

Article

Chatting Ain't Diagnosing: Diagnostic Variability and Fundamental Errors in Multimodal LLM Interpretation in Radiology

Sungjoon Hong ¹, Mihir Matalia ² and Milan Toma ^{1,*}

¹ Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA; shong09@nyit.edu

² Academic Technologies Group, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA; mihir.matalia@nyit.edu

* Correspondence: mtoma@nyit.edu or tomamil@tomamil.com

Abstract

Recent studies investigating the diagnostic capabilities of large language models (LLMs) have attracted significant media attention, often resulting in headlines claiming that AI systems can match or even outperform physicians. As LLMs have rapidly proliferated, this has fueled a widespread misconception that they represent the cutting edge of artificial intelligence in all contexts. This narrative tends to overshadow the continued importance of task-specific machine learning models, which were developed and validated for particular diagnostic applications well before the rise of LLMs. This single-case study evaluated the reliability of five leading multimodal LLMs (GPT-5, Gemini 3 Pro, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended) for radiological image interpretation by presenting each model with an identical non-contrast head CT demonstrating intracranial pathology, complemented by a novel cross-evaluation protocol wherein each model graded all responses. The deliberate use of a straightforward case (rather than diagnostically challenging pathology) aimed to establish minimum competency thresholds; if LLMs cannot reliably interpret obvious pathology, their deployment on ambiguous cases becomes indefensible. The study intentionally excluded human radiologist ground truth to avoid generating comparative accuracy metrics that could be selectively cited for commercial purposes, focusing instead on demonstrating class-wide limitations rather than ranking individual products. Results revealed a 20% rate of fundamental diagnostic error, with one model misidentifying ischemic stroke as intracerebral hemorrhage with incorrect lateralization. Even among concordant models, clinically meaningful variability persisted in acuity characterization, anatomical localization, and differential diagnoses. Cross-evaluation exposed ground truth disagreement between models, self-evaluation bias, inconsistent grading stringency, and divergent evaluation philosophies. Only one model included appropriate safety disclaimers. These findings demonstrate that current multimodal LLMs exhibit unacceptable diagnostic variability and evaluative inconsistency for autonomous clinical deployment. The appropriate clinical role for LLMs should be distinguished by deployment context: autonomous diagnosis requires validated task-specific models; decision support applications demand rigorous radiologist oversight protocols; and educational summarization represents the most appropriate current use case, with mandatory disclaimers. Healthcare applications requiring reliable image interpretation should prioritize validated, task-specific machine learning systems over general-purpose language models.



Academic Editors: Pietro Hiram Guzzi and Guanqiu Qi

Received: 7 January 2026

Revised: 6 February 2026

Accepted: 24 February 2026

Published: 25 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: large language models; multimodal artificial intelligence; medical image interpretation; radiology; computed tomography; diagnostic accuracy; inter-model variability; ischemic stroke; artificial intelligence in healthcare; AI safety

1. Introduction

The public discourse surrounding artificial intelligence (AI) has become increasingly synonymous with large language models (LLMs), i.e., systems such as ChatGPT, Claude, Gemini, and their contemporaries that have captured global attention since 2022. This conflation, while understandable given the unprecedented visibility these tools have achieved, obscures a fundamental reality: AI as a field has been under active development for over half a century, and LLMs represent merely the most recent (and in many respects, most narrowly specialized) manifestation of a far broader technological landscape.

1.1. A Brief History of Artificial Intelligence

Understanding why large language models are ill-suited for medical image interpretation requires situating them within the broader landscape of artificial intelligence, a field of which seven-decade history has produced diverse approaches to machine cognition, each with distinct strengths and limitations. The current public fascination with LLMs obscures this rich technological heritage and, more critically, the specialized diagnostic tools that emerged from it.

The foundations of artificial intelligence were laid in the 1950s, with Alan Turing's seminal 1950 paper "Computing Machinery and Intelligence" [1] and the 1956 Dartmouth Conference [2] that formally established AI as a field of academic inquiry. In the decades that followed, AI research progressed through multiple paradigms: symbolic reasoning and expert systems in the 1960s–1980s, neural network revivals in the 1980s–1990s, and the machine learning renaissance of the 2000s–2010s driven by increased computational power and data availability [3–5]. Deep learning architectures, particularly convolutional neural networks (CNNs), achieved breakthrough performance in image recognition tasks by the early 2010s [6], years before the transformer architecture that underlies modern LLMs was even conceived.

The transformer architecture was introduced in 2017, and foundational models such as BERT and GPT emerged in 2018 [7,8]. However, it was not until GPT-3 in 2020 (with its 175 billion parameters) that LLMs began to demonstrate the conversational fluency that would later captivate public attention [9,10]. The release of ChatGPT in November 2022 marked the inflection point at which LLMs transitioned from technical curiosities to mainstream phenomena [11]. In this context, the entire history of publicly prominent LLMs spans barely three years; which is a tiny fraction of AI's seven-decade evolution [11,12].

1.2. The Misconception: LLMs as the Totality of AI

The explosive visibility of LLMs has fostered a pervasive misconception: that "AI" and "LLM" are interchangeable terms, and that these language-centric systems represent the state of the art across all AI applications. Such conflation complicates ongoing efforts to define AI's transformative potential in healthcare and navigate its ethical implications [13]. Misunderstanding the distinction can have serious consequences, particularly in domains where task-specific machine learning models (developed and validated long before the LLM era) remain the appropriate technological choice.

The very nomenclature is instructive. Large language models are, by design and training objective, optimized for linguistic tasks: text generation, summarization, translation,

and conversational interaction. Their architecture (i.e., the transformer) was developed specifically for sequence-to-sequence modeling in natural language processing. While multimodal variants have emerged that can process images alongside text, with applications extending to domains such as orthodontic biomechanical reasoning [14], these capabilities represent extensions grafted onto a fundamentally language-centric architecture, not systems purpose-built for visual analysis.

The viral launch of ChatGPT in November 2022 marked an inflection point in public awareness of artificial intelligence capabilities (Figure 1). Within months, large language models transitioned from specialized research tools to household technologies, fundamentally altering public expectations regarding AI performance across diverse domains. This rapid mainstreaming, however, has fostered a pervasive misconception: that general-purpose LLMs represent the state of the art across all artificial intelligence applications, including medical image interpretation.

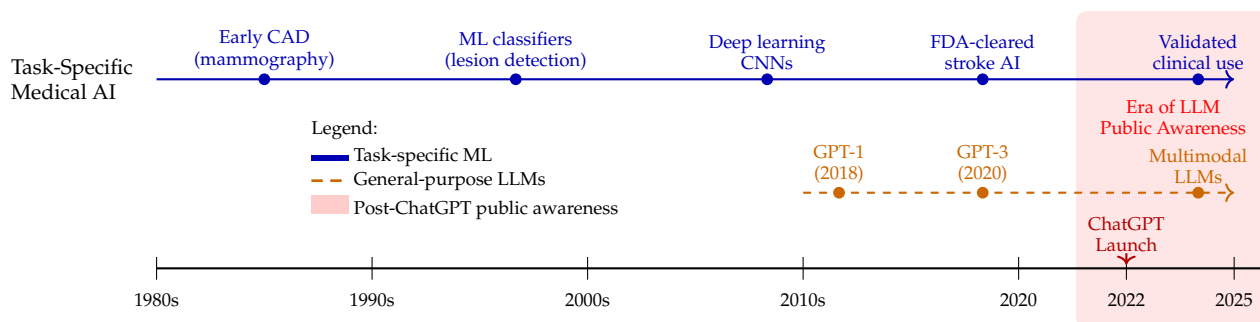


Figure 1. Timeline illustrating the divergent development trajectories of task-specific medical machine learning and general-purpose large language models. Task-specific ML systems for medical imaging (blue) have undergone decades of development, validation, and regulatory clearance, culminating in FDA-cleared diagnostic AI deployed in clinical practice. General-purpose LLMs (orange, dashed) emerged more recently and achieved widespread public awareness following the ChatGPT launch in November 2022 (shaded region).

As illustrated in Figure 1, this perception obscures a critical distinction. Task-specific machine learning models (developed, validated, and in many cases regulatory-cleared for defined diagnostic imaging tasks) have been advancing through detailed clinical validation pathways for decades. These systems remain the appropriate technological choice for applications requiring reliable, reproducible diagnostic accuracy. The conflation of LLM capabilities with validated medical AI risks inappropriate deployment of unvalidated tools in contexts where patient safety demands proven performance.

Recent studies examining LLM diagnostic capabilities have garnered substantial media attention, often generating headlines that proclaim AI systems can match or exceed physician performance [15,16]. Figure 2 illustrates findings from a randomized controlled trial by Goh et al., in which GPT-4 alone achieved a diagnostic reasoning score of 92%, i.e., significantly higher than physicians using conventional resources (74%, $p = 0.03$) [15]. Such results, when reported without context, fuel public perception that LLMs represent reliable diagnostic tools ready for clinical deployment. However, it is necessary to note that the Goh study evaluated LLM performance on text-based clinical vignettes, i.e., a task that leverages the core linguistic reasoning capabilities for which these models were designed. Medical diagnosis in practice frequently requires interpretation of imaging studies, ranging from cross-sectional imaging to specialized reports such as visual field tests [17], a fundamentally different task that tests the multimodal visual capabilities of these systems rather than their text-processing strengths.

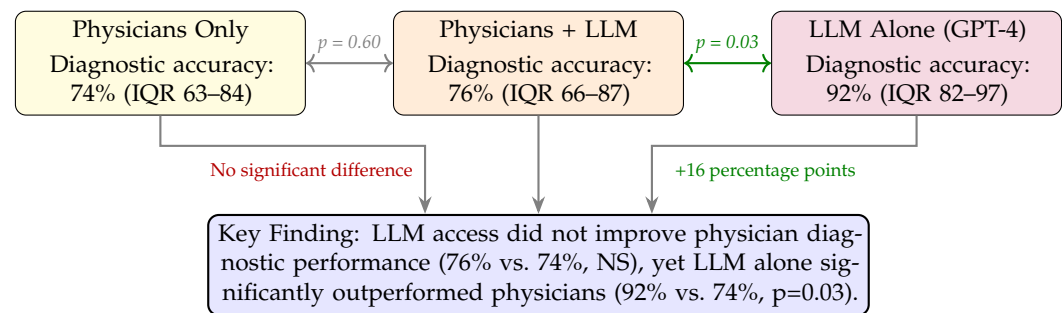


Figure 2. Summary of key findings from a randomized controlled trial examining physician diagnostic reasoning with and without LLM access [15]. Despite the LLM alone achieving significantly higher scores (92%) than physicians using conventional resources (74%, $p = 0.03$), physician access to the LLM produced no significant improvement (76% vs. 74%, $p = 0.60$).

1.3. Machine Learning in Radiology: A Parallel History

The application of machine learning to medical imaging predates the LLM era by decades, as illustrated in Figure 1. Foundational work began in the early 1960s, when Lodwick et al. introduced the first computational analysis of chest X-rays for lung cancer classification using manual feature coding and rule-based computer analysis [18]. Lodwick subsequently coined the term “computer-aided diagnosis” (CAD) and outlined a research plan for automated classification of radiological images [19]. Throughout the 1970s, these methods advanced significantly: Kruger et al. applied automated feature extraction and statistical classification to chest radiographs for cardiac analysis [20], Toriwaki et al. employed image processing and pattern recognition to automatically measure the cardiothoracic ratio [21], and further work by Kruger et al. advanced pattern recognition techniques for automated abnormality detection in chest radiographs [22,23]. These early efforts laid the groundwork for CAD systems for mammography, which received FDA approval as early as 1998 [24]. Deep learning approaches to radiological image analysis began demonstrating expert-level performance in the mid-2010s, with landmark studies showing CNN-based systems matching or exceeding dermatologist accuracy in skin lesion classification (2017) [25], radiologist performance in chest X-ray interpretation (2017–2018) [26], and ophthalmologist accuracy in diabetic retinopathy screening (2016) [27]. Recent comprehensive reviews have further documented the expanding applications of machine learning across medical imaging modalities [28,29].

These systems share a critical characteristic: they are task-specific. A CNN trained to detect acute ischemic stroke on non-contrast head CT has been optimized exclusively for that task, with training datasets comprising thousands to millions of expert-annotated images of that specific pathology in that specific imaging modality [30]. This focused training enables the model to learn the precise visual features (subtle hypodensities, loss of gray–white differentiation, insular ribbon signs, hyperdense vessel signs) that distinguish pathology from normalcy in that defined context. Similar validation strategies have been applied successfully to automated MRI-based classification tasks in neurological diagnosis, demonstrating the efficacy of task-specific approaches [31].

By contrast, multimodal LLMs acquire their image interpretation capabilities as a byproduct of training on heterogeneous image–text pairs scraped from diverse internet sources. Medical imaging content constitutes a small and potentially unrepresentative fraction of this training corpus [32]. The resulting systems possess broad but shallow capabilities: they can describe what they “see” in an image using medically plausible language, but their visual representations have not been optimized for the discriminative features that matter for specific diagnostic tasks. Comparative studies examining multi-

modal LLMs against purpose-built feature extraction methods in clinical contexts, such as autism prediction from home videos, further illuminate these architectural limitations [33].

1.4. The Critical Distinction: Language Generation Versus Diagnostic Accuracy

The core competency of an LLM is generating contextually appropriate text [34]. When presented with a medical image, a multimodal LLM generates a textual description that is linguistically coherent and stylistically consistent with radiological reporting conventions. This output may appear authoritative (e.g., formatted in bullet points, employing correct anatomical terminology, offering differential diagnoses) yet the relationship between this generated text and the actual pixel-level content of the image is fundamentally different from that of a purpose-trained diagnostic model.

A task-specific stroke detection algorithm processes the image through layers of learned convolutional filters optimized to detect the visual signatures of ischemic injury. Standardized validation frameworks for clinically actionable machine learning models emphasize the importance of detailed performance characterization across diverse datasets [35]. Its output (i.e., stroke present/absent, with localization) is a direct function of learned visual features validated against confirmed diagnoses. An LLM, by contrast, generates a textual response that is statistically consistent with the kinds of reports that accompanied similar-seeming images in its training data. The LLM is not “diagnosing” in any meaningful sense; it is predicting what text a radiologist might have written, based on pattern matching against its training distribution.

This distinction has profound implications for reliability. A well-validated task-specific model has quantified performance characteristics: known sensitivity, specificity, positive and negative predictive values, and characterized failure modes. Its errors are systematic and discoverable through rigorous validation. An LLM’s errors, by contrast, are emergent properties of a system not designed for the task at hand, potentially unpredictable, and difficult to characterize systematically.

1.5. The Present Study: Empirical Demonstration of LLM Limitations

This study was designed to empirically demonstrate the limitations of current multimodal LLMs for medical image interpretation, thereby illustrating why the broader AI ecosystem (particularly task-specific machine learning models) remains essential for clinical diagnostic applications.

We evaluated five leading multimodal LLMs (GPT-5, Gemini 3 Pro, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended) on a standardized radiological interpretation task: analysis of a non-contrast head CT demonstrating intracranial pathology. Beyond assessing diagnostic accuracy, we implemented a novel cross-evaluation protocol in which each LLM graded the responses of all five models, enabling assessment of evaluative consistency and the reliability of LLM-based quality assurance.

Our findings reveal that current LLMs exhibit unacceptable variability in diagnostic interpretation, including fundamental errors (misidentifying ischemic stroke as hemorrhage with incorrect lateralization) and categorical disagreements about basic radiological findings (acute versus chronic infarction). Even among models that reached the correct general diagnosis, clinically meaningful variability persisted in acuity characterization, anatomical localization, and differential diagnosis generation. The cross-evaluation protocol further revealed that LLMs cannot reliably agree on radiological ground truth, undermining proposals for their use in automated quality assurance or educational assessment.

Importantly, this study was designed to demonstrate the general unreliability of multimodal LLMs for diagnostic imaging interpretation, not to rank or compare the performance of individual models against one another or against human radiologists. The authors have

no relationship (financial, professional, or otherwise) with any AI system or developer, including those whose products were evaluated (OpenAI, Google DeepMind, Meta AI, xAI, or Anthropic), nor with any other AI systems or developers not included in this study. We deliberately avoided comparative accuracy assessments against human radiologist ground truth because such comparisons risk shifting focus toward determining “which LLM performs best,” which could inadvertently serve as product endorsement. Our objective is not to identify superior products within an unreliable class of tools. Instead, we aim to demonstrate that the entire class of general-purpose multimodal LLMs exhibits fundamental limitations that render them unsuitable for autonomous diagnostic imaging interpretation.

