LLM BASED CONSULTATION AND TRIAGE SYSTEM FOR APPOINTMENT SCHEDULING

BY

AFOLABI OLAJIDE SAMUEL
(IFS/19/0592)

SUBMITTED TO

DEPARTMENT OF INFORMATION SYSTEMS,

SCHOOL OF COMPUTING,

FEDERAL UNIVERSITY OF TECHNOLOGY, AKURE.

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF A BACHELOR OF TECHNOLOGY (B. TECH) DEGREE IN INFORMATION SYSTEMS

May, 2025

CERTIFICATION

This is to certify that this project was carried out by **AFOLABI OLAJIDE SAMUEL** with matric number IFS/19/0592 and submitted to the Department of Information Systems, school of computing, Federal University of Technology Akure, Ondo State.

……. …….

AFOLABI OLAJIDE SAMUEL **DATE**

(IFS/19/0592)

This is to certify that this project report was carried out by **AFOLABI OLAJIDE SAMUEL** with matriculation number IFS/19/0592 of the Department of Information Systems, school of computing, Federal University of Technology Akure, Ondo State.

…….. …….

DR. (MRS.) S. P. AKINRINWA **DATE**

Project Supervisor

ACKNOWLEDGMENT

I would like to God almighty all the glory for His grace upon my life to successfully complete my bachelor's degree in The Federal University of Technology, Akure. God has been there for me all through my undergraduate journey, supplying all my needs in Christ Jesus. I would like to express my deepest gratitude to my supervisor, Dr. (Mrs.) S. P. Akinrinwa, for her invaluable guidance, support, and encouragement throughout this project. Her expertise and insightful feedback were instrumental in shaping the direction and completion of this work. I am truly grateful for her patience and dedication, which have greatly contributed to my academic growth and success. I would also like to express my heartfelt appreciation to my Head of Department, Prof. Olufemi Akinyede, and the entire staff of the Department of Information Systems for their unwavering support and encouragement in my pursuit of academic excellence. Their valuable feedback and guidance have played a key role in the successful completion of this project.

I would also like to acknowledge my parents, Mr. Olajide Sunday Afolabi and Mrs. Pricilla Afolabi, for their unwavering support and encouragement throughout my academic journey. Their belief in my abilities has been a constant source of motivation. I am truly thankful for their understanding and for providing me with the environment and support needed to complete this project successfully. My gratitude also goes to my sister, Abimbola Afolabi.

Lastly, I would also like to extend my heartfelt thanks to my colleagues and friends for their constant support and for providing helpful suggestions. Their feedback was critical in refining the user experience and ensuring that the project met its intended goals. Worthy of note are ….

DEDICATION

This project is dedicated to the Almighty God, the giver and custodian of all grace, mercy, knowledge and wisdom who has helped me through my journey in the Federal University of Technology, Akure.

ABSTRACT

The healthcare sector faces persistent challenges in patient consultation, triage, and appointment scheduling, including prolonged wait times, and inefficient resource allocation, exacerbated by manual processes and limited scalability. This paper proposes an intelligent healthcare management platform leveraging Large Language Models (LLMs) to enhance outpatient care delivery across multiple healthcare organizations. The system automates initial patient assessments, prioritizes cases based on clinical urgency, and optimizes appointment scheduling by processing unstructured symptom narratives in real-time. Designed using Next.js and powered by LLMs such as Palmyra-med-70b and OpenAI's GPT model, the platform integrates a modular architecture comprising patient interaction, LLM processing, and scheduling components. Real-time data collection during consultations, coupled with advanced natural language processing, enables accurate triage and dynamic appointment allocation. The system is evaluated using standard metrics, including triage accuracy, waiting time reduction, and user satisfaction, targeting significant improvements over traditional methods. Expected contributions include enhanced triage precision, reduced no-show rates, and a scalable framework for multi-entity healthcare coordination. By addressing inefficiencies and ethical challenges, such as data privacy and algorithmic bias, this research aims to foster a patient-centered, equitable approach to healthcare delivery, with implications for operational efficiency and care quality.

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER ONE

INTRODUCTION

1.0 Background of the Study

The healthcare sector is on the brink of a transformative change, driven by artificial intelligence (AI) and its sophisticated variant, large language models (LLMs), which are expected to enhance patient care and streamline operational efficiency across diverse settings. This paper proposes an intelligent healthcare management platform that harnesses LLMs to transform patient consultation and triage systems, with a central focus on optimizing appointment scheduling across multiple healthcare organizations. The urgency for such innovation is palpable amid escalating challenges: emergency departments (EDs) grapple with surging patient volumes and resource constraints, particularly during pandemics (Garrido *et al*., 2024, Preiksaitis *et al*., 2024), while outpatient departments (OPDs) face inefficiencies such as disordered information flow and prolonged wait times, stifling effective service delivery (Nwankwo *et al*., 2023, Ramdurai, 2025).

A 2018 Beryl Institute study highlights that 91% of patients prioritize patient experience, with 88% willing to switch providers due to poor encounters (Ramdurai, 2025). Compounding these issues, missed appointments or "no-shows" disrupt resource allocation, with 31% of patients canceling late or failing to attend, intensifying delays and administrative burdens (Komarneni *et al*., 2025). Platforms like Doctolib demonstrate that AI-driven solutions can reduce no-show rates by 25% through automated reminders and real-time scheduling adjustments, offering a model for tackling such inefficiencies (Deepa P. *et al.*, 2024). Moreover, inefficiencies in outpatient clinics often stem from a mismatch between demand and capacity, exacerbated by heterogeneous service times (i.e., consultation durations), which traditional scheduling methods struggle to address (Feng *et al*., 2024).

Patient experience, encompassing ease of scheduling, reduced wait times, and timely information access, is a cornerstone of healthcare satisfaction (Ramdurai, 2025). Surveys of over 230 providers reveal that 50% of patients value shorter wait times, and 41% seek simpler booking processes (Ramdurai, 2025). Traditional systems, reliant on manual processes and phone-based communication, falter under these demands, contributing to inefficiencies and clinician overload

(Nwankwo *et al*., 2023, Komarneni *et al*., 2025). In emergency contexts, AI-driven tools like the extreme gradient boosting (XGB) algorithm demonstrate superior performance, achieving 91.61% balanced accuracy in predicting mortality during the COVID-19 pandemic, surpassing traditional triage systems in speed and precision (Garrido *et al*., 2024). Similarly, LLMs enhance ED triage by synthesizing electronic health records (EHRs) in real time, alleviating cognitive burdens (Preiksaitis *et al*., 2024), while in OPDs, NLP techniques like the Word Rank algorithm summarize patient records, streamlining consultation workflows (Nwankwo *et al*., 2023). Predictive models leveraging datasets like the Medical Appointment No Shows (110,527 visits) further mitigate no-shows by enabling proactive scheduling adjustments (Komarneni *et al*., 2025), with Doctolib's predictive analytics reducing cancellations by 30% through demand forecasting (Deepa P. *et al*., 2024). Complementing these efforts, data-driven approaches like the Cluster-Predict-Schedule (CPS) system use machine learning to group patients by service time heterogeneity, achieving up to 15% cost reductions over first-call, first-appointment (FCFA) methods and 4.7% savings over New/Return classifications, enhancing outpatient efficiency without sacrificing fairness (Feng *et al*., 2024). These advancements align with a digital transformation, with 80% of healthcare CIOs prioritizing AI solutions to elevate patient experience (Ramdurai, 2025).

LLMs, powered by transformer architectures, excel at processing vast unstructured data, offering a versatile foundation for healthcare innovation (Preiksaitis *et al*., 2024). In pandemics, XGB-driven triage systems rapidly identify high-risk patients using variables like procalcitonin, age, and oxygen saturation, optimizing resource allocation and blocking transmission chains (Garrido *et al*., 2024). In outpatient settings, LLMs integrate speech-to-text capabilities, real-time availability updates, and personalized reminders, enhancing patient-provider communication (Nwankwo *et al*., 2023). Doctolib exemplifies this potential, boosting telemedicine usage by 40% and improving patient satisfaction by 30% through secure video consultations and 24/7 AI chatbot support (Deepa P. *et al*., 2024). Similarly, CPS leverages unsupervised and supervised machine learning to adaptively cluster patients and generate cost-minimizing appointment templates, balancing implementation simplicity with the complexity of individual service time variations (Feng *et al*., 2024). This dual functionality reduces wait times, improves decision-making, and maximizes resource use across EDs and OPDs (Komarneni *et al*., 2025). Yet,

challenges such as output reliability, data privacy, and ethical integration evident in both emergency and outpatient applications demand careful navigation (Preiksaitis *et al*., 2024, Nwankwo *et al*., 2023). Moreover, the global unpreparedness for future pandemics, as noted by the World Health Organization, underscores the need for scalable AI tools (Garrido *et al*., 2024).

This paper explores how an LLM-based platform can bridge these gaps, delivering an equitable, efficient solution that enhances accessibility, accuracy, and care quality across healthcare ecosystems. It also aims to develop an LLM based consultation and triage system specifically designed for intelligent appointment scheduling. The proposed system will leverage large language models to automate initial patient assessments, prioritize cases based on clinical urgency, and optimize appointment allocation. By doing so, the project seeks to establish a more responsive, efficient, and patient-centered approach to outpatient healthcare services.

1.1 Project Motivation

There is a need to develop an intelligent healthcare management platform that leverages Large Language Models (LLMs) to enhance patient consultation and triage processes, optimizing appointment scheduling across multiple healthcare organizations. This platform will address the inefficiencies of traditional scheduling systems, the limited integration of real-time consultation and triage capabilities, and the lack of scalability in existing AI-driven healthcare models, which often fail to process unstructured patient data or coordinate across diverse healthcare entities.

The limitations of Kumar and Sulaiman (2024), Deepa *et al*. (2024), Feng *et al*. (2024), Taylor *et al*. (2024), Akinode and Oloruntoba (2017), Garrido *et al*. (2024), Komarneni *et al*. (n.d.), and Nwankwo *et al*. (2023) are the key motivations for this research work. These include:

I. Failure to incorporate advanced LLM-driven consultation and triage capabilities, relying instead on structured inputs or basic NLP, which limits adaptability to diverse patient needs and real-time prioritization (Kumar and Sulaiman, 2024, Deepa *et al*., 2024, Nwankwo *et al*., 2023).

II. Inability to provide a scalable model that coordinates appointment scheduling across multiple healthcare organizations, restricting applications to single-facility contexts (Feng *et al*., 2024, Taylor *et al*., 2024, Komarneni *et al*., n.d.).

III. Lack of integration between consultation data processing and dynamic triage or scheduling optimization, leaving gaps in proactive patient care and resource allocation (Akinode and Oloruntoba, 2017, Garrido *et al*., 2024).

IV. Insufficient use of unstructured patient inputs (e.g., symptom narratives) for triage and scheduling, limiting the ability to enhance accuracy and responsiveness beyond predefined workflows or historical data (Deepa *et al*., 2024; Kumar and Sulaiman, 2024, Nwankwo *et al*., 2023).

Hence, this research aims to provide a comprehensive solution to the challenges of patient consultation, triage, and appointment scheduling, enhancing healthcare efficiency, accessibility, and scalability through an LLM-based approach.

1.2 Aim and Objectives

The aim of this research work is to develop an intelligent healthcare management platform that leverages artificial intelligence by using LLMs to enhance patient consultation and triage systems to optimize appointment scheduling across multiple healthcare organizations.

Specific objectives include:

a) Design an LLM based consultation and triage system for appointment scheduling.

b) Implement the system in (a) using Next.js, and also using LLMs like Palmyra-med-70b and OpenAI gpt model for AI functionalities.

c) Evaluate the system in (b) using standard evaluation metrics.

