

Original Paper

Discordance Between Textual Reasoning and Visual Interpretation in Large Language Models for Low Back Pain: Cross-Sectional Quantitative Evaluation and Exploratory Multimodal Stress Test

Ziyu Zhang^{1,2*}, BMed; Longhao Chen^{1,2,3*}, MD; Zhizhen Lv^{1,2,3*}, MD; Hanzhe Lv^{1,2}, BMed; Wei Sheng^{1,2}, BMed; Zicheng Wei⁴, MMed; Binghao Wang^{1,2}, BMed; Ying Shen^{1,2}, BMed; Yu Tian^{1,2,3}, MMed; Jingwen Hu⁵, MA; Zhifang Shen⁶, BMed; Lijiang Lv^{1,2,3}, Prof Dr

¹The Third Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China

²The Third School of Clinical Medicine (School of Rehabilitation Medicine), Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China

³Research Institute of Tuina (Spinal disease), Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China

⁴Zhejiang Provincial Hospital of Traditional Chinese Medicine, Hangzhou, Zhejiang, China

⁵Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China

⁶Jiaxing Hospital of Traditional Chinese Medicine, Jiaxing, Zhejiang, China

*these authors contributed equally

Corresponding Author:

Lijiang Lv, Prof Dr

The Third Affiliated Hospital of Zhejiang Chinese Medical University

No.219 Moganshan Road, West Lake District

Hangzhou, Zhejiang 310005

China

Phone: +86 18958077865

Email: lvlijiang0288@163.com

Abstract

Background: Large language models (LLMs) are rapidly evolving from text-based agents to multimodal systems capable of interpreting medical images. While their textual reasoning has improved, the safety implications of this shift remain underexplored, specifically regarding the alignment between visual interpretation and textual advice in low back pain (LBP) management.

Objective: This study aims to evaluate the performance of 10 commonly used and highly representative LLMs, including paid models such as ChatGPT 5 Plus (OpenAI) and reasoning-oriented models like Qwen3-Thinking (Alibaba/Qwen team) in the context of LBP. Specifically, it assesses their clinical accuracy, readability, and the “asymmetric evolution” in cross-modal diagnostic consistency.

Methods: We conducted a cross-sectional quantitative evaluation integrated with an exploratory multimodal stress test. First, 10 LLMs, including ChatGPT 5 and ChatGPT 5 Plus, Gemini 3 Pro (Google DeepMind), Claude 4.5 Sonnet (Anthropic), Grok 4 (xAI), DeepSeek, Kimi K2 (Moonshot AI), Doubao (ByteDance Seed [Doubao Team]), Qwen3 (Base), and Qwen3-Thinking, addressed 25 standardized LBP inquiries. A multidisciplinary expert panel extracted 1816 individual recommendations by segmenting the models’ point-by-point responses. These were independently coded for clinical accuracy against international guidelines on a 5-point Likert scale. Outcomes also included readability (Flesch Reading Ease [FRE] and Simple Measure of Gobbledygook [SMOG]), understandability (Patient Education Materials Assessment Tool [PEMAT]), and the presence of disclaimers. Second, to probe multimodal capabilities, we conducted an exploratory stress test (N=5) using a curated set of challenging clinical cases, including representative but complex LBP pathologies to assess baseline image recognition, and adversarial cases with deliberate clinical-radiological mismatches to evaluate cross-modal alignment.

Results: In textual tasks, models achieved an overall accuracy of 88.4% (1605/1816). ChatGPT 5 Plus demonstrated near-perfect performance, with zero serious errors. Within the only directly matched reasoning comparison, Qwen3-Thinking showed a lower serious-error rate than Qwen3 (Base). However, textual advice remained difficult to read, although understandability scores were robust. In contrast, multimodal performance in the exploratory stress test exhibited a negative divergence.

In pure imaging tasks featuring complex but clinically representative pathologies, the average diagnostic score was poor, with models failing to identify core pathologies. In comprehensive diagnostic analysis tasks (Q29-Q30), models exhibited “textual masking,” aligning visual findings with text cues rather than imaging evidence. Crucially, safety disclaimer coverage dropped precipitously from 82.8% (207/250) in textual tasks to 28% (14/50) in multimodal interactions.

Conclusions: Under adversarial stress conditions, current LLMs exhibit a capability mismatch: expert-level textual accuracy masks vulnerable visual interpretation. This “high-confidence trap” may cause users to misapply trust in textual logic to visual diagnostics. Although the matched Qwen3 comparison suggested a lower serious-error rate in the reasoning-enhanced setting, poor cross-modal alignment and the systemic absence of medical disclaimers in visual tasks pose substantial risks. Thus, current multimodal LLMs remain unready for direct patient use in LBP management.

JMIR Med Inform 2026;14:e93522; doi: [10.2196/93522](https://doi.org/10.2196/93522)

Keywords: lower back pain; large language models; AI; patient education; multimodal imaging

Introduction

Low back pain (LBP) dominates global disability charts, yet the most consequential shift in its management is unfolding outside the clinic walls [1]. Driven toward self-management, patients are increasingly incorporating AI as an accessible “first stop” for health information and preliminary triage guidance [2]. The tools people rely on have undergone major changes. Large language models (LLMs) have transitioned from passive information retrievers typical of traditional search engines to active “virtual consultants” [3].

Recent foundational studies, such as the benchmark by Scaff et al [4], established a critical early baseline for assessing text-based AI responses in LBP management. However, due to the rapid iteration of large model architectures and variations in prompt engineering, simple quantitative comparisons can no longer capture the current technological frontier [5]. Our study builds on the conceptual foundation of research by Scaff et al [4] to upgrade the evaluation framework, specifically addressing the newly evolved chain-of-thought reasoning capabilities [6] and nascent multimodal interactions of LLMs.

The core clinical question has shifted accordingly: for these models to effectively serve as “virtual advisors,” they must not only deliver high accuracy in text-based interactions but also maintain diagnostic reliability when visual information is introduced [7]. Do reasoning models truly reduce the risk of errors on the textual front [8]? And does the exploratory introduction of visual inputs create new pathways for misleading advice? The primary concern is “asymmetric evolution,” where sophisticated textual logic might mask underlying unreliability in visual interpretations.

To address this, we conducted a comprehensive assessment of 10 representative agents (including international models like ChatGPT 5 Plus (OpenAI) [9] and Chinese domestic models like DeepSeek [10] and Qwen [Alibaba/Qwen team] [11]), anchored by a robust text-based evaluation and supplemented by an exploratory multimodal probe. By building upon a systematic quantitative assessment of standardized textual queries and integrating a small-scale, exploratory multimodal case series, we aim to characterize the cross-modal performance gap between textual reasoning and visual processing and its potential implications

for patient use [12]. Ultimately, this study seeks to identify whether these “experts” exhibit capability imbalances, thereby providing a necessary caveat for digital health practices.

Methods

Study Design and Question Development

We constructed a cross-sectional stress test to evaluate the current capabilities of 10 LLMs [13]. To mimic real-life clinical scenarios, the initial question pool was generated by simulating a comprehensive patient perspective across 10 different LLMs. The specific prompt parameters used were “If you are a patient with low back pain and want to use a large language model to seek consultation and ask questions in order to gain relevant information, knowledge, and advice, please list the 20 most likely questions you would ask from the perspective of a patient (Note: Each question must contain only one inquiry point).” With each model providing 20 questions, this generated a total of 200 potential inquiries. Subsequently, for content validation, a multidisciplinary team of 6 senior experts (including clinical spine surgeons and researchers) categorized, screened, and further refined the language of these questions to ensure high quality and comprehensive multidimensional coverage. To verify that these inquiries authentically mimicked real-life phrasing and aligned with clinical realities, we consulted 10 patients with LBP from diverse backgrounds to evaluate the conversational tone and appropriateness of the questions. The items were iteratively revised based on their feedback until each question achieved consensus approval from at least 8 of the 10 patients. This rigorous process culminated in a final test set comprising 25 standardized textual questions spanning 3 key domains: self-management, risk factors, and treatment strategies. To probe the models’ multimodal capabilities, we designed an exploratory multimodal “stress test” using a purposively sampled series of clinical cases (N=5). Rather than aiming for statistical representativeness—which typically requires large-scale benchmarking datasets—this component was explicitly designed as an exploratory quantitative adversarial probe (akin to AI “red teaming”) to expose potential systemic vulnerabilities and cross-modal hallucinations. This test comprised 2 distinct challenges: (Q26-Q28) clinically common yet moderately complex LBP imaging cases, designed to assess baseline image

recognition capabilities; and (Q29-Q30) adversarial cases featuring deliberate clinical-radiological mismatches (where the clinical history text contradicted the imaging findings, requiring comprehensive synthesis to avoid misdiagnosis if relying solely on one modality), designed to evaluate the LLMs’ cross-modal analytical capabilities. Full details and deidentified images are provided in [Multimedia Appendix 1](#). Ultimately, this process culminated in a comprehensive final test set of 30 standardized questions, spanning both text-only inquiries and multimodal diagnostic tasks.

