MMATH: A Multilingual Benchmark for Mathematical Reasoning

Wenyang Luo[1], **Wayne Xin Zhao**[1] ✉,
Jing Sha[2], **Shijin Wang**[2] ✉, **Ji-Rong Wen**[1]
[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] iFLYTEK Research (Central China), iFLYTEK Co., Ltd.
wenyang_luo@outlook.com batmanfly@gmail.com {jingsha,sjwang3}@iflytek.com

Abstract

The advent of large reasoning models, such as OpenAI o1 and DeepSeek R1, has significantly advanced complex reasoning tasks. However, their capabilities in multilingual complex reasoning remain underexplored, with existing efforts largely focused on simpler tasks like MGSM. To address this gap, we introduce **MMATH**, a benchmark for multilingual complex reasoning spanning 374 high-quality math problems across 10 typologically diverse languages. Using MMATH, we observe that even advanced models like DeepSeek R1 exhibit substantial performance disparities across languages and suffer from a critical *off-target* issue—generating responses in unintended languages. To address this, we explore strategies including prompting and training, demonstrating that reasoning in English and answering in target languages can simultaneously enhance performance and preserve target-language consistency. Our findings offer new insights and practical strategies for advancing the multilingual reasoning capabilities of large language models. Our code and data could be found at https://github.com/RUCAIBox/MMATH.

1 Introduction

Large language models (LLMs) have shown surprising reasoning ability in many areas, such as mathematical reasoning and logical reasoning, with the advancement of chain-of-thought (CoT). Recent research such as OpenAI o1 (Jaech et al., 2024) and DeepSeek R1 (Guo et al., 2025), has further improved the ability through longer CoT with intermediate plan actions and engaging in trial and error exploration, ultimately improving their performance on complex tasks.

The imbalance in reasoning performance across different languages has drawn increasing attention since the early development (Shi et al., 2022).

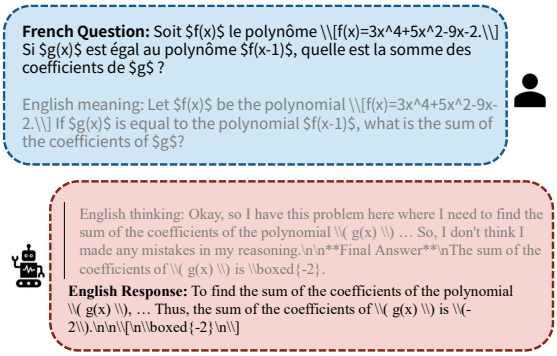


Figure 1: A demonstration of off-target generation. The text with a blue background shows a French question, while the red text represents LLMs' English thinking and response, highlighting a language inconsistency.

To enhance the multilingual reasoning abilities of LLMs, prior research has generally followed two main approaches: prompt-based techniques and training-based interventions. Prompt-based methods typically guide models to leverage their stronger reasoning abilities in English, while training-based methods aim to align multilingual inputs with English-centric reasoning capabilities through supervised learning (Huang et al., 2023; Zhu et al., 2024). However, most of these efforts have focused on relatively *simple benchmarks* such as MGSM (Shi et al., 2022) and MSVAMP (Chen et al., 2023b), while more complex benchmarks, like AIME, remain largely monolingual. This gap has limited progress in understanding and improving *challenging multilingual reasoning tasks*. Moreover, the *off-target issue*—where models respond in unintended languages—fails to meet the needs of monolingual users (see Figure 1), yet remains a significant but overlooked problem in previous research.

To address this gap, we introduce **MMATH**, a new benchmark specifically designed for multilingual complex reasoning. MMATH comprises 374 carefully selected math problems from high-quality

✉ Corresponding authors

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 11187–11202
November 4-9, 2025

sources including AIME, CNMO, and MATH-500, and covers ten typologically and geographically diverse languages. Each problem is translated and validated through a rigorous pipeline that combines frontier LLMs with human verification, ensuring semantic consistency.

Building on the MMATH benchmark, we analyze the behavior of advanced LLMs and identify a prevalent issue: off-target phenomena, where models generate responses in unintended languages. To quantify this, we introduce a metric called language consistency ratio (LCR), which measures the degree of language alignment between input and output. Our investigation centers around two key research questions: **(1) Can LLMs solve non-English questions by reasoning in English? (2) Can LLMs generate answers in the target language?** The first question explores whether reasoning in English—a high-resource language—can enhance performance on non-English tasks, while the second addresses the practical usability of ensuring outputs are in the user's language. For the first question, we find that reasoning in English shows consistently better performance when asked in low-resource languages. For the second question, our prompting skills reveal that large models can be explicitly prompted to generate responses in the desired language, while smaller models frequently fail to retain this control. And moderate thinking intervention can greatly improve language consistency. Finally, we show that after being trained with English reasoning traces and answers in target languages, models can get substantial increases in both answer accuracy and language consistency. Qwen2.5-32B-Instruct with 3K data achieves comparable performance (66.72) with Distill-Qwen-32B (67.01), and answering LCR grows to 97.61, much more higher than reasoning models like QwQ-32B (58.94).

Our contributions are listed as follows:

• We propose **MMATH**, a new benchmark for evaluating multilingual complex reasoning covering 374 high-quality math problems across 10 typologically and geographically diverse languages.

• Our prompting techniques show that moderate thinking intervention could greatly mitigate the off-target problem, and the results may reveal more truthful ability of multilingual models.

• We demonstrate that training on English reasoning traces with multilingual answers could significantly improve answer accuracy and language consistency simultaneously.

2 The MMATH benchmark

In this section, we introduce the construction process of the MMATH benchmark, a new multilingual dataset designed to evaluate complex mathematical reasoning across ten languages. We begin by describing the source data used to build the English portion of the benchmark, followed by the language selection and translation methodology.

Source data. To get high quality English mathematical reasoning benchmark, we choose the following three datasets as the data source.

• *AIME.* American Invitational Mathematics Examination (AIME)[1] is a challenging math contest for top high school students, requiring high logical thinking.

• *CNMO.* China National Mathematical Olympiad (CNMO)[2] is a high-level math competition in China, used to help select students for the national IMO team.

• *MATH-500.* MATH-500 (Lightman et al., 2023) is a benchmark of 500 math problems covering topics such as algebra and calculus.

We collect 30 problems from AIME 2024, 15 from AIME 2025, 18 from CNMO, and 311 filtered problems from MATH-500, resulting in a total of 374 English examples. Most answers in MMATH are written as single LaTeX formulas or plain Arabic numerals. Since some MATH-500 problems are purely textual and may introduce bias when translated (*e.g.,* name results may have different translations), we filter them out and retain only those with numerical answers.

Language Selection. We select a total of 10 languages, including different language families. In addition to English (en), the selected languages are Chinese (zh), Arabic (ar), Spanish (es), French (fr), Japanese (ja), Korean (ko), Portuguese (pt), Thai (th), and Vietnamese (vi), resulting in a total of 3,740 examples in our MMATH benchmark.

Construction Process. To build high-quality multilingual translations of mathematical problems, we develop a three-stage pipeline that combines the strengths of large language models (LLMs) and human expertise. Our process begins with initial LLM-based translation, followed by iterative refinement through cross-model verification, and

[1] https://maa.org/maa-invitational-competitions/
[2] https://www.cms.org.cn/

concludes with manual revision by certified human annotators. The pipeline is illustrated in Figure 4.

• *Stage I: Initial LLM Translation.* We begin by translating the mathematical problems into the target languages using a powerful large language model, such as GPT-4 (Achiam et al., 2023). As for prompt design, we explicitly instruct the model to preserve all mathematical formulas unchanged and to avoid generating unnecessary text such as "The translation is: xxx.". Additionally, we include a one-shot example to better elicit the model's translation capabilities. The full prompt used in this stage is provided in Table 9.

• *Stage II: Iterative LLM Revision.* After that, we use GPT-4 (Achiam et al., 2023), Claude-3.5 sonnet[3] and Grok-3[4] to analyze the translation results and iteratively improve them. In one iteration, if a model approves the translation results, it will be marked as "correct" by that model. If not, it will replace the original translation with an improved version, and all marks already given will be removed. We repeat this for several iterations until all models agree with the translation results. The prompt is shown in Table 10. After 3 iterations, only 15% of the translation results are changed until all models agree.

• *Stage III: Final Human Revision.* Finally each translation undergoes manual revisions, and the evaluation details are shown in Table 11. In this stage, only 3% of the translation results are modified, mainly to correct subtle linguistic inaccuracies or ensure cultural appropriateness.

After all stages, we get the final results and use them as our MMATH benchmark. The difference between previous work and our benchmark is shown in Table 1, highlighting its linguistic diversity and mixed difficulty.

Benchmark	#Languages	#Problems	Difficulty
AIME 24	1	30	Competition level
AIME 25	1	15	Competition level
CNMO 24	1	18	Competition level
MATH-500	1	500	Undergraduate level
MGSM	10	250 × 10	Grade school
MSVAMP	10	500 × 10	Grade school
MMATH (Ours)	10	374 × 10	Mixed

Table 1: Comparison between our MMATH and other mathematical reasoning benchmarks. Our MMATH provides a balanced set with mixed difficulty across multiple languages.

[3] https://www.anthropic.com/news/claude-3-5-sonnet

[4] https://x.ai/news/grok-3

3 Experiments

In this section, we evaluate the multilingual reasoning abilities of popular LLMs on our MMATH benchmark.

3.1 Experimental Setups

Models. We conduct comprehensive evaluations on several popular models. For open-source complex reasoning models, we include QwQ-32B (Qwen-Team, 2025), DeepSeek-R1 (Guo et al., 2025), and various sizes of its distilled versions. For closed-source models, we consider OpenAI's o3-mini. And for comparison, we also include chat models not specifically designed for complex reasoning, such as Gemma3-27B-IT (Gemma-Team, 2025) and Qwen2.5-32B-Instruct (Yang et al., 2024).

Evaluation Prompts. To elicit the potential of models' reasoning ability, we prompt models with native languages (Shi et al., 2022) and ask models to provide the final answer in a specified format (*e.g.,* within a box), enabling reliable rule-based verification of correctness. Figure 5 shows the native prompt of different languages.

Evaluation Setups. By default, we generate outputs using a temperature of t = 0.6, a top-p value of 0.95, and a maximum output length of 32,768 tokens. To obtain a more reliable estimate of reasoning accuracy, each evaluation is repeated 4 times, and the average result is recorded. Given the varying complexity of each benchmark subset, we report the final score using macro-average metric instead of micro-average. We adopt two metrics answer accuracy and language consistency ratio to assess the multilingual reasoning ability of models, as defined below:

Answer Accuracy. Answer accuracy measures the proportion of instances in which the model produces the correct final answer. To extract this answer, we employ the math extraction tool from OpenCompass (2023), which identifies boxed answers. If no boxed output is found, the final numerical value is extracted as a fallback. The extracted answers are then verified against the ground truth using `math_verify` [5].

Language Consistency Ratio. The language consistency ratio (LCR) quantifies how consistently a list of detected languages matches a reference list. In our work, we use fastText (Joulin et al., 2016)

[5] https://github.com/huggingface/Math-Verify

for automatic language identification, and compute LCR to evaluate whether (question, thinking) and (question, answering) are expressed in the same language. We further validate its reliability by manually inspecting 100 randomly selected samples, reaching a 95% correct ratio, which is consistent with existing work (Wyawhare, 2023),

3.2 Main Results

Table 2 presents the overall results, with detailed subset-level outcomes shown in Table 13. We observe a consistent pattern of linguistic inconsistency across all benchmarks and model sizes: models perform significantly better on high-resource languages (e.g., English, Chinese) than on low-resource ones (e.g., Arabic, Thai). This disparity underscores the ongoing difficulty of achieving robust cross-lingual generalization. Interestingly, the performance gap between chat and reasoning models varies by language. In high-resource languages, reasoning models show clear advantages, while in low-resource settings, the gap narrows—suggesting that language modeling ability remains a key bottleneck.

When comparing model types, chat models perform reasonably well on simpler tasks like MATH-500 but struggle on more complex reasoning benchmarks. In contrast, reasoning models consistently outperform them, especially on harder tasks. Notably, smaller reasoning models such as DeepSeek-R1-Distill-Qwen-7B rival or surpass larger chat models, demonstrating the value of targeted reasoning supervision. Among all evaluated models, o3-mini, DeepSeek-R1, and QwQ-32B emerge as the top performers, establishing strong baselines for multilingual mathematical reasoning.

3.3 Further Analysis

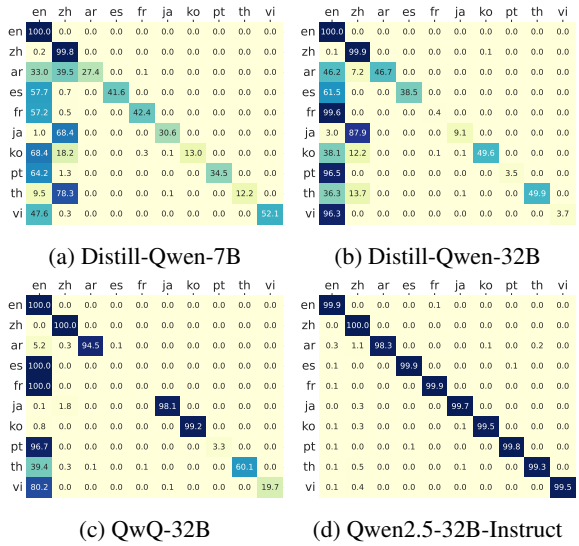
In this section, we mainly focus on two research questions related to language consistency: **(1) Can LLMs solve non-English questions by reasoning in English? (2) Can LLMs generate answers in the target language?** In detail, we start by investigating the normal behavior of LLMs in Section 3.3.1. Then we introduce different methods to explicitly elicit target-language responses in Section 3.3.2. Finally, we train LLMs with English reasoning traces and target language answers, which proves helpful for both answer accuracy and language consistency in Section 3.3.3.

3.3.1 The Off-target Problem of Complex Reasoning Models

Previous study (Chen et al., 2023a) has observed the issue of off-target in some multilingual scenarios, which means the question and response language are mismatched. In this section, we investigate whether this happens in the multilingual complex reasoning area. Given that reasoning models often generate both internal thinking steps and final answers, a fundamental question emerges: What language do models exactly use during thinking and answering?

To examine this, we employ fastText as discussed in Section 3.1 to detect the language and calculate the LCR results at both parts. We analyze three reasoning models DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-32B, and QwQ-32B and compare their behavior with a chat model Qwen2.5-32B-Instruct.

Off-Target Answering Exists. The LCR results are shown in Table 3. As we can observe, though reasoning models have a high performance in accuracy as discussed in 3.2, they have a much lower answering LCR (lower than 60%) compared with chat models like Qwen2.5-32B-Instruct (nearly 100%).



(a) Distill-Qwen-7B (b) Distill-Qwen-32B

(c) QwQ-32B (d) Qwen2.5-32B-Instruct

Figure 2: The percentage to think in each language. The vertical is the source language and the horizontal is the target language. For Qwen2.5-32B-Instruct, as its response doesn't contain <think>, we use the whole response language instead.

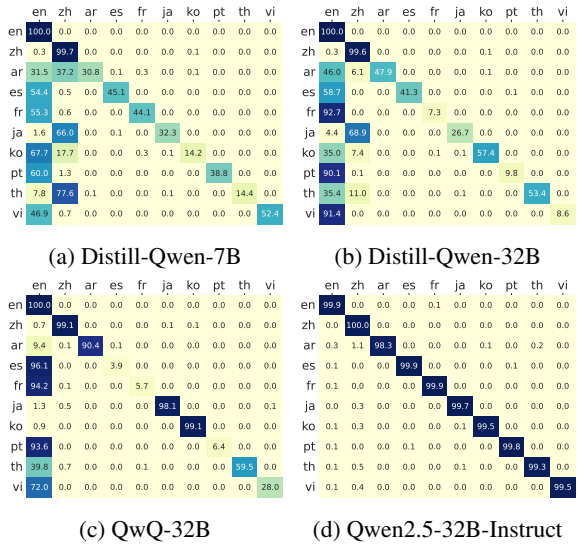
Figure 2 and 3 show some more specific results, the former shows how often a language is used in thinking steps and the latter shows the same statistics for the answer part. As we can see, for

Model	EN	ZH	AR	ES	FR	JA	KO	PT	TH	VI	AVG
Chat LLMs											
Qwen2.5-32B-Instruct	38.43	29.38	27.03	31.13	29.48	25.94	26.44	31.17	27.76	27.37	29.41
Gemma3-27B-IT	50.55	46.39	43.82	46.09	46.95	43.01	43.69	43.36	42.90	42.06	44.88
Reasoning LLMs (distilled)											
DeepSeek-R1-Distill-Qwen-1.5B	45.41	37.59	34.50	40.40	42.08	35.08	34.40	35.49	28.06	36.89	36.99
DeepSeek-R1-Distill-Qwen-7B	63.90	58.53	56.50	62.81	61.58	50.90	59.90	62.72	48.97	58.53	58.44
DeepSeek-R1-Distill-Llama-8B	56.31	45.70	33.68	52.44	54.51	39.21	36.21	55.42	30.19	48.20	45.19
DeepSeek-R1-Distill-Qwen-14B	71.88	55.09	64.71	69.17	65.28	55.65	61.04	66.46	62.36	66.85	63.85
DeepSeek-R1-Distill-Qwen-32B	73.94	61.69	65.02	71.96	70.88	60.29	59.23	72.68	63.31	71.12	67.01
Reasoning LLMs											
QwQ-32B	79.43	74.72	71.10	80.27	79.04	64.38	68.56	78.65	73.43	77.28	74.69
Deepseek-R1	78.81	74.03	72.59	79.54	76.05	72.69	71.38	79.09	75.54	77.43	75.72
o3-mini	82.18	80.95	82.06	79.53	79.52	78.21	73.75	83.74	77.66	81.37	79.90

Table 2: Evaluation results on our MMATH. AVG represents the average answer accuracy score across languages.

Model	Thinking LCR	Answering LCR
Distill-Qwen-7B	45.31	47.11
Distill-Qwen-32B	40.12	45.13
QwQ-32B	57.47	58.94
Qwen2.5-32B-Instruct	99.51	99.51

Table 3: Language consistency ratio for different models. Thinking LCR measures the match ratio between detected thinking language and question language; Answering LCR measures for the answer language.



(a) Distill-Qwen-7B (b) Distill-Qwen-32B

(c) QwQ-32B (d) Qwen2.5-32B-Instruct

Figure 3: The percentage to answer in each language.

high-resource questions like English and Chinese, all models tend to answer in the native language. However, for low-resource languages like Arabic, DeepSeek-R1-Distill-Qwen-7B shows a tendency to answer in either English, Chinese, or Arabic in equal probability. This phenomenon varies between languages, as Thai tends to answer in Chinese while Vietnamese tends to English. QwQ-32B demonstrates relatively better language consistency, except in the cases of Spanish, French, and Portuguese, where it often defaults to English in both thinking and answering steps. We hypothesize that this behavior stems from post-training processes that heavily emphasize high-resource languages.

Off-target Thinking Increases Accuracy. To assess whether this off-target thinking (Figure 2) actually improves the accuracy of mathematical reasoning, we compare the results between thinking in the target language and thinking in off-target languages in Table 4.

From the results, we observe a consistent trend: off-target thinking often yields comparable or even superior accuracy compared to reasoning strictly in the target language. This phenomenon is even more observable in low-resource languages. For instance, in the DeepSeek-R1-Distill-Qwen-7B model, when tackling Arabic, the model completely fails when reasoning in the target language, yet achieves a substantial improvement (0.44) through off-target reasoning. A similar pattern appears in Thai.

In high-resource languages, however, the benefit of off-target thinking is less pronounced. For example, for English, which often serves as the backbone language in pretraining, target-language reasoning yields the highest accuracies across all models (0.57, 0.72, and 0.77, respectively) and no off-target reasoning is recorded.

3.3.2 Explicit Language Elicitation

In this section, we try different methods to elicit models to respond in the target language, aiming at mitigating the off-target problem.

Type	EN	ZH	AR	ES	FR	JA	KO	PT	TH	VI	AVG
DeepSeek-R1-Distill-Qwen-7B											
Target	57	45	0	60	33	0	N/A	50	0	25	30
Off-target	N/A	0	44	53	53	28	47	50	29	51	39
DeepSeek-R1-Distill-Qwen-32B											
Target	72	48	42	48	N/A	0	30	N/A	22	0	33
Off-target	N/A	0	60	69	67	45	67	69	66	69	57
QwQ-32B											
Target	77	65	67	N/A	N/A	56	59	33	65	71	62
Off-target	N/A	N/A	64	81	77	0	50	79	70	76	62

Table 4: The accuracy between thinking in target language and off-target language. N/A means the model has no sample thinking in that language.

Can Models Answer in the Target Language with Explicit Prompts? We begin by examining whether models can be guided to answer in the target language using explicit prompts. To leverage the models' internal English reasoning capabilities, we append a multilingual version of the instruction "please think in English and answer in [target language]" after the native language prompts. We refer to this strategy as the Answer-in-Target Prompt (ATP) as illustrated in Figure 6.

As shown in Table 5 and 6, applying ATP slightly influences the accuracy results and answering LCR. For example, DeepSeek-R1-Distill-Qwen-32B and QwQ-32B have a 1% accuracy increase, while DeepSeek-R1-Distill-Qwen-7B even experiences a performance decline. Considering LCR, we find that DeepSeek-R1-Distill-Qwen-7B has a comparable thinking and answering LCR, which indicates it may has already lost the ability to follow multilingual instructions, explaining why its accuracy decreases. Furthermore, the answering LCR of different models are greatly enhanced, indicating reasoning models are naturally possible to answer in the target language to some extent.

Can Thinking Intervention Mitigate Off-target Problem? Recent research on complex reasoning (Wu et al., 2025; Ma et al., 2025) has proved that thinking intervention could provide a more fine-grained control over models' behavior. In this part, we collect several multilingual thinking patterns we observed in models' original thinking responses and see whether this could mitigate the off-target issue.

• *Discourse-Initiated Thinking (DIT).* When asked in English, the model tends to start their thinking with discourse markers like "Alright" or "Okay", we observe a similar pattern in multilingual scenarios as shown in Figure 7. To leverage this behavior, we extract these markers from native prompt responses. When multiple candidates are available, one is randomly selected and appended after the <think> token. This approach encourages models to initiate their reasoning using discourse cues as entry points into the thinking process.

• *Question-Restatement Thinking (QRT).* Another common pattern observed is that models often restate the question before engaging in actual reasoning. We replicate this behavior by explicitly inserting a restated version of the question at the beginning of the thinking process, as illustrated in Figure 9. This intervention encourages the model to frame the problem before attempting to solve it.

As shown in Table 5, DIT and QRT lead to serious performance drops for the two distilled models, especially on languages except English and Chinese. Compared with them, QwQ shows a relatively better result, which might be attributed to the multilingual CoT training. However, answering LCR results in Table 6 have shown great increases, which means the models' off-target problem could be effectively mitigated with moderate thinking intervention, and these results may show more truthful multilingual ability for multilingual models.

3.3.3 Training with English reasoning traces

In this section, we manually create multilingual training datasets and further prove that training with English thinking could help increase performance while maintaining answering LCR.

Datasets Creation. To construct a moderately sized dataset for complex reasoning, we use the 3K-example subset of Light-R1 (Wen et al., 2025). The questions are sourced from recent benchmarks such as Open-R1 (Face, 2025), LIMO (Ye et al., 2025), and S1 (Muennighoff et al., 2025), while the answers are generated via knowledge distillation from DeepSeek-R1 (Guo et al., 2025). This dataset is filtered to retain 3,000 examples based on reasoning complexity, with all questions, thought processes, and answers presented in English.

Given the lack of multilingual datasets for complex reasoning, we translate this English dataset into 10 languages using *GPT-4o-mini* (OpenAI, 2024). To reduce translation inconsistencies in long-form content, we segment the reasoning process into paragraphs and translate each step-by-step.

