IS CONTINUOUS COT BETTER SUITED FOR MULTILINGUAL REASONING?

Ali Hamza Bashir[1, 2], Behzad Shomali[1, 3], Markus Frey[1, 2, 3], Mehdi Ali[1, 2], Rafet Sifa[1, 2, 3] & David Berghaus[1, 2]
Lamarr Institute[1], Fraunhofer IAIS[2], University of Bonn[3]
david.berghaus@iais.fraunhofer.de

ABSTRACT

We investigate whether performing reasoning in a continuous latent space leads to more robust multilingual capabilities. We compare Continuous Chain-of-Thought (using the CODI framework) against standard supervised fine-tuning across five typologically diverse languages: English, Chinese, German, French, and Urdu. Our experiments on GSM8k and CommonsenseQA demonstrate that continuous reasoning significantly outperforms explicit reasoning on low-resource languages, particularly in zero-shot settings where the target language was not seen during training. Additionally, this approach achieves extreme efficiency, compressing reasoning traces by approximately $29\times$ to $50\times$. These findings indicate that continuous latent representations naturally exhibit greater language invariance, offering a scalable solution for cross-lingual reasoning.

1 INTRODUCTION & RELATED WORK

While LLMs have demonstrated impressive reasoning capabilities, their performance varies dramatically across languages. Low-resource languages in particular suffer from substantially degraded results compared to high-resource languages like English (Lai & Nissim, 2024; Barua et al., 2025; Liu et al., 2026). This disparity poses a fundamental challenge: how can we enable more equitable reasoning performance across linguistically diverse settings?

Existing approaches have explored several strategies to address this challenge. One common approach translates multilingual prompts into a high-resource pivot language like English before applying chain-of-thought (CoT) reasoning (Shi et al., 2022; Ahuja et al., 2023). However, this creates a bottleneck at the translation step, where subtle linguistic nuances may be lost (Nicholas & Bhatia, 2023). More recent work has taken a different direction by directly fine-tuning LLMs on multilingual CoT data (Lai & Nissim, 2024; Barua et al., 2025), showing improvements for low-resource languages. Yet this approach faces scalability issues (covering hundreds of languages during fine-tuning is impractical) and risks catastrophic forgetting as more languages are added.

We explore a different angle: what if the reasoning process itself could operate in a more language-agnostic representation space? Recent advances in continuous latent reasoning (Hao et al., 2024; Shen et al., 2025) offer a promising direction. Rather than verbalizing each reasoning step in natural language tokens, these methods compress reasoning into continuous hidden representations. This approach is motivated by classic findings showing that word embeddings across different languages tend to occupy similar geometric spaces (Mikolov et al., 2013). If reasoning happens in continuous space rather than through explicit language tokens, the representations might naturally exhibit greater language invariance.

2 SETUP

To investigate whether continuous latent reasoning enables more language-agnostic representations, we conduct experiments across 5 typologically diverse languages: **English**, **Chinese**, **German**, **French**, and **Urdu**. This selection spans diverse branches diverse branches (Germanic, Romance,

Indo-Iranian) and families (Sino-Tibetan) (Lewis, 2009) and writing systems (Latin, Chinese characters, Arabic script), providing a rigorous test of cross-lingual generalization.

We compare the base model with two fine-tuning strategies: (i) Classical CoT-SFT (see section 2.2.3) (ii) Continuous CoT using the training recipe of (Shen et al., 2025) (see section 2.2.2). Both strategies are used in various multilingual configurations (see sections 3.1-3.2) to investigate which strategy achieves the best performance among different languages.

We use the same hyperparameters as (Shen et al., 2025) and keep them constant for all configurations to ensure comparability between both fine-tuning approaches. Complete hyperparameter details are provided in Appendix B

2.1 Data

2.1.1 Datasets

We use the following datasets to perform our experiments:

GSM8k-Aug-NL This dataset is based on the GSM8k grade-school math reasoning benchmark (Cobbe et al., 2021). We use the augmented version from (Deng et al., 2023), which expands the original 7.5k training examples to 385k training samples and 1.3k test samples and contains CoT reasoning traces in natural language.

