

mCoT: Multilingual Instruction Tuning for Reasoning Consistency in Language Models

Huiyuan Lai and Malvina Nissim
Center for Language and Cognition (CLCG)
University of Groningen / The Netherlands
{h.lai, m.nissim}@rug.nl

Abstract

Large language models (LLMs) with Chain-of-thought (CoT) have recently emerged as a powerful technique for eliciting reasoning to improve various downstream tasks. As most research mainly focuses on English, with few explorations in a multilingual context, the question of how reliable this reasoning capability is in different languages is still open. To address it directly, we study multilingual reasoning consistency across multiple languages, using popular open-source LLMs. First, we compile the first large-scale multilingual math reasoning dataset, mCoT-MATH, covering eleven diverse languages. Then, we introduce multilingual CoT instruction tuning to boost reasoning capability across languages, thereby improving model consistency. While existing LLMs show substantial variation across the languages we consider, and especially low performance for lesser resourced languages, our 7B parameter model mCoT achieves impressive consistency across languages, and superior or comparable performance to close- and open-source models even of much larger sizes.

1 Introduction

Recent progress on language models shows that they can achieve surprising performance on complex reasoning tasks in natural language processing (NLP), such as symbolic reasoning (Wei et al., 2022a; Kojima et al., 2022), math word problem (Cobbe et al., 2021; Wei et al., 2022a), and commonsense reasoning (Wei et al., 2022a; Kojima et al., 2022). Most of the research focuses on prompting large language models (LLMs), where the LLMs are conditioned on a few examples or instructions describing the target task (Wei et al., 2022b; Fu et al., 2023b; Zhou et al., 2023; Kojima et al., 2022; Chuanyang et al., 2023).

While most previous works focus on reasoning with LLMs in English, Shi et al. (2023) recently have extended it to a multilingual setting leveraging a few-shot prompting approach. However,

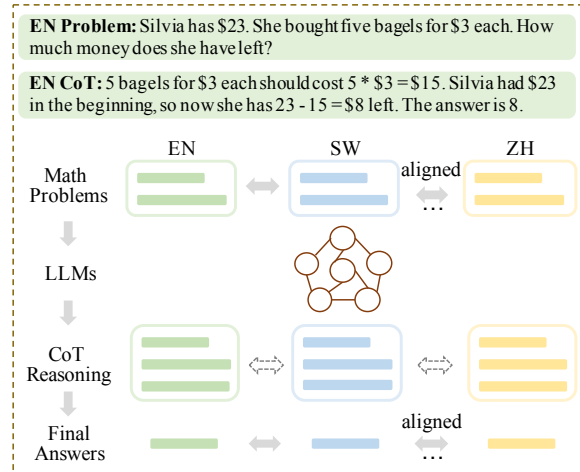


Figure 1: Overview of multilingual reasoning; LLMs are expected to have consistent reasoning capabilities across different languages when given the same problem which has the same answer. Shown in picture are three example languages: English (EN), Swahili (SW), and Chinese (ZH). For EN, we show the problem formulation, and the Chain-of-Thought (CoT) reasoning.

performance for lesser resourced languages still lags behind, in a similar way as generalization of factual knowledge has been shown to vary widely across languages (Fierro and Søgaard, 2022; Qi et al., 2023), mainly due to the fact that most languages are not well represented in LLMs. The work we present here has the twofold aim of (i) better understanding and evaluating the general reasoning capabilities of LLMs beyond just English, and (ii) providing lesser resourced languages with capable but manageable models which can be used for reasoning tasks.

To this end, we propose to measure reasoning consistency across multiple languages. As shown in Figure 1, LLMs are expected to produce logically similar reasoning solutions and consistent final results for inputs which are semantically equivalent but expressed in different languages. Based on our findings, we aim to enhance multilingual

reasoning abilities through instruction tuning, yielding a model that can solve reasoning tasks in various languages, with similar reasoning capabilities across those languages. Specifically, we focus on math word problems and empirically investigate the reasoning consistency of current open-source state-of-the-art LLMs across multiple languages. The model’s reasoning consistency is evaluated toward the final answer (the final results should be the same across languages). On the basis of preliminary results showing substantial reasoning gaps between different languages, we propose a multilingual Chain-of-Thought reasoning (mCoT) framework using instruction tuning, which aims to boost reasoning capability across languages, thereby improving consistency. So, first we construct the multilingual instruction training dataset (mCoT-MATH) by automatically translating English data into multiple languages, and then we use it for LLM finetuning.

In summary, our contributions are:¹

- We propose to study reasoning consistency of LLMs across different languages, providing insights into (the evaluation of) this ability of LLMs.
- We compile and distribute mCoT-MATH, the first large-scale multilingual math CoT reasoning dataset containing around 6.3 million samples for 11 diverse languages.
- Based on mCoT-MATH, we train and make available a 7B parameter model mCoT for multilingual math reasoning, which achieves impressive consistency across languages, and superior or comparable performance to closed and open-source models.

2 Background

Math Reasoning with LLMs In recent years, LLMs such as GPT-3 (Brown et al., 2020) have shown impressive performance in various NLP tasks. In particular, LLMs with CoT-based method exhibit the emergent ability to perform complex math reasoning tasks (Wei et al., 2022b; Wang et al., 2023). One popular line is prompt engineering, which aims to elicit reasoning capability of LLMs by exploring various prompts, such as basic CoT prompting (Wei et al., 2022b), complex CoT (Fu et al., 2023b), auto-CoT (Zhang et al., 2023), self-consistency CoT (Wang et al., 2023),

multilingual CoT (Shi et al., 2023), least-to-most prompting (Zhou et al., 2023), progressive-hint prompting (Chuanyang et al., 2023), residual connection prompting (Jiang et al., 2024), and using specific phrases like “Let’s think step by step” (Kojima et al., 2022). In addition to guiding the model through prompting methods, several works propose to control the reasoning path during inference through verifier (Cobbe et al., 2021; Khalifa et al., 2023) and decoding method (O’Brien and Lewis, 2023). Another research line is a tailor-designed reasoning model that improves the mathematical reasoning ability of LLMs through instruction tuning on reasoning data, including reinforcement learning (Uesato et al., 2022; Luo et al., 2023), knowledge distillation (Fu et al., 2023a; Hsieh et al., 2023; Magister et al., 2023; Shridhar et al., 2023; Yue et al., 2024), and data augmentation (Huang et al., 2023; Zelikman et al., 2022; Ni et al., 2023; Zhu et al., 2023; Yu et al., 2024). In this work we extend the multilingual reasoning carried out by Shi et al. (2023) in GPT-3 and PaLM (Chowdhery et al., 2022) to current open-source popular LLMs, and study multilingual reasoning consistency. Additionally, following previous works (Chen et al., 2023a; Chai et al., 2024), we employ machine translation to translate existing English data into other languages for multilingual reasoning, but we do so at a very large scale, with substantial gains in performance.

Consistency in Language Models Consistency is one of the core qualities of language models, which refers to models behaving consistently on semantically equivalent inputs (Elazar et al., 2021; Fierro and Søgaard, 2022). Intra-language consistency has been studied across different tasks, such as language inference (Li et al., 2019; Mitchell et al., 2022), explanation generation (Camburu et al., 2020), fill-in-the-blank phrases (Ravichander et al., 2020), and math reasoning (Wang et al., 2023). These works mainly consider the English language, although there is some recent research on factual consistency in a multilingual scenario (Fierro and Søgaard, 2022; Qi et al., 2023). To our knowledge, the present work is the first systematic analysis of multilingual reasoning consistency for LLMs, measuring the extent to which language models reason about the same answer to the same question written in different languages. In this context, we introduce multilingual CoT instruction tuning to boost reasoning capability across

¹Data, code, and model are available at <https://github.com/laihuiyuan/mcot>.

multiple languages, thereby improving model consistency.

