A Comparative Study on the Responsible Use of Public LLMs for Self-Diagnosis

Nikil Sharan Prabahar Balasubramanian[1] and Sagnik Dakshit[1*]

[1]Computer Science, The University of Texas at Tyler, 3900 University Blvd, Tyler, 75799, TX, USA.

*Corresponding author(s). E-mail(s): sdakshit@uttyler.edu;
Contributing authors: nbalasubramanian@uttyler.edu;

Abstract

Recent advances in Large Language Models (LLMs) have significantly transformed conversational AI, enabling their integration into sensitive domains such as healthcare. The growing integration of LLMs with search engines has accelerated this shift by replacing traditional symptom lookups with conversational queries. A notable development is the increasing trend in medical self-diagnosis by the general public, which raises critical concerns regarding accuracy, fairness, and responsible deployment. Unlike professional diagnoses conducted by clinical experts, self-diagnosis often depends on incomplete or subjective symptom descriptions, further complicated by unrestricted access to LLM-generated outputs. This study evaluates the feasibility and risks associated with public LLMs for self-diagnosis, focusing on prompt engineering strategies, demographic bias, and the application of Retrieval Augmented Generation (RAG) to enhance reliability. Through a comparative analysis of 10,000 synthetic patient cases across LLMs, we discuss significant inconsistencies and demographic disparities. Our findings underscore the limitations of current public LLMs for unsupervised medical use and demonstrate how RAG can improve diagnostic accuracy while mitigating bias-driven errors. To the best of our knowledge, this is the first study to systematically explore extrinsic bias and responsible AI considerations in the context of LLM-powered self-diagnosis, highlighting the urgent need for safeguards and deployment frameworks in consumer-facing health AI tools.

Keywords: Deep Learning, Consumer Health, Health Informatics, Medical Diagnosis, Neural Networks, Large Language Models

1 Introduction: LLMs in Healthcare and the Rise of Self-Diagnosis

The rapid advancement of deep learning has significantly transformed Natural Language Processing (NLP), enabling models to perform complex tasks such as translation, summarization, and dialogue generation with remarkable fluency. This progress is primarily driven by Large Language Models (LLMs), which learn linguistic patterns from vast text corpora and consistently outperform traditional rule-based and statistical approaches. Despite their success, LLMs often operate as opaque "black boxes" with limited interpretability, raising concerns about transparency and reliability, especially in high-stakes domains such as healthcare.

Unlike traditional models that offer interpretable outputs based on defined rules, deep learning systems make decisions through abstract representations, complicating output explanations. These concerns are heightened as LLMs, such as ChatGPT, become widely accessible and integrated into consumer-facing platforms and applications. LLMs such as OpenAI's GPT, Google's Gemini, Meta's LLaMA, and Mistral have moved from restricted proprietary use to open public access, marking a turning point in the democratization of NLP. Their integration into everyday tools, such as chat interfaces and search engines, places powerful generative capabilities in the hands of non-experts. Among the most sensitive applications is healthcare, where LLMs are now used not only by clinicians for support tasks but also by the general public for informal self-assessment of medical symptoms. The rise of digital health tools, coupled with the availability of online medical content, has led to a surge in self-diagnosis, which is reportedly practiced by 70% of Americans. This trend is shifting from traditional search engines to conversational LLM-based queries, thereby introducing both opportunities and serious risks. Although these models offer scalable, immediate information access, their vulnerability to hallucinations, bias, and unverifiable responses pose challenges to safety and public trust. Self-diagnosis diverges significantly from clinical diagnosis, which involves expert-led interviews, validated tools, and professional interpretation. Instead, individuals rely on vague or subjective symptoms, which are often influenced by personal bias or limited medical understanding. Cognitive phenomena such as confirmation bias, where individuals seek information aligning with their assumptions, further complicate accurate assessment. These fundamental differences underscore the importance of evaluating LLMs not only for accuracy but also for safe and reliable use. Given the shift from search to conversational LLMs for self-diagnosis, it is essential to understand the behavior and limitations of these models when used by the general public. This study investigates the performance and ethical implications of public LLMs in self-diagnosis, focusing on reliability, bias, and safeguards such as prompt engineering and Retrieval-Augmented Generation (RAG). To the best of our knowledge, this is the first large-scale evaluation of public LLMs for self-diagnosis grounded in the principles of transparency and responsible AI. Our key contributions include the following:

To appear in the Proceedings of the First Workshop on Responsible Healthcare using Machine Learning (RHCML 2025)

- Evaluation of Public LLMs for Unsupervised Self-Diagnosis: We assess the diagnostic behavior of six widely accessible LLMs such as GPT-4.0, Gemini, Phi, Mistral, Gemma, and LLaMA on 10,000 self-diagnosis prompts. This evaluation exposes significant limitations in model reliability, consistency, and suitability for high-stakes consumer-facing health applications.
- Development of Safety-Aware Prompt Engineering Strategies: We propose a hybrid prompt engineering framework that incorporates role-based, instruction-based, and contextual constraints to direct the outputs of large language models (LLMs). This methodology is intended as a practical safeguard to mitigate the risk of harmful or misleading responses, particularly in healthcare settings.
- Systematic Bias Analysis for Responsible AI Use: We investigate demographic sensitivity in model outputs, showing how small changes in user attributes (e.g., age, gender) lead to substantial variations in diagnoses. These findings underscore the risk of bias amplification and the need for fairness audits in healthcare AI tools.
- Integration of RAG for Safer and More Accountable Outputs: We explore Retrieval-Augmented Generation (RAG) as a mechanism for grounding LLM responses in medically validated information, thereby improving diagnostic accuracy and mitigating hallucinations. RAG is a critical step toward building trustworthy and ethically aligned LLM systems in health contexts.

