ISSN: 3065-7628

Research Article

Journal of Future Medicine and Healthcare Innovation

The Reliability of LLMs For Medical Diagnosis: An Examination of Consistency, Manipulation, and Contextual Awareness

Krishna Subedi*

Independent Researcher, India

***Corresponding Author**
Krishna Subedi, Independent Researcher, India.

Submitted: 2025, Mar 10; **Accepted:** 2025, Apr 11; **Published:** 2025, Apr 16

Citation: Subedi, K. (2025). The Reliability of LLMs For Medical Diagnosis: An Examination of Consistency, Manipulation, and Contextual Awareness. *J Future Med Healthcare Innovation, 3*(1), 01-34.

Abstract

Universal healthcare access remains a critical unmet need, especially in resource-limited settings. Large Language Models (LLMs) hold immense promise for democratizing healthcare globally, offering sophisticated diagnostic tools even in remote areas. However, responsible clinical deployment, especially in resource-scarce and trust-dependent environments, demands comprehensive reliability evaluation. This must go beyond accuracy to encompass diagnostic consistency, manipulation resilience, and intelligent contextual integration, ensuring the safe and ethical application of LLMs for universal healthcare.

This study strictly evaluated the diagnostic reliability of leading LLMs, focusing on: (1) evaluating their diagnostic consistency across repeated queries and minor demographic variations of identical clinical cases; (2) examining their susceptibility to diagnostic manipulation through prompt engineering, narrative shifts, and irrelevant information insertion; and (3) evaluating the extent of their contextual awareness and ability to incorporate patient history and lifestyle factors into diagnostic reasoning.

We employed a controlled experimental methodology utilizing a dataset of 52 original patient cases, each expanded into multiple variants. These variants included demographic alterations (age, gender, race, country), rewording of symptom descriptions, and slight modifications to physical examination de- tails while maintaining core diagnostic markers unchanged. Susceptibility to manipulation was tested by strategically inserting misleading narratives and irrelevant details into diagnostic prompts. Contextual awareness was evaluated by comparing diagnoses generated with and without supplementary patient history and lifestyle information. We analyzed both quantitative diagnostic change rates and qualitative patterns in LLM responses across these manipulations.

Both LLMs demonstrated ***perfect (100%) diagnostic consistency*** *for identical clinical information, reflecting their deterministic nature and focus on core data. However, significant* ***susceptibility to manipulation*** *emerged: Gemini exhibited a* ***40%*** *diagnosis change rate, and Chat GPT* ***30%*** *when irrel- evant details were added. While Chat GPT showed a higher* ***context influence rate (77.8% vs. Gemini's 55.6%)*** *quantitatively, qualitative analysis revealed limitations in clinically subtle contextual integration for both. Both models exhibited* ***anchoring bias****, prioritizing salient clinical data, superfi- cially incorporating context, and sometimes overemphasizing demographics, and medical history while underweighting contradictory evidence.*

Despite remarkable consistency in controlled settings, LLMs' ***demonstrated susceptibility to manipulation*** *and limitations in sophisticated contextual understanding pose critical challenges for real-world clinical deployment. Specifically, LLMs exhibit weaknesses in contextual awareness and are highly susceptible to input manipulation, unlike human clinicians who leverage iterative questioning, critical evaluation, and comprehensive contextual integration. Human clinicians also express uncertainty and seek validation, contrasting with LLMs' tendency to overstate diagnostic certainty. These findings strongly emphasize the urgent need for* ***domain-specific architectures, reliable input safeguards, and***

__careful validation frameworks__ to ensure ethical and reliable LLM application in healthcare. Until these fundamental vulnerabilities are decisively overcome, broad clinical implementation of LLMs outside of highly controlled, human-supervised research settings would be premature, ethically questionable, and potentially harmful.

Due to their inability to critically evaluate input validity or request clarifying information, LLMs are demonstrably more susceptible to manipulation than clinicians.

While more susceptible to manipulation and less sophisticated in contextual reasoning than clinicians, LLMs, when used responsibly under human over- sight, can still enhance diagnostics. Future research must prioritize improving LLMs' manipulation resistance and contextual reasoning to responsibly realize their promise for global healthcare democratization.

I. Introduction

Pursuing universally accessible healthcare is a cornerstone of global equity and social justice, deeply intertwined with the fundamental human right to health. Yet, despite unprecedented advancements in medical science, vast disparities persist, particularly in resource-constrained regions where access to timely and accurate diagnostics remains a formidable challenge. Imagine a transformative paradigm shift in a world where even the most geographically remote and underserved communities possess access to sophisticated diagnostic capabilities that are readily available and free of prohibitive costs. Large Language Models (LLMs), sophisticated artificial intelligence systems capable of understanding and generating human-like text, emerge as a revolutionary force poised to democratize healthcare on a global scale. By offering the potential to automate and augment complex clinical tasks, LLMs hold the promise of bridging critical gaps in healthcare access, especially in settings lacking specialist expertise or advanced infrastructure. From assisting in differential diagnosis and interpreting medical imaging to personalizing patient education and streamlining administrative processes, the scope of LLM applications in medicine appear boundless, heralding a new era of enhanced efficiency and broader reach. However, accurate evaluation of LLM performance in clinical practice is crucial. While initial enthusiasm focuses on diagnostic accuracy, safe and ethical real-world deployment necessitates evaluating broader **reliability** dimensions. In medicine, reliability transcends accuracy, encompassing consistency, manipulation resilience, and contextual integration factors determining clinical appropriateness and trustworthiness. Understanding these reliability parameters is essential for LLMs to become dependable clinical tools. Currently, the research landscape surrounding LLMs in healthcare, while rapidly expanding, reveals a significant gap in our understanding of these critical reliability dimensions, particularly within the specific domain of medical diagnostics. Much of the existing work has focused on benchmarking diagnostic accuracy against expert clinicians in highly controlled settings, demonstrating encouraging, albeit preliminary, results. However, a systematic investigation into the consistency of LLM diagnostic recommendations, their potential susceptibility to manipulation through adversarial prompts or misleading information, and their capacity for contextual reasoning remain critically under- explored. This knowledge gap is not merely academic; it carries profound implications for the responsible and ethical integration of LLMs into clinical workflows. Imagine a scenario where an LLM, while accurate under ideal conditions, exhibits inconsistent diagnoses when presented with slightly varied patient demographics or clinical narratives. Or consider the vulnerability of a system susceptible to manipulation through engineered prompts, potentially leading to misdiagnoses based on subtle, yet clinically irrelevant, input alterations. Furthermore, a failure to effectively integrate crucial contextual factors such as patient history, lifestyle, or socio-economic background could lead to diagnostic errors rooted in a superficial understanding of the patient's holistic clinical picture. Such shortcomings, if unaddressed, could erode clinician trust, compromise patient safety, and ultimately undermine the very goal of democratizing healthcare that these technologies aspire to achieve.

To address these critical gaps in our understanding of LLM reliability in medical diagnostics, this study is guided by the following core research questions:

- How consistent are the diagnostic recommendations provided by LLMs for the same clinical scenario presented on different occasions, particularly when minor demographic or clinical variations are introduced?
- How susceptible are LLMs to manipulation through carefully crafted prompts or the introduction of irrelevant information, and how does this susceptibility compare to the known resilience and critical evaluation skills of human clinicians?
- How effectively do LLMs utilize contextual information (e.g., patient history, lifestyle factors) in making diagnostic recommendations, and how does this contextual integration compare to the comprehensive, patient-centered approach of human clinicians?

To answer these questions and enhance understanding of LLM **reliability** in diagnostics, this study empirically evaluates the diagnostic consistency, susceptibility to manipulation, and con-textual awareness of LLMs. Utilizing a controlled experimental methodology, this research aims to provide critical insights into the strengths and limitations of these AI tools, informing comprehensive validation and responsible translation of LLM technology for safe, ethical, and democratized clinical practice, ensuring patient safety and clinical trust remains a prime concern [1].

2. Method

This section details the methodological framework designed for a three-dimensional evaluation of Large Language Models (LLMs) in medical diagnosis. To ensure transparency and facilitate reproducibility, we provide a comprehensive account encompassing experimental design, synthetic data generation, standardized model interaction protocols, and analytical techniques for examining diagnostic reliability. Specifically, this chapter delineates the procedures used to evaluate LLM performance across three critical dimensions: diagnostic consistency, susceptibility to manipulation, and contextual awareness. For each dimension, we outline scenario modification strategies, prompting structures, data collection, and the mixed-methods data analysis approach integrating quantitative metrics and qualitative physician review.

2.1. Study Design

This study employed a controlled experimental and comparative design to evaluate the di- agnostic reliability of Large Language Models (LLMs) in medical diagnostic scenarios. The primary objective was to evaluate and compare the performance of prominent LLMs across three critical dimensions of diagnostic reliability: diagnostic consistency, susceptibility to manipulation, and contextual awareness [2]. The experimental nature of the study was achieved through systematic manipulation of various prompt parameters, including demographic alterations, rewording of symptom descriptions, and the introduction of irrelevant details, allowing for a direct investigation into how these factors influence the diagnostic outputs of the LLMs. The study involved the systematic manipulation of clinical scenarios to examine how variations in input affected the LLMs' diagnostic outputs. The experimental framework was structured to isolate causal relationships between input modifications and diagnostic responses, ensuring controlled evaluation conditions.

2.2. Primary Objectives

The primary objectives of this study were to evaluate the diagnostic reliability of Large Language Models (LLMs) across three key dimensions in medical contexts. Specifically, the study aimed to:

- **Quantify Diagnostic Consistency:** To examine the consistency of diagnostic recommendations provided by LLMs when presented with identical clinical scenarios across multiple queries, including variations in demographic details and presentation phrasing.
- **Measure Susceptibility to Manipulation:** To determine the extent to which LLM di- agnostic outputs are susceptible to manipulation through adversarial prompt engineering and the insertion of irrelevant clinical information.
- **Evaluate Contextual Awareness:** To evaluate the effectiveness of LLMs in integrating and utilizing contextual patient information, such as medical history and lifestyle factors, in their diagnostic reasoning processes.
- **Compare Model Performance:** To directly compare the diagnostic reliability of Google's Gemini and Open AI's ChatGPT across the aforementioned dimensions, aiming to identify model-specific strengths and limitations in clinical diagnostic tasks.
- **Qualitatively Evaluate Clinical Appropriateness:** To qualitatively evaluate the clinical appropriateness of LLM diagnostic reasoning in context by employing expert physician review to examine the validity and clinical soundness of LLM diagnostic changes in response to relevant contextual patient information.

2.3. LLM Selection and Configuration

Two commercially available, state-of-the-art Large Language Models (LLMs) were selected for evaluation in this study:

- **Google Gemini 2.0 Flash** (Accessed between February 2 and February 7, 2025)
- **Open AI ChatGPT-4o** (Accessed between February 2 and February 7, 2025)

Rationale for Model Selection

- Clinical Relevance: Both models demonstrate excellent reasoning capabilities in biomedical domains, with documented applications in symptom analysis and differential diagnosis.
- Their increasing exploration and integration into healthcare applications, such as diagnostic support tools and clinical decision assistance, underscored their relevance for this study, which aimed to gauge their utility and reliability in real-world clinical settings.

All interactions in this study were conducted via these direct chat interfaces to simulate realistic user interaction and to reflect how clinicians or healthcare professionals might typically engage with these tools in practice.

2.4. Clinical Scenarios

To evaluate the diagnostic reliability of Large Language Models (LLMs), this study employed a constructed dataset of 52 clinical scenarios. These scenarios were developed *de novo*, specifically for this research, ensuring direct alignment with the study's objectives and enabling precise experimental control over case characteristics and manipulations. The dataset was designed to comprehensively evaluate three key dimensions of LLM reliability: diagnostic consistency, susceptibility to manipulation, and contextual awareness [4].

2.4.1. Scenario Characteristics

The dataset comprises **52 clinical cases,** originating from a carefully curated selection of **39 unique medical conditions**. This approach, utilizing a slightly larger number of scenarios than unique conditions, allowed for the creation of varied presentations and refined cases within similar diagnostic categories, enhancing the reliability of the evaluation. The medical conditions were chosen to represent a **broad spectrum of clinical presentations, medical specialities, and levels of diagnostic complexity,** effectively mirroring the diversity encountered in real-world clinical practice and aligning with standards commonly used in medical education and clinical training.

To ensure comprehensive coverage, the scenarios spanned **eight major medical specialties:**

- **Cardiovascular Diseases:** Including conditions such as Myocardial Infarction (STEMI), Angina Pectoris, Stage I and Stage II Hypertension, and Heart Failure (both HFrEF and HFpEF subtypes).
- **Pulmonary Diseases:** Encompassing conditions like Acute Bronchitis, Acute Exacerbation of COPD, COPD (GOLD Stages II and IV), Asthma (Mild Persistent, Exacerbation, General presentations), COVID-19 (Mild and Moderate severity), and Influenza A (Un- complicated and with Pneumonia).
- **Neurological Diseases:** Covering Ischemic Stroke, Transient Ischemic Attack (TIA), Subarachnoid Hemorrhage, Migraine (With and Without Aura), Diabetic Neuropathy, and Generalized Anxiety Disorder.
- **Endocrine & Metabolic Disorders:** Including Type 1 and Type 2 Diabetes Mellitus, and Hypothyroidism.
- **Gastrointestinal Diseases:** Featuring Peptic Ulcer Disease (Duodenal and Gastric sub- types), Acute Appendicitis, Acute Diverticulitis, Viral Gastroenteritis, Functional Constipation, and Chronic Constipation.
- **Musculoskeletal Diseases:** Representing Lumbar Muscle Strain, Lumbar Osteoarthritis, Knee Osteoarthritis, Hip Osteoarthritis, Rheumatoid Arthritis, and Distal Radius Fracture (Colles' Fracture).
- **Infectious Diseases:** Including Urinary Tract Infection (Cystitis and Pyelonephritis subtypes), Common Cold, Strep Throat, Viral Exanthem (with possible reference to Varicella), and Acute Otitis Media.
- **Pain & Miscellaneous:** Covering Musculoskeletal Pain and Functional Constipation (also listed in Gastrointestinal due to dual classification in some medical contexts).

The **complexity of the scenarios** was intentionally varied to evaluate LLM performance across different levels of diagnostic difficulty. Cases ranged from **common, relatively straight-forward outpatient presentations** (e.g., Viral Gastroenteritis, Uncomplicated Influenza, Strep Throat) to **acute, complex, or intricate conditions** requiring more intricate diagnostic reasoning (e.g., STEMI, Subarachnoid Hemorrhage, Acute Exacerbation of COPD, Heart Failure subtypes, COPD staging, UTI subtypes). This range ensured that the LLMs were tested on both routine and more challenging diagnostic problems [5].

2.4.2. Dataset Construction and Structure

The clinical accuracy and medical validity of the scenarios were supreme. The dataset was constructed by drawing upon **evidence-based and authoritative medical resources.** The primary sources for clinical details, including **symptoms, vital signs, diagnostic criteria, risk factors, differential diagnoses, final diagnoses, and treatment recommendations,** were the point-of-care medical databases **UpToDate** and **DynaMed**. These resources were chosen for their comprehensive coverage of clinical topics, their continuous updating to reflect current medical knowledge, and their focus on providing practical, evidence-based information for clinical decision-making. To further enrich the contextual accuracy and depth of the medical information, especially for more complex or specialized cases, **Clinical Key** and **Pub- Med/ MEDLINE** were utilized as supplementary resources. This multi-faceted approach to data sourcing ensured a high degree of medical fidelity and clinical relevance for each scenario. Each of the 52 clinical scenarios was structured to ensure consistency and completeness, using a standardized set of parameters to document patient information and clinical findings
comprehensively. Each scenario record included the following structured fields:

- **Patient Information: patient_id** (unique identifier), **age, gender.**
- **Medical Background: medical_history, current_ medications.**
- **Presenting Complaint: presenting_complaint** (the patient's primary reason for seeking medical attention).
- **Symptoms:** A detailed profile of symptoms, with each symptom described using granular attributes including: **name, severity** (e.g., mild, moderate, severe), **character** (e.g., sharp, dull, crushing), **associated_symptoms, exacerbating_ factors, relieving_factors**, and **type** (e.g., productive cough).
- **Vital Signs:** Physiological measurements including **heart_ rate** (beats per minute), **blood_pressure** (systolic/diastolic in mmHg), **temperature** (in Fahrenheit), and **respiratory_rate** (breaths per minute).
- **Physical Examination:** A concise summary of pertinent physical examination findings.
- **Diagnostic Test Results:** Results from relevant diagnostic tests, including but not limited to ecg (electrocardiogram findings), troponin levels, cbc (complete blood count findings), cxr and other condition-specific laboratory or imaging results as clinically in- dicated.
- **Differential Diagnosis**: A curated list of possible conditions considered in the diagnostic process.
- **Final Diagnosis:** The confirmed or most likely diagnosis for the presented scenario.
- **Treatment:** Recommended medications, therapies, or management strategies aligned with clinical guidelines and best practices.
- **Additional Notes:** A field for supplementary clinical observations, relevant contextual details, or follow-up information as needed.

2.4.3. Data Modification and Open Access

While the foundational medical knowledge within the scenarios was thoroughly derived from authoritative sources, certain non-clinical parameters and scenario variations were intentionally modified for experimental purposes. Specifically, elements such as patient_id, age, gender, and details within the medical_history and presenting_complaint fields were manually created and adjusted to create diverse patient presentations and to facilitate the testing of consistency, susceptibility, and contextual awareness. These modifications were carefully implemented to maintain clinical plausibility while achieving the experimental objectives.

