

# 000 WHEN VERIFIABLE REWARDS SWITCH THE LAN- 001 GUAGE: CROSS-LINGUAL COLLAPSE IN CHAIN-OF- 002 THOUGHT

006 **Anonymous authors**

007 Paper under double-blind review

## 011 ABSTRACT

013 Reinforcement learning with verifiable reward (RLVR) has been instrumental in  
 014 eliciting strong reasoning capabilities from large language models (LLMs) via  
 015 long chains of thought (CoT). During RLVR training, we **formalize and system-**  
 016 **ically study** an empirical phenomenon whereby a multilingual model’s CoT re-  
 017verts to its dominant pre-training language (e.g., English) even when prompted in  
 018 another language, which we term Cross-lingual Collapse. Because the long-CoT  
 019 regime magnifies exposure to linguistic priors, the underlying trade-off between  
 020 maximizing reasoning depth and preserving target-language fidelity has remained  
 021 under-characterized. To examine this trade-off, we train LLMs with Group-  
 022 Relative Policy Optimization (GRPO) on translated versions of math datasets  
 023 widely used to elicit long-CoT reasoning. Throughout training, we track both task  
 024 accuracy and the language consistency of reasoning chains. Our experiments yield  
 025 three findings: (i) under RLVR, CoT in LLMs systematically drifts toward the pre-  
 026 training dominant language as reasoning performance rises; (ii) English-centric  
 027 priors, long-CoT GRPO optimization, task difficulty, and high-entropy decoding  
 028 jointly amplify this drift, and the pattern persists beyond mathematics; and (iii) in-  
 029 terventions that favor target-language traces—via a language-consistency reward,  
 030 decoding-time controls, or more balanced backbones—mitigate collapse but re-  
 031veal a persistent performance-fidelity trade-off.

## 032 1 INTRODUCTION

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

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

050 To systematically analyze this performance-fidelity trade-off, we study target-language reasoning  
 051 under reinforcement learning with verifiable reward (RLVR). We instantiate Group-Relative Policy  
 052 Optimization (GRPO) (Shao et al., 2024) on an English-centric backbone (OLMo et al., 2024) and  
 053 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 et al., 2021),

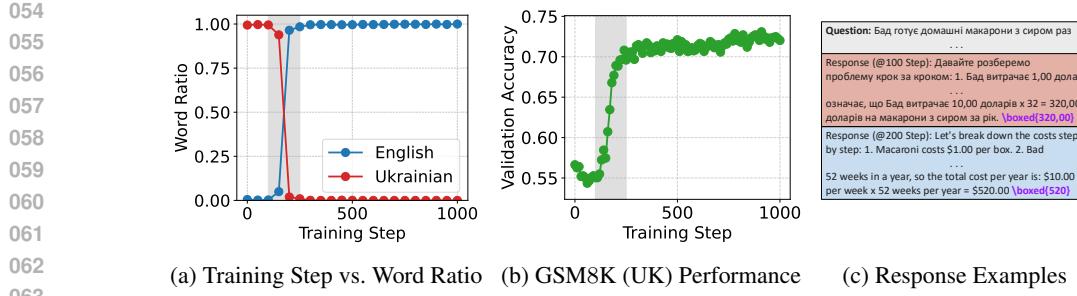


Figure 1: Illustration of **Cross-lingual Collapse**. We train Llama-3.2-3B Instruct with GRPO on a fully Ukrainian translation of GSM8K, seeking Ukrainian-only reasoning. **(a)** Chain-of-thought word-ratio in reward warding rollouts over training steps. In the grey band, the share of Ukrainian tokens plummets, while English abruptly dominates, signaling a language switch within the rollout 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.

SimpleRL-Zoo (Zeng et al., 2025) translated into **five target languages**. 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 formalize and operationalize Cross-lingual Collapse as a phenomenon characterized by rising accuracy with systematic drift of CoT into the pre-training dominant language, quantified via accuracy and word ratios on English and a target language.
- **Amplifiers and mechanisms.** We show that English-dominant LLMs 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 math.
- **Mitigations and limits.** We evaluate **intuitive** interventions (language-consistency reward, decoding controls, and multilingual mixing) that partially alleviate collapse, revealing a **persistent performance–fidelity trade-off** rather than a one-size-fits-all solution.

## 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 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

108 languages, chiefly English, even though the prompts remained Ukrainian. The trend is visualized in  
 109 Figure 1. We dub this behavior **Cross-lingual Collapse** in reasoning models: a systematic collapse  
 110 of target-lingual chains-of-thought toward the model’s dominant pre-training language.  
 111

112 **In response,** this work aims to establish and explain Cross-lingual Collapse under RLVR: we  
 113 corroborate the phenomenon across translated long-CoT settings, identify its causal drivers and trig-  
 114 gering conditions, and examine how it can be mitigated and to what extent.  
 115

### 116 3 EXPERIMENTS

#### 117 3.1 EXPERIMENTAL SETTINGS

118 **Base models.** To investigate the influence of foundation model design on reasoning in a target lan-  
 119 guage, we categorized base models into two groups: (1) english-dominant LLMs, (2) non-English  
 120 dominant LLMs. We selected OLMo2-1B Instruct as an english-dominant LLM (OLMo et al.,  
 121 2024), Llama-3.2 3B Instruct (Grattafiori et al., 2024) and Qwen-2.5 1.5B Instruct(Team, 2024)  
 122 as representative non-English dominant LLMs.<sup>1</sup> This setup allows us to investigate how the in-  
 123 trinsic prior of languages shape the emergence of non-English reasoning abilities when the models are  
 124 prompted to reason in a variety of language.  
 125

126 **Training configuration.** To enhance the reasoning capability of LLMs, we train the base mod-  
 127 els with GRPO, a representative RLVR algorithm shown to strengthen reasoning. We used  
 128 GSM8K training dataset, the community’s most widely utilized dataset for mathematical prob-  
 129 lems (Shao et al., 2024; DeepSeek-AI et al., 2025). Training was conducted within a verl frame-  
 130 work (Sheng et al., 2024), using a slightly modified hyperparameter configuration from the Sim-  
 131 pleRL project Zeng et al. (2025), which are proven effective for this task. To evaluate the improve-  
 132 ment of reasoning ability of a trained language, we translated the entire training corpus into Korean  
 133 (KO), Ukrainian (UK), Chinese (ZH), **Thai(TH)**, **Japanese(JA)** using GPT-4o. These languages span  
 134 distinct scripts and a gradient of pre-training resource levels (high-resource: ZH/JA, mid-resource:  
 135 KO, lower-resource: TH/UK) and are intended as a representative case rather than exhaustive cover-  
 136 age. The quality of the translated data was ensured using quality filtering (Guerreiro et al., 2024), as  
 137 detailed in Appendix B. We excluded 15% of training dataset for validation. Unless otherwise noted,  
 138 we train GRPO models on the translated GSM8K corpus only. In Sec. 3.3, when studying the effect  
 139 of task difficulty, we utilize a mix of a 7k subset of translated SimpleRL-Zoo dataset.  
 140

141 **Evaluation dataset.** We evaluated our model on the translated GSM8K and MATH500 (Lightman  
 142 et al., 2024) test sets across multiple languages. In order to compute the accuracy, we utilize math-  
 143 verify library<sup>2</sup> for obtaining robust mathematical expression.  
 144

