Original Article

# Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach

**Zhixiang Wang[1,2], Zhen Zhang[2], Alberto Traverso[2], Andre Dekker[2], Linxue Qian[1], Pengfei Sun[1]**

[1]Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, Beijing, China; [2]Department of Radiation Oncology (Maastro), GROW-School for Oncology, Maastricht University Medical Centre+, Maastricht, The Netherlands

*Contributions:* (I) Conception and design: Z Wang, Z Zhang; (II) Administrative support: L Qian, P Sun; (III) Provision of study materials or patients: P Sun; (IV) Collection and assembly of data: A Traverso; (V) Data analysis and interpretation: A Dekker; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

*Correspondence to:* Linxue Qian, PhD; Pengfei Sun, MM. Department of Ultrasound, Beijing Friendship Hospital, Capital Medical University, No. 95, Yong'an Road, Xicheng District, Beijing 100050, China. Email: qianlinxue2002@163.com; spfjlmpc@163.com.

**Background:** As artificial intelligence (AI) becomes increasingly prevalent in the medical field, the effectiveness of AI-generated medical reports in disease diagnosis remains to be evaluated. ChatGPT is a large language model developed by open AI with a notable capacity for text abstraction and comprehension. This study aimed to explore the capabilities, limitations, and potential of Generative Pre-trained Transformer (GPT)-4 in analyzing thyroid cancer ultrasound reports, providing diagnoses, and recommending treatment plans.

**Methods:** Using 109 diverse thyroid cancer cases, we evaluated GPT-4's performance by comparing its generated reports to those from doctors with various levels of experience. We also conducted a Turing Test and a consistency analysis. To enhance the interpretability of the model, we applied the Chain of Thought (CoT) method to deconstruct the decision-making chain of the GPT model.

**Results:** GPT-4 demonstrated proficiency in report structuring, professional terminology, and clarity of expression, but showed limitations in diagnostic accuracy. In addition, our consistency analysis highlighted certain discrepancies in the AI's performance. The CoT method effectively enhanced the interpretability of the AI's decision-making process.

**Conclusions:** GPT-4 exhibits potential as a supplementary tool in healthcare, especially for generating thyroid gland diagnostic reports. Our proposed online platform, "ThyroAIGuide", alongside the CoT method, underscores the potential of AI to augment diagnostic processes, elevate healthcare accessibility, and advance patient education. However, the journey towards fully integrating AI into healthcare is ongoing, requiring continuous research, development, and careful monitoring by medical professionals to ensure patient safety and quality of care.

**Keywords:** ChatGPT; thyroid cancer; diagnosis; artificial intelligence (AI); ultrasound

Submitted Aug 20, 2023. Accepted for publication Nov 30, 2023. Published online Jan 11, 2024.
doi: 10.21037/qims-23-1180
**View this article at:** https://dx.doi.org/10.21037/qims-23-1180

## Introduction

Thyroid nodules are the most common endocrine tumors, and in recent years, the incidence of thyroid cancer has been increasing (1). Ultrasound is often the preferred imaging modality for the thyroid due to its non-invasive nature, absence of radiation, and affordability (2). It is used for the staging of the disease, based on Thyroid Imaging-Reporting and Data System proposed by the American College of Radiology (ACR TI-RADS) (3). The accuracy of thyroid cancer diagnosis largely depends on the doctor's experience, as physicians with varying expertise may conclude with inconsistent results with low repeatability (4).

Last year, open artificial intelligence (AI) introduced ChatGPT (5), a large language model with a notable capacity for text abstraction and comprehension. AI has shown capabilities that are close to or even exceed human performance in abstract and logical reasoning tasks. For instance, AI has achieved remarkable performance in tasks like DeepMind's AlphaGo mastering the game of Go, Generative Pre-trained Transformer (GPT)-3's natural language understanding, IBM Watson's success on Jeopardy, and MuZero's ability to learn board games without prior knowledge (6-9). AI has significant potential to improve diagnostic accuracy and efficiency while reducing human errors (10). The 2023 release of GPT-4 (11) represents the most powerful language model to date, sparking interest in applying AI to the medical field, particularly for automated medical report analysis (12). However, the limitations of AI models in specialized vertical domains, such as medical diagnosis, include a scarcity of training data in these specific fields, which can hinder the model's ability to generalize and provide accurate predictions. Additionally, AI models may exhibit poor robustness, leading to increased vulnerability to adversarial examples or noise, which can significantly impact their performance and reliability in complex medical diagnostic tasks (13).

Our goal is to explore the real capabilities, limitations, and potential of GPT-4 in analyzing thyroid cancer ultrasound reports, providing diagnoses, and suggesting treatment plans.

The application of GPT-4 in medical diagnosis, particularly for evaluating thyroid cancer reports, is crucial as it can improve the accuracy, efficiency, and overall patient outcomes of diagnosis. This technology has the potential to save time, reduce costs, reduce human errors, and improve the accessibility of healthcare in remote areas. Furthermore, it encourages continuous learning and promotes collaboration among healthcare professionals, ultimately incorporating AI into clinical practice to enhance the level of patient care.

## Methods

### *Data*

In this study, we assessed GPT-4 performance using 109 diverse thyroid cancer cases from Beijing Friendship Hospital, Capital Medical University, between January 2022 and April 2023. These patients underwent thyroid ultrasound examinations. Their diagnoses were confirmed by fine needle aspiration (FNA) or surgical pathology, and the reports were written in Chinese according to the ACR TI-RADS (4) in a free-text style. Ultrasound examinations were conducted using a 3–12 linear probe (RS80A with Prestige, Samsung Medison, Co., Ltd., Seoul, South Korea) and a SL15–4 multi-frequency linear probe (SuperSonic Imagine, Aix-en-Provence, France). The cases included patients of different ages, genders, medical histories, and conditions, such as papillary (102 cases), follicular (3 cases), undifferentiated (2 cases), and medullary carcinoma (2 cases). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Ethical approval for this study was obtained from the Medical Ethics Committee of Beijing Friendship Hospital, Capital Medical University. This was not a study of human subjects. Therefore, informed consent was not required. Forty doctors, of which 13 junior (1–5 years of experience), 17 intermediate (6–10 years), and 10 senior (>10 years), participated in the study. They represented four hospitals—Beijing Friendship Hospital, Capital Medical University, Beijing Children's Hospital affiliated with Capital Medical University, Binzhou Central Hospital in Shandong Province, and Department of Ultrasound, Peking University Third Hospital, Beijing—ensuring a comprehensive and objective assessment of the performance.

### *GPT-4 report generation and analysis and platform development*

In this study, we hypothesize that an online platform based on GPT-4 can be built to provide evidence-based diagnosis and treatment recommendations based on thyroid ultrasound description reports. We input the ultrasound

