

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 LEARN GLOBALLY, SPEAK LOCALLY: BRIDGING THE GAPS IN MULTILINGUAL REASONING

Anonymous authors

Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have achieved strong performance in domains like mathematics, factual question answering, and code generation, yet their ability to reason on these tasks in different languages remains underdeveloped. Especially for low-resource languages such as Swahili or Thai, LLMs can often misinterpret prompts or default to reasoning in English. This implicit bias toward high-resource languages undermines factual accuracy, interpretability, and trust. We propose M2A, a novel method that combines multi-scale multilingual alignment with language-consistency rewards on machine-translated questions, training models to reason directly and accurately in the target language. Furthermore, existing multilingual benchmarks only evaluate on final answers, overlooking whether reasoning occurs in the intended language. To close this gap, we introduce GEOFACT-X, a geography-based multilingual factual reasoning benchmark together with reasoning traces in five languages: English, Hindi, Japanese, Swahili, and Thai. Our results show that M2A significantly enhances multilingual reasoning fidelity in both mathematical and factual reasoning tasks, highlighting that reasoning-aware multilingual reinforcement learning is crucial for robust cross-lingual generalization.

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, or Thai (Cahyawijaya et al., 2024; Nguyen et al., 2023). This performance disparity undermines the *trustworthiness* of LLMs in these languages since users cannot check the reasoning traces to verify the answer. For this purpose, both the final answer and the intermediate reasoning should ideally be expressed in the question language to ensure *interpretability*, *i.e.*, users can directly follow the reasoning in their own language. The central issue is not only whether LLMs can provide correct answers, but whether they can *think* in the language of the question. When they cannot, translation of reasoning traces offers only a partial solution and one that often fails on cultural and linguistic nuance. Recent studies (Aggarwal et al., 2025; Yao et al., 2024) suggest that both LLMs and machine translation systems struggle with cultural and linguistic nuances. For example, culturally grounded concepts such as Chinese *Guānxì* (系), Japanese *Wa* (和), and Korean *Jeong* (정) remain difficult to capture faithfully.

In this work, we conduct the first comprehensive study of **multilingual reasoning**: 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 found 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 a language. Ensuring both accuracy and language-consistent reasoning is essential for globally inclusive and interpretable AI.

To tackle this challenge, we introduce M2A (Multi-Scale Multilingual Alignment), an efficient approach that explicitly enforces language-consistent reasoning while retaining factual correctness.

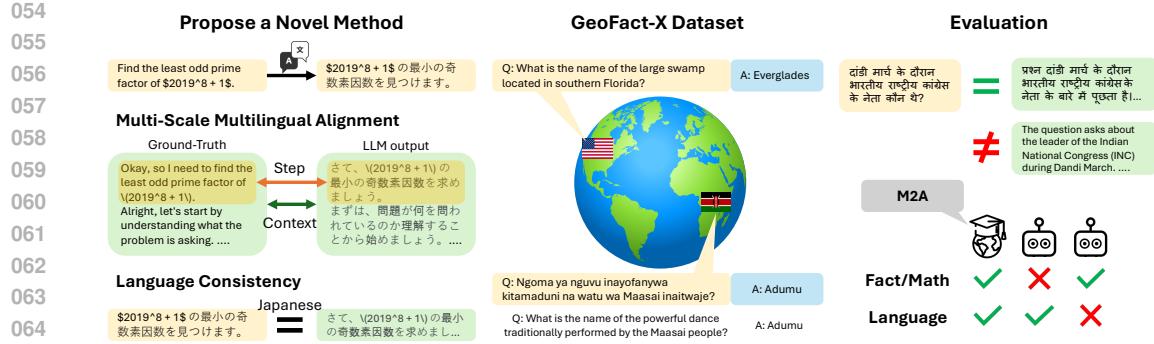


Figure 1: **Illustration of Contributions** We propose M2A, a new method that utilizes multi-scale multilingual alignment and language consistency rewards from a given machine-translated question, enabling reasoning in the question language. We also introduce GEOFACT-X, a new multilingual factual reasoning benchmark which includes training datasets and step-by-step reasoning traces across five languages. We propose 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.

The key idea is to combine multi-scale multilingual reasoning alignment with a language-consistency reward, providing reinforcement-learning signals that encourage reasoning traces to remain in the question language, enabling reasoning capabilities to be learned without ground-truth supervision in that language. We jointly employ supervised fine-tuning and group relative policy optimization (GRPO) to integrate supervised learning on ground-truth reasoning traces with reinforcement-based refinement. Unlike prior work (Guo et al., 2025; Liu et al., 2025; Ranaldi & Pucci, 2025) that optimizes only for correctness or formality, our method targets the *alignment of reasoning itself*.

Despite recent progress in multilingual LLMs (Ahuja et al., 2023; Qin et al., 2025), their ability to perform factual reasoning 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, Swahili, Thai). By grounding evaluation in country-specific knowledge, GEOFACT-X enables systematic assessment of whether LLMs can reason faithfully within linguistically and culturally contextualized spaces.

We train M2A on the s1K-1.1 (Muennighoff et al., 2023) dataset and GEOFACT-X train set for mathematical and factual reasoning, respectively. Our experiments demonstrate that M2A yields significant improvement on multilingual reasoning in both cases while reasoning in the question languages. Figure 1 summarizes our key contributions. Together with the release of our code, data, and evaluation protocols, our work provides a foundation for future work on multilingual reasoning.

2 RELATED WORK

2.1 MULTILINGUALISM IN LARGE LANGUAGE MODELS

Recent advances in multilingual large language models (LLMs) mark a shift from monolingual dominance to more inclusive cross-lingual capabilities. Two perspectives frame much of this discourse. The scaling view holds that increasing the volume and diversity of multilingual data during pretraining enhances cross-lingual generalization (Xue et al., 2021; Conneau et al., 2020; Chang et al., 2023; Gurgurov et al., 2024), though it faces the *curse of multilinguality*, where accommodating many languages dilutes performance in individual ones due to limited parameter capacity. The optimization view emphasizes careful fine-tuning as a means of preserving and amplifying multilingual knowledge (Devlin et al., 2019; Conneau et al., 2020; Luo et al., 2023; Zhai et al., 2023), yet aggressive post-training risks catastrophic forgetting by overwriting deeply embedded linguistic priors. Beyond scaling and optimization, Schut et al. (2025) and Zhong et al. (2025) analyzed the internal representations of multilingual models, while Yong et al. (2025) proposed test-time strategies for improving cross-lingual reasoning. Notably, Schut et al. (2025) found that multilingual LLMs often perform intermediate reasoning in English, but their scope was limited to internal representation analysis rather than generated reasoning traces. More recently, post-training alignment techniques

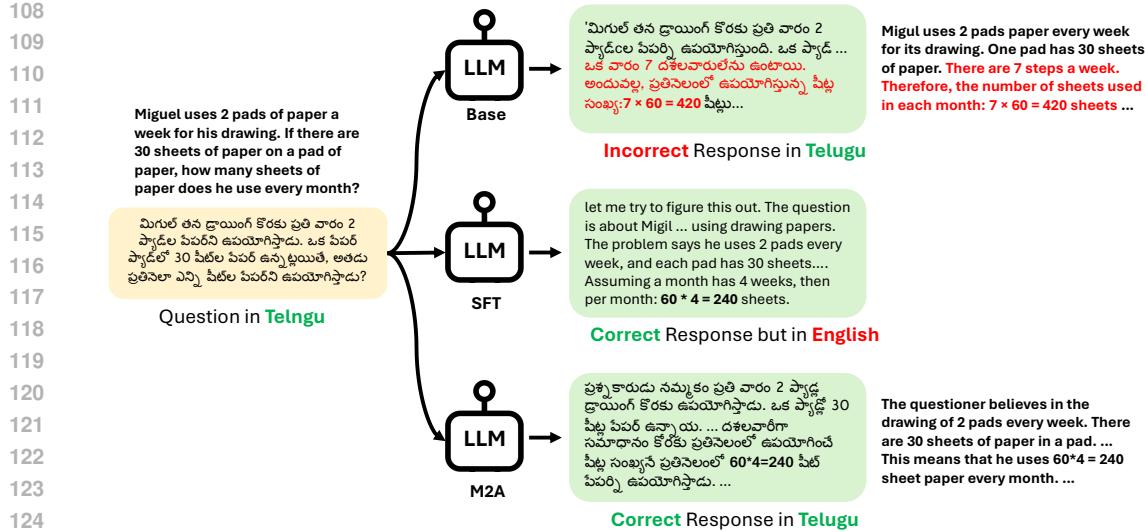


Figure 2: **Example of outputs from a Telugu question from three different models.** Base LLM and supervised fine-tuned models (SFT) are correct only in either answer or language, whereas our M2A is correct in both answer and language.

such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) have been applied to embed multilingual reasoning (Dang et al., 2024; Ranaldi & Pucci, 2025). Similarly, we employ Group Relative Policy Optimization (GRPO) (Guo et al., 2025; Shao et al., 2024) with multi-scale multilingual alignment and language consistency rewards, leveraging explicit reasoning traces in the original languages to build stronger multilingual reasoning capabilities.