These findings support the position that LLMs, despite their impressive linguistic capabilities, are not appropriate tools for medical image interpretation requiring diagnostic accuracy. The appropriate AI technology for such applications remains task-specific machine learning models: systems designed, trained, and validated specifically for defined imaging modalities and pathological conditions. LLMs may serve valuable roles in healthcare (clinical documentation, patient communication, literature synthesis) but the pattern recognition demands of diagnostic radiology require purpose-built solutions that the current generation of language models cannot reliably provide.

The sections that follow detail our methodology, present our findings, and discuss the implications of this work for the responsible integration of AI technologies into clinical practice. We argue that the current enthusiasm for LLMs must be tempered by recognition of their fundamental limitations and that healthcare institutions, AI developers, and regulatory bodies must maintain clear distinctions between general-purpose language systems and validated diagnostic AI tools.

2. Materials and Methods

To empirically evaluate the reliability of multimodal large language models for radiological image interpretation, we designed a two-phase experimental protocol. The first phase assessed diagnostic performance by presenting five leading multimodal LLMs with an identical medical imaging task and systematically comparing their outputs across clinically relevant dimensions. The second phase implemented a novel cross-evaluation protocol wherein each LLM graded the responses generated by all five models (including its own, presented anonymously), enabling assessment of evaluative consistency, self-evaluation bias, and inter-model agreement on radiological ground truth. This dual-phase approach allowed examination of LLM reliability from two complementary perspectives: their capacity to generate accurate diagnostic interpretations and their capacity to evaluate the quality of such interpretations; both of which are prerequisite competencies for any proposed clinical or educational deployment. The following subsections detail the selection criteria for models and test images, the standardized prompts employed, the grading rubric and cross-evaluation methodology, and the analytical framework used to assess inter-model variability and agreement.

2.1. Selection of Large Language Models

Five LLMs with vision capabilities were selected for this comparative analysis. The selection criteria included: (1) demonstrated ability to process and interpret medical imaging data, (2) public availability or API access for research purposes, (3) representation among the most widely used and cited multimodal AI systems as of late 2025, and (4) diverse model architectures and training paradigms to capture variability across different development approaches. The selected models were: GPT-5 (OpenAI), Gemini 3 Pro (Google DeepMind), Llama 4 Maverick (Meta AI), Grok 4 (xAI), and Claude Opus 4.5 Extended (Anthropic).

2.2. Image Selection and Prompt Design

A non-contrast axial CT scan of the head demonstrating intracranial pathology was selected as the test image (Figure 3). No external radiologist ground truth labeling was incorporated into the study design. This methodological choice was deliberate: the selected image depicts unambiguous intracranial pathology that does not require subspecialty expertise to recognize as abnormal. The relevant clinical question is not whether expert radiologists can correctly interpret this straightforward case (they can), but whether state-of-the-art LLMs can achieve consistent, accurate interpretations when presented with a case designed to be easy. By selecting an obviously pathological image rather than a diagnostically challenging case, we aimed to establish a minimum competency threshold. If LLMs cannot reliably interpret straightforward pathology, their deployment on ambiguous or challenging cases becomes indefensible. Furthermore, incorporating human radiologist grading would have introduced a comparative dimension (LLM vs. human accuracy) that risks framing certain models as “better” or “closer to human performance,” which is a characterization we specifically sought to avoid given our lack of commercial relationships with any AI developer and our objective of demonstrating class-wide limitations rather than identifying superior products.



Figure 3. Non-contrast axial CT scan of the head at the level of the lateral ventricles and basal ganglia, demonstrating intracranial pathology. This image served as the standardized test case for all five LLM evaluations. Ground truth interpretation and clinical implications are discussed in the Results section.

The same standardized prompt was submitted to all five LLMs to ensure consistency in evaluation parameters. The prompt was designed to elicit responses across multiple clinically relevant dimensions:

A standardized prompt was submitted to all five LLMs instructing them to analyze the image as a board-certified radiologist across five domains: imaging modality, anatomical region, primary diagnosis, supporting features, and differential diagnoses. The complete prompt text is provided in Appendix A.1, and all technical parameters relevant to reproducibility are documented in Section 2.5.

2.3. Data Collection and Analysis

Responses from each LLM were collected and systematically compared across the following categories: (1) imaging modality identification, (2) anatomical localization, (3) primary diagnosis, (4) supporting radiological features, and (5) differential diagnoses. Inter-model agreement was assessed qualitatively, with particular attention to clinically significant discrepancies that could impact patient management.

2.4. Cross-Model Evaluation Protocol

To further assess the reliability and consistency of LLM-generated radiological interpretations, a cross-evaluation protocol was implemented wherein each of the five LLMs was tasked with grading the responses generated by all five models (including its own output, presented anonymously). This approach enabled assessment of inter-model agreement not only in diagnostic interpretation but also in the evaluation of diagnostic quality, which is a critical consideration if LLMs are to be used for quality assurance or educational feedback in clinical settings.

2.4.1. Grading Prompt Design

A standardized grading prompt was developed based on principles of medical education assessment, structured evaluation frameworks, and prompt engineering best practices. The prompt instructed each LLM to assume the role of an expert medical educator and radiologist, systematically evaluating each of the five anonymized analyses against standard radiological criteria. The complete grading prompt was as follows:

A standardized grading prompt was developed instructing each LLM to assume the role of an expert medical educator and radiologist, systematically evaluating each of the five anonymized analyses against standard radiological criteria. The complete grading prompt, including all rubric details and stepwise instructions, is reproduced in Appendix A.2.

2.4.2. Grading Rubric and Scoring Domains

A structured grading rubric was provided to each LLM to ensure standardized evaluation across all assessments. Five domains were evaluated, each with domain-specific scoring criteria (Table 1). Complete detailed scoring criteria for all domains are provided in Appendix B.

Table 1. Condensed grading rubric for cross-model evaluation. Complete scoring criteria are provided in Appendix B.

Domain	Scale	Brief Description
Imaging Modality	0–3	Correct identification and specificity of imaging modality
Anatomical Region and Organ	0–3	Accurate localization of anatomical structures
Most Likely Diagnosis	0–4	Correctness and clinical appropriateness of primary diagnosis
Key Image Features	0–4	Completeness and accuracy of radiological feature descriptions
Differential Diagnoses	0–4	Plausibility, prioritization, and justification of alternatives
Maximum Total Score	18 points	

2.4.3. Stepwise Evaluation Process

To promote systematic reasoning and reduce variability, the grading prompt instructed each LLM to follow a stepwise chain-of-thought evaluation process. Rather than providing an external reference standard, each grading LLM used its own expert knowledge to establish the expected correct answer for each domain. The complete stepwise instructions, quality assurance guidelines, and rationale for this self-referential grading paradigm are provided in Appendix C.

2.4.4. Cross-Evaluation Score Matrices

To systematically organize and analyze the cross-evaluation data, five separate 5×5 scoring matrices were generated; one for each evaluation domain. In each matrix, rows represent the grading LLM (GPT-5, Gemini 3 Pro, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended), and columns represent the graded analysis (Analysis 1 through Analysis 5, corresponding to the anonymized outputs of the five models). Each cell M_{ij} contains the score assigned by grading model i to analysis j for the respective domain.

The five domain-specific matrices are defined as follows:

- $\mathbf{M}_{\text{modality}}$: Imaging Modality scores (0–3 scale)
- $\mathbf{M}_{\text{anatomy}}$: Anatomical Region/Organ scores (0–3 scale)
- $\mathbf{M}_{\text{diagnosis}}$: Most Likely Diagnosis scores (0–4 scale)
- $\mathbf{M}_{\text{features}}$: Key Image Features scores (0–4 scale)
- $\mathbf{M}_{\text{differential}}$: Differential Diagnoses scores (0–4 scale)

An example matrix structure is illustrated in Table 2.

Table 2. Example structure of a 5×5 cross-evaluation scoring matrix for a single domain (see Appendix B for complete scoring criteria). Rows represent grading LLMs; columns represent graded analyses.

	Analysis 1	Analysis 2	Analysis 3	Analysis 4	Analysis 5
GPT-5	$M_{1,1}$	$M_{1,2}$	$M_{1,3}$	$M_{1,4}$	$M_{1,5}$
Gemini 3 Pro	$M_{2,1}$	$M_{2,2}$	$M_{2,3}$	$M_{2,4}$	$M_{2,5}$
Llama 4 Maverick	$M_{3,1}$	$M_{3,2}$	$M_{3,3}$	$M_{3,4}$	$M_{3,5}$
Grok 4	$M_{4,1}$	$M_{4,2}$	$M_{4,3}$	$M_{4,4}$	$M_{4,5}$
Claude Opus 4.5 Ext.	$M_{5,1}$	$M_{5,2}$	$M_{5,3}$	$M_{5,4}$	$M_{5,5}$

This matrix structure enables several analytical approaches:

- Column analysis: Aggregating scores within each column reveals the consensus evaluation of each analysis across all grading models, highlighting which responses were consistently rated highly or poorly.
- Row analysis: Examining score distributions within each row reveals grading tendencies of individual LLMs, such as systematic leniency or stringency.
- Diagonal analysis: The diagonal elements (M_{ii}) represent self-evaluation scores (i.e., instances where a model graded its own anonymized response) enabling detection of self-favoring bias.
- Off-diagonal variance: Variability among off-diagonal elements within a column indicates inter-grader disagreement for a given analysis, while variability within a row indicates inconsistency in a grader's standards across analyses.

Additionally, a sixth aggregate matrix ($\mathbf{M}_{\text{total}}$) was computed by summing the corresponding cells across all five domain matrices, yielding total scores (0–18 scale) for each grader–analysis pair.

2.4.5. Bias Minimization and Quality Assurance

Several measures were implemented to minimize bias and enhance the reliability of cross-model evaluations:

- Anonymization: All five analyses were presented to each grading LLM without identification of the originating model, labeled only as “Analysis 1” through “Analysis 5.”
- Self-Evaluation Inclusion: Each LLM graded its own response (presented anonymously) alongside the four other responses, enabling assessment of potential self-favoring bias through diagonal matrix analysis.

- **Standardized Quality Assurance Instructions:** Explicit instructions were embedded in the grading prompt to promote consistency and objectivity across all grading LLMs; the complete quality assurance guidelines are provided in Appendix C.1.

2.4.6. Data Analysis

Cross-model grading data were analyzed to assess:

1. **Inter-grader agreement:** The degree to which different LLMs assigned concordant scores to the same analyses, quantified using intraclass correlation coefficients (ICC) and percent agreement.
2. **Rank-order consistency:** Whether LLMs produced consistent rankings of the five analyses from best to worst.
3. **Self-evaluation bias:** Whether LLMs systematically assigned higher scores to their own (anonymized) responses compared to scores assigned by other models.
4. **Detection of critical errors:** Whether grading LLMs appropriately identified and penalized the fundamentally incorrect diagnosis provided by Grok 4 (intracerebral hemorrhage with incorrect lateralization).

This cross-evaluation methodology enabled a multi-dimensional assessment of LLM reliability, examining not only the capacity for accurate image interpretation but also the consistency and objectivity of quality assessment; both of which are essential competencies if such systems are to be integrated into clinical education or quality assurance workflows.

2.5. Reproducibility and Technical Parameters

Table 3 summarizes the technical parameters and access conditions for each model evaluated in this study. All diagnostic analyses and cross-evaluation grading sessions were conducted in January 2026. The complete analysis and grading prompts are reproduced verbatim in Appendix A.

Table 3. Reproducibility parameters for model evaluations. All models were accessed via their respective web interfaces using default inference settings.

Model	Version/Build	Access Method	Date Accessed
GPT-5	gpt-5-2024-12-01	Web UI (ChatGPT)	January 2026
Gemini 3 Pro	gemini-3-pro-001	Web UI (Gemini)	January 2026
Llama 4 Maverick	llama-4-maverick	Web UI (Meta AI)	January 2026
Grok 4	grok-4	Web UI (X/Grok)	January 2026
Claude Opus 4.5 Ext.	claude-opus-4.5-extended-2024-12-01	Web UI (Claude)	January 2026

Several methodological considerations regarding reproducibility warrant explicit acknowledgment. All evaluated models employ stochastic sampling during text generation. Even with identical prompts and images, different users or the same user at different times may receive different outputs. The temperature parameter, which controls output randomness, was not manually adjusted and remained at each platform's default setting throughout all evaluations.

Commercial LLM providers frequently update their models without public announcement or version increment. The model versions reported in Table 3 reflect the identifiers available at the time of access, but underlying model weights, safety filters, or system prompts may have changed between the evaluation dates and subsequent access attempts. The system prompts prepended by each platform to user queries are not publicly disclosed and may vary across time, user accounts, or geographic regions. The prompts documented in Appendix A represent only the user-visible portion of the full prompt context.

Each platform implements content moderation and safety filters that may influence responses to medical queries. These filters are not standardized across providers and may be updated without notice. Similarly, the preprocessing applied to uploaded images, including compression, resizing, and format conversion, is not disclosed by any platform and may affect model interpretation. All images in this study were uploaded in their original format without user-side preprocessing.

These factors collectively limit the strict reproducibility of the findings reported herein. Researchers attempting to replicate this study should anticipate variability in model outputs and may observe different specific responses while potentially observing similar patterns of inter-model variability and diagnostic inconsistency. This inherent unpredictability of LLM outputs is itself a finding relevant to their clinical deployment: systems that cannot guarantee reproducible outputs under identical conditions pose fundamental challenges for medical applications requiring consistency and auditability.

3. Results

The findings from both phases of the experimental protocol are presented below. First, we report the diagnostic performance of the five multimodal LLMs across the core evaluation domains: imaging modality identification, anatomical region localization, primary diagnosis, supporting radiological features, and differential diagnoses. This analysis reveals both areas of inter-model consensus and clinically significant variability, including one instance of fundamental diagnostic error. We then examine the patterns of variability among models that reached concordant general diagnoses, highlighting differences in acuity characterization, anatomical extent, pathophysiological terminology, and incidental finding detection that would carry meaningful implications for patient management. Subsequently, we present the results of the cross-evaluation protocol, wherein each LLM graded all five anonymized responses. The cross-evaluation findings are organized by domain-specific scoring matrices, inter-grader agreement metrics, rank-order consistency, self-evaluation bias analysis, detection and penalization of critical errors, and (most consequentially) the discovery of fundamental ground truth disagreement between grading models. Finally, we assess the presence or absence of safety disclaimers across model outputs, a finding with important implications for responsible AI deployment in medical contexts.

3.1. *Imaging Modality and Anatomical Region Identification*

All five LLMs correctly identified the imaging modality as computed tomography (CT) of the brain. Minor variations in terminology were observed, with some models specifying “non-contrast CT” (GPT-5, Claude Opus 4.5 Extended) while others used the general term “CT scan” (Gemini 3 Pro, Llama 4 Maverick, Grok 4). All models accurately identified the anatomical region as the brain at the level of the lateral ventricles and/or basal ganglia.

3.2. *Primary Diagnosis*

Substantial disagreement was observed in the primary diagnosis among the five LLMs (Table 4). Four models (GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Claude Opus 4.5 Extended) identified the primary finding as an acute or acute-to-subacute ischemic stroke involving the left middle cerebral artery (MCA) territory. In contrast, Grok 4 provided a fundamentally different diagnosis of intracerebral hemorrhage localized to the right cerebral hemisphere, which is a discrepancy affecting both the pathological process (ischemic vs. hemorrhagic) and the lateralization of findings (left vs. right hemisphere).

Table 4. Primary diagnoses provided by each large language model.

Model	Primary Diagnosis	Lateralization
GPT-5	Acute large territorial ischemic infarct, left MCA distribution	Left
Gemini 3 Pro	Large acute to subacute ischemic stroke, left MCA territory	Left
Llama 4 Maverick	Large left MCA infarct (ischemic stroke)	Left
Grok 4	Intracerebral hemorrhage	Right
Claude Opus 4.5 Extended	Acute-to-subacute ischemic stroke, left MCA territory	Left

3.3. Key Imaging Features

The models identifying ischemic stroke cited similar supporting features, including hypodensity/hypoattenuation in the affected region, loss of gray–white matter differentiation, mass effect with sulcal effacement, and midline shift. Claude Opus 4.5 Extended and GPT-5 additionally noted incidental bilateral calcifications (choroid plexus and basal ganglia, respectively). Grok 4, in contrast, described a “hyperdense lesion” with mass effect and surrounding edema, i.e., findings more consistent with acute hemorrhage rather than ischemic infarction and representing a fundamental misinterpretation of the image density characteristics.

3.4. Differential Diagnoses

Differential diagnoses varied considerably across models (Table 5). Common alternatives among the ischemia-diagnosing models included brain tumor (glioma or metastasis), encephalitis, and cerebral venous thrombosis. Grok 4’s differential diagnoses were consistent with its erroneous primary diagnosis, listing hemorrhagic stroke, brain tumor with intratumoral hemorrhage, and arteriovenous malformation with acute bleed.

