Reasoning Under a Hierarchy of Perturbations: A Systematic Study with Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have proved able to address complex reasoning tasks, yet they are highly sensitive to input variations, often leading to divergent solution paths and final answers. Despite the awareness of this issue, we still lack a systematic understanding of how different types of input perturbations affect reasoning behaviour. To address this gap, we introduce a hierarchy of perturbations affecting the input with increasing modification depth: order-level (the order of shots is varied), phrasal-level (the input is rephrased), and language-level (the input is presented in different languages). This hierarchy enables a fine-grained analysis of how variations across representational levels influence reasoning. Experiments on multiple math datasets show that reasoning consistency declines as perturbations operate at deeper levels, revealing the fragility of current LLM reasoning. Additionally, since different perturbation types interact in non-trivial ways, leveraging their combination can improve reasoning accuracy. In particular, we propose leveraging cross-lingual consistency derived from language perturbation, extending self-consistency to the multilingual setting and yielding consistent performance gains. These findings provide insights into how input perturbations shape reasoning behaviour in LLMs, and suggest practical strategies for improving robust reasoning.

1 Introduction

It is by now established that Large Language Models (LLMs) have impressive abilities in addressing a variety of complex reasoning tasks (Brown et al., 2020; Bommasani et al., 2022). Notably, 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). Building on this reasoning paradigm, a variety of (long) Chain-of-X approaches have been proposed to address challenges across a wide range of domains (Xia et al., 2025; Guo et al., 2025).

Yet, LLMs have also proved to be sensitive and somewhat brittle with respect to input variations (Zhao et al., 2021; Lu et al., 2022). Minor changes, such as context structure and wording, might lead to different reasoning paths and even different final answers. This sensitivity raises a fundamental question about the nature of reasoning: *How do different levels of perturbations affect both reasoning stability and performance in LLMs?*

Prior work has explored specific forms of variation. For instance, in a few-shot setting, just altering the order in which the example shots are provided might lead to different reasoning answers (Wang et al., 2022); the same can happen if different formulations of the same math 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). These studies, however, focus on a single perturbation type in isolation and on possibly different datasets. Particularly, not all input variations operate at the same structural level: some changes merely adjust formatting, while others transform the encoding of the problem more structurally and substantially. As a result, we lack a systematic understanding of how reasoning consistency behaves across perturbations of different structural depth, and how such perturbations differ in their impact on reasoning stability and performance.

In this paper, we present a comprehensive empirical study of reasoning behavior in LLMs under three systematically controlled perturbation levels: (i) order-level perturbation, induced by varying the order of few-shot exemplars; (ii) phrasal-level perturbation, induced by paraphrasing the problem within the same language; and (iii) language-level perturbation, induced by expressing the same prob-

lem in different natural languages.

These three perturbation levels form a natural hierarchy of increasing structural depth: from changes in input formatting to linguistic reformulation, and then to full linguistic system variation. By studying how reasoning behaviour evolves as perturbations move from surface to deeper levels, at the same time and on multiple datasets, we provide a unified empirical perspective on robustness and diversity in LLM reasoning.

We conduct extensive experiments on both monolingual and multilingual mathematical reasoning benchmarks, covering models of varying sizes (7B–72B). Our results and analysis yield the following key findings:

- **Reasoning consistency declines with perturbation depth.** Models are highly stable under order-level changes, moderately sensitive to rephrasing, and substantially less consistent under cross-lingual variation.
- **Consistency is influenced by dataset difficulty and model characteristics.** Models exhibit lower consistency on more challenging datasets, while larger models generally achieve higher agreement rates.
- **Consistency correlates with accuracy, and aggregating perturbations improves performance.** Perturbation agreement across ordering and phrasing aligns well with accuracy, serving as a practical proxy for reasoning reliability. Furthermore, we propose leveraging cross-lingual consistency derived from language perturbation to improve model performance, especially when combined with other types.

Our findings are expected to provide a foundation for future research on perturbation-based robustness, multilingual reasoning, and consistency-driven analysis of LLM reasoning.

2 Related Work

Math Reasoning in LLMs. Mathematical reasoning has garnered great interest 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 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' capability, Wei et al. (2022) propose Chain-of-Thought prompting that break down a problem into multiple intermediate steps, gradually arriving at the final answer. Given its success, a series of related methods have been proposed to improve reasoning performance, such as complex CoT (Fu et al., 2023), auto-CoT (Zhang et al., 2023), multilingual CoT (Shi et al., 2023), least-to-most prompting (Zhou et al., 2023), progressive-hint prompting (Chuanyang et al., 2023), and residual connection prompting (Jiang et al., 2024).

