### RSNA Press Release

# GPT-4 Matches Radiologists in Detecting Errors in Radiology Reports

Released: April 16, 2024

OAK BROOK, Ill. — Large language model GPT-4 matched the performance of radiologists in detecting errors in radiology reports, according to research published today in *Radiology*, a journal of the Radiological Society of North America (RSNA).

download photo

Roman J. Gertz, M.D.

Errors in radiology reports may occur due to resident-to-attending discrepancies, speech recognition inaccuracies and high workload. Large language models, such as GPT-4, have the potential to enhance the report generation process.

“Our research offers a novel examination of the potential of OpenAI’s GPT-4,” said study lead author Roman J. Gertz, M.D., resident in the Department of Radiology at University Hospital of Cologne, in Cologne, Germany. “Prior studies have demonstrated potential applications of GPT-4 across various stages of the patient journey in radiology: for instance, selecting the correct imaging exam and protocol based on a patient’s medical history, transforming free-text radiology reports into structured reports or automatically generating the impression section of a report.”

However, this is the first study to distinctively compare GPT-4 and human performance in error detection in radiology reports, assessing its capabilities against radiologists of varied experience levels in terms of accuracy, speed and cost-effectiveness, Dr. Gertz noted.

Dr. Gertz and colleagues set out to assess GPT-4’s effectiveness in identifying common errors in radiology reports, focusing on performance, time and cost-efficiency.

For the study, 200 radiology reports (X-rays and CT/MRI imaging) were gathered between June 2023 and December 2023 at a single institution. The researchers intentionally inserted 150 errors from five error categories (omission, insertion, spelling, side confusion and “other”) into 100 of the reports. Six radiologists (two senior radiologists, two attending physicians and two residents) and GPT-4 were tasked with detecting these errors.

Researchers found that GPT-4 had a detection rate of 82.7% (124 of 150). The error detection rates were 89.3% for senior radiologists (134 out of 150) and 80.0% for attending radiologists and radiology residents (120 out of 150), on average.

In the overall analysis, GPT-4 detected less errors compared with the best performing senior radiologist (82.7% vs 94.7%). However, there was no evidence of a difference in the percentage of average performance in error detection rate between GPT-4 and all the other radiologists.

GPT-4 required less processing time per radiology report than even the fastest human reader, and the use of GPT-4 resulted in lower mean correction cost per report than the most cost-efficient radiologist.

“This efficiency in detecting errors may hint at a future where AI can help optimize the workflow within radiology departments, ensuring that reports are both accurate and promptly available,” Dr. Gertz said, “thus enhancing the radiology department’s capacity to deliver timely and reliable diagnostics.”

Dr. Gertz notes that the study’s findings are significant for their potential to improve patient care by enhancing the accuracy of radiology reports through GPT-4 assisted proofreading. Demonstrating that GPT-4 can match the error detection performance of radiologists—while significantly reducing the time and cost associated with report correction—this research shows the potential benefits of integrating AI into radiology departments.

“The study addresses critical health care challenges such as the increasing demand for radiology services and the pressure to reduce operational costs,” he said. “Ultimately, our research provides a concrete example of how AI, specifically through applications like GPT-4, can revolutionize health care by boosting efficiency, minimizing errors and ensuring broader access to reliable, affordable diagnostic services—fundamental steps toward improving patient care outcomes.”

“Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy.” Collaborating with Dr. Gertz were Thomas Dratsch, M.D., Alexander Christian Bunck, M.D., Simon Lennartz, M.D., Andra-Iza Iuga, M.D., Martin Gunnar Hellmich, Ph.D., Thorsten Persigehl, M.D., Lenhard Pennig, M.D., Carsten Herbert Gietzen, M.D., Philipp Fervers, M.D., David Maintz, M.D., Robert Hahnfeldt, M.D., and Jonathan Kottlors, M.D.

*Radiology* is edited by Linda Moy, M.D., New York University, New York, N.Y., and owned and published by the Radiological Society of North America, Inc. (https://pubs.rsna.org/journal/radiology)

RSNA is an association of radiologists, radiation oncologists, medical physicists and related scientists promoting excellence in patient care and health care delivery through education, research and technologic innovation. The Society is based in Oak Brook, Illinois. (RSNA.org)

For patient-friendly information on radiology reports, visit *RadiologyInfo.org*.

