x1: Learning to Think Adaptively Across Languages and Cultures

Anonymous ACL submission

Abstract

Languages encode distinct abstractions and inductive priors, yet most large language models (LLMs) overlook this diversity by reasoning in a single dominant language. In this work, we introduce *x1*, a family of reasoning models that can *adaptively* reason in an advantageous language on a per-instance basis. To isolate the effect of reasoning-language choice, *x1* is constructed without expanding the model's knowledge boundaries and is trained by contrasting linguistically distinct reasoning trajectories for the same input. Our extensive experiments demonstrate the benefits of adaptive multilingual reasoning in multilingual mathematical and culturally grounded tasks. Moreover, our results challenge a simplistic view of scaling law: while scaling reduces cross-lingual disparities in procedural domains such as math reasoning, it does not eliminate the advantages of reasoning in culture-associated languages in cultural scenarios, as we empirically show that such reasoning enables more efficient and accurate cultural knowledge recall. Overall, our findings establish language choice as a functional component of reasoning, with implications for building more generalist and globally competent reasoning models.

1 Introduction

"The limits of my language mean
the limits of my world."

— Ludwig Wittgenstein, 1922 (Wittgenstein, 2023)

"Reasoning" endows large language models (LLMs) with the ability to go beyond surface-level pattern matching and to tackle complex tasks such as competition-level mathematics, logical reasoning, and multi-hop question answering (Wei et al., 2022; Snell et al., 2024; Brown et al., 2024). However, most existing LLMs, like OpenAI-o1/o3 (Jaech et al., 2024) and Qwen3 (Yang et al., 2025), reason predominantly in high-resource languages (such as English, Chinese). While reasoning in a dominant language is often effective, it implicitly assumes a single linguistic perspective to be universally optimal. However, languages encode distinct abstractions, cultural priors, and modes of expression, which can subtly shape how problems are decomposed, interpreted, and solved (Goddard, 2003; Kovecses, 2006). Consequently, confining reasoning to a single language may underutilize this diversity, particularly in multilingual and culturally grounded scenarios.

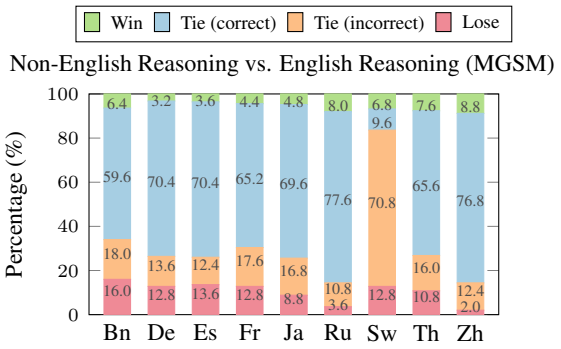


Figure 1: Win/tie/lose rates of non-English vs. English reasoning pathways for *Qwen3-4B* on *MGSM*. Reasoning in contrast languages is enabled via Step 1 (introduced in §1). "Win": cases where non-English reasoning outperforms English reasoning; "Tie" (correct/incorrect): both reasoning pathways yield correct/incorrect answers; "Lose": cases where English reasoning outperforms non-English reasoning.

Recently, Yong et al. (2025) conducted a detailed analysis of cross-lingual reasoning and demonstrate that, in multilingual mathematics problems, English is not always the optimal language for reasoning. In Figure 1, we quantify the advantages of non-English reasoning by measuring the win/tie/lose rates of non-English reasoning compared to English reasoning, revealing that in *MGSM* (Shi et al., 2022), roughly 3%–9% (varying across languages) of instances can be solved by non-English reasoning paths where English reasoning fails. Beyond mathematical reasoning, Yong et al. (2025)

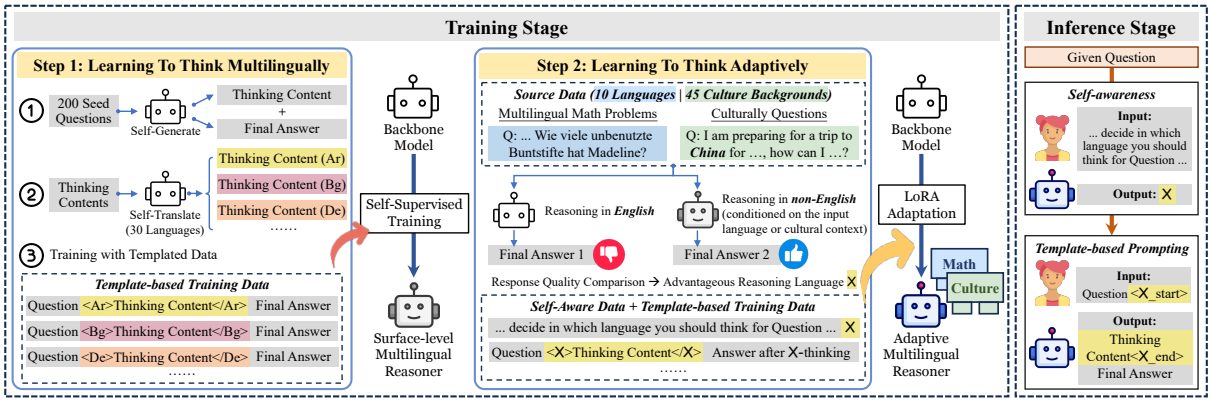


Figure 2: Overview of the construction process and inference usage of *x1* models.

observed that test-time scaling of English reasoning (i.e., extending the reasoning length) does not necessarily improve performance in culture-aware scenarios, where success may require more knowledge recall or culturally grounded reasoning. This limitation further motivates our investigation into non-English reasoning pathways, particularly for culturally situated reasoning, where the choice of thinking language may play a decisive role.

In this paper, we take a step toward exploiting the diversity of reasoning perspectives offered by English and non-English languages. Building on existing open-source reasoning models, we introduce ***x1***, a family of models that can *adaptively select* the more advantageous thinking language on a per-instance basis. Crucially, this is achieved *without expanding the model's knowledge boundaries*, allowing us to isolate the effect of reasoning-language choice in a controlled and comparable setting, rather than emphasizing absolute performance gains. Specifically, we first equip the model with the ability to reason in a *specified language* across multiple languages through self-supervised training, and then induce adaptive reasoning behavior by contrasting linguistically distinct reasoning trajectories for the same input.

We investigate adaptive multilingual reasoning in two deeply language-intertwined scenarios: multilingual mathematical reasoning and cultural reasoning. Our findings suggest that reasoning in LLMs is not language-agnostic, but shaped by the reasoning language choice. By enabling adaptive multilingual reasoning, we show that linguistic diversity offers systematic benefits beyond reliance on a single language. Importantly, our results challenge a simplistic view of scaling laws: while scaling reduces cross-lingual disparities in procedural domains such as mathematical reasoning, it fails to eliminate the intrinsic advantages of reasoning in culture-associated languages for culturally grounded tasks, which leads to more efficient and accurate cultural knowledge recall. Together, these findings indicate that language choice is not merely a surface realization of thought, but a functional component of reasoning itself, with important implications for building truly generalist and globally competent reasoning models.

2 Constructing *x1*

To achieve reasoning in a advantageous language without expanding the model's knowledge boundaries, we propose a two-stage training strategy that leverages self-generated contrastive signals to induce adaptive multilingual reasoning.

2.1 Step 1: To be a Multilingual Reasoner

The first step aims to open up the model's multilingual reasoning space, enabling it to deliberately reason in a specified language rather than being implicitly anchored to its default thinking language.

To achieve this, we adopt a lightweight self-supervised training strategy that teaches the model to perform reasoning in a required language while ensuring that the final answer remains consistent with the prompting language[1]. In practice, we first query the backbone model $\mathcal{M}$ with 200 seed questions sampled from *Flan_v2* (Wei et al., 2021), collecting its generated reasoning traces $\{\mathrm{t}_i\}_{i=1}^{200}$ and final answers $\{\mathrm{a}_i\}_{i=1}^{200}$. We then instruct $\mathcal{M}$ itself to translate each reasoning trace into 30 different languages $\{l_j\}_{j=1}^{30}$ (listed in Appendix A.1), yielding a multilingual set of reasoning traces $\{\mathrm{t}_i^{l_j}\}_{i=1}^{200}$ and $j = 1, \ldots, 30$. To ensure linguistic fidelity, we

[1]For example, if the prompt is in English but the reasoning language is constrained to German, the final response should still be generated in English.

apply quality filtering using COMET[2] scores and discard translations of evidently poor quality.

