



## The Humanity Analysis of ChatGPT in Swahili Language Processing

---

Hassan Juma, Ibrahim Mohammed and Sarah Zhao

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

August 20, 2024

# The Humanity Analysis of ChatGPT in Swahili Language Processing

Hassan Juma 1, Ibrahim Mohammed 2, Sarah Zhao 3

1. University of Dodoma (UDOM)
2. Muhimbili University of Health and Allied Sciences (MUHAS)
3. Swiss Hotel Management School (SHMS)

## abstract

This paper explores the impact of ChatGPT on Swahili language processing, encompassing its capabilities in understanding, generating, and translating Swahili text. It examines the social implications of using ChatGPT in Swahili-speaking communities, evaluates how the technology enhances efficiency, and addresses challenges related to bias and ethical considerations. The study reveals that while ChatGPT significantly improves communication and access to information in Swahili, ongoing efforts are necessary to mitigate biases and ensure equitable representation. The findings highlight the importance of integrating human-centered design principles to maximize the benefits of AI technologies in diverse linguistic contexts.

## Introduction

### Background

Swahili, a Bantu language spoken by millions across East Africa, serves as a lingua franca in several countries, including Kenya, Tanzania, and Uganda. As artificial intelligence (AI) technologies advance, their application in natural language processing (NLP) has broadened, with models like ChatGPT offering new possibilities for enhancing language-related tasks. ChatGPT, developed by OpenAI, has demonstrated considerable proficiency in various languages, yet its impact on Swahili requires thorough examination to understand its effectiveness and implications.

### Objectives

This paper aims to:

1. Analyze ChatGPT's capabilities in understanding and generating Swahili language content.
2. Evaluate the model's performance in translating between Swahili and other languages.
3. Investigate the social impact of ChatGPT on Swahili-speaking communities.
4. Explore the potential biases in the model's outputs and suggest measures for improvement.

### Significance

Understanding ChatGPT's interaction with Swahili is crucial for optimizing AI tools to better serve diverse linguistic populations. The findings will contribute to the broader discourse on AI ethics, inclusivity, and the role of technology in linguistic and cultural preservation.

## **Discussion**

### **Swahili Language Understanding and Generation**

ChatGPT exhibits a notable ability to understand and generate Swahili text. The model can accurately interpret context and produce coherent sentences, demonstrating its proficiency in handling the language's syntactic and semantic features. However, challenges remain, particularly with idiomatic expressions and culturally specific references, which may require additional contextual training to enhance accuracy.

### **Translation Capabilities**

In translation tasks, ChatGPT performs well in converting Swahili to other major languages and vice versa. The quality of translations generally meets acceptable standards, although nuances and regional dialects occasionally pose difficulties. Comparative studies with other translation models suggest that while ChatGPT is competitive, there is room for improvement in capturing the subtleties of Swahili.

### **Social Impact and Efficiency**

The integration of ChatGPT in Swahili-speaking communities has led to increased efficiency in communication and access to information. Educational tools powered by ChatGPT have facilitated learning and information dissemination, while customer service applications have improved user experiences. The technology has empowered individuals by bridging language barriers and providing more inclusive digital interactions.

### **Bias and Ethical Considerations**

Despite its advancements, ChatGPT is not immune to biases inherent in its training data. These biases can manifest in outputs that perpetuate stereotypes or exclude marginalized groups. To address these issues, it is crucial to implement regular audits and updates to the training data, ensuring a more balanced representation of diverse linguistic and cultural contexts.

### **Mitigation Strategies**

Strategies to mitigate biases include diversifying training datasets, incorporating feedback from native Swahili speakers, and employing human oversight in critical applications. Continuous monitoring and refinement of the model are essential to enhance fairness and accuracy.

## **Conclusion**

ChatGPT has made significant strides in processing Swahili, offering valuable tools for understanding, generating, and translating Swahili text. Its impact on efficiency and communication in Swahili-speaking communities is largely positive, contributing to greater accessibility and information dissemination. However, challenges related to bias and cultural sensitivity remain. Addressing these issues requires a concerted effort to refine the model and ensure equitable representation. Future research should focus on enhancing the

model's ability to handle linguistic nuances and incorporating feedback from diverse user groups to improve overall performance and fairness.

## Reference

- [1] Polomé E C. Swahili language handbook[J]. 1967.
- [2] Whiteley W. Swahili: the rise of a national language[M]. Taylor & Francis, 2023.
- [3] Mazrui A M, Mazrui A. Swahili beyond the boundaries: Literature, language, and identity[M]. Ohio University Press, 2007.
- [4] Zhong Q, Ding L, Liu J, et al. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert[J]. arXiv preprint arXiv:2302.10198, 2023.
- [5] Steere E. A handbook of the Swahili language: as spoken at Zanzibar[M]. Society for promoting Christian knowledge, 1906.
- [6] Peng K, Ding L, Zhong Q, et al. Towards making the most of chatgpt for machine translation[J]. arXiv preprint arXiv:2303.13780, 2023.
- [7] Shikali C S, Mokhosi R. Enhancing African low-resource languages: Swahili data for language modelling[J]. Data in brief, 2020, 31: 105951.
- [8] Ding L, Wu D, Tao D. The USYD-JD Speech Translation System for IWSLT 2021[J]. arXiv preprint arXiv:2107.11572, 2021.
- [9] Ding L, Wang L, Tao D. Self-attention with cross-lingual position representation[J]. arXiv preprint arXiv:2004.13310, 2020.
- [10] Zan C, Ding L, Shen L, et al. Building Accurate Translation-Tailored LLMs with Language Aware Instruction Tuning[J]. arXiv preprint arXiv:2403.14399, 2024.
- [11] Wanjawa B, Wanzare L, Indede F, et al. Kencorpus: A kenyan language corpus of swahili, dholuo and luhya for natural language processing tasks[J]. arXiv preprint arXiv:2208.12081, 2022.
- [12] Hsu C, Zan C, Ding L, et al. Prompt-learning for cross-lingual relation extraction[C]//2023 International Joint Conference on Neural Networks (IJCNN). IEEE, 2023: 1-9.
- [13] Zhou L, Ding L, Duh K, et al. Self-guided curriculum learning for neural machine translation[J]. arXiv preprint arXiv:2105.04475, 2021.