Consistency under Input Perturbations. In principle, language models could be expected to yield consistent answers in semantically equivalent contexts. In practice, this is often not the case. Recent studies have shown that LLM predictions can vary under minor input change (Zhao et al., 2021; Lu et al., 2022). For instance, Wang et al. (2023) treat variation as a stochastic tool, sampling diverse solution paths and then selecting the most consistent answer. Beyond sampling strategies, Wang et al. (2022) investigate prompt sensitivity, demonstrating that reordering few-shot exemplars yield different outputs. Zhou et al. (2024) explore question rephrasing as a means to improve performance via answer aggregation. In the multilingual context, Lai and Nissim (2024) show that expressing the same problem in different languages may induce distinct reasoning paths. For perturbation analysis, Chatziveroglou et al. (2025) study LLM reasoning under prompt variations on a single dataset, focusing on irrelevant or pathological context.

Our work differs from prior research in three aspects. First, we propose a hierarchical perturbation framework that affects the input with increasing modification depth. Second, we perform a controlled empirical comparison of different perturbation types across different model scales and task difficulties. Finally, we suggest strategies for integrating answers across different perturbation levels to enhance overall performance.

3 Perturbation Framework

Figure 1 shows our overall perturbation framework. Using systematic input perturbations, the model generates the corresponding solutions and answers for a given question, which are used to measure the model's consistency and performance.

3.1 Formalising Perturbation Depth

To systematically study how input variations affect reasoning behaviour, we formalise the notion of perturbation depth. Intuitively, perturbations can

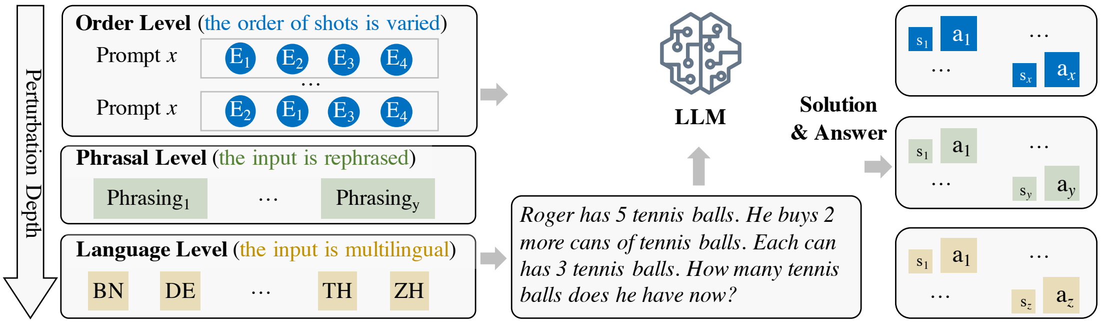


Figure 1: Overview of our multi-level perturbation framework: order-level (the order of shots is varied), phrasal-level (the question is rephrased), and language-level (the question is presented in different languages).

modify the representation of an input to different extents while preserving the semantic content of the underlying reasoning task. We define perturbation depth as the degree of representational change induced by a perturbation.

Let $x \in \mathcal{X}$ denote a model input, and let $P(x)$ denote a semantically equivalent perturbation of x. We operationalise perturbation depth using the embedding distance between the original and perturbed inputs. Specifically, we encode both x and $P(x)$ and compute their cosine distance:

$$D(P) = \mathbb{E}_{x \sim \mathcal{X}} \left[1 - \cos(x, P(x)) \right], \quad (1)$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. Higher values of $D(P)$ indicate larger representational shifts induced by the perturbation operator. This allows us to empirically compare perturbation levels through a unified representational metric.

3.2 Perturbation Hierarchy

We introduce the following three perturbations levels: order, phrasal, and language.

Order-level Perturbation. Problem phrasing stays the same, but the model sees the shots in different orders. 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 exemplars in few-shot prompting (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 (`question`:q, `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).

Phrasal-level Perturbation. Problem gets rewritten with different phrasings. In addition to the order of the exemplars in the prompt, the wording and expression form of the math problem itself can also have an impact on the performance of the model. Therefore, we assess the model's consistency in terms of *phrasing*. Differently from Zhou et al. (2024), who select 'good' paraphrases based on solving rate improvement as the few-shot exemplars, we directly prompt the language model to rewrite the question with the goal of making it easier for itself to solve. As shown in Table 1, we use two main settings: rewrite-without-solve and rewrite-then-solve, which finally yield the four different settings when combined with the original question.

