INTERNSHIP TRAINING

At

Jivi.ai

Gurgaon

Study/Project Title

A Comparative Study: Evaluating the Large Language Model (LLM) For AI-Generated Medical Response

By

Dr Aayushi Singh

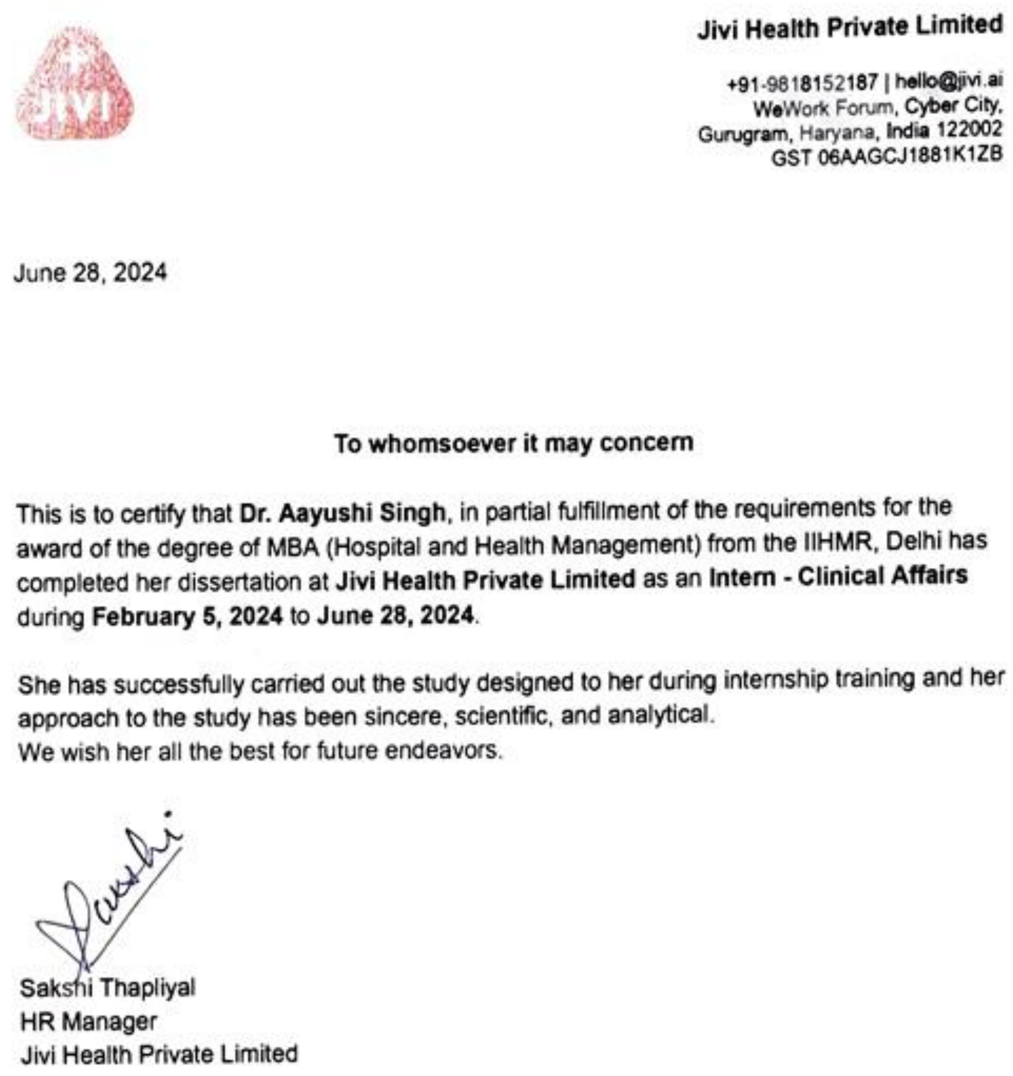
PG/22/002

UNDER THE GUIDANCE OF

Dr Divya Aggarwal

PGDM (Hospital and Health Management) 2022-2024

International Institute of Health Management Research New Delhi



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 **Dr. Aayushi Singh**, 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 5, 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 Thapliyal
HR Manager
Jivi Health Private Limited

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **Dr Aayushi Singh student** of **PGDM (Hospital & Health Management) from the International Institute of Health Management Research**, New Delhi has undergone internship training at **Jivi. Ai** from Feb 2024 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 fulfilment of the course requirements.

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

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

Aggarwal
Dr Divya Aggarwal
Associate Professor
IIHMR, New Delhi

Certificate of Approval

The following dissertation titled "**A Comparative Study: Evaluating the Large Language Model (LLM) for AI-generated Medical Response**" at "**Jivi.Ai**" 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	
Dr. Anandhi	

Ekta Saroha

Ekta Saroha

Certificate from Dissertation Advisory Committee

This is to certify that **Dr. Aayushi Singh**, a graduate student of the PGDM (Hospital & Health Management) has worked under our guidance and supervision. She is submitting this dissertation titled "**A Comparative Study: Evaluating the Large Language Model (LLM) for AI-generated Medical Response**" at "**Jivi. Ai**" 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.

Dr Divya Aggarwal
Associate Professor
IIHMR, New Delhi

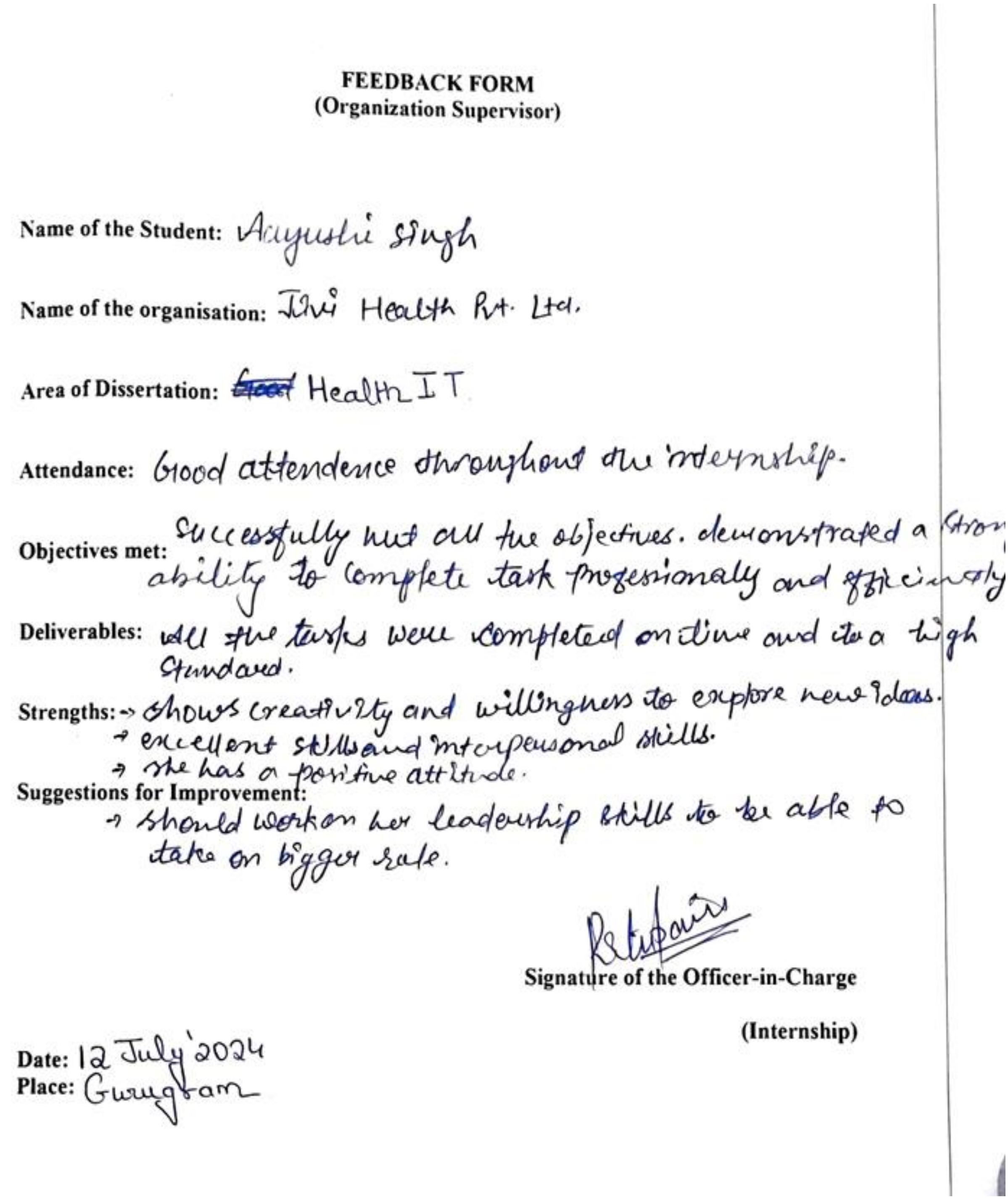
Organization Mentor
Dr. Gurukiran Babu Tumma
VP Clinical Services
Jivi.Ai

NEW DELHI

CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled **A Comparative Study: Evaluating the Large Language Model (LLM) for AI-generated Medical Response** and submitted by **Dr. Aayushi Singh** Enrolment No. **PG/22/002** under the supervision of **Dr. Divya Aggarwal** 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: Aayushi Singh

Name of the organisation: Jivi Health Pvt. Ltd.

Area of Dissertation: ~~Good~~ Health IT

Attendance: Good attendence throughout the internship.

Objectives met: Successfully met all the objectives. demonstrated a strong ability to complete task professionally and efficiently

Deliverables: All the tasks were completed on time and to a high standard.

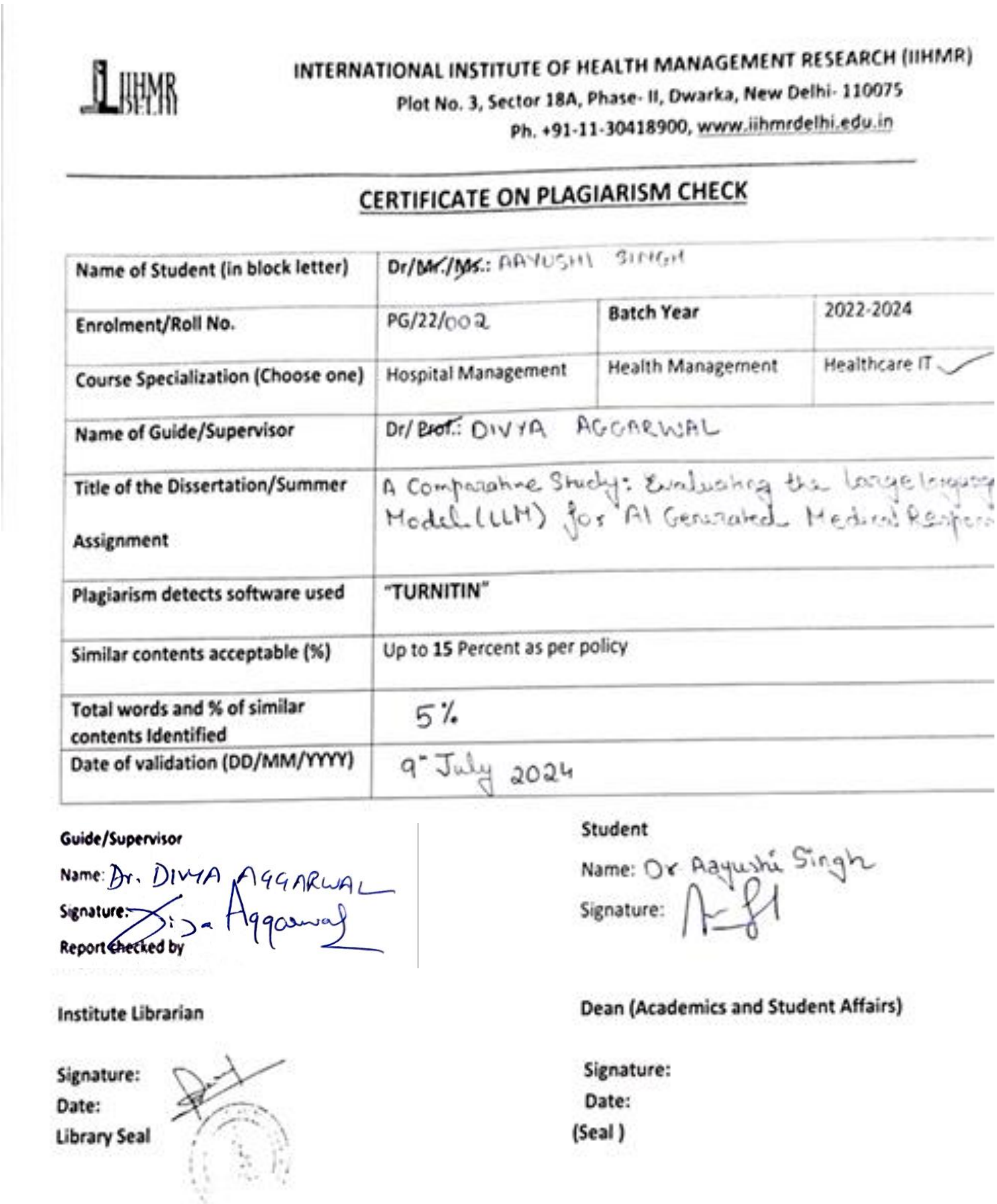
Strengths:
→ shows creativity and willingness to explore new ideas.
→ excellent skills and interpersonal skills.
→ she has a positive attitude.

