Poster No.
PW 10.09

International Surgical Week
The World's Congress of Surgery
isw2024.org
Kuala Lumpur, Malaysia
25-29 August 2024

Can AI Large Language Models (LLMs) Provide Accurate Information for the Management of Thyroid Disease?

Rajam Raghunathan, MD, PhD,[1] Anna R. Jacobs MD, MBA,[2] Jason Prescott, MD,[1] John Allendorf, MD,[2] Kepal N. Patel, MD,[1] Insoo Suh, MD[1]

[1] NYU Grossman School of Medicine; [2] NYU Grossman Long Island School of Medicine

NYU Langone Health

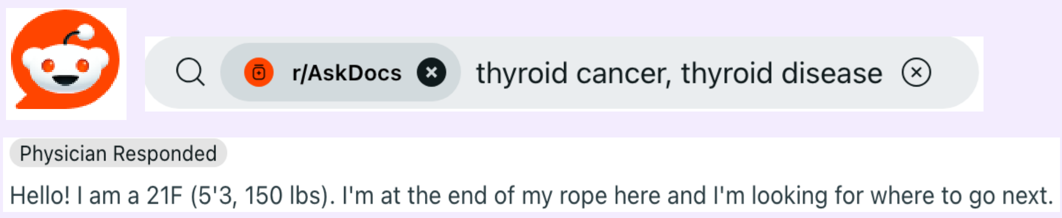
INTRODUCTION

Large language models (LLMs), like ChatGPT and GPT-4, **are becoming first-line sources of information and decision-guidance in the diagnosis and management of thyroid disease but their accuracy, quality, and reliability of their recommendations remains unknown.** This cross-sectional study assesses GPT-4's recommendations for the diagnosis and management of thyroid disease.

METHOD

Part I: Diagnosis and Management Questions

33 randomly selected patient-questions were sourced from an online forum (Reddit/askdocs) using a "thyroid+disease" and "thyroid+cancer" search.


r/AskDocs thyroid cancer, thyroid disease

Physician Responded

Hello! I am a 21F (5'3, 150 lbs). I'm at the end of my rope here and I'm looking for where to go next.

- All questions had responses provided by site-verified physicians and additional responses were generated using a fresh session of GPT-4.
- *Responses were randomized, anonymized and graded by 7 blinded healthcare providers with expertise in endocrine disease* on a 4-point Likert scale based on:

Accuracy	**Quality**
Dangerous and false information	Irrelevant response that does not answer the question
Less than 50% true information	Response partially answers the question
Greater then 50% true information	Response completely answers the question
Completely accurate information	Response provides additional information beyond what was asked

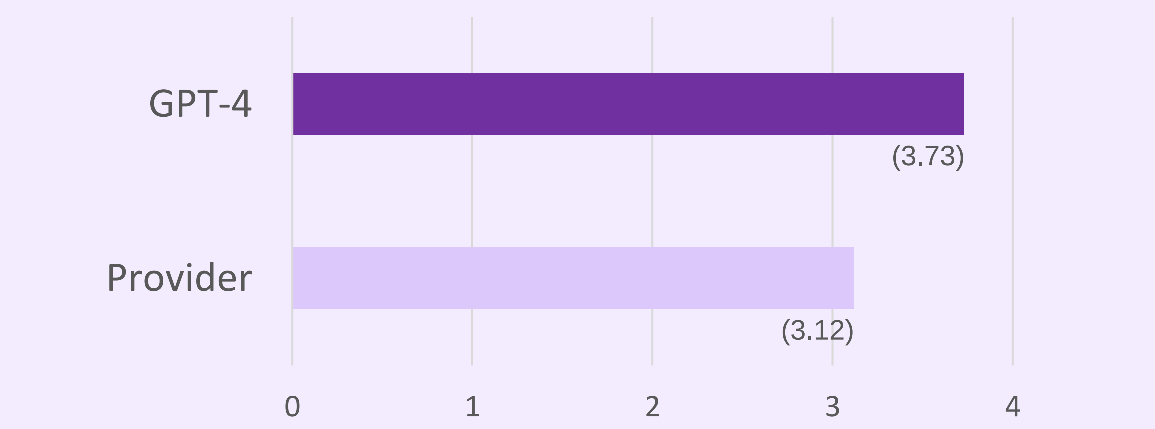
Part II: Treatment Recommendations

- Thyroid cancer diagnoses specifying cancer subtype (PTC, FTC, medullary, anaplastic) and disease stage (DTC: I-IVB, medullary I-IVC, anaplastic I-III) were submitted to a fresh session of GPT-4 using a standard prompt: e.g. "For Stage I Papillary Thyroid Cancer, what is the best treatment?"
- Treatment recommendations made by GPT-4 were evaluated by 5 blinded providers with expertise in endocrine disease for accuracy and quality.
- Part I and Part II results were analyzed using single-factor ANOVA; a t-test was also used for Part II results.