CommonsenseQA-CoT Derived from CommonsenseQA (Talmor et al., 2019), a multiple-choice question answering dataset requiring commonsense reasoning. Since the original dataset lacks CoT annotations, we've used the annotations provided by Shen et al. (2025). This results in approximately 8.1k training and 1.2k validation examples.

2.1.2 Multilingual Data Construction

For GSM8k-Aug-NL, we create multilingual versions by translating questions into our target languages while preserving mathematical expressions and CoT structure. We employ `Llama-3.3-70B-Instruct` (Touvron et al., 2023) for German and French, `Qwen2.5-72B-Instruct` (Team, 2024) for Chinese, and `GPT-5-mini` for Urdu, all with low temperature (0.1) to ensure consistent translations. Translation guidelines include: (i) preserving all numerical values exactly, (ii) keeping mathematical expressions unchanged, (iii) adapting person names to culturally appropriate alternatives in target languages, and (iv) ensuring natural, native-sounding output. Samples with inadequate translation quality were discarded. For CommonsenseQA, we use `GPT-5-mini` for all language translations. All Urdu translations were verified and corrected by a native speaker. To enable rigorous cross-lingual evaluation, we ensure zero overlap between languages meaning that each unique problem appears in exactly one language, preventing any data leakage during training. Dataset statistics are provided in Appendix C.

2.2 Model

2.2.1 Base Architecture

We build upon `LLaMA3.2-1B-Instruct` (Grattafiori et al., 2024) as our base language model. This model officially supports 8 different languages (Grattafiori et al., 2024), but since it has been pre-trained on a large-scale web corpus containing numerous languages, it often performs reasonably well for languages that are not officially supported. This becomes apparent in table 5 where the base model achieves similar performance in German & French (which are officially supported) and Chinese (which is not officially supported).

2.2.2 Continuous Latent Reasoning via CODI

To enable reasoning in continuous latent space, we employ the CODI training framework (Shen et al., 2025). CODI trains a single shared model to jointly optimize two distinct reasoning modes: explicit token-based reasoning (teacher task) and implicit continuous reasoning (student task). This

self-distillation setup allows the model to compress verbose natural language reasoning steps into compact continuous representations while preserving reasoning capability.

Teacher Task: Explicit CoT Reasoning The teacher task learns standard chain-of-thought generation by predicting a sequence of explicit reasoning tokens c followed by the final answer y. The training objective is a standard cross-entropy loss:

$$\mathcal{L}_{\text{teacher}} = -\frac{1}{N}\sum_{i=1}^{N} \log P(r_i \mid r_{1:i-1}, Q), \tag{1}$$

where Q represents the input question, $r = [c, y]$ is the concatenated sequence of CoT and answer tokens, and N is the total sequence length.

Student Task: Continuous Thought Generation The student task generates reasoning in a continuous latent space by autoregressively propagating hidden states $Z = \{z_1, \ldots, z_K\}$ between learnable <bot> and <eot> tokens. An MLP projects these states between steps to distinguish latent reasoning from token embeddings. The objective is to predict the answer y given these thoughts:

$$\mathcal{L}_{\text{student}} = -\frac{1}{M}\sum_{i=1}^{M} \log P(y_i \mid y_{1:i-1}, Q, Z). \tag{2}$$

Knowledge Distillation Mechanism To transfer reasoning capabilities and prevent the continuous representations from drifting too far from the initial language representations, CODI aligns the student's hidden activations with the teacher's at the token immediately preceding the answer (e.g., ":"). This anchors the latent reasoning to the explicit trace via an L1 loss:

$$\mathcal{L}_{\text{KD}} = \frac{1}{L}\sum_{l=1}^{L} \|\text{sg}[\mathbf{h}^l_{\text{teacher}}] - \mathbf{h}^l_{\text{student}}\|_1, \tag{3}$$

where sg$[\cdot]$ is a stop-gradient ensuring unidirectional knowledge transfer.