2.2 EVALUATION OF MULTILINGUAL REASONING 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), XWino-grad (Tikhonov & 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 reasoning 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 M2A: MULTI-SCALE MULTILINGUAL ALIGNMENT

We propose a new method to enrich existing English-based reasoning models with multilingual reasoning capabilities. Our approach combines the complementary strengths of supervised fine-tuning (SFT) and reinforcement learning (RL). While SFT enables base reasoning capabilities in English, we find that these models struggle with multilingual reasoning during test time as shown in Figure 2. To that end, we propose a new test-time RL method for multilingual reasoning. Key to this new approach is defining the right set of rewards that incentivize the model to reason consistently across different languages. We first translate each question into multiple languages by using Google Translate, allowing the model to generate outputs conditioned on the translated inputs. We then define a set of multi-scale rewards across different multilingual granularities. A context alignment reward measures multilingual alignment across the entire reasoning context with the original ground-truth reasoning trace. This is followed by a reasoning-step alignment reward that aligns each individual reasoning step to capture fine-grained correspondence. Finally, a language consistency reward explicitly enforces reasoning in the question language.

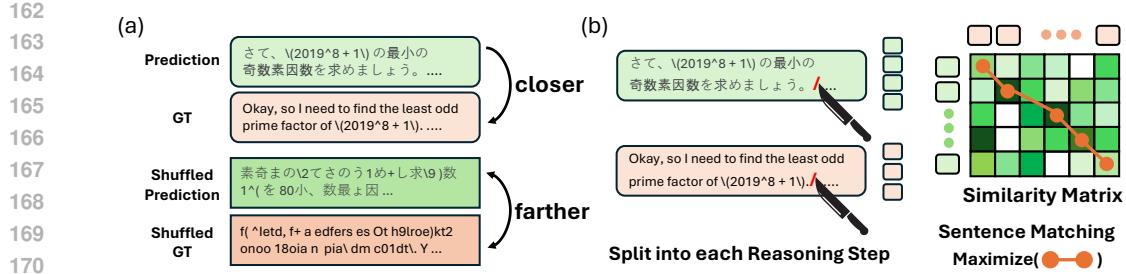


Figure 3: **Overview of M2A.** (a) *Multilingual Context Alignment* enforces global similarity between generated and reference responses while discouraging trivial matches via shuffled negatives. (b) *Multilingual Reasoning-Step Alignment* provides finer-grained supervision by aligning individual reasoning steps with ground-truth traces using dynamic programming.

Problem Setup. Given the question sampled from the question dataset, $q \sim P(Q)$ and its corresponding ground-truth response y , we translate each question to the target language l , q' by using a machine translator (e.g., Google Translate). GRPO generates a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the translated question. The reward r_t is calculated based on each output o_t .

Multilingual Context Alignment. We first encode both output and the ground-truth with the encoder, ϕ , $z_o = \phi(o)$ and $z_y = \phi(y)$. We utilize mT5 (Xue et al., 2021) for encoding texts. The alignment reward can be the cosine similarity between two embeddings, $\cos(z_o, z_y)$. However, it is maximized when the generated output o is identical to y , ignoring the question language. To address this, we introduce negative samples by shuffling both outputs and ground-truth responses with the same permutation, $\tilde{z}_o = \phi(\psi(o))$ and $\tilde{z}_y = \phi(\psi(y))$, where ψ denotes the shuffle function. Inspired by Schroff et al. (2015), the multilingual context alignment maximizes similarity between positive samples and minimizes similarity between negative samples, enforcing a margin, α , between these similarities:

$$\cos(\tilde{z}_o, \tilde{z}_y) + \alpha < \cos(z_o, z_y). \quad (1)$$

The final context alignment reward is defined as:

$$r_{\text{context-align}} = \max(\cos(z_o, z_y) - \cos(\tilde{z}_o, \tilde{z}_y) + \alpha, 0), \quad (2)$$

where α denotes the margin, set to 1, the maximum possible value of cosine similarity.

Multilingual Reasoning-Step Alignment. We further introduce a multilingual reasoning-step alignment to provide finer-grained matching. Given the split output sentences, $\mathbf{o} = (o^{(1)}, \dots, o^{(N)})$ and ground-truth sentences $\mathbf{y} = (y^{(1)}, \dots, y^{(M)})$, each output sentence, $o^{(i)}$ is aligned with a ground-truth sentence, $y^{(j_i)}$. Since the number of output and reference sentences (N and M) may differ, we use dynamic programming to maximize the total similarity between the pairs while preserving order:

$$\max_{1 \leq j_1 \leq \dots \leq j_N \leq M} \sum_{i=1}^N \mathbf{C}_{i,j_i}, \quad (3)$$

where $\mathbf{C} \in \mathbb{R}^{N \times M}$ is the similarity matrix, and $\mathbf{C}_{i,j}$ denotes the similarity score between embeddings $z_o^{(i)}$ and $z_y^{(j)}$. We use the same function used in Eq. (2) for \mathbf{C} . The multilingual reasoning-step alignment reward is then defined as the average similarity across aligned pairs:

$$r_{\text{step-align}} = \frac{1}{N} \sum_{i=1}^N \mathbf{C}_{i,j_i} = \frac{1}{N} \sum_{i=1}^N \max(\cos(z_o^{(i)}, z_y^{(j_i)}) - \cos(\tilde{z}_o^{(i)}, \tilde{z}_y^{(j_i)}) + \alpha, 0). \quad (4)$$

Language Consistency. We also define a language consistency reward for giving a more direct incentive to reason in the question language. Given the output o , and language detector f (e.g., Google Translate, langid (Lui & Baldwin, 2012)), the language consistency reward is defined as 1 if the detected language in response matches the target language l , 0 if it does not:

$$r_{\text{lang}} = \delta[f(o) = l_t], \quad (5)$$

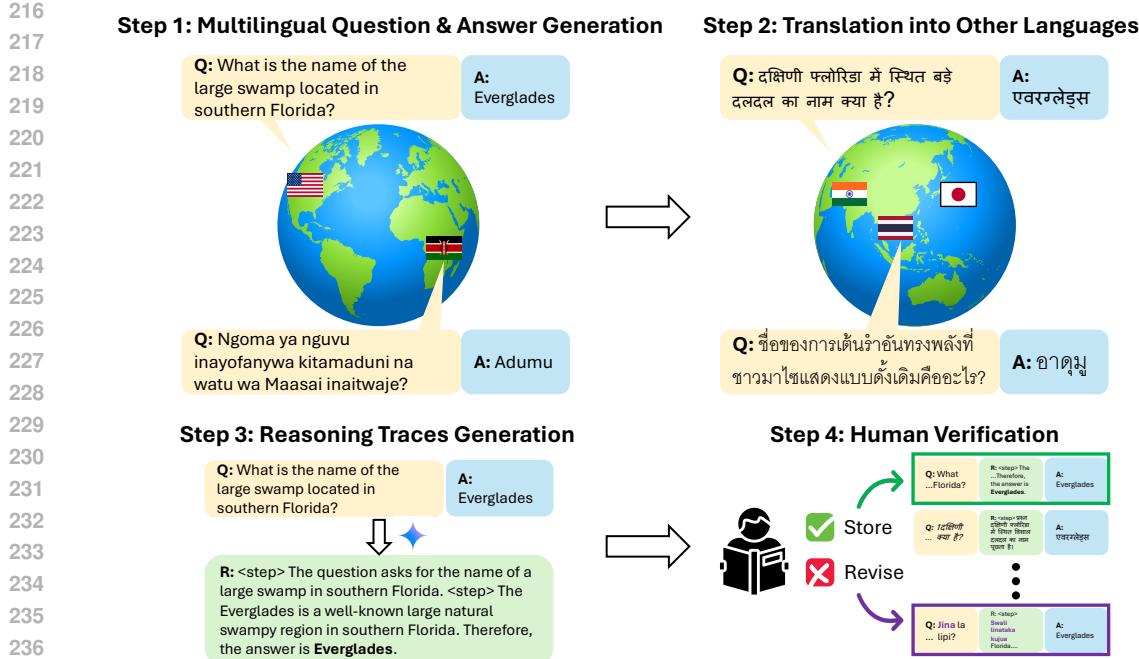


Figure 4: **Illustration of GEOFACT-X benchmark construction.** (1) Geography-aware multilingual questions and answers are generated by Gemini 2.0 Flash. (2) The data is translated into other languages, verifying whether it is back-translatable. (3) The reasoning trace for each question and answer pair 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	Reasoning Eval.
XStoryCloze (Lin et al., 2022)	1872	11		✓	
KWINO (Tikhonov & Ryabinin, 2021)	3961	6			
XCOPA(Ponti et al., 2020)	6600	11		✓(English only)	
X-FaKT(Agarwal et al., 2025)	2362	13			
XLQA (Roh et al., 2025)	3000	8	✓		
GEOFACT-X (ours)	12780	5	✓	✓	✓

where $\delta[\cdot]$ denotes the indicator function. The final reward is defined as the sum of individual reward:

$$r = r_{\text{context-align}} + r_{\text{step-align}} + r_{\text{lang}}. \quad (6)$$

4 GEOFACT-X: GEOGRAPHY-BASED FACTUAL REASONING BENCHMARK

Despite advances in multilingual LLMs (Ahuja et al., 2023; Qin et al., 2025), robust evaluation of factual reasoning across cultures remains underexplored. 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). Our goal is to capture country-specific factual knowledge, encouraging language models to reason effectively within culturally contextualized knowledge spaces. Table 1 compares GEOFACT-X with existing multilingual factual reasoning benchmarks. Our geography-aware multilingual benchmark has a training set and reasoning evaluations compared to other benchmarks.

Figure 4 illustrates the process of the dataset construction. We adopt a two-stage validation pipeline to ensure factual accuracy and dataset quality. Rule-based filters and cross-language checks remove incorrect or inconsistent pairs. Specifically, we verify cross-language answer consistency by translating each answer into English via the Google Translate API to identify mismatches. Gemini 2.0

270 Flash (Team et al., 2023) then generates structured chain-of-thought reasoning traces for each item,
 271 enhancing interpretability and providing supervision signals. We split the dataset into a train set
 272 (85%) and a test set (15%), ensuring no semantic overlap across splits, even across languages. All test
 273 samples are manually verified by native or C1-level speakers for factual correctness and linguistic
 274 clarity. Figure 6 illustrates an example multilingual question with its reasoning trace and final answer.

275 For evaluation, we measure answer accuracy and reasoning score. Answer accuracy is computed
 276 by comparing predictions against the reference answers. The reasoning score is assessed by Qwen-
 277 2.5-72B-Instruct (Qwen et al., 2025) as an LLM-as-a-judge, comparing model-generated reasoning
 278 traces against the human-verified reasoning traces in the test set. If a reasoning trace is produced
 279 in a language different from the question, identified by a language detector (Lui & Baldwin, 2012),
 280 its score is set to zero. We validate the reliability of this metric via a human agreement study in
 281 Appendix H. Detailed curation, distribution, and evaluation procedures are provided in Appendix A.

282

283 5 REVISITING MULTILINGUAL MATHEMATICAL REASONING BENCHMARK 284

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

290 To address this, we introduce *language accuracy*, which measures whether the generated reasoning
 291 matches the intended question language. Formally, given the language identifier (e.g., Google
 292 Translate, langid (Lui & Baldwin, 2012)), f , language accuracy, A_{lang} is defined as follows:

$$294 A_{\text{lang}} = \frac{1}{N} \sum_n^N \delta[f(o_n) = l_n], \quad (7)$$

295 where N denotes the number of samples in the dataset, and $\delta[\cdot]$ is indicator function. o_n , and l_n
 296 mean the generated output and the target question language, respectively. Then, we defined the joint
 297 accuracy of mathematics and language, A_{joint} as follows:

$$300 A_{\text{joint}} = \frac{1}{N} \sum_n^N (\delta[f(o_n) = l_n] \cdot \delta[\hat{a}_n = a_n]), \quad (8)$$

302 where \hat{a}_n and a_n indicate predicted and ground-truth answers for n -th sample, respectively.

303 We evaluate various recent large language models, including Qwen2.5 (Hui et al., 2024),
 304 Llama3 (Grattafiori et al., 2024), Gemma3 (Team et al., 2025), and DeepSeek-R1 (Guo et al.,
 305 2025), on MGSM (see Appendix B.3 for the full list). Figure 5 illustrates average mathematical
 306 accuracy against joint accuracy across different languages. Ideally, both metrics should be the same
 307 (grey dashed line), yet models such as Qwen2.5-72B-Math-Instruct and Llama-3-70B-Instruct show
 308 large gaps, indicating frequent reasoning in the wrong language. Moreover, the s1 models (orange),
 309 fine-tuned from Qwen2.5-Instruct (green), notably degrade language accuracy while improving math-
 310 ematical performance. These results demonstrate that mathematical accuracy alone overestimates
 311 multilingual reasoning ability, and joint evaluation is essential for assessing true language-consistent
 312 reasoning.

313

314 6 EXPERIMENTS 315

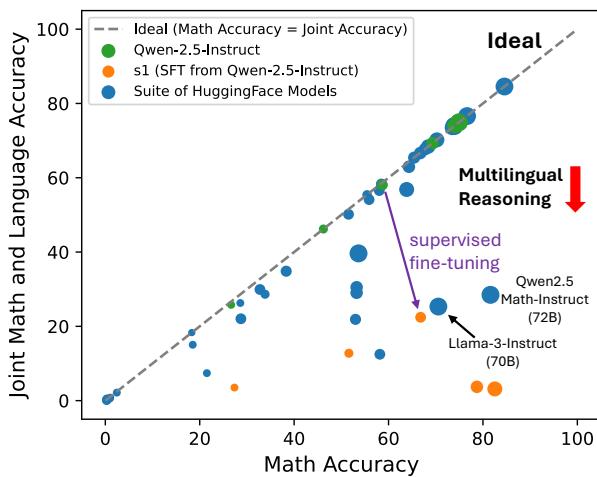
316 We use Qwen-2.5-7B-Instruct as the backbone for all experiments on mathematical and factual
 317 reasoning. Training and evaluation are conducted on 4 NVIDIA A100 GPUs with DeepSpeed (Rasley
 318 et al., 2020). We use three random seeds to calculate the mean and standard error. Please refer to
 319 Appendix B and the attached codebase for the implementation and training details.

320

321 6.1 DATASET

322 **Mathematical Reasoning.** The s1K-1.1 dataset (Muennighoff et al., 2025) contains 1,000 curated
 323 math questions with chain-of-thought traces, selected for difficulty, diversity, and quality. To test

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339



340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377

Figure 5: 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, fall significantly below this line, indicating they solve the problem correctly but fail to adhere to the language prompt.

Table 2: Accuracy of Qwen2.5-7B-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 in each column.