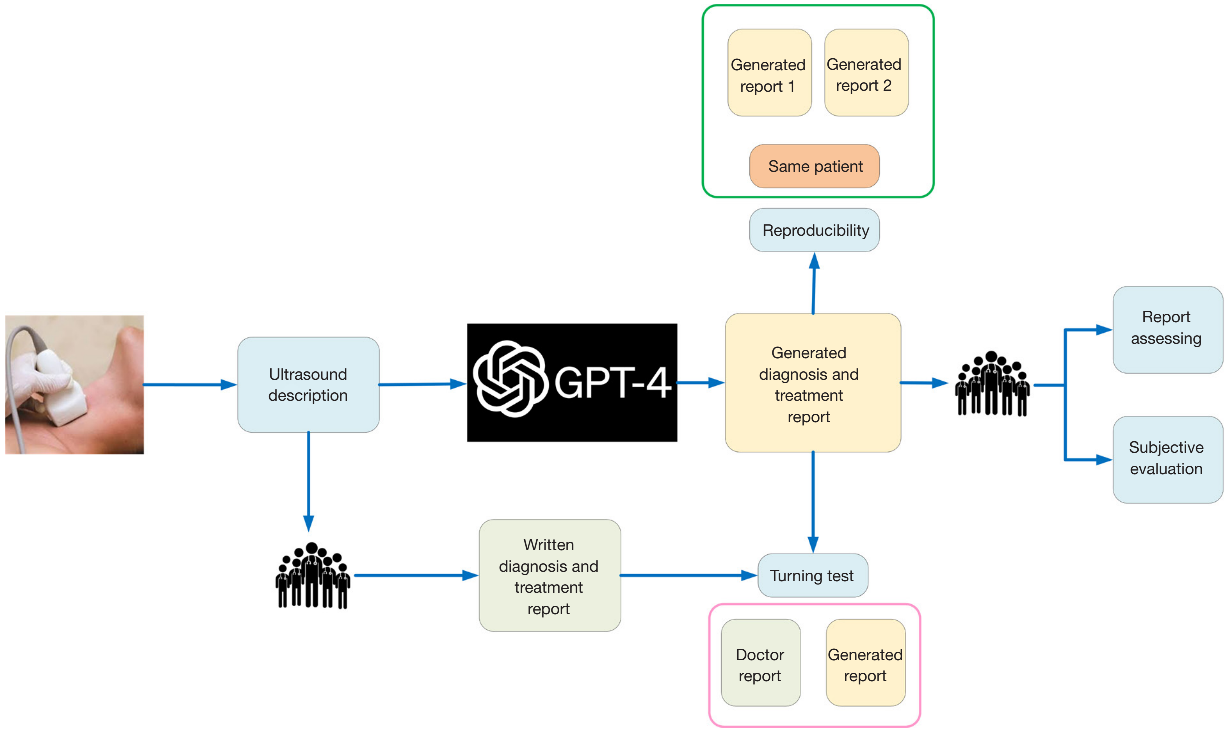


**Figure 1** Experimental workflow diagram. The process begins with data collection and obtaining ultrasound reports, followed by using GPT-4 to generate diagnoses and treatment outcomes. Subsequently, the results are subjected to three experiments: (I) doctors rate the generated reports; (II) a Turing Test is used to compare doctor-generated and AI-generated reports; (III) model reproducibility experiment, where reports are generated twice and differences are compared. GPT, Generative Pre-trained Transformer; AI, artificial intelligence.

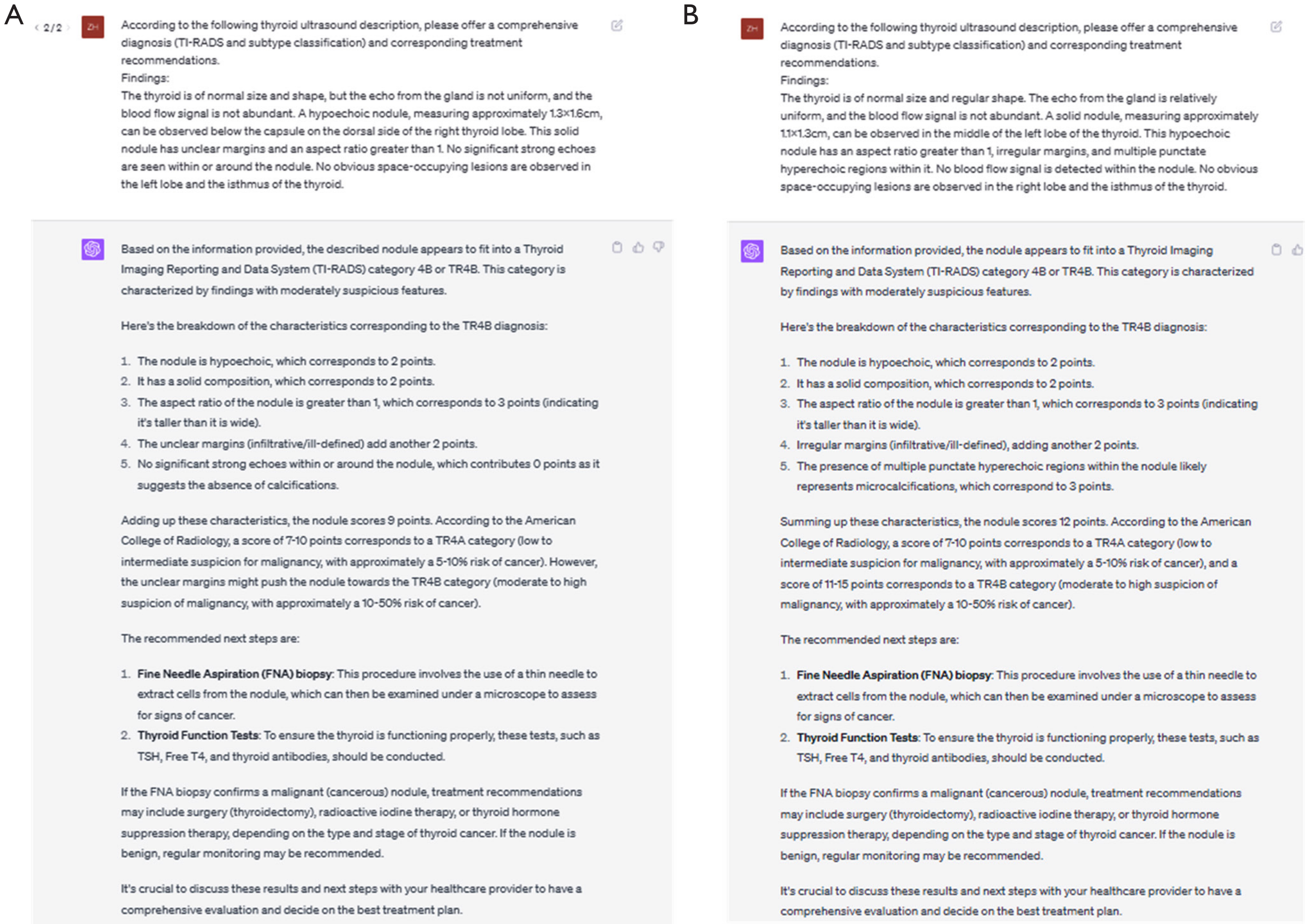
description reports into the GPT-4 model to generate structured diagnostic reports and treatment plan suggestions for patients.

The GPT series, developed by OpenAI, represents a cutting-edge pre-training language model capable of handling highly complex Natural Language Processing tasks, such as article generation, code generation, machine translation, and questions and answers (Q&A) (14). GPT-4, released in March 2023, is the latest upgrade of the architecture, widely acknowledged as the most advanced AI-generated content (AIGC) model available (7). The research workflow is shown in *Figure 1*. During our report analysis process, we input ultrasound medical descriptions onto ChatGPT and prompted it to generate diagnosis and treatment recommendations. We subsequently collected the output reports for evaluation. *Figure 2* exhibits the examples of collecting questions and answers. The detailed prompt provided was: "*According to the following thyroid ultrasound description*, *please offer a comprehensive diagnosis (TI-RADS classification) and corresponding treatment recommendations*".

In parallel, to enhance the practicality of our research, we employed the Flask framework (15) to create an intelligent diagnostic platform named "ThyroAIGuide" (*Figure 3*). ThyroAIGuide is an intuitive online diagnostic tool designed for preliminary thyroid cancer diagnosis. Users are required to input pertinent clinical data and ultrasound image descriptions, after which the platform harnesses the power of the GPT-4 model to generate comprehensive diagnostic reports. These reports include TI-RADS grading, risk assessment, and recommendations on whether a biopsy is necessary. Access to this system is available via the following link: http://47.101.207.206:8000/thyro. Details for "ThyroAIGuide" can be found in Appendix 1.

### *Assessing GPT-4's performance in report analysis*

To gain a deeper understanding of ChatGPT's accuracy level in generating reports, we collected and analyzed the ratings of the generated reports. Evaluation criteria



According to the following thyroid ultrasound description, please offer a comprehensive diagnosis (TI-RADS and subtype classification) and corresponding treatment recommendations.
Findings:
The thyroid is of normal size and shape, but the echo from the gland is not uniform, and the blood flow signal is not abundant. A hypoechoic nodule, measuring approximately 1.3×1.6cm, can be observed below the capsule on the dorsal side of the right thyroid lobe. This solid nodule has unclear margins and an aspect ratio greater than 1. No significant strong echoes are seen within or around the nodule. No obvious space-occupying lesions are observed in the left lobe and the isthmus of the thyroid.

Based on the information provided, the described nodule appears to fit into a Thyroid Imaging Reporting and Data System (TI-RADS) category 4B or TR4B. This category is characterized by findings with moderately suspicious features.

Here's the breakdown of the characteristics corresponding to the TR4B diagnosis:

1. The nodule is hypoechoic, which corresponds to 2 points.
2. It has a solid composition, which corresponds to 2 points.
3. The aspect ratio of the nodule is greater than 1, which corresponds to 3 points (indicating it's taller than it is wide).
4. The unclear margins (infiltrative/ill-defined) add another 2 points.
5. No significant strong echoes within or around the nodule, which contributes 0 points as it suggests the absence of calcifications.

