Crosslingual Reasoning through Test-Time Scaling

 1 Brown University 2 MBZUAI 3 Stanford University 4 University of Tübingen 5 Capital One 6 Cohere Labs contact.yong@brown.edu

Abstract

Reasoning capabilities of large language models are primarily studied for English, even when pretrained models are multilingual. In this work, we investigate to what extent English reasoning finetuning with long chain-of-thoughts (CoTs) can generalize across languages. First, we find that scaling up inference compute for English-centric reasoning language models (RLMs) improves multilingual mathematical reasoning across many languages including low-resource languages, to an extent where they outperform models twice their size. Second, we reveal that while English-centric RLM's CoTs are naturally predominantly English, they consistently follow a *quote-and-think* pattern to reason about quoted non-English inputs. Third, we discover an effective strategy to control the language of long CoT reasoning, and we observe that models reason better and more efficiently in high-resource languages. Finally, we observe poor out-of-domain reasoning generalization, in particular from STEM to cultural commonsense knowledge, even for English. Overall, we demonstrate the potentials, study the mechanisms and outline the limitations of crosslingual generalization of English reasoning test-time scaling. We conclude that practitioners should let English-centric RLMs reason in high-resource languages, while further work is needed to improve reasoning in low-resource languages and out-of-domain contexts.

1 Introduction

Scaling up compute at test-time can maximize model performance and output quality [1, 2, 3, 4], but it has been understudied in multilingual settings. In particular, reasoning language models (RLMs), such as Deepseek's r1 [5] and OpenAI's o1 or o3 models [6, 7], strongly benefit from added inference compute to their long chain-of-thoughts (long CoTs) [8]. However, this advantage has primarily been explored in English contexts, such as in recent work that combined small-scale reasoning finetuning with scaled up thinking tokens at test time [9, 10]. State-of-the-art RLMs rely on reasoning training data that contain long CoTs, which is currently most available for English [11]. Thus, these RLMs are English-centric [9, 12, 13, 14, 15, 11]. Given that their base models are often multilingual models such as Qwen models [16], does reasoning finetuning in English give them multilingual reasoning abilities?

In this work, we investigate how much test-time compute can improve multilingual reasoning abilities of English-centric RLMs. In particular, our research questions are as follows:

^{*}Core contributors. See Appendix A for further details.

- RQ1. **Crosslingual test-time scaling**: How effective is test-time scaling of English-centric RLMs on multilingual reasoning tasks? (Section 4)
- RQ2. **Language-mixing behaviors**: What kind of language-mixing patterns do English-centric RLMs exhibit when they interact with non-English prompts? (Section 5)
- RQ3. **Language forcing**: How well do English-centric RLMs perform when being forced to think in non-English languages? (Section 6)
- RQ4. **Cross-domain generalization:** How well does crosslingual reasoning generalize beyond the original STEM domain, such as humanities and social sciences? (Section 7)

We experiment with s1 models [9] as our English-centric RLMs for crosslingual generalization study. They are multilingual Qwen2.5-Instruct models [16] supervised finetuned on 1k training samples of English STEM reasoning tasks and achieve state-of-the-art performance on English math reasoning benchmarks [9]. Our most significant contributions are as follows: ²

- 1. We provide evidence that larger models benefit from crosslingual test-time scaling, which *contrasts* with contemporary work [17] that reports early negative findings based on 1.5B models. Crosslingual test-time scaling is not only effective for both high-resource and low-resource languages, but it also allows an RLM to outperform models twice its size on multilingual math reasoning tasks.
- 2. We report a dominant language-mixing pattern where RLMs quote non-English phrases related to the question prompts in quotation marks in the thinking process. This *quote-and-think* pattern suggests that model's multilingual capability to parse and understand questions enables crosslingual generalization of English reasoning finetuning.
- 3. We discover an effective strategy to control the reasoning language of RLMs, and we find that forcing RLMs to think in high-resource languages yields substantially better reasoning performance than in low-resource languages. Furthermore, the long CoTs for high-resource languages are more token-efficient at test time.
- 4. We show that while reasoning finetuning may benefit tasks in certain domains, we *do not* observe consistent benefits of scaling up thinking tokens for non-STEM tasks that are outside of the STEM domain of reasoning finetuning data. In domains requiring cultural commonsense knowledge, test-time scaling can even hurt performance.

Our work shows that test-time scaling of English-centric RLMs can serve as a strong multilingual reasoning baseline. Furthermore, for English-centric reasoning finetuning, it is advisable to finetune models through data-efficient training (such as with s1's data [9]) to minimize catastrophic forgetting of multilingual capability. We recommend letting the English-centric RLMs reason in high-resource languages such as English and Chinese for optimal performance and inference-compute efficiency. Future work is needed for enabling RLMs to generalize to out-of-domain tasks or to reason in low-resource languages.

2 Background and Related Work

Reasoning language models (RLMs) Recent advancements of reasoning language models (RLMs) such as OpenAI-o1 [6, 7] and DeepSeek-R1 [5] builds on LLMs' capability to perform intermediate reasoning steps, which is commonly referred to as chain-of-thought reasoning [18]. Prior work demonstrates that these intermediate computation steps can significantly improve the correctness for final answer outputs [18, 19, 20, 21, 22]. Furthermore, extending the lengths of these computation steps, thereby creating long chain-of-thoughts (long CoTs), can allow the model to backtrack on incorrect reasoning steps and self-correct its final answer [8, 23, 5, 12, 24]. In our work, we focus on RLMs with long CoTs capability, which is an emerging research area. These models are created through *distilling* long English-only reasoning chains from larger RLMs [25, 10, 26, 27, 28] to finetune multilingual pretrained models like Llama [29] or Qwen models [16]. There is a limited understanding how pretrained models' multilingual capability enables crosslingual reasoning of long CoTs, which our work aims to address.

 $^{^2}$ We release our code and artifacts at https://github.com/BatsResearch/crosslingual-test-time-scaling.

Test-time scaling and s1 Test-time scaling is a new scaling paradigm where more computation budget is allocated for LLMs at inference time before committing to an answer [1]. One example of test-time scaling is scaling up model generation length, particularly for RLMs with long CoTs capability [1, 30, 6, 5]. In particular, the s1 work [9] demonstrates the effectiveness of a simple test-time scaling recipe: reasoning finetuning on small amount of training data with long CoTs (specifically 1k samples distilled from larger RLMs such as DeepSeek-R1) and scaling up inference budget at test time. Through test-time scaling of a 32B-parameter model, the authors achieve the state-of-the-art mathematical reasoning performance, and their models even rival industry-grade RLMs such as o1-mini [6]. Nonetheless, similar to aforementioned RLMs literature, exploration of test-time scaling paradigm mostly evaluate on English math benchmarks [1, 11, 4, 15, 3, 9]. Here, our work focuses on understanding how effective test-time scaling of English-centric RLMs, specifically s1 models, in multilingual settings for various domains.

Multilingual reasoning Multilingual reasoning encompasses the ability of language models to perform complex reasoning tasks across different languages. Early work has demonstrated that chain-of-thought prompting in English can significantly improve performance on multilingual mathematical reasoning tasks [31], which suggests that LLMs might rely on dominant languages like English as a pivot language for complex reasoning. Follow-up work explores several strategies such as translating the multilingual queries to English [32, 33, 34], aligning of latent representation spaces [35, 36] and reasoning outputs [37, 38] across languages, or expanding language coverage of reasoning training data [39]. Our work focuses on understanding how controlling the length of long CoTs and their reasoning language at test time affects multilingual reasoning. One similar work [17] experimented with controlling generation lengths of finetuned DeepSeek-R1-1.5B [5] but reported *negative* results: increasing thinking tokens leads to minimal performance gains for mathematical reasoning in non-English languages. We believe that their negative findings are due to constrained model parameters, as we show that larger models *can benefit* from crosslingual test-time scaling.

3 Experimental Setup

Models We use s1 models [9] as our English-centric RLMs. In particular, we work with the s1.1 variants, which are multilingual Qwen2.5-Instruct models finetuned on 1k English-only reasoning data generated by DeepSeek-R1. We choose s1 models for two reasons: (1) their English mathematical reasoning capability reaches state-of-the-art performance, and (2) the models and the training data are fully open-sourced. We experiment with s1 models at different scales—specifically with 1.5B, 3B, 7B, 14B, and 32B parameters.

Budget forcing Budget forcing refers to techniques for controlling inference budget for long CoTs [9], which can be done in two ways: (1) **truncation**, which *cuts off* long CoTs after they reach maximum thinking tokens, or (2) **extrapolation**, which *adds* tokens such as "Wait" at the end of CoTs to force the model continue reasoning. In our extrapolation setup, we experiment with adding "Wait" only once since we do not observe significant performance gains from lengthening CoTs.

Evaluation data For research questions RQ1, RQ2, and RQ3, we use the **Multilingual Grade School Math (MGSM)** benchmark [31]. It contains 250 grade-school math problems from the GSM8K dataset that have been manually translated into ten languages—Bengali (bn), German (de), Spanish (es), French (fr), Japanese (ja), Russian (ru), Swahili (sw), Telugu (te), Thai (th), and Mandarin Chinese (zh). We check for 8-gram overlap between s1's training samples and MGSM (including the English subset) following standard practice [40], and we observe no train—test overlap.

For RQ4, addressed in section 7, we evaluate on three different cross-domain benchmarks:

- 1. **Global-MMLU** [41]: a multilingual translated version of the original MMLU dataset [42] that spans different subject topics, including STEM, business, humanities, medical, social sciences, etc. Additionally, it also contains annotations that identify whether the question is culturally agnostic or specific, allowing for more fine-grained analysis.
- 2. **Food ORiented cultural commonsense Knowledge (FORK)** [43]: a manually-curated set of commonsenseQA-style [44] questions in English for probing cultural biases and assumptions on food-related customs around the world.

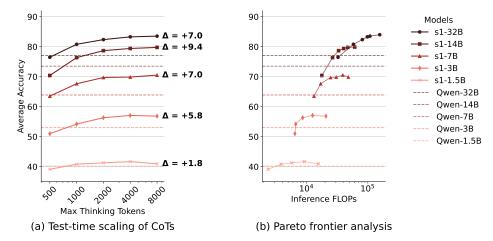


Figure 1: Crosslingual test-time scaling of s1 and Qwen models on the MGSM benchmark (*excluding English*) across different model sizes. In subfigure (a) we enforce a hard limit of maximum thinking token, and in (b) we measure their inference FLOP compute for a Pareto frontier analysis. Δ measures the absolute difference between average accuracy scores at 0.5k and 8k maximum thinking tokens. Dash lines indicate the best few-shot prompting baseline performance of Qwen.

