Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–9

DOI: 10.1177/20552076251326019
journals.sagepub.com/home/dhj

Sage

An assessment of ChatGPT in error detection for thyroid ultrasound reports: A comparative study with ultrasound physicians

Zhirong Xu[1,*], Jiayi Ye[2,*], Weiwen Luo[3], Lina Han[1], Hui Yin[1], Yanru Li[1], Qichen Su[1], Shanshan Su[1], Guorong Lyu[1] and Shaohui Li[1]

Abstract

Background: This study evaluates the performance of GPT-4o in detecting errors in ACR TIRADS ultrasound reports and its potential to reduce report generation time.

Methods: A retrospective analysis of 200 thyroid ultrasound reports from the Second Affiliated Hospital of Fujian Medical University was conducted, with reports categorized as correct or containing up to three errors. GPT-4o's performance was compared with ultrasound physicians of varying experience levels in error detection and processing time.

Results: GPT-4o detected 90.0% (180/200) of errors, slightly less than the best-performing senior ultrasound physician's 93.0% (186/200) with no significant difference ($p = 0.281$). GPT-4o's error detection rate was comparable to that of ultrasound physicians overall ($p = 0.098$ to 0.866). It outperformed Resident 2 in diagnostic errors (87% vs. 69%). Reader agreement was low (Cohen's kappa = 0 to 0.31). GPT-4o reviewed reports significantly faster than all ultrasound physicians (0.79 vs. 1.8 to 3.1 h, $p < 0.001$), making it a reliable and efficient tool for error detection in medical imaging.

Conclusions: GPT-4o is comparable to experienced ultrasound physicians in error detection and significantly improves report processing efficiency, offering a valuable tool for enhancing diagnostic accuracy and aiding junior residents.

Keywords

Artificial intelligence, ChatGPT, ultrasonic diagnosis, diagnostic errors, ACR TIRADS

Submission date: 31 August 2024; Acceptance date: 18 February 2025

Introduction

Thyroid ultrasound is essential for evaluating thyroid nodules, with the ACR Thyroid Imaging Reporting and Data System (ACR TIRADS) serving as a widely adopted framework for standardized assessment. However, implementing the ACR TIRADS classification can be complex, particularly for less experienced ultrasound physicians, as it demands meticulous evaluation of multiple features to derive a final score.[1,2] Although automated reporting systems have streamlined the process by assigning feature scores and identifying basic discrepancies, they remain limited in managing the inherent

[1]Department of Ultrasound, Second Affiliated Hospital of Fujian Medical University, Quanzhou, China
[2]Department of Nuclear Medicine, Second Affiliated Hospital of Fujian Medical University, Quanzhou, China
[3]Department of Medical Ultrasound, Zhangzhou Municipal Hospital of Fujian Province and Zhangzhou Affiliated Hospital of Fujian Medical University, Zhangzhou, China

[*]These authors contributed equally to this work as first authors.

Corresponding author:
Shanshan Su, Department of Ultrasound, Second Affiliated Hospital of Fujian Medical University, Quanzhou, Fujian, China.
Email: susan@fjmu.edu.cn

complexity of the ACR TIRADS classification. The system's nuanced categorization and detailed scoring requirements can result in errors, especially in high-volume or time-constrained scenarios.

The workload of ultrasound physicians continues to be a critical factor influencing reporting accuracy. Even with automated systems, ultrasound physicians frequently work in high-pressure environments with limited time for assessments, heightening the risk of errors. Research indicates a strong correlation between fatigue, extended working hours, and increased error rates in radiological reporting.[3,4] Common errors, such as positional and descriptive inaccuracies,[5] often go unnoticed by conventional automated systems, which primarily focus on basic error detection. Standard proofreading tools are effective for identifying typographical errors but cannot resolve complex issues like misclassification of nodule features or scoring inconsistencies.

In many clinical settings, particularly in regions where automated or structured reporting templates are not fully adopted, free-text reporting remains the prevailing practice. Free-text reporting allows ultrasound physicians to document findings in an unstructured format, offering flexibility for individualized observations. However, this approach has significant limitations, including heightened error risk, variability in report quality, and increased workload compared to structured reporting systems. The continued reliance on free-text reporting, particularly in high-volume settings, underscores the pressing need for tools capable of enhancing error detection and improving reporting accuracy.

Advances in artificial intelligence, particularly through models such as GPT-4, demonstrate considerable potential to enhance the accuracy and efficiency of medical report generation.[6] GPT-4, with its capability to process natural language and interpret complex medical terminology, has been successfully utilized in radiology for error detection and improving reporting consistency.[7] Although prior studies have examined GPT's role in general medical error detection, limited research has focused on how GPT-4o addresses common errors in Chinese-language thyroid ultrasound reports.

