# Correct Answers, Wrong Languages: A Study of Multilingual Reasoning and Retrieval

**Anonymous ACL submission**

## Abstract

This paper comprehensively studies the multilingual performance of today's large language models along two complementary dimensions: multilingual reasoning and multilingual knowledge retrieval. For reasoning, we evaluate on GSM8K and MGSM, finding that models often default to English reasoning traces even when prompted in other languages. For knowledge retrieval, no existing benchmark jointly evaluates cultural grounding and language-consistent retrieval, so we curate GEOFACT-X, a culturally grounded, human-verified test set spanning five languages and five countries. Across both dimensions, we evaluate several post-training approaches: SFT, GRPO, joint SFT+GRPO, and a range of reward functions from local (step-level alignment) to global (context-level alignment and a language-consistency reward). The analysis surfaces an accuracy and language-consistency trade-off under standard recipes—SFT raises GSM8K joint accuracy from 81.2 to 87.2 but collapses MGSM language consistency from 99% to 31%, while training on translated multilingual data reverses the trade-off and degrades math to 45%. Dense cross-lingual alignment paired with a language-consistency reward (M2A) is the only design that breaks the trade-off, matching SFT-level math (87.3% on GSM8K) while preserving 98% language consistency, and improving the joint reasoning score on GEOFACT-X from 30.4 to 48.5.

## 1 Introduction

Large Language Models (LLMs) have made remarkable progress in reasoning tasks, such as mathematics (Liu et al., 2024; Shao et al., 2024), code generation (Jain et al., 2024; Team et al., 2025), and factual QA (Achiam et al., 2023; Guo et al., 2025; Qwen et al., 2025; Wang et al., 2024), primarily in English. Yet, their reasoning capabilities remain underdeveloped in low-resource languages such as Swahili, Marathi, and Thai (Cahyawijaya et al., 2024; Nguyen et al., 2023). This performance disparity undermines the *trustworthiness* of LLMs in these languages: users cannot easily evaluate reasoning traces to verify an answer. While reasoning traces are not guaranteed to faithfully reflect a model's internal computation (Turpin et al., 2023), they remain the primary artifact a user reads to scrutinize an answer, and a trace in a language the user does not read offers no opportunity for scrutiny at all.

In this work, we conduct the first comprehensive study of **multilingual reasoning and knowledge retrieval**: assessing whether LLMs can not only answer questions correctly, but also *reason in the same language as the question*. Prior multilingual benchmarks primarily assess the final accuracy (Ponti et al., 2020; Shi et al., 2023; Xuan et al., 2025), overlooking the language of the reasoning traces. By evaluating reasoning traces directly on MGSM (Shi et al., 2023), we find that models often revert to English reasoning even under non-English prompts. This gap between the prompt language and the reasoning process underscores a broader problem that LLMs can appear correct while failing to reason in the target language. Ensuring both accuracy and language-consistent reasoning is essential for globally inclusive and interpretable AI. Figure 1 illustrates this problem: given a Telugu question, GRPO reasons in English, SFT answers incorrectly despite using the target language, and only one strategy achieves both.

Beyond reasoning, multilingual performance also depends on culturally grounded factual retrieval. Despite recent progress in multilingual LLMs (Ahuja et al., 2023; Qin et al., 2025), their ability to perform factual retrieval across cultural contexts remains largely unevaluated. We introduce GEOFACT-X, a benchmark of culturally grounded questions localized to five countries (USA, India, Japan, Kenya, Thailand) in their predominant languages (English, Hindi, Japanese,

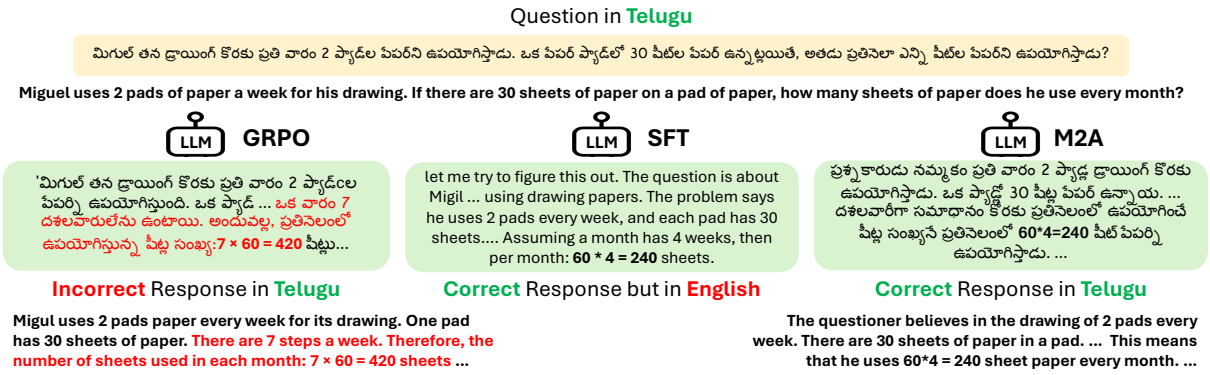


Figure 1: **Example outputs from a Telugu question across three post-training strategies.** The base LLM defaults to reasoning in English. SFT produces a response in the target language but arrives at an incorrect answer. The third strategy achieves both language-consistent reasoning and correctness.

Swahili, Thai). By grounding evaluation in country-specific knowledge, GEOFACT-X tests whether LLMs can retrieve and explain culturally grounded facts in the question language.

We compare post-training strategies: SFT, group relative policy optimization (GRPO) (Shao et al., 2024), joint SFT+GRPO, SFT on machine-translated data, and multi-scale multilingual alignment (M2A), which trains SFT and GRPO jointly with additional context- and step-level cross-lingual alignment rewards plus a language-consistency reward.

On mathematical reasoning (s1K-1.1 (Muennighoff et al., 2025), evaluated on MGSM) and GEOFACT-X (factual), we find an accuracy and language-consistency trade-off: SFT raises accuracy but collapses language consistency; translated-data training preserves language but hurts accuracy; GRPO barely moves either metric. Dense cross-lingual alignment paired with a language-consistency reward breaks the trade-off. Together with the release of our code, data, and evaluation protocols, our work provides a foundation for future research on multilingual reasoning. Our contributions are as follows:

- We introduce GEOFACT-X, a new multilingual factual retrieval benchmark which includes training data and step-by-step reasoning traces across five languages.
- We design an automatic evaluation protocol to assess whether a model reasons in the question language and the correctness of reasoning via language identifier or LLM-as-a-judge.
- We systematically analyze post-training strategies and surface key findings for enabling language-consistent multilingual reasoning.

## 2 Related Work

**Multilingualism in large language models.** Enhancing multilingual large language models (LLMs) traditionally relies on scaling pretraining data (Xue et al., 2021; Conneau et al., 2020; Chang et al., 2023; Gurgurov et al., 2024) or optimizing fine-tuning (Devlin et al., 2019; Conneau et al., 2020; Luo et al., 2023; Zhai et al., 2023). However, scaling faces the *curse of multilinguality*, while aggressive fine-tuning risks catastrophic forgetting. Beyond these paradigms, others explore test-time strategies (Yong et al., 2025) and internal representations (Schut et al., 2025; Zhong et al., 2025). Notably, Schut et al. (2025) found multilingual LLMs often reason internally in English, though they did not analyze generated reasoning traces. More recently, post-training alignment like Direct Preference Optimization (DPO) (Rafailov et al., 2023) has been used to embed multilingual reasoning (Dang et al., 2024; Ranaldi and Pucci, 2025). In this work, we systematically study several post-training strategies, including Group Relative Policy Optimization (GRPO) (Guo et al., 2025; Shao et al., 2024) with multi-scale alignment and language consistency rewards, to understand which approaches effectively build cross-lingual reasoning capabilities.

**Evaluation of multilingual capabilities.** Instruction-tuning datasets such as Bactrian (Li et al., 2023), Aya (Singh et al., 2024), Multilingual Alpaca (Chen et al., 2023), and SphinX (Ahuja et al., 2024) have improved performance across

high- and low-resource languages by emphasizing diversity and cultural specificity. Complementary to these efforts, benchmarks like MEGA (Ahuja et al., 2023) provide broad multilingual coverage across 70 languages and 16 NLP tasks, but primarily evaluate task-level accuracy rather than reasoning processes. Other multilingual reasoning benchmarks, including XCOPA (Ponti et al., 2020), XWinograd (Tikhonov and Ryabinin, 2021), and XStoryCloze (Lin et al., 2022), adopt multiple-choice formats that permit shallow guessing and suffer from translation artifacts (Li et al., 2024). In contrast, our benchmark directly targets multilingual knowledge retrieval by evaluating free-form, step-by-step generation with explicit reasoning traces. This design enables more faithful assessment of both reasoning quality and language alignment, providing a sharper diagnostic tool for multilingual LLMs.

## 3 Evaluating Multiligual Performance

### 3.1 Multilingual reasoning with MGSM

We investigate whether strong performance on multilingual reasoning benchmarks reliably reflects reasoning in the question language. As a case study, we use MGSM (Shi et al., 2023), which evaluates multilingual mathematical reasoning in ten diverse languages and provides chain-of-thought prompts (Naive-CoT) in each language to enforce reasoning in the language. MGSM reports only mathematical accuracy, implicitly assuming that high accuracy implies language-consistent reasoning.

To address this, we introduce *language accuracy*, which measures whether the generated reasoning matches the intended question language, and *joint accuracy*, which requires both a correct answer and reasoning in the target language:

$$A_{\text{lang}} = \frac{1}{N}\sum_{n}^{N} \delta[f(o_n) = l_n]\,, \quad (1)$$

$$A_{\text{joint}} = \frac{1}{N}\sum_{n}^{N} \delta[f(o_n) = l_n] \cdot \delta[\hat{a}_n = a_n]\,, \quad (2)$$

where $f$ is a language identifier, $o_n$ is the generated output, $l_n$ is the target language, and $\hat{a}_n$ and $a_n$ are predicted and ground-truth answers, respectively.

Figure 2 illustrates mathematical accuracy against joint accuracy across recent LLMs; the $y = x$ line marks the ideal where a model always reasons in the target language. While some lie near it, several widely-used families—math-specialized

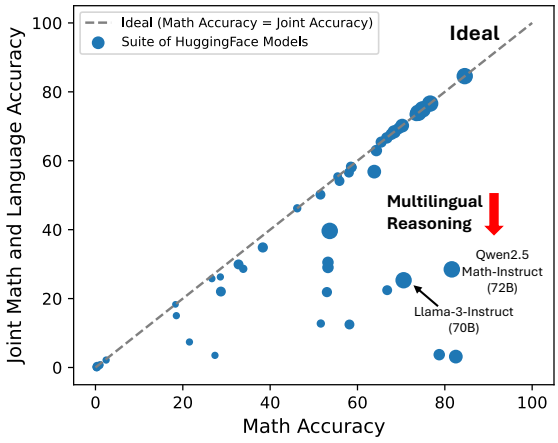


Figure 2: **Mathematical accuracy and the joint accuracy of mathematics and language of various LLMs on MGSM with native Chain-of-Thought.** Circle size is proportional to the number of parameters. The $y = x$ line represents ideal performance, where a model always uses the target question language in reasoning. Many models fall significantly below this line, indicating they solve the problem correctly but fail to adhere to the language prompt.

variants (e.g., Qwen2.5-72B-Math-Instruct), large general-purpose instruct models (e.g., Llama-3-70B-Instruct), and SFT-fine-tuned variants such as s1—fall well below; even consistent base models lose this consistency under standard reasoning SFT (Section 5.1: language accuracy drops from 99% to 31%). A full per-model analysis is provided in Appendix E.

Joint accuracy exposes the language gap on mathematical reasoning, but MGSM is culturally neutral; to test reasoning over cultural and regional knowledge, we introduce GEOFACT-X.

### 3.2 Multilingual Knowledge with GEOFACT-X (Geography-Based Factual Retrieval Benchmark)

Robust evaluation of factual retrieval across cultures remains underexplored (Ahuja et al., 2023; Qin et al., 2025). We introduce GEOFACT-X, a benchmark of 3,000 culturally grounded questions (about 600 per country) spanning history, politics, geography, art, and culture, localized to the USA, India, Japan, Kenya, and Thailand in their predominant languages (English, Hindi, Japanese, Swahili, and Thai). Unlike existing multilingual factual benchmarks, GEOFACT-X provides both a training set and reasoning-trace evaluations (Table 1).