145 **Target Word Ratio (Target WR).** To assess whether GRPO training preserves input-output lan-  
 146 guage consistency, we computed the word ratio for both the target language and English. We first  
 147 remove all LaTeX expressions (e.g.,  $\$ \dots \$$ ,  $\backslash begin\{ \dots \}$ ,  $\backslash end\{ \dots \}$ ) from the model’s out-  
 148 put. The remaining text is tokenized using simple regular-expression rules, using Multi-bleu<sup>3</sup>, so that  
 149 punctuation, brackets, and quotes are properly separated. Tokens that consist purely of math expres-  
 150 sions, special symbols, or backslash commands are discarded. For each remaining token, we exam-  
 151 ine its characters to determine whether they belong exclusively to one of several script ranges, such  
 152 as Hangul (U+AC00–U+D7A3), Latin alphabets (A–Z, a–z), CJK characters (U+4E00–U+9FFF,  
 153 etc.), or Cyrillic (U+0400–U+04FF). We calculate the *Target word ratio* of a given language by di-  
 154 viding its token count by the total token count. Any token that mixes English letters with another  
 155 script is labeled as a code-switching token, whose ratio is similarly tracked. This uniform preprocess-  
 156 ing and detection pipeline thus enables a quantitative assessment of how models maintain linguistic  
 157 fidelity in multilingual output. Additionally, we also denote English word ratio as EN WR.  
 158

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

<sup>2</sup><https://github.com/huggingface/Math-Verify>

<sup>3</sup><https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl>

162  
 163 Table 1: Accuracy and target-language word ratio for models fine-tuned with GRPO on translated  
 164 GSM8K. We evaluate on the translated GSM8K and MATH500 test sets. Language codes: **EN** =  
 165 English, **ZH** = Chinese, **KO** = Korean, **UK** = Ukrainian, **TH**: Thai, **JA**: Japanese. **Base Models**:  
 166 **OLMo 2** = OLMo-2-0425-1B-Instruct, **Llama** = Llama-3.2-3B Instruct, **Qwen** = Qwen-2.5-1.5B  
 167 Instruct. **Numbers in parentheses indicate the change relative to these base models. Accuracy (Acc)**  
 168 **and target-language word ratio (WR) are reported for each language-model pair.**

Language	Model	GSM8K			MATH500		
		Target Acc (%)	Target WR (%)	EN WR (%)	Target Acc (%)	Target WR (%)	EN WR (%)
ZH	OLMo2	59.8 ( <b>+34.3</b> )	0.3 ( <b>-75.5</b> )	80.8 ( <b>+73.7</b> )	17.6(+1.9)	26.3 (-10.0)	71.0 ( <b>+8.4</b> )
	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 ( <b>+39.9</b> )	14.3 ( <b>-79.4</b> )	83.5 ( <b>+78.3</b> )	12.2 (+5.2)	0.1 ( <b>-45.1</b> )	73.0 ( <b>+51.3</b> )
	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 ( <b>+37.8</b> )	0.3 ( <b>-75.5</b> )	85.3 ( <b>+79.3</b> )	13.0 ( <b>+5.6</b> )	0.1 ( <b>-52.3</b> )	72.7 ( <b>+56.1</b> )
	Llama	70.9 ( <b>+17.1</b> )	<b>0.3</b> ( <b>-82.6</b> )	96.8 ( <b>+80.8</b> )	47.6 ( <b>+12.0</b> )	<b>5.6</b> ( <b>-72.7</b> )	93.4 ( <b>+73.1</b> )
	Qwen	39.7 (+4.9)	99.3 (+0.5)	0.5 (-0.2)	23.4 (+4.0)	82.8 (-9.8)	9.9 ( <b>+8.5</b> )
TH	OLMo2	29.7 ( <b>+27.2</b> )	1.4 ( <b>-80.3</b> )	90.5 ( <b>+81.0</b> )	9.8 (+5.2)	12.6 (-78.4)	84.7 (+76.7)
	Llama	74.6 ( <b>+16.3</b> )	84.1 ( <b>-10.1</b> )	16.6 (+10.5)	50.9 (+5.8)	75.4 (-8.1)	16.8 (+10.0)
	Qwen	68.3 ( <b>+31.7</b> )	8.3 ( <b>-78.9</b> )	89.9 ( <b>+80.5</b> )	43.8 (+23.9)	13.1 ( <b>-76.3</b> )	83.0 ( <b>+75.5</b> )
JA	OLMo2	52.6 ( <b>+45.4</b> )	2.3 ( <b>-83.9</b> )	97.0 ( <b>+80.1</b> )	17.0 (+2.2)	7.4 (-75.5)	90.3 (+76.9)
	Llama	62.6 (+6.6)	96.4 (-0.6)	3.3 (+0.5)	37.1 (+5.2)	92.1 (+4.1)	5.5 (+1.8)
	Qwen	43.1 (+7.5)	95.8 (-0.7)	2.5 (+1.0)	38.4 (+5.7)	98.9 (-0.4)	1.0 (+0.2)

### 185 3.2 EXPERIMENTAL VERIFICATIONS OF CROSS-LINGUAL COLLAPSE

187 For the main study, to verify Cross-lingual Collapse and analyze its behavior, we examine how  
 188 GRPO-trained models behave on mathematical benchmarks in terms of both *accuracy* and *language*  
 189 *fidelity* across five non-English target languages.

191 **Performance-fidelity trade-off.** Table 1 demonstrates the fine-tuning results across languages  
 192 (Chinese, Korean, Japanese, Thai, and Ukrainian) and backbones, where selected languages repre-  
 193 sent distinct alphabet scripts and different pre-training resource levels. RLVR consistently improves  
 194 target-language accuracy relative to the before fine-tuning (i.e., SFT models). We can also observe  
 195 gains in English accuracy, which remains noticeably higher than that of the target languages (Ta-  
 196 ble 6). However, these improvements in the target language often come at the cost of target-language  
 197 fidelity. The English-dominant LLM, OLMo2-1B, shows the sharpest trade-off across languages.  
 198 RLVR training yields large accuracy gains but drives the Target word ratio (Target WR) almost to  
 199 zero and the English Word Ratio (EN WR) toward 80–97%.

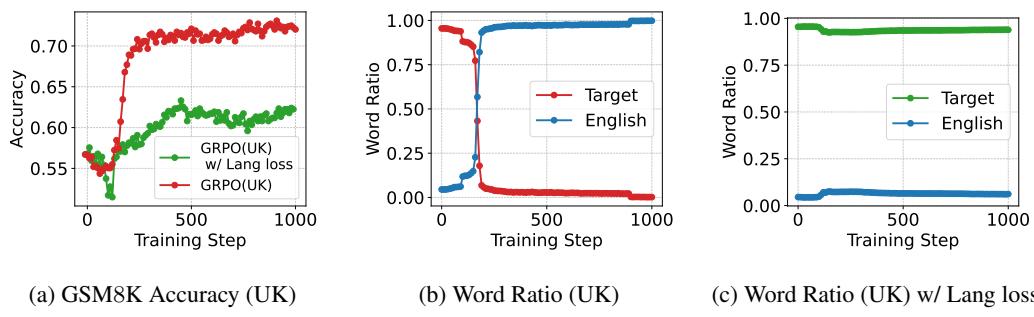
200 On the other hands, multilingual backbones (i.e., Llama-3.2-3B, Qwen-2.5-1.5B) show a resource-  
 201 sensitive pattern: high-resource Chinese and Japanese preserve language fidelity (Target WR  $\geq$  92%,  
 202 EN WR  $\leq$  8%), mid-resource Korean shows a moderate drop in Target WR. Low-resource Thai  
 203 and Ukrainian exhibit distinctive behavior, with significant language drift occurring inconsistently  
 204 across backbones. Even where Target WR remains high, EN WR is often non-trivial ( $\sim$ 15–20%),  
 205 which our qualitative analysis traces to short English scaffolding and Latin-script technical tokens  
 206 embedded in target-language CoTs (Appendix D). This performance-fidelity trade-off persists at  
 207 larger scales, as detailed in Appendix E, implying that the phenomenon is robust across scales.

