DOI: 10.7860/JCDR/2026/85628.23103

Original Article

Surgery Section

Diagnostic Concordance of Large Language Models with Postoperative Surgeon Diagnoses: A Comparative Analysis of ChatGPT, Claude, and Gemini

TASNEEM SHAIK ZAHEERUDEEN[1], NEMBIAN RAJA RAJAN[2]

ABSTRACT

Introduction: Large Language Models (LLMs) are increasingly explored for clinical diagnostic support, yet their real-world performance compared to clinician diagnoses remains inadequately characterised.

Aim: To evaluate the diagnostic concordance of three leading LLMs, ChatGPT, Claude and Gemini against postoperative surgeon diagnoses.

Materials and Methods: The present single centre retrospective observational comparative study including a total of 106 surgical cases from SRM Medical College and Hospitals, Kattankulathur, Chennai, Tamil Nadu, India, performed between March and September 2025, with documented presenting symptoms and confirmed postoperative diagnoses were analysed. Standardised prompts containing patient demographics, symptoms and clinical history were submitted to ChatGPT (GPT-4, version GPT-4-0613), Claude (Claude 3 Opus, version claude-3-opus-20240229) and Gemini (Gemini Pro, version gemini pro-1.0) between March 25, 2025 and September 24, 2025. Each LLM-generated diagnosis was independently categorised as concordant or discordant with the surgeon's diagnosis by two investigators. Model accuracy was calculated with Clopper-Pearson 95% confidence intervals. Pair-wise differences were assessed using McNemar's exact test and inter-model agreement was quantified with Cohen's kappa. Qualitative analysis identified strengths and limitations of Artificial Intelligence (AI) diagnostic reasoning.

Results: Mean patient age was 42.6±16.0 years; 70 (66.0%) were male. Diagnostic accuracy was 80.2% (95% CI: 71.3-87.3%) for ChatGPT (85/106 concordant), 83.0% (95% CI: 74.5-89.6%) for Claude (88/106 concordant) and 80.2% (95% CI: 71.3-87.3%) for Gemini (85/106 concordant). Pair-wise comparisons revealed no statistically significant differences (ChatGPT vs Claude, p=0.581; ChatGPT vs Gemini, p=1.000; Claude vs Gemini, p=0.375). Inter-model agreement was moderate to high (κ=0.592-0.843). All three models correctly diagnosed 78 cases (73.6%), while all three failed in 13 cases (12.3%). Qualitative analysis revealed that LLMs excelled at pattern recognition for classic presentations (e.g., appendicitis, hernias) but struggled with atypical symptoms, rare conditions and cases lacking imaging data.

Conclusion: ChatGPT, Claude and Gemini demonstrated comparable diagnostic concordance with postoperative surgeon diagnoses in this retrospective cohort. While LLMs show promise as adjunctive diagnostic tools, significant limitations in handling atypical presentations, contextual interpretation and accountability concerns preclude independent clinical deployment. These findings underscore the necessity of human oversight rather than replacement role of AI in surgical decision-making.

Keywords: Artificial intelligence, Diagnostic accuracy, Generative pretrained transformers, Surgical diagnosis

INTRODUCTION

The integration of Artificial Intelligence (AI) into healthcare has accelerated dramatically, with LLMs such as ChatGPT (OpenAI), Claude (Anthropic) and Gemini (Google) demonstrating capabilities in clinical reasoning, medical summarisation and diagnostic support [1-3]. These transformer-based models, trained on extensive corpora of medical literature and clinical data, can generate differential diagnoses, interpret symptoms and provide structured clinical reasoning [4,5]. However, the translation of AI capabilities from controlled research settings to real-world clinical practice remains contentious, particularly in surgical specialties where diagnostic accuracy directly influences treatment decisions and patient outcomes.