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

multilingual generalization, we additionally construct s1K-X, a multilingual version of s1K-1.1 obtained by translating into ten typologically diverse languages via Google Translate, used for baseline SFT results. For evaluation, we report results on GSM8K (Cobbe et al., 2021) and its multilingual counterpart MGSM (Shi et al., 2023) with Native CoT prompts. We also evaluate language accuracy introduced in Section 5.

Factual Reasoning. We utilize GEOFACT-X (Section 4), which contains culturally grounded factual QA pairs across five countries (USA, India, Japan, Kenya, Thailand) in five local languages (English, Hindi, Japanese, Swahili, Thai). Models are trained on the train split and evaluated on the test split.

6.2 MATHEMATICAL REASONING

Table 2 presents the performance of the base model, Qwen2.5-7B-Instruct, and models with post-training methods, supervised fine-tuning (SFT), GRPO, and M2A. Supervised fine-tuning on s1K-1.1 improves mathematical reasoning performance on GSM8K and MGSM but substantially degrades multilingual performance in MGSM, leading to lower joint accuracy. Training on the translated multilingual dataset (s1K-X) preserves language accuracy on MGSM but reduces mathematical accuracy. GRPO, in contrast, produces little change, likely due to sparse rewards. For instance, Figure 11 shows that GRPO outputs are identical to the base model, whereas SFT produces an English response to a Russian query.

M2A outperforms baselines in all metrics on GSM8K and achieves large gains in joint accuracy on MGSM compared to SFT. Unlike SFT, it preserves reasoning in the query language while

378 Table 3: Comparison of model performance on average reasoning score (%), language accuracy (%),
 379 and answer accuracy (%) on GEOFACT-X test set, evaluated across all examples and split by whether
 380 the language is associated with the country (‘Assoc.’) or not (‘Non-Assoc.’). Bold means the best
 381 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 (ours)	48.5 ± 0.4	52.6 ± 0.5	47.5 ± 0.3	32.0 ± 0.6	41.3 ± 1.0	29.7 ± 0.5
M2A (ours, Thai only)	49.8 ± 0.3	53.4 ± 0.2	48.8 ± 0.5	32.2 ± 0.4	39.9 ± 0.4	30.2 ± 0.3

393 Table 4: Machine-translated performance of each model on GEOFACT-X test set. Google Translate is
 394 used to translate the generated output into the question language. Bold means the best performance.
 395

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-Instruct	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 (ours)	48.7 ± 0.4	52.8 ± 0.5	47.7 ± 0.4	31.8 ± 1.0	41.3 ± 1.2	29.4 ± 0.9
M2A (ours, Thai only)	50.1 ± 0.2	53.1 ± 0.3	49.4 ± 0.3	30.6 ± 0.8	39.1 ± 0.3	28.5 ± 0.9

406
 407 still improving mathematical correctness. In effect, M2A learns mathematical reasoning without
 408 sacrificing multilingual fidelity, whereas other methods either fail to learn reasoning (GRPO) or lose
 409 multilingualism (SFT). Appendix E further examines a variant of M2A trained with translation into
 410 a single language instead of multiple languages, and detailed per-language results are provided in
 411 Appendix F.

413 6.3 FACTUAL REASONING

415 Table 3 summarizes the performance of the base model (Qwen-2.5-Instruct) and the gains obtained
 416 after post-training with GRPO, supervised fine-tuning (SFT), and M2A on GEOFACT-X. For com-
 417 parison, we also illustrate the performance of other pretrained LLMs (Cohere et al., 2025; Guo et al.,
 418 2025). We report reasoning score and answer accuracy. Results are additionally split by whether the
 419 language is associated with the country (*assoc.*) or not (*non-assoc.*); for instance, Thai is associated
 420 with Thailand but not with the USA. All pretrained models perform better in associative settings, likely
 421 because pretraining corpora contain more paired examples where language and country co-occur.
 422 This gap underscores the challenge of aligning reasoning across languages and contexts, motivating
 423 methods that explicitly enforce language consistency.

424 M2A achieves the strongest reasoning performance compared to both pretrained and post-trained
 425 baselines, and it is reinforced when only using Thai for translation. Notably, M2A improves both
 426 settings at a similar rate (4–6% in reasoning score and 20–21% in answer accuracy). A per-language
 427 and per-country breakdown is provided in Appendix G. Figure 12 further illustrates model outputs:
 428 although all systems reason in the question language (Swahili), only M2A predicts the correct answer.

429 Finally, we apply machine translation as a post-hoc strategy. Table 4 shows that translation via Google
 430 Translate offers no substantial improvements over the direct setting (Table 3), reflecting their weaker
 431 multilingual alignment. This suggests that post-hoc translation provides, at most, a superficial fix and
 fails to address the core challenge of multilingual reasoning.

432 Table 5: Contribution of individual reward functions to M2A. The evaluation is performed on GSM8K
 433 and MGSM. Bold means the best performance. Lang: Language Consistency, CA: Context Alignment,
 434 RA: Reasoning-Step Alignment.

M2A Variants			GSM8K			MGSM		
Lang	CA	RA	Math.	Lang.	Joint	Math.	Lang.	Joint
✓			86.9 ± 0.0	100.0 ± 0.0	86.9 ± 0.0	54.2 ± 0.1	98.3 ± 0.1	53.8 ± 0.1
✓	✓		84.7 ± 0.1	100.0 ± 0.0	84.7 ± 0.1	57.8 ± 0.1	99.5 ± 0.1	57.5 ± 0.1
✓	✓	✓	87.3 ± 0.1	100.0 ± 0.0	87.3 ± 0.1	59.0 ± 0.3	97.8 ± 0.2	58.1 ± 0.4

441 Table 6: Comparison of reward formulation for multilingual alignment rewards of M2A. The evalua-
 442 tion is performed on GSM8K and MGSM. Bold means the best performance.

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

448 6.4 ABLATION STUDY

449 **Contribution of Individual Reward Functions.** We analyze the effectiveness of individual reward
 450 functions in M2A on the mathematical reasoning task. Table 5 shows that context alignment (CA)
 451 improves multilingual performance on MGSM but slightly lowers GSM8K accuracy, as enforcing
 452 global embedding similarity adds constraints unnecessary for English-only tasks. Reasoning-step
 453 alignment (RA) provides finer supervision by aligning individual reasoning steps, which boosts
 454 multilingual performance and mitigates the small degradation from CA. The full model, combining
 455 language consistency, CA, and RA, achieves the best results on both benchmarks, confirming that
 456 the reward functions are complementary: CA promotes global cross-lingual alignment, while RA
 457 enforces stepwise reasoning fidelity.

458 **Reward Formulations.** We compare different formulations of the multilingual alignment reward
 459 used in Eq. (2) and Eq. (4). Table 6 reports results on GSM8K and MGSM. Using vanilla cosine
 460 similarity yields weaker performance, while adding a negative-sample term improves MGSM but
 461 slightly reduces GSM8K. Our margin-based hinge formulation achieves the best results across all
 462 metrics, demonstrating the benefit of combining negative samples with a margin to stabilize alignment.

465 7 DISCUSSION

466 We conducted a comprehensive study of whether large language models (LLMs) reason in the
 467 language of the input question. Our findings show that many LLMs predominantly reason in English
 468 or Chinese, even when prompted in other languages, undermining multilingual reasoning quality and
 469 limiting their applicability in culturally and linguistically diverse settings.

470 To overcome this limitation, we introduce a novel method, M2A, which enforces language-consistent
 471 reasoning while preserving factual correctness. By combining multi-scale alignment rewards with
 472 a language-consistency objective, M2A aligns outputs with ground-truth reasoning traces at both
 473 context and reasoning step levels, encouraging reasoning to remain in the query language.

474 Robust evaluation of multilingual reasoning is itself difficult, since most benchmarks focus on final
 475 answers rather than reasoning quality or language alignment. We therefore propose GEOFACT-X,
 476 a geography-based factual reasoning benchmark spanning five diverse languages, paired with step-
 477 by-step reasoning traces and a reasoning evaluation protocol including logical structure, factual
 478 correctness, and language consistency.