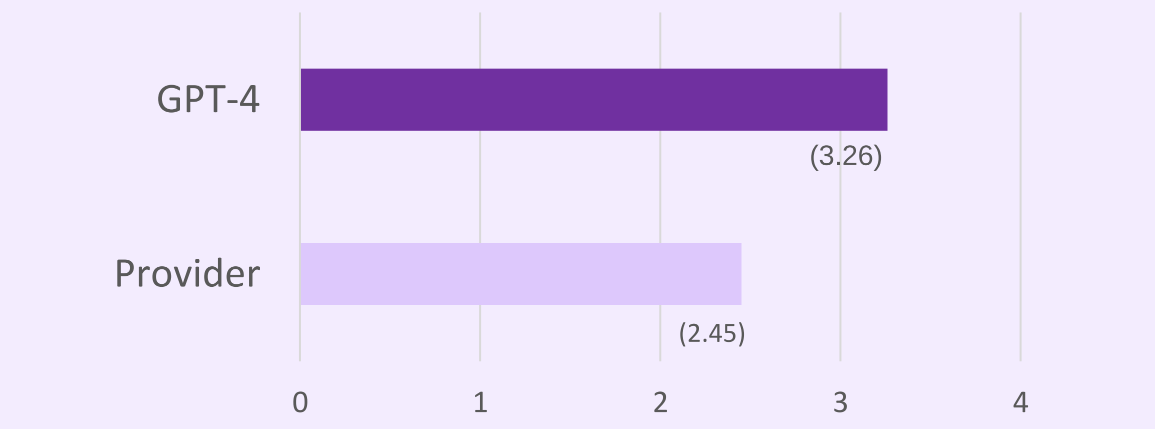
RESULTS

Part I: Diagnosis and Management Questions

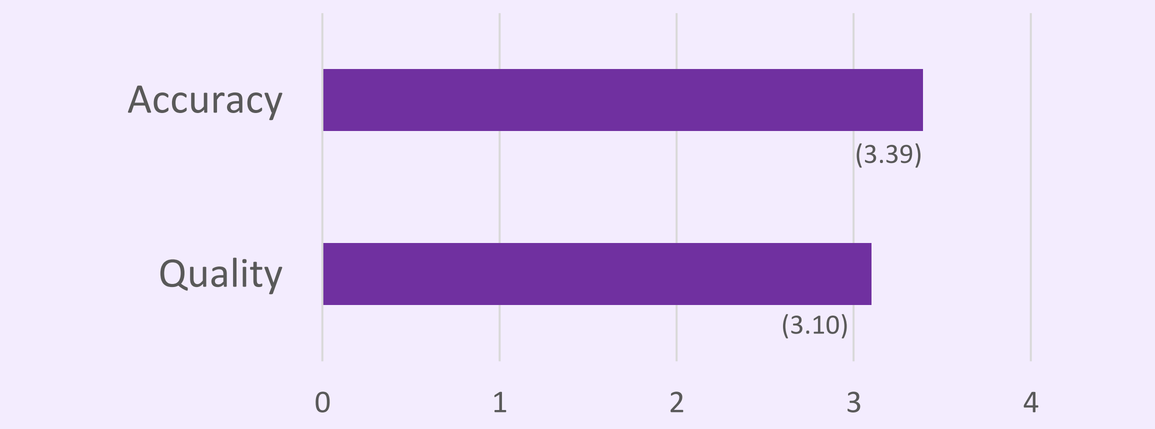
Accuracy: GPT-4 Scores Significantly Higher (p<0.01)



Quality: GPT-4 Scores Significantly Higher (p<0.01)



Part II: GPT-4 Treatment Recommendations



"Dangerous and False Information"

- Online physician responses to patient diagnosis and management questions contained "dangerous and false information" 11% of the time vs. 1% of GPT-4 responses.
- "Dangerous and false information was identified in 0% of GPT-4 treatment recommendation responses

CONCLUSIONS

- **LLM responses** to queries about thyroid disease diagnosis and **management were more accurate and complete than online physician responses**.
- LLM thyroid cancer **treatment recommendations were rated as consistent with guidelines and complete**.
- The rate of false or dangerous information provided by LLMs was minimal.