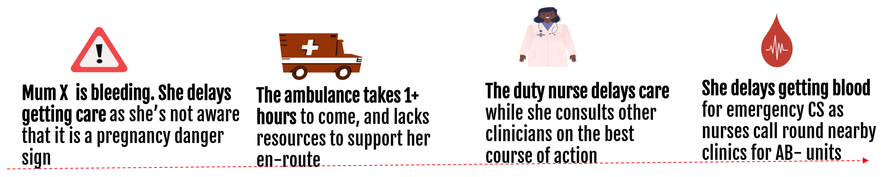
UlizaMama

In-Domain Adaptation of Large Language Models for Maternal and New-Born Health Questions-Answering in Low Resource and Code-Mixed Settings

Stanslaus Mwongela, Jay Patel, Sathy Rajasekharan, Anneka Wickramanayake, Rachael Alldian, Laura Wotton, Lyvia Lusiji,
Gilles Hacheme, Francesco Picinno, Mfoniso Ukwak, Ellen Sebestian, Julius Butime, Bernard Shibwabo

Motivation

- **33%** of maternal deaths are caused by delays in care seeking.
- Maternal mortality ratio in Kenya stands at 342 per 100,000 live births, significantly higher than in developed nations.



Problem Setting

Open source LLMs?

Table 1: **In context Learning (5-Shot)** Evaluationwith Jacaranda MNH Swahili

Model	BertScore -F1	BLEURT	METEOR	BLEU
Gemma 7b-it	0.61	0.05	0.11	0.01
Mistral-7b -Instruct	0.57	0.11	0.07	0.01
Llama2-7b-hf	0.6	0.09	0.1	0.01
Llama3-8B	0.64	0.14	0.15	0.01

Off-the-Shelf LLMs?

1. Not Quite Good Enough with Code-Mixed Local Dialects using informal Swahili.
2. Expensive to Scale.
3. No full control over user data – compliance with local privacy laws

