Japanese Journal of Gastroenterology and Hepatology

Research Article

ISSN: 2435-1210 | Volume 10

Does ChatGPT Pass the LIRADS Test? Comparing Quality of AI Generated Impressions to Human Reports

Perchik J*, Godwin R, West J, Summerlin D, Zahid M and Smith A

Department of Diagnostic Radiology, University of Alabama at Birmingham, USA

***Corresponding author:**

Jordan Perchik,
Department of Diagnostic Radiology, University of Alabama at Birmingham, 619 19th Street South, Birmingham, AL 35294, USA

Received: 28 Sep 2023
Accepted: 01 Nov 2023
Published: 10 Nov 2023
J Short Name: JJGH

Citation:

Perchik J. Does ChatGPT Pass the LIRADS Test? Comparing Quality of AI Generated Impressions to Human Reports. J Gastro Hepato. 2023; V10(5): 1-5

Keywords:

Artificial intelligence; Large language model

1. Abstract

1.1. Purpose: Few artificial intelligence (AI) applications have garnered the same level of attention and excitement as ChatGPT. This large language model (LLM) chatbot has been touted for its potential applications in business, law, and healthcare. One such application in radiology is automated report conclusion generation. AI natural language processing (NLP) applications typically perform best when using a defined lexicon for a specific disease process, such as LI-RADS for reporting of hepatocellular carcinoma. Several commercially available NLP applications are already utilized in radiology, and in this study, we evaluate the report quality of a ChatGPT based impression generator compared to human generated conclusions and a commercially available radiology NLP application.

1.2. Methods: Abdomen MRIs performed between March 1 and March 28, 2023 and containing the term "LIRADS" were collected. The reports from 30 exams were randomly selected, exported and anonymized. The human generated conclusions were exported and AI generated reports were acquired from an application utlilizing ChatGPT-3.5 and a commercially available radiology NLP application. Three subspecialty trained abdominal imaging radiologists graded the quality of the conclusions on a scale of 1 to 10 with 10 being an attending level report requiring no edits, 5 being a correct report requiring substantial edits, and 1 being an incorrect or unusable report. Readers were also asked to select the highest quality report, the lowest quality report, and select which report they thought was human generated for each exam. Results were compared using a student t-test.

1.3. Results: Human generated reports (average = 9/10) were of a significantly higher quality than those generated by ChatGPT-based model (average = 5.3/10, $p<0.001$) and by the commercially available radiology NLP application (average=5.7/10, $p<0.001$). There was no significant difference in the quality of the AI generated conclusions between ChatGPT-based model and the commercially available application ($p=0.64$). Human reports were consistently selected to be the highest quality of the three options (69/90). The lowest quality exam was nearly exclusively AI generated (88/90) with the ChatGPT-based model representing the majority (48/90). The readers correctly identified the human generated report in the majority of cases (70/90) with a minority belonging to the radiology NLP application (17/90) and to the ChatGPT model (3/90).

1.4. Conclusion: AI generated radiology report conclusions, including those generated by a ChatGPT-based model and by a current commercially available NLP application, are of significantly lower quality than human generated conclusions.

1.5. Clinical Significance: To assess the performance of ChatGPT and a current commercially available radiology NLP application to generate human quality radiology reports.

2. Introduction

Radiology has been a leading medical specialty for the integration of artificial intelligence (AI) in clinical practice. Radiology has experienced exponential growth in research and publication since 2000 and radiology represents over two thirds of FDA cleared AI appli-

cations in healthcare [1-3]. Radiology, as a specialty, is particularly well-suited to AI integration due to its integral relationship with information technology and imaging informatics, the large volume of imaging data available for algorithm training, and the numerous steps involved in exam acquisition, interpretation, and reporting that can be optimized by AI integration [4, 5]. Because of radiology's position at the forefront in medicine, it can sometimes be viewed as a proving ground for novel techniques and technologies in AI, with one example being the application of large language models (LLMs) in the medical field.

Few AI applications have garnered the same level of attention and excitement as ChatGPT. This LLM chatbot has been touted for its potential applications in business, law, and healthcare. ChatGPT has been shown to perform well in recommending imaging protocols for emergency department radiology exams, provide patient information for breast cancer prevention and screening, and achieved a near passing score on a radiology board-style certification exam [6-8]. ChatGPT could also be integrated into radiology exam reporting, with potential applications including automated generation of exam impressions, providing a layperson summary of the exam report, and providing quality control checks to ensure that there are not report errors on laterality (right or left sided findings) or patient demographics (name, date of birth, sex, gender, medical record number, etc.) [9]. Automated generation of report impression has a benefit of not only increasing radiologist efficiency, but also decreasing the potential for dictation error and decreasing the radiologist's cognitive load [10].