1.3 Methodology

The methodology for developing this LLM-Based Consultation and Triage System for Appointment Scheduling outlines a systematic approach to designing, implementing, and evaluating an intelligent healthcare platform that leverages Large Language Models (LLMs) to optimize patient consultation, triage, and appointment scheduling across multiple healthcare

organizations. It features a System Design that describes a modular architecture with Patient Interface, LLM Processing, and Doctor Interface Modules, formalizing the application flow patient registration, symptom input, follow-up questions, triage assignment, and appointment booking. The Technology Stack and Tools include Next.js for the frontend, Node.js with MongoDB for the backend, and LLMs (Palmyra-med-70b and OpenAI's GPT model) for AI functionalities, and AWS for scalability. Data Collection occurs in real-time during consultations, capturing symptom narratives and responses, stored as anonymized JSON objects, and supplemented by synthetic datasets. The Implementation Process follows an Agile methodology, with phases for, system development, testing, and AWS deployment using Docker. The Evaluation Strategy employs usability testing (User Acceptance Testing, System Usability Scale), performance testing (response time, throughput), end-to-end testing (error rate, data integrity), and LLM-specific evaluations (triage accuracy, embedding quality), benchmarking against systems like Doctolib and XIAO YI. This methodology ensures a scalable, patient-centered solution, addressing technical and ethical challenges for enhanced healthcare delivery.

1.4 Expected Contribution to Knowledge

At the end of this research, a LLM based consultation and triage system for appointment scheduling will be developed. This system will enhance healthcare management by automating consultation and triage, improving triage accuracy, optimizing scheduling efficiency, and providing a scalable framework for multi-entity healthcare delivery.

1.5 Organization of the Project

This project report is organized into five chapters, each focusing on specific aspects of the research and development process. Chapter one which is the introduction provides the background of the study, project motivation, objectives, scope, and organization of the project. Chapter Two takes a deep dive into reviewing the existing body of knowledge related to LLM based consultation and triage systems for appointment scheduling. It covers the theoretical foundations, existing systems, and relevant research findings. Chapter Three focuses on the

System Design and Methodology which outlines the system architecture, components, and data flow. It details the design choices and implementation strategies for the various modules of the system. Chapter Four which is the implementation and Results describes the implementation process, including the selection of programming languages, frameworks, and tools. It presents the results of the system evaluation, including performance metrics and user feedback. Chapter Five summarizes the key findings of the project, discusses the limitations, and provides recommendations for future research and development.

CHAPTER TWO

LITRARURE REVIEW

1.0 Overview

This literature review explores the integration of Large Language Models (LLMs) with healthcare consultation, triage, and appointment scheduling systems, emphasizing their theoretical foundations and practical applications in optimizing healthcare workflows. The review synthesizes current research to inform the development of LLM-based solutions that enhance efficiency, patient satisfaction, and care equity in outpatient settings.

This review commences by examining the current state of the outpatient care ecosystem, focusing on inefficiencies such as patient no-shows, prolonged wait times, and scheduling challenges that disrupt resource allocation and compromise care quality. These limitations significantly impact patient experience and healthcare delivery, necessitating innovative solutions. It then traces the evolution of digital technologies in healthcare, highlighting the progression from early rule-based systems to advanced natural language processing (NLP) frameworks that have shaped modern LLMs (Alowais et al., 2023; Vuong, 2024).

Recent advancements in LLM capabilities are critically evaluated, with a focus on their transformative applications in clinical settings. The review synthesizes empirical evidence from implementation studies, such as the XIAO YI system, which leverages NLP to streamline consultation, triage, and scheduling, reducing wait times and costs (Li et al., 2021). LLMs demonstrate proficiency in tasks like medical history taking, preliminary diagnosis, and predictive scheduling, enhancing accuracy, reliability, and user acceptance (Zhakhina et al., 2023; Borkowski & Ben-Ari, 2024). Particular attention is given to their role in facilitating patient communication and supporting clinical decision-making through real-time data analysis (Mumtaz et al., 2025; Shalko et al., 2024).

This review also addresses critical challenges in implementing LLM-based healthcare systems. Ethical concerns, including data privacy, algorithmic bias, and transparency, demand robust governance frameworks to ensure patient safety and trust (Alowais et al., 2023; Preiksaitis et al., 2024). Technical barriers, such as interoperability with electronic health record (EHR) systems

and reliance on high-quality datasets, pose significant hurdles (Adler-Milstein et al., 2022; Borkowski & Ben-Ari, 2024). Additionally, practical considerations, including the digital divide and clinician acceptance, highlight the need for inclusive design and comprehensive training (Zhakhina et al., 2023; Vuong, 2024).

By synthesizing these interconnected areas, this review establishes a comprehensive knowledge base to guide the design, implementation, and evaluation of LLM-based consultation and triage systems for appointment scheduling. It underscores the potential of LLMs to revolutionize healthcare delivery while identifying key areas for future research to address ethical, technical, and practical challenges.

2.1.1 The Outpatient Care Ecosystem: Consultation, Triage, and Scheduling

The outpatient care ecosystem integrates consultation, triage, and scheduling processes, each critical for optimizing healthcare delivery and resource allocation. Consultation and triage serve as pivotal entry points, where accurate assessment of patient needs determines the urgency and type of care required, directly influencing scheduling efficiency. However, inefficiencies such as patient no-shows ranging from 3% to 80% depending on service type and demographics disrupt this balance, increasing costs and compromising care quality (Toker *et al*., 2024). Traditional strategies like overbooking and phone reminders often fail to address root causes of no-shows, including demographic factors, prior appointment patterns, or logistical barriers (Toker *et al*., 2024). Moreover, scheduling challenges are exacerbated by limited provider availability and patient access issues, particularly for under-served populations, which can lead to prolonged wait times and reduced patient satisfaction (Woodcock, 2022). Patient surveys indicate that 65% rate clinic wait times as unsatisfactory, with 50% prioritizing shorter wait times as a key factor for improving their experience (Ramdurai, 2020).

Advancements in artificial intelligence (AI) have introduced transformative solutions to these challenges. Machine learning models, such as decision trees and regression analysis, utilize patient demographics and historical data to predict no-show behavior, enabling proactive scheduling adjustments (Toker et al., 2024). AI-assisted systems like XIAO YI, implemented at Shanghai Children's Medical Center, enhance consultation and triage by employing natural language processing (NLP) to extract clinical features from electronic medical records and

recommend preliminary tests or imaging before physician consultation (Li et al., 2021). This system reduced median waiting times from 1.97 hours in conventional workflows to 0.38 hours in AI-assisted workflows, demonstrating improved patient flow and resource utilization (Li et al., 2021). Additionally, XIAO YI's integration with mobile platforms like WeChat enables patients to complete pre-consultation steps remotely, reducing in-hospital congestion and enhancing accessibility (Li et al., 2021). AI-driven predictive analytics further optimize patient flow by forecasting wait times and appointment delays, allowing providers to dynamically adjust schedules and prioritize patients with special needs, such as senior citizens or those with urgent conditions, thereby enhancing care equity and satisfaction (Ramdurai, 2020).

The adoption of automated patient self-scheduling further addresses these inefficiencies by empowering patients to book appointments directly via web or mobile platforms, bypassing traditional phone-based systems. Woodcock (2022) highlights that self-scheduling offers significant advantages, including labor savings, improved patient satisfaction, and reduced no-show rates, with studies reporting 34% to 51% of appointments booked outside office hours, optimizing early morning slots and staff resources. However, barriers to implementation persist, particularly in the outer setting, such as disparities in technology access for rural or low-socioeconomic-status patients, and in the inner setting, where physician concerns about losing control over schedules and the complexity of integrating self-scheduling with existing systems pose challenges (Woodcock, 2022). Ramdurai (2020) emphasizes that ease of appointment booking and check-in processes, such as AI-powered kiosks that streamline patient identification and queue management, are critical to reducing patient frustration and improving experience. These systems can also provide navigational guidance within large facilities, addressing a commonly overlooked factor that impacts patient satisfaction (Ramdurai, 2020).

Large language models (LLMs) amplify these capabilities by enabling dynamic, data-driven appointment management. By analyzing patient complaints and medical histories, LLMs can enhance triage accuracy and personalize scheduling recommendations, addressing both no-show risks and prolonged waiting times. The cost-effectiveness of such systems is evident, as XIAO YI not only shortened waiting times but also reduced total outpatient costs from 364.58 CNY to 335.97 CNY by prioritizing non-invasive, low-cost tests (Li *et al.*, 2021). Despite these advancements, Woodcock (2022) notes that organizational uptake remains low due to gaps in

understanding inner setting dynamics, such as physician resistance and implementation processes. Ramdurai (2020) adds that the infancy of AI application in healthcare necessitates further research to refine algorithms and validate their impact on large datasets, particularly for predictive wait-time models and resource optimization. AI-driven innovations, combined with self-scheduling, thus hold immense potential to mitigate the multifaceted challenges of outpatient care, enhancing capacity utilization, patient satisfaction, and financial efficiency in healthcare systems

2.1.2 Evolution of Digital Technologies in Healthcare

Historical Foundations (1950s–1990s)

The journey of digital technologies in healthcare began in the 1950s with exploratory AI efforts, such as Christopher Strachey's 1951 AI program and the formalization of the term "Artificial Intelligence" at the 1956 Dartmouth Conference (Alowais et al., 2023). Early systems like the Dendral project in the 1960s, which analyzed chemical components, and MYCYIN in the 1970s, designed to diagnose bacterial infections, laid the groundwork for medical decision-making tools (Vuong, 2024; Preiksaitis et al., 2024). The 1980s and 1990s saw a shift toward machine learning (ML) and neural networks, driven by increased computational power, enabling systems like INTERNIST-I and QMR to assist in complex diagnoses (Vuong, 2024; Preiksaitis et al., 2024).

Emergence of Digital Solutions (2000s)

By the 2000s, advancements in natural language processing (NLP) and the advent of LLMs facilitated the development of virtual assistants and chatbots capable of processing unstructured clinical data, improving patient-provider communication, and streamlining administrative tasks (Alowais et al., 2023; Vuong, 2024; Preiksaitis et al., 2024). Concurrently, web-based appointment systems marked a pivotal shift toward patient-centered care, allowing patients to browse and select convenient appointment slots through intuitive, calendar-like interfaces (Zhao et al., 2017). These systems, available in asynchronous and real-time modes, reduced no-show

rates, decreased waiting times from 98 minutes to 7 minutes in some cases, and enhanced patient satisfaction by enabling after-hours access and minimizing staff labor (Zhao et al., 2017).

Modern Advancements with LLMs (2010s–2020s)

The 2010s and 2020s witnessed a transformative leap with LLMs, characterized by their ability to process vast volumes of text with remarkable speed and accuracy (Mumtaz et al., 2025). Trained on extensive datasets, LLMs demonstrate proficiency in understanding and generating human-like text, enabling applications from clinical decision support to patient education (Mumtaz et al., 2025). They have been pivotal in processing complex clinical notes, interpreting medical imaging reports, and facilitating patient-provider communication, thus streamlining workflows and enhancing diagnostic precision (Mumtaz et al., 2025). LLMs also leverage predictive analytics and NLP to optimize resource allocation and streamline triage processes, particularly in emergency medicine settings (Vuong, 2024; Preiksaitis et al., 2024).

