Large Reasoning Models Struggle to Transfer Parametric Knowledge Across Scripts

Lucas Bandarkar[§] **Alan Ansell** **Trevor Cohn**[ϕ]
Google Research
[§]University of California, Los Angeles [ϕ]University of Melbourne

Abstract

In this work, we analyze shortcomings in cross-lingual knowledge transfer in large, modern reasoning LLMs. We demonstrate that the perceived gap in knowledge transfer is primarily a script barrier. First, we conduct an observational data analysis on the performance of thinking models on two datasets with local knowledge from around the world, ECLeKTic and MultiLoKo. Our regression analysis shows that script match—not language or family—is the primary predictor of knowledge transfer failure once model capability and question difficulty are accounted for. We further this finding by providing the LLMs with the key entities of the questions in their source language and find that this disproportionately improves cross-script questions. We then posit that these LLMs could be reasoning better at test-time. To evaluate this, we develop a synthetic generation pipeline to design SFT samples to encourage the model to better reason about transliteration ambiguities when trying to fetch parametric knowledge at inference-time. We show that teaching two models to reason better reduces the cross-script transfer gap. As a result, we conclude that there is potential to improve cross-lingual parametric knowledge transfer during post-training.

1 Introduction

The recent scaling of parameters and training data has increased the multilingual capabilities of Large Language Models (LLMs), leading to remarkable cross-lingual transfer across domains and tasks (Hu et al., 2025b; He et al., 2025). While these modern models exhibit a huge capacity to learn, retain, and retrieve factual global knowledge, issues regarding inconsistency across languages remain (Romanou et al., 2025; Wang et al., 2025; Guo et al., 2026).

In this work, we analyze the shortcomings in cross-lingual knowledge transfer within large, modern reasoning LLMs and find that language script remains a notable barrier. Inconsistency is higher when there is a mismatch between the script of the query and the script of origin of the knowledge (e.g. a query in English about a Serbian monastery). Utilizing two local knowledge QA evaluation datasets, ECLeKTic (Goldman et al., 2025) and MultiLoKo (Hupkes and Bogoychev, 2025), we first conduct an observational data analysis across numerous thinking models in Section 4. We find that when controlling for the statistical difficulty of questions and the model's inherent capability (using reading comprehension results as a proxy), cross-lingual gaps are statistically insignificant when the source and target languages share a script. However, languages not sharing a script leads to a 13% lower accuracy, implying significant transfer limitations when retrieving knowledge across scripts. In Section 5, we identify cross-lingual mapping of entities as one of the drivers of this script barrier. We show that prefixing a model's English thinking trace with the original-language entity disproportionately improves performance on cross-script questions, highlighting a specific failure in cross-script entity alignment.

Finally, positing that reasoning LLMs possess the latent capacity to better reason across scripts, we develop a synthetic generation pipeline to design Supervised Fine-Tuning (SFT) samples. We demonstrate that teaching the model to explicitly reason about transliteration ambiguities during inference significantly reduces the cross-script transfer gap. This suggests a large proportion of the gap from cross-script limitations can cheaply be solved.

2 Related Work

To address constraints to model vocabulary, numerous works proposed different uses of transliteration as test-time solutions for cross-script transfer limitations (Amrhein and Sennrich, 2020; Muller et al., 2021; Purkayastha et al., 2023; J et al., 2024). More recent works on LLMs identify persistent cross-lingual inconsistency, especially in knowl-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 29519–29529
October 24-29, 2026 ©2026 Association for Computational Linguistics

edge (Wang et al., 2025), with some noting script's high influence (Beniwal et al., 2024; Nie et al., 2025). Solutions often involve better aligning representations to enhance cross-lingual sharing and transfer (Lu et al., 2025; Sun et al., 2026). In analyses of knowledge representations, both Saji et al. (2025) and Ifergan et al. (2025) identify script-specific representations.