Model Selection and Configuration

Ten widely available models have been selected, covering international flagship LLMs such as ChatGPT 5 and ChatGPT 5 Plus, Grok 4 (xAI) [14], Gemini 3 Pro (Google DeepMind), and Claude 4.5 Sonnet (Anthropic) [15], as well as representative domestic LLMs from China, including DeepSeek V3.1, Kimi K2 (Moonshot AI), Doubao 1.6 Pro (ByteDance Seed [Doubao team]), and Qwen3-235B-A22B-2507 (encompassing both general and reasoning modes) [16]. (Nomenclature note: the exact versions evaluated are Qwen3-235B-A22B-2507 and its reasoning-enhanced variant Qwen3-235B-A22B-2507-Thinking. For visual clarity and to accommodate spatial constraints in certain figures and data-dense tables, these models are standardized and abbreviated globally as Qwen3 [Base] and Qwen3-Thinking [Alibaba/Qwen team] throughout the

text and visualizations.) All tests were conducted within a unified Edge browser environment using English prompts. Each query was initiated in a new chat session to minimize carryover effects.

To ensure representativeness and sufficient user accessibility, this study included 10 LLMs with high visibility and widespread usage both internationally and within China (comprising 5 international flagship models and 5 representative domestic Chinese models). We purposively established several internal comparative groups, such as comparing ChatGPT 5 with its paid counterpart, ChatGPT 5 Plus, and comparing the open-source Qwen3 (Qwen3-235B-A22B-2507) in its standard “Fast” mode—Qwen3 (Base)—against its “Thinking” (reasoning-enhanced) mode. Given the current paradigm shift where most state-of-the-art LLMs integrate advanced reasoning capabilities, activating these specific configurations (eg, “Thinking,” “Expert,” or “Extended Thinking” modes) generally yields superior logical deduction. Therefore, with the exception of the standard Qwen3 “Fast” model, which served as a general baseline, we prioritized choosing the enhanced reasoning modes for all applicable models during testing. The specific model versions, selected operational modes, and the exact testing date ranges (conducted systematically throughout November 2025) are thoroughly detailed in [Table 1](#).

Table 1. Overview of the evaluated large language models (LLMs), including specific versions, test dates, and operational modes.

Model name	Version ^a	Test date	Selected mode
ChatGPT	5.0	Nov 1-5, 2025	Thinking
ChatGPT	5.0 Plus	Nov 1-5, 2025	Thinking
Gemini	3 Pro	Nov 25-30, 2025	Pro
Grok	4	Nov 5-10, 2025	Expert
Claude	Sonnet 4.5	Nov 25-30, 2025	Extended Thinking
Doubao	1.6 Pro	Nov 10-15, 2025	Thinking
DeepSeek	V3.1	Nov 10-15, 2025	Deep Thinking
Kimi	K2	Nov 5-10, 2025	Thinking
Qwen	Qwen3-235B-A22B-2507	Nov 20-25, 2025	Fast
Qwen	Qwen3-235B-A22B-2507	Nov 20-25, 2025	Thinking

^aThe version column reports the technical or product version identifiers used for reproducibility; therefore, these entries may differ from the standardized display names used elsewhere in the manuscript.

To clarify the use of reasoning capabilities, our evaluation was not limited to Qwen3-Thinking. As shown in [Table 1](#), given that the vast majority of current LLMs offer modes with enhanced reasoning capabilities, we activated these advanced reasoning modes across the included models (with the exception of Qwen3 [Base], which served as a baseline control) to assess their performance under optimal conditions. This included the “Thinking” mode for ChatGPT 5, ChatGPT 5 Plus and Doubao, the “Pro” mode for Gemini 3 Pro, “Extended Thinking” for Claude 4.5 Sonnet, “Expert” mode for Grok 4, and “Deep Thinking” for DeepSeek V3.1. Throughout the manuscript, the term “reasoning models” serves as an umbrella term encompassing all these specific advanced configurations, rather than referring exclusively to Qwen.

Expert Recruitment and Selection

To ensure rigorous and objective evaluation across different domains, we established specialized expert panels for question design, multimodal evaluation, and linguistic assessment. For the clinical and multimodal assessments (including question refinement, textual accuracy coding, and imaging evaluation), the recruited experts (comprising spine surgeons, researchers, and senior radiologists) were required to meet strict inclusion criteria: (1) holding an associate senior professional title or higher; and (2) possessing a minimum of 10 years of active clinical or research experience in spine surgery, rehabilitation medicine, or radiology. For the readability and understandability assessment, the linguistic panel was assembled based on advanced English proficiency,

comprising native English-speaking researchers or nonmedical professionals holding a master's degree or higher in English, alongside 1 bilingual medical expert. To practically maintain strict blinding and prevent potential unblinding caused by the models' self-identifying tendencies, 2 research team members manually audited all generated outputs prior to evaluation. They systematically redacted all model-specific identifiers, explicit self-identifications, and brand-suggestive phrasing to preserve text purity. The fully anonymized text blocks were then distributed to the expert panel for independent grading.

Evaluation of Accuracy

Reference standards were established based on authoritative international guidelines, including Clinical guidelines for the diagnosis and treatment of nonspecific LBP in China [17], the National Institute for Health and Care Excellence (NICE) Guideline [18], American College of Physicians Clinical Practice Guideline [19], and The Lancet LBP Series [20-22].

To ensure output uniformity and facilitate precise coding by the expert panel, a consistent instructional prompt was appended to each textual question (Q1-Q25): "Please provide your answer point-by-point, ensuring that each point contains only a single, distinct recommendation." This instruction established a clear operational boundary, allowing the model outputs to be automatically segmented into bullet points.

A panel of 5 experts independently coded each individual recommendation extracted from model responses into 1 of 6 categories: correct recommendation or rejection (guideline-supported), incorrect recommendation or rejection (guideline-opposed), and unclear (not mentioned in the guidelines). To evaluate potential clinical safety risks, we specifically identified "serious errors," defined as recommendations that are life-threatening or could cause permanent functional loss. Interrater reliability was assessed using Fleiss κ , with values exceeding 0.60 interpreted as substantial agreement.

Given that most responses from LLMs encompass multiple recommendations, a 5-point Likert scale was used to quantify the overall quality of each model's response:

- 5 (excellent): 100% accuracy in all suggestions.
- 4 (good): inaccurate or unclear suggestions $\leq 25\%$; no serious errors.
- 3 (average): inaccurate or unclear suggestions $> 25\%$ but $\leq 50\%$, with no serious errors.
- 2 (poor): inaccurate or unclear suggestions $> 50\%$ but $\leq 75\%$, or contains 1 serious error.
- 1 (very poor): inaccurate or unclear suggestions $> 75\%$, or contains > 1 serious errors.

Given the known limitations of current vision-language models, we adopted an exploratory quantitative case series approach to the multimodal assessment. We selected 5 adversarial "probe cases" specifically designed to trigger potential hallucinations, rather than aiming for statistical representativeness.

