WHEN VERIFIABLE REWARDS SWITCH THE LANGUAGE: CROSS-LINGUAL COLLAPSE IN CHAIN-OF-THOUGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

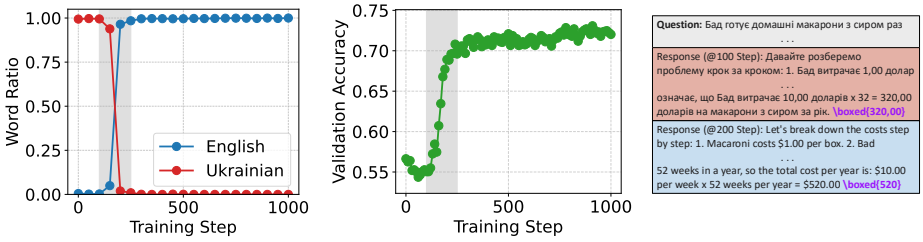
Reinforcement learning with verifiable reward (RLVR) has been instrumental in eliciting strong reasoning capabilities from large language models (LLMs) via long chains of thought (CoT). During RLVR training, we identify an empirical phenomenon—a systematic drift whereby a multilingual model's CoT reverts to its dominant pre-training language (e.g., English) even when prompted in another language—which we term Cross-lingual Collapse. Because the long-CoT regime magnifies exposure to linguistic priors, the underlying trade-off between maximizing reasoning depth and preserving target-language fidelity has remained uncharacterized. To examine this trade-off, we train LLMs with Group-Relative Policy Optimization (GRPO) on translated versions of math datasets widely used to elicit long-CoT reasoning. Throughout training, we track both task accuracy and the language consistency of reasoning chains. Our experiments yield three findings: (i) under RLVR, CoT in LLMs systematically drifts toward the pre-training dominant language as reasoning performance rises; (ii) English-centric priors, long-CoT GRPO optimization, task difficulty, and high-entropy decoding jointly amplify this drift, and the pattern persists beyond mathematics; and (iii) interventions that favor target-language traces—via a language-consistency reward, decoding-time controls, or more balanced backbones—mitigate collapse but reveal a persistent performance–fidelity trade-off.

1 INTRODUCTION

Large language models (LLMs) trained with long chain-of-thought (CoT) supervision have demonstrated impressive performance across mathematically demanding problems, code generation tasks, and multi-step logical reasoning benchmarks (Wei et al., 2022; Shao et al., 2024; Yu et al., 2025; DeepSeek-AI et al., 2025). These models' strengthened reasoning capabilities not only enable human-level performance on challenging tasks but also facilitate monitoring of intermediate reasoning traces, thereby improving interpretability and enabling more reliable auditing.

Although multilingual competence has been studied during pre-training and instruction tuning (Shaham et al., 2024; Zhong et al., 2024; Kew et al., 2024; Wang et al., 2025), reasoning-centric models remain comparatively underexplored. We posit an inherent *trade-off*: pushing for deeper, verification-driven reasoning with long CoT can come at the expense of *target-language* fidelity. Mechanistically, long CoT increases exposure to pre-training priors; when those priors are English-dominant—as is the case for most open-source foundation models (OLMo et al., 2024; Grattafiori et al., 2024; Yoo et al., 2024b; Yang et al., 2025; Team et al., 2025)—reward-seeking optimization can preferentially route the reasoning trace through English even under non-English prompts. We refer to the resulting drift as **Cross-lingual Collapse**: the chain-of-thought reverts to the pre-training dominant language while task performance continues to rise.

To systematically analyze this performance–fidelity trade-off, we study target-language reasoning under reinforcement learning with verifiable reward (RLVR). We instantiate Group-Relative Policy Optimization (GRPO) (Shao et al., 2024) on an English-centric backbone (OLMo et al., 2024) and non-English-centric backbones (Grattafiori et al., 2024; Yang et al., 2025), using standard math word-problem corpora widely used to elicit long-CoT reasoning (e.g., GSM8K (Cobbe



(a) Training Step vs. Word Ratio (b) GSM8K (UK) Performance (c) Response Examples

Figure 1: Illustration of **Cross-lingual Collapse**. We train Llama-3.2-3B Instruct with GRPO on a fully Ukrainian translation of GSM8K, seeking Ukrainian-only reasoning. **(a)** Chain-of-thought word-ratio in reward warding roll-outs over training steps. In the grey band, the share of Ukrainian tokens plummets, while English abruptly dominates, signaling a language switch within the roll-out reasoning trace. **(b)** Accuracy on the Ukrainian GSM8K. The sharp rise in accuracy aligns with the same 100–250-step window, showing that the model scores higher once its reasoning drifts into English. **(c)** Representative responses at steps 100 and 200 (answer spans highlighted in purple). When the model reasons in Ukrainian it produces an incorrect answer, but after switching to English it solves the problem correctly, exemplifying the collapse from target-language reasoning to the pre-training-dominant language. The word ratio is measured during training from the rollout samples.

et al., 2021), SimpleRL-Zoo (Zeng et al., 2025)) translated into three target languages (Chinese, Korean, Ukrainian). Our evaluation tracks (i) task accuracy and (ii) a target-language word ratio over training, enabling us to quantify language drift alongside performance. Beyond measurement of Cross-Lingual Collapse, we interrogate both the amplifiers and mechanisms and the mitigations and limits of this behavior. Our novelty is three-fold:

- **Phenomenon.** We identify Cross-Lingual Collapse—the systematic reversion of chain-of-thought to the pre-training dominant language as reasoning performance rises—and operationalize it via accuracy and a target-language word ratio.
- **Amplifiers and mechanisms.** We show that English-dominant language model and long-CoT GRPO optimization steer reward toward dominant-language traces, and that task difficulty and high-entropy decoding further exacerbate the drift; the pattern persists beyond mathematics.
- **Mitigations and limits.** We evaluate interventions (language-consistency reward signal, decoding-time controls, and multilingual mixing) that preserve target-language fidelity to varying degrees, revealing a persistent performance–fidelity trade-off.