Table 5. Differential diagnoses provided by each large language model.

Model	Differential Diagnoses
GPT-5	High-grade glioma/metastasis; Encephalitis (HSV); Cerebral venous thrombosis
Gemini 3 Pro	Primary brain tumor (low-grade glioma); Cerebral abscess; Encephalitis
Llama 4 Maverick	Hemorrhagic transformation of infarct; Trauma or infection-related edema
Grok 4	Hemorrhagic stroke; Brain tumor with hemorrhage; AVM with acute bleed
Claude Opus 4.5 Extended	Cerebral edema (trauma/metabolic); Low-grade glioma; Encephalitis

3.5. Summary of Inter-Model Variability

While 80% of models (4/5) reached consensus on the general diagnosis of left MCA territory ischemic stroke, one model (Grok 4) provided a fundamentally incorrect diagnosis, i.e., intracerebral hemorrhage lateralized to the wrong hemisphere. This represents a critical diagnostic error with significant clinical implications, as the management of ischemic stroke differs fundamentally from that of intracerebral hemorrhage. However, even among the four concordant models (GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Claude Opus 4.5 Extended), clinically meaningful variability was observed across multiple domains.

3.5.1. Variability in Acuity Characterization

The temporal characterization of the infarct differed among models. GPT-5 definitively classified the stroke as “acute,” while both Gemini 3 Pro and Claude Opus 4.5 Extended described it as “acute to subacute.” Llama 4 Maverick notably omitted any explicit acuity designation. This inconsistency is clinically consequential, as the determination of stroke acuity directly influences eligibility for time-sensitive interventions such as intravenous thrombolysis (typically within 4.5 h of symptom onset) and mechanical thrombectomy (up

to 24 h in select patients). A radiologist's assessment of acuity on imaging, in conjunction with clinical history, is integral to these therapeutic decisions.

3.5.2. Variability in Anatomical Extent

The models differed in their description of the anatomical structures involved. GPT-5 described involvement of "left cortical and subcortical regions," while Gemini 3 Pro and Llama 4 Maverick both specified involvement of the "frontal, temporal, and parietal lobes." In contrast, Claude Opus 4.5 Extended described only the "left frontal-parietal region," omitting reference to temporal lobe involvement. Accurate delineation of the affected territory has prognostic implications, as involvement of specific regions (e.g., language areas, motor cortex, insular cortex) correlates with distinct functional deficits and rehabilitation trajectories.

3.5.3. Variability in Pathophysiological Terminology

Subtle but potentially significant differences emerged in the pathophysiological language employed. Claude Opus 4.5 Extended specifically attributed the hypodensity to "cytotoxic edema" (i.e., the hallmark of acute ischemic injury) whereas other models used nonspecific descriptors such as "hypoattenuation" (GPT-5) or "hypodensity" (Gemini 3 Pro, Llama 4 Maverick) without mechanistic attribution. While these terms describe the same imaging finding, the explicit identification of cytotoxic edema demonstrates a deeper inference about the underlying pathophysiology, which may reflect differences in model training or reasoning depth.

3.5.4. Variability in Identification of Incidental Findings

Only two models reported incidental findings. GPT-5 noted "punctate bilateral basal ganglia calcifications," while Claude Opus 4.5 Extended identified "bilateral choroid plexus calcifications." Gemini 3 Pro and Llama 4 Maverick made no mention of incidental findings. Although these calcifications are typically benign, the inconsistent reporting raises questions about the completeness of image analysis performed by each model and whether clinically relevant incidental pathology might be similarly overlooked.

3.5.5. Variability in Differential Diagnoses

Substantial heterogeneity was observed in the differential diagnoses proposed by the four concordant models (Table 5). While all four included some form of neoplastic or infectious/inflammatory process, the specific entities varied considerably:

- Tumor characterization: GPT-5 suggested "high-grade glioma or metastasis," whereas Gemini 3 Pro and Claude Opus 4.5 Extended proposed "low-grade glioma." The distinction between high-grade and low-grade neoplasms carries markedly different prognostic and therapeutic implications.
- Infectious etiologies: GPT-5 specifically mentioned herpes simplex virus (HSV) encephalitis, Gemini 3 Pro uniquely included cerebral abscess, while Claude Opus 4.5 Extended and Llama 4 Maverick used the nonspecific term "encephalitis" or "infection."
- Unique differential considerations: GPT-5 was the only model to suggest cerebral venous thrombosis with venous infarction. Llama 4 Maverick uniquely raised the possibility of hemorrhagic transformation of the infarct and provided only two differential diagnoses rather than the requested three.

These discrepancies in differential diagnosis generation could lead to divergent downstream clinical workups. For example, consideration of HSV encephalitis would prompt empiric acyclovir therapy and lumbar puncture, while suspicion for cerebral abscess would necessitate neurosurgical consultation and prolonged antibiotic therapy.

3.5.6. Clinical Implications of Inter-Model Variability

The observed variability among the four models that correctly identified ischemic stroke (encompassing differences in acuity assessment, anatomical localization, pathophysiological interpretation, incidental finding detection, and differential diagnosis formulation) demonstrates that even when LLMs converge on a general diagnosis, the clinical nuance embedded in their responses varies substantially. Such inconsistencies could lead to different clinical decisions if these outputs were used to guide patient care without expert radiologist oversight.

3.6. Cross-Evaluation Results

The cross-evaluation protocol yielded a rich dataset enabling multi-dimensional assessment of LLM evaluative behavior. Each of the five models graded all five anonymized analyses across the five rubric domains, generating 125 individual domain scores that were aggregated into domain-specific and total score matrices. The findings presented below reveal patterns of inter-grader agreement and disagreement, systematic differences in grading stringency, evidence of self-evaluation bias, and (most consequentially) fundamental divergence in the interpretation of radiological ground truth between grading models. We first present the aggregate total score matrix to establish overall patterns, then examine domain-specific scoring to identify where agreement was strongest and where disagreement was most pronounced. Subsequent sections analyze inter-grader reliability metrics, rank-order consistency, self-evaluation bias, detection and penalization of critical diagnostic errors, and the implications of the ground truth disagreement revealed by Gemini 3 Pro's anomalous grading behavior.

3.6.1. Total Score Matrix

Table 6 presents the aggregate cross-evaluation scoring matrix ($\mathbf{M}_{\text{total}}$), with total scores (0–18 scale) for each grader–analysis pair. Figure 4 provides a heatmap visualization of these scores normalized to percentages, enabling rapid identification of scoring patterns through color-coded intensity mapping.

Table 6. Cross-evaluation total score matrix ($\mathbf{M}_{\text{total}}$). Rows represent grading LLMs; columns represent graded analyses (Analysis 1 = GPT-5, Analysis 2 = Gemini 3 Pro, Analysis 3 = Llama 4 Maverick, Analysis 4 = Grok 4, Analysis 5 = Claude Opus 4.5 Extended). Diagonal elements represent self-evaluations.

	A1	A2	A3	A4	A5	Row Mean
GPT-5	18	17	15	5	18	14.6
Gemini 3 Pro	13	13	12	6	13	11.4
Llama 4 Maverick	18	17	17	9	18	15.8
Grok 4	18	16	15	8	17	14.8
Claude Opus 4.5 Ext.	18	15	13	4	18	13.6
Column Mean	17.0	15.6	14.4	6.4	16.8	

Column analysis reveals strong consensus that Analysis 1 (GPT-5) and Analysis 5 (Claude Opus 4.5 Extended) represented the highest-quality interpretations (mean scores 17.0 and 16.8, respectively), while Analysis 4 (Grok 4) was consistently rated lowest (mean 6.4). As visually evident in Figure 4, the A4 column exhibits uniformly low percentage scores (red/orange shading, 22–50%) across all graders, reflecting unanimous recognition of its fundamental diagnostic error, while the A1 and A5 columns display consistently high scores (green shading, 72–100%). The intermediate coloring of Gemini 3 Pro's row (Row 2)

illustrates its systematically lower scoring pattern attributable to its divergent ground truth interpretation, with no score exceeding 72%.

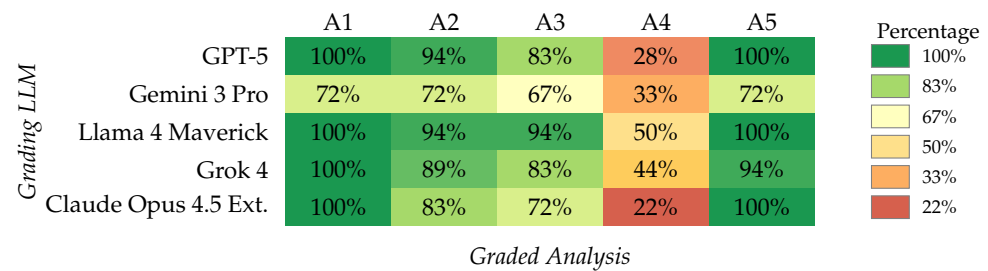


Figure 4. Heatmap visualization of the cross-evaluation total score matrix (M_{total}) expressed as percentages of maximum possible score (18 points). Cell color intensity reflects percentage magnitude (green $\geq 80\%$, yellow 50–79%, red $< 50\%$). Analysis 4 (Grok 4) received consistently low percentage scores (22–50%) across all graders due to its fundamental diagnostic error.

3.6.2. Domain-Specific Score Matrices

Beyond the aggregate total scores, analysis of domain-specific scoring matrices reveals important patterns in inter-grader agreement and disagreement. The five domain matrices (M_{modality} , M_{anatomy} , $M_{\text{diagnosis}}$, M_{features} , and $M_{\text{differential}}$) are presented below with accompanying analysis.

Imaging Modality (M_{modality})

Table 7 presents the cross-evaluation scores for imaging modality identification. This domain demonstrated the highest inter-grader agreement among all five evaluation categories, with column means ranging from 2.6 to 3.0 on a 0–3 scale.

Table 7. Cross-evaluation imaging modality score matrix (M_{modality}). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–3. Diagonal elements represent self-evaluations.

	A1	A2	A3	A4	A5	Row Mean
GPT-5	3	3	3	3	3	3.0
Gemini 3 Pro	3	3	3	3	3	3.0
Llama 4 Maverick	3	3	3	3	3	3.0
Grok 4	3	3	3	2	3	2.8
Claude Opus 4.5 Ext.	3	2	2	2	3	2.4
Column Mean	3.0	2.8	2.8	2.6	3.0	

High agreement was observed, with most graders assigning scores of 2–3 across analyses. Three graders (GPT-5, Gemini 3 Pro, and Llama 4 Maverick) awarded full marks (3/3) to all five analyses, reflecting consensus that all models correctly identified the imaging modality as computed tomography. The only notable deviations were observed in Claude Opus 4.5 Extended, which assigned scores of 2 (rather than 3) to Analyses 2, 3, and 4 for lacking specificity regarding contrast status and acquisition plane, and Grok 4, which similarly assigned 2/3 to Analysis 4 (its own response) for incomplete modality characterization. These deviations reflect a more stringent interpretation of the rubric criterion requiring “full correctness and specificity.”

Column analysis reveals that Analyses 1 (GPT-5) and 5 (Claude Opus 4.5 Extended) achieved perfect mean scores of 3.0, indicating unanimous recognition of their comprehensive modality descriptions (“non-contrast axial head CT”). Analysis 4 (Grok 4) received the lowest column mean (2.6), penalized by two graders for describing only “CT” without specifying contrast status or imaging plane.

Row analysis demonstrates minimal variability in grading stringency for this domain. GPT-5, Gemini 3 Pro, and Llama 4 Maverick each assigned a row mean of 3.0, while Grok 4 (2.8) and Claude Opus 4.5 Extended (2.4) applied slightly more rigorous standards. This pattern suggests that Claude Opus 4.5 Extended employed the most exacting criteria for modality specification.

Anatomical Region/Organ (M_{anatomy})

Table 8 presents the cross-evaluation scores for anatomical region identification. This domain showed moderate inter-grader agreement, with column means ranging from 2.4 to 3.0 on a 0–3 scale.

Agreement was uniformly high for Analyses 1, 2, and 5, which all achieved perfect column means of 3.0. All five graders recognized that these analyses provided specific and accurate anatomical localization, identifying the brain at the level of the lateral ventricles and/or basal ganglia with appropriate hemispheric specification.

In contrast, Analyses 3 and 4 received lower scores (column means of 2.4), with three graders (GPT-5, Grok 4, and Claude Opus 4.5 Extended) assigning 2/3 for insufficient anatomical specificity. Analysis 3 (Llama 4 Maverick) was penalized for describing only “brain, transverse section through the cerebral hemispheres” without specifying the slice level or ventricular structures. Analysis 4 (Grok 4) similarly received deductions for the vague descriptor “brain (cerebral hemispheres)” without further anatomical detail.

Row analysis reveals divergent grading stringency. Gemini 3 Pro and Llama 4 Maverick demonstrated lenient grading patterns, assigning 3/3 to all five analyses (row means of 3.0). In contrast, GPT-5, Grok 4, and Claude Opus 4.5 Extended applied stricter criteria (row means of 2.6), penalizing analyses that lacked specificity regarding slice level and lateralization.

Notably, the diagonal elements reveal that both Llama 4 Maverick and Grok 4 self-evaluated their anatomical descriptions more favorably than other graders assessed them. Llama 4 Maverick assigned itself 3/3 while receiving a mean of 2.33 from other graders; Grok 4 assigned itself 2/3, matching the consensus assessment.

Table 8. Cross-evaluation anatomical region score matrix (M_{anatomy}). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–3. Diagonal elements represent self-evaluations.

	A1	A2	A3	A4	A5	Row Mean
GPT-5	3	3	2	2	3	2.6
Gemini 3 Pro	3	3	3	3	3	3.0
Llama 4 Maverick	3	3	3	3	3	3.0
Grok 4	3	3	2	2	3	2.6
Claude Opus 4.5 Ext.	3	3	2	2	3	2.6
Column Mean	3.0	3.0	2.4	2.4	3.0	

Most Likely Diagnosis ($M_{\text{diagnosis}}$)

Table 9 presents the cross-evaluation scores for diagnostic accuracy. This domain showed the greatest inter-grader variance among all five evaluation categories, driven primarily by fundamental disagreement regarding the radiological ground truth. Figure 5 provides a percentage-normalized heatmap visualization of this matrix, rendering the scoring patterns immediately apparent.

Table 9. Cross-evaluation most likely diagnosis score matrix ($M_{\text{diagnosis}}$). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–4. Diagonal elements represent self-evaluations. Note: Gemini 3 Pro assigned 2/4 to Analyses 1, 2, 3, and 5 based on its divergent interpretation that the image depicted chronic rather than acute infarction.

	A1	A2	A3	A4	A5	Row Mean
GPT-5	4	4	4	0	4	3.2
Gemini 3 Pro	2	2	2	0	2	1.6
Llama 4 Maverick	4	4	4	0	4	3.2
Grok 4	4	4	4	0	4	3.2
Claude Opus 4.5 Ext.	4	4	4	0	4	3.2
Column Mean	3.6	3.6	3.6	0.0	3.6	

Four graders (GPT-5, Llama 4 Maverick, Grok 4, and Claude Opus 4.5 Extended) assigned full marks (4/4) to Analyses 1, 2, 3, and 5 for correctly identifying left MCA ischemic infarction. These graders interpreted the image as demonstrating an acute or acute-to-subacute infarct with cytotoxic edema and mass effect, i.e., findings consistent with the majority diagnostic interpretation.

In contrast, Gemini 3 Pro assigned only 2/4 to all four of these analyses, asserting that the image depicted a chronic infarct with encephalomalacia and ex-vacuo ventricular dilatation rather than an acute/subacute infarct with mass effect. Gemini’s ground truth interpretation described “a large, well-defined area of marked hypodensity (CSF density)” with “ex-vacuo dilatation of the ipsilateral (left) lateral ventricle and prominence of the overlying sulci”, which are features indicative of chronic tissue loss rather than acute swelling. This represents a fundamental disagreement about the radiological ground truth that propagated systematically through all of Gemini’s evaluations.

All five graders unanimously assigned 0/4 to Analysis 4 (Grok 4) for its incorrect diagnosis of intracerebral hemorrhage. This consensus demonstrates that despite disagreements about infarct chronicity, all graders recognized the categorical error of misidentifying hypodensity (dark) as hyperdensity (bright) and the consequent misdiagnosis of hemorrhage versus infarction.

