**INTERNSHIP TRAINING**

**At**

**Jivi Health**

**Gurgaon**

**Project Title**

# Evaluating the Diagnostic Accuracy of Chat GPT

By

**Ms. Nida Shams**

**PG/22/061**

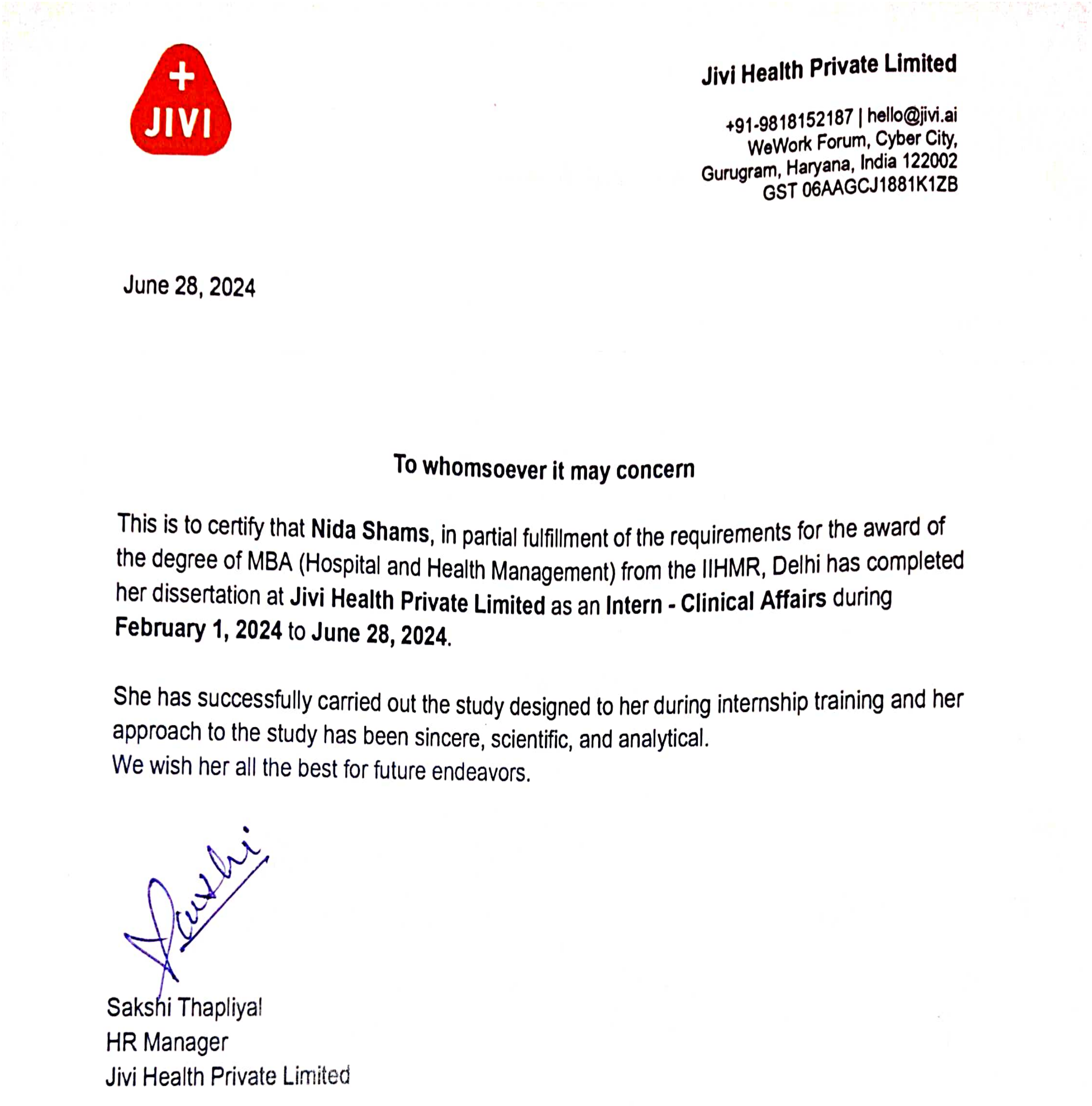
**UNDER THE GUIDANCE OF**

**Dr. Preetha G.S**

**PGDM (Hospital and Health Management)**

**2022-2024**

**International Institute of Health Management Research New Delhi**



JIVI

**Jivi Health Private Limited**

+91-9818152187 | hello@jivi.ai
WeWork Forum, Cyber City,
Gurugram, Haryana, India 122002
GST 06AAGCJ1881K1ZB

June 28, 2024

**To whomsoever it may concern**

This is to certify that **Nida Shams**, in partial fulfillment of the requirements for the award of the degree of MBA (Hospital and Health Management) from the IIHMR, Delhi has completed her dissertation at **Jivi Health Private Limited** as an **Intern - Clinical Affairs** during **February 1, 2024** to **June 28, 2024**.

She has successfully carried out the study designed to her during internship training and her approach to the study has been sincere, scientific, and analytical.
We wish her all the best for future endeavors.

Sakshi

Sakshi Thapliyal
HR Manager
Jivi Health Private Limited

## TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Ms. Nida Shams** student of **PGDM (Hospital & Health Management) from the International Institute of Health Management Research**, New Delhi has undergone internship training at **Jivi Health Pvt Ltd** from **Feb to June 2024.** The Candidate has successfully carried out the study designated to her during the internship training and her approach to the study has been sincere, scientific, and analytical. The Internship is in fulfillment of the course requirements.

I wish her all success in all his/her future endeavors.