2 Motivation

Recent reinforcement learning with verifiable reward (RLVR) methods such as Group-Relative Policy Optimization (GRPO) (DeepSeek-AI et al., 2025) unlock state-of-the-art reasoning by having the model speak its thoughts aloud: each answer is preceded by a multi-step chain-of-thought that can be several hundred tokens long. With this drastic increase in utterance length, the burden on the model's linguistic competence also multiplies for every step of the trace.

In non-english contexts, this burden is even greater Marchisio et al. (2024). For English-centric LLMs, a single error introduced during an early non-English step can propagate through the entire chain of reasoning, ultimately compromising the final answer. Early work (Shaham et al., 2024; Kew et al., 2024) demonstrated that even target-language-centric supervised fine-tuning (SFT) (Ouyang et al., 2022) on a single language can still coax a model into showing modest generalization beyond English. However, current evidence is sparse on how reasoning-driven training like GRPO affects these cross-lingual gains—do they hold steady, or do they shift?

We therefore ran a pilot experiment on the Llama-3.2-3B Instruct, giving it target-language reasoning supervision through GRPO. Concretely, we fine-tuned the model on the GSM8K grade-school arithmetic corpus, translated into Ukrainian so that all intermediate chain-of-thought steps as well as

Table 1: Accuracy and target-language word ratio for models fine-tuned with GRPO on translated GSM8K. We evaluate on the translated GSM8K and MATH500 test sets. Language codes: **EN** = English, **ZH** = Chinese, **KO** = Korean, **UK** = Ukrainian. Model keys: **OLMo 2** = OLMo-2-0425-IB-Instruct, **Llama** = Llama-3.2-3B Instruct, **Qwen** = Qwen-2.5-1.5B Instruct. Numbers in parentheses indicate the change relative to the corresponding non-fine-tuned baseline. Accuracy (Acc) and target-language word ratio (WR) with languages and models arranged as rows.

Language	Model	GSM8K			MATH500		
		Target Acc (%)	Target WR (%)	EN WR (%)	Target Acc (%)	Target WR (%)	EN WR (%)
ZH	OLMo2	59.8 (+34.3)	0.3 (-75.5)	80.8 (+73.7)	17.6(+1.9)	26.3 (-10.0)	71.0 (+8.4)
	Llama	69.4 (+7.4)	94.1 (-1.4)	8.3 (-0.5)	38.8 (+1.2)	77.5 (-0.4)	18.8 (+0.1)
	Qwen	63.4 (+1.3)	92.9 (+0.6)	7.0 (-0.9)	41.9 (+4.7)	79.8 (+0.4)	19.5 (-0.7)
KO	OLMo2	46.5 (+39.9)	14.3 (-79.4)	83.5 (+78.3)	12.2 (+5.2)	0.1 (-45.1)	73.0 (+51.3)
	Llama	61.3 (+14.5)	82.4 (-8.1)	14.7 (+7.1)	28.5 (+7.2)	70.9 (-17.8)	21.8 (+16.0)
	Qwen	42.2 (+3.5)	94.3 (-2.4)	3.1 (+1.9)	27.0 (+6.8)	88.0 (-8.0)	10.1 (+7.7)
UK	OLMo2	45.2 (+37.8)	0.3 (-75.5)	85.3 (+79.3)	13.0 (+5.6)	0.1 (-52.3)	72.7 (+56.1)
	Llama	70.9 (+17.1)	**0.3** (-82.6)	96.8 (+80.8)	47.6 (+12.0)	**5.6** (-72.7)	93.4 (+73.1)
	Qwen	39.7 (+4.9)	99.3 (+0.5)	0.5 (-0.2)	23.4 (+4.0)	82.8 (-9.8)	9.9 (+8.5)

the final answer were presented in a low-resource language (relatively lower than English (Wenzek et al., 2020)). As training progressed, however, the chains gradually drifted back to high-resource languages, chiefly English, even though the prompts remained Ukrainian. The trend is visualized in Figure 1. We dub this behaviour **Cross-lingual Collapse** in reasoning models: a systematic collapse of target-lingual chains-of-thought toward the model's dominant pre-training language.

In response, this work aims to establish and explain Cross-lingual Collapse under RLVR: we corroborate the phenomenon across translated long-CoT settings, identify its causal drivers and triggering conditions, and examine how it can be mitigated and to what extent.

3 Experiment

3.1 Experimental Settings

Base models. To investigate the influence of foundation model design on reasoning in a target language, we categorized base models into two groups: (1) english-dominant LLMs, (2) non-english dominant LLMs. We selected OLmo2-1B Instruct as an english-dominant LLM (OLMo et al., 2024), Llama-3.2 3B Instruct (Grattafiori et al., 2024) and Qwen-2.5 1.5B Instruct(Team, 2024) as representative non-english dominant LLMs. [1] This setup allows us to investigate how the intrinsic prior of languages shape the emergence of non-english reasoning abilities when the models are prompted to reason in a variety of language.

