All You Need Is Context: Clinician Evaluations of various iterations of a Large Language Model-Based First Aid Decision Support Tool in Ghana.

Paulina Boadiwaa Mensah
SnooCODE Red Development Team
SnooCODE
Accra, Ghana
ORCID: 0000-0002-0570-5662

Nana Serwaa Quao
Accident and Emergency Centre
Korle Bu Teaching Hospital
SnooCODE
Accra, Ghana
ORCID: 0000-0001-8476-5999

Sesinam Dagadu
SnooCODE Red
SnooCODE
Accra, Ghana
s.dagadu@snoocode.com

Project Genie Clinician Evaluation Group1
Ghana
projectgenie314@gmail.com

***Abstract*— As advancements in research and development expand the capabilities of Large Language Models (LLMs), there is a growing focus on their applications within the healthcare sector, driven by the large volume of data generated in healthcare. There are a few medicine-oriented evaluation datasets and benchmarks for assessing the performance of various LLMs in clinical scenarios; however, there is a paucity of information on the real-world usefulness of LLMs in contextspecific scenarios in resource-constrained settings. In this work, 5 iterations of a decision support tool for medical emergencies using 5 distinct generalized LLMs were constructed, alongside a combination of Prompt Engineering and Retrieval Augmented Generation techniques. Quantitative and qualitative evaluations of the LLM responses were provided by 12 physicians (general practitioners) with an average of 2 years of practice experience managing medical emergencies in resource-constrained settings in Ghana.**

Keywords*— *SnooCODE, Clinical Decision Support, Large Language Models, First Aid, Emergency Medical Services, Medical Emergencies, Clinical Context, Clinician Evaluation, Resource-Constrained Settings, Gemini 1.5 Pro, GPT 4, Claude Sonnet

I. INTRODUCTION

"Provide a cool mist humidifier or take the infant into a steamy bathroom to help loosen mucus."* –** this was First Aid Step no.3 provided by Claude 3 Sonnet for managing possible Bronchiolitis or Asthma Exacerbations – two conditions that cause breathing problems. While this may be valuable advice, it might not be applicable to a child living on a rural cattle farm in Akobo, South Sudan. When this particular location is added to the prompt, the response makes no mention of mist humidifiers and steamy bathrooms. Rather the first step provided by the model is to ***"Move the infant to an area with fresh air and away from any dust/irritants." This shows the importance of considering the background contexts of prompts in evaluating the performance of Large Language Models (LLMs). Amongst the popular biomedical Natural Language Processing (NLP) datasets for evaluating LLMs, none of them have been specifically prepared for resource-constrained settings as found in Low-and Low-Middle-Income countries (LMICs)[2]. Thus, though a few models achieve high scores when evaluated on these datasets, their translational value in everyday clinical scenarios in LMICs cannot be readily ascertained.

In this work we aim to add to the limited knowledge base on LLM applications for clinical scenarios in LMICs. Specifically, we aim to evaluate the appropriateness of some selected generalized LLMs for use in clinical decision support tools in LMICs and to provide a reference for future, more expansive research. After conducting several experiments, we found that when generalized LLMs are given prompts that aim to generate first aid advice for medical emergencies, their outputs differ significantly when additional context-specific location is provided[3]. Thus, we provided context-specific prompts and asked clinicians with substantial familiarity with those contexts and clinical scenarios to evaluate the outputs. This work is part of a research and development process to eventually deploy LLM-based Clinical Decision Support tools for managing medical emergencies in resource-constrained settings.

II. METHODOLOGY

A. LLM Selection

We selected Open AI's GPT-4 Turbo Preview, both via the Assistant Application Programming Interface (API) and the Chat Completions API. We evaluated these separately as the temperature of the model is almost impossible to be tweaked when using the OpenAI Assistant. In addition we selected Gemini 1.5 Pro and Claude Sonnet. These models were selected based on performance on popular benchmarks[2], availability of API and ease-of-access. We did not select open medical LLMs such as Meditron-70B because of the

computational resources required to host/access them, for example, advanced GPUs. We then tested a combination of prompt-engineering and Retrieval Augmented Generation (RAG) techniques to produce outputs/responses from the various LLMs as follows:

- GPT 4-Turbo Preview via Open AI Assistant API + Prompt Engineering = ***Response A***
- Gemini 1.5 Pro + Prompt Engineering = ***Response B***
- Claude Sonnet + Prompt Engineering = ***Response C***
- GPT4-Turbo Preview via Open AI Chat Completions API + Prompt Engineering + RAG = ***Response D***
- Claude Sonnet + Prompt Engineering + RAG = ***Response E***