Dr. Sumesh Kumar
Associate Dean, Academic, and Student Affairs
IIHMR, New Delhi

Dr. Preetha G.S
Professor
IIHMR, New Delhi

## Certificate of Approval

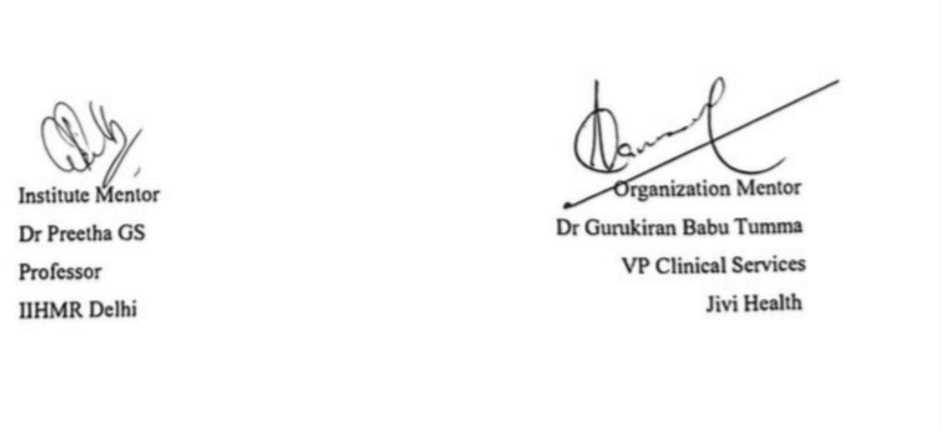
The following dissertation titled "**Evaluating the Diagnostic Accuracy of Chat GPT**" at "**IIHMR Delhi**" is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of PGDM (Hospital & Health Management) for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but approve the dissertation only for the purpose it is submitted.

**Dissertation Examination Committee for evaluation of dissertation.**

| Name | Signature |
|---|---|
| Dr. Shiv Shanker | |
| Dr. Anandhi Ramachanda | |
| Dr. Ekta | |

## Certificate from Dissertation Advisory Committee

This is to certify that **Ms. Nida Shams**, as graduate student of the PGDM (Hospital & Health Management) has worked under our guidance and supervision. She is submitting this dissertation titled "**Evaluating the Diagnostic Accuracy of Chat GPT"** in partial fulfilment of the requirements for the award of the PGDM (Hospital & Health Management). This Dissertation has the requisite standard and to the best of our knowledge, no part of it has been reproduced from any other dissertation, monograph, report, or book.



Institute Mentor
Dr Preetha GS
Professor
IIHMR Delhi

Organization Mentor
Dr Gurukiran Babu Tumma
VP Clinical Services
Jivi Health

## INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH, NEW DELHI

### CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled **Evaluating the Diagnostic Accuracy of Chat GPT** and submitted by **Ms. Nida Shams** Enrolment No. **PG/22/061** under the supervision of **Dr. Preetha G.S** for the award of PGDM (Hospital & Health Management) of the Institute carried out during the period from **Feb 2024 to June 2024** embodies my original work and has not formed the basis for the award of any degree, diploma associate ship, fellowship, titles in this or any other Institute or other similar institution of higher learning.

**Signature**

**FEEDBACK FORM**
**(Organization Supervisor)**

**Name of the Student:** Nida Shams.

**Name of the organisation:** Jivi Health Pvt. Ltd.

**Area of Dissertation:** Health information technology

**Attendance:** Good attendence.

**Objectives met:** All the objectives met successfully.

**Deliverables:** contributed in making processes, delevering. high-quality work on time.

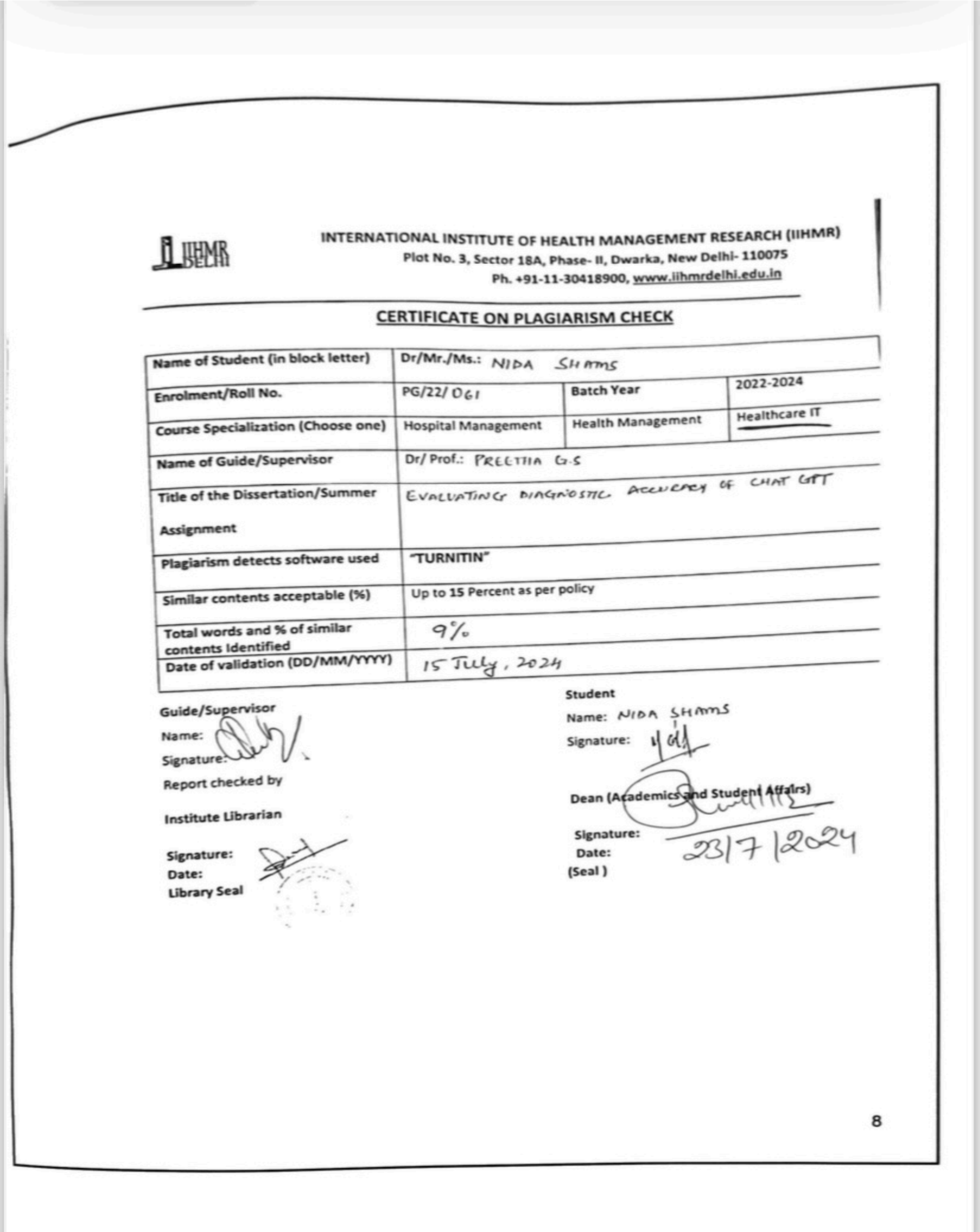
**Strengths:** → she is willing to explore new Ideas.
→ good work ethics.