Training Objective The complete training loss combines all three components:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{student}} + \beta\mathcal{L}_{\text{KD}} + \gamma\mathcal{L}_{\text{teacher}}, \tag{4}$$

2.2.3 CoT-SFT Baseline

As a baseline, we use equation 1 to train a standard Chain-of-Thought Supervised Fine-Tuning (CoT-SFT). Note that equation 1 excludes the prompt in the training objective, meaning that the model only receives gradients for predicting the reasoning trace.

3 Experiments

3.1 Training on English only

To establish a baseline performance, we train both CODI and CoT-SFT on English data (meaning that both the question and the CoT traces are in English). Our results can be found in table 1. Similar to (Shen et al., 2025) we found that CODI performs worse on GSM8K and better on CommonsenseQA when evaluated in English. Moreover, in line with (Barua et al., 2025; Lai & Nissim, 2024) we found that the performance gets significantly worse for both models in other languages. CODI performs however better than CoT-SFT in low-resource languages for both datasets.

3.2 Training on a Mixture of Languages

To improve the multilingual performance we investigate both CODI and CoT-SFT in two multilingual training configurations. In the first setup, we train on English, German, French, and Chinese, deliberately excluding Urdu to assess zero-shot generalization to a low-resource language. In the

Table 1: Accuracy (%) on GSM8K-AUG-NL and CommonsenseQA (test splits) comparing CODI vs. CoT-SFT when trained on English.

Test Language	**GSM8K-AUG-NL**		**CommonsenseQA**	
	CODI	**CoT-SFT**	**CODI**	**CoT-SFT**
German	31.79	36.09	58.48	38.41
French	32.35	38.21	57.66	46.03
Chinese	30.46	27.37	57.58	48.40
English	46.08	55.04	71.42	63.23
Urdu	9.58	5.38	32.57	21.87

Note: CODI preprocessing removes the final sentence from reference CoTs. In GSM8K (English train), **28.30%** of CoTs contain only one sentence, resulting in empty references which disadvantages CODI on that specific dataset.

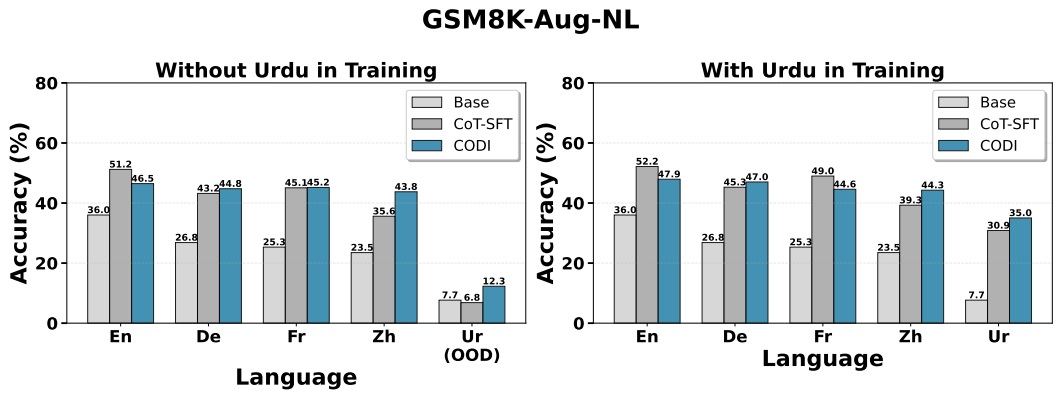

Figure 1: Performance of `LLaMA3.2-1B-Instruct` trained on multi-lingual GSM8k-Aug-NL data. Both models perform similarly, with CoT-SFT performing better for high-resource languages and CODI performing better for low-resource languages. Moreover, CODI performs significantly better than CoT-SFT on Urdu when it is Out-of-distribution (OOD) (i.e., when Urdu was not part of the fine-tuning data).

