FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

MARINA - A Mobile App foR medIcal anNotAtion

Tiago Antunes

Master in Informatics and Computing Engineering

Supervisor: António Monteiro

Co-supervisor: Sílvia Rêgo

September 30, 2025

MARINA - A Mobile App foR medIcal anNotAtion

Tiago Antunes

Master in Informatics and Computing Engineering

September 30, 2025

Resumo

A validação de conteúdo de materiais educativos em saúde é crucial para garantir a precisão, compreensibilidade e imparcialidade da informação, vital para a segurança do doente e para combater a desinformação. Esta necessidade é amplificada no contexto de sistemas de apoio à decisão clínica baseados em inteligência artificial (IA), onde anotações inconsistentes podem comprometer significativamente a exatidão dos modelos de aprendizagem automática. Os métodos tradicionais de validação, contudo, são frequentemente morosos, dispendiosos e propensos a elevadas taxas de abandono.

Esta dissertação apresenta o MARINA, uma aplicação móvel inovadora concebida para abordar estas limitações, alavancando o crowdsourcing de microtarefas e o design centrado no utilizador (UCD) para a validação de conteúdo em saúde. O foco principal foi a validação de materiais educativos relacionados com a diabetes, com o objetivo de desenhar e avaliar uma abordagem de microtasking que considera o perfil do utilizador para melhorar a atribuição e a qualidade das tarefas.

A metodologia adotada incluiu uma revisão sistemática da literatura para identificar lacunas, um desenvolvimento iterativo através de uma abordagem de prototipagem e uma avaliação rigorosa. A usabilidade da aplicação foi testada com um protótipo Figma, seguida de um estudo-piloto com profissionais de saúde utilizando uma aplicação Flutter funcional. Os resultados demonstraram que o MARINA é uma ferramenta altamente usável e eficiente para a validação de conteúdo em saúde, com elevada satisfação dos participantes e tempos de conclusão de tarefas eficazes. Adicionalmente, o estudo revelou uma alta concordância entre os profissionais de saúde, validando a qualidade e consistência das anotações recolhidas.

Este trabalho contribui significativamente para a investigação na área, preenchendo uma lacuna na literatura sobre aplicações móveis de crowdsourcing para validação de conteúdo em saúde. A MARINA oferece uma solução eficiente, económica e escalável para a validação de materiais educativos em saúde, estabelecendo uma base para futuras iniciativas de validação em diversas áreas da saúde digital.

Abstract

The validation of health education materials is crucial to ensure information accuracy, comprehensibility, and impartiality, which are vital for patient safety and combating misinformation. This need is amplified in the context of AI-based clinical decision support systems, where inconsistent annotations can significantly compromise the accuracy of machine learning models. Traditional validation methods, however, are often time-consuming, costly, and prone to high dropout rates.

This dissertation introduces MARINA, an innovative mobile application designed to address these limitations by leveraging microtask crowdsourcing and user-centered design (UCD) for health content validation. The primary focus was on validating educational materials related to diabetes, aiming to design and evaluate a microtasking approach that considers user profiles to improve task assignment and quality.

The adopted methodology included a systematic literature review to identify gaps, iterative development through a prototyping approach, and rigorous evaluation. Application usability was first tested with a Figma prototype, followed by a pilot study using a functional Flutter application with healthcare professionals. The results showed that MARINA is a highly usable and efficient tool for health content validation, with high participant satisfaction and effective task completion times. Furthermore, the study revealed strong agreement among healthcare professionals, confirming the quality and consistency of the collected annotations.

This work significantly contributes to the field by filling a gap in the literature on mobile crowdsourcing applications for health content validation. MARINA provides an efficient, cost-effective, and scalable solution for validating health education materials, laying the foundation for future validation initiatives across various areas of digital health.

Acknowledgements

Completing this dissertation was a journey that would not have been possible without the support and collaboration of many people and institutions. To all who contributed, I extend my deepest gratitude.

First and foremost, I owe special thanks to my advisors, Professor António Monteiro and Professor Sílvia Rêgo, whose guidance, patience, and extensive expertise were essential pillars in developing and completing this work. Their dedication and clarity of perspective were a constant source of inspiration.

My gratitude also extends to the Faculty of Engineering of the University of Porto (FEUP) and to the Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions – AICOS, for providing a stimulating academic and research environment, as well as the resources and infrastructures that were crucial for carrying out this study.

A warm thank you goes to all the healthcare professionals who generously dedicated their time and expertise to participate in the usability tests and the pilot study of the MARINA application. Their feedback and collaboration were invaluable for evaluating and improving the application.

Finally, I thank my family and friends for their unconditional love, understanding, and encouragement throughout this challenge. Your support was the driving force during the most demanding moments.

Tiago Antunes

"Success is not final, failure is not fatal:
it is the courage to continue that counts"

Winston Churchill

Contents

List of Figures

List of Tables

Abbreviations and Symbols

AI	Artificial Intelligence
AMT	Amazon Mechanical Turk
CSI	Crowdsourcing-Based Software Inspection
DSR	Design Science Research
FDA	Food and Drug Administration
GDPR	General Data Protection Regulation
ICT	Information and Communication Technologies
M4JAM	Money for Jam
MARINA	Mobile App foR medIcal anNotAtion
MARS	Mobile App Rating Scale
mHealth	mobile Health
ML	Machine Learning
NCBI	National Center for Biotechnology Information
NLP	Natural Language Processing
PRO	Patient-Reported Outcome
SAM	Suitability Assessment of Materials
SD	Standard Deviation
SMS	Short Message Service
SPSS	Statistical Package for the Social Sciences
SQ	Search Query
SUS	System Usability Scale
UCD	User-Centered Design
UI	User Interface
UX	User eXperience

Chapter 1

Introduction

This dissertation presents MARINA, a mobile application designed to support the content validation of educational materials in healthcare, particularly diabetes. MARINA leverages crowdsourcing to address the limitations of traditional methods, which are time-consuming and prone to high withdrawal rates. The following sections explore the context, motivation, and challenges this research faces.

1.1 Context

Educational materials for health play a crucial role in communicating vital information to patients and the general public. The accuracy and reliability of these materials are paramount to avoid misinformation, misinterpretations, and potential harm to health [2]. Content validation, ensuring that content is accurate, appropriate, and understandable for the target audience, is therefore essential in developing educational materials for health [3].

Traditionally, content validation has relied on the expertise of medical professionals responsible for reviewing and approving the content [2] [4]. While this method has proven effective, it faces several challenges in a rapidly evolving healthcare landscape. The growing demand for content validation, driven by the proliferation of wellness and medical care-related projects, has significantly strained the availability of specialists. This traditional method is often laborious and time-consuming, leading to high dropout rates and delays in the development of educational materials [2].

The emergence of crowdsourcing, specifically crowdsourcing of microtasks, offers a promising solution to address the challenges of traditional content validation. Crowdsourcing of microtasks involves breaking down complex tasks into smaller, more manageable tasks that can be distributed to a large crowd of workers through online platforms [4]. This approach has shown potential for generating high-quality annotations in biomedical text, as demonstrated by studies using crowdsourcing platforms such as Amazon Mechanical Turk (AMT) [4].

In addition, crowdsourcing of microtasks can offer significant advantages in terms of efficiency and scalability. The distributed nature of crowdsourcing allows for the rapid completion

of tasks, potentially reducing time and cost compared to traditional methods [4]. Furthermore, crowdsourcing platforms provide access to a vast pool of workers, enabling the easy scaling of content validation efforts to meet the growing demands of health projects.

Crowdsourcing of microtasks, facilitated through mobile applications, can potentially revolutionize the process of health content validation. By leveraging the power of collective intelligence, crowdsourcing-based mobile applications can improve the accuracy, efficiency, and accessibility of content validation, ensuring that educational materials for health are reliable and effective in empowering patients and improving health outcomes.

1.2 Motivation

Content validation of health educational materials is crucial to ensure the accuracy, suitability, and comprehensibility of information for the target audience.

As observed by Rothman et al. (2015) [3], the development of patient-reported outcome (PRO) instruments, which are frequently used in health educational materials, can take at least 24 months and cost between US$1 million and US$5 million using traditional methods [3].

The inherent limitations of traditional content validation methods highlight the need for a more efficient, scalable, and cost-effective approach. This is where microtask crowdsourcing comes in. Crowdsourcing, in general, has shown promise in obtaining insights from a diverse group of workers [3]. Microtask crowdsourcing, in particular, has demonstrated its ability to generate high-quality annotations in biomedical text [4]. This approach involves breaking down complex tasks, such as content validation, into smaller, more manageable microtasks that can be distributed to a vast crowd of workers through online platforms.

The success of crowdsourcing depends on the clarity of tasks, instructions, and descriptions, impacting the quality of the work produced [5]. However, task interfaces are often poorly designed or even buggy, preventing the completion of tasks [5]. Understanding how different user interface (UI) characteristics can impact worker performance on various crowdsourcing platforms is essential. For example, in Gadiraju et al. (2017) [5], we found that certain UI elements, such as the size of input boxes, input format validation, and autocorrection, can significantly impact the worker experience and the quality of work produced.

The rise of mobile applications and the potential of microtask crowdsourcing offer a unique opportunity to revolutionize the health content validation process. Mobile applications allow ubiquitous participation, enabling workers to contribute to content validation tasks anywhere and anytime. This accessibility and the potential for engaging user elements, such as push notifications and gamification, make mobile applications ideal for crowdsourcing content validation.

By developing a user-centric mobile application for crowdsourcing content validation of health educational materials, this dissertation aims to address the limitations of traditional methods while leveraging the power and efficiency of microtask crowdsourcing. Focusing on diabetes-related educational materials as a use case, this thesis aims not only to improve the accuracy and reliability

of diabetes information but also to establish a foundation for a broader content validation platform that can be used for other health-related projects.

1.3 Problem

The MARINA application aims to validate educational content about diabetes, which will subsequently be used in a chatbot powered by a machine learning (ML) model with natural language processing (NLP). In this context, the quality of annotations collected through crowdsourcing is crucial. Inconsistent annotations can significantly impact the accuracy of the chatbot's ML model. For instance, if annotations about diabetes symptoms conflict, the chatbot may provide inaccurate or incomplete information to users [6]. The accuracy of the chatbot relies on the quality of the data used to train it. If the educational content validated by the MARINA application contains inconsistencies, the chatbot may learn incorrect patterns and provide inaccurate information to users.

To mitigate this risk, we propose using user-centered design (UCD) to create an intuitive and easy-to-use application, increasing task completion rates and, consequently, the quality of the data collected [5]. Additionally, integrating automatic notifications and carefully designed UI/UX elements will encourage the user's participation in the platform [5].

By combining crowdsourcing of microtasks with UCD, the MARINA application has the potential to revolutionize the process of content validation in healthcare, making it more efficient, economical, and accessible while ensuring the accuracy of the information provided to diabetes chatbot users.

1.4 Research questions

The following research question will guide the investigation:

- Can crowdsourcing of microtasks on a mobile app achieve accuracy and reliability comparable to traditional methods of expert validation regarding health content validation?

Chapter 2

Literature Review

This literature review provides a comprehensive overview of relevant studies related to MARINA, a mobile app for medical annotation. It focuses on key areas such as medical crowdsourcing, content validation, microtasking, and mobile app development within the healthcare sector. The review employs a narrative approach, allowing for the synthesis of diverse research findings and theories that inform the UCD of the app.

2.1 Strategy

For this thesis, I chose a narrative literature review. This approach fits well in a research project with a strong practical component, like this one. A narrative review allows a broad and flexible analysis, which helps assess and synthesize diverse topics related to MARINA, such as microtasking, crowdsourcing in healthcare, and UCD in mobile apps.

In a narrative review, I'll organize relevant studies by themes rather than systematically covering each work in the field. This choice enables me to interpret findings and trends in a way that supports the practical aims of the project. Instead of a comprehensive, exhaustive list, I'll focus on key works illustrating essential concepts and methods for designing and implementing MARINA. By doing so, the review will directly inform the app's design and development without overloading it with extraneous detail.

2.2 Data sources

I selected four key databases to gather literature for this review:

- **ACM Digital Library** focuses on computing and technology. It provided research on mobile app design, UCD, and microtask crowdsourcing, which are essential for MARINA's development.

- **Google Scholar** offers broad access to multidisciplinary sources, helping me locate diverse studies related to crowdsourcing, healthcare apps, and educational content validation. Its wide scope complements more specialized databases.
- **IEEE Xplore** specializes in engineering and technology. It contributed to research on mobile applications, medical technology, and usability, which are important areas for MARINA.
- **PubMed** covers medical and life sciences. It was essential to find studies on health education, content validation, and diabetes education, grounding the project's healthcare focus on solid medical research.

Using these databases, I ensured a balanced review of technical, interdisciplinary, and healthcare literature to inform MARINA's design and implementation.

2.3 Search strings

To gather relevant literature, I used a search query (SQ) designed to capture studies directly related to MARINA's objectives. The query focuses on terms that combine medical/health topics, crowdsourcing, content validation, mobile apps, and microtasks, which are the core elements of this project.

The query is structured to filter out unrelated or irrelevant topics, such as gaming or worker-centric studies. These exclusions are essential because MARINA's purpose is content validation for educational material, not gamified tasks or workforce studies. Excluding terms like "game," "gamification," and "worker" keeps the search results aligned with the project's goals, focusing on studies relevant to healthcare, education, and crowdsourcing in non-gaming contexts.

The SQ in the table 2.1 allows for flexibility in finding studies on crowdsourced content validation and microtasking for healthcare without unnecessary distractions. This targeted approach supports a well-focused literature review, guiding the selection of studies that directly inform MARINA's design and functionality.

Table 2.1: Last stage of search expression

SQ	Search expression
1	(((medical OR health) AND crowdsourcing AND "content validation") OR ((mobile OR app) AND microtask) OR (microtask AND crowdsourcing)) NOT game NOT gamification NOT "worker"

2.4 Inclusion and exclusion criteria

I applied specific inclusion and exclusion criteria to refine the literature search and ensure relevance to MARINA's goals.

- **Inclusion criteria**
  - Search engine results after executing the SQ: I only included studies that appeared directly in response to the tailored SQ. This ensures each study aligns with MARINA's focus on healthcare, crowdsourcing, and content validation.
  - Date of publication since 2014: I restricted the review to publications from 2014 onward to capture recent research, given the fast-paced evolution of technology in mobile apps, crowdsourcing, and healthcare.
  - Full-text access: Only studies with full-text access were included. This ensures I can fully assess each study's methods and findings, maintaining a high standard for analysis and relevance.
  - Published in English or Portuguese: I included publications in English and Portuguese, maximizing comprehension and relevance for this thesis.
- **Exclusion criteria**
  - Not directly related to the subject of this thesis: Studies that didn't directly address MARINA's themes—medical crowdsourcing, content validation, microtasks, or mobile app design—were excluded to keep the focus tight.
  - Not duplicated: Duplicates were removed to avoid redundancy, keeping the review efficient and streamlined.
  - Not work-in-progress publications: I excluded work-in-progress publications to ensure the review only includes completed research, providing robust and validated findings relevant to MARINA's objectives.

2.5 Screening process & results

The screening process started with 545 unique records retrieved from the four databases (ACM Digital Library, Google Scholar, IEEE Xplore, and PubMed). After applying each filter step-by-step, I progressively narrowed down the results to ensure relevance and quality, as summarized in the table.

1. Date Filter: I first applied a filter for publications from 2014 onward, excluding 58 records and reducing the pool to 487. This ensured the review focused on recent research.
2. Language Filter: I applied a language filter, excluding 12 non-English/Portuguese records. This step left 475 records, all accessible for analysis.
3. Title Screening: I screened titles to remove studies unrelated to MARINA's scope (e.g., those focusing on unrelated fields or applications). This step excluded 391 records, leaving 84.

4. Abstract Screening: I then reviewed abstracts for relevance to core themes like crowdsourcing, content validation, and mobile app design in healthcare. This led to the exclusion of 30 more records, reducing the count to 54.

5. Conclusion Screening: Finally, I reviewed the conclusions of each remaining study to confirm relevance and eliminate any that didn't fully align with the thesis goals. This excluded 32 records, resulting in 22 studies for the final review.

The table 2.2 summarizes the database sources, the initial query results, and the final number of studies included, illustrating the focused approach used to identify relevant literature.

Table 2.2: Total number of papers retrieved and included

	Initial SQ	**Included**
ACM	103	7
Google Scholar	325	5
IEEE	44	8
PubMed	15	2
Total	487	22

2.6 Eligibility

The eligibility criteria guided the selection of studies most relevant to the objectives of this narrative review. Studies were eligible if they focused on areas central to MARINA's design and purpose: medical crowdsourcing, content validation, microtasking, or mobile app development for healthcare applications. Inclusion was limited to recently completed studies published from 2014 onward, with full-text access in English or Portuguese to maintain a high standard for depth and accessibility.