Training configuration. To enhance the reasoning capability of LLMs, we train the base models with GRPO, a representative RL-from-verification (RLVR) algorithm shown to strengthen chain-of-thought reasoning. We used GSM8K training dataset, the community's most widely utilized dataset for mathematical word problems (Shao et al., 2024; DeepSeek-AI et al., 2025). Training was conducted within a verl framework (Sheng et al., 2024), using a slightly modified hyperparameter configuration from the SimpleRL project Zeng et al. (2025), which are proven effective for this task. To assess how these reasoning ability improved in a trained language, we translated the entire training corpus into Korean (KO), Ukrainian (UK), Chinese (ZH) using GPT-4o. The quality of the translated data was ensured using quality filtering Guerreiro et al. (2024), as detailed in Appendix B.We excluded 15% of training dataset for validation. Additionally, in order to ablate the model's training dynamics under the challenging dataset, we sampled 7K dataset from the SimpleRL-Zoo dataset (Zeng et al., 2025) with various difficulty and its translated dataset as more challenging math dataset than GSM8K.

[1] Our classification is based on the models' technical reports and cards in Huggingface. The OLMo 2 report only focuses on its English performance, having been trained predominantly on English data. Conversely, the reports for Qwen-2.5 and Llama-3.2 explicitly detail their multilingual capabilities.

Evaluation dataset. We evaluated our model on the translated GSM8K and MATH500 Lightman et al. (2024) test sets across multiple languages. In order to compute the accuracy, we utilize math-verify library [2] for obtaining robust mathematical expression.

Target Word Ratio (Target WR). To assess whether GRPO training preserves input-output language consistency, We computed both the word ratio. We first remove all LaTeX expressions (e.g., `$...$`, `\begin{...}`, `\end{...}`) from the model's output. The remaining text is tokenized using simple regular-expression rules, using Multi-bleu [3], so that punctuation, brackets, and quotes are properly separated. Tokens that consist purely of math expressions, special symbols, or backslash commands are discarded. For each remaining token, we examine its characters to determine whether they belong exclusively to one of several script ranges, such as Hangul (U+AC00–U+D7A3), Latin alphabets (A–Z, a–z), CJK characters (U+4E00–U+9FFF, etc.), or Cyrillic (U+0400–U+04FF). We calculate the *Target word ratio* of a given language by dividing its token count by the total token count. Any token that mixes English letters with another script is labeled as a code-switching token, whose ratio is similarly tracked. This uniform preprocessing and detection pipeline thus enables a quantitative assessment of how models maintain linguistic fidelity in multilingual output. Additionally, we also denote English word ratio as EN WR.

3.2 Experimental Verifications of Cross-lingual Collapse

For the main study, to verify Cross-lingual Collapse and establish its baseline behavior, we examine how GRPO-trained models behave on mathematical benchmarks in terms of both *accuracy* and *language fidelity*. Table 1 summarizes results across languages and base models.

Accuracy. Finetuning LLMs with GRPO generally increases accuracy on our mathmatical benchmarks, though the size of the gain depends on language, backbone models and evaluation dataset. On the translated GSM8K, fine-tuning Llama-3.2-3B Instruct with GRPO raises accuracy by $+7.4$ pp in Chinese, $+14.5$ pp in Korean, and $+17.1$ pp in Ukrainian. MATH500 shows the same upward trend (e.g. $+12.0$ pp for Ukrainian). We can see the same improvement across the models. Interestingly, we observed a significant improvement across all OLMo2 cases and in the Llama model fine-tuned on a Ukrainian dataset. Additionally, Table 6 reports English-language accuracy for each trained model, revealing a persistent gap between English and target-language reasoning even when training is conducted in the target language.

Language fidelity. The accuracy gains come at the cost of target-language fidelity. This pattern is exemplified by OLMo 2. For example, the improvement of target Accuracy in GSM8k in each languages is above -34.0, and target word rate is dropped almost -75.0. In open-source multilingual LLMs, for high-resource Chinese, the target-language word ratio stays above 90% (a modest -1.4 pp drift). For mid-resource Korean the drop is larger (-8.1 pp), while for low-resource Ukrainian the collapse is catastrophic: the share of Ukrainian tokens plummets from 98% to 0.3% on GSM8K (-97.6 pp change) and to 0.9% on MATH500. On the other hand, when we track the English Word Ratio (EN WR) in the table, it moves in the opposite direction: wherever Target WR declines, the English Word Ratio (EN WR) in Table 1 increases where Target WR decreases, indicating that the model's reasoning trace increasingly shifts into English as training progresses.

From above trends, these trends reveal a clear trade-off between accuracy and language fidelity under RLVR: accuracy rises while Target WR falls and EN WR rises. We refer to this joint pattern as **Cross-lingual Collapse**—the chain of thought reverts to the pre-training dominant language as reasoning performance increases.

3.3 Triggering Cross-lingual collapse

Building on the trade-off established above, we now unpack *how* the collapse is mechanistically induced, *when* it emerges during training, and *where* it shows up beyond mathematics.

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

[3] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

Table 2: Harder training triggers Cross-lingual Collapse in Korean. Qwen-2.5-1.5B Instruct trained on Korean GSM8K alone (*Base*, 1K/2K) preserves target-language fidelity, whereas mixing SimpleRL-Zoo (*Base+Hard*, 2K) collapses Korean word ratio (Target WR) to 14.5%(GSM8K) and 2.1% (MATH500), with accuracy rising to 47.5% and 46.7%. On GSM8K, English word ratio (EN WR) also increases, indicating drift toward English.

