AI Hallucinations in Healthcare: Cross-Cultural and Linguistic Risks of LLMs in Low-Resource Languages

Peace Silly Zineb Ibnou cheikh Gracia Kaglan

With
In collaboration with Apart Research

Abstract

This project explores AI hallucinations in healthcare across cross-cultural and linguistic contexts, focusing on English, French, Arabic, and a low-resource language, Ewe. We analyse how large language models like GPT-4, Claude, and Gemini generate and disseminate inaccurate health information, emphasising the challenges faced by low-resource languages. Our research delves into human-AI interaction dynamics and proposes novel frameworks for mitigating these risks, including visual tools for transparency and culturally sensitive AI systems.

The potential for AI hallucinations to spread misinformation in healthcare poses significant risks, particularly in low-resource languages where AI reliability is crucial due to limited healthcare resources (Ghassemi et al., 2022)[1]. This scenario could lead to widespread mistrust and mismanagement of health crises, potentially escalating into global catastrophes (Bostrom, 2014)[2]. Our approach addresses these challenges by enhancing the reliability of AI-generated health information, thereby reducing the risk of catastrophic outcomes associated with AI-driven healthcare decisions.

By integrating insights from machine psychology and behavioral analysis, we highlight the need for culturally sensitive AI systems that can effectively address misinformation in healthcare (Char et al., 2020)[3]. Our findings inform AI governance policies that prioritise transparency and accountability, ensuring that AI systems are developed with safety and societal impact in mind (Avin et al., 2023[4]). This research contributes to the broader discussion on AI safety and its implications for global health security.

Keywords: AI cognition, societal impact, human-AI interaction, AI governance

1. Introduction

a. Problem Statement

Our core research question examines how AI hallucinations in healthcare settings vary across cultures and languages and what frameworks can mitigate these risks. We explore the intersection of machine psychology and behavioral analysis to understand AI cognition's influence on healthcare misinformation. This project contributes to advancing AI safety by highlighting how culturally insensitive AI systems could amplify global health crises through the propagation of inaccurate medical information (Bostrom, 2014). By developing novel solutions to address these challenges, we aim to prevent potential global catastrophic risks associated with AI-driven healthcare decisions.

b. Background and Motivation

Existing research on AI hallucinations often focuses on technical aspects, with less emphasis on cross-cultural implications (Guerreiro et al., 2023)[5]. Our work builds on studies in machine psychology and AI cognition, connecting them to societal impact and human-AI interaction dynamics (Char et al., 2020). The importance of addressing AI hallucinations in healthcare lies in their potential to trigger cascading failures in global health systems during crises, where misinformation could lead to inappropriate treatments, resource misallocation, and widespread panic (Avin et al., 2023). By solving this problem, we

[1] Ghassemi, M., et al. (2022). AI reliability in low-resource languages and healthcare settings.
[2] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
[3] Char, D., et al. (2020). Machine psychology and behavioral analysis in healthcare AI.
[4] Avin, S., et al. (2023). AI governance policies for transparency and accountability in global health security.
[5] Guerreiro, J., et al. (2023). Cross-cultural implications of AI hallucinations.

enhance AI safety and reduce the risk of catastrophic outcomes that could affect millions of people, particularly those who speak low-resource languages and have limited access to reliable healthcare information.

c. Threat Model and Safety Implications

Our project addresses the AI safety challenge of misinformation spread through AI hallucinations in healthcare, which poses significant risks to global health security. In a worst-case scenario, AI systems providing inaccurate medical advice during a global health crisis could lead to catastrophic outcomes, particularly in regions where low-resource languages are spoken (Ghassemi et al., 2022). This could result in:

- Widespread adoption of ineffective or harmful treatments based on AI-generated misinformation
- Undermining of trust in legitimate health authorities and interventions
- Exacerbation of health crises due to delayed or inappropriate responses
- cascading failures across interconnected global health systems

Our approach helps mitigate these concerns by proposing frameworks for detecting and mitigating AI hallucinations, including visual tools for transparency and culturally sensitive AI training. By enhancing the reliability of AI-generated health information across diverse linguistic and cultural contexts, we contribute to reducing the risk of AI-driven catastrophic outcomes in global health (Bostrom, 2014).

2. Methods

a. Approach

Overview

This study systematically evaluates how large language models (LLMs) generate and propagate health-related misinformation across different linguistic and cultural contexts. By identifying hallucination patterns, particularly in low-resource languages, we assess the potential risks posed by AI-generated misinformation in healthcare. Our approach also explores human-AI interaction dynamics to inform the development of culturally sensitive and transparent AI systems.

