applied sciences

MDPI

Article

Exploring the Effectiveness of Large Language Models in Generating Clinical Differential Diagnoses: A Case Study of ICU Patients

Mohammad Alsulmi

Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia; malsulmi@ksu.edu.sa

Abstract

Recent advances in large language models (LLMs) have demonstrated their potential for a wide range of applications, including biomedical and clinical domains. Their ability to process and interpret large amounts of data, reason over clinical cases, and generate contextually relevant responses has created opportunities for their use in healthcare. In clinical settings, LLMs may assist clinicians in decision-making by generating and ranking potential differential diagnoses based on individual patient cases, thereby helping them consider a broader range of diagnostic possibilities. However, the rapid development of LLMs has resulted in an increasingly diverse range of models with different architectures, sizes, and capabilities, making it challenging to determine which models are most suitable for clinical diagnostic assistance tasks. Furthermore, clinical cases vary considerably in their characteristics and diagnostic complexity, potentially leading to differences in model performance across clinical conditions. This article evaluates 11 LLMs using 60 intensive care unit (ICU) patient cases covering a range of clinical conditions. Specifically, we assess the models' ability to generate ranked differential diagnoses that align with reference diagnoses and compare their diagnostic performance across models and cases. The results demonstrate that LLMs, particularly recent-generation models, can generate differential diagnoses that align with the reference diagnoses under the evaluated benchmark conditions, while revealing notable differences in performance across models and cases. These findings highlight the importance of systematically evaluating and carefully selecting models when investigating their potential application in clinical diagnosis.

Keywords: LLM evaluation; LLMs for healthcare; clinical decision support; artificial intelligence in healthcare; clinical case reasoning

Check for updates

Academic Editors: Paolo Renna and Arkadiusz Gola

Received: 2 October 2026
Revised: 30 October 2026
Accepted: 3 November 2026
Published: 5 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The application of computational methods to biomedical and clinical problems has a long history, with early efforts to support medical diagnosis predating the emergence of modern artificial intelligence (AI). Researchers explored mathematical, statistical, and probabilistic approaches, including Bayes' theorem, to support medical decision-making [1,2]. These efforts were followed by the development of computer-based systems for processing clinical information and supporting diagnostic decisions. As AI emerged as a distinct field, research increasingly focused on applying AI methods to represent medical knowledge and reason about clinical problems. This led to the development of knowledge-based and rule-based systems, commonly known as expert systems, that attempted to capture aspects

of expert clinical reasoning. These systems represented an important transition from purely statistical and algorithmic approaches toward systems that used inference mechanisms to generate diagnostic assessments based on stored domain knowledge [3,4].

Although knowledge-based expert systems demonstrated the feasibility of AI for clinical decision support, their adoption was constrained by several limitations. These systems typically relied on manually encoded rules and domain specific knowledge, requiring substantial effort from medical experts to develop, maintain, and update their knowledge bases. Their ability to generalize beyond the conditions and knowledge for which they were explicitly designed was also limited [5]. As increasing amounts of clinical data became available, particularly through the growing adoption of electronic health records (EHRs), machine learning (ML) introduced data-driven approaches that could learn patterns from clinical data rather than relying exclusively on manually specified rules [6,7]. This transition enabled the development of increasingly advanced models for a variety of clinical decision support tasks.

The emergence of deep learning further advanced these capabilities by enabling models to learn complex representations from large scale data. In particular, the development of large language models (LLMs) has expanded the potential applications of AI in healthcare. LLMs can process clinical text and generate responses that integrate information from different parts of a clinical case [8,9]. Their ability to process and interpret large amounts of information, reason over clinical cases, and generate contextually relevant responses has attracted increasing attention in biomedical and clinical research. LLMs have the potential to support clinical decision-making by generating and ranking potential differential diagnoses from a patient's clinical case [10–12]. Such capabilities may help clinicians consider a broader range of diagnostic possibilities. This capability is particularly relevant in complex clinical environments, where patients may present with multiple interacting conditions and symptoms associated with a wide range of diseases. However, the rapid emergence of LLMs has resulted in an increasingly diverse range of models with different architectures, sizes, and capabilities, making it challenging to determine which models are most suitable for clinical diagnostic support [10]. Furthermore, clinical cases vary considerably in their characteristics and diagnostic complexity, potentially leading to differences in model performance across models and clinical conditions.

To address these challenges, this article evaluates a diverse set of LLMs for differential diagnosis generation, including recent-generation models and older, relatively lightweight models suitable for local deployment. We use 60 intensive care unit (ICU) patient cases from the TREC Clinical Decision Support (CDS) 2014 and 2015 tracks [13,14], covering a range of clinical conditions. We assess the models' ability to generate differential diagnoses that align with the provided ground truth diagnoses and compare their performance across models and clinical cases. The results demonstrate that LLMs can generate differential diagnoses consistent with the reference diagnoses under the evaluation settings considered, while also revealing notable differences in performance across models and cases. These findings provide insights into the capabilities and limitations of LLMs for differential diagnosis generation and motivate further investigation of their potential role in clinical decision support.

The remainder of this article is organized as follows. Section 2 provides background on computational methods for clinical diagnosis and reviews related research employing AI based approaches. Section 3 describes the methodology and introduces the LLMs used in this work. Section 4 presents the results of evaluating the selected LLMs and provides an analysis and discussion of the findings. Finally, Section 5 extends our discussion and concludes the article by outlining directions for future research.

2. Related Work

2.1. Early Work on Computer-Based Clinical Diagnosis

The application of computational methods to clinical diagnosis predates the emergence of modern AI approaches. We note the early work of Ledley and Lusted [1,2], which conducted pioneering work examining how symbolic logic, probability theory, and value theory could provide formal foundations for medical diagnostic reasoning. Their work also highlighted the potential of computers to assist physicians in processing clinical information and generating diagnoses based on a set of patient symptoms. This work provided an important foundation for the following research in computer-assisted diagnosis and clinical decision support systems (CDSSs).

Later studies explored knowledge-based expert systems for clinical diagnosis. Shortliffe et al. [3] developed MYCIN, a rule-based system that used patient information and expert-derived rules to diagnose bacterial infections and recommend appropriate antimicrobial therapy. Evaluation studies, such as [4], reported that MYCIN could produce diagnostic and therapeutic recommendations comparable to those of physicians in selected cases. Similarly, Miller et al. [15] extended knowledge-based approaches by developing INTERNIST-I, a system designed to provide diagnostic capabilities across a broader range of general internal medicine cases. Its development demonstrated the potential of knowledge-based systems for complex clinical reasoning while also highlighting challenges associated with acquiring, representing, maintaining, and applying large-scale medical knowledge.

Researchers also investigated probabilistic approaches to medical diagnosis. Bayesian methods provided a mathematical framework for combining prior knowledge about diseases with observed clinical findings and updating diagnostic probabilities as new evidence became available [16]. For instance, de Dombal et al. [17] investigated the computer-assisted diagnosis of acute abdominal pain using clinical information and probabilistic methods. Their findings demonstrated that such methods could achieve diagnostic performance comparable to, and in some cases better than, that of clinicians for this specific medical condition.

These early studies established important foundations for clinical AI while also revealing challenges associated with the adoption of such systems. Particularly, knowledge-based systems typically relied on manually encoded rules and domain specific knowledge, requiring substantial effort from medical experts to develop, maintain, and update their knowledge bases [5]. These limitations motivated the development of data-driven ML approaches for clinical diagnosis as discussed in the following section.

2.2. The Emergence of ML and LLM Methods

Various studies have investigated the use of data-driven ML approaches for clinical diagnostic tasks. Many of these approaches have focused on diagnosing specific diseases or groups of diseases. For example, Palaniappan and Awang [18] investigated several ML algorithms, including naïve Bayes, decision trees, and neural networks, for heart disease diagnosis by estimating the likelihood of disease based on a set of patient findings. Similarly, Ferreira et al. [19] examined several ML techniques for diagnosing jaundice in newborns and reported that these models could be effectively applied to this diagnostic task. More recently, Liu et al. [20] investigated liver disease prediction using a range of ML methods, including neural networks, convolutional neural networks (CNNs), long short-term memory (LSTM) networks, support vector machines (SVMs), and XGBoost. Their results demonstrated the effectiveness of several of these approaches, with XGBoost achieving particularly strong performance for liver disease diagnosis. This category of research has been later extended through applying conventional ML methods to the diagnosis of a wide range of diseases, including cancer [21], respiratory syndromes [22], and cognitive impairments [23].

Although the studies in [18–23] demonstrated the potential of data-driven ML methods, many of these approaches were designed for relatively well-defined diagnostic tasks, typically focusing on distinguishing among a subset of predefined diseases. This limitation motivated research into more general approaches capable of considering a broader range of clinical information and generating multiple possible diagnoses. The recent emergence of LLMs has introduced a new paradigm for clinical diagnostic support, in which models can process clinical text and generate candidate diagnoses or differential diagnoses in natural language. Recent advances in LLMs, including models such as GPT [24], Gemini [25], and Claude [26], have demonstrated strong capabilities in understanding and generating natural language, enabling them to process complex and unstructured clinical information.

Several studies have investigated the application of LLMs to clinical diagnosis tasks. Mykhalko et al. [27] conducted an empirical evaluation of ChatGPT-3.5 using 50 clinical cases and reported a diagnostic accuracy of 66.0%. Similarly, Eriksen et al. [28] evaluated GPT-4 on complex medical cases and reported that the model correctly identified the diagnosis in 57% of the cases considered. Because these studies used different clinical cases, evaluation settings, and definitions of diagnostic correctness, their reported accuracy values are not directly comparable. Other studies have focused specifically on differential diagnosis rather than requiring an LLM to provide a single diagnosis. For instance, Hirosawa et al. [10] evaluated GPT-3.5 and GPT-4 for generating differential diagnoses and introduced top-k diagnostic accuracy as an evaluation measure, assessing whether the correct diagnosis appeared among the model's top-k candidate diagnoses. Likewise, Ríos-Hoyo et al. [11] examined the ability of GPT-4 to generate differential diagnoses and reported that the correct diagnosis appeared within the generated differential diagnoses for approximately two-thirds of the clinical cases considered.

