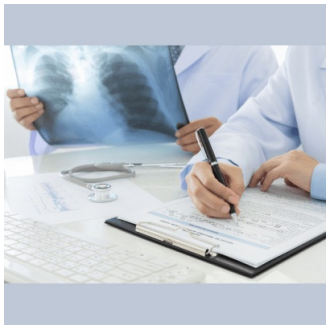
HealthManagement
Promoting Management and Leadership

Can GPT-4 Enhance Radiology Workflows by Proofreading Reports?



GPT-4, developed by OpenAI, has shown promising potential in addressing challenges faced in radiology, particularly in detecting errors in radiology reports. These reports are crucial for accurate diagnosis and treatment planning but are prone to errors due to various factors like high workload and unreliable speech recognition. Common errors include confusion in laterality and misregistration by speech recognition. GPT-4, as an autoregressive large language model, has been demonstrated to offer solutions in transforming free-text reports into structured formats, generating impression sections, and producing competent radiology reports. It has also been tested in radiology board examinations and shown capabilities in simplifying reports, extracting information, and providing screening recommendations. Utilising GPT-4 for proofreading radiology reports could reduce the workload of supervising radiologists and serve as an educational resource for residents. This study aims to evaluate GPT-4's performance in detecting common errors in radiology reports and estimate its potential in reducing time and cost.

Evaluating GPT-4 Performance in Detecting Radiology Report Errors

In this retrospective study, ethical approval was obtained, and informed consent was waived due to the study's retrospective nature. The study utilised 200 original radiology reports from both radiography (89 reports) and cross-sectional imaging (CT and MRI; 111 reports) obtained from University Hospital Cologne, Germany, covering a range of pathologic abnormalities. The reports were collected between June 2023 and December 2023 and were randomised into two sets: correct and incorrect, each containing 100 reports. Within the incorrect set, 150 errors were deliberately introduced by a radiology resident. Each report had a maximum of three errors, categorised into five types: omission, insertion, spelling errors, side confusion (e.g., right instead of left), and other errors (e.g., incorrect date entries, mistakes in units of measurement). These error categories were defined based on prior research to encompass the most common types of errors found in radiology reports. To establish a reference standard, only the errors intentionally inserted into the text were considered. Three readers with varying years of experience (5, 5, and 3 years) independently verified the reports to ensure accuracy. Any discrepancies were resolved through consensus reading involving all three readers. This rigorous process ensured the reliability of the dataset for evaluating the performance of GPT-4 in detecting errors and discrepancies in radiology reports.

Comparing GPT-4's Performance with Human Radiologists

In the overall analysis, GPT-4 detected fewer errors compared to the best-performing senior radiologist (82.7% vs. 94.7% detection rate, respectively). However, there was no significant difference in the average performance of error detection rate between GPT-4 and all other radiologists per report. In the subgroup analysis based on imaging modalities, GPT-4 showed comparable performance to radiologists in detecting errors in radiography reports, but detected fewer errors in CT and MRI reports compared to the best-performing senior radiologist. However, there was no significant difference in error detection rates between GPT-4 and other radiologists per report. Specifically, GPT-4's performance in detecting side confusion errors was worse than the best-performing radiologist, but there was no significant difference in error detection rates compared to other radiologists. Regarding other error categories, there was no significant difference in error detection rates between GPT-4 and radiologists. GPT-4 incorrectly labelled eight reports as containing errors when they did not, but there was no significant difference in incorrectly flagged reports compared to radiologists. In terms of reading time, GPT-4 significantly outperformed radiologists, with a mean reading time per report much faster than the fastest radiologist. In terms of cost, the estimated average cost for proofreading by human readers was substantially higher compared to using GPT-4. GPT-4 demonstrated a significantly lower mean cost per report compared to the radiologist with the lowest mean cost per report.

This study aimed to assess the effectiveness of GPT-4, a large language model, in detecting errors and discrepancies in radiology reports and to evaluate its potential to reduce time and cost. Results showed that GPT-4's performance in proofreading radiology reports was comparable to human readers across different levels of experience. Although GPT-4 detected fewer errors than the most experienced radiologist, it outperformed the fastest human reader in terms of reading time. Additionally, using GPT-4 for proofreading was found to be more cost-effective compared to employing human readers. However, the study had limitations, including its experimental nature, potential biases, and the need for human validation of GPT-4's detections. Despite these limitations, the findings suggest that GPT-4 could be a valuable tool in radiology workflows, but human supervision remains essential. Future research should explore ways to integrate GPT-4 into clinical settings while addressing legal and privacy concerns and optimising its performance for specific radiology tasks.

Source: **RSNA Radiology**

Image Credit: **iStock**

Published on : Thu, 18 Apr 2024