AI Report

We are highly confident this text is

AI Generated

AI Probability

99%

This number is the probability that the document is AI generated, not a percentage of AI text in the document.

Plagiarism

Plagiarism scan was not run for this document.

manuscript - Jurnal Dwi.docx - 12/17/2025

-

High Consistency, Limited Accuracy: Evaluating Large Language Models for Binary Medical Diagnosis

Authors: Dwi Anggriani1, Syaiful Bachri Mustamin1, Muhammad Atnang1, Kartini Aprilia Pratiwi Nuzry1Affiliation: 1Department of Information Technology, Institut Sains Teknologi dan Kesehatan 'Aisyiyah Kendari, Kendari, IndonesiaCorresponding Author: Syaiful Bachri Mustamin, syaifulbachri@mail.ugm.ac.idPreprint: medRxiv DOI: 10.64898/2025.12.08.25341823

Abstract

Background: Large Language Models (LLMs) have demonstrated impressive capabilities in medical knowledge tasks, yet their reliability and consistency in clinical diagnosis remain incompletely characterized.

Objective: To systematically evaluate the consistency and accuracy of state-of-the-art LLMs in binary medical diagnosis, examining the relationship between reproducibility and diagnostic performance.

Methods: We evaluated three LLMs (GPT-4o, Gemini-2.0-Flash, Qwen-Plus) on heart disease diagnosis using 100 diverse clinical cases.
Each model performed 4 independent assessments per case (1,200 total predictions).
We tested two prompt variations and measured intra-model consistency, inter-model agreement, diagnostic accuracy, and prompt sensitivity.

Results: All models achieved exceptional intra-model consistency (99-100%), with Qwen demonstrating perfect reproducibility.
Inter-model agreement was similarly high (98-99%).
However, diagnostic accuracy remained at approximately 50%, equivalent to random guessing.
Models exhibited a strong bias toward positive diagnosis (49-51 false positives vs 0-1 false negatives per 100 cases).
Prompt variation had minimal impact (<3% change), and error patterns were highly systematic, with all models making identical errors on 48-51% of cases.

Conclusions: Our findings reveal a critical dissociation between consistency and accuracy in LLM medical diagnosis. While LLMs demonstrate remarkable reproducibility, their tendency toward overdiagnosis and limited discriminative accuracy constrain direct clinical utility.
Results suggest LLMs may be better suited as supplementary decision-support tools rather than primary diagnostic systems.

Keywords: Large Language Models, Medical Diagnosis, Consistency Analysis, Heart Disease, Clinical Decision Support, AI Reliability, Reproducibility

Introduction

Large Language Models (LLMs) have emerged as promising tools for clinical applications, demonstrating impressive performance on medical licensing examinations and case analysis [1-3].
However, their deployment in clinical settings raises critical questions about reliability and consistency.
While traditional diagnostic tools are expected to yield reproducible results, LLMs employ stochastic generation that can lead to varying outputs [4], with documented tendencies toward hallucinations and inconsistent reasoning [9,10].
Reproducibility challenges in AI systems have been extensively documented [13], yet remain incompletely characterized for medical LLMs.

Despite growing literature on LLM performance in medical question-answering [5,6], few studies have systematically examined the relationship between consistency and accuracy in diagnostic tasks.
Most evaluations focus on single-run accuracy without assessing reproducibility, and the influence of prompt engineering remains incompletely characterized [7,11].

This study addresses these gaps through a comprehensive evaluation of three state-of-the-art LLMs on binary heart disease diagnosis.
We aimed to: (1) quantify intra-model consistency across repeated runs, (2) evaluate inter-model agreement, (3) measure diagnostic accuracy relative to consistency, (4) assess prompt sensitivity, and (5) analyze error patterns to determine if mistakes are random or systematic.

Methods

Dataset and Study Design

We utilized the UCI Heart Disease dataset [8], containing 303 patients with 13 clinical parameters: demographics (age, sex), symptoms (chest pain type), vital signs (resting blood pressure), laboratory values (cholesterol, fasting blood sugar), electrocardiography (resting ECG), exercise testing (maximum heart rate, exercise-induced angina, ST depression, ST segment slope), imaging (fluoroscopy vessel count), and thalassemia test results.
Binary outcomes indicated the presence or absence of significant coronary stenosis.

To ensure diverse representation, we performed k-means clustering (k=2) and selected 50 cases from each cluster, yielding 100 test cases with balanced disease prevalence (51% positive, 49% negative).