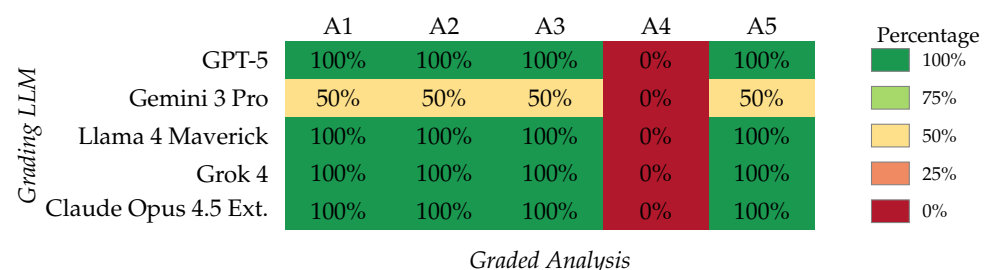


Figure 5. Heatmap visualization of the cross-evaluation diagnosis score matrix ($M_{\text{diagnosis}}$) expressed as percentages of maximum possible score (4 points). Cell color intensity reflects percentage magnitude (green = 100%, yellow = 50%, red = 0%). Analysis 4 (Grok 4) received 0% from all graders for its fundamental misdiagnosis of hemorrhage. Gemini 3 Pro’s row (Row 2) shows uniform 50% scores for Analyses 1, 2, 3, and 5, reflecting its divergent interpretation that the image depicted chronic rather than acute infarction.

Column analysis reveals a stark bimodal distribution, visually striking in Figure 5: the A4 column is uniformly red (0%) across all graders, while columns A1, A2, A3, and A5 display predominantly green shading (100%) with the exception of Gemini 3 Pro’s row. Analyses 1, 2, 3, and 5 all achieved column means of 3.6 (reduced from 4.0 solely by Gemini’s 2/4 scores), while Analysis 4 received a column mean of 0.0, reflecting universal recognition of its diagnostic error.

Row analysis demonstrates that Gemini 3 Pro’s divergent ground truth interpretation resulted in the lowest row mean (1.6), compared to uniform row means of 3.2 for the other four graders. As evident in Figure 5, Gemini’s entire row displays yellow shading (50%) for Analyses 1, 2, 3, and 5, i.e., a distinctive horizontal band reflecting its systematic penalization of all analyses describing acute findings. This 1.6-point differential in row means (equivalent to a full letter grade in many assessment systems) illustrates the profound impact of ground truth disagreement on evaluation outcomes.

Key Image Features (M_{features})

Table 10 presents the cross-evaluation scores for the description of key image features. This domain showed substantial inter-grader variance, again driven primarily by Gemini 3 Pro’s divergent ground truth interpretation, with additional variability arising from differences in completeness standards. Figure 6 provides a percentage-normalized heatmap visualization of this matrix, highlighting both the systematic patterns and the increased variability compared to the diagnosis domain.

Table 10. Cross-evaluation key image features score matrix (M_{features}). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–4. Diagonal elements represent self-evaluations. Note: Gemini 3 Pro assigned 2/4 to Analyses 1, 2, 3, and 5 for describing “mass effect” and “sulcal effacement,” which Gemini considered inconsistent with its interpretation of chronic atrophy.

	A1	A2	A3	A4	A5	Row Mean
GPT-5	4	4	4	0	4	3.2
Gemini 3 Pro	2	2	2	0	2	1.6
Llama 4 Maverick	4	4	4	0	4	3.2
Grok 4	4	3	4	1	4	3.2
Claude Opus 4.5 Ext.	4	3	3	0	4	2.8
Column Mean	3.6	3.2	3.4	0.2	3.6	

Gemini 3 Pro’s divergent interpretation propagated directly to feature scoring. Gemini assigned 2/4 to Analyses 1, 2, 3, and 5 for describing “mass effect” and “sulcal effacement,” i.e., features that Gemini considered hallucinated findings inconsistent with the chronic atrophy it perceived. In Gemini’s assessment, the correct features should have included “ex-vacuo dilatation” and “sulcal prominence” rather than effacement. This systematic penalization reflects not a difference in grading stringency but a fundamental disagreement about what the image actually demonstrates. As visible in Figure 6, Gemini’s entire row (Row 2) displays uniform yellow shading (50%) for Analyses 1, 2, 3, and 5, i.e., the same pattern observed in the diagnosis domain.

Among the four graders who interpreted the image as acute/subacute infarction, minor variability was observed in feature scoring. GPT-5 and Llama 4 Maverick assigned uniform 4/4 scores to Analyses 1, 2, 3, and 5, indicating satisfaction with the completeness and accuracy of feature descriptions. Claude Opus 4.5 Extended applied stricter criteria, assigning 3/4 to Analyses 2 and 3 for incomplete feature descriptions (omitting sulcal effacement or incidental calcifications). Grok 4 similarly assigned 3/4 to Analysis 2 for not mentioning incidental findings. This variability is evident in Figure 6, where light green cells (75%) appear in the A2 and A3 columns for Claude and Grok, contrasting with the uniform green (100%) assigned by GPT-5 and Llama.

Notably, Grok 4 was the only grader to award any credit (1/4) to Analysis 4’s feature description. Grok’s reasoning acknowledged that some features (mass effect, midline shift) were partially described correctly in directional terms, even though the fundamental characterization of lesion density was wrong. This lenient approach contrasts with the zero

scores assigned by all other graders, who considered the misidentification of hypodensity as hyperdensity to be a disqualifying error. This unique partial credit is visible as the single orange cell (25%) in the otherwise red A4 column of Figure 6.

Column analysis reveals that Analyses 1 and 5 achieved the highest mean scores (3.6), followed by Analysis 3 (3.4), Analysis 2 (3.2), and Analysis 4 (0.2). The near-zero score for Analysis 4 reflects near-unanimous agreement that its feature description was fundamentally incorrect.

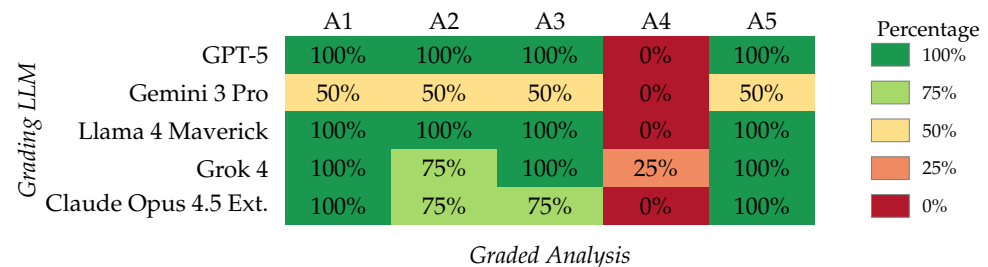


Figure 6. Heatmap visualization of the cross-evaluation key image features score matrix (M_{features}) expressed as percentages of maximum possible score (4 points). Cell color intensity reflects percentage magnitude (green = 100%, light green = 75%, yellow = 50%, orange = 25%, red = 0%). Analysis 4 (Grok 4) received near-zero scores from four graders for its fundamental mischaracterization of lesion density. Gemini 3 Pro’s row (Row 2) displays uniform 50% scores for Analyses 1, 2, 3, and 5, reflecting its penalization of descriptions of “mass effect” and “sulcal effacement” as inconsistent with its interpretation of chronic atrophy. Grok 4 uniquely awarded 25% to Analysis 4 for partially correct directional descriptions of mass effect.

Differential Diagnoses ($M_{\text{differential}}$)

Table 11 presents the cross-evaluation scores for differential diagnoses. This domain revealed the widest variability in grading philosophy, with fundamental disagreement about whether differentials should be evaluated for internal consistency with the stated diagnosis or for relevance to the actual imaging findings. Figure 7 provides a percentage-normalized heatmap visualization of this matrix, revealing patterns of variability distinct from the previous domains.

Table 11. Cross-evaluation differential diagnoses score matrix ($M_{\text{differential}}$). Rows represent grading LLMs; columns represent graded analyses. Scale: 0–4. Diagonal elements represent self-evaluations. Note: Llama 4 Maverick and Grok 4 awarded partial credit (3/4) to Analysis 4’s differentials despite their irrelevance to the actual imaging findings, reflecting divergent evaluation philosophies.

	A1	A2	A3	A4	A5	Row Mean
GPT-5	4	3	2	0	4	2.6
Gemini 3 Pro	3	3	2	0	3	2.2
Llama 4 Maverick	4	3	3	3	4	3.4
Grok 4	4	3	2	3	3	3.0
Claude Opus 4.5 Ext.	4	3	2	0	4	2.6
Column Mean	3.8	3.0	2.2	1.2	3.6	

Column analysis demonstrates a clear quality hierarchy. Analysis 1 (GPT-5) received the highest mean score (3.8), with four graders awarding 4/4 for providing three plausible, prioritized, and justified differentials (high-grade glioma/metastasis, HSV encephalitis, cerebral venous thrombosis). Only Gemini 3 Pro assigned a lower score (3/4), noting that the differentials were biased toward acute pathology consistent with Gemini’s interpretation that the image showed chronic changes.

Analysis 5 (Claude Opus 4.5 Extended) achieved the second-highest mean score (3.6), with similar recognition for well-reasoned differentials. Analysis 2 (Gemini 3 Pro) received

a mean of 3.0, with deductions primarily for lacking prioritization or justification. Analysis 3 (Llama 4 Maverick) received the lowest mean among the correctly-diagnosing analyses (2.2), penalized by multiple graders for providing only two differential categories and including the vague descriptor “other causes of edema/encephalomalacia.” This pattern is visible in Figure 7, where the A3 column displays uniform yellow shading (50–75%) rather than the predominantly green shading seen in A1 and A5.

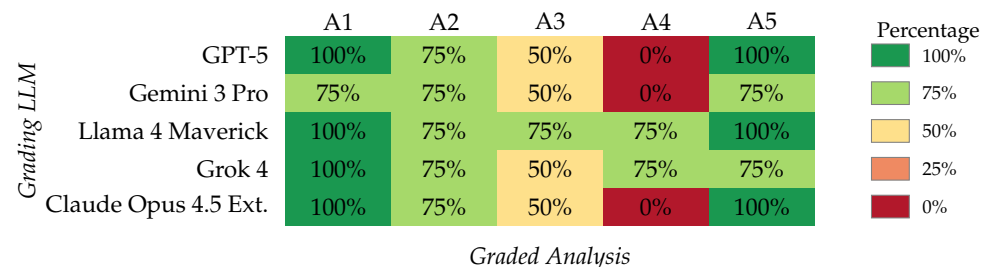


Figure 7. Heatmap visualization of the cross-evaluation differential diagnoses score matrix ($M_{\text{differential}}$) expressed as percentages of maximum possible score (4 points). Cell color intensity reflects percentage magnitude (green = 100%, light green = 75%, yellow = 50%, red = 0%). Column A4 reveals divergent evaluation philosophies: GPT-5, Gemini 3 Pro, and Claude Opus 4.5 Extended assigned 0% (red) for differentials predicated on an incorrect diagnosis, while Llama 4 Maverick and Grok 4 awarded 75% (light green) for internally consistent but externally irrelevant differentials. Column A3 (Llama 4 Maverick’s response) shows uniform yellow/light green shading (50–75%), reflecting penalization for incomplete differential lists.

The most striking finding in this domain was the divergent treatment of Analysis 4’s differentials. Analysis 4 (Grok 4) listed hemorrhagic stroke, brain tumor with intratumoral hemorrhage, and arteriovenous malformation with acute bleed, i.e., differentials that were internally consistent with its (incorrect) primary diagnosis of intracerebral hemorrhage but entirely irrelevant to the actual imaging findings of ischemic infarction.

Three graders (GPT-5, Gemini 3 Pro, and Claude Opus 4.5 Extended) assigned 0/4, reasoning that differentials predicated on a fundamentally incorrect primary diagnosis have no clinical utility and should receive no credit. These graders applied an “actual findings” standard: differentials are only meaningful if they represent plausible alternatives to the correct diagnosis.

In contrast, Llama 4 Maverick and Grok 4 awarded 3/4, evaluating the differentials for internal consistency with the analysis’s stated diagnosis rather than for relevance to the true imaging findings. Grok 4’s self-evaluation explicitly acknowledged that “3 plausible options for hemorrhage” were provided, awarding credit despite recognizing that the primary diagnosis was incorrect. This “internal consistency” standard represents a fundamentally different evaluation philosophy.

This philosophical divergence, whether to assess differentials relative to radiological ground truth or relative to the analysis’s internal logic, represents a significant inconsistency in evaluation standards with important implications for the use of LLMs in educational assessment or quality assurance contexts. The divergence is visualized in Figure 7: the A4 column displays a distinctive split pattern, with three red cells (0%) from GPT-5, Gemini 3 Pro, and Claude Opus 4.5 Extended, and two light green cells (75%) from Llama 4 Maverick and Grok 4. Unlike the diagnosis and features domains (where A4 was uniformly red) the differential diagnosis domain reveals the clearest visual evidence of divergent evaluation philosophies among graders.

Row analysis reveals corresponding differences in grading stringency for this domain. Llama 4 Maverick demonstrated the most lenient pattern (row mean 3.4), followed by Grok 4 (3.0), GPT-5 and Claude Opus 4.5 Extended (2.6 each), and Gemini 3 Pro (2.2). The 1.2-point spread in row means indicates substantial inter-grader variability in differen-

tial diagnosis evaluation standards. As evident in Figure 7, Row 3 (Llama 4 Maverick) is the only row containing no red cells, reflecting its consistently lenient approach (including the controversial 75% awarded to Analysis 4).

Summary of Domain-Specific Patterns

Table 12 consolidates the column means across all five domain-specific scoring matrices, providing a comprehensive overview of analysis quality by domain. Figure 8 visualizes these scores as a stacked bar chart, illustrating how each domain contributes to the total score for each analysis.

Table 12. Summary of column means across all five domain-specific scoring matrices. Values represent mean scores assigned by all five grading LLMs to each analysis.

Domain (Scale)	A1	A2	A3	A4	A5
Imaging Modality (0–3)	3.0	2.8	2.8	2.6	3.0
Anatomical Region (0–3)	3.0	3.0	2.4	2.4	3.0
Most Likely Diagnosis (0–4)	3.6	3.6	3.6	0.0	3.6
Key Image Features (0–4)	3.6	3.2	3.4	0.2	3.6
Differential Diagnoses (0–4)	3.8	3.0	2.2	1.2	3.6
Total (0–18)	17.0	15.6	14.4	6.4	16.8

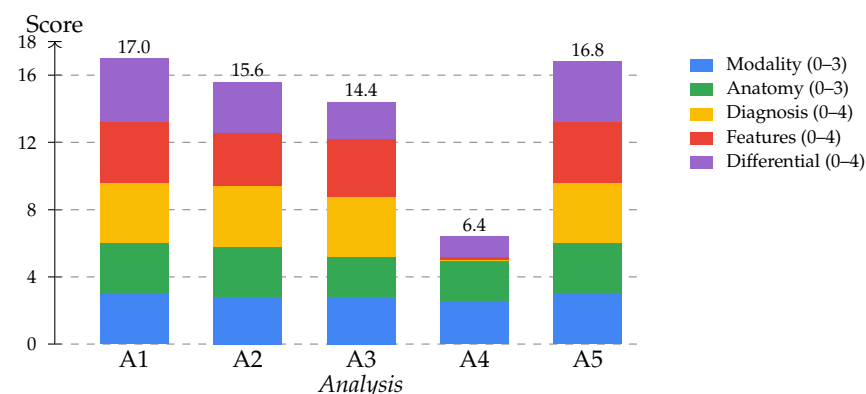


Figure 8. Stacked bar chart showing domain-specific contributions to total cross-evaluation scores. Analysis 4 (Grok 4) demonstrates dramatically reduced scores in diagnosis, features, and differential domains due to its fundamental diagnostic error, while maintaining comparable scores for modality and anatomical identification.

Several patterns emerge from this summary:

1. Modality and anatomy showed highest agreement: These foundational domains demonstrated the least inter-grader variability, with column means ranging from 2.4 to 3.0 (on 0–3 scales). This suggests that LLMs can reliably evaluate basic image characterization tasks.
2. Diagnosis and features showed greatest variability: These clinically critical domains were most affected by Gemini 3 Pro’s divergent ground truth interpretation, with systematic 2-point deductions (on 0–4 scales) for analyses describing acute findings.
3. Differential diagnosis revealed philosophical divergence: The 1.2-point column mean for Analysis 4 (vs. 0.0 for diagnosis) reflects the split between graders who applied “internal consistency” versus “actual findings” evaluation standards.
4. Analysis 4 was penalized across all domains: Even in modality and anatomy (where Analysis 4 was not fundamentally incorrect), it received the lowest or tied-lowest column means, suggesting a halo effect wherein initial diagnostic errors may have influenced subsequent domain evaluations. As strikingly illustrated in Figure 8,

Analysis 4's total score (6.4) is dominated by the blue and green segments (modality and anatomy), with near-absent contributions from diagnosis (yellow), features (red), and differential (purple); in contrast to the fuller stacked profiles of Analyses 1 and 5.