Using these translated reasoning traces, we fine-tune the backbone model $\mathcal{M}$ with a templated format that explicitly specifies a reasoning language as follows: (see complete examples in Figure 6)

// Example of reasoning in **Arabic**
Input: {Question$_i$}
Output: <think>\n<Arabic_start>\n\n{t_i^{Ar}}\n\n
<Arabic_end>\n</think>\n\n{a_i}

The resulting "sibling" model, $\mathcal{M}_{\text{surface}}$, can be viewed as a *surface-level multilingual reasoner*: it is capable of following simple templates to reason in a specified language X by appending a simple template "*<think>\n<X_start>*" at inference time, while avoiding undesirable language drift. Notably, this capability operates purely at the level of reasoning expression: it exposes alternative language-conditioned reasoning paths and does not introduce any external knowledge or capability expansion (such as Google Translation, external knowledge bases, or stronger auxiliary models).

2.2 Step 2: To be an Adaptive Reasoner

The goal of Step 2 is to elevate multilingual reasoning from language following to adaptively choose the more advantageous reasoning language.

Leveraging $\mathcal{M}_{\text{surface}}$, we are able to generate paired reasoning trajectories in English and non-English languages for each input[3]. These paired trajectories can be viewed as *counterfactual reasoning paths* induced by different language choices, providing a basis for comparative learning.

Training Data Source. We study adaptive reasoning in two language-intertwined scenarios. **(1) Multilingual Math Problems** (D_{math}): We use the *MGSM8KInstruct* dataset (Chen et al., 2023), which contains math problems paired with their correct answers. We sample 200 samples across 10 languages, resulting in 2,000 training instances. **(2) Culture-related Problems** (D_{culture}): We use the *CultureBank* dataset (Shi et al., 2024), which provides cultural questions along with their underlying cultural knowledge. We sample 100-200 samples from each of the 25 language groups (depending on availability), covering cultural questions from 45 countries/regions, resulting in 4,413 samples in total. (Detailed statistics are in Appendix A.2)

[2]Unbabel/wmt23-cometkiwi-da-xxl (Rei et al., 2023)

[3]If $\mathcal{M}$ naturally reasons in English, we require $\mathcal{M}_{\text{surface}}$ to reason in a non-English language; conversely, we require $\mathcal{M}_{\text{surface}}$ to reason in English.

Advantageous Language Identification. For each question in D_{math} or D_{culture}, we first obtain the paired reasoning trajectories in both English and non-English with $\mathcal{M}$ and $\mathcal{M}_{\text{surface}}$[3]. We identify the advantageous reasoning language based on the quality of the final answers, under the principle that stronger reasoning should manifest in superior outputs. For multilingual math problems, answer quality is determined via exact numerical matching. For culture-related questions, we employ *LLM-as-a-Judge* (Zheng et al., 2023) to score how well the response entails the relevant cultural knowledge. The reasoning language associated with the higher-scoring trajectory is then regarded as the advantageous thinking language for that instance. (Detailed implementations for comparison are provided in Appendix A.2.)

Training Data Construction. For each instance, we retain only the reasoning trace and answer corresponding to the advantageous language, and discard cases where English and non-English reasoning receive tied scores, to enforces a sharper contrast between alternative reasoning pathways and provides a clear signal for language preference.

We train the backbone model $\mathcal{M}$ using a templated format (similar to Step 1) that explicitly specifies the selected reasoning language X, implemented via parameter-efficient LoRA finetuning. For a given question Q, if language X is identified as the advantageous reasoning language, with corresponding reasoning trace T and final answer A, we construct the training instance as follows:

Input: {Q}
Output: <think>\n<X_start>\n\n{T}\n\n<X_end>\n
</think>\n\n{A}

Additionally, we introduce auxiliary self-awareness data that externalizes the model's language-selection decision as an explicit prediction task. This encourages the model to internalize reasoning-language choice as a deliberate component of its reasoning strategy rather than as a fixed or implicit heuristic. The data format is illustrated below (see complete example in Figure 7):

Input: ...decide in which language you should internally
think...for question {Q}\n\nThinking Language:
Output: <think>\n\n</think>\n\n{X}

All training data in Step 2 are self-generated by either $\mathcal{M}$ or $\mathcal{M}_{\text{surface}}$. This stage relies on instance-level comparisons between alternative reasoning-language choices, enabling the backbone models to select and exploit its latent multilingual reasoning capabilities in an adaptive manner.

Models	Multilingual Math Reasoning				Cultural Reasoning			
	MGSM		MT-AIME		FORK		CulturalBench	
	Non-Think	***Think***	*Non-Think*	***Think***	*Non-Think*	***Think***	*Non-Think*	***Think***
Top-tiered Reasoning Models								
o4-mini-high	–	82.32	–	75.33	–	74.46	–	82.31
DeepSeek-V3.2	79.24	76.32	53.00	85.67	72.83	78.80	84.60	89.16
Open-sourced Reasoning Models								
Qwen3-4B	70.21	76.59	12.89	21.78	74.64	73.73	68.76	70.85
Qwen3-14B	77.64	82.56	19.33	29.22	74.46	73.91	75.50	78.24
Qwen3-32B	80.52	83.98	21.83	33.89	78.26	81.88	78.76	81.26
DeepSeek-R1-Distill-Qwen-7B	54.76	60.05	8.33	25.83	40.04	55.25	29.58	38.63
DeepSeek-R1-Distill-Llama-8B	38.17	40.36	2.67	14.44	37.86	74.09	35.02	57.19
x1 Series Models								
	+ Math				***+ Culture***			
x1-Qwen3-4B	70.30	77.69 (↑ 1.10)	13.56	22.83 (↑ 1.05)	75.18	78.08 (↑ 4.35)	68.46	72.74 (↑ 1.89)
x1-Qwen3-14B	77.38	83.64 (↑ 1.08)	19.44	33.11 (↑ 3.89)	73.12	76.81 (↑ 2.90)	76.07	81.58 (↑ 3.34)
x1-Qwen3-32B	80.12	84.43 (↑ 0.45)	22.11	34.50 (↑ 0.61)	79.17	82.61 (↑ 0.73)	80.47	81.99 (↑ 0.73)
x1-DeepSeek-R1-Distill-Qwen-7B	54.52	63.24 (↑ 3.19)	9.00	27.00 (↑ 1.17)	54.89	58.70 (↑ 3.45)	44.82	49.04 (↑ 10.4)
x1-DeepSeek-R1-Distill-Llama-8B	38.01	52.17 (↑ 11.8)	2.89	17.00 (↑ 2.56)	57.07	76.27 (↑ 2.18)	51.13	60.88 (↑ 3.69)

Table 1: Mean@3 results that are the averages of the performance across all languages / cultures for each dataset.

3 Experiments

3.1 Setup

Models. We perform training on five representative reasoning models. Qwen3-series models (Yang et al., 2025): (1) *Qwen3-4B*, (2) *Qwen3-14B*, (3) *Qwen3-32B* and DeepSeek-R1-distilled models (Guo et al., 2025): (4) *DeepSeek-R1-Distill-Qwen-7B* and (5) *DeepSeek-R1-Distill-Llama-8B*.

Training. Our training processes are conducted on *8×A800-SXM4-80GB* GPUs. In step-1 self-supervised training, we perform full-parameter fine-tuning (Ouyang et al., 2022) for all models to ensure effective acquisition of the designed templates, except for *Qwen3-32B*, for which we adopt LoRA (Hu et al., 2022) training due to its large model size. In step-2, we apply LoRA training separately for multilingual math and cultural reasoning scenarios, aiming to minimally modify the backbone models while enabling lightweight and flexible scenario adaptation. More training details and hyperparameters are provided in Appendix A.

Evaluations. We conduct experiments on four benchmarks, which can be categorized into:

- **Multilingual Math Reasoning:** (1) *MGSM* (Shi et al., 2022), the multilingual version of grade-school math problems translated by human annotators and (2) *MT-AIME* (Son et al., 2025), the multilingual version of challenging mathematical problems from American Invitational Mathematics Examination (AIME). For multilingual math reasoning, our main experiments cover 10 languages: *Bn, De, En, Es, Fr, Ja, Ru, Sw, Th, Zh*.
- **Cultural Reasoning:** (1) *FORK* (Palta and Rudinger, 2023), a manually-curated set of English questions for probing cultural biases present in commonsense reasoning, with a specific focus on food-related customs, covering 10 global regions and (2) *CulturalBench* (Chiu et al., 2025), a set of 1,696 human written and human-verified English questions to assess LMs' cultural knowledge, covering 45 global regions.