Multimodal capabilities were assessed across 2 dimensions using a consensus standard defined by 3 senior radiologists and surgeons:

- Imaging recognition (Q26-Q28): the models generated radiology reports from computed tomography (CT) and magnetic resonance imaging (MRI) slices. Performance was evaluated using two metrics: (1) diagnostic consistency, scored on a 1-4 scale, where 4 indicates complete concordance with the reference diagnosis and 1 indicates $> 75\%$ error or severe hallucination; (2) descriptive accuracy, defined as the model's ability to identify core pathological findings specified in the gold standard, calculated as: descriptive accuracy = (number of key findings correctly mentioned by the AI/total number of key findings in the gold standard) $\times 100\%$.
- Comprehensive diagnostic analysis (Q29-Q30): the models integrated history, labs, and imaging for diagnosis and treatment planning. Scoring followed a 5-point scale:
 - 5 (excellent) correct primary diagnosis, all key secondary diagnoses correct, accurate clinical-imaging correlation, 0% noncompliant treatments, and no serious errors.
 - 4 (good) correct primary diagnosis, secondary diagnoses and clinical-imaging correlation correct or only minor or incomplete without changing clinical management, noncompliant treatments $\leq 25\%$, and no serious errors. This includes responses with correct primary and secondary diagnoses but $> 0\%$ and $\leq 25\%$ noncompliant treatments.
 - 3 (moderate) correct primary diagnosis but one or more key secondary diagnoses or clinical-imaging correlations missed or wrong, or noncompliant treatments $> 25\%$ and $\leq 50\%$, or 1 serious error.
 - 2 (poor) wrong primary diagnosis with partially correct secondary findings, or correct primary diagnosis with missed or wrong key secondary diagnoses plus noncompliant treatments $> 50\%$ and $\leq 75\%$, or more than 1 serious error.
 - 1 (very poor) both primary and key secondary diagnoses wrong or missing, or noncompliant treatments $> 75\%$.

The same 3 experts independently graded all LLM-generated multimodal outputs against this consensus standard. Interrater reliability was assessed using Fleiss κ , with values exceeding 0.60 interpreted as substantial agreement.

Readability, Disclaimer Coverage, and Statistical Analysis

Readability was quantified using the Flesch Reading Ease (FRE) [23] and the Simple Measure of Gobbledygook (SMOG) [24]. Understandability was evaluated using the Understandability section of the Patient Education Materials Assessment Tool-Print Materials (PEMAT-P) scale [25,26]. The actionability domain of the PEMAT-P was excluded based on two methodological considerations: (1) our analysis was conducted at the microlevel of 1816 individualized,

segmented recommendations rather than whole, intact educational brochures, meaning that applying actionability criteria to isolated sentences would introduce systemic floor effects; and (2) the primary focus of this linguistic assessment was to strictly isolate the local clarity and wording of the information, whereas the clinical appropriateness and safety of the actionable content had already been independently and thoroughly evaluated by the expert panel using a 5-point Likert scale. Language scoring was performed by a mixed panel comprising 3 nonmedical English professionals or native English-speaking researchers, along with 1 medical expert. Consensus on divergent items was achieved through panel discussions. Furthermore, the presence or absence of medical disclaimers (eg, “The above information is for reference only; please consult a professional doctor” or similar statements) was evaluated as an indicator of system-level guardrails, and the coverage rate for each model was subsequently calculated [27].

Ethical Considerations

This study was conducted as a substudy within the approved parent project, which was reviewed and approved by the Ethics Committee of The Third Affiliated Hospital of Zhejiang Chinese Medical University (Approval ZSLL-KY-2024-080-01). The parent project provided the ethical approval framework for the patient materials used in the present manuscript. The consultation with 10 patients with LBP for question-tone evaluation was covered under the approved protocol, and written informed consent was obtained from these participants. The retrospective imaging materials used in the exploratory multimodal stress test were deidentified before analysis, and individual informed consent for this retrospective imaging component was waived under the approved protocol. Before resubmission, all image panels included in [Multimedia Appendix 1](#) were re-reviewed at full resolution; patient names, patient IDs and medical record numbers, accession numbers, dates and times, age and sex information, and other patient-identifying

overlay information were removed. Nonidentifying anatomical orientation markers and lumbar level labels, such as R/L, L3/4, L4/5, and L5/S1, were retained only when necessary for anatomical interpretation. This manuscript represents an AI and LLM evaluation substudy based on materials from the approved parent project and did not involve additional patient intervention.

Results

Accuracy of Textual Responses

The interrater reliability assessment for text-based clinical accuracy coding yielded a Fleiss κ of 0.79, indicating substantial agreement among the expert panel regarding the clinical accuracy of the 1816 coded recommendations. In textual tasks, the models demonstrated a qualitative leap in performance compared to previous benchmarks, achieving an overall accuracy rate of 88.4% (1605/1816; [Figure 1](#)). The paid reasoning model, ChatGPT 5 Plus, exhibited near-perfect performance with 98.96% accuracy and zero instances of “inaccurate” or “serious errors” recommendations. Other top-tier models (ChatGPT 5 Free and Claude 4.5 Sonnet) also maintained high standards (>96% accuracy, ChatGPT 5 Free: 190/196, 96.9%; Claude 4.5 Sonnet: 194/202, 96.0%). It is noteworthy that in the English context, China’s domestic models, such as Doubao and DeepSeek, exhibit an accuracy rate below 80% (Doubao: 154/198, 77.8%; DeepSeek: 145/182, 79.7%), indicating a descriptive difference compared to top-tier models ([Figure 1](#)). This performance gap may be partially influenced by the English-language constraint, as these models were tested exclusively in a nonnative language, introducing a cross-lingual processing variable. In the within-model comparison of Qwen3, the reasoning or thinking mode (Qwen3-Thinking) showed a lower rate of serious errors than the standard mode, Qwen3 (Base; 1.64% vs 4.91%), suggesting, to some extent, the potential value of “slow thinking” in clinical contexts.

Figure 1. Proportions of accurate, inaccurate, serious errors, and unclear recommendations in text-based question answering across various large language models (LLMs). In terms of the overall quality scores across all responses, the performance of each model is illustrated in Table 2, while the question-level distribution of quality scores is illustrated in Figure 2. ChatGPT 5 Plus achieved the highest average overall quality score of 4.92 (SD 0.28), demonstrating exceptionally high accuracy in its responses. Following closely behind are ChatGPT 5 with a score of 4.68 (SD 0.75) and Claude 4.5 Sonnet at 4.56 (SD 1.00). Conversely, Doubao received the lowest average overall quality score (3.44, SD 1.26), followed by Qwen3 (Base; 3.60, SD 1.35), indicating relatively lower response quality characterized by inaccuracies or unclear content.

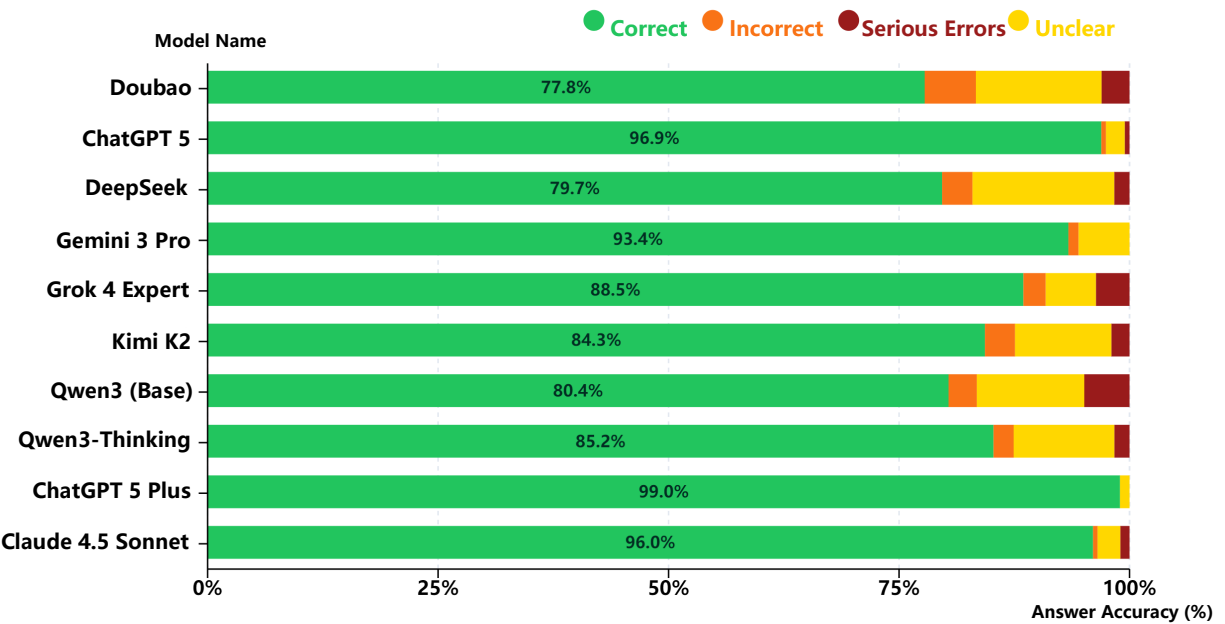


Table 2. Mean (SD) of text question answering scores for various large language models (LLMs) and disclaimer coverage rates across all questions.