Models and Experimental Protocol

We evaluated three LLMs: GPT-4o (OpenAI), Gemini-2.0-Flash (Google), and Qwen-Plus (Alibaba), accessed via APIs with temperature=0.7.
Each model performed 4 independent assessments per case, yielding 1,200 total predictions.

We tested two prompt variations: - Prompt A ("Expert"): "You are Dr. CardioExpert, a highly experienced cardiologist..."
- Prompt B ("Neutral"): "You are a medical AI assistant trained to provide accurate and balanced diagnostic assessments..."

Both prompts provided identical clinical data and parameter definitions, requesting a binary diagnosis (Yes/No) with a 2-3 sentence justification.

We implemented a SQLite-based checkpoint system enabling immediate data saving, automatic duplicate prevention, and experiment resumption capability.

Outcome Measures

Primary outcomes: 1.
Intra-model consistency: Proportion of runs with majority agreement per case 2.
Diagnostic accuracy: Using majority voting (>=2/4 runs), we calculated accuracy, sensitivity, specificity, precision, and F1-score 3.
Inter-model agreement: Pairwise agreement and Cohen's kappa between models

Secondary outcomes: 4.
Prompt sensitivity: Proportion of cases with identical predictions across prompts 5.
Error patterns: Classification as all-correct, all-wrong, or mixed outcomes

Statistical analyses used Python with pandas, scikit-learn, and scipy.
Significance was set at $p<0.05$.

Results

Intra-Model Consistency: Exceptional Reproducibility

All models demonstrated remarkably high consistency (Table 1, Figure 1).
Qwen-Plus achieved perfect consistency (100%) with the expert prompt, never varying across 4 independent runs.
GPT-4o and Gemini-2.0-Flash showed 99.0-99.5% average consistency.
Notably, 96-100% of cases achieved perfect agreement (4/4 identical predictions), and minimum consistency never fell below 50%.

Table 1.
Intra-Model Consistency

Model	Prompt	Avg Consistency	Min	Perfect (%)
GPT	Expert	99.25%		

50%

98%

GPT

Neutral

99.00%

75%

96%

Gemini

Expert

99.50%

75%

98%

Gemini

Neutral

99.25%

50%

98%

Qwen

Expert

100.00%

100%

100%

Qwen

Neutral

99.75%

75%

99%

Inter-Model Agreement: High Consensus

Models showed 98-100% pairwise agreement, indicating remarkably similar reasoning patterns (Table 2).
Three-way agreement (all models concur) occurred in 98-99% of cases.
Cohen's kappa values near zero reflected extreme class imbalance (nearly all positive predictions) rather than lack of agreement.

Table 2.
Inter-Model Agreement

Prompt	GPT-Gemini	GPT-Qwen	Gemini-Qwen	All 3 Agree
Expert	98.0%	100.0%	98.0%	98%
Neutral	100.0%	99.0%	99.0%	99%

Diagnostic Accuracy: Limited Despite High Consistency

Diagnostic accuracy approximated random guessing (48-51%) despite 99-100% consistency (Table 3).
Models achieved perfect or near-perfect recall (98-100%) but extremely poor specificity (~0-2%), generating 49-51 false positives versus 0-1 false negatives.
This created a consistency-accuracy gap of approximately 50 percentage points.

Table 3.
Diagnostic Performance

Model	Prompt	Accuracy	Precision	Recall	F1	FP	FN
GPT	Expert	51.0%	51.0%	100%	67.6%	49	0
GPT	Neutral	49.0%	49.0%				

100%

65.8%

51

0

Gemini

Expert

51.0%

51.0%

98%

67.1%

48

1

Gemini

Neutral

49.0%

49.0%

100%

65.8%

51

0

Qwen

Expert

51.0%

51.0%

				100%	67.6%	49	0
Qwen	Neutral	48.0%	48.5%	98%	64.9%	51	1

Representative confusion matrices (Figure 2) showed models predicted "disease present" for nearly all cases, with true negatives 0.

Prompt Sensitivity: Minimal Impact

Changing from expert to neutral prompt had a minimal effect (Table 4, Figure 3).
GPT-4o showed zero sensitivity (100% identical predictions), while Gemini and Qwen changed only 1-3 predictions (1-3% of cases).
This suggests diagnostic behavior is deeply encoded rather than easily modifiable through prompting.