Another direction of research has examined LLM performance in comparison with physicians or other clinical decision-support approaches. Ito et al. [29] compared GPT-4 with three physicians on 45 clinical cases involving diagnosis and triage and reported comparable performance between the LLM and physicians on the diagnostic task considered. Kim et al. [30] examined ChatGPT-4.0 for identifying obsessive–compulsive disorder and compared its performance with that of medical and mental-health professionals. Their results indicated that ChatGPT-4.0 achieved higher performance than the professionals included in their evaluation. In contrast to studies that primarily evaluate whether LLMs can match or exceed clinician performance, McDuff et al. [31] introduced AIME, a fine-tuned LLM designed for medical applications, and investigated the use of LLMs as assistive tools for clinicians. Their findings demonstrated improvements in clinicians' diagnostic accuracy when they were assisted by the LLM. Finally, Feldman et al. [32] compared ChatGPT-4.0 and Gemini 1.5 with a traditional clinical decision-support expert system using a set of clinical cases. Their results indicated that the evaluated LLMs did not outperform the expert system, leading the authors to advocate for hybrid LLM-CDSS approaches rather than relying on standalone LLMs for clinical diagnostic tasks.

Despite the growing body of research on LLMs for clinical diagnosis, several aspects remain insufficiently explored. First, many previous studies have evaluated only one or a small number of LLMs, limiting comparisons across different models and providers. In contrast, this work evaluates 11 LLMs, including recent-generation models accessed through cloud-based services and relatively lightweight, locally deployable models. The inclusion of locally deployable models is particularly relevant to clinical applications, as local deployment may help support data privacy by allowing clinical information to be processed within a local environment. Second, this article evaluates 60 clinical cases covering a range of conditions and considers both long and short versions of the same cases, enabling an assessment of model performance with different levels of clinical detail. Third, beyond examining whether

the correct diagnosis appears among the top-k predictions, this work uses Mean Reciprocal Rank (MRR) to assess the ranking of correct diagnoses and Differential Diagnosis Coverage (DDC), where applicable, to measure the extent to which models identify the multiple reference diagnoses associated with each case. Finally, all models are evaluated under a common experimental setting, facilitating a systematic comparison of their differential diagnosis performance across clinical cases and levels of detail.

3. Methodology

This section presents the methodology used to evaluate the effectiveness of LLMs for generating differential diagnoses in response to patient clinical cases. First, we describe the data set used in the work. We then introduce the LLMs included in the evaluation, followed by a detailed description of the data processing procedures and prompt formulation. Finally, we describe the evaluation measures used to assess the performance of the models.

3.1. Dataset

We use the clinical cases provided by the TREC Clinical Decision Support (CDS) track for the 2014 and 2015 tasks [13,14]. Each year includes 30 clinical cases involving intensive care unit (ICU) patients and covering a wide range of medical conditions, including tumor-related, neurological, abdominal, obstetric and gynecological, physical, and infectious conditions. The TREC CDS track uses these cases as queries representing clinicians' information needs when searching for relevant medical literature. Based on their underlying information needs, the cases are classified into three types: diagnosis, test, and treatment. In total, the dataset comprises 60 clinical cases.

For the 2014 cases, explicit diagnostic answers were not included with the original clinical cases, as the primary objective of the track was to evaluate information retrieval systems in clinical settings. However, we located the corresponding differential diagnoses in supplementary data provided by the track organizers. For most cases, physicians provided a list of differential diagnoses containing up to seven diagnoses, although some cases had only one reference diagnosis. On average, each case had 2.8 reference diagnoses.

In contrast, the 2015 dataset provides a single correct diagnosis for 20 of the 30 cases. For the remaining 10 cases, which are classified as diagnosis-type cases, we identified reference diagnoses using articles from PubMed Central (PMC) whose relevance had been assessed by physician assessors in the TREC CDS track. Specifically, we considered only articles assigned relevance label 2 (definitely relevant). We examined the articles associated with each case and identified diseases mentioned repeatedly across the relevant articles. To maintain consistency with the single-diagnosis format used for the other 2015 cases, we deliberately selected only one diagnosis per case. To support the verification of each selected diagnosis, we identified at least three PMC articles associated with the corresponding case that mentioned the same disease.

Although this procedure may not capture all possible differential diagnoses for a case, it provides a systematic approach to identifying a diagnosis mentioned across multiple highly relevant articles. The authors did not independently infer or generate additional diagnoses; instead, the selected diagnoses were based on recurring mentions in the examined articles. This identification process was conducted independently by the authors, rather than by physicians. The PMC identifiers of the supporting articles are provided in the supplementary data described in the Available Data section. Although these 10 cases are included in the overall evaluation presented in Section 4, their results are reported separately under the diagnosis case type in Table A5 in the Appendix A.

All clinical cases considered in this work are available in two formats: a long version containing detailed health related information about the admitted patient and a shorter

version containing key clinical information. Both versions are included in our evaluation to examine the impact of the amount of clinical information provided on model performance. Table 1 presents an example of these clinical cases along with a corresponding set of differential diagnoses.

Table 1. Clinical cases set provided in two formats long and short with a set of differential diagnosis (one or more diagnoses).

No	Clinical Case (Long)	Clinical Case (Short)	Differential Diagnosis
1	A 2-year-old boy is brought to the emergency department by his parents for 5 days of high fever and irritability. The physical exam reveals conjunctivitis, strawberry tongue, inflammation of the hands and feet, desquamation of the skin of the fingers and toes, and cervical lymphadenopathy with the smallest node at 1.5 cm. The abdominal exam demonstrates tenderness and enlarged liver. Laboratory tests report elevated alanine aminotransferase, white blood cell count of 17,580/mm, albumin 2.1 g/dL, C-reactive protein 4.5 mg, erythrocyte sedimentation rate 60 mm/h, mild normochromic, normocytic anemia, and leukocytes in urine of 20/mL with no bacteria identified. The echocardiogram shows moderate dilation of the coronary arteries with possible coronary artery aneurysm.	2-year-old boy with fever and irritability for 5 days. Physical exam findings include conjunctivitis, strawberry tongue, and desquamation of the fingers and toes. Lab results include low albumin, elevated white blood cell count and C-reactive protein, and urine leukocytes. Echo shows moderate dilation of the coronary arteries.	1. Kawasaki disease 2. Infectious exanthemas of childhood 3. reactions to drugs
2	A 28-year-old female with neck and shoulder pain and left hand and arm paresthesias three weeks after returning from a trip to California where she attended a stray animal recovery campaign. Her physical exam was unremarkable except for slight tremors and almost imperceptible spasticity. She was prescribed NSAIDS and a topical muscle relaxant. She was brought in to the ER three days later with spastic arm movements, sweating, increasing agitation and anxiety, malaise, difficul swallowing and marked hydrophobia, and was immediately hospitalized.	28-year-old female with neck pain and left arm numbness 3 weeks after working with stray animals. Physical exam initially remarkable for slight left arm tremor and spasticity. Three days later she presented with significant arm spasticity, diaphoresis, agitation, difficulty swallowing, and hydrophobia.	1. Rabies 2. Encephalitis 3. Delerium tremens 4. Diphtheria 5. Drug ingestion 6. Plant ingestion 7. Tetanus

3.2. Large Language Models Set

We consider a diverse set of LLMs from multiple vendors in our experiments. The models are divided into two groups based mainly on their generation and model size, which influence their capabilities, computational requirements, deployment options, and practical applicability to clinical environments. The first group contains recent-generation, relatively large-scale models accessed through cloud-based services. These models typically require substantial computational resources for inference and, in many cases, have closed weights, making local deployment infeasible. This group includes six models: Claude Fable 5.1 [26], GPT-6 Astra [24], DeepSeek V4 Pro [33], Gemini 3.8 Flash [25], Grok-4.6 [34], and Qwen 3.8 Max [35]. These models are accessed through cloud-based APIs using the LangChain library.

The second group consists of lightweight, open-weight and previous generation models that can be deployed locally on a user personal computer. Compared to the models in first

group, these models generally have more modest computational requirements, making local deployment more feasible. This group allows us to examine models that provide a more accessible deployment option for clinical environments where transferring sensitive patient information to external cloud services may be restricted by privacy regulations, institutional policies, or approval requirements. Local deployment allows clinical data to remain within the local computing environment, although appropriate privacy, security, and data-governance procedures are still required. Thus, the two groups allow us to investigate how difference model generation and size, along with their deployment options, relate to their performance and effectiveness in differential diagnoses generation task.

The second group includes DeepSeek R [36], Llama 3.2 [37], Qwen 3.0 [38], Granite 4.1 [39], and MedGemma 1.5 [40]. It is worth noting that MedGemma 1.5 is based on the Gemma 1.5 architecture and is specifically trained for medical applications. All models in this group were deployed locally using the Ollama framework [41]. All models in both groups were used with their default configuration settings. Table 2 summarizes the models included in this work, including their parameter sizes and weight status. Additionally, a summary of the default configurations used for both cloud-based and locally deployed models is provided in Table A1 in Appendix A.

Table 2. The LLMs considered in this work, showing their parameter sizes, deployment settings, and weight status. "n/d" indicates that the number of parameters was not disclosed by the vendor.

Model	Parameters	Deployment	Weight Status
Claude Fable 5.1	n/d	cloud	closed weights
GPT-6 Astra	n/d	cloud	closed weights
DeepSeek V4 Pro	49B (active)	cloud	open weights
Gemini 3.8 Flash	n/d	cloud	closed weights
Grok 4.6	n/d	cloud	closed weights
Qwen 3.8 Max	95B (active)	cloud	closed weights
DeepSeek-R1	8B	local	open weights
Llama 3.2	3B	local	open weights
MedGemma 1.5	4B	local	open weights
Granite 4.1	3B	local	open weights
Qwen3	0.6B	local	open weights

3.3. Data Parsing and Prompt Formulation

Both the 2014 and 2015 clinical case sets are provided in XML format. We wrote a program code using the W3C Document Object Model (DOM) to parse the XML files and extract the clinical cases. For each case, we extracted the long version from the <description> tag, the shorter version from the <summary> tag, and the corresponding case identifier from the <topic number> tag. Each clinical case was evaluated by every LLM twice: once using the long version and once using the short version. Thus, each of the 60 clinical cases was evaluated 22 times across the 11 models and two case versions.

Our prompt formulation follows a zero-shot prompting approach, in which the LLMs are provided only with the clinical case and the relevant evaluation guidelines, without any examples or demonstrations. Specifically, the prompt instructs each model to analyze the given clinical case and generate a ranked list of potential differential diagnoses. Following an approach similar to that of Savage et al. [42], we also ask each model to assign a confidence score between 0 and 100 to each diagnosis to indicate its confidence in that prediction. These scores are primarily used as a validation aid to check whether the generated diagnoses are ranked in decreasing order of relevance to the given clinical case, as intended. However, they are not assumed to represent calibrated probabilities of diagnostic correctness and are therefore excluded from the quantitative evaluation. In addition, because some models tend

to provide explanations or reasoning for their diagnostic suggestions, the prompt explicitly instructs them not to include such explanations. This instruction simplifies the processing and analysis of the generated outputs. Figure 1a,b show examples of outputs generated by two of the evaluated models, DeepSeek V4 Pro and GPT-6 Astra, for the clinical case no. 1 "A 2-year-old boy is brought to the emergency ...," which was presented previously in Table 1.