Suggestions for Improvement:
→ should work on her leadership skills to be able to take on bigger role.

Signature of the Officer-in-Charge

(Internship)

Date: 12 July'2024
Place: Gurugram



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.~~: AAYUSHI SINGH		
Enrolment/Roll No.	PG/22/002	Batch Year	2022-2024
Course Specialization (Choose one)	Hospital Management	Health Management	Healthcare IT ✓
Name of Guide/Supervisor	Dr/ ~~Prof.~~: DIVYA AGGARWAL		
Title of the Dissertation/Summer Assignment	A Comparative Study: Evaluating the large language Model (LLM) for AI Generated Medical Respon		
Plagiarism detects software used	"TURNITIN"		
Similar contents acceptable (%)	Up to **15** Percent as per policy		
Total words and % of similar contents Identified	5%		
Date of validation (DD/MM/YYYY)	9th July 2024		

Guide/Supervisor
Name: Dr. DIVYA AGGARWAL
Signature: Divya Aggarwal
Report checked by

Student
Name: Dr Aayushi Singh
Signature:

Institute Librarian

Signature:
Date:
Library Seal

Dean (Academics and Student Affairs)

Signature:
Date:
(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. Divya Aggarwal, 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 a 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 Transformer
3	NLP	Natural Language Processing
4	AI	Artificial Intelligence
5	RAG	Retrieval Augmented Generation
6	ENT	Ear Nose Throat

Table of Contents

Figures, Tables, and Graphs

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

Background

LLMs are neural network architectures trained on massive amounts of text data. This allows them to analyse and synthesize information, generate different creative text formats, and answer questions in an informative way. Large Language Models (LLMs) such as GPT-4 have created new opportunities for AI applications in the medical domain, offering possible assistance to medical practitioners in the form of precise and thorough medical responses. To ascertain the efficacy and practicality of three top LLMs—GPT-4, Anthropic, and Cohere—in various medical specializations, this study assesses their performance.

Objectives

This study aims to assess the accuracy, completeness, and user satisfaction of GPT-4, Anthropic, and Cohere's AI-generated medical replies. The study also seeks to determine each model's advantages and disadvantages to facilitate its adoption into healthcare environments.

Methods

Creating medical answers for a series of pre-formulated questions covering eighteen different medical fields was the evaluation's task. Using a Likert scale, responses were evaluated for accuracy, completeness, and user satisfaction. A statistical study was done to compare the performance of the three LLMs, and healthcare professionals evaluated the responses to provide satisfaction ratings.

Key Findings

- Overall Performance: In 16 of the 18 fields of medicine, GPT-4 fared better than Anthropic and Cohere.
- Accuracy: With a mean accuracy score of 5.4, GPT-4 was the most accurate, followed by Anthropic; Cohere continuously displayed poorer accuracy.
- Completeness: With a mean score of 2.6, both GPT-4 and Anthropic received good marks for completeness, whereas Cohere received a lesser score of 1.7.
- Satisfaction: The GPT-4 had the highest level of satisfaction, which is indicative of improved usage and acceptance among medical professionals.

- Response Length: GPT-4 successfully struck a balance between brevity and detail, while Anthropic occasionally gave answers that were overly verbose and lengthy.

Conclusion

The best LLM for producing medical responses is GPT-4, which offers the most accuracy, comprehensiveness, and user experience. Even while Anthropic performs competitively in some areas, its lengthy replies might need to be optimized. On the other hand, Cohere requires a lot of work in terms of accuracy and thoroughness. The study emphasizes how LLMs might improve patient education and clinical decision assistance, but it also emphasizes the necessity for ethical standards, data privacy protections, and ongoing model improvement.

Limitations

- Limited sample size and specialization scope.
- Subjectivity in ratings of user satisfaction.
- Controlled environment assessments that do not incorporate the real world.
- Updates and ongoing assessments of LLM capabilities are required.

Implications and Future Research

Ensuring ethical use and overcoming practical issues are necessary for the effective integration of LLMs into healthcare. Subsequent investigations ought to concentrate on practical testing, enhancing contextual comprehension, and examining the enduring effects on patient results and healthcare efficacy.

Section -1

Introduction

Large language models (LLMs) are becoming a significant milestone as the demand for comprehensive natural language processing skills keeps rising. The way we engage with text has changed tremendously because of the rapid advancement of AI technology. LLMs enable us to communicate, evaluate, and create knowledge with an array of sophistication that was previously not achievable.

The potential for significantly enhancing patient and healthcare provider access to medical information emerges with the integration of natural language processing (NLP) models. Machine learning methods called large language models (LLMs) can comprehend and produce text that mimics that of a person.

Large language models can learn from data with greater efficiency than traditional supervised deep learning models thanks to a two-stage training process that involves self-supervised learning on enormous amounts of unannotated data in the first stage and fine-tuning on smaller, task-specific, annotated datasets in the second stage so they can function on end-user-specified tasks.

Chat-Generative Pre-Trained Transformer (ChatGPT), a conversational chatbot built on top of Generative-Pre-Trained Transformer-3.5 (GPT-3.5), an LLM featuring about 175 billion parameters, is one such AI-powered technology that is currently becoming highly discussed. The training data used by ChatGPT comes from a variety of online resources, including books, papers, and web pages.

By employing reinforcement learning from human feedback to refine ChatGPT for conversational tasks, the algorithm grows capable of considering the complex nature of users' intentions and can manage a range of end-user tasks, including queries related to medicine.

Considering the growing volume of medical data and the intricacy of clinical decision-making, NLP technologies may be able to help doctors make timely, well-informed judgments, thereby enhancing the general effectiveness and quality of healthcare.

Without any specific training, ChatGPT performed at or close to the USMLE passing criteria, indicating its potential for clinical decision assistance and medical education.

Furthermore, the democratization of knowledge generated by technological improvements means that patients are no longer primarily dependent on medical professionals for information. In fact, as easy-to-use and readily available sources of medical information, people are increasingly using search engines and, more recently, AI chatbots. Along with other newly announced chatbots, ChatGPT conducts dialogues and answers complex medical questions authoritatively.

Nevertheless, despite its promise, ChatGPT frequently generates outputs that appear reliable but are inaccurate, thus its use in medical practice and research should be approached with caution. It has not been determined whether these engines are accurate and reliable, especially when it comes to open-ended medical queries that patients and doctors are likely to ask.

Here in this research, we have used 3 large language models- Anthropic, Cohere, and GPT.

1) Anthropic

Claude, Anthropic's language model, has an array of remarkable characteristics and unique ethical considerations. Claude, created by a group of former OpenAI researchers, uses a novel "Constitutional AI" technique to highlight AI safety and alignment.

Features:

Constitutional AI: A "constitution" of moral and human rights-based precepts is incorporated into Claude's training. Creating interactions that are constructive and safe, directs the AI's self-evaluation and response production. Long Context Handling: For applications demanding long-term coherence and in-depth document analysis, Claude 3 models can handle a broad context window of up to 1 million tokens.

Various Models:

Claude 3 comes in three different forms:

1) **Haiku:** Quick and economical, good for easy jobs.
2) **Sonnet:** Perfect for most enterprise applications, it strikes a balance between cost and intelligence.
3) **Opus:** A top-tier model with sophisticated thinking and innovative problem-solving ability for intricate jobs.

Accuracy and Reliability: The most recent Claude models show notable increases in accuracy and a decrease in false information. To improve trustworthiness features like citation support are being implemented.

Document and Data Processing: Claude is capable of reading and analyzing lengthy papers. He also provides features that are helpful for enterprises and research, like summary and in-depth data extraction.

Limitations on Ethics:

Alignment Tax: The model's usability may occasionally be hampered by the emphasis on ethical alignment. For instance, Claude could decline to help with certain programming questions if they seem dangerous, which some users find restricting.

Bias and Fairness: Although biases have been reduced, it is still difficult to strike a balance between objectivity and usefulness. Although the model has improved in terms of bias benchmarks, it is still developing in this regard.

Safety and Transparency: Claude is made to be safe and transparent while reducing hazards like false information and privacy concerns. But occasionally, these safety measures encourage conservative conduct, which can restrict their usefulness in particular situations.

2) **<u>Cohere</u>**

The language models (LLMs) from Cohere have been developed to provide strong, adaptable natural language understanding and generating capabilities, meeting the requirements of many business and research applications.

Principal Elements of the Cohere LLM Command Models:

1) The command-following models from Cohere, such as Command R and Command R+, are designed to generate text, carry out conversational activities, and manage intricate workflows like tool use and retrieval-augmented generation (RAG) and tool use (Cohere Enterprise Group).

2) Multilingual Support: Aya along with other Cohere models is multilingual. Specifically, Aya is a massively multilingual model that supports 101 languages and concentrates on underutilized languages to improve tasks like translation and summarization.

3) Flexibility in Deployment: Cohere's models can be implemented on multiple platforms, such as Microsoft Azure, Oracle Cloud, and Amazon SageMaker. considering their versatility, enterprises can easily incorporate these models into their current cloud infrastructure.

4) Enhanced Retrieval and Ranking: By employing sophisticated embedding and re-ranking strategies, the models perform exceptionally well in retrieval-augmented generation. This is achieved by greatly enhancing the relevance and precision of generated replies from large datasets.

5) Tool Integration: Command models are useful for automating business operations since they are made to work with databases and software applications to streamline workflows.

Limitations and Considerations in Ethics:

1) Bias and Fairness: Comparable to many LLMs, the generated material may contain biases resulting from the data used to train the models. Cohere (Cohere Enterprise Group) recognizes these drawbacks and works to address them via policies for responsible use and ongoing model evaluation.

2) Data Security and Privacy: It is essential to guarantee the security and privacy of the data that these models process. Although Cohere has strong security mechanisms in place, users still need to exercise caution when entering any kind of data into the models (Cohere Enterprise Group).

3) Environmental Impact: There are environmental risks associated with the massive computational resources required for training and deploying large models. To solve this, Cohere (Cohere Enterprise Group) optimizes its models for effectiveness and transparency in reporting its environmental impact.