3. **COPAL-ID** [45]: a manually-curated datasets following COPA's causal reasoning format [46] in Indonesian languages to evaluate causal reasoning with Indonesian cultural nuances.

We use the lm-evaluation-harness library [47] as the main evaluation framework.³. We evaluate with greedy decoding and report task accuracy, which is equivalent to pass@1.

Baselines Qwen2.5-Instruct models, which are the base models of s1, are our main baseline models. For MGSM, we apply two prompting strategies. The first is zero-shot prompting, and the second is few-shot CoT prompting using the provided eight CoT examples in MGSM train set. For the latter, we follow [31] and experiment with prompting with few-shot CoT in English (EN-CoT) and in the same language as input prompt (native-CoT). In addition, for MGSM, we benchmark against prior state-of-the-art models [48, 37, 49, 35, 50] and comparable English-centric RLMs such as R1-distilled-Qwen models [5] open-sourced by DeepSeek. For cross-domain benchmarks, we simply compare against zero-shot prompting of Qwen2.5-Instruct.

4 Crosslingual Test-Time Scaling

In RQ1, we explore test-time scaling in a *zero-shot crosslingual setting*, where English-centric reasoning models are applied to math problems in different languages. These models were finetuned from multilingual pretrained models in the same domain, therefore carrying high potential for crosslingual transfer.

4.1 Effectiveness of Crosslingual Test-Time Scaling

Crosslingual generalization of reasoning training and test-time scaling We report two main obsevations from Figure 1 (a). First, we observe that s1 outperforms Qwen's few-shot prompting baseline across languages in MGSM (excluding English) when given high inference thinking budget. Second, crosslingual test-time scaling is effective for models with 3B parameters and above, with s1 at 14B size having the largest accuracy gain of $+\Delta9.4\%$ when maximum inference budget increases from 0.5k to 8k thinking tokens. We want to highlight that *sufficient model capacity* is necessary for effective crosslingual test-time scaling, as test-time scaling only yields minimal benefits at 1.5B

³For cross-domain benchmarks, we use the configuration <code>generate_until</code> tasks instead of the default <code>multiple_choices</code> setting for cross-domain benchmarks and extract answers from generations using GPT-4o-mini (see Appendix E.1). For Global-MMLU, we use the <code>default</code> subset of the benchmark and pick 7 languages that overlap with the MGSM benchmark.

Table 1: MGSM performance comparison against 14B-sized s1 model with maximum 8k thinking tokens. We report the language-breakdown accuracy from cited papers if available; otherwise, we reproduce using their open-sourced models without any inference budget constraint. We report the average length of the generations (avg. len) and the relative accuracy difference (green text) between s1-14B under extrapolation budget forcing and its baseline Qwen2.5-14B-Instruct. We **bold** both s1 performance and baseline models that outperform s1.

Models	avg len	bn	de	en	es	fr	ja	ru	sw	te	th	zh	AVG
Qwen2.5-14B-Instruct [16]	413.1	74.0	77.6	82.0	77.6	67.6	70.4	76.4	40.4	50.8	78.8	84.0	70.9
+ 8-Shot EN-CoT [31]	316.5	77.2	75.2	87.6	86.0	68.4	76.8	76.4	45.6	52.0	79.2	84.4	73.5
+ 8-Shot Native-CoT [31]	365.2	79.2	77.2	88.0	87.2	68.4	76.0	75.6	46.8	53.2	80.4	83.6	74.1
s1-14B (truncation)	1912.9	82.0	84.8	92.8	88.4	85.2	83.6	86.8	55.6	59.6	85.2	86.4	80.9
s1-14B (extrapolation)	2352.3	82.8	86.8	92.4	86.4	83.2	83.2	88.8	57.2	58.0	84.8	87.6	81.0
Relative accuracy difference (%)		+11.9%	+11.9%	+12.7%	+11.3%	+23.1%	+18.2%	+16.2%	+41.6%	+14.2%	+7.6%	+4.3%	+14.2%
MetaMath-13B [52]	529.8	6.8	64.4	70.4	63.6	65.2	47.6	60.0	11.6	0.8	4.8	50.8	40.5
MetaMathOctopus-13B [37]	545.8	41.6	60.1	66.8	61.1	60.8	57.3	59.1	50.9	3.6	52.1	53.1	51.5
MAPO-DPO-13B [37]	552.4	54.7	69.5	70.5	70.6	71.3	69.0	68.2	62.9	4.0	64.7	68.2	61.2
SLAM-13B [49]	101.5	45.6	62.8	71.2	67.6	65.2	54.0	64.4	46.4	2.4	47.6	58.8	53.3
MetaMath-LB-15B [35]	93.2	50.0	63.6	67.6	63.2	61.6	42.0	60.0	41.6	36.4	52.8	48.0	53.5
MetaMath-LB-20B [35]	93.1	52.8	64.0	66.4	60.4	64.0	45.2	58.8	49.2	47.2	53.6	52.4	55.8
R1-Distill-Qwen-14B [5]	1030.7	66.0	77.2	83.6	80.4	74.4	78.4	82.4	22.4	22.4	74.8	79.6	67.4
R1-Distill-Qwen-32B [5]	1353.8	77.6	82.8	85.2	85.6	79.6	83.2	84.8	38.0	14.4	82.4	85.6	72.7
Gemma-3-12B-it [50]	238.2	55.6	74.4	83.2	81.2	64.8	74.0	74.8	71.2	73.2	78.4	79.2	73.6
Gemma-3-27B-it [50]	461.7	64.8	83.2	88.4	84.0	72.4	79.2	83.2	78.0	76.0	84.4	84.4	79.8
Qwen3-14B [50]	1575.2	85.2	83.6	94.8	88.4	87.2	77.6	94.0	63.6	80.0	86.8	85.2	84.2

size. The model only experiences $+\Delta1.8\%$ accuracy gain with performance peak at 4000 maximum thinking tokens.

Our findings contrast the conclusion drawn by recent work [17] that "test-time scaling may not generalize as effectively to multilingual tasks." Our results suggest that the limitation observed by [17] is due to their usage of 1.5B models in their experiments, instead of an inherent weakness of test-time scaling. Furthermore, [17] posit that multilingual generalization of test-time scaling would occur for "significantly larger" models with at least 70B parameters, but we discover a substantially smaller parameter threshold at 3B parameters, above which models consistently benefit from crosslingual test-time scaling.

Pareto frontier analysis Figure 1 (b) illustrates the performance-efficiency trade-off across different sizes of the s1 model family on MGSM. We follow prior test-time scaling work [1, 51] and compute the inference cost using the approximation FLOPs = $2ND_{\rm inference}$ where N represents model parameters and $D_{\rm inference}$ the total number of tokens generated at inference time, and we average across different languages. The figure further demonstrates the model capacity constraint on test-time scaling: The 32B and 14B models can reach accuracy levels that are simply unattainable for the smaller models, no matter how much test-time compute is allocated to those smaller models. Even at their maximum performance ceiling, the 3B and 7B models cannot match what the larger models achieve with moderate computational resources. While we observe accuracy-to-computation tradeoffs (i.e., better performance comes with using larger models and higher test-time compute), the 14B model offers a compelling compromise by achieving above 80% accuracy with substantially lower inference FLOPs than the 32B model, representing a "sweet spot" on the Pareto frontier for practical applications.

4.2 Performance Comparison on MGSM Benchmark

Comparison against Qwen baselines. Table 1 shows that with cross-lingual test-time scaling, s1 gains substantial accuracy increase as compared to different baselines with Qwen2.5 models. Furthermore, crosslingual test-time scaling benefits both high-resource and low-resource languages. For instance, fr receives a significant $+\Delta23.1\%$ relative accuracy increase, whereas sw—the worst-performing language for the base model Qwen—receives $+\Delta41.6\%$ relative accuracy improvement. Lastly, we observe similar performance for both truncation and extrapolation budget forcing strategies. This is because s1 models are already generating extensive reasoning chains so further lengthening the CoTs have minimal benefits. Similar trends of language-specific improvements are also observed in other s1 model sizes (Appendix B).

Box 1. Example of *quote-and-think* language-mixing pattern demonstrated by s1 where non-English phrases are mentioned in quotation marks and used for reasoning.

Question: ローブを作成するには、青色の繊維を2巻分、白色の繊維をその半分用いる必要があります。全体で何巻必要ですか?

(To make a robe, you need 2 rolls of blue fiber and half of the white fiber. How many rolls are needed total?)

s1's Thinking: Okay, let's see. ... The problem says "白色の繊維をその半分用いる" which is white fibers the half amount. "その" refers to the blue fibers, right? So half of the blue fibers, which is 2 rolls, so half is 1. So 2 + 1 is 3. Yeah, that seems right. Wait, sometimes these problems can be tricks, but I don't think so. If it had said "白色の繊維を2巻分とその半分" that would be different, but no. It's "青色の繊維を2巻分、白色の繊維をその半分". So white is half of blue. So calculation is straightforward. Blue 2, white 1, total 3. ...

Comparison against state-of-the-art models. Table 1 shows that crosslingual test-time scaling of s1 models can serve as a strong multilingual baseline for MGSM, as it outperforms all prior state-of-the-art models that involve finetuning on multilingual data such as MetaMath, MAPO, and [52, 37, 49, 35]. We believe this is because these prior studies use Llama as their base models, which generate significantly shorter reasoning traces and lack sophisticated reasoning behaviors such as verification and backtracking compared to Qwen models [23].

Surprisingly, 14B-sized s1 can even outperform recent state-of-the-art reasoning models twice its size, namely DeepSeek's R1-Distill-Qwen-32B [5] and Google's Gemma-3-27B-it [50]. We observe that R1-Distill-Qwen has substantially poorer performance on sw and te, suggesting that their 800k samples of English and Chinese training data [5] leads to *catastrophic forgetting* of lower-resource languages. In contrast, s1 is only trained with 1k English samples for only 5 epochs [9], which leads to minimal forgetting and better crosslingual generalization. While the multilingual Gemma-3 models outperform s1 on low-resource languages, probably due to these languages being incorporated during reasoning finetuning, its performance gap against s1 on high-resource languages may be attributed to the shorter reasoning thinking time. Qwen3 [53] is the most performant model, probably due to its long reasoning capability and extensive multilingual training data.