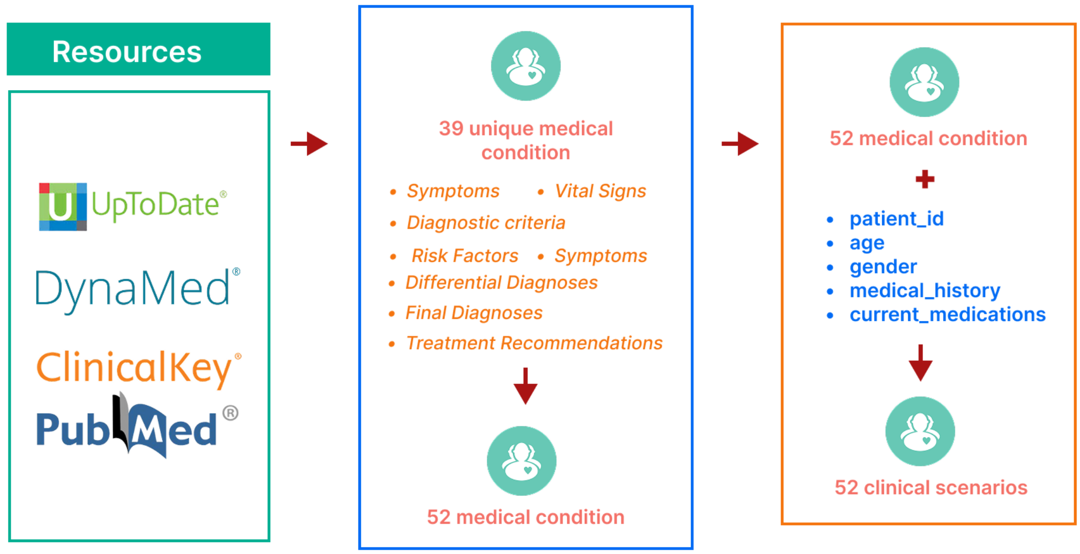

Figure 1: Dataset Construction Methodology

To ensure transparency, promote reproducibility, and contribute to the broader research com- munity, the **complete dataset of all 52 clinical scenarios will be made openly accessible** in the Supplementary Materials. This includes detailed structured data for each case and will be provided via a link to a public repository or as a direct appendix to this publication, serving as a valuable resource for future investigations in clinical decision-making with AI and medical education.

2.5. Consistency Assessment Protocol

To evaluate the diagnostic consistency of Large Language Models (LLMs), a structured and detailed protocol was designed and implemented. The core objective was to determine whether LLMs would generate stable and consistent diagnostic recommendations when presented with clinically equivalent but superficially varied patient scenarios. This evaluation was essential to evaluate the reliability of these models in delivering consistent clinical judgments, even when faced with minor, clinically inconsequential variations in patient case presentation.

2.5.1. Clinical Scenario Variations for Consistency Testing

For each of the 52 baseline clinical scenarios, a systematic approach was used to generate four distinct variations. These variations were designed to introduce minor, clinically irrelevant alterations, ensuring that the underlying diagnostic essence of each case remained unchanged [6]. The intention was to test the LLMs' sensitivity and reliability to superficial changes in input while maintaining the core clinical problem. The three specific types of variations created were:

- **Patient Demographics:** To examine for potential demographic biases and sensitivity to patient characteristics irrelevant to diagnosis, the following demographic parameters were systematically altered across the four variations:
- **Country:** Patient's country of origin.
- **Race:** Patient's self-identified race.
- **Ethnicity:** Patient's ethnicity.
- **Age:** Patient's age.
- **Gender:** Patient's gender identity.

These demographic modifications were carefully chosen to be superficial, ensuring they did not introduce clinically significant differences in disease prevalence, presentation, or risk factors for the specific medical conditions represented in the scenarios.

- **Presenting Complaint Phrasing Variations:** To evaluate the LLMs' reliable to variations in natural language and phrasing, the **presenting_complaint** field was subtly reworded in each variation. This involved the use of synonyms and similar sentence structures to express the same core complaint in slightly different linguistic forms. This manipulation tested the LLMs' ability to recognize and process clinically equivalent in- formation expressed through different linguistic forms.

- **Minor Rephrasing in Medical History and Physical Exam:** To examine sensitivity to subtle changes in descriptive language within clinical narratives, minor rephrasing was applied to the **medical_history** and **physical_exam** sections. These modifications consisted of minor rephrasing of descriptive terms and sentence structure, ensuring that the core clinical meaning and findings remained consistent across the variations. The intent was to evaluate whether LLMs were sensitive to superficial textual changes in background clinical information.

2.5.2. Standardized Prompt Structure and Procedure

To ensure consistent interaction with the LLMs, a standardized prompt structure was employed for presenting each scenario variation for all LLMs. The prompt was designed to be direct and unambiguous, instructing the LLMs to focus on providing a diagnostic conclusion based on the clinical information provided:

You are a medical doctor. Diagnose the following patient case:
[Patient Data]

Provide patient_id and the corresponding diagnosis for that patient. Provide only your final diagnosis. Do not include explanations or differential diagnoses.

This consistent prompt phrasing was used for all 624 diagnostic queries per LLM (52 base scenarios x 4 variations x 3 repetitions), eliminating prompt variability as a potential confounding factor [7].

2.5.3. Diagnostic Consistency Testing Protocol

Each of the four variations derived from a single baseline clinical scenario was presented to LLMs on **three independent occasions.** This resulted in a total of **624 diagnostic queries per LLM** (52 base scenarios × 4 variations × 3 repetitions). For each query, the LLM's diagnostic response was automatically recorded for subsequent analysis. The three repetitions per variation were designed to account for potential minor stochasticity in LLM outputs, even for identical inputs, and to ensure the reliability of the consistency assessment.

2.5.4. Criteria for Determining Diagnostic Consistency

Diagnostic consistency was evaluated at the level of the original 52 baseline clinical scenarios.

A hierarchical set of criteria was applied to determine consistency:

- **Internal Variation Consistency:** For each of the four variations of a given baseline scenario and for each LLM, the three repeated diagnostic outputs were examined. If all three outputs for a given variation were identical (clinically same diagnosis), that variation was classified as **internally consistent**.
- **Baseline Scenario Consistency:** A baseline clinical scenario was classified as yielding a **consistent diagnosis** for a specific LLM only if *all four of its variations* (across demo- graphic, phrasing, and minor rephrasing manipulations) resulted in clinically equivalent final diagnoses. Conversely, if *any one* of the four variations produced a diagnostic output that was not clinically equivalent to the others for a given baseline scenario, that scenario was classified as **inconsistent** for that LLM.
- Consistency was defined as the **LLM providing clinically equivalent diagnoses** across all four variations of a given base clinical scenario. Diagnostic outputs were considered equivalent if they matched **exactly** or were determined to be **clinically synonymous** (e.g., **gastroesophageal reflux disease** vs. **acid reflux**). If any single variation produced a diagnosis that deviated from the others in a way that altered clinical interpretation, the scenario was deemed inconsistent.

This multi-layered protocol, incorporating systematic scenario variations, repeated trials, a standardized prompt, and a precise definition of consistency, allowed for a thorough and distinct evaluation of the diagnostic consistency of Gemini and ChatGPT under experimentally controlled conditions [8].

2.6. Susceptibility to Manipulation Protocol

To evaluate the susceptibility of Large Language Models (LLMs) to diagnostic manipulation, a structured experimental protocol was designed. This protocol examined whether the diagnostic outputs of Gemini and ChatGPT could be influenced by the introduction of irrelevant, yet superficially plausible, information into the clinical prompts. The study aimed to simulate real-world clinical scenarios where patient narratives often include extraneous details, subjective interpretations, and information not directly pertinent to the core diagnosis [9]. The main objective was to determine if LLMs could effectively prioritize clinically relevant data while demonstrating reliability against diagnostically inconsequential noise, thus evaluating their suitability for use in complex, real-world clinical environments.

2.6.1. Manipulation Strategy

The manipulation strategy focused on enriching a subset of clinical scenarios with six distinct categories of irrelevant, non-clinical information. Ten original clinical scenarios were randomly selected from the 52-case baseline dataset to undergo manipulation. For each of these ten scenarios, a manipulated counterpart was created by systematically embedding extraneous details, while crucially ensuring that all diagnostically critical information (e.g., symptoms, medical his- tory, vital signs, test results) remained unchanged. The six categories of irrelevant information, designed to mimic the types of **noise** encountered in clinical practice, were as follows [10].

Example of Original vs. Manipulated Prompt

To illustrate how manipulation alters a prompt while retaining core clinical information, consider the following representative example (Case ID 27):

Original Prompt (Case ID 27, Hypothetical Example)

A 45-year-old female presents with substernal chest pain radiating to the left arm, worsened by exertion. History of hypertension and hyperlipidemia. ECG shows ST-segment elevation. Troponin levels elevated.

Manipulated Prompt (Case ID 27, Hypothetical Example)

A 45-year-old female (who collects vintage teacups and **always knows when it will rain**) presents with substernal chest pain like a lightning bolt, radiating to the left arm. Pain worsens with exertion, but she's **more worried about missing yoga class**. Insists her recent Sedona retreat **cleared her chakras**. History of hypertension. Patient dismisses cholesterol as a **myth** and prefers **ginger supplements**. ECG shows ST-segment elevation. Troponin levels elevated.

As demonstrated in this example, the manipulated prompt retains all core clinical information (age, symptoms, medical history, ECG, troponin) while including elements from the irrelevant information categories (figurative language, alternative medicine beliefs, irrelevant lifestyle de- tails, subjective claims) [11, 12].

Category	Manipulation Strategy	Examples of Introduced Irrelevant Information
Whimsical/Figurative Language	Use of metaphorical, humorous, or exaggerated symptom descriptions.	**Abdominal pain rated 9/10, like stepping on a Lego barefoot; Fatigue feels like wading through treacle; Chest pain described as a balloon popping in my chest.**
Alternative Medicine Beliefs	References to unproven therapies or dismissive attitudes toward conventional medicine.	**Patient insists apple cider vinegar is a universal cure; Prefers chakra alignment therapy over conventional pain relievers; Believes high cholesterol is a modern myth.**
Anecdotal/Subjective Claims	Inclusion of unverified patient intuitions, personal theories, or irrelevant boasts.	**Claims to predict rain based on joint pain; Boasts of being healthy as a horse; Attributes mood changes to the aurora borealis seen online.**
Cultural/Regional References	Location-specific idioms, habits, or travel history not relevant to diagnosis.	**Uses British idioms such as getup- and-go just got up and went; Mentioned returning from a yoga retreat in Sedona; Observed wearing a woolly scarf indoors despite mild temperatures.**
Patient Demeanor/ Psychology	Descriptions of emotional states, personality traits, or consultation behaviors.	**Sighed wistfully throughout the consultation; More concerned about missing a yoga class than chest pain; Used excessive air quotes and wellness jargon.**
Irrelevant Lifestyle Details	Unrelated hobbies, family history, or personal preferences.	**Collects vintage teacups as a hobby; Father is an avid vinyl record collector; Takes ginger supplements daily for an energy boost; Enjoys knitting and watching period dramas.**

Table 1: Examples of Manipulation Strategies and Irrelevant Information Introduced in Diagnostic Scenarios

2.6.2. Prompt Structure and Susceptibility Testing Procedure

Both the original and manipulated versions of the ten selected clinical scenarios were presented to Gemini and ChatGPT using a standardized prompt:

You are a medical doctor. Diagnose the following patient case:
[Patient Data]
Provide patient_id and the corresponding diagnosis for that patient. Provide only your final diagnosis. Do not include explanations or differential diagnoses.

For each of the ten selected scenarios, both the original and the manipulated versions were presented once to each LLM. The diagnostic outputs were recorded and then compared.

2.6.3. Criteria for Determining Susceptibility to Manipulation

To evaluate whether manipulation led to a diagnostically significant change, the diagnoses generated by each LLM for the original and manipulated versions of each scenario were com- pared. The criteria for determining susceptibility were as follows:

- **Diagnosis Change (Susceptible to Manipulation):** Susceptibility was defined as occurring when the manipulation resulted in a **clinically distinct diagnosis**. This was determined by evaluating if the diagnosis provided for the manipulated scenario represented a different medical condition from the diagnosis provided for the original scenario, reflecting a change that could alter clinical interpretation or management. For example, a change from *acute appendicitis* in the original scenario to *irritable bowel syndrome* in the manipulated scenario would be considered a clinically distinct diagnosis change, indicating susceptibility.
- **No Diagnosis Change (Reliability to Manipulation):** It was defined as occurring when the LLM's diagnostic outputs for the original and manipulated scenarios remained **clinically equivalent**, despite the introduction of irrelevant information.

By quantifying the number of scenarios for which each LLM exhibited a change in diagnosis following manipulation, we calculated the overall susceptibility rate for all LLMs. This thorough protocol, using categorized irrelevant information and a direct comparison of diagnostic outputs, provided a reliable measure of the LLMs' vulnerability to diagnostic manipulation in clinically realistic contexts.

This detailed protocol, which incorporated systematically designed categories of irrelevant information, controlled comparison of original and manipulated scenarios, and clinically in- formed criteria for evaluating diagnosis change, provided a precise framework for evaluating the susceptibility of Gemini and ChatGPT to diagnostic manipulation [13].

2.7. Contextual Awareness Assessment Protocol

This protocol was designed to systematically evaluate the contextual awareness of Large Language Models (LLMs) in medical diagnostics. Specifically, it aimed to determine how effectively LLMs could incorporate clinically relevant contextual information such as patient demographics, medical history, lifestyle factors, and intricate clinical presentations into their diagnostic reasoning processes. The core of this assessment was to compare LLM performance against clinically expected diagnostic shifts, induced by intentionally crafted contextual modifications to baseline clinical scenarios.

2.7.1. Approach to Scenario Modification

To evaluate contextual awareness, we employed a targeted approach using a subset of clinical scenarios from the larger dataset. Two original patient cases were randomly selected from the 52 baseline scenarios to serve as the foundation for this evaluation.

- **Dataset Creation**
- **Case 1:** One baseline case was modified to create **four** contextually varied versions.
- **Case 2:** A second baseline case was modified to create **five** contextually varied versions.
- **Total:** Nine contextually varied scenarios derived from two distinct baseline cases.
- **Types of Contextual Modifications:** Clinically meaningful modifications were systematically introduced, altering key contextual parameters known to influence diagnostic probabilities and clinical decision-making. These modifications spanned several categories:
- **Demographics:** Patient demographics were varied, including:
- **Age, gender, and race/ethnicity** – to account for variations in disease prevalence, typical presentations, and risk factors.
- **Country of origin** – primarily to explore potential biases or sensitivities related to geographical context.
- **Clinical Presentation (Symptoms and Exam Findings):** Substantial alterations were made to the core clinical presentation, modifying:
- **Presenting complaint**
- **Symptom characteristics** (severity, duration, location, character, associated symptoms, relieving/aggravating factors)
- **Physical examination findings**

These changes created clinically distinct yet related symptom complexes and exam findings to steer diagnostic reasoning.

- **Medical History and Medications:** Adjustments ensured clinical coherence with altered presentations by:
- Adding, removing, or modifying pre-existing conditions.
- Altering medication regimens to reflect different diagnostic possibilities.
- **Diagnostic Test Results:** Key test results (e.g., ECG findings, lab values, imaging results) were systematically altered to align with varied clinical presentations, guiding the LLMs toward different, clinically appropriate diagnoses [14].

2.7.2. Standardized Prompt Structure

To ensure consistent interaction and minimize variability due to prompting, all LLMs (Gemini 2.0 Flash and ChatGPT GPT-4o) received identical prompts with a standardized structure:

You are a medical doctor. Diagnose the following patient case:
[Manipulated Patient Data]
Provide patient_id and the corresponding diagnosis for that patient. Provide only your final diagnosis. Do not include explanations or differential diagnoses.

This prompt format was designed to elicit concise final diagnoses, focusing the LLMs on the diagnostic task and minimizing extraneous outputs.

2.7.3. Context-Rich and Context-Absent Scenarios

To clarify the nature of contextual variation, Table 2 provides examples of how contextual layers were manipulated to create **context-rich** scenarios compared to a hypothetical **context-absent** baseline.

Context Layer	Context-Rich Example	Context-Absent Example
Demographics	35-year-old South Asian male	Adult patient
Medical History	History of hypertension, GERD, currently on pantoprazole	No significant past medical history reported
Symptoms	Epigastric pain radiating to back, worsens after spicy meals	Abdominal pain
Lifestyle/Geographic	Current smoker, works night shifts in a factory setting	[Omitted - no lifestyle/geographic Details]
Diagnostic Test Results	H. pylori stool antigen test is positive	Routine lab results are within normal limits

Table 2: Examples of Contextual Layers and Variations

2.7.4. Illustrative Scenario Modifications

To further illustrate the contextual modifications, consider these examples derived from a hypothetical **Original Case 1** (presented in a context-absent format): ***A 60-year-old male presents with chest pressure. ECG shows ST elevation. Troponin elevated.***

- **Context-Rich Variation 1 Example:**
  - Contextual Additions: Demographic shift to a 35-year-old female; addition of anxiety disorder and GERD history; presenting complaint changed to upper back discomfort related to emotional stress; ECG reported as normal and H. pylori positive test result added [15].
  - Resulting Scenario Fragment: 35-year-old female with anxiety disorder and GERD history reports upper back discomfort after emotional stress. Normal ECG. H. pylori positive.
  - Intended Contextual Shift: Away from acute cardiac event towards musculoskeletal or gastrointestinal etiology.

- Context-Rich Variation 2 Example:
  - Contextual Additions: Demographic shift to 70-year-old Black male; addition of atherosclerotic risk factors: diabetes, hypertension, and smoking history; presenting complaint described as crushing substernal chest pain; ECG showing ST depression; and LDL 190 mg/dL reported.
  - Resulting Scenario Fragment: 70-year-old Black male with diabetes, hypertension, and smoking history describes crushing substernal chest pain. ECG shows ST depression. LDL 190 mg/dL.