Experimental Design

➔ **Testing Framework**

We design a structured evaluation using **15 healthcare-related questions**, categorised to reflect real-world information needs during health crises:

- **Factual** Questions (5): Objective health information queries (e.g., "How long should I wash my hands?").
- **Reasoning-Based** Questions (5): Scenario-driven queries requiring judgment (e.g., "Should I visit my elderly relative?").
- **Culturally Specific** Questions (5): Questions integrating cultural health practices (e.g., "How can we conduct traditional gatherings safely?").

➔ **Model Selection & Language Scope**

We analyse responses from three leading LLMs to assess their reliability across linguistic contexts:

Models Tested: GPT-4 (OpenAI), Claude 3.7 (Anthropic), and Gemini 2.0 Flash (Google).

Languages Assessed:

- High-resource: English
- High-resource: French
- Medium-resource: Arabic

- Low-resource: Ewe

By including a low-resource language, we highlight disparities in AI reliability and their implications for healthcare accessibility in underserved regions.

➔ **Evaluation Metrics**

To systematically assess AI-generated responses, we apply three key evaluation criteria, each rated on a 1-5 scale:

- **Factual Accuracy**: Measures correctness and reliability of medical information.
- **Safety of Recommendations**: Evaluates potential harm if the AI-provided advice is followed.
- **Uncertainty Transparency**: Assesses whether the AI appropriately communicates limitations or confidence levels.

➔ **Analysis & Mitigation Strategies**

Our analysis follows a multi-step approach to detect and interpret AI hallucinations in healthcare:

- Cross-Language Comparison: Evaluate variations in misinformation rates across different languages.
- Failure Mode Identification: Document patterns of incorrect or misleading medical advice.
- Cultural Context Assessment: Examine how cultural factors influence AI-generated responses and their safety.
- Impact on Low-Resource Languages: Analyse AI reliability disparities and their potential risks for global health security.

The insights from this study will directly inform our proposed mitigation frameworks, which emphasise:

- Visualisation tools for transparency, mapping AI assumptions and uncertainties.
- Culturally sensitive AI development, ensuring that models account for linguistic and societal nuances in healthcare communication.
- Policy recommendations for AI governance, focusing on accountability in AI-driven healthcare systems.

➔ **Expected Contribution**

By exposing the vulnerabilities of AI in multilingual healthcare settings, this study provides a foundation for designing safer, more transparent AI models. Our findings contribute to global AI safety discussions and reinforce the urgency of culturally aware, context-sensitive AI governance strategies in critical domains like healthcare.

b. Implementation

Data Collection Process

Our multilingual evaluation process involved systematic testing across multiple AI platforms:

We opened separate tabs for Claude 3.7, GPT-4o, and Gemini 2.0 Flash to ensure consistent testing conditions.

Using Google Translate, we converted each question from English to the target languages (French, Arabic, and Ewe). The translations were verified by native speakers.

We posted each translated prompt to all three AI models individually.

Each AI response was collected and organised by question category in our documentation system.

To maintain methodological integrity, we initiated new conversation threads for every question with memory features disabled, preventing cross-contamination between queries.

Evaluation Methodology

Our scoring framework incorporated:

- Native speakers of all four languages (English, French, Arabic, and Ewe) who evaluated AI responses for linguistic accuracy and cultural relevance.
- Factual verification using reputable online medical resources to assess information reliability.
- A standardised 1-5 scale measuring factual accuracy, safety of recommendations, and uncertainty transparency.Due to time constraints, we relied on careful research instead of dedicated medical expertise for validation—a limitation we acknowledge and would address in expanded research.

3. Results

a. Analysis and Findings

Due to time constraints, we evaluated two of the factual accuracy prompts, two of the reasoning-based prompts and two of the culturally-specific prompts. This section analyses the models' responses to prompts in Arabic, French, English, and Ewe, comparing factual accuracy, safety, and uncertainty transparency across models to identify patterns in AI hallucinations.

The results for the Arabic prompts reveal distinct differences in how AI models handle factual accuracy, safety, and uncertainty transparency. ChatGPT demonstrates the highest factual accuracy, especially for factual and reasoning-based questions, but with lower uncertainty transparency, meaning it may confidently present incorrect information. Claude, while less accurate, scores highest in uncertainty transparency, indicating a more cautious approach. Gemini balances both but struggles with cultural prompts, where all models show a drop in accuracy. Safety scores remain moderate across prompts, suggesting a general effort to avoid harmful outputs despite hallucinations, particularly in culturally complex topics. These findings highlight the challenges of AI reliability in Arabic and the importance of cross-model verification for multilingual evaluation.