Language-level Perturbation. Problem gets rewritten in different languages. 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). Cross-lingual variation induces changes in the linguistic system beyond the surface forms. These transformations may lead to substantially different internal representations and reasoning decompositions. However, given the same math problem in different languages, the generated reasoning solutions could still be expected to arrive at the same final answer if the language model is multilingual.

Setting	Description	Rewrite?
RwS	The model first rewrites the question without solving it. Then a second prompt asks for the solution to the rewritten question.	Yes
RwS+	The original question and the rewritten version above are concatenated before prompting the model to generate the solution.	No
RtS	The model rewrites the question to make it easier to solve and immediately provides the corresponding solution in the same prompt.	Yes
RtS-	The model is prompted to generate the solution for the rewritten question in the 'RtS' setting.	No

Table 1: Different settings used in phrasal-level perturbation. Examples are in Appendix A.1.

These perturbations form a hierarchy from shallow to deeper input modifications, corresponding to variations at the ordering, phrasing, and language levels. This hierarchy provides a unified framework for analysing the effect of perturbations on LLM reasoning behaviour. We hypothesise that perturbations with greater structural depth introduce larger variation in internal reasoning representations, which may reduce model consistency but increase diversity.

3.3 Consistency and Aggregation

We introduce two complementary perspectives: reasoning consistency and reasoning aggregation, to study the impact of perturbations.

Measuring Reasoning Consistency. To quantify reasoning stability under perturbations, we propose to measure reasoning consistency (RC). Formally, given a set of math problems $\mathcal{Q}$, each consisting of a two-tuple (`question`:q, `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 perturbation level. 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.

$$\mathrm{RC} = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{\max_j |\mathcal{S}_j|}{n}, \qquad (2)$$
$$\mathcal{S}_j = \{s_i \in \{s_1, \ldots, s_n\} | f(s_i) = a_j\},$$

where $f(s_i)$ maps solution s_i to the final answer. This metric can capture the model's stability under controlled variation and enable direct comparison across perturbations, models, and datasets.

Leveraging Reasoning Aggregation. We further examine how aggregation across perturbations affects model performance. To this end, we leverage answer consistency across three perturbation levels to improve overall reasoning performance. For each question, the language model produces the corresponding solution set $\{s_1, \ldots, s_n\}$ and the final answer set $\{a_1, \ldots, a_m\}$. We select the most consistent answer in n solution paths as the final answer $\hat{a}$, which is obtained through majority voting:

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

where $\mathcal{A}$ denotes the set of candidate answers and $\mathbb{I}(\cdot)$ is the indicator function. Aggregation serves not only as a test-time technique but also as an analytical tool to study the trade-off between stability and accuracy. By comparing individual perturbation results with aggregated outcomes, we can evaluate how different levels of structural variation contribute to complementary reasoning signals.

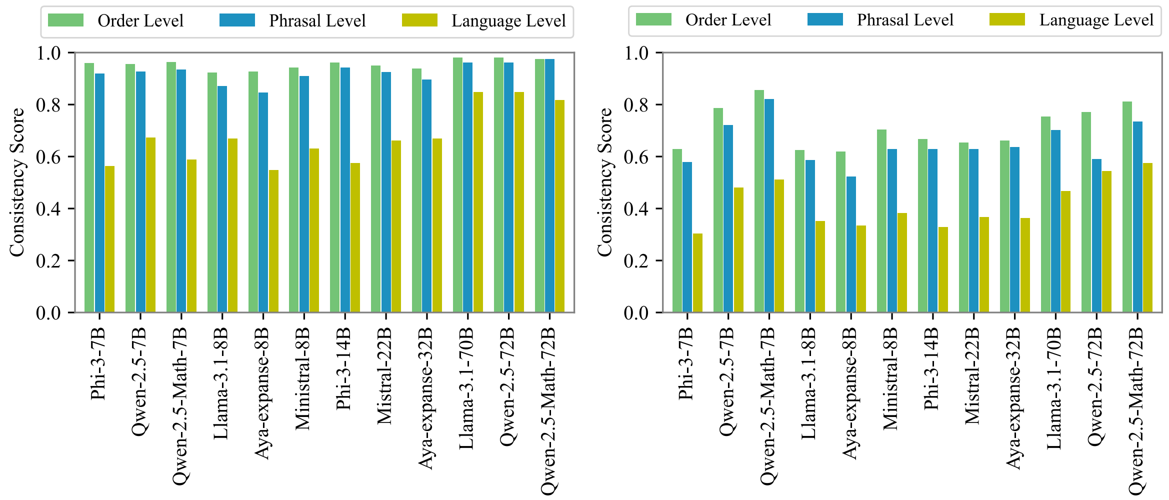
4 Experimental Setup

Datasets. To comprehensively study our framework at different perturbation depths, 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., 2023). To test language-level perturbation on the more challenging dataset, we machine-translate questions from MATH 500 into 10 languages in MGSM.[1]

Models. We select a range of open-source LLMs in varying scales: (i) 7-8B; (ii) 14-35B; and (iii) 70-72B.[2] We use Qwen3-Embedding-8B (Zhang et al., 2025) to measure the perturbation depth.