Adding up these characteristics, the nodule scores 9 points. According to the American College of Radiology, a score of 7-10 points corresponds to a TR4A category (low to intermediate suspicion for malignancy, with approximately a 5-10% risk of cancer). However, the unclear margins might push the nodule towards the TR4B category (moderate to high suspicion of malignancy, with approximately a 10-50% risk of cancer).

The recommended next steps are:

1. **Fine Needle Aspiration (FNA) biopsy**: This procedure involves the use of a thin needle to extract cells from the nodule, which can then be examined under a microscope to assess for signs of cancer.
2. **Thyroid Function Tests**: To ensure the thyroid is functioning properly, these tests, such as TSH, Free T4, and thyroid antibodies, should be conducted.

If the FNA biopsy confirms a malignant (cancerous) nodule, treatment recommendations may include surgery (thyroidectomy), radioactive iodine therapy, or thyroid hormone suppression therapy, depending on the type and stage of thyroid cancer. If the nodule is benign, regular monitoring may be recommended.

It's crucial to discuss these results and next steps with your healthcare provider to have a comprehensive evaluation and decide on the best treatment plan.

According to the following thyroid ultrasound description, please offer a comprehensive diagnosis (TI-RADS and subtype classification) and corresponding treatment recommendations.
Findings:
The thyroid is of normal size and regular shape. The echo from the gland is relatively uniform, and the blood flow signal is not abundant. A solid nodule, measuring approximately 1.1×1.3cm, can be observed in the middle of the left lobe of the thyroid. This hypoechoic nodule has an aspect ratio greater than 1, irregular margins, and multiple punctate hyperechoic regions within it. No blood flow signal is detected within the nodule. No obvious space-occupying lesions are observed in the right lobe and the isthmus of the thyroid.

Based on the information provided, the nodule appears to fit into a Thyroid Imaging Reporting and Data System (TI-RADS) category 4B or TR4B. This category is characterized by findings with moderately suspicious features.

Here's the breakdown of the characteristics corresponding to the TR4B diagnosis:

1. The nodule is hypoechoic, which corresponds to 2 points.
2. It has a solid composition, which corresponds to 2 points.
3. The aspect ratio of the nodule is greater than 1, which corresponds to 3 points (indicating it's taller than it is wide).
4. Irregular margins (infiltrative/ill-defined), adding another 2 points.
5. The presence of multiple punctate hyperechoic regions within the nodule likely represents microcalcifications, which correspond to 3 points.

Summing up these characteristics, the nodule scores 12 points. According to the American College of Radiology, a score of 7-10 points corresponds to a TR4A category (low to intermediate suspicion for malignancy, with approximately a 5-10% risk of cancer), and a score of 11-15 points corresponds to a TR4B category (moderate to high suspicion of malignancy, with approximately a 10-50% risk of cancer).

The recommended next steps are:

1. **Fine Needle Aspiration (FNA) biopsy**: This procedure involves the use of a thin needle to extract cells from the nodule, which can then be examined under a microscope to assess for signs of cancer.
2. **Thyroid Function Tests**: To ensure the thyroid is functioning properly, these tests, such as TSH, Free T4, and thyroid antibodies, should be conducted.

If the FNA biopsy confirms a malignant (cancerous) nodule, treatment recommendations may include surgery (thyroidectomy), radioactive iodine therapy, or thyroid hormone suppression therapy, depending on the type and stage of thyroid cancer. If the nodule is benign, regular monitoring may be recommended.

It's crucial to discuss these results and next steps with your healthcare provider to have a comprehensive evaluation and decide on the best treatment plan.

**Figure 2** GPT-4 assisted thyroid diagnosis and treatment suggestions. The examples of collecting questions and answers. The detailed prompt provided was: "*According to the following thyroid ultrasound description, please offer a comprehensive diagnosis (TI-RADS classification) and corresponding treatment recommendations*". GPT, Generative Pre-trained Transformer; TI-RADS, Thyroid Imaging-Reporting and Data System.

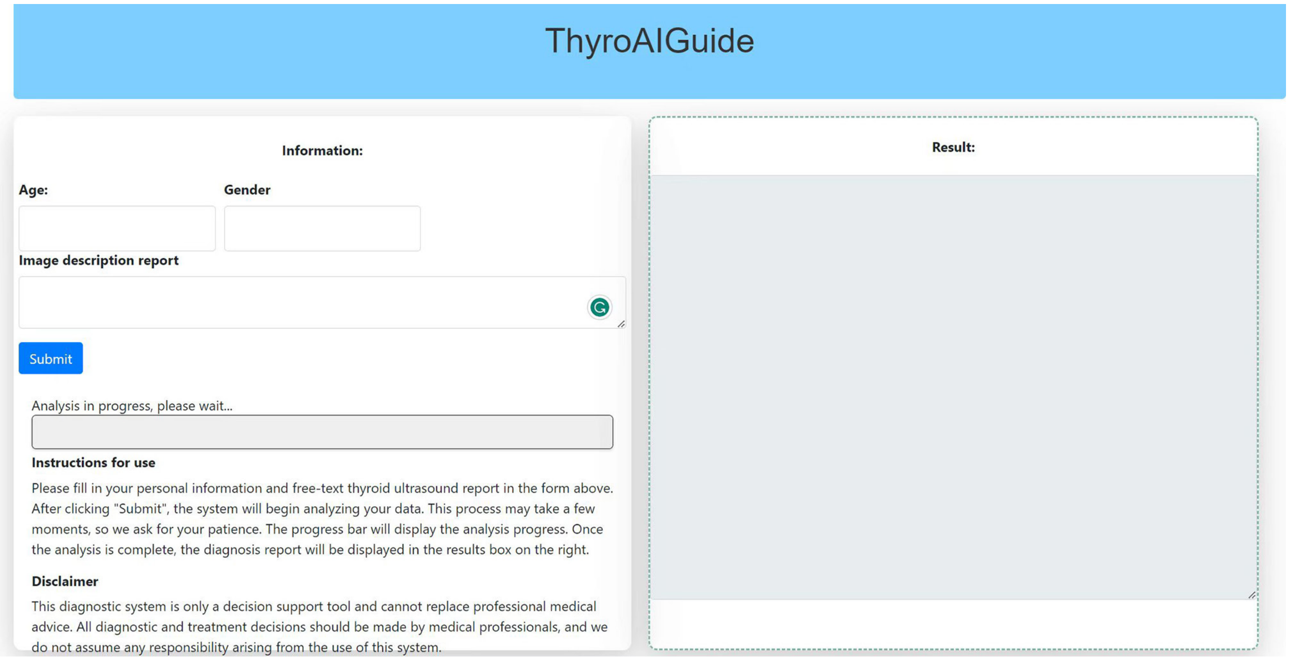
included diagnostic accuracy (TI-RADS Classification), report structure, use of professional terminology, and clarity of expression. Details of the score table are shown in the Appendix 2.

In addition, we explored the level of expertise that AI-generated reports can achieve by comparing their ratings with those of doctors with different levels of experience. We also conducted an in-depth investigation of erroneous reports.

### *Turing Test for generated report evaluation*

To assess doctors' ability to discern between human-written and AI-generated reports (16), we randomly included 50% of the AI-generated reports into the evaluation set. The report source was confidential and randomized. Doctors were asked to read the reports and rate the likelihood of human authorship (1= very low, 5= very high). While no specific instructions were provided, doctors typically assessed reports based on language quality, coherence, and medical context relevance. These factors influenced the assigned scores, with higher scores signifying a closer resemblance to human-authored reports.

We collected their judgments and calculated the proportion of correct determinations. If doctors' accuracy exceeded random guessing (50%), it would imply GPT-4 passed the Turing Test, suggesting its generated reports are nearly indistinguishable from human-written ones.



ThyroAIGuide

Information:

Age:

Gender

Image description report

Submit

Analysis in progress, please wait...

Instructions for use

Please fill in your personal information and free-text thyroid ultrasound report in the form above. After clicking "Submit", the system will begin analyzing your data. This process may take a few moments, so we ask for your patience. The progress bar will display the analysis progress. Once the analysis is complete, the diagnosis report will be displayed in the results box on the right.