Figure 3 illustrates dataset construction. A two-stage validation pipeline removes incorrect or in-

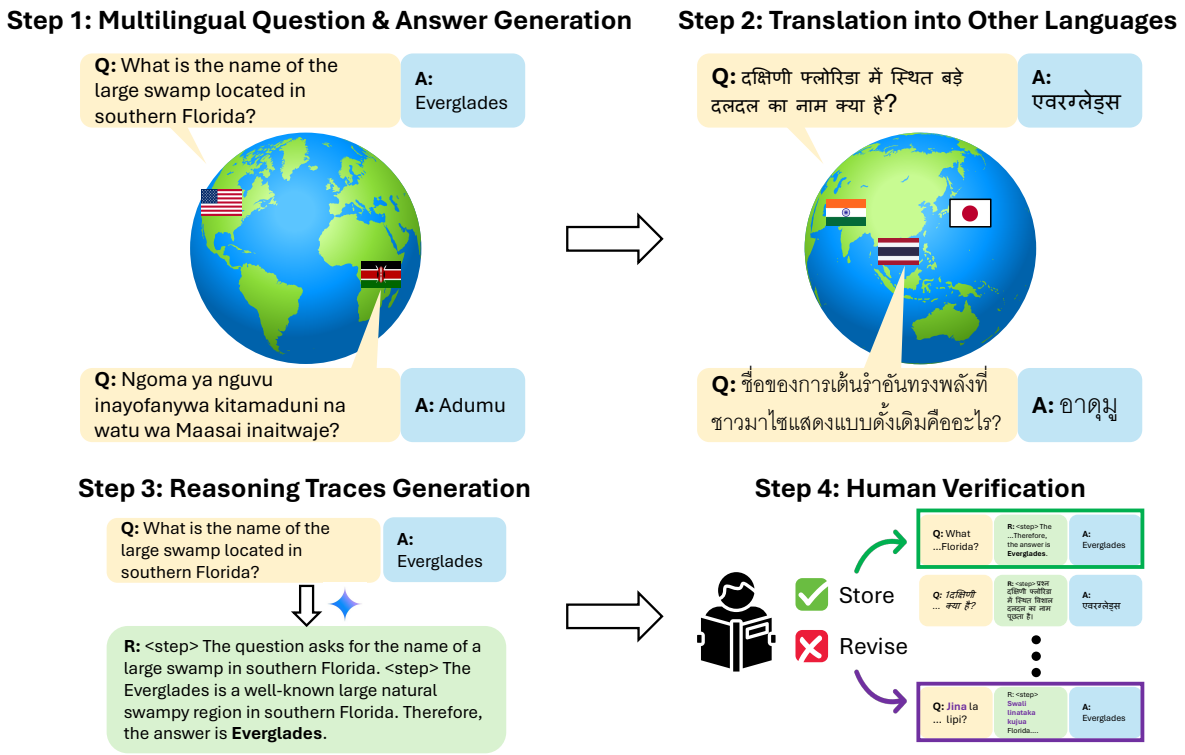


Figure 3: **Illustration of GEOFACT-X benchmark construction.** (1) Geography-aware multilingual questions and answers are generated by Gemini 2.0 Flash. (2) The data is translated into other languages, verifying whether it is back-translatable. (3) The reasoning trace for each question and answer pair is generated. (4) Native or C1-level speakers verify each data and revise it if needed.

Table 1: Comparison between existing multilingual factual or common-sense benchmark and GEOFACT-X.

| Benchmark | Size | #Lang. | Geo-Aware | Train Set | Knowledge Retrieval Eval. |
|---|---|---|---|---|---|
| XStoryCloze (Lin et al., 2022) | 1872 | 11 | | ✓ | |
| XWINO (Tikhonov and Ryabinin, 2021) | 3961 | 6 | | | |
| XCOPA (Ponti et al., 2020) | 6600 | 11 | | ✓(English only) | |
| X-FaKT (Aggarwal et al., 2025) | 2362 | 13 | | | |
| XLQA (Roh et al., 2025) | 3000 | 8 | ✓ | | |
| GEOFACT-X (ours) | 12565 | 5 | ✓ | ✓ | ✓ |

consistent pairs via rule-based filters and cross-language checks. Specifically, we verify cross-language answer consistency by translating each answer into English via the Google Translate API to identify mismatches. Gemini 2.0 Flash (Team et al., 2023) then generates structured chain-of-thought reasoning traces for each item. We split the dataset into a train set (85%) and a test set (15%), ensuring that semantically identical examples result in the same split. All test samples are manually verified by native or C1-level[1] speakers for factual correctness and linguistic clarity. Figure 6 in Appendix illustrates an example multilingual question with its reasoning trace and final answer.

For evaluation, we measure answer accuracy and reasoning score. Answer accuracy is computed by comparing predictions against the reference answers. The reasoning score is assessed by Qwen-2.5-72B-Instruct (Qwen et al., 2025) as an LLM-as-a-judge, comparing model-generated reasoning traces against the human-verified reasoning traces in the test set. If a reasoning trace is produced in a language different from the question, identified by a language detector (Lui and Baldwin, 2012), its score is set to zero. Detailed curation, distribution, and evaluation procedures are provided in Appendix A. We validate the reliability of this metric via a human agreement study in Appendix B.

[1]Proficient users who can use language flexibly and effectively for social, academic and professional purposes (CouncilOfEurope, 2001).

## 4 Post-Training Strategies for Multilingual Reasoning

To comprehensively study post-training for language-consistent multilingual reasoning, we evaluate three families of approaches. SFT is the standard reasoning-distillation recipe but uses English-only traces, providing no signal about the language of the model's output. GRPO adds reinforcement signal but typically relies on a sparse correctness reward together with format and language-consistency rewards. Joint SFT+GRPO combines the two to overcome GRPO's sparse-reward limitation, and provides a natural place to introduce additional reward functions inspired by traditional machine-translation evaluation—from local (step-level alignment) to global (context-level alignment). We refer to our instantiation of this joint approach with multi-scale alignment as M2A.

**Problem Setup.** Given a training question $q$ with ground-truth response $y$, we translate $q$ into target language $l$ via machine translation to obtain $q'$. SFT minimizes the negative log-likelihood of $y$ given $q$; GRPO (Shao et al., 2024) samples a group of $G$ outputs $\{o_1, \ldots, o_G\}$ from the policy $\pi_\theta$ conditioned on $q'$ and updates $\pi_\theta$ using per-output advantages $\hat{A}_i$ computed from the group's reward.

### 4.1 Supervised Fine-Tuning (SFT)

SFT minimizes the negative log-likelihood of a ground-truth reasoning trace $y$ given a question $q$. For mathematical reasoning we follow the s1K-1.1 recipe (Muennighoff et al., 2025), whose traces are predominantly English—providing no signal about the language in which the model should reason on non-English prompts. For factual retrieval, we use the (multilingual) GEOFACT-X training split.

### 4.2 GRPO

Following Guo et al. (2025), the baseline GRPO setup uses three rewards: correctness $r_{\text{correct}}$, format $r_{\text{format}}$, and a language-consistency reward defined as

$$r_{\text{lang}} = \delta[f(o) = l], \tag{3}$$

where $\delta[\cdot]$ is the indicator function, $f$ is a language detector (*e.g.*, langid (Lui and Baldwin, 2012)), and $l$ is the target language. These rewards are sparse—$r_{\text{correct}}$ is binary and $r_{\text{lang}}$ saturates once any output is in the target language—leaving little gradient to improve the *quality* of multilingual reasoning.

### 4.3 Joint SFT+GRPO and Multi-Scale Alignment Rewards (M2A)

**M2A** combines SFT supervision with GRPO and adds dense alignment rewards that exploit reference traces—signals the sparse GRPO baseline ignores. Drawing on classical work in machine translation, we introduce two complementary rewards: a **context-level** reward for whole-trace similarity (Zhang et al., 2020), and a **step-level** reward for sentence-by-sentence correspondence (Gale and Church, 1991; Thompson and Koehn, 2019).

**Multilingual context alignment.** The output and ground-truth are encoded with the encoder $\phi$ as $z_o = \phi(o)$ and $z_y = \phi(y)$. A naive alignment reward could be the cosine similarity $\cos(z_o, z_y)$, but this is maximized when $o = y$, ignoring the question language. To address this, we introduce negative samples by shuffling outputs and ground-truth responses with the same permutation, $\tilde{z}_o = \phi(\psi(o))$ and $\tilde{z}_y = \phi(\psi(y))$, where $\psi$ denotes the shuffle function. Inspired by Triplet Loss (Schroff et al., 2015), the context-alignment reward maximizes similarity between positive samples and minimizes similarity between negative samples, enforcing a margin $\alpha$:

$$r_{\text{context-align}} = \max(\cos(z_o, z_y) - \cos(\tilde{z}_o, \tilde{z}_y) + \alpha, 0), \tag{4}$$

where $\alpha = 1$, the maximum possible value of cosine similarity.

**Multilingual reasoning-step alignment.** At a finer granularity, we align individual reasoning steps. Given split output sentences $\mathbf{o} = (o^{(1)}, \ldots, o^{(N)})$ and ground-truth sentences $\mathbf{y} = (y^{(1)}, \ldots, y^{(M)})$, each output sentence $o^{(i)}$ is aligned to a ground-truth sentence $y^{(j_i)}$. Since $N$ and $M$ may differ, we use dynamic programming (DP) to maximize the total similarity while preserving order:

$$\max_{1 \le j_1 \le \ldots \le j_N \le M} \sum_{i=1}^{N} \mathbf{C}_{i,j_i}, \tag{5}$$

where $\mathbf{C} \in \mathbb{R}^{N \times M}$ is the similarity matrix, with each $\mathbf{C}_{i,j}$ defined as in Eq. (4). The step-alignment

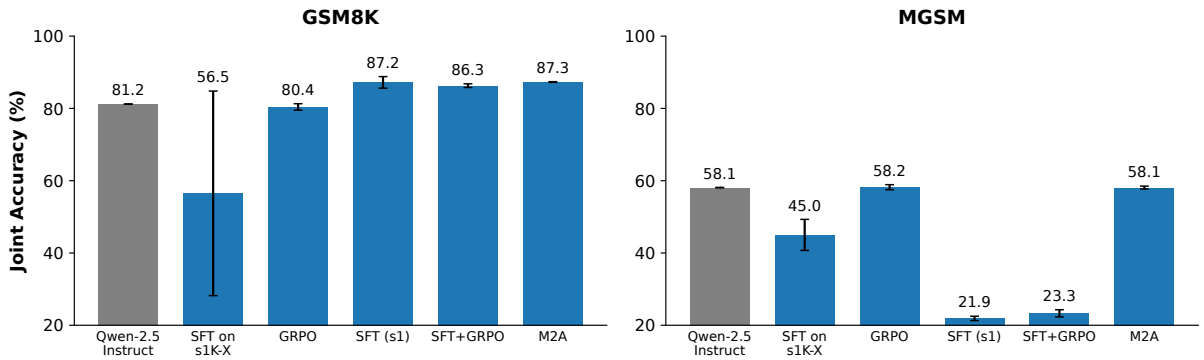


Figure 4: **Joint Accuracy of Qwen2.5-7B-Instruct and post-training methods on GSM8K (English) and MGSM (10 languages).** Results are reported with mean and standard error (see Table 7 for full performance).

reward is the average over aligned pairs:

$$
\begin{aligned}
r_{\text{step-align}} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{C}_{i,j_i} \\
&= \frac{1}{N}\sum_{i=1}^{N} \max\big(\cos(z_o^{(i)}, z_y^{(j_i)}) \\
&\quad - \cos(\tilde{z}_o^{(i)}, \tilde{z}_y^{(j_i)}) + \alpha,\ 0\big). \qquad (6)
\end{aligned}
$$

The DP operates over *sentences* (not tokens) with cost $O(NM)$; for our trace lengths (tens to a few hundred sentences) this completes in milliseconds.

**Final reward.** M2A's final reward sums the GRPO baseline rewards ($r_{\text{correct}}, r_{\text{format}}, r_{\text{lang}}$) and the two alignment rewards ($r_{\text{context-align}}, r_{\text{step-align}}$).