This study seeks to bridge this gap by assessing GPT-4o's performance in detecting errors in ACR TIRADS thyroid ultrasound reports, with a focus on reducing errors and report generation time. Moreover, this study emphasizes GPT-4o's potential to adapt to diverse clinical documentation practices, including free-text reporting, addressing the varied requirements of global healthcare settings.

Materials and methods

Study design and data acquisition

This study performed a retrospective analysis. It included 200 original thyroid ultrasound reports from the Second Affiliated Hospital of Fujian Medical University, collected between February 2023 and February 2024, that were randomly ordered. All reports analyzed in this study were written in free-text format, following the ACR TIRADS classification guidelines. The exclusion criteria were as follows: (a) ultrasound examinations that did not involve the evaluation of thyroid nodules; (b) ultrasound reports lacking key image feature descriptions or diagnostic conclusions; (c) reports in which the ACR TIRADS classification was not used. The data were randomly divided into two groups: correct and incorrect, each containing 100 reports. In the incorrect group (n = 100), 200 errors were intentionally introduced, with a maximum of three errors per report (Figure 1). Specifically, 40 reports had one error each, 20 reports had two errors each, and 40 reports had three errors each. Each report was assigned only one type of errors. Errors were introduced by a separate team (Jiayi Ye, Lina Han and Hui Yin) to ensure objectivity in subsequent evaluations. All reports were de-identified prior to processing by GPT-4o to ensure compliance with ethical and privacy standards.

Based on previous studies,[5] we defined five categories of errors encompassing the most common types of errors in ultrasound reports: (a) Omission: The omission of relevant words or expressions, including the deletion of words, units, values, or punctuation. (b) Insertion: The unintentional insertion of incorrect words or expressions, including inappropriate or incorrect word substitutions. (c) Spelling error: Results from typing errors or inaccurate text selection by the ultrasound physician during manual text processing. (d) Positioning error: The incorrect identification of anatomical locations. (e) Diagnostic error: Inconsistencies between the ACR TIRADS classification in the "Findings" and "Impression" sections.

This study was approved by the Ethics Committee of the Second Affiliated Hospital of Fujian Medical University (2022 Ethical Review No. 332), and the requirement for written informed consent was waived owing to the retrospective nature of the study.

Report writing time calculation

A team of ultrasound physicians with varying levels of clinical experience evaluated ultrasound imaging reports for potential errors. The team included two seniors (Qichen Su and Shaohui Li, with 17 and 16 years of experience, respectively); two attendings (Shanshan Su and Yanru Li, with 11 and 9 years of experience, respectively); and two residents (Zhirong Xu and Weiwen Luo, with 7 and 2 years of experience, respectively). The time taken to evaluate each report was recorded using a stopwatch.

Using the "Temporary Chat" function of the GPT-4o (Version 13 May 2024) online interface, attach the ACR TIRADS guide and enter the following prompt in Chinese: "You are a professional ultrasound physician, and you will provide me with accurate information using the ACR TIRADS guide I sent you." Next, submit a thyroid ultrasound report written according to the ACR TIRADS guidelines, divided into "findings" and