5 Language-Mixing Behaviors

We notice that s1 models can mix languages in their CoT reasoning under inference scaling. Given a lack of systematic study of language mixing behaviors in crosslingual reasoning in prior work [5, 11], we present an analysis of linguistic behaviors of s1 in multilingual math reasoning tasks to address RQ2.

In particular, we are interested in two sub-questions of RQ2. First, how does the *dominant lan-guage*—the primary language of generated response, also known as *matrix language*—change after English-centric reasoning finetuning? Second, what are the language-mixing patterns exhibited by s1 in its reasoning outputs? We focus on four languages, namely ja, ru, th, and zh, and we refer our readers to Appendix C.1 for detailed methodology.

5.1 Dominant Language in Model Outputs

Figure 2 shows that s1 and its base model Qwen use entirely different dominant languages in their overall responses in multilingual settings. Qwen models generate responses in the same language as the questions, especially for zero-shot and native CoT prompting settings. Surprisingly, even in crosslingual input settings where the 8-shot CoT reasoning samples are provided in English, Qwen models *still stick to* generating outputs in the same language as the question. In contrast, s1 generates reasoning and final answers with English as the dominant language, which suggests that finetuning on only 1k English reasoning data can sufficiently change model behavior from multilingual to English-dominant. Similar behaviors are observed for models at smaller parameter sizes (Figure 8).

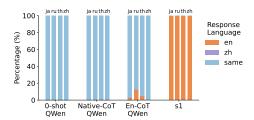


Figure 2: Proportion of dominant languages in models' entire responses when queried with multilingual math questions. "same" indicates that the response language is the same as query language.

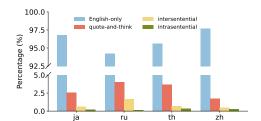


Figure 3: Breakdown of language-mixing patterns in s1's reasoning. Percentage indicates the probability of a sentence being English only, quoting non-English phrases (quote-and-think), entirely being in a different language (intersentential), or mixing different languages within the same sentence (intrasentential).

5.2 Language-Mixing Patterns During Reasoning

Dominant mixing pattern Figure 3 shows that s1's reasoning is predominantly in English for all four languages of study. Specifically, at least 92.5% of the sentences in s1's CoTs are in English only and do not mix languages. However, in the remaining cases when s1 mixes languages during reasoning, it primarily follows a sophisticated pattern to which we refer as **quote-and-think**. Particularly, s1 will first quote certain words or phrases, often from the input question, and then interpret their meanings and implications during its thinking process. This is demonstrated by the quoted phrase "白色の繊維をその半分用いる" and s1's literal translation "white fibers the half amount" in Box 1. In linguistics, this type of language-mixing is known as foreign-language quotation [54]. This language-mixing behavior happens due to crosslingual generalization of the quoting-and-thinking reasoning characteristic in s1's English finetuning data (see Appendix C.4).

We want to emphasize that the *quote-and-think pattern goes beyond simple translation*. As demonstrated in Box 1, s1 builds upon the extracted phrase and synthesizes a new multilingual setting where if the question had asked "白色の繊維を2巻分とその半分" (two and a half rolls of white fiber) it would have arrived at a different answer. Here, the model shows an understanding of how the syntactic structure in Japanese affects the semantic meaning of the math problem, which suggests that s1 is genuinely parsing and reasoning about the mathematical relationships expressed in Japanese and not merely translating the content to English before processing. This suggests that the multilingual capability of the base models is preserved for natural language understanding and allows s1 to reason about what it has understood about the question.

Other language-mixing patterns Figure 3 shows that, compared to other languages, Russian exhibits *intersentential language-mixing*—mixing of sentences of different languages—the most, where English sentences are switched into Russian sentences and then back to English in the middle of the reasoning. This is undesirable [5] because it can create confusion for users expecting a coherent response in a single language and thus impacting readability. For *intrasentential language-mixing*—mixing of words of different languages in a single sentence—we perform a deeper analysis (see Appendix C.5) and report that at least 70% of cases resemble quote-and-think as the non-English phrases are extracted from the question prompt, but no quotation marks were provided. This further confirms that quote-and-think is the dominant pattern of language mixing in s1's reasoning process, even when not explicitly marked, and represents a natural strategy the model has developed to handle multilingual mathematical reasoning.

6 Language Forcing

When a multilingual user interacts with LLMs, it is natural to expect the LLMs to respond in the language consistent with the user's query. Therefore, in RQ3, we are interested in understanding if

Table 2: Performance comparison of different language forcing strategies where we force s1 to reason in the same language as the question with 8k maximum thinking tokens. Languages are categorized into high-resource (HRL: de, en, es, fr, ru, ja, zh) and low-resource (LRL: bn, sw, te, th) groups. We use \uparrow or \downarrow to indicate if the language forcing method outperforms or underperforms the baseline of thinking in English.

Model	Method	ALL	Accuracy HRL	LRL	Langu ALL	age Con HRL	npliance LRL
s1-32B	Baseline	87.4	91.2	80.8	0.0	0.0	0.0
	translated_wait	87.8 ↑	91.3 ↑	81.7 ↑	19.4	30.4	0.2
	prefix	85.6 ↓	91.3 ↑	75.7 ↓	96.2	99.4	90.7
	system	86.0 ↓	90.3↓	79.3 ↓	47.7	71.4	6.2
	combined	84.3 ↓	90.5↓	73.4 ↓	98.6	99.7	96.6
s1-14B	Baseline	84.4	90.3	73.9	0.1	0.2	0.0
	translated_wait	85.0 ↑	90.6 ↑	75.2 ↑	25.4	26.5	0.3
	prefix	82.0 ↓	90.7 ↑	66.9 ↓	96.4	99.6	89.2
	system	83.8 ↓	89.5 ↓	73.8 ↓	35.4	53.0	4.6
	combined	80.4 ↓	89.0 ↓	65.4 ↓	97.5	99.9	93.1
s1-7B	Baseline translated_wait prefix system combined	74.8 74.1 \(\) 69.5 \(\) 73.1 \(\) 69.9 \(\)	86.9 85.8 ↓ 84.0 ↓ 84.8 ↓ 84.5 ↓	53.6 53.6 44.2 ↓ 52.7 ↓ 44.4 ↓	0.0 27.1 89.8 70.5 96.0	0.0 41.0 96.9 96.6 98.0	0.0 2.7 77.3 24.9 92.6

we can perform *language forcing*—controlling an English-centric RLM to generate reasoning in a particular language—and if the difference in reasoning language affects performance.

6.1 Methodology: Language Forcing Techniques

We experiment with the following language forcing techniques to control s1's reasoning language:

- Translated Wait (translated_wait): Building upon extrapolation budget forcing strategy that explicitly extends reasoning traces [9], once the model finishes English reasoning, we append a translated "Wait" token as an intervention strategy to force the model to switch language and continue reasoning in our chosen language.
- **Prefix** (prefix): We appended a prefix string translation-equivalent of "Okay, let me try to figure this out." at the beginning of the reasoning generation in order to guide the model's generation in our chosen language. We also apply the translated_wait strategy and append the translated "Wait" token.
- **System Prompt** (system): We use a system prompt to control the language use in model generation. Specifically, we translate the system prompt "You are a helpful assistant." into our chosen language and add the translation-equivalent of the instruction "You must think and answer only in {language}".
- **Combined** (combined): This method uses all the techniques above to maximize control over the model's reasoning language.

We experiment with language forcing on the MGSM benchmark with K=11 languages. Given a question in $k\in K$ language (i.e., query language), we force the model to reason in $k'\in K$ language (i.e., reasoning language). We compute two evaluation metrics: task accuracy and language compliance. The former follows the setup described in Section 3, whereas the latter measures the proportion of tokens generated in the intended reasoning language relative to tokens produced in other languages. In other words, language compliance quantifies the effectiveness of s1 in following explicit language constraints. To only measure the effect of reasoning language and eliminate the confounding factors such as introduction of prefix string, our baseline is to let s1 models think in its dominant language English using the combined setup.

⁴We remove the part of "You are Qwen, created by Alibaba Cloud." because English proper nouns like 'Qwen' and 'Alibaba' do not have translation equivalents in many non-English languages.

Table 3: Performance scores across different reasoning languages given query language. We use 11 color codes to *rank each row* to highlight the high- (blue) and low-performing (red) reasoning language given a query language. We also **bold** the best-performing reasoning language. Lastly, we use \(\sigma \) to indicate the average accuracy when the reasoning language is the same as query language (i.e., average of the diagonals).

Query					Reason	ing La	nguage	;				Range
Language	bn	de	en	es	fr	ja	ru	\mathbf{sw}	te	th	zh	(max - min)
bn	79.2	85.2	86.8	84.4	81.6	81.2	83.6	62.4	75.6	80.8	81.2	24.4
de	88.4	89.2	90.4	88.8	90.8	90.0	87.6	75.6	78.4	88.0	89.6	15.2
en	93.2	94.4	94.4	95.2	94.8	94.4	93.2	84.0	84.0	94.8	96.8	12.8
es	86.4	92.4	93.6	93.6	92.4	90.8	93.2	76.6	82.8	90.0	90.8	17.0
fr	87.2	87.2	88.4	87.2	88.0	89.6	88.4	72.8	77.6	87.2	88.0	16.8
ja	79.2	84.8	83.6	81.6	85.6	82.0	84.8	71.6	74.0	85.6	83.6	14.0
ru	89.2	91.2	92.4	89.6	93.6	92.0	92.4	77.6	80.8	90.0	91.2	16.0
sw	45.6	58.8	59.6	55.2	55.6	47.6	48.4	44.4	32.4	45.2	52.0	27.2
te	53.2	56.4	60.0	56.4	60.0	57.2	55.2	34.8	54.4	53.6	52.8	25.2
th	80.8	88.4	89.2	88.4	91.2	87.2	87.2	66.4	69.2	86.4	88.8	24.8
zh	85.2	86.8	89.6	87.2	86.8	88.8	90.8	73.6	77.2	86.0	89.2	17.2
AVG	78.9	83.2	84.4	82.6	83.7	81.9	82.3	67.3	71.5	80.7	82.2	≥81.2

6.2 In-Language Forcing

In-language forcing refers to the setting where the reasoning language matches the query language, i.e., k' = k. Here, we focus on analyzing if we should perform in-language forcing for English-centric RLMs or let models reason in their natural reasoning language (i.e., English for s1). We showcase an example of s1 thinking in Japanese due to language forcing in Appendix D.1.