Training Setups. Based on the constructed dataset, we propose three supervised fine-tuning strategies designed to enhance multilingual rea-

Model	EN	ZH	AR	ES	FR	JA	KO	PT	TH	VI	AVG
Distill-Qwen-7B	63.90	58.53	56.50	62.81	61.58	50.90	59.90	62.72	48.97	58.53	58.44
Distill-Qwen-7B-ATP	62.64	56.61	51.67	64.97	62.59	40.62	58.66	62.64	49.31	51.61	56.13
Distill-Qwen-7B-DIT	62.63	55.88	56.38	52.61	49.68	24.53	52.58	50.30	40.22	38.53	48.34
Distill-Qwen-7B-QRT	62.48	57.23	56.70	51.81	49.74	29.73	37.97	55.88	33.19	40.48	47.52
Distill-Qwen-32B	73.94	61.69	65.02	71.96	70.88	60.29	59.23	72.68	63.31	71.12	67.01
Distill-Qwen-32B-ATP	73.35	59.06	68.56	69.38	72.51	66.18	69.91	72.77	66.18	70.08	68.80
Distill-Qwen-32B-DIT	72.29	61.99	58.71	65.79	63.48	37.85	49.42	62.72	51.78	47.04	57.11
Distill-Qwen-32B-QRT	71.68	58.15	56.21	61.40	61.22	48.01	48.35	62.67	49.36	58.74	57.58
QwQ-32B	79.43	74.72	71.10	80.27	79.04	64.38	68.56	78.65	73.43	77.28	74.69
QwQ-32B-ATP	78.95	72.85	79.27	78.52	78.80	66.07	68.57	78.31	78.23	73.78	75.34
QwQ-32B-DIT	78.34	75.47	69.68	76.00	74.12	68.60	67.56	75.54	68.39	71.64	72.53
QwQ-32B-QRT	77.86	74.12	71.90	74.73	76.80	66.90	66.18	75.45	67.00	72.50	72.34

Table 5: Evaluation results under different strategies. ATP denotes prompting models to answer in the target language. DIT introduces multilingual discourse markers to guide the models' reasoning language. QRT imitates the models' behavior of repeating questions before reasoning about solutions.

Model	Thinking LCR	Answering LCR
Distill-Qwen-7B	45.31	47.11
Distill-Qwen-7B-ATP	56.54	56.48
Distill-Qwen-7B-DIT	75.61	74.10
Distill-Qwen-7B-QRT	81.41	78.04
Distill-Qwen-32B	40.12	45.13
Distill-Qwen-32B-ATP	29.38	74.55
Distill-Qwen-32B-DIT	97.02	96.57
Distill-Qwen-32B-QRT	97.73	97.62
QwQ-32B	57.47	58.94
QwQ-32B-ATP	35.98	68.27
QwQ-32B-DIT	98.40	96.20
QwQ-32B-QRT	99.88	97.86

Table 6: LCR results for different elicitation strategies.

soning and maintain language consistency. Each method utilizes 3K different examples with the same prompts used in Section 3.1.

• *EN-SFT.* We directly fine-tune on the original English dataset, where the question, thinking, and answering are all in English.

• *Native-Think.* To ensure comparability with the English setup, we randomly divide the dataset indices into 10 parts (corresponding to the target languages), assigning each part to a specific language. This results in 300 examples per language. In this setting, the question, thinking, and answering are all in the respective native language.

• *EN-Think.* As discussed in Section 3.3.1, using English reasoning may enhance multilingual understanding while preserving answering LCR. Based on the Native-Think setup, we substitute the thinking component with its English version while retaining the native language for the question and answer parts.

We fine-tune Qwen2.5-32B-Instruct for 3 epochs on 8 H100 GPUs, employing a cosine learning rate schedule with a peak learning rate of 1×10^{-5}, a warm-up ratio of 0.1, and a total batch size of 8. And we report the results from the final checkpoint.

Results. The performance results are summarized in Table 7, LCR scores are shown in Table 8, and the language usage for reasoning and answering is illustrated in Figure 10. As shown, all three training strategies yield substantial improvements over the base model (Qwen2.5-32B-Instruct). Specifically, the EN-SFT strategy improves performance to 62.38, demonstrating the effectiveness of supervised fine-tuning on English data. The Native-Think variant, which encourages the model to reason in the question's native language, achieves a comparable average score of 61.46. Notably, the EN-Think variant—where the model reasons in English regardless of the input language—achieves the highest average score of 66.72, outperforming all other configurations and approaching the performance of Distill-Qwen-32B (67.01).

In terms of language consistency, EN-SFT yields the lowest answering LCR, suggesting that English-only training contributes to off-target responses. While Native-Think increases answering LCR, it does not lead to notable performance gains. In contrast, EN-Think maintains a high answering LCR (97.61) while significantly boosting performance. These findings provide strong evidence that reasoning in English while answering in the target language is an effective strategy for enhancing multilingual complex reasoning.

Model	EN	ZH	AR	ES	FR	JA	KO	PT	TH	VI	AVG
Base	38.43	29.38	27.03	31.13	29.48	25.94	26.44	31.17	27.76	27.37	29.41
EN-SFT	65.35	65.09	52.20	67.65	66.08	59.54	54.18	64.93	62.93	65.83	62.38
Native-Think	65.38	59.82	61.58	65.18	66.01	56.93	52.86	65.48	58.29	63.04	61.46
EN-Think	66.00	66.31	65.82	68.44	66.37	65.90	67.87	66.11	66.29	68.10	66.72

Table 7: Evaluation results of different training strategies on our benchmark based on Qwen2.5-32B-Instruct. AVG represents the average score across languages.

Model	Thinking LCR	Answering LCR
EN-SFT	57.69	59.20
Native-Think	99.85	99.76
EN-Think	10.04	97.61

Table 8: Language Consistency Ratio (LCR) for different training setting for Qwen2.5-32B-Instruct.

4 Related Work

Multilingual Reasoning. Multilingual reasoning with large language models (LLMs) has received increasing attention, driven by the need for equitable performance across languages. For example, Shi et al. (2022) build the first multilingual mathmatical reasoning benchmark, MGSM, based on GSM8k (Cobbe et al., 2021) and provides several prompting strategies like EN-CoT, which asks the model to predict the chain of thought in English.

Based on the benchmark, more and more techniques have been developed such as prompting and fine-tuning (Chen et al., 2023b; Yoon et al., 2024; She et al., 2024). For example, XLT (Huang et al., 2023) prompts models to translate the question into English and solve it step-by-step, while Liu et al. (2024) leverage multilingual models like NLLB (Costa-Jussà et al., 2022) to improve translation quality. Chen et al. (2023b) further fine-tune models on multilingual data by training them to answer questions either in the same language or across different languages. Also, multilinguality has been proved helpful for empowering other abilities (Mu et al., 2024; Gao et al., 2025).

Despite the advancement of previous research, most of them are based on MGSM, which appears too easy for temporary reasoning models (*e.g.,* Claude 3.5 Sonnet 91.6). To address the limitation, concurrent work like MCLM (Son et al., 2025) and PolyMath (Wang et al., 2025) introduce new reasoning benchmarks across diverse languages. Compared with our work, we not only provide a multilingual complex reasoning benchmark with several subsets and diverse languages, but also focus on the problem of off-target, which is overlooked by previous work. Based on the analysis, we propose several strategies to balance the performance and the off-target phenomenon.

Complex Reasoning. Solving complex reasoning tasks with LLMs is advancing rapidly, driven by methods that enhance test-time computation and learning dynamics. One line of work introduces step-level feedback through process reward models, which score intermediate reasoning steps (Yuan et al., 2024; Snell et al., 2024). Another adopts planning-based techniques such as Monte Carlo tree search to explore and optimize reasoning paths (Feng et al., 2023; Qi et al., 2024; Guan et al., 2025). DeepSeek-R1 (Guo et al., 2025) shows that LLMs can develop strong reasoning skills through reinforcement learning with simple rule-based rewards, without intermediate supervision. Follow-up studies (Hu et al., 2025; Face, 2025) extend this approach to open-source models. Despite these advances, most work remains focused on English benchmarks like AIME and MATH-500 (Lightman et al., 2023), overlooking the multilingual aspect. This work addresses that gap by introducing MMATH, a benchmark for complex reasoning across diverse languages.