208 Beyond word ratios, we also analyze cross-lingual consistency for a collapsed Chinese GRPO  
 209 model. As shown in Appendix F, GRPO increases the proportion of problems that the model solves  
 210 correctly in both English and non-English while reducing cases that are solvable only in English.  
 211 Combined with the language-fidelity trends in Table 1, this supports the interpretation that the ac-  
 212 curacy gains in collapsed models are driven by routing target-language prompts through a stronger  
 213 English reasoning mode, rather than by genuinely improved target-language reasoning.

214 These observations reveal a clear trade-off between accuracy and language fidelity under RLVR:  
 215 accuracy rises while Target WR falls and EN WR rises. We refer to this joint pattern as **Cross-**  
 216 **lingual Collapse**—the chain of thought reverts to the pre-training dominant language.

216  
 217  
 218  
 219  
 220  
 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	<b>14.5</b>	80.1	46.7	<b>2.1</b>	87.4



227  
 228  
 229  
 230  
 231  
 232  
 233  
 234  
 235  
 Figure 2: Figures 2a–2c compare Llama-3.2-3B Instruct trained with GRPO on the *Ukrainian*-  
 236 translated GSM8K with and without the language-consistency reward (Lang loss). The language-  
 237 consistency reward reliably preserves the target-language word ratio, yet it also *dampens* the  
 238 accuracy gains that GRPO would otherwise deliver. In particular, Figures 2a–2c show that the reward  
 239 almost completely prevents cross-lingual collapse in the Ukrainian run—though at the cost of a  
 240 modest drop in performance

### 245 3.3 TRIGGERING CROSS-LINGUAL COLLAPSE

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

248  
 249  
 250 **Difficulty triggers collapse beyond GSM8K.** We verify our hypothesis that problem difficulty ac-  
 251 celerates Cross-lingual Collapse even in mid-resource languages (e.g., KO). We adopt SimpleRL-Zoo  
 252 as a challenging complement to GSM8K. This increased difficulty widens the reasoning between En-  
 253 glish and the target language, causing the policy to quickly converge to the more effective English  
 254 reasoning path. Concretely, for Qwen-2.5-1.5B trained on the *Korean* translation, keeping GSM8K  
 255 only preserves target-language fidelity after 2K updates (Target WR: GSM8K 94.0%, MATH500  
 256 86.5%; Table 2). Introducing the harder SimpleRL-Zoo subset collapses the chain-of-thought into  
 257 English by 2K steps: Target WR falls to 14.5% on GSM8K (−79.5%) and to 2.1% on MATH500  
 258 (−84.4%), while accuracy rises to 47.5% on GSM8K and 46.7% on MATH500. Furthermore, we  
 259 compare SFT approaches in Appendix I. However, SFT approach still show less performance than  
 260 GRPO approach and the same performance-fidelity trade off.

261  
 262 **Cross-lingual Collapse is initiated during exploration at rollout generation.** Advantage-weighted credit under a correctness-only reward systematically favors English  
 263 reasoning trajectories, creating a self-reinforcing drift. Figure 3 illustrates for Qwen-2.5-1.5B on  
 264 Korean GSM8K: exploration often uncovers English CoT continuations that solve the problem  
 265 more reliably than staying in the target language. Each time such an off-target (English) trajectory  
 266 succeeds, its advantage is positive, increasing the log-probability of its tokens and shifting future  
 267 rollouts toward English-Target WR declines while English WR increases. The resulting regime  
 268 shift—English traces dominating despite non-English prompts—constitutes the rollout-level  
 269 mechanism behind Cross-lingual Collapse and foreshadows the accuracy jump and fidelity drop  
 observed under harder curricula and high-entropy decoding.

270  
 271 Table 3: Global MMLU-Lite (KO) accuracy and Korean word ratio (Target WR) of CoT outputs for  
 272 Qwen-2.5-1.5B Instruct trained on GSM8K (KO) under three settings: **Base** (GSM8K only), **Base**  
 273 (**w/ Lang loss**) (GSM8K + language-consistency reward), and **Base + Hard** (GSM8K + SimpleRL-  
 274 Zoo hard curriculum). The hard-curriculum variant achieves the highest accuracy but shows the  
 275 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

281  
 282 **Beyond math: domain-general drift.** The other question is whether Cross-lingual Collapse is  
 283 confined to the mathematical reasoning domain or is a general phenomenon. To investigate this,  
 284 we evaluated trained models on the Korean question and answer pairs of Global MMLU-Lite (Singh  
 285 et al., 2024). Specifically, we evaluate three fine-tuning variants of the Qwen2.5-1.5B Instruct model:  
 286 (1) training with GSM8K (KO), (2) training with GSM8K with a language-consistency loss (Lang  
 287 loss), and (3) a cross-lingual-collapse setting training with GSM8K and a hard-curriculum dataset  
 288 (GSM8K + SimpleRL).

289 As shown in Table 3, the results show a pattern similar to our primary findings on mathematical  
 290 benchmarks. The cross-lingual-collapse model, fine-tuned with the harder curriculum (GSM8K +  
 291 SimpleRL), not only achieves the highest performance on MMLU-Lite but also suffers the most  
 292 severe language drift, with the Korean token ratio in its outputs falling to 23.4%. Conversely, adding  
 293 the language-consistency reward (Lang loss) preserves a higher Korean token ratio (75.2%) at the  
 294 cost of a minor dip in performance (31.0). This demonstrates that the trade-off between task  
 295 accuracy and linguistic fidelity is not confined to mathematics; rather, the pressure to revert to En-  
 296 glish reasoning for performance gains appears to be a domain-general effect that also holds for  
 297 general-knowledge tasks.

### 298 3.4 MITIGATING CROSS-LINGUAL COLLAPSE

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

307 **Language consistency reward.** Following DeepSeek-AI et al. (2025), we augment the verifi-  
 308 cation reward with an auxiliary signal that favors target-language CoT tokens, [as detailed in Appendix H](#).  
 309 As shown in Figure 2, we add additional reward in which Llama-3.2-3B is training with  
 310 GRPO on the Ukrainian GSM8K, once with the language-consistency reward and once without it. In  
 311 the vanilla setting (Figures 2a–2c, solid line) the model undergoes a full cross-lingual collapse: the  
 312 share of Ukrainian tokens in its chain of thought drops to almost zero while accuracy rises sharply.  
 313 Adding the language-consistency reward (dashed line) prevents that collapse—the Ukrainian word  
 314 ratio stays high—yet the accuracy gain is noticeably smaller. This shows that forcing GRPO to keep  
 315 the reasoning trace in the target language safeguards linguistic fidelity at the cost of some per-  
 316 formance. [Furthermore, we compare SFT approaches in Appendix I. However, SFT baselines yield](#)  
 317 [lower performance than GRPO, reinforcing the inherent trade-off between accuracy and fidelity.](#)

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

322 **Adjusting rollout sampling parameters.** Our experiments reveal a consistent dominant-language  
 323 reversion in chain-of-thought: even under target-language prompts, the word ratio briefly rises and

324  
 325 Table 4: Impact of rollout entropy on Llama3.2-3B with GSM8k(UK) through adjusting top  
 326 p (Top P) and temperature (Temp.) parameters. The default high-entropy setting ( $\text{top\_p}=1.0$ ,  
 327  $\text{Temp}=1.0$ ) maximizes accuracy by allowing the model to revert to high-yield English reason-  
 328 ing paths. Restricting the decoding space with lower  $\text{top\_p}$  or temperature effectively prevents  
 329 this language drift, but at the cost of a 5–12 percentage-point drop in accuracy.