Reasoning in high-resource languages As shown in Table 2, reasoning in high-resource languages (HRLs) achieve similar scores—typically within a 1-2 points difference compared to baseline—regardless of language forcing strategy. One possible explanation is that these languages are highly represented within pretraining of Qwen, which allows for effective crosslingual transfer from English reasoning finetuning.

Reasoning in low-resource languages Table 2 highlights that, in contrast to HRLs, for low-resource languages (LRLs), combined strategy substantially underperforms the baseline. This suggests that English reasoning finetuning does not transfer to LRLs as effectively, and thus s1 is relatively incapable of reasoning in LRLs. Strategies that permit a mix of English and the target language—particularly translated_wait and system—outperform strict in-language approaches such as combined. This is because s1 can lean on its English reasoning ability while incorporating the target language, which appears to be beneficial in scenarios with limited language-specific training data. For example, in (te) and (sw), the 14B and 32B models using translated_wait consistently outperform those using prefix or combined strategies by 4–5 accuracy points (see Appendix D.3 for detailed results).

Trade-off between accuracy and language compliance Table 2 shows that we can successfully control the reasoning language of RLMs in their long CoTs, but that usually come at a cost of task performance. For instance, for both 32B- and 14B-sized s1 models, the combined strategy achieves near 100% language compliance for both HRLs and LRLs, but it yields the worst accuracy compared to other language forcing strategies. For 7B-sized models, we observe that we need a combination of all strategies (i.e., combined) to achieve high language compliance. This suggests that it is non-trivial to control the reasoning language, as different choices of language forcing strategy can significantly affect how consistently s1 adhere to our specified reasoning language.

6.3 Crosslingual Language Forcing

Given our success in controlling the reasoning language of s1 with combined strategy, we explore if there is a particular language that is best served as reasoning language for s1. Particular, for each

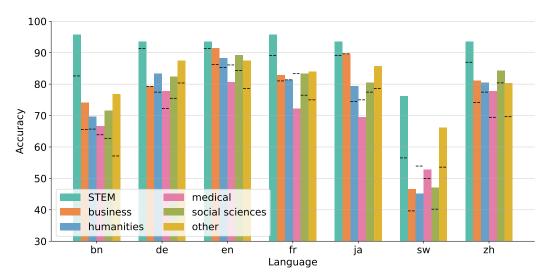


Figure 4: Language and domain breakdown for Global-MMLU benchmark. Dashed lines indicate the performance of zero-shot prompting of Qwen-32B-Instruct models.

query language $m \in M$, we force the model to all M possible languages, resulting in an exhaustive $M \times M$ query-reasoning language-pair analysis.⁵

Performance comparison of reasoning languages Table 3 shows that reasoning in HRLs such as en, fr, or de yield similarly high performance (the accuracy difference is within 1 to 2 points), with English being the most performant reasoning language and French being the close second. We discover two surprising findings: first, even though the Qwen2.5 base model is highly pretrained in Chinese [16], it is not necessarily the best reasoning language, even when the question is asked in zh; second, neither reasoning in en nor in query language necessarily yields the best performance—quite the contrary, even reasoning in languages that are usually less represented in pretraining data [55] such as ru and th can achieve the best performance for query languages in other families such as ja. Lastly, we observe that languages that are considered as slightly less-resourced [55] such as th and bn still achieve nearly 80% overall accuracy, but further lower-resourced languages such as sw or te result in substantially lower overall accuracy.

Inference cost analysis Our analysis of inference costs across reasoning languages in Figure 5 reveals a significant negative correlation (-0.811) between token count and mathematical problem-solving accuracy. Reasoning in LRLs not only underperform their HRL counterpart (with accuracy below 80%), but they also demand substantially more computational resources at test-time—often exceeding 3,000 tokens in their long CoTs for bn, sw, and te. This token efficiency disparity at test-time translates directly to computational costs; for instance, reasoning in Swahili requires approximately 3.5 times more compute than French for the same tasks. This is very likely due to the well-known tokenization disparity for LRLs, where tokenized lengths for an equivalent sentence in different languages are much larger for LRLs [56, 57].

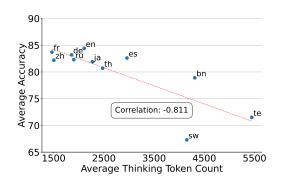


Figure 5: MGSM accuracy against number of thinking tokens in s1 models' outputs in different reasoning languages.

⁵This analysis is is computationally heavy, so we only focus on 14B-sized s1 models.

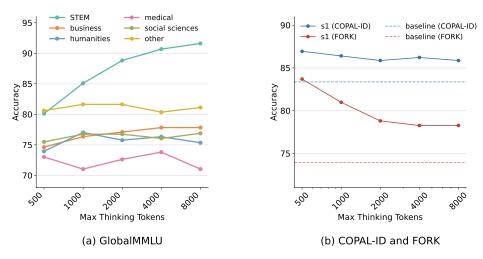


Figure 6: Effects of thinking time for s1 models on different domains of Global-MMLU benchmark (subfigure (a)) and cultural commonsense knowledge (FORK) and reasoning (COPAL-ID) benchmarks (subfigure (b)). Similar to Figure 4, we added dashed lines as zero-shot prompting of Qwen-32B as baselines in (b).

Choice of query language Table 3 sheds light on whether we should translate inputs into HRLs such as English for reasoning tasks, which has proven to be an effective strategy [32, 33]. Our results are consistent with prior work: merely translating the question from Swahili to French can boost the accuracy from 59.6 to 90.8 even when the model reasons in French—a language that s1 is not trained to reason in. Besides, based on the range column, which measures difference between the best and worst reasoning languages for a particular query language, the model is less sensitive to query language in HRLs than in LRLs as exhibited by the smaller range. In other words, querying s1 in HRLs increases the model's consistency in achieving the same correct answer with different reasoning languages.

7 Cross-Domain Generalization

Since s1 models obtain strong crosslingual math performance with English-only training, a natural question to ask is whether such generalization extends to other non-math domains that may require knowledge recall or cultural reasoning. We address this research question RQ4 using the cross-domain benchmarks described in Section 3.

7.1 In-domain generalization

STEM domains Figure 6 (a) shows that performance on STEM subject domain in Global-MMLU improves with test-time scaling of thinking tokens (cyan line). Furthermore, s1 outperforms its base model Qwen2.5-32B-Instruct (black dashed line for cyan blocks in Figure 4) by a large margin, especially for bn and sw where both gain $+\Delta19.8\%$ ($56.5\% \rightarrow 76.1\%$) and $+\Delta13.1\%$ ($82.6\% \rightarrow 95.7\%$) accuracy respectively. Given that s1 training data includes OlympicArena dataset [58] that encompasses various STEM subject knowledge such as biology and astronomy, our results here further corroborates Section 4 findings on strong crosslingual in-domain generalization.

7.2 Out-of-domain generalization

non-STEM domains We report two main observations. First, Figure 4 shows that the benefits of reasoning finetuning are domain-specific. For instance, business and social sciences domains gain slight improvements, but the medical domain experiences up to nearly 10% decrease in accuracy even for HRLs such as en and fr compared to the baseline Qwen model. Second, we report **minimal cross-domain generalization of test-time scaling** from Figure 6 (a). Domains such as medicine do not benefit from scaling up thinking tokens, as increasing maximum thinking tokens from 0.5k

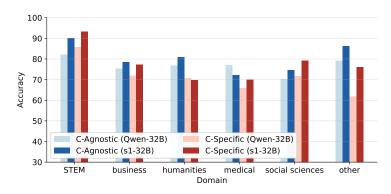


Figure 7: Performance comparison between s1 and zero-shot prompting of Qwen baseline for culturally-agnostic (C-Agnostic) and culturally-specific (C-Specific) questions in different domains. Results are average across languages.

to 4k tokens merely improves accuracy by only $+\Delta0.8\%$ ($73.0\% \rightarrow 73.8\%$), and further scaling to 8000 thinking tokens even reduces accuracy by $-\Delta2.0\%$ ($73.0\% \rightarrow 71.0\%$). Out of all non-STEM domains, business benefits the most from test-time scaling ($+\Delta3.2\%$), but the accuracy gain still lags behind STEM domain ($+\Delta11.5\%$) by a huge margin.

Cultural-specific knowledge and reasoning For cultural commonsense knowledge and reasoning benchmarks, we observe similar findings that there is minimal benefits of test-time scaling of s1. Figure 6 (b) shows that while reasoning finetuning improves overall model performance over Qwen baselines (dashed lines), scaling up test-time thinking compute does not improve performance. In fact, for the English FORK benchmark, increasing thinking tokens leads to substantially poorer performance. This is also known as *overthinking* [59] where reasoning models expend excessive compute in their long CoTs and lead to worse performance [60, 61, 62].

As for Global MMLU, which contains additional annotations for whether questions belong to culture-specific or culture-agnostic types, Figure 7 shows that there is no obvious pattern explaining which type benefits the most from the s1-training. For the STEM category, the improvements are almost equivalent. However, in other domains, we observe inconsistent performance change. For instance, we observe that performance drops in the humanities subject mostly belong to the culture-specific type while the opposite for the medical domain. We also experiment with language forcing (using the combined strategy from Section 6) to see if reasoning in the same language as the question helps with performance, but we observe minimal and often inconsistent performance gains from language forcing (see Appendix E.2 for more details).

8 Discussion and Future Work

Data-efficient English reasoning finetuning One notable finding from our work is that s1 remains capable of generating text in different languages and experiences minimal catastrophic forgetting—a phenomenon where the model loses its ability to generate fluent text in other languages after language-specific supervised finetuning [63, 64]. In contrast, R1-Distill-Qwen baseline experiences significant catastrophic forgetting for low-resource languages. This suggests that data-efficient finetuning with a small number of reasoning finetuning steps (s1 is only trained with 1k English samples for 5 epochs) is advisable for English-centric reasoning finetuning to preserve multilingual capability.