Images (JPG, TIF):

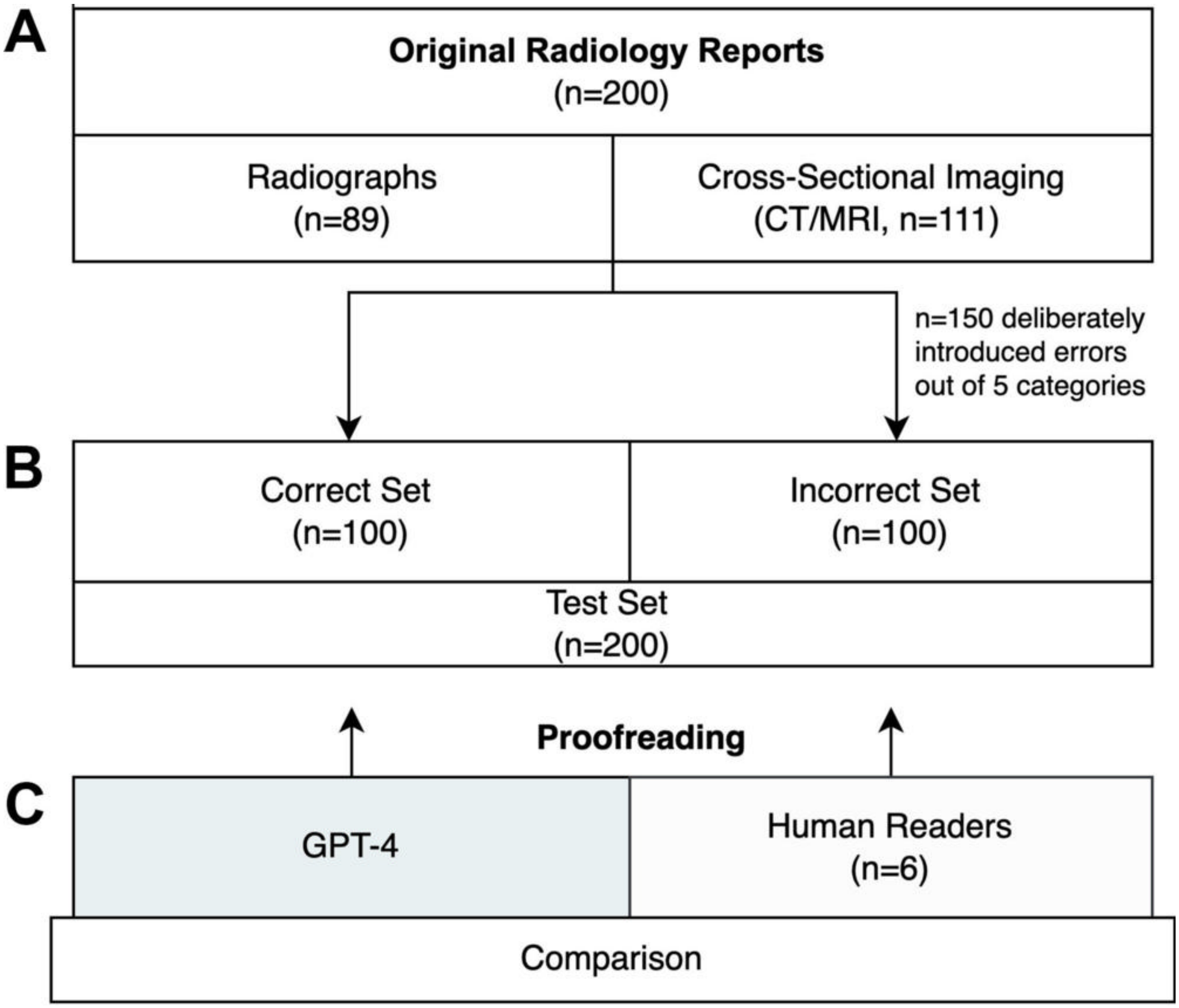


**Figure 1.** Study flowchart. **(A)** Initially, 200 original radiology reports from radiographs and cross-sectional imaging (CT and MRI) studies were selected. **(B)** These were then randomized into two sets: a correct set and an incorrect set, each containing 100 radiology reports. Within the incorrect set, 150 errors across five categories (omission, insertion, spelling errors, side confusion, other errors) were deliberately introduced, with a maximum of three errors per case. **(C)** GPT-4 and six radiologists were tasked with evaluating each radiology report to identify potential errors, allowing for a comparative analysis of their performance.