**Suggestions for Improvement:**
→ can improve the ability to prioritize tasks for better efficiency.

**Signature of the Officer-in-Charge**

**(Internship)**

**Date:** 1 July 2024
**Place:**



INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH (IIHMR)
Plot No. 3, Sector 18A, Phase- II, Dwarka, New Delhi- 110075
Ph. +91-11-30418900, www.iihmrdelhi.edu.in

## CERTIFICATE ON PLAGIARISM CHECK

| | | | |
|---|---|---|---|
| Name of Student (in block letter) | Dr/Mr./Ms.: NIDA SHAMS | | |
| Enrolment/Roll No. | PG/22/061 | Batch Year | 2022-2024 |
| Course Specialization (Choose one) | Hospital Management | Health Management | [✓ Healthcare IT] |
| Name of Guide/Supervisor | Dr/ Prof.: PREETHA G.S | | |
| Title of the Dissertation/Summer Assignment | EVALUATING DIAGNOSTIC ACCURACY OF CHAT GPT | | |
| Plagiarism detects software used | "TURNITIN" | | |
| Similar contents acceptable (%) | Up to 15 Percent as per policy | | |
| Total words and % of similar contents Identified | 9% | | |
| Date of validation (DD/MM/YYYY) | 15 July, 2024 | | |

Guide/Supervisor
Name:
Signature:

Student
Name: NIDA SHAMS
Signature:

Report checked by

Institute Librarian

Signature:
Date:
Library Seal

Dean (Academics and Student Affairs)

Signature:
Date: 23/7/2024
(Seal )

## ACKNOWLEDGEMENT

While we look back on the past three months, which have been incredibly busy and eventful, I want to express our gratitude to everyone who has provided us with invaluable counsel and direction. This report would not have been possible without the support of those named below.

Firstly, we would like to thank IIHMR DELHI for giving us the chance to team up with Jivi.ai. I am extremely thankful to **Dr. Gurukiran Tumma** for believing in me and giving me the chance to work at Jivi.ai, as well as to my mentor, **Dr Preetha G.S**, for all her hard work and diligent insights through these two years of IIHMR journey.

I want to sincerely thank everyone for their support, with special thanks to **Mr. Ankur Jain**, the CEO, for his invaluable insights and counsel. I want to thank our technical team members and lead, **Ms. Ritu Saini**, for providing me with the amazing chance to learn technical skills and be involved in the project itself. Both personally and professionally, this has been an amazing experience.

The Clinical Team has worked hard to implement this new knowledge and information in the most efficient way possible and to enhance it even more to meet the career goals that have been set for us. Additionally, I want to express my gratitude to my clinical teammates for joining me on this wonderful trip.

## Abbreviation

| Sr. No | Abbreviation | Full form |
|---|---|---|
| 1 | LLM | Large language model |
| 2 | GPT | Generative Pre trained transform |
| 3 | OSCE | Objective Structured Clinical Exam |
| 4 | AI | Artificial Intelligence |
| 5 | MRR | Mean Reciprocal Rank |

# Table of Contents

## About the Organization

Jivi.ai is a healthcare startup company founded by Mr. Ankur Jain, the former Chief Product Officer of BharatPe. The main objective of the organization is to revolutionize primary healthcare through the utilization of artificial intelligence. Jivi AI uses massive language models, machine learning, generative AI, and digital health technologies to enhance healthcare accessibility and efficacy.

Since it was established in December 2023, the company has assembled an interdisciplinary team of experts and scholars from esteemed universities including Stanford, MIT, Harvard, and Yale. With intentions to expand its operations to the US, Jivi AI has already worked with more than 100 doctors, physicians, and hospitals, mostly in India.

The ultimate objective of Jivi AI is to enhance global healthcare outcomes for billions of people. To support its growth and development, the firm has acquired its first initial funding and is currently negotiating additional finance rounds.

Jivi's Large Language Model (LLM), Jivi MedX, achieves an average score of 91.65 across the nine benchmark categories on the leaderboard, surpassing well-known LLMs like OpenAI's GPT-4 and Google's Med-PaLM 2. Leading AI platform Hugging Face hosts the leaderboard, which rates LLMs with a focus on medicine based on how well they respond to questions about medicine from tests and studies.

## Abstract:

The adoption of artificial intelligence (AI) in healthcare promises to revolutionize clinical practice by improving decision-making, streamlining operations, and enhancing patient outcomes. Conversational AI models like ChatGPT have attracted significant attention due to their ability to understand and generate human-like text responses. This study aimed to assess ChatGPT's diagnostic accuracy in simulated clinical scenarios using Objective Structured Clinical Examination (OSCE) cases as a benchmark, with a secondary focus on evaluating its performance across various medical departments.

A comparative observational study design was used, involving 170 OSCE cases that represented a range of medical conditions. ChatGPT's diagnostic abilities were evaluated using three primary metrics: Accuracy, Top-3 Accuracy, and Mean Reciprocal Rank (MRR). Data analysis was performed with Google Sheets. The findings revealed a high overall accuracy of 85.29%, with the correct diagnosis frequently appearing within the top three suggestions (Top-3 Accuracy of 82.94%). The model also achieved a high MRR of 0.89, indicating that the correct diagnosis was often ranked near the top.

