Long-Context and Short-Context Supplemented GPT-4 Surpass Base GPT-4 in Adhering to the American College of Radiology Appropriateness Criteria for Ordering Neuroradiology Imaging

Yasmine Eichbaum, Medical Student, Thomas Jefferson University
Julietta Gervase; Harish Appiakanna, MD; Rishi Gadepally, MD; Adam Flanders, MD, CIIP, FSIIM

Introduction

The American College of Radiology (ACR) Appropriate Use Criteria (AUC) are extensive, making them difficult to use efficiently. Foundational models have shown promise in providing appropriate recommendations, but the role of context length and order of information remains unclear.

Hypothesis

Compared to the base ChatGPT (GPT 4, 12/2023) which is not supplemented with AUC, our context supplemented versions of ChatGPT will provide more accurate imaging recommendations for neuroradiology clinical vignettes. Models with greater context will outperform ones with shorter context.

Methods

Four experimental models were created through the ChatGPT custom assistant web interface. Two were supplemented with the full AUC corpus (FC), while two were supplemented with tables only (TO). Two versions of each (FC and TO) were made by varying the order of AUC documents provided (FC1, FC2, TO1, TO2). These four, context-supplemented models plus the ChatGPT4 base model each processed fifty-one neurological clinical vignettes. Outputs were scored in accordance with the grading schema (Figure 1), with the final score being an average of the three runs per scenario. Kruskal-Wallis test was used to evaluate performance between models.

Results

All context-supplemented models performed significantly better against the base model (Figure 2) in terms of percent correct: FC1 (73%), FC2 (68.8%), TO1 (70.6%), or TO2 (73.3%) versus base (42.6%) ($p < 0.001$ for each). There were no significant differences between the FC and TO models ($p = 1.000$) or between the two version orders ($p = 1.000$).

Conclusion

Customized context models supplemented with AUC guidelines significantly outperformed the base model in providing appropriate imaging recommendations. There was no significant difference between the FC and TO models, nor between the models with varying orders of provided context.

Figure(s)

Grading Method	
Imaging Modality Classification in AUC Documentation	**Assigned Point Value**
"Usually Appropriate"	+1 point
"May Be Appropriate"	+0.5 points
"Usually Not Appropriate"	0 points
For each clinical vignette, the AUC-recommended imaging was scored per the criteria above and used to calculate the "maximum score." Then, model-generated responses were reviewed using the same criteria grading each imaging recommendation. The model's "final score" for each vignette was the sum of total points divided by the maximum score.	

Figure 1. Grading schema for model outputs

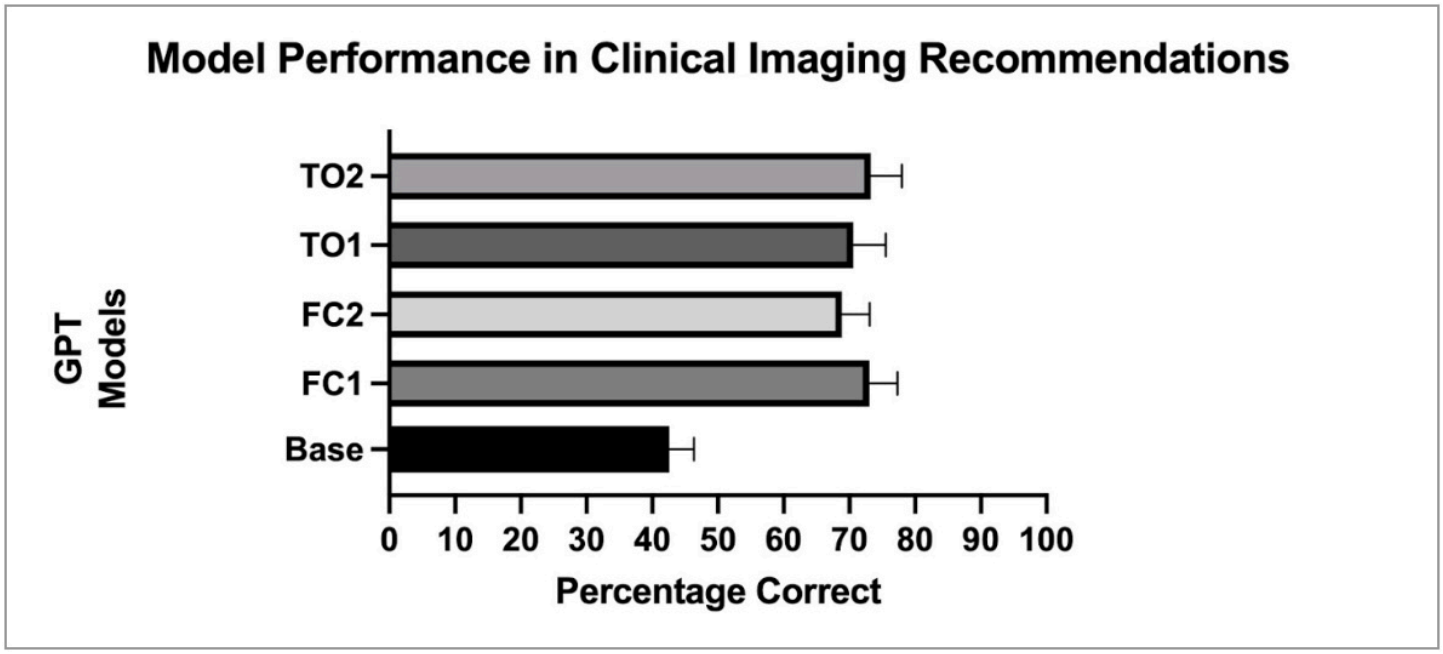

Figure 2. Comparing performance between the base, FC1, FC2, TO1, and TO2 models

Keywords

Applications; Artificial Intelligence/Machine Learning; Emerging Technologies; Imaging Research