Disclaimer

This diagnostic system is only a decision support tool and cannot replace professional medical advice. All diagnostic and treatment decisions should be made by medical professionals, and we do not assume any responsibility arising from the use of this system.

Result:

**Figure 3** Interface of ThyroAIGuide platform. ThyroAIGuide is an intuitive online diagnostic tool designed for preliminary thyroid cancer diagnosis. Users are required to input pertinent clinical data and ultrasound image descriptions, after which the platform harnesses the power of the GPT-4 model to generate comprehensive diagnostic reports. These reports include TI-RADS grading, risk assessment, and recommendations on whether a biopsy is necessary. TI-RADS, Thyroid Imaging-Reporting and Data System; GPT, Generative Pre-trained Transformer.

### *Report reproducibility experiment restated*

To assess the consistency of GPT-4 answers, we compared two responses generated by different temporary model instances to analyze their similarities. For each question, we examined the scores assigned to both responses. We then evaluated the statistical differences between the two sets of answers to determine whether there were significant discrepancies between the two experiments. This evaluation provides insight into the consistency of GPT-4 performance when addressing similar diagnosis.

### *Subjective evaluation by doctors*

Subjective evaluation of clinical diagnostic reports by doctors is crucial for assessing AI-generated report effectiveness (17). We collected feedback from doctors, analyzed it, and classified it into positive (praises AI-generated report for accuracy, clarity, or usefulness), negative (highlights concerns, such as inaccuracies, ambiguities, or irrelevance), or neutral (neither praises nor criticizes; offers general observations or suggestions) categories. To better understand the results, we calculated the proportions of each feedback category, shedding light on the overall sentiment towards the AI-generated reports. To complement the analysis, we employed word cloud methods to visualize high-frequency keywords from the doctors' feedback. This graphical representation allowed us to quickly identify common themes and areas of concern or praise in the AI-generated reports (18). This approach enabled comprehensive assessment of the clinical impact of AI-generated reports (AIGC reports). Examining doctors' perspectives helps identify the AI system's strengths and areas needing refinement to better serve medical professionals' needs.

### *GPT Chain of Thought (CoT) visualization*

The CoT method breaks down the decision-making process of the GPT model into several stages, each symbolizing a link in the thought chain (19). These chains were then gathered and depicted as a flowchart, offering a clear and insightful way to scrutinize the model's decision-making patterns. To provide a more detailed walkthrough of the CoT, we took a patient as an example, with the GPT-4 illustrating its thought chain during the diagnostic process. Initially, the CoT tracked how GPT-4 interacted with the given input, following the AI's process of interpreting the ultrasound description and extracting key information. The visual representation of the thought chain elucidates the decisions made by GPT-4, enhancing our understanding of its decision-making process. This transparency of the AI's reasoning process also provides clinicians with valuable insight into the AI's suggestions, fostering trust and better integration of AI advice in the clinical process.

### *Statistical analyses*

The statistical analysis was carried out using the Mann-Whitney *U* test method (20), provided by the scipy package (21), with all code written in Python 3.8. A P value lower than 0.05 was deemed to be statistically significant.

## Results

In this Results section, we summarize our findings on GPT-4's performance in generating thyroid gland diagnosis reports. We assessed its performance through report scoring, comparison with doctors of varying experience levels, a Turing Test, and consistency analysis. The following sections present the outcomes of these evaluations.

### *GPT-4 report generation performance evaluation result*

We evaluated GPT-4's performance in creating Chinese medical reports by having doctors of native speaker analyzing them based on diagnostic accuracy, report organization, utilization of professional terminology, clarity of language, and overall assessment. The results are presented in *Figure 4A*. Based on a 5-point Likert scale, where higher scores indicate better performance, GPT-4 had a diagnostic accuracy mean score of 3.68 [95% confidence interval (CI): 3.52–3.85] and a report structure mean score of 4.28 (95% CI: 4.15–4.40). The AI's mean score for using professional terminology was 4.28 (95% CI: 4.17–4.38), and for clarity of expression, it achieved a mean score of 4.26 (95% CI: 4.16–4.37). In the general evaluation, GPT-4's mean score was 3.80 (95% CI: 3.67–3.94). The AI's performance in report structure, professional terminology, and clarity of expression exceeded a score of 4. The radar chart in *Figure 4B* displays the average indicators for each group. The chart shows that GPT-4 performed particularly well in report structure, professional terminology, and clarity of expression, with scores exceeding 4. However, the diagnostic accuracy score was slightly lower, at 3.68, indicating room for improvement in this area. Overall, GPT-4's general evaluation score was 3.80, suggesting that its performance was generally satisfactory. This visualization allows for a quick and comprehensive understanding of the AI's strengths and areas requiring further development.

### *Turing Test results*

*Figure 5A* presents the proportion of correct judgments made by doctors when assessing the origin of the reports. Doctors assigned scores over 3 to 37% of the reports, indicating a higher likelihood of the report being human-written. On the other hand, 71% of the AI-generated reports received scores over 3, with 60.8% of the scores falling in the highest category (class 5). *Figure 5B* supports this trend, showing the majority of human reports scored in class 1, while AI-generated reports were predominantly scored in class 5. These findings suggest that the AI-generated reports have a high probability of being perceived as human-written.

### *Report reproducibility experiment results*

The box diagram in *Figure 6* displays the distribution of features for assessing the consistency between two sets of answers generated by GPT-4. The distribution outlooks were observed for accuracy, structure, terminology, clarity, and general evaluation. The P values for each aspect were as follows—accuracy: 0.028, structure: 0.095, terminology: 0.110, clarity: 0.060, and general evaluation: 0.019. These P values revealed varying levels of consistency in the AI-generated reports across the accuracy and general evaluation aspects.

### *Subjective feedback*

The analysis of collected feedback (*Figure 7*) showed the proportion of negative comments (65%) to be higher than positive (14%) and neutral (20%) evaluations. The most common positive feedback keywords were "detailed" and "clear", while the prevalent negative feedback terms included "cumbersome", "misdiagnose", and "unclear". The term "professional" frequently appeared in neutral evaluations. Some doctors may have found the AI-generated reports detailed and clear, leading to positive feedback.