Implementation. In the multilingual scenario, we use 4-shot for all languages except TE which uses 2-shot, since a 4-shot prompt would exceed the default maximum length, due to tokenisation

[1]See Appendix A.2 for more details on datasets.
[2]See Appendix A.3 for more details on models.



(a) Reasoning consistency on MGSM.

(b) Reasoning consistency on MATH 500.

Figure 2: Comparison of reasoning consistency across three perturbation levels.

Perturbation	MGSM	MATH500
Order Level	0.025	0.031
Phrasal Level	0.043	0.055
Language Level	0.169	0.163

Table 2: Average perturbation depth scores.

issues unfavourable to this language (Ahia et al., 2023). We report the final answer accuracy for all experiments except the consistency score.[3]

5 Results and Analysis

5.1 A Hierarchy of Input Perturbations

Before analysing models' reasoning behaviour in different perturbation levels, we first verify whether the proposed perturbation hierarchy corresponds to increasing representational change in practice.

Table 2 reports average perturbation depth scores across perturbation levels on datasets MGSM and MATH500.[4] We observe a clear increasing trend in representational distance from order to phrasal and finally language-level perturbations. These results empirically support our perturbation hierarchy and validate the use of perturbation depth as a meaningful representational measure. More broadly, they suggest that the three perturbation types correspond to progressively deeper transformations of the input space, rather than merely qualitatively different forms of variation.

[3]See Appendix A.4 for detailed implementation.

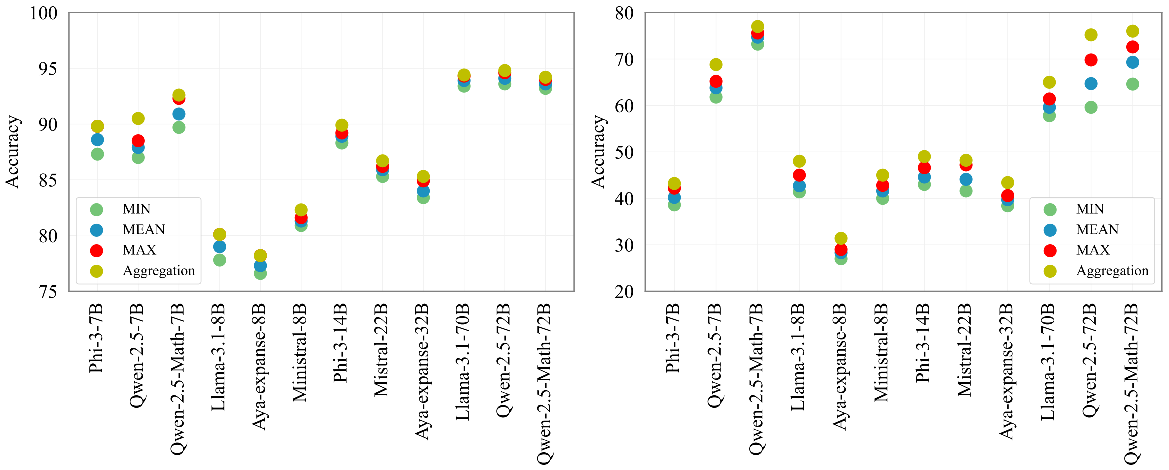
[4]See Appendix A.5 for detailed results on perturbation depth.

5.2 Consistency Landscape Across Perturbation Depth

Figure 2 shows the agreement rate among model predictions when the same problem is presented through different perturbed inputs.

Consistency decreases as perturbations become structurally deeper. When only the order of the shot is varied, models maintain very high consistency scores, typically above 0.9. This suggests that reasoning behaviour is largely stable under superficial changes. In contrast, phrasal perturbations, i.e., rephrasing the math problem, lead to slight decreases in agreement. Different phrasings may alter syntactic structures to highlight distinct aspects of the problem, which can affect how the model performs the reasoning process. The largest variability arises when the same problem is expressed in different languages. Language perturbations induce changes in the linguistic system beyond the surface forms, potentially leading to different internal representations and reasoning trajectories. As a result, predictions under these perturbations are substantially less consistent.