- ◦ Intended Contextual Shift: Reinforces high-risk cardiac scenario, shifting to- wards possible angina or non-ST elevation myocardial infarction.

2.7.5. Example Prompts: Context-Absent vs. Context-Rich

To provide concrete examples of the prompts used, Table 3 presents illustrative prompts for both context-absent and context-rich variations, derived from the two baseline cases used in this assessment.

Note:- The patient data presented in Table 3 is only for demonstration and does not represent the experimental data.

Case	VariationType	Prompt Example
Case 1	Context- Absent	You are a medical doctor. Diagnose the following patient case: A patient presents with chest discomfort. ECG abnormal. Provide diagnosis. Provide ONLY your final diagnosis. Do not provide explanations.
Case 1	Context- Rich	You are a medical doctor. Diagnose the following patient case: A 28-year-old Nepalese female with no cardiac history reports sharp left-sided chest pain worsening with deep inspiration. Recent travel to a high-altitude region. D-dimer 0.3 µg/mL. ECG normal. Diagnose. Provide ONLY your final diagnosis. Do not provide explanations.
Case 2	Context- Absent	You are a medical doctor. Diagnose the following patient case: Patient complains of abdominal pain and nausea. Endoscopy shows ulcer. Provide diagnosis. Provide ONLY your final diagnosis. Do not provide explanations.
Case 2	Context- Rich	You are a medical doctor. Diagnose the following patient case: A 65-year-old Japanese male with daily NSAID use for osteoarthritis presents with melena and epigastric tenderness. H. pylori negative. Hemoglobin 9.2 g/dL. Diagnose. Provide ONLY your final diagnosis. Do not provide explanations.

Table 3: Comparison of Context-Absent and Context-Rich Variations in Diagnostic Prompts

2.7.6. Methodology for Diagnosis Comparison and Contextual Awareness Scoring Quantitative and Qualitative Analysis

To evaluate the LLMs' contextual awareness, a mixed-methods approach was used, combining quantitative and qualitative analyses of their diagnostic outputs.

2.7.6.1. Quantitative Diagnostic Match Rate

For each contextually varied scenario, we quantitatively examined whether the LLM's generated diagnosis **matched the clinically expected diagnosis** given the introduced contextual modifications. The **clinically expected diagnosis** was determined *a priori* by the physician reviewers during the scenario design phase. This pre-determined diagnosis represented the clinically appropriate diagnostic shift that a human physician would be expected to make in response to the specific contextual alterations in each scenario variation. Essentially, for each context-rich scenario, there was a target **correct** diagnosis based on the contextual cues [16].

To calculate the **Diagnostic Match Rate**, we employed the following process:

- **Diagnosis Categorization:** For each contextually varied scenario presented to each LLM (Gemini and ChatGPT), the generated diagnosis was recorded.
- **Comparison to Clinically Expected Diagnosis:** Each LLM-generated diagnosis was then compared to the pre-defined **clinically expected diagnosis** for that specific context-rich scenario.
- **Binary Match Determination:** A binary **match** was determined based on clinical equivalence. A diagnosis was considered a **match** if it was either:
- **An Exact Match:** Identical to the pre-defined clinically expected diagnosis (e.g., LLM diagnosed **Gastric Ulcer** and the expected diagnosis was **Gastric Ulcer**).
- **Clinically Equivalent:** Synonymous or clinically interchangeable with the expected diagnosis, as determined by physician reviewers. Clinical equivalence ac- counted for minor variations in terminology or phrasing that did not alter the clinical meaning (e.g., LLM diagnosed **Unstable Angina Pectoris** and the expected diagnosis was **Unstable Angina**). Disagreements on clinical equivalence, if any, would have been resolved through physician consensus, although in this study, such disagreements were minimal due to the clear nature of the expected diagnostic shifts.
- **Calculation of Match Rate:** For each LLM (Gemini and ChatGPT), the **Diagnostic Match Rate** was calculated as the percentage of contextually varied scenarios for which the LLM's generated diagnosis was deemed a **match** (either exact or clinically equivalent) to the pre-defined clinically expected diagnosis. This was calculated as:

$$\text{Diagnostic Match Rate} = \left(\frac{\text{Number of Matched Diagnoses}}{\text{Total Number of Contextually Varied Scenarios}}\right) \times 100\% \tag{1}$$

Where:

- **Number of Matched Diagnoses** refers to the count of scenarios where the LLM's diagnosis was deemed a match (exact or clinically equivalent) to the expected diagnosis.
- **Total Number of Contextually Varied Scenarios** represents the total number of clinical variations analyzed.

Interpretation: A higher Diagnostic Match Rate indicates greater alignment between the LLM's diagnostic reasoning and expected clinical decision-making, suggesting a reliable ability to integrate contextual modifications. Conversely, lower match rates suggest susceptibility to contextual misinterpretation or diagnostic instability [17].

2.7.6.2. Qualitative Review of Contextual Appropriateness

In addition to the quantitative Diagnostic Match Rate, a **qualitative review** was con- ducted by two independent board-certified physicians to gain deeper insights into the **nature** and **clinical appropriateness** of the LLMs' responses to contextual information. This qualitative analysis aimed to understand *why* the LLMs achieved or failed to achieve a diagnostic match in certain scenarios and to identify patterns in their contextual reasoning. The qualitative review process involved the following steps:

- **Physician Review of LLM Diagnoses:** Two independent board-certified physicians were provided with:
- The **context-absent baseline scenario.**
- Each **context-rich variation** of that baseline scenario, including the specific con- textual modifications introduced.
- The **diagnoses generated** by both Gemini and ChatGPT for each context-rich scenario.
- The **pre-defined clinically expected diagnosis** for each context-rich scenario.
- **Categorization of Diagnostic Changes:** For each context-rich scenario and for each LLM, the physicians independently categorized the diagnostic shift (or lack thereof) from the baseline scenario to the context-rich scenario based on its clinical appropriateness, using pre-defined categories:
- **Appropriate Change:** The LLM's diagnostic shift from the baseline to the context- rich scenario was deemed clinically justified, evidence-based, and aligned with the intended contextual shift. It indicated that the LLM effectively utilized the contextual information to refine the diagnosis in a clinically sound manner.
- **Inappropriate Change:** The LLM's diagnostic shift was considered clinically un- justified, erroneous, illogical, or mis interpretive of the contextual information. It indicated a failure in contextual reasoning, leading to a clinically invalid or less accurate diagnosis.
- **Ambiguous Change:** The clinical appropriateness of the diagnostic shift was not definitively clear-cut based on the provided information alone and required further clinical detail or investigation. These cases represented scenarios where the con- textual modification introduced genuine clinical ambiguity, or where the LLM's response, while not clearly incorrect, was not definitively the most clinically appropriate shift.
- **Assessment of Inter-Rater Reliability and Resolution of Discrepancies:** To ensure the reliability of the qualitative categorization, the agreement between the two physicians' independent assessments was quantified using **Cohen's kappa (κ)**. The calculated Cohen's kappa coefficient (κ = 0.85) indicated near-perfect agreement between the physician reviewers, demonstrating a high degree of consistency in their clinical judgments [18]. Any cases where the initial physician categorizations differed (ambiguous cases and disagreements) underwent secondary review and discussion, involving a third senior clinician, to reach a final consensus categorization and resolve any discrepancies.

2.7.7. Rationale for Protocol Design

This mixed-methods protocol, combining quantitative Diagnostic Match Rates with in- depth qualitative physician review, was specifically designed to provide a comprehensive and clinically meaningful assessment of LLM contextual awareness. The quantitative metric pro- vided an overall measure of how frequently the LLMs' diagnostic outputs aligned with clinically expected shifts, while the qualitative analysis provided essential insights into the *clinical validity* and *reasoning processes* underlying these responses. This combined approach allowed for a sound and subtle evaluation of the extent to which Gemini and ChatGPT could effec- tively emulate human-like clinical judgment in integrating and responding to clinically relevant contextual information within complex diagnostic scenarios [19].

2.8. Data Analysis

The diagnostic outputs from all LLMs across experimental conditions (consistency, susceptibility to manipulation, and contextual awareness) were systematically analyzed using mixed methods to comprehensively evaluate diagnostic reliability.

2.8.1. Quantitative Analysis: Performance Metrics for Diagnostic Reliability

To quantitatively examine the LLMs' performance across the three key dimensions of diagnostic reliability, we calculated specific metrics for each dimension. These metrics were designed to provide objective, numerical measures of consistency, susceptibility to manipulation, and the influence of clinically relevant context on diagnostic outputs.

2.8.1.1. Consistency Rates: Measuring Output Stability

- **Definition:** The **Consistency Rate** was defined as the percentage of baseline clinical scenarios for which an LLM produced an *identical* final diagnosis across three repeated presentations of the *same unmanipulated clinical scenario* with different variations. This metric evaluates the internal stability and reproducibility of the LLMs' diagnostic outputs when provided with the same input multiple times.

- **Calculation Formula:**

$$\text{Consistency Rate} = \left(\frac{\text{Number of Baseline Scenarios with Consistent Diagnoses}}{\text{Total Number of Baseline Scenarios}} \right) \times 100\% \tag{2}$$

Where:

- **Number of Baseline Scenarios with Consistent Diagnoses** refers to the count of baseline scenarios (out of 52) for which all three diagnostic outputs from an LLM were clinically identical.
- **Total Number of Baseline Scenarios** is the total number of original, unmanipulated clinical scenarios used in the consistency assessment (n = 52).

Interpretation: A higher Consistency Rate indicates greater stability and reliability in the LLM's diagnostic outputs under identical input conditions, suggesting a reliable and deterministic response to consistent clinical information. Conversely, lower Consistency Rates would suggest potential instability or variability in the LLM's diagnostic reasoning process [20].

2.8.1.2. Susceptibility Rates: Evaluating Vulnerability to Manipulation

• **Definition:** The **Susceptibility Rate** quantifies the percentage of manipulated scenarios in which an LLM's final diagnosis *changed* when compared to the diagnosis generated for the corresponding *original, unmanipulated prompt* of the same clinical scenario. This metric evaluates the LLMs' vulnerability to diagnostically irrelevant but superficially plausible information introduced into the prompt.

Calculation Formula:

$$\text{Susceptibility Rate} = \left(\frac{\begin{array}{c}\text{Number of Manipulated Scenarios}\\ \text{Resulting in a Changed Diagnosis}\end{array}}{\text{Total Number of Manipulated Scenarios}} \right) \times 100\% \tag{3}$$

Where:

- **Number of Manipulated Scenarios Resulting in a Changed Diagnosis** is the count of manipulated scenarios (out of 10 tested) for which the LLM's diagnosis differed from the baseline diagnosis of the original, unmanipulated scenario.
- **Total Number of Manipulated Scenarios** is the total number of scenarios subjected to manipulation testing (n=10).

Interpretation: A higher Susceptibility Rate indicates a greater vulnerability of the LLM to manipulation, suggesting that its diagnostic output is more easily swayed by irrelevant information. This would highlight a potential weakness in the model's ability to filter noise and focus on core clinical data. A lower Susceptibility Rate suggests greater stability against such manipulation [21].

2.8.1.3. Context Influence Rates: Evaluating Responsiveness to Contextual Variations

- **Definition:** The **Context Influence Rate** measures the percentage of contextually varied scenario sets in which the LLM's final diagnosis *changed* across the different contextually modified versions of a single, original clinical case. This metric evaluates the LLMs' responsiveness to clinically meaningful contextual factors, such as changes in demographics, medical history, or clinical presentation.

- **Calculation Formula:**

$$\text{Context Influence Rate} = \left(\frac{\begin{array}{c}\text{Number of Contextually Varied Scenario}\\ \text{Sets Showing Diagnostic Changes}\end{array}}{\begin{array}{c}\text{Total Number of Contextually Varied}\\ \text{Scenario Sets}\end{array}} \right) \times 100\% \tag{4}$$

Where:

- **Number of Contextually Varied Scenario Sets Showing Diagnostic Changes** is the count of original scenario sets for which the LLM produced different diagnoses across the contextually varied versions.
- **Total Number of Contextually Varied Scenario Sets** is the total number of original scenarios used as the basis for contextual variation.

Interpretation: The Context Influence Rate should be interpreted with detail. While it quantifies the LLMs' responsiveness to contextual changes, a higher rate is not inherently indicative of better performance. It merely suggests a greater degree of diagnostic variability in response to contextual shifts. The *clinical appropriateness* and *validity* of these diagnostic changes—whether they represent clinically justified adaptations to context or inappropriate over-sensitivity—were critically evaluated through the qualitative analysis described below.

2.8.2. Qualitative Analysis: Clinician-Informed Review of Diagnostic Changes

To provide a deeper understanding of the clinical relevance and appropriateness of the diagnostic changes observed—particularly in the Contextual Awareness experiments—a structured qualitative review was conducted. This analysis aimed to move beyond simple numerical metrics and evaluate the clinical meaningfulness of the LLMs' diagnostic variability.

2.8.3. Qualitative Review Process:

- **Case Selection for Review:** As described previously, qualitative review focused on cases exhibiting diagnostic variability. Specifically, all instances where LLM diagnoses differed across compared prompts (original vs. manipulated, or across contextually varied prompts) were flagged for in-depth clinical evaluation. This targeted approach ensured that qualitative analysis concentrated on scenarios where the *nature* of diagnostic changes was most pertinent.
- **Establishment of Reference Standards:** To provide a benchmark for evaluating the clinical validity of LLM diagnoses, we established **ground-truth** diagnoses for each clinical scenario. These reference standard diagnoses were derived from **UpToDate** and **DynaMed,** two widely respected and evidence-based clinical decision support systems.
- **Independent Review by Board-Certified Physicians:** The qualitative review of di- agnostic changes was performed by two board-certified physicians, who remained anonymous to protect their privacy and ensure unbiased assessments. These physicians possessed expertise in internal medicine and relevant subspecialties, aligning with the clinical domains covered by the study's scenarios. The physicians independently reviewed each flagged case, provided with the original and compared prompts (manipulated or contextually varied), and the corresponding diagnostic outputs from Gemini and ChatGPT.

- **Categorization of Clinical Appropriateness:** Physicians categorized each diagnostic change based on clinical validity using pre-defined categories: **Clinically Appropriate Diagnostic Change, Clinically Inappropriate Diagnostic Change, and Ambiguous Diagnostic Change.**

- **Assessment of Inter-Rater Reliability and Resolution of Discrepancies:** The agreement between the two physicians' assessments was measured using Cohen's kappa (($\kappa = 0.85$), showing near-perfect consistency. Disagreements were resolved through secondary review and discussion with a third senior clinician to reach a consensus.

2.8.4. Integration of Qualitative and Quantitative Findings

The insights gained from the qualitative analysis were directly integrated with the quantitative metrics (Context Influence Rates) to provide a more holistic and clinically grounded interpretation of the LLMs' diagnostic performance. Specifically:

- **Contextual Awareness Cases:** The qualitative review was essential for determining whether the diagnostic shifts observed in contextually varied scenarios were genuinely *contextually appropriate*—reflecting clinically sound adaptations to new contextual in- formation—or represented erroneous or clinically illogical changes. This allowed us to interpret the Context Influence Rates in terms of true contextual *awareness* versus mere diagnostic variability.

By combining quantitative metrics with the in-depth qualitative clinical review, this comprehensive data analysis approach provided a reliable and clinically meaningful evaluation of the diagnostic reliability and reasoning capabilities of Gemini and ChatGPT across the different dimensions.

2.9. Ethical Considerations and Limitations

2.9.1. Data Privacy and Anonymization

This study employed de novo synthetic patient data for all experiments, eliminating direct privacy risks associated with real-world Protected Health Information (PHI). While synthetic data reduces re-identification risks, we upheld ethical data handling and security best practices throughout the study, acknowledging the sensitive healthcare context [22].

2.9.1. Potential for Bias and Fairness in Diagnostic AI

LLMs are trained on massive datasets that may reflect societal biases, potentially leading to disparities in diagnostic accuracy across demographic groups. While not directly investigated here, we acknowledge algorithmic bias as a critical ethical concern in healthcare AI, potentially exacerbating health inequities [24]. Future research must prioritize bias evaluation and mitigation to ensure equitable AI-augmented healthcare for all.

2.9.2. Transparency, Explainability, and the Black Box Challenge

Transparency and explainability are ethically essential in high-stakes diagnostics. Clinician trust and accountability depend on understanding AI reasoning. This study recognizes the inherent limitations in the explainability of current LLMs like Gemini and ChatGPT. Future efforts must prioritize enhancing AI system transparency, utilizing techniques like attention mechanisms and explainable AI, to foster responsible clinical integration and oversight.

2.9.3. Responsible Clinical Use and the Imperative of Clinician Oversight

This research underscores the ethical imperative for LLMs to serve as adjunctive tools, augmenting, not replacing, human clinical expertise. Ethical AI deployment in healthcare requires reliable clinician oversight, including critical evaluation of AI outputs and safeguarding human clinical judgment in all patient care aspects. Over-reliance on unvalidated AI diagnoses risks clinician deskilling, inappropriate delegation, and compromised patient safety.

2.9.4. Potential for Misinformation, Misdiagnosis, and Impact on Patient Outcomes

Our findings, particularly regarding manipulation susceptibility and limited contextual awareness, highlight the real risk of misinformation and misdiagnosis if LLMs are prematurely deployed without safeguards. Erroneous LLM outputs could negatively impact patient out- comes, causing treatment delays, inappropriate care, and eroded trust. This study aims to inform responsible deployment strategies, emphasizing comprehensive validation, monitoring, and integration under clinician direction to mitigate these risks [25].