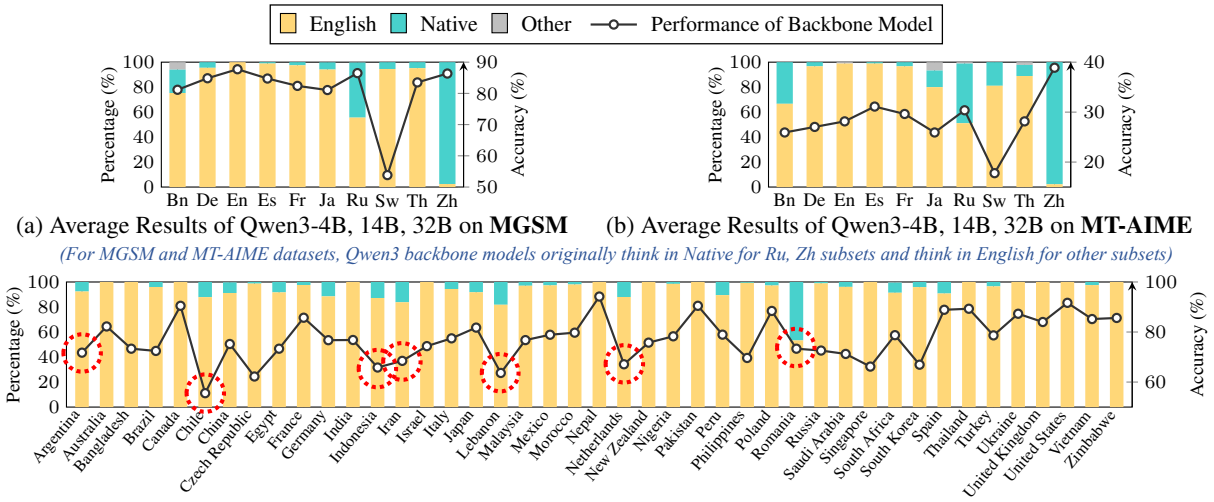
We follow the default official sampling parameters for each model and we report the Mean@3 (average results of three runs) performance in our main experiments. We set the maximum number of new tokens to 32,768 to ensure sufficient output length. To prevent meaningless repetitive reasoning during inference, we adopt a repetition-detection truncation mechanism to reduce unnecessary token generation and computational overhead. The implementations are provided in the Appendix A.3.

3.2 Overall Performance

The average Mean@3 results over different languages / cultures are presented in Table 1. The results for each language are in §A.4 Table 9. We also compare and anaylze the performance of *vanilla finetuning* and *majority voting* in Appendix §A.5.

(1) Evidence of Decoupling in Math and Cultural Reasoning. We observe a decoupling between a model's proficiency in math reasoning and culturally situated tasks. For instance, *DS-R1-Distill-Qwen-7B* substantially outperforms *DS-R1-Distill-Llama-8B* on *MGSM* (+19.69%) and *MT-AIME* (+11.39%), yet trails significantly in

| *Data Format: (Thinking Phase \| Answer Phase \| Both Phases)* | | |



(a) Average Results of Qwen3-4B, 14B, 32B on **MGSM** (b) Average Results of Qwen3-4B, 14B, 32B on **MT-AIME**

(For MGSM and MT-AIME datasets, Qwen3 backbone models originally think in Native for Ru, Zh subsets and think in English for other subsets)

(c) Average Results of Qwen3-4B, 14B, 32B on **CulturalBench**

(For Culturalbench dataset, Qwen3 backbone models originally think in English)

Figure 3: Distribution of thinking languages frequency by *x1* models along with the performance of their backbones across different languages/cultures subsets. All statistics and results are averaged over *Qwen3-4B*, *14B* and *32B*.

reasoning pathway becomes unreliable, the model increases thinking-language switching as a recovery mechanism, attempting to exploit language-specific knowledge or alternative reasoning pathways encoded in non-English representations.

4.2.2 Net Benefits of Language Switching

To assess the true utility of *x1*'s adaptive thinking mechanism, we focus on the subset of samples where a thinking-language switch occurred. We define the *Benefit Rate* as the proportion of cases where the backbone fails but *x1* succeeds after switching (computed over all samples), and the *Harm Rate* as the opposite case. Their difference reflects the *net benefit* of adaptive switching. As shown in Figure 4, the benefit consistently outweighs the harm, providing clear evidence for the effectiveness of thinking-language switching. Besides, the analysis further reveals several insights:

(1) Larger Cross-Lingual Disparity Leads to Greater Net Benefits from Thinking-Language Switching. From Figure 4 (a), we observe that in *MGSM* (grade school level), the net benefit of thinking-language switching decreases steadily as model size scales. But for the more challenging benchmark *MT-AIME* shown in (b), the net benefits remains low and nearly unchanged.

This trend aligns with the standard deviation (*Std*) statistics in §B Table 6, where a larger *Std* indicates greater cross-lingual performance disparity. We find a clear positive correlation: the greater this disparity, the larger the net benefit of switching.

For *MGSM*, the Qwen3 backbone becomes increasingly balanced across languages as model size grows (Std: 4B-16.96, 14B-7.446, 32B-4.665), naturally reducing the benefit of switching. Intuitively, if performance were identical across all languages, switching would offer no gain. For *MT-AIME*, although overall accuracy improves with scale, the performance imbalances remains low and stable (Std: 4B-6.576, 14B-8.308, 32B-7.495), explaining the consistently small and almost scale-invariant net benefit observed in Figure 4(b).

(2) Thinking in Culture-associated Language Offers Inherent and Persistent Advantages. Figure 4 (c)(d) shows that the net benefit of reasoning in culture-associated language remains stable across model scales, and is consistently larger than that observed in multilingual math reasoning. This indicates that the advantage of reasoning in culture-associated language is not a transient effect of limited model capacity, but instead reflects a deeper entanglement between language and culture.

This persistence also reveals a fundamental challenge to "Scaling Law": cultural knowledge is tightly intertwined with the linguistic forms through which it is acquired, represented, and invoked. Consequently, reasoning in English cannot fully substitute for reasoning in the culture-associated language, even as model size and overall capability continue to increase. This points to an intrinsic upper bound on English-centric reasoning in cultural contexts—one that is difficult to overcome through scaling alone.

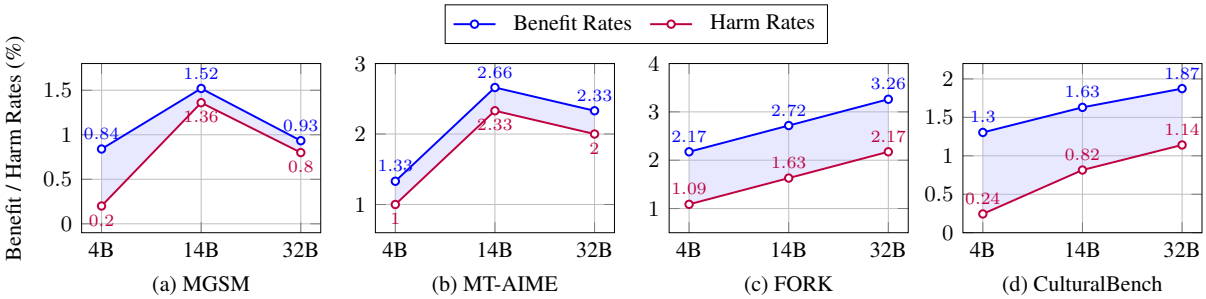


Figure 4: Benefit and harm rates of thinking-language switching on *MGSM*, *MT-AIME*, *FORK* and *CulturalBench* datasets across *Qwen3-4B*, *14B* and *32B*. The shaded area (benefit rate minus harm rate) represents the net benefits.

Models	Avg. Cultural Knowledge Recall Count per Thought (Efficiency)	Avg. Cultural Knowledge Recall Accuracy (%) (Accuracy)
Qwen3-4B	3.366	47.699
x1-Qwen3-4B	2.662 (↓ 0.704)	53.439 (↑ 5.740)
Qwen3-14B	3.353	49.123
x1-Qwen3-14B	2.765 (↓ 0.588)	55.319 (↑ 6.196)
Qwen3-32B	3.561	45.813
x1-Qwen3-32B	3.000 (↓ 0.561)	56.725 (↑ 7.902)

Table 3: Statistics of cultural knowledge recall frequency and accuracy on the *CulturalBench* dataset.

4.3 Cultural Knowledge Recall Behavior

Section 4.2.2 demonstrates the persistent advantages of thinking in culture-associated language. But what underlies this advantage? As illustrated in Appendix §C Figure 8, both backbone and *x1* models exhibit explicit *cultural knowledge recall* during cultural reasoning, actively retrieving relevant cultural facts or norms as part of their reasoning process. While such recall alone does not guarantee correctness, it plays a critical role in shaping the final answer. Motivated by this observation, we compare the cultural knowledge recall behavior of *x1* and its backbone models. Specifically, we employ *GPT-5-mini* to identify recalled cultural knowledge that supports or justifies the golden answer and to verify its correctness (implementations are in Appendix §C). The results are summarized in Table 3, leading to the following finding:

Thinking in Culture-associated Language Enables More Efficient and Accurate Cultural Knowledge Recall. As shown in Table 3, across all model scales, *x1* consistently recalls fewer but more accurate cultural knowledge instances compared to its backbone model. *x1* reduces the average recall count per thought by 0.56-0.70, while improving recall accuracy by 5.74%-7.90%. This indicates that reasoning in culture-associated languages promotes more targeted and relevant knowledge retrieval, rather than broader but noisier recall, providing a concrete mechanistic explanation for the persistent gains observed in cultural tasks.