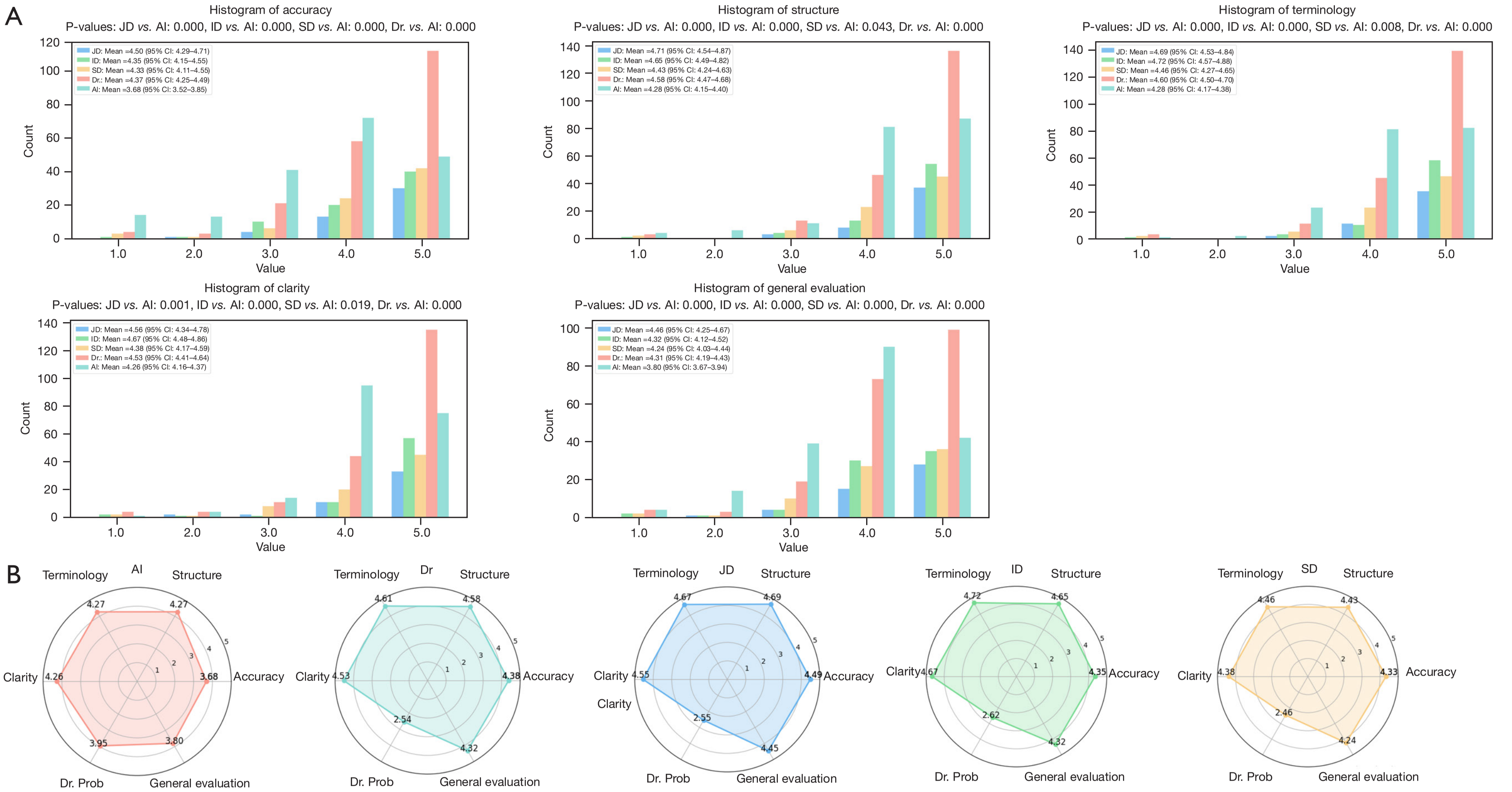


**Figure 4** Report quality and accuracy comparison. (A) Distribution of ratings for accuracy and other evaluation metrics across JD, ID, SD, Dr., and AI. (B) Radar chart presenting the average ratings for each evaluation metric among different levels of doctors and AI-generated reports. JD, junior doctors; ID, intermediate doctors; SD, senior doctors; Dr., the overall doctor cohort; AI, GPT-4 generated reports; CI, confidence interval; AI, artificial intelligence; GPT, Generative Pre-trained Transformer.

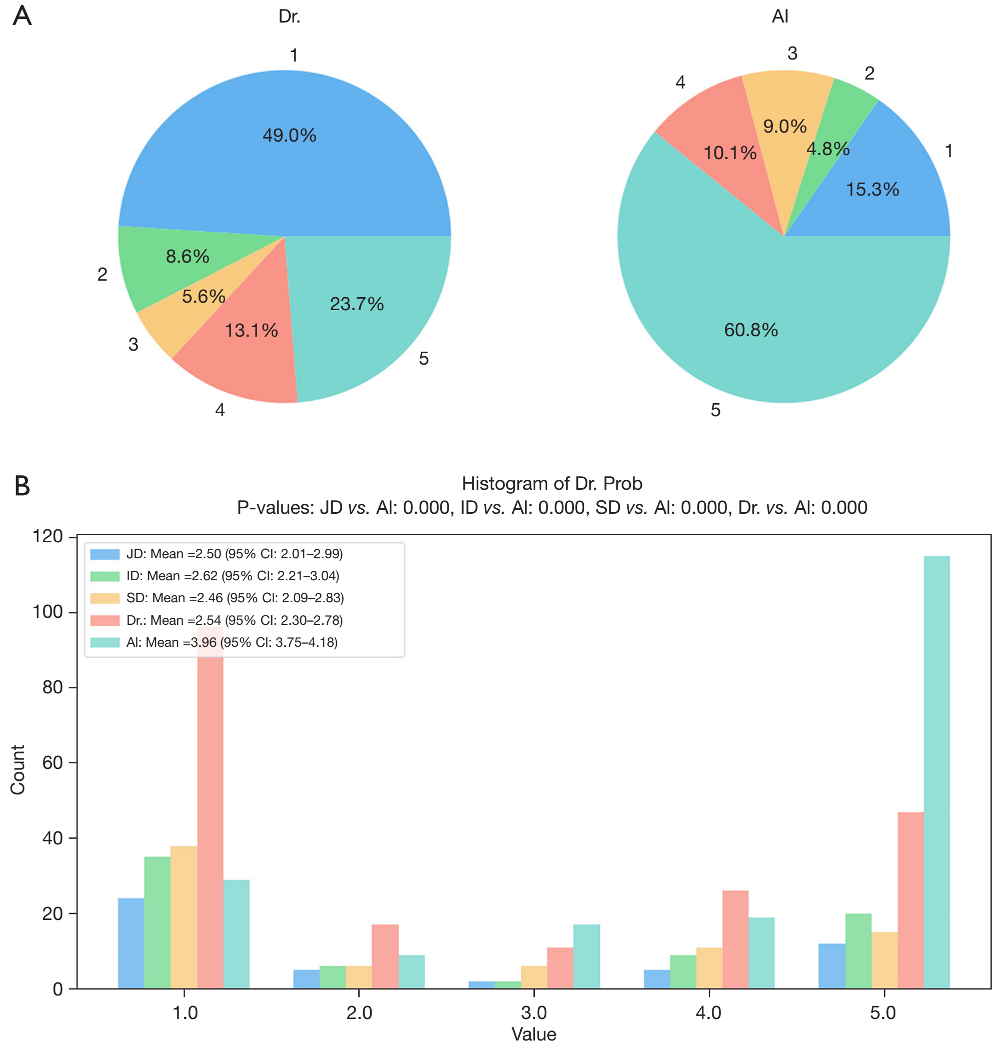


**Figure 5** Perceived human authorship assessment. (A) Pie chart depicting the proportion of Turing Test scores for AI-generated reports and doctor-authored reports. (B) Histogram showing the distribution of probabilities that reports were assessed as written by doctors across JD, ID, SD, and Dr. 1–5 scores present the likelihood levels assigned by doctors, ranging from 'strongly AI-generated' [1] to 'strongly human-written' [5]. JD, junior doctors; ID, intermediate doctors; SD, senior doctors; Dr., the overall doctor cohort; AI, GPT-4 generated reports; CI, confidence interval; AI, artificial intelligence; GPT, Generative Pre-trained Transformer.

However, others might have encountered cumbersome or unclear sections, causing them to provide negative feedback. Additionally, some doctors might have focused on the professional terminology used, contributing to the neutral evaluations.

### *Visualization CoT result*

The CoT method was applied in three steps to assess thyroid nodule malignancy risk and propose treatment recommendations (*Figure 8*). In Step 1, the model extracted and scored data about the thyroid nodules using the TI-RADS standard. In Step 2, it combined demographic data and lymph node information with the TI-RADS score to calculate an overall cancer risk score. In Step 3, the model matched the patient's risk level with treatment guidelines to propose a suitable treatment plan. Additional factors like medical history, nodule count and location, and symptoms were also considered for a comprehensive risk assessment. The CoT method proved effective in enhancing AI decision-making interpretability in this medical context.

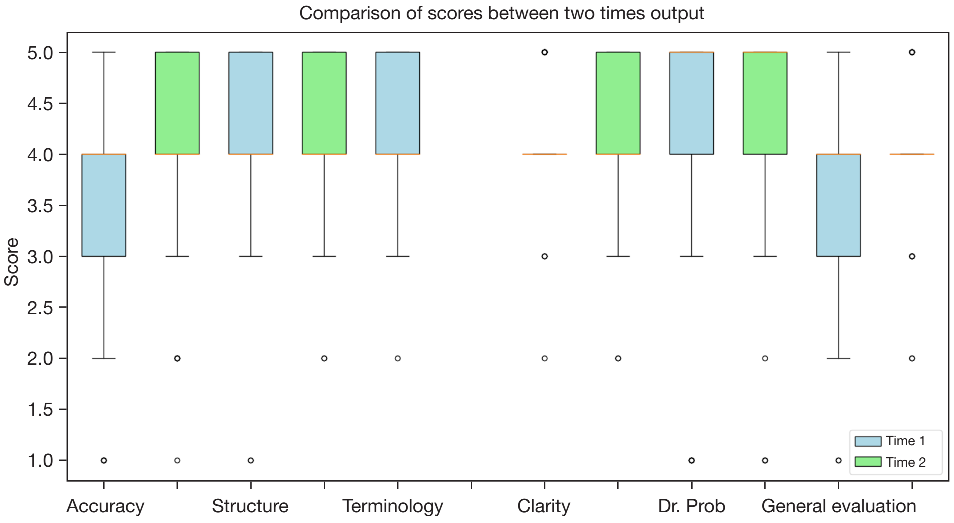


**Figure 6** Distribution of GPT-4 generated report scores. Boxplot showing the distribution of scores for reports generated by GPT-4 for the same patient at different time windows. P value differences are as follows—accuracy: 0.02801, structure: 0.09455, terminology: 0.10964, clarity: 0.06006, and general evaluation: 0.01931. Dr. Prob, probabilities that reports were assessed as written by doctors; GPT, Generative Pre-trained Transformer.