3.6.3. Inter-Grader Agreement

Substantial variability in grading stringency was observed across models (Table 6, Row Mean column). Llama 4 Maverick demonstrated the most lenient grading pattern (mean score assigned: 15.8), while Gemini 3 Pro was the most stringent (mean: 11.4). The remaining three graders occupied intermediate positions: Grok 4 (14.8), GPT-5 (14.6), and Claude Opus 4.5 Extended (13.6). This 4.4-point spread in row means (equivalent to approximately 24% of the total scale) indicates that absolute scores from different LLM graders are not directly comparable without calibration.

Excluding Analysis 4 (which contained critical errors), inter-grader agreement for the remaining four analyses was moderate. The coefficient of variation (CV) for total scores ranged from 6.2% (Analysis 1) to 14.8% (Analysis 3), indicating reasonable but imperfect consensus on analysis quality. Analysis 1 achieved the tightest agreement, with all five graders assigning scores between 13 and 18. Analysis 3 showed the widest dispersion, receiving scores ranging from 12 (Gemini 3 Pro) to 17 (Llama 4 Maverick and Grok 4).

The most significant source of disagreement stemmed from Gemini 3 Pro's alternative ground truth interpretation. When Gemini's scores are excluded, agreement among the remaining four graders increases substantially, with total scores for Analyses 1, 2, 3, and 5 showing coefficients of variation below 8%. This finding suggests that the primary driver of inter-grader disagreement was not differences in evaluative stringency or rubric interpretation, but rather a fundamental disagreement about what the image actually demonstrated.

3.6.4. Rank-Order Consistency

Despite differences in absolute scoring, four of five graders produced identical rank orderings: Analyses 1 and 5 tied for first place, followed by Analysis 2 in third, Analysis 3 in fourth, and Analysis 4 in last place (Table 13). This consensus ranking emerged despite substantial differences in the magnitude of scores assigned.

Table 13. Rank orderings assigned by each grading LLM (1 = highest quality). Ties are indicated in parentheses.

	A1	A2	A3	A4	A5
GPT-5	1 (tie)	3	4	5	1 (tie)
Gemini 3 Pro	1 (tie)	1 (tie)	4	5	1 (tie)
Llama 4 Maverick	1 (tie)	3 (tie)	3 (tie)	5	1 (tie)
Grok 4	1	3	4	5	2
Claude Opus 4.5 Ext.	1 (tie)	3	4	5	1 (tie)

Grok 4 produced a nearly identical ranking, with the sole deviation being the placement of Analysis 5 in second place rather than tied for first. This minor discrepancy arose from Grok 4 assigning Analysis 5 a score of 17/18 (deducting one point for differential diagnoses lacking explicit prioritization) while awarding Analysis 1 a perfect 18/18.

Gemini 3 Pro's ranking deviated more substantially from this consensus. Due to its divergent ground truth interpretation, Gemini assigned identical scores of 13/18 to Analyses 1, 2, and 5, creating a three-way tie for first place. Analysis 3 received 12/18 (fourth place), and Analysis 4 received 6/18 (last place). This ranking compression resulted from Gemini's uniform penalization of all analyses that described acute findings; since Gemini considered these descriptions to be hallucinated, the distinctions among them became compressed.

The preservation of rank-order consistency despite absolute score variability has practical implications. If LLMs were deployed for comparative assessment (e.g., ranking trainee responses or identifying outliers), the results might be more reliable than if used for absolute scoring (e.g., determining whether a response meets a fixed threshold). However, the Gemini 3 Pro anomaly demonstrates that rank-order consistency can break down when graders disagree about fundamental ground truth.

3.6.5. Self-Evaluation Bias Analysis

Diagonal analysis of M_{total} (Table 6) enabled assessment of self-favoring bias. Self-evaluation scores (diagonal elements) were compared against the mean scores assigned by other graders to the same analysis (Table 14). A positive Δ indicates that a model assigned its own analysis a higher score than the mean assessment by other graders; a negative Δ indicates the opposite. Figure 9 visualizes these bias values as a diverging bar chart, with positive deviations (green) extending rightward and negative deviations (red) extending leftward from the zero line.

Table 14. Self-evaluation bias analysis. Positive Δ indicates self-favoring bias; negative Δ indicates self-penalizing bias.

Model	Self-Score	Others' Mean	Δ
GPT-5	18	16.75	+1.25
Gemini 3 Pro	13	16.25	−3.25
Llama 4 Maverick	17	13.75	+3.25
Grok 4	8	6.0	+2.0
Claude Opus 4.5 Ext.	18	16.5	+1.5

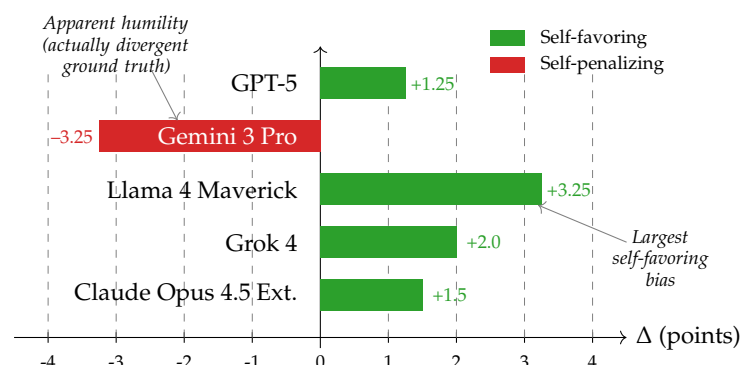


Figure 9. Diverging bar chart showing self-evaluation bias ($\Delta = \text{Self-Score} - \text{Others' Mean}$). Positive values (green) indicate self-favoring bias; negative values (red) indicate apparent self-penalizing bias. Llama 4 Maverick demonstrated the largest self-favoring deviation (+3.25 points), while Gemini 3 Pro's negative bias (−3.25 points) reflects consistent application of its divergent ground truth interpretation rather than genuine self-criticism.

Three distinct patterns emerged from this analysis:

1. **Modest self-favoring bias:** GPT-5, Grok 4, and Claude Opus 4.5 Extended assigned themselves scores 1.25–2.0 points higher than the mean of other graders. This modest positive bias suggests mild self-favoring tendencies but falls within the range of normal inter-grader variability observed for other analyses. Notably, even Grok 4 (whose analysis contained fundamental diagnostic errors) assigned itself a higher score than other graders, suggesting that self-evaluation bias persists even when objective performance is poor.
2. **Pronounced self-favoring bias:** Llama 4 Maverick demonstrated the largest self-favoring deviation, assigning its own analysis a score of 17 compared to a mean of

13.75 from other graders ($\Delta = +3.25$). As illustrated by the longest green bar in Figure 9, this 3.25-point discrepancy exceeds one standard deviation of the inter-grader score distribution for Analysis 3, suggesting a systematic tendency toward self-preferential evaluation. Examination of domain-specific scores reveals that Llama 4 Maverick awarded itself full marks (3/3) for anatomical description and (3/4) for differentials, whereas other graders assigned 2.33/3 and 2.0/4 on average for these domains.

3. Apparent negative self-bias: Gemini 3 Pro uniquely assigned its own analysis a lower score (13) than the mean assigned by other graders (16.25), yielding a Δ of -3.25 . This anomaly is immediately apparent in Figure 9 as the sole red bar extending leftward, i.e., visually distinguishing Gemini from the four green bars representing self-favoring models. Superficially, this pattern might suggest humility or conservative self-assessment. However, examination of Gemini's reasoning reveals that this apparent self-penalization reflects its divergent ground truth interpretation rather than genuine self-criticism. Gemini penalized its own analysis for describing "acute" findings and "mass effect," i.e., the same features it penalized in all other analyses. Thus, Gemini's negative Δ does not indicate superior self-awareness but rather consistent application of an internally coherent (though externally discordant) reference standard.

The presence of self-evaluation bias, even in anonymized conditions, has implications for proposed uses of LLMs in automated peer review or self-assessment contexts. If models systematically favor their own outputs, ensemble approaches that include self-evaluation could produce inflated quality estimates.

3.6.6. Detection of Critical Errors

A key test of LLM grading reliability was whether graders appropriately identified and penalized the critical diagnostic errors in Analysis 4, which misidentified hypodensity as hyperdensity and consequently diagnosed intracerebral hemorrhage with incorrect lateralization (right instead of left hemisphere).

All five grading LLMs assigned Analysis 4 the lowest total score, confirming unanimous detection of the critical error. This finding demonstrates that current LLMs possess sufficient radiological knowledge to recognize fundamental diagnostic mistakes, at least in cases of clear-cut pathology misidentification. However, the magnitude of penalization varied considerably across graders:

- Claude Opus 4.5 Extended applied the most severe penalization, assigning 4/18. This score reflected 0/4 for diagnosis, 0/4 for features, and 0/4 for differentials, with full credit awarded only for modality (2/3) and anatomy (2/3). The accompanying reasoning explicitly identified the fundamental misinterpretation: "The analysis describes a 'hyperdense lesion in the right cerebral hemisphere' with 'midline shift to the left.' This is entirely wrong: the lesion is hypodense (not hyperdense), located in the left hemisphere (not right), and the midline shifts to the right (not left)."
- GPT-5 assigned 5/18, with an identical scoring pattern: zero for diagnosis, features, and differentials, and partial credit for modality and anatomy.
- Gemini 3 Pro assigned 6/18, also awarding zero for the three core clinical domains but providing full marks (3/3) for both modality and anatomy identification.
- Grok 4 (self-evaluating) assigned 8/18 to its own analysis. While acknowledging the incorrect diagnosis, Grok awarded partial credit: 1/4 for features (noting that "mass effect and midline shift are partially correct" in directional terms) and 3/4 for differentials (stating "3 plausible options for hemorrhage" were provided). This lenient self-assessment reflects an "internal consistency" evaluation philosophy wherein

differentials are judged relative to the stated diagnosis rather than the actual imaging findings.

- Llama 4 Maverick was most lenient, assigning 9/18. This included full marks for modality (3/3) and anatomy (3/3), zero for diagnosis and features, but 3/4 for differentials; i.e., the same rationale applied by Grok 4.

The variation in Analysis 4 scores (range: 4–9) reveals inconsistent standards for penalizing critical diagnostic errors. The 5-point spread represents 28% of the total scale and could translate to meaningful differences in pass/fail determinations if such systems were used for high-stakes assessment. Notably, the split between graders who applied “actual findings” versus “internal consistency” standards for differential diagnosis evaluation (discussed in Section 3.6.2) accounted for much of this variability.

3.6.7. Divergent Ground Truth: The Gemini 3 Pro Anomaly

The most significant and consequential finding from the cross-evaluation protocol was Gemini 3 Pro’s fundamentally different interpretation of the radiological ground truth. While four graders established that the image showed an acute/subacute ischemic infarct with mass effect, Gemini 3 Pro concluded the image demonstrated a chronic left MCA infarction with encephalomalacia.

Gemini’s ground truth interpretation was characterized by the following findings, each directly contradicting the consensus interpretation:

- Ex-vacuo dilatation of the left lateral ventricle (vs. consensus: ventricular compression)
- Sulcal prominence overlying the lesion (vs. consensus: sulcal effacement)
- Absence of mass effect (vs. consensus: mass effect present)
- Central or slightly leftward midline position due to volume loss (vs. consensus: rightward midline shift due to mass effect)

This interpretive divergence is not a matter of nuance or differential emphasis; it represents a categorical disagreement about the pathophysiological state depicted in the image. Acute infarction with cytotoxic edema produces tissue swelling, sulcal effacement, ventricular compression, and contralateral midline shift. Chronic infarction with encephalomalacia produces tissue loss, sulcal prominence, ex-vacuo ventricular dilatation, and ipsilateral midline shift (or no shift). These are opposite findings with opposite clinical implications.

Gemini’s divergent interpretation propagated systematically through all of its evaluations. In its grading rationale, Gemini explicitly stated:

“Common Error: Most analyses incorrectly identified this as ‘Acute’ and described ‘mass effect/swelling,’ which is the opposite of the actual finding (atrophy/volume loss).”

Consequently, Gemini assigned 2/4 for diagnosis and 2/4 for features to all analyses (including its own) that described acute findings with mass effect. This uniform penalization compressed Gemini’s score distribution, creating the three-way tie for first place noted in Table 13 and producing a row mean (11.4) substantially lower than other graders.

The clinical implications of this ground truth disagreement are profound. If the image truly depicts acute infarction (as four graders concluded), then a patient would require emergent stroke management: blood pressure optimization, consideration of reperfusion therapy, neurological monitoring, and evaluation for decompressive craniectomy if malignant edema develops. If the image depicts chronic encephalomalacia (as Gemini concluded), no acute intervention is indicated, and management focuses on secondary prevention and rehabilitation.

That two sophisticated multimodal LLMs could arrive at diametrically opposed interpretations of the same image (interpretations with opposite therapeutic implications) raises fundamental questions about the reliability of these systems for autonomous clinical

deployment. Furthermore, this finding has direct implications for the use of LLMs in quality assurance roles: if grading LLMs cannot agree on what constitutes radiological ground truth, their evaluations of diagnostic accuracy become unreliable.

3.7. Summary of Cross-Evaluation Findings

The cross-evaluation protocol revealed several key insights regarding the reliability and consistency of LLM-based radiological interpretation and evaluation:

1. Rank-order consistency was high but not universal: Four of five graders produced identical or near-identical rankings, with Analyses 1 and 5 consistently rated highest and Analysis 4 unanimously rated lowest. However, Gemini 3 Pro's divergent ground truth interpretation produced a compressed, discordant ranking that would yield different conclusions about relative analysis quality.
2. Critical error detection was universal but penalization was inconsistent: All five graders correctly identified Analysis 4 as containing fundamental diagnostic errors, demonstrating that LLMs can detect clear-cut misinterpretations. However, the severity of penalization varied from 4/18 to 9/18, reflecting divergent evaluation philosophies regarding partial credit for internally consistent but externally incorrect responses.
3. Self-evaluation bias was present but heterogeneous: Four of five models demonstrated positive self-evaluation bias (+1.25 to +3.25 points), with Llama 4 Maverick showing the largest deviation. Gemini 3 Pro's apparent negative self-bias (−3.25 points) reflected consistent application of its divergent ground truth interpretation rather than genuine self-critical assessment.
4. Ground truth disagreement was the dominant source of variance: Gemini 3 Pro's interpretation of chronic versus acute infarction created systematic scoring discrepancies that exceeded inter-grader variation attributable to differences in stringency or rubric interpretation. This finding has profound implications for LLM-based quality assurance: if evaluator models cannot agree on fundamental radiological facts, their utility as automated graders is substantially compromised.
5. Grading stringency varied substantially: Mean scores assigned ranged from 11.4 (Gemini 3 Pro) to 15.8 (Llama 4 Maverick), a 4.4-point spread representing 24% of the total scale. This variability indicates that absolute scores from different LLM graders are not directly comparable and would require calibration for standardized assessment applications.
6. Evaluation philosophy diverged on key questions: Graders differed on whether differential diagnoses should be evaluated relative to actual imaging findings (awarding zero credit when predicated on incorrect diagnoses) or relative to internal consistency (awarding partial credit for plausible differentials given the stated diagnosis). This philosophical divergence contributed meaningfully to score variability, particularly for Analysis 4.

These findings collectively suggest that while LLMs demonstrate promising capabilities for radiological interpretation, their use as automated evaluators or quality assurance tools requires substantial methodological development. The combination of ground truth disagreement, self-evaluation bias, variable stringency, and divergent evaluation philosophies creates an assessment landscape that lacks the reliability and consistency expected for high-stakes educational or clinical applications.

4. Discussion

The results of this study reveal that even when multiple LLMs arrive at the same general diagnostic conclusion, significant variability persists in the clinical details of their

interpretations. Furthermore, the cross-evaluation protocol uncovered fundamental disagreements not only in diagnostic interpretation but also in the standards by which LLMs assess the quality of radiological analyses. These findings have important implications for the potential integration of LLM-based tools into radiological practice.

4.1. Concordance Does Not Imply Clinical Equivalence

Among the four models that correctly diagnosed left MCA territory ischemic stroke, no two responses were identical. Differences in acuity characterization (acute vs. acute-to-subacute vs. unspecified) could influence time-sensitive treatment decisions. In acute ischemic stroke, the differentiation between hyperacute, acute, and subacute phases is critical for determining eligibility for reperfusion therapies. The observation that GPT-5 characterized the stroke as definitively “acute” while Gemini 3 Pro and Claude Opus 4.5 Extended hedged with “acute to subacute” (and Llama 4 Maverick avoided temporal characterization entirely) suggests that current LLMs lack the consistency required for treatment-guiding acuity assessments.