330 Top P	331 Temp.	332 GSM8K (UK)			333 MATH500 (UK)		
		334 Accuracy (%)	335 Target WR (%)	336 EN WR(%)	337 Accuracy (%)	338 Target WR (%)	339 EN WR(%)
332 1.0	333 1.0	334 70.9	335 0.3	336 96.8	337 47.6	338 5.6	339 93.4
332 0.8	333 1.0	334 64.2	335 81.9	336 11.2	337 35.8	338 83.2	339 15.5
332 0.6	333 1.0	334 63.5	335 80.6	336 15.0	337 36.1	338 82.5	339 14.5
332 1.0	333 0.8	334 65.6	335 81.2	336 16.0	337 37.4	338 81.0	339 16.9

340  
 341 then abruptly flips to the pre-training dominant language (English), coinciding with a sharp accuracy  
 342 jump—what we term Cross-lingual Collapse. This pattern suggests that reward optimization exploits  
 343 English as a higher-yield reasoning path in English-centric LLMs. In light of evidence that general  
 344 language confusion peaks at high-entropy (Marchisio et al., 2024), large-nucleus decoding points  
 345 and is partially mitigated by lowering temperature and nucleus size, we posit that collapse is a  
 346 sampling-gated manifestation of the same bias: structural but partially controllable at inference.

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

349 Table 5: Effect of multilingual GRPO training with mix of languages. We train Llama-3.2-3B In-  
 350 struct on GSM8K with three mixes—UK only, UK+KO, and UK+KO+ZH+EN—and evaluate on  
 351 Ukrainian GSM8K and MATH500, reporting accuracy and the Target word ratio. Adding Korean  
 352 alone leaves the model collapsed (near-zero Target WR), whereas a four-language mix largely re-  
 353 stores Ukrainian CoT but lowers accuracy.

354 Languages	355 GSM8K(UK)			356 MATH500(UK)		
	357 Accuracy (%)	358 Target WR (%)	359 EN WR(%)	360 Accuracy (%)	361 Target WR (%)	362 EN WR(%)
354 UK	355 70.9	356 0.3	357 96.8	358 47.6	359 5.6	360 93.4
354 UK, KO	355 72.1	356 0.0	357 98.7	358 42.0	359 6.9	360 91.7
354 UK, KO, ZH, EN	355 63.5	356 79.6	357 19.0	358 33.2	359 77.5	360 17.1

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

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

## 377 4 DISCUSSION

### 378 4.1 CROSS-LINGUAL COLLAPSE

379 The evidence assembled so far paints a coherent picture: (1) Universal Drift. GRPO pushes all  
 380 models toward the dominant pre-training language, but the speed and severity of that drift scale with

378 resource level based on the prior work (Wenzek et al., 2020): minimal in high-resource (e.g., ZH,  
 379 JA), moderate in mid-resource (e.g., KO), catastrophic in low-resource (e.g., TH, UK) (Table 1). (2)  
 380 Difficulty as a Trigger. A mid-resource model that is stable on GSM8K alone collapses after we  
 381 inject a harder curriculum (Table 2), showing that *task difficulty*, tilts the optimizer toward English  
 382 reasoning. (3) Reward Design Matters, but Costs Accuracy. Mitigate algorithms partially prevent  
 383 collapse (Figure 2) yet remove much of GRPO’s accuracy gain, implying that the model *strategically*  
 384 chooses English traces to maximize reward under pressure.

385 These findings confirm our central claim: GRPO amplifies the linguistic prior that best optimizes  
 386 reward, and the gap between high- and lower-resource languages widens as tasks grow harder.  
 387

#### 388 4.2 FUTURE RESEARCH DIRECTION

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

393 **Persistent accuracy-fidelity trade-off.** Lowering rollout entropy (e.g., via temperature or top- $p$ )  
 394 curbs cross-lingual collapse but also suppresses exploration and hurts accuracy, while higher-entropy  
 395 sampling does the opposite. This aligns with evidence that broad, diversified search improves rea-  
 396 soning when paired with multi-sample selection or structured exploration—e.g., self-consistency  
 397 voting and tree-structured search (Wang et al.; Yao et al., 2023)—and with maximum-entropy prin-  
 398 ciples in reinforcement learning that stabilize learning via entropy regularization (Haarnoja et al.,  
 399 2018; Cui et al., 2025). At the same time, high entropy increases language confusion in multilingual  
 400 models (Marchisio et al., 2024). A promising direction is therefore to redesign exploration mecha-  
 401 nism to keep exploration broad in the semantic space while constraining surface form to the target  
 402 language.

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

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

## 421 5 RELATED WORKS

### 423 5.1 LONG CHAIN-OF-THOUGHT GENERATION

425 DeepSeek-AI et al. (2025) push the envelope on reinforcement-learning-based reasoning by in-  
 426 troducing DeepSeek-R1-Zero, the open-source model trained with pure RL, specifically Group-  
 427 Relative Policy Optimization (GRPO), without any supervised warm-up, and its follow-up  
 428 DeepSeek-R1, which adds a small cold-start SFT stage and multi-stage RL to further boost per-  
 429 formance. Their study demonstrates that large-scale GRPO can elicit impressive gains on mathematics  
 430 and coding benchmarks, and that the resulting reasoning patterns can be distilled into much smaller  
 431 dense models. Notably, the authors briefly report undesirable “language mixing” and readability is-  
 432 sues that emerge during RL, suggesting that reward-driven optimization may inadvertently disrupt

432 linguistic fidelity. However, DeepSeek-R1 focuses almost exclusively on English prompts and does  
 433 not quantify the extent, or direction, of its language drift. Our work complements these findings by  
 434 conducting a systematic, multilingual analysis of GRPO and revealing a pronounced *Cross-lingual*  
 435 *Collapse*: as RL progresses, chain-of-thought reasoning reverts to the pre-training-dominant lan-  
 436 guage, catastrophically eroding performance in low-resource languages.

## 437 5.2 MULTILINGUAL INSTRUCTION TUNING

440 Recent work shows that even a pinch of multilingual data during instruction tuning can unlock  
 441 substantial cross-lingual generalization in otherwise English-centric LLMs. Shaham et al. (2024)  
 442 demonstrate that fine-tuning with as few as two to three languages is “necessary and sufficient”  
 443 to elicit target-language responses across five downstream tasks, with the marginal benefit largely  
 444 determined by how well that language was covered in pre-training. Complementing this, Kew et al.  
 445 (2024) find that injecting only 40 non-English instruction-response pairs, or diversifying the tun-  
 446 ing mix to merely 2–4 languages, yields instruction-following quality on a par with (or exceeding)  
 447 monolingual baselines while slashing per-language data by an order of magnitude. Yoo et al. (2024a)  
 448 demonstrate that incorporating a sufficient amount of code-switched data (combining English and  
 449 the target language) can effectively adapt an English-centric model, allowing the model to transfer  
 450 its English-based knowledge into the target. Those studies therefore argue that massive multilingual  
 451 corpora are not a prerequisite for broad cross-lingual utility; rather, strategically chosen seed lan-  
 452 guages can act as effective “anchors” that bootstrap transfer to unseen languages. Crucially, neither  
 453 paper probes how reinforcement-learning-based reasoning objectives interact with this minimalist  
 454 recipe, leaving open the question of whether such scarce multilingual supervision can withstand the  
 455 linguistic pressures we observe under GRPO.

