Evaluating ChatGPT and Alpaca on the ACR Appropriateness Criteria to Help Ordering Physicians

Samuel Shenoi BS[1], Lorece Harris BS[1], Ashley Bancroft BS[1], Quan Nguyen MD[1]

[1] Department of Radiology, Baylor College of Medicine, 1504 Ben Taub Loop Houston, Tx-77030

BACKGROUND

In November 2022, OpenAI released ChatGPT, a bot utilizing a large language model (LLM) to interpret user entered text and output a response. ChatGPT has since been employed in healthcare settings, including in clinical decision support. Rau et al. utilized the ACR Appropriateness Criteria to develop an appropriateness criteria contexted chatbot, and were able to show non-inferiority amongst all their models as compared to radiologists. While the performance of their models were impressive, the authors noted some limitations such as the inability for the models to recommend no imaging as well as the significant cost in using these models. In this study, we evaluated the ability of ChatGPT and a open source LLM (Alpaca) to select appropriate medical imaging as compared to physician decision makers.

RESEARCH OBJECTIVES

- Do AI large language models have the ability to accurately choose appropriate medical imaging with similar accuracy to physicians?
- If similar in performance, what AI model shows the least amount of discrepancy to physician decision making?

STUDY DESIGN AND METHODS

- **Data Collection**
  - Cases (**n=42**) were collected from ACR Cortex. Cases from various specialties were selected.
  - ACR Appropriateness Criteria was downloaded from the American College of Radiology on 4/27/2023. Two hundred and twelve records were downloaded.
- **LLMs Used**
  - ChatGPT 3.5 (text-davinci-003) online demo version
  - Alpaca 7B deployed on a Google Cloud Compute Instance with 15 GB memory
- **New Model Development**
  - Langchain with a ChromaDB instance was used to create a modified ChatGPT and Alpaca 7B model. Models were run usin llama.cpp. Both models were discarded due to poor results and cost considerations (data not shown).
- **Testing**
  - All models were tested using the records from the collected data set. The first imaging modality output by the model was used fo comparison.
  - Responses were graded on a scale of 1-3 based on ACR appropriateness criteria guidelines.
  - A one-way ANOVA test was run in R (v 4.0.1) comparing the physician decision maker's imaging decisions to Alpaca and ChatGPT LLMs.
  - T-test comparing the outputs each model to each other, the expected results, and the physician decision maker's imaging decisions was run in R (v 4.0.1).

Number of Cases for Each Specialty					
Medicine	Surgery	Ob/Gyn	Neurology	Family	Pediatrics
7	10	9	10	11	10

Table 1: Number of cases collected from ACR Cortex per specialty. Cases were distributed equally between specialties.

Sample Cases Used

"A 62-year-old woman was admitted due to progressive dysphagia over the course of 3 months. The patient had lost 8 kg, but experienced no diarrhea, hematochezia, or abdominal distension. The patient had no history of malignancy. A physical examination found no abnormalities."

"A 30-year-old man presents with a 1-day history of nausea, vomiting, and crampy, nonradiating abdominal pain. He has a history of diabetes mellitus."

"A 27-year-old (gravida 0 para 0) woman presents to the emergency department with a 1-day history of worsening left lower quadrant abdominal pain with nausea and vomiting. On arrival, she is afebrile with a normal leukocyte count, with moderate blood and negative beta human chorionic gonadotropin on urinalysis. Physical examination reveals a tender left lower quadrant, and pelvic examination is normal."

Figure 1: Heatmap displaying the ACR Appropriateness Criteria scores for each case was built in R using pheatmap package (R v 4.0.1). The physician decision maker's (original) imaging decisions, ChatGPT's output for each case, Alpaca's output for each case, and the expected imaging decision based on the appropriateness criteria were all scored. Results show that no modality did as expected by ACR criteria.