5 Conclusion

In this paper, we introduce MMATH, a new multilingual benchmark to evaluate models' complex reasoning ability. MMATH is an extension of the widely used benchmark including AIME, CNMO and MATH-500. It contains 374 examples in 10 typologically diverse languages. Then we present a comprehensive analysis of the multilingual complex reasoning abilities of large reasoning models. We find that temporary reasoning models still show a gap in low-resource language scenarios. Finally, we propose several strategies like prompting, thinking intervention, and training, revealing the possibility to utilize models' English reasoning ability to enhance their multilingual performance while maintaining language consistency.

Limitations

In this paper, we introduce a new multilingual benchmark for complex reasoning and conduct empirical studies on its effectiveness. Nonetheless, several challenges remain as limitations of our work. First, although we explore training-free approaches, balancing accuracy and language consistency remains a significant challenge. Further investigation is needed to develop strategies that optimize both aspects. Secondly, synthesizing multilingual reasoning data remains a challenging problem. Our translation-based approach represents a preliminary attempt in this direction. Future work can explore additional methods for generating native multilingual reasoning data, encompassing both multilingual reasoning processes and corresponding answers. Furthermore, our work focuses on mathematical reasoning. However, it leaves various tasks (*e.g.,* coding and STEM) within the context of multilingualism for large reasoning models unexplored.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China under Grant No. 92470205 and 62222215, Beijing Natural Science Foundation under Grant No. L233008 and Beijing Municipal Science and Technology Project under Grant No. Z231100010323009. Xin Zhao and Shijin Wang are the corresponding authors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023a. On the off-target problem of zero-shot multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9542–9558, Toronto, Canada. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2023b. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. 2023. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*.

Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. 2025. Could thinking multilingually empower llm reasoning? *arXiv preprint arXiv:2504.11833*.

Gemma-Team. 2025. Gemma 3.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. 2025. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning models can be effective without thinking. *arXiv preprint arXiv:2504.09858*.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, and JingBo Zhu. 2024. Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6976–6997, Miami, Florida, USA. Association for Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence.

OpenCompass. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.

Qwen-Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. 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, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. 2025. Linguistic generalizability of test-time scaling in mathematical reasoning. *arXiv preprint arXiv:2502.17407*.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. 2025. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*.

Tong Wu, Chong Xiang, Jiachen T Wang, and Prateek Mittal. 2025. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*.

Arinjay Wyawhare. 2023. Comparative analysis of multilingual text classification & identification through deep learning and embedding visualization. *arXiv preprint arXiv:2312.03789*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

A Prompts in Benchmark Creation

To construct a reliable multilingual benchmark, we designed prompts that guide models to translate mathematical questions accurately and evaluate the quality of these translations. This section presents the prompts used for both translation and translation revision, as shown in Table 9 and Table 10.

I have the following mathematical question written in English, which contains LaTeX formatting. Please translate it into {target_lang} while preserving the LaTeX format and mathematical notation. Ensure that the translation remains accurate and the mathematical expressions do not change. Do not add anything else. ### Example input: How many positive whole-number divisors does 196 have? ### Example output: 196有多少个正整数因子? ### Input: {text} ### Output:

Table 9: The prompt to translate English questions into other languages.

You are given two versions of a mathematical question: one in English and one in {target_lang} language (translated version). Your task is to evaluate whether the translated version is an accurate representation of the original English version. If the translation is correct, confirm that the translation is accurate and return 'Correct'. If the translation is incorrect or the language is wrong or there is unnessary parts, return 'Incorrect' and provide a corrected translation between <trans> and </trans>. Analyse step by step. ### Input: English question: {text_en} Translated {target_lang} question: {text} ### Output:

Table 10: The prompt to judge translation results and give better feedback.

B Human Evaluation Details

To ensure the correctness, fluency, and cultural appropriateness of translations in our multilingual benchmark, we conducted a comprehensive human evaluation. We recruited qualified validators with strong linguistic backgrounds to review the model-generated translations. Specifically, we engaged native speakers for Chinese, and university students for Vietnamese and Portuguese. For other languages, we selected individuals with corresponding language certifications. The validator details are summarized in Table 11.

Validators were instructed to evaluate the translations based on the accuracy of mathematical meaning, correctness of LaTeX formatting, and naturalness of language usage. Each validator was compensated with 1 $ per example, and the whole evaluation process lasted 8 hours. Finally, we manually reviewed and carefully consolidated their assessments to ensure high-quality results.

Language	Language Certification or Identity
Chinese	Native Speaker
Japanese	JLPT N1, TEM8
French	TCF C1, TEM8
Arabic	TEM8
Spanish	MCER B2, TEM8
Korean	TOPIK II
Vietnamese	Vietnam National University Student
Portuguese	Universidade NOVA de Lisboa Student
Thai	CUTFL Chula Superior

Table 11: Languages and corresponding certifications (or identity) of translation validators.

C Benchmark Creation Process

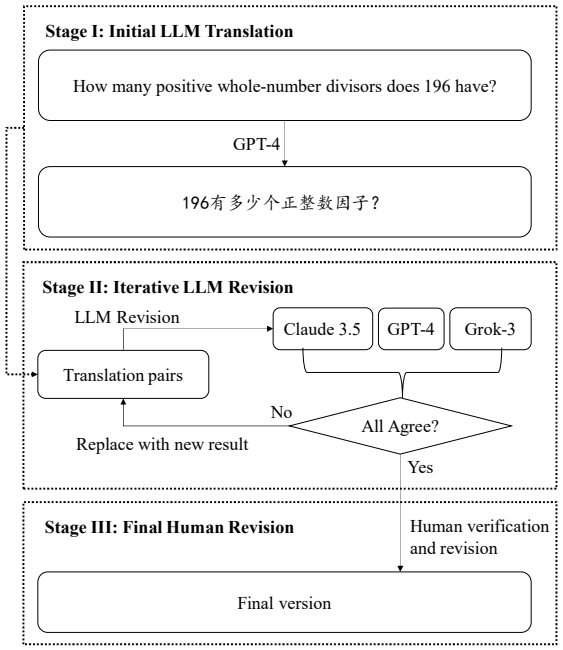


Figure 4: The demonstration of our benchmark construction process.

D Case study

E Other Results

Question

朝、アヤは9キロメトルの散に出かけてからコヒショップに立ち寄ります。彼女がsキロメトルの一定の速度でくと、コヒショップでごすt分を含めて散には4かかります。彼女が$s+2$キロメトルでくと、散には224分かかり、その中にはコヒショップでごすt分が含まれます。アヤが$s+\frac{1}{2}$キロメトルでくと定すると、コヒショップでごすt分を含めた散にかかるを分で求めなさい。

Response

<think>好，我现在来解决这个问题。首先，我要理清楚题目中的各个条件和变量，然后逐步分析，找到正确的解题方法。题目说，阿雅每天早上散步9公里，然后去咖啡店。她的速度是每小时s公里，包括在咖啡店停留的t分钟，整个过程需要4个小时。当她以s+2公里每小时的速度行走时，整个过程用了2小时24分钟，同样包括t分钟在咖啡店。现在，假设她以s+1/2公里每小时的速度行走，求包括t分钟在内的总时间，用分钟表示。...现在，第三种情况：s+0.5=3公里/小时，步行时间=9/3=3小时=180分钟，加上t=24分钟，总时间=204分钟。所以，答案应该是204分钟。 </think>