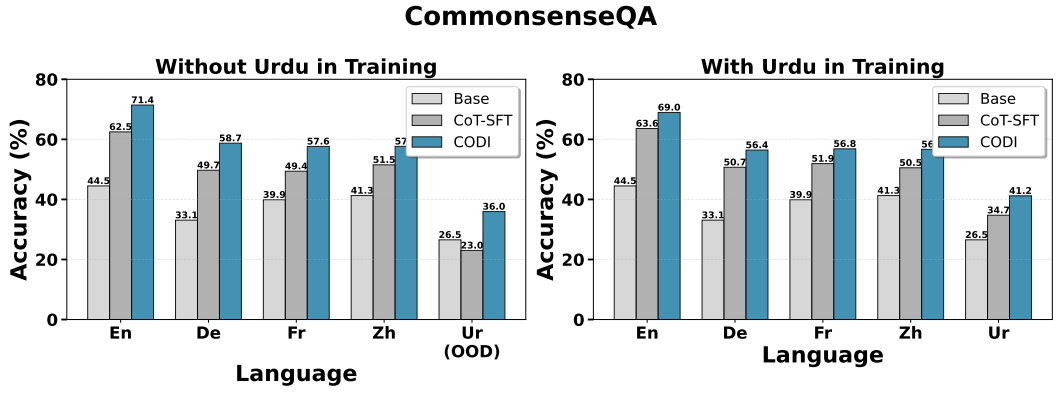

Figure 2: Performance of `LLaMA3.2-1B-Instruct` trained on multi-lingual CommonsenseQA data. CODI performs better across all languages.

second setup, we include Urdu during training to examine performance when the low-resource language is present in the training data.

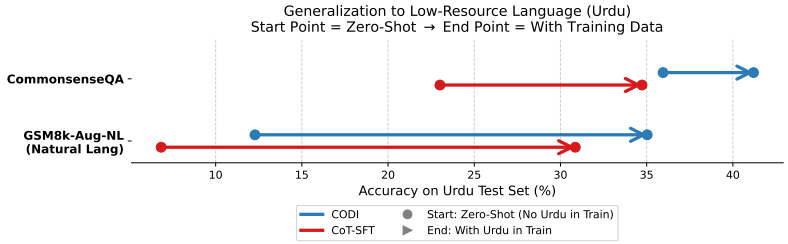

Figure 3: Performance on a low-resource language (Urdu). The latent reasoning approach works significantly better than CoT-SFT in all cases.

Both approaches substantially outperform the base model `LLaMA3.2-1B-Instruct` across all languages and also improve the multilingual performance compared to english-only training, as evident in figures 1 & 2. Moreover, contrary to the single-language results in section 3.1 (and the results reported in (Shen et al., 2025)) where CODI performs worse on GSM8K, we find that CODI matches the performance of classical CoT-SFT much better, despite using significantly less output tokens (see section 3.3). To be more precise, our results for GSM8K show that CODI performs noticeably better than COT-SFT for low-resource languages and slightly worse for high-resource languages. For COMMONSENSEQA, CODI performs better among all languages.

Remarkably, CODI performs significantly better on Urdu even if it has not been trained on it during fine-tuning (see figure 3). To emphasize, even when Urdu is out-of-distribution, CODI performs better on COMMONSENSEQA (35.95%) than CoT-SFT which had Urdu in the fine-tuning data (34.73%). This indicates that the latent thinking of CODI is much more capable of generalizing to multiple languages compared to classical CoT-SFT.