4.4 Language-Mixing in *x1*

In multilingual reasoning, models often interleave multiple languages within a single reasoning trajectory, a phenomenon commonly known as *language mixing* (Yong et al., 2025) (we present a example in Figure 9). Unlike explicit thinking-language switching, which operates at the context level, language mixing captures a finer-grained, sentence-level adaptation that can occur without explicit switching. To isolate this effect, we focus on samples where *x1* does not perform explicit thinking-language switching. For these samples, we measure the change in sentence-level language-mixing strength of *x1*'s responses relative to its backbone and analyze its correlation with the resulting net benefits aggregated over all samples.

Increased Language Mixing Is Associated with Positive Outcomes in Multilingual Reasoning. As we do not introduce any design about language mixing, we observe that on both *MGSM* and *MT-AIME*, the proportions of samples with increased and decreased language mixing rates are broadly balanced. However, those with increased language mixing rates consistently yield higher net benefits than those with decreased mixing (0.26% vs. 0.21% on *MGSM*, 0.85% vs. 0.11% on *MT-AIME*).

These results indicate that the gains of *x1* cannot be fully attributed to explicit language switching. Even in the absence of such switching, higher sentence-level language mixing rates is positively correlated with improved performance. Language mixing thus appears to act as a fine-grained mechanism that integrates complementary linguistic representations within a single reasoning trajectory, and may further provide a useful guidance signal for model optimization (e.g., in reward design).

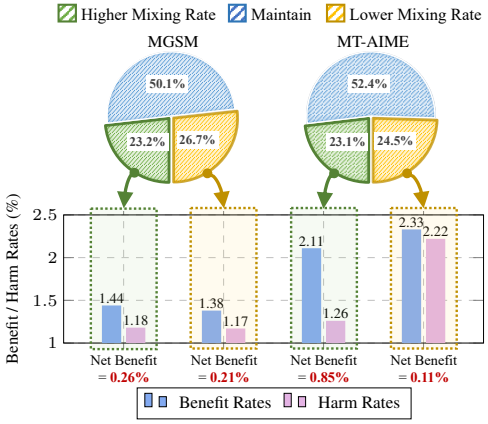


Figure 5: Proportions of samples with changed language mixing strength in *x1*'s responses relative to its backbone, and the correlation with the resulting net benefits.

4.5 Preference Optimization with DPO

The instance-level contrastive signals naturally induce implicit preferences between alternative reasoning-language choices. In this subsection, we further apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) as an auxiliary enhancement, treating reasoning trajectories in the advantageous language as preferred samples and those in the less advantageous language as dispreferred ones[4]. The results are shown in Appendix §E Table 8, where we observe that DPO further improves the performance of *x1* models, achieving a second jump in multilingual math reasoning.

5 Related Work

Multi- & Cross-lingual Reasoning. Recent studies examined how LLMs perform reasoning across languages. Some approaches utilize response alignment, latent connection and language separability to bridge the gap between languages (Zhu et al., 2024; Ye et al., 2025b,a). Besides, Ko et al. (2025) anchor reasoning in English to minimize the gaps, whereas SLAM (Fan et al., 2025) maintains reasoning capabilities with lower computational costs. Further refinement is achieved through reasoning-focused tuning like mCoT (Lai and Nissim, 2024), preference-based alignment (She et al., 2024), and process reward modeling (Wang et al., 2025). Some previous works focus on prompting strategies (Qin et al., 2023). Researchers proposed $\mathcal{X}$Transplant for cross-lingual complementarity (Ye et al., 2024b), dictionary insertion prompting (Lu et al., 2024) that embeds key English terms into native prompts and program-based demonstrations (Ranaldi et al., 2025) for structured logic transfer. Recently, Yong et al. (2025) conducted a detailed analysis, revealing the potentials and limitations of cross-lingual test-time scaling.

Cultural Reasoning. Early efforts primarily focused on constructing culturally grounded knowledge bases, such as CANDLE (Nguyen et al., 2023) for normative assertions, CultureBank (Shi et al., 2024) for community-driven cultural descriptors, MAPS (Liu et al., 2024) for figurative expressions, and GlobeSumm (Ye et al., 2024a) for global perspectives on international news. From a methodological perspective, existing work explores how to elicit or control cultural perspectives in LLMs (Kovač et al., 2023), enhance cultural reasoning by explicitly recognizing cultural context in prompts (Wang et al., 2024), integrate moral reasoning across normative ethics frameworks (Rao et al., 2023), and complement English reasoning with non-English cultural features via cross-lingual transplantation (Ye et al., 2024b).

Rather than enforcing consistency across languages, we investigate whether different languages inherently provide complementary reasoning advantages for different instances. By framing reasoning-language choice as a self-aware decision process, our method bridges multilingual and cultural reasoning from a unified perspective, positioning language as an active component of reasoning rather than a passive carrier of content.

6 Conclusion

In this work, we presented *x1* models that enable adaptive multilingual reasoning. By constructing *x1* without expanding the model's knowledge boundaries, we isolated the effect of reasoning-language choice and demonstrated that linguistic diversity can be systematically leveraged to improve both multilingual mathematical reasoning and culturally grounded reasoning. Beyond empirical gains, our findings challenge a simplistic interpretation of scaling laws, revealing that even as models grow larger and more capable, the advantages of reasoning in culture-associated languages persist rather than vanish. This highlights reasoning-language choice as a functional component of reasoning rather than a superficial artifact of generation. We hope this work encourages future research to move beyond monolithic reasoning paradigms toward building LLMs that are more globally and culturally competent.

[4]Reusing the same data produced in Section 2 Step 2.

Limitations

This work exhibits several limitations worth noting. First, to enable a clean comparison between alternative reasoning pathways, for each instance, we intentionally restrict the candidate thinking languages to English and one non-English language (which is determined by the query language or cultural background). Although this design choice facilitates controlled analysis, it limits the exploration of richer multilingual combinations and linguistic diversity. Extending adaptive reasoning to a larger pool of languages can be a further exploration. Second, while our analysis reveals the potential positive impact of "language mixing" for reasoning, we do not pursue this direction further due to space constraints. Exploring principled ways to incorporate mixed-language reasoning, for example as a training signal or an optimization objective can be great future work. Thirdly, for culture-related tasks, we employ an LLM-as-a-Judge framework to assess response quality. Although this approach enables scalable evaluation, it may inherit the biases and limitations of the evaluator model. We mitigate this risk by using evaluation only for relative comparison between reasoning-language choices, rather than for absolute or normative judgments.

Ethical Considerations

This work studies multilingual and culturally grounded reasoning in large language models. All model outputs, especially those related to cultural knowledge or perspectives, are generated by the models themselves and do not represent the views or opinions of the authors. Our method does not introduce new knowledge sources or external supervision, and therefore does not increase the risk of generating sensitive or private information beyond the capabilities of the backbone models. While reasoning in culture-associated languages may surface different cultural expressions or viewpoints, such outputs should not be interpreted as authoritative or normative cultural judgments. In this paper, we use Gemini to correct grammatical errors.

References

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

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

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. Slam: Towards efficient multilingual reasoning via selective language alignment. *arXiv preprint arXiv:2501.03681*.

Cliff Goddard. 2003. Thinking across languages and cultures: Six dimensions of variation. *Cognitive linguistics*, 14(2/3):109–140.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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

Hyunwoo Ko, Guijin Son, and Dasol Choi. 2025. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv preprint arXiv:2501.02448*.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Zoltan Kovecses. 2006. *Language, mind, and culture: A practical introduction*. Oxford University Press.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039.

Hongyuan Lu, Zixuan Li, and Wai Lam. 2024. Dictionary insertion prompting for multilingual reasoning on multilingual large language models. *arXiv preprint arXiv:2411.01141*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM web conference 2023*, pages 1907–1917.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

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. *Advances in neural information processing systems*, 36:53728–53741.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025. When natural language is not enough: The limits of in-context learning demonstrations in multilingual reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7369–7396.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, Daan Van Stigt, Marcos Treviso, Luísa Coheur, José GC de Souza, André FT Martins, et al. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. *URL https://arxiv. org/abs/2408.03314*, 20.

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

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Demystifying multilingual chain-of-thought in process reward modeling. *arXiv preprint arXiv:2502.12663*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ludwig Wittgenstein. 2023. Tractatus logico-philosophicus.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yangfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024a. GlobeSumm: A challenging benchmark towards unifying multi-lingual, cross-lingual and multi-document news summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10803–10821, Miami, Florida, USA. Association for Computational Linguistics.

Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Lei Huang, Weitao Ma, Qichen Hong, Yunfei Lu, Duyu Tang, Dandan Tu, and Bing Qin. 2025a. Langgps: Language separability guided data pre-selection for joint multilingual instruction tuning. *arXiv preprint arXiv:2511.10229*.

Yangfan Ye, Xiaocheng Feng, Xiachong Feng, Libo Qin, Yichong Huang, Lei Huang, Weitao Ma, Qichen Hong, Zhirui Zhang, Yunfei Lu, et al. 2024b. Exploring cross-lingual latent transplantation: Mutual opportunities and open challenges. *arXiv preprint arXiv:2412.12686*.

Yangfan Ye, Xiaocheng Feng, Zekun Yuan, Xiachong Feng, Libo Qin, Lei Huang, Weitao Ma, Yichong Huang, Zhirui Zhang, Yunfei Lu, Xiaohui Yan, Duyu Tang, Dandan Tu, and Bing Qin. 2025b. CC-tuning: A cross-lingual connection mechanism for improving joint multilingual supervised fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19036–19051, Vienna, Austria. Association for Computational Linguistics.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *arXiv preprint arXiv:2405.01345*.

A Experiment Details

A.1 Step 1: To be a Multilingual Reasoner

Language List. The 31 languages (English included) involved in step 1 training are as follows:

```
["Arabic","Bulgarian","Bengali","German","Greek",
"English","Spanish","Finnish","French","Hebrew",
"Hindi","Hungarian","Indonesian","Italian",
"Japanese","Korean","Malay","Dutch","Polish",
"Portuguese","Romanian","Russian","Swedish",
"Swahili","Thai","Tagalog","Turkish","Ukrainian",
"Urdu","Vietnamese","Chinese"]
```

Quality Filtering. To ensure linguistic fidelity, we apply quality filtering using COMET (Unbabel/wmt23-cometkiwi-da-xxl) scores and discard translations of evidently poor quality (whose scores fall below 0.4).

Data Format. We adopt the official training templates provided by the corresponding backbone models. For instance, when training *Qwen3-4B*, *14B*, and *32B*, we use the official *"Qwen3"* template to ensure consistency with the original training and inference protocols. Beyond the official templates, we introduce a customized template tailored for multilingual reasoning control. This template explicitly specifies the language used in the reasoning process while keeping the final answer aligned with the prompt language. Real examples are shown in Figure 6. Note that, due to the diversity of X and to preserve the generality of the template, we do not register "<X_*start*>" series as special tokens in the tokenizer.

Training Hyperparameters. We perform full-parameter finetuning for all models, except for *Qwen3-32B*, for which we adopt LoRA training due to its large model size.

```
# for all models except Qwen3-32B
per_device_train_batch_size: 1
gradient_accumulation_steps: 16
learning_rate: 1.0e-5
num_train_epochs: 3.0
bf16: true

# for Qwen3-32B
lora_rank: 8 # default lora_alpha: lora_rank * 2
lora_target: all
per_device_train_batch_size: 1
gradient_accumulation_steps: 16
learning_rate: 1.0e-4
num_train_epochs: 3.0
bf16: true
```

A.2 Step 2: To be an Adaptive Reasoner

Training Data Statistics.

- (1) Multilingual Math Problems: We use the *MGSM8KInstruct* dataset (Chen et al., 2023), which contains math problems paired with their correct answers. We sample 200 questions across 10 languages, resulting in 2,000 training instances. Detailed statistics are as follows:

```
{   // language: sample size
    "Bengali": 200,
    "German": 200,
    "English": 200,
    "Spanish": 200,
    "French": 200,
    "Japanese": 200,
    "Russian": 200,
    "Swahili": 200,
    "Thai": 200,
    "Chinese": 200
}
// Total: 2000
```

- (2) Culture-related Problems: We use the *CultureBank* dataset (Shi et al., 2024), which provides cultural questions along with the underlying cultural knowledge they reflect. We sample 100-200 questions from each of the 25 language groups (depending on availability), covering cultural questions from 45 countries/regions, resulting in 4,413 samples in total. Detailed statistics are as follows:

```
{   // language group: sample size
    "Arabic": 200,
    "Danish": 122
    "German": 200,
    "Greek": 100,
    "English": 200,
    "Spanish": 200,
    "Finnish": 142,
    "French": 200,
    "Irish": 157,
    "Scottish Gaelic": 125,
    "Hindi": 200,
    "Indonesian": 200,
    "Italian": 200,
    "Japanese": 200,
    "Korean": 200,
    "Maori": 141,
    "Malay": 130,
    "Dutch": 200,
    "Norwegian": 173,
    "Polish": 123,
    "Portuguese": 200,
    "Russian": 200,
    "Swedish": 200,
    "Tagalog": 200,
    "Chinese": 200
}
// Total: 4413
```

Each language group is associated with a list of countries/regions:

```
{   // language group: list of
    // involved countries/regions
    "Arabic":["Algeria","Arab","Egypt",
    "Iraq","Lebanon","Syria",
    "Morocco","Tunisia","Jordan"],
    "Danish":["Denmark"],
    "German":["Germany","Austria",
    "Switzerland"],
    "Greek":["Greece","Cyprus"],
    "English":["United States"],
    "Spanish":["Argentina","Spain",
    "Cuba","Chile","Colombia",
    "Dominican Republic","Mexico",
    "Peru"],
    "Finnish":["Finland"],
    "French":["France"],
    "Irish":["Ireland"],
    "Scottish Gaelic":["Scotland"],
    "Hindi":["India"],
    "Indonesian":["Indonesia"],
    "Italian":["Italy"],
    "Japanese":["Japan"],
    "Korean":["Korea"],
    "Maori":["New Zealand"],
    "Malay":["Malaysia"],
    "Dutch":["Netherlands","Belgium"]
```

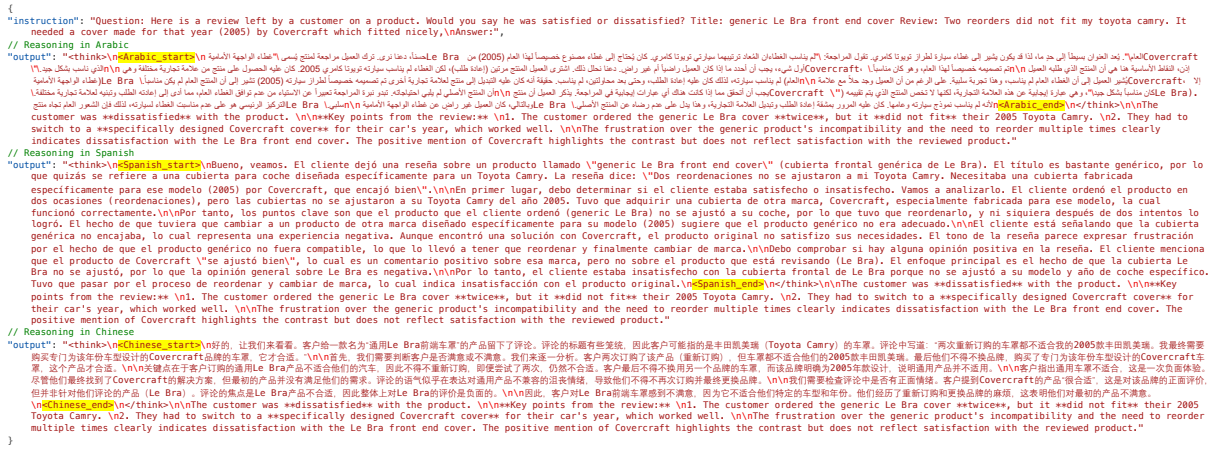


Figure 6: Examples of Step 1 constructed training data (reasoning in Arabic, Spanish and Chinese).

Advantageous Language Identification. For each question in D_{math} or D_{culture}, we first obtain the paired reasoning trajectories in both English and non-English with $\mathcal{M}$ and $\mathcal{M}_{\text{surface}}$[3]. We identify the advantageous reasoning language based on the quality of the final answers, under the principle that stronger reasoning should manifest in superior outputs. For multilingual math problems, answer quality is determined via numerical matching.

$$score_{\text{math}}(\hat{y}, y) = \begin{cases} 10, & \text{if } \hat{y} \text{ matches } y, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $\hat{y}$ denotes the model-predicted numerical answer and y is the ground-truth answer.

For culture-related questions, we employ *GPT-5-mini* to score how well the response entails the relevant cultural knowledge.

$$score_{\text{culture}}(\hat{y}, n) = \text{Entailment}(Q, \hat{y}, n, LLM) \in [0, 10] \quad (2)$$

where Q denotes the culture-related question, $\hat{y}$ denotes the model-predicted response and n represents the relevant cultural knowledge or norm that the model is expected to comply with. The prompt for scoring is shown in Table 4.

The reasoning language associated with the higher-scoring trajectory is then regarded as the advantageous thinking language for that instance.

