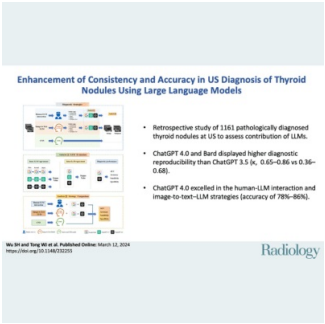
HealthManagement
Promoting Management and Leadership

Advancing US Diagnosis of Thyroid Nodules Using Large Language Models



The rapid progress of large language models (LLMs) presents significant opportunities. LLMs, like ChatGPT, possess capabilities in language understanding and generation, which can be applied to tasks such as translation, question answering, and diagnostic advice in healthcare. In medical image analysis, combining LLMs with image-to-text models offers the potential for more precise and tailored medical services. However, understanding the strengths and limitations of LLMs is crucial. Current discussions mostly focus on language generation capabilities, lacking comprehensive evaluations against objective standards like pathologic examinations. Concerns also arise regarding biases perpetuated by LLMs. A recent paper published in the journal Radiology aims to assess the utility of LLMs in diagnosing thyroid nodules based on standardised reporting, potentially reducing interpretational variability and workload for radiologists.

Assessment of Large Language Models in Diagnosing Thyroid Nodules

This retrospective study, approved by the medical ethics committee of the First Affiliated Hospital of Sun Yat-sen University, assessed the agreement and diagnostic performance of three large language models (LLMs)—OpenAI's ChatGPT 3.5, ChatGPT 4.0, and Google's Bard (later renamed Gemini)—in diagnosing thyroid nodules based on TI-RADS criteria. The study also evaluated three model deployment strategies: human-LLM interaction, image-to-text–LLM, and convolutional neural network (CNN). US images of thyroid nodules from 725 patients (mean age, 42.2 years ± 14.1 [SD]; age range, 20–71 years; 516 women) were collected, with definitive pathologic results confirming 498 benign and 663 malignant cases. Each image represented an independent thyroid nodule, and all lesions were confirmed based on histopathologic results.

Enhancing Consistency and Accuracy in Thyroid Nodule Diagnosis

The study aimed to integrate large language models (LLMs) into the diagnostic process for thyroid nodules to improve consistency and accuracy. ChatGPT 4.0 and Bard showed substantial to almost perfect intra-LLM agreement, while ChatGPT 3.5 exhibited fair to substantial agreement. In terms of deployment strategies, ChatGPT 4.0 demonstrated higher accuracy and sensitivity compared to Bard, with an accuracy range of 78%–86% and a sensitivity range of 86%–95%. The image-to-text–LLM strategy with ChatGPT 4.0 performed similarly to human-LLM interaction involving two senior readers and one junior reader, and better than human-LLM interaction involving only one junior reader.

Exploring the Feasibility and Implications of Large Language Models in Medical Diagnosis

The study highlighted the scarcity of research on LLMs' feasibility in handling reasoning questions associated with medical diagnosis using a reference standard like pathology. The authors evaluated reproducibility, showing ChatGPT 3.5 had the poorest reproducibility among the three LLMs. The study emphasized understanding the variations among LLMs' responses for broader implications and potential applications. It was noted that LLMs possess emotional intelligence, potentially aiding in patient equity, although studies on their logical reasoning ability are rare. While LLMs cannot interpret images directly, they enhance diagnostic understanding and provide transparency in decision-making processes. The study underscored the importance of evaluating LLMs against a reference standard like pathology, which diverges from previous research relying on subjective judgments. Although LLMs combined with TI-RADS signs were less effective than CNN models in diagnosing thyroid nodules, they provided clearer insights into diagnostic steps, essential for trust and medical education.

Despite its limitations, including a focus solely on TI-RADS signs for diagnosis and potential inaccuracies in the voting mechanism, the study demonstrated the potential of combining LLMs with image-to-text approaches to enhance medical imaging and diagnostics. ChatGPT 4.0 emerged as the optimal choice for consistency and diagnostic accuracy. The study suggested that integrating image-to-text models and LLMs could advance medical imaging and diagnostics, informing secure deployment for enhanced clinical decision-making. Further research is warranted to explore applicability across different models, techniques, and medical image types.

Source & Image Credit: RSNA Radiology

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.

Published on : Tue, 19 Mar 2024

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.