3 Setup

We work with five state-of-the-art reasoning LLMs, Qwen3-30B-A3B[1], Qwen3-235B-A22B[2] (Yang et al., 2025), GLM-4.5-Air (Team et al., 2025), GPT-OSS-120B (OpenAI: et al., 2025), and Olmo-3.1-32B[3] (Olmo: et al., 2025). While the Qwen3 series is the most multilingual, the 30-32B models naturally lag the larger models in performance.

We focus on two cross-lingual knowledge QA datasets, MultiLoKo (Hupkes and Bogoychev, 2025) and ECLeKTic (Goldman et al., 2025). While designed differently, both datasets attempt to ask questions about rare knowledge likely only present in one or very few languages in an LLMs' training corpus. Each question has one labeled source language, which allows us to study the cross-lingual transfer dynamics in directions beyond Eng->X. We use *cross-script questions* to refer to questions where the language of the question differs from the source language. We note that across many runs with slightly perturbed conditions (prompts, hyperparameters, etc.), there was significantly more variance for individual questions in low-resource languages, in line with Piratla et al. (2025).

4 Observational Data Analysis

To determine which factors most limit a model's ability to use its knowledge across different languages, we conduct a regression analysis on ECLeKTic and MultiLoKo with our table of model results for each question. For features, we start with indicators for the model and the source language to account for different model capabilities and question subset difficulties. Rather than relying on categorical indicators for the language of the question (i.e., target language), we use the model's accuracy ($\in [0,1]$) on the Belebele reading comprehension task (Bandarkar et al., 2024) as a continuous feature. This serves as a direct proxy for the model's "raw" linguistic capability in that language[4], reducing the feature space and enhancing interpretability. We argue this captures cross-lingual proficiency more granularly than language-specific fixed effects all while increasing degrees of freedom for our regression model. And finally, we add binary variables to represent the relationship between the source and target languages: whether they are the same language family[5], script, or identity. Note that since we do not use text-level features, many questions shared the same set of features. For ECLeKTic, this resulted in 144 unique set of features with an average of 32 questions each, while for MultiLoKo, this resulted in 91 unique sets with exactly 250 questions each.

Table 1: OLS Coefficients (β) when trying to predict the binary score of a question, **with no scaling**. *n.s.* denotes statistically insignificant (β); $p > 0.05$.

Variable	**ECleKTic**	**MultiLoKo**
OLS Intercept	−0.350	−0.439
Belebele Score	0.999	0.937
Script Match	0.121	0.133
Language Family	*n.s.*	*n.s.*
In-Language	*n.s.*	*n.s.*
Model Indicators (vs. Qwen3-30B)		
GPT-OSS-120B	0.119	0.065
Olmo-3.1-32B	0.061	0.091
GLM-4.5-Air	0.040	*n.s.*
Qwen3-235B	0.038	0.031
Src Lang Indicators[†] (vs. EN)	[-0.300, -0.059]	[*n.s.*, 0.330]

[†] Range of β for source-language subset indicators used to account for subset difficulty.

Using these features, we fit an Ordinary Least Squares (OLS) regression model to predict the binary feature of the actual model score (1.0 or 0.0) on a given question given the features to disentangle the drivers of cross-lingual gaps (See full details in Appendix B, including a discussion on OLS vs. logistic regression). Although we fit separate models for ECLeKTic and MultiLoKo, the results are remarkably consistent as displayed in

[1]Specifically, Qwen3-30B-A3B-Thinking-2507
[2]Specifically, Qwen3-235B-A22B-Thinking-2507
[3]Specifically, Olmo-3.1-32B-Think

[4]The task of reading comprehension requires no external knowledge or reasoning, since the context is provided. As a result, it is solely dependent on textual understanding.
[5]Using the classification in FLORES (Goyal et al., 2022)

Table 1. The Belebele score—our proxy for "raw" linguistic capability—has the highest coefficient but also is the only continuous variable (with a much smaller standard deviation), meaning it is not directly comparable to the indicator variables. However, in Table 4 we provide coefficients with all features standard-scaled, and show that Belebele score captures the most variance in general model capabilities. The fact that simple linguistic understanding is the dominant predictor supports the continued utility of translation evaluations as proxies for multilingual task benchmarks (Issaka et al., 2026).