Although LLMs and other natural language processing (NLP) models can technically be used for any exam type, AI applications typically perform best on report types with a defined reporting system and a defined lexicon, such as breast cancer (Breast Imaging Reporting and Data System), thyroid nodules (Thyroid Imaging Reporting and Data System), and hepatocellular carcinoma (Liver Imaging Reporting and Data System) [4]. Cirrhosis is a leading cause of death worldwide (2.4% of deaths in 2019) with a 1-6% incidence of hepatocellular carcinoma per year [11, 12]. Automated report generating AI applications are already available for use in radiology, however the performance of ChatGPT has not been compared to existing NLP models and human users. In this study, we evaluate the quality of the automatically generated reports generated by a ChatGPT based impression tool compared to human generated conclusions and a commercially available radiology NLP application.

3. Methods

This IRB approved study was performed at a single, tertiary care university hospital in the United States. Abdomen MRIs performed between March 1 and March 28, 2023 and containing the term "LI-RADS" were collected. From the total aggregate of exam reports, 30 exams were randomly selected, exported and anonymized. The human generated conclusions were exported and AI generated reports were acquired from an application utilizing ChatGPT-3.5 and a commercially available radiology NLP application. The ChatGPT conclusions were generated by uploading the report findings for each exam along with structured prompts to increase conclusion quality, also known as prompt engineering [13]. These prompts included "please provide numbered conclusions using the LIRADS lexicon", "limit conclusions to five", "provide management recommendations for critical findings", and "group incidental findings". Conclusions were randomized and distributed to three subspecialty trained Abdominal Imaging Radiologists (2-6 years of independent practice). A document containing the original exam findings for each report was provided to the readers to give additional clinical background, but no patient identifying information or exam images were included in the reports (Figure 1).

The reviewers graded the quality of the conclusions on a scale of 1 to 10 with 10 being an attending level report requiring no edits, 5 being a generally correct report that required substantial edits, and 1 being an incorrect or unusable report. Readers were also asked to select the highest quality report, the lowest quality report, and select which report they thought was human generated for each exam. Results were compared using a student t-test.

4. Results

A total of 36 total Abdomen MRI reports containing the term "LI-RADS" were performed during the study period, 30 of which were randomly selected for further analysis. Three reviewers reviewed 30 cases, resulting in a total of 90 cases.

Human generated reports (average quality score = 9/10) were of a significantly higher quality than those generated by ChatGPT-based model (average quality score = 5.3/10; $p<0.001$) and by the commercially available radiology NLP application (average=5.7/10, $p<0.001$) (Figure 2). There was no significant difference in the quality of the AI generated conclusions between ChatGPT-based model and the commercially available application ($p=0.64$). Human reports were consistently selected to be the highest quality of the three options (69/90). The lowest quality exam was nearly exclusively AI generated (88/90) with the ChatGPT-based model representing the majority (48/90). The readers correctly identified the human generated report in the majority of cases (70/90) with a minority belonging to the radiology NLP application (17/90) and to the ChatGPT model (3/90).

Case 1
EXAM: Outside MR Image Interpretation

CLINICAL INFORMATION: Liver lesions. Per review of outside records, the patient underwent laparoscopic biopsy of these liver lesions, with outside pathology demonstrating hepatocellular carcinoma in both lesions.

COMPARISON: Outside PET/CT 12/22/2022, outside MRI abdomen without and with contrast 12/28/2022.

FINDINGS:

STRUCTURED REPORT: MR HCC Screening

IMAGE QUALITY: Satisfactory

LOWER CHEST:
LUNG BASES / PLEURA: No significant abnormality.
HEART / VESSELS: No significant abnormality.

ABDOMEN:

LIVER: Cirrhotic. No steatosis.

LIVER LESIONS:
- Observation number: 1
- Focality: Multiple lesions.
- Description: Two similar lesions in the right hepatic dome demonstrate ill-defined nonrim arterial hyperenhancement with washout and intralesional microscopic fat.
- Location: Segment(s) 7 and 8
- Size: 2.4 x 2.3 cm (segment 7 observation on series 804, image 17) and 2.8 x 2.3 cm (segment 8 cm on series 804, image 8).
- Enhancement: Nonrim arterial phase hyperenhancement
- Vascular invasion: No
- Additional major features present: 1
  - Enhancing "capsule": Not present.
  - Nonperipheral "washout": Present.
  - Threshold growth (>= 50% in <= 6 months): Not present.
- Ancillary features:
  - Favoring HCC: Fat in mass, more than adjacent liver.
  - Favoring malignancy: Restricted diffusion.
  - Favoring benignity: None.
- LI-RADS: Not applicable, pathologically proven hepatocellular carcinoma.

