



FEDERAL UNIVERSITY OF CEARÁ
TECHNOLOGY CENTER
DEPARTMENT OF ELECTRICAL ENGINEERING
GRADUATE PROGRAM IN ELECTRICAL ENGINEERING
DOCTORATE IN ELECTRICAL ENGINEERING

MARCELLO CARVALHO DOS REIS

INTEROPERABLE WEB PLATFORM BASED ON LARGE LANGUAGE MODELS
FOR MEDICAL DATA ANALYSIS

FORTALEZA

2024

MARCELLO CARVALHO DOS REIS

INTEROPERABLE WEB PLATFORM BASED ON LARGE LANGUAGE MODELS FOR
MEDICAL DATA ANALYSIS

Thesis presented to the Graduate Program in
Electrical Engineering of the Technology Center
at the Federal University of Ceará, as a partial
requirement for obtaining the title of Doctor in
Electrical Engineering. Area of Concentration:
Electrical Engineering.

Advisor: Prof. Dr. Victor Hugo Costa de
Albuquerque.

FORTALEZA

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

- R311i Reis, Marcello Carvalho dos.
Interoperable web plataform based on large language models for medicals data analasys / Marcello Carvalho dos Reis. – 2024.
82 f. : il. color.
- Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Fortaleza, 2024.
Orientação: Prof. Dr. Victor Hugo Costa de Albuquerque.
1. Interoperability. 2. Artificial Intelligence. 3. FHIR Standard. 4. Medical Data Integration. 5. Load Testing. I. Título.

CDD 621.3

MARCELLO CARVALHO DOS REIS

INTEROPERABLE WEB PLATFORM BASED ON LARGE LANGUAGE MODELS FOR
MEDICAL DATA ANALYSIS

Thesis presented to the Graduate Program
in Electrical Engineering of the Technology
Center at the Federal University of Ceará, as
a partial requirement for obtaining the title
of Doctor in Electrical Engineering. Area of
Concentration: Electrical Engineering.

Approved on: 13/12/2024.

EXAMINING COMMITTEE

Prof. Dr. Victor Hugo Costa de
Albuquerque (Advisor)
Universidade Federal do Ceará (UFC)

Prof. Dr. Bruno Riccelli dos Santos Silva
Universidade Federal do Ceará (UFC)

Prof. Dr. Placido Rogério Pinheiro
Universidade de Fortaleza (UNIFOR)

Prof. Dr. Renato William Rodrigues de Souza
Instituto Federal de Educação, Ciência e
Tecnologia do Ceará (IFCE)

Francisco Nauber Bernardo Gois
Controladoria Geral do Estado - Examinador Externo à Instituição

To my parents, Heloisa and José Luiz, for their unwavering faith in me and for investing in every step of this journey. Rafaelly, your presence was my source of security and the certainty that I was never alone on this path.

ACKNOWLEDGEMENTS

Completing this doctorate represents the realization of a dream, and this moment was made possible only thanks to the support and contribution of essential people in my life. To all of them, I express my deepest gratitude.

To my parents, who have always encouraged me to follow my dreams and have been my greatest examples of perseverance and integrity. Wherever they may be, know that without your love and dedication, this moment would not have come to fruition. My eternal thanks.

To my life partner, Rafaelly Rios, for her patience, resilience, and for being my constant inspiration. Your strength makes me believe even more that, with Education and Innovation, we can truly transform the world. Thank you for being by my side at every stage of this journey.

To my academic partner and sister, Patricia Reis, for her constant encouragement and for helping me see the potential we can achieve together. Our partnership and friendship were key in pushing me to always bring out the best in myself.

To my advisor, Victor Hugo, for his guidance, patience, and trust in me. Your advice was crucial to my academic and professional development, and I am grateful for your generosity in sharing knowledge and wisdom throughout this journey.

ABSTRACT

This Thesis presents an interoperable web platform for managing medical data, prioritizing security and integration of information from multiple sources using FHIR (Fast Healthcare Interoperability Resources). The objective is to optimize medical record analysis with artificial intelligence (AI) and machine learning, offering automatic alerts and preventive recommendations while complying with the General Data Protection Law (LGPD). The platform facilitates efficient sharing of data among hospitals, clinics, remote devices, and healthcare systems, improving diagnostic and treatment accuracy. The methodology involved creating a secure, LGPD-compliant web platform, the integration of data through FHIR to ensure interoperability. AI algorithms analyze medical data, generate alerts, and provide personalized recommendations. Performance was assessed in controlled and stress tests, focusing on scalability and security. Results highlighted promising performance of the Retrieve Augmentation Generation (RAG) technique with BAAI/bge-small-en embedding models. Metrics such as BertF1, BertP, and BertR ranged from 0.389 to 0.538, averaging 0.43, indicating moderate consistency. The average Bleu score was 0.442, reflecting diverse response quality, while Rouge metrics averaged 0.326, indicating lower precision. Performance with Chest X-rays and MedQA datasets showed better results with Chest X-rays, achieving higher scores but higher perplexity ($3.635e4$), indicating challenges in generating clinical text. MedQA showed greater response diversity (0.807) but lower precision. In qualitative analysis, Chest X-rays demonstrated higher semantic similarity (mean 0.767) compared to MedQA (mean 0.754). During load testing, the platform remained stable as user numbers increased, but response times grew under stress, suggesting bottlenecks in high-demand scenarios. In conclusion, the platform is a promising tool for integrating medical data and supporting clinical decisions. The FHIR standard ensured interoperability, while AI effectively analyzed records and issued alerts. Adjustments are needed in response times under heavy loads and improvements in infrastructure and mobile experience to encourage greater patient adoption.

Palavras-chave: Interoperability; Artificial Intelligence; FHIR Standard; Medical Data Integration; Load Testing

ABSTRACT

Esta tese apresenta uma plataforma web interoperável para gestão de dados médicos, priorizando segurança e integração de informações de diversas fontes, utilizando o padrão FHIR (Fast Healthcare Interoperability Resources). O objetivo é otimizar a análise de prontuários médicos com inteligência artificial (IA) e aprendizado de máquina, oferecendo alertas automáticos e recomendações preventivas, em conformidade com a Lei Geral de Proteção de Dados (LGPD). A plataforma promove o compartilhamento eficiente de dados entre hospitais, clínicas, dispositivos remotos e sistemas de saúde, melhorando a precisão no diagnóstico e tratamento. A metodologia envolveu a criação de uma plataforma web segura e compatível com a LGPD, integrando dados por meio do FHIR para garantir interoperabilidade. Algoritmos de IA analisam dados médicos, geram alertas e fornecem recomendações personalizadas. O desempenho foi avaliado em testes controlados e de estresse, com foco na escalabilidade e segurança. Os resultados destacaram o desempenho promissor da técnica Retrieve Augmentation Generation (RAG) com modelos de embedding BAAI/bge-small-en. Métricas como BertF1, BertP e BertR variaram de 0,389 a 0,538, com média de 0,43, indicando consistência moderada. A média do score Bleu foi 0,442, refletindo diversidade na qualidade das respostas, enquanto as métricas Rouge apresentaram média de 0,326, indicando menor precisão. No desempenho com os conjuntos Chest X-rays e MedQA, Chest X-rays obteve scores mais altos, mas maior perplexidade ($3,635e4$), evidenciando desafios na geração de textos clínicos. MedQA mostrou maior diversidade de respostas (0,807), mas menor precisão. Na análise qualitativa, Chest X-rays apresentou maior similaridade semântica (média 0,767) em comparação a MedQA (média 0,754). Nos testes de carga, a plataforma manteve estabilidade com aumento de usuários, mas o tempo de resposta cresceu sob estresse, indicando gargalos em cenários de alta demanda. Conclui-se que a plataforma é uma ferramenta promissora para integrar dados médicos e apoiar decisões clínicas. O padrão FHIR garantiu interoperabilidade, enquanto a IA foi eficaz na análise de registros e emissão de alertas. São necessários ajustes em tempos de resposta sob alta carga e melhorias na infraestrutura e experiência móvel para maior adesão de pacientes.

Palavras-chave: Interoperabilidade; Inteligência Artificial; Padrão FHIR; Integração de Dados Médicos; Teste de Carga.

LIST OF FIGURES

Figure 1 – Example of architecture for disease identification.	24
Figure 2 – Web Interface	31
Figure 3 – Text Analysis Interface	32
Figure 4 – Summarized process of decision making	33
Figure 5 – Illustrative example of our collaborative Multi Agent approach to medical decision making using the Chest-X-Ray dataset (reports only)	34
Figure 6 – System architecture	44
Figure 7 – Evaluations, containing the classification and the explanation of the classifi- cation, by two reviewers regarding how correct the answer is in relation to the question. Additionally, the feedback from the Agent Judge containing the classification and the explanation of the classification.....	46
Figure 8 – Graphic results of values for Pearson correlation and Hit Rate metrics and RMSE and MAE errors.	47
Figure 9 – Distribution of metrics results for Motivational technique of engineering prompt.....	51
Figure 10 – GPU stress utilization during workflow of MDT pipeline	54
Figure 11 – Costs of 1 million input tokens for different language model APIs	55
Figure 12 – Costs of 1 million output tokens for different language model APIs	56
Figure 13 – API cost analysis from a production perspective	56
Figure 14 – The Chain-of-thought prompt for analyzing the complexity of a consultation made by a general practitioner	60
Figure 15 – The standard prompt for analyzing the complexity of a consultation made by a specialist physician.	60
Figure 16 – Prompt used for the Moderator Agent	61
Figure 17 – Prompt used for the Recruiter Agent	62
Figure 18 – Prompt used for the Checker Agent.....	62
Figure 19 – Prompt used for the Specialist Agent.....	63
Figure 20 – Short prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.	63
Figure 21 – Zero-shot prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.	64

Figure 22 – Motivational prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.	65
Figure 23 – Few-shot prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.	66
Figure 24 – Few-shot + Motivational prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.	67
Figure 25 – Distribution of request times in load testing	68
Figure 26 – Request sampling table by endpoint in load testing	68
Figure 27 – Active user distribution graph in load testing.....	69
Figure 28 – Histogram of response times during load testing	69
Figure 29 – Distribution of request times in stress testing	70
Figure 30 – Request sampling table by endpoint in stress testing	70
Figure 31 – Active user distribution graph in stress testing.....	71
Figure 32 – Histogram of response times during stress testing	71

LIST OF TABLES

Table 1 – Summary of references on LLMs in various application areas.....	29
Table 2 – Quantitative intra-class analysis involving both reviewers and their respective feedback on answer quality	49
Table 3 – ICC calculation results for two human reviewers.....	50
Table 4 – Results table using the Pearson correlation, RMSE, and MAE metrics, based on two human reviewers.....	50
Table 5 – Results table using the Pearson correlation, RMSE, MAE, and Hit Rate metrics for two types of prompts: motivational prompt and motivational combined with few-shot prompt.....	52
Table 6 – Quantitative intra-class analysis involving both reviewers and their respective feedback on answer quality using motivational and combination of motiva- tional and few-shot prompts.	52
Table 7 – Average cost of building a similar infrastructure	57
Table 9 – Analyze comparative of results obtained from retrieval metrics.....	59
Table 10 – Specifications of the resources used in the development environment and performance testing.....	66

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANPD	National Data Protection Authority
APIs	Application Programming Interfaces
CDSS	Clinical Decision Support System
CoT	Representational State Transfer
CoT-SC	Self-Consistency with Chain of Thought
CPT	Current Procedural Terminology
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
EHRs	Electronic Health Records
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
HL7	Health Level Seven
ICD-10	International Classification of Diseases
IoT	Internet of Things
LGPD	General Data Protection Law
LLM	Large Language Model
LOINC	Logical Observation Identifiers Names and Codes
LSTM	Long Short-Term Memory
MDT	Multi-disciplinary Team
NLP	Natural Language Processing
NSHA	National Supplementary Health Agency
PCC	Primary Care Clinician
PIS	Private Insurance Superintendence
REST	Representational State Transfer
SC	Self-Consistency
SHIE	Supplementary Health Information Exchange
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
UHS	Unified Health System
UTSH	Unified Terminology of Supplementary Health

CONTENTS

1	INTRODUCTION.....	13
1.1	Research Questions	15
1.2	Objectives.....	15
1.2.1	<i>General Objective</i>	15
1.2.2	<i>Specific Objectives</i>	15
1.3	Contributions.....	16
1.4	Organization	17
2	LITERATURE REVIEW	18
2.1	Healthcare Market	18
2.2	Interoperability in Healthcare.....	19
2.2.1	<i>Health Level Seven International</i>	20
2.2.2	<i>Fast Healthcare Interoperability Resources</i>	20
2.2.3	<i>Digital Imaging and Communications in Medicine</i>	21
2.2.4	<i>Systematized Nomenclature of Medicine – Clinical Terms.....</i>	21
2.2.5	<i>Logical Observation Identifiers Names and Codes.....</i>	22
2.3	Artificial Intelligence and LLM	22
2.4	LGPD and Health Data Protection in Brazil	25
2.4.1	<i>Compliance Requirements and Informed Consent</i>	25
2.4.2	<i>Security of Sensitive Personal Data</i>	25
2.4.3	<i>Impact of LGPD on the Development of Digital Solutions</i>	25
2.5	Applications of Language Models in Healthcare: Advances and Limitations	26
3	METHODOLOGY.....	30
3.1	Technologies Used.....	30
3.2	Web Interface.....	31
3.3	Dataset.....	31
3.4	Agent Pipeline.....	32
3.5	Multi-collaborative Agent: Medical Decision-making	33
3.5.1	<i>Query Complexity</i>	35
3.5.2	<i>Medical Agent Roles</i>	35
3.6	Database Description	37
3.6.1	<i>Data Cleaning.....</i>	37

3.7	Validation Metrics	38
3.7.1	<i>System Evaluation</i>	38
3.7.2	<i>RAG Evaluation</i>	39
3.8	BackEnd	42
3.8.1	<i>Credential Verification</i>	43
3.8.2	<i>Architecture</i>	43
3.8.3	<i>Libraries and Tools Used</i>	44
4	RESULTS	45
4.1	Agent-LLM Based Judge: Is it a Good Judge?	45
4.1.1	<i>Intra-class correlation coefficient (ICC)</i>	48
4.1.1.1	<i>Types of ICC analyses</i>	48
4.2	Motivation is all you need	50
4.3	System Evaluation	53
4.3.1	<i>GPU Utilization</i>	53
4.3.2	<i>Cost of LLM API calls</i>	55
4.3.3	<i>Infrastructure Cost</i>	57
4.4	RAG Evaluation	57
4.4.1	<i>Comparative Quantitative Analysis</i>	58
4.4.2	<i>Comparative Qualitative Analysis</i>	59
4.5	Prompt Templates	59
4.5.1	<i>Medical Agents</i>	59
4.5.1.1	<i>General Practioner</i>	59
4.5.1.2	<i>Specialist, Moderator, Recruiter</i>	61
4.5.2	<i>LLM Judge</i>	63
4.6	Performance backend tests	65
4.7	Load Testing	67
4.8	Stress Test	69
5	CONCLUSIONS AND FUTURE WORK	72
	BIBLIOGRAPHY	74

1 INTRODUCTION

The storage, management, and sharing of health information represent complex challenges in the current healthcare landscape (CHUKWU; GARG, 2020; GHUGARE *et al.*, 2024; CHEN *et al.*, 2024). As medical care is provided across various institutions, including hospitals, diagnostic clinics, laboratories, and doctor's offices, a single patient's information is often recorded and stored in different information systems (VARDHAN *et al.*, 2024). In many cases, these systems are not interoperable, which hinders the efficient retrieval and sharing of data (LIU *et al.*, 2021).

The absence of a universal electronic health record prevents healthcare professionals from quickly accessing critical patient information during care, which can compromise diagnosis and treatment, especially in emergency situations (CHENG *et al.*, 2022). Limited access to a complete medical history remains a significant barrier, even in a scenario where digital technologies could potentially facilitate the collection of essential information for more effective diagnosis and treatment (FANG *et al.*, 2022).

Recent research has focused on solutions to mitigate the limitations faced in connected healthcare. For example, Zhou et al. (ZHOU *et al.*, 2024) investigated the management of health equipment in Computed Tomography (CT) through a remote self-care network integrated with the Internet of Things (IoT). This study stood out for its application of Artificial Intelligence (AI) to predict failures in medical equipment, aiming to reduce downtime and improve maintenance. Other works, such as those by (LIU *et al.*, 2024b; WU *et al.*, 2024; MCPEAK *et al.*, 2024; PAN *et al.*, 2024), explored the analysis of large volumes of textual data and the use of data mining to identify patterns in online medical records, improving disease detection. These studies demonstrated the potential of models based on large transformers, such as Large Language Model (LLM), and integration with knowledge graphs to enhance diagnostic accuracy and efficiency.

Data security in the storage and management of health information is a critical issue. Works such as (AGARWAL; PAL, 2024; MAMTA *et al.*, 2024; KUMAR *et al.*, 2024; AL-NBHANY *et al.*, 2024; KALA; PRIYA, 2024; ALIJOYO *et al.*, 2024) have proposed solutions using blockchain and fog computing to create sharing and storage systems on health IoT devices, promoting greater protection against unauthorized access and balancing privacy with data availability for research. These approaches enhance scalability, security, and computational efficiency in digital health systems, especially in hospital and telemedicine environments where

the protection of sensitive data is crucial.

Other studies highlighted solutions for the secure sharing of medical data, using different technologies, such as blockchain, access control based on ontologies, and deep learning. Liu et al. (LIU *et al.*, 2024a) proposed a distributed security model emphasizing scalability and data privacy. Alijoyo et al. (ALIJOYO *et al.*, 2024) explored secure sharing algorithms based on Ethereum Blockchain 2.0 and elliptic curve cryptography. Nakajima et al. (NAKAJIMA *et al.*, 2024) addressed the challenges of real-time sharing during the COVID-19 pandemic, highlighting the importance of zero-trust networks. Renuka and Guruprakash (RENUKA; GURUPRAKASH, 2024) proposed an access control model based on ontologies and encryption with the enhanced BB84 protocol, while Deepthika et al. (DEEPTHIKA *et al.*, 2024) integrated blockchain and Long Short-Term Memory (LSTM) networks to ensure privacy and security in data consent and sharing. These works illustrate how the combination of these technologies can offer robust solutions for the security of medical data.

Despite technological advances, patients' access to their own health information remains fragmented and limited. The creation of an efficient, secure, and user-friendly web platform is essential to centralize this data and provide patients with more integrated and accessible control over their health. Such a solution would not only facilitate the monitoring of diagnoses and treatments but also encourage patient autonomy and active participation in managing their own well-being.

Therefore, this research aims to analyze the feasibility of developing a web platform for processing medical data that centralizes, stores, and shares health information from various sources on demand. The proposal includes the use of artificial intelligence and machine learning technologies for the automated analysis of medical records, pattern identification, and issuing alerts in critical situations. Additionally, the platform may provide personalized preventive medicine content based on collected data. It will integrate data entered by patients themselves, validated by a technical team, and information from partner systems through Application Programming Interfaces (APIs) and connected medical devices.

To ensure the effectiveness and security of the solution, interoperability standards such as Fast Healthcare Interoperability Resources (FHIR) (POPE *et al.*, 2023) will be adopted, facilitating integration with other health systems. Data storage and management will be performed in cloud environments using advanced cybersecurity technologies, encryption, and compliance with regulations such as General Data Protection Law (LGPD) and General Data

Protection Regulation (GDPR), ensuring privacy and protection of patient data.

The proposed platform aims to offer patients an integrated tool that allows easy and centralized access to their health information, in addition to the secure and efficient sharing of these data with healthcare professionals. This initiative reflects a vision of a more agile, accessible, and preventive healthcare system, where the patient is empowered to manage their health with the support of advanced technologies and smart data.

1.1 Research Questions

This research seeks to analyze and reflect on the following research questions:

- What are effective strategies to ensure interoperability between different health systems using standards such as FHIR, enabling the efficient sharing of medical data between hospitals, clinics, and remote monitoring devices?
- How can AI and machine learning be applied to optimize the analysis of medical records, generate automatic alerts in critical situations, and contribute to providing preventive health recommendations?
- What are the best practices for developing an accessible and intuitive user interface that allows patients to securely and efficiently manage and share their health data?

1.2 Objectives

1.2.1 General Objective

Develop an interoperable web platform that centralizes, stores, and shares health information securely, integrating data from various sources and offering support for medical diagnosis and treatment through AI and machine learning, in compliance with LGPD.

1.2.2 Specific Objectives

- Ensure interoperability between different health systems using standards like FHIR, to facilitate data sharing between hospitals, clinics, laboratories, and remote monitoring devices.
- Ensure compliance with LGPD, ensuring that the processing of sensitive personal data follows the principles of consent, security, and transparency, respecting patients' rights to

their information.

- Implement AI and machine learning to analyze medical records and issue alerts in critical situations, as well as provide personalized preventive medicine recommendations.
- Develop an accessible and intuitive interface that allows patients full control over their data, facilitating the secure sharing of information with healthcare professionals and authorized third parties.

1.3 Contributions

This research makes significant contributions to the computational field, particularly regarding system integration, promoting interoperability between different platforms, and adopting standards like FHIR. These advancements are crucial for the progress of computing in healthcare, enabling scalable and more secure solutions. Furthermore, the use of AI and machine learning, enabling the real-time automatic analysis of large volumes of medical data, opens new opportunities for computational innovations. This results in the optimization of complex information processing and the development of new algorithms tailored to the healthcare sector.

From the user's perspective, the contribution lies in providing centralized and simplified access to their health data. The proposal of an integrated platform guarantees patients direct and easy control over their information, also enabling the secure and efficient sharing of these data with healthcare professionals. Moreover, security and privacy, aligned with LGPD compliance, ensure that patients' personal information is protected and will be used ethically with proper consent.

For healthcare professionals, the highlight is quick and consolidated access to patient health histories. The platform facilitates obtaining a complete medical history, improving diagnostic accuracy and the choice of appropriate treatments. More importantly, integration with AI and machine learning provides real-time insights and alerts regarding critical conditions, enhancing clinical decision-making and speeding up medical response in emergency situations.

Finally, for healthcare companies, operational efficiency is enhanced through the integration of different systems and data sources into a single platform. This optimizes internal processes and reduces the time needed for data collection and analysis. Compliance with LGPD not only mitigates legal risks but also strengthens public trust in the digital solutions offered. Companies adopting cutting-edge technologies for managing medical data position themselves competitively in the market, offering more efficient and secure services while contributing

to the development of a cloud infrastructure that enables interoperability among various web applications focused on healthcare. The focus will be on creating a robust authentication protocol for IoT applications and user interfaces.