Most notably, however, whether or not the source and target language were the same, or even in the same family, has a statistically *in*significant coefficient β[6] when regressed alongside script match. This indicator has a very high β, implying a 12.1% and 13.3% increase in probability of correct response if the target and source languages share a script. This demonstrates significant transfer limitations in retrieving knowledge across scripts and that, when controlling for this, transfer across languages within the same script, or even family is only limited by the model's "raw" performance in the target language. For now, this is purely observational and so we explore this further.

5 Prefixing Thinking

We evaluate the hypothesis that cross-script gaps stem in part from a model's inability to correctly link key entities across languages. We do so by prefilling the beginning of each model's thinking trace and providing the key entity in the source language. This eliminates the potential source of retrieval error from failed cross-lingual entity mappings.

We start by using Gemini-3-Pro (DeepMind, 2025) to identify important entities (notably named entities) in all questions in their original source language. We prompt each model and force its reasoning to begin with a specific prefix similar to "<think> Okay, the key entity of the question in its source language is {src_entity}". As a result, the model is provided the pre-translation entities central to the question. Note that, irrespective of the language of the question, all 5 models almost always produce English thinking tokens. GLM and Qwen3 models, however, sometimes produce Chinese thinking when prompted in Chinese. In this experiment, we implicitly force it to think in English. As a control, we also evaluate the model with that entire prefix minus the actual "{src_entity}" at the end. To maintain the natural distribution, we stylize the prefix in accordance with the wording typically used by each model to start its thinking (provided in Appendix C). By providing the entity as an anchor, we test whether the primary bottleneck is the cross-lingual entity mapping.

As expected, providing the pre-translation entity of the question resulted in positive Δ on ECLeKTic and MultiLoKo accuracy. On all types of questions, we notice a small uptick on average, indicating that cross-lingual mapping of the entities was a limitation for such cross-lingual knowledge retrieval, as displayed in Table 2.

Condition	All	Non-Latin	Latin	Script $\neq$	Script =
GPT-OSS-120B					
Original	20.6	16.3	23.6	18.5	24.6
Prefix, Control	21.2	17.9	23.5	19.3	24.8
Prefix w/ src Ent.	22.2	19.4	24.2	20.7	25.1
Δ	+1.0	+1.5	+0.7	+1.4	+0.2
Qwen3-235B-A22B					
Original	28.3	24.2	31.2	26.0	32.8
Prefix, Control	28.3	24.5	31.1	26.1	32.7
Prefix w/ src Ent.	30.0	27.1	32.2	28.6	32.7
Δ	+1.7	+2.6	+1.1	+2.5	+0.0
GLM-4.5-Air					
Original	25.5	21.2	28.6	22.6	29.6
Prefix, Control	25.2	20.5	28.6	22.6	30.1
Prefix w/ src Ent.	26.9	22.8	29.9	24.7	31.2
Δ	+1.7	+2.3	+1.3	+2.1	+1.1
Average Δ	+1.5	+2.1	+1.0	+2.1	+0.5

Table 2: Impact of our think-prefixing on ECLeKTic accuracy ↑. Script $\neq$ denotes cross-script; Script = denotes same-script.

However, this sort of prefixing seems to disproportionately impact cross-script questions. For example, for Qwen3-235B-A22B, the impact from prefixing the pre-translation entity is negligible for all other questions, while there is a 2.5-point boost for cross-script questions. Averaged across three models, the Δ for cross-script questions is $+2.1$ while being just $+0.5$ for the rest. This further con-

[6]Concretely, the null hypothesis $\beta = 0$ cannot be rejected.

firms our hypothesis that script is a barrier for these models, negatively impacting knowledge transfer.

6 Reducing Gaps using SFT

Based on qualitative and quantitative analyses of thinking traces in Sections 4 and 5, we hypothesize that these LLMs have the capacity to better reason over the potential transliteration ambiguities in the question. For example, we see in Table 2 that the prefix *without* the source entity provided (labeled as "Prefix, Control", see Appendix C) often leads to some increase in performance for GPT-OSS-120B. This simple modification to the beginning of the thinking trace implies there is a way to generally improve these model's reasoning.