Another seemingly natural configuration would be to receive the question in the target language, perform the thinking in a high-resource pivot language (i.e. English), and then respond in the target language again. This approach has already been investigated by (Barua et al., 2025) and (Lai & Nissim, 2024), which both concluded that this performs worse than performing CoT in the target language. We therefore did not investigate this further here.

3.3 NUMBER OF THINKING TOKENS: COT-SFT VS. CODI

One of the main benefits of CODI is that it requires much fewer thinking tokens compared to classical CoT-SFT (Shen et al., 2025). This becomes obvious in table 2 where, even when trained on all target languages ("Multi-Lingual-with-Urdu" setup), standard latent reasoning achieves compression ratios of roughly $29\times$ and $50\times$.

Table 2: Efficiency comparison between explicit and continuous reasoning. Average CoT lengths are calculated across all 5 test languages using the *Multi-Lingual-with-Urdu* training setup (see Tables 8 & 9 in the appendix).

Dataset	**CoT-SFT (Avg Tokens)**	**CODI (Latent Tokens)**	**Compression Ratio**
GSM8K-Aug-NL	$\sim$176	6	$\sim$29$\times$
CommonsenseQA	$\sim$299	6	$\sim$50$\times$

4 CONCLUSION

Performing CoT reasoning in continuous latent space instead of token space offers several advantages, such as requiring fewer thinking tokens (Shen et al., 2025) and performing several thoughts in parallel (Hao et al., 2024). In this work we found empirical evidence that continuous CoT also learns more language-agnostic representations which allows it to generalize better to new languages and perform better in low-resource languages.

In future work, we plan to directly analyze the learned continuous representations to verify their language-agnostic properties. Additionally, we aim to scale this investigation to larger models and a broader range of datasets to determine if these findings hold across different architectures and domains.

5 ACKNOWLEDGMENTS

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

REFERENCES

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGA: Multilingual evaluation of generative AI. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.258. URL https://aclanthology.org/2023.emnlp-main.258/.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. Long chain-of-thought reasoning across languages. *arXiv preprint arXiv:2508.14828*, 2025.

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

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2024. URL https://arxiv.org/abs/2412.06769.

Huiyuan Lai and Malvina Nissim. mcot: Multilingual instruction tuning for reasoning consistency in language models. *arXiv preprint arXiv:2406.02301*, 2024.

M. Paul Lewis (ed.). *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition, 2009.

Yihong Liu, Raoyuan Zhao, Hinrich Schütze, and Michael A Hedderich. Large reasoning models are (not yet) multilingual latent reasoners. *arXiv preprint arXiv:2601.02996*, 2026.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

Gabriel Nicholas and Aliya Bhatia. Lost in translation: Large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*, 2023.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation, 2025. URL https://arxiv.org/abs/2502.21074.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

A APPENDIX

Table 3: Accuracy (%) on GSM8K-Aug-NL (test split): CODI vs. COT-SFT across training setups.

Training Dataset	**Train Setup**	**Test Language**	**Accuracy (%)**	
			CODI	**COT-SFT**
GSM8K-Aug-NL	Multi-Lingual	German	44.76	43.20
		French	45.19	45.07
		Chinese	43.75	35.60
		English	46.47	51.20
		Urdu	12.28	6.84
	Multi-Lingual-with-Urdu	German	47.01	45.26
		French	44.58	48.98
		Chinese	44.27	39.27
		English	47.93	52.16
		Urdu	35.03	30.86

Table 4: Accuracy (%) on CommonSenseQA (test split) across test languages for different training setups (CODI vs. COT-SFT).

Training Dataset	**Train Setup**	**Test Language**	**Accuracy (%)**	
			CODI	**COT-SFT**
CommonSenseQA	Multi-Lingual	German	58.72	49.71
		French	57.65	49.39
		Chinese	57.66	51.52
		English	71.42	62.49
		Urdu	35.95	23.01
	Multi-Lingual-with-Urdu	German	56.43	50.70
		French	56.84	51.92
		Chinese	56.67	50.53
		English	68.96	63.61
		Urdu	41.20	34.73

Table 5: Base model accuracy (%) on GSM8K-Aug-NL by test language (independent of fine-tuning setup).