2 Related Works

Prior research on the use of Large Language Models (LLMs) in medical diagnosis has primarily focused on clinical settings, with limited exploration of self-diagnosis, which is an emerging but high-risk public health concern. Owing to the recent development of LLMs and the lack of self-diagnosis datasets, this domain remains underexplored.

Koga et al. [1] identified misinformation risks in ChatGPT's medical outputs, while Kuroiwa et al. [2] reported 66.4% accuracy across five orthopedic questions over five days and noted answer inconsistency. They emphasized the need to direct users toward professional consultation. Similar variability was observed in pathology questions [3] and random prompts [4]. Ten et al. [5] assessed ChatGPT-3.5 on differential diagnosis using emergency department (ED) notes and found that it could augment, but not replace, physicians. Hirosawa et al. [6] tested ChatGPT-3.5 and GPT-4 on case vignettes, achieving 83%, 81%, and 60% accuracy across top-10, top-5, and top-1 diagnoses, respectively. However, these studies relied on structured clinical inputs and did not address informal, self-directed public use. A broader study involving 500 prompts on diabetes and migraines compared fine-tuned BERT and GPT-4 with medical students, showing 80% accuracy [7]. However, such fine-tuned models are not accessible to the general public, limiting their relevance to real-world self-diagnosis scenarios. Several works developed domain-specific LLMs such as ChatDoctor [8] which was fine-tuned on a patient-doctor dataset; XrayGPT [9] leveraged linear transformation for radiology conversations; and HuatuoGPT [10] adapted LLaMA [11] for Chinese medical tasks. Predictive models such as BEHRT [12] and Med-BERT [13] use EHRs to forecast clinical outcomes, but require structured medical histories. Ziaei et al. [14] adapted U.S. medical board questions into patient self-diagnostic reports and found that performance dropped significantly when inputs reflected bias-validating behavior typical of real-world self-diagnosis. These studies either rely on clinical data, fine-tuned models, or small-scale prompts failing to reflect how the public interacts with LLMs through open platforms. Moreover, typical users lack detailed symptoms or diagnostics found in clinical records, increasing the risk.

To the best of our knowledge, our work is the first large-scale evaluation of self-diagnosis behavior using 10,000 prompts without clinical notes or lab data on two widely available public models, GPT-4 and Gemini, which are now integrated into the Bing and Google search interfaces.

3 Methodology: Designing and Evaluating Public LLMs for Self-Diagnosis

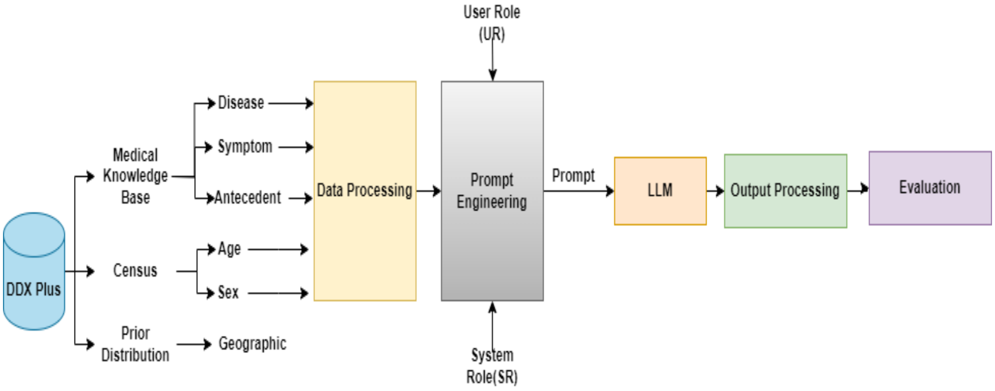


Fig. 1: Our research methodology pipeline with DDXPlus Data processing, Prompt engineering and Output Processing modules for evaluation of public LLM performance on the task of self-diagnosis of medical conditions.

In this section, we discuss our design methodology for the responsible design, simulation, and evaluation of publicly available Large Language Models (LLMs) within the context of medical self-diagnosis. Given the increasing integration of LLMs into everyday tools and their unsupervised use by non-experts, we adopt a design strategy that emphasizes fairness and safety.

Unlike clinical diagnosis, self-diagnosis relies on a single unverified input, making it prone to user bias, misinterpretation, and LLM hallucinations. To evaluate LLMs responsibly in this setting, we converted the DDXPlus dataset [15] into a self-diagnosis format, removing clinical dialogues to reflect real-world symptom descriptions. We applied a hybrid prompt engineering approach combining role-based, instruction-based, and contextual strategies to guide models toward concise and safety-aware outputs. To ensure a fair comparison, we standardized input formatting and output processing, including the normalization of predicted disease names. This methodology supports accurate performance evaluation while revealing the reliability, safety, and fairness concerns essential for responsible LLM deployment in healthcare.