1.4 Organization

In Chapter 1, Introduction, we will present the context and relevance of creating an interoperable platform, highlighting the limitations of current systems such as data fragmentation and the lack of patient information access. The importance of applying AI in data analysis and ensuring compliance with LGPD will also be discussed. The objectives and research questions, such as ensuring interoperability and LGPD compliance, will be outlined to guide the study's development.

In Chapter 2, Literature Review, we will explore previous works and relevant theories on interoperability, AI in healthcare, and data privacy. Standards such as FHIR, applications of AI in medical diagnosis, and the impacts of LGPD on managing sensitive data will be discussed, offering theoretical and practical foundations for the research.

In Chapter 3, Methodology, we will explain how the platform will be developed, detailing the selection of technologies and interoperability standards adopted, such as FHIR. The integration of systems and the AI algorithms to be employed will also be described. Additionally, this chapter will address the security measures implemented to ensure compliance with the General Data Protection Law (LGPD), including the use of encryption and other protective technologies, guaranteeing that patients retain control over their data.

In Chapter 4, Discussion of Expected Results, we will analyze the anticipated contributions of the research, such as technological innovations in the interoperability of healthcare systems and the use of AI to enhance diagnostic accuracy. For patients, centralized and secure access to their data is expected, while healthcare professionals will benefit from real-time information and decision-making support. For healthcare companies, expected outcomes include reduced operational costs and compliance with LGPD regulations.

Finally, in Chapter 5, Conclusion, we will summarize the platform's main contributions, emphasizing its technological innovation and benefits for the healthcare system. Research limitations and challenges, such as implementation costs and adapting existing systems, will also be discussed. Recommendations for future studies, including expanded use of AI and machine learning, will be presented, suggesting directions for continuous research advancement.

2 LITERATURE REVIEW

2.1 Healthcare Market

The aging population, the shortage of qualified healthcare professionals, and rising costs present substantial challenges for healthcare systems worldwide (MCGRATH; CHASLER, 2024). The growing demand for medical assistance due to the aging population increases pressure on medical resources, especially in high-income countries, which creates additional pressure on already overwhelmed healthcare infrastructures due to rising costs and demand for specialized care. Furthermore, the financial burden of long-term care is shifting to personal savings and pension funds, further complicating access to healthcare for older individuals (MCGRATH; CHASLER, 2024; ROESLER *et al.*, 2024a).

At the same time, digital innovations in healthcare, such as wearable devices, big data, and AI, are transforming the way healthcare is delivered, enhancing efficiency and personalization, and offering some relief by improving care delivery (Núcleo de Informação e Coordenação do Ponto BR, 2024). However, this vast and complex global healthcare market encompasses medical services, pharmaceuticals, healthcare equipment, insurance, and technologies targeted at the sector (ROESLER *et al.*, 2024a). It is important to note that many countries adopt mixed models, where public and private healthcare systems coexist, with regulations and attempts at interoperability that vary by country, depending on the adopted health policy.

In Brazil, specifically, the healthcare market is characterized by the coexistence of two systems: the public and universal Unified Health System (UHS) (Ministério da Saúde do Brasil, 2024) and the supplementary healthcare sector, consisting of private health plans (Núcleo de Informação e Coordenação do Ponto BR, 2024).

UHS, established by the Federal Constitution of 1988 and regulated by Laws No. 8.080/1990 and 8.142/1990 (Ministério da Saúde do Brasil, 2024), is based on the principles of universality, comprehensiveness, and equity, with the goal of ensuring that all Brazilian citizens have access to free healthcare. On the other hand, the private sector covers approximately 25% of the Brazilian population, using private health plans and insurance, regulated by National Supplementary Health Agency (NSHA) and Private Insurance Superintendence (PIS) (MCGRATH; CHASLER, 2024).

Although UHS faces challenges such as underfunding and long waiting lists, there is increasing adoption of telemedicine technologies and electronic health records (MAHADIK

et al., 2024). The quality of healthcare services in Brazil varies by region, with major urban centers offering better services compared to rural and poorer areas (Núcleo de Informação e Coordenação do Ponto BR, 2024). The private sector, in turn, heavily invests in top-tier hospitals and diagnostic technologies. To regulate the supplementary healthcare sector in Brazil, NSHA was created by Law No. 9.961/2000, which includes private health plans (Agência Nacional de Saúde Suplementar (ANS), 2024). Its role is to ensure the financial balance of health plan operators, guarantee the quality of services provided, and protect consumer rights, particularly defining standards for Data Exchange in Supplementary Healthcare to attempt to ensure interoperability between service providers, health plan operators, and NSHA itself. To this end, standards such as Supplementary Health Information Exchange (SHIE) and Unified Terminology of Supplementary Health (UTSH) were developed (SARACCO *et al.*, 2020).

2.2 Interoperability in Healthcare

Health data interoperability refers to the ability of different healthcare systems, devices, and applications to exchange, interpret, and use information in an integrated and efficient manner (MEYER; WILSON, 2023; PATEL; KIM, 2023). This is crucial to ensure that data generated in one healthcare system can be accessed, understood, and used in another, without the need for manual intervention or format conversion (SANTOS; ZHAO, 2024; NGUYEN; ALI, 2024).

Health data interoperability consists of three types: First, Technical Interoperability, which is the ability of different systems to exchange data through common standards, such as communication protocols and data formats, for example, Health Level Seven (HL7), FHIR, and Digital Imaging and Communications in Medicine (DICOM) (KUMAR; SMITH, 2024; MEYER; WILSON, 2023). This allows data to be sent and received between healthcare systems; Second, Semantic Interoperability, which ensures that the exchanged information is interpreted in the same way by all systems. This requires the use of standardized terminologies and taxonomies, such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) and Logical Observation Identifiers Names and Codes (LOINC) (PATEL; KIM, 2023; SANTOS; ZHAO, 2024), so that health data have the same meaning across different systems; Lastly, Organizational Interoperability, which involves collaboration between different healthcare organizations, ensuring that policies, standards, and procedures are aligned so that data exchange, for example, SHIE and UTSH, happens smoothly, respecting privacy and security norms (NGUYEN; ALI,

2024; KUMAR; SMITH, 2024).

2.2.1 Health Level Seven International

HL7 is an organization dedicated to developing standards for the exchange, integration, sharing, and retrieval of electronic health information (CHEN; GARCIA, 2024; MARTINEZ; ALI, 2024). These standards are widely used to ensure that different healthcare information systems can communicate effectively, promoting interoperability among hospitals, clinics, laboratories, and other healthcare providers (KUMAR; SMITH, 2024).

The standardized framework provided by HL7 enables the exchange of clinical and administrative data, reducing ambiguities and improving accuracy in communication between systems (MARTINEZ; ALI, 2024). The HL7 standards facilitate the integration of various healthcare information systems, such as Electronic Health Records (EHRs), laboratory systems, applications, and APIs (CHEN; GARCIA, 2024).

HL7 facilitates interoperability both at the syntactic level (how the data is formatted) and at the semantic level (what the data means), ensuring that systems understand the exchanged information consistently (KUMAR; SMITH, 2024). Through this effective exchange of information, HL7 standards contribute to better care coordination and more informed clinical decisions (CHEN; GARCIA, 2024; MARTINEZ; ALI, 2024).

2.2.2 Fast Healthcare Interoperability Resources

FHIR is a standard developed by HL7 that establishes a standardized way to represent and exchange health information (NGUYEN; LEE, 2023; CHEN; GARCIA, 2024). By using modern web technologies such as RESTful APIs, JSON, and XML, FHIR facilitates interoperability between healthcare information systems, allowing different applications to communicate efficiently and quickly (MARTINEZ; ALI, 2024).

By enabling the fast and secure exchange of information, FHIR promotes continuity of care and improves communication between healthcare providers (KUMAR; SMITH, 2024). With its modular approach, it allows the integration of various systems, such as EHRs, laboratory systems, and telemedicine applications. Additionally, the use of APIs and modern data formats makes it easier to implement new healthcare solutions, such as mobile apps and real-time data analytics tools (CHEN; GARCIA, 2024). Structured data via FHIR can also be applied in clinical research, aiding in generating insights and improvements in medical practice (NGUYEN; LEE,

2023).

2.2.3 Digital Imaging and Communications in Medicine

DICOM is an international standard that defines the exchange, storage, and presentation of medical images and related information. This standard is crucial for ensuring interoperability between different devices and healthcare information systems, especially in areas such as radiology, cardiology, and other specialties that use medical imaging (PATEL; KIM, 2023).

DICOM provides a standardized format for medical images, such as X-rays and MRIs, allowing different devices and software to interpret and display these images consistently. This standardization enables healthcare institutions to share images and reports efficiently, improving continuity of care and collaboration among professionals (SMITH; LEE, 2023).

Moreover, DICOM facilitates the integration of medical imaging systems with EHRs, providing access to complete patient information. Medical images encoded in DICOM can be used in research and clinical analysis, enabling the development of machine learning algorithms for diagnoses (CHEN; GARCIA, 2024; MARTINEZ; ALI, 2024). The interoperability provided by DICOM improves the efficiency of clinical workflows and reduces the risk of errors related to data exchange (NGUYEN; ALI, 2024; WANG; ALI, 2024).

2.2.4 Systematized Nomenclature of Medicine – Clinical Terms

SNOMED CT is a comprehensive coding system that provides a standardized language for representing clinical concepts in electronic health records. It allows clinical data to be accurately and consistently coded, facilitating communication between healthcare professionals, information systems, and institutions (MEYER; LUTZ, 2021; SANTOS; OLIVEIRA, 2022).

The main contributions of SNOMED CT to health data interoperability come through standardizing terminology and codes, ensuring consistency and quality in the description of diagnoses, procedures, and health conditions, which reduces ambiguities (SMITH; LEE, 2023; KIM; PARK, 2024). The use of SNOMED CT enables the integration of data from different sources, such as hospitals, clinics, and laboratories, facilitating large-scale data analysis. Finally, data coded with SNOMED CT can be used in clinical research, epidemiological studies, and public health analysis, providing insights into trends and patterns (SANTOS; OLIVEIRA, 2022; SMITH; LEE, 2023).

2.2.5 *Logical Observation Identifiers Names and Codes*

LOINC is a standardized coding system that enables the identification and exchange of information on laboratory, clinical, and other healthcare observations (PATEL; KIM, 2023; MARTINEZ; ALI, 2024). This system ensures that health data can be consistently represented, regardless of the healthcare institution or information system used.

LOINC codes cover a wide range of clinical and laboratory tests, measurements, and observations, such as blood pressure readings, cholesterol levels, and radiology reports (CHEN; GARCIA, 2024; PATEL; KIM, 2023). This allows healthcare providers to exchange data about patients' medical conditions in a standardized format, reducing errors and improving the accuracy of diagnoses and treatment decisions (SMITH; LEE, 2023; MARTINEZ; ALI, 2024).

2.3 **Artificial Intelligence and LLM**

Conceptually, LLM is a deep learning model architecture composed of an extremely large number of parameters, often in the billions or even trillions. These models are trained on pre-training tasks such as Masked Language Modeling (MLM) and Auto-Regressive Prediction (ARP) (BASILE *et al.*, 2024). These models are fed massive corpora of text to capture linguistic, semantic, and syntactic patterns (AL., 2024).

The LLM must be able to capture complex semantic relationships and syntactic dependencies in sentences and large texts. This includes the ability to understand anaphoras and linguistic ambiguity, such as pronoun resolution (ROESLER *et al.*, 2024b). The model uses its internal mechanisms to encode the contextual meanings of tokens in a nonlinear way, allowing it to understand not only the literal meaning of words but also their implicit meanings, metaphors, and other high-level linguistic aspects.

The ability to generate fluent and coherent text that mimics the structure and fluency of human language is one of the central metrics for evaluating the quality of an LLM (BASILE *et al.*, 2024). Text generation involves the probabilistic selection of the most appropriate tokens, ensuring not only grammatical correctness but also contextual relevance. This is achieved through the probabilistic modeling of hidden Markov chains within the model's latent space, where transition probabilities between tokens are adjusted by the context and semantic relationships learned during pre-training.

Highly effective LLMs are trained on a wide range of topics and domains, enabling them to maintain sophisticated contextual awareness in knowledge-intensive areas such as medicine, law, and science (AL., 2024). This requires the use of specialized datasets to adapt the model to technical fields. The ability to adjust the generated text to be coherent and relevant in very specific contexts, maintaining terminological and factual rigor, is achieved through techniques such as supervised fine-tuning and supervised instruction.

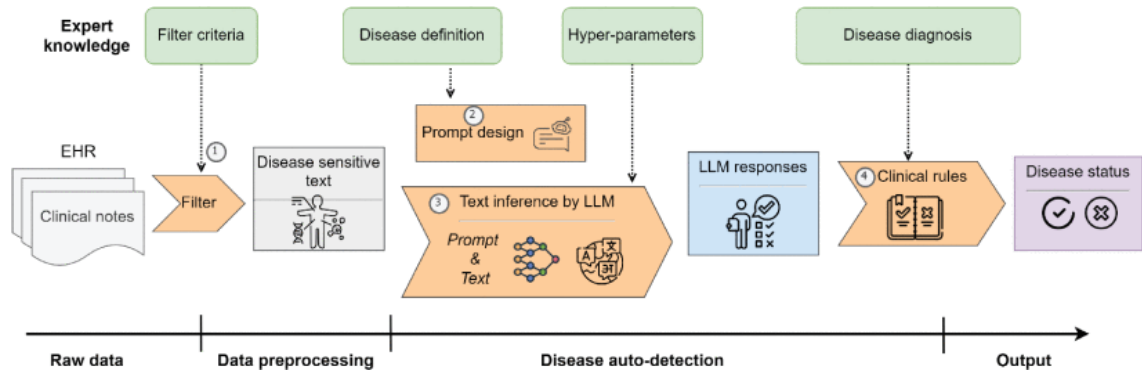
The ability to follow instructions efficiently and accurately, an increasingly desired feature in LLMs, is obtained through methods such as Reinforcement Learning with Human Feedback (RLHF) (ROESLER *et al.*, 2024b). During this process, the model is refined based on human interactions, learning to align its output with the specific intentions and directions provided by human users. This makes LLMs effective in tasks involving problem-solving, decision-making, and executing complex instructions, particularly useful in customer support, assisted programming, and other industrial applications.

LLMs, with their advanced ability to understand, generate, and interpret natural language, have a broad range of applications in the healthcare sector. Due to their ability to handle large volumes of textual data and complex contexts, these models are being implemented in various areas of healthcare to improve care, clinical decision-making, and research (BASILE *et al.*, 2024).

For example, in Clinical Decision Support, LLMs can act as Clinical Decision Support Systems (Clinical Decision Support System (CDSS)), assisting doctors and other healthcare professionals in making more informed decisions, such as text-based diagnosis, treatment suggestions based on clinical guidelines, medical literature, and patient history, and predicting clinical outcomes, as LLMs can process large volumes of data to predict clinical outcomes based on historical patterns, such as the risk of postoperative complications or the probability of hospital readmission (AL., 2024; PAN *et al.*, 2024).

Another application of LLMs is in the role of Virtual Assistants for Patients and Professionals, as they can be used to create intelligent virtual assistants that interact with patients and healthcare professionals (ROESLER *et al.*, 2024b). An important application is in managing health data, where in the Analysis of Electronic Health Records and Natural Language Processing (NLP), LLMs play a crucial role in extracting information from Electronic Health Records (EHRs), which contain vast amounts of unstructured textual data. For example, in automatic coding, LLMs can perform the automated coding of diagnoses and medical procedures

Figure 1 – Example of architecture for disease identification.



from clinical notes, facilitating the generation of classification codes such as International Classification of Diseases (ICD-10) and Current Procedural Terminology (CPT), which are essential for billing and reporting (BASILE *et al.*, 2024). Additionally, in detecting adverse events and clinical alerts: by processing admission notes and exam reports, LLMs can identify early signs of complications or adverse events, such as dangerous drug interactions, and issue alerts to the medical team. Another key point is in summarizing medical records: language models can generate concise summaries of complex electronic health records, highlighting the most relevant points for healthcare professionals during consultations, optimizing time, and facilitating decision-making (AL., 2024).

Still, in data handling, LLMs are important in Automatic Documentation and Writing of Clinical Reports, where they are effective in automating medical reports and clinical documentation, helping to relieve the workload of healthcare professionals and improving the accuracy of documentation. Using another type of file, such as DICOM, we can see its use in automatically generating image reports: LLMs can transform the analysis of medical images, such as X-rays or MRIs, into structured reports, assisting radiologists in quick and accurate documentation. Finally, in transcribing and summarizing medical consultations: using LLMs, it is possible to automatically generate transcriptions of medical consultations and summarize them into clinical notes, integrating this information directly into the electronic health record. This reduces the time doctors spend on administrative tasks and allows more focus on the patient (ROESLER *et al.*, 2024b).

2.4 LGPD and Health Data Protection in Brazil

LGPD is particularly relevant in the healthcare sector, where the processing of sensitive personal data is frequent and critical (Brasil, 2018). Protecting this information is essential to ensure patient privacy and trust in healthcare institutions.

2.4.1 Compliance Requirements and Informed Consent

Informed consent is a central aspect of LGPD, especially in the context of health data ((ANPD), 2021a). In 2023 and 2024, hospitals and clinics must ensure that information about data collection, such as medical histories, exam results, and treatment information, is clearly explained, including how this information will be used, such as for clinical research or scheduling appointments (Brasil. Conselho Federal de Medicina, 2018); also, regarding consent for sharing, informing and obtaining agreement when data is shared with laboratories or other entities for research or treatment purposes, the consent must be explicit, ensuring that the patient is aware of and agrees with this use. For this purpose, National Data Protection Authority (ANPD) has been closely monitoring consent practices, highlighting cases where lack of clarity resulted in penalties ((ANPD), 2021b).

2.4.2 Security of Sensitive Personal Data

Health data is highly sensitive and requires stringent security measures (LEME; BLANK, 2020). For example, a healthcare device might implement encryption systems to protect patient information stored in electronic records. The use of multifactor authentication is another recommended practice to ensure that only authorized professionals have access to this data. In case of a breach, it is essential to have incident response plans that include clear protocols for notifying patients in the event of data leaks, in compliance with LGPD requirements (Brasil. Conselho Nacional de Saúde, 2012). These practices help prevent cyberattacks, which have become more frequent in the healthcare sector, with data being targeted by ransomware and other types of attacks.

2.4.3 Impact of LGPD on the Development of Digital Solutions

The impact of LGPD directly reflects on the development of digital solutions in healthcare, such as health apps. When developing health monitoring apps, such as those collecting

data on eating habits or chronic conditions, companies must incorporate data protection features from the start, such as clear notices about data collection and consent options (ZAGANELLI; FILHO, 2023). Additionally, telemedicine platforms must ensure that patient data, such as diagnoses and prescriptions, are handled in accordance with LGPD. This includes using secure systems that ensure confidentiality of consultations. Compliance with LGPD not only helps protect patient data but also becomes a competitive differentiator, as it increases user trust in using these technologies.

LGPD imposes a regulatory framework that requires healthcare institutions to not only ensure legal compliance but also uphold a strong ethical stance in data handling. Implementing practices that respect patient data privacy and security not only protects patients' rights but also strengthens the integrity of healthcare institutions in an increasingly digital landscape (NETO *et al.*, 2023).

2.5 Applications of Language Models in Healthcare: Advances and Limitations

The use of Large Language Models (LLMs) has been increasingly integrated into various fields, bringing innovation and enhancement to how complex tasks are performed. In the healthcare sector, these models have stood out particularly in tasks such as diagnosis, triage, and clinical support. For example, a comprehensive evaluation conducted by Jahan *et al.* analyzed the effectiveness of LLMs, such as ChatGPT, PaLM, and LLaMA, in biomedical contexts. The results indicate that these models have demonstrated robust performance in processing medical texts, with significant potential to contribute to diagnostics and clinical care optimization (JAHAN *et al.*, 2024).

Moreover, the adaptation of LLMs for specific clinical needs has been a major research focus. In a study focused on mental health, Taylor *et al.* developed LLMs tailored for digital triage, an application that can significantly enhance the personalization of care. By considering the particularities of each patient, these models help improve diagnostic accuracy and efficiency, ensuring a more individualized approach to care (TAYLOR *et al.*, 2024). A similar application of LLMs in medical diagnosis was explored by Song *et al.*, who created PneumoLLM, a model specialized in diagnosing pneumoconiosis. Besides being a valuable tool for physicians, this model also enhances the accuracy of respiratory disease diagnoses from medical images (SONG *et al.*, 2024).

The potential of LLMs extends beyond clinical diagnosis, encompassing care man-

agement and complex data analysis. An innovative study by Chien et al. demonstrated how LLMs can be used to identify overburdened informal caregivers in long-term care settings. The combination of these technologies with attention visualization techniques provides clearer insights into workload patterns, contributing to more efficient human resource management in this sector (CHIEN *et al.*, 2024). In ophthalmology, Xue et al. developed Xiaoqing, an LLM-based question-and-answer system specialized in glaucoma diagnosis. This system not only improves patient and healthcare professional education but also serves as a support tool for clarifying clinical doubts (XUE *et al.*, 2024).

In parallel with these innovations in conventional healthcare, traditional medicine is also benefiting from LLM applications. Tan et al. fine-tuned a dialogue model for consultations on traditional Chinese medicine, creating MedChatZH. This model exemplifies the flexibility of LLMs in various cultural and medical contexts, expanding their application possibilities (TAN *et al.*, 2024). The versatility of LLMs was also highlighted by Wu et al., who combined LLMs with Transformers in a hybrid system for medical diagnosis. By integrating fuzzy models and context selectors, this system improved diagnostic accuracy in hospital settings, making the process more reliable (WU *et al.*, 2024).

Another significant advancement in the LLM field is the ability to handle incomplete data and human emotions. Zhang et al. proposed a multi-teacher knowledge distillation technique for emotion recognition, using LLMs to manage uncertainty and data scarcity in emotional health contexts. This application shows how LLMs can be used to enhance emotional intelligence in healthcare and patient support systems (ZHANG *et al.*, 2024). Additionally, the application of LLMs in fields outside healthcare is also expanding. Sarzaeim et al. explored how these models can be used in smart policing systems to predict crimes, employing few-shot and zero-shot prompting techniques. This opens new possibilities for automating public safety (SARZAEIM *et al.*, 2024).

The impact of LLMs is not limited to diagnostics and security but is also felt in areas such as clinical data analysis and process optimization. Kim et al. developed PhenoFlow, an LLM-based system that facilitates the visual analysis of stroke data, reducing the cognitive load on healthcare professionals and aiding in quick and accurate decision-making (KIM *et al.*, 2025). Similarly, Chen et al. utilized LLMs to optimize energy dispatch in space-air-ground networks, demonstrating how these models can be effectively applied in engineering and resource optimization (CHEN *et al.*, 2024).