Dataset	Steps	GSM8K (KO)			MATH500 (KO)		
		Accuracy (%)	Target WR (%)	EN WR (%)	Accuracy (%)	Target WR (%)	EN WR (%)
Base	1K	42.3	94.3	3.1	25.7	88.0	10.1
	2K	43.1	94.0	3.6	27.1	86.5	10.9
Base + Hard	2K	47.5	**14.5**	80.1	46.7	**2.1**	87.4

Table 3: Global MMLU-Lite (KO) accuracy and Korean word ratio (Target WR) of CoT outputs for Qwen-2.5-1.5B Instruct trained on GSM8K (KO) under three settings: **Base** (GSM8K only), **Base (w/ Lang loss)** (GSM8K + language-consistency reward), and **Base + Hard** (GSM8K + SimpleRL-Zoo hard curriculum). The hard-curriculum variant achieves the highest accuracy but shows the language drift (lowest Target WR).

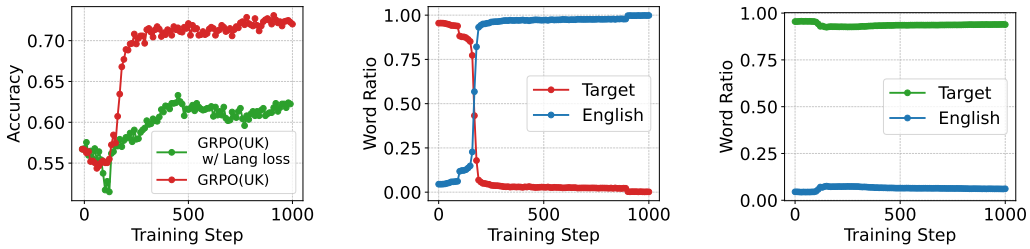
	Base	**Base (w/ Lang loss)**	**Base + Hard**
Global MMLU	31.5	31.0	33.4
Target Word ratio	71.6	75.2	23.4
English Word ratio	27.7	20.3	68.3

Difficulty triggers collapse. To validate our assumption that Harder problems trigger and accelerate Cross-lingual Collapse even in mid-resource languages, . Mixing SimpleRL-Zoo into the replay buffer widens the success gap between English and the target language and the policy quickly exploits the higher-yield English pathway. Concretely, for Qwen-2.5-1.5B trained on the *Korean* translation, keeping GSM8K only preserves target-language fidelity after 2K updates (Target WR: GSM8K 94.0%, MATH500 86.5%; Table 2). Introducing the harder SimpleRL-Zoo subset collapses the chain-of-thought into English by 2K steps: Target WR falls to **14.5%** on GSM8K (-79.5 pp) and to **2.1%** on MATH500 (-84.4 pp), while accuracy *rises* to 47.5% on GSM8K ($+4.4$ pp) and 46.7% on MATH500 ($+19.6$ pp).

Cross-lingual Collapse is initiated during exploration at rollout generation. Advantage-weighted credit under a correctness-only reward systematically favors English reasoning trajectories, creating a self-reinforcing drift. Figure 4 illustrates for Qwen-2.5-1.5B on Korean GSM8K: exploration often uncovers English CoT continuations that solve the problem more reliably than staying in the target language. Each time such an off-target (English) trajectory succeeds, its advantage is positive, increasing the log-probability of its tokens and shifting future rollouts toward English-Target WR declines while English WR increases. The resulting regime shift—English traces dominating despite non-English prompts—constitutes the rollout-level mechanism behind Cross-lingual Collapse and foreshadows the accuracy jump and fidelity drop observed under harder curricula and high-entropy decoding.

Beyond math: domain-general drift. The other question is whether Cross-lingual Collapse is confined to the mathematical reasoning domain or is a general phenomenon. To investigate this, we evaluated trained models on the Korean question and answer pairs of Global MMLU-Lite (Singh et al., 2024). Specifically, we evaluate three fine-tuning variants of the Qwen2.5-1.5B Instruct model: (1) training with GSM8K (KO), (2) training with GSM8K with a language-consistency loss (Lang loss), and (3) a cross-lingual-collapse setting training with GSM8K and a hard-curriculum dataset (GSM8K + SimpleRL).

As shown in Table 3, the results show a pattern similar to our primary findings on mathematical benchmarks. The cross-lingual-collapse model, fine-tuned with the harder curriculum (GSM8K + SimpleRL), not only achieves the highest performance on MMLU-Lite but also suffers the most severe language drift, with the Korean token ratio in its outputs falling to 23.4%. Conversely, adding the language-consistency reward (Lang loss) preserves a higher Korean token ratio (75.2%) at the



(a) GSM8K Accuracy (UK) (b) Word Ratio (UK) (c) Word Ratio (UK) w/ Lang loss

Figure 2: Figures 2a–2c compare Llama-3.2-3B Instruct trained with GRPO on the *Ukrainian*-translated GSM8K with and without the language-consistency reward (Lang loss). The language-consistency reward reliably preserves the target-language word ratio, yet it also *dampens* the accuracy gains that GRPO would otherwise deliver. In particular, Figures 2a–2c show that the reward almost completely prevents cross-lingual collapse in the Ukrainian run—though at the cost of a modest drop in performance

cost of a minor dip in performance (31.0). This demonstrates that the trade-off between task accuracy and linguistic fidelity is not confined to mathematics; rather, the pressure to revert to English reasoning for performance gains appears to be a domain-general effect that also holds for general-knowledge tasks.

