Repeatability, Reproducibility, and Diagnostic Accuracy of a Commercial Large Language Model (ChatGPT) to Perform Disaster Triage Using the Simple Triage and Rapid Treatment (START) Protocol

Jeffrey Michael Franc[1,2], Attilla Hertelendy [3], Lenard Cheng[3], Ryan Hata[3], Manuela Verde[2]
[1]University of Alberta, Edmonton, AB, Canada, [2]Universita' del Piemonte Orientale, Novara, Italy [3]Harvard Beth Israel Deaconess Medical Center, Boston, MA, USA

INTRODUCTION

Engaging with chat-oriented large language models (LLMs) involves a straightforward, conversational interaction: the user provides text input (the *prompt*), and the model responds with text output. The release of ChatGPT in November 2022 drastically reduced the barrier to using artificial intelligence with an intuitive web-based interface to a large language model. Chat-based LLMs are powerful and intuitive, but they are not without shortcoming and have two major faults:

1. Hallucinations/Confabulations: outputs that are syntactically correct and superficially plausible but factually inaccurate or entirely fabricated.
2. Prompt Dependence: slight changes in phrasing or specifying different contexts can yield varying outputs, allowing for some level of "tuning" of the responses by modifying the prompt, but also introducing challenges related to consistency and reliability of the outputs. The study and practice of crafting and optimizing prompts to extract desired outputs from LLMs is sometimes referred to as "prompt engineering" or "promptology."

One potential situation where ChatGPT could be useful is in triaging patients at the site of a disaster. Although many disaster triage methods exist, the Simple Triage and Rapid Treatment (START) protocol is most common. This study addressed the research problem: "Can ChatGPT adequately triage simulated disaster patients using START?"

STUDY DESIGN

This study was based on a crossed gauge repeatability and reproducibility (gauge R and R) study with comparison to a standard.

Gauge R and R studies use repeated measurement of multiple parts by several appraisers using the same equipment, ensures that the output is the same as the input and that the same measurements occur over time.

In this study, the patient vignettes are the different "parts" while the various prompts are the "appraisers."

METHODS

This study used ChatGPT to triage simulated patient vignettes and assessed three major outcomes:

1.Repeatability - Variation in response with repeated use of the same prompt with the same patient vignette
2.Reproducibility - Variation in response with use of different prompts with the same patient vignette
3.Diagnostic Accuracy - Overall accuracy of triage when compared to a documented reference standard

Nine unique prompts were developed by five emergency physicians with disaster medicine training. A Python script queried ChatGPT Version 4 with each of the prompts combined with 391 validated patient vignettes. Ten repetitions of each combination were performed: 35190 simulated triages. Data analysis was performed using Stat59 (Stat59 Services LTD, Edmonton, AB, Canada)

Sample Patient Vignette

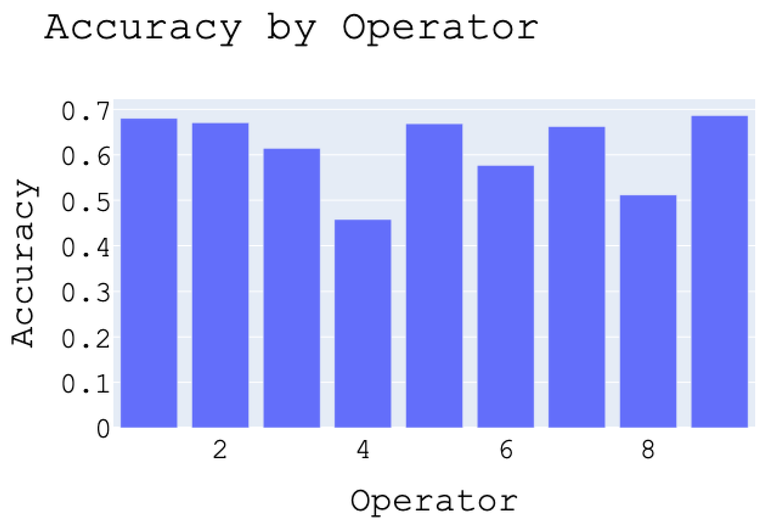
History: Highway speed MVC. Vehicle burst into flames. Extricated self through window. Wearing seatbelt. Airbag deployed. Complains of low back pain, pain in right ribs, and pain in left knee. Unable to Ambulate. *Examination*: Alert, oriented, no apparent distress. Pulse: 70. Respiratory rate 24. Blood Pressure: 152/75. GCS: 15