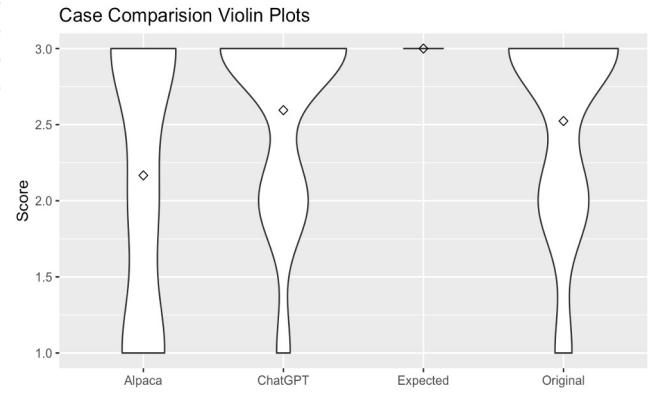

Figure 2: Violin plots displaying summary statistics built using R ggplot package (R v 4.0.1). Results show that the ACR Appropriateness Criteria score for the cases followed a similar distribution between physician decision makers (original) and ChatGPT. The mean score between ChatGPT and physician decision makers was also similar. The score distribution and mean of Alpaca differed from the other two.

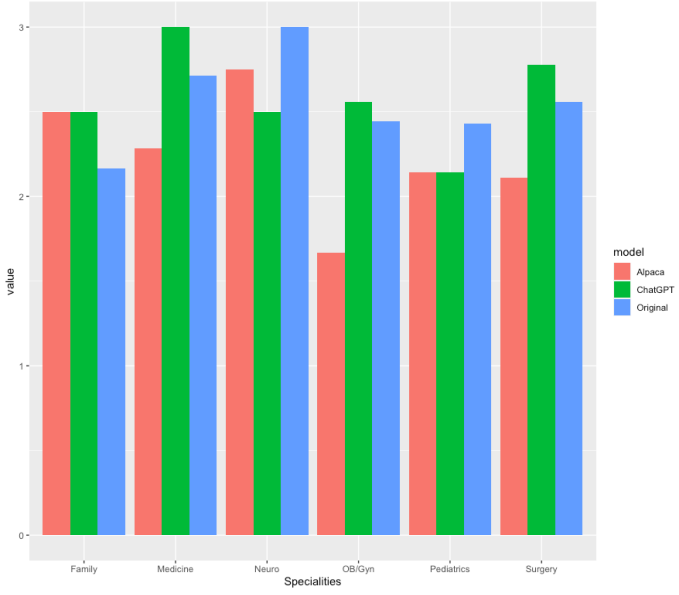

Figure 3: Bar graph representing the average score per model for each specialty was built in R using ggplot package. The results show that variance exists in models' performance between specialties.

RESULTS

- ANOVA results were as follows. The results displayed a clear difference between the means of the three groups.
  - Degrees of Freedom: 3
  - Sum Squares: 14.71
  - Mean Squares: 4.905
  - F value: 12.11
  - P-value: 3.36e-7
- Difference between ChatGPT and physician decision makers was not statistically significant (p=0.6157, CI -0.2105 0.3534)
- Alpaca was statistically significantly less accurate than physician decision makers (p=0.03999, CI -0.69752 -0.01675)
- All categories performed statistically significantly different from ACR Gold Standard criteria
  - ChatGPT was statistically significantly different from Gold Standard ACR criteria (p=0.0001475, CI -0.6001 -0.2093)
  - Alpaca was statistically significantly different from Gold Standard ACR criteria (p=2.835e-07, CI -1.1079 -0.5587)
  - Physician decision makers were statistically significantly different from Gold Standard ACR criteria (p=4.056e-05, CI -0.6853 -0.2669)
- ChatGPT had the highest mean score (2.595238) compared to physician decision makers (2.523810), and Alpaca (2.166667)

CONCLUSIONS

- ChatGPT and physicians performed with similar accuracy in determining appropriate imaging.
- Both ChatGPT and physicians performed superiorly to Alpaca.
- All categories showed significant deviation from ACR Appropriateness Criteria. This indicates that significant modification must be made to the AI LLMs before clinical implementation is appropriate.

FUTURE DIRECTIONS

- Future studies should focus on training AI models specifically to the ACR appropriateness criteria, not done in this study due to cost limitation.
- Further discussion is necessary to determine the minimum accuracy of AI large language models necessary for acceptable clinical implementation.
- Additional work is also needed to assess the accuracy of models in different clinical settings.
- It may be beneficial to analyze LLM performance by specialty to determine if there is a discrepancy between fields.

REFERENCES

All cases pulled from https://cortex.acr.org/CiP/Pages/CaseArchive

Rau A, Rau S, Fink A, et al. *A Context-Based Chatbot Surpasses Trained Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines*. Radiology and Imaging; 2023. doi:10.1101/2023.04.10.23288354