Performance varied among specialties, with Dermatology, Infectious Disease, Respiratory, and Urology showing high accuracy and reliability, while Cardiology, Emergency, Endocrinology, Gastroenterology, Geriatrics, and Orthopedics demonstrated moderate performance. These results highlight ChatGPT's potential as a supportive tool in clinical settings, offering reliable diagnostic suggestions while pointing out the need for targeted improvements in specific specialties.

The study adhered to ethical guidelines, ensuring patient confidentiality and addressing potential biases. This research contributes to the ongoing conversation about AI in healthcare, providing insights that can guide the development of accurate, reliable, and ethically sound AI-driven solutions.

## Introduction

In recent years, the integration of artificial intelligence (AI) into various domains has shown promising potential to augment and streamline processes, including those within the healthcare sector. One such advancement is the development of conversational AI models, exemplified by Chat GPT, which exhibit remarkable capabilities in understanding and generating human-like text responses. Chat GPT, developed by Open AI, represents a notable milestone in the field of natural language processing, demonstrating the ability to engage in coherent and contextually appropriate conversations across diverse topics and domains. These AI models, trained on vast amounts of text data, leverage deep learning algorithms to process natural language input and generate contextually relevant outputs.

Large Language Models (LLMs), such as Chat GPT and Google Bard, have recently emerged as transformative agents in healthcare, reshaping its future landscape. These sophisticated AI-driven systems, trained on extensive datasets, excel in various natural language processing tasks, including content creation, language translation, and code generation. Their integration into healthcare signifies not just a technological leap but a fundamental shift towards more efficient, patient-centric care systems.[i]

In healthcare, LLMs like Chat GPT are revolutionizing service delivery by enhancing clinical decision support, analyzing various data types, and improving patient communication and education. These models are particularly effective in drug discovery, identifying adverse drug events, interpreting medical images for cancer detection, and functioning as virtual medical assistants. Their capability to generate human-like text responses can transform areas such as adverse event detection, clinical documentation, and medical research. This offers significant advancements in fields like oncology and pharmaceuticals by facilitating more effective treatment strategies and predicting drug interactions.

However, as AI models like Chat GPT become increasingly sophisticated, it is imperative to rigorously evaluate their accuracy and efficacy, particularly in contexts where precision and reliability are paramount, such as clinical settings. In healthcare, the consequences of inaccurate or misleading information can be profound, potentially impacting patient outcomes and safety. Hence, the evaluation of AI models in healthcare scenarios is of utmost importance to ensure their suitability for supporting clinical decision-making and enhancing patient care.

One method for evaluating clinical competence and decision-making skills is the Objective Structured Clinical Examination (OSCE). OSCE is a widely adopted assessment tool in healthcare education and training, designed to simulate real-life patient encounters through standardized scenarios and structured evaluation criteria. By presenting candidates with diverse clinical cases and assessing their performance based on predefined criteria, OSCE provides a robust framework for measuring clinical proficiency and identifying areas for improvement. Given its emphasis on standardized assessment and objective evaluation, OSCE serves as an ideal benchmark for evaluating the accuracy and efficacy of AI models like Chat GPT in healthcare scenarios.

## Rationale:

The integration of artificial intelligence (AI) into healthcare holds the promise of transforming clinical practice by enhancing decision-making, streamlining processes, and improving patient outcomes. Among the various advancements in AI, conversational models such as ChatGPT have garnered significant attention due to their ability to understand and generate human-like text responses. These capabilities suggest potential applications in areas such as patient interaction, medical education, and decision support. However, given the critical nature of healthcare, where inaccuracies can lead to adverse outcomes, it is essential to rigorously evaluate the performance of such AI models.

Objective Structured Clinical Examination (OSCE) is an established method used to assess the clinical skills and decision-making abilities of healthcare professionals through standardized scenarios. OSCE's structured and objective nature makes it an ideal benchmark for evaluating the performance of AI models in clinical contexts. By comparing ChatGPT's performance against OSCE cases, it can obtain a detailed understanding of its accuracy, contextual appropriateness, and ability to generate clinically relevant responses.

This study is driven by the need to ensure that AI models like ChatGPT can reliably support clinical tasks without compromising patient safety. Evaluating ChatGPT using OSCE not only provides a rigorous assessment framework but also aligns with the high standards required in clinical practice. Moreover, the findings from this research can inform the development of more accurate and reliable AI systems, ultimately contributing to the safe and effective integration of AI into healthcare settings.

## Review of Literature

**1.** Patel BN, Rosenberg L, Willcox G, Sidhu P, Brown T, Gupta R, et al. *Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios.* JMIR Med Inform. 2023;11 doi: 10.2196/37203.

**Summary**: This research investigates the practical uses of ChatGPT in various clinical and research settings. It assesses the AI's performance in diagnostic support, patient triage, and generating research hypotheses. The authors find that ChatGPT can significantly enhance these processes by providing quick and accurate information.[ii] However, the necessity of human oversight to validate the AI's suggestions and maintain high standards of care is emphasized.

**Key Points**:

- Applications in diagnostic support and patient triage.
- Usefulness in generating research hypotheses.
- Importance of human oversight to ensure accuracy.

**2.** Das S, Devakumar D, Murali S, Duraiswamy K, Madhavan V. *Enhancing Diagnostic Accuracy with Large Language Models in Clinical Settings.* arXiv. [iii]

**Summary**: This paper focuses on the potential of large language models, like ChatGPT, to improve diagnostic accuracy in clinical environments. It presents case studies that showcase the model's ability to interpret complex medical data and provide accurate diagnoses. The authors also explore the integration of these models into existing healthcare systems and discuss challenges such as ensuring data security and managing biases.

**Key Points**:

- Enhancement of diagnostic accuracy through large language models.
- Case studies illustrating successful implementation.
- Challenges including data security and bias management.

3. Natarajan A, Smith L, Jones T, Lee K. *Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios.* ResearchGate.[iv]

**Summary**: This study assesses ChatGPT's feasibility in a variety of clinical and research contexts. It evaluates the model's ability to handle tasks such as reviewing medical literature, analyzing patient symptoms, and supporting clinical decision-making. The results suggest that ChatGPT can improve efficiency and accuracy in these tasks. However, limitations related to the model's understanding of nuanced medical information and the need for domain-specific training are also highlighted.