阿ヤが$s+\frac{1}{2}$キロメトルでく合、コヒショップでの滞在を含めた散にかかるは204分です。

Table 12: An example of cross-lingual thinking and answering. The blue text is Japanese, and the red text is Chinese.

```
'en': "{question}\nPlease reason step by step, and put your final answer within \\boxed{}.",
'es': "{question}\nPor favor, razona paso a paso y pon tu respuesta final dentro de \\boxed{}.",
'fr': "{question}\nVeuillez raisonner étape par étape et mettre votre réponse finale dans \\boxed{}.",
'zh': "{question}\n请逐步推理，并将您的最终答案放在 \\boxed{} 中。",
'ja': "{question}\nステップバイステップで推論し、最終的な答えを \\boxed{} の中に入れてください。",
'th': "{question}\nกรุณาเหตุผลขั้นตอนต่อขั้นตอนและใส่คำตอบสุดท้ายของคุณใน \\boxed{}",
'ko': "{question}\n단계별로 추론하고 최종 답변을 \\boxed{} 안에 넣어주세요.",
'pt': "{question}\nPor favor, raciocine passo a passo e coloque sua resposta final dentro de \\boxed{}.",
'vi': "{question}\nVui lòng lý giải từng bước và đặt câu trả lời cuối cùng của bạn trong \\boxed{}.",
'ar': "{question}\n\\ داخل النهائية إجابتك ووضع بخطوة، خطوة المنطق يرجىboxed{}."
```

Figure 5: Multilingual native language prompts for different languages.

```
'en': "{question}\nPlease reason step by step, and put your final answer within \\boxed{{}}. Please think in English and answer in English.",
'es': "{question}\nPor favor, razona paso a paso y pon tu respuesta final dentro de \\boxed{{}}. Por favor, piensa en inglés y responde en español.",
'fr': "{question}\nVeuillez raisonner étape par étape et mettre votre réponse finale dans \\boxed{{}}. Veuillez réfléchir en anglais et répondre en français.",
'zh': "{question}\n请逐步推理，并将您的最终答案放在 \\boxed{{}} 中。请用英文思考并用中文作答。",
'ja': "{question}\nステップバイステップで推論し、最終的な答えを \\boxed{{}} の中に入れてください。英語で考えて、日本語で答えてください。",
'th': "{question}\nกรุณาเหตุผลเป็นขั้นตอนและใส่คำตอบสุดท้ายของคุณใน \\boxed{{}}. กรุณาคิดเป็นภาษาอังกฤษและตอบเป็นภาษาไทย.",
'ko': "{question}\n단계별로 추론하고 최종 답변을 \\boxed{{}} 안에 넣어주세요. 영어로 사고하고 한국어로 답변해주세요.",
'pt': "{question}\nPor favor, raciocine passo a passo e coloque sua resposta final dentro de \\boxed{{}}. Por favor, pense em inglês e responda em português.",
'vi': "{question}\nVui lòng lý giải từng bước và đặt câu trả lời cuối cùng của bạn trong \\boxed{{}}. Vui lòng suy nghĩ bằng tiếng Anh và trả lời bằng tiếng Việt.",
'ar': "{question}\n\\ داخل النهائية الإجابة ووضع بخطوة خطوة التفكير يرجىboxed{{}}. باللغة التفكير يرجى العربية باللغة والإجابة الإنجليزية".
```

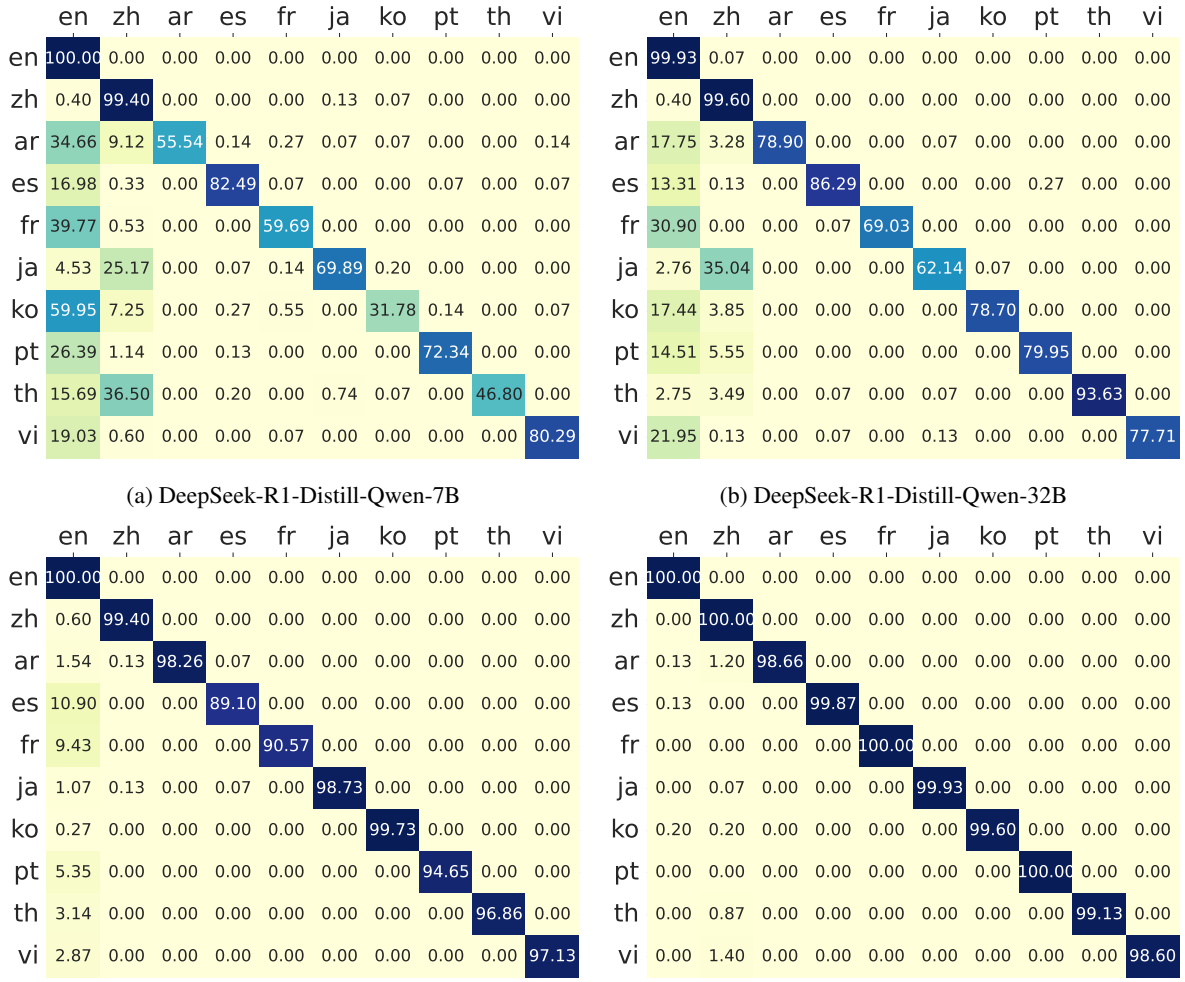
Figure 6: Our ATP prompts, used to ask LLMs to explicitly answer in the target language.