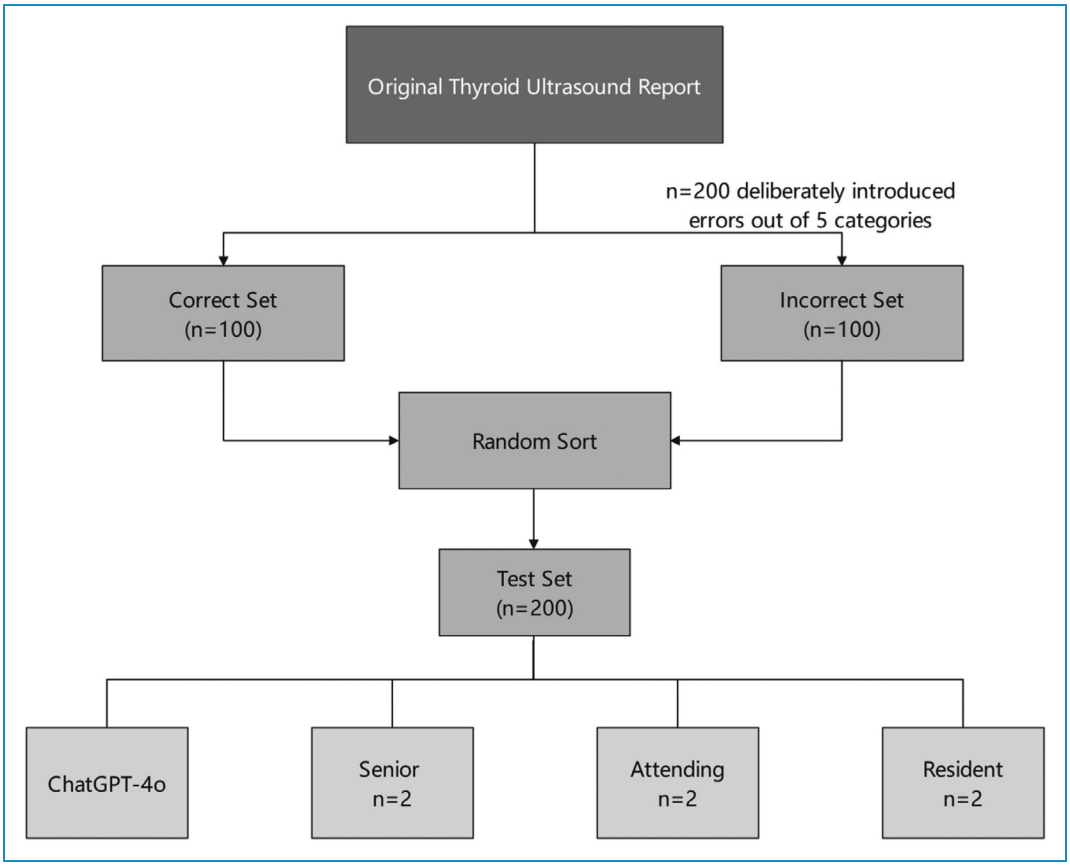

Figure 1. Study flowchart.

"impressions." Please evaluate the report for errors (including spelling, units of measure, punctuation, and consistency between "findings" and "impressions") according to the guide. If there are any errors, highlight them. Then enter the Chinese version of the thyroid ultrasound report, and record the results and time of each report.

The original reports were written in Chinese and input into GPT-4o in their native language. Supplementary Materials include translated English examples for illustrative purposes.

Statistical analysis

All analyses were performed using IBM SPSS Statistics Version 25.0 and RStudio Version 4.3.3. The number of reports with correctly detected errors and the time taken to correct each thyroid ultrasound report were used as outcome indicators. The number of errors detected by GPT-4o was compared with those detected by ultrasound physicians using Fisher's exact test, and the 95% confidence interval (CI) was calculated using Wilson's method.[8] The paired sample t-test was used to compare the average time taken by GPT-4o to correct ultrasound reports with that taken by the ultrasound physicians. A two-sided p-value of <0.05 indicated a statistically significant difference. Agreement among evaluators was assessed using Cohen's kappa, defined as follows: 0.01 to 0.20: none to slight agreement; 0.21 to 0.40: fair agreement; 0.41 to 0.60: moderate agreement; 0.61 to 0.80: substantial agreement; 0.81 to 1.00: almost perfect agreement.[9]

Results

Performance in detecting errors

A comparative example of GPT-4o and the ultrasound physicians shows erroneous ultrasound reports, their respective errors and error types, and the corresponding proofreading results (Table 1). Among the 200 errors, GPT-4o detected fewer errors compared to the best-performing senior ultrasound physician (detection rate: 90.0% [180 of 200; 95% CI: 85.0%, 93.8%] vs. 93.0% [186 of 200; 95% CI: 88.5%, 96.1%], $p = 0.281$). The error detection rate of GPT-4o was between that of senior ultrasound physicians and attending ultrasound physicians; however, there was no significant difference in error detection rates between GPT-4o and all ultrasound physicians (p-value range: 0.098 to 0.866) (Table 1). This suggests that GPT-4o is a reliable tool for error detection in medical imaging, performing on par comparably to human readers.

Across different error types, there was no evidence that GPT-4o was superior in detecting errors compared to the best-performing ultrasound physicians (p-value range, 0.339 to 0.99). GPT-4o demonstrated comparable performance to human readers in detecting most types of errors, with statistically significant differences observed only in the category of diagnostic errors compared to Resident 2 (detection rate: 87% [39 of 45; 95% CI: 73%, 95%] vs 69% [32 of 45; 95% CI: 53%, 82%]). These findings suggest that GPT-4o has the potential to support and enhance human ultrasound physicians in error detection (Figure 2; Table 2). For a more detailed example of a report, please refer to Supplementary Table.

