5	90.5	90.8	91.0	92.6
Llama-3.1-8B	79.7	78.8	79.0	79.0	80.1	79.1	77.8	78.7	80.1
Aya-expanse-8B	76.7	77.3	78.2	76.6	76.9	77.6	77.3	77.4	78.2
Ministral-8B	81.2	81.4	80.9	81.4	80.9	81.3	81.4	81.6	82.3
14-32B									
Phi-3-14B	89.2	88.9	88.6	88.3	88.9	89.0	89.1	88.9	89.9
Mistral-22B	85.8	85.8	85.9	85.3	86.1	86.1	86.1	86.2	86.7
Aya-expanse-32B	83.8	83.4	84.6	83.6	84.9	84.3	83.5	83.9	85.3
70-72B									
Llama-3.1-70B	94.0	94.0	94.3	94.1	93.9	93.9	93.4	93.8	94.4
Qwen2.5-72B	94.6	93.9	93.9	93.6	94.1	94.1	94.2	94.2	94.8
Qwen2.5-Math-72B	94.0	93.6	93.6	93.7	93.8	93.6	93.2	93.4	94.2

Table 6: Reasoning accuracy of prompts in different orders on GSM8K compared to COC.

Orders (4-shot)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**COC**
7-8B									
Phi-3-7B	40.4	41.0	40.0	40.8	39.0	42.2	40.0	38.6	43.2
Qwen2.5-7B	63.6	64.6	64.8	64.4	63.6	65.2	61.8	62.4	68.8
Qwen2.5-Math-7B	75.2	74.4	75.2	75.6	73.2	75.4	74.6	74.2	77.0
Llama-3.1-8B	42.6	44.0	41.8	41.4	42.4	41.6	45.0	42.6	48.0
Aya-expanse-8B	28.0	28.8	29.0	27.4	28.6	28.8	27.0	29.0	31.4
Ministral-8B	42.8	42.2	40.0	41.4	42.0	41.2	41.8	41.2	45.0
14-32B									
Phi-3-14B	44.2	43.0	44.2	45.8	44.4	46.6	44.0	44.4	49.0
Mistral-22B	42.6	47.2	43.0	44.0	44.0	44.6	41.6	45.6	48.2
Aya-expanse-32B	38.6	40.0	39.4	38.4	40.4	40.0	40.4	40.6	43.4
70-72B									
Llama-3.1-70B	58.6	57.8	60.0	59.6	59.0	60.2	60.2	61.4	65.0
Qwen2.5-72B	63.2	65.2	62.2	59.6	67.2	61.4	69.8	68.6	75.2
Qwen2.5-Math-72B	66.6	64.6	70.2	72.2	68.8	69.6	72.6	69.8	76.0

Table 7: Reasoning accuracy of prompts in different orders on MATH500 compared to COC.

A.4 Examples of questions written in different languages

Figure 9 provides examples of questions written in different languages.

A.5 Models

We select a range of open-source state-of-the-art LLMs in varying scales. For all models, we only consider instruction-tuned versions.

7-8B: Phi-3-7B (128k) (Abdin et al., 2024); Qwen2.5-7B (Yang et al., 2024a); Qwen2.5-7B-Math (Yang et al., 2024b); Llama-3.1-8B (Grattafiori et al., 2024); Aya-expanse-8B (Üstün et al., 2024); Ministral-8B[5].