4) Licensing and Usage Restrictions: Special licensing agreements may be essential for commercial use, even if some models, such as Command R+, are available for research reasons. This guarantees that the models are applied in commercial applications responsibly and morally.

Chat GPT 4 (Generative-Pre-Trained Transformer)

Natural language processing has advanced to a new level with OpenAI's GPT (Generative Pre-trained Transformer) series, which includes GPT-4. GPT models are adaptable tools for a range of applications since they are made to produce text that appears human-like in response to input cues.

Language Understanding and Generation: GPT-4 can understand and produce text in a variety of languages, just like its predecessors. It is excellent at delivering responses that are both coherent and culturally appropriate, which makes it helpful for jobs like summarizing, translating, and creating content.

Enhanced Capabilities: GPT-4 is more accurate, fluent, and versatile than previous iterations. It has been optimized for enhanced efficiency in low-resource languages and exhibits sophisticated reasoning and instruction compliance. Because of this, it can be used for a variety of purposes, including customer service and education.

Applications:

Language Translation: Facilitates precise yet speedy language translation.

Customer service: Enables virtual assistants and chatbots to deliver prompt, conversational responses.

Programming Assistance: Provides code snippets, debugs, and ideas for code enhancement to developers.

Ethical Considerations:

Like other large language models, GPT-4 has significant ethical challenges despite its great capabilities.

1) Fairness and Bias: Due to the biases in its training data, GPT-4 may produce content that is unsuitable or prejudiced. Thus, continual efforts are required to identify and reduce these biases to guarantee impartial and fair results.

2) Misinformation: The model may generate information that is misleading or factually inaccurate. While GPT-4 is intended to lessen these kinds of incidents, ongoing observation and development are necessary to increase its dependability.

3) Privacy Concerns: A significant portion of the publicly accessible text used as training data for GPT models may unintentionally contain sensitive or private information. To deploy and use these models, it is imperative to ensure data security and privacy.

4) Misuse Potential: GPT-4's potent text-generating powers could be exploited for malicious ends, such as generating spam, creating deepfakes, or launching social engineering scams. Strong usage guidelines and oversight procedures must be put in place to prevent such types of usage.

In the current digital era, big language models are essential for giving precise, thorough, and morally sound medical responses that improve patient education, healthcare delivery, and medical research. With continued development and enhancement, these models can fundamentally transform the methods of obtaining, comprehending, and utilizing medical information, thereby contributing to the improvement of worldwide health and welfare

Section -2

Rationale

Large language models (LLMs) have great potential to improve patient outcomes, help medical practitioners, and improve healthcare delivery when incorporated into medical practice. To ensure safety, accuracy, and relevance, this potential must be carefully evaluated and balanced.

The following important factors serve as the foundation for the study's reasoning:

1) **Developing Need for AI in Medical Fields**:

The volume and complexity of medical data are expanding, and healthcare practitioners are looking more and more to AI-driven solutions for support. Workflows may be streamlined, illnesses may be diagnosed with the help of LLMs, and advice based on solid evidence may be given. Assessing their precision and comprehension of context is essential to properly utilizing their strengths.

2) **Importance of Reliable and Accurate Medical Information**: Accurate information is crucial while making medical decisions. Any errors in AI-generated replies may result in incorrect diagnoses, ineffective treatment strategies, and even patient injury. By reducing the possibility of errors, this work seeks to guarantee that LLMs can deliver trustworthy and accurate medical information.

3) **Evaluation of Contextual Understanding**: Various clinical departments face different needs and difficulties. LLM's applicability in a variety of specializations depends on its capacity to comprehend and produce contextually appropriate replies. By examining whether LLMs can adjust to the subtleties of various medical specialties, this study will assist ensure their wider applicability.

4) **Evaluation of Strengths and Limitations**: Though LLMs have demonstrated remarkable skills to produce content that resembles that of a human, it is crucial to comprehend their unique strengths and limitations in medical situations. This information will direct enhancements to the application, fine-tuning, and training of the models, improving their overall dependability and performance.

5) **Ensuring Patient Safety and Privacy**: Patient safety and privacy must come first when integrating LLMs into healthcare. This research will examine potential hazards and challenges related to the use of LLMs in medical settings and provide suggestions to

address these problems. For AI technologies to be widely used and successful, they must adhere to privacy laws and ethical norms.

6) **Recommendations for Successful Integration**: By offering doable suggestions grounded in empirical data, healthcare institutions will be able to successfully incorporate Life Cycle Managers (LCM) into their operations. The operational, ethical, and technological aspects of AI technology adoption in healthcare will be covered by this guidance, promoting a secure and effective implementation.

The overall goal of this research is to close the knowledge gap between LLM potential and real-world healthcare use, making sure that their use improves rather than degrades patient care. Through a thorough evaluation of these models and focused recommendations, the research aims to support the ethical, safe, and successful integration of AI in medicine.

Primary Objective

1. To evaluate and compare the accuracy of three different LLMs in generating medical responses.

Secondary Objective

1. To assess the contextual understanding and relevance of AI-generated medical responses in different clinical departments.
2. To identify the strengths and limitations of different models.
3. To provide recommendations for the integration of LLMs into medical applications, addressing challenges and ensuring patient safety and privacy.

The primary goal of this study is to evaluate and compare the accuracy of three different large language models (LLMs) in generating medical responses. This involves assessing how well each model performs in delivering precise and correct medical information, which is critical for ensuring the reliability and effectiveness of these AI systems in medical contexts.

While considered as an entire, these objectives seek to give a comprehensive assessment of LLMs in the medical area, stressing their advantages, pointing out their drawbacks, and providing helpful advice for their safe and efficient application in hospital environments.

Section -3

Review of Literature

This study's literature review discusses many facets of using large language models (LLMs) in healthcare, with an emphasis on user happiness, accuracy, and completeness while producing medical responses across a range of specializations.

The Origins and Expansion of LLMs

Natural language processing has advanced significantly thanks to large language models like GPT-4, Anthropic, and Cohere. These models may produce writing that is human-like and has applications in a variety of industries, including healthcare. They have been trained on large datasets. Their capacity to comprehend and produce complicated medical information has been demonstrated in earlier research that examined their potential in patient interaction and medical education.

Leveraging Apps in Specialized Medicine

The study looks at how well these models work in eighteen different medical disciplines, ranging from surgery and general medicine to more specialized areas like neurology and dermatology. Since each specialty has its own set of diagnostic standards and treatment guidelines, it is imperative to evaluate the models' performance in its entirety. Previous research has demonstrated that LLMs can offer precise medical advice, but there is heterogeneity based on the complexity and specificity of the inquiries (Bernstein et al., 2023; Mu et al., 2023).

Accuracy and Reliability

For LLM-generated responses to be accepted in therapeutic contexts, their correctness is essential. Studies have utilized a variety of criteria, like the Likert scale used in this study, to assess accuracy. According to research by Shen et al. (2023) and Kung et al. (2023), LLMs can generate accurate answers, but how well they work differs depending on the type of medicine. Because of this heterogeneity, a thorough comparison is required to pinpoint the advantages and disadvantages.

Satisfaction and Completeness

Completeness is the degree to which answers address every important facet of a medical question. This study used approaches akin to those described in earlier studies, measuring completeness on a 3-point scale. Healthcare practitioners' subjective assessments of replies are captured via the 5-point Likert scale used to measure satisfaction. Previous research, such as those of Giannakopoulos et al. (2023) and Ray (2023), indicates that although LLMs can be very fulfilling for users, verbosity and relevancy are issues.

Practical and Ethical Aspects to Consider

When incorporating AI into healthcare, ethical issues are crucial. In light of worries expressed in the literature regarding possible abuse and data privacy issues, this study highlights the importance of transparency, accuracy, and privacy. Studies conducted by Tam et al. (2024) and Borger et al. (2023) emphasize the significance of developing clinical standards and ethical guidelines to guarantee that AI tools improve patient care rather than worsen it.

Comparative Effectiveness

This study's comparative analysis demonstrates that, across most disciplines, GPT-4 performs better than other models. But Anthropic outperforms GPT-4 in several domains, such as urology and obstetrics and gynecology, which is consistent with Wilhelm et al. (2023) findings. This implies that while a single model might perform well overall, depending on the needs, different models might be more advantageous for specialized applications.

Prospective Courses

Subsequent investigations ought to concentrate on evaluating LLMs in actual clinical environments to confirm their effectiveness and effect on patient results. It is advised to improve contextual understanding and conduct long-term research on patient satisfaction and healthcare delivery. As Artsi et al. (2024) and Fournier et al. (2023) have argued, the specialization of LLMs for medical disciplines and the establishment of defined clinical guidelines for their usage will be vital.

In summary the evaluation of the literature emphasizes the benefits and drawbacks of employing LLMs in healthcare. The project intends to close the gap between AI capabilities and useful healthcare apps by analyzing their performance across specialties and guaranteeing that their integration improves patient care while upholding ethical norms.

Methodology

This research study employed a comparative analysis of three LLM models (Anthropic, Cohere, and GPT4). It included testing medical questions by different healthcare professionals. The questions were divided into 18 different medical specialties.

Study Design: Comparative Study

Duration of Study: 3 months

Data Type: Simulated data (Data collected from relevant medical Databases and customized based on most frequently asked questions by users)

Database- Marrow database (licensed database for the simulated data)

Tools- Large Language Models- Anthropic, Cohere, and Gpt and Excel

Data Analysis: Descriptive analysis of each specialty for all three large language models

Data Validation- Data was validated based on the 4 criteria mentioned (accuracy, completeness, satisfaction, and word count) and cross-referenced from the source Up-to-date (licensed medical database) and validated by the Expert medical professional.

Table 1: - Representation of specialty and number of questions in each

Sr. No	**Specialty**	**No. of Questions**
1	General Medicine	11
2	General Surgery	12
3	Emergency	13
4	Paediatrics	14
5	Neurology	8
6	Sexual Health	7
7	Dental	5

8	Vitals	6
9	Orthopaedics	18
10	Urology	14
11	Laboratory	28
12	Obstetrics and Gynaecology	11
13	Ophthalmology	11
14	ENT	13
15	Psychiatry	8
16	Respiratory	18
17	Cardiology	9
18	Dermatology	3

Each question from every specialty was tested on 3 Large Language Models- Anthropic, Cohere, and Gpt. The questions were selected to ensure complete comprehensive coverage of each specialty, including common conditions, diagnostic criteria, treatment protocols, and recent advancements.

The models have generated responses to each question without human intervention to ensure the objectivity of the analysis.

Reference responses were developed by compiling data from reliable medical sources, such as reputable medical textbooks and peer-reviewed research publications. In this research, the scoring under each parameter is mentioned in the criteria for assessment.
These reference answers were cross-verified by an Up-to-date system.

The **Up-to-date system** is an evidence-based clinical resource is the UpToDate system. It contains several medical calculators, access to Lexicomp drug monographs and drug-to-drug interactions, and a compilation of medical and patient data.

More than 7,100 doctors write for UpToDate as authors, editors, and peer reviewers. It can be accessed offline on mobile devices or personal computers as well as online.

Mean and median were calculated for each specialty comparing the 3 large language models on the following parameters.