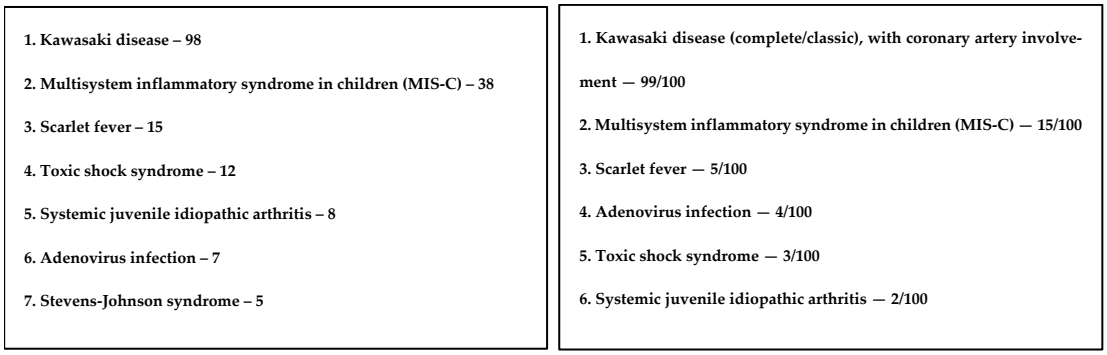


Figure 1. Sample differential diagnosis output generated by two LLMs: (**a**) DeepSeek V4 Pro, and (**b**) GPT-6 Astra for a clinical case. In addition to condition/disease name, each model outputs a score (0–100) indicating the model prediction confidence for that diagnosis.

3.4. Evaluation Measures and Settings

Once an LLM generates a list of differential diagnoses for a clinical case, each generated diagnosis is matched against the corresponding reference diagnoses provided in the dataset. We implemented a simple matching procedure to determine whether each generated diagnosis corresponds to a reference diagnosis or an accepted abbreviation or alternative name listed in the TREC CDS reference data. The resulting matches identify correct predictions and provide the basis for calculating the evaluation measures described below.

Some cases required additional consideration to ensure accurate matching. For example, an LLM might use an abbreviated disease name instead of its full name, include brackets or special characters, omit apostrophes, combine two distinct diagnoses into a single entry, or list both the full and abbreviated forms of the same diagnosis as separate entries in the ranked list. Such variations can complicate automated matching and result in incorrect matches or missed correspondences. To address these issues, we conducted an additional manual validation step to review ambiguous cases and verify the matches between model-generated diagnoses and the reference diagnoses. Nevertheless, some matching decisions may involve interpretive judgment, and residual errors cannot be entirely ruled out. This limitation should be considered when interpreting the evaluation results.

Following prior work [31,32], we first evaluate the models using a top-*k* measure, which assesses whether at least one ground truth diagnosis appears among the model's top *k* generated diagnoses. For a clinical case *c*, let G_c denote the set of ground truth diagnoses and P_c^k denote the first *k* diagnoses predicted (i.e., generated) by the model. The Top-*k* measure for a clinical case *c* can be defined as:

$$\text{Top-}k(c) = \begin{cases} 1, \ \textit{if } G_c \cap P_c^k \neq \emptyset, \\ 0, \ \textit{otherwise}. \end{cases} \quad (1)$$

The overall Top-*k* performance is then calculated as the proportion of clinical cases for which at least one ground truth diagnosis appears within the top *k* generated diagnoses:

$$Top - k = \frac{1}{N}\sum_{c=1}^{N} Top - k(c) \tag{2}$$

where N is the total number of clinical cases evaluated. We consider $k = 1$, $k = 3$, $k = 5$, and $k = max$, where max represents the maximum number of diagnoses generated by the model for case c.

Although Top-k accuracy indicates whether a correct diagnosis appears within the top-k predictions, it does not account for its exact position in the ranked list. For example, a correct diagnosis ranked first receives the same Top-5 result as one ranked fifth, provided that both appear within the top five. To account for the ranking position of correct diagnoses, we use Mean Reciprocal Rank (MRR) [43,44]. First, the Reciprocal Rank (RR) for a clinical case c is defined as the reciprocal of the rank position of the first correct diagnosis in the model's generated list. Let r_c denote the rank of the first generated diagnosis that matches any diagnosis in G_c. Then RR is defined as:

$$RR(c) = \frac{1}{r_c} \tag{3}$$

If no ground-truth diagnosis appears in the generated list, then RR = 0. MRR is calculated as the mean reciprocal rank across all N clinical cases:

$$MRR = \frac{1}{N}\sum_{c=1}^{N} RR(c) \tag{4}$$

Thus, MRR assigns higher scores when the first correct diagnosis appears near the top of the list and penalizes models when it appears further down.

While Top-k accuracy and MRR evaluate whether and how highly a correct diagnosis is ranked, they do not indicate how comprehensively the model covers the full set of differential diagnoses associated with a clinical case. We therefore use Differential Diagnosis Coverage (DDC), inspired by diversity and novelty measures in information retrieval and ranking [45–48], to assess the extent to which a model's generated differential diagnosis list covers the reference diagnoses. DDC is conceptually similar to recall, as both measure the proportion of reference diagnoses identified by a system. However, DDC emphasizes coverage of distinct diagnostic alternatives and accounts for synonymous terms referring to the same underlying condition, preventing such terms from artificially inflating the coverage score. For a clinical case c, let Gd_c denote the set of distinct ground truth diagnoses, and Pd_c denote the set of distinct diagnoses generated by the model. DDC is defined as:

$$DDC(c) = \frac{|Gd_c \cap Pd_c|}{|Gd_c|} \tag{5}$$

The overall DDC is calculated as the mean coverage across all clinical cases:

$$DDC = \frac{1}{N}\sum_{c=1}^{N} \frac{|Gd_c \cap Pd_c|}{|Gd_c|} \tag{6}$$

A DDC value of 1 indicates that the model identified all ground-truth differential diagnoses for a case, whereas a value of 0 indicates that none were identified. However, similar to recall, DDC may increase with the number of diagnoses generated, as longer lists provide more opportunities to cover the reference diagnoses. Therefore, a higher DDC should not be interpreted in isolation as evidence of better diagnostic performance. Instead, it should be considered alongside other evaluation measures, such as Top-k and MRR, to provide a more comprehensive assessment of diagnostic performance.

4. Results and Analysis

Having described our methodology in the previous section, this section presents the results of our evaluation of LLMs in generating differential clinical diagnoses. As described in Section 3.1, we consider 60 clinical cases from the TREC CDS 2014 and 2015 tracks. Tables 3 and 4 present the results for the 2014 and 2015 case sets, respectively, using several evaluation measures: Top-1, Top-3, Top-5, Top-max, Mean Reciprocal Rank (MRR), and Differential Diagnosis Coverage (DDC). Tables A2–A5 in the Appendix A provide additional details on these results. Because the ground-truth data for the 2015 case set contain only a single diagnosis for each case, DDC is not applicable to this set and is therefore omitted. To guide our exploratory analysis, we investigate the following research questions (RQs):

- **RQ1**: How effectively do the evaluated LLMs generate clinically relevant differential diagnoses for the clinical cases?
- **RQ2**: How comprehensively do these LLMs cover the reference diagnoses provided by physicians?
- **RQ3**: How does diagnostic performance vary across the evaluated LLMs?
- **RQ4**: How does the amount of clinical information provided to the LLMs affect their diagnostic performance?

Table 3. Effectiveness of LLMs for generating differential diagnoses on the 30 TREC CDS 2014 clinical cases, evaluated using Top-*k*, Differential Diagnosis Coverage (DDC), and Mean Reciprocal Rank (MRR). Underlined values indicate the highest-performing model(s). Additional analyses, including confidence intervals and performance by topic type, are provided in the Appendix A (Tables A2 and A4).

Recent Generation LLMs							
LLM Model	**Case Type**	**Top-1**	**Top-3**	**Top-5**	**Top-Max**	**MRR**	**DCC**
Claude Fable 5.1	Long	0.63	0.90	0.97	<u>1.00</u>	0.78	0.82
Claude Fable 5.1	short	0.70	0.90	<u>1.00</u>	<u>1.00</u>	0.82	0.82
GPT-6 Astra	Long	0.70	0.87	0.90	0.93	0.79	0.75
GPT-6 Astra	short	<u>0.80</u>	0.87	0.90	0.90	<u>0.84</u>	0.73
DeepSeek V4 Pro	Long	0.73	0.93	0.97	0.97	0.83	0.75
DeepSeek V4 Pro	short	0.67	0.83	0.90	0.97	0.76	0.74
Gemini 3.8 Flash	Long	0.63	0.87	0.93	0.93	0.76	0.71
Gemini 3.8 Flash	short	0.70	0.87	0.93	0.93	0.79	0.78
Grok 4.6	Long	0.77	0.87	0.93	0.93	0.83	0.69
Grok 4.6	short	0.77	<u>0.93</u>	0.97	0.97	<u>0.84</u>	0.76
Qwen 3.8 Max	Long	0.63	<u>0.93</u>	0.93	0.97	0.77	<u>0.83</u>
Qwen 3.8 Max	short	0.77	0.90	0.93	0.93	<u>0.84</u>	<u>0.83</u>
Previous Generation LLMs							
DeepSeek-R1	Long	0.23	0.30	0.30	0.30	0.27	0.21
DeepSeek-R1	short	0.13	0.27	0.30	0.30	0.21	0.18
Llama 3.2	Long	0.57	0.70	0.77	0.77	0.64	0.52
Llama 3.2	short	0.70	0.80	0.83	<u>0.90</u>	0.77	<u>0.65</u>
MedGemma 1.5	Long	<u>0.73</u>	<u>0.90</u>	<u>0.90</u>	<u>0.90</u>	<u>0.81</u>	0.62
MedGemma 1.5	short	0.60	0.80	0.80	0.87	0.71	0.60
Granite 4.1	Long	0.60	0.77	0.80	0.80	0.68	0.52
Granite 4.1	short	0.63	0.80	0.80	0.80	0.71	0.59
Qwen3	Long	0.43	0.50	0.57	0.57	0.48	0.37
Qwen3	short	0.47	0.73	0.73	0.73	0.59	0.48

Table 4. Effectiveness of LLMs for generating differential diagnoses on the 30 TREC CDS 2015 clinical cases, evaluated using Top-k, and Mean Reciprocal Rank (MRR). Underlined values indicate the highest-performing model(s). Additional analyses, including confidence intervals and performance by topic type, are provided in the Appendix A (Tables A3 and A5).