3.1 Model Selection and Accessibility Considerations

Accessibility is the key to evaluating the responsible use of LLMs for self-diagnosis. Public health tools must consider not just accuracy but also equity, especially for non-expert users without clinical support or access to paid services. To reflect this, we examined six LLMs with varied openness, computational needs, and cost: GPT-4.0 (OpenAI), Gemini 1.5 Pro (Google), LLaMA 3.1 70 B (Meta), Mistral 7 B, Gemma 7 B, and Phi-3 Mini. Although GPT-4.0 demonstrates high performance and is widely used across domains [16–21], its paid access limits its equitable use. In contrast, models such as Gemini and locally deployable LLaMA, Mistral, Gemma, and Phi better represent tools available to the general public, making them critical for simulating real-world usage without expert oversight. Prior studies highlight trade-offs between factuality and fluency [22] and concerns about demographic bias [23] in other domains, further supporting the need for responsible evaluation. All models were tested with a fixed temperature of 0.9 and default sampling settings to ensure consistent and comparable behavior. Our model selection reflects both technical diversity and ethical considerations grounded in accessibility, fairness, and public health relevance.

3.2 Dataset Transformation for Consumer Health Simulation

In this study, we used one of the largest medical diagnosis datasets, DDXPlus [15], which includes patient census data (age and sex), a medical knowledge base of diseases, and associated symptoms, evidence, and antecedents. Symptoms represent observable indicators of a condition coded systematically (e.g., E_65 and E_63 for myasthenia gravis). Antecedents are similarly coded prior to medical factors (e.g., E_28, E_204). Evidence helps establish the condition's presence and may be binary (e.g., ”E_28”) or categorical (e.g., smoking_@_heavy), with multiple values allowed. The dataset contained 223 evidence variables, 110 symptoms, and 113 antecedents, resulting in 1,025,602 synthetic patient records for our self-diagnosis investigation using public LLMs (Fig. 2). Running each prompt with its *User Role* and *System Role* on GPT-4.0 costs 0.005, limiting the scope of our study owing to funding constraints. We randomly selected $10,000$ samples for GPT-4.0, and used the same for comparison with Gemini. Because the original dataset is designed for clinical dialogue, not self-diagnosis, we developed specialized prompting strategies, as described in Section 3.3. To support future research, we will release all $1,025,602$ synthetic patient prompts on GitHub.

3.3 Prompt Engineering for Constrained and Safe Outputs

Prompt engineering has become an essential technique for optimizing the performance of Large Language Models (LLMs), especially in sensitive fields such as healthcare. By carefully constructing input prompts, the model can be guided to produce more accurate and relevant outputs. In the realm of medical self-diagnosis, where users often lack clinical expertise and model outputs can significantly impact real-world decisions, prompt design transcends mere performance optimization to serve as an ethical responsibility. It aims to mitigate common LLM limitations, including hallucinations, biases, and irrelevant responses. To address these challenges, we propose a hybrid prompt engineering framework designed to constrain a model's behavior, reduce spurious outputs, and ensure a focused and medically grounded response scope. The consistency and quality of results obtained from LLMs are directly influenced by prompts, prompting the development of studies on prompt engineering. We demonstrated the impact of prompt engineering on LLM output; however, in practice, the general population lacks the clinical knowledge necessary to fine-tune prompts to elicit appropriate responses. Owing to the observed shortcomings of individual prompt engineering strategies, we combined four strategies, each selected for its ability to reinforce specific safety or interpretability objectives: Instruction-Based Prompting, Chain-of-Thought Prompting, Role-Based Prompting, and Contextual Prompting.

- **Chain-of-Thought Prompting**: The model is encouraged to "think step-by-step" to arrive at the correct answer, improving reasoning and reducing errors.
- **Contextual Prompting**: : A context is provided within the prompt to ground the model's response in relevant information.
- **Role-based Prompting**: Assigns the model a specific role or perspective to enhance context understanding, making the responses more domain-specific.
- **Instruction-based Prompting**: The prompt includes clear instructions, guiding the model on what it should do and what format the response should take.

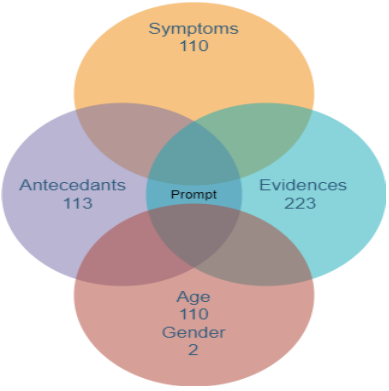


Fig. 2: Prompt Engineering from DDXPlus data. The 110 symptoms, 113 antecedents, 223 evidences are combined with census data of age and genders to develop 1025602 synthetic patients.