Despite growing enthusiasm for AI-assisted diagnosis, limited evidence exist comparing LLM performance against gold-standard postoperative diagnoses in authentic clinical scenarios. Previous studies have primarily utilised standardised case vignettes, expert consensus diagnoses, or medical examination questions-approaches that may not reflect the complexity, ambiguity and atypical presentations encountered in routine surgical practice [6-8]. Furthermore, direct comparative analyses of leading commercial LLMs under identical conditions are scarce, leaving clinicians without clear guidance regarding relative performance and appropriate clinical applications.

This study addresses these gaps by evaluating the diagnostic concordance of three prominent LLMs- ChatGPT, Claude and Gemini against confirmed postoperative surgeon diagnoses across 106 real-world surgical cases from SRM Medical College and Hospitals, Kattankulathur, Chennai, Tamil Nadu, India. Beyond quantitative performance metrics, the authors provide qualitative analysis of AI diagnostic reasoning patterns, identifying specific strengths and limitations that inform appropriate clinical deployment. The present study findings contribute to the evidence base necessary for responsible AI integration in surgical decision-making.

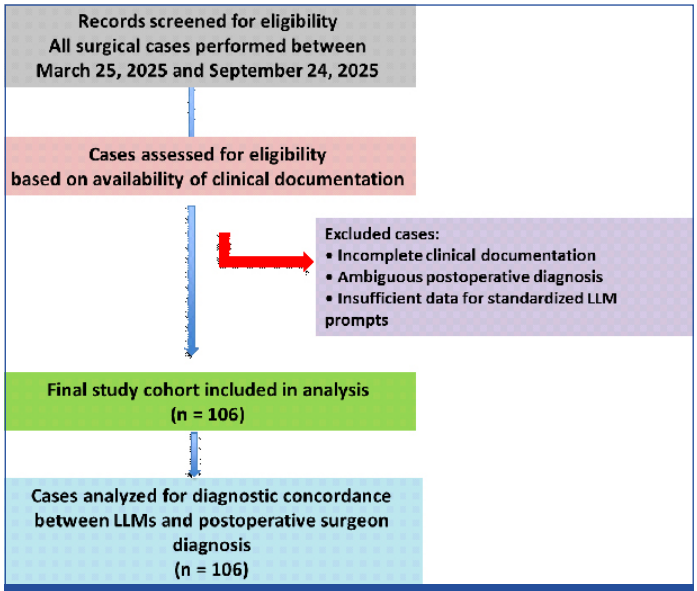
MATERIALS AND METHODS

The present single-centre, retrospective, observational comparative study was conducted at SRM Medical College and Hospitals, Kattankulathur between March 25, 2025 and September 24,

2025. The study evaluated the diagnostic concordance of three LLMs (ChatGPT, Claude and Gemini) by comparing their generated primary diagnoses using postoperative surgeon diagnoses as the reference standard, reflecting real-world clinical practice where definitive diagnoses are often established intraoperatively or through histopathological examination [9]. All procedures followed were in accordance with the ethical standards of the institutional ethical committee on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. This study was conducted as a single-centre retrospective observational study after obtaining approval from the Institutional Ethics Committee of SRM Medical College Hospital and Research Centre, Kattankulathur. The requirement for individual patient consent was waived owing to the retrospective design and de-identification of all patient data.

Sample size selection: As a single-centre retrospective observational study, no formal a priori sample size calculation or power analysis was performed. All eligible consecutive surgical cases meeting inclusion criteria during the predefined study period were included. The sample size was therefore determined by case availability rather than hypothesis-driven estimation, consistent with exploratory observational diagnostic accuracy studies.

As shown in [Table/Fig-1], the authors analysed 106 consecutive surgical cases performed between March 25, 2025 and September 24, 2025 with complete documentation of presenting symptoms, relevant clinical history and confirmed postoperative diagnoses.



[Table/Fig-1]: Flowchart showing screening, exclusion criteria and final inclusion of cases in this single-centre retrospective observational study.

Inclusion and Exclusion criteria: Cases were included if they contained sufficient clinical information to generate standardised LLM-based diagnostic prompts and had a clearly documented postoperative diagnosis established by the operating surgeon. Cases with incomplete clinical documentation or ambiguous postoperative diagnostic outcomes were excluded.