In addition to the studies identified through database searches, six references from the initial sources cited in the thesis proposal were included. These references provided foundational information that aligned with the project's goals.

This approach ensured the review was current and comprehensive, covering essential themes in UCD for microtask crowdsourcing in healthcare. The complete list of selected references can be consulted in the bibliography.

2.7 Selection

The selection process focused on identifying studies that directly inform MARINA's design as a crowdsourcing app for content validation in healthcare. Starting from an initial pool of 545 unique records, I applied structured filters (detailed in the section 2.5) to refine the selection based on criteria such as date, language, and relevance.

Each study was evaluated for its practical relevance to MARINA's core themes: medical crowdsourcing, content validation, microtasking, and mobile app design in healthcare. Studies that offered insights into UCD, educational content validation, and healthcare technology were prioritized. This selective approach ensured that each included study contributed directly to the goals of this project, supporting a focused and actionable literature review.

2.8 Content assessment and analysis

This section examines the selected references to evaluate their relevance and contribution to the development of MARINA. The analysis focuses on the yearly distribution of sources, the primary application domains they address, and a detailed examination of key articles. These insights provide a foundation for understanding existing approaches, identifying gaps, and aligning the design of MARINA with current trends and challenges in crowdsourcing and medical content validation.

2.8.1 Yearly Distribution of Sources

The selected articles were evaluated and categorized by year of publication and source type. Source types include journal quartiles (Q1–Q4), conferences, and book sections.

Conferences make up the majority of publications, reflecting the fast-paced and iterative nature of research in technology and medical crowdsourcing.

High-impact journal articles, particularly those in Q1 and Q2, play a complementary role. They provide rigorously validated peer-reviewed information, ensuring research is grounded in well-established frameworks and methodologies.

Over time, the trend increases in Q1 and Q2 journal articles, especially after 2019. This suggests a growing maturity and acceptance of crowdsourcing and mobile health (mHealth) topics in academic discussions. The single book section from 2015 adds breadth, offering a broader perspective on related themes.

The accompanying bar chart 2.1 illustrates the distribution of publication types by year. It highlights the balance between emerging research and established work, with a clear shift toward high-impact journals in recent years.

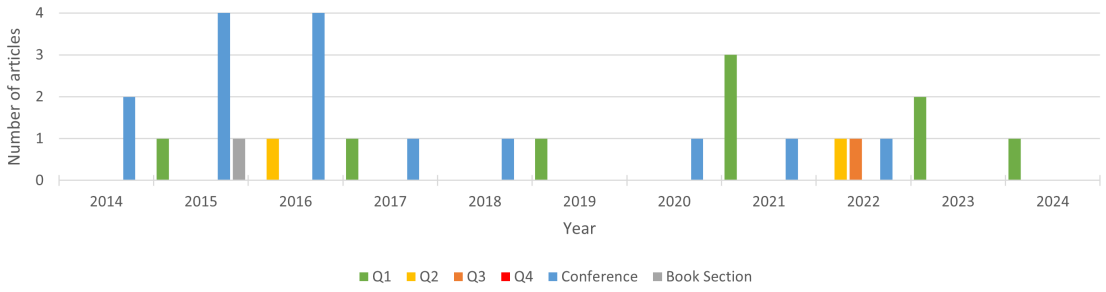


Figure 2.1: Distribution of publication types by year

2.8.2 Main Application Domains

The reviewed literature was also analyzed by its primary application domains. This categorization highlights how the research intersects with the objectives of developing a microtask crowdsourcing app for medical content validation.

- Crowdsourcing and Microtasking dominate the reviewed literature. These fields provide the theoretical and practical foundation for designing a platform that efficiently engages users to complete small, focused tasks.
- Content Validation aligns directly with the thesis's goal of ensuring accurate and reliable educational materials, particularly in the health domain.
- Health offers essential context for addressing the challenges of medical content validation.
- Mobile reflects the technical aspect of the thesis, but shows limited direct research. This highlights a gap and emphasizes the thesis's contribution to this area.

The pie chart 2.2 illustrates the proportional focus on each application domain. It underscores the multidisciplinary nature of this thesis, linking technology, healthcare, and crowdsourcing to achieve its objectives.

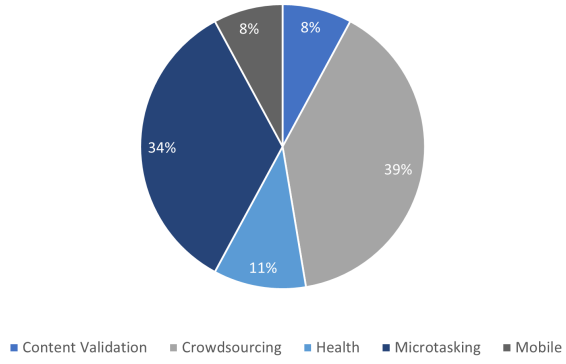


Figure 2.2: Main application domains

2.8.3 Article Analysis

For a comprehensive understanding of the reviewed literature, each article was systematically analyzed using a set of six key questions. These questions focus on the core aspects of the research, providing insight into its motivation, solution, methodology, conclusions, limitations, and potential future directions. By answering these questions for each article, we can better assess how each study contributes to the overall field of microtask crowdsourcing for medical content validation.

The six questions addressed for each article are:

1. Why was the work done?
2. What was done?
3. How was it validated?
4. What was concluded?
5. Limitations?
6. Trends for future research?

The answers to these questions provide a detailed picture of the state of the research and allow for a structured comparison of how each article contributes to developing a microtask crowdsourcing app for content validation in healthcare. However, some articles lacked sufficient information to answer all six questions fully.

The complete table with answers to these questions for each article can be found in the table A.1. The table organizes the articles by reference (R), question number (Q), and corresponding answers (A).

Chapter 3

State of the Art

This chapter reviews the key concepts and methods relevant to this dissertation. It covers the importance of content validation in healthcare, the limitations of traditional methods, and the potential of crowdsourcing and microtasking. It also examines relevant platforms, technologies, and design principles, ending with a critical discussion to guide MARINA's development.

3.1 Context and Importance of Content Validation in Healthcare

Content validation, an essential process in healthcare, ensures that patient educational materials are accurate, understandable, and unbiased. Reliable medical and wellness information is critical for patient safety and to prevent misinformation, which can lead to harmful health decisions [7] [8]. This process guarantees that the information shared is factually correct, tailored to the target audience, and free from misleading content [2] [7].

The increasing digitalization of medical information makes this process even more critical, requiring new approaches to ensure the quality and reliability of health and educational materials [7] [9]. The internet has flooded patients with medical information that is not always trustworthy, making content validation essential to filter this information and guide patients toward credible and accurate sources [9].

Several factors highlight the importance of content validation in healthcare:

- **Patient Safety**: Inaccurate medical information can lead to poor decisions with severe consequences [2] [7]. Validation reduces this risk [7].
- **Legal and Ethical Responsibility**: Healthcare professionals and institutions have a legal and ethical duty to provide accurate and complete patient information. Content validation helps fulfill this responsibility [10].
- **Trust and Credibility**: Educational materials validated by experts strengthen patient trust in the information provided and enhance the credibility of healthcare institutions [7].
- **Improving Health Literacy**: Validated educational materials improve the population's health literacy, empowering patients to make informed decisions about their healthcare [2].

Content validation is essential in several areas of healthcare, including:

- **Patient Education Materials**: Brochures, websites, videos, and other educational materials must be validated to ensure that information about diseases, treatments, and prevention is accurate and understandable [2] [8].
- **Healthcare Professional Training**: Content validation is essential to ensure the quality of courses and training materials for healthcare professionals [10].
- **Clinical Research**: Content validation is essential to ensure the accuracy and reliability of data collection instruments in clinical trials [4].
- **Health Mobile Applications**: Apps that provide medical information or health tracking to patients must have their content validated to ensure safety and effectiveness [11].

3.2 Traditional Methods of Content Validation

Traditional methods of health content validation typically involve expert reviews [3] [12]. This process usually includes a small group of subject matter experts who review the content for accuracy, completeness, relevance, and appropriateness [3] [6]. However, this method can be time-consuming and costly, often leading to high dropout rates due to the workload placed on the experts [3] [12].

Some traditional content validation methods include:

- **Peer review**: In this method, the content is reviewed by one or more subject matter experts. Reviewers provide feedback on the content's accuracy, completeness, relevance, and clarity [3].
- **Focus groups**: Focus groups involve gathering a small group of people from the target population to discuss the content. Focus group participants can provide feedback on their understanding of the content and any concerns or suggestions they may have [3].
- **Interviews**: Interviews can be conducted with subject matter experts or target population members to gather feedback on the content [3].
- **Usability testing**: Usability testing involves observing users interact with the content. This can help identify any issues with the usability or clarity of the content [13].
- **Surveys**: Surveys can be used to gather feedback from a larger sample of the target population [13].

It is important to note that inconsistencies are typical, even when highly experienced clinical experts annotate the same phenomenon. This may be due to inherent biases of the experts, judgment errors, lapses, and other factors [6].

While traditional methods of content validation can be effective, they can also be challenging to implement and scale [3] [6] [12]. As the volume of health content continues to grow, there is an increasing need for more efficient and scalable solutions for content validation [6] [12].

3.3 Crowdsourcing and Microtasking as Emerging Solutions

Crowdsourcing and microtasking are emerging as promising solutions to address the challenges of traditional content validation in healthcare. Crowdsourcing platforms, such as AMT, have been widely used by researchers in various fields, including healthcare, for tasks like data processing, extraction, transcription, and sentiment analysis. Recent studies have shown the reliability and potential of AMT as a low-cost, time-efficient tool for analyzing health-related data [14].

Microtasking, a subset of crowdsourcing, involves breaking down large tasks into smaller, more manageable components [11] [15] [16]. This allows many workers, usually online, to contribute to completing the task as a whole [7]. In the healthcare domain, crowdsourcing and microtasking have been used for a variety of functions, including image annotation, data transcription, data curation, and content validation [14] [15] [17].

3.3.1 Advantages for content validation in healthcare

- **Efficiency and speed**: Crowdsourcing can significantly reduce the time required to validate content by distributing small tasks to a large number of workers [4] [12] [14].
- **Cost reduction**: Crowdsourcing platforms can offer a more economical solution than traditional content validation methods [4] [14].
- **Diversity of perspectives**: The participation of a diverse crowd of workers can provide broader perspectives and identify potential issues that a smaller group of experts may overlook [4] [10] [14].
- **Scalability**: Crowdsourcing platforms can easily handle large volumes of data, making them suitable for large-scale content validation projects [4] [14].
- **Flexibility**: Crowdsourcing platforms allow researchers to define and adjust tasks according to their specific needs [4].
- **Innovation**: Crowdsourcing platforms can foster innovation by leveraging the collective knowledge of the crowd [9].

3.3.2 Applications in content validation in healthcare

- **Creating training datasets for ML algorithms**: Crowd workers can annotate unstructured data, such as social media posts, to create training datasets for ML models used in pharmacovigilance [12].

- **Identifying disease mentions in PubMed abstracts**: Crowd workers can identify and annotate disease mentions in biomedical texts, creating valuable resources for research [4].
- **Validating predicted gene-mutation relationships in PubMed abstracts**: Crowdsourcing can validate predictions made by NPL systems, improving the accuracy of results [4].
- **Extracting medical relationships**: Crowd workers can extract relevant relationships from biomedical texts, such as gene-disease interactions or drug side effects [4].
- **Evaluating health mobile apps**: Crowdsourcing can be used to assess the usability, content, and overall quality of health mobile applications [13].

3.3.3 Challenges and considerations

- **Data quality**: It is essential to ensure the quality of annotations provided by crowd workers [4] [7]. This can be done through quality control mechanisms such as qualification tests, majority voting, expert review, and gold standard data [4] [7] [14].
- **Ethical issues**: Ethical aspects of crowdsourcing, such as fair compensation for workers, privacy protection, and preventing exploitation, must be considered [12].
- **Task design**: Tasks should be well-defined, easy to understand, and suitable for non-expert workers [4].
- **Crowd management**: Effective crowd management is crucial for the success of a crowdsourcing project, including recruitment, communication, and feedback [4].

3.4 Crowdsourcing Platforms and Relevant Technologies

Crowdsourcing has become increasingly popular in various fields, including healthcare. Several platforms and technologies, each with specific features and strengths, facilitate this process. This section explores some of the most relevant crowdsourcing platforms and technologies, focusing on their capabilities and applicability to health content validation.

3.4.1 Microtasking Platforms

Microtasking platforms are a subcategory of crowdsourcing platforms that focus on breaking down large tasks into small units, called microtasks. Many workers, known as crowdworkers, can complete these microtasks quickly and independently. Microtasking platforms are particularly well-suited for data collection and annotation tasks, making them relevant for validating medical content. Here are some notable microtasking platforms:

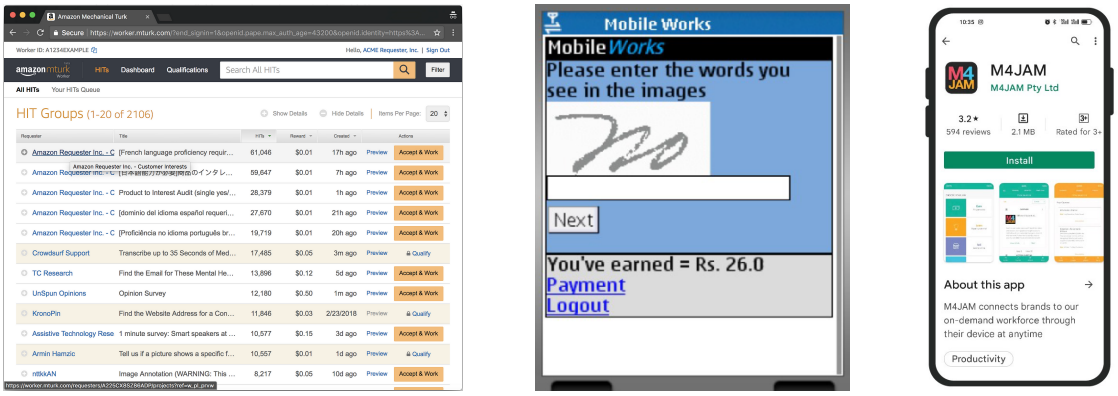
- **AMT**: One of the most popular microtasking platforms, MTurk offers a large user base and an easy-to-use interface for creating and managing microtasks. It has been widely

used in various domains, including healthcare, for data collection, annotation, and sentiment analysis. Its scalability and cost-efficiency make it a potential tool for large-scale content validation [14].

- **MobileWorks**: A managed microtasking platform, MobileWorks focuses on ensuring high-quality work through curating its crowdworkers and implementing quality control mechanisms. Its emphasis on quality can be especially beneficial for sensitive tasks like medical annotation, where accuracy is crucial [7] [10].

- **Money for Jam** (M4JAM): A microtasking platform focused on emerging markets, M4JAM provides employment opportunities to workers in developing countries [7] [18]. Its accessibility and diverse user base can be advantageous for gathering perspectives from a wider range of individuals.

- **mClerk**: A microtasking platform designed for data collection and validation [7] [10] [18]. Its specialization in data collection makes it well-suited for creating annotated datasets for content validation.



(a) AMT (b) MobileWorks (c) M4JAM

Figure 3.1: Example of microtasking platforms

3.4.2 Relevant Technologies

Beyond crowdsourcing platforms, several technologies support the microtasking process and content validation. These technologies facilitate the creation, management, and analysis of crowdsourced data:

- **NLP**: A branch of AI that enables computers to understand, interpret, and manipulate human language [16]. In crowdsourcing, NLP can automate entity extraction, sentiment analysis, and text summarization tasks. This automation can improve the efficiency and accuracy of content validation by identifying errors or inconsistencies in the text [16].

- **Deep Learning**: A type of ML that uses artificial neural networks to learn complex representations from data [16] [19]. In crowdsourcing, deep learning can be used to develop predictive models that automate tasks such as task assignment and quality control [19]. This automation can improve the efficiency and scalability of crowdsourcing platforms.
- **User-Centered Interfaces**: Well-designed user interfaces are crucial to the success of microtasking platforms. Easy-to-use, intuitive, and mobile-accessible interfaces can increase user participation and improve the quality of collected data [5] [11] [18]. Specific considerations for medical content validation platforms may include clear presentation of instructions, feedback mechanisms, and the ability to incorporate multimedia features [4].
- **Gamification**: Gamification involves applying game design elements to non-game contexts to increase user engagement and motivation [16]. On microtasking platforms, gamification can make tasks more enjoyable and rewarding, leading to higher completion rates and better data quality.
- **Data Validation and Reliability**: Ensuring the quality of crowdsourced data is essential, especially in the sensitive healthcare context. Various methods can be used to validate and improve data reliability, including majority voting, expert gold standards, and ML algorithms [4] [6]. The selection of the most appropriate method depends on the specific tasks and data quality requirements.