## 456 5.3 MULTILINGUAL REASONING

457 Mechanistic analyses show that multilingual LLMs are not language-neutral: logit-lens (Schut et al.,  
 458 2025) studies find models like Llama-3.1 route concepts through an English-centered space even for  
 459 non-English prompts, and steering vectors learned in English transfer more robustly; circuit trac-  
 460 ing of Claude 3.5 Haiku reveals language-agnostic subcircuits cooperating with language-specific  
 461 pathways, yet English often dominates when languages compete (Lindsey et al., 2025). Building on  
 462 this asymmetry, two families of methods explicitly leverage English reasoning to boost multilingual  
 463 performance: (i) pivot-translation approaches translate questions or intermediate steps into English  
 464 to exploit stronger reasoning priors and tools, then map solutions back to the target language (Zhu  
 465 et al., 2024; Chen et al., 2024a; Yoon et al., 2024); and (ii) cross-lingual preference alignment aligns  
 466 step-level choices across languages via preference optimization (She et al., 2024). These works  
 467 chiefly optimize outcomes rather than explain failure modes. In contrast, we identify when and why  
 468 Cross-lingual Collapse emerges in RL-based reasoning and link it to English-biased latent computa-  
 469 tion, offering a diagnostic lens complementary to cross-lingual consistency work and clarifying how  
 470 language-specific reasoning abilities emerge—and sometimes fail—under optimization pressure.

## 471 6 CONCLUSION

473 This study uncovers and characterizes **Cross-lingual Collapse**: when trained with reinforcement  
 474 learning with verifiable reward and long chain-of-thought, large language models (LLMs) increas-  
 475 ingly route their reasoning through the pre-training-dominant language as accuracy rises. **Across five**  
 476 **target languages and multiple backbones, we observe a clear resource-sensitivity gradient: negligible**  
 477 **drift in high-resource Chinese/Japanese, moderate in mid-resource Korean, and severe collapse in**  
 478 **low-resource Thai/Ukrainian, with English-centric backbones collapsing fastest.** The effect persists  
 479 beyond mathematics. A language-consistency reward, entropy reduction at rollout time (e.g., lower  
 480 temperature), and multilingual RLVR all preserve target-language traces to varying degrees, but  
 481 each incurs a measurable accuracy cost; even broad multilingual mixes largely restore on-language  
 482 CoT while lowering scores. These results reveal a persistent *performance-fidelity* trade-off. We view  
 483 this phenomenon as a natural consequence of English-dominant pre-training and argue that securing  
 484 linguistic diversity during pre-training is a necessary (though not always sufficient) condition for  
 485 maintaining language fidelity in long CoT settings.

486 7 REPRODUCE STATEMENT  
487488 In order to ensure the reproduceability of the project, we describe details hyperparameter config-  
489 urations and dataset creation pipeline described in Sec. 3.1. We will release the datasets and code  
490 ,including configuration files and reproduction scripts, in a public GitHub repository upon publica-  
491 tion to enable end-to-end replication of our results.  
492493 REFERENCES  
494495 Frederic Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno Miguel Guerreiro, Diptesh Kanodia,  
496 José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. Findings of  
497 the wmt 2023 shared task on quality estimation. In *Eight conference on machine translation*, pp.  
498 629–653. Association for Computational Linguistics, 2023.499 Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. Breaking lan-  
500 guage barriers in multilingual mathematical reasoning: Insights and observations. In Yaser  
501 Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Com-  
502 putational Linguistics: EMNLP 2024*, pp. 7001–7016, Miami, Florida, USA, November 2024a.  
503 Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.411. URL  
504 <https://aclanthology.org/2024.findings-emnlp.411/>.  
505506 Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth  
507 Heafield. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In Yvette  
508 Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics:  
509 EACL 2024*, pp. 1347–1356, St. Julian’s, Malta, March 2024b. Association for Computational  
510 Linguistics. URL <https://aclanthology.org/2024.findings-eacl.90/>.511 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,  
512 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John  
513 Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*,  
514 2021.515 Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen  
516 Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for  
517 reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.  
518519 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,  
520 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,  
521 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao  
522 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,  
523 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,  
524 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,  
525 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang  
526 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai  
527 Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,  
528 Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang,  
529 Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,  
530 Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,  
531 R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye,  
532 Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing  
533 reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.  
534 48550/ARXIV.2501.12948. URL <https://doi.org/10.48550/arXiv.2501.12948>.535 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad  
536 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd  
537 of models. *arXiv preprint arXiv:2407.21783*, 2024.538 Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT  
539 Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection.  
*Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.

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

542

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

548

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

553

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

560

561

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

567

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

576

577 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin,  
 578 Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser  
 579 Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan  
 580 Leike, and Ryan Lowe. Training language models to follow instructions with human feedback.  
 581 In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Ad-  
 582 vances in Neural Information Processing Systems 35: Annual Conference on Neural Infor-  
 583 mation Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December  
 584 9, 2022*, 2022. URL [http://papers.nips.cc/paper\\_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html).

585

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

588

589 Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Mul-  
 590 tilingual instruction tuning with just a pinch of multilinguality. In Lun-Wei Ku, Andre Martins,  
 591 and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL  
 592 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 2304–2317. Associa-  
 593 tion for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.136. URL  
<https://doi.org/10.18653/v1/2024.findings-acl.136>.

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

598 Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen.  
 599 MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference  
 600 optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd*  
 601 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,  
 602 pp. 10015–10027, Bangkok, Thailand, August 2024. Association for Computational Linguis-  
 603 tics. doi: 10.18653/v1/2024.acl-long.539. URL <https://aclanthology.org/2024.acl-long.539/>.

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

607 Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel  
 608 Vila-Suero, Peerat Limkongchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Ray-  
 609 mond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre  
 610 F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis,  
 611 and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in  
 612 multilingual evaluation, 2024. URL <https://arxiv.org/abs/2412.03304>.

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

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

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

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

623 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi,  
 624 Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language  
 625 models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Ad-  
 626 vances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/forum?id=\\_VjQ1MeSB\\_J](https://openreview.net/forum?id=_VjQ1MeSB_J).

627 Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán,  
 628 Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from  
 629 web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri,  
 630 Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mari-  
 631 ani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the*  
 632 *Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May  
 633 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.494/>.

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

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

648 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R  
649 Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In  
650 *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=5Xc1ecx01h>.  
651

652 Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. Code-switching curriculum  
653 learning for multilingual transfer in llms. *arXiv preprint arXiv:2411.02460*, 2024a.  
654

655 Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook  
656 Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. Hyperclova x technical report. *arXiv*  
657 *preprint arXiv:2404.01954*, 2024b.  
658

659 Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo.  
660 LangBridge: Multilingual reasoning without multilingual supervision. In Lun-Wei Ku, Andre  
661 Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association*  
662 *for Computational Linguistics (Volume 1: Long Papers)*, pp. 7502–7522, Bangkok, Thailand, Au-  
663 gust 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.405. URL  
664 <https://aclanthology.org/2024.acl-long.405/>.  
665

666 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong  
667 Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at  
668 scale. *arXiv preprint arXiv:2503.14476*, 2025.  
669

670 Weihsao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-  
671 zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025.  
672 URL <https://arxiv.org/abs/2503.18892>.  
673

674 Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Mu-  
675 rawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual lan-  
676 guage models think in? *CoRR*, abs/2408.10811, 2024. doi: 10.48550/ARXIV.2408.10811. URL  
677 <https://doi.org/10.48550/arXiv.2408.10811>.  
678

679 Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Ques-  
680 tion translation training for better multilingual reasoning. In Lun-Wei Ku, Andre Martins,  
681 and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL*  
682 2024, pp. 8411–8423, Bangkok, Thailand, August 2024. Association for Computational Linguis-  
683 tics. doi: 10.18653/v1/2024.findings-acl.498. URL <https://aclanthology.org/2024.findings-acl.498/>.  
684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702 A MAIN TABLE WITH ENGLISH-LANGUAGE ACCURACY  
703704  
705 Table 6: Accuracy and target-language word ratio for models fine-tuned with GRPO on translated  
706 GSM8K. We evaluate on the translated GSM8K and MATH500 test sets. Language codes: **EN** =  
707 English, **ZH** = Chinese, **KO** = Korean, **UK** = Ukrainian. Model keys: **OLMo 2** = OLMo-2-0425-1B-  
708 Instruct, **Llama** = Llama-3.2-3B Instruct, **Qwen** = Qwen-2.5-1.5B Instruct. Numbers in parentheses  
709 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.  
710