1) Accuracy

2) Completeness

3) Satisfaction

Criteria for Assessment:

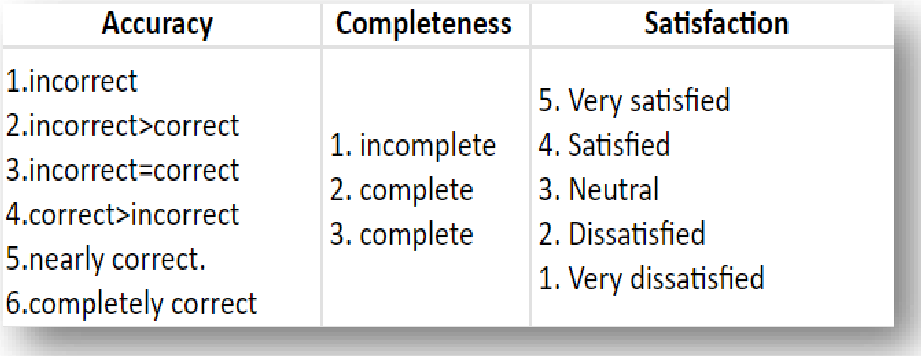
1) **Accuracy** is measured using a 6-point Likert scale, where 1 represents a full error and 6 represents a complete correction.

2) **Completeness:** Ranked from 1 (incomplete) to 3 (complete) on a 3-point Likert scale.

3) **Satisfaction**: Ranked from 1 (very unsatisfied) to 5 (very satisfied) on a 5-point Likert scale.

4) To evaluate the verbosity and conciseness of the responses, the word count of each response will also be noted.

The results are analyzed using relevant statistical tools by calculating the mean and median of each specialty based on the above-mentioned criteria.



Accuracy	Completeness	Satisfaction
1.incorrect 2.incorrect>correct 3.incorrect=correct 4.correct>incorrect 5.nearly correct. 6.completely correct	1. incomplete 2. complete 3. complete	5. Very satisfied 4. Satisfied 3. Neutral 2. Dissatisfied 1. Very dissatisfied

Ethical Considerations:

- There is absolutely no harm done to research. courtesy toward the dignity.
- Student Review Board clearance from IIHMR Delhi.
- Maintaining the privacy and anonymity of the medical professionals taking part in the assessment.
- Endeavor to eliminate any potential biases in the questions and reference answers chosen to preserve transparency in the approach and guarantee the objectivity of the analysis.

Section -4

Results:

The results were cross-referenced with Up-to-date and expert medical professionals.

The comparative analysis concludes that GPT4 has the highest performance score in most of the departments (16) except in Urology, Obstetrics, and Gynaecology in all the parameters of accuracy, completeness, and satisfaction.

Representations of Specialty Performance on each Parameter: -

1) **General Surgery: -**

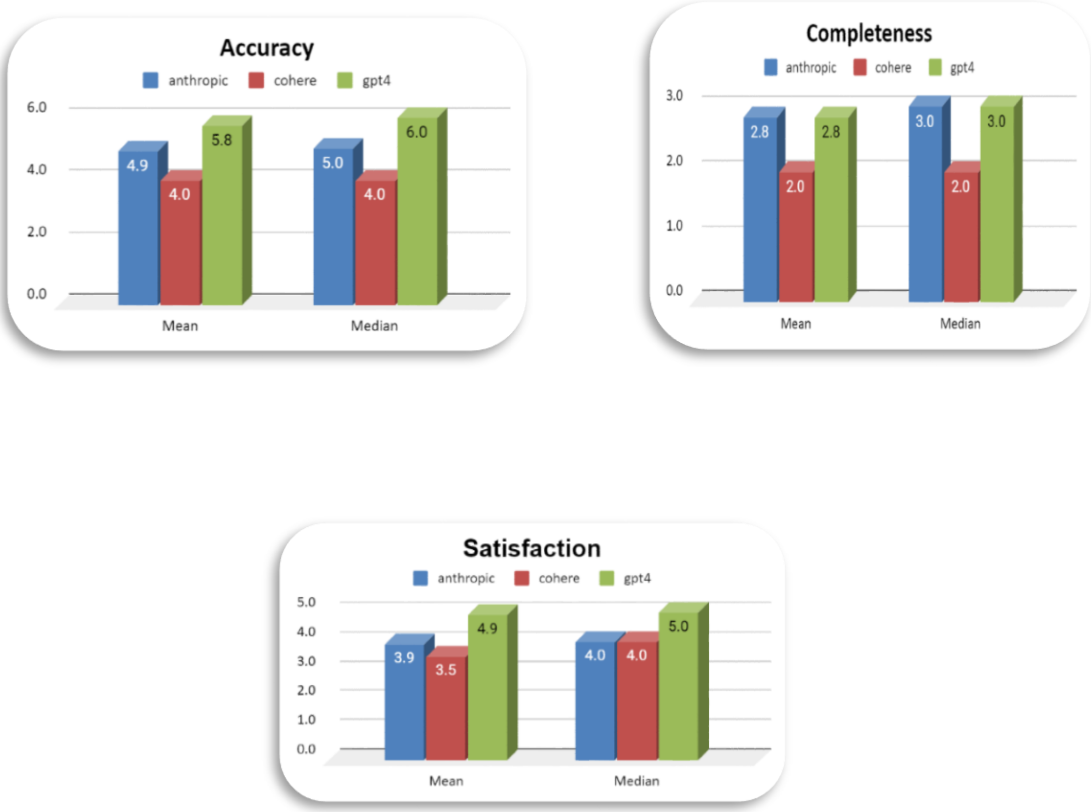


Figure 1 Shows Mean for General Surgery including all the parameters of the research

In General Surgery GPT 4 has shown the highest performance with a mean accuracy score of 5.8 and an average word count of 1307 words.

2) **General Medicine: -**

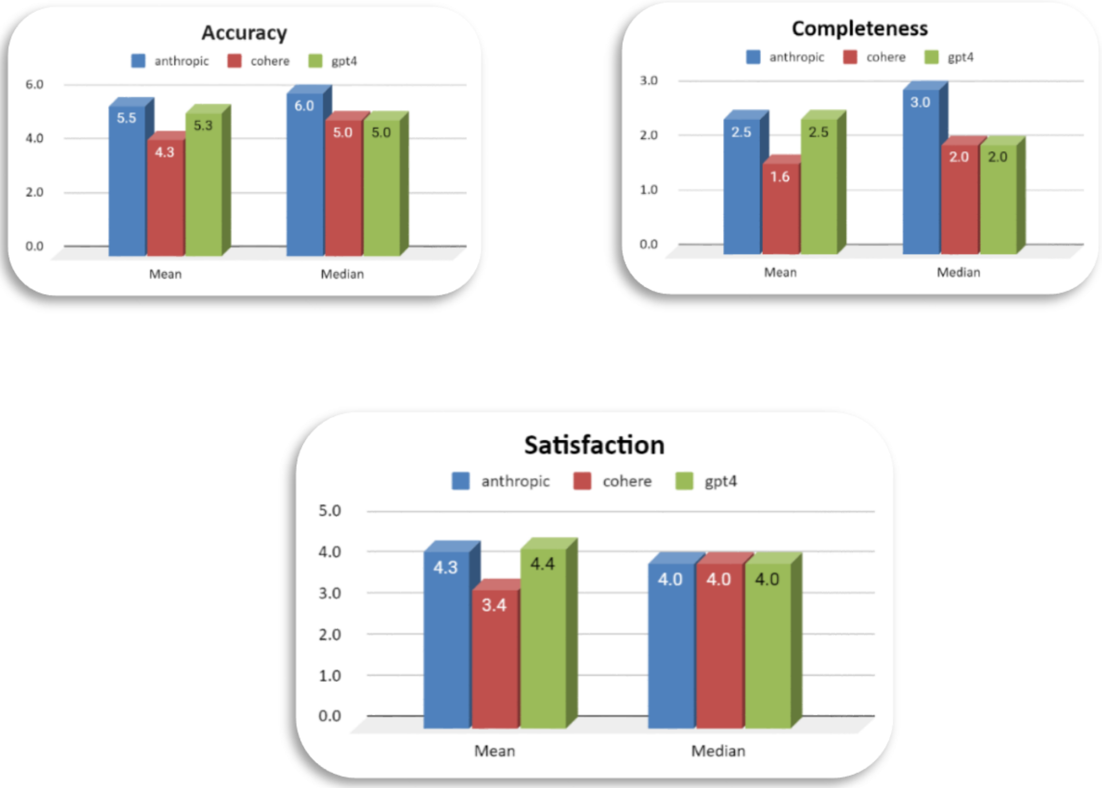
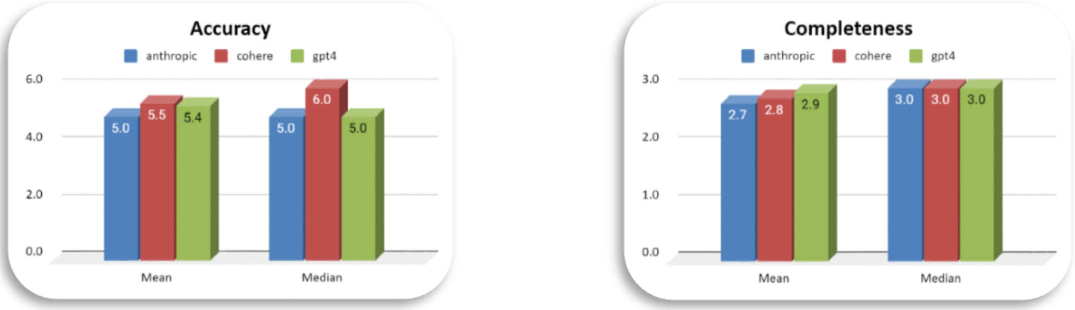


Figure 2 Shows Mean for General Medicine including all the parameters of the research

In General Medicine, GPT 4 has shown the highest performance with a mean accuracy score of 5.3 and an average word count of 1224 words.

3) **Ophthalmology: -**



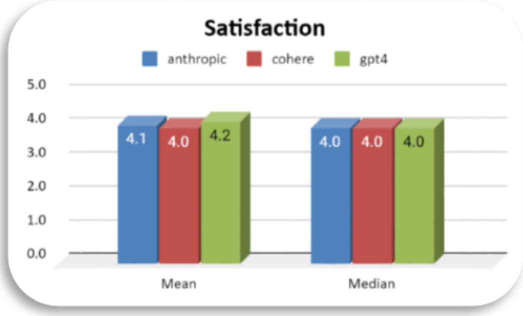


Figure 3 Shows Mean for Ophthalmology including all the parameters of the research

In Ophthalmology GPT 4 has shown the highest performance with a mean accuracy score of 5.4 and an average word count of 1226 words.

4) **Laboratory**: -

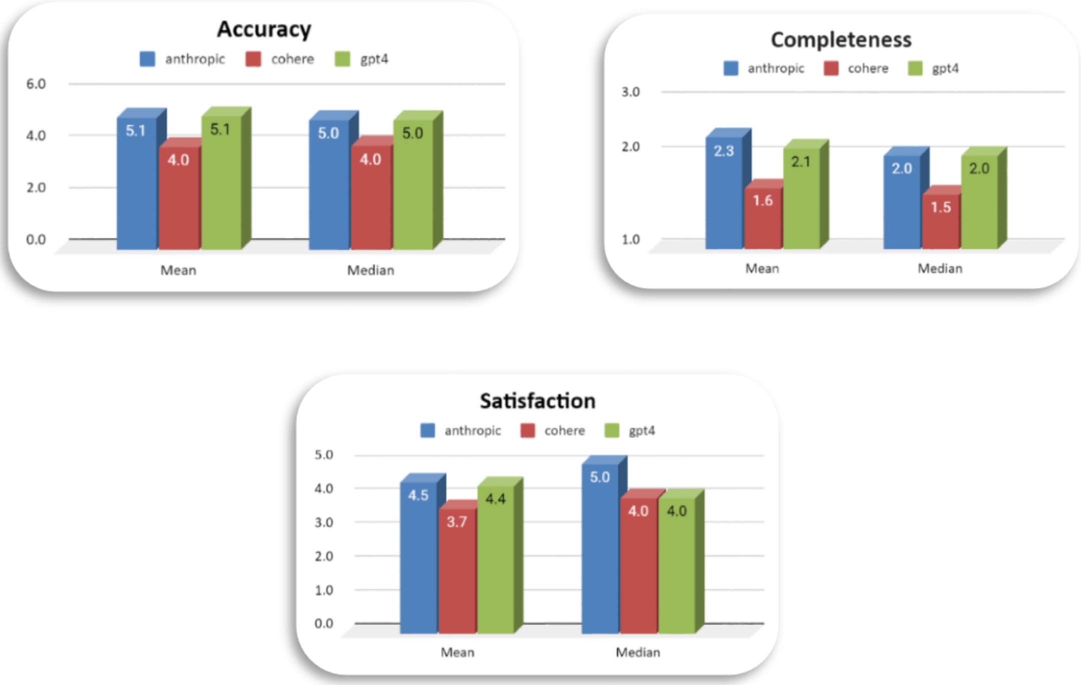


Figure 4 shows the Mean for the Laboratory including all the parameters of the research

In Laboratory GPT 4 has shown the highest performance with a mean accuracy score of 5.1 and an average word count of 1002 words.