3.4 Mitigating cross-lingual Collapse

Our analyses in §3.3 indicate that cross-lingual collapse is driven by a language-agnostic (accuracy-only) verification reward and exploratory rollouts that discover and reinforce dominant-language reasoning. This observation suggests three complementary mitigation ideas that act at different: (1) **reward shaping** to inject language fidelity into the objective itself; (2) **rollout sampling controls** that constrain exploration so English-only trajectories are less accessible during rollouts; and (3) **training with mixture of multiple languages** that regularize the model's internal arbitration across languages by aligning training with a more balanced linguistic prior.

Language consistency reward. Following DeepSeek-AI et al. (2025), we augment the verification reward with an auxiliary signal that favors target-language CoT tokens. As summarized in Figure 2, we add additional reward in which Llama-3.2-3B is training with GRPO on the Ukrainian GSM8K, once with the language-consistency reward and once without it. In the vanilla setting (Figures 2a–2c, solid line) the model undergoes a full cross-lingual collapse: the share of Ukrainian tokens in its chain of thought drops to almost zero while accuracy rises sharply. Adding the language-consistency reward (dashed line) prevents that collapse—the Ukrainian word ratio stays high—yet the accuracy gain is noticeably smaller. This shows that forcing GRPO to keep the reasoning trace in the target language safeguards linguistic fidelity at the cost of some performance.

These results suggest that during GRPO the model actively probes alternative reasoning paths and, when allowed, gravitates toward high-resource English to maximize reward. Constraining the trace to a non-english language blocks that shortcut, preserving the intended language but sacrificing part of the accuracy gain.

Adjusting rollout sampling parameters. Our experiments reveal a consistent dominant-language reversion in chain-of-thought: even under target-language prompts, the word ratio briefly rises and then abruptly flips to the pre-training dominant language (English), coinciding with a sharp accuracy jump—what we term Cross-lingual Collapse. This pattern suggests that reward optimization exploits English as a higher-yield reasoning path in English-centric LLMs. In light of evidence that general language confusion peaks at high-entropy Marchisio et al. (2024), large-nucleus decoding points and is partially mitigated by lowering temperature and nucleus size, we posit that collapse is a sampling-gated manifestation of the same bias: structural but partially controllable at inference.

Table 4: Impact of rollout entropy on Llama3.2-3B with GSM8k(UK) through adjusting top p (Top P) and temperature (Temp.) parameters. The default high-entropy setting (`top_p=1.0`, `Temp=1.0`) maximizes accuracy by allowing the model to revert to high-yield English reasoning paths. Restricting the decoding space with lower `top_p` or `temperature` effectively prevents this language drift, but at the cost of a 5–12 percentage-point drop in accuracy.

Top P	Temp.	GSM8K (UK)			MATH500 (UK)		
		Accuracy (%)	Target WR (%)	EN WR(%)	Accuracy (%)	Target WR (%)	EN WR(%)
1.0	1.0	70.9	0.3	96.8	47.6	5.6	93.4
0.8	1.0	64.2	81.9	11.2	35.8	83.2	15.5
0.6	1.0	63.5	80.6	15.0	36.1	82.5	14.5
1.0	0.8	65.6	81.2	16.0	37.4	81.0	16.9

As shown in Table 4, reducing temperature or top-p attenuates reversion for Llama-3.2-3B on Ukrainian, though stabilized runs still trail the adding a language consistency reward.

Table 5: Effect of multilingual GRPO training with mix of languages. We train Llama-3.2-3B Instruct on GSM8K with three mixes—UK only, UK+KO, and UK+KO+ZH+EN—and evaluate on Ukrainian GSM8K and MATH500, reporting accuracy and the Target Word Ratio. Adding Korean alone leaves the model collapsed (near-zero Target WR), whereas a four-language mix largely restores Ukrainian CoT but lowers accuracy.

Languages	GSM8K(UK)			MATH500(UK)		
	Accuracy (%)	Target WR (%)	EN WR(%)	Accuracy (%)	Target WR (%)	EN WR(%)
UK	70.9	0.3	96.8	47.6	5.6	93.4
UK, KO	72.1	0.0	98.7	42.0	6.9	91.7
UK, KO, ZH, EN	63.5	79.6	19.0	33.2	77.5	17.1

Training with multiple languages. Prior work shows that adding a small set of languages during instruction tuning is more effective than monolingual insturuction tuning (Kew et al., 2024; Chen et al., 2024b; Shaham et al., 2024). We test whether the same idea mitigates Cross-lingual Collapse under RLVR framework. Concretely, we train Llama-3.2-3B Instruct with GRPO on three GSM8K training mixes: (1) Ukrainian only (UK), (2) bilingual (UK+KO), and (3) four-language (UK+KO+ZH+EN). We then evaluate on Ukrainia GSM8K and Ukrainian MATH500, reporting accuracy and the target word ratio of Ukrainian.

As shown in Table 5, adding a single additional language (UK+KO) leaves the model in a collapsed regime on GSM8K. In contrast, training on four languages largely *restores* input–output language consistency on Ukrainian (Target WR $\approx 80\%$ on both test sets), but it *reduces* accuracy relative to the collapsed Ukrainian only (GSM8K: -7.4 pp; MATH500: -14.4 pp). Thus, multilingual training acts as a crude regularizer against collapse, but introduces a pronounced performance–fidelity trade-off, making it a suboptimal mitigation compared to targeted interventions such as a language-consistency reward and rollout sampling controls.