Of the 200 ultrasound reports, GPT-4o incorrectly labeled six accurate reports as incorrect. However, there was no evidence of a difference between GPT-4o and the ultrasound physicians in the frequency of incorrectly labeled ultrasound reports ($p > 0.99$) (Table 3).

Table 1. Comparison of error detection rates between GPT-4o and the ultrasound physicians.

Reader	Detection rate	CI lower	CI upper	p-value
Senior 1	90.5% (181/200)	85.6%	94.2%	0.866
Senior 2	93.0% (186/200)	88.5%	96.1%	0.281
Attending 1	91.0% (182/200)	86.1%	94.6%	0.733
Attending 2	88.0% (176/200)	82.7%	92.2%	0.522
Resident 1	87.5% (175/200)	82.1%	91.7%	0.428
Resident 2	84.5% (169/200)	78.7%	89.2%	0.098
GPT-4o	90.0% (180/200)	85.0%	93.8%	

Note: Data in parentheses are numerators/denominators.
The number of correctly detected errors by GPT-4o was compared with the ultrasound physicians by using Fisher's exact test. CI: confidence interval.

Agreement among readers

The reader agreement between GPT-4o and ultrasound physicians, as well as among ultrasound physicians, ranged from none to fair (Cohen's kappa = 0 to 0.31), suggesting that GPT-4o reports error detection without any specific pattern (Figure 3).

Reading time

The total reading time for all 200 reports by GPT-4o was 0.79 h; the fastest ultrasound physician read 200 reports in 1.8 h, while the slowest ultrasound physician needed 3.1 h. GPT-4o required less time than all ultrasound physicians ($p < 0.001$) (Figure 4). Additionally, GPT-4o's average reading time per thyroid ultrasound report was faster than that of the fastest ultrasound physician ($p < 0.001$) (Figure 5).

Discussion

This study emphasizes that the thyroid ultrasound reports analyzed were in a free-text format. Although structured

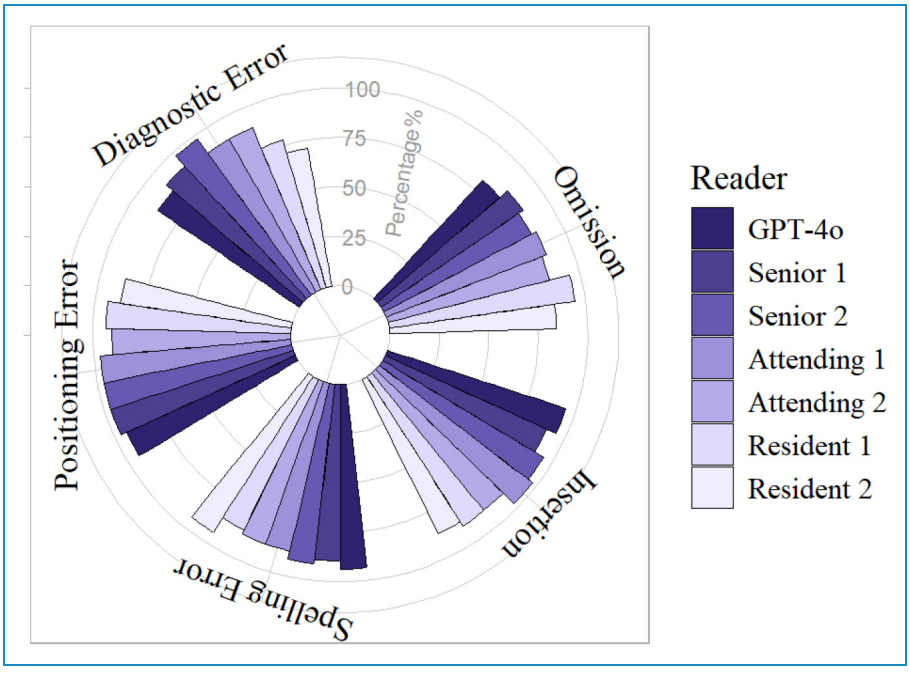

Figure 2. Radial column chart shows comparison of detection rates for different error types in ultrasound reports.

Table 2. Comparison of detection rates for different error types in ultrasound reports.