To evaluate this, we experiment with teaching the model to reflect carefully about what entity the translated question could be referring to. This is done via LoRA Supervised-Finetuning (SFT) on two of the smaller (but still large) models Qwen3-30B-A3B and GPT-OSS-120B.

6.1 Synthetic Data Generation

To create SFT data, we use a multi-step pipeline that uses Gemini to generate the reasoning logic that we want to teach the models. However, due to the sensitivity of SFT to formatting (Zhou et al., 2023; Chu et al., 2025; Matsutani et al., 2026), we additionally use style transfer to ensure the training samples resemble each model's own style and formatting. This stylistic alignment ensures the model can learn the reasoning patterns without the distraction of unfamiliar formatting.

1. We prompt Gemini to create multilingual knowledge questions and answers. Alongside, we instruct Gemini to create a thinking trace where the reasoner reflects carefully about the potential transformations during translation that could have led to the resulting entity in the question. We iterate through prompts and filtering mechanisms until qualitative checks of the generated questions reveal quality samples. In the end, we use a variety of prompts to induce more variation (an example is provided in Appendix E).

2. Using only the question generated by Gemini, we prompt each of the three models to answer the questions, without any special instructions.

3. Then, we provide the Gemini question, thinking, answer in context and prompt the target model to regenerate the thinking trace in its own style. We provide its own unadulterated thinking trace on that question to demonstrate the difference in style between it and Gemini. We provide the exact instruction in Appendix F.

We provide additional details on this data generation pipeline in Appendix D. The resulting data is in a variety of languages, but as specified in our instructions, the question and answer are the same language and the thinking is English.

6.2 SFT Results

Then, we perform LoRA SFT using Llama-Factory (Zheng et al., 2024) using these training samples. We provide specifics in Appendix H. Our qualitative analysis of the thinking traces reveals a change in the first few sentences of the thinking trace to reflect over possible transliterations, as desired.

As displayed in Table 3, we can achieve modest improvement across both models, evidenced on both ECLeKTic and MultiLoKo. Through a simple SFT of just a few hundred samples and 8 epochs, we are able to teach the model to better stop and think about the potential entity of origin of the question. While this does not help the model gain more knowledge, the fact that simply reasoning differently enables a few point boost shows that its cross-lingual reasoning behavior is not optimal.

		ECLeKTic (12 languages)							**MultiLoKo** (31 languages)						
			Latin		Resource		Script			Latin		Resource		Script	
Model	**Run**	**All**	**No**	**Latin**	**Low**	**High**	$\neq$	$=$	**All**	**No**	**Latin**	**Low**	**High**	$\neq$	$=$
GPT-OSS-120B	Base	20.6	16.3	23.6	12.6	23.9	18.5	24.6	24.4	18.0	31.2	17.0	27.8	21.6	27.3
	LoRA	21.6	18.6	23.8	15.3	24.0	20.1	24.7	25.0	18.7	32.0	17.7	28.3	22.3	27.4
Qwen3-30B-A3B	Base	18.1	15.1	20.1	12.2	20.0	16.9	20.3	25.6	21.4	30.1	19.6	27.1	21.1	28.0
	LoRA	18.9	16.3	20.8	12.8	18.2	18.0	20.6	26.0	21.9	30.4	20.2	27.4	21.7	28.2

Table 3: Comprehensive evaluation of Base vs. SFT across two datasets. We compare performance across Latinity (Non-Latin vs. Latin), Resource levels (Low vs. High), and Script matching ($\neq$ vs. $=$).

Notably, the performance increase happens significantly more with low-resource languages and on cross-script questions, as was hypothesized. This suggests that these model's barrier around script is not necessarily a fundamental one, but one relating to its internal retrieval mechanisms.