4 Discussion

4.1 Cross-lingual Collapse

The evidence assembled so far paints a coherent picture: (1) Universal Drift. GRPO pushes all models toward the dominant pre-training language, but the speed and severity of that drift scale with resource level: minimal in Chinese, moderate in Korean, catastrophic in Ukrainian (Table 1). (2) Difficulty as a Trigger. A mid-resource model that is stable on GSM8K alone collapses after we inject a harder curriculum (Table 2), showing that *task difficulty*, tilts the optimizer toward English reasoning. (3) Reward Design Matters, but Costs Accuracy. The three mitigate algorithms partially prevents collapse (Figure 2) yet removes much of GRPO's accuracy gain, implying that the model *strategically* chooses English traces to maximize reward under pressure.

These findings confirm our central claim: GRPO amplifies the linguistic prior that best optimizes reward, and the gap between high- and lower-resource languages widens as tasks grow harder.

4.2 Future research direction

Building on the identification and analysis in Sec. 3.2and Sec. 3.3, we designed and evaluated several mitigation algorithms; nevertheless, important limitations persist. Taken together, the experimental results in Sec. 3.4 motivate three research questions to guide future work.

Persistent accuracy–fidelity trade-off. Lowering rollout entropy (e.g., via temperature or top-p) curbs cross-lingual collapse but also suppresses exploration and hurts accuracy, while higher-entropy sampling does the opposite. This aligns with evidence that broad, diversified search improves reasoning when paired with multi-sample selection or structured exploration—e.g., self-consistency voting and tree-structured search (Wang et al.; Yao et al., 2023)—and with maximum-entropy principles in reinforcement learning that stabilize learning via entropy regularization (Haarnoja et al., 2018; Cui et al., 2025). At the same time, high entropy increases language confusion in multilingual models (Marchisio et al., 2024). A promising direction is therefore to redesign exploration mechanism to keep exploration broad in the semantic space while constraining surface form to the target language.

Drift is merely incidental or actually the optimizer's "best path" under current objectives. Our findings are consistent with a reward-shortcut hypothesis under RLVR: high-yield English trajectories discovered during exploration receive positive advantage and become reinforced (Shao et al., 2024; DeepSeek-AI et al., 2025). Rather than fixing a global weight on language fidelity, we propose casting training as constrained or multi-objective RL that explicitly traces the Pareto frontier between accuracy and target-language consistency. Adaptive Lagrangian or primal–dual methods can strengthen the constraint when early warning signals (e.g., a drop in target-language ratio) are detected and relax it otherwise, aiming to block the English shortcut without needlessly sacrificing performance.

Reconsidering the purpose of interpretable CoT in multilingual settings. When, if ever, is it acceptable to sacrifice on-language reasoning traces to gain accuracy, and what do we lose in interpretability, auditability, education, and localization when we do? One promising compromise is latent reasoning with target-language summaries: the model reasons internally but must emit concise, on-language plans or explanations for human inspection. Establishing evaluation protocols that jointly reward task accuracy and on-language interpretability will clarify when fidelity should dominate and when performance gains justify off-language traces.

5 Related Works

5.1 Long Chain-of-Thought Generation

DeepSeek-AI et al. (2025) push the envelope on reinforcement-learning–based reasoning by introducing DeepSeek-R1-Zero, the open-source model trained with pure RL, specifically Group-Relative Policy Optimization (GRPO), without any supervised warm-up, and its follow-up DeepSeek-R1, which adds a small cold-start SFT stage and multi-stage RL to further boost performance. Their study demonstrates that large-scale GRPO can elicit impressive gains on mathematics and coding benchmarks, and that the resulting reasoning patterns can be distilled into much smaller dense models. Notably, the authors briefly report undesirable "language mixing" and readability issues that emerge during RL, suggesting that reward-driven optimization may inadvertently disrupt linguistic fidelity. However, DeepSeek-R1 focuses almost exclusively on English prompts and does not quantify the extent, or direction, of its language drift. Our work complements these findings by conducting a systematic, multilingual analysis of GRPO and revealing a pronounced *Cross-lingual Collapse*: as RL progresses, chain-of-thought reasoning reverts to the pre-training-dominant language, catastrophically eroding performance in low-resource languages.

5.2 Multilingual Instruction tuning

Recent work shows that even a pinch of multilingual data during instruction tuning can unlock substantial cross-lingual generalisation in otherwise English-centric LLMs. Shaham et al. (2024) demonstrate that fine-tuning with as few as two to three languages is "necessary and sufficient" to elicit target-language responses across five downstream tasks, with the marginal benefit largely determined by how well that language was covered in pre-training. Complementing this, Kew et al. (2024) find that injecting only 40 non-English instruction–response pairs, or diversifying the tuning mix to merely 2–4 languages, yields instruction-following quality on a par with (or exceeding) monolingual baselines while slashing per-language data by an order of magnitude. Yoo et al. (2024a) demonstrate that incorporating a sufficient amount of code-switched data (combining English and the target language) can effectively adapt an English-centric model, allowing the model to transfer its English-based knowledge into the target. Those studies therefore argue that massive multilingual corpora are not a prerequisite for broad cross-lingual utility; rather, strategically chosen seed languages can act as effective "anchors" that bootstrap transfer to unseen languages. Crucially, neither paper probes how reinforcement-learning–based reasoning objectives interact with this minimalist recipe, leaving open the question of whether such scarce multilingual supervision can withstand the linguistic pressures we observe under GRPO.

5.3 Multilingual Reasoning

Mechanistic analyses show that multilingual LLMs are not language-neutral: logit-lens (Schut et al., 2025) studies find models like Llama-3.1 route concepts through an English-centered space even for non-English prompts, and steering vectors learned in English transfer more robustly; circuit tracing of Claude 3.5 Haiku reveals language-agnostic subcircuits cooperating with language-specific pathways, yet English often dominates when languages compete Lindsey et al. (2025). Building on this asymmetry, two families of methods explicitly leverage English reasoning to boost multilingual performance: (i) pivot-translation approaches translate questions or intermediate steps into English to exploit stronger reasoning priors and tools, then map solutions back to the target language (Zhu et al., 2024; Chen et al., 2024a; Yoon et al., 2024); and (ii) cross-lingual preference alignment aligns step-level choices across languages via preference optimization She et al. (2024). These works chiefly optimize outcomes rather than explain failure modes. In contrast, we identify when and why Cross-lingual Collapse emerges in RL-based reasoning and link it to English-biased latent computation, offering a diagnostic lens complementary to cross-lingual consistency work and clarifying how language-specific reasoning abilities emerge—and sometimes fail—under optimization pressure.

6 Conclusion

This study uncovers and characterizes **Cross-lingual Collapse**: when trained with reinforcement learning with verifiable reward (RLVR) and long chain-of-thought (CoT), large language models (LLMs) increasingly route their reasoning through the pre-training–dominant language as accuracy rises. Across Chinese, Korean, and Ukrainian and multiple backbones, we observe a clear resource-sensitivity gradient—negligible drift in high-resource Chinese, moderate in mid-resource Korean, and severe in low-resource Ukrainian—with English-centric backbones collapsing fastest. Harder curricula and high-entropy rollouts precipitate the shift, and rollout analyses show a correctness-only advantage signal that repeatedly reinforces higher-yield English trajectories. The effect persists beyond mathematics. A language-consistency reward, entropy reduction at rollout time (e.g., lower temperature), and multilingual RLVR all preserve target-language traces to varying degrees, but each incurs a measurable accuracy cost; even broad multilingual mixes largely restore on-language CoT while lowering scores. These results reveal a persistent *performance–fidelity* trade-off. We view this phenomenon as a natural consequence of English-dominant pre-training and argue that securing linguistic diversity during pre-training is a necessary (though not always sufficient) condition for maintaining language fidelity in long CoT settings.

7 REPRODUCE STATEMENT

In order to ensure the reproduce-ability of the project, we describe details hyperparameter configurations and dataset creation pipeline described in Sec. 3.1. We will release the datasets and code ,including configuration files and reproduction scripts, in a public GitHub repository upon publication to enable end-to-end replication of our results.

REFERENCES

Frederic Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno Miguel Guerreiro, Diptesh Kanojia, José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. Findings of the wmt 2023 shared task on quality estimation. In *Eight conference on machine translation*, pp. 629–653. Association for Computational Linguistics, 2023.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7001–7016, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.411. URL https://aclanthology.org/2024.findings-emnlp.411/.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1347–1356, St. Julian's, Malta, March 2024b. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.90/.

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

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

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.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Tannon Kew, Florian Schottmann, and Rico Sennrich. Turning english-centric llms into polyglots: How much multilinguality is needed? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 13097–13124. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-emnlp.766.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6653–6677, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.380. URL https://aclanthology.org/2024.emnlp-main.380/.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious. 2024. URL https://arxiv.org/abs/2501.00656.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *CoRR*, abs/2502.15603, 2025. doi: 10.48550/ARXIV.2502.15603. URL https://doi.org/10.48550/arXiv.2502.15603.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 2304–2317. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.136. URL https://doi.org/10.18653/v1/2024.findings-acl.136.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

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 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10015–10027, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.539. URL https://aclanthology.org/2024.acl-long.539/.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

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. URL https://arxiv.org/abs/2412.03304.

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.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Investigating and scaling up code-switching for multilingual language model pre-training. *arXiv preprint arXiv:2504.01801*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494/.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*, 2024.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=5Xc1ecxO1h.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. Code-switching curriculum learning for multilingual transfer in llms. *arXiv preprint arXiv:2411.02460*, 2024a.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*, 2024b.

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 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7502–7522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.405. URL https://aclanthology.org/2024.acl-long.405/.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025. URL https://arxiv.org/abs/2503.18892.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? *CoRR*, abs/2408.10811, 2024. doi: 10.48550/ARXIV.2408.10811. URL https://doi.org/10.48550/arXiv.2408.10811.

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 (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8411–8423, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.498. URL https://aclanthology.org/2024.findings-acl.498/.

A Main table with English-language accuracy

Table 6: Accuracy and target-language word ratio for models fine-tuned with GRPO on translated GSM8K. We evaluate on the translated GSM8K and MATH500 test sets. Language codes: **EN** = English, **ZH** = Chinese, **KO** = Korean, **UK** = Ukrainian. Model keys: **OLMo 2** = OLMo-2-0425-IB-Instruct, **Llama** = Llama-3.2-3B Instruct, **Qwen** = Qwen-2.5-1.5B Instruct. Numbers in parentheses indicate the change relative to the corresponding non-fine-tuned baseline. Accuracy (Acc) and target-language word ratio (WR) with languages and models arranged as rows.