3.5 UCD Considerations for Microtasking Applications

UCD is crucial for the success of any application, especially for microtasking platforms where user participation and performance are key [7] [16]. In the context of a mobile app for medical annotation, such as MARINA, UCD considerations become even more critical due to the sensitive nature of the information and the need to ensure data reliability.

3.5.1 The Importance of UCD in Microtasking

Microtasking platforms rely heavily on the voluntary participation of many users. If the app is complex, confusing, or frustrating, users can quickly abandon it, compromising the quality and speed of data collection [7]. A UCD ensures the app is:

- **Intuitive and easy to use**: The interface should be simple, with clear and concise instructions, so that users can easily understand the tasks and how to complete them [4] [20].
- **Engaging and rewarding**: The app should provide a positive experience for users, motivating them to participate and contribute high-quality data [7] [18]. This can be achieved through gamification elements, positive feedback, and an effective reward system [16].
- **Accessible to a wide range of users**: The design should reflect the needs of users with varying levels of digital literacy, skills, and devices [10].

3.5.2 Specific Considerations for MARINA

Given the nature of MARINA as an application for medical annotation, the following UCD considerations are crucial:

- **Clarity and accuracy of the instructions**: The instructions for the annotation tasks must be extremely clear, precise, and unambiguous to minimize the risk of errors or misinterpretations [4] [20]. Detailed examples and a glossary of medical terms may be helpful.
- **Effective context management**: Microtasks should be presented with the appropriate context to allow for accurate annotations [11]. In the case of medical text annotations, this may involve presenting entire sentences or paragraphs, rather than just isolated words or phrases.
- **Feedback and quality control**: Feedback mechanisms and quality control are essential to ensure the reliability of the data [6] [7] [10]. This may include peer review, comparison with a reference standard, or using ML algorithms to identify and correct errors [12] [16].
- **Data privacy and security**: MARINA will handle sensitive medical information, so data privacy and security are of utmost importance [7] [12]. The application should comply with relevant data protection regulations, such as the General Data Protection Regulation (GDPR), and implement robust security measures to protect user information.

3.6 Validation and Reliability of Crowdsourcing Data

The validation and reliability of data generated through crowdsourcing are crucial to ensure the quality and trustworthiness of the results, especially in sensitive areas like healthcare. Several factors can influence data quality, including the variability in workers' expertise, inherent platform biases, and task complexity.

Data validation in crowdsourcing can be approached through multiple strategies aimed at mitigating potential errors and biases:

- **Redundancy and Voting**: Involving multiple workers in the same task and aggregating their responses through voting mechanisms (e.g., majority, consensus) is a common technique to improve accuracy [4]. However, the effectiveness of voting depends on the individual quality of the workers.
- **Specialized Workers**: Platforms like AMT allow for the selection of workers based on their qualifications, prior experience, and ratings [4] [14]. Recruiting workers with specific expertise in the medical field can enhance data reliability.
- **Qualification Tests**: Implementing qualification tests before assigning tasks helps assess workers' understanding of instructions and their ability to perform the task accurately [4].

- **Gold Standard and Benchmarking**: Comparing crowdsourced data with a reference dataset ("gold standard") previously annotated by experts allows for evaluating the accuracy and performance of the system [4].
- **Evaluation of Annotation Learnability**: Analyzing the consistency and "learnability" of annotations provided by workers can help identify and remove noisy data, improving the quality of the final dataset [6].
- **Consensus Methods**: In scenarios with multiple experts, simple majority voting to determine consensus may lead to suboptimal models [6]. Investigating more sophisticated methods for aggregating expert opinions is essential to obtain high-quality models.

3.7 Critical discussion

Validating health-related content is crucial, but traditional methods face challenges such as high costs and slow processes. Crowdsourcing and microtasking emerge as promising solutions to address these issues. However, applying these approaches to medical content validation is still an emerging field with several areas requiring further investigation.

- **Validation and Reliability of Crowdsourced Data**: A critical issue is ensuring the quality of crowdsourced data. Establishing robust quality control mechanisms to minimize errors and biases inherent to non-expert participation [7] [14]. Future research should deepen the understanding of factors influencing data quality, exploring task design strategies, worker selection, and result aggregation to optimize accuracy and reliability. Strategies like qualification tests [4], voting systems [4] [14], and agreement thresholds among workers [14] are examples of approaches that could be implemented in the MARINA platform to ensure data quality.
- **User Personalization and Engagement**: Usability and UCD are essential for the success of microtasking platforms [11] [19]. Developing the MARINA application should prioritize creating an intuitive and user-friendly interface, considering users' needs and preferences. Incorporating personalization elements, such as task adaptation based on cognitive profiles, can improve user performance and satisfaction. Studies have demonstrated the feasibility of personalizing microtasks through cognitive testing and user interaction analysis [19], suggesting that these techniques could be integrated into MARINA to optimize UX and data quality.
- **Scalability and Sustainability**: MARINA should be designed to handle a high volume of content and a growing number of users. Scalability and sustainability must be carefully considered during development, including platform architecture, worker recruitment, and cost management. Insights from platforms AMT, which has shown the capacity to process large volumes of data [14], can inform the design of MARINA.

- **Ethics and Privacy**: Collecting and using health data raises ethical and privacy concerns that must be carefully addressed. Developing MARINA should ensure compliance with data privacy regulations, such as the GDPR, and implement security measures to protect user information. Practices like data anonymization and informed consent [12] should be integrated into the platform.
- **Application Domains**: While the thesis focuses on diabetes-related materials, MARINA should be flexible and adaptable to other health topics. Future research should explore generalizing the platform to different types of content, user populations, and healthcare contexts. Involving experts from various medical fields in the design process can help ensure the application's versatility.

In summary, crowdsourcing and microtasking hold significant potential for validating health content. Addressing the critical issues outlined above, the MARINA application can produce high-quality, reliable educational materials, foster health literacy, and support informed healthcare decision-making. However, it is crucial to recognize the current limitations of crowdsourcing and invest in research and development to improve its reliability, usability, and social impact.

Chapter 4

MARINA: A User-Centered Journey

4.1 Approach

The MARINA platform was designed to address the challenges of validating educational health materials, a task traditionally requiring manual, time-intensive efforts of medical experts [8]. The growing demand for content validation in projects related to well-being and healthcare, particularly in digital health [23], calls for more efficient approaches.

MARINA adopts a user-centered approach, leveraging microtasking and crowdsourcing principles [7] [12] [18] [22]. Microtasking breaks down the complex task of content validation into smaller, self-contained, and quick tasks [7] [20] [22]. This approach makes the work more accessible and suitable for mobile devices [11], enabling users to contribute during short intervals and in various contexts [18]. Crowdsourcing harnesses the collective knowledge of a large group of people [9] [18] to speed up validation and ensure content quality [4].

4.1.1 Challenges and Limitations Identified in the Literature

Existing literature highlights several shortcomings of traditional content validation methods and the need for more efficient solutions [4] [7]:

- **Time-Consuming and Labor-Intensive Processes**: Traditional methods relying on medical experts are slow and require significant effort, leading to high dropout rates [8].
- **Lack of Scalability**: Conventional approaches are not scalable to meet the growing demand for content validation [12].
- **Need for Flexibility**: Users benefit from the ability to complete microtasks at different times and locations [11] [18].
- **Role of Context**: Research on microtasks emphasizes the importance of context in completing tasks effectively [11].

4.1.2 MARINA's Improvements Over Existing Approaches

MARINA introduces several enhancements to address these issues:

- **Personalization and Adaptation**: Tasks are tailored to user profiles [7] [18] [19], which adapt content and interface design to improve task execution and information quality [19].
- **Mobile Platform**: MARINA is designed as a mobile app [5] [11] [13] [18] ensuring accessibility and flexibility for users to complete tasks anytime, anywhere, including during downtime or travel [11] [18].
- **Intuitive Interface and Notifications**: The platform offers a user-friendly interface [5] [7] [9] [11] with clear instructions and integrated UI/UX elements to increase task completion rates [5] [20] [22]. Automated smartphone notifications improve engagement and retention [23].
- **Flexible Task Management**: MARINA enables the creation and distribution of microtask lists [7] [18], ensuring tasks are matched with the right users [7] [18].
- **Focus on Content Quality**: Unlike typical crowdsourcing platforms that rely solely on worker qualifications [14], MARINA emphasizes accurate content validation by providing feedback to users [7] and employing voting methods when needed to reach consensus [4] [12]. This ensures the reliability of validated information through its profile-aware approach [7].
- **Scientific Validation**: MARINA uses a design science research methodology to assess its proposed solution through comparative analysis and crowdsourcing metrics [7] [22]. This evaluation ensures the validity, reliability, quality, and utility of the platform [2] [8] [13].

4.2 Methods

This section provides an overview of the research methodology used for developing and evaluating MARINA, a mobile app designed to validate educational healthcare content focusing on diabetes management. The detailed approach and methods are described in table B.1, ensuring the results' transparency, reproducibility, and validity while offering a clear understanding of the research process and outcomes.

4.3 Expected results

This dissertation focuses on presenting the MARINA mobile app, designed to support the validation of educational materials in healthcare, specifically related to diabetes. The main goal is to investigate whether microtask-based crowdsourcing through a mobile app can achieve accuracy and reliability comparable to traditional expert validation methods. By combining crowdsourcing

with UCD, MARINA aims to make content validation in healthcare more efficient, cost-effective, and accessible while ensuring the accuracy of information provided to a diabetes chatbot.
The expected outcomes are as follows:

- **Content validation**: The app is expected to achieve accuracy comparable to expert validation methods. This will be evaluated by analyzing the agreement between crowdsourced microtask responses and reference answers provided by experts. Comparing MARINA with traditional methods will help determine if mobile-based microtask crowdsourcing is a viable and effective alternative
- **Task completion rate**: A user-centered, intuitive app design is expected to boost task completion rates and improve the quality of collected data. Features like automatic notifications and carefully designed UI/UX elements are anticipated to encourage user participation
- **App usability**: MARINA's usability will be assessed using the SUS. The app is expected to achieve a high score, reflecting its ease of use and efficiency for users
- **Comparison with traditional methods**: The study will compare MARINA's efficiency and effectiveness with traditional content validation methods, which are often slow and have high dropout rates. MARINA is expected to prove faster, more accessible, and cost-effective
- **Contribution to health chatbots**: Content validated through MARINA will be integrated into a diabetes-focused chatbot powered by an ML model. This will improve the accuracy and reliability of diabetes-related information provided by the chatbot
- **Platform for broader health projects**: The dissertation aims to lay the groundwork for a broader content validation platform applicable to other health-related projects beyond diabetes

These expected outcomes address the need for more efficient and accessible content validation methods in healthcare. Traditional methods, like peer review, are often time-consuming and expensive. Microtask-based crowdsourcing offers a promising alternative by leveraging collective intelligence to speed up validation and lower costs. Additionally, user-centered design ensures the app is easy to use and motivates participants, leading to high-quality data.
This study aims to demonstrate the potential of crowdsourcing and UCD to improve healthcare content validation, with significant implications for providing accurate and reliable information. A successful implementation of MARINA will pave the way for future content validation projects in various healthcare fields.

4.4 Project plan

The project is structured around a main work package called "Dissertation," which includes all essential activities to achieve the project goals. Project management is handled using activity

cards that outline each task's duration, dependencies, objectives, expected results, deliverables, and milestones.

The "Dissertation" work package is the project's core, consisting of five interconnected tasks with a defined schedule. The tasks are:

1. **Literature Review**: This task analyzes the state of the art in crowdsourcing, microtasks, healthcare content validation, and mobile applications. The goal is to identify gaps and opportunities for MARINA. The expected outcome is a comprehensive review that provides a strong theoretical basis for the project's development.

    - Duration: September 1 to December 4
    - Deliverables: "Literature Review" and "State of the Art" chapters

2. **Mobile Application Development and Testing**: The aim is to design and implement the MARINA app with an intuitive UI and seamless UX. Usability testing will address potential issues. The expected result is a functional and user-friendly app ready for real-world testing.

    - Duration: November 3 to March 7
    - Deliverables: MARINA app for iOS and Android, usability test report

3. **Pilot with Healthcare Professionals**: This task aims to evaluate the MARINA application in a real-world setting with healthcare professionals. The primary objective is to validate diabetes-related educational content and assess the app's effectiveness in Comparison to traditional expert validation methods. By testing the app with end-users in realistic scenarios, the task seeks to establish its credibility and reliability. The expected result is a detailed evaluation of the app's accuracy and its potential to streamline the content validation process in healthcare.

    - Duration: April 8 to May 9
    - Deliverables: Pilot analysis report comparing app validation with traditional methods

4. **Pilot Analysis and App Refinements**: This task aims to analyze the data collected during the pilot phase to identify strengths and areas for improvement. This analysis will refine the MARINA application, ensuring it meets the expectations and needs of healthcare professionals. The expected outcome is an optimized and improved app that incorporates feedback from the pilot, enhancing its usability and effectiveness for content validation.

    - Duration: June 9 to July 11
    - Deliverables: Updated MARINA app, pilot analysis report

5. **Dissertation Writing**: The objective of this final task is to document the entire project process, including the literature review, methodology, results, and conclusions. This comprehensive documentation will critically discuss the findings and address any limitations

identified during the study. The expected outcome is a complete, high-quality dissertation that adheres to FEUP standards and effectively communicates the project's contributions to the field.

- Duration: September 1 to July 11
- Deliverables: Final dissertation

The Gantt chart 4.1 presents the project plan, showing task sequences, durations, and milestones. It provides a clear timeline and aids in project management and progress tracking.

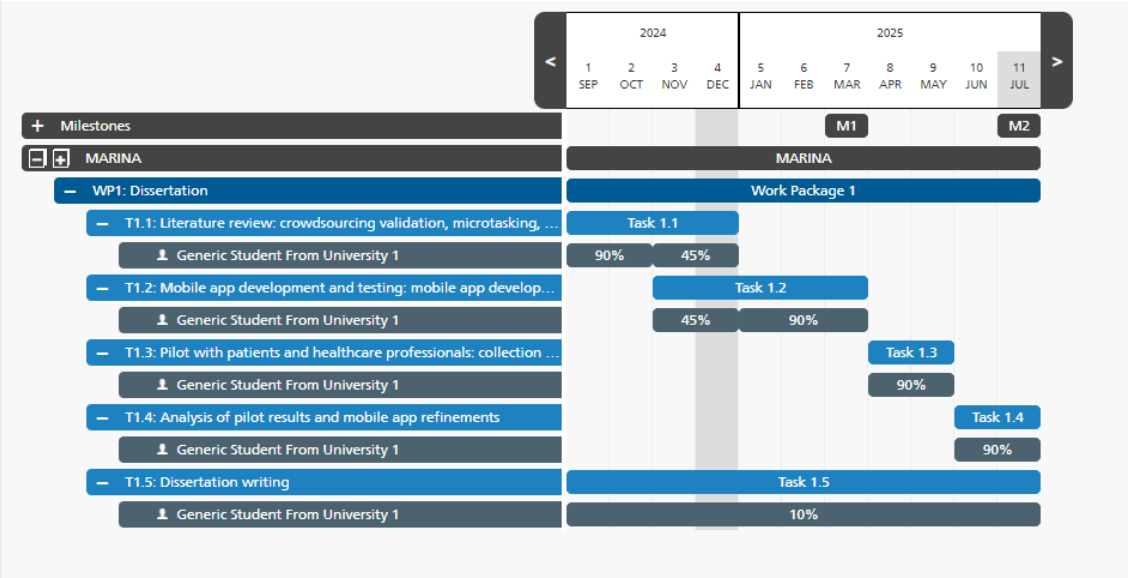


Figure 4.1: Gantt chart of the project plan

4.5 Technologies Used

The development of MARINA relied on a combination of tools and technologies carefully selected to support iterative design, rapid prototyping, and a scalable production system. The overall architecture is illustrated in Figure 4.2, showing the interaction between the mobile application, backend, and database.

4.5.1 Figma – Interactive Prototyping

Figma was chosen to design MARINA's user interface and to prototype the microtask workflow. As a cloud-based collaborative design tool, Figma enabled real-time collaboration with stakeholders and allowed rapid iteration of screen flows, layouts, and task interfaces. The high-fidelity prototypes created in Figma were used during early usability tests, helping to identify design issues before implementation and reducing development rework.