7 Discussion and Conclusion

Our investigation into the cross-lingual capabilities of state-of-the-art reasoning LLMs reveals that the "language barrier" in knowledge retrieval is often, more accurately, a "script barrier", most notably defined in Section 4. Intuitively, this could be expected. When a query and the relevant parametric knowledge share a script (e.g., Spanish to Indonesian), there is likely subword token overlap that allow for easy recall. Conversely, in cross-script scenarios, disjoint token sets force the model to rely entirely on abstract, semantic mappings. Even so, the persistence of this barrier in massive reasoning models is surprising. Current literature suggests that as models scale, they develop language-agnostic representations, parameterizing knowledge redundantly in a way that should be robust to surface-level shifts (Lim et al., 2025; Chen et al., 2025; Dumas et al., 2025; Bandarkar et al., 2026). In addition, these models are all generating English thinking traces. Even so, our results suggest that knowledge tasks remain subject to syntax-induced latent processes (Hu et al., 2025a), as our think prefixing experiments in Section 5 show cross-script entity mappings are a cause for this gap.

Crucially, our experiments reveal that this struggle is not merely one of representation, but of reasoning. The success of our simple SFT runs in Section 6 demonstrates that LLMs sometimes possess the requisite parametric knowledge but fail to access it during reasoning. By simply teaching the model to reflect more carefully about transliteration ambiguities and entity origins, we achieved modest improvements across models.

Consequently, our results suggest that the most accessible solution for reducing cross-lingual knowledge gaps lies in post-training reasoning improvements. While we utilized SFT due to resource constraints, we hypothesize that Reinforcement Learning (RL) could more effectively elicit these specific reasoning behaviors. However, relying solely on inference-time reasoning is likely insufficient; fundamental improvements must also come from ensuring knowledge representations are more distant from token-level forms, an area requiring further research. Ultimately, we identify that the largest potential for gain lies at the intersection of low-resource languages and cross-script settings, where simple reasoning interventions can unlock significant latent knowledge.

Limitations

Real-World Transliteration Ambiguities A factor potentially weighing on cross-script issues is from the evaluation dataset construction. In the translation of questions, it is possible that translation ambiguity and even errors are higher for cross-script translations. This could arise from rare entities from one part of the world not having transliterations that are well-agreed upon. However, this would affect our *observational* regression analysis in Section 4. Our next two controlled experiments analyses in Sections 5 and 6, analyze the Δ and find it to be higher to cross-script questions, which is separate from inherent question difficulty due to transliteration ambiguities.

SFT rather than RL For Section 6, we acknowledge that Reinforcement Learning methods are currently the state-of-the-art for eliciting particular reasoning behavior, but we elect LoRA Supervised-Finetuning (SFT) due to resource constraints.

Use of AI Assistants

We used Gemini for assistance with drafting language in the paper, code development, and research ideation. Any output used was reviewed and edited by the authors, who remain fully responsible for the content.

Acknowledgments

We acknowledge Vihari Piratla, Eleftheria Briakou, Guido Zuccon, and Raj Dabre for their feedback and support during this work.

References

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

J.D. Angrist and J.S. Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. 2026. Multilingual routing in mixture-of-experts. In *The Fourteenth International Conference on Learning Representations*.

Himanshu Beniwal, Kowsik D, and Mayank Singh. 2024. Cross-lingual editing in multilingual language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2078–2128, St. Julian's, Malta. Association for Computational Linguistics.

Yuxin Chen, Yiran Zhao, Yang Zhang, An Zhang, Kenji Kawaguchi, Shafiq Joty, Junnan Li, Tat-Seng Chua, Michael Qizhe Shieh, and Wenxuan Zhang. 2025. The emergence of abstract thought in large language models beyond any language. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

Google DeepMind. 2025. Gemini 3 pro model card. Technical report, Google.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. A framework for few-shot language model evaluation.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *Preprint*, arXiv:2502.21228.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Pei-Fu Guo, Yun-Da Tsai, Chun-Chia Hsu, Kai-Xin Chen, Ya An Tsai, Kai-Wei Chang, Nanyun Peng, Mi-Yen Yeh, and Shou-De Lin. 2026. LiveCLKTBench: Towards reliable evaluation of cross-lingual knowledge transfer in multilingual LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15193–15208, San Diego, California, United States. Association for Computational Linguistics.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. 2025. Scaling laws for multilingual language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4257–4273, Vienna, Austria. Association for Computational Linguistics.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025a. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.