2.9.5. Broader Ethical and Societal Landscape of AI in Healthcare

AI integration raises broader ethical, legal, and societal questions beyond this study's scope. These include accountability for AI errors, evolving patient-physician relationships, impacts on healthcare equity, and the need for proactive policy. Ongoing multi-stakeholder ethical reflection—involving clinicians, patients, ethicists, policymakers, and developers—is crucial to responsibly guide AI development in healthcare and ensure it enhances beneficence, non- maleficence, autonomy, and justice.

In conclusion, this study endeavors to contribute to the growing body of empirical research critically exploring the capabilities and limitations of LLMs in the complex domain of health- care. We strongly advocate for continued and independent evaluation of these technologies, coupled with proactive and comprehensive ethical consideration, to ensure that AI-driven tools are developed and implemented in a manner that demonstrably enhances, rather than potentially compromising, the paramount goals of patient safety, clinical quality, equitable access, and the fundamental trust that underpins the patient-physician relationship.

3. Result

The detailed findings for each dimension are presented below, highlighting both the promising capabilities and key limitations of these LLMs in clinical diagnosis.

3.1. Diagnostic Consistency

The first dimension of our investigation systematically evaluated the diagnostic consistency of Gemini and ChatGPT. This assessment focused on the reproducibility of their diagnostic outputs across variations of clinically equivalent scenarios and repeated trials, adhering to a comprehensive hierarchical methodology designed to comprehensively evaluate diagnostic stability [26]. The most significant finding of this analysis is that both LLMs demonstrated **perfect diagnostic consistency** across all 52 baseline clinical scenarios when evaluated against our defined criteria for clinical equivalence .

LLM	**Baseline Scenario Consistency Rate (%) Gemini 100.0**
Gemini	100.0
ChatGPT	100.0

Table 4: Baseline Scenario Consistency Rates of LLMs

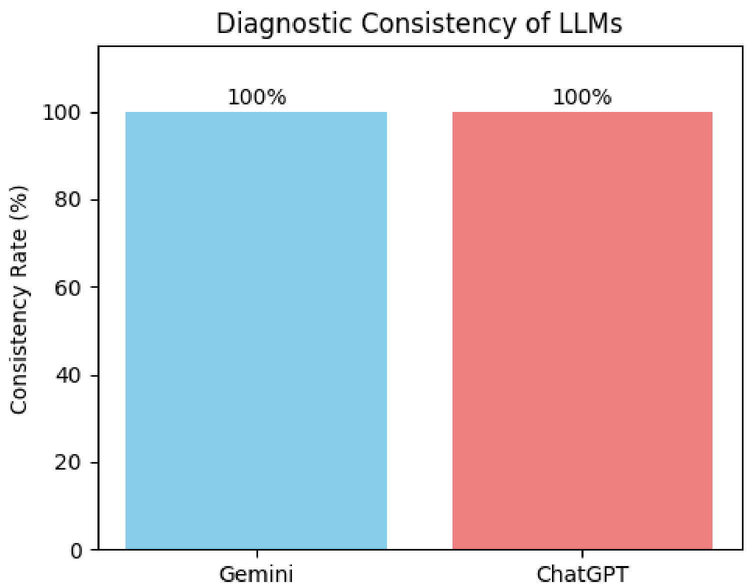

Figure 2: Bar Chart Visualization of Baseline Scenario Consistency of LLMs

- **Quantitative Analysis: 100% Baseline Scenario Consistency Rate for Both LLMs:** As numerically summarized in **Table 4** and visually represented in **Figure 2**, both Gemini and ChatGPT achieved a **100% Baseline Scenario Consistency Rate**. This key quantitative result signifies that, for each of the 52 distinct baseline clinical scenarios, when subjected to our defined consistency evaluation process, the models were consistently deemed to provide clinically equivalent diagnoses across all tested variations and repetitions [27]. This perfect consistency rate was uniformly observed across the entire set of baseline scenarios for both LLMs, demonstrating a high level of diagnostic stability as defined by our methodology.
- **A consistent diagnosis for a baseline scenario was defined as the LLM pro- viding clinically equivalent diagnoses across all four variations (demographic, phrasing, rephrasing) and their repeated trials**. This means that to achieve a **consistent** classification for a baseline scenario, the LLM had to demonstrate clinical equivalence across all minor alterations we introduced to that base case. Clinical equiva- lence was defined as exact matches or clinically synonymous diagnoses, as determined by physician reviewers.
- **Interpretation of Perfect Baseline Scenario Consistency:** The unequivocally high Baseline Scenario Consistency Rates achieved by both Gemini and ChatGPT demonstrate a significant degree of diagnostic stability and reproducibility under the controlled conditions of our study. This finding indicates that, when evaluated through the lens of our multi-layered methodology incorporating scenario variations and repeated trials, both LLMs exhibit a reliable ability to provide clinically equivalent diagnostic outputs for a given baseline clinical scenario, across minor alterations in input phrasing and demo- graphics. This suggests a foundational level of algorithmic reliability in their diagnostic reasoning when examined for consistency across clinically similar presentations of the same underlying clinical problem.

3.2. Susceptibility to Manipulation

The second dimension of our investigation explored the susceptibility of LLMs to diagnostic manipulation. This assessment specifically examined their vulnerability to the introduction of clinically irrelevant, but superficially plausible, information into the clinical prompts, mirroring the presence of noise often encountered in real-world clinical narratives. Our findings revealed that while both models exhibited a degree of reliability by retaining their original diagnoses in the majority of manipulated scenarios, neither LLM was entirely impervious to manipulation, and they demonstrated distinct susceptibility profiles [28].

- **Quantitative Susceptibility Analysis: Gemini Exhibits Numerically Higher Susceptibility Rate:** The quantitative analysis of susceptibility, summarized in **Table 5** and visualized in **Figure 3**, revealed differential vulnerability between the two LLMs. **Gemini exhibited a Susceptibility Rate of 40.0%**, indicating that in 4 out of 10 manipulated clinical scenarios, the introduction of irrelevant information led to a change in its primary diagnosis. In comparison, **ChatGPT demonstrated a numerically lower Susceptibility Rate of 30.0%**, with diagnostic alterations occurring in 3 out of 10 manipulated cases.

LLM	Cases with Changed Diagnosis (%)
Gemini	40.0
ChatGPT	30.0

Table 5: Susceptibility Rates (Diagnosis Change) of LLMs Under Manipulation

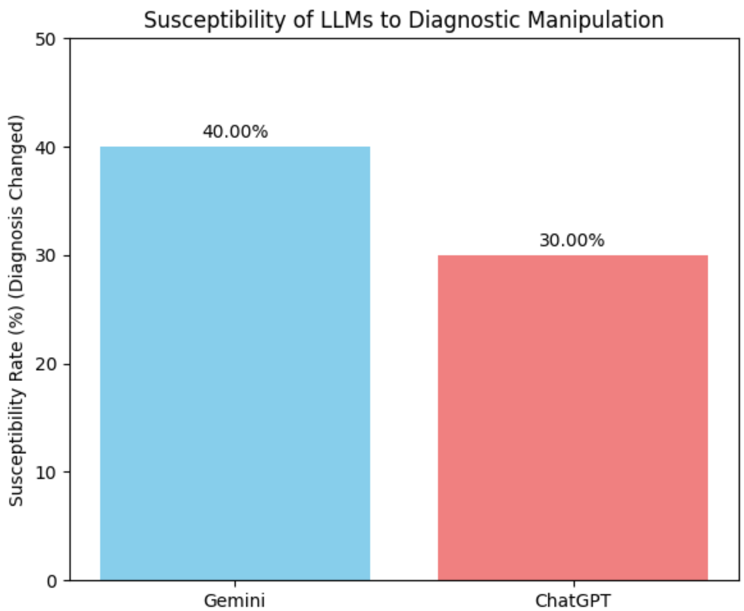

Figure 3: Bar Chart Comparing Susceptibility Rates (Diagnosis Change) of LLMs

- **Manipulation Protocol and Irrelevant Information Categories:** To systematically examine susceptibility, we selected a subset of ten original clinical scenarios from our 52-case dataset for manipulation. For each of these ten scenarios, we created a manipulated counterpart by embedding clinically **irrelevant, non-clinical information across** different distinct categories, while crucially maintaining all diagnostically critical clinical data unchanged [29]. These distinct categories of irrelevant information, designed to simulate **noise** in clinical settings, were: 1) Whimsical/Figurative Language, 2) Alternative Medicine Beliefs, 3) Anecdotal/Subjective Claims, 4) Cultural/Regional References, 5) Patient Demeanor/Psychology, and 6) Irrelevant Lifestyle Details (detailed examples of each category are provided in the Methodology section). This structured approach ensured a systematic and clinically plausible form of manipulation.
- **Criteria for Determining Susceptibility:** Susceptibility to manipulation was deter- mined by comparing the diagnoses generated by each LLM for the original and manipulated versions of each of the ten scenarios. A diagnosis change, indicative of susceptibility, was defined as occurring when the manipulation resulted in a **clinically distinct diagnosis,** meaning the diagnosis for the manipulated scenario represented a different medical condition from the original diagnosis, and reflected a change that could potentially alter clinical interpretation or subsequent patient management. Conversely, reliability was defined as the LLM maintaining clinically equivalent diagnostic outputs for both the original and manipulated scenarios, despite the introduction of irrelevant information [30].

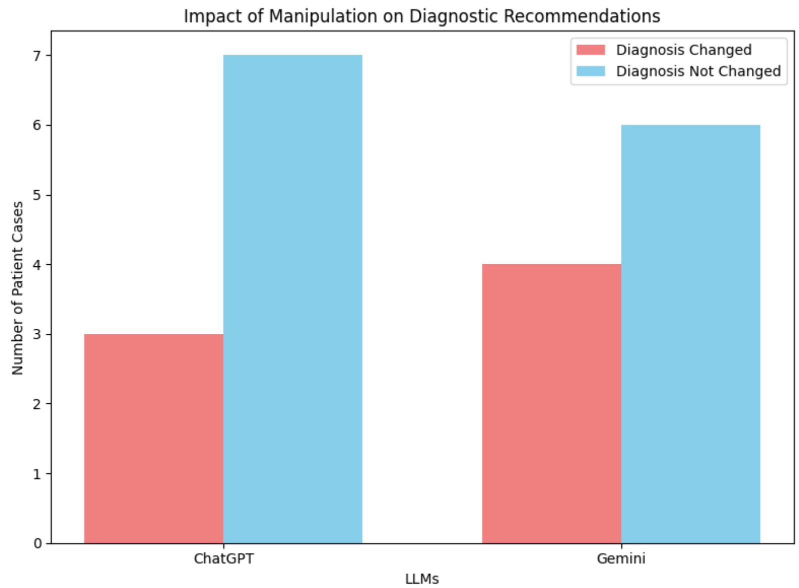

Figure 4: Grouped Bar Chart of Diagnosis Change Status under Manipulation

- **Interpretation of Quantitative Susceptibility Findings:** The quantitative Susceptibility Rates reveal that both Gemini and ChatGPT are vulnerable to diagnostic shifts when presented with clinically irrelevant, yet plausible, input modifications, although Gemini exhibited a numerically higher rate of such susceptibility. This suggests that while both models can be influenced by diagnostically inconsequential **noise**, ChatGPT demonstrated slightly greater overall reliability against this specific form of manipulation based on the aggregate rates of diagnostic change. However, the clinical significance of these changes requires further examination through qualitative analysis. The numerical susceptibility rates, while providing an overview of the frequency of diagnostic changes, do not fully capture the clinical implications of these shifts. A Susceptibility Rate of 30-40% indicates a non-negligible vulnerability to manipulation, suggesting that in a significant minority of cases, the introduction of irrelevant information *can* sway the diagnostic output of these LLMs. This raises concerns about the potential for these models to be influenced by extraneous details in real-world clinical settings, where patient narratives are often complex and may include clinically unimportant information. Furthermore, the differential susceptibility rates between Gemini and ChatGPT suggest potential architectural or training differences influencing their reliability to noisy inputs.

3.3. Contextual Awareness

This third dimension of our investigation evaluated the contextual awareness of Gemini and ChatGPT, specifically examining their ability to appropriately adapt their diagnostic outputs in response to the introduction of clinically relevant contextual patient information. Our findings indicate that while both LLMs are influenced by context, ChatGPT demonstrated a higher degree of responsiveness in terms of overall diagnostic change, but this was not unequivocally associated with superior clinical appropriateness.

- **Quantitative Context Influence Rates: ChatGPT Demonstrates a Higher Rate of Diagnostic Modification with Contextual Enrichment:** As numerically summarized in **Table 6** and visually compared in **Figure 5**, our analysis revealed that **Chat- GPT exhibited a Context Influence Rate of 77.8%**, which is significantly higher than **Gemini's Context Influence Rate of 55.6%**. This core quantitative finding indicates that across the nine contextually varied scenarios tested, ChatGPT modified its initial diagnosis in a greater proportion of cases (approximately 78%) compared to Gemini (approximately 56%) when presented with clinically pertinent contextual information.

LLM	Context Influence Rate (%)
Gemini	55.6
ChatGPT	77.8

Table 6: Context Influence Rates (Diagnosis Change due to Contextual Information) of LLMs

- **Concise Methodological Foundation for Contextual Awareness Assessment:** The Context Influence Rates were derived from a targeted protocol (detailed in Methodology) using nine contextually varied scenarios, developed from two baseline cases. Specifically, **four contextually varied versions** were created from one baseline case and **five** from another, totaling nine enriched scenarios. These variations systematically incorpo- rated clinically meaningful modifications across key parameters: **demographics, clinical presentation, medical history/medications**, and **diagnostic test results**. These modifications were designed to create clinically distinct scenarios by altering contextually relevant factors [31].

- **Concise Calculation of Context Influence Rate:** The Context Influence Rate quantifies the *frequency* of diagnostic change due to contextual modifications. For each of the nine scenario sets (from two baseline cases), we compared the LLM's baseline diagnosis to diagnoses from each contextually enriched version. A **diagnosis change** was registered if *any* alteration occurred between the baseline and a contextually enriched scenario within a set. The Context Influence Rate is the percentage of these scenario sets exhibiting at least one such diagnosis change.

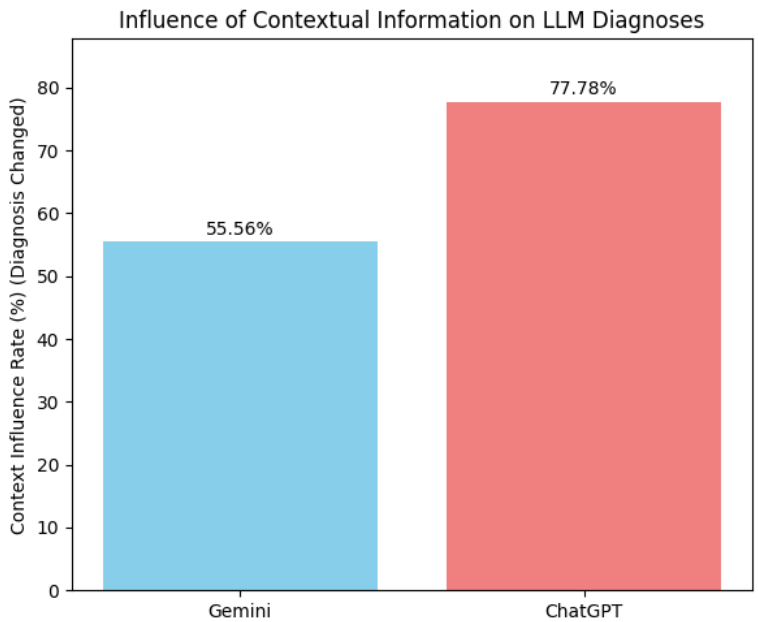

Figure 5: Bar Chart Comparing Context Influence Rates (Diagnosis Change due to Context) of LLMs

- Interpretation of Differential Context Influence Rates: The observed Context Influence Rates indicate that both Gemini and ChatGPT demonstrate a capacity to alter their diagnostic outputs when presented with clinically relevant contextual information. However, ChatGPT's substantially higher Context Influence Rate (77.8% vs. Gemini's 55.6%) suggests a significantly greater tendency to modify its diagnoses in response to the types of contextual variations introduced in our study. This quantitative finding implies that ChatGPT's diagnostic algorithm is, in purely numerical terms, more sensitive and responsive to clinically relevant context, leading to diagnostic changes more frequently than Gemini. It is crucial to emphasize, however, that this quantitative metric alone does not determine the *clinical quality* or *appropriateness* of these context-driven diagnostic changes. A higher rate of change does not automatically equate to improved diagnostic accuracy or clinical reasoning. The subsequent qualitative analysis will critically evaluate the *clinical validity* and *appropriateness* of these context-driven diagnostic adaptations to determine whether this heightened responsiveness translates to clinically sound diagnostic refinement or potentially introduces new forms of error or clinically inappropriate over- sensitivity to context.

3.3.1. Contextual Awareness: Qualitative Analysis

To gain a clinically detailed understanding of the LLMs' contextual awareness, a thorough qualitative review was performed by two independent, board-certified physicians (inter-rater reliability: Cohen's Kappa = 0.85). These expert clinicians evaluated the diagnostic shifts triggered by the introduction of clinically relevant contextual modifications (such as updated lab results, refined patient history, or new imaging findings). Each diagnostic change was then categorized, based on pre-defined criteria and reference standards derived from **UpToDate** and **DynaMed**, as either *Appropriate*, *Inappropriate*, or *Ambiguous*. This process, detailed in the Methodology section, focused on cases exhibiting diagnostic variability across contextually enriched prompts and involved resolving discrepancies through consensus with a third senior clinician to ensure reliable and clinically valid categorizations.