## 5 Experiments

We use Qwen-2.5-7B-Instruct as the backbone for all post-training strategies. mT5 (Xue et al., 2021) is used as the encoder $\phi$ for its cross-lingual embedding properties (Asai et al., 2021; Yoon et al., 2024; Wen-Yi and Mimno, 2023) (see Appendix C). Google Translate is used for machine translation (Agarwal et al., 2023; Aurpa et al., 2024; Li et al., 2021; Subedi and Bal, 2022; Tu et al., 2024).

For mathematical reasoning, we train on s1K-1.1 (Muennighoff et al., 2025) and evaluate on GSM8K (Cobbe et al., 2021) and MGSM (Shi et al., 2023) with Native CoT prompts across ten typologically diverse languages (Bengali, German, Spanish, French, Japanese, Russian, Swahili, Telugu, Thai, Chinese). For factual retrieval, we train and evaluate on GeoFact-X (Section 3.2), which covers five languages chosen for typological/script diversity and native-speaker verification availability. We additionally construct s1K-X, a multilingual version of s1K-1.1 translated into ten languages via Google Translate, to test whether translated training data suffices. All results use three random seeds to report mean and standard error. Training and implementation details are in Appendix D.

### 5.1 Mathematical Reasoning

Figure 4 and Table 7 present each post-training strategy alongside the base model (Qwen2.5-7B-Instruct). SFT on s1K-1.1 improves math but collapses MGSM language consistency (99% to 31%, joint accuracy 58.1 to 21.9; Figure 1 shows the resulting English-trace failure on a Telugu question), and joint SFT+GRPO inherits the same collapse (MGSM joint 23.3%). Training on s1K-X reverses the trade-off—near-perfect MGSM language accuracy (99.7%) at the cost of math (45.2%)—while GRPO barely moves either metric (Figure 13).

M2A is the only configuration on the Pareto frontier of the accuracy and language-consistency trade-off, reaching 87.3% joint accuracy on GSM8K and 58.1% on MGSM at 97.8% language consistency. Appendix G examines a single-translation-language variant; per-language MGSM results are in Appendix H.

### 5.2 Factual Retrieval

The same trade-off pattern appears on GeoFact-X (Figure 5, left): GRPO leaves performance near the base, SFT and joint SFT+GRPO each fail to maintain language-consistent reasoning, and only M2A improves the joint reasoning score (from 30.4 to 48.5) and the answer accuracy (from 26% to 32%). Complete tabular results, including average answer accuracy and the performance of additional pretrained LLMs (Cohere et al., 2025; Guo et al., 2025), are in Tables 10 and 11 in Appendix.

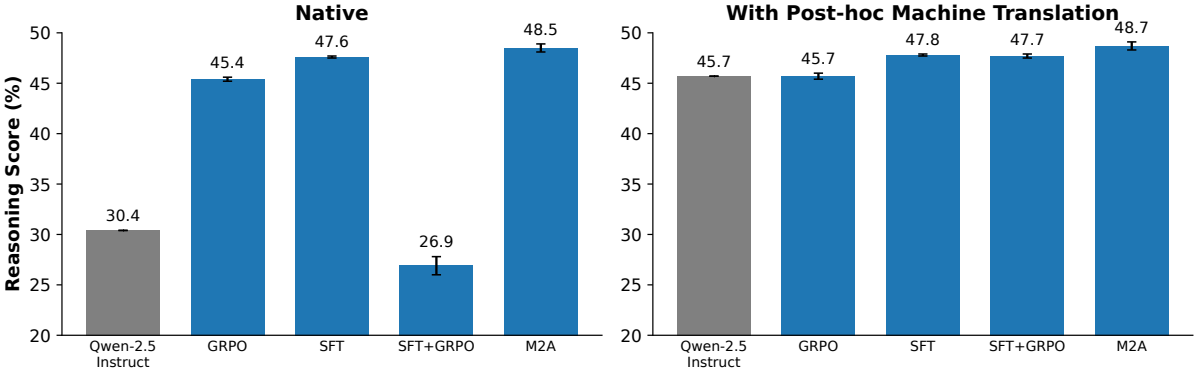


Figure 5: **Average Reasoning Scores evaluated without (left) and with (right) post-hoc machine translation on GEOFACT-X.** Error bars indicate standard error. M2A achieves the highest native performance, effectively bridging the reasoning gap typically observed in models that do not rely on external translation tools.

Table 2: Contribution of individual reward functions to M2A. The evaluation is performed on GSM8K and MGSM. Bold means the best performance. Reward components (under "Reward Variants"): **GRPO**: GRPO baseline rewards (correctness, format, and language-consistency, following Guo et al. (2025)); **CA**: Context Alignment reward; **RA**: Reasoning-Step Alignment reward. Metrics: **Math.** = mathematical answer accuracy; **Lang.** = language accuracy; **Joint** = joint accuracy (Section 3.1).

| Reward Variants | | | GSM8K | | | MGSM | | |
|---|---|---|---|---|---|---|---|---|
| GRPO | CA | RA | Joint | Math. | Lang. | Joint | Math. | Lang. |
| ✓ | | | 86.9 ± 0.0 | 86.9 ± 0.0 | **100.0 ± 0.0** | 53.8 ± 0.1 | 54.2 ± 0.1 | 98.3 ± 0.1 |
| ✓ | ✓ | | 84.7 ± 0.1 | 84.7 ± 0.1 | **100.0 ± 0.0** | 57.5 ± 0.1 | 57.8 ± 0.1 | **99.5 ± 0.1** |
| ✓ | ✓ | ✓ | **87.3 ± 0.1** | **87.3 ± 0.1** | **100.0 ± 0.0** | **58.1 ± 0.4** | **59.0 ± 0.3** | 97.8 ± 0.2 |

Splitting results by whether the question language is associated with the content's country (*e.g.*, Thai with Thailand) reveals a consistent baseline gap—models do better on associated pairs, likely from pretraining co-occurrence (Table 10). M2A improves both settings at similar rates: +4–6% reasoning score, +20–21% answer accuracy.

Post-hoc machine translation (Figure 5, right) inflates the base model but yields no gains for post-trained models, indicating they natively internalize cross-lingual alignment. Qualitatively (Figure 14): on a Swahili question, baselines reason in the target language but reach the wrong answer; only M2A succeeds.

M2A shows consistent improvement on different backbone (Cohere et al., 2025) in Appendix F.

### 5.3 Ablation Study

**Contribution of individual reward functions.** Table 2 analyzes the effectiveness of individual reward components on mathematical reasoning. Context alignment (CA) improves multilingual performance on MGSM but slightly lowers GSM8K accuracy, as enforcing global embedding similarity adds constraints unnecessary for English-only tasks. Reasoning-step alignment (RA) provides finer supervision by aligning individual reasoning steps, which boosts multilingual performance and mitigates the small degradation from CA. The full configuration, combining language consistency, CA, and RA, achieves the best results on both benchmarks, confirming the components are complementary: CA promotes global cross-lingual alignment, while RA enforces stepwise reasoning fidelity.

**Reward formulations.** We compare different formulations of the multilingual alignment reward used in Eq. (4) and Eq. (6). Table 3 reports results on GSM8K and MGSM. Vanilla cosine similarity yields weaker performance, while adding a negative-sample term improves MGSM but slightly reduces GSM8K. The margin-based hinge formulation achieves the best results across all joint and math metrics, demonstrating the benefit of combining negative samples with a margin to stabilize alignment.

**Translation methods.** We utilize Google Translate as a question translator for its computational scalability and broad language coverage (> 100

Table 3: Comparison of reward formulation for multilingual alignment rewards of M2A. The evaluation is performed on GSM8K and MGSM. Bold means the best performance.

| Reward Formulation | GSM8K | | | MGSM | | |
|---|---|---|---|---|---|---|
| | Joint | Math. | Lang. | Joint | Math. | Lang. |
| $\cos(z_o, z_y)$ | $84.1 \pm 0.1$ | $84.1 \pm 0.1$ | $100.0 \pm 0.0$ | $56.0 \pm 0.2$ | $57.4 \pm 0.6$ | $97.4 \pm 0.1$ |
| $\cos(z_o, z_y) - \cos(\tilde{z}_o, \tilde{z}_y)$ | $83.6 \pm 0.1$ | $83.6 \pm 0.1$ | $100.0 \pm 0.0$ | $57.4 \pm 0.1$ | $57.6 \pm 0.1$ | $\mathbf{99.6 \pm 0.1}$ |
| $\max(\cos(z_o, z_y) - \cos(\tilde{z}_o, \tilde{z}_y) + \alpha, 0)$ (ours) | $\mathbf{87.3 \pm 0.1}$ | $\mathbf{87.3 \pm 0.1}$ | $\mathbf{100.0 \pm 0.0}$ | $\mathbf{58.1 \pm 0.4}$ | $\mathbf{59.0 \pm 0.3}$ | $97.8 \pm 0.2$ |

Table 4: Model performance with different translation methods on GEOFACT-X. *Assoc.*: the question language is natively associated with the content's country (e.g., Thai for Thailand); *Non-Assoc.*: no such association. *All* averages across both. Bold values indicate top performance.

| Translation | All | Assoc. | Non-Assoc. |
|---|---|---|---|
| *Reasoning Score (%)* | | | |
| Google Translate | $48.5 \pm 0.4$ | $\mathbf{52.6 \pm 0.6}$ | $47.5 \pm 0.3$ |
| ChatGPT 4.1 | $48.8 \pm 0.1$ | $52.2 \pm 0.3$ | $48.0 \pm 0.2$ |
| ChatGPT 5 | $\mathbf{49.0 \pm 0.1}$ | $\mathbf{52.6 \pm 0.5}$ | $\mathbf{48.1 \pm 0.0}$ |
| *Answer Accuracy (%)* | | | |
| Google Translate | $32.0 \pm 0.6$ | $41.3 \pm 1.0$ | $29.7 \pm 0.5$ |
| ChatGPT 4.1 | $\mathbf{33.1 \pm 0.3}$ | $41.3 \pm 1.3$ | $\mathbf{31.0 \pm 0.4}$ |
| ChatGPT 5 | $32.7 \pm 0.3$ | $\mathbf{41.4 \pm 1.0}$ | $30.5 \pm 0.2$ |

Table 5: Model performance with different multilingual encoders $\phi$ on GEOFACT-X. *Assoc.*: the question language is natively associated with the content's country (e.g., Thai for Thailand); *Non-Assoc.*: no such association. *All* averages across both. Bold values indicate top performance.

| Encoder $\phi$ | All | Assoc. | Non-Assoc. |
|---|---|---|---|
| *Reasoning Score (%)* | | | |
| mT5 (default) | $48.5 \pm 0.4$ | $52.6 \pm 0.5$ | $47.5 \pm 0.3$ |
| BGE-M3 | $48.2 \pm 0.2$ | $52.3 \pm 0.2$ | $47.2 \pm 0.2$ |
| IBM Granite-r2 | $\mathbf{50.7 \pm 0.2}$ | $\mathbf{54.4 \pm 0.4}$ | $\mathbf{49.7 \pm 0.3}$ |
| *Answer Accuracy (%)* | | | |
| mT5 (default) | $32.0 \pm 0.6$ | $41.3 \pm 1.0$ | $29.7 \pm 0.5$ |
| BGE-M3 | $33.2 \pm 1.6$ | $39.9 \pm 2.2$ | $31.6 \pm 1.5$ |
| IBM Granite-r2 | $\mathbf{35.2 \pm 0.4}$ | $\mathbf{42.1 \pm 0.4}$ | $\mathbf{33.4 \pm 0.4}$ |

languages), enabling researchers to reproduce M2A without expensive LLM APIs or high-end GPUs. We compare against LLM-based translators (GPT-4.1 and GPT-5) on GEOFACT-X. Table 4 demonstrates comparable downstream performance across all methods, confirming M2A remains robust regardless of the translator.

**Encoders** We default to mT5 (Xue et al., 2021) for comparability with prior work (Asai et al., 2021; Yoon et al., 2024; Wen-Yi and Mimno, 2023). However, M2A is encoder-agnostic: substituting BGE-M3 (Chen et al., 2024) matches mT5 on GEOFACT-X, while IBM Granite-Embedding-r2 (Awasthy et al., 2026) improves all metrics (Table 5). This highlights significant room for further performance gains simply by leveraging better multilingual embedding functions.

