



# Accuracy of Large Language Models in Thyroid Nodule-Related Questions Based on the Korean Thyroid Imaging Reporting and Data System (K-TIRADS)

Esat Kaba, Nur Hürsoy, Merve Solak,  
Fatma Beyazal Çeliker

Department of Radiology, Recep Tayyip Erdogan University, Rize, Turkey

**Keywords:** Large language models; Thyroid; Nodules; K-TIRADS

We read with great pleasure the review article “Updated Primer on Generative Artificial Intelligence and Large Language Models in Medical Imaging for Medical Professionals” by Kim et al. [1] which was published online in the *Korean Journal of Radiology* in February. The authors impressively presented a very comprehensive overview of generative artificial intelligence, and also discussed the background and working principles of large language models (LLMs). Inspired by this article, we would like to present this letter, in which we investigate the performance of LLMs on questions related to thyroid nodules based on the Korean Thyroid Imaging Reporting and Data System (K-TIRADS).

K-TIRADS was most recently updated in 2021 and consists of consensus recommendations for imaging-based management of thyroid nodules compiled by the Korean

Society of Thyroid Radiology [2]. The latest update includes significant revisions in biopsy criteria, ultrasound (US) criteria for extrathyroidal extension, thyroid computed tomography protocol, and recommendations for US follow-up of thyroid nodules [2]. To evaluate the accuracy and reliability of LLMs’ knowledge regarding K-TIRADS, we prepared 15 multiple-choice questions based on the latest version of K-TIRADS (Supplement). We used Open AI’s ChatGPT-3.5 and 4 (<https://chat.openai.com>), Google’s Gemini (<https://gemini.google.com/app>), and Perplexity (<https://www.perplexity.ai/>) chatbots with default parameters in March 2024. Our initial prompt was, “As a 25-year highly experienced radiologist, answer questions based on the Korean Society of Thyroid Radiology Thyroid Imaging Reporting and Data System (K-TIRADS); there is only one correct option.” ChatGPT-3.5, ChatGPT-4, Gemini, and Perplexity respectively yielded 73% (11/15), 93% (14/15), 80% (12/15), and 87% (13/15) accuracy. ChatGPT-4 outperformed the other LLMs.

LLMs offer potential benefits in many domains of radiology, including reporting, diagnostic support, and creating educational material for patients [3,4]. Many studies have emphasized that LLMs have the potential to generate patient-friendly language and improve physician-patient communication [5,6]. Our preliminary results indicate that some LLMs also have the potential to provide educational material for patients related to diagnosis and management of thyroid nodules in the future, although large validation studies are needed to test their accuracy.

## Supplement

The Supplement is available with this article at <https://doi.org/10.3348/kjr.2024.0229>.

## Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

## Author Contributions

Conceptualization: all authors. Investigation: all authors. Methodology: all authors. Software: all authors. Writing—original draft: all authors. Writing—review & editing: all authors.

**Received:** March 6, 2024 **Revised:** March 10, 2024

**Accepted:** March 15, 2024

**Corresponding author:** Esat Kaba, MD, Department of Radiology, Recep Tayyip Erdogan University, Vadi Street, No. 9, Rize 53800, Turkey

• E-mail: [esatkaba04@gmail.com](mailto:esatkaba04@gmail.com)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

**ORCID IDs**

Esat Kaba

<https://orcid.org/0000-0001-7464-988X>

Nur Hürsoy

<https://orcid.org/0000-0001-5059-2268>

Merve Solak

<https://orcid.org/0000-0003-3466-7260>

Fatma Beyazal Çeliker

<https://orcid.org/0000-0002-5420-9825>**Funding Statement**

None

**REFERENCES**

1. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. *Korean J Radiol* 2024;25:224-242
2. Ha EJ, Chung SR, Na DG, Ahn HS, Chung J, Lee JY, et al. 2021 Korean thyroid imaging reporting and data system and imaging-based management of thyroid nodules: Korean Society of Thyroid Radiology consensus statement and recommendations. *Korean J Radiol* 2021;22:2094-2123
3. Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. *AJR Am J Roentgenol* 2023;221:373-376
4. Kim S, Lee CK, Kim SS. Large language models: a guide for radiologists. *Korean J Radiol* 2024;25:126-133
5. Haver HL, Gupta AK, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. *Radiol Imaging Cancer* 2024;6:e230086
6. Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, et al. Enhancing patient communication with chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J Am Coll Radiol* 2024;21:353-359