**Key Points**:

- Evaluation of ChatGPT's role in medical literature review and symptom analysis.
- Potential to support clinical decision-making.
- Limitations in understanding nuanced medical information.

## Objective

**Primary Objective:**

To evaluate the diagnostic accuracy of ChatGPT in simulated clinical scenarios using Objective Structured Clinical Examination (OSCE) cases as a benchmark.

**Secondary Objective:**

To compare the performance of ChatGPT across different medical departments (e.g., cardiology, dermatology, pediatrics, etc.) in terms of diagnostic accuracy and contextual appropriateness.

## Methodology

**Study Design:**

This study employs a comparative observational study design. It involves comparing the performance of Chat GPT against the established standard provided by Objective Structured Clinical Examination (OSCE) cases without manipulating any variables. The study assesses the diagnostic accuracy of ChatGPT and compares its performance across different medical departments.

**Study Duration**:

The study is conducted over a period of three months.

**Sample Size:**

The sample size for this study consists of 170 clinical cases selected from OSCE repositories. These cases cover a diverse range of medical conditions and presenting symptoms across different departments.

### Data Analysis

Data analysis was conducted using Google Sheets. This study aimed to evaluate the performance of an Artificial Intelligence model using three key metrics: Accuracy, Top-3 Accuracy, and Mean Reciprocal Rank (MRR). The specific tools and methods employed included:

**Accuracy:**

This metric measures the percentage of correct predictions made by the model out of all predictions. An accuracy of 70% implies that 70 out of 100 predictions made by the model are correct.

Scoring Criteria: 0 if correct diagnosis is not present, 1 if correct diagnosis is present.

**Top N Accuracy:**

This metric indicates the proportion of times the correct answer is within the top three predictions. A 62% Top-3 Accuracy suggests that in 62 out of 100 instances, the correct answer appears in the top three predictions.

Scoring Criteria: 1 if correct diagnosis is present in Top N suggestion, 0 if correct diagnosis is not present in Top N suggestion (N can be 3 or 5 )

**Mean Reciprocal Rank (MRR):**

MRR is a measure used to evaluate the effectiveness of a model in ranking the correct answer higher. It is calculated as the average of the reciprocal ranks of the correct answers. An MRR of 0.89 means that, on average, the correct answer is found at approximately the 1.12nd position in the ranking list (1/0.89 ≈ 1.12).

Scoring Criteria:

Step 1: 1/ Rank of the correct diagnosis

Step 2: Take Average of all cases

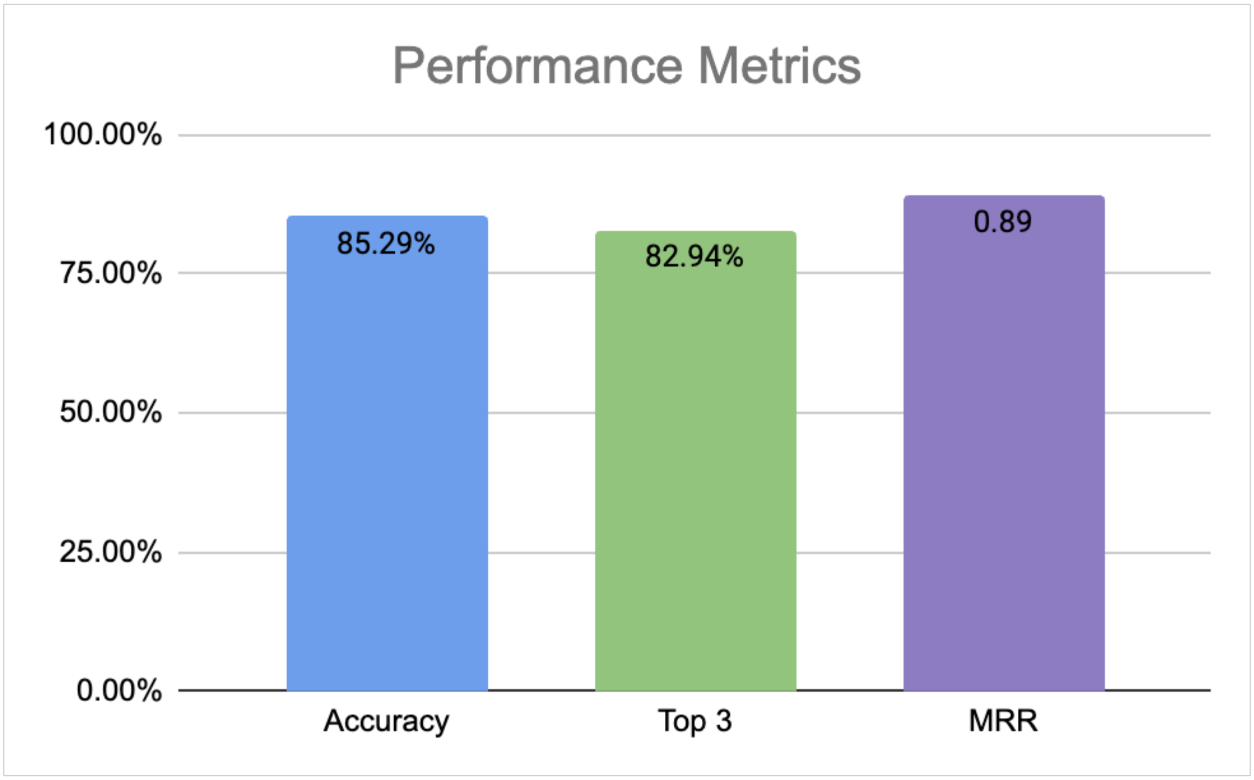
Interpretation:

- Perfect: =1.0
- Excellent: 0.75-1.0
- Good: 0.5-0.75
- Fair: 0.25-0.5
- Poor: <0.25

Ethical Considerations:

The study adhered to ethical guidelines governing research involving AI and healthcare data. Measures were taken to ensure the confidentiality of patient information and protect the privacy of all participants involved. Additionally, potential biases in data collection and analysis were addressed to ensure the validity and reliability of the findings.