B. Parameter Tuning

The temperature was set at 0 for generating Responses C to E. This was to get deterministic responses as often as possible due to the critical nature of the proposed use case. For Response A, the default temperature used in Open AI Assistant was maintained as it was difficult to ascertain and tweak. For Response B, the default temperate of 2 set in the Google AI Studio was maintained as it was also difficult to tweak. An output length of 4000 was set in Google AI Studio for assessing Gemini 1.5 Pro to provide an ample window for the extent of generated responses. Similarly, the max tokens parameter was set at 4000 for assessing Claude Sonnet to provide an ample window for the extent of generated responses.

C. Prompt Engineering

We employed in-context learning using one-shot inference. The prompt consisted of three parts, the system message/prompt/instructions, an example conversation and the input message. Here is an example of the input message for one of the prompts:

“

Location: rural area, Bongo, Ghana. There is a chemist 300m away and a district hospital 1km away. Patient's age as: 5 months, sex as: male. Description of medical emergency: fall from stool, vomiting. 1. PATIENT CAN TALK NORMALLY 2. PATIENT CAN BREATHE NORMALLY 3. PATIENT HAS A NORMAL PULSE 4. PATIENT IS NOT VISIBLY BLEEDING 5. PATIENT IS AWAKE AND ALERT 6. PATIENT DOES NOT HAVE A VISIBLE TRAUMATIC INJURY, ANIMAL BITE OR RASH 7. PATIENT HAS NO KNOWN ALLERGIES 8. THE PATIENT HAS TAKEN PARACETAMOL 9. PATIENT
HAS NO KNOWN PAST MEDICAL HISTORY 10.THE TIME OF LAST MEAL WAS 30 MINUTES AGO

”

D. Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) has been touted as a highly promising approach to improving factuality, reasoning, and interpretability of LLM outputs[4,5]. We provided a free manual for first aid instruction geared towards settings in sub-Saharan Africa titled “Basic First Aid for Africa” published in 2017 by the Belgian Red Cross (and produced in conjunction with African experts). The text from the module was divided into chunks and vector embeddings generated by the “text-embedding-3-large” embedding model from Open AI. This embedding model works well with both GPT4Turbo and Claude Sonnet for retrievals but does not work as well with Gemini 1.0 Pro, thus RAG was not tested with the Gemini model. The vector embeddings were stored in the open-source vector database, ChromaDB.

E. Clinician Selection

Clinician evaluators were selected via a local clinician network, from diverse practice locations within Ghana and based on their familiarity with the locations, contexts, and clinical scenarios. Clinicians selected were verified to be in good standing with the Ghana Medical and Dental Council and had valid licenses to practice. Clinicians were asked to input their completed number of years of practice, and to round up surplus months to one year, for 10+ months and to round down to 0 years if less than 10 months. On average, clinician evaluators had 2 completed years of experience, as the first point-of-call in the hospital in managing medical emergencies in Ghana. It is expected that they possess sufficient knowledge and skills to deliver, at a minimum, first aid in the selected medical scenarios.

F. Selection of Medical Scenarios

“Love how it span over all major disciplines” – A clinician evaluator.

Six simulated clinical scenarios were provided in the format shown in Section above. The scenarios featured a wide range of demographics with the youngest simulated patient being 6 months old, and the oldest being 85 years old. The clinical scenarios cut across all major clinical specialties. There was an equal distribution of male and female patients in the scenarios represented.

G. Response Evaluation and Ranking

Each simulated scenario produced 5 responses making a total of 30 responses. At the end of each response, a 10-point Likert scale was provided for ranking the response. An evaluator had to select a number from 0 to 10, with 0 representing “Totally Unsatisfactory” and “Totally Satisfactory”. At the end of each Scenario-Responses pair, a comment box is provided for clinicians to input any additional comment about the scenario and accompanying 5 responses. Each of the 12 physicians ranked all 5 responses for every scenario, thus across the 6 scenarios, each response was ranked 72 times. A total of 360 rankings were then analyzed.

H. Collection and Analysis of Evaluation Reports Evaluation reports were collected via an online form. Quantitative analysis and associated visualizations were performed in

Microsoft Excel Version 16.83. The Real Statistics Resource Pack[6] was used for Interrater Reliability Analysis. For qualitative analysis, evaluators' comments were compiled as text in a document and coding was performed using Taguette 1.4.1-40-gfea8597[7]. Thematic analysis and visualization were performed in Python 3.11[8].

