Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263762

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Prospective quantitative analysis of hyperparameter and input optimization in GPT-5: comparative contribution to radiologist performance in abdominal radiology

Eren Çamur[1]
Turay Cesur[2]
Yasin Celal Güneş[3]
Muhammed Batuhan Gökhan[4]
Rıza Sarper Ökten[5]

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

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

[3]Yüksek İhtisas Hospital, Clinic of Radiology, Kırıkkale, Türkiye

[4]Yozgat Yerköy State Hospital, Clinic of Radiology, Yozgat, Türkiye

[5]Ankara Bilkent City Hospital, Clinic of Radiology, Ankara, Türkiye

PURPOSE

This study aims to evaluate the effect of input format and hyperparameter settings on GPT-5 and explore the contribution of GPT-5 assistance to radiologists' performance in abdominal cases.

METHODS

In this prospective study, 86 abdominal cases were evaluated, with GPT-5 evaluated in two deployment contexts: browser-based GPT-5 (default, non-configurable sampling settings) and GPT-5 accessed via the OpenAI application programming interface (API) with different temperature and top-p settings. First, the diagnostic and differential diagnosis performance of browser-based GPT-5 in these cases was assessed using two different input formats: "only visual" and "visual with imaging findings and clinical presentation." Subsequently, its performance was evaluated at varying temperature (0, 0.5, 1, 1.5) and top-p (0, 0.5, 1) values; the values at which the model performs best are considered "optimal settings." Finally, two junior radiologists evaluated the same cases with and without GPT-5 assistance with washout periods. Their performances were compared internally and with that of an abdominal radiologist.

RESULTS

With the "only visual" input format, browser-based GPT-5 achieved a diagnostic accuracy of 12%, increasing to 58% when imaging findings and clinical presentations were provided ($P < 0.001$). Hyperparameter optimization further improved GPT-5 performance, with diagnostic accuracy increasing to 73% at the optimal settings (temperature: 1.5, top-p: 1) and mean differential diagnosis scores improving from 3.44 to 3.84. The radiologists' diagnostic accuracy increased from 73% and 71% without assistance to 87% and 86% with browser-based GPT-5 assistance and further to 94% with GPT-5 with optimal settings assistance. Differential diagnosis performance similarly improved from median scores of 4 (range: 3–5) without assistance to 5 (range: 4–5) with GPT-5 (with optimal settings) assistance.

CONCLUSION

Using hyperparameter and input optimization settings with GPT-5 could improve its clinical utility.

CLINICAL SIGNIFICANCE

This study evaluates GPT-5 performance in a single-source, open-access abdominal case set. In this study, GPT-5 performance improved with structured text inputs and API-based hyperparameter optimization, and large language model (LLM) assistance was associated with improved diagnostic and differential diagnosis performance among junior radiologists. These findings suggest that documenting and standardizing hyperparameter settings (e.g., temperature and top-p) may be important for future LLM-based decision-support applications.

KEYWORDS

ChatGPT, abdomen, artificial intelligence, temperature, hyperparameter, optimization

Corresponding author: Eren Çamur

E-mail: eren.camur@outlook.com

Received 21 November 2025; revision requested 21 December 2025; accepted 31 January 2026.

Epub: 16.02.2026

DOI: 10.4274/dir.2026.263762

You may cite this article as: Çamur E, Cesur T, Güneş YC, Gökhan MB, Ökten RS. Prospective quantitative analysis of hyperparameter and input optimization in GPT-5: comparative contribution to radiologist performance in abdominal radiology. *Diagn Interv Radiol*. 16 February 2026 DOI: 10.4274/dir.2026.263762 [Epub Ahead of Print].

The development of large language models (LLMs) and, more recently, large multimodal models (LMMs) and vision-language models (VLMs) marks a major advance in artificial intelligence (AI), with the potential to transform various fields.[1] LLMs are developed using extensive datasets and sophisticated algorithms, enabling them to generate human-like text accurately. This capability has garnered considerable interest from both academia and industry, particularly for complex decision-making in data-rich fields such as medicine.[2] Within healthcare, LLMs hold promise for enhancing clinical workflows, offering diagnostic insights, supporting documentation, and facilitating patient education.[3,4] Unlike text-only LLMs, LMMs can jointly process multiple input modalities such as text and images. A major subgroup of LMMs are VLMs, which integrate image understanding with natural language processing to enable multimodal reasoning.[5]

Main points

- This study evaluated the performance of GPT-5 in abdominal radiology by assessing its diagnostic accuracy and differential diagnosis performance with two different input formats (only visual vs. visual with imaging findings and clinical presentation) using browser-based GPT-5 (default settings) and GPT-5 via the OpenAI application programming interface with controlled hyperparameters (temperature/top-p), demonstrating the potential contribution of GPT-5 assistance to abdominal radiology.
- The diagnostic accuracy of GPT-5 improved markedly when supplemented with imaging findings and clinical presentations (from 12% to 58%). Hyperparameter optimization further enhanced the diagnostic and differential diagnosis performance of the model [from 58% to 73% and from 3.44 ± 1.10, 4 (4–3) to 3.84 ± 0.81, 4 (4–3)].
- Systematic hyperparameter optimization revealed that the optimal settings significantly enhanced both diagnostic and differential diagnosis performance compared with different settings and browser-based settings, demonstrating the critical role of hyperparameter optimization in large language model (LLM) performance.
- Radiologists' performance was markedly improved with GPT-5 assistance; unassisted diagnostic accuracy rates of 73% and 71% increased to 87% and 86% with browser-based GPT-assistance and further to 94% with optimal settings assistance, aligning their performance with that of an experienced abdominal radiologist [92%, Likert score: 5 (5–49)].
- These findings underscore the potential of optimized LLMs as effective decision-support tools in abdominal radiology while emphasizing the need for further research to validate their utility in diverse clinical settings.

Radiology, a specialty that relies heavily on precision and detail to establish a correct diagnosis, stands to benefit immensely from integrating multimodal models (LMMs/VLMs). These models have already demonstrated notable potential across radiologic subspecialties, excelling in tasks such as generating multiple-choice questions and crafting patient-friendly report impressions.[6,7] Previous studies have underscored their high performance in analyzing publicly available cases across subspecialties, especially when detailed patient histories and imaging findings are provided.[8-13]

For clinical applications, optimizing the performance of these models requires adjustment of the hyperparameters, such as temperature and top-p, which influence response variability and creativity.[14,15] Temperature controls the randomness and creativity of model responses—a higher temperature results in more diverse and creative outputs, whereas a lower value makes the output more focused and deterministic.[16,17] Top-p limits the model by considering only the most probable tokens whose cumulative probability meets a specified threshold.[15] Together, these parameters help balance creativity and precision in generating outputs.[14-16] Unlike deterministic models (e.g., convolutional neural networks), LLMs are inherently variable, producing diverse outputs even for identical inputs. This variability can affect reliability, a critical factor in radiology.[9] Adjusting these hyperparameters enables the model to balance creativity with consistency, fostering dependable LLM-based applications in imaging interpretation.[18]

In a recent study, Suh et al.[19] investigated how temperature settings (0, 0.5, and 1) affect the diagnostic accuracy of multimodal LLMs, specifically ChatGPT-4V and Gemini Pro Vision, across radiological cases spanning multiple subspecialties. They reported modest improvements in diagnostic accuracy with increasing temperature for both models (GPT-4V: 41% to 49%; Gemini Pro Vision: 29% to 39%).[19] This study underscores the influence of temperature settings on potential improvements in the diagnostic performance of models.

To the best of our knowledge, no prior study has evaluated the effect of top-p and temperature settings on GPT-5 performance in abdominal radiology. This study addresses this gap and explores the impact of GPT-5 assistance on radiologists' performance.

Methods

Study design

This experimental prospective study used a cross-sectional design with three steps and was conducted between July 2024 and November 2025. The cases used in this study were obtained from the Eurorad database, established by the European Society of Radiology in April 2000.[20] Cases with multisystemic etiologies were included if abdominal imaging constituted the primary diagnostic focus, reflecting the spectrum of conditions routinely encountered in abdominal radiology practice. Each included case had "imaging findings," "clinical presentation," "correct diagnosis," and "differential diagnosis" sections.

The performance of GPT-5 was evaluated in two deployment contexts: browser-based GPT-5 (https://chatgpt.com), which uses default and undisclosed hyperparameter settings, and GPT-5 accessed via the OpenAI application programming interface (API; https://platform.openai.com), which allows explicit control of the temperature and top-p settings. The performances of browser-based GPT-5 and GPT-5 at different temperature and top-p settings using the OpenAI API were compared using the same cases. The temperature and top-p settings at which GPT-5 performs best are called the "optimal settings." The responses of browser-based GPT-5 and GPT-5 with the optimal settings for each case were noted. Finally, to demonstrate the contribution of GPT-5 assistance to radiologists' performance, Radiologist 1 (R1) and Radiologist 2 (R2), both European Diploma in Radiology certified with 7 years of experience in general radiology, evaluated the same cases with and without GPT-5 assistance, and their performances were compared with that of an abdominal radiologist (AR) with 25 years of experience in abdominal radiology. Radiologist 3 (R3) and Radiologist 4 (R4), both also EDiR-certified radiologists with 7 years of experience in general radiology, were responsible for case selection, the verification of input quality, and consensus-based evaluation of the model and radiologist responses throughout the study.

The study methodology adhered to the Standards for Reporting Diagnostic Accuracy Studies statement and was based on relevant items from the TRIPOD-LLM reporting framework for studies involving LLMs and

LMMs/VLMs, including explicit reporting of the model version, input structure, hyperparameter configuration, and radiologist–AI interaction.[21] This study was not registered; the protocol, TRIPOD-LLM checklist, and analysis plan were finalized prior to model querying and are provided in Supplementary Table 1. All data supporting the findings of this study are available within the text and Supplementary Table 2.

The flowchart of the study is illustrated in Figure 1.

Data collection

From the "Abdominal Imaging" section of the Eurorad database (https://www.eurorad.org/advanced-search?filter%5B0%5D=section%3A37), 100 cases were randomly selected by R3 and R4, who assessed the responses together. A total of 14 cases were excluded based on predefined criteria to ensure the integrity and diagnostic neutrality of the dataset, including 4 cases excluded due to the absence of a specified differential diagnosis and 6 cases removed because the correct diagnosis was explicitly mentioned within the imaging findings section, potentially biasing model interpretation. An additional three cases were excluded owing to the lack of descriptive findings in the imaging findings section, and one case was omitted because it contained only a video file without an accompanying static radiologic image suitable for model analysis. All the included cases were considered abdominal radiology cases, comprising both primary abdominal pathologies and multisystemic diseases with predominant abdominal organ involvement (the diagnostic reasoning was driven primarily by abdominal imaging findings).

For each case, multiple radiological images were available online. From these, R3 selected the two most diagnostically relevant images, excluding those with visual cues (e.g., arrows or annotations). Images were captured as unmodified JPEG screenshots and uploaded to GPT-5. The clinical presentation and imaging findings were transcribed verbatim without alteration or preprocessing, as all cases were already anonymized and publicly accessible. No additional enhancement or formatting was performed. None of the models were pretrained with any specific information by authors prior to the study. To minimize memory contamination, all cases were entered on the same day using seven independent accounts across browser-based environments, each with distinct temperature and top-p settings. In addition, to assess potential data contamination, we screened model outputs for verbatim case-title matches and for near-exact reproduction of the reference diagnosis phrasing. No verbatim title matches were identified, and near-exact phrasing overlap was observed in 2/86 cases (2.3%), which were retained but flagged in the sensitivity analyses.

Prompting and the hyperparameter tuning process

A single structured, zero-shot following prompt was used uniformly across all GPT-5 models:

"As a highly experienced radiology professor with 25 years of expertise in abdominal radiology, you will help me to solve abdominal cases. I will give you radiological images and sometimes give their imaging findings and clinical presentations of the cases. Your task is to analyze the images, imaging findings, and clinical presentations and then combine them to obtain the most likely diagnosis for the patient. Give me also the best four differential diagnoses."

The prompt employed role-based contextualization to emulate senior radiologist reasoning, aiming to enhance clinical relevance and foster detailed differentials. To prevent bias from variable session prompts, one standardized format was applied without iterative refinement. Prior to the study, the GPT-5 models received no additional training or author-provided information that could influence outcomes.

As the internal settings of browser-based GPT-5 are undisclosed and non-configurable, its performance was assessed as a baseline. To overcome this limitation, additional experiments were conducted via OpenAI's API, enabling manual adjustment of the tempera-

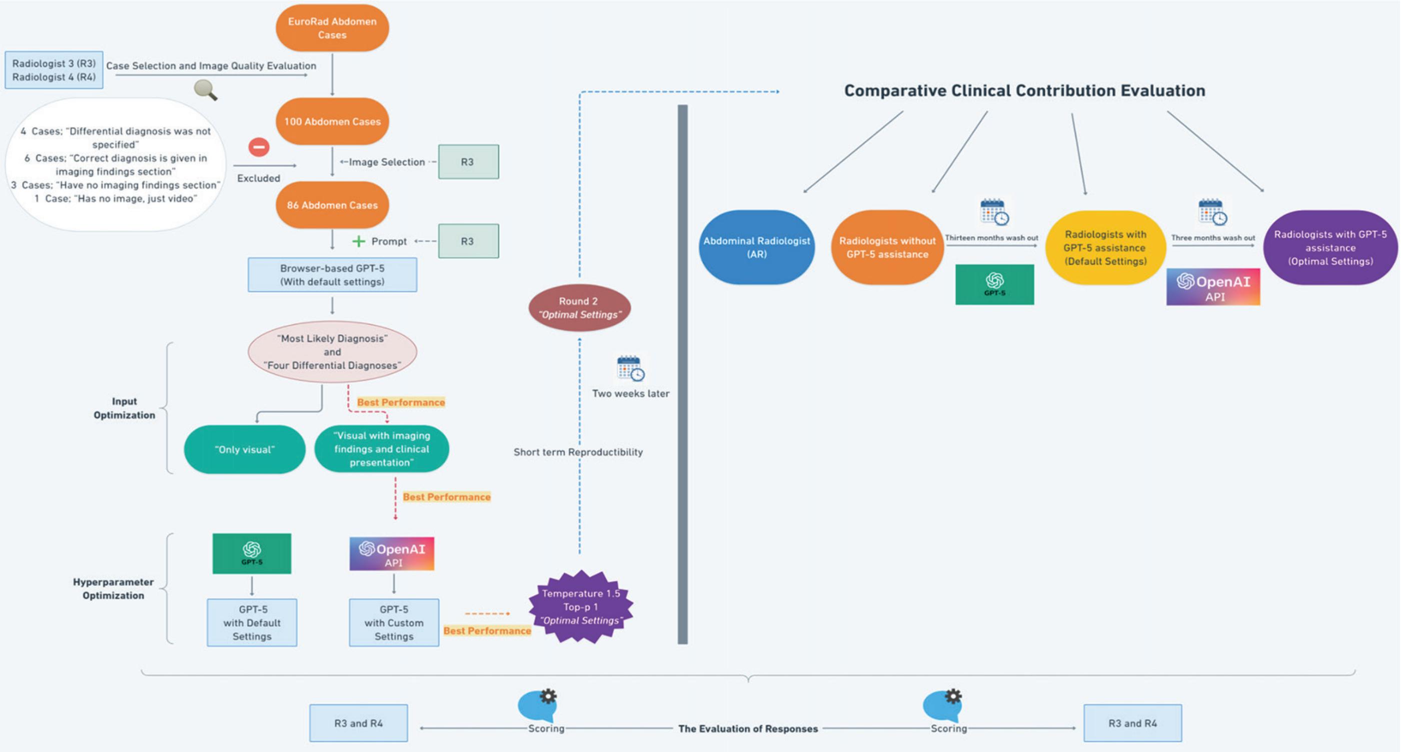

EuroRad Abdomen Cases
Radiologist 3 (R3)
Radiologist 4 (R4)
Case Selection and Image Quality Evaluation
100 Abdomen Cases
4 Cases: "Differential diagnosis was not specified"
6 Cases: "Correct diagnosis is given in imaging findings section"
3 Cases: "Have no imaging findings section"
1 Case: "Has no image, just video"
Excluded
Image Selection
R3
86 Abdomen Cases
Prompt
R3
Browser-based GPT-5 (With default settings)
"Most Likely Diagnosis" and "Four Differential Diagnoses"
Input Optimization
Best Performance
"Only visual"
"Visual with imaging findings and clinical presentation"
Best Performance
Hyperparameter Optimization
GPT-5
OpenAI API
GPT-5 with Default Settings
GPT-5 with Custom Settings
Best Performance
Temperature 1.5
Top-p 1
"Optimal Settings"
Round 2
"Optimal Settings"
Two weeks later
Short term Reproductibility
Comparative Clinical Contribution Evaluation
Abdominal Radiologist (AR)
Radiologists without GPT-5 assistance
Thirteen months wash out
GPT-5
Radiologists with GPT-5 assistance (Default Settings)
Three months wash out
OpenAI API
Radiologists with GPT-5 assistance (Optimal Settings)
R3 and R4
Scoring
The Evaluation of Responses
Scoring
R3 and R4

Figure 1. Study flowchart. API, application programming interface.

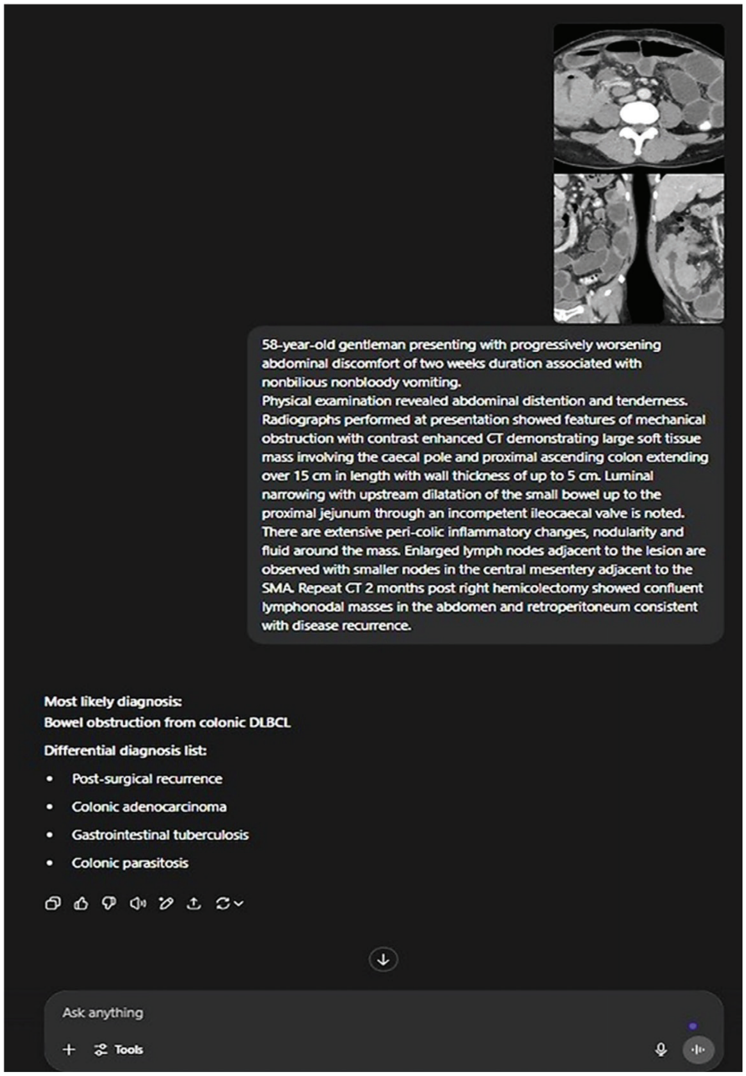
ture and top-p settings, which influence response determinism and diversity. Separate accounts were used for each setting to avoid contamination or memorization. The models were first tested across temperature values (0, 0.5, 1, 1.5) with a fixed top-p value—higher temperature values produced unevaluable outputs. The optimal temperature was then combined with the three top-p values (0, 0.5, 1) for further performance optimization.

Browser-based GPT-5

In the first phase of the study, browser-based GPT-5 (https://chatgpt.com; GPT-5-2025-08-08), which does not allow user control over the hyperparameter settings, was evaluated in two input formats: "only visual" and "visual with imaging findings and clinical presentation." To mitigate version drift, all browser-based queries were performed on the same day using the same displayed model build (GPT-5-2025-08-08), and API queries were executed using a fixed model identifier. For each case, the model was asked to provide the most likely diagnosis and four differentials (Figure 2). All cases were entered by R3 in a single tab and session to avoid memory contamination. Although this minimized inter-session variability, it may have introduced contextual carryover effects between cases due to the continuous chat structure.

OpenAI's application programming interface (GPT-5 with different hyperparameter settings)

In the second step, using OpenAI's API (https://platform.openai.com) and allowing the explicit adjustment of hyperparameters, the same cases were uploaded again to GPT-5 at different temperature and top-p settings, which are the two most important hyperparameters affecting the randomness, creativity, and precision of GPT-5 responses. In this step, cases were uploaded to GPT-5 in "visual with imaging findings and clinical presentation" format. Temperature was evaluated at settings of 0, 0.5, 1, and 1.5. At temperature: 2, GPT-5 does not respond correctly as text, often generating codes instead of text and producing answers that cannot be evaluated; therefore, no evaluation was performed in this setting. The temperature setting at which GPT-5 has the highest diagnostic accuracy was determined, and the cases were uploaded again at this temperature setting with different top-p settings (0, 0.5, and 1). Although these settings were selected to optimize diagnostic performance by balancing consistency and creativity, the same cases were uploaded again 2 weeks later using the optimal settings and the same format to assess short-term response stability and reproducibility; this was referred to as "Round 2."



58-year-old gentleman presenting with progressively worsening abdominal discomfort of two weeks duration associated with nonbilious nonbloody vomiting.
Physical examination revealed abdominal distention and tenderness. Radiographs performed at presentation showed features of mechanical obstruction with contrast enhanced CT demonstrating large soft tissue mass involving the caecal pole and proximal ascending colon extending over 15 cm in length with wall thickness of up to 5 cm. Luminal narrowing with upstream dilatation of the small bowel up to the proximal jejunum through an incompetent ileocaecal valve is noted. There are extensive peri-colic inflammatory changes, nodularity and fluid around the mass. Enlarged lymph nodes adjacent to the lesion are observed with smaller nodes in the central mesentery adjacent to the SMA. Repeat CT 2 months post right hemicolectomy showed confluent lymphonodal masses in the abdomen and retroperitoneum consistent with disease recurrence.

Most likely diagnosis:
Bowel obstruction from colonic DLBCL

Differential diagnosis list:

- Post-surgical recurrence
- Colonic adenocarcinoma
- Gastrointestinal tuberculosis
- Colonic parasitosis

Figure 2. Chat session example with browser-based GPT-5. CT, computed tomography.

Contribution of GPT-5 assistance to radiologist performance

In July 2024, R1 and R2 independently evaluated all cases offline using R3's internet-isolated computer without access to ChatGPT. For each case, they recorded the most likely diagnosis and four differential diagnoses. All evaluations by R1, R2, and AR were performed under blinded conditions, using personal computers without internet access. Each case included clinical history and corresponding radiological images.

Following a 13-month washout, in August 2025, R1 and R2 reassessed the cases in randomized case order but this time reviewing anonymized responses from browser-based GPT-5 (comprising the model's most likely and differential diagnoses) without knowing the source model. After a 3-month washout, in November 2025, they repeated the evaluation with GPT-5 outputs generated under the optimal settings, again blinded to the model identity.

At no step were R1 or R2 informed of the correct diagnoses. All model outputs were formatted into standardized digital folders containing anonymized clinical data, radiological images, and GPT-5 responses. These were securely transferred via encrypted drives and reviewed on offline systems by R3, ensuring complete blinding and data isolation. Finally, to better demonstrate the contribution of GPT-5's assistance to radiologists, AR evaluated the same cases without GPT-5 assistance in July 2024, and AR's performance was compared with those of R1 and R2.

Fundamentals of response evaluation

Through consensus, R3 and R4 assessed the responses of all the models and radiologists (R1, R2, and AR), categorizing the "most likely diagnosis" as either correct (1) or incorrect (0) and the "differential diagnosis" according to a 5-point Likert score:

1: 0/4 differentials are correct

2: 1/4 differentials are correct

3: 2/4 differentials are correct

4: 3/4 differentials are correct

5: 4/4 differentials are correct

The order of the differential diagnoses was not considered in the scoring. The responses were evaluated by R3 and R4 through consensus using only the differentials listed in the Supplementary Tables 1 and 2, and R3 and R4 independently scored a randomly selected subset (25/86 cases; 29.1%) prior to reaching consensus. Inter-assessor agreement was substantial for diagnostic correctness (Cohen's κ: 0.82) and moderate to substantial for the 5-point differential diagnosis score (weighted κ: 0.84). No additional diagnoses were accepted as "correct" beyond those provided in the answers. However, synonyms of the terms used (e.g., "celiac disease" vs. "gluten-sensitive enteropathy" or "volvulus" vs. "intestinal torsion") were scored as correct when medically equivalent.

This binary scoring (correct or incorrect) approach was chosen because each case had a clearly defined "most likely" (correct) diagnosis derived from the dataset, allowing objective evaluation. By contrast, a differential diagnosis inherently reflects a graded spectrum of alternatives and cannot be categorized only through binary scoring; therefore, a 5-point Likert scale was used to assess the degree of overlap between the responses and the reference differential list. The model inputs, responses, scoring criteria, and reader–AI interaction workflows were predefined and documented in accordance with key TRIPOD-LLM recommendations to enhance reproducibility and interpretability.

Statistical analysis

Descriptive statistics included mean, median, standard deviation, range, and frequencies with percentages. Normality was assessed using the Kolmogorov–Smirnov test. Diagnostic accuracy differences were evaluated using the McNemar test, and the Likert scores for GPT-5 and radiologists were compared using the Wilcoxon signed-rank test. Because comparisons across multiple inference settings were exploratory, we report unadjusted *P* values and interpret findings in the context of potential multiplicity. "Most likely diagnosis" and "differential diagnosis" Likert scores were not normally distributed (Kolmogorov–Smirnov, *P* < 0.001). Agreement between GPT-5 with the optimal settings and the Round 2 results was assessed using Cohen's kappa. Analyses were performed using SPSS software (version 26.0; IBM Corp., Armonk, NY, USA). Statistical significance was set at *P* < 0.05.

Results

Diagnostic and differential diagnosis performance of browser-based GPT-5 and across different hyperparameter settings

Browser-based GPT-5 demonstrated limited diagnostic accuracy in the "only visual" format, correctly answering 12% of cases [10/86; 95% confidence interval (CI), 6%–21%]. When the imaging findings and clinical presentations were added, performance markedly improved to 58% (50/86; 95% CI: 48%–68%) (*P* = 0.0006; Table 1). Differential diagnosis performance similarly improved from a mean of 1.85 (95% CI: 1.65–2.05) to 3.44 (95% CI: 3.25–3.63) (*P* = 0.0004).

At temperature settings of 0, 0.5, 1, and 1.5, diagnostic accuracies were 53% (46/86; 95% CI: 42%–63%), 64% (55/86; 95% CI: 54%–74%), 66% (57/86; 95% CI: 56%–76%), and 73% (63/86; 95% CI: 63%–82%), respectively. A significant improvement over browser-based performance was observed only at temperature: 1.5 (*P* = 0.007; Table 2). In the differential diagnoses, higher temperatures correlated with improved performance, with scores increasing from 3.05 (95% CI: 2.88–3.22; temperature: 0) to 3.84 (95% CI: 3.66–4.02; temperature: 1.5). Similarly, tuning the top-p parameter revealed a trend toward improved differential diagnoses. In addition, GPT-5 achieved a mean Likert score of 3.19 (95% CI: 3.01–3.37) at top-p: 0, 3.45 (95% CI: 3.27–3.63) at top-p: 0.5, and 3.84 (95% CI: 3.66–4.02) at top-p: 1, with top-p: 1 significantly outperforming top-p: 0 (*P* = 0.043; Table 3).

Overall, the optimal settings were defined as temperature: 1.5 and top-p: 1, which enabled GPT-5 to achieve its best diagnostic and differential diagnosis performance. Although there was a minor difference in the

Table 1. Diagnostic accuracy rates of browser-based GPT-5 with different evaluation formats

Format	Accuracy rate (%)	*P*
Only visual	12	*P* 0.0006
Visual with imaging findings and clinical presentation	58	

P values obtained from McNemar test.

Table 2. Comparison of the diagnostic performances of GPT-5 at different temperature settings

	Temperature 0	Temperature 0.5	Temperature 1	Temperature 1.5	Diagnostic accuracy (%)
Temperature 0	–	0.004	0.013	0.0009	53
Temperature 0.5	0.004	–	0.727	0.057	64
Temperature 1	0.013	0.727	–	0.070	66
Temperature 1.5	0.0009	0.057	0.070	–	73

P values obtained from Wilcoxon test.

Table 3. Comparison of the diagnostic performances of GPT-5 at different top-p settings

	Top-p 0	Top-p 0.5	Top-p 1	Mean Likert score
Top-p 0	–	0.004	0.043	3.19
Top-p 0.5	0.004	–	0.048	3.45
Top-p 1	0.043	0.048	–	3.84

P values obtained from Wilcoxon test.

GPT-5 responses with the optimal settings in Round 2, no significant difference was observed in diagnostic accuracy (*P* = 0.41) or differential diagnosis performance (*P* = 0.36). Agreement between GPT-5 responses generated with the optimal settings at baseline and in the Round 2 evaluation was high, with Cohen's κ: 0.882 (95% CI: 0.80–0.96) for the most likely diagnosis and κ: 0.816 for differential diagnosis performance, indicating strong agreement.

The performance of the browser-based GPT-5 and across different hyperparameter settings are provided in Figures 3 and 4.

Diagnostic and differential diagnosis performance of radiologists with and without GPT-5 assistance

Without GPT-5 assistance, R1 and R2 demonstrated diagnostic accuracies of 73% (63/86; 95% CI: 63%–82%) and 71% (61/86; 95% CI: 61%–80%), respectively. With browser-based GPT-5 assistance, their accuracy significantly improved to 87% (75/86; 95% CI: 79%–94%) and 86% (74/86; 95% CI: 78%–93%), respectively (*P* = 0.001, *P* = 0.001). Using GPT-5 with the optimal settings further improved their accuracy to 94% (81/86; 95% CI: 88%–99%) (*P* = 0.031, *P* = 0.028). Comparatively, AR achieved an accuracy of 92% (79/86; 95% CI: 85%–97%), which was superior to the accuracy of R1 and R2 without GPT-5 assistance (*P* < 0.001) but not significantly different from their performances with both browser-based GPT-5 and GPT-5 with the optimal settings (Table 4).

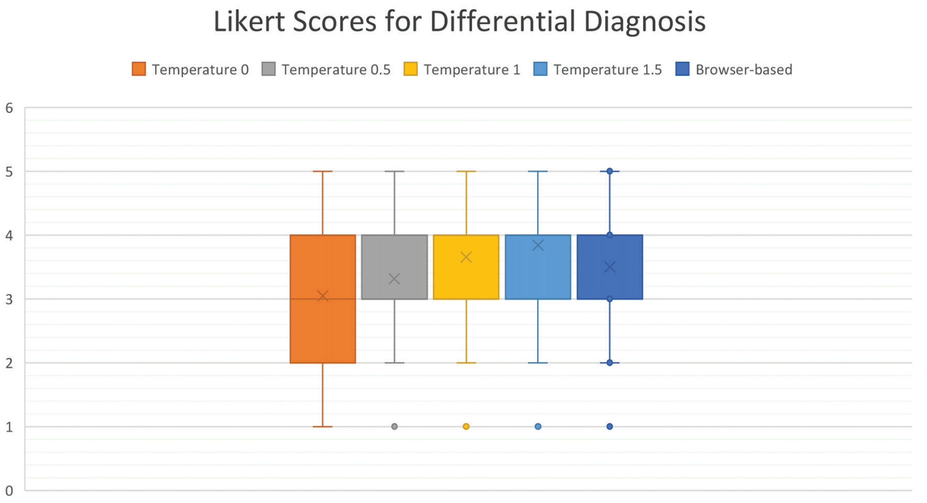


Figure 3. Differential diagnosis performances at different temperature settings and with browser-based GPT-5.

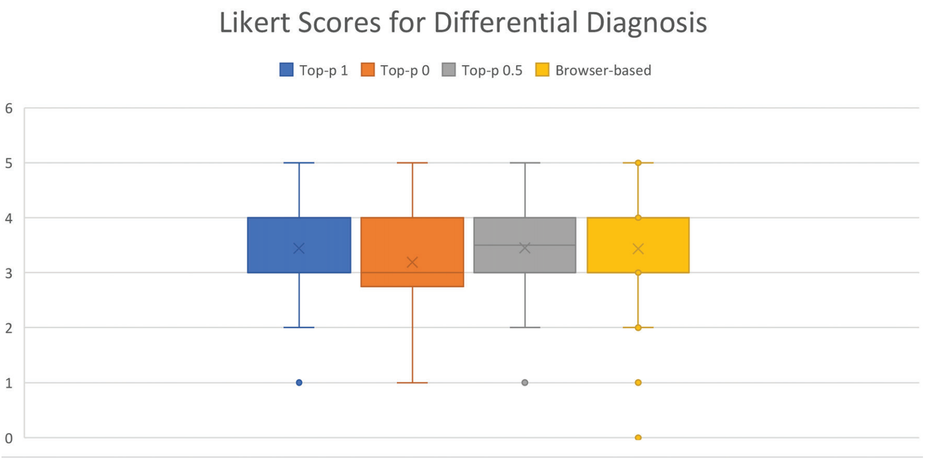


Figure 4. Differential diagnosis performances at different top-p settings and with browser-based GPT-5.

Table 4. Comparison of the diagnostic performances of radiologists with/without GPT-5 assistance and with an abdominal radiologist

	Abdominal radiologist	R1 without GPT-5 assistance	R1 with browser-based GPT-5 assistance	R1 with GPT-5 optimal setting assistance	R2 without GPT-5 assistance	R2 with browser-based GPT-5 assistance	R2 with GPT-5 optimal setting assistance	Diagnostic accuracy (%)
Abdominal radiologist	–	0.0006	0.219	0.687	0.0005	0.180	0.125	92
R1 without GPT-5 assistance	0.0006	–	0.0008	0.0005	0.227	0.001	0.0006	73
R1 with browser-based GPT-5 assistance	0.219	0.0008	–	0.031	0.0007	1	0.004	87
R1 with GPT-5 optimal setting assistance	0.687	0.0005	0.031	–	0.0002	0.016	0.250	94
R2 without GPT-5 assistance	0.0005	0.227	0.0007	0.0002	–	0.0006	0.0002	71
R2 with browser-based GPT-5 assistance	0.180	0.001	1	0.016	0.0006	–	0.002	86
R2 with GPT-5 optimal setting assistance	0.125	0.0006	0.004	0.250	0.0002	0.002	–	94

P values obtained from Wilcoxon test. R1, radiologist 1; R2, radiologist 2.

The differential diagnosis performance (mean Likert score) of R1 and R2 improved from 3.85 (95% CI: 3.67–4.03) and 3.90 (95% CI: 3.72–4.08) to 4.20 (95% CI: 4.05–4.35) and 4.24 (95% CI: 4.09–4.39), respectively, with browser-based GPT-5 assistance and further to 4.56 (95% CI: 4.43–4.69) and 4.49 (95% CI: 4.35–4.63), respectively, with the assistance of the optimal settings (*P* < 0.001, *P* < 0.001) (Table 5). Although AR achieved a mean Likert score of 4.40 (95% CI: 4.25–4.55), both radiologists outperformed AR in terms of differential diagnosis performance when assisted by GPT-5 with the optimal settings (*P* = 0.002, *P* = 0.001; Table 6).

Discussion

The most striking result of our study was that optimizing GPT-5's hyperparameters significantly enhanced both diagnostic accuracy and differential diagnosis performance in abdominal radiology. This improvement has important clinical implications, as a well-structured differential diagnosis list directly aids radiologists and clinicians by narrowing diagnostic considerations, reducing uncertainty, and potentially helping patient management decisions. The improved model performance following hyperparameter optimization likely reflects enhancements in the model's diagnostic reasoning processes. Rather than simply generating more creative responses, fine-tuning the temperature and top-p settings may encourage the model to explore a broader but still clinically plausible range of diagnostic possibilities, improving its capacity to generate comprehensive differential diagnoses.[14-16] This is particularly beneficial in scenarios involving ambiguous imaging findings, overlapping disease presentations, or rare pathologies, where rigid, pattern-based outputs may fall short. When calibrated appropriately, these settings help the model prioritize salient features, propose relevant alternatives, and support diagnostic reasoning akin to expert human thinking—ultimately translating into increased diagnostic confidence for the radiologist.

Table 5. Distribution of differential diagnosis Likert scores across GPT-5 conditions and radiologists

	1 Point (%)	2 Point (%)	3 Point (%)	4 Point (%)	5 Point (%)
Browser-based GPT-5 (only visual)	62	20	3	4	11
Browser-based GPT-5 (visual with imaging findings and clinical presentation)	7	7	32	40	14
GPT-5 with optimal settings	3	5	20	55	17
R1 without GPT-5 assistance	0	8	21	45	26
R1 with browser-based GPT-5 assistance	0	8	7	38	47
R1 with GPT-5 optimal setting assistance	0	0	8	22	70
R2 without GPT-5 assistance	0	11	15	46	27
R2 with browser-based GPT-5 assistance	0	9	6	36	49
R2 with GPT-5 optimal setting assistance	0	0	8	28	74
Abdominal radiologist	0	0	12	37	51

Likert score definition; 1: 0/4 correct differentials, 2: 1/4, 3: 2/4, 4: 3/4, 5: 4/4. R1, radiologist 1; R2, radiologist 2.

Table 6. Comparison of the differential diagnosis performances of radiologists with/without GPT-5 assistance and with an abdominal radiologist

	Abdominal radiologist	R1 without GPT-5 assistance	R1 with browser-based GPT-5 assistance	R1 with GPT-5 optimal setting assistance	R2 without GPT-5 assistance	R2 with browser-based GPT-5 assistance	R2 with GPT-5 optimal setting assistance	Mean Likert score
Abdominal radiologist	–	0.0006	0.002	0.002	0.0005	0.036	0.117	4.40
R1 without GPT-5 assistance	0.0006	–	0.0008	0.0002	0.285	0.0005	0.0003	3.85
R1 with browser-based GPT-5 assistance	0.006	0.0008	–	0.0004	0.0004	0.046	0.0006	4.20
R1 with GPT-5 optimal setting assistance	0.002	0.0002	0.0004	–	0.0002	0.0007	0.109	4.56
R2 without GPT-5 assistance	0.0005	0.285	0.0004	0.0002	–	0.0006	0.0003	3.90
R2 with browser-based GPT-5 assistance	0.036	0.0005	0.046	0.0007	0.0006	–	0.001	4.24
R2 with GPT-5 optimal setting assistance	0.117	0.0003	0.0006	0.109	0.0003	0.001	–	4.49

P values obtained from Wilcoxon test. R1, radiologist 1; R2, radiologist 2.

The study by Suh et al.[19] suggested that temperature may influence multimodal diagnostic outputs, although reported improvements were modest and not consistently significant. Our study extends these findings by systematically evaluating both the temperature and top-p settings by focusing specifically on abdominal radiology and assessing impacts on diagnostic accuracy, differential diagnosis quality, and the contribution to radiologists' performance.

Another important result of our study is that GPT-5 provides more accurate responses for the most likely diagnoses and differentials when imaging findings and clinical presentations are provided in addition to radiological images. Previous studies have focused on the visual performance of ChatGPT.[22-26] Dehdab et al.[24] reported a diagnostic accuracy of 56% for ChatGPT-4V in the interpretation of chest computed tomography (CT); however, performance improved markedly to 83.3% in cases of diffuse Coronavirus Disease 2019 involvement, likely due to more conspicuous imaging features.Similarly, Kuzan et al.[26] demonstrated that the model exhibited high accuracy in identifying magnetic resonance imaging (MRI) sequences (approximately 89%) and reasonable sensitivity (79.6%) for detecting diffusion restriction in acute stroke imaging. Conversely, Ren et al.[23] observed limited diagnostic performance in the detection of osteosarcoma on radiographs, with an overall accuracy of only 20%, underscoring the current limitations of LLMs in direct image interpretation.Horiuchi et al.[25] further compared the diagnostic capabilities of ChatGPT-4 (text-based input) and ChatGPT-4V (image-based input) in musculoskeletal imaging, concluding that the text-based model demonstrated superior diagnostic accuracy. Our study is unique in that it reveals how the performance of ChatGPT changes when supported by imaging findings and clinical presentations. LLMs use natural language processing as a starting point; because of their nature, it is likely that when imaging findings are described and clinical presentations provided, LLMs are better able to evaluate and analyze this information in a text-based manner.

Previous studies have evaluated the diagnostic performance of ChatGPT in various sections. Horiuchi et al.[25] evaluated the diagnostic performance of ChatGPT-4 on 100 consecutive cases from the American Journal of Neuroradiology "Case of the Week," reporting that the diagnostic accuracy of ChatGPT-4 in these cases was 50%.[13] Moreover, Kahalian et al.[27] uploaded 52 radiological images to ChatGPT-4 in their study on oral and maxillofacial pathologies, reporting that the model had a diagnostic accuracy of 56.9% when given a hint of an imaging finding in addition to these images. Similarly, our results demonstrate that browser-based GPT-5 has a diagnostic accuracy of 58%.

As a diagnostic adjunct, ChatGPT's recommendations may reinforce radiologist confidence by serving as an AI-driven cognitive checklist, potentially reducing diagnostic omissions. Alignment between the model's differential output and the radiologist's impression encourages broader deliberation while allowing final synthesis within the clinical, laboratory, and imaging context. Future research should assess radiologist performance under varying conditions—receiving only the top-ranked diagnosis, only the full differential list, or both—to clarify the respective contributions of broad versus focused AI support. Such comparative analyses will guide the optimal integration of LLMs into radiologic practice, determining whether expansive reasoning or targeted guidance provides the greatest diagnostic benefit.

We noted that the diagnostic performance of radiologists improved with GPT-5 assistance both with browser-based and optimal settings. Similarly, Siepmann et al.[28] evaluated the influence of ChatGPT-4 assistance on radiological interpretation by asking six radiologists with varying levels of experience to assess 40 different radiological images—including X-ray, CT, MRI, and angiographic examinations—in both unassisted and ChatGPT-4-assisted sessions. ChatGPT-4 assistance slightly increased the diagnostic accuracy of the radiologists, as evidenced by an improvement from 75.4% to 78.3%, but this difference was not statistically significant.[28] This study is the first to demonstrate that GPT-5 assistance not only enhances diagnostic accuracy but also facilitates more comprehensive differential diagnosis formulation, thereby contributing to improved diagnostic reasoning.

This study has several limitations. First, it was based on a single open-access dataset, which raises concerns about data contamination and limits the generalizability of the results to the broader, heterogeneous patient populations encountered in actual radiologic practice. Additionally, no real patient data were used, which further constrains the applicability of the findings to clinical settings that involve complex comorbidities and subspecialty-specific nuances.

Second, since this study focuses on hyperparameter optimization, evaluation with different prompts/prompt engineering effects was not tested. Different prompts may affect the success of GPT-5 by providing different results in this regard. Further studies evaluating the effect of different prompts are needed.

Third, GPT-5 responded to the same cases in different time periods in this study. This may have caused differences in the responses due to both the nature of ChatGPT itself and the different responses resulting from stochasticity and updates made at different times within the model; models can improve themselves over time with new knowledge and updates. We did not conduct controlled experiments or probing to infer the default hyperparameter settings used by browser-based GPT-5. As such, potential variability stemming from undocumented or evolving internal configurations may cause performance differences at different times. Therefore, the browser-based GPT-5 responses included in the study reflect the responses at the time of the study. It should be noted that these responses may improve and change in the future.

Fourth, despite long washout periods and a randomized case order, R1 and R2 continued routine clinical practice during the 13-month interval, and a natural learning curve may have contributed to improved performance. Therefore, the observed gains cannot be attributed solely to GPT-5 assistance, and residual confounding from time-related improvement cannot be fully excluded.

Finally, although the radiologists were blinded to the model generating each response, the same cases were evaluated at different stages of the study. Despite implementing a 3-month washout period to minimize recall bias, the potential for learning bias cannot be fully excluded. Future studies should consider alternative designs or longer washout intervals to further reduce this risk.

In conclusion, GPT-5 performance varied strongly with different input formats and hyperparameter settings. Adding imaging findings and clinical presentations increased browser-based diagnostic accuracy from 12% to 58%, and API tuning improved it further to 73% at the optimal setting (temperature: 1.5, top-p: 1) with better differential diagnosis quality. Radiologist accuracy improved from 73%/71% unaided to 87%/86% with browser-based assistance and to 94%/94% with optimized assistance, approaching the AR benchmark (92%). These findings indicate a need for further studies with standardized, documented hyperparameter configurations to develop LLM-assisted decision-support systems and improve their contribution to radiologist performance.

Acknowledgments

Eurorad cases used in this study have been approved for use in this study by "Ms. Stefanie Muzik," who has granted the necessary permission and confirmed that these cases may be used.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Table 1 Link: https://d2v96fxpocvxx.cloudfront.net/90a4190a-90d9-41a4-a9c9-d78d3fa8efda/content-images/591c3248-ae15-4ca8-bb66-1924aff4d612.pdf

Supplementary Table 2 Link: https://d2v96fxpocvxx.cloudfront.net/90a4190a-90d9-41a4-a9c9-d78d3fa8efda/documents/DIR-2026.263762-supplement-table-2.xlsx

References

1. Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274. [Crossref]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. [Crossref]
3. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res*. 2023;25:e48659. [Crossref]
4. Kuckelman IJ, Yi PH, Bui M, Onuh I, Anderson JA, Ross AB. Assessing AI-powered patient education: a case study in radiology. *Acad Radiol*. 2024;31(1):338-342. [Crossref]
5. Nakaura T, Ito R, Ueda D, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol*. 2024;42(7):685-696. [Crossref]
6. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice questions. *Acad Radiol*. 2024;31(9):3872-3878. [Crossref]
7. López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: a feasibility study. *Int J Med Inform*. 2024;187:105443. [Crossref]
8. Almeida LC, Farina EMJM, Kuriki PEA, Abdala N, Kitamura FC. Performance of ChatGPT on the Brazilian radiology and diagnostic ımaging and mammography board examinations. *Radiol Artif Intell*. 2024;6(1):e230103. [Crossref]
9. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):e232756. [Crossref]
10. Toyama Y, Harigai A, Abe M, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. *Jpn J Radiol*. 2024;42(2):201-207. [Crossref]
11. Ariyaratne S, Jenko N, Mark Davies A, Iyengar KP, Botchu R. Could ChatGPT pass the UK radiology fellowship examinations? *Acad Radiol*. 2024;31(5):2178-2182. [Crossref]
12. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and bard in a text-based radiology knowledge assessment. *Can Assoc Radiol J*. 2024;75(2):344-350. [Crossref]
13. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology*. 2024;66(1):73-79. [Crossref]
14. Lee JH, Shin J. How to optimize prompting for large language models in clinical research. *Korean J Radiol*. 2024;25(10):869-873. [Crossref]
15. Cheat Sheet: Mastering Temperature and Top_p in ChatGPT API [Internet]. OpenAI Developer Forum. [Crossref]
16. Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of ChatGPT: modifying creativity for clinical research. *JMIR Hum Factors*. 2024;11:e53559. [Crossref]
17. Akamine A, Hayashi D, Tomizawa A, et al. Effects of temperature settings on information quality of ChatGPT-3.5 responses: a prospective, single-blind, observational cohort study. *medRxiv*. 2024 Jun 12. Preprint. [Crossref]
18. Suh CH, Yi J, Shim WH, Heo H. Insufficient transparency in stochasticity reporting in large language model studies for medical applications in leading medical journals. *Korean J Radiol*. 2024;25(11):1029-1031. [Crossref]
19. Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. *Radiology*. 2024;312(1):e240273. [Crossref]
20. Eurorad [Internet]. Vienna: European Society of Radiology. [Crossref]
21. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277(3):826-32. [Crossref]
22. Kaba E, Solak M, Beyazal M. Evaluating ChatGPT-4o in diffusion-weighted imaging interpretation: is it useful? *Acad Radiol*. 2025;32(1):591-593. [Crossref]
23. Ren Y, Guo Y, He Q, et al. Exploring whether ChatGPT-4 with image analysis capabilities can diagnose osteosarcoma from X-ray images. *Exp Hematol Oncol*. 2024;13(1):71. [Crossref]
24. Dehdab R, Brendlin A, Werner S, e al. Evaluating ChatGPT-4V in chest CT diagnostics: a critical image interpretation assessment. *Jpn J Radiol*. 2024;42(10):1168-1177. [Crossref]
25. Horiuchi D, Tatekawa H, Oura T, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol*. 2025;35(1):506-516. [Crossref]
26. Kuzan BN, Meşe İ, Yaşar S, Kuzan TY. A retrospective evaluation of the potential of ChatGPT in the accurate diagnosis of acute stroke. *Diagn Interv Radiol*. 2025;31(3):187-195. [Crossref]
27. Kahalian S, Rajabzadeh M, Öçbe M, Medisoglu MS. ChatGPT-4.0 in oral and maxillofacial radiology: prediction of anatomical and pathological conditions from radiographic images. *Folia Med (Plovdiv)*. 2024;66(6):863-868. [Crossref]
28. Siepmann R, Huppertz M, Rastkhiz A, et al. The virtual reference radiologist: comprehensive AI assistance for clinical image reading and interpretation. *Eur Radiol*. 2024;34(10):6652-6666. [Crossref]