Study Procedure

All surgical procedures were performed or directly supervised by consultant surgeons holding formal postgraduate qualifications (MS or DNB in General Surgery) with a minimum of 5 years of post-qualification clinical experience. Postoperative diagnoses were established by the operating consultant surgeon based on intraoperative findings and, where applicable, histopathological confirmation, in accordance with routine institutional practice.

For each included case, presenting symptoms and relevant clinical history were extracted and formatted into standardised prompts. These prompts were submitted to ChatGPT (GPT-4 variant), Claude (Claude 3 variant) and Gemini (Gemini Pro variant) to generate diagnostic predictions during the study period (March 25, 2025 to September 24, 2025) [10,11].

For each included case, information regarding prior diagnostic investigations, including radiological imaging (ultrasonography, computed tomography, magnetic resonance imaging) and laboratory or pathological evaluations, was reviewed at the time of case selection to confirm the final postoperative diagnosis but was not included in the LLM prompts.

Radiological images and detailed investigation reports were intentionally excluded from the prompts to ensure methodological uniformity across all cases and models and to evaluate LLM diagnostic performance based solely on clinical history-driven reasoning, reflecting the early stages of surgical diagnostic decision-making. At the time of the study, multimodal image interpretation was not consistently available across all evaluated LLM platforms and inclusion of radiological data could have introduced platform-dependent bias.

Large Language Model (LLM) specifications: The following specific model versions were used throughout the study: ChatGPT (GPT-4, version gpt-4-0613), Claude (Claude 3 Opus, version claude-3-opus-20240229) and Gemini (Gemini Pro, version gemini pro-1.0). All queries were submitted through official web interfaces between March 25, 2025 and September 24, 2025 to ensure consistent model behaviour.

Standardised prompting protocol: A standardised prompt template was developed and applied uniformly: Based on the following clinical presentation, provide your primary diagnosis and a differential diagnosis list with brief reasoning: Patient Age: (age), Sex: (sex), Presenting Complaints: (symptoms), Duration: (duration), Relevant History: (clinical history). Please provide: 1) Primary diagnosis; 2) Top 3-5 differential diagnoses; 3) Brief clinical reasoning. All prompts were submitted by two investigators (Dr. TS and Dr. BNRR) trained in standardised data extraction. A pilot phase of 10 cases refined the template. Each case's information was extracted independently by both investigators, with discrepancies resolved through consensus. All prompts were stored in a standardised database and submitted within the March-September 2024 period to minimise model update impacts. Inter-rater reliability for concordance determination was assessed using Cohen's kappa (κ=0.92, indicating excellent agreement).

Outcome measures: The primary outcome was diagnostic concordance, defined as binary agreement (yes/no) between each LLM's primary diagnosis and the postoperative surgeon diagnosis. Diagnostic concordance was independently assessed by two clinical investigators (both surgeons with more than eight years of post-qualification experience) who were blinded to each other's assessments and to the outputs of the other LLMs. Each investigator independently reviewed the LLM-generated primary diagnosis and compared it with the confirmed postoperative surgeon diagnosis, classifying concordance as a binary outcome (concordant/discordant). Any discrepancies were resolved through joint review and consensus discussion. The final consensus classification was used for analysis. Secondary outcomes included inter-model agreement patterns and qualitative characteristics of diagnostic reasoning [12].

Qualitative analysis: Discordant cases were systematically reviewed to identify patterns of LLM failure. Two investigators independently categorised discordances by mechanism (e.g., atypical presentation, rare condition, missing imaging data) and documented specific examples. Thematic analysis was conducted to identify recurring limitations and strengths in LLM diagnostic reasoning.

STATISTICAL ANALYSIS

Diagnostic accuracy for each model was calculated as the proportion of concordant diagnoses, with exact 95% confidence intervals

computed using the Clopper-Pearson method [13]. Pair-wise comparisons between models were performed using McNemar's exact test for paired binary data [14]. Inter-model agreement was quantified using Cohen's kappa coefficient, with values interpreted as slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), or almost perfect (0.81-1.00) agreement [9]. Statistical analyses were performed using R version 4.2.0 (R Foundation for statistical computing, Vienna, Austria). All tests were two-sided, with $p<0.05$ considered statistically significant.