In evidence-based medicine, Li et al. conducted a study on the application of LLMs in tasks such as prompt engineering and clinical question answering. These advancements not only aid in diagnostic automation but also improve therapy efficiency, making the medical process faster and more accessible (LI *et al.*, 2024). Furthermore, Yan et al. integrated LLMs with knowledge graphs for the KNOWNET system, a health information search tool that helps minimize misinformation, ensuring safer and more efficient exploration of biomedical data (YAN *et al.*, 2024).

Finally, Bonfigli et al. emphasized the importance of fine-tuning for the success of LLMs in biomedical contexts, demonstrating how these models can be adjusted for specific tasks such as diagnosis and medical triage. This type of adjustment allows LLMs to achieve superior performance, improving the accuracy and efficiency of medical services in various contexts (BONFIGLI *et al.*, 2024). Together, these contributions reveal that LLMs are becoming indispensable tools in multiple fields, with transformative potential for healthcare, security, and beyond, driving significant advancements in how we interact with artificial intelligence.

Authors	Goals	Methods	Highlights	Limitations
Jahan et al. (JAHAN <i>et al.</i> , 2024)	Evaluate LLM performance in biomedical tasks	ChatGPT, PaLM, LLaMA	High accuracy in biomedical tasks	Limited focus on biomedical domain
Taylor et al. (TAYLOR <i>et al.</i> , 2024)	Develop LLMs for digital screening in mental health	LLM fine-tuned for clinical screening	Improvement in personalized care	Need for additional clinical data
Song et al. (SONG <i>et al.</i> , 2024)	Pneumoconiosis diagnosis with LLMs	LLMs in image diagnostics	High accuracy in radiological diagnoses	Dependence on high-quality images
Chien et al. (CHIEN <i>et al.</i> , 2024)	Identify overloaded informal caregivers	NLP, Visual attention	Visualization of workload patterns	Limited long-term application
Xue et al. (XUE <i>et al.</i> , 2024)	Create a Q&A system for glaucoma	Q&A with LLMs	Improvement in medical education	Exclusive focus on glaucoma
Tan et al. (TAN <i>et al.</i> , 2024)	Fine-tune LLMs for traditional Chinese medicine consultations	Fine-tuning, Medical dialogue	Integration of traditional medicine and LLMs	Need for greater generalization
Wu et al. (WU <i>et al.</i> , 2024)	Hybrid medical diagnosis with LLMs	Transformers, Fuzzy selection	Better performance in complex diagnoses	Requires continuous model tuning
Zhang et al. (ZHANG <i>et al.</i> , 2024)	Emotion recognition with LLMs and multi-teacher distillation	Multi-teacher Knowledge Distillation	Accuracy in emotional recognition	Limitation in heterogeneous data
Sarzaeim et al. (SARZAEIM <i>et al.</i> , 2024)	Smart policing with LLMs	Few-shot, Zero-shot prompting	High efficiency in crime prediction	Dependence on reliable criminal data
Kim et al. (KIM <i>et al.</i> , 2025)	Visual analysis of stroke with LLMs	PhenoFlow, Visual analysis	Reduced cognitive load in diagnoses	Dependence on human-AI interaction
Chen et al. (CHEN <i>et al.</i> , 2024)	Coordination for energy dispatch with connected vehicles	LLMs, SAGVNs	Cost optimization and energy decision-making	Complexity in integrated vehicular networks
Li et al. (LI <i>et al.</i> , 2024)	Benchmarking LLMs in evidence-based medicine	Prompt engineering, Clinical NLP	Contribution to automated diagnostics	Need for more clinical samples
Yan et al. (YAN <i>et al.</i> , 2024)	Integration of LLMs with knowledge graphs for health	KNOWNET, Knowledge graphs	Reduction in biomedical misinformation	Need for comprehensive data
Bonfigli et al. (BONFIGLI <i>et al.</i> , 2024)	Adapt LLMs to the biomedical domain	Fine-tuning, Probing	Success in diagnostics and screening	High dependence on specialized data

Table 1 – Summary of references on LLMs in various application areas.

3 METHODOLOGY

In this chapter, the process of developing the tool will be presented, covering the technologies and frameworks used, as well as the capacity and stress tests. The methodology adopted for the development of the web interface and its integration with our multimodal LLMOps pipeline involves the use of open-source LLMs (Large Language Models) and SLMs (Small Language Models), such as Ollama models, Microsoft-Phi, among others. The combination of these language models using Ensemble RAG (Retriever Augmentation Generation) will also be discussed, along with the incorporation of Autonomous Agents based on SLMs. The platform was built with a software architecture combining Flutter for the frontend and Flask API for the backend. The interface was designed to allow users to input text with questions and receive responses generated by the natural language model.

3.1 Technologies Used

The development process of the tool was divided into two main modules: the user interface and the resource server. The user interface, called FrontEnd, is responsible for consuming data provided by the server, which consists of an API and a MySQL database. Communication between the interface and the server was implemented using the Representational State Transfer (REST) standard, a widely adopted model to ensure system interoperability, scalability, and flexibility.

The FrontEnd was developed using the Flutter framework, which supports the development of responsive applications for both web and mobile devices. The choice of Flutter was made because of its ability to create a rich user experience with a single codebase, simplifying system maintenance and evolution.

The resource server, or BackEnd, was developed using Flask API, a lightweight and flexible microframework in Python. Flask was chosen for its simplicity in creating RESTful APIs and its ability to integrate with external libraries, such as machine learning tools. It also facilitates communication with the database and the organization of application routes.

Additionally, the system was designed to ensure the security of user data. The login interface was implemented to validate credentials and verify access to the application, ensuring that only authorized users can use it.

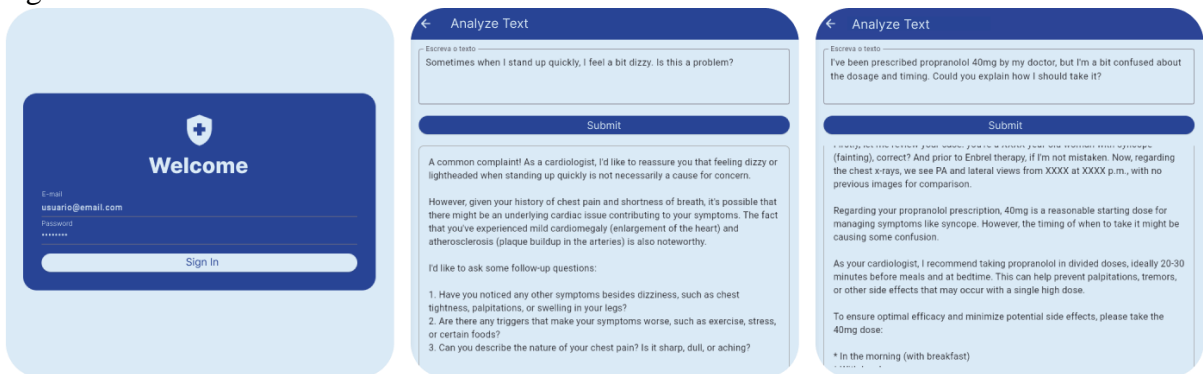
3.2 Web Interface

The navigation flow begins with the login page, which checks the user's credentials before granting access to the application. This process is essential to ensure the security and privacy of the medical data processed by the tool.

- Login Page: Contains two input fields (email and password) and a login button. After the backend verifies the credentials, the user is redirected to the main page of the application, provided the user is authorized.
- Text Input Field: On the text interaction page, the user inputs their question or comment related to the exam.
- Submit Button: Triggers the request to the backend, sending the text for processing.
- Response Area: Displays the response generated by the Ollama model.

The layout was designed with a focus on simplicity and usability, ensuring that interactions are direct and efficient. In figure 6, the web application pages can be seen.

Figure 2 – Web Interface



Source: Created by the author.

In figure 3, we can observe in detail the processing of a text. The analyzed text contained 100 characters.

3.3 Dataset

Our platform was designed to be fed with patient data as they use the platform or through public or private institutions that wish to acquire the system. Thus, the platform supports user registration, report creation, the patient's medical history, scheduled appointments, and other types of patient information.

Initially, we proposed feeding the platform with a dataset obtained from the National

Figure 3 – Text Analysis Interface

The interface is titled "Analyze Text" with a back arrow. It contains a text input field with the placeholder "Escreva o texto" and the question: "How would you interpret an abnormal ECG result with ST elevation in a 50-year-old patient?". Below the input is a blue "Submit" button. The response area shows: "A great question!", a paragraph about the patient's symptoms and history, a paragraph about potential causes of ST elevation, and a numbered list of two potential interpretations: 1. **Acute Coronary Syndrome (ACS)** and 2. **Atrial Fibrillation**.

Source: Created by the author.

Library of Medicine website at the University of India. This dataset contains two tabular data files and 7,470 chest X-ray images in DICOM format, with different orientations (frontal and lateral). The tabular data includes information on various medical reports of patients and their respective diseases. In addition, this tabular data contains complementary information that aids in medical diagnosis, such as: images associated with the patient, indications, comparisons, findings, and impressions.

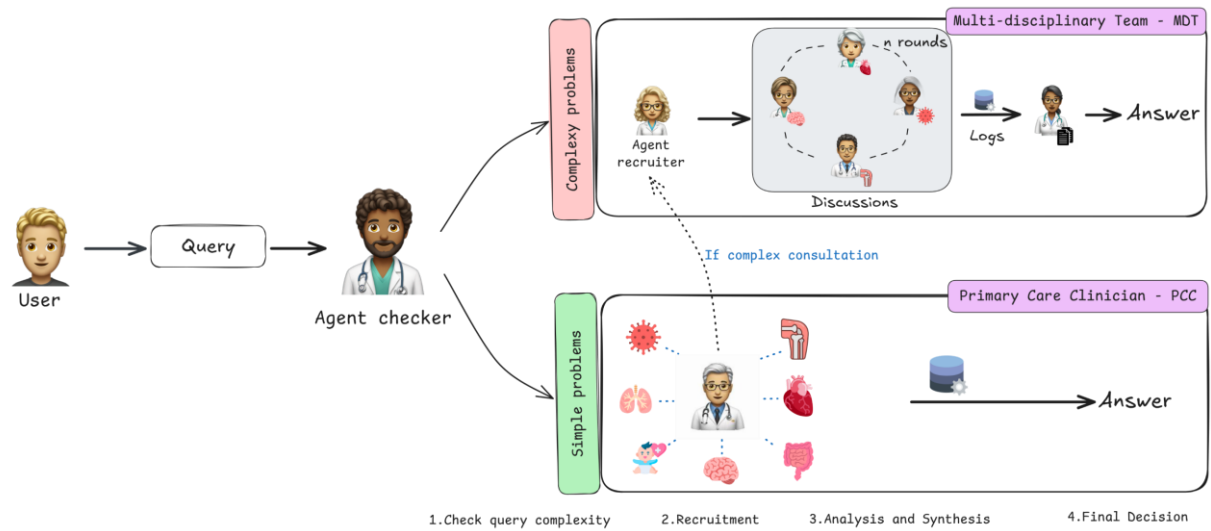
The database used has already undergone basic preprocessing to optimize performance during the processes carried out on it. The images were converted from DICOM to PNG, adjusting pixel values to the 0-255 range and resized to a limit of 2048 pixels on the shortest side.

3.4 Agent Pipeline

The pipeline for medical decision-making proposed by us consists of a collaborative approach involving multi-agent systems, covering four stages:

Sending the query — Through our platform, the user will have access to our workflow

Figure 4 – Summarized process of decision making



Source: Prepared by autor

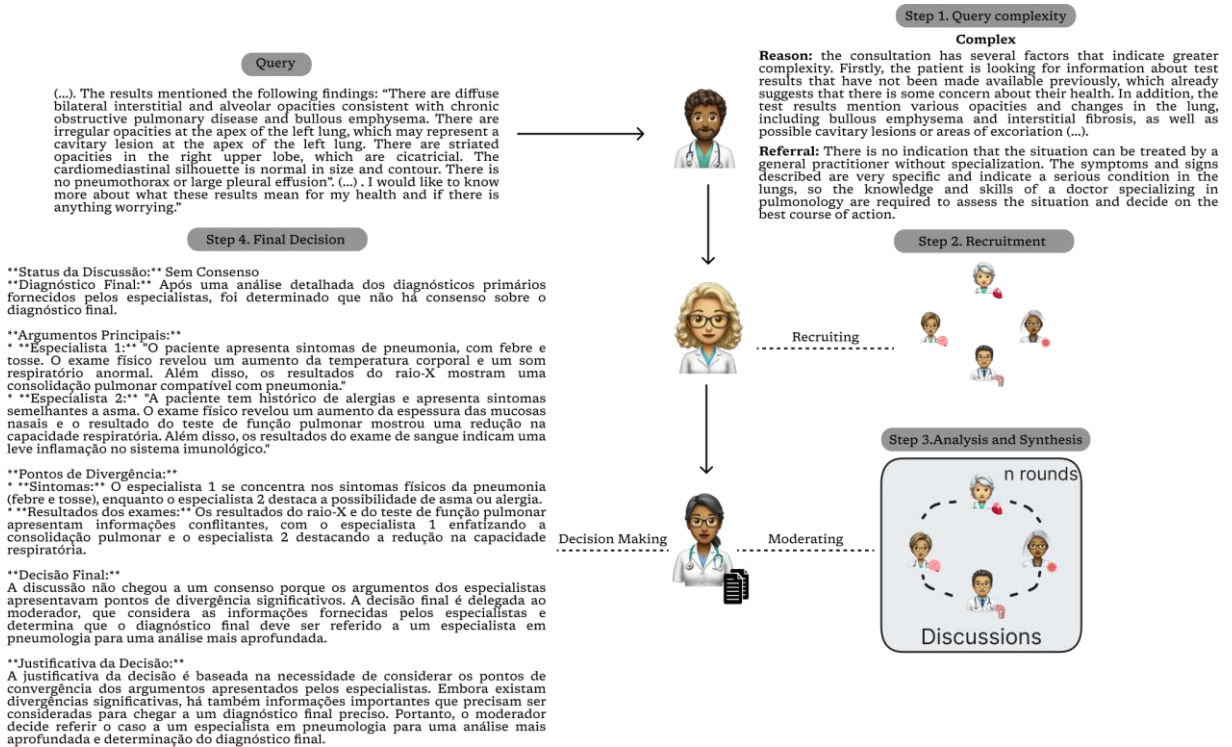
that uses large natural language models to perform a complete decision-making process involving medical experts based on Agents. **Query complexity check** — The system classifies the user’s query into two complexity levels: low and high, based on computational cost and context criteria. **Recruitment of Expert Agents** — Based on the complexity level of the user’s query, the system directs the Primary Care Clinician (PCC) approach for simple problems or uses the Multi-disciplinary Team (MDT) approach for complex problems. **Analysis and Synthesis** — Simple queries use Representational State Transfer (CoT) (WEI *et al.*, 2022; ZHANG *et al.*, 2022; SAHOO *et al.*, 2024), Self-Consistenc (SC) (SAHOO *et al.*, 2024), and the combination of Self-Consistency with Chain of Though (CoT-SC) (WANG *et al.*, 2022) techniques. Complex queries, falling under the MDT approach, involve multiple LLM and SLM agents that form a consensus and make the decision. **Medical Decision-making** — The final stage involves analyzing the user’s query and providing a response based on clinical medical knowledge.

3.5 Multi-collaborative Agent: Medical Decision-making

The entire flow can be followed in detail through the algorithm 1.

Initially, the user submits an initial query via the chat prompt. The text is received by the API via a POST request and sent to the Agent Checker, the agent responsible for evaluating the complexity level of the query, as detailed in the section below. After determining the complexity of the query, it is directed to one of two approaches: PCC or MDT. Regardless of the selected approach, a new check is performed through the General Clinical Agent (CG), in order

Figure 5 – Illustrative example of our collaborative Multi Agent approach to medical decision making using the Chest-X-Ray dataset (reports only)



Source: Prepared by autor

to simulate the process in detail as done in practice (DAMARELL *et al.*, 2020), and is initially triggered to analyze the user's query. After the analysis, the General Clinical Agent, if capable, will respond to the query or forward it to the specialist team.

When the query is sent to the group of specialists, the process of recruiting experts begins through the Recruiting Agent. This agent is responsible for selecting specialists from each area based on the indications provided by the General Clinical Agent (CG). After recruitment, the MDT starts the collaborative discussion process, where specialists from each area will present their technical views on the problem until all or at least most of the expert agents reach a consensus. This process may occur over up to N rounds.

For each round $n \in N$, the Agents A_i , where $i \in 1, 2, 3, \dots, M$, participate in the discussions. If any Agent does not respond, the Moderator Agent will point out the need for the specialist to make their comments. During the discussion session, a turn-based speaking system is adopted. A turn ends when all specialists reach a consensus on an important aspect of the patient's query, although they may not be fully convinced about the specific problem. If there is no consensus and the agents agree to continue, a new round of discussion is initiated based on the previous conversations.

For each round, consensus within the MDT is determined by analyzing and comparing the experts' opinions, which is ensured by the Moderator Agent. In case of disagreement, the Moderator Agent analyzes the current round of conversation and previous discussions, pointing out the facts discussed earlier to each disagreeing agent. If the disagreement persists, the discussion round is closed. The Moderator Agent then summarizes all the information raised by the specialists and makes a final decision.

3.5.1 *Query Complexity*

To analyze the complexity of the medical query requested by the patient, a specialized verifier agent is triggered to direct the query to the appropriate Agentic workflow based on the complexity of the query. The verifier agent is based on an architecture of small natural language models. Tests were conducted using Qwen 2.5 (with its 3 billion parameter version) from Alibaba. The agent receives instructions on how the complexity of a medical query should be defined and classifies it according to the given instructions, assigning the patient's query to one of the complexity levels:

1. **Simple** — The patient's query is classified as simple when the patient has a predictable and well-organized medical history (KILIAN *et al.*, 2020), clear and widely known symptoms, the presence of visual data such as CT scans, X-rays, ultrasound, among others. Direct, clear, well-defined questions with a limited context (up to 50 words) during the consultation.
2. **Complex** — The patient's medical query is assigned to this level due to vague symptoms presented during the consultation, multiple comorbidities or complex and rare conditions, patients with a tendency toward chronic diseases, lack of medical history, or disorganized data. Patients with communication issues, queries with large contexts (over 50 words).

3.5.2 *Medical Agent Roles*

General Doctor — An agent based on large language models that acts as a general practitioner, initially analyzing the patient's query forwarded by the verifier agent and responding if it's a simple issue or directing it to the recruiter agent to recruit a team of specialist doctors if it is a complex question.

Recruiter — An agent based on small language models responsible for recruiting the team of specialist doctors after the prior analysis of the patient's query conducted by the

Algoritmo 1: Medical Decision-making Pipeline

```

1: Function RECEIVEQUERY(userInput)
2:   POST to API with userInput
3:   Receive query text
4:   Return text
5: End
6: Function EVALUATECOMPLEXITY(query)
7:   COMPLEXITY_LEVEL  $\leftarrow$  AGENTCHECKER(query)
8:   Return COMPLEXITY_LEVEL
9: End
10: Function DIRECTQUERY(query, LEVEL)
11:   If LEVEL = "simple"
12:     APPROACH  $\leftarrow$  PCC
13:     GENERALCLINICALAGENT(query)
14:   Else
15:     APPROACH  $\leftarrow$  MDT
16:   Return APPROACH
17: End
18: Function GENERALCLINICALAGENT(query)
19:   Analyze query
20:   If capable then Respond
21:   Else                                     {#Forward to specialist team (MDT APPROACH)}
22:     LIST  $\leftarrow$  RECRUITSPECIALISTS()
23:     DISCUSSWITHSPECIALISTS(LIST, query)
24:   End
25: Function RECRUITSPECIALISTS()
26:   LIST  $\leftarrow$  Recruiting Agent recruits specialists
27:   Return LIST
28: End
29: Function DISCUSSIONS(LIST, query)
30:   For round from 1 to N do
31:     For each Ai in LIST
32:       If Ai commented then Add contribution
33:       Else Notify Ai
34:       If consensus then Move to next
35:       Else If continue then New round
36:       Else End round
37:   End For
38: End
39: query  $\leftarrow$  RECEIVEQUERY(userInput)
40: level  $\leftarrow$  EVALUATECOMPLEXITY(query)
41: APPROACH  $\leftarrow$  DIRECTQUERY(query, level)
42: If APPROACH = MDT then
43:   LIST  $\leftarrow$  RECRUITSPECIALISTS()
44:   DISCUSSIONS(LIST, query)
45: If APPROACH = PCC then
46:   GENERALCLINICALAGENT(query)
=0

```

general practitioner.

Specialist Doctor — These agents are specialists based on a specific context observed by the general practitioner, recruited by the recruiter agent. These specialist agents work in a group with other specialists, discuss among themselves, reach a consensus, and ultimately make the medical decision regarding the patient’s query, complementing with the provision of technical reports on possible anomalies and pathologies related to the patient’s query.

Moderator — This agent moderates discussions, summarizes information, and encourages opinions among the specialist agents in a multidisciplinary team setting. The moderator is mainly responsible in cases of disagreement among the specialists, mediating discussions divided into rounds and speaking turns. The moderator records the MDT meeting report, highlighting all the important information addressed during the meeting.

3.6 Database Description

The databases used in this work were Chest-X-Ray and MedQA. The first dataset is a large two-part dataset, one with tabular data (medical reports) and images. The images in the dataset are from MRI scans of the thorax and the tabular data is the data we are using in this work. While the second dataset is a multi-choice question and answer dataset, this dataset is more focused on correct and well-defined answers while the first is more of a clinical case study. In addition, the first dataset contains very valuable information for clinical cases, which are divided into: patient problem, images (the tests carried out), indication, comparison, findings (of the tests) and impression. It’s worth remembering that not all of these data are present for all patients, so there’s a need to treat these data as well.

3.6.1 Data Cleaning

A deeper investigation was conducted on the tabular data of the addressed dataset, identifying that it is more likely to contain noise. It was observed that additional preprocessing was required, beyond what was applied to the original dataset. To ensure that the Ensemble RAG works effectively, it was necessary to analyze in detail the files *indiana_reports.csv* and *indiana_projections.csv* of the dataset, in order to eliminate the noise present in these data. A traditional data cleaning approach was proposed, distributed in three stages: noise removal, normalization, and stopword removal.

Noise Removal — Removal of special characters (such as: #!@\$%^`_+*), emojis, and meaning-less repeated characters, such as: XXXX and x-XXXX.

Normalization — Conversion of all words in a sentence to lowercase letters.

Stopword Removal — Discarding common or repeated words that do not add real meaning to the sentence, such as “i”, “a”, “and”, “of”, among others.