Dataset difficulty and model scale also play important roles. Unsurprisingly, compared to MGSM, models achieve lower consistency scores on the more challenging dataset MATH 500. Larger models generally exhibit higher agreement scores across perturbation levels, suggesting that increased capacity improves reasoning stability. However, on dataset MATH 500, Qwen-2.5-Math-7B achieves the highest score, suggesting that fac-



(a) Model performance on GSM8K. (b) Model performance on MATH 500.

Figure 3: Reasoning accuracy of 4-shot for 8 different exemplars orders. Note that MIN, MEAN, and MAX denote the minimum, mean, and maximum scores, respectively.

tors beyond sheer model size contribute to performance; additionally, there is a bigger gap in language perturbation for different models, even within the same scale, e.g., Phi-3-7B vs Qwen-2.5-7B, indicating that multilingual robustness remains uneven across models.

Overall, these results reveal a consistency hierarchy aligned with perturbation depth, where deeper transformations of the input representation lead to more diverse reasoning outcomes.

5.3 Performance Gains from Perturbation Aggregation

We evaluate whether perturbation-induced diversity can improve reasoning accuracy when predictions from multiple perturbed inputs are aggregated, including single perturbation and across types.

Order-level Perturbation. Figure 3 reports the reasoning performance on datasets GSM8K and MATH 500, where we use 8 different exemplar orders for the 4-shot prompt.[5] The first observation is that aggregation consistently improves reasoning performance across all models compared to vanilla CoT prompting. When looking at each dataset individually, on GSM8K the accuracy obtained using the aggregated results exceeds the average of the 8 different prompt orderings for all models and reaches the highest scores for most models. On the more challenging MATH 500 dataset, aggregation attains the highest scores across all models. This indicates that while most models exhibit strong inherent robustness to prompt ordering (as shown in the consistency score in Section 5.2), shallow prompt variation still provides a consistent improvement across models and task difficulties.

Phrasal-level Perturbation. Table 3 reports results on phrasal-level perturbation. We observe that 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 vanilla prompting and tend to achieve higher scores in the rewrite-then-solve setting (RtS). The latter observation suggests that asking the model to first simplify the question before attempting a solution can help improve reasoning effectiveness, likely by reducing ambiguity and guiding the model's thinking. Lastly, aggregation consistently enhances reasoning performance beyond these settings. These findings suggest that phrasal perturbations can alter the model's interpretation of the problem while maintaining logical equivalence, thereby generating complementary reasoning trajectories.

Language-level Perturbation. Table 4 presents the result of language-level perturbation on machine-translated MATH 500.[6] Across models, reasoning performance is generally highest when problems are presented in English. Despite these disparities, aggregating answers across languages (cross-lingual consistency) improves overall accu-

[5] Complete results are in Appendix A.6.

[6] Complete results on MGSM are provided in Appendix A.7. Appendix A.9 presents example questions and model outputs.

	GSM8K						MATH500					
Models	**CoT**	**RwS**	**RwS+**	**RtS-**	**RtS**	**AGG**	**CoT**	**RwS**	**RwS+**	**RtS-**	**RtS**	**AGG**
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 3: Reasoning accuracy on the benchmarks GSM8K and MATH 500, where AGG is obtained by aggregating vanilla CoT prompting and 4 different question rewriting settings. The best result for each model across settings is bolded.

Models	**BN**	**DE**	**EN**	**ES**	**FR**	**JA**	**RU**	**SW**	**TE**	**TH**	**ZH**	**AGG**
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 4: Reasoning accuracy on machine-translated MATH 500, where AGG is obtained by aggregating answers from different languages. The best result for each model across languages and AGG is bolded.

racy relative to a single language, e.g, a 12.0% percentage point relative improvement over that of English on Qwen-2.5-72B), highlighting its effectiveness in handling more complex tasks and scaling with model size. [7]

Interaction Between Perturbation Types. We analyse how different perturbation types interact when applied jointly. Table 5 reports results for pairwise and full combinations of perturbation dimensions. Across models and datasets, phrasal perturbations tend to produce the strongest individual improvements, while the results from order and language levels are mixed. However, language perturbations exhibit strong synergy when combined with others. In particular, combining phrasal and language perturbations frequently produces the best performance among pairwise combinations. This suggests that language variation can amplify the diversity introduced by rewriting, leading to complementary reasoning signals. Overall, these results highlight that perturbation levels interact in non-additive ways. The effectiveness of perturbation-based aggregation depends not only on individual types but also on how their induced reasoning variations complement each other.

[7] See Appendix A.7 for the zero-shot setting, where the model is prompted to generate responses in different target languages, and Appendix A.8 for a comparison between self-consistency and cross-lingual consistency, as well as their accuracy–efficiency trade-off.