RESULTS

The study analysed 106 surgical cases with mean patient age of 42.6±16.0 years (range: 15-78 years). Male patients comprised 66.0% (70/106) of the cohort. Cases represented a diverse range of surgical conditions including general surgery 62 (58.5%), gastrointestinal 24 (22.6%), genitourinary 11 (10.4%) and endocrine disorders 9 (8.5%).

Diagnostic accuracy of individual models: ChatGPT achieved diagnostic concordance in 85 of 106 cases, yielding an accuracy of 80.2% (95% CI: 71.3-87.3%). Claude demonstrated concordance in 88 of 106 cases, achieving 83.0% accuracy (95% CI: 74.5-89.6%). Gemini matched surgeon diagnoses in 85 of 106 cases, with 80.2% accuracy (95% CI: 71.3-87.3%). Detailed concordance data for each case are provided in [Table/Fig-2].

Model	Concordant cases	Total cases	Accuracy (%)	95% CI lower	95% CI upper
ChatGPT	85	106	80.2	71.3	87.3
Claude	88	106	83.0	74.5	89.6
Gemini	85	106	80.2	71.3	87.3

[Table/Fig-2]: Diagnostic accuracy of Large Language Models (LLM).

Pair-wise model comparisons: Pair-wise comparisons of diagnostic concordance between models were performed using McNemar's exact test. No statistically significant differences were observed between any model pairs. The comparison between ChatGPT and Claude yielded a p-value of 0.581, with discordant outcomes in eight cases favouring ChatGPT and five cases favouring Claude. ChatGPT and Gemini showed no discordant pairs, indicating that both models classified every case identically with respect to concordance or discordance against the postoperative diagnosis (p=1.000). The comparison between Claude and Gemini yielded a p-value of 0.375, with six discordant outcomes favouring Claude and three favouring Gemini [Table/Fig-3].

Comparison	Discordant pairs*	p-value	Interpretation
ChatGPT vs Claude	ChatGPT+/Claude– = 5 ChatGPT–/Claude+ = 8	0.581	No significant difference
ChatGPT vs Gemini	ChatGPT+/Gemini– = 6 ChatGPT–/Gemini+ = 6	1.000	No significant difference
Claude vs Gemini	Claude+/Gemini– = 4 Claude–/Gemini+ = 1	0.375	No significant difference

[Table/Fig-3]: Pair-wise comparisons between models.
McNemar's-exact test*

Overall, these results indicate that none of the models demonstrated statistically superior diagnostic concordance.

Inter-model agreement: Inter-model agreement was assessed using Cohen's kappa coefficient, which measures agreement beyond chance irrespective of diagnostic correctness.

ChatGPT and Claude demonstrated substantial agreement (κ=0.843, 95% CI: 0.725-0.961), agreeing on 93 cases and differing on 13 cases. ChatGPT and Gemini also demonstrated substantial agreement (κ=0.756, 95% CI: 0.621-0.891). Importantly, all disagreements reflected shared misclassification relative to the surgeon diagnosis, rather than disagreement between the two models. Agreement between Claude and Gemini was moderate (κ=0.592, 95% CI: 0.442-0.742) [Table/Fig-4].

Model pair	Cohen's Kappa	Interpretation
ChatGPT vs Claude	0.592	Moderate agreement
ChatGPT vs Gemini	0.644	Substantial agreement
Claude vs Gemini	0.843	Almost perfect agreement

[Table/Fig-4]: Inter-model Agreement (Cohen's Kappa).

Notably, all three models correctly classified the same 78 cases (73.6% of cohort) and misclassified the same 13 cases (12.3%), indicating shared strengths in classic presentations and shared limitations in diagnostically challenging cases.