Pre-Consultation and Triage Systems

Historically, medical history collection relied on time-consuming face-to-face interviews, limited by incomplete patient recall and physician time constraints, which could compromise diagnostic accuracy (Zhakhina et al., 2023). Digital solutions, such as computerized history-taking programs pioneered by institutions like the Mayo Clinic, enabled structured, patient-driven data collection (Zhakhina et al., 2023). AI and NLP integration has revolutionized these systems, allowing sophisticated analysis of patient-reported data, identification of critical health indicators, and contextually relevant follow-up questions, thereby optimizing triage and consultation efficiency (Zhakhina et al., 2023). Tablet-based pre-consultation systems reduce preclinical and consultation times compared to paper-based methods, while virtual triage systems, enhanced by AI, demonstrate high efficacy, with 75% of users finding them helpful in determining appropriate care levels (Zhakhina et al., 2023).

AI-Assisted Tools

AI-assisted programs like Smart-doctor, leveraging deep learning-driven NLP, emulate clinical reasoning and recommend diagnostic tests, significantly reducing queuing times in pediatric outpatient settings from a median of 21.81 minutes to 8.78 minutes, with a 17.53% increase in patient satisfaction scores (Li et al., 2022). These systems handle repetitive tasks, such as pre-

diagnosis inquiries, allowing healthcare providers to focus on complex cases and improving resource allocation (Li et al., 2022).

Appointment Scheduling

LLMs are poised to revolutionize appointment scheduling by enabling automated, context-aware systems that prioritize patient needs and align appointments with clinical requirements across healthcare organizations (Mumtaz et al., 2025; Zhao et al., 2017). These systems enhance scheduling flexibility by processing complex patient inputs, addressing limitations of traditional telephone or in-person scheduling, such as scheduler availability and phone line constraints (Zhao et al., 2017).

Challenges and Future Directions

Despite these advancements, challenges persist, including data privacy, ethical considerations, cybersecurity, and the digital divide, which necessitate robust measures and equitable access (Mumtaz et al., 2025; Alowais et al., 2023; Zhakhina et al., 2023). Limited system flexibility, safety concerns in triaging urgent conditions, and slow adoption due to provider and patient reluctance highlight the need for intelligent, scalable solutions (Zhao et al., 2017). Seamless integration with existing clinical workflows and ongoing collaboration between healthcare professionals and technologists are critical to ensuring effective implementation (Vuong, 2024; Preiksaitis et al., 2024).

The evolution of digital technologies in healthcare, from rule-based systems to sophisticated AI frameworks and LLMs, has transformed patient care delivery, consultation, triage, and appointment scheduling. These advancements offer scalable solutions to meet the growing demands of healthcare systems, paving the way for efficient, patient-centered platforms. Continued refinement and collaboration will be essential to address challenges and fully realize the potential of these technologies in optimizing healthcare delivery.

2.1.3 LLMs in Healthcare: Current Applications and Trends

Large Language Models (LLMs), a subset of artificial intelligence utilizing deep learning and vast datasets, have emerged as transformative tools in healthcare. By leveraging advanced natural language processing (NLP) capabilities, LLMs enhance clinical workflows, patient care,

and administrative efficiency, particularly in consultation, triage systems, and appointment scheduling. Their ability to process diverse data types text, imaging, and patient narratives supports timely and informed clinical decisions, fostering patient-centered care delivery.

Consultation and Triage Systems

LLMs are increasingly utilized to streamline consultation and triage processes, including diagnostic decision-making, patient triage, and generating patient-friendly reports. In clinical settings, LLMs facilitate human-in-the-loop decision-making, where healthcare providers integrate model-generated insights with clinical expertise to improve triage accuracy and patient outcomes. For instance:

I. **Oncology:** LLMs like ChatGPT have been explored as decision-support tools for breast tumor boards, achieving a 70% concordance rate with expert decisions in treatment recommendations (Mumtaz et al., 2025). In emergency settings, ChatGPT 4.0 demonstrates high sensitivity (95.7%) in triaging patients with metastatic prostate cancer, offering concise and comprehensive diagnoses (Mumtaz et al., 2025).

II. **Dermatology:** Multimodal LLMs like SkinGPT-4 analyze skin disease images, achieving a 78.76% validation rate by dermatologists, addressing specialist shortages and complex image interpretation (Mumtaz et al., 2025).

III. **Mental Health and Nephrology:** Models like Med-PaLM 2 and GPT-4 show high accuracy in assessing psychiatric conditions and answering nephrology-related queries, respectively (Mumtaz et al., 2025).

IV. **Emergency Departments:** LLMs enable real-time triage, prioritizing high-risk cases to reduce wait times and enhance patient flow (Alowais et al., 2023; Preiksaitis et al., 2024).

V. **Digital Assistants:** NLP-driven chatbots, such as the Florence Chatbot used by the UK's National Health Service (NHS) and the Babylon Health Chatbot for the NHS 111 hotline, analyze symptoms, guide care decisions, and reduce the burden on call centers (Shalko et al., 2024; Vuong, 2024).

LLMs also support patient self-triage by analyzing symptoms and guiding care decisions, enabling laypeople to navigate care options. However, their effectiveness in precise medical

triage varies, with specialized symptom-assessment applications (SAAs) outperforming models like ChatGPT, which showed no significant improvement in self-triage accuracy (Kopka et al., 2025).

Administrative Efficiency

LLMs enhance administrative processes by automating clinical documentation, such as generating discharge summaries and interpreting diagnostic studies, reducing administrative burdens (Shalko et al., 2024). Their ability to process unstructured data from electronic health records (EHRs) supports efficient information management, allowing healthcare providers to focus on patient care (Preiksaitis et al., 2024). Additionally, LLMs assist with patient inquiries, prescription reminders, and appointment scheduling, improving operational efficiency and patient engagement (Vuong, 2024).

Medical Education

In medical education, models like GPT-4 and Med-PaLM 2 process large datasets to provide objective evaluations, improving learning outcomes and supporting staff training (Shalko et al., 2024).

Recent trends highlight the growing role of LLMs in healthcare:

i. **Multimodal LLMs:** There is a shift toward integrating visual, textual, and auditory inputs for holistic patient assessments. For example, in dentistry, frameworks combine X-rays, audio, and text for comprehensive diagnostics (Mumtaz et al., 2025). Integration with diverse data sources, such as wearable devices and EHRs, supports personalized healthcare delivery (Shalko et al., 2024).

ii. **Predictive Analytics for Scheduling:** LLMs optimize appointment scheduling by predicting patient urgency and demand, prioritizing urgent cases and improving resource allocation across healthcare organizations (Alowais et al., 2023; Preiksaitis et al., 2024; Vuong, 2024).

iii. **Personalized Medicine and Population Health:** LLMs are increasingly applied to personalized medicine and population health management, enhancing care delivery through tailored interventions and optimized resource use (Alowais et al., 2023).

iv. **Human-Centered Applications:** Ongoing developments focus on human-centered LLM applications, emphasizing integration with clinical expertise and compatibility with existing systems to foster trust and compliance (Shalko et al., 2024; Vuong, 2024).

Challenges

Despite their potential, LLMs face several barriers to widespread adoption:

i. **Diagnostic Accuracy:** Ensuring diagnostic accuracy remains a challenge, particularly for precise medical triage, where LLMs may underperform compared to specialized SAAs (Kopka et al., 2025; Shalko et al., 2024).

ii. **Data Privacy and Ethical Concerns:** Robust cybersecurity measures and ethical frameworks are needed to address data privacy and potential biases (Alowais et al., 2023; Preiksaitis et al., 2024; Vuong, 2024).

iii. **System Integration:** Seamless integration with existing EHR systems and healthcare platforms is critical for effective implementation (Preiksaitis et al., 2024; Vuong, 2024).

iv. **Explainability and Personalization:** The lack of explainability and perceived personalization limits LLMs' autonomous use in clinical decision-making, necessitating human oversight and robust data bases (Kopka et al., 2025; Shalko et al., 2024).

v. **Validation and Safety:** Robust validation is essential to ensure patient safety, as LLMs may not fully tailor responses to individual patient characteristics (Shalko et al., 2024).

LLMs are revolutionizing healthcare by enhancing consultation, triage, and administrative processes, with emerging applications in optimizing appointment scheduling and supporting medical education. Current trends toward multimodal integration, predictive analytics, and human-centered applications underscore their potential to transform care delivery. However, addressing challenges such as diagnostic accuracy, data privacy, ethical concerns, and system integration is critical to ensuring their effective and widespread adoption. As LLMs evolve, their role in creating intelligent healthcare management platforms is poised to foster more efficient, responsive, and patient-centered care delivery.

2.1.4 LLMs for Patient Consultation: History Taking and Preliminary Diagnosis

The integration of Large Language Models (LLMs) into patient consultation processes, particularly for history taking and preliminary diagnosis, represents a transformative advancement in healthcare delivery. LLMs, leveraging their natural language processing (NLP) capabilities, enable efficient and accurate collection of comprehensive medical histories, support preliminary diagnostic assessments, and optimize triage and appointment scheduling. By addressing limitations in traditional methods, such as incomplete data and time constraints, LLMs enhance patient-centered care while necessitating rigorous validation to ensure reliability, accessibility, and ethical implementation.

LLMs integrated into pre-consultation systems allow patients to provide detailed medical histories through structured digital questionnaires, improving the accuracy and completeness of collected data (Zhakhina et al., 2023). These systems interpret patient-reported information, generate contextually relevant follow-up questions, and capture sensitive details, such as smoking or psychosocial issues, often underreported due to patient forgetfulness or embarrassment (Zhakhina et al., 2023). For example, the Smart-doctor system, an AI-based medical assistant, facilitates comprehensive history taking via a mobile interface, capturing chief complaints like cough or diarrhea with high accuracy (Li et al., 2022). This approach reduces the likelihood of omitting critical details, a common challenge in face-to-face consultations due to communication barriers or time constraints, and empowers patients to complete questionnaires at their own pace, fostering engagement and ownership in their healthcare process (Zhakhina et al., 2023).

LLMs enhance preliminary diagnostic assessments by analyzing integrated patient data in real time. The Diagnostic Agent, optimized through fine-tuning on medical corpora, applies clinical criteria to identify patterns indicative of specific conditions, such as sepsis, and adapts to rare or atypical presentations using techniques like few-shot learning (Borkowski & Ben-Ari, 2024). Similarly, the Smart-doctor system demonstrates diagnostic accuracy of 92% for respiratory diseases and 85% for gastrointestinal diseases, recommending appropriate examinations and mimicking clinical reasoning (Li et al., 2022). Studies show that AI-based systems can reduce diagnostic errors to as low as 11%, compared to 15-30% in physician-led assessments, by

prioritizing critical information and classifying patients based on symptoms and medical history (Zhakhina et al., 2023).

LLMs streamline triage by prioritizing cases based on severity and determining appropriate care levels, aligning with recommended care pathways in 75% of virtual triage cases (Zhakhina et al., 2023). The Smart-doctor system significantly reduces queuing times, with a median of 8.78 minutes compared to 21.81 minutes in conventional settings, enhancing patient throughput and satisfaction by 17.53% due to reduced waiting times and increased staff attention (Li et al., 2022). By automating history taking and preliminary assessments, LLMs reduce consultation time, save significant physician time annually, and optimize resource allocation, thereby improving appointment scheduling across healthcare organizations (Zhakhina et al., 2023; Li et al., 2022). Remote pre-consultation assessments also reduce hospital congestion and nosocomial transmission risks, particularly during pandemics (Li et al., 2022).