3 Multilingual Reasoning Consistency

This section provides details on the task, the data, and the experimental setup we use to investigate the LLMs’ reasoning capability in different languages and their reasoning consistency.

3.1 Task, Dataset, and Setup

Task Definition Generally, each math problem is fed to an LLM along with a set of manually written CoT exemplars, which is expected to generate all necessary intermediate steps up to and including the final answer. Given a set of math problems \mathcal{M} , each problem consists of a triplet (question: q_x , reasoning steps: r_x , final answer: a), where the problem and intermediate steps are written in a natural language $x \in \mathcal{L}$. We define the multilingual reasoning consistency of an LLM as the extent to which it reasons to the same answer for the same question asked in different languages, including correct consistency and incorrect consistency. For a given language pair (x, y) , correct consistency (CC) is the percentage of identical math problems written in that language pair for which the correct answer is predicted. Formally:

$$CC(x, y) = \frac{\sum_{i=1}^{|M|} \mathbb{I}(\hat{a}_i^x = \hat{a}_i^y = a_i)}{|M|} \quad (1)$$

Where $\mathbb{I}(\cdot)$ is the indicator function. a_i represents the gold answer corresponding to the i -th math question, \hat{a}_i^x and \hat{a}_i^y are the predicted answers to the i -th question written in languages x and y , respectively. For incorrect consistency (IC), we calculate the proportion of predicted answers that are incorrect but identical in the language pair out of the total number of incorrect answers in each respective language, and take the average of the two languages as the final result:

$$IC(x, y) = \left(\frac{\sum_{i=1}^{|M|} \mathbb{I}(\hat{a}_i^x = \hat{a}_i^y \neq a_i)}{\sum_{i=1}^{|M|} \mathbb{I}(\hat{a}_i^x \neq a_i)} + \frac{\sum_{i=1}^{|M|} \mathbb{I}(\hat{a}_i^x = \hat{a}_i^y \neq a_i)}{\sum_{i=1}^{|M|} \mathbb{I}(\hat{a}_i^y \neq a_i)} \right) / 2 \quad (2)$$

Reasoning Dataset As defined above and depicted in Figure 1, multilingual reasoning consistency requires both questions and reasoning steps to be written in the same language, for multiple

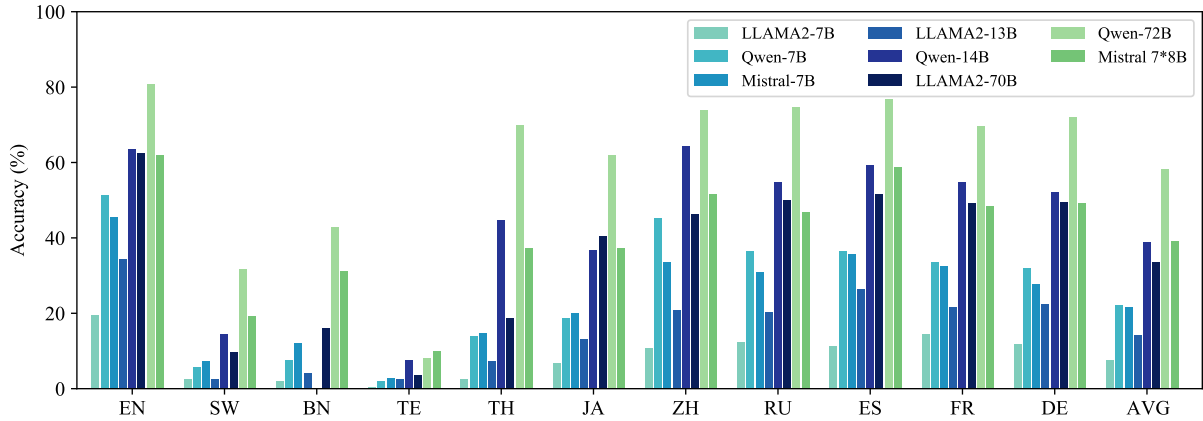
languages. GSM8K (Cobbe et al., 2021) is an English (EN) dataset which includes high-quality grade school math word problems, each involving basic arithmetic operations (addition, subtraction, multiplication, and division) that usually require two to eight steps to solve according to the official, provided solution. It contains approximately 7,500 and 1,319 samples for training and testing, respectively. Based on GSM8K, Shi et al. (2023) create MGSM, extending math reasoning into a multilingual setting. To do so, they select the first 250 problems from GSM8K and manually translate them into ten different languages: Bengali (BN), Chinese (ZH), French (FR), German (DE), Japanese (JA), Russian (RU), Spanish (ES), Swahili (SW), Telugu (TE) and Thai (TH). SW, BN, TE, and TH are usually heavily underrepresented languages in pre-trained language models; in PaLM (Chowdhery et al., 2022) they account for less than 0.1% of the pretraining data. We conduct experiments on these ten languages plus English, and measure the reasoning consistency between any two languages.

Model Setup We select a range of open-source state-of-the-art LLMs in three different sizes:

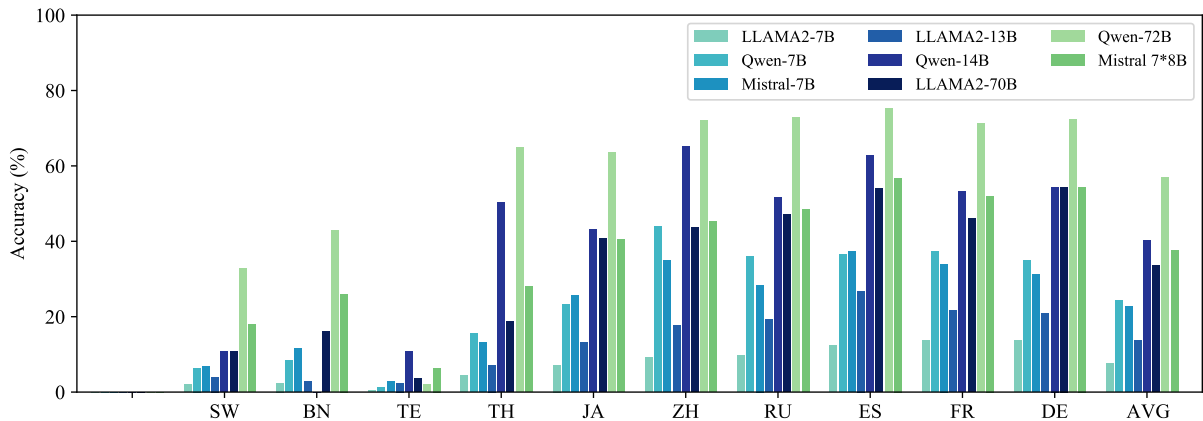
- 7B: LLAMA2 (Touvron et al., 2023); Qwen (Bai et al., 2023); Mistral (Jiang et al., 2023).
- 13-14B: LLAMA2-13B (Touvron et al., 2023); Qwen-14B (Bai et al., 2023).
- 56-72B: Mistral-8×7B²; LLAMA2-70B (Touvron et al., 2023); Qwen-72B (Bai et al., 2023).