479 Our results show that M2A consistently improves multilingual mathematical and factual reasoning
 480 capability while maintaining strong English performance. While our experiments were conducted
 481 on 7B-parameter models, the approach is scalable and provides a practical alternative to massive
 482 multilingual instruction tuning. More broadly, our contributions establish a foundation for training
 483 and evaluating LLMs that reason faithfully across languages, advancing the goal of globally inclusive,
 484 culturally grounded, and interpretable AI.

486 ETHICS STATEMENT
487488 We read the ICLR Code of Ethics before the submission. Our paper focuses on multilingual reasoning
489 capabilities of large language models (LLMs), emphasizing knowledge transfer between languages
490 and highlighting limitations faced by low-resource languages. We believe this work encourages the
491 research community to address these limitations, ultimately contributing toward equitable access to
492 high-performing LLMs, regardless of the user’s language. However, our work also shares similar
493 negative societal concerns with standard large language model research (e.g., biased toward high-
494 resource languages and hallucination).495
496 REPRODUCIBILITY STATEMENT
497498 We specify the experimental setting in Appendix B and attached the codebase as supplementary
499 materials.500 REFERENCES
501503 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
504 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.
505 *arXiv preprint arXiv:2303.08774*, 2023.506 Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang.
507 Language models’ factuality depends on the language of inquiry. *arXiv preprint arXiv:2502.17955*,
508 2025.510 Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay
511 Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. Mega: Multilingual evaluation of
512 generative ai. *arXiv preprint arXiv:2303.12528*, 2023.513 Sanchit Ahuja, Kumar Tanmay, Hardik Hansrajbhai Chauhan, Barun Patra, Kriti Aggarwal, Luciano
514 Del Corro, Arindam Mitra, Tejas Indulal Dhamecha, Ahmed Awadallah, Monojit Choudhary, et al.
515 sphinx: Sample efficient multilingual instruction fine-tuning through n-shot guided prompting.
516 *arXiv preprint arXiv:2407.09879*, 2024.517 Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. Llms are few-shot in-context low-resource
518 language learners. *arXiv preprint arXiv:2403.16512*, 2024.520 Tyler A Chang, Catherine Arnett, Zhiwen Tu, and Benjamin K Bergen. When is multilingual-
521 quality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint
522 arXiv:2311.09205*, 2023.523 Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth
524 Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *arXiv
525 preprint arXiv:2309.08958*, 2023.527 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
528 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve
529 math word problems. *arXiv preprint arXiv:2110.14168*, 2021.530 Team Cohere, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Yazeed Alnumay, Sophia
531 Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara
532 Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia,
533 Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt
534 Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander
535 Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash
536 Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier,
537 Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru,
538 Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas,
539 Ben Cyrus, Daniel D’souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche
Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit

540 Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün,
 541 Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias
 542 Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Seraphina
 543 Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan,
 544 Manoj Govindassamy, Nathan Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian
 545 Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom
 546 Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline
 547 Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia
 548 Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov,
 549 Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia
 550 Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin
 551 Luong, Raymond Ma, Lukas Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez,
 552 Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic
 553 McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynihan, Maximilian
 554 Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham
 555 Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu,
 556 Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus,
 557 Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan,
 558 Pierre Richermond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Laura Ruis, Louise
 559 Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam,
 560 Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne,
 561 Sylvie Chang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivasan, Inna Shteingub,
 562 Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina,
 563 Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone,
 564 Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat
 565 Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White,
 566 Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern,
 567 Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language
 568 model, 2025.

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

572 John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. Rlhf
 573 can speak many languages: Unlocking multilingual preference optimization for llms. In *EMNLP*,
 574 2024.

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

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

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

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

585 Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang,
 586 Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*,
 587 2024.

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

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

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

597

598 Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice
 599 questions really be useful in detecting the abilities of llms? *arXiv preprint arXiv:2403.17752*,
 600 2024.

601

602 Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuhui Chen, Daniel Simig, Myle Ott,
 603 Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative
 604 language models. In *Proceedings of the 2022 conference on empirical methods in natural language*
 605 *processing*, pp. 9019–9052, 2022.

606

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

610

611 Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min
 612 Lin. Understanding r1-zero-like training: A critical perspective. In *COLM*, 2025.

613

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

616

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

620

621 Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven
 622 Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual
 623 generalization through multitask finetuning. In *ACL*, 2023.

624

625 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke
 626 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time
 627 scaling. *arXiv preprint arXiv:2501.19393*, 2025.

628

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

632

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

635

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

639

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

642

643 Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan
 644 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
 645 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
 646 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi
 647 Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,
 648 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

649

650 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
 651 Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*,
 652 2023.

653

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

648 Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*, pp. 649 3505–3506, 2020.

650

651 Keon-Woo Roh, Yeong-Joon Ju, and Seong-Whan Lee. Xlqa: A benchmark for locale-aware 652 multilingual open-domain question answering. *arXiv preprint arXiv:2508.16139*, 2025.

653

654 Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face 655 recognition and clustering. In *CVPR*, 2015.

656

657 Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual LLMs think in english? In *ICLR* 658 *Workshop on Building Trust in Language Models and Applications*, 2025.

659

660 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, 661 Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical 662 reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

663

664 Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, 665 Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language 666 models are multilingual chain-of-thought reasoners. In *ICLR*, 2023.

667

668 Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin 669 Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. Aya dataset: An 670 open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*, 2024.

671

672 Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, 673 Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly 674 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

675

676 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, 677 Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical 678 report. *arXiv preprint arXiv:2503.19786*, 2025.

679

680 Alexey Tikhonov and Max Ryabinin. It's all in the heads: Using attention heads as a baseline for 681 cross-lingual transfer in commonsense reasoning. *arXiv preprint arXiv:2106.12066*, 2021.

682

683 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming 684 Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging 685 multi-task language understanding benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2024.

686

687 Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao 688 Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian 689 Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. Mmlu-prox: A multilingual benchmark for 690 advanced large language model evaluation, 2025.

691

692 Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya 693 Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In 694 *NAACL*, 2021.

695

696 Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. Benchmarking machine translation 697 with cultural awareness. In *EMNLP Findings*, 2024.

698

699 Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, 700 Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 701 Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*, 2025.

702

703 Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating 704 the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 705 2023.

706

707 Chengzhi Zhong, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, 708 and Sadao Kurohashi. What language do non-english-centric large language models think in? In 709 *ACL Findings*, 2025.