Despite their potential, LLMs face challenges that must be addressed for widespread adoption. Ensuring data quality, mitigating bias, and maintaining interpretability are critical to realizing their full potential in clinical settings (Borkowski & Ben-Ari, 2024). Potential inaccuracies in patient-reported data and accessibility issues for populations with limited digital literacy necessitate user-friendly interfaces and inclusive design strategies (Zhakhina et al., 2023). The Smart-doctor system, while effective for common conditions, is limited to prescribing simple tests, which may not suffice for complex or rare diseases, potentially leading to misdiagnoses (Li et al., 2022). Patient engagement can also be hindered by unfamiliarity with digital interfaces, as evidenced by a 15.41% withdrawal rate due to distrust or operational difficulties (Li et al., 2022). Privacy concerns require robust data security measures to protect sensitive patient information, particularly given the online nature of these systems (Zhakhina et al., 2023; Li et al., 2022).

The application of LLMs in patient consultation, history taking, and preliminary diagnosis offers significant potential to revolutionize healthcare delivery. By streamlining data collection, enhancing diagnostic accuracy, and optimizing triage and appointment scheduling, LLMs foster patient-centered care and operational efficiency. However, challenges related to data accuracy, digital accessibility, disease complexity, and privacy must be addressed through targeted improvements, user-friendly designs, and rigorous validation. With careful management of these

limitations, LLMs can transform patient consultation processes, ensuring equitable access and reliable outcomes while prioritizing patient safety and ethical considerations.

2.1.5 LLMs for Triage and Prioritization

Integrating Large Language Models (LLMs) into triage and prioritization systems is revolutionizing appointment scheduling in healthcare settings by enhancing efficiency, patient-centered care, and resource allocation. LLMs, as core components of AI agents, excel in processing complex and unstructured clinical data, such as electronic health records (EHRs), patient-reported symptoms, and clinical notes, to facilitate real-time, evidence-based decision-making.

LLMs enable comprehensive data collection and analysis, streamlining triage processes. In a multi-agent AI system, Borkowski and Ben-Ari (2024) describe a Data Collection and Integration Agent that normalizes and organizes patient data from diverse sources, formatting it for both human and machine consumption to provide actionable insights for timely prioritization. Similarly, Zhakhina et al. (2023) emphasize LLMs' role in pre-consultation history-taking systems, where they analyze patient responses via web-based questionnaires to identify critical information based on symptom severity and medical history. This capability enhances the accuracy of triage decisions and supports patient engagement by allowing individuals to provide detailed medical histories through user-friendly platforms.

LLMs support risk stratification and prioritization by calculating severity scores and predicting patient outcomes. Borkowski and Ben-Ari (2024) highlight a Risk Stratification Agent that leverages LLMs to prioritize appointments based on clinical urgency, integrating consultant notes with clinical guidelines for contextually relevant decisions. Vuong (2024) notes that NLP-driven LLMs process unstructured EHR data to identify high-priority cases, improving operational efficiency and reducing clinician time on administrative tasks. These systems ensure appointments are allocated based on urgency and resource availability, minimizing wait times and optimizing resource utilization.

Virtual triage systems powered by LLMs have shown significant promise. Zhakhina et al. (2023) report that 75% of users in a multinational survey found such systems helpful in determining care needs, classifying patients according to appropriate care levels. However, Kopka et al.

(2025) indicate that while specialized Symptom-Assessment Applications (SAAs) improve self-triage accuracy significantly (from 53.2% to 64.5%, OR = 2.52, $p < .001$), LLMs like ChatGPT show limited improvement (54.8% pre vs. 54.2% post, $p = .79$). This suggests that LLMs may not consistently enhance triage accuracy for laypeople due to limited explainability and quantifiable uncertainty, which can reduce user trust. Users often rely on heuristics or seek additional verification when faced with conflicting or generic LLM advice.

Despite their potential, LLMs face challenges in healthcare deployment. Vuong (2024) underscores the need to address data privacy, mitigate biases in LLM outputs, and integrate these systems with existing workflows. Zhakhina et al. (2023) highlight the digital divide and data accuracy concerns, necessitating careful system design to ensure accessibility and reliability for diverse patient populations. Kopka et al. (2025) suggest that LLMs require higher personalization and explainability to improve their role in prioritization and ensure accurate care allocation.

2.1.6 LLMs for Intelligent Appointment Scheduling

The integration of large language models (LLMs) into healthcare management systems for intelligent appointment scheduling offers a transformative approach to optimizing patient flow, reducing waiting times, and enhancing resource allocation in outpatient clinics. By leveraging LLMs such as Palmyra-med-70b or OpenAI's GPT models, consultation and triage systems can process unstructured patient queries, predict demand patterns, and dynamically assign appointments based on clinical needs, patient preferences, and resource availability.

LLMs enhance consultation and triage systems by interpreting complex patient inputs, such as clinical notes or chief complaints, to facilitate pre-consultation diagnostics and prioritize appointment slots. A retrospective cohort study by Li et al. (2021) demonstrated the efficacy of an AI-assisted system, XIAO YI, which used natural language processing to automate the ordering of imaging examinations or laboratory tests, reducing median waiting times from 1.97 hours to 0.38 hours ($p < 0.05$) and lowering costs (median: 335.97 CNY vs. 364.58 CNY, $p < 0.05$) (Li et al., 2021). This system prioritized non-invasive, low-cost tests, with backstage doctors reviewing AI recommendations to ensure clinical appropriateness. LLMs can also predict service duration and patient arrival patterns, enabling strategies like overlapping appointment

scheduling (OLAS) to minimize both patient waiting times and provider idle times, as explored by Ala and Chen (2022).

The study by Munavalli *et al*. (2020) provides a framework for real-time scheduling through a multi-agent system (MAS) model at Aravind Eye Hospital, which used a hybrid ant-agent algorithm and Takt time management to reduce waiting times by 51.6% and improve resource utilization by 8.3% (Munavalli *et al*., 2020). LLMs can extend this paradigm by enabling context-aware negotiations between virtual agents representing patients and resources, transforming push-based systems into pull-based systems. Gupta and Denton (2008) emphasize Industrial Engineering and Operations Research (IE/OR) techniques to address variability in arrival and service times, patient preferences, and capacity reservation for urgent cases, enabling LLMs to balance provider utilization and accommodate same-day appointments (Gupta & Denton, 2008).

LLMs support fairness in scheduling by modeling patient and provider preferences, as discussed by Ala and Chen (2022). Panaviwat *et al*. (2014) demonstrated that a multiple block, fixed-interval appointment system with 30-minute intervals reduced weighted average waiting times by approximately 44% (from 96.94 minutes to 53.97–54.05 minutes) by prioritizing patients with laboratory test requirements and adapting to no-shows or unpunctual arrivals (Panaviwat *et al*., 2014). LLMs can predict patient punctuality and dynamically adjust schedules to ensure equitable waiting time distribution, as shown by Panaviwat *et al*.'s first-come, first-served (FCFS) approach, which reduced waiting times for appointment patients by up to 68%.

LLMs enable automated patient self-scheduling systems, which improve access and efficiency. Woodcock (2022) highlighted that self-scheduling systems allow 34% to 51% of appointments to be booked after hours, optimizing early morning slots and reducing no-show rates (Woodcock, 2022). However, barriers like socioeconomic disparities and limited internet access require LLMs to incorporate adaptive algorithms, such as multilingual support or simplified interfaces. Integration with electronic medical record systems via unique patient identifiers ensures seamless data exchange and maintains provider trust.

The success of systems like XIAO YI, which assisted over 270,000 visits, underscores the potential for LLMs to create adaptive, patient-centered scheduling systems. These systems can be evaluated using metrics such as waiting time reduction, patient satisfaction, no-show rates,

resource utilization, and access equity, aligning with the objectives of developing intelligent healthcare management platforms.

2.1.7 Impact of LLM-Based Systems on Patient Experience and Healthcare Access

Large Language Model (LLM)-based consultation and triage systems have significantly transformed patient experience and healthcare access by streamlining clinical workflows, optimizing appointment scheduling, and enhancing pre-consultation processes. These systems leverage artificial intelligence (AI) and natural language processing (NLP) to address operational inefficiencies, reduce wait times, and improve patient engagement, while also tackling challenges related to equitable access and data privacy.

LLM-based systems enhance patient experience by optimizing patient flow and reducing wait times. For example, the XIAO YI AI-assisted module automates the ordering of imaging examinations or laboratory tests based on patients' chief complaints, reducing median outpatient waiting times from 1.97 hours to 0.38 hours (Li et al., 2021). Similarly, AI-driven triage systems prioritize appointments based on symptom severity, ensuring timely care for critical cases (Ramdurai, 2020; Preiksaitis et al., 2024). Automated rescheduling systems, such as Fast Pass, utilize SMS and online portals to reduce appointment wait times by an average of 15 days for primary care and 24 days for specialty care, while also decreasing no-show rates by 38% through appointment reminders (Chung et al., 2020). These advancements demonstrate the potential of LLMs to improve operational efficiency and patient satisfaction in high-demand healthcare settings.

LLM-based systems improve healthcare access by enhancing the ease and flexibility of appointment scheduling. Automated self-scheduling platforms allow patients to book appointments at their convenience, with 34% to 51% of appointments scheduled after hours, accommodating patients with demanding schedules (Woodcock, 2022). Intuitive interfaces, such as kiosks or mobile applications, mitigate frustrations associated with traditional telephone-based or interactive voice response systems (Ramdurai, 2020). Additionally, multi-agent AI frameworks dynamically coordinate appointments based on patient acuity and resource availability, prioritizing vulnerable groups like senior citizens or patients with special needs to

ensure equitable access (Borkowski & Ben-Ari, 2024). Integration with electronic health records (EHRs) further enables real-time data processing, contributing to seamless and personalized patient interactions.

LLM-driven systems streamline pre-consultation processes by facilitating comprehensive data collection through digital questionnaires, allowing patients to provide detailed medical histories prior to appointments (Zhakhina et al., 2023; Alowais et al., 2023; Preiksaitis et al., 2024). This preemptive data gathering reduces consultation time spent on history-taking, enabling clinicians to focus on targeted discussions and personalized care, which enhances patient satisfaction and engagement. For instance, such systems have been shown to reduce patient wait times by approximately half, improving compliance and rapport (Zhakhina et al., 2023). Additionally, LLMs in chatbots like Florence and Babylon Health assess symptoms and direct patients to appropriate care levels, reducing the burden on call centers and reception desks, particularly in high-demand or resource-constrained settings (Shalko et al., 2024).

LLM-based systems contribute to cost-effective healthcare delivery without compromising service quality. The XIAO YI system, for example, reduced median total costs from 364.58 CNY to 335.97 CNY, making healthcare more affordable (Li et al., 2021). Furthermore, LLMs personalize patient experiences by analyzing EHRs and wearable device data to generate tailored health recommendations and interpret complex medical information in a patient-friendly manner (Shalko et al., 2024). Multilingual communication capabilities also enhance healthcare equity by breaking down language barriers, particularly for underserved populations.

Despite their transformative potential, LLM-based systems face challenges that must be addressed to ensure equitable access and maintain trust. Populations with limited digital literacy, internet connectivity, or socioeconomic resources such as older adults, those with comorbidities, or residents of low-income or remote areas may face barriers to adoption (Woodcock, 2022; Chung et al., 2020; Zhakhina et al., 2023). Data privacy and the risk of biases in AI-driven algorithms also pose concerns, potentially exacerbating disparities if not mitigated through robust cybersecurity measures and human oversight (Alowais et al., 2023; Borkowski & Ben-Ari, 2024). Additionally, LLMs may provide generalized responses that fail to account for individual patient characteristics, which could impact care quality in complex cases (Shalko et al., 2024). To address these issues, systems must incorporate user-friendly interfaces, alternative access

methods, and targeted outreach to underserved populations while ensuring data quality and security.