## 6 Discussion

We conducted a comprehensive study of the multilingual performance of large language models (LLMs) along two complementary dimensions: **multilingual reasoning** and **multilingual knowledge retrieval**. Our findings show that many LLMs predominantly reason in English or Chinese even when prompted in other languages, undermining multilingual reasoning quality and limiting applicability in culturally and linguistically diverse settings. Robust evaluation of multilingual reasoning is itself difficult, since most benchmarks focus on final answers rather than reasoning quality or language alignment. We therefore introduce GEOFACT-X, a geography-based factual retrieval benchmark spanning five diverse languages, paired with step-by-step reasoning traces and a reasoning evaluation protocol including logical structure, factual correctness, and language consistency. Our analysis surfaces an accuracy and language-consistency trade-off under standard post-training—SFT and joint SFT+GRPO collapse language consistency, translated-data SFT degrades math, and GRPO barely moves either metric. Only dense cross-lingual alignment paired with a language-consistency reward (M2A) breaks the trade-off, suggesting that improving multilingual reasoning requires training signals that target the language of the reasoning trace, not just answer correctness. Our contributions establish a foundation for training and evaluating LLMs that reason faithfully across languages, advancing the goal of globally inclusive, culturally grounded, and interpretable AI.

## Limitations

We focus our evaluation on 7B-parameter backbones to establish the efficacy of M2A due to resource constraints; validating the scaling laws of this approach on larger models and additional low-resource languages remains an important direction for future work. Furthermore, we used LLM-as-a-judge with Qwen2.5-72B-Instruct. While we demonstrate that this proxy exhibits high reliability with 82.9% human agreement Appendix B, evaluating M2A with stronger judge families could yield further insights.

## Ethics Statement

This work studies multilingual reasoning and knowledge retrieval in large language models (LLMs), with the broader goal of improving equitable performance across languages. GEOFACT-X test items were verified by native or C1-level speakers from each of the five regions to mitigate cultural misrepresentation; the dataset's coverage is nonetheless limited to five languages and five countries, and may inherit biases from the Gemini-2.0 Flash model used in generation. Our work shares standard LLM concerns such as hallucination.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anmol Agarwal, Jigar Gupta, Rahul Goel, Shyam Upadhyay, Pankaj Joshi, and Rengarajan Aravamudhan. 2023. Cst5: Data augmentation for code-switched semantic parsing. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pages 1–10.

Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. 2025. Language models' factuality depends on the language of inquiry. *arXiv preprint arXiv:2502.17955*.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Awadallah, Monojit Choudhary, and 1 others. 2024. sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting. *arXiv preprint arXiv:2407.09879*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. *NeurIPS*.

Tanjim Taharat Aurpa, Samiha Maisha Jeba, Md Shoaib Ahmed, Mohammad Aman Ullah, Maria Mehzabin, and Md Musfique Anwar. 2024. Bangla_mer: a unique dataset for bangla mathematical entity recognition. *Data in Brief*, 54:110407.

Parul Awasthy, Aashka Trivedi, Yushu Yang, Ken Barker, Yulong Li, Bhavani Iyer, Martin Franz, Juergen Bross, Meet Doshi, Vignesh P, Vishwajeet Kumar, Todd Ward, Abraham Daniels, Madison Lee, Luis Lastras, Jaydeep Sen, and Radu Florian. 2026. Granite embedding multilingual r2 models.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2023. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint arXiv:2311.09205*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2023. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *arXiv preprint arXiv:2309.08958*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, and 207 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

CouncilOfEurope. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. Rlhf can speak many languages: Unlocking multilingual preference optimization for llms. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 177–184.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Daniil Gurgurov, Tanja Bäumel, and Tatiana Anikina. 2024. Multilingual large language models and curse of multilinguality. *arXiv preprint arXiv:2406.10602*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Diederik P Kingma. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, and 1 others. 2021. Multi-domain task completion dialog challenge ii at dstc9. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence, Ninth Dialog System Technology Challenge Workshop*.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of llms? *arXiv preprint arXiv:2403.17752*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2022. Few-shot learning with multilingual generative language models. In *EMNLP*.

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. 2024. Are your llms capable of stable reasoning? *arXiv preprint arXiv:2412.13147*.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *ACL*, pages 25–30.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

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. *arXiv preprint arXiv:2501.19393*.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372*.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Qwen, :, 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, and 25 others. 2025. Qwen2.5 technical report.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*.

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. In *NAACL*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over

100 billion parameters. In *SIGKDD*, pages 3505–3506.

Keon-Woo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. Xlqa: A benchmark for locale-aware multilingual open-domain question answering. *arXiv preprint arXiv:2508.16139*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR Workshop on Building Trust in Language Models and Applications*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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 *ICLR*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, and 1 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Bipesh Subedi and Bal Krishna Bal. 2022. Cnn-transformer based encoder-decoder model for nepali image captioning. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 86–91.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. In *ACL Findings*.

Lifu Tu, Jin Qu, Semih Yavuz, Shafiq Joty, Wenhao Liu, Caiming Xiong, and Yingbo Zhou. 2024. Efficiently aligned cross-lingual transfer learning for conversational tasks using prompt-tuning. In *EACL*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *NeurIPS*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *NeurIPS Datasets and Benchmarks Track*.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot llms: Cross-lingual interpretability in token embeddings. In *EMNLP*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision. In *ACL*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *ICLR*.

Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What language do non-english-centric large language models think in? In *ACL Findings*.

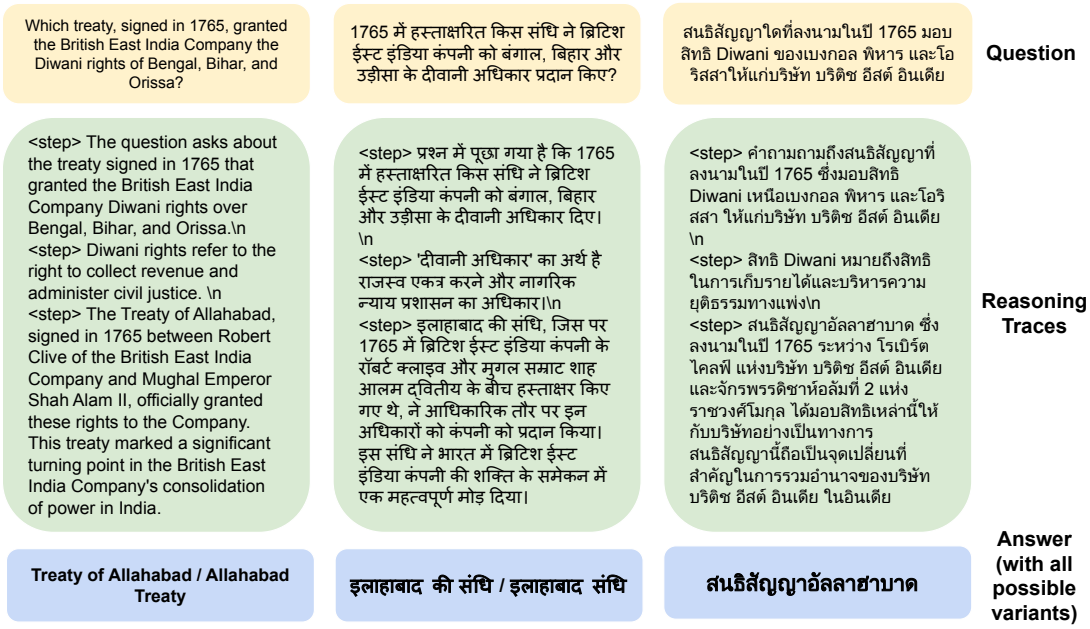


Figure 6: **A sample from GEOFACT-X in English, Hindi, and Thai.** Each presents the same factual question and answer content translated across languages. These multilingual and semantically equivalent traces serve as reference reasoning for benchmarking the reasoning quality of other language models in our evaluation framework.

Generate {num_questions} factual questions about {country} focused on the topic of {topic} where the answer type is {answer_type}.
Requirements:
1. Each question must have a SINGLE, DEFINITE answer (not subjective or opinion-based).
2. Focus on facts that are well-established and locally known in {country}.
3. For each answer, provide ALL possible correct variants (e.g., full names, common abbreviations, alternative names).
4. DO NOT include any ambiguous questions where the answer could be interpreted in multiple ways.
5. Each question should be translated into exactly these languages: English, Hindi, Japanese, Swahili, and Thai.

CRITICAL TRANSLATION REQUIREMENTS:
- Ensure HIGHEST QUALITY translations in all languages. Translations must be accurate and natural-sounding.
- For proper nouns, provide BOTH the transliterated version AND the commonly accepted translation in each language.
- Pay special attention to terms that have specific cultural meaning or context.
- Maintain consistent terminology across all translations of the same question/answer.
- For Hindi translations: Follow modern standard Hindi conventions and proper transliteration standards.
- For Japanese translations: Use appropriate kanji, hiragana, and katakana. Include both kanji and phonetic readings where appropriate.
- For Swahili translations: Use standard Swahili spelling and grammar conventions.
- For Thai translations: Use proper Thai script and formal Thai language.
- When translating names of people, places, or specific terms, include commonly recognized translations in each language.
6. For EACH language version, provide ALL possible correct answer variants in that language.
7. Questions should be DIVERSE within the selected topic - avoid redundant or very similar questions.
8. Ensure the answers are SPECIFIC and PRECISE - avoid phrases or long explanations as answers.

Return the data in the following JSON format:
```
{{
  "question_id": "unique_incrementing_number",
  "languages": [
  {{
  "language_code": "en",
  "language_name": "English",
  "question": "The exact question text in English",
  "answers": ["Primary answer", "Alternative form 1", "Alternative form 2"]
  }},
  {{
  "language_code": "hi",
  "language_name": "Hindi",
  "question": "The exact question text in Hindi",
  "answers": ["Primary answer in Hindi", "Alternative form 1 in Hindi", "Alternative form 2 in Hindi"]
  }},
  // Repeat for Japanese (ja), Swahili (sw), and Thai (th)
],
"topic": "{topic}",
"answer_type": "{answer_type}",
"region": "{country}"
}}
```
IMPORTANT: Return ONLY valid JSON without any explanations, formatting, or additional text outside the JSON structure. Ensure all apostrophes, quotation marks, and special characters are properly escaped in the JSON.

Figure 7: **Prompt for Generating Multilingual Factual Questions and Answers in GeoFact-X.** This prompt instructs the LLM to generate diverse, unambiguous factual questions about a specific country and topic, each with a single, definite answer. The questions and their answers are provided in five languages, English, Hindi, Japanese, Swahili, and Thai, with strict requirements for high-quality translations, consistent terminology, and inclusion of all valid answer variants in each language.

# A Details of GeoFact-X

## A.1 Dataset Collection

We constructed a multilingual factual QA dataset using Gemini 2.0 Flash. For each country–language pair (USA–English, India–Hindi, Japan–Japanese, Kenya–Swahili, Thailand–Thai), we generated 600 unique QA pairs (3,000 examples in total) by using prompt templates shown in Figure 7. The topics spanned ten high-level domains: History, Geography, Politics, Literature, Arts & Culture, Science & Technology, Sports, Food & Cuisine, Language, and Religion with subcategories such as *Person*, *Date*, and *Place* (Figure 8). For each subcategory, 20 questions were generated per country. Translations were produced with Google Translate, and semantic fidelity was checked via back-translation. Reasoning traces were generated by Gemini 2.0 Flash using Chain-of-Thought prompting (Figure 9), with each step explicitly tagged by a '<step>' token.

The dataset is split into training (85%) and test (15%) sets, with no semantic overlap across splits or languages. Ten percent of the training data and all test data were manually verified by the authors through cross-referencing with Wikipedia and Google Search. In addition, all test samples were reviewed by native or C1-level speakers to ensure factual correctness and linguistic clarity and modify the samples if needed.

Our human verification process was comprehensive, covering both linguistic fidelity and the quality of the reasoning traces, rather than being limited to a simple grammar check. The protocol involved a rigorous two-stage process. First, we verified the factual correctness of the original English questions and answers by cross-referencing them with external sources, including online searches and encyclopedias. Second, native speakers evaluated the translated questions, the generated reasoning traces, and the final answers against these verified facts (Figure 10).

