Multilingual Chain-of-Thought Compression via Cross-Lingual Distillation

Jiarui Wan[1,2], Songming Zhang[1,2], Yufeng Chen[1,2,*]
[1] Key Laboratory of Big Data & Artificial Intelligence in Transportation
(Beijing Jiaotong University), Ministry of Education
[2] School of Computer Science and Technology, Beijing Jiaotong University
{24120380}@bjtu.edu.cn

Abstract

Chain-of-thought reasoning improves the performance of large language models on complex tasks but often produces overly verbose outputs, leading to increased inference cost. This issue is exacerbated in multilingual settings, where differences in tokenization and linguistic structure result in inconsistent compression performance across languages. Existing methods are largely English-centric and tend to suffer from accuracy degradation, especially in low-resource languages. We propose **M**ultilingual **C**hain-of-thought Compression via Cross-lingual **D**istillation **(MCD)**, a unified framework that addresses these challenges through both data construction and optimization. MCD builds a cross-lingually aligned dataset using a translation-with-verification pipeline and difficulty-aware sampling, and employs a reinforcement training strategy that combines supervised fine-tuning with direct preference optimization to encourage concise yet sufficient reasoning. Experiments on multilingual mathematical benchmarks show that MCD consistently reduces reasoning length while maintaining competitive accuracy, and significantly improves robustness in low-resource languages.

1 Introduction

Chain-of-Thought (CoT) reasoning has become a key technique for improving the ability of large language models (LLMs) to solve complex multi-step problems, particularly in mathematical reasoning. By explicitly generating intermediate steps, CoT significantly improves accuracy across a wide range of benchmarks (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025; Team et al., 2025). However, these gains come at the cost of increased inference overhead, as CoT often produces overly verbose reasoning traces with redundant steps, a phenomenon known as overthinking (Wu et al., 2025; Kumar et al., 2025).

Language	Metric	Original	Compressed CoT
EN	Acc.	95.76	93.92↓ (−1.84)
	Tok.	2126	1660↑ (+0.22)
BN	Acc.	82.72	79.52↓ (−3.20)
	Tok.	2645	2486↑ (+0.06)
SW	Acc.	32.48	31.44↓ (−1.04)
	Tok.	4371	4194↑ (+0.04)

Table 1: Performance of the SFT compression method(Huang et al., 2025a) based on Qwen3-4B across languages, evaluated on the MGSM (Shi et al., 2022) dataset. "Acc." denotes the accuracy and "Tok." denotes the average generated response token numbers.

This inefficiency is further exacerbated in multilingual settings. As models are predominantly pretrained on English-dominant corpora, they exhibit higher generative uncertainty when reasoning in non-English languages, often requiring longer reasoning traces to arrive at the correct answer, which increases both inference latency and cost. More importantly, a compression strategy that works well in English may not transfer effectively to other languages, either causing significant accuracy degradation or yielding minimal token reduction, due to differences in tokenization granularity, morphological complexity, and the availability of training signals.

Despite its importance, multilingual CoT compression remains underexplored. Existing methods exhibit two closely related limitations. **First, they are predominantly English-centric** (Xia et al., 2025; Kang et al., 2025; Yuan et al., 2025; Li et al., 2025; Huang et al., 2025a) and fail to account for cross-lingual variation in reasoning behavior, leading to disproportionately large performance degradation in low-resource languages. This issue is reflected in the larger compression rate collapses observed in lower-resource languages such as Ben-

*Yufeng Chen is the corresponding author.

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 83–91

gali and Swahili in Table 1. This disparity in training signal quality directly exacerbates **the second limitation: SFT alone is insufficient** to preserve correctness under aggressive compression, as uneven multilingual supervision makes it particularly difficult for the model to distinguish necessary reasoning steps from redundant verbosity. This difficulty is especially severe in low-resource languages where training signals are sparse and inconsistent.

To address these challenges, we propose **M**ultilingual **C**hain-of-thought Compression via Cross-lingual **D**istillation (**MCD**), a unified framework that tackles the above limitations from both data and optimization perspectives. The framework consists of two steps:

Cross-lingual aligned data construction. We construct a high-quality multilingual CoT compression dataset covering five typologically diverse languages (German, French, Japanese, Russian, and Chinese) based on DeepMath-103K (He et al., 2025). Using Qwen3.5-27B (Qwen Team, 2026), we adopt a translation-with-verification strategy to ensure both linguistic fidelity and answer correctness. We further apply difficulty-aware sampling to preserve the original difficulty distribution across languages.

Cross-lingual distillation via reinforcement learning. We introduce a Cross-lingual distillation framework that combines SFT as a cold start with Direct Preference Optimization (DPO) (Rafailov et al., 2023). SFT establishes a compressed reasoning style, while DPO explicitly encourages concise yet sufficient reasoning by preferring compressed traces over verbose ones. This design enables the model to distinguish essential reasoning from redundancy, improving both efficiency and robustness.