LLM-based consultation and triage systems offer a scalable and transformative approach to enhancing patient experience and healthcare access. By streamlining workflows, optimizing scheduling, and improving pre-consultation processes, these systems reduce wait times, enhance patient satisfaction, and promote cost-effective, patient-centered care. However, equitable access requires addressing technological and socioeconomic barriers through inclusive design and robust infrastructure. With continued refinement, LLM-based systems hold significant promise for fostering a more efficient and inclusive healthcare environment, particularly in resource-constrained settings.

2.1.8 Ethical, Technical, and Practical Challenges of LLMs in Healthcare

The integration of Large Language Models (LLMs) into healthcare consultation and triage systems for appointment scheduling presents a complex array of ethical, technical, and practical challenges that must be addressed to ensure safe, effective, and equitable implementation.

Ethical Challenges

Safeguarding patient data privacy and security is a paramount concern, as LLMs process sensitive clinical information, such as medical histories and clinical notes, making them vulnerable to cyberattacks and data breaches (Mumtaz et al., 2025; Adler-Milstein et al., 2022; Borkowski & Ben-Ari, 2024; Alowais et al., 2023; Vuong, 2024; Preiksaitis et al., 2024). Robust anonymization, compliance with regulations like HIPAA and GDPR, and advanced cybersecurity measures, such as federated learning, are essential to protect patient confidentiality and maintain trust (Mumtaz et al., 2025; Alowais et al., 2023; Preiksaitis et al., 2024).

The potential for LLMs to generate inconsistent, inaccurate, or biased outputs raises significant ethical risks. Inaccurate triage recommendations, as seen in cases like ChatGPT's unreliable responses to breast cancer screening queries or low specificity in triaging metastatic prostate cancer, could lead to patient harm or inequitable treatment (Mumtaz et al., 2025; Adler-Milstein et al., 2022; Borkowski & Ben-Ari, 2024). Bias in LLM outputs, often stemming from flawed or unrepresentative training data, risks discriminatory outcomes, particularly for underrepresented

populations (Alowais et al., 2023; Vuong, 2024; Preiksaitis et al., 2024). The phenomenon of "hallucination," where LLMs produce convincing but incorrect information, further undermines clinical reliability (Alowais et al., 2023). These issues necessitate continuous human oversight, rigorous validation, and transparent decision-making to align outputs with clinical standards (Mumtaz et al., 2025; Borkowski & Ben-Ari, 2024; Preiksaitis et al., 2024).

The legal and ethical implications of AI-driven decisions also pose challenges, as the distribution of responsibility between clinicians and LLMs remains undefined, complicating accountability (Preiksaitis et al., 2024). Addressing these ethical concerns requires robust governance frameworks to ensure patient safety, autonomy, and equitable care.

Technical Challenges

The accuracy and reliability of LLMs depend heavily on high-quality, comprehensive, and representative datasets, which are often scarce in healthcare due to restricted access to current medical records, particularly in specialized fields like dentistry (Mumtaz et al., 2025; Adler-Milstein et al., 2022; Borkowski & Ben-Ari, 2024; Alowais et al., 2023; Vuong, 2024). Incomplete or inconsistent datasets, compounded by up to 60% of electronic records containing errors or omissions, can introduce biases, impair diagnostic precision, and hinder triage reliability (Zhakhina et al., 2023; Preiksaitis et al., 2024). For instance, errors in patient-specific therapy suggestions for breast cancer and limited diagnostic accuracy in glioma classification highlight the risk of misdiagnoses (Mumtaz et al., 2025).

Integrating LLMs into existing healthcare IT infrastructures, such as Electronic Health Records (EHRs), presents significant interoperability challenges. Seamless functionality requires secure APIs, standardized protocols like HL7 FHIR, and standardized systems for reporting collected data to avoid workflow disruptions (Adler-Milstein et al., 2022; Borkowski & Ben-Ari, 2024; Alowais et al., 2023; Vuong, 2024; Zhakhina et al., 2023). The opaque nature of LLMs, characterized by complex architectures and obscure fine-tuning processes, further complicates transparency and explainability, making it difficult to identify and correct errors, which can erode clinician trust (Preiksaitis et al., 2024).

Additionally, the accuracy of LLMs hinges on patients' ability to understand medical terminology and provide reliable inputs. Usability issues, such as irrelevant questions or

interfaces that fail to capture nuanced patient data, can lead to diagnostic errors (Zhakhina et al., 2023). These technical challenges demand advancements in data quality, system interoperability, and model transparency to optimize LLM performance in healthcare settings.

Practical Challenges

The implementation of LLM-based systems requires substantial financial investment in infrastructure, training, maintenance, and advanced cybersecurity protocols, which can strain the budgets of healthcare organizations, particularly those with limited resources (Adler-Milstein et al., 2022; Borkowski & Ben-Ari, 2024; Zhakhina et al., 2023; Alowais et al., 2023). The digital divide poses a significant barrier, as older adults, low-income individuals, and those in remote areas may lack access to technology or digital literacy, potentially exacerbating health disparities and widening healthcare access gaps (Zhakhina et al., 2023; Alowais et al., 2023; Vuong, 2024).

Clinician trust and acceptance are critical practical challenges. The "black box" nature of LLMs, combined with potential medicolegal risks and the absence of standardized frameworks for evaluating LLM performance, may deter adoption (Adler-Milstein et al., 2022; Borkowski & Ben-Ari, 2024; Preiksaitis et al., 2024). Extensive clinician training is needed to enhance AI literacy and prevent overreliance, which could exacerbate automation bias (Preiksaitis et al., 2024). Similarly, patient trust and engagement vary, with concerns over privacy, lack of personalization, or interface accessibility particularly among elderly patients who may prioritize human interaction limiting compliance with LLM-based systems (Zhakhina et al., 2023; Alowais et al., 2023; Vuong, 2024; Preiksaitis et al., 2024).

Operational inefficiencies, such as those caused by low specificity in triage processes, can disrupt appointment scheduling workflows, requiring validation and alignment with clinical practices (Mumtaz et al., 2025). Addressing these practical challenges necessitates inclusive design, transparent communication, comprehensive training, and patient education to foster confidence and ensure equitable access.

The deployment of LLMs in healthcare consultation and triage systems for appointment scheduling offers transformative potential but requires careful navigation of ethical, technical, and practical challenges. A multidisciplinary approach combining robust ethical governance, technical standardization, and inclusive implementation strategies is essential to optimize LLM

integration, enhance healthcare delivery, and mitigate risks while prioritizing patient safety, equity, and trust.

2.1.9 Related Works

Kumar and Sulaiman (2024) in a study titled "AI-Based Booking Doctor Appointment for Holistic Health Care," focused on the development and evaluation of an AI-driven web system designed to streamline doctor appointment scheduling. The research aimed to tackle inefficiencies inherent in traditional manual booking processes, patient registration, and delays stemming from diagnostic procedures. By integrating advanced AI technologies, specifically Convolutional Neural Networks (CNNs) employing Generative Adversarial Networks (GANs) for iris image enhancement, the system sought to improve pre-diagnostic accuracy and overall administrative efficiency, ultimately enhancing the patient experience in a holistic healthcare context. The study's objective was to design, implement, and test this system, with a focus on user acceptance to determine its practical effectiveness, aligning partially with broader goals of intelligent healthcare management platforms that incorporate automation and diagnostics. The methodology employed by Kumar and Sulaiman (2024) utilized the Waterfall model, a linear and sequential system development framework consisting of distinct phases: requirement analysis, system design, implementation, testing, deployment, and maintenance. An object-oriented design approach was adopted, supported by Unified Modeling Language (UML) diagrams including use case, sequence, and class diagrams to map out system interactions and architecture. The system was built using PHP and MySQL for the core web framework, with the Django framework facilitating the integration of a GAN-based CNN, trained on Google Colab, to enhance iris images for pre-diagnosis. The testing phase encompassed system testing to validate functionalities like appointment booking and AI-driven image enhancement, alongside User Acceptance Testing (UAT) divided into alpha testing (conducted internally) and beta testing (involving seven patients) to gauge functionality and user satisfaction. The results of study indicated that the system effectively met its intended goals. It successfully automated manual processes, significantly reducing appointment scheduling delays and improving patient acquisition and administrative workflows at the holistic healthcare facility. The system featured a range of modules, including advertisement management, user login, appointment scheduling,

supplement inventory tracking, prescription management, digital receipt generation, and data analytics. The AI-enhanced iris imaging, powered by GANs, proved functional for pre-diagnostic purposes, though the authors noted it required further refinement for optimal performance. System testing confirmed the reliability of core features such as booking, canceling, and viewing appointments with all test cases passing successfully. User Acceptance Testing revealed strong approval: doctors rated the system 8/10, receptionists gave it a 9/10, and patients provided positive feedback with mean satisfaction scores ranging from 3.29 to 4.86 on a 5-point scale. Users suggested enhancements, such as improving the AI image enhancer's precision and refining the prescription management module, pointing to areas for future development. Despite its successes, the study faced implicit limitations that warrant consideration. The system's design relied heavily on structured inputs and predefined workflows, lacking the flexibility of advanced natural language processing or conversational AI capabilities. The AI functionality was confined to image enhancement via GAN-based CNNs, without extending to consultation or triage features that could interpret patient symptoms or prioritize appointments based on medical urgency. This restricted its adaptability to diverse patient needs and broader healthcare contexts, as it was tailored specifically to a single facility employing iridology-based diagnostics. The use of the Waterfall methodology, while structured, may have constrained iterative refinements by limiting real-time incorporation of user feedback during development. These factors suggest a need for further research to explore more dynamic, scalable solutions such as those incorporating Large Language Models (LLMs) for consultation and triage which could address unstructured patient inputs and optimize scheduling across multiple organizations, offering a more comprehensive approach to healthcare management beyond the localized, image-focused system presented by Kumar and Sulaiman (2024).

Deepa *et al*. (2024) in "AI-Powered Doctolib: Revolutionizing Healthcare Appointment Management," developed and evaluated an AI-driven platform, Doctolib, to enhance healthcare appointment management. The researchers aimed to address inefficiencies in traditional systems such as appointment delays and no-shows by integrating advanced features including real-time scheduling, predictive analytics to forecast patient attendance, a 24/7 AI chatbot for constant accessibility, and secure telemedicine capabilities to expand virtual care options. The study focused on improving patient engagement through easier access to services, optimizing resource allocation for healthcare providers by streamlining workflows, and reducing operational burdens,

with a particular emphasis on scalability and accessibility to serve underserved regions effectively. They adopted a methodology called the Design Thinking approach, structured around five iterative stages: Empathy, Define, Ideate, Prototype, and Test. The Empathy phase involved conducting surveys and interviews with patients and healthcare providers to identify key pain points, such as scheduling delays and limited access to virtual care. The Define stage translated these findings into specific system requirements, prioritizing a user-friendly interface and an efficient backend. During the Ideate phase, the team selected Doctolib's AI-driven features like the chatbot, predictive analytics, and telemedicine portal over less scalable alternatives after brainstorming potential solutions. The Prototype stage resulted in a functional system built with React.js for the frontend to ensure a responsive user interface, Node.js paired with Express.js for a robust backend, and MongoDB for flexible data storage. The AI chatbot leveraged natural language processing (NLP) and machine learning (ML) to handle bookings and respond to queries. Testing included unit tests to verify individual components, integration tests to ensure system cohesion, and performance evaluations using both simulated datasets and real-world pilot clinic deployments, with key performance indicators (KPIs) such as scalability, response time, and user satisfaction measured throughout. The results of the study highlights Doctolib's effectiveness in transforming appointment management. Predictive analytics reduced no-shows by 30% through timely reminders and flexible rescheduling options, while real-time scheduling and automation decreased the administrative workload by 20%, allowing staff to focus on patient care. Telemedicine usage surged by 40%, indicating strong patient acceptance of virtual consultations, and user satisfaction surveys reported an 85% approval rate, with patients praising the chatbot's responsiveness and providers appreciating workflow improvements. Scalability tests confirmed the platform could handle over 10,000 concurrent users without performance issues, reinforcing its potential for large-scale deployment across diverse healthcare settings. These outcomes underscore the value of AI in appointment systems, offering a foundation for this research, where LLMs could further enhance consultation accuracy and triage efficiency. A notable limitation is the lack of advanced capabilities within Doctolib's AI framework. The chatbot, while effective for booking appointments and answering FAQs using basic NLP and ML, does not possess the deep contextual understanding or reasoning power of LLMs, limiting its ability to perform complex tasks like assessing symptom severity or prioritizing appointments based on medical urgency. The platform's design also only caters to

individual healthcare settings rather than facilitating cross-organizational coordination. This gap highlights the need for this research, which could leverage LLMs to process nuanced patient inputs, provide intelligent triage, and enable seamless appointment coordination across diverse healthcare organizations, extending beyond Doctolib's scope of general scheduling and basic patient interaction.