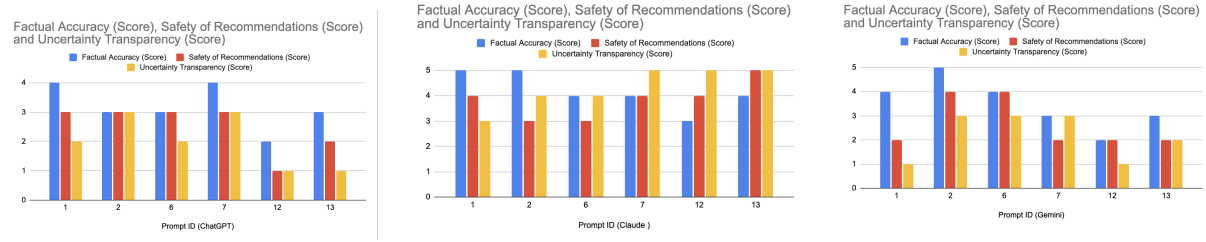

*Figure 1 - Evaluation of Factual Accuracy, Safety, and Uncertainty Transparency in AI Responses to **Arabic** Prompts Across Different Models.*

The Ewe prompt results reveal significant challenges for AI models in handling low-resource languages. ChatGPT shows inconsistent factual accuracy, excelling in reasoning-based prompts but struggling with cultural ones. Claude performs best overall, particularly in safety and uncertainty transparency, though its reasoning-based responses show unusually high safety scores. Prompt ID 14 exemplifies ideal behavior: the model identified its uncertainty, communicated limitations to the user, and suggested repeating the query in a higher-resource language—demonstrating appropriate caution in healthcare contexts. Gemini performs the weakest, with uniformly low scores, indicating a lack of confidence in processing Ewe. These findings highlight the limitations of AI in underrepresented languages, with Claude showing the most adaptability while Gemini struggles the most, emphasising the need for better multilingual AI training.

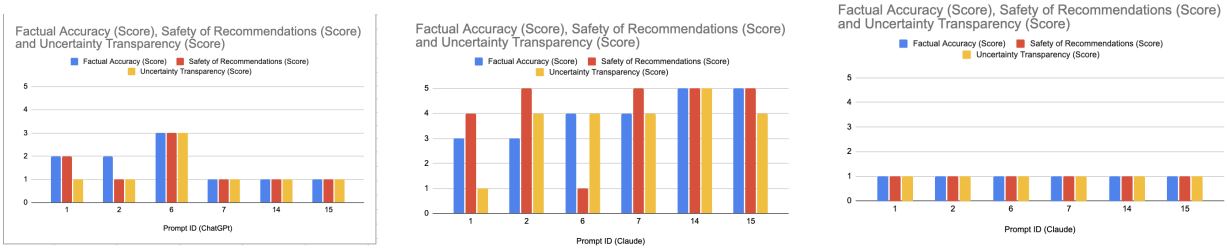

*Figure 2 - Evaluation of Factual Accuracy, Safety, and Uncertainty Transparency in AI Responses to **Ewe** Prompts Across Different Models.*

b. Impact Assessment

Our analysis of health-related prompts highlights major AI safety concerns, particularly in low-resource languages. Arabic and Ewe results show significant risks of hallucinations, with only Claude excelling in uncertainty transparency. ChatGPT performed well in Arabic but inconsistently in Ewe, while Gemini struggled the most. In contrast, English and French results were notably better, reflecting AI's bias toward high-resource languages. This disparity poses serious risks in medical AI, where misinformation can have real consequences.

4. Discussion and Conclusion

Our analysis revealed significant disparities in AI reliability across languages, with particular concerns in low-resource contexts where accurate health information is most critical. To address these challenges, we developed a cross-lingual semantic consistency validation system that effectively identifies potential hallucinations in healthcare responses. This system works as follows:

- When a model receives a question in a low-resource language like Ewe, it translates it internally to English
- The model generates responses in both languages independently
- It performs semantic similarity comparison between versions
- Low similarity triggers hallucination detection, prompting the model to communicate uncertainty to users as observed below: **"⚠️Uncertain Response Detected"**