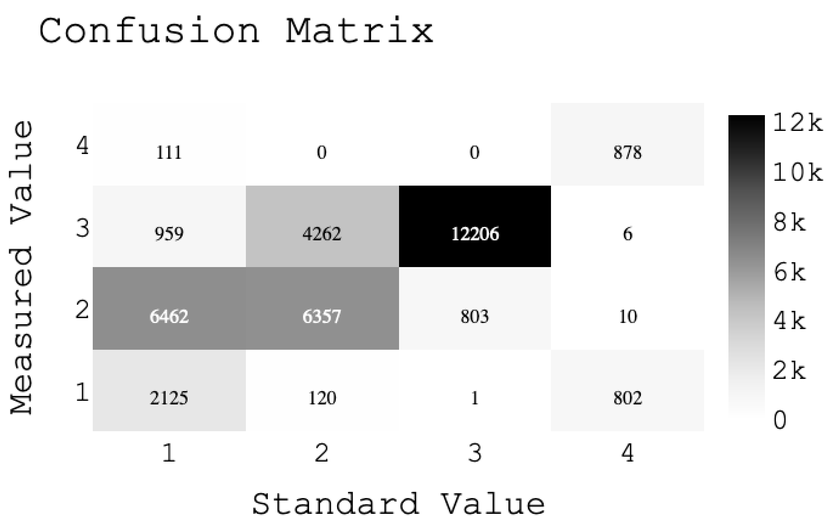
Sample prompt

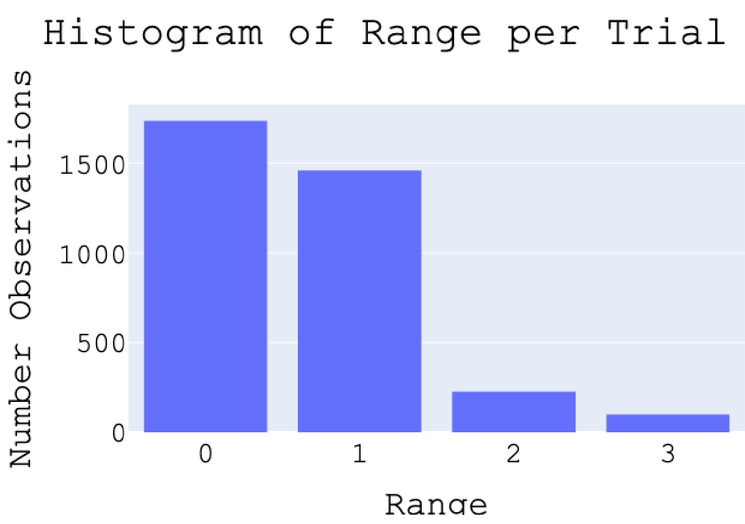
Take the role of a trained paramedic. Strictly use the "START (Simple triage and rapid treatment)" algorithm. Do not deviate from the START algorithm. Do not use other triage methods. Pay attention to walking status, breathing status, respiratory rate, capillary refill time, presence of radial pulse, and ability to obey commands. Triage the following patient by assigning to one of the following four categories: Red, Yellow, Green, Black. Respond only with the color code and no other text.

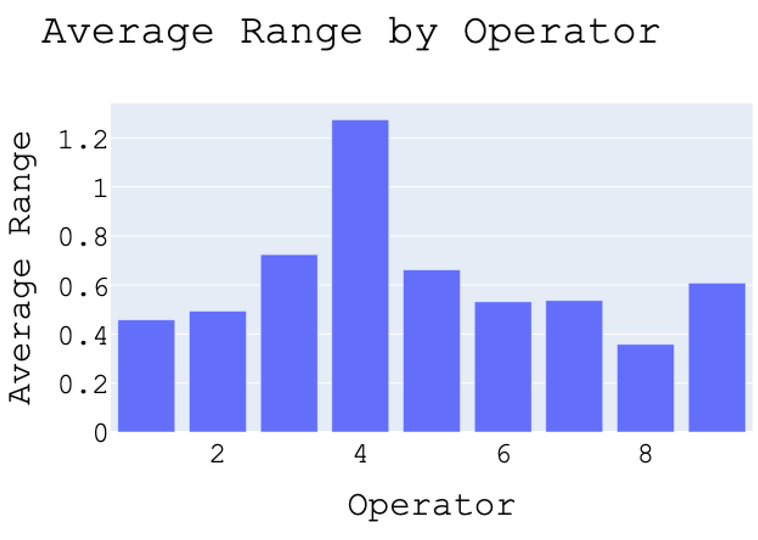
RESULTS

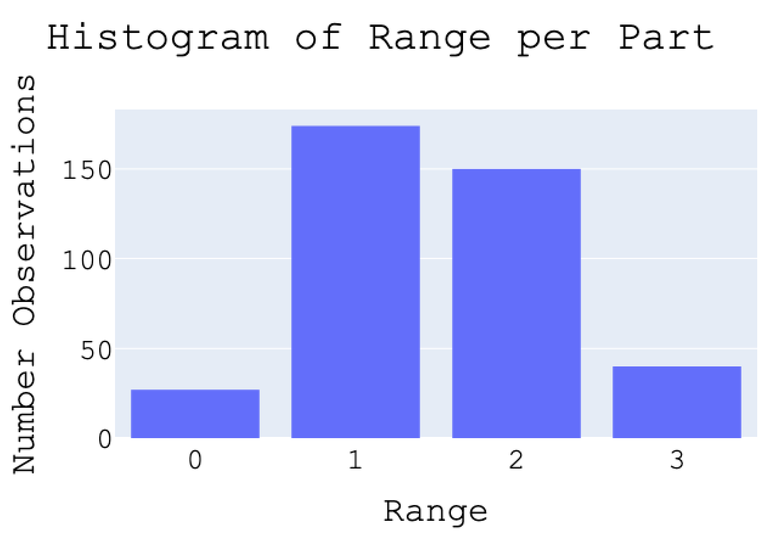
In 35102 (99.7%) of queries a valid START score was returned. However, there was considerable variability in the results
Repeatability (use of the same prompt repeatedly) was responsible for 14.0% of overall variation.
Reproducibility (use of different prompts) was responsible for 4.1% of overall variation.
Overall accuracy of ChatGPT for START was 63.9% with a 32.9% over-triage rate and a 3.1% under-triage rate. Accuracy varied by prompt with a maximum accuracy of 71.8% and a minimum accuracy of 46.7%.











CONCLUSIONS

This study suggests that the current ChatGPT large language model is not sufficient for triage of simulated patients using START due to poor repeatability and accuracy.
Medical practitioners should be aware that while ChatGPT can be a valuable tool, it may lack consistency and may frequently provide false information.

Disclosure Statement: JMF and MV are employees of Stat59.