Datasets	MGSM							MATH500						
Models	O	P	L	O&P	O&L	P&L	AGG	O	P	L	O&P	O&L	P&L	AGG
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 5: Accuracy on interaction between perturbation types. Notes: (i) O: order; P: phrasal; L: Language; (ii) best result for each model is bolded.

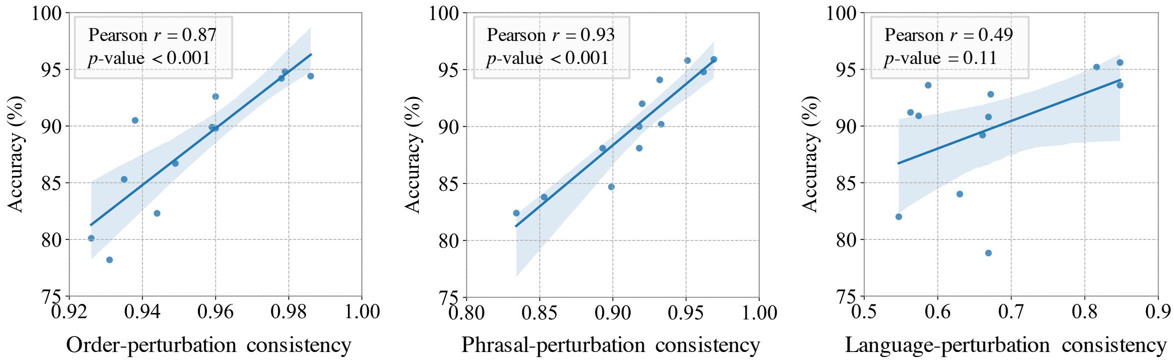


Figure 4: Pearson correlation between models' accuracy and consistency scores under different perturbation depths.

5.4 Correlation Between Consistency and Accuracy

To examine the relationship between reasoning stability and performance, we analyse the correlation between accuracy and consistency under perturbations (Figure 4). We observe that accuracy exhibits a strong positive correlation with consistency under order and phrasal perturbations, with Pearson correlation coefficients of 0.87 and 0.93, respectively. By contrast, the correlation under language perturbation is markedly weaker (Pearson $r = 0.49$) and not statistically significant (p-value $= 0.11$). This weaker relationship can likely be attributed to large performance disparities between high-resource and underrepresented languages, which introduce variability not directly related to reasoning correctness. These findings show that consistency under order and phrasal perturbations appears to be a useful proxy for model reliability without gold answers, whereas this does not hold for language perturbations. Nevertheless, language perturbations may still provide insights into multilingual reasoning and could be leveraged to improve accuracy.

6 Conclusion

We systematically investigated perturbation-based reasoning in LLMs through introducing a hierarchy of perturbations affecting the input with increasing modification depth. We revealed that reasoning consistency decreases as perturbations affect deeper language levels, while aggregating predictions across perturbed inputs consistently improves accuracy. Our findings highlight the value of analysing model behaviour across controlled input variations to better understand and improve the stability and reliability of LLM reasoning.

For future work, it would be interesting to extend this analysis to additional reasoning domains, such as logical reasoning. Another valid extension would be *integrating* the different dimensions (e.g., surface perturbation per language). While this presents an explosion of possibilities, strategic selection based on empirical results, especially cross-all accuracy, might mitigate this challenge.

7 Limitations

While we investigated how different perturbation levels influence reasoning behaviour, several promising directions remain for future exploration. We mainly focus on the perturbation 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.

Giannis Chatziveroglou, Richard Yun, and Maura Kelleher. 2025. Exploring llm reasoning through controlled prompt variations. *arXiv preprint arXiv:2504.02111*.

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.

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*.

Daya Guo, Dejian Yang, Haowei Zhang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, page 633–638.

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.

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.

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.

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*.

Qwen Team. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qwen Team. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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*.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. 2025. Beyond chain-of-thought: A survey of chain-of-X paradigms for LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10795–10809, Abu Dhabi, UAE. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

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 Example Prompts

Figures 5-8 present example prompts used in our experiments, including order-level perturbation, phrasal-level perturbation, and zero-/few-shot language-level perturbation.

A.2 Datasets

We perform experiments and analysis on three math reasoning datasets.

GSM8K (Cobbe et al., 2021), an English dataset of grade school math word problems (about 7,500 for training and 1,319 for testing).

MGSM (Shi et al., 2023), 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.