711 Language	712 Model	713 <b>GSM8K</b>			714 <b>MATH500</b>		
		715 Target Acc (%)	716 Target WR (%)	717 EN Acc (%)	718 Target Acc (%)	719 Target WR (%)	720 EN Acc (%)
721 ZH	722 OLMo2	723 59.8 ( <b>+34.3</b> )	724 0.3 ( <b>-75.5</b> )	725 74.8 ( <b>+4.0</b> )	726 17.6(+1.9)	727 26.3 (-10.0)	728 21.4 ( <b>+0.7</b> )
	Llama	69.4 ( <b>+7.4</b> )	94.1 ( <b>-1.4</b> )	83.5 ( <b>+3.4</b> )	38.8 ( <b>+1.2</b> )	77.5 ( <b>-0.4</b> )	50.3 ( <b>+1.8</b> )
	Qwen	63.4 ( <b>+1.3</b> )	92.9 ( <b>+0.6</b> )	77.9 ( <b>+4.0</b> )	41.9 ( <b>+4.7</b> )	79.8 ( <b>+0.4</b> )	55.7 ( <b>+7.5</b> )
729 KO	730 OLMo2	731 46.5 ( <b>+39.9</b> )	732 14.3 ( <b>-79.4</b> )	733 73.1 ( <b>+2.3</b> )	734 12.2 (+5.2)	735 0.1 ( <b>-45.1</b> )	736 22.2 ( <b>+1.5</b> )
	Llama	70.9 ( <b>+17.1</b> )	82.4 ( <b>-8.1</b> )	81.6 ( <b>+1.5</b> )	28.5 ( <b>+7.2</b> )	70.9 ( <b>-17.8</b> )	49.6 ( <b>+1.1</b> )
	Qwen	42.2 ( <b>+3.5</b> )	94.3 ( <b>-2.4</b> )	74.1 ( <b>+0.2</b> )	27.0 ( <b>+6.8</b> )	80.3 ( <b>-12.3</b> )	54.1 ( <b>+5.9</b> )
743 UK	744 OLMo2	745 45.2 ( <b>+37.8</b> )	746 0.3 ( <b>-75.5</b> )	747 73.7 ( <b>+2.9</b> )	748 13.0 ( <b>+5.6</b> )	749 29.8 ( <b>-57.4</b> )	750 21.6 ( <b>+0.7</b> )
	Llama	751 70.9 ( <b>+17.1</b> )	752 0.3 ( <b>-97.6</b> )	753 80.8 ( <b>+0.6</b> )	754 47.6 ( <b>+12.0</b> )	755 <b>5.6</b> ( <b>-72.7</b> )	756 51.2 ( <b>+1.7</b> )
	Qwen	757 39.7 ( <b>+4.9</b> )	758 99.3 ( <b>+0.5</b> )	759 75.4 ( <b>+1.6</b> )	760 23.4 ( <b>+4.0</b> )	761 82.8 ( <b>-9.8</b> )	762 51.2 ( <b>+3.0</b> )

721  
722 B TRANSLATED DATASET DETAILS  
723724  
725 To ensure high translation quality, we re-translated the English source with GPT-4o, a model that  
726 exhibits near-professional performance across many language pairs (Yan et al., 2024; Blain et al.,  
727 2023). After each pass we filtered candidates with xCOMET (Guerreiro et al., 2024): only hypothe-  
728 ses that exceeded the Flores dev-set mean for the target language were retained; sentences falling  
729 below the threshold were re-translated.730  
731 Moreover, to validate training data quality, we conducted an additional sanity check. We normalized  
732 all numerals across languages to Arabic digits and verified via regex that every number in the English  
733 source appeared unchanged in the translation. This process ensures that all translations properly  
734 preserve the original Arabic digits and equations.735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

756 C ROLLOUT GENERATION EXAMPLE  
757758  
759  
760  
761  
762  
763

## 764 Question (Korean)

765 비고의 나이는 그의 남동생이 2살일 때 남동생 나이의 두 배보다 10살 더 많았습니다. 만약 그의 남동생이 현재 10살이라면,  
766 (Vigo was 10 years older than twice his younger brother's age when the brother was 2 years old. If the younger brother is now 10)  
767 그들 의 나이 합은 얼마입니까? 차근차근 단계별로 논리적으로 풀어주세요, 최종 답은  $\boxed{\cdot}$  안에 넣어주세요.  
768 (what is the sum of their current ages? Please solve the problem step by step, showing your reasoning logically, and place the final answer  
769 inside  $\boxed{\cdot}$ .)

## 770 Rollout Examples during Training

771 1. 남동생의 현재 나이를 확인합니다. 문제에서 남동생이 현재 10살이라고 주어졌습니다.  
772 (1. Confirm the younger brother's current age. According to the problem, he is now 10 years old.)  
773 [...]  
774 남동생이 지금 10살이므로, 비고는  $10 + 12 = 22$ 살입니다. 4. 두 사람의 현재 나이 합을 구합니다.  $10 + 22 = 32$   
775 따라서, 두 사람의 나이 합은  $\boxed{32}$   
(Since the younger brother is 10, Vigo is  $10 + 12 = 22$  years old. 4. Calculate the sum of their current ages:  
776  $10 + 22 = 32$ . Therefore, the sum of their ages is  $\boxed{32}$ )→ Target-language  
Reward777 Let's break this down step by step:  
778 The problem states that the younger brother is currently 10 years old.  
779 [...]  
780 The younger brother is 10 years old.  
781 The older brother is 22 years old.  
782 The sum of their ages is  $\boxed{32}$ .  
783 Therefore, the sum of their ages is  $\boxed{32}$ → Non-target-language  
Reward784 Figure 3: Rollout examples from GRPO training of Qwen-2.5 1.5B on the Korean-translated  
785 GSM8K. Observe that the model often arrives at the right answer via English reasoning (*non-*  
786 *target language*); because any correct answer earns full reward, repeated reinforcement of such  
787 off-language traces gradually shifts the chain-of-thought word ratio away from Korean.793 D QUALITATIVE ANALYSIS OF LANGUAGE FIDELITY IN REASONING  
794 TRACES795  
796  
797  
798799 We analyze the inference results of fine-tuned models to study patterns in reasoning traces regarding  
800 language fidelity aspect. We categorize these distinct patterns: (1) **Cross-lingual Collapse**, where  
801 the model generates both the reasoning trace and the final response in the dominant language (i.e.,  
802 English) regardless of the input language; (2) **English Scaffolding and Terminology**, where the  
803 model generates the response in the target language but keeps English for structural markers (e.g.,  
804 “Step 1”) or technical terminologies; and (3) **Latin Alphabet Tokens**, where Latin-script represents  
805 algorithm names, function identifiers, or unit abbreviations.806 Here, patterns (2) and (3) are the primary reasons why the English Word Ratio (EN WR) remains  
807 non-zero. Unlike cross-lingual collapse, these patterns result in a mixture of the target language and  
808 English words. In contrast, models that maintain high multilingual fidelity (e.g., Qwen on Ukrainian  
809 tasks) successfully localize both scaffolding and technical terms, thereby achieving a near-zero EN  
810 WR.

810

## Inference Example 1: Cross-lingual Collapse

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

**Question :** 花瓶にバラが4本あります. ダリアはバラより7本多いです. 花瓶の中に花は全部で何本ありますか?

**Response :**

To determine the total number of flowers in the vase, we need to find the number of each type of flower and then add them together. Let's break it down step by step.

1. **Number of roses in the vase:** - According to the problem, there are 4 roses in the vase.

