BIICK-Bench: A Bengali Benchmark for Introductory Islamic Creed Knowledge in Large Language Models

Umar Hasan, Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

GitHub Repository

1. Introduction

The Problem: LLMs increasingly serve as information sources, yet their proficiency in specialized, non-English domains remains untested.

The Contribution: BIICK-Bench, the first Bengali benchmark for Introductory Islamic Creed Knowledge.

- **50** Multiple Choice Questions (MCQA).
- **14** Open-source models evaluated (1.5B to 8B).
- **Key Finding:** 13 of 14 models failed, performing near random chance.

2. BIICK-Bench: Design & Curation

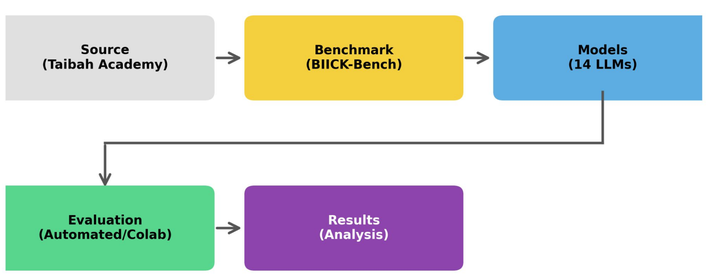
Design Principles

- **Native Bengali:** Created directly in Bengali for native speakers; no translation artifacts.
- **Theologically Consistent:** Derived from a single, mainstream Islamic curriculum.
- **Automated Evaluation:** Standardized 4-option MCQA format for objective testing.

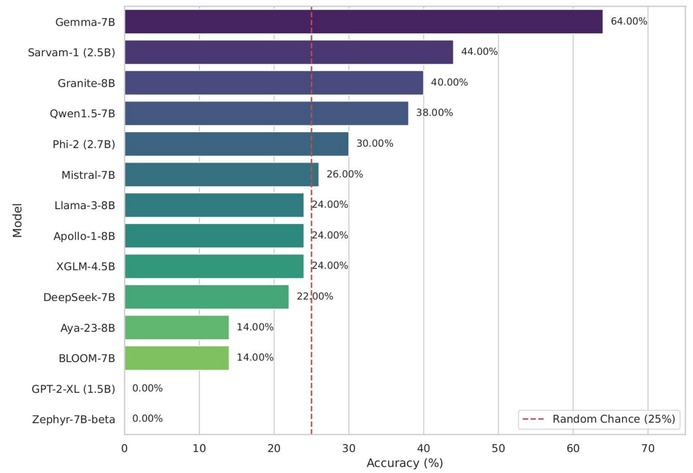
Data Source

- **Curriculum:** Expert-validated final exam from **Taibah Academy** (“Introduction to Islamic Creed”).
- **Quality:** Expert-verified and relevant to the target community.

3. Experimental Setup



4. Results & Analysis



Analysis

Performance is poor across most models. **Gemma-7B (64%)** is a clear outlier. Leading models like **Llama-3-8B (24%)** and **Mistral-7B (26%)** perform at or near the random-chance baseline. Massively multilingual models (Aya-23, BLOOM) also failed.

5. Discussion & Conclusion

The low performance demonstrates that LLMs **cannot be considered reliable sources** for general Islamic knowledge in Bengali. This validates the need for community-specific benchmarks to hold developers accountable and inform users of the risks of uncritical AI adoption.

Limitations & Future Work

Limitations include a small (50-question) benchmark and evaluation on free-tier compute. Future work will focus on expanding the benchmark, translating it to other under-resourced languages, and conducting qualitative analysis of failure modes.

A Critical Delineation

Not for Fatwas: This research evaluates LLMs as a repository of *general knowledge*, **not** as a source for issuing religious verdicts (*fatwas*). We firmly maintain that seeking *fatwas* from AI is impermissible; this role must remain with qualified human scholars.

References

Abdelrahman Abdallah et al. (2024). “ArabicaQA: A comprehensive dataset for arabic question answering.” In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24. Association for Computing Machinery, pp. 2049-20591.

Abubakar Abid, Maheen Farooqi, and James Zou (2021). “Persistent anti-muslim bias in large language models.” In: CoRR, abs/2101.057832.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka (2023). “HAQA and QUQA: Constructing two Arabic question-answering corpora for the Quran and Hadith.” In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. Ed. by Ruslan Mitkov and Galia Angelova. INCOMA Ltd., pp. 90-973.

Dan Hendrycks et al. (2021). “Measuring massive multitask language understanding.” In: International Conference on Learning Representations.

Stephanie Lin, Jacob Hilton, and Owain Evans (2021). “TruthfulQA: Measuring how models mimic human falsehoods.” In: CoRR, abs/2109.079585.

Thomas Wolf et al. (2020). “Transformers: State-of-the-art natural language processing.” In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Ed. by Qun Liu and David Schlangen. Association for Computational Linguistics, pp. 38-456.

Rowan Zellers et al. (2019). “Hellaswag: Can a machine really finish your sentence?” In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, pp. 4791-48007.

My LinkedIn Profile