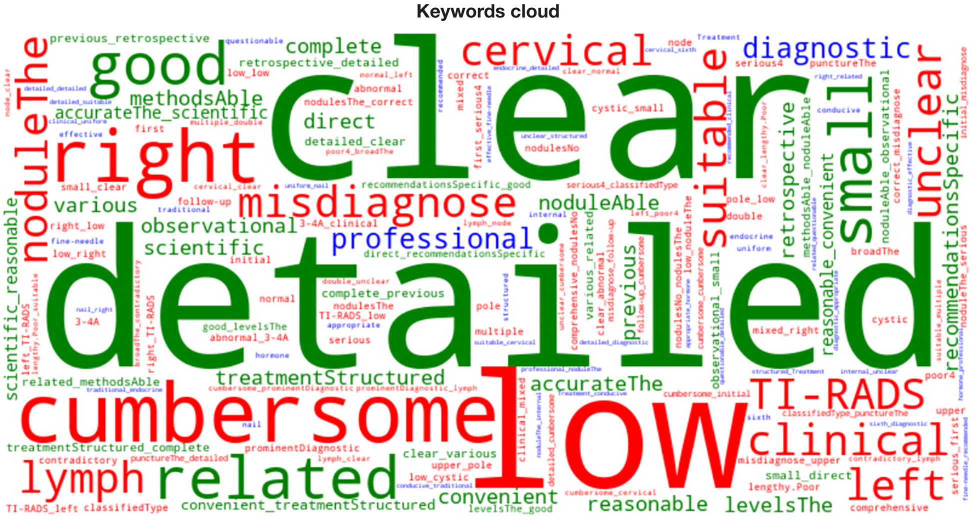


**Figure 7** Doctors' subjective feedback word cloud. Green represents positive feedback, red indicates negative feedback, and blue signifies neutral feedback. Font size corresponds to frequency of occurrence.

## Discussion

In this study, we evaluated GPT-4's performance in generating thyroid gland diagnosis reports, focusing on report scoring, quality comparison by doctors with different experience levels, Turing Test results, and consistency analysis. Our findings provide valuable insights into GPT-4's current capabilities, as well as areas for potential improvement and practical applications in the medical field (22,23).

In fact, due to their complex architecture and opaque operations, large language models like GPT are often described as "black boxes" (24). This opacity becomes particularly challenging in the field of medical decision-making, where the interpretability of a model is crucial for gaining the trust of clinicians and effectively integrating AI

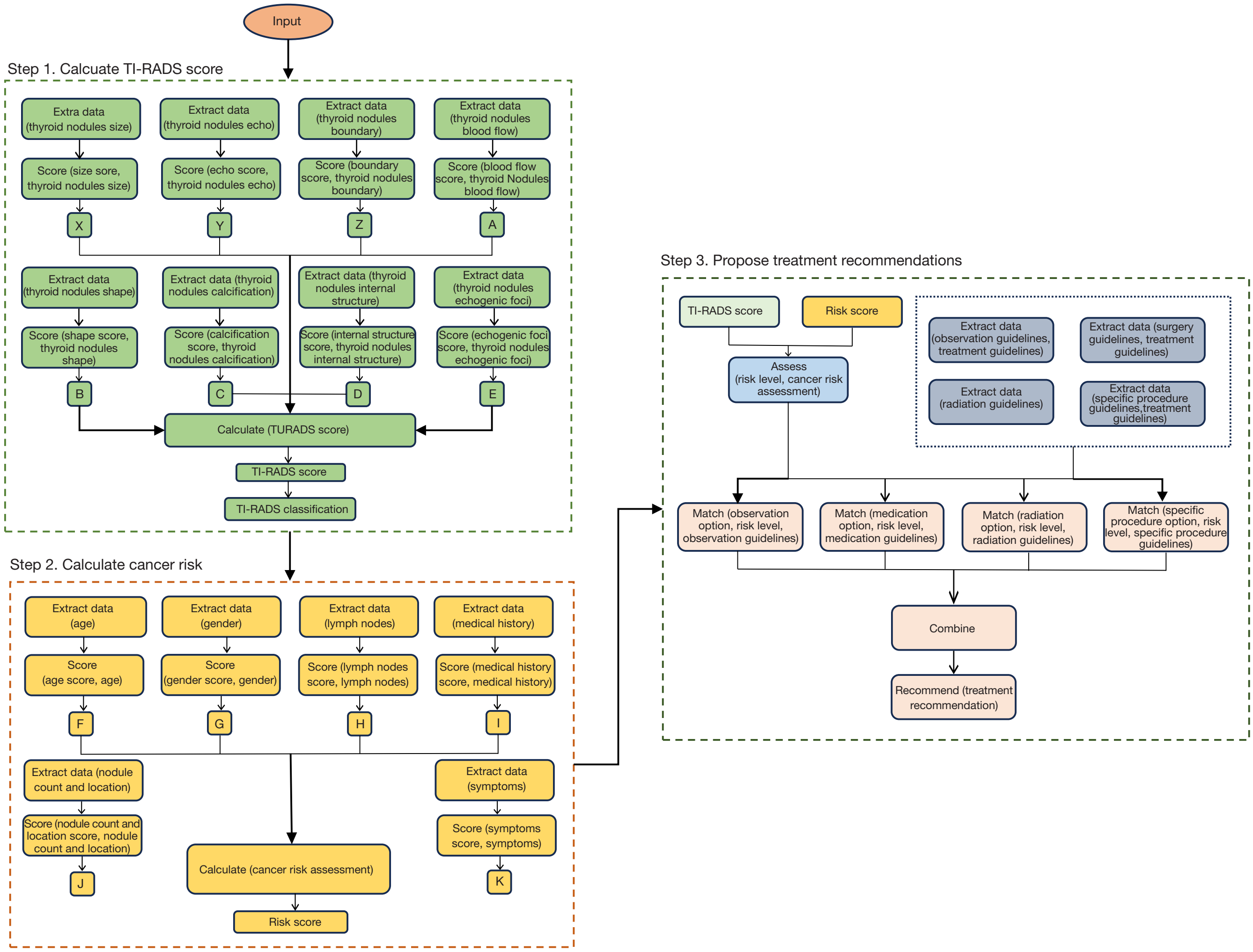

Input
Step 1. Calcuate TI-RADS score
Extra data (thyroid nodules size)
Extract data (thyroid nodules echo)
Extract data (thyroid nodules boundary)
Extract data (thyroid nodules blood flow)
Score (size sore, thyroid nodules size)
Score (echo score, thyroid nodules echo)
Score (boundary score, thyroid nodules boundary)
Score (blood flow score, thyroid Nodules blood flow)
X
Y
Z
A
Extract data (thyroid nodules shape)
Extract data (thyroid nodules calcification)
Extract data (thyroid nodules internal structure)
Extract data (thyroid nodules echogenic foci)
Score (shape score, thyroid nodules shape)
Score (calcification score, thyroid nodules calcification)
Score (internal structure score, thyroid nodules internal structure)
Score (echogenic foci score, thyroid nodules echogenic foci)
B
C
D
E
Calculate (TURADS score)
TI-RADS score
TI-RADS classification
Step 2. Calculate cancer risk
Extract data (age)
Extract data (gender)
Extract data (lymph nodes)
Extract data (medical history)
Score (age score, age)
Score (gender score, gender)
Score (lymph nodes score, lymph nodes)
Score (medical history score, medical history)
F
G
H
I
Extract data (nodule count and location)
Extract data (symptoms)
Score (nodule count and location score, nodule count and location)
Score (symptoms score, symptoms)
J
K
Calculate (cancer risk assessment)
Risk score
Step 3. Propose treatment recommendations
TI-RADS score
Risk score
Assess (risk level, cancer risk assessment)
Extract data (observation guidelines, treatment guidelines)
Extract data (surgery guidelines, treatment guidelines)
Extract data (radiation guidelines)
Extract data (specific procedure guidelines,treatment guidelines)
Match (observation option, risk level, observation guidelines)
Match (medication option, risk level, medication guidelines)
Match (radiation option, risk level, radiation guidelines)
Match (specific procedure option, risk level, specific procedure guidelines)
Combine
Recommend (treatment recommendation)

**Figure 8** CoT visualization for GPT model. There are three main steps: Step 1, calculate TI-RADS score and classification; Step 2, calculate cancer risk; Step 3, propose treatment recommendations. TI-RADS, Thyroid Imaging-Reporting and Data System; CoT, Chain of Thought; GPT, Generative Pre-trained Transformer.

suggestions into diagnostic and treatment plans. To address this issue, we adopted the CoT method (19), a technique aiming at enhancing the transparency and interpretability of AI model reasoning processes. We uniquely applied the CoT method to deconstruct the decision-making chain of the GPT model in diagnosing thyroid cancer cases. This allows us to track, elucidate, and understand the decision-making process and reasoning logic of the GPT model step by step, thereby revealing the "black box" of AI and systematically dissecting the reasoning path of AI to enhance the interpretability of the model.