5) **Paediatrics: -**

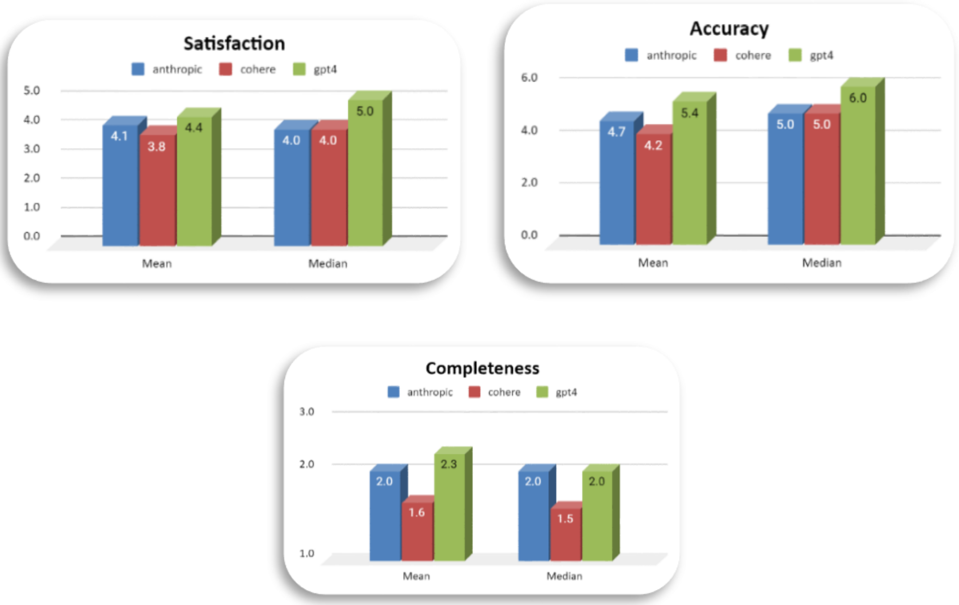
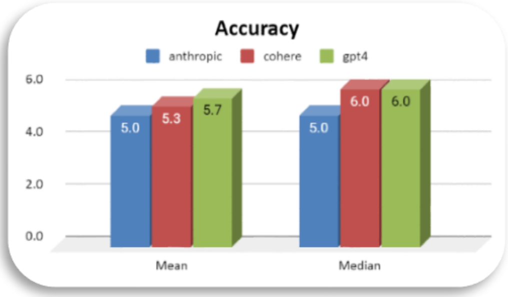
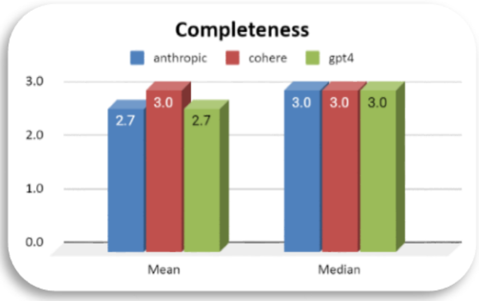


Figure 5 shows the Mean for the Laboratory including all the parameters of the research

In Paediatrics GPT 4 has shown the highest performance with a mean accuracy score of 5.4 and an average word count of 1259 words.

6) **Dermatology: -**





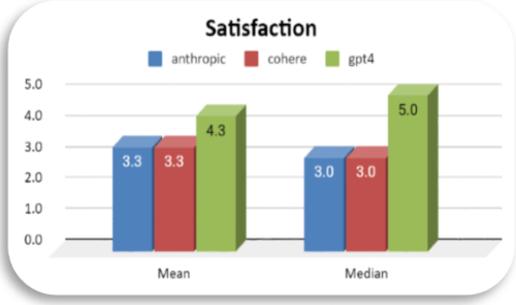


Figure 6 shows the Mean for the Dermatology including all the parameters of the research

In Dermatology GPT 4 has shown the highest performance with a mean accuracy score of 5.7 and an average word count of 1426 words.

7) **Dental:**

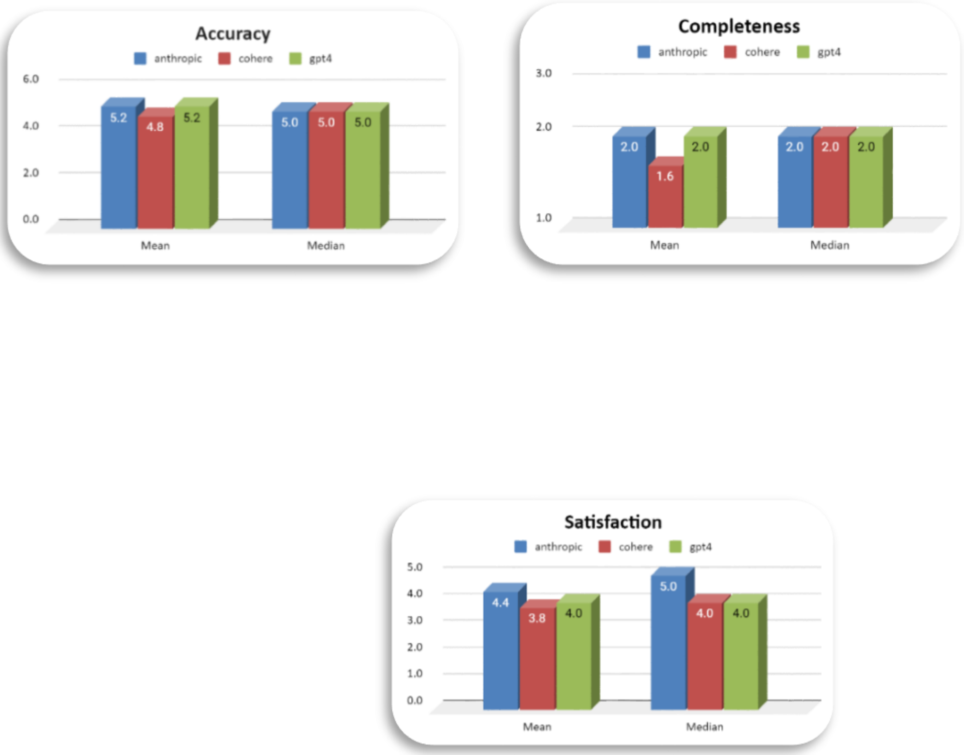


Figure 7 shows the Mean for the Dental including all the parameters of the research

In Dental GPT 4 has shown the highest performance with a mean accuracy score of 5.2 and an average word count of 744 words.

8) **Respiratory: -**

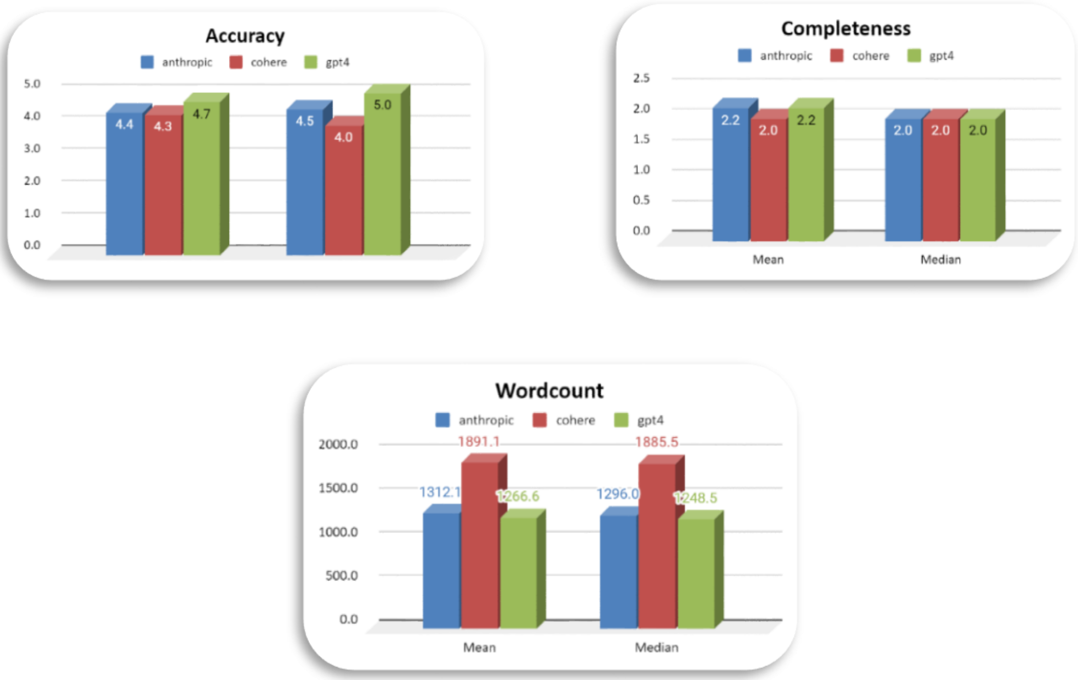
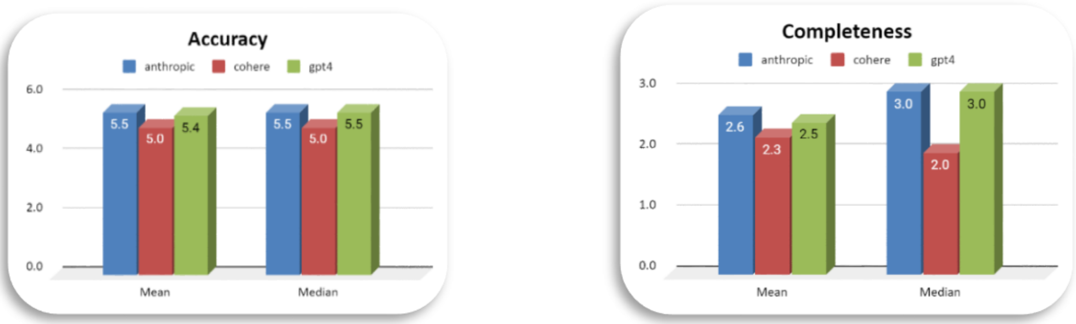


Figure 8 shows the Mean for the Respiratory including all the parameters of the research

In Respiratory GPT 4 has shown the highest performance with a mean accuracy score of 4.7 and an average word count of 1266 words.

9) **Neurology: -**



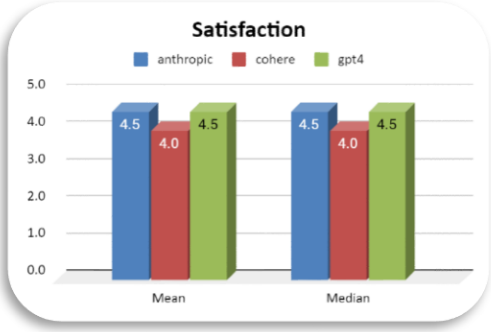


Figure 9 shows the Mean for the Respiratory including all the parameters of the research

In Neurology GPT 4 has shown the highest performance with a mean accuracy score of 5.4 and an average word count of 1320 words.