Model	EN	ZH	AR	ES	FR	JA	KO	PT	TH	VI	AVG
AIME2024											
Qwen2.5-32B-Instruct	16.67	15.83	10.00	12.50	9.17	8.33	7.50	12.50	11.67	10.83	11.50
Gemma3-27B-IT	32.50	24.17	21.67	27.50	27.50	20.83	23.33	24.17	18.33	15.83	23.58
DeepSeek-R1-Distill-Qwen-1.5B	28.33	20.83	13.33	25.00	25.00	12.50	19.17	24.17	15.00	25.83	20.92
DeepSeek-R1-Distill-Qwen-7B	56.67	44.17	43.33	53.33	50.83	27.50	46.67	50.00	28.33	49.17	45.00
DeepSeek-R1-Distill-Llama-8B	43.33	25.83	23.33	45.00	39.17	18.33	23.33	40.83	16.67	40.00	31.58
DeepSeek-R1-Distill-Qwen-14B	69.17	45.83	64.17	62.50	64.17	44.17	54.17	60.00	59.17	61.67	58.50
DeepSeek-R1-Distill-Qwen-32B	71.67	47.50	55.00	65.00	66.67	44.17	48.33	69.17	50.83	68.33	58.67
QwQ-32B	76.67	65.00	66.67	80.83	76.67	55.83	59.17	77.50	68.33	75.00	70.17
Deepseek-R1	76.67	70.00	73.33	79.17	76.67	67.50	65.00	78.33	70.00	78.33	73.50
o3-mini	80.83	75.83	75.83	76.67	77.50	76.67	70.83	80.00	71.67	78.33	76.42
AIME2025											
Qwen2.5-32B-Instruct	15.00	8.33	6.67	11.67	11.67	10.00	8.33	15.00	6.67	13.33	10.67
Gemma3-27B-IT	25.00	31.67	30.00	30.00	26.67	26.67	30.00	21.67	31.67	23.33	27.67
DeepSeek-R1-Distill-Qwen-1.5B	28.33	16.67	18.33	25.00	20.00	18.33	16.67	13.33	8.33	23.33	18.83
DeepSeek-R1-Distill-Qwen-7B	38.33	40.00	36.67	41.67	40.00	38.33	43.33	46.67	25.00	40.00	39.00
DeepSeek-R1-Distill-Llama-8B	28.33	33.33	13.33	23.33	35.00	18.33	21.67	35.00	8.33	26.67	24.33
DeepSeek-R1-Distill-Qwen-14B	50.00	28.33	38.33	51.67	45.00	33.33	35.00	45.00	33.33	46.67	40.67
DeepSeek-R1-Distill-Qwen-32B	55.00	45.00	45.00	56.67	51.67	50.00	40.00	55.00	40.00	53.33	49.17
QwQ-32B	66.67	63.33	50.00	66.67	66.67	40.00	51.67	61.67	56.67	61.67	58.50
Deepseek-R1	65.00	55.00	55.00	65.00	56.67	56.67	55.00	61.67	61.67	60.00	59.17
o3-mini	71.67	71.67	75.00	70.00	63.33	75.00	66.67	73.33	66.67	71.67	70.50
CNMO											
Qwen2.5-32B-Instruct	36.11	12.50	13.89	16.67	13.89	9.72	12.50	13.89	16.67	6.94	15.28
Gemma3-27B-IT	51.39	38.89	34.72	34.72	41.67	36.11	33.33	36.11	33.33	38.89	37.92
DeepSeek-R1-Distill-Qwen-1.5B	37.50	30.56	43.06	37.50	45.83	40.28	34.72	33.33	29.17	29.17	36.11
DeepSeek-R1-Distill-Qwen-7B	65.28	58.33	61.11	63.89	65.28	54.17	65.28	62.50	56.94	61.11	61.39
DeepSeek-R1-Distill-Llama-8B	61.11	38.89	34.72	51.39	54.17	44.44	36.11	56.94	31.94	44.44	45.42
DeepSeek-R1-Distill-Qwen-14B	72.22	52.78	63.89	69.44	61.11	55.56	66.67	68.06	66.67	68.06	64.44
DeepSeek-R1-Distill-Qwen-32B	72.22	61.11	66.67	70.83	69.44	54.17	56.94	70.83	72.22	68.06	66.25
QwQ-32B	76.39	73.61	72.22	76.39	75.00	68.06	69.44	77.78	75.00	76.39	74.03
Deepseek-R1	76.39	73.61	66.67	76.39	73.61	70.83	69.44	79.17	75.00	75.00	73.61
o3-mini	79.17	79.17	80.56	73.61	79.17	65.28	61.11	83.33	76.39	79.17	75.69
MATH500											
Qwen2.5-32B-Instruct	85.93	80.87	77.57	83.68	83.20	75.72	77.41	83.28	76.05	78.38	80.21
Gemma3-27B-IT	93.33	90.84	88.91	92.12	91.96	88.42	88.10	91.48	88.26	90.19	90.36
DeepSeek-R1-Distill-Qwen-1.5B	87.46	82.32	63.26	74.12	77.49	69.21	67.04	71.14	59.73	69.21	72.10
DeepSeek-R1-Distill-Qwen-7B	95.34	91.64	84.89	92.36	90.19	83.60	84.32	91.72	85.61	83.84	88.35
DeepSeek-R1-Distill-Llama-8B	92.44	84.73	63.34	90.03	89.71	75.72	63.75	88.91	63.83	81.67	79.41
DeepSeek-R1-Distill-Qwen-14B	96.14	93.41	92.44	93.09	90.84	89.55	88.34	92.77	90.27	91.00	91.78
DeepSeek-R1-Distill-Qwen-32B	96.86	93.17	93.41	95.34	95.74	92.85	91.64	95.74	90.19	94.77	93.97
QwQ-32B	97.99	96.95	95.50	97.19	97.83	93.65	93.97	97.67	93.73	96.06	96.05
Deepseek-R1	97.19	97.51	95.34	97.59	97.27	95.74	96.06	97.19	95.50	96.38	96.58
o3-mini	97.03	97.11	96.86	97.83	98.07	95.90	96.38	98.31	95.90	96.30	96.97
MMATH											
Qwen2.5-32B-Instruct	38.43	29.38	27.03	31.13	29.48	25.94	26.44	31.17	27.76	27.37	29.41
Gemma3-27B-IT	50.55	46.39	43.82	46.09	46.95	43.01	43.69	43.36	42.90	42.06	44.88
DeepSeek-R1-Distill-Qwen-1.5B	45.41	37.59	34.50	40.40	42.08	35.08	34.40	35.49	28.06	36.89	36.99
DeepSeek-R1-Distill-Qwen-7B	63.90	58.53	56.50	62.81	61.58	50.90	59.90	62.72	48.97	58.53	58.44
DeepSeek-R1-Distill-Llama-8B	56.31	45.70	33.68	52.44	54.51	39.21	36.21	55.42	30.19	48.20	45.19
DeepSeek-R1-Distill-Qwen-14B	71.88	55.09	64.71	69.17	65.28	55.65	61.04	66.46	62.36	66.85	63.85
DeepSeek-R1-Distill-Qwen-32B	73.94	61.69	65.02	71.96	70.88	60.29	59.23	72.68	63.31	71.12	67.01
QwQ-32B	79.43	74.72	71.10	80.27	79.04	64.38	68.56	78.65	73.43	77.28	74.69
Deepseek-R1	78.81	74.03	72.59	79.54	76.05	72.69	71.38	79.09	75.54	77.43	75.72
o3-mini	82.18	80.95	82.06	79.53	79.52	78.21	73.75	83.74	77.66	81.37	79.90

Table 13: Evaluation results of different models on various subsets. Scores of MMATH are calculated with macro-average metric.

en: Alright, Okay
zh: 嗯, 好
ar: حسنًا
es: Buneo
fr: Bon
ja: まず
ko: 좋아
pt: Ok, Bem
th: โอเค
vi: Được rồi, Đầu tiên

Figure 7: Multilingual discourse marks used in our DIT thinking intervention method. These are collected from our observations about LLMs' native responses.



(a) DeepSeek-R1-Distill-Qwen-7B

(b) DeepSeek-R1-Distill-Qwen-32B

(c) QwQ-32B

(d) Qwen2.5-32B-Instruct

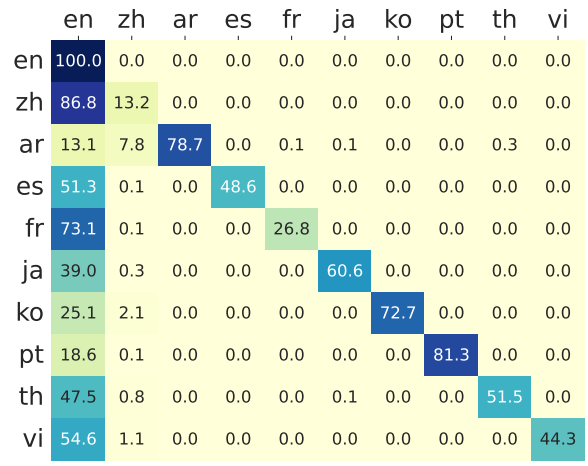
Figure 8: The percentage to answer in each language when prompted with ATP (Figure 6).

