

Alberto da Rocha Miranda, Beatriz Hirasaki Leite e Gabriel Caetano Nhoncanse

FastTriage

SÃO PAULO
2025

Alberto da Rocha Miranda, Beatriz Hirasaki Leite e Gabriel Caetano Nhoncanse

FastTriage

Final Course Project submitted to the Institute of Technology and Leadership (INTELI), to obtain a bachelor's degree in Software Engineering

Advisor: Prof. Afonso Lelis Brandão

SÃO PAULO
2025

Cataloging in Publication
Library and Documentation Service
Institute of Technology and Leadership (INTELLI)
Data entered by the author.

(Cataloging record with international cataloging data, according to NBR 14724. The record will be completed later, after approval and before the final version is deposited. The completion of the cataloging record is the responsibility of the institution's library.)

Sobrenome, Nome

Título do trabalho: subtítulo / Nome Sobrenome do autor; Nome e Sobrenome do orientador. – São Paulo, 2025.
nº de páginas : il.

Trabalho de Conclusão de Curso (Graduação) – Curso de [Ciência da Computação] [Engenharia de Software] [Engenharia de Hardware] [Sistema de Informação] / Instituto de Tecnologia e Liderança.

Bibliografia

1. [Assunto A]. 2. [Assunto B]. 3. [Assunto C].

CDD. 23. ed.

Acknowledgments

We would like to express our deepest gratitude to the Institute of Technology and Leadership - Inteli for providing the environment and resources necessary for the development of this project.

We thank our advisor, Prof. Afonso, for the guidance, patience, and valuable feedback throughout the modules.

A special thanks goes to our Dra. Cristina, medical partner from Hospital das Clínicas (USP), whose insights were crucial in validating our triage flows and ensuring our solution met real-world clinical needs.

We also extend our gratitude to our families and friends for their unwavering support and understanding during the intense weeks of development. Finally, we thank each member of the FastTriage team for their dedication, collaboration, and resilience in overcoming the technical challenges faced during this journey.

Resumo

Caetano Nhoncanse, Gabriel; Hirasaki Leite, Beatriz; Da Rocha Miranda, Alberto.
FastTriage: Automação de Triagem Hospitalar com LLM. 2025. 16 f. TCC
(Graduação) – Curso de Engenharia de Software, Instituto de Tecnologia e
Liderança, São Paulo, 2025.

O setor de saúde enfrenta desafios crescentes relacionados à sobrecarga de serviços e ineficiência operacional, especialmente nos processos de triagem que são majoritariamente manuais. Este trabalho apresenta o FastTriage, uma solução tecnológica baseada em *Large Language Models* (LLM) projetada para automatizar a coleta inicial de sintomas de pacientes via chatbot. O objetivo é reduzir o tempo de espera, diminuir a carga de trabalho da equipe de enfermagem e gerar relatórios clínicos estruturados para apoio à decisão médica. A metodologia de desenvolvimento seguiu uma abordagem ágil dividida em módulos, abrangendo a validação do modelo de negócios B2B SaaS, o desenvolvimento de um MVP técnico utilizando Python (FastAPI) e MongoDB, e a validação de fluxos com profissionais de saúde. Os resultados demonstram a viabilidade técnica da solução, apesar dos desafios de integração com APIs de mensageria, e indicam um potencial significativo de economia financeira e otimização de fluxo para instituições hospitalares.

Palavras-chave: Triagem Hospitalar; Inteligência Artificial; LLM; Automação em Saúde; Chatbot.

ABSTRACT

Caetano Nhoncanse, Gabriel ;Hirasaki Leite, Beatriz ;Da Rocha Miranda, Alberto.

FastTriage: Hospital Triage Automation with LLM. 2025. 16 p. Final Course Project (Bachelor). Course Software Engineering, Institute of Technology and Leadership , São Paulo, 2025.

The healthcare sector faces growing challenges related to service overload and operational inefficiency, particularly in triage processes that are predominantly manual. This paper presents FastTriage, a technological solution based on Large Language Models (LLM) designed to automate the initial collection of patient symptoms via a chatbot . The objective is to reduce waiting times, decrease the workload of nursing staff, and generate structured clinical reports to support medical decision-making. The development methodology followed an agile approach divided into modules, covering the validation of the B2B SaaS business model, the development of a technical MVP using Python (FastAPI) and MongoDB, and flow validation with healthcare professionals. The results demonstrate the technical feasibility of the solution, despite integration challenges with messaging APIs, and indicate significant potential for financial savings and workflow optimization for hospital institutions.

Keywords: Hospital Triage; Artificial Intelligence; LLM; Healthcare Automation; Chatbot.

List of Abbreviations and Acronyms

API – Application Programming Interface

B2B – Business to Business

BMC – Business Model Canvas

CAC – Customer Acquisition Cost

CRUD – Create, Read, Update, Delete

EHR – Electronic Health Records

KPI – Key Performance Indicator

LGPD – Lei Geral de Proteção de Dados (General Data Protection Law)

LLM – Large Language Model

LTV – Lifetime Value

MVP – Minimum Viable Product

SaaS – Software as a Service

SAM – Serviceable Available Market

SOM – Serviceable Obtainable Market

SWOT – Strengths, Weaknesses, Opportunities, Threats

TAM – Total Addressable Market

USP – Universidade de São Paulo

Summary

1 Introduction	9
1.1 Context and Motivation	9
2 Solution Development	11
2.1 Definition of Market Assumptions and Hypotheses	12
2.2 Market Sizing and Analysis	12
2.3 Competitive Analysis and Differentiators	13
2.4 Technological Solution	13
2.5 The Business Plan	16
2.6 Validation and Results	19
3 Conclusion	20
4 References	21

1 Introduction

In this work, we present FastTriage, an innovative technological solution we developed to automate the initial stage of hospital triage using Large Language Models (LLM). In a scenario where we observe that healthcare systems face growing demand and operational overload of medical teams compromises care agility, our project aims to validate a B2B SaaS business model and develop a functional technical MVP capable of optimizing patient flow. By combining accessible conversational interfaces with generative artificial intelligence, we seek not only to accelerate care but also to reduce operational costs and provide structured clinical support to healthcare professionals, establishing a relevant context for innovation in digital health .

1.1 Context and Motivation:

The healthcare industry is currently facing a dual challenge: the exponential increase in demand for services and the burnout of specialized staff. Traditional triage systems rely heavily on nursing professionals to manually collect patient history, a repetitive and time-consuming task that diverts their attention from direct patient care. This manual dependency creates a linear bottleneck; the speed of triage is strictly limited by the number of available staff. Furthermore, the pressure to process patients quickly can lead to inconsistencies in data collection and errors in case prioritization.

Our motivation for developing FastTriage stems from the recent advancements in Artificial Intelligence, specifically Generative AI. We identified a unique opportunity to deploy LLMs not as diagnostic tools—which entails complex ethical and legal risks—but as highly efficient administrative assistants capable of interacting with patients in natural language. By automating the "interview" phase of triage, we aim to free up human professionals to focus on clinical assessment and critical care, thereby improving the overall efficiency of the healthcare system.

1.2 Problem Definition and Value Proposition:

The core problem identified is the inefficiency and lack of scalability in manual triage. In the current model, patients often face long waiting times before they even speak to a healthcare professional. This delay negatively impacts the patient experience and, in severe cases, can pose health risks. For hospital administration, this inefficiency translates into high operational costs and reduced patient throughput.

Value Proposition: FastTriage proposes a paradigm shift from manual data entry to automated, intelligent data collection. Our solution offers three primary value drivers:

1. **Operational Efficiency:** By automating the initial interaction, we significantly reduce the time required for triage, allowing the system to operate continuously (24/7) with minimal human intervention.
2. **Clinical Support:** The system processes unstructured patient dialogue into structured clinical reports, providing doctors with a clear, standardized summary of symptoms, pain levels, and red flags before the consultation begins.
3. **Scalability:** Unlike human staff, the software can handle hundreds of simultaneous patient interactions without performance degradation, making it highly scalable for large institutions.

1.3 Objectives of the Work:

- General: To design, implement, and validate a B2B SaaS platform that automates hospital triage using LLMs, ensuring technical feasibility, data security, and business viability.
- Specifics:
 - Strategic Planning (Module 1): To define the product roadmap, validate the problem hypothesis through personas and risk matrices, and design the initial high-level architecture and low-fidelity prototypes .
 - Technical Implementation (Module 2): To develop a robust Minimum Viable Product (MVP) using a microservices architecture. This includes setting up a FastAPI backend, a MongoDB database for dynamic data

storage, and integrating messaging APIs (WhatsApp/Messenger) to enable real-time user interaction .

- Validation and Recovery (Module 3): To validate the system's logic with healthcare professionals and address critical technical debt. Specifically, to implement a "Recovery Plan" to overcome integration challenges with the Meta API, ensuring the reliable delivery of message persistence and AI-generated summaries .

1.4 Justification and Contributions:

This work is justified by its potential to generate tangible economic and social value. Financially, our analysis indicates that despite the initial costs of cloud infrastructure and API usage, the solution becomes profitable in the medium term by reducing the man-hours required for triage. Technologically, this project contributes to the field of Software Engineering applied to Health by demonstrating a practical architecture for integrating probabilistic AI models into deterministic hospital workflows, ensuring reliability and compliance with data protection laws (LGPD).

1.5 Work Structure:

The remainder of this document is organized as follows: Chapter 2 details the end-to-end development of the solution, covering market analysis, competitive positioning, and a deep dive into the technical architecture and sprint-based implementation. Chapter 3 presents the conclusion, summarizing our achievements, the lessons learned regarding third-party API dependencies, and the roadmap for future development.

2 Solution Development

To guide development, we established fundamental hypotheses that we needed to validate:

2.1 Definition of Market Assumptions and Hypotheses:

Before commencing technical development, we established a set of core hypotheses to guide our strategic decisions:

2.1.1 Problem Hypothesis

We assumed that the current manual triage process is the primary bottleneck in patient flow and that hospital administrators are actively seeking automation to reduce costs and improve patient satisfaction.

2.1.2 Solution Hypothesis

We hypothesized that an LLM-based chatbot is the most effective interface for this task because it offers the accessibility of a familiar app (WhatsApp) combined with the analytical power to structure complex medical complaints .

2.1.3 Value Hypothesis

We posited that a B2B SaaS model is the optimal commercial strategy, as it aligns costs with usage (scalability) and lowers the barrier to entry for hospitals compared to expensive on-premise software.

2.2 Market Sizing and Analysis:

Our market analysis indicates a robust opportunity within the digital health sector.

2.2.1 Market Size (TAM, SAM, SOM):

The Total Addressable Market (TAM) includes the global healthcare sector undergoing digital transformation. Our Serviceable Available Market (SAM) focuses specifically on hospitals and urgent care centers that report high overcrowding rates. The Serviceable Obtainable Market (SOM) for our initial phase comprises institutions open to innovation pilots and academic partnerships .

2.2.2 Customer Segmentation and Profiling

We identified three distinct personas:

- The Buyer (Hospital Manager): Focused on reducing operational costs (OPEX) and improving efficiency metrics.
- The User (Healthcare Professional): Needs accurate, structured information to make faster decisions; fears workload overload.
- The End-User (Patient): Desires quick attention and clear communication, often frustrated by long waiting room times.

2.3 Competitive Analysis and Differentials:

While there are no direct competitors offering our exact feature set (WhatsApp Triage + LLM Structuring), we face indirect competition:

- Manual Triage: The status quo. It is familiar but highly inefficient and expensive.
- Rule-Based Chatbots: Existing market solutions often use rigid decision trees ("Press 1 for fever"). These lack the flexibility to understand nuanced patient complaints, leading to frustration.
- Our Differentiators: FastTriage distinguishes itself by using Generative AI to understand context. Unlike rigid bots, our system can parse descriptions like "a stabbing pain in my left side" and map it to clinical severity scales, offering a "Standardized Clinical Report" that enhances, rather than replaces, the doctor's assessment .

2.4 Technological Solution

This section details the engineering journey, the system architecture, and the technical decisions made to ensure the solution is scalable, reliable, and compliant with data protection standards.

2.4.1 Requirements and Specifications:

The system was designed to automate clinical triage over WhatsApp, aiming to reduce latency in patient care. The core functional requirements include:

- Inbound Handling: The system must receive HTTPS POST webhooks from Twilio containing WhatsApp messages.
- Asynchronous Processing: To handle bursty traffic without blocking the web server, message processing must be decoupled from message reception using a message broker.
- AI Integration: The system must utilize an enterprise-grade LLM (IBM Watson) to transform raw text into structured triage JSON and human-readable summaries.
- Persistence: All triage records must be stored in a schema-flexible database (MongoDB/DocumentDB), and generated artifacts (PDFs) must be stored in object storage (Amazon S3).

2.4.2 Architecture and Technology:

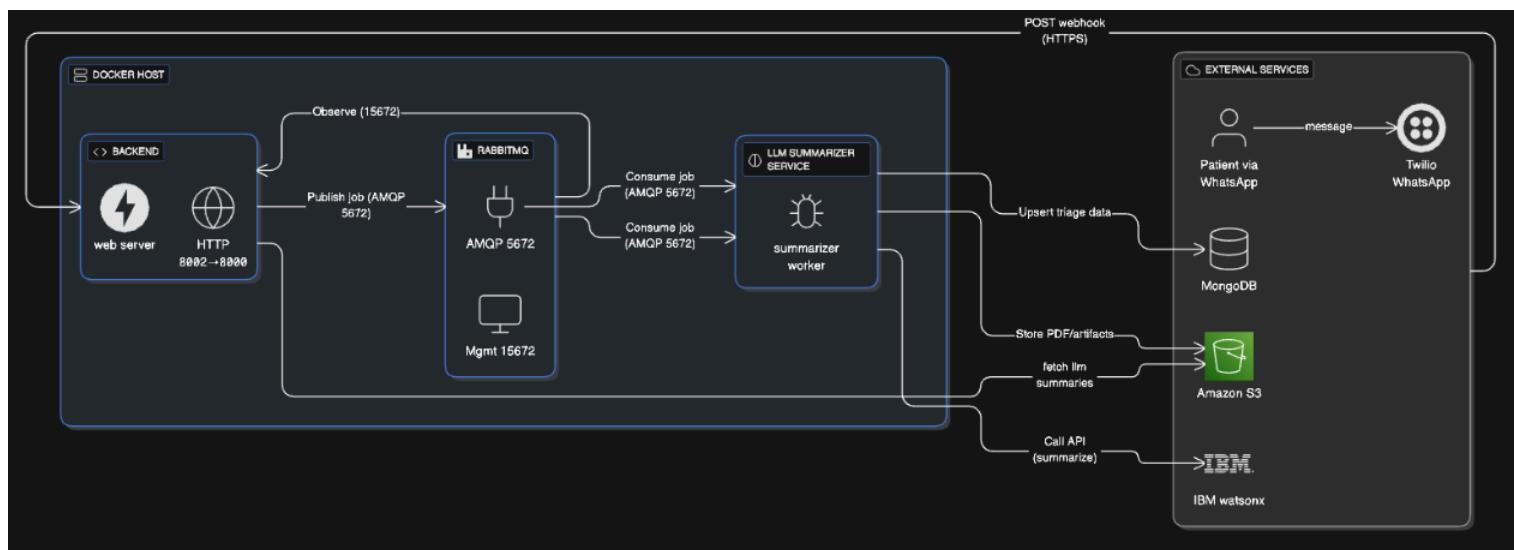
The architecture follows a modular microservices pattern orchestrated via Docker Compose, ensuring a reproducible environment. The system is divided into Internal Services (Docker Host) and External Services.

Internal Services:

- Backend (API & Webhook): Developed in Python (FastAPI), this component serves as the entry point. It validates/normalizes payloads from Twilio and publishes triage jobs to the message broker. It exposes port 8000 internally.
- RabbitMQ (Message Broker): Acts as the backbone for asynchronous communication (AMQP port 5672). It queues jobs to ensure reliability and back-pressure management during high-traffic periods, preventing system overload.
- LLM Summarizer Service (Worker): A dedicated worker that consumes jobs from RabbitMQ. It is responsible for calling the AI models (IBM Watson), generating PDF artifacts, and upserting data into the database.

External Services:

- Twilio: Manages the WhatsApp interface, forwarding user messages via webhooks and delivering replies.
- IBM watson: Provides the managed LLM runtime for summarization and structured extraction, ensuring governance and project-scoped quotas.
- Amazon S3: Used for durable, low-cost storage of generated PDF summaries and audit artifacts.
- MongoDB / AWS DocumentDB: A NoSQL store chosen for its ability to handle evolving triage schemas and nested fields.



2.4.3 End-to-End Data Flow:

The data flow is designed to be non-blocking and event-driven:

- Ingestion: The patient sends a message via WhatsApp, which Twilio forwards to the Backend via an HTTPS POST webhook.
- Queuing: The Backend validates the payload and immediately publishes a job to the RabbitMQ exchange. This ensures the webhook returns quickly, providing a fast user experience.
- Processing: The LLM Summarizer Service consumes the job from the queue. It calls the IBM watsonx API to process the text.
- Persistence: The worker stores the generated human-readable PDF summary in Amazon S3 and upserts the structured triage JSON into MongoDB.

- Completion: The system updates the triage status, making the report available for the clinical team.

2.4.4 Design Decisions and Justifications

- Asynchronous Processing (RabbitMQ): We selected RabbitMQ because WhatsApp traffic is bursty. A synchronous model would make the webhook handler fragile; the queue allows for buffering and independent scaling of the Summarizer service.
- Separation of Concerns: Splitting the I/O-bound Backend from the CPU/LLM-bound Worker ensures that heavy AI processing does not degrade the responsiveness of the web server.
- LLM Offloading (IBM watsonx): Using a managed enterprise runtime allows for central governance, auditability, and faster iteration on prompts compared to managing local GPU infrastructure.

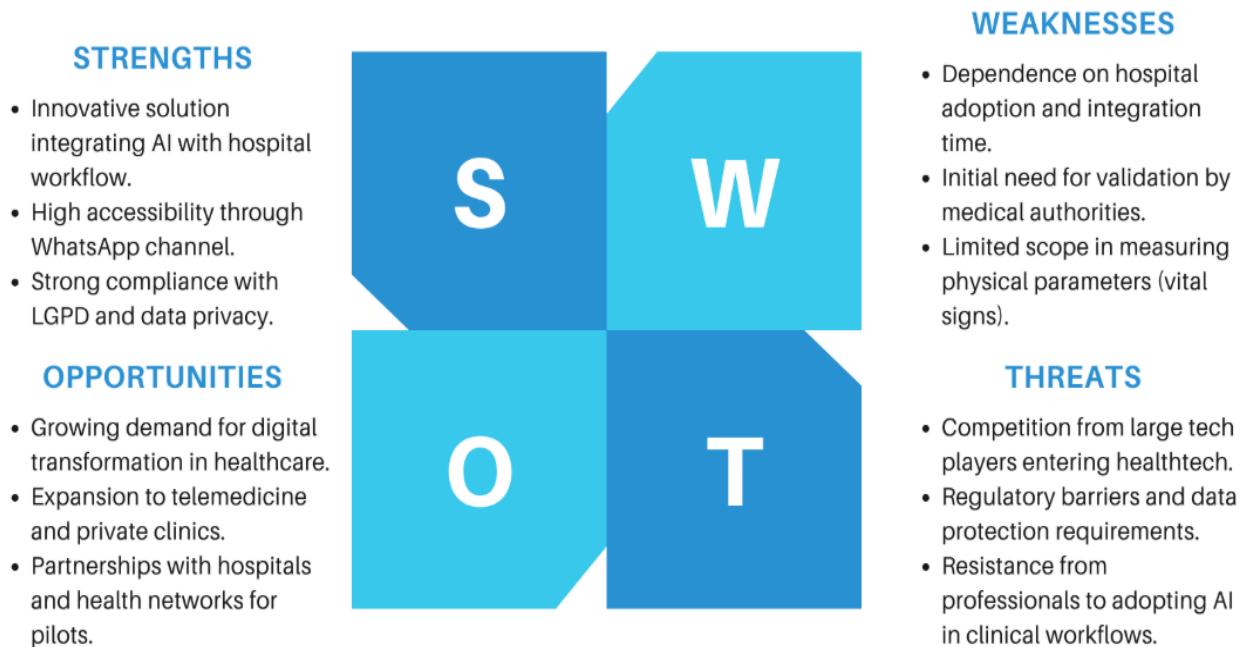
2.5 The Business Plan

2.5.1 Market and Competitor Analysis:

The healthtech sector in Brazil is expanding rapidly. According to the Distrito HealthTech Report 2024, there are over 1,000 active startups in the ecosystem. However, adoption of solutions specifically targeting hospital triage automation remains limited, creating a significant market opportunity.

- Target Audience: Our primary focus is on medium and large hospitals (Public/SUS and Private) with high emergency room volumes, as well as healthtech system integrators.
- Competitor Analysis:
 - Direct Competitors: We identified T.C. (Triage Chatbot), Infermedica, and Ada Health as key players. While they offer symptom mapping and pre-diagnosis, few have deep localization for the Brazilian context or seamless integration with WhatsApp, which is crucial for our market.

- Indirect Competitors: General-purpose AI bots (e.g., IBM Watson Assistant, Microsoft Health Bot) and manual digital forms. FastTriage differentiates itself through its specific clinical adaptive logic and strict LGPD compliance.
- SWOT Analysis:
 - Strengths: Innovative AI integration, high accessibility via WhatsApp, and strong LGPD compliance.
 - Weaknesses: Dependence on hospital adoption time and the need for validation by medical authorities.
 - Opportunities: Growing demand for digital transformation and expansion into telemedicine.
 - Threats: Competition from big tech players and regulatory barriers.



2.5.2 Business Model Canvas - BMC:

FastTriage operates on a B2B SaaS (Software as a Service) model.

- Mission: To improve the speed, accuracy, and efficiency of hospital triage through intelligent automation that supports, but does not replace, healthcare professionals.
- Vision: To significantly improve intelligent healthcare automation in Brazil, reducing patient waiting times and health risks.
- Values: Ethics and patient safety; Transparency in data handling; Collaboration with medical professionals; Technological innovation.
- Revenue Streams: Subscription-based licensing, priced according to the volume of patient triages or hospital size. Future streams include Enterprise customizations and EHR integrations.

2.5.3 Marketing and Sales Strategy:

Our strategy focuses on B2B direct sales and strategic partnerships.

- Positioning: We position FastTriage not just as software, but as an "Intelligent Triage Assistant" that augments the medical team.
- Channels: Direct sales to hospital boards, partnerships with telemedicine platforms, and participation in healthtech hubs.
- Competitive Advantages: Our main differentiators are the integration with WhatsApp (familiarity), adaptive LLM questioning (personalization), automatic generation of clinical summaries, and a strong emphasis on privacy and data security (LGPD).

2.5.4 Financial Projection and Feasibility:

The financial structure is designed for scalability.

- Cost Structure: Primary costs involve IBM Cloud infrastructure (hosting services), LLM token usage (variable cost via IBM Watson), and Twilio/WhatsApp Business API fees.
- Viability: The SaaS model allows for recurring revenue. The reduction in operational costs for hospitals—specifically fewer nursing hours spent on manual data entry—justifies the subscription investment, ensuring a positive ROI for the client and sustainability for FastTriage.

2.6 Validation and Results

2.6.1 Validation Methodology:

Validation was not merely theoretical. We maintained a continuous feedback loop with a medical specialist from Hospital das Clínicas (USP). This partner reviewed our triage decision trees, validated the terminology used by the chatbot, and ensured that the "Red Flags" logic aligned with actual emergency protocols.

2.6.2 Market Validation Results:

The validation process revealed that while the *concept* was sound, the *implementation* of the messaging channel needed adjustment. The feedback loop highlighted that relying solely on direct Meta integration was a single point of failure. Consequently, we pivoted to a more agnostic backend architecture capable of switching between messaging providers (e.g., Twilio, Zenvia) without rewriting the core business logic. This pivot was crucial for ensuring the project's deliverability within the semester's timeframe .

2.6.3 Key Performance Indicators (KPIs):

To ensure quality, we established strict technical KPIs:

- Reliability: Persistence error rate must be < 0.5%.
- Performance: The LLM must generate the clinical summary in < 5 seconds to ensure the doctor doesn't wait.
- Usability: The average time for a doctor to create/edit a triage form via the chatbot interface must be under 3 minutes.

2.6.4 Risks and Mitigation Plan:

We utilized a Risk Matrix to categorize and address potential pitfalls:

Risk: Users (patients) failing to understand the chatbot -> Mitigation:

Implementation of buttons and pre-defined options instead of open text for complex questions.

Risk: Data leakage -> Mitigation: Full encryption of data at rest and strict anonymization of logs used for debugging.

3 Conclusion

The development of FastTriage over the course of four academic modules represents a comprehensive journey from problem identification to the delivery of a technically validated solution and a structured business model.

In Module 1, we laid the strategic foundation of the project. We validated the "Triage Bottleneck" as a critical pain point in healthcare and defined the initial product vision. The creation of personas and the Risk Matrix allowed us to anticipate challenges related to user adoption and data privacy, setting the stage for a user-centric development .

In Module 2, we transitioned from strategy to core engineering. We successfully built a scalable backend infrastructure using FastAPI and MongoDB, proving the feasibility of orchestrating complex conversational flows. The integration of Docker ensured our solution was deployable in any cloud environment, a key requirement for the proposed B2B SaaS model .

In Module 3, we focused on system resilience and overcoming external blockers. Facing significant challenges with the WhatsApp API integration, we implemented a robust "Recovery Plan" that prioritized data integrity. We refined the logic for message persistence and validated the critical "LLM Summary Flow," demonstrating that Artificial Intelligence could effectively parse unstructured patient data into actionable clinical insights .

Finally, in Module 4, we consolidated our technical and business efforts. We finalized the integration of the MVP components and structured the comprehensive Business Plan, validating the financial viability of the solution. The project concludes with a fully functional product and a clear roadmap for future steps, including deep integration with Electronic Health Record (EHR) systems and the execution of large-scale clinical pilots .

4. REFERENCES

FASTAPI. *FastAPI Framework Documentation*. Available at:

<https://fastapi.tiangolo.com/>. Accessed on: 18 Dec. 2025.

META. *WhatsApp Business Platform API*. Available at:

<https://developers.facebook.com/docs/whatsapp/>. Accessed on: 18 Dec. 2025.

MONGODB. *MongoDB Documentation*. Available at:

<https://www.mongodb.com/docs/>. Accessed on: 18 Dec. 2025.

SEBRAE. *Monte um plano de negócio fácil e simples*. Portal Sebrae, [s.d.]. Available at:

<https://sebrae.com.br/sites/PortalSebrae/ufs/ap/artigos/monte-um-plano-de-negocio-facil-e-simples,17f2850c4d8f2610VgnVCM1000004c00210aRCRD>. Accessed on: 18 Dec. 2025.

TYLER, Samantha et al. Use of Artificial Intelligence in Triage in Hospital Emergency Departments: A Scoping Review. *Cureus*, v. 16, n. 5, e59906, 8 May 2024. DOI: 10.7759/cureus.59906. Available at:

<https://pmc.ncbi.nlm.nih.gov/articles/PMC11158416/>. Accessed on: 18 Dec. 2025.