



OPEN ACCESS

EDITED BY

Rashid Ibrahim Mehmood,
Islamic University of Madinah, Saudi Arabia

REVIEWED BY

Cihan Atar,
Osmaniye State Hospital, Türkiye
Esat Kaba,
Recep Tayyip Erdoğan University, Türkiye

*CORRESPONDENCE

Luzeng Chen
✉ chenluzeng@126.com

RECEIVED 16 October 2024

ACCEPTED 10 February 2025

PUBLISHED 05 March 2025

CITATION

Wei B, Zhang X, Shao Y, Sun X and
Chen L (2025) Comparison of the accuracy of
GPT-4 and resident physicians in
differentiating benign and malignant thyroid
nodules.
Front. Artif. Intell. 8:1512438.
doi: 10.3389/frai.2025.1512438

COPYRIGHT

© 2025 Wei, Zhang, Shao, Sun and Chen. This
is an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited,
in accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Comparison of the accuracy of GPT-4 and resident physicians in differentiating benign and malignant thyroid nodules

Boxiong Wei, Xiumei Zhang, Yuhong Shao, Xiuming Sun and
Luzeng Chen*

Department of Ultrasound, Peking University First Hospital, Beijing, China

Objective: To assess the diagnostic performance of the GPT-4 model in comparison to resident physicians in distinguishing between benign and malignant thyroid nodules using ultrasound images.

Methods: This study analyzed 1,145 ultrasound images, including 632 malignant and 513 benign nodules. Both the GPT-4 model and two resident physicians independently classified the nodules using ultrasound images. The diagnostic accuracy of the resident physicians was determined by calculating the average of the individual accuracy rates of the two physicians and this was compared with the performance of the GPT-4 model.

Results: The GPT-4 model correctly identified 367 out of 632 malignant nodules (58.07%) and 343 out of 513 benign nodules (66.86%). Resident physicians identified 467 malignant (73.89%) and 383 benign nodules (74.66%). There was a statistically significant difference in the classification of malignant nodules ($p < 0.001$) and benign nodules ($p = 0.048$) between the GPT-4 model and residents. GPT-4 performed better for larger nodules (>1 cm) at 65.38%, compared to 53.77% for smaller nodules (≤ 1 cm, $p = 0.004$). The AUC for GPT-4 was 0.67, while residents achieved 0.75.

Conclusion: The GPT-4 model shows potential in classifying thyroid nodules, but its diagnostic accuracy remains significantly lower than that of resident physicians, particularly for smaller malignant nodules.

KEYWORDS

GPT-4, large language models, thyroid nodules, ultrasound, artificial intelligence

Introduction

Recent advancements in artificial intelligence (AI), especially models such as Generative Pre-trained Transformers (GPT), show promise in the medical field, including the analysis of ultrasound images (Wang et al., 2024; Choi et al., 2017). The new version, GPT-4, with its improved natural language processing and image analysis skills, appears to be a valuable tool for interpreting complex medical data (Law et al., 2024; Fink et al., 2023; Brin et al., 2024).

Although GPT-4 has demonstrated potential in ultrasound imaging tasks, it still has certain limitations. GPT-4V, the vision-enabled version of GPT-4, extends its capabilities by processing both text and visual inputs, allowing it to interpret radiological images. GPT-4V's ability to interpret radiological images was examined, including ultrasound (Brin et al., 2024; Jiang et al., 2024). It was found that while GPT-4V could identify imaging modalities, its accuracy in diagnosing different pathologies varied significantly across different types of

images. This indicates that GPT-4's capabilities in ultrasound analysis may require further refinement. Additionally, another study was conducted and GPT-4V's capabilities across a broader range of medical imaging tasks were evaluated, not limited to but including ultrasound (Li et al., 2023). Their results pointed out that, although GPT-4V was proficient in generating descriptive reports and identifying question types within medical visual question answering, significant issues were observed, particularly in the areas of disease diagnosis and visual grounding in ultrasound images. The challenges still remain, making the potential less straightforward.

Using GPT for ultrasound analysis is still new, but early studies suggest it might be beneficial. GPT4MIA, a framework utilizing GPT-3 for medical image analysis was proposed. Their findings suggest that GPT can effectively detect prediction errors and improve prediction accuracy in conjunction with vision-based models (Zhang and Chen, 2023). Similarly, the application of deep learning, including GPT models, in ultrasound image analysis was reviewed (Liu et al., 2019). They have pointed out the need for advanced automatic analysis methods to enhance diagnostic accuracy and make assessments more objective.

Another notable application of GPT in medical imaging was explored, GPT-4V's performance in multimodal medical diagnosis across various human body systems and imaging modalities was assessed (Wu et al., 2023). While GPT-4 V demonstrated proficiency in modality and anatomy recognition, but still face challenges in disease diagnosis and report generation, highlighting the need for further refinement. Additionally, the ability of GPT-4 on medical competency examinations and benchmark datasets was evaluated, finding that GPT-4 significantly outperformed earlier models in text-based medical tasks, though its image analysis capabilities still require improvement (Nori et al., 2023).

Furthermore, the effectiveness of real-time analysis of ultrasound images using GPU-accelerated techniques was demonstrated, showing significant speed improvements in image processing, which is crucial for practical applications in clinical settings (Eidheim et al., 2005). Similarly, a study highlighted the potential parallel computation that is GPU-based for the process of ultrasound computed tomography, this can not only enhance speed but also improve image quality, in detection of early breast cancer (Sun et al., 2018).

Despite the promising initial results, there is currently no large-scale study that evaluates the accuracy of GPT models in classifying thyroid nodules through ultrasound. This study aims to address this gap by comparing the performance of GPT-4 with ultrasound resident physicians in distinguishing benign from malignant thyroid nodules. The results could pave the way for future applications of LLMs in medical diagnostics, and provide directions for improvements.

Materials and methods

Patients and data set

For this study, we retrospectively selected 1,145 ultrasound images from 923 thyroid nodules, obtained at our center between June 2021 and June 2024, involving patients who were aged 18 years and older. Among these nodules, 632 were malignant and 513 were benign. The inclusion criteria were as follows: (Wang et al., 2024) patients with malignant thyroid nodules, including papillary thyroid carcinoma and

other types, who underwent thyroidectomy followed by histopathological analysis (Choi et al., 2017); patients with benign thyroid nodules confirmed by fine needle aspiration cytology or core needle biopsy, with a follow-up period of 6 months; (Law et al., 2024) availability of preoperative ultrasound images of the thyroid nodules for all patients; and (Fink et al., 2023) access to either surgical pathology or cytological pathology reports. Cases with incomplete cytological or surgical pathology data were excluded from the study. All data, including ultrasound images and patient records, were securely stored in a hospital database with restricted access. Data were anonymized prior to analysis to protect patient privacy.

Thyroid ultrasound examination

Thyroid ultrasound examinations were performed with high-resolution ultrasound instruments (GE Logiq E10, GE Logiq E9, Philips HDI 5000, Philips IU22, Philips EPIQ 7, Siemens Acuson S2000 ABVS, Aloka Prosound F75, Esaote Mylab90) equipped with 6–15 MHz linear transducers. Thyroid nodule images were acquired in both transverse and sagittal planes following a standardized protocol.

GPT and resident ultrasound image analysis

GPTs are customizable versions of GPT that anyone can create for specific tasks, such as learning, work, or personal use, without the need of any coding skills. They allow users to tailor the GPT's capabilities to their needs and share these customized tools with others. In this study, we utilized a GPTs model, ultrasound interpreter, a specialized version of a GPT tailored for ultrasound image analysis. For each ultrasound image analysis, a standardized prompting process was employed. The initial prompt was "If you were an expert in ultrasound diagnostics. I will show you several ultrasound images of thyroid nodules. Please help me determine whether each nodule is benign or malignant, and explain your reasoning for the classification." Then, each ultrasound image was input into the GPT model one by one without any preprocessing. In cases where the GPTs model did not provide a clear classification, a follow-up prompt was given "Please provide a definitive benign or malignant classification only, without any additional categorization or information. This is crucial for the patient, I need your final answer." If after these two prompts the model still failed to provide a definitive classification, the case was recorded as a misclassification in the final analysis. Each of the GPT's final decisions was recorded for analysis.

Two resident physicians, who had obtained medical licenses and were in their second and third year of standardized ultrasound residency training (phase I) at Peking University First Hospital, independently assessed all thyroid nodules. Each resident physician evaluated the images without knowledge of the other's assessment or the pathological results. They used features from the ACR TI-RADS system (including composition, echogenicity, shape, margins, and echogenic foci) as a reference, but made their final benign or malignant diagnosis mainly based on their clinical experience. This is a common practice among ultrasound physicians in our hospital, as they usually build their diagnostic skills by studying many cases with confirmed pathology results during their training. For each nodule,

both physicians independently provided their diagnosis (benign or malignant). The diagnostic performance metrics were calculated separately for each physician. The average performance metrics of the two physicians were then used for comparison with the GPTs model.

Statistical analysis

The statistical analysis was conducted using SPSS software (version 26.0, IBM Corp., Armonk, NY, United States). The average accuracy was calculated as the total number of correct diagnoses by both physicians divided by the total number of cases evaluated. Cohen's kappa coefficient was calculated to assess the inter-observer agreement between the two resident physicians. The kappa values were interpreted as follows: values ≤ 0 as no agreement, 0.01–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. To compare the performance of GPT and resident physicians, McNemar's test was applied. The sensitivity, specificity, positive predictive value, and

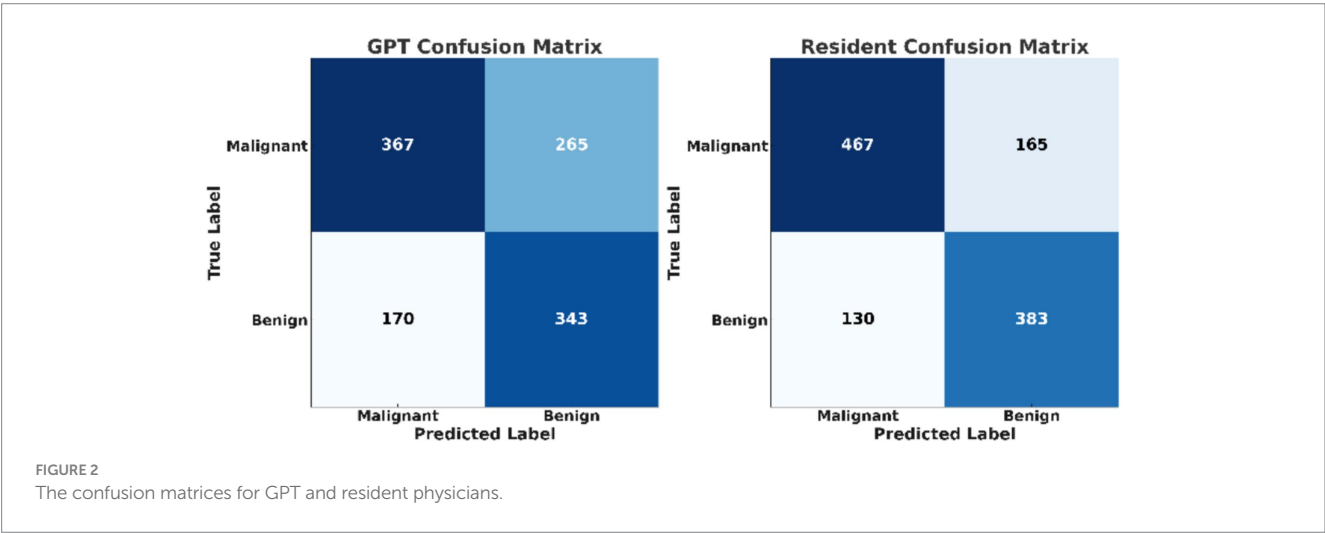
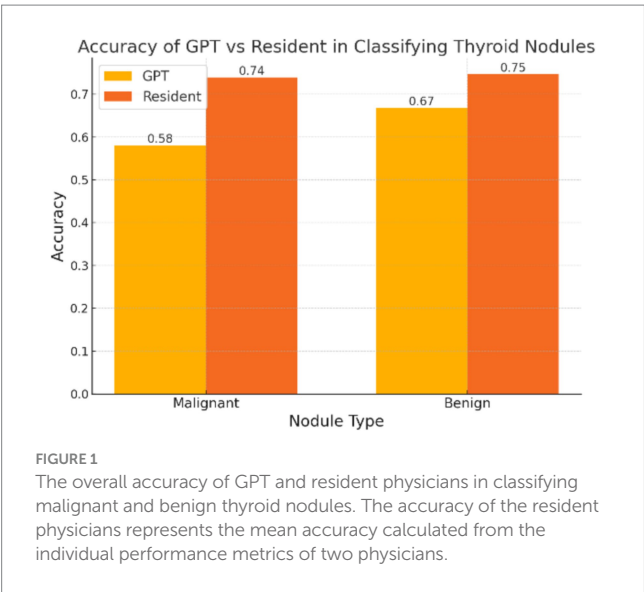
negative predictive value of both the GPT model and resident physicians were calculated. Confusion matrices were used to evaluate the performance of both classifiers. A receiver operating characteristic (ROC) curve analysis was performed to compare the diagnostic accuracy, and the area under the ROC curve (AUC) was calculated as a measure of overall diagnostic performance for both classifiers. A *p*-value of less than 0.05 was considered statistically significant.

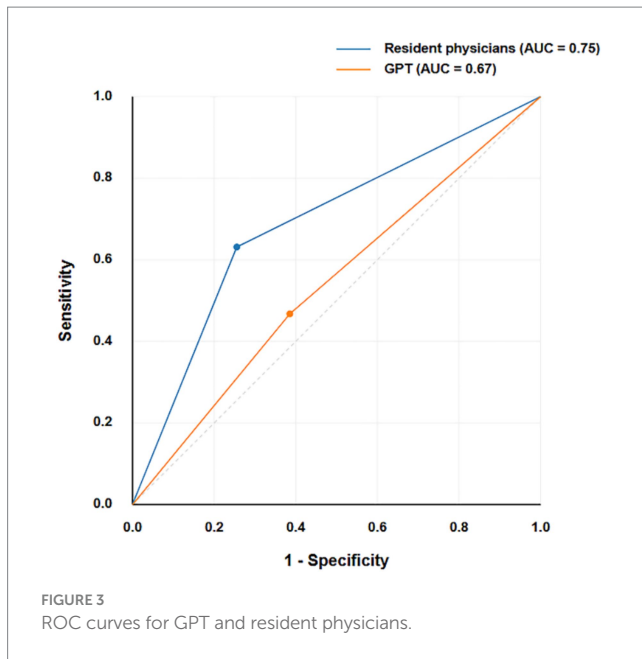
Results

The diagnostic performance of both resident physicians and the GPT model was analyzed (Figure 1). For malignant nodules, Resident A correctly identified 487 of 632 cases (77.1%) while Resident B identified 448 cases (70.9%), with an average accuracy of 73.89%. For benign nodules, Resident A correctly classified 378 of 513 cases (73.7%) and Resident B classified 388 cases (75.6%), with an average accuracy of 74.66%. The inter-observer agreement between the two residents was moderate, with a Cohen's kappa coefficient of 0.41. In comparison, GPT correctly identified 367 out of 632 malignant nodules (58.07%) and 343 out of 513 benign nodules (66.86%). The test results indicated a statistically significant difference in classification of malignant nodules between GPT and the average performance of resident physicians ($p < 0.001$). For benign nodules, although to a lesser extent, the difference was still significant ($p = 0.048$).

Among the 632 malignant nodules, 398 had a maximum diameter of 1 cm or less, and GPT correctly identified 214 nodules, with an accuracy of 53.77%. For the 234 nodules with a maximum diameter greater than 1 cm, GPT correctly identified 153 nodules, resulting in an accuracy of 65.38%. The two groups had a significant difference ($p = 0.004$), indicating that the GPT model performed better at identifying larger nodules than smaller ones.

The confusion matrices provide a detailed breakdown of true positives, true negatives, false positives, and false negatives for both classifiers (Figure 2). The confusion matrices reveal that resident physicians outperformed GPT in correctly classifying both malignant and benign nodules, particularly with a significant reduction in false negatives for malignant cases. As shown in Figure 3, the AUC for GPT





was 0.67, whereas for resident physicians, the AUC was 0.75. This indicates that resident physicians have a higher ability in distinguishing between malignant and benign nodules compared to GPT.

Discussion

The comparative analysis between the ultrasound interpreter GPTs model and resident physicians in classifying thyroid nodules highlights both the potential and limitations of AI in clinical diagnostics. The study demonstrated that while the GPTs model is capable of providing consistent interpretations, it still falls short when compared to human clinicians, particularly in identifying malignant nodules less than 1 cm.

Overall, the GPTs correctly identified 367 out of 632 malignant nodules, achieving an accuracy of 58.07%, whereas the resident physicians achieved a significantly higher accuracy of 73.89%. This performance gap suggests that while AI can assist in diagnostic processes, it is still far from being able to replace humans, especially in critical areas such as cancer detection using ultrasound. For benign nodules, the GPTs performed better, with an accuracy of 66.86%, but still lagged behind the residents' accuracy of 74.66%.

One of the findings of this study was the difference in the GPTs model's performance based on the size of the nodules. The model correctly identified 53.77% of malignant nodules with a maximum diameter of 1 cm or less, compared to 65.38% for nodules larger than 1 cm. The significant difference ($p = 0.004$) suggests that the GPT model encounters greater challenges when analyzing smaller nodules. This difficulty is likely attributed to the more subtle image features, which are harder to interpret accurately. This observation is consistent with findings in the existing literature, where AI models often face difficulties in handling smaller or more complex cases (He et al., 2022; Wang et al., 2022). This highlights the need for further refinement in both training datasets and model algorithms to improve performance.

These limitations of the GPTs model can be attributed to several factors. First, the current AI algorithms used in GPTs are highly dependent on the quality and diversity of the training data. If the training datasets are not sufficiently representative of all potential clinical scenarios, the AI model may struggle to generalize effectively to new cases (Willemink et al., 2020; Wildman-Tobriner et al., 2019). Another significant factor is that the GPT model used in this study was a custom-tailored version, optimized for ultrasound image interpretation. However, this particular GPT is not open-source, and we do not have insight into how it was specifically configured or tuned, nor do we know the amount and quality of the original data used for training. Consequently, it is unclear to what extent this model leverages the full potential of GPTs for image recognition. This lack of transparency makes it difficult to assess the model's true capabilities (Savadjiev et al., 2019). In future studies, we could design and fine-tune our own GPTs to potentially achieve better results by controlling these variables.

An analysis of the GPT model's errors shows that it often mimics human reasoning by relying on common ultrasound features like nodule size, margins, echogenicity, shape, and vascularity (Figure 4). However, the model is not accurate in identifying key features of benign and malignant thyroid nodules, such as taller than wide shape and very low echogenicity, which are clear signs of malignancy, as well as coarse, strong echogenicity, which are often associated with benign lesions. While this approach seems logical, it may miss the more detailed and complex patterns that advanced imaging techniques can detect. Adding radiomic features, which measure the texture of an image, or deep learning features, which capture more complex details, could greatly improve the model's ability to tell benign nodules from malignant ones (Lambin et al., 2017).

Despite these current limitations, the potential for AI in medical diagnostics remains vast. One of the greatest strengths of AI, including GPTs models, is their ability to rapidly evolve and improve. As more diverse and comprehensive datasets become available, and as AI algorithms are refined, the performance gap between AI and human clinicians is expected to narrow. In fact, with sufficient data and appropriate training methodologies, it is conceivable that AI could surpass the diagnostic accuracy of average clinicians in the near future.

To achieve this, future developments should focus on enhancing the transparency and interpretability of AI models, integrating them more effectively into clinical workflows, and continuously updating the models with new data to ensure their accuracy and relevance. The rapid pace of advancements in AI suggests that these improvements are not only possible but imminent (Hosny et al., 2018; Sheth and Giger, 2020).

This study has several limitations that should be considered. First, the dataset used in this study consists of thyroid nodules with pathological results, most of which are TI-RADS 4 or higher. As a result, the dataset may not represent all types of thyroid nodules seen in clinical practice, particularly those with lower risk (TI-RADS 2 and 3). The performance of GPT models in identifying these lower-risk nodules has not been thoroughly explored. Additionally, the images in the dataset come from a single hospital, which may introduce sample bias. The study also did not account for differences in image quality or variations in image acquisition methods, factors that could influence the model's performance in



Detailed Analysis of the Thyroid Nodule

- Hypoechoic Nature:** The nodule appears hypoechoic compared to the surrounding thyroid tissue. Hypoechoic nodules can be seen in both benign and malignant conditions; thus, it requires careful assessment of additional features.
- Shape and Orientation:** The nodule is well-defined and relatively round. Irregular margins and taller-than-wide orientation are more concerning for malignancy, but this nodule does not exhibit these features.
- Microcalcifications:** There are no evident microcalcifications in the nodule. Microcalcifications are often associated with papillary thyroid carcinoma.
- Vascularity:** The image does not show Doppler information, but based on the grayscale image alone, there is no obvious increased vascularity, which is sometimes a feature of malignancy.
- Comet Tail Artifact:** There appears to be a comet tail artifact, which can be indicative of a benign colloid nodule.
- Cystic Components:** The nodule appears solid without cystic components. Purely cystic nodules are often benign, while solid nodules require further evaluation.

Conclusion

Based on the sonographic features of being hypoechoic, well-defined margins, absence of microcalcifications, and presence of a comet tail artifact, the nodule appears to be benign. However, it is essential to correlate these findings with clinical and laboratory data, and possibly follow up with fine-needle aspiration biopsy if clinically indicated.



The nodule in this ultrasound image is most likely malignant. Here is the detailed analysis:

- Boundary Characteristics:** The nodule has an ill-defined, irregular boundary, suggesting possible infiltration into surrounding tissues. Malignant nodules often present with such characteristics due to invasive growth patterns.
- Shape:** The nodule appears irregular in shape. Malignant nodules commonly have irregular, lobulated shapes as opposed to the smooth, round shapes seen in benign nodules.
- Internal Structure:** There is heterogeneous echotexture within the nodule. Malignant nodules typically have a heterogeneous appearance due to areas of necrosis, fibrosis, or calcification.
- Size:** The dimensions provided (D1: 4.23 cm, D2: 2.76 cm) show a relatively large nodule. While size alone is not diagnostic, larger nodules warrant further investigation and are more suspicious for malignancy.
- Echogenicity:** The nodule's echogenicity appears hypoechoic relative to the surrounding tissue, which is a common feature in malignant nodules.

Based on these ultrasound features, the nodule displays several characteristics commonly associated with malignancy. Further evaluation, including a biopsy, would be necessary to confirm the diagnosis.

FIGURE 4

Two typical examples of GPT's misjudgment process and results in nodule classification. The left nodule is malignant, and the right nodule is benign. Both ultrasound resident physicians correctly identified the nodules, while GPT made incorrect classifications, failing to recognize the clinical significance of the taller-than-wide shape, comet-tail sign, and coarse calcifications.

real-world clinical settings. Moreover, the study involved two resident physicians from the same hospital, but their varying stages of training were not explicitly controlled. Furthermore, because the residents are still in training, their diagnostic reasoning may differ from that of more experienced physicians. The residents also had access to cases with pathological results in daily work, while many primary care hospitals may not have such opportunities for training. As a result, the performance of human physicians in this study should be interpreted as a reference, rather than representing the capabilities of ultrasound physicians in general clinical practice. Future research could address these limitations by including a wider range of nodules, more diverse datasets, and a larger group of residents or more experienced clinicians.

Conclusion

In conclusion, although the GPTs ultrasound interpreter model shows potential, it currently falls short of matching clinicians in accurately identifying malignant from benign thyroid nodules. However, with the continuous advancements in AI technology, supported by increasingly large and diverse datasets, this gap is expected to narrow. There is potential for the models to equal or even surpass human clinicians in diagnostic accuracy, potentially reshaping the future of medical diagnostics.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving humans were approved by Peking University First Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

BW: Conceptualization, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Writing – original draft, Writing – review & editing. XZ: Data curation, Investigation, Validation, Writing – review & editing. YS: Data curation, Validation, Writing – review & editing. XS: Investigation,

Writing – review & editing. LC: Resources, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Brin, D., Sorin, V., Barash, Y., Konen, E., Glicksberg, B. S., Nadkarni, G. N., et al. (2024). Assessing GPT-4 multimodal performance in radiological image analysis. *Eur. Radiol.* doi: 10.1007/s00330-024-11035-5
- Choi, Y., Baek, J., Park, H., Shim, W., Kim, T. Y., Shong, Y., et al. (2017). A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: initial clinical assessment. *Thyroid* 27, 546–552. doi: 10.1089/thy.2016.0372
- Eidheim, O. C., Skjermo, J., and Aurdal, L. (2005). Real-time analysis of ultrasound images using GPU. *Int. Congr. Ser.* 1281, 284–289. doi: 10.1016/j.ics.2005.03.187
- Fink, M. A., Bischoff, A., Fink, C. A., Moll, M., Kroschke, J., Dulz, L., et al. (2023). Potential of ChatGPT and GPT-4 for data Mining of Free-Text CT reports on lung Cancer. *Radiology* 308:e231362. doi: 10.1148/radiol.231362
- He, L. T., Chen, F. J., Zhou, D. Z., Zhang, Y. X., Li, Y. S., Tang, M. X., et al. (2022). A comparison of the performances of artificial intelligence system and radiologists in the ultrasound diagnosis of thyroid nodules. *Curr. Med. Imaging* 18, 1369–1377. doi: 10.2174/1573405618666220422132251
- Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510. doi: 10.1038/s41568-018-0016-5
- Jiang, Y., Chen, C., Nguyen, D., Mervak, B., and Tan, C. (2024). GPT-4V cannot generate radiology reports yet. *ArXiv*. doi: 10.48550/arXiv.2407.12176
- Lambin, P., Leijenaar, R. T. H., Deist, T. M., Peerlings, J., de Jong, E. E. C., van Timmeren, J., et al. (2017). Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 14, 749–762. doi: 10.1038/nrclinonc.2017.141
- Law, S., Oldfield, B., and Yang, W. (2024). ChatGPT/GPT-4 (large language models): opportunities and challenges of perspective in bariatric healthcare professionals. *Obes. Rev.* 25:e13746. doi: 10.1111/obr.13746
- Li, Y., Liu, Y., Wang, Z., Liang, X., Liu, L., Wang, L., et al. (2023). A comprehensive study of GPT-4V's multimodal capabilities in medical imaging. *medRxiv*. doi: 10.1101/2023.11.03.23298067
- Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., et al. (2019). Deep learning in medical ultrasound analysis: a review. *Engineering* 5, 261–275. doi: 10.1016/j.eng.2018.11.020
- Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *CoRR*. doi: 10.48550/arXiv.2303.13375
- Savadjiev, P., Chong, J., Dohan, A., Vakalopoulou, M., Reinhold, C., Paragios, N., et al. (2019). Demystification of AI-driven medical image interpretation: past, present and future. *Eur. Radiol.* 29, 1616–1624. doi: 10.1007/s00330-018-5674-x
- Sheth, D., and Giger, M. L. (2020). Artificial intelligence in the interpretation of breast cancer on MRI. *J. Magn. Reson. Imaging* 51, 1310–1324. doi: 10.1002/jmri.26878
- Sun, X., Wang, S., Song, J., Zhou, L., Peng, Y., Ding, M., et al., editors. Toward parallel optimal computation of ultrasound computed tomography using GPU. In: Proceedings of SPIE 10580, Medical Imaging 2018: Ultrasonic Imaging and Tomography; (2018).
- Wang, B., Wan, Z., Li, C., Zhang, M., Shi, Y., Miao, X., et al. (2022). Identification of benign and malignant thyroid nodules based on dynamic AI ultrasound intelligent auxiliary diagnosis system. *Front. Endocrinol. (Lausanne)* 13:1018321. doi: 10.3389/fendo.2022.1018321
- Wang, Z., Zhang, Z., Traverso, A., Dekker, A., Qian, L., and Sun, P. (2024). Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. *Quant. Imaging Med. Surg.* 14, 1602–1615. doi: 10.21037/qims-23-1180
- Wildman-Tobriner, B., Buda, M., Hoang, J. K., Middleton, W. D., Thayer, D., Short, R. G., et al. (2019). Using artificial intelligence to revise ACR TI-RADS risk stratification of thyroid nodules: diagnostic accuracy and utility. *Radiology* 292, 112–119. doi: 10.1148/radiol.2019182128
- Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., et al. (2020). Preparing medical imaging data for machine learning. *Radiology* 295, 4–15. doi: 10.1148/radiol.2020192224
- Wu, C., Lei, J., Zheng, Q., Zhao, W., Lin, W., Zhang, X., et al. (2023). Can GPT-4V(ision) serve medical applications? Case studies on GPT-4V for multimodal medical diagnosis. *CoRR*. doi: 10.48550/arXiv.2310.09909
- Zhang, Y., and Chen, D. (2023). GPT4MIA: Utilizing Geneative pre-trained transformer (GPT-3) as a plug-and-play Transductive model for medical image analysis. *arXiv*. doi: 10.48550/arXiv.2302.08722

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.