Training Methodology

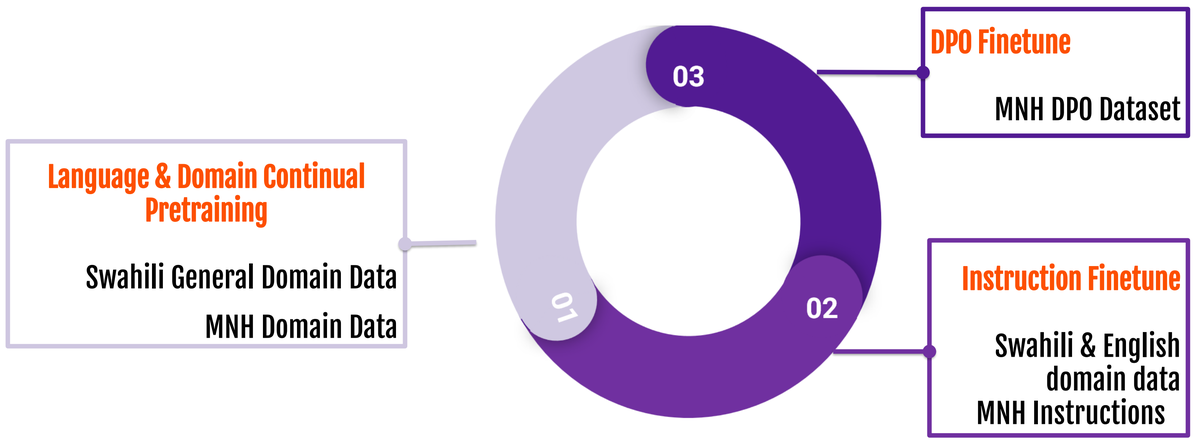


Figure 1 : Illustration of UlizaMama Training Methodology

Evaluation Methodology

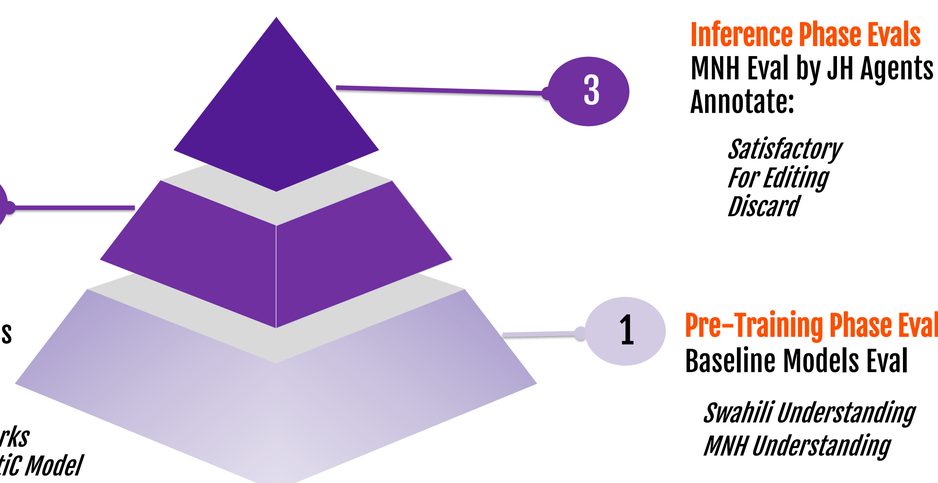


Figure 2 : Illustration of our Evaluation Methodology

Results

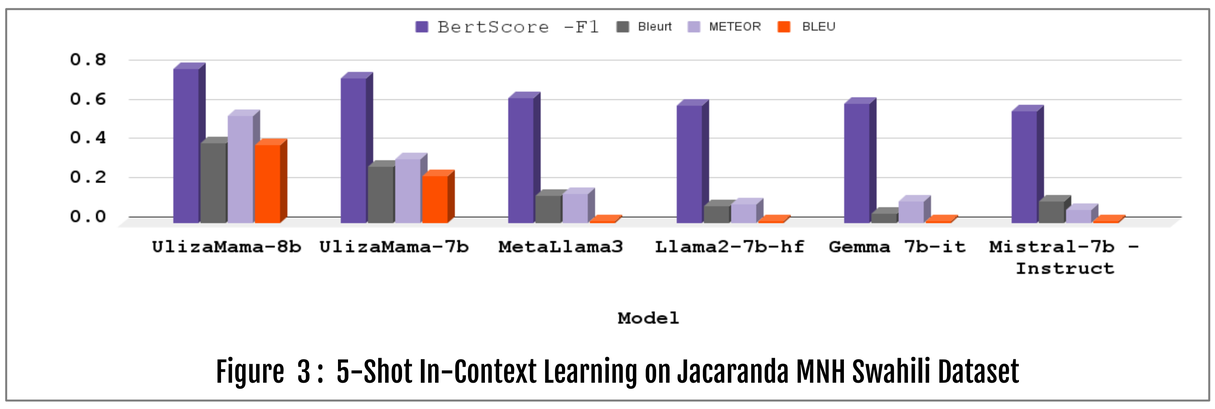


Figure 3 : 5-Shot In-Context Learning on Jacaranda MNH Swahili Dataset

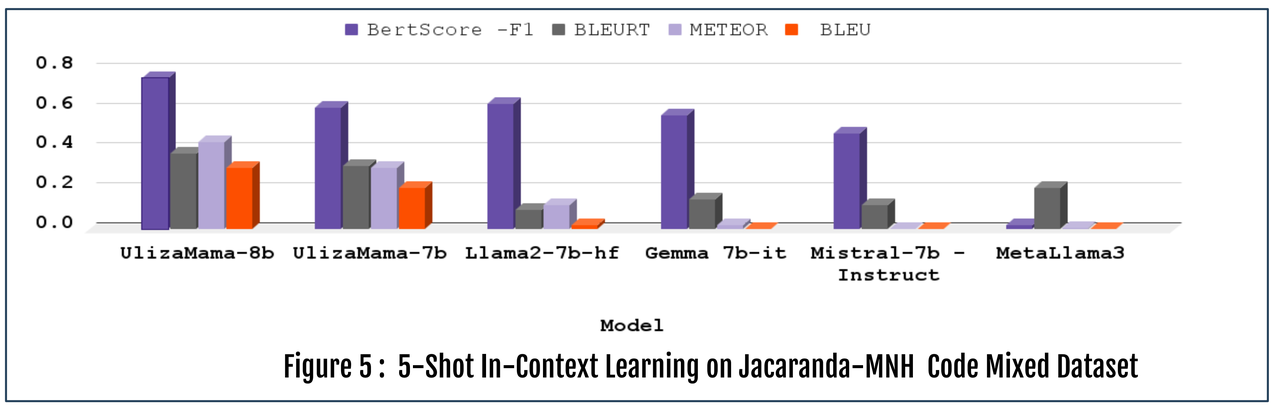


Figure 5 : 5-Shot In-Context Learning on Jacaranda-MNH Code Mixed Dataset

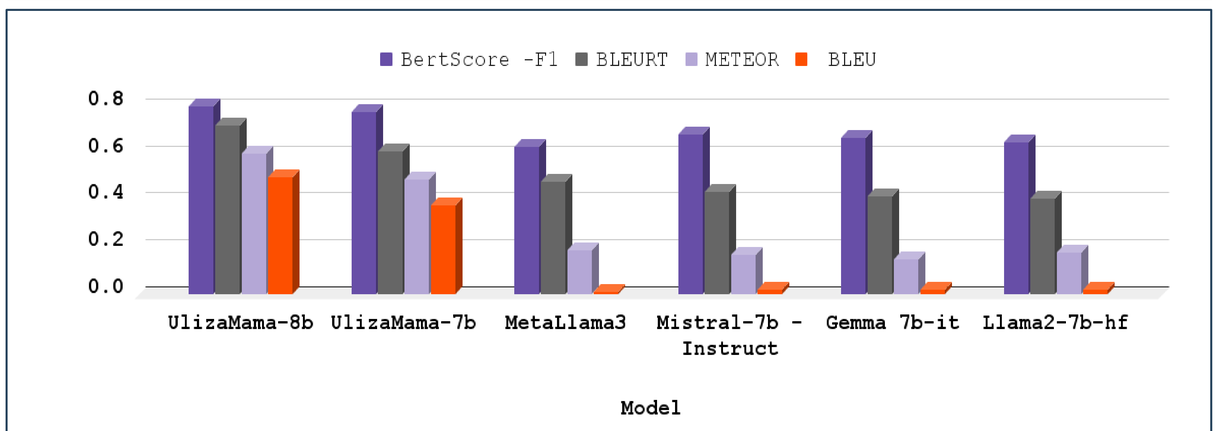


Figure 4 : 5-Shot In-Context Learning on Jacaranda MNH English Dataset

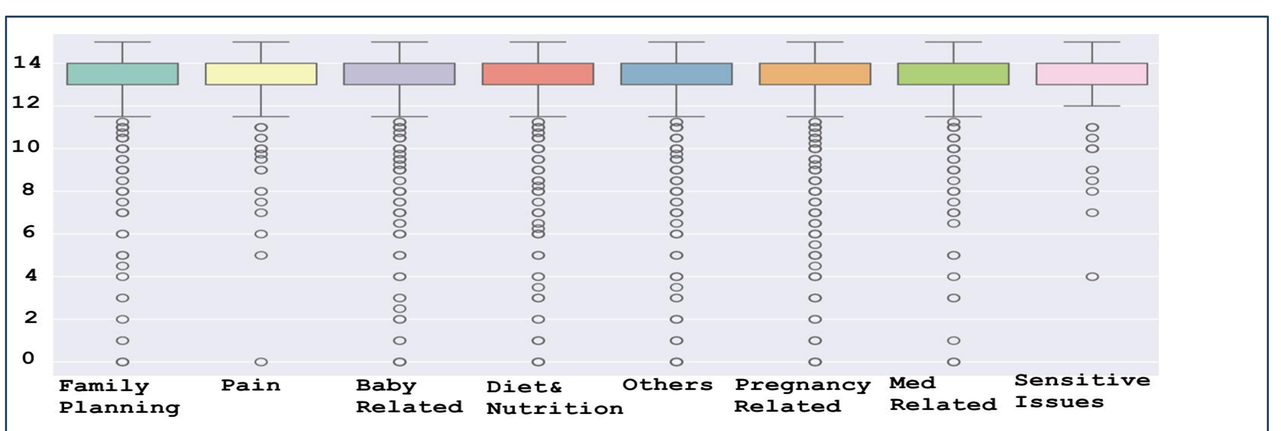


Figure 6: GCM-GPT Distribution of Audit Scores- UlizaMama 8b

Key Findings

Insights:

1. Improved performance of UlizaMama Models can be attributed to; tailored continual pretraining on blended datasets and tailored instruction tuning and DPO for MNH.

2. Development of the GCM-GPT Assisted Score provides a nuanced evaluation framework for medical content quality, scoring generated text (0-15) on three key aspects; Grammar (correct usage and structure), Communication efficacy (clarity and coherence), Medical accuracy (correctness and relevance of medical information).

3. Direct preference optimization with AI feedback yields high quality responses though it compromises human responses alignment.

Future Work:

1. Compare different Direct Preference Optimization techniques with UlizaMama.

2. Direct Preference Optimization finetuning of UlizaMama with Human Feedback.

3. Explore more hybrid and novel approaches to pretraining, finetuning and human alignment.

4. Extend UlizaMama to 5 African Languages: Hausa, Yoruba, isiXhosa, isiZulu, Wolof.

Google.org · Patrick J McGovern Foundation · aws · Strathmore University · Jacaranda Health · Jacaranda Health · @jacarandakenya
smwongela@jacarandahealth.org
jpatel@jacarandahealth.org
srajan@jacarandahealth.org