Recent Generation LLMs						
LLM Model	**Case Type**	**Top-1**	**Top-3**	**Top-5**	**Top-Max**	**MRR**
Claude Fable 5.1	Long	0.90	0.97	0.97	1.00	0.94
Claude Fable 5.1	short	0.83	0.90	0.97	0.97	0.88
GPT-6 Astra	Long	0.90	0.97	0.97	0.97	0.93
GPT-6 Astra	short	0.80	0.90	0.97	0.97	0.86
DeepSeek V4 Pro	Long	0.87	0.93	0.97	0.97	0.90
DeepSeek V4 Pro	short	0.80	0.90	0.93	0.93	0.85
Gemini 3.8 Flash	Long	0.87	0.97	0.97	0.97	0.91
Gemini 3.8 Flash	short	0.80	0.90	0.97	0.97	0.87
Grok 4.6	Long	0.87	0.93	0.97	0.97	0.91
Grok 4.6	short	0.80	0.87	0.97	0.97	0.85
Qwen 3.8 Max	Long	0.90	0.93	0.97	0.97	0.93
Qwen 3.8 Max	short	0.83	0.93	0.97	1.00	0.89
Previous Generation LLMs						
DeepSeek-R1	Long	0.10	0.17	0.17	0.17	0.13
DeepSeek-R1	short	0.23	0.30	0.30	0.30	0.26
Llama 3.2	Long	0.43	0.57	0.67	0.73	0.52
Llama 3.2	short	0.67	0.73	0.77	0.80	0.71
MedGemma 1.5	Long	0.47	0.70	0.77	0.87	0.59
MedGemma 1.5	short	0.60	0.73	0.73	0.87	0.68
Granite 4.1	Long	0.40	0.67	0.73	0.73	0.53
Granite 4.1	short	0.53	0.73	0.80	0.80	0.64
Qwen3	Long	0.27	0.37	0.43	0.43	0.33
Qwen3	short	0.40	0.47	0.53	0.53	0.44

4.1. RQ1: The Effectiveness of LLMs in Generating Relevant Diagnoses

RQ1 examines the effectiveness of the evaluated LLMs in generating relevant diagnoses for the given clinical cases. As shown in Tables 3 and 4, the results indicate that the considered LLMs were generally able to identify at least one correct diagnosis for many of the clinical cases, particularly when considering the Top-3 and Top-5, Top-*max* measures. For the recent-generation models, the average Top-1 performance across the long and short case versions was approximately 0.71 for the 2014 cases and increased to 0.85 for the 2015 cases. When considering the Top-5 measure (i.e., moving down a bit in the ranked list), the average performance increased to approximately 0.94 and 0.96 for the 2014 and 2015 cases, respectively. When all diagnoses generated for each case were considered (Top-max), the performance reached approximately 0.95 for the 2014 cases and 0.97 for the 2015 cases. Notably, Claude Fable 5.1 identified at least one correct diagnosis for all cases in both the 2014 and 2015 sets. Overall, these results indicate that, for most clinical cases, at least one reference diagnosis appeared among the first few diagnoses generated by the recent-generation models.

The MRR results provide an additional perspective on the quality of the generated ranked diagnoses. MRR accounts for the rank position of the first correct diagnosis and introduces a penalty if it appears down in the ranked list. Across the long and short case versions, the average MRR for the recent-generation models was approximately 0.80 for the 2014 cases and 0.89 for the 2015 cases. These results suggest that the first correct diagnosis was generally ranked relatively high in both case sets, and particularly in the 2015 cases. Although the models achieved high Top-5 and Top-max scores, their lower Top-1 scores

indicate that the first prediction (i.e., the first generated diagnosis) was not always correct, and additional predictions were sometimes needed to identify a correct diagnosis. MRR complements these measures by distinguishing between cases in which the first correct diagnosis appears near the top of the list and those in which it appears at a lower rank.

In comparison, the lightweight earlier-generation models exhibited substantial variation in performance across both models and clinical cases. Their average Top-1 performance was approximately 0.51 for the 2014 cases and 0.41 for the 2015 cases, while the corresponding Top-5 performance increased to approximately 0.68 and 0.59, respectively. When all generated diagnoses were considered (Top-max), the average performance reached approximately 0.69 for the 2014 cases and 0.62 for the 2015 cases. Their average MRR was approximately 0.59 for the 2014 cases and 0.48 for the 2015 cases, indicating that the first correct diagnosis tended to appear lower in the generated rankings than for the recent generation models. DeepSeek-R1 performed particularly poorly in this group, with MRR values of 0.27 and 0.21 for the long and short versions of the 2014 cases, respectively, and 0.13 and 0.26 for the corresponding 2015 cases. Thus, although these models identified relevant diagnoses for a substantial subset of the cases, their performance was less consistent than that of the recent generation models. DeepSeek-R1, in particular, exhibited consistently weak performance across both case sets relative to the other models in its group.

The results also reveal a noticeable difference between the two TREC CDS case sets. The recent-generation models generally achieved higher performance on the 2015 cases than on the 2014 cases. This difference is evident in both Top-1 accuracy, which increased from approximately 0.71 to 0.85, and MRR, which increased from approximately 0.80 to 0.89. In contrast, the previous generation models exhibited lower average performance on the 2015 cases than on the 2014 cases, with both Top-1 accuracy and MRR values decreasing. These differences suggest that the effectiveness of LLM-generated differential diagnoses may vary depending on both the characteristics of the clinical cases and the capabilities of the models being evaluated.

Compared with prior studies evaluating LLMs for clinical diagnosis generation [27–29, 31,32], our results indicate comparatively high performance, particularly among the recent-generation models. Several factors may contribute to this difference, including variations in the clinical case sets, evaluation protocols, and models examined across studies. In addition, advances in LLM capabilities since the publication of earlier studies may have contributed to the higher performance observed in our experiments, as many previously evaluated models belong to earlier generations. Nevertheless, direct comparisons with prior studies should be interpreted cautiously because of differences in datasets, evaluation settings, and model selection. The performance differences observed between the two TREC CDS case sets in our own experiments further highlight the importance of accounting for case characteristics when comparing diagnostic performance.

4.2. RQ2: The Comprehensiveness of the Generated Differential Diagnoses

This RQ examines how comprehensively the evaluated LLMs cover the physician-provided differential diagnoses. In other words, we aim to quantify how well these models identify the complete set of reference diagnoses provided for each clinical case. To address this question, we use the Differential Diagnosis Coverage (DDC) measure described in Section 3.3. Unlike the Top-*k* and MRR measures considered in RQ1, which determine whether at least one physician-provided diagnosis appears among the model's generated diagnoses, DDC measures the proportion of physician-provided differential diagnoses identified by the model for each case. This analysis is conducted using the TREC CDS 2014 cases, for which multiple physician-provided differential diagnoses are available.

The DDC results in Table 3 show that the recent-generation models generally provided substantially more comprehensive differential diagnoses than the previous generation models. Across the six models in first group and considering both case versions, the average DDC was approximately 0.77. In comparison, the corresponding average for the five LLMs in second group was approximately 0.47. Thus, on average, the recent-generation models identified more than three-quarters of the physician-provided reference diagnoses, whereas the lightweight, earlier-generation models identified fewer than half.

Among the recent-generation models, Qwen 3.8 Max and Claude Fable 5.1, achieved relatively high coverage, with DDC values of 0.83 and 0.82, respectively, for both case versions. In contrast, Grok 4.6 obtained the lowest DDC among this group of LLMs for the long case version, with a value of 0.69. Among the lightweight, earlier-generation models, DDC values ranged from 0.18 to 0.65 across the two case versions, indicating considerable variation in their ability to cover the physician-provided reference diagnoses.

However, DDC should also be interpreted while considering the number of diagnoses generated by each LLM. Table 5 reports the average and maximum numbers of generated diagnoses for the TREC CDS 2014 and 2015 cases. Focusing on 2014 cases, the recent generation models are shown to generate an average of approximately 6.9 diagnoses per case, compared with approximately 5.6 for the previous generation models. A similar pattern was observed for the 2015 cases, with averages of approximately 7.0 and 5.6 diagnoses per case, respectively. Nevertheless, output length varied substantially among individual models. For example, Llama 3.2 generated an average of 7.1 diagnoses for the 2014 cases, exceeding the averages for several recent generation models, including Grok 4.6, GPT-6 Astra, DeepSeek V4 Pro, and Gemini 3.8 Flash. These results suggest that differences in output length may contribute to differences in DDC, but list length alone does not fully explain the observed variation. Because longer lists provide more opportunities to identify reference diagnoses, DDC should not be interpreted in isolation as evidence of superior diagnostic performance.

Table 5. Average and maximum numbers of differential diagnoses generated by each LLM for the TREC CDS 2014 and 2015 clinical cases.

Model	TREC CDS 2014		TREC CDS 2015	
	Avg Diagnoses	**Max. Diagnoses**	**Avg Diagnoses**	**Max. Diagnoses**
Claude Fable 5.1	9.8	19	9.8	18
GPT-6 Astra	5.5	12	5.7	10
DeepSeek V4 Pro	7.0	13	7.1	13
Gemini 3.8 Flash	5.8	10	6.0	12
Grok 4.6	4.7	9	5.0	8
Qwen 3.8 Max	8.5	20	8.5	16
DeepSeek-R1	4.6	15	4.4	25
Llama 3.2	7.1	11	7.5	17
MedGemma 1.5	7.9	20	8.0	21
Granite 4.1	4.6	8	4.5	12
Qwen3	3.6	7	3.7	7

It is also worth noting that, given the nature of the diagnostic task, the differential diagnosis lists generated in this work were generally relatively short. Although some models occasionally produced longer lists (e.g., Claude Fable 5.1 generated as many as 19 differential diagnoses for a single case), these list lengths are not comparable to those used in conventional information retrieval and recommendation tasks, where ranked lists typically contain a much larger number of items (e.g., 100 or more).

The DDC results also reveal that identifying at least one relevant diagnosis does not necessarily imply comprehensive coverage of the differential diagnosis. For instance, MedGemma 1.5 (from the second group) achieved a Top-1 score of 0.73 and MRR of 0.81 for the long case versions but a DDC of only 0.62. Conversely, Qwen 3.8 Max (from the first group) achieved a Top-1 score of 0.63 and MRR of 0.77 for the long cases while obtaining a substantially higher DDC score of 0.83. This difference illustrates the complementary nature of these measures: Top-*k* can capture whether a model identifies at least one relevant diagnosis within a specified rank, MRR accounts for the rank of the first correct diagnosis in the generated list, whereas DDC can capture the proportion of the physician-provided differential diagnoses covered by the model's generated output.