MATH 500 (Hendrycks et al., 2021) is a benchmark of competition, which contains 500 math problems of varying difficulty. We machine-translate questions from this dataset into 10 languages in MGSM.[8] As shown in Table 8, the machine-translated data achieves cosine similarity scores comparable to those of manually translated data (i.e., MGSM), suggesting that the semantic content is largely preserved across translations.

A.3 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 (Team, 2024a); Qwen2.5-7B-Math (Team, 2024b); Qwen3-8B (Team, 2025); Llama-3.1-8B (Grattafiori et al., 2024); Aya-expanse-8B (Üstün et al., 2024); Ministral-8B[9].

14-35B: Phi-3-14B (Abdin et al., 2024); Mistral-22B[10]; Aya-expanse-32B (Üstün et al., 2024), Qwen3.5-35B-A3B (Team, 2025).

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

A.4 Implementation

We perform inference experiments with greedy decoding on 4 × NVIDIA H100 94GB GPUs using the library vLLM (Kwon et al., 2023), without training or fine-tuning language models. 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 tokenisation issues unfavourable to this language (Ahia et al., 2023). All prompt exemplars we use are released by Shi et al. (2023) and An et al. (2024). We report the final answer accuracy for all experiments except the consistency score.

A.5 Perturbation Depth Scores

For order-level perturbations, embeddings are computed over the full prompted input including few-shot exemplars, since the question itself remains unchanged while the contextual structure varies. For phrasal- and language-level perturbations, embeddings are computed over the perturbed question itself. Tables 6-8 presents the detailed perturbation depth score across different levels.

A.6 Results for Order-level Perturbation

Table 9 and 10 report reasoning results of aggregation(AGG) compared to vanilla CoT prompting on GSM8K and MATH 500.

A.7 Language-level Perturbation

Table 11 reports reasoning results of aggregation(AGG) compared to vanilla CoT prompting on the benchmark MGSM.

To further investigate the role of language in solutions, we perform an additional experiment in a zero-shot setting where the model is prompted to answer in different target languages. The results are reported in Figure 9. It is interesting to see that models perform better when answering in a language different from the question language in some cases, suggesting that reasoning may occur in an internal representation that is partially decoupled from the surface language used for generation. These observations highlight that language perturbations affect reasoning at multiple stages: both the encoding of the math problem and the language

[8] We use Google Translate for the data translation.

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

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

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 5: Example prompts for order-level perturbation.

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 6: Example prompts for phrasal-level perturbation.

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 7: Example prompts (few-shot) for language-level perturbation.

EN Prompt

Roger has 5 tennis balls. **[omit 15 words]**. How many tennis balls does he have now?
Please reason step by step in English, and put your final answer within \boxed{}.

...

DE Prompt

Roger has 5 tennis balls. **[omit 15 words]**. How many tennis balls does he have now?
Bitte begründen Sie Ihre Antwort Schritt für Schritt auf Deutsch und setzen Sie Ihre endgültige Antwort in \boxed{}.

Figure 8: Example prompts (zero-shot) for language-level perturbation.

Orders (4-shot)	2	3	4	5	6	7	8	AVG
MGSM	0.003	0.005	0.002	0.005	0.005	0.077	0.079	0.025
MATH500	0.007	0.014	0.016	0.002	0.014	0.066	0.067	0.031

Table 6: Order-level perturbation depth scores computed against the original prompt (order 1).

used for reasoning. This further explains why language perturbations tend to introduce greater variability compared to order or phrasal perturbations.

Models	1	2	3	4	5	6	7	8	9	10	11	12	AVG
MGSM	0.047	0.042	0.054	0.048	0.037	0.032	0.045	0.060	0.038	0.041	0.040	0.035	0.043
MATH500	0.052	0.055	0.061	0.049	0.046	0.045	0.050	0.066	0.056	0.059	0.060	0.060	0.055

Table 7: Phrasal-level perturbation depth scores computed against the original math problems.

Language	BN	DE	ES	FR	JA	RU	SW	TE	TH	ZH	AVG
MGSM	0.193	0.119	0.119	0.149	0.139	0.154	0.232	0.212	0.192	0.186	0.169
MATH500	0.173	0.136	0.125	0.138	0.156	0.140	0.222	0.182	0.170	0.187	0.163

Table 8: Language-level perturbation depth scores computed against the English math problems.

A.8 Accuracy–efficiency Trade-off

In this work, we conceive cross-lingual consistency (CLC) as a multilingual extension of monolingual self-consistency, going beyond the typical sampling-based strategy. Figure 10 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 11). 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 $\times$ 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.

A.9 Examples Questions and Outputs

Table 12 provides examples of questions written in different languages. Table 13 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 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.