High-res (TIF) version

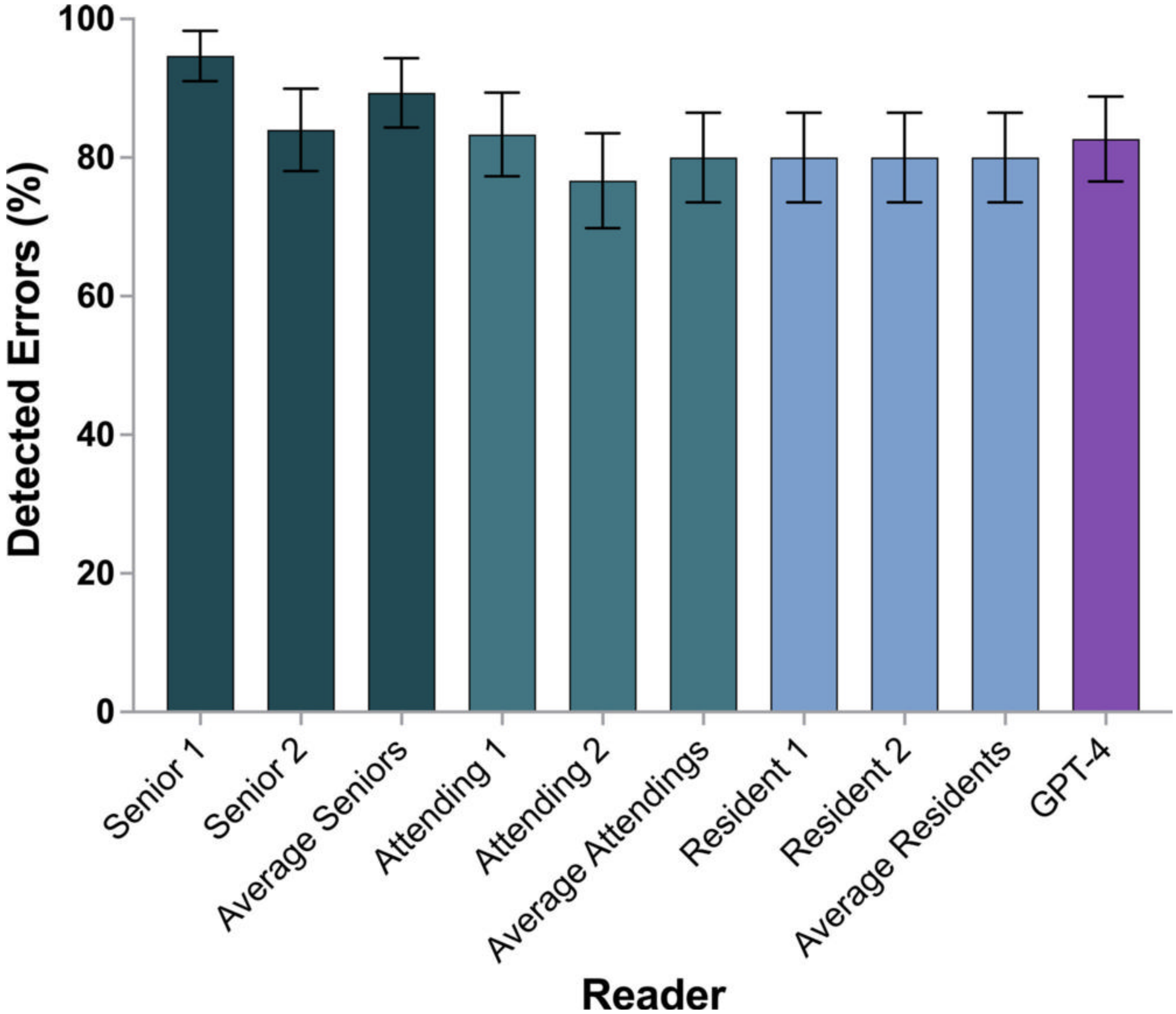