2. **Number of daisies in the vase:** - The problem states that the number of daisies is 7 more than the number of roses. Since there are 4 roses, we need to add 7 to find the number of daisies.

$$\text{Number of daisies} = 4 + 7 = 11$$

3. **Total number of flowers in the vase:**

- Now, we need to find the total number of flowers by adding the number of roses and the number of daisies.

$$\text{Total number of flowers} = \text{Number of roses} + \text{Number of daisies} = 4 + 11 = 15$$

Therefore, the total number of flowers in the vase is 15.

## Inference Example 2: Scaffolding or terminology inside CoT

**Question :** 카를로스는 레몬 나무를 심고 있습니다. 나무를 심는 데에는 90달러가 듭니다. 이 나무에서는 매년 7개의 레몬이 열리고, 레몬은 개당 1.5달러에 팔 수 있습니다. 나무에 물과 비료를 주는 데에는 매년 3달러가 듭니다. 카를로스가 이 나무에서 순수익을 내기 시작하려면 몇 년이 걸릴까요?

**Response :**

1. **Initial Investment:** 카를로스는 레몬 나무를 심는데 90달러가 들었습니다.

2. **Annual Revenue:** 매년 7개의 레몬이 열리고, 각각 1.5달러에 팔 수 있습니다. 따라서 매년 수익은  $7 \times 1.5 = 10.5$ 달러입니다.

3. **Annual Expenses:** 매년 3달러의 비용이 듭니다.

4. **Net Annual Profit:** 매년 수익에서 비용을 뺀 결과는  $10.5 - 3 = 7.5$ 달러입니다.

5. **Time to Break Even:** 초기 비용을 빼고, 매년 수익을 빼면, 카를로스가 수익을 내기 시작하려면  $90/7.5 = 12$ 년이 걸립니다.

최종 답: 12 년

864  
865

## Inference Example 3: Using english Abbreviation term

866  
867**Question :** 如果任意一点  $A$   $B$   $C$  或  $D$  出, 其他三点各一次, 最短的行程距是多少?868  
869**Response :**870  
871  
872到了到任意一点  $A$   $B$   $C$  或  $D$  出, 其他三点各一次的最短行程距, 我可以考化一, 其中每点代表一点, 每代表一距 我需要到任意一点出, 其他三点, 最后回到原点的最短路。873  
874  
875  
876  
877  
878  
879  
880

根据目出的, 我可以看到:

- $A$  到  $B$  的距是 3
- $A$  到  $C$  的距是 6
- $A$  到  $D$  的距是 5
- $B$  到  $C$  的距是 6
- $B$  到  $D$  的距是 4
- $C$  到  $D$  的距是 5

881  
882  
883

到了到最短路, 我可以使用中的最短路算法, 如Dijkstra算法 但是, 由于是一的, 我可以直接算。

884  
885  
886  
887  
888  
889首先, 我  $A$  出,  $B$   $C$  和  $D$  各一次, 然后回到  $A$  最短路可以通以下步算:

1.  $A$  到  $B$  的距是 3。
2.  $B$  到  $C$  的距是 6。
3.  $C$  到  $D$  的距是 5。
4.  $D$  到  $A$  的距是 5。

890  
891因此, 最短路是  $A \rightarrow B \rightarrow C \rightarrow D \rightarrow A$ , 距是  $3 + 6 + 5 + 5 = 19$ 。892  
893  
894所以, 最短的行程距是 19。895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

918  
 919 Table 7: GRPO training results with 7B or 8B backbones on translated GSM8K dataset. We report  
 920 accuracy, target-language word ratio (Target WR), and English word ratio (EN WR). Numbers in  
 921 parentheses denote the change relative to the corresponding backbone.

Model	Language	Accuracy (%)	Target WR (%)	EN WR (%)
OLMo-2-7B-SFT	ZH	83.3 (+33.6)	0.3 (-80.9)	87.7 (+83.6)
OLMo-2-7B-SFT	KO	71.1 (+58.3)	0.2 (-82.1)	89.1 (+81.8)
Llama-3.1-8B-Instruct	KO	76.2 (+7.9)	92.4 (-4.7)	6.8 (+4.3)
Llama-3.1-8B-Instruct	ZH	91.3 (+8.3)	90.8 (+6.8)	7.8 (-5.4)

## E LARGER MODEL EXPERIMENTS

931 To verify whether our findings extend to larger models, we conduct additional experiments on two  
 932 larger backbones: OLMo-2-7B-SFT and Llama-3.1-8B-Instruct. We follow the same experimental  
 933 settings and evaluation metrics as described in Section 3.1, training on translated GSM8K in Chinese  
 934 (ZH) and Korean (KO).

935 Table 7 demonstrate the results. For OLMo-2-7B, GRPO dramatically improves GSM8K accuracy  
 936 in both Chinese and Korean, but Target WR drifts to almost zero while the English word ratio rises to  
 937 nearly 90%. This reproduces the collapse behavior we observed for the smaller OLMo backbone in  
 938 Table 1. On the other hands, Llama-3.1-8B-Instruct have improvement while preserving high target-  
 939 language fidelity: the target-language word ratio remains above 90% in both Chinese and Korean,  
 940 and English WR retains below 10%. These results indicate that cross-lingual collapse persists at  
 941 larger scales and strongly depends on the backbone.

942  
 943  
 944  
 945  
 946  
 947  
 948  
 949  
 950  
 951  
 952  
 953  
 954  
 955  
 956  
 957  
 958  
 959  
 960  
 961  
 962  
 963  
 964  
 965  
 966  
 967  
 968  
 969  
 970  
 971

972  
 973 Table 8: Cross-lingual consistency between English and Chinese on OLMo-2-7B with Chinese  
 974 dataset. Each column corresponds to one of three correctness patterns on the same math prob-  
 975 lem:  $(\checkmark, \checkmark)$  for both English and Chinese correct,  $(\checkmark, \times)$  for English-only success, and  $(\times, \checkmark)$  for  
 976 Chinese-only success. Larger proportion of the agreement case  $(\checkmark, \checkmark)$  and smaller proportion of  
 977 the disagreement cases  $(\times, \checkmark)$  and  $(\checkmark, \times)$  indicates stronger cross-lingual consistency.

	Cross-lingual Consistency		
	$(\checkmark, \checkmark) \uparrow$	$(\checkmark, \times) \downarrow$	$(\times, \checkmark) \downarrow$
OLMo-2(7B) + GRPO (ZH)	<b>79.7</b>	<b>9.9</b>	3.6
OLMo-2(7B)	46.4	35.8	<b>3.1</b>

## F CROSS-LINGUAL CONSISTENCY ANALYSIS IN COLLAPSED MODELS

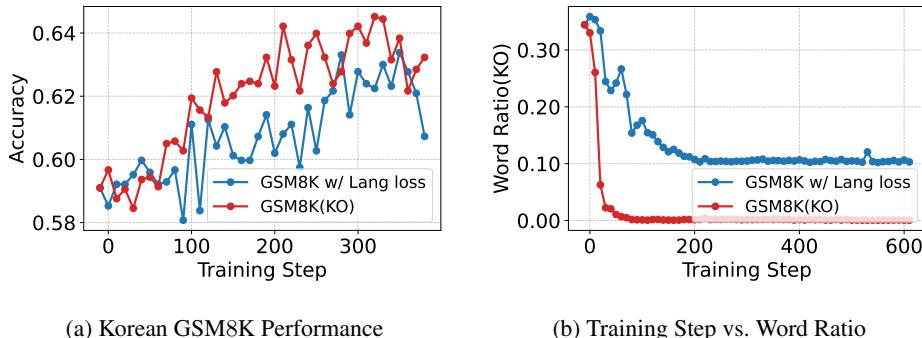
To quantify how GRPO affects alignment between English and Chinese under the Cross-lingual collapse case, we measure cross-lingual consistency on GSM8K for OLMo-2-7B-SFT before and after GRPO training on GSM8K(ZH). For each test item, we query the model twice: once with the original English question and once with the Chinese translation, using the same evaluation. We then categorize each item according to whether the model is correct ( $\checkmark$ ) or incorrect ( $\times$ ) under the English and Chinese prompts, computing the proportion of three cases:  $(\checkmark, \checkmark)$ ,  $(\checkmark, \times)$  and  $(\times, \checkmark)$ .