4.2. Anatomical Precision and Prognostic Implications

The inconsistent reporting of anatomical involvement has direct relevance to prognosis communication and rehabilitation planning. Claude Opus 4.5 Extended’s omission of temporal lobe involvement, which was noted by the other three concordant models, could affect predictions regarding language function (given the temporal lobe’s role in Wernicke’s area) and memory. In a clinical context, such omissions could lead to incomplete counseling of patients and families regarding expected deficits and recovery trajectories.

4.3. Differential Diagnosis Heterogeneity and Downstream Testing

Perhaps most consequential from a resource utilization and patient safety standpoint is the heterogeneity in differential diagnoses. The differential diagnosis list generated by a radiologist often guides the ordering of subsequent investigations. The divergence between models (with GPT-5 recommending consideration of cerebral venous thrombosis (necessitating CT or MR venography), Gemini 3 Pro suggesting cerebral abscess (prompting infectious workup), and others proposing varying tumor grades (affecting urgency of contrast-enhanced MRI and oncological referral)) illustrates how reliance on any single LLM could bias clinical decision-making in unpredictable directions.

Furthermore, the inconsistency in tumor grade suggested in the differentials (high-grade vs. low-grade glioma) is not a trivial distinction. High-grade gliomas are aggressive malignancies requiring urgent multidisciplinary oncological management, whereas low-grade gliomas may be observed or managed less urgently. An LLM that systematically biases toward one characterization could influence the perceived urgency of follow-up imaging and specialist referral.

4.4. Completeness and Reliability of Image Analysis

The selective identification of incidental findings by only two of four concordant models raises concerns about the thoroughness of LLM-based image analysis. While choroid plexus and basal ganglia calcifications are typically benign, the failure to consistently detect and report all visible findings suggests that more clinically significant incidental pathology (such as aneurysms, masses, or vascular malformations) might similarly escape detection. This inconsistency undermines confidence in using LLMs as comprehensive image analysis tools.

4.5. The Insufficiency of Majority Consensus

It might be tempting to propose that multi-model consensus could mitigate the risks of individual LLM errors, as demonstrated by the 4:1 agreement on ischemic stroke in this study. However, our detailed analysis reveals that such consensus is superficial. The four “agreeing” models differed in ways that could materially affect patient care. Moreover, a consensus-based approach assumes that errors across models are independent and randomly distributed; which is an assumption that may not hold if models share similar training data, architectures, or systematic biases. The possibility of correlated errors across LLMs remains an underexplored risk in medical AI deployment.

4.6. Cross-Evaluation Reveals Deeper Inconsistencies

The cross-evaluation protocol provided a novel lens through which to examine LLM reliability, extending the analysis beyond diagnostic accuracy to encompass evaluative consistency. Several findings from this protocol warrant detailed discussion.

4.6.1. Self-Evaluation Bias: A Predictable but Concerning Pattern

The observation that most LLMs assigned higher scores to their own anonymized outputs than did other graders aligns with findings from human peer-review research, where self-assessment bias is well-documented. However, the magnitude of self-favoring bias varied considerably across models, with Llama 4 Maverick demonstrating the largest deviation (+3.25 points above others’ mean assessment). This finding suggests that some LLMs may be more prone to self-preferential evaluation than others, which is a consideration that becomes critical if such systems are deployed for automated quality assurance or peer-review applications.

The anomalous case of Gemini 3 Pro (which assigned its own analysis a lower score than other graders) initially appears to represent humility or conservative self-assessment. However, closer examination reveals that this negative bias stemmed from Gemini’s divergent interpretation of the ground truth rather than genuine self-criticism. Gemini penalized its own analysis (and all others) for describing “acute” findings that contradicted Gemini’s internal reference standard of chronic infarction. This finding illustrates how apparent self-penalization can mask fundamental interpretive disagreements rather than reflect true evaluative objectivity.

4.6.2. Ground Truth Disagreement: The Fundamental Challenge

Perhaps the most consequential finding from the cross-evaluation protocol was Gemini 3 Pro’s fundamentally different interpretation of the radiological ground truth. While four graders established that the image depicted an acute/subacute ischemic infarct with mass effect, Gemini concluded the image demonstrated chronic encephalomalacia with atrophy, i.e., the opposite pathophysiological state.

This disagreement is not a matter of nuance or stylistic variation; it represents a categorical divergence in the interpretation of core radiological findings. The clinical implications of this disagreement are profound:

- An acute infarct with mass effect may require emergent intervention, including osmotic therapy, decompressive craniectomy evaluation, and close neurological monitoring.
- A chronic infarct with encephalomalacia requires no acute intervention and carries fundamentally different prognostic implications.

That two sophisticated multimodal LLMs could arrive at diametrically opposed interpretations of the same image raises serious questions about the reliability of these systems for autonomous clinical deployment. Furthermore, this disagreement propagated systematically through Gemini’s grading behavior, leading to uniformly lower scores for analyses

that described findings Gemini considered “hallucinated” but that four other models (and, presumably, human expert consensus) would consider accurate.

4.6.3. Inconsistent Standards for Penalizing Critical Errors

While all five grading LLMs correctly identified Analysis 4 (Grok 4) as the lowest-quality interpretation, the severity of penalization varied from 4/18 (Claude Opus 4.5 Extended) to 9/18 (Llama 4 Maverick). This range reflects divergent standards for how severely critical diagnostic errors should be penalized.

The decision to award partial credit for differentials that were internally consistent with a hemorrhage diagnosis but entirely irrelevant to the actual imaging findings (as observed in both Grok 4’s self-evaluation and Llama 4 Maverick’s grading) represents a philosophically distinct approach to evaluation. One perspective holds that differentials should be judged purely on internal consistency with the stated diagnosis; another holds that differentials are only meaningful if the primary diagnosis is correct and thus should receive zero credit when predicated on a fundamentally erroneous interpretation.

This lack of consensus on evaluation philosophy has practical implications for any proposed use of LLMs as automated graders. Without standardized rubric interpretation and calibration protocols, LLM-generated scores would exhibit unacceptable variability, undermining their utility for formative or summative assessment purposes.

4.6.4. Grading Stringency as a Confounding Variable

The substantial variation in mean scores assigned by different grading LLMs (ranging from 11.4 for Gemini 3 Pro to 15.8 for Llama 4 Maverick) demonstrates that LLMs possess different baseline stringencies independent of the quality of analyses being evaluated. This finding parallels observations in human educational assessment, where inter-rater reliability requires calibration exercises and standardization protocols.

If LLMs were to be deployed as automated graders for radiological education or quality assurance, this variability would necessitate either: (a) normalization procedures to adjust for model-specific stringency, (b) ensemble approaches that aggregate scores across multiple LLM graders, or (c) calibration protocols that align LLM scoring to human expert standards. None of these solutions are trivial to implement, and each introduces additional complexity and potential failure modes.

4.7. Distinguishing Clinical Deployment Contexts

The findings of this study have different implications depending on the intended deployment context for LLM-based medical image interpretation. Three distinct use cases warrant separate consideration: autonomous diagnosis, decision support under radiologist oversight, and educational summarization.

4.7.1. Autonomous Diagnosis

Autonomous diagnosis refers to deployment contexts wherein LLMs generate diagnostic interpretations without mandatory human expert review prior to clinical action. Our findings unequivocally demonstrate that current multimodal LLMs are unsuitable for such applications. The 20% rate of fundamental diagnostic error (misidentifying ischemic stroke as intracerebral hemorrhage with incorrect lateralization), the categorical disagreement about infarct acuity between Gemini 3 Pro and four other models (chronic versus acute), and the persistent clinically meaningful variability even among ostensibly concordant interpretations collectively establish that these systems lack the reliability required for autonomous clinical deployment. The cross-evaluation findings further reveal that LLMs cannot agree on radiological ground truth, undermining any proposal for autonomous quality assurance roles. Given these limitations, autonomous diagnostic deployment of

general-purpose multimodal LLMs for medical imaging interpretation is not supported by the evidence presented in this study.

4.7.2. Decision Support Under Radiologist Oversight

Decision support applications represent a middle ground wherein LLM outputs serve as preliminary screening tools, second-reader aids, or workflow optimization components, but remain subject to verification by expert radiologists who retain ultimate diagnostic authority. In such contexts, LLMs might offer value for specific constrained tasks: preliminary triage of normal versus abnormal studies, flagging of potentially urgent findings for expedited review, or generation of structured preliminary reports that radiologists revise. However, several findings from this study suggest caution even in supervised contexts. First, the confident, authoritative tone of LLM outputs (complete with technical terminology and structured formatting) may foster cognitive anchoring, wherein radiologists' interpretations are inappropriately influenced by the preliminary AI assessment. Second, the cross-evaluation findings demonstrate that LLMs cannot reliably serve as automated evaluators or quality assurance tools, limiting their utility for double-reading applications or resident trainee assessment. Third, the absence of safety disclaimers in 80% of evaluated models suggests inadequate safeguards against inappropriate reliance. If decision support applications are pursued, institutions must implement rigorous protocols ensuring that radiologist review is genuinely independent and that LLM outputs are treated as preliminary information requiring verification rather than as reference standards.

4.7.3. Educational Summarization

Educational applications represent the most appropriate current use case for LLMs in radiology, provided appropriate disclaimers and supervision are maintained. LLMs may assist in generating teaching cases, summarizing imaging findings for learner discussion, or providing general explanatory content regarding radiological concepts. In such contexts, the goal is not to produce clinically actionable diagnoses but rather to facilitate learning and discussion. However, even educational applications require safeguards. The finding that only Claude Opus 4.5 Extended included explicit limitations disclaimers while four other models presented their interpretations without cautionary language underscores the need for mandatory safety messaging in educational contexts. Furthermore, educators must be aware that LLM-generated teaching content may contain the same diagnostic errors and inconsistencies documented in this study, necessitating expert review of AI-generated educational materials prior to dissemination. Educational summarization applications should be positioned as adjuncts to expert instruction rather than replacements for faculty teaching, with explicit acknowledgment of AI limitations integrated into the learning experience.

4.7.4. Implications for Development and Deployment

These distinctions carry implications for responsible AI development and deployment strategies. Autonomous diagnostic applications should prioritize validated, task-specific machine learning models trained for defined imaging modalities and pathological conditions, as detailed in Section 4.4 and Table 15. General-purpose LLMs, despite their impressive linguistic capabilities, are not optimized for the pixel-level accuracy and validated performance characteristics required for autonomous diagnosis. Decision support applications require explicit human-in-the-loop workflows with safeguards against cognitive anchoring and inappropriate reliance on preliminary AI outputs. Educational applications represent the most appropriate near-term use case for general-purpose LLMs in radiology, but require mandatory disclaimers, expert oversight, and integration into structured learning environments wherein AI limitations are explicitly taught alongside AI capabilities.

Table 15. Comparison of general-purpose multimodal LLMs versus task-specific machine learning models for medical image interpretation.

Characteristic	General-Purpose LLMs	Task-Specific ML Models
Training Paradigm	Massive, heterogeneous corpora spanning diverse domains; medical imaging content constitutes a small, potentially unrepresentative fraction	Curated datasets comprising thousands to millions of expert-annotated examples for a specific imaging modality and defined pathological conditions
Feature Optimization	Visual representations not optimized for specific diagnostic features; broad but shallow capabilities across many domains	Learned representations directly optimized for discriminative features relevant to the diagnostic task (e.g., hypodensity vs. hyperdensity, sulcal effacement vs. prominence)
Performance Characteristics	Largely unknown and unvalidated for radiological interpretation; 20% fundamental diagnostic error rate observed in this study; failure modes poorly characterized	Quantified sensitivity, specificity, and predictive values; validated against histopathological or clinically confirmed diagnoses; systematic failure mode analysis
Regulatory Status	Not validated or cleared for diagnostic imaging interpretation; no FDA authorization for autonomous clinical use	Multiple systems FDA-cleared (510(k) or De Novo) for specific indications; defined performance benchmarks and intended use populations
Interpretability	Text-based explanations generated post-hoc; may not reflect actual visual features driving interpretation (cf. Grok 4 describing “hyperdense” findings on hypodense lesion)	Direct visual attention mapping (CAM, Grad-CAM) highlighting image regions driving predictions; enables verification of reasoning alignment
Clinical Integration	No established workflows; outputs lack standardization for PACS integration; variable disclaimer inclusion	Purpose-built for clinical workflow integration; standardized outputs formatted for radiologist review and PACS compatibility
Appropriate Applications	Report summarization, clinical documentation assistance, natural language interfaces, literature synthesis, workflow orchestration	Diagnostic triage, lesion detection and localization, quantitative analysis, second-reader applications within validated scope

4.8. General-Purpose LLMs Versus Task-Specific Machine Learning Models

The diagnostic variability, fundamental interpretive disagreements, and critical errors documented in this study raise a broader question: are general-purpose multimodal LLMs the appropriate tool for medical image interpretation, or should clinical applications prioritize purpose-built machine learning models trained specifically for defined imaging tasks and pathological conditions [36]? Table 15 summarizes the key distinctions between these two paradigms across seven critical dimensions, providing a framework for the discussion that follows.

4.8.1. Architectural and Training Paradigm Differences

General-purpose multimodal LLMs such as those evaluated in this study are trained on massive, heterogeneous corpora spanning diverse domains, including but not limited to medical content. While this broad training enables impressive generalization across tasks, it fundamentally differs from the paradigm underlying specialized diagnostic ML systems. Task-specific models are trained on curated datasets comprising thousands to millions of labeled examples for a single imaging modality (e.g., non-contrast head CT) and a defined set of pathological conditions (e.g., acute ischemic stroke, intracranial hemorrhage, mass lesions), following established principles of predictive modeling in medicine [37]. This focused training enables optimization of feature representations directly relevant to the diagnostic task at hand.

The distinction is analogous to that between a general practitioner and a subspecialist radiologist. While the generalist possesses broad knowledge applicable across many domains, the subspecialist's concentrated expertise in a narrow field typically yields superior diagnostic accuracy for conditions within that specialty. Similarly, a convolutional neural network (CNN) or vision transformer trained exclusively on head CT images for stroke detection can optimize its learned representations for the specific visual features that distinguish ischemic from hemorrhagic pathology (hypodensity versus hyperdensity, cytotoxic edema versus vasogenic edema, sulcal effacement versus sulcal prominence) without the representational overhead required for unrelated tasks.

4.8.2. Evidence for Superior Performance of Task-Specific Models

A substantial body of literature supports the superior performance of task-specific ML models over general-purpose systems for defined medical imaging applications. Deep learning models trained specifically for stroke detection on non-contrast CT have demonstrated sensitivities exceeding 90% and specificities above 85% in multicenter validation studies, with performance approaching or matching that of expert neuroradiologists [38]. Cross-dataset validation approaches have similarly demonstrated robust generalization in ophthalmologic imaging applications [39]. Critically, these models undergo rigorous validation against histopathologically or clinically confirmed diagnoses, establishing ground truth in ways that the self-referential evaluation paradigm of the present study could not achieve.

In contrast, the LLMs evaluated here demonstrated a 20% rate of fundamental diagnostic error (1/5 models misdiagnosing ischemic stroke as hemorrhage), with an additional model (Gemini 3 Pro) demonstrating categorical disagreement about the acuity of the infarct (chronic versus acute). Such error rates and interpretive inconsistencies would be unacceptable in validated diagnostic AI systems intended for clinical deployment.

4.8.3. The Training Data Advantage

Task-specific ML models benefit from training datasets that are:

1. Curated for quality: Expert radiologists review and annotate each case, ensuring accurate labels that directly supervise model learning.
2. Balanced for pathology: Training sets are constructed to include representative examples of the full spectrum of disease presentations, including atypical cases and mimics, enabling robust differential performance.
3. Enriched for edge cases: Challenging cases that commonly cause diagnostic errors (e.g., subtle early infarcts, small hemorrhages, stroke mimics) can be oversampled to improve model performance on these clinically critical presentations.
4. Annotated with granular labels: Beyond binary classification (disease present/absent), task-specific datasets often include detailed annotations for lesion location, size, acuity, and associated features—enabling models to learn the nuanced radiological vocabulary required for comprehensive interpretation.

General-purpose LLMs, by contrast, derive their medical imaging capabilities from whatever medical content happened to be included in their broad training corpora. This content may be biased toward common presentations described in textbooks and review articles, underrepresenting the atypical or challenging cases that most benefit from AI assistance. Furthermore, the image–text pairs available for multimodal training may not include the granular annotations required for nuanced interpretation.