702	Which treaty, signed in 1765, granted the British East India Company the Diwani rights of Bengal, Bihar, and Orissa?	703	1765 में हस्ताक्षरित किस संधि ने ब्रिटिश ईस्ट इंडिया कंपनी को बंगाल, बिहार और उड़ीसा के दीवानी अधिकार प्रदान किए?	704	1765 में हस्ताक्षरित किस संधि ने ब्रिटिश ईस्ट इंडिया कंपनी को बंगाल, बिहार और उड़ीसा के दीवानी अधिकार प्रदान किए?	705	1765 में हस्ताक्षरित किस संधि ने ब्रिटिश ईस्ट इंडिया कंपनी को बंगाल, बिहार और उड़ीसा के दीवानी अधिकार प्रदान किए?	706	Question
707	<step> The question asks about the treaty signed in 1765 that granted the British East India Company Diwani rights over Bengal, Bihar, and Orissa.\n<step> Diwani rights refer to the right to collect revenue and administer civil justice. \n<step> The Treaty of Allahabad, signed in 1765 between Robert Clive of the British East India Company and Mughal Emperor Shah Alam II, officially granted these rights to the Company. This treaty marked a significant turning point in the British East India Company's consolidation of power in India.	708	<step> प्रश्न में पूछा गया है कि 1765 में हस्ताक्षरित किस संधि ने ब्रिटिश ईस्ट इंडिया कंपनी को बंगाल, बिहार और उड़ीसा के दीवानी अधिकार प्रदान किए। \n<step> दीवानी अधिकार का अर्थ है राजस्व एकत्र करने और नागरिक न्याय प्रशासन का अधिकार। \n<step> इलाहाबाद की संधि, जिस पर 1765 में ब्रिटिश ईस्ट इंडिया कंपनी के रॉबर्ट क्लाइव और मुगल सम्राट शाह आलम द्वितीय के बीच हस्ताक्षर किए गए थे, ने आधिकारिक तौर पर इन अधिकारों को कंपनी को प्रदान किया। इस संधि ने भारत में ब्रिटिश ईस्ट इंडिया कंपनी की शक्ति के समेकन में एक महत्वपूर्ण माझ दिया।	709	<step> कामाक्षमताएँ सन्दर्भित किस संधि ने 1765 में हस्ताक्षरित किस संधि ने ब्रिटिश ईस्ट इंडिया कंपनी को बंगाल, बिहार और उड़ीसा के दीवानी अधिकार प्रदान किए। \n<step> दीवानी अधिकार का अर्थ है राजस्व एकत्र करने और नागरिक न्याय प्रशासन का अधिकार। \n<step> इलाहाबाद की संधि, जिस पर 1765 में ब्रिटिश ईस्ट इंडिया कंपनी के रॉबर्ट क्लाइव और मुगल सम्राट शाह आलम द्वितीय के बीच हस्ताक्षर किए गए थे, ने आधिकारिक तौर पर इन अधिकारों को कंपनी को प्रदान किया। इस संधि ने भारत में ब्रिटिश ईस्ट इंडिया कंपनी की शक्ति के समेकन में एक महत्वपूर्ण माझ दिया।	710	Reasoning Traces		
711	Treaty of Allahabad / Allahabad Treaty	712	इलाहाबाद की संधि / इलाहाबाद संधि	713	सन्दर्भित किस संधि ने 1765 में हस्ताक्षरित किस संधि ने ब्रिटिश ईस्ट इंडिया कंपनी को बंगाल, बिहार और उड़ीसा के दीवानी अधिकार प्रदान किए।	714	Answer (with all possible variants)		
715		716		717		718			

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.

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 (Fig. 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.

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 & Baldwin, 2012), its score is set to zero. Figure 10 illustrates the prompt structure used for the LLM-as-a-Judge, including the evaluation instructions and rules applied to model outputs.

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

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

783 Return the data in the following JSON format:
 784

```
{
  785   "question_id": "unique_increasing_number",
  786   "languages": [
  787     {
  788       "language_code": "en",
  789       "language_name": "English",
  790       "question": "The exact question text in English",
  791       "answers": ["Primary answer", "Alternative form 1", "Alternative form 2"]
  792     },
  793     {
  794       "language_code": "hi",
  795       "language_name": "Hindi",
  796       "question": "The exact question text in Hindi",
  797       "answers": ["Primary answer in Hindi", "Alternative form 1 in Hindi", "Alternative form 2 in Hindi"]
  798     },
  799     // Repeat for Japanese (ja), Swahili (sw), and Thai (th)
  800   ],
  801   "topic": "{topic}",
  802   "answer_type": "{answer_type}",
  803   "region": "{country}"
  804 }
```


 805 IMPORTANT: Return ONLY valid JSON without any explanations, formatting, or additional text outside
 806 the JSON structure. Ensure all apostrophes, quotation marks, and special characters are properly
 807 escaped in the JSON.

808
 809 Figure 7: **Prompt for Generating Multilingual Factual Questions and Answers in GEOFACT-X.**
 810 This prompt instructs the LLM to generate diverse, unambiguous factual questions about a specific
 811 country and topic, each with a single, definite answer. The questions and their answers are provided in
 812 five languages, English, Hindi, Japanese, Swahili, and Thai, with strict requirements for high-quality
 813 translations, consistent terminology, and inclusion of all valid answer variants in each language.

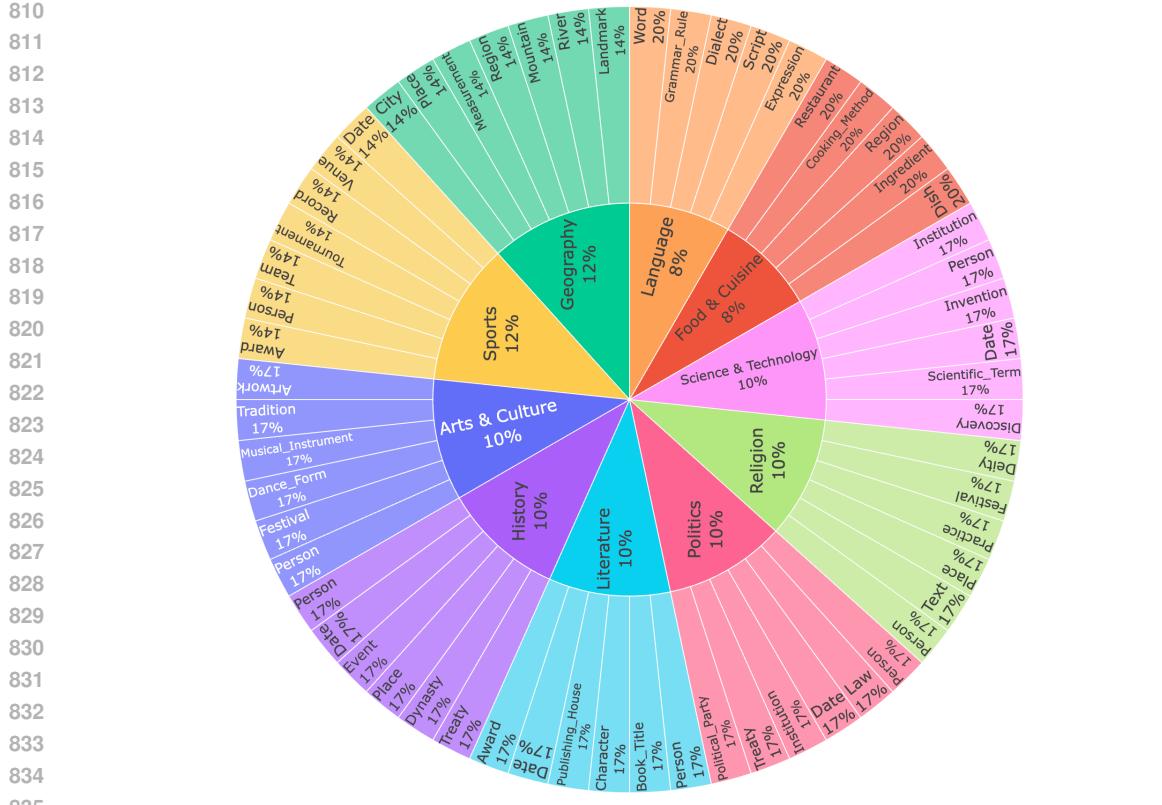


Figure 8: **Illustration of the hierarchical distribution of generated factual question categories by topic and subcategory.** Each colored wedge represents a major topic (e.g., History, Geography), and its outer segments represent specific subcategories (e.g., Person, Place, Treaty). The size of each 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.

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 EXPERIMENTAL DETAILS

B.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 reasoning parameters in case of difference). We also use a loss coefficient for GRPO as 0.01 for mathematical reasoning and 0.5 for factual reasoning. 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,

```

864 You are a multilingual reasoning assistant. For each of the following factual questions about the country,
865 generate structured output in the following format for all the 5 questions:
866 {
867   "question": "",
868   "answer_list": [""],
869   "reasoning": "<step> Step-by-step reasoning in the same language as the question, each step starting
870   with <step>",
871   "corrected_answer_list": [""],
872   "topic": "",
873   "region": "",
874   "answer_type": ""
875 }
876 Instructions:
877 1. The reasoning must be written strictly in the same language as the question.
878 2. Each reasoning step must begin with <step>
879 (e.g., <step> माउंट केन्या एक ज़्यालामुखी पर्वत है ...).
880 3. Start from relevant background knowledge or interpretation of the question and proceed step-by-step
881 toward the correct answer.
882 4. For the corrected_answer_list:
883   - Review the provided answer_list.
884   - Remove duplicates (case-insensitive, spacing-normalized).
885   - Add valid alternative phrasings or translations if any are missing (e.g., transliterations, local variants).
886 5. The reasoning must be comprehensive and detailed, including:
887   - Relevant background information and definitions of key terms or entities.
888   - Historical, cultural, geographical, or scientific context if applicable.
889   - Logical deductions and connections to prior knowledge.
890   - Contrast with similar or confusing facts (e.g., common misconceptions).
891   - Justifications for why incorrect options are incorrect (if multiple answers are possible).
892   - Step-by-step elimination or validation of answer candidates.
893 6. Each <step> should be at least 1–2 full sentences and contribute meaningfully to building up the
894 answer. Do not skip intermediate steps, even if obvious. Think like a teacher explaining to a curious
895 student
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

```

Figure 9: **Structured prompt for multilingual factual reasoning 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.