Reasoning finetuning with multilingual data Given the limitations of crosslingual test-time scaling—poor reasoning in low-resource languages and in cross-domain settings—one potential solution would be to curate multilingual reasoning training data with wide language and domain coverage. However, the current practice of generating multilingual reasoning training data simply through translation technology [39, 17] is insufficient, as it is well-established that translation models including LLMs still suffer from poor cultural alignment and Western-centric bias [65, 66] as well as poor translation performance with low-resource languages [67, 68, 69]. Future work

should systematically explore the effectiveness of multilingual augmentation techniques such as back-translation or synthetic data generation [70, 71].

Reasoning in low-resource languages (LRLs) Our results indicate poor performance of using LRLs as reasoning language. Furthermore, unfairness in tokenization for LRLs [56] leads to significantly higher inference costs for test-time scaling of RLMs. This disparity creates accessibility barriers for multilingual communities, as deploying reasoning models in LRLs becomes prohibitively expensive. Future work should focus on developing more equitable tokenization strategies for reasoning across diverse languages [72, 73, 74].

Crosslingual generalization for small RLMs Our observation of successful crosslingual test-time scaling happening at 3B parameter threshold corroborates contemporary work that shows small models below 3B parameters struggle to learn from long CoTs [75] and generalize in multilingual settings [17]. While there are ongoing work creating English-centric RLMs smaller than 3B size [28], we argue that future work on small RLMs should benchmark on multilingual reasoning tasks rather than solely on English data. This would not only advance the development of multilingual reasoning models but also help identify factors that enable crosslingual reasoning.

9 Conclusion

Our work studies the effectiveness of test-time scaling of English-centric reasoning language models for multilingual math reasoning, and we perform pareto frontier analyses of different model sizes and inference thinking compute. Furthermore, we document the language-mixing patterns naturally exhibited by reasoning models and report the novel "quote-and-think" pattern. We also investigate the effects of the choice of reasoning languages as well as query languages, demonstrating that successful generalization of English reasoning finetuning to other high-resource languages. Finally, we observe minimal cross-domain generalization of test-time scaling, thus highlighting the need for future work to explore more robust and domain-agnostic reasoning strategies, particularly for multilingual applications.

Acknowledgments and Disclosure of Funding

We thank Nihal Nayak and Peilin Yu from BATS research group at Brown University for helpful discussions. Disclosure: Stephen Bach is an advisor to Snorkel AI, a company that provides software and services for data-centric artificial intelligence.

References

- [1] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- [2] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.
- [3] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv* preprint arXiv:2408.00724, 2024.
- [4] Noam Levi. A simple model of inference scaling laws. arXiv preprint arXiv:2410.16377, 2024.
- [5] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [6] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv* preprint arXiv:2412.16720, 2024.

- [7] OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, April 2025.
- [8] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.
- [9] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- [10] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.
- [11] Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. The multilingual mind: A survey of multilingual reasoning in language models, 2025.
- [12] Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. Advancing language model reasoning through reinforcement learning and inference scaling. *arXiv* preprint arXiv:2501.11651, 2025.
- [13] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv* preprint *arXiv*:2412.06769, 2024.
- [14] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. In *ICLR*, 2024.
- [15] Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, et al. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-though. arXiv preprint arXiv:2501.04682, 2025.
- [16] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [17] Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. Linguistic generalizability of test-time scaling in mathematical reasoning. *arXiv preprint arXiv:2502.17407*, 2025.
- [18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.
- [19] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [20] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [21] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv* preprint arXiv:2112.00114, 2021.
- [22] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. In *The Twelfth International Conference on Learning Representations*, 2024.

- [23] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.
- [24] Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. Evolving deeper llm thinking. arXiv preprint arXiv:2501.09891, 2025.
- [25] Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? arXiv preprint arXiv:2411.16489, 2024.
- [26] Bespoke Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoningdistillation, 2025. Accessed: 2025-01-22.
- [27] Sathwik Tejaswi Madhusudhan, Shruthan Radhakrishna, Jash Mehta, and Toby Liang. Millions scale dataset distilled from r1-32b. https://huggingface.co/datasets/ServiceNow-AI/R1-Distill-SFT, 2025.
- [28] Shangshang Wang, Julian Asilis, Ömer Faruk Akgül, Enes Burak Bilgin, Ollie Liu, and Willie Neiswanger. Tina: Tiny reasoning models via lora. *arXiv preprint*, April 2025.
- [29] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [30] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*, 2024.
- [31] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023.
- [32] Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore, December 2023. Association for Computational Linguistics.
- [33] Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [34] Hyunwoo Ko, Guijin Son, and Dasol Choi. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv* preprint arXiv:2501.02448, 2025.
- [35] Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. LangBridge: Multilingual reasoning without multilingual supervision. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [36] Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. Mindmerger: Efficiently boosting LLM reasoning in non-english languages. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

- [37] Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [38] Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Language imbalance driven rewarding for multilingual self-improving. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [39] Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv* preprint *arXiv*:2310.20246, 2023.
- [40] Andy K Zhang, Kevin Klyman, Yifan Mai, Yoav Levine, Yian Zhang, Rishi Bommasani, and Percy Liang. Language model developers should report train-test overlap. *arXiv* preprint *arXiv*:2410.08385, 2024.
- [41] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024.
- [42] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [43] Shramay Palta and Rachel Rudinger. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [44] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [45] Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. COPAL-ID: Indonesian language reasoning with local culture and nuances. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1404–1422, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
- [46] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95, 2011.
- [47] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024.
- [48] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv* preprint arXiv:2309.05653, 2023.

- [49] Yuchun Fan, Yongyu Mu, YiLin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. SLAM: Towards efficient multilingual reasoning via selective language alignment. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9499–9515, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.
- [50] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- [51] Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. arXiv preprint arXiv:2401.00448, 2023.
- [52] Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
- [53] Qwen Team. Qwen3: Think deeper, act faster, 4 2025. 2036 words, 10 min read.
- [54] Philippe De Brabanter. Foreign-language quotations and code-switching: The grammar behind. In ESSE Conference (European Society for the Study of English), 2004.
- [55] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.
- [56] Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990, 2023.
- [57] Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.
- [58] Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, Advances in Neural Information Processing Systems, volume 37, pages 19209–19253. Curran Associates, Inc., 2024.
- [59] Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.
- [60] Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*, 2025.
- [61] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv* preprint arXiv:2412.21187, 2024.

- [62] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Hanjie Chen, Xia Hu, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.
- [63] Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [64] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024.
- [65] Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. Translating across cultures: LLMs for intralingual cultural adaptation. In Libby Barak and Malihe Alikhani, editors, *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 400–418, Miami, FL, USA, November 2024. Association for Computational Linguistics.
- [66] Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *arXiv preprint arXiv:2502.16534*, 2025.
- [67] Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Transla*tion, pages 392–418, Singapore, December 2023. Association for Computational Linguistics.
- [68] Sara Court and Micha Elsner. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [69] Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo Gentile, Radu State, Tegawendé F Bissyandé, and Jacques Klein. Is Ilm the silver bullet to low-resource languages machine translation? arXiv preprint arXiv:2503.24102, 2025.
- [70] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. LLM-powered data augmentation for enhanced cross-lingual performance. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore, December 2023. Association for Computational Linguistics.
- [71] Zheng Xin Yong, Cristina Menghini, and Stephen Bach. LexC-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association* for Computational Linguistics: EMNLP 2024, pages 13990–14009, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [72] Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore, December 2023. Association for Computational Linguistics.
- [73] HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, and Huda Khayrallah. Adapters for altering LLM vocabularies: What languages benefit the most? In *The Thirteenth International Conference on Learning Representations*, 2025.

- [74] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.
- [75] Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. arXiv preprint arXiv:2502.12143, 2025.
- [76] Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. How do multilingual models remember? investigating multilingual factual recall mechanisms. arXiv preprint arXiv:2410.14387, 2024.
- [77] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Contents

1	Intr	oduction	1
2	Bacl	kground and Related Work	2
3	Exp	erimental Setup	3
4	Cros	sslingual Test-Time Scaling	4
	4.1	Effectiveness of Crosslingual Test-Time Scaling	4
	4.2	Performance Comparison on MGSM Benchmark	5
5	Lan	guage-Mixing Behaviors	6
	5.1	Dominant Language in Model Outputs	6
	5.2	Language-Mixing Patterns During Reasoning	7
6	Lan	guage Forcing	7
	6.1	Methodology: Language Forcing Techniques	8
	6.2	In-Language Forcing	9
	6.3	Crosslingual Language Forcing	9
7	Cros	ss-Domain Generalization	11
	7.1	In-domain generalization	11
	7.2	Out-of-domain generalization	11
8	Disc	oussion and Future Work	12
9	Con	clusion	13
A	Auth	nors' Contributions	22
В	Furt	ther Details on Crosslingual Test-Time Scaling	22
C	Furt	ther Details on Language-Mixing Behaviors	22
	C.1	Methodology	22
	C.2	Annotation for Language-Mixing Patterns in s1 Reasoning	23
	C.3	Dominant Language in 14B-Sized Model Outputs	23
	C.4	Quote-And-Think Pattern in s1's Training Data	23
	C.5	Fine-Grained Analysis of s1-32B's Intrasentential Language Mixing	24
D	Furt	ther Details on Language Forcing	24
	D.1	Example of CoTs after Language Forcing	24
	D.2	Language Compliance	24
	D.3	In-Language Budget Forcing Reasoning	24

\mathbf{E}	Furt	ther Details on Cross-Domain Generalization	25
	E.1	Evaluation Details	25
	E.2	Language Forcing for GlobalMMLU	26

Appendix

A Authors' Contributions

Zheng-Xin Yong led the project and ran most of the experiments. Muhammad Farid Adilazuarda and Jonibek Mansurov worked on language forcing. Ruochen Zhang worked on cross-domain generalization with GlobalMMLU. Niklas Muennighoff trained s1 models at different model sizes.

Alham Fikri Aji, Stephen Bach, Julia Kreutzer, Genta Indra Winata, and Niklas Muennighoff advised the project at the early stage. All authors contributed to the paper writing.

B Further Details on Crosslingual Test-Time Scaling

Table 4, Table 5, Table 6, and Table 7 shows the s1 performance against Qwen baselines on the MGSM benchmark. Relative accuracy difference measures the relative gains for s1 under extrapolation budget forcing compared to zero-shot prompting Qwen models (first row), except for Table 7 where the relative accuracy is measured for s1 under truncation budget forcing.

Table 4: MGSM performance comparison against 32B-sized s1 model with maximum 8k thinking tokens.

