Assessing the Performance of ChatGPT-4 Omni in Classifying Fracture and Non-Fracture X-ray Images

Nitin Chetla, Medical Student, University of Virginia School of Medicine
Swapna Vaja; Tamer Hage; Alper Turgut, MD; Tejas Sekhar; Joseph Chang; Mihir Tandon; Jorge Chahla, MD, PhD; Arun Krishnaraj, MD

Introduction

Radiographic assessment is the standard of care for diagnosing suspected fractures. While LLMs like ChatGPT-4 Omni (ChatGPT-4o) show promise in augmenting radiographic workflows, supporting evidence is limited. This study aims to evaluate ChatGPT-4o's performance in identifying various fracture types from the FracAtlas database.

Hypothesis

ChatGPT-4 Omni (ChatGPT-4o) will demonstrate high sensitivity but low specificity in classifying X-ray images as fracture or non-fracture, and its diagnostic accuracy will significantly vary depending on the prompt structure and wording used during the classification process.

Methods

The experiment evaluated ChatGPT-4 Omni (ChatGPT-4o) by classifying 1,000 X-ray images from the FracAtlas database (500 fracture, 500 non-fracture) of various body parts (hand, leg, hip, shoulder). Using a Python-based recursive loop, four prompts were tested: the first two (Test 1) asked if the image showed a fracture, with answer options reversed to test order effects. The other two prompts (Test 2) used different wording to test variations in response. Accuracy, precision, sensitivity, specificity, and F1 scores were calculated. Indecisive, incomplete, or refused responses were excluded from the analysis.

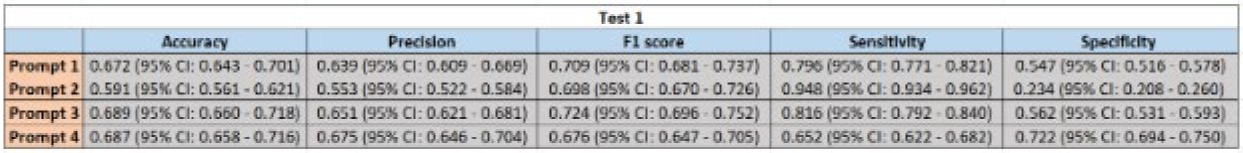
Results

The study found that ChatGPT-4 Omni (ChatGPT-4o) tends to over-identify fractures, with low to moderate specificity but high sensitivity. Changing the order of answer choices in prompts led to significant differences in accuracy and specificity. Prompt 1 achieved 0.672 accuracy and 0.547 specificity, while Prompt 2's accuracy dropped to 0.591, with specificity falling to 0.234. These results suggest ChatGPT-4o's outputs are dependent on prompt structure.Test 2 demonstrates similar differences across metrics of sensitivity, specificity, and F1 score.

Conclusion

Our study demonstrated statistically significant differences in intra- and inter-test accuracy, precision, sensitivity, specificity, and F1 score, suggesting that prompt outputs are dependent on user prompt input. Given ChatGPT-4o's generally moderate to high sensitivity but low specificity, the tool poses a significant risk of false positives, with more serviceable use cases being highly dependent on prompt input.

Figure(s)



Test 1					
	Accuracy	Precision	F1 score	Sensitivity	Specificity
Prompt 1	0.672 (95% CI: 0.643 - 0.701)	0.639 (95% CI: 0.609 - 0.669)	0.709 (95% CI: 0.681 - 0.737)	0.796 (95% CI: 0.771 - 0.821)	0.547 (95% CI: 0.516 - 0.578)
Prompt 2	0.591 (95% CI: 0.561 - 0.621)	0.553 (95% CI: 0.522 - 0.584)	0.698 (95% CI: 0.670 - 0.726)	0.948 (95% CI: 0.934 - 0.962)	0.234 (95% CI: 0.208 - 0.260)
Prompt 3	0.689 (95% CI: 0.660 - 0.718)	0.651 (95% CI: 0.621 - 0.681)	0.724 (95% CI: 0.696 - 0.752)	0.816 (95% CI: 0.792 - 0.840)	0.562 (95% CI: 0.531 - 0.593)
Prompt 4	0.687 (95% CI: 0.658 - 0.716)	0.675 (95% CI: 0.646 - 0.704)	0.676 (95% CI: 0.647 - 0.705)	0.652 (95% CI: 0.622 - 0.682)	0.722 (95% CI: 0.694 - 0.750)

Table 1. Relative effectiveness results for the correct identification of the presence of orthopedic fracture.

Keywords

Artificial Intelligence/Machine Learning