LIVER VASCULATURE AND COLLATERALS:
- Hepatic artery patency and anatomy: Patent with conventional anatomy.
- Portal venous system: Patent intra- and extra-hepatic portal venous system.
- Hepatic veins: Patent right, middle and left hepatic veins.
- Esophageal varices: None.
- Other varices or collaterals: None.

LIVER RELATED FINDINGS:
BILIARY DUCTS: No biliary dilatation.
GALLBLADDER: Absent.
LYMPH NODES: None enlarged.
SPLEEN: Normal size and appearance.
PERITONEUM / ASCITES: No ascites.

OTHER FINDINGS:
PANCREAS: Normal.
ADRENALS: Normal.
KIDNEYS: Normal.
STOMACH / SMALL BOWEL: No abnormality.
COLON: No abnormality.
MESENTERY: Normal.
RETROPERITONEUM: Normal.
OTHER VESSELS: Circumaortic left renal vein.
BODY WALL: Postsurgical changes to the ventral abdominal wall.

MUSCULOSKELETAL: No aggressive osseous abnormality is identified. Multilevel degenerative changes in the lumbar spine.

(a)

Option 1	Option 2	Option 3
1.Two arterially hyperenhancing lesions with washout in the right hepatic dome are compatible with pathologically proven multifocal hepatocellular carcinoma. No evidence of metastatic disease in the abdomen. 2. Hepatic cirrhosis without sequela of portal venous hypertension.	1.Cirrhotic liver with two similar lesions in the right hepatic dome demonstrating ill-defined nonrim arterial hyperenhancement with washout and intralesional microscopic fat. Pathologically proven hepatocellular carcinoma. 2. Patent intra- and extra-hepatic portal venous system, patent right, middle and left hepatic veins. 3. Postsurgical changes to the ventral abdominal wall and multilevel degenerative changes in the lumbar spine.	1. No steatosis. 2. Ill-defined foci in the right hepatic dome demonstrates irregular hyperenhancement with washout and intralesional microscopic fat. Favoring malignancy. Restricted diffusion in this region. 3.Known HCC: None. 4.Multiple metastatic foci are seen throughout the abdomen and pelvis. Restricted hepatocellular carcinoma is considered less likely. Ill-defined foci in the right hepatic dome demonstrates irregular hyperenhancement with washout and intralesional microscopic fat. Favoring malignancy. Restricted diffusion in this region. Known HCC: None.

Quality of report (1-10) 1.___10____ 2. __7_____ 3. ___5____ (10-attending quality, 7-minor edits, 5-substantial edits, 1-unusable, incorrect conclusions; Incorrect use of LIRADS = 1)

Best report: ___1_____ Worst report: ____3_____ Human report: _____1_____

(b)

Figure 1: Example of distributed materials with original report findings (a) and human, commercially available NLP application, and ChatGPT generated conclusions (b). In this case, Option 1 is human generated, option 2 is generated by the commercially available NLP application, and Option 3 is generated by ChatGPT. A completed score sheet for this case is included in b).

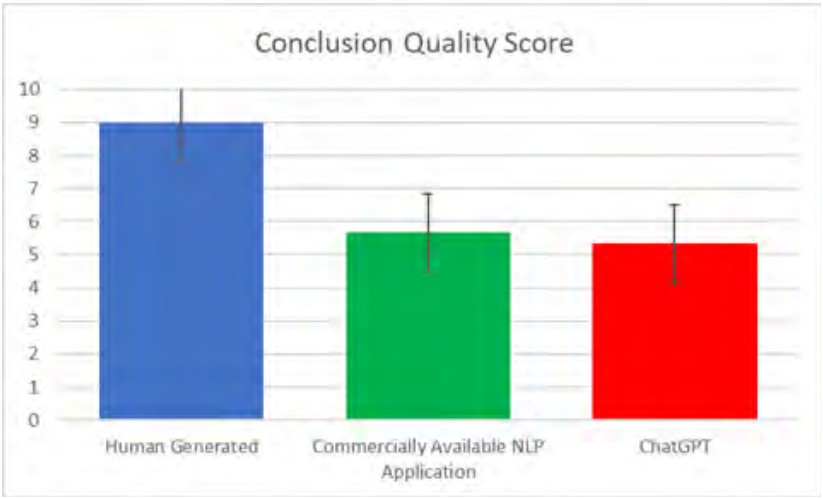

Figure 2: Average impression quality score of human generated conclusions and conclusions generated by commercially available NLP (natural language processing) application and ChatGPT. The human generated conclusions were of a significantly higher quality than both the commercially available AI application and ChatGPT ($p<0.001$ and $p<0.001$). There was no significant difference in quality between the conclusion quality of the commercially available NLP application and ChatGPT.