Experiments on MGSM (Shi et al., 2022) and MMATH (Luo et al., 2025b) show that MCD consistently reduces reasoning length while maintaining competitive accuracy. Notably, it significantly improves robustness in low-resource languages and mitigates the cross-lingual performance gap observed in prior methods.

Our contributions are summarized as follows:

- We construct a cross-lingually aligned CoT compression dataset with verified long and compressed reasoning traces.
- We propose MCD, a cross-lingual distillation framework that enables concise yet logically sufficient reasoning.
- We demonstrate strong cross-lingual robustness and substantial token reduction across multiple models and benchmarks.

2 Related Work

Chain-of-thought (CoT) reasoning (Kojima et al., 2022) and large reasoning models (LRMs) (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025) have substantially improved model performance by enabling explicit multi-step thinking during inference. However, this comes at the cost of verbose reasoning traces with redundant steps, a phenomenon known as overthinking (Wu et al., 2025; Kumar et al., 2025), which inflates inference latency and computational cost.

CoT compression addresses this by reducing reasoning length while preserving correctness. Existing approaches span several strategies: reducing the granularity of intermediate steps via token-level or step-level pruning (Xia et al., 2025; Yuan et al., 2025), constraining generation length through short CoT supervision (Kang et al., 2025), and leveraging self-generated outputs with adaptive filtering to distill concise reasoning traces (Huang et al., 2025a). Some methods further modulate reasoning depth according to problem difficulty (Luo et al., 2025a). Despite this progress, these methods are predominantly designed and evaluated in English, leaving the multilingual setting largely unexplored.

In multilingual LLMs, disparities in syntax, morphology, and tokenization lead to significant variation in reasoning structure and information density across languages. Benchmarks such as MGSM (Shi et al., 2022) and MMATH (Luo et al., 2025b) have revealed substantial cross-lingual performance gaps, particularly for low-resource languages. While recent work shows that reinforcement learning can improve cross-lingual reasoning transfer (Huang et al., 2025b), compression strategies that account for such variation remain underexplored. A method effective in English may yield minimal token reduction or severe accuracy degradation in morphologically complex or low-resource languages. Our work addresses this gap by constructing a cross-lingually aligned dataset with difficulty-aware sampling and adopting a reinforcement training framework using DPO (Rafailov et al., 2023) algorithm that encourages concise yet logically sufficient reasoning across typologically

diverse languages.

3 Method

We present a cross-lingual distillation approach to refining language model reasoning efficiency in multilingual mathematical settings. Our method leverages a difficulty-aware data curation and compression pipeline to promote concise and stable inference across diverse linguistic groups. We first detail the construction of a high-quality multilingual corpus based on the Deepmath-103K dataset, then design a reinforcement learning paradigm consisting of pattern imitation and preference alignment. Finally, we define a hybrid training objective that utilizes Supervised Fine-Tuning (SFT) to instill short-form reasoning patterns and Direct Preference Optimization (DPO) to explicitly penalize redundant tokens while preserving logical rigor.

3.1 Multilingual Data Construction

To facilitate the study of multilingual chain-of-thought (CoT) compression, we construct a high-quality, difficulty-aware dataset spanning five languages: German (de), French (fr), Japanese (ja), Russian (ru), and Chinese (zh). For more language selection details, please refer to Appendix C. The construction pipeline consists of three stages: (1) multilingual expansion, (2) difficulty-aware multilingual sampling, and (3) CoT compression.

Multilingual Expansion. We start from the DeepMath-103K (He et al., 2025) dataset, where each instance is represented as a triplet (q, d, gt), denoting the problem, its difficulty level, and the ground-truth answer, respectively. To obtain multilingual data, we employ Qwen3.5-27B (Qwen Team, 2026) as the translation backbone. Instead of performing direct translation, we adopt a translation-with-verification strategy to ensure both linguistic fidelity and semantic correctness. Specifically, for each target language $\ell \in$ de, fr, ja, ru, zh, the model is prompted (Appendix B) to jointly translate the query and generate a reasoning trace $c^{(\ell)}$ leading to an answer $\hat{gt}^{(\ell)}$. We retain a translated instance only if:

$$\hat{gt}^{(\ell)} = gt. \tag{1}$$

This rejection sampling process filters out erroneous translations that may distort the underlying mathematical semantics, resulting in a high-quality multilingual dataset:

$$D_{\text{multi}} = (q^{(\ell)}, d, c^{(\ell)}, gt). \tag{2}$$

Difficulty-Aware Sampling. DeepMath-103K provides difficulty annotations aligned with AoPS standards. Preserving this distribution is crucial for preventing bias toward either overly simple or excessively difficult problems. Let $P(d)$ denote the empirical difficulty distribution of the original dataset. For each language ℓ, we construct a subset $\mathcal{D}^{(\ell)}$ via stratified sampling such that:

$$P_{D^{(\ell)}}(d) \approx P(d). \tag{3}$$

This ensures that each language-specific subset maintains the same difficulty profile as the original dataset. The final multilingual dataset is then obtained by aggregating all language subsets:

$$\mathcal{D}_{\text{final}} = \bigcup_{\ell} \mathcal{D}^{(\ell)}. \tag{4}$$

CoT Compression. To train models for concise reasoning, we further compress the reasoning traces using Qwen3.5-27B with the prompt template described in Appendix A. Given an original reasoning trace c, the model generates a compressed version c_T that preserves essential logical steps while removing redundancy $c \to c_T$. We then condition on c_T to regenerate the final answer $c_T \to c_A$, ensuring that the compressed reasoning remains sufficient for correct problem solving. The final dataset $\mathcal{D}_{\text{final}}$ is organized as a set of tuples:

$$(q, d, c_T, T, c_A, gt), \tag{5}$$

where T denotes the original reasoning trace.

3.2 Hybrid Training Strategy

To bridge the gap between verbose chain-of-thought reasoning and concise multilingual inference, we propose a two stage optimization framework. This strategy first establishes a foundational reasoning style through Supervised Fine-Tuning (SFT) and subsequently refines the model's efficiency via Direct Preference Optimization (DPO).

Pattern Imitation via SFT. The primary objective of the first stage is to instill the target model with the structural patterns of compressed reasoning across multiple languages. Given the multilingual dataset $\mathcal{D}_{\text{final}}$, we fine-tune the student model to maximize the conditional likelihood of the compressed reasoning traces. For each instance $(q, c_T, c_A) \in \mathcal{D}_{\text{final}}$, the SFT loss is defined as:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(q,c_T,c_A)\sim\mathcal{D}_{\text{final}}}\Big[\sum_{t=1}^{|c_T|} \log P_\theta(c_{T,t} \mid q, c_{T,<t})\Big] \tag{6}$$

where θ denotes the model parameters. By training exclusively on c_T rather than the original trace T, the model learns to internalize a more token-efficient reasoning trajectory while maintaining the cross-lingual mapping between the query q and the final answer c_A.

Preference Alignment via DPO. While SFT effectively transfers the format of compressed CoT, it may struggle to distinguish between necessary logical steps and redundant verbosity when faced with complex multilingual queries. To address this, we employ Direct Preference Optimization (Rafailov et al., 2023) to explicitly penalize linguistic redundancy. We construct preference pairs from our dataset where the compressed trace c_T serves as the preferred completion (y_w) and the original, verbose trace T serves as the rejected completion (y_l). This setup forces the model to recognize that while both c_T and T may lead to the correct ground-truth gt, the more concise path is superior. The DPO objective is formulated as:

$$\mathcal{L}_{\text{DPO}}(\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(q,c_T,T)\sim\mathcal{D}_{\text{final}}}\Bigg[\log \sigma\Big(\beta \log \frac{\pi_\theta(c_T \mid q)}{\pi_{\text{ref}}(c_T \mid q)} - \beta \log \frac{\pi_\theta(T \mid q)}{\pi_{\text{ref}}(T \mid q)}\Big)\Bigg] \tag{7}$$

where π_{ref} is the reference model (typically the checkpoint resulting from Stage I), and β is a hyperparameter controlling the deviation from the reference policy. By optimizing this objective, the model learns a relative preference for brevity. Crucially, because $\mathcal{D}_{\text{final}}$ is constructed via difficulty-aware sampling, the preference pairs presented to DPO span the full spectrum of problem complexity. For harder problems, the gap between the compressed trace c_T and the original trace T is naturally smaller, as aggressive compression of complex reasoning tends to produce incorrect answers and is therefore filtered out during data construction. This implicitly discourages the model from over-compressing difficult problems, without requiring any explicit difficulty conditioning at training or inference time.

4 Experiments

4.1 Experimental Setup

Training Data. We use the dataset $\mathcal{D}_{\text{final}}$ constructed in Section 3.1 for training. For a fair comparison, all methods use this mixed dataset.

Baselines. SEER (Huang et al., 2025a) is a self-enhancing supervised fine-tuning (SFT) compression framework that integrates Best-of-N sampling with adaptive, dataset-specific length filtering to distill concise yet accurate reasoning traces from self-generated outputs. To further disentangle the contribution of each training phase in our pipeline, we also compare against two ablated variants: **SFT-Only**, which fine-tunes the model solely on $\mathcal{D}_{\text{final}}$ using supervised learning without any preference optimization; and **DPO-Only**, which applies Direct Preference Optimization directly to the base model without the preceding SFT stage. These variants allow us to isolate the effect of each component and validate the necessity of combining both stages.

Backbone Models. We conduct experiments using widely adopted Large Reasoning Models (LRMs), including Qwen3-1.7B, Qwen3-4B and Qwen3-8B (Yang et al., 2025). These models have been extensively used in prior studies and span different parameter scales, enabling a fair and comprehensive comparison. Training details are provided in Appendix E.