We perform few-shot CoT reasoning following Wei et al. (2022a) and Shi et al. (2023), who improve the reasoning task by augmenting few-shot examples with intermediate steps. We use 8-shot for all languages except TE which only uses 2-shot due to the maximum number of input tokens. All CoT prompts in different languages are sourced from the original multilingual CoT reasoning paper (Shi et al., 2023). On the other hand, assuming that we can not access existing math problems with the corresponding reasoning solutions in some languages, a natural and simple way is to use machine translation to translate existing data (e.g., English data) into the target languages. Therefore, we compare not only the consistency between languages, but also the consistency between human-translated (HT) prompts and machine-translated

²<https://mistral.ai/news/mixtral-of-experts/>



(a) Reasoning accuracy (%) using human-written prompt.



(b) Reasoning accuracy (%) using machine-translated prompt.

Figure 2: Accuracy (%) on MGSM of different models with the few-shot method. All machine-translated prompts are translated from English data using Google Translate.

(MT) prompts to understand the impact of translated data on model performance.

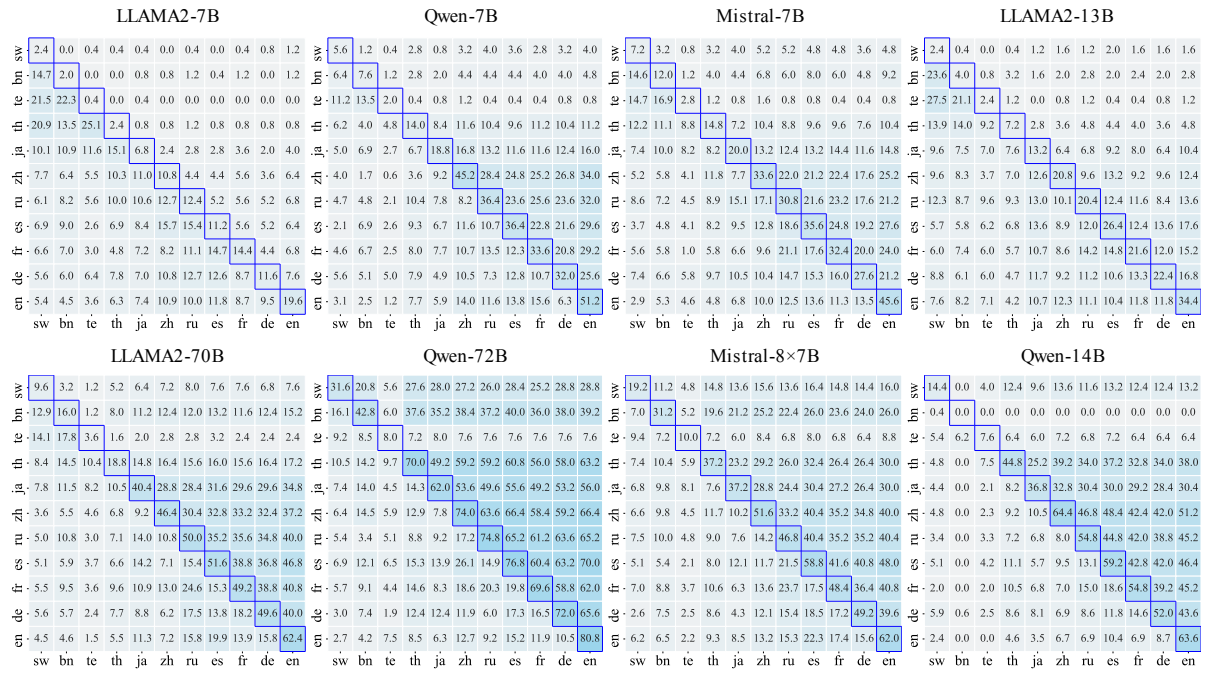
3.2 Results

Model Performance We first compare the performance of different models, as shown in Figure 2.³ Unsurprisingly, and similar to the results reported by Shi et al. (2023), the performance generally improves for all models across all languages as the models scale up in size. However, in contrast to their findings for the large-scale model PaLM-540B, open-source models still yield a performance gap between underrepresented languages (which cover less than 0.1% of the training corpora in PaLM) and high-resource languages. For instance, Qwen-7B scores above 30% in most high-resource languages (e.g., ES and FR) and below 10% in underrepresented languages (e.g., SE, BN, and TE). Particularly, Mistral-8×7B and Qwen-72B achieve

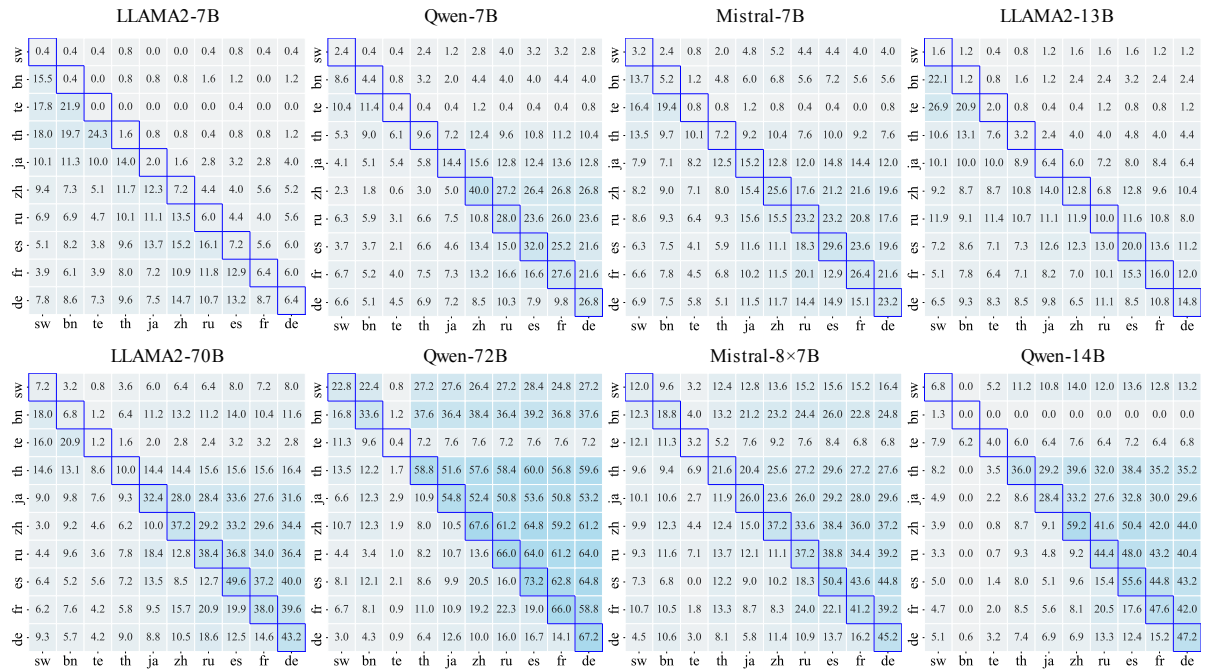
similar or higher scores than PaLM-540B in high-resource languages, while PaLM-540B has better results in underrepresented languages. Regarding HT and MT prompts, we compare all languages except EN, and observe that the results using these two prompts are very close for all models and languages we considered.

Reasoning Consistency We illustrate multilingual reasoning consistency results for different models in Figure 3. When looking at reasoning consistency between language pairs as presented in Figure 3(a), consistent with the trend in model performance, we find overall consistency to be lower for smaller-scale models (7B and 13B) and underrepresented languages (SW, BN, and TE). The correct consistency for all models improves with increasing language representation in pre-training, which is not surprising as higher-resource languages have better accuracy. Incorrect consistency, however, shows a different trend, with higher

³Detailed results are in Appendix A.2.



(a) Reasoning consistency (%) between language pairs in LLMs.