German, Spanish, French, Japanese, Russian, Swahili, Telugu, Thai, and Chinese. Both codebase and datasets are attached to the supplementary material for reproducibility.

B.2 EVALUATION

For mathematical reasoning, we employed the lm-evaluation-harness¹ 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 & Baldwin, 2012) is used to evaluate language correctness.

B.3 MODELS EVALUATED ON MGSM

Table 7 provides the complete list of models evaluated in Figure 5. All models are sourced from HuggingFace², a public repository of large language models. We include Qwen2.5 (Qwen et al., 2025), s1 (Muenninghoff et al., 2025), Llama (Grattafiori et al., 2024), Gemma (Team et al., 2025), and DeepSeek-R1 (Guo et al., 2025), each with a range of model sizes.

¹<https://github.com/EleutherAI/lm-evaluation-harness>

²<https://huggingface.co/>

```

918 # Reasoning Quality Evaluation
919 You are an expert reasoning evaluator tasked with comparing an LLM's reasoning trace against a ground
920 truth reasoning trace. Your evaluation must be fair, consistent, and based solely on the quality of
921 reasoning, not on superficial similarities.
922
923 ## Input:
924 - Question: {question}
925 - Answer List: {answer_list}
926 - Ground Truth Reasoning: {ground_truth_reasoning}
927 - LLM Response: {llm_generation}
928
929 ## Evaluation Criteria:
930 Assess the quality of the LLM's reasoning compared to the ground truth on a scale from 0-10 based on
931 the following:
932
933 1. Logical Structure (40%):
934 - How well does the reasoning follow a clear, step-by-step logical progression?
935 - Are the steps in a sensible order that builds toward the answer?
936
937 2. Key Insights (30%):
938 - Does the reasoning identify the same critical insights as the ground truth?
939 - Are the important clues from the question properly recognized and utilized?
940
941 3. Factual Correctness (20%):
942 - Is the reasoning free from factual errors?
943 - Does it avoid adding irrelevant information or missing necessary information?
944
945 4. Conclusion Validity (10%):
946 - Does the reasoning correctly lead to the answer?
947 - Is the link between the reasoning and the conclusion clear?
948
949 ## Scoring Guide:
950 0-1: Completely irrelevant or fundamentally flawed reasoning
951 2-3: Major logical errors or missing critical insights
952 4-5: Contains some correct elements but misses important aspects
953 6-7: Good reasoning with minor gaps or imperfections
954 8-9: Very good reasoning, almost matching ground truth quality
955 10: Perfect reasoning, capturing all key insights with proper structure
956
957 ## Your Response (FORMAT STRICTLY REQUIRED):
958 REASONING_SCORE: [integer between 0-10]
959 JUSTIFICATION: [Brief explanation of your evaluation, highlighting strengths and weaknesses]
960
961

```

Figure 10: **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.

C ABLATION STUDY OF TRANSLATION METHOD

To ensure the accessibility and scalability of our pipeline, we prioritize computational efficiency and broad language coverage. We select Google Translate as our primary translation tool because it supports over 100 languages—including many low-resource ones—and remains computationally efficient. This design choice allows researchers to reproduce our method without requiring access to expensive LLM APIs or high-end GPUs.

To validate this choice, we conduct an ablation study comparing Google Translate against LLM-based translation methods (specifically GPT-4.1 and GPT-5). As shown in Table 8, we observe no statistically significant difference in downstream model performance between the translation methods.

972 Table 7: List of all models and sizes evaluated on MGSM in Figure 5. All models are sourced from
 973 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

993 Table 8: Comparison of model performance with different translation methods on average reasoning
 994 score (%), language accuracy (%), and answer accuracy (%) on GEOFACT-X test set, evaluated
 995 across all examples and split by whether the language is associated with the country (‘Assoc.’) or not
 996 (‘Non-Assoc.’). Bold means the best performance.

Translation Method	Average Reasoning Score (%)			Average Answer Accuracy (%)		
	All	Assoc.	Non-Assoc.	All	Assoc.	Non-Assoc.
Google Translate	48.5 ± 0.4	52.6 ± 0.6	47.5 ± 0.3	32.0 ± 0.6	41.3 ± 1.0	29.7 ± 0.5
ChatGPT 4.1	48.8 ± 0.1	52.2 ± 0.3	48.0 ± 0.2	33.1 ± 0.3	41.3 ± 1.3	31.0 ± 0.4
ChatGPT 5	49.0 ± 0.1	52.6 ± 0.5	48.1 ± 0.0	32.7 ± 0.3	41.4 ± 1.0	30.5 ± 0.2

1003
 1004 These results confirm that our framework is robust to the choice of translator and that the performance
 1005 gains stem principally from the M2A objective rather than translation artifacts.
 1006

1008 D M2A WITH DIFFERENT BACKBONE ON GEOFACT-X

1009
 1010 We use Command R7B (Cohere et al., 2025) as a backbone network for comparing M2A other
 1011 baselines. As shown in Table 9, the results are consistent with our main Qwen findings. M2A achieves
 1012 the highest performance in both reasoning score and answer accuracy, outperforming SFT and
 1013 standard GRPO. This confirms that our method is model-agnostic and effective across different
 1014 multilingual architectures.

1016 E M2A WITH DIFFERENT LANGUAGES

1017
 1018 In the main paper, we present M2A trained with random translations drawn from the ten MGSM
 1019 languages (Bengali, German, Spanish, French, Japanese, Russian, Swahili, Telugu, Thai, and Chinese).
 1020 Here, we examine the effect of using a single fixed translation language, as shown in Table 10.
 1021 We select Japanese and Swahili as representative examples of high- and low-resource languages,
 1022 respectively, following the categorization of Nicholas & Bhatia (2023).

1023 Across both choices, we observe only minor decreases in GSM8K mathematical accuracy and
 1024 MGSM language accuracy relative to the multi-language setting. This finding indicates that training
 1025 with a single language—even a low-resource one—can still induce strong multilingual reasoning
 1026 ability. Nevertheless, randomizing translations across multiple languages yields the strongest overall

1026 Table 9: Comparison of model performance on average reasoning score (%), language accuracy (%),
 1027 and answer accuracy (%) on GEOFACT-X test set, evaluated across all examples and split by whether
 1028 the language is associated with the country (‘Assoc.’) or not (‘Non-Assoc.’). Bold means the best
 1029 performance.

Model	Average Reasoning Score (%)			Average Answer Accuracy (%)		
	All	Assoc.	Non-Assoc.	All	Assoc.	Non-Assoc.
Command R7B	44.2	48.5	43.1	25.0	34.1	22.7
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

1038 Table 10: Accuracy of M2A trained with various languages. All languages denote that the language
 1039 translator randomly translates the question language into ten different languages used in MGSM.
 1040 Bold denotes the best performance for each metric.

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

1042 Table 11: Accuracy of base (Qwen2.5-7B-Instruct) model and models fine-tuned with each post-
 1043 training method on MGSM. Standard error is not included for readability. Bold means the best
 1044 performance.

Model	MGSM	Question Language									
		bn	de	es	fr	ja	ru	sw	te	th	zh
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
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	100.0	99.7	99.9	100.0	100.0	100.0	98.0
M2A	97.8	99.7	98.0	98.5	94.8	99.9	99.2	88.3	99.9	99.3	100.0
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 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

1073 results, suggesting that language diversity provides additional regularization benefits for cross-lingual
 1074 alignment.