### *GPT-4 report generation performance*

The assessment of GPT-4's performance in generating medical reports reveals both strengths and areas for improvement spanning various aspects of report generation. GPT-4 demonstrated strong capabilities in report structuring, use of professional terminology, and clarity of expression, with scores exceeding 4 in each of these categories. However, GPT-4's diagnostic accuracy was significantly lower than that of human doctors, as evidenced by the P values and a mean score of 3.68. While the AI showed some diagnostic capability with an 85% accuracy rate for generated answers with scores ≥3, it still falls short of the performance of human doctors in this crucial aspect of medical report generation (25). Conversely, GPT-4 excels in report structuring, using professional terminology, and ensuring clarity of expression: the AI can produce well-structured reports with appropriate professional language and clear communication, which are essential components of effective medical reporting (26).

### *Turing Test*

GPT-4 excels at generating medical reports resembling human-written ones, with potential applications in diagnostics, telemedicine, and patient education. However, medical professionals should monitor AIGC for patient safety and care quality. GPT-4's performance, though impressive, still reveals occasional discrepancies, such as overly detailed reports. Further development can refine the language model for more accurate and nuanced outputs, increasing the difficulty of distinguishing between human and AI-generated reports (27).

### *Report reproducibility*

The report reproducibility experiment results shed light on GPT-4's strengths and weaknesses in generating medical reports. GPT-4 demonstrates consistent performance in producing reports with coherent structure, appropriate professional terminology, and clear expression. However, the AI encounters challenges in maintaining diagnostic accuracy and overall evaluation consistency, as evidenced by statistically significant P values below 0.05 for both aspects. These findings emphasize the need to refine GPT-4's capabilities, focusing on enhancing the reliability and precision of its generated reports for better utility in medical practice (28).

### *Subjective feedback*

Feedback analysis highlights areas for improvement to better meet medical professionals' needs and expectations. Although GPT-4 excels in generating detailed and clear reports, the prevalence of negative comments suggests challenges remain. Terms like "cumbersome", "benign", and "unclear" imply reports may be overly complex, hard to interpret, or lack clarity. Addressing these concerns is vital for enhancing report quality and utility. The term "professional" in neutral evaluations indicates GPT-4's acceptable professionalism. Refining performance in areas identified by negative feedback will strengthen GPT-4's role in the medical field, promoting efficient and accurate diagnosis processes (29).

### *GPT-4 CoT*

As demonstrated in this study, the CoT method offers a systematic and transparent approach to understanding the decision-making process of AI models. This method is particularly valuable in a clinical setting where interpretability and trust are crucial for integrating AI into healthcare. From a clinical perspective, the CoT method allows for a comprehensive assessment of thyroid cancer risk. By considering a wide range of factors, including nodule characteristics, patient demographics, lymph node information, as well as other factors such as medical history, nodule count and location, and symptoms, the model can provide a more comprehensive and personalized risk assessment. This could lead to more accurate diagnoses and more targeted treatment plans, thereby improving patient outcomes. Moreover, the CoT method highlights the potential of AI in enhancing the diagnostic process. By automating the extraction and scoring of data and the calculation of risk scores, the model can save clinicians' time and reduce the possibility of human error. This could be particularly beneficial in high-volume or resource-limited settings where efficiency is paramount. In terms of insights, this study showcases the potential of AI, particularly large language models like GPT-4, in the field of medical diagnosis. However, it also emphasizes the importance of interpretability in these models. As AI continues to advance and become increasingly integrated into healthcare, methods like CoT are crucial for ensuring these models are transparent, trustworthy, and clinically useful (30).

### *Clinical relevance*

Addressing the rising prevalence of thyroid nodules and the need for accurate ultrasound-based diagnoses is crucial, especially given the variability in diagnostic accuracy and repeatability due to doctors' differing experience levels. This study investigates GPT-4's potential in the medical field, with a focus on thyroid cancer ultrasound report analysis, to enhance diagnostic outcomes by supporting human expertise. We examine GPT-4's capabilities in improving diagnostic accuracy, efficiency, and minimizing human error in the diagnostic process. A clear use case for

GPT-4 in this setting is assisting radiologists in reviewing and interpreting ultrasound description, providing a second opinion to support decision-making. This AI-assisted approach could foster collaboration among healthcare teams, facilitate continuous learning, and improve healthcare quality and accessibility in remote or underserved areas. Through this investigation, we aim to contribute to the development of a reliable, accessible, and efficient diagnostic tool for clinical practice, ultimately benefiting healthcare professionals and patients by promoting better patient care and outcomes (31). Furthermore, the integration of GPT-4 and the CoT method into the clinical process presents valuable opportunities for improving the diagnostic process. By providing a second opinion in the interpretation of ultrasound descriptions, GPT-4 can aid in identifying critical details that may be overlooked and in preventing diagnostic errors. Meanwhile, the CoT method provides a comprehensive insight into GPT-4's decision-making process. Understanding how the AI arrived at its conclusions not only enables clinicians to verify the AI's recommendations but also to incorporate AI insights into their diagnosis more confidently. This transparency ultimately leads to better-informed medical decisions, fostering trust in AI-assisted diagnostic processes. Additionally, understanding the AI's thought process could provide a basis for continuous learning and improvement, facilitating the development of more refined diagnostic strategies (32).

### *Online platform*

The proposed online platform "ThyroAIGuide", leveraging AI capabilities, holds substantial potential to transform healthcare delivery, especially in diagnostics. Its accessibility allows patients from remote or underserved areas to access preliminary diagnostic information anytime, anywhere, thereby identifying those requiring urgent care and guiding them towards appropriate resources. The platform's efficiency lies in its automation of data extraction, scoring, and risk calculation, providing rapid diagnostic insights that save time for both patients and healthcare providers, and potentially hastening the initiation of suitable treatment. Furthermore, the AI-powered platform offers consistency in evaluations, reducing variability that can occur with different healthcare providers, which is particularly beneficial in complex or subjective areas of medicine like imaging study interpretation. Additionally, the platform serves as an educational tool, helping patients better understand their condition and risk factors, thereby empowering them to take a more active role in their healthcare and make informed treatment decisions.

### *Limitations and future directions*

This study highlights GPT-4's limitations in thyroid diagnosis and suggests future research on specialized models, advanced techniques, and medical partnerships to improve performance. Investigating the AI's comprehension of medical terminology and context is crucial. One of the more tangible concerns raised is the clinical utility of AI-generated reports. While the system may reduce the time required to generate initial drafts of reports, the added time for proofreading, especially given the model's current inaccuracies, often offsets this advantage. This dual-edged nature of AI assistance—where time saved in one aspect might be spent in another—needs to be carefully addressed in subsequent versions and applications of the model. Enhancing accuracy is not merely for the sake of reducing errors but also to genuinely save time and effort for medical professionals. Future work should also consider image analysis, time-cost efficiency, and specific diagnostic models to provide a comprehensive evaluation and develop more precise, efficient tools for medical professionals. Our study did not include the process of structuring or digitizing the free text data. Instead, we directly generated results without clear analytical processes, resembling black boxes, which may have led to lower clinical credibility. Moreover, as we transition into a more AI-assisted clinical landscape, it is paramount to ensure that these "black box" models gain the trust of the medical community. Transparency in how the AI reaches its conclusions or the potential integration of explainable AI mechanisms can be pivotal in future iterations. The ultimate goal is to align the efficiency gains from AI with the rigorous demands and standards of medical practice. In future studies, we aim to explore the potential application of GPT in structuring and organizing the data, addressing the limitations identified in our current research.