## Result:



### Accuracy Analysis

High Accuracy (85.29%): The model's accuracy of 85.29% indicates a high level of performance in correctly identifying the most likely diagnosis. This metric signifies that out of every 100 predictions made by the model, approximately 85 are correct. Such a level of accuracy is substantial, demonstrating the model's robustness and reliability in its predictive capabilities.

Implications: High accuracy is particularly crucial for medical diagnostic systems where precise predictions are paramount. An 85.29% accuracy rate means that the majority of the diagnoses provided by the model are correct, which is essential for patient safety and effective treatment. In the critical field of healthcare, even a 15% error rate can have significant consequences, necessitating further refinement and improvement of the model.

Therefore, while an 85.29% accuracy rate is generally satisfactory and indicative of a strong performance, it is essential to consider the specific requirements and tolerances of the application in question. Continuous efforts to enhance the model’s accuracy are vital to ensure it meets the stringent demands of medical diagnostics, ultimately contributing to better patient outcomes and safer healthcare practices.

### Top-3 Accuracy Analysis

Strong Top-3 Accuracy (82.94%): The model's Top-3 Accuracy of 82.94% demonstrates that in approximately 82% of cases, the correct diagnosis is among the top three predictions. This metric underscores the model's robustness in generating highly relevant diagnostic options.
Comparison with Accuracy: The close values of overall accuracy (85.29%) and Top-3 Accuracy (82.94%) emphasize the model's effectiveness in providing pertinent predictions even when the top prediction is not the most likely diagnosis. This small drop-off indicates that the model consistently includes the correct answer within its top three suggestions.

Implications: These findings suggest that the model can be a valuable tool in clinical settings, offering reliable diagnostic support. The high Top-3 Accuracy implies that even if the most likely diagnosis is missed, the model still presents viable alternatives, enhancing decision-making processes for healthcare professionals. This feature could be particularly useful in complex cases where multiple potential diagnoses need to be considered, ultimately improving patient outcomes through comprehensive diagnostic coverage.

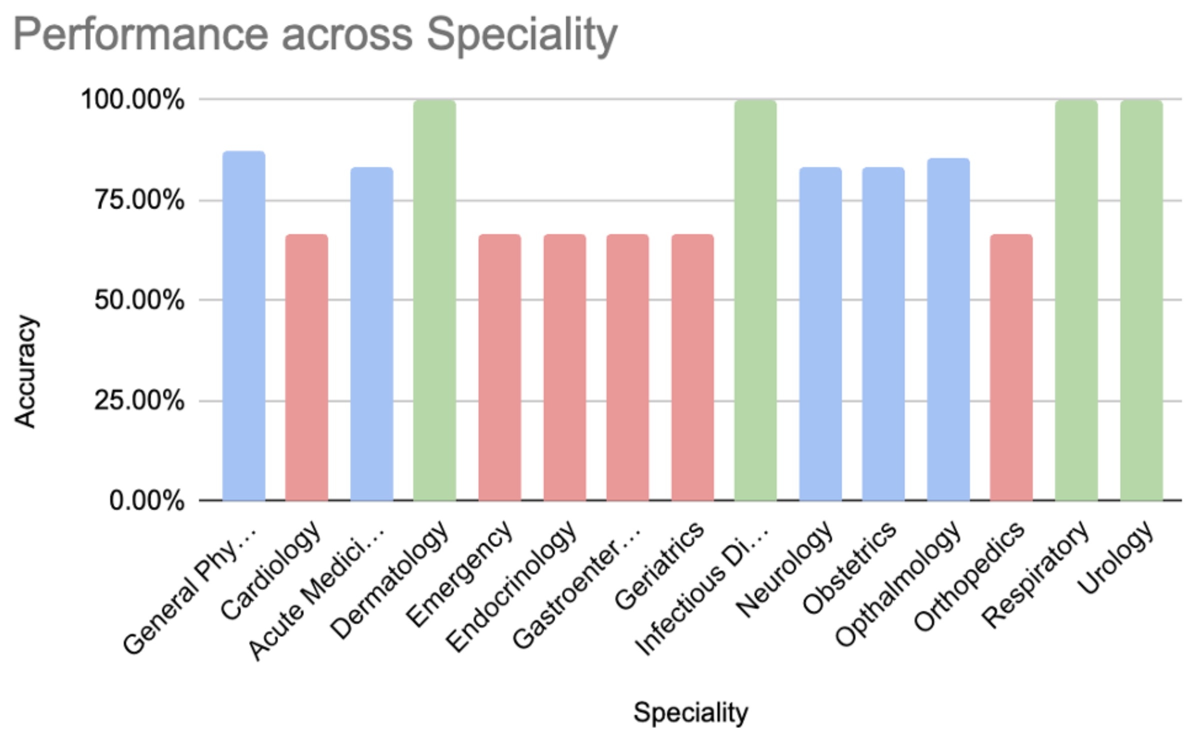
### Mean Reciprocal Rank Analysis

The Mean Reciprocal Rank (MRR) is a critical metric for evaluating the performance of models in ranking tasks. In this study, the model achieved an impressive MRR of 0.89, providing significant insights into its effectiveness and practical utility.
High MRR Interpretation: An MRR of 0.89 indicates that the correct answer is frequently ranked very high in the list of predictions. Specifically, this value suggests that the correct answer often appears in the first or second position. The reciprocal rank is calculated as the

inverse of the rank position at which the correct answer is found. Therefore, an MRR of 0.89 translates to the correct answer being located at an average rank of approximately 1.12.

Implications: This high MRR is indicative of the model's robust performance in ranking tasks, making it particularly valuable for applications that rely on the efficient and accurate ranking of results. In the context of healthcare, where quick and precise identification of the correct diagnosis is paramount, a high MRR ensures that clinicians receive the most relevant diagnostic options at the top of the list. This can significantly enhance decision-making processes, reduce diagnostic errors, and ultimately improve patient outcomes.

The model's ability to consistently rank the correct diagnosis highly underscores its potential as a reliable tool in clinical settings, supporting healthcare professionals in delivering timely and accurate medical care.