To sum up, the RQ2 analysis suggests that the evaluated LLMs differ not only in their ability to identify at least one relevant diagnosis but also in the comprehensiveness of the differential diagnoses they generate. The relatively higher DDC values observed for the recent-generation models indicate broader coverage of physician-provided diagnostic alternatives, while the greater variation among the lightweight, earlier-generation models highlights differences in their ability to cover these alternatives. These findings should be interpreted alongside the number of diagnoses generated and the results of Top-k accuracy and MRR to provide a more balanced assessment of diagnostic performance.

4.3. RQ3: Diagnostic Performance Across LLMs

RQ3 examines differences in diagnostic performance among the evaluated LLMs. In particular, this RQ focuses on differences among individual models and between the two model groups. We consider both the Top-*k* and MRR results across the 2014 and 2015 clinical case sets and, where applicable, the DDC results for the 2014 cases.

The results show substantial variation in diagnostic performance across the evaluated LLMs. This variation is particularly noticeable between the recent-generation models and the previous-generation models. Across both case sets, the recent generation models generally achieve higher performance for Top-*k*, MRR, and DDC, whereas the lightweight previous generation models show greater variability in their results. However, performance also varies among models within each group, indicating that model generation alone does not fully explain the observed differences.

To examine performance variation within each model group, we apply two-way ANOVA without replication, treating the LLMs as columns and the clinical cases as rows. The results for MRR and DCC are reported in Tables 6 and 7, respectively. For the recent-generation models, the ANOVA results show no statistically significant differences among models in MRR for either the 2014 or 2015 case set, regardless of whether the long or short case descriptions are used. These findings suggest that the evaluated models exhibit broadly comparable performance in ranking correct diagnoses, as measured by the reciprocal rank of the first correct diagnosis. Thus, with respect to MRR, the choice among these recent-generation models appears to have a relatively limited effect on performance for the cases examined.

However, the DCC results reveal a more different picture. For the 2014 cases, the differences among recent-generation models are statistically significant for the long case descriptions, but not for the short versions (as shown in Table 7). This finding suggests that differences in diagnostic coverage among these models may depend on the amount of clinical information provided. However, this interpretation cannot be directly examined using the 2015 cases because their reference data do not contain multiple differential diagnoses for each case. Qwen 3.8 Max and Claude Fable 5.1 achieve the highest DDC values, suggesting that these models provide broader coverage of the physician-provided differential diagnoses than some of the other recent-generation models for the given cases.

Table 6. Two-way ANOVA without replication results for the MRR measure across the TREC CDS 2014 and 2015 clinical cases.

MRR Analysis	Recent Generation Models		Previous Generation Models	
	Long	Short	Long	Short
TREC CDS 2014				
#LLMs compared	6	6	5	5
#clinical cases	30	30	30	30
degrees of freedom	5, 145	5, 145	4, 116	4, 116
F-statistic	0.88	1.17	9.1	16.4
p-value	0.50	0.33	2.11×10^{-6}	1.08×10^{-10}
conclusion	not significant	not significant	significant	significant
TREC CDS 2015				
#LLMs compared	6	6	5	5
#clinical cases	30	30	30	30
degrees of freedom	5, 145	5, 145	4, 116	4, 116
F-statistic	1.16	0.78	10.6	9.42
p-value	0.33	0.57	2.38×10^{-7}	1.24×10^{-6}
conclusion	not significant	not significant	significant	significant

Table 7. Two-way ANOVA without replication results for the DCC measure across the TREC CDS 2014 clinical cases.

DCC Analysis	Recent Generation Models		Previous Generation Models	
	Long	Short	Long	Short
#LLMs compared	6	6	5	5
#clinical cases	30	30	30	30
degrees of freedom	5, 145	5, 145	4, 116	4, 116
F-statistic	5.73	1.74	9.63	16.70
p-value	7.45×10^{-5}	0.13	9.30×10^{-7}	7.79×10^{-11}
Conclusion	significant	not significant	significant	significant

In contrast, the earlier-generation models exhibit statistically significant differences in diagnostic performance across all the examined measures. For MRR, the differences among models are statistically significant for both the 2014 and 2015 case sets, using both long and short case versions as shown in Table 6. The same pattern is observed for DDC in the 2014 cases as shown in Table 7. Thus, unlike the recent-generation models, the previous generation models exhibit statistically significant differences in diagnostic performance across the evaluated measures. These findings indicate greater heterogeneity in the diagnostic performance of these models, including their ability to rank correct diagnoses near the top of the generated lists. Within this group, models such as MedGemma 1.5 and Llama 3.2 demonstrate relatively strong performance compared with their group counterparts. Nevertheless, their performance remains below that of the recent generation models on several of the evaluated measures.

These findings are consistent with previous research, such as Sandmann et al. [49], which highlights the importance of empirically evaluating individual LLMs for clinical decision-making rather than assuming that particular model characteristics guarantee better performance. Specifically, Sandmann et al. found that selected open-weight models performed comparably to proprietary, closed-weight models on the clinical tasks they

evaluated. However, differences in datasets, model selection, and evaluation methods should be considered when interpreting findings across studies.

Overall, the findings indicate that differences among LLMs depend on both the model group and the evaluation measure. The recent-generation models exhibit statistically comparable MRR performance, whereas their DCC performance differs significantly for long cases. The lightweight, earlier-generation models, by comparison, show statistically significant differences across both MRR and DDC. These findings highlight the importance of selecting models based on their performance across multiple measures and clinical cases. Model selection may be particularly consequential among the lightweight, earlier-generation models, while differences in diagnostic coverage also warrant consideration when choosing among recent-generation models.

4.4. RQ4: Amount of Clinical Information vs. Diagnostic Performance

This RQ examines whether the amount of clinical information provided to the LLMs affects their diagnostic performance. Specifically, we investigate whether providing longer, more detailed clinical cases improve diagnostic effectiveness compared with shorter versions of the same cases.

As shown in Table 3, the results for the 2014 case set reveal no consistent advantage for either long or short case descriptions. Among the recent-generation models, short descriptions result in higher Top-1 accuracy for several models, while the differences at higher k values are generally smaller. The MRR results follow a similar pattern: four of the six recent-generation models achieve higher MRR with short descriptions, whereas DeepSeek V4 Pro performs better with long descriptions. A similar pattern can be observed for the previous generation models, where four of the five models achieved higher performance with short cases. MedGemma 1.5 is the exception, achieving higher performance across all measures with long descriptions.

The results for the 2015 case set, presented in Table 4, show a clearer distinction among the two model groups. All six recent-generation models achieve higher MRR with long descriptions than with short descriptions. Their Top-1 accuracy also generally favors long descriptions, although the differences at higher k values are smaller. In contrast, all five earlier-generation models achieve higher MRR with short descriptions. Their Top-k results show a similar pattern, with short descriptions generally yielding better performance.

The DDC results for the 2014 cases provide an additional perspective. Across the recent-generation models, the average DDC was 0.76 for the long case versions and 0.78 for the short versions. Similarly, the previous generation models achieved an average DDC of 0.45 for the long cases and 0.50 for the short cases. Thus, providing longer versions of the clinical cases did not result in higher average differential diagnosis coverage. Instead, the effect varied across models, with some models achieving higher coverage using the long versions and others using the short versions.

To summarize, these findings suggest that providing additional clinical information does not necessarily improve diagnostic performance. The effect appears to vary depending on the model and case set, with longer clinical descriptions providing benefits for some models and cases but not for others. Thus, the relationship between the amount of clinical information and LLM diagnostic performance appears to be model and case dependent rather than uniform across clinical cases.

An important consideration when interpreting these findings is that the short-form clinical cases used in this study are curated summaries that retain key clinical information. They do not represent the limited information that may be available during the early stages of a real-world clinical assessment. Consequently, good performance on short-form cases does not necessarily indicate that less clinical information is sufficient for clinical diagnosis.

5. Further Discussions and Concluding Remarks

The findings of this work demonstrate that LLMs, particularly recent-generation models, can generate differential diagnoses that align with ground-truth diagnoses across a range of ICU clinical cases under the considered experiment settings. Across the evaluated cases, these models were generally able to generate at least one correct diagnosis within their top-ranked outputs and achieved relatively high coverage of the reference diagnoses provided. These results suggest that LLMs have considerable potential for interpreting complex clinical information and generating plausible diagnostic alternatives. At the same time, the observed variation across models and case sets indicates that their performance is not uniform, highlighting the importance of evaluating individual models rather than treating LLMs as interchangeable diagnostic systems.

An important observation is that the recent-generation models generally outperformed the lightweight, earlier-generation models across the evaluated measures. This difference may reflect the continuous evolution of LLMs through advances in model architecture, scale, training, and data. Although this work does not isolate the contribution of these individual factors, the results are consistent with the broader improvements in LLM capabilities across successive generations.

However, a notable limitation of this work is the relatively small sample of clinical cases included in the evaluation. Although the TREC CDS 2014 and 2015 collections are well-established benchmarks for clinical information retrieval systems, our evaluation relies on only 60 ICU cases. Consequently, this sample may not adequately represent the diversity of clinical conditions, patient demographics, and diagnostic complexities encountered in broader clinical settings.

Another limitation concerns the nature of the clinical cases used in this work. These cases represent clinical situations in which an initial assessment has already taken place and additional information has been obtained through physical examinations, laboratory investigations, imaging, and other diagnostic procedures. Consequently, the information provided to the LLMs may represent a relatively advanced stage of the diagnostic process rather than the limited information initially available when a patient first presents. Furthermore, some of the reported clinical findings in those cases can be highly suggestive of, or closely associated with, the eventual diagnosis. Such findings may provide strong diagnostic clues that make it easier for an LLM to identify the corresponding condition, potentially resulting in an optimistic assessment of its diagnostic capabilities.

The examples presented in Table 1 illustrate this issue. By the time the complete case information is presented to an LLM, it includes findings obtained through multiple stages of clinical assessment, including physical examinations and diagnostic investigations. Some of these findings may be particularly informative or characteristic of the underlying condition, potentially making the correct diagnosis more apparent. However, obtaining such information is itself part of the diagnostic process. The attending physician must recognize which findings are relevant, decide which examinations to perform, and determine which laboratory or imaging investigations should be requested. The LLM, by contrast, receives the resulting clinical information and is then asked to generate a differential diagnosis. Thus, the model is evaluated primarily on its ability to interpret an already collected set of clinical evidence rather than on its ability to determine what information should be collected in the first place.