Model	Answer accuracy score, mean (SD)	Disclaimer coverage rate, n/N (%)	Textual disclaimer coverage (Q1-Q25), n/N (%)	Multimodal disclaimer coverage (Q26-Q30), n/N (%)
ChatGPT 5 Plus	4.92 (0.28)	21/30 (70)	21/25 (84)	0/5 (0)
ChatGPT 5	4.68 (0.75)	28/30 (93.3)	25/25 (100)	3/5 (60)
Claude 4.5 Sonnet	4.56 (1.00)	25/30 (83.3)	25/25 (100)	0/5 (0)
Gemini 3 Pro	4.56 (0.51)	12/30 (40)	9/25 (36)	3/5 (60)
Grok 4 Expert	4.12 (1.27)	21/30 (70)	21/25 (84)	0/5 (0)
Qwen3-Thinking	4.00 (0.96)	24/30 (80)	23/25 (92)	1/5 (20)
Kimi K2	4.00 (1.00)	15/30 (50)	14/25 (56)	1/5 (20)
DeepSeek	3.88 (1.17)	29/30 (96.7)	25/25 (100)	4/5 (80)
Qwen3 (Base)	3.60 (1.35)	21/30 (70)	21/25 (84)	0/5 (0)
Doubao	3.44 (1.26)	25/30 (83.3)	23/25 (92)	2/5 (40)
Ensemble average ^a	4.18 (0.95)	221/300 (73.7)	207/250 (82.8)	14/50 (28)
Chinese models ^a	3.78 (1.16)	114/150 (76)	106/125 (84.8)	8/25 (32)
International models ^a	4.57 (0.86)	107/150 (71.3)	101/125 (80.8)	6/25 (24)

^aAggregated SDs for the “Ensemble average,” “Chinese models,” and “International models” rows were calculated based on pooled individual responses across the respective models.

Figure 2. Heatmap of response scores for each text question across various large language models (LLMs).



Multimodal Capability Assessment

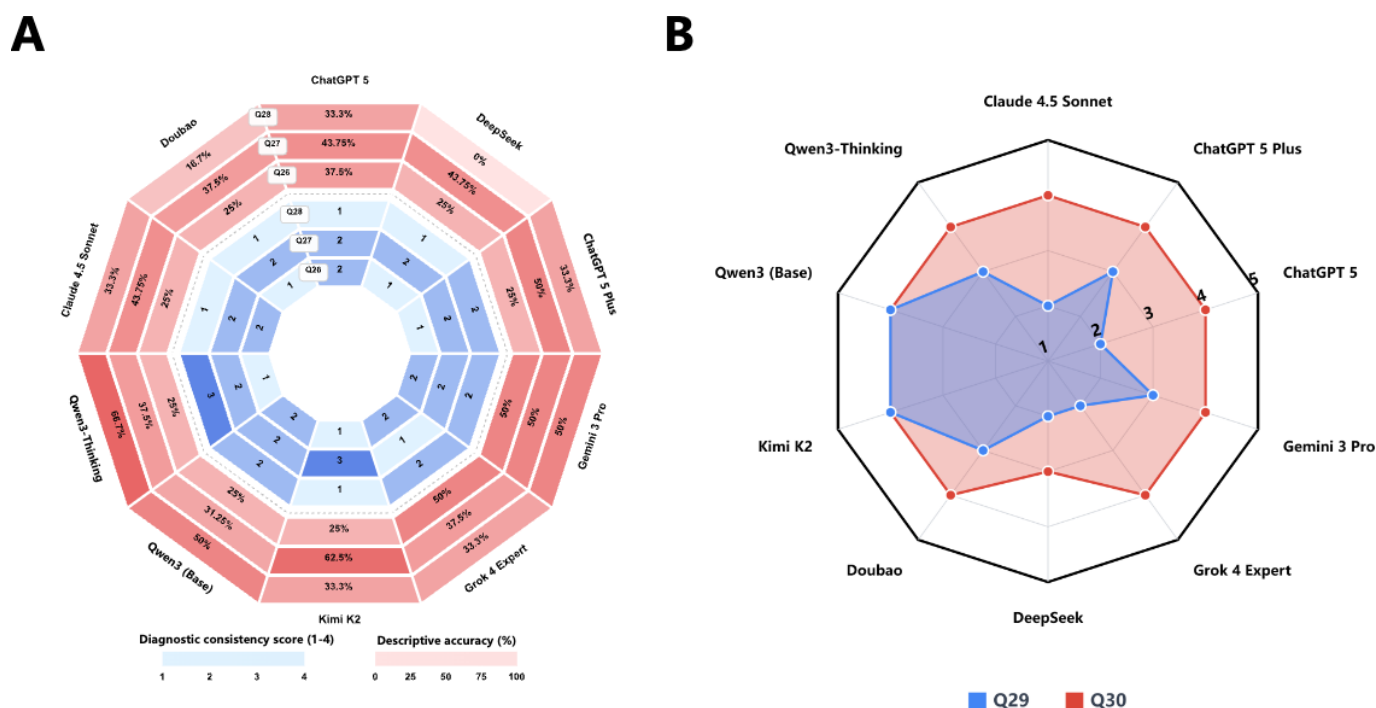
In stark contrast to the high proficiency observed in textual tasks, the exploratory adversarial stress test (N=5) suggested a potential “vision-language Mismatch” under these constrained test conditions. The interrater reliability for the multimodal assessment yielded a Fleiss κ of 0.82, indicating substantial agreement among the 3 expert raters.

Pure Imaging Tasks (Q26-Q28)

Performance in isolated visual interpretation was notably poor (Figure 3A). The average diagnostic score was 1.70/4.0

(below “partially accurate”). Models frequently failed to identify core pathological signs on CT and MRI slices. A distinct phenomenon of “textual-cue-driven guessing” was observed: in some cases, models appeared to extrapolate from minimal nondiagnostic textual cues, including the phrase “low back pain” in the task prompt and residual spinal-level labels retained in the deidentified images, rather than relying solely on visual evidence. For instance, when the visual input was ambiguous, DeepSeek generated plausible but hallucinated radiological descriptions that appeared to infer pathology from these limited contextual cues rather than from identifiable imaging findings.

Figure 3. Multimodal performance of the evaluated large language models (LLMs) in the exploratory stress test. (A) Performance on pure imaging recognition tasks (Q26-Q28). The circular heatmap presents diagnostic consistency in the inner blue rings (1-4 scale; higher scores indicate greater concordance with the reference diagnosis) and descriptive accuracy in the outer red rings (percentage of core pathological findings correctly identified). Q26, Q27, and Q28 denote the individual imaging test cases, and the labels around the plot identify the evaluated models. (B) Performance on comprehensive diagnostic analysis tasks (Q29-Q30). The radar chart shows model-level scores for Q29 (blue) and Q30 (red) on the 1-5 scoring scale, with higher values indicating better comprehensive diagnostic performance.



Comprehensive Diagnostic Analysis (Q29-Q30)

Performance showed an illusory improvement (mean score 3.35/5.0, SD 0.81; Figure 3B). However, a detailed breakdown based on our 5-point scoring rubric and case-level analysis indicated that this improvement was likely attributable to textual cues (eg, lab results, physical exam descriptions) acting as “anchors” rather than genuine visual understanding. A distinct pattern emerged in the model responses: models frequently secured baseline points by correctly identifying the primary diagnosis, driven by the clinical history, but consistently lost points (typically scoring 3 or 4) due to missing or misdiagnosing secondary conditions, which required genuine extraction of contradictory visual evidence. For example, models successfully identified the primary pathology in the ankylosing spondylitis “trap case” (Q30) because the provided laboratory data (HLA-B27 positive) guided their reasoning. This effectively bypassed their visual deficit, creating a “Clever Hans” effect in which the model appeared to diagnose the image but was actually diagnosing the text.