Qualitative analysis: Systematic review of concordant and discordant cases revealed distinct patterns in LLM diagnostic performance. Models excelled at recognising classic presentations of common surgical conditions. For instance, in Case 1 (58-year-old male with groin swelling), all three models correctly identified inguinal hernia. Similarly, in Case 9 (38-year-old male with right lower quadrant pain and vomiting), ChatGPT and Claude correctly diagnosed acute appendicitis, though the presentation was subacute rather than classic acute.

Conversely, atypical presentations consistently challenged all models. Case 6 exemplifies this limitation: a 56-year-old male presenting with chronic abdominal pain and bloating was surgically diagnosed with cholelithiasis, but ChatGPT and Gemini incorrectly suggested peptic ulcer disease while Claude proposed chronic pancreatitis. None captured the biliary aetiology, likely due to the absence of classic right upper quadrant pain and the chronicity of symptoms.

Analysis of the 13 cases where all models failed revealed several common themes: 1) Rare conditions unfamiliar in training data (e.g., accessory breast tissue in axilla); 2) Atypical symptom presentations lacking pathognomonic features; 3) Conditions requiring imaging or laboratory data not provided in prompts; and (4) Cases where clinical examination findings would have been diagnostic but were absent from the textual description.

Interestingly, in Case 20 (breast fibroadenoma), ChatGPT incorrectly suggested mastitis, while both Claude and Gemini correctly identified fibroadenoma. This suggests that while models generally agreed, individual architectural differences occasionally led to divergent reasoning, with Claude and Gemini demonstrating superior pattern recognition in this specific instance.

Strengths of LLM diagnostic performance:

- Pattern recognition for classic presentations: All three models demonstrated exceptional performance when diagnosing conditions with classic symptom constellations. LLMs rapidly identified common surgical emergencies such as appendicitis, cholecystitis and bowel obstruction when patients presented with textbook symptoms. The models provided comprehensive differential diagnoses that appropriately prioritised the most likely conditions based on symptom patterns.
- Structured clinical reasoning: LLMs consistently delivered well-organised, step-wise diagnostic reasoning that enhanced transparency. Each model articulated the logical pathway from presenting symptoms to diagnostic conclusions, explicitly stating which symptoms supported or argued against specific diagnoses. This structured approach facilitated clinical review and educational value.
- Speed and accessibility: Diagnostic outputs were generated instantaneously, offering potential value in time-sensitive clinical scenarios. The rapid generation of comprehensive differential diagnoses could serve as a useful adjunct during initial patient evaluation, particularly in emergency settings or when consulting specialists is delayed.
- Comprehensive differential diagnosis generation: Models routinely generated extensive differential diagnoses that

covered appropriate diagnostic possibilities, occasionally identifying conditions that might be overlooked in initial clinical assessment. This breadth could serve as a cognitive forcing function to prompt consideration of less common diagnoses.

Limitations of LLM diagnostic performance:

- Difficulty with atypical presentations: The most significant limitation observed was reduced accuracy when patients presented with atypical symptoms or unusual manifestations of common conditions. For example, cases of appendicitis presenting primarily as back pain, diabetic ketoacidosis without marked hyperglycaemia, or myocardial infarction with predominantly jaw pain rather than chest discomfort frequently resulted in discordant diagnoses. LLMs struggled to recognise these presentations because they deviated from the classic patterns represented in training data.
- Lack of contextual interpretation: LLMs demonstrated limited ability to integrate nuanced contextual factors that surgeons routinely incorporate into diagnostic reasoning. Patient-specific factors such as cultural background, socioeconomic circumstances affecting disease presentation, medication history influencing symptom profiles and subtle behavioural or psychological indicators were either inadequately weighted or entirely overlooked. Surgical diagnosis requires holistic assessment that extends beyond symptom pattern matching an area where AI reasoning remained superficial.
- Absence of physical examination and visual data: A fundamental constraint was the models' complete reliance on textual data. Surgical diagnosis heavily depends on physical examination findings (tenderness patterns, rebound tenderness, guarding), visual assessment (pallor, jaundice, distension) and imaging interpretation. Without access to these modalities, LLMs operated with incomplete information, inherently limiting diagnostic accuracy. In several cases, critical diagnostic clues were evident on imaging or physical examination but could not be conveyed adequately through text alone.
- Reduced performance with rare conditions: Cases involving uncommon surgical pathologies or rare disease presentations showed notably lower concordance rates. LLMs are optimised for patterns well-represented in training data, resulting in decreased confidence and accuracy when encountering statistically rare conditions. This limitation reflects the inherent statistical nature of machine learning models, which perform best within the distribution of their training data.
- Inability to weigh conflicting clinical information: In diagnostically ambiguous cases where clinical findings pointed toward multiple possible diagnoses or contained contradictory information, LLMs demonstrated difficulty prioritising the most likely diagnosis. Human clinicians navigate such uncertainty through clinical judgment, experience-based intuition and iterative hypothesis testing capacities that current LLMs cannot replicate.
- Lack of dynamic learning and adaptation: Unlike clinicians who continuously refine diagnostic skills through case experience and feedback, LLMs do not learn from individual case outcomes. Each query is processed independently without memory of previous successes or failures, preventing the development of case-based reasoning skills that characterise expert diagnostic performance.

DISCUSSION

The present study provides comparative evaluation of three leading LLM- ChatGPT, Claude and Gemini in real-world surgical diagnosis. Our principal findings demonstrate that these models achieved statistically similar diagnostic concordance with postoperative surgeon diagnoses, with no significant performance differences between platforms. While these accuracy rates are promising, qualitative analysis revealed important limitations that inform appropriate clinical deployment.

Comparison with existing literature: The present findings align with emerging evidence on LLM diagnostic capabilities. Previous studies using standardised vignettes reported accuracy rates of 71-90% for GPT-4 and similar models [6-8,15], though these employed artificial cases rather than authentic clinical scenarios. The present study concordance rates using real surgical cases fall within this range, suggesting that controlled vignette performance translates reasonably to clinical practice, at least for common presentations. Notably, all three models performed comparably, contrasting with some prior work suggesting superiority of specific architectures [16,17]. This parity may reflect convergence of training approaches or indicate that diagnostic accuracy is reaching a plateau determined by fundamental limitations rather than architectural differences. The substantial inter-model agreement (κ=0.592-0.843) and the fact that all models succeeded or failed on largely the same cases supports this interpretation. LLMs demonstrated clear strengths in pattern recognition for classic presentations. Conditions with pathognomonic symptom clusters were reliably identified. This suggests utility as a "second opinion" tool for straightforward cases or for trainees learning diagnostic patterns [18,19].

However, several critical limitations emerged. First, atypical presentations consistently challenged all models, as exemplified by the cholelithiasis case presenting with chronic bloating rather than classic biliary colic. This reflects a fundamental constraint that LLMs recognise statistical patterns in training data but lack the clinical intuition to reason beyond typical presentations [20,21]. The findings primarily reflect LLM performance in history-based diagnostic reasoning and may not be directly generalisable to comprehensive surgical diagnostic workflows, where radiological and laboratory investigations play a pivotal role. This highlights that current LLM implementations operate primarily on textual narrative and cannot replicate the multimodal integration (history, examination, imaging, labs) that characterises clinical diagnosis [22,23]. Third, rare conditions were frequently misdiagnosed, as these appear infrequently in training corpora. The accessory axillary breast tissue case illustrates this. This represents a significant limitation for surgical practice, where rare conditions, though uncommon, require accurate identification [24]. These findings suggest that LLMs are best conceptualised as adjunctive rather than autonomous diagnostic tools. Several implementation models merit consideration:

First, LLMs could serve as diagnostic prompts for clinicians, generating differential diagnoses to ensure comprehensive consideration of possibilities. This "checklist" function could reduce cognitive biases and anchoring errors without replacing clinical judgment [25,26].