## A.2 Evaluation Protocol

The benchmark has three metrics, answer accuracy, and reasoning score. Answer accuracy is computed by checking whether the model prediction appears in the list of reference answers provided for each test instance. Reasoning score is evaluated with Qwen-2.5-72B-Instruct (Qwen et al., 2025) as an LLM-as-a-Judge, which compares model-generated reasoning traces against human-verified references to measure the quality of generated reasoning. If a reasoning trace is written in a different language from the question, detected by a language identifier (Lui and Baldwin, 2012), its score is set to zero. Figure 11 illustrates the prompt structure used for the LLM-as-a-Judge, including the evaluation instructions and rules applied to model outputs.

## A.3 License

We release the dataset under the MIT License, which permits reuse, modification, and distribution, provided that the original license and copyright notice are included.

# B Human Agreement for LLM-as-a-Judge

We validate the alignment between reasoning score and human judgment through a Human-LLM agreement study on the English-USA subset of GeoFact-X. A human evaluator (native English speaker) performs pairwise A/B testing, selecting the superior reasoning trace from two model outputs and assigning a confidence score on a 1–5 scale. The results demonstrate strong alignment between the Qwen-based judge and human preference, achieving an agreement rate of 82.9%.

We further observe a strong positive correlation between human reported confidence and the reasoning score difference assigned by the LLM. Notably, human confidence is significantly higher when agreeing with the LLM (average confidence 3.3/5) compared to cases of disagreement (average confidence 2.1/5). This indicates that divergences primarily occur in ambiguous "close calls," whereas the metric remains highly reliable for clear-cut distinctions. Future work will expand this validation across all five languages with native speakers to confirm these findings multilingually.

# C mT5 Embedding Robustness

We utilize mT5 to extract multilingual features. To validate its robustness on our dataset, we evaluate matching accuracy across 10,000 GeoFact-X training samples (three trials). A match succeeds if the semantic distance between ground-truth translation pairs is strictly less than that of unrelated pairs. mT5 achieves $83.0\% \pm 0.0$ accuracy, demonstrating sufficient cross-lingual semantic equivalence for our pipeline.

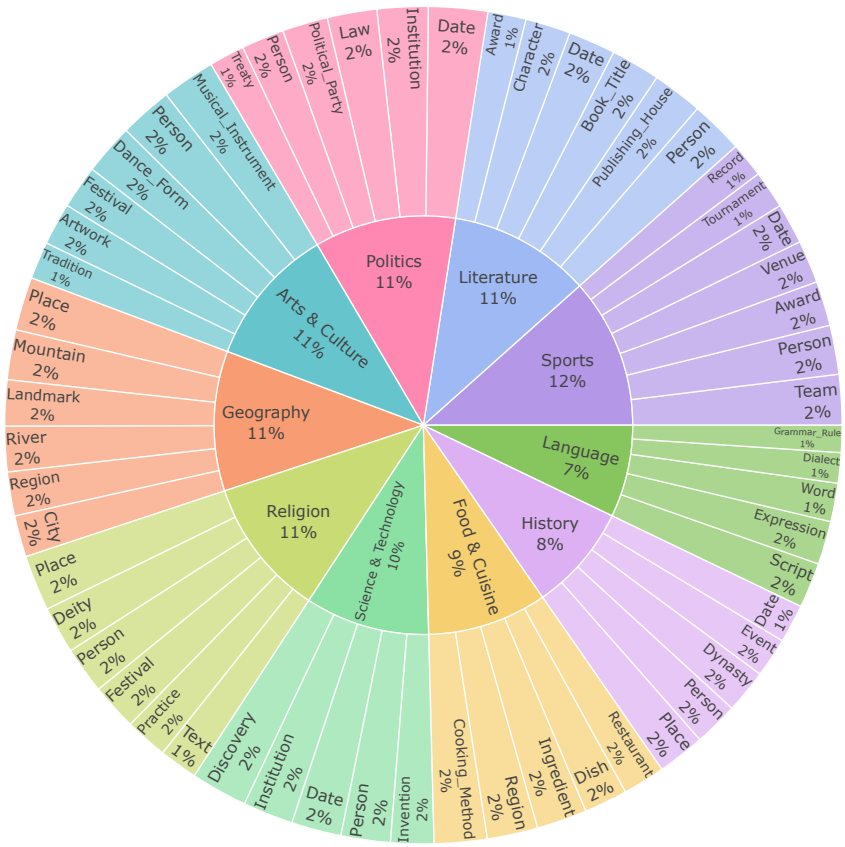


Figure 8: **Illustration of the hierarchical distribution of generated factual question categories by topic and subcategory.** Each colored wedge represents a major topic (e.g., History, Geography), and its outer segments represent specific subcategories (*e.g.*, Person, Place, Treaty). The size of each segment reflects the proportion of questions allocated to that subcategory within its topic. This generation schema was applied uniformly across five countries, and all question sets were translated into five different languages.

# D Experimental Details

## D.1 Training

We adopt the hyperparameter configuration from s1 (Muennighoff et al., 2025), with the exception of batch size, which we reduced from 16 to 8 due to resource constraints. Specifically, the training hyperparameters are as follows: learning rate of $10^{-5}$, minimum learning rate of 0, weight decay of $10^{-4}$, total batch size of 8, training conducted for 5 epochs, and a cosine learning rate scheduler with warmup ratio of 0.05. Adam (Kingma, 2015) is used with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The maximum sequence and token lengths are set to 20,000. GRPO uses accuracy and format rewards following Guo et al. (2025). M2A uses a maximum completion length of 1,024 (256) tokens, generates 2 (8) completions per prompt, and sets the maximum step to 10 due to the resource constraints (parentheses denote factual retrieval parameters in case of difference). We also use a loss coefficient for GRPO as 0.01 for mathematical reasoning and 0.5 for factual retrieval. We train the model with three different random seeds to calculate the standard error. For the s1K-X dataset, we use Google Translate to translate the s1K-1.1 dataset into multiple languages used in the MGSM benchmark: Bengali, German, Spanish, French, Japanese, Russian, Swahili, Telugu, Thai, and Chinese. Both codebase and datasets are attached to the supplementary material for reproducibility. Training and evaluation are conducted on 4 NVIDIA A100 GPUs with DeepSpeed (Rasley et al., 2020).

## D.2 Evaluation

For mathematical reasoning, we employed the lm-evaluation-harness[2] library to evaluate each model. Specifically, we used the MGSM (Shi et al., 2023) Native-CoT setting and the MMLU-ProX (Xuan et al., 2025) Math category with a 5-shot chain-of-thought prompt to ensure the model reasons in its native language. langid (Lui and

[2] https://github.com/EleutherAI/lm-evaluation-harness

You are a multilingual reasoning assistant. For each of the following factual questions about the country, generate structured output in the following format for all the 5 questions:
{
"question": "",
"answer_list": [""],
"reasoning": "<step> Step-by-step reasoning in the same language as the question, each step starting with <step>",
"corrected_answer_list": [""],
"topic": "",
"region": "",
"answer_type": ""
}
Instructions:
1. The reasoning must be written strictly in the same language as the question.
2. Each reasoning step must begin with <step>
(e.g., <step> माउंट केन्या एक ज् वालामुखी पर्वत है ...).
3. Start from relevant background knowledge or interpretation of the question and proceed step-by-step toward the correct answer.
4. For the corrected_answer_list:
- Review the provided answer_list.
- Remove duplicates (case-insensitive, spacing-normalized).
- Add valid alternative phrasings or translations if any are missing (e.g., transliterations, local variants).
5. The reasoning must be comprehensive and detailed, including:
- Relevant background information and definitions of key terms or entities.
- Historical, cultural, geographical, or scientific context if applicable.
- Logical deductions and connections to prior knowledge.
- Contrast with similar or confusing facts (e.g., common misconceptions).
- Justifications for why incorrect options are incorrect (if multiple answers are possible).
- Step-by-step elimination or validation of answer candidates.
6. Each <step> should be at least 1–2 full sentences and contribute meaningfully to building up the answer. Do not skip intermediate steps, even if obvious. Think like a teacher explaining to a curious student

Figure 9: **Structured prompt for multilingual factual retrieval trace generation using Gemini 2.0 Flash on GeoFact-X.** This prompt guides the model to generate step-by-step reasoning and corrected answers for factual questions about a country, using the same language as the input question. The output consists of five JSON object strings for the same factual question, each in a different language.

Baldwin, 2012) is used to evaluate language correctness.

## E Full Analysis of Task 1: Revisiting MGSM

We provide a detailed analysis of the MGSM revisit introduced in Section 3.1. We evaluate a broad set of recent large language models, including Qwen2.5 (Qwen et al., 2025), Llama3 (Grattafiori et al., 2024), Gemma3 (Team et al., 2025), and DeepSeek-R1 (Guo et al., 2025), on MGSM with native Chain-of-Thought prompts (Naive-CoT) provided in each language to enforce reasoning in the target language. All models are sourced from HuggingFace[3], a public repository of large language models. See Table 6 for the complete list of models evaluated.

[3] https://huggingface.co/

Figure 12 plots the average mathematical accuracy against joint accuracy across languages. Ideally, both metrics should be identical (grey dashed line), meaning a model always reasons in the target language when it answers correctly. However, we observe substantial gaps across model families:

- **Math-specialized models** such as Qwen2.5-72B-Math-Instruct and Llama-3-70B-Instruct show among the largest gaps, indicating that math-focused training actively suppresses multilingual reasoning by defaulting to English.
- **SFT degrades language consistency.** The s1 models (orange), fine-tuned from Qwen2.5-Instruct (green), notably degrade language accuracy while improving mathematical performance. This demonstrates that supervised fine-tuning on English reasoning data can

Thank you for accepting our request.

All questions and answers in English are already verified. We then paired these with the Thai questions and answers. However, the [LANG] answers and reasoning traces have not been verified yet.

Please check that the [LANG] questions and answers match the English versions and that the reasoning traces are correct.

Below is an example of pairs:

Q: What is the traditional Thai silk weaving technique known for its intricate patterns?

A: Mudmee Silk

Q: เทคนิคการทอผ้าไหมของไทยแบบใดที่มีชื่อเสียงด้านลวดลายที่ซับซ้อน?"<step> คำถามถามถึงเทคนิคการทอผ้าไหมไทยแบบดั้งเดิมที่มีชื่อเสียงด้านลวดลายที่ซับซ้อน

Reasoning: <step> ""มัดหมี่"" เป็นเทคนิคการย้อมสีแบบผูกที่ใช้ในประเทศไทยเพื่อสร้างลวดลายที่ซับซ้อนบนผ้าไหม ซึ่งสอดคล้องกับคำอธิบายของคำถามอย่างลงตัว เป็นการทอผ้าอิกัตชนิดหนึ่งที่แพร่หลายในประเทศไทย
<step> ""Ikat"" เป็นคำที่กว้างกว่าซึ่งหมายถึงเทคนิคการย้อมสีที่ใช้ในการทำลวดลายสิ่งทอซึ่งใช้กระบวนการย้อมสีแบบต้านทานที่คล้ายกับการย้อมสีแบบผูกบนเส้นใยยืนหรือเส้นใยพุ่งก่อนการทอ แม้ว่ามัดหมี่จะเป็นอิกัตชนิดหนึ่ง แต่คำถามก็ถามถึงเทคนิคของไทย*โดยเฉพาะ*
<step> ดังนั้นคำตอบที่ดีที่สุดคือ ""มัดหมี่""ผ้าไหมมัดหมี่ (Mudmee Silk) โดยเฉพาะจากภาคตะวันออกเฉียงเหนือของไทย

A: ผ้าไหมมัดหมี่ (Mudmee Silk) โดยเฉพาะจากภาคตะวันออกเฉียงเหนือของไทย

Figure 10: **Instruction to Thai native speaker for human verification.** All human evaluators are fluent in both English and their native languages.