	Omission		Insertion		Spelling error		Positioning error		Diagnostic error	
Reader	Detection rate(%)	*p*-value	Detection rate(%)	*p*-value	Detection rate(%)	*p*-value	Detection rate(%)	*p*-value	Detection rate(%)	*p*-value
Senior 1	87(66,92) [33/38]	0.223	90(76,97) [35/39]	>0.99	89(77,96) [42/47]	0.292	97(83,100) [30/31]	>0.99	91(79,98) [41/45]	0.448
Senior 2	84(66,92) [32/38]	0.302	95(83,99) [37/39]	0.101	91(80,98) [43/47]	>0.99	97(89,100) [30/31]	>0.99	98(92,100) [44/45]	>0.99
Attending 1	87(66,92) [33/38]	0.223	97(87,100) [38/39]	0.051	87(74,95) [41/47]	0.343	97(89,100) [30/31]	>0.99	89(76,96) [40/45]	>0.99
Attending 2	84(66,92) [32/38]	>0.99	90(76,97) [35/39]	0.197	87(74,95) [41/47]	>0.99	90(74,98) [28/31]	0.187	89(76,96) [40/45]	>0.99
Resident 1	95(66,92) [36/38]	0.339	90(76,97) [35/39]	>0.99	85(72,94) [40/47]	>0.99	94(79,99) [29/31]	>0.99	78(63,89) [35/45]	0.113
Resident 2	84(66,92) [32/38]	0.063	87(73,96) [34/39]	0.243	94(82,99) [44/47]	>0.99	87(70,96) [27/31]	0.245	69(53,82) [32/45]	0.005
GPT-4o	82(66,92) [31/38]		95(83,99) [37/39]		94(82,99) [44/47]		94(79,99) [29/31]		87(73,95) [39/45]	

Note: Data in parentheses are 95% CIs; data in brackets are numerators/denominators.
The number of correctly detected errors by GPT-4o was compared with the sonographers by using Fisher's exact test.

Table 3. Comparison of correctly reported mislabeling errors between GPT-4o and ultrasound physicians.

Reader	Rate of correctly reported mislabeling errors(%)	*p*-value
Senior 1	1 (2/200)	>0.99
Senior 2	0.5(1/200)	>0.99
Attending 1	0.5 (1/200)	>0.99
Attending 2	1 (2/200)	>0.99
Resident 1	1 (2/200)	>0.99
Resident 2	2 (4/200)	>0.99
GPT-4o	3 (6/200)	

Note: Data in parentheses are numerators/denominators.
The number of correctly detected errors by GPT-4o was compared with the sonographers by using Fisher's exact test.

reporting systems and automated templates are becoming more prevalent in modern clinical practice, free-text reporting remains widely utilized in certain regions and institutions.[10] While this format provides flexibility in documentation, it is associated with higher error rates and greater workloads compared to structured formats. The use of free-text reports in this study reflects real-world clinical practices in settings that have not transitioned to standardized templates. By focusing on free-text reporting, this research addresses a critical need in such settings, demonstrating GPT-4o's potential to significantly enhance reporting accuracy and efficiency in clinical workflows.

The implementation of ACR TI-RADS in thyroid ultrasound reporting has significantly improved diagnostic accuracy and consistency.[11] Ultrasonographic features are assigned specific scores to assess nodules, where higher scores indicate greater suspicion. However, the complexity of feature assignment and the steep learning curve associated with TI-RADS classification pose challenges to its implementation in clinical practice. Reporting error rates in routine imaging practice range from 3% to 5% and increase substantially in high-intensity work environments.[12,13] The integration of large language models, such as GPT-4o,[14] provides a promising solution to address these challenges. GPT-4o can detect and correct grammatical and spelling errors while interpreting semantics and contextual nuances, enabling more precise detection and analysis.[15]

GPT-4o demonstrated a comparable capacity for error detection but occasionally overlooked specific issues. For omission errors, GPT-4o reliably identified