'en': 'OK, so the problem is {question}. Let me think in English. First',
'zh': '好的，问题是{question}。让我用中文思考一下。首先',
'ar': 'حسنًا، المشكلة هي{question} . ,’أولا. أفكر باللغة العربية دعني ',
'es': 'Bien, el problema es {question}. Déjame pensar en español. Primero',
'fr': 'D\'accord, donc le problème est {question}. Laissez-moi réfléchir en français. D\'abord',
'ja': 'わかりました。問題は{question}です。日本語で考えさせてください。まず',
'ko': '좋습니다. 문제는 {question}입니다. 한국어로 생각해 보겠습니다. 먼저',
'pt': 'Ok, então o problema é {question}. Deixe-me pensar em português. Primeiro',
'th': 'ตกลง ดังนั้นปัญหาคือ {question} ให้ฉันคิดเป็นภาษาไทย ก่อนอื่น',
'vi': 'Được rồi, vấn đề là {question}. Hãy để tôi nghĩ bằng tiếng Việt. Đầu tiên'

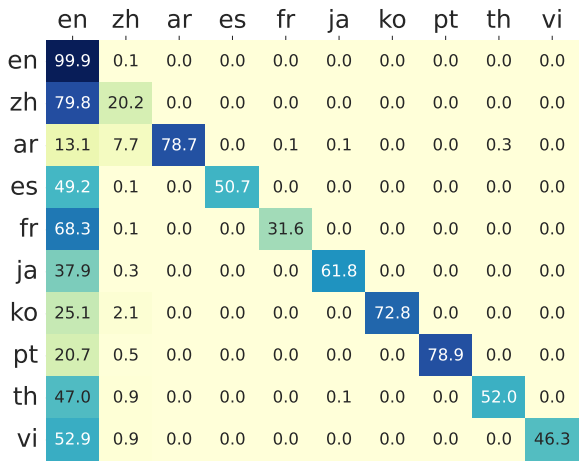
Figure 9: Our QRT thinking intervention, which imitates LLMs' behavior about repeating questions before actually thinking about how to solve it.

Model	EN	ZH	AR	ES	FR	JA	KO	PT	TH	VI	AVG
AIME2024											
EN-SFT	47.50	53.33	43.33	59.17	50.00	49.17	39.17	50.00	48.33	57.50	49.75
Native-Think	50.00	43.33	45.00	55.00	50.83	33.33	32.50	51.67	38.33	43.33	44.33
EN-Think	61.67	59.17	58.33	58.33	58.33	56.67	60.83	56.67	58.33	60.00	58.83
AIME2025											
EN-SFT	46.67	46.67	30.00	51.67	51.67	38.33	41.67	48.33	50.00	43.33	44.83
Native-Think	46.67	43.33	45.00	45.00	45.00	43.33	35.00	50.00	43.33	53.33	45.00
EN-Think	43.33	43.33	50.00	48.33	40.00	45.00	50.00	41.67	45.00	51.67	45.83
CNMO											
EN-SFT	72.22	65.28	50.00	66.67	68.06	61.11	47.22	68.06	65.28	70.83	63.47
Native-Think	69.44	58.33	63.89	66.67	73.61	61.11	52.78	66.67	62.50	62.50	63.75
EN-Think	63.89	68.06	62.50	72.22	72.22	68.06	66.67	70.83	69.44	66.67	68.06
MATH500											
EN-SFT	95.02	95.10	85.45	93.09	94.61	89.55	88.67	93.33	88.10	91.64	91.45
Native-Think	95.42	94.29	92.44	94.05	94.61	89.95	91.16	93.57	88.99	93.01	92.75
EN-Think	95.10	94.69	92.44	94.86	94.94	93.89	93.97	95.26	92.36	94.05	94.16
MMATH											
EN-SFT	65.35	65.09	52.20	67.65	66.08	59.54	54.18	64.93	62.93	65.83	62.38
Native-Think	65.38	59.82	61.58	65.18	66.01	56.93	52.86	65.48	58.29	63.04	61.46
EN-Think	66.00	66.31	65.82	68.44	66.37	65.90	67.87	66.11	66.29	68.10	66.72

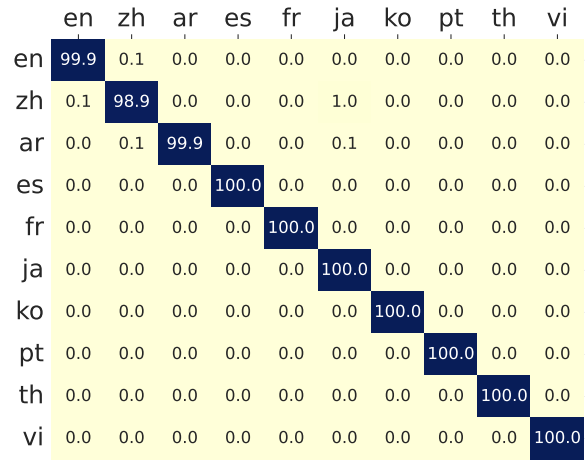
Table 14: Evaluation results of different training methods on Qwen2.5-32B-Instruct: EN-SFT (fully English fine-tuning), Native-Think (full native-language reasoning), and EN-Think (English reasoning with native questions and answers).



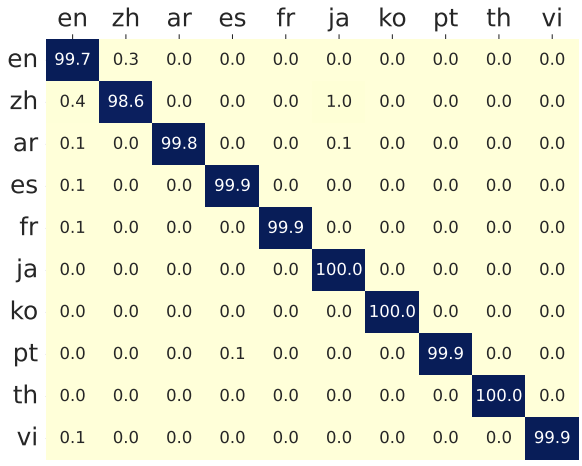
(a) Qwen2.5-32B-Instruct-EN-SFT



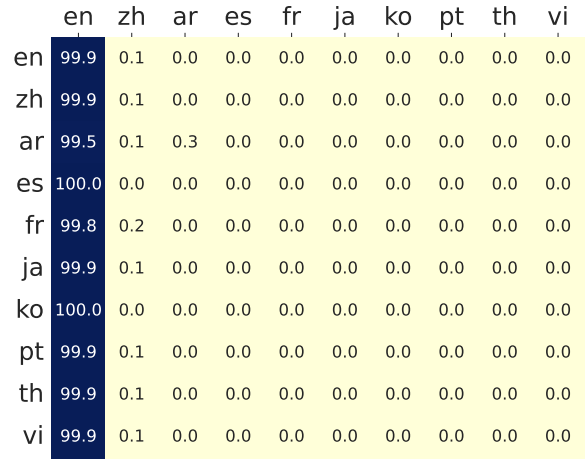
(b) Qwen2.5-32B-Instruct-EN-SFT



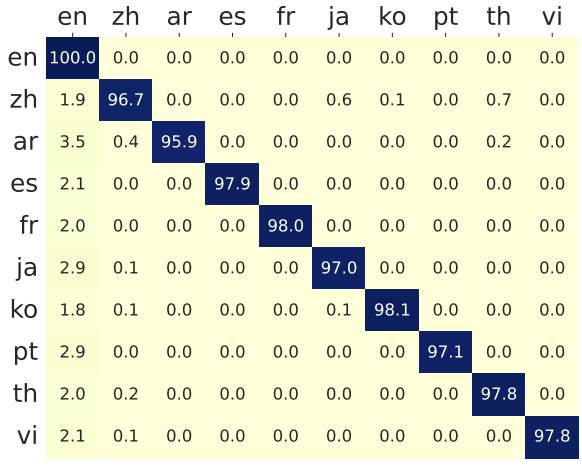
(c) Qwen2.5-32B-Instruct-Native-Think



(d) Qwen2.5-32B-Instruct-Native-Think



(e) Qwen2.5-32B-Instruct-EN-Think



(f) Qwen2.5-32B-Instruct-EN-Think

Figure 10: The percentage to think and answer in each language for our training methods: EN-SFT (fully English fine-tuning), Native-Think (full native-language reasoning), and EN-Think (English reasoning with native questions and answers). The left column is the percentage of thinking, and the right column is answering.