10) **Sexual Health**: -

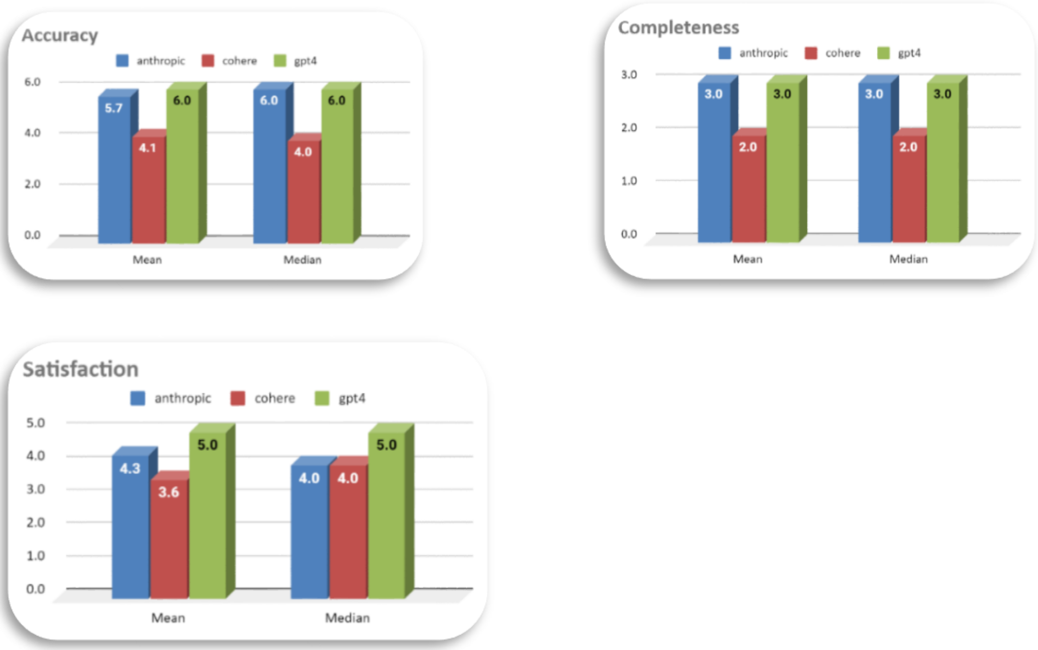


Figure 10 shows the Mean for Sexual Health including all the parameters of the research

In Sexual Health GPT 4 has shown the highest performance with a mean accuracy score of 6.0 and an average word count of 900 words.

11) **Vital Signs: -**

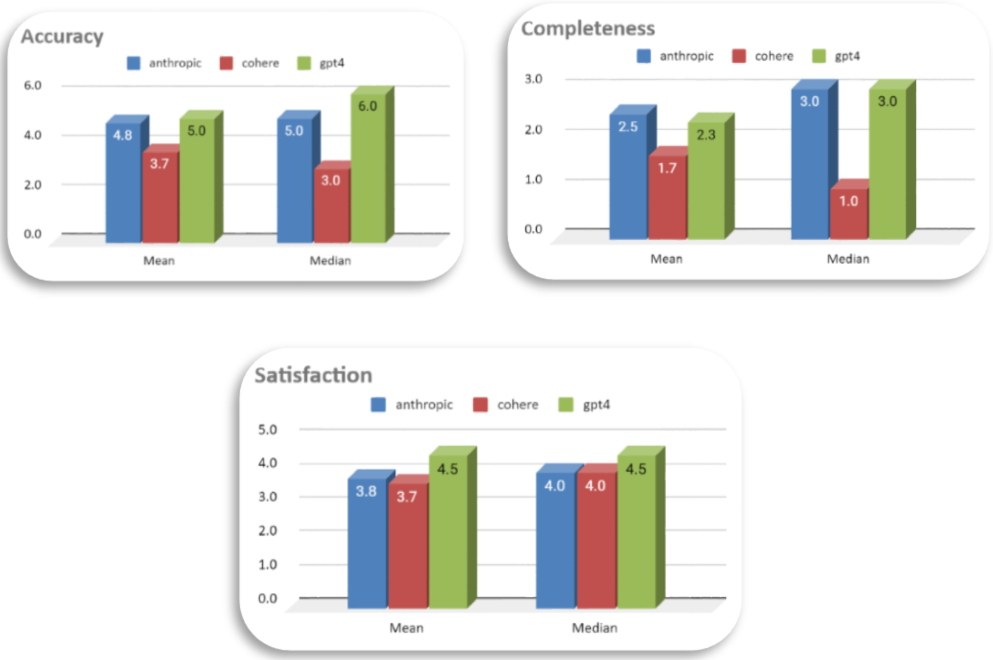


Figure 11 shows the Mean for Vital Signs including all the parameters of the research

In Vital Signs, GPT 4 has shown the highest performance with a mean accuracy score of 5 and an average word count of 987 words.

12) **Psychiatry: -**

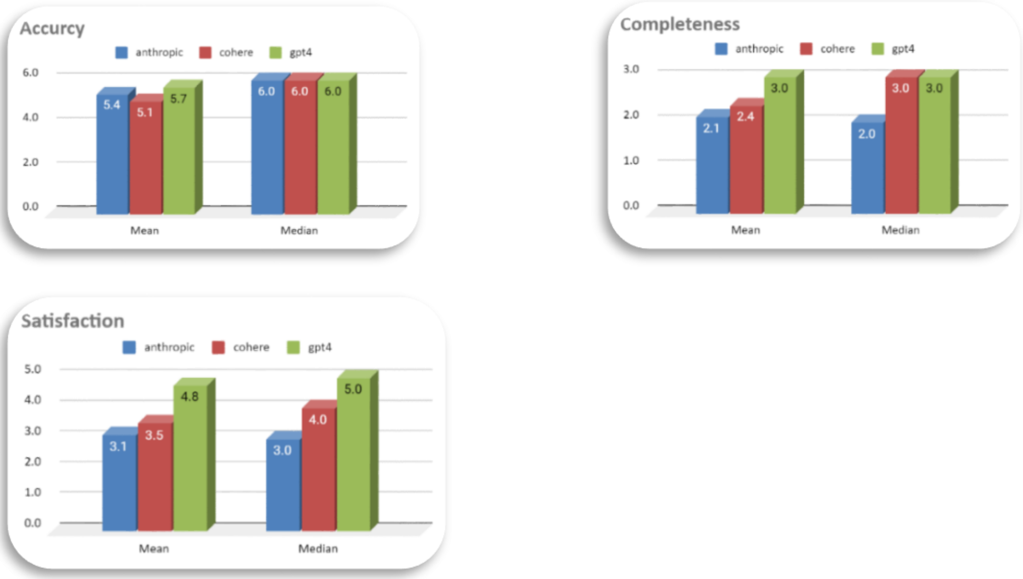


Figure 12 shows the Mean for Psychiatry including all the parameters of the research

In Psychiatry GPT 4 has shown the highest performance with a mean accuracy score of 5.8 and an average word count of 1206 words.

13) **Orthopaedics: -**

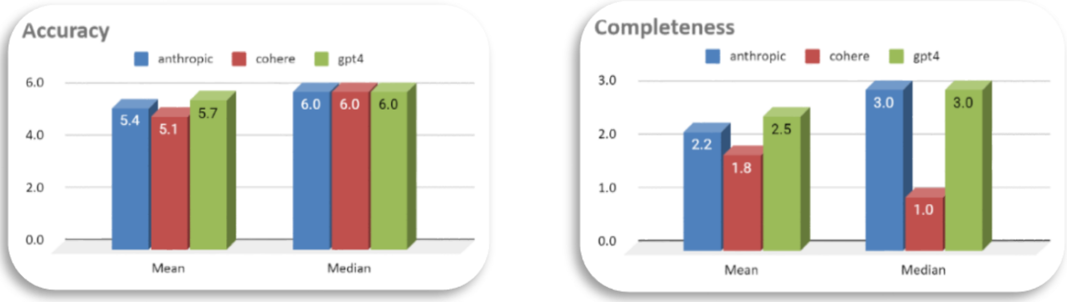


Figure 13 shows the Mean for Orthopaedics including all the parameters of the research

In Orthopaedics GPT 4 has shown the highest performance with a mean accuracy score of 5.7 and an average word count of 1201 words.

14) **ENT: -**

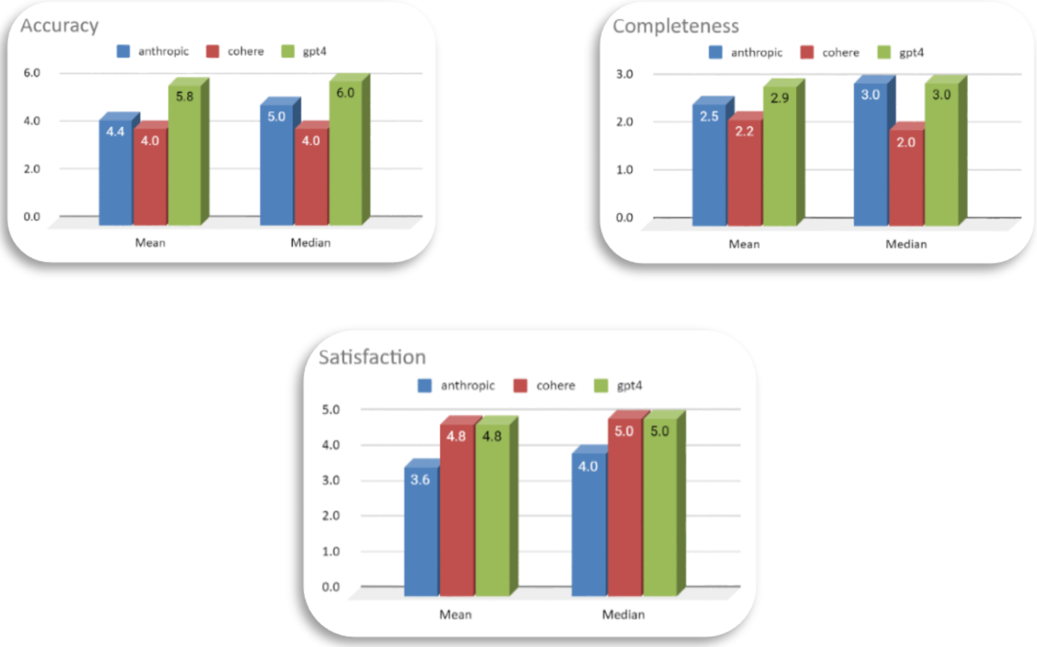


Figure 14 shows the Mean for ENT including all the parameters of the research

In ENT GPT 4 has shown the highest performance with a mean accuracy score of 5 and an average word count of 1290 words.