Models	avg len	bn	de	en	es	fr	ja	ru	sw	te	th	zh	AVG
Qwen-32B-Instruct	365.6	82.0	79.6	84.0	80.0	69.2	80.8	77.2	54.0	56.4	84.4	84.0	75.6
+ 8-Shot EN-CoT	264.7	82.0	80.4	89.6	84.8	66.8	85.2	77.6	56.8	55.6	84.8	84.8	77.1
+ 8-Shot Native-CoT	179.9	82.4	78.0	90.0	87.2	66.8	80.8	74.0	57.2	59.6	87.2	88.0	77.4
s1-32B (truncation)	1682.1	89.2	86.8	94.8	90.0	84.8	84.8	91.2	66.8	65.6	87.2	88.8	84.5
s1-32B (extrapolation)	2610.1	89.2	88.8	93.2	89.6	83.6	84.8	88.8	70.8	68.4	88.0	88.0	84.8
Relative accuracy difference (%)		+8.8%	+11.6%	+11.0%	+12.0%	+20.8%	+5.0%	+15.0%	+31.1%	+21.3%	+4.3%	+4.8%	+12.2%

Table 5: MGSM performance comparison against 7B-sized s1 model with maximum 8k thinking tokens.

Models	avg len	bn	de	en	es	fr	ja	ru	sw	te	th	zh	AVG
Qwen-7B-Instruct	537.7	59.2	69.2	78.0	72.8	66.4	67.2	71.2	13.6	33.2	68.8	79.6	61.7
+ 8-Shot EN-CoT	537.8	62.0	72.4	86.4	78.8	62.4	68.0	76.8	15.2	33.2	70.4	76.8	63.9
+ 8-Shot Native-CoT	480.1	65.2	74.4	90.4	76.4	65.2	71.6	68.8	18.4	20.8	69.6	76.8	63.4
s1-7B (truncation)	3767.1	65.2	82.8	88.8	86.0	82.0	78.8	86.4	21.6	38.8	80.0	83.6	72.2
s1-7B (extrapolation)	4363.5	70.8	84.0	90.4	83.6	84.4	74.8	84.4	19.2	36.4	78.4	82.8	71.7
Relative accuracy difference (%)		+19.6%	+21.4%	+15.9%	+14.8%	+27.1%	+11.3%	+18.5%	+41.2%	+9.6%	+14.0%	+4.0%	+16.2%

Table 6: MGSM performance comparison against 3B-sized s1 model with maximum 8k thinking tokens.

Models	avg len	bn	de	en	es	fr	ja	ru	sw	te	th	zh	AVG
Qwen-3B-Instruct + 8-Shot EN-CoT + 8-Shot Native-CoT	1023.3 281.3 1657.0	37.6 48.0 36.0	58.8 67.6 63.2	74.0 79.2 80.0	66.0 71.2 70.8	54.4 65.2 58.4	54.8 58.0 52.0	64.8 70.4 62.0	9.2 12.4 9.6	7.6 14.8 9.6	56.8 60.4 59.2	68.4 68.4 70.4	50.2 56.0 51.9
s1-3B (truncation) s1-3B (extrapolation) Relative accuracy difference (%)	4813.3 5367.1	56.8 55.2 +46.8%	66.8 65.6 +11.6%	82.0 81.6 +10.3%	74.4 76.4 +15.8%	69.6 71.6 +31.6%	60.4 60.8 +10.9%	72.4 74.4 +14.8%	10.4 9.6 +4.3%	16.8 20.0 +163.2%	68.8 68.8 +21.1%	72.0 74.0 +8.2%	59.1 59.8 +19.1%

C Further Details on Language-Mixing Behaviors

C.1 Methodology

To filter out language-mixed sentences, we first identify the *dominant language*, also known as matrix language, of the generated response using the state-of-the-art language identification library lingua. Then, we use the NLP library stanza to perform sentence segmentation according to the matrix language and obtain individual sentences. Finally, we use lingua to annotate the language label of each sentence and of each individual word token in the sentence.

Table 7: MGSM performance comparison against 1.5B-sized s1 model with maximum 8k thinking tokens. We didn't run extrapolation budget forcing since without it, s1 already generates extremely long CoTs.

Models	avg len	bn	de	en	es	fr	ja	ru	sw	te	th	zh	AVG
Qwen-1.5B-Instruct	2991.7	10.4	35.6	66.0	52.8	41.2	31.2	43.6	2.0	1.2	31.2	56.0	33.7
+ 8-Shot EN-CoT	1100.3	21.6	46.4	70.0	58.0	55.2	37.2	51.6	2.8	6.4	41.2	52.0	40.2
+ 8-Shot Native-CoT	1729.9	14.0	44.4	71.6	52.4	41.6	34.4	39.2	3.2	1.6	33.2	54.0	35.4
s1-1.5B (truncation)	8227.2	27.6	51.2	66.8	62.8	56.0	43.2	55.6	1.6	6.4	46.4	58.8	43.3
Relative accuracy difference (%)		+165.4%	+43.8%	+1.2%	+18.9%	+35.9%	+38.5%	+27.5%	-20.0%	+433.3%	+48.7%	+5.0%	+28.5%

We classify language-mixing patterns into three categories: (1) quote-and-think, where words or phrases in foreign language are quoted in quotation marks; (2) intersentential, where the entire sentence is in a language entirely different from generation dominant language, and (3) intrasentential, where words, phrases or clauses of different languages are present in the same sentence. We refer our readers to Appendix C.2 for our annotation procedures.

We focus on four languages, namely Japanese (ja), Russian (ru), Thai (th), and Mandarin Chinese (zh), as they are readily supported by the libraries stanza and lingua. We avoid languages with Latin scripts due to their possible shared vocabulary with English and results in incorrect language classification. For instance, the German word 'also' (therefore) in the sentence "Sie isst 3 Eier zum Frühstück und verwendet 4 Eier für Muffins, also verwendet sie insgesamt 3 + 4 = 7 Eier pro Tag." was misclassified as English word.

C.2 Annotation for Language-Mixing Patterns in s1 Reasoning

We collect the language label with the highest probability assigned to *the entire sentence*, and we label a sentence belonging to "intersentential" language-mixing for s1 if the sentence is non-English, as the dominant language of s1's overall output is English. We then check language labels for individual word tokens. If there are mixing of different languages *within the same sentence*, and quotation marks are present around the non-English words or phrases, then the sentence is assigned with the "quote-and-think" label. Otherwise, if quotation marks are not present, the sentence is assigned with the "intrasentential" label.

C.3 Dominant Language in 14B-Sized Model Outputs

Figure 8 shows the dominant language distribution in model outputs when MGSM questions are asked in Japanese (ja), Russian (ru), Thai (th), and Mandarin Chinese (zh).

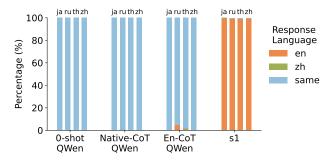


Figure 8: Proportion of dominant languages used by 14B-sized models' responses when queried with Japanese (ja), Russian (ru), Thai (th), and Mandarin Chinese (zh) languages. "same" indicates that the response language is the same as query language.

C.4 Quote-And-Think Pattern in s1's Training Data

Among 1k English training samples of s1 models, 68.3% of the samples exhibit the quote-and-think pattern, among which at least half of them involves directly copying from the question prompts. This suggests that the quote-and-think language-mixing pattern is due to crosslingual transfer of the

original s1 model's learned behavior of quoting phrases from question prompts during its long CoTs thinking process.

C.5 Fine-Grained Analysis of s1-32B's Intrasentential Language Mixing

We perform human annotations on the intrasententially language-mixed sentences during reasoning and classify if each sentence belongs to one of the following categories: (1) **extract-and-explain**, where the non-English phrases are taken directly from the original input prompt but *without* quotation marks given (this resembles quote-and-think but no quotation marks are generated around the non-English phrases); (2) **insertional code-switching**, where non-English lexical items (usually nouns) are inserted into the morphosyntactic frame of the English sentence (an example would be "I want to eat *nasi goreng*" where *nasi goreng* is a Malay word for fried rice), and (3) **clause-level code-switching**, where switching between two languages within a single sentence structure at clausal level (an example would be "I want to go to the library *dan bersedia untuk peperiksaan*" where *dan bersedia untuk peperiksaan* is a Malay clause for "and prepare for the exam".) Figure 9 demonstrates the distribution of each category, with extract-and-explain being the dominant language-mixing patterns.

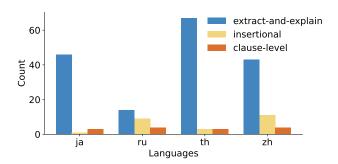


Figure 9: Breakdown of s1-32B's finegrained intrasentential language mixing patterns.

D Further Details on Language Forcing

D.1 Example of CoTs after Language Forcing

Box 2 show the example of CoTs generated by s1 from the combined language-forcing strategy. This example contrasts Box 1 where Box 1 showcases s1's natural CoTs without language forcing.

Note that we still observe the sophisticated *quote-and-think* pattern where here, English phrases are quoted for reasoning (the dominant language is Japanese). Specifically, in that sentence where quote-and-think occurs, it is translated to 'if this question was in English, it would be "how many rolls are needed in total?"

D.2 Language Compliance

Detailed results for language compliance can be found in Tables 8, 9 and 10 for 32B, 14B and 7B, respectively.

D.3 In-Language Budget Forcing Reasoning

Detailed results for in-language budget forcing reasoning performance on MGSM are presented in Table 11.

Box 2. Example of CoTs prdocued from combined language forcing strategy.

Question: ローブを作成するには、青色の繊維を2巻分、白色の繊維をその半分用いる必要があります。全体で何巻必要ですか?