Table 6: List of all models and sizes evaluated on MGSM in Figure 2. All models are sourced from HuggingFace.

| Model Name | Model Sizes |
|---|---|
| Qwen2.5 | 1.5B, 3B, 7B, 14B, 32B |
| Qwen2.5-Instruct | 1.5B, 3B, 7B, 14B, 32B |
| Qwen2.5-Instruct-GPTQ-Int4 | 1.5B, 3B, 7B, 14B, 32B |
| Qwen2.5-Instruct-GPTQ-Int8 | 1.5B, 3B, 7B, 14B, 32B |
| Qwen2.5-Instruct-AWQ | 7B, 14B, 32B |
| Qwen2.5-Instruct-1M | 7B, 14B |
| Qwen2.5-Math | 1.5B, 7B, 72B |
| Qwen2.5-Math-Instruct | 1.5B, 7B, 72B |
| s1 | 1.5B, 3B, 7B, 14B, 32B |
| Llama-3-Instruct | 8B, 70B |
| Llama-3.3-Instruct | 70B |
| Gemma-3-PT | 1B, 4B, 12B, 27B |
| Gemma-3-IT | 1B, 4B, 12B, 27B |
| DeepSeek-R1-Distill-Qwen | 1.5B, 3B, 7B, 14B, 32B |
| DeepSeek-R1-Distill-Llama | 8B, 70B |

harm multilingual capabilities even when the base model supports them.

- **Scale does not resolve the gap.** Larger models do not consistently close the gap between mathematical and joint accuracy, suggesting that the English-reasoning bias is a training-objective issue rather than a capacity limitation.

These results demonstrate that mathematical accuracy alone overestimates multilingual reasoning

```
# Reasoning Quality Evaluation
You are an expert reasoning evaluator tasked with comparing an LLM's reasoning trace against a ground truth reasoning trace. Your evaluation must be fair, consistent, and based solely on the quality of reasoning, not on superficial similarities.

## Input:
  - Question: {question}
  - Answer List: {answer_list}
  - Ground Truth Reasoning: {ground_truth_reasoning}
  - LLM Response: {llm_generation}


## Evaluation Criteria:
Assess the quality of the LLM's reasoning compared to the ground truth on a scale from 0-10 based on the following:

  1. Logical Structure (40%):
     - How well does the reasoning follow a clear, step-by-step logical progression?
     - Are the steps in a sensible order that builds toward the answer?

  2. Key Insights (30%):
     - Does the reasoning identify the same critical insights as the ground truth?
     - Are the important clues from the question properly recognized and utilized?

  3. Factual Correctness (20%):
     - Is the reasoning free from factual errors?
     - Does it avoid adding irrelevant information or missing necessary information?

  4. Conclusion Validity (10%):
     - Does the reasoning correctly lead to the answer?
     - Is the link between the reasoning and the conclusion clear?

## Scoring Guide:
  0-1: Completely irrelevant or fundamentally flawed reasoning
  2-3: Major logical errors or missing critical insights
  4-5: Contains some correct elements but misses important aspects
  6-7: Good reasoning with minor gaps or imperfections
  8-9: Very good reasoning, almost matching ground truth quality
  10: Perfect reasoning, capturing all key insights with proper structure

## Your Response (FORMAT STRICTLY REQUIRED):
  REASONING_SCORE: [integer between 0-10]
  JUSTIFICATION: [Brief explanation of your evaluation, highlighting strengths and weaknesses]
```

Figure 11: **Prompt for LLM-as-a-Judge to Evaluate Reasoning Traces Using Gemini 2.0 Flash.** This prompt guides the evaluation of an LLM-generated reasoning trace against a ground truth using specific criteria such as logical structure, key insights, factual correctness, and conclusion validity. The evaluation is performed by Qwen2.5-72B Instruct, acting as the LLM-as-judge, and includes scoring, language mismatch detection, and answer validation.

ability, and joint evaluation is essential for assessing true language-consistent reasoning. This motivates both our benchmark (GEOFACT-X, which evaluates reasoning traces directly) and our analysis of post-training strategies that explicitly target language consistency.

## F M2A with Different Backbone on GEOFACT-X

We use Command R7B (Cohere et al., 2025) as a backbone network for comparing M2A other baselines. As shown in Table 8, the results are consistent with our main Qwen findings. M2A achieves the highest performance in both reasoning score and answer accuracy, outperforming SFT and stan-

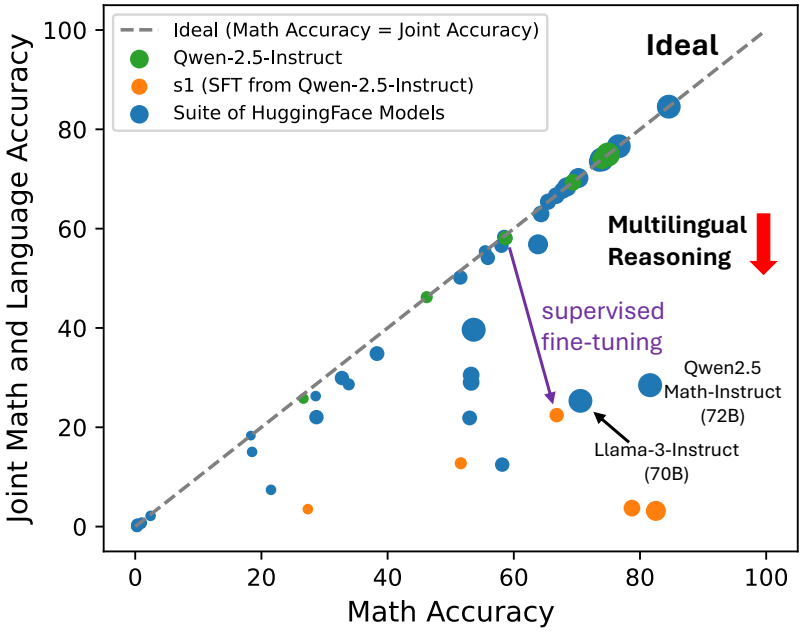


Figure 12: **Mathematical accuracy and the joint accuracy of mathematics and language of various LLMs on MGSM with native Chain-of-Thought.** Circle size is proportional to the number of parameters. The $y = x$ line represents ideal performance, where a model always uses the target question language in reasoning. Many models, especially the supervised fine-tuned model, s1 (Muennighoff et al., 2025), fall significantly below this line, indicating they solve the problem correctly but fail to adhere to the language prompt.

Table 7: Accuracy of Qwen-2.5-7B-Instruct and post-training methods in GSM8K (English) and MGSM (ten languages). Results are reported for mathematical accuracy (Math.), language accuracy (Lang.), and joint accuracy (Joint). Bold indicates the best performance.

| Method | GSM8K | | | MGSM | | |
|---|---|---|---|---|---|---|
| | Joint | Math. | Lang. | Joint | Math. | Lang. |
| Qwen-2.5-Instruct | 81.2 | 81.2 | **100** | 58.1 | 58.7 | 99.0 |
| GRPO | 80.4 ± 0.9 | 80.4 ± 0.9 | **100.0 ± 0.0** | **58.2 ± 0.7** | 58.8 ± 0.4 | 95.9 ± 2.9 |
| SFT (s1) | 87.2 ± 1.6 | 87.2 ± 1.6 | **100.0 ± 0.0** | 21.9 ± 0.6 | **66.7 ± 0.1** | 31.0 ± 0.5 |
| SFT on s1K-X | 56.5 ± 28.3 | 84.3 ± 1.1 | 66.7 ± 33.3 | 45.0 ± 4.3 | 45.2 ± 4.1 | **99.7 ± 0.1** |
| SFT + GRPO | 86.3 ± 0.5 | 86.3 ± 0.5 | 99.9 ± 0.1 | 23.3 ± 1.0 | 62.9 ± 0.4 | 35.1 ± 0.6 |
| M2A | **87.3 ± 0.1** | **87.3 ± 0.1** | **100.0 ± 0.0** | 58.1 ± 0.4 | 59.0 ± 0.3 | 97.8 ± 0.2 |

dard GRPO. This confirms that M2A generalizes across different multilingual architectures.

## G M2A with Different Languages

In the main paper, we present M2A trained with random translations drawn from the ten MGSM languages (Bengali, German, Spanish, French, Japanese, Russian, Swahili, Telugu, Thai, and Chinese). Here, we examine the effect of using a single fixed translation language, as shown in Table 9. We select Japanese and Swahili as representative examples of high- and low-resource languages, respectively, following the categorization of Nicholas and Bhatia (2023).

Across both choices, we observe only minor decreases in GSM8K mathematical accuracy and MGSM language accuracy relative to the multi-language setting. This finding indicates that training with a single language—even a low-resource one—can still induce strong multilingual reasoning ability. Nevertheless, randomizing translations across multiple languages yields the strongest overall results, suggesting that language diversity provides additional regularization benefits for cross-lingual alignment.

Table 8: Comparison of model performance on average reasoning score (%), language accuracy (%), and answer accuracy (%) on GeoFact-X test set. Results are partitioned into Assoc. (the language is natively associated with the content's country) and Non-Assoc. (no native association). Bold values indicate top performance.

| Model | Average Reasoning Score (%) | | | Average Answer Accuracy (%) | | |
|---|---|---|---|---|---|---|
| | All | Assoc. | Non-Assoc. | All | Assoc. | Non-Assoc. |
| Command R7B | 33.1 | 40.3 | 31.2 | 25.8 | 33.7 | 23.8 |
| GRPO | 38.3 ± 0.4 | 42.4 ± 0.9 | 37.2 ± 0.4 | 23.4 ± 3.2 | 30.5 ± 4.2 | 21.6 ± 3.0 |
| SFT | 46.6 ± 0.0 | **51.3 ± 0.3** | 45.4 ± 0.1 | 28.9 ± 0.3 | 38.2 ± 0.6 | 26.6 ± 0.2 |
| SFT + GRPO | 44.1 ± 1.4 | 48.8 ± 1.9 | 42.9 ± 1.2 | 21.9 ± 6.8 | 29.0 ± 8.9 | 20.1 ± 6.3 |
| M2A | **46.8 ± 0.8** | 50.1 ± 0.9 | **46.0 ± 0.7** | **30.9 ± 0.9** | **38.5 ± 1.3** | **29.1 ± 0.8** |

Table 9: Accuracy of M2A trained with various languages. All languages denote that the language translator randomly translates the question language into ten different languages used in MGSM. Bold denotes the best performance for each metric.

| Translation Language | GSM8K | | | MGSM | | |
|---|---|---|---|---|---|---|
| | Joint | Math. | Lang. | Joint | Math. | Lang. |
| All | **87.3 ± 0.1** | **87.3 ± 0.1** | **100.0 ± 0.0** | **58.1 ± 0.4** | **59.0 ± 0.3** | **97.8 ± 0.2** |
| Japanese | 86.8 ± 0.1 | 86.8 ± 0.1 | 100.0 ± 0.0 | 56.5 ± 0.1 | 59.0 ± 0.2 | 90.7 ± 0.4 |
| Swahili | 85.3 ± 0.7 | 85.3 ± 0.7 | 100.0 ± 0.0 | 56.6 ± 0.7 | 58.4 ± 0.3 | 96.2 ± 1.0 |

Table 10: Comparison of model performance on average reasoning score (%), and answer accuracy (%) on GeoFact-X test set. Results are partitioned into Assoc. (the language is natively associated with the content's country) and Non-Assoc. (no native association). Bold values indicate top performance. Bold means the best performance.

| Model | Average Reasoning Score (%) | | | Average Answer Accuracy (%) | | |
|---|---|---|---|---|---|---|
| | All | Assoc. | Non-Assoc. | All | Assoc. | Non-Assoc. |
| DeepSeek-R1-Distill-Llama-8B | 13.8 | 16.7 | 13.1 | 8.4 | 10.8 | 7.8 |
| DeepSeek-R1-Distill-Qwen-7B | 13.8 | 16.7 | 13.1 | 7.2 | 10.3 | 6.4 |
| Command R7B | 33.1 | 40.3 | 31.2 | 25.8 | 33.7 | 23.8 |
| Qwen-2.5-Instruct | 30.4 | 38.5 | 28.3 | 26.2 | 33.7 | 24.3 |
| GRPO | 45.4 ± 0.2 | 48.1 ± 0.1 | 44.8 ± 0.3 | **32.1 ± 0.3** | 37.6 ± 0.2 | **30.7 ± 0.3** |
| SFT | 47.6 ± 0.1 | 50.7 ± 0.5 | 46.9 ± 0.1 | 29.3 ± 0.2 | 37.1 ± 1.0 | 27.3 ± 0.3 |
| SFT + GRPO | 26.9 ± 0.9 | 29.2 ± 0.9 | 26.4 ± 0.9 | 10.7 ± 0.9 | 14.7 ± 1.7 | 9.7 ± 0.7 |
| M2A | **48.5 ± 0.4** | **52.6 ± 0.5** | **47.5 ± 0.3** | 32.0 ± 0.6 | **41.3 ± 1.0** | 29.7 ± 0.5 |