15) **Emergency: -**

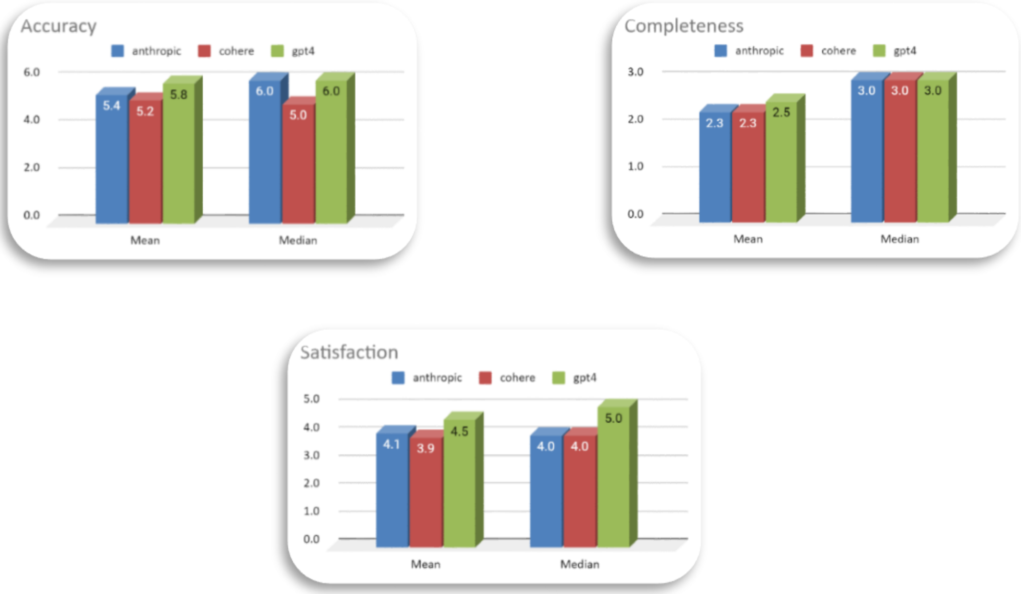


Figure 15 shows the Mean for Emergency including all the parameters of the research

Emergency GPT 4 has shown the highest performance with a mean accuracy score of 5.8 and an average word count of 1186 words.

16) **Cardiology: -**

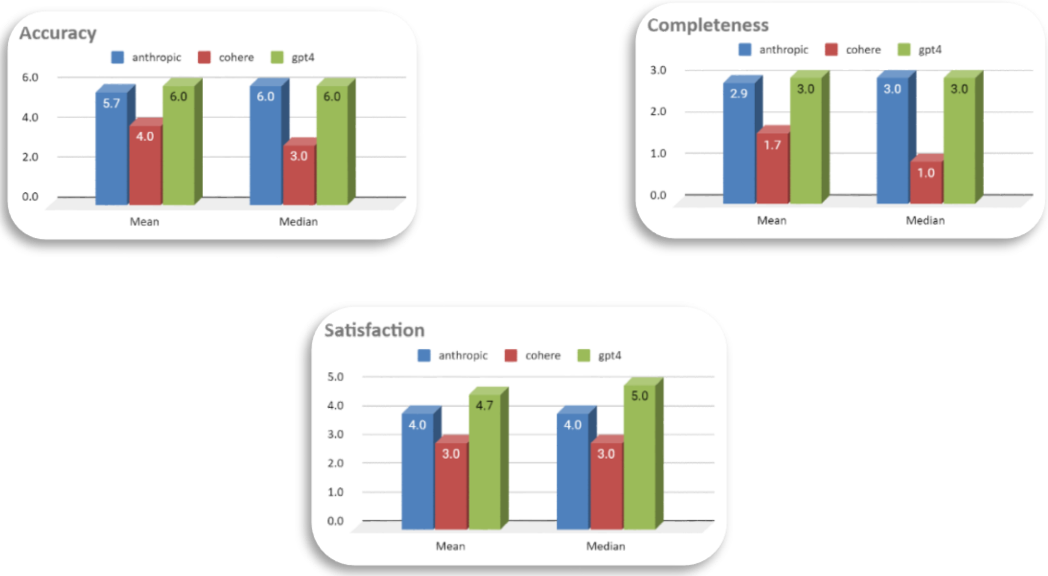


Figure 16 shows the Mean for Cardiology including all the parameters of the research

In Cardiology GPT 4 has shown the highest performance with a mean accuracy score of 6 and an average word count of 1089 words.

The above graphs show a department or specialty-wise performance of the large language models on different sets of questions.

Overall Performance of the Large Language Models:

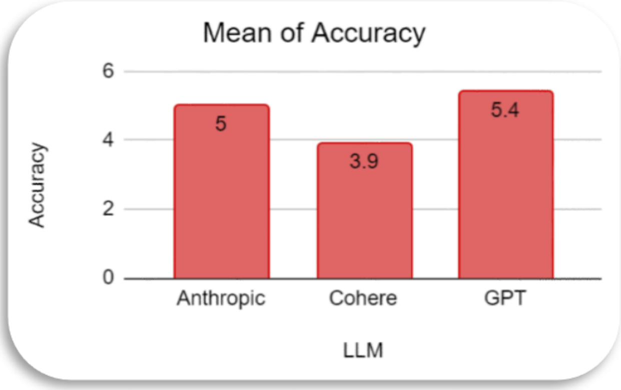


Figure 17 Showing the Accuracy of Anthropic, Cohere, and GPT4

As shown in Fig.17 the accuracy of GPT4 is 5.4 which represents the highest accuracy among all the three large language models and the lowest is performed by the Cohere model which means that the most accurate medical responses were generated by GPT.

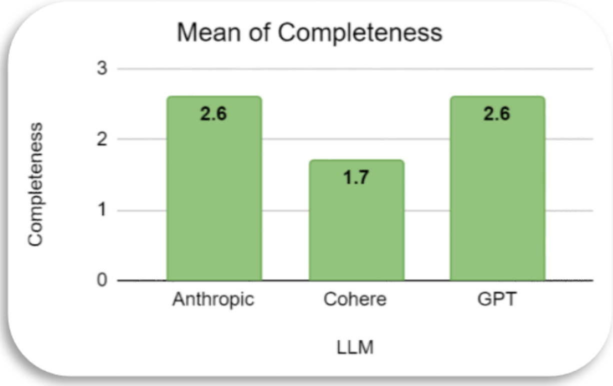


Figure18 Mean of Completeness of Anthropic, Cohere, and GPT4

Fig.18 represents the mean of completeness of all three models thereby, showing GPT4 and Anthropic at 2.6 and Cohere at 1.7. Completeness of the medical responses is determined by the Up-to-date system and other published studies. Though, GPT has performed well in all the parameters even anthropic-generated medical responses gave an equal contesting mean.

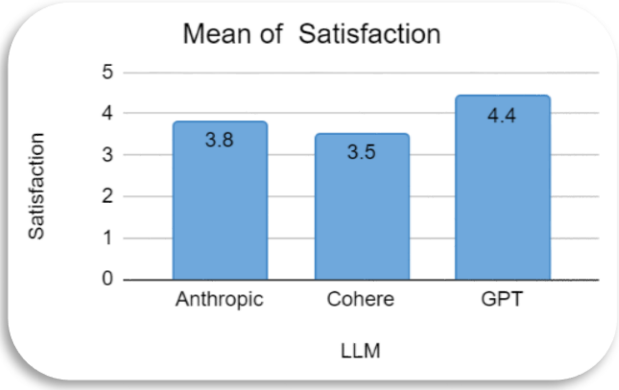


Figure 19 Mean of Satisfaction of Anthropic, Cohere, and GPT4

Fig 19 shows the mean of satisfaction of the three large language models scored on the Likert scale from the range of 1 to 5 scored by the healthcare professionals according to the level of satisfaction of the medical responses.

In Urology, Obstetrics, and Gynaecology the large language models-Anthropic has surpassed the other models.

Obstetrics and Gynaecology:

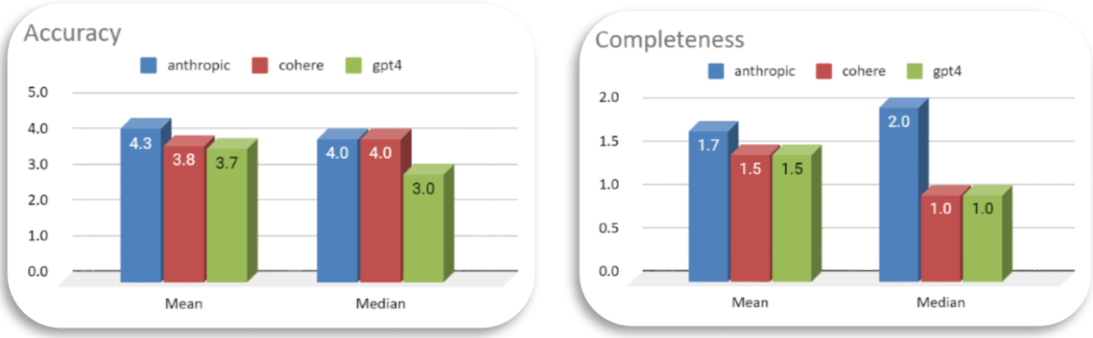


Figure 20 Shows the Accuracy and Completeness

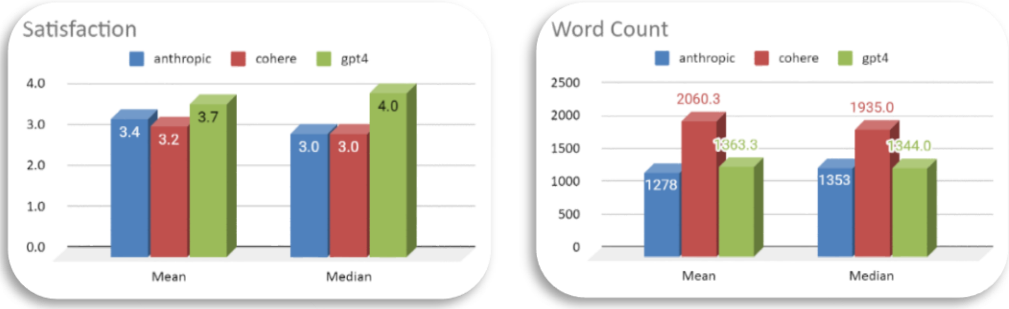


Figure 21 Showing the Satisfaction and Word Count

Fig. 20 and 21 show that the Anthropic model has surpassed GPT 4 giving a higher mean for the mentioned parameters.

Also, Anthropic has shown equal accuracy scores to GPT 4 in some subjects but it has provided a higher average word count than GPT in those subjects.

From all the above-calculated results it has been concluded that GPT4 has overall performed better than Anthropic and Cohere in almost 16 specialties out of 18.

Although, Anthropic has also generated accurate levels of medical responses and completeness in the answers due to more word count of the medical responses the comprehensive of the responses generated were more extended.

Cohere performance was poor in all department in all departments as the model gives less accurate answers with more words.

Key Findings:

- In 16 of the 18 categories of medicine, GPT-4 fared better than the other two models (Anthropic and Cohere).
- Comparing Anthropic to GPT-4 and Cohere, Anthropic offered superior accuracy in obstetrics, gynaecology, and urology.

Accuracy

- With a mean accuracy score of 5.4, GPT-4 achieved the greatest accuracy scores out of all the models, suggesting the most correct medical responses.
- Cohere's accuracy was consistently poorer in every specialty.

Completeness

- With mean scores of 2.6, both GPT-4 and Anthropic demonstrated a high level of completeness in their medical responses.
- Cohere's mean completeness score was 1.7, which was lower than the others.

Satisfaction

- A Likert scale was implemented to assess levels of satisfaction, and GPT-4 once more received the highest score, indicating higher levels of satisfaction among medical professionals.

- Anthropic came in second, however, it was noticed that its responses were longer.

  Word Count

- The average word count a person can read silently is 238 words per minute according to (How many Words Do We read per minute? A review and Meta-analysis of reading rate a 2019 Study https://www.researchgate.net/publication/332380784_How_many_words_do_we_read_per_minute_A_review_and_meta-analysis_of_reading_rate from the 3 models GPT4 has given the average of 1344 which the least from among the 3.

- GPT-4 balanced completeness with brevity by providing succinct yet thorough responses.
- According to the expert opinion the criteria of assessment for the study validates that the GPT4 model gives a better medical response than Anthropic and Cohere, though they also agree that Anthropic can still perform better after training the model.

