Supplementary

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