Evaluation Benchmarks. We evaluate on two benchmarks: **MGSM** (Shi et al., 2022) (250 math problems across 10 languages) and **MMATH** (Luo et al., 2025b) (374 problems across 10 languages). We report accuracy (Acc) and token count (Tok), with maximum generation lengths of 16,384 and 32,768 tokens respectively. Following Yang et al. (2025), we use temperature 0.6 and top-p 0.95, averaging results over five samples.

4.2 Main Results

The main experimental results presented in Table 2 and Table 3 provide a robust evaluation of our method, specifically addressing the core limitations of multilingual overthinking and imbalanced proficiency identified in Section 1. While prior chain of thought compression techniques are predominantly English centric and often lead to significant performance degradation in non English languages, our proposed framework maintains competitive accuracy across diverse linguistic settings while substantially improving inference efficiency.

Across the three model scales of Qwen3-1.7B, 4B, and 8B, the experimental data reveals a consistent pattern of effective token reduction. For the smaller Qwen3-1.7B model, our approach achieves substantial compression with only marginal impacts on accuracy, suggesting that the preference alignment stage successfully regularizes the

Method	bn		de		en		es		fr		ja		ru		sw		te		th		zh		ΔAcc↑	Comp.↑
	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓	Acc↑	Tok↓		
Qwen3-1.7B																								
Original	63.36	3384	83.60	1490	85.84	2338	86.08	1574	80.56	1433	71.84	1952	80.80	1515	7.76	5511	40.72	4839	76.24	2483	79.84	2360	–	–
SFT	35.60	4228	65.12	1340	75.44	**1381**	67.28	1203	62.32	1169	53.12	1453	63.60	1614	2.08	11175	22.64	7429	53.84	1790	66.16	**1267**	-17.22	0.00
DPO	64.48	3345	**82.48**	1400	85.76	2205	**85.68**	1509	**80.24**	1361	**72.80**	1920	**81.84**	1542	**7.52**	5752	40.40	5017	76.32	2413	79.60	2292	**0.04**	0.02
SEER	62.72	2835	81.84	1333	**89.44**	1558	**85.68**	1321	80.08	1263	70.40	1648	80.08	1573	6.80	6101	40.00	4679	**76.96**	1944	79.44	2465	-0.29	0.10
MCD (ours)	**64.96**	**1908**	80.00	**858**	85.28	1415	84.24	**903**	77.44	**857**	68.64	**1103**	77.84	**868**	6.40	**4495**	**42.08**	**3017**	76.00	**1409**	**80.00**	1369	-1.25	**0.40**
Qwen3-4B																								
Original	82.72	2645	90.40	1321	95.76	2126	91.68	1293	85.20	1422	83.68	1817	91.12	1548	32.48	4371	72.96	3405	87.20	2068	88.08	2176	–	–
SFT	67.52	2300	77.68	1400	88.72	**944**	81.68	1110	77.28	**1050**	74.24	**1171**	80.56	1346	23.28	3720	59.52	3007	76.48	**1135**	80.24	**895**	-10.37	**0.26**
DPO	**81.44**	2484	**89.12**	1289	93.44	1996	**90.64**	1265	**84.88**	1376	82.88	1777	**90.16**	1600	**31.92**	4308	71.36	3300	84.88	1970	**87.12**	2241	**-1.22**	0.02
SEER	79.52	2486	88.88	1249	93.92	1660	90.00	1298	83.52	1306	82.16	1744	88.32	1850	31.44	4194	68.16	3423	**85.28**	1843	82.64	2720	-2.49	0.01
MCD (ours)	79.12	**1944**	87.60	**1020**	**94.24**	1543	89.92	**982**	84.32	1059	**83.20**	1343	89.52	**1166**	30.40	**3555**	**71.84**	**2560**	85.20	1571	86.96	1673	-1.72	0.24
Qwen3-8B																								
Original	87.52	2382	90.56	1511	96.56	2096	91.76	1523	88.32	1531	85.12	1842	91.84	1616	60.88	3306	80.40	3081	89.60	2180	89.04	2583	–	–
SFT	76.40	**1587**	79.76	1309	90.48	**1078**	83.12	**1149**	78.64	**1071**	77.12	**1077**	82.64	**1194**	45.36	2746	67.76	**2235**	80.40	**1336**	82.56	**922**	-9.76	**0.33**
DPO	86.40	2310	**89.20**	1487	91.52	2084	91.52	1539	86.80	1570	85.12	1826	**91.04**	1617	58.64	3416	**80.72**	3091	**89.60**	2098	87.44	2584	-1.24	0.00
SEER	61.52	7472	89.04	1468	**94.80**	1826	**91.92**	1527	86.08	1567	**85.36**	1783	90.24	1729	60.64	3422	51.52	7824	88.64	1909	85.52	2679	-6.03	-0.32
MCD (ours)	**86.72**	1863	89.04	**1149**	92.00	1691	91.36	1235	**86.96**	1192	85.20	1437	90.48	1284	**60.88**	**2714**	79.92	2530	89.28	1644	**87.60**	2098	**-1.11**	0.21

Table 2: Results on MGSM benchmarks across three LRMs. "Comp." denotes the compression ratio $(l - l')/l$ where l and l' are the token counts of the baseline and the proposed method, respectively. "Tok" denotes the model's generated token counts. The best results in each language are in **bold**, and the second-best results are underlined.