Feng *et al*. (2024) explored "An Adaptive Decision Support System for Outpatient Appointment Scheduling with Heterogeneous Service Times," with the primary objective of developing a data-driven decision support system called Cluster-Predict-Schedule (CPS) to improve outpatient appointment scheduling by tackling the challenge of varying consultation durations among patients. The study sought to enhance operational efficiency in outpatient clinics by minimizing patient wait times, reducing periods when physicians were not seeing patients, and cutting down on extra hours needed beyond regular schedules, all while ensuring fair service delivery across patients; unlike traditional approaches that group patients into fixed categories like new or returning or assume uniform consultation times, CPS uses machine learning to dynamically classify patients based on historical data and create optimized schedules tailored to real-world clinic needs. The methodology unfolded in three stages: first, the Cluster stage analyzed past patient service time records to group patients into categories based on how long their consultations typically lasted, using a clustering technique followed by a decision tree to assign patients to these groups; next, the Predict stage applied a supervised machine learning model to estimate how long new patients' consultations might take based on their individual traits, feeding these predictions into the grouping process; finally, the Schedule stage generated scheduling templates using a candidate rules generator algorithm, which were tested through simulations to find the best balance of wait times, idle periods, and extra hours, with real-world clinical data used to compare CPS against benchmarks like the First-Call, First-Appointment method and the traditional New/Return classification with standard sequencing rules, assuming a single physician working in a single session with equal time slots and punctual patient arrivals. Results showed CPS significantly outperformed these benchmarks, cutting overall scheduling inefficiencies by 15.0% compared to the First-Call, First-Appointment approach and by 4.7% against the New/Return method with traditional rules, while also reducing disparities in how long patients waited across different appointment slots and proving adaptable to various clinic priorities such as emphasizing shorter waits for patients over minimizing physician downtime

demonstrated by testing different group sizes like two or three categories; the system's simple scheduling rules closely matched ideal outcomes from more exhaustive methods, making it practical for diverse settings. However, a key limitation is that CPS focuses narrowly on scheduling efficiency based on predicted consultation times from historical data and basic patient characteristics, lacking the ability to handle real-time patient inputs like symptom descriptions or to perform advanced consultation and triage tasks, and it is designed for a single-physician, single-session context rather than coordinating across multiple healthcare organizations.

Taylor *et al*. (2024) explored "Bespoke Large Language Models for Digital Triage Assistance in Mental Health Care," with the objective of developing and assessing tailored large language models (LLMs) to support clinicians in triaging mental health referrals within the UK's National Health Service (NHS) by analyzing unstructured narrative text from electronic health records (EHRs), aiming to streamline the inefficient manual triage process that handles 370,000–470,000 monthly referrals in 2023 and reduce referral bouncing by suggesting appropriate secondary care teams. Using a decade's worth of 8 million de-identified clinical notes from 200,000 patients at Oxford Health NHS Foundation Trust (OHFT), their methodology involved a classification task to pinpoint accepting triage teams, employing a 14-day activity cutoff heuristic to label referrals and comparing three LLM approaches: a brute force method processing individual truncated documents with a RoBERTa-based model and aggregating via majority voting, a concatenated truncated method combining documents into a single input for RoBERTa or Clinical-Longformer, and a segment-and-batch method that segmented concatenated documents into chunks processed with a custom RoBERTa-OHFT model fine-tuned on a single NVIDIA Tesla T4 GPU using Low-Rank Adaptation (LoRA) for efficiency, evaluated by accuracy, F1 score, precision, and recall across five mental health sub-specialties like eating disorders and psychosis intervention. Results showed the segment-and-batch approach excelling with an accuracy of 0.981 and F1 score of 0.938, surpassing brute force (accuracy 0.935, F1 0.846) and concatenated methods (e.g., Clinical-Longformer: accuracy 0.975, F1 0.924), with LoRA cutting trainable parameters to under 1% while retaining strong performance (accuracy 0.968, F1 0.924), proving feasible for resource-limited NHS settings and improving with longer text sequences for richer clinical insights. Moreover, its limitations comes from its confinement to triage within one NHS trust, lacking real-time consultation capabilities or multi-organizational scheduling integration, as it processes historical EHR data without incorporating live patient inputs like symptom

descriptions or extending beyond team recommendations to optimize appointments across multiple entities.

Akinode and Oloruntoba (2017) investigated "Design and Implementation of a Patient Appointment and Scheduling System," aiming to create a web-based system to improve healthcare delivery efficiency in developing countries like Nigeria by automating patient appointment scheduling, with the goal of reducing patient waiting times, minimizing overtime for doctors and nurses, and easing peak workloads for counter staff to enhance patient satisfaction and optimize outpatient clinic resources. Their methodology utilized AngularJS as a JavaScript MVC framework to build a responsive single-page application (SPA) for the front end, paired with PHP for server-side logic, AJAX for seamless client-server communication, and a dual-database backend of SQLite3 and MySQL for data management, supported by Apache server software, it also involved designing use case diagrams to map interactions for patients, doctors, and receptionists, an Entity Relationship Diagram and relational data model to structure the database, and implementing features like patient registration, appointment booking, viewing, and rejection, alongside doctor schedule editing and visibility into booked, free, and pending slots. The results showed the system effectively cut waiting times and eliminated long queues by allowing patients to book appointments online based on available slots, with doctors and administrators managing schedules efficiently through a modular, cross-browser-compatible SPA and a robust backend, delivering timely healthcare access as demonstrated by appointment status and schedule editing modules, though specific quantitative reductions in waiting times were not detailed, the authors highlighted improved efficiency and satisfaction for all users, but the system's focuses on basic scheduling automation within a single facility, lacking integration with consultation or triage processes and being unable to coordinate across multiple organizations, as it depends on predefined slots and manual inputs without leveraging advanced AI like LLMs to interpret patient needs or prioritize based on urgency.

Garrido *et al*. (2024) explored "Innovation Through Artificial Intelligence in Triage Systems for Resource Optimization in Future Pandemics," with the objective of demonstrating how artificial intelligence, specifically through a machine learning system using the extreme gradient boosting (XGB) algorithm, can improve triage in hospital emergency departments during pandemics like COVID-19 by quickly identifying patients at high risk of mortality and infection

to enable early interventions and optimize resource use, aiming to enhance decision-making speed and accuracy over traditional methods for better patient outcomes and disease containment. Their methodology involved a single-center, observational, cross-sectional study at Virgen de la Luz Hospital in Cuenca, Spain, from March 2 to April 30, 2020, analyzing 708 adult patients with confirmed COVID-19, collecting 89 variables demographics, comorbidities, symptoms, vital signs, lab results, imaging, and treatments from emergency records, they preprocessed data by addressing missing values, encoding categories, normalizing numbers, and balancing classes, then trained the XGB model with a 70-30 train-test split and 5-fold cross-validation, fine-tuning it with Bayesian optimization, and compared it against models like Random Forest and SVM using metrics such as accuracy and AUC. Results showed XGB achieving a balanced accuracy of 91.61% and an AUC of 0.92, surpassing Random Forest (89.16%) and AdaBoost (88.53%), rapidly pinpointing key mortality predictors like procalcitonin, age, oxygen saturation, lactate dehydrogenase, C-reactive protein, chest X-ray findings, and D-dimer, while emphasizing oxygen therapy's role in resource planning, thus reducing emergency department strain and optimizing care during pandemics. Its over reliance on structured clinical data for mortality prediction within one emergency department poses a major limitation, the system also lacks integration of real-time consultation narratives or scheduling across multiple organizations, as it does not process unstructured patient inputs like symptom descriptions, restricting its scope to immediate triage and leaving room for a LLM-based platform to enhance triage operations with consultation analysis and enable broader, multi-entity scheduling optimization.

Komarneni *et al*. (n.d.) investigated "Optimizing Doctor Availability and Appointment Allocation in Hospitals Through Digital Technology and AI Integration," with the objective of tackling patient no-shows by developing a machine learning model, specifically a Support Vector Machine (SVM) classifier, to predict missed appointments using the "Medical Appointment No Shows" dataset of 110,527 visits, aiming to identify patterns in patient behavior to enhance scheduling efficiency, reduce wait times, and improve resource utilization and patient satisfaction in hospitals through digital technology integration. Their methodology employed a three-tier architecture database, application logic, and user interface built with Django, MySQL, HTML, CSS, and JavaScript, focusing on preprocessing the dataset by handling missing values, encoding categorical variables like gender and neighborhood, scaling numerical features like age, and engineering new features like wait periods, they balanced the dataset's 80% "Show" versus

20% "No-show" classes with under-sampling and oversampling, tested multiple supervised learning models including Decision Trees, Linear and Non-Linear SVM, Logistic Regression, and KNN using the SK-Learn API with grid search for hyperparameter tuning, applied feature selection and dimensionality reduction techniques like PCA, and developed a patient mobile app and admin web interface for scheduling and management. The Appointment Booking and Intuitive Management (ASIM) system achieved an accuracy and F1-score of 80%, revealing that SMS reminders boost attendance, scholarship recipients are more likely to miss appointments, and women are nearly twice as likely as men to attend, enabling real-time availability updates, shorter wait times, and an enhanced patient experience via online booking and reminders. The system's reliance on structured data without integrating unstructured consultation narratives or real-time triage capabilities exposes a limitation, and its single-hospital focus, lacking coordination across multiple organizations, highlights the potential for a LLM-based platform to excel by processing consultation text for improved triage and extend scheduling optimization across a broader healthcare network.

Nwankwo *et al*. (2023) investigated "Web Based Medical Consulting Information Flow for Hospital Out-Patients Using Machine Learning Techniques," its objective was to develop a natural language-based medical consulting system for hospital outpatients in Nigeria, utilizing machine learning and natural language processing (NLP) with a Word Rank algorithm to summarize patient case reports, aiming to improve healthcare delivery efficiency by addressing outpatient department (OPD) management bottlenecks like disordered information flow, reducing physician effort, enhancing consultations, and cutting patient waiting times for integration into the Nigerian Electronic Health Records (NEHR). Their methodology featured a web-based system with a two-part architecture: a user interface (UI) designed with Universal Modeling Language (UML) tools for patient and physician interaction, supporting speech-to-text conversion via NLP, and a backend integrating clinical decision support (CDS) systems and knowledge bases, it employed the Word Rank algorithm to tokenize, lemmatize, and rank words in patient reports for summaries, used machine learning for heart disease diagnosis, captured patient data like symptoms and vital signs through forms and voice inputs, and evaluated summary accuracy with cosine similarity against traditional OPD flows. Results showed a 90% accuracy in heart disease diagnosis, effective voice-to-text translation for faster documentation, reduced physician recording time, and enhanced information flow with summarized reports,

CDS suggestions, and SMS reminders, leading to shorter patient wait times and better consultation efficiency. However, the single-hospital OPD focus without mechanisms for multi-organizational appointment scheduling or real-time triage prioritization beyond follow-ups exposes a limitation that needs to be addressed, as it processes consultation data but lacks scalability across healthcare entities.

