Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.253619

ARTIFICIAL INTELLIGENCE AND INFORMATICS

LETTER TO THE EDITOR

©Copyright 2026 Author(s) - Available online at dirjournal.org.
Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Letter to the Editor: Challenges of applying large language models to image-based interpretation in abdominal radiology

Eren Çamur[1]
Turay Cesur[2]
Yasin Celal Güneş[3]

[1]Ministry of Health Ankara 29 Mayıs State Hospital, Clinic of Radiology, Ankara, Türkiye

[2]Ankara Mamak State Hospital, Clinic of Radiology, Ankara, Türkiye

[3]Kırıkkale High Specialization Hospital, Clinic of Radiology, Kırıkkale, Türkiye

Corresponding author: Eren Çamur

E-mail: eren.camur@outlook.com

Received 24 August 2025; accepted 31 August 2025.

Epub: 04.11.2025

Publication date: 02.11.2026

DOI: 10.4274/dir.2025.253619

Dear Editor,

We read with great interest the study by Elek et al.[1] This pioneering study deserves recognition for moving beyond natural language applications to directly evaluate image-based interpretation, thereby opening an important avenue for translational research in abdominal radiology.

The authors demonstrated that Bing ChatGPT-4 performed well in basic categorization, correctly identifying modality, anatomical region, and imaging plane. However, performance declined significantly when greater domain-specific expertise was required, such as differentiating magnetic resonance imaging pulse sequences, characterizing contrast enhancement, or recognizing pathology. These findings underscore the gap between surface-level pattern recognition and the deeper integrative reasoning central to radiologic diagnosis. Unlike convolutional neural networks or vision transformers trained on pixel-level radiology data, large language models (LLMs) are fundamentally language-based systems. Their image analysis relies on multimodal encoders that generate coarse visual embeddings, which are inadequate for detecting subtle grayscale variations, texture details, or multiparametric signal patterns. Diagnostic interpretation also requires the dynamic assessment of multiphase acquisitions, volumetric navigation across planes, and the integration of clinical metadata—all beyond the scope of single-image input models.

Considering that the authors did not employ radiology-specific pretraining, the favorable results demonstrated by relatively older models may reflect selection bias or stereotypical features (e.g., modality-specific appearances, anatomical landmarks) rather than genuine interpretive capacity. Prior exposure to similar images during training could also explain the model's apparent accuracy. Moreover, due to the black box nature of these systems, performance in image interpretation may vary across contexts. Comparing outcomes with the interpretations of radiologists with differing levels of experience would further clarify the true clinical relevance of such models.

The model's restriction to single-image inputs represents a major limitation, since radiologists rarely interpret isolated slices. Accurate diagnosis typically requires the analysis of image series across multiple planes and sequences. Future model development should therefore focus on sequential or volumetric image ingestion, approximating human workflow more closely.

Another important point is the absence of structured prompts. While this absence provided a fair "out-of-the-box, browser-based" evaluation, it may have underestimated the model's performance. Prior studies show that carefully designed prompts and stepwise reasoning frameworks can markedly enhance accuracy.[2-4] Collaborative efforts between artifical intelligence (AI) scientists and radiologists to develop structured, validated prompt libraries for specific diagnostic tasks may represent a key step forward.

Similarly, while the exclusion of patient history was methodologically justified to isolate image interpretation, it created a synthetic setting that diverges from clinical reality. Radiologists nearly always integrate clinical data into diagnostic reasoning. Excluding such information

You may cite this article as: Çamur E, Cesur T, Güneş YC. Letter to the editor: challenges of applying large language models to image-based interpretation in abdominal radiology. *Diagn Interv Radiol*. 2026;32(6):765-766.

may underestimate the potential of LLMs, which are fundamentally language-processing systems. Ultimately, an effective AI assistant must synthesize both imaging and clinical data.

In conclusion, Elek et al.[1] provide valuable early evidence on the opportunities and limitations of LLMs in abdominal radiology. Their study initiates a critical discussion and highlights important directions for future studies in this emerging field.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Elek A, Ekizalioğlu DD, Güler E. Evaluating Microsoft Bing with ChatGPT-4 for the assessment of abdominal computed tomography and magnetic resonance images. *Diagn Interv Radiol.* 2025;31(3):196-205. [Crossref]
2. Kaba E, Solak M, Çeliker FB. The role of prompt engineering in radiology applications of generative AI. *Acad Radiol*. 2024;31(6):2641. [Crossref]
3. Nguyen D, MacKenzie A, Kim YH. Encouragement vs. liability: how prompt engineering influences ChatGPT-4's radiology exam performance. *Clin Imaging*. 2024;115:110276. [Crossref]
4. Alam S, Rahman A, Sohail SS. Optimizing ChatGPT-4's radiology performance with scale-invariant feature transform and advanced prompt engineering. *Clin Imaging*. 2025;118:110368. [Crossref]