Table 11: Machine-translated performance of each model on GeoFact-X test set. Google Translate is used to translate the generated output into the question language. Bold means the best performance.

| Model with Machine Translation | Average Reasoning Score (%) | | | Average Answer Accuracy (%) | | |
|---|---|---|---|---|---|---|
| | All | Assoc. | Non-Assoc. | All | Assoc. | Non-Assoc. |
| DeepSeek-R1-Distill-Llama-8B | 27.6 | 29.2 | 27.2 | 7.9 | 11.1 | 7.1 |
| DeepSeek-R1-Distill-Qwen-7B | 33.2 | 33.9 | 33.1 | 8.8 | 11.1 | 8.2 |
| Command R7B | 44.2 | 48.5 | 43.1 | 25.0 | 34.1 | 22.7 |
| Qwen-2.5-Insturct | 45.7 | 49.2 | 44.8 | 28.9 | 36.4 | 27.0 |
| GRPO | 45.7 ± 0.3 | 48.3 ± 0.3 | 45.1 ± 0.3 | 31.9 ± 0.3 | 37.6 ± 0.8 | 30.5 ± 0.2 |
| SFT | 47.8 ± 0.1 | 50.9 ± 0.4 | 47.1 ± 0.1 | 27.2 ± 0.1 | 35.8 ± 1.1 | 25.0 ± 0.3 |
| SFT + GRPO | 47.7 ± 0.2 | 51.1 ± 0.2 | 46.8 ± 0.3 | **34.0 ± 0.1** | 39.1 ± 0.7 | **32.7 ± 0.1** |
| M2A | **48.7 ± 0.4** | **52.8 ± 0.5** | **47.7 ± 0.4** | 31.8 ± 1.0 | **41.3 ± 1.2** | 29.4 ± 0.9 |

## H MGSM Evaluation in Each Language

Table 12 shows individual math and language accuracy change compared to the base model (Qwen2.5-

Table 12: Accuracy of base (Qwen2.5-7B-Instruct) model and models fine-tuned with each post-training method on MGSM. Standard error is not included for readability. Bold means the best performance.

| Model | MGSM | Question Language | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | bn | de | es | fr | ja | ru | sw | te | th | zh |
| Joint Performance | | | | | | | | | | | |
| Qwen2.5-Instruct | 58.1 | **61.2** | 72.0 | 72.8 | 62.4 | 65.2 | 65.6 | **14.0** | **29.6** | **68.8** | 69.2 |
| GRPO | **58.2** | 59.6 | 69.7 | 75.9 | 65.6 | 66.9 | 71.7 | 7.7 | 27.7 | 64.5 | 72.4 |
| SFT (s1) | 21.9 | 5.6 | 62.8 | 68.6 | 4.6 | 2.4 | 1.2 | 0.0 | 5.8 | 18.6 | 49.0 |
| SFT+GRPO | 23.3 | 3.6 | 55.7 | 74.4 | 1.6 | 16.3 | 2.1 | 0.0 | 3.9 | 11.3 | 63.9 |
| SFT on s1K-X | 45.0 | 34.7 | 51.2 | 60.8 | 69.3 | 44.0 | 52.5 | 10.4 | 8.8 | 64.7 | 53.7 |
| M2A | 58.1 | 53.2 | **74.1** | **74.3** | **71.2** | **66.3** | **80.4** | 2.7 | 14.9 | 66.0 | **78.3** |
| Math Performance | | | | | | | | | | | |
| Qwen2.5-Instruct | 58.7 | 61.2 | 72.0 | 72.8 | 62.4 | 70.4 | 65.6 | 14.0 | 29.6 | 69.6 | 69.2 |
| GRPO | 58.8 | 59.7 | 69.7 | 75.9 | 65.7 | 66.9 | 71.7 | 12.8 | 27.7 | 64.9 | 72.4 |
| SFT (s1) | **66.7** | **66.4** | **77.4** | **78.0** | **79.6** | **73.6** | **83.6** | **18.6** | **35.2** | **74.6** | **79.8** |
| SFT on s1K-X | 45.2 | 34.7 | 51.2 | 60.8 | 69.3 | 44.0 | 52.5 | 10.4 | 8.8 | 64.7 | 55.1 |
| SFT+GRPO | 62.9 | 62.1 | 75.6 | 75.5 | 75.3 | 69.1 | 79.5 | 17.9 | 31.5 | 71.7 | 70.5 |
| M2A | 59.0 | 53.3 | 75.6 | 75.2 | 75.7 | 66.3 | 80.5 | 3.7 | 14.9 | 66.5 | 78.3 |
| Language Performance | | | | | | | | | | | |
| Qwen2.5-Instruct | 99.0 | **100.0** | **100.0** | **100.0** | **100.0** | 91.6 | **100.0** | **100.0** | **100.0** | 98.8 | **100.0** |
| GRPO | 95.9 | 99.9 | **100.0** | **100.0** | 99.9 | **100.0** | **100.0** | 60.0 | **100.0** | 99.6 | **100.0** |
| SFT (s1) | 31.0 | 13.6 | 81.4 | 88.6 | 5.8 | 6.0 | 2.6 | 0.0 | 22.8 | 27.8 | 62.2 |
| SFT on s1K-X | **99.7** | **100.0** | 99.9 | 99.9 | 99.9 | **100.0** | 99.7 | 99.9 | **100.0** | **100.0** | 98.0 |
| SFT+GRPO | 35.1 | 11.2 | 74.4 | 98.3 | 1.7 | 29.2 | 3.7 | 0.0 | 18.7 | 23.1 | 91.2 |
| M2A | 97.8 | 99.7 | 98.0 | 98.5 | 94.8 | 99.9 | 99.2 | 88.3 | 99.9 | 99.3 | **100.0** |

7B-Instruct) in each language. As we mentioned in Section 5, supervised fine-tuning on s1K-1.1 improves math performance while losing language performance. Conversely, GRPO rarely changes performance on both mathematics and language across all languages. M2A generally maintains language performance. However, its German performance is much lower, which might be related to the language sampled for translating the question.

We also illustrate an MGSM Russian example response from Qwen2.5-7B-Instruct and SFT, GRPO, M2A fine-tuned models on Figure 13. All models generate the correct answer from a given question written in Russian. However, the SFT model uses English instead of Russian, while others reason in Russian. GRPO has an almost identical reasoning process to the base model, which might explain why it performs almost the same in any metrics as the base model.

# I GeoFact-X Evaluation in Each Language and Country Pair

Tables 13, and 14 present reasoning score, language accuracy, and answer accuracy for each language–country pair. Diagonal entries correspond to associative pairs, while off-diagonal entries correspond to non-associative pairs. Associative countries consistently achieve higher performance across all metrics compared to non-associative ones. English (a high-resource language) generally outperforms other languages. Among non-associative cases, performance for the United States exceeds that of Kenya and Thailand, which may reflect differences in pre-training and fine-tuning data exposure. As discussed in Section 5.2, all methods substantially increase language accuracy as the training dataset includes all five languages. However, no consistent trend is observed in reasoning score and answer accuracy across language–country pairs.

Table 13: Average reasoning score (%) by language and region. Reasoning quality is assessed using an LLM-as-a-judge framework, which evaluates model-generated justifications against reference Gemini 2.0 Flash reasoning traces in the GeoFact-X dataset. Higher scores indicate more coherent, relevant, and logically sound reasoning. The gray diagonal entries represent associated language–country pairs. Bold means the best performance in each pair.

| Language | Model | USA | India | Japan | Kenya | Thailand |
|---|---|---|---|---|---|---|
| English | Qwen2.5-Instruct | 67.5 | 55.7 | **62.8** | 56.6 | 51.4 |
| | GRPO | **70.3 ± 1.2** | 61.8 ± 1.3 | 58.5 ± 1.2 | 57.2 ± 0.7 | **51.9 ± 0.6** |
| | SFT | 69.2 ± 0.6 | 62.9 ± 0.7 | 57.6 ± 0.6 | 57.4 ± 0.1 | 51.2 ± 0.7 |
| | M2A | 69.3 ± 0.5 | **63.0 ± 0.7** | 57.8 ± 0.7 | **57.4 ± 0.2** | 51.5 ± 0.8 |
| Hindi | Qwen2.5-Instruct | **39.4** | **39.0** | **39.0** | 35.0 | 38.6 |
| | GRPO | 38.7 ± 1.4 | 36.6 ± 0.9 | 38.2 ± 0.7 | 35.5 ± 0.7 | 38.2 ± 0.9 |
| | SFT | 36.3 ± 1.0 | 36.0 ± 0.5 | 37.7 ± 1.4 | 42.4 ± 0.5 | **40.8 ± 0.9** |
| | M2A | 36.5 ± 1.1 | 36.0 ± 0.4 | 38.2 ± 1.1 | **42.5 ± 0.5** | 40.8 ± 0.9 |
| Japanese | Qwen2.5-Instruct | **51.4** | 43.1 | **53.2** | 45.2 | 40.3 |
| | GRPO | 50.1 ± 1.4 | **43.2 ± 0.6** | 52.3 ± 0.1 | 47.0 ± 1.2 | 43.5 ± 0.6 |
| | SFT | 45.3 ± 1.1 | 43.0 ± 0.6 | 52.3 ± 1.3 | 47.7 ± 0.7 | 43.0 ± 0.2 |
| | M2A | 45.5 ± 1.1 | 43.0 ± 0.5 | 52.6 ± 1.4 | **48.1 ± 0.7** | **43.5 ± 0.3** |
| Swahili | Qwen2.5-Instruct | 41.4 | 43.3 | 39.5 | 41.3 | 34.9 |
| | GRPO | 34.6 ± 0.5 | 33.2 ± 1.4 | 31.4 ± 0.6 | 35.6 ± 1.2 | 33.8 ± 0.4 |
| | SFT | 47.9 ± 0.2 | **49.8 ± 1.3** | 47.4 ± 0.2 | 49.6 ± 0.4 | 47.3 ± 0.7 |
| | M2A | **48.0 ± 0.3** | **49.8 ± 1.3** | **47.5 ± 0.2** | **49.7 ± 0.4** | **47.3 ± 0.8** |
| Thai | Qwen2.5-Instruct | **56.4** | 44.0 | **49.0** | 43.5 | 45.3 |
| | GRPO | 54.1 ± 1.4 | **47.5 ± 0.3** | 46.7 ± 1.6 | **47.8 ± 0.6** | 45.0 ± 0.6 |
| | SFT | 40.3 ± 0.8 | 40.2 ± 0.3 | 47.2 ± 0.6 | 45.1 ± 0.1 | 45.5 ± 0.4 |
| | M2A | 40.3 ± 0.6 | 40.5 ± 0.3 | 47.5 ± 0.6 | 45.5 ± 0.2 | **45.6 ± 0.4** |

Table 14: Average answer accuracy by language and region. The gray diagonal entries represent associated language–country pairs. Bold means the best performance in each pair.

