Appendix for Paper "Reasoning Transfer for an Extremely Low-resource and Endangered Language: Bridging Languages through Sample-Efficient Language Understanding", Id 20089

Anonymous submission

Detailed Experiment Setup

Baselines. We select r1-distill-Llama-8B (DeepSeek-AI 2025) and DeepHermes-3-Llama-3-8B (Teknium et al. 2025) due to their state-of-the-art performance and open-source availability. We perform fine-tuning on these models following our proposed approach (*English-Pivoted CoT Training*). We leverage the same set of hyperparameters for English-Pivoted CoT Training and the baseline approach of multilingual reasoning fine-tuning (denote *Native CoT Training*) (Anand et al. 2024; Lai and Nissim 2024): maximum input sequence length of 16,384 tokens, AdamW optimizer (Loshchilov and Hutter 2019), training for 3 epochs, learning rate of 1×10^{-5}. Training is distributed across two Nvidia A100 GPUs (80GB each), with a total batch size of 24. Each training run can be completed under 12 hours.

For TIES-merging, follow (Tao et al. 2024), we set density for math-specific model to 0.2, and 1.0 for language-specific model. For layer-swap, as we benchmark on the exact same 32-layer Llama 3.1 8B architecture used in the original paper, we leverage their best found setting (e.g., 5 bottom layers and 2 top layers). For reproducibility, both model editing techniques are performed between the two open-weight models: r1-distill-Llama-8B which focuses of mathematical reasoning tasks, and the general language-specific model Llama-3.1-8B-Instruct (Grattafiori et al. 2024).

Evaluation setup. We adopt each benchmark's default hyperparameters, enabling sampling a decoding temperature of 0.6, top-p sampling with $p = 0.95$, limiting generation outputs to 16,384 tokens. Moreover, for reliable performance estimation, we generate 64 responses per benchmark for AIME2024, Irish AIME2024, and LC2024, and report the average accuracy across them (DeepSeek-AI 2025).

Additional Results

Improved language understanding capability. Table 1 compares all methods on the two LC2024 splits. Baseline results are inconsistent, for example, few-shot prompting with Irish CoT boosts r1-distill-Llama-8B on contexts & applications but harms DeepHermes-3-Llama-3-8B. In contrast, our approach, English-Pivoted CoT Training delivers consistent gains on both splits and across all models.

Diverged reasoning traces when forced to reason in different languages. We extend our test-time CoT-language control experiments to DeepHermes-3-Llama-3-8B. As shown in Table 2, models prompted to generate CoTs in English consistently outperform those forced to use Irish across all baselines. This confirms that in data-scarce settings, steering the model to "think" in English remains the most effective strategy for reliable multilingual reasoning.

Ethics Statement

Our work advances language technologies for low-resource and endangered languages, with a particular focus on Irish. By leveraging English-Pivoted CoT Training, we aim to enhance the accessibility and usability of these languages, thereby promoting linguistic diversity and preserving cultural heritage.

All training and evaluation data are drawn from publicly available sources. Our newly released LC2024 dataset is based on Regulations on Re-use of Public Sector Information 2005, SI 279 of 2005, and is licensed for non-commercial research and educational use only.

Limitations

While English-Pivoted CoT Training demonstrates strong gains on mathematical reasoning benchmarks, its efficacy beyond non-verifiable domain (e.g., commonsense) remains to be validated. Additionally, our experiments are restricted to bilingual settings across three resource regimes (low, medium, high). An interesting future research direction can be in exanding our approach to multilingual scenarios.

References

Anand, A.; Prasad, K.; Kirtani, C.; Nair, A. R.; Nema, M. K.; Jaiswal, R.; and Shah, R. R. 2024. Multilingual Mathematical Reasoning: Advancing Open-Source LLMs in Hindi and English. arXiv:2412.18415.

Bandarkar, L.; Muller, B.; Yuvraj, P.; Hou, R.; Singhal, N.; Lv, H.; and Liu, B. 2025. Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Model	Setting	concepts & skills	contexts & applications
r1-distill-Llama-8B	Zero-shot Prompting	**84.85**	31.82
r1-distill-Llama-8B	Few-shot Prompting - Irish CoT	37.50	56.52
r1-distill-Llama-8B	Few-shot Prompting - English CoT	28.13	39.13
r1-distill-Llama-8B	TIES-merging (Yadav et al. 2023; Tao et al. 2024)	40.63	26.09
r1-distill-Llama-8B	Layer Swap (Bandarkar et al. 2025)	28.13	34.78
r1-distill-Llama-8B	Native CoT Training (Anand et al. 2024; Lai and Nissim 2024)	47.50	31.82
r1-distill-Llama-8B	English-Pivoted CoT Training (OURS)	82.82	**59.09**
DeepHermes-3-Llama-3-8B	Zero-shot Prompting	57.58	45.45
DeepHermes-3-Llama-3-8B	Few-shot Prompting - Irish CoT	21.88	26.09
DeepHermes-3-Llama-3-8B	Few-shot Prompting - English CoT	34.38	30.43
DeepHermes-3-Llama-3-8B	TIES-merging (Yadav et al. 2023; Tao et al. 2024)	18.75	8.70
DeepHermes-3-Llama-3-8B	Layer Swap (Bandarkar et al. 2025)	25.00	26.09
DeepHermes-3-Llama-3-8B	Native CoT Training (Anand et al. 2024; Lai and Nissim 2024)	57.58	13.64
DeepHermes-3-Llama-3-8B	English-Pivoted CoT Training (OURS)	**57.58**	**50.00**

Table 1: Comparisons of low-resource language understanding on LC2024, where the exam is split into 2 parts: concepts & skills and contexts & applications. The latter part has additional contextual and real-world descriptions, requiring a higher level of Irish language understanding to comprehend the input. Bold scores indicate the best performance per benchmark.

Model	Training Setting	Prompting Setting	AIME2024	LC2024 - concepts&skills	LC2024 - contexts&applications
r1-distill-Llama-8B	Mixed CoT Language	Irish CoT	33.33	3.33	54.55
	Mixed CoT Language	English CoT	33.33	30.00	61.82
	Two-stage	Irish CoT	40.00	6.67	45.45
	Two-stage	English CoT	33.33	33.33	58.18
DeepHermes-3-Llama-3-8B	Mixed CoT Language	Irish CoT	0.00	0.00	20.00
	Mixed CoT Language	English CoT	0.00	3.33	34.55
	Two-stage	Irish CoT	3.33	0.00	20.00
	Two-stage	English CoT	3.33	3.33	47.27

Table 2: Irish task performance by CoT language for models trained on both English and Irish CoTs.

Grattafiori, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Lai, H.; and Nissim, M. 2024. mCoT: Multilingual Instruction Tuning for Reasoning Consistency in Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12012–12026. Bangkok, Thailand: Association for Computational Linguistics.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Tao, M.; Zhang, C.; Huang, Q.; Ma, T.; Huang, S.; Zhao, D.; and Feng, Y. 2024. Unlocking the Potential of Model Merging for Low-Resource Languages. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 8705–8720. Miami, Florida, USA: Association for Computational Linguistics.

Teknium; Jin, R.; Guang, C.; Suphavadeeprasit, J.; and Quesnelle, J. 2025. DeepHermes 3 Preview.

Yadav, P.; Tam, D.; Choshen, L.; Raffel, C.; and Bansal, M. 2023. TIES-Merging: Resolving Interference When Merging Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.