Performance of GPT-4 on the American College of Radiology In-Service Examination: Evaluating Accuracy, Model Drift, and Fine-tuning

David L. Payne[1,2], Kush Purohit[1] , Walter Morales Borrero[1] , Katherine Chung[1] , Max Hao[1] , Mutshipay Mpoy[1] , Michael Jin[1] , Prateek Prasanna[2] , Virginia Hill[3]

1: Stony Brook University Hospital Department of Radiology, 2: Stony Brook University Department of Biomedical Informatics, 3: Northwestern University Feinberg School of Medicine Department of Radiology

Introduction

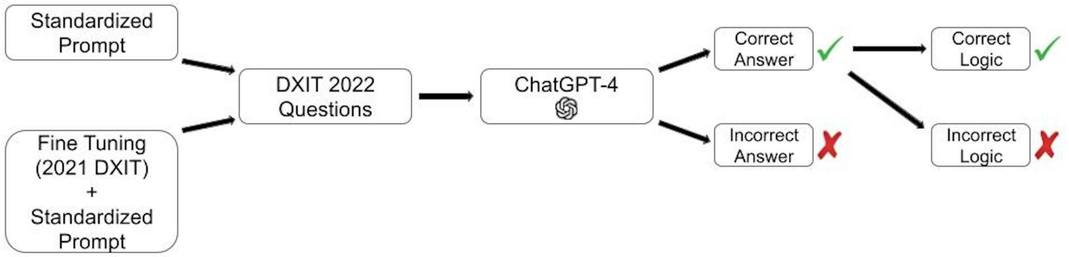
- To our knowledge, no previous work has evaluated the ability of GPT-4 to answer image-rich DR board-style questions or tested ChatGPT for model drift in its ability for medical image interpretation.
- In our study we evaluate GPT-4's performance on the American College of Radiology (ACR) 2022 Diagnostic Radiology In-Training Examination (DXIT). We perform multiple experiments across time points to assess for model drift, as well as after fine-tuning to assess for differences in accuracy.

Methods

- The 2022 and 2021 DXIT questions, publicly available online, were used in our study.
- Questions were sequentially input into GPT-4 with a standardized prompt. Each answer was recorded and overall accuracy was calculated, as was logic-adjusted accuracy, and accuracy on image-based questions. This experiment was repeated several months later to assess for model drift, then again after the performance of fine-tuning to assess for changes in GPT's performance.

Results

- GPT-4 achieved 58.5% overall accuracy, lower than the PGY-3 average (61.9%) but higher than the PGY-2 average (52.8%). Adjusted accuracy was 52.8%. GPT-4 showed significantly higher ($p = 0.012$) confidence for correct answers (87.1%) compared to incorrect (84.0%). Performance on image-based questions was significantly poorer ($p < 0.001$) at 45.4% compared to text-only questions (80.0%), with adjusted accuracy for image questions of 36.4%. When the questions were repeated, GPT-4 chose a different answer 25.5% of the time and there was no change in accuracy. Fine-tuning did not improve accuracy.



Questions were input in November-December 2023, February 2024, and March 2024

Conclusion

- GPT-4 performed between PGY-2 and PGY-3 levels on the 2022 DXIT, but significantly poorer on image-based questions, and with large variability in answer choices across time points, but without improvement with fine-tuning. This study underscores the potential and risks of using minimally-prompted general AI models in interpreting radiologic images as a diagnostic tool. Implementers of general AI radiology systems should exercise caution given the possibility of spurious yet confident responses.

References

- Please use the QR Code below to access our preprint on bioRxiv, which contains a full list of references

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY