EURASIAN JOURNAL OF EMERGENCY MEDICINE
Original Article
DOI: 10.4274/eajem.galenos.2026.23356
Eurasian J Emerg Med. 2026;25: 411-8

Evaluating the Performance of Large Language Models in Radiological Decision-making: A Comparative Study Using ACR Appropriateness Criteria for Acute Abdomen

Abdullah Enes Ataş[1], Halil İbrahim Şara[1], Fatih Cemal Tekin[2]

[1]Necmettin Erbakan University Faculty of Medicine, Department of Radiology, Konya, Türkiye
[2]Konya City Hospital, Clinic of Emergency Medicine, Konya, Türkiye

Abstract

Aim: To evaluate the accuracy of four large language models in assessing the appropriateness of imaging tests for patients with suspected acute abdomen, using the American College of Radiology (ACR) Appropriateness Criteria as the gold standard.

Materials and Methods: This cross-sectional study evaluated ChatGPT 5.5, Gemini 3.1 Pro, Claude Opus 4.8, and DeepSeek V4 Pro across 30 clinical scenarios of acute abdominal pain. Models were tested using zero-shot and one-shot prompting strategies, yielding a total of 240 responses. Responses were compared to the ACR categories of "Usually Appropriate, "May Be Appropriate," and "Usually Not Appropriate." Performance was measured using exact match accuracy, linearly weighted Cohen's kappa, and major discordance rates.

Results: The overall accuracy of the pooled models was 86.3%, with a weighted kappa of 0.78. Claude Opus 4.8 achieved the highest accuracy at 95.0% and a kappa of 0.91. Clinically risky major discordance rates remained low across all models, ranging from 3.3% to 8.3%. Utilizing a one-shot prompting strategy increased overall accuracy from 81.7% to 90.8% compared to zero-shot prompting.

Conclusion: Large language models successfully align imaging recommendations for acute abdominal conditions with established ACR guidelines. Providing a guideline-formatted example prior to the task improves model performance and highlights their potential as reliable radiological decision-support tools. By supporting faster, guideline-concordant imaging selection, such tools may help reduce unnecessary ionizing radiation exposure and alleviate decision-making pressure in high-volume emergency settings.

Keywords: Radiology, acute abdomen, artificial ıntelligence, clinical decision support systems, natural language processing

Introduction

Acute abdominal pain is a very broad and complex clinical condition that physicians most frequently encounter in emergency departments and outpatient clinics of hospitals, with underlying causes ranging from simple and self-limiting disorders to life-threatening conditions requiring emergency surgical intervention (1). While the patient's history, physical examination, and basic laboratory tests help physicians narrow down the list of possible diseases, radiological imaging methods are often needed to make a definitive diagnosis and create the correct treatment plan (2). However, choosing the right imaging method within this wide range of differential diagnoses is not always easy. Ordering unnecessary or clinically inconsistent imaging tests exposes patients to ionizing radiation, increases healthcare system costs, and slows diagnostic processes by causing inefficient use of limited resources in emergency departments (3,4).

To overcome these challenges and ensure standardization in clinical decision-making, the American College of Radiology (ACR) developed the "Appropriateness Criteria," an evidence-based guideline (1,5). These criteria, updated annually by expert panels using current data from the medical literature, guide physicians in selecting the most appropriate radiological imaging or treatment method for a given clinical scenario (5,6). In fact, in the United States, under regulations aimed at improving the

Corresponding Author: Abdullah Enes Ataş MD, Necmettin Erbakan University Faculty of Medicine, Department of Radiology, Konya, Türkiye
E-mail: aenesatas@gmail.com **ORCID ID:** orcid.org/0000-0001-6623-3024

Cite this article as: Ataş AE, Şara Hİ, Tekin FC. Evaluating the performance of large language models in radiological decision-making: a comparative study using ACR appropriateness criteria for acute abdomen. Eurasian J Emerg Med. 2026;25: 411-8

Received: 14.07.2026
Accepted: 16.08.2026
Published: 04.09.2026

quality of healthcare, consulting such clinical decision support systems is encouraged when ordering advanced imaging studies (3,6). Despite this, the use of these guidelines in daily and intensive clinical practice remains generally limited. The main reason for this is that establishing a direct link between complex patient notes kept in free text and guidelines with strict rules and structured templates is time-consuming and difficult (7).

This gap between guideline complexity and everyday practice is most pronounced in the emergency department, where physicians must select an imaging strategy within minutes, under diagnostic uncertainty, and often outside regular working hours when subspecialty radiology consultation may be less readily available. In this setting, an unnecessary computed tomography scan not only exposes the patient to avoidable ionizing radiation but also prolongs the patient's length of stay in the emergency department and increases downstream imaging costs. A decision-support tool capable of instantly cross-referencing a free-text clinical presentation against ACR-based recommendations could, therefore, help emergency physicians select the correct initial imaging test more consistently, reduce unnecessary radiation exposure, and support timely, evidence-based triage decisions.

Recent advancements in artificial intelligence, particularly in large language models (LLMs), hold significant potential to address this problem (8). These systems, trained to understand and generate human language, promise the ability to rapidly analyze patient data in free-text format and function as evidence-based clinical decision support systems (8). Research in the medical literature indicates that LLMs show promise in radiology for interpreting clinical scenarios and recommending examinations. Furthermore, it has been found that an AI model specifically supported by ACR guidelines can outperform radiologists and standard language models in selecting examinations (9). Similarly, it has been proven that LLMs can predict the correct imaging tests with high agreement in various radiology subspecialties (10). However, given the sensitivity of medical data, the risk of hallucinations and the capacity of these models to meet clinical safety standards remain obstacles to their widespread adoption (11). While the current medical literature has demonstrated the potential of LLMs for general radiological decision-making, direct comparative studies assessing the extent to which the newest, most advanced language models align with the ACR Appropriateness Criteria in field-specific scenarios are lacking.

The aim of this research is to address this gap in the literature by evaluating how accurately four current LLMs, which are increasingly used in the medical field, can assess the appropriateness of imaging tests for patients with suspected acute abdomen.

Materials and Methods

Study Design

This cross-sectional, comparative study is designed to measure the ability of LLMs to assess the appropriateness of radiological examinations. The main focus of the study is to determine the extent to which the responses generated by four LLMs regarding the appropriateness of a selected imaging test for a patient presenting with acute abdominal pain are consistent with reference guidelines. All clinical scenarios were completed using the most current versions of the LLMs dated between June 1 and July 1, 2026.

The ACR Appropriateness Criteria® published by the ACR were used as the reference and gold standard in the study. These criteria, prepared by expert panels, evaluate each imaging or treatment option for a given clinical scenario on a comprehensive 9-point scale. For ease of use in daily practice, this scale is generally divided into three categories: "Usually Appropriate" (green category), "May Be Appropriate" (yellow category), and "Usually Not Appropriate" (red category). Throughout the study, the accuracy of the models' responses was tested by comparing them to the gold standard template consisting of these three categories.

Ethical Considerations

This study was completed using current versions of publicly available LLMs, drawing on ACR's publicly available guidelines (12). Because no patient data, hospital records, or images were used, approval from the ethics committee was not required.

Clinical Scenarios

For the study, thirty (30) clinical case scenarios were prepared, directly corresponding to specific variants defined under the heading of acute abdominal pain in the ACR Appropriateness Criteria. These scenarios were selected to cover nine basic clinical presentations frequently encountered by physicians in daily clinical and emergency practice. These are: non-localized generalized abdominal pain, right lower quadrant pain (suspected appendicitis), right upper quadrant pain, left lower quadrant pain (suspected diverticulitis), left upper quadrant pain, epigastric pain, suspected small bowel obstruction, suspected acute mesenteric ischemia, and acute pyelonephritis.

For each prepared scenario, a specific imaging test (e.g., non-contrast computed tomography or abdominal ultrasonography) was selected, and the correct ACR guideline category for this scenario-test pairing was recorded as the gold-standard response. The 30 selected scenarios were balanced to reflect that, in actual clinical guidelines, the tests deemed appropriate outnumber

those deemed inappropriate. In this context, 17 of the scenarios were created from test matches belonging to the "Usually Appropriate" category, 6 from "May Be Appropriate", and 7 from "Usually Not Appropriate."

All 30 vignettes were drafted by a board-certified radiologist with 10 years of experience and were based directly on the illustrative clinical scenarios provided in the corresponding ACR Appropriateness Criteria variant. Each vignette and its assigned gold-standard category were then independently checked against the published ACR document by a second radiologist with 8 years' experience and an emergency medicine physician; disagreements were resolved by consensus. The vignettes were written as concise, single-issue presentations rather than as verbatim excerpts from real patient records, since no patient data were used in this study; this simplification is discussed as a limitation.

LLMs and Prompting Strategy

All models were accessed via a web interface between June 1 and July 1, 2026, using the following snapshots: ChatGPT 5.5, Gemini 3.1 Pro, Claude Opus 4.8, and DeepSeek V4 Pro. Default vendor settings were used for all models. Each of the 60 prompts per model (30 scenarios × 2 prompting strategies) was submitted as a new, independent conversation with no shared context, to ensure that no model's response to one scenario was influenced by its response to a previous one.

Four advanced LLMs were selected for evaluation in this study: ChatGPT 5.5, Gemini 3.1 Pro, Claude Opus 4.8, and DeepSeek V4 Pro. Each model was presented with 30 prepared case scenarios, using two different prompting strategies. This allowed each model to respond to a total of 60 queries (30 scenarios × 2 prompting strategies). The two strategies used to guide the models are:

• Zero-shot prompting: In this method, the model is given only a the task brief instruction explaining, followed by the clinical scenario and the suggested imaging test that it must evaluate directly. No examples of what the expected response should be are shown to the model beforehand.

• One-shot prompting: In this method, the model is given a single task instruction and a solved example is shown before the actual questions. This example illustrates that the study included a scenario outside the 30 cases, a test, the category deemed correct according to the guideline, and a brief justification for that classification. Thus, the model was able to see exactly what format and logical framework for the response were expected of it before proceeding to the actual task.

Across all methods employed, models were asked to provide responses in one of the three specified categories ("Usually Appropriate", "May Be Appropriate", or "Usually Not Appropriate") and a brief justification consisting of one or two sentences supporting their preference. As a result, 240 model responses were obtained by presenting 30 scenarios to 4 models using 2 strategies. Each of these responses was independently compared with the gold standard ACR category to which it belonged.

The complete wording of the fixed instruction, the one-shot worked example, and the clinical scenario and imaging study text for all 30 cases are provided in Supplementary Material S1, allowing independent replication of the prompting procedure

Evaluation of Responses

Each of the 240 responses obtained from the models was scored at various levels by comparing it to the established gold standard.

• Perfect match (accuracy): Whether the category suggested by the model exactly matched the correct category in the ACR guideline (yes/no) was evaluated.

• Ordered distance (absolute ordered error): Categories were assigned numerical values according to severity (Red = 1, Yellow = 2, Green = 3). The absolute difference between the numerical value chosen by the model and the gold standard value was calculated. A value of 0 indicates a perfect match; a value of 1 indicates that the model deviated by only one step (e.g., evaluating a test that is "Usually Appropriate" as "May Appropriate"); and a value of 2 indicates that the model chose a completely opposite category.

• Major discordance is the most clinically dangerous type of error, with the potential to cause the greatest harm to the patient. This occurs when the model classifies an examination as "Usually Appropriate" when it should be "Usually Not Appropriate," or vice versa. This type of error has been analyzed separately from the overall accuracy assessment, as it can result in a patient being subjected to an unnecessary test or being denied an urgent test.

Statistical Analysis

There is a natural ranking and a significance level among the response categories (red, yellow, green) in the study. An error of one category is clinically more acceptable than an error of two categories (exact contrast). Therefore, the linearly-weighted Cohen's kappa test was used as the primary indicator of agreement between each model and the ACR gold standard. Unlike simple accuracy percentages, the weighted kappa is considered a more reliable metric because it statistically corrects for chance agreement and awards the model partial credit in situations where it "narrowly missed" (e.g., predicting yellow instead of green). The Landis and Koch criteria, considered standard in this field, were used to interpret the obtained kappa

values. Accordingly, scores were categorized as follows: below 0, weak; 0.00-0.20, weak; 0.21-0.40, fair; 0.41-0.60, moderate; 0.61-0.80, substantial; and 0.81-1.00, almost perfect. For kappa analyses, 95% confidence intervals (CIs) were calculated using a bootstrap method in which the data were resampled 2,000 times.

In addition to complex statistics, the simple accuracy rate (percentage of exact matches) and the mean absolute rank error were also presented as descriptive measures to facilitate interpretation. In inter-method comparisons, the McNemar's exact test, which is suitable for paired data since the same scenarios were evaluated twice by the same model, was used to test the significance of the difference between zero-shot and one-shot methods. To test differences in success among the four models, Cochran's Q test (a paired-group test) was used, while the chi-square test was used to compare large discordance between models.

In clinical practice, preventing unnecessary imaging and avoiding missed diagnoses are vital. Therefore, the capacity of each model to correctly identify "Usually Not Appropriate" (red) tests (sensitivity) and its ability to avoid mistakenly labeling an appropriate test as inappropriate (specificity) were also calculated using the binary classification method. All data were analyzed using python (the pandas, SciPy, and statsmodels libraries), and statistical significance was defined as a two-sided p-value <0.05.

Results

Overall Agreement with ACR Goodness-of-Fit Criteria

When the total of 240 model responses obtained in the study was evaluated as a pooled sample, the overall perfect agreement (accuracy) rate between the LLMs and the ACR gold standard was 86.3% (207 of 240 responses). The linearly weighted overall kappa was 0.78 (95% CI, 0.70-0.85), corresponding to a clinically substantial level of agreement according to the Landis and Koch criteria. The mean absolute rank error was low across all analyzed responses. These data show that even when the models fail to match the ACR category perfectly, they generally deviate only into a neighboring category (one unit away) rather than giving a completely opposite response.

Inter-model Comparison

When the overall capacity of each model was examined by combining results from different request methods, the accuracies of the four models ranged from 81.7% to 95.0% (Table 1). The Claude Opus 4.8 model exhibited the highest performance, being the only model to fall into the "almost perfect" fit band with 95.0% accuracy and a weighted kappa value of 0.91 (95% CI: 0.77-1.00). The other three models, ChatGPT 5.5, Gemini 3.1 Pro, and DeepSeek V4 Pro, showed significant clinical agreement with kappa values of 0.72, 0.70, and 0.79, respectively.

Despite these observed percentage differences, no significant difference in accuracy was found when the models were tested using the Cochran Q test. Cochran's Q test showed no statistically significant difference in accuracy among the four models for either prompting condition (zero-shot: $p=0.082$; one-shot: $p=0.958$). This statistical result indicates that the apparent percentage differences in performance between the models may be attributable to the relatively small sample size in this study and should therefore be interpreted cautiously.

Examination of the "major discordance" rates, which indicate the frequency of clinically risky decisions made by the models, showed that all models kept these rates quite low. This rate was 3.3% for the Claude Opus 4.8 and DeepSeek V4 Pro, while it reached its highest value of 8.3% for the Gemini 3.1 Pro. The chi-square test revealed no statistically significant difference in the frequency of these clinical errors between the models ($p=0.533$).

The Effect of the Prompt Strategy

One important finding of the study is that the way the question is asked affects model performance. Presenting the models with a previously solved example (single-shot method), instead of simply describing the task, significantly increased the success rate compared to presenting no example (zero-shot method).

Table 1. Overall performance metrics of LLMs: Exact match accuracy, weighted Kappa, and Levels of agreement (combined zero-shot and one-shot prompting, n=60 responses per model)

Model	Exact match (accuracy)	Weighted kappa (95% CI)	Agreement level	Major discordance
ChatGPT 5.5	83.3%	0.72 (0.54-0.87)	Substantial	6.7%
Gemini 3.1 Pro	81.7%	0.70 (0.51-0.85)	Substantial	8.3%
Claude Opus 4.8	95.0%	0.91 (0.77-1.00)	Almost perfect	3.3%
DeepSeek V4 Pro	85.0%	0.79 (0.64-0.91)	Substantial	3.3%

CI: Confidence interval, LLMs: Large language models, agreement level based on Landis and Koch criteria. Major discordance = a two-category error ("Usually Appropriate" scored as "Usually Not Appropriate," or the reverse)

When responses from all models were combined, overall accuracy increased from 81.7% in the zero-shot setting to 90.8% in the one-shot setting. Similarly, the overall kappa agreement increased from 0.71 ("significant") to 0.85 ("almost perfect").

Examination of individual models indicated that this positive effect was consistently observed in three of the four models (Table 2). The McNemar test, applied to paired data, approached but did not reach the statistical significance threshold of $p<0.05$ (observed $p=0.071$). This situation reveals that a single example confers a real structural benefit to the models, but larger sets of scenarios are needed to statistically validate this benefit. Only the Claude Opus 4.8 model performed slightly better when asked without an example than when asked with an example; however, this model remained within a very high fit range under both methods.

Notably, both Cochran's Q results indicate that one-shot prompting narrowed the performance gap between models: correct-answer counts under zero-shot prompting ranged from 22 to 29 out of 30 across the four models ($Q=6.69$, $p=0.082$), whereas under one-shot prompting, this range narrowed to 27-28 out of 30 ($Q=0.31$, $p=0.958$). This suggests that providing a single worked example not only improved average accuracy but also reduced variability between different LLMs.

Identifying Inappropriate and Appropriate Tests

In clinical practice, failing to identify a test in the "Usually Not Appropriate" category can result in unnecessary testing, while failing to recommend a "Usually Appropriate" test can lead to missed diagnoses. To examine the safety profile at these two extremes, the sensitivity and specificity values of the models were calculated (Table 3). All four models correctly identified most inappropriate tests, with sensitivity rates ranging from 79% (ChatGPT 5.5) to 100% (Claude Opus 4.8). Specificity—the ability to avoid mistakenly labeling an appropriate test as "inappropriate"—was high, with models' performance ranging from 89% to 96%.

For "Usually Appropriate" tests, the calculated sensitivity values ranged from 82% to 91%, and the specificity values ranged from 88% to 100%. Claude Opus 4.8 demonstrated the best performance across all safety thresholds in both extreme scenarios.

Table 2. Comparative analysis of model performance by prompting strategy: evaluating accuracy and weighted Kappa for zero-shot versus one-shot approaches (n=30 responses per cell)

Model	Prompting strategy	Accuracy	Weighted kappa
ChatGPT 5.5	Zero-shot	76.7%	0.61
ChatGPT 5.5	One-shot	90.0%	0.84
Gemini 3.1 Pro	Zero-shot	73.3%	0.59
Gemini 3.1 Pro	One-shot	90.0%	0.81
Claude Opus 4.8	Zero-shot	96.7%	0.92
Claude Opus 4.8	One-shot	93.3%	0.89
DeepSeek V4 Pro	Zero-shot	80.0%	0.73
DeepSeek V4 Pro	One-shot	90.0%	0.85

Claude Opus 4.8 was the only model that performed marginally better with zero-shot prompting than with one-shot prompting, although both results indicate very good agreement

Table 3. Diagnostic reliability of AI models: sensitivity and specificity in classifying "Usually Not Appropriate" and "Usually Appropriate" clinical guideline categories

Model	"Not Appropriate" sensitivity	"Not Appropriate" specificity	"Appropriate" sensitivity	"Appropriate" specificity
ChatGPT 5.5	79%	96%	82%	88%
Gemini 3.1 Pro	86%	89%	82%	88%
Claude Opus 4.8	100%	96%	91%	100%
DeepSeek V4 Pro	93%	93%	82%	96%

ACR: American College of Radiology, "Not Appropriate" = ACR "Usually Not Appropriate" category; "Appropriate" = ACR "Usually Appropriate" category. Values are one-versus-rest sensitivity/specificity for each model detecting that category correctly

Distribution According to Clinical Presentations

The fit of the models to the ACR standards was not equally distributed across the nine clinical scenarios considered. Descriptive data, which were not subjected to formal statistical testing because of the small sample size, indicate that the models achieved their greatest success in cases with very clear diagnostic boundaries, such as suspected acute mesenteric ischemia (100%) and left upper quadrant pain (93.8%). In contrast, a significant decrease in model accuracy was observed in scenarios such as suspected small bowel obstruction (79.2%) and right lower quadrant pain/suspected appendicitis (80.0%), which present numerous gray areas in clinical practice. This table appears directly related to the clarity of disease markers in the models' training data.

Major-discordance responses were unevenly distributed across scenarios: 13 of the 240 responses (5.4%) met this definition and clustered in three clinical panels—right lower quadrant pain/suspected appendicitis, right upper quadrant pain, and epigastric pain/pancreatitis—which together accounted for 8 of the 13 major-discordance responses. These are precisely the presentations in which the correct ACR category depends on secondary features (pregnancy status, presence of a complication, intensive care unit setting, an equivocal prior ultrasound) rather than on the primary symptom alone, suggesting that current models are more likely to overlook a qualifying clinical detail than to misjudge the primary presentation itself. The complete case-by-case response matrix for all 30 scenarios and all four models is provided in Supplementary Material S2.

Discussion

The findings of this research reveal that AI-based LLMs have reached a highly competent level in interpreting complex medical texts and matching them with evidence-based clinical guidelines. The overall accuracy rate of 86.3% and the weighted kappa score of 0.78, both achieved by pooling data from four different language models, demonstrate that the models successfully placed a patient history in text format within the strict clinical boundaries set by the ACR.

As noted in extensive studies by Zaki et al. (10) or reports by Rau et al. (9), supporting physicians with an algorithmic system when ordering radiological examinations both increases diagnostic success and prevents unnecessary examination costs. The available data also reinforce the claims in the medical literature that these current language models provide a solid infrastructure for their integrated use in radiology clinical decision support systems. Detailed comparison of the data across models shows that all systems exceed a certain level of clinical competence. Statistical tests did not show dramatic differences between the models. Nevertheless, it is noteworthy that the Claude Opus 4.8 model achieved both a high accuracy rate and high agreement values compared to other current models tested. Other analyses comparing various models in radiology cases indicate that Claude-series models sometimes produce more consistent responses when interpreting clinical event sequences than ChatGPT and Gemini derivatives (13). This difference may be related to how intensively the models are exposed to the semantic structure contained in medical journals, case reports, or guidelines such as ACR during their initial training (8). Medical reasoning processes are not simply about concatenating words; the model needs to capture the most critical detail in the patient's clinical presentation and assign it to the correct guideline category. In our study, the Claude Opus 4.8 model appears to perform this synthesis in a more refined manner.

Another finding of our study concerns how a clinical scenario is presented to the AI. Adding a single clinical-scenario example to the question stem significantly improved the models' overall performance and shifted the level of fit from the "significant" range to the "nearly perfect" range. The generalizing nature inherent in LLMs can sometimes lead to models producing overly creative or vague medical advice (11). Just as the human mind understands the framework by looking at a solved example before executing a complex instruction, language models similarly narrow their large vocabulary in response to a single presented example and align themselves with the desired triple "ACR category format." Notably, Claude Opus 4.8 demonstrated high accuracy without disrupting its structure, even when presented without an example. This suggests that the medical question-answer logic may already be preserved within the model's internal parameters, albeit within very distinct limits. As highlighted in the research of Yao et al. (14), it is possible to improve the clinical fit of models from the level of ordinary physicians to the level of experts by applying the correct prompt to the models or providing them with a structural context (8).

Another outcome of our study is that, when viewed through the lens of the most fundamental rule of clinical practice, "first, do no harm," the frequency of models producing dangerous and erroneous decisions remained in the single-digit percentages across all evaluated models. The fact that errors with the potential to harm patients (for example, stating that an essential imaging procedure is unnecessary or recommending a test that should not be performed) constituted a very small minority of the 240 cases examined proves that models can be reliable assistants. An AI capable of identifying the "Usually Not Appropriate" category with 100% accuracy could act as a safety filter in emergency departments, preventing unnecessary tests that might be

overlooked by busy and fatigued physicians (11). However, the decline in the performance of language models in cases where diagnostic boundaries are ambiguous and physicians face the greatest uncertainty, for example right lower quadrant pain that could indicate appendicitis or small bowel obstruction, indicates that AI has not yet fully achieved the delicate balance that an experienced clinician achieves by combining physical examination findings and intuition. Therefore, AI should be positioned not as an authority that replaces the physician but as a consultant that illuminates their blind spots.

An additional consideration is automation bias, the tendency of a busy clinician to accept a system's recommendation without independent verification, particularly under time pressure. If an LLM incorrectly labels a necessary test as "Usually Not Appropriate," and this recommendation is accepted uncritically, the result is not merely a statistical error, but a missed diagnosis with direct medical and legal consequences; responsibility for the final imaging decision and for any resulting harm remains with the ordering physician regardless of the AI input. This underscores that, at the current stage of development, LLM output should be positioned as one input among several, to be reviewed by a clinician before it changes patient management, rather than as an autonomous gatekeeper for imaging decisions.

Regarding the strengths of our study, the latest generation of LLMs was compared impartially using direct, simultaneous, and paired data. To support the clinical relevance of the findings, the analysis was expanded to include not only with simple accuracy percentages but also with weighted kappa statistics, which numerically correct for the severity of errors. The ability to explain how demand strategies affect the model through clinical decision-making mechanisms distinguishes this study from fundamental research in the literature.

Study Limitations

In addition to these strengths, the study has some limitations. The most significant limitation is the small number of cases tested. The small number of scenarios reduced statistical power in the subgroup analyses, preventing clear differentiation of performance between models. The second major limitation is that the clinical cases used were specially prepared, cleaned, and noise-free "ideal" scenarios for training purposes. In real-world hospital settings, physician notes in patient information systems are often disorganized, contain many abbreviations, are incomplete, or contain spelling errors. It would not be accurate to assume that the success of models in these organized, idealized texts will be identical when applied to real-life data. Because the vignettes were constructed to closely mirror the illustrative scenarios in the ACR documents, the scenario text may share phrasing patterns with material to which the models were plausibly exposed during training. This could inflate model performance relative to genuinely novel, real-world clinical narratives. Independent validation using clinician-authored or retrospective cases is an important next step. This focuses only on a small, specific branch of medicine. Since ACR criteria may involve more complex clinical specialties such as traumatology, oncologic follow-up, or neuroradiology, the success levels of the models in those scenarios are unknown.

In light of these considerations, it is essential to take several steps in future clinical and academic studies. First, to measure the true clinical performance of the models, large-scale studies should be designed using real retrospective electronic health records while ensuring patient confidentiality, rather than relying on synthetic cases. Second, analyses should be expanded to include thousands of cases, rather than hundreds, thereby increasing statistical power to cover all sub-branches of radiology. Furthermore, the performance of generic models should be compared with specific models that have been fine-tuned directly with medical data or that have their information confirmed by instantly connecting to external guidelines using techniques such as "Retrieval-Augmented Generation." Beyond fine-tuning or retrieval augmentation, future systems could be designed to reference multiple evidence sources simultaneously, for example, by cross-checking a recommendation against both the ACR Appropriateness Criteria and other relevant guideline bodies, and to present the supporting evidence profile alongside the recommended category, rather than providing a single categorical answer. Such transparency would allow the ordering physician to see not just what the model recommends, but why, which is likely to be a prerequisite for clinical trust and adoption. Finally, the integration of multimodal systems—where language models can analyze both text and a patient's previous X-ray or ultrasound, in alignment with ACR guidelines— should be the next logical focus of medical technologies.

Conclusion

Our study shows that modern LLMs can accurately align radiological imaging recommendations for complex acute abdominal clinical conditions with the scientific standards set by the ACR. Demonstrating the solution to a sample problem in guideline format, rather than simply giving instructions, increased model performance. By helping physicians reach a guideline-concordant imaging decision more quickly, such support has the potential to shorten decision times in the emergency department and reduce patients' exposure to unnecessary ionizing radiation. However, algorithms have not yet reached the level of clinicians'

experience in complex disease presentations, and more extensive research using real-world hospital automation data is needed.

Ethics

Ethics Committee Approval: This study was completed using current versions of publicly available large language models, drawing on ACR's publicly available guidelines. Since no patient data, hospital records, or images were used, ethical committee approval was not required.

Informed Consent: This is cross-sectional study.

Data availability: The prompt texts and clinical scenarios used in the study are provided as supplementary material.

Footnotes

Authorship Contributions

Concept: A.E.A., Design: A.E.A., H.İ.Ş., Data Collection or Processing: H.İ.Ş., F.C.T., Analysis or Interpretation: A.E.A., F.C.T., Literature Search: H.İ.Ş., Writing: A.E.A.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

Supplementary Material: https://d2v96fxpocvxx.cloudfront.net/72b4fafb-5db4-4d61-bedf-c8100360bb15/content-images/6633bd7b-5abf-47ef-90f1-aae68082fdb9.pdf

References

1. Scheirey CD, Fowler KJ, Therrien JA, Kim DH, Al-Refaie WB, Camacho MA, et al. ACR Appropriateness Criteria(®) acute nonlocalized abdominal pain. J Am Coll Radiol. 2018;15:S217-31.
2. Cartwright SL, Knudson MP. Diagnostic imaging of acute abdominal pain in adults. Am Fam Physician. 2015;91:4529.
3. Russo GK, Zaheer A, Kamel IR, Porter KK, Archer-Arroyo K, Bashir MR, et al. ACR Appropriateness Criteria® Right Upper Quadrant Pain: 2022 Update. J Am Coll Radiol. 2023;20:S211-23.
4. Üçdal M, Ekingen E. Multidisciplinary AI-assisted diagnosis and management of acute abdominal pain syndrome: a pilot study. Eurasian J Emerg Med. 2026;25:272-4.
5. Chang KJ, Marin D, Kim DH, Fowler KJ, Camacho MA, Cash BD, et al. ACR Appropriateness Criteria® Suspected Small-Bowel Obstruction. J Am Coll Radiol. 2020;17:S305-14.
6. Kambadakone AR, Santillan CS, Kim DH, Fowler KJ, Birkholz JH, Camacho MA, et al. ACR Appropriateness Criteria® Right Lower Quadrant Pain: 2022 Update. J Am Coll Radiol. 2022;19:S445-61.
7. Pambudi S, Menolascina F. Bridging clinical narratives and ACR appropriateness guidelines: a multi-agent RAG system for medical imaging decisions. arXiv. 2025.
8. Papageorgiou PS, Christodoulou RC, Pitsillos R, Petrou V, Vamvouras G, Kormentza EV, et al. The role of large language models in improving diagnostic-related groups assignment and clinical decision support in healthcare systems: an example from radiology and nuclear medicine. Appl Sci. 2025;15:9005.
9. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A Context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology. 2023;308:e230970.
10. Zaki HA, Aoun A, Munshi S, Abdel-Megid H, Nazario-Johnson L, Ahn SH. The application of large language models for radiologic decision making. J Am Coll Radiol. 202;21:1072-8.
11. Tziakouri A, Menolascina F. Reinforcement learning for clinical reasoning: aligning LLMs with ACR imaging appropriateness criteria. arXiv [Preprint]. 2025;arXiv:2510.05194.
12. American College of Radiology. ACR Appropriateness Criteria® [Internet]. [cited 2026 Jul 1]. Available from: https://acsearch.acr.org/list.
13. Sonoda Y, Kurokawa R, Nakamura Y, Kanzawa J, Kurokawa M, Ohizumi Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "diagnosis please" cases. Jpn J Radiol. 2024;42:1231-5.
14. Yao MS, Chae A, Saraiya P, Kahn CE Jr, Witschey WR, Gee JC, et al. Evaluating acute image ordering for real-world patient cases via language model alignment with radiological guidelines. Commun Med (Lond). 2025;5:332.