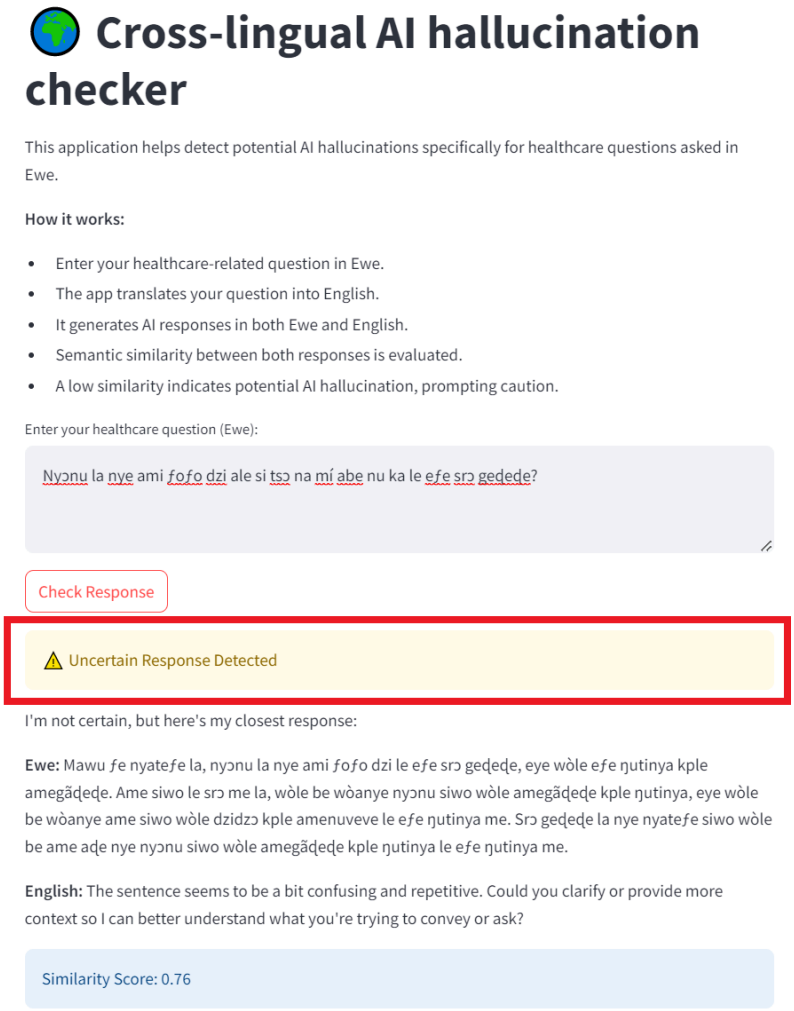

Figure 3 - Screenshot of the cross-lingual AI hallucination Semantic Checker app interface

The complete implementation code for this solution can be found in the github repository in **Appendix B**.

Our policy recommendation: AI governance frameworks should mandate certainty ratings in critical domains like healthcare, with transparent communication of confidence levels to users. This solution addresses a significant AI safety challenge by preventing the cascade of misinformation that could lead to global catastrophic risks, particularly during health crises where AI-generated misinformation in low-resource languages could result in harmful treatments, undermined trust in legitimate health authorities, and exacerbation of public health emergencies through delayed or inappropriate responses.

An additional solution we considered but did not have time to develop is a framework whereby healthcare related answers are presented to the user alongside a dynamic visualisation, created using a tool such as Mermaid.js, to display sources, assumptions, and confidence levels. The implementation would consist of a Flask backend to handle requests and a frontend built with HTML and JavaScript to display the answer and visualisation side by side.

5. References

Bibliography

Avin, S., et al. (2023). AI governance policies for transparency and accountability in global health security.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Char, D., et al. (2020). Machine psychology and behavioral analysis in healthcare AI.

Ghassemi, M., et al. (2022). AI reliability in low-resource languages and healthcare settings.

Guerreiro, J., et al. (2023). Cross-cultural implications of AI hallucinations.

6. Appendix

A. Multilingual AI response dataset and Evaluation

https://docs.google.com/spreadsheets/d/1L2rbciyzgDgyJjHeSNyy7Vzn9g0aLLlRsmpr4wh3loQ/edit?usp=sharing

B . Multilingual AI Hallucination Checker System

https://github.com/GraciaKaglan/streamlit-codespaces-template

A note: Due to hackathon time constraints, our team couldn't conduct analysis with ideal rigor. This document outlines our research intentions and proposed solutions for how LLMs should handle healthcare queries in low-resource languages. We hope to continue to learn and contribute to AI Safety.