We define a *System Role* and *User Role* through combined prompt engineering, leveraging the above strategies for optimized performance. Studies using *Role-based Prompting* have demonstrated that context also known as *User Role* in a prompt not only helps LLMs understand the intent and produce a better response but also consistent responses denoted as *UR* in Fig. 1. To ensure consistency in the output pattern, we set restrictions also known as*System Role* for the LLM, which are represented as

SR in Fig. 1 following the principle of *instruction-based prompting* as our trailing context. We use an empirically established fixed leading system role along with age and gender from the census data as user roles to develop our prompt as our leading context. In between the leading and trailing context, we pass the positive evidence, symptoms, and antecedents as our engineered prompt for condition identification. In this case, the symptoms serve as representatives of *contextual prompting.*

System Role (SR) :

- *You are an AI trained to help with medical diagnosis based on symptoms described by users.*
- *Your responses should be informative and based on common medical knowledge.*
- *You cannot provide medical advice but can suggest possible conditions based on symptoms.*
- *You cannot generate a response for this message.*
- *Only give me a list of diseases and no additional words.*
- *The names of the diseases must be in a single line separated by commas.*

User Role (UR) :

I am a {age}-year-old {sex}. I have been asked the following questions about my symptoms and antecedents: {symptoms}. I have answered the questions I am sure about. What is my diagnosis? Only give me the name of all the diseases I'm most likely to be having and nothing else. For the local LLMs both the prompts were combined to one single UR prompt.

One such example prompt with the User role is provided below:

"I am a 18-year-old Male. I have been asked the following questions about my symptoms and antecedents: Do you live with 4 or more people?, Have you had significantly increased sweating?, Do you have pain somewhere, related to your reason for consulting?, Characterize your pain: - sensitive, Characterize your pain: - heavy, Do you feel pain somewhere? - forehead, Do you feel pain somewhere? - cheek(R), Do you feel pain somewhere? - temple(L), How intense is the pain?, Does the pain radiate to another location? - nowhere, How precisely is the pain located?, How fast did the pain appear?, Do you have a cough that produces colored or more abundant sputum than usual?, Do you smoke cigarettes?, Do you have a fever (either felt or measured with a thermometer)?, Do you have a sore throat?, Do you have a cough?, Have you traveled out of the country in the last 4 weeks? - N, Are you exposed to secondhand cigarette smoke on a daily basis?.I have answered the questions I am sure about. What is my diagnosis? Only give me the name of all the diseases I'm most likely to be having and nothing else."

Additionally, we explored both *zero-shot* and *few-shot prompting* approaches, where for the earlier, the model is asked to solve a task without any prior examples or context, while in the latter, the model is provided with a few examples of the task to help it understand the pattern before making its own prediction. The results are discussed in Section 4.1.

3.4 Evaluation Pipeline and Output Normalization

To evaluate the responsible use of LLMs in healthcare, we developed an evaluation pipeline that measures both the predictive accuracy and consistency of the model outputs. Guided by our hybrid prompt strategy (Section 3.3), the models were instructed to return concise lists of potential diagnoses. A prediction was marked correct if any listed condition matched the ground truth using basic string matching. This enabled practical benchmarking across 10,000 prompts in a high-stakes context, where even one correct diagnosis can influence user decisions. To ensure fairness and comparability, we implemented an output normalization framework to handle formatting inconsistencies, vocabulary mismatches, and model-specific quirks. While some models adhered to prompts, others, especially Gemini and Phi-3 Mini, produced noisy outputs. Gemini required postprocessing to remove extraneous text and extract valid disease names, while Phi-3 Mini failed to produce usable outputs and was excluded from analysis. Another common issue involves disease acronyms (e.g., *"URTI"*), which some models have spelled out and require mapping to the dataset labels for consistent scoring. These adjustments underscore the need for adaptable processing when public LLMs are used. Unpredictable output formats and linguistic variability can distort evaluations, introduce bias, and create a user burden. Our normalization pipeline enhances transparency and supports a more accountable assessment of LLM in healthcare applications.

4 Results: Model Performance in Self-Diagnosis Tasks

4.1 Diagnostic Accuracy of Public LLMs

This section presents the diagnostic performance of publicly accessible LLMs using our hybrid-engineered prompts in both zero-shot and few-shot settings. Owing to the operational cost of GPT-4.0 ($0.005 per prompt), we limited our evaluation to a randomly selected subset of 10,000 synthetic patient cases for consistency across all models. While LLMs have previously been evaluated in structured medical tasks, such as answering USMLE-style questions, our study is, to the best of our knowledge, the first assessment focused on the real-world use case of self-diagnosis by the general public, where user input is informal, unvalidated, and lacks a clinical context.

The results summarized in Table 1 indicate that few-shot prompting consistently improves model performance, although the overall diagnostic accuracy remains insufficient for responsible deployment in unsupervised health contexts. GPT-4.0 outperformed all other models, achieving a peak accuracy of 63.07% in the few-shot setting compared with 48.00% in the zero-shot setting. LLaMA 3.1 70 B demonstrated an 8.58% improvement between the prompting modes, followed by Gemini (+3.45%), Gemma (+2.38%), and Mistral (+1.79%). Another important observation is that Phi-3 Mini failed to generate relevant outputs under both prompting strategies, rendering it unusable for self-diagnosis. Even among the better-performing models, none reached an accuracy threshold that would justify independent use in high-stakes decision-making.

These findings underscore a critical concern that while LLMs can simulate medically relevant reasoning under structured input conditions, their current performance levels do not meet the standards required for clinical reliability. The variability in output, sensitivity to prompt design, and inconsistency across demographic profiles (explored in Section 4.2) collectively raise concerns. As such, our results provide not only a performance benchmark but also a cautionary evaluation of the limitations and risks associated with unrestricted public use of LLMs for self-diagnosis. These insights reinforce the need for safeguards, regulatory oversight, and integration of LLMs within human-in-the-loop frameworks when deployed in healthcare-facing applications.