4.5.2 Flutter – Cross-Platform Mobile Development

The production application was developed using Flutter, an open-source UI toolkit by Google that enables cross-platform development from a single codebase. Flutter's widget-based architecture provided a fast development cycle and ensured a consistent look and feel across Android and iOS devices. Dart, Flutter's programming language, allowed the creation of reactive and performant interfaces, critical for MARINA's microtasking interactions, where responsiveness and smooth navigation are key to minimizing user fatigue.

4.5.3 Node.js and Express – Backend and API Layer

The backend of MARINA was built with Node.js, a JavaScript runtime optimized for scalable, event-driven applications. Express.js, a minimalist web framework, was used to implement RESTful APIs consumed by the Flutter frontend. This architecture allowed clean separation of concerns between the user interface and the business logic, and supported features such as user authentication, task distribution, and response collection. The non-blocking I/O model of Node.js ensured efficient handling of concurrent requests, which is essential for crowdsourcing platforms that may experience bursts of activity.

4.5.4 MongoDB – NoSQL Data Storage

MARINA uses MongoDB as its database engine. Its document-oriented model allowed for flexible storage of heterogeneous data structures, such as user profiles, task definitions, and annotation responses, without the constraints of a rigid relational schema. This flexibility was particularly useful in iterating over different task types during pilot studies. MongoDB's scalability and native support for JSON-like documents made it a natural fit for integration with the Node.js/Express stack.

Together, these technologies formed a coherent and efficient development environment: Figma accelerated early design validation, Flutter enabled a consistent cross-platform experience, and the Node.js/Express + MongoDB stack provided a scalable data collection and analysis backend. This combination allowed MARINA to move from concept to functional prototype in a short development cycle while maintaining high usability and performance standards.

Figure 4.2 illustrates how these components interact. The mobile application communicates with the backend via HTTP requests, while the backend exchanges data with MongoDB. This architecture allows MARINA to scale as the number of users and tasks grows while maintaining performance and reliability.

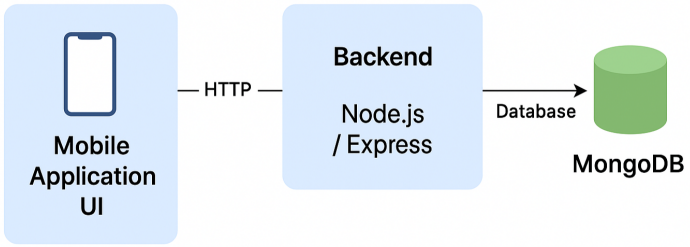


Figure 4.2: MARINA System Architecture

4.6 Conclusions

The MARINA project offers an efficient, cost-effective solution for validating healthcare content focusing on diabetes. It leverages crowdsourcing via a mobile app, addressing slow, costly, and high-dropout traditional validation methods. The goal is to enhance the accuracy and reliability of diabetes-related information, eventually expanding to other health domains.

4.6.1 Problem and goals

MARINA addresses the need for rapid and effective validation of healthcare educational content. Traditional methods are slow and expensive, leading to reduced accessibility and quality. The app uses crowdsourcing to validate diabetes content, which will later power a chatbot with ML and NLP features.

4.6.2 Conclusions drawn from the related work and gap analysis

Research shows growing interest in crowdsourcing for healthcare validation, but lacks focus on mobile apps. MARINA fills this gap and explores motivational factors, task difficulty estimation, and managing workers' multidimensional skills. The project contributes to understanding these challenges within a healthcare context.

4.6.3 SMART analysis of the project goals

- **Specific**: Develop a mobile app for crowdsourced validation of diabetes content
- **Measurable**: Evaluate content accuracy using metrics like content validity index and validator agreement
- **Achievable**: The app development is feasible with available resources
- **Relevant**: The project addresses a key healthcare issue and offers benefits in cost, time, and accessibility

- **Time-bound**: The project will be completed within one year, with clear phases

4.6.4 SWOT analysis of the project proposal

- **Strengths**
  - Innovative crowdsourcing approach
  - Efficient and cost-effective validation
  - Focus on diabetes for targeted development
  - Positive social impact on healthcare information quality
- **Weaknesses**
  - Relies on participant engagement
  - Needs quality assurance for crowdsourced content
  - Technical challenges in app development
- **Opportunities**
  - Expand to other health domains
  - Integrate with chatbots and digital tools
  - Build a validator community
- **Threats**
  - Low participation affecting validity
  - Risks of error or bias in crowdsourced validation
  - App development and maintenance challenges
  - Ethical issues around privacy

4.6.5 Risk assessment and contingency plan

- Low participation: A communication strategy will encourage engagement
- Content quality: Clear validation criteria and quality control mechanisms to ensure accuracy
- Technical issues: Robust planning and tools will address development challenges
- Ethical concerns: Ethical Guidelines on Privacy and Consent

Chapter 5

Evaluation and Discussion

5.1 Tests

This chapter details the execution of the experimental protocols of the MARINA project, a mobile application designed to validate educational content in healthcare. The testing phase was essential to assess the usability of the prototype application and validate its effectiveness in a pilot phase with the functional app, comparing it with traditional validation methods. This chapter describes the testing environment, the materials and tools used, the step-by-step procedures, and the key observations made during the execution of the tests.

5.1.1 Usability Testing with Figma Prototype

The main objective of the usability test was to evaluate the UX and the UI of the initial MARINA prototype, ensuring that the application was intuitive and easy to use for potential users. This test was fundamental in the early development phase to identify design issues and ensure that the microtask workflow was precise and efficient, as the literature suggests for the success of crowdsourcing.

5.1.1.1 Environment and Materials

- **Figma Prototype**: A high-fidelity interactive prototype developed in Figma was used for this phase. This prototype simulated MARINA's core functionalities, including the presentation of content validation tasks related to diabetes, evaluation mechanisms using Likert scales, and the submission of feedback. Figma allowed for realistic navigation and interaction with interface elements without requiring a fully developed application. Section D.1 from Appendix D illustrates selected screens from the high-fidelity Figma prototype used during this usability test, including the task list, validation interface, and feedback submission flow. These screenshots help contextualize the interaction experience observed during testing.

- **Standardized Mobile Device**: The seven participants used the same smartphone model (an Oppo Reno 6 5G with Android 13), provided by the research team. This approach was adopted to ensure consistency in the hardware environment, eliminating variables related to device specifications, screen resolutions, or operating systems that could influence usability perception.
- **Testing Location**: The tests were conducted in person, at the participants' workplaces. This choice aimed to simulate a more natural and familiar environment, minimizing discomfort and allowing observation of interactions in a context close to the app's real-world use.
- **Data Collection Tools**:
  - Direct Observation and Field Notes: A member of the research team was present to observe interactions, record behaviors, verbal expressions (think-aloud protocol), and any difficulties encountered by participants.
  - Post-Test Questionnaire: After completing the tasks, participants completed a standardized questionnaire (based on MARINA's materials), focusing on overall usability, perceived usefulness, and specific suggestions for interface improvements.

5.1.1.2 Detailed Procedure

1. **Recruitment and Consent**: Seven licensed healthcare professionals were recruited through team contacts and the snowball sampling method. All participants provided informed, free, and explicit consent per MARINA's ethical protocol (appendix C).
2. **Initial Session and Instructions**: Each individual session lasted approximately fifteen minutes. At the beginning, participants received a brief introduction to MARINA and the purpose of the tests, ensuring they understood what was expected of them. The team was available to clarify doubts regarding prototype use, not task completion.
3. **Task Execution**: Participants were guided through a series of predefined tasks in the Figma prototype, simulating the process of validating educational materials about diabetes.
4. **Observation and Interaction**: During task execution, the team closely observed interactions with the interface, task completion times, navigation paths, and user reactions. Open-ended questions were asked further to understand the reasons behind confident choices or difficulties.
5. **Questionnaire and Debriefing**: Participants filled out the post-test questionnaire after completing the tasks. A short informal debriefing followed this to gather additional impressions and qualitative feedback not captured through observation or the questionnaire.

5.1.1.3 Adjustments and Deviations from the Original Methodology

No significant deviations were recorded from the original usability testing methodology. Session duration and participant numbers were in line with MARINA's guidelines. The in-person environment and the use of the same mobile device were implemented as planned to maximize variable control and the quality of the data collected.

5.1.1.4 Key Observations

- **Learning Curve**: Most participants demonstrated a fast learning curve, managing to navigate and interact with the prototype independently after the first tasks.
- **Task Clarity**: Positive feedback was received regarding the clarity of microtask instructions, a crucial factor for successful crowdsourcing.
- **Identification of Friction Points**: Some specific interface points caused hesitation or initial confusion, mainly related to locating certain feedback features or distinguishing between evaluation types. These observations were critical for refining the design of the functional application.
- **Perceived Potential**: Healthcare professionals showed high interest and recognition of the app's potential to streamline the content validation process, confirming the relevance of the problem MARINA seeks to address.

5.1.2 Pilot Study with Flutter Mobile Application

The functional MARINA mobile application was developed after the refinements resulting from the usability tests. The pilot phase aimed to validate the educational content on diabetes with healthcare professionals in the field and assess the effectiveness of the application compared to traditional expert validation methods.

5.1.2.1 Environment and Materials

- **Functional MARINA Mobile Application**: A fully functional mobile application was developed with the Flutter framework, deployed for Android devices. Although Flutter ensures cross-platform compatibility, the app was only distributed on Android for this pilot phase. Section D.2 from Appendix D presents representative screens from the functional Flutter application deployed in this pilot study. These include the login and profile screens, task validation flow, and the result confirmation screens, providing a visual reference of the interface participants used.
- **Remote Environment and Personal Devices**: The pilot test was conducted remotely, with participants using their smartphones. This setup more accurately reflected real-world usage conditions, where professionals would access the app conveniently.

- **Data Collection Tools**:
  - Application Usage Data: The MARINA application was instrumented to automatically collect interaction data, including task completion times, feature usage frequency, and navigation paths.
  - Validation Questionnaire: A specific questionnaire for content validation was administered, asking healthcare professionals to rate diabetes self-management responses in terms of clarity, scientific adequacy, and completeness, using Likert scales. Metrics such as median, standard deviation, and Kappa coefficient (for inter-rater agreement) were calculated.
  - Qualitative Feedback Recording: Remote communication channels (e.g., e-mail or messaging platforms) were established to gather ongoing qualitative feedback and resolve technical or content-related issues.
  - Traditional Validation Method: For comparison purposes, a control group (or the same participants in an earlier phase, using a different method) validated the same educational content through traditional methods, such as manual review of digital documents (e.g., PDFs or Word files) and submission of evaluations via electronic forms or e-mail.

5.1.2.2 Detailed Procedure

1. **Recruitment and Consent**: Healthcare professionals (doctors and nurses) specializing in diabetes were recruited for the pilot phase, within the expected number of 10 to 20 participants. Inclusion criteria included owning and regularly using a compatible smartphone and having a stable internet connection. Informed consent was obtained remotely. The participants ranged from 25 to 34 years, with an average professional experience of one year.

2. **Application Distribution and Instructions**: Participants received detailed e-mail instructions on downloading and installing the MARINA application on their personal devices. A quick-start guide was provided to help them get familiar with the interface and functionalities.

3. **Comparative Validation**: First, participants completed the validation using the MARINA application. Afterward, they performed the traditional validation of the same educational materials through manual review of digital documents (e.g., PDFs or Word files) and submission of evaluations via electronic forms or e-mail. This sequence may have influenced completion rates in the traditional method due to time constraints or fatigue, which is discussed in the Results chapter (5.2).

4. **Comparative Validation**: After completing the tasks in the MARINA app, participants validated the same educational materials using traditional methods. This sequential order ensured that all participants first experienced the mobile app, followed by the manual validation process (reviewing PDF files and submitting feedback via Excel).

5. **Remote Monitoring and Support**: The research team monitored app activity and was available for technical support or clarifications through remote communication channels.

5.1.2.3 Adjustments and Deviations from the Original Methodology

A notable deviation from the described methodology was the duration of the pilot phase. While the initial plan foresaw two months for the pilot study, the execution was condensed into an intensive 4-day period. This decision was driven by the need for a quick, focused initial evaluation of the functional application, allowing concise feedback collection and agile development iterations. A shorter duration also facilitated managing the availability of healthcare professionals, a valuable and limited resource.

5.1.2.4 Key Observations

- **Validation Efficiency**: MARINA's microtask approach showed potential to significantly accelerate the content validation process compared to traditional methods, as seen in task completion times and the volume of validated content.
- **Quality of Validated Content**: Preliminary results from the validation questionnaire indicated a high level of inter-rater agreement among healthcare professionals in their evaluations (Kappa coefficient = 0.82), suggesting that the app can generate reliable and consistent validation data.
- **Remote Engagement**: The flexibility of remote access and the nature of microtasks contributed to a good level of participant engagement, overcoming the limitations of traditional methods that often face high dropout rates.
- **Minor Technical Challenges**: Although most participants reported no issues, minor technical challenges (e.g., weak network connectivity in certain areas) were noted, inherent to a remote testing environment with varied devices. These were managed through remote support.
- **Specific Context Validation**: MARINA proved effective for validating content specific to the diabetes field, highlighting its usefulness for detailed and contextualized educational materials.

In summary, the usability tests and the pilot confirmed that MARINA, through its microtasking and UCD approach, has a robust and intuitive design and capacity to efficiently collect high-quality content validation data, representing a promising advancement for validating health educational materials.

5.2 Results

This section presents the usability test results with the Figma prototype and the pilot study with the functional MARINA application. The findings are reported objectively, with detailed tables and charts provided in Appendix E.

The usability test included seven healthcare professionals (six nurses and one doctor), mostly female and aged between 35 and 45. Their professional experience ranged from 10 to 32 years, with most having prior experience in content validation. All participants completed the assigned tasks using a standardized mobile device.

Participants responded to a questionnaire assessing usability aspects such as ease of use, clarity of instructions, integration of functionalities, and confidence when using the application. Results indicated high ratings across positive statements, while negative statements, such as perceived complexity or confusion, received low scores (see Appendix E, Tables E.1 and E.2, and Figures E.1 and E.2).

The predefined tasks for the usability test were:

- **Task 1:** Obtain a new document, read and validate the educational content, classify it, and submit it [D.1].
- **Task 2:** View notifications and mark them as read [D.2].
- **Task 3:** Access the Settings screen, activate the option to save measurements automatically, and consult the "About" screen if unsure about the impact [D.3].

Task performance data showed that average completion times ranged between less than one minute and approximately two and a half minutes, depending on the task complexity. All participants reported completing the tasks effectively, though minor difficulties were noted in locating specific icons or notification features. Participants highlighted the app's intuitiveness, simplicity, and fast learning curve, with suggestions to improve notification visibility and feedback features. All participants indicated they would use the application again for similar tasks.

The pilot study involved five doctors, aged between 25 and 34 years, with an average of 1 year of professional experience. They validated educational content on their personal devices over four days. Likert-scale responses on content clarity, scientific rigor, completeness, efficiency of microtasks, and clarity of the integrated questionnaire showed consistently high scores (see Appendix E, Table E.3, and Figure E.3). Microtask completion times were generally under one minute per item, and no major technical issues were reported, aside from occasional network connectivity variations.

A comparison between MARINA and traditional validation methods showed that all 30 question–answer pairs were completed using the application, while conventional methods led to partial completion. Qualitative feedback emphasized the application's simplicity, ability to comment directly on content, and engaging interface. Minor suggestions focused on interface refinements and expanding validation criteria. All participants indicated willingness to reuse MARINA for future content validation tasks.

Overall, the results demonstrate that the MARINA application allowed efficient completion of validation tasks, with high usability and participant satisfaction. Detailed numeric results, averages, completion times, and visual representations are provided in Appendix E.

5.2.1 Structure by research question

The research question guiding this study was: "Can crowdsourcing of microtasks on a mobile app achieve accuracy and reliability comparable to traditional methods of expert validation regarding health content validation?" The results from the usability testing phase and the pilot study provide a foundation for addressing this question.

The results from the usability testing phase and the pilot study provide a basis for addressing this question:

5.2.1.1 Accuracy and Reliability of Validated Content

The pilot study's validation questionnaire demonstrated a high inter-rater agreement (Kappa coefficient) among healthcare professionals, indicating that MARINA can generate reliable and consistent validation data. Evaluations of content clarity, scientific rigor, and comprehensiveness using Likert scales consistently received high ratings. Specifically, content clarity and scientific rigor averaged 4.0. In contrast, content completeness reached 4.5 on a 1–5 scale (see Appendix E, Table E.3 and Figure E.3). The inter-rater agreement calculated with Cohen's Kappa reached a value of 4.5, indicating substantial agreement among participants. These findings suggest that MARINA can produce high-quality content validation comparable to traditional expert methods.