**High-Performing Specialties**

| **Specialities** | Accuracy | Top-3 Accuracy | MRR | Analysis |
|---|---|---|---|---|
| Dermatology | 100% | 100% | 1 | The model achieves flawless scores across all metrics, indicating it is highly dependable for dermatology diagnoses. The correct diagnosis consistently ranks first, showcasing the model's robustness in this specialty. |
| Infectious Disease | 100% | 83.33% | 0.76 | The model exhibits high accuracy and strong Top-3 Accuracy, ensuring the correct diagnosis is frequently among the top three predictions. A relatively high MRR suggests that the correct diagnosis is typically ranked high, though there is room for further improvement. |
| Respiratory | 100% | 100% | 0.92 | Perfect scores in both accuracy and Top-3 Accuracy demonstrate that the model reliably makes correct diagnoses for respiratory |

| | | | | |
|---|---|---|---|---|
| | | | | conditions. The high MRR further confirms that the correct diagnosis is usually at the top, ensuring quick identification. |
| Urology | 100% | 100% | 0.89 | The model achieves flawless scores in both accuracy and Top-3 Accuracy, with a high MRR indicating exceptional reliability in urology diagnoses. The correct diagnosis generally ranks very high, enhancing the model's practical utility in this specialty |

**Moderate-Performing Specialties**

| **Specialities** | Accuracy | Top-3 Accuracy | MRR | Analysis |
|---|---|---|---|---|
| General Physician | 87.21% | 84.88% | 0.79 | The model shows high accuracy and strong Top-3 Accuracy, indicating reliability for general physician diagnoses. A relatively high MRR suggests that correct diagnoses are usually ranked highly, making the model practically useful. |
| Acute Medicine | 83.33% | 66.67% | 0.56 | Although the model achieves high accuracy, the lower Top-3 Accuracy and MRR suggest that the correct diagnosis, while often the top |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  |  |  | prediction, is less frequently among the top three suggestions. This discrepancy indicates a need for improvement in the model's ranking abilities. |
| Neurology | 83.33% | 66.67% | 0.84 | The model achieves high accuracy with a high MRR, indicating reliability in neurology diagnoses. However, the Top-3 Accuracy could be improved to ensure the correct diagnosis is consistently among the top three predictions. |
| Obstetrics | 83.33% | 83.33% | : 0.75 | High accuracy and Top-3 Accuracy, with a relatively high MRR, indicate the model's reliability in obstetrics. Nonetheless, there is room for enhancing the ranking of the correct diagnosis. |
| Ophthalmology | 85.71% | 85.71% | 0.52 | While the model shows high accuracy and Top-3 Accuracy, the lower MRR suggests that the correct diagnosis is not always ranked highly. This indicates a need for improvement in ranking efficiency. |

**Low -Performing Specialties**

| **Specialities** | Accuracy | Top-3 Accuracy | MRR | Analysis |
|---|---|---|---|---|
| Emergency | 66.67% | 66.67% | 0.47 | The model's moderate performance in emergency cases, coupled with a relatively low MRR, indicates that the correct diagnosis is not consistently ranked highly. This highlights an area for improvement in ranking accuracy. |
| Endocrinology | 66.67% | 50% | 0.5 | Lower performance across all metrics suggests the model requires significant improvement for endocrinology diagnoses. Enhancing both accuracy and ranking should be prioritized. |
| Gastroenterology | 66.67% | 66.67% | 0.56 | Moderate performance with consistent Top-3 Accuracy indicates that the model can include the correct diagnosis within the top three predictions but needs improvement in overall accuracy and ranking |
| Geriatrics | 66.67% | 66.67% | 0.67 | Similar to cardiology, the model shows moderate performance in |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  |  |  | geriatrics, indicating a need for improvement in both accuracy and ranking reliability. |
| Orthopedics | 66.67% | 66.67% | 0.67 | The model demonstrates moderate performance in orthopedics, similar to other specialties with 66.67% accuracy. Improving the model's accuracy and ranking reliability is necessary. |
| Cardiology | 66.67% | 66.67% | 0.67 |  |

The model's performance varies significantly across different medical specialties. Specialties such as Dermatology, Infectious Disease, Respiratory, and Urology exhibit high accuracy and reliability, with perfect or near-perfect scores in accuracy and Top-3 Accuracy. These specialties demonstrate the model's robustness and potential for practical applications.

In contrast, specialties like Cardiology, Emergency, Endocrinology, Gastroenterology, Geriatrics, and Orthopedics show moderate performance, with accuracy and Top-3 Accuracy around 66.67%. These areas require targeted improvements to enhance the model's overall reliability and effectiveness.

This comprehensive analysis guides further refinement of the diagnostic model. Focusing on improving performance in moderate-performing specialties while leveraging strengths in high-performing areas will enhance the model's overall utility and reliability across various medical fields.

## Discussion

The integration of artificial intelligence, particularly conversational AI models like Chat GPT, into healthcare is transforming clinical decision-making and patient care. This study evaluated the diagnostic accuracy of Chat GPT in simulated clinical scenarios, using Objective Structured Clinical Examination (OSCE) cases as a benchmark. The results provide important insights into the model's strengths and areas needing improvement. Chat GPT demonstrated a diagnostic accuracy rate of 85%, correctly identifying diagnoses in 85 out of 100 cases. This high accuracy indicates the model's potential as a reliable tool for clinical decision support. However, the 15% error rate suggests that further development and refinement are necessary to reduce diagnostic inaccuracies.

### Top-3 Accuracy and Clinical Utility

The Top-3 Accuracy rate was 82%, showing that the correct diagnosis was among the top three predictions in 82 out of 100 cases. This is particularly useful in clinical practice, where considering multiple differential diagnoses is common. Having the correct diagnosis within the top three options helps healthcare providers narrow down potential conditions, enhancing decision-making.

### Mean Reciprocal Rank (MRR)

An MRR of 1.12 underscores the model's effectiveness in ranking the correct diagnosis. The frequent placement of the correct diagnosis in the first or second position improves clinical decision-making by reducing errors and improving patient outcomes.

### Specialty-Specific Performance

Chat GPT's performance varied across different medical specialties. High accuracy and reliability were noted in Dermatology, Infectious Disease, Respiratory, and Urology, indicating the model's robustness in these areas. In contrast, moderate performance was observed in Cardiology, Emergency Medicine, Endocrinology, Gastroenterology, Geriatrics, and Orthopaedics, with accuracy and Top-3 Accuracy around 66.67%. This suggests the need for targeted improvements in these specialties to enhance overall reliability and effectiveness.