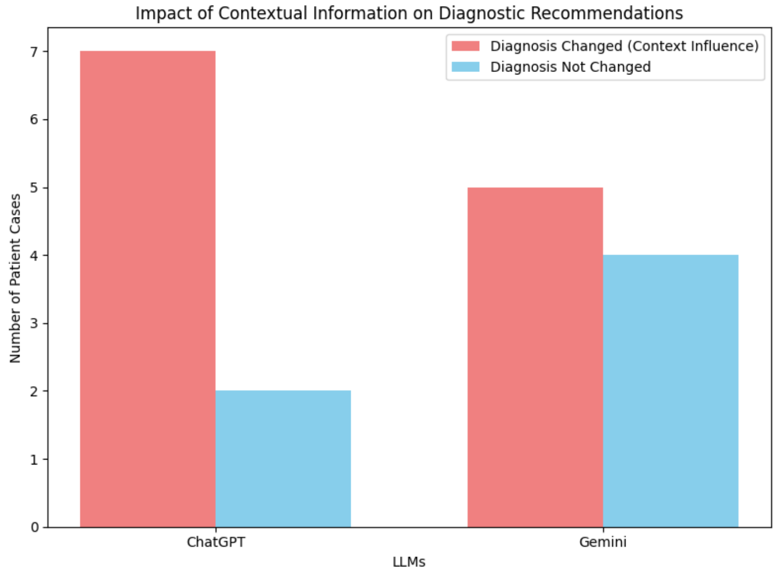

Figure 6: Grouped Bar Chart of Diagnosis Change Status with Contextual Information

3.3.2. Inter-Rater Agreement and Physician Consensus

The qualitative review process was marked by exceptionally strong inter-rater agreement between the two physician reviewers, as evidenced by a Cohen's Kappa coefficient of 0.85. This near-perfect agreement underscores the reliability and consistency of their clinical judgments and the reliability of the qualitative framework employed. Furthermore, in all cases, the physicians reached identical categorizations, reinforcing the consensus-driven nature of the qualitative findings. **Table 7** presents the physician-consensus categorization of context-driven diagnostic changes for Gemini and ChatGPT, providing an overview of the clinical appropriate- ness distribution. **Table 8** offers a more granular view, detailing the individual categorizations (counts and percentages) provided by Physician X and Physician Y for each LLM.

Category	Gemini (%)	ChatGPT (%)
Appropriate Changes	66.7	55.6
Inappropriate Changes	22.2	33.3
Ambiguous Changes	11.1	11.1

Table 7: Consensus Categorization of Context-Driven Diagnostic Changes

Model	Physician	Appropriate	Inappropriate	Ambiguous
Gemini	X	6 (66.67%)	2 (22.22%)	1 (11.11%)
Gemini	Y	6 (66.67%)	2 (22.22%)	1 (11.11%)
ChatGPT	X	5 (55.56%)	3 (33.33%)	1 (11.11%)
ChatGPT	Y	5 (55.56%)	3 (33.33%)	1 (11.11%)

Table 8: Physician-Specific Categorization (Counts/Percentages)

3.3.3. Key Trends in Clinical Appropriateness of Context-Driven Changes

The qualitative analysis revealed distinct trends in the clinical appropriateness of diagnostic modifications driven by contextual information for both Gemini and ChatGPT, highlighting a critical trade-off between responsiveness and clinical soundness.

3.3.4. Prevalence of Appropriate Context-Driven Changes

A significant finding was the demonstration by both LLMs of Appropriate Context- Driven Changes. These were defined as diagnostic refinements that were clinically justifiable, aligned with evidence-based guidelines, and demonstrably enhanced diagnostic accuracy or specificity. As shown in **Table 7** and **Table 8**, Appropriate Changes constituted the largest proportion of context-driven modifications for both models. Gemini exhibited Appropriate Changes in 66.7% of context-driven diagnostic shifts, while ChatGPT demonstrated Appropriate Changes in 55.6% of cases, according to physician consensus.

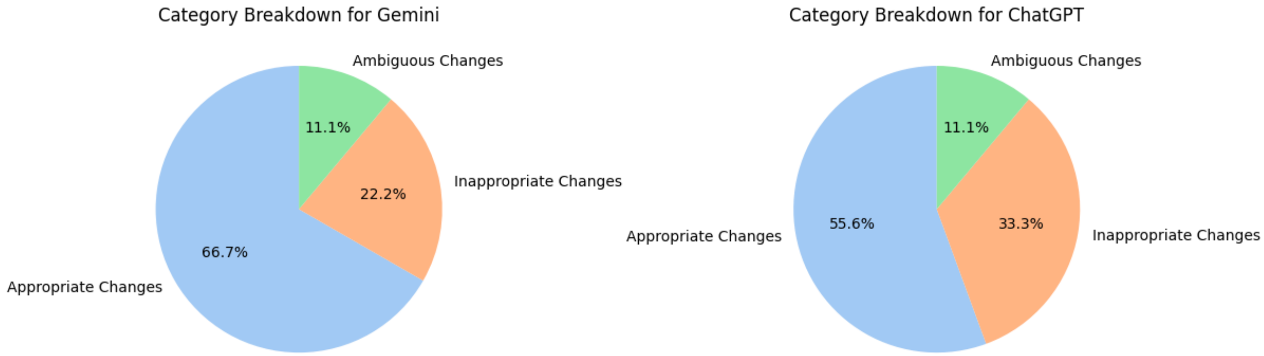

Figure 7: Diagnosis Category Distribution per LLM - pie chart

These findings indicate a fundamental capability in both LLMs to integrate and utilize clinically pertinent contextual information to refine their diagnoses in a clinically meaningful direction. For example, both models effectively demonstrated guideline-concordant diagnostic refinement in scenarios involving acute coronary syndromes. When presented with a base- line case of ***Angina Pectoris***, and then provided with contextual ECG findings indicative of ST-segment depression, both **Gemini and ChatGPT appropriately upgraded the diag- nosis to *Unstable Angina***. This diagnostic shift aligns with established clinical guidelines from organizations like the ACC/AHA, which emphasize ECG findings in risk stratification and management of acute chest pain. This example showcases the LLMs' ability to integrate key diagnostic findings (ECG changes) into the clinical context and appropriately modify their diagnosis towards a more specific and clinically actionable category. Similarly, in scenarios involving infectious diseases, both LLMs showed capacity for appropriate refinement. For in- stance, when given a baseline case suggestive of a ***Viral Upper Respiratory Infection*** and then provided with contextual data indicating ***prolonged fever*** and ***purulent nasal discharge,*** **both Gemini and ChatGPT appropriately shifted their diagnoses to- wards *Bacterial Sinusitis***. This context-driven change reflects sound clinical reasoning by incorporating key symptom characteristics (duration, nature of nasal discharge) to differentiate between viral and bacterial etiologies of upper respiratory complaints, demonstrating an ability to move beyond a general diagnosis to a more specific and clinically relevant bacterial infection diagnosis when warranted by the contextual clinical picture.

3.3.5. Risk of Inappropriate Context-Driven Changes

Despite the demonstrated capacity for appropriate context utilization, a critical finding was the identification of **Inappropriate Context-Driven Changes** in both LLMs. These changes were defined as diagnostic shifts that were clinically unjustified, erroneous, illogical, or led by misinterpretation of the contextual information, ultimately diminishing diagnostic accuracy or clinical relevance. Crucially, **ChatGPT exhibited a higher proportion of Inappropriate Context-Driven Changes (33.3%) compared to Gemini (22.2%)**, according to physician consensus (Table 7 and Table 8). This disparity suggests that while ChatGPT is quantitatively more responsive to contextual cues (as indicated by its higher Context Influence Rate), this heightened responsiveness is accompanied by a greater propensity to make clinically unsound diagnostic modifications.

For example, in one illustrative case (Case ID: 024_4), **ChatGPT erroneously shifted its diagnosis from *Acute Bronchitis* to *Asthma Exacerbation* based solely on the contextual information of a history of childhood asthma**. Physician reviewers deemed this shift clinically inappropriate because it misclassified an *acute infectious process* (bronchitis, typically triggered by viral or bacterial infection) as a *chronic inflammatory condition exacerbation* (asthma), based predominantly on a remote historical factor (childhood asthma history) and neglecting the acute clinical presentation. This inappropriate shift occurred de- spite the absence of key clinical findings typically associated with asthma exacerbation in adults, such as wheezing or spirometric evidence of airflow obstruction in the presented clinical scenario. This example highlights a potential over-reliance of ChatGPT on historical context, even when it leads to clinically illogical diagnostic categorization based on the overall clinical picture. Similarly, **Gemini, in one instance, inappropriately downgraded a diagnosis of *Community-Acquired Pneumonia* to *Bronchiolitis* in an adult patient**, despite contextual chest X-ray findings clearly demonstrating ***consolidation***. This downgrade represents a clinically inappropriate diagnostic shift as ***Bronchiolitis*** is predominantly a diagnosis in infants and young children, characterized by bronchiolar inflammation, and is not clinically consistent with the presence of lobar or segmental consolidation on chest X-ray in an adult patient. This example suggests a potential misapplication of pediatric diagnostic criteria or an under-weighting of critical objective findings (chest X-ray consolidation) in Gemini's con- textual reasoning in this particular scenario. These instances of **Inappropriate Changes** underscore a critical caveat: while contextual awareness is essential for sophisticated clinical reasoning, over-sensitivity to context, especially without reliable clinical validation, can lead to diagnostically unsound and potentially clinically misleading outputs.

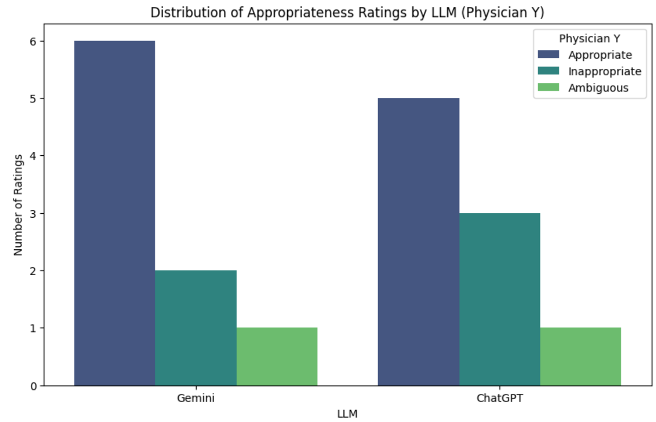

Figure 8: Distribution of Appropriateness Ratings by LLM (Physician Y)

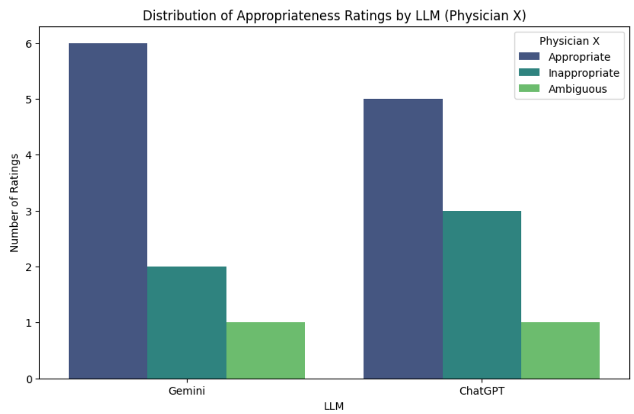

Figure 9: Distribution of Appropriateness Ratings by LLM (Physician X)

3.3.6. Ambiguous Context-Driven Changes

A smaller subset of context-driven diagnostic changes were categorized as **Ambiguous Changes** for both LLMs (11.1% for both Gemini and ChatGPT according to physician con- sensus). These **Ambiguous Changes** represented cases where the clinical appropriateness of the diagnostic shift was not definitively clear-cut based solely on the information provided within the scenario. These were often cases involving subtle clinical presentations where further clinical details, more detailed patient history, or a higher degree of clinical judgment would be necessary to conclusively classify the change as definitively appropriate or inappropriate [32].

For example, Gemini's diagnostic shift from ***Gastroesophageal Reflux Disease (GERD)***

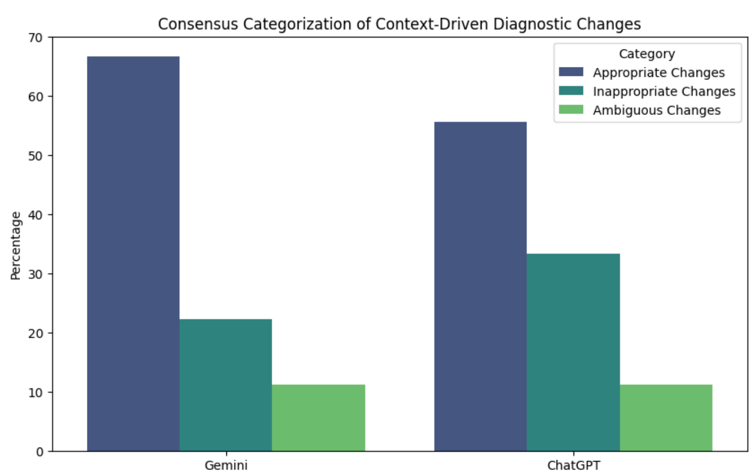

Figure 10: Consistency Comparison: Physician Category Distribution by LLM

to ***Functional Dyspepsia*** in a patient presenting with overlapping symptoms of both conditions was classified as **Ambiguous**. Physician reviewers noted that differentiating between GERD and Functional Dyspepsia can be clinically challenging in certain presentations, often re- quiring detailed dietary history, response to acid suppression therapy, or further investigations. In the absence of such detailed contextual information in the scenario, the appropriateness of this diagnostic shift remained ambiguous, highlighting the inherent limitations of diagnostic AI in scenarios requiring highly subtle clinical data or longitudinal patient assessment.

3.3.7. Differential Handling of Contextual Information

The qualitative analysis also allowed for the identification of patterns in the types of contextual information that were generally well-handled by both LLMs, as well as areas where they exhibited weaknesses or inconsistencies in contextual reasoning.

- **Contexts Generally Well-Handled:**

Integration of Biomarker Trends: Both Gemini and ChatGPT reliably integrated and appropriately responded to trends in biomarker data. For example, in scenarios involving acute kidney injury, both models appropriately escalated diagnostic concern when presented with contextual information **indicating rising crea- tinine levels**, demonstrating effective use of longitudinal biomarker data to refine diagnostic assessment.

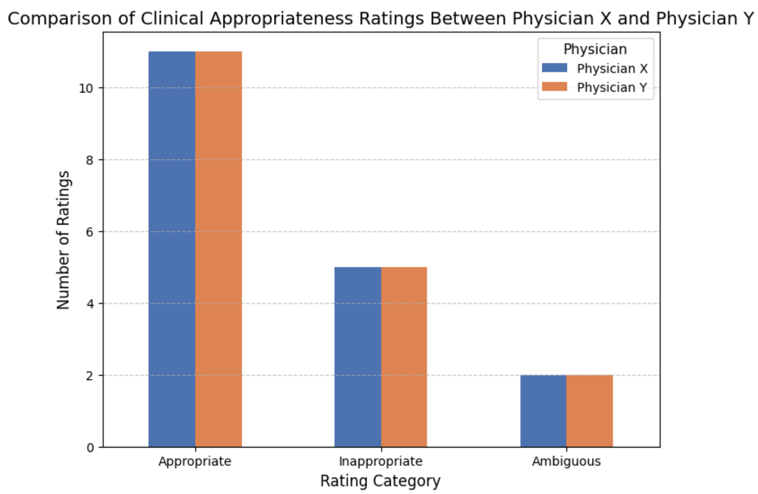

Figure 11: Comparison of Physician X and Physician Y Ratings

- **Utilization of Imaging Correlates:** Both LLMs effectively utilized imaging findings to guide diagnostic shifts in clinically appropriate directions. For instance, in cases presenting with potential pulmonary embolism, the provision of contextual **CT evidence of pulmonary embolism** appropriately prompted diagnostic escalation towards pulmonary embolism as the primary diagnosis for both models, showcasing effective integration of objective imaging data into their diagnostic reasoning process.

- **Contexts Less Reliably Handled:**
  - **Overemphasis on Social History:** In some scenarios, contextual **social history** information, such as **homelessness**, appeared to lead to unwarranted diagnostic shifts towards potentially stigmatizing diagnoses, such as **substance-related psychosis**. This suggests a potential bias or over-reliance on social determinants of health context, leading to diagnostic choices that may not be clinically justified based on the core clinical presentation and objective findings.
  - **Overinterpretation of Non-Specific Vital Signs:** Both models, in some in- stances, exhibited a tendency to **overinterpret non-specific vital signs**, such as **tachycardia**, as strong indicators of specific conditions like **sepsis**, even in the absence of other corroborating clinical data or specific infectious signs. This suggests a potential over-sensitivity to isolated abnormal vital signs, without sufficient weighting of overall clinical context and the need for more specific evidence to support high-acuity diagnoses like sepsis [33].

3.3.8. Balancing Diagnostic Responsiveness and Clinical Soundness

The qualitative physician review underscores a critical insight: while ChatGPT demonstrated a quantitatively higher Context Influence Rate, indicating greater overall responsiveness to contextual information, this heightened responsiveness was accompanied by a proportion- ally higher rate of clinically *inappropriate* diagnostic changes compared to Gemini. Conversely, Gemini, while exhibiting a lower overall quantitative responsiveness to context, demonstrated a comparatively higher proportion of clinically *appropriate* and clinically justified context-driven diagnostic refinements.