4.8.4. Validation and Regulatory Pathways

A critical distinction between task-specific ML models and general-purpose LLMs lies in their validation and regulatory status. Purpose-built diagnostic AI systems intended for clinical use undergo systematic validation against predefined performance benchmarks, typically including:

- Prospective validation on held-out datasets from multiple institutions.
- Comparison against expert consensus or histopathological ground truth.
- Subgroup analyses to assess performance across demographic groups and disease subtypes.
- Failure mode analysis to characterize the types of errors the system makes.
- Regulatory review (e.g., FDA 510(k) clearance [40] or De Novo authorization [41,42] in the United States).

Several task-specific AI systems for stroke detection and triage have received regulatory clearance, with published evidence supporting their clinical utility in reducing time to treatment and improving patient outcomes. These systems are deployed with defined indications for use, known performance characteristics, and established workflows for integration with radiologist oversight.

General-purpose LLMs, including those evaluated in this study, have not undergone equivalent validation for radiological interpretation. Their performance characteristics are largely unknown, their failure modes are poorly characterized, and their outputs lack the standardization required for reliable clinical integration. The present study represents an early effort to characterize LLM performance on a radiological task, and the findings underscore the gap between current capabilities and clinical-grade reliability.

4.8.5. Interpretability and Explainability

Task-specific ML models increasingly incorporate interpretability features that enhance clinical trust and utility. Techniques such as class activation mapping (CAM), gradient-weighted class activation mapping (Grad-CAM), and attention visualization enable these systems to highlight the image regions driving their predictions. A stroke detection algorithm can indicate the specific area of hypodensity it identified as ischemic tissue, enabling the reviewing radiologist to verify alignment between the model's reasoning and the actual imaging findings.

The LLMs evaluated in this study provided text-based descriptions of their reasoning, which superficially resembles interpretability. However, these textual explanations are generated post-hoc by the language model and may not accurately reflect the visual features that drove the underlying interpretation. The case of Grok 4 is illustrative: despite describing “hyperdense” findings consistent with hemorrhage, the image actually demonstrated hypodensity consistent with infarction. The model's explanation was internally coherent but fundamentally disconnected from the actual image content—a failure mode that would be more readily detected in a system with direct visual attention mapping.

4.8.6. The Case for Specialized Models in Clinical Practice

The findings of this study support the position that clinical applications requiring reliable medical image interpretation should prioritize validated, task-specific ML models over general-purpose LLMs. This recommendation is grounded in several considerations:

1. Superior accuracy: Task-specific models trained and validated for defined diagnostic tasks consistently outperform general-purpose systems on those tasks.

2. Known performance characteristics: Validated models have quantified sensitivity, specificity, and failure modes, enabling appropriate clinical deployment and interpretation of outputs.
3. Regulatory oversight: Cleared diagnostic AI systems have undergone independent review of safety and effectiveness, providing a layer of quality assurance absent for general-purpose LLMs.
4. Workflow integration: Purpose-built systems are designed for integration into clinical workflows, with outputs formatted for radiologist review and PACS integration.
5. Defined scope: Task-specific models operate within clearly defined boundaries, reducing the risk of inappropriate application to tasks outside their validated capabilities.

These distinctions are consolidated in Table 15, which contrasts general-purpose LLMs and task-specific ML models across training paradigm, feature optimization, performance characteristics, regulatory status, interpretability, clinical integration, and appropriate applications.

This is not to suggest that general-purpose LLMs have no role in healthcare. Their capabilities in natural language processing, information synthesis, and clinical documentation may offer substantial value in appropriate applications. However, for the specific task of medical image interpretation (where diagnostic accuracy directly impacts patient safety) the evidence favors specialized systems purpose-built for defined imaging applications.

4.9. Implications for LLM-Based Quality Assurance

The cross-evaluation findings substantially complicate proposals for using LLMs as automated quality assurance tools in radiology. While recent work has explored LLM applications in systematic review tasks such as risk of bias assessment [43], our findings suggest that fundamental ground truth instability may limit such applications in contexts requiring pixel-level image interpretation. The key challenges identified include:

1. Ground truth instability: LLMs may disagree on fundamental interpretive facts, not merely stylistic or descriptive nuances. A quality assurance system that cannot reliably establish ground truth cannot meaningfully assess deviation from it.
2. Evaluation philosophy divergence: LLMs apply inconsistent standards for partial credit, error severity weighting, and the relevance of internally consistent but externally incorrect responses.
3. Self-evaluation bias: If LLMs are used to evaluate outputs from the same or related systems, self-preferential scoring may systematically inflate quality assessments.
4. Stringency variation: Without calibration, raw scores from different LLM graders are not directly comparable, complicating aggregation and benchmarking efforts.

These challenges do not preclude all uses of LLMs in quality assurance contexts but do suggest that significant methodological development is required before such applications can be considered reliable.

Implications for AI Development Strategy

The comparison between general-purpose LLMs and task-specific ML models carries implications for AI development strategy in healthcare. While the allure of a single, general-purpose system capable of addressing all medical AI applications is understandable from an efficiency standpoint, the findings of this study suggest that such an approach may sacrifice the reliability required for clinical-grade performance.

A more prudent strategy may involve:

- Specialized diagnostic engines: Task-specific ML models developed, validated, and cleared for defined imaging applications, deployed as reliable diagnostic aids within their validated scope.

- LLMs as integration layers: General-purpose LLMs serving as natural language interfaces, clinical summarization tools, or workflow orchestrators, i.e., tasks that leverage their linguistic capabilities without requiring pixel-level diagnostic accuracy.
- Hybrid architectures: Systems that combine specialized diagnostic modules with LLM-based reporting and communication features, leveraging the strengths of each paradigm while mitigating their respective limitations.

Such an architecture would enable healthcare organizations to benefit from advances in both specialized ML and general-purpose AI while maintaining the reliability standards essential for patient safety.

4.10. Implications for Clinical Practice

These findings support the position that current LLMs, despite their impressive capabilities, are not yet suitable for independent clinical use in radiological interpretation. The variability documented here (even among models reaching the “correct” general diagnosis) underscores the continued necessity of expert human oversight. LLMs may serve as useful adjuncts for education, preliminary screening, or workflow optimization, but the clinical nuances required for safe and effective patient care demand the judgment of trained radiologists who can synthesize imaging findings with clinical context in ways that current AI systems cannot reliably replicate.

The cross-evaluation findings further suggest that LLMs are not yet suitable for autonomous quality assurance or peer-review functions. The combination of ground truth disagreement, inconsistent error penalization, and variable stringency creates an evaluation landscape that lacks the reliability required for high-stakes assessment applications.

Future research should focus on developing standardized benchmarks for evaluating not only diagnostic accuracy but also the clinical completeness and consistency of LLM-generated radiological interpretations. Additionally, methodologies for calibrating and validating LLM-based evaluation systems against human expert consensus are urgently needed. Until such benchmarks demonstrate robust inter-model reliability across the full spectrum of clinically relevant output dimensions, caution should govern the integration of these tools into diagnostic and evaluative workflows.

4.11. Variability in Safety Disclaimers and Ethical Safeguards

Regardless of diagnostic accuracy, inconsistency in the provision of safety disclaimers across LLM responses was identified. Of the five models evaluated, only Claude Opus 4.5 Extended included an explicit warning regarding the limitations of AI-generated medical interpretations. The disclaimer stated:

“This is an AI-assisted educational analysis only and should not be used for clinical decision-making. Always consult a qualified radiologist and clinical team for official interpretation.”

In contrast, GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Grok 4 provided their diagnostic interpretations without any accompanying cautionary language. This discrepancy raises significant concerns regarding the responsible deployment of LLMs in medical contexts.

4.11.1. Clinical and Ethical Implications

The absence of safety disclaimers in 80% of tested models (4/5) is particularly concerning given the authoritative tone with which these systems present their analyses. As the ethical landscape and public perspectives surrounding AI in healthcare continue to evolve [13], the responsibility to provide appropriate disclaimers becomes increasingly critical. When prompted with the phrase “You are a board-certified radiologist,” all models complied by generating responses styled as expert radiological interpretations. However, only Claude Opus 4.5 Ex-

tended contextualized its output with an appropriate disclaimer, explicitly acknowledging the boundary between AI-assisted analysis and clinical decision-making.

This finding has several implications:

1. **Risk of Overreliance:** Users unfamiliar with the limitations of LLMs may interpret undisclaimed outputs as authoritative medical opinions. The confident, structured format of responses from GPT-5, Gemini 3 Pro, Llama 4 Maverick, and Grok 4 (complete with bullet points, differential diagnoses, and technical terminology) could foster unwarranted trust in the accuracy and completeness of these interpretations.
2. **Medicolegal Considerations:** In jurisdictions where AI-assisted diagnostic tools are subject to regulatory oversight, the absence of disclaimers could expose developers, healthcare institutions, and individual practitioners to liability. The failure to communicate the investigational or educational nature of AI outputs may constitute a breach of informed consent principles if patients are affected by decisions influenced by these tools.
3. **Inconsistent Safety Culture Across Developers:** The variability in disclaimer inclusion reflects divergent approaches to safety and responsible AI deployment among leading AI developers. While Anthropic's Claude model demonstrated alignment with principles of transparency and user protection, other models (including those from organizations with substantial resources and stated commitments to AI safety) did not incorporate similar safeguards in this use case.
4. **Prompt Vulnerability:** The observation that role-playing prompts (e.g., "You are a board-certified radiologist") can elicit authoritative medical responses without automatic safety guardrails in most models suggests a vulnerability in current LLM design. Responsible deployment in healthcare settings may require mandatory, non-overridable disclaimers when medical content is detected, regardless of prompt framing.

4.11.2. The Disclaimer as a Minimum Standard

It is important to note that the inclusion of a disclaimer does not validate an LLM for clinical use; rather, it represents a minimum ethical safeguard. Claude Opus 4.5 Extended's diagnostic interpretation, while accompanied by appropriate cautionary language, still contained the same types of variability observed in other models (e.g., omission of temporal lobe involvement, characterization as "acute-to-subacute"). The disclaimer appropriately signals that such outputs require expert verification; i.e., a message absent from the other four models.

4.12. The Correlated Error Problem

A fundamental assumption underlying multi-model consensus approaches is that errors across different LLMs are statistically independent. However, several factors suggest this assumption may be violated:

1. **Shared Training Data:** Many LLMs are trained on overlapping corpora, including publicly available medical literature, textbooks, and imaging databases. Systematic biases or errors present in these shared sources may propagate to multiple models.
2. **Architectural Similarities:** Despite differences in specific implementations, many leading LLMs share fundamental architectural principles (transformer-based architectures, attention mechanisms, similar tokenization strategies). These shared design choices may lead to correlated failure modes.
3. **Benchmark Contamination:** Models are increasingly optimized against publicly available benchmarks, which may lead to convergent behavior on tasks similar to benchmark items while masking divergent performance on novel or atypical presentations.

4. Knowledge Cutoff Effects: LLMs with similar training cutoff dates may share the same knowledge gaps regarding recent developments, guideline updates, or emerging pathological entities.

The observation that four of five models converged on the same general diagnosis (ischemic stroke) while one model (Grok 4) made a fundamental error (hemorrhage) might initially suggest that majority consensus is protective. However, the cross-evaluation findings (particularly Gemini 3 Pro's divergent ground truth interpretation) demonstrate that even among "agreeing" models, fundamental interpretive differences may exist. Furthermore, a scenario in which correlated errors lead multiple models to converge on the same incorrect diagnosis cannot be excluded based on the present study and would represent a particularly dangerous failure mode for consensus-based systems.

4.13. Recommendations

Based on these findings, we recommend that:

- For clinical diagnostic imaging applications requiring reliable automated interpretation, healthcare institutions prioritize validated, task-specific machine learning models over general-purpose multimodal LLMs. Purpose-built diagnostic AI systems trained and cleared for defined imaging modalities and pathological conditions offer superior accuracy, known performance characteristics, regulatory oversight, and established clinical workflows compared to general-purpose LLMs, which demonstrate unacceptable variability and error rates for autonomous diagnostic use.
- General-purpose LLMs, if deployed in healthcare settings, should be restricted to applications that leverage their linguistic capabilities (such as report summarization, clinical documentation assistance, or natural language interfaces) rather than pixel-level diagnostic interpretation, for which they are not optimized.
- AI developers implement mandatory, context-aware disclaimers for medical content that cannot be suppressed by user prompts.
- Healthcare institutions establishing policies for LLM use require disclaimer verification as a prerequisite for any permitted applications.
- Regulatory bodies consider disclaimer requirements as part of pre-market approval processes for AI tools intended for healthcare-adjacent applications.
- Users of LLMs for medical education or preliminary analysis be trained to recognize the limitations of AI-generated content, regardless of whether disclaimers are present.
- Research into LLM-based quality assurance systems prioritize the development of calibration protocols, ground truth validation methods, and standardized rubric interpretation guidelines before deployment in educational or clinical settings.
- Multi-model consensus approaches, if pursued, be designed with explicit consideration of correlated error risks and validated against diverse, externally curated reference standards.
- Healthcare AI development strategies consider hybrid architectures that combine specialized diagnostic ML modules (for accurate image interpretation) with LLM-based components (for reporting, communication, and workflow integration), thereby leveraging the respective strengths of each paradigm while mitigating their limitations.

4.14. Limitations

The five models evaluated were selected to represent major Western AI developers and their flagship multimodal products. Notable models from other developers were not included due to practical constraints on study scope. Future research should examine whether the patterns of diagnostic variability and cross-evaluation inconsistency observed here extend to models developed outside the organizational contexts represented in this study.

This study utilized a single non-contrast head CT case depicting unambiguous intracranial pathology. This design choice was intentional: by selecting a straightforward case, we aimed to establish a minimum competency threshold for LLM diagnostic performance. However, we acknowledge that a single case cannot capture the full spectrum of stroke presentations, imaging artifacts, anatomical variants, or diagnostically challenging edge cases that these systems would encounter in clinical practice. Our study design prioritized depth of analysis over breadth of cases, but future research employing multi-case datasets with varying levels of diagnostic difficulty would provide valuable complementary evidence regarding the generalizability of these findings across diverse clinical presentations.

This study did not include human radiologist ground truth labeling or grading of LLM responses. This omission was intentional: the authors, as researchers at a medical college with no financial or advisory relationships with any AI developer, sought to avoid producing comparative accuracy data that could be selectively cited to market specific LLM products. The selected image depicts pathology sufficiently obvious that radiologist confirmation would add little beyond affirming what is clinically apparent from the image itself, while potentially enabling claims such as “Model X achieved Y% agreement with expert radiologists.” Our objective was to demonstrate that even on straightforward cases, leading LLMs exhibit fundamental diagnostic errors and irreconcilable interpretive disagreements, i.e., findings that argue against the entire class of tools rather than favoring any particular product. Future research incorporating blinded radiologist evaluation may provide complementary evidence, but such studies should be designed with explicit safeguards against selective commercial citation of model-specific accuracy figures.

Finally, the inherent stochasticity of LLM inference limits the strict reproducibility of the findings reported in this study. Commercial LLM providers do not disclose system prompts, safety filter configurations, or image preprocessing pipelines, and these parameters may change without notice. The model versions and access dates are documented in Table 3, but researchers attempting replication should anticipate output variability. This irreproducibility is itself clinically relevant: medical applications typically require auditable, consistent outputs that current LLM architectures cannot guarantee.

5. Conclusions

This study demonstrates that current multimodal large language models are not reliable tools for autonomous medical image interpretation. By deliberately selecting a straightforward case depicting unambiguous intracranial pathology, we established a minimum competency threshold: if state-of-the-art LLMs cannot consistently interpret obvious pathology, their deployment on diagnostically challenging cases cannot be justified. A 20% rate of fundamental diagnostic error (misidentifying ischemic stroke as intracerebral hemorrhage with incorrect lateralization), persistent clinically meaningful variability among ostensibly concordant models, and irreconcilable ground truth disagreement between grading LLMs collectively establish that these systems lack the consistency and accuracy required for autonomous clinical deployment. The cross-evaluation protocol further revealed that LLMs cannot serve as reliable automated evaluators: self-evaluation bias, divergent grading philosophies, and categorical disagreement about radiological findings undermine any proposed use in quality assurance or educational assessment contexts.

The implications of these findings vary by deployment context. For autonomous diagnostic applications, healthcare institutions should recognize that general-purpose LLMs (despite their linguistic sophistication) are architecturally unsuited for tasks demanding pixel-level accuracy and validated performance characteristics. For applications requiring reliable automated image interpretation, validated task-specific machine learning models trained and cleared for defined imaging modalities and pathological conditions remain the

appropriate technological choice. Decision support applications, wherein LLMs serve as preliminary screening tools subject to expert verification, require explicit human-in-the-loop workflows with safeguards against cognitive anchoring and inappropriate reliance. Educational summarization represents the most appropriate current use case for LLMs in radiology, provided that mandatory disclaimers and expert oversight are maintained, and that AI limitations are explicitly integrated into the learning experience.