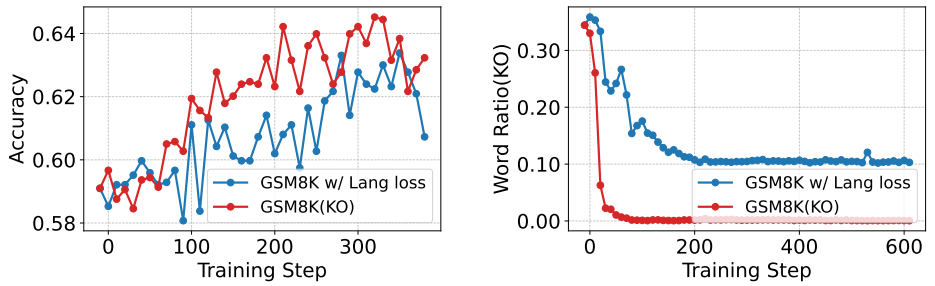
Language	Model	GSM8K			MATH500		
		Target Acc (%)	Target WR (%)	EN Acc (%)	Target Acc (%)	Target WR (%)	EN Acc (%)
ZH	OLMo2	59.8 (+34.3)	0.3 (-75.5)	74.8 (+4.0)	17.6(+1.9)	26.3 (-10.0)	21.4 (+0.7)
	Llama	69.4 (+7.4)	94.1 (-1.4)	83.5 (+3.4)	38.8 (+1.2)	77.5 (-0.4)	50.3 (+1.8)
	Qwen	63.4 (+1.3)	92.9 (+0.6)	77.9 (+4.0)	41.9 (+4.7)	79.8 (+0.4)	55.7 (+7.5)
KO	OLMo2	46.5 (+39.9)	14.3 (-79.4)	73.1 (+2.3)	12.2 (+5.2)	0.1 (-45.1)	22.2 (+1.5)
	Llama	61.3 (+14.5)	82.4 (-8.1)	81.6 (+1.5)	28.5 (+7.2)	70.9 (-17.8)	49.6 (+1.1)
	Qwen	42.2 (+3.5)	96.1 (-2.4)	74.1 (+0.2)	27.0 (+6.8)	80.3 (-12.3)	54.1 (+5.9)
UK	OLMo2	45.2 (+37.8)	0.3 (-75.5)	73.7 (+2.9)	13.0 (+5.6)	29.8 (-57.4)	21.6 (+0.7)
	Llama	70.9 (+17.1)	**0.3** (-97.6)	80.8 (+0.6)	47.6 (+12.0)	**5.6** (-72.7)	51.2 (+1.7)
	Qwen	39.7 (+4.9)	99.3 (+0.5)	75.4 (+1.6)	23.4 (+4.0)	82.8 (-9.8)	51.2 (+3.0)

B Translated Dataset details

To ensure high translation quality, we re-translated the English source with GPT-4, a model that exhibits near-professional performance across many language pairs (Yan et al., 2024; Blain et al., 2023). After each pass we filtered candidates with xCOMET Guerreiro et al. (2024): only hypotheses that exceeded the Flores dev-set mean for the target language were retained; sentences falling below the threshold were re-translated.

C Further training of Distilled LRMs

As depicted in Figure 3, we apply a second round of GRPO to the DeepSeek-R1–Distilled Qwen to test whether continued fine-tuning can correct the entrenched reasoning bias. The results reveal a steep decline in the target-language word ratio, indicating that the phenomenon is difficult to reverse.



(a) Korean GSM8K Performance

(b) Training Step vs. Word Ratio

Figure 3: We continued GRPO fine-tuning of the DeepSeek-R1-Distill Qwen model on the Korean-translated GSM8K dataset to encourage Korean chain-of-thought reasoning. As Figure 3b shows, the distilled model still exhibits cross-lingual collapse during training.

D ROLLOUT GENERATION EXAMPLE

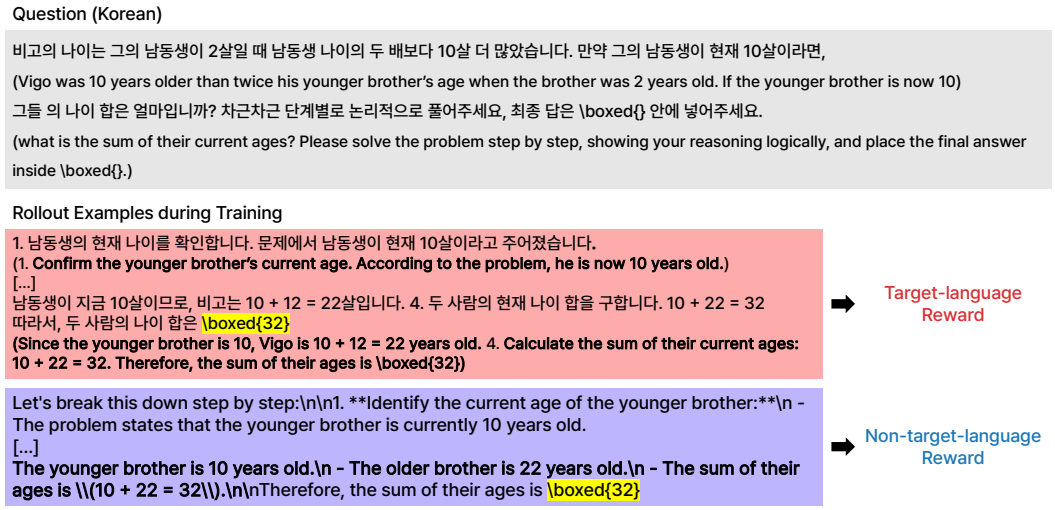


Figure 4: Rollout examples from GRPO training of Qwen-2.5 1.5B on the Korean-translated GSM8K. Observe that the model often arrives at the right answer via English reasoning (*non-target language*); because any correct answer earns full reward, repeated reinforcement of such off-language traces gradually shifts the chain-of-thought word ratio away from Korean.