III. Results

A. Quantitative Analysis

Table 1 shows the ranking scores of the 12 evaluators labelled "1" to "12" for each of the responses labelled "A to E". These rankings are from the arithmetic mean of each evaluator's ranking of the 5 responses across the 6 prompts/scenarios, rounded up to the nearest whole number for ease of readability. The overall mean ranking was 6.6 with a standard deviation of 0.4.

TABLE I. Ranking Scores Per Evaluator.

Response	**A**	**B**	**C**	**D**	**E**
Evaluator	***Ranking scores***				
1	8	7	8	8	7
2	8	8	6	5	5
3	6	7	7	5	6
4	5	6	5	5	5
5	7	8	8	7	7
6	7	7	7	7	7
7	7	7	7	7	6
8	8	8	8	7	8
9	8	7	6	5	6
10	7	7	6	6	7
11	7	7	7	6	7
12	7	8	6	5	6
Mean± s.d[a]	6.8±0.4	7.2±0.7	6.6±0.6	6±0.6	6.4±1.2

a. standard deviation

Gemini 1.5 Pro + Prompt Engineering (Response B) elicited the highest rating scores: 7 or 8 out of 10, at least 90% of the time and it's lowest mean rating was 6. GPT 4-Turbo Preview via Open AI Assistant API + Prompt Engineering (Response A) had the second highest ratings: 7 or 8 out of 10, at least 80% of the time. Claude Sonnet + Prompt Engineering + RAG (Response E) had a score of 7 or 8 out of 10, 50% of the time. The worst ranked was GPT4-Turbo Preview via Open AI Chat Completions API + Prompt Engineering + RAG (Response D) with a score of 5 out of 10, 40% of the time. None of the responses had a mean rating below 5 (Figure 1).

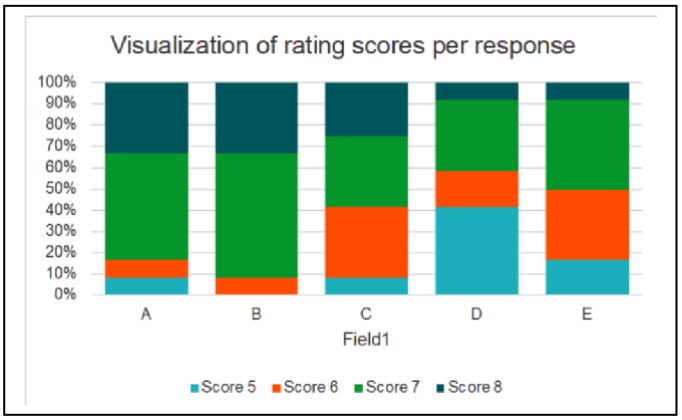

Fig. 1. Distribution of rating scores per Response category.

Gwet's AC_2 score using ordinal weights and a significance level (alpha) of 0.05 was calculated as a measure of interrater reliability. As seen in Table 2, there was a high level of agreement between evaluators, reflected by a Gwet's AC_2 score of 0.89.

TABLE II. Interrater Reliability Analysis

Statistic	***Score***
Gwet's AC2	0.89
Standard error for subjects	0.02
C.I.[a] lower end	0.83
C.I. [a] upper end	0.96
Total standard error (s.e.)	0.04
C.I. [a] lower end (s.e. accounted for)	0.77
C.I. [a] upper end (s.e. accounted for)	1

a. Confidence Interval

B. Qualitative Analysis

8 codes were generated representing recurring viewpoints expressed. Table 3 shows the 8 codes and their descriptions.

TABLE III. DESCRIPTION OF CODES

Code	***Description***
QuickTransfer	Emphasis on transferring or referring the casualty quickly to a health facility
ResponseSatisfaction	Expresses positive sentiments about the response
MissedDiagnosis	Mentions a diagnosis the response did not provide
NotConcise	Expresses dissatisfaction that response is not concise.
Concise	Expresses satisfaction that the response is concise
UnsureAboutCapabilityOfFacility	Expresses uncertainty and lack of confidence in nearby facilities
UnsureOfDiagnosis	Expresses low confidence in the outputted diagnosis
DisagreesOnPlan	Disagrees or is unsure of the first aid plan suggested.

ResponseSatisfaction was the most frequently occurring code, indicating numerous instances where the responses were considered satisfactory.

Concise and ***QuickTransfer*** also had significant occurrences, suggesting that the importance of conciseness in responses and the importance of quick transfers were often emphasized. ***MissedDiagnosis*** and ***NotConcise*** were less frequent but notable, indicating areas where responses may have missed critical diagnoses or were not concise enough. Figure 2 outlines the distribution of the codes.