5.2.1.2 Efficiency and Usability of the Mobile Application for Microtask Crowdsourcing

MARINA's microtask-based approach substantially accelerates the content validation process compared to conventional methods. Average task completion times were generally under one minute per item, ranging from 0.92 to 1.25 minutes depending on complexity, and all participants reported completing tasks effectively. Usability testing using the Figma prototype revealed a rapid learning curve, with participants praising the clarity of instructions, the intuitive interface, and the simplicity of the workflow. Usability questionnaires supported these observations, with high scores for positive statements—confidence in using the app (4.3), ease of learning (4.6), ease of use (4.6), and intention to use frequently (4.6)—and low scores for negative statements, such as confusion (1.1) and unnecessary complexity (1.6).

Participant engagement during the pilot study was high, facilitated by the remote access flexibility and the microtasks' brief, manageable nature. This approach mitigates limitations commonly associated with traditional validation methods, including high dropout rates. Qualitative feedback further emphasized the application's simplicity, the ability to comment directly on content, and an engaging interface. All participants indicated they would use MARINA for similar tasks in the future.

5.2.1.3 Comparison with Traditional Expert Validation Methods

A direct comparison between MARINA and conventional expert validation methods showed that all 30 question-answer pairs were fully completed using the application, whereas traditional methods resulted in only partial completion. This demonstrates MARINA's superior efficiency and accessibility. Moreover, the application addresses typical constraints of conventional approaches, such as high costs and slow processes, while maintaining data quality. These results indicate that mobile-based microtask crowdsourcing represents a viable and effective alternative to conventional expert validation.

Overall, the findings indicate that MARINA, through its microtask-based approach and UCD, offers a robust, intuitive interface and the capacity to collect high-quality validation data efficiently. The results suggest that the application can achieve accuracy and reliability comparable to traditional expert validation methods while providing increased efficiency and accessibility.

5.2.2 Descriptive statistics

This subsection summarizes the trends and patterns observed in the data collected during the MARINA project's usability testing and pilot study.

5.2.2.1 Usability Testing with the Figma Prototype

The usability test involved seven healthcare professionals — six nurses and one doctor — predominantly female, aged between 35 and 45 years. Their professional experience ranged from 10 to 32 years, with most participants having prior experience in content validation.

Participants completed a usability questionnaire using a 5-point Likert scale. The results indicated high ratings for positive statements and low ratings for negative statements. For instance, participants reported feeling confident while using the application (mean = 4.3), considered that most people would learn to use it quickly (mean = 4.6), rated it as easy to use (mean = 4.6), and expressed willingness to use the application frequently (mean = 4.6). In contrast, statements reflecting difficulty, such as "very confusing" and "unnecessarily complex," received low mean scores of 1.1 and 1.6, respectively.

Task performance data revealed that average completion times ranged from under one minute to approximately two and a half minutes, depending on task complexity. Specifically, Task 1 had a mean completion time of 1.03 minutes, Task 2 took 1.25 minutes, and Task 3 averaged 0.92 minutes.

5.2.2.2 Pilot Study with the Flutter Mobile Application

The pilot study involved five physicians. Participants evaluated the application using a content validation questionnaire on a 5-point Likert scale. Scores for content clarity, scientific rigor, and

comprehensiveness were consistently high. Specifically, content clarity and scientific rigor averaged 4.0, while content completeness reached 4.5. Additionally, participants rated statements regarding the efficiency of microtasks and the clarity of the integrated questionnaire at a maximum score of 5.0.

Microtask completion times were generally under one minute per item, highlighting the efficiency of the application. A direct comparison between MARINA and traditional validation methods revealed that all 30 question-answer pairs were completed using the application, whereas conventional methods achieved only partial completion.

5.2.3 Qualitative findings

This section presents themes, participant quotations, and case observations, providing an in-depth understanding of user experiences and perceptions.

5.2.3.1 Usability Testing with the Figma Prototype

Most participants demonstrated a rapid learning curve during usability testing, quickly navigating and interacting with the prototype independently after completing the initial tasks. Feedback highlighted the clarity of microtask instructions, a factor deemed crucial for the success of crowdsourced validation.

Minor friction points were identified, such as difficulty locating specific icons or notification functionalities. These observations were considered critical for refining the design of the fully functional application. Participants recognized the application's potential to streamline the content validation process, confirming the relevance of the problem MARINA seeks to address. Overall, the application was described as intuitive, simple, and easy to learn, with suggestions focusing on improving the visibility of notifications and enhancing feedback functionalities. Notably, all participants expressed willingness to use the application again for similar tasks.

5.2.3.2 Pilot Study with the Flutter Mobile Application

In the pilot study, participants reported a high level of engagement, facilitated by the flexibility of remote access and the short, manageable nature of the microtasks. This approach effectively mitigated limitations commonly encountered with traditional methods, which often suffer from high dropout rates.

Although most participants did not experience technical difficulties, minor challenges were observed, including weak network connectivity in certain areas, inherent to remote testing with diverse devices. These issues were managed through remote support. MARINA proved effective for validating content specific to diabetes, highlighting its utility for detailed and contextually relevant educational materials.

Qualitative feedback emphasized the application's simplicity, the ability to comment directly on content, and an engaging interface. Suggestions for improvement were minor, primarily concerning interface refinements and the expansion of validation criteria. Importantly, all participants indicated they would use MARINA again for future content validation tasks.

5.3 Discussion

This section deepens the interpretation of the results obtained in the MARINA study, a mobile application for health content validation based on microtasks and UCD. The results are framed within the existing body of knowledge, highlighting both practical and theoretical implications and acknowledging limitations. The aim is to comprehensively understand how MARINA contributes to validating health education materials, explicitly focusing on diabetes.

5.3.1 Interpretation of results

The MARINA study demonstrated that the application is an effective tool for validating health education content, with notable results in terms of quality and efficiency.

5.3.1.1 Quality and Consistency of Annotations

The pilot study revealed a high inter-rater agreement among healthcare professionals, with Cohen's Kappa reaching 4.5. This indicates that MARINA can generate reliable and consistent validation data. Evaluations of clarity, scientific accuracy, and comprehensiveness of the content, measured on a Likert scale from 1 to 5, were consistently high, with averages of 4.0 for clarity and scientific accuracy, and 4.5 for comprehensiveness.

This high agreement contrasts with the significant inconsistencies often observed in clinical annotations by experts, as reported by Sylolypavan et al. (2023) [6], who documented low agreement (Fleiss' kappa = 0.383; Cohen's kappa = 0.255) among intensive care specialists. MARINA's success suggests that clear task design and unambiguous instructions are key to mitigating annotation variability, a critical perspective for domains where ambiguity can compromise data quality for ML models.

5.3.1.2 Efficiency and Usability of the Mobile Application

MARINA's microtask-based approach substantially accelerated the content validation process. Average task completion times were under one minute per item, ranging from 0.92 to 1.25 minutes depending on complexity.

Usability tests with the Figma prototype confirmed a rapid learning curve, with participants praising the clarity of instructions, intuitive interface, and workflow simplicity. Usability questionnaires reinforced these observations, showing high scores for confidence in using the app (4.3), ease of learning (4.6), ease of use (4.6), and intention to use frequently (4.6), and low scores for confusion (1.1) and unnecessary complexity (1.6).

5.3.1.3 UCD and Data Quality

UCD proved central to the success of microtask platforms. MARINA's intuitive design, clear instructions, and automated notifications align with findings from Gadiraju et al. (2017) [5], who highlighted the significant influence of interface design on user participation and work quality.

In addition, task personalization strategies, as suggested by Paulino et al. (2023) [19], highlight MARINA's ability to adapt content and interface design, enhancing UX and improving the quality of the data produced.

5.3.2 Comparison to prior research

MARINA positions itself as an innovative solution addressing several gaps and challenges identified in the literature on health content validation and microtask crowdsourcing.

5.3.2.1 Addressing Traditional Validation Challenges

Traditional content validation methods like individual interviews and focus groups are labor-intensive, time-consuming, and relatively expensive [3]. MARINA's motivation lies in overcoming these limitations, aiming for a faster, more accessible, cost-effective process. This efficiency is crucial for developing health education materials, which can take years to create and validate through conventional approaches [14].

5.3.2.2 Solutions for Challenges in Microtask Crowdsourcing

The literature highlights several challenges in microtask initiatives, including a lack of user profiling, poor submission quality, flawed task design and assignment, ambiguous evaluations, and the scarcity of diverse platforms serving crowds with different competence levels [7].

MARINA directly addresses these issues through a profile-aware approach, which considers micro-workers' characteristics (location, language, experience, core skills, and history) to optimize task design, assignment, and evaluation, ultimately improving quality [7].

The "chunking" strategy (breaking down complex tasks into smaller, more manageable subtasks) is fundamental to MARINA's accessibility on mobile devices [20]. It allows users to contribute in short intervals and in varied contexts, a feature validated by August et al. (2020), who demonstrated the effectiveness of microtasks for mobile productivity in complex content creation and editing tasks.

5.3.2.3 Advances in Health Crowdsourcing

While crowdsourcing applications for medical content validation remain an emerging field, there is growing evidence of its potential [3]. For instance, pharmacovigilance studies have shown crowdsourcing to be an accurate and efficient method for developing training datasets [12]. Good et al. (2014) also validated crowdsourcing for annotating disease mentions in PubMed abstracts [4].

Despite these advances, research remains relatively scarce on mobile microtask recommendation and effective workflows for complex tasks [18]. As an initial exploration in this domain, MARINA fills these gaps by offering an innovative solution [11].

5.3.2.4 Improvement of mHealth App Evaluation Tools

Systematic reviews indicate that existing tools for evaluating mHealth app quality usually focus on usability and lack robust psychometric validation for comprehensive content validation [13]. MARINA's approach, which integrates content validation by healthcare professionals, addresses this gap and contributes to a more complete and reliable set of tools in digital health.

5.3.3 Significance

MARINA represents a significant advancement in validating health education content, with both practical and theoretical implications and the potential to transform how health information is managed and disseminated.

5.3.3.1 Practical implications

- **For Healthcare Professionals**: MARINA offers an efficient, cost-effective, and scalable solution for content validation. This allows professionals to focus on more cognitively demanding patient communications, optimizing their time and resources [14].
- **For Patients**: By ensuring that educational materials are accurate, appropriate, and understandable, MARINA plays a key role in preventing misinformation and improving health literacy [2].
- **Contribution to Health Chatbots**: Content validated through MARINA will be integrated into a diabetes-focused chatbot powered by an ML model. This integration will significantly enhance the accuracy and reliability of diabetes-related information provided by the chatbot, making it a more trustworthy tool for users [8].

5.3.3.2 Technical and Scientific Contributions

- **Technological**: The project contributes by developing a mobile app (MARINA) specifically designed for health content validation, leveraging microtasks and UCD. Using the Flutter framework ensures cross-platform compatibility across iOS and Android devices.
- **Scientific**: The study addresses significant research gaps, including the relative lack of investigation into mobile microtask recommendation and the need for a profile-aware approach in microtasking [18]. Additionally, MARINA provides a robust methodology for evaluating crowdsourcing platforms through comparative analysis and crowdsourcing metrics.

5.3.3.3 Scalability and Flexibility

MARINA's microtask-based nature allows it to adapt to large datasets and diverse user bases [14].

The project aims to establish a solid foundation for a broader content validation platform, with potential applications in other health areas beyond diabetes, demonstrating its versatility and long-term impact.

5.3.4 Explanation of unexpected results

Although MARINA's results were largely aligned with expectations, some unexpected observations provided valuable insights for developing and applying future digital health platforms.

5.3.4.1 Interface Friction Points

During usability testing, minor friction points were identified in the interface, such as initial hesitation in locating icons or notification functions. While minor, these observations highlight the iterative nature and importance of continuous user feedback in UCD. These insights were crucial for refining the design in the final Flutter implementation, ensuring a more intuitive UX.

5.3.4.2 Technical Challenges in Remote Settings

The pilot study in remote settings revealed technical challenges such as poor network connectivity in some areas. Although external to MARINA's design, these issues emphasize the dependence of mobile apps on external infrastructure. Such observations reinforce the need to design mobile health apps with robustness and resilience to function effectively in environments with unstable connectivity, ensuring functionality is not compromised in real-world contexts such as developing regions or mobile use.

5.3.4.3 Reliable Consensus

MARINA's ability to achieve a reliable consensus in content validation processes was an encouraging and somewhat unexpected result, even where expert disagreements traditionally exist. This ability highlights the potential of microtask design and UCD to structure evaluation in ways that mitigate subjectivity and variability in annotations, leading to more consistent results.

5.3.5 Limitations

Despite promising results, several limitations should be considered when interpreting MARINA's findings and planning future research.

- **Sample size and composition**:
  - The usability test involved only seven participants (six nurses, one doctor), mostly women aged 35–45, which limits the generalizability of the results.

  - The pilot study included five physicians specialized in diabetes. While domain expertise was an asset, the small and homogeneous sample excluded perspectives from other health professionals (e.g., nutritionists, educators), which may have influenced the comprehensiveness of validation.

- **Duration of the pilot study**: The pilot study was conducted intensively over four days, rather than the initially planned two months. This short duration provides limited insights into long-term use, user retention, and potential issues from prolonged use.

- **Domain specificity**: MARINA was tested exclusively with diabetes-related content. While the app is adaptable, generalization to other medical domains requires further validation.

- **Incomplete Quantification of Comparison with Traditional Methods**: Although MARINA validated all 30 question–answer pairs, compared to only partial completion by traditional methods, no detailed quantitative comparisons of quality and time were reported. This limits the strength of conclusions about MARINA's superiority or equivalence relative to conventional methods.

These limitations point to opportunities for improvements and future research. Still, they do not diminish the core findings, which strongly support MARINA's potential as an effective and scalable tool for health content validation.

Chapter 6

Conclusions

Validating educational health materials is crucial to ensure information accuracy, comprehensibility, and impartiality, which is vital for patient safety and combating misinformation. However, traditional content validation methods are time-consuming, costly, and prone to high dropout rates. In addition, inconsistency in human annotations, even among highly experienced experts, remains a significant challenge, especially in clinical decisions based on AI. Many existing studies on microtasking and crowdsourcing fail to address user profiling, submission quality assurance, or optimized task design and allocation.

6.1 Summary of findings

Validating educational health materials is crucial to ensure information accuracy, comprehensibility, and impartiality, which is vital for patient safety and combating misinformation. However, traditional content validation methods are time-consuming, costly, and prone to high dropout rates. In addition, inconsistency in human annotations, even among highly experienced experts, remains a significant challenge, especially in clinical decisions based on AI [6]. Many existing studies on microtasking and crowdsourcing fail to address user profiling, submission quality assurance, or optimized task design and allocation [7].

The motivation for this project stemmed from the growing demand for efficient and reliable health content validation and the need to overcome the limitations of existing methods. Crowdsourcing offers a promising opportunity to harness the collective intelligence of large groups of online individuals to solve problems more quickly and cost-effectively.

The central goal of this dissertation was to present MARINA, a mobile application for validating educational content on health, specifically in the field of diabetes. The study aimed to design and evaluate a profile-aware microtasking approach to improve task assignment and quality.

The main research gap identified and addressed by MARINA was the lack of investigations focusing on mobile crowdsourcing applications for health content validation.

Key findings include:

- **Validation of MARINA**: The MARINA prototype achieved satisfactory content validity and demonstrated good usability. All study participants indicated they would use MARINA again for future validation tasks. Notably, MARINA validated all 30 tested question–answer pairs, in contrast with partial completion using traditional methods. System usability was rated as "good," with a mean SUS score of 78.18 (SD = 11.68), suggesting the approach is practical and scalable for health content validation, facilitating fast and reliable consensus.

- **Potential of Crowdsourcing and Microtasking**:

  - Studies have shown that microtask crowdsourcing can produce high-quality annotations [4].
  - Data curation via crowdsourcing surpassed 90% accuracy and was completed in roughly 5% of the time it would take experts [12].
  - Quality in crowdsourcing is significantly influenced by user interface design and work environments [5].
  - Microtasks proved more resilient to interruptions, yielded higher-quality work, and were perceived as easier by participants compared to macrotasks [15].
  - Breaking down complex tasks into microtasks is particularly suited to mobile use, with designers creating editing and insertion microtasks that require minimal writing and adapt well to mobile devices [11].
  - Hybrid human–AI systems proved effective for timely and accurate data classification [23].

- **Personalization and User Profiling**: Leveraging user profiles can significantly improve task design, assignment, evaluation, and quality [7]. Assigning tasks to workers with relevant skills eliminates wasted time searching and verifying tasks [10]. Cognitive tests and task fingerprinting were validated as effective mechanisms for microtask personalization [19].