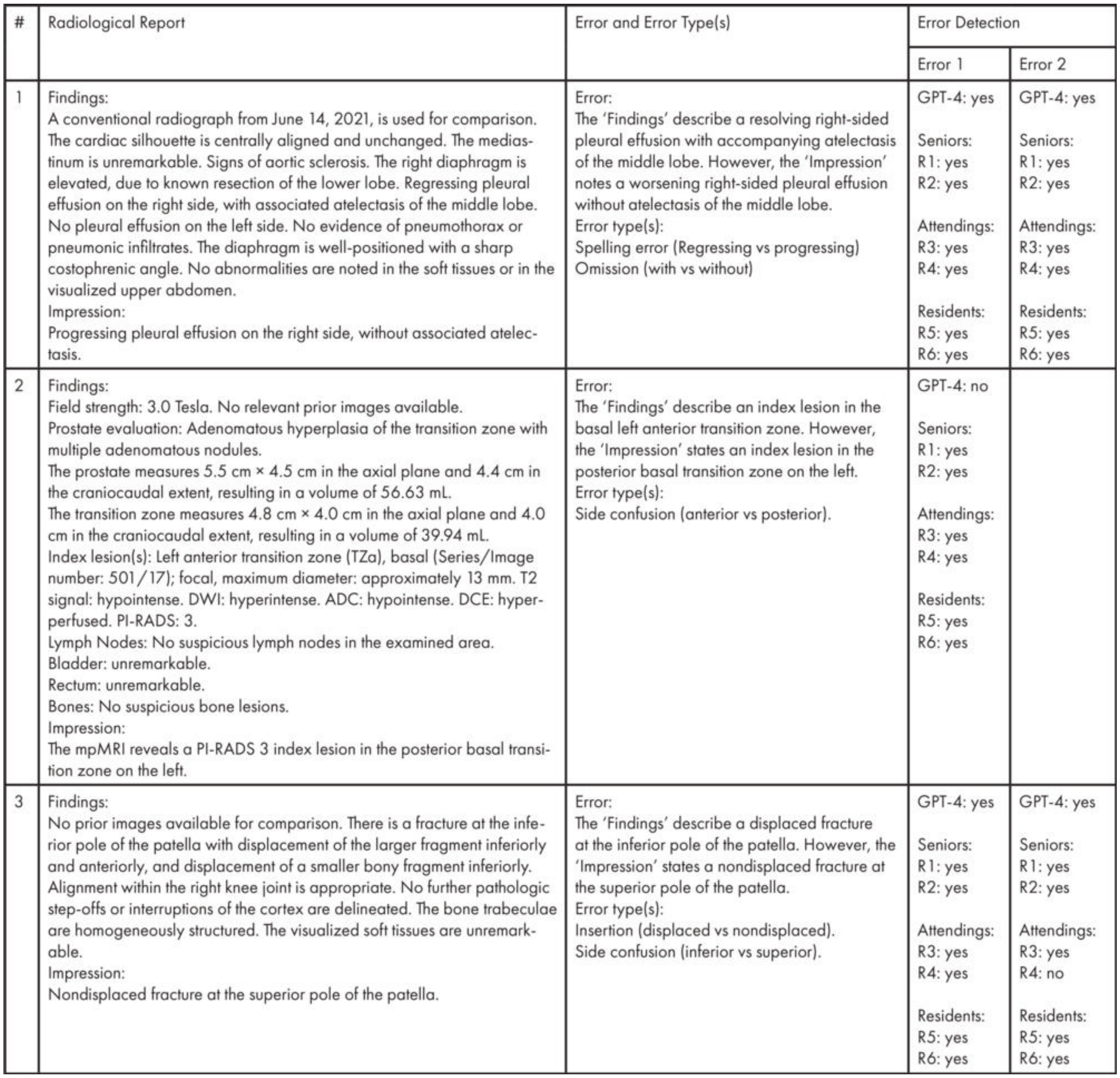