Readability and Understandability

The readability of all model-generated English texts is relatively high on average, failing to meet the recommended

standards for public health communication. The average FRI index is 46.15 (SD 16.39), falling into the “fairly difficult” range. The average SMOG index is 11.54 (SD 1.94; Table 3), indicating a reading level equivalent to approximately 11th-12th grade (US education system). ChatGPT 5 Plus produced the most readable texts (average FRI: 56.86, SD 15.24; SMOG: 8.75, SD 1.05; Table 3), while Grok 4 Expert generated the least readable content (average FRI: 37.7, SD 16.93; SMOG: 14.04, SD 2.61; Table 3). In terms of understandability, although the language complexity is high, the models perform moderately well in information organization and clarity of expression. At the segment level, the relatively high Patient Education Materials Assessment Tool (PEMAT) understandability scores appeared to reflect local clarity features, including explicit wording, single-topic recommendation units, and consistent point-by-point phrasing in the source responses, rather than macrolevel action tools or comprehensive document structure. The average PEMAT understandability score was 85.61 (SD 7.51; Table 3). Qwen3-Thinking had the highest score (89.60, SD 5.64; Table 3), while Grok 4 Expert had the relatively lowest understandability score (82.72, SD 7.21; Table 3).

Table 3. Mean (SD) of readability scores (Flesch Reading Ease [FRE] and Simple Measure of Gobbledygook [SMOG]) and understandability scores (Patient Education Materials Assessment Tool [PEMAT]) across all questions for various large language models (LLMs).

Model	FRE score, mean (SD)	SMOG score, mean (SD)	Understandability (PEMAT) score, mean (SD)
ChatGPT 5 Plus	56.86 (15.24)	8.75 (1.05)	83.44 (9.58)
ChatGPT 5	45.78 (18.63)	11.97 (2.84)	84.56 (8.80)
Claude 4.5 Sonnet	41.94 (19.08)	12.34 (2.09)	84.48 (8.26)
Gemini 3 Pro	54.56 (13.96)	10.86 (1.69)	86.60 (7.93)
Grok 4 Expert	37.72 (16.93)	14.04 (2.61)	82.72 (7.21)
Qwen3-Thinking	40.22 (17.11)	12.10 (2.14)	89.60 (5.64)
Kimi K2	42.33 (18.02)	11.74 (2.26)	84.40 (5.81)
DeepSeek	51.76 (13.20)	11.31 (1.53)	86.12 (7.54)
Qwen3 (Base)	39.67 (16.99)	11.77 (1.72)	86.52 (8.31)
Doubao	50.62 (15.14)	10.53 (1.44)	87.64 (6.05)
Ensemble average ^a	46.15 (16.39)	11.54 (1.94)	85.61 (7.51)
International models ^b	47.37 (8.17)	11.59 (1.96)	84.36 (1.47)
Chinese models ^b	44.92 (5.82)	11.49 (0.61)	86.86 (1.93)

^aSD for the “Ensemble average” was calculated based on pooled individual responses.
^bSDs for “International models” and “Chinese models” were calculated based on the model-level means.

Disclaimer Coverage

We observed marked descriptive variation and potential safety risks in the presentation of health disclaimers. Among all 300 responses, the overall inclusion rate was 73.67% (221/300). In pure text tasks (Q1-Q25), safety guardrails were robust: ChatGPT 5, Claude 4.5 Sonnet, and DeepSeek achieved a perfect 100% coverage rate, bringing the overall average coverage for text-based tasks to 82.8% (207/250).

However, a concerning red flag emerges when analyzing cross-modal performance (Table 2). When tasks transitioned from pure text to multimodal diagnostics (Q26-Q30), the overall disclaimer coverage plummeted to 28% (14/50), including 24% (6/25) for international models and 32% (8/25) for Chinese models. Notably, top-performing flagship models like ChatGPT 5 Plus, Claude 4.5 Sonnet, and Grok 4 Expert, which maintained exceptionally high textual disclaimer coverage (21/25, 84% to 25/25, 100%), saw their coverage drop entirely to 0.0% when responding to multimodal queries.

Discussion

Principal Findings

Our data confirm a leap in the development of AI over the past year [28]. Performance data indicate that the latest advancements in LLM capabilities may signify a quantum leap rather than merely iterative linear improvements [29]. Compared with the benchmarks studied by Scaff et al [4] a year ago, the new generation not only performs better but also possesses different functionalities. With overall accuracy reaching 88.4% (1605/1816), and specific architectures such as ChatGPT 5 Plus exceeding 98.96% [30], these systems appear to be transitioning beyond the role of passive information retrieval [31]. This represents a direct

manifestation of the architectural shift: moving away from the previous simplistic model that focused solely on rapid response and instead offering a reasoning mode with “long thinking” capability that implements the chain-of-thought protocol to address more complex logical problems [32]. Earlier versions displayed a more “impulsive” nature. In contrast, modern language models, such as Qwen3-Thinking [33,34], simulate an analytical reasoning process that closely resembles human “slow thinking.” They deconstruct clinical vignettes, explicitly weighing risk factors and synthesizing contradictory guidelines before committing to an output [35]. In the Qwen3 matched comparison, Qwen3-Thinking showed a lower serious-error rate than Qwen3 (Base; 1.64% vs 4.91%). This finding is best interpreted as a hypothesis-generating, within-model observation that reasoning-enhanced operation may reduce “serious errors” in this specific paired comparison, rather than as proof of a universal advantage of reasoning modes across all models. Therefore, longer “thinking” time should be considered a potential safety-relevant design feature that requires further paired validation, not a guaranteed safeguard.

Yet this textual mastery constructs a perilous facade. Our investigation exposes a profound “asymmetric evolution.” A sharp divergence emerged. While textual logic scores compare favorably with senior consultant levels, the visual diagnostic capabilities did not track; they remained, by contrast, unreliable and largely undeveloped in our testing [36]. This creates a “high-confidence trap.” The danger lies not in the incompetence itself but in the masking of that incompetence. In our multimodal case series, models frequently failed to identify even elementary pathological signs in pure imaging tasks (Q26-Q28), yielding a dismal diagnostic score of 1.70/4.00. If such results were to come from a radiologist, they would undoubtedly be deemed unqualified; however, LLMs present their errors with the

authoritative cadence of an expert. If unaddressed, this asymmetry could introduce new vectors for public health risks when patients interact with multimodal AI systems. The realism of our multimodal design, where patients present raw CT and MRI scans to LLMs, is deeply rooted in current digital health infrastructures and common clinical “waiting gaps.” Specifically, two pervasive real-world scenarios drive this patient behavior:

1. The follow-up gap via mobile portals: patients often access their imaging slices through hospital mobile apps days before their scheduled follow-up appointment, frequently downloading them due to health anxiety,
2. The radiology report is delayed: immediately after a scan, images are transmitted to the clinical workstation, allowing anxious patients to acquire image captures (eg, via smartphone photos of the screen) before the formal text report is finalized by a radiologist.

During these reportless waiting periods, driven by anxiety and eager for immediate answers, patients increasingly use LLMs to “preread” or interpret these fragmented image slices as an accessible “second opinion.” Therefore, evaluating this specific multimodal interaction is highly clinically relevant.

A patient, seduced by the model’s brilliant, nuanced answers to textual questions (eg, “How do I titrate my analgesics?”), will naturally extend that trust to the visual domain (eg, “Does this protrusion mean I have a herniated disc?”). They cross a capability cliff without a warning sign. This discrepancy may lead to misunderstandings, as users might mistakenly believe that the “brain” responsible for generating text responses in LLMs operates similarly to the “brain” that analyzes images. However, our exploratory studies demonstrate that this is not the case; in fact, the models were unable to accurately identify lesion sites in the lumbar vertebrae.

The “Clever Hans” effect [37] is a narrative coercion. The mechanism of this failure is perhaps more complex than simple “visual blindness.” Our “trap cases” (Q29-Q30) revealed a phenomenon of “textual masking.” The models appeared to succeed in comprehensive diagnosis not because they integrated the visual data but rather because they successfully ignored it.

When provided with a rich clinical history—such as keywords like “HLA-B27 (human leukocyte antigen B27) positive” or “morning stiffness”—the vast majority of models correctly identified the primary diagnosis of ankylosing spondylitis. However, their performance was markedly lower in pure image-recognition tasks. We have reason to infer that they were not diagnosing the patient; they were diagnosing the text. The observed mechanism mirrors the “Clever Hans” phenomenon, albeit computationally. The models appeared to function correctly on the surface, but in reality, they were likely reacting to textual cues (such as “HLA-B27”) rather than processing the visual pathology itself. The models hallucinated visual confirmations to align with the textual narrative.