| Lang | Model | USA | India | Japan | Kenya | Thailand |
|---|---|---|---|---|---|---|
| English | Qwen2.5-Instruct | 56.3 | 32.7 | 41.9 | 31.6 | 24.4 |
| | GRPO | **74.7 ± 1.1** | **62.9 ± 2.1** | **52.0 ± 2.4** | **49.1 ± 2.1** | **38.4 ± 0.7** |
| | SFT | 70.9 ± 1.9 | 60.2 ± 2.0 | 47.0 ± 1.3 | 45.3 ± 0.6 | 34.5 ± 1.0 |
| | M2A | 65.6 ± 2.2 | 59.0 ± 3.0 | 51.7 ± 2.2 | 43.1 ± 5.6 | 31.1 ± 2.3 |
| Hindi | Qwen2.5-Instruct | 19.5 | **23.3** | **16.1** | 14.3 | 20.0 |
| | GRPO | **22.8 ± 1.5** | 22.1 ± 1.8 | 11.9 ± 2.0 | 13.9 ± 1.0 | **20.8 ± 0.8** |
| | SFT | 11.4 ± 1.6 | 14.7 ± 2.1 | 10.0 ± 0.4 | 18.3 ± 0.4 | 15.0 ± 1.4 |
| | M2A | 15.1 ± 1.8 | 19.7 ± 1.7 | 13.3 ± 1.9 | **19.8 ± 1.2** | 14.5 ± 2.4 |
| Japanese | Qwen2.5-Instruct | **36.8** | 23.0 | 39.3 | 20.2 | 18.9 |
| | GRPO | 34.6 ± 2.7 | **25.3 ± 1.8** | 40.5 ± 2.1 | 29.0 ± 1.4 | **26.6 ± 2.4** |
| | SFT | 32.9 ± 1.3 | 20.3 ± 0.8 | 42.1 ± 4.0 | **29.4 ± 0.8** | 19.8 ± 1.2 |
| | M2A | 30.8 ± 2.6 | 25.1 ± 0.6 | **42.5 ± 2.0** | 27.9 ± 3.1 | 20.3 ± 2.0 |
| Swahili | Qwen2.5-Instruct | 24.3 | 32.6 | 17.6 | 22.6 | 13.5 |
| | GRPO | 29.3 ± 0.9 | 25.6 ± 1.8 | 15.5 ± 0.4 | 19.6 ± 1.7 | 18.9 ± 0.0 |
| | SFT | 32.4 ± 2.1 | **34.5 ± 2.5** | **21.7 ± 0.8** | 26.7 ± 1.4 | **25.7 ± 2.1** |
| | M2A | **33.3 ± 3.5** | 33.3 ± 1.1 | 21.1 ± 1.8 | **28.3 ± 3.4** | 22.2 ± 0.7 |
| Thai | Qwen2.5-Instruct | 31.9 | 23.8 | 18.5 | 20.7 | 25.7 |
| | GRPO | **39.6 ± 1.7** | **29.2 ± 0.8** | **29.2 ± 1.1** | **30.1 ± 2.3** | 29.3 ± 3.2 |
| | SFT | 15.0 ± 0.5 | 19.6 ± 1.1 | 20.6 ± 0.4 | 19.5 ± 0.7 | **29.3 ± 0.5** |
| | M2A | 28.6 ± 3.4 | 23.3 ± 0.0 | 27.8 ± 3.6 | 28.6 ± 2.4 | 24.3 ± 1.8 |

**Question**

'<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are a helpful assistant.<|im_end|>
<|im_start|>user\nЗадача: Утки Дженет несут по 16 яиц в день. Она ест три на завтрак каждое утро и печет кексы для своих друзей каждый день, используя четыре. Остаток она  аждый день продает на фермерском рынке по 2 $ за свежее утиное яйцо. Сколько в долларах она зарабатывает на фермерском рынке каждый день?\nПошаговоерешение:<|im_end|>
<|im_start|>assistant

**Qwen2.5-7B-Instruct**

Давайте решим эту задачу пошагово:
1. Сначала определим, сколько яиц остается после того, как утки Дженет съедают и используют их:
- Всего яиц в день: 16
- Яиц на завтрак: 3
- Яиц для кексов: 4
- Остаток: 16 - 3 - 4 = 9 яиц

2. Теперь определим, сколько Дженет продает на фермерском рынке:
- Она продает все оставшиеся яйца: 9 яиц

3. Посчитаем, сколько денег она зарабатывает:
- Цена за яйцо: 2 $
- За 9 яиц: 9 * 2 = 18 $

Таким образом, Дженет зарабатывает **18** долларов на фермерском рынке каждый день.

**Supervised Fine-Tuning (s1)**

Okay, let's try to figure out how much money Janet's ducks make each day from selling eggs at the farmer's market. Let me break down the problem step by step.

First, the problem states that Janet's ducks lay 16 eggs each day. That's the total production. Then, Janet eats 3 eggs every morning for breakfast. So, those 3 eggs are consumed by her. Additionally, she uses 4 eggs each day to bake cakes for her friends. So, both her consumption and baking take a total of 3 + 4 = 7 eggs per day.

**...**

But just to make sure I didn't misread the question. Let me re-read it:

"Утки **...** день?"

Which translates to: "Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes cakes for her friends using four eggs every day. The remainder she sells at the farmer's market for $2 per fresh duck egg each day. How much does she earn at the farmer's market each day in dollars?"

Yes, the question is asking about her daily earnings from the farmer's market. So, the calculation is correct.

Therefore, the final answer is \boxed{18}.
<|im_start|>answer
Answer: Janet's ducks lay 16 eggs each day. She eats 3 eggs every morning and uses 4 eggs to bake cakes for her friends. The total number of eggs used each day is 3 + 4 = 7. The remaining eggs are 16 - 7 = 9. She sells each egg for $2, so her daily earnings are 9 * $2 = $18.

**\boxed{18}**

**GRPO**

Давайте решим эту задачу пошагово:

1. Сначала определим, сколько яиц остается после того, как утки Дженет съедают и используют их для кексов:
- Всего яиц в день: 16
- Яиц, съеденных на завтрак: 3
- Яиц, использованных для кексов: 4
- Остаток яиц: 16 - 3 - 4 = 9

2. Теперь определим, сколько денег Дженет зарабатывает на продаже оставшихся яиц:
- Цена за одно яйцо: 2 $
- Количество продаваемых яиц: 9
- Общая сумма за день: 9 * 2 = 18 $

Таким образом, Дженет зарабатывает **18** долларов на фермерском рынке каждый день.

**M2A**

Okay, давайте решим эту задачу шаг за шагом. У нас есть информация о том, что утки Дженет несут 16 яиц в день. Она сама ест три яйца на завтрак каждый день и использует еще четыре яйца для печения кексов для своих друзей. Остаток яиц она продает на фермерском рынке по 2 доллара за каждое свежее утиное яйцо. Нам нужно определить, сколько она зарабатывает на фермерском рынке каждый день.

**Шаг 1: Вычислить количество яиц, которые Дженет использует на свой завтрак и для кексов.**

Дженет ест 3 яйца на завтрак каждый день. Кроме того, она использует 4 яйца для кексов для друзей каждый день. Всего яиц, которые она использует сама или дает друзьям, будет 3 + 4 = 7 яиц в день.

**Шаг 2: Вычислить количество оставшихся яиц, которые она может продать.**

Утки несут 16 яиц в день. Если из них 7 яйца используются Дженет и ее друзьями, то оставшихся будет 16 - 7 = 9 яиц. Эти 9 яиц она продает на фермерском рынке.

**Шаг 3: Вычислить доход от продажи оставшихся яиц.**

Каждое яйцо продаётся по 2 доллара. Поэтому, если она продает 9 яиц по 2 доллара каждое, то доход составит 9 * 2 = 18 долларов.

**Ответ:** Дженет зарабатывает 18 долларов на фермерском рынке каждый день.

**Проверка:** Давайте проверим свои вычисления. Яйца в день: 16. Используется на завтрак и кексы: 3 + 4 = 7. Осталось: 16 - 7 = 9. Цена: 9 * 2 = 18. Все верно. Ответ: 18 долларов.

**Конечный ответ:** \boxed{18}
<|im_start|>answer
Ответ: Утки Дженет несут 16 яиц в день. Она использует 3 яйца на завтрак и 4 яйца для кексов, что в сумме составляет 3 + 4 = 7 яиц. Остаток яиц, который она продает на фермерском рынке, составляет 16 - 7 = 9 яиц. По 2 долларов за каждое яйцо она зарабатывает 9 * 2 = 18 долларов в день.

**\boxed{18}**

Figure 13: **Example question and response of each model on MGSM Russian.** The question in English is "Janet's ducks lay sixteen eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for two dollars per fresh duck egg. How much in dollars does she make every day at the farmers' market?" All models correctly provide the answer (18), but the SFT model conducts its reasoning in English. GRPO's reasoning process is almost identical to that of the base model (Qwen2.5-7B-Instruct).

| Model | Q: Safari ya Lewis na Clark ilianza wapi? (Where did Lewis and Clark's journey begin?) A: St. Louis |
|---|---|
| Base | Swali linauliza kuhusu mahali ambapo safari ya Lewis na Clark ilianza.<br>>Safari ya Lewis na Clark ilikuwa safari ya kuvuka juu ya maji ya Kimarekani. Ililianza mnamo 1804 na ilishuhudia kipindi cha kwanza cha Marekani kushoto mashariki.<br>>Safari hii ilianza Fort Atkinson, ambayo ilikuwa fortuoni ya watalii huko Council Bluffs, Iowa, na Omaha, Nebraska. Kwa hivyo, majibu yote mawili yanakubalika.<br><answer>Fort Atkinson</answer><br><br>**Translated Output:**<br>The question asks where the Lewis and Clark expedition began.<br>>The Lewis and Clark expedition was a voyage across American waters. It began in 1804 and marked the first American expedition eastward.<br>>The expedition began at Fort Atkinson, which was a tourist fort in Council Bluffs, Iowa, and Omaha, Nebraska. Therefore, both answers are acceptable.<br><answer>Fort Atkinson</answer> |
| GRPO | Safari ya Lewis na Clark ilianza wapi?<br>Hii inatafsiriwa kama safari ya kawaida kwa ajili ya kujenga kazi kwa kawaida na kujenga mazingira na kawaida.<br>Iliopiga safari hili, Lewis na Clark inatafsiriwa kama kawaida kwa ajili ya kujenga mazingira na kawaida na kujenga kazi kwa kawaida.<br>Safari hili ina kujenga mazingira na kawaida kwa ajili ya kujenga kazi kwa kawaida.<br><answer> Safari ya Lewis na Clark ilianza kwa ajili ya kujenga mazingira na kawaida na kujenga kazi kwa kawaida. </answer><br><br>**Translated Output:**<br>The Lewis and Clark expedition began in 1804 and 1806.<br>To help with births and births and births.<br>To learn if there are countries with births and births and births and long periods, Lewis and Clark began in 1804 and 1806.<br>To learn if there are countries with births and births and long periods, Lewis and Clark began in 1804 and 1806.<br><answer> The Lewis and Clark expedition began in 1804 and 1806. </answer> |
| SFT | Swali linauliza kuhusu mahali ambapo safari ya Lewis na Clark ilianza. Safari hii ilikuwa safari ya kuvuka juu ya maji ya Katika Marekani baada ya ununuzi wa Louisiana.<br><step> Meriwether Lewis alikuwa mwanamke mtawalia ambaye aliongoza safari ya Lewis na Clark. Alikuwa pia vice presidenta wa Theodore Roosevelt.<br><step> Safari ya Lewis na Clark ilianza Washington, D.C. Washington, D.C. ndio eneo kuu la bandari nchini Marekani na linajulikana kwa kuwa kituo cha kikatiba cha Marekani.<br><step> Jiji la Washington ni jina lingine la Washington, D.C., hivyo majibu yote mawili ni sahihi.<br><answer>Washington, D.C.</answer><br><br>**Translated Output:**<br>The question asks where the Lewis and Clark expedition began. This expedition was a journey across the waters of the United States after the Louisiana Purchase.<br><step> Meriwether Lewis was a woman who led the Lewis and Clark expedition. She was also Theodore Roosevelt's vice president.<br><step> The Lewis and Clark expedition began in Washington, D.C. Washington, D.C. is the largest port in the United States and is known for being the constitutional capital of the United States.<br><step> The city of Washington is another name for Washington, D.C., so both answers are correct.<br><answer>Washington, D.C.</answer> |
| M2A | Swali linauliza kuhusu mahali ambapo safari ya Lewis na Clark ilianza.<br>Safari ya Lewis na Clark ilikuwa safari ya kuvuka juu ya ardhi ya Amerika Kaskazini ili kuashiria maji ya Missouri na kubadilisha mapema ya Marekani.<br>Ilianzishwa mwaka 1804 na ilishirikisha Merikani.<br>Utafiti wa kihistoria na kuvunjika kwenye muktadha wa safari huo unaonyesha kwamba ilianza St. Louis, Missouri.<br>St. Louis iko katika jimbo la Missouri, makao ya ardhi ya Missouri.<br><answer>St. Louis</answer><br><br>**Translated Output:**<br>The question asks where the Lewis and Clark expedition began.<br>The Lewis and Clark expedition was a journey across the land of North America to chart the course of the Missouri River and the early development of the United States.<br>It began in 1804 and involved Americans. Historical research and contextual analysis of the expedition indicate that it began in St. Louis, Missouri.<br>St. Louis is located in the state of Missouri, home to the Missouri Territory.<br><answer>St. Louis</answer> |

Figure 14: **Generated outputs from a given question written in Swahili on GEOFACT-X.** All models use Swahili, but only M2A generates the correct answer, St Louis.