Test Language	Base Accuracy (%)
German	26.84
French	25.32
Chinese	23.50
English	36.01
Urdu	7.66

Table 6: Base model accuracy (%) on CommonSenseQA (test split) by test language (independent of fine-tuning setup).

Test Language	Base Accuracy (%)
German	33.09
French	39.89
Chinese	41.28
English	44.49
Urdu	26.54

B Training Hyperparameters

C Dataset Construction Details

C.1 GSM8k-Aug-NL Multilingual Dataset

We construct multilingual versions of GSM8k-Aug-NL by translating the English dataset into German, French, Chinese, and Urdu. Each translated sample maintains a reference to its English original, enabling cross-language deduplication to ensure zero overlap, each mathematical problem appears in exactly one language.

For the five-language configuration, we prioritize Urdu (4,619 samples) to maximize low-resource coverage, then distribute remaining problems across high-resource languages based on translation quality.

C.2 CommonsenseQA Multilingual Dataset

We apply the same methodology to CommonsenseQA, ensuring zero overlap across languages.

Table 7: Training hyperparameters for CODI and CoT-SFT. All parameters are kept constant across both GSM8k-Aug-NL and CommonsenseQA datasets to ensure fair comparison.

Hyperparameter	**Value**
Shared Parameters (CODI & CoT-SFT)	
Learning Rate	8×10^{-4}
LR Scheduler	Cosine
Warmup Ratio	0.03
Weight Decay	0.1
Max Gradient Norm	2.0
Training Epochs	10
Max Sequence Length	512
Precision	BF16
Random Seed	11
Effective Batch Size	128
LoRA Configuration	
Rank (r)	128
Alpha (α)	32
Dropout	0.1
Target Modules	q, k, v, o, gate, up, down
CODI-Specific Parameters	
Number of Latents	6
Projection Dimension	2048
Distillation Loss Factor	20
Normalize Distillation Loss	Yes

Table 8: Average Chain-of-Thought (CoT) length (tokens) on GSM8K_AUG_NL across training setups and test languages (CoT-SFT).

Training Dataset	**Train Setup**	**Test Language**	**Avg CoT Length (tokens)**
GSM8K_AUG_NL	English	German	172.2
		French	161.6
		Chinese	171.4
		English	155.4
		Urdu	512.0
	Multi-Lingual	German	239.6
		French	211.7
		Chinese	175.9
		English	156.3
		Urdu	503.0
	Multi-Lingual-with-Urdu	German	187.9
		French	186.8
		Chinese	151.0
		English	154.0
		Urdu	199.3

Table 9: Average Chain-of-Thought (CoT) length (tokens) on CommonsenseQA across training setups and test languages (CoT-SFT).

Training Dataset	Train Setup	Test Language	Avg CoT Length (tokens)
CommonsenseQA	English	German	455.1
		French	329.1
		Chinese	203.8
		English	205.5
		Urdu	512.0
	Multi-Lingual	German	296.8
		French	286.7
		Chinese	228.4
		English	208.0
		Urdu	512.0
	Multi-Lingual-with-Urdu	German	303.5
		French	294.7
		Chinese	215.2
		English	207.6
		Urdu	476.0

Table 10: GSM8k-Aug-NL multilingual dataset statistics.

Language	Samples	% of Total
English	95,131	24.5%
German	96,139	24.8%
French	95,868	24.7%
Chinese	96,178	24.8%
Urdu	4,619	1.2%
Total	**387,935**	100%

Table 11: CommonsenseQA multilingual dataset statistics.

Language	Samples	% of Total
English	1,940	20.0%
German	1,940	20.0%
French	1,940	20.0%
Chinese	1,940	20.0%
Urdu	1,940	20.0%
Total	**9,700**	100%