This overreliance on textual cues is strongly validated by our scoring rubric for Q29-Q30. According to the predefined

5-point scale, a score of 3 or 4 indicates a correct primary diagnosis but errors or omissions in the secondary diagnosis. In our deliberately designed adversarial cases, the primary diagnosis was strongly suggested by the text (simulating clinical scenarios where imaging findings contradict the main symptoms, necessitating careful physical examination and comprehensive analysis), whereas the secondary diagnosis required a cross-modal synthesis incorporating the imaging. The models’ systematic failure to achieve perfect scores of 5 mathematically indicates that while their “textual brain” correctly parsed the history, their “visual eyes” failed to capture the radiological findings required to determine the secondary diagnosis.

In controlled testing, this approach yields a seemingly accurate label. However, in practice, it presents important practical safety challenges. For example, if a patient uploads an image of a vertebral fracture while describing symptoms resembling muscle strain, a text-reliant model constrained by the narrative may completely overlook the fracture. An auxiliary consultation tool that cannot prioritize visual evidence over patient statements serves not as a diagnostic aid but rather as an “echo chamber” [38].

Most alarming is the collapse of the safety architecture itself. We observed a large descriptive drop in safety-disclaimer coverage, from 82.8% (207/250) in textual tasks to 28% (14/50) in multimodal settings. This pattern suggests an observed gap in the model’s response protocols rather than simple random error. It signals a fundamental, structural flaw in the alignment of multimodal LLMs.

We hypothesize that this is “Cross-Modal Misalignment” [39]. The industry’s safety training—reinforcement learning from human feedback—has been predominantly textual. The model’s “superego”—its refusal mechanisms, its hedged language, its ethical guardrails—is triggered by textual tokens (eg, “diagnose,” “tumor”). Visual inputs, however, arrive via a separate pathway (eg, a vision transformer encoder) and are projected directly into the language space [40]. These visual embeddings appear to bypass the textual sentries. The model “sees” the pixels of a spine but fails to trigger the “I am an AI, not a doctor” protocol that the word “spine” would instantly activate. This “ethical bypass” leaves the user vulnerable precisely when the task is most complex and the model’s competence is at its nadir.

Consistent with the concerns raised by Scaff et al [4], the linguistic output of these models remains inaccessible to the population that needs it most. With a reading level hovering at grade 11-12 (versus the recommended grade 6-8), the advice is accurate but opaque. This is the tension of the training data: models fed on high-impact medical literature inevitably adopt an academic register.

However, the differences between international and domestic models suggest potential remedies. The Chinese models (Qwen and Doubao), while demonstrating relatively lower accuracy in English contexts, achieve superior PEMAT understandability scores. Their secret lies in their distinctive presentation formats. These models tend to use clearer local organizational cues, such as explicit labels, short, single-topic

units, and point-by-point phrasing. These features may improve segment-level understandability, but they should not be interpreted as macrolevel actionability tools or as comprehensive document structure [41]. If we cannot dumb down the vocabulary without losing precision, we must structure the layout. Future AI must prioritize this “structural readability,” ensuring that actionable advice is visually isolated from the dense reasoning that supports it.

Based on the aforementioned observations, there is an urgent need to transform passive monitoring into concrete interventions. For clinicians, they must actively uncover the “asymmetric” nature of these tools. The application of LLMs with patients must transition from a binary “trust and distrust” framework to a more nuanced operational guidance. Physicians can promote the use of LLMs for managing lifestyle and educational reinforcement—areas in which they excel—while issuing stern and explicit warnings against relying on them for image interpretation. The conclusion is clear; leverage the “brain” of LLMs, but never trust their “eyes.” For patients, “digital triage” is the new literacy. Users must understand that these agents are “virtual textbooks,” not “virtual radiologists.” A model’s confident description of an MRI and CT scan is often a linguistic fabrication—a statistical guess dressed in a lab coat. “Read the text, but ignore the diagnosis” must become the cardinal rule of AI self-help. For developers, the “safety gap” is an important technical challenge, not a backlog item. Future reinforcement learning from human feedback pipelines must incorporate multimodal adversarial examples. Models must be trained to establish a functional equivalence: an image of a bone must activate the same guardrails as the text string “diagnose this bone,” enforcing either an immediate refusal or a nonnegotiable disclaimer.

Implications of “Serious Errors” for Patient Safety

While the primary methodological focus of this study is the evaluation of clinical accuracy and cross-modal reliability, it is critical to recognize that, in the medical domain, profound inaccuracy inherently translates into patient safety risks. Our findings reveal a concerning dynamic: current LLMs demonstrate near-expert accuracy in textual reasoning, effectively minimizing “serious errors” (life-threatening advice). However, when transitioning to visual interpretation tasks, their diagnostic accuracy drops precipitously.

Crucially, this vulnerability is compounded by the systemic absence of medical disclaimers in multimodal tasks (dropping to 14/50, 28%). In these adversarial scenarios, this stark contrast represents a potential “high-confidence trap.” Patients, seduced by the model’s highly accurate textual advice, may mistakenly extend their trust to its fabricated visual interpretations. In a direct-to-patient application scenario, an AI that confidently hallucinates a benign condition when presented with a complex LBP imaging pathology—without issuing a disclaimer to consult a human doctor—could dangerously delay critical medical intervention. Therefore, while our metrics measure “accuracy,” the

profound lack thereof in multimodal settings poses a potential safety risk for patient self-management.

Limitations

First, while our study is anchored in a text-based evaluation, the sample size of the multimodal component (N=5) limits the statistical generalizability of the conclusions. We must emphasize that this component was designed exclusively as an exploratory adversarial probe rather than a statistically powered benchmark. Future studies with larger sample sizes should be conducted for more comprehensive research. Consequently, these multimodal findings reflect specific vulnerabilities under adversarial stress conditions rather than a comprehensive assessment of the models’ average baseline capabilities. Therefore, our results serve as a preliminary caveat, and the specific performance of LLMs in complex imaging tasks urgently requires validation through dedicated future studies using large-scale datasets. Furthermore, our comprehensibility scores may be relatively high; our raters consist of either English majors or native English-speaking researchers with advanced degrees, whose cognitive thresholds for academic terminology surpass those of the average patient. The PEMAT scale, while reliable, risks reducing the chaotic experience of comprehension to a single, sterile metric. We also acknowledge a linguistic bias; testing exclusively in English may have obscured the localization advantages of domestic models in their native Chinese contexts. Furthermore, because the FRE and SMOG indices are calibrated specifically for English syntax, their application to outputs from Chinese domestic models may not perfectly reflect true readability due to potential nonnative phrasing. Finally, we are chasing a moving target. Given the hyper-velocity of LLM iteration, these results are a snapshot frozen in time. The discrepancy between these findings and the model version available tomorrow is not a flaw in the study; it is the unavoidable condition of research in this field.

Future Work and Recommendations

Future research should focus on expanding the multimodal stress-test dataset to encompass a broader spectrum of musculoskeletal disorders and evaluating these models within real-world, prospective clinical workflows. To mitigate the risk of serious errors observed in our study, we recommend developing localized, medically fine-tuned multimodal LLMs—following validation via large-scale, statistically powered benchmark datasets—or integrating a dedicated “medical mode” into existing LLMs. Furthermore, in any future clinical or health guidance applications involving LLMs, it is imperative that AI-generated diagnostic recommendations be consistently verified by a team of qualified medical professionals, or that comprehensive coverage of disclaimers is guaranteed, to safeguard patient safety.

Conclusions

This study evaluated the performance of 10 representative LLMs in LBP management through a cross-sectional textual assessment and an exploratory multimodal adversarial stress test. In textual tasks, models achieved an overall accuracy of 88.4% (1605/1816). Within the matched Qwen3 pair,

Qwen3-Thinking showed a lower serious-error rate than Qwen3 (Base; 1.64% vs 4.91%), suggesting a possible safety benefit that requires confirmation in broader paired comparisons. In contrast, under exploratory multimodal adversarial stress conditions, models exhibited a marked decline in performance: diagnostic scores in pure imaging tasks were poor (mean 1.70/4.00), and models frequently relied

on textual cues rather than integrating contradictory visual evidence. Additionally, medical disclaimer coverage dropped from 82.8% (207/250) in textual tasks to 28% (14/50) in multimodal interactions. These findings indicate that current multimodal LLMs remain unready for direct patient use in visual diagnostic tasks.

Acknowledgments

The authors declare that no generative AI or AI-assisted technologies were used in the writing, editing, or preparation of this manuscript. The large language models (eg, ChatGPT, Claude, and DeepSeek) evaluated and discussed in this article were strictly the subjects of the experimental research and were not used to generate any portion of the manuscript's text.