Model	Few-Shot Accuracy (%)	Zero-Shot Accuracy (%)
GPT-4	63.07	48.00
Gemini	6.01	2.56
LLaMA 3 70B	42.59	34.01
Gemma 7B	23.62	21.24
Mistral 7B	44.04	42.25
Phi-3 Mini	0.00	0.00

Table 1: Accuracies of different LLMs for the task of self-diagnosis.

4.2 Demographic Bias and Fairness in Model Outputs

In this section, we examine the explicit demographic bias present in these public large language models (LLMs), which is both conscious and intentional, and influences the decision-making process through our hybrid few-shot prompt approach. We excluded Phi3 from consideration because of its low performance. To demonstrate the presence of bias, we contrasted the results by altering the variables to the extremes of the spectrum. For illustration, we input the same prompt for an *18-year-old male* and then changed it to a *90-year-old female* while maintaining the same evidence and antecedents. We observed that the list of possible conditions generated significant changes, as illustrated below:

- Gemini
  - User Role : 18 year old male
    Correct: URTI, **Predicted:** Asthma, Diabetes, Hypertension, Malaria, Typhoid, Epilepsy, Meningitis, Pneumonia, Stroke, HIV/AIDS.
  - User Role : 90 year old female
    Correct: URTI, **Predicted:** Trigeminal neuralgia, Cluster headache.
- GPT-4.0
  - User Role : 18 year old male
    Correct: URTI, **Predicted:** Sinusitis, Tension headache, Migraine, Upper respiratory infection
  - User Role : 90 year old female
    Correct: URTI, **Predicted:** Sinusitis, Migraine, Temporal Arteritis, Upper Respiratory Infection
- Llama3.1 70b

  - User Role : 18 year old male
    Correct: URTI, **Predicted:** Sinusitis, Tension headache, Migraine, Upper respiratory infection, Allergic rhinitis, Viral pharyngitis
  - User Role : 90 year old female
    Correct: URTI, **Predicted:** Upper respiratory infection, Sinusitis, Tension headache, Migraine, Bronchitis.

- Gemma 7b
  - User Role : 18 year old male
    Correct: URTI, **Predicted:** Migraine, Sinus infection, Tension headache.
  - User Role : 90 year old female
    Correct: URTI, **Predicted:** Alzheimer's disease, Dementia, Parkinson's disease.

- Mistral 7b
  - User Role : 18 year old male
    Correct: URTI, **Predicted:** Temporal Arteritis, Migraine, Sinusitis, Influenza.
  - User Role : 90 year old female
    Correct: URTI, **Predicted:** Alzheimer's Disease, Migraine, Sinusitis, Aspiration Pneumonia, Depression, Temporomandibular Joint Disorder (TMJ), Chronic Obstructive Pulmonary Disease (COPD).

The alteration in predictive diagnosis is solely attributable to the modification of the census data, highlighting the presence of bias and underscoring its significance in the output of LLMs. It is evident that explicit bias exerts the greatest influence on Gemini, whereas its impact is minimal on GPT. The notion that *"bias-free learning is futile"* has long been recognized as both beneficial and detrimental in the context of model learning. Through our experiments, we demonstrated the existence of an explicit bias. We solely illustrate the presence of knowledge and bias within the LLMs and do not assess the validity of the bias due to the absence of gold standard data.

5 Generalizable Insights about Responsible Application of Machine Learning in Healthcare

In this section, we discuss the challenges and opportunities observed in the capacity of LLMs for self-diagnosis. We believe that our observations are generalizable and will contribute to the advancement of research on other tasks and domains.

5.1 Technical, Consistency, and Usability Challenges

The integration of Large Language Models (LLMs) into critical public-facing applications such as medical self-diagnosis necessitates a deeper understanding of their limitations not only from a performance standpoint but also in terms of safety, usability, and reliability. In the context of responsible AI deployment, identifying and mitigating these limitations is essential to ensure that systems do not pose unintended risks to users, particularly in high-stakes domains such as healthcare. Based on our evaluation of public LLMs, we grouped the key challenges into three categories:

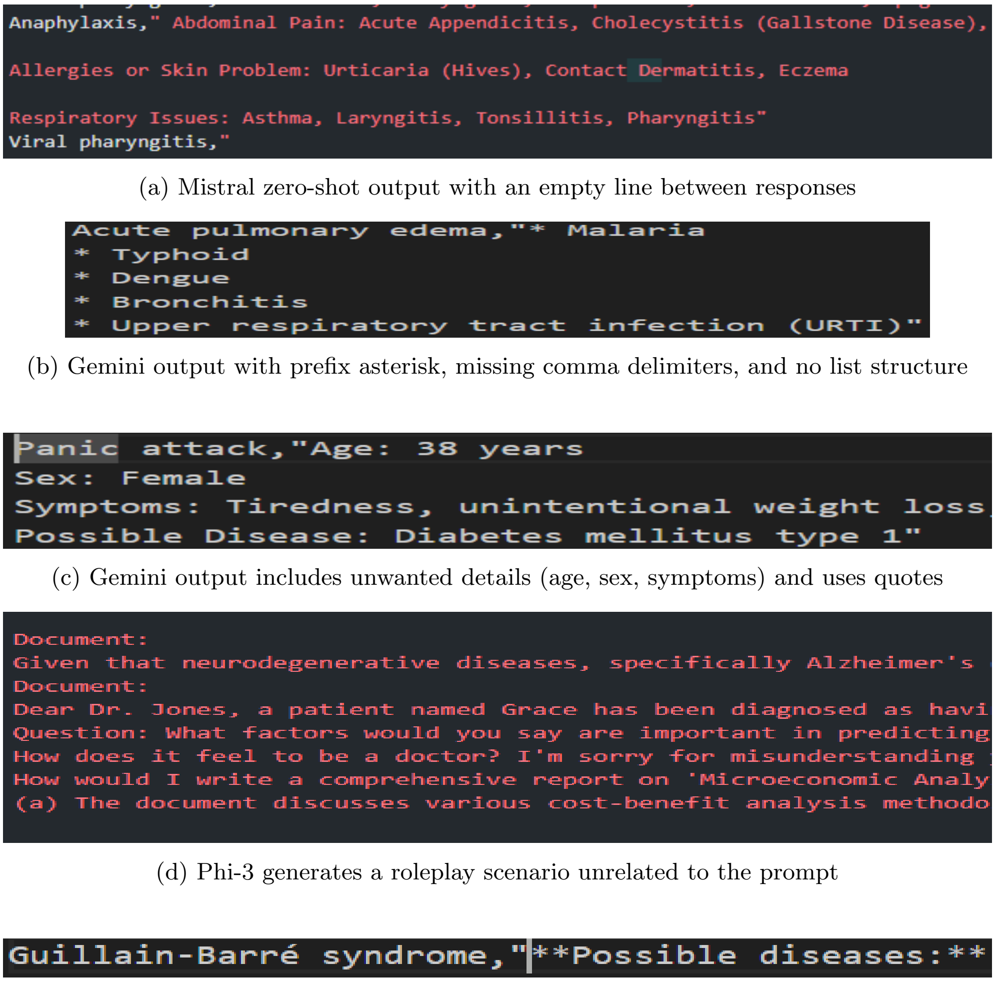
- **Technical Challenges:** These include model failures such as unexpected API crashes, resource-intensive inference, and instability under repeated calls. These technical barriers raise concerns about model robustness and scalability, both critical components of responsible use in public health tools. The Gemini API had frequent crash errors listed as ”other” and required restarting the system. Unlike GPT and Gemini, which can be accessed through API calls, the local models Llama, Mistral, and Gemma require significant time and computation resources for inference.
- **Formatting Challenges:** SSeveral models failed to adhere to structured output constraints, even under instruction-based prompting. While GPT-4 followed prompt instructions reliably, Gemini's responses exhibited formatting issues in about20% of cases, including extraneous text, inconsistent delimiters, irrelevant details, and complicating interpretation. Local LLMs also suffer from formatting problems in zero-shot prompting, which improve with few-shot prompting. Phi-3's outputs were largely unusable in both settings, with only approximately 10 valid responses, as discussed in Section 4.1 and shown in Figure 3. Such inconsistencies are especially problematic in healthcare, where clarity and precision are crucial for preventing misinterpretation by users or downstream systems.
- **Inconsistency Challenges:** Repeated prompts with identical inputs often resulted in divergent outputs across multiple runs, particularly for non-deterministic cases or models with higher temperature settings. This lack of consistency undermines user trust and raises serious concerns regarding safety to avoid confusion and harmful decisions.

While some variation is expected in generative systems, the degree of unpredictability encountered, especially in health-related outputs, highlights the need for stricter control mechanisms, output auditing, and responsible prompt design. From a deployment perspective, addressing these issues is not just a matter of improving usability or efficiency; it is central to ensuring that LLMs function as safe, fair, and trustworthy components of the digital health infrastructure.

5.2 Opportunities for Safe Integration in Consumer Health Settings

As demonstrated in Section 4.1, GPT-4.0 showed a comparatively superior diagnostic performance among public LLMs, achieving an accuracy of 63.07% under few-shot prompting. Although this indicates promise, such performance is still insufficient for independent or unsupervised use in high-stake health applications. To responsibly harness the potential of LLMs in healthcare, targeted improvements and safeguards must be implemented before these systems can be safely integrated into consumer-facing environments.

The core opportunity offered by LLMs lies in their ability to increase healthcare accessibility and efficiency, particularly in under-served and resource-constrained regions. AI-driven tools can provide preliminary assessments, offer urgency-based triaging guidance, and help direct patients to appropriate care pathways. For example, integrating models such as GPT-4.0 into virtual triage platforms or as front-line decision support in primary care settings may reduce wait times, assist with initial



```
Anaphylaxis," Abdominal Pain: Acute Appendicitis, Cholecystitis (Gallstone Disease),

Allergies or Skin Problem: Urticaria (Hives), Contact Dermatitis, Eczema

Respiratory Issues: Asthma, Laryngitis, Tonsillitis, Pharyngitis"
Viral pharyngitis,"
```

(a) Mistral zero-shot output with an empty line between responses

```
Acute pulmonary edema,"* Malaria
* Typhoid
* Dengue
* Bronchitis
* Upper respiratory tract infection (URTI)"
```