(To make a robe, you need 2 rolls of blue fiber and half of the white fiber. How many rolls are needed total?)

s1's Thinking: よし、解いてみよう。問題はローブを作るために必要な繊維の巻数を計算するものだね。青色の繊維が2巻必要で、白色の繊維はその半分、つまり1巻必要だ。だから合計で3巻か。でも待って、問題文をもう一度確認しよう。

「ローブを作成するには、青色の繊維を2巻分、白色の繊維をその半分用いる必要があります。全体で何巻必要ですか?」そうだ、青色が2巻、白色がその半分。つまり青色の半分は1巻だから、白色は1巻必要。だから合計3巻。これでいいのかな。でも、もしかして青色の2巻と白色の半分を足すと、2+0.5 = 2.5巻って意味じゃないのだろうか。でも問題の言い回しが少しややこしいな。

待って、原文をもう一度。「青色の繊維を2巻分、白色の繊維をその半分用いる」というのは、青色が2巻、そして白色が「その半分」。つまり、「青色の半分」なのか、それとも「2巻の半分」なのか。日本語の文法で考えると、「その半分」は直前の数詞、「2巻分」を指している可能性が高い。つまり、青色2巻、白色は2巻の半分なので1巻。だから合計3巻。でも、もし「青色の繊維を2巻分、白色の繊維をその半分」というのが、青色が2巻、白色が青色の半分、つまり1巻、という意味なら、同じく3巻。結局同じ結果になるのか。

それに、問題が「全体で何巻必要ですか?」と聞いているので、必要な総巻数を聞かれているのだろう。つまり、青色2巻、白色1巻、合計3巻。これが答えだと思われる。でも、もしこの問題が英語だと、 \ddot{h} ow many rolls are needed in total?"だと仮定すると、 \ddot{h} or rollsが答え。でも、もしかして青色2巻、白色半分、つまり2.5巻という解釈もある。でも、問題文の日本語の表現からすると、青色2巻、白色1巻、合計3巻が妥当だと思う。

...

Box 3. GlobalMMLU Eval Prompt Template

System: Your task is to extract the answer choice from the Response field. Do not attempt to answer the question in the Question field yourself.

Task Prompt: Your task is to extract the answer (A, B, C, or D) from the generated response based on the question and the option choices.

Question: {question} Answer choices: {answer_options}

Response: {response}

E Further Details on Cross-Domain Generalization

E.1 Evaluation Details

For GlobalMMLU, we use GPT-4o-mini to parse the final answer since their formats can vary in the reasoning traces. We set max_tokens to 2 to check if the extracted answer contains any of the four options (A, B, C or D). See the prompts used for evaluation in Box 3.

Table 8: Results for s1-32B.

Strategy	Reasoning Language	avg len	Language Distribution	Language Compliance
	bn	3073.3	ENGLISH: 99.6, BENGALI: 0.4	0.4
	de	2405.3	ENGLISH: 51.6, GERMAN: 48.4	48.4
	en	1833.1	ENGLISH: 100.0	100.0
	es	2401.5	ENGLISH: 88.4, SPANISH: 11.6	11.6
	fr	2379.7	ENGLISH: 90.8, FRENCH: 8.8, CHINESE: 0.4	8.8
translated_wait	ja	2515.1	ENGLISH: 70.8, JAPANESE: 29.2	29.2
	ru	2601.1	ENGLISH: 90.4, RUSSIAN: 9.2, CHINESE: 0.4	9.2
	sw	2611.4	ENGLISH: 100.0	0.0
	te	3821.3	ENGLISH: 99.6, TELUGU: 0.4	0.4
	th	1894.7	ENGLISH: 99.6, CHINESE: 0.4	0.0
	zh	1776.3	ENGLISH: 94.4, CHINESE: 5.6	5.6
	bn	3320.0	BENGALI: 98.8, ENGLISH: 1.2	98.8
	de	1747.2	GERMAN: 98.8, CHINESE: 1.2	98.8
	en	1729.7	ENGLISH: 100.0	100.0
	es	2790.6	SPANISH: 99.6, ENGLISH: 0.4	99.6
	fr	1822.9	FRENCH: 100.0	100.0
prefix	ja	2321.6	JAPANESE: 98.8, ENGLISH: 1.2	98.8
	ru	1564.2	RUSSIAN: 98.8, ENGLISH: 0.8, CHINESE: 0.4	98.8
	sw	3083.9	SWAHILI: 91.2, ENGLISH: 8.4, JAPANESE: 0.4	91.2
	te	6912.7	TELUGU: 95.2, ENGLISH: 4.8	95.2
	th	2126.3	THAI: 77.6, CHINESE: 20.4, ENGLISH: 2.0	77.6
	zh	1150.1	CHINESE: 99.6, ENGLISH: 0.4	99.6
	bn	2693.8	ENGLISH: 90.4, BENGALI: 9.6	9.6
	de	1979.3	GERMAN: 100.0	100.0
	en	1728.3	ENGLISH: 100.0	100.0
	es	2241.6	ENGLISH: 69.6, SPANISH: 30.4	30.4
	fr	2346.0	ENGLISH: 68.4, FRENCH: 31.2, CHINESE: 0.4	31.2
system	ja	1805.7	JAPANESE: 99.2, ENGLISH: 0.8	99.2
-	ru	2180.0	ENGLISH: 59.6, RUSSIAN: 39.2, CHINESE: 1.2	39.2
	sw	2721.2	ENGLISH: 98.8, SWAHILI: 0.8, CHINESE: 0.4	0.8
	te	3869.3	ENGLISH: 93.6, TELUGU: 6.4	6.4
	th	1930.3	ENGLISH: 90.8, THAI: 8.0, CHINESE: 1.2	8.0
	zh	1162.1	CHINESE: 100.0	100.0
	bn	3507.6	BENGALI: 99.6, CHINESE: 0.4	99.6
	de	1845.2	GERMAN: 100.0	100.0
	en	1582.6	ENGLISH: 100.0	100.0
	es	2604.3	SPANISH: 100.0	100.0
	fr	1726.5	FRENCH: 99.2, CHINESE: 0.8	99.2
combined	ja	2127.4	JAPANESE: 100.0	100.0
	ru	1523.6	RUSSIAN: 98.8, ENGLISH: 0.4, GERMAN: 0.4, CHINESE: 0.4	98.8
	sw	3161.1	SWAHILI: 98.4, ENGLISH: 1.6	98.4
	te	7036.0	TELUGU: 97.6, ENGLISH: 2.4	97.6
	th	2043.3	THAI: 90.8, CHINESE: 8.8, GERMAN: 0.4	90.8
	zh	1188.7	CHINESE: 100.0	100.0

E.2 Language Forcing for GlobalMMLU

We further study if forcing s1 to reason in the same language as the question helps with performance, as different languages may have different knowledge recall mechanism [76], and retrieving the right knowledge will help with MMLU-style questions [77]. Following the findings in Section 6, we use the combined language forcing strategy that yields the highest language compliance in multilingual generations.

Table 13 shows marginal performance improvements for reasoning language in HRLs with Latin scripts, such as German (de), English (en) and French (fr). We want to highlight that the performance gain from language forcing is inconsistent: for instance, in-language reasoning for German business subject questions improves accuracy by around 3% but decreases accuracy for humanities subject by nearly 5%. For Chinese, even though it is one of the highly represented languages during pretraining of the base model Qwen, reasoning in Chinese introduces a minor performance degradation even for STEM domain. For the even lower-resourced languages like Bengali (bn) and Swahili (sw), language forcing leads to performance degradation—there is around 6% decrease on average across domains. In general, the patterns for non-math-specific domains are similar that reasoning in HRLs would be more beneficial compared to LRLs; nonetheless the latter could be more friendly in a user-facing setting and could potentially help with easier verification for more cultural-specific tasks.

Table 9: Results for s1-14B.

Strategy	Reasoning Language	avg len	Language Distribution	Language Compliance
	, bn	2457.1	ENGLISH: 100.0	0.0
	de	1940.1	ENGLISH: 76.4, GERMAN: 23.6	23.6
	en	1638.2	ENGLISH: 100.0	100.0
	es	1966.1	ENGLISH: 94.4, SPANISH: 5.6	5.6
	fr	2062.6	ENGLISH: 95.2, FRENCH: 4.4, CHINESE: 0.4	4.4
translated_wait	ja ja	2162.1	ENGLISH: 86.8, JAPANESE: 13.2	13.2
	ru	1937.9	ENGLISH: 98.8, RUSSIAN: 0.8, CHINESE: 0.4	0.8
	sw	3044.8	ENGLISH: 99.6, SWAHILI: 0.4	0.4
	te	3852.4	ENGLISH: 98.8, CHINESE: 0.4, TELUGU: 0.8	0.8
	th	1767.3	ENGLISH: 100.0	0.0
	zh	1438.9	ENGLISH: 62.0, CHINESE: 38.0	38.0
	bn	4570.1	BENGALI: 92.8, ENGLISH: 7.2	92.8
	de	1621.8	GERMAN: 100.0	100.0
	en	1572.2	ENGLISH: 100.0	100.0
	es	2343.2	SPANISH: 99.6, ENGLISH: 0.4	99.6
	fr	1354.3	FRENCH: 99.6, ENGLISH: 0.4	99.6
prefix	ja ja	2194.3	JAPANESE: 99.2, ENGLISH: 0.8	99.2
	ru	1470.3	RUSSIAN: 98.8, CHINESE: 0.8, ENGLISH: 0.4	98.8
	sw	4442.6	SWAHILI: 85.2, ENGLISH: 14.4, TAGALOG: 0.4	85.2
	te	6041.8	TELUGU: 90.4, ENGLISH: 9.2, CHINESE: 0.4	90.4
	th	2448.6	THAI: 88.4, CHINESE: 10.8, ENGLISH: 0.8	88.4
	zh	1149.3	CHINESE: 100.0	100.0
	bn	2196.2	ENGLISH: 99.2, BENGALI: 0.8	0.8
	de	1643.4	GERMAN: 92.8, ENGLISH: 7.2	92.8
	en	1664.3	ENGLISH: 100.0	100.0
	es	1766.7	ENGLISH: 84.4, SPANISH: 15.6	15.6
	fr	1841.7	ENGLISH: 91.6, FRENCH: 8.4	8.4
system	ja ja	2036.4	JAPANESE: 17.2, ENGLISH: 82.8	17.2
	ru	1631.3	RUSSIAN: 49.2, ENGLISH: 50.8	49.2
	sw	2984.6	ENGLISH: 98.8, SWAHILI: 1.2	1.2
	te	3942.8	ENGLISH: 91.6, TELUGU: 8.4	8.4
	th	1775.4	ENGLISH: 91.6, THAI: 8.0, CHINESE: 0.4	8.0
	zh	1295.8	ENGLISH: 12.0, CHINESE: 88.0	88.0
	bn	4823.5	BENGALI: 98.0, ENGLISH: 2.0	98.0
	de	1479.8	GERMAN: 100.0	100.0
	en	1553.2	ENGLISH: 100.0	100.0
	es	2503.3	SPANISH: 100.0	100.0
	fr	1269.8	FRENCH: 100.0	100.0
combined	ja ja	1941.4	JAPANESE: 100.0	100.0
	ru	1577.6	RUSSIAN: 99.6, CHINESE: 0.4	99.6
	sw	4525.5	SWAHILI: 88.4, ENGLISH: 11.6	88.4
	te	6046.9	TELUGU: 93.6, ENGLISH: 6.4	93.6
	th	2244.8	THAI: 92.4, CHINESE: 7.2, ENGLISH: 0.4	92.4
	zh	1118.8	CHINESE: 100.0	100.0

Table 10: Results for s1-7B.