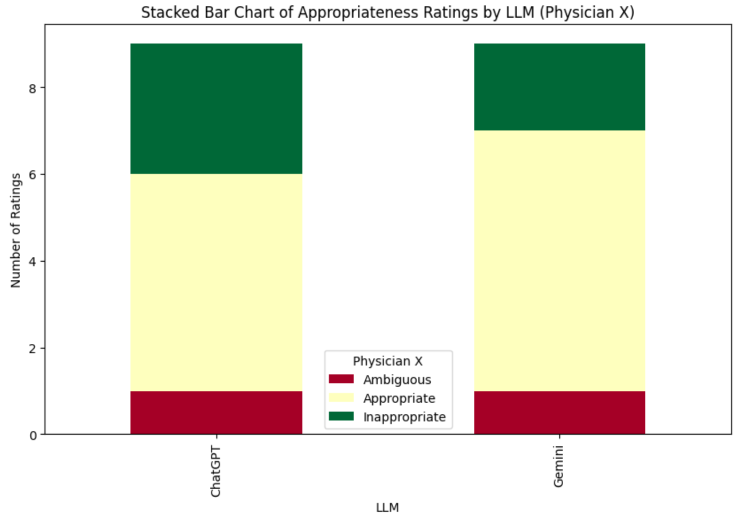

Figure 12: Stacked Bar Chart of Appropriateness Ratings by LLM (Physician X)

This divergence highlights a fundamental trade-off in the design and application of LLMs for clinical diagnostic support: **achieving an optimal balance between diagnostic responsiveness and clinical soundness**. While contextual awareness and adaptability are crucial for sophisticated clinical reasoning and emulating human-like diagnostic judgment, excessive sensitivity to contextual cues, particularly without reliable mechanisms for clinical validation and evidence-based weighting, can inadvertently amplify clinically irrelevant variability and increase the risk of diagnostically inaccurate or even misleading outputs.

The findings from this qualitative analysis strongly emphasize the critical need for incorporating **contextual guardrails** in the development and deployment of LLMs for clinical diagnostic applications. These guardrails are essential to ensure that diagnostic adaptability and responsiveness are tightly coupled with evidence-based clinical reasoning and guideline adherence, preventing over-interpretation of peripheral context and promoting diagnostic shifts that are not only responsive to new information but also consistently valid and clinically jus- tifiable. Future research and development efforts should prioritize the exploration of hybrid approaches that effectively integrate the benefits of contextual awareness with stricter validity checks, clinical reasoning constraints, and mechanisms to filter clinically irrelevant or misleading contextual noise. This will be crucial to harness the potential of LLMs for enhanced clinical decision support while mitigating the risks associated with unchecked contextual sensitivity and ensuring that LLM-driven diagnostic adaptations consistently improve, rather than com- promise, the quality and safety of patient care [34, 35].

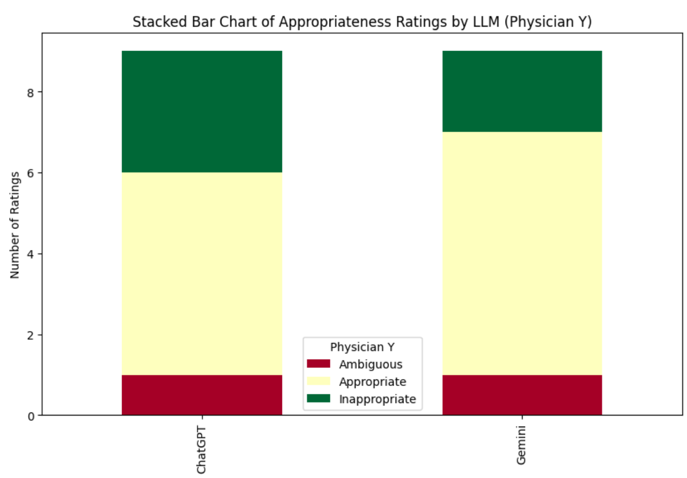

Figure 13: Stacked Bar Chart of Appropriateness Ratings by LLM (Physician Y)

In conclusion, our comprehensive, three-dimensional evaluation of Gemini and ChatGPT's diagnostic reliability reveals an intricate performance profile with critical implications for their clinical application. Across **Diagnostic Consistency**, both LLMs achieved a **perfect 100% Baseline Scenario Consistency Rate**, demonstrating reliable reproducibility in providing clinically equivalent diagnoses for consistent clinical inputs. However, in the **Susceptibility to Manipulation** assessment, we observed differential vulnerability. While ChatGPT exhibited a numerically lower **Susceptibility Rate of 30.0%**, Gemini showed a higher rate at **40.0%**, indicating that both models are susceptible to diagnostically irrelevant noise, albeit to varying degrees. Most significantly, the **Contextual Awareness** analysis highlighted a crucial trade-off. Quantitatively, ChatGPT displayed a higher **Context Influence Rate of 77.8%** compared to Gemini's **55.6%**, suggesting greater responsiveness to contextual cues. However, the qualitative physician review revealed that this heightened responsiveness in ChatGPT came at the cost of clinical soundness. While Gemini demonstrated **Appropriate Context-Driven Changes** in a higher proportion of cases (**66.7%**) compared to ChatGPT (**55.6%**), ChatGPT exhibited a notably elevated rate of **Inappropriate Context-Driven Changes (33.3%)**, exceeding Gemini's rate of **22.2%**. This qualitative divergence, evidenced by physician consensus and a near-perfect inter-rater reliability (Cohen's Kappa = 0.85), underscores a key finding: ChatGPT's greater quantitative responsiveness to context is accompanied by a higher risk of clinically unjustified diagnostic modifications. In contrast, Gemini, while less frequently altering diagnoses in response to context, demonstrated a greater tendency for clinically appropriate and justified contextual refinements. Taken together, these results indicate that while both Gemini and ChatGPT possess strengths in diagnostic consistency and context integration, their vulnerabilities to manipulation and, more critically, the differential clinical appropriateness of their context-driven adaptations—with ChatGPT showing a higher risk of clinically unsound changes despite greater responsiveness—necessitate cautious interpretation and implementation in clinical settings. Future applications of these LLMs for diagnostic support must carefully weigh the trade-off between contextual sensitivity and clinical validity, prioritizing strategies that maximize clinically justified adaptations while mitigating the risk of inappropriate diagnostic shifts driven by noise or over-sensitivity to non-critical contextual factors.

4. Discussion

The aim of this study was to evaluate the diagnostic reliability of two prominent Large Language Models (LLMs), Gemini and ChatGPT, by examining their diagnostic consistency, susceptibility to manipulation, and contextual awareness in simulated clinical scenarios [36]. This section interprets the key findings, discusses their implications for the application of LLMs in healthcare, acknowledges the study's limitations, and suggests directions for future research.

4.1. Diagnostic Consistency

Both LLMs demonstrated **100% consistency** across 52 clinical scenarios, reflecting their deterministic nature and reproducibility under controlled conditions. This outcome highlights a fundamental characteristic of these Large Language Models (LLMs): their capacity to consistently produce identical diagnostic outputs when presented with clinically equivalent input data repeatedly. This level of unwavering consistency warrants careful consideration in the context of clinical applications.

- **Implications of Perfect Diagnostic Consistency:** The achievement of 100% diagnostic consistency by both Gemini and ChatGPT underscores the inherent reliability and exceptional reproducibility of their underlying diagnostic algorithms, at least within the controlled parameters of our experimental design. This stability suggests that, under conditions of consistent and clinically equivalent but superficially varied patient input data, these LLMs operate in a deterministic manner, processing information and arriving at diagnostic conclusions with remarkable fidelity each time the clinically equivalent scenario is presented. From an engineering standpoint, such reliable reproducibility is generally viewed as a highly desirable attribute, indicating a predictable and stable system behavior, essential for dependable application.
- **Crucial Caveat: Consistency Does Not Equate to Accuracy:** While the perfect consistency observed is a notable strength, it is critically important to emphasize that di- agnostic consistency should not be misconstrued as a guarantee of diagnostic *accuracy*. An LLM could, with perfect consistency, reproducibly generate an *incorrect* diagnosis. Our study, in this initial phase, deliberately focused on evaluating the reliability of LLMs, specifically their consistency, rather than examining the absolute correctness or clinical validity of their diagnostic outputs [37]. The crucial next step in evaluating the clinical utility of these models necessitates thorough, accuracy-focused validation studies, comparing LLM diagnoses against established clinical gold standards and expert physician consensus. Furthermore, the deterministic nature that underpins this consistency also implies that any biases or inaccuracies embedded within the LLMs' training data will also be consistently propagated in their outputs.
- **Possible Explanations for Uniform Consistency:** The observed perfect diagnostic consistency is likely attributable to the inherent deterministic nature of contemporary LLMs, particularly when functioning within the controlled and well-defined parameters of a simulated experimental environment. Unlike human clinicians, whose diagnostic judgments can exhibit inherent variability influenced by factors such as cognitive load, subjective interpretation of detailed clinical data, potential biases, and even transient states of fatigue, LLMs, under these controlled conditions, operate primarily through algorithmic processing. Given identical input prompts, the fixed parameters and algorithmic processes within these models predictably lead to identical outputs in each instance, resulting in the 100% consistency we observed. This algorithmic determinism is a core characteristic of their current architecture and operational mode.
- **Contrasting LLM Consistency with Human Clinician Variability:** When considering diagnostic consistency, a significant contrast emerges between LLMs and human clinicians. While direct, head-to-head comparisons are complex and context-dependent, it is well-established that human clinician diagnostic consistency is not absolute and exhibits inherent variability. This variability arises from numerous factors, including the complexity and ambiguity of clinical cases, differences in individual clinician experience and expertise, evolving interpretations of clinical guidelines

and evidence, and unavoidable inter-observer variations in the subjective assessment of clinical signs and symptoms. Indeed, studies on inter-rater reliability in clinical diagnosis, even among expert physicians in well-defined specialities, often reveal agreement rates ranging from approximately 60% to 80% for complex or detailed conditions. In this context, the 100% consistency demonstrated by Gemini and ChatGPT in our study represents a level of output stability and reproducibility that may surpass that achievable by human clinicians in similar, repeated diagnostic tasks when provided with strictly identical information. However, it is paramount to remember that human diagnostic expertise is far more multifaceted than simply applying consistent algorithms. It encompasses critical elements absent in current LLMs, including subtle clinical judgment honed by years of experience, adaptive reasoning in the face of novel or incomplete information, crucial ethical considerations in patient care, and the capacity for empathy and effective patient communication.

4.2. Susceptibility to Manipulation

Our investigation into the susceptibility of LLMs to manipulation, through the introduction of irrelevant yet plausible information into clinical prompts, revealed that both models are vulnerable to such influence, albeit to differing degrees. Quantitatively, **Gemini exhibited a higher susceptibility rate of 40%, compared to ChatGPT's 30%**. These rates, while seemingly moderate, indicate a clinically significant vulnerability, highlighting a potential fragility in the diagnostic reasoning of these LLMs when confronted with subtly altered or noisy input data.

- **Clinical Significance of Susceptibility Rates and Diagnostic Examples:** Even a 30-40% susceptibility rate to manipulation is not inconsequential in a clinical context. It implies that in a non-trivial proportion of cases, seemingly minor or irrelevant modifications to patient histories while keeping the core clinically relevant information consistent, can lead to changes in the LLM's diagnostic output. These changes are not merely academic; they can represent shifts with profound clinical implications. For example, in our study, manipulation led Gemini to revise a diagnosis of ***Acute ST-elevation my-ocardial infarction (STEMI)*** – a critical cardiac emergency – to ***Angina pectoris*** – a condition requiring different and less urgent management. Similarly, ChatGPT, un- der manipulation, shifted a diagnosis of ***Acute Exacerbation of COPD to Heart Failure Exacerbation*** – two distinct respiratory conditions with divergent therapeutic pathways. These concrete examples underscore that manipulation-induced diagnostic changes are not always subtle refinements; they can involve clinically significant alterations with direct implications for patient care pathways and treatment decisions.
- **Qualitative Insights into LLM Reasoning Vulnerabilities: The qualitative analysis of diagnosis changes induced by manipulation provided valuable insights into the underlying reasoning vulnerabilities of the LLMs. Several key patterns emerged from the physician review:**
- **Over-Reliance on Key Phrases and Underweighting of Holistic Con- text:** A recurrent theme was both models' apparent **over-reliance on specific keywords or phrases** within the clinical prompts, sometimes at the expense of a more holistic and balanced consideration of the entire clinical context. For instance, the presence of phrases like **chest pain radiating to the left arm** seemed to disproportionately anchor the LLMs towards cardiac diagnoses, even when other clinical features might have suggested alternative etiologies. Conversely, subtle contextual intricacies, which a human clinician would typically integrate into a broader clinical picture, appeared to be underweighted or missed by the LLMs.
- **Manifestation of Anchoring Bias:** The manipulation experiments also revealed a form of **anchoring bias** in both LLMs' diagnostic processes. The manipulated, often irrelevant, inputs seemed to **anchor** the LLMs to peripheral details or subtly altered aspects of the clinical presentation. This phenomenon mirrors the well-documented cognitive bias in human clinicians where initial pieces of information can disproportionately influence subsequent judgments, hindering objective re-evaluation. However, unlike human clinicians who possess metacognitive abilities and can be trained to recognize and mitigate anchoring biases, the LLMs in our study demonstrated less capacity for self-correction or critical appraisal of the manipulated inputs, treating all prompt information as equally valid and relevant.
- **LLM Susceptibility Compared to Human Clinician Strengths and Weak- nesses:** When comparing LLM susceptibility to manipulation with the known vulnerabilities of human clinicians, several key distinctions emerge, as summarized in the table below:

Factor	LLM Weaknesses	Human Clinician Strengths	Human Clinician Weaknesses
Anchoring Bias	**Highly susceptible** to leading or anchoring prompts.	**Resists bias** through clinical protocols and awareness.	**Susceptible**, but can be mitigated with training and awareness.
Contextual Gaps	**Struggles** with missing, conflicting, or irrelevant data.	**Probes actively** for missing details (history, tests, etc.).	May make assumptions based on incomplete or conflicting data.
Red Herrings (Irrelevant Info)	**Overweights irrelevant information** in prompts.	**Filters irrelevant information** effectively using clinical reasoning and judgment.	Can be distracted by salient but irrelevant findings.
Diagnostic Certainty/- Calibration	May **overstate confidence** without sufficient evidence.	**Expresses uncertainty** appropriately and orders further tests when needed.	May exhibit overconfidence or premature closure without sufficient data.

Input Validity Assessment	**Lacks ability to question input validity**; treats all input as factual.	**Critically evaluates input validity,** seeks corroboration, and questions inconsistencies.	May be influenced by source credibility bias.

Table 9: Comparison of LLM and Human Clinician Di-agnostic Factors

As highlighted in this comparison, LLMs exhibit a pronounced weakness in their inability to critically evaluate the validity or relevance of input information. Human clinicians, through training and experience, develop crucial metacognitive skills that allow them to identify inconsistencies, question the reliability of data sources, and actively seek clarifying information when faced with ambiguous or potentially mis- leading clinical presentations. LLMs, in their current form, lack this critical appraisal ability. They process all prompt information as factual and relevant, rendering them particularly vulnerable to adversarial or unintentional manipulation through the in- troduction of irrelevant or misleading details. In contrast, human clinicians, while also susceptible to cognitive biases, possess the capacity for critical reasoning, contextual filtering, and iterative questioning that can mitigate the impact of manipulation and enhance the reliability of their diagnostic judgments.

- **Implications for Trust, Safety, and Responsible Clinical Integration:** The demonstrated susceptibility of LLMs to manipulation has significant implications for trust and safety considerations surrounding their potential integration into clinical diagnostic workflows. If even subtle, irrelevant manipulations can induce clinically meaningful diagnostic changes, it underscores the critical imperative for caution and reliable safeguards. **Uncritical reliance on LLM diagnostic outputs, without close clinician oversight and validation, would be imprudent and potentially unsafe**. Furthermore, the vulnerability to manipulation raises concerns about the potential for malicious exploitation. Intentionally crafted, subtly misleading prompts could be designed to steer LLMs towards erroneous diagnoses, potentially for nefarious purposes. Therefore, ensuring the safe and responsible application of LLMs in healthcare necessitates the implementation of multi-layered safeguards, in- cluding careful input data validation and pre-processing; output verification and clinical plausibility checks by human experts; reliable clinician-in-the-loop oversight throughout the diagnostic process; and ongoing research into methods for enhancing LLM reliability against both intentional and unintentional manipulation. LLMs, in this context, should be conceptualized and deployed as powerful *tools* to augment, rather than replace, the essential critical thinking, contextual understanding, and ethical judgment of human clinicians in medical diagnosis.

4.3. Contextual Awareness

Our investigation into contextual awareness, examining how Gemini and ChatGPT utilize clinically relevant contextual variations to modify their diagnoses, revealed a subtle landscape of capabilities and limitations. Quantitatively, **ChatGPT exhibited a significantly higher Context Influence Rate (77.8%) compared to Gemini (55.6%)**. This finding under- scores fundamental differences in their contextual reasoning and raises important questions about the clinical utility of context integration in current LLM diagnostic applications.

- **Interpreting Context Influence Rates: Responsiveness vs. Clinical Validity:** The higher Context Influence Rate for ChatGPT initially suggests a greater sensitivity to contextual information. However, it is crucial to understand that a high Influence Rate, in isolation, does not equate to better diagnostic quality. It merely indicates a greater *propensity* for diagnostic change in response to context. The critical question is whether these context-driven changes are clinically *appropriate* and lead to more accurate or safer diagnoses. Our qualitative physician review directly addressed this question of clinical validity, revealing a more complex picture than the quantitative rates alone suggest. As summarized in **Table 10**, a higher Context Influence Rate for ChatGPT was accompanied by a *lower* percentage of Appropriate Changes and a *higher* percentage of Inappropriate Changes compared to Gemini.

Feature	Gemini (55.56% CIR)	ChatGPT (77.78% CIR)	Interpretation
Context Influence Rate	55.56%	77.78%	ChatGPT is quantitatively more responsive to contextual changes.
% Clinically Appropriate Changes	66.7%	55.6%	Gemini demonstrates a higher proportion of clinically justified context driven refinements, despite a lower overall influence rate.
% Clinically Inappropriate Changes	22.2%	33.3%	ChatGPT exhibits a higher proportion of clinically inappropriate diagnostic shifts in response to context, suggesting a potential trade-off between responsiveness and clinical soundness.