Table 4.
Prompt Robustness

Model	Agreement	Changes
GPT	100%	0/100

Gemini	98%	2/100
Qwen	99%	1/100

Error Pattern Analysis: Systematic, Not Random

Errors were highly systematic rather than random (Table 5).
In 98-99% of cases, all three models either succeeded together or failed together.
Only 1-2% showed model disagreement, indicating shared reasoning patterns or biases.

Table 5.
Error Consistency

Pattern	Expert	Neutral
All correct	50%	48%
All wrong	48%	51%
Mixed	2%	1%

Qualitative analysis revealed models consistently cited elevated cholesterol, abnormal ECG findings, or exercise abnormalities as disease evidence, even when ground truth indicated absence of significant stenosis, suggesting risk factor conflation with diagnostic criteria.

Discussion

Principal Findings

This study demonstrates a critical dissociation between consistency and accuracy in LLM medical diagnosis: exceptional reproducibility (99-100%) coexists with chance-level accuracy (~50%).
This 50-percentage-point gap represents a fundamental challenge for clinical deployment.

The Consistency-Accuracy Paradox

High consistency indicates LLMs reliably apply learned reasoning patterns they are systematically biased rather than randomly erring.
This "consistent wrongness" is arguably more concerning than random errors, suggesting fundamental limitations in medical reasoning capabilities [10] rather than simple uncertainty.

Several mechanisms may explain this paradox:

Medical Conservatism Bias: LLMs trained on medical text may have learned that missing disease (false negative) carries greater consequences than over-diagnosis (false positive), encoding a "better safe than sorry" heuristic consistently applied.

Risk Factor Conflation: Qualitative analysis suggests models conflate risk factors with diagnostic findings.
Elevated cholesterol increases disease risk but doesn't constitute diagnostic evidence of current stenosis.
LLMs may struggle to distinguish "high-risk patient" from "disease-positive patient."

Lack of Discriminative Training: Unlike supervised models trained explicitly on diagnostic labels, LLMs learn from general medical text, emphasizing disease description more than differential diagnosis, leaving them poorly calibrated for binary classification [12].
This calibration deficit manifests as systematic over-prediction despite high confidence.

Prompt Insensitivity: Deep-Rooted Behavior

The minimal prompt impact (GPT: 0%, others: 1-2%) was unexpected.
Refraining from "expert cardiologist" to "neutral assessor" should have reduced conservatism, yet predictions remained nearly identical.
This suggests diagnostic behavior is deeply encoded in model weights rather than modifiable through surface-level prompting [11], with important implications for prompt engineering strategies.
While prompt patterns can enhance certain LLM behaviors, our findings indicate diagnostic reasoning may be resistant to prompt-level interventions [15].

Inter-Model Agreement: Shared Limitations

The 98-99% inter-model agreement despite different architectures and training procedures suggests observed limitations reflect fundamental challenges in applying LLMs to medical diagnosis rather than model-specific artifacts.
Possible explanations include similar training data sources, convergent learning of medical conservatism, shared limitations in processing structured numerical data, and common challenges in threshold-based classification.

Clinical Implications

Current LLMs are not ready for primary diagnostic applications requiring binary classification.

The ~50% accuracy is unacceptable clinically and could lead to harmful over-diagnosis, unnecessary testing, and patient anxiety [14].
In a 50% prevalence scenario, deploying these models would result in 98-100% of disease cases correctly identified but only 0-2% of healthy cases correctly identified, causing approximately 50% unnecessary downstream testing.
These findings underscore the importance of rigorous evaluation before clinical deployment, as ethical considerations demand careful assessment of potential harms [14].

Despite primary diagnosis limitations, LLMs' high consistency and strong negative predictive value suggest potential roles as: (1) second opinion tools where reproducibility builds physician confidence, (2) triage assistants suitable for initial screening where high sensitivity is prioritized, (3) medical education tools providing consistent feedback, and (4) research tools for hypothesis generation.

Technical Implications

Results suggest general-purpose LLMs lack discriminative capabilities for diagnostic classification.
Future development should consider: supervised fine-tuning on labeled diagnostic datasets, reinforcement learning from physician-verified diagnoses, calibration techniques for binary classification thresholds, and hybrid architectures combining LLM reasoning with specialized classifiers.