Prompt for Scoring Cultural Knowledge Entailment
You will be given a culture-related question, a related cultural knowledge and a response from an agent. Your task is to provide a 'total rating' scoring how well the response from an agent answers the culture-related question and entails the related cultural knowledge. Give your answer as a float on a scale of 0 to 10, where 0 means that the response from agent is not helpful at all, and 10 means that the answer completely and helpfully addresses the question. Your output should be a float score only. Question: {question} Cultural Knowledge: {cultural knowledge description} Response: {model response} Total rating (a float on a scale of 0 to 10):

Table 4: Prompts designed for scoring cultural knowledge entailment.

Data Format. Besides the data format introduced in Figure 6, we also construct self-awareness data that externalizes the model's language-selection decision as an explicit prediction task and the format is shown in Figure 7.

Training Hyperparameters. We apply LoRA training separately for the multilingual math and cultural reasoning scenarios. For *MGSM8KInstruct*, we set the LoRA rank to 4, as we empirically observe that using a larger rank leads to unexpected and substantial drops in performance.

```yaml
# for MGSM8KInstruct
lora_rank: 4 # default lora_alpha: lora_rank * 2
lora_target: all
per_device_train_batch_size: 1
gradient_accumulation_steps: 16
learning_rate: 1.0e-4
bf16: true

# for CultureBank
```

```
{
"instruction": "You are an AI assistant trained to
  decide in which language you should internally
  think in order to maximize the probability of
  correctly answering a given question.\n\nGiven a
  question, choose the single language (e.g.,
  English, Chinese, Arabic, etc.) you believe will
  give you the highest chance of producing a correct
  answer based on your own knowledge.\n\nYour output
  should be only the name of the language without
  any additional words or explanations. Even if you
  are uncertain, you must pick exactly one
  language.\n\nQuestion: I'm planning a trip to
  Egypt, and I'm really excited to learn more about
  the local culture and history. I've heard that
  there are so many incredible historical sites to
  visit. How can I best engage with the local
  community and learn about their deep-rooted
  appreciation for their own history during my
  travels? I want to make sure I'm respectful and
  open to their insights.\n\nThinking Language:",
"output": "<think>\n\n</think>\n\nArabic"
}
```


Figure 7: Example of self-awareness data, where "Arabic" is the advantageous reasoning language.


```
lora_rank: 16 # default lora_alpha: lora_rank * 2
lora_target: all
per_device_train_batch_size: 1
gradient_accumulation_steps: 16
learning_rate: 1.0e-4
bf16: true
```


A.3 Repetition-Detection Truncation Mechanism

To prevent the model from producing meaningless repetitive reasoning during inference, we implement a repetition-detection truncation mechanism based on a custom `StoppingCriteria`. The method monitors the generated text in fixed-size blocks and terminates decoding if a newly generated block has appeared earlier in the output, indicating repetitive looping behavior.

Algorithm 1 Repetition-Detection Truncation Mechanism

Require: Tokenizer $\mathcal{T}$, prompt length L, block size B, model $\mathcal{M}$
Initialize generated sequence $X \leftarrow$ model prompt
flag $\leftarrow$ False
while not flag **do**
  $x \leftarrow \mathcal{M}.\texttt{generate_next}()$
  Append x to X
  $T \leftarrow \mathcal{T}.\texttt{decode}(X[L:])$ ▷ decode newly generated text
  if $|T| < 2B$ **then**
    continue ▷ insufficient text length
  end if
  last_block $\leftarrow T[|T| - B : |T|]$
  prefix $\leftarrow T[0 : |T| - B]$
  if last_block $\in$ prefix **then**
    flag $\leftarrow$ True ▷ repetition detected
    break
  end if
end while
return X

Models	Multilingual Math Reasoning			
	MGSM		MT-AIME	
	Non-Think	***Think***	*Non-Think*	***Think***
Backbone Models				
Qwen3-4B	70.21	76.59	12.89	21.78
Qwen3-14B	77.64	82.56	19.33	29.22
Qwen3-32B	80.52	83.98	21.83	33.89
+ Vanilla Finetuning				
Qwen3-4B	50.80		1.667	
Qwen3-14B	64.24		2.000	
Qwen3-32B	72.68		0.333	
Backbone + Majority Voting				
Qwen3-4B	87.46	92.00	16.67	35.56
Qwen3-14B	89.33	92.93	28.89	47.78
Qwen3-32B	89.20	93.07	34.44	50.00
x1 Series Models				
x1-Qwen3-4B	70.30	77.69	13.56	22.83
x1-Qwen3-14B	77.38	83.64	19.44	33.11
x1-Qwen3-32B	80.12	84.43	22.11	34.50
x1 Series Models + Majority Voting				
x1-Qwen3-4B	88.80	93.33	20.00	35.56
x1-Qwen3-14B	90.27	93.73	31.11	48.89
x1-Qwen3-32B	90.00	93.20	36.67	50.00

Table 5: Supplementary experimental Mean@3 results that are the averages of the performance across all languages / cultures involved for each dataset.

A.4 Results For Each Subset

We present the results for each language subset for *MGSM* and *MT-AIME* in Table 9.

A.5 Supplementary Experimental Results

We present the supplementary experimental results of (1) *vanilla finetuning* and (2) *majority voting* in Table 5. Since the *CultureBank* dataset lacks reference answers for supervised training, and *FORK* and *CulturalBench* are not designed to support multilingual evaluation, we restrict the analysis to multilingual mathematical reasoning tasks.

- **Vanilla Finetuning:** We perform vanilla finetuning (lora) for the backbone models with *MGSM8KInstruct* dataset.
- **Majority Voting:** For each math problem, we query the model using 10 parallel versions of the same question in different languages and aggregate the resulting answers via majority voting, taking the most frequent answer as the final prediction. In this setting, we ignore constraints on response-language consistency and the voting outcome serves as an approximate indicator of the potential gains from linguistic diversity.

We observe that continuous finetuning on already strong *Qwen3*-series models leads to a noticeable degradation in performance. This degradation

Models	Mean@3 Performance on MGSM											
	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Average	**Standard Deviation**
Qwen3-4B	75.33	82.53	87.60	84.67	81.6	78.67	84.67	26.80	79.33	84.67	76.59	**16.96**
Qwen3-14B	83.60	85.60	86.60	84.60	82.80	82.20	87.00	60.80	86.00	86.40	82.56	**7.446**
Qwen3-32B	84.00	86.60	89.20	84.20	83.20	82.80	86.80	71.80	84.20	87.00	83.98	**4.665**
Models	Mean@3 Performance on MT-AIME											
	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Average	**Standard Deviation**
Qwen3-4B	22.22	25.56	25.56	24.44	23.33	22.22	17.78	6.67	25.56	24.44	21.78	**6.576**
Qwen3-14B	26.67	26.67	32.22	25.56	34.44	24.44	34.44	18.89	23.33	45.56	29.22	**8.308**
Qwen3-32B	28.89	28.89	26.67	43.33	31.11	31.11	38.89	27.78	35.56	46.67	33.89	**7.495**

Table 6: Performance of backbone models on *MGSM* and *MT-AIME* along with the average and standard deviation. The larger the standard deviation, the greater the cross-lingual performance disparity.

is likely due to the limited difficulty and suboptimal quality of the reasoning traces in *MGSM8KInstruct*, which can substantially disrupt the models' existing reasoning capabilities.

By contrasting "*Backbone + Majority Voting*" with "*x1 + Majority Voting*", we find that the upper bound afforded by linguistic diversity remains remarkably high: aggregating predictions across 10 languages substantially surpasses the models' original performance. Moreover, the *x1* series consistently exhibits a higher majority-voting upper bound than their backbone models, suggesting that *x1* more effectively integrates and aligns reasoning signals across languages, thereby enabling linguistic diversity to be exploited more coherently.

B Performance Standard Deviation Across Languages

The statistics of performance standard deviation (*Std*) across languages for backbone models in *MGSM* and *MT-AIME* is presented in Table 6. Larger *Std* indicates greater cross-lingual performance disparity. The statistics provides important support for the analysis in Section 4.2.2 (1).

C Cultural Knowledge Recall Analysis

We observe that both backbone and *x1* models exhibit explicit *cultural knowledge recall* behavior during cultural reasoning, actively retrieving relevant cultural facts or norms as part of their reasoning process. While such recall alone does not guarantee correctness, it plays a critical role in shaping the final answer. As illustrated in Figure 8, we present a pair of examples of cultural knowledge recall behavior produced by *Qwen3-32B* (thinking in English) and *x1-Qwen3-32B* (thinking in Chinese) on a question from *CulturalBench* dataset. We use high-light to highlight the cultural knowledge recall behavior, where the green parts are the key recall behaviors that directly lead to the model's correct answer.

Motivated by this observation and to further analyze such behavior, we employ *GPT-5-mini* to identify recalled cultural knowledge that supports or justifies the golden answer and to verify its correctness. The designed prompts are presented in Table 7.