Table 10: Quantitative Context Influence vs. Qualitative Appropriateness

• **Qualitative Analysis:** The qualitative physician review revealed that both models demonstrated the capability for **Clinically Appropriate Context-Driven Diagnostic Refinements**. These instances, detailed with examples in **Table 11**, showcase how LLMs can utilize context to enhance diagnostic specificity and, in some cases, align with clinical guideline recommendations.

LLM	Case ID	Original Diagnosis	Contextual Information Added	Context- Driven Diagnosis Refinement	Clinical Benefit
Gemini	**001_3_2**	Angina Pectoris	ECG findings suggestive of instability	Unstable Angina	More accurate risk stratification and appropriate escalation of care.
ChatGPT	**024_1**	Peptic Ulcer Disease	Context of gastrointestinal bleeding	Gastric Ulcer with Gastrointestinal Bleeding	Improved specificity, guiding appropriate management of potential complications.
Gemini ChatGPT	**Multiple**	Initial Broad Diagnosis	Lifestyle factors, demographic details	More Specific Subtype or Clinical Qualifier Added	Enhanced diagnostic precision, potentially informing personalized treatment strategies.

Table 11: Examples of Clinically Appropriate Context- Driven Diagnostic Refinements

This table provides illustrative examples of cases where both Gemini and ChatGPT appropriately refined their initial diagnoses based on added contextual information. These examples demonstrate the potential of LLMs to leverage context for clinically justifiable diagnostic enhancement, such as improving diagnostic specificity and aligning with clinical guidelines.

LLM	Case ID	Original Diagnosis	Contextual Information Added	Context-Driven Diagnosis Shift	Clinical Concern
Gemini	**001_3_3**	Atypical MI (NSTEMI)	Recent history of "heavy lifting"	Musculoskeletal Back Pain	Dangerous shift from a critical cardiac condition to a musculoskeletal issue, disregarding cardiac markers.
ChatGPT	**024_4**	Acute Bronchitis	Childhood asthma history	Asthma Exacerbation	Misclassification of infection as chronic inflammation, potentially leading to incorrect treatment.
Gemini	**024_3**	Peptic Ulcer Disease	Vague symptoms, no confirmatory tests	Possible Malignancy (Gastric Cancer)	Overconfidence in ambiguity, suggesting serious diagnosis without biopsy emphasis.
Both	**Multiple**	Diagnosis requiring test results	Scenarios designed to necessitate data	Diagnosis without requesting tests	**Critical Data Handling Deficiency:** LLMs attempt diagnoses without essential test results, unlike clinicians.

Table 12: Examples of Clinically Inappropriate Shifts and Diagnostic Overconfidence

This table provides illustrative examples of clinically concerning context-driven diagnostic shifts observed in Gemini and ChatGPT. These examples highlight critical limitations, including illogical diagnostic changes (system-hopping errors), diagnostic overconfidence in ambiguous scenarios, and a failure to request essential test results before making diagnoses—raising concerns about the reliability and safety of LLM contextual reasoning in clinical settings.

- However, a critical limitation emerged in the form of **Clinically Inappropriate Shifts, System-Hopping Errors, and Diagnostic Overconfidence**. As quantified and exemplified in **Table 10** and **Table 12**, both models, particularly ChatGPT, tended to diagnostically illogical or clinically unsupported changes in response to context. Further- more, a significant concern was the observation of **Diagnostic Overconfidence**, where LLMs rendered definitive diagnoses even in the absence of crucial information, such as essential test results, instead of appropriately requesting further data.
- **LLM Contextual Understanding vs. Human Clinical Expertise:** These qualitative findings point to fundamental differences between LLM contextual reasoning and human clinical expertise. Human clinicians integrate context in a hierarchical manner, prioritizing clinically significant data and filtering irrelevant noise. They also exhibit crucial **diagnostic restraint**, explicitly recognizing and addressing data gaps, particularly the need for essential test results before reaching definitive diagnoses. In contrast, as summarized in **Table 13**, current LLMs demonstrate limitations in hierarchical weighting, noise filtering, and, critically, data handling [37].

Aspect	LLMs	Human Clinicians
Data Reliance	Static training data; no real-time updates.	Dynamic integration of evolving guidelines, patient feedback, and new evidence.
Hierarchical Weighting	Inconsistent prioritization; may conflate relevant and irrelevant details.	Expert prioritization; e.g., lab results carry more weight than social history in cardiac cases.
Clinical Prioritization	Focuses on terminology and statistical associations.	Prioritizes life-threatening conditions (e.g., ruling out ACS before GERD).
Contextual Integration	**Flat** attention mechanisms; lacks domain-specific heuristics.	Hierarchical integration; leverages experiential learning and domain expertise.
Red Herring Filtering	Limited ability to filter distractions; may overemphasize irrelevant details.	Effective filtering; pattern recognition helps prioritize relevant data.
Bias Resistance	Vulnerable to leading prompts or misinformation.	Uses protocols to minimize cognitive biases.
Input Assessment	Accepts all inputs uncritically.	Validates data via history-taking and tests.
Ambiguity Handling	Generates speculative diagnoses without flagging uncertainty.	Explicitly acknowledges ambiguity and orders confirmatory tests.
Ambiguity Resolution	Generates guesses without flagging gaps.	Proactively seeks missing data (e.g., labs).
Error Patterns	System-hopping errors (e.g., cardiac → musculoskeletal).	Anchoring biases or overreliance on experience.
Error Correction	Cannot self-audit; repeats training patterns.	Revises diagnoses iteratively with new data.
Diagnostic Appropriateness	Higher rate of inappropriate diagnostic adjustments (ChatGPT in this study).	Clinically validated contextual adjustments; guided by protocols and experience.
Data Handling Diagnostic Restraint	Deficient: Often diagnose without essential test results, lack request for data.	Reliable: Defer diagnosis pending essential data, explicitly request and prioritize test results.

Table 13: Comparison of LLMs and Human Clinicians in Contextual Reasoning

- **Implications for Clinical Translation:** The limitations highlighted, particularly the data handling deficiencies and the propensity for clinically inappropriate shifts, underscore the critical need for targeted development and strict validation before LLMs can be responsibly deployed in clinical diagnostic settings. Future research and development must prioritize enhancing the **discriminative contextual reasoning** of LLMs, enabling them to: hierarchically weight contextual information; filter irrelevant noise; recognize and address data gaps; and demonstrate appropriate diagnostic restraint, including the explicit request for and integration of essential test results. **Until these critical limitations are substantially mitigated, LLMs should be considered investigational assistive tools only, requiring stringent clinician oversight and serving to augment, not replace, human clinical judgment in diagnostic workflows.** The focus must shift towards developing LLMs that not only respond to context quantitatively but also integrate and utilize contextual information in a manner that is demonstrably safe, clinically valid, and aligned with the principles of expert human diagnostic reasoning.

4.4. Integrated Discussion and Broader Implications

This study undertook a comprehensive, multi-dimensional evaluation of the diagnostic re- liability of LLMs, examining their consistency, susceptibility to manipulation, and contextual awareness within simulated clinical scenarios. Synthesizing our findings across these three critical axes reveals a complex and paradoxical performance profile: while LLMs exhibit remarkable algorithmic consistency, this very determinism paradoxically underpins vulnerabilities in reliability, contextual reasoning, and ultimately, clinical appropriateness. Our integrated analysis illuminates both the promising avenues and the inherent perils of deploying current LLM technology in safety-critical diagnostic applications, underscoring the crucial imperative for cautious, human-guided, and ethically grounded implementation.

4.5. Synthesizing Across Consistency, Susceptibility, and Context

The most salient overarching finding is the **fundamental decoupling of diagnostic consistency from clinical validity and reliability**. The 100% diagnostic consistency demonstrated by both Gemini and ChatGPT across repeated presentations of identical clinical scenarios initially suggests a bedrock of reliability. However, this algorithmic stability proves to be a double-edged sword, inadvertently amplifying inherent vulnerabilities rather than guaranteeing clinical safety.

- **Consistency Paradox:** The observed perfect consistency, while highlighting the impressive reproducibility of LLM algorithms, simultaneously unmasks a critical paradox: **deterministic outputs can magnify inherent vulnerabilities**. As our manipulation experiments revealed, even seemingly minor alterations, such as the introduction of irrelevant but plausible symptoms, could induce clinically significant diagnostic shifts in a substantial proportion of cases (30-40%). This underscores that the very algorithmic determinism that ensures consistency also renders LLMs less adaptable and resilient to noisy or adversarial inputs. Unlike human clinicians who can leverage experience and contextual understanding to filter out noise and re-evaluate initial assumptions, LLMs, bound by their deterministic frameworks, tend to process all inputs with equal weight, lacking the sophisticated critical appraisal necessary to discern valid signals from irrelevant perturbations. Thus, consistency, in isolation, offers no guarantee of diagnostic safety and may, in fact, become a liability by consistently propagating errors embedded in training data or introduced via manipulation.

- **Contextual Responsiveness:** Our investigation into contextual awareness reveals a detailed trade-off in diagnostic performance. ChatGPT's higher Context Influence Rate (77.8% vs. Gemini's 55.6%) suggests superior contextual integration. However, this increased responsiveness comes at a cost.

While ChatGPT effectively refined diagnoses based on context—such as escalating **Pep- tic Ulcer Disease to Gastric Ulcer with Bleeding** when warranted—it also demonstrated a higher frequency of clinically *inappropriate shifts*. For instance, ChatGPT's overconfident diagnosis of **Asthma Exacerbation** for a bronchitis case, based solely on a history of childhood asthma, illustrates the risks of uncalibrated contextual sensitivity.

In contrast, Gemini, with its more conservative Context Influence Rate, paradoxically exhibited a higher percentage of clinically *appropriate* context-driven changes. This sug- gests a critical trade-off: **Greater sensitivity to context, without a corresponding ability to prioritize information hierarchically and apply clinically sound judgment, can increase the risk of diagnostic errors and inappropriate variability**.

The **flat** attention mechanisms of current LLMs—lacking the domain-specific heuristics and hierarchical weighting of clinical information that human clinicians develop through experiential learning—likely contribute to this challenge.

- The Fragility Triad - A Framework for Understanding LLM Limitations: Synthesizing across these dimensions, we identify a Fragility Triad that characterizes the limitations of current LLMs in clinical diagnosis:
  - Deterministic Consistency Entrapment: Algorithmic consistency, while ensuring reproducibility, can entrench biases, amplify vulnerabilities, and hinder adaptive reasoning in the face of novel or noisy data.
  - Manipulation Susceptibility Reflects Input Blindness: Vulnerability to manipulation, even through subtle and clinically irrelevant input changes, reveals a fundamental lack of critical input validity assessment and an over-reliance on surface- level prompt information.
  - Contextual Reasoning Gaps: Noise Amplification and Hierarchical Blind- ness: Limitations in contextual awareness stem from a failure to effectively discriminate between clinically relevant and irrelevant contextual details, often overemphasizing less critical cues while under-weighting or disregarding diagnostically vital information. This **flat"** attention and lack of hierarchical contextual integration lead to both missed opportunities for valid diagnostic refinement and increased risks of inappropriate, context-driven diagnostic shifts.

This **Fragility Triad** underscores that deploying current LLMs in clinical settings involves navigating the inherent risks of applying rigid, algorithm-driven systems within the inherently dynamic, ambiguous, and noise-rich environment of real-world medical practice.

4.6. Implications for Clinical Use

Our findings carry significant implications for the responsible translation of LLMs into clinical diagnostic workflows, highlighting a careful balance between potential benefits and inherent risks.

- **Potential Benefits: Streamlining Routine Tasks and Enhancing Standardization:** The reliable diagnostic consistency demonstrated by LLMs suggests valuable applications in **streamlining standardized diagnostic tasks**, particularly in high-volume, low-complexity scenarios. LLMs could contribute to **standardizing diagnostic terminology**, enhancing **guideline adherence in routine cases**, and assisting in **preliminary triage** by efficiently processing and categorizing straightforward clinical presentations. In settings with high clinician workloads or resource constraints, LLMs could potentially augment diagnostic efficiency in handling stable or uncomplicated cases, freeing up clinician time for more complex or ambiguous patient encounters.
- **Key Risks: Misdiagnosis, Overconfidence, and Erosion of Clinical Judgment:** However, the identified vulnerabilities and limitations present substantial risks that out- weigh potential efficiency gains if LLMs are deployed without strict safeguards and a clear understanding of their current limitations. **The susceptibility to manipulation poses a direct patient safety risk, highlighting the potential for both unintentional errors and malicious exploitation to induce clinically significant misdiagnoses**. The observed diagnostic overconfidence in ambiguous scenarios, coupled with a failure to request essential test results, further amplifies the risk of premature diagnostic closure and potentially inappropriate treatment decisions. Moreover, the documented instances of clinically inappropriate context-driven shifts, particularly the **system-hopping** errors and misclassification of serious conditions, underscore the potential for LLMs to intro- duce new error modes into the diagnostic process, potentially delaying or undermining life-saving interventions. Uncritical reliance on LLM outputs, particularly in complex or detailed cases, could inadvertently **erode critical clinical judgment skills** and create a false sense of security, ultimately compromising patient safety.
- **Practical Considerations for Responsible Deployment: The Imperative of Hu- man Oversight and Reliable Safeguards:** Responsible and ethical deployment of LLMs in clinical diagnosis demands a cautious, incremental, and thoroughly controlled approach, prioritizing patient safety

and maintaining the centrality of human clinical expertise. Key practical considerations include:

- ◦ **Clinician-in-the-Loop Validation and Oversight: Strict clinician-in-the- loop oversight is not merely advisable, but an absolute prerequisite.** LLM outputs must undergo mandatory validation and contextualization by human clinicians, particularly in all but the most routine and low-risk scenarios. Safeguards must be implemented to flag inconsistencies, overconfidence, data gaps, and potentially inappropriate diagnostic shifts.
- ◦ **Reliable Input Sanitization and Adversarial Defense:** Developing and implementing reliable input sanitization and pre-processing tools to detect and filter irrelevant, misleading, or adversarial inputs is crucial to mitigate manipulation risks. Research into adversarial training and other reliability-enhancing techniques is essential to strengthen LLM resilience against malicious or unintentional prompt perturbations. **Transparency and Explainability as Cornerstones of Trust:** Transparency regarding LLM limitations, potential error modes, and the inherent uncertainty in AI-driven diagnostic suggestions is paramount for building appropriate trust and guiding responsible utilization. Developing Explainable AI (XAI) approaches to elucidate LLM reasoning pathways and provide clinicians with insights into the factors driving diagnostic outputs is crucial for fostering informed clinician oversight and promoting appropriate calibration of trust. Clear documentation of LLM capabilities and, critically, their data handling deficiencies (e.g., inability to actively request tests) is essential to prevent over-reliance and misuse.
- ◦ **Focus on Augmentation, Not Autonomous Replacement:** LLMs should be conceptualized, developed, and deployed solely as **assistive tools to augment and enhance human clinical expertise**. The focus must remain on leveraging LLMs' computational strengths to support, rather than supplant, the irreplaceable facets of human clinical judgment, ethical reasoning, empathy, and comprehensive patient care. The aim is to create collaborative clinician-AI diagnostic ecosystems where the strengths of both human and artificial intelligence are synergistically combined, with human clinicians retaining ultimate authority and responsibility for diagnostic accuracy and patient safety.

4.7. Limitations of the Study and Future Research Directions

This study, while providing critical insights into LLM diagnostic reliability, is inherently bounded by certain limitations that inform the direction of future research [38].

- **Scope of Clinical Scenarios and Complexity:** Our evaluation employed a focused set of 52 simulated clinical scenarios, primarily within specific medical domains. Future research must **significantly expand the breadth, depth, and complexity of clinical scenarios** to encompass the full spectrum of medical specialities, varying levels of disease severity and diagnostic ambiguity, rarer conditions, multi-morbidity, and the de- tailed, longitudinal nature of real-world patient presentations. Incorporating cases with higher levels of clinical uncertainty, conflicting data points and evolving symptom pat- terns is essential to more comprehensively evaluate LLM performance in ecologically valid settings.
- **Extend of Manipulation and Context Types:** We explored specific, plausible text- based manipulations (irrelevant symptoms) and contextual variations (demographics, history). Future investigations should **broaden the spectrum of manipulation tactics**, encompassing more subtle semantic distortions, biased framing, adversarial attacks de- signed to specifically exploit known LLM vulnerabilities and real-world misinformation scenarios. Similarly, exploring a **richer array of clinically relevant contextual fac- tors**, including social determinants of health, cultural and linguistic diversity, patient- specific values and preferences, and longitudinal electronic health record data, is crucial for a more holistic assessment of LLM contextual reasoning in diverse clinical populations.
- **Model Diversity and Evolution:** Our study focused on two prominent but proprietary LLMs, Gemini and ChatGPT, at specific points in their development. Future research must **expand the evaluated model landscape** to include a wider array of LLM architectures, open-source models, specialized medical LLMs (such as Med-PaLM), and continuously evaluate the rapidly evolving performance of new model versions and interations. Comparative evaluations across diverse model types and developmental stages are essential to understand the generalizability of our findings and track the trajectory of LLM capabilities and limitations in medical diagnosis.
- **Need for thorough Accuracy Benchmarking and Real-World Outcomes Data:** This study prioritized reliability and reliability metrics (consistency, susceptibility, con- textual influence). While these are crucial foundational assessments, **future research must thoroughly prioritize diagnostic accuracy evaluation**, directly comparing LLM diagnostic outputs against established diagnostic gold standards, expert physician consensus, and, critically, real-world clinical outcomes data (e.g., treatment appropriate- ness, time to diagnosis, patient safety outcomes, cost-effectiveness). Large-scale, prospective studies comparing LLM-augmented vs. standard clinical diagnostic workflows are essential to definitively ascertain their clinical utility and impact on patient care.
- **Translational Research in Real-World Clinical Settings:** The controlled experimental nature of our study necessitates translational research to evaluate LLM performance in **live clinical environments**. Future studies must examine their integration into real-world clinical workflows, their impact on clinician-patient interactions, their usability and acceptability to clinicians and patients, their effects on clinician decision-making processes in complex, time-constrained clinical settings, and their broader impact on healthcare system efficiency, equity, and access to care across diverse healthcare set- tings and patient populations.
- **Developing Reliability-Enhancing and Clinically-Grounded AI Techniques:** An essential direction for future research lies in the **development and strict evaluation of**

novel AI techniques specifically designed to address the identified limitations of current LLMs in clinical diagnosis.