(b) Reasoning consistency (%) between human-translated (row) and machine-translated (column) prompts in LLMs.

Figure 3: Multilingual reasoning consistency. The triangle above the marked diagonal shows the consistency of the models on the correct answers; the triangle below the diagonal contains the consistency between the language pairs where the final answer is the same but incorrect.

scores between underrepresented languages and between high-resource languages in most models, and this occurs even in larger-scale models such as Llama2-70B and Qwen-72B. This suggests that incorrect reasoning knowledge in LLMs is similar to a certain extent between these languages.

To further analyse the impact of using machine-translated prompts, we present reasoning consistency between HT and MT prompts in Figure 3(b). The consistency trend is similar to those in Figure 3(a), with high-resource languages having high consistency on correct answers and underrepre-

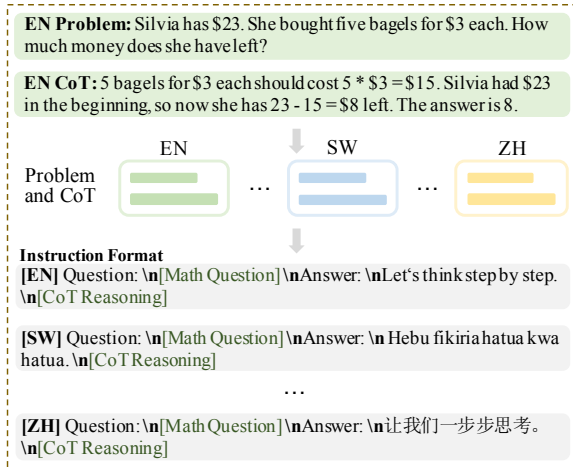


Figure 4: Overview of multilingual CoT reasoning data. English data is first automatically translated into target languages, and then inserted into the templates to construct multilingual instruction data.

sented languages having high consistency on incorrect answers. On the other hand, consistency scores of the larger models are closer to the reasoning accuracy of using human-written prompts, indicating that using machine-translated prompts has less impact on these models.

4 mCoT Instruction Tuning

Based on our findings in multilingual reasoning consistency, we propose a multilingual instruction tuning framework mCoT to supervise the reasoning process, aiming to generate similar reasoning and the same results on inputs expressed in different languages but semantically equivalent. Formally, given a question $q = \{q_1, \dots, q_m\}$ in language x and its corresponding answer – always an integer in this work – the model can generate the step-by-step reasoning solution $s = \{s_1, \dots, s_n\}$ written in language x with the final answer.

4.1 mCoT-MATH

Source Data We leverage two English datasets as source data: MetaMathQA (Yu et al., 2024), which augments the training data of GSM8K and MATH with a question bootstrapping method that rewrites questions using both forward and backward reasoning paths and leverages LLMs to reformulate the question text; and MathInstruct (Yue et al., 2024), which is based on seven existing math rationale datasets annotated by humans or GPT-4 (OpenAI, 2023). This dataset contains two prompt formats: Chain-of-Thought (CoT, Wei et al. 2022b) and Program-of-Thought (PoT, Chen et al.

2023b). We select MathInstruct’s CoT samples and combine them with MetaMathQA, resulting in approximately 580,000 samples.

Automatic Translation Following Shi et al. (2023), we select the ten different languages included in MGSM as the target languages for translation. The overall framework is illustrated in Figure 4. First, we machine-translate⁴ all EN data into the target languages. After translation, we use the instruction format to reformulate all the data to obtain mCoT-MATH, the first large-scale multilingual math CoT dataset, containing around 6.3 million samples. This is expected to facilitate an exhaustive exploration of model reasoning consistency across languages.

4.2 Implementation

We use Mistral-7B as base model, training our mCoT instruction tuning framework using HuggingFace Transformers (Wolf et al., 2020) and DeepSpeed (Rasley et al., 2020). During training, we set the maximum length of the input sequence to 1024, thus reducing GPU memory consumption and improving training speed. We train our model using AdamW optimiser (Loshchilov and Hutter, 2019) with a maximum learning rate of $5e-6$ and a 3% learning rate warmup. The batch size is set to 32 and gradients are accumulated in 4 update steps. We train our model on $4 \times$ NVIDIA A100 40GB GPUs for around 10 days. We report the final answer accuracy for all experiments.

4.3 Evaluation Data

We select two popular multilingual math reasoning datasets, in which each sample contains a question and the corresponding final answer. Specifically, in addition to MGSM, we also include MSVAMP (Chen et al., 2023a), which is constructed based on English data SVAMP (Patel et al., 2021). Chen et al. (2023a) use Google Translate to transform 1,000 questions from the SVAMP test set into nine languages: SW, BN, TH, JA, ZH, RU, ES, FR, and DE. To ensure the translation quality, they back-translate the translated text into English and ask three professional annotators to check semantic consistency manually.

4.4 Baselines

We consider state-of-the-art models of different sizes, both closed and open source.

⁴<https://translate.google.com/>.

Model	SW	BN	TE	TH	JA	ZH	RU	ES	FR	DE	EN
Lang. Freq. (%)	<0.1	<0.1	<0.1	<0.1	0.4	0.4	0.5	2.1	3.3	3.5	78.0
Close-Source Models											
GPT-3 few-shot	11.2	6.4	0.4	0.8	26.0	40.0	28.4	40.4	37.6	36.0	53.6
GPT-3.5-En 2-shot	40.0	7.6	-	15.6	46.8	52.8	50.4	61.2	59.2	62.0	67.2
GPT-4-En 2-shot	64.4	17.6	-	40.4	71.6	70.0	64.0	71.2	72.0	73.6	80.0
PaLM-540B few-shot	35.2	46.0	45.6	52.8	40.0	46.8	48.4	56.8	46.4	49.2	62.4
Open-source Models											
7B Models											
WizardMath (Luo et al., 2023)	3.4	2.0	-	4.0	24.0	22.4	30.8	34.8	30.4	30.4	47.6
MathOctopus (Chen et al., 2023a)	38.4	33.2	-	36.4	35.6	45.2	48.4	45.2	38.0	43.6	54.8
MathOctopus-Mistral (Chen et al., 2023a)	51.6	44.0	-	48.8	48.0	51.6	49.6	53.2	47.2	50.0	58.4
xCoT (Chai et al., 2024)	48.4	40.4	42.8	49.2	50.0	50.0	50.0	48.8	49.6	47.2	48.4
13B Models											
WizardMath (Luo et al., 2023)	5.6	6.4	-	5.6	22.0	28.0	34.4	45.6	42.0	40.4	52.8
MathOctopus (Chen et al., 2023a)	46.0	42.0	-	46.0	39.6	51.2	47.6	53.2	49.6	49.2	51.6
xCoT (Chai et al., 2024)	51.6	50.0	47.2	50.0	49.6	54.0	56.8	54.8	46.4	52.4	54.4
mCoT-7B (ours)	67.2	65.6	62.4	67.6	65.2	64.8	66.8	68.4	63.8	61.2	71.6

Table 1: Multilingual evaluation results (final answer accuracy:%) on the MGSM benchmark. Notes: (i) Lang. Freq. (%) is the language frequency in PaLM training data; (ii) the results of GPT-3 and PaLM-540B are from Shi et al. (2023), while those for GPT-3.5 and GPT-4 are from Chen et al. (2023a); and (iii) in boldface best results per language among closed models and among open models.