Method	ar		en		es		fr		ja		ko		pt		th		vi		zh		ΔAcc ↑	Comp. ↑
	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓	Acc ↑	Tok ↓		
Qwen3-1.7B																						
original	79.14	5589	85.09	7138	85.76	6016	84.96	5850	79.88	6037	79.81	5804	84.02	6181	81.68	6072	82.35	5527	79.01	7256	–	–
SFT	46.12	7405	54.08	9457	51.00	7889	50.74	7598	46.39	7690	46.66	6563	51.67	8271	49.06	6780	48.80	7179	51.14	8122	-32.60	-0.25
DPO	**78.34**	5479	**85.56**	6954	84.49	5905	**83.76**	5849	**79.68**	5781	79.88	5764	**84.22**	5675	**79.95**	6139	**81.15**	5475	**78.88**	7010	**-0.58**	0.02
SEER	77.54	5148	85.43	6232	**84.76**	5469	82.62	5281	79.21	5489	**80.28**	5268	83.89	5445	**79.95**	5401	81.08	5132	76.27	8230	-1.07	0.07
MCD (ours)	75.40	**3686**	81.89	**4911**	80.95	**3885**	80.88	**3697**	75.74	**3913**	77.14	**3780**	80.88	**3931**	78.07	**4054**	78.14	**3633**	76.07	**4436**	-3.65	**0.35**
Qwen3-4B																						
original	87.17	5471	92.65	6502	91.71	5449	92.05	5395	90.17	5748	89.30	5718	92.31	5382	88.70	5944	90.78	5386	86.83	7362	–	–
SFT	62.63	6093	69.79	7186	68.25	6543	67.45	5841	62.10	6340	63.84	6506	68.52	6100	63.24	6251	65.24	5741	67.98	6433	-24.26	-0.09
DPO	**87.37**	5362	**93.18**	6229	91.38	5376	**91.98**	5320	**90.78**	5629	87.70	5689	**92.11**	5395	**89.04**	5761	**89.77**	5311	86.97	7121	**-0.14**	0.02
SEER	85.90	5036	92.25	5868	**91.84**	4999	91.11	5033	89.44	5483	88.50	5476	90.78	4874	87.57	5146	89.44	5012	84.16	7725	-1.07	0.07
MCD (ours)	85.36	**4399**	92.18	**5368**	89.71	**4383**	90.64	**4417**	88.90	**4691**	**89.10**	**4588**	91.38	**4377**	88.64	**4717**	89.51	**4350**	**87.63**	**5574**	-0.86	**0.20**
Qwen3-8B																						
original	89.17	5855	93.72	6708	92.78	5907	92.78	5766	90.11	6256	91.44	5798	93.11	5825	91.11	6215	91.78	5848	88.77	7448	–	–
SFT	65.04	6691	69.18	7499	69.05	6866	69.72	6584	65.04	6521	65.51	6256	69.72	6594	65.31	6580	65.98	6566	67.45	7285	-24.28	-0.10
DPO	**89.64**	5833	93.92	6677	**92.58**	5860	92.58	5614	**90.04**	6096	**91.84**	5860	92.51	5901	90.24	6233	**92.25**	5785	88.44	7798	**-0.07**	0.07
SEER	87.77	5499	93.72	5910	**92.58**	5403	92.45	5361	89.91	5605	91.58	5431	**93.65**	5333	**90.64**	5571	91.31	5360	85.90	8126	-0.63	0.07
MCD (ours)	88.10	**4904**	**94.05**	**5660**	92.11	**5032**	**93.18**	**4734**	89.77	**5225**	91.58	**4849**	92.51	**4951**	90.44	**5343**	91.91	**4809**	**89.24**	**6120**	-0.19	**0.16**

Table 3: Results on MMATH benchmarks across three LRMs. "Comp." denotes the compression ratio $(l - l')/l$ where l and l' are the token counts of the baseline and the proposed method, respectively. "Tok" denotes the model's generated token counts. The best results in each language are in **bold**, and the second-best results are underlined.

model's reasoning trajectories. As the parameter count increases to 4B and 8B, the models exhibit even greater stability in reasoning quality under compression, indicating that larger models possess a more resilient internal logic that can be effectively distilled into more concise forms without sacrificing correctness. Furthermore, the comparison between SFT Only and our full two stage pipeline demonstrates that preference optimization is essential for preventing the over compression and accuracy loss often associated with simple imitation learning.