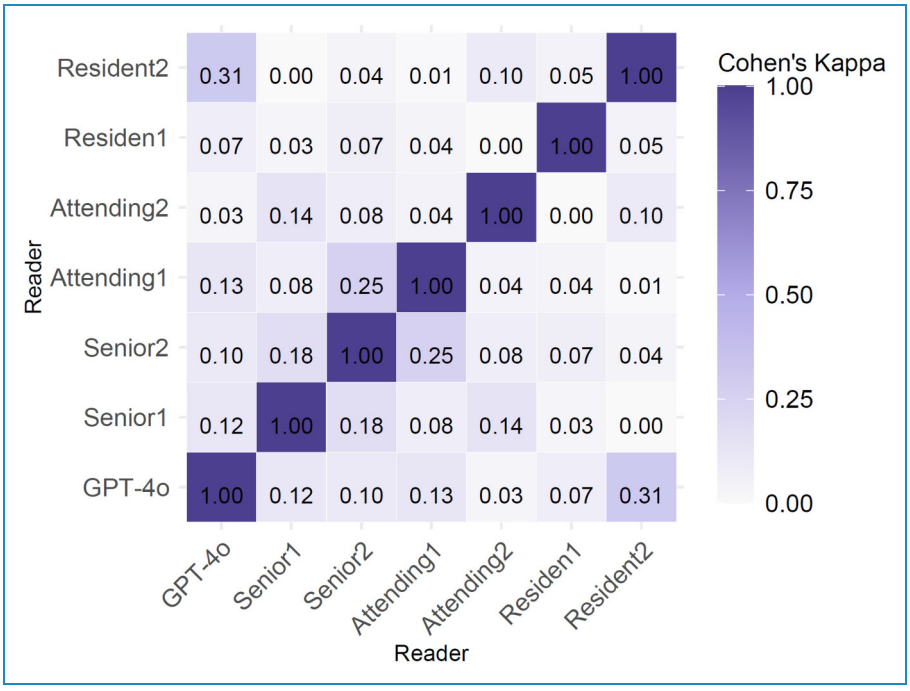

Figure 3. Heatmap of reader agreement between GPT-4o and the ultrasound physicians.
Note: Data are Cohen's kappa values (0.01–0.20, none to slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement).

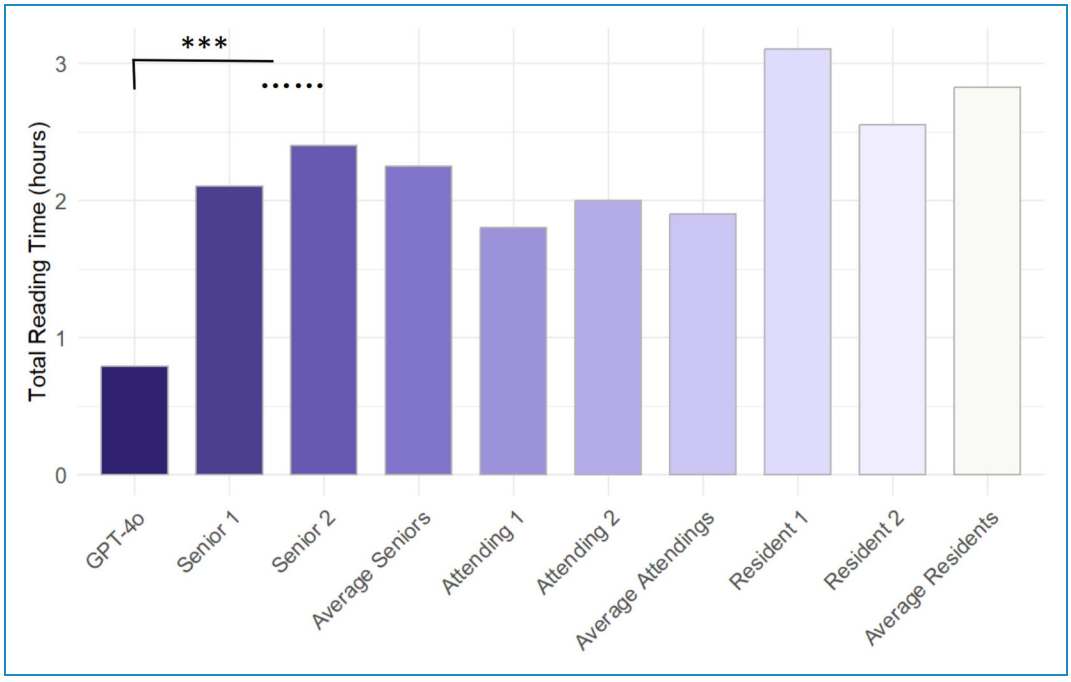

Figure 4. Bar graph shows total reading time in seconds.
Note: *** $p < 0.001$.

missing values or units for most nodules, although a small subset of errors remained undetected. This indicates occasional inconsistencies in identifying missing information. In one case, the size of a nodule was correctly reported in millimeters (mm), but GPT-4o incorrectly converted the unit to centimeters (cm). This