Section -5

Discussion:

1) Practical and Ethical Impacts:

GPT-4 has great accuracy and completeness, which makes it a useful tool for improving patient care and assisting medical professionals.

Strict usage policies and controls are required to stop abuse, which includes spreading false information and participating in unethical activities like deepfakes.

2) Model Restrictions:

Even while GPT-4 performed better, the study notes that LLMs still need to be continuously improved to overcome issues like context awareness and specialty-specific adaptability.

It may be possible to balance verbosity and informativeness by optimizing Anthropic's propensity for longer responses.

3) Integration with Medical Practices:

To successfully incorporate LLMs such as GPT-4 into clinical practice, issues of patient safety, data privacy, and ethical use must be resolved.

Creating explicit guidelines for the employment of LLMs in the medical profession and providing them with appropriate training are among the recommendations.

4) Prospective Studies:

It is advised that more studies be done to examine how well LLMs work in actual medical situations and assess how they affect patient outcomes.

Further research needs to concentrate on strengthening LLMs' comprehension of complex medical circumstances and their capacity to produce succinct but thorough solutions.

All things considered, the study offers a comprehensive assessment of the capacities of several LLMs in producing medicinal reactions, identifying GPT-4 as the best model and detailing the procedures that must be followed to ensure their safe and efficient application in the medical field.

Section -6

Conclusion:

1) Superior GPT-4 Performance:

- Accuracy: Across most disciplines, GPT-4 showed the highest accuracy in medical responses. It outperformed both Anthropic and Cohere, with a mean accuracy score of 5.4 on a 6-point Likert scale, demonstrating its remarkable ability to produce accurate and precise medical information.
- Completeness: GPT-4 scored highly on measures of completeness, demonstrating their exceptional ability to provide thorough answers. This guarantees that the material produced is comprehensive and addresses every facet of medical inquiries.
- Satisfaction: The GPT-4 demonstrated the highest degree of satisfaction among healthcare professionals, indicating its efficaciousness in fulfilling users' requirements and expectations. This measure demonstrates the GPT-4's acceptability and practicality in a clinical context.

2) Competitive Edge of Anthropic:

- Specialized Strengths: Anthropic performed competitively, especially in the fields of obstetrics, gynaecology, and urology. In these domains, it yielded remarkably precise and pertinent answers, indicating its possible areas of expertise.
- Response Length: Although Anthropic's answers were frequently longer and more thorough, they occasionally provided additional depth, which is advantageous in complicated medical situations. To preserve readability and user interest, this verbosity must be controlled.

3) Underperformance of Cohere:

- Reduced Accuracy and Completeness: Cohere's accuracy and completeness were below par. Compared to GPT-4 and Anthropic, its responses were less thorough and accurate, which made it less appropriate for essential medical applications.
- Satisfaction Levels: The lower user satisfaction with Cohere suggests that considerable model modifications are required before it can be used in medicine.

4) Implications for Medical Practice:

- Improving Clinical Decision Support: GPT-4 can be a useful tool for clinical decision support, providing medical practitioners with correct information and recommendations, as indicated by the high accuracy and completeness of its responses.

- Patient Education: GPT-4 can be used to ensure that patients obtain accurate information about their problems and treatments by providing clear and concise information.

5) Ethical and Practical Considerations:

- Ethical Usage: The study emphasizes how crucial ethical standards and supervision are while using LLMs. The models must be continuously monitored and improved to reduce the possibility of producing false information.
- Data Security and Privacy: When incorporating LLMs into healthcare, it is essential to guarantee the security and privacy of patient data. This entails abiding by stringent ethical and data protection guidelines.

  To sum everything up, GPT-4 stands out as the top LLM for producing medical replies since it provides excellent accuracy, comprehensiveness, and satisfaction. Even if Cohere needs a lot of work and Anthropic has some promising areas, the study indicates how LLMs can revolutionize the healthcare industry. However, resolving ethical, privacy, and practical issues through focused suggestions and continued study will be necessary for their successful integration.

Prospective Routes for Research:

- Testing in the Real World: The study recommends testing LLMs in clinical settings to confirm their effectiveness and influence on patient outcomes.
- Enhancing Contextual Understanding: To make sure LLMs can properly answer intricate and nuanced medical concerns, future research should concentrate on improving LLMs' contextual understanding.
- Long-Term Impact: To fully grasp the potential and constraints of LLM integration, it will be imperative to look at the long-term effects on patient satisfaction and healthcare delivery.
- Specialized Training: To improve the efficacy and dependability of LLMs, the study suggests that they be further specialized and fine-tuned for medical specialties.
- Clinical standards: To ensure that LLMs are used in a way that supports rather than impedes medical treatment, it is important to establish defined clinical standards for their usage in healthcare settings.

References:

1. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model [Internet]. 2023 Feb 28; Available from: https://assets.researchsquare.com/files/rs-2566942/v1/5c64b009-ab48-47a7-bd66-afc5c46d97af.pdf?c=1677623849

2.Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology. 2023 Jan 26;307(2).

3.Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. Dagan A, editor. PLOS Digital Health. 2023 Feb 9;2(2):e0000198.

4. Ray PP. ChatGPT: a Comprehensive Review on background, applications, Key challenges, bias, ethics, Limitations and Future Scope. Internet of Things and Cyber-Physical Systems [Internet]. 2023 Apr;3(1):121–54. Available from: https://www.sciencedirect.com/science/article/pii/S266734522300024X

5. Navigating the LLM Landscape: A Comparative Analysis of Leading Large Language Models [Internet]. DEV Community. 2023 [cited 2024 Jun 11]. Available from: https://dev.to/mindsdb/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models-1ocn?comments_sort=oldest

6. Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. Journal of Medical Internet Research [Internet]. 2023 Dec 28;25(1):e51580. Available from: https://www.jmir.org/2023/1/e51580/

7. Yaara Artsi, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC medical education. 2024 Mar 29;24(1).

8. Borger JG, Ng AP, Anderton H, Ashdown GW, M. Elaine Auld, Blewitt ME, et al. Artificial intelligence takes center stage: exploring the capabilities and implications of

ChatGPT and other AI-assisted technologies in scientific research and education. Immunology and Cell Biology. 2023 Sep 18;

9. Fournier A, C. Fallet, F. Sadeghipour, N. Perrottet. Assessing the Applicability and Appropriateness of ChatGPT in Answering Clinical Pharmacy Questions. Annales Pharmaceutiques Françaises. 2023 Nov 1;

10. Request Rejected [Internet]. ieeexplore.ieee.org. Available from: https://ieeexplore.ieee.org/abstract/document/10221755

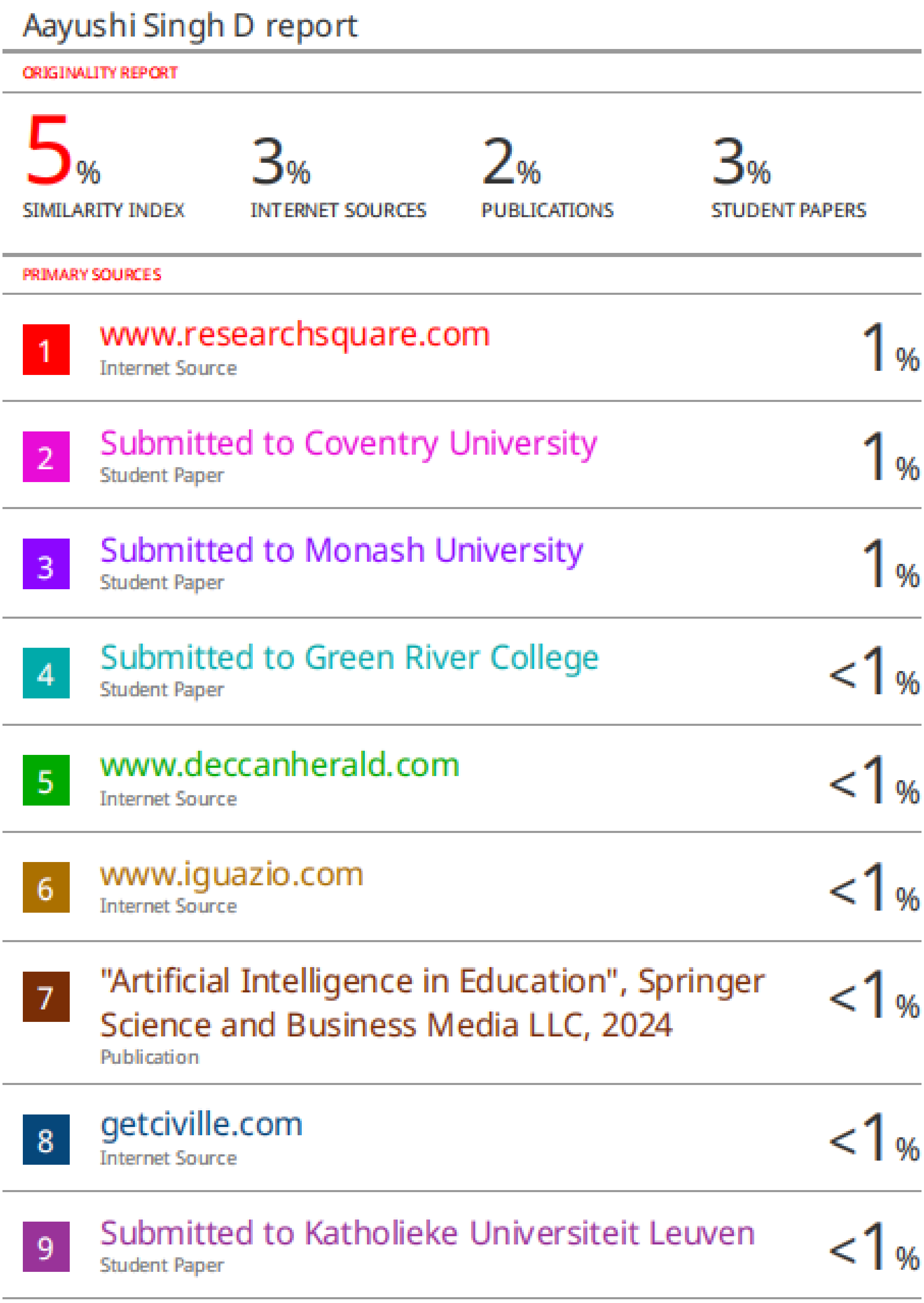
11. Wilhelm TI, Roos J, Kaczmarczyk R. Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study. Journal of Medical Internet Research [Internet]. 2023 Oct 30 [cited 2024 Jan 22];25(1):e49324. Available from: https://www.jmir.org/2023/1/e49324/

12. Bernstein IA, Zhang Y, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. JAMA network open. 2023 Aug 22;6(8):e2330320–0.

13. Mu X, Lim B, Seth I, Xie Y, Jevan Cevik, Foti Sofiadellis, et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. Skin health and disease. 2023 Nov 28;

14. Adi Lahat, Shachar E, Avidan B, Glicksberg BS, Klang E. Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? Diagnostics. 2023 Jun 2;13(11):1950–0.

15. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A Literature Review and Framework for Human Evaluation of Generative Large Language Models in Healthcare [Internet]. arXiv.org. 2024. Available from: https://arxiv.org/abs/2405.02559



Aayushi Singh D report

ORIGINALITY REPORT

5%	3%	2%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchsquare.com Internet Source	1%
2	Submitted to Coventry University Student Paper	1%
3	Submitted to Monash University Student Paper	1%
4	Submitted to Green River College Student Paper	<1%
5	www.deccanherald.com Internet Source	<1%
6	www.iguazio.com Internet Source	<1%
7	"Artificial Intelligence in Education", Springer Science and Business Media LLC, 2024 Publication	<1%
8	getciville.com Internet Source	<1%
9	Submitted to Katholieke Universiteit Leuven Student Paper	<1%