Close-Source Models We include six models from three different companies as foundational benchmarks: (i) OpenAI’s GPT-3 (Brown et al., 2020), GPT-3.5⁵, and GPT-4 (OpenAI, 2023); (ii) Anthropic’s Claude-2⁶; and (iii) Google’s PaLM 2 (Chowdhery et al., 2022) and Flan-PaLM (Anil et al., 2023).

Open-Source Models We also compare our model mCoT with several best-performing open-source models for the sake of fairness (including size comparison): (i) WizardMath (Luo et al., 2023); (ii) MathOctopus (Chen et al., 2023a); (iii) xCoT (Chai et al., 2024); and (iv) MetaMath (Yu et al., 2024).

4.5 Results

Evaluation on MGSM Table 1 reports results on the MGSM benchmark. The first observation is that most existing models, including close- and open-source, perform poorly on underrepresented languages such as SW, BN, and TE. For close-source models, GPT models achieve higher accuracy in high-resource languages, with GPT-4 scoring the highest; PaLM-540 achieves competitive performance in all languages, especially reaching the highest score in low-resource languages BN, TE and TH. For open-source models, we observe that

WizardMath achieves less than 7% accuracy on low-resource languages which is explained by the fact that this model is trained on English data, while both MathOctopus and xCoT gain strong improvement with the help of the multilingual instruction dataset. When looking at our model, we see that mCoT significantly outperforms all previous strong baselines, and even outperforms GPT-4 in under-represented languages such as SW, BN, TE, and TH. Particularly, mCoT has higher accuracy scores than PaLM-540B across all languages.

Evaluation on MSVAMP Table 2 reports results on the MSVAMP benchmark. We can observe that GPT-4 with 2-shot achieves the best performance in all languages except BN, where our model achieves the best results. When looking only at open-source models, similar to the observations on MGSM, WizardMath performs poorly in low-resource languages, while our model mCoT shows the highest scores across the board. In particular, mCoT outperforms MathOctopus-Mistral, the one also based on Mistral-7B, confirming that leveraging our dataset mCoT-MATH can yield substantial gains in performance. Finally, we observe that mCoT scores are very close across all languages, suggesting a lesser dependency on the low- vs high-resource aspect.

Reasoning Consistency To further evaluate mCoT’s reasoning capability in different languages, in Figure 5 we present its reasoning consistency

⁵<https://openai.com/blog/chatgpt>.

⁶<https://www.anthropic.com/index/claude-2>.

Model	SW	BN	TH	JA	ZH	RU	ES	FR	DE	EN	AVG
Lang. Freq. (%)	<0.1	<0.1	<0.1	0.4	0.4	0.5	2.1	3.3	3.5	78.0	-
Close-Source Models											
GPT-3.5-En zero-shot	63.2	3.1	24.4	63.3	72.4	62.3	69.5	71.9	66.7	76.1	57.3
GPT-3.5-En 2-shot	68.4	14.4	46.0	74.0	78.4	70.9	74.6	78.2	73.9	81.2	66.0
GPT-4-En 2-shot	75.7	31.2	68.1	74.8	78.9	77.9	81.5	83.9	78.1	80.1	73.0
Open-source Models											
7B Models											
WizardMath (Luo et al., 2023)	10.3	16.1	6.3	26.7	26.8	33.7	42.9	39.9	39.6	45.1	27.0
MathOctopus (Chen et al., 2023a)	42.3	32.8	40.5	43.2	43.2	42.1	44.5	45.3	43.1	46.8	42.4
MathOctopus-Mistral (Chen et al., 2023a)	41.2	36.7	40.2	41.5	43.1	44.0	47.0	49.0	46.4	49.7	43.9
13B Models											
WizardMath (Luo et al., 2023)	12.5	13.7	16.3	29.5	37.0	43.8	50.4	49.4	48.7	56.3	35.8
MathOctopus (Chen et al., 2023a)	43.4	34.2	39.5	43.1	46.4	48.2	48.2	49.9	47.7	44.6	44.5
mCoT-7B (ours)	55.0	53.7	56.4	58.8	58.2	58.1	58.9	58.8	61.1	58.3	57.7

Table 2: Multilingual evaluation results (final answer accuracy:%) on the MSVAMP benchmark. Notes: (i) Lang. Freq. (%) is the language frequency in PaLM training data; (ii) the results of GPT-3.5 and GPT-4 are from Chen et al. (2023a); and (iii) in boldface best results per language among closed models and among open models.

results on the dataset MGSM. After instruction tuning on our dataset mCoT-MATH, correct consistency shows a strong improvement as our model’s reasoning accuracy improves, and in particular we observe that: (i) this is especially true for underrepresented languages; and (ii) the scores for all language pairs are very close, with most being above 50%. It is also interesting to see an increasing trend in incorrect consistency for most language pairs, including low- and high-resource languages, confirming that our model exhibits similar reasoning capabilities across languages. Overall, these observations underscore the efficacy of our method in improving the model’s reasoning consistency.

5 Analysis

Case Study Table 3 shows reasoning solutions generated by our model mCoT for the same math question expressed in different languages. In this case, mCoT incorrectly reasons in the second step in TE: *Yogurts cost \$5.00 each, so Terry spends $60 * 5.00 = 300.00$ on yogurts in 30 days* (EN translation of red background part), which leads to the wrong reasoning and final answer. For other languages, we observe that mCoT generates the correct final answers, while the solutions might be logically different. For example, the focus of FR is to first reason about how many packs of yogurt are needed and then calculate the total cost based on the price of each pack, while for other languages such as EN, the total cost is calculated through the total amount of yogurt and the unit price of each yogurt. Overall, these evidences demonstrate

sw	67.6	56.0	54.4	56.4	56.0	54.8	56.4	60.0	54.4	53.2	60.0
bn	20.4	65.6	52.8	53.6	53.2	54.4	53.6	58.0	52.0	49.6	56.8
te	14.9	12.2	62.4	51.6	50.0	50.4	51.6	55.6	48.8	50.0	56.0
th	18.5	13.2	13.8	67.6	54.4	52.0	56.0	60.0	53.2	50.8	59.2
ja	16.7	13.9	7.8	14.3	65.2	52.0	53.2	56.4	51.6	50.0	57.6
zh	19.0	18.4	12.1	17.8	13.7	64.8	54.4	56.8	50.0	50.8	58.0
ru	13.4	11.8	7.9	19.5	8.2	14.1	66.8	57.6	53.6	52.4	59.6
es	16.2	15.8	14.0	25.0	9.7	18.0	16.1	68.4	55.6	52.8	62.8
fr	15.2	18.1	14.1	21.0	13.5	11.2	23.0	21.3	63.6	50.0	55.2
de	19.3	15.4	14.7	17.0	12.0	15.2	15.7	18.4	20.2	61.2	56.4
en	21.1	14.1	9.9	17.2	14.1	15.3	17.0	25.4	17.5	19.5	71.6
	sw	bn	te	th	ja	zh	ru	es	fr	de	en

Figure 5: Multilingual reasoning consistency of mCoT. The triangle above the marked diagonal shows the consistency of the models on the correct answers; the triangle below the diagonal contains the consistency between the language pairs where the final answer is the same but incorrect.

that mCoT has a good reasoning capability across different languages.