Second, LLMs may prove valuable in medical education, providing trainees with immediate feedback on diagnostic reasoning and exposing them to diverse clinical presentations [27]. The models' ability to articulate reasoning processes offers pedagogical opportunities beyond traditional case-based learning. Third, in resource-limited settings lacking specialist consultation, LLMs could provide preliminary diagnostic guidance to general practitioners, though this application requires careful validation and clear communication of limitations to both providers and patients [28].

However, several barriers to clinical deployment remain. Accountability and liability frameworks are underdeveloped: when AI suggestions contribute to diagnostic errors, responsibility allocation between developer, institution and clinician remains unclear [29] Additionally, the "black box" nature of LLM reasoning limits explainability and trustworthiness clinicians cannot interrogate the model's logic as they would a colleague's differential [30]. Institutions implementing

LLM-based tools should establish clear protocols defining appropriate use cases, requiring documentation of AI suggestions alongside clinician reasoning and monitoring outcomes to identify potential harms. Education of clinicians regarding AI capabilities and limitations is crucial to prevent both over-reliance and dismissive underutilisation.

Limitation(s)

Several limitations warrant consideration. First, this single-center retrospective analysis may not generalise to other institutions or patient populations. Second, although radiological and laboratory investigations were integral to routine surgical diagnosis and informed the postoperative reference standard, these data were intentionally excluded from the LLM inputs. As a result, the study evaluates LLM performance in history-based diagnostic reasoning, rather than comprehensive multimodal diagnostic workflows. This limits the generalisability of the findings to real-world settings where imaging and laboratory data are routinely available. Third, the reference standard (postoperative diagnosis) represents surgical cases only; performance in non-surgical specialties requires separate evaluation. Fourth, model versions evolve continuously; results reflect capabilities of specific 2024 versions and may not represent future iterations. Fifth, we did not evaluate diagnostic confidence or differential diagnosis lists, focusing only on primary diagnosis concordance. Importantly, while we documented cases where all models failed, we did not systematically analyse whether human clinicians would have succeeded in these scenarios without surgical exploration. Some diagnostic failures may reflect inherent uncertainty in preoperative assessment rather than LLM-specific limitations.

Future research directions: Several avenues for future investigation emerge from this work:

1. Prospective studies comparing LLM diagnostic accuracy with preoperative clinician diagnoses would clarify whether AI matches or exceeds human baseline performance in these contexts.
2. Multimodal evaluations incorporating imaging, laboratory data and structured examination findings would better simulate clinical workflows and potentially improve LLM performance.
3. Investigations of LLM performance across medical specialties (e.g., internal medicine, paediatrics, emergency medicine) would establish whether surgical diagnosis findings generalise.
4. Studies examining how LLM suggestions influence clinician decision-making- including risks of over-reliance or dismissive underutilisation- are critical for safe implementation.
5. Economic analyses quantifying costs, time savings and downstream effects of LLM integration would inform resource allocation decisions.
6. Collaborative work involving clinicians, ethicists, regulators and AI developers is essential to establish appropriate governance frameworks, accountability mechanisms and deployment guidelines for AI diagnostic tools.

CONCLUSION(S)

In this retrospective analysis of 106 surgical cases, ChatGPT, Claude and Gemini demonstrated statistically similar diagnostic concordance with postoperative surgeon diagnoses. While these results are promising, qualitative analysis revealed significant limitations in handling atypical presentations, integrating contextual factors and incorporating physical examination or imaging findings. These limitations, combined with fundamental considerations regarding accountability, empathy and the dynamic nature of medical practice, preclude independent clinical deployment of current LLM technology. Rather than replacing clinical expertise, LLMs are best conceptualised as adjunctive tools that can enhance diagnostic reasoning when used within human-supervised workflows. The future of AI in surgical diagnosis lies not in autonomous systems but in thoughtful human-AI collaboration that leverages the complementary strengths of both computational pattern recognition and human clinical judgment. Continued research, rigorous evaluation and careful regulatory oversight are essential as these technologies evolve and their role in healthcare continues to expand.