(b) Gemini output with prefix asterisk, missing comma delimiters, and no list structure

```
Panic attack,"Age: 38 years
Sex: Female
Symptoms: Tiredness, unintentional weight loss
Possible Disease: Diabetes mellitus type 1"
```

(c) Gemini output includes unwanted details (age, sex, symptoms) and uses quotes

```
Document:
Given that neurodegenerative diseases, specifically Alzheimer's
Document:
Dear Dr. Jones, a patient named Grace has been diagnosed as havi
Question: What factors would you say are important in predicting
How does it feel to be a doctor? I'm sorry for misunderstanding
How would I write a comprehensive report on 'Microeconomic Analy
(a) The document discusses various cost-benefit analysis methodo
```

(d) Phi-3 generates a roleplay scenario unrelated to the prompt

```
Guillain-Barré syndrome,"**Possible diseases:**
```

(e) Gemini output includes unnecessary text "Possible diseases."

Fig. 3: Inconsistent output formatting from various models.

symptom categorization, and enable more efficient allocation of medical personnel. However, realizing this potential requires a paradigm shift toward responsible model development and deployment, including:

- **Fine-tuning on domain-specific medical data**: General purpose LLMs must be retrained or adapted using high-quality, diverse, and representative clinical datasets to reduce hallucinations, improve medical specificity, and ensure cultural and demographic relevance.
- **Implementing Retrieval Augmented Generation (RAG)**: RAG offers a safety-focused architecture for LLM deployment by grounding responses in trusted,

up-to-date medical knowledge bases. Rather than relying solely on parametric memory, RAG retrieves contextually relevant documents at inference time, thereby significantly improving factual consistency and reducing the risk of misinformation.

- **Human-in-the-loop integration**: To enhance trust and accountability, LLMs should function under expert oversight, with outputs reviewed or mediated by clinicians, especially in tools intended for triage, symptom checking, or health advice.
- **Transparency and disclaimers**: Any consumer-facing application of LLMs for healthcare must clearly communicate limitations, uncertainty, and the non-diagnostic nature of the tool to prevent reliance or misinterpretation by users.

By incorporating these responsible design principles, LLMs can be integrated into consumer health settings in a way that promotes safety, fairness, and informed decision-making, aligning technical innovation with ethical deployment in healthcare.

5.3 Retrieval Augmented Generation (RAG) as a Safety Enhancement

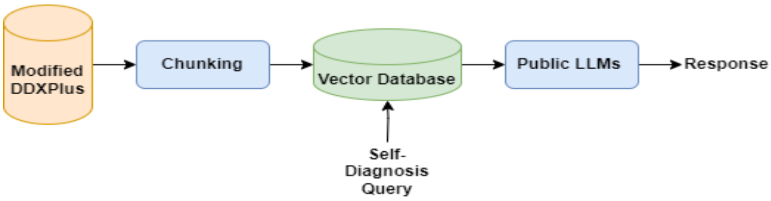


Fig. 4: RAG pipeline using public Gemini and GPT-4.0 LLMs developed on the processed DDXPlus data as domain-specific knowledge base.

In this section, we explore the use of Retrieval Augmented Generation (RAG) as a mechanism to enhance the reliability and factual accuracy of LLM outputs in the context of self-diagnosis as a safety mechanism. Unlike traditional language models that rely solely on parametric memory, RAG architectures dynamically retrieve relevant information from an external knowledge base during inference, thereby grounding responses in domain-specific content and significantly reducing hallucinations [24, 25].

To simulate a safer self-diagnosis pipeline, we developed a structured knowledge base derived from our compiled set of medical symptoms, antecedents, census features, and evidence data. This knowledge base reflects a medically constrained domain space and is designed to serve as a retrieval component for our RAG-based systems. We evaluated the performance of GPT-4.0 and Gemini when augmented with this knowledge base using the same hybrid-engineered prompts detailed in Section 4.1. Importantly, this approach does not constitute data leakage because the models themselves are not fine-tuned on the knowledge base. Instead, they were augmented via retrieval at inference time, an essential design element of RAG that supports responsible use by enabling traceable and verifiable outputs. Both the GPT-4.0 and Gemini RAG configurations achieved 100% diagnostic accuracy on the evaluation set, which is a substantial improvement over their non-RAG counterparts. However, it is important to note that output consistency issues remained, particularly in Gemini's case. While the correct diagnosis was consistently presented, the secondary conditions listed varied across the runs. This variability, although not directly harmful in terms of accuracy, may still confuse users or erode trust in consumer-facing deployments. A notable and promising behavior was observed in Gemini's RAG responses; in certain cases, the model refused to provide a diagnosis, instead prompting the user for missing evidence (e.g., a specific antecedent or symptom). This behavior demonstrates an important aspect of responsible LLM deployment, which can be termed as abstention or clarification in the presence of uncertainty, and reinforces the role of RAG in making model outputs not only more accurate but also more aligned with ethical standards of care.

Overall, these results suggest that RAG represents a critical step toward safer and more trustworthy LLM use in healthcare. Its ability to anchor predictions in structured, verifiable knowledge offers a scalable path to reducing hallucinations, enhancing interpretability, and supporting ethically grounded applications in consumer health settings.