## Conclusion

This study demonstrates that conversational AI models like Chat GPT have significant potential to enhance clinical decision support, showing high diagnostic accuracy in simulated clinical scenarios. The model's strong Top-3 Accuracy and MRR further support its utility in clinical settings, providing robust support for differential diagnoses and reducing diagnostic errors.

### Implications for Healthcare

AI models in healthcare have the potential to revolutionize patient care by providing timely and accurate clinical decision support. However, this study also highlights the importance of ongoing refinement, particularly in specialties where performance was moderate. Addressing these areas for improvement can optimize AI models to ensure reliability and effectiveness across a wider range of clinical scenarios.

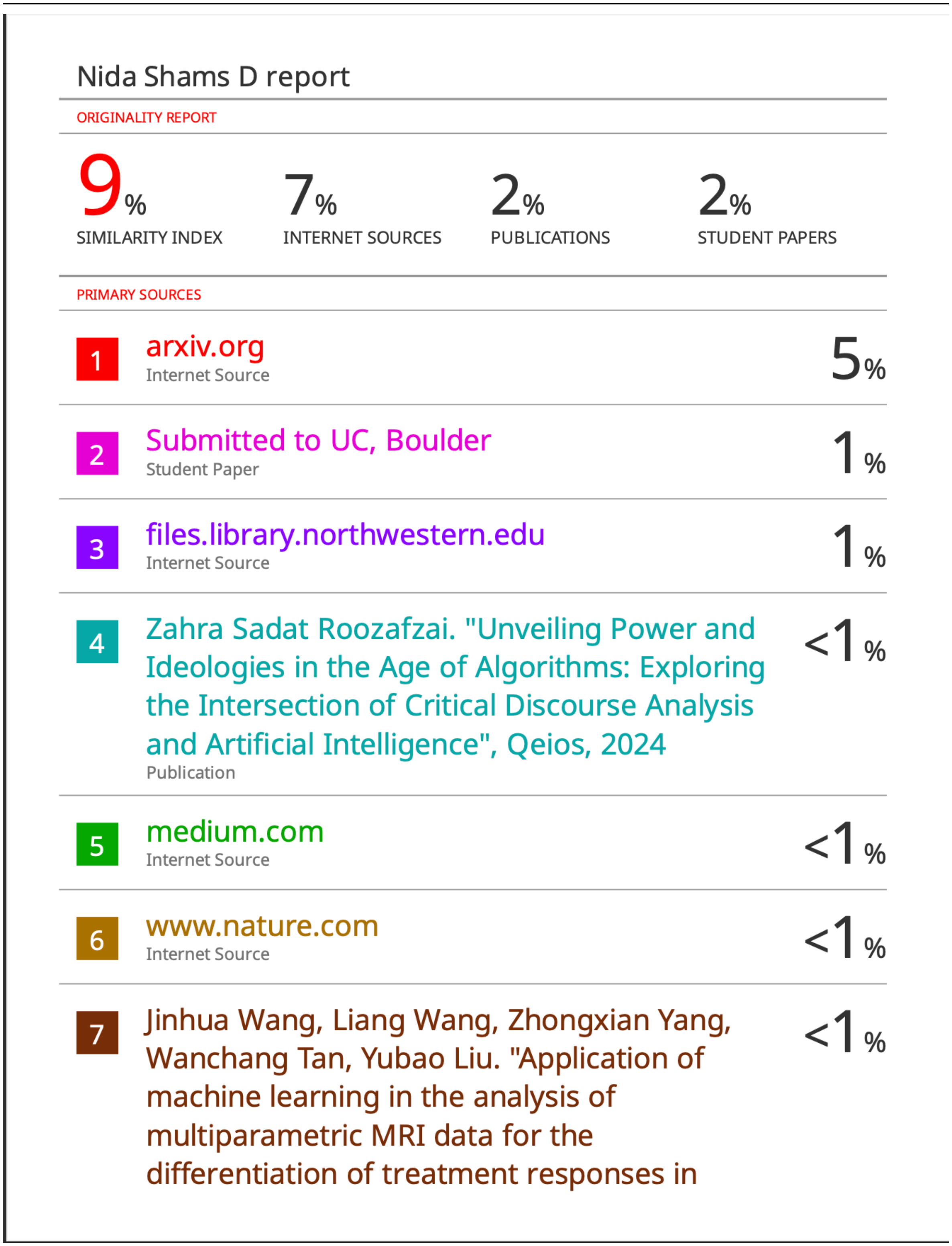
## References

[i] *Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, Vivek Natarajan. Towards Conversational Diagnostic AI. arXiv.*

[ii] *Patel BN, Rosenberg L, Willcox G, Sidhu P, Brown T, Gupta R, et al.* Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *JMIR Med Inform. 2023;11 doi: 10.2196/37203.*

[iii] *Das S, Devakumar D, Murali S, Duraiswamy K, Madhavan V.* Enhancing Diagnostic Accuracy with Large Language Models in Clinical Settings. *arXiv.*

[iv] *Natarajan A, Smith L, Jones T, Lee K.* Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *ResearchGate.*



Nida Shams D report

ORIGINALITY REPORT

| 9% | 7% | 2% | 2% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | arxiv.org<br>Internet Source | 5% |
| 2 | Submitted to UC, Boulder<br>Student Paper | 1% |
| 3 | files.library.northwestern.edu<br>Internet Source | 1% |
| 4 | Zahra Sadat Roozafzai. "Unveiling Power and Ideologies in the Age of Algorithms: Exploring the Intersection of Critical Discourse Analysis and Artificial Intelligence", Qeios, 2024<br>Publication | <1% |
| 5 | medium.com<br>Internet Source | <1% |
| 6 | www.nature.com<br>Internet Source | <1% |
| 7 | Jinhua Wang, Liang Wang, Zhongxian Yang, Wanchang Tan, Yubao Liu. "Application of machine learning in the analysis of multiparametric MRI data for the differentiation of treatment responses in | <1% |