6 Conclusion

We studied multilingual reasoning consistency across multiple languages for popular open-source LLMs such as Mistral and LLAMA2, which provides insights into the evaluation of LLMs. Our findings show that there is a substantial variation across the languages, with lesser resourced ones

Language	CoT Reasoning
Question	Terry eats 2 yogurts a day. They are currently on sale at 4 yogurts for \$5.00. How much does he spend on yogurt over 30 days? Answer: Let's think step by step.
EN	Terry eats 2 yogurts a day, so in 30 days, he will eat $2 \times 30 = 60$ yogurts. The yogurts are on sale at 4 for \$5.00, so each yogurt costs $\$5.00 / 4 = \1.25 . Therefore, Terry will spend $\$1.25 \times 60 = \75.00 on yogurt over 30 days. The answer is 75.
SW	Terry hunywa maziwa 2 ya mtindi kwa siku, kwa hiyo katika siku 30 atakunywa $2 \times 30 = 60$ maziwa ya mtindi. Kwa sasa kuna ofa ya maziwa 4 ya mtindi kwa \$5.00, kwa hivyo kila moja inagharimu $\$5.00 / 4 = \1.25 . Kwa muda wa siku 30, Terry atatumia $\$1.25 \times 60 = \75.00 kwa maziwa ya mtindi. Jibu ni 75.
BN	তেরি দিনে ২টি ইয়োগার্ট খায়, তাই ৩০ দিনের মধ্যে, সে $2 \times 30 = 60$ টি ইয়োগার্ট খাবে। ইয়োগার্ট ৪ প্যাকেজের জন্য \$5.00 এর জন্য বিক্রি হচ্ছে, তাই প্রতিটি ইয়োগার্টের দাম $\$5.00 / 4 = \1.25 । ৩০ দিনের মধ্যে, তেরি ইয়োগার্টের জন্য $60 \times \$1.25 = \75.00 খরচ করবে। উত্তর হল 75।
TE	తెరి ప్రతిరోజూ 2 యోగర్టులు తింటాడు, కాబట్టి 30 రోజుల్లో అతను $2 * 30 = 60$ యోగర్టులు తింటాడు. యోగర్టు ధర ఒక్కొక్కటి \$5.00, కాబట్టి తెరి 30 రోజుల్లో యోగర్టుల కోసం $60 * \$5.00 = \300.00 ఖర్చు చేస్తాడు. సమాధానం 300.
TH	เทรีกินโยเกิร์ต 2 ถ้วยต่อวัน ดังนั้นใน 30 วัน เขาจะกินโยเกิร์ต $2 \times 30 = 60$ ถ้วย โยเกิร์ตถ้วยละ 4 ถ้วยลดราคาอยู่ที่ 5.00 ดอลลาร์ต่อ 4 ถ้วย ดังนั้นเทรีจะจ่ายเงินรวม 5.00 ดอลลาร์ต่อ 4 ถ้วย / 4 = 1.25 ดอลลาร์ต่อถ้วย ดังนั้นเทรีจะจ่าย 60 x 1.25 ดอลลาร์ = 75.00 ดอลลาร์ คำตอบคือ 75
JA	テリーは 1 日に 2 つのヨーグルトを食べるので、30 日後には $2 \times 30 = 60$ 個のヨーグルトを食べることになります。ヨーグルトは 4 個で 5 ドルのセール中なので、ヨーグルト 1 個の価格は $5 \text{ ドル} / 4 = 1.25 \text{ ドル}$ となります。したがって、テリーは 30 日分のヨーグルトに $60 \times 1.25 \text{ ドル} = 75 \text{ ドル}$ を費やすことになります。答えは 75 です。
ZH	特里每天吃 2 份酸奶，因此 30 天他将吃 $2 \times 30 = 60$ 份酸奶。酸奶打折，4 份酸奶售价 5 美元，因此 1 份酸奶售价 $5/4 = 1.25$ 美元。Terry 将在 30 天内花费 $60 \times 1.25 \text{ 美元} = 75.00 \text{ 美元}$ 购买酸奶。答案是 75。
RU	Терри съедает 2 йогурта в день, значит, за 30 дней он съест $2 \times 30 = 60$ йогуртов. Йогурты продаются по цене 5 долларов за 4 штуки, поэтому каждый йогурт стоит $5 \text{ долларов} / 4 = 1,25 \text{ доллара}$. За 30 дней Терри потратит на йогурты $60 \times 1,25 \text{ доллара} = 75 \text{ долларов}$. Ответ: 75.
ES	Terry come 2 yogures al día, por lo que en 30 días comerá $2 \times 30 = 60$ yogures. Actualmente hay una oferta de 4 yogures por \$5, por lo que cada yogur cuesta $\$5/4 = \$1,25$. Durante 30 días, Terry gastará $60 \times \$1,25 = \75 en yogur. La respuesta es 75.
FR	Terry mange 2 yaourts chaque jour, donc sur 30 jours, il mangera $2 \times 30 = 60$ yaourts. Les yaourts sont vendus au prix de 4 pour 5,00 \$, donc Terry devra acheter $60/4 = 15$ ensembles de yaourts. Chaque ensemble coûte 5,00 \$, donc Terry dépense $15 \times 5,00 \$ = 75,00 \$$ en yaourt sur 30 jours. La réponse est 75.
DE	Terry isst 2 Joghurts pro Tag, also isst sie in 30 Tagen $2 \times 30 = 60$ Joghurts. Die Joghurts sind für 5,00 \$ für 4 Stück im Angebot, also kostet 1 Joghurt $5,00 \$/4 = 1,25 \$$. In 30 Tagen wird Terry $60 \text{ Joghurts zu je } 1,25 \$$ ausgeben, also $60 \times 1,25 \$ = 75,00 \$$. Die Antwort ist 75.

Table 3: Case study in the test set of MSGM. Note that here we only show an EN question, the questions corresponding to each output are written in their respective languages.

substantially underperforming. To address this issue, we constructed the first large-scale multilingual math reasoning instruction dataset mCoT-MATH, with around 6.3 million samples in eleven diverse languages. We then introduced a multilingual reasoning instruction tuning framework to train our model mCoT on mCoT-MATH. Evaluation on two multilingual benchmark datasets shows that our 7B parameter model achieves impressive reasoning consistency across all languages, and comparable or superior performance to close- and open-source state-of-the-art models even of much larger sizes.

7 Limitations

In this work we investigated multilingual reasoning consistency across 11 languages and eight open-source models in different sizes, but there are very many more languages and LLMs that can still bring substantial challenges and insights if considered.

In addition, when considering more languages in the future, it will also be interesting to consider language families as factor. On the other hand, this work focused on reasoning consistency based on the final answer, while the consistency of intermediate reasoning steps is definitely an interesting direction. Reasoning solutions are not necessarily consistent across languages (even within the same language), since they might be logically different but still result in the same and correct answer. Therefore, automatically assessing intermediate steps is challenging and requires more explorations in the future. Regarding the dataset, while mCoT-MATH boosts the reasoning capability across languages, screening high-quality machine-translated data could further improve the model. Finally, the full potential of our approach could be further explored by for example extending instruction tuning with reward learning to encourage models to generate more diverse solutions.

Acknowledgments

The anonymous reviewers of ACL Rolling Review provided us with useful comments which contributed to improving this paper and its presentation, so we're grateful to them. We would also like to thank the SURF organisation and the Center for Information Technology of the University of Groningen for their support and for providing access to the high-performance computing clusters Snellius and Hábrók, respectively.