Orders (4-shot)	1	2	3	4	5	6	7	8	AGG
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

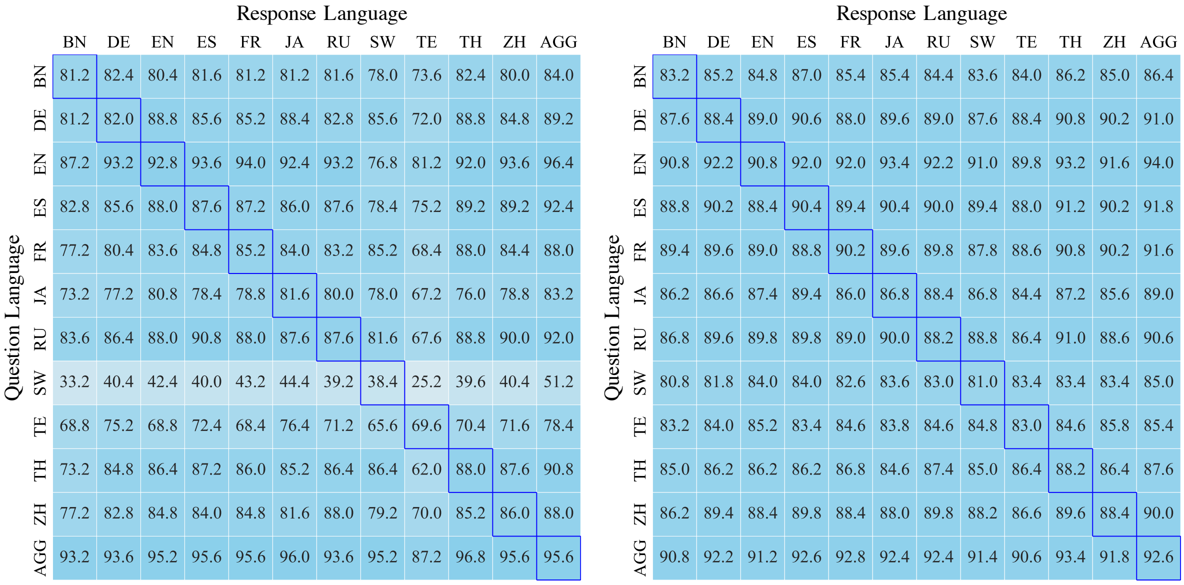
Table 9: Reasoning accuracy of prompts in different orders on GSM8K compared to aggregation (AGG).

Orders (4-shot)	1	2	3	4	5	6	7	8	AGG
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 10: Reasoning accuracy of prompts in different orders on MATH 500 compared to aggregation (AGG).

Models	BN	DE	EN	ES	FR	JA	RU	SW	TE	TH	ZH	AGG
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 11: Reasoning accuracy on the benchmark MGSM. Note that bold numbers indicate the best result for each model among different languages and aggregation (AGG).



(a) Accuracy of Qwen3-8B on MGSM.

(b) Accuracy of Qwen3.5-35B on MATH 500.

Figure 9: Model performance under cross-lingual query-response settings.

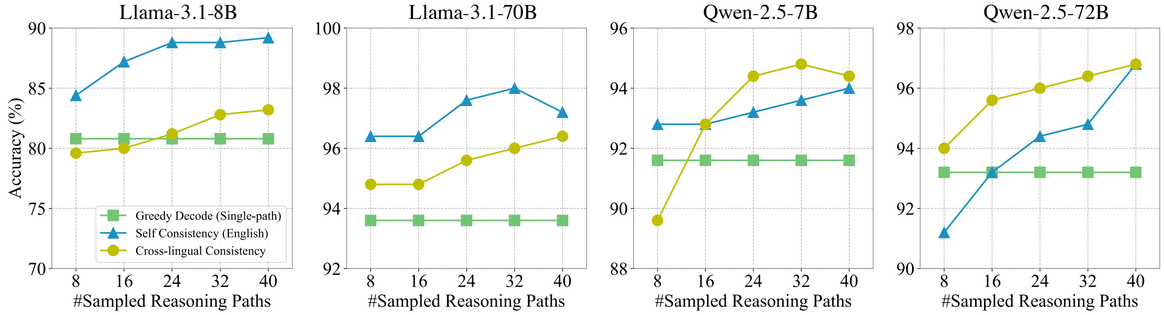


Figure 10: Reasoning accuracy of using varying numbers of reasoning paths on GSM8K.

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？

Table 12: Examples of questions written in different languages.

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 ప్రతిరోజు 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 13: Case study in the test set of MSGM, where the solutions are generated by Qwen2.5-72B for a question written in different languages.