Songbo Hu, Ivan Vulić, and Anna Korhonen. 2025b. Quantifying language disparities in multilingual large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4003–4018, Suzhou, China. Association for Computational Linguistics.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *CoRR*, abs/2504.10356.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2025. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4630–4644, Albuquerque, New Mexico. Association for Computational Linguistics.

Sheriff Issaka, Erick Rosas Gonzalez, Lieqi Liu, Evans Kofi Agyei, Lucas Bandarkar, Nanyun Peng, David Ifeoluwa Adelani, Francisco Guzmán, and Saadia Gabriel. 2026. Translation as a scalable proxy for multilingual evaluation. *Preprint*, arXiv:2601.11778.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of large language

models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2025. Language-specific latent process hinders cross-lingual performance. *Preprint*, arXiv:2505.13141.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15066–15096, Suzhou, China. Association for Computational Linguistics.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2026. RL squeezes, SFT expands: A comparative study of reasoning LLMs. In *The Fourteenth International Conference on Learning Representations*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Ercong Nie, Bo Shao, Mingyang Wang, Zifeng Ding, Helmut Schmid, and Hinrich Schuetze. 2025. BMIKE-53: Investigating cross-lingual knowledge editing with in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16357–16374, Vienna, Austria. Association for Computational Linguistics.

Team Olmo:, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

OpenAI:, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vihari Piratla, Purvam Jain, Darshan Singh, Partha Talukdar, and Trevor Cohn. 2025. Rethinking cross-lingual gaps from a statistical viewpoint. *Preprint*, arXiv:2510.15551.

Sukannya Purkayastha, Sebastian Ruder, Jonas Pfeiffer, Iryna Gurevych, and Ivan Vulić. 2023. Romanization-based large-scale adaptation of multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7996–8005, Singapore. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2025. RomanLens: The role of latent Romanization in multilinguality in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26410–26429, Vienna, Austria. Association for Computational Linguistics.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Chenghao Sun, Zhen Huang, Yonggang Zhang, Xinmei Tian, Xu Shen, and Jieping Ye. 2026. Bridging the language gap: Uncovering and aligning shared circuits for multi-hop reasoning in multilingual llms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33091–33099.

GLM-4.5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze.

2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Evaluation Setup Details

ECLeKTic and MultiLoKo are evaluated using an auto-rater setup, using vLLM (Kwon et al., 2023) for inference. We prompt the questions with the system instruction "In just a few words, answer in the language of the question." and each of the model's individuals chat formats. GPT-OSS-120B is set to "Reasoning: medium". Under certain conditions, such as the model over-thinking before the max_model_len is exhausted, we reformat the thinking trace and prompt the model to just generate the answer. Then, the model output (and not its thinking trace) is provided to an auto-rater (implemented with Qwen3-30B-A3B) along with the golden answer. Our auto-rater system instruction requires a one-word answer ("YES" or "NO") and provides context about the question and answer. Our auto-rater system instructions are thoroughly tuned via qualitative analysis to ensure reliability.

Belebele (used in Section 4) is evaluated using the lm-evaluation-harness (Gao et al., 2023) in python using the vLLM backend.

B OLS Details

OLS vs. Logistic Regression Despite the predictor variable being a binary outcome (1.0 for correct answer, 0.0 for incorrect), we elect OLS for interpretability of the coefficients. In OLS, the coefficients represent direct probability changes rather than log-odds ratios. This approach is standard when predicted probabilities remain well within [0,1] (Angrist and Pischke, 2009), which they do for all rows in the table. We have additional ran a logistic regression model and calculated average marginal effects. The average marginal effects (AME) remains very similar for the script match indicator: 0.124 for ECLeKTic an 0.137 for MultiLoKo, in comparison to the OLS β of 0.121 and 0.133 reported. Importantly, the coefficients for language match and family match are also statistically insignificant (cannot reject null hypothesis that they are zero).