[5] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

Language	Math Question
EN	On Mondays, Wednesdays, and Fridays, college student Kimo has three 1-hour classes each day. On Tuesdays and Thursdays, he has two 2-hour classes each day. In one semester, there are 16 weeks of school. In one semester, how many hours does Kimo spend attending classes?
BN	কলেজ ছাত্র কিমোর প্রতি সোমবার, বুধবার ও শুক্রবারে তিনটি 1-ঘণ্টার ক্লাস থাকে। প্রতি মঙ্গলবার ও বৃহস্পতিবারে, তার দুটি 2-ঘণ্টার ক্লাস থাকে। একটি সেমিস্টার 16 সপ্তাহের স্কুল থাকে। একটি সেমিস্টারে, কিমোকে ক্লাসে কত ঘণ্টা ব্যয় করতে হয়?
DE	Montags, mittwochs und freitags hat College-Student Kimo drei 1-st\u00fcndige Kurse pro Tag. Dienstags und Donnerstags hat er zwei 2-st\u00fcndige Kurse pro Tag. Ein Semester hat 16 Schulwochen. Wie viele Stunden verbringt Kimo in einem Semester mit Kursbesuchen?
ES	Los lunes, miércoles y viernes, el estudiante universitario Kimo tiene tres clases de 1 hora por día. Los martes y jueves, tiene dos clases de 2 horas por día. En un semestre, hay 16 semanas de clases. En un semestre, ¿cuántas horas pasa Kimo en clases?
FR	Les lundis, mercredis et vendredis, l'étudiant Kimo a trois cours d'une heure par jour. Les mardis et jeudis, il a deux cours de 2 heures chaque jour. S'il y a 16 semaines d'école par semestre, combien d'heures Kimo consacre-t-il à ses cours durant un semestre ?
JA	月曜、水曜、金曜に、大学生のキモは1時間の3つの講義を各日受ける。火曜と木曜に、彼は2時間の講義を2つ各日受ける。1セメスターで、学校は16週間ある。1セメスターで、キモは何時間講義に出席して過ごす?
RU	Каждый понедельник, среду и пятницу у студента колледжа Кимо три 1-часовых занятия. Каждый вторник и четверг у него два 2-часовых занятия. В одном семестре 16 учебных недель. Сколько часов Кимо проводит на занятиях за один семестр?
SW	Siku za Jumatatu, Jumatano, na Ijumaa, Kimo mwanafunzi wa chuo huwa na masomo matatu ya saa 1 kila siku. Siku za Jumanne na Alhamisi, huwa ana masomo mawili ya saa 2 kila siku. Katika muhula moja, kuna wiki 16 za shule. Katika muhula moja, Kimo huwa anatumia saa ngapi kuhudhuria masomo?
TE	సోమవారాలు, బుధవారాలు, మరియు శుక్రవారాల నాడు, కాలేజీ విద్యార్థి కిమోకు ప్రతిరోజు మూడు 1-గంట క్లాసులు ఉన్నాయి. మంగళవారం మరియు గురువారాల్లో, అతడికి ప్రతిరోజు రెండు 2-గంటల క్లాసులున్నాయి. ఒక సెమిస్టర్‌లో, స్కూలులో 16 వారాలు ఉన్నాయి. ఒక సెమిస్టర్‌లో, కిమో క్లాసులకు హాజరు కావడానికి ఎన్ని గంటలు గడిపాడు?
TH	ในวันจันทร์ วันพุธ และวันศุกร์ ไคโมซึ่งเป็นนักศึกษาเข้าเรียนสามคาบต่อวัน โดยแต่ละคาบใช้เวลา 1 ชั่วโมง ส่วนในวันอังคารและวันพฤหัสบดี เขาเข้าเรียนสองคาบต่อวัน ซึ่งแต่ละคาบใช้เวลา 2 ชั่วโมง ในหนึ่งภาคการศึกษามีการเรียนการสอน 16 สัปดาห์ ไคโมใช้เวลากี่ชั่วโมงไปกับการเข้าเรียน
ZH	周一、周三和周五，大学生基莫每天有 3 节 1 小时的课。周二和周四，他每天有 2 节 2 小时的课。一个学期中有 16 周的上学时间。在一个学期中，基莫用多少时间上课?

Figure 9: Examples of questions written in different languages.

14-32B: Phi-3-14B (Abdin et al., 2024); Mistral-22B[6]; Aya-expanse-32B (Üstün et al., 2024).

70-72B: Qwen2.5-72B (Yang et al., 2024a); Qwen2.5-72B-Math (Yang et al., 2024b); Llama-3.1-70B (Grattafiori et al., 2024).

A.6 Implementation

We perform inference experiments on 4 × NVIDIA H100 94GB GPUs using the library vLLM (Kwon et al., 2023), without training or fine-tuning language models. During inference, we use few-shot prompts covering the 11 languages released by Shi et al. (2023a). In the multilingual scenario, we use 4-shot for all languages except TE which only uses 2-shot, since a 4-shot prompt would exceed the default maximum length, due to tokenization issues unfavourable to this language (Ahia et al., 2023). We use greedy decoding unless otherwise specified. For all experiments we report the final answer accuracy except the reasoning consistency score.

A.7 Correlation

Figure 10 shows the correlations of models' accuracy with the three consistency scores. COC and CPC have high correlations with reasoning accuracy, while CLC has a weak and non-significant one. This suggests that we can use COC and CPC to assess the model's uncertainty in its generated solutions without using gold answers. While CLC does not seem to be a reasonable metric to assess models' accuracy, it can still be used to evaluate models from a multilingual perspective.

[6] https://huggingface.co/mistralai/Mistral-Small-Instruct-2409

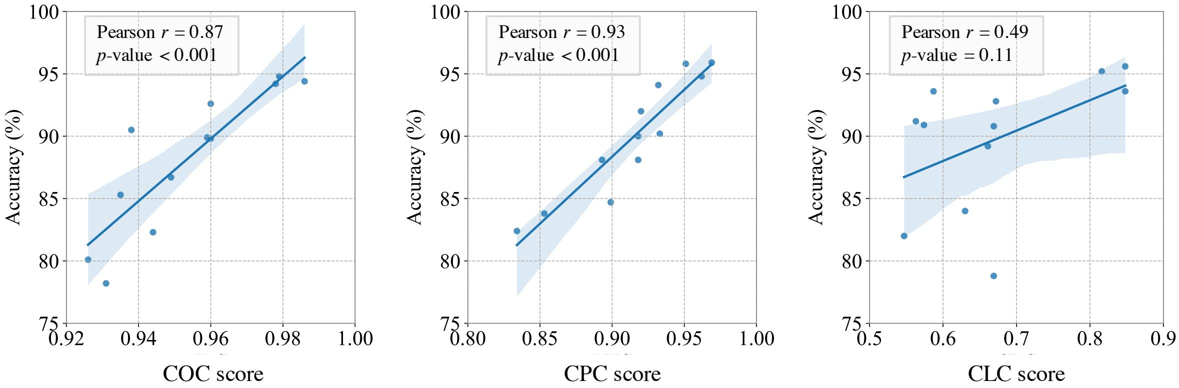

Figure 10: Pearson correlation between models' accuracy and different consistency scores.