References

- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. [Palm 2 technical report](#). *arXiv preprint arXiv:2305.10403*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Sheng-guang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *arXiv preprint arXiv:2309.16609*.
- 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.
- Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. [xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning](#). *arXiv preprint arXiv:2401.07037*.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023a. [Breaking language barriers in multilingual mathematical reasoning: Insights and observations](#). *arXiv preprint arXiv:2310.20246*.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023b. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#). *Transactions on Machine Learning Research*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim,

- Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#). *arXiv preprint arXiv:2204.02311*.
- Zheng Chuanyang, Liu Zhengying, Xie Enze, Li Zhen-guo, 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*.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhishava Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- 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, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023a. [Specializing smaller language models towards multi-step reasoning](#). In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 10421–10430.
- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023b. [Complexity-based prompting for multi-step reasoning](#). In *The Eleventh International Conference on Learning Representations*.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. [Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.
- Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. [Large language models can self-improve](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *arXiv preprint arXiv:2310.06825*.
- 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*.
- Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. [GRACE: Discriminator-guided chain-of-thought reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15299–15328, Singapore. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Sriku-mar. 2019. [A logic-driven framework for consistency of neural models](#). 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 3924–3935, Hong Kong, China. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. [Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct](#). *arXiv preprint arXiv:2308.09583*.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. [Teaching small language models to reason](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. 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.
- Ansong Ni, Jeevana Priya Inala, Chenglong Wang, Alex Polozov, Christopher Meek, Dragomir Radev, and Jianfeng Gao. 2023. [Learning math reasoning from self-sampled correct and partially-correct solutions](#). In *The Eleventh International Conference on Learning Representations*.
- Sean O’Brien and Mike Lewis. 2023. [Contrastive decoding improves reasoning in large language models](#). *arXiv preprint arXiv:2309.09117*.
- OpenAI. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP models really able to solve simple math word problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. [Cross-lingual consistency of factual knowledge in multilingual language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. [Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 3505–3506, New York, NY, USA. Association for Computing Machinery.
- 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.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, 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. 2023. [Language models are multilingual chain-of-thought reasoners](#). In *The Eleventh International Conference on Learning Representations*.
- Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. [Distilling reasoning capabilities into smaller language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. [Solving math word problems with process- and outcome-based feedback](#). *arXiv preprint arXiv:2211.14275*.
- 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. 2022a. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

- Quentin Lhoest, and Alexander Rush. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [Meta-math: Bootstrap your own mathematical questions for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. 2024. [MAMmoTH: Building math generalist models through hybrid instruction tuning](#). In *The Twelfth International Conference on Learning Representations*.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [STar: Bootstrapping reasoning with reasoning](#). In *Advances in Neural Information Processing Systems*.
- 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*.
- 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*.
- Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaying Zhang, and Yujie Yang. 2023. [Solving math word problems via cooperative reasoning induced language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4471–4485, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Prompt Examples

Question: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
Step-by-Step Answer: 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. Five more computers were installed each day, from Monday to Thursday. How many computers are now in the server room?

Step-by-Step Answer: There are 4 days from Monday to Thursday. 5 computers were added each day. That means in total $4 * 5 = 20$ computers were added. There were 9 computers in the beginning, so now there are $9 + 20 = 29$ computers. The answer is 29.

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

Step-by-Step Answer: Leah had 32 chocolates and Leah's sister had 42. That means there were originally $32 + 42 = 74$ chocolates. 35 have been eaten. So in total they still have $74 - 35 = 39$ chocolates. The answer is 39.

Question: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

Step-by-Step Answer: He has 5 toys. He got 2 from mom, so after that he has $5 + 2 = 7$ toys. Then he got 2 more from dad, so in total he has $7 + 2 = 9$ toys. The answer is 9.

Question: Michael had 58 golf balls. On Tuesday, he lost 23 golf balls. On Wednesday, he lost 2 more. How many golf balls did he have at the end of Wednesday?

Step-by-Step Answer: Michael started with 58 golf balls and lost 23, so he has $58 - 23 = 35$. After he lost 2 more, he has $35 - 2 = 33$ balls now. The answer is 33.

Question: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

Step-by-Step Answer: 5 bagels for \$3 each should cost $5 * 3 = 15$ dollars. Olivia had \$23 in the beginning, so now she has $23 - 15 = 8$ dollars left. The answer is 8.

Question: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

Step-by-Step Answer: Jason started with 20 lollipops, but now he only has 12, so he gave Denny $20 - 12 = 8$ lollipops. The answer is 8.

Question: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

Step-by-Step Answer: There are 3 cars in the beginning, 2 more arrive, so now there should be $3 + 2 = 5$ cars. The answer is 5.

Question: [Math Question]
Step-by-Step Answer:

Frage: Roger hat 5 Tennisbälle. Er kauft noch 2 Dosen Tennisbälle. In jeder Dose sind 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Schritt-für-Schritt-Antwort: Roger begann mit 5 Bällen. 2 Dosen von jeweils 3 Tennisbällen macht 6 Tennisbälle. $5 + 6 = 11$. Die Antwort ist 11.

Frage: Es waren neun Computer im Serverraum. Von Montag bis Donnerstag wurden jeden Tag noch fünf Computer installiert. Wie viele Computer sind jetzt im Serverraum?

Schritt-für-Schritt-Antwort: Von Montag bis Donnerstag sind es 4 Tage. Jeden Tag kamen 5 neue Computer hinzu. Das macht insgesamt $4 * 5 = 20$ Computer, die hinzugefügt wurden. Am Anfang waren es 9 Computer, also sind es jetzt $9 + 20 = 29$ Computer. Die Antwort lautet 29.

Frage: Leah hat 32 Pralinen und ihre Schwester hat 42. Wenn sie 35 essen, wie viele sind dann insgesamt noch übrig?

Schritt-für-Schritt-Antwort: Leah hat 32 Pralinen und Leahs Schwester hat 42. Das bedeutet, dass es ursprünglich $32 + 42 = 74$ Pralinen waren. 35 wurden gegessen. Also haben sie insgesamt noch $74 - 35 = 39$ Pralinen übrig. Die Antwort lautet 39.

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten hat er von seiner Mama und seinem Papa jeweils zwei Spielzeuge bekommen. Wie viele Spielzeuge hat er jetzt?

Schritt-für-Schritt-Antwort: Er hat 5 Spielzeuge. Er hat 2 von seiner Mama bekommen, sodass er nun $5 + 2 = 7$ Spielzeuge hat. Dann hat er noch 2 von seinem Papa bekommen, also hat er insgesamt $7 + 2 = 9$ Spielzeuge. Die Antwort lautet 9.

Frage: Michael hat 58 Golfbälle. Am Dienstag hat er 23 Golfbälle verloren. Am Mittwoch hat er 2 weitere verloren. Wie viele Golfbälle hat er Mittwoch am Ende des Tages?

Schritt-für-Schritt-Antwort: Michael hatte anfangs 58 Golfbälle und hat 23 verloren, sodass er $58 - 23 = 35$ hat. Nachdem er 2 weitere verloren hat, hat er jetzt $35 - 2 = 33$ Bälle. Die Antwort lautet 33.

Frage: Olivia hat 23 US-Dollar. Sie hat fünf Bagels für 3 US-Dollar pro Stück gekauft. Wie viel Geld hat sie übrig?

Schritt-für-Schritt-Antwort: 5 Bagels für 3 US-Dollar pro Stück kosten $5 * 3 = 15$ Dollar. Olivia hat anfangs 23 US-Dollar, also hat sie jetzt $23 - 15 = 8$ Dollar übrig. Die Antwort lautet 8.

Frage: Jason hatte 20 Lutscher. Er hat Denny einige Lutscher gegeben. Jetzt hat Jason 12 Lutscher. Wie viele Lutscher hat Jason Denny gegeben?

Schritt-für-Schritt-Antwort: Jason hat mit 20 Lutschern angefangen, aber jetzt hat er nur 12, also hat er Denny $20 - 12 = 8$ Lutscher abgegeben. Die Antwort lautet 8.

Frage: Wenn 3 Autos auf dem Parkplatz stehen und 2 weitere Autos ankommen, wie viele Autos stehen dann auf dem Parkplatz?