In addition, during the verifier agent call, data cleaning and structuring in JSON format are applied to facilitate the understanding of the patient’s query for the subsequent stages of medical decision-making.

3.7 Validation Metrics

Considering all stages of the platform and the processes that compose it, three types of evaluations were proposed to assess the system in detail, as well as its parts, at different stages of our platform: 1) System Evaluation e 2) RAG Evaluation.

3.7.1 System Evaluation

In the context of Big Data, the novelty and uncertainties inherent in many language model-based resources are inevitable. Therefore, it is crucial to maintain data privacy standards and social responsibility. Currently, large data platforms and models are trying to apply LLM-LOps (Large Language Models Operations, a niche of MLOps, Machine Learning Operations), being heavily conditioned to keep these platforms running continuously and monitoring their life cycles (WANG *et al.*, 2024).

Thus, it is essential to analyze and measure some aspects of the usage of these large systems focused on natural language models, such as: monitoring the structure of the data provided for training these models, quality of LLMs, SLMs, RAGs, autonomous agents, and the entire life cycle of the models. These analytics enable developers and companies to gain valuable insights into the real-world usage of these technologies, ensuring the reliability and efficiency of LLMs through controlled and automated evaluations.

This work, therefore, focused on online evaluations, which are conducted in production environment scenarios exposed to the real world. This process involves automatic evaluators triggered by new log entries derived from live production. Online evaluation is excellent for reflecting the complexities of real-world use and integrates valuable user feedback, making it

ideal for continuous performance monitoring. The following metric categories were analyzed: Cost metrics and Performance metrics.

Cost Metrics

- **GPU Utilization** This metric aims to measure the GPU usage configured in the platform backend based on the total number of tokens and the number of responses received.
- **LLM Calls Cost** We decided measures the cost price of a commercial API (i.e., Anthropic, OpenAI) or when opting to purchase a machine with excellent configurations to run the full server.
- **Infrastructure Cost** This is an unconventional metric but quite important for the LLM serving deployment context. The cost of your application’s infrastructure becomes a crucial factor in production contexts. The system addressed here has not been tested in production with numerous users using the platform simultaneously and continuously, as it should be. Therefore, this metric aims to demonstrate a base cost for maintaining a high-level LLMOps-based structure in an early production version (2/3 simultaneous users on the platform) for one month. Costs for vector storage (pinecone), backend server (AWS EC2 instance — t3.medium or machine local), storage (logs, patient data, results, LLM models) and network will be evaluated.

3.7.2 RAG Evaluation

For the evaluation of the RAG technique employed in this work, it was initially divided into two key conventional areas widely used in RAG systems and another alternative to conventional methods. Evaluations based on RAG systems are made through retrieval evaluation (ES *et al.*, 2023), generation evaluation (ES *et al.*, 2023), and finally, evaluation by an LLM-as-a-Judge (ZHENG *et al.*, 2023).

Evaluation Retrieval is a crucial step to identify the potential and robustness of the system during the generation of responses based on the retrieval of documents stored in the database. Nowadays, two key metrics are employed to measure these aspects: context relevancy and context recall.

- **Semantic Textual Similarity (STS)** This metric uses modern language models to generate vector representations (embeddings) to compare the semantic similarities between the retrieved documents and the user’s query. The comparison between the embeddings is performed by calculating cosine similarity (RAHUTOMO *et al.*, 2012). The similarity

formula is as follows:

$$\text{CosineSimilarity} = \frac{\vec{v}_1 \cdot \vec{v}_2}{||\vec{v}_1|| ||\vec{v}_2||} \quad (3.1)$$

Evaluation Generation is the second step after retrieving data from the database. Once the correct information is retrieved from the patient's query, it is necessary to evaluate not only if the natural language model was able to respond to the query, but also how clear and accurate the response is. Therefore, it is crucial to evaluate these key points.

- **BLEU** is a metric that compares n-gram precision, a sentence of n words, and applies penalties to sentences that are either too long or too short, known as the brevity penalty. It takes the geometric mean of the modified precision scores from the test corpus and then multiplies the result by an exponential brevity penalty factor, BP . First, the geometric mean of the modified n-gram precisions, p_n , is calculated using n-grams up to length N and positive weights w_n summing to one. Then, let c be the length of the candidate translation and r the effective length of the reference corpus (PAPINENI *et al.*, 2002). Initially, the calculation of BP is formally described by:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases} \quad (3.2)$$

Then, from the equation of BP , we obtain the equation for the BLEU metric:

$$BLEU = BP * \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (3.3)$$

- **Rouge** or Recall-Oriented Understudy for Gisting Evaluation is actually a set of metrics, widely used for automatic summary evaluation, meaning it focuses more on recall, the ability to retrieve n-grams present in the reference. The set of metrics compares a summary or translation automatically produced with reference summaries or translations (gold-standard high-quality, human-produced). For this work, ROUGE-1 (unigram), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence - LCS) were used (LIN, 2004). The calculation for ROUGE-N, for n-grams, is given by:

$$ROUGE - N = \frac{\sum_{S \in \text{ReferenceSummaries}} \sum_{gram_n \in S} \text{Count}_{match}(gram_n)}{\sum_{S \in \text{ReferenceSummaries}} \sum_{gram_n \in S} \text{Count}_{gram_n}} \quad (3.4)$$

Meanwhile, ROUGE-L differs from ROUGE-N, as it is based on the longest common subsequence (LCS) between the model's output and the reference, i.e., the longest sequence

of words shared between both. A longer shared sequence should indicate greater similarity between the two sequences (LIN, 2004). The formal notation of this metric is:

$$F_{lcs} = \frac{(1 + b^2)R_{lcs}P_{lcs}}{R_{lcs} + b^2P_{lcs}} \quad (3.5)$$

- **BERTScore** is a widely used metric based on modern natural language models, such as BERT(DEVLIN, 2018). Unlike n-gram-based approaches like Rouge(LIN, 2004) and BLEU(PAPINENI *et al.*, 2002), it aims to analyze semantic representations by generating what we call word *embeddings*, obtained from pre-trained language models or fine-tuned models, in order to calculate the similarities between the generated response and the gold-standard (ZHANG *et al.*, 2019). In this work, we chose to use the "bert-base-uncased" model and recorded the metrics calculated for recall, precision, and F1.
- **Perplexity** is a key metric widely used in probabilistic language models. It is used to evaluate how well the language model can predict a sequence of words in a response, i.e., it measures the perplexity or confusion of the model during the generation of the next token (BENGIO *et al.*, 2000). Perplexity is calculated as follows:

$$Perplexity = 2^{H(p)} = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i | w_1, \dots, w_{i-1}) \right) \quad (3.6)$$

Where $P(w_i | w_1, \dots, w_{i-1})$ is the predicted probability for the word w_i given the previous sequence.

LLM-as-a-Judge is one of the most commonly used unconventional techniques for evaluation, offering two key benefits: scalability and explainability (ZHENG *et al.*, 2023). In the context of Big Data, large volumes of data are routinely stored in data lakes, making manual annotation unfeasible in these scenarios. Therefore, the need arises to train LLMs to act as judges and automatically assess how accurate and precise the responses from other LLMs are. Additionally, LLM-as-a-Judge provides not only scores but also explanations, making their results interpretable (ZHENG *et al.*, 2023), as shown in 7.

As used in other stages of this work, we employ this technique to analyze the response from our RAG-based system in order to generate analytics about our system. It was proposed to use an LLM-as-a-Judge for single-response classification with a few-shot motivational prompt, meaning the LLM judge is asked to directly assign a score to a single response. The prompt used for this scenario is shown in ???. Few-shot prompts use examples to help improve classification

response accuracy, usually between 20 and 30 examples to base the analysis on, and then compare the results provided by the prompt with those generated by the system to analyze if they align. After the analysis, it returns feedback, scoring from 1 to 4 on how accurate the response is when compared to the gold standard. Additionally, it calculates the Pearson correlation between the response generated by the LLM and the gold standard. The classification levels generated by the LLM Judge are:

1. The system's response is awful: completely irrelevant to the question asked or highly biased;
2. The system's response is mostly unhelpful: does not address some important aspects of the question;
3. The system's response is mostly helpful: provides support but can still be improved;
4. The system's response is excellent: relevant, direct, detailed, and addresses all the issues raised in the patient's query.

3.8 BackEnd

The backend of the application was developed using Flask, a lightweight framework that allows for the rapid creation of APIs. One of the central features of the API is the endpoint `/api/analyze_text`, which processes user text queries using the Ollama model in conjunction with a Retrieval-Augmented Generation (RAG) approach.

- Prompt Template: The model's prompt is configured to provide context and guide the model's behavior. The template includes information about the model's role, for example, "You are a very experienced cardiologist. You need to answer the question related to a patient's symptoms."
- RAG (Retrieval-Augmented Generation): To improve the quality of the responses, we use a RAG approach. This technique combines document retrieval with text generation. The implementation loads a database of medical reports from the dataset `indian_projections.csv` (an open collection of chest X-rays from Indiana University).
 - The data was pre-processed to optimize performance by converting DICOM images to PNG, adjusting pixel values to 0-255, and resizing the images to a maximum of 2048 pixels on the shortest side.
 - The pipeline loads these data as vectors using the embedding technique generated by the Hugging Face model "BAAI/bge-small-en." The embeddings are indexed with

FAISS, an optimized library for vector search, enabling the application to quickly retrieve the most relevant documents based on the user's query.

- Query Pipeline: The pipeline performs retrieval and text generation in parallel. The most relevant information is retrieved from the database (using FAISS), and the context is then combined with the Ollama language model, which generates a final response based on the query.
- Result Storage: All queries, along with execution time, the model used, and generated responses, are saved in a CSV file for auditing and further analysis.

3.8.1 Credential Verification

Before accessing the text analysis functionality with the Ollama model, users go through a verification process on the login page. This process ensures that only authenticated users can interact with the tool. The endpoint `/api/login` receives the user's credentials (email and password) via a POST request, and the data is verified against a user database stored in MySQL.

The password is verified against a stored hash, and, upon success, the user's information, such as ID and email, is returned in the response. If the credentials are incorrect, an error message is sent. Only after successful authentication can the user access the analysis and query functionalities using the Ollama model with RAG. This mechanism ensures security and access control for the tool.

3.8.2 Architecture

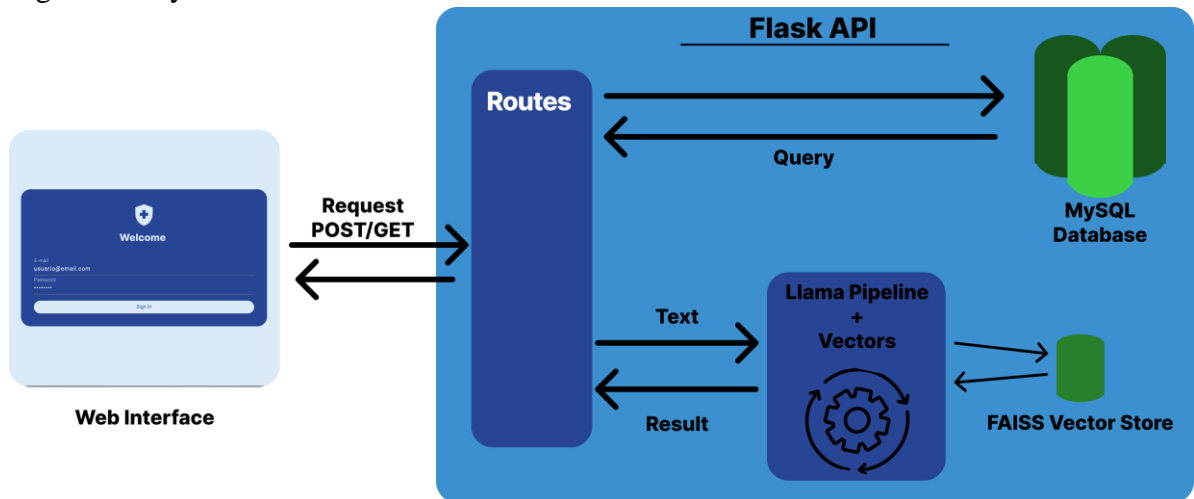
The backend architecture was structured in a REST model, and the figure below shows the graphical representation of the backend. The architecture follows a modular pattern, with specific functions for loading vectors, pipelines, and processing. The separation between the model logic, vectors, and endpoints contributes to a layered system, a characteristic of the REST model. The architecture is marked by several components: Flask API, Model Layer, Vector Database, Pipeline Process, and Result Storage.

- Model Layer (Llama): Performs the text analysis using a model called Llama. The model uses a prompt template that guides the model's response.
- Vector Database (FAISS + .csv Files): A repository that stores vectors for semantic similarity based on pre-processed data. The vectors are loaded from the configured path `(/modelo_llama/vectors/llama3_indiana-reports_.CSV)` and `(/modelo_llama/vectors/llama3_medQ`

and the vectors generated with FAISS are saved.

- Pipeline Process: The pipeline process involves executing the text analysis with the LLM. The Vector Retriever retrieves relevant documents from the vectors, inserts the data into the template, and invokes the model, returning the response to the client.

Figure 6 – System architecture



Source: Elaborated by the author.

3.8.3 Libraries and Tools Used

Among the key libraries for the development of the endpoint responsible for the model integration, we have Langchain, FAISS, Hugging Face Embeddings, and Ollama.

- Langchain: Used to create modular and scalable pipelines for retrieval and text generation. Langchain allows for easy connection between the natural language model and document retrieval with FAISS.
- FAISS: A library developed by Facebook for high-performance similarity search. In this project, FAISS was used to index the embeddings of the medical reports and retrieve relevant documents quickly.
- Hugging Face Embeddings: The embeddings were generated using the "BAAI/bge-small-en" model, optimizing query processing in an environment with or without a GPU.
- Ollama: The natural language model that answers user queries, using a specific model (llama3) trained to generate high-quality natural language responses.

4 RESULTS

This chapter is divided into two parts. The first part addresses the results related to our pipeline, which integrates multi-Agents based on LLMs/SLMs collaborative approach, while the second part focuses on analyzing the robustness of the platform, from the perspective of the application infrastructure and the performance tests conducted in load and stress scenarios. The application snapshots were taken in a local development environment, using Python 3.11, version 0.3.12 of Ollama, and a client-server architecture based on Flask 3.0.3 for the backend and Flutter 3.22.3 for the frontend.

4.1 Agent-LLM Based Judge: Is it a Good Judge?

The use of assistants based on large language models, initially proposed by Zheng *et al.* (ZHENG *et al.*, 2023), has emerged as a promising practice. These language models can perform various functions, ranging from conversational interactions to serving as medical specialists. Although there are several benchmarks for evaluating language models, most focus on closed questions with short answers, requiring well-defined reference standards. In situations where reference answers are unavailable, traditional evaluation metrics, which rely on the similarity between results and reference answers, such as ROUGE (LIN, 2004) and BLEU (PAPINENI *et al.*, 2002), become ineffective. The LLM Judge drastically reduces the need for human intervention in evaluations, enabling scalable benchmarks and rapid iterations. Additionally, LLM-based judges not only provide scores but also explanations, making their results interpretable, as illustrated in Figure 7.

Therefore, to evaluate the robustness and accuracy of our LLM-based Agent Judge, we calculated the Pearson correlation metrics to measure how close or correlated the result of the final answer classification proposed by our Agent Judge was in relation to the correct answer, the hit rate of correct classifications (all correct classifications in relation to all classifications) and we measured the error related to the Pearson correlation values. To do this, we performed a calculation involving *MAE* and *RMSE*, which measure the average difference between the correct classification and the classification predicted by our Agent and the other calculates the average difference between the predicted value and the actual value, respectively. The results of these metrics can be viewed on table ??.

The results of all results can be summarized through the following graphics presented

Figure 7 – Evaluations, containing the classification and the explanation of the classification, by two reviewers regarding how correct the answer is in relation to the question. Additionally, the feedback from the Agent Judge containing the classification and the explanation of the classification.

Q: In what ways does covid 19 increase the risk of domestic violence?

A: Violence against women is a grave violation of women's human rights, no matter when, where, or how it takes place. There are many forms of violence against women, which have many potential negative health consequences for women and their children. Violence against women can result in injuries and serious physical, mental, sexual and reproductive health problems, including sexually transmitted infections, HIV, and unplanned pregnancies. In extreme cases, violence against women can result in death. The effects of violence are very often long-lived. Violence - in all its forms - can have an impact on a woman's well-being throughout the rest of her life. This is unacceptable, and the risks of violence that women and their children face during the current COVID-19 crisis cannot be ignored. When we are able to prevent violence, or to support women survivors of violence, we help to safeguard women's human rights, and promote physical and mental health and well-being for women throughout their lives. This also helps to alleviate pressure on already stretched essential public services, including the health system.

Reviewer 1:

Rating:
Bad

Explanation:
Condemns violence against women but doesn't explain why the risk increases during the pandemic.

Reviewer 2:

Rating:
Bad

Explanation:
This answer does not provide any information on ways Covid-19 can increase the risk of domestic violence on women. Instead, it provides explanation and information on the impact and results of violence against women on their mental and physical health.

Mixtral-8x7B-Instruct:

Total rating:
Bad

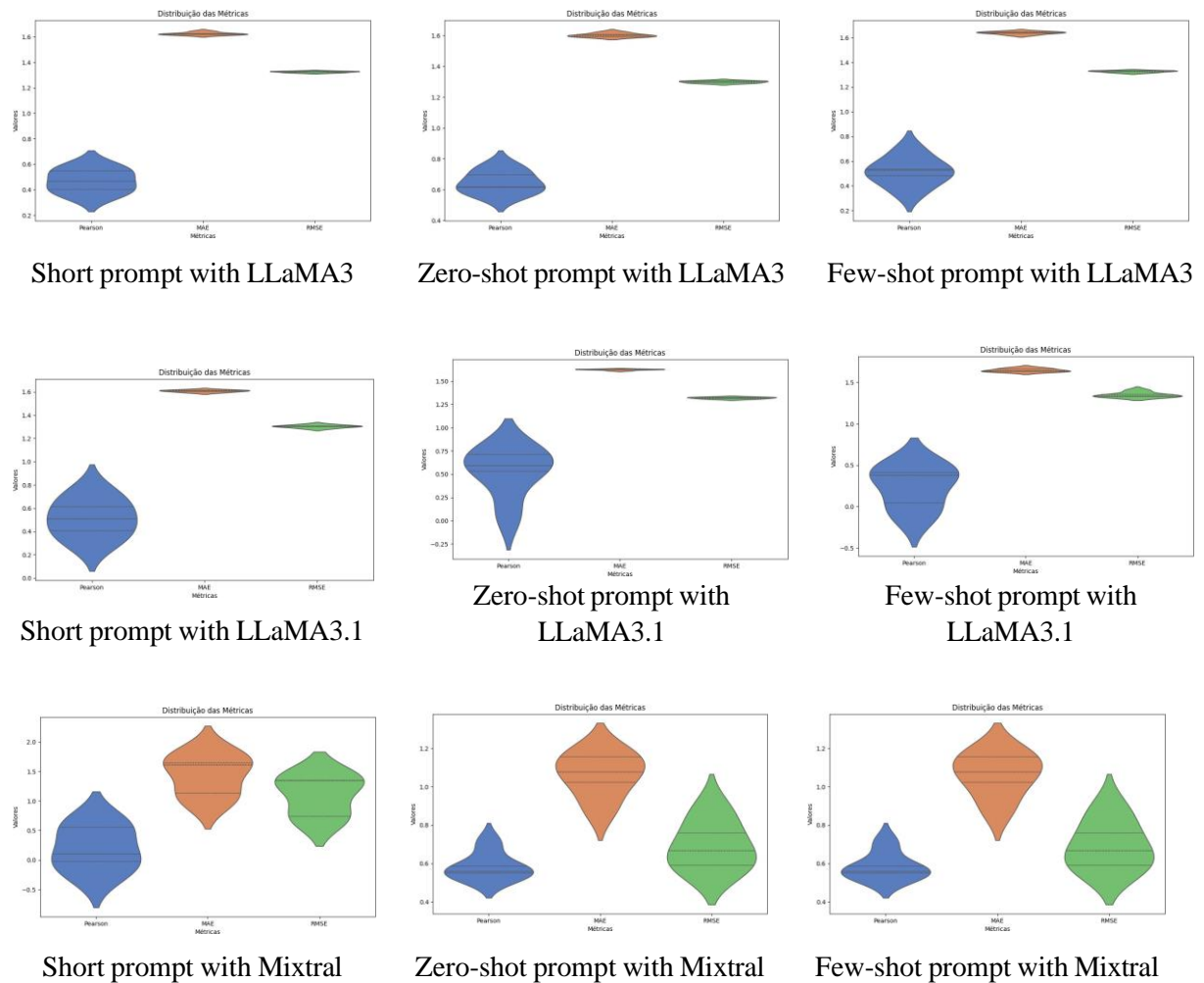
Evaluation:
The system's response is terrible: completely irrelevant to the question. The question asks about the ways in which COVID-19 increases the risk of domestic violence, but the system's response does not address this at all. Instead, it focuses on the general issue of violence against women, which is not specific to the current pandemic. The response is also excessively long and repetitive, making it difficult to extract any useful information."

Source: Prepared by autor

on figure 8.

The violin-shaped graph can provide important information about the behavior of the model during the evaluation of our LLM-based judgment agent. The element of the graph with the larger "belly" represents that there is a set of close values, representing low error variability (MAE and RMSE), accompanied by maximum and minimum values delimited by the lower and upper tail of the violin, which generally indicates an optimal signal when accompanied with the position of this element (violin) on the graph. The figure 8 in question is able to provide valuable information about our model instructed through prompts (exemplified in section 4.5) using various prompt engineering techniques, such as: Few-shot, Zero-shot and Short. The figure provides a visualization of the distribution of the values obtained from the metrics calculations (Spearman correlation, MAE, RMSE).

Figure 8 – Graphic results of values for Pearson correlation and Hit Rate metrics and RMSE and MAE errors.



Source: Prepared by autor

Looking at the graphs in the figures, we can see a high degree of error variability when comparing the LLaMA3 and LLaMA3.1 models in relation to Mixtral. On the other hand, the Mixtral model showed a certain superiority in relation to Pearson's correlation metric, i.e. it can be inferred that the Mixtral model, despite obtaining a higher value for the correlation metric and, as observed in the table (cite table), the hit rate, shows a lot of inconsistency in its classifications, which can be very harmful in medical contexts. The LLaMA models as well as Mixtral showed Pearson correlation values very close to 100%, however, these were accompanied by slightly variable error values, which makes a model consistent, but not necessarily the best, because unlike Mixtral, the error values are higher. I believe that the great impasse in the choice of model and for our analysis would be to know which would be the best trade-off between consistency and trends.

4.1.1 *Intra-class correlation coefficient (ICC)*

A quantitative intra-class analysis was performed, with the calculation of the Intra-class Correlation Coefficient (ICC), due to the central role it plays in the analysis of consistency and reliability between reviewers and the Agent-LLM Judge, especially when multiple individuals or systems evaluate the same issues. The goal is to quantify the degree of agreement and reveal the accuracy of evaluations in subjective judgment tasks, such as the analysis of answer quality, evaluation scores, or any other type of qualitative classification between the Agent-LLM Judge and the reviewers (ZHENG *et al.*, 2023; CHATURVEDI; SHWETA, 2015).

The ICC analysis in this context seeks to answer fundamental questions such as: "To what extent do the model and the reviewer evaluate consistently?" and "Could the averages of these evaluations be trusted as a final judgment metric?". Each type of ICC provides distinct and complementary information about the consistency and reliability of the evaluations made.

4.1.1.1 *Types of ICC analyses*

ICC1 — Absolute Reliability of Individual Evaluators

ICC1 measures the reliability of individual evaluators in absolute terms, i.e., it assesses how consistent each evaluator is individually when judging the items. In contexts where evaluators are unique and the rating scale is crucial for judgment, it is essential to determine whether the evaluators maintain standardization in their evaluations.

ICC2 — Reliability of Random Evaluators

The ICC2 metric is used to estimate reliability when it is assumed that the evaluators are representative of a larger and random set of evaluators, that is, when the identity of the evaluators is not relevant to the consistency of the judgment. This metric indicates whether the observed consistency would be maintained if other reviewers replaced the LLM or the human.

ICC3 — Reliability of Fixed Evaluators

ICC3 is used when the consistency of specific evaluators (in this case, the Agent-LLM Judge and the human reviewer) is of interest and there is no expectation to replace these evaluators. By measuring reliability between fixed evaluators, ICC3 allows understanding the consistency of evaluations between these specific evaluators.

ICC1k, ICC2k, and ICC3k — Mean Metrics of Evaluations

The index k in ICC1, ICC2, and ICC3 indicates the average ICC when there is a combination of

reviewer evaluations, i.e., it is the scenario where the average of evaluations is used as a final measure, which can be useful in cases of high variability between individual evaluations.

We performed the calculation by simulating 3 different LLMs, all with the same number of parameters (llama3 — 8B, llama3.1 — 8B, and Mixtral — 8B) and open-sourced, initially designing 3 prompt techniques: few-shot, zero-shot, and short. The example of each prompt can be seen in Figures 23, 21, and 20, respectively. As mentioned earlier, the ICC calculation was performed only on a subsample of 28 questions, which reflect the agreement between the human reviewers themselves (first and second). In this scenario, an Agent-LLM Judge is introduced to infer its judgment based on this subsample. One of the two human reviewers is removed, and a quantitative analysis is performed based on 6 intra-class scenarios (ICC1, ICC2, ICC3, ICC1k, ICC2k, ICC3k) between the Agent-LLM Judge and the human reviewer.

Table 2 – Quantitative intra-class analysis involving both reviewers and their respective feedback on answer quality.

Judge	Prompt	ICC1	ICC2	ICC3	ICC1k	ICC2k	ICC3K
Mixtral	short	0.136	0.164	0.182	0.320	0.370	0.400
	zero-shot	0.280	0.353	0.507	0.539	0.621	0.755
	few-shot	0.224	0.323	0.522	0.463	0.588	0.766
LLaMA3	short	0.234	0.331	0.532	0.479	0.597	0.773
	zero-shot	0.244	0.340	0.551	0.492	0.607	0.786
	few-shot	0.218	0.321	0.531	0.456	0.587	0.772
LLaMA3.1	short	0.521	0.525	0.537	0.766	0.768	0.777
	zero-shot	0.246	0.343	0.560	0.495	0.611	0.792
	few-shot	0.599	0.597	0.589	0.817	0.816	0.811

Source: Prepared by autor

As shown in Table 2 above, the Agent-LLM based on LLaMA3.1 achieved the best intra-class results in all scenarios involving ICC calculation. These indicators demonstrate the robustness of this model, making it fully viable to use it as a judge to assess complex questions. The first scenario analyzes the absolute reliability of individual evaluators, and the results obtained confirm the relatively high degree of reliability of our Agent-LLM and the synchronization maintained between it and the human reviewer. This reflects the accuracy and consistency of both in evaluating each question properly. The second scenario analyzes the reliability of randomly selected evaluators, showing high values for the LLaMA3.1 model. Unlike the Mixtral and LLaMA3 models, the order of who evaluates, whether it is the Agent-LLM Judge or the human reviewer, does not affect the consistency and accuracy of the results. The

third important scenario analyzes the reliability of fixed evaluators, i.e., in settings where the evaluators and the Agent-LLM Judge do not change. As seen in the results table above, the results remained consistent with scenarios 1 and 2, emphasizing high reliability, consistency, and accuracy in this intra-class evaluation. Finally, the indicators from the last three scenarios confirm that the average of the evaluations is a robust measure, providing a reliable consensus metric.

Quantitatively speaking, these analyses have a significant impact on speed, reliability, and scalability, as an Agent-LLM Judge has proven to be more capable of evaluating questions properly than human reviewers specialized in the field. The two tables 3, 4 below show the same calculations performed but comparing the two human reviewers based on the complete dataset, which contains 5660 samples of questions, answers, evaluations from each reviewer, and their explanations of their evaluations (LI *et al.*,).

Table 3 – ICC calculation results for two human reviewers.

Method	ICC1	ICC2	ICC3	ICC1k	ICC2k	ICC3K
Intra-reviewer	0.483	0.486	0.493	0.651	0.654	0.660

Source: Prepared by autor

As shown earlier, these results demonstrate greater variability compared to those of the Agent-LLM Judge, proving that at least one of the two human reviewers is in disagreement with the other, showing inconsistency in the responses. Next, the calculation involving the following metrics: Pearson Correlation, RMSE, and MAE is observed.

Table 4 – Results table using the Pearson correlation, RMSE, and MAE metrics, based on two human reviewers.

Method	Pearson	RMSE	MAE
Intra-reviewer	0.563	1.165	0.765

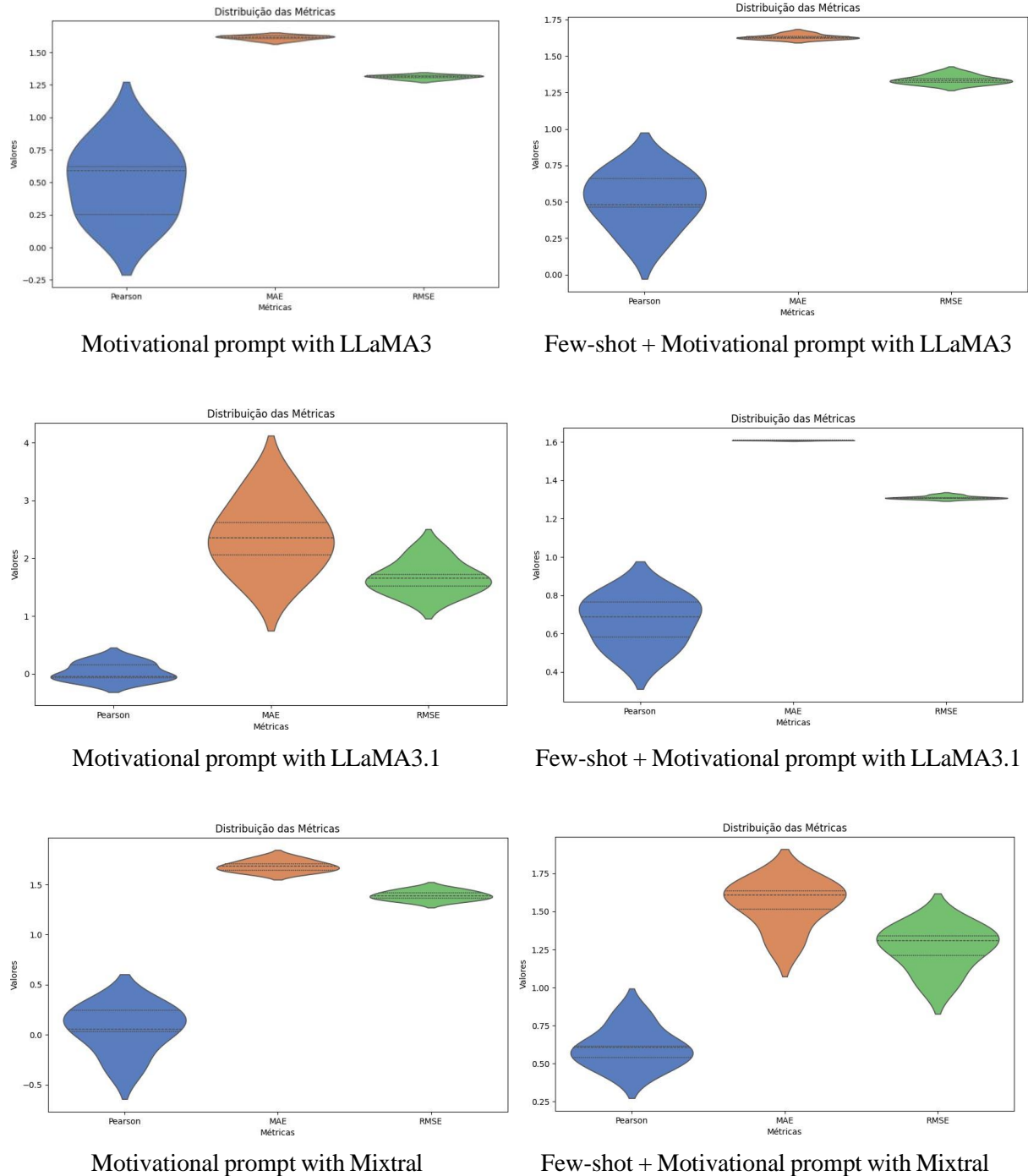
Source: Prepared by autor

4.2 Motivation is all you need

An additional study proposed in this work focused on improving the accuracy of LLM responses through prompt engineering, a technique that blends knowledge from four major areas that define prompt engineering: the combination of linguistics, creative writing, artificial intelligence, and cognitive psychology (SINGH *et al.*, 2024; SAHOO *et al.*, 2024). In particular,

during experimentation and the use of prompt engineering techniques (SINGH *et al.*, 2024; SAHOO *et al.*, 2024; WEI *et al.*, 2022; ZHANG *et al.*, 2022; WANG *et al.*, 2022), it was observed that the final response of Agent Judge based on LLM, the responses significantly improved in cases where creative prompts were used. Some results obtained from these experiments can be seen in figure 9 and table 5 below.

Figure 9 – Distribution of metrics results for Motivational technique of engineering prompt.



Source: Prepared by autor

The figure above graphically presented of figure 9 formalizes the potential of the results obtained from combining traditional prompts with the addition of creative text, the motivation, generating consistent and accurate responses. Figure 9 shows that one of its results reached nearly 100% for the correlation calculation using the Spearman method, which alone demonstrates the potential of the technique. Furthermore, notably, the errors (MAE, RMSE) are low, and the success rate exceeds those seen with traditional prompts and advanced prompt engineering techniques such as CoT and SC-CoT.

Table 5 – Results table using the Pearson correlation, RMSE, MAE, and Hit Rate metrics for two types of prompts: motivational prompt and motivational combined with few-shot prompt.

Judge	Prompt	Pearson	RMSE	MAE	Hit Rate
Mixtral	motivational	0.268	1.628	1.364	9/28
	few-shot + motivational	0.806	1.686	1.408	12/28
LLaMA 3	motivational	0.561	1.599	1.296	14/28
	few-shot + motivational	0.413	1.641	1.346	13/28
LLaMA 3.1	motivational	0.205	2.449	1.567	13/28
	few-shot + motivational	0.678	1.604	1.289	12/28

Source: Prepared by autor

As observed in the table above, excellent results were obtained, especially when considering the set of metrics from using the motivational prompt or the combination of few-shot + motivational prompt compared to the purely traditional prompts (zero-shot, few-shot, short) discussed earlier.

The use of creative writing can elevate the standards we know today in the development of generative artificial intelligence. Using motivation to inspire language models sparks interest, making it worthwhile to explore this prompt engineering technique in a detailed and thorough manner. It is worth conducting experiments testing various contexts of questions (QA, queries, clinical exams, among others) and other domains beyond healthcare.

Table 6 – Quantitative intra-class analysis involving both reviewers and their respective feedback on answer quality using motivational and combination of motivational and few-shot prompts.

Judge	Prompt	ICC1	ICC2	ICC3	ICC1k	ICC2k	ICC3K
Mixtral	motivational	0.176	0.281	0.453	0.391	0.539	0.713
	few-shot + motivational	-0.346	0.056	0.529	-3.358	0.150	0.771
LLaMA 3	motivational	0.249	0.343	0.548	0.500	0.610	0.784
	few-shot + motivational	0.223	0.324	0.531	0.463	0.590	0.772
LLaMA 3.1	motivational	0.271	0.353	0.533	0.527	0.620	0.774
	few-shot + motivational	0.249	0.344	0.555	0.499	0.611	0.789

Source: Prepared by autor

However, unlike the ICC results obtained from responses generated by traditional methods, creative prompts (using motivation) showed some bias. Using the LLaMA3.1 model, we obtained the best results in all scenarios when compared to other models (LLaMA3 and Mixtral). Although we did not achieve unanimous results in all scenarios, both approaches, motivational and few-shot + motivational, were impressive, with a slight advantage for the motivational-only prompt, as it won 4 out of 6 scenarios, even though it did not outperform the few-shot-only approach. The response using the motivational prompt was more robust in the ICC1, ICC2, ICC1k, and ICC2k scenarios, while the combined prompt approach performed better in the ICC3 and ICC3k scenarios. Based on each metric, it can be concluded that the first approach is superior, with a high degree of reliability in the Judge-based Agent and maintaining synchronization between randomly chosen human reviewers, as also shown in the superior ICC2 scenario, confirmed by the ICC1k and ICC2k scenarios.

The second approach showed greater robustness only in the third and last scenario, which concerns when both human evaluators and the Agent are fixed. Overall, the first approach proved to be more robust, not only in the number of scenarios it won but also in real scenarios that more closely reflect ICC1 and ICC2, where reviewers are chosen randomly.

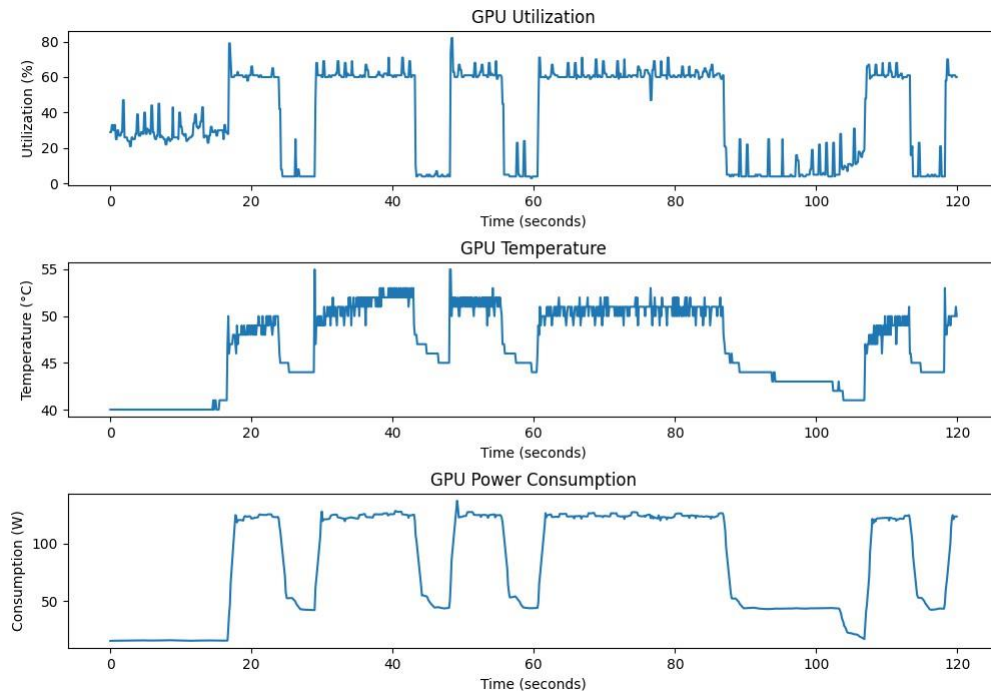
4.3 System Evaluation

The evaluation of the system was based on some of the key factors to consider when trying to build a platform like this from scratch. We divided the analysis of this section into 3 parts, 1) Analysis of the use of the GPU during the execution of our pipeline, as well as monitoring its temperature and power consumption, 2) Analysis of the costs of the most popular commercial APIs for large language models and 3) Costs related to the operational infrastructure as a whole (use of data storage services, vector storage, network and others).

4.3.1 GPU Utilization

When creating a high-level platform, it is essential to consider a good GPU for running the experiments, especially when the idea is to maintain a local server. Therefore, we decided to list some important information analyzed in real time during a 2-minute pipeline execution section. Data was collected every second during the execution period, measuring factors such as: GPU usage (%), GPU temperature (°C) and GPU consumption (W).

Figure 10 – GPU stress utilization during workflow of MDT pipeline



Source: Prepared by autor

Figure 10 shows in detail the behavior of our GPU that was used during the execution process. It is important to note that the GPU used was the Nvidia RTX 3070 TI 12GB model. With regard to temperature, we had sharp increases in temperature at approximately the same time as we had high levels of GPU consumption, which would be the expected behavior and, consequently, the power consumption figures also follow the same behavior.

Although our GPU model isn't the best, we managed to run the entire pipeline, with the Ollama server running the Llama3.1 model, integrating the agents into the entire workflow without using the entire GPU, only at a few points in the execution did we have execution problems (we got close to 100%), which represents the start of the generation of responses between the agents. The behavior of the graph visually resembles a step function, where the beginning represents the start of an agent's execution, hence the "almost periodic" behavior, as it is the behavior of an agent's response generation, then the time it takes to send the response from one to the other until it starts generating it, and so on until we finish the flow. Therefore, we were able to show that our entire pipeline can be run even with a video card that is not even in the top 5 video cards intended for this type of experiment, indicating the cost-effectiveness of

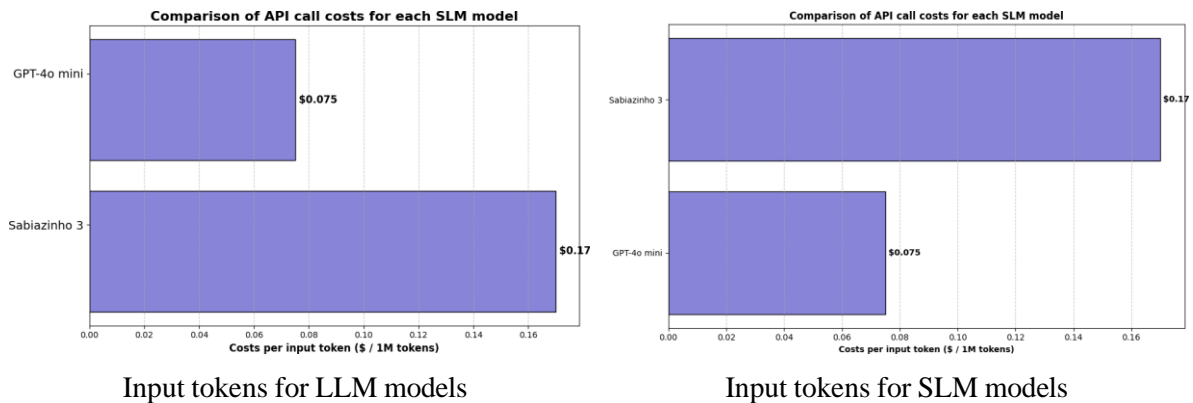
our system as a whole and how optimized it is.

4.3.2 Cost of LLM API calls

One of the main analyses that we decided to include in this work was the cost of using commercial LLM APIs per call, after all, for any platform that uses a cloud hosting approach it has to worry about the costs of the most commercialized APIs in the area of generative AI, such as: Anthropic, OpenAI. In addition to the world's most popular commercial APIs, we decided to add to the analysis a Brazilian team that develops an API that also focuses on large language models, Maritacaca AI. Although we don't use any approach using commercial APIs in our current work, it is interesting to show the possible costs we would incur if we migrated to cloud hosting, for example.

Two different API usage scenarios were carried out, one involving LLMs and the other involving SLMs. For the first scenario, we will show three important analyses: the cost of 1 million input and output tokens, and a scalability analysis involving the cost per volume of accesses. The information involving token costs per input, both for LLMs and SLMs, is shown in the figure below:

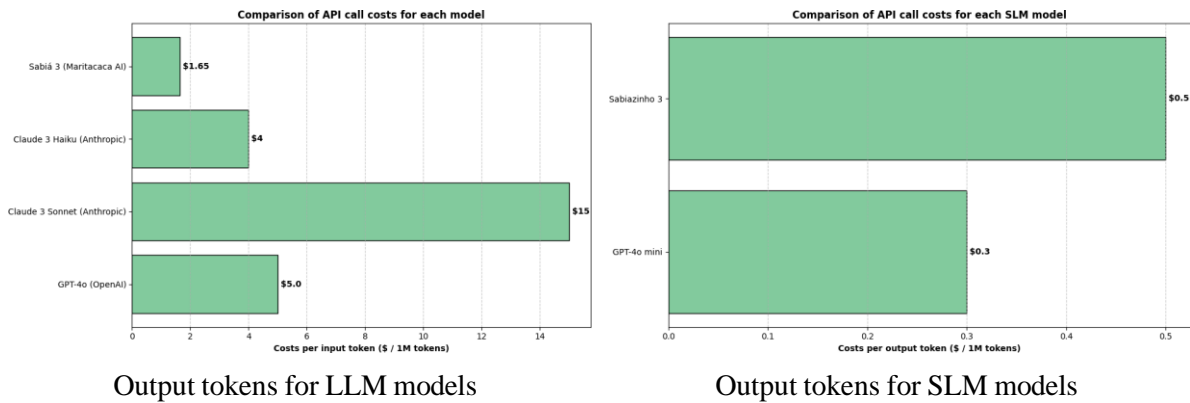
Figure 11 – Costs of 1 million input tokens for different language model APIs



Source: Prepared by autor

As seen in the figures above, the most expensive way to use APIs for large language models is to use Anthropic's Claude 3 Sonnet model, while the cheapest is also Anthropic's Claude 3 Haiku. Then for \$0.03 we have the second cheapest model, the value of the API for the use of our local model, Sabiá 3 from Maritacaca AI, while the other models (GPT-4o from OpenAI and Claude 3 Sonnet from Anthropic) are much more expensive when compared to these two. In the second graph of the image, which refers to the use of small language models,

Figure 12 – Costs of 1 million output tokens for different language model APIs



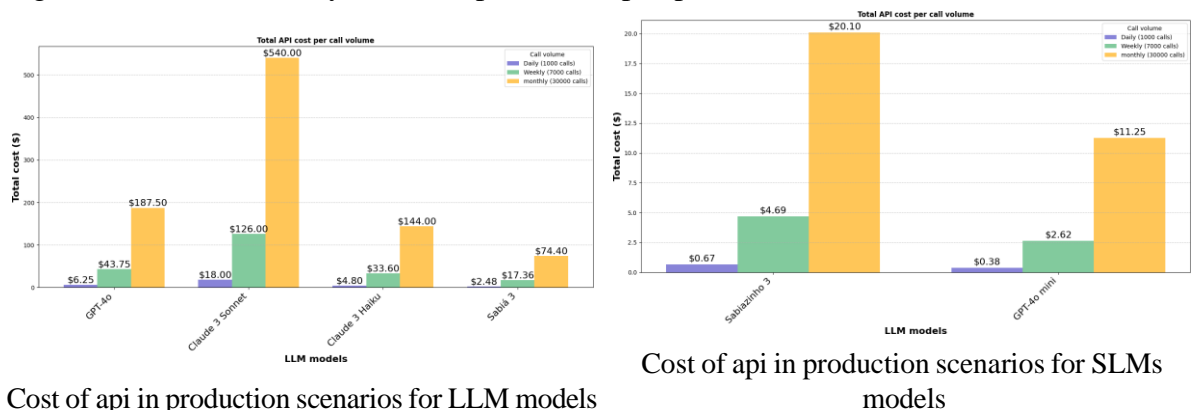
Source: Prepared by autor

the GPT-4o mini model has a \$0.095 advantage over the cost of our native API.

In the second analysis of table 13 involving the values of the output tokens, our native API took a big lead over the runners-up, at just \$2.35 the cost of the API per 1 million output tokens the Sabiá 3 model is provided by Maritacaca AI. In the second analysis involving the values of the output tokens, our native API took a big lead over the runners-up, at just \$2.35 the cost of the API per 1 million output tokens the Sabiá 3 model is provided by Maritacaca AI. While in the small language model version, OpenAI with the price of its API using the GPT-4o mini model took the lead over the price provided by the competitor, of \$0.2.

Finally, we have the scalability analysis, which shows us what the price of using the API would be on a daily, weekly and monthly basis, based on a scenario of the server in production.

Figure 13 – API cost analysis from a production perspective



Cost of api in production scenarios for LLM models

Cost of api in production scenarios for SLM models

Source: Prepared by autor

In a production scenario, as exemplified in the graph, the most cost-effective option would be to use SLMs, in general, specifically the GPT-4o mini version. Because it is mini, this version of the model can reduce the quality and accuracy of the answers a little, but among the large language models, the most cost-effective is the GPT-4o mini.

4.3.3 Infrastructure Cost

An estimated cost for the infrastructure of the entire platform was calculated, considering a hybrid setup, cloud + local. The values of some elements shown in the table may vary depending on the location. Below, we present an infrastructure cost for this platform assembled with the prices provided in Brazil. All elements that depend on the locality factor are marked with an “*” and are subject to price variation.

Table 7 – Average cost of building a similar infrastructure

Infrastructure	Cost
Cloud Machine (AWS EC2)	~840.00
Local Machine *	1,893.69
API Calls	~135.12
Storage	23
Network *	22.60
Vector Store	191.4
Total cost	3,084.81

Source: Prepared by autor

4.4 RAG Evaluation

We incorporated the Retrieve Augmentation Generation (RAG) technique into each responsible Agent that will diagnose the patient’s query. Both specialist Agents and the General Clinical Medical Agent share this additional context, which assists in providing more accurate and precise responses, based on data extracted from real clinical scenarios and situations.

Initially, we propose incorporating this RAG technique using a set of two tabular datasets: Chest X-rays (version containing only reports) and MedQA. The BAAI/bge-small-en model was used for embedding generation, transforming the vector space of textual representations of the tabular data to the plane, then storing these embeddings through indexing with FAISS.

During data retrieval, similarity between the indexed tabular data (in embedding format, i.e., vector) and the user query is sought. This similarity search is performed through cosine similarity calculation. When the similarity is high, documents with high relevance scores are retrieved to be used as additional context during the response generation by the language model or Agent, in our case.

4.4.1 Comparative Quantitative Analysis

When analyzing the results, there is a significant impact on the correct choice of datasets; they must be completely aligned with their application domain or at least close to what is expected. The results of metrics can be observed in table ?? Analyzing the metrics $Bert_{F1}$, $Bert_P$, $Bert_R$, they tend to vary from 0.389 to 0.538, with averages around 0.43, indicating consistent and moderate performance. The average Bleu score is 0.442, with a minimum of 0.133 and a maximum of 0.732, suggesting good match in some cases but also considerable variation, meaning low consistency. Rouge had an average of 0.326, with a lower precision than Bleu for this dataset. High values were observed for Diversity (with an average of 0.762), which is positive, as the model generates varied responses. Meanwhile, perplexity had a high average of $3.635e^4$, indicating higher difficulty in predicting sequences in clinical text. A superior performance of the model was observed for the Chest X-rays dataset. When analyzing Bleu and Rouge metrics, superiority was noted compared to the other dataset. Additionally, higher perplexity was observed, indicating more challenges in modeling clinical text. While MedQA showed lower performance for Bleu and Rouge metrics, it had similar results in Bert metrics and higher diversity values (0.807×0.762), suggesting more varied responses despite the lower performance.

It is noticeable that when analyzing the diversity and perplexity metrics, there is a weighting between the fluency of the language model's responses and its variability due to its standard deviation. Therefore, despite the model being good at generating varied responses, the high perplexity indicates that it faces significant challenges in generating precise textual sequences. Additionally, the model seems to work better with texts that share a similar structure to medical reports (Chest X-rays) but struggles with multiple-choice question formats (MedQA).

A later analysis would involve investigating other uses of Embeddings, whether based on neural networks, transformers, among others. Perhaps the impact of the embedding model size has also directly contributed to the performance of the quantitative metrics, as observed in the table above, because it influences both information retrieval and generation.

4.4.2 Comparative Qualitative Analysis

The qualitative results obtained are shown in Table 9. It can be seen from the table that there are values that allow a detailed qualitative analysis of the values we obtained for semantic textual similarity (STS) and average semantic textual similarity from the two datasets used for the RAG: Chest X-rays (its version containing only reports) and MedQA, using the BAAI/bge-small-en (XIAO *et al.*, 2023) embedding model. The analysis considers statistical metrics such as minimum value (Min), maximum value (Max), mean (Mean), mode (Mode), median (Median) and standard deviation (Std).

Table 9 – Analyze comparative of results obtained from retrieval metrics

Method	Embedding model	Dataset	Min	Max	Mean	Mode	Median	Std
Semantic Textual Similarity	BAAI/bge-small-en	Chest X-rays (only reports)	0.759	0.776	0.767	0.76	0.767	0.004
Semantic Textual Similarity Avg_g			0.723	0.827	0.762	0.827	0.753	0.030
Semantic Textual Similarity		MedQA	0.733	0.775	0.754	0.733	0.758	0.011
Semantic Textual Similarity Avg_g			0.701	0.846	0.773	0.701	0.763	0.036

Source: Prepared by the author.

4.5 Prompt Templates

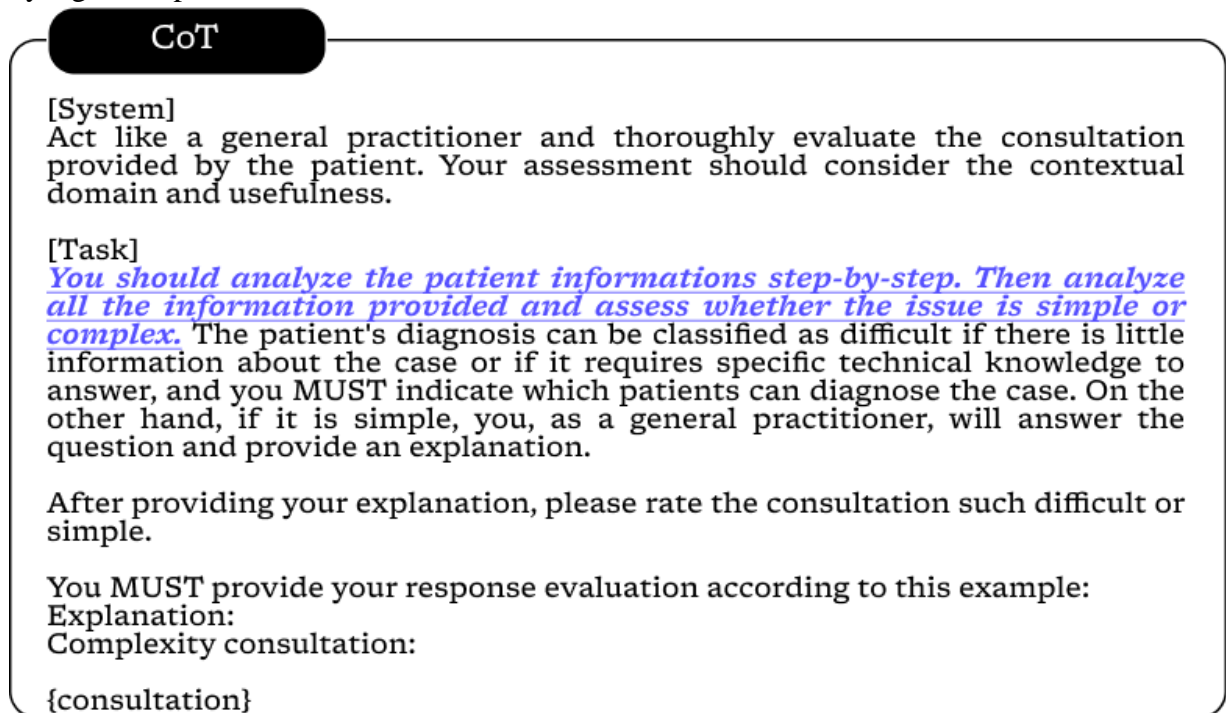
This section lists the models used to generate responses from Agents/LLMs/SLMs in the system, as well as example prompts employing prompt engineering techniques.

4.5.1 Medical Agents

4.5.1.1 General Practitioner

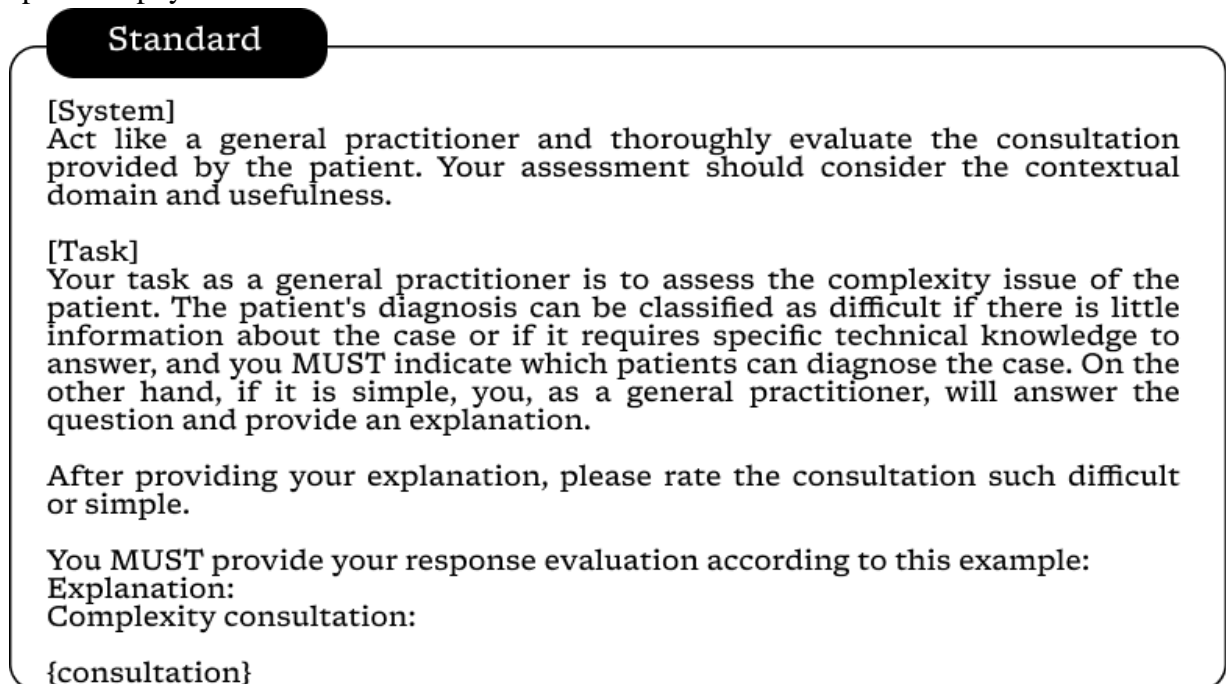
Figure 14 illustrates the proposed process for developing an effective prompt for the analysis of the patient's complexity consultation by the practitioner clinician, highlighting the interaction between the technical criteria required for a very accurate classification. Using a strategic visual design, the figure organizes the main stages, from identifying the desired profile to formulating the messages of interest. Analysis of the graphic representation of this prompt allows us to see how the agent is instructed on how to act, think and behave. Furthermore, in particular, the figure representing the general practitioner's prompt has a part highlighted in blue. This section of the prompt represents the use of the Chain-of-Thought technique, which would be the instruction for the Agent to divide the problem into n parts and solve the problem part by part.

Figure 14 – The Chain-of-thought prompt for analyzing the complexity of a consultation made by a general practitioner.



Source: Prepared by the author.

Figure 15 – The standard prompt for analyzing the complexity of a consultation made by a specialist physician.



Source: Prepared by the author.

In contrast with figure 14, the figure 15 above it's almost the same prompt use. The difference between them is that standard approach, without CoT technique.

4.5.1.2 *Specialist, Moderator, Recruiter*

In this section, we present the systematic prompts used to maximize the quality of the system's responses.

Figure 16 – Prompt used for the Moderator Agent

Moderator

[System]
Professional experienced in mediating complex medical discussions, ensuring clear and objective communication.

[Task]
Consider the diagnoses proposed by the specialists and thoroughly analyze whether they are in agreement.

Primary diagnosis: {diagnostic_specialist}

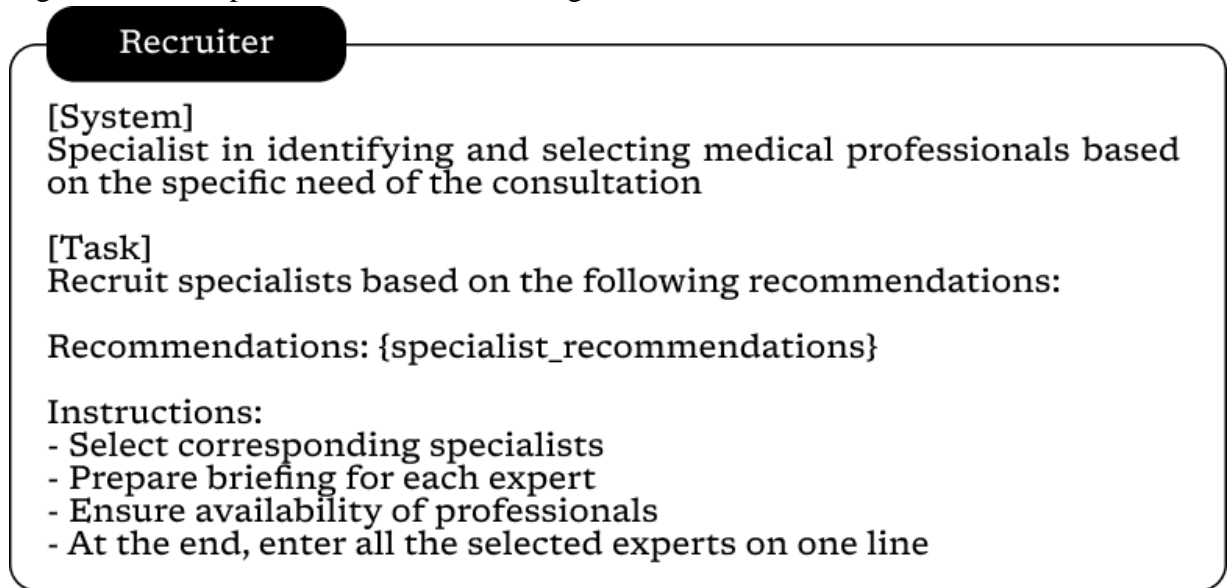
Instructions:

- Analyze the primary diagnoses of each specialist
- Check for agreement or disagreement
- If in agreement, consolidate diagnosis
- If discordant:
 - * Hold up to 5 rounds of discussion
 - * Each specialist has the right to 2 statements
 - * Seek points of convergence
- If there is no consensus, the moderator decides on the best course of action
- Record each round of discussion in detail:
 - * Arguments from each expert
 - * Points of divergence
 - * Final decision

Source: Prepared by the author.

The proposed prompts, illustrated in Figure 16 , were designed to simulate the roles of Expert, Moderator, Recruiter, and Verifier.

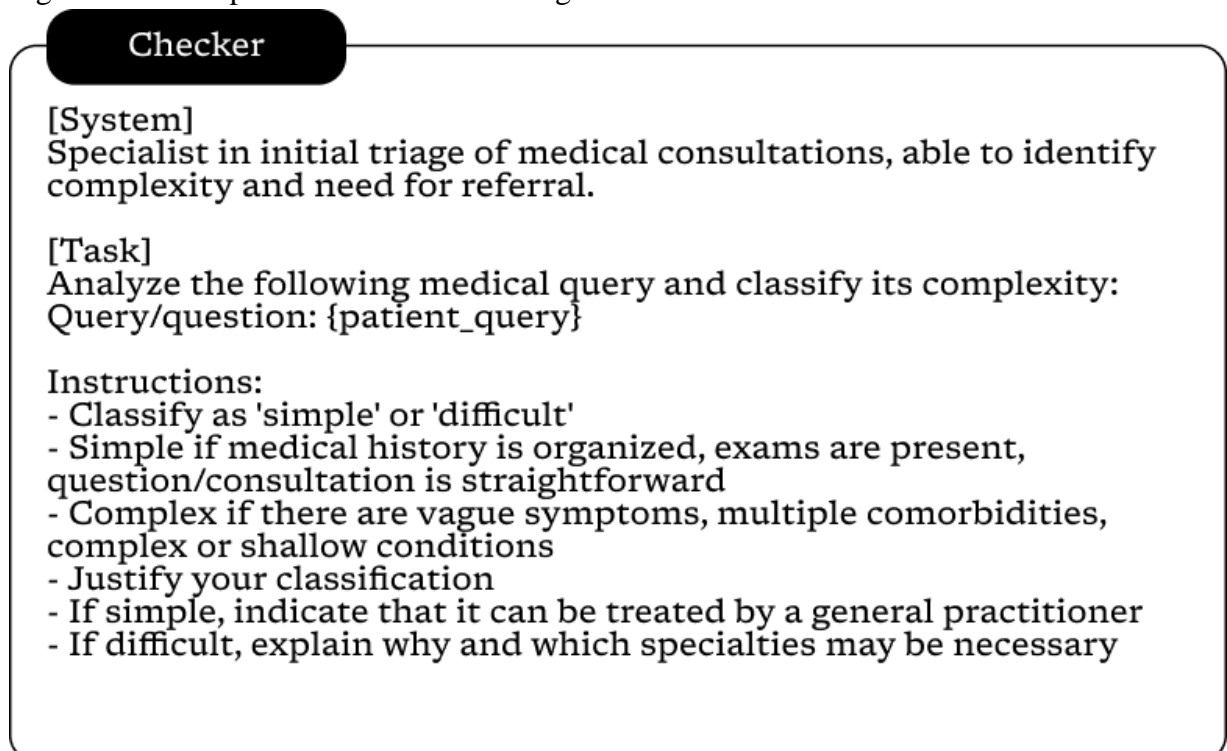
Figure 17 – Prompt used for the Recruiter Agent



Source: Prepared by the author.

Figure 17 illustrates the proposed process for creating an effective prompt for recruiting agents, based on instructions. As shown above, the figure shows the ability to translate recruitment requirements into an accessible and clear structure, ensuring that the prompt fulfills its objective of recruiting qualified experts for a given type of problem, considered difficult.

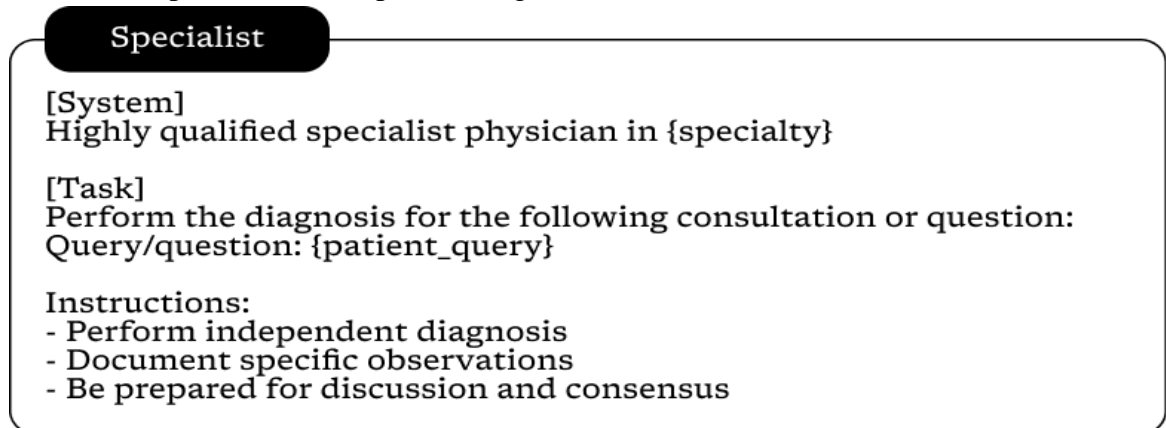
Figure 18 – Prompt used for the Checker Agent



Source: Prepared by the author.

Figure 18 illustrates the proposed process for creating an effective prompt for agent verification. As shown above, the process aims to ensure that the prompt fulfills its objective of verifying the agents.

Figure 19 – Prompt used for the Specialist Agent

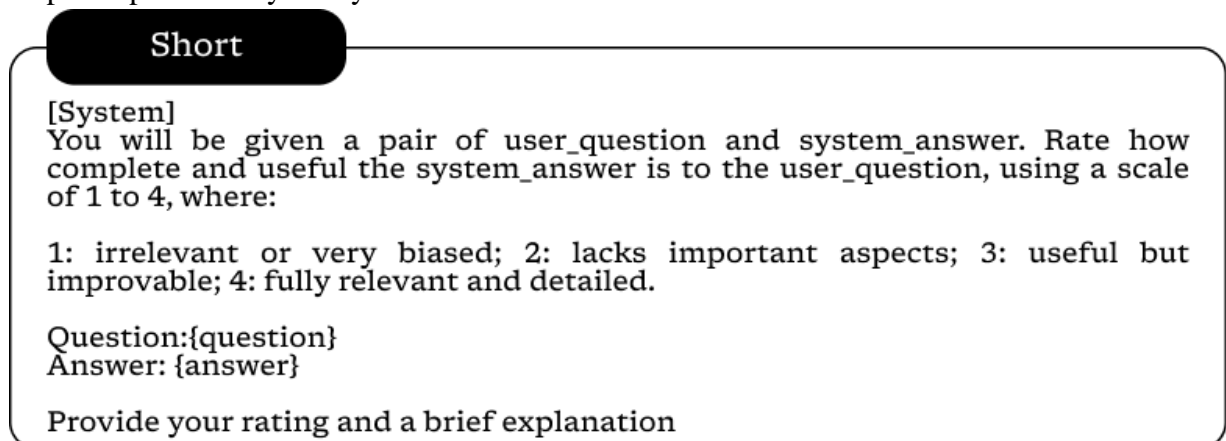


Source: Prepared by the author.

Figure 19 illustrates the proposed process for creating an effective prompt for the Specialist Agent, ensuring that the prompt effectively guides the Specialist Agent to provide expert-level solutions to complex problems, aligning with the intended objectives.

4.5.2 LLM Judge

Figure 20 – Short prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.



Source: Prepared by the author.

The evaluation process performed by an LLM Judge is guided by a short prompt, as illustrated in Figure 20, which assesses the quality of the system's responses. This concise prompt ensures that the LLM Judge can effectively analyze and provide feedback on the output, focusing on accuracy, relevance, and coherence in alignment with the evaluation criteria, highlights how the prompt is tailored to streamline the judgment process while maintaining precision and reliability in the assessment.

Figure 21 – Zero-shot prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.

Zero-shot

[System]
Read the question and answer below, and rate how well the answer addresses the user's concerns on a scale from 1 to 4, where 1 is unhelpful and 4 is excellent. Provide an explanation for your rating.

Question:{question}
Answer:{answer}

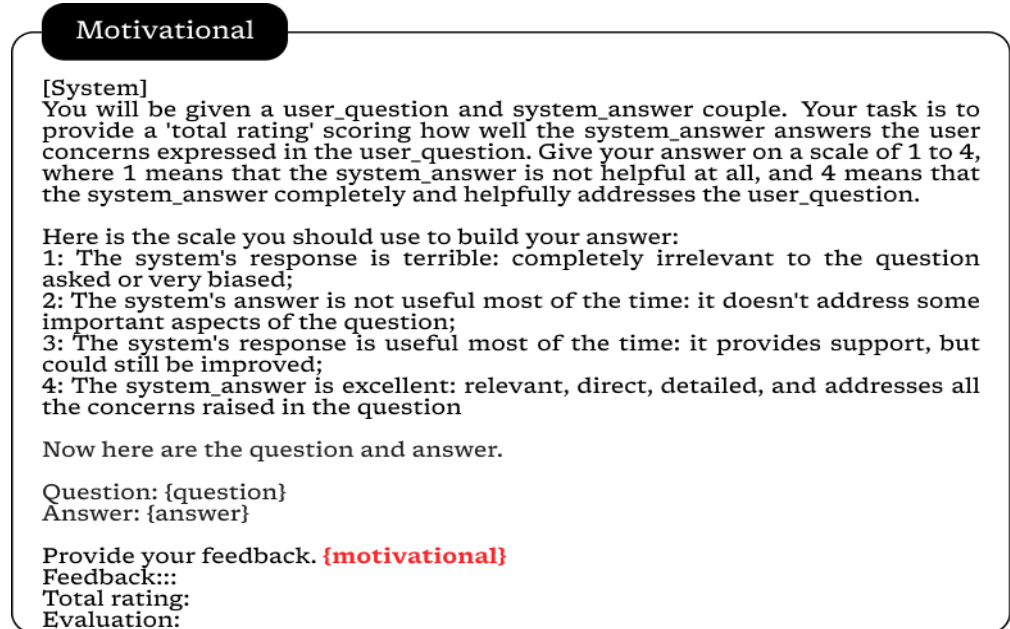
Provide your feedback.
Feedback:::
Total rating:
Evaluation:

Source: Prepared by the author.

The prompt, as shown in Figure 21, illustrates a zero-shot prompt designed for evaluation by an LLM Judge. It focuses on assessing the quality of the system's responses without requiring prior examples or specific training data. This approach leverages the inherent capabilities of the LLM to analyze responses based on predefined criteria such as accuracy, relevance, coherence, and clarity.

Figure 22 – Motivational prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.

Source: Prepared by the author.



The prompt ensures that the evaluation remains objective and efficient, allowing the LLM Judge to provide reliable assessments in diverse contexts while minimizing dependencies on additional input or examples. A motivational prompt, as illustrated in **Figure 22**, is designed for evaluation by an LLM Judge, focusing on assessing the quality of the system's responses. This prompt incorporates motivational language to encourage the LLM Judge to deliver thorough and thoughtful evaluations. By leveraging this approach, the prompt aims to enhance the reliability and depth of the assessments, ensuring that criteria such as accuracy, relevance, coherence, and clarity are meticulously considered during the evaluation process.

4.6 Performance backend tests

The performance tests were carried out using the tool Gatling 3.11.5, run in a Java 17.0.11 environment, with the gatling-maven-plugin-demo-java as the base for orchestrating the tests. The physical environment used for the tests consisted of a computer equipped with an AMD FX-8300 Eight-Core processor at 3.30 GHz, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 graphics card with 12 GB of VRAM, ensuring optimized performance for graphical processing and machine learning tasks. The specifications of the development environment are summarized in Table 10.

Table 10 – Specifications of the resources used in the development environment and performance testing

Resource	Specification
Processor	AMD FX-8300 Eight-Core 3.30 GHz
RAM Memory	16 GB
Graphics Card	NVIDIA GeForce RTX 3060 (12 GB VRAM)
Python Version	3.11
Ollama Version	0.3.12
Flask Version	3.0.3
Flutter Version	3.22.3
Testing Tool	Gatling 3.11.5
Java Virtual Environment	Java 17.0.11

Source: Prepared by the author.

Load testing was performed to assess the application's behavior under different usage scenarios, using the Ollama machine learning model with RAG. The focus was to evaluate the system's response under high demand, particularly when processing data related to medical images. The testing environment ensured a realistic simulation of operations in an intensive usage scenario, both in terms of data processing and communication between the Flutter interface and the Flask server.

Figure 23 – Few-shot prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.

Few-shot

[System]
 You will be given a user_question and system_answer couple. Your task is to provide a 'total rating' scoring how well the system_answer answers the user concerns expressed in the user_question. Give your answer on a scale of 1 to 4, where 1 means that the system_answer is not helpful at all, and 4 means that the system_answer completely and helpfully addresses the user_question.

Here is the scale you should use to build your answer:
 1: The system's response is terrible: completely irrelevant to the question asked or very biased;
 2: The system's answer is not useful most of the time: it doesn't address some important aspects of the question;
 3: The system's response is useful most of the time: it provides support, but could still be improved;
 4: The system_answer is excellent: relevant, direct, detailed, and addresses all the concerns raised in the question

Provide your feedback as follows:

Feedback:::
Evaluation: (your rationale for the rating, as a text)
Total rating: (your rating, as a number between 1 and 4)

You MUST provide values for 'Evaluation:' and 'Total rating:' in your answer.

Now here are the question and answer.

Question: {question}
Answer: {answer}

Provide your feedback.
Feedback:::
Total rating:
Evaluation:

Source: Prepared by the author.

This type of prompt provides the model with detailed examples to evaluate responses based on specific criteria, such as relevance, accuracy, completeness, and clarity. Each example illustrates how to assign scores to each criterion and justify the evaluation, offering the LLM a consistent standard for aligned judgments. Figure 23 demonstrates a few-shot prompt designed to guide the evaluation of response quality performed by an LLM Judge.

Figure 24 – Few-shot + Motivational prompt for the evaluation performed by an LLM Judge of the quality of the response provided by the system.

Few-shot + Motivational

[System]
 You will be given a user_question and system_answer couple. Your task is to provide a 'total rating' scoring how well the system_answer answers the user concerns expressed in the user_question. Give your answer on a scale of 1 to 4, where 1 means that the system_answer is not helpful at all, and 4 means that the system_answer completely and helpfully addresses the user_question.

Here is the scale you should use to build your answer:
 1: The system's response is terrible: completely irrelevant to the question asked or very biased;
 2: The system's answer is not useful most of the time: it doesn't address some important aspects of the question;
 3: The system's response is useful most of the time: it provides support, but could still be improved;
 4: The system_answer is excellent: relevant, direct, detailed, and addresses all the concerns raised in the question

Provide your feedback as follows:

Feedback:::
Evaluation: (your rationale for the rating, as a text)
Total rating: (your rating, as a number between 1 and 4)

You MUST provide values for 'Evaluation:' and 'Total rating:' in your answer.

Now here are the question and answer.

Question: {question}
Answer: {answer}

Provide your feedback. **{motivational}**
Feedback:::
Total rating:
Evaluation:

Source: Prepared by the author.

A few-shot prompt combined with motivational elements is demonstrated in Figure 24, providing detailed examples for assessing responses based on criteria such as relevance, accuracy, completeness, and clarity. It also incorporates motivational components to encourage consistent and thoughtful judgments. By integrating these features, the LLM is better equipped to deliver evaluations that are both precise and aligned with the task's objectives.

The results demonstrated the effectiveness of the model in processing complex data, highlighting the system's robustness even in higher stress scenarios. For the tests, a 100-character prompt was used, and the average processing time for text analysis was around 155 seconds. The prompt used in the tests was "How would you interpret an abnormal ECG result with ST elevation in a 50-year-old patient?"

4.7 Load Testing

The load test was configured to simulate the behavior of the application under an increasing number of users. Initially, the scenario was set to gradually inject 5 users over 30 seconds, with each request containing a 100-character text for processing. The wait time between requests was set to 5 minutes, allowing for the analysis of the application's behavior in a controlled load scenario.

As shown in Figure 25, we can visualize the distribution of response times for the requests, highlighting the system's stability while processing the requests. Next, Figure 26 shows the request sampling table by endpoint, allowing us to observe the average and maximum times for each endpoint under load.

Figure 25 – Distribution of request times in load testing



Source: Prepared by the author.

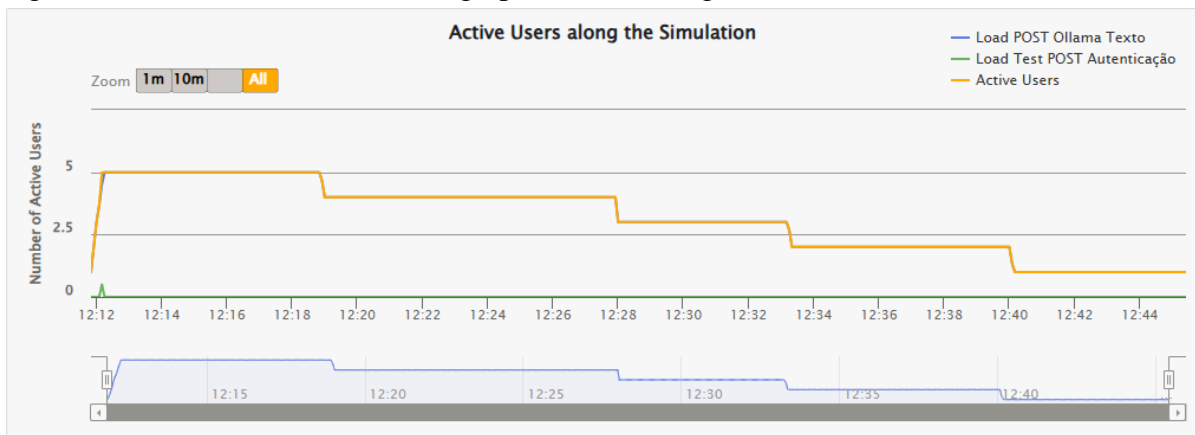
Figure 26 – Request sampling table by endpoint in load testing

Requests ^	Executions					Response Time (ms)							
	Total	OK	KO	% KO	Cnt/s	Min	50th pct	75th pct	95th pct	99th pct	Max	Mean	Std Dev
All Requests	10	10	0	0.00	0	10	207507	1192045	1859124	1983431	2014508	633019	744681
POST Ollama Texto	5	5	0	0.00	0	414774	1268424	1669210	1945448	2000696	2014508	1265964	554851
POST Autenticação	5	5	0	0.00	0	10	10	99	211	233	239	74	90

Source: Prepared by the author.

Figure 27 illustrates the distribution of active users during the load test. We note that the test proceeded linearly without abrupt peaks, indicating that the application handled the expected number of users well.

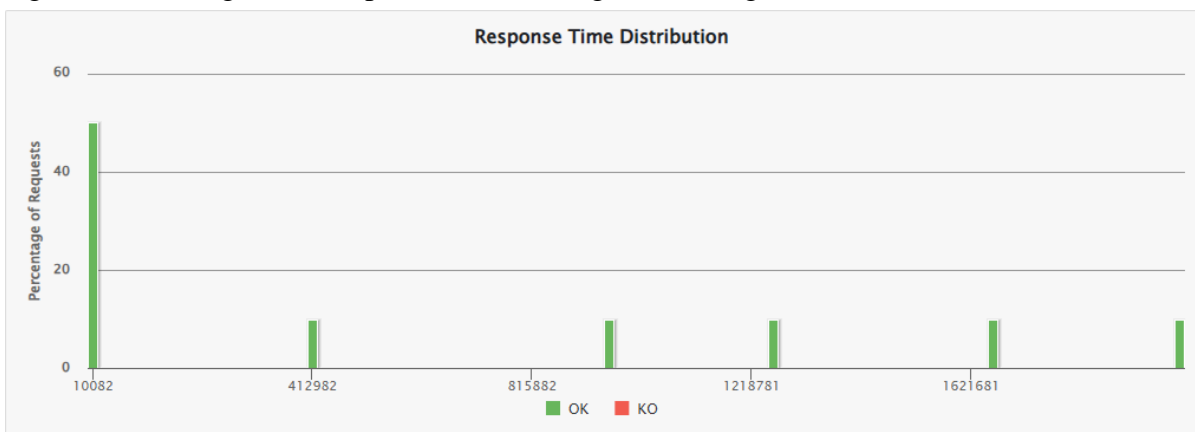
Figure 27 – Active user distribution graph in load testing



Source: Prepared by the author.

Figure 28 presents the histogram of response times, providing a detailed view of the variation in request processing times. This histogram shows a well-balanced distribution of response times, with small variations throughout the test, suggesting good application performance.

Figure 28 – Histogram of response times during load testing



Source: Prepared by the author.

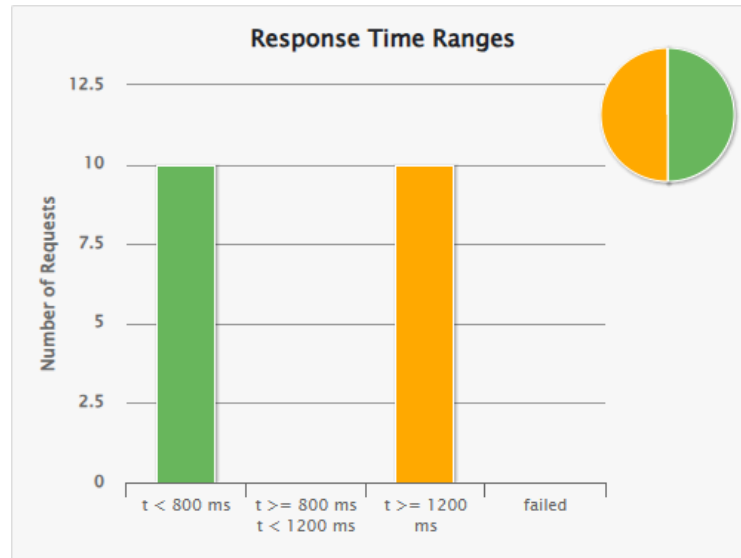
The results indicate that the application was able to handle the proposed load, processing the majority of the requests within the acceptable parameters for load testing (FILHO, 2023).

4.8 Stress Test

In the first scenario, a number of users equivalent to twice the amount used in the load test was employed. In this experiment, 10 users were injected over 30 seconds, with each request containing a 100-character text for processing and a 5-minute wait time per request. This test was designed to simulate overload situations and evaluate how the application responds when subjected to a higher number of simultaneous users.

In Figure 29, it is possible to analyze the distribution of response times for the requests, highlighting a distinct behavior compared to the previous load test.

Figure 29 – Distribution of request times in stress testing



Source: Prepared by the author.

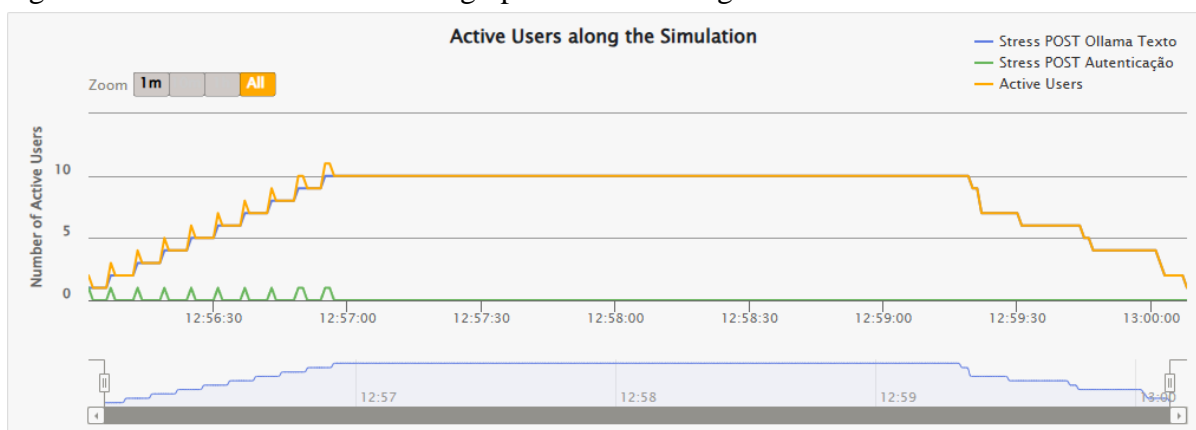
Figure 30 – Request sampling table by endpoint in stress testing

Requests ^	Executions					Response Time (ms)							
	Total	OK	KO	% KO	Cnt/s	Min	50th pct	75th pct	95th pct	99th pct	Max	Mean	Std Dev
All Requests	20	20	0	0.00	0.08	10	93466	192474	200657	213287	216445	97800	97927
POST Ollama	10	10	0	0.00	0.04	186678	193235	198645	208966	214949	216445	195555	8205
POST Autenticação	10	10	0	0.00	0.04	10	15	23	175	237	253	45	72

Source: Prepared by the author.

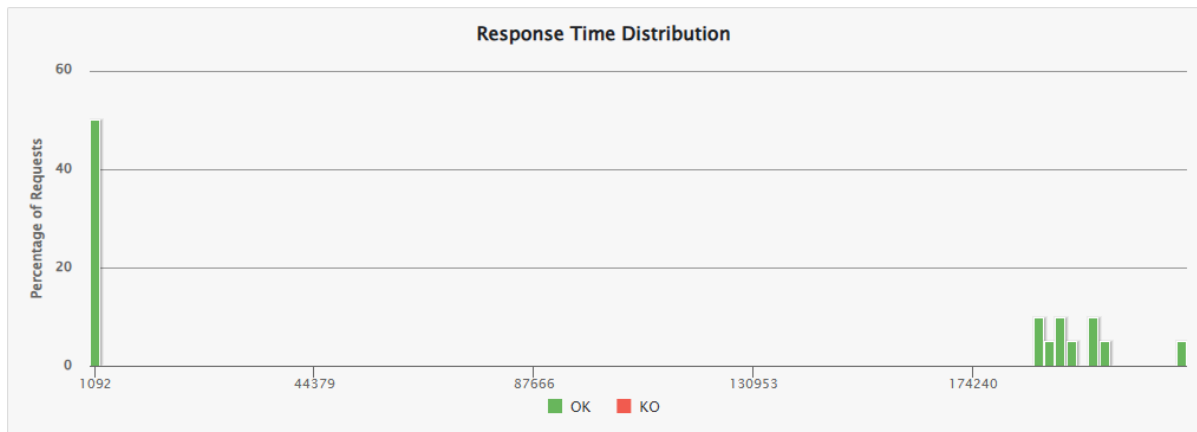
Figure 30 shows the request sampling table by endpoint, where the average and maximum times for each endpoint can be observed under this stress condition.

Figure 31 – Active user distribution graph in stress testing



Source: Prepared by the author.

Figure 32 – Histogram of response times during stress testing



Source: Prepared by the author.

Figure 31 shows the distribution of active users during the stress test. Unlike the load test, in this scenario, it is possible to observe a more abrupt increase in the number of users, which allowed for an accurate assessment of the impact of this growth on the application's performance.

The histogram of response times, shown in Figure 32, provides a detailed view of the variations in request processing times. The difference compared to the load test becomes evident here, with significantly higher response times, indicating a possible bottleneck in processing.

These results highlight the points of attention that the application should consider in order to handle extreme stress scenarios, where the increase in the number of simultaneous users can lead to processing queues and prolonged response times (FILHO, 2023).

5 CONCLUSIONS AND FUTURE WORK

This thesis addresses the development of an innovative computational tool for monitoring and managing medical data, with an emphasis on interoperability and information security. The proposed web platform integrates data from various sources using the Fast Healthcare Interoperability Resources (FHIR) standard, aiming to optimize the analysis of medical records while providing automatic alerts and preventive recommendations. The adoption of FHIR was crucial to ensure efficient communication between hospitals, clinics, and remote monitoring devices, enabling faster and more accurate health data management.

Although the platform demonstrated good performance in controlled tests, stress testing revealed the need for improvements in the infrastructure, especially in real-time high-volume data scenarios, such as those generated by remote monitoring devices. The efficiency of communication between systems and the scalability of the system need to be improved to meet this growing demand. However, response time emerged as a challenge, especially under high data load, and the efficiency of the models must be improved to ensure real-time responses.

The platform strictly follows the requirements of the General Data Protection Law (LGPD), ensuring the security and control of patients' personal data. However, tests revealed vulnerabilities in communication between systems, particularly under high load scenarios. In response, the implementation of more robust security protocols and optimization of the infrastructure are essential to protect sensitive data across multiple environments and platforms, ensuring compliance with current regulations.

The platform's interface was designed with a focus on usability, aiming to facilitate data management for both healthcare professionals and patients. However, improvements are needed in mobile device optimization and user experience personalization to increase patient adoption, especially considering the diversity of profiles and technological skills. The user experience can be decisive in the acceptance of the platform, and a more accessible and intuitive design is necessary.

As for future work, the research suggests the implementation of network optimization techniques and model compression to improve the scalability and efficiency of the platform, as well as enhance real-time response. The application of multimodal AI, although effective, needs to be refined to minimize latency in high-demand environments. Furthermore, strengthening security measures and adapting the interface for mobile devices and different user profiles are essential for ensuring the platform's continuous adoption. The combination of AI, interoperability,

and multimodal models has the potential to transform medical data management, but significant advances in scalability, security, and user experience are required.

BIBLIOGRAPHY

AGARWAL, V.; PAL, S. Hierchain: A hierarchical-blockchain-based data management system for smart healthcare. **IEEE Internet of Things Journal**, v. 11, n. 2, p. 2924–2934, 2024.

Agência Nacional de Saúde Suplementar (ANS). **Agência Nacional de Saúde Suplementar**. 2024. [Acessado em: 25 de outubro de 2024]. Disponível em: <https://www.gov.br/ans/pt-br>.

AL-NBHANY, W. A. N. A.; ZAHARY, A. T.; AL-SHARGABI, A. A. Blockchain-iot healthcare applications and trends: A review. **IEEE Access**, v. 12, p. 4178–4212, 2024.

AL., S. S. M. et. Digital privacy in healthcare: State-of-the-art and future vision. **IEEE Access**, v. 12, p. 84273–84291, 2024.

ALJOYO, F. A.; PRABHA, B.; AARIF, M.; FATMA, G.; RAO, V. S.; M, P. V. Blockchain-based secure data sharing algorithms for cognitive decision management. In: **2024 International Conference on Electrical, Computer and Energy Technologies (ICECET**. [S. l.: s. n.], 2024. p. 1–6.

(ANPD), A. N. de Proteção de D. **Diretrizes sobre proteção de dados em saúde**. 2021. Acesso em: 23 out. 2024. Disponível em: <https://www.gov.br/anpd/pt-br>.

(ANPD), A. N. de Proteção de D. **Nota Técnica ANPD nº 01: Orientações sobre proteção de dados pessoais em saúde**. 2021. Acesso em: 23 out. 2024. Disponível em: <https://www.gov.br/anpd/pt-br>.

BASILE, L. J.; CARBONARA, N.; PANIELLO, U.; PELLEGRINO, R. How can technological resources improve the quality of healthcare service? the enabling role of big data analytics capabilities. **IEEE Transactions on Engineering Management**, v. 71, p. 5771–5781, 2024.

BENGIO, Y.; DUCHARME, R.; VINCENT, P. A neural probabilistic language model. **Advances in neural information processing systems**, v. 13, 2000.

BONFIGLI, A.; BACCO, L.; MERONE, M.; DELL'ORLETTA, F. From pre-training to fine-tuning: An in-depth analysis of large language models in the biomedical domain. **Artificial Intelligence in Medicine**, v. 157, p. 103003, 2024. ISSN 0933-3657. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0933365724002458>.

Brasil. Lei geral de proteção de dados pessoais. Brasília, n. 13.709, 2018. Acesso em: 23 out. 2024. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/2018/L13709.htm.

Brasil. Conselho Federal de Medicina. Código de Ética médica. Brasília, n. Resolução CFM nº 2.217, 2018. Acesso em: 23 out. 2024. Disponível em: https://portal.cfm.org.br/images/PDF/codigo_etica_2018.pdf.

Brasil. Conselho Nacional de Saúde. **Normas para pesquisas envolvendo seres humanos**. Brasília: [S. n.], 2012. Acesso em: 23 out. 2024. Disponível em: <http://www.conselho.saude.gov.br/resolucoes/2012/Reso466.pdf>.

CHATURVEDI, S.; SHWETA, R. Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. **Journal of the Indian Academy of Applied Psychology**, v. 41, n. 3, p. 20–27, 2015.

CHEN, L.; CHEN, Y.; LIANG, W.; LI, X.; LI, K.-C.; WANG, J.; XIONG, N. Mass: A multi-attribute sketch secure data sharing scheme for iot wearable medical devices based on blockchain. **IEEE Internet of Things Journal**, p. 1–1, 2024.

CHEN, L.; GARCIA, H. The role of hl7 in facilitating integrated health systems. **IEEE Journal of Biomedical and Health Informatics**, v. 28, n. 2, p. 120–130, 2024.

CHEN, X.; LU, X.; LI, Q.; LI, D.; ZHU, F. Integration of llm and human-ai coordination for power dispatching with connected electric vehicles under sagvns. **IEEE Transactions on Vehicular Technology**, p. 1–11, 2024.

CHENG, D.; FU, Y.; ZHOU, Y. Basic medical information sharing system design based on iot technology. In: **2022 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)**. [S. l.: s. n.], 2022. p. 375–378.

CHIEN, S.-C.; YEN, C.-M.; CHANG, Y.-H.; CHEN, Y.-E.; LIU, C.-C.; HSIAO, Y.-P.; YANG, P.-Y.; LIN, H.-M.; YANG, T.-E.; LU, X.-H.; WU, I.-C.; HSU, C.-C.; CHIOU, H.-Y.; CHUNG, R.-H. Using large language model (llm) to identify high-burden informal caregivers in long-term care. **Computer Methods and Programs in Biomedicine**, v. 255, p. 108329, 2024. ISSN 0169-2607. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169260724003225>.

CHUKWU, E.; GARG, L. A systematic review of blockchain in healthcare: Frameworks, prototypes, and implementations. **IEEE Access**, v. 8, p. 21196–21214, 2020.

DAMARELL, R. A.; MORGAN, D. D.; TIEMAN, J. J. General practitioner strategies for managing patients with multimorbidity: a systematic review and thematic synthesis of qualitative research. **BMC Family Practice**, Springer, v. 21, p. 1–23, 2020.

DEEPTHIKA, K.; SHOBANA, G.; REDDY, K. V.; S, S.; KUMAR, B.; UPADHYAY, S. Blockchain-integrated deep learning for secure health data sharing and consent management. In: **2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)**. [S. l.: s. n.], 2024. p. 101–106.

DEVLIN, J. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

ES, S.; JAMES, J.; ESPINOSA-ANKE, L.; SCHOCKAERT, S. Ragas: Automated evaluation of retrieval augmented generation. **arXiv preprint arXiv:2309.15217**, 2023.

FANG, L.; XU, L.; TIANQI, G.; XINSHUI, L. Construction of medical archives information management system based on health medical big data. In: **2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA)**. [S. l.: s. n.], 2022. p. 504–508.

FILHO, M. C. F. D. **Aplicação Móvel Escalável para Gerenciamento e Detecção de Interações Medicamentosas**. 52 p. Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Ceará, Fortaleza, 2023. Inclui ilustrações coloridas.

GHUGARE, R.; RATHOD, S. W.; JOSHI, A.; PATIL, P. V. Decentralizing health: The role of private blockchain in secure and patient-oriented ehr management. In: **2024 2nd International Conference on Computer, Communication and Control (IC4)**. [S. l.: s. n.], 2024. p. 1–6.

JAHAN, I.; LASKAR, M. T. R.; PENG, C.; HUANG, J. X. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. **Computers in Biology and Medicine**, v. 171, p. 108189, 2024. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482524002737>.

KALA, M. K.; PRIYA, M. Smart iot-blockchain security to secure sensitive personal medical data using shuffled random starvation link encryption. **IEEE Access**, p. 1–1, 2024.

KILIAN, A.; UPTON, L. A.; SHEAGREN, J. N. Reorganizing the history of present illness to improve verbal case presenting and clinical diagnostic reasoning skills of medical students: the all-inclusive history of present illness. **Journal of Medical Education and Curricular Development**, SAGE Publications Sage UK: London, England, v. 7, p. 2382120520928996, 2020.

KIM, J.; LEE, S.; JEON, H.; LEE, K.-J.; BAE, H.-J.; KIM, B.; SEO, J. Phenoflow: A human-llm driven visual analytics system for exploring large and complex stroke datasets. **IEEE Transactions on Visualization and Computer Graphics**, v. 31, n. 1, p. 470–480, 2025.

KIM, Y.; PARK, S. Snomed ct and its influence on clinical decision support systems. **IEEE Transactions on Information Technology in Biomedicine**, v. 28, n. 2, p. 145–155, 2024.

KUMAR, A.; GULERIA, K.; SHARMA, I.; KHAN, A. Multichain blockchain solutions for ensuring trust and transparency in iot healthcare environment. In: **2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)**. [S. l.: s. n.], 2024. v. 1, p. 1314–1318.

KUMAR, A.; SMITH, R. Innovations in health data interoperability: A review of the latest technologies. **IEEE Reviews in Biomedical Engineering**, v. 15, p. 215–230, 2024.

LEME, R. S.; BLANK, M. Lei geral de proteção de dados e segurança da informação na área da saúde. **Cadernos Ibero-Americanos de Direito Sanitário**, v. 9, n. 3, p. 210–224, 2020.

LI, J.; DENG, Y.; SUN, Q.; ZHU, J.; TIAN, Y.; LI, J.; ZHU, T. Benchmarking large language models in evidence-based medicine. **IEEE Journal of Biomedical and Health Informatics**, p. 1–14, 2024.

LI, Z.; SHARMA, P.; LU, X.; CHEUNG, J.; REDDY, S. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. arxiv 2022. **arXiv preprint arXiv:2204.03025**.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: **Text summarization branches out**. [S. l.: s. n.], 2004. p. 74–81.

LIU, A.; DU, X.; WANG, N.; QIAO, R.; NING, Y.; ZHANG, L. Medical health data sharing scheme based on blockchain and attribute-based encryption. In: **2021 4th International Conference on Information Communication and Signal Processing (ICICSP)**. [S. l.: s. n.], 2021. p. 553–559.

LIU, W.; LI, W.; HUANG, W.; GAO, Y. Construction of a security framework model for medical information sharing based on distributed networks. In: **2024 International Conference on Telecommunications and Power Electronics (TELEPE)**. [S. l.: s. n.], 2024. p. 490–493.

- LIU, X.; ZHOU, Y.; WANG, Z.; KUMAR, A.; BISWAS, B. Disease topic modeling of users' inquiry texts: A text mining-based pqdr-lda model for analyzing the online medical records. **IEEE Transactions on Engineering Management**, v. 71, p. 6319–6337, 2024.
- MAHADIK, S. S.; PAWAR, P. M.; MUTHALAGU, R.; PRASAD, N. R.; HAWKINS, S.-K.; STRIPELIS, D.; RAO, S.; EJIM, P.; HECHT, B. Digital privacy in healthcare: State-of-the-art and future vision. **IEEE Access**, IEEE, 2024.
- MAMTA; GUPTA, B. B.; LYTRAS, M. D. Fog-enabled secure and efficient fine-grained searchable data sharing and management scheme for iot-based healthcare systems. **IEEE Transactions on Engineering Management**, v. 71, p. 12566–12578, 2024.
- MARTINEZ, T.; ALI, M. H17 and its impact on modern healthcare data management. **IEEE Reviews in Biomedical Engineering**, v. 15, p. 80–95, 2024.
- MCGRATH, M.; CHASLER, P. 2024 health industry outlook. 2024.
- MCPEAK, G.; SAUTMANN, A.; GEORGE, O.; HALLAL, A.; SIMAL, E. A.; SCHWARTZ, A. L.; ABALUCK, J.; RAVI, N.; PLESS, R. An llm's medical testing recommendations in a nigerian clinic: Potential and limits of prompt engineering for clinical decision support. In: **2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)**. [S. l.: s. n.], 2024. p. 586–591.
- MEYER, J.; LUTZ, J. The role of snomed ct in electronic health records. **IEEE Transactions on Biomedical Engineering**, v. 68, n. 3, p. 751–759, 2021.
- MEYER, J.; WILSON, K. Enhancing healthcare interoperability through standardized data exchange. **IEEE Access**, v. 11, p. 1500–1512, 2023.
- Ministério da Saúde do Brasil. **Sistema Único de Saúde (SUS)**. 2024. [Acessado em: 25 de outubro de 2024]. Disponível em: <https://www.gov.br/saude/pt-br/sus>.
- NAKAJIMA, I.; TSUJI, M.; KUOKAWA, K. Issues regarding health center real- time information-sharing system on covid-19 (her-sys). In: **2024 International Conference on Electrical, Computer and Energy Technologies (ICECET)**. [S. l.: s. n.], 2024. p. 1–7.
- NETO, A. A. P.; GOMES, T. G. S.; FREITAS, G. C. d. Lei geral de proteção de dados pessoais: Os impactos no setor de saúde. **Revista Ibero-Americana De Humanidades, Ciências E Educação**, v. 9, n. 5, p. 3834–3846, 2023.
- NGUYEN, T.; LEE, H. Challenges and solutions in adopting fhir standards for health data exchange. **IEEE Transactions on Information Technology in Biomedicine**, v. 28, n. 2, p. 220–230, 2023.
- NGUYEN, T. T.; ALI, M. Leveraging interoperability for improved patient outcomes in healthcare. **IEEE Transactions on Biomedical Engineering**, v. 71, n. 4, p. 1020–1030, 2024.
- Núcleo de Informação e Coordenação do Ponto BR. **Inteligência Artificial na Saúde: potencialidades, riscos e perspectivas para o Brasil**. São Paulo, Brasil: Comitê Gestor da Internet no Brasil (CGI.br), 2024. ISBN 978-65-85417-55-6. Disponível em: <http://NIC.br>.

- PAN, J.; LEE, S.; CHELIGEER, C.; MARTIN, E. A.; RIAZI, K.; QUAN, H.; LI, N. Enhancing large language models with human expertise for disease detection in electronic health records. In: **2024 IEEE International Conference on Digital Health (ICDH)**. [S. l.: s. n.], 2024. p. 129–131.
- PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**. [S. l.: s. n.], 2002. p. 311–318.
- PATEL, R.; KIM, H. Interoperability in electronic health records: Challenges and solutions. **IEEE Transactions on Information Technology in Biomedicine**, v. 28, n. 3, p. 350–362, 2023.
- POPE, T.; PATOOGHY, A.; SARRAFZADEH, A. On the trustworthiness of fhir-based internet-of-things digital health systems. In: **2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS)**. [S. l.: s. n.], 2023. p. 279–283.
- RAHUTOMO, F.; KITASUKA, T.; ARITSUGI, M. *et al.* Semantic cosine similarity. In: UNIVERSITY OF SEOUL SOUTH KOREA. **The 7th international student conference on advanced science and technology ICAST**. [S. l.], 2012. v. 4, n. 1, p. 1.
- RENUKA, S.; GURUPRAKASH, C. D. Reliable ontology based data access control mechanism for improved security in healthcare data. In: **2024 2nd World Conference on Communication Computing (WCONF)**. [S. l.: s. n.], 2024. p. 1–7.
- ROESLER, J.; BUDDE, L.; FRIEDLI, T.; OGHAZI, P.; PALMIÉ, M. Scaling digital solutions in healthcare: Paradoxical tensions in provider–customer relations and coping strategies. **IEEE Transactions on Engineering Management**, v. 71, p. 12228–12242, 2024.
- ROESLER, J.; BUDDE, L.; FRIEDLI, T.; OGHAZI, P.; PALMIÉ, M. Scaling digital solutions in healthcare: Paradoxical tensions in provider–customer relations and coping strategies. **IEEE Transactions on Engineering Management**, v. 71, p. 12228–12242, 2024.
- SAHOO, P.; SINGH, A. K.; SAHA, S.; JAIN, V.; MONDAL, S.; CHADHA, A. A systematic survey of prompt engineering in large language models: Techniques and applications. **arXiv preprint arXiv:2402.07927**, 2024.
- SANTOS, F.; OLIVEIRA, L. Health data standards and snomed ct for interoperability. **IEEE Journal of Biomedical and Health Informatics**, v. 26, n. 7, p. 2710–2718, 2022.
- SANTOS, F.; ZHAO, J. Data interoperability in health systems: A systematic review of recent advancements. **IEEE Journal of Biomedical and Health Informatics**, v. 28, n. 1, p. 55–67, 2024.
- SARACCO, R.; AUTIOSALO, J.; KERCKHOVE, D. de; FLAMMINI, F.; NISIOTIS, L. Personal digital twins and their role in epidemics control: An iieee digital reality white paper. Institute of Electrical and Electronics Engineers (IEEE), 2020.
- SARZAEIM, P.; MAHMOUD, Q. H.; AZIM, A. A framework for llm-assisted smart policing system. **IEEE Access**, v. 12, p. 74915–74929, 2024.
- SINGH, A.; EHTESHAM, A.; GUPTA, G. K.; CHATTA, N. K.; KUMAR, S.; KHOEI, T. T. Exploring prompt engineering: A systematic review with swot analysis. **arXiv preprint arXiv:2410.12843**, 2024.

- SMITH, J.; LEE, C. Advancements in dicom standards for medical imaging interoperability. **IEEE Access**, v. 11, p. 2450–2463, 2023.
- SONG, M.; WANG, J.; YU, Z.; WANG, J.; YANG, L.; LU, Y.; LI, B.; WANG, X.; WANG, X.; HUANG, Q.; LI, Z.; KANELAKIS, N. I.; LIU, J.; WANG, J.; WANG, B.; YANG, J. Pneumollm: Harnessing the power of large language model for pneumoconiosis diagnosis. **Medical Image Analysis**, v. 97, p. 103248, 2024. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1361841524001737>.
- TAN, Y.; ZHANG, Z.; LI, M.; PAN, F.; DUAN, H.; HUANG, Z.; DENG, H.; YU, Z.; YANG, C.; SHEN, G.; QI, P.; YUE, C.; LIU, Y.; HONG, L.; YU, H.; FAN, G.; TANG, Y. Medchatzh: A tuning llm for traditional chinese medicine consultations. **Computers in Biology and Medicine**, v. 172, p. 108290, 2024. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482524003743>.
- TAYLOR, N.; KORMILITZIN, A.; LORGE, I.; NEVADO-HOLGADO, A.; CIPRIANI, A.; JOYCE, D. W. Model development for bespoke large language models for digital triage assistance in mental health care. **Artificial Intelligence in Medicine**, v. 157, p. 102988, 2024. ISSN 0933-3657. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0933365724002306>.
- VARDHAN, M.; NATHANI, D.; VARDHAN, S.; AGGARWAL, A.; SIMINI, F. Large language models as synthetic electronic health record data generators. In: **2024 IEEE Conference on Artificial Intelligence (CAI)**. [S. l.: s. n.], 2024. p. 804–810.
- WANG, L.; ALI, M. Leveraging dicom for improved data interoperability in medical imaging. **IEEE Access**, v. 12, p. 300–312, 2024.
- WANG, X.; WEI, J.; SCHUURMANS, D.; LE, Q.; CHI, E.; NARANG, S.; CHOWDHURY, A.; ZHOU, D. Self-consistency improves chain of thought reasoning in language models. **arXiv preprint arXiv:2203.11171**, 2022.
- WANG, Y.; CHEN, Y.; LI, Z.; TANG, Z.; GUO, R.; WANG, X.; WANG, Q.; ZHOU, A. C.; CHU, X. Towards efficient and reliable llm serving: A real-world workload study. **arXiv preprint arXiv:2401.17644**, 2024.
- WEI, J.; WANG, X.; SCHUURMANS, D.; BOSMA, M.; XIA, F.; CHI, E.; LE, Q. V.; ZHOU, D. *et al.* Chain-of-thought prompting elicits reasoning in large language models. **Advances in neural information processing systems**, v. 35, p. 24824–24837, 2022.
- WU, D.; NIE, L.; MUMTAZ, R. A.; AGARWAL, K. A llm-based hybrid-transformer diagnosis system in healthcare. **IEEE Journal of Biomedical and Health Informatics**, p. 1–12, 2024.
- XIAO, S.; LIU, Z.; ZHANG, P.; MUENNIGHOFF, N. **C-Pack: Packaged Resources To Advance General Chinese Embedding**. 2023.
- XUE, X.; ZHANG, D.; SUN, C.; SHI, Y.; WANG, R.; TAN, T.; GAO, P.; FAN, S.; ZHAI, G.; HU, M.; WU, Y. Xiaoqing: A qa model for glaucoma based on llms. **Computers in Biology and Medicine**, v. 174, p. 108399, 2024. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482524004839>.
- YAN, Y.; HOU, Y.; XIAO, Y.; ZHANG, R.; WANG, Q. Knownet: Guided health information seeking from llms via knowledge graph integration. **IEEE Transactions on Visualization and Computer Graphics**, p. 1–11, 2024.

- ZAGANELLI, M. V.; FILHO, D. L. B. O sigilo médico e os dados sensíveis na telemedicina à luz da lei geral de proteção de dados. **Revista Eletrônica De Comunicação, Informação Inovação Em Saúde**, v. 17, n. 3, 2023.
- ZHANG, T.; KISHORE, V.; WU, F.; WEINBERGER, K. Q.; ARTZI, Y. Bertscore: Evaluating text generation with bert. **arXiv preprint arXiv:1904.09675**, 2019.
- ZHANG, Y.; LIU, H.; XIAO, Y.; AMOON, M.; ZHANG, D.; WANG, D.; YANG, S.; QUEK, C. Llm-enhanced multi-teacher knowledge distillation for modality-incomplete emotion recognition in daily healthcare. **IEEE Journal of Biomedical and Health Informatics**, p. 1–11, 2024.
- ZHANG, Z.; ZHANG, A.; LI, M.; SMOLA, A. Automatic chain of thought prompting in large language models. **arXiv preprint arXiv:2210.03493**, 2022.
- ZHENG, L.; CHIANG, W.-L.; SHENG, Y.; ZHUANG, S.; WU, Z.; ZHUANG, Y.; LIN, Z.; LI, Z.; LI, D.; XING, E. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. **Advances in Neural Information Processing Systems**, v. 36, p. 46595–46623, 2023.
- ZHOU, H.; LI, Z.; WU, T.; WANG, C.; LI, K. Prognostic and health management of ct equipment via a distance self-attention network using internet of things. **IEEE Internet of Things Journal**, v. 11, n. 19, p. 31338–31354, 2024.