**Figure 2.** Bar graph shows the percentage of detected errors for GPT-4 and the radiologists. The error bars are 95% CIs.
High-res (TIF) version



| # | Radiological Report | Error and Error Type(s) | Error Detection | |
|---|---|---|---|---|
| | | | Error 1 | Error 2 |
| 1 | Findings:<br>A conventional radiograph from June 14, 2021, is used for comparison. The cardiac silhouette is centrally aligned and unchanged. The mediastinum is unremarkable. Signs of aortic sclerosis. The right diaphragm is elevated, due to known resection of the lower lobe. Regressing pleural effusion on the right side, with associated atelectasis of the middle lobe. No pleural effusion on the left side. No evidence of pneumothorax or pneumonic infiltrates. The diaphragm is well-positioned with a sharp costophrenic angle. No abnormalities are noted in the soft tissues or in the visualized upper abdomen.<br>Impression:<br>Progressing pleural effusion on the right side, without associated atelectasis. | Error:<br>The 'Findings' describe a resolving right-sided pleural effusion with accompanying atelectasis of the middle lobe. However, the 'Impression' notes a worsening right-sided pleural effusion without atelectasis of the middle lobe.<br>Error type(s):<br>Spelling error (Regressing vs progressing)<br>Omission (with vs without) | GPT-4: yes<br>Seniors:<br>R1: yes<br>R2: yes<br>Attendings:<br>R3: yes<br>R4: yes<br>Residents:<br>R5: yes<br>R6: yes | GPT-4: yes<br>Seniors:<br>R1: yes<br>R2: yes<br>Attendings:<br>R3: yes<br>R4: yes<br>Residents:<br>R5: yes<br>R6: yes |
| 2 | Findings:<br>Field strength: 3.0 Tesla. No relevant prior images available.<br>Prostate evaluation: Adenomatous hyperplasia of the transition zone with multiple adenomatous nodules.<br>The prostate measures 5.5 cm × 4.5 cm in the axial plane and 4.4 cm in the craniocaudal extent, resulting in a volume of 56.63 mL.<br>The transition zone measures 4.8 cm × 4.0 cm in the axial plane and 4.0 cm in the craniocaudal extent, resulting in a volume of 39.94 mL.<br>Index lesion(s): Left anterior transition zone (TZa), basal (Series/Image number: 501/17); focal, maximum diameter: approximately 13 mm. T2 signal: hypointense. DWI: hyperintense. ADC: hypointense. DCE: hyperperfused. PI-RADS: 3.<br>Lymph Nodes: No suspicious lymph nodes in the examined area.<br>Bladder: unremarkable.<br>Rectum: unremarkable.<br>Bones: No suspicious bone lesions.<br>Impression:<br>The mpMRI reveals a PI-RADS 3 index lesion in the posterior basal transition zone on the left. | Error:<br>The 'Findings' describe an index lesion in the basal left anterior transition zone. However, the 'Impression' states an index lesion in the posterior basal transition zone on the left.<br>Error type(s):<br>Side confusion (anterior vs posterior). | GPT-4: no<br>Seniors:<br>R1: yes<br>R2: yes<br>Attendings:<br>R3: yes<br>R4: yes<br>Residents:<br>R5: yes<br>R6: yes | |
| 3 | Findings:<br>No prior images available for comparison. There is a fracture at the inferior pole of the patella with displacement of the larger fragment inferiorly and anteriorly, and displacement of a smaller bony fragment inferiorly. Alignment within the right knee joint is appropriate. No further pathologic step-offs or interruptions of the cortex are delineated. The bone trabeculae are homogeneously structured. The visualized soft tissues are unremarkable.<br>Impression:<br>Nondisplaced fracture at the superior pole of the patella. | Error:<br>The 'Findings' describe a displaced fracture at the inferior pole of the patella. However, the 'Impression' states a nondisplaced fracture at the superior pole of the patella.<br>Error type(s):<br>Insertion (displaced vs nondisplaced).<br>Side confusion (inferior vs superior). | GPT-4: yes<br>Seniors:<br>R1: yes<br>R2: yes<br>Attendings:<br>R3: yes<br>R4: yes<br>Residents:<br>R5: yes<br>R6: yes | GPT-4: yes<br>Seniors:<br>R1: yes<br>R2: yes<br>Attendings:<br>R3: yes<br>R4: no<br>Residents:<br>R5: yes<br>R6: yes |

**Figure 3.** Comparative proofreading examples by GPT-4 and the human readers show incorrect radiology reports with the respective errors and error types and the corresponding proofreading results. ADC = apparent diffusion coefficient, DCE = dynamic contrast enhanced, DWI = diffusion-weighted imaging, mpMRI = multiparametric MRI, PI-RADS = Prostate Imaging Reporting and Data System, R1 = reader 1, R2 = reader 2, R3 = reader 3, R4 = reader 4.
High-res (TIF) version