## Conclusions

This study underscores GPT-4's potential as an auxiliary tool in healthcare, particularly in generating thyroid gland diagnosis reports. Despite its proficiency in report structuring and clarity of expression, the need for further refinement is evident due to limitations in diagnostic accuracy. The proposed online platform "ThyroAIGuide" illustrates the potential of AI in enhancing diagnostic processes and improving healthcare accessibility. The CoT method, applied in this study, provides a systematic and transparent approach to understanding the decision-making process of AI models, thereby addressing the 'black box' nature of such large models. However, the journey towards fully integrating AI into healthcare is ongoing, requiring continuous research, development, and careful monitoring by medical professionals to ensure patient safety and quality of care.

## Acknowledgments

*Funding:* This work was supported by the Capital's Funds for Health Improvement and Research (No. 2022-1-2022).

## Footnote

*Conflicts of Interest:* All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-1180/coif). A.D. reports honoraria for consultancy from the following companies Varian, Janssen, Philips, BMS, Mirada Medical, Medical Data Works B.V. These conflicts of interest did not interfere with the submitted publication. The other authors have no conflicts of interest to declare.

*Ethical Statement:* The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Ethical approval for this study was obtained from the Medical Ethics Committee of Beijing Friendship Hospital, Capital Medical University. This was not a study of human subjects. Therefore, informed consent was not required.

## References

1. Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The Diagnosis and Management of Thyroid Nodules: A Review. JAMA 2018;319:914-24.
2. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest 2009;39:699-706.
3. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, Schuff KG, Sherman SI, Sosa JA, Steward DL, Tuttle RM, Wartofsky L. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid 2016;26:1-133.
4. Moon WJ, Baek JH, Jung SL, Kim DW, Kim EK, Kim JY, Kwak JY, Lee JH, Lee JH, Lee YH, Na DG, Park JS, Park SW; Korean Society of Thyroid Radiology (KSThR); Korean Society of Radiology. Ultrasonography and the ultrasound-based management of thyroid nodules: consensus statement and recommendations. Korean J Radiol 2011;12:1-14.
5. OpenAI TJO. Chatgpt: Optimizing language models for dialogue. 2022. Available online: https://autogpt.net/chatgpt-optimizing-language-models-for-dialogue/
6. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484-9.
7. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in Neural Information Processing Systems 2020;33:1877-901.
8. Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, Lally A, Murdock JW, Nyberg E, Prager J, Schlaefer N, Welty C. Building Watson: An overview of the DeepQA project. AI Magazine 2010;31:59-79.
9. Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, Guez A, Lockhart E, Hassabis D, Graepel T, Lillicrap T, Silver D. Mastering Atari, Go, chess and shogi by planning with a learned model. Nature

2020;588:604-9.

10. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med 2019;380:1347-58.
11. Koubaa A. GPT-4 vs. GPT-3.5: A Concise Showdown. Preprints 2023, 2023030422. Available online: https://doi.org/10.20944/preprints202303.0422.v1
12. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 2018;22:1589-604.
13. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nat Med 2019;25:24-9.
14. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017;2:230-43.
15. Grinberg M. Flask web development: developing web applications with python[M]. O'Reilly Media, Inc.; 2018.
16. French RM. The Turing Test: the first 50 years. Trends Cogn Sci 2000;4:115-22.
17. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309:1351-2.
18. He W, Zha S, Li L. Social media competitive analysis and text mining: A case study in the pizza industry. Int J Inf Manage 2013;33:464-72.
19. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Part of Advances in Neural Information Processing Systems 35 (NeurIPS 2022); 2022.
20. McKnight PE, Najab JJTCeop. Mann-Whitney U Test. 2010:1-1. Available online: https://doi.org/10.4135/9781412961288.n228
21. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261-72.
22. Gebrael G, Sahu KK, Chigarira B, Tripathi N, Mathew Thomas V, Sayegh N, Maughan BL, Agarwal N, Swami U, Li H. Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0. Cancers (Basel) 2023.
23. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021;21:125.
24. Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Can J Cardiol 2022;38:204-13.
25. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer? Urology 2023;180:35-58.
26. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, Thun S. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform 2022;10:e35724.
27. Myong Y, Yoon D, Kim BS, Kim YG, Sim Y, Lee S, Yoon J, Cho M, Kim S. Evaluating diagnostic content of AI-generated chest radiography: A multi-center visual Turing test. PLoS One 2023;18:e0279349.
28. Sohn E. The reproducibility issues that haunt health-care AI. Nature 2023;613:402-3.
29. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, McCoy AB, Sittig DF, Wright A. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30:1237-45.
30. Joyce DW, Kormilitzin A, Smith KA, Cipriani A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. NPJ Digit Med 2023;6:6.
31. Frasca F, Piticchio T, Le Moli R, Tumino D, Cannavò S, Ruggeri RM, Campennì A, Giovanella L. Early detection of suspicious lymph nodes in differentiated thyroid cancer. Expert Rev Endocrinol Metab 2022;17:447-54.
32. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial Intelligence Applications in Health Care Practice: Scoping Review. J Med Internet Res 2022;24:e40238.

**Cite this article as:** Wang Z, Zhang Z, Traverso A, Dekker A, Qian L, Sun P. 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 2024;14(2):1602-1615. doi: 10.21037/qims-23-1180

## Appendix 1 "ThyroAIGuide" Platform Details

We have developed an intelligent diagnostic platform named "ThyroAIGuide" using the Flask framework to translate our research into a practical application. ThyroAIGuide is a user-friendly online tool designed to provide a preliminary diagnosis of thyroid disorders, making it accessible to a wide range of users.

The interaction with the platform is straightforward. Users are required to input relevant clinical information and descriptions of ultrasound images. Upon receiving this data, the platform leverages the advanced capabilities of the GPT-4 model to analyze the information.

The main functionality of ThyroAIGuide lies in its ability to generate comprehensive diagnostic reports. These reports include an assessment of the thyroid nodule's size and characteristics, a risk assessment of thyroid cancer, and recommendations treatment plan, all derived from the user-provided information.

The practical application of ThyroAIGuide is significant. It serves as a valuable tool in the preliminary diagnosis of thyroid disorders, offering insights that can guide subsequent medical consultations and decisions. This platform, therefore, stands as a testament to the potential of AI in enhancing healthcare delivery and patient outcomes.

## Appendix 2 Score Table

The evaluation form we developed aims to assess the quality of AI-generated medical reports, focusing on accuracy, structure, terminology, clarity, doctor-like writing probability, and overall evaluation. To construct this form, we gathered data on variables such as ID, gender, age, reason for consultation, description, and ultrasound conclusion, which serve as the foundation for evaluating the generated reports.

Intended for use by medical professionals (e.g., doctors or radiologists), the form enables them to rate the generated reports based on the specified criteria. Each criterion receives a numerical score, and a section for subjective comments allows evaluators to offer additional feedback or insights.

Utilizing this evaluation form, our goal is to better understand the strengths and weaknesses of AI-generated medical reports, including those produced by GPT-4. The feedback collected through this form can be employed to enhance the AI model's performance, increasing its reliability and accuracy in generating medical reports.

The score table for assessing the generated report is shown below in *Table S1*.

Accuracy:

a. Please rate the diagnostic accuracy of the report (1 point = completely incorrect, 5 points = completely correct)

Structure:

a. Please rate the structure of the report (1 point = very poor, 5 points = excellent)

Terminology:

a. Please rate the use of professional terminology in the report (1 point = very poor, 5 points = excellent)

Clarity:

a. Please rate the clarity of expression in the report (1 point = very poor, 5 points = excellent)

Dr.Prob:

a. Please rate the likelihood that the report was written by a doctor (1 point = very low, 5 points = very high)

General evaluation:

a. Please provide an overall rating for the entire report (1 point = very poor, 5 points = excellent)

**Table S1** The score table for assessing the generated report

| ID | Gender | Age | Reason for Consultation | Description | Ultrasound Conclusion | Accuracy | Structure | Terminology | Clarity | Dr.Prob | General evaluation | Subjective comment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | |

https://dx.doi.org/10.21037/qims-23-1180