Table 8 reports the resulting ratios. For the Instruct backbone, only 46.4% of items are solved in both languages, while 35.8% are correct in English but wrong in Chinese. After fine-tuning with GRPO, the collapsed model displays much stronger cross-lingual alignment: the  $(\checkmark, \checkmark)$  category increases to 79.7% and the  $(\checkmark, \times)$  category drops to 9.9%, while the  $(\times, \checkmark)$  category remains small (3-4%) for both models. Thus, GRPO largely eliminates cases where the model can solve a problem only in English but not in the target language.

Consistent with this view, among the problems that the GRPO model solves under Chinese prompts, **95.7%** are also solved correctly when the same model is run on the English input. Combined with the language-fidelity results in Table 1, this supports the interpretation that accuracy gains in the collapsed model come from routing the target language prompts through the model’s stronger English reasoning, rather than from improved target-language reasoning capabilities

## G FURTHER TRAINING OF DISTILLED LRM

As depicted in Figure 4, 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.



1007  
 1008 Figure 4: We continued GRPO fine-tuning of the DeepSeek-R1-Distill Qwen model on the Korean-  
 1009 translated GSM8K dataset to encourage Korean chain-of-thought reasoning. As Figure 4b shows,  
 1010 the distilled model still exhibits cross-lingual collapse during training.  
 1011  
 1012  
 1013  
 1014  
 1015  
 1016  
 1017  
 1018  
 1019  
 1020  
 1021  
 1022  
 1023  
 1024  
 1025

1026 Table 9: **Ablation of language-consistency interventions for Llama-3.2-3B Instruct on Ukrainian**  
 1027 **GSM8K. All three approaches substantially improve target-language word ratio (Target WR) relative**  
 1028 **to vanilla GRPO, at the cost of lower accuracy.**

Model / Setting	GSM8K (Acc %)	Target WR (%)	EN WR (%)
GRPO (UK)	70.9	0.3	96.8
GRPO + Language proportion reward	62.5	81.6	16.4
GRPO + Language threshold reward	62.1	85.0	11.9
GRPO + Rollout-level filtering	63.8	88.0	13.0

## H ABLATIONS OF LANGUAGE-CONSISTENCY INTERVENTIONS

To analyze how different language-consistency interventions affect cross-lingual collapse at the rollout level, we start from the Ukrainian GSM8K setting in Sec. 3.1. For each sampled rollout  $y$  we compute its target-language word ratio  $\text{TargetWR}(y)$  using the script-based preprocessing pipeline described in Sec. 3.1.

**Language-consistency reward.** We study two ways of injecting language fidelity into the scalar reward:

1. **Proportion reward.** We add the target-language ratio directly to the correctness reward:

$$r_{\text{total}}(y) = r_{\text{corr}}(y) + \lambda \cdot \text{TargetWR}(y),$$

where  $\lambda$  represents the weight of the language reward. We set  $\lambda$  to 0.5 in all experiments.

2. **Threshold reward.** We add a fixed bonus of 0.5 if the rollout is mostly in the target language and sufficiently long:

$$r_{\text{total}}(y) = r_{\text{corr}}(y) + \lambda \cdot \mathbf{1}[\text{TargetWR}(y) \geq 0.5 \wedge \text{len}(y) > 10],$$

where  $\text{len}(y)$  counts non-LaTeX tokens. The length constraint avoids degenerate short but the target language responses. We also set  $\lambda$  to 0.5 in the experiment.

We denote the proportion reward and threshold reward as Language proportion reward and Lang threshold reward in Table 9, respectively. Note that the model with the language reward in Fig. 2 utilizes the proportion reward as an auxiliary reward.

**Rollout-level filtering.** We also evaluate an alternative that acts on rollout sampling level rather than on the reward. For each trajectory  $y$  we compute the English word ratio  $\text{ENWR}(y)$ . If  $\text{ENWR}(y) \geq 0.5$ , we discard  $y$  and resample until the number of valid samples is filled. This procedure filters out trajectories whose reasoning has already collapsed into predominantly English. We denote this approach as rollout-level filtering in Table 9

**Results.** Table 9 reports results for Llama-3.2-3B Instruct trained with GRPO on Ukrainian GSM8K. All three interventions substantially increase the target-language word ratio (from 0.3% under vanilla GRPO to 82–88%) and reduce the English word ratio to around 12–16%. However, they also lower GSM8K accuracy from 70.9% to 62–64%, i.e., by roughly 7–9 percentage points relative to the correctness-only baseline. Among the three, English-response detection offers the best accuracy–fidelity trade-off, but it does not remove the underlying performance–fidelity tension highlighted in Sec. 3.4.

1080  
 1081 Table 10: Qwen-2.5-1.5B Instruct on Korean GSM8K and MATH500 under GRPO and super-  
 1082  
 1083  
 1084  
 1085  
 1086  
 1087  
 1088  
 1089  
 1090  
 1091  
 1092  
 1093  
 1094  
 1095  
 1096  
 1097  
 1098  
 1099  
 1100  
 1101  
 1102  
 1103  
 1104  
 1105  
 1106  
 1107  
 1108  
 1109  
 1110  
 1111  
 1112  
 1113  
 1114  
 1115  
 1116  
 1117  
 1118  
 1119  
 1120  
 1121  
 1122  
 1123  
 1124  
 1125  
 1126  
 1127  
 1128  
 1129  
 1130  
 1131  
 1132  
 1133

Dataset	GSM8K (KO)			MATH500 (KO)		
	Accuracy (%)	Target WR (%)	EN WR (%)	Accuracy (%)	Target WR (%)	EN WR (%)
GRPO (Base)	43.1	94.0	3.6	27.1	86.5	10.9
GRPO (Base + Hard)	47.5	<b>14.5</b>	80.1	46.7	<b>2.1</b>	87.4
SFT(Base)	40.9	97.0	2.8	24.0	94.2	4.1
SFT(Base + Hard)	42.5	96.8	3.1	30.9	94.5	5.4

## I ADDITIONAL SUPERVISED FINE-TUNING BASELINES

To complement our RLVR results, we conduct an additional supervised fine-tuning (SFT) study on Korean. Our goal is to test whether enriched target-language math supervision can narrow the accuracy gap to GRPO-based models and mitigate Cross-lingual Collapse.

**Experimental setup.** For a fair comparison with the GRPO models, we reuse the same math datasets as in the main experiments: GSM8K and SimpleRL-Zoo. We build a Korean SFT corpus by translating the original English CoTs of the dataset into Korean, following exactly the same translation pipeline in Appendix B.

We consider two SFT configurations for Qwen-2.5-1.5B Instruct: (i) SFT on Korean GSM8K only(i.e.,SFT(Base)), and (ii) SFT on the union of Korean GSM8K and the hard curriculum (SimpleRL-Zoo) with their translated responses (SFT(Base+ Hard)). In both configurations, we train for two epochs to mitigate overfitting, using the same optimizer, learning rate, and other hyperparameters as in the GRPO runs.

**Results.** Table 10 demonstrates that the accuracy of SFT variants increase while keeping Target WR high, but they still underperform the GRPO-based models on GSM8K dataset. Moreover, GRPO (Base + Hard) model attains the highest accuracy at the cost of reduced Target WR and increased EN WR. This result still shows the accuracy–language fidelity trade-off.