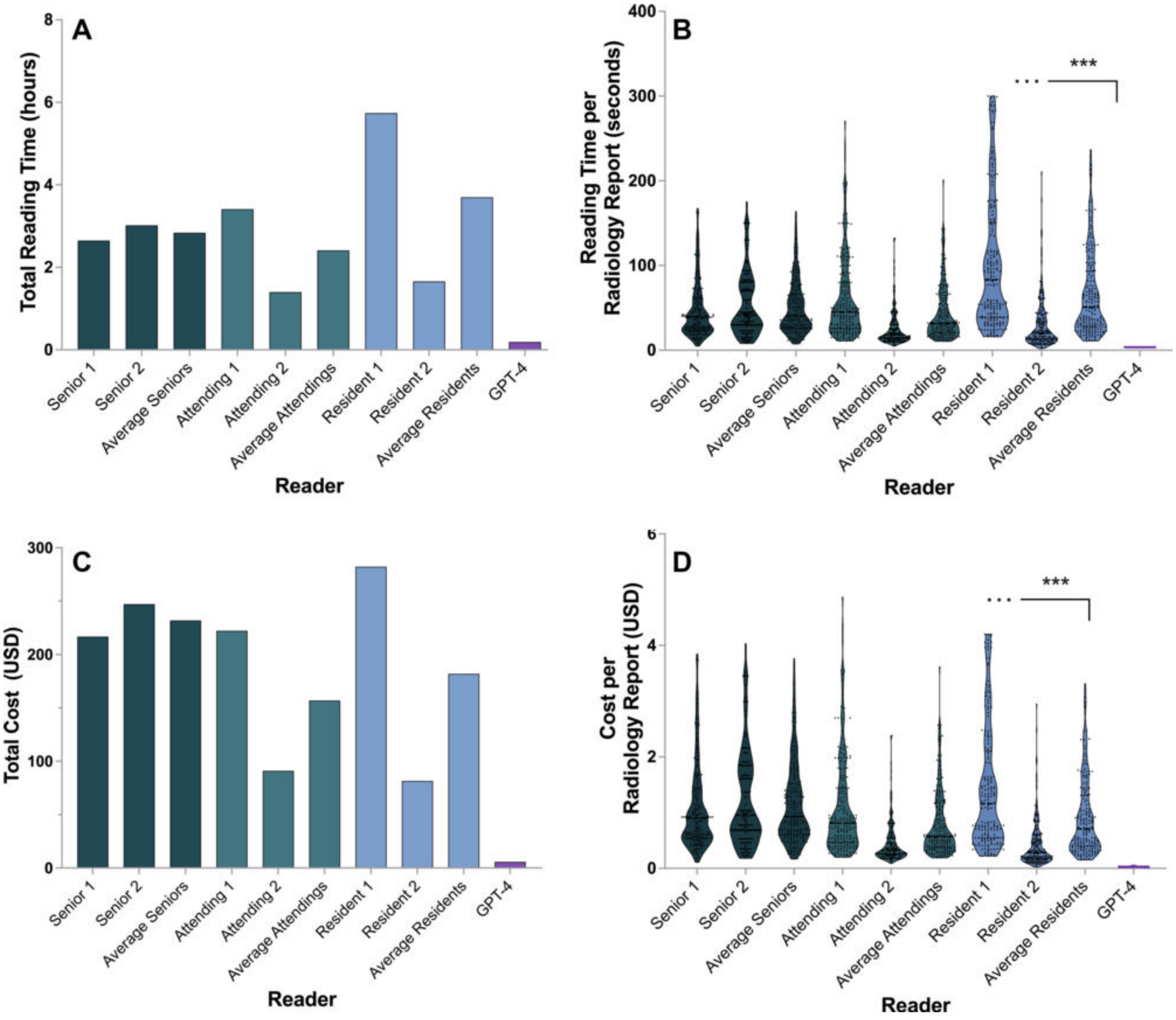


**Figure 4.** **(A)** Bar graph shows total reading time in seconds, **(B)** Violin plot shows reading time per radiology report in seconds, **(C)** bar graph shows total cost in U.S. dollars, and **(D)** violin plot shows cost per radiology report in U.S. dollars. Dashed lines are the medians and dotted lines are quartiles. *** *P* < .001.="">

High-res (TIF) version

Resources:

Editorial
Study abstract