As shown in Table 3, across all model scales, *x1* consistently recalls fewer but more accurate cultural knowledge instances compared to its backbone model. *x1* reduces the average recall count per thought by 0.56-0.70, while improving recall accuracy by 5.74%-7.90%. This indicates that reasoning in culture-associated languages promotes more targeted and relevant knowledge retrieval, rather than broader but noisier recall, providing a concrete mechanistic explanation for the persistent gains observed in cultural tasks.

D Language Mixing Behavior

We present a example of language mixing behavior in Figure 9, where the default reasoning in English is interleaved with reasoning in Japanese.

E Further Enhancement with Direct Preference Optimization

In our main experiments, our trainining strategy only leverages the signals from the advantageous language. However, the instance-level contrastive signals between the advantageous and less advantageous languages naturally induce implicit preferences between alternative reasoning-language choices. In this subsection, we further apply Direct Preference Optimization (DPO) as an auxiliary enhancement, treating reasoning trajectories in the advantageous language as preferred samples and those in the less advantageous language as dispreferred ones. The results are shown in Table 8.

Question:

In Chinese culture, what should be avoided during Ghost Festival in Summertime in July?\n(A) Dinning out\n(B) Going haircut\n(C) Go swimming\n(D) Going out

Qwen3-32B's thinking content:

Okay, so I need to figure out what should be avoided during the Ghost Festival in Chinese culture during the summertime in July. The options are dining out, going for a haircut, going swimming, or going out. Let me think about what I know about the Ghost Festival, also known as the Hungry Ghost Festival. \n\nI remember that this festival is a time when it's believed that ghosts and spirits are released from the underworld. People perform rituals to appease them, offer food, and avoid certain activities to prevent offending the spirits. There are some common taboos during this period. For example, I think it's not advisable to go out at night because the spirits are more active. Also, some people might avoid activities like getting a haircut because it's thought that the spirits might take your hair, or that cutting hair could be seen as a bad omen. \n\nLooking at the options, (B) is going for a haircut. I've heard that getting a haircut during this time is discouraged. But wait, the question is about the summertime in July. The Ghost Festival is usually in the seventh lunar month, which can fall in July or August in the Gregorian calendar. The main thing is the seventh lunar month. \n\nAnother point: some people avoid dining out because the ghosts might be around, and you don't want to invite them to your meal. However, the answer might be more specific. Also, going swimming could be risky if the spirits are near water. But I'm not sure. \n\nI think the most common advice is to avoid going out at night, but option (D) is "Going out," which is a bit vague. But maybe during the Ghost Festival, people are advised to stay indoors, especially at night. However, the options here are (D) Going out, which could mean in general. Alternatively, maybe cutting hair is more specific. \n\nWait, I recall that during the Ghost Festival, it's considered bad luck to get a haircut because the spirits might take your hair, or because the act of cutting hair could be associated with bad omens. So (B) Going haircut might be the correct answer. But I need to confirm. \n\nAlternatively, some sources say that during the Ghost Festival, people avoid certain activities like getting married, moving house, or getting a haircut. So (B) is likely the answer. Let me check again. The question is about what should be avoided. So the answer is (B) Going haircut.

***x1*-Qwen3-32B's thinking content:**

好的，我现在需要回答这个问题。问题是在中国传统文化中，夏季七月的鬼节期间应该避免做什么。选项是：(A) 外出吃饭，(B) 理发，(C) 游泳，(D) 外出。\n\n首先，我得回忆一下鬼节的相关习俗。鬼节，也称为中元节，在农历七月十五日庆祝。这个节日与亡灵有关，人们会举行各种仪式来安抚游魂野鬼。在鬼节期间，有一些特定的禁忌需要遵守。\n\n接下来，我需要逐一分析选项。选项A是外出吃饭。我听说在鬼节期间，有些人可能会避免在户外或公共场所吃饭，因为担心食物会吸引鬼魂。不过，这可能因地区而异，我不确定这是否是普遍的禁忌。\n\n选项B是理发。在中国传统文化中，理发通常与某些节日或时期有关。例如，在春节期间，人们会避免在年初一理发，因为“发”与“福”谐音，而年初一理发可能被认为会剪掉福气。但鬼节期间是否也有类似的禁忌呢？我不太确定这一点。\n\n选项C是游泳。在鬼节期间，人们可能会避免游泳，因为水边是鬼魂容易出没的地方。此外，有些人认为在鬼节游泳可能会遇到溺水的鬼魂，从而带来危险。这可能是常见的禁忌。\n\n选项D是外出。这个选项比较笼统。在鬼节期间，有些人可能会避免夜间外出，尤其是独自外出，因为相信鬼魂会在那时游荡。不过，选项D说的是“外出”，并没有说明时间，所以可能不够具体。\n\n现在我需要确认这些习俗。根据我所知，鬼节期间常见的禁忌包括避免游泳、夜间外出、在户外点火或做饭，以及避免穿着红色衣物。其中，游泳是一个常见的禁忌，因为水边是鬼魂容易出没的地方。因此，选项C可能是正确答案。\n\n再检查其他选项。选项B，理发，可能在某些地区有禁忌，但我不确定是否普遍。选项A和D可能不是主要的禁忌。因此，最可能的正确答案是选项C：游泳。

(English version of above Chinese thinking content for better understandability)

Okay, now I need to answer this question. The question is, in traditional Chinese culture, what should one avoid doing during the Ghost Festival in the seventh lunar month? The options are: (A) eating out, (B) getting a haircut, (C) swimming, (D) going out. First, I need to recall the relevant customs of the Ghost Festival. The Ghost Festival, also known as the Zhongyuan Festival, is celebrated on the 15th day of the seventh lunar month. This festival is related to the spirits of the dead, and people hold various rituals to appease wandering ghosts. During the Ghost Festival, there are some specific taboos to observe. Next, I need to analyze the options one by one. Option A is eating out. I've heard that during the Ghost Festival, some people may avoid eating outdoors or in public places because they are worried that the food will attract ghosts. However, this may vary from region to region, and I'm not sure if this is a universal taboo. Option B is getting a haircut. In traditional Chinese culture, getting a haircut is usually associated with certain festivals or periods. For example, during the Spring Festival, people avoid getting a haircut on the first day of the new year because "hair" (发) is a homophone for "fortune" (福), and getting a haircut on the first day of the new year may be considered as cutting away good fortune. But are there similar taboos during Ghost Festival? I'm not sure. Option C is swimming. During Ghost Festival, people might avoid swimming because watersides are considered haunted by ghosts. Additionally, some believe that swimming during Ghost Festival might lead to encounters with drowning ghosts, thus bringing danger. This is likely a common taboo. Option D is going out. This option is rather general. During Ghost Festival, some people might avoid going out at night, especially alone, because they believe ghosts roam at that time. However, Option D says "going out" without specifying the time, so it might not be specific enough. Now I need to confirm these customs. As far as I know, common taboos during Ghost Festival include avoiding swimming, going out at night, lighting fires or cooking outdoors, and avoiding wearing red clothing. Swimming is a common taboo because watersides are considered haunted by ghosts. Therefore, Option C is likely the correct answer. Let's check the other options. Option B, getting a haircut, might be taboo in some areas, but I'm not sure if it's widespread. Options A and D are probably not major taboos. Therefore, the most likely correct answer is option C: swimming.

Figure 8: Examples of cultural knowledge recall behavior.

Question:

ジョンは時速60マイルで3時間車を運転した後に、自宅に非常に大事なものを忘れたことに気づいたため引き返します。4時間以内に自宅に戻ろうとしますが、最初の2時間は大渋滞で動けません。次の30分間を時速30マイルで運転し、その後時速80マイルで自宅に向かいました。4時間が経過したタイミングで、ジョンは自宅から何マイル離れていることになりますか?

Qwen3-32B's thinking content:

Okay, let me try to figure out this problem step by step. So, John drives for 3 hours at 60 mph, then realizes he forgot something and has to go back home. He wants to get back within 4 hours total, but the first 2 hours after turning around are stuck in traffic. Then he drives 30 minutes at 30 mph, then the rest at 80 mph. We need to find out how far he is from home after 4 hours total.\n\nFirst, let's break down the timeline. Initially, John drives away from home for 3 hours at 60 mph. So, the distance he covers in that time is speed multiplied by time, which is 60 * 3 = 180 miles. So he's 180 miles away from home when he realizes he forgot something.\n\nNow, he needs to return home. The total time he has from the moment he starts returning is 4 hours minus the initial 3 hours he already spent driving away. Wait, hold on. Wait, the problem says he wants to get back home within 4 hours. Wait, does the 4 hours start from when he realized he forgot the item, or from the beginning?\n\nWait, let me check the problem again: \"4時間以内に自宅に戻ろうとしますが、最初の2時間は大渋滞で動けません。次の30分間を時速30マイルで運転し、その後時速80マイルで自宅に向かいました。4時間が経過したタイミングで...\"\n\nSo, he starts trying to return home, and within 4 hours (from the time he starts returning?), but the first 2 hours are stuck in traffic …

Figure 9: Example of language mixing behavior.