This experimental design makes the evaluated task less challenging than generating a differential diagnosis at initial patient presentation, which may lead to an optimistic assessment of LLM diagnostic performance. Accordingly, these findings should be interpreted in the context of this experimental setting. While they support further investigation of LLMs as assistive tools for interpreting clinical evidence, this work did not evaluate LLM use in

prospective clinical practice, clinician–AI interactions, decision-making changes, workflow impacts, or patient outcomes. Consequently, these findings cannot establish whether integrating LLMs into clinical practice would improve diagnostic accuracy, healthcare delivery, or patient care.

Another important consideration is clinical safety. While the evaluation assesses diagnostic accuracy in terms of the inclusion of correct diagnoses, it does not systematically account for the potential harm associated with incorrect diagnostic suggestions. Such suggestions may lead to unnecessary investigations, unsuitable treatments, or increased clinical workload, particularly when serious conditions are suggested without sufficient supporting evidence. Therefore, high performance in generating correct diagnoses does not necessarily imply clinical safety. Although evaluating the clinical consequences of incorrect LLM-generated diagnoses is beyond the scope of this work, these risks should be considered when interpreting the findings and warrant further investigation.

Future research should therefore examine LLMs under more realistic, progressively revealed information settings. For instance, an LLM could initially receive limited patient information, identify relevant follow-up questions and examinations or tests, and then update its differential diagnosis as additional information becomes available. Such an interactive approach would provide a more realistic assessment of LLMs as clinical information gathering and decision support tools, while also evaluating their ability to recognize missing information, select relevant investigations, and revise their diagnostic suggestions as new evidence emerges. In addition, future work involving clinicians is needed to determine whether LLM-generated differential diagnoses improve diagnostic decisions, how clinicians interpret and use model suggestions, and whether their integration affects workflow, efficiency, diagnostic errors, or patient outcomes. Such evaluations would help establish whether the capabilities observed in our benchmarked experiments translate into meaningful benefits in clinical practice.

To sum up, this article demonstrates the potential of LLMs, particularly recent-generation models, in generating differential diagnoses for ICU clinical cases. However, the results should not be interpreted as evidence that LLMs can independently perform clinical diagnosis. Their performance depends on the model, the clinical case, and the information provided to them, and the cases evaluated in this work may contain substantially more diagnostic information than would be available at the beginning of a real clinical encounter. Furthermore, the evaluation was limited to 60 clinical cases from the TREC CDS 2014 and 2015 datasets, which may limit the statistical precision and generalizability of the findings. Evaluations on larger and more diverse clinical benchmarks could strengthen the evidence regarding the diagnostic capabilities of LLMs. The findings therefore support further investigation of LLMs as assistive technologies integrated into a clinician diagnostic workflow rather than as standalone diagnostic systems.

Funding: This research was funded by the Ongoing Research Funding program (ORF-2026-1799), King Saud University, Riyadh, Saudi Arabia.

Data Availability Statement: The data, source code, and LLM-generated outputs used in this study are publicly available in our GitHub repository at https://github.com/malsulmi8-eng/LLMDiagnosis (accessed on 2 October 2026).

Acknowledgments: The author would like to thank the Ongoing Research Funding program (ORF-2026-1799), King Saud University, Riyadh, Saudi Arabia for supporting this work.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

Appendix A.1

As described in Section 3.2, all models considered in this work were run using their default platform configurations. Table A1 summarizes the configuration parameters for both cloud-based and locally deployed models.

Table A1. Default platform configurations of the cloud-based and locally deployed LLMs evaluated in this study.

LLM Model	Default Running Configuration
Claude Fable 5.1	Version = claude-fable-5-1, Effort = high, Context window = 1M tokens, Max output = 128K tokens, number of inferences = 2, Access date = 20 September 2026
GPT-6 Astra	Version = gpt-6-astra, Effort = medium, Context window = 1M tokens, Max output = 128K tokens, number of inferences = 2, Access date = 20 September 2026
DeepSeek V4 Pro	Version = DeepSeek-V4-Pro-0813, Effort = high, Context window = 1M tokens Max output = 384K tokens, number of inferences = 2, Access date = 21 September 2026
Gemini 3.8 Flash	Version = gemini-3.8-flash, Effort = medium, Context window = 1M tokens, Max output = 65K tokens, number of inferences = 2, Access date = 21 September 2026
Grok 4.6	Version = grok-4.6, Effort = high, Context window = 500K tokens, Max output = 64K tokens, number of inferences = 2, Access date = 21 September 2026
Qwen 3.8 Max	Version = qwen3.8-max-0902, Effort = extra high, Context window = 1M tokens, Max output = 64K tokens, number of inferences = 2, Access date = 21 September 2026
DeepSeek-R1	Version = DeepSeek-R1-0528, Temperature = 0.8, Top-p = 0.9, Context window = 8K tokens (ollama), Max output = 32K tokens, number of inferences = 2
Llama 3.2	Version = Llama-3.2-3B, Temperature = 0.8, Top-p = 0.9, Context window = 8K tokens (ollama), Max output = 128K tokens, number of inferences = 2
MedGemma 1.5	Version = google/medgemma-1.5-4b, Temperature = 0.8, Top-p = 0.9, Context window = 8K tokens (ollama), Max output = 8K tokens, number of inferences = 2
Granite 4.1	Version = granite-4.1-3b-base, Temperature = 0.8, Top-p = 0.9, Context window = 8K tokens (ollama), Max output = 8K tokens, number of inferences = 2
Qwen3	Version = Qwen3-0.6B, Temperature = 0.8, Top-p = 0.9, Context window = 8K tokens (ollama), Max output = 32k tokens, number of inferences = 2

Appendix A.2

To further assess the performance results presented in Section 4, we calculated 95% confidence intervals (CIs) for each LLM's performance on Top-1 accuracy, mean reciprocal rank (MRR), and differential diagnosis coverage (DDC), where applicable. These intervals provide additional information about the uncertainty associated with the reported performance estimates. The analysis was conducted separately for the TREC CDS 2014 and 2015 clinical case sets. The results for the 2014 and 2015 datasets are reported in Tables A2 and A3, respectively.

Table A2. Performance of LLMs on the 30 clinical cases from the TREC CDS 2014 dataset, measured using Top-1 accuracy, MRR, and DDC with corresponding 95% confidence intervals (CIs).

LLM Model	Case Type	Top-1	MRR	DCC
Claude Fable 5.1	Long	0.63, 95% CI [0.46, 0.81]	0.78, 95% CI [0.67, 0.89]	0.82, 95% CI [0.72, 0.90]
Claude Fable 5.1	short	0.70, 95% CI [0.53, 0.87]	0.82, 95% CI [0.72, 0.92]	0.82, 95% CI [0.72, 0.90]

Table A2. *Cont.*

LLM Model	Case Type	Top-1	MRR	DCC
GPT-6 Astra	Long	0.70, 95% CI [0.53, 0.87]	0.79, 95% CI [0.67, 0.91]	0.75, 95% CI [0.62, 0.86]
GPT-6 Astra	short	0.80, 95% CI [0.65, 0.95]	0.84, 95% CI [0.72, 0.96]	0.73, 95% CI [0.60, 0.85]
DeepSeek V4 Pro	Long	0.73, 95% CI [0.57, 0.89]	0.83, 95% CI [0.72, 0.94]	0.75, 95% CI [0.64, 0.86]
DeepSeek V4 Pro	short	0.67, 95% CI [0.50, 0.84]	0.76, 95% CI [0.63, 0.89]	0.74, 95% CI [0.62, 0.85]
Gemini 3.8 Flash	Long	0.63, 95% CI [0.46, 0.81]	0.76, 95% CI [0.64, 0.88]	0.71, 95% CI [0.58, 0.82]
Gemini 3.8 Flash	short	0.70, 95% CI [0.53, 0.87]	0.79, 95% CI [0.66, 0.91]	0.78, 95% CI [0.66, 0.88]
Grok 4.6	Long	0.77, 95% CI [0.61, 0.92]	0.83, 95% CI [0.71, 0.95]	0.69, 95% CI [0.56, 0.80]
Grok 4.6	short	0.77, 95% CI [0.61, 0.92]	0.84, 95% CI [0.73, 0.95]	0.76, 95% CI [0.64, 0.86]
Qwen 3.8 Max	Long	0.63, 95% CI [0.46, 0.81]	0.77, 95% CI [0.66, 0.89]	0.83, 95% CI [0.72, 0.92]
Qwen 3.8 Max	short	0.77, 95% CI [0.61, 0.92]	0.84, 95% CI [0.72, 0.95]	0.83, 95% CI [0.71, 0.93]
DeepSeek-R1	Long	0.23, 95% CI [0.08, 0.39]	0.27, 95% CI [0.11, 0.42]	0.21, 95% CI [0.07, 0.34]
DeepSeek-R1	short	0.13, 95% CI [0.01, 0.26]	0.21, 95% CI [0.08, 0.34]	0.18, 95% CI [0.06, 0.31]
Llama 3.2	Long	0.57, 95% CI [0.39, 0.75]	0.64, 95% CI [0.49, 0.80]	0.52, 95% CI [0.36, 0.66]
Llama 3.2	short	0.70, 95% CI [0.53, 0.87]	0.77, 95% CI [0.63, 0.90]	0.65, 95% CI [0.51, 0.77]
MedGemma 1.5	Long	0.73, 95% CI [0.57, 0.89]	0.81, 95% CI [0.69, 0.93]	0.62, 95% CI [0.49, 0.75]
MedGemma 1.5	short	0.60, 95% CI [0.42, 0.78]	0.71, 95% CI [0.57, 0.85]	0.60, 95% CI [0.46, 0.73]
Granite 4.1	Long	0.60, 95% CI [0.42, 0.78]	0.68, 95% CI [0.53, 0.83]	0.52, 95% CI [0.37, 0.66]
Granite 4.1	short	0.63, 95% CI [0.46, 0.81]	0.71, 95% CI [0.56, 0.86]	0.59, 95% CI [0.44, 0.73]
Qwen3	Long	0.43, 95% CI [0.25, 0.61]	0.48, 95% CI [0.31, 0.65]	0.37, 95% CI [0.22, 0.52]
Qwen3	short	0.47, 95% CI [0.29, 0.65]	0.59, 95% CI [0.44, 0.75]	0.48, 95% CI [0.34, 0.63]

Table A3. Performance of LLMs on the 30 clinical cases from the TREC CDS 2015 dataset, measured using Top-1 accuracy and MRR with corresponding 95% confidence intervals (CIs). DDC is not reported because reference differential diagnoses were unavailable for this dataset.