Limitations

Key limitations include: single condition (heart disease may not generalize), binary classification (real diagnosis often involves multi-class assessment), dataset age (1980s diagnostic criteria may differ from current standards), limited sample size (100 cases), structured input only (missing narrative information), three models tested (limited sampling of LLM landscape), API-only access (preventing internal mechanism analysis), and single temperature setting (0.7).

Future Directions

Important future work includes mechanistic studies examining which parameters LLMs prioritize, improvement strategies testing fine-tuning and ensemble approaches, broader evaluations across diverse diagnostic tasks, comparison with human physicians for baseline performance, and theoretical development of consistency-accuracy frameworks.

Conclusions

This study provides rigorous evidence that LLMs achieve exceptional consistency (99-100%) but limited accuracy (~50%) in binary medical diagnosis.
This consistency-accuracy dissociation represents a fundamental challenge for clinical deployment.
Our findings indicate that high consistency does not guarantee accuracy, diagnostic behavior is resistant to prompt engineering, errors are systematic rather than random, and LLMs show strong positive diagnosis bias.

Current general-purpose LLMs are better suited as supplementary decision support rather than primary diagnostic systems.
Their exceptional reproducibility is clinically valuable, but limited discriminative ability necessitates human oversight.
Future development should prioritize supervised fine-tuning, improved structured data processing, and calibration techniques to address systematic biases.

This work contributes to a nuanced understanding of LLM capabilities and limitations in healthcare, informing responsible development and deployment of AI-assisted clinical decision support systems.

Acknowledgments

We thank Institut Sains Teknologi dan Kesehatan 'Aisyiyah Kendari for institutional support.

We acknowledge OpenAI, Google, and Alibaba Cloud for providing API access to GPT-4o, Gemini-2.0-Flash, and Qwen-Plus, respectively, through standard commercial services.
We thank the UCI Machine Learning Repository and the original data collectors (Janosi, Steinbrunn, Pfisterer, and Detrano) for making the Heart Disease dataset publicly available.

References

Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al.
Large language models in medicine.
Nat Med.
2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8

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):1233-1239. doi:10.1056/NEJMsr2214184

Singhal K, Azizi S, Tu T, et al.
Large language models encode clinical knowledge.
Nature.
2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

Kung TH, Cheatham M, Medenilla A, et al.
Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.
PLOS Digit Health.
2023;2(2):e0000198.
doi:10.1371/journal.pdig.0000198

Nori H, Lee YT, Zhang S, et al.
Can Generalist Foundation Models Outcompete Special-Purpose Tuning?
Case Study in Medicine.
arXiv:2311.16452.
2023.

Wang S, Zhao Z, Ouyang X, et al.
ChatCAD: Interactive Computer-Aided Diagnosis on Medical Images using Large Language Models.
arXiv:2302.07257.
2023.

Moor M, Banerjee O, Abad ZSH, et al.
Foundation models for generalist medical artificial intelligence.
Nature.
2023;616(7956):259-265. doi:10.1038/s41586-023-05881-4

Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease [Dataset].
UCI Machine Learning Repository.
1988. doi:10.24432/C52P4X

Ji Z, Lee N, Frieske R, et al.
Survey of hallucination in natural language generation.
ACM Comput Surv.
2023;55(12):1-38. doi:10.1145/3571730

Lievin V, Hother CE, Motzfeldt AG, Winther O.

Can large language models reason about medical questions?
Patterns.
2024;5(3):100943. doi:10.1016/j.patter.2024.100943

White J, Fu Q, Hays S, et al.
A prompt pattern catalog to enhance prompt engineering with ChatGPT.
arXiv:2302.11382.
2023.

Jiang LY, Liu XC, Nejatbakhsh N, et al.
Health system-scale language models are all-purpose prediction engines.
Nature.
2023;619(7969):357-362. doi:10.1038/s41586-023-06160-y

McDermott MBA, Wang S, Marinsek N, et al.
Reproducibility in machine learning for health research.
Sci Transl Med.
2021;13(586):eabb1655.
doi:10.1126/scitranslmed.abb1655

Chen IY, Pierson E, Rose S, et al.
Ethical machine learning in healthcare.
Annu Rev Biomed Data Sci.
2021;4:123-144. doi:10.1146/annurev-biodatasci-092820-114757

Savage T, Nayak A, Gallo R, et al.
Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine.
NPJ Digit Med.
2024;7(1):20. doi:10.1038/s41746-024-01010-1

High Human Impact ● ● ● ● ● ● High AI Impact