The results across different language resource levels further validate the cross lingual robustness of our method. For high resource languages such as English and Chinese, the model preserves nearly all of its original reasoning capability while significantly reducing the generated token counts. In medium resource settings including German, French, Russian, and Japanese, we observe a similar trend where the compressed trajectories remain logically rigorous across typologically diverse structures. Crucially, for low resource languages such as Bengali, Swahili, and Telugu, our method successfully mitigates the severe performance drops typically seen in existing compression frameworks. This demonstrates that by explicitly accounting for cross lingual variations in information density, our approach enables more equitable reasoning efficiency regardless of the language's resource level.

4.3 Analysis of Difficulty-Aware Token Reduction

To investigate whether MCD adapts its compression behavior to problem difficulty, we analyze token counts across difficulty levels defined by the AoPS scale on the MGSM dataset using Qwen3-4B. As shown in Figure 1, MCD consistently reduces token generation compared to the original model across nearly all languages and difficulty levels, demonstrating broad compression effective-

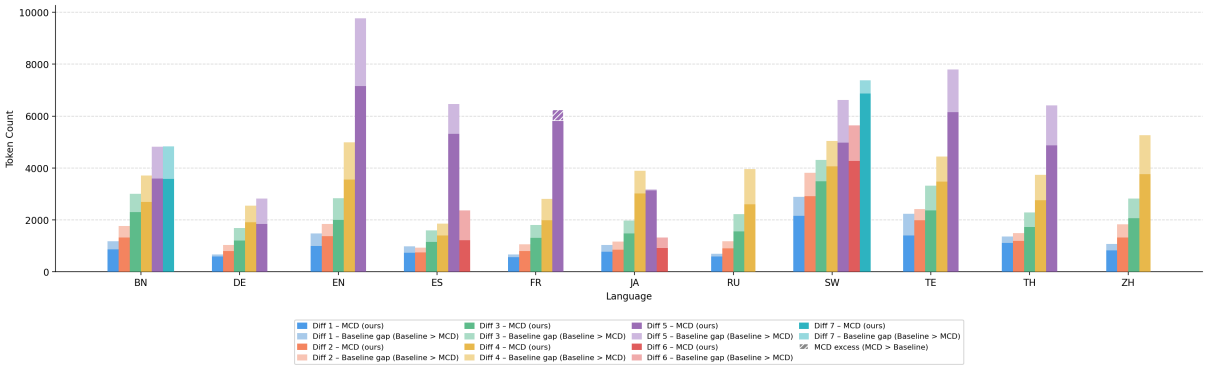

Figure 1: Token counts by language and difficulty (AoPS scale) on MGSM for MCD (ours) and the original Qwen3-4B. Each group represents a language. Solid bars show MCD tokens; lighter stacked segments indicate additional tokens used by the Baseline (Baseline - MCD), while hatched segments mark cases where MCD exceeds the Baseline (MCD - Baseline). Missing levels indicate no available instances.

ness. Crucially, the magnitude of token reduction tends to be larger for lower-difficulty problems and diminishes as difficulty increases, indicating that the model learns to allocate reasoning budget in a difficulty-sensitive manner. For the most challenging problems at higher AoPS levels, MCD occasionally matches or slightly exceeds the token count of the baseline, suggesting that the model appropriately preserves reasoning depth when the problem demands it. This behavior is consistent with our difficulty-aware sampling strategy during data construction, which ensures that the training distribution reflects the full spectrum of problem complexity. Together, these results suggest that MCD does not apply uniform compression indiscriminately, but instead modulates reasoning length according to the inherent difficulty of each problem, thereby achieving an effective balance between token efficiency and solution correctness across diverse linguistic and mathematical contexts. The results for Qwen3-1.7B and Qwen3-8B are reported in Appendix D.

4.4 Ablation Studies

Tables 2 and 3 confirm that both training stages are necessary. The SFT-only variant incurs substantial accuracy drops across all model scales, most severely in low-resource languages such as Swahili, Telugu, and Bengali, reflecting the inability of imitation learning alone to preserve correctness under aggressive compression. The DPO-only variant largely maintains baseline accuracy but achieves only marginal token reduction, indicating that preference optimization without prior exposure to compressed patterns fails to induce concise reasoning behavior. MCD resolves these complementary weaknesses by combining pattern imitation with preference alignment. Relative to SFT-only, MCD recovers most of the lost accuracy while retaining substantial compression gains, demonstrating that DPO effectively prevents over-compression. Relative to DPO-only, MCD yields significantly higher compression ratios, confirming that SFT provides a necessary inductive bias toward brevity. Smaller models such as Qwen3-1.7B benefit more from the RL design in terms of accuracy recovery, while larger models such as Qwen3-8B show stronger robustness under DPO-only but still require SFT to achieve competitive compression efficiency. Consistent trends across both MGSM and MMATH further suggest that the framework generalizes well across datasets of varying difficulty.