This includes:

- Developing and validating methods to enhance LLM reliability against manipulation and adversarial attacks, potentially through adversarial training, input sanitization, and uncertainty quantification frameworks.
- Engineering more sophisticated contextual reasoning mechanisms that move beyond flat attention to incorporate hierarchical weighting of clinical information, explicit data gap recognition and handling, and clinically grounded causal reasoning approaches.
- Exploring techniques for imbuing LLMs with diagnostic restraint and uncertainty awareness, enabling them to appropriately defer definitive diagnoses pending essential data and to transparently communicate diagnostic uncertainty to clinicians.
- Developing Explainable AI (XAI) methodologies customized to the clinical domain to enhance the transparency and interpretability of LLM diagnostic reasoning, facilitating clinician trust, oversight, and error detection.
- Designing and evaluating clinician-AI collaborative interfaces that optimize the synergistic integration of human clinical expertise and LLM computational strengths, fostering effective human-AI teamwork in diagnostic workflows.
- Establishing reliable ethical frameworks and regulatory guidelines to govern the development, validation, deployment, and ongoing monitoring of LLM-based diagnostic tools, ensuring patient safety, data privacy, algorithmic fairness, and equitable access to the benefits of AI-augmented medical diagnosis [39].

This comprehensive evaluation of Gemini and ChatGPT offers a sophisticated and critical perspective on the current state of LLM-based medical diagnosis. While highlighting the re- markable algorithmic consistency of these models and their potential for standardization and efficiency gains, our findings simultaneously illuminate significant vulnerabilities to manipulation, limitations in clinically sound contextual reasoning, and critical data handling deficiencies. **These limitations unequivocally underscore that current LLMs are not yet suited for autonomous diagnostic decision-making in clinical practice**. Instead, our results advocate for a **cautious and ethically grounded approach, emphasizing responsible augmentation rather than premature automation**. LLMs, in their current form, should be viewed as investigational assistive tools, deployed under stringent clinician oversight, within well-defined use cases, and with continuous monitoring and strict validation in real-world clinical settings [40]. The path forward requires a sustained and concerted research effort, prioritizing the enhancement of reliability, contextual discrimination, clinical validity, and, above all, patient safety, to responsibly unlock the potential of AI to augment and elevate, rather than inadvertently compromise, the essential art and science of human-centred medical diagnosis. The ultimate measure of success will not be algorithmic sophistication alone, but the demonstrable improvement of patient outcomes, the enhancement of clinician expertise, and the equitable advancement of healthcare for all [41].

5. Conclusion

This study undertook a comprehensive, three-dimensional evaluation of Large Language Models (LLMs), specifically Gemini and ChatGPT, as applied to the critical domain of medical diagnostics. Our investigation, designed to examine beyond superficial accuracy metrics, examined the essential dimensions of **diagnostic consistency, susceptibility to manipulation**, and **contextual awareness.** Employing a mixed-methods approach, we integrated quantitative performance metrics with qualitative physician review to furnish a clinically sophisticated and ethically grounded assessment of these transformative technologies.

While our findings reveal a reassuring **algorithmic diagnostic consistency**, with both Gemini and ChatGPT achieving a **perfect 100% Baseline Scenario Consistency Rate**, this isolated strength is overshadowed by clinically critical vulnerabilities uncovered across other reliability dimensions. The **Susceptibility to Manipulation** assessment demonstrated that both models exhibit a concerning **fragility in the face of diagnostically irrelevant noise**, with susceptibility rates reaching 40.0% for Gemini and 30.0% for ChatGPT. This input blindness reveals a fundamental lack of robust input validation mechanisms, allowing superficial prompt alterations to induce clinically significant diagnostic shifts.

Most critically, our in-depth **Contextual Awareness** analysis illuminated not merely quantitative differences but a qualitative divergence in clinical reasoning. While ChatGPT exhibited a higher **Context Influence Rate (77.8% vs. Gemini's 55.6%)**, indicating a greater tendency to alter diagnoses in response to contextual cues, the qualitative physician review exposed a **disturbing trade-off between responsiveness and clinical soundness**. Gemini demon- strated a higher proportion of **Appropriate Context-Driven Changes (66.7%)**, signifying clinically justified diagnostic refinements. In stark contrast, ChatGPT, despite its greater quantitative responsiveness, exhibited a notably elevated rate of **Inappropriate Context- Driven Changes (33.3%)**, exceeding Gemini's 22.2%. This qualitative divergence, validated by near-perfect inter-rater reliability (Cohen's Kappa = 0.85) among expert physician reviewers, underscores a critical insight: **ChatGPT's enhanced quantitative responsiveness to context is significantly undermined by a higher propensity for clinically unjustified and potentially erroneous diagnostic modifications**. These findings, taken together, re- veal a **Fragility Triad deterministic consistency, input blindness, and hierarchical contextual gaps**—demonstrating that reliability in controlled, simplified settings does not translate to clinical safety in complex, real-world diagnostic scenarios.

The implications of these findings are significant and carry a warning for the field. While LLMs offer unprecedented algorithmic consistency and the potential for standardized diagnostic processes, our research demonstrates that their **demonstrated susceptibility to manipulation and fundamental limitations**

in sophisticated, clinically valid contextual reasoning render current LLMs unsuitable for autonomous diagnostic decision-making in clinical practice. These are not minor limitations to be incrementally addressed; they are **systemic deficiencies** that, if unmitigated, pose **unacceptable risks** to patient safety, particularly in resource-constrained settings where robust clinician oversight may be compromised.

Unlike experienced clinicians who iteratively examine, critically weigh evidence, and deeply integrate complex patient contexts, current LLMs consistently **overstate diagnostic certainty without proactively seeking essential clarifying information**, exhibiting a concerning lack of clinical judgment.

Therefore, this study strongly advocates for a **paradigm of cautious and ethically grounded augmentation, rather than premature automation**. LLMs, in their current form, should be viewed as **investigational assistive tools**, deployed solely under clinician oversight, within defined use cases, and subject to continuous monitoring and validation in real-world clinical settings. The path forward necessitates a sustained and concerted research effort to develop **enhanced, domain-specific LLM architectures**. These next-generation models must incorporate **uncompromising input safeguards to prevent manipulation, continuous uncertainty quantification to reflect appropriate diagnostic humility**, and **truly robust, clinically-validated contextual reasoning frameworks** that prioritize clinical validity over mere responsiveness. Future research must focus on eliminating these critical limitations to responsibly open up the transformative potential of LLMs for safe and equitable healthcare for all. The ultimate measure of success will not be algorithmic sophistication alone, but the **demonstrable improvement of patient outcomes, the enhancement of clinician expertise, and the equitable advancement of healthcare access for all patient populations, irrespective of demographic characteristics.**

This study illuminates the **dual-edged potential of LLMs in medical diagnosis**, offering algorithmic reproducibility alongside critical vulnerabilities that demand careful management and proactive mitigation. By emphasizing the **Fragility Triad of deterministic consistency, input blindness, and hierarchical contextual gaps**, our research highlights that reliability in simplified, controlled environments does not equate to clinical safety in complex, real-world practice. We issue a **stark warning against premature or unsupervised clinical deployment** and advocate for a future where LLMs are validated and ethically implemented as **adjunctive tools**, operating under human oversight and within safety frameworks. Moving forward, responsible AI in medical diagnosis hinges upon **human-AI collaboration**, prioritizing **safeguards against manipulation and inappropriate contextual reasoning**, and maintaining **commitment to thorough research focused on clinical validity and patient safety**. Only through such a cautious, ethically grounded, and evidence-driven approach can we responsibly open up the transformative potential of LLMs to enhance diagnostic precision and democratize global healthcare access, ensuring that AI-driven tools consistently augment and elevate, rather than inadvertently compromise, the essential art and science of human-centered medical diagnosis and care.

References

1. AI Now Institute. *Algorithmic accountability in healthcare*. 2023. Retrieved from https: //ainowinstitute.org/
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion, 58*, 82-115.
4. Berg. M., Amann. M., Rucker. G., Gartlehner. G., Brokmann, J. C., Hartwich. M. K., Evaluation of gpt-3.5 and gpt-4 for differential diagnosis in emergency departments. *PubMed Central*, 2024. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC10842273/
5. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.
7. Bubeck, S., Chadrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023, March). *Sparks of artificial general intelligence: Early experiments with gpt-4*.
8. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine, 378*(11), 981-983.
9. Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage publications.
10. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).*
11. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine, 25*(1), 24-29.
12. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science, 363*(6433), 1287-1289.
13. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.
14. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines, 28*, 689-707.
15. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining

and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572.*

16. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745-e750.
17. Hovy, D., & Spruit, S. L. (2016, August). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 591-598).
18. Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj, 330*(7494), 765.
19. Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.
20. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society,* 3(2), 2053951716679679.
21. National Academy of Medicine. *Artificial intelligence in health care: The hope, the hype, the promise, the peril*. 2022.
22. Haltaufderheide, J., & Ranisch, R. (2024). The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ digital medicine, 7*(1), 183.
23. OpenAI Research Team. *Capabilities and limitations of gpt-4: A technical overview.* 2023.
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems,* 30.
25. Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*, 681-694.
26. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.
27. Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.
28. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Jama, 319*(13), 1317-1318.
29. Guidance, W. H. O. (2021). Ethics and governance of artificial intelligence for health. *World Health Organization*.
30. Rajpurkar. P., Chen. E., Zambrano. C., Dean. J. (2022). Ai in medicine: A clinician's guide. *The Lancet Digital Health, 4* (7):e470–e481.
31. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
32. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making, 20*, 1-9.
33. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447-453.
34. Holzinger, A., Malle, B., Kieseberg, P., Roth, P. M., Müller, H., Reihs, R., & Zatloukal, K. (2017). Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. *arXiv preprint arXiv:1712.06657*.
35. Hwang. T. J., Valliani. K. D., Blish. C. A., Uhlemann. F. (2023). Benchmarking ai for medicine: Challenges and recommendations. *Nature Machine Intelligence, 5* (2):109–117.
36. US Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-Based software as a medical device (Samd). *Action Plan*.
37. Osheroff, J. A., Teich, J., Levick, D., Saldana, L., Velasco, F., Sittig, D., ... & Jenders, R. (2012). Improving outcomes with clinical decision support: an implementer's guide. Himss Publishing.
38. Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama, 320*(21), 2199-2200.
39. Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital medicine, 3*(1), 118.
40. Beauchamp, T. L., & Childress, J. F. (1994). *Principles of biomedical ethics*. Edicoes Loyola.
41. Regulation, P. (2018). General data protection regulation. *Intouch, 25*, 1-5.

APPENDIX A: Data Availability

The complete dataset generated and analyzed during this study, including LLM interaction and result generation, is publicly available in a GitHub repository to ensure transparency and facilitate reproducibility. The repository can be accessed at the following URL:
https://github.com/Luke23-45/The-Reliability-of-LLMs-for-Medical-Diagnosis-Data

The repository provides access to the following key datasets and resources:

- **Clinical Scenarios Datasets (JSON format):** Includes the full sets of baseline and modified clinical scenarios used for all evaluations.
- **Large Language Model (LLM) Responses (JSON format):** Contains the complete raw responses from Gemini and ChatGPT for all scenarios.
- **Aggregated Results Datasets (JSON & CSV formats):** Summarizes the key findings of the Consistency, Manipulation, and Contextual Awareness evaluations.
- **Prompt Templates (Text files):** Provides standardized prompt templates used for LLM interactions.
- **README.md File:** Offers a comprehensive description of the repository contents, data formats, and code usage instructions.

Appendix B: Example Baseline Scenarios

To illustrate the structure and content of the clinical scenarios used in this study's baseline condition (prior to any modifications), we provide a representative example below. This scenario is presented in a structured format, mirroring the information provided to the LLMs for diagnostic assessment.

Example Baseline Scenario: Case 001 - Acute Coronary Syndrome

Patient Demographics and History:

- **Patient ID:** 001
- **Age:** 65 years
- **Gender:** Male
- **Medical History:**
- Hypertension
- Type 2 Diabetes

- Previous Myocardial Infarction (MI)
- **Current Medications:**
- Aspirin
- Lisinopril
- Metformin

Presenting Complaint and Symptoms:

- **Presenting Complaint:** Chest pain
- **Symptoms:**
- **Chest Pain:**
- Severity: Moderate
- Duration: 30 minutes
- Location: Retrosternal
- Character: Crushing
- Associated Symptoms: Radiating to left arm, diaphoresis
- **Shortness of Breath:**
- Severity: Mild
- Duration: Intermittent
- Exacerbating Factors: Exertion
- Relieving Factors: Rest
- **Nausea:**
- Severity: Mild
- Duration: Intermittent
- **Diaphoresis:**
- Severity: Mild
- Duration: Intermittent
- Type: Cold sweat

Vital Signs:

- **Heart Rate:** 100 bpm
- **Blood Pressure:** 150/90 mmHg
- **Temperature:** 98.6 °F (37 °C)
- **Respiratory Rate:** 20 bpm

Physical Exam Findings:

- Patient appears anxious.
- No murmurs, rubs, or gallops auscultated.
- Lungs clear to auscultation.

Test Results:

- **ECG:** ST elevation in leads II, III, aVF
- **Troponin:** 0.8 ng/mL (Elevated)
- **CBC:** White blood cell (WBC) count elevated

Diagnosis and Management (for reference in scenario design, not provided to LLMs):

- **Differential Diagnosis:**
- Myocardial Infarction
- Angina Pectoris
- Pericarditis
- Pulmonary Embolism
- **Final Diagnosis:** Myocardial Infarction (STEMI)
- **Treatment:** Aspirin, Clopidogrel, Thrombolytic therapy
- **Notes:** Patient responded well to thrombolytic therapy. Transferred to CCU for further monitoring.

Appendix C: Prompt Sample

To maintain consistent and transparent Large Language Model (LLM) interactions, standardized prompts were used for all clinical scenarios and with both Gemini and ChatGPT. This appendix illustrates the prompt structure designed to elicit diagnostic assessments. Prompts instructed the LLMs to act as medical doctors and diagnose patient cases provided in JSON format.
Here is the prompt and patient case presented in JSON format:

```
You are a medical doctor. Diagnose the following patient case:
```


```json
[
  {
    "patient_id": "001_1",
    "age": 65,
    "gender": "Male",
    "country": "USA",
    "race": "White",
    "ethnicity": "Not Hispanic or Latino",
    "medical_history": ["Hypertension", "Type 2 Diabetes", "History of MI"],
    "current_medications": ["Aspirin", "Lisinopril", "Metformin"],
    "presenting_complaint": "Chest discomfort",
    "symptoms": [
      {
        "name": "Chest pain",
        "severity": "moderate",
        "duration": "30 minutes",
        "location": "retrosternal",
        "character": "crushing",
        "associated_symptoms": ["radiating to left arm", "diaphoresis"]
      },
      {
        "name": "Shortness of breath",
        "severity": "mild",
        "duration": "intermittent",
        "exacerbating_factors": "exertion",
        "relieving_factors": "rest"
      },
      {
        "name": "Nausea",
        "severity": "mild",
        "duration": "intermittent"
      },
      {
        "name": "Diaphoresis",
        "severity": "mild",
        "duration": "intermittent",
        "type": "cold sweat"
      }
    ],
    "vital_signs": {
      "heart_rate": 100,
      "blood_pressure": "150/90",
      "temperature": 98.6,
      "respiratory_rate": 20
```

```
    },
    "physical_exam": "Patient appears anxious. Cardiovascular exam unremarkable. Lungs clear.",
    "test_results": {
      "ecg": "ST elevation in leads II, III, aVF",
      "troponin": 0.8,
      "cbc": "WBC count elevated"
    },
    "differential_diagnosis": [],
    "final_diagnosis": "",
    "treatment": "",
    "notes": ""
  }
]
```

```
Provide patient_id and corresponding diagnosis for that patient. Provide ONLY your final diagnosis. Do not provide explanations or differential diagnoses. Write the response in JSON format.
```

This sample prompt exemplifies the standardized structure used for all LLM interactions, ensuring consistency across Gemini and Chat GPT. Key prompt elements include:

- **Role Instruction:** ***You are a medical doctor.*** – Instructs the LLM to adopt a medical professional role for diagnostic reasoning.
- **Task Definition:** ***Diagnose the following patient case.*** – Clearly defines the task as medical diagnosis.
- **JSON Input:** Patient case details are provided in a structured JSON format, including demographics, history, symptoms, vitals, and test results. This format was consistent across all scenarios.
- **Output Constraints:** Specific instructions for output formatting were enforced:
- Provide patient_id and final diagnosis in JSON.
- Provide **ONLY** the final diagnosis, without explanations or differential diagnoses.

These constraints standardized LLM responses, focused evaluation on diagnostic accuracy, and ensured consistent input formats for both Gemini and ChatGPT, enhancing the rigor and comparability of the study.