

Multilingual AI Safety Benchmark: Cross-Lingual Reliability Analysis

Executive Summary

This project evaluates whether a frontier-scale Large Language Model exhibits different reliability, hallucination behavior, and refusal patterns when answering the same questions across four languages: English, Hindi, Tamil, and Vietnamese. The benchmark was designed to investigate multilingual safety and performance degradation in lower-resource languages.

Project Objective

Measure cross-lingual robustness by comparing model behavior on identical tasks translated into multiple languages. The benchmark focuses on: Accuracy Hallucination Rate Refusal Rate Cross-language consistency

System Architecture

1. FastAPI inference server using NVIDIA Inference API.
2. Model evaluated: Meta Llama 3.3 70B Instruct.
3. Dataset contains multilingual benchmark questions spanning science, mathematics, health, and government domains.
4. Each question is evaluated in English, Hindi, Tamil, and Vietnamese.
5. Responses are collected, logged, and analyzed automatically.
6. Refusal detection is performed using rule-based pattern matching.

Dataset Characteristics

Questions evaluated: 50

Languages: 4

Total model evaluations: 200

Categories: Science, Mathematics, Health, Government

Difficulty Levels: Easy, Medium, Hard

Benchmark Results

English: Accuracy 83.4%, Hallucination 35.0%, Refusal 0.0%

Hindi: Accuracy 81.0%, Hallucination 32.5%, Refusal 0.0%

Tamil: Accuracy 82.4%, Hallucination 29.5%, Refusal 0.0%

Vietnamese: Accuracy 85.4%, Hallucination 33.5%, Refusal 0.0%

Overall Accuracy: 83.05%

Overall Hallucination Rate: 32.62%

Overall Refusal Rate: 0.0%

Key Findings

- No significant performance collapse was observed in lower-resource languages.
- Vietnamese achieved the highest measured accuracy despite being a lower-resource language.
- Hallucinations remained present across all languages, indicating that factual reliability issues are not limited to any single language.
- Refusal behavior was effectively zero across the benchmark, suggesting strong willingness to answer user queries.
- Accuracy variance across languages remained relatively small, demonstrating reasonable cross-lingual robustness.

Engineering Assessment

From an engineering perspective, the benchmark successfully establishes a reproducible pipeline for multilingual AI safety evaluation. The architecture separates inference, orchestration, and analysis components, enabling future scaling to additional languages, models, and safety metrics. The project demonstrates practical benchmarking methodology under limited time and resource constraints.

Future Improvements

- Increase dataset size beyond 50 questions.
- Add Bengali, Tagalog, Bahasa Indonesia, and Swahili.
- Replace rule-based hallucination detection with LLM-as-a-Judge evaluation.
- Introduce adversarial safety prompts and sycophancy measurements.
- Compare multiple open-source and frontier models.

Conclusion

This benchmark provides evidence that multilingual performance differences exist but are smaller than commonly assumed for the tested model. While accuracy remained above 80% across all languages, hallucinations remained a persistent challenge. The project establishes a foundation for future research into multilingual AI safety, reliability, and evaluation methodologies.