Strategy	Reasoning Language	avg len	Language Distribution	Language Compliance
	, bn	3355.8	ENGLISH: 100.0	0.0
	de	2490.5	ENGLISH: 68.0, GERMAN: 32.0	32.0
	en	2050.4	ENGLISH: 100.0	100.0
	es	3156.8	ENGLISH: 83.6, SPANISH: 16.4	16.4
	fr	2544.1	ENGLISH: 90.4, FRENCH: 9.2, CHINESE: 0.4	9.2
translated_wait	ja ja	3381.9	ENGLISH: 72.0, JAPANESE: 28.0	28.0
	ru	2742.3	ENGLISH: 96.0, RUSSIAN: 2.8, CHINESE: 1.2	2.8
	sw	5381.5	ENGLISH: 100.0	0.0
	te	5232.8	ENGLISH: 97.6, TELUGU: 2.4	2.4
	th	2654.8	ENGLISH: 89.6, THAI: 8.4, CHINESE: 2.0	8.4
	zh	1471.1	CHINESE: 99.2, ENGLISH: 0.8	99.2
	bn	3814.5	BENGALI: 91.6, ENGLISH: 7.2, CHINESE: 1.2	91.6
	de	2183.4	GERMAN: 98.0, CHINESE: 1.6, ENGLISH: 0.4	98.0
	en	2405.3	ENGLISH: 100.0	100.0
	es	3468.3	SPANISH: 99.2, CHINESE: 0.4, ENGLISH: 0.4	99.2
	fr	1712.5	FRENCH: 99.2, CHINESE: 0.8	99.2
prefix	ja	4976.8	JAPANESE: 99.6, ENGLISH: 0.4	99.6
	ru	2242.7	RUSSIAN: 82.4, CHINESE: 13.6, ENGLISH: 4.0	82.4
	sw	7653.5	SWAHILI: 55.2, ENGLISH: 43.6, TAGALOG: 1.2	55.2
	te	6649.8	TELUGU: 82.4, ENGLISH: 17.2, CHINESE: 0.4	82.4
	th	3239.8	THAI: 80.0, CHINESE: 17.6, ENGLISH: 2.4	80.0
	zh	1675.4	CHINESE: 100.0	100.0
	bn	3508.5	ENGLISH: 96.4, BENGALI: 3.2, CHINESE: 0.4	3.2
	de	2093.8	GERMAN: 97.2, CHINESE: 2.8	97.2
	en	2721.8	ENGLISH: 100.0	100.0
	es	4099.5	SPANISH: 100.0	100.0
	fr	1872.1	FRENCH: 99.6, CHINESE: 0.4	99.6
system	ja	2513.9	JAPANESE: 100.0	100.0
	ru	2166.5	RUSSIAN: 79.2, CHINESE: 9.2, ENGLISH: 11.6	79.2
	sw	5310.8	ENGLISH: 100.0	0.0
	te	5250.4	ENGLISH: 96.8, TELUGU: 3.2	3.2
	th	3438.9	THAI: 93.2, CHINESE: 6.8	93.2
	zh	1528.5	CHINESE: 100.0	100.0
	bn	4094.8	BENGALI: 97.6, CHINESE: 0.8, ENGLISH: 1.6	97.6
	de	2046.7	GERMAN: 100.0	100.0
	en	2857.6	ENGLISH: 100.0	100.0
	es	3907.6	SPANISH: 100.0	100.0
	fr	1657.1	FRENCH: 99.6, CHINESE: 0.4	99.6
combined	ja ja	4804.9	JAPANESE: 100.0	100.0
	ru	2450.9	RUSSIAN: 86.4, CHINESE: 13.2, ENGLISH: 0.4	86.4
	sw	7393.2	SWAHILI: 81.6, ENGLISH: 18.4	81.6
	te	6403.0	TELUGU: 98.4, ENGLISH: 1.6	98.4
	th	3609.9	THAI: 92.8, CHINESE: 7.2	92.8
	zh	1734.0	CHINESE: 100.0	100.0

Table 11: Performance comparison of different language forcing strategies across multiple model sizes and languages on fixed 8k thinking tokens. Languages are categorized into high-resource (HRL: de, en, es, fr, ru, ja, zh) and low-resource (LRL: bn, sw, te, th) groups.

Model	Method	bn	de	en	es	fr	ja	ru	sw	te	th	zh	ALL	HRL	LRL
s1.1-32B	Baseline	90.8	90.8	96.0	93.2	89.6	87.6	93.2	72.4	68.4	91.6	88.0	87.4	91.2	80.8
	translated_wait	91.2	90.4	94.8	93.2	89.2	89.2	92.0	73.2	70.8	91.6	90.0	87.8	91.3	81.7
	prefix	85.2	90.4	95.6	92.8	90.4	84.8	94.0	65.2	63.6	88.8	90.8	85.6	91.3	75.7
	system	87.6	90.0	96.4	91.2	86.8	85.2	92.8	71.2	67.2	90.8	89.6	86.0	90.3	79.3
	combined	82.8	89.2	95.2	91.6	88.8	85.2	92.8	58.4	63.2	89.2	90.4	84.3	90.5	73.4
	Baseline	86.8	90.4	94.4	93.6	88.4	83.6	92.4	59.6	60.0	89.2	89.6	84.4	90.3	73.9
	translated_wait	85.6	90.0	96.8	93.6	86.8	85.2	92.4	63.2	61.2	90.8	89.2	85.0	90.6	75.2
s1.1-14B	prefix	81.2	90.4	95.2	92.0	90.4	82.8	92.8	44.4	55.2	86.8	91.2	82.0	90.7	66.9
	system	84.0	88.8	95.2	91.2	87.2	82.8	91.2	58.8	62.0	90.4	90.4	83.8	89.5	73.8
	combined	81.2	90.0	93.2	92.0	86.4	82.4	90.4	36.8	54.4	89.2	88.4	80.4	89.0	65.4
	Baseline	72.0	87.6	92.4	88.8	83.2	82.4	88.0	24.0	36.8	81.6	86.0	74.8	86.9	53.6
s1.1-7B	translated_wait	69.2	84.0	93.2	89.2	87.2	76.4	87.2	24.0	37.2	84.0	83.2	74.1	85.8	53.6
	prefix	64.0	82.8	93.6	86.8	87.2	68.0	85.6	14.4	24.0	74.4	84.0	69.5	84.0	44.2
	system	71.6	84.0	90.8	92.0	82.8	74.4	87.6	25.6	36.8	76.8	82.0	73.1	84.8	52.7
	combined	60.8	84.0	92.8	88.0	83.6	72.8	86.4	14.8	27.6	74.4	83.6	69.9	84.5	44.4

Table 12: Number of average thinking tokens for each reasoning language when the MGSM task questions are asked in a particular query language.

Query	Reasoning Language									AVG		
Language	bn	de	en	es	fr	ja	ru	sw	te	th	zh	AVG
bn	4,559	1,974	2,190	3,473	1,640	2,276	1,977	4,900	5,461	2,852	1,710	3,001
de	3,838	1,642	1,910	2,838	1,159	1,794	1,617	3,510	5,073	2,025	1,327	2,430
en	3,429	1,239	1,467	2,253	1,075	1,625	1,341	3,431	5,154	1,703	1,075	2,162
es	3,736	1,388	1,779	2,512	1,225	1,665	1,467	3,600	5,323	1,875	1,133	2,336
fr	3,868	1,562	1,886	2,886	1,218	2,103	1,577	4,077	5,033	2,218	1,263	2,517
ja	3,994	1,623	2,028	2,627	1,366	2,033	1,616	3,957	5,334	2,285	1,333	2,563
ru	4,105	1,392	1,713	2,695	1,170	2,007	1,469	4,055	4,818	2,132	1,098	2,423
sw	5,958	2,945	2,955	3,801	2,144	4,114	3,108	3,867	6,836	3,976	2,705	3,855
te	5,761	3,341	3,528	4,160	2,473	3,574	3,337	5,934	6,277	3,879	2,557	4,074
th	4,166	1,687	2,000	2,801	1,300	1,993	1,756	4,107	5,247	2,279	1,241	2,597
zh	4,017	1,650	1,737	2,521	1,365	1,822	1,649	4,233	5,344	2,009	1,117	2,496
AVG	4,311	1,858	2,108	2,960	1,466	2,273	1,901	4,151	5,445	2,475	1,505	>2,585

Table 13: Model performances comparison with and without using combined strategy. Green cells show positive gains compared to no-forcing setting and red for the opposite.

Model		Domain	Languages								
	Method		bn	de	en	fr	ja	sw	zh		
	none	STEM	95.65	93.48	93.48	95.65	93.48	76.09	93.48		
		business	74.14	79.31	91.38	82.76	89.66	46.55	81.03		
		humanities	69.61	83.33	88.24	81.37	79.41	45.10	80.39		
		medical	66.67	77.78	80.56	72.22	69.44	52.78	77.78		
		other	76.79	87.50	87.50	83.93	85.71	66.07	80.36		
		social sciences	71.57	82.35	89.22	83.33	80.39	47.06	84.31		
s1.1-32B		AVG	75.74	83.96	88.39	83.21	83.02	55.61	82.89		
	combined	STEM	89.13	95.65	97.83	97.83	86.96	67.39	91.30		
		business	67.24	82.76	87.93	81.03	82.76	32.76	86.21		
		humanities	67.65	78.43	87.25	79.41	74.51	49.02	79.41		
		medical	63.89	83.33	80.56	72.22	77.78	44.44	75.00		
		other	66.07	80.36	94.64	87.50	76.79	64.29	76.79		
		social sciences	64.71	84.31	88.24	81.37	82.35	41.18	80.39		
		AVG	69.78	84.14	89.41	83.23	80.19	49.85	81.52		