Funding

This work was supported by the General Program of the National Natural Science Foundation of China (82305426), General Program of the National Natural Science Foundation of China (82474668), Key Discipline Project of High-level Traditional Chinese Medicine of the National Administration of Traditional Chinese Medicine (GJXK2023-85), National Advantage Specialty of Traditional Chinese Medicine (2024-90), Zhejiang Provincial Rising Medical Star (ytxx202402), and Natural Science Foundation of Zhejiang Province (LQN26H270009). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

Data are available upon reasonable request.

Authors' Contributions

Conceptualization: ZZ, LC, ZW

Data curation: LC, YS, BW

Formal analysis: LC, WS, BW

Funding acquisition: ZL, YT, LL

Methodology: ZZ, HL, JH

Project administration: ZL, LL

Supervision: YT

Validation: HL, YS, YT

Visualization: ZZ

Writing – original draft: ZZ, HL

Writing – review & editing: LC, ZL, ZS, LL

LL and ZS (email: shenzf08@163.com; phone: +86 15858362992) are co-corresponding authors.

ZZ, LC, and ZL contributed equally to this work and share first authorship.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details of the exploratory multimodal stress test (N=5), including deidentified images, structured large language model prompts, and gold-standard reference answers.

[[DOCX File \(Microsoft Word File\), 44396 KB-Multimedia Appendix 1](#)]

References

1. Ferreira ML, de Luca K, Haile L, et al. Global, regional, and national burden of low back pain, 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *Lancet Rheumatol*. 2023;5:e316–e329. [doi: [10.1016/S2665-9913\(23\)00098-X](https://doi.org/10.1016/S2665-9913(23)00098-X)]
2. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of self-diagnosis health chatbots in real-world settings: case study. *J Med Internet Res*. Jan 6, 2021;23(1):e19928. [doi: [10.2196/19928](https://doi.org/10.2196/19928)] [Medline: [33404508](https://pubmed.ncbi.nlm.nih.gov/33404508/)]
3. Idan D, Einav S. Primer on large language models: an educational overview for intensivists. *Crit Care*. Jun 12, 2025;29(1):238. [doi: [10.1186/s13054-025-05479-4](https://doi.org/10.1186/s13054-025-05479-4)] [Medline: [40506762](https://pubmed.ncbi.nlm.nih.gov/40506762/)]
4. Scaff SPS, Reis FJJ, Ferreira GE, Jacob MF, Saragiotto BT. Assessing the performance of AI chatbots in answering patients' common questions about low back pain. *Ann Rheum Dis*. Jan 2025;84(1):143–149. [doi: [10.1136/ard-2024-226202](https://doi.org/10.1136/ard-2024-226202)] [Medline: [39874229](https://pubmed.ncbi.nlm.nih.gov/39874229/)]

5. Matheny ME, Yang J, Smith JC, et al. Enhancing postmarketing surveillance of medical products with large language models. *JAMA Netw Open*. Aug 1, 2024;7(8):e2428276. [doi: [10.1001/jamanetworkopen.2024.28276](https://doi.org/10.1001/jamanetworkopen.2024.28276)] [Medline: [39150707](https://pubmed.ncbi.nlm.nih.gov/39150707/)]
6. Dai G, Liao J, Zhao S, et al. Large language model enhanced logic tensor network for stance detection. *Neural Netw*. Mar 2025;183:106956. [doi: [10.1016/j.neunet.2024.106956](https://doi.org/10.1016/j.neunet.2024.106956)] [Medline: [39637822](https://pubmed.ncbi.nlm.nih.gov/39637822/)]
7. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. Apr 2023;616(7956):259-265. [doi: [10.1038/s41586-023-05881-4](https://doi.org/10.1038/s41586-023-05881-4)] [Medline: [37045921](https://pubmed.ncbi.nlm.nih.gov/37045921/)]
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. Aug 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
9. Arslan HÇ, Arslan K, Gün M, Karkın PO, Arslanoğlu T. Evaluation of ChatGPT-5 responses in obstetric and gynecological emergencies: concordance, readability, and clinical reliability. *BMC Emerg Med*. Oct 30, 2025;25(1):220. [doi: [10.1186/s12873-025-01396-5](https://doi.org/10.1186/s12873-025-01396-5)] [Medline: [41163202](https://pubmed.ncbi.nlm.nih.gov/41163202/)]
10. Vural Camalan B, Doluoglu S, Taraf NH, Gunay MM, Ozlugedik S. ChatGPT versus DeepSeek in head and neck cancer staging and treatment planning: guideline-based study. *Eur Arch Otorhinolaryngol*. Sep 2025;282(9):4815-4824. [doi: [10.1007/s00405-025-09524-4](https://doi.org/10.1007/s00405-025-09524-4)] [Medline: [40523995](https://pubmed.ncbi.nlm.nih.gov/40523995/)]
11. Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models. *Natl Sci Rev*. Dec 2024;11(12):nwae403. [doi: [10.1093/nsr/nwae403](https://doi.org/10.1093/nsr/nwae403)] [Medline: [39679213](https://pubmed.ncbi.nlm.nih.gov/39679213/)]
12. Xue J, Wang Z, Chen N, et al. Evaluating multimodal ChatGPT for emergency decision-making of ocular trauma cases. *Front Cell Dev Biol*. 2025;13:1564054. [doi: [10.3389/fcell.2025.1564054](https://doi.org/10.3389/fcell.2025.1564054)] [Medline: [40213397](https://pubmed.ncbi.nlm.nih.gov/40213397/)]
13. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. Aug 3, 2023;620(7972):172-180. [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
14. Wu X, Cai G, Guo B, et al. A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. *BMC Oral Health*. Jul 28, 2025;25(1):1272. [doi: [10.1186/s12903-025-06619-6](https://doi.org/10.1186/s12903-025-06619-6)] [Medline: [40721763](https://pubmed.ncbi.nlm.nih.gov/40721763/)]
15. Sonoda Y, Kurokawa R, Nakamura Y, et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "diagnosis please" cases. *Jpn J Radiol*. Nov 2024;42(11):1231-1235. [doi: [10.1007/s11604-024-01619-y](https://doi.org/10.1007/s11604-024-01619-y)] [Medline: [38954192](https://pubmed.ncbi.nlm.nih.gov/38954192/)]
16. Tripathi A, Waqas A, Schabath MB, Yilmaz Y, Rasool G. HONeYBEE: enabling scalable multimodal AI in oncology through foundation model-driven embeddings. *NPJ Digit Med*. Oct 23, 2025;8(1):622. [doi: [10.1038/s41746-025-02003-4](https://doi.org/10.1038/s41746-025-02003-4)] [Medline: [41131352](https://pubmed.ncbi.nlm.nih.gov/41131352/)]
17. Gu R, Wang Y, Chen BH, Qiu Y. Clinical guidelines for the diagnosis and treatment of nonspecific low back pain in China. *Chin J Spine Spinal Cord*. 2022;32(3):258-268. [doi: [10.3969/j.issn.1004-406X.2022.03.09](https://doi.org/10.3969/j.issn.1004-406X.2022.03.09)]
18. Bernstein IA, Malik Q, Carville S, Ward S. Low back pain and sciatica: summary of NICE guidance. *BMJ*. Jan 14, 2017;356:i6748. [doi: [10.1136/bmj.n1627](https://doi.org/10.1136/bmj.n1627)] [Medline: [28062522](https://pubmed.ncbi.nlm.nih.gov/28062522/)]
19. Qaseem A, Wilt TJ, McLean RM, et al. Noninvasive treatments for acute, subacute, and chronic low back pain: a clinical practice guideline from the American College of Physicians. *Ann Intern Med*. Apr 4, 2017;166(7):514-530. [doi: [10.7326/M16-2367](https://doi.org/10.7326/M16-2367)] [Medline: [28192789](https://pubmed.ncbi.nlm.nih.gov/28192789/)]
20. Hartvigsen J, Hancock MJ, Kongsted A, et al. What low back pain is and why we need to pay attention. *Lancet*. Jun 2018;391(10137):2356-2367. [doi: [10.1016/S0140-6736\(18\)30480-X](https://doi.org/10.1016/S0140-6736(18)30480-X)]
21. Buchbinder R, van Tulder M, Öberg B, et al. Low back pain: a call for action. *Lancet*. Jun 9, 2018;391(10137):2384-2388. [doi: [10.1016/S0140-6736\(18\)30488-4](https://doi.org/10.1016/S0140-6736(18)30488-4)] [Medline: [29573871](https://pubmed.ncbi.nlm.nih.gov/29573871/)]
22. Clark S, Horton R. Low back pain: a major global challenge. *Lancet*. Jun 9, 2018;391(10137):2302. [doi: [10.1016/S0140-6736\(18\)30725-6](https://doi.org/10.1016/S0140-6736(18)30725-6)] [Medline: [29573869](https://pubmed.ncbi.nlm.nih.gov/29573869/)]
23. Flesch R. A new readability yardstick. *J Appl Psychol*. Jun 1948;32(3):221-233. [doi: [10.1037/h0057532](https://doi.org/10.1037/h0057532)] [Medline: [18867058](https://pubmed.ncbi.nlm.nih.gov/18867058/)]
24. Pominville R, Tay K, Callegari M, et al. Evaluating the readability of online testosterone search results. *Int J Impot Res*. Jun 2024;36(4):399-402. [doi: [10.1038/s41443-023-00682-9](https://doi.org/10.1038/s41443-023-00682-9)] [Medline: [36864203](https://pubmed.ncbi.nlm.nih.gov/36864203/)]
25. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns*. Sep 2014;96(3):395-403. [doi: [10.1016/j.pec.2014.05.027](https://doi.org/10.1016/j.pec.2014.05.027)] [Medline: [24973195](https://pubmed.ncbi.nlm.nih.gov/24973195/)]
26. Bonner C, Fajardo MA, Hui S, Stubbs R, Trevena L. Clinical validity, understandability, and actionability of online cardiovascular disease risk calculators: systematic review. *J Med Internet Res*. Feb 1, 2018;20(2):e29. [doi: [10.2196/jmir.8538](https://doi.org/10.2196/jmir.8538)] [Medline: [29391344](https://pubmed.ncbi.nlm.nih.gov/29391344/)]
27. Gallifant J, Fiske A, Levites Strekalova YA, et al. Peer review of GPT-4 technical report and systems card. *PLOS Digit Health*. Jan 2024;3(1):e0000417. [doi: [10.1371/journal.pdig.0000417](https://doi.org/10.1371/journal.pdig.0000417)] [Medline: [38236824](https://pubmed.ncbi.nlm.nih.gov/38236824/)]