5. Discussion

Neither the ChatGPT based tool nor the commercially available NLP application approached the quality of human generated reports in radiology. This case series was highly standardized, which theoretically should generate an optimal performance for a clinical AI tool. This case series consisted of a single exam type, on a single modality, on a single region of anatomy, and utilized a standardized reporting lexicon (LIRADS), however the performance of AI generated conclusions was of significantly lower quality compared to human-generated reports. It is expected that this difference would be even more pronounced with more diverse data covering a wider range of pathologies and study types. Not only were the AI generated conclusions, including both the commercial application and the ChatGPT generated conclusions, consistently rated as the lowest quality on each case (88/90), but reviewers were also able to correctly identified the human generated report in the majority of cases (70/90). This infers that in addition to a lower overall quality, the AI tools were not able to consistently pass the so-called "Turing Test" and provide a convincingly human response.

There was not significant difference in the quality of reports generated by the commercially available NLP application and the ChatGPT based model, however the average score of the commercially availa-

ble application was slightly higher (5.7/10 and 5.3/10, respectively). ChatGPT can be accessed for free through OpenAI (San Francisco, California), and in the absence of a statistical difference in exam quality, may be considered as an alternative to paid conclusion generating applications. It should be noted that ChatGPT is not inherently HIPPA compliant and private health information should not be used in the OpenAI platform [14]. There are also considerations with data ownership and responsible use of information that must be considered. Crucially, the radiologist maintains final responsibility for the exam report, meaning that they are responsible for the quality of the report that is submitted to the medical record and are liable for any errors or adverse outcomes that occur as a result of their AI assistant.

A small amount of prompt engineering (providing additional instructions to guide algorithm output) was utilized for the ChatGPT portion of the data acquisition. It is possible that the quality of the ChatGPT output could have been improved with more robust prompt engineering [13], however the purpose of this study was to test the performance of ChatGPT in a more naïve state compared to a commercially available NLP counterpart. Additional limitations of the study include the small number of exams and reviewers and the subjective assessment of report quality. Future iterations of this study could include a larger, multi-institution data set and reader cohort.

6. Conclusion

AI generated radiology report conclusions, including those generated by a ChatGPT-based model and by a current commercially available NLP application, are of significantly lower quality than human generated conclusions. These technologies will require continued training and oversight to reach clinical utility.

7. Author Contributions

All authors substantially contributed to the conception and design of the work and the writing and revision of the manuscript. All authors approve the final version of the manuscript and are accountable for its contents. The authors declare that they had full access to all of the data in this study and the authors take complete responsibility for the integrity of the data and the accuracy of the data analysis.

References

1. West E, Mutasa S, Zhu Z, ha R. Global trend in artificial intelligence-based publications in Radiology from 2000 to 2018. American Journal of Roentgenology. 2019; 213: 1204-120.
2. Perera N, Perchik JD, Perhcik MC, Tridandapani S. Trends in medical artificial intelligence publications from 2000-2020: Where does radiology stand? Open Journal of Clinical and Medical Images. 2022.
3. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices.
4. Perchik JD, Rothenberg SA, Tridandapani S. Artificial intelligence in Body Imaging: an overview of commercially available tools. Seminars in Roentgenology. 2023; 58(2): 196-202.
5. Yang L, Ene IC, Belaghi RA. Sakeholders' perspecives on the future of artificial intelligence in radiology: a scoping review. European Radiology. 2022; 32: 1477-1495.
6. Barash Y, Klang E, Konen E, Sorin V. ChatGPT-4 Assistance in optimizing Emergency Department Radiology referrals and imaging selection. Journal of the American College of Radiology. 2023.
7. Haver HL, Ambinder EB, Bahl M. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology. 2023; 307(4): e230424.
8. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into current strengths and limitations. Radiology. 2023; 307(5): e230582.
9. Elkassem AA, Smith AD. Potential use caess for ChatGPT in Radiology reporting. American Journal of Roentgenology. 2023.
10. Zhang Y, Ding DY, Qian T. Learning to Summarize Radiology Findings. In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis. 2018; 204–213. Brussels, Belgium. Association for Computational Linguistics.
11. Lllovet JM, Kelley RK, Villanueva A. Hepatocelular carcinoma. Nature Reviews Disease Primers. 2021; 7(6).
12. Huang DQ, Terrault NA, Tacke F. Global epidemiology of cirrhosis - aetiology trends and predictions. Nature Reviews Gastroenterology and Hepatology. 2023; 20: 388-398.
13. Giray L. Prompt engineering with ChatGPT: A guide for academic writers. Annals of Biomedical Engineering. 2023.
14. Brande A. Using ChatGPT and generative AI in a HIPAA-compliant way.