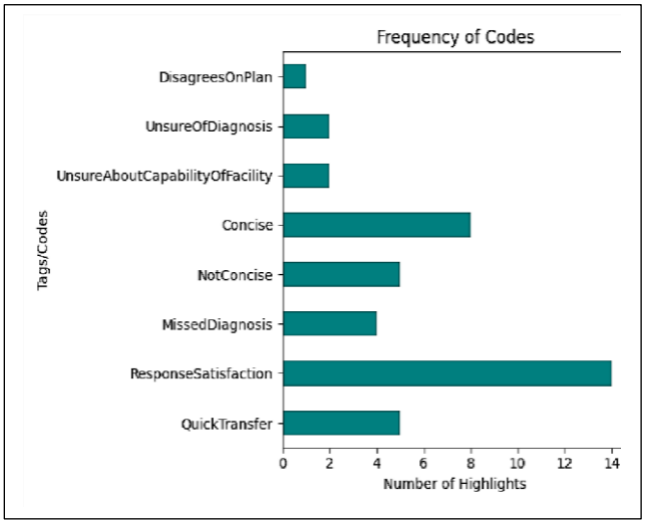

Fig. 2. Frequency of Codes in Analysis of Evaluators' Comments

The most commonly occurring codes were grouped into the following themes showing what clinicians considered most in evaluating scenarios and accompanying responses, arranged in descending order of frequency:

- Theme 1. **Clarity and Efficiency of Communication:** Includes ResponseSatisfaction, Concise, and NotConcise.
- Theme 2. **Diagnostic and Management Accuracy:** Includes MissedDiagnosis, UnsureOfDiagnosis and DisagreesOnPlan.
- Theme 3. **Urgency and Efficiency in Patient Transfer:** Includes QuickTransfer and UnsureAboutCapabilityOfFacility.

Table 4 details some of the clinicians' comments under each of these themes.

TABLE IV. EXAMPLES OF EVALUATORS' COMMENTS UNDER THE VARIOUS THEMES

Theme	***Supporting Quotes***
1	"Good responses provided overall." "B: First aid measures concise and accurate enough."
2	"Completely missed epistaxis as a diagnosis." "…too early to be considering asthma as first diagnosis."
3	"Don't wait till the patient deteriorates before you try and transfer to the nearest facility." "Transfer to the nearest hospital should be paramount."

The contexts surrounding the most frequently occurring codes expressing dissatisfaction with responses were further analyzed in a word cloud to identify areas of improvement. The larger the word, the more often it appears in the evaluators' comments. As shown in Figure 3, evaluators commented often that an emphasis should be placed on not waiting for Emergency Medical Services (EMS) but rather transferring the patient to the nearest facility. There was also a substantial number of complaints about some responses not being concise enough and thus not appropriate as first aid measures.

Fig. 3. Context analysis of "NotConcise" and "QuickTransfer" codes

IV. Discussion

Evaluators were generally satisfied with the diagnosis and first aid instructions outputted by the best performing generalized LLMs combined with moderate prompt engineering as indicated by both the quantitative and qualitative analysis results. This performance by the LLMs is notable considering that they had not had any prior pretraining or finetuning geared for the tasks. Also, the prompting strategy implemented was amongst the simplest with only one-shot inference. Past studies have shown that more sophisticated prompting strategies on generalized LLMs can lead to performances that out-perform state-of-the art, medical LLMs[9]. The best performing model in our study, achieved a mean ranking score of 7/10 which is encouraging. This is a positive finding for resourceconstrained settings where the ability to create more specialized, domain-specific models and/or to run them is greatly limited. If generalized models which are often more accessible to wider groups of people, can be made to perform at par/or better than specialized medical LLMs using simpler techniques, then developers in resource-constrained settings can take advantage to develop effective yet cost-efficient applications. An example of such applications is the SnooCODE Red app being developed in Ghana[10]. Figure 4 shows a version of the app in development.

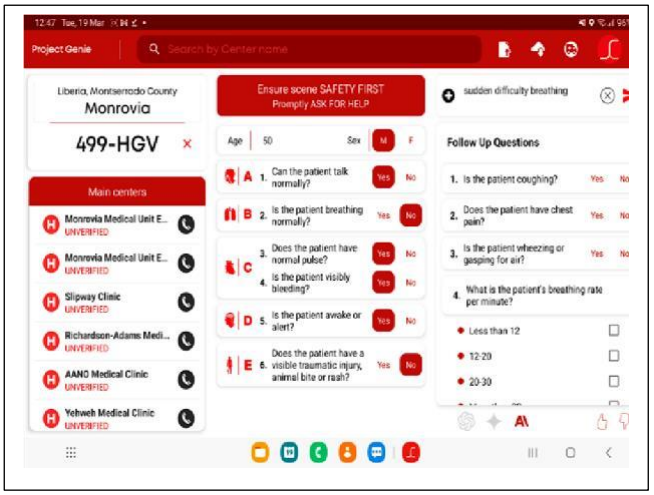

Fig. 4. A screenshot of the SnooCODE Red app under development

Though many studies have demonstrated the benefits of RAG in boosting the performance of generalized LLMs in domain specific tasks[4,5,11,12], more emphasis must be placed on RAG technique. Though internal experiments revealed that the addition of RAG can achieve better performance that prompt engineering alone, the findings from this study is that no RAG is better than RAG not done properly. Beyond, the embedding model used and the embedding and retrieval techniques, the content and formatting of the retrieval document can have a significant impact on the final model performance. This is a lesson that developers must pay attention to in the development of LLM-based applications.

The study also sheds more light on the importance of considering context in the evaluation of LLM performance. This is an area that human evaluators might beat machine evaluators. Clinician evaluators were not satisfied with responses that did not demonstrate a higher sense of urgency in the transfer of casualties to nearby health facilities even though in the prompt instruction, all the models were informed that EMS was on the way. Responses that instructed that the patient be transported to the nearest health facility even as first aid steps were being instituted were rated as more satisfactory. In contexts with better access to resources, evaluators might not have expressed such a strong concern about waiting for EMS. In many of the rural settings provided in the scenarios with meagre resources, this expression of concern was warranted. This underscores the huge importance of considering contexts in developing LLM-based clinical decision support tools . It is not enough that LLMs pass general medical benchmarks, their performance in different contexts must be evaluated, otherwise responses considered helpful in some settings may not only be unhelpful in other settings, but also harmful.

There are obvious limitations in this study. Firstly, a larger cohort of responses could have been evaluated. Also, a more

comprehensive evaluation framework could have been employed. We hope that the feedback obtained can be used to improve LLM outputs for the provided scenarios. We also hope that the insights derived can provide some direction in implementing more detailed and extensive studies of LLM outputs in resource-constrained settings.

V. Conclusion

LLM-based first aid assistants have the potential to provide clinically useful instructions in medical emergencies. This is especially helpful in resource-constrained settings where timely access to well-equipped health facilities is often difficult. This potential should be explored further to build applications which may prove life-saving in real-world settings

References

1. Project Genie Clinician Evaluation Group (March, 2023) https://bit.ly/clinician-evaluators-project-genie
2. Zhou, H., Gu, B., Zou, X., Li, Y., Chen, S.S., et al. (2023). A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. ArXiv, abs/2311.05112.
3. SnooCODE Red Team, "CONTEXT MATTERS: DIFFERENCES IN AI FIRST AID ASSISTANT OUTPUTS IN VARIOUS CONTEXTS." https://bit.ly/snoocodered-context-matters
4. Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A Survey on Retrieval-Augmented Text Generation. ArXiv, abs/2202.01110.
5. Anantha, R., Bethi, T., Vodianik, D., & Chappidi, S. (2023). Context Tuning for Retrieval Augmented Generation. ArXiv, abs/2312.05708.
6. Real Statistics Resource Pack (n.d.). Retrieved March 19, 2024, from https://real-statistics.com/free-download/real-statistics-resourcepack/
7. Rampin, R., Rampin, V. (2021). Taguette: open-source qualitative data analysis. Journal of Open Source Software, 6(68), 3522, https://doi.org/10.21105/joss.03522
8. Python Software Foundation. Python Language Reference, version 3.11. Available at http://www.python.org
9. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., et al. (2023). Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. ArXiv, abs/2311.16452.
10. SnooCODE. (n.d.). SnooCODE RED. Retrieved March 21, 2024, from https://snoocode.com/red
11. Soman, K., Rose, P.W., Morris, J.H., Akbas, R.E., Smith, B., et al. (2023). Biomedical knowledge graph-enhanced prompt generation for large language models. ArXiv, abs/2311.17330.
12. Gao, Y., Li, R., Croxford, E., Tesch, S., To, D., Caskey, J., Patterson, B., Churpek, M., Miller, T., Dligach, D., & Afshar, M. (2023). Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction. . https://doi.org/10.1101/2023.11.24.23298641.