F MGSM EVALUATION IN EACH LANGUAGE

1075 Table 11 shows individual math and language accuracy change compared to the base model (Qwen2.5-
 1076 7B-Instruct) in each language. As we mentioned in Section 6, supervised fine-tuning on s1K-1.1

Table 12: 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 13: 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

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 11. 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.

1134 **G GEOFACT-X EVALUATION IN EACH LANGUAGE AND COUNTRY PAIR**
11351136 Tables 12, and 13 present reasoning score, language accuracy, and answer accuracy for each lan-
1137 guage–country pair. Diagonal entries correspond to associative pairs, while off-diagonal entries
1138 correspond to non-associative pairs. Associative countries consistently achieve higher performance
1139 across all metrics compared to non-associative ones. English (a high-resource language) generally
1140 outperforms other languages. Among non-associative cases, performance for the United States ex-
1141 ceeds that of Kenya and Thailand, which may reflect differences in pre-training and fine-tuning data
1142 exposure. As discussed in Section 6.3, all methods substantially increase language accuracy as the
1143 training dataset includes all five languages. However, no consistent trend is observed in reasoning
1144 score and answer accuracy across language–country pairs.
11451146 **H HUMAN AGREEMENT FOR LLM-AS-A-JUDGE**
11471148 We validate the alignment between reasoning score and human judgment through a Human-LLM
1149 agreement study on the English-USA subset of GeoFact-X. A human evaluator (native English
1150 speaker) performs pairwise A/B testing, selecting the superior reasoning trace from two model
1151 outputs and assigning a confidence score on a 1–5 scale. The results demonstrate strong alignment
1152 between the Qwen-based judge and human preference, achieving an agreement rate of 82.9%.1153 We further observe a strong positive correlation between human reported confidence and the reasoning
1154 score difference assigned by the LLM. Notably, human confidence is significantly higher when
1155 agreeing with the LLM (average confidence 3.3/5) compared to cases of disagreement (average
1156 confidence 2.1/5). This indicates that divergences primarily occur in ambiguous "close calls," whereas
1157 the metric remains highly reliable for clear-cut distinctions. Future work will expand this validation
1158 across all five languages with native speakers to confirm these findings multilingually.
11591160 **I LLM USAGE**
11611162 We used LLM to help with paper writing for improving grammar and wording.
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

1188		
1189		
1190		
1191		
1192		
1193	Question	
1194	< im_start >system	
1195	You are Qwen, created by Alibaba Cloud. You are a helpful	
1196	assistant.< im_end >	
1197	< im_start >user\nЗадача: Утки Дженет несут по 16 яиц в день. Она	
1198	ест три на завтрак каждое утро и печет кексы для своих друзей	
1199	каждый день, используя четыре. Остаток онааждый день продает	
1200	на фермерском рынке по 2 \$ за свежее утиное яйцо. Сколько в	
1201	долларах она зарабатывает на фермерском рынке каждый	
1202	день? \nПошагово: < im_end >	
1203	< im_start >assistant	
1204		
1205	Qwen2.5-7B-Instruct	
1206	Давайте решим эту задачу пошагово:	
1207	1. Сначала определим, сколько яиц остается после того, как утки	
1208	Дженет съедают и используют их:	
1209	- Всего яиц в день: 16	
1210	- Яиц на завтрак: 3	
1211	- Яиц для кексов: 4	
1212	- Остаток: 16 - 3 - 4 = 9 яиц	
1213	2. Теперь определим, сколько Джленет продает на фермерском	
1214	рынке:	
1215	- Она продает все оставшиеся яйца: 9 яиц	
1216	3. Посчитаем, сколько денег она зарабатывает:	
1217	- Цена за яйцо: 2 \$	
1218	- За 9 яиц: 9 * 2 = 18 \$	
1219	Таким образом, Джленет зарабатывает 18 долларов на фермерском	
1220	рынке каждый день.	
1221	Supervised Fine-Tuning (s1)	
1222	Okay, let's try to figure out how much money Janet's ducks make each	
1223	day from selling eggs at the farmer's market. Let me break down the	
1224	problem step by step.	
1225	First, the problem states that Janet's ducks lay 16 eggs each day. That's	
1226	the total production. Then, Janet eats 3 eggs every morning for	
1227	breakfast. So, those 3 eggs are consumed by her. Additionally, she	
1228	uses 4 eggs each day to bake cakes for her friends. So, both her	
1229	consumption and baking take a total of $3 + 4 = 7$ eggs per day.	
1230	...	
1231	But just to make sure I didn't misread the question. Let me re-read it:	
1232	"Утки *** день?"	
1233	Which translates to: "Janet's ducks lay 16 eggs per day. She eats three	
1234	for breakfast every morning and bakes muffins for her friends using four	
1235	eggs every day. The remainder she sells at the farmer's market for \$2	
1236	per fresh duck egg each day. How much does she earn at the farmer's	
1237	market each day in dollars?"	
1238	Yes, the question is asking about her daily earnings from the farmer's	
1239	market. So, the calculation is correct.	
1240	Therefore, the final answer is \boxed{18} .	
1241	< im_start >answer	
1242	Answer: Janet's ducks lay 16 eggs each day. She eats 3 eggs every	
1243	morning and uses 4 eggs to bake cakes for her friends. The total	
1244	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 .	
1245	\boxed{18}	

Figure 11: **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).

1242
1243
1244
1245
1246
1247
1248
1249

1250	Model	Q: Safari ya Lewis na Clark ilianza wapi? (Where did Lewis and Clark's journey begin?) A: St. Louis
1251		Swali linauliza kuhusu mahali ambapo safari ya Lewis na Clark ilianza. >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. >Safari hii ilianza Fort Atkinson, ambayo ilikuwa fortuoni ya watalii huko Council Bluffs, Iowa, na Omaha, Nebraska. Kwa hivyo, majibu yote mawili yanakubalika. <answer>Fort Atkinson</answer>
1252	Base	Translated Output: The question asks where the Lewis and Clark expedition began. >The Lewis and Clark expedition was a voyage across American waters. It began in 1804 and marked the first American expedition eastward. >The expedition began at Fort Atkinson, which was a tourist fort in Council Bluffs, Iowa, and Omaha, Nebraska. Therefore, both answers are acceptable. <answer>Fort Atkinson</answer>
1253		
1254	GRPO	Safari ya Lewis na Clark ilianza wapi? Hii inatafsiriwa kama safari ya kawaida kwa ajili ya kujenga kazi kwa kawaida na kujenga mazingira na kawaida. Iliopiga safari hili, Lewis na Clark inatafsiriwa kama kawaida kwa ajili ya kujenga mazingira na kawaida na kujenga kazi kwa kawaida. Safari hili ina kujenga mazingira na kawaida kwa ajili ya kujenga kazi kwa kawaida. <answer> Safari ya Lewis na Clark ilianza kwa ajili ya kujenga mazingira na kawaida na kujenga kazi kwa kawaida.</answer>
1255		
1256		
1257		
1258		
1259	SFT	Translated Output: The Lewis and Clark expedition began in 1804 and 1806. To help with births and births and births. To learn if there are countries with births and births and births and long periods, Lewis and Clark began in 1804 and 1806. To learn if there are countries with births and births and long periods, Lewis and Clark began in 1804 and 1806. <answer> The Lewis and Clark expedition began in 1804 and 1806. </answer>
1260		
1261		
1262		
1263		
1264		
1265		
1266		
1267	M2A	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. <step> Meriwether Lewis alikuwa mwanamke mtawalimba aliongoza safari ya Lewis na Clark. Alikuwa pia vice presidenta wa Theodore Roosevelt. <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. <step> Jiji la Washington ni jina lingine la Washington, D.C., hivyo majibu yote mawili ni sahihi. <answer>Washington, D.C.</answer>
1268		
1269		
1270		
1271		
1272		
1273		
1274		
1275		
1276		
1277		
1278		
1279		
1280		
1281		
1282		
1283		
1284		
1285		
1286		
1287		
1288		
1289		
1290		
1291		
1292		
1293		
1294		
1295		

Figure 12: 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.