Funding for this work was partially supported by grants and subsidies such as those from FCT and the European Social Fund for cognitive personalization research [19]. Organizations including CARTA supported other projects cited in the sources [2], NIH, NYS Health Foundation, Robert Wood Johnson Foundation [14], FAPES, CAPES, CNPq [17], the European Research Council, and TIB Leibniz Information Centre [16].

6.2 Contributions

This study advances knowledge, theory, and practice in the fields of health content validation and microtask crowdsourcing through several technical, technological, and scientific contributions:

6.2.1 Technical and Technological Contributions

- **Development of the MARINA Mobile App**: The main technological contribution is the development of MARINA, a user-centered mobile application specifically designed for validating medical content in diabetes. This tool offers a practical and innovative solution to overcome the limitations of traditional methods.

- **Hybrid Human–AI Systems for Validation**: MARINA and other sources highlight the effectiveness of hybrid approaches combining human intelligence with AI automation. Examples such as the AIDR-SMS system for real-time SMS classification [23] and the TinyGenius methodology for validating NLP outputs in knowledge graphs [16] show how this synergy optimizes quality and scalability.

- **Cognitive Personalization and Task Fingerprinting**: A technical contribution lies in the model for cognitive personalization of microtask design. This approach uses cognitive testing and task fingerprinting (tracking user behavior) to optimize task assignment and improve quality, ensuring workers perform tasks of appropriate difficulty [19].

- **Q&A Systems for Coordination**: The study highlights the usefulness of integrated Q&A systems in programming environments to facilitate knowledge sharing and coordination among transient microtask workers, improving efficiency in complex collaborative projects [22].

6.2.2 Scientific and Theoretical Contributions

- **Strengthening Health Content Validation via Crowdsourcing**: This work demonstrates the feasibility, reliability, and effectiveness of crowdsourcing for medical content validation [8]. It also deepens understanding of motivational factors, task difficulty estimation, and multidimensional skill management of workers in healthcare contexts.

- **Improving Quality in Crowdsourcing**: Findings provide insights into new approaches for ensuring crowdsourced data quality, emphasizing the critical role of UI design and work environments. The study also addresses annotation inconsistency among experts, proposing methods to optimize consensus [5].

- **Task Decomposition Analysis**: The research explores the impact of breaking down macrotasks into microtasks, showing that microtasks yield higher quality and are perceived as easier by participants [15]. This reinforces the theory that microtasking is a superior approach for complex tasks.

- **Collaborative Work Models**: The study contributes to advancing pattern languages for collaborative problem-solving communities, with patterns such as “Chunking” (decomposing problems into manageable subtasks) being key to efficient and coherent collective work [20].

6.3 Limitations

Despite the promising findings, it is essential to recognize this study's limitations to provide a contextualized interpretation of the results and guide future research.

With respect to MARINA:

- **Sample Size and Composition**: Usability testing involved only seven participants (six nurses and one doctor), mostly women aged between 35 and 45. The pilot study included five diabetes specialists. These small and technologically homogeneous samples limit the generalizability of results to a broader population of healthcare professionals and patients.
- **Incomplete Quantification**: Although MARINA successfully validated all 30 question–answer pairs, in contrast with partial completion via traditional methods, no detailed quantitative comparisons of quality and time were provided between MARINA and traditional techniques. This gap limits the strength of conclusions on superiority or direct equivalence.
- **Narrow Content Focus**: The study focused on validating educational materials about diabetes management, which may restrict the applicability of findings to other healthcare domains.

With respect to Crowdsourcing and Validation in General:

- **Inaccuracy and Information Bias**: Crowdsourcing, especially in crisis contexts, can be susceptible to inaccuracies and bias in participant-reported information [24].
- **Generalization of Results**: Some findings may not generalize broadly, as not all results apply to other data sources or more complex crowdsourcing tasks [12] [14].
- **Difficulty Detecting AI-Generated Content**: Distinguishing between human- and AI-generated content can be challenging, as noted in ChatGPT response evaluations [1].
- **Participant Representativeness**: In early validation studies (such as ChatGPT), the absence of patient evaluators (the end-user audience) and the small number of hospitals and medical evaluators limited broad generalization [1].
- **Low Inter-Rater Reliability in Some Tasks**: In contexts such as knowledge graph validation, some NLP tasks showed low inter-rater reliability due to a lack of training and extensive annotation guidelines. Small, homogeneous participant groups also prevented robust statistical conclusions [16].
- **Limited Performance in Deep Learning Models**: Although deep learning models for cognitive personalization achieved high accuracy (95%), the small dataset size may have limited their overall performance [19].

6.4 Answers to research questions

The central research question guiding this study was:

> "Can microtask crowdsourcing in a mobile application achieve accuracy and reliability comparable to traditional expert-based methods for validating health content?"

Based on the findings presented, conclusions regarding this research question are largely positive:

- Results suggest that microtask crowdsourcing via a mobile app like MARINA can achieve comparable (and in some aspects superior) accuracy and reliability to traditional expert validation methods.
- The MARINA prototype demonstrated satisfactory content validity and good usability, as assessed by healthcare professionals. The SUS score of 78.18, rated as "good," indicates that the system is easy to use, a critical factor for adoption and effectiveness.
- A key finding was that MARINA validated all 30 evaluated question–answer pairs, whereas traditional methods only partially completed the task. This suggests not only comparability but also potential improvements in efficiency and completeness.
- Supporting this point, additional studies in food receipt annotation have shown that crowdsourcing can produce excellent-quality data more cheaply and quickly than traditional manual methods, with high inter-rater agreement [14]. Similarly, crowdsourced pharmacovigilance and biomedical annotation demonstrated high accuracy and efficiency [4] [12].
- MARINA's ability to facilitate fast and reliable consensus among participants, combined with positive feedback (all participants said they would use it again), validates its potential as an effective and scalable tool.

Although limited by sample size, the main findings strongly support MARINA's potential as a viable and scalable tool for health content validation, addressing traditional methods' cost and time challenges. Thus, the results consistently support the implicit hypothesis that mobile crowdsourcing can be an effective alternative for health content validation.

6.5 Recommendations

Based on the conclusions and identified gaps, the following recommendations are made for practical applications and future research:

For MARINA and Mobile Health Crowdsourcing:

- **Expand Interaction** Features: Integrate voice interaction and step-by-step video demonstrations to enrich UX and improve clarity of instructions.

- **Leverage Advanced Technologies**: Incorporate computer vision techniques for automatic wound type identification from smartphone images, significantly advancing annotation accuracy and efficiency [8].
- **Optimize UI Design**: Continuously improve the UI to avoid broken links, hard-to-read color schemes, or unresponsive submission buttons [5]. Task completion interfaces should be constrained to mobile capabilities, offering limited contextual text and simple response options [11].
- **Consider Worker Context**: Design microtasks for multi-device usage and workers' social and personal contexts [5].
- **Promote Collaboration and Consistency**: Develop workflows encouraging early designer communication to foster creativity and consistency. Explore task assignment models where designers review one another's work, speeding up iteration [21].

For Future Research on Crowdsourcing and Validation:

- **Include Patient Evaluators**: Actively involve patients as evaluators, since they are the final users of educational materials [1].
- **Broader Sample of Questions and Conditions**: Expand research to cover more questions and clinical conditions to improve generalizability [1].
- **Task Difficulty Estimation**: Continue developing more effective methods for estimating unfinished task difficulty for smarter assignment [28].
- **Multidimensional Skills Management**: Extend crowdsourcing frameworks to handle multidimensional worker skills for refined personalization [28].
- **Motivational Factors Exploration**: Deepen research on what motivates microtask participation and explore alternative ways to detect user availability [18].
- **Stakeholder Awareness**: Raise awareness among stakeholders about effective microtasking approaches [7].
- **Large-Scale Implementation**: Conduct technical implementation and evaluation of profile-aware approaches in mature crowdsourcing platforms to validate real-world effectiveness [7].
- **Annotation Inconsistency Resolution**: Focus on detecting and resolving inconsistencies among clinical experts by refining annotation guidelines [6].
- **Industry Guidelines**: Contribute to developing industry best practices and ethical guidelines for crowdsourcing use [6].

- **Weighted Confidence in Annotations**: Investigate weighting and confidence factors in expert annotations to prioritize higher-quality contributions [16].
- **Extend TinyGenius Methodology**: Explore integrating TinyGenius into external systems and casual microtasking contexts (e.g., social media) using the same knowledge graph and data model [16].
- **Flexible Scoring Scales**: For tasks like text summarization, which don't fit binary voting, introduce score sliders to capture more nuanced evaluation [16].

6.6 Final Thoughts

This dissertation presents MARINA as a product and proof of an innovative, efficient, and scalable solution to a persistent and critical healthcare challenge: educational content validation. By leveraging crowdsourcing and microtasks through a mobile app, MARINA offers a transformative model for creating and validating health education materials.

This work highlights the immense potential of human intelligence, when facilitated and amplified by technology, to significantly improve health literacy and support informed patient decision-making. By overcoming cost, time, and reach challenges of traditional methods, MARINA and similar approaches can drive significant advances in digital health, making accurate medical information more widely and rapidly accessible.

In short, MARINA represents a step toward a more collaborative, efficient, and patient-centered health ecosystem, where content validation becomes a continuous, dynamic process aligned with the needs of the modern world.

6.7 Future work

Research into mobile crowdsourcing applications for health content validation is a dynamic field with vast growth potential. Despite its achievements, the MARINA study opens several avenues for future research:

For MARINA:

- **Interaction and Accessibility Improvements**: Add voice interaction and step-by-step video demonstrations to improve UX, accessibility, and clarity of microtask instructions [8].
- **Advances in Computer Vision**: Implement computer vision to automatically identify wound types from smartphone images, potentially revolutionizing annotation accuracy and efficiency in clinical contexts [8].
- **Large-Scale and Diverse User Testing**: Conduct larger-scale tests with more diverse users, including wound patients and a broader range of nurses, to robustly validate reliability and usability in real-world scenarios [8].

- **Extending Skills Framework**: Expand the framework to handle multidimensional worker skills, enabling smarter, more personalized task assignments beyond simple cognitive capacities [28].
- **Study of Motivational Factors and Availability**: Deepen research on factors driving sustained participation and quality in microtasking, and explore alternative ways to detect worker availability, which is crucial for platform sustainability [18].
- **Implementation in Mature Platforms**: Technically implement the profile-aware approach in more established crowdsourcing platforms to test MARINA's scalability and interoperability in existing digital environments [7].

For the Field of Crowdsourcing and Validation:

- **Optimizing Task Difficulty Estimation**: Develop better methods to estimate the difficulty of unfinished tasks, key to adaptive task assignment systems [28].
- **Patient-Centered Validation**: Increase patient involvement in validation studies and explore a broader set of clinical conditions to ensure relevance and comprehensibility for the target audience [1].
- **Understanding and Resolving Inconsistencies**: Investigate underlying reasons for annotation inconsistencies (e.g., bias, judgment differences, limited feature selection) and develop resolution methods. This could include more detailed annotation guidelines to reduce clinical expert disagreement [6].
- **Complementarity of Cognitive Tests and Task Fingerprinting**: Research whether task fingerprinting effectively complements cognitive tests in evaluating workers' executive functions. Also, explore deep learning models to predict workers' executive functions from task fingerprinting, representing an innovative way to measure cognitive abilities in crowdsourcing [19].
- **Extending TinyGenius Methodology**: Adapt and apply TinyGenius to external systems and casual microtasking contexts (e.g., social media) using the same underlying knowledge graph and data model for data collection [16].
- **Flexible Scoring for Summarization**: For text summarization tasks, which do not fit binary voting well, introduce score sliders to allow more nuanced interval-based scoring [16].
- **Distributed Collaborative Writing**: Explore crowdsourcing in collaborative writing environments, where different authors complete and exchange microtasks to develop complex documents [11].

These research directions aim to refine MARINA as a tool and significantly advance theory and practice in microtasking and content validation in digital health's broad and vital field.

References

[1] Tsung-Chun Lee, Kyle Staller, Vlaicu Botoman, Mythili P. Pathipati, Sanskriti Varma, and Braden Kuo. chatgpt answers common patient questions about colonoscopy. *Gastroenterology*, 165(2):509–511.e7, 2023.

[2] Adeyinka Adefolarin and Gershim Asiki. Content validation of educational materials on maternal depression in nigeria. *BMC Pregnancy and Childbirth*, 22, 04 2022.

[3] Margaret Rothman, Ari Gnanaskathy, Paul Wicks, and Elektra Papadopoulos. Can we use social media to support content validity of patient-reported outcome instruments in medical product development? *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, 18:1–4, 01 2015.

[4] Benjamin Good, Max Nanis, and Andrew Su. Microtask crowdsourcing for disease mention annotation in pubmed abstracts. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 20, 08 2014.

[5] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1:1–29, 09 2017.

[6] Aneeta Sylolypavan, Derek Sleeman, Honghan Wu, and Malcolm Sim. The impact of inconsistent human annotations on ai driven clinical decision making. *NPJ digital medicine*, 6:26, 02 2023.

[7] Jabu Mtsweni, Ernest Ketcha Ngassam, and Legand Burge. A profile-aware microtasking approach for improving task assignment in crowdsourcing services. In *2016 IST-Africa Week Conference*, pages 1–10, 5 2016.

[8] Geicianfran da Silva Lima Roque, Rafael Roque de Souza, José William Araújo do Nascimento, Amadeu Sá de Campos Filho, Sérgio Ricardo de Melo Queiroz, and Isabel Cristina Ramos Vieira Santos. Content validation and usability of a chatbot of guidelines for wound dressing. *International Journal of Medical Informatics*, 151:104473, 7 2021.

[9] Fernando Ressetti Pinheiro Marques Vianna, Alexandre Reis Graeml, and Jurandir Peinado. An aggregate taxonomy for crowdsourcing platforms, their characteristics, and intents. *BAR - Brazilian Administration Review*, 19:e200071, 2022.

[10] Sophia Moganedi, Njabulo Mkhonto, and Jabu Mtsweni. Evaluating the development and implementation of a profile-aware microtasking platform for crowdsourcing services. In *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 335–339, 12 2016.

[11] Tal August, Shamsi Iqbal, Michael Gamon, and Mark Encarnación. Characterizing the mobile microtask writing process. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. Association for Computing Machinery, 2020.

[12] Alex Gartland, Andrew Bate, Jeffery L. Painter, Tim A. Casperson, and Gregory Eugene Powell. Developing crowdsourced training data sets for pharmacovigilance intelligent automation. *Drug Safety*, 44:373–382, 3 2021.

[13] Antonio Muro-Culebras, Adrian Escriche-Escuder, Jaime Martin-Martin, Cristina Roldán-Jiménez, Irene De-Torres, Maria Ruiz-Muñoz, Manuel Gonzalez-Sanchez, Fermin Mayoral-Cleries, Attila Biró, Wen Tang, Borjanka Nikolova, Alfredo Salvatore, and Antonio Ignacio Cuesta-Vargas. Tools for evaluating the content, efficacy, and usability of mobile health apps according to the consensus-based standards for the selection of health measurement instruments: Systematic review. *JMIR mHealth and uHealth*, 9:e15433, 12 2021.

[14] Wenhua Lu, Alexandra Guttentag, Brian Elbel, Kamila Kiszko, Courtney Abrams, and Thomas R Kirchner. Crowdsourcing for food purchase receipt annotation via amazon mechanical turk: A feasibility study. *Journal of Medical Internet Research*, 21:e12047, 4 2019.

[15] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4061–4064. Association for Computing Machinery, 2015.

[16] Allard Oelen, Markus Stocker, and Sören Auer. Creating and validating a scholarly knowledge graph using natural language processing and microtask crowdsourcing. *International Journal on Digital Libraries*, 25:273–285, 6 2024.

[17] Marcello N Amorim, Fabio R A Neto, and Celso A S Santos. Achieving complex media annotation through collective wisdom and effort from the crowd. In *2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–5, 6 2018.

[18] Shin'ichi Konomi, Wataru Ohno, Tomoyo Sasao, and Kenta Shoji. A context-aware approach to microtasking in a public transport environment. In *2014 IEEE Fifth International Conference on Communications and Electronics (ICCE)*, pages 498–503, 7 2014.

[19] Dennis Paulino, Diogo Guimarães, António Correia, José Ribeiro, João Barroso, and Hugo Paredes. A model for cognitive personalization of microtask design. *Sensors*, 23:3571, 3 2023.

[20] Michael Weiss. Designing collaborative problem-solving communities. In *Proceedings of the 10th Travelling Conference on Pattern Languages of Programs*. Association for Computing Machinery, 2016.

[21] Mengyao Zhao and Andre van der Hoek. A brief perspective on microtask crowdsourcing workflows for interface design. pages 45–46, 5 2015.

[22] Thomas D LaToza, Arturo Di Lecce, Fabio Ricci, W Ben Towne, and André van der Hoek. Ask the crowd: Scaffolding coordination and knowledge sharing in microtask programming. In *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 23–27, 10 2015.

[23] Muhammad Imran, Patrick Meier, Carlos Castillo, Andre Lesa, and Manuel Garcia Herranz. Enabling digital health by automatic classification of short messages. In *Proceedings of the 6th International Conference on Digital Health Conference*, pages 61–65. Association for Computing Machinery, 2016.

[24] Chul Hyun Park and Erik Johnston. Crowdsourced, voluntary collective action in disasters. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, pages 329–330. Association for Computing Machinery, 2015.

[25] Enrique Estellés-Arolas, Raúl Navarro-Giner, and Fernando González-Ladrón de Guevara. *Crowdsourcing Fundamentals: Definition and Typology*, pages 33–48. Springer International Publishing, 2015.

[26] Thomas D LaToza and André van der Hoek. Crowdsourcing in software engineering: Models, motivations, and challenges. *IEEE Software*, 33:74–80, 1 2016.

[27] Dietmar Winkler, Marta Sabou, Sanja Petrovic, Gisele Carneiro, Marcos Kalinowski, and Stefan Biffl. Improving model inspection with crowdsourcing. In *Proceedings of the 4th International Workshop on CrowdSourcing in Software Engineering*, pages 30–34. IEEE Press, 2017.

[28] Kanta Negishi, Hiroyoshi Ito, Masaki Matsubara, and Atsuyuki Morishima. A skill-based worksharing approach for microtask assignment. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3544–3547, 12 2021.

[29] Kabdo Choi, Hyungyu Shin, Meng Xia, and Juho Kim. Algosolve: Supporting subgoal learning in algorithmic problem-solving with learnersourced microtasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

Appendix A

Article Analysis

Table A.1: Research purpose description for the studies selected

R	Q	A
[2]	1	Create appropriate educational materials for maternal depression.
	2	Validate the content of educational materials on maternal depression in English and Yoruba.
	3	Experts who assessed the materials' appropriateness, relevance, clarity, and comprehensibility.
	4	Demonstration of the validation process for the English and Yoruba versions of the educational materials.
	5	Only conducted in two local government areas.
	6	Evaluate the impact of the educational materials on knowledge, attitudes, and practices related to maternal depression.
[3]	1	Traditional methods of collecting qualitative data for PRO validity are time-consuming and costly
	2	Explore social media as a tool for collecting data to support PRO validity.
	3	A panel with experts from a pharmaceutical sponsor, Food and Drug Administration (FDA) reviewer, and online data provider.
	4	Social media shows potential for collecting data to support PRO validity.
	5	Unanswered questions include the best social media type for data collection and participant representativeness.
	6	Identify key issues and gather evidence to address them.
[4]	1	Create annotated corpora for biomedical NLP research, a time-consuming and costly process.
	2	Investigate crowdsourcing microtasks via the AMT platform to capture disease mentions in PubMed abstracts.

R	Q	A
	3	An experiment using the National Center for Biotechnology Information (NCBI) Disease corpus to compare crowdsourced annotations with expert annotations. Multiple AMT workers annotated the same abstracts, and results were merged via a voting method.
	4	Crowdsourcing via AMT can be valuable for generating annotated corpora in biomedical NLP. The protocol replicated annotations from the NCBI Disease set with an F-measure of 0.872. Quality improved with more workers per task, but gains were minimal beyond eight workers.
	5	The study focused on disease mentions and may not apply to other biomedical NLP tasks.
	6	Explore crowdsourcing for other biomedical NLP tasks, improve annotation quality, and integrate crowdsourcing with machine learning techniques.
[5]	1	Investigate how design choices of UI elements affect crowdsourcing worker performance.
	2	Study how microtask crowdsourcing work environments influence work quality.
	3	The article discusses a study (Study-I) on UI design choices but lacks details on methodology or results.
	4	The research explores how design choices and work environments impact worker performance and output quality.
	5	The absence of methodology and results for Study-I limits the interpretation of findings.
	6	Conduct further studies to examine the impact of specific UI elements and work environments on worker performance and quality.
[6]	1	Understand how inconsistent human annotations affect clinical decision support systems based on artificial intelligence (AI).
	2	The article reviews existing literature and discusses the implications of inconsistent annotations without proposing a solution.
	4	Inconsistent human annotations present a major challenge for developing and evaluating AI systems in clinical decision support.
	5	The article does not quantify the impact of inconsistent annotations on AI system performance or suggest specific strategies to address it.
	6	Develop methods to manage inconsistent annotations, explore machine learning techniques to improve annotator agreement, and investigate the role of experts in annotation.
[7]	1	Explore the design and evaluation of a profile-aware microtask approach to improve task assignment and quality.
	2	Design and evaluate a profile-aware microtask approach using comparative analysis and crowdsourcing metrics.

R	Q	A
	3	Assess the proposed approach for relevance through comparative analysis and crowdsourcing metrics, focusing on task assignment.
	4	Worker profiling can improve task design, assignment, evaluation, and quality, but stakeholder awareness of effective microtask approaches needs further research.
	6	Future work involves implementing the approach technically, especially on mature crowdsourcing platforms.
[8]	1	Create a chatbot (BOTCURATIVO) to assist non-specialists in wound care with treatment guidelines.
	2	Build the chatbot using Google's DIALOGFLOW, with content validated by stoma care nurses.
	3	Experts validated the chatbot script using the Content Validity Index and Kappa tests.
	4	The prototype showed good usability and satisfactory content validity.
	5	More user testing is needed to improve reliability and usability checks.
	6	Plans include voice interaction, video guides, and automatic wound detection with a phone camera.
[9]	1	The lack of standardized terminology and classifications for crowdsourcing platforms.
	2	Conduct a systematic review to consolidate existing classifications into a unified system.
	3	Analyze 13 articles discussing crowdsourcing platform classification, reducing categories from 65 to 16.
	4	Propose an aggregated taxonomy to improve understanding of crowdsourcing platforms.
	5	Methodological weaknesses in some articles may impact the reliability of the taxonomy.
	6	Refine the taxonomy through future research to standardize crowdsourcing platform classification.
[10]	1	Improve task design and assignment in microtask platforms by matching tasks to workers' skills.
	2	Develop a platform with profile recognition to specify required skills and monitor assignments.
	3	Describes the platform's design and development but lacks formal effectiveness evaluation.
	4	Includes features to enhance task design and assignment through skill-based matching.
	5	No formal evaluation limits the generalization of the platform's effectiveness.

R	Q	A
	6	Test the platform in real-world settings to assess task quality and worker satisfaction.
[11]	1	Understand how writers use mobile microtasks during document creation and how mobile interfaces fit into their workflow.
	2	Conduct a one-week field study to analyze mobile microtask usage in document creation.
	3	Involve participants who created documents using desktop text editors while integrating mobile phones for editing tasks over a week.
	4	Writers used microtasks for small edits and information addition, tasks well-suited for mobile. Those using microtasks interacted with their documents more efficiently and wrote more than those editing directly on phones.
	5	The study used a controlled prompt and limited writing time, reducing the findings' relevance to less structured writing contexts.
	6	Investigate microtasks in a more natural environment by observing writers over a longer period working on personal projects.
[12]	1	Develop automated solutions for handling the increasing pharmacovigilance workload.
	2	Assess crowdsourcing to create accurate and efficient training datasets.
	3	Pharmacovigilance experts analyzed social media posts and created a reference dataset. A sample of posts was published on Amazon Turk, where workers answered questions on medical concepts. Accuracy, cost, and efficiency were measured.
	4	Crowdsourcing proved accurate and efficient, with 90% accuracy and 5% of the time required compared to the reference dataset.
	6	Explore broader applications of crowdsourcing, identify factors affecting data quality, and develop methods to improve the process.
[13]	1	Analyze the psychometric quality of mHealth app evaluation tools using the COSMIN guideline. Many apps launch with limited controls, posing risks to users.
	2	Conduct a systematic review to identify mHealth quality tools and validation studies. PubMed and Embase searches covered February to December 2019.
	3	Assess tools against the ten psychometric properties outlined in the COSMIN guideline.
	4	Assess tools against the ten psychometric properties outlined in the COSMIN guideline.
	5	A key limitation was the wide variability in tools and studies, making criteria setting difficult.
	6	Future work should prioritize creating better tools or improving validation for existing ones, especially MARS.

R	Q	A
[14]	1	Record food purchase receipts for nutritional analysis, which is costly and time-consuming manually.
	2	Use crowdsourcing via the AMT platform to annotate receipts.
	3	A consensus task where multiple workers work on the same assignment, and their agreement is verified.
	4	The study shows that crowdsourcing can annotate food receipts accurately.
	5	The study used a small sample size and focused on limited data extraction.
	6	Explore machine learning to automate the process and improve crowdsourcing quality.
[15]	1	Large tasks can feel overwhelming because they often have a fixed structure.
	2	Study the pros and cons of breaking down macrotasks into microtasks for arithmetic, classification, and transcription.
	3	An experiment with 110 participants compared macrotasks to microtasks, with and without interruptions.
	4	Microtasks took longer but gave better results, a better experience, and handled interruptions well.
	5	The study only covered simple tasks, which may not apply to complex ones.
	6	Test task decomposition on complex tasks and use cognitive models to study its effects.
[16]	1	Organize and represent academic knowledge in a machine-readable format due to the growing number of publications.
	2	Develop TinyGenius, a methodology for creating and validating an academic knowledge graph with NLP and crowdsourcing..
	3	Evaluate data performance using a subset of the arXiv corpus and assess usability and label quality through user evaluation.
	4	TinyGenius shows promise for validating academic knowledge through NLP and microtasks. The triple store handles data volume well, and the system's usability is strong.
	5	Participant agreement varies for microtasks, and there may be bias in selecting popular articles for evaluation.
	6	Investigate using machine learning to enhance NLP accuracy and explore methods for handling complex instructions requiring domain knowledge.
[17]	1	Improve task allocation and quality in microtask environments, particularly in developing countries.
	2	Design and evaluate a profile-aware microtasking approach to enhance task assignment using worker profiles.
	3	Use design science research (DSR) methodology, comparative analysis, and crowdsourcing metrics to assess the solution's relevance.

R	Q	A
	4	Exploring microworker profiles can improve task design, allocation, evaluation, and quality.
	5	Stakeholder awareness of effective microtasking approaches need further research.
	6	Investigate stakeholder awareness of effective microtasking approaches.
[18]	1	Collecting meaningful data requires an integrated platform aligned with human activities.
	2	Explore mobile app designs for recommending microtasks in public spaces like public transport.
	3	Conduct field observations and surveys of public transport users' activities.
	4	Age and occupation influence activity choices during short leisure times, enabling personalization.
	6	Develop a system for a specific domain and study motivation, availability detection, and context-based strategies.
[19]	1	Lack of cognitive personalization in microtasks reduces performance and work quality.
	2	Create a model using cognitive testing and task fingerprinting for personalization.
	3	A case study tested four microtask types with and without personalization.
	4	Results showed improved accuracy and better task adaptation to worker abilities.
	5	Controlled conditions may limit the real-world applicability of the findings.
	6	Apply the model to more tasks, integrate machine learning, and study its impact on motivation.
[20]	1	Design effective collaborative problem-solving communities.
	2	Use design patterns like "Starting in a Niche" and "Chunking" to structure and manage them.
	3	Show examples of real communities applying these patterns in practice.
	4	Applying these patterns can make communities more productive.
	6	Test these patterns with case studies or experiments.
[21]	1	Microtask crowdsourcing workflows struggle with complex interface design tasks.
	2	Three experiments explore workflow design for interface design via microtask crowdsourcing.
	3	The experiments investigate task decomposition, flexibility versus consistency, and task reviews.
	4	They aim to identify workflows that address challenges in interface design tasks.
	5	The paper describes the experiments but does not provide results or conclusions.
	6	Results from the experiments will guide the creation of improved workflows.
[22]	1	Share knowledge among transient workers in microtask programming.
	2	Integrate a Q&A system into the programming environment for code-specific questions.

R	Q	A
	3	Observe 20 crowdsourcing workers using the system over 30 hours.
	4	The system helped coordinate work but had issues like delays and duplicate questions.
	5	The study's limited scope may not reflect large-scale microtask programming.
	6	Improve response speed, merge similar questions, and link decisions to code.
[24]	1	Disasters demand coordinated relief efforts.
	2	Information and Communication Technologies (ICT) helps mobilize volunteers, report crises, and map information online.
	3	Examples include the 2011 Japan tsunami and the 2010 Haiti earthquake.
	4	ICT supports crowdsourced voluntary action in crises.
	5	Risks include inaccuracies, privacy issues, and volunteer burnout.
	6	Future research must address these risks and challenges.
[25]	1	No validated or standardized tools exist to assess mobile health app quality.
	2	Review and analyze mHealth app quality tools using COSMIN guidelines.
	3	Systematic review of PubMed and Embase studies identifying tools and their validation.
	4	Most tools lack proper psychometric validation for mobile applications.
	5	Heterogeneity of tools and studies make searches and criteria selection difficult.
	6	Develop and validate transparency, privacy, and data security tools.
[26]	1	Crowdsourcing is growing in software engineering, requiring insights into its methods and potential.
	2	Examine crowdsourcing models in software development and highlight future opportunities.
	4	Crowdsourcing could transform traditional software development methods.
	6	Study how crowdsourcing models can evolve and influence software development further.
[23]	1	Respond effectively to health-related SMS messages in disaster scenarios.
	2	Develop a crowdsourcing platform to label and categorize SMS messages.
	3	Describes the system architecture but lacks specific evaluations or case studies.
	4	Aims to improve response speed and effectiveness through expert categorization.
	5	Missing details on implementation and evaluation limit assessing scalability.
	6	Test the platform in real-world scenarios, integrate it with disaster tools, and use machine learning for automation.

R	Q	A
[27]	1	Traditional software model inspection struggles with large-scale artifacts and lacks proper tool support.
	2	Introduce a Crowdsourcing-Based Software Inspection (CSI) process with tools to detect defects early in development.
	3	A controlled study comparing defect detection effectiveness between CSI and traditional pen-and-paper inspection.
	4	CSI showed promising results in defect detection performance.
	6	Explore CSI team organization, assess its impact on model quality assurance, and address confidentiality challenges.
[28]	1	Increase job opportunities for many workers.
	2	Match task assignments to workers' skills and task difficulty.
	3	A preliminary experiment showed the approach increases the number of qualified workers.
	4	The approach works, but estimating task difficulty is tough.
	5	The experiment may not reflect real-world complexities.
	6	Further research is needed to improve task difficulty estimation and expand the framework to multidimensional skills.
[29]	1	Help students improve subgoal label quality and enhance solution-planning skills.
	2	Develop AlgoSolve, a platform for peer-curated subgoal labels using microtasks.
	3	Conduct a between-subjects study comparing AlgoSolve to a baseline interface with expert feedback.
	4	AlgoSolve's learnersourcing workflow gathered high-quality labels, improving understanding of solution techniques.
	5	Limited participant number and focus on one problem-solving technique.
	6	Explore applying the approach to other techniques and adding expert explanations for feedback.

R - Reference, Q - Question, A - Answer

Appendix B

Approach Methods

Table B.1: Description of methods

Phase	Description
Research Design	A mixed-methods approach will be used, combining quantitative and qualitative methods. The quantitative aspect evaluates MARINA's usability and effectiveness through usage data analysis and surveys. The qualitative component explores participants' experiences through interviews and feedback analysis, offering a comprehensive view of the impact of the app. The study is exploratory and descriptive. Initially, it explores the content validation process through a mobile platform. Then, it describes the platform's functionality and user responses in detail.
Research Context	The study will take place in real-world conditions with MARINA users, including patients, caregivers, and healthcare professionals, both with and without prior experience with diabetes. Participants will use the app on their devices, which allows data collection in diverse environments. This ensures the relevance and applicability of the findings.
Population and sampling	The study participants will include healthcare professionals (e.g., nurses and doctors). A balanced and diverse sample will be selected, and the size will be determined based on statistical and practical considerations. Participants must be over 18, provide informed consent, own a compatible smartphone, and have a stable internet connection. Those without a compatible device, stable connection, or informed consent, as well as those with exclusion factors, will not be included.

Phase	Description
Data Collection	The tools and instruments for data collection include the following: • App Usage Data: Automatically logs user interactions, such as task completion times and feature usage frequency • Usability Surveys: Employs the validated System Usability Scale (SUS) • Content Evaluation Survey: Assesses the clarity, scientific accuracy, and completeness of MARINA's responses • Interviews/Feedback: Collects detailed insights into user experiences and satisfaction Participants will use MARINA for a defined period to complete predefined tasks. Usage data will be collected anonymously. Surveys will be administered at specific intervals, and participants will provide feedback through interviews or comments. Data collection tools will be validated. For content evaluation, experts will review a questionnaire adapted from the Suitability Assessment of Materials (SAM). The SUS has already been validated for usability studies.
Variables and Measurements	Key Variables: • Usability: Measured through SUS and user surveys • Effectiveness: Assessed through task completion times, feature usage frequency, and satisfaction survey results • Content Quality: Evaluated via the content survey and user feedback Measurements: • Quantitative methods (e.g., scales, counts) • Qualitative methods (e.g., content analysis, feedback)
Analysis Methods	Quantitative analysis will involve descriptive statistics (e.g., means, standard deviations) and inferential statistics (e.g., t-tests, ANOVA) using SPSS software. Qualitative analysis will focus on identifying recurring themes and patterns through content analysis. Finally, quantitative and qualitative data will be integrated to understand the data comprehensively.
Ethical Considerations	Participants will be informed of the study's goals, risks, and benefits and will provide voluntary, informed consent. Data will be collected and processed anonymously to protect participants' privacy, securely stored, and used only for research purposes. Participants will have the freedom to withdraw consent at any time without consequences.

Phase	**Description**
Limitations and Mitigation Plan	The sample may not represent all users of mobile health apps, limiting generalizability. Data collection tools might not capture all relevant aspects of content validation. Specific samples or environmental factors may also influence results. A diverse sample will be used to address these limitations, and results will be interpreted cautiously. Complementary instruments will be employed, and results will be integrated from multiple perspectives. A detailed context analysis will also be conducted, and the implications of findings for various settings will be discussed.

Appendix C

Informed Consent Form, Free and Clear

Este formulário de consentimento contém informação acerca de um estudo no âmbito do projeto MARINA - A Mobile App foR medIcal anNotAtion. Queremos que esteja informado sobre o objetivo deste estudo e o que a sua participação implica. O projeto iniciou a 15 de setembro de 2024 e termina 1 ano depois. As entidades envolvidas neste projeto são FEUP (Faculdade Engenharia da Universidade do Porto) e Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions - AICOS.

Objetivo do estudo

Este estudo tem como objetivo desenvolver uma aplicação móvel para suportar a validação de conteúdo de aplicações de saúde, com foco em materiais educativos relacionados à diabetes. A aplicação deve ser fácil de usar e permitir que médicos possam completar tarefas de validação de conteúdo rapidamente e de forma eficiente, através do seu telemóvel.

Materiais usados

- MARINA: um aplicativo móvel funcional desenvolvido para plataformas iOS e Android.
- Questionários: um questionário pós-estudo para recolher feedback sobre a usabilidade, utilidade e sugestões de melhoria da aplicação e um questionário para recolher medidas de validação do conteúdo educacional para auto-gestão da diabetes.

Procedimentos

Este estudo envolve a realização de uma série de tarefas no telemóvel, com registo das interações do utilizador através de uma aplicação de gravação de ecrã. Será também solicitado que responda a um questionário breve para recolher feedback qualitativo e quantitativo sobre a sua experiência na aplicação. Além disso, será pedido que valide conteúdo educacional sobre auto-gestão da diabetes utilizando tanto a aplicação móvel como métodos tradicionais. Serão recolhidos

dados sociodemográficos para caracterizar o perfil dos participantes. Este procedimento será preferencialmente presencial, no local de trabalho dos participantes ou nas instalações da Fraunhofer AICOS, mas também poderá ser realizado remotamente.

Fatores de exclusão

- Incapacidades visuais conhecidas que possam afetar a conclusão da tarefa.

Os seus dados pessoais serão analisados pelos investigadores da FEUP e Fraunhofer AICOS, e os dados de interação em formato de vídeo recolhido pela aplicação móvel destruídos no final do estudo. Os restantes dados serão armazenados numa base de dados, de forma anonimizada, para que possam ser utilizados para fins de investigação. Os dados recolhidos são confidenciais. A FEUP e a Fraunhofer AICOS tomarão todas as medidas necessárias à salvaguarda e proteção dos dados recolhidos por forma a evitar que venham a ser acedidos por terceiros não autorizados. Os seus direitos no âmbito da proteção de dados serão sempre assegurados, tendo em consideração as normas específicas para o desenvolvimento da atividade de investigação. Informações adicionais poderão ser prestadas pelo Investigador Responsável abaixo identificado. Gostaríamos de contar com a sua participação. A participação não envolve qualquer prejuízo ou dano material e não haverá lugar a qualquer pagamento. O material necessário para este estudo será apenas o seu telemóvel pessoal. A participação não terá custos para o participante. A sua participação é voluntária, podendo em qualquer altura cessá-la sem qualquer tipo de consequência. Também poderá pedir a retificação ou destruição da informação recolhida a qualquer momento, exceto dos dados anonimizados que serão integrados numa base de dados. Agradece-mos muito o seu contributo, fundamental para a nossa investigação!

O participante:
Declaro ter lido e compreendido este documento, bem como as informações verbais fornecidas e aceito participar nesta investigação. Declaro ainda não ter nenhuma das condições listadas acima que constituem fatores de exclusão deste estudo. Permito a utilização dos dados que forneço de forma voluntária, confiando que apenas serão utilizados para investigação e com as garantias de confidencialidade e anonimato que me são dadas pelo investigador. Autorizo a comunicação de dados de forma anónima a outras entidades que estabeleçam parceria com a FEUP e a Fraunhofer AICOS para fins académicos e de investigação científica, no âmbito do projeto MARINA. Autorizo a comunicação de meus dados pessoais (nome e idade) para ativação de um seguro de acidentes pessoais, exclusivamente durante a realização das atividades de resposta aos questionários e tarefas nas instalações da Fraunhofer AICOS, em cumprimento dos procedimentos internos da Fraunhofer AICOS.

Nome do participante:
Assinatura do participante:
Data:__ /__ /____

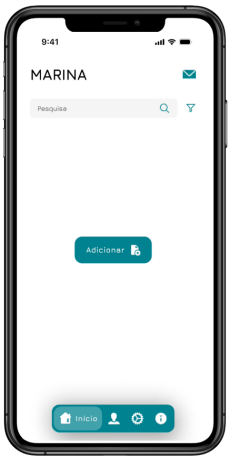
Estudante responsável pela tese:
Nome: Tiago Antunes
Telemóvel: (+351) 910 239 456
E-mail: tiagoantunes850@gmail.com
Assinatura:

Co-Orientadora:
Nome: Sílvia Rêgo
Telemóvel: (+351) 924 490 807
E-mail: silvia.rego@aicos.fraunhofer.pt
Assinatura:

Appendix D

MARINA Screenshots

D.1 Figma Prototype Screens

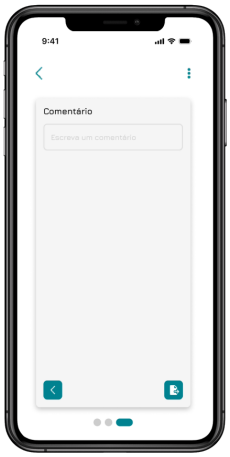


(a) Home screen without microtasks

(b) Home screen with one microtask unread

(c) Screen displaying the content of a microtask

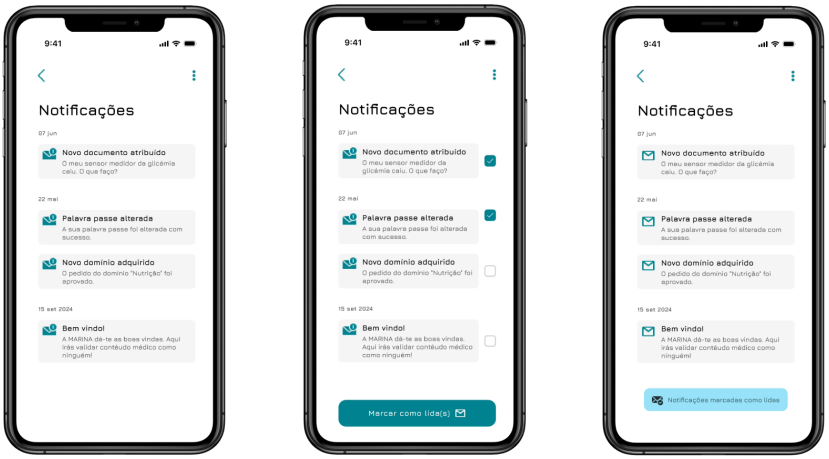
(d) Screen used for validating the microtask



(e) Screen for adding comments on the microtask

(f) Home screen with one microtask submitted

Figure D.1: Screens of Task 1



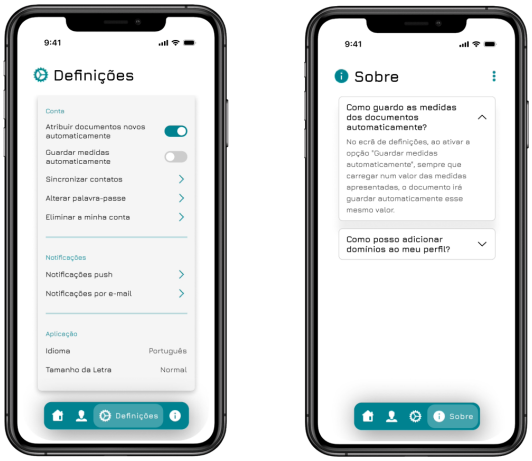
(a) Notifications screen with unread messages

(b) Notifications selection screen

(c) Notifications screen after marking messages as read

Figure D.2: Screens of Task 2

S



(a) Settings screen

(b) About screen

Figure D.3: Screens of Task 3

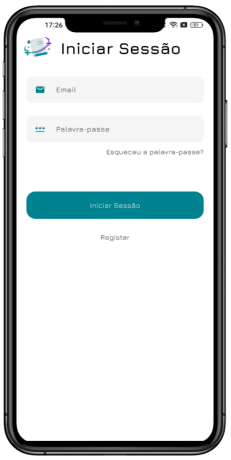
D.2 Flutter Application Screens

(a) First Onboarding page

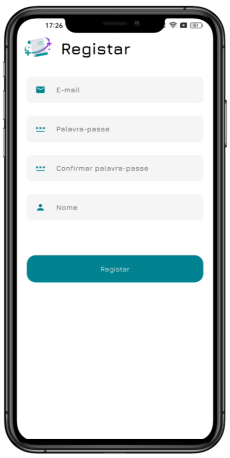
(b) Second Onboarding page

(c) Third Onboarding page

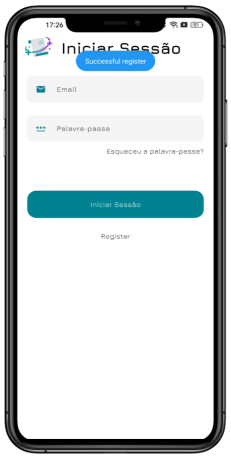
Figure D.4: Onboarding screen



(a) Sign In screen

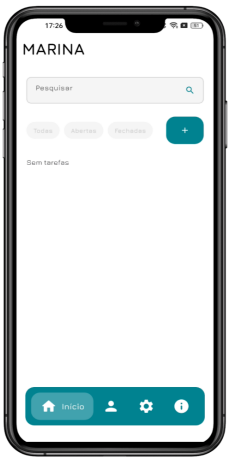


(b) Sign Up screen



(c) Sign-in screen displayed after completing registration

Figure D.5: Sign In and Sign Up screens



(a) Home screen without microtasks

(b) Home screen with multiple microtasks

(c) Home screen displayed after a microtask is submitted

Figure D.6: Home screen

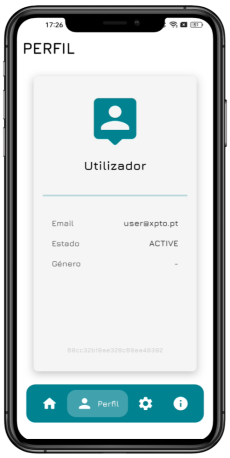


Figure D.7: Profile Screen

(a) Screen displaying the content of a microtask

(b) Screen used for validating the microtask (empty values)

(c) Screen used for validating the microtask (values filled)

(d) Screen for adding comments on the microtask

(e) Feedback displayed when attempting to submit a microtask with empty fields

Figure D.8: Microtask Screens

Appendix E

Detailed Results: Usability Test and Pilot Study

E.1 Tables

Table E.1: Mean scores from usability questionnaire (Likert 1–5)

Q	Question	Mean
1	Needed to learn a lot first	2.3
2	Felt confident using	4.3
3	Very confusing	1.1
4	Most people would learn quickly	4.6
5	Too inconsistent	1.3
6	Functionalities well integrated	4.0
7	Need technical support	2.0
8	Easy to use	4.6
9	Unnecessarily complex	1.6
10	Use app frequently	4.6

Table E.2: Average task completion times (minutes)

Task	Average
1	1,03
2	1,25
3	0,92

Table E.3: Mean scores from content validation questionnaire (Likert 1–5)

	Mean
Content easy to understand	4.0
Scientific rigor adequate	4.0
Content complete and comprehensive	4.5
Microtasks improved efficiency	5.0
Integrated questionnaire clear	5.0

E.2 Figures

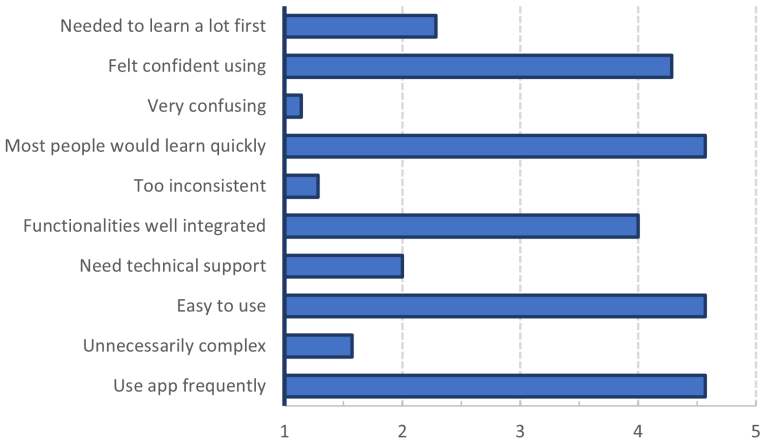


Figure E.1: Main application domains

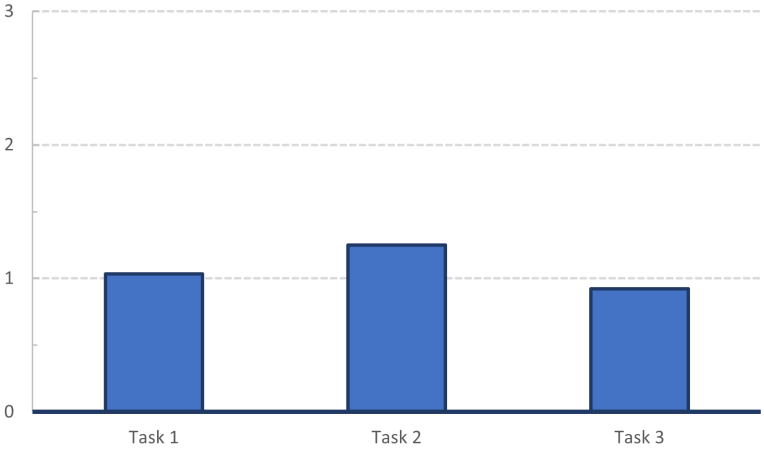


Figure E.2: Average Task Completion Times

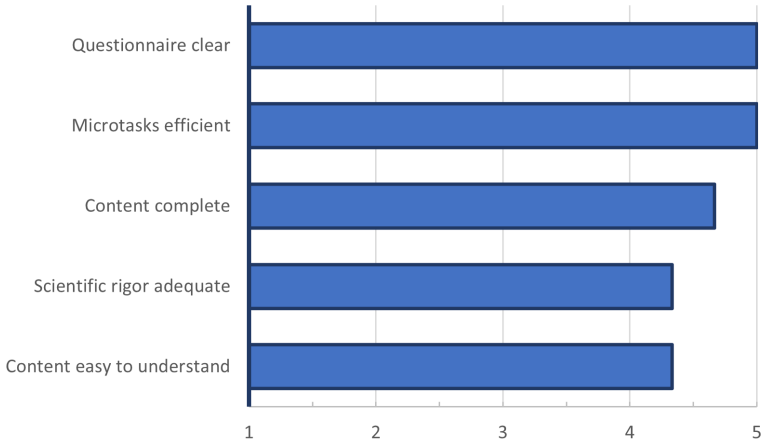


Figure E.3: Pilot questionnaire results