Schritt-für-Schritt-Antwort: Anfangs sind 3 Autos da, 2 weitere kommen an, also sind jetzt $3 + 2 = 5$ Autos da. Die Antwort lautet 5.

Frage: [Math Question]
Schritt-für-Schritt-Antwort:

Frage: Roger hat 5 Tennisbälle. Er kauft noch zwei Dosen Tennisbälle. Jede Dose enthält 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt?

Schritt-für-Schritt-Antwort: Roger begann mit 5 Bällen. 2 Dosen mit je 3 Tennisbällen ergeben 6 Tennisbälle. $5 + 6 = 11$. Die Antwort ist 11.

Frage: Im Serverraum befanden sich neun Computer. Jeden Tag, von Montag bis Donnerstag, wurden fünf weitere Computer installiert. Wie viele Computer stehen jetzt im Serverraum?

Schritt-für-Schritt-Antwort: Es gibt 4 Tage von Montag bis Donnerstag. Jeden Tag kamen 5 Computer hinzu. Das bedeutet, dass insgesamt $4 * 5 = 20$ Computer hinzugefügt wurden. Am Anfang gab es 9 Computer, jetzt sind es $9 + 20 = 29$ Computer. Die Antwort ist 29.

Frage: Leah hatte 32 Pralinen und ihre Schwester hat 42. Wenn sie 35 gegessen hätten, wie viele Stücke wären dann insgesamt übrig geblieben?

Schritt-für-Schritt-Antwort: Leah hatte 32 Pralinen und Leahs Schwester hatte 42. Das bedeutet, dass es ursprünglich $32 + 42 = 74$ Pralinen waren. 35 wurden gegessen. Insgesamt haben sie also noch $74 - 35 = 39$ Pralinen. Die Antwort ist 39.

Frage: Shawn hat fünf Spielzeuge. Zu Weihnachten bekam er von seiner Mutter und seinem Vater jeweils zwei Spielzeuge. Wie viele Spielsachen hat er jetzt?

Schritt-für-Schritt-Antwort: Er hat 5 Spielzeuge. Er hat 2 von Mama bekommen, also hat er danach $5 + 2 = 7$ Spielzeuge. Dann bekam er noch zwei weitere von Papa, also hat er insgesamt $7 + 2 = 9$ Spielzeuge. Die Antwort ist 9.

Frage: Michael hatte 58 Golfbälle. Am Dienstag verlor er 23 Golfbälle. Am Mittwoch verlor er zwei weitere. Wie viele Golfbälle hatte er am Ende des Mittwochs?

Schritt-für-Schritt-Antwort: Michael begann mit 58 Golfbällen und verlor 23, also hat er $58 - 23 = 35$. Nachdem er 2 weitere verloren hat, hat er jetzt $35 - 2 = 33$ Bälle. Die Antwort ist 33.

Frage: Olivia hat 23 \$. Sie kaufte fünf Bagels für jeweils 3 Dollar. Wie viel Geld bleibt ihr übrig?

Schritt-für-Schritt-Antwort: 5 Bagels für jeweils 3 US-Dollar sollten $5 * 3 = 15$ US-Dollar kosten. Olivia hatte am Anfang 23 Dollar, jetzt hat sie also $23 - 15 = 8$ Dollar übrig. Die Antwort ist 8.

Frage: Jason hatte 20 Lutscher. Er gab Denny ein paar Lutscher. Jetzt hat Jason 12 Lutscher. Wie viele Lutscher hat Jason Denny gegeben?

Schritt-für-Schritt-Antwort: Jason begann mit 20 Lutschern, aber jetzt hat er nur noch 12, also gab er Denny $20 - 12 = 8$ Lutscher. Die Antwort ist 8.

Frage: Wenn 3 Autos auf dem Parkplatz stehen und 2 weitere Autos ankommen, wie viele Autos sind dann auf dem Parkplatz?

Schritt-für-Schritt-Antwort: Am Anfang stehen 3 Autos, 2 weitere kommen hinzu, also sollten es jetzt $3 + 2 = 5$ Autos sein. Die Antwort ist 5.

Frage: [Math Question]
Schritt-für-Schritt-Antwort:

(a) English CoT prompt.

(b) HT German CoT prompt.

(c) MT German CoT prompt.

Figure 6: CoT prompt template: mathematical questions are inserted in square brackets, and the model generates corresponding CoT reasoning.

A.2 Multilingual reasoning results on MGSM

Language Model	Prompt	EN	SW	BN	TE	TH	JA	ZH	RU	ES	FR	DE	AVG
Lang. Freq. (%)	-	78.0	<0.1	<0.1	<0.1	<0.1	0.4	0.4	0.5	2.1	3.3	3.5	-
COMET Score	-	-	84.6	87.3	89.7	81.3	86.3	89.2	86.3	87.9	88.5	88.8	-
7B Models													
LLAMA2	HT	19.6	2.4	2.0	0.4	2.4	6.8	10.8	12.4	11.2	14.4	11.9	7.5
	MT	-	2.0	2.4	0.4	4.4	7.2	9.2	9.6	12.4	13.6	13.6	7.5
Qwen	HT	51.2	5.6	7.6	2.0	14.0	18.8	45.2	36.4	36.4	33.6	32.0	22.1
	MT	-	6.4	8.4	1.2	15.6	23.2	44.0	36.0	36.4	37.2	34.8	24.3
Mistral	HT	45.6	7.2	12.0	2.8	14.8	20.0	33.6	30.8	35.6	32.4	27.6	21.7
	MT	-	6.8	11.6	2.8	13.2	25.6	34.8	28.4	37.2	34.0	31.2	22.6
13-14B Models													
LLAMA2-13B	HT	34.4	2.4	4.0	2.4	7.2	13.2	20.8	20.4	26.4	21.6	22.4	14.1
	MT	-	4.0	2.8	2.4	7.2	13.2	17.6	19.2	26.8	21.6	20.8	13.6
Qwen-14B	HT	63.6	14.4	0.0	7.6	44.8	36.8	64.4	54.8	59.2	54.8	52.0	38.9
	MT	-	10.8	0.0	10.8	50.4	43.2	65.2	51.6	62.8	53.2	54.4	40.2
>65B Models													
LLAMA2-70B	HT	62.4	9.6	16.0	3.6	18.8	40.4	46.4	50.0	51.6	49.2	49.6	33.5
	MT	-	10.8	16.0	3.6	18.8	40.8	43.6	47.2	54.0	46.0	54.4	33.5
Qwen-72B	HT	80.8	31.6	42.8	8.0	70.0	62.0	74.0	74.8	76.8	69.6	72.0	58.2
	MT	-	32.8	42.8	2.0	64.8	63.6	72.0	72.8	75.2	71.2	72.4	57.0
Mistral-8×7B	HT	62.0	19.2	31.2	10.0	37.2	37.2	51.6	46.8	58.8	48.4	49.2	39.0
	MT	-	18.0	26.0	6.4	28.0	40.4	45.2	48.4	56.8	52.0	54.4	37.6
PaLM-540B [†]	HT	62.4	35.2	46.0	45.6	52.8	40.0	46.8	48.4	56.8	46.4	49.2	48.1

Table 4: Accuracy (%) on MGSM of different models with the few-shot method. Notes: (i) Lang. Freq. (%) is the language frequency in PaLM training data; (ii) we report COMET (Rei et al., 2020) score between the HT and MT prompts; (iii) average (AVG) scores do not include EN results; (iv) [†]: Results from Shi et al. (2023).