LLM Model	Case Type	Top-1	MRR
Claude Fable 5.1	Long	0.90, 95% CI [0.79, 1.00]	0.94, 95% CI [0.86, 1.00]
Claude Fable 5.1	short	0.83, 95% CI [0.70, 0.97]	0.88, 95% CI [0.78, 0.98]
GPT-6 Astra	Long	0.90, 95% CI [0.79, 1.00]	0.93, 95% CI [0.86, 1.00]
GPT-6 Astra	short	0.80, 95% CI [0.65, 0.95]	0.86, 95% CI [0.76, 0.96]
DeepSeek V4 Pro	Long	0.87, 95% CI [0.74, 0.99]	0.90, 95% CI [0.82, 1.00]
DeepSeek V4 Pro	short	0.80, 95% CI [0.65, 0.95]	0.85, 95% CI [0.74, 0.96]
Gemini 3.8 Flash	Long	0.87, 95% CI [0.74, 0.99]	0.91, 95% CI [0.82, 1.00]
Gemini 3.8 Flash	short	0.80, 95% CI [0.65, 0.95]	0.87, 95% CI [0.76, 0.97]
Grok 4.6	Long	0.87, 95% CI [0.74, 0.99]	0.91, 95% CI [0.86, 1.00]
Grok 4.6	short	0.80, 95% CI [0.65, 0.95]	0.85, 95% CI [0.76, 0.97]
Qwen 3.8 Max	Long	0.90, 95% CI [0.79, 1.00]	0.93, 95% CI [0.84, 1.00]
Qwen 3.8 Max	short	0.83, 95% CI [0.70, 0.97]	0.89, 95% CI [0.80, 0.98]
DeepSeek-R1	Long	0.10, 95% CI [0.00, 0.21]	0.13, 95% CI [0.02, 0.24]

Table A3. *Cont.*

LLM Model	Case Type	Top-1	MRR
DeepSeek-R1	short	0.23, 95% CI [0.08, 0.39]	0.26, 95% CI [0.11, 0.41]
Llama 3.2	Long	0.43, 95% CI [0.25, 0.61]	0.52, 95% CI [0.37, 0.68]
Llama 3.2	short	0.67, 95% CI [0.50, 0.84]	0.71, 95% CI [0.56, 0.87]
MedGemma 1.5	Long	0.47, 95% CI [0.29, 0.65]	0.59, 95% CI [0.45, 0.74]
MedGemma 1.5	short	0.60, 95% CI [0.42, 0.78]	0.68, 95% CI [0.53, 0.83]
Granite 4.1	Long	0.40, 95% CI [0.22, 0.58]	0.53, 95% CI [0.38, 0.68]
Granite 4.1	short	0.53, 95% CI [0.35, 0.71]	0.64, 95% CI [0.49, 0.79]
Qwen3	Long	0.27, 95% CI [0.11, 0.43]	0.33, 95% CI [0.17, 0.48]
Qwen3	short	0.40, 95% CI [0.22, 0.58]	0.44, 95% CI [0.27, 0.61]

Appendix A.3

As described in Section 3.1, the clinical cases used in this work are drawn from the TREC CDS track and serve as queries representing clinicians' information needs when searching for relevant medical literature. Based on the underlying information needs, the cases are classified into three types: diagnosis, test, and treatment. In this section, we report the effectiveness of the evaluated LLMs across these three case types. Tables A4 and A5 present the results for the 2014 and 2015 datasets, respectively.

Table A4. Performance of LLMs on the 30 clinical cases from the TREC CDS 2014 dataset, measured using Top-*k* accuracy, MRR, and DDC. The results are categorized by clinical case type.

LLM Model	Case Type	Top-1	Top-3	Top-5	Top-Max	MRR	DCC
Claude Fable 5.1	Diagnosis	0.80	0.90	1.00	1.00	0.87	0.76
Claude Fable 5.1	Test	0.50	0.80	0.90	1.00	0.68	0.68
Claude Fable 5.1	Treatment	0.60	1.00	1.00	1.00	0.77	1.00 [1]
GPT-6 Astra	Diagnosis	0.70	0.90	0.90	1.00	0.82	0.74
GPT-6 Astra	Test	0.80	0.80	0.80	0.80	0.80	0.49
GPT-6 Astra	Treatment	0.60	0.90	1.00	1.00	0.76	1.00
DeepSeek V4 Pro	Diagnosis	0.90	0.90	1.00	1.00	0.92	0.76
DeepSeek V4 Pro	Test	0.50	0.90	0.90	0.90	0.68	0.48
DeepSeek V4 Pro	Treatment	0.80	1.00	1.00	1.00	0.88	1.00
Gemini 3.8 Flash	Diagnosis	0.80	0.90	1.00	1.00	0.87	0.62
Gemini 3.8 Flash	Test	0.50	0.80	0.80	0.80	0.63	0.48
Gemini 3.8 Flash	Treatment	0.60	0.90	1.00	1.00	0.78	1.00
Grok 4.6	Diagnosis	0.90	0.90	0.90	0.90	0.90	0.60
Grok 4.6	Test	0.60	0.70	0.90	0.90	0.70	0.44
Grok 4.6	Treatment	0.80	1.00	1.00	1.00	0.88	1.00
Qwen 3.8 Max	Diagnosis	0.70	0.90	0.90	1.00	0.80	0.78
Qwen 3.8 Max	Test	0.50	0.90	0.90	0.90	0.68	0.67

Table A4. *Cont.*

LLM Model	Case Type	Top-1	Top-3	Top-5	Top-Max	MRR	DCC
Qwen 3.8 Max	Treatment	0.70	1.00	1.00	1.00	0.83	1.00
DeepSeek-R1	Diagnosis	0.20	0.30	0.30	0.30	0.25	0.10
DeepSeek-R1	Test	0.10	0.10	0.10	0.10	0.10	0.02
DeepSeek-R1	Treatment	0.40	0.50	0.50	0.50	0.45	0.50
Llama 3.2	Diagnosis	0.30	0.50	0.70	0.70	0.45	0.34
Llama 3.2	Test	0.60	0.70	0.70	0.70	0.63	0.30
Llama 3.2	Treatment	0.80	0.90	0.90	0.90	0.85	0.90
MedGemma 1.5	Diagnosis	0.80	0.90	0.90	0.90	0.85	0.51
MedGemma 1.5	Test	0.60	0.90	0.90	0.90	0.73	0.45
MedGemma 1.5	Treatment	0.80	0.90	0.90	0.90	0.85	0.90
Granite 4.1	Diagnosis	0.60	0.80	0.90	0.90	0.72	0.37
Granite 4.1	Test	0.50	0.60	0.60	0.60	0.53	0.27
Granite 4.1	Treatment	0.70	0.90	0.90	0.90	0.78	0.90
Qwen3	Diagnosis	0.50	0.60	0.60	0.60	0.55	0.16
Qwen3	Test	0.30	0.30	0.40	0.40	0.33	0.26
Qwen3	Treatment	0.50	0.60	0.70	0.70	0.58	0.70

[1] The clinical cases under "Treatment" category have a single diagnosis on their physician-provided differential diagnoses.

Table A5. Performance of LLMs on the 30 clinical cases from the TREC CDS 2015 dataset, measured using Top-k accuracy and MRR. The results are categorized by clinical case type.

LLM Model	Case Type	Top-1	Top-3	Top-5	Top-Max	MRR
Claude Fable 5.1	Diagnosis	0.80	0.90	0.90	1.00	0.86
Claude Fable 5.1	Test	0.90	1.00	1.00	1.00	0.95
Claude Fable 5.1	Treatment	1.00	1.00	1.00	1.00	1.00
GPT-6 Astra	Diagnosis	0.80	0.90	0.90	0.90	0.85
GPT-6 Astra	Test	0.90	1.00	1.00	1.00	0.95
GPT-6 Astra	Treatment	1.00	1.00	1.00	1.00	1.00
DeepSeek V4 Pro	Diagnosis	0.70	0.80	0.90	0.90	0.775
DeepSeek V4 Pro	Test	0.90	1.00	1.00	1.00	0.95
DeepSeek V4 Pro	Treatment	1.00	1.00	1.00	1.00	1.00
Gemini 3.8 Flash	Diagnosis	0.70	0.90	0.90	0.90	0.78
Gemini 3.8 Flash	Test	0.90	1.00	1.00	1.00	0.95
Gemini 3.8 Flash	Treatment	1.00	1.00	1.00	1.00	1.00
Grok 4.6	Diagnosis	0.70	0.90	0.90	0.90	0.80
Grok 4.6	Test	0.90	0.90	1.00	1.00	0.92
Grok 4.6	Treatment	1.00	1.00	1.00	1.00	1.00
Qwen 3.8 Max	Diagnosis	0.80	0.90	0.90	0.90	0.85

Table A5. *Cont.*

LLM Model	Case Type	Top-1	Top-3	Top-5	Top-Max	MRR
Qwen 3.8 Max	Test	0.90	0.90	1.00	1.00	0.93
Qwen 3.8 Max	Treatment	1.00	1.00	1.00	1.00	1.00
DeepSeek-R1	Diagnosis	0.00	0.20	0.20	0.20	0.08
DeepSeek-R1	Test	0.00	0.00	0.00	0.00	0.00
DeepSeek-R1	Treatment	0.30	0.30	0.30	0.30	0.30
Llama 3.2	Diagnosis	0.30	0.40	0.40	0.50	0.35
Llama 3.2	Test	0.40	0.40	0.60	0.70	0.46
Llama 3.2	Treatment	0.60	0.90	1.00	1.00	0.76
MedGemma 1.5	Diagnosis	0.40	0.60	0.70	0.70	0.49
MedGemma 1.5	Test	0.40	0.70	0.70	0.90	0.57
MedGemma 1.5	Treatment	0.60	0.80	0.90	1.00	0.72
Granite 4.1	Diagnosis	0.40	0.50	0.60	0.60	0.47
Granite 4.1	Test	0.20	0.60	0.70	0.70	0.41
Granite 4.1	Treatment	0.60	0.90	0.90	0.90	0.72
Qwen3	Diagnosis	0.30	0.40	0.50	0.50	0.38
Qwen3	Test	0.20	0.30	0.40	0.40	0.26
Qwen3	Treatment	0.30	0.40	0.40	0.40	0.35