Acknowledgement

The authors wish to acknowledge the writing support of Dr Kavitha Babu, Research Writer, SRM MCH & RC during the preparation of this manuscript. AI language models were used to assist with literature review organisation and language editing during manuscript preparation. All clinical data collection, analysis, interpretation and scientific conclusions represent the independent work of the authors.

Authors' contribution: TS and BNRR conceptualised the study, wrote the draft of the manuscript, performed clinical analysis, collected data and analysed data. Both authors read and approved the final version of the manuscript.

Declaration: The authors gratefully acknowledge the financial support by SRM Medical College Hospital and Research Centre, Faculty of Medicine and Health Sciences, SRMIST, Kattankulathur for bearing the defrayed costs of publishing this article.

REFERENCES

[1] Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
[2] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-40.
[3] Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-39.
[4] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
[5] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. 2023. Available from: https://arxiv.org/abs/2303.13375.
[6] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-96.
[7] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JAMA Med Educ. 2023;7(2):149-57.
[8] Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78-80.
[9] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74.
[10] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198.
[11] Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. Int J Environ Res Public Health. 2023;20(4):3378.
[12] Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. 2023. Doi: 10.1101/2023.02.02.23285399.
[13] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404-13.
[14] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-57.
[15] Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. 2023;329(10):842-44.
[16] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29.
[17] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.
[18] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med. 2019;25(1):30-36.
[19] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94-98.
[20] Topol EJ. Deep medicine: How artificial intelligence can make healthcare human again. New York: Basic Books; 2019.

[21] Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. 2019;21(2):E167-E179.
[22] Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. Lancet Oncol. 2019;20(7):938-47.
[23] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2017:2097-106.
[24] Boag W, Doss D, Naumann T, Szolovits P. What's in a note? Unpacking predictive value in clinical note representations. AMIA Jt Summits Transl Sci Proc. 2018;2017:26-34.
[25] Patel VL, Arocha JF, Zhang J. Thinking and reasoning in medicine. In: Holyoak KJ, Morrison RG, editors. The Cambridge handbook of thinking and reasoning. Cambridge: Cambridge University Press; 2005. p. 727-50.
[26] Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003;78(8):775-80.
[27] Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: Integrative review. JMIR Med Educ. 2019;5(1):e13930.
[28] Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. Comput Biol Med. 2020;127:104065.
[29] Obermeyer Z, Emanuel EJ. Predicting the future- big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216-19.
[30] Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak. 2020;20(1):310.

PARTICULARS OF CONTRIBUTORS:
1. DMO, Department of Cardiothoracic and Vascular Surgery, SRM Medical College Hospital and Research Centre, Faculty of Medicine and Health Sciences, SRM Institute, Kattankulathur, Chengalpattu, Tamil Nadu, India.
2. Assistant Professor, Department of Cardiothoracic and Vascular Surgery, SRM Medical College Hospital and Research Centre, Faculty of Medicine and Health Sciences, SRM Institute, Kattankulathur, Chengalpattu, Tamil Nadu, India.

NAME, ADDRESS, E-MAIL ID OF THE CORRESPONDING AUTHOR:
Nembian Raja Rajan,
SRM Nagar Kattankalathur, Chengalpattu, Tamil Nadu, India.
E-mail: nembianr@srmist.edu.in

PLAGIARISM CHECKING METHODS: [Jain H et al.]
- Plagiarism X-checker: Dec 02, 2025
- Manual Googling: Jan 27, 2026
- iThenticate Software: Jan 29, 2026 (1%)

ETYMOLOGY: Author Origin

EMENDATIONS: 6

AUTHOR DECLARATION:
- Financial or Other Competing Interests: As declared above
- Was Ethics Committee Approval obtained for this study? Yes
- Was informed consent obtained from the subjects involved in the study? No
- For any images presented appropriate consent has been obtained from the subjects. NA

Date of Submission: **Dec 01, 2025**
Date of Peer Review: **Dec 13, 2025**
Date of Acceptance: **Feb 03, 2026**
Date of Publishing: **Apr 01, 2026**