5 Conclusion

We propose MCD, a multilingual chain-of-thought compression framework that addresses the limitations of English-centric approaches. Through cross-lingually aligned data construction via translation with verification and difficulty-aware sampling, combined with a cross-lingual distillation training pipeline, MCD achieves concise yet sufficient reasoning across diverse languages. Experiments on MGSM and MMATH across three model scales demonstrate consistent token reduction with competitive accuracy, and improved robustness in low-resource languages such as Bengali, Swahili, and Telugu. Ablations confirm that both stages are essential: SFT learns compressed reasoning patterns, while DPO prevents over-compression by preserving correctness.

Acknowledgments

The research work descried in this paper has been supported by the National Nature Science Foundation of China (No. 62476023, 61976016, 62376019, 61976015), and the authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning.

Kerui Huang, Shuhan Liu, Xing Hu, Tongtong Xu, Lingfeng Bao, and Xin Xia. 2025a. Reasoning efficiently through adaptive chain-of-thought compression: A self-optimizing framework. *arXiv preprint arXiv:2509.14093*.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. 2025b. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms. *arXiv preprint arXiv:2509.23657*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. C3ot: Generating shorter chain-of-thought without compromising effectiveness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24312–24320.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. 2025. Overthink: Slowdown attacks on reasoning llms. *arXiv preprint arXiv:2502.02542*.

Chengzhengxu Li, Xiaoming Liu, Zhaohan Zhang, Shaochu Zhang, Shengchao Liu, Guoxin Ma, Yu Lan, and Chao Shen. 2025. Upfront chain-of-thought: A cooperative framework for chain-of-thought compression. *arXiv preprint arXiv:2510.08647*.

Haotian Luo, Haiying He, Yibo Wang, Jinluan Yang, Rui Liu, Naiqiang Tan, Xiaochun Cao, Dacheng Tao, and Li Shen. 2025a. Adar1: From long-cot to hybrid-cot via bi-level adaptive reasoning optimization. *arXiv e-prints*, pages arXiv–2504.

Wenyang Luo, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2025b. Mmath: A multilingual benchmark for mathematical reasoning. *arXiv preprint arXiv:2505.19126*.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. 2025. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Hang Yuan, Bin Yu, Haotian Li, Shijun Yang, Christina Dan Wang, Zhou Yu, Xueyin Xu, Weizhen Qi, and Kai Chen. 2025. Not all tokens are what you need in thinking. *arXiv preprint arXiv:2505.17827*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Reasoning Process Compression Prompt

The following is the prompt used for compressing CoT reasoning.

CoT Compression Prompt

Task Description
You are an expert in compressing "Thought Processes." Please compress the provided "Thought Process" according to the following requirements:

1. Refer to the input information below (the related Question, Thought Process, and Answer). You must analyze the relationship between the Question and the Answer, and compress the Thought Process. It is crucial that you do not alter the original style or meaning of the thought process. The compressed thought process must serve as a logical bridge between the Question and the Answer, ensuring coherence.
2. While compressing the Thought Process, avoid excessive compression. Strive to retain the most critical content of the thought process.
3. The first sentence of the original Thought Process must remain unchanged.
4. Use the same language as the thought process.

Input Information
Question
{query}
Thought Process
{think}
Answer
{answer}
Output Format
Compressed Thought Process:

B Dataset Translation Prompt

We translate DeepMath-103K dataset using Qwen3.5-27B. Here is the translation prompt. {target_language} is replaced with the corresponding language during translation.

Translation Prompt

You are a professional mathematics translator. Translate the following math problem from English into {target_language}.
Rules:

1. Preserve ALL mathematical expressions, formulas, and symbols exactly as-is (LaTeX, Unicode, or plain notation).
2. Use standard mathematical terminology in {target_language} as defined in official textbooks or curricula.
3. Maintain the logical structure and sentence flow — do NOT rearrange problem conditions.
4. Keep variable names unchanged (e.g., x, y, n).
5. If a term has multiple valid translations, choose the most common one used in {target_language} academic contexts.
6. Do NOT add explanations or solve the problem.
7. Output ONLY the translated text, nothing else.

Source (English): {math_problem}
Translation ({target_language}):

C Discussion of Training Language Selection

The selection of languages for multilingual group sampling aims to achieve a balance between signal reliability and linguistic diversity. We construct a set of high-resource languages with diverse linguistic characteristics, including German, French, Japanese, Russian, and Chinese. These languages span different syntactic and morphological families, offering a stable and diverse reference space for reasoning compression. In addition, Huang et al. (2025b) report that the generalization performance of reinforcement learning algorithms varies across languages, with several non-English languages surpassing English. Based on this observation, English is excluded from our language set.