References

1. Ledley, R.S.; Lusted, L.B. Reasoning Foundations of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason. *Science* **1959**, *130*, 9–21.
2. O'Connor, G.T.; Sox, H.C., Jr. Bayesian Reasoning in Medicine: The Contributions of Lee B. Lusted, MD. *Med. Decis. Mak.* **1991**, *11*, 107–111.
3. Shortliffe, E.H.; Axline, S.G.; Buchanan, B.G.; Merigan, T.C.; Cohen, S.N. An Artificial Intelligence Program to Advise Physicians Regarding Antimicrobial Therapy. *Comput. Biomed. Res.* **1973**, *6*, 544–560. [CrossRef]
4. Yu, V.L.; Buchanan, B.G.; Shortliffe, E.H.; Wraith, S.M.; Davis, R.; Scott, A.C.; Cohen, S.N. Evaluating the Performance of a Computer-Based Consultant. *Comput. Programs Biomed.* **1979**, *9*, 95–102. [CrossRef]
5. Sheikhtaheri, A.; Sadoughi, F.; Hashemi Dehaghi, Z. Developing and Using Expert Systems and Neural Networks in Medicine: A Review on Benefits and Challenges. *J. Med. Syst.* **2014**, *38*, 110. [CrossRef]
6. Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1589–1604. [CrossRef]
7. Si, Y.; Du, J.; Li, Z.; Jiang, X.; Miller, T.; Wang, F.; Zheng, W.J.; Roberts, K. Deep Representation Learning of Patient Data from Electronic Health Records (EHR): A Systematic Review. *J. Biomed. Inform.* **2021**, *115*, 103671. [CrossRef]
8. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large Language Models in Medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef]
9. Clusmann, J.; Kolbinger, F.R.; Muti, H.S.; Carrero, Z.I.; Eckardt, J.-N.; Laleh, N.G.; Löffler, C.M.L.; Schwarzkopf, S.-C.; Unger, M.; Veldhuizen, G.P.; et al. The Future Landscape of Large Language Models in Medicine. *Commun. Med.* **2023**, *3*, 141. [CrossRef]
10. 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 Med. Inform.* **2023**, *11*, e5471. [CrossRef]
11. Ríos-Hoyo, A.; Shan, N.L.; Li, A.; Pearson, A.T.; Pusztai, L.; Howard, F.M. Evaluation of Large Language Models as a Diagnostic Aid for Complex Medical Cases. *Front. Med.* **2024**, *11*, 1380148. [CrossRef]
12. Mizuta, K.; Hirosawa, T.; Harada, Y.; Shimizu, T. Can ChatGPT-4 Evaluate Whether a Differential Diagnosis List Contains the Correct Diagnosis as Accurately as a Physician? *Diagnosis* **2024**, *11*, 321–324. [CrossRef]

13. Simpson, M.S.; Voorhees, E.M.; Hersh, W. Overview of the TREC 2014 Clinical Decision Support Track. In Proceedings of the TREC 2014, Gaithersburg, MD, USA, 19 November 2014.
14. Roberts, K.; Simpson, M.S.; Voorhees, E.M.; Hersh, W. Overview of the TREC 2015 Clinical Decision Support Track. In Proceedings of the TREC 2015, Gaithersburg, MD, USA, 17 November 2015.
15. Miller, R.A.; Pople, H.E., Jr.; Myers, J.D. INTERNIST-I, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. *N. Engl. J. Med.* **1982**, *307*, 468–476. [CrossRef]
16. Horrocks, J.C.; McCann, A.P.; Staniland, J.R.; Leaper, D.J.; de Dombal, F.T. Computer-Aided Diagnosis: Description of an Adaptable System, and Operational Experience with 2,034 Cases. *Br. Med. J.* **1972**, *2*, 5–9. [CrossRef]
17. de Dombal, F.T.; Leaper, D.J.; Staniland, J.R.; McCann, A.P.; Horrocks, J.C. Computer-Aided Diagnosis of Acute Abdominal Pain. *Br. Med. J.* **1972**, *2*, 9–13. [CrossRef]
18. Palaniappan, S.; Awang, R. Intelligent Heart Disease Prediction System Using Data Mining Techniques. In Proceedings of the 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, Qatar, 31 March 2008.
19. Ferreira, D.; Oliveira, A.; Freitas, A. Applying Data Mining Techniques to Improve Diagnosis in Neonatal Jaundice. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, 143. [CrossRef]
20. Liu, Y.-X.; Liu, X.; Cen, C.; Li, X.; Liu, J.-M.; Ming, Z.-Y.; Yu, S.-F.; Tang, X.-F.; Zhou, L.; Yu, J.; et al. Comparison and Development of Advanced Machine Learning Tools to Predict Nonalcoholic Fatty Liver Disease: An Extended Study. *Hepatobiliary Pancreat. Dis. Int.* **2021**, *20*, 409–415. [CrossRef]
21. Wagholikar, K.B.; MacLaughlin, K.L.; Henry, M.R.; Greenes, R.A.; Hankey, R.A.; Liu, H.; Chaudhry, R. Clinical Decision Support with Automated Text Processing for Cervical Cancer Screening. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 833–839. [CrossRef]
22. Schmickl, C.N.; Owens, W.B.; Keating, M.; Herasevich, V.; Gajic, O. Decision Support Tool for Early Differential Diagnosis of Acute Lung Injury and Cardiogenic Pulmonary Edema in Medical Critically Ill Patients. *Chest* **2011**, *140*, 1102–1111. [CrossRef]
23. Zhang, X.; Hu, B.; Ma, X.; Moore, P.; Chen, J. Ontology Driven Decision Support for the Diagnosis of Mild Cognitive Impairment. *Comput. Methods Programs Biomed.* **2014**, *113*, 781–791. [CrossRef]
24. OpenAI GPT-6 Astra: A New Generation of Intelligence. Available online: https://openai.com/index/gpt-6-astra/ (accessed on 1 October 2026).
25. Google Gemini 3.8 Flash. Available online: https://ai.google.dev/gemini-api/docs/models/gemini-3.8-flash (accessed on 1 October 2026).
26. Anthropic Claude Fable 5.1—Claude Platform Docs. Available online: https://platform.claude.com/docs/en/models/fable-5-1/overview (accessed on 1 October 2026).
27. Mykhalko, Y.; Kish, P.; Rubtsova, Y.; Kutsyn, O. From Text to Diagnose: ChatGPT's Efficacy in Medical Decision-Making. *Wiad. Lek.* **2023**, *76*, 2345–2350. [CrossRef]
28. Eriksen, A.V.; Möller, S.; Ryg, J. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI* **2024**, *1*, AIp2300031. [CrossRef]
29. Ito, N.; Kadomatsu, S.; Fujisawa, M.; Fukaguchi, K.; Ishizawa, R.; Kanda, N.; Kasugai, D.; Nakajima, M.; Goto, T.; Tsugawa, Y. The Accuracy and Potential Racial and Ethnic Biases of GPT-4 in the Diagnosis and Triage of Health Conditions: Evaluation Study. *JMIR Med. Educ.* **2023**, *9*, e47532. [CrossRef]
30. Kim, J.; Leonte, K.G.; Chen, M.L.; Torous, J.B.; Linos, E.; Pinto, A.; Rodriguez, C.I. Large Language Models Outperform Mental and Medical Health Care Professionals in Identifying Obsessive-Compulsive Disorder. *npj Digit. Med.* **2024**, *7*, 193. [CrossRef]
31. McDuff, D.; Schaekermann, M.; Tu, T.; Palepu, A.; Wang, A.; Garrison, J.; Singhal, K.; Sharma, Y.; Azizi, S.; Kulkarni, K.; et al. Towards Accurate Differential Diagnosis with Large Language Models. *Nature* **2025**, *642*, 451–457. [CrossRef]
32. Feldman, M.J.; Hoffer, E.P.; Conley, J.J.; Chang, J.; Chung, J.A.; Jernigan, M.C.; Lester, W.T.; Strasser, Z.H.; Chueh, H.C. Dedicated AI Expert System vs Generative AI with Large Language Model for Clinical Diagnoses. *JAMA Netw. Open* **2025**, *8*, e2512994. [CrossRef]
33. DeepSeek-V4 Preview: Entering the Era of Affordable Million-Token Context. Available online: https://www.deepseek.com/en/news/v4-preview/ (accessed on 1 October 2026).
34. SpaceXAI. Introducing Grok 4.6. Available online: https://x.ai/news/grok-4-6 (accessed on 1 October 2026).
35. Alibaba Cloud Qwen3.8-Max Model Info. Available online: https://www.alibabacloud.com/help/en/model-studio/qwen3-8-max (accessed on 1 October 2026).
36. DeepSeek-R1 Release. Available online: https://www.deepseek.com/en/news/deepseek-r1/ (accessed on 1 October 2026).
37. Meta Open-Source AI Models for Any Application|Llama 3. Available online: https://dev.meta.ai/llama/models/llama-3 (accessed on 1 October 2026).
38. Qwen-3: Alibaba Cloud's Next-Gen Open Source LLM. Available online: https://qwen-3.com/ar (accessed on 1 October 2026).
39. IBM Granite 4.1. Available online: https://www.ibm.com/granite/docs/models/granite4-1 (accessed on 1 October 2026).
40. Google DeepMind MedGemma. Available online: https://deepmind.google/models/gemma/medgemma/ (accessed on 1 October 2026).
41. Ollama Model Library. Available online: https://ollama.com/search (accessed on 1 October 2026).

42. Savage, T.; Wang, J.; Gallo, R.; Boukil, A.; Patel, V.; Safavi-Naini, S.A.A.; Soroush, A.; Chen, J.H. Large Language Model Uncertainty Proxies: Discrimination and Calibration for Medical Diagnosis and Treatment. *J. Am. Med. Inform. Assoc.* **2025**, *32*, 139–149. [CrossRef]
43. Voorhees, E.M. The TREC-8 Question Answering Track Report. In Proceedings of the TREC-8, Gaithersburg, MD, USA, 17 November 1999.
44. Radev, D.R.; Qi, H.; Wu, H.; Fan, W. Evaluating Web-Based Question Answering Systems. In Proceedings of the LREC'02, Las Palmas, Spain, 27 May–2 June 2002.
45. Carbonell, J.; Goldstein, J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In Proceedings of the ACM SIGIR '98, Melbourne, Australia, 24 August 1998.
46. Zhai, C.; Cohen, W.W.; Lafferty, J. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In Proceedings of the ACM SIGIR '2003, Toronto, ON, Canada, 28 July 2003.
47. Clarke, C.L.A.; Kolla, M.; Cormack, G.V.; Vechtomova, O.; Ashkan, A.; Büttcher, S.; MacKinnon, I. Novelty and Diversity in Information Retrieval Evaluation. In Proceedings of the ACM SIGIR '2008, Singapore, 20 July 2008.
48. Chapelle, O.; Ji, S.; Liao, C.; Velipasaoglu, E.; Lai, L.; Wu, S.-L. Intent-Based Diversification of Web Search Results: Metrics and Algorithms. *Inf. Retr.* **2011**, *14*, 572–592. [CrossRef]
49. Sandmann, S.; Hegselmann, S.; Fujarski, M.; Bickmann, L.; Wild, B.; Eils, R.; Varghese, J. Benchmark Evaluation of DeepSeek Large Language Models in Clinical Decision-Making. *Nat. Med.* **2025**, *31*, 2546–2549. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.