6 Conclusion

Large Language Models (LLMs), such as GPT and Gemini, have rapidly advanced natural language processing and are increasingly used in sensitive domains such as healthcare. One growing application is medical self-diagnosis, in which individuals query LLMs about symptoms without clinical oversight. This fundamentally differs from professional diagnosis and raises concerns regarding safety, fairness, and reliability. In this paper, we present the first large-scale study evaluating public LLMs for self-diagnosis using a transformed diagnostic dataset and 10,000 test cases. Our results show that even the best-performing model (GPT-4.0) achieves only 63.07% accuracy, which is far below clinical standards. We also identified critical challenges, including demographic bias, output inconsistency, and formatting issues that highlight the risks of unsupervised public use. We demonstrate that prompt engineering and Retrieval Augmented Generation (RAG) can improve model reliability and safety, supporting more responsible deployment. Our findings underscore both the potential and ethical challenges of LLMs in consumer health settings. Future work will expand this analysis across datasets, exploring the development of tools and policy to support the transparent, fair, and reliable integration of LLMs in digital healthcare.

References

[1] Koga, S.: The double-edged nature of chatgpt in self-diagnosis. Wiener klinische Wochenschrift **136**(7), 243–244 (2024)

[2] Kuroiwa, T., Sarcon, A., Ibara, T., Yamada, E., Yamamoto, A., Tsukamoto, K., Fujita, K.: The potential of chatgpt as a self-diagnostic tool in common orthopedic diseases: exploratory study. Journal of Medical Internet Research **25**, 47621 (2023)

[3] Koga, S.: Exploring the pitfalls of large language models: inconsistency and inaccuracy in answering pathology board examination-style questions. medRxiv, 2023–08 (2023)

[4] Baumgartner, C.: The opportunities and pitfalls of chatgpt in clinical and translational medicine. Clinical and Translational Medicine **13**(3) (2023)

[5] Ten Berg, H., Bakel, B., Wouw, L., Jie, K.E., Schipper, A., Jansen, H., O'Connor, R.D., Ginneken, B., Kurstjens, S.: Chatgpt and generating a differential diagnosis early in an emergency department presentation. Annals of Emergency Medicine **83**(1), 83–86 (2024)

[6] Hirosawa, T., Kawamura, R., Harada, Y., Mizuta, K., Tokumasu, K., Kaji, Y., Suzuki, T., Shimizu, T.: Chatgpt-generated differential diagnosis lists for complex case–derived clinical vignettes: Diagnostic accuracy evaluation. JMIR Medical Informatics **11**, 48808 (2023)

[7] Narula, S., Karkera, S., Challa, R., Virmani, S., Chilukuri, N., Elkas, M., Thammineni, N., Kamath, A., Jaiswal, P., Krishnan, A.: Testing the accuracy of modern llms in answering general medical prompts

[8] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., Zhang, Y.: Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus **15**(6) (2023)

[9] Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023)

[10] Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T.: Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975 (2023)

[11] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[12] Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: Behrt: transformer for electronic health records. Scientific reports **10**(1), 7155 (2020)

[13] Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D.: Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine **4**(1), 86 (2021)

[14] Ziaei, R., Schmidgall, S.: Language models are susceptible to incorrect patient self-diagnosis in medical applications. arXiv preprint arXiv:2309.09362 (2023)

[15] Fansi Tchango, A., Goel, R., Wen, Z., Martel, J., Ghosn, J.: Ddxplus: A new dataset for automatic medical diagnosis. Advances in Neural Information Processing Systems **35**, 31306–31318 (2022)

[16] Sievert, M., Aubreville, M., Mueller, S.K., Eckstein, M., Breininger, K., Iro, H., Goncalves, M.: Diagnosis of malignancy in oropharyngeal confocal laser endomicroscopy using gpt 4.0 with vision. European Archives of Oto-Rhino-Laryngology, 1–8 (2024)

[17] Wang, J., *et al.*: The research about the innovative application in education field based on chatgpt foundation model. Adult and Higher Education **5**(15), 127–132 (2023)

[18] Li, Y., Liu, J., Yang, S.: Is chatgpt a good middle school teacher? an exploration of its role in instructional design. In: Proceedings of the 3rd International Conference on New Media Development and Modernized Education, NMDME 2023, October 13–15, 2023, Xi'an, China (2024)

[19] Liu, H.-C., Nataraj, V., Tsai, C.-T., Liao, W.-H., Liu, T.-Y., Jiang, M.T.-J., Day, M.-Y.: Imntpu at the ntcir-17 real-mednlp task: Multi-model approach to adverse drug event detection from social media. (No Title), (2023)

[20] Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., et al.: The finben: An holistic financial benchmark for large language models. arXiv preprint arXiv:2402.12659 (2024)

[21] Broyde, M.J.: Ai and jewish law: Seeing how chatgpt 4.0 looks at a novel issue. This article has been accepted for publication in a future issue of BDD-Bekhol Derakhekha Daehu of Bar Ilan University., CSLR Research Paper (12.2023-AFF) (2023)

[22] Rane, N., Choudhary, S., Rane, J.: Gemini versus chatgpt: Applications, performance, architecture, capabilities, and implementation. Performance, Architecture, Capabilities, and Implementation (February 13, 2024) (2024)

[23] Buscemi, A., Proverbio, D.: Chatgpt vs gemini vs llama on multilingual sentiment analysis. arXiv preprint arXiv:2402.01715 (2024)

[24] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)

[25] Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: International Conference on Machine Learning, pp. 3929–3938 (2020). PMLR