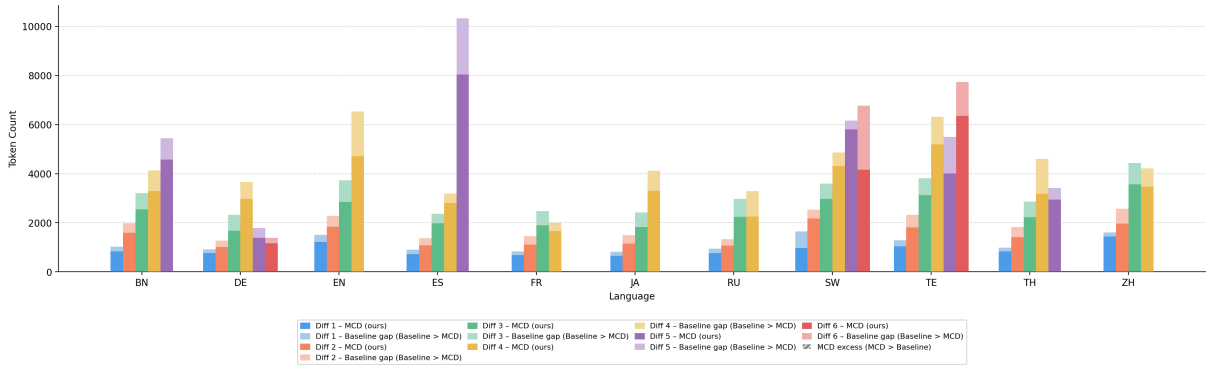

Figure 2: Token counts per language and difficulty level on the MGSM dataset for MCD (ours) and the original Qwen3-8B model.

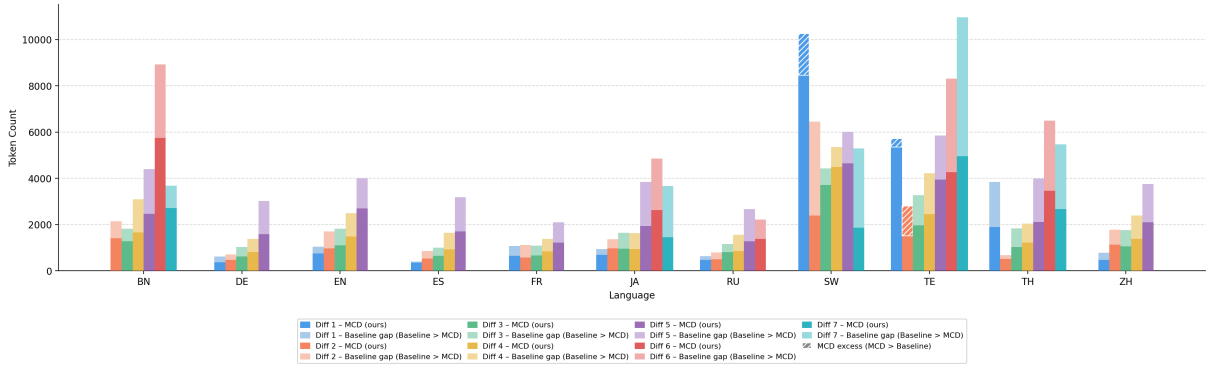

Figure 3: Token counts per language and difficulty level on the MGSM dataset for MCD (ours) and the original Qwen3-1.7B model.

D Additional

As shown in Figures 2 and 3, MCD exhibits consistent compression behavior on Qwen3-8B and Qwen3-1.7B, corroborating the findings reported for Qwen3-4B in Section 4.3. For Qwen3-8B, MCD achieves substantial token reduction across nearly all languages and difficulty levels, with compression margins narrowing as problem difficulty increases; at the highest AoPS levels, MCD occasionally matches or marginally exceeds the baseline token count, indicating that the larger model appropriately preserves reasoning depth when the problem demands it. For Qwen3-1.7B, overall compression remains pronounced, though the gap between MCD and the baseline is relatively larger for high-difficulty problems in low-resource languages such as Swahili and Telugu, reflecting the higher generative uncertainty of smaller models under aggressive compression. These results show that MCD's difficulty-aware compression generalizes across model scales, reducing more tokens on easier problems while preserving sufficient reasoning for harder ones.

E Implementation Details

We use open-source framework `LlamaFactory` (Zheng et al., 2024) as the training framework. For SFT, the learning rate is set to 3×10^{-4} for Qwen3-4B, Qwen3-1.7B and 2×10^{-4} for Qwen3-8B. For DPO, the learning rate is set to 7×10^{-6} and 5×10^{-6}, respectively.

Hyperparameter	**SFT**	**DPO**
LoRA rank	16	16
LoRA target	all	all
Cutoff length	4096	4096
Per device batch size	2	1
Gradient accumulation steps	8	8
Training epochs	3.0	3.0
Pref beta	–	0.1
Pref loss	–	sigmoid
LR scheduler type	–	cosine
Warmup ratio	–	0.05

Table 4: Hyperparameters for SFT and DPO training phases.