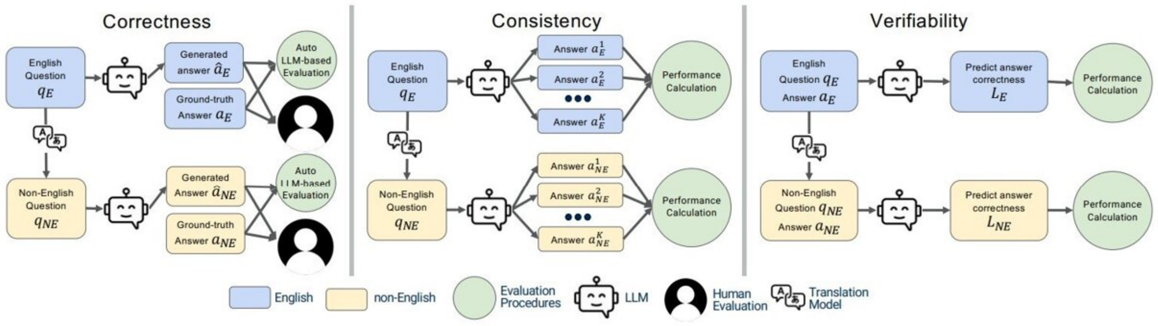
Chatbots are poor multilingual health care consultants, study finds

May 28 2024


Correctness
English Question q_E
Generated answer $\hat{a}_E$
Ground-truth Answer a_E
Auto LLM-based Evaluation
Non-English Question q_{NE}
Generated Answer $\hat{a}_{NE}$
Ground-truth Answer a_{NE}
Auto LLM-based Evaluation
Consistency
English Question q_E
Answer a_E^1
Answer a_E^2
Answer a_E^K
Performance Calculation
Non-English Question q_{NE}
Answer a_{NE}^1
Answer a_{NE}^2
Answer a_{NE}^K
Performance Calculation
Verifiability
English Question q_E Answer a_E
Predict answer correctness L_E
Performance Calculation
Non-English Question q_{NE} Answer a_{NE}
Predict answer correctness L_{NE}
Performance Calculation
English
non-English
Evaluation Procedures
LLM
Human Evaluation
Translation Model

Evaluation pipelines for correctness, consistency, and verifiability criteria in the XLingEval framework. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2310.13132

Georgia Tech researchers say non-English speakers shouldn't rely on chatbots like ChatGPT to provide valuable health care advice.

A team of researchers from the College of Computing at Georgia Tech has developed a framework for assessing the capabilities of large language models (LLMs). Ph.D. students Mohit Chandra and Yiqiao (Ahren) Jin are the co-lead authors of the paper "Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Health care Queries." The paper is published on the *arXiv* preprint server.

Their paper's findings reveal a gap between LLMs and their ability to answer health-related questions. Chandra and Jin point out the limitations of LLMs for users and developers but also highlight their potential.

Their XLingEval framework cautions non-English speakers from using chatbots as alternatives to doctors for advice. However, models can improve by deepening the data pool with multilingual source material such as their proposed XLingHealth benchmark.

"For users, our research supports what ChatGPT's website already states: chatbots make a lot of mistakes, so we should not rely on them for critical decision-making or for information that requires high accuracy," Jin said.

"Since we observed this language disparity in their performance, LLM developers should focus on improving accuracy, correctness, consistency, and reliability in other languages," Jin said.

Using XLingEval, the researchers found chatbots are less accurate in Spanish, Chinese, and Hindi compared to English. By focusing on correctness, consistency, and verifiability, they discovered:

- Correctness decreased by 18% when the same questions were asked in Spanish, Chinese, and Hindi.
- Answers in non-English were 29% less consistent than their English counterparts.
- Non-English responses were 13% overall less verifiable.

XLingHealth contains question-answer pairs that chatbots can reference, which the group hopes will spark improvement within LLMs.

The HealthQA dataset uses specialized health care articles from the popular health care website Patient. It includes 1,134 health-related question-answer pairs as excerpts from original articles. LiveQA is a second dataset containing 246 question-answer pairs constructed from frequently asked questions (FAQs) platforms associated with the U.S. National Institutes of Health (NIH).

For drug-related questions, the group built a MedicationQA component. This dataset contains 690 questions extracted from anonymous consumer queries submitted to MedlinePlus. The answers are sourced from medical references, such as MedlinePlus and DailyMed.

In their tests, the researchers asked over 2,000 medical-related questions to ChatGPT-3.5 and MedAlpaca. MedAlpaca is a health care question-answer chatbot trained in medical literature. Yet, more than 67% of its responses to non-English questions were irrelevant or contradictory.

"We see far worse performance in the case of MedAlpaca than ChatGPT," Chandra said. "The majority of the data for MedAlpaca is in English, so it struggled to answer queries in non-English languages. GPT also struggled, but it performed much better than MedAlpaca because it had some sort of training data in other languages."

Ph.D. student Gaurav Verma and postdoctoral researcher Yibo Hu co-authored the paper.

Jin and Verma study under Srijan Kumar, an assistant professor in the School of Computational Science and Engineering, and Hu is a postdoc in Kumar's lab. Chandra is advised by Munmun De Choudhury, an associate professor in the School of Interactive Computing.

The team presented their paper at The Web Conference, occurring May 13-17 in Singapore. The annual conference focuses on the future direction of the internet. The group's presentation is a complementary match, considering the conference's location.

English and Chinese are the most common languages in Singapore. The group tested Spanish, Chinese, and Hindi because they are the world's most spoken languages after English. Personal curiosity and background played a part in inspiring the study.

"ChatGPT was very popular when it launched in 2022, especially for us computer science students who are always exploring new technology," said Jin. "Non-native English speakers, like Mohit and I, noticed early on that chatbots underperformed in our native languages."

More information: Yiqiao Jin et al, Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries, *arXiv* (2023). DOI: 10.48550/arxiv.2310.13132

Provided by Georgia Institute of Technology

Citation: Chatbots are poor multilingual health care consultants, study finds (2024, May 28) retrieved 20 June 2024 from https://medicalxpress.com/news/2024-05-chatbots-poor-multilingual-health.html