LLMs may offer value in healthcare applications that leverage their linguistic capabilities, such as clinical documentation, report summarization, or patient communication, but expert radiologist oversight remains non-negotiable for all diagnostic imaging interpretation. This study was designed to demonstrate class-wide limitations of multimodal LLMs for diagnostic imaging rather than to rank individual products, reflecting the authors' position that comparative performance assessments risk serving as inadvertent product endorsements. The intentional exclusion of human radiologist ground truth avoided generating accuracy metrics that could be selectively cited for commercial purposes while focusing attention on the fundamental reliability concerns that affect the entire category of general-purpose multimodal language models. Until LLMs demonstrate reproducible accuracy validated against external reference standards across diverse clinical presentations, their outputs should be treated as unverified preliminary information rather than actionable clinical guidance. The inherent stochasticity of LLM inference and the opacity of commercial model updates further limit the reproducibility and auditability that medical applications demand, reinforcing the case for task-specific models in contexts requiring consistent, validated performance.

Author Contributions: Conceptualization, S.H., M.M. and M.T.; methodology, S.H., M.M. and M.T.; software, S.H., M.M. and M.T.; validation, S.H., M.M. and M.T.; formal analysis, S.H., M.M. and M.T.; investigation, S.H., M.M. and M.T.; resources, S.H., M.M. and M.T.; data curation, S.H., M.M. and M.T.; writing—original draft preparation, S.H., M.M. and M.T.; writing—review and editing, S.H., M.M. and M.T.; visualization, S.H., M.M. and M.T.; supervision, S.H., M.M. and M.T.; project administration, S.H., M.M. and M.T.; funding acquisition, S.H., M.M. and M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated for this study are contained within the article. The complete set of LLM-generated radiological interpretations, cross-evaluation scoring matrices, and analytical results are presented in the manuscript. Further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the author(s) used AI-assisted tools for the purposes of grammar review and language editing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The authors have no financial, advisory, employment, or other relationships with OpenAI, Google DeepMind, Meta AI, xAI, Anthropic, or any other AI developer, whether or not their products were evaluated in this study. No AI developer was consulted during study design, provided funding, or had any input into data analysis or manuscript preparation. The selection of models for evaluation was based solely on technical criteria (multimodal capabilities, availability, and market prominence) without consideration of commercial interests. The study was conducted independently with the sole objective of evaluating the reliability of multimodal LLMs for medical image interpretation, not to rank or endorse any particular product.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
AVM	Arteriovenous Malformation
CT	Computed Tomography
CV	Coefficient of Variation
HSV	Herpes Simplex Virus
ICC	Intraclass Correlation Coefficient
LLM	Large Language Model
MCA	Middle Cerebral Artery
MRI	Magnetic Resonance Imaging

Appendix A. Complete Prompts Used in This Study

This appendix provides the complete, verbatim prompts used for both the diagnostic analysis phase and the cross-evaluation grading phase of the study. These prompts were submitted identically to all five LLMs to ensure consistency in evaluation parameters.

Appendix A.1. Diagnostic Analysis Prompt

The following prompt was submitted to all five LLMs along with the test image (Figure 3):

You are a board-certified radiologist. Please analyze the attached image and answer the following in a concise, bullet-point format:

1. *What type of imaging modality is shown? (e.g., X-ray, CT, MRI, ultrasound)*
2. *Which anatomical region or organ is depicted?*
3. *What is the most likely diagnosis based on the image findings?*
4. *Briefly describe the key image features that support your diagnosis.*
5. *List up to three possible differential diagnoses.*

No additional context, patient history, or clinical information was provided beyond what is contained in the prompt above. The image was uploaded directly through each platform's native image upload functionality without preprocessing or modification.

Appendix A.2. Cross-Evaluation Grading Prompt

The following prompt was submitted to each LLM for the cross-evaluation phase, along with the original test image and all five anonymized analyses labeled as Analysis 1 through Analysis 5:

You are an expert medical educator and radiologist tasked with systematically evaluating and grading radiological image analyses. You will assess 5 medical image interpretations, focusing on accuracy, completeness, and clinical appropriateness.

Evaluation Context

Image: [Medical image provided]

Original Prompt: [Original prompt provided]

Task: Grade 5 analyses (provided) for accuracy across 5 domains detailed in the original prompt.

The grading prompt was followed by the complete grading rubric (Appendix B), the stepwise evaluation instructions (Appendix C), and the quality assurance guidelines

(Appendix C.1). The five anonymized analyses were appended sequentially, with no identification of the originating model.

Appendix B. Detailed Grading Rubric

This appendix provides the complete grading rubric used in the cross-evaluation protocol. Each of the five LLMs used these criteria to systematically evaluate all five anonymized analyses across five domains.

Table A1. Complete grading rubric for cross-model evaluation of radiological image analyses.

Domain	Scale	Detailed Scoring Criteria
Imaging Modality	0–3	0 = Incorrect identification of modality; 1 = Vague or partially correct (e.g., “scan” without specifying type); 2 = Correct modality but lacks specificity (e.g., “CT” without mentioning contrast status or plane); 3 = Fully correct and specific (e.g., “non-contrast axial CT”)
Anatomical Region and Organ	0–3	0 = Incorrect anatomical localization; 1 = Vague or partially correct (e.g., “body” or “head” without further specification); 2 = Correct general region but lacks anatomical specificity (e.g., “brain” without specifying structures or slice level); 3 = Fully correct and specific with appropriate anatomical detail (e.g., “brain at the level of the lateral ventricles and basal ganglia”)
Most Likely Diagnosis	0–4	0 = Incorrect or irrelevant diagnosis; 1 = Vague, overly broad, or tangentially relevant diagnosis; 2 = Plausible diagnosis but not the most likely given the findings; 3 = Correct diagnosis but lacks clinical detail or precision; 4 = Fully correct, specific, and clinically appropriate diagnosis with relevant qualifiers (e.g., acuity, lateralization, vascular territory)
Key Image Features	0–4	0 = Absent, incorrect, or contradictory feature descriptions; 1 = Incomplete or vague descriptions that lack clinical utility; 2 = Some correct features identified but missing key findings; 3 = Mostly complete and accurate feature descriptions with minor omissions; 4 = Complete, accurate, and well-described key features that comprehensively support the diagnosis
Differential Diagnoses	0–4	0 = Absent, irrelevant, or implausible differential diagnoses; 1 = Only one plausible alternative provided; 2 = Two plausible alternatives provided; 3 = Three plausible alternatives provided but not prioritized or justified; 4 = Three plausible alternatives that are appropriately prioritized and clinically justified
Maximum Total Score		18 points

Each domain was scored independently, and total scores were calculated by summing across all five domains. The rubric was designed to balance objectivity (through clear anchors for each score level) with flexibility to accommodate different but clinically equivalent descriptions. Grading LLMs were instructed to apply these criteria systematically to all five analyses without knowledge of which model generated each response.

Appendix C. Grading Protocol: Stepwise Instructions and Quality Assurance Guidelines

To promote systematic reasoning and reduce variability, the grading prompt instructed each LLM to follow a stepwise chain-of-thought evaluation process. Notably, rather than providing an external reference standard, each grading LLM was instructed to use its own expert knowledge to establish the expected correct answer for each domain, i.e., an approach that enables assessment of whether LLMs hold consistent internal standards for radiological accuracy.

The stepwise instructions were as follows:

1. For each of the five analyses, follow these steps:
 - (a) For each domain (modality, anatomy, diagnosis, features, differential):
 - i. Use your expert knowledge to determine the most accurate and specific answer based on the image and context.
 - ii. Compare the analysis to your expert expectation for that domain.
 - iii. Assign a score based on the rubric.
 - iv. Provide a brief justification for the score, referencing your clinical reasoning.
 - (b) After all domains are graded:
 - i. Calculate and state the total score for the analysis.
 - ii. Provide a concise summary of the analysis's strengths and weaknesses.
2. Repeat this process for all five analyses.
3. At the end, rank the analyses from highest to lowest total score.

This self-referential grading paradigm (wherein each LLM serves as both participant and evaluator using its own internal knowledge as the reference standard) introduces an additional dimension of analysis. Discrepancies in scores assigned by different grading LLMs to the same analysis may arise not only from differences in grading stringency but also from fundamental disagreements about what constitutes the correct radiological interpretation. This design choice enables examination of whether LLMs possess internally consistent and externally concordant standards for radiological accuracy, a prerequisite for their use in automated quality assurance or educational assessment applications.

Appendix C.1. Quality Assurance Instructions

Additionally, the grading prompt included explicit quality assurance instructions to promote consistency and objectivity:

- Base all judgments strictly on expert clinical knowledge and the information provided in the analysis.
- Avoid speculation or assumptions beyond the evidence presented in the analysis.
- If information is missing or ambiguous, note this explicitly in the reasoning.
- Be concise, objective, and systematic in all grading.

These instructions were designed to standardize the evaluative approach across all grading LLMs, reducing variability attributable to differences in interpretive style or inferential tendencies. By requiring graders to flag missing or ambiguous information rather than making assumptions, the protocol aimed to distinguish between analyses that were incomplete versus those that were incorrect, which is a clinically important distinction.

References

1. Turing, A.M. I.—Computing Machinery and Intelligence. *Mind* **1950**, *LIX*, 433–460. [CrossRef]
2. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Mag.* **2006**, *27*, 12. [CrossRef]
3. Newell, A.; Simon, H.A. Computer science as empirical inquiry: Symbols and search. *Commun. ACM* **1976**, *19*, 113–126. [CrossRef]
4. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
5. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 6000–6010. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (accessed on 14 December 2025).
8. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey. *ACM Comput. Surv.* **2022**, *55*, 1–28. [CrossRef]
9. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. Available online: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf (accessed on 14 December 2025).
10. Chen, K.; Shao, A.; Burapachee, J.; Li, Y. Conversational AI and equity through assessing GPT-3’s communication with diverse social groups on contentious topics. *Sci. Rep.* **2024**, *14*, 1561. [CrossRef]
11. Denecke, K.; May, R.; Rivera-Romero, O. Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks. *J. Med. Syst.* **2024**, *48*, 23. [CrossRef]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1 (Long and Short Papers), pp. 4171–4186. [CrossRef]
13. Bekbolatova, M.; Mayer, J.; Ong, C.W.; Toma, M. Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. *Healthcare* **2024**, *12*, 125. [CrossRef] [PubMed]
14. Arisan, A.; Genç, C.; Duran, G.S. Orthodontic Biomechanical Reasoning with Multimodal Language Models: Performance and Clinical Utility. *Bioengineering* **2025**, *12*, 1165. [CrossRef] [PubMed]
15. Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J.A.; Kanjee, Z.; Parsons, A.S.; Ahuja, N.; et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw. Open* **2024**, *7*, e2440969. [CrossRef] [PubMed]
16. Urda-Cimpean, A.E.; Leucuța, D.C.; Drugan, C.; Dutu, A.G.; Calinici, T.; Drugan, T. Assessing the Accuracy of Diagnostic Capabilities of Large Language Models. *Diagnostics* **2025**, *15*, 1657. [CrossRef]
17. Tan, J.C.K. Coherent Interpretation of Entire Visual Field Test Reports Using a Multimodal Large Language Model (ChatGPT). *Vision* **2025**, *9*, 33. [CrossRef]
18. Lodwick, G.S.; Keats, T.E.; Dorst, J.P. The Coding of Roentgen Images for Computer Analysis as Applied to Lung Cancer. *Radiology* **1963**, *81*, 185–200. [CrossRef]
19. Lodwick, G.S. Computer-aided Diagnosis in Radiology: A Research Plan. *Investig. Radiol.* **1966**, *1*, 72–80. [CrossRef]
20. Kruger, R.P.; Townes, J.R.; Hall, D.L.; Dwyer, S.J.; Lodwick, G.S. Automated Radiographic Diagnosis via Feature Extraction and Classification of Cardiac Size and Shape Descriptors. *IEEE Trans. Biomed. Eng.* **1972**, *BME-19*, 174–186. [CrossRef]
21. Toriwaki, J.I.; Suenaga, Y.; Negoro, T.; Fukumura, T. Pattern recognition of chest X-ray images. *Comput. Graph. Image Process.* **1973**, *2*, 252–271. [CrossRef]
22. Kruger, R.P.; Thompson, W.B.; Turner, A.F. Computer Diagnosis of Pneumoconiosis. *IEEE Trans. Syst. Man Cybern.* **1974**, *SMC-4*, 40–49. [CrossRef]
23. Turner, A.F.; Kruger, R.P.; Thompson, W.B. Automated Computer Screening of Chest Radiographs for Pneumoconiosis. *Investig. Radiol.* **1976**, *11*, 258–266. [CrossRef] [PubMed]
24. Fenton, J.J.; Taplin, S.H.; Carney, P.A.; Abraham, L.; Sickles, E.A.; D’Orsi, C.; Berns, E.A.; Cutter, G.; Hendrick, R.E.; Barlow, W.E.; et al. Influence of Computer-Aided Detection on Performance of Screening Mammography. *N. Engl. J. Med.* **2007**, *356*, 1399–1409. [CrossRef] [PubMed]
25. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]

26. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **2018**, *15*, e1002686. [\[CrossRef\]](#)
27. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402. [\[CrossRef\]](#)
28. Husain, G.; Mayer, J.; Bekbolatova, M.; Vathappallil, P.; Matalia, M.; Toma, M. Machine learning for medical image classification. *Acad. Med.* **2024**, *1*. [\[CrossRef\]](#)
29. Toma, M.; Husain, G. Algorithm Selection and Data Utilization in Machine Learning for Medical Imaging Classification. In *2024 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [\[CrossRef\]](#)
30. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [\[CrossRef\]](#)
31. Franco, L.; Toma, M. Validation strategies for automated MRI-based classification of Alzheimer’s disease using deep feature extraction and machine learning. *Artif. Intell. Health* **2025**, 025360073. [\[CrossRef\]](#)
32. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#)
33. Azizian, P.; Honarmand, M.; Jaiswal, A.; Kline, A.; Dunlap, K.; Washington, P.; Wall, D.P. Multimodal LLM vs. Human-Measured Features for AI Predictions of Autism in Home Videos. *Algorithms* **2025**, *18*, 687. [\[CrossRef\]](#)
34. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2021; pp. 610–623. [\[CrossRef\]](#)
35. Nasef, D.; Nasef, D.; Sher, M.; Toma, M. A Standardized Validation Framework for Clinically Actionable Healthcare Machine Learning with Knee Osteoarthritis Grading as a Case Study. *Algorithms* **2025**, *18*, 343. [\[CrossRef\]](#)
36. Toma, M. *AI-Assisted Medical Diagnostics: A Clinical Guide to Next-Generation Diagnostics*; Dawning Research Press: New York, NY, USA, 2025. Available online: <https://openlibrary.org/works/OL44048041W/> (accessed on 14 December 2025).
37. Toma, M.; Wei, O.C. Predictive Modeling in Medicine. *Encyclopedia* **2023**, *3*, 590–601. [\[CrossRef\]](#)
38. Chilamkurthy, S.; Ghosh, R.; Tanamala, S.; Biviji, M.; Campeau, N.G.; Venugopal, V.K.; Mahajan, V.; Rao, P.; Warier, P. Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. *Lancet* **2018**, *392*, 2388–2396. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Remyes, D.; Nasef, D.; Remyes, S.; Tawfello, J.; Sher, M.; Nasef, D.; Toma, M. Clinical Applicability and Cross-Dataset Validation of Machine Learning Models for Binary Glaucoma Detection. *Information* **2025**, *16*, 432. [\[CrossRef\]](#)
40. Muehlematter, U.J.; Bluethgen, C.; Vokinger, K.N. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510(k) predicate networks. *Lancet Digit. Health* **2023**, *5*, e618–e626. [\[CrossRef\]](#)
41. Merk, D.; Grisoni, F.; Friedrich, L.; Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **2018**, *1*, 68. [\[CrossRef\]](#)
42. Aboy, M.; Crespo, C.; Stern, A. Beyond the 510(k): The regulation of novel moderate-risk medical devices, intellectual property considerations, and innovation incentives in the FDA’s De Novo pathway. *npj Digit. Med.* **2024**, *7*, 29. [\[CrossRef\]](#)
43. Leucuța, D.C.; Urda-Cîmpean, A.E.; Istrate, D.; Drugan, T. Risk of Bias Assessment of Diagnostic Accuracy Studies Using QUADAS 2 by Large Language Models. *Diagnostics* **2025**, *15*, 1451. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.