CHAPTER THREE

METHODOLOGY

3.1 System Design

The system design integrates consultation, triage, and appointment scheduling functionalities, driven by LLMs to process unstructured patient inputs, assign triage levels, and optimize appointment allocation. The architecture comprises three primary modules: the Patient Interface Module, the LLM Processing Module, and the Doctor Interface Module, as depicted in the system architecture diagram in Figure 1 below. The Patient Interface Module facilitates user registration and symptom input via a web-based interface. The LLM Processing Module employs LLM models such as Palmyra-med-70b and OpenAI's GPT model to analyze symptoms, generate follow-up questions, assign triage levels and book appointment with available doctors. The Doctor Interface Module allows doctors to view and interact with appointment details such as viewing symptoms, follow up questions, triage level and adding notes where necessary.

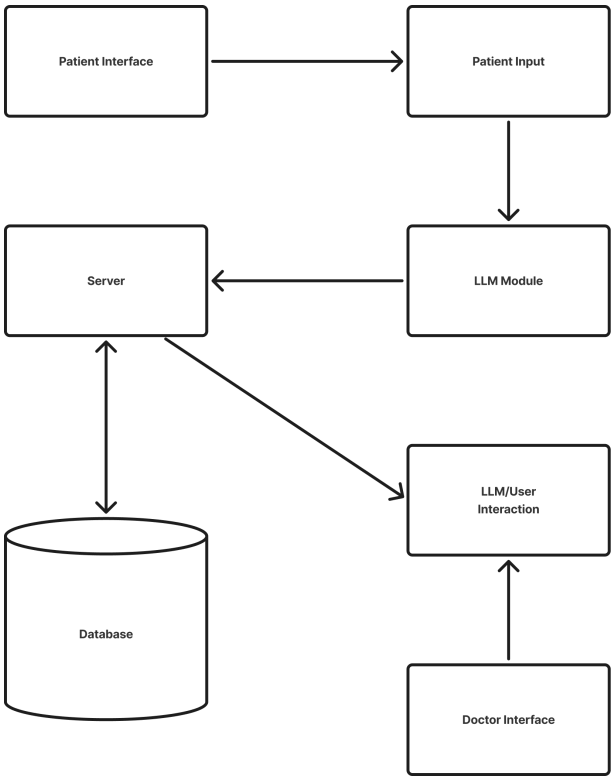

Figure 1. System Architecture

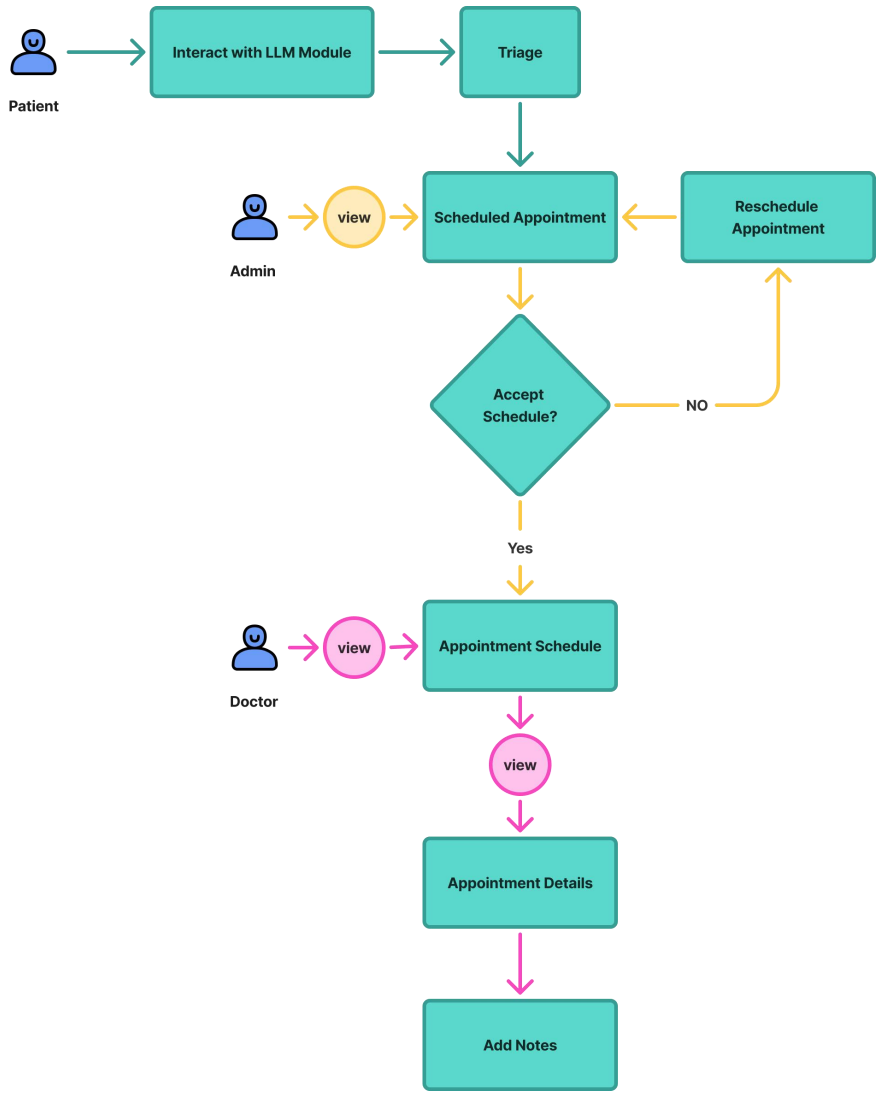

Figure 2. User conceptual flow diagram

The application flow is formalized as a sequential process. A patient registers and inputs symptoms as shown in Figure 1 and Figure 2:

$$S=\{s_1, s_2, \ldots, s_n\}$$

where s_i represents individual symptom descriptions.

The LLM Processing Module vectorizes symptoms into a high-dimensional embedding space:

$$v_S \in \mathbb{R}^d$$

using transformer-based embeddings (e.g., Palmyra-med-70b).

Follow-up questions are generated based on symptom analysis:

$$Q=\{q_1, q_1, \ldots, q_m\}$$

and patient responses are collected as:

$$R=\{r_1, r_2, \ldots, r_m\}$$

These responses are integrated to refine the symptom vector, yielding:

$$v_{S+R}$$

The triage level is inferred as:

$$T \in \{T_1, T_2, \ldots, T_k\}$$

(e.g., self-care, urgent, non-urgent, critical), and is computed using a classification function:

$$T = f(v_{S+R}; \theta)$$

where f is the LLM's inference model parameterized by θ.

The appointment scheduling optimizes slot allocation based on triage level, provider availability, and organizational constraints, modeled as an optimization problem:

$$\min \sum_{i=1}^{N} \omega_i . t_i$$

Subject to the constraints:

$$t_i \geq a_i, \qquad \sum_{i \in P_j} t_i \leq C_j, \quad T_i \, matches \, P_j$$

where:

- t_i: assigned time slot for patient *i*
- ω_i: priority weight based on triage level
- a_i: earliest available time

- C_j: capacity of provider j
- P_j: set of patients assigned to provider j

3.2 Technology Stack and Tools

The system is implemented using a modern web technology stack to ensure scalability, responsiveness, and seamless integration of LLM functionalities. The frontend will be developed with Next.js, a React-based framework that supports server-side rendering and static site generation, optimizing user experience and performance. The backend will leverage Node.js with Express.js for API development, interfacing with a MongoDB database for flexible storage of patient profiles, symptom data, and appointment records.

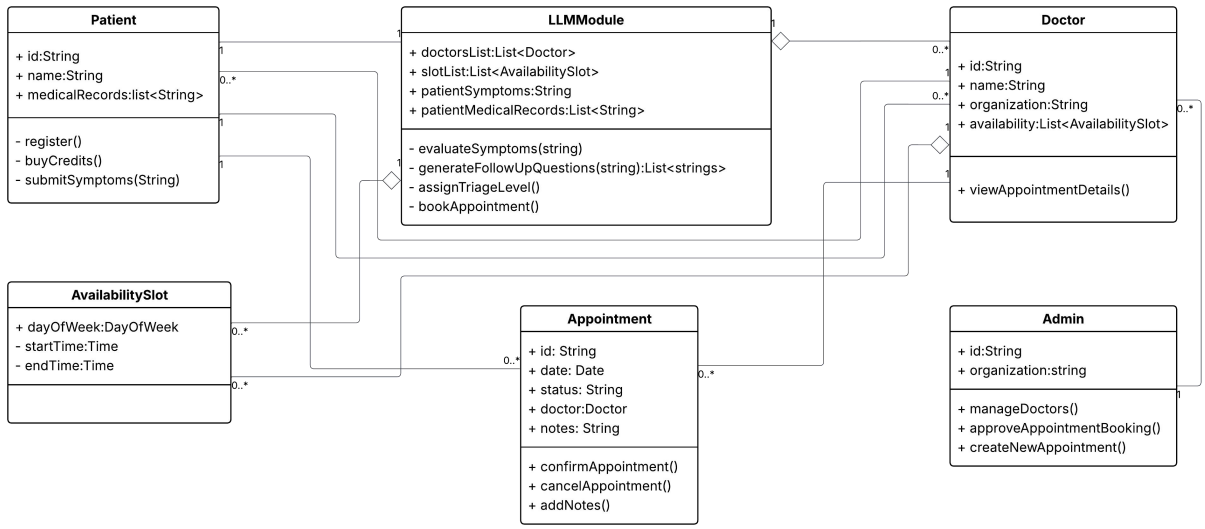

Figure 3. System class diagram

For AI functionalities, two LLMs will be employed: Palmyra-med-70b, a domain-specific model optimized for medical text processing, and OpenAI's GPT model (e.g., GPT-4) for general-purpose language understanding and generation. Palmyra-med-70b excels in clinical feature extraction from unstructured symptom narratives, while GPT-4 supports conversational follow-up questions and patient-friendly outputs. These models are accessed via secure APIs, with fine-

tuning performed on a curated medical corpus to enhance triage accuracy and accurate appointment booking.

Additional tools include Docker for containerization, ensuring consistent deployment across environments, and Git for version control. Cloud infrastructure, such as AWS, supports scalability and real-time processing, critical for handling concurrent users across multiple organizations.

3.3 Data Collection

Data collection occurs in real-time during the consultation phase, aligning with the application flow. Patients register via the web interface, providing demographic details (e.g., age, gender) and symptom narratives through text processed by the LLM model. The LLM Processing Module analyzes initial symptoms and generates follow-up questions to elicit additional context, such as symptom duration or severity. Responses are stored as structured JSON objects in MongoDB, with sensitive data anonymized to comply with HIPAA and GDPR regulations.

The dataset comprises:

i. Symptom Narratives: Unstructured text inputs, e.g., “persistent cough for three days.”

ii. Patient Responses: Answers to follow-up questions, e.g., “Is the cough accompanied by fever?”

iii. Triage Levels: Assigned categories based on clinical urgency (e.g., self-care, urgent, non-urgent, critical).

iv. Appointment Metadata: Provider availability, time slots, and organizational constraints.