Implementation We implement our regression models in both statsmodels (Seabold and Perktold, 2010) and scikit-learn (Pedregosa et al., 2011), which provides us a sanity check that the fitted model is unambiguous and deterministic. The final predictor variable and all the indicator features are binary variables with values 1.0 and 0.0. The Belebele score is the only continuous variable, and most values are in the range 0.50 to 0.90.

Feature Collinearity While the features were certainly not independent and showed expected inter-correlations (e.g., between the cross-script and cross-language indicator), we verified the stability of our estimates using Variance Inflation Factors (VIF), ensuring that multicollinearity did not undermine the reliability of the observed coefficients. However, this leads to counter-intuitive coefficients such as Qwen3-235B-A22B having the lowest β other than Qwen3-30B-A3B despite it being the best model by quite a bit on both datasets. This is explained by the correlation between model capability and Belebele score, which likely captures most of the variance in model capabilities.

Interestingly, in ECLeKTic, the English-source subset seemed to be the easiest (the coefficient for all other languages is negative), while for MultiLoKo the English-source subset was the most difficult (coefficient for all other languages is positive).

C Think Prefixing Prompts

Note that in certain questions in ECLeKTic and MultiLoKo, there may be more than one entity. So we create a prompt for both scenarios:

For Qwen3 and Olmo-3.1-32B:

Table 4: OLS Coefficients (β) when trying to predict the binary score of a question, **after features are standard-scaled**. *n.s.* denotes statistically insignificant coefficients; $p > 0.05$. The values represent the expected increase in probability from an increase in one standard deviation of the feature. It allows for comparison between continuous features (e.g. Belebele score) and the rest. The value is essentially a metric for how much variance the feature encapsulates.

Variable	ECleKTic	MultiLoKo
OLS Intercept	0.452	0.569
Belebele Score	0.088	0.100
Script Match	0.056	0.064
Language Family	*n.s.*	*n.s.*
In-Language	*n.s.*	*n.s.*
Model Indicators (vs. Qwen3-30B)		
GPT-OSS-120B	0.046	0.026
Olmo-3.1-32B	0.022	0.034
GLM-4.5-Air	0.016	*n.s.*
Qwen3-235B	0.016	0.013
Src Lang Indicators† (vs. EN)		
	[−0.064, −0.023]	[0.007, 0.048]

† Range of β for source-language subset indicators used to account for subset difficulty.

"Okay, the key entity of the question in its source language is "

"Okay, the important entities in the question in their source language are "

For GLM-4.5-Air:

"The question is asking about the entity (in English) "

"The question is asking about the entities (in English) "

For GPT-OSS-120B, where {lang_tgt} is the English name for the language of the question:

"The user asks in {lang_tgt} about an entity, that in its source language would be "

"The user asks in {lang_tgt} a question in multiple entites, that would be, in their source language, "

D Additional SFT Data Generation Details

We provide here details for our SFT data generation pipeline discussed in Section 6. We first use the prompts described in Appendix E to create candidate question/thinking/answer trios with Gemini. While a quick check of some samples revealed that Gemini was creating valid answers for the questions. However, we do not particularly care if the answer is right, as our goal is not to teach the model facts. The goal is to teach it a reasoning style. We then prompt the target model with just the question, and again, do not care if its answer is coherent with Gemini's provided answer as all we care about here is the thinking *style*.

We then use the prompt detailed in Appendix F to get the target model to provide a style-adhering version of the thinking trace. We are forced to discard many samples during each step if there are errors in the formatting (i.e., requested JSON formatting). Because of the higher amount of samples dropped for Qwen3-30B-A3B during the style transfer process, we end up with 599 training samples for GPT-OSS-120B and only 387 training samples for Qwen3-30B-A3B.

E Sample Gemini Prompt

Below we provide a sample prompt used for generating question/thinking/answer trios. We check that Gemini is generating a thinking trace that does as we ask. To settle on the final prompt, it took numerous attempts to get the reasoning logic that we desire. Even after settling on this prompt, we do continuous sanity checks that the generation is coherent and logical. We evaluated a few dozen before scaling the generation to about 700 thinking traces. To induce more variability, we modify the wording of the following prompt, but the details and instructions remain the same.