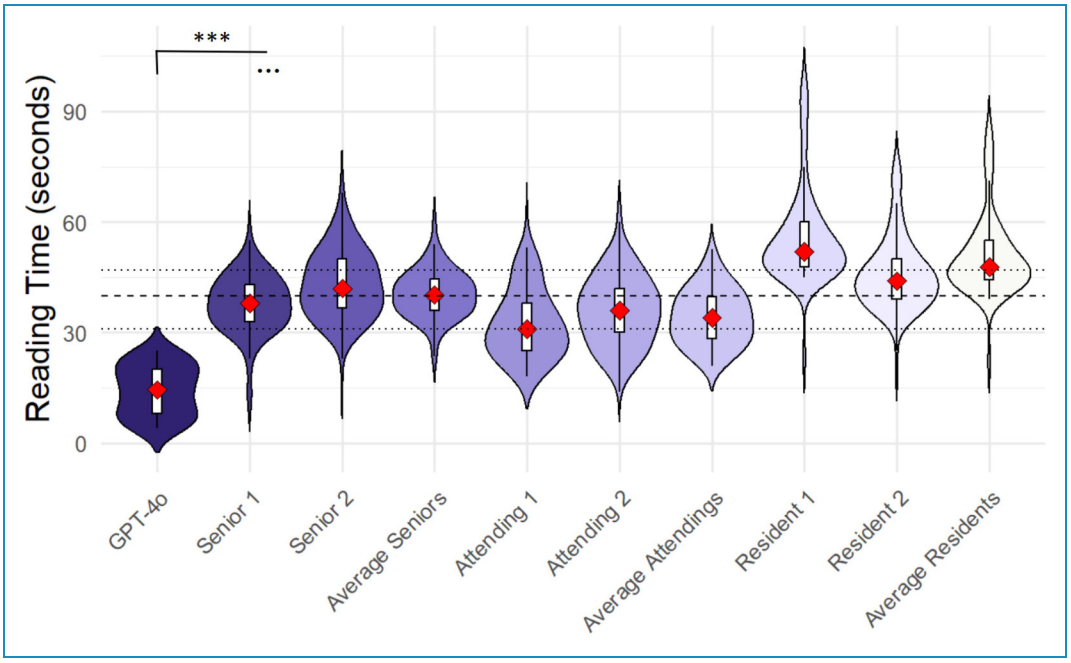

Figure 5. Violin plot shows reading time per radiology report in seconds.
Note: *** $p < 0.001$.

highlights GPT-4o's occasional challenges with unit conversions, which can lead to inaccuracies in specific contexts.

In diagnostic errors, GPT-4o misclassified a mixed cystic and solid nodule in the right thyroid lobe as "TR 2" instead of "TR 3," based on the "Findings" section. This misclassification underscores the need for further refinement to ensure the model's diagnostic accuracy adheres to clinical guidelines. Additionally, GPT-4o failed to identify an inconsistency in one report (Supplementary Table, Ultrasound Report 4), where cystic nodules described in the middle right thyroid lobe in the "Findings" section were omitted from the "Impression" section. This highlights the need for improvements to ensure coherence between sections of the report.

In clinical practice, GPT-4o could be integrated immediately after dictation or the initial drafting of reports. This workflow allows for the identification and correction of errors before finalizing reports. By serving as an error-checking tool within reporting systems, GPT-4o offers immediate feedback to ultrasound physicians, enhancing report accuracy while maintaining clinical efficiency.

Our findings show that GPT-4o achieves an error detection rate comparable to both primary and advanced ultrasound physicians. Senior ultrasound physicians achieved an error detection rate of 93.0%, while GPT-4o achieved a rate of 90.0%. Analysis shows that GPT-4o performs on par with experienced ultrasound physicians in detecting errors in ultrasound reports ($p > 0.05$). In the category of diagnostic errors, GPT-4o's detection rate exceeded that of Resident 2 ($p = 0.005$). Furthermore, GPT-4o did not differ significantly from ultrasound physicians in detecting most error types, including omissions, insertions, spelling errors, positional errors, and diagnostic errors. These findings suggest that GPT-4o has the potential to support and enhance ultrasound physicians in error detection and may be particularly effective in identifying certain critical error categories, especially for less experienced physicians.

In addition, GPT-4o demonstrates considerable advantages in processing time for reports. It processed 200 reports in just 0.79 h, whereas the fastest physician required 1.8 h ($p < 0.001$). This substantial time savings indicates that GPT-4o can greatly improve reporting efficiency, saving physicians valuable time and reducing error rates associated with fatigue.

This study highlights several areas for further research and improvement. Future studies should simulate real-world, high-pressure environments to evaluate GPT-4o's performance under such conditions. One notable limitation of this study is the relatively small sample size of 200 reports, which may affect the generalizability of the findings. Large-scale studies are essential to validate these results and provide a more robust evaluation of GPT-4o's capabilities. Additionally, this study primarily focused on textual errors in ultrasound reports, without addressing the consistency between ultrasound images and corresponding reports. Future research could build upon this by examining image-report consistency, thereby offering a more comprehensive assessment of GPT-4o's clinical utility.