To ensure data quality, inputs are validated for completeness, and erroneous entries (e.g., ambiguous symptoms) are flagged for manual review. A synthetic dataset, derived from publicly available medical corpora (e.g., MIMIC-III), is used for initial model training and testing, supplemented by real-world pilot data collected during evaluation.

3.4 Implementation Process

The implementation follows an iterative Agile methodology, structured in three phases: system development, testing, and deployment. This approach contrasts with the Waterfall model used in prior work (Kumar & Sulaiman, 2024), enabling continuous refinement based on user feedback.

1. **System Development:** The system is developed in sprints, with the frontend built using Next.js for responsive interfaces and the backend implemented with Node.js and MongoDB. The LLM Processing Module integrates Palmyra-med-70b and GPT-4 via APIs, to process patient symptoms, generate follow up questions, assign triage and book appointment. Appointment scheduling employs a rule-based algorithm augmented with optimization logic, as described earlier.

2. **Testing:** Unit tests validate individual components (e.g., symptom parsing, triage assignment), while integration tests ensure module cohesion. Performance tests assess scalability under simulated loads (e.g., 10,000 concurrent users), following Deepa *et al*. (2024).

3. **Deployment:** The system will be deployed on AWS using Docker containers, with continuous integration/continuous deployment (CI/CD) pipelines for updates. Administrative verification is implemented via a secure dashboard, allowing real-time appointment approvals.

The implementation prioritizes modularity and extensibility, enabling future integration of additional LLMs or scheduling algorithms.

3.5 Evaluation Strategy

The system is evaluated using a comprehensive strategy encompassing usability testing, performance testing, end-to-end testing, responsiveness, and UX assessment, with specific techniques for the LLM Processing Module. Evaluation metrics align with industry standards and prior healthcare AI studies (Li *et al*., 2021; Deepa *et al*., 2024).

Usability Testing

Usability is assessed through User Acceptance Testing (UAT) involving patients, clinicians, and administrators. A cohort of 50 participants (20 patients, 20 clinicians, 10 administrators) interacts with the system in a controlled pilot study. Metrics include:

i. **Task Success Rate:** Percentage of successfully completed tasks (e.g., symptom input, appointment booking).

ii. **System Usability Scale (SUS):** A standardized questionnaire yielding a score from 0 to 100, targeting a minimum of 80 for high usability.

iii. **User Satisfaction:** Likert-scale surveys (1–5) on ease of use, responsiveness, and trust, aiming for a mean score ≥ 4.

Qualitative feedback is collected via semi-structured interviews to identify interface improvements, addressing barriers noted in prior studies (Woodcock, 2022).

Performance Testing

Performance is evaluated under varying loads to ensure scalability and responsiveness:

I. **Response Time:** Average time for symptom analysis and triage assignment, targeting ≤ 2 seconds.

II. **Throughput:** Number of concurrent users supported without degradation, aiming for ≥ 10,000 users.

III. **Resource Utilization:** CPU and memory usage during peak loads, monitored using AWS CloudWatch.

Stress testing simulates high-demand scenarios (e.g., pandemic surges) to verify robustness, aligning with scalability tests in Deepa *et al*. (2024).

End-to-End Testing

End-to-end testing validates the entire application flow, from patient registration to administrative verification. Automated test scripts, developed using Cypress, simulate user journeys and verify data consistency across modules. Key metrics include:

i. **Error Rate:** Percentage of failed transactions (e.g., incorrect triage assignments), targeting < 1%.

ii. **Data Integrity:** Consistency of symptom and appointment data in MongoDB, validated via checksums.

Responsiveness and UX

Responsiveness is tested across devices (desktop, tablet, mobile) using BrowserStack, ensuring compatibility with major browsers (Chrome, Firefox, Safari). UX metrics include:

I. **Page Load Time:** Targeting ≤ 3 seconds for initial load.

II. **Mobile Accessibility:** Compliance with WCAG 2.1 standards, verified using Lighthouse.

III. **Navigation Efficiency:** Average clicks to complete tasks, aiming for ≤ 5 clicks.

LLM Module Evaluation

The LLM Processing Module is evaluated for triage accuracy and conversational effectiveness:

i. Triage Accuracy: Precision, recall, and F1-score for triage level assignments, benchmarked against ground-truth labels from clinicians. The target is an F1-score ≥ 0.90, comparable to prior studies (Li *et al*., 2022).

ii. Embedding Quality: Cosine similarity between symptom embeddings and reference medical texts, targeting ≥ 0.85.

iii. Question Relevance: Percentage of follow-up questions deemed clinically relevant by experts, aiming for ≥ 95%.

iv. Explainability: Use of attention visualization to interpret LLM decisions, ensuring transparency (Preiksaitis *et al*., 2024).

A confusion matrix analyzes triage misclassifications, and ROC-AUC assesses model performance across triage levels. Synthetic and real-world datasets will be used to mitigate bias, with results validated by a panel of medical experts.

Comparative Analysis

The system is benchmarked against existing solutions, such as Doctolib (Deepa *et al*., 2024) and XIAO YI (Li *et al*., 2021), using metrics like waiting time reduction (target: ≥ 50%), no-show rate reduction (target: ≥ 30%), and patient satisfaction (target: ≥ 85%).

This evaluation strategy ensures the system meets functional, performance, and usability requirements while addressing ethical and technical challenges, positioning it as a scalable, patient-centered solution for healthcare management.

REFEENCES

Adler-Milstein, J., Aggarwal, N., Ahmed, M., Castner, J., Evans, B. J., Gonzalez, A. A., ... & Williams, A. (2022). Meeting the moment: addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis. *NAM perspectives*, *2022*.

Akinode, J. L., & Oloruntoba, S. A. (2017). Design and implementation of a patient appointment and scheduling system. *Department of Computer Science, Federal Polytechnic Ilaro Nigeria*.

Ala, A., & Chen, F. (2022). Appointment scheduling problem in complexity systems of the healthcare services: A comprehensive review. *Journal of Healthcare Engineering*, *2022*(1), 5819813.

Alowais, S.A., Alghamdi, S.S., Alsuhebany, N. *et al.* Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* **23**, 689 (2023). https://doi.org/10.1186/s12909-023-04698-z

Borkowski, A., & Ben-Ari, A. (2024). Multi-Agent AI Systems in Healthcare: Technical and Clinical Analysis.

Chung, S., Martinez, M. C., Frosch, D. L., Jones, V. G., & Chan, A. S. (2020). Patient-centric scheduling with the implementation of health information technology to improve the patient experience and access to care: retrospective case-control analysis. *Journal of Medical Internet Research*, *22*(6), e16451.

Deepa, P., Badrinath, P., Mohanaprasanth, N., Pranesh, S., & Sriram, E. (2024). AI-Powered Doctolib: Revolutionizing Healthcare Appointment Management. *Asian Journal of Basic Science & Research*, *6*(4), 90-102.

Feng, H., Jia, Y., Huang, T., Zhou, S., & Chen, H. (2024). An adaptive decision support system for outpatient appointment scheduling with heterogeneous service times. *Scientific Reports*, *14*(1), 27731.

Garrido, N. J., González-Martínez, F., Losada, S., Plaza, A., del Olmo, E., & Mateo, J. (2024). Innovation through Artificial Intelligence in Triage Systems for Resource Optimization in Future Pandemics. *Biomimetics*, *9*(7), 440. https://doi.org/10.3390/biomimetics9070440

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, *40*(9), 800-819.

Komarneni, P., Kalakoti, T. K., Narla, P. K., Alla, S. P., & Bomma, R. (nd) OPTIMIZING DOCTOR AVAILABILITY AND APPOINTMENT ALLOCATION IN HOSPITALS THROUGH DIGITAL TECHNOLOGY AND AI INTEGRATION.

Kopka, M., Wang, S. M., Kunz, S., Schmid, C., & Feufel, M. A. (2024). Technology-Supported Self-Triage Decision Making: A Mixed-Methods Study. *medRxiv*, 2024-09.

Kumar, T. M., & Sulaiman, N. L. (2024). AI-Based Booking Doctor Appointment for Holistic Health Care. *Applied Information Technology And Computer Science*, *5*(2), 1121-1140.

Li, X., Tian, D., Li, W. *et al.* Artificial intelligence-assisted reduction in patients' waiting time for outpatient process: a retrospective cohort study. *BMC Health Serv Res* **21**, 237 (2021). https://doi.org/10.1186/s12913-021-06248-z

Li, X., Tian, D., Li, W., Hu, Y., Dong, B., Wang, H., ... & Liu, S. (2022). Using artificial intelligence to reduce queuing time and improve satisfaction in pediatric outpatient service: a randomized clinical trial. *Frontiers in Pediatrics*, *10*, 929834.

Mumtaz, U., Ahmed, A., & Mumtaz, S. (2023). LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. *arXiv preprint arXiv:2311.12882*.

Munavalli, J. R., Rao, S. V., Srinivasan, A., & van Merode, G. G. (2020). An intelligent real-time scheduler for out-patient clinics: A multi-agent system model. *Health Informatics Journal*, *26*(4), 2383-2406.

Nwankwo, U. C., Ngene, N. J., Ezekeke, L. C., Onuora, J. N., & Obi, J. N. (2023). Web Based Medical Consulting Information Flow for Hospital Out-Patients Using Machine Learning Techniques.

Panaviwat, C., Lohasiriwat, H., & Tharmmaphornphilas, W. (2014, June). Designing an appointment system for an outpatient department. In *IOP Conference Series: Materials Science and Engineering* (Vol. 58, No. 1, p. 012010). IOP Publishing.

Preiksaitis, C., Ashenburg, N., Bunney, G., Chu, A., Kabeer, R., Riley, F., ... & Rose, C. (2024). The role of large language models in transforming emergency medicine: scoping review. *JMIR medical informatics*, *12*, e53787.

Ramdurai, B. How AI (Artificial Intelligence) can improve Patient Experience in OPD (Out-Patient Dept.).

Ramdurai, B. (2021). Use of artificial intelligence in patient experience in OP. *Computer Science and Engineering*, *11*(1), 1-8.

Shalko, M., Domina, O., Korobko, I., Melnyk, D., & Andriushchenko, A. (2024). The transformative impact of large language models in healthcare. *Technology audit and production reserves*, *6*(4 (80)).

Taylor, N., Kormilitzin, A., Lorge, I., Nevado-Holgado, A., & Joyce, D. W. (2024). Bespoke Large Language Models for Digital Triage Assistance in Mental Health Care. *arXiv preprint arXiv:2403.19790*.

Toker, K., Ataş, K., Mayadağlı, A., Görmezoğlu, Z., Tuncay, I., & Kazancıoğlu, R. (2024, October). A Solution to Reduce the Impact of Patients' No-Show Behavior on Hospital Operating Costs: Artificial Intelligence-Based Appointment System. In *Healthcare* (Vol. 12, No. 21, p. 2161). MDPI.

Vuong, Q. P. (2024). The potential for artificial intelligence and machine learning in healthcare: the future of healthcare through smart technologies.

Woodcock, E. W. (2022). Barriers to and facilitators of automated patient self-scheduling for health care organizations: scoping review. *Journal of Medical Internet Research*, *24*(1), e28323.

Zhakhina, G., Tapinova, K., Kainazarov, T., & Kanabekova, P. (2023). Pre-consultation history taking systems and their impact on modern practices: Advantages and limitations. *Journal of Clinical Medicine of Kazakhstan*, *20*(6), 26-35.

Zhao, P., Yoo, I., Lavoie, J., Lavoie, B. J., & Simoes, E. (2017). Web-based medical appointment systems: a systematic review. *Journal of medical Internet research*, *19*(4), e134.