25 - Multilingual Long Chain-of-Thoughts with Small Reasoning Models

https://huggingface.co/mutilingual-long-cot

https://github.com/Multilingual-long-COT


Swati Rajwal
Captain
Presenter


Marek Suppa


Yadnyesh. C


Allison Yang


Ira Salsabila


Shayekh Islam
Captain


Drishti Sharma


Azmine Toushik


Morteza Kashani

Cohere Labs

15th May 2025

Motivation🎯

- Majority of STEM resources/benchmarks are available in English/high-resource languages.
- This limits access to quality scientific assessments for non-English speakers [1,2].
- LLMs can help democratize complex reasoning tasks like long-form CoT across languages.
- Our work targets critical gaps by translating and evaluating CoT reasoning.
- We focus on diverse set of non-english languages such as: Hindi, Indonesian, Spanish, Bengali, Marathi and others.

* Chain-of-Thought

Evaluation Data Construction 🚧
Multilingual-GPQA

- Seed Dataset GPQA*-Diamond
- Subjects: Biology, Physics, Chemistry
- 198 MCQ samples by domain experts[3]
- Manually labeled options: 1 = translate, 0 = don't
- MQM** prompting for translation with Claude 3.7 [4]
- Human post-edit (Ongoing)

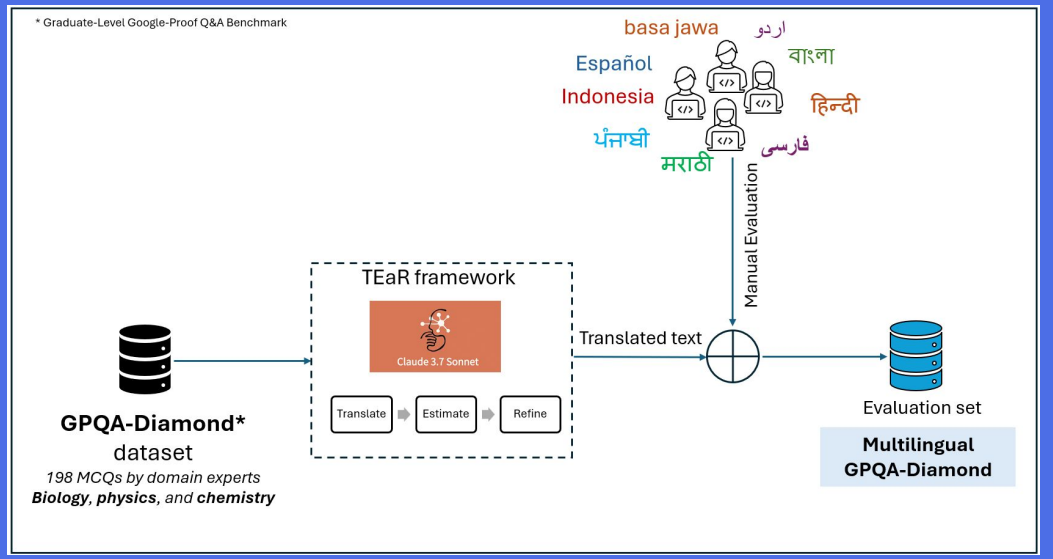


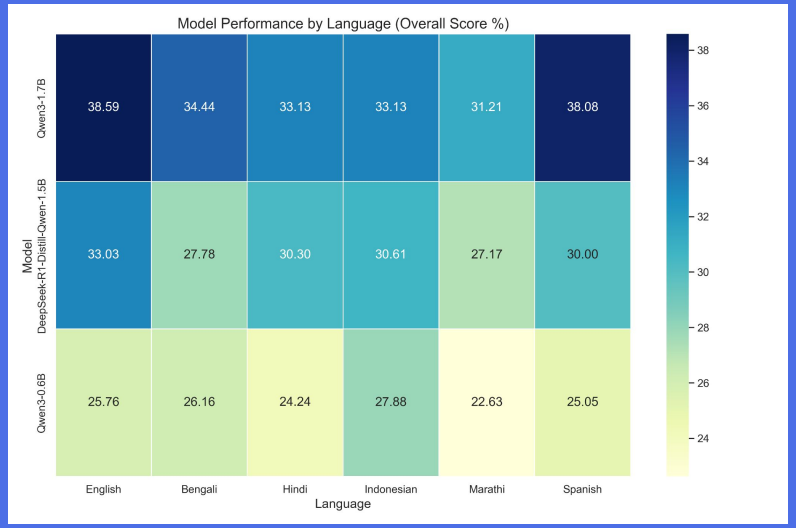
Figure 1: Flowchart for evaluation data curation.