Although free-text reporting is becoming less prevalent in some advanced healthcare systems, it remains widely

used in many regions. Consequently, future research should explore GPT-4o's performance within structured reporting systems and its potential integration into automated workflows. Given that the reports in this study were written in Chinese, GPT-4o's performance may have been influenced by its proficiency in processing medical terminology in this language. This limitation highlights the critical need for optimizing AI models for multilingual environments to ensure consistent performance across diverse clinical settings.

A collaborative approach between ultrasound physicians and AI tools is encouraged. GPT-4o should act as a supplementary tool, providing real-time feedback and error detection for validation by physicians. This approach could leverage AI's efficiency alongside human expertise to improve both accuracy and workflow efficiency. Future research should explore human-AI collaboration strategies, including the development of user-friendly interfaces and specialized training programs. While GPT-4o was chosen for its advanced capabilities, local AI models like GPT-2 might offer better privacy protections. Future research should evaluate the feasibility of using local AI models for similar applications.[16]

Finally, ethical and legal considerations, such as accountability for errors and transparency in AI decision-making, remain paramount. Further research is needed to ensure the safe and reliable use of AI tools in healthcare.[17–19]

Conclusion

The error detection rate of GPT-4o in ultrasound reports is comparable to that of ultrasound physicians, offering the potential for significant time savings. However, legal and privacy considerations, along with the challenges GPT-4o encounters in interpreting specialized medical terminology, highlight the necessity of continued human oversight in the report generation process.

Acknowledgements: The authors disclose the use of the following AI model in the writing of this manuscript: GPT-4o (OpenAI) was used to check spelling and grammar.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Institutional Review Board approval was obtained.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Fujian Province Young and Middle-aged Teacher Education Research Program of China (Grant No. JAT210112).

ORCID iDs: Zhirong Xu https://orcid.org/0000-0002-2928-468X
Jiayi Ye https://orcid.org/0009-0009-4003-1796
Shanshan Su https://orcid.org/0000-0003-3850-3968

Supplemental material: Supplemental material for this article is available online.

References

1. Zhou J, Yin L, Wei X, et al. 2020 Chinese guidelines for ultrasound malignancy risk stratification of thyroid nodules: the C-TIRADS. *Endocrine* 2020; 70: 256–279.
2. Tessler FN, Middleton WD, Grant EG, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *J Am Coll Radiol* 2017; 14: 587–595.
3. Berlin L. Radiologic errors and malpractice: a blurry distinction. *Am J Roentgenol* 2007; 189: 517–522.
4. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging* 2017; 8: 171–182.
5. Vosshenrich J, Nesic I, Cyriac J, et al. Revealing the most common reporting errors through data mining of the report proofreading process. *Eur Radiol* 2021; 31: 2115–2125.
6. Miah MSU, Kabir MM, Sarwar TB, et al. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Sci Rep* 2024; 14: 9603.
7. Gertz RJ, Dratsch T and Bunck AC. Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy. *Radiology* 2024; 311: e232714.
8. Wallis S. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *J Quant Linguist* 2013; 20: 178–208.
9. Kvalseth TO. A coefficient of agreement for nominal scales: an asymmetric version of Kappa. *Educ Psychol Meas* 1991; 51: 95–101.
10. Ernst BP, Dörsching C, Bozzato A, et al. Structured reporting of head and neck sonography achieves substantial interrater reliability. *Ultrasound Int Open* 2023; 09: E26–E32.
11. Kang YJ, Ahn HS, Stybayeva G, et al. Comparison of diagnostic performance of two ultrasound risk stratification systems for thyroid nodules: a systematic review and meta-analysis. *Radiol Med* 2023; 128: 1407–1414.
12. Onder O, Yarasir Y, Azizova A, et al. Errors, discrepancies and underlying bias in radiology with case examples: a pictorial review. *Insights Imaging* 2021; 12: 51.
13. Lamoureux C, Hanna TN, Callaway E, et al. Radiologist age and diagnostic errors. *Emerg Radiol* 2023; 30: 577–587.
14. Siepmann R, Huppertz M, Rastkhiz A, et al. The virtual reference radiologist: comprehensive AI assistance for clinical image reading and interpretation. *Eur Radiol* 2024; 34: 6652–6666.
15. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology* 2023; 308: e231362.
16. Zhang S and Song J. A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model. *Sci Rep* 2024; 14: 17118.

17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25: 44–56.
18. Price WN 2nd and Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019; 25: 37–43.
19. Dave T, Athaluri SA and Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023; 6: 1169595.