28. McGinness L, Baumgartner P. Large language models imitate logical reasoning, but at what cost? In: Liu M, Yu X, Xu C, Song Y, editors. AI 2025: Advances in Artificial Intelligence: 38th Australasian Joint Conference on Artificial Intelligence, AI 2025, Canberra, ACT, Australia, December 1–5, 2025, Proceedings, Part I. Springer; 2025:80-96. [doi: [10.1007/978-981-95-4969-6_7](https://doi.org/10.1007/978-981-95-4969-6_7)]
29. Hagendorff T, Fabi S. Beyond chains of thought: benchmarking latent-space reasoning abilities in large language models. arXiv. Preprint posted online on Apr 14, 2025. [doi: [10.48550/arXiv.2504.10615](https://doi.org/10.48550/arXiv.2504.10615)]
30. Bird JJ, Ekárt A, Faria DR. Chatbot interaction with artificial intelligence: human data augmentation with T5 and language transformer ensemble for text classification. J Ambient Intell Human Comput. Apr 2023;14(4):3129-3144. [doi: [10.1007/s12652-021-03439-8](https://doi.org/10.1007/s12652-021-03439-8)]
31. Saeid Y, Neüburger F, Krügl S, Hüster H, Kopinski T, Lanwehr R. Harnessing large language models: fine-tuned BERT for detecting charismatic leadership tactics in natural language. Presented at: 2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS); Aug 26-28, 2024; Debrecen, Hungary. [doi: [10.1109/CITDS62610.2024.10791373](https://doi.org/10.1109/CITDS62610.2024.10791373)]
32. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Presented at: NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems; Nov 28 to Dec 9, 2022; New Orleans, Louisiana, USA. [doi: [10.52202/068431-1800](https://doi.org/10.52202/068431-1800)]
33. Ding J, Long S, Pu C, et al. NL2Repo-bench: towards long-horizon repository generation evaluation of coding agents. arXiv. Preprint posted online on Dec 14, 2025. [doi: [10.48550/arXiv.2512.12730](https://doi.org/10.48550/arXiv.2512.12730)]
34. Gong EJ, Bang CS, Lee JJ, Baik GH. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. J Med Internet Res. Dec 1, 2025;27:e84120. [doi: [10.2196/84120](https://doi.org/10.2196/84120)] [Medline: [41325597](https://pubmed.ncbi.nlm.nih.gov/41325597/)]
35. Yung KK, Ardern CL, Serpiello FR, Robertson S. Judgement and decision making in clinical and return-to-sports decision making: a narrative review. Sports Med. Aug 2024;54(8):2005-2017. [doi: [10.1007/s40279-024-02054-9](https://doi.org/10.1007/s40279-024-02054-9)] [Medline: [38922556](https://pubmed.ncbi.nlm.nih.gov/38922556/)]
36. Sepehri MS, Fabian Z, Soltanolkotabi M, Soltanolkotabi M. MediConfusion: can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models. Presented at: The Thirteenth International Conference on Learning Representations (ICLR 2025); Apr 24-28, 2025; Singapore. URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/b37c2e26b75ee02fcabd65a2a0367136-Abstract-Conference.html [Accessed 2026-07-24]
37. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun. Mar 11, 2019;10(1):1096. [doi: [10.1038/s41467-019-08987-4](https://doi.org/10.1038/s41467-019-08987-4)] [Medline: [30858366](https://pubmed.ncbi.nlm.nih.gov/30858366/)]
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. Feb 2020;128(2):336-359. [doi: [10.1007/s11263-019-01228-7](https://doi.org/10.1007/s11263-019-01228-7)]
39. Chakraborty T, Shayegani E, Cai Z, et al. Can textual unlearning solve cross-modality safety alignment? In: Al-Onaizan Y, Bansal M, Chen YN, editors. Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics; 2024:9830-9844. [doi: [10.18653/v1/2024.findings-emnlp.574](https://doi.org/10.18653/v1/2024.findings-emnlp.574)]
40. Fei N, Lu Z, Gao Y, et al. Towards artificial general intelligence via a multimodal foundation model. Nat Commun. Jun 2, 2022;13(1):3094. [doi: [10.1038/s41467-022-30761-2](https://doi.org/10.1038/s41467-022-30761-2)] [Medline: [35655064](https://pubmed.ncbi.nlm.nih.gov/35655064/)]
41. Balakrishnan V, Chandy Z, Hsieh A, Bui TL, Verma SP. Readability and understandability of online vocal cord paralysis materials. Otolaryngol Head Neck Surg. Mar 2016;154(3):460-464. [doi: [10.1177/0194599815626146](https://doi.org/10.1177/0194599815626146)] [Medline: [26861227](https://pubmed.ncbi.nlm.nih.gov/26861227/)]

Abbreviations

CT: computed tomography
FRE: Flesch Reading Ease
LBP: low back pain
LLM: large language model
MRI: magnetic resonance imaging
NICE: National Institute for Health and Care Excellence
PEMAT: Patient Education Materials Assessment Tool
SMOG: Simple Measure of Gobbledygook
PEMAT-P : Patient Education Materials Assessment Tool–Print Materials

Edited by Andrew Coristine; peer-reviewed by Pei-fu Chen, Peng Zhao; submitted 13.Feb.2026; final revised version received 09.Jul.2026; accepted 13.Jul.2026; published 20.Aug.2026

*Please cite as:**Zhang Z, Chen L, Lv Z, Lv H, Sheng W, Wei Z, Wang B, Shen Y, Tian Y, Hu J, Shen Z, Lv L**Discordance Between Textual Reasoning and Visual Interpretation in Large Language Models for Low Back Pain: Cross-Sectional Quantitative Evaluation and Exploratory Multimodal Stress Test**JMIR Med Inform 2026;14:e93522*URL: <https://medinform.jmir.org/2026/1/e93522>doi: [10.2196/93522](https://doi.org/10.2196/93522)

© Ziyu Zhang, Longhao Chen, Zhizhen Lv, Hanzhe Lv, Wei Sheng, Zicheng Wei, Binghao Wang, Ying Shen, Yu Tian, Jingwen Hu, Zhifang Shen, Lijiang Lv. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 20.Aug.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.