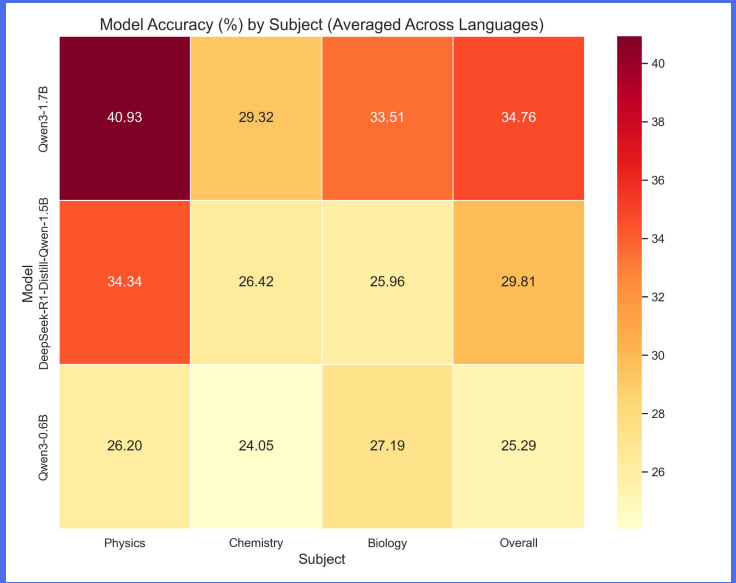
Cohere Labs

* Graduate-Level Google-Proof Q&A Benchmark
** Multi-dimensional Quality Metrics

Evaluating Small Reasoning Models on M-GPQA

- Qwen3-1.7B performs best among three.
- Poor performance in low-resource languages, like Marathi
- Chemistry is the hardest subject of the three.

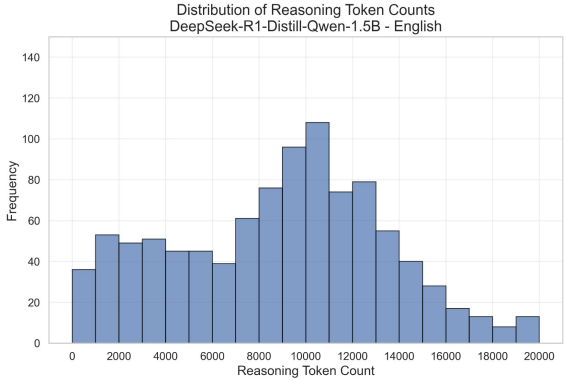


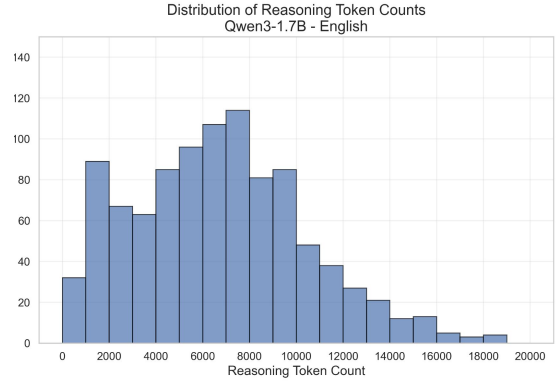


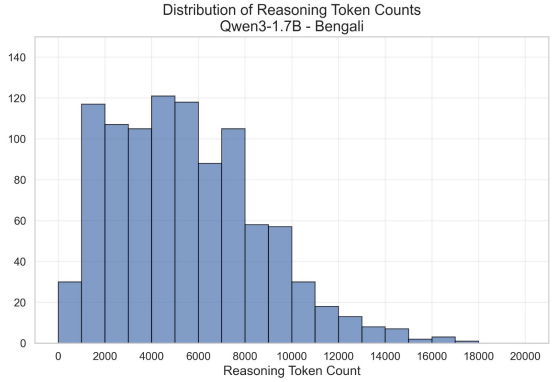
Cohere Labs

Evaluating Small Reasoning Models on M-GPQA

- Qwen3-1.7B can achieve better performance with less amount of reasoning tokens vs Deepseek-R1-Distill
- Reasoning Models spend less reasoning tokens for non-English questions







Cohere Labs

Training Small Reasoning Models for Multilingual

- Dataset
  - Seed dataset: Camel*-AI (Training set)
  - Subjects: Biology, Physics, Chemistry
  - 1.2k Deepseek-R1 traces per subject [5]
  - Simple Translation Prompting with Claude 3.7
  - Only translating final CoT/answer (keeping <think> part in English)
  - Initially working on Hindi, Indonesian, Bengali, Spanish and Marathi.
- Working on fine-tuning Qwen3-1.7B

Cohere Labs

* Communicative Agents for “Mind” Exploration of Large Scale Language Model Society

Key Takeaways 📍

- Small Reasoning Models performs poorly on lower resource languages.
- Moreover, they “Think” less for multilingual questions.
- By enabling multilingual models to perform on par with English models in complex STEM domains, we aim to empower global education and inclusion in scientific advancement

References

1. Northwestern University McCormick School of Engineering. (2021, September). Lack of non-English languages in STEM publications hurts diversity. https://www.mccormick.northwestern.edu/news/articles/2021/09/lack-of-non-english-languages-in-stem-publications-hurts-diversity/

2. Le Pichon, E., Ye, R., & Kang, S. H. (2024). Enhancing equitable access to education for English language learners: evaluating the impact of a digital multilingual STEM resource in Canada. International Journal of Multilingualism, 1-19.

3. Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., ... & Bowman, S. R. (2024). Gpqa: A graduate-level google-proof q&a benchmark. In First Conference on Language Modeling.

4. Feng, Z., Zhang, Y., Li, H., Wu, B., Liao, J., Liu, W., ... & Liu, Z. (2024). Tear: Improving llm-based machine translation with systematic self-refinement. arXiv preprint arXiv:2402.16379.

5. Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). Camel: Communicative agents for" mind" exploration of large scale language model society.

6. Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., ... & Wei, J. (2022). Language models are multilingual chain-of-thought reasoners. arXiv preprint arXiv:2210.03057.