Prompt for Identifying Cultural Knowledge Recall Behaviors

(Input:)
Given the following QA pair, I will provide you with my reasoning process.
Your task is to determine whether my reasoning includes any recall of cultural norms that directly support or justify the Golden Answer.
If such culturally relevant recall exists, extract only the portions that are directly tied to the Golden Answer and return them in a Python list.

Question: {question}

Golden Answer: {answer}

Reasoning Process: {reasoning}

(Output:)
{Response of Extracted Cultural Knowledge Recalls}

Prompt for Verifying Cultural Knowledge Recall Behaviors

(Input:)
Given the following QA pair, I will provide you with a cultural statement.
Your task is to determine whether this statement provides a decisive and indispensable contribution to arriving at the Golden Answer.
Mere relevance or weak association does not count.
Return only True or False.

Question: {question}

Golden Answer: {answer}

Cultural Statement: {norm}

(Output:)
{True or False}

Table 7: Prompts for identifying and verifying cultural knowledge recall behaviors.

Models	Multilingual Math Reasoning		Cultural Reasoning	
	MGSM *Think*	MT-AIME *Think*	FORK *Think*	CulturalBench *Think*
Open-sourced Reasoning Models				
Qwen3-4B	76.59	21.78	73.73	70.85
Qwen3-14B	82.56	29.22	73.91	78.24
Qwen3-32B	83.98	33.89	81.88	81.26
x1 Series Models				
	+ Math		***+ Culture***	
x1-Qwen3-4B	77.69	22.83	78.08	72.74
x1-Qwen3-14B	83.64	33.11	76.81	81.58
x1-Qwen3-32B	84.43	34.50	82.61	81.99
x1 Series Models + DPO				
	+ Math		***+ Culture***	
x1-Qwen3-4B	78.23	22.00	77.72	72.51
x1-Qwen3-14B	83.56	33.33	77.90	81.34
x1-Qwen3-32B	84.56	38.67	83.70	81.05

Table 8: Results after applying DPO to *x1* models.

	MGSM										
	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Average
Backbone Models											
Qwen3-4B (non-think)	68.67	76.27	83.87	75.73	74.40	73.07	81.20	15.87	74.00	79.07	70.21
Qwen3-4B (think)	75.33	82.53	87.60	84.67	81.60	78.67	84.67	26.80	79.33	84.67	76.59
Qwen3-14B (non-think)	80.20	81.20	86.00	80.60	78.40	78.60	82.80	43.80	81.60	83.20	77.64
Qwen3-14B (think)	83.60	85.60	86.60	84.60	82.80	82.20	87.00	60.80	86.00	86.40	82.56
Qwen3-32B (non-think)	80.8	80.8	85.6	82	77.6	80.8	84.4	65.2	83.6	84.4	80.52
Qwen3-32B (think)	84.00	86.60	89.20	84.20	83.20	82.80	86.80	71.80	84.20	87.00	83.98
DeepSeek-R1-Distill-Qwen-7B (non-think)	46.00	61.00	80.40	69.00	63.40	45.60	59.20	5.00	44.20	73.80	54.76
DeepSeek-R1-Distill-Qwen-7B (think)	50.40	67.07	80.00	71.07	68.80	51.87	71.87	7.07	52.13	80.27	60.05
DeepSeek-R1-Distill-Llama-8B (non-think)	12.13	46.00	65.73	49.60	45.60	35.33	47.33	3.73	19.87	56.40	38.17
DeepSeek-R1-Distill-Llama-8B (think)	11.73	45.07	66.40	51.87	47.07	33.87	44.93	5.07	25.47	72.13	40.36
x1 Series Models											
x1-Qwen3-4B (non-think)	66.20	76.40	84.80	78.00	75.60	71.40	78.80	17.80	76.60	77.40	70.30
x1-Qwen3-4B (think)	77.47	84.53	91.07	85.47	81.87	78.93	85.47	27.33	80.13	84.67	77.69
x1-Qwen3-14B (non-think)	78.00	79.20	86.40	79.00	77.40	80.00	83.00	44.40	84.40	82.00	77.38
x1-Qwen3-14B (think)	82.00	86.00	88.80	86.20	84.00	83.00	87.60	61.20	88.40	89.20	83.64
x1-Qwen3-32B (non-think)	82.00	82.40	87.60	82.40	76.80	80.00	84.80	62.00	78.40	84.80	80.12
x1-Qwen3-32B (think)	85.47	84.80	89.87	85.07	83.07	84.40	87.33	73.33	84.27	86.67	84.43
x1-DeepSeek-R1-Distill-Qwen-7B (non-think)	50.00	57.20	80.40	68.00	60.80	50.80	60.40	4.80	43.60	69.20	54.52
x1-DeepSeek-R1-Distill-Qwen-7B (think)	57.60	71.40	80.20	72.40	67.80	60.40	74.00	9.80	59.00	79.80	63.24
x1-DeepSeek-R1-Distill-Llama-8B (non-think)	10.67	46.00	65.60	49.60	45.60	35.33	47.33	3.73	19.87	56.40	38.01
x1-DeepSeek-R1-Distill-Llama-8B (think)	26.40	60.53	76.53	62.80	62.67	49.73	64.93	12.93	37.20	68.00	52.17
	MT-AIME										
	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Average
Backbone Models											
Qwen3-4B (non-think)	6.67	16.67	17.78	12.22	23.33	7.78	15.56	4.44	7.78	16.67	12.89
Qwen3-4B (think)	22.22	25.56	25.56	24.44	23.33	22.22	17.78	6.67	25.56	24.44	21.78
Qwen3-14B (non-think)	16.67	26.67	25.00	25.00	21.67	18.33	21.67	5.00	13.33	20.00	19.33
Qwen3-14B (think)	26.67	26.67	32.22	25.56	34.44	24.44	34.44	18.89	23.33	45.56	29.22
Qwen3-32B (non-think)	23.33	21.67	26.67	21.67	21.67	25.00	26.67	10.00	21.67	20.00	21.83
Qwen3-32B (think)	28.89	28.89	26.67	43.33	31.11	31.11	38.89	27.78	35.56	46.67	33.89
DeepSeek-R1-Distill-Qwen-7B (non-think)	2.22	6.67	16.67	11.11	7.78	3.33	10.00	3.33	6.67	15.56	8.33
DeepSeek-R1-Distill-Qwen-7B (think)	23.33	31.67	33.33	25.00	30.00	28.33	26.67	8.33	20.00	31.67	25.83
DeepSeek-R1-Distill-Llama-8B (non-think)	1.11	2.22	1.11	4.44	3.33	1.11	3.33	0.00	4.44	5.56	2.67
DeepSeek-R1-Distill-Llama-8B (think)	11.11	16.67	15.56	18.89	17.78	12.22	15.56	5.56	11.11	20.00	14.44
x1 Series Models											
x1-Qwen3-4B (non-think)	6.67	18.89	23.33	15.56	16.67	8.89	15.56	5.56	7.78	16.67	13.56
x1-Qwen3-4B (think)	16.67	25.00	25.00	25.00	23.33	21.67	23.33	10.00	25.00	33.33	22.83
x1-Qwen3-14B (non-think)	14.44	21.11	23.33	23.33	24.44	17.78	27.78	7.78	18.89	15.56	19.44
x1-Qwen3-14B (think)	20.00	34.44	38.89	35.56	27.78	36.67	32.22	20.00	33.33	52.22	33.11
x1-Qwen3-32B (non-think)	18.89	21.11	24.44	25.56	16.67	23.33	22.22	18.89	25.56	24.44	22.11
x1-Qwen3-32B (think)	31.67	35.00	35.00	33.33	36.67	40.00	35.00	23.33	30.00	45.00	34.50
x1-DeepSeek-R1-Distill-Qwen-7B (non-think)	7.78	8.89	12.22	11.11	7.78	4.44	10.00	6.67	10.00	11.11	9.00
x1-DeepSeek-R1-Distill-Qwen-7B (think)	23.33	26.67	23.33	26.67	30.00	23.33	26.67	23.33	33.33	33.33	27.00
x1-DeepSeek-R1-Distill-Llama-8B (non-think)	2.22	3.33	3.33	2.22	2.22	3.33	2.22	0.00	2.22	7.78	2.89
x1-DeepSeek-R1-Distill-Llama-8B (think)	3.33	23.33	23.33	20.00	23.33	20.00	16.67	3.33	16.67	20.00	17.00

Table 9: Results for each subsets in *MGSM* and *MT-AIME* across different languages.