> I have this cross-lingual knowledge database with questions and want to fine-tune the model to reason better. The questions are often about entities and they exist in a bunch of languages in parallel. However the (human) translation of the entities is sometimes a bit ambiguous because they are very rare/obscure entities and may not have a proper written format in other scripts. Given this, can you produce seven sample model thinking traces where the model reasons about the entity and what the original entity could possibly be (potentially from different languages/scripts). Then once the model thinks it has found the entity, it provides

its answers to the question in the original language ?

Please generate nothing but a JSON, where each item in the list is a dictionary containing question, thinking, and model_response. Please don't include formatted markdown text in the thinking trace, simply a few paragraphs.

F Style Transfer Prompt

The prompt provided to Qwen3-30B-A3B for "style transfer" from Gemini's thinking trace to its style. The prompt for GPT-OSS-120B was worded differently because it was not following these instructions well. {gemini_thinking_trace} and {style_reference} are placeholders for the question-specific thinking trace provided by Gemini and the target model, respectively.

I am performing style transfer of a thinking trace to create SFT data for a thinking model.

<instruction>

1. Read the logic in <logic_to_rewrite>. This is the factual chain of events you must keep.

2. Read the writing style in <style_reference>. Note the tone, verbosity, and language structure. They are sometimes long and so therefore cutoff at 500 characters.

3. REWRITE the content of <logic_to_rewrite> using the voice/persona found in <style_reference>.

4. CRITICAL:

- You must write in ENGLISH.

- Don't copy the text from <style_reference> directly; only copy its *style*.

- You must reach the same final answer as the <logic_to_rewrite>.

- Please respond in the same JSON format with question, thinking, and model_response keys.

- The model_response should be in the same language as the question, irrespective of the <style_reference>.

5. If you think the initial question and/or answer are flawed, please respond simply "BAD" (no JSON).

</instruction>

\n

<logic_to_rewrite>

{gemini_thinking_trace}

</logic_to_rewrite>

\n

<style_reference>

{style_reference}

</style_reference>

G Example SFT Training Sample

The example of a synthetic reasoning trace, written in the typical thinking style of Qwen3, which often starts with "Okay, the user is asking ...", identifies the language of the question, and uses "Hmmm".

Okay, the user is asking 'Melyik évben adták ki a 'Dzeelda' which is Hungarian for 'In which year was 'Dzeelda' released?'. Hmm, 'Dzeelda' doesn't sound quite right—I'm thinking this might be a phonetic misspelling. The iconic Nintendo franchise is actually 'The Legend of Zelda', and in Hungarian contexts, it's commonly shortened to 'Zelda' but sometimes gets mangled in pronunciation. Let me verify this.

First, I should confirm which specific game they're referencing since 'Zelda' alone could mean multiple titles. Given how the question is phrased, they're likely asking about the most culturally significant entry—Ocarina of Time, ...

Note that this sample is not necessarily a cross-script question, it's meant simply to nudge the LLM to reason more carefully.

H LoRA SFT Details

We detail here the configurations for the LoRA SFT with Llama-Factory. For all models, we use the adafactor optimizer and cosine LR scheduling with warmup. In the table below, "All Attn" signifies that LoRA modules were applied to the Q_{proj}, K_{proj}, V_{proj}, and O_{proj} weights in the attention block, but not to any FFN parameters. At the time of the experiments, Llama-Factory does not have the capability to apply LoRA modules to the Qwen3 MoE implementation of the FFN layers.

Table 5: Hyperparameters for the LoRA training runs for which results are reported in Table 3.

Hyperparam.	**GPT-OSS-120B**	**Qwen3-30B-A3B**
Num. Samples	599	387
Target Modules	All Attn	All Attn
LoRA Rank (r)	32	32
LoRA Dropout	0	0.05
Learning Rate